Université de Lille
LEM — Lille Économie Management — CNRS 9221
École Doctorale SÉSAM

Innovations basées sur les données dans le secteur des services financiers

Data-driven Innovations in the Financial Services Sector

Stephanie Beyer Díaz

Thèse en vue de l'obtention du titre de docteure en sciences économiques, écrit sous la direction du Prof. Dr. Kristof Coussement et du Prof. Arno De Caigny.

Soutenu le 14 juin 2024.

Membres du Jury:

Directeurs de thèse:
Dr. Kristof Coussement, Professor, IÉSEG School of Management
Dr. Arno De Caigny, Professor, IÉSEG School of Management

Président du Jury:
Dr. Dominique Crié, Professor, Université de Lille

Rapporteurs:
Dr. Wouter Verbeke, Professor, KU Leuven
Dr. Vera Miguéis, Professor, Universidade do Porto

Examinateur:
Dr. Mathias Kraus, Professor, FAU Erlangen-Nürnberg School
Dr. Carla Vairetti, Professor, Universidad de los Andes

L'université de Lille n'entend donner aucune approbation ni improbation aux opinions émises dans cette thèse. Ces opinions doivent être considérées comme propres à leur auteur.

LABORATOIRE DE RATTACHEMENT :
Lille Economie Management (LEM - UMR CNRS 9221), Laboratoire de recherche rattaché à l'IAE de Lille et à la Fédération Universitaire Polytechnique de Lille (FUPL).
Préparation de la thèse sur le site de l'IESEG School of Management, 3 Rue de la Digue, 59000 Lille.

Acknowledgments

I had heard before that it takes a village to finish a PhD, which I initially, and mistakenly, dismissed as a dad joke. Now, I find myself with a long list of people and organizations to thank. These names have contributed in different ways to my progress, both directly and indirectly. Without their support, I would not have been able to complete this project.

The jury and examiners also have my thanks for reviewing my doctoral thesis, for providing insights and feedback, encouraging me to improve my research output. I am also grateful to my thesis director Kristof Coussement and co-director Arno De Caigny for the freedom they allowed me in pursuing my interest in Deep Learning. To the co-authors of my first paper, Professor Stefan Creemers and Luis Fernando Pérez Armas, thank you both for your excellent comments and helpful disposition.

Moreover, I thank Crédit Agricole Nord de France, l'Université Catholique de Lille, and IÉSEG School of Management for granting me the opportunity to be the Chair of Big Data and CRM marketing. This experience has also allowed me to present my work at international conferences, alongside experienced academics whose works I have read and which I greatly admire. I am truly grateful for this enriching experience, both in a professional and personal sense.

I further highlight Crédit Agricole Nord de France for providing the necessary funding and access to their vasts amount of data to work on relevant, applicable and interesting projects. Angelo Caria et Aurelie Duba, je vous remercie vivement de m'avoir accordé la liberté d'explorer différentes bases de données et de vous proposer des projets adaptés à mes intérêts de recherche. Je me permets également de remercier Angelo encore une fois pour nos conversations enrichissantes, le partage de son temps et de ses connaissances approfondies en matière de données, ainsi que pour son accueil et son soutien constants face à mes questions ou problèmes. J'ai grandement apprécié notre collaboration.

Further, a big thanks to KU Leuven and Bart Baesens for kindly allowing me to participate in a highly productive feedback session with Professor María Óskarsdóttir, to whom I remain indebted for her invaluable feedback. This played a significant role in improving the overall storyline, as well as strengthening the experimental setting and conclusions from my research.

I am grateful for the support and camaraderie I have received from my fantastic colleagues Philipp Borchert, Khaoula Idbenjra, Emil Guliyev, Juliana Sánchez Ramírez, and Minh Phan. I would also like to thank Edson Souza and Umur Cengiz for their cheerful disposition and generosity. Best of luck to all of you with your future endeavors!

To my sister, my dad, my brother, and my mom (listed in alphabetical order by first name, to avoid any complaints): you have my thanks for lovingly listening to my interests, my successes, and my failures. I treasure the moments we shared over video calls, the "chismes", and the snacks sent overseas. 😄

Steph

Doctoral Committee

Directeurs de thèse:

Dr. Kristof Coussement, Professor, IÉSEG School of Management

Dr. Arno De Caigny, Professor, IÉSEG School of Management

Président du Jury:

Dr. Dominique Crié, Professor, Université de Lille

Rapporteurs:

Dr. Wouter Verbeke, Professor, KU Leuven

Dr. Vera Miguéis, Professor, Universidade do Porto

Examinateurs:

Dr. Mathias Kraus, Professor, FAU Erlangen-Nürnberg School

Dr. Carla Vairetti, Professor, Universidad de los Andes

List of Publications

This dissertation is based on three different studies:

1. **Stephanie Beyer Díaz**, Kristof Coussement, Arno De Caigny, Luis Fernando Pérez & Stefan Creemers. Do the US president's tweets better predict oil prices? An empirical examination using long short-term memory networks. *International Journal of Production Research*. Received 14 November 2022, Accepted 28 April 2023, Published online: 01 June 2023. Preliminary results were also presented at:
    - 6th Analytics for Management and Economics Conference (2020), held online.
    - 51st Annual Conference of the Decision Sciences Institute (2020), held online.

2. **Stephanie Beyer Díaz**, Kristof Coussement, Arno De Caigny. Improved Decision-Making Through Life Event Prediction: A Case Study in the Financial Services Industry. Working paper submitted to *Decision Support Systems*. Preliminary results were also presented at:
    - 8th Analytics for Management and Economics Conference (2021), held online.
    - 32nd European Conference on Operational Research (2022), Espoo, Finland.

3. **Stephanie Beyer Díaz**, Kristof Coussement, Arno De Caigny. Longitudinal Data for Recommender Systems in the Financial Services Industry. Working paper submitted to *European Journal of Operational Research*. Preliminary results were also presented at:
    - 52nd Annual Conference of the Decision Sciences Institute (2021), held online.
    - 31st European Conference on Operational Research (2021), held online.

Table of Contents

1 Introduction **3**
- 1.1 Résumé . . . 3
  - 1.1.1 Résumé Général . . . 3
  - 1.1.2 Résumé Détaillé . . . 8
- 1.2 Abstract . . . 15
  - 1.2.1 General Abstract . . . 15
  - 1.2.2 Detailed Abstract . . . 19
- 1.3 References . . . 24

2 Do the US President's Tweets Better Predict Oil Prices? An Empirical Examination Using Long Short-Term Memory Networks **29**
- 2.1 Introduction . . . 30
- 2.2 Related Work . . . 32
- 2.3 Experimental Setup . . . 34
  - 2.3.1 Data . . . 34
  - 2.3.2 NLP techniques . . . 35
  - 2.3.3 LSTM . . . 36
  - 2.3.4 Benchmark Models . . . 36
  - 2.3.5 Evaluation Metrics and Statistical Tests . . . 36
- 2.4 Results . . . 38
- 2.5 Discussion . . . 40
- 2.6 Conclusions and Future Research . . . 42
- 2.7 References . . . 43
- 2.8 Appendix . . . 48
  - 2.8.1 Appendix A. Methodology . . . 48

3 Improved Decision-Making Through Life Event Prediction: A Case Study in the Financial Services Industry **59**
- 3.1 Introduction . . . 60
- 3.2 Related Research . . . 62
- 3.3 Methodology . . . 65
  - 3.3.1 Long Short-Term Memory . . . 65
  - 3.3.2 Featurization for Longitudinal Data . . . 66
  - 3.3.3 Integrated Gradients . . . 67
- 3.4 Experimental Setup . . . 69
  - 3.4.1 Data & Data Preprocessing . . . 69
  - 3.4.2 LSTM Architecture . . . 72
  - 3.4.3 Hyperparameter Tuning . . . 73
- 3.5 Results . . . 74
- 3.6 Discussion . . . 79
- 3.7 Conclusions and Further Research . . . 80
- 3.8 References . . . 82
- 3.9 Appendix . . . 86
  - 3.9.1 Additional Experimental Results . . . 86

4 Longitudinal Data for Recommender Systems in the Financial Services Industry 92
4.1 Introduction . . . 93
4.2 Related Work . . . 95
4.2.1 Recommendations for financial services . . . 95
4.2.2 Longitudinal data for recommendations . . . 96
4.3 Methodology . . . 98
4.3.1 Data . . . 98
4.3.2 Featurization . . . 100
4.3.3 Multi-label Classification Techniques . . . 101
4.3.4 Traditional RS . . . 102
4.3.5 SHAP for Interpretable Results . . . 103
4.4 Experimental Setup . . . 104
4.5 Results . . . 104
4.5.1 Does the use of various RS methods lead to significantly distinct levels of performance? . . . 104
4.5.2 Does the utilization of various featurization methods for longitudinal data yield significantly different levels of performance? . . . 105
4.5.3 Does the incorporation of longitudinal data using state-of-the-art DL models significantly outperform other recommendation approaches? . . . 108
4.5.4 Can relevant features be identified for product recommendations in the financial services industry? . . . 109
4.6 Conclusions and Future Research . . . 111
4.7 References . . . 112
4.8 Appendix . . . 115
4.8.1 Appendix A. Additional Data . . . 115
4.8.2 Appendix B. Additional Results . . . 115

5 Conclusions 122
5.1 Conclusions . . . 122
5.1.1 Conclusions générales . . . 122
5.1.2 Limitations et perspectives de recherche future . . . 123
5.2 Conclusions . . . 125
5.2.1 General conclusions . . . 125
5.2.2 Limitations and future research . . . 126

List of Figures 129

List of Tables 130

Chapter 1: Introduction

CHAPTER 1

Introduction

1.1 Résumé

1.1.1 Résumé Général

Parmi l'intensification de la concurrence, avoir des relations solides avec les clients est considéré comme un facteur clé pour maintenir un avantage concurrentiel. Ainsi, les entreprises ont investi de manière significative dans le développement de stratégies de gestion de la relation client (CRM) ces dernières années (Dalla Pozza et al., 2018). Théoriquement, le CRM se compose de plusieurs initiatives différentes, chacune pouvant être classée en quatre dimensions distinctes (Kumar and Reinartz, 2006) : (i) l'alignement organisationnel, (ii) la gestion des clients, (iii) la technologie et (iv) la mise en œuvre de la stratégie CRM. La première dimension, l'alignement organisationnel, se réfère à la refonte et à l'alignement des processus existants, avec l'objectif ultime de placer les clients au centre (Dalla Pozza et al., 2018). La deuxième dimension, la gestion des clients, fonctionne selon le principe de traiter les clients de manière différenciée, en s'adaptant à leurs besoins, préférences et priorités (Reinartz et al., 2004). La troisième dimension, la technologie, englobe le degré auquel les applications CRM analytiques, opérationnelles et collaboratives sont mises en œuvre pour collecter des informations sur les clients à travers les points de contact et pour faciliter la diffusion et l'analyse des informations (Dalla Pozza et al., 2018). Enfin, la dimension de mise en œuvre de la stratégie CRM nécessite une approche clairement orientée vers le client, incluant le soutien de la direction, des métriques de performance orientées client et une vision globale du client à travers toute l'organisation (Palmatier et al., 2007).

Une mise en œuvre réussie des stratégies CRM aborde simultanément les quatre dimensions. Cependant, cette thèse se concentre fortement sur la dimension technologique. Cette dimension offre l'opportunité d'exploiter des méthodes innovantes pour mieux comprendre le secteur des services financiers, les facteurs qui l'affectent, et des insights utiles pour ses décideurs. En fin de compte, l'utilisation d'outils tels que des technologies innovantes et de nouvelles sources de données est proposée comme moyen d'obtenir une compréhension plus approfondie du contexte dans lequel les prestataires de services financiers interagissent avec leurs clients.

La figure 1.1 présente une illustration schématique des insights basés sur les données qui peuvent être utilisés pour une meilleure compréhension du contexte dans lequel les prestataires de services financiers interagissent avec leurs clients. Plus précisément, les insights provenant de facteurs externes peuvent être utiles pour comprendre les fluctuations du marché et les conditions économiques qui peuvent affecter à la fois les priorités des clients et les opportunités de croissance d'un prestataire de services financiers. Les insights provenant des prestataires eux-mêmes, représentés par l'anneau central, se réfèrent aux analyses des produits, services et offres disponibles, ainsi que des caractéristiques qui peuvent différencier un prestataire de ses concurrents, telles que les canaux de communication à la disposition de leurs clients, la qualité des services ou l'emplacement des différentes agences, entre autres. Enfin, les insights provenant des clients incluent des

informations sur leur comportement, leurs caractéristiques démographiques et l'intensité de leur relation avec les prestataires, entre autres. Ainsi, ces perspectives de insights sont modélisées comme des anneaux concentriques en raison de leur interdépendance. Dans cette thèse, nous explorons des applications pour tous les différents insights basés sur les données afin de mieux comprendre le secteur financier et de développer des applications innovantes qui peuvent être exploitées pour continuer à renforcer les trois autres dimensions du CRM.

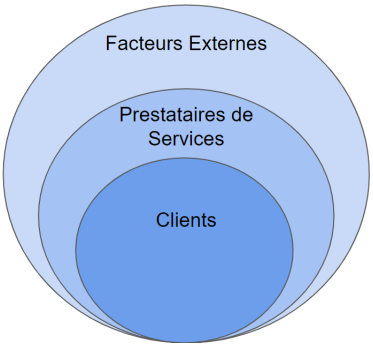


Figure 1.1. Illustration schématique des insights basés sur les données dans le secteur des services financiers

La motivation derrière l'exploration des différentes sources de données est liée aux avancées spectaculaires survenues ces dernières années dans le secteur technologique. Ces développements incluent la capacité de traiter de grands volumes de données grâce à des modèles améliorés pour les tâches prédictives, tels que le deep learning (DL), ce qui offre l'opportunité d'explorer des sources de données nouvelles et complexes.

Pour explorer comment l'intérêt de la recherche autour de ces technologies a évolué, nous examinons le domaine de la recherche sur les systèmes de recommandation, choisi pour être un exemple populaire de publications liées au CRM, avec de fortes implications managériales pour le secteur des services financiers. En effet, le nombre d'articles sur les systèmes de recommandation est bien plus élevé que pour d'autres tâches prédictives dans le CRM, telles que le churn, la segmentation des clients ou la prévision des ventes.

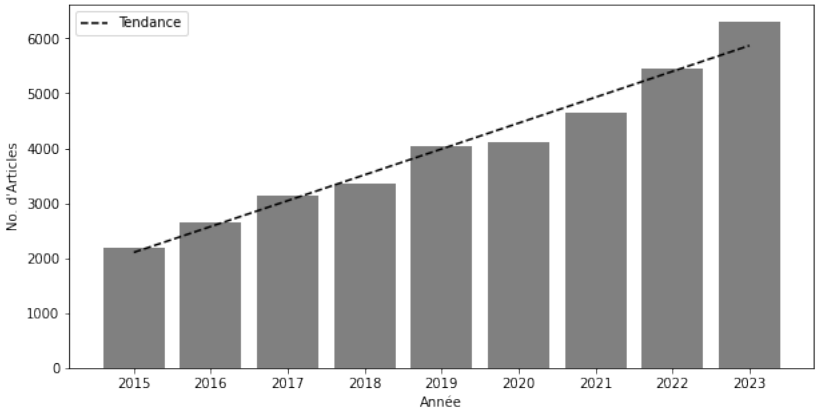


Figure 1.2. Nombre d'articles de recherche sur les systèmes de recommandation publiées par année

En utilisant Scopus pour entrer les mots-clés « système de recommandation », « recommender system », « collaborative filtering », « content-based recommender » ou « content-based recommendation », un total de 38 171 articles et communications de conférences sont trouvés en anglais, publiés entre 2015 et 2023. L'intérêt pour les systèmes

de recommandation n'a cessé d'augmenter, avec de plus en plus d'articles publiés chaque année, comme illustré dans la Figure 1.2.

Ces documents peuvent être segmentés en deux groupes : les études utilisant le deep learning (DL) et celles se concentrant sur les systèmes de recommandation traditionnels. Le premier groupe, DL, inclut des publications trouvées grâce à des mots-clés tirés de Goodfellow et al. (2016), tels que « deep learning », « neural networks », « MLP », « CNN », « LSTM », « RNN », « GRU », « GNN », « Boltzmann », « autoencoders », « transformers » et « representation learning », entre autres. Le second groupe exclut les publications du premier groupe ainsi que celles utilisant des approches hybrides. De ces deux groupes, plusieurs conclusions peuvent être tirées. Premièrement, les études utilisant les systèmes traditionnels de recommandation (TR) restent très populaires et leur volume continue d'augmenter. Deuxièmement, l'intérêt pour les nouvelles technologies a crû à un rythme accéléré, comme le montre la Figure 1.3. En fait, 2023 est la première année où la recherche sur le DL appliqué aux systèmes de recommandation a dépassé l'utilisation des TR.

Ces changements soulignent la nécessité pour les entreprises de se tenir à jour avec l'évolution du paysage pour maintenir un avantage concurrentiel soutenu. Ainsi, cette thèse contribue aux applications technologiques innovantes dans le secteur des services financiers, en utilisant de nouvelles sources de données, en déployant des méthodologies de pointe pour la modélisation prédictive, et en mettant en œuvre des techniques avancées d'interprétabilité, avec un accent sur la fourniture d'insights quantitatifs pour les décideurs.

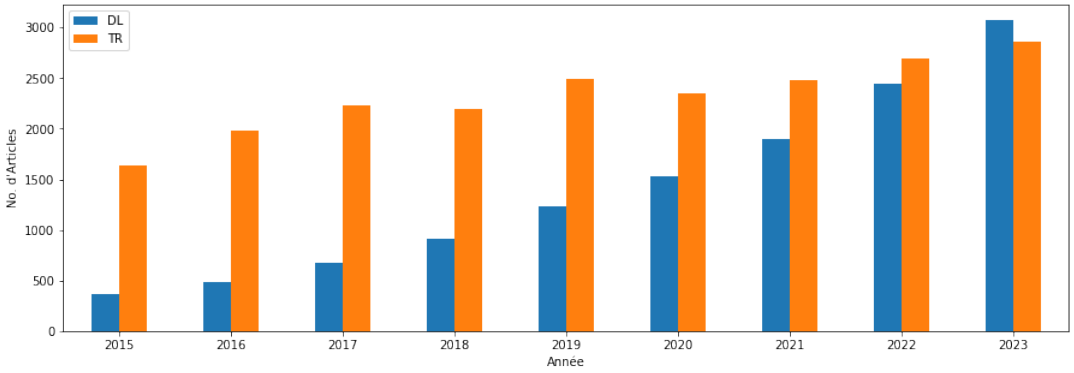


Figure 1.3. Nombre d'articles de recherche utilisant des algorithmes DL et CF par année

La modélisation prédictive dans cette thèse se concentre sur l'utilisation du deep learning (DL), souvent critiqué pour son opacité ou son statut de « boîte noire ». Notamment, la recherche a attiré l'attention sur le fait que les modèles de boîte noire causent déjà des problèmes dans des industries sensibles, telles que la santé et la justice pénale (Rudin, 2019). Cependant, beaucoup de ces problèmes, notamment le manque de transparence et de responsabilité des modèles prédictifs, sont contrôlés en Europe grâce à des réglementations telles que le RGPD. Par exemple, l'article 15 du RGPD garantit la transparence et la responsabilité des organisations en permettant aux individus de demander l'accès à leurs données personnelles. Cette disposition permet également aux individus d'accéder aux informations concernant les décisions automatisées, telles que le profilage ou le ciblage. Ces informations doivent également inclure des explications sur la logique sous-jacente et l'impact de ces traitements de données sur l'individu. Ainsi, les industries plus surveillées, telles que le secteur des services financiers, peuvent être plus hésitantes à adopter de tels

modèles de boîte noire.

Néanmoins, il est toujours utile de poursuivre la recherche sur ces modèles pour obtenir une compréhension plus complète de leurs avantages et inconvénients. Par exemple, les modèles de boîte noire montrent souvent une performance supérieure, ce qui sert de référence précieuse pour que les modèles de boîte blanche les surpassent chaque fois que possible. De plus, la recherche sur l'explicabilité a progressé depuis la publication de Rudin (2019), avec un exemple étant la méthode neuron Shapley, qui améliore la transparence et la granularité dans l'analyse DL (Ghorbani and Zou, 2020). Par conséquent, les modèles de boîte noire dépassent fréquemment les modèles de boîte blanche dans certaines applications, et les techniques d'explicabilité continuent d'être affinées pour les modèles plus opaques. En essence, les modèles de boîte noire sont encore en développement, justifiant des efforts de recherche continus pour comprendre la valeur qu'ils peuvent ajouter à différentes applications, que ce soit comme références pour les modèles de boîte blanche ou pour découvrir des insights supplémentaires basés sur les données.

Ces insights sont particulièrement importants dans le secteur des services financiers, car il est très difficile d'attirer de nouveaux clients (Knott et al., 2002). En même temps, la fidélité des clients existants fluctue en raison de facteurs externes impactant leurs préférences de consommation et priorités, tels que les événements géopolitiques, les tendances du marché ou les changements dans leur vie (Mathur et al., 2008). Néanmoins, les entreprises peuvent atténuer l'effet de ces facteurs externes en comprenant les préoccupations et intérêts de leurs propres clients, puis en prenant des actions marketing appropriées (Jackson, 1985).

Une compréhension approfondie du comportement des clients peut aider une entreprise à identifier plusieurs problèmes qui détermineront les décisions d'achat futures. Cela peut aider à comprendre si un client priorisera des préoccupations plus immédiates, comme un prêt à court terme, ou des enjeux à long terme, comme passer d'un produit à un autre au fil du temps (Jackson, 1985). En raison des coûts de changement plus élevés pour les clients qui priorisent les enjeux à long terme (Jackson, 1985), les facteurs externes auront un impact moindre. Ainsi, la dimension temporelle est essentielle pour une compréhension holistique à la fois du comportement des clients et de l'impact potentiel des actions marketing, permettant aux marketeurs de concevoir des adaptations pour les clients le long du spectre des préoccupations à court et à long terme (Jackson, 1985).

Par conséquent, la compréhension des clients et de leurs préférences dans le temps permet une amélioration constante des relations client-entreprise (Roos and Gustafsson, 2007). Cela peut être réalisé en surveillant et en analysant le comportement des clients et en concevant des actions marketing spécialisées (Sin et al., 2005), permettant finalement une allocation des ressources plus efficace en ciblant les clients les plus pertinents (Sin et al., 2005).

Dans ce contexte, cette thèse explore différentes avenues de recherche appliquée, exploitant des données séquentielles variant dans le temps pour différentes tâches : prédiction du prix du pétrole, prédiction d'événements de vie et systèmes de recommandation. Trois contributions clés sont faites : incorporer de nouvelles sources de données séquentielles pour améliorer la performance prédictive, appliquer des méthodologies de pointe pour optimiser l'utilisation des données séquentielles et déployer des techniques d'explicabilité pour explorer comment les données séquentielles contribuent à une tâche prédictive. Chacun des trois chapitres à venir contient une étude différente utilisant des données séquentielles, tandis que le dernier chapitre conclut avec des conclusions générales

et les limites de la recherche.

Les données séquentielles se réfèrent à des informations ordonnées qui aident un modèle à détecter des motifs, pour résoudre des tâches de différentes complexités. Plusieurs des modèles précédemment listés comme mots-clés pour la Figure 1.7 ont été utilisés pour apprendre à partir des données séquentielles. On peut dire que les données séquentielles ont percé grâce aux grands modèles de langage tels que ChatGPT. Ici, des séquences de mots et de phrases sont utilisées pour entraîner de grands modèles, basés sur des architectures DL, afin de trouver des solutions à différents problèmes liés au texte. Néanmoins, les données séquentielles existent dans plusieurs domaines, avec quelques exemples illustrés dans la Figure 1.8. Par exemple, une séquence d'ADN peut être utilisée pour prédire l'expression des gènes (Žiga Avsec et al., 2021). De même, une séquence d'actifs financiers mensuels peut être utilisée pour prédire des actifs futurs, le churn, le défaut de paiement de prêts ou d'autres tâches liées à la classification des clients. Les données séquentielles peuvent donc être utilisées par des modèles tels que DL pour des tâches prédictives pertinentes dans plusieurs industries, pour des problèmes de classification et de régression. Cette thèse se concentre spécifiquement sur l'utilisation de données chronologiquement ordonnées provenant de l'industrie financière pour la prédiction.

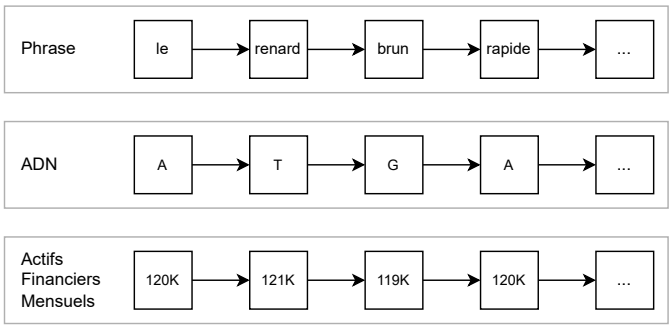


Figure 1.4. Exemples de données séquentielles

Le chapitre 2 utilise les prix du pétrole antérieurs et les données textuelles comme entrée séquentielle pour prédire les prix du pétrole à venir. Le chapitre 3 utilise des données démographiques et comportementales séquentielles des clients, fournies par un prestataire de services financiers, pour prédire 10 différents événements de la vie. Le chapitre 4 utilise des données démographiques des clients et l'historique d'achats sous la forme de caractéristiques RFM séquentielles.

En résumé, les trois études présentent des résultats dérivés de données réelles, incorporant de nouvelles sources de données séquentielles, mettant en œuvre des méthodologies de pointe pour évaluer l'utilisation des données séquentielles susmentionnées et déployant des techniques d'explicabilité pour analyser davantage les résultats. Chaque étude met en avant une valeur académique et une pertinence commerciale notable, à travers les trois contributions précédemment énumérées.

Mots clés : Big data, gestion de la relation client, analyse marketing, machine learning, réseau de neurones, analyse textuelle, prix du pétrol, prédiction des moment de vie, moteur de recommandation.

1.1.2 Résumé Détaillé

Cette section résume les trois articles contenus dans cette thèse, en examinant comment ils abordent les trois contributions générales précédemment décrites comme le fil conducteur de cette thèse. Ces contributions sont : (i) l'incorporation de nouvelles sources de données, (ii) l'évaluation de l'utilisation des données séquentielles, et (iii) le déploiement de techniques d'explicabilité. Ces contributions sont en phase avec les défis posés par le contexte actuel, faisant des trois études des exemples utiles d'approches innovantes pour améliorer le processus de prise de décision dans l'industrie financière. Ainsi, cette section contextualise davantage ces études à travers les trois contributions générales et l'état actuel de la recherche et de la littérature existante.

Cette thèse a été développée dans un contexte où les modèles basés sur l'apprentissage profond (DL) et les recherches connexes progressent à une vitesse incroyable. Par exemple, ChatGPT change continuellement le paysage de l'analyse des données, car il a à la fois accéléré la recherche autour des applications DL et multiplié les questions soulevées par les régulateurs et le grand public (Cauffman and Goanta, 2021). Cette surveillance ne se limite pas à ChatGPT et s'applique également aux algorithmes DL en général. Cela souligne la pertinence commerciale de l'équilibre entre la performance du modèle et l'explicabilité, car des études existantes montrent que la perception de la sécurité, de la vie privée et de la compatibilité avec les valeurs des consommateurs peut influencer la décision d'adopter des applications et services innovants, en particulier dans l'industrie financière (Luo et al., 2010; Hoehle et al., 2012).

Il est donc essentiel de développer des applications explicables, quel que soit le modèle sélectionné, pour garantir la transparence des résultats du modèle. Il est également important de permettre à toutes les parties prenantes de comprendre rapidement les informations fournies par un modèle, tout en permettant aux chercheurs de vérifier la valeur des sources de données complexes. En tant que telles, ces questions deviennent une partie intégrante des stratégies globales de gestion de la relation client (CRM) et de la recherche, dans le cadre d'un cadre d'analyse responsable (De Bock et al., 2023).

Nous postulons que les techniques d'explicabilité, combinées à l'utilisation de nouvelles sources de données séquentielles, sont des outils puissants pour améliorer la transparence des modèles innovants. De plus, ces techniques permettent une compréhension multidimensionnelle du comportement des clients, ajoutant un facteur de variation temporelle à l'analyse avec les mêmes données déjà utilisées dans les applications CRM existantes (De Caigny et al., 2020; Bogaert et al., 2019). Cela signifie que des sources de données supplémentaires, qui pourraient être perçues comme plus intrusives en termes de vie privée, ne sont pas nécessaires pour améliorer encore les performances. Au lieu de cela, les données historiques peuvent être utilisées séquentiellement, permettant une compréhension plus détaillée du comportement des clients. Cette approche respecte les directives de confidentialité et les cadres réglementaires, sans sacrifier l'innovation ou les performances. Ce contexte est exploré plus en détail en décrivant comment chaque étude s'aligne avec les principales contributions de cette thèse.

Incorporation de nouvelles sources de données

Tout au long de cette thèse, nous résolvons différents problèmes grâce à l'utilisation de nouvelles sources de données. En particulier, cela est réalisé de deux manières différentes. Premièrement, en utilisant des données comme une séquence chronologiquement ordon-

née, à comparer avec des données transversales, ce qui reste une zone sous-étudiée de la littérature prédictive pour les applications CRM (Óskarsdóttir et al., 2018). Deuxièmement, en incorporant des sources de données qui n'ont pas été évaluées auparavant dans la littérature pertinente.

En particulier, le chapitre 2 intègre une nouvelle source de données textuelles, utilisée séquentiellement comme entrée pour prédire les prix du pétrole à venir, ce qui aborde les perspectives concernant les facteurs externes, comme le montre la Figure 1.1. Les recherches utilisant des données textuelles pour la prédiction des prix du pétrole restent rares (Beyer Díaz et al., 2023), bien qu'elles offrent encore des opportunités pour la recherche prédictive dans le secteur financier (Huang et al., 2020). De plus, la source spécifique des données textuelles utilisée, les tweets publiés par Donald Trump pendant son mandat présidentiel, n'avait pas été étudiée auparavant (Beyer Díaz et al., 2023). En outre, cette source de données est à la fois nouvelle et pertinente, car le prix du pétrole est influencé par les événements géopolitiques mondiaux (Monge et al., 2017), la politique et le sentiment du marché (Alvarez-Ramirez et al., 2003), et les annonces publiques (Singleton, 2014). Enfin, plusieurs études dans le secteur financier prouvent que l'utilisation de données textuelles pour les prévisions est bénéfique (Kraus and Feuerriegel, 2017).

Le chapitre 3 utilise des données démographiques et comportementales des clients fournies par un prestataire de services financiers, pour prédire 10 différents événements de la vie pour les applications de gestion de la relation client (CRM). Les données originales peuvent donc être comprises comme contribuant à l'anneau intérieur des perspectives basées sur les données de la Figure 1.1. Ces données sont longitudinales, mais elles sont transformées par agrégation en données transversales pour être comparées à l'utilisation de données séquentielles comme entrée. La recherche sur les événements de la vie s'est principalement concentrée sur des données transversales (De Caigny et al., 2020), ce qui place l'utilisation de données séquentielles comme une nouveauté. De plus, les recherches existantes pour les applications CRM soulignent l'importance des données longitudinales des clients pour capturer le comportement dynamique (Óskarsdóttir et al., 2018) et garantir la validité des résultats dans le temps (Boulding et al., 2005), positionnant cette étude comme une contribution précieuse à la littérature sur les événements de la vie. Enfin, cette étude propose également de nouveaux événements de la vie à prédire, à savoir l'achat de résidence principale, l'achat de résidence secondaire et l'achat de résidence en location, ce qui signifie que les étiquettes cibles comprennent également des données nouvelles.

Le chapitre 4 analyse les données démographiques des clients et l'historique des achats, représentées sous forme de caractéristiques RFM longitudinales. Ces données, provenant du même prestataire de services financiers que dans le chapitre 3, sont utilisées pour fournir des perspectives basées sur les données à la fois à partir des anneaux intermédiaires et intérieurs de la Figure 1.1. Les données sont utilisées pour développer divers systèmes de recommandation, y compris des systèmes de recommandation traditionnels (RS), des classificateurs multi-étiquettes basés sur l'apprentissage automatique (MLC) et des modèles d'apprentissage profond (DL). Bien que les RS et MLC capturent le comportement des clients à travers des données transversales, négligeant les changements temporels (You et al., 2019), les modèles DL offrent une alternative en exploitant les données séquentielles. Des applications réussies des modèles DL pour modéliser les actions en ligne pour générer des recommandations ont été notées (Tan et al., 2016; You et al., 2019). Le comportement des clients financiers peut également être modélisé comme une

série d'actions, qui mènent finalement à un achat (Prinzie and Van den Poel, 2006). La possession de produits passés et l'ordre d'acquisition des produits sont des prédicteurs fiables pour les achats futurs dans le secteur des services financiers (Kamakura et al., 1991). De plus, les caractéristiques RFM sont pertinentes pour générer des recommandations (Bogaert et al., 2019) et excellent dans la capture des comportements d'achat (Chen et al., 2016). Néanmoins, peu de recherches existent sur l'utilisation des données séquentielles dans cette industrie, par rapport aux biens durables ou aux plateformes en ligne (Prinzie and Van den Poel, 2006). De plus, les caractéristiques RFM évoluent au fil du temps et sont naturellement dynamiques, mais leur utilisation comme entrée séquentielle pour des tâches prédictives reste une zone de recherche peu développée (Mena et al., 2023). Par conséquent, l'utilisation des entrées séquentielles pour les recommandations reste une voie de recherche précieuse, car elle constitue une approche de données novatrice dans le secteur des services financiers.

De plus, les chapitres 3 et 4 impliquent des données client sous forme séquentielle pour les applications CRM, tout en fournissant des perspectives innovantes et exploitables pour la stratégie de l'entreprise. Une tâche clé dans l'industrie des services financiers est la ciblage des clients pertinents (Geuens et al., 2018), car les clients existants sont plus rentables que les nouveaux (Knott et al., 2002). Cela rend l'utilisation de nouvelles sources de données très importante car elle permet aux praticiens d'aller au-delà du profil comportemental factuel et de l'historique des clients, pour développer des stratégies qui font sentir au client qu'il est écouté (Crié and Micheaux, 2006).

En résumé, les trois études contenues dans cette thèse utilisent des sources de données nouvelles et des approches pour leurs tâches prédictives respectives. En outre, toutes les études exploitent les données séquentielles comme moyen d'améliorer les performances, ce qui reste une approche peu étudiée dans l'industrie financière. Enfin, les nouvelles sources de données sont un élément essentiel de la recherche, car elles permettent aux praticiens de tirer de nouvelles informations du comportement de leurs clients pour construire des stratégies CRM innovantes.

Évaluation de l'utilisation des données séquentielles

Au cours des dernières années, les données séquentielles ont été exploitées pour réaliser d'incroyables applications. En particulier, ChatGPT est souvent mentionné comme un exemple réussi de formation d'un modèle sur des données séquentielles. Son architecture est basée sur une forme de réseaux neuronaux (Brown et al., 2020), avec la capacité de traiter des données séquentielles et de capturer des dépendances à long terme (Vaswani et al., 2017). Le succès de ChatGPT a inévitablement suscité l'intérêt pour l'exploration et l'adaptation d'autres formes de données séquentielles pour différents domaines, étendant leur utilité au-delà des tâches liées au langage.

Par conséquent, cette thèse examine la valeur de l'utilisation de données séquentielles pour évaluer son impact réel dans l'industrie des services financiers et pour apporter de l'innovation basée sur les données dans le processus de prise de décision. Pour cela, différentes approches de travail avec des données séquentielles sont examinées, pour garantir l'extraction d'informations utiles pour les tâches prédictives est optimisée.

Plus précisément, l'étude présentée dans le chapitre 2 utilise des données séquentielles sous forme de prix du pétrole historiques et de publications sur les réseaux sociaux. L'inclusion de ces publications est comparée à l'exclusion des données textuelles, pour évaluer si la performance de la prédiction des prix du pétrole s'améliore ou non. De plus,

différentes techniques de traitement des données textuelles sont explorées pour évaluer laquelle fonctionne le mieux, y compris TF-IDF, Word2vec, Doc2Vec, GloVe et BERT. Enfin, une analyse d'exclusion est déployée, confirmant davantage la valeur des données textuelles lorsqu'elles sont utilisées séquentiellement.

Le chapitre 3 compare la performance des données caractérisées, transformant ainsi les données longitudinales en données transversales, par rapport aux données séquentielles. Les données caractérisées peuvent être traitées par des classificateurs statistiques et d'apprentissage automatique, tandis que les données séquentielles sont traitées par un modèle DL. Les résultats montrent que les modèles DL surpassent les autres approches avec des différences de performance statistiquement significatives. Par conséquent, on peut conclure que les données séquentielles ajoutent en effet de la valeur à la prédiction des événements de la vie.

De même, le chapitre 4 compare la performance des données transversales par rapport aux données séquentielles pour la production de recommandations. Les données transversales sont traitées par des algorithmes RS et MLC, tandis que les données séquentielles sont traitées par un modèle DL. Le modèle DL utilisant des données séquentielles fonctionne mieux que les algorithmes RS et MLC, avec des différences de performance statistiquement significatives pour les recommandations.

Dans l'ensemble, toutes les études montrent que les données séquentielles fournissent des informations précieuses pour différentes tâches prédictives. De plus, les chapitres 3 et 4 montrent que les données séquentielles à travers les modèles DL surpassent les autres approches pour leurs tâches respectives. Par conséquent, l'utilisation de données client longitudinales est particulièrement précieuse pour les applications CRM, par opposition à l'utilisation de données caractérisées, semblable à une approche transversale. Ces résultats sont en accord avec des recherches existantes, qui soulignent la valeur des données longitudinales.

Déploiement de techniques d'explicabilité pour les données séquentielles

Les réseaux neuronaux ont démontré un succès significatif dans diverses applications, dépassant fréquemment les capacités prédictives des modèles d'apprentissage automatique conventionnels (Kraus et al., 2020). Notamment, la structure adaptable des architectures DL permet la création de modèles pouvant manipuler différents types de données d'entrée, en particulier les données séquentielles, avec peu de prétraitement (De Bock et al., 2023). Cependant, ces modèles sont critiqués comme étant des boîtes noires, en raison de la complexité impliquée lors de l'explication de la relation entre les données d'entrée et la sortie (De Bock et al., 2023). Ceci devient un point de recherche particulièrement pertinent lorsque l'on considère le contexte des réglementations changeantes, telles que le Règlement général sur la protection des données (RGPD), la Loi sur les services numériques (DSN) et la Loi sur les marchés numériques (DMN) (Cauffman and Goanta, 2021).

De plus, la montée de nouvelles technologies a conduit de nombreuses entreprises à collecter de vastes volumes de données client, sans nécessairement en tirer parti (Aina Turillazzi and Casolari, 2023). Ainsi, la conceptualisation et l'application de méthodes avancées pour transformer les données en informations présentant des résultats performants, interprétables et responsables dans le cadre réglementaire, sont essentielles pour améliorer le processus de prise de décision (De Bock et al., 2023).

Dans cette thèse, nous mettons en œuvre des modèles DL en utilisant des données conformes au RGPD, pour développer des modèles innovants et performants dans

l'industrie financière. Plus précisément, nous avons collaboré avec le service juridique d'un prestataire de services financiers pour obtenir le consentement des clients, conformément à l'article 15 du RGPD. Nous explorons différentes techniques d'explicabilité pour tirer des informations des données, améliorer la transparence du modèle et aider davantage dans le processus de prise de décision.

Notamment, dans le chapitre 2, l'utilisation de l'analyse d'exclusion permet de capturer l'impact sur le prix du pétrole de certains mots-clés ou sujets. Par exemple, une analyse d'exclusion partielle, dans laquelle les mots-clés liés au pétrole sont supprimés des données textuelles, révèle une baisse des performances. De même, une exclusion totale, où tous les tweets contenant ces mots-clés sont entièrement supprimés, présente une baisse plus drastique des performances. De plus, une analyse de changement structurel, où les données textuelles menant à une déviation dans les valeurs des prix du pétrole sont examinées de près, révèle la présence de mots-clés liés à des événements géopolitiques et des concepts. Ces résultats sont en accord avec la littérature existante, indiquant que le prix du pétrole est influencé par des événements géopolitiques mondiaux, la politique et le sentiment du marché, ainsi que les annonces publiques. Aucune de ces approches n'a été utilisée auparavant dans la littérature sur la prédiction des prix du pétrole. Ainsi, ces analyses éclairent davantage la source de la puissance prédictive pour la prédiction des prix du pétrole à partir du modèle DL mis en œuvre.

Le chapitre 3 utilise la méthode d'attribution des gradients intégrés (IG) (Sundararajan et al., 2017), qui n'avait pas été appliquée auparavant à la prédiction des événements de la vie. Cette méthode quantifie la contribution de chaque entrée par rapport à la sortie du modèle, représentant ainsi la pertinence d'une caractéristique pour la variable cible. Les résultats révèlent que (i) les mois les plus proches de la période de prédiction ont un poids plus important en termes de performance prédictive, (ii) l'importance des caractéristiques diffère pour chaque événement de la vie, et (iii) les caractéristiques disponibles séquentiellement ont une plus grande influence sur la prédiction des événements de la vie que les caractéristiques transversales. Du point de vue de la prise de décision en marketing, le premier point concerne le moment optimal pour contacter les clients, ce qui pourrait permettre une allocation plus efficace des ressources. Par exemple, un client ayant une forte probabilité d'achat d'une résidence principale pourrait être plus réceptif lorsqu'il est contacté quelques mois seulement avant que l'événement de vie ne se produise. Comme le révèle également une recherche antérieure, l'allocation correcte et opportune des ressources marketing, reflétant une compréhension précise du comportement dynamique du client, a un impact positif sur la fidélité du client. De plus, le deuxième point permet aux décideurs d'utiliser les prédictions des événements de la vie comme un outil pour améliorer la segmentation, personnaliser les services qu'ils offrent, détecter de nouvelles opportunités de vente croisée et améliorer leurs recommandations de produits. Enfin, le troisième point souligne la valeur des données longitudinales. Ainsi, ces résultats mettent en évidence l'importance de transformer des données complexes en informations exploitables.

Enfin, le chapitre 4 déploie les valeurs SHapley Additive exPlanations (SHAP) pour analyser davantage les différences d'importance des caractéristiques lors de la production de recommandations. SHAP attribue une valeur d'importance à chaque entrée, identifiant ainsi les caractéristiques fortement corrélées à la sortie d'un modèle. Les résultats montrent que les caractéristiques séquentielles ont un impact plus important que les caractéristiques statiques dans toutes les catégories de produits. De plus, les caractéristiques

liées aux produits d'assurance, en particulier les données de fréquence et monétaires, ont un impact élevé dans toutes les catégories de produits. De plus, les schémas comportementaux diffèrent selon la catégorie de produit. Dans l'ensemble, ces résultats signalent que les clients qui investissent dans les assurances sont de bons candidats à cibler pour des initiatives de vente croisée, tandis que les campagnes marketing pourraient bénéficier d'une adaptation au comportement des clients de manière longitudinale.

D'autres techniques sont encore à l'étude, car il est encore nécessaire d'évaluer l'impact de différentes méthodes d'explicabilité, pour comparer leurs forces et leurs faiblesses dans différentes applications. Par conséquent, les techniques déployées ne sont en aucun cas exhaustives, mais elles contribuent à approfondir la recherche appliquée dans l'industrie des services financiers, en utilisant des données réelles avec des informations exploitables pour les décideurs.

Résumé des Contributions

En résumé, les trois études contenues dans cette thèse utilisent des sources de données et des approches novatrices pour leurs tâches prédictives respectives. De plus, toutes les études montrent que les données séquentielles fournissent des informations précieuses pour différentes tâches prédictives. Enfin, les techniques d'explicabilité déployées élargissent la recherche appliquée dans l'industrie financière, en utilisant des données réelles avec des informations exploitables pour les décideurs. Ces contributions sont approfondies dans chaque article de recherche, résumé dans les paragraphes suivants.

Le chapitre 2 utilise les prix du pétrole antérieurs et les données textuelles comme entrée séquentielle pour prédire les prix futurs du pétrole. Les contributions sont (i) l'inclusion d'une nouvelle source de données textuelles pour la prédiction des prix du pétrole, (ii) l'utilisation d'un large éventail de techniques de traitement du langage naturel (NLP) pour extraire des informations contextuelles, et (iii) l'incorporation d'analyses supplémentaires pour élucider davantage la sortie du modèle. Les techniques de NLP consistent en une approche basée sur l'espace vectoriel, des modèles d'encastrement et une technique basée sur les transformateurs, appelée Bidirectional Encoder Representations from Transformers (BERT). Les résultats montrent que BERT est la technique supérieure pour extraire des informations pertinentes des données textuelles pour la prédiction des prix du pétrole. De plus, les techniques d'explicabilité révèlent des mots-clés liés à des événements géopolitiques lors de changements structurels dans les prix du pétrole, en accord avec la littérature existante.

Le chapitre 3 utilise des données démographiques et comportementales des clients, fournies par un prestataire de services financiers, pour prédire 10 événements de vie différents. Dans cette étude, les données séquentielles sont disponibles longitudinalement sous forme de données comportementales des clients. Les contributions consistent en (i) l'incorporation de nouveaux événements de vie pour la prédiction, (ii) la comparaison de la performance prédictive des données séquentielles et transversales, et (iii) la fourniture d'informations pour la prise de décision en marketing. Les résultats révèlent que les données séquentielles sont plus performantes que les données transversales, avec des données comportementales ayant un impact plus important sur la prédiction des événements de vie que les caractéristiques démographiques. En général, les données les plus récentes de l'entrée séquentielle ont une influence plus importante sur la prédiction du modèle. Enfin, l'occurrence d'un événement de vie a un impact clair sur les taux de rétention.

Le chapitre 4 utilise des données démographiques des clients et l'historique des achats

sous forme de caractéristiques RFM. Ici, les données séquentielles proviennent des caractéristiques RFM, disponibles longitudinalement. Ces données proviennent du même prestataire de services financiers et sont utilisées pour construire différents systèmes de recommandation. Les contributions sont (i) la comparaison d'algorithmes de recommandation, y compris des modèles DL de pointe dans un scénario réel, (ii) la comparaison des techniques de featurisation pour évaluer si l'incorporation de données longitudinales améliore la performance des recommandations, (iii) l'évaluation des données longitudinales en tant qu'entrée séquentielle, grâce à l'utilisation de modèles d'apprentissage en profondeur, (iv) l'application de techniques d'explicabilité pour améliorer la compréhension des décideurs et des spécialistes du marketing lors du déploiement de systèmes de recommandation.

1.2 Abstract

1.2.1 General Abstract

Amid rising competition intensity, having strong customer relationships is considered a key driver for maintaining a competitive edge. Thus, companies have been investing significantly in the development of customer relationship management (CRM) strategies in recent years (Dalla Pozza et al., 2018). Theoretically, CRM consists of several different initiatives, each of which can be classified into four different dimensions (Kumar and Reinartz, 2006): (i) organizational alignment, (ii) customer management, (iii) technology, and (iv) CRM strategy implementation. The first dimension, organizational alignment, refers to the redesigning and aligning of existing processes, with the ultimate objective of placing customers at the center (Dalla Pozza et al., 2018). The second dimension, customer management, operates under the principle of treating customers differently, by adapting to their needs, preferences, and priorities (Reinartz et al., 2004). The third dimension, technology, encompasses the degree to which analytical, operative, and collaborative CRM applications are implemented to collect customer information across the touch points and to facilitate information dissemination and analysis (Dalla Pozza et al., 2018). Lastly, the CRM strategy implementation dimension requires a clear customer-oriented approach, including top management support, customer-oriented performance metrics, and a comprehensive view of the customer across the entire organization (Palmatier et al., 2007).

A successful implementation of CRM strategies addresses all four dimensions simultaneously. However, this thesis strongly focuses on the dimension of technology. This dimensions offers the opportunity of harnessing innovative methods for a better understanding of the financial services industry, the factors affecting it, and insights useful for its decision-makers. Ultimately, the use of tools such as innovative technology and novel data sources are proposed as means to achieve a deeper understanding of the context in which financial services providers interact with their customers.

Figure 1.5 displays a schematic illustration of data-driven insights that can be used for a deeper understanding of the context in which financial services providers interact with their customers. Specifically, insights from external factors can be useful to understand market fluctuations and economic conditions that can affect both customer priorities and a financial services provider's growth opportunities. Insights from the providers themselves, represented by the middle ring, refers to analyses about the products, services, and offers available, as well as characteristics that may differentiate a provider from its competitors, such as the communication channels at the disposal of their customers, the quality of the services, or the location of different branches, among others. Finally, the insights from customers includes information about their behavior, demographics, and intensity of the relationship with the providers, among others. Thus, these perspectives of insights are modeled as concentric rings due to their interdependence. In this thesis, we explore applications for all different data-driven insights to better understand the financial sector and develop innovative applications which may be harnessed to continue strengthening the three other CRM dimensions.

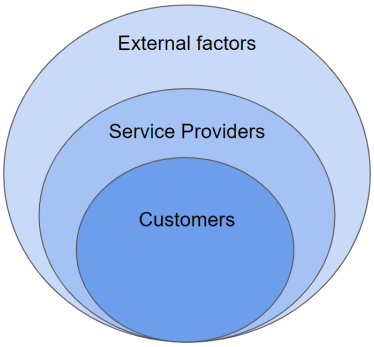


Figure 1.5. Schematic illustration of data-driven insights in the financial services industry

The motivation behind exploring the different sources of data is related to the spectacular advances that have occurred during the last few years in the technological sector. These developments include the ability to process large volumes of data through improved models for predictive tasks, such as deep learning (DL), which grants the opportunity to explore novel and complex sources of data.

To explore how research interest around these technologies has evolved, we review the area of recommendation systems research, chosen for being a popular example of CRM-related publications, with strong managerial implications for the financial services industry. In fact, the number of recommendation systems articles is far higher than other predictive tasks in CRM, such as churn, customer segmentation, or sales forecasting.

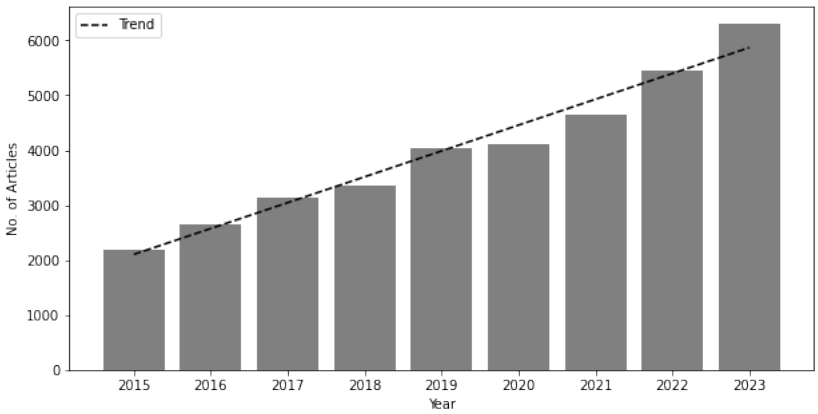


Figure 1.6. Number of recommendation systems studies published per year

Using Scopus to enter the keywords “recommender system”, “recommendation system”, “collaborative filtering”, “content-based recommender”, or “content-based recommendation”, a total of 38.171 articles and conference papers are found to be published in English between 2015 and 2023. The interest in recommender systems has been steadily increasing, with more articles being published every year, as illustrated on Figure 1.6.

These documents can further be segmented into studies that use DL and studies that focus on traditional recommendation systems. The first group, DL, includes publications found through keywords sourced from Goodfellow et al. (2016), such as “deep learning”, “neural networks”, “MLP”, “CNN”, “LSTM”, “RNN”, “GRU”, “GNN”, “Boltzmann”, “autoencoders”, “transformers”, and “representation learning”, among others. The second group excludes publications both from the first group and that use hybrid approaches. From these two groups, several conclusions can be drawn. Firstly, studies using traditional recommenders (TR) remain highly popular and the volume has continued to in-

crease. Secondly, the interest in newer technologies has been growing at an accelerated rate, as shown in Figure 1.7. In fact, 2023 is shown to be the first year where DL research on recommendation systems has surpassed the use of TR.

These changes highlight the need for companies to keep up to date with the changing landscape to maintain a sustained competitive advantage. As such, this thesis contributes to innovative technology applications in the financial services industry, by using novel sources of data, deploying state-of-the art methodologies for predictive modeling, and implementing advanced interpretability techniques, with a focus on providing quantitative insights for decision-makers.

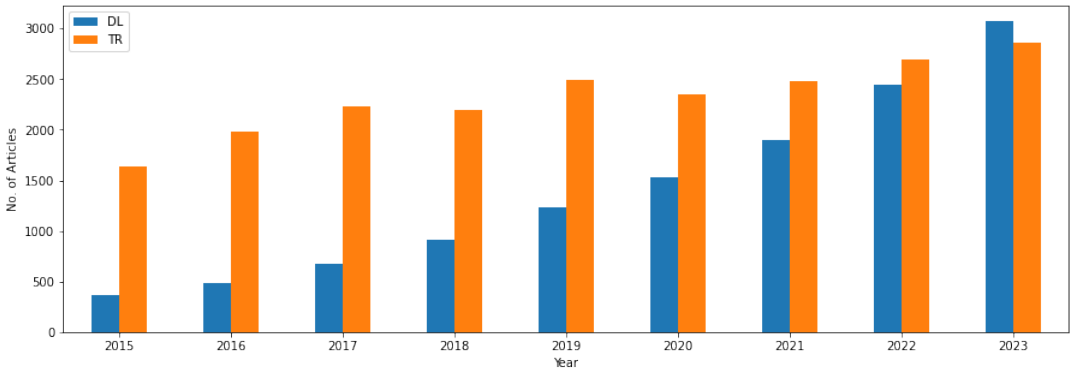


Figure 1.7. Number of studies using DL and CF algorithms per year

The predictive modeling on this thesis focuses on the use of DL, which have been frequently criticized for being opaque or black box models. Notably, research has called attention to the fact that black box models are already causing problems in sensitive industries, such as healthcare and criminal justice (Rudin, 2019). However, many of these problems, namely lack of transparency and accountability of predictive models, are being controlled in Europe through regulation such as the GDPR. For instance, Article 15 of the GDPR guarantees transparency and responsibility from organizations, by allowing individuals to request access to their personal data. This provision also allows individuals access to information regarding automated decision-making, such as profiling or targeting. This information must also include explanations of the underlying logic and the impact of such data processing on the individual. Thus, industries that are under more scrutiny, such as the financial services industry, may be more hesitant to adopt such black box models. Nonetheless, there is still merit in persisting with research on these models, ultimately to gain a more comprehensive understanding of their pros and cons. For instance, black box models often demonstrate superior performance, which serves as a valuable benchmark for white box models to surpass whenever feasible. Moreover, research on explainability has advanced since the publication of Rudin (2019), with an example being the neuron Shapley method, which enhances transparency and granularity in DL analysis (Ghorbani and Zou, 2020). Consequently, black box models frequently outshine white box models in specific applications, and techniques for explainability are still being refined for more opaque models. In essence, black box models are still being developed, warranting ongoing research efforts to understand the value they can add to different applications, whether as benchmarks for white box models or for uncovering additional data-driven insights.

These insights are particularly important within the financial services industry, as it is highly difficult to attract new clients (Knott et al., 2002). Simultaneously, the loy-

alty of existing customers fluctuates from external factors impacting their consumption preferences and priorities, such as geopolitical events, market trends, or changes in their lives (Mathur et al., 2008). Nonetheless, companies can dampen the effect of these external factors by understanding their own customers' concerns and interests, to then take appropriate marketing actions (Jackson, 1985).

A profound understanding of customer behavior can help a company identify multiple issues that will determine future purchase decisions. It can help understand whether a customer will prioritize more immediate concerns, like a short-term loan, or long-term issues, like upgrading from one product to another over time (Jackson, 1985). Due to the higher switching costs for customers that prioritize long-term issues (Jackson, 1985), external factors will have a lower impact. Thus, the dimension of time is essential for a holistic understanding of both the customer behavior as well as the potential impact of marketing actions, allowing marketers to design adaptions for customers along the short-term and long-term concerns spectrum (Jackson, 1985).

Therefore, the understanding of customers and their preferences in time allows for a constant improvement of customer-firm relationships (Roos and Gustafsson, 2007). This can be achieved by monitoring and analyzing customer behavior and designing specialized marketing actions (Sin et al., 2005), ultimately allowing for a more efficient resource allocation by targeting the most relevant customers (Sin et al., 2005).

In this context, this thesis explores different avenues of applied research, harnessing time-varying sequential data for different tasks: oil price prediction, life event prediction, and recommendation systems. Three key contributions are made: incorporating novel sequential data sources to improve predictive performance, applying state-of-the-art methodologies to optimize the use of sequential data, and deploying explainability techniques to explore how sequential data contributes to a predictive task. Each of the three upcoming chapters contains a different study using sequential data, while the last chapter closes with general conclusions and research limitations.

Sequential data refers to ordered information that helps a model detect patterns, to solve tasks of different complexities. Multiple of the models previously listed as keywords for Figure 1.7 have been used to learn from sequential data. Arguably, sequential data has had its breakthrough into celebrity status thanks to large language models such as ChatGPT. Here, sequences of words and sentences are used to train huge models, based on DL architectures, to find solutions for different problems related to text. Nonetheless, sequential data exists across multiple fields, with some examples displayed on Figure 1.8. For instance, a sequence of DNA can be used for predicting gene expression (Žiga Avsec et al., 2021). Similarly, a sequence of monthly financial assets can be used for predicting future assets, churn, loan default, or other tasks related to customer classification. Sequential data can thus be used through models such as DL for predictive tasks relevant to several industries, for both classification and regression problems. This thesis specifically focuses on the use of chronologically ordered data from the financial industry for prediction.

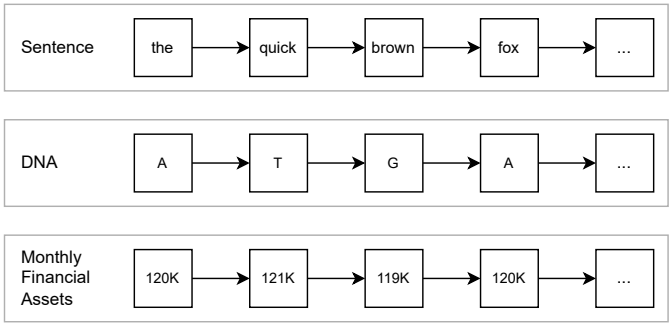


Figure 1.8. Examples of sequential data

Chapter 2 uses previous oil prices and textual data as a sequential input to predict upcoming oil prices. Chapter 3 uses customer demographic and sequential behavioral data, supplied by a financial services provider, to predict 10 different life events. Chapter 4 uses customer demographic data and purchase history in the form of sequential RFM features.

In sum, all three studies present results derived from real-world data, incorporating novel sequential data sources, implementing state-of-the-art methodologies for assessing the use of the aforementioned sequential data, and deploying explainability techniques to further analyze the results. Each study showcases academic value and notable business relevance, through the three previously listed contributions.

Keywords: Big data, customer relationship management, marketing analysis, machine learning, neural networks, text analysis, oil price, life event prediction, recommendation system.

1.2.2 Detailed Abstract

This section summarizes all three papers contained in this thesis, examining how they approach the three general contributions previously outlined as the common thread of this thesis. These contributions are namely: (i) incorporating novel data sources, (ii) assessing the use of sequential data, and (iii) deploying explainability techniques. These contributions are in line with the challenges brought about by the current context, making the three studies useful examples of innovative approaches for improving the decision-making process within the financial industry. Thus, this section further contextualizes these studies through the three general contributions and the current state of research and existing literature.

This thesis has been developed in a context where models based on DL and related research are advancing at incredible speed. For instance, ChatGPT is continually changing the data analysis landscape, as it has both accelerated research around DL applications, as well as multiplied the questions arising from regulators and the general public (Cauffman and Goanta, 2021). This scrutiny is not restricted to ChatGPT and, in fact, also applies to DL algorithms in general. This points towards the business relevance of balancing model performance with explainability, as existing studies show that the perception of security, privacy, and the compatibility with consumer values can impact the decision to adopt innovative applications and services, particularly in the financial industry (Luo et al., 2010; Hoehle et al., 2012).

Therefore, it is highly important to develop explainable applications, regardless of

the model selected, to ensure the transparency of model output. It is also important to allow for all stakeholders to achieve a rapid understanding of insights provided by a model, while also allowing researchers to verify the value of complex data sources. As such, these issues are becoming an integral part of overall CRM strategies and research, as part of a responsible analytics framework (De Bock et al., 2023).

We posit that explainability techniques, combined with the use of novel and sequential data sources, are powerful tools for improving the transparency of innovative models. Moreover, these techniques allow for a multi-dimensional understanding of customer behavior, adding a time-varying factor of analysis with the same data already used in existing CRM applications (De Caigny et al., 2020; Bogaert et al., 2019). This means that additional sources of data, which may be perceived as more privacy-invasive, are not required for further improving performance. Instead, historical data can be used sequentially, achieving a more detailed understanding of customer behavior. This approach respects privacy guidelines and regulatory frameworks, without sacrificing innovation or performance. This context is further explored by detailing how each study aligns with the main contributions of this thesis.

Incorporating novel data sources

Throughout this thesis, we solve different problems through the use of novel data sources. In particular, this is achieved in two different ways. Firstly, by using data as a chronologically ordered sequence, to be compared against cross-sectional data, which remains an under-researched area of predictive literature for CRM applications (Óskarsdóttir et al., 2018). Secondly, by incorporating data sources that have not been previously evaluated within relevant literature.

In particular, chapter 2 incorporates a novel textual data source, used sequentially as input to predict upcoming oil prices, which addresses the insights regarding external factors, as shown on Figure 1.5. Research leveraging textual data for oil price prediction remains rare (Beyer Díaz et al., 2023), despite it still offering opportunities for predictive research in the financial sector (Huang et al., 2020). Moreover, the specific source of textual data used, tweets published by Donald Trump during his presidential term, had not been studied before (Beyer Díaz et al., 2023). Further, this data source is both novel and relevant, as oil price is influenced by global geopolitical events (Monge et al., 2017), politics and market sentiment (Alvarez-Ramirez et al., 2003), and public announcements (Singleton, 2014). Finally, multiple studies within the financial sector prove that using textual data for forecasting is beneficial (Kraus and Feuerriegel, 2017).

Chapter 3 uses customer demographic and behavioral data from a financial services provider, to predict 10 different life events for customer relation management (CRM) applications. The original data thus can be understood as contributing to the innermost data-driven insights ring from Figure 1.5. This data is longitudinal, but it is transformed through aggregation into cross-sectional data to be compared against the usage of sequential data as input. Life event research has mainly focused on cross-sectional data (De Caigny et al., 2020), which places the use of sequential data as a novelty. Further, existing research for applications in CRM emphasize the importance of longitudinal customer data to capture dynamic behavior (Óskarsdóttir et al., 2018) and to guarantee the validity of results over time (Boulding et al., 2005), positioning this study as a valuable contribution to life event literature. Lastly, this study also proposes new life events to be predicted, i.e. primary residence purchase, secondary residence purchase, and rental

residence purchase, which means the target labels also encompass novel data.

Chapter 4 analyzes customer demographic data and purchase history, represented as longitudinal RFM features. This data, sourced from the same financial services provider as in chapter 3, is used to provide data-driven insights from both the middle and the innermost rings of Figure 1.5. The data is employed to develop various recommender systems, including traditional recommender systems (RS), machine learning multi-label classifiers (MLC), and deep learning (DL) models. While RS and MLC capture customer behavior through cross-sectional data, neglecting temporal changes (You et al., 2019), DL models offer an alternative by leveraging sequential data. Successful applications of DL models in modeling online actions for generating recommendations have been noted (Tan et al., 2016; You et al., 2019). Financial customer behavior can also be modeled as a series of actions, which eventually lead to a purchase (Prinzie and Van den Poel, 2006). Past product ownership and the order of product acquisition are reliable predictors for future purchases in the financial services industry (Kamakura et al., 1991). Further, RFM features are relevant for generating recommendations (Bogaert et al., 2019) and excel at capturing purchasing behaviors (Chen et al., 2016). Nonetheless, limited research exists on the use of sequential data in this industry, compared to consumer durable goods or online platforms (Prinzie and Van den Poel, 2006). Additionally, RFM features evolve over time and are naturally dynamic, but their use as sequential input for predictive tasks remains an underdeveloped area of research (Mena et al., 2023). Consequently, the use of sequential input for recommendations remains a valuable research path, as it constitutes a novel data approach within the financial services industry.

Additionally, both chapters 3 and 4 involve customer data in sequential form for CRM applications, while also providing innovative and actionable insights for company strategy. A key task in the financial services industry is the targeting of relevant customers (Geuens et al., 2018), as existing customers are more profitable than new ones (Knott et al., 2002). This makes the use of novel data sources highly important as it enables practitioners to go beyond the factual behavioral profile and customer history, to develop strategies which make the customer feel like they are being listened to (Crié and Micheaux, 2006).

In sum, all three studies contained in this thesis employ novel data sources and approaches for their respective predictive tasks. Furthermore, all studies leverage sequential data as a means of improving performance, which remains a scarcely researched approach within the financial industry. Finally, novel data sources are an essential element of research, as it enables practitioners to draw new insights from their customers' behavior to construct innovative CRM strategies.

Assessing the use of sequential data

In the last few years, sequential data has been harnessed to achieve incredible applications. In particular, ChatGPT is often mentioned as a successful example of training a model on sequential data. Its architecture is based on a form of neural networks (Brown et al., 2020), with the ability to handle sequential data and capture long-range dependencies (Vaswani et al., 2017). ChatGPT's success inevitably lead to interest in exploring and adapting other forms of sequential data for different domains, expanding their utility beyond language-related tasks.

Therefore, this thesis examines the value of using sequential data to assess its real-life impact within the financial services industry and bring about data-driven innovation in the decision-making process. To achieve this, different approaches of working with

sequential data are examined, to ensure the extraction of useful information for predictive tasks is optimized.

Specifically, the study outlined in chapter 2 uses sequential data in the form of historical oil prices and social media posts. The inclusion of these posts are compared against the exclusion of textual data, to assess whether the performance of oil price prediction improves or not. Further, different techniques to process the textual data are explored, to evaluate which one performs better, including TF-IDF, Word2vec, Doc2Vec, GloVe, and BERT. Finally, exclusion analysis is deployed, further confirming the value of textual data when used sequentially.

Chapter 3 compares the performance of featurized data, thus transforming longitudinal data into cross-sectional data, against sequential data. Featurized data can be processed by statistical and machine learning classifiers, while sequential data is processed by a DL model. The results show that DL models outperform other approaches with statistically significant performance differences. Therefore, it can be concluded that sequential data does indeed add value for life event prediction.

Similarly, chapter 4 compares the performance of cross-sectional data against sequential data for producing recommendations. The cross-sectional data is processed by RS and MLC algorithms, while the sequential data is processed by a DL model. The DL model using sequential data performs better than RS and MLC algorithm, with statistically significant differences for recommendations.

Overall, all studies show that sequential data does provide valuable information for different predictive tasks. Moreover, both chapters 3 and 4 show that sequential data through DL models outperform other approaches for their respective tasks. Therefore, using longitudinal customer data is particularly valuable for CRM applications, as opposed to using featurized data, akin to a cross-sectional approach. These findings are in line with existing research, which highlight the value of longitudinal data (Óskarsdóttir et al., 2018).

Deploying explainability techniques for sequential data

Neural networks have demonstrated significant success across diverse applications, frequently surpassing the predictive capabilities of conventional machine learning models (Kraus et al., 2020). Notably, the adaptable structure of DL architectures enables the creation of models that can handle different types of input data, particularly sequential data, with little preprocessing (De Bock et al., 2023). However, these models are criticized as being black boxes, due to the complexity involved when explaining the relation between the input data and the output (De Bock et al., 2023). This becomes a particularly relevant point of research when considering the context of changing regulations, such as the General Data Protection Regulation (GDPR), the Digital Services Act (DSA) and the Digital Markets Act (DMA) (Cauffman and Goanta, 2021).

Additionally, the rise of new technologies has led many companies to collect vast volumes of customer data, without necessarily leveraging their benefits (Aina Turillazzi and Casolari, 2023). As such, the conceptualization and application of advanced methods for transforming data into insights that exhibit high-performance, interpretable results, and are responsible within the regulatory framework, is key to enhancing the decision-making process (De Bock et al., 2023).

In this thesis, we implement DL models using data that follows GDPR requirements, to develop innovative and high-performing models within the financial industry. Specif-

ically, we have collaborated with the legal department of a financial services provider to obtain consent from customers, following Article 15 of GDPR. We explore different explainability techniques to derive insights from the data, improve model transparency, and further aid in the decision-making process.

Namely, in chapter 2, the use of exclusion analysis allows to capture the impact on the price of oil of certain keywords or topics. For instance, a partial exclusion analysis, wherein oil-related keywords are removed from the textual data source, reveals a drop in performance results. Similarly, full exclusion, where all tweets containing these keywords are fully removed, exhibit a more drastic drop in performance. Further, a structural change analysis, where the textual data leading up to a deviation in the oil price values is closely examined, reveals the presence of keywords relevant to geopolitical events and concepts. These findings are in line with existing literature, pointing towards oil price being influenced by global geopolitical events (Monge et al., 2017), politics and market sentiment (Alvarez-Ramirez et al., 2003), and public announcements (Singleton, 2014). Neither of these approaches have been previously used within the oil price prediction literature. Thus, these analyses further elucidate the source of predictive power for oil price prediction from the implemented DL model.

Chapter 3 uses integrated gradients (IG) attribution method (Sundararajan et al., 2017), which has not been previously applied to life event prediction. This method quantifies the contribution of each input in relation to the output of the model, therefore representing a feature's relevance for the target variable. The results reveal that (i) the months closer to the prediction period exert greater weight in terms of predictive performance, (ii) the feature importance differs for each life event, and (iii) features that are available sequentially have a larger influence on life event prediction than cross-sectional features. From a marketing decision-making perspective, the first point relates to the optimal moment to contact customers, which could enable more efficient resource allocations. For example, a customer with a high probability of a primary residence purchase may be most responsive when contacted just a few months before the life event occurs. As previous research also reveals, the correct and timely allocation of marketing resources, reflecting an accurate understanding of dynamic customer behavior, has a positive impact on customer loyalty (Han and Anderson, 2022). Further, the second point allows decision-makers to use life event predictions as a tool to improve segmentation, to personalize the services they offer, detect new cross-selling opportunities, and improve their product recommendations. Finally, the third point echos the value of longitudinal data (Óskarsdóttir et al., 2018). Thus, these findings highlighting the importance of transforming complex data into actionable insights (De Bock et al., 2023).

Finally, chapter 4 deploys SHapley Additive exPlanations (SHAP) values to further analyze the differences in feature importance when producing recommendations. SHAP assigns an importance value for each input, thus identifying features strongly correlated with a model's output (Notz and Pibernik, 2024). The results show that sequential features have a higher impact than static features across product categories. Furthermore, features related to Insurance products, particularly frequency and monetary data, have a high impact across product categories. Moreover, behavioral patterns differ by product category. Overall, these findings signal that customers that invest in insurance are good candidates to target for cross-selling initiatives, while marketing campaigns could benefit from adapting to customer behavior longitudinally.

Additional techniques are still being researched, as it is still required to assess the im-

pact of various explainability methods, to compare their strengths and weaknesses across different applications. Therefore, the deployed techniques are by no means exhaustive, but they do contribute in furthering applied research in the financial services industry, using real-life data with actionable insights for decision-makers.

Summary of Contributions

In sum, all three studies contained in this thesis employ novel data sources and approaches for their respective predictive tasks. Moreover, all studies show that sequential data does provide valuable information for different predictive tasks. Finally, the deployed explainability techniques expand applied research in the financial industry, using real-life data with actionable insights aimed at decision-makers. These contributions are further deepened on each research article, summarized in the following paragraphs.

Chapter 2 uses previous oil prices and textual data as a sequential input to predict upcoming oil prices. The contributions are (i) the inclusion of a novel text data source for oil price prediction, (ii) the use of a wide array of natural language processing (NLP) techniques for extracting contextual information, and (iii) the incorporation of additional analyses to further elucidate the model's output. The NLP techniques consist of a vector space-based approach, embedding models, and a transformer-based technique, known as Bidirectional Encoder Representations from Transformers (BERT). Results show BERT is the superior technique to extract relevant information from textual data for oil price prediction. Further, explainability techniques reveal keywords related to geopolitical events during structural changes in oil prices, in line with existing literature.

Chapter 3 uses customer demographic and behavioral data, supplied by a financial services provider, to predict 10 different life events. In this study, sequential data is available longitudinally, in the form of customer behavioral data. The contributions consist of (i) incorporating novel life events for prediction, (ii) comparing the predictive performance of sequential and cross-sectional data, and (iii) delivering insights for decision-making in marketing. The findings reveal that sequential data performs better than cross-sectional data, with behavioral data impacting life event prediction more than demographic features. In general, the more recent data from the sequential input has a heavier influence on the model's prediction. Lastly, the occurrence of a life event has a clear impact on retention rates.

Chapter 4 uses customer demographic data and purchase history in the form of RFM features. Here, the sequential data stems from RFM features, available as longitudinal data. This data stems from the same financial services provider and is used to construct different recommender systems. The contributions are (i) the comparison recommendation algorithms, including state-of-the-art DL models in a real-life scenario, (ii) the contrast of featurization techniques to assess if the incorporation of longitudinal data improves the performance of recommendations, (iii) the evaluation of longitudinal data as sequential input, through the use of deep learning models, (iv) the application of explainability techniques to improve the understanding of decision-makers and marketers when deploying recommendation systems.

1.3 References

Aina Turillazzi, Mariarosaria Taddeo, L.F., Casolari, F., 2023. The digital services act: an analysis of its ethical, legal, and social implications. Law, Innovation and Technology 15, 83–106. doi:10.1080/

17579961.2023.2184136.

Alvarez-Ramirez, J., Soriano, A., Cisneros, M., Suarez, R., 2003. Symmetry/anti-symmetry phase transitions in crude oil markets. Phys. A Stat. Mech. its Appl. 322, 583–596. doi:10.1016/S0378-4371(02)01831-9.

Žiga Avsec, Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R., 2021. Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods 18, 1196–1203. doi:10.1038/s41592-021-01252-x.

Beyer Díaz, S., Coussement, K., De Caigny, A., Pérez, L.F., Creemers, S., 2023. Do the us president's tweets better predict oil prices? an empirical examination using long short-term memory networks. Int. J. Prod. Res. 0, 1–18. doi:10.1080/00207543.2023.2217286.

Bogaert, M., Lootens, J., Van den Poel, D., Ballings, M., 2019. Evaluating multi-label classifiers and recommender systems in the financial service sector. Eur. J. Oper. Res. 279, 620–634. doi:10.1016/j.ejor.2019.05.037.

Boulding, W., Staelin, R., Ehret, M., Johnston, W.J., 2005. A customer relationship management roadmap: What is known, potential pitfalls, and where to go. J. Mark. 69, 155–166. doi:10.1509/jmkg.2005.69.4.155.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 1877–1901.

Cauffman, C., Goanta, C., 2021. A new order: The digital services act and consumer protection. Eur. J. Risk Regul. 12, 758–774. doi:10.1017/err.2021.8.

Chen, Z.Y., Fan, Z.P., Sun, M., 2016. A multi-kernel support tensor machine for classification with multitype multiway data and an application to cross-selling recommendations. Eur. J. Oper. Res. 255, 110–120. doi:doi.org/10.1016/j.ejor.2016.05.020.

Crié, D., Micheaux, A., 2006. From customer data to value: What is lacking in the information chain? J. Database Mark. Cust. Strategy Manag. 13, 282–299. doi:10.1057/palgrave.dbm.3240306.

Dalla Pozza, I., Goetz, O., Sahut, J.M., 2018. Implementation effects in the relationship between crm and its performance. J. Bus. Res. 89, 391–403. doi:10.1016/j.jbusres.2018.02.004.

De Bock, K.W., Coussement, K., Caigny, A.D., Słowiński, R., Baesens, B., Boute, R.N., Choi, T.M., Delen, D., Kraus, M., Lessmann, S., Maldonado, S., Martens, D., Óskarsdóttir, M., Vairetti, C., Verbeke, W., Weber, R., 2023. Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. Eur. J. Oper. Res. doi:10.1016/j.ejor.2023.09.026.

De Bock, K.W., Coussement, K., De Caigny, A., Slowiński, R., Baesens, B., Boute, R.N., Choi, T.M., Delen, D., Kraus, M., Lessmann, S., et al., 2023. Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. Eur. J. Oper. Res. doi:10.1016/j.ejor.2023.09.026.

De Caigny, A., Coussement, K., De Bock, K.W., 2020. Leveraging fine-grained transaction data for customer life event predictions. Decis. Support Syst. 130, 113232. doi:10.1016/j.dss.2019.113232.

Geuens, S., Coussement, K., De Bock, K.W., 2018. A framework for configuring collaborative filtering-based recommendations derived from purchase data. Eur. J. Oper. Res. 265, 208–218. doi:10.1016/j.ejor.2017.07.005.

Ghorbani, A., Zou, J.Y., 2020. Neuron shapley: Discovering the responsible neurons, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 5922–5932. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/41c542dfe6e4fc3deb251d64cf6ed2e4-Paper.pdf.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org.

Han, S., Anderson, C.K., 2022. The dynamic customer engagement behaviors in the customer satisfaction survey. Decis. Support Syst. 154, 113708. doi:10.1016/j.dss.2021.113708.

Hoehle, H., Scornavacca, E., Huff, S., 2012. Three decades of research on consumer adoption and utilization of electronic banking channels: A literature analysis. Decis. Support Syst. 54, 122–132. doi:10.1016/j.dss.2012.04.010.

Huang, S., Potter, A., Eyers, D., 2020. Social media in operations and supply chain management: State-of-the-art and research directions. Int. J. Prod. Res. 58, 1893–1925. doi:10.1080/00207543.2019.1702228.

Jackson, B.B., 1985. Build customer relationships that last. Harv. Bus. Rev. 63 (November-December), 120–128.

Kamakura, W.A., Ramaswami, S.N., Srivastava, R.K., 1991. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. Int. J. Res. Mark. 8, 329–349. doi:10.1016/0167-8116(91)90030-B.

Knott, A., Hayes, A., Neslin, S.A., 2002. Next-product-to-buy models for cross-selling applications. J. Interact. Mark. 16, 59–75.

Kraus, M., Feuerriegel, S., 2017. Decision support from financial disclosures with deep neural networks and transfer learning. Decis. Support Syst. 104, 38–48. doi:10.1016/j.dss.2017.10.001.

Kraus, M., Feuerriegel, S., Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. Eur. J. Oper. Res. 281, 628–641. doi:10.1016/j.ejor.2019.09.018.

Kumar, V., Reinartz, W.J., 2006. Customer relationship management: A databased approach. Wiley Hoboken.

Luo, X., Li, H., Zhang, J., Shim, J., 2010. Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services. Decis. Support Syst. 49, 222–234. doi:10.1016/j.dss.2010.02.008.

Mathur, A., Moschis, G.P., Lee, E., 2008. A longitudinal study of the effects of life status changes on changes in consumer preferences. J. Acad. Mark. Sci. 36, 234—-246. doi:10.1007/s11747-007-0021-9.

Mena, G., Coussement, K., De Bock, K., De Caigny, A., Lessmann, S., 2023. Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. Ann. Oper. Res. 53, 80–95. doi:10.1007/s10479-023-05259-9.

Monge, M., Gil-Alana, L.A., Pérez de Gracia, F., 2017. Crude oil price behaviour before and after military conflicts and geopolitical events. Energy 120, 79–91. doi:10.1016/j.energy.2016.12.102.

Notz, P.M., Pibernik, R., 2024. Explainable subgradient tree boosting for prescriptive analytics in operations management. Eur. J. Oper. Res. 312, 1119–1133. doi:doi.org/10.1016/j.ejor.2023.08.037.

Palmatier, R.W., Scheer, L.K., Houston, M.B., Evans, K.R., Gopalakrishna, S., 2007. Use of relationship marketing programs in building customer–salesperson and customer–firm relationships: Differential influences on financial outcomes. Int. J. Res. Mark. 24, 210–223.

Prinzie, A., Van den Poel, D., 2006. Investigating purchasing-sequence patterns for financial services using markov, mtd and mtdg models. Eur. J. Oper. Res. 170, 710–734. doi:10.1016/j.ejor.2004.05.004.

Reinartz, W., Krafft, M., Hoyer, W.D., 2004. The customer relationship management process: Its measurement and impact on performance. J. Mark. Res. 41, 293–305.

Roos, I., Gustafsson, A., 2007. Understanding frequent switching patterns. Journal of Service Research 10, 93–108. doi:10.1177/1094670507303232.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–201. doi:10.1038/s42256-019-0048-x.

Sin, L., Tse, A., Yim, F., 2005. CRM: conceptualization and scale development. Eur. J. Mark. 39, 1264–1290. doi:10.1108/03090560510623253.

Singleton, K.J., 2014. Investor flows and the 2008 boom/bust in oil prices. Manage. Sci. 60, 300–318. doi:10.1287/mnsc.2013.1756.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, PMLR. p. 3319–3328.

Tan, Y.K., Xu, X., Liu, Y., 2016. Improved recurrent neural networks for session-based recommendations, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Association for Computing Machinery, New York, NY, USA. p. 17–22. doi:10.1145/2988450.2988452.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Curran Associates Inc., New York, NY, USA. p. 6000–6010.

You, J., Wang, Y., Pal, A., Eksombatchai, P., Rosenburg, C., Leskovec, J., 2019. Hierarchical temporal convolutional networks for dynamic recommender systems, in: The World Wide Web Conference, Association for Computing Machinery, New York, NY, USA. p. 2236–2246. doi:10.1145/3308558.3313747.

Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., Vanthienen, J., 2018. Time series for early churn detection: Using similarity based classification for dynamic networks. Expert Syst. Appl. 106, 55–65. doi:10.1016/j.eswa.2018.04.003.

Chapter 2: Do the US President's Tweets Better Predict Oil Prices? An Empirical Examination Using Long Short-Term Memory Networks

CHAPTER 2

Do the US President's Tweets Better Predict Oil Prices? An Empirical Examination Using Long Short-Term Memory Networks

Abstract.

The price of oil is highly complex to predict as it is impacted by global demand and supply, geopolitical events, and market sentiment. The accuracy of such predictions, however, has far-reaching implications for supply chain performance, portfolio management, and expected stock market returns. This paper contributes to the oil price prediction literature by evaluating the predictive impact of the US President's communication on Twitter, while benchmarking various Natural Language Processing (NLP) techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, Doc2Vec, Global Vectors for Word Representation (GloVe), and Bidirectional Encoder Representations from Transformers (BERT). These techniques are combined with a deep neural network Long Short-Term Memory (LSTM) architecture using a five-day lag for both the oil price and the textual Twitter data. The data was collected during the term of US President Donald Trump, resulting in 1,449 days of crude oil price prediction and a total of 16,457 tweets. The study is validated for Brent and West Texas Intermediate blends, using the daily price of a barrel of crude oil as the target feature. The results confirm that including the US President's tweets significantly increases the predictive power of oil price prediction models, and that an LSTM architecture with BERT as NLP technique has the best performance.

Keywords: Analytics, Oil price prediction, LSTM, BERT, NLP, US President.

2.1 Introduction

Oil price prediction is an important prediction task in operations management because of its economical impact. Higher predicted oil prices often result in lower economic activity, especially when the price increase is expected to persist (Heath, 2019). Higher oil prices also upend the way companies manage their supply chain (Simchi-Levi et al., 2008b,a), with oil spot trading being helpful for coordinating supply chains (Mendelson and Tunca, 2007). Crude oil prices also play an important role in stock returns (Sadorsky, 1999) as they impact stock market expected returns and cash flows (Jones and Kaul, 1996). They also can lead to financial losses across the business sector (Li et al., 2019c) and provide valuable information for predicting stock market volatility (Kim and Won, 2018). Oil price forecasting is also important for operations related to oil and gas firms, portfolio diversification, and portfolio management (Antonakakis et al., 2018).

However, oil prices are notoriously difficult to predict for a variety of reasons. First, they show a high sensitivity to global demand and supply which might result in high short-term volatility (Beckers and Beidas-Strom, 2015). Additionally, global geopolitical events have a significant influence on the oil price (Monge et al., 2017), which is also strongly affected by politics and market sentiment (Alvarez-Ramirez et al., 2003). To further understand oil prices, its volatility has been explored from different angles, with researchers reporting findings such as volatility spillovers between oil prices and stock sector returns (El Hedi Arouri et al., 2011), and a persistent lead-lag relationship between S&P500 index and expectations for crude oil (Kyrtsou et al., 2016). Finally, trading patterns of investors learning about economic fundamentals, both from public announcements and market prices, contribute to highly volatile oil prices (Singleton, 2014).

Methodologically, oil price prediction is part of financial time series forecasting, where prediction models based on Recurrent Neural Networks (RNN) have emerged as the dominant technique (see Sezer et al. (2020) for an overview). RNNs are easily adapted to multiple forecasting problems and exhibit good prediction performance. However, oil price prediction literature uses models that rely on structured, lagged oil price data. A limited number of papers have considered textual data to predict oil prices (e.g., Li et al. (2019c); as discussed in detail in Section 2.2), leaving a research gap for the use of oil-relevant textual data through complex natural language processing (NLP) techniques.

We source textual content from the twitter communication of former US President Donald Trump. In previous years, much has been discussed about the impact of his tweets on oil prices. Trump himself claims responsibility for low oil prices, but the press is divided on the subject. According to Bloomberg and Forbes, there is at least a short-term drop in prices after an intervention of Trump. Yet CNBC claims oil prices rose after Trump tweeted he "called up" OPEC and told them to bring down costs. A few examples of notable tweets of former President Trump are included on Table 2.1. Regardless, there is evidence that communication on Twitter (and on social media in general) can have a significant impact. For instance, Schmidt et al. (2020) have shown that increased Twitter activity after supply chain glitches is linked to stock market disruptions. Twitter data can also predict stock market fluctuations (Bollen et al., 2011) and improve the performance of supply chain metrics (Maheshwari et al., 2021). In addition, Ilk et al. (2020) use NLP techniques to interpret chat textual data to improve the operational performance of a contact center. Additionally, there is precedent in tweets by Elon Musk increasing Tesla's stock price by over six percent (Craig and Amernic, 2020). Lastly, according to the US Energy Information Administration (EIA), the US is currently the biggest oil producer.

As such, we assume the tweets of the US President may impact the price of oil. This assumption is supported by studies that show world leaders have a significant causal influence on the economy (Jones and Olken, 2005) and they matter for economic growth (Berry and Fowler, 2021).

Table 2.1. Notable oil tweets by former President Trump

Date	Text
September 20, 2018	We protect the countries of the Middle East, they would not be safe for very long without us, and yet they continue to push for higher and higher oil prices! We will remember. The OPEC monopoly must get prices down now!
February 25, 2019	Oil prices getting too high. OPEC, please relax and take it easy. World cannot take a price hike - fragile!
April 2, 2020	Just spoke to my friend MBS (Crown Prince) of Saudi Arabia, who spoke with President Putin of Russia, & I expect & hope that they will be cutting back approximately 10 Million Barrels, and maybe substantially more which, if it happens, will be GREAT for the oil & gas industry!

This study seeks to confirm the predictive power of a US president's tweets for oil prices and to contribute to the overall methodology, by comparing several NLP techniques:

1. a *vector space-based approach* due to its good performance across areas of research (Coussement and Van den Poel, 2008),
2. *embedding models* for their effective adaptability to capture subtle semantic similarities (Hirschberg and Manning, 2015), and
3. a *transformer-based approach* as this architecture is the current state-of-the art technique (Borchert et al., 2022).

The vector space-based approach is implemented using Term Frequency–Inverse Document Frequency (TF-IDF), while the embedding models used in this study are Word2Vec, Doc2Vec, and Global Vectors for word representation (GloVe). The Bidirectional Encoder Representations from Transformers (BERT) is the transformer-based technique mentioned earlier. These techniques are combined with a Long Short-Term Memory (LSTM) architecture. LSTM is able to learn efficiently while holding on to information for long periods (Bengio et al., 1994). It can also be adapted to different tasks and has been successfully applied in different areas of financial time series forecasting (Fischer and Krauss, 2018; Flori and Regoli, 2021), particularly in cases too complex for traditional forecasting methods (Shen and Sun, 2021). Predictive modeling with its time dimension is a valuable actor in the digital space. This study uses a five day lag for both the oil price and the textual Twitter data following the lag order selection criteria (Ivanov and Kilian, 2005). The data includes the daily crude oil price over a period of almost four years, or 1,449 days, plus a total of 16,457 tweets by the 45th US President. The study is validated for Brent and West Texas Intermediate (WTI) blends, using the daily price of a barrel of crude oil as the target feature. We conclude that the tweets of former President Trump have significant predictive power, and that an LSTM architecture with BERT as a NLP technique has the best performance.

Therefore, the contributions of this study are the use of a novel text data source for oil price prediction, the use of an exhaustive array of NLP techniques, including state-of-the-art methodologies which are more powerful for extracting contextual information, plus the incorporation of additional analyses to further elucidate the model's output. Our results show BERT is an excellent NLP technique for extracting relevant information for oil price prediction. Additionally, we confirm previous findings in existing literature, which suggest oil price is influenced by global geopolitical events (Monge et al., 2017), politics and market sentiment (Alvarez-Ramirez et al., 2003), and public announcements (Singleton, 2014). Finally, we perform exclusion and structural break analysis to further back up these conclusions, as well as contributing towards deep learning model explainability. We structure this study through the following research questions (RQs):

RQ1: Can the tweets of the US President be used to better predict oil prices?

RQ2: What is the impact of different NLP techniques on prediction performance?

RQ3: Does LSTM outperform other benchmark forecasting models to predict oil prices?

2.2 Related Work

This section reviews the literature for oil price prediction using deep learning models and discusses the use of textual data with applications in the financial industry. Table 2.2 offers an overview of the type of oil price data, the methods, NLP techniques, the type of textual data, and the evaluation metrics used in the respective studies.

The conclusions are the following. First, it is clear that the WTI and Brent crude oil prices are the most popular oil price data sources. Second, the table lists several methods that are able to handle sequential data (e.g., SVR, VMD, and the autoregressive-based models). Third, only one study has incorporated text as a data source, leaving the impact of various NLP techniques yet to be explored. Thus, textual data from social media still offers opportunities for research in forecasting in the financial sector (Huang et al., 2020), with applications in decision-making and operations management (Chan et al., 2017). Further, other studies within the financial sector show evidence of the benefits of using textual data through state-of-the-art NLP techniques for forecasting, for instance, stock prices (Kraus and Feuerriegel, 2017), business failure (Borchert et al., 2022), and economic trends (Buczkowski, 2017).

Moreover, Sezer et al. (2020) show that LSTM models have dominated the financial time series forecasting area, and can incorporate textual data from financial news and stock market data sources. On the other hand, Nguyen et al. (2022) suggest LSTM as an advanced time series technique to continue improving upon forecasting accuracy. LSTM has also been used in operations for complex tasks such as assembly sequence planning (Wu et al., 2022) and product configuration (Wang et al., 2022), among others. This shows LSTM is a flexible model that can be adapted to multiple tasks and remains a viable opportunity for research, constituting the motivation for our study setup.

This study incorporates the use of a five-day lag, following the lag order selection criteria (Ivanov and Kilian, 2005). However, there seems to be no consensus, as per our literature review. Therefore, additional research remains to verify whether different lag lengths will impact performance in a positive way.

Table 2.2. Literature review: oil price prediction using deep learning models

Study	Oil Data	Method	NLP	Text Data	Evaluation
Azevedo and Campos (2016)	Brent, WTI	ARIMA			MAPE
Mostafa and El-Masry (2016)	WTI	GEP, NN			MAE, MSE, RMSE, R^2
Zhao et al. (2017)	WTI	SDAE			MAPE, RMSE
Sun et al. (2018)	Brent, WTI	INN			ARV, Theil's U
Wang et al. (2018)	WTI	HTW-MBPNN			MAE, MAPE, MSE
Cheng et al. (2019)	Brent, WTI	VEC-NARNN			MAE, MSE, RMSE
Li et al. (2019b)	Brent, WTI	VMD-AI			MAPE, RMSE
Li et al. (2019c)	Brent, WTI	CNN	Sentiment analysis, LDA	Online news	MAE, RMSE
Ramyar and Kianfar (2019)	Brent, WTI	MLP			MSE, R^2, Adjusted R^2
Álvarez-Díaz (2020)	Brent	NARNN			NMSE
Wang et al. (2020)	WTI	LR-SVR-ANN			MAE, MAPE, RMSE
Huang and Deng (2021)	WTI	VMD-LSTM-MW			MAPE, RMSE
He et al. (2022)	WTI	VMD-SVR			MAPE, RMSE
Karasu and Altan (2022)	Brent, WTI	LSTM			MAPE, Theil's U
This study	Brent, WTI	LSTM	BERT, W2V, D2V, GloVe, TF-IDF	Twitter	MAPE, RMSE

AI = Artificial Intelligence, ANN = Artificial Neural Network, ARV = Average Relative Variance, CNN = Convolutional Neural Network, D2V = Doc2Vec, GEP = Gene Expression Programming, HTW–MBPNN = Harr a Trous Wavelet Multilayer Back Propagation Neural Network, ICA = Independent Component Analysis, INN = Interval Neural Network, LDA = Latent Dirichlet Allocation, LR = Logistic Regression, LSTM = Long Short-term Memory, MAE = Mean Absolute Error, MAPE = Mean Absolute Percentage Error, MSE = Mean Squared Error, MW = Moving Window, NARNN = Nonlinear Autoregressive Neural Network, NMSE = Normalized Mean Square Error, RMSE = Root Mean Squared Error, SDAE = Stacked Denoising Autoencoders, SVR = Support Vector Regression, VEC = Vector Error Correction, VMD = Variational Mode Decomposition, W2V = Word2Vec, WTI = West Texas Intermediate crude oil

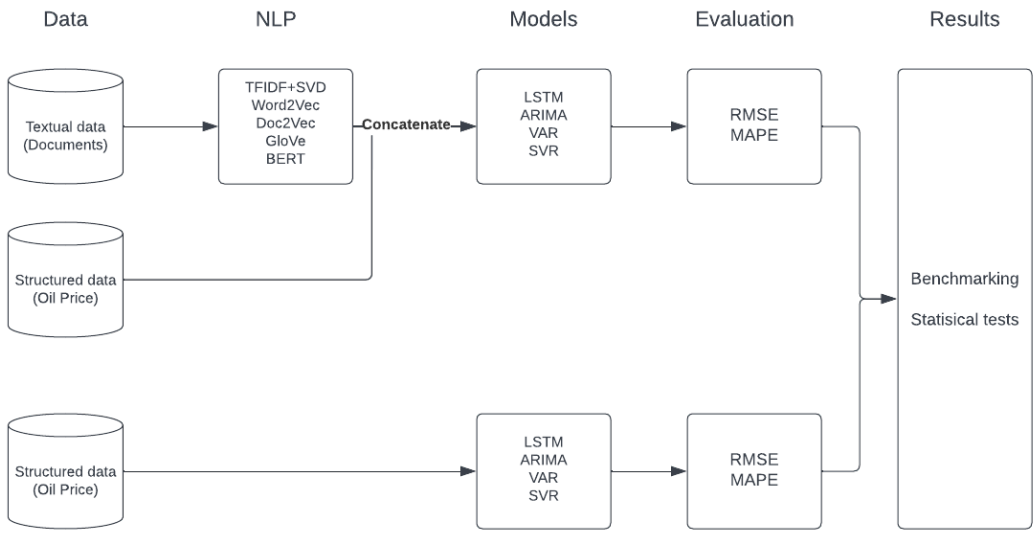


Figure 2.1. Schematic illustration of experimental setup

2.3 Experimental Setup

In this Section, we discuss the experimental setup that is used to: (1) show that the tweets of former President Trump have significant predictive power when predicting WTI and Brent oil prices, and (2) benchmark various models and NLP techniques. Figure 2.1 summarizes the experimental setup. In what follows, we briefly explain the different aspects of the experiment. More details may be found in Appendix.

2.3.1 Data

The data consists of structured oil price data and textual tweet data starting from January 20, 2017, i.e., the beginning of the term of the 45th US President Donald Trump, and ending with his ban from the social media platform, on January 8, 2021. This time span contains trading and non-trading days. The *structured data* is represented by the daily spot price of a barrel of crude oil on trading days, resulting in 1,006 Brent and 1,025 WTI crude oil observations. Non-trading days are treated as missing values with the largest gap of consecutive non-trading days being five days around Christmas 2018.

The *textual data* consists of every tweet the 45th US President has posted during the observed time span, including deleted tweets, but excluding retweeted or reposted tweets from his account @realDonaldTrump. This results in 16,457 tweets, amounting to an average of 11 tweets per day.

The data is split in training, validation, and test sets, as shown in Figure 2.2. Training data contains 60% of the data from January 2017 to early June 2019. Then, the validation data spans an additional 20%, until March 2020. The test set contains the last 20% of the data, until January 2021, remaining unused until final model evaluation and comparisons, as recommended by Granger (1993).

The training and validation sets are used to optimize a given model using lagged oil price data. The best performing model for each approach is then used on the test set, once using only structured data and once concatenating structured and textual data. The

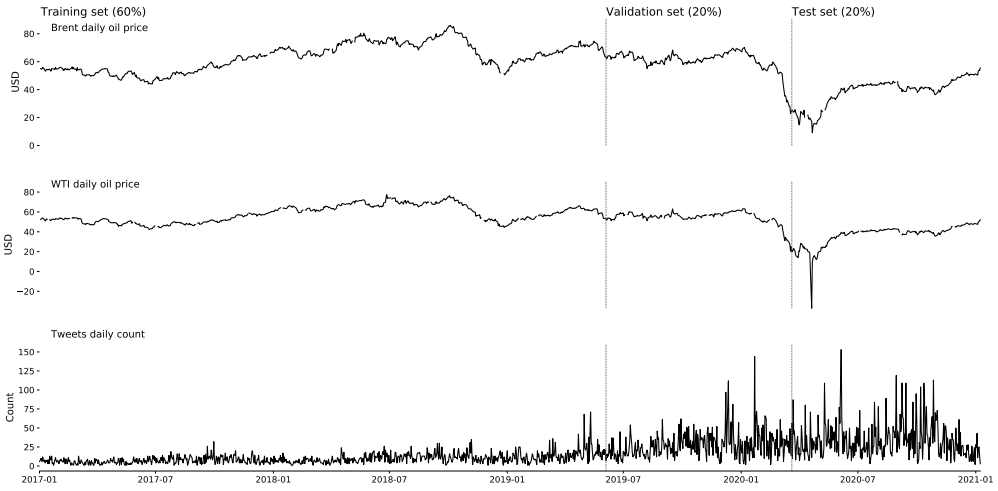


Figure 2.2. Timeline showing how the data is split

results are compared using Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and a post-hoc test, further detailed in Section 2.3.5.

The different types of data can be used together because all NLP techniques transform the text into numerical representations, which can then be concatenated to the structured dataset. To match the number of observations, first the tweets are grouped per day as part of data preprocessing. Each day can contain multiple tweets and is thus referred to as a document henceforth.

2.3.2 NLP techniques

TF-IDF transforms the dataset of documents into a matrix summarizing the relative frequency of each word, losing the order of each sentence contained in the documents. To reduce the matrix size, Singular Value Decompostion (SVD) is implemented (Zhu and Ghodsi, 2006).

GloVe and Word2Vec represent each word, or token, through a list of numbers, also known as numerical vectors. These vectors can be used, for instance, to calculate similarity between words. Using a similar technique as Word2Vec, Doc2Vec instead creates numerical vectors for each document.

BERT uses wordpiece embeddings, preserving more information by further fragmenting a word to specify plurals or verb endings (Wu et al., 2016). BERT's pre-trained embeddings can be accessed and further tuned on a specific task, in this case oil price prediction.

To prepare the data for these techniques, certain preprocessing steps are first performed. Examples are raw text cleaning, case conversion, term filtering, and tokenization. Although the preprocessing is very similar across NLP techniques, there are some differences depending on the dimensionality and contextual information that each approach can handle. Further explanation of the preprocessing steps for each NLP technique is included in Section 2.8.1 (Appendix A.2).

2.3.3 LSTM

LSTM is a neural network which is particularly apt at using sequential data for predictive tasks. LSTM requires passing a succession of ordered data as input, where each element is the data of a given timestep. In this study, five consecutive days of data are used for predicting the oil price of the sixth day, each day being a timestep. More details about LSTM are included in Section 2.8.1 (Appendix A.3).

2.3.4 Benchmark Models

As benchmark models, we use Auto Regressive Integrated Moving Average (ARIMA), a Vector Auto-Regression (VAR), and Support Vector Regression (SVR). **ARIMA** has been shown to yield excellent performance when forecasting oil prices (Azevedo and Campos, 2016). It has also been used with exogenous features (ARIMAX) for oil price prediction (Elshendy et al., 2018), where, in our case, the exogenous features correspond to textual features. A non-seasonal ARIMA(p, d, q) model has parameters for the aspects it accounts for, with p referring to the number of lagged observations included, d being given by the differencing order between the observations, and q by the size of the moving average window (Hyndman and Khandakar, 2008). **VAR** is included as a benchmark, because it yields accurate predictions of oil-related KPIs (Allegret et al., 2015) and has been used before to leverage information both from textual and structured data (Nguyen et al., 2022). Its ability to capture the relationships between multiple time series has also been considered a strength, and it is a popular benchmark for oil price prediction (Ramyar and Kianfar, 2019). **SVR** has also been successfully used as a benchmark in oil price prediction (Ribeiro and dos Santos Coelho, 2020). Further details for each model are included in Section 2.8.1.

2.3.5 Evaluation Metrics and Statistical Tests

Choosing the metric for evaluation is not a straightforward task, as all of them have advantages and disadvantages. For example, MAE, being in the same unit as the target variable, is straightforward to grasp and is robust against outliers. However, its failure to penalize large errors proportionately to small ones might distort the model's true accuracy. Additionally, it is not differentiable at zero, posing challenges for optimization with gradient-based methods like LSTM.

In contrast, MSE penalizes large errors more, offering a better reflection of a model's true accuracy. It's differentiable across its domain, allowing for optimization with gradient-based techniques. However, MSE's drawbacks include sensitivity to outliers and a unit mismatch with the target variable, making comparisons and interpretation harder.

RMSE combines advantages from both MAE and MSE, such as sharing the target variable's unit and allowing gradient-based optimization. It penalizes large errors more than MSE does, but it remains sensitive to outliers. Thus, RMSE strikes a balance between the robustness of MAE and the accuracy reflection of MSE, making it a favorable choice in many scenarios.

MAPE provides a percentage measure of how inaccurate the forecast is on average. It gives an indication of the size of the error relative to the actual value being forecasted. This makes it a widely used metric, due to its simplicity and intuitive interpretation, despite not being a reliable accuracy indicator. In fact, among its limitations are that it can

be heavily influenced by extreme values, particularly when the actual values are close to zero. Additionally, MAPE cannot be used when actual values are zero or close to zero, as it involves division by the actual value, leading to undefined or infinite values. Therefore, we optimize for RMSE, adding MAPE as a complementary measure for analysis.

RMSE and MAPE have been frequently used to evaluate energy price forecasting performance (Lu et al., 2021), and are selected due to their ease of interpretation. RMSE is on the same units as the target feature and is prioritized when comparing results. RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}.$$

MAPE is expressed as a percentage and is defined as follows:

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.$$

For both metrics, y_t and $\hat{y}_t$ represent respectively the actual and the predicted values, with N being the number of predicted values.

The Diebold-Mariano (DM) test is used as a post-hoc test to confirm the results (Diebold and Mariano, 2002). This test is frequently used to verify statistically significant forecasting results (Zhao et al., 2017). It is also the best-known approach to establish differences exist between forecasting method results (De Gooijer and Hyndman, 2006).

The original DM statistic was developed for comparing forecasts from two different models (Diebold and Mariano, 2002). It is an asymptotic z-test, with the null hypothesis being that the expected loss differential is zero (Diebold, 2015). Having a model i and a model j for a time series $\{y_t\}_1^T$, this can be expressed as follows:

$$H_0 : E(d_{ijt}) = E(e_{it} - e_{jt}) = 0,$$

where d_{ijt} is a series of loss differential at timestep t, e_{it} are residuals from forecast i, and e_{jt} are residuals from forecast j (Hyndman and Khandakar, 2008). From this, the sample mean of the loss differential, $\bar{d}$, can be denoted as:

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$$

Finally, the original DM statistic, proposed to test equal forecast accuracy as the null hypothesis, is expressed as:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}},$$

where $\hat{V}$ corresponds to the asymptotically estimated variance for $\bar{d}$. However, Harvey et al. (1997) propose a modified DM statistic to account for small dataset sizes, that is calculated as follows:

$$DM^* = DM\sqrt{\frac{t + 1 - 2h + t^{-1}h(h-1)}{t}},$$

where h refers to the forecast horizon ahead. In this paper, we use Equation 2.3.5 for all DM statistic results. Finally, sequential Holm-Bonferroni correction (Holm, 1979) is applied for an overall family-wise error rate equal or lower to a threshold α.

2.4 Results

The RMSE and MAPE of the LSTM models are summarized in Table 2.3. First, we compare the LSTM model that only uses structured data as input ($LSTM_{Struct}$) and the LSTM models that incorporate structured and textual data (the subscripts refer to the NLP technique used). We can observe that it is always beneficial to include textual data if we consider the RMSE, regardless of the method that was used to represent textual data, and for both Brent and WTI oil prices. Also if we look at MAPE, this finding holds and models that include textual data outperform the LSTM model without textual data. We additionally follow Zhao et al. (2017), reaching the same conclusions using directional accuracy (DA), as presented in Appendix B.

Table 2.3. Performance of LSTM models, comparing structured data to LSTM models including both structured and textual data

Model	Brent Oil Price		WTI Oil Price	
	RMSE	MAPE	RMSE	MAPE
$LSTM_{Struct}$	5.389	2.516	4.132	3.528
$LSTM_{TFIDF}$	5.206	2.411	4.121	2.730
$LSTM_{GloVe}$	3.146	2.428	3.145	2.545
$LSTM_{W2V}$	3.041	2.435	3.020	2.594
$LSTM_{D2V}$	3.072	2.486	3.029	2.620
$LSTM_{BERT}$	**2.989**	**1.370**	**2.997**	**1.526**

To verify whether the predictions of the different models differ significantly, we perform a modified DM test. The results of this test are summarized in Table 2.4. For both Brent and WTI oil prices, Table 2.4 lists the result of the modified DM test when comparing model i and model j, with $i, j \in$ {BERT, W2V, D2V, GloVe, TFIDF} and i $\neq$ j.

Table 2.4. Modified DM test p-values for LSTM models with structured and textual data

		Model j				
Oil	Model i	$LSTM_{Struct}$	$LSTM_{TFIDF}$	$LSTM_{GloVe}$	$LSTM_{D2V}$	$LSTM_{W2V}$
Brent	$LSTM_{BERT}$	0.000***	0.000***	0.000***	0.000***	0.000***
	$LSTM_{W2V}$	0.000***	0.002**	0.002**	0.003**	-
	$LSTM_{D2V}$	0.001***	0.004**	0.004**	-	-
	$LSTM_{GloVe}$	0.038*	0.046*	-	-	-
	$LSTM_{TFIDF}$	0.008**	-	-	-	-
WTI	$LSTM_{BERT}$	0.004**	0.004**	0.000***	0.000***	0.000***
	$LSTM_{W2V}$	0.006**	0.006**	0.000***	0.004**	-
	$LSTM_{D2V}$	0.007**	0.007**	0.000***	-	-
	$LSTM_{GloVe}$	0.039*	0.048*	-	-	-
	$LSTM_{TFIDF}$	0.038*	-	-	-	-

***p-value<.001, **p-value<.01, *p-value<.05.

For both Brent and WTI oil prices, Table 2.3 and Table 2.4 show that all LSTM models incorporating NLP significantly outperform the LSTM model using only structured data.

Hence, we can conclude that the tweets of US President Donald Trump help to better predict the Brent and WTI oil prices.

Next, we compare the impact, on the predictive performance, of the five NLP techniques with each other. The results show that $\text{LSTM}_{\text{BERT}}$ performs significantly better than all other LSTM models. Within models using text features, LSTM_{W2V} significantly outperforms LSTM_{D2V}, $\text{LSTM}_{\text{GloVe}}$, and $\text{LSTM}_{\text{TFIDF}}$. LSTM_{D2V} significantly performs above $\text{LSTM}_{\text{GloVe}}$ and $\text{LSTM}_{\text{TFIDF}}$. $\text{LSTM}_{\text{GloVe}}$ only outperforms $\text{LSTM}_{\text{TFIDF}}$, suggesting other techniques may capture more contextually valuable information for this task. Lastly, the vector-space based approach ($\text{LSTM}_{\text{TFIDF}}$) consistently has the worse performance, hinting that the order of words may add value to the predictive performance.

The best LSTM model is benchmarked against ARIMA(X), VAR, and SVR models. The results of this benchmark are presented in Table 2.6. Table 2.5 provides an overview of the predictive performance of the traditional forecasting benchmarks using only structured data (having subscript Struct) and structured combined with textual data (having subscript Text). Furthermore, the LSTM model with significantly best performance, $\text{LSTM}_{\text{BERT}}$, is added as a reference. From Table 2.5 and Table 2.6, it can be seen that among the benchmark models, SVR shows an improvement in performance when incorporating textual data as features, for both Brent and WTI oil prices, which is in line with our findings above. On the other hand, both autoregressive models, which include ARIMA models and VAR, show a decrease in performance when adding textual features, when compared to the use of only structured data. Overall, SVR including textual is the best performing benchmark model, but all benchmark models are outperformed by $\text{LSTM}_{\text{BERT}}$, suggesting LSTM is better at extracting meaningful information for oil price prediction from the concatenated textual and structured data.

Table 2.5. Performance of benchmark models & $\text{LSTM}_{\text{BERT}}$

Model	Brent Oil Price		WTI Oil Price	
	RMSE	MAPE	RMSE	MAPE
$\text{ARIMA}_{\text{Struct}}$	5.768	4.539	5.913	4.651
$\text{ARIMAX}_{\text{Text}}$	5.962	4.672	5.989	4.962
$\text{VAR}_{\text{Struct}}$	5.789	4.678	5.981	4.794
VAR_{Text}	5.936	4.764	5.983	4.731
$\text{SVR}_{\text{Struct}}$	5.784	4.546	5.327	4.876
SVR_{Text}	5.747	4.528	5.297	4.762
$\text{LSTM}_{\text{BERT}}$	**2.989**	**1.370**	**2.997**	**1.526**

Table 2.6. Modified DM test p-values from comparing $\text{LSTM}_{\text{BERT}}$, the best performing LSTM model, to the benchmark models

	Model j					
Oil	$\text{ARIMA}_{\text{Struct}}$	$\text{ARIMAX}_{\text{Text}}$	$\text{VAR}_{\text{Struct}}$	VAR_{Text}	$\text{SVR}_{\text{Struct}}$	SVR_{Text}
Brent	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
WTI	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***

***p-value<.001, **p-value<.01, *p-value<.05.

2.5 Discussion

This section further discusses additional insights regarding the impact of the textual data through a brief review of oil-related keywords, as outlined by Hyne (2015). To assess the impact of these keywords, the best performing model is reran twice, first after removing oil-related keywords from the textual data (partial exclusion), and a second time after excluding all tweets containing any oil-related keyword (full exclusion). The results are displayed on Table 2.7.

Table 2.7. $\text{LSTM}_{\text{BERT}}$ comparison with oil-related keyword exclusion

Text usage	Brent Oil Price		WTI Oil Price	
Full text	2.989	1.370	2.997	1.526
Partial exclusion	3.009	2.376	3.000	2.534
Full exclusion	3.198	2.507	3.158	2.546

Further, we perform a modified DM test to evaluate the significance of the results. The results of this test are summarized in Table 2.8. The test reveals there are statistically significant differences for both exclusion levels, with 95% confidence level. This is in line with our previously discussed findings, regarding $\text{LSTM}_{\text{BERT}}$ being able to capture relevant contextual information from the text documents included as data. Thus, the

Table 2.8. DM test p-values for comparing $\text{LSTM}_{\text{BERT}}$ with the full text against different variations of oil-related keywords exclusion

	Model j	
Oil	Partial Exclusion	Full Exclusion
Brent	0.049*	0.043*
WTI	0.048*	0.041*

***p-value<.001, **p-value<.01, *p-value<.05.

results show exclusion does affect the model performance, with the removal of oil-related keywords marginally affecting the results, while fully removing tweets with oil-related keywords performs the lowest for both Brent and WTI oil prices. Around 23% of tweets contain oil-related keywords, with several being used multiple times. Examples are 62 uses of the word *oil*, 12 uses of *OPEC*, 35 uses of *gas*, 30 uses of *Saudi Arabia*, and 2 uses of *commodities*. Although these frequencies are not enough to position these keywords among the top tokens, displayed in Figure 2.3, the top tokens can also be described as related to geopolitical events. Thus, the impact on performance likely can be explained beyond the presence of oil-related keywords.

To further explore text beyond oil-related terms, we perform a structural change analysis of both Brent and WTI oil prices (Zeileis et al., 2002). The objective of this analysis is to detect any deviations from a stable classical linear regression model. This is implemented on the oil price data, by assuming it follows a classical linear regression model ($y_i = x_i^\top \beta + u_i$). After a structural change, the coefficients for a linear regression model change. Therefore, m structural changes results in $m + 1$ segments with stable regression. Thus, for a segment j, the approach with the model can be rewritten as:

$$y_i = x_i^\top \beta_j + u_i \qquad (i = i_{j-1} + 1, ..., i_j, \quad j = 1, ..., m + 1),$$

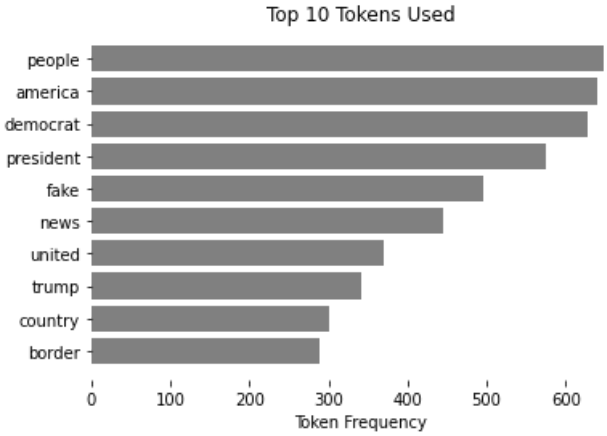


Figure 2.3. Top 10 token frequency from Donald Trump's twitter account

with the number of structural changes being estimated by minimizing the residual sum of squares (RSS) for the previous equation (Zeileis et al., 2002).

For each structural change, the top five words present in the previous five days are examined; these words are expected to contribute most to the deviation from the linear regression model. In our data, we detected structural change at several points in time (also referred to as "breakpoints"). These breakpoints are displayed in Figure 2.4.

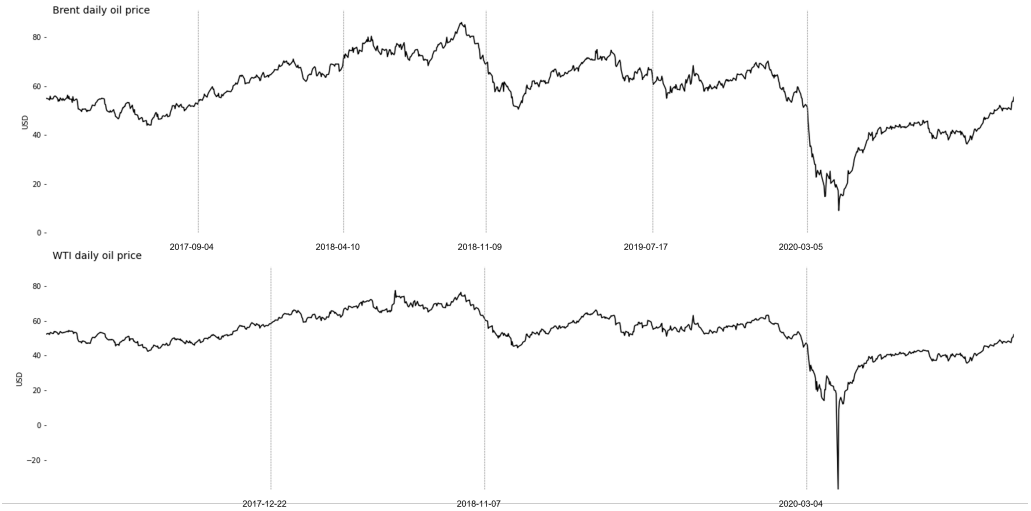


Figure 2.4. Timeline of Brent and WTI oil prices with breakpoints resulting from structural change analysis

In addition, Figure 2.5 displays the top five words for each of the detected breakpoints.

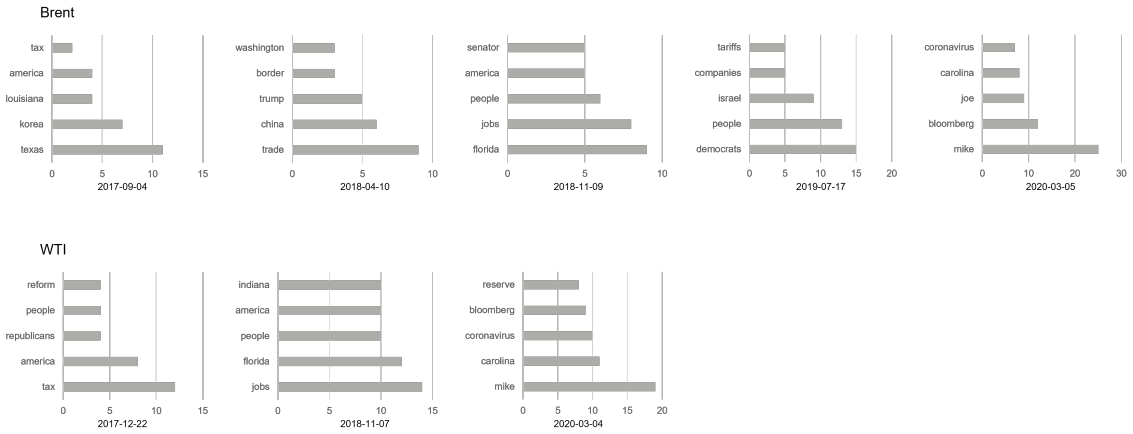


Figure 2.5. Breakpoints from structural change analysis

Although the number of breakpoints is different for Brent and WTI oil price datasets, a similar pattern emerges regarding the presence of keywords relevant to geopolitical events. Particular examples are the mention of taxes and tariffs, different countries, and even the coronavirus. Certain first names and state names appear within the top words as well, indicating the occurrence of political events such as elections or election campaigns. These exploratory findings are in line with previous literature stating oil price is influenced by global geopolitical events (Monge et al., 2017), politics and market sentiment (Alvarez-Ramirez et al., 2003), and public announcements (Singleton, 2014). Additional research, however, can still be done. For instance, it would be interesting to see which text features are better at predicting the direction of oil price.

2.6 Conclusions and Future Research

Oil price prediction remains a challenging task and is an ongoing area of research. Extant literature has investigated a variety of model architectures and combinations of structured data, as discussed in Section 2.2. This study extends the oil price prediction field by investigating novel data enrichment possibilities through the incorporation of textual data. In particular, we investigated the added value of tweets from the 45th US President for predicting the daily barrel price of Brent and WTI crude oil. The use of global leaders' voices as sources of forecasting data within the realm of financial-related deep learning is still in its infancy. This study adds textual data into a LSTM model, while comparing various NLP techniques including the vector-space approach (TF-IDF), embeddings (Word2Vec, Doc2Vec, GloVe), and BERT as a transformer-based method to answer the following RQs:

RQ1: Can the tweets of the US President be used to better predict oil prices?

RQ2: What is the impact of different NLP techniques on prediction performance?

RQ3: Does LSTM outperform other benchmark forecasting models to predict oil prices?

Our results indicate that adding the US President's tweets to the structured lagged oil price data leads to superior oil price prediction performance. Furthermore, the LSTM models outperform benchmark models like ARIMA, VAR, and SVR, thus being the most accurate model for oil price prediction, both with and without the incorporation of textual data. Finally, our results indicate that BERT is the best NLP technique for both Brent

and WTI crude oil price prediction.

Moreover, this study provides a framework to combine structured and textual data, using different textual techniques. Possibilities for future research include the application of this framework by examining additional sources of relevant textual data, such as the information available in the OPEC press room website, containing news items, press releases, and official speeches delivered by different members of the OPEC.

Lastly, this study also fosters model explainability by providing insights into what textual features are important for oil price prediction. Specifically, we find that the removal of oil-related keywords and tweets affect predictive performance. Furthermore, exploratory research in the form of structural change analysis reveals the presence of keywords related to geopolitical events before deviations from a stable classical linear regression model.

Additional research can yet be accomplished through different ways. First, through comparing different methods for explaining deep learning model results, such as attribution. Second, by incorporating an attention layer into the model architecture, for clearer visualization of feature importance. Third, by altering the current oil price prediction target into a binary feature, such as price direction, to clarify how textual data interacts with oil prices. These changes may reveal complementary insights into how LSTM enhances performance using text, in the context of oil price prediction.

2.7 References

Allegret, J.P., Mignon, V., Sallenave, A., 2015. Oil price shocks and global imbalances: Lessons from a model with trade and financial interdependencies. Econ. Model. 49, 232–247. doi:10.1016/j.econmod.2015.04.009.

Álvarez-Díaz, M., 2020. Is it possible to accurately forecast the evolution of Brent crude oil prices? An answer based on parametric and nonparametric forecasting methods. Empir. Econ. 59, 1285–1305. doi:10.1007/s00181-019-01665-w.

Alvarez-Ramirez, J., Soriano, A., Cisneros, M., Suarez, R., 2003. Symmetry/anti-symmetry phase transitions in crude oil markets. Phys. A Stat. Mech. its Appl. 322, 583–596. doi:10.1016/S0378-4371(02)01831-9.

Antonakakis, N., Cunado, J., Filis, G., Gabauer, D., Perez de Gracia, F., 2018. Oil volatility, oil and gas firms and portfolio diversification. Energy Econ. 70, 499–515. doi:10.1016/j.eneco.2018.01.023.

Azevedo, V.G., Campos, L.M., 2016. Combination of forecasts for the price of crude oil on the spot market. Int. J. Prod. Res. 54, 5219–5235. doi:10.1080/00207543.2016.1162340.

Beckers, B., Beidas-Strom, S., 2015. Forecasting the Nominal Brent Oil Price with VARs-One Model Fits All? IMF Work. Pap. 2015, 261–266. doi:10.5089/9781513524276.001.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Trans. Neural Networks 5, 157–166. doi:10.1109/72.279181.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inf. Sci. 191, 192–213. doi:10.1016/j.ins.2011.12.028.

Berry, C.R., Fowler, A., 2021. Leadership or luck? Randomization inference for leader effects in politics, business, and sports. Sci. Adv. 7, eabe3404. doi:10.1126/sciadv.abe3404.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. J. Comput. Sci. 2, 1–8. doi:10.1016/j.jocs.2010.12.007.

Borchert, P., Coussement, K., De Caigny, A., De Weerdt, J., 2022. Extending business failure prediction models with textual website content using deep learning. Eur. J. Oper. Res. doi:10.1016/j.ejor.2022.06.060.

Brown, B., . Trump Twitter Archive. URL: http://www.trumptwitterarchive.com/.

Buczkowski, P., 2017. Predicting stock trends based on expert recommendations using GRU/LSTM neural networks, in: Lect. Notes Comput. Sci., pp. 708–717. doi:10.1007/978-3-319-60438-1_69.

Chan, H.K., Lacka, E., Yee, R.W., Lim, M.K., 2017. The role of social media data in operations and production management. Int. J. Prod. Res. 55, 5027–5036. doi:10.1080/00207543.2015.1053998.

Cheng, F., Li, T., Wei, Y.m., Fan, T., 2019. The VEC-NAR model for short-term forecasting of oil prices. Energy Econ. 78, 656–667. doi:10.1016/j.eneco.2017.12.035.

Chollet, F., et al., 2015. Keras. Accessed June, 2022. URL: https://keras.io.

Coussement, K., Van den Poel, D., 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. Decis. Support Syst. 44, 870–882. doi:10.1016/j.dss.2007.10.010.

Craig, R., Amernic, J., 2020. Benefits and pitfalls of a CEO's personal Twitter messaging. Strateg. Leadersh. 48, 43–48. doi:10.1108/SL-10-2019-0154.

Dai, A.M., Olah, C., Le, Q.V., 2015. Document embedding with paragraph vectors. CoRR abs/1507.07998. doi:10.48550/arXiv.1507.07998.

De Gooijer, J.G., Hyndman, R.J., 2006. 25 years of time series forecasting. Int. J. Forecast. 22, 443–473. doi:10.1016/j.ijforecast.2006.01.001.

Denning, L., 2020. Trump's 10 million barrel tweet is performance art. Bloomberg URL: https://www.bloomberg.com/opinion/articles/2020-04-02/trump-s-10-million-oil-barrel-tweet-is-performance-art.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR doi:10.48550/arXiv.1810.04805.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., p. 4171–4186. doi:10.48550/arXiv.1810.04805.

Diebold, F.X., 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. J. Bus. Econ. Stat. 33, 1–1. doi:10.1080/07350015.2014.983236.

Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. J. Bus. Econ. Stat. 20, 134–144. doi:10.1198/073500102753410444.

El Hedi Arouri, M., Jouini, J., Nguyen, D.K., 2011. Volatility spillovers between oil prices and stock sector returns: Implications for portfolio management. J. Int. Money Financ. 30, 1387–1405. doi:10.1016/j.jimonfin.2011.07.008.

Elshendy, M., Fronzetti Colladon, A., Battistoni, E., Gloor, P.A., 2018. Using four different online media sources to forecast the crude oil price. J. Inf. Sci. 44, 408–421. doi:10.1177/0165551517698298.

Fischer, T., Krauss, C., 2018. Deep learning with long short-term memory networks for financial market predictions. Eur. J. Oper. Res. 270, 654–669. doi:10.1016/j.ejor.2017.11.054.

Flori, A., Regoli, D., 2021. Revealing Pairs-trading opportunities with long short-term memory networks. Eur. J. Oper. Res. 295, 772–791. doi:10.1016/j.ejor.2021.03.009.

Granger, C.W., 1993. Strategies for Modelling Nonlinear Time-Series Relationships. Econ. Rec. 69, 233–238. doi:10.1111/j.1475-4932.1993.tb02103.x.

Graves, A., 2013. Generating sequences with recurrent neural networks. CoRR abs/1308.0850. doi:10.48550/arXiv.1308.0850.

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. Int. J. Forecast. 13, 281–291. doi:10.1016/S0169-2070(96)00719-4.

He, H., Sun, M., Li, X., Mensah, I.A., 2022. A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features. Energy 244, Part A, 122706. doi:10.1016/j.energy.2021.122706.

Heath, D., 2019. Macroeconomic factors in oil futures markets. Manage. Sci. 65, 3949–4450. doi:10.1287/mnsc.2017.3008.

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349, 261–266. doi:10.1126/science.aaa8685.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Huang, S., Potter, A., Eyers, D., 2020. Social media in operations and supply chain management: State-of-the-art and research directions. Int. J. Prod. Res. 58, 1893–1925. doi:10.1080/00207543.2019.1702228.

Huang, Y., Deng, Y., 2021. A new crude oil price forecasting model based on variational mode decomposition. Knowl.-Based Syst. 213, 106669. doi:10.1016/j.knosys.2020.106669.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. J. Stat. Softw. 27, 1–22. doi:10.18637/jss.v027.i03.

Hyne, N.J., 2015. Dictionary of petroleum exploration, drilling & production. Choice Rev. Online doi:10.5860/choice.187346.

Ilk, N., Shang, G., Goes, P., 2020. Improving customer routing in contact centers: An automated triage design based on text analytics. J. Oper. Manag. 66, 553–577. doi:10.1002/joom.1084.

Imbert, F., Stevens, P., 2020. Oil surges 24% for best day on record after trump tells cnbc saudis, russia reach agreement. CNBC URL: https://www.cnbc.com/2020/04/02/oil-rallies-10percent-after-trump-says-he-expects-saudi-arabia-russia-feud-to-end-soon.html.

Ivanov, V., Kilian, L., 2005. A practitioner's guide to lag order selection for VAR impulse response analysis. Stud. Nonlinear Dyn. Econom. 9, 1–34. doi:10.2202/1558-3708.1219.

Jones, B.F., Olken, B.A., 2005. Do leaders matter? National leadership and growth since world war II. Q. J. Econ. 120, 835–864. doi:10.1093/qje/120.3.835.

Jones, C.M., Kaul, G., 1996. Oil and the Stock Markets. J. Finance 51, 463–491. doi:10.2307/2329368.

Karasu, S., Altan, A., 2022. Crude oil time series prediction model based on lstm network with chaotic henry gas solubility optimization. Energy 242, 122964. doi:10.1016/j.energy.2021.122964.

Kim, H.Y., Won, C.H., 2018. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Syst. Appl. 103, 25–37. doi:10.1016/j.eswa.2018.03.002.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015. doi:10.48550/arXiv.1412.6980.

Koopman, S.J., 1997. Exact initial kalman filtering and smoothing for nonstationary time series models. J. Am. Stat. Assoc. 92, 1630–1638. doi:10.1080/01621459.1997.10473685.

Kraus, M., Feuerriegel, S., 2017. Decision support from financial disclosures with deep neural networks and transfer learning. Decis. Support Syst. 104, 38–48. doi:10.1016/j.dss.2017.10.001.

Kraus, M., Feuerriegel, S., Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. Eur. J. Oper. Res. 281, 628–641. doi:10.1016/j.ejor.2019.09.018.

Kyrtsou, C., Mikropoulou, C., Papana, A., 2016. Does the S&P500 index lead the crude oil dynamics? A complexity-based approach. Energy Econ. 56, 239–246. doi:10.1016/j.eneco.2016.02.001.

Lau, J.H., Baldwin, T., 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, in: Proceedings of the 1st Workshop on Representation Learning for NLP, p. 78–86. doi:10.18653/v1/w16-1609.

Le, Q.V., Mikolov, T., 2014. Distributed representations of sentences and documents, in: 31st Int. Conf. Mach. Learn. ICML 2014, pp. 1188–1196. doi:10.48550/arXiv.1405.4053.

Li, B., Drozd, A., Guo, Y., Liu, T., Matsuoka, S., Du, X., 2019a. Scaling Word2Vec on Big Corpus. Data Sci. Eng. 4, 157–175. doi:10.1007/s41019-019-0096-6.

Li, J., Zhu, S., Wu, Q., 2019b. Monthly crude oil spot price forecasting using variational mode decomposition. Energy Econ. 83, 240–253. doi:10.1016/j.eneco.2019.07.009.

Li, X., Shang, W., Wang, S., 2019c. Text-based crude oil price forecasting: A deep learning approach. Int. J. Forecast. 35, 1548–1560. doi:10.1016/j.ijforecast.2018.07.006.

Li, X., Wu, P., Wang, W., 2020. Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Inf. Process. Manag. 57, 102212. doi:10.1016/j.ipm.2020.102212.

Loshchilov, I., Hutter, F., 2017. Fixing weight decay regularization in adam. CoRR abs/1711.05101. doi:10.48550/arXiv.1711.05101.

Lu, H., Ma, X., Ma, M., Zhu, S., 2021. Energy price prediction using data-driven models: A decade review. Comput. Sci. Rev. 39, 100356. doi:10.1016/j.cosrev.2020.100356.

Lynch, M., 2019. Why do tweets from trump move oil prices? Forbes URL: https://www.forbes.com/sites/michaellynch/2019/04/28/why-do-tweets-from-trump-move-oil-prices.

Maheshwari, S., Gautam, P., Jaggi, C.K., 2021. Role of big data analytics in supply chain management: current trends and future perspectives. Int. J. Prod. Res. 59, 1875–1900. doi:10.1080/00207543.2020.1793011.

Maldonado, S., González, A., Crone, S., 2019. Automatic time series analysis for electric load forecasting via support vector regression. Appl. Soft Comput. J. 83, 105616. doi:10.1016/j.asoc.2019.105616.

Mendelson, H., Tunca, T.I., 2007. Strategic spot trading in supply chains. Manage. Sci. 53, 742–759. doi:10.1287/mnsc.1060.0649.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. CoRR doi:10.48550/arXiv.1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013b. Distributed Representations of Words and Phrases and their Compositionality. CoRR abs/1310.4546. doi:10.48550/arXiv.1310.4546.

Monge, M., Gil-Alana, L.A., Pérez de Gracia, F., 2017. Crude oil price behaviour before and after military conflicts and geopolitical events. Energy 120, 79–91. doi:10.1016/j.energy.2016.12.102.

Mostafa, M.M., El-Masry, A.A., 2016. Oil price forecasting using gene expression programming and artificial neural networks. Econ. Model. 54, 40–53. doi:10.1016/j.econmod.2015.12.014.

Nguyen, A., Pellerin, R., Lamouri, S., Lekens, B., 2022. Int. J. Prod. Res. 0, 1–12. doi:10.1080/00207543.2022.2070044.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. doi:10.48550/arXiv.1201.0490.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global vectors for word representation, in: EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 1532–1543. doi:10.3115/v1/d14-1162.

Porter, M.F., 1980. An algorithm for suffix stripping. Program 14, 130–137. doi:10.1108/eb046814.

Ramyar, S., Kianfar, F., 2019. Forecasting Crude Oil Prices: A Comparison Between Artificial Neural Networks and Vector Autoregressive Models. Comput. Econ. 53, 743–761. doi:10.1007/s10614-017-9764-7.

Rehurek, R., Sojka, P., 2011. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3, 2.

Ribeiro, M.H.D.M., dos Santos Coelho, L., 2020. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl. Soft Comput. 86, 105837. doi:10.1016/j.asoc.2019.105837.

Sadorsky, P., 1999. Oil price shocks and stock market activity. Energy Econ. 21, 449–469. doi:10.1016/S0140-9883(99)00020-1.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523. doi:10.1016/0306-4573(88)90021-0.

Schmidt, C.G., Wuttke, D.A., Ball, G.P., Heese, H.S., 2020. Does social media elevate supply chain importance? an empirical examination of supply chain glitches, twitter reactions, and stock market returns. J. Oper. Manag. 66, 646–669. doi:10.1002/joom.1087.

Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M., 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Appl. Soft Comput. J. 90, 106181. doi:10.1016/j.asoc.2020.106181.

Shen, Z.M., Sun, Y., 2021. Strengthening supply chain resilience during covid-19: A case study of jd.com. J. Oper. Manag. doi:10.1002/joom.1161.

Simchi-Levi, D., Nelson, D., Mulani, N., Wright, J., 2008a. Crude calculations – why high oil prices are upending the way companies should manage their supply chains. The Wall Street Journal URL: https://www.wsj.com/articles/SB122160061166044841.

Simchi-Levi, D., Nelson, D., Mulani, N., Wright, J., 2008b. The impact of oil price on supply chain strategies: From static to dynamic. Technical Report. Massachusetts Institute of Technology.

Singleton, K.J., 2014. Investor flows and the 2008 boom/bust in oil prices. Manage. Sci. 60, 300–318. doi:10.1287/mnsc.2013.1756.

Sun, S., Sun, Y., Wang, S., Wei, Y., 2018. Interval decomposition ensemble approach for crude oil price forecasting. Energy Econ. 76, 274–287. doi:10.1016/j.eneco.2018.10.015.

US Energy Information Administration, EIA. International Energy Statistics, total oil (petroleum and other liquids) production. Accessed June 21, 2022, https://www.eia.gov/international/data/world/petroleum-and-other-liquids/annual-petroleum-and-other-liquids-production. .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Curran Associates Inc., New York, NY, USA. p. 6000–6010.

Wang, J., Athanasopoulos, G., Hyndman, R.J., Wang, S., 2018. Crude oil price forecasting based on internet concern using an extreme learning machine. Int. J. Forecast. 34, 665–677. doi:10.1016/j.ijforecast.2018.03.009.

Wang, J., Zhou, H., Hong, T., Li, X., Wang, S., 2020. A multi-granularity heterogeneous combination approach to crude oil price forecasting. Energy Econ. 91, 104790. doi:10.1016/j.eneco.2020.104790.

Wang, Y., Li, X., Zhang, L.L., Mo, D., 2022. Configuring products with natural language: a simple yet effective approach based on text embeddings and multilayer perceptron. Int. J. Prod. Res. 60, 5394–5406. doi:10.1080/00207543.2021.1957508.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J., 2019. Huggingface's transformers: State-of-the-art natural language processing. CoRR abs/1910.03771. doi:10.48550/arXiv.1910.03771.

Wu, W., Huang, Z., Zeng, J., Fan, K., 2022. A decision-making method for assembly sequence planning with dynamic resources. Int. J. Prod. Res. 60, 4797–4816. doi:10.1080/00207543.2021.1937748.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR doi:10.48550/ARXIV.1609.08144.

Zeileis, A., Leisch, F., Hornik, K., Kleiber, C., 2002. strucchange: An r package for testing for structural change in linear regression models. J. Stat. Softw. 7, 1–38. doi:10.18637/jss.v007.i02.

Zhao, Y., Li, J., Yu, L., 2017. A deep learning ensemble approach for crude oil price forecasting. Energy Econ. 66, 9–16. doi:10.1016/j.eneco.2017.05.023.

Zhu, M., Ghodsi, A., 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. Comput. Stat. Data Anal. 51, 918–930. doi:10.1016/j.csda.2005.09.010.

2.8 Appendix

2.8.1 Appendix A. Methodology

This section provides an overview of the methodology. Figure 2.6 visualizes the main components of the focal model's system design, which contains four subsequent steps, i.e., (i) the raw data retrieval step for the structured and textual data (or tweets), (ii) the preprocessing step, (iii) the incorporation of NLP techniques for the textual data, and (iv) the modeling and evaluation step.

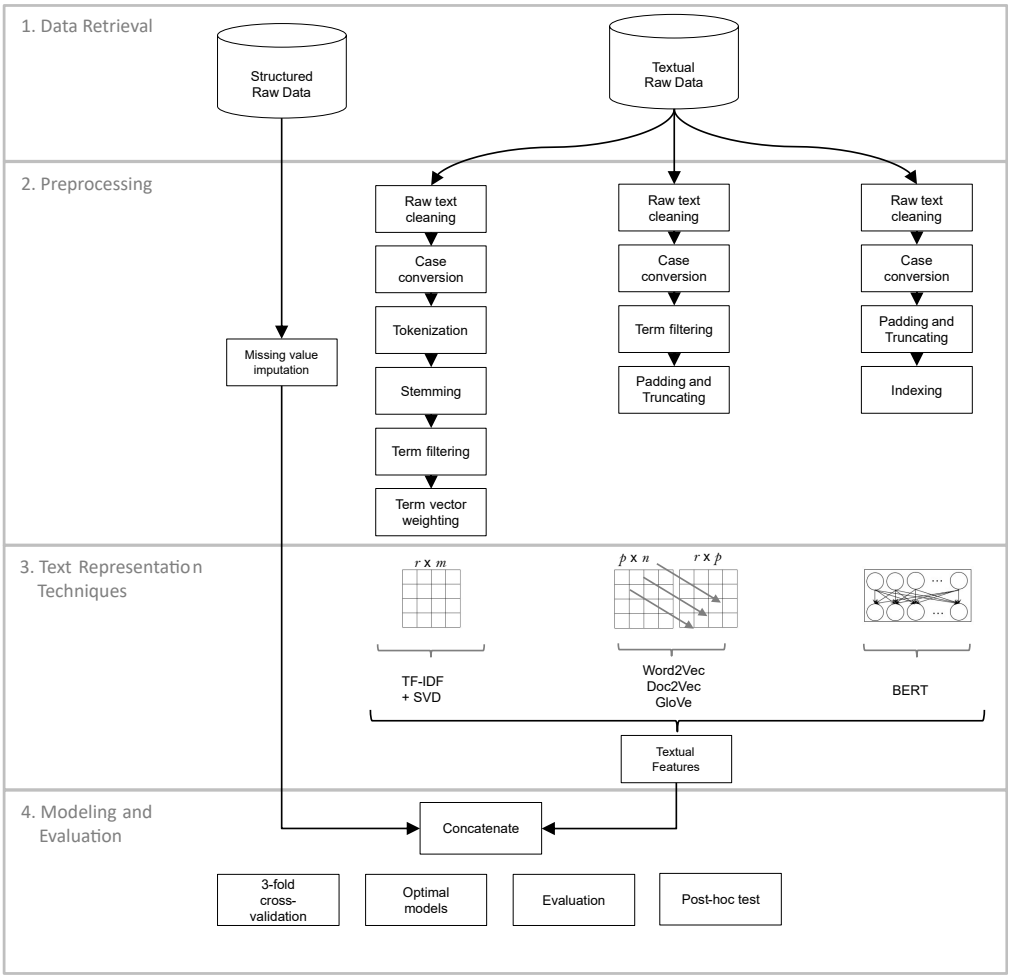


Figure 2.6. System design showing four consecutive steps

Appendix A.1. Data Retrieval and Preprocessing. The data consists of structured oil price data and textual data in the form of tweets. The *structured data* is represented by the daily spot price of a barrel of crude oil on trading days as sourced from the EIA. Daily oil prices for both Brent and WTI crude oil barrels are retrieved. Missing values occur for the structured data during non-trading days. These are imputed using a Kalman filter (Koopman, 1997), commonly used in forecasting due to its precision for missing value estimations even when the precise nature of the modeled system is unknown.

For the *textual data*, tweets are downloaded from the Trump Twitter Archive website (Brown), accessed on June 21, 2022. Multiple tweets can be posted on a certain date. Therefore, tweets produced on the same day are grouped and ordered according to the posting time to create one document per day. This results in an average length of 152 tokens (or words) per document, with the longest document having 933 tokens after data cleaning by converting the text to lowercase and removing punctuation, special characters, and numbers. There are 14 days without any tweets posted, none of which are consecutive, which is around 1% of the total textual data. No imputation is implemented as days with no tweets are coded as zero-vectors. Additional preprocessing steps are deployed depending on the NLP technique following previous literature, as shown in Figure 2.6, with further details listed below.

- TF-IDF: raw text cleaning, by removing non-alphabetic characters; case conversion, by transforming everything to lowercase; term filtering, by removing rare words, stop words, and words with a length below 3 characters; stemming, by transforming words to their root form; tokenization, by using spaces to split a document into tokens; term-vector weighting, by assigning a weight to each token according to their frequency. (Porter, 1980; Pedregosa et al., 2011)
- Word2Vec, Doc2Vec, and GloVe: raw text cleaning, case conversion, term filtering, padding and truncating (where documents that are shorter than a given dimension size are padded with zeroes, while those that are longer are truncated, ensuring the same vector length for all representations), and tokenization. (Rehurek and Sojka, 2011)
- BERT: raw text cleaning, case conversion, padding and truncation, wordpiece tokenization (splitting words into their smallest possible unit), and indexing (assigning an integer number to each token, to match to the corresponding embedding). (Vaswani et al., 2017; Devlin et al., 2018; Wolf et al., 2019)

The differences in preprocessing are due to the different characteristics of each technique. For example, stemming and term filtering are important for TF-IDF to reduce the dimensionality size of the resulting matrix, thus reducing sparsity. Further, BERT uses a unique type of tokenization to capture more contextual information. For instance, using word endings such as -s or -ing as separate tokens allow to represent plurals and verbs.

Appendix A.2. NLP techniques. All methods transform the corpus, i.e. the collection of tweets, into different numerical representations which are then concatenated to the structured data to be used by the LSTM model (see also step 4 in Figure 2.6).

First, the *TF-IDF method* is used as a vector space-based approach weighting each token (or term) depending on how relevant the token is to a document relative to the corpus (Salton and Buckley, 1988). TF-IDF results in a high-dimensional term-by-document matrix where each cell in the matrix is the product of the term frequency (TF) and inverse document frequency (IDF) (Salton and Buckley, 1988). Following Pedregosa et al. (2011), TF-IDF is calculated for each token-document pair as:

$$\text{TF-IDF}(s,d) = \text{TF}(s,d) \times log\left(\frac{1+n}{1+\text{DF}(s)} + 1\right),$$

where $\text{TF}(s,d)$ is the number of times token s is present in document d, $\text{DF}(s)$ is the number of documents in the corpus containing token s, and n is the total number of documents in the corpus. This results in a sparse high-dimensional term by document matrix. Latent Semantic Indexing (LSI) using truncated Singular Value Decompostion (SVD) is implemented to reduce the dimensionality of this matrix with the optimal number of dimensions selected through profile log-likelihood as proposed by Zhu and Ghodsi (2006).

Second, this study incorporates NLP techniques that generate *word embeddings*, i.e., Word2Vec, Doc2Vec, and GloVe. **Word2Vec** is a shallow neural network model trained to learn linguistic word contexts as numerical vector representations (Mikolov et al., 2013a). Its flexibility and performance are the reasons for including it in this study (Li et al., 2019a). Two variations of Word2Vec exist, i.e., the Skip-gram and Continuous Bag-Of-Words (CBOW) model, and both are included for comparison reasons.

- The Skip-gram model is trained with the objective of learning word vector representations that are good at predicting nearby tokens. Following Mikolov et al. (2013b), for a sequence of M training tokens $s_1, s_2, ..., s_M$, Skip-gram seeks to maximize the average log probability:

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{-q\leq j\leq q, j\neq 0} logP(s_{m+j}|s_m),$$

  where s_m is the center token and q is the number of neighbors.

- The CBOW training objective is to learn representations that are helpful for predicting a given token, using q neighboring context tokens (Mikolov et al., 2013a). The objective function seeks to maximize the average log probability of token s_m occurring, given q neighboring tokens:

$$\frac{1}{M}\sum_{m=1}^{M} logP(s_m|s_{m-q}, ..., s_{m-1}, s_{m+1}, ..., s_{m+q}).$$

Doc2Vec or Paragraph Vector is a generalization of Word2Vec and represents larger pieces of text instead of tokens. This technique inherits Word2Vec parameters, but learns fixed-length representations from variable-length pieces of texts, such as sentences, paragraphs, and documents (Le and Mikolov, 2014). Similar to Word2Vec, two model architectures are available, i.e., a Distributed Bag-Of-Words version of Paragraph Vector (PV-DBOW) and a Distributed Memory model of Paragraph Vector (PV-DM). Like CBOW, PV-DM is trained with the objective to learn vectors that contribute to predicting the next token in the document using the context tokens (Dai et al., 2015). PV-DBOW ignores the order of context tokens in the input to predict tokens randomly sampled from the document as output comparable to Skip-gram (Lau and Baldwin, 2016). **GloVe** trains on global word co-occurrence counts within a corpus and therefore overcomes the drawback of training on separate local context windows which might result in the model poorly utilizing the statistics of the complete corpus (Pennington et al., 2014). Following Pennington et al. (2014), a weighted least squares objective function minimizes the difference between the dot product of the vectors for two tokens and the logarithm of their co-occurrences:

$$\sum_{j,k=1}^{V} f(G_{jk})\left(w_j^{\intercal}\tilde{w}_k + b_j + \tilde{b}_k - log(G_{jk})\right)^2,$$

where V is the size of the vocabulary, G_{jk} is the number of times token j co-occurs with token k, $f(G_{jk})$ is a weighting function to avoid overweighted rare and frequent co-occurrences, w_j and b_j are respectively the vector and bias for token j, while $\tilde{w}_k$ and $\tilde{b}_k$ are the vector and bias for context token k. The vectors result in a matrix of tokens, where the number of dimensions is consistent for all word vectors. This matrix is used to assign weights to the tokens present on each document through an embedding layer (Chollet et al., 2015).

Third, *BERT* is included as a transformer-based approach, because it is the first unsupervised deeply bidirectional system for pretraining NLP tasks, and it has achieved great results in various language-related tasks, such as question-answering and comparing semantic meaning similarity, among others (Devlin et al., 2019). Its architecture is based on complex neural networks. BERT includes an encoder and decoder connected

through an attention mechanism which reduces the required training time and makes the model more parallelizable when compared to other models in the literature (Vaswani et al., 2017). BERT was created to solve sequence-to-sequence NLP tasks and to learn contextual relations between tokens in a text using attention mechanisms to handle long-range dependencies (Vaswani et al., 2017). During training, a sequence of tokens is fed into BERT with information learned from both the left and right side of a token incorporating context to learn general purpose language representations. For each input token, an output sequence of vectors is returned, where the size of the vector depends on the model used. The general pretrained model, BERT-Base, can be used effectively on new applications without substantial task-specific architecture modifications (Devlin et al., 2019).

TF-IDF and Doc2Vec allow new features to be extracted from text on a document level, thus allowing for immediate concatenation with the structured data. For GloVe and Word2Vec, an embedding layer is used with the vectors as the corresponding weights. This allows to combine and reshaped the indices and weight matrices to allow concatenation with the structured data. Similarly, BERT requires reshaping for concatenation.

Appendix A.3. Long Short-Term Memory Networks. LSTM networks were introduced to learn long-term dependencies and to improve upon problems present in previous neural networks, such as vanishing and exploding gradients (Hochreiter and Schmidhuber, 1997). As if it were a network with multiple copies of itself, information persists in a chain-like fashion from one copy to the next through different layers interacting between them.

The general structure of LSTM networks consists of an input layer, one or multiple hidden layers, and an output layer. The input layer has the same number of neurons as input features used, while the output layer of a single neuron represents the target feature or the daily oil price in our study context. The hidden layers contain memory cells, each with three gates to adjust the information in its cell state by using a sigmoid function (σ) and a point-wise multiplication operation ($\odot$) to produce an output between zero and one.

At timestep t, the cell state (C_t) can be thought of as the central part of a memory cell, where information passes through, regulated by the forget gate (f_t), the input gate (i_t), and the output gate (o_t), defined below:

1. Forget gate layer (f_t): decides what information is discarded from the cell state by looking at the previous hidden state (h_{t-1}) and the current observation (x_t).

$$f_t = \sigma(W_{f,x}x_t + W_{f,h}h_{t-1} + b_f).$$

2. Input gate layer (i_t): selects which values to update, and creates a vector of new candidate values ($\tilde{C}_t$) to update the cell state. The hyperbolic tangent function ($tanh$) additionally distributes the gradients in order to prevent the vanishing or exploding of gradients.

$$i_t = \sigma(W_{i,x}x_t + W_{i,h}h_{t-1} + b_i).$$
$$\tilde{C}_t = tanh(W_{\tilde{C},x} + W_{\tilde{C},h}h_{t-1} + b_{\tilde{C}}).$$

3. Update step: the previous cell state C_{t-1} is updated into the new cell state C_t, depending on the outputs of the previous gate layers defining the amount of earlier information dropped and new information added.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t.$$

4. Output gate layer (o_t): the updated cell state value gets filtered by σ and then fed into a hyperbolic tangent function, defining if the information in the current cell state C_t is visible or not.

$$o_t = \sigma(W_{o,x}x_t + W_{o,h}h_{t_1} + b_o).$$

$$h_t = o_t \odot tanh(C_t).$$

The input data is a sequence of timesteps to predict the next trading day. At timestep t, each gate will have a weight matrix W, a bias term b, the corresponding input element x_t, and the output of the previous timestep h_{t-1}, if applicable. A memory cell, shown schematically in Figure 2.7, is updated at every timestep t (Graves, 2013).

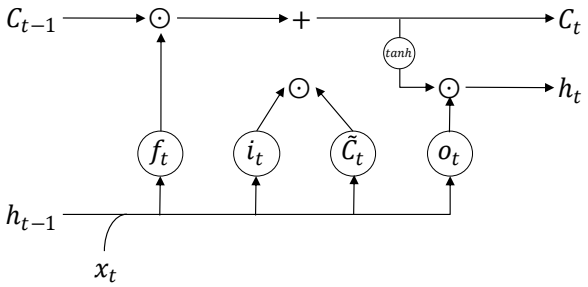


Figure 2.7. LSTM memory cell

Appendix A.4. Model Architecture. The LSTM model architecture requires the concatenation of textual and structured data. The text is first converted to a numerical representation, which differs per NLP technique. For TF-IDF, the SVD features are passed as inputs similar to the structured features. For Word2Vec and GloVe, the text is first tokenized and then turned into integer indices, which are then passed into an embedding layer and weighted through the pertinent word vectors. Doc2Vec issues vectors weighted by document, rather than by token, which are used similarly to TF-IDF values. These vectors were previously prepared by training models on their corresponding integer indices. These vectors are loaded into the LSTM architecture. For BERT, the pretrained model is stacked onto the model architecture to fine-tune the language model for oil price prediction, also using token indices as input.

The concatenated inputs are then fed into an LSTM layer that returns the full output sequence to pass through another dropout layer. This output is passed into a second LSTM layer that only returns the last output to be processed by a batch normalization layer, a dropout layer, and a final dense layer with Rectified Linear Units (ReLU) activation. Both batch normalization and dropout are useful to avoid overfitting and speeding up the model training (Chollet et al., 2015).

Appendix A.5. Hyperparameter Tuning and Selection. Hyperparameter tuning and selection is performed through the combination of a grid search and a 3-fold cross-validation method as proposed by Li et al. (2020). 3-fold cross-validation is applied on the training set by re-splitting at fixed time intervals to create alternate training and validation subsets as shown in Figure 2.8. For each fold, the validation observations occur at a later time period than the training observations, with each training set being a combination of the previous splits with no shuffling to ensure robust model creation (Bergmeir and Benítez,

2012). The best hyperparameter combinations are selected according to the performance on the validation subsets.

For the LSTM model, the validation set is used for early stopping Li et al. (2020). This tracks the losses in the validation data set, optimizing models without overfitting to in-sample data. Depending on the loss, the number of epochs for training is adjusted, halting when the loss is no longer reducing after 10 consecutive epochs (Fischer and Krauss, 2018). The best models are then scored out on the test sets.

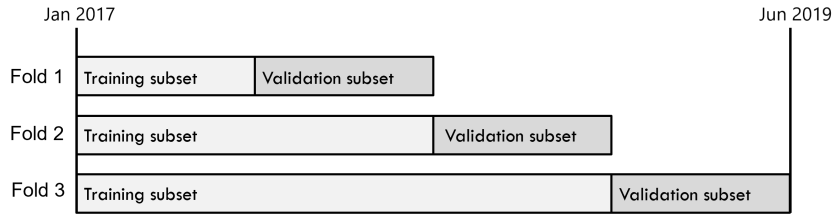


Figure 2.8. 3-fold cross-validation on the original training set, or 60% of the data

LSTM hyperparameter values, shown in Table 2.9, are selected using grid search in line with Li et al. (2020), for both Brent and WTI oil prices. These are the number of LSTM layers, number of neurons (dimension of hidden state), forget bias (initial bias vector of the LSTM layer), dropout rate (rate of randomly ignored outputs), kernel initializer (initializer of the weight matrix of input features with the exception of the hidden state), and kernel regularizer (regularization function applied to the weight matrix of input features). Other values are fixed beforehand based on existing literature and therefore not included in Table 2.9. Specifically, a minibatch size of 256 (Kraus et al., 2020), a training duration of 500 epochs (Li et al., 2020; Zhao et al., 2017), and Adam with weight decay as an optimizer (Kingma and Ba, 2014; Loshchilov and Hutter, 2017).

Different hyperparameters are fine-tuned for each NLP technique. For TF-IDF ($\text{LSTM}_{\text{TFIDF}}$), 9 dimensions are set before hyperparameter selection, following Zhu and Ghodsi (2006). For GloVe ($\text{LSTM}_{\text{GloVe}}$), pretrained word vectors with different numbers of dimensions are used to compare performance. For Word2Vec (LSTM_{W2V}) and Doc2Vec (LSTM_{D2V}), the number of dimensions follow $\text{LSTM}_{\text{GloVe}}$, while the rest of the hyperparameters arise from Mikolov et al. (2013b) and Mikolov et al. (2013a). Finally, the textual data for BERT ($\text{LSTM}_{\text{BERT}}$) obeys the base model requirements (Devlin et al., 2019), being truncated at 512 tokens per document and 768 dimensions.

The optimized LSTM models are compared to a set of benchmark models, with hyperparameter values tuned through the same process as the LSTM models. For ARIMA, the optimal order of (p,d,q) is selected based on structured data ($\text{ARIMA}_{\text{Struct}}$), following Elshendy et al. (2018). This order is retained when incorporating textual features ($\text{ARIMAX}_{\text{Text}}$). Similarly, VAR with only structured data ($\text{VAR}_{\text{Struct}}$) determines order p under VAR lag order selection criteria (Ivanov and Kilian, 2005), keeping the same value for textual features (VAR_{Text}). SVR is tuned for gamma, cost, and epsilon hyperparameters, following Maldonado et al. (2019) for the model using only structured data ($\text{SVR}_{\text{Struct}}$). Then, the different vectors from each NLP technique are incorporated onto this optimized model, displaying only the results for the best performing version (SVR_{Text}). An augmented Dickey-Fuller test reveals non-stationary behavior with first-order differencing for ARIMA, ARIMAX, and VAR, before model tuning. Features from NLP are added depending on the optimal parameters selected, indicating the optimal lag order. The hyperparameter values are displayed on Table 2.10.

Table 2.9. LSTM hyperparameter tuning summary

Model	Name	Hyperparameters	Values Evaluated	Values Selected	
				Brent	WTI
LSTM	$LSTM_{Struct}$	Layers	1, 2, 3, 4, 5	3	3
		Neurons	20, 50, 100, 200	200	200
		Forget bias	True, False	False	False
		Dropout rate	0.2, 0.35, 0.5	0.35	0.2
		Kernel initializer*	RN, RU, GN, GU	GN	RU
		Kernel regularizer	None, L2	None	None
LSTM + TF-IDF	$LSTM_{TFIDF}$	Dimensions	9	9	9
LSTM + GloVe	$LSTM_{GloVe}$	Dimensions	25, 50, 100, 200	100	200
		Minimum count	1	1	1
LSTM + Word2Vec	$LSTM_{W2V}$	Dimensions	25, 50, 100, 200	50	100
		Architecture	Skip-gram, CBOW	CBOW	CBOW
		Epochs	1, 3	1	1
		Context window	4, 5	5	5
		Minimum count	1	1	1
		Negative sampling	0, 5, 10, 15, 20	5	10
		Downsampling	0, 0.00001	0	0.00001
LSTM + Doc2Vec	$LSTM_{D2V}$	Dimensions	25, 50, 100, 200	50	100
		Architecture	PV-DM, PV-DBOW	PV-DBOW	PV-DBOW
		Epochs	1, 3	1	1
		Context window	4, 5	5	5
		Minimum count	1	1	1
		Negative sampling	0, 5, 10, 15, 20	5	10
		Downsampling	0, 0.00001	0	0.00001
LSTM + BERT	$LSTM_{BERT}$	Dimensions	768	768	768

*Kernel initializer values refer to random normal (RN), random uniform (RU), glorot normal (GN), and glorot uniform (GU).

Table 2.10. Benchmark model hyperparameter tuning summary

Model	Name	Hyperparameters	Values Evaluated	Values Selected	
				Brent	WTI
ARIMA	$\text{ARIMA}_{\text{Struct}}$	(p, d, q)	(1, 1, 1), (1, 1, 2)	(1, 1, 1)	(1, 1, 2)
ARIMA + Text	$\text{ARIMAX}_{\text{Text}}$	NLP Technique	BERT, W2V, D2V, GloVe, TF-IDF	TF-IDF	GloVe
VAR	$\text{VAR}_{\text{Struct}}$	p	2	2	2
VAR + Text	VAR_{Text}	NLP Technique	BERT, W2V, D2V, GloVe, TF-IDF	TF-IDF	GloVe
SVR	$\text{SVR}_{\text{Struct}}$	gamma	$2e-15, ..., 2e+15$	$2e-03$	$2e-03$
		cost	$2e-15, ..., 2e+15$	$2e+02$	$2e+02$
		epsilon	0.01, 0.1, 0.5, 1.0	0.01	0.01
SVR + Text	SVR_{Text}	NLP Technique	BERT, W2V, D2V, GloVe, TF-IDF	TF-IDF	TF-IDF

2.8.2 Appendix B. Additional Results

We include directional accuracy (DA) for complementary insights with our previously displayed results. Following Zhao et al. (2017), DA is defined as:

$$DA = \frac{1}{N}\sum_{t=1}^{N} a(t) \times 100\%,$$

where N is the number of predictions, a(t) is 1 if $(y(t+1) - y(t))(\hat{y}(t+1) - y(t)) \geq 0$, else it is 0. As before, the results displayed on Table 2.11 show that the inclusion of text outperforms LSTM without text, while BERT continues to be the superior NLP technique.

Table 2.11. Directional accuracy of LSTM models

Model	**Brent Oil Price**	**WTI Oil Price**
$\text{LSTM}_{\text{Struct}}$	0.289	0.286
$\text{LSTM}_{\text{TFIDF}}$	0.221	0.211
$\text{LSTM}_{\text{GloVe}}$	0.346	0.228
LSTM_{W2V}	0.641	0.635
LSTM_{D2V}	0.572	0.586
$\text{LSTM}_{\text{BERT}}$	**0.789**	**0.703**

Chapter 3: Improved Decision-Making Through Life Event Prediction: A Case Study in the Financial Services Industry

CHAPTER 3

Improved Decision-Making Through Life Event Prediction: A Case Study in the Financial Services Industry

Abstract.

Life event prediction is an important tool for customer relationship management (CRM), because life events shift customers' preferences towards different products and services. Existing life event research mainly uses cross-sectional data, whereas in the CRM field, incorporating longitudinal data is increasingly common. Because longitudinal data can capture the dynamics of customer behavior, opportunities arise to benchmark the power of longitudinal customer data for predictions of cross-sectional versus longitudinal life events. Therefore, this study compares statistical and machine learning (SaML) classifiers, such as logistic regression, random forest, and XGBoost, with long- and short-term memory networks (LSTM), using data represented in both cross-sectional and longitudinal setups for life event prediction. Through a real-life longitudinal customer data set from a European bank, the authors represent the longitudinal data in a cross-sectional data format, using featurization in the form of aggregation. The available data cover 42 end-of-month snapshots for 760,438 unique customers. For marketing decision-making literature, this article (1) introduces three novel life events (i.e., primary, secondary, and rental residence purchases) to life event predictions; (2) offers guidance for how to leverage longitudinal customer data, according to the comparison of various featurization approaches and benchmarking SaML classifiers against LSTM; and (3) clarifies the importance of features and timing for improving marketing decision-making dynamically. The results show that aggregating features over time is preferable as a featurization approach for cross-sectional modeling using SaML classifiers. Furthermore, LSTM can capture behavioral changes over time, unlike SaML classifiers. It also performs significantly better than SaML classifiers on the area under curve and F1 metrics. Insights into the uses of integrated gradients reveal that feature importance changes over time. An integrated gradients method can assist decision-makers in their efforts to plan effective communication with customers in advance, such as by allocating more resources to customers who exhibit high probabilities of a particular life event occurrence.

Keywords: Life event prediction, decision-making, deep learning, LSTM, explainability.

3.1 Introduction

Customer relationship management (CRM) is a critical factor for successful business performance (Sin et al., 2005). Constant improvement of customer-firm relationships is an important goal of CRM, which can be achieved by monitoring customer behavior (Sin et al., 2005). However, customers' needs change over time, making these relationships challenging to maintain (De Caigny et al., 2020). Thus, firms seek to detect triggers that indicate upcoming changes in customers' behavior and needs, such as *life events*, which are important moments in every customer's life (De Caigny et al., 2020).

Life event prediction contributes to marketing decision-making as it is linked with life-cycle theory (De Caigny et al., 2020), which implies that customers change their behavior in accordance to disruptions in their lives. Following lifecycle theory, customers behave similarly when they are in the same stage, purchasing analogous products within the financial services industry (Antonides and Van Raaij, 1999). Thus, managers have come to understand that identifying consumption drivers can lead to cross-selling opportunities (Verhoef and Donkers, 2001). Successful cross-selling in turn increases the profitability of a customer (Knott et al., 2002), increased switching costs (Kumar et al., 2008), and better retention rates (Kamakura et al., 2003). Therefore, life event prediction adds value to CRM research by serving as an additional tool for improving customer segmentation and targeting, more profound customer relations, and detecting potential cross-selling opportunities in advance.

In particular, life events lead to a reevaluation of consumption priorities (Mathur et al., 2008), alter preferences for different products and services (Sahoo et al., 2012), usher people towards specific products, and foreshadow their future behavior (Malthouse, 2007). Research shows customer behavioral changes arise from moving, getting married or divorced, starting a job, or having a child, among others (Andreasen, 1984). Further, customers adjust their financial priorities according to their life events, such as focusing on buying a home after marriage, covering different insurance needs after forming a family, or investing after retirement (Kamakura et al., 1991). Thus, life event prediction contributes to a better understanding of customer behavior and needs, which improves firms' retention efforts, their understanding of marketing segments (Andreasen, 1984), and their efforts to develop highly personalized services over time. This places life event prediction as a critical element in CRM strategies (Kumar et al., 2021).

Some of the life events we study have not been documented before, which might inspire marketers to pursue novel applications for their decision-making processes. These novel life events are primary residence purchase, secondary residence purchase, and rental residence purchase, which are closely related to specific financial services products and thus are also highly relevant for decision-making. Previous literature shows that setting up and confirming a new mortgage loan is a process prone to delays that can hinder a company's business opportunities (Brahma et al., 2021). This highlights how any improvement in managing this process, such as by accurately targeting relevant customers in advance, can present an opportunity for developing a competitive advantage (Brahma et al., 2021). In addition, customers will also exhibit varying financial goals and capabilities depending on their different mortgage loans or housing situation (Bunnell et al., 2020), requiring products and services tailored to their characteristics. For example, a customer who is gearing towards the purchase of a primary residence may have financial goals more oriented towards managing debt (Bunnell et al., 2020). The purchase of a secondary residence could be by a customer working towards financial independence

(Bunnell et al., 2020). A rental residence purchase may be more aligned with broad financial legacy objectives, such as estate planning (Bunnell et al., 2020). Therefore, these life events present distinct business applications, as well as valuable information for providing highly personalized services within the CRM context.

From a methodological perspective, extant life event literature tends to focus on a limited set of life events and rely mainly on cross-sectional modeling strategies, which cannot reflect behavioral changes over time (Boulding et al., 2005). Predictive modeling applications in CRM also emphasize the need to incorporate longitudinal customer data to capture dynamic behavior (Óskarsdóttir et al., 2018) and ensure the validity of results over time (Boulding et al., 2005). Longitudinal customer data might be transformed into cross-sectional data through *featurization*. However, a research gap remains on whether featurization is an effective approach for life event prediction.

To featurize longitudinal data, we aggregate customer data according to the mean, standard deviation, coefficient of variation (CV), or sum, depending on the type of information (Gattermann-Itschert and Thonemann, 2021). As a baseline, we transpose longitudinal customer data into a tabular data set (Chen et al., 2012). Both data sets allow conventional statistical and machine learning (SaML) classifiers to predict life events. As a key contribution to extant literature, we evaluate the performance of these data approaches for life event prediction.

Since the predictive power of longitudinal data also depends on the evolution of the focal feature over time (Bagnall et al., 2017), we also model the data as a chronologically ordered sequence of features, or sequential input. Deep neural networks (DNN), especially recurrent neural networks, are excellent at extracting key information from sequential input for prediction (Wang et al., 2021, 2023; Zhong et al., 2023). Long short-term memory (LSTM) is well-suited for deriving predictions from longitudinal data, represented in a sequential ordered input, and is particularly popular in the CRM field Cheng and Chen (2022). We thus provide an initial evaluation of life event prediction modeling with longitudinal data, as sequential input for LSTM. Following the explosive growth of DNN applications, attribute-based explanation methods also have been proposed to improve understanding (Sundararajan et al., 2017). However, we do not find DNN or related techniques in existing life event prediction literature. As such, we contribute an evaluation of the LSTM model results through the use of attribute-based explainability techniques, in the form of integrated gradients (IG).

Thus, we contribute to the life event prediction literature by incorporating novel life events, offering a predictive performance comparison of longitudinal and cross-sectional approaches through aggregation featurization, and investigating whether life event detection can be improved by using longitudinal data as sequential input for LSTM. Further, we develop a framework that is generalizable for classification problems using consumer data. Moreover, we apply a state-of-the-art interpretability technique to deliver insights into how prediction drivers vary in time and by life event, which is relevant information for decision-makers seeking to improve the personalization of their provided services.

To establish these contributions, we leverage a real-world data set from a large European financial services firm that provided longitudinal data, in the form of 12 end-of-month behavioral snapshots per customer and 10 life events, 3 of which are novel. The data contain detailed information for a diverse pool of 760,438 customers, 10% of which have experienced at least one life event occurrence. The customers are aged between 18 and 67 years and exhibit an overall average relationship length of 19 years; 80% of

customers name this firm as their main financial services provider. We gather 245 customer features, 169 of which are dynamic features from longitudinal data (e.g., number of monthly credit card transactions, paid fees). These features then can be transposed for our baseline featurization approaches or aggregated for our SaML classifiers. They also provide the ordered input sequence for the LSTM model. The remaining 76 features are static, and we obtain the latest customer information available as additional input (e.g., age, civil status).

Furthermore, our three main contributions reflect the motivation for this research, which we summarize in the following research questions (RQ):

RQ1. Which featurization approach for representing longitudinal customer data improves cross-sectional life event prediction performance for conventional SaML classifiers?

RQ2. Does LSTM improve life event prediction performance, relative to SaML classifiers, when longitudinal customer data are available?

RQ3. Is it possible to identify life event drivers that are useful for marketing decision-making?

To answer RQ1, we compare the life event prediction performance of the aggregation featurization approach against the transposed longitudinal data set, as baseline input for SaML classifiers. We follow extant literature and deploy logistic regression (LR) (Hosmer Jr et al., 2013), random forest (RF) (Breiman, 2001), and XGBoost (XGB)(Friedman, 2001). All these SaML classifiers commonly inform CRM predictive tasks, due to their stellar performance (Huang and Meng, 2019; Coussement and Benoit, 2021; Yi et al., 2023). For RQ2, we compare the best performing combination of featurization approach and SaML classifier against the performance of the LSTM model, which is particularly well-suited for longitudinal, sequentially ordered input data Cheng and Chen (2022). Finally, for RQ3, we specify the best performing approach for delivering interpretable, actionable insights for different life events, according to feature type and time step.

Therefore, we compare the incorporation of longitudinal data for life event prediction in multiple ways. First, we transform the data through aggregation, explored in RQ1 and chosen for being the most commonly used method for life event prediction within previous research. Second, the longitudinal data is used as sequential input, therefore without the need of aggregation. These two different approaches are compared against each other on RQ2, ultimately evaluating whether or not longitudinal data, when used sequentially through LSTM, is of value for life event prediction.

In the next section, we review extant life event prediction literature to identify some research gaps. Section 3.3 introduces the LSTM model and featurization approaches for longitudinal data. Then in Section 3.4, we present the experimental setup and zoom into the data and data preprocessing steps, the LSTM architecture, and the hyperparameter tuning process. After presenting the predictive performance results in Section 5, we offer managerial insights in Section 3.6, as they pertain to marketing decision-making. Section 3.7 concludes.

3.2 Related Research

Life event prediction is an important task for improving CRM strategies and customer-centered decision-making (De Caigny et al., 2020). Table 3.1 presents an overview of extant life event prediction literature, in terms of life events that have been investigated (*Life Events*) and the types of data they use (*Data Usage*). In addition to classifying

whether they use cross-sectional (CSD) and/or longitudinal data (LD) to predict life events, we note whether each study includes a prediction performance benchmark. We also indicate the featurization approach (*Feat.*), if applicable. Finally, we examine the modeling approach adopted by each study (*Modelization*). We evaluate the presence of SaML classifiers and DNN models, as well as whether explainability techniques (Exp.) provide further insights from the model's results.

As Table 3.1 reveals, the literature on life event prediction offers several contributions. First, the most researched life events are the start of a personal relationship (RelS) and birth of a child (BoC), both of which appear in five previous studies. Moving (Mov) and personal relationship end (RelE) are covered by two studies; job market entry (JobE) and car purchase (CarP) are addressed by one study each. Some other life events, such as primary residence purchase (PrimR), secondary residence purchase (SecR), or rental residence purchase (RentR), have not been studied before. We seek to contribute to life event prediction literature by considering their viability in our study.

Second, we review the data usage characteristics of the extant literature. Life events are predicted with CSD, LD, or both. From Table 3.1, we conclude that CSD is the more popular data type, yet we know of no comparison between the prediction performance of CSD and LD. Furthermore, aggregation is the most common way to featurize longitudinal data. Therefore, we investigate whether SaML classifiers benefit from aggregating longitudinal customer data, compared with using the time-dependent information directly, while also benchmarking the contributions of CSD and LD to total prediction performance.

Third, we review the use of machine learning techniques for the purpose of life event prediction. We find that Khodabakhsh et al. (2018) and De Caigny et al. (2020) apply machine learning techniques, with other studies opting for analyzing Pearson correlation between life events and a consumption variable. Further, DNN is also rare within life event prediction literature, in spite of their high predictive performance, particularly when using longitudinal data as a sequential input (Wang et al., 2021, 2023; Zhong et al., 2023). In fact, within the reviewed studies, we find that only one applies a DNN (Khodabakhsh et al., 2018). Two other studies include LD but no DNN, opting for transforming the data through aggregation instead (Mathur et al., 2008; Koschate-Fischer et al., 2018). We also find that any performance benchmark relative to cross-sectional SaML is rare. Thus, the use of DNN, specifically in combination with sequential data, and its evaluation against cross-sectional algorithms remains vastly unexplored in life event prediction.

Fourth, Table 3.1 reveals that no explainability techniques have been applied to life event prediction, leaving a notable gap in extant research.

Briefly then, we contribute to the life event prediction literature by incorporating novel life events, offering a predictive performance comparison of longitudinal and cross-sectional approaches through aggregation featurization, and investigating whether life event detection can be improved by using longitudinal data as sequential input for LSTM. Further, we deliver insights into how life event prediction drivers can improve marketing decision-making.

Table 3.1. Life event prediction literature review

Study	Life Event*										Data Usage**				Modelization***		
	Mov	RelS	RelE	JobE	BoC	CarP	PrimR	SecR	RentR	Ret	CSD	LD	Vs.	Feat.	SaML	DNN	Exp.
Andreasen (1984)	✓	✓	✓	✓	✓	-	-	-	-	-	✓	-	-	Agg	NA	-	-
Lee et al. (2001)	✓	✓	✓	✓	✓	-	-	-	-	✓	✓	-	-	Agg.	NA	-	-
Mathur et al. (2003)	✓	✓	✓	✓	✓	-	-	-	-	✓	✓	-	-	Agg.	NA	-	-
Mathur et al. (2008)	✓	✓	✓	✓	✓	-	-	-	-	✓	✓	✓	-	Agg.	NA	-	-
Khodabakhsh et al. (2018)	✓	✓	✓	-	✓	-	-	-	-	-	✓	✓	-	NA	SVM, RF, GBT	✓	-
Koschate-Fischer et al. (2018)	-	-	-	✓	✓	-	-	-	-	✓	-	✓	-	Agg	NA	-	-
De Caigny et al. (2020)	✓	✓	✓	-	✓	-	-	-	-	-	✓	-	-	Agg	LR	-	-
This study	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Agg., SI	LR, RF, XGB	✓	✓

*Life events: moving (Mov), relationship start (RelS), relationship end (RelE), job market entry (JobE), birth of a child (BoC), car purchase (CarP), primary residence purchase (PrimR), secondary residence purchase (SecR), rental residence purchase (RentR), and retirement (Ret).

**Data usage featurization abbreviations: Agg (aggregation), SI (sequential input).

**Modelization abbreviations: SVM (support vector machines), RF (random forest), LR (logistic gegression), GBT (gradient boosting tree).

3.3 Methodology

3.3.1 Long Short-Term Memory

We compare the predictive performance of the LSTM model against the best performing SaML classifier to evaluate the beneficial impact of longitudinal data usage. For this purpose, LSTM is particularly interesting, because it can handle longitudinal data as sequences of chronologically ordered inputs (Cheng and Chen, 2022). It incorporates gate-like structures to regulate the flow of information efficiently (Hochreiter and Schmidhuber, 1997). The structures include a forget gate, an input gate and an output gate, all of which use a logistic sigmoid function to select the information preserved in the cell state (C_t) (Hochreiter and Schmidhuber, 1997). The gate structures process the output of the previous timestep (h_{t-1}) and the input from the current timestep (x_t), storing the key information on C_t before moving onto the next timestep, $t+1$. Therefore, the cell state can be understood as the long-term memory capacity of an LSTM, which allows it to adequately harness longitudinal data for prediction. The LSTM structure is schematically depicted on Figure 3.1.

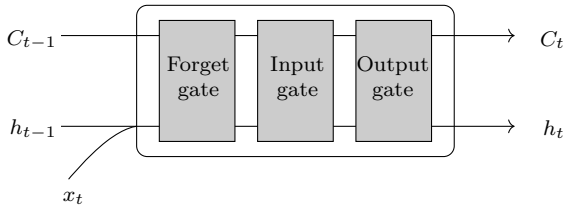


Figure 3.1. Schematic illustration of an LSTM layer

The performance of a standard LSTM model can generally be improved by incorporating attention mechanisms. In particular, attention is useful when working with complex data sources, such as textual data (Wang et al., 2016) or time series data (Cheng and Chen, 2022). Extant research also shows that attention assigns input sequence importance and improves LSTM prediction performance effectively, thus counterbalancing the difficulties it has with regard to capturing long-term dependencies (Wang et al., 2021). Therefore, we incorporate a self-attention layer into the LSTM architecture as an experimental parameter. Attention improves the robustness of DNN models by adding more weight to meaningful information from among a large number of features (Wang et al., 2021). Following Vaswani et al. (2017), we express the attention mechanism as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent matrices of a set of input vectors for queries, keys, and values, respectively, and d_k represents the dimensions for matrix K. In self-attention, the input vectors are the output of the previous layer.

A multi-head attention mechanism of h heads linearly projects the queries, keys, and values h times. Both queries and keys are linearly projected onto d_k dimensions, and values are projected onto d_v dimensions. For each projection, we compute scaled dot-product attention, expressed as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V).$$

The concatenated results get projected through a feed-forward layer, and the concatenation of h heads is:

$$MultiHead(Q, K, V) = Concat_i(head_i)W^O,$$

where W_i and W^O represent parameter matrices. The loss reflects the binary cross-entropy function, which leads to faster training and better generalization for classification tasks (Bishop, 2006). Multiple studies have combined LSTM with attention for different applications (Wang et al., 2021, 2023; Zhong et al., 2023).

LSTM can capture temporal relationships present in the data (Wang et al., 2023), with previous research showing its ability to harness chronologically ordered sequential input data Cheng and Chen (2022). Furthermore, adding an attention layer improves the robustness of LSTM and has successfully been used for various applications with different types of data, such as for fraud detection (Wang et al., 2023), for cryptocurrency price prediction(Zhong et al., 2023; Subramanian et al., 2024), and sales prediction (Lin et al., 2023). These examples show that LSTM can be adapted to learn from different data types, as well as capturing complex patterns of customer behavior for prediction (Lin et al., 2023).

3.3.2 Featurization for Longitudinal Data

In this section, we zoom in on featurization approaches to represent longitudinal data. Traditionally, longitudinal data contain static and dynamic features. The static features remain unchanged for a long period, as with age. The latter vary constantly; to capture these features, studies often take a monthly snapshot, showing the status of each customer in the database, such as the number of credit card transactions. Depending on the availability of static and dynamic features, life event prediction might follow a cross-sectional approach using conventional SaML classifiers or a longitudinal approach using LSTM. Two options exist to incorporate the longitudinal data into cross-sectional SaML classifiers: using a transposed tabular format with monthly snapshots as independent features in the model (baseline) or aggregating the monthly snapshot data (aggregation). Both static and dynamic data is included into SaML and LSTM models.

Featurizing Longitudinal Data for SaML

Two options exist to featurize longitudinal data for SaML. In a baseline setup, the dynamic, monthly snapshot features represented in tabular data format are used directly in the SaML classifiers. The most recent static features are concatenated, to represent the final input data set. This approach represents the baseline. Alternatively, different aggregations can be computed over the various monthly snapshots for each feature. Similar to Gattermann-Itschert and Thonemann (2021), we use the mean, standard deviation, and coefficient of variation (CV) for continuous features; the sum is used for ordinal features. Once the data are aggregated, they can be concatenated to the static features and used as input for the classification, to produce a probability of the occurrence of a given life event, as depicted in Figure 3.2.

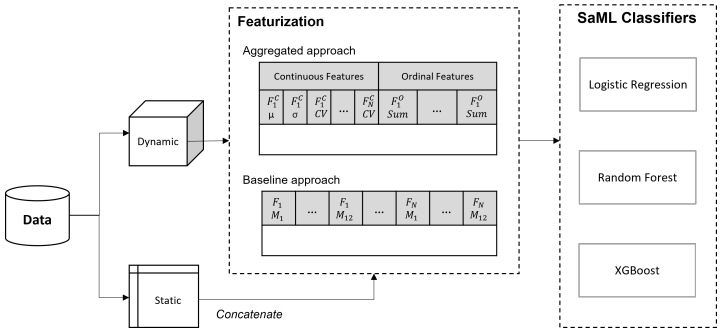


Figure 3.2. Featurization for SaML classifiers. The aggregation approach transforms continuous (F^C) and ordinal (F^O) features differently. The baseline approach transposes all features (F), resulting in a column per month ($M_1, ..., M_{12}$) for each feature.

Featurizing Longitudinal Data for LSTM

For LSTM, the dynamic features from longitudinal data are input in chronological order, as a sequence of 12 monthly snapshots for each feature. Thus, the longitudinal data are not transformed, but rather are reshaped, such that the LSTM can capture temporal relationships present in the data (Wang et al., 2023). Once the relevant information has been extracted by the standard LSTM and attention layers, the data is concatenated with the static data to reveal the probability of a given life event, as summarized in Figure 3.3.

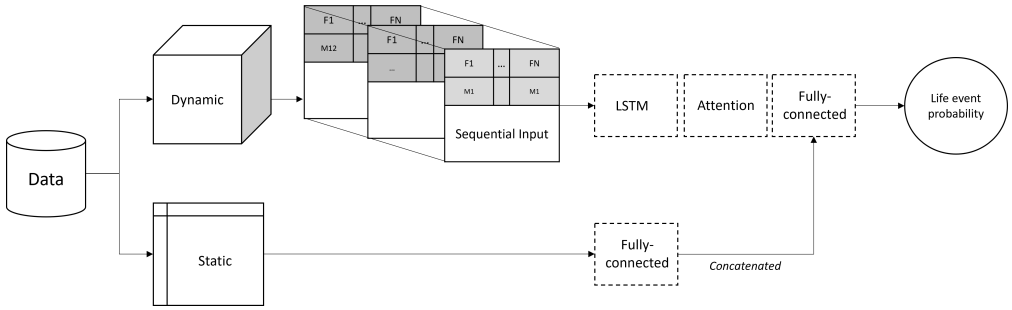


Figure 3.3. Feature use for LSTM

LSTM can capture temporal relationships present in the data (Wang et al., 2023), with previous research showing its ability to harness chronologically ordered sequential input data Cheng and Chen (2022). Furthermore, adding an attention layer improves the robustness of LSTM and has successfully been used for various applications with different types of data, such as for fraud detection (Wang et al., 2023), for cryptocurrency price prediction(Zhong et al., 2023; Subramanian et al., 2024), and sales prediction (Lin et al., 2023). These examples show that LSTM can be adapted to learn from different data types, as well as capturing complex patterns of customer behavior for prediction (Lin et al., 2023).

3.3.3 Integrated Gradients

LSTM is a blackbox model, and as such we use the IG attribution method (Sundararajan et al., 2017) for further clarity on its results. For our study context, the given function $F : R^n \rightarrow [0, 1]$ represents our LSTM model. The function's input is represented by

$x \in R^n$, while a baseline input is represented by $x' \in R^n$. The value of this baseline varies depending on the type of input data used. For example, a zero embedding vector could serve as the baseline for text models, while a black image could be used for image networks (Sundararajan et al., 2017). For our data, we follow missing value imputation guidelines to define a baseline representing the absence of information for each data type (De Caigny et al., 2020), operating under the assumption that absence of information would result in a prediction value of zero, or no life event occurrence. Then, a straight-line path can be drawn from baseline x' to the input x, from which the integral of gradients can be computed. This integral represents the accumulated gradients of the model's prediction with respect to each input feature, measuring how the prediction changes as each feature departs from its baseline value to the actual input value. By integrating these gradients along the path, Integrated Gradients (IG) capture the cumulative effect of feature changes on the model's prediction. Formally, the IG along the i^th dimension for an input x and baseline x' are expressed as follows:

$$IG_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

Where $IG_i(x)$ represents the Integrated Gradient for the i^{th} feature of input x, x_i represents the input value of the i^{th} feature, x_i' represents the baseline value of the i^{th} feature, F represents the model's prediction function, $\frac{\partial F}{\partial x_i}$ represents the partial derivative of the model's prediction with respect to the i^{th} feature, α represents the interpolation parameter, varying from 0 to 1 along the integrated path. Therefore, this equation computes the contribution from the i^{th} feature to the model's prediction by integrating the partial derivative of the model's prediction with respect to that feature along the straight-line path from the baseline to the input. The difference $(x_i - x_i')$ scales the contribution by the change in the feature's value from the baseline to the actual input. Integrating over α ensures that the attributions capture the cumulative effect of feature changes along the path. Computing this equation for each input feature provides a comprehensive understanding of feature attributions, helping to interpret and explain the model's predictions. In other words, IG are simple to interpret, as these attributions can be linked directly to the target output, a property known as "completeness", particularly in cases where the baseline corresponds to a zero prediction value of zero (Sundararajan et al., 2017). Thus, features with positive values have a positive influence on the prediction, while negative values indicate that an increase in that feature's value contributes negatively to the model's prediction. Further, the magnitude of the IG value reflects the importance or influence of the corresponding feature on the model's prediction. A larger magnitude indicates a stronger influence, while a smaller magnitude suggests a weaker influence. In addition, IG require no modification to the original LSTM network and have been tested with different types of data (Guidotti et al., 2018), including longitudinal data (Turbé et al., 2023). Overall, IG provide a quantitative measure of feature importance or contribution to the model's prediction, aiding in the interpretation and explanation of the model's behavior, while producing robust results from longitudinal data (Turbé et al., 2023). These values help identify which features are most influential in driving the model's decisions and can guide further analysis or model refinement.

To the best of our knowledge though, it has not been applied to life event prediction. In doing so, we anticipate three relevant insights. First, we seek to establish the importance of feature information throughout the 12-month time frame. Therefore, we calculate a

normalized mean attribution value across all features by month to identify changes over time. Second, the IG method can reveal the overall importance of features for predicting life events. The normalized mean, applied for each feature, can indicate the 10 features that contribute most to predictive model performance. Third, we can gain insights into the overall contributions of dynamic and static features, analyzed as a percentage of the total attributions.

3.4 Experimental Setup

Figure 3.4 depicts the experimental setup of this study. We first discuss the context, data set, and data preprocessing, after which we elaborate on the featurization approaches. Next, we detail the LSTM model architecture, the hyperparameter tuning and cross-validation strategies used, and the evaluation metrics and statistical hypothesis tests that we have implemented.

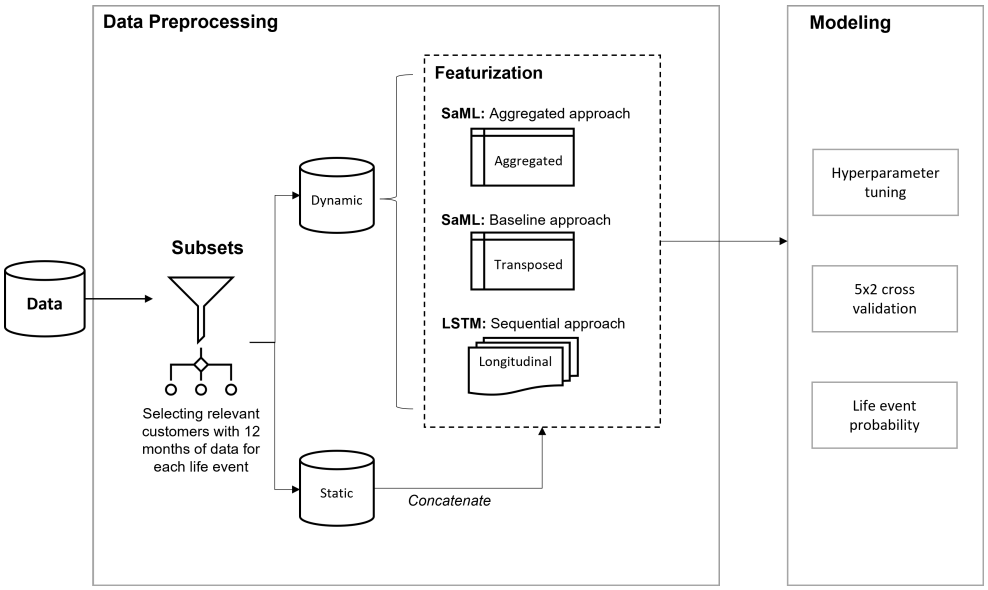


Figure 3.4. Experimental setup

3.4.1 Data & Data Preprocessing

We obtained data from a large, European financial services firm. This data set supports predictions of 10 life events and includes both static and dynamic features. The data comprise 760,438 unique customers, at least 18 years of age, with an open account at the time of the data dump.

The company delivers 42 monthly snapshots for each customer, of which it uses 12 snapshots to create the prediction features and then tracks, over the subsequent six months, whether a life event occurs or not, resulting in a binary dependent variable. Thus, with the overall data dump, we can construct three feature windows by rolling forward through the data (Fischer and Krauss, 2018; Krauss et al., 2017), as illustrated in Figure 3.5. These three 12-month sequences (i.e., S1, S2, and S3) are consolidated in one data set.

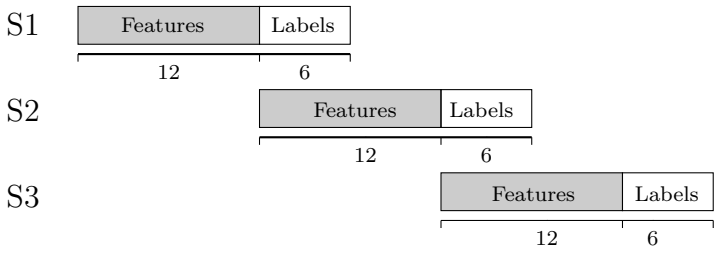


Figure 3.5. Rolling window design

The use of data in a rolling window manner, following existing literature (Fischer and Krauss, 2018; Krauss et al., 2017), allows to avoid issues arising from seasonality in a standard manner across all life events. In this study, we present the results from a particular window of time, due to constraints from the availability, consistency, and quality of the data. We propose the models to be retrained on a monthly basis, in line with traditional marketing practices. Multiple methods of accounting for time-varying factors exist, but these are out of the scope of this study, and we thus propose a comparative framework as a possible avenue for future research.

In total, there are 245 features before data preprocessing De Caigny et al. (2020), involving transaction data, customer demographics, and customer behavior features, as well as relationship data, information related to client-firm communication, loyalty, and relationship strength with the service provider. First, the transaction data include 46 features, all of which are dynamic, such as the monthly amount of business travel expenses, the monthly number of credit-related transactions, or the monthly amount of credit- and debit-related fees and charges. Second, we leverage 45 static demographic features, like age, civil status, living situation, and socio-professional group. Third, 117 customer behavior features relate to product ownership. Of these, 105 are dynamic features, such as product-related RFM variables, the number of accounts or active products, and total asset amounts. The remaining 12 features are static, such as the customer's investment risk profile. Fourth, relationship data include the communication channels used by a customer, frequency of appointments and exchanges, and time spent on web and app platforms, for a total of 37 features, 18 of which are dynamic. The static relationship variables are aspects like opt-in information for email and phone and the type of agent or advisor managing the customer's assets.

We can plug the static features directly into the life event prediction models, without needing to aggregate or sequence them. As noted, they are the most recent information of a client. We transform categorical features into dummy variables.

The dynamic features require preprocessing, depending on the featurization approach used for the 12 months of available customer data. The data are centered and scaled by subtracting the mean from each customer value and dividing it by the standard deviation for each feature (LeCun et al., 2012; Crone et al., 2006).

With regard to the life events, detected in the subsequent six months, we use the definitions provided by the firm. If a customer fulfills the life event criteria during the six months subsequent to the independent variable period, a binary variable for that life event occurrence equals 1, and 0 otherwise. The life events definitions are displayed on Table 3.2.

With these definitions, we divide the entire data set into subsets of unique customers that are relevant for each life event, whether they experienced that event at the end of the

Table 3.2. Life event definitions

Life Event	Label	Definition	Self-reported
Moving	Mov	The customer's address has been modified or added to their personal profile	✓
Car Purchase	CarP	The date of vehicle registration, which is mandatory information for customers to provide when applying for insurance or auto loans	
Primary Residence Purchase	PrimR	Loan or insurance application, where the type of residence is mandatory information and has been marked as a primary residence	
Secondary Residence Purchase	SecR	Loan or insurance application, where the type of residence is mandatory information and has been marked as a secondary residence	
Rental Residence Purchase	RentR	Loan or insurance application, where the type of residence is mandatory information and has been marked as a rental residence	
Relationship Start	RelS	Civil status modified to civil union or marriage	✓
Relationship End	RelE	Civil status modified to divorce or separation	✓
Job Market Entry	JobE	A customer is aged between 18 and 30 years and has been receiving salary payments on their account for 4 or more months; the first payment is considered the actual life event	
Birth of a Child	BoC	A new date of birth is registered for the same household	✓
Retirement	Ret	A customer starts receiving pension fund payments into their account, or their economic activity is self-reported as "retired"; only the earliest such occurrence is considered	✓

independent variable period, or not. For example, to predict if a relationship will end, we only include customers already in a relationship in that life event data set. Furthermore, we apply age range restrictions per life event, creating customer subsets to be used in the prediction of each life event. These age ranges are defined using robust estimators of location and scale with z-scores as thresholds (Rousseeuw and Hubert, 2011). In other words, different thresholds are calculated, following the age distribution of each life event. The creation of these subsets allows to select customers within the relevant age groups for each life event, which accounts for the low incidence and reduces the negative impact on model performance of data imbalance (Coussement et al., 2017). Further, these subset rules are already applied by the data provider, and as such, are based on both the business know-how, as well as proven techniques for data preparation. Table 3.3 summarizes the age subset rules applied by the data provider, along with the number of customers in the life event data set and the incidences.

Table 3.3. Life event subset definitions

Life Event	Subset Rules	Customers	Incidence
Mov	Age 19-42	300K	7.67%
CarP	Age 19-42	300K	6.40%
PrimR	Age 19-42	300K	2.11%
SecR	Age 19-42	300K	0.02%
RentR	Age 19-42	300K	0.35%
RelS	Age 25-45, civil status single	270K	1.72%
RelE	Age 29-45, civil status in a relationship	220K	0.82%
JobE	Age 19-28, not previously flagged as JobE	110K	4.52%
BoC	Age 25-38	180K	0.30%
Ret	Age 55-67, not previously flagged as retired	70K	12.12%

Following a responsible analytics framework (De Bock et al., 2023), this study complies with legal, ethical, and financial requirements. The data is anonymized to ensure customer privacy, with the financial services provider legal team having secured consent for research purposes. GDPR guidelines are followed throughout the study and the features are in line with previous CRM research in the financial services industry (Bogaert et al., 2019; De Caigny et al., 2020; Idbenjra et al., 2024). The use of longitudinal data presents the opportunity of time-varying analysis of the model results, allowing a multi-dimensional understanding of customer behavior. This additional layer of analysis improves the transparency of LSTM models and longitudinal data, which we consider relevant for the growing literature of ethical considerations for analytical applications (De Bock et al., 2023). Further, ethical issues are an essential point for companies to reflect on, especially as they can negatively impact customer's adoption of different products. For instance, existing studies show that the perception of security, privacy, and the compatibility with consumer values can impact the decision to adopt innovative banking applications and services (Luo et al., 2010; Hoehle et al., 2012). As such, our experimental setup and data use follows responsible analytics since conception, to ensure organizations eventually plan and manage the deployment of such projects ethically as well (De Bock et al., 2023).

3.4.2 LSTM Architecture

We deploy LSTM with an I-H-O structure, denoting the number of input neurons (I), neurons in hidden layers (H), and output neurons (O) (Krauss et al., 2017). The input layer matches the number of features, while H is tested as an experimental parameter with

values between 1 and 3. With multiple hidden layers, the first layer takes a size of 128 (Kim et al., 2020), which gets reduced by half for each subsequent layer, to introduce a bottleneck that forces decreased dimensionality (Dixon et al., 2015). Per (Srivastava et al., 2014), on each hidden layer output, we apply a 0.5 dropout rate to enable faster model training and less overfitting (Chollet et al., 2015). We also introduce an attention layer as an experimental parameter into the architecture, optimized during the hyperparameter tuning process. This attention layer has eight heads (Vaswani et al., 2017). The final layer produces the life event probabilities.

3.4.3 Hyperparameter Tuning

In addition to LSTM, we test LR, RF, and XGB as SaML classifiers, considering their popularity and good performance across multiple domains (Hosmer Jr et al., 2013; Breiman, 2001; Chen and Guestrin, 2016). We use cost-sensitive versions for each classifier for training, using balanced-class weights (Gattermann-Itschert and Thonemann, 2021). All classifiers are tuned through a 5x2-fold cross-validation, using the hyperparameters and candidate values displayed in Table 3.4. The data set is split in five folds, and each fold contains a training and a holdout set (Pedregosa et al., 2011). The training set also gets split in half, to create a validation set and obtain the best hyperparameter values (Borchert et al., 2022). The best performing combination of hyperparameters from the validation set is used to refit the model on the entire training set, then predict the holdout set (Dietterich, 1998). To obtain the reported evaluation metrics, we average the 10 values from the 5x2-fold cross-validation results on the holdout data for each life event.

For LSTM, we implement a binary cross-entropy loss function, with a weight assigned to positive examples to account for the imbalanced data (Paszke et al., 2019). The number of input and output neurons correspond to the number of features and number of classes, respectively, with a minibatch size of 256 (Kraus et al., 2020). The maximum number of epochs is 100, with an early stop mechanism during training if no improvement in validation loss occurs after 5 consecutive epochs (Kraus et al., 2020). The optimization algorithm, AdamW, adds decoupled weight decay regularization to the Adam optimizer (Kingma and Ba, 2015), which improves generalization performance (Loshchilov and Hutter, 2018).

Table 3.4. Hyperparameter tuning summary

Model	Hyperparameter	Values Evaluated	Reference
LR	Penalty	L2 regularization	Gattermann-Itschert and Thonemann (2021)
	Solver	Liblinear	Gattermann-Itschert and Thonemann (2021)
	Regularization C	10^{-5},...,10^{2}	Gattermann-Itschert and Thonemann (2021)
RF	Estimators	50, 100, 150, 500	Gunnarsson et al. (2021)
	Min. Samples Leaf	2	Gunnarsson et al. (2021)
	Max. Depth	1, 2, 3, 10	Gunnarsson et al. (2021)
XGB	Estimators	50, 100, 150, 500	Gunnarsson et al. (2021)
	Max. Tree Depth	1,2,3	Gunnarsson et al. (2021)
	Learning Rate	0.30, 0.40	Gunnarsson et al. (2021)
	Columns Sampled	0.60, 0.80	Gunnarsson et al. (2021)
	Rows Sampled	0.50, 0.75, 1.00	Gunnarsson et al. (2021)
LSTM	Hidden Layers	1, 2, 3	Kraus et al. (2020)
	Learning Rate	0.001, 0.005, 0.01, 0.05	Kraus et al. (2020)
	Multi-head attention	None, 8 heads	(Vaswani et al., 2017)

For this study, we use the top decile lift (TDL), area under the receiver operating characteristics curve (AUC), and F1 metric as evaluation criteria for the hyperparameter tuning and final holdout set validation, in line with extant life event prediction literature (De Caigny et al., 2020). To test for any significant differences across model results, we use the non-parametric Wilcoxon signed-ranks test or Friedman test with Nemenyi post hoc assessments (Demšar, 2006).

3.5 Results

The experimental results we report herein align with our three research questions. First, we investigate which featurization approach to represent the longitudinal data improves cross-sectional life event prediction for SaML classifiers (RQ1). We compare the life event prediction performance of two featurization approaches: the baseline (BL) and aggregation approach (AA), for all SaML classifiers under investigation. The comparative predictive performance results per life event for the SaML classifiers across various evaluation metrics appear in Table 3.5 for TDL, Table 3.6 for AUC, and Table 3.7 for F1 scores. The values in bold represent the best performing approach per life event.

Several conclusions arise from Tables 3.5, 3.6, and 3.7. In particular, life event prediction is feasible and helps building proactive marketing strategies. Most SaML classifiers in both featurization approaches exceed the random benchmark values in terms of TDL, AUC, and F1. Furthermore, the AA featurization strategy is superior to BL. With a Wilcoxon signed-ranks test, in which we compare the BL and AA approaches for a given SaML classifier, we find that for XGB, the null hypothesis is rejected at a 99% confidence level for the TDL, AUC, and F1 metric. For LR and RF, the null hypothesis is also rejected with high significance with regard to TDL and AUC, with p <0.01 in both cases. Therefore, in the following analyses, we rely on featurization of longitudinal data. Noting the superiority of the AA approach, we also investigate which SaML classifier performs best when the dynamic features from longitudinal data are represented by the aggregation approach. Using a non-parametric Friedman test of the SaML classifiers, we test whether significant differences exist across the SaML classifiers. It indicates significant differences in AUC performance ($\chi^2_F = 5.600, F(2,18) = 3.500, p = 0.061 < 0.1$) and F1 performance ($\chi^2_F = 9.800, F(2,18) = 8.647, p = 0.007 < 0.01$), but it reveals mixed results for TDL ($\chi^2_F = 4.200, F(2,18) = 2.392, p = 0.122$). The results of Nemenyi post hoc tests on AUC and F1, as depicted in Figure 3.6 and Figure 3.7, respectively, reveal that SaML classifiers with distances below the CD exhibit no significant differences and are joined by a line. Figure 3.6 and Figure 3.7 also affirm the superior performance of XGB over LR and RF, even if the difference is not statistically significant for the F1 metric.

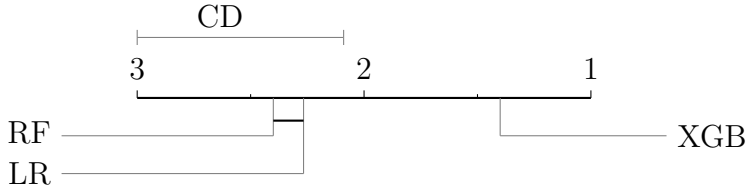


Figure 3.6. Nemenyi post hoc test for SaML classifiers with featurized data, using AUC. Notes: Classifiers that are not significantly different at $p = 0.10$ are connected.

Table 3.5. TDL results for SaML classifiers

Life Event	LR		RF		XGB	
	BL	AA	BL	AA	BL	AA
Mov	1.529	2.653	1.578	2.866	1.497	**3.406**
CarP	1.137	2.623	1.205	3.007	1.062	**3.224**
PrimR	1.301	2.845	1.368	3.331	1.449	**3.352**
SecR	0.971	**2.743**	0.972	1.771	0.914	1.343
RentR	1.536	**2.880**	1.360	2.097	1.486	2.816
RelS	1.237	2.652	1.180	2.049	1.072	**2.864**
RelE	1.067	**3.011**	1.117	2.263	1.033	2.873
JobE	1.701	2.176	1.729	2.657	1.942	**2.983**
BoC	1.085	**3.004**	1.103	2.768	1.085	2.779
Ret	1.110	3.292	1.115	3.512	1.081	**3.700**

Table 3.6. AUC results for SaML classifiers

Life Event	LR		RF		XGB	
	BL	AA	BL	AA	BL	AA
Mov	56.232	60.387	56.380	61.700	52.631	**66.730**
CarP	51.952	60.662	51.997	61.378	50.107	**62.178**
PrimR	53.458	60.924	52.126	64.420	52.819	**65.555**
SecR	50.000	50.150	50.000	50.100	50.000	**61.097**
RentR	53.427	53.485	51.152	**63.325**	50.239	54.565
RelS	52.111	60.165	50.725	52.916	50.005	**61.602**
RelE	51.147	**60.183**	50.217	51.356	50.023	59.649
JobE	56.747	60.562	56.252	61.660	58.311	**66.580**
BoC	50.542	**64.383**	50.163	51.718	50.006	57.867
Ret	51.690	65.387	50.742	66.000	50.115	**68.101**

Table 3.7. F1 results for SaML classifiers

Life Event	LR		RF		XGB	
	BL	AA	BL	AA	BL	AA
Mov	45.094	50.003	50.071	**57.203**	52.282	53.331
CarP	41.510	51.001	48.747	55.486	49.644	**55.625**
PrimR	41.004	43.112	**48.934**	46.506	25.336	28.679
SecR	49.055	40.938	49.992	49.994	49.994	**50.062**
RentR	44.862	40.313	49.489	47.984	**50.109**	49.738
RelS	38.805	42.492	49.412	47.310	49.756	**52.111**
RelE	41.717	42.244	49.426	48.387	49.910	**51.941**
JobE	45.585	44.238	51.646	**51.940**	30.777	35.349
BoC	47.682	40.085	49.677	48.573	49.989	**50.906**
Ret	44.433	56.439	50.479	63.327	48.440	**64.696**

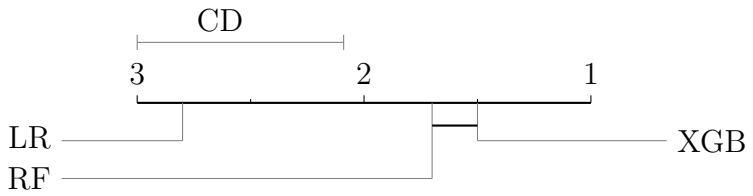


Figure 3.7. Nemenyi post hoc test for SaML classifiers with featurized data, using F1. Notes: Classifiers that are not significantly different at $p = 0.10$ are connected.

Thus, life event prediction delivers trustworthy input for proactive marketing decisions. Our results establish the added value of representing longitudinal data through an aggregation featurization approach and also identify XGB as the most preferred SaML classifier.

Second, we investigate whether LSTM, which uses sequence-ordered customer input directly, delivers superior performance relative to our best SaML, that is, XGB with aggregated featurization (XGB-AA) (RQ2). Table 3.8 summarizes the predictive performance results for each life event for both XGB-AA and LSTM, showing the best result for each life event and evaluation metric in bold.

Table 3.8. Predictive performance results for XGB-AA and LSTM

Life Event	TDL		AUC		F1	
	XGB-AA	LSTM	XGB-AA	LSTM	XGB-AA	LSTM
Mov	3.406	**3.958**	66.730	**67.581**	53.331	**61.818**
CarP	**3.224**	2.368	62.178	**66.436**	55.625	**57.692**
PrimR	3.352	**5.429**	**65.555**	64.703	28.679	**66.529**
SecR	1.343	**5.329**	61.097	**61.719**	50.062	**60.831**
RentR	2.816	**5.739**	54.565	**61.938**	49.738	**54.967**
RelS	**2.864**	2.252	61.602	**63.351**	52.111	**61.757**
RelE	**2.873**	1.762	59.649	**61.966**	51.941	**60.872**
JobE	**2.983**	2.441	66.580	**79.076**	35.349	**62.403**
BoC	2.779	**3.244**	57.867	**62.038**	50.906	**61.004**
Ret	**3.700**	1.709	68.101	**74.925**	64.696	**76.197**

As Table 3.8 reveals, LSTM outperforms XGB-AA, with average performance gains across the life events of 34% for TDL, 6% for AUC, and 33% for F1. For AUC and F1, LSTM outperforms XGB-AA on all life events (cf. PrimR, where XGB-AA performs better than LSTM on AUC); the TDL results are more mixed. Similarly, the Wilcoxon signed-ranks tests only reveal significant differences at 95% confidence levels for AUC and F1. These results confirm the beneficial impact of using sequentially ordered data with an LSTM model over a cross-sectional SaML classifier with optimal featurization.

We also note that the metrics are in line with previous life event prediction literature, as shown in De Caigny et al. (2020); Khodabakhsh et al. (2018). This also applies for other popular classification problems within decision support literature in the financial services industry, particularly under scenarios of high data imbalance, such as for bankruptcy prediction (Kou et al., 2021) and fraud detection (Baesens et al., 2021). However, any additional gains in model performance can strongly benefit a company's profits, as financial services providers are firms with a large volume of customers. Table 3.9 displays the p-values for all metrics and life events, using Wilcoxon Signed-Rank test to evaluate the performance differences between a random classifier and the best performing model in our

study, LSTM. LSTM outperforms these results by 241.15% for TDL, 32.68% for AUC, and 24.75% for F1, when averaging across life events. Finally, the table shows the higher performance from LSTM across all life events and metrics is statistically significant from the random classifier.

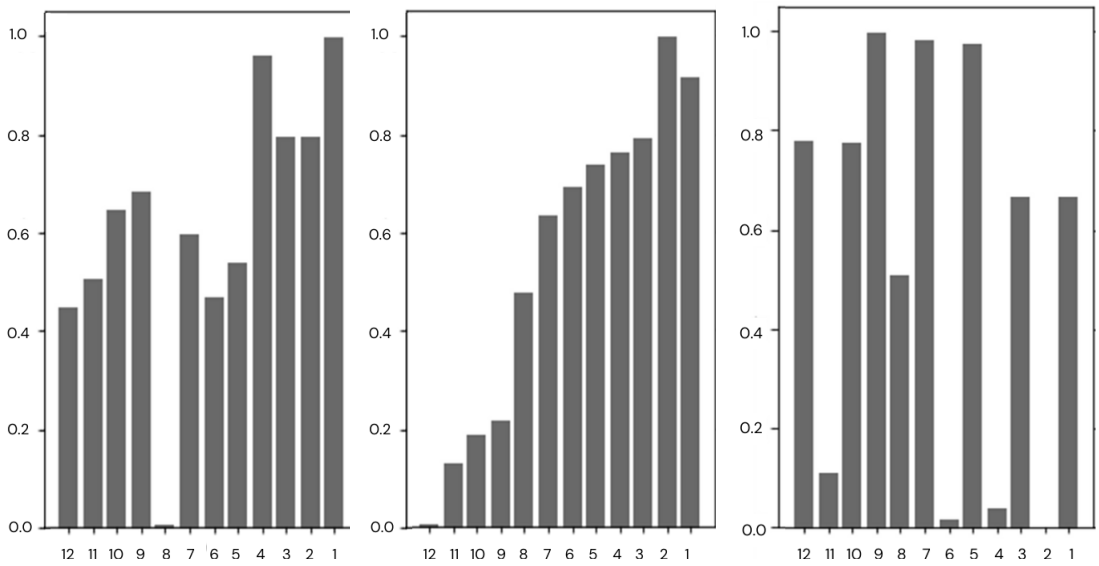
Table 3.9. Wilcoxon Signed-Rank test results for random classifier against LSTM

Life Event	TDL	AUC	F1
Mov	.002***	.002***	.002***
CarP	.004***	.002***	.084*
PrimR	.004***	.002***	.002***
SecR	.009***	.002***	.002***
RentR	.004***	.002***	.084*
RelS	.002***	.002***	.002***
RelE	.002***	.002***	.002***
JobE	.002***	.002***	.002***
BoC	.002***	.002***	.002***
Ret	.002***	.002***	.002***

***p-value<.01, **p-value<.05, *p-value<.10

We also investigate the drivers of LSTM, the best performing life event prediction model, using the IG attribution method (Sundararajan et al., 2017). The objective is to understand the importance of life event drivers to answer RQ3, particularly as additional tools that may be useful for marketing decision-making. Among the many types of life events, we focus on two in detail, according to their superior F1 values: Ret and PrimR. Discussions of the other life events are available in 3.9.1 and 3.9.1. First, with regard to the contributions of each time step to final life event prediction, Figure 3.8a displays the mean attributions per time step for customers undergoing retirement (Ret); Figure 3.8b contains the mean attributions for customers buying their first residence (PrimR); and Figure 3.8c shows the attributions for customers without either life event. The Y-axis represents the mean attribution value; the closer an attribution value is to 1, the higher its contribution for predicting the given life event. The X-axis represents time steps, where 1 is the closest and 12 is the furthest month relative to the prediction period. In line with other life events, as exhibited in 3.9.1, attribution values increase for time steps closer to the prediction period. In contrast, the absence of life events does not establish a generalizable pattern. From a marketing decision-making perspective, such results would reveal the optimal moment to contact customers, which could enable more efficient resource allocations. The features to predict the purchase of a primary residence have more weight the closer to the occurrence of the event. This suggests the possibility of progressive changes in customer behavior. Previous research shows the correct and timely allocation of marketing resources, through accurate understanding of dynamic customer behavior, has a positive impact on customer loyalty (Han and Anderson, 2022). Therefore, it may be of interest to evaluate customer response when using life event predictions as a targeting tool.

Second, we consider the normalized mean attribution values. The 10 features with the highest attribution values display different patterns for Ret and PrimR, as shown in Figures 3.9a and 3.9b. In detail, the top Ret features suggest that an intense relationship encourages the life event occurrence, as expressed in the main bank feature (main bank [Y/N]), number of insurance-related accounts (insurance savings, housing savings accounts), and payment amounts (online transfers, estimated recurring expenses). But PrimR prediction assigns more importance to asset amounts (high risk credit, active

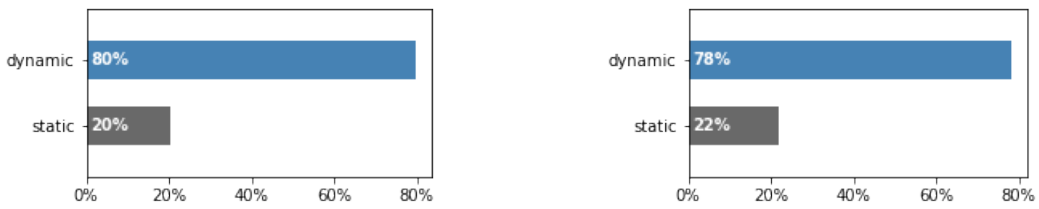


(a) Life event: Ret **(b)** Life event: PrimR **(c)** No life event

Figure 3.8. Integrated gradients showing mean normalized attributions per time step

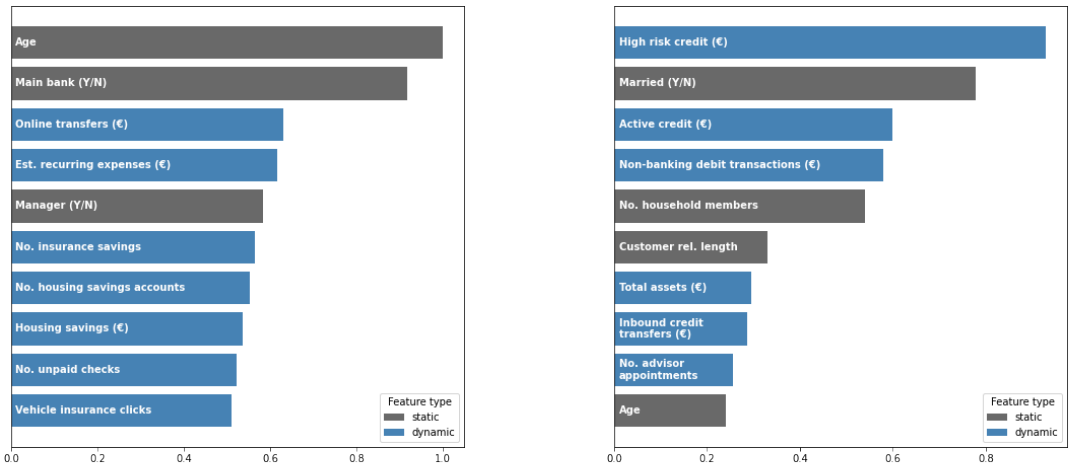
credit), the demographic situation (married [Y/N], household members), and transaction data (non-banking debit transactions, inbound credit transfers). With these insights, marketing decision-makers can use their good life event predictions to personalize the services they offer, detect new cross-selling opportunities, and improve their product recommendations.

Third, we present the mean attributions by static and dynamic features, to investigate their respective importance. Overall, static and dynamic features respectively represent around 30% and 70% of the total features, but for both life events, dynamic features are over-represented compared with the average. That is, for Ret, dynamic features are around 80% of the average attribution, and static features account for 20%. Similarly, PrimR reveals that 78% of attributions stem from dynamic features, versus 22% from static features. Thus, both static and dynamic features contribute to life event prediction, but the dynamic features have a relatively greater impact.



(a) Life Event: Ret **(b)** Life Event: PrimR

Figure 3.10. Integrated gradients showing the distribution of mean normalized attributions per feature type (%)



(a) Life Event: Ret **(b)** Life Event: PrimR

Figure 3.9. Integrated gradients showing mean normalized attributions for the top 10 features

3.6 Discussion

IG is a flexible technique, that allows for aggregation of different variables to uncover which features drive a model's prediction for particular life events. Furthermore, these feature importances can be paired against retention rates for further managerial insights, particularly in terms of establishing differences between customer behavior per life event cohort. Thus, we propose that the observed differences in retention rates are further proof of life events being a valuable source of information for CRM applications. In particular, they suggest customers potentially need to be addressed differently, depending on the life event experienced.

We present relative differences in retention rates, 12 months after our experimental period as defined in Section 3.4. The retention rate for customers who did not experience a life event is the baseline, shown as 0% in Figure 3.11.

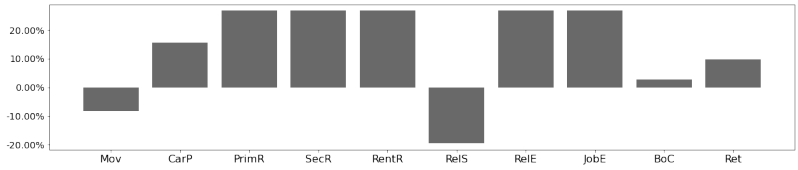


Figure 3.11. Retention rates per life event (%), using no life event occurrence as a baseline (0%).

The occurrence of a life event reveals a higher retention rate than this baseline. Specifically, life events such as CarP, JobE, PrimR, SecR, RentR, BoC, RelE, and Ret drive loyalty and retention. However, customers who experience Mov or RelS are at risk of attrition; their retention rates are lower than those of customers without life events for the year after their life event occurrence. For managers, such life event predictions offer a way to improve the overall customer experience and provide a better, more personalized

service in a timely manner. In particular, life event prediction might inform real-time adaptations of product recommendations offered through digital channels. Alternatively, this information could be harnessed by a bank advisor, proactively proposing a meeting with customers exhibiting a high life event probability in the upcoming months. Life event prediction also opens up additional avenues of research, where life event occurrences can be used to evaluate their impact or causality onto other tasks. For instance, life events can be studied alongside the purchase of specific products, thus evaluating their impact on cross-selling measures, uncovering possible causal relationships with purchase intention. Life events could also be evaluated as a complementary tool for churn prevention measures, to ultimately assess whether churn rates can be decreased or whether customers experiencing certain life events are more responsive to specific retention actions. Finally, life events can also be developed as a tool to develop more complex customer segments, to further personalize CRM initiatives.

3.7 Conclusions and Further Research

This study contributes to life event prediction literature by introducing novel life events and analyzing their impact on relevant dimensions for decision-makers, on the basis of data provided by a European financial service provider. Furthermore, we compare different approaches for incorporating longitudinal customer data into a cross-sectional SaML classification setup, and we implement a state-of-the-art LSTM. We structure this study, and our discussion of its conclusions, according to our three central research questions.

RQ1. Which featurization approach for representing longitudinal customer data improves cross-sectional life event prediction performance for conventional SaML classifiers?

We find aggregation as a featurization approach performs better than using monthly snapshot data (baseline), for all SaML classifiers across all life events. Furthermore, XGB with aggregated dynamic features offers the best predictive performance.

RQ2. Does LSTM improve life event prediction performance over SaML classifiers when longitudinal customer data are available?

Our results establish that the LSTM model outperforms the best cross-sectional approach, namely, the XGB model with aggregated features.

RQ3. Is it possible to identify life event drivers that are useful for marketing decision-making?

The months closer to the prediction period exert greater weight in terms of prediction performance. We offer specific actionability insights for the top 10 performing drivers for each life event. Dynamic features influence life event prediction more than static features. In turn, life event prediction is an important tool for customer relationship management, due to its notable effect on retention rates.

In addition to contributing to life event prediction literature, this study reveals various paths for further research. First, we demonstrate the importance of explaining model outcomes for decision-makers that rely on life event prediction model using an attribution method. Yet, additional research is needed to determine the impacts of various attribution methods to further assist decision-makers with model explanations in the life event prediction field. Second, tracking various customer behaviors across life events may help marketers further improve their personalized marketing strategies, which should evoke stronger customer engagement and customer responses (Han and Anderson, 2022). For example, researchers might run A/B tests on highly personalized recommendations for

customers approaching life events to gain a dynamic understanding of those behaviors (De Caigny et al., 2020; Cheng and Chen, 2022). Third, we show that the dynamic features derived from longitudinal data contain important information for life event prediction. Hence, exploring longer periods of historical customer data might reveal additional information relevant for deploying marketing actions, especially when related to specific life events farther in advance. Furthermore, applying our framework onto different data sources from multiple companies or industries is an interesting future avenue of research, to either strengthen the generalization of our findings or improve upon our conclusions.

3.8 References

Andreasen, A.R., 1984. Life status changes and changes in consumer preferences and satisfaction. J. Consum. Res. 11, 784–794. doi:10.1086/209014.

Antonides, G., Van Raaij, W.F., 1999. Consumer behaviour:: A european perspective. Eur J Mark 33, 1–2.

Baesens, B., Höppner, S., Verdonck, T., 2021. Data engineering for fraud detection. Decis Support Syst 150, 113492. doi:10.1016/j.dss.2021.113492. interpretable Data Science For Decision Making.

Bagnall, A.J., Bostrom, A., Large, J., Lines, J., 2017. The great time series classification bake off: An experimental evaluation of recently proposed algorithms. Data Min. Knowl. Disc. 31, 606–660. doi:10.1007/s10618-016-0483-9.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY, USA.

Bogaert, M., Lootens, J., Van den Poel, D., Ballings, M., 2019. Evaluating multi-label classifiers and recommender systems in the financial service sector. Eur J Oper Res 279, 620–634. doi:10.1016/j.ejor.2019.05.037.

Borchert, P., Coussement, K., De Caigny, A., De Weerdt, J., 2022. Extending business failure prediction models with textual website content using deep learning. Eur. J. Oper. Res. doi:10.1016/j.ejor.2022.06.060.

Boulding, W., Staelin, R., Ehret, M., Johnston, W.J., 2005. A customer relationship management roadmap: What is known, potential pitfalls, and where to go. J. Mark. 69, 155–166. doi:10.1509/jmkg.2005.69.4.155.

Brahma, A., Goldberg, D.M., Zaman, N., Aloiso, M., 2021. Automated mortgage origination delay detection from textual conversations. Decis Support Syst 140, 113433. doi:10.1016/j.dss.2020.113433.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324.

Bunnell, L., Osei-Bryson, K.M., Yoon, V.Y., 2020. Finpathlight: Framework for an multiagent recommender system designed to increase consumer financial capability. Decis Support Syst 134, 113306. doi:10.1016/j.dss.2020.113306.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. p. 785–794. doi:10.1145/2939672.2939785.

Chen, Z.Y., Fan, Z.P., Sun, M., 2012. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. Eur. J. Oper. Res. 223, 461–472. doi:10.1016/j.ejor.2012.06.040.

Cheng, L.C., Chen, K., 2022. Mining longitudinal user sessions with deep learning to extend the boundary of consumer priming. Decis. Support Syst. 162, 113864. doi:10.1016/j.dss.2022.113864.

Chollet, F., et al., 2015. Keras. Accessed June, 2023. URL: https://keras.io.

Coussement, K., Benoit, D.F., 2021. Interpretable data science for decision making. Decis. Support Syst. 150, 113664. doi:10.1016/j.dss.2021.113664.

Coussement, K., Lessmann, S., Verstraeten, G., 2017. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decis. Support Syst. 95, 27–36.

Crone, S.F., Lessmann, S., Stahlbock, R., 2006. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. Eur. J. Oper. Res. 173, 781–800. doi:10.1016/j.ejor.2005.07.023.

De Bock, K.W., Coussement, K., De Caigny, A., Slowiński, R., Baesens, B., Boute, R.N., Choi, T.M., Delen, D., Kraus, M., Lessmann, S., et al., 2023. Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. Eur J Oper Res .

De Caigny, A., Coussement, K., De Bock, K.W., 2020. Leveraging fine-grained transaction data for customer life event predictions. Decis. Support Syst. 130, 113232. doi:10.1016/j.dss.2019.113232.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30. URL: http://jmlr.org/papers/v7/demsar06a.html.

Dietterich, T.G., 1998. Approximate statistical test for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923. doi:10.1162/089976698300017197.

Dixon, M., Klabjan, D., Bang, J.H., 2015. Implementing deep neural networks for financial market prediction on the Intel Xeon Phi, in: Proceedings of the 8th Workshop on High Performance Computational Finance, ACM, New York, NY, USA. pp. 1–6. doi:10.1145/2830556.2830562.

Fischer, T., Krauss, C., 2018. Deep learning with long short-term memory networks for financial market predictions. Eur. J. Oper. Res. 270, 654–669. doi:10.1016/j.ejor.2017.11.054.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29, 1189–1232. doi:10.1214/aos/1013203451.

Gattermann-Itschert, T., Thonemann, U.W., 2021. How training on multiple time slices improves performance in churn prediction. Eur. J. Oper. Res. 295, 664–674. doi:10.1016/j.ejor.2021.05.035.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51. doi:10.1145/3236009.

Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., Lemahieu, W., 2021. Deep learning for credit scoring: Do or don't? Eur. J. Oper. Res. 295, 292–305. doi:10.1016/j.ejor.2021.03.006.

Han, S., Anderson, C.K., 2022. The dynamic customer engagement behaviors in the customer satisfaction survey. Decis. Support Syst. 154, 113708. doi:10.1016/j.dss.2021.113708.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. doi:10.1162/neco.1997.9.8.1735.

Hoehle, H., Scornavacca, E., Huff, S., 2012. Three decades of research on consumer adoption and utilization of electronic banking channels: A literature analysis. Decis Support Syst 54, 122–132. doi:10.1016/j.dss.2012.04.010.

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. volume 398. John Wiley & Sons. doi:10.1002/9781118548387.

Huang, Y., Meng, S., 2019. Automobile insurance classification ratemaking based on telematics driving data. Decis. Support Syst. 127, 113156. doi:10.1016/j.dss.2019.113156.

Idbenjra, K., Coussement, K., De Caigny, A., 2024. Investigating the beneficial impact of segmentation-based modelling for credit scoring. Decis Support Syst 179, 114170. doi:10.1016/j.dss.2024.114170.

Kamakura, W.A., Ramaswami, S.N., Srivastava, R.K., 1991. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. Int. J. Res. Mark. 8, 329–349. doi:10.1016/0167-8116(91)90030-B.

Kamakura, W.A., Wedel, M., de Rosa, F., Mazzon, J.A., 2003. Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction. Int J Res Mark 20, 45–65. doi:10.1016/S0167-8116(02)00121-0.

Khodabakhsh, M., Kahani, M., Bagheri, E., Noorian, Z., 2018. Detecting life events from twitter based on temporal semantic features. Knowl.-Based Syst. 148, 1–16. doi:10.1016/j.knosys.2018.02.021.

Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M.C., Johnson, J., 2020. Can deep learning predict risky retail investors? a case study in financial risk behavior forecasting. Eur. J. Oper. Res. 283, 217–234. doi:10.1016/j.ejor.2019.11.007.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, pp. 1–13.

Knott, A., Hayes, A., Neslin, S.A., 2002. Next-product-to-buy models for cross-selling applications. J. Interact. Mark. 16, 59–75.

Koschate-Fischer, N., Hoyer, W.D., Stokburger-Sauer, N.E., Engling, J., 2018. Do life events always lead to change in purchase? the mediating role of change in consumer innovativeness, the variety seeking tendency, and price consciousness. J. Acad. Mark. Sci. 46, 516–536. doi:10.1007/s11747-017-0548-3.

Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., Kou, S., 2021. Bankruptcy prediction for smes using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140, 113429. doi:10.1016/j.dss.2020.113429.

Kraus, M., Feuerriegel, S., Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. Eur. J. Oper. Res. 281, 628–641. doi:10.1016/j.ejor.2019.09.018.

Krauss, C., Do, X.A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. Eur. J. Oper. Res. 259, 689–702. doi:10.1016/j.ejor.2016.10.031.

Kumar, V., George, M., Pancras, J., 2008. Cross-buying in retailing: Drivers and consequences. J Retail 84, 15–27.

Kumar, V., Ramachandran, D., Kumar, B., 2021. Influence of new-age technologies on marketing: A research agenda. J. Bus. Res. 125, 864–877. doi:10.1016/j.jbusres.2020.01.007.

LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R., 2012. Efficient BackProp. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-35289-8_3.

Lee, E., Moschis, G.P., Mathur, A., 2001. A study of life events and changes in patronage preferences. J. Bus. Res. 54, 25–38. doi:10.1016/S0148-2963(00)00116-8.

Lin, Q., Jia, N., Chen, L., Zhong, S., Yang, Y., Gao, T., 2023. A two-stage prediction model based on behavior mining in livestream e-commerce. Decis Support Syst 174, 114013. doi:10.1016/j.dss.2023.114013.

Loshchilov, I., Hutter, F., 2018. Fixing weight decay regularization in Adam, in: ICLR 2018, Vancouver, BC, Canada.

Luo, X., Li, H., Zhang, J., Shim, J., 2010. Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services. Decis Support Syst 49, 222–234. doi:10.1016/j.dss.2010.02.008.

Malthouse, E.C., 2007. Mining for trigger events with survival analysis. Data Min. Knowl. Discov. 15, 383–402. doi:10.1007/s10618-007-0074-x.

Mathur, A., Moschis, G., Lee, E., 2003. Life events and brand preference changes. J. Consum. Behav. 3, 129–141. doi:10.1002/cb.128.

Mathur, A., Moschis, G.P., Lee, E., 2008. A longitudinal study of the effects of life status changes on changes in consumer preferences. J. Acad. Mark. Sci. 36, 234—-246. doi:10.1007/s11747-007-0021-9.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

Rousseeuw, P.J., Hubert, M., 2011. Robust statistics for outlier detection. Wiley interdisciplinary reviews: Data Min. Knowl. Discov. 1, 73–79. doi:10.1002/widm.2.

Sahoo, N., Singh, P.V., Mukhopadhyay, T., 2012. A hidden Markov model for collaborative filtering. Manag. Inf. Syst. Q. 36, 1329–1356. doi:10.2307/41703509.

Sin, L., Tse, A., Yim, F., 2005. CRM: conceptualization and scale development. Eur. J. Mark. 39, 1264–1290. doi:10.1108/03090560510623253.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958. URL: https://jmlr.org/papers/v15/srivastava14a.html.

Subramanian, H., Angle, P., Rouxelin, F., Zhang, Z., 2024. A decision support system using signals from social media and news to predict cryptocurrency prices. Decis Support Syst 178, 114129. doi:10.1016/j.dss.2023.114129.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, PMLR. p. 3319–3328.

Turbé, H., Bjelogrlic, M., Lovis, C., Mengaldo, G., 2023. Evaluation of post-hoc interpretability methods in time-series classification. Nat. Mach. Intell. 5, 250–260. doi:10.1038/s42256-023-00620-w.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Curran Associates Inc., New York, NY, USA. p. 6000–6010.

Verhoef, P.C., Donkers, B., 2001. Predicting customer potential value an application in the insurance industry. Decis Support Syst 32, 189–199. doi:10.1016/S0167-9236(01)00110-5. decision Support Issues in Customer Relationship Management and Interactive Marketing for E-Commerce.

Wang, G., Ma, J., Chen, G., 2023. Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. Decis. Support Syst. 167, 113913. doi:10.1016/j.dss.2022.113913.

Wang, P., Li, J., Hou, J., 2021. S2SAN: A sentence-to-sentence attention network for sentiment analysis of online reviews. Decis. Support Syst. 149, 113603. doi:10.1016/j.dss.2021.113603.

Wang, Y., Huang, M., Zhu, X., Zhao, L., 2016. Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 606–615. doi:10.18653/v1/D16-1058.

Yi, Z., Liang, Z., Xie, T., Li, F., 2023. Financial risk prediction in supply chain finance based on buyer transaction behavior. Decis. Support Syst. 170, 113964. doi:10.1016/j.dss.2023.113964.

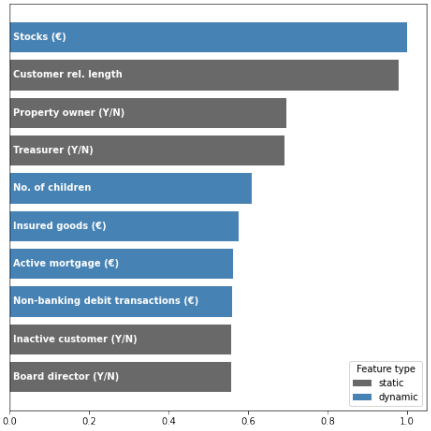
Zhong, C., Du, W., Xu, W., Huang, Q., Zhao, Y., Wang, M., 2023. LSTM-ReGAT: A network-centric approach for cryptocurrency price trend prediction. Decis. Support Syst. 169, 113955. doi:10.1016/j.dss.2023.113955.

Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., Vanthienen, J., 2018. Time series for early churn detection: Using similarity based classification for dynamic networks. Expert Syst. Appl. 106, 55–65. doi:10.1016/j.eswa.2018.04.003.

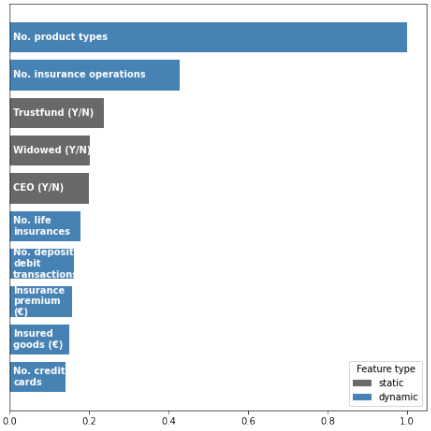
3.9 Appendix

3.9.1 Additional Experimental Results

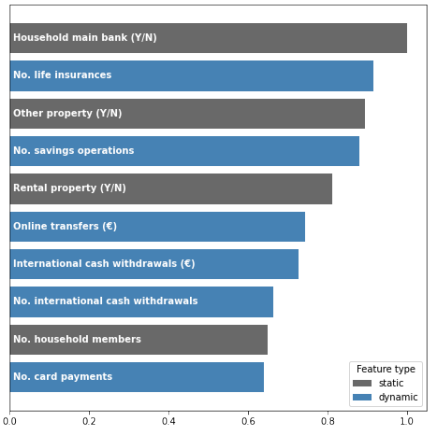
Top 10 features, by importance, per life event



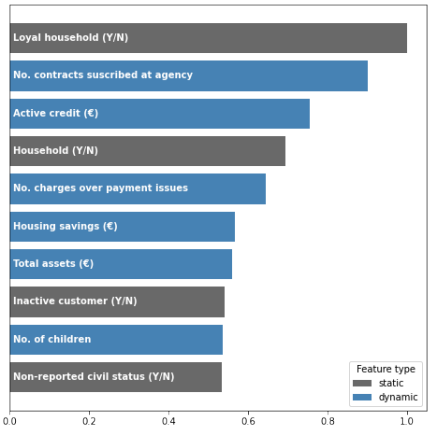
(a) Mov



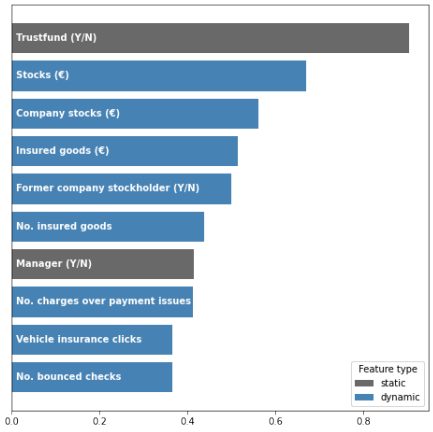
(b) CarP



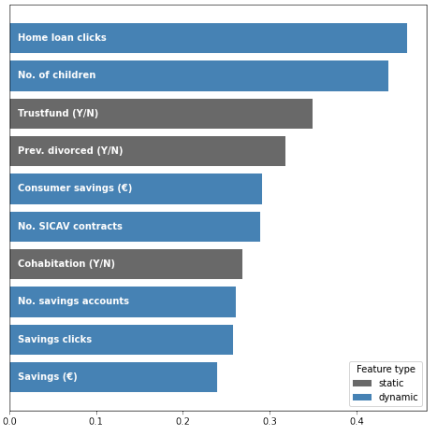
(a) SecR



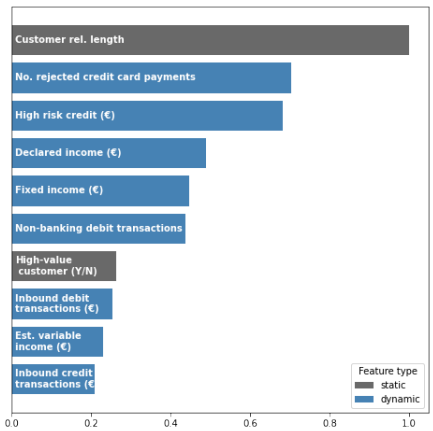
(b) RentR



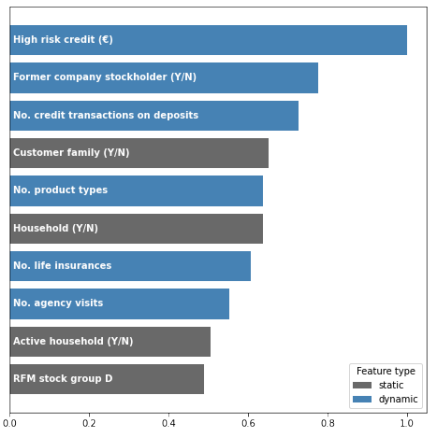
(a) RelS



(b) RelE

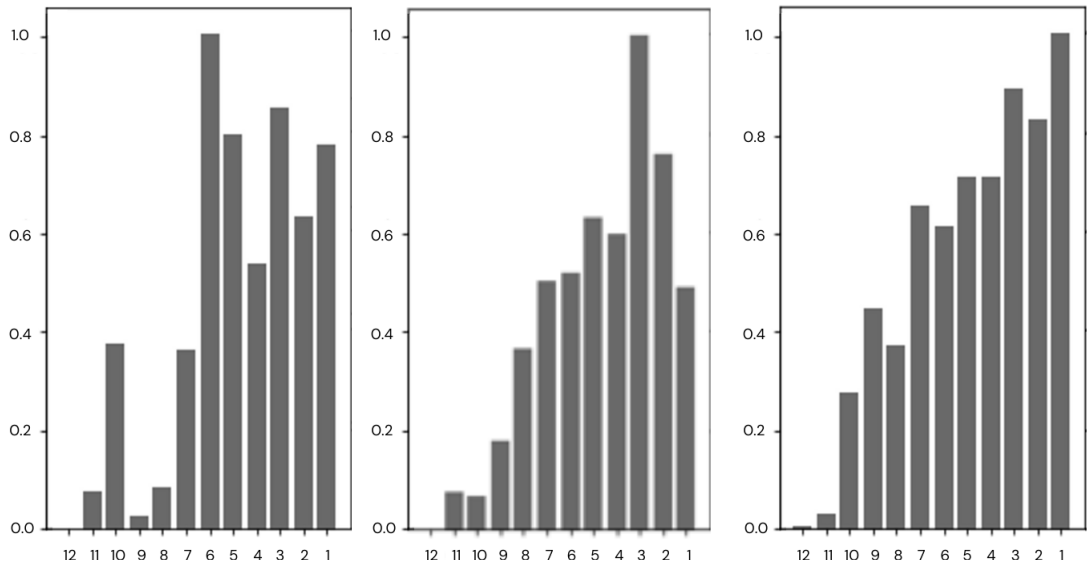


(a) JobE

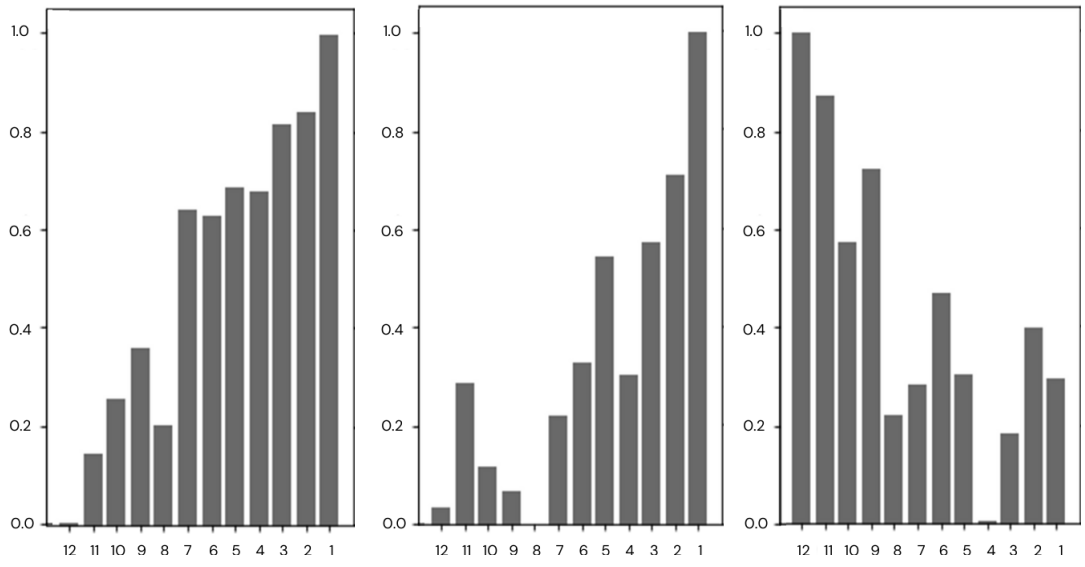


(b) BoC

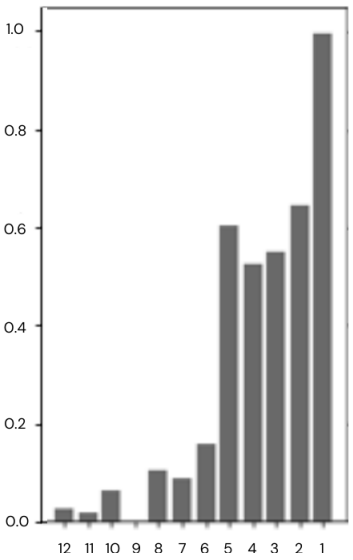
Time importance per life event



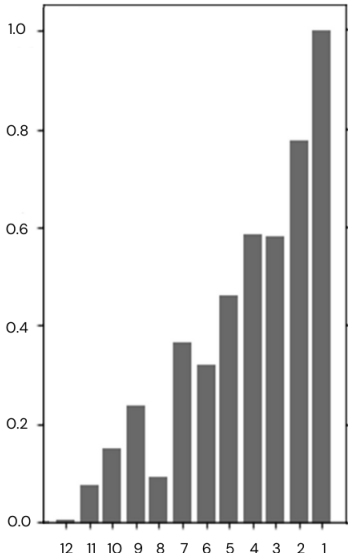
(a) Mov **(b)** CarP **(c)** SecR



(a) RentR **(b)** RelS **(c)** RelE


1.0
0.8
0.6
0.4
0.2
0.0
12 11 10 9 8 7 6 5 4 3 2 1

(a) JobE


1.0
0.8
0.6
0.4
0.2
0.0
12 11 10 9 8 7 6 5 4 3 2 1

(b) BoC

Chapter 4: Longitudinal Data for Recommender Systems in the Financial Services Industry

CHAPTER 4

Longitudinal Data for Recommender Systems in the Financial Services Industry

Abstract.

Recommender systems (RS) are highly relevant for multiple domains, allowing to construct personalized suggestions for consumers. Previous studies have strongly focused on collaborative filtering approaches, but the inclusion of longitudinal data (LD) has received limited attention. To address this gap, we investigate the impact of incorporating LD for recommendations, comparing traditional collaborative filtering approaches, machine learning multi-label classifier (MLC) algorithms, and a deep learning model (DL) in the form of gated recurrent units (GRU). Additional analysis for the best performing model is provided through SHapley Additive exPlanations (SHAP), to uncover relations between the different recommended products and features. Thus, this article contributes to operational research literature by (1) comparing several MLC techniques and RS, including state-of-the-art DL models in a real-life scenario, (2) the comparison of various featurization techniques to assess the impact of incorporating LD on MLC performance, (3) the evaluation of LD as sequential input through the use of DL models, (4) offering interpretable model insights to improve the understanding of RS with LD. The results uncover that DL models are capable of extracting information from longitudinal features for overall higher and statistically significant performance. Further, SHAP values reveal that LD has the higher impact on model output and managerial relevant temporal patterns emerge across product categories.

Keywords: Analytics, Recommender Systems, Deep Learning, Longitudinal Data.

4.1 Introduction

Recommender systems (RS) are a dynamic area of research with multiple applications. They excel in offering personalized suggestions based on consumers' needs, interests and preferences. This enhances the decision-making process by optimizing resource allocation and targeting relevant customers (Geuens et al., 2018). This translates to increased switching costs for customers (Kumar et al., 2008) and improved retention rates (Kamakura et al., 2003). This is crucial for the financial-services industry, as existing customers are more profitable than new ones (Knott et al., 2002). Therefore, further RS research adds value by improving customer relations and the effectiveness cross-selling opportunities in advance.

Traditionally categorized into collaborative filtering, content-based, and hybrid models (Khanal et al., 2020), RS rely on large matrices for similarity calculations. Collaborative filtering computes recommendations based on past user behavior or ratings of available items, content-based recommendations exploits item characteristics to identify similarities with a user's historical purchasing behavior; and hybrid models combine both approaches.

Personalized recommendations, extensively studied in retail, pose challenges in the financial sector due to a lower product count, longer renewal cycles, and limited user interactions (Musto et al., 2015). However, recent OR literature confirms that recommendations can also be performed using machine learning multi-label classifiers (MLC).

MLC fall into two main categories: problem transformation methods (PT) and algorithm adaptation methods (AA) (Bogaert et al., 2019). Within PT, there are binary relevance (BR) and classifier chain (CC) algorithms. BR involves building a separate model for each product label, while CC makes iterative predictions for each product, incorporating new predictions into subsequent classifications. AA involves modifying algorithms to suit the multi-label task, enabling them to output probabilities for all product labels simultaneously.

Both RS and MLC capture a customer's behavior using cross-sectional data, without reflecting evolving interests in time (You et al., 2019). Customer behavior, however, is a series of decisions leading to purchase sequences (Prinzie and Van den Poel, 2006). Thus, exploring longitudinal data (LD) is a valuable research path to extend RS literature in OR. Deep learning (DL) models are particularly effective in modeling customer behavior as a sequence of actions or events (Tan et al., 2016; You et al., 2019). While this approach has been applied to evaluate recommendation effectiveness over time (Zhang et al., 2020; Ferraro et al., 2020), its performance in comparison to other recommendation methods remains unexplored.

Our data is provided by a large European financial services provider. The data contains product ownership data for 35 distinct products and their respective recency, frequency, and monetary (RFM) information. Recency refers to the number of days since the last product purchase, frequency considers the total number of distinct products, and monetary reflects the overall income from product ownership and use.

Demographic data is also available for 384,859 customers. This data hase been previously cleaned to select relevant customers and products with low sparsity, using different data combinations and recommendation approaches to find the best performing techniques.

Previous research has shown that static RFM features are relevant for generating recommendations (Bogaert et al., 2019). Moreover, RFM features evolve in time and are naturally dynamic, but are excellent at capturing purchasing behaviors (Chen et al.,

2016). Therefore, their use for predictive tasks remains an ongoing field of research (Mena et al., 2023).

Literature shows that a research gap remains in comparing RS, MLC, and DL, especially in leveraging the benefits of LD. We extend Bogaert et al. (2019) by evaluating the impact of incorporating LD, in the form of RFM features, on recommendation performance. Furthermore, state-of-the-art DL algorithms are included in the benchmark framework, to evaluate their performance using LD in the financial services sector. DL models have the advantage of being able to process LD as a sequential input, without the need of data aggregation, providing an advantage over alternatives like feature aggregation, as demonstrated in previous research (Mena et al., 2023).

Finally, we deploy SHapley Additive exPlanations (SHAP) to better understand the output of our best performing models. SHAP assigns an importance value for each feature and is adept at handling MLC algorithms (Lundberg and Lee, 2017). SHAP is a popular feature-attribution mechanism for interpretable AI, with applications in diverse operational research (OR) predictive tasks, such as credit scoring (Chen et al., 2024), process monitoring (Stevens and De Smedt, 2023), and corporate default risk (Sigrist and Leuenberger, 2023). Notably, SHAP maintains properties of local accuracy and consistency, unlike other interpretability methods (Notz and Pibernik, 2024). Thus, SHAP aids in identifying features strongly correlated with each product.

In sum, we perform an exhaustive benchmark to identify the best approach for cross-sell analysis in the financial services sector. This is done by comparing MLC techniques against RS through the comparison of different models, outlined in Section 4.3. Additionally, we methodologically contribute to the RS literature in OR by using DL models to handle longitudinal financial features, incorporating them into the benchmarking framework.

Thus, the research questions we address in this paper are the following:

RQ1. Does the use of various RS methods lead to significantly distinct levels of performance?

RQ2. Does the utilization of various featurization methods for longitudinal data yield significantly different levels of performance?

RQ3. Does the incorporation of longitudinal data using state-of-the-art DL models significantly outperform other recommendation approaches?

RQ4. Can we identify relevant features for product recommendations in the financial services industry?

First, we follow previous research (Bogaert et al., 2019), by comparing traditional RS approaches in the form of (i) item-based collaborative filtering (IBCF), (ii) user-based collaborative filtering (UBCF), (iii) association rule-based recommender (AR), (iv) popular items recommender (PR), and (v) random items recommender (RR). We also implement post-hoc significance testing to confirm the best performing approach.

Secondly, we transform the data using different featurization methods to uncover the best performing approach for MLC models, by comparing no data transformation against singular value decomposition (SVD) and principal component analysis (PCA), which are frequently used in recommendation settings (Geuens et al., 2018; Coussement et al., 2022). These comparisons are performed using multiple algorithms under BR, CC, and AA approaches. Both BR and CC include (i) adaboost (AB), (ii) shallow forward neural network (FNN), (iii) naive Bayes (NB), (iv) random forest (RF), and (v) xgboost (XGB). Algorithm adaptation includes (i) FNN, (ii) RF, and (iii) XGB. We also analyze

whether significant differences in performance are found between AA, BR, and CC MLC approaches.

Thirdly, we compare the top performing MLC and RS approaches against DL models which are capable of incorporating LD as sequential input, with no additional data transformation. The DL models are gated recurrent units (GRU), which is capable of producing multiple outputs, as are AA approaches. This allows us to assess whether the use of DL adds value when generating recommendations in the financial services industry, as well as uncovering the best approach and structure for the task at hand.

Fourth, we implement SHAP on our best performing model, to extract the features with the highest impact on the recommendations. This has not been previously done for recommendations in the financial services industry, constituting a novel application of great interest for constructing managerial insights.

Therefore, our contributions are fourfold. First, the comparison of several MLC techniques and RS, including state-of-the-art DL models in a real-life scenario. Second, the comparison of various featurization techniques to assess if the incorporation of longitudinal data improves MLC performance. Third, the evaluation of LD as a sequential input, through the use of DL models. Fourth, the application of interpretability techniques to improve the understanding of decision-makers and marketers when deploying RS. Overall, these contributions uncover a better understanding on how to pre-process data and generate recommendations in the financial services sector.

4.2 Related Work

We evaluate MLC against traditional RS in the financial services sector, using LD for cross-selling purposes. As such, this literature review covers two areas of OR research. First, we review studies with cross-selling applications with real-world data from the financial services industry in Section 4.2.1. Second, we analyze the use of LD for recommendations in Section 4.2.2.

These algorithms produce multiple outputs. However, next-product-to-buy (NPTB) models prioritize one product per customer (Knott et al., 2002; Prinzie and Van den Poel, 2008). Therefore, they are an inadequate comparison against RS approaches and are excluded from our literature review. Similarly, when examining session-based recommendations, which use a sequence of interactions from an online session, we exclude studies that only predict the next event or interaction (Hidasi et al., 2015).

4.2.1 Recommendations for financial services

Several studies have explored different techniques regarding the best way to work with financial services data to output recommendations. New clients have a lower profitability than existing customers who continue to use existing products, as well as subscribing to new ones (Knott et al., 2002). However, a big challenge lies in the low rate of customer response to solicitations (Li et al., 2011). Thus, the inclusion of LD has proven essential to detect changes in customer behavior in time (Li et al., 2011). This allows for a more effective resource allocation for marketing actions by selecting customers with a higher probability of responding to solicitations.

For instance, Li et al. (2005) assess changes in customer demand for different financial products over time, finding different demographic characteristics affect the speed with which financial needs evolve. They incorporate LD as input for a MLC algorithm and

find that the ownership of multiple products drives the purchase of other additional products. Musto et al. (2015) also use demographic characteristics, as well as financial assets, and risk and investment profiles, to recommend a set of portfolios to a target user. This data, in cross-sectional form, is used to calculate similarity between clients. They compare multiple MLC algorithms, without using LD. Additionally, Bogaert et al. (2019) compare several MLC techniques against collaborative filtering RS. Using cross-sectional data, they output recommendations for multiple banking products. Their results show MLC significantly outperforms RS, with adaboost being the best performing algorithm. Boulenger et al. (2022) propose using DL models with attention mechanisms to process product ownership as LD, as well as demographic information. They follow a MLC approach, recommending multiple products such as debit, credit cards or term deposits. Chou et al. (2022) extends the collaborative filtering framework (RS), incorporating LD from historical transaction records as input for their model. Different relations from the transactional data are represented through a graph structure, combined with DL, for more efficient data usage, lower computational cost, and recommendation performance improvements.

Table 4.1 summarizes the previously described studies to support our conclusions. The first column contains the relevant study. The following columns use a check-mark to respectively indicate if a study uses a traditional RS approach; if MLC algorithms are included; if DL models are evaluated; if LD is incorporated as input; and if interpretability techniques (IT) are used to analyze the model output.

From Table 4.1, we find a research gap remains with regards to recommendations using LD in the financial services sector. This is an interesting area of research as previous studies have shown LD can be used to capture a customer's behavioral changes through time (Li et al., 2011), while models that do not account for temporal ordering of products tend to exhibit lower performances (Li et al., 2005).

In sum, we conclude that personalized recommendations in the financial services industry have been evaluated with both MLC and RS approaches, although it remains rare to find studies that compare both. Further, it remains unclear how DL models perform in comparison to MLC and RS. Moreover, LD has been shown to be relevant for capturing customer behavior changing through time (Li et al., 2005, 2011), but it remains to be compared against cross-sectional approaches commonly used for RS and MLC (Bogaert et al., 2019). Finally, no studies have incorporated interpretability techniques to further elucidate a model's results or to analyze the interactions between different features.

Table 4.1. Literature review for studies on RS for financial services

Study	RS	MLC	DL	LD	IT
Li et al. (2005)		✓		✓	
Musto et al. (2015)		✓			
Bogaert et al. (2019)	✓	✓			
Boulenger et al. (2022)		✓	✓	✓	
Chou et al. (2022)	✓		✓	✓	
This study	✓	✓	✓	✓	✓

4.2.2 Longitudinal data for recommendations

Research shows that customers change their purchasing behavior in time, depending on their financial maturity (Li et al., 2005; Kamakura et al., 1991). Moreover, the use of LD as input is particularly useful for capturing evolving customer interests and to provide

highly personalized recommendations (Quadrana et al., 2018). As such, we review recent studies that include LD for recommendation, to develop an adequate experimental setup.

Rendle et al. (2010) propose generating transition matrices for each user, applying matrix factorization onto them to be used as input for Markov chains. The approach adapts collaborative filtering, thus falling under a traditional RS setting, while Markov chains use sequential input.

Tan et al. (2016) propose the use of recurrent neural networks (RNN) for session-based recommendations (SBR). As an output, they produce a ranking over all the next items that may occur in that session, with the top-k items being recommended. Therefore, their study uses a DL model and sequential input, capable of producing multiple outputs, as in a MLC setting.

You et al. (2019) also use DL models in the form of RNNs. As data, they use a sequence of items and interactions by users from several sessions, to create recommendations that adapt to users' preferences in real-time. Thus, both the input and output are sequences of data.

Zhang et al. (2020) evaluates the use of sequential data for collaborative filtering models, thus choosing a RS approach. Their approach is capable of producing a sequence of outputs by iteratively rolling forward the relevant time window for each period. Therefore, it can be considered an adaptation of the CC approach of MLC.

Ferraro et al. (2020) deploy several algorithms for SBR, including DL models such as RNNs. Thus, their framework also considers sequences of data as input. Their algorithms produce multiple outputs and, as such, are akin to MLC algorithms.

Chou et al. (2022) follow a collaborative filtering framework, but do not apply traditional RS. Instead, they opt for a BR approach (MLC), using transactional LD to be fed into a graph-structured DL model. The results suggest that both static and longitudinal features are necessary to capture customer behavior patterns and to achieve the best performance.

Table 4.2 summarizes the aforementioned information. The table includes, respectively, the cited study, followed by columns indicating if a RS is evaluated (RS); if a MLC technique is included (MLC); if a DL model is proposed (DL); if their suggested method uses a sequential input (SI); and if interpretability techniques (IT) are applied.

Table 4.2. Literature review for studies using longitudinal data

Study	RS	MLC	DL	SI	IT
Rendle et al. (2010)	✓			✓	
Tan et al. (2016)		✓	✓	✓	
You et al. (2019)		✓	✓	✓	
Ferraro et al. (2020)		✓	✓	✓	
Zhang et al. (2020)	✓	✓		✓	
Chou et al. (2022)		✓	✓	✓	
This study	✓	✓	✓	✓	✓

Among the studies using LD for recommendations, DLs are frequently deployed, with RNNs being the most popular type. Furthermore, sequences of events, with no additional transformations, are typically used as input data. Several studies recognize limitations when using DL models, in terms of output transparency. However, the incorporation of interpretability techniques for recommendations remains rare.

Overall, the literature review reveals a research gap regarding the comparison between RS and MLC techniques, particularly when LD is available. Further, DLs are frequently employed with LD, with RNNs being the most popular approach. However, performance

comparisons against other approaches reveal a research opportunity. The use of LD as input, especially when used as a sequence of features, remains a scarcely explored area of research. The same is true for deploying interpretability techniques for recommendations.

4.3 Methodology

4.3.1 Data

The data has been provided by an European financial services provider. This data contains all available customer information between the start of May 2021 and the end of April 2022, with their respective product acquisition between the start of June 2022 and the end of April 2023. Therefore, our independent period ends on the last day of April 2022, while the following year is used as the dependent period. The data contains demographic information for each customer, such as age, sex, and commune of residence. These features are static, and reflect the last information available at the end of the independent period. Further, we include 13 different products from the company portfolio. For each of these products, the RFM features are calculated for each customer on an end-of-month basis. These features are sequential, as they consist of a chronologically ordered sequence of 12 end-of-month snapshots, with the last one coinciding with the end of the independent period.

Overall, 55 independent features are available, distributed in 16 static and 39 sequential features. Sequential features are used without featurization as input for DL models, while only the last snapshot is usable for the remaining techniques. In particular, the MLC techniques require features in the form of tabular data, while matrices summarizing the products purchased by the users are created for the RS models. The independent features are summarized on Table 4.3. These features are used to predict the acquisition of new products during the dependent period, modeled as 13 binary labels, or one per product.

Table 4.3. Definitions of available independent features

Feature	Definition	Type
Age	Customer's age	Static
Commune	Customer's commune of residence	Static
Sex_Ind	Indicates if customer is male	Static
Owner_Ind	Indicates if customer owns property	Static
Lib_Ind	Indicates if customer is independent	Static
Use_Email	Indicates if customer provided a usable email address	Static
Use_Mobile	Indicates if customer provided a usable mobile phone number	Static
Use_Phone	Indicates if customer provided a usable phone number	Static
Optin_Email	Indicates if customer allows emailing from the company	Static
Optin_Phone	Indicates if customer allows calls from the company	Static
Main_Ind	Indicates if company is customer's main financial service provider	Static
LOR_days	Length of relationship in number of days	Static
Employee_Ind	Indicates if customer is employed by the company	Static
HDG_Ind	Indicates if customer has assets over €50K	Static
Hb_Ind	Indicates if customer has access to home banking	Static
Cz_Ind	Indicates if customer has activated customer zone	Static
Recency (x13)	Number of days since last purchase	Sequential
Frequency (x13)	Number of products owned	Sequential
Monetary (x13)	Profit from product usage	Sequential
	Total available features:	55

RFM features are excellent at capturing historical customer behaviors, resulting in an improved predictive performance when used within the financial services domain and marketing applications (De Caigny et al., 2020; Mena et al., 2023). Moreover, RFM

features contain highly relevant domain information, contributing to understandability, justifiability, and actionability from model results (De Bock et al., 2023). Plus, these features are naturally time-evolving, while offering considerable freedom when designing an experimental setup (Mena et al., 2023). As such, we further study RFM features to explore whether these same characteristics translate to RS, especially when used sequentially.

The cutoff for the independent period and the dependent features are the same for all models and approaches. Nonetheless, RS models can be understood as relying on the frequency information derived from RFM, while MLC models are able to include all RFM features and additional static information. Lastly, DL models incorporate additional information by processing RFM features sequentially. The independent period data setup is schematically represented in Figure 4.1, outlining the difference between sequential features and static features, i.e a chronologically ordered sequence of monthly snapshots against the latest available snapshot, respectively.

Figure 4.1. Schematic representation of the features throughout the independent period

Finally, data preprocessing steps are performed to ensure data quality, following existing research (Prinzie and Van den Poel, 2006; Bogaert et al., 2019). Only customers active in the both the independent and dependent periods are included, meaning we select non-deceased individuals aged over 18, with at least a checking account open with the financial services provider. We also remove data from contracts that are delayed, blocked, canceled, or generally unavailable to be used by a customer. Additionally, products with a sparsity below 0.25% are discarded to avoid any issues during cross-validation, resulting in the 13 products in the data. Customers who own products with high sparsity levels or missing values are also excluded. Lastly, we remove customers with less than 5 distinct products. This results in 384,859 customers being valid, in line with previous research (Prinzie and Van den Poel, 2006; Bogaert et al., 2019).

Table 4.12 displays summary statistics on the distinct number of products owned, before and after dropping those customers with less than 5 distinct products. The table shows that with no restrictions, around 25% of customers would not receive relevant recommendations when using traditional RS. This is because their predictive stage generates recommendations for customers based on g randomly given items. Therefore, selecting customers with a distinct number of products larger than the number of items g allows the recommended products can be compared to the products that the customers possess (Bogaert et al., 2019). We use the largest value of g as a threshold, i.e. ownership of at least 5 products.

Thus, dropping these observations is a way to mitigate the cold-start problem, where the algorithms may have insufficient data to make accurate recommendations. These decisions can thus improve overall model performances, as high sparsity observations

often contain little information. High sparsity may not contribute significantly to model training and could also introduce noise. Furthermore, reducing the dataset size leads to faster training times and lower computational costs.

4.3.2 Featurization

Our data contains both static and sequential features. The sequential features are longitudinal, as they are available as a sequence of monthly snapshots on a customer level. These features can also be transformed to be used cross-sectionally, as is the case for the static data. LD is commonly available to financial services providers, making it important to assess whether its use improves the performance of personalized recommendations.

LD has the advantage of preserving dynamic information which may help identify ruptures in customer behavior (Sarkar and De Bruyn, 2021). However, MLC traditionally requires the use of tabular data, for which the transformation of LD is necessary. As such, we deploy different featurization techniques to transform the data in the most efficient ways for information preservation.

Featurization has been shown to positively impact performance (Geuens et al., 2018; Nilashi et al., 2018), but it is also used to reduce the size of large volumes of data, while preserving valuable information, better scalability, and model efficiency (Nilashi et al., 2018). The same has shown to be true when applying featurization onto LD (Rendle et al., 2010).

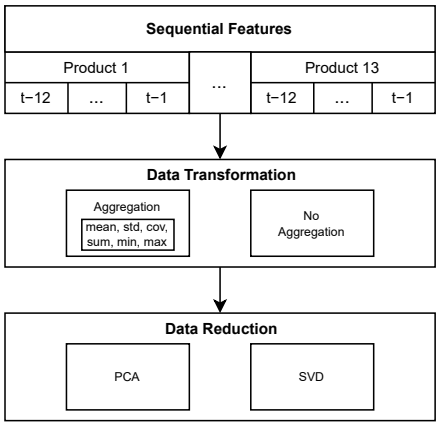


Figure 4.2. Schematic representation of featurization process

We include common techniques for recommendations that are apt for our data, namely PCA and SVD (Coussement et al., 2022; Geuens et al., 2018). We also construct a baseline, using the latest available data and no featurization. We assess whether featurization benefits from transforming the data beforehand through aggregation, using the mean, standard deviation, coefficient of variation (CoV), sum, minimum, and maximum values for all sequential features (Gattermann-Itschert and Thonemann, 2021), available for each product as a sequence of 12 monthly RFM snapshots. Thus, our featurization process consists of two steps, schematically illustrated on Figure 4.2. The first step is Data Transformation. For aggregation, the LD is transformed by calculating the mean, standard deviation, CoV, sum, minimum, and maximum across the 12 months per product for each user. No aggregation results in the RFM snapshots per month for each product being unchanged. The second step is Data Reduction, which decreases data dimensions,

resulting in less features with either PCA or SVD.

4.3.3 Multi-label Classification Techniques

The MLC techniques applied are divided into PT methods and AA methods, through the sklearn package on Python (Pedregosa et al., 2011). These techniques require the input to be in the form of tabular data.

We focus on BR and CC algorithms for the PT method. For BR, we require a binary classifier for each product, treating each label as a separate problem. For CC, each previously predicted label is included as an additional feature for the following prediction, modeling label dependencies through iteration. Therefore, when using CC, a given label can only be predicted after the labels that precede it an iteration have been output. In line with Bogaert et al. (2019), we include AB, FNN, NB, and RF, adding XGB because it excels in a variety of analytics problems (Gunnarsson et al., 2021). We also contribute by adding GRU to the comparison framework, as it is apt for incorporating LD as sequential input for recommendations (Sun et al., 2019) as extracting information from financial data (Shen et al., 2018), while being relatively fast to train (Chung et al., 2014). AA methods require modifying an algorithm for implementing a MLC problem, in this case implemented through multi-label random forest, multi-label FNN, and multi-label GRU.

Table 4.4 summarizes the models used and the respective hyperparameters evaluated for each approach. The information is grouped by method, i.e. BR and CC for PT, plus AA. The last column displays a check-mark if the input data is sequential.

Table 4.4. Hyperparameter tuning summary for MLC

Method	Algorithm	Hyperparameter	Values Evaluated	Sequential
Problem Transformation Binary Relevance	AB	iterations	100	
		max depth	15	
	FNN	learning rate	0.05, 0.01, 0.1	
	NB	var smoothing	$1e-9, 1e-7, 1e-5, 1e-3, 0$	
	RF	mtry	$\sqrt{features}$	
		N of trees	100	
		node size	5	
	XGB	tree method	exact, approx, hist	
Problem Transformation Classifier Chains	AB	iterations	100	
		max depth	15	
	FNN	learning rate	0.05, 0.01, 0.1	
	NB	var smoothing	$1e-9, 1e-7, 1e-5, 1e-3, 0$	
	RF	mtry	$\sqrt{features}$	
		N of trees	100	
		node size	5	
	XGB	tree method	exact, approx, hist	
Algorithm Adaptation	FNN	learning rate	0.05, 0.01, 0.1	
	GRU	Epochs	25, 50, 100, 150, 200	✓
		Batch size	200, 500	
		Layers	1, 3, 5	
		Learning rate	0.05, 0.01, 0.1	
		Dropout	0, 0.2	
		Bidirectional	True, False	
	RF	mtry	$\sqrt{features}$	
		N of trees	100	
		node size	5	
	XGB	tree method	exact, approx, hist	

As evidenced in this table, GRU is the only model which can handle sequential data. GRU are a type of RNN which have (i) successfully been used for recommendations (Sun et al., 2019), and (ii) proven useful when extracting information from a vast array of financial sequences of data (Shen et al., 2018). Further, GRU has been shown to perform

as well as other more complex models, with the advantage of being faster to train (Chung et al., 2014).

GRU is faster to train due to its simplified structure, which uses the same mechanism to update and discard information, rather than a separate one for each task (Chung et al., 2014). These mechanisms are referred to as a reset gate and an update gate, as illustrated in Figure 4.3. Both gate mechanisms use the inputs from the current timestep (x_t), as well as the previous hidden state (h_{t-1}), to update the weights used to learn a given task.

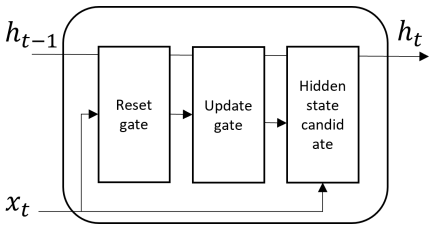


Figure 4.3. Schematic representation of a GRU layer

We use GRU layers to process sequential features, while static features are fed into a dense layer. The outputs from these layers are concatenated to produce the probability of acquiring each product per customer. The output follows the AA approach, running a single model to produce a vector of probabilities as an output. BR and CC approaches are excluded as computationally they are much more costly and time consuming. Figure 4.4 shows a schematic representation of how the GRU architecture has generally been constructed.

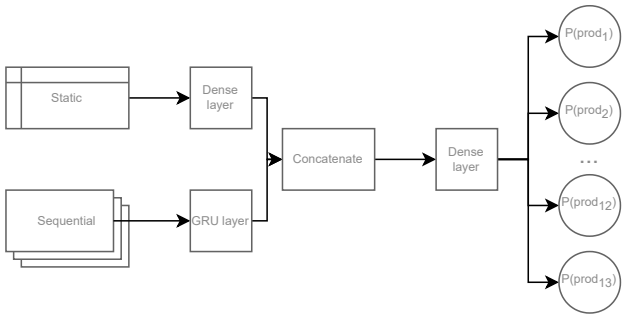


Figure 4.4. Schematic representation of GRU for recommendations

4.3.4 Traditional RS

The RS we deployed are item-based (IBCF) and user-based (UBCF) collaborative filtering, association rules (AR), popular recommender (PR) method, and the random recommender (RR). The first approach recommends items similar to a user's previously purchased items, the second recommends items based on other similar users, the third determines sets of items that are closely correlated in a transaction database (Aggarwal, 2016), the fourth simply recommends the most popular items, and the fifth recommends random items.

The collaborative filtering methods require the use of the Jaccard index, apt for binary product purchase data, to calculate similarities on which to base recommendations (Geuens et al., 2018). This measure seeks to match the positive events, or product purchases, between user I and user J, rather than a rated preference. Thus, the Jaccard

index is appropriate in a financial context. The measure is defined as:

$$sim_{Jaccard}(I, J) = \frac{|I \cap J|}{|I \cap J|}$$

Finally, Table 4.5 summarizes the hyperparameter evaluated for each RS technique, under a 5x2-fold cross-validation setup (Dietterich, 1998).

Table 4.5. Hyperparameter tuning summary for RS

Technique	Hyperparameter	Values Evaluated
AR	g	2, 3, 4, 5
	n	2, 3, 4, 5, 6, 7, 8
	s	0.01, 0.05, 0.1
	c	0.6, 0.7, 0.8, 0.9
IBCF	g	2, 3, 4, 5
	n	2, 3, 4, 5, 6, 7, 8
	k	10, 20, 30, 40, 50
PR	g	2, 3, 4, 5
	n	2, 3, 4, 5, 6, 7, 8
RR	g	2, 3, 4, 5
	n	2, 3, 4, 5, 6, 7, 8
UBCF	g	2, 3, 4, 5
	n	2, 3, 4, 5, 6, 7, 8
	nn	25, 50, 75, 100

4.3.5 SHAP for Interpretable Results

SHAP is an approach based on game theory, created for interpreting a machine learning model output. This approach is model agnostic, thus assigning an importance to each feature used regardless of the model. These importance values will be positive for features with a positive impact on a prediction, while those with negative values will have a negative impact (Lundberg and Lee, 2017). A SHAP value is calculated as follows:

$$\widehat{\phi_j} = \frac{1}{K} \sum_{k=1}^{K} ((\hat{g}\left(x_{+j}^{m}\right) - \hat{g}\left(x_{-j}^{m}\right))$$

where $\hat{g}(x_{+j}^{m})$ is the prediction for x, with a random number of feature values.

The benefits of this technique are its ease of interpretability and flexibility. We take advantage of these characteristics by aggregating the resulting SHAP values to implement a generalizable analysis by product type. The products are grouped, as per the financial services provider practices, into the following categories: credit, insurance, savings, and services. Then, the absolute SHAP values for each feature can be summed to produce its global importance within a product category. To account for the different category sizes, the global importance for each category is divided by the number of products present. The number of products per category is displayed on Table 4.6.

The features can either be static or sequential, the latter which can be further classified into recency, frequency, and monetary features for each product category. Further, sequential features can also be examined by timesteps, revealing temporal patterns.

Table 4.6. Product categories

Product category	Product names	Count
Credit	C02, C07, C08, C09	4
Insurance	A01, A02	2
Savings	E02, E04, E05, E06	4
Services	S01, S02, S03	3
	Total products:	13

4.4 Experimental Setup

We perform an exhaustive comparison of different approaches for producing personalized recommendations. These comparisons are structured, following the previously outlined research questions, for which we (i) create matrices of items and users to use as features for traditional RS approaches, (ii) contrast different featurization techniques, using several MLC approaches with different configurations, (iii) evaluate the incorporation of LD through DL state of the art models, and (iv) assess the interpretability of LD for recommendations, from the best performing DL model.

All models are tuned using a 5x2-fold cross-validation setup (Dietterich, 1998). This requires splitting the data set 5 different times into two non-overlapping halves, i.e. training and test data sets. This ensures all models are trained and tested over all observations.

Therefore, the reported evaluation metrics correspond to the mean out of the 10 values from the 5x2-fold cross-validation, with the objective of ensuring the same training and test samples are used across all models. Commonly used evaluation metrics, including F1 and G-mean, are deployed for 5x2-fold cross-validated evaluation in line with Bogaert et al. (2019).

Further, statistical significance tests are implemented to confirm the results, a non-parametric Friedman test, a Bonferroni-Dunn post hoc test, and Wilcoxon signed ranks test to compare different classifiers against the best performing algorithm (Demšar, 2006).

4.5 Results

4.5.1 Does the use of various RS methods lead to significantly distinct levels of performance?

The previously defined RS approaches are compared, using user and item matrices as input, by predicting new product acquisitions for the next year. The results are displayed on Table 4.7, with their respective standard deviation values and the best performing metrics in bold.

Table 4.7. 5x2 cross-validation mean results for traditional RS

Algorithm	F1	G-mean
AR	39.784 ± 0.015	41.553 ± 0.016
IBCF	40.824 ± 0.057	42.648 ± 0.061
PR	8.522 ± 0.054	8.795 ± 0.057
RR	5.564 ± 0.089	5.791 ± 0.096
UBCF	**41.116** ± 0.044	**42.970** ± 0.047

The results show that UBCF performs better on both F1 and G-mean metrics, suggesting a better ability to identify items that are both relevant and correctly identified as being likely purchases. IBCF and AR perform closely behind to UBCF, while PR and RR

perform vastly lower. This suggests that popular items are unlikely to be repurchased, in line with the financial services context.

To assess whether these differences are significant, a Friedman test is applied. The test indicates significant differences across approaches for both F1 ($\chi^2_F = 40.000, p < 0.000$) and G-mean ($\chi^2_F = 40.000, p < 0.000$). Thus, we proceed with a Bonferroni-Dunn post hoc test, to assess whether all algorithms exhibit significantly different performances. These results are available in Figure 4.5 for both F1 and G-mean, representing the average rank for each RS approach across all cross-validation folds. Each plot displays the RS approaches represented by a horizontal line, starting at their respective average ranking value, with a length equal to the critical difference value at a 95% confidence value, or 2.498 for both metrics (Bogaert et al., 2019). Further, a vertical line on each plot indicates the critical difference value added to the average ranking of the best performing model. Thus, the vertical line serves as a threshold, where an average rank beyond this line differ significantly from the best performing approach, UBCF. Further, horizontal lines that overlap indicate no statistically significant differences in performance. The figures show that the ranking of the algorithms is sustained across metrics. Moreover, UBCF performs significantly different from PR, and RR, but not from IBCF or AR for both metrics.

(a) F1 **(b)** G-mean

Figure 4.5. RS Bonferroni-Dunn post hoc tests, for a 95% confidence level

4.5.2 Does the utilization of various featurization methods for longitudinal data yield significantly different levels of performance?

We employ different featurization techniques to uncover the best performing MLC models under different configurations, including AA, BR, and CC methods. For each configuration, the highest scoring model and featurization combination is selected to be compared against baseline versions, which use the last available information with no additional transformation. We exclude DL models for this research question, as they can incorporate LD as a sequential input with no featurization required.

We first display the results by metric and featurization approach, where Table 4.8 contains the F1 outcomes and Table 4.9 those for G-mean. Both tables summarize the results by algorithm (Algo.), featurization approach (Feat.), and best MLC configuration, i.e. AA, BR, or CC. Within Feat., we compare the use of previously aggregated data against the use of data without aggregation. Therefore, with these results we evaluate whether data transformation improves the predictive performance of featurization. Lastly, the best performing results for each MLC are compared against a baseline on Table 4.10, using the data without aggregation or featurzation as input. On all tables, the results displayed are the mean from the 5x2 cross-validation approach, with the best results shown in bold. For simplicity, the results are discussed using $m_{(f,a)}$ as an abbreviation, where m refers to the model name, f to the type of featurization used, and a to the aggregation form. For instance, $\text{AB}_{(PCA,A)}$ refers to an AB model with PCA featurization

and aggregated data, while $AB_{(SVD,NA)}$ is an AB model with SVD featurization and no aggregation.

Table 4.8. 5x2 cross-validation mean F1 results for MLC algorithms with featurization

		F1 (No Aggregation)			F1 (Aggregation)		
Algo.	Feat.	AA	BR	CC	AA	BR	CC
AB	PCA	-	53.522 ±0.015	52.588 ±0.014	-	53.540 ±0.016	52.501 ±0.022
FNN	PCA	56.074 ±0.009	-	-	56.074 ±0.009	-	-
NB	PCA	-	52.152 ±0.015	53.078 ±0.014	-	52.115 ±0.014	53.021 ±0.012
RF	PCA	53.140 ±0.010	52.388 ±0.016	50.001 ±0.010	53.213 ±0.014	52.530 ±0.010	49.775 ±0.020
XGB	PCA	53.568 ±0.014	53.568 ±0.014	52.359 ±0.008	53.609 ±0.008	53.609 ±0.008	52.373 ±0.018
AB	SVD	-	53.657 ±0.014	52.507 ±0.015	-	**54.715** ±0.016	52.930 ±0.014
FNN	SVD	**56.075** ±0.009	-	-	56.074 ±0.009	-	-
NB	SVD	-	52.355 ±0.017	**53.080** ±0.016	-	52.196 ±0.018	52.939 ±0.013
RF	SVD	53.094 ±0.017	52.442 ±0.022	49.922 ±0.014	53.220 ±0.013	52.572 ±0.014	49.886 ±0.012
XGB	SVD	53.525 ±0.018	53.505 ±0.018	52.236 ±0.010	53.652 ±0.014	53.652 ±0.014	52.324 ±0.020

The F1 results show that the best performing model under AA configuration corresponds to $FNN_{(SVD,NA)}$. This result is an exception as in general, MLC models under AA configuration tend to perform better for both SVD and PCA when incorporating aggregation. Under BR configuration, the best result is also achieved with SVD, but using AB with aggregated data ($AB_{(SVD,A)}$). Further, NB performs better with no aggregation for both SVD and PCA, although all other MLC models increase their performance with aggregation. CC performs best for NB, again with SVD but with no aggregated data ($NB_{(SVD,NA)}$). Moreover, RF under CC performs better with no aggregation on both featurization approaches, but the scores are lower than the previously analyzed configurations, which indicates the possibility of overfitting. Similar conclusions can be drawn for AB and XGB with CC possibly resulting in overfitting. NB consistently performs better under CC and without aggregation. Lastly, $FNN_{(SVD,NA)}$ results in the highest overall F1 score.

Table 4.9. 5x2 cross-validation mean G-mean results for MLC algorithms

		G-mean (No Aggregation)			G-mean (Aggregation)		
Algo.	Feat.	AA	BR	CC	AA	BR	CC
AB	PCA	-	55.858 ±0.015	55.150 ±0.014	-	55.870 ±0.016	55.064 ±0.022
FNN	PCA	58.421 ±0.008	-	-	58.421 ±0.008	-	-
NB	PCA	-	54.489 ±0.014	**55.725** ±0.014	-	54.441 ±0.014	55.647 ±0.012
RF	PCA	55.503 ±0.010	54.738 ±0.016	52.569 ±0.011	55.579 ±0.014	54.875 ±0.010	52.336 ±0.020
XGB	PCA	55.913 ±0.014	55.913 ±0.014	54.994 ±0.009	55.950 ±0.007	55.950 ±0.007	55.007 ±0.019
AB	SVD	-	55.988 ±0.014	55.061 ±0.015	-	**57.027** ±0.017	55.537 ±0.014
FNN	SVD	**58.422** ±0.008	-	-	58.421 ±0.008	-	-
NB	SVD	-	54.691 ±0.017	55.720 ±0.016	-	54.521 ±0.018	55.558 ±0.013
RF	SVD	55.451 ±0.016	54.797 ±0.022	52.487 ±0.014	55.579 ±0.013	54.929 ±0.014	52.455 ±0.012
XGB	SVD	55.866 ±0.019	55.846 ±0.018	54.861 ±0.010	55.989 ±0.014	55.990 ±0.014	54.957 ±0.022

Table 4.9 reveals that G-mean follows the same patterns as seen with F1, showing that the results are stable across metrics. Therefore, once again the best result for AA configuration is achieved by $FNN_{(SVD,NA)}$. Similarly, the best result for BR is also stable, with $AB_{(SVD,A)}$ being the top performer. Lastly, the best CC result is still from NB with no data aggregation, but with PCA instead of SVD featurization ($NB_{(PCA,NA)}$). Overall, when available, AA generally performs at least as well as BR, while CC shows a propensity towards overfitting.

To further examine the MLC performances, we apply a Friedman non-parametric test. Significant differences are found among MLC models under AA configuration for F1 ($\chi_F^2 = 80.000, p < 0.000$) and G-mean ($\chi_F^2 = 80.000, p < 0.000$), under BR for F1 ($\chi_F^2 = 110.430, p < 0.000$) and G-mean ($\chi_F^2 = 109.920, p < 0.000$), and under CC for F1 ($\chi_F^2 = 120.000, p < 0.000$) and G-mean ($\chi_F^2 = 118.830, p < 0.000$). A Bonferroni-Dunn test with a 95% confidence level is implemented for each configuration per metric, with graphic representations available on 4.8.2. Under AA, $\text{FNN}_{(SVD,NA)}$ performs better than other AA algorithms by 3.30% on average, while the Bonferroni-Dunn test reveals statistically significant performance differences against RF and XGB for both metrics. For BR configuration, $\text{AB}_{(SVD,A)}$ outperforms other BR algorithms by 3.10% on average, while having statistically significant differences in performance for both metrics against RF and NB, but not XGB. For CC, $\text{NB}_{(PCA,NA)}$ and $\text{NB}_{(SVD,NA)}$ are tied in terms of ranking for both metrics. However, both have statistically significant results against AB, XGB and RF. Further, $\text{NB}_{(PCA,NA)}$ performs better than other CC approaches by 2.18% against 2.17% for $\text{NB}_{(SVD,NA)}$.

For further analysis, the best performers for each configuration are compared against their respective baseline versions. Specifically, $\text{FNN}_{(SVD,NA)}$ is used for AA configuration, $\text{AB}_{(SVD,A)}$ for BR configuration, and $\text{NB}_{(PCA,NA)}$ for CC, while each baseline version uses the unaltered data as input. These results are displayed on Table 4.10, with the best scores per metric shown in bold. The remaining baseline results are available on 4.14. The results show that all configurations perform better with featurization across metrics. On average, AA outperforms its baseline version by 1.99%, BR by 1.51%, and CC by 7.08%. Further, a Friedman test confirms statistically significant differences in performances for F1 ($\chi_F^2 = 50.000, p < 0.000$) and G-mean ($\chi_F^2 = 50.000, p < 0.000$), when comparing featurized algorithms against the baseline versions.

Table 4.10. 5x2 cross-validation mean results for best MLC algorithms and baseline versions

		F1		G-mean	
Config.	Featurization	Featurized	Baseline	Featurized	Baseline
AA	$\text{FNN}_{(SVD,NA)}$	**56.075** ±0.009	54.960 ±0.006	**58.422** ±0.008	57.307 ±0.005
BR	$\text{AB}_{(SVD,A)}$	54.715 ±0.016	53.891 ±0.012	57.027 ±0.017	56.192 ±0.012
CC	$\text{NB}_{(PCA,NA)}$	53.078 ±0.014	47.256 ±0.074	55.725 ±0.014	54.713 ±0.010

Additionally, Figure 4.6 shows the Bonferroni-Dunn test results for a 95% confidence level, with a critical difference value is of 2.782 . The best performing approach, FNN with featurization, is significantly different from all baseline approaches. Further, FNN performs better than AB with featurization by 2.47% and by 5.24% against NB with featurization, when averaging for both metrics.

(a) F1 **(b)** G-mean

Figure 4.6. MLC Bonferroni-Dunn post hoc tests, for a 95% confidence level

From these results we draw several conclusions. First, we observe that SVD generally performs better than PCA, particularly with no aggregation. Thus, further research can further optimize the incorporation of longitudinal data into MLC algorithms. Second, CC is a likely source of overfitting, as AA and BR consistently perform better, with the exception of NB. Third, AA configuration with FNN consistently outperforms other MLC algorithms. This is in line with previous research, where FNN achieves higher predictive performance, with reduced bias on the input data (Borchert et al., 2022).

4.5.3 Does the incorporation of longitudinal data using state-of-the-art DL models significantly outperform other recommendation approaches?

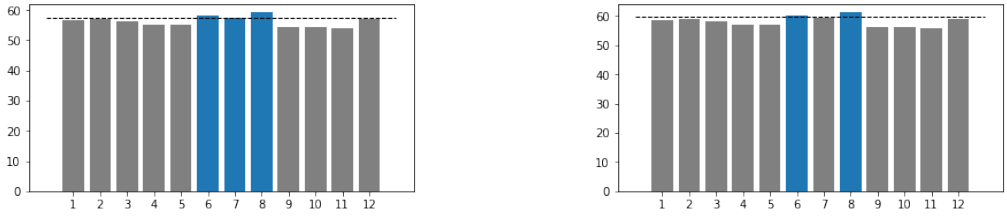
We compare a tuned DL model against the best performing algorithms from the previous results, as displayed on Table 4.11. The DL model, in the form of GRU, is the only one able to process sequential data. Further, we use an asterisk to indicate statistically significant results, using a Wilcoxon signed ranks test at a 95% confidence level.

Table 4.11. 5x2 cross-validation mean results for GRU and best performing MLC and RS algorithms

Approach	Algorithm	Featurization	F1	G-mean
DL (AA)	GRU	None	**57.218*** ±0.021	**59.707*** ±0.020
MLC (AA)	FNN	SVD	56.075* ±0.009	58.422* ±0.008
RS	UBCF	None	41.116* ±0.044	42.970* ±0.047

By performing parwise comparisons for all algorithms, we find statistically significant differences for both metrics between GRU and FNN, as well as between GRU and UBCF. Thus, UBCF is outperformed by both GRU and FNN with statistically significant differences in performance. This is expected, due to the complexity of financial services purchase patterns and the additional data processed by MLC and DL models. Additionally, GRU also outperforms FNN, thus showing GRU is able to harness sequences of data to achieve optimal performance. A possible reason for this is that GRU, unlike other recommender algorithms used, has been particularly developed for learning from sequences of data. On the other hand, the featurization algorithms used were developed for data reduction without the consideration of temporal patterns, which could explain a loss of additional predictive information. Furthermore, consumer behavior within the financial industry is relatively sparse, with the purchase of new products remaining relatively rare. As such, the incorporation of time-dependent data may pose a larger challenge when extracting relevant information for MLC algorithms than for GRU. Specifically, the ability of GRUs to maintain a hidden state that can capture long-term dependencies in the data allows them to carry information across timesteps (Chung et al., 2014), information which is not taken into consideration by MLCs.

Finally, we retrain the same GRU model to produce recommendations on a monthly basis, using the same training data as before. Figure 4.7 contains plots for F1 and G-mean, showing the respective 5x2 CV scores per month, where month 1 is the first timestep of the dependent period. A horizontal line indicates the best-performing values from Table 4.11, where bars with a superior performance are displayed in blue, with the objective of analyzing the performance changes in time. The plots show that monthly performance is often lower than the scores previously reported. However, between months 6 and 8, the performance improves. This suggests that recommendations produced in a sequential manner may have an optimal length, which we propose as a future avenue of research.



(a) F1 values per timestep

(b) G-mean values per timestep

Figure 4.7. Mean absolute SHAP values by product category

In sum, we find that GRU provides the best performing recommendations, for both F1 and G-mean, with statistically significant differences. Therefore, the DL model outperforms both MLC and RS approaches. Moreover, we explore the use of GRU to produce sequential recommendations, which points towards the possibility of an optimal length. This requires further research, as sequential recommendations have been reported to suffer from decreased diversity (Zhang et al., 2020).

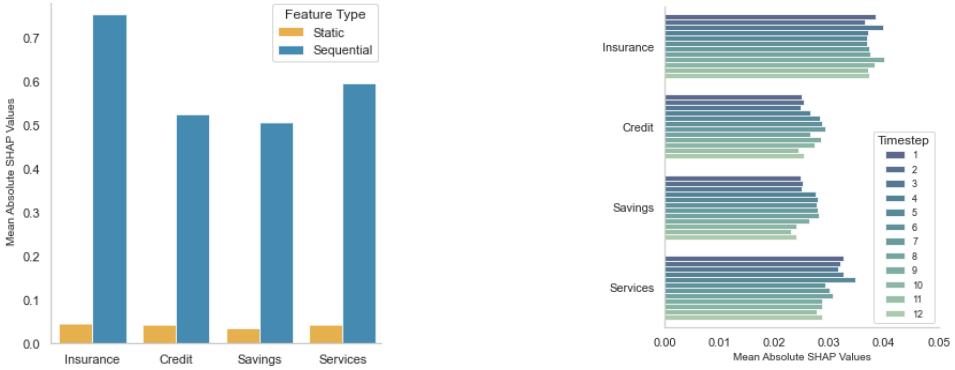
4.5.4 Can relevant features be identified for product recommendations in the financial services industry?

We examine the features using SHAP values and the top performing model. We group the products by type, as defined on Section 4.3.5, for generalizable conclusions. The original SHAP plots before aggregation are displayed on 4.8.2.

First, the absolute SHAP values for all features are added. For a fair comparison, the mean is calculated over the number of distinct products per category. Then, the SHAP values are grouped, as shown on Figure 4.8, into (i) static and sequential type features, to assess the impact from each feature type, and (ii) timestep, to uncover temporal patterns within the sequential features.

Figure 4.8a shows the same pattern emerges across all product categories, with sequential features far surpassing the overall impact of static features. Static features show similar magnitude in their impact across product categories, close to 0.04 out of a maximum of 1. For sequential features, the impact is higher for Insurance products, followed by Services, Credit, and Savings. Product categories with a higher sum of absolute values, such as Insurance, can be explained by more features impacting both the purchase and non-purchase of a product, resulting in a higher overall magnitude.

Figure 4.8b shows different temporal patterns for each product category. For example, Insurance products show a relatively stable impact across timesteps. Both Credit and Savings products evidence higher values around the middle timesteps. Finally, Services shows a trend of increasing impact for more recent timesteps. These patterns show customer behavior shifts in time, captured by RFM features, depending on the product suscribed. This suggests action is taken by customers before a new product purchase and may be worth considering when implementing marketing campaigns.



(a) SHAP values per feature type

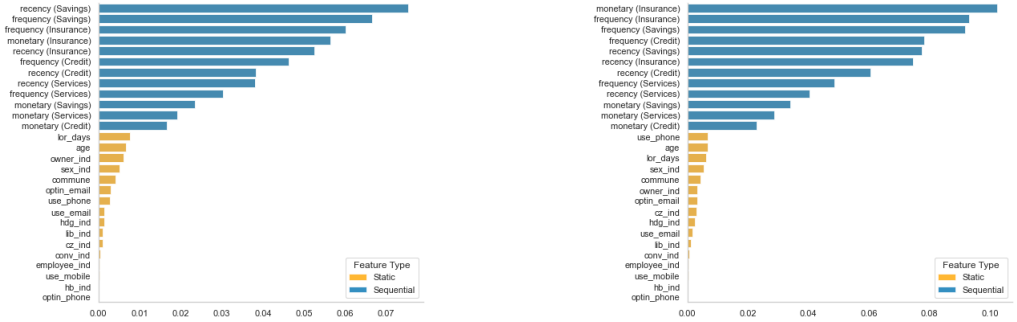
(b) SHAP values per Timestep

Figure 4.8. Mean absolute SHAP values by product category

To further identify the features producing an impact towards recommendations of a particular product category, we examine the top contributors per type, i.e. static and sequential. Since sequential features reflect the historical RFM values for the same products as the output, we apply the same grouping strategy. Thus, we analyze the relation between the different product categories in terms of magnitude. Nonetheless, additional details regarding the directional impact are available on 4.8.2.

Credit products, depicted in Figure 4.9a, are primarily influenced by the recency and frequency of Savings products, followed by the frequency, monetary, and recency features for Insurance products. Monetary features for Savings, Services, and Credit have the least impact as sequential features. Thus, the quantity of Savings products exhibits a stronger correlation with Credit products than any monetary values across different product categories. The length of the relationship and the age contribute the most among all static features, although they are all less impactful than sequential features. Consequently, ownership of Savings products may be crucial information when targeting customers for Credit solicitations.

Similarly, sequential features have a bigger impact than all static features for Insurance products, on Figure 4.9b. These are mostly impacted by the monetary value and frequency of other Insurance products, followed by the frequency of Savings products. The lowest impact from sequential features is exhibited by the monetary features from all other product categories. This suggests that customers already in possession of Insurance products, particularly those with may be noteworthy targets.



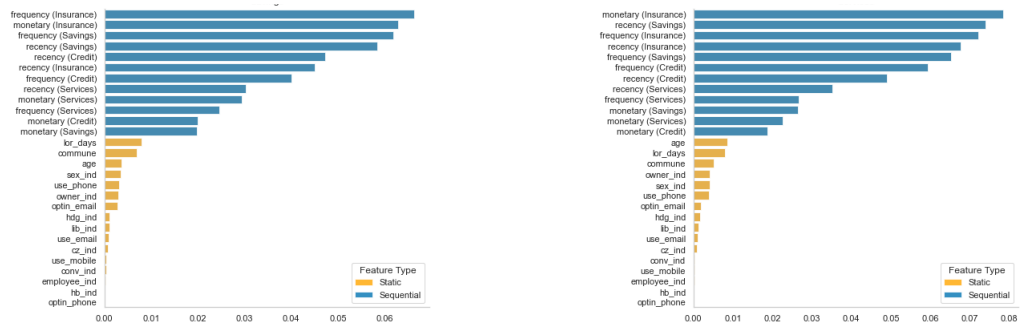
(a) Credit

(b) Insurance

Figure 4.9. Mean Absolute SHAP Values

Insurance products also are among the most effective for recommendations of both

Savings and Services, seen on 4.10. In particular, Insurance frequency and monetary features appear as the most impactful for Savings products. In the case of Services products, monetary features for Insurance products and recency for Savings products are at the top.



(a) Savings **(b)** Services

Figure 4.10. Mean Absolute SHAP Values

In sum, all sequential features have a higher impact than static features, across product categories. Furthermore, Insurance products features, particularly frequency and monetary data, have a high impact across product categories. Finally, behavioral patterns differ by product category. Overall, these findings signal that customers that invest in insurance are good candidates to target for cross-selling initiatives, while marketing campaigns could benefit from taking into account customer behavior longitudinally.

4.6 Conclusions and Future Research

We incorporate data from the financial industry to evaluate different RS. The results show that longitudinal data does improve performance when using a DL model. In addition, SHAP values reveal that sequential features far outweigh static features in terms of model output impact, across product categories. Moreover, different temporal patterns emerge per product category, confirming sequential features indeed contribute valuable information for recommendations. Finally, we find Insurance product ownership has a strong impact on other product category recommendations, which reveals interesting information for marketers and decision-makers, especially for resource allocation and customer targeting.

Therefore, we offer novel contributions, valuable for both the financial services industry and RS research. First, the comparison of several MLC techniques and RS, including state of the art DL models, in a real life scenario. Second, the comparison of different featurization techniques to assess if the incorporation of LD improves MLC performance. Third, the evaluation of LD as sequential input, through the use of DL models. Fourth, the use of interpretability techniques for managerial insights. Overall, these contributions uncover a better understanding on how to preprocess sequences of historical data for recommendations in the financial services industry.

We also describe limitations that may be of interest for future research. Firstly, other avenues for incorporating LD are yet to be explored, for instance through the use of additional featurization approaches or incorporating diverse DL models. Secondly, a broader array of predictor features could have resulted in better perfoming MLC models

and additional managerial insights. As examples, the incorporation of marketing actions by the financial services provider, as well as customer responses, could further enrich the model results and overall analysis. This data is particularly relevant to consider longitudinally, especially to improve the tailoring of campaigns to each user in time. Thirdly, the incorporation of textual data related to product descriptions could have allowed for a content-based RS and the evaluation of different methodologies for text embeddings. Examples of textual data include brief product descriptions, such as that found on a provider's website, or longer textual information, such as that from a contract. The performance of content-based RS is still uncertain within financial services, but it is worth researching to further uncover insights for cross-selling strategies. Fourthly, an empirical experiment, where the DL model results are used to serve the actual clientele of the financial services provider, would assess its ability to improve cross-selling efficacy, as well as prove its broader generalizability. Finally, previous research has reported that recommendations eventually result in a decrease in the diversity of items proposed (Zhang et al., 2020; Ferraro et al., 2020; Quadrana et al., 2017). Thus, a pending avenue of research would be to assess the impact of the aforementioned data sources on improving the diversity of recommendations, or of uncovering an optimal length of recommendations.

4.7 References

Aggarwal, C.C., 2016. Model-Based Collaborative Filtering. Springer International Publishing, Cham. pp. 71–138.

Bogaert, M., Lootens, J., Van den Poel, D., Ballings, M., 2019. Evaluating multi-label classifiers and recommender systems in the financial service sector. Eur. J. Oper. Res. 279, 620–634. doi:10.1016/j.ejor.2019.05.037.

Borchert, P., Coussement, K., De Caigny, A., De Weerdt, J., 2022. Extending business failure prediction models with textual website content using deep learning. Eur. J. Oper. Res. doi:10.1016/j.ejor.2022.06.060.

Boulenger, A., Liu, D., Farajalla, G.P., 2022. Sequential banking products recommendation and user profiling in one go, in: Proceedings of the Third ACM International Conference on AI in Finance, Association for Computing Machinery, New York, NY, USA. p. 317–324. doi:10.1145/3533271.3561697.

Chen, Y., Calabrese, R., Martin-Barragan, B., 2024. Interpretable machine learning for imbalanced credit scoring datasets. Eur. J. Oper. Res. 312, 357–372. doi:doi.org/10.1016/j.ejor.2023.06.036.

Chen, Z.Y., Fan, Z.P., Sun, M., 2016. A multi-kernel support tensor machine for classification with multitype multiway data and an application to cross-selling recommendations. Eur. J. Oper. Res. 255, 110–120. doi:doi.org/10.1016/j.ejor.2016.05.020.

Chou, Y.C., Chen, C.T., Huang, S.H., 2022. Modeling behavior sequence for personalized fund recommendation with graphical deep collaborative filtering. Expert Syst. Appl. 192, 116311. doi:doi.org/10.1016/j.eswa.2021.116311.

Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555. arXiv:1412.3555.

Coussement, K., De Bock, K., Geuens, S., 2022. A decision-analytic framework for interpretable recommendation systems with multiple input data sources: a case study for a european e-tailer. Ann. Oper. Res. 315, 671—694. doi:10.1007/s10479-021-03979-4.

De Bock, K.W., Coussement, K., Caigny, A.D., Słowiński, R., Baesens, B., Boute, R.N., Choi, T.M., Delen, D., Kraus, M., Lessmann, S., Maldonado, S., Martens, D., Óskarsdóttir, M., Vairetti, C., Verbeke, W., Weber, R., 2023. Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. Eur. J. Oper. Res. doi:doi.org/10.1016/j.ejor.2023.09.026.

De Caigny, A., Coussement, K., De Bock, K.W., 2020. Leveraging fine-grained transaction data for customer life event predictions. Decis. Support Syst. 130, 113232. doi:doi.org/10.1016/j.dss.2019.113232.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30. URL: http://jmlr.org/papers/v7/demsar06a.html.

Dietterich, T.G., 1998. Approximate statistical test for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923. doi:10.1162/089976698300017197.

Ferraro, A., Jannach, D., Serra, X., 2020. Exploring longitudinal effects of session-based recommendations. CoRR abs/2008.07226.

Gattermann-Itschert, T., Thonemann, U.W., 2021. How training on multiple time slices improves performance in churn prediction. Eur. J. Oper. Res. 295, 664–674. doi:10.1016/j.ejor.2021.05.035.

Geuens, S., Coussement, K., De Bock, K.W., 2018. A framework for configuring collaborative filtering-based recommendations derived from purchase data. Eur. J. Oper. Res. 265, 208–218. doi:10.1016/j.ejor.2017.07.005.

Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., Lemahieu, W., 2021. Deep learning for credit scoring: Do or don't? Eur. J. Oper. Res. 295, 292–305. doi:10.1016/j.ejor.2021.03.006.

Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D., 2015. Session-based recommendations with recurrent neural networks. doi:10.48550/ARXIV.1511.06939.

Kamakura, W.A., Ramaswami, S.N., Srivastava, R.K., 1991. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. Int. J. Res. Mark. 8, 329–349. doi:10.1016/0167-8116(91)90030-B.

Kamakura, W.A., Wedel, M., de Rosa, F., Mazzon, J.A., 2003. Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction. Int. J. Res. Mark. 20, 45–65. doi:10.1016/S0167-8116(02)00121-0.

Khanal, S.S., Prasad, P., Alsadoon, A., Maag, A., 2020. A systematic review: machine learning based recommendation systems for e-learning. Educ. Inf. Technol. 25, 2635–2664.

Knott, A., Hayes, A., Neslin, S.A., 2002. Next-product-to-buy models for cross-selling applications. J. Interact. Mark. 16, 59–75.

Kumar, V., George, M., Pancras, J., 2008. Cross-buying in retailing: Drivers and consequences. J. Retail. 84, 15–27.

Li, S., Sun, B., Montgomery, A.L., 2011. Cross-selling the right product to the right customer at the right time. J. Mark. Res. 48, 683–700. URL: http://www.jstor.org/stable/23033447.

Li, S., Sun, B., Wilcox, R.T., 2005. Cross-selling sequentially ordered products: An application to consumer banking services. J. Mark. Res. 42, 233–239. doi:10.1509/jmkr.42.2.233.62288.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mena, G., Coussement, K., De Bock, K., De Caigny, A., Lessmann, S., 2023. Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. Ann. Oper. Res. 53, 80–95. doi:10.1007/s10479-023-05259-9.

Musto, C., Semeraro, G., Lops, P., de Gemmis, M., Lekkas, G., 2015. Personalized finance advisory through case-based recommender systems and diversification strategies. Decis. Support Syst. 77, 100–111. doi:10.1016/j.dss.2015.06.001.

Nilashi, M., Ibrahim, O., Bagherifard, K., 2018. A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. Expert Syst. Appl. 92, 507–520.

Notz, P.M., Pibernik, R., 2024. Explainable subgradient tree boosting for prescriptive analytics in operations management. Eur. J. Oper. Res. 312, 1119–1133. doi:doi.org/10.1016/j.ejor.2023.08.037.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Prinzie, A., Van den Poel, D., 2008. Random forests for multiclass classification: Random multinomial logit. Expert Syst. Appl. 34, 1721–1732.

Prinzie, A., Van den Poel, D., 2006. Investigating purchasing-sequence patterns for financial services using markov, mtd and mtdg models. Eur. J. Oper. Res. 170, 710–734. doi:10.1016/j.ejor.2004.05.004.

Quadrana, M., Cremonesi, P., Jannach, D., 2018. Sequence-aware recommender systems. CoRR abs/1802.08452. URL: http://arxiv.org/abs/1802.08452, arXiv:1802.08452.

Quadrana, M., Karatzoglou, A., Hidasi, B., Cremonesi, P., 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA. p. 130–137. doi:10.1145/3109859.3109896.

Rendle, S., Freudenthaler, C., Schmidt-Thieme, L., 2010. Factorizing personalized markov chains for next-basket recommendation, in: Proceedings of the 19th International Conference on World Wide Web, Association for Computing Machinery, New York, NY, USA. p. 811–820. doi:10.1145/1772690.1772773.

Sarkar, M., De Bruyn, A., 2021. LSTM response models for direct marketing analytics: Replacing feature engineering with deep learning. J. Interact. Mark. 53, 80–95. doi:10.1016/j.intmar.2020.07.002.

Shen, G., Tan, Q., Zhang, H., Zeng, P., Xu, J., 2018. Deep learning with gated recurrent unit networks for financial sequence predictions. Procedia Comput. Sci. 131, 895–903. doi:doi.org/10.1016/j.procs.2018.04.298. recent Advancement in Information and Communication Technology:.

Sigrist, F., Leuenberger, N., 2023. Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities. Eur. J. Oper. Res. 305, 1390–1406. doi:doi.org/10.1016/j.ejor.2022.06.035.

Stevens, A., De Smedt, J., 2023. Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. Eur. J. Oper. Res. doi:doi.org/10.1016/j.ejor.2023.09.010.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., Jiang, P., 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. p. 1441–1450. doi:10.1145/3357384.3357895.

Tan, Y.K., Xu, X., Liu, Y., 2016. Improved recurrent neural networks for session-based recommendations, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Association for Computing Machinery, New York, NY, USA. p. 17–22. doi:10.1145/2988450.2988452.

You, J., Wang, Y., Pal, A., Eksombatchai, P., Rosenburg, C., Leskovec, J., 2019. Hierarchical temporal convolutional networks for dynamic recommender systems, in: The World Wide Web Conference, Association for Computing Machinery, New York, NY, USA. p. 2236–2246. doi:10.1145/3308558.3313747.

Zhang, J., Adomavicius, G., Gupta, A., Ketter, W., 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. Info. Sys. Research 31, 76–101. doi:10.1287/isre.2019.0876.

4.8 Appendix

4.8.1 Appendix A. Additional Data

Appendix A.1. Summary Statistics Table 4.12 contains summary statistics on the distinct number of products owned. The first row, no restrictions, shows product ownership information for all customers. The second row summarizes ownership for clients with 5 or more distinct products. The table shows that with no restrictions, close to 25% of customers would not receive relevant recommendations when using traditional RS.

Table 4.12. Summary statistics for distinct number of products owned per customer

Distinct products	Min	Q1	Median	Mean	Q3	Max
No restriction	1.00	5.00	7.00	6.89	9.00	18.00
≥ 5	5.00	6.00	7.00	7.79	9.00	18.00

4.8.2 Appendix B. Additional Results

Table 4.13. Summary for dimensionality reduction

Distinct products	Min	Q1	Median	Mean	Q3	Max
No restriction	1.00	5.00	7.00	6.89	9.00	18.00
≥ 5	5.00	6.00	7.00	7.79	9.00	18.00

Appendix B.1. PCA and SVD dimensionality

Appendix B.2. MLC Bonferroni-Dunn tests This section displays graphic representations for the Bonferroni-Dunn test with 95% confidence level for MLC algorithms for both metric on a configuration level. As before, a vertical line on each plot indicates the critical difference value added to the average ranking of the best performing model. For each configuration, the comparison includes algorithms using different featurization methods.

Figure 4.11 shows that for each model, no significant differences in performance are detected among the featurization approaches for both metrics. However, XGB and RF, regardless of the featurization approach, are performing significantly different from the best performing FNN, which is SVD with no aggregation for F1 ($\text{FNN}_{(SVD,NA)}$), and FNN with PCA and aggregation under G-mean ($\text{FNN}_{(PCA,A)}$). However, $\text{FNN}_{(SVD,NA)}$ better than all other AA approaches by 3.30% on average, compared to 3.29% for $\text{FNN}_{(PCA,A)}$. As such, we continue further analysis using $\text{FNN}_{(SVD,NA)}$.

(a) F1 **(b)** G-mean

Figure 4.11. Bonferroni-Dunn test for AA configuration

Figure 4.12 reveals that in general there are no significant differences in BR performance between AB and XGB. Nonetheless, RF and NB exhibit statistically significant differences in performances from the best approach, which is AB with SVD and aggregation for both metrics. This model outperforms other BR approaches by 3.10% on average and we continue its use for additional insights.

(a) F1 **(b)** G-mean

Figure 4.12. Bonferroni-Dunn test for BR configuration

Figure 4.13 reveals that in general there are no significant differences in CC performance between NB with different featurization approaches. Additionally, NB with SVD is tied in terms of top ranking with NB with PCA. Nonetheless, AB, XGB, and RF exhibit statistically significant differences in performances from both of these models for both metrics. However, when calculating average difference in results, NB with PCA performs higher by 2.18%, compared to 2.17% for NB with SVD. Thus, further analysis is performed using NB with PCA. Finally, the overall ranking for NB under CC configuration improves vastly when compared to NB under BR, which suggests that the dependence of labels modeled in CC positively contributes to NB performance, while negatively affecting AB, RF, and XGB.

(a) F1 **(b)** G-mean

Figure 4.13. Bonferroni-Dunn test for CC configuration

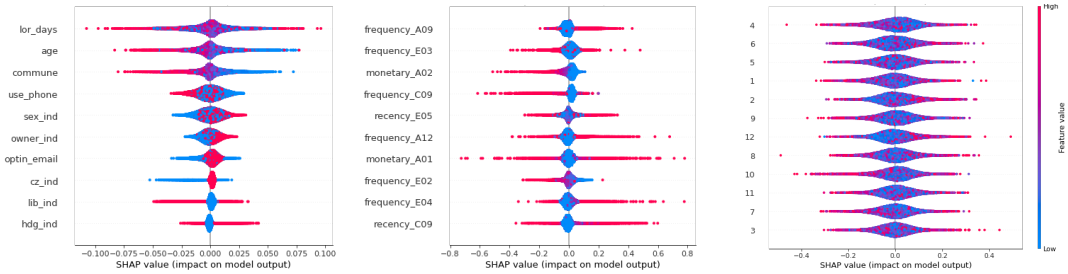
Overall, the top algorithms under AA and CC configurations exhibit statistically significant performance differences from other algorithms, with the exception of BR. Further, results are consistent across metrics, with some minor differences in the ranking order. However, no clear patterns within featurization emerge.

Appendix B.3. MLC Baseline Results This section contains baseline results for all MLC, displayed on Table 4.14 algorithms. These are obtained from using the data without aggregation or featurization, using each monthly snapshot as an additional feature. As before, the results stem from the 5x2 CV mean values, with the best performance per metric and configuration shown in bold text.

Table 4.14. 5x2 cross-validation mean results for MLC algorithms without featurization

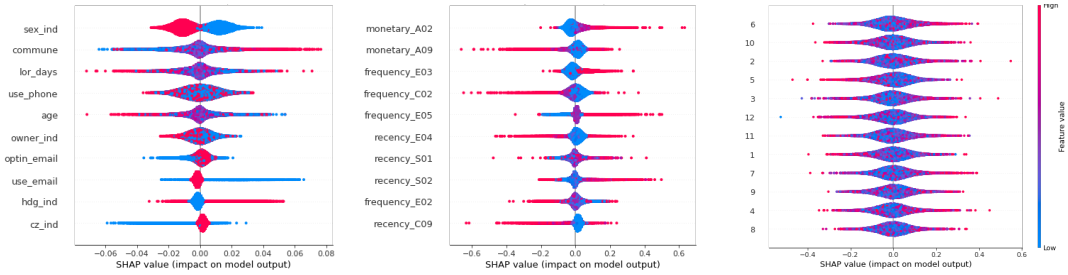
	F1			G-mean		
Algo.	AA	BR	CC	AA	BR	CC
AB	-	**53.891** ±0.012	**52.085** ±0.010	-	**56.192** ±0.012	**54.713** ±0.010
FNN	**54.960** ±0.006	-	-	**57.307** ±0.005	-	-
NB	-	47.254 ±0.074	47.256 ±0.074	-	49.573 ±0.075	49.575 ±0.075
RF	52.046 ±0.013	51.597 ±0.011	49.260 ±0.008	54.411 ±0.013	53.928 ±0.011	51.827 ±0.007
XGB	52.133 ±0.009	52.133 ±0.009	50.887 ±0.012	54.484 ±0.009	54.484 ±0.009	53.507 ±0.012

Appendix B.4. SHAP Values per Product This section contains summary plots for SHAP values per product, for the best performing GRU model. First, we present one figure for static features and one figure for sequential features for each product. Each figure contains the top 10 features with most impact on the model output. Second, we include one figure showing the total SHAP values per product, for each of the 12 timesteps, where 1 is the most recent timestep. This allows a visual understanding of the changes in time, from the sequential features impacting the model output. For all plots the values are displayed in descending order, showing the most impactful features at the top. These values can be aggregated to construct the figures presented in Section 4.5.4.



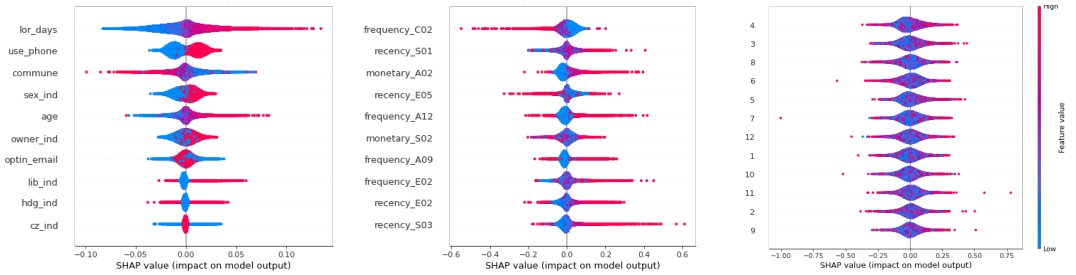
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.14. SHAP values for insurance product A01



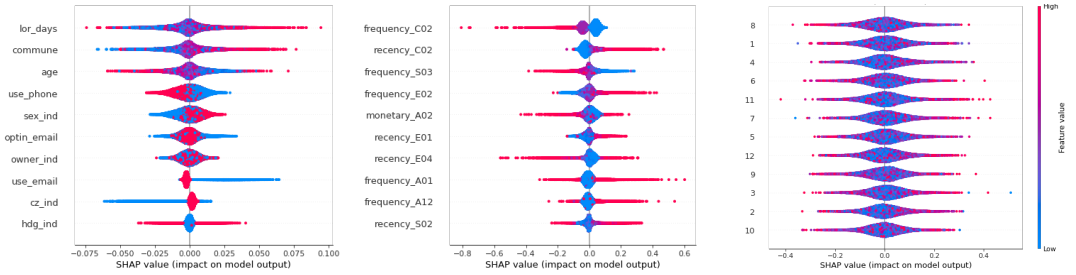
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.15. SHAP values for insurance product A02



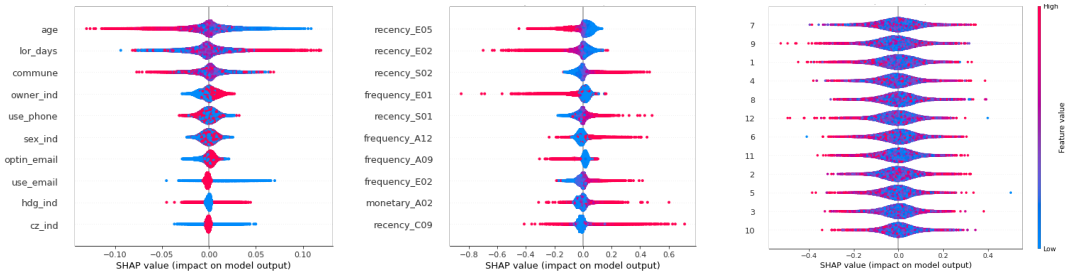
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.16. SHAP values for credit product C02



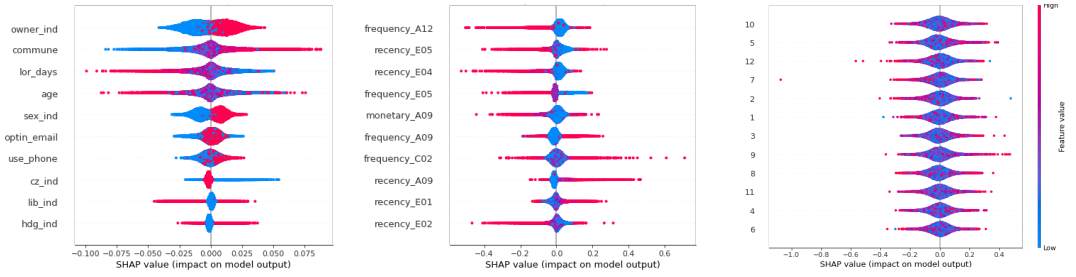
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.17. SHAP values for credit product C07



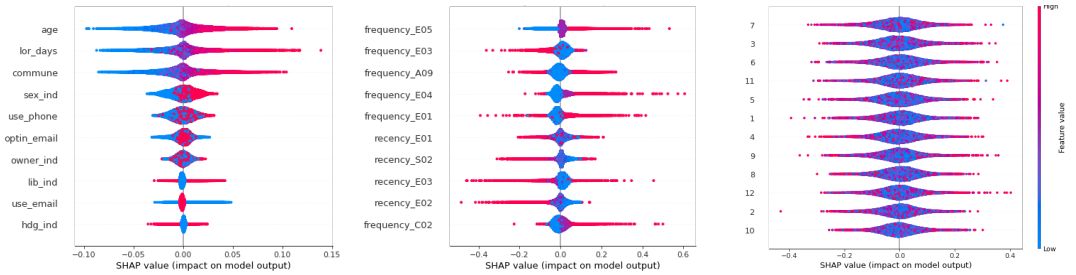
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.18. SHAP values for credit product C08



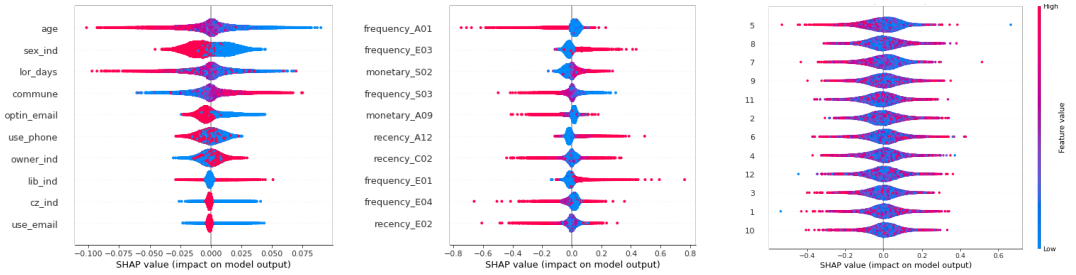
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.19. SHAP values for credit product C09



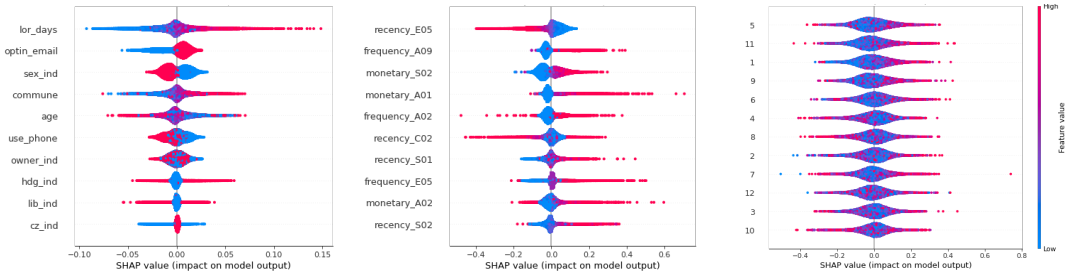
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.20. SHAP values for savings product E02



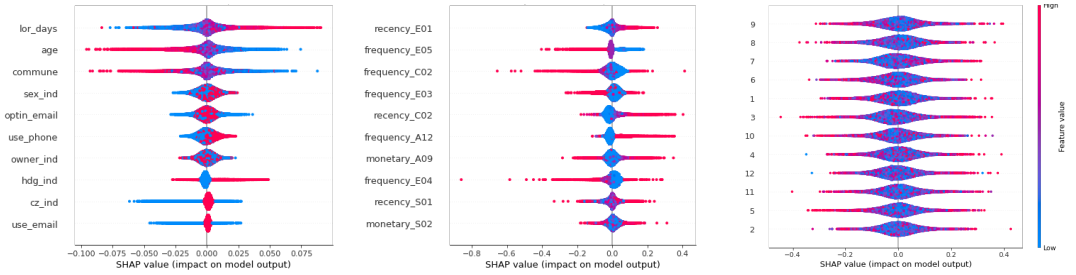
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.21. SHAP values for savings product E04



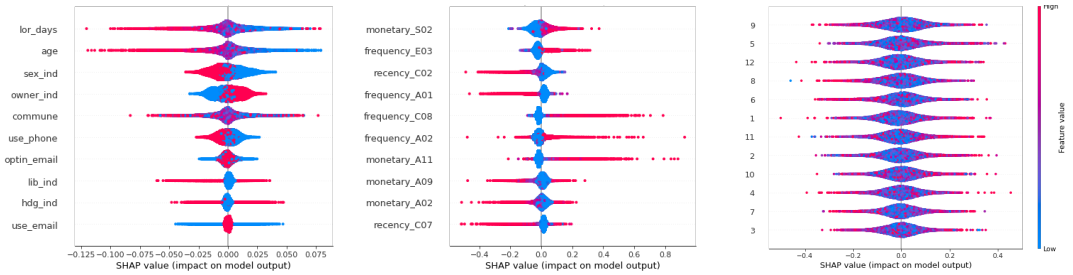
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.22. SHAP values for savings product E05



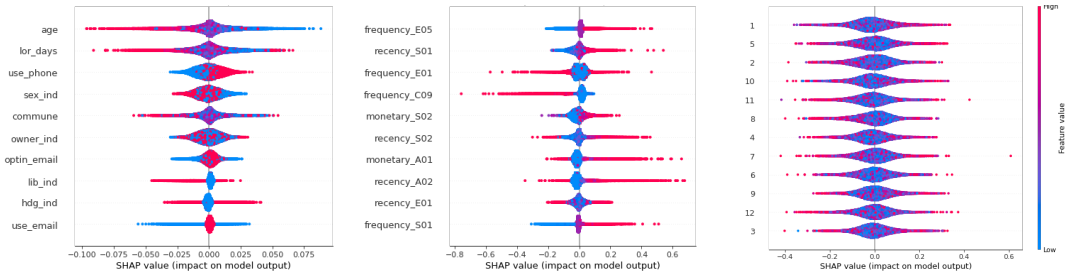
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.23. SHAP values for savings product E06



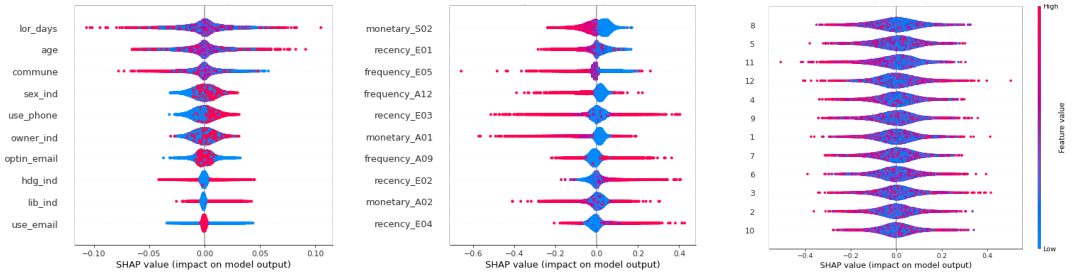
(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.24. SHAP values for services product S01



(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.25. SHAP values for services product S02



(a) Static features **(b)** Sequential features **(c)** Timesteps

Figure 4.26. SHAP values for services product S03

Chapter 5: Conclusions

CHAPTER 5

Conclusions

5.1 Conclusions

5.1.1 Conclusions générales

L'objectif principal de recherche de cette thèse est de réaliser des innovations axées sur les données dans le secteur des services financiers. Trois principales voies d'innovation ont été exposées dans le chapitre 1 : (i) l'incorporation de nouvelles sources de données, (ii) l'application de méthodologies de pointe, et (iii) l'application de techniques d'explicabilité. La recherche se compose de trois études, trouvées des chapitres 2 à 4, qui contribuent à ces voies.

Innovation à travers des sources de données novatrices. Les trois études utilisent des sources de données novatrices et emploient des techniques d'explicabilité pour mieux comprendre leur impact. Globalement, chaque source de données est sélectionnée pour fournir des perspectives différentes de l'industrie financière, comme représenté dans la Figure 1, déployant ainsi des applications sous différents angles. La nouveauté réside également dans l'utilisation de données comme entrée séquentielle avec des modèles robustes pour différentes tâches prédictives. La première étude prédit les prix du pétrole en incorporant des données textuelles provenant des tweets de Donald Trump pendant sa présidence. La deuxième étude utilise des informations démographiques et comportementales des clients, sous forme de données transactionnelles, pour prédire 10 événements de vie distincts, dont certains n'ont pas été précédemment étudiés. La troisième étude introduit une nouveauté dans l'industrie des services financiers en utilisant des informations sur les produits longitudinales sous forme de caractéristiques RFM pour construire un système de recommandation.

Innovation à travers la méthodologie. Cette thèse explore trois tâches prédictives, en employant des modèles d'apprentissage profond (DL) sous forme de RNN pour leur efficacité dans le traitement de données longitudinales en tant qu'entrée séquentielle. Les chapitres 2 et 3 utilisent LSTM, tandis que le chapitre 4 emploie GRU. Bien que la sélection de modèle suive la littérature existante, l'innovation réside dans les architectures globales, capables de traiter plusieurs sources de données et qui n'ont pas été précédemment appliquées à leurs domaines respectifs. Dans le chapitre 2, LSTM traite à la fois des données textuelles et numériques, le chapitre 3 applique LSTM aux données statiques et séquentielles, et le chapitre 4 utilise GRU, également pour des entrées statiques et séquentielles.

Innovation à travers l'explicabilité. L'utilisation de nouvelles sources de données de manière séquentielle en entrée pour les modèles DL pose un défi pour obtenir des résultats explicables et transmettre des informations exploitables à diverses parties prenantes. Chaque chapitre emploie des techniques distinctes alignées sur son objectif spécifique. Par exemple, le chapitre 2 utilise une analyse des ruptures structurelles pour repérer l'émergence de mots-clés lors de moments critiques dans les variations des prix du pétrole, indépendamment de la direction. Dans le chapitre 3, les gradients intégrés mettent en lumière les changements dans le comportement des clients précédant les événements de vie. Le chapitre 4 utilise les valeurs SHAP pour mettre en évidence les changements dans le comportement des clients au fil du temps, agrégeant par catégorie de produit pour des insights plus généralisables.

En résumé, la recherche incluse dans cette thèse est très pertinente d'un point de vue décisionnel et managérial. Les implications managériales spécifiques pour chaque étude ont déjà été discutées dans leurs chapitres respectifs. Cependant, des conclusions générales supplémentaires peuvent être tirées. Premièrement, toutes les recherches ont été réalisées sur des données réelles, ce qui signifie que les

méthodologies peuvent être transférées à d'autres applications commerciales. Deuxièmement, les défis présents dans l'industrie des services financiers, tels que l'allocation des ressources et la gestion efficace de la relation client, peuvent être abordés grâce aux méthodologies déployées dans ces études. Ainsi, les décideurs et les gestionnaires devraient envisager d'ajouter la modélisation DL, les données séquentielles et les techniques d'explicabilité comme outils supplémentaires pour améliorer continuellement le ciblage de leur clientèle. Troisièmement, les chapitres 3 et 4 utilisent des données provenant de la même source pour la prédiction des événements de vie et les systèmes de recommandation, respectivement. En tant que tel, les gestionnaires peuvent combiner les deux approches pour améliorer encore le ciblage des clients, l'allocation des ressources et les services personnalisés, surtout au bon moment, afin de renforcer leurs stratégies de CRM globales.

5.1.2 Limitations et perspectives de recherche future

Tout au long de cette thèse, nous avons clairement énoncé les contributions de notre travail à l'innovation axée sur les données dans l'industrie des services financiers. Chaque chapitre contient déjà une section brève sur les limitations pour une étude particulière. Cependant, pour des raisons de complétude, cette section discute des lacunes et des opportunités pour des recherches supplémentaires à travers les trois études différentes.

Le chapitre 2 utilise des données textuelles des médias sociaux et des prix du pétrole retardés dans LSTM pour prédire les prix futurs. Des insights ont été tirés à l'aide d'une analyse des ruptures structurelles, mais une meilleure compréhension peut être obtenue par différents moyens. Premièrement, différents modèles DL peuvent être évalués, ainsi que l'utilisation de méthodes d'attribution pour des insights plus détaillés. Deuxièmement, l'amélioration de l'architecture générale avec une couche d'attention peut fournir une visualisation plus claire de l'importance des caractéristiques. Troisièmement, la transformation de la cible de prédiction des prix du pétrole en une variable binaire, comme la direction des prix, pourrait clarifier l'interaction entre les données textuelles et les prix du pétrole. Enfin, l'utilisation d'un horizon de données historiques plus long, englobant les tweets de tous les présidents américains précédents, peut offrir des insights généralisables sur l'impact du leadership. Ces modifications peuvent révéler des insights complémentaires sur la manière dont LSTM exploite le texte pour améliorer les performances de prédiction des prix du pétrole, si les différences de leadership sont capturées à travers le texte et comment elles affectent les changements de prix du pétrole.

Le chapitre 3 contribue à la littérature sur la prédiction des événements de vie, mais révèle également diverses voies pour des recherches supplémentaires. Premièrement, des études supplémentaires sont nécessaires pour évaluer les impacts de différentes méthodes d'attribution, aidant les décideurs à comprendre les explications du modèle pour la prédiction des événements de vie. Deuxièmement, le suivi de divers comportements des clients pendant les événements de vie peut améliorer les stratégies personnalisées des spécialistes du marketing, favorisant un engagement client plus fort et des réponses (Han and Anderson, 2022). Par exemple, mener des tests A/B sur des recommandations très personnalisées pour les clients approchant des événements de vie peut fournir des insights dynamiques sur leurs comportements (De Caigny et al., 2020; Cheng and Chen, 2022). Troisièmement, l'exploration de périodes de données client historiques plus longues peut révéler des insights sur les moments clés pour déployer des actions marketing bien à l'avance.

Le chapitre 4 contribue à la recherche sur les systèmes de recommandation dans l'industrie des services financiers, mais comporte également des limitations pour de futures considérations. Premièrement, un éventail plus large de caractéristiques prédictrices aurait pu conduire au développement de modèles MLC plus performants. Par exemple, des données socio-démographiques plus complexes (par exemple, revenu, taille du ménage et niveau d'éducation), des détails de consommation plus complets (par exemple, allocation des dépenses, canaux de distribution des ventes) et des indicateurs de satisfaction client (par

exemple, données textuelles des plaintes des clients) pourraient enrichir davantage les résultats du modèle. Deuxièmement, les spécifications détaillées des produits, telles que les descriptions de produits sur le site Web, auraient pu faciliter le déploiement d'un système de recommandation basé sur le contenu. Malgré la performance incertaine d'un système de recommandation basé sur le contenu dans le cadre des services financiers, son exploration présente un mérite substantiel, notamment dans le domaine de la vente croisée. Il peut également être intégré, en conjonction avec le filtrage collaboratif, dans le cadre d'une approche hybride. De plus, l'utilisation d'approches hybrides permet également d'explorer le problème du démarrage à froid (Fernandes et al., 2023).

Pour les chapitres 3 et 4, une voie de recherche à venir consiste en une expérience empirique où le modèle DL est appliqué à la clientèle réelle du fournisseur de services financiers. Cette approche vise à corroborer si les modèles envisagés améliorent l'efficacité de la vente croisée. De telles expériences sont très complexes, car elles nécessitent des initiatives marketing judicieuses et une durée de test adéquate. Cependant, des résultats réussis présenteraient un avantage clé, car on peut en déduire que nos résultats ont une plus grande généralisabilité, et peuvent également être pertinents pour développer des métriques de profit importantes.

De même, des horizons de données longitudinales plus longs et des informations sur les offres des concurrents exposeraient le modèle aux changements des conditions du marché, des cycles économiques et des taux d'intérêt qui ont également une plus grande généralisabilité. Les informations internes, telles que les sollicitations ou les campagnes marketing, pourraient impacter l'utilisation du compte, le solde, la rétention, et sont donc des informations intéressantes à ajouter à la recherche sur les services financiers.

Dans l'ensemble, l'inclusion de données supplémentaires pour des recherches futures conduit également à une autre étude en suspens : la scalabilité de ces approches. Par conséquent, une dernière voie de recherche que nous suggérons est la construction d'un cadre clair, en fonction des contraintes de ressources, de la disponibilité des données et de la taille de la clientèle.

5.2 Conclusions

5.2.1 General conclusions

The main research objective of this thesis is to achieve data-driven innovations in the financial services sector. Three main pathways to innovation were outlined in Chapter 1: (i) the incorporation of novel data sources, (ii) the application of state-of-the-art methodologies, and (iii) the application of explainability techniques. The research consists of three studies, found from Chapters 2 to 4, that contribute to these pathways.

Innovation through novel data sources. All three studies use novel data sources and employ explainability techniques to better comprehend their impact. Overall, each data source is selected to provide different data-driven insights from the financial industry, as represented on Figure 1.5, thus deploying applications from different perspectives. The novelty also lies in utilizing data as a sequential input with robust models for diverse predictive tasks. The first study predicts oil prices by incorporating textual data from Donald Trump's tweets during his presidency. The second study utilizes customer demographic and behavioral information, in the form of transactional data, to predict 10 distinct life events, some of which have not been previously researched. The third study introduces novelty in the financial services industry by using longitudinal product information in the form of RFM features to construct a recommendation system.

Innovation through methodology. This thesis explores three predictive tasks, employing deep learning (DL) models in the form of RNNs for their proficiency in handling longitudinal data as sequential input. Chapters 2 and 3 utilize LSTM, while Chapter 4 employs GRU. Although model selection follows existing literature, innovation lies in the overarching architectures, which are capable of processing multiple data sources and have not been previously applied to their respective domains. In Chapter 2, LSTM processes both textual and numerical data, Chapter 3 applies LSTM to static and sequential data, and Chapter 4 utilizes GRU, also for static and sequential inputs.

Innovation through explainability. Utilizing novel data sources sequentially as input for DL models poses a challenge for obtaining explainable results and conveying actionable insights to various stakeholders. Each chapter employs distinct techniques aligned with its specific objective. For instance, Chapter 2 employs structural break analysis to pinpoint keywords' emergence during critical moments in oil price shifts, irrespective of direction. In Chapter 3, integrated gradients spotlight changes in customer behavior preceding life events. Chapter 4 utilizes SHAP values to highlight shifts in customer behavior over time, aggregating per product category for more generalizable insights.

In sum, the research included in this thesis is highly relevant from a decision-maker and managerial perspective. The specific managerial implications for each study have already been discussed in their respective chapters. However, additional general conclusions can further be drawn. First, all research was performed on real-world data, meaning the methodologies can be transferred to other business applications. Second, the challenges present in the financial services industry, of resource allocation and effective customer relationship management, can be addressed through the methodologies deployed in these studies. Thus, decision-makers and managers should consider adding DL modeling, sequential data, and explainability techniques as additional tools to continually improve their customer targeting. Third, Chapters 3 and 4 use data from the same source for life event prediction and recommender systems, respectively. As such, managers can combine both approaches to further improve customer targeting, resource allocation, and personalized services, especially at the right moment in time, to strengthen their overall CRM strategies.

5.2.2 Limitations and future research

Throughout this thesis, we have clearly stated the contributions of our work towards data-driven innovation in the financial services industry. Each chapter already contains a brief section of limitations for a particular study. However, for the sake of completeness, this section discusses shortcomings and opportunities for additional research across the three different studies.

Chapter 2 employs social media textual data and lagged oil prices in LSTM to predict future prices. Insights were drawn using structural break analysis, but further understanding may be achieved through different ways. Firstly, different DL models can be evaluated, as well as using attribution methods for more detailed insights. Secondly, enhancing the overall architecture with an attention layer may provide clearer visualization of feature importance. Thirdly, transforming the oil price prediction target into a binary variable, such as price direction, could clarify the interaction between textual data and oil prices. Lastly, using a longer historical data horizon, encompassing tweets from all previous US presidents, may offer generalizable insights into the impact of leadership. These modifications may uncover complementary insights into how LSTM leverages text for improved oil price prediction performance, whether leadership differences are captured through text, and how they affect oil price changes.

Chapter 3 contributes to life event prediction literature, but also reveals various paths for further research. Firstly, additional studies are needed to assess the impacts of various attribution methods, aiding decision-makers in understanding model explanations for life event prediction. Secondly, tracking diverse customer behaviors during life events can enhance marketers' personalized strategies, fostering stronger customer engagement and responses (Han and Anderson, 2022). For instance, conducting A/B tests on highly personalized recommendations for customers approaching life events can provide dynamic insights into their behaviors (De Caigny et al., 2020; Cheng and Chen, 2022). Thirdly, exploring longer historical customer data periods may unveil insights into key moments for deploying marketing actions well in advance.

Chapter 4 contributes to research on recommendation systems in the financial services industry, but also contains limitations for future consideration. Firstly, a broader array of predictor features could have led to the development of more proficient MLC models. For instance, more intricate socio-demographic data (e.g., income, household size, and educational attainment), more comprehensive consumption details (e.g., allocation of expenditures, sales distribution channels), and indicators of customer satisfaction (e.g. textual data from customer complaints) could further enrich the model results. Secondly, detailed product specifications, such as website product descriptions, could have facilitated the deployment of a content-based recommender system. Despite the uncertain performance of a content-based recommender system within the ambit of financial services, its exploration holds substantive merit, particularly in the domain of cross-selling. It may also be integrated, in conjunction with collaborative filtering, as part of a hybrid approach. Moreover, the use of hybrid approaches also allow the opportunity to further research the cold-start problem (Fernandes et al., 2023).

For both Chapters 3 and 4, a subsequent avenue for upcoming research entails the execution of an empirical experiment wherein the DL model is applied to the actual clientele of the financial services provider. This approach seeks to corroborate whether the envisioned models improve the cross-selling efficacy. Such experiments are highly complex, as they require judicious marketing initiatives and an adequate length of testing horizon. However, successful results would pose a key advantage, as it may be inferred that our results bear broader generalizability, and may also be relevant to develop important profit metrics.

Similarly, longer longitudinal data horizons and information from competitors' offers would expose the model to changes in market conditions, economic cycles, and interest rates that also bear broader generalizability. Inner information, such as solicitations or marketing campaigns, could impact account usage, balance, retention, and as such are interesting information to add to financial services research.

Overall, the inclusion of additional data for future research also inevitably leads to another pending study: the scalability of these approaches. Therefore, a final avenue of research we suggest is the construction of a clear framework, depending on resource constraints, data availability, and size of the clientele.

5.3 References

Cheng, L.C., Chen, K., 2022. Mining longitudinal user sessions with deep learning to extend the boundary of consumer priming. Decis. Support Syst. 162, 113864. doi:10.1016/j.dss.2022.113864.

De Caigny, A., Coussement, K., De Bock, K.W., 2020. Leveraging fine-grained transaction data for customer life event predictions. Decision Support Systems 130, 113232. doi:doi.org/10.1016/j.dss.2019.113232.

Fernandes, L., Miguéis, V., Pereira, I., e Oliveira, E., 2023. Towards hyper-relevance in marketing: Development of a hybrid cold-start recommender system. Applied Sciences 13. doi:10.3390/app132312749.

Han, S., Anderson, C.K., 2022. The dynamic customer engagement behaviors in the customer satisfaction survey. Decis. Support Syst. 154, 113708. doi:10.1016/j.dss.2021.113708.

List of Figures

1.1 Illustration schématique des insights basés sur les données dans le secteur des services financiers . . . 4
1.2 Nombre d'articles de recherche sur les systèmes de recommandation publiées par année . 4
1.3 Nombre d'articles de recherche utilisant des algorithmes DL et CF par année . . . 5
1.4 Exemples de données séquentielles . . . 7
1.5 Schematic illustration of data-driven insights in the financial services industry . . . 16
1.6 Number of recommendation systems studies published per year . . . 16
1.7 Number of studies using DL and CF algorithms per year . . . 17
1.8 Examples of sequential data . . . 19

2.1 Schematic illustration of experimental setup . . . 34
2.2 Timeline showing how the data is split . . . 35
2.3 Top 10 token frequency from Donald Trump's twitter account . . . 41
2.4 Timeline of Brent and WTI oil prices with breakpoints resulting from structural change analysis . . . 41
2.5 Breakpoints from structural change analysis . . . 42
2.6 System design showing four consecutive steps . . . 49
2.7 LSTM memory cell . . . 53
2.8 3-fold cross-validation on the original training set, or 60% of the data . . . 54

3.1 Schematic illustration of an LSTM layer . . . 65
3.2 Featurization for SaML classifiers. The aggregation approach transforms continuous (F^C) and ordinal (F^O) features differently. The baseline approach transposes all features (F), resulting in a column per month ($M_1, ..., M_{12}$) for each feature. . . . 67
3.3 Feature use for LSTM . . . 67
3.4 Experimental setup . . . 69
3.5 Rolling window design . . . 70
3.6 Nemenyi post hoc test for SaML classifiers with featurized data, using AUC. Notes: Classifiers that are not significantly different at $p = 0.10$ are connected. . . . 74
3.7 Nemenyi post hoc test for SaML classifiers with featurized data, using F1. Notes: Classifiers that are not significantly different at $p = 0.10$ are connected. . . . 76
3.8 Integrated gradients showing mean normalized attributions per time step . . . 78
3.10 Integrated gradients showing the distribution of mean normalized attributions per feature type (%) . . . 78
3.9 Integrated gradients showing mean normalized attributions for the top 10 features . . . 79
3.11 Retention rates per life event (%), using no life event occurrence as a baseline (0%). . . . 79

4.1 Schematic representation of the features throughout the independent period . . . 99
4.2 Schematic representation of featurization process . . . 100
4.3 Schematic representation of a GRU layer . . . 102
4.4 Schematic representation of GRU for recommendations . . . 102
4.5 RS Bonferroni-Dunn post hoc tests, for a 95% confidence level . . . 105
4.6 MLC Bonferroni-Dunn post hoc tests, for a 95% confidence level . . . 107
4.7 Mean absolute SHAP values by product category . . . 109
4.8 Mean absolute SHAP values by product category . . . 110
4.9 Mean Absolute SHAP Values . . . 110
4.10 Mean Absolute SHAP Values . . . 111
4.11 Bonferroni-Dunn test for AA configuration . . . 116
4.12 Bonferroni-Dunn test for BR configuration . . . 116
4.13 Bonferroni-Dunn test for CC configuration . . . 116
4.14 SHAP values for insurance product A01 . . . 117
4.15 SHAP values for insurance product A02 . . . 117

4.16 SHAP values for credit product C02 118
4.17 SHAP values for credit product C07 118
4.18 SHAP values for credit product C08 118
4.19 SHAP values for credit product C09 118
4.20 SHAP values for savings product E02 119
4.21 SHAP values for savings product E04 119
4.22 SHAP values for savings product E05 119
4.23 SHAP values for savings product E06 119
4.24 SHAP values for services product S01 120
4.25 SHAP values for services product S02 120
4.26 SHAP values for services product S03 120

List of Tables

2.1 Notable oil tweets by former President Trump . . . 31
2.2 Literature review: oil price prediction using deep learning models . . . 33
2.3 Performance of LSTM models, comparing structured data to LSTM models including both structured and textual data . . . 38
2.4 Modified DM test p-values for LSTM models with structured and textual data . . . 38
2.5 Performance of benchmark models & $\text{LSTM}_{\text{BERT}}$ . . . 39
2.6 Modified DM test p-values from comparing $\text{LSTM}_{\text{BERT}}$, the best performing LSTM model, to the benchmark models . . . 39
2.7 $\text{LSTM}_{\text{BERT}}$ comparison with oil-related keyword exclusion . . . 40
2.8 DM test p-values for comparing $\text{LSTM}_{\text{BERT}}$ with the full text against different variations of oil-related keywords exclusion . . . 40
2.9 LSTM hyperparameter tuning summary . . . 55
2.10 Benchmark model hyperparameter tuning summary . . . 56

3.1 Life event prediction literature review . . . 64
3.2 Life event definitions . . . 71
3.3 Life event subset definitions . . . 72
3.4 Hyperparameter tuning summary . . . 73
3.5 TDL results for SaML classifiers . . . 75
3.6 AUC results for SaML classifiers . . . 75
3.7 F1 results for SaML classifiers . . . 75
3.8 Predictive performance results for XGB-AA and LSTM . . . 76
3.9 Wilcoxon Signed-Rank test results for random classifier against LSTM . . . 77

4.1 Literature review for studies on RS for financial services . . . 96
4.2 Literature review for studies using longitudinal data . . . 97
4.3 Definitions of available independent features . . . 98
4.4 Hyperparameter tuning summary for MLC . . . 101
4.5 Hyperparameter tuning summary for RS . . . 103
4.6 Product categories . . . 104
4.7 5x2 cross-validation mean results for traditional RS . . . 104
4.8 5x2 cross-validation mean F1 results for MLC algorithms with featurization . . . 106
4.9 5x2 cross-validation mean G-mean results for MLC algorithms . . . 106
4.10 5x2 cross-validation mean results for best MLC algorithms and baseline versions . . . 107
4.11 5x2 cross-validation mean results for GRU and best performing MLC and RS algorithms 108
4.12 Summary statistics for distinct number of products owned per customer . . . 115
4.13 Summary for dimensionality reduction . . . 115
4.14 5x2 cross-validation mean results for MLC algorithms without featurization . . . 117