UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

NETWORK REPRESENTATION LEARNING FOR CREDIT SCORING

TESIS PARA OPTAR AL GRADO DE DOCTOR EN SISTEMAS DE INGENIERÍA

RICARDO LUIS MUÑOZ CANCINO

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

PROFESOR CO-GUÍA:
CRISTIÁN BRAVO ROMÁN

MIEMBROS DE LA COMISIÓN:
MANUEL GRAÑA ROMAY
MARCOS ORCHARD CONCHA

Este trabajo ha sido parcialmente financiado por ANID, Beca FOLIO: 21190345

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE DOCTOR EN SISTEMAS DE INGENIERÍA
POR: RICARDO LUIS MUÑOZ CANCINO
FECHA: 2023
PROF. GUÍA: SEBASTIÁN RÍOS PÉREZ

APRENDIZAJE DE REPRESENTACIONES EN REDES COMPLEJAS PARA CLASIFICACIÓN CREDITICIA

El acceso al crédito juega un papel crucial en la sociedad y beneficia la economía. Permite a las personas alcanzar sus metas como adquirir vivienda, acceder a atención médica y obtener educación. Además, brinda a los emprendedores y empresas el capital necesario para iniciar o expandir operaciones, creando empleos y fomentando el crecimiento económico. Es necesario contar con mecanismos para medir el riesgo de incumplimiento crediticio, ya que ayudan mantener a la estabilidad del sistema financiero y protegen a los prestatarios de préstamos impagables, asegurando que no enfrenten riesgos financieros innecesarios. La investigación en credit scoring busca mejorar la discriminación de los modelos mediante mejores algoritmos e incorporando datos alternativos como redes o grafos. Estos datos capturan interacciones familiares, sociales y económicas de los individuos y ha demostrado ser especialmente útil con prestatarios con historial crediticio limitado o inexistente. Esta tesis explora el valor de integrar datos de grafos en modelos de credit scoring, con tres objetivos específicos, cada uno vinculado a una publicación diferente. El primer objetivo se centra en el uso de network representation learning en modelos de credit scoring. Se presenta un framework que combina atributos generados manualmente, graph embeddings y atributos obtenidos de redes neuronales de grafos. El estudio valida el uso de datos de redes en préstamos corporativos y de consumo, y revela que el impacto de la información de grafos varía según el prestatario, ya sean personas o empresas. Este es el primer estudio que considera el comportamiento crediticio de todo un país utilizando diversas relaciones sociales y económicas. Nuestros resultados resaltan el valor de los datos de redes para abordar los desafíos que enfrentan particularmente para las empresas con historial crediticio limitado o nulo, facilitando su inclusión en el sistema financiero. El segundo objetivo busca comprender el impacto de los datos de redes en el desempeño de los modelos a medida que el comportamiento de pago adquiere relevancia. Este trabajo desafía la división actual del proceso de gestión de riesgo de crédito al examinar etapas intermedias entre application credit scoring y behavioral credit scoring. Al centrarnos en el prestatario en lugar del proceso comercial, encontramos información valiosa sobre la dinámica del desempeño de los modelos a medida que evoluciona el historial crediticio. Además, investigamos la influencia de los atributos de redes y observamos que su valor decrece en presencia de atributos de comportamiento. En nuestro tercer objetivo, presentamos una metodología para entrenar un modelo en datos sintéticos y luego aplicarlo a datos reales. Los resultados muestran que es posible entrenar un modelo con datos sintéticos que funcione bien en situaciones reales. Sin embargo, observamos que al aumentar el número de atributos, disminuye la calidad de los datos sintéticos. Además, identificamos un costo en el desempeño asociado con trabajar en un entorno que preserva la privacidad. Este costo es una reducción del poder predictivo, que en nuestro estudio fue de un 3% en el área bajo la curva ROC y un 6% en el estadístico de Kolmogorov-Smirnov. Los hallazgos de esta tesis aportan a una comprensión integral de los modelos de credit scoring, destacando la importancia de considerar los datos de redes y las oportunidades para la investigación de behavioral credit scoring mediante el aumento de datos de entrenamiento a partir de datos sintéticos.

NETWORK REPRESENTATION LEARNING FOR CREDIT SCORING

Access to credit plays a vital role in society and significantly benefits the economy. It enables individuals to fulfill essential life goals, including acquiring housing, obtaining healthcare, and pursuing education. Moreover, it gives entrepreneurs and businesses access to the necessary capital to initiate or expand operations, create jobs, and promote the economy's growth. Considering its crucial role, it is important to have mechanisms to quantify the risk of loan default. These mechanisms serve the dual purpose of maintaining the financial system's stability and protecting borrowers from being granted loans they cannot afford, ensuring they are not exposed to unnecessary financial risks. Credit scoring research has recently focused on enhancing model discriminatory power through improved assessment algorithms and incorporating alternative data. Regarding the alternative data, we focus on utilizing network or graph representations. This form of data captures the individual's familial, social, and economic interactions. It has proven especially useful with borrowers with limited or nonexistent credit history. This doctoral thesis explores the value of integrating network data into credit scoring models. The research is driven by three specific aims, each corresponding to a distinct publication. The first objective concerns network representation learning, different methods to extract knowledge from networks, and their effect on credit scoring models. This work introduces a framework that combines traditional hand-engineered features with graph embeddings and graph neural network features for credit scoring. The study validates the use of graph data in corporate and consumer lending, revealing that the impact of graph information varies depending on the borrower being analyzed, whether individuals or companies. It is the first study to consider the credit behavior of an entire country using various social and economic relationships such as parents, spouses, business owners, employers, employees, and transactional services. Our results highlight the significant value of graph data in addressing the credit scoring challenges faced by thin-file borrowers, particularly for companies with limited or no credit history. This valuable information can facilitate such entities' entry into the financial system. The second objective deepens the previous results and seeks to understand the impact of graph data on performance as the payment behavior gains relevance. This work challenges the current division of the credit risk management process by examining the intermediate stage between application scoring and behavioral scoring. By turning the focus onto the borrower instead the business process, we found valuable insights into the performance dynamics of credit scoring models as the borrower's credit history evolves. Furthermore, we investigate the influence of graph-data features and observe their diminishing value in the presence of behavioral attributes. Finally, in our third aim, we introduce a framework that enables training a model on synthetic data and then applying it to real-world data. Additionally, we investigate the model's ability to handle data drift by evaluating its performance on real-world data gathered one year later. Our findings demonstrate that training a model on synthetic data that perform well in real-world situations is possible. However, we observed that as the number of features increases, the quality of the synthesized data decreases. Furthermore, we identified a performance cost associated with working in a privacy-preserving environment. This cost corresponds to a reduction in predictive power, which in our study was approximately 3% when measured using the area under the curve and 6% in the Kolmogorov-Smirnov statistics. This thesis's findings contribute to a more comprehensive understanding of credit scoring, highlighting the importance of considering graph data and the possibilities for behavioral scoring research using data augmentation from synthetic data.

Sé que desde el cielo
me observas con orgullo en este día.

Table of Content

List of Tables

List of Figures

Chapter 1

Introduction

This introductory chapter presents the motivation to study network representation for credit scoring. Section 1.1 starts with the research's motivation and introduces the research area addressed in this thesis. Section 1.2 presents the research problem and the thesis's general and specific objectives. Then, in Section 1.3, we show the research methodology to continue with the contributions of this research in Section 1.4, and its results, in the form of publications, are detailed in the Section 1.5. Finally, the structure of the thesis is presented in Section 1.6.

1.1. Motivation

Access to credit is widely regarded as one of the economy's main engines, a key driver for economic growth (Rajan & Zingales, 1996; Banu, 2013; Van, Vo, Nguyen, & Vo, 2021). Credit enables individuals to fulfill their personal aspirations, life goals and achieve financial success. Goals such as access to housing, healthcare, and education are often achievable only through credit financing (Hurley & Adebayo, 2017; Aziz & Dowling, 2019). Additionally, it allows people and companies to access the necessary capital to create or expand businesses, create jobs, and contribute to the economy's growth (Diallo & Al-Titi, 2017). The main players in the lending ecosystem are banks and financial institutions, which provide credit access to individuals and businesses. Lending money is not a risk-free act, which is why credit risk is one of the primary sources of uncertainty faced by banks and financial institutions (Apostolik, Donohue, & Went, 2009). This credit risk refers to the likelihood that a borrower will default on a loan, which can result in economic losses for the lender (The Basel Committee on Banking Supervision, 2000). For this reason, financial institutions have used techniques to estimate and manage this risk since its beginning. Although these techniques are as old as the concept of lending itself, it was only in the mid-20th century that these concepts gained relevance with the advent of computing (Anderson, 2022). Banks and financial institutions use sophisticated mathematical and statistical models to assess the borrower's creditworthiness. These models are applied throughout the credit life cycle, application, repayment, and collection. Among these models, application scoring and behavioral scoring are particularly noteworthy.

Application scoring is part of the loan granting decision, and it is used to evaluate the creditworthiness of loan applicants and assess the risk of lending them money (Anderson, 2022). With this information, lenders decide on loan approval, interest rate applied, and other conditions based on the borrower's risk level. On the other hand, behavioral scoring is applied in the portfolio management process once the borrower is already part of the credit portfolio of the financial institution (Paleologo, Elisseeff, & Antonini, 2010; Anderson, 2022).

This scoring enables lenders to address borrowers with a high risk of default proactively. For instance, lenders can help alleviate the financial burden of borrowers struggling to keep up with payment schedules and other obligations by implementing payment arrangements or loan term restructuring.

Application scoring and behavioral scoring are intensive in the use of borrower information. Application scoring uses mainly the information provided in the application form, while behavioral scoring incorporates the borrower's past behavior, such as loan repayment history, credit utilization, and borrower historical data. In both cases, borrower knowledge is essential because the more information is provided, the more accurate the creditworthiness assessment will be. It is due to the above that these scoring models have some limitations. For example, it may not fully capture a borrower's creditworthiness if they have a limited credit history and borrower data (Cusmano, 2018; Hurley & Adebayo, 2017; Baidoo, 2020; Djeundje, Crook, Calabrese, & Hamid, 2021). Moreover, behavioral scoring models require significant data for effective training. Historical information from borrowers is crucial in this process. However, the availability of such data poses a significant challenge, as it tends to be scarce and financial institutions are often hesitant to share it (Liu, 2001; Goh & Lee, 2019). This data scarcity frequently hampers research efforts focused on these models (Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013).

Several methods have been developed to deal with these problems when the borrower's information is scarce, their credit behavior is insufficient, or they do not have it. On the one hand, the business model based on microcredits, where the borrower's payment behavior is observed from a limited initial exposure. Moreover, in the presence of good behavior, the lender is willing to increase the credit limits granted to that borrower. However, this model is neither cost-effective nor matches borrowers' needs (Hurley & Adebayo, 2017; Baidoo, 2020). In contrast, there is widespread consensus that enhancing the efficacy and effectiveness of credit scoring models is a feasible strategy, which can be achieved by improving the creditworthiness prediction algorithms, integrating alternative data, or utilizing a hybrid approach that combines both methods. The use cases of alternative data are varied, telephone call data (Óskarsdóttir et al., 2017; Óskarsdóttir, Bravo, Vanathien, & Baesens, 2018a; Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019), written risk assessments (Stevenson, Mues, & Bravo, 2021), data generated by an app-based marketplace (Roa, Correa-Bahnsen, et al., 2021; Roa, Rodríguez-Rey, Correa-Bahnsen, & Valencia, 2021), social media data (Tan & Phan, 2018; Cnudde et al., 2019; Putra, Joshi, Redi, & Bozzon, 2020), network information (Ruiz, Gomes, Rodrigues, & Gama, 2017), behavioral and psychological surveys (Goel & Rastogi, 2021), fund transfers datasets (Shumovskaia, Fedyanin, Sukharev, Berestnev, & Panov, 2020; Sukharev, Shumovskaia, Fedyanin, Panov, & Berestnev, 2020), and psychometric data (Rabecca, Atmaja, & Safitri, 2018; Djeundje, Crook, Calabrese, & Hamid, 2021; Rathi, Verma, Jain, Nayyar, & Thakur, 2022). These sources share the common use of social-interaction information gathered from the graph formed by the interactions among individuals recorded in alternative data sources. This is why social relationships also play an important role in determining a borrower's creditworthiness. In particular, family, friends, and financial relationships can affect an individual's credit risk by shaping their behavior and financial decisions. In addition, family and friends can provide financial assistance in times of financial need. Similarly, economic relationships such as business partnerships or financial transactions can significantly impact the credit risk of individuals. Sometimes, it is difficult to differentiate the financial management of a small business from the personal finances of its owner, or many times the business partner's financial difficulties can threaten the solvency of the company or its other owners. The study of how these relationships influence the evaluation of creditworthiness has gained strength in recent years thanks to advances in graph analysis, deep learning, and graph convolutional networks. The most significant advances in this area have been in showing the value of incorporating this alternative data in the credit scoring problem and testing different methods to incorporate the information from the graphs into the traditional credit scoring problem.

Despite extensive research on credit scoring, remaining challenges need to be tackled. This thesis aims to investigate and analyze various strategies for integrating graph information into conventional credit scoring problems. Furthermore, it seeks to identify the stages in the credit life cycle where incorporating such alternative information proves most valuable. Lastly, this study aims to explore methods for facilitating research in behavioral credit scoring models. The following section will outline this thesis's objectives and describe how we intend to address the points mentioned earlier. The related research questions and publications will be presented.

1.2. Research Problem

1.2.1. General Objective

The main objective of this thesis is to extend the general knowledge of how to build and train credit scoring models by incorporating network data.

1.2.2. Specific Aims

The idea is to investigate the value of incorporating network data in credit risk management. The research will be structured with three specific aims to achieve this goal. Each specific aim in this thesis is driven by its own research questions, which have been thoroughly investigated and documented in their corresponding articles. These articles have been published in peer-reviewed journals and conferences, providing insights and findings for further research.

1.2.2.1. Aim 1: On the combination of graph data

Context:

This specific aim seeks to address two gaps in the literature regarding using graph data in credit scoring. The first gap is related to the data sources utilized. Many studies have used partial social networks that fail to capture the overall picture of the borrower's social interactions. Secondly, the network knowledge extraction has mainly relied on both handmade feature engineering (Freedman & Jin, 2017; Ruiz, Gomes, Rodrigues, & Gama, 2017; Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019; Niu, Ren, & Li, 2019) and, in recent years, through graph neural networks (Roa, Rodríguez-Rey, Correa-Bahnsen, & Valencia, 2021) that are no improvement over the traditional feature-engineering approach. This specific aim will investigate combining different representation learning techniques with complex graph structures instead of observing them in isolation.

Research Questions:

- When combining different graph representation learning (GRL) techniques over complex graph structures, is there a performance improvement compared to merely applying hand-crafted feature engineering or graph neural networks?

- What insights are obtained into the combined network features, and what value do these insights add to credit risk assessment?
- Where does social information help the most? Is the most significant performance enhancement obtained in personal credit scoring or business credit scoring? What can we gather from this information? Does it influence which network and which features are the most relevant?

1.2.2.2. Aim 2: On the dynamics of graph data features and their impact on performance

Context:

We know the impact of repayment behavior and social-interaction data on application and behavioral scoring problems. The effect of repayment behavior is gaining relevance as the borrower-lender relationship becomes entrenched: the more data on borrower's behavior is gathered, the more accurately borrower's creditworthiness can be predicted. As for social interaction data, at some point, its relevance decreases in the presence of the borrower's behavior and repayment history. Both relationships require careful study. Research into credit scoring has primarily focused on the initial stages (application scoring) and at some point during the loan payment schedule (behavioral scoring). Therefore, examining these dynamics enhances our understanding of credit scoring.

Research Questions:

- Knowing that borrowers' repayment history increases creditworthiness assessment performance, at which point in time since the loan is granted, does this information become meaningful? For how long do we need to observe borrowers' repayment history to assess their creditworthiness accurately?
- Knowing that social-interaction data contributes more value to application scoring, that is when behavioral information is scarce. For how long is it beneficial to rely on these sources of information?
- What insights and benefits to credit risk management are obtained from studying the dynamics of both the creditworthiness assessment performance and the value of alternative data sources?

1.2.2.3. Aim 3: On the training of credit scoring models using synthetic data

Context:

Despite all the years of research on credit scoring, behavioral scoring models have received relatively less attention because it requires large volumes of data and a relevant historical depth (Goh & Lee, 2019; Kennedy et al., 2013). In addition, financial institutions are often hesitant to collaborate in this type of research due to worries about data security and individuals' privacy protection. Currently, the use of synthetic data in credit scoring is mainly restricted to balancing the minority class using the traditional SMOTE (Gicić & Subasi, 2019) or variational autoencoders (Wan, Zhang, & He, 2017), and lately, generative adversarial networks

(Fiore, De Santis, Perla, Zanetti, & Palmieri, 2019; Lei et al., 2020; Ngwenduna & Mbuvha, 2021). Therefore, we want to study if it is possible to train a model on synthetic data and then apply it to real-world data, achieving performance as good as a model trained on real-world data.

Research Questions:

- Can a model trained on synthetic data perform well in real-world scenarios?
- How does increasing the features impact synthetic data quality?
- Is there a performance cost for working in a privacy-preserving environment?

1.3. Proposed Methodology

The proposed methodology to address the specific aims outlined in this doctoral thesis is thoroughly explained in each of the corresponding articles. However, several common key elements are applied across all the proposed methodologies that we can summarize as shown in Figure 1.1. This methodology is an adaptation of traditional data mining approaches, such as the Knowledge Discovery in Databases (KDD) methodology (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) and CRISP-DM Cross-Industry Standard Process for Data Mining (Wirth & Hipp, 2000), but with specific modifications and adjustments to address the challenges and objectives raised in this research.

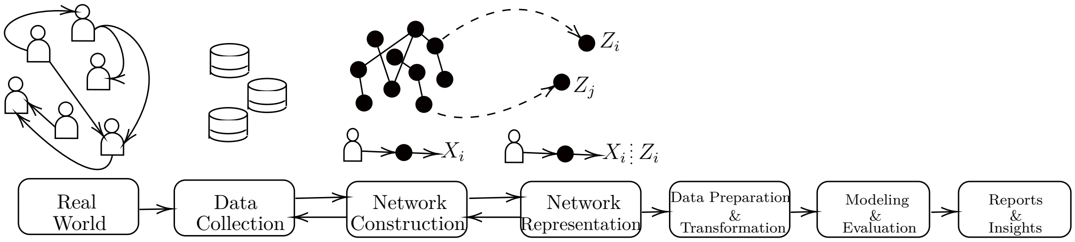


Figure 1.1: Proposed methodology

The first step in the methodology involves a broad exploration of the real-world context, where individuals, including borrowers, interact with each other. These interactions are reflected through multiple relationships recorded in numerous computer systems. We carefully select specific records from these systems that can influence the problem of predicting creditworthiness. We have chosen those observed relationships that allow us to build social and economic networks to study how they affect creditworthiness assessment. In particular, in all three studies, we hold information regarding the credit behavior of an entire country, as well as data pertaining to familial relationships, corporate structures, and networks between employers and workers. By leveraging these specific data sources, we can gain a more complete and accurate view of borrowers' financial situation, thus enhancing the creditworthiness assessment.

Once the data has been collected, it is necessary to define how this information is structured in the form of graphs in the network construction stage. Subsequently, several network representation learning methods, the ways of extracting knowledge from the graph data, are studied to include this knowledge within the problem of traditional credit scoring problem. Although the experimental configuration depends on each study in the modeling stage, the process is structured based on K-fold cross-validation and bootstrap sampling to obtain conclusions supported by statistical tests. These conclusions are obtained by comparing state-of-the-art models with the particular proposed models.

Regarding validation metrics, standard industry metrics are also used, such as the area under the ROC curve and the Kolmogorov-Smirnov statistic. In the last stage, the results are analyzed, and relevant knowledge is obtained about each research question presented. In summary, this methodology runs from the conceptualization, collection, and structure of data in the form of graphs to its representation in tabular format, modeling, and statistical validation.

1.4. Contributions and outline

Our research extends the understanding of credit scoring by examining the use of social network data. Firstly, we proposed a framework that combines traditional hand-engineered features, graph embeddings, and graph neural network features to generate a single credit score, enabling a simpler decision-making process in credit approval. Furthermore, our research challenges the conventional division between application and behavioral scoring, focusing on the borrower's credit history evolution and investigating the contribution of graph data features. Our massive dataset enables us to extend the behavioral credit scoring research by characterizing individuals and companies from granting their first loan and capturing subsequent credit history, repayment behavior, and social network data. It enables a comprehensive study of credit assessment performance dynamics and the value of graph data throughout the study period of each borrower. Moreover, we present a framework that uses synthetic data to train credit scoring models. We evaluated their performance on real-world data and examined their stability to data drift. Our findings show the feasibility of achieving good model performance using synthetic data, although the quality of synthesized data decreases when the number of synthesized features increase. We also observe a performance cost associated with these privacy-preserving practices, which reduces predictive power when synthetic data is used to train models. Finally, we validate and test the efficacy of graph data in both corporate and consumer lending contexts, highlighting its varying impact on different borrower types and the predictive power enhancement this alternative data offers. Our study pioneers the analysis of the credit behavior of an entire country, incorporating diverse social and economic relationships such as parental, marital, business ownership, and employment networks, thereby expanding the understanding of credit scoring with extensive social interaction data.

Our research contributes to credit risk management by proposing an integrated framework and analyzing the impact of credit history, repayment behavior, and social network features on the dynamics of creditworthiness assessment performance. It also incorporates social network data comprehensively and demonstrates the viability of training credit scoring models using synthetic data. These contributions enhance our understanding of credit risk assessment and improve credit decision-making processes.

This thesis consists of five chapters. Chapter 2 presents the publication *"On the combination of graph data for assessing thin-file borrowers' creditworthiness,"* where the use of graph data for the credit scoring problem is investigated. In particular, it seeks to identify the most appropriate way to incorporate this information into the traditional credit scoring problem. In order to meet this objective, three ways are studied to convert social network data in the form of graphs into tabular data. This work is published in the *Expert Systems* *with Applications* journal.

Chapter 3 presents the publication *"On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance"*. This study focuses on understanding the dynamics of creditworthiness assessment performance and its relationship with credit history, repayment behavior, and social network features. The findings indicate that including social network features significantly impacts loan application scoring, lasting for approximately six months in individual scoring and persisting throughout the study period in business scoring. This work is published in the *Expert Systems with Applications* journal.

Chapter 4 presents the conference paper *"Assessment of creditworthiness models privacy-preserving training with synthetic data"*. Where the problem of low data availability for training behavioral credit scoring models is addressed, this article shows that it is possible to train behavior models using synthetic data and thus preserve clients' privacy. The results show that models trained with synthetic data lose predictive power compared to models trained with real data. However, these losses of predictive power are 3% of AUC and 6% of KS. Thus there is a trade-off between accessing more significant volumes of information and the predictive power of the associated models. This work is published in the *Hybrid Artificial Intelligent Systems 2022* conference.

Finally, a general conclusion regarding all the work developed is presented in Chapter 5.

1.5. Publications

These articles are a direct outcome of the research conducted in this thesis and were produced and published as part of the Ph.D. program.

- Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña, On the combination of graph data for assessing thin-file borrowers' creditworthiness, Expert Systems with Applications, Volume 213, Part A, 2023, 118809, ISSN 0957-4174 (Muñoz-Cancino, Bravo, Ríos, & Graña, 2023a)

- Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña, On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance, Expert Systems with Applications, Volume 218, 2023, 119599, ISSN 0957-4174 (Muñoz-Cancino, Bravo, Ríos, & Graña, 2023b)

- Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña (2022). Assessment of Creditworthiness Models Privacy-Preserving Training with Synthetic Data. In: , et al. Hybrid Artificial Intelligent Systems. HAIS 2022. Lecture Notes in Computer Science(), vol 13469. Springer, Cham. (Muñoz-Cancino, Bravo, Ríos, & Graña, 2022)

Other articles were developed during the doctoral studies; however, they do not pertain to the subject matter of this research.

- Muñoz-Cancino, R., Rios, S. A., Goic, M., & Graña, M. (2021). Non-Intrusive Assessment of COVID-19 Lockdown Follow-Up and Impact Using Credit Card Information: Case Study in Chile. International Journal of Environmental Research and Public Health, 18(11), 5507. (Muñoz-Cancino, Rios, Goic, & Graña, 2021)

- Muñoz-Cancino, R., Rios, S. A., & Graña, M. (2023). Clustering Cities over Features Extracted from Multiple Virtual Sensors Measuring Micro-Level Activity Patterns Allows One to Discriminate Large-Scale City Characteristics. Sensors. 2023; 23(11):5165. (Muñoz-Cancino, Ríos, & Graña, 2023)

- Ricardo Muñoz-Cancino, Sebastián A. Ríos, Manuel Graña. Predicting innovative cities using spatio-temporal activity pattern. Hybrid Artificial Intelligent Systems. HAIS 2023. Under Review.

1.6. Structure of the thesis

This work is structured in chapters, which are detailed below:

- **Chapter 2**: presents the publication *"On the combination of graph data for assessing thin-file borrowers' creditworthiness,"*

- **Chapter 3**: presents the publication *"On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance"*

- **Chapter 4**: presents the publication *"Assessment of creditworthiness models privacy-preserving training with synthetic data"*

- **Chapter 5**: Summarizes the main conclusions and findings from the three studies presented in this doctoral thesis.

Chapter 2

On the combination of graph data for assessing thin-file borrowers' creditworthiness[1]

Abstract

Thin-file borrowers are customers for whom a creditworthiness assessment is uncertain due to their lack of credit history. To address missing credit information, many researchers have used borrowers' social interactions as an alternative data source. Exploiting social networking data has traditionally been achieved by hand-crafted feature engineering, but lately, graph neural networks have emerged as a promising alternative. Here we introduce an information-processing framework to improve credit scoring models by blending several methods of graph representation learning: feature engineering, graph embeddings, and graph neural networks. In this approach, we aggregate the methods' outputs to be fed to a gradient boosting classifier to produce a final creditworthiness score. We have validated this framework over a unique multi-source dataset that characterizes the relationships, interactions, and credit history for the entire population of a Latin American country, applying it to credit risk models, application, and behavior. It also allows us to study both individuals and companies. Our results show that the methods of graph representation learning should be used as complements; they should not be seen as self-sufficient methods, as it is currently done. We improve the creditworthiness assessment performance in terms of the measures of Area Under the ROC Curve (AUC) and Kolmogorov-Smirnov (KS), outperforming traditional methods of exploiting social interaction data. In the area of corporate lending, where the potential gain is much higher, our results confirm that the evaluation of a thin-file company cannot solely consider the company's own characteristics. The business ecosystem in which these companies interact with their owners, suppliers, customers, and other companies provides novel knowledge that enables financial institutions to enhance their creditworthiness assessment. Our results let us know when and on which population to use graph data and the expected effects on performance. They also show the enormous value of graph data on the credit scoring problem for thin-file borrowers, mainly to help companies with thin or no credit history to enter the financial system.

Keywords: credit scoring; machine learning; social network analysis; network data; graph

[1] The following is a copy of the paper published at the Expert Systems with Applications Journal. Please cite this paper as follows: Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña, On the combination of graph data for assessing thin-file borrowers' creditworthiness, Expert Systems with Applications, Volume 213, Part A, 2023, 118809, ISSN 0957-4174

neural networks

2.1. Introduction

A large part of the population requires access to credit to achieve their life goals: social mobility, owning a home, and financial success. Moreover, access to financial services and a proper credit evaluation can facilitate and are often necessary to obtain a job, rent a home, buy a car, start a new business, or pursue a college education (Hurley & Adebayo, 2017; Aziz & Dowling, 2019). At the macroeconomic level, access to credit is a major driver for local economic growth, especially in developing economies (Diallo & Al-Titi, 2017). Financial institutions play a significant social role in facilitating access to credit and facing the entailed risks of lending money. To manage this credit risk, financial institutions have applied credit scoring models to assess the creditworthiness of their borrowers, that is, to distinguish between good and bad payers and delivering loans to those who are most likely to repay. To build a credit scoring model, financial institutions often use personal information, banking data, and payment history to estimate creditworthiness and the probability of default. Despite being the standard mechanism in the industry for credit-granting decisions and the management of the loan's life cycle (L. Thomas, Crook, & Edelman, 2017), this ubiquitous tool still does not ensure adequate access to credit and to the financial system.

The World Bank estimates that more than 1.4 billion adults remain unbanked, without access to the financial system (The Global Financial Index, 2022). This number only considers those who do not have a bank account through either a financial institution or mobile banking. If we included underbanked people, that is, those who have an account but cannot apply for a loan, this number would be much larger. Being unbanked or underbanked raises the issue of those who lack a credit history, also known as thin-file borrowers: people who have no access to a loan not because they are bad payers but because they lack the attributes evaluated by traditional credit scoring models (Cusmano, 2018; Hurley & Adebayo, 2017; Baidoo, 2020; Djeundje, Crook, Calabrese, & Hamid, 2021).

In this scenario, lenders have tried different ways to reach this population; we highlight two business models here. In the traditional business model, the higher risk assumed due to the lack of information is compensated by applying higher interest rates. Alternatively, granting microcredits has been used as a strategy to assess the client's payment behavior under limited exposure. However, neither of these solutions has proven to be cost-effective in addressing the credit needs of this population (Hurley & Adebayo, 2017; Baidoo, 2020).

For this reason, financial institutions, fintech, and researchers have looked in recent years for business-model innovations and better decision-making with the available information. This search is done via developing better scoring algorithms and using alternative data sources to improve credit scoring models. Regarding the use of alternative information, graph data has gained high visibility because it allows an improvement of credit scoring models' performance (Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019; Roa, Rodríguez-Rey, Correa-Bahnsen, & Valencia, 2021).

We have identified two main gaps, which are addressed in this work. The first gap is the data sources employed. Most of the studies are carried out with partial social networks that fail to capture the overall picture of the client's social interactions. These networks are limited by the data provider. Our study uses social networking data to characterize the interactions of the country's entire population, encompassing the complete financial system. Secondly, the network knowledge extraction is mainly done both through hand-made feature engineering (Freedman & Jin, 2017; Ruiz, Gomes, Rodrigues, & Gama, 2017; Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019; Niu, Ren, & Li, 2019) and, in recent years, through graph neural networks (Roa, Rodríguez-Rey, Correa-Bahnsen, & Valencia, 2021) that do are no improvement over the traditional feature-engineering approach.

Our work will investigate the combination of different representation learning techniques with complex graph structures instead of observing them in isolation. Hence, we formulate the following research questions:

1. When combining different graph representation learning (GRL) techniques over complex graph structures, is there a performance improvement compared to merely applying hand-crafted feature engineering or graph neural networks?

2. What insights are obtained into the combined network features, and what value do these insights add to credit risk assessment?

3. Where does social information help the most? Is the most significant performance enhancement obtained in personal credit scoring or business credit scoring? What can we gather from this information? Does it influence which network and which features are the most relevant?

This study challenges traditional hand-crafted feature engineering and the novel approach of graph neural networks (GNNs) by combining multiple GRL methods. In particular, our work contributes to the following aspects.

- We introduce a framework to combine traditional hand-engineered features with graph embeddings and GNN features. This framework produces a single score, helping its users decide whether to approve or reject a credit.

- Our results are the first to validate and test graph data regarding both corporate and consumer lending, showing that the information from graphs has a different effect depending on the analyzed borrower, people, or companies. These effects are reflected both in the predictive power enhancement and in the features relevant in each problem, letting us know not only when and on which population to use social-interaction data but also which effects on creditworthiness prediction performance to expect.

- To the best of our knowledge, this is the first study that considers the credit behavior of an entire country, together with social networks that allow the characterizing of its entire population and consolidate multiple types of social and economic relationships, for example, parental, spouses, business owners, employers and employees, or transactional services.

- This paper also contributes to the growing literature in credit scoring and network data, proposing a mechanism to achieve better results than the popular hand-crafted feature engineering and the novel GNN approach.

This paper is structured as follows. Section 2.2 presents a review of credit risk management, credit scoring and social networks. The GRL methods are presented in Section 2.3. Section 2.4 describes the data sources and features extracted for classification. Section 2.5 shows the proposed information-processing methodology and the adopted experimental design. Section 2.6 presents the results obtained. The conclusions and future work that originated from this research are presented in Section 2.7.

2.2. Background and Related Work

2.2.1. Credit Risk Management

Banks' core business is granting loans to individuals and companies. Granting a loan is not risk-free; in fact, banks are heavily exposed to credit risk (Anderson, 2022), originating from the potential loss due to the debtors' default or their inability to comply with the agreed conditions (The Basel Committee on Banking Supervision, 2000). Banking risk management focuses on detecting, measuring, reporting, and managing all sources of risk. Banks define strategies, policies, and procedures to limit the assumed risk. These strategies encourage and integrate the use of mathematical models for the early detection of potential risks. Credit scoring is widely used for managing credit risk, handling large volumes of data, and capturing complex patterns that are difficult to express as simple business rules. This instrument became popular and ubiquitous in the 1980s, mainly due to advances in computing power and to the growth of financial markets, which made it almost impossible to manage large credit portfolios without this kind of tool (L. Thomas, Crook, & Edelman, 2017).

The regulatory framework also endorses the use of credit scoring models; in fact, the Basel Accords allow banks to manage credit risk with internal ratings. Specifically, banks develop internal models for assessing the expected loss. This assessment can be divided into three components: the probability of default (PD), the loss given default (LGD), and the exposure at default (EAD). The PD is a key component, because it is used to define the credit granting policies and for portfolio management. The general approach to estimating the PD and assessing the borrower's creditworthiness is through classification techniques using demographic features and payment history as explanatory variables.

Over the years, lenders have explored multiple ways to improve creditworthiness assessment, novel machine learning techniques (Moscato, Picariello, & Sperlí, 2021), and non-traditional data sources (Aziz & Dowling, 2019). Multiple lines of research have been established; some of them attempt to understand the characteristics of defaulters (Bravo, Thomas, & Weber, 2015), the feature selection process (Kozodoi, Lessmann, Papakonstantinou, Gatsoulis, & Baesens, 2019; Maldonado, Pérez, & Bravo, 2017), or the transformation of the feature space (Carta, Ferreira, Reforgiato Recupero, & Saia, 2021). However, the most significant improvements have been obtained by the exploitation of alternative data sources such as telephone call data (Óskarsdóttir et al., 2017; Óskarsdóttir, Bravo, Vanathien, & Baesens, 2018a; Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019), written risk assessments (Stevenson, Mues, & Bravo, 2021), data generated by an app-based marketplace (Roa, Correa-Bahnsen, et al., 2021; Roa, Rodríguez-Rey, Correa-Bahnsen, & Valencia, 2021), social media data (Tan & Phan, 2018; Cnudde et al., 2019; Putra, Joshi, Redi, & Bozzon, 2020), network information (Ruiz, Gomes, Rodrigues, & Gama, 2017), behavioral and psychological surveys (Goel & Rastogi, 2021), fund transfers datasets (Shumovskaia, Fedyanin, Sukharev, Berestnev, & Panov, 2020; Sukharev, Shumovskaia, Fedyanin, Panov, & Berestnev, 2020), and psychometric data (Rabecca, Atmaja, & Safitri, 2018; Djeundje, Crook, Calabrese, & Hamid, 2021; Rathi, Verma, Jain, Nayyar, & Thakur, 2022). All these studies have in common the use of social-interaction information, the graph formed of the interactions among individuals recorded in alternative data sources.

There are multiple taxonomies of credit scoring problems. One that has been widely adopted by academics and practitioners distinguishes between application scoring and behavior scoring. On the one hand, application scoring corresponds to a credit scoring system for new customers, where the available information is often scarce and limited. On the other hand, behavioral scoring is a credit scoring system for borrowers with available credit and repayment history. In the current study, both of these credit scoring types and their differences between personal and business clients are explored.

2.2.2. Credit Risk and Social Networks

The inclusion of alternative data in the credit scoring problem has gained relevance in recent years. We define *graph data* as any information that records the relationships or interactions among entities that can be represented by a set of nodes in which edges connect pairs of nodes. We refer to a network as a *Social Network* when nodes are people or companies, and edges denote any social interaction, such as among friends, acquaintances, neighbors, colleagues, or affiliations with the same group (Romero, Uzzi, & Kleinberg, 2019). Mathematically, we describe a network through a graph $G(V, E, A)$, where V is the set of nodes, and E is the set of edges. Let $V = \{v_1, \ldots, v_N\}$ where $|V| = N$ is the number of nodes, and the adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$ with $A_{ij} = 1$ if there is an edge e_{ij} from v_i to v_j, $A_{ij} = 0$ otherwise. Additionally, the graph can be associated with a matrix of node attributes $X \in \mathbb{R}^{N\times F}$, where $X_i \in \mathbb{R}^F$ represents the feature vector of node v_i.

The emerging literature on credit scoring and network data has focused on incorporating hand-crafted features into a traditional credit scoring problem (Óskarsdóttir et al., 2017; Óskarsdóttir, Bravo, Verbeke, Baesens, & Vanthienen, 2018). The authors incorporate network information into the formulation of the customer churn problem, using eight telco datasets originating from around the world. This series of studies outlines the foundations for the incorporation of network data in credit scoring. This framework is applied to the credit scoring problem by (Óskarsdóttir et al., 2018a, 2018b; Óskarsdóttir et al., 2019), where the authors introduce a methodology to enhance smartphone-based credit scoring models' predictive power through feature engineering from a pseudo-social network, combining social-network analysis and representation learning. According to this research, it is feasible to increase the performance of micro-lending smartphone applications, generating high helping potential for financial inclusion.

An extension of this work is to measure the temporal and topological dynamics of credit risk, that is, how it evolves and spreads over the graph representation of the social network. For instance, (Bravo & Óskarsdóttir, 2020) implemented modifications to the PageRank algorithm to quantify this phenomenon. This methodology allows them to quantify the risk of the different entities in a multilayer network. Their results show how the risk of default spreads and evolves over a network of agricultural loans. Then, (Óskarsdóttir & Bravo, 2021) analyzed how to build the multilayer network, interpret the variables derived from it, and incorporate this knowledge into credit risk management. Their results reveal that the default risk increases as a debtor presents links with many defaulters; however, this effect is mitigated by the size of each individual's neighborhood. These results are significant because they indicate that default and financial-stability risk spread through the network.

Other works have used an approach based on a graph convolutional network (GCN) for this purpose. (Shumovskaia et al., 2020) present one of the first empirical works with massive graphs created from transactions between clients of a large Russian bank. They propose a framework to estimate links using SEAL (Zhang & Chen, 2018) and recurrent neural networks, the SEAL-RNN framework. One of the advantages of using SEAL is that it focuses only on the link's neighborhood to be predicted, and it does not use the entire graph as in GCN. This framework permits the analysis to be scaled to massive graphs of

86 million nodes and 4 billion edges. Although the framework is not a methodology for default prediction, Shumovskaia et al. extend the scope of their research and apply it to a credit scoring problem. (Sukharev et al., 2020) propose a method to predict the default from a money transfer network and the historical information of transactions. To work with both datasets, they propose a methodology based on GCN and recurrent neural networks to handle network data and transactional data, respectively. As baseline models, they train a model with 7000 features; however, they achieve an increase of 0.4% AUC when comparing the proposed model with the best baseline model. Finally, (Roa, Correa-Bahnsen, et al., 2021) present a methodology for using alternative information in a credit scoring model. Models are estimated using data generated by an app-based marketplace. This information is precious for low-income segments and young individuals, who are often not assessed well by traditional credit scoring models. The authors compare a model with hand-crafted features versus models from GCNs. However, GCNs do not achieve better results than do hand-crafted features in terms of predictive power.

2.3. Representation Learning on Networks

The machine learning subfield that works on graph-structured data is known as graph representation learning or GRL (Hamilton, Ying, & Leskovec, 2017). Unlike the traditional tabular data, network data imposes a challenge to conventional machine learning algorithms because it is not possible to use them directly, forcing changes either on the algorithms or on the data representation. These challenges are required because the network information is, in essence, unstructured. In fact, operations that are easy to calculate on other data types, such as convolutions on images, cannot be applied directly to graphs because each node has a variable number of neighbors. Researchers have proposed many methodologies to extract knowledge for networks; here, we present a nomenclature and the characteristics of the most popular methods.

2.3.1. Feature Engineering

Data preparation is one of the most critical steps in any analytical project before training any machine learning model. Formulating accurate and relevant features is critical to improving model performance (Nargesian, Samulowitz, Khurana, Khalil, & Turaga, 2017). Regarding the use of graph data, the traditional feature engineering approach consists of characterizing each node either based on the aggregation of its neighborhood's features or the node's statistics within the network.

2.3.2. Network Embeddings

Network embedding methods are unsupervised learning techniques aiming to learn a Euclidean representation of networks in a much lower dimension. Each node is mapped into a Euclidean space through the optimization of similarity functions. The distance between network nodes in the new space is a surrogate for the node's closeness within the network structure. Node embedding techniques often replace feature engineering processes.

Formally, a network is represented by a graph $G(V, E, A)$ defined by a set of nodes V, a set of edges E, and an adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$. The embedding of a node is a function $f : G(V, E, A) \to \mathbb{R}^d$ that maps each node $v \in V$ to an embedding vector $\{Z_v\}_{v \in V} \in \mathbb{R}^d$ (Arsov & Mirceva, 2019), preserving the adjacency in the graph. The embedding vectors of pairs of nodes that are connected by an edge are closer than those that are disconnected. Let $Z \in \mathbb{R}^{|V| \times d}$ denote the node-embedding matrix, where $d \ll |V|$ for scalability purposes. The most popular network embedding method is Node2vec (Grover & Leskovec, 2016), which is an algorithmic framework for learning low-dimensional network representation. This algorithm maximizes the probability of preserving the neighborhood of the nodes in the embedding subspace. The algorithm optimizes using stochastic gradient descent, a network-based objective function, and produces samples for neighborhoods of nodes through second-order random walks. The key feature of Node2vec is the use of biased-random walks, providing a trade-off among two network search methods: breadth-first search (BFS) and depth-first search (DFS). This trade-off creates more informative network embeddings than other competing methods.

2.3.3. Graph Neural Networks (GNN)

Graph-structured data has arbitrary structures that can vary significantly between networks or within different nodes of the same network. Their support domain is not a uniformly discretized Euclidean space. For this reason, the convolution operator that is often used for signal processing cannot be directly applied to graph-structured data. Geometric deep learning (GDL) and graph neural networks (GNN) aim to modify, adapt and create deep learning techniques for non-Euclidean data. The proposed GDL computational schemes are an adaptation of deep autoencoders, convolutional networks, and recurrent networks to this particular data domain. In this study, we will be applying graph convolutional networks (GCN) and graph autoencoders (GAE).

2.3.3.1. Graph Convolutional Networks

The Graph convolutional networks generalize the convolution operation to networks formalized as graphs. The GCNs aim to produce a node's representation Z_v by adding its attributes or feature vector X_v and neighbors $\{X_u\}_{u \in N(v)}$, where $N(v)$ is the ego network of node n, that is, the subgraph composed of the nodes to whom node n is connected. This study uses the spectral-based GCN, also known as the Chebyshev spectral convolutional neural network, proposed by (Defferrard et al., 2016), which defines the graph convolution operator as a filter from graph signal processing. In particular, we use the specific GCN proposed by (Kipf & Welling, 2016a), which uses as a filter a first-order approximation of the Chebyshev polynomial of the eigenvalues' diagonal matrix. This graph convolution operator follows the expression:

$$X_i * g_\theta = \theta_0 X_i - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X_i, \tag{2.1}$$

where X_i is the feature vector, g_θ is a function of the eigenvalues of the normalized graph, Laplacian matrix $L = \mathcal{I}_n - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, A is the adjacency matrix, $D_{ij} = \sum_j A_{i,j}, \ \forall i, j \in V : i = j$, and θ is the vector of the Chebyshev coefficients. In the following, section we summarize the derivation of this first-order approximation (Kipf & Welling, 2016a).

2.3.3.2. Derivation of GCN from Spectral Methods

Spectral methods are founded on a solid theoretical basis defined for methods of graph signal processing developed essentially from the Laplacian matrix properties. To build up the graph convolution operator, we start from the normalized graph Laplacian matrix defined as follows:

$$L = \mathcal{I}_n - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{2.2}$$

where A is the adjacency matrix, and $D_{ij} = \sum_j A_{i,j}, \ \forall i, j \in V : i = j$. Because L is a real, symmetric, and positive semi-defined matrix, we can rewrite L as a function of its eigenvector matrix U and its eigenvalues λ_i, that is, $L = U \Lambda U^T$, where $U \in \mathbb{R}^{N \times N}$, and $\Lambda_{ij} = \lambda_i, \ \forall i, j \in V : i = j$.

The next step is to define the graph Fourier transform and its inverse. The graph Fourier transform $\mathcal{F}$ of the feature vector $X_i \in X$ is defined as follows:

$$\mathcal{F}(X_i) = U^T X_i, \tag{2.3}$$

where X is the matrix of node attributes, and the inverse Fourier transform of a graph is defined as follows:

$$\mathcal{F}^{-1}(\hat{X}_i) = U \hat{X}_i, \tag{2.4}$$

where $\hat{X}_i$ are the coordinates of the nodes in the new space. Therefore, the feature vector can be written as $X_i = \sum_{jinV} \hat{X}_i u_j$. Finally, the graph convolution of feature vector X_i with filter $g \in \mathbb{R}^N$, using the element-wise product $\odot$, is defined as follows:

$$X_i * g = \mathcal{F}^{-1}(\mathcal{F}(X_i) \odot \mathcal{F}(g)) \tag{2.5}$$

One of the most popular filters is the Chebyshev polynomial of the eigenvalues' diagonal matrix, that is, $g_\theta = diag(U^T g) = \sum_K \theta_k T_k(\hat{\Lambda})$, where $\hat{\Lambda} = 2\lambda/\lambda_{max} - \mathcal{I}$ and the polynomials T_k are defined as $T_k(x) = 2xT_{k-1} - T_{k-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$. Therefore, the GCN, defined as the Chebyshev Spectral CNN (Defferrard, Bresson, & Vandergheynst, 2016), takes the following form:

$$X_i * g_\theta = \sum_K \theta_k T_k(\hat{L}) X_i, \tag{2.6}$$

where $\hat{L} = 2L/\lambda_{max} - \mathcal{I}$. Despite being a graph convolution simplification, this convolution is computationally expensive for large graphs. To solve this problem, (Kipf & Welling, 2016a) present a first-order approximation of the Chebyshev Spectral CNN. Assuming $K = 1$ and $\lambda_{max} = 2$, the Equation 2.5 takes the following form:

$$X_i * g_\theta = \theta_0 X_i - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X_i \tag{2.7}$$

2.3.3.3. Graph Autoencoders (GAEs)

Graph autoencoders (GAEs) are an unsupervised method to obtain a low-dimensional representation of the network. The objective of the GAE is to reconstruct the original network using the same network for this task but encoding it, reducing its dimensionality, and then decoding it to reconstruct the network. The encoded representation is used as the network embedding. (Wu et al., 2020) distinguish two main uses of GAEs, namely graph generation and network embedding. This research will use GAEs to obtain a lower-dimensional vector, preserving the network topology (network embedding).

The previously defined GCN is the building block of the GAE architecture and allows the simultaneous encoding of the network topology and the attributes of the nodes. The GAE

(Kipf & Welling, 2016b) calculates the network embedding matrix Z and the reconstruction of the original network adjacency matrix $\hat{A}$ as follows:

$$\hat{A} = \sigma(ZZ^T), \ with \ Z = GCN(X, A), \tag{2.8}$$

where X is the matrix of node attributes, and A is the network's adjacency matrix.

2.4. Data Description

The data used in this paper encompasses several datasets provided by a large Latin American bank. Some datasets contain information from their customers, while others concern the entire population of the country.

2.4.1. Ethical and Privacy Protection Considerations

The datasets contain anonymized information and do not compromise the identity of any customer or their personal information in any way. The datasets share an anonymized customer's ID allowing us to merge multiple sources. Regarding the value, importance, and sensitivity of the data, we have applied multiple actions to ensure its security, integrity, and confidentiality. Customer identifiers and any personal data were removed before starting the analysis, and there is no possibility that this investigation can leak any personal private information. In addition, any final data produced as a result of this research does not compromise customers' privacy.

2.4.2. Social-Interaction Data

The information collected by the financial institution to construct the social network background information of the thin file borrower originates from varied sources and can be cataloged as follows:

- [**WeddNet**] **Network of marriages:** This network is built from the information of marriages recorded by the bureau of vital statistics from 1938 to December 2015. It includes the anonymized identifiers of the husband and the wife and the wedding date.
- [**TrxSNet**] **Transactional services network:** The primary source of this network comes from transactional services data, primarily payroll services and the transfers of funds between two entities. We have access to monthly data from January 2017 to December 2019.
- [**EnOwNet**] **Enterprise's ownership network:** This network is built from the information on companies' ownership structure. For each firm, we have information concerning their owners, be they individuals or other firms. We have quarterly information from January 2017 to November 2019.
- [**PChNet**] **Parents and children network:** This network corresponds to parental relationships. For every person born between January 1930 and June 2018, we have the anonymized identifiers of their parents.
- [**EmpNet**] **Employment network:** This network is built from multiple sources and connects people with their employers. We have monthly data from January 2017 to December 2019.

2.4.3. Financial Data

The node dataset contains information on the consolidated indebtedness of each debtor in the financial system from January 2018 until March 2020, reporting monthly the debt decomposition from 7.65 million people and 245,000 firms. We refer to the features extracted from this dataset as node features. Additionally, for every person and firm in the previous dataset, we have access to the BenchScore, which corresponds to the probability of default for the coming 12 months. This probability was assessed and provided by the financial institution, and it is our benchmark to contrast the performance of our models.

2.4.4. Network Construction

It is possible to build a network from each of the data sources indicated in Section 2.4.2. However, they share characteristics that allow them to be grouped. For this reason, we combine the networks into two primary data sources, from which we construct networks that characterize people and businesses.

[FamilyNet] Family network: This network is formed through the combination of the network of marriages (WeddNet) and the parent and children network (PChNet). For the construction of this network, we use the historical information available until the beginning of the analysis period; no further information is included. In this way, the network remains unvarying throughout the study. We call this type of network a static network.

[EOWNet] Enterprise's ownership Network and Workers: This network is composed of the fusion of the transactional services network (TrxSNet), the enterprise's ownership network (EnOwNet), and the employment network (EmpNet). This network attempts to represent the business ecosystem in which companies, business owners, and employees interact. Based on these data sources, a series of 24 networks are generated, one for each of the 24 months available, including the information collected up to the last day of the corresponding month. We call this type of network a temporal network, because nodes and edges change over time.

2.5. Experimental Design and Methodology

2.5.1. Datasets

The credit scoring models are built with information about the financial system for 24 months. However, the models are trained over 23 months because the first month is left out for the feature extraction process (presented in Section 2.5.3) in order to avoid target leakage (Kaufman, Rosset, Perlich, & Stitelman, 2012). For the unbanked application scoring model, individuals and companies are considered only in the month that they enter the financial system. In contrast, for the behavioral scoring model individuals and companies are considered six or more months after entering the financial system. Table 2.1 summarizes the scoring application, the model trained and the size of the dataset used.

2.5.2. Target

The target event was "becoming a defaulter during the period of observation". Therefore, we only took into account individuals or businesses that were non-defaulters at the start of the period of observation; we dismissed entities that were defaulters at the very beginning of

Tabla 2.1: Description of dataset

Scoring application	Model	Observations	# Features
Unbanked Application Scoring	Business Credit Score	29,044	687
	Personal Credit Score	192,942	1,283
Behavioral Scoring	Business Credit Score	931,910	687
	Personal Credit Score	1,978,664	1,283

the observation. In the current study, a person or company was considered a defaulter when they had payments past the due date for 90 or more days within 12 months starting from the observation point. Otherwise, they were considered non-defaulters. The target vector, denoted by y_{def}, contained the actual information about the target event.

2.5.3. Traditional and Graph Representation Learning Features

Combining the node information from Section 2.4.3 and the network data from Section 2.4.2 makes it possible to generate a set of new characteristics through a feature extraction process. The sets of characteristics generated are detailed below:

- [**NodeStats**] **Node Statistics:** This dataset collects node centrality statistics, namely, its degree, degree centrality, number of triads, PageRank score, authority and hub score given by the Hits algorithm (Kleinberg, 1999), and an indicator of whether the node is an articulation point.

- [**EgoNet**] **EgoNetwork Agreggation:** In this dataset, each node is characterized by the information of other nodes connected to it (ego network). We refer to the dataset as **egoNet aggregation features** when we apply some aggregation function to the characteristics of the nodes included in the ego network. Specifically, for each attribute in the **NodeStats** feature set we apply the mean and SD in this study as in (Nargesian et al., 2017; Roa, Correa-Bahnsen, et al., 2021). Figure 2.1 shows how this process would be within the network; the black node corresponds to the target node, and the gray nodes belong to its Ego Network. Computing the egoNet aggregation features assumes that each connection in the ego network has the same importance; however, connections in ego networks can be weighted. For this reason, we compute the **egoNet weighted aggregation features** where the features of the neighboring nodes are weighted according to a measure of the relationship intensity measured by the weighted average and SD of the **NodeStats** attributes.

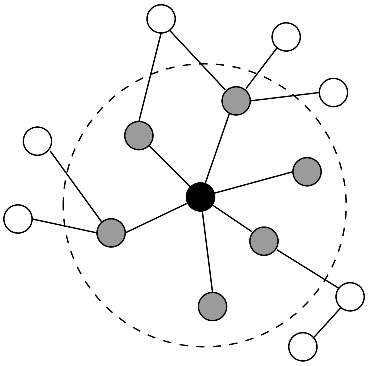


Figure 2.1: Example of Network Features

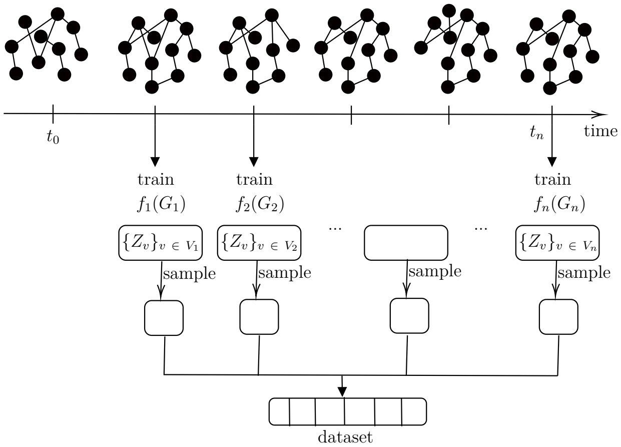


Figure 2.2: Node2Vec to Features

- **[N2V] Node2Vec Features:** Node2Vec is an unsupervised method that only uses the network structure to generate the graph embedding. For the static network **FamilyNet**, Node2Vec is applied only once. A node is characterized by this embedding regardless of the moment it was sampled in the dataset. For temporal networks (**EOWNet**), Node2Vec has to be recomputed every period because of the network changes. Each node is characterized by the embedding corresponding to the month in which it was sampled in the dataset. Figure 2.2 shows the process through which to obtain the embedding features by applying Node2Vec. Each period, a new model is trained, and the resulting embedding is consolidated to characterize the sample dataset that will allow us to train the final credit scoring models.

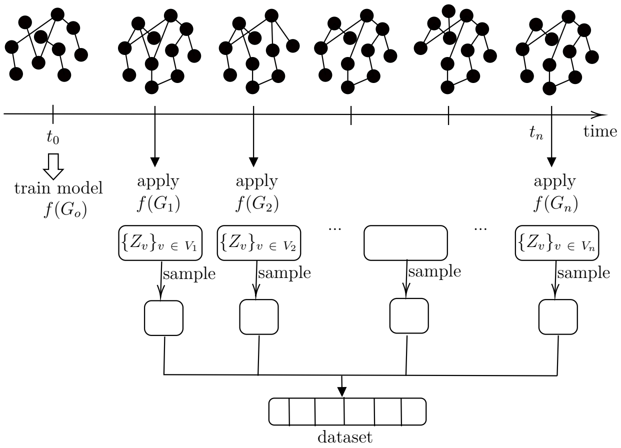


Figure 2.3: Graph Convolutional Networks and Graph Autoencoders to Features

- **[GNN] Graph Neural Network Features:** We used either GCN or GAE for the extraction of GNN features. These methods carry out the graph convolution of the network structure through a node feature vector in order to generate as output the feature vector for ensuing classification by machine learning approaches. In this study, GNN input feature vectors are in the **Node** feature dataset. Regardless of whether the network is static or temporal, the feature vector is dynamic and changes during each observation period, that is, each month. To avoid target leakage (Kaufman, Rosset, Perlich, & Stitelman, 2012), we train the GCNs with the first available data period; these models are applied in the subsequent months while the network or the feature vectors change. This approach does not handle new entrants to the social interaction networks; however, due to the sources of social interaction information employed in this study, most thin-file borrowers are taken into account in the training dataset, despite having temporal networks. Therefore, new entrants do not affect our results. Under other circumstances, for instance when working with a partial network, our recommendation is to train a new model for every period, taking precautions not to incur in target leakage on the population of the credit scoring model. Depending on the dynamism of the network, one solution is to calculate local connection updates as suggested in (Vlasselaer et al., 2015) in the interim between model training phases. The GCNs are trained on the entire network (either FamilyNet or EOWNet) regardless of whether the nodes belong to our training dataset or not. The output of the GCN is the label of the nodes in the network, which can be either defaulter, non-defaulter, or unbanked; therefore, each GCN solves a multi-classification problem by providing the a posteriori probabilities of each label for each node. Finally, each node is characterized by the GCN Features resulting from the application of the GCN on the network and on the feature vector of the month in which the node was sampled, that is, in which it entered the banking system. Figure 2.3 illustrates the feature-engineering process to extract the GCN features. Regarding GAE, the feature engineering process is similar. In this case, the embedding corresponds to the bottleneck hidden layer representing the encoder section's output. For consistency, we apply the same data selection applied for the GCN to the training of the GAE models, although their training is unsupervised.

2.5.4. Feature Subsets

The datasets for machine learning training and validation experiments are composed of the following subsets of features.

- Subset **A** : X_{Node} is the dataset of node characteristics.
- Subset **B** : $X_{BenchScore}$ is the benchmark score; this attribute is also used as a benchmark to quantify the performance of our proposed approach.
- Subset **C** : $X_{NodeStats}$ is composed of the statistics obtained from the position of the node within the network.
- Subset **D** : X_{EgoNet} includes the egoNet aggregation and the egoNet weighted aggregation features that are calculated in three scenarios, considering the entire network, considering only those edges that are bridges, and considering those edges that are not bridges[2].
- Subset **E** : $X_{GNN,N2V}$ corresponds to the features created by applying GNNs and Node2Vec. People are characterized by features from both networks (EOWNet and FamilyNet), while companies are characterized only by EOWNet.

For this study, we aggregated these feature subsets into eight increasingly larger datasets for training and validation, each defining a different experimental setting. The details of the feature sets are presented below in Table 2.2.

Tabla 2.2: Experiments Setup

Experiment Id	Feature Group
A	$X = \{X_{Node}\}$
A+B	$X = \{X_{Node} + X_{BenchScore}\}$
A+B+C	$X = \{X_{Node} + X_{BenchScore} + X_{NodeStats}\}$
A+B+D	$X = \{X_{Node} + X_{BenchScore} + X_{EgoNet}\}$
A+B+E	$X = \{X_{Node} + X_{BenchScore} + X_{GNN,N2V}\}$
A+B+C+D	$X = \{X_{Node} + X_{BenchScore} + X_{NodeStats} + X_{EgoNet}\}$
A+B+C+E	$X = \{X_{Node} + X_{BenchScore} + X_{NodeStats} + X_{GNN,N2V}\}$
A+B+C+D+E	$X = \{X_{Node} + X_{BenchScore} + X_{NodeStats} + X_{EgoNet} + X_{GNN,N2V}\}$

2.5.5. Evaluation Metrics

The AUC is a popular metric used to evaluate the model performance in classification problems (Zeng & Zeng, 2019). It ranges between 0.5 and 1. Values closer to 1 indicate a better discriminatory capacity, while a value of 0.5 indicates a performance equivalent to a chance decision. In the context of credit scoring, the AUC can be easily interpreted as follows: For a randomly selected defaulter and non-defaulter pair, the AUC corresponds to the probability that the classification model assigns a higher score to the defaulter.

[2] An edge connecting the nodes u and v is called a **bridge**; if removing this edge, there is no longer a path connecting u and v

Another extensively utilized performance measure is the Kolmogorov-Smirnov (KS) statistic, which measures the distance separating the cumulative distributions of defaulters ($P_D(t)$) and non-defaulters ($P_{ND}(t)$) (Fang & Chen, 2019). The KS statistic is defined as:

$$KS = \max_t |P_D(t) - P_{ND}(t)| \tag{2.9}$$

and KS ranges between 0 and 1, and a higher KS indicates a higher prediction performance.

2.5.6. Methodology

First, we carry out a *feature engineering* process that seeks to create attributes to characterize the nodes from the network. Then, in *the train-test split* step, the available dataset is divided into a training dataset of which 30% consists of the samples used to estimate the model's hyper-parameters, and the remaining 70% consists of the samples used to train and validate models according to an N-Fold Cross-Validation scheme.

Before estimating the best hyper-parameters, we apply a feature selection process. In this step, the intention is to choose a low-correlated subset of features with high predictive power. Three selection levels are formulated. The first is a bivariate selection that only considers one feature at a time and the target vector to build a prediction model, evaluating its performance. The selection process applies a threshold to this feature's predictive power. Only those variables are selected such that $KS > KS_{min}$ and $AUC > AUC_{min}$, where KS_{min} and AUC_{min} are threshold parameters.

Next, a multivariate selection is applied following a simple but effective method to drop correlated features that we have developed. This algorithm starts with an empty list $\mathcal{S}$. We iterate over a set of features $\mathcal{P}$ in decreasing order of predictive power and append to $\mathcal{S}$ those features whose absolute value of the correlation with each and all the features in $\mathcal{S}$ is less than a threshold ρ set to avoid high correlated features (Akoglu, 2018). The algorithm stops when all the features have been visited. The first feature in $\mathcal{P}$ is added to $\mathcal{S}$ without correlation comparison.

For this study, the multivariate selection process is applied twice. In the first application, the process selects low-correlated features for each group of attributes $\mathcal{P} \in \{X_{Node} \cup X_{BenchScore}, X_{NodeStats}, X_{EgoNet}, X_{GNN,N2V}\}$. Secondly, it is applied globally to all the remaining features $\mathcal{P} = \{X_{Node} \cup X_{BenchScore} \cup X_{NodeStats} \cup X_{EgoNet} \cup X_{GNN,N2V}\}$. In both cases, a threshold ρ is used, and the features are ordered by the features' AUC, from higher to lower.

Finally, at the N-Fold Cross-Validation stage, the dataset is partitioned into N subsets of equal size. Each subset is used alternatively as the test dataset, while the remaining folds are used to train the classification model. The hyper-parameters used in each iteration are those estimated in the previous stage. Additionally, in each of these iterations, multiple models are trained with different feature sets and stored to be used later to compare the models.

2.5.7. Experimental Setup

The parameters of the univariate selection are set at $KS_{min} = 0.01$ and $AUC_{min} = 0.53$; for the multivariate selection process, $\rho = 0.7$ in both processes to avoid high correlated features (Akoglu, 2018). The N-Fold Cross-Validation stage is carried out considering $N = 10$, and in each iteration, the results of gradient boosting (Friedman, 2001) models are displayed. Other classification models such as regularized logistic regression and Random Forest (Breiman, 2001) were trained. However, gradient boosting consistently delivered better results.

2.6. Results and Discussion

In this section, we present the results obtained. We begin with the technical implementation details; then, we analyze the execution times. Subsequently, we display the model's performance in three scenarios: the impact on performance using traditional features, the impact on performance using the different graph representation methods, and the advantages of combining these methods. Finally, an analysis is presented of the main features, traditional and network-based, for the creditworthiness assessment.

2.6.1. Implementation Details

In this work, we used the Python implementations Networkx v2.6.3 (Hagberg, Swart, & SChult, 2008) and Stanford Network Analysis Platform (SNAP) v5.0.0 (Leskovec & Sosič, 2016) in the hand-crafted feature engineering process ($X_{NodeStats}$, X_{EgoNet}); for Node2Vec, GCN and GAE ($X_{GNN,N2V}$) we used PyTorch v1.6.0 (Paszke et al., 2019) and PyTorch Geometric v2.0.1 (Fey & Lenssen, 2019).

To conduct the experiments, we used a laptop with an Intel 8-Core i7 CPU and 32 GB of RAM for network construction and hand-crafted feature engineering. For the Node2Vec, GCN, GAE, and model training phases, we used a server with a driver node with 140 GB of RAM and 20 CPU cores and between two and eight auto-scaling worker nodes with 112GB of RAM and 16 CPU cores.

2.6.2. Execution Time

Below we detail the execution time of the most critical stages of our work, the implementation of the GRL methods, and the models' training.

- [NodeStats] **Node Statistics:** This process corresponds to the computation of the metrics defined in Section 2.5.3. This process is carried out only once for the static network FamilyNet, and it is calculated for all the available periods (24) of the EOWNet. The computation of all the metrics for a network takes, on average, 25 minutes. The total execution time of this stage was 625 minutes.

- [EgoNet] **EgoNetwork Aggregation:** This process is calculated once per network type (FamilyNet and EOWNet). The total execution time of this stage was 300 minutes.

- [N2V] **Node2Vec Features:** This process is carried out only once for the static network FamilyNet, and it is calculated for all the available periods (24) of the EOWNet. The Node2Vec training for a network takes, on average, 300 minutes. The total execution time of this stage was 7,500 minutes.

- [GNN] **Graph Neural Network Features:** In this stage, eight models are trained using each network and eight different feature vectors. The training of each model is carried out over the first available period data; afterward, the trained model is applied to the data of the remaining periods. Table 2.3 shows the execution time by GNN type and by network type. The total execution time of this stage was 6,920 minutes.

- **Gradient boosting training:** Four models were trained using the methodology described in Section 2.5.6 for the scenarios defined in Section 2.4, predicting application and behavioral scoring for individuals and companies. The complete training for each

Tabla 2.3: GCN and GAE Execution Time

GNN Type	Network Type	Unit Train Time (min)	Total Training Time (min)	Total Apply Time (min)
GCN	EOWNet	25	200	600
	FamilyNet	60	480	600
GAE	EOWNet	140	1,120	1,200
	FamilyNet	190	1.520	1.200
Total Execution Time (min)			3,320	3,600

scenario takes, on average, 40 minutes. The total execution time of this step was 160 minutes.

The total execution time of the application of the proposed methodology to the datasets is 15,500 minutes. Although a large part of these executions was parallelized using the server described in Section 2.6.1, the high computational cost is mainly due to two factors, the volume of data and the complexity of the algorithms. Regarding the large volume of data, the FamilyNet has 20 million nodes and 30 million edges, and the EOWNet has 8.6 million nodes and 26 million edges. These massive dataset sizes directly impact the algorithms used, because the complexity depends on the nodes $|V|$, edges $|E|$ and the embedding dimension d; the algorithmic complexity for this particular case is the following: Node2Vec: $\mathcal{O}(|V|d)$ (Grover & Leskovec, 2016), Graph Convolutional Networks: $\mathcal{O}(|E|d)$ (Kipf & Welling, 2016a) and Graph Autoencoder: $\mathcal{O}(|V|^2d)$ (Kipf & Welling, 2016b).

2.6.3. Model Performance Results

The model training process was conducted for each Experiment ID using different Feature Set combinations as specified in Section 2.5.1. The results are presented in Tables 2.4 and 2.5. These tables show the relative improvement in AUC and KS achieved by each model over the baseline BenchScore, measured as $\frac{row_{AUC}-BenchScore_{AUC}}{BenchScore_{AUC}}$ and $\frac{row_{KS}-BenchScore_{KS}}{BenchScore_{KS}}$, respectively. Each row corresponds to an experiment defined in Section 2.5.4, and each column displays the four credit scoring scenarios illustrated in Section 2.5.1.

The reported results correspond to the average of 10-fold cross-validation, as indicated in Section 2.5. A t-test is applied to establish the statistical significance of the performance differences obtained using the different feature sets, according to the suggestions proposed by (Flach, 2012).

2.6.3.1. Model Performance Using Traditional Features

The first step toward reaching a conclusion on the contribution of network data is to understand whether our methodology enables us to obtain equal or better results than the current decision-making scheme in the financial institution. For this, we compare the BenchScore with the results obtained by the feature set **A+B**. A comparison with only the feature set **A** is not entirely accurate, considering that we do not have access to all the features used in the BenchScore training.

The results show that our methodology obtains equal or greater performance, measured in terms of AUC and KS statistics, in all four scenarios; three of them are greater, with statistically significant differences.

The performance enhancements in Behavioral Business Credit Scoring are 0.58% and

Tabla 2.4: Improvement in AUC relative to the benchmark model (mean and std). We only report results when the equal performance hypothesis is rejected, with a confidence level of 95%; otherwise, we display *. The best performance in each column is shown in bold; more than one bold value indicates that the hypothesis of equal performance between those models cannot be rejected.

Feature Set	Business Credit Score		Personal Credit Score	
	Application	Behavioral	Application	Behavioral
A	-3.52% ± 2.87%	-0.90% ± 0.21%	-0.74% ± 0.63%	-0.63% ± 0.09%
A+B	*	0.58% ± 0.06%	1.45% ± 0.39%	0.95% ± 0.06%
A+B+C	*	1.13% ± 0.12%	2.02% ± 0.49%	1.07% ± 0.06%
A+B+D	**8.96**% ± 3.37%	2.33% ± 0.15%	2.31% ± 0.64%	1.25% ± 0.08%
A+B+E	3.92% ± 2.03%	1.77% ± 0.13%	3.17% ± 0.55%	1.96% ± 0.04%
A+B+C+D	**9.00**% ± 3.47%	2.37% ± 0.16%	2.39% ± 0.60%	1.32% ± 0.08%
A+B+C+E	4.25% ± 1.84%	1.94% ± 0.16%	3.26% ± 0.48%	2.03% ± 0.05%
A+B+C+D+E	**8.43**% ± 2.83%	**2.80**% ± 0.16%	**3.58**% ± 0.61%	**2.18**% ± 0.04%

Tabla 2.5: Improvement in KS relative to the benchmark model (mean and std). We only report results when the equal performance hypothesis is rejected, with a confidence level of 95%; otherwise, we display *. The best performance in each column is shown in bold; more than one bold value indicates that the hypothesis of equal performance between those models cannot be rejected.

Feature Set	Business Credit Score		Personal Credit Score	
	Application	Behavioral	Application	Behavioral
A	*	-4.15% ± 0.94%	-5.25% ± 2.40%	-2.39% ± 0.46%
A+B	*	1.56% ± 0.40%	4.38% ± 1.19%	1.95% ± 0.35%
A+B+C	*	3.21% ± 0.71%	6.27% ± 1.02%	2.23% ± 0.39%
A+B+D	**20.69**% ± 16.73%	7.69% ± 0.92%	6.79% ± 1.36%	2.69% ± 0.47%
A+B+E	**12.22**% ± 10.89%	5.83% ± 0.74%	8.64% ± 2.13%	4.68% ± 0.28%
A+B+C+D	**21.28**% ± 17.10%	8.09% ± 0.95%	7.12% ± 1.52%	2.83% ± 0.52%
A+B+C+E	**12.88**% ± 10.11%	6.33% ± 0.70%	8.93% ± 1.98%	4.93% ± 0.26%
A+B+C+D+E	**19.32**% ± 14.77%	**9.45**% ± 0.85%	**10.83**% ± 1.98%	**5.15**% ± 0.42%

1.56% for AUC and KS, respectively. In terms of Personal Credit Scoring, the performance enhancements are AUC: 1.45%, KS: 4.38% for Application Scoring and AUC: 0.95%, KS: 1.95% for Behavioral Scoring. These results indicate that using our methodology and training a model with similar features to the benchmark model can obtain better results than the current decision scheme applied by the financial institution.

2.6.3.2. Model Performance Using Graph Representation Learning Features

In Application Business Credit Scoring, a nearly 9% AUC increase over the BenchScore is achieved. These results are obtained by a model that incorporates three feature sets: A+B+D, A+B+C+D, and A+B+C+D+E. Note that in these three sets, the common attributes correspond to the traditional features **A+B** : $X_{Node} \cup X_{BenchScore}$ and EgoNet Aggregation Features **D** : X_{EgoNet}. The performance comparison between all feature sets is shown in Table 2.6; we marked as * those comparisons with no statistically significant differences.

When performance is measured in terms of KS, the maximum is obtained with five feature sets, with GRL features, but no method for feature extraction predominates. Although differences are observed in the values presented, these are not statistically significant. From these results, it is necessary to highlight at least one GRL method in the best feature sets. The complete comparison is presented in Table 2.7

Tabla 2.6: Performance comparison of Business Application Scoring. Performance is measured by the relative increase in AUC ($\frac{row_{AUC} - column_{AUC}}{column_{AUC}}$).

	BENCH	A	A+B	A+B+C	A+B+D	A+B+E	A+B+C+D	A+B+C+E	A+B+C+D+E
BENCH	*	3.65%	*	*	-8.23%	-3.77%	-8.26%	-4.08%	-7.77%
A	-3.52%	*	-2.92%	-4.55%	-11.45%	-7.16%	-11.49%	-7.45%	-11.01%
A+B	*	3.01%	*	-1.68%	-8.79%	-4.37%	-8.83%	-4.67%	-8.34%
A+B+C	*	4.77%	1.71%	*	-7.23%	-2.73%	-7.26%	-3.04%	-6.77%
A+B+D	8.96%	12.94%	9.64%	7.79%	*	4.85%	*	4.52%	*
A+B+E	3.92%	7.71%	4.57%	2.81%	-4.63%	*	-4.66%	*	-4.15%
A+B+C+D	9.00%	12.98%	9.68%	7.83%		4.89%	*	4.56%	*
A+B+C+E	4.25%	8.05%	4.90%	3.13%	-4.32%	*	-4.36%	*	-3.85%
A+B+C+D+E	8.43%	12.38%	9.10%	7.26%	*	4.33%	*	4.00%	*

Tabla 2.7: Performance comparison of Business Application Scoring. Performance is measured by the relative increase in KS ($\frac{row_{KS} - column_{KS}}{column_{KS}}$).

	BENCH	A	A+B	A+B+C	A+B+D	A+B+E	A+B+C+D	A+B+C+E	A+B+C+D+E
BENCH	*	*	*	*	-17.14%	-10.89%	-17.55%	-11.41%	-16.19%
A	*	*	*	*	-19.94%	-13.90%	-20.34%	-14.40%	-19.02%
A+B	*	*	*	*	-16.16%	-9.83%	-16.57%	-10.36%	-15.20%
A+B+C	*	*	*	*	-16.02%	-9.68%	-16.43%	-10.21%	-15.06%
A+B+D	20.69%	24.91%	19.28%	19.08%	*	*	*	*	*
A+B+E	12.22%	16.14%	10.90%	10.72%	*	*	*	*	*
A+B+C+D	21.28%	25.53%	19.86%	19.67%	*	*	*	*	*
A+B+C+E	12.88%	16.83%	11.56%	11.37%	*	*	*	*	*
A+B+C+D+E	19.32%	23.49%	17.92%	17.73%	*	*	*	*	*

The best performance is observed in other scenarios when combining traditional features and all the GRL features; this corresponds to the Feature Set **A+B+C+D+E**. The best performance is achieved in AUC (see Table 2.4) and KS (see Table 2.5).

These results are of great importance because they indicate that the methods combined by our methodology are complementary, and none is significantly better than the others. Both methods, namely hand-crafted feature engineering and GNNs, have until now been treated in the literature as independent in addressing the credit scoring problem.

When comparing the results of Application and Behavioral Credit Scoring, it is observed that the most significant increase in performance, regardless of the metric, is achieved in Application Credit Scoring. Network-related features complement the least availability of information, such that the relationships that a person or company has are relevant when predicting their creditworthiness. These results are of high interest for lenders and in terms of their strategies for the unbanked. The improvement in predictive performance implies that more borrowers can be serviced without increasing the portfolio default rate.

Regarding Behavioral Credit Scoring, traditional attributes are already good predictors of creditworthiness; the borrower's credit behavior is a good indicator of default. For this reason, the increase in predictive performance is more limited, although still significant.

2.6.3.3. The Advantages of Blending Graph Representation Learning

The previous sections have shown that our approach allows us to enhance the discrimination power of our benchmark in terms of AUC and KS. Through the incorporation of the graph data by means of the GRL methods, this increase is even more significant. Now, we are interested in discovering the contribution of each of these methods. The performance comparison between the A+B+C+D+E feature set and each method by itself is shown in Tables 2.8 and 2.9, for AUC and KS respectively; we marked as * those comparisons with no statistically significant differences. In each table, the results are presented for each credit scoring scenario and the comparison using the X_{EgoNet} (A + B + D) and $X_{GNN,N2V}$ (A + B + E) features; In both cases, the models trained with the $X_{NodeStats}$ features are also included.

Tabla 2.8: Blended Graph Representation Learning performance. The performance enhancement of training a model using all GRL methods (A+B+C+D+E) is measured as the relative increase in AUC given by $(\frac{[A+B+C+D+E]_{AUC}-column_{AUC}}{column_{AUC}})$.

Scoring	Model	Feature Set			
		A+B+D	A+B+C+D	A+B+E	A+B+C+E
Application Scoring	Business Credit Score	*	*	4.33%	4.00%
	Personal Credit Score	1.23%	1.16%	0.39%	0.31%
Behavioral Scoring	Business Credit Score	0.47%	0.43%	1.02%	0.85%
	Personal Credit Score	0.92%	0.84%	0.22%	0.15%

Tabla 2.9: Blended Graph Representation Learning performance. The performance enhancement of training a model using all GRL methods (A+B+C+D+E) is measured as the relative increase in KS $(\frac{[A+B+C+D+E]_{KS}-column_{KS}}{column_{KS}})$.

Scoring	Model	Feature Set			
		A+B+D	A+B+C+D	A+B+E	A+B+C+E
Application Scoring	Business Credit Score	*	*	*	*
	Personal Credit Score	3.79%	3.47%	2.02%	1.75%
Behavioral Scoring	Business Credit Score	1.68%	1.31%	3.47%	2.99%
	Personal Credit Score	2.40%	2.26%	0.45%	0.21%

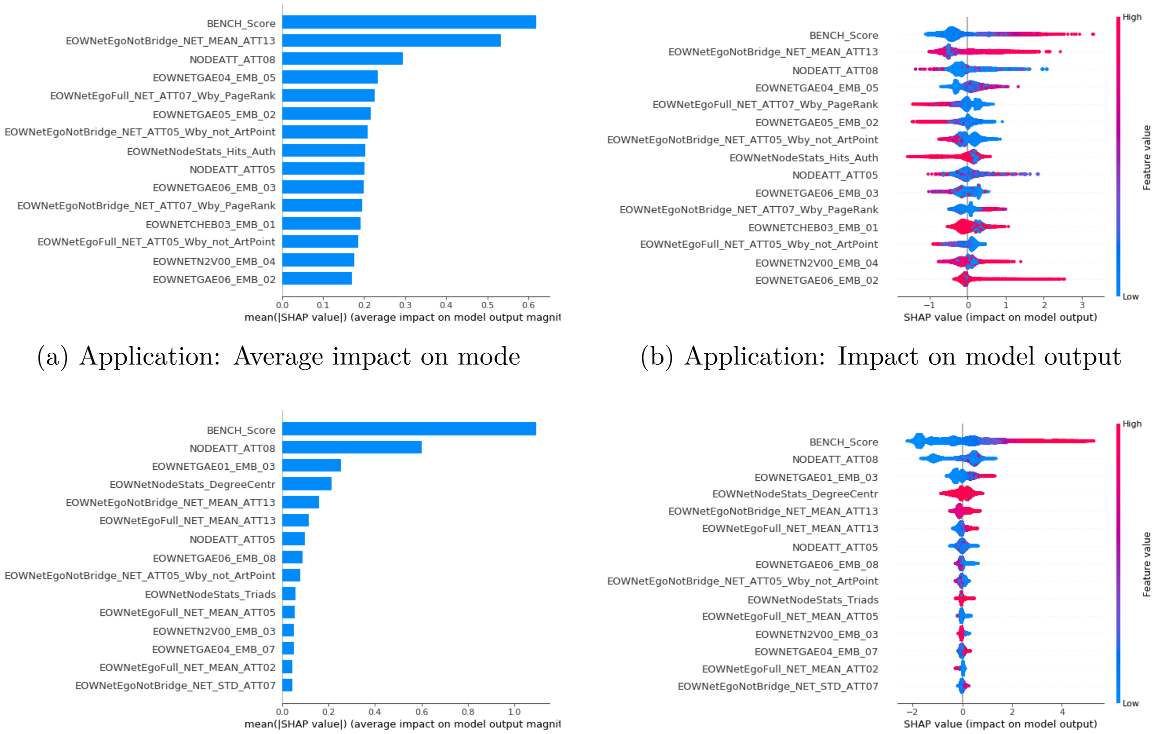
The results show that combining the GRL methods always generates better or similar results than using each method independently. An equal performance is only obtained for the Business Application Credit Score, where the only statistically significant increase, in AUC terms, occurs when using the $X_{GNN,N2V}$ features. However, this feature subset does not produce an increment compared to using only the X_{EgoNet} features. On the other hand, in all other scenarios, the GRL combination generates statistically significant increments, independent of the method used and whether or not the $X_{NodeStats}$ features are incorporated. In this way, our approach allows us to increase discriminatory power in assessing creditworthiness, generate more accurate models, and use graph data better through a framework that combines multiple methods of GRL.

2.6.4. Feature Importance Analysis

To determine the importance of each feature, we utilize SHAP: SHapley Additive exPlanations (Lundberg & Lee, 2017), an approach based on game theory that calculates each attribute's importance by comparing the model predictions with and without the attribute. The global feature importance is examined with regard to the four scenarios described earlier, that is, the prediction of Application or Behavioral Scoring for individuals or for businesses. All analyses are conducted with the feature set **A+B+C+D+E**, which incorporates all the features and is the one that reports the best results.

2.6.4.1. Business Credit Scoring

In Figure 2.4, the importance of the attributes incorporated into the model is presented. Figures 2.4(a) and 2.4(c) show each attribute's importance for Application and Behavioral Scoring, respectively; we define it as the average of the absolute values of the SHAP values. Figures 2.4(b) and 2.4(d) show each feature's impact on the model output; only the 15 most relevant attributes are displayed in both figures. Figures 2.4(b) and 2.4(d) allow us to understand how the value of a particular feature affects the probability of default.



(a) Application: Average impact on mode

(b) Application: Impact on model output

(c) Behavioral: Average impact on mode

(d) Behavioral: Impact on model output

Figure 2.4: Business Credit Scoring: Feature Importance

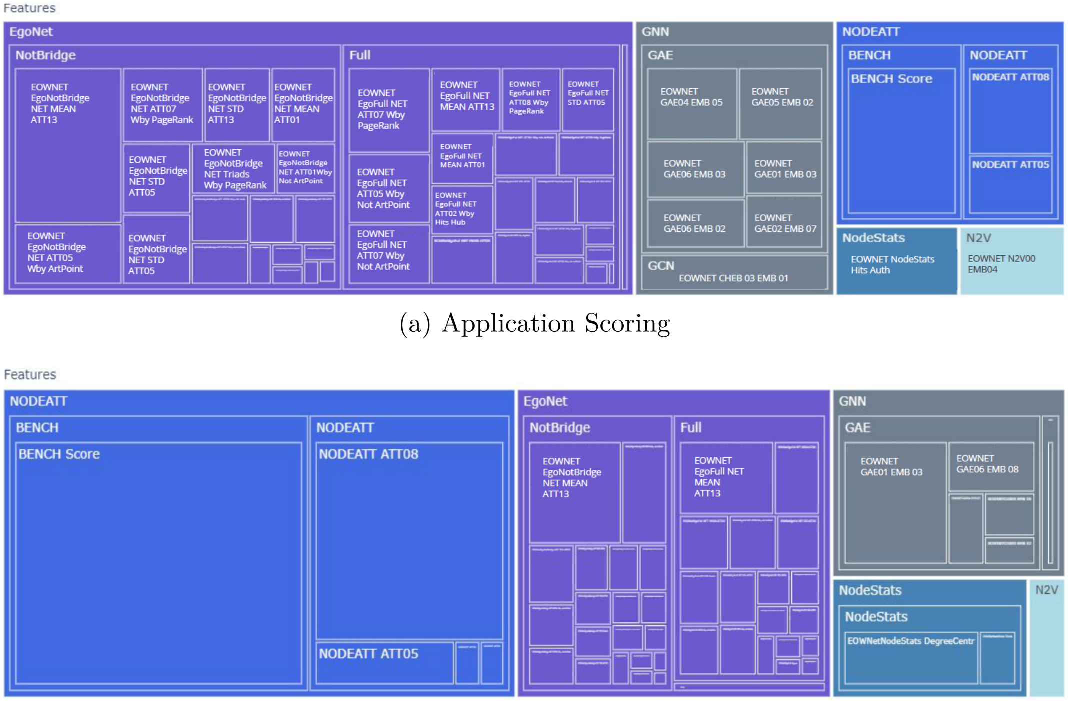
As expected, the most significant contribution to the model is the BenchScore, which already summarizes valuable information about each company that allows the estimation of its creditworthiness. This influence occurs in both scenarios, Application, and Behavioral. However, its importance is more significant in Behavioral Scoring.

Among these 15 relevant attributes in both scenarios, only the BenchScore, commercial debt amount (NODEATT_08), and unused revolving credit amount (NODEATT_05) correspond to the business-related characteristics. The remaining top features correspond to Network-related features.

An additional relevant feature is the average BenchScore of the company's ego network, including only the non-bridge edges. This result indicates the creditworthiness of the company's neighborhood is also highly predictive of the company's creditworthiness. In Application Scoring, this feature is practically as relevant as the BenchScore. See Table 2.10 for more detail on the description of the most relevant variables. This table presents the taxonomy of the features used in the current study, giving the necessary specifications for the correct interpretation of the feature attributes and the nomenclature used for the management of the datasets.

Further, we find attributes whose influence corresponds to people related to the company, including its owners, for instance, the attribute generated from a Graph Autoencoder trained with the consumer debt of the EOWNet Network. The consumer-debt effect of the ego network is also observed in the attribute corresponding to the consumer debt weighted by the PageRank of the node's neighborhood. The presence of consumer debt as a relevant network-related feature in Business Scoring is highly significant, especially in SMEs. The EgoNet's short-term personal debt, mainly on the part of the owners, accounts for the often blurred separation between personal finances and company finances. The owner's default can affect the company and vice versa. This hypothesis requires a more detailed investigation and will be addressed for future work.

To quantify the usefulness, impact and importance of the different feature sets on the output model, Figure 2.5 presents a Treemap based on the average of the absolute values of the SHAP values; the complete list of the model features is displayed, and the different colors indicate that they belong to different feature sets.



(a) Application Scoring

(b) Behavioral Scoring

Figure 2.5: Business Credit Scoring: Treemap of Feature Importance, the Average Impact on Model Output

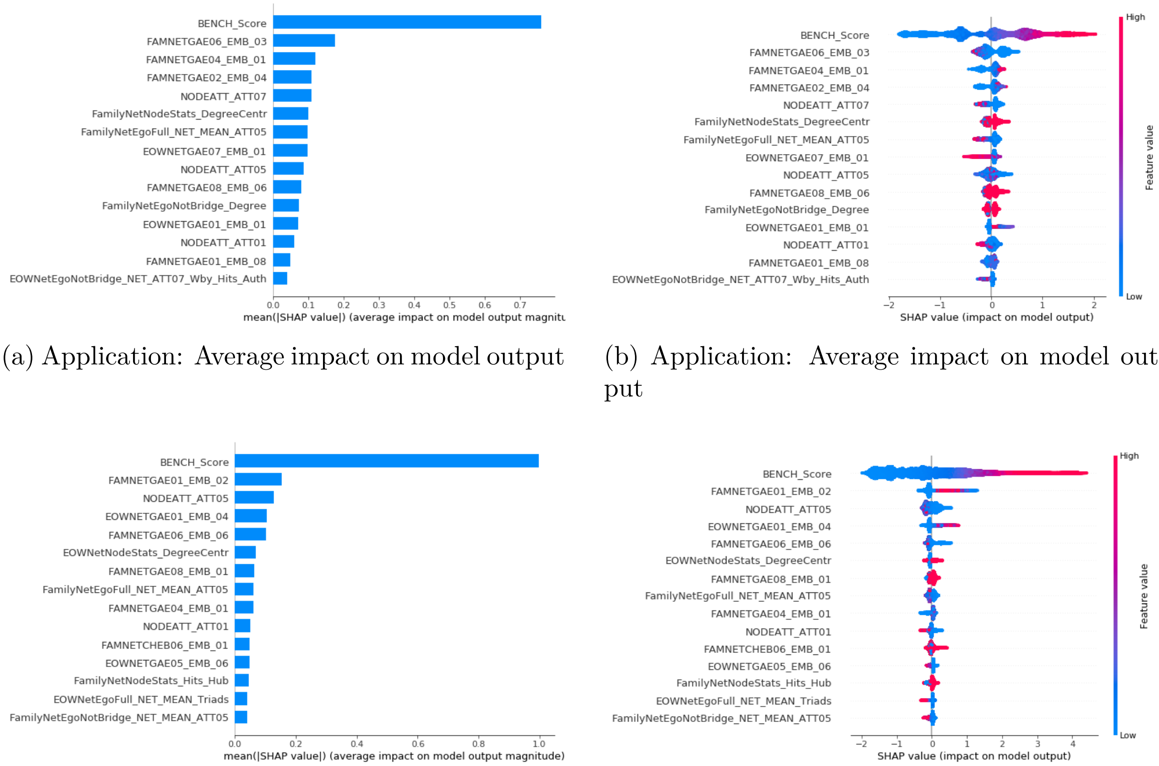
In Figure 2.5(a), it is shown that the feature set X_{EgoNet} (**D**) contributes, in Application Scoring, 60% of the model's overall impact. The feature set $X_{GNN,N2V}$ (**E**) contributes 21%, of which 19% correspond to GNN features, while 2% correspond to Node2Vec. The low importance of Node2Vec features is likely the reason for the limited research on Node2Vec to enhance the prediction of creditworthiness.

However, in Business Behavioral Scoring, the traditional characteristics now represent 48 % of the total impact of the model. In contrast, in Business Application Scoring, they represent only 16% (See Figure 2.5(b)). Indeed, the BenchScore attribute alone represents 29% of the total impact. The feature set X_{EgoNet} (**D**) represents 30% of the total impact, the average BenchScore of the ego network being the most relevant attribute.

2.6.4.2. Personal Credit Scoring

In Personal Scoring, the person's characteristics ($X_{Node} + X_{BenchScore}$) produce a more meaningful impact than the business score. The person's attributes represent 37% and 47% of the total impact for Application and Behavioral Scoring, respectively.

Besides the BenchScore, other relevant features are the amount of consumer debt (NODEATT_07) and amount of unused revolving credit, and total debt amount (NODEATT_01) (see Figures 2.6(a) and 2.6(c)).



(a) Application: Average impact on model output

(b) Application: Average impact on model output

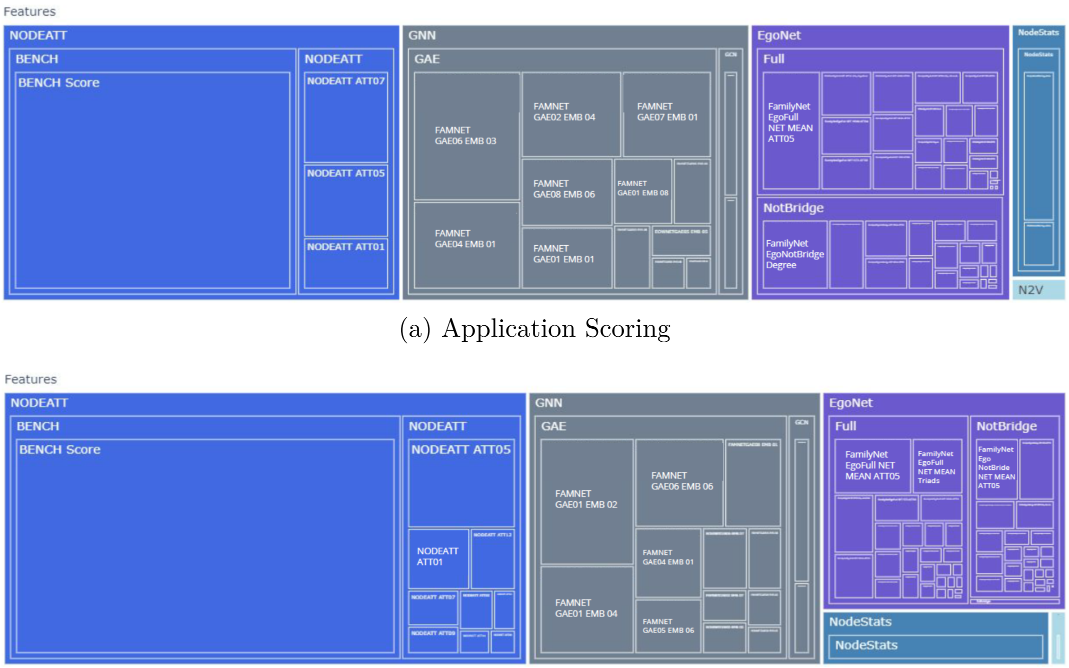
(c) Behavioral: Average impact on model output

(d) Behavioral: Average impact on model output

Figure 2.6: Personal Credit Scoring: Feature Importance

The combined network features also play an essential role in the final score; the feature sets X_{EgoNet} (**D**) and $X_{GNN,N2V}$ (**E**) represent 25% and 33.4% in Application Scoring (See Figure 2.7(a)), while the impact in Behavioral Scoring (See Figure 2.7(b)) are 18% and 28.3% respectively. In both cases, the contribution of Node2Vec features is negligible. The network feature with the highest impact is, as the average neighborhood's amount, the network's amount of unused revolving credit.

Personal Credit Scoring includes attributes generated with both FamilyNet and EOWNet networks. When analyzing the network-related features, almost all of them, in both scenarios, are FamilyNet features. These results show us both the suitability of the network used to characterize borrowers and the importance of the type of relationship used to build the network. In this study, family ties are the most appropriate to characterize borrowers as regards the problem of individual credit scoring.



(a) Application Scoring

(b) Behavioral Scoring

Figure 2.7: Personal Credit Scoring: Treemap of Feature Importance, the average impact on model output

2.7. Conclusions

This study presents an information processing methodology that allows us to assess the additional value of social-interaction data to approach the credit scoring problem for thin-file clients. This framework is applied in four scenarios arising from the consideration of all combinations of Application and Behavioral Scoring of individual and business lending. Additionally, this methodology allows the evaluation of different GRL approaches to feature extraction from social networks: hand-crafted feature engineering, Node2Vec, and Graph Neural Networks. The results show an improvement in creditworthiness assessment performance when different GRL approaches are combined. Specifically, two of the three GRL methods significantly enhance credit scoring models, namely the Hand-crafted Feature Engineering and Graph Neural Networks, which have the greatest impact when used together. We believe this to be very relevant to the community because, until now, these two methods have been used independently. On the other hand, we have found that the contribution of Node2Vec is negligible. This result seems to justify the limited research conducted on Node2Vec as a feature-engineering method for credit scoring.

As a baseline, we use a credit scoring model developed by a financial institution. This model, the BenchScore, already outperforms the credit bureau model they obtain from the credit bureau offices, and our methodology allows us to obtain better results in each of the four scenarios.

The highest value of the proposed approach is found in Unbanked Application Scoring. The unbanked applicants, individuals, and companies, lack behavioral information, which, as it turns out, is one of the best predictors of creditworthiness. Our approach overcomes the lack of behavioral information and delivers a proper credit risk assessment using graph data. In this way, applicants have greater access to the financial system. In the case of the

Behavior Scoring models, our methodology also improves performance. In both cases, the maximum improvement in predictive performance is achieved when these GRL methods are used together.

Explanatory measures, such as SHAP values, allow us to understand each attribute's contribution. If the impact on the output model is measured in this way, the baseline model (BenchScore), although it continues to be an essential attribute, has a diminished effect because it is in the presence of other good predictors. This feature importance analysis allows us to understand that we cannot solely examine a company's characteristics to evaluate the company, especially if it is unbanked. We also have to understand that they are part of an ecosystem in which the owners, suppliers, clients, and related companies are essential. The business ecosystem information allows us to improve the creditworthiness assessment performance. A similar situation occurs in Personal Credit Scoring, although with less intensity. The network data allows us to address the scarcity of information and achieve a better credit risk assessment.

Our research shows that there is still room for improvement in incorporating network information into the credit scoring problem. This methodology goes in the right direction, improving the performance of creditworthiness assessment, and it has great value for unbanked and under-banked people and even in the management of portfolio's credit risk.

Acknowledgments

This work would not have been accomplished without the financial support of CONICYT-PFCHA / DOCTORADO BECAS CHILE / 2019-21190345. The second author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery Grant RGPIN-2020-07114]. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. The last author thanks the partial support of FEDER funds through the MINECO project TIN2017-85827-P and the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 777720.

Appendix A

A1 Feature Description

Tabla 2.10: Taxonomy of features used in the experiments and nomenclature that we used for the management of the data in the experiments.

Feature Set	Nomenclature	Prefix/suffix	Description
Node Features (X_{Node})	Identifier	$NODEATT$	Feature subset identifier
	Borrower feature identifier	$ATT01, \cdots, ATT04$	The debt situation characterized by the delinquency level
		$ATT05, \cdots, ATT08$	The debt type: revolving, consumer, commercial, or mortgage
		$ATT09, \cdots, ATT13$	Other aspects of the customer's debt, payments in arrears, and the time in the financial system
		$Bench\_Score$	The benchmark score
Node Statistics ($X_{NodeStats}$)	Identifier	NodeStats	Feature subset identifier
	Statistic identifier	$DegreeCentr$	Degree centrality
		$Triads$	Number of triads
		$PageRank$	PageRank Algorithm
		$ArtPoint$	Articulation point
		$Hits\_Auth$	Hits algorithm Authority score
		$Hits\_Hub$	Hits algorithm Hub score
	Network identifier	EOWNET	EOWNet Network
		FamilyNet	Family Network
EgoNetwork Agreggation Features (X_{EgoNet})	Borrower feature identifier	$ATT01, \cdots, ATT13$	Borrower Feature
	Network identifier	EOWNET	EOWNet Network
		FamilyNet	Family Network
	Aggregation Function	MEAN	Mean
		STD	Standard Deviation
	Edges	Full	All edges
		NotBridge	Edges that are not bridges
		IsBridge	Edges that are bridges
	Weighted Aggregations	$Wby + Feature$	Suffix for weighted aggregations
Node2Vec Features (X_{EgoNet})	Identifier	N2V	Feature subset identifier
	Embedding Identifier	$EMB\_01, \cdots, EMB\_08$	The embedding number
	Network identifier	EOWNET	EOWNet Network
		FamilyNet	Family Network
Graph Neural Network Features (X_{GNN})	GNN Identifier	CHEB	Graph Convolutional Network (GCN)
		GAE	Graph Autoencoder
	Borrower feature identifier	$01, \cdots, 13$	Borrower Feature
	Embedding Identifier	$EMB\_01, \cdots, EMB\_n$	The embedding number, where $n = 3$ for CHEB and $n = 8$ for GAE
	Network identifier	EOWNET	EOWNet Network
		FAMNET	Family Network

Chapter 3

On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance[3]

Abstract

For more than a half-century, credit risk management has used credit scoring models at each of the well-defined stages of credit risk management. Application scoring is used to decide whether to grant a loan or not, while behavioral scoring is used mainly for portfolio management and to take preventive actions in case of default signals. In both cases, social network data has recently been shown to be valuable to increase the predictive power of these models, especially when the borrower's historical data is scarce or not available. This study aims to understand the dynamics of creditworthiness assessment performance and how it is influenced by credit history, repayment behavior, and social network features. To accomplish this, we build up a machine learning classification framework demonstrating its value analyzing 97,000 individuals and companies from the moment they obtained their first loan up to 12 months afterward. Our original and massive dataset allowed us to characterize each borrower according to its credit behavior, and socioeconomic relationships. Our study finds that credit scoring based on borrowers' history improves performance at a decreasing rate during the first six months and then stabilizes. The most notable effect on the performance of credit scoring based on social network features occurs in loan applications; for personal scoring, this effect prevails for approximately six months, while for business scoring, social network features add value throughout the entire study period. These findings are of great value to improve credit risk management and optimize the combined use of both the traditionally exploited information and new alternative data sources.

Keywords: behavioral credit scoring; application credit scoring; machine learning; social network data

[3] The following is a copy of the paper published at the Expert Systems with Applications Journal. Please cite this paper as follows: Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña, On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance, Expert Systems with Applications, Volume 218, 2023, 119599, ISSN 0957-4174

3.1. Introduction

Financial institutions operate in a complex and dynamic environment, in which they are exposed to multiple risk sources, with credit risk being the most significant (Apostolik, Donohue, & Went, 2009). Credit risk management is vital and should be part of each lending decision; adequate risk management helps avoid financial losses and is a crucial element for the profitability and well-being of the financial institution and its borrowers (Brown & Moles, 2014). One of the main objectives of credit risk management is to predict whether the borrower will repay a loan meeting the agreed-upon terms (The Basel Committee on Banking Supervision, 2000). It requires policies, procedures, experience, and the expertise to extract knowledge from massive data sources (Brown & Moles, 2014). Researchers and practitioners have defined various types of credit scoring problems to manage credit risk, depending on the circumstances and background of each borrower (Paleologo, Elisseeff, & Antonini, 2010; L. Thomas, Crook, & Edelman, 2017). In this research, we are interested in both application and behavioral scoring models and procedures. Application scoring supports the loan granting decision. Its objective is to assess the creditworthiness of new applicants combining the applicant's demographic information, loan repayment history, borrower historical data, and credit bureau data, along with data collected in the application form (Anderson, 2022). Credit risk management in application scoring tries to grant loans to those borrowers who will be able to pay and avoid granting credit to those who will not. Similarly, behavioral scoring models are used in credit risk management, but they are applied only to existing customers (Paleologo, Elisseeff, & Antonini, 2010; Anderson, 2022) for whom all of their loan payment behavior is available. This enables lenders to develop an active portfolio management process and to take preventive actions on borrowers with high default likelihood, such as reducing the financial burden of those borrowers who have difficulties complying with the payment schedule and established obligations.

Research on credit scoring is extensive but mainly focused on application scoring. Some researchers describe behavioral scoring knowledge as limited and scarce (Liu, 2001; Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013; Goh & Lee, 2019). We are interested in delving into what we already understand about application scoring and behavioral scoring. First, both strategies combine borrower demographic data, historical information, and features obtained from multiple data sources. Repayment history emerges as one of the main creditworthiness predictors. The effect of this feature set is seen mainly in behavioral scoring; in application scoring, the payment behavior often is not available, or the applicant does not have it (Muñoz-Cancino, Rios, Goic, & Graña, 2021).The payment history features are built by observing the borrower's payment behavior during a specific period; some authors suggest that this period should be 12 months (Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013; Nikolaidis, Doumpos, & Zopounidis, 2017), while most assume its duration as a given. Second, better performance of credit scoring models leads to more accurate decision-making and allows more efficient and profitable credit risk management (Verbraken, Bravo, Weber, & Baesens, 2014; Djeundje, Crook, Calabrese, & Hamid, 2021). To increase the prediction power of these models, financial institutions have used alternative data, especially information from borrowers' relationships and interactions (Ruiz, Gomes, Rodrigues, & Gama, 2017; Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019; Roa, Correa-Bahnsen, et al., 2021; Muñoz-Cancino, Rios, Goic, & Graña, 2021). This type of information adds value to both types of credit scoring. However, it is in application scoring that it achieves the most significant performance enhancement, especially with applicants whose repayment history is not available.

We know the effect of repayment behavior and social-interaction data on application and behavioral scoring problems based on the above. The effect of repayment behavior, is gaining relevance as the relationship between borrower and lender becomes entrenched: the more information on borrower's behavior is collected, the more accurately borrower's creditworthiness can be predicted. In the case of social interaction data, at some point, it becomes irrelevant given the borrower's behavior and repayment history. Both relationships require carefull study. However, to date, we are not aware of any studies on the dynamics of this phenomenon. Research into credit scoring has only examined what occurs at the beginning (application scoring) and at some point during the loan payment schedule (behavioral scoring).

Consequently, this work endeavors to answer the following research questions:

1. Knowing that borrowers' repayment history increases creditworthiness assessment performance, at which point in time since the loan is granted, does this information become meaningful? For how long do we need to observe borrowers' repayment history to assess their creditworthiness accurately?

2. Knowing that social-interaction data contributes more value to application scoring, that is when behavioral information is scarce. For how long is it beneficial to rely on these sources of information?

3. What insights and benefits to credit risk management are obtained from studying the dynamics of both the creditworthiness assessment performance and the value of alternative data sources?

To this end, we gathered and curated a massive multi-source credit dataset containing borrower information and social interaction data in the form of graphs. Then, we conducted a computational experiment in which we selected individuals and companies when they obtained their first loan. And then observed their financial behavior for the first 24 months of those loans. The results were analyzed by considering credit history, repayment behavior, and alternative data and their impact on the creditworthiness assessment performance.

This work contributes to the growing knowledge on credit scoring and the use of social network data. Through our analysis, we challenge the current division of the credit risk management process by investigating what happens between application scoring and behavioral scoring. Focusing the analysis on the borrower rather than on the business process lets us discover how the credit scoring models' performance varies as the borrower credit history is growing. Additionally, we analyzed the contribution of social-interaction features and how their value decreases in the presence of behavioral attributes.

Furthermore, our dataset is novel because it characterizes individuals and companies using information from the moment they obtain their first loan, their subsequent credit history and repayment behavior, and social network data. It overcomes the low availability of data for behavioral models research noted by (Kennedy et al., 2013) and (Goh & Lee, 2019). It allows us to carry out what we believe is the first study on credit assessment performance dynamics.

This paper is structured as follows. Section 3.2 presents a review of the literature on application and behavioral scoring models. The proposed methodology is presented in Section 3.3. Section 3.4 describes the experimental design and datasets. Section 3.5 discusses the results and their implications. The last section provides the conclusions, research findings, and suggestions for future work.

3.2. Background and Related Work

Credit scoring models enable and support credit risk management in financial institutions. For more than half a century, they have been part of decisions throughout the credit risk management cycle (L. Thomas, Crook, & Edelman, 2017). Today, no decisions about whom to grant a loan to, portfolio management, preventive collection actions, or even pricing are made without the support of credit scoring models (Ntwiga, 2016; Anderson, 2022). Academics and practitioners have developed different credit scoring tools to support the different decisions at each stage of the credit risk management cycle. Application scoring is used to decide whether to grant a loan to a new applicant entering the financial system. In contrast, behavioral scoring allows lenders to characterize those borrowers who have already been granted a loan, and it is used mainly for portfolio management. Finally, collection scoring allows optimizing policies and strategies for the collection and recovery (Paleologo, Elisseeff, & Antonini, 2010).

Application scoring models are used to decide whether to grant a loan. In this way, they are understood as the gateway to the lending institution and the financial system. Their correct usage allows the implementation of risk policies and defines the applicant population within which to operate. Therefore, it is important to develop models that allow lenders to correctly quantify the borrower's risk level, predicting with high certainty whether the applicant will default. One of the most commonly used approaches to enhance creditworthiness assessment performance is to improve the modeling techniques, from the traditional logistic regression to other computational intelligence techniques, such as support vector machines (Huang, Chen, & Wang, 2007), Bayesian models (Kao, Lin, & Yu, 2021), genetic algorithms (Kozeny, 2015), ensemble classifiers (García, Marqués, & Sánchez, 2019; Radović, Marinković, & Radojičić, 2021; Moscato, Picariello, & Sperlí, 2021), and deep learning models (West, 2000; Gunnarsson, vanden Broucke, Baesens, óskarsdóttir, & Lemahieu, 2021) including deep belief networks (Luo, Wu, & Wu, 2017; Gunnarsson, vanden Broucke, Baesens, óskarsdóttir, & Lemahieu, 2021). Another approach to improving the creditworthiness assessment is the inclusion of alternative data sources. More and better data leads to better decisions, and in the case of application scoring, there is an increasing body of knowledge analyzing the contribution of alternative data sources such as satellite and geospatial data (Simumba, Okami, Kodaka, & Kohtake, 2021), psychometric data (Rabecca, Atmaja, & Safitri, 2018; Djeundje, Crook, Calabrese, & Hamid, 2021), mobile phone data and communication networks (Óskarsdóttir, Bravo, Vanathien, & Baesens, 2018a; Óskarsdóttir, Bravo, Sarraute, Vanthienen, & Baesens, 2019), network data (Wei, Yildirim, Van den Bulte, & Dellarocas, 2016; Masyutin, 2015; Freedman & Jin, 2017; Cnudde et al., 2019; Giudici, Hadji-Misheva, & Spelta, 2020), and written risk assessments (Stevenson, Mues, & Bravo, 2021). The common grounds of all these studies are that most of the increase in creditworthiness assessment performance occurs when an applicant's traditional information is scarce or unavailable.

Behavioral scoring is used mainly for risk management — that is, understanding what happens after credit has been granted. These models assess actual customers' creditworthiness and enable lenders to take preventive actions with borrowers with a high default likelihood. Unlike the application models, there is no extended research about behavioral scoring (Liu, 2001; Goh & Lee, 2019). (Kennedy et al., 2013) suggest that the reason for the scarcity of research on behavioral scoring is the large volume of data required and the difficulty of accessing the data. However, the lines of research into increasing the performance of these models are the same as for application scoring. (Putra et al., 2020) investigated the value of social network data for predicting bankruptcy, and (Letizia & Lillo, 2019) examined a corporate payments network to assess an internal rating provided by a financial institution. Our previous research (Muñoz-Cancino, Rios, Goic, & Graña, 2021) shows that use of social network data produces a much more significant performance enhancement in application scoring than in behavioral scoring when considering the same population and features. Moreover, this result is consistent when the credit scoring model is applied to both individuals (personal scoring) and companies (business scoring).

Among other studies that address the behavioral scoring problem, the work of (Hsieh, 2004) developed a behavioral scoring model to manage credit card customers using an RFM-based segmentation model and then defined marketing strategies for each group using association rules. (Biron & Bravo, 2014) studied what happens when the logistic regression independence assumptions in behavioral scoring are violated. (Kao et al., 2021) concluded that increasing the APR (annual percentage rate) significantly increases the probability of default using a credit cardholder database from Taiwan.

Behavioral scoring models include information such as repayment behavior, and credit history (L. C. Thomas, 2000) that is not necessarily available in application scoring. The time span in which repayment behavior and banking data are observed is defined as the performance period (L. C. Thomas, 2000). Aspects of the behavior during this period are added as features-for instance, number of missed payments and average balance. There is still no consensus on the optimal performance period length. (L. C. Thomas, 2000; Liu, 2001) used 12 months as an example, and (Djeundje et al., 2021) stated "Behavioral scoring models are applied to accounts that have been open for a sufficient period" (p.2), but without providing details of the sufficient period. (Kennedy et al., 2013) analyzed the performance period length by comparing windows of six months, 12 months, and 18 months. They concluded that the best performance is achieved using a 12-month performance window but limited to shorter outcome windows; in longer outcome windows, it is harder to find optimum performance windows. Therefore, the selection of performance window only affects the short-term creditworthiness assessment.

In Table 3.1, we present a literature review of previous studies dealing with the setting of to the performance period (or observation window). Due to the scarcity of articles that explicitly study this topic, we include articles, books, or reviews that indicate what its length should be as well as those that incorporate features aggregating the borrower's behavior during a specific period. In the table, we list the authors, the entry type, the length of the performance period, and the conclusions obtained in the study. Additionally, we show the details of the experiment carried out: the dataset, algorithms, methodology, and whether the researchers used alternative data. When the performance period is commented upon but was not specifically studied, we display only the comment without the details related to the experiment. Regarding the algorithms, we use the following abbreviations: LR, logistic regression; DT, decision tree; RF, random forest; SVM, support vector machines; GB, gradient boosting; KNN, K-nearest neighbor; ANN, artificial neural networks; and MLP, multi-layer perceptron.

Despite all of this, we do not fully understand how long the performance period should be and how the predictive power varies as more knowledge on repayment behavior becomes available. Additionally, the role and contribution of network data in the shift from application to behavioral scoring remains an open question.

Tabla 3.1: Literature review of performance window analysis for behavioral credit scoring

Authors	Entry	Performance Period (Months)	Conclusions on the Performance Period	Datasets	Algorithms	Methodology	Alternative Data
(Kennedy et al., 2013)	Journal Article	6, 12, 18	The 12-month window achieved the best performance but was limited to some scenarios	2,500 customers from the Irish Credit Bureau	LR	Train Test Split & Bootstrapping	
(Neto, Jorge Adeodato, & Carolina Salgado, 2017)	Journal Article	6, 12, 24		Two datasets: 682 records from the PKDD1999 Challenge and 30,000 customers from Brazilian retail	MLP, ANN, RF, KNN	K-Fold Cross Validation	
(Nikolaidis et al., 2017)	Incollection	1, 3, 6, 12	The 12-month window achieved the best performance but only slightly better	20,000 borrowers and their 86,082 credit lines	LR, SVM	K-Fold Cross Validation	
(Ruiz et al., 2017)	Inproceedings	12		Two datasets: first loans and all loans granted	LR, SVM	K-Fold Cross Validation	Mobile phone network usage
(Óskarsdóttir et al., 2019)	Journal Article	1, 3		22,000 observations. Customers from a telecommunications operator and a commercial bank	LR, DT, RF	Train Test Split	Mobile phone and graph data
(Djeundje et al., 2021)	Journal Article	3,6		Two datasets: 1,826 records from Mexico and 16,358 from Nigeria; both were supplied by Lenddo	LR, ANN, GB	Train Test Split	Psychometric data and customer's email activity
(Kyeong, Kim, & Shin, 2022)	Journal Article	6		200,000 records from KakaoBank in Korea	LR	Train Test Split & Bootstrapping	Log data recorded by the mobile application
(L. C. Thomas, 2000)	Journal Article	12	The 12-month window is used as an example				
(Liu, 2001)	Tech Report	12	The 12-month window is used as an example				
(Siddiqi, 2012)	Book	6, 12	(Siddiqi, 2012) stated, "For behavior scorecard development, accounts are chosen at one point in time, and their behavior analyzed over, typically, a 6- or 12-month period."				
(Bhalla, 2016)	Blog Entry	1	(Bhalla, 2016) stated that "No fixed window for all the models. Depends on the type of model."				
(Mashanovich, 2017)	Blog Entry	12	(Mashanovich, 2017) stated that "The length of the observation and performance windows will depend on the industry sector for which the model is being designed."				

3.3. Creditworthiness Assessment Methodology

We use an approach based on machine learning classification models to analyze the dynamics of creditworthiness assessment performance and how performance is affected by credit history, repayment history, and social network features.

The proposed methodology is presented in Figure 3.1. The diagram begins with a dataset and a feature engineering process. However, because our study largely depends on how datasets are built, these processes are explained in detail in Section 3.4. Specifically, Section 3.4.2 explains how datasets are built, and Section 3.4.4 explains the feature engineering process. This section explains the model training process to assess creditworthiness given a dataset. This process is applied 12 times, once for each dataset built. The first objective of our proposed methodology for training a model is to maximize generalization capability and avoid model overfitting. For this purpose, some authors use the traditional split train-test validation scheme assessing generalization through bootstrapping (Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013; Kyeong, Kim, & Shin, 2022), while others use K-fold cross-validation (Neto, Jorge Adeodato, & Carolina Salgado, 2017; Nikolaidis, Doumpos, & Zopounidis, 2017; Ruiz, Gomes, Rodrigues, & Gama, 2017). However, using the same K-fold cross validation procedure for both hyperparameter optimization and model selection can lead to a biased assessment of the model's performance. For this reason, we apply a hold-out validation methodology that partitions the dataset in two (see Train Split in Fig.3.1 trying to avoid such bias. The first dataset, which contains 30% of the original dataset, is used for feature selection and hyperparameter optimization. The remaining 70% of the dataset is used to train and validate the final models, using the features and hyperparameters previously selected. The results and conclusions are based on the average of 10-fold cross-validation (see K-Fold Cross-Validation in Fig.3.1). The comparison between models' performance results is made using a t-test, according to the recommendations given in (Flach, 2012).

This study uses gradient boosted trees (GB) because they have consistently shown state-of-the-art performance over different problems (Friedman, 2001; Chen & Guestrin, 2016; Muñoz-Cancino, Rios, Goic, & Graña, 2021). Additionally, to quantify the performance,

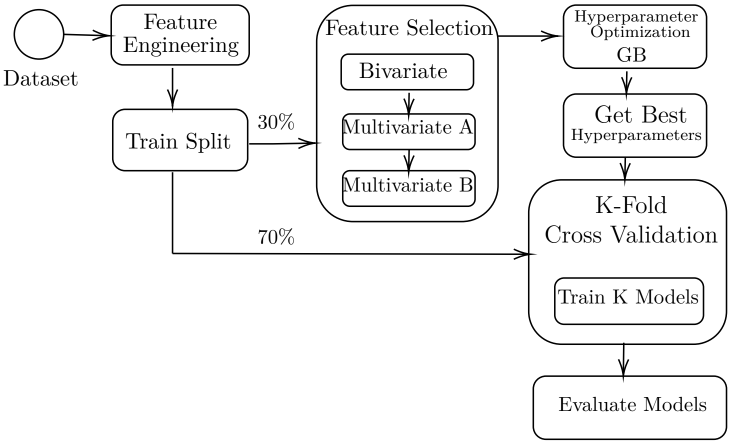


Figure 3.1: Proposed methodology

we use the area under the receiver operating characteristic curve (Bradley, 1997) and the Kolmogorov-Smirnov statistic (Hodges, 1958) as performance measures.

Feature selection (see Feature Selection in Fig.3.1) and hyperparameter optimization (see Hyperparameter Optimization in Fig.3.1) start by discarding those features with low or almost null predictive power; to do so, we calculate the KS and AUC in a univariate (See in Fig.3.1: Bivariate) way and discarded those attributes with a $KS <= 0.01$ or an $AUC <= 0.53$ and then apply a method to drop out highly correlated features. This method begins by selecting the feature with the highest predictive power and then discards those for which the absolute value of the correlation is greater than a parameter $\rho = 0.7$; this process is repeated until all target features are evaluated. We use this method twice, first considering attributes that belong only to the feature sets defined in Section 3.4.4 (shown in Fig.3.1 as Multivariate A); thus, we ensure a representative mix of attributes for each dimension analyzed. Then, we apply it again by considering all remaining features (shown in Fig.3.1 as Multivariate B). Finally, to find the best hyperparameters, we apply an exhaustive search over three specific parameters: first, the number of boosted trees to fit; second, their learning rate; and third, the minimum data needed in leaves. The best combination of hyperparameters is obtained by averaging the performance in a 5-fold cross-validation. The outputs from this stage are a subset of features and a combination of hyperparameters that maximize gradient boosting performance.

3.4. Experimental Setup

3.4.1. Dataset Overview

Experimental data was provided by a Latin American bank. The information was anonymized to protect customer confidentiality and not compromise any customer's identification or relationships; we have taken all precautions to ensure that there is no possibility that this study can leak private data. The experimental dataset includes 97,044 individuals and companies who obtained their first loan between January 2018 and December 2018. The dataset contained information about borrowers' repayment behavior until December 2019.

The data sources used in this research have already been used to develop credit scoring models (Muñoz-Cancino, Rios, Goic, & Graña, 2021). Application and behavioral scoring models were trained. In both cases, the use of borrower information and social interaction data provided a statistically significant improvement to creditworthiness assessment performance. Additionally, in this study, we are interested in analyzing credit scoring models according to the type of borrower. We will refer to personal credit scoring when borrowers are individuals and business credit scoring when borrowers are companies. This classification is complementary to the one previously defined. In this way, it is possible to assess either individuals' or companies' creditworthiness at the time application or afterwards (that is, application or behavioral scoring, respectively).

3.4.1.1. Target

The borrower characteristic used as the target variable for the creditworthiness assessment problem was the event of becoming a defaulter.

Each borrower in the dataset is labeled as a defaulter or non-defaulter. For a borrower to be considered a defaulter, the loan must be more than 90 days past due within the subsequent 12 months from when the borrower was observed; borrowers who are never more than 90 days past due during this period are considered non-defaulters.

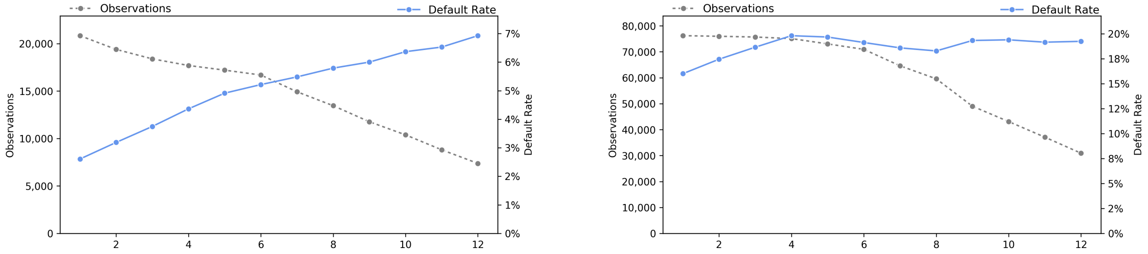
3.4.1.2. Dataset Description

Table 3.2 describes the available information and summarizes the distribution of borrowers, which is grouped into business credit scoring data and personal credit scoring data to distinguish between companies and individuals. It shows the number of borrowers of each type. The number of features corresponds to those provided by the financial institution and those built for this research.

Tabla 3.2: Dataset description. Borrowers correspond to the total number of individuals and companies that are part of our analysis, which will be observed from the moment they obtain a loan until 12 months later.

Model	Borrowers	Features
Business Credit Score	20,835	585
Personal Credit Score	76,209	936

Figure 3.2 shows the number of borrowers and the default rate for each month elapsed since the first loan granting, grouped by business scoring and personal scoring. Both charts start with borrowers at the moment of application: 20,835 and 76,209 borrowers for business and personal scoring, respectively. The number of observations decreases for two reasons, either the borrower is declared in default or the loan is paid in full. The information available in each of the 12 months since the loan granting was used to generate 12 datasets. In Section 3.4.2, we give the details of the construction of these 12 datasets. These datasets are mutually dependent because they contain information on the same borrowers but with diverse progress in repaying their loans. This allows us to gain insights into performance dynamics from independently trained models.



(a) Business Scoring

(b) Personal Scoring

Figure 3.2: Dataset Statistics. The X-axis displays the number of months elapsed since the first loan granting. The left Y-axis shows the observation samples, and the right Y-axis the default rate.

Figure 3.2(a) shows a change in the slope of the decrease in the number of observations in the business scoring problem after the seventh month since the loan granting. At the same time, Fig. 3.2(b) shows an atypical movement of the default rate between the fifth and eighth months since the loan was granted. Our hypothesis, which might be analyzed in future work, is that this may be due to a change in the expected loss estimation model for commercial loans. This change affected firms and individuals because student loans are reported as commercial loans. We believe these changes were due to an adjustment in the bank's client portfolio originating from a change in the formula for calculating the expected loss. This change was implemented in July 2019, seven months after the last record in our sample. Therefore, although this change does not affect our study population, there is the possibility that it could affect our target feature.

3.4.2. Dataset Engineering Pipeline

The original dataset is decomposed into 12 datasets for both companies and individuals. Each one contains information about the borrowers when i months ($i = 1, \ldots, 12$) have elapsed since the granting of the loan. The i-th dataset contains the information of the borrowers available at the i-th month after their first loan. The calendar date of the first loan is not the same for all borrowers in the i-th dataset. The i-th dataset contains information on borrowers whose first loan could have been granted between Jan-18 and Dec-18. Moreover, although the date of granting could be in this period, the i-th dataset information reflects the borrower's financial situation after i months have elapsed since the granting of his first credit.

An overview of the construction process of the datasets can be seen in Figure 3.3. The lower timeline shows how many months have passed since the first credit was granted, in other words it is a virtual timeline. The upper timeline corresponds to calendar line. Hence, $dataset_1$ contains the information relative to the first month of all borrowers who have obtained their first loan between Jan-2018 and Dec-2018, regardless of the date of loan granting. Similarly, $dataset_2$ contains the information of the borrowers during the second month after obtaining their first loan. An so on until month 12. As the months go by, borrowers can be excluded from the ensuing datasets due to default or to full credit payment.

Additionally, to build the target variable, the behavior of each borrower is observed during the 12 months following the moment it was sampled. This definition means the target variable for borrowers in the i-dataset is built from observing these borrowers' payment behavior between the $i + 1$-month and the $i + 12$-month since the first loan was granted.

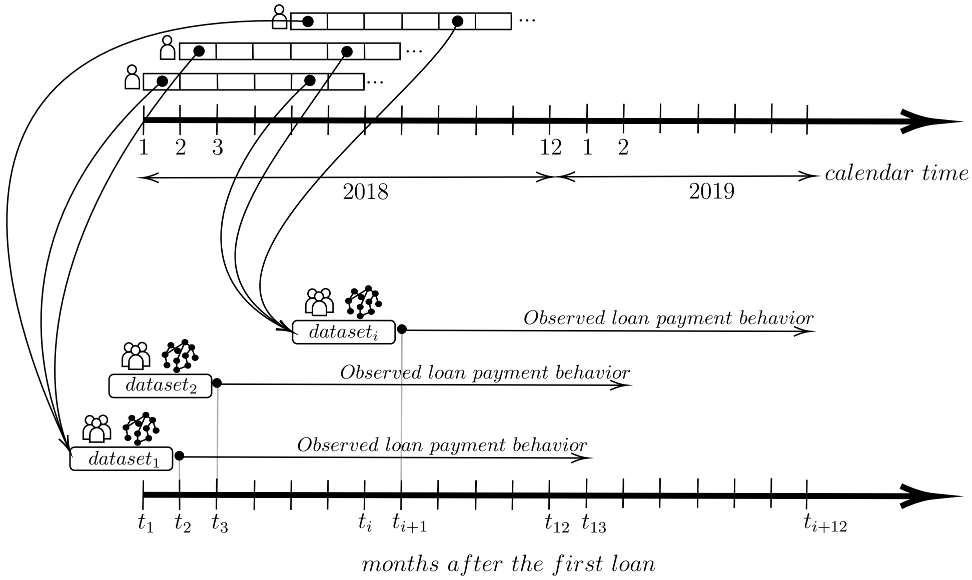

1
2
3
12
1
2
2018
2019
calendar time
$dataset_i$
Observed loan payment behavior
$dataset_2$
Observed loan payment behavior
$dataset_1$
Observed loan payment behavior
t_1
t_2
t_3
t_i
t_{i+1}
t_{12}
t_{13}
t_{i+12}
months after the first loan

Figure 3.3: Dataset construction. Upper timeline corresponds to calendar dates. Lower timeline corresponds to the relative time from first loan granting.

It is necessary to point out that although the models will be trained independently over each dataset, the datasets are not statistically independent between them. In this way, it is possible to obtain insights into performance dynamics through multiple static models.

Borrowers are described using the same feature set in all datasets. However, these features reflect diverse behaviors as the borrower repays the loan or shows signs of credit deterioration.

3.4.3. Data Sources

3.4.3.1. Traditional Data Sources: Borrower Data

To describe the borrowers, we have a massive background dataset with the financial information of 7.65 million people and almost a quarter-million companies from the period between January 2018 and March 2020. The financial information provided in this dataset corresponds to the monthly debt decomposition by type and by days-past-due grouped into buckets. This particular dataset includes all of the study subjects described in Table 3.2.

3.4.3.2. Alternative Data Sources: Social Interaction Data

This study characterizes the companies using network data information originating from their economic and social interactions. The network used for this purpose is composed of transactional services, the enterprise's ownership, and the company's employees. This network builds an ecosystem in which companies, business owners, and employees interact. We call this network EOWNet. On the other hand, individuals are mainly characterized by combining marriages, parents, and children. We call this network FamilyNet. The EOWNet is also used in the personal scoring problem because many of these borrowers are part of the EOWNet. However, due to the partial coverage of this dataset, it is expected to have limited added value, as we observed in (Muñoz-Cancino et al., 2021). The EOWNet is a dynamic network because the interactions that constitute it change monthly, while the FamilyNet is a static network, fixed at the beginning of the study period.

3.4.4. Borrower and Social Network Features

The people and companies whose data is used in this study are characterized by features created through a feature engineering process. This process combines the borrower information, the repayment history discussed in Section 3.4.3.1, and the network data discussed in 3.4.3.2. We classify these features into the following subsets:

- **Borrower's Financial Features:** These correspond to borrower features based on the information provided by the financial institution and allow us to characterize the financial situation of each borrower. The data contains the debt decomposition by type (consumer, commercial, and mortgage) and delinquency situation and the amount in revolving loans. This feature set is called X_{Fin}. It exclusively represents the borrower's situation at the moment of observation. Additionally, we include a feature set with the borrower's historical information and their repayment history; we call this feature set $X_{FinHist}$. This borrower's historical features include the mean and SD for each X_{Fin} feature for the last three and six months. In this way, both X_{Fin} and $X_{FinHist}$ feature sets describe the borrower's financial situation. However, X_{Fin} describes the borrower's credit situation at the observation point, and $X_{FinHist}$ summarizes the borrower's historical financial situation for the last three and six months.

- **Node Statistics:** This feature set considers each borrower as a node within the network. Therefore, these features correspond to nodes' statistics based on their positions and characteristics within the network. For each node in the network, we calculate the degree, degree centrality, number of triads, PageRank score, authority and hub scores from the Hyperlink-Induced Topic Search (HITS) algorithm and an indicator of whether the node is an articulation point. We call this feature set $X_{NodeStats}$, and it is one of the features derived from alternative data sources.

- **Social Interaction Features:** We utilize the borrower's social interactions to characterize each borrower based on their neighborhood's financial information-that is, the individuals and companies to which they are connected. Formally, we use the borrower ego network (egonet) to characterize a borrower using social network data corresponding to all the nodes the borrower is connected to. We aggregate the egonet financial features in the X_{SocInt} feature set, calculating the mean and SD for the nodes' features in the borrower's egonet (Nargesian, Samulowitz, Khurana, Khalil, & Turaga, 2017; Roa, Correa-Bahnsen, et al., 2021). As we did with borrowers' features, we aggregate historical social interaction features by calculating the mean and SD of the last six and three months. We call this additional feature set $X_{SocIntHist}$.

In Table 3.3 we present examples of the definition of attributes constructed based on the above definitions. These variables were constructed from the financial information of each borrower, specifying the decomposition of their debt in the financial system by type and delinquency status and attributes obtained from the borrower's position in the network. The companies in the business credit scoring problem were characterized only based on the EOWNet, while the individuals were characterized based on both the FamilyNet and the EOWNet; this explains why the dataset for personal scoring contains a greater number of attributes. In this work, we are interested in the impact of each feature subset more than each particular attribute and how these impact creditworthiness performance dynamics; therefore, the results were analyzed at the level of each feature subset.

Tabla 3.3: Feature description. Due to the large number of generated features, we describe only three examples per feature subset.

Feature Source	Feature Subset	Feature Name	Description
Borrower's Financial Features	X_{Fin}	NODEATT ATT05	The total amount of revolving loans
		NODEATT ATT07	The total amount of consumer debt
		NODEATT ATT08	The total amount of commercial debt
	$X_{FinHist}$	NODEATT ATT05 MEAN3	The average amount of revolving loans in the last three months
		NODEATT ATT07 MEAN6	The average amount of consumer debt in the last six months
		NODEATT ATT08 STD6	The standard deviation of the amount of commercial debt in the last six months
Node Statistics	$X_{NodeStats}$	EOWNET NodeStats PageRank	Borrower's PageRank score in the EOWNet
		FAMNET NodeStats Hits Auth	Borrower's authority score in the FamilyNet
		FAMNET NodeStats ArtPoint	Whether the borrower is an articulation point in the FamilyNet
Social Interaction Features	X_{SocInt}	EOWNET EgoFull NET MEAN ATT05	The average of the borrower's ego network amount of revolving loans using the complete EOWNet
		EOWNET EgoFull NET MEAN ATT05 Wby Hits Auth	The dot product of the borrower's ego network amount of revolving loans and their authority score using the complete EOWNet
		FAMNET EgoFull NET MEAN ATT07	The average of the borrower's ego network amount of consumer debt using the complete FamilyNet Network
	$X_{SocIntHist}$	EOWNET EgoFull NET MEAN3 ATT05	The average of the borrower's ego network amount of revolving loans in the last three months using the complete EOWNet
		EOWNET EgoFull NET MEAN6 ATT05 Wby Hits Auth	The dot product of the borrower's ego network amount of revolving loans and their Authority Score in the last six months using the complete EOWNet
		FAMNET EgoFull NET MEAN6 ATT07	The average of the borrower's ego network amount of consumer debt in the last six months using the complete FamilyNet

3.4.5. Experiments

We devised a series of experiments to analyze the effects on performance dynamics of credit history, repayment history, and social network features. For this purpose, we generated different sets of characteristics detailed in Table 3.4. With each of these feature sets trained twelve independent 12 models, each model trained over one of the datasets described in Section 3.4.2.

Tabla 3.4: Experiments setup

Experiment Id	Feature Group
E1	$X = \{X_{Fin}\}$
E2	$X = \{X_{Fin} + X_{FinHist}\}$
E3	$X = \{X_{Fin} + X_{FinHist} + X_{NodeStats} + X_{SocInt} + X_{SocIntHist}\}$

We trained a gradient boosting model for each of these experiments according to the methodology outlined in Section 3.3. The optimization of hyperparameters chooses the best model within all the combinations of the following parameters:

- Number of boosting iterations: {50, 100, 250, 500}
- Learning rate: {0.01, 0.05, 0.1}
- Minimal amount of data in one leaf (as % of the training dataset): {2%, 4%, 6%}

For each hyperparameters combination, a gradient boosting model was trained using a 5-fold cross validation. According to the configuration defined in the Experimental Setup section, 6,840 models were trained, of which, 6,480 corresponded to the hyperparameter optimization (36 combinations × 5 folds × 12 datasets × 3 experiments) and 360 corresponded to the final models (10 folds × 12 datasets × 3 experiments). To carry out this study, we used a server with the following specifications:

- Driver node: 140 GB of RAM and 20 CPU cores
- Auto-scaling worker nodes (between 2 and 8)
- Worker node: 112 GB of RAM and 16 CPU cores

3.5. Results and Discussion

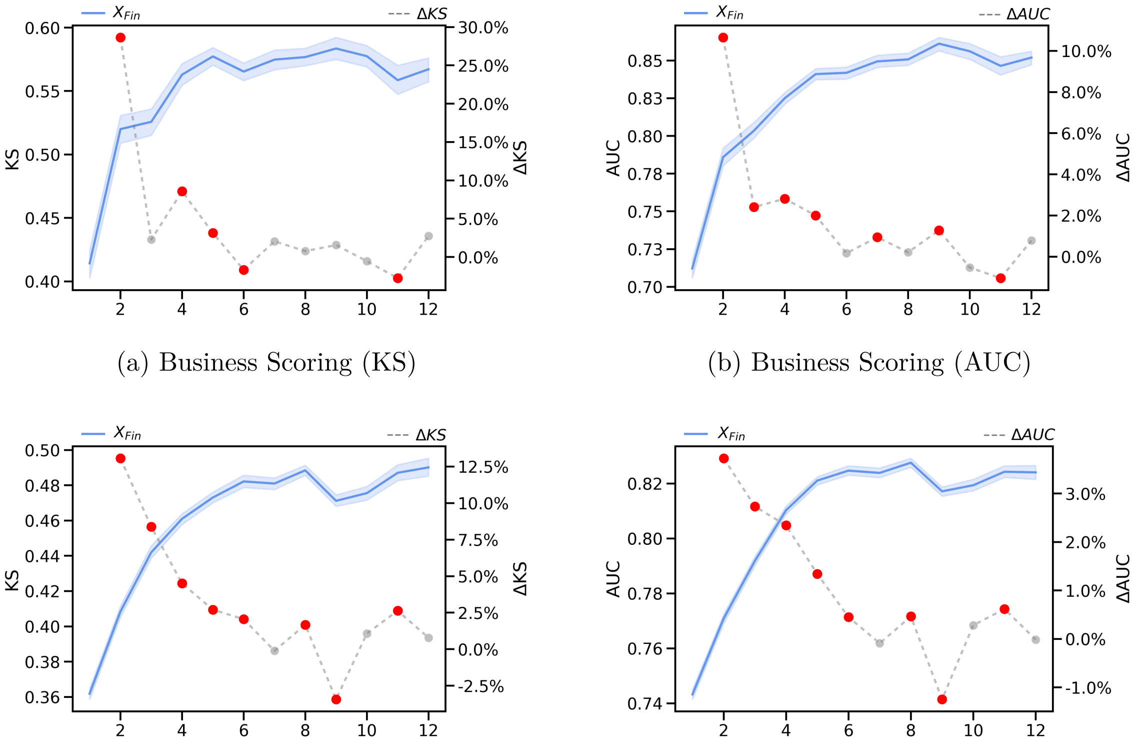
This section presents the results obtained after applying our methodology to study the dynamics of creditworthiness assessment performance. First, we present the effect of the borrower's credit history on the model's performance (experiment **E1**). Then we show how this effect changes when the repayment features are incorporated into the analysis (experiment **E2**). Finally, we study the effect on the model's performance of incorporating the social interaction features, the borrower's credit history and repayment features (experiment **E3**).

A further relevant analysis is to understand how much the social interaction features influence the creditworthiness assessment compared to the borrower's features and how this impact varies over time. The results of this analysis are presented in Section 3.5.4.

3.5.1. Experiment E1: Borrower Credit History

The first goal was to understand how the borrower's credit history affects creditworthiness assessment performance. To this end, we analyzed the behavior of the borrowers' financial features X_{Fin} over time. Figures 3.4(a) and 3.4(b) show this effect for the business scoring problem, and Figures 3.4(c) and 3.4(d) for personal scoring. For each problem, performance was evaluated using the KS and AUC scores.

For each elapsed month since the loan granting, the initial feature set at the beginning of the training was the same. However, after applying the methodology defined in Section 3.3, the final variables could vary between one period and another because we selected those that produced more improvement in default assessment. The discriminatory power increased as the borrower's credit history increased, and the rate of the improvement decreased over time. The increase ceased to be consistently statistically significant after six months, meaning that the gains in discriminatory power were relevant in the first six months.



(a) Business Scoring (KS)

(b) Business Scoring (AUC)

(c) Personal Scoring (KS)

(d) Personal Scoring (AUC)

Figure 3.4: KS and AUC scores for the business scoring and personal scoring problems. The X-axis displays the number of months elapsed since the loan granting. The blue line shows the creditworthiness assessment performance (left Y-axis) for experiment **E1**, using only X_{Fin}: borrower features. The dotted gray line (right Y-axis) shows the percentage increment between consecutive periods; when this increment is statistically significant, the dots are colored red. Otherwise, they are colored gray.

In business scoring, we observed a substantial increase in the second month, greater than 25% in KS and 10% in AUC. However, additional credit history produced relatively minor increases. In personal scoring, on the other hand, the performance increases were smaller but remained consistent in the first six months.

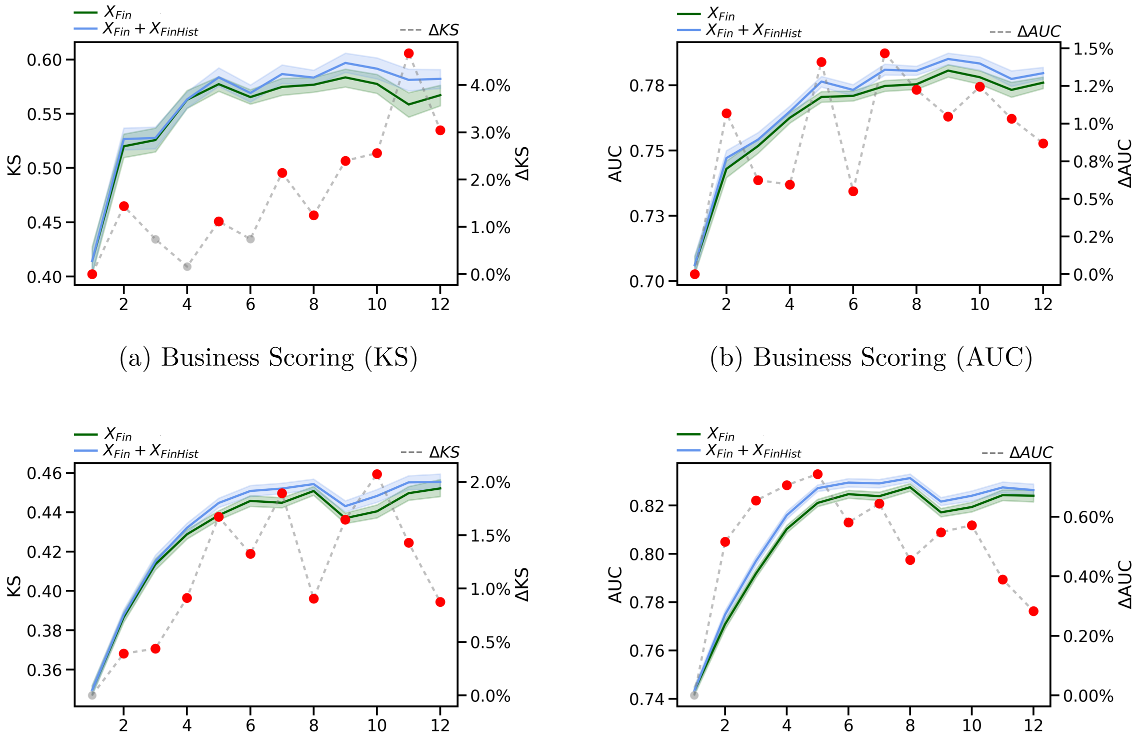
These results confirm what academics and practitioners already know: the importance of borrower credit history in the creditworthiness assessment. Furthermore, they allow us to partially answer the first research question and understand how long we need to observe the borrower's credit history to improve the creditworthiness assessment performance. The value of these results is that they reveal the discrimination power dynamics produced by the availability of borrower history. The credit history produces increases in performance at a decreasing rate. After six months, the gains are marginal; this suggests that the transition from an application scoring problem to a behavioral scoring problem, in terms of discriminatory power, occurs in these six months. Our results challenge the previously proposed definitions of the performance window in behavioral scoring from a "sufficient period" (Djeundje, Crook, Calabrese, & Hamid, 2021) or 12 months (L. C. Thomas, 2000; Liu, 2001; Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013) and instead suggest that the six-month performance window observed in this study is ideal.

A smaller performance window allows reduction in the volume of information necessary to research behavioral scoring models, which, as we have seen, is a limitation in this area (Kennedy, Mac Namee, Delany, O'Sullivan, & Watson, 2013; Goh & Lee, 2019). Furthermore, these results broaden the adequate target population to which these models can be applied; borrowers with six months or more of credit history can now be evaluated rather than only including, borrowers with 12 months or more. The third point favoring a six-month performance window is that it allows faster recovery for defaulters. That is, the credit re-evaluation for borrowers who had negative events in the past and now exhibit good payment behavior can occur earlier, which can greatly help these borrowers in terms of financial inclusion and access to lower interest rates. Finally, other benefits include more straightforward technological implementations, reduction in storage costs, and generation of behavior models that quickly capture the portfolio's trends and shifts.

3.5.2. Experiment E2: Borrower Credit History and Repayment Features

Another advantage of using the borrower's credit history is that allows attributes to be built that reflect the temporal evolution of its characteristics. To do so, we created a set of features that summarized the credit information from the last three and six months. In the first period of analysis, these attributes did not add value because there was no previous history; however, these attributes contributed more as the months passed since loan granting.

We call the repayment history features $X_{FinHist}$, as mentioned in Section 3.4.4. Figure 3.5, compares the results of experiments **E2** and **E1** to analyze the effect of incorporating the attributes $X_{FinHist}$ into the creditworthiness assessment process.



(a) Business Scoring (KS)

(b) Business Scoring (AUC)

(c) Personal Scoring (KS)

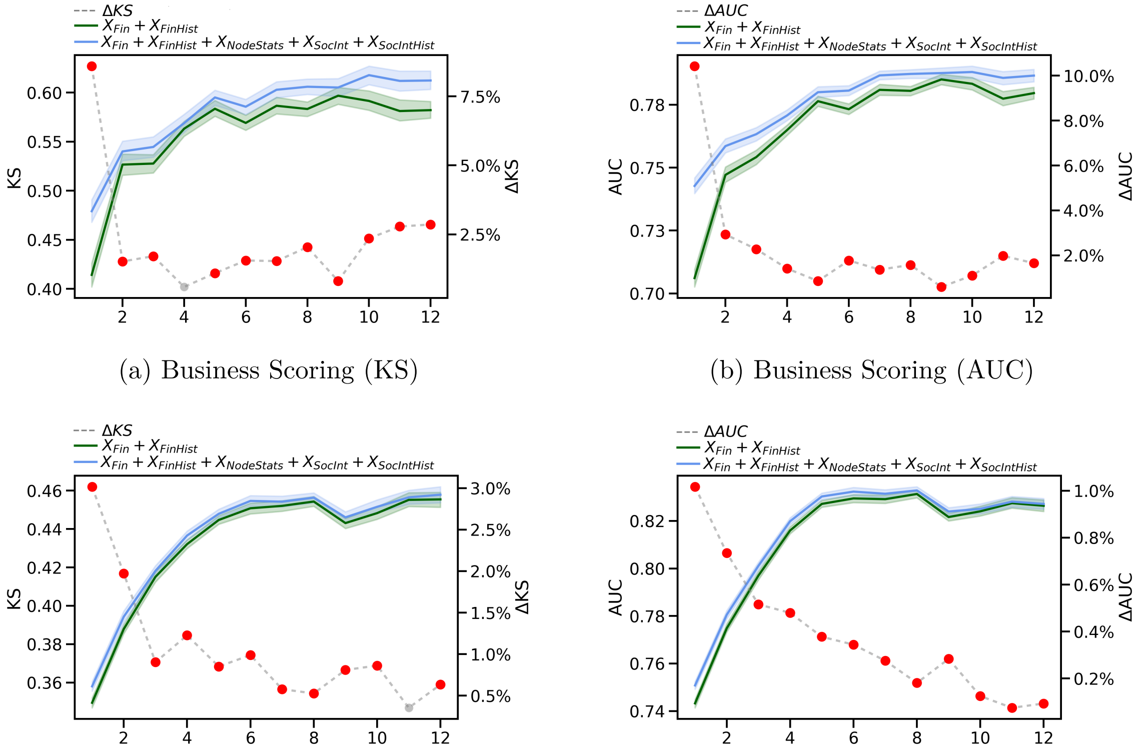
(d) Personal Scoring (AUC)

Figure 3.5: KS and AUC scores for the business scoring and personal scoring problems. The X-axis displays the number of months elapsed since the granting of the loan. The blue line and the green line show the creditworthiness assessment performance (left Y-axis) for experiment **E2** and **E1**, respectively. The dotted gray line (right Y-axis) shows the percentage increment between **E2** and **E1**; when this increment is statistically significant, the dots are colored red. Otherwise, they are colored gray.

The repayment history features affected creditworthiness assessment performance. Discrimination power, measured as KS, increased as borrower history increased and repayment features reflected the borrower's payment behavior. The most meaningful improvements in personal scoring and business scoring occurred six months after credit was initially granted. When performance was measured based on AUC, the relationship between AUC and borrower history was not clear, and the benefits of incorporating repayment history features were observed from the second month onward. The preceding confirms the importance of incorporating repayment history features. These results allow us to complement the answer to the first research question and understand how the credit history and the repayment features impact creditworthiness performance as time passes from the granting of credit. The credit history adds the most significant increase in performance, and when this contribution stabilizes after the first six months, the contribution of the repayment features begins to be noticed. However, it does not change the conclusions obtained above and our suggestion based on the results of the six-month performance window.

3.5.3. Experiment E3: borrower credit history, repayment features and social interaction features

The second research question that motivated this study this study was to determine the value delivered by social interaction features and their impact on the dynamics of creditworthiness assessment performance. Figure 3.6 compares the results of experiments **E3** and **E2**; this comparison allows us to analyze the added value of social interaction features as the borrower's credit history and repayment behavior become available.



(a) Business Scoring (KS)

(b) Business Scoring (AUC)

(c) Personal Scoring (KS)

(d) Personal Scoring (AUC)

Figure 3.6: KS and AUC scores for the business scoring and personal scoring problems. The X-axis displays the number of months elapsed since the granting of the loan. The blue line and the green line show the creditworthiness assessment performance (left Y-axis) for experiment **E3** and **E2**, respectively. The dotted gray line (right Y-axis) shows the percentage increment between **E3** and **E2**; when this increment is statistically significant, the dots are colored red. Otherwise, they are colored gray.

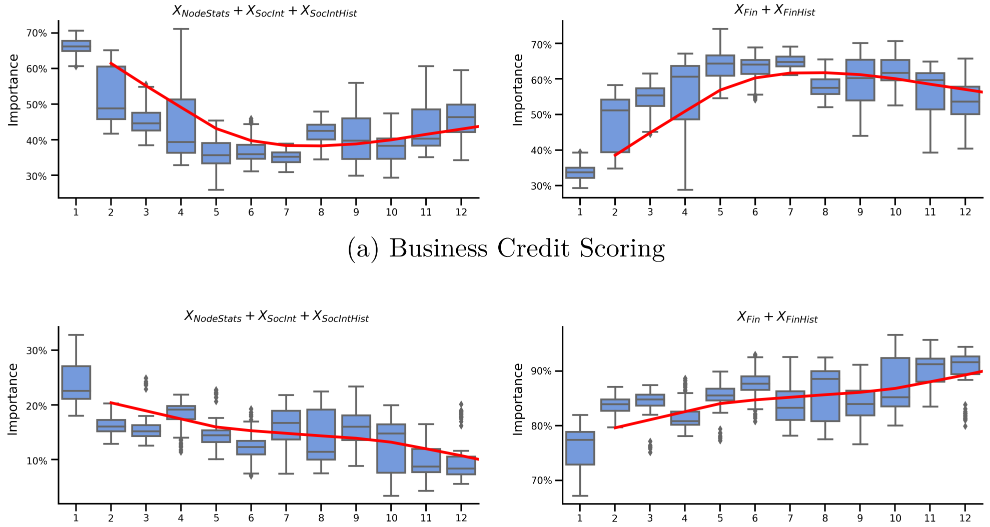
Incorporating alternative data also improves creditworthiness assessment performance. The most significant improvement is observed in the first month, when lenders face an application scoring problem and the borrower's credit information is not available or does not exist.

In personal scoring, social interaction features increased the discrimination power during the 12 months of observation. However, their impact decreased as the borrower's credit history became available, and repayment features were a better predictor of the borrower's payment behavior. In business scoring, social interaction features increased performance by approximately 8% and 10% for KS and AUC, respectively. This significant enhancement in discrimination power occurred in loan applications, a clear sign that the creditworthiness assessment of a firm should not rely only on its features. The assessment should also consider the behavior and attributes of the firm's owners and analyze its supply chain with customers, suppliers, and employees.

3.5.4. Importance of Social Interaction Features Over Time

This section analyzes the borrower's ego network characteristics (that is social interaction features) and how their impact varies as credit history and repayment behavior become available. To do this, we considered the relative importance of each attribute to predict creditworthiness. This relative importance was grouped into two categories: borrower features ($X_{Fin} + X_{FinHist}$) and the features obtained from social network data (that is, node statistics and social interaction features) ($X_{NodeStats} + X_{SocInt} + X_{SocIntHist}$). For each feature, the importance was calculated as the average of the Shapley values from a subset of the dataset and then aggregated according to the two categories previously defined. The Shapley values were obtained using a tree-based SHAP explainer (Lundberg & Lee, 2017).

Figures 3.7(a) and (b) show feature importance using Shapley values for business and personal scoring, respectively.



(a) Business Credit Scoring

(b) Personal Credit Scoring

Figure 3.7: Feature Importance Analysis using Shapley Values. Figure (a) presents the Business Scoring problem and Figure (b) Personal Scoring Problem, in both using the Experiment **E3** feature set. The features are grouped into two categories, the borrower's features ($X_{Fin} + X_{FinHist}$), and the social interaction features ($X_{NodeStats} + X_{SocInt} + X_{SocIntHist}$). The X-axis displays the number of months elapsed since the loan granting. The Y-axis shows the relative feature importance. The boxplots show the feature importance in the 10-fold cross-validation, and the red line is a LOWESS regression fitted using these results.

The importance of network features in business scoring was 63.8% in the first month, the same month in which this information generated its maximum discrimination power enhancement. A similar effect was observed in personal scoring; however, network information was less important, and its increase in discrimination power was correspondingly smaller.

Figure 3.7(a) shows the importance of social interaction features in the creditworthiness assessment. When a company is applying for a loan, this alternative data contributed the most to the credit evaluation. Its value decreased as the company's information became available during the following months; despite this decrease, the importance of social interaction features stabilized around 40% after the first six months. This result confirms what is already known by practitioners: the credit evaluation of a firm, especially in small and medium companies, must consider its owners and the business ecosystem in which the company interacts.

On the other hand, in personal scoring, the importance of social interaction features diminished almost linearly as time passed, as illustrated in Fig. 3.7(b). These features were most important at the time of application, but their importance was considerably smaller than in business scoring, as was their increase in discrimination power. Parental relationships and marriages did not have the same impact on creditworthiness assessment of individual borrowers as transactional and economic relationships had on business scoring. Despite this, social interaction features increase the power of discrimination. They provide fundamental support in the financial inclusion of those people whom traditional credit scoring models cannot evaluate since they do not have a credit history.

An interesting relationship to analyze is the one presented in Fig. 3.8. This figure shows on the same scale the increase in discrimination power measured in KS and AUC and the importance of social interaction features. A high correlation was observed between the increase in discrimination power and the importance of social interaction features in both cases. In business scoring, this correlation was almost perfect during the first six months of the study, and in personal scoring, a strong pseudo-linear correlation was observed. This relationship shows that the contribution of the social interaction features to the creditworthiness assessment directly translated to an increase in discrimination power.

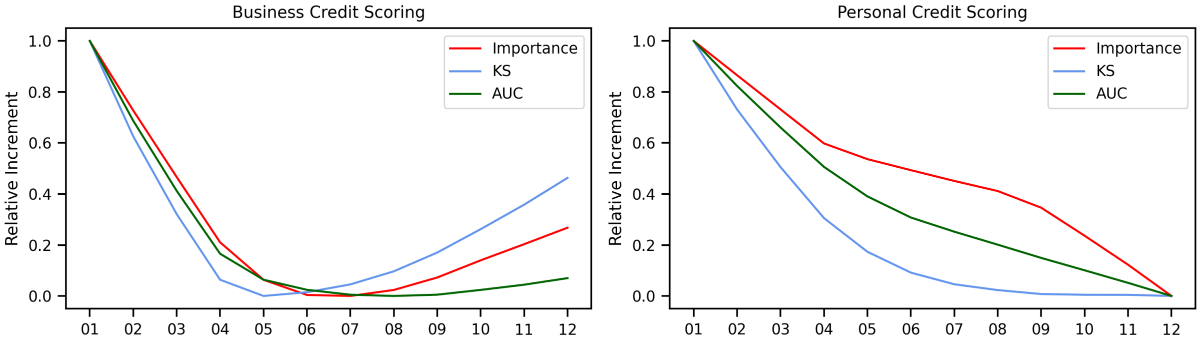


Figure 3.8: Feature importance and predictive power relationship. (Left) Business scoring problem; (right) Personal scoring problem. Both using the Experiment **E3** feature set. The blue and green lines show the relative increase between **E3** and **E2** experiments for the KS and the AUC. The red line is the social interaction features' importance in experiment **E3**. A *MinMaxScaler* was applied to all series to limit the results between 0 and 1. The X-axis corresponds to the months elapsed after the first loan granting. The Y-axis shows the relative increment.

3.6. Conclusions and Future Work

This study analyzed how credit history, loan repayment features, and social network data influence the performance of credit scoring models. We used traditional financial data and graph data originating from borrowers' economic and social interactions. Additionally, our novel dataset allowed us to analyze all of the financial behavior of individuals and companies from the moment they obtained their first loan until 24 months afterward. Furthermore, we analyzed performance dynamics based on the results of multiple independent creditworthiness assessment models trained with time-dependent datasets. These models were trained with features representing the borrower's credit history, repayment behavior, and social interactions. The performance of these models was measured in terms of AUC and KS; the feature importance was quantified using Shapley values.

Our findings showed that as more borrowers' credit history became available, creditworthiness assessment performance increased at a decreasing rate. This effect was observed up to six months from the loan granting, when it stabilized. This finding is meaningful because it reduces the temporal extent of the datasets necessary to train and research behavioral credit scoring models. It also increases the population that can be evaluated using these models from borrowers with 12 months or more of credit history to those with only six months. Furthermore, it enhances financial inclusion and leverages second-chance banking, allowing those borrowers with good credit behavior but with a negative credit history to be reintegrated into the financial system sooner. An additional noteworthy finding is that the features that summarize the borrower's repayment behavior-the repayment history features $X_{FinHist}$-enhanced the creditworthiness assessment performance, especially after the first six months, and they consequently increased performance when the contribution of credit history stabilized. Finally, the social interaction features improved performance, and they added the most value when the borrower was applying for the loan. In personal scoring, this effect decreased to nearly zero as the customer's history became more available. In business scoring, the increment in discrimination power gained by incorporating social network features remained stable, at least during the first year.

The results obtained allow us to analyze the dynamics of creditworthiness assessment performance and how it is influenced by the borrower's credit history, repayment behavior, and social interaction features. These results are important since they provide support for a six-month performance window, reducing it from the current recommendations of 12 months. In addition, they show us how the importance and discrimination power enhancement of social interaction features changes over time. Both insights allow us to improve credit risk management by establishing when and for how long to use social network data; similar conclusions are drawn about the performance windows and the contribution of repayment features.

These results also help answer the third research question raised in this study. Our work is beneficial and has implications to credit risk management in the academic field and real-world applications. In the academic field, our work extends and deepens current knowledge about the impact of credit history and repayment behavior. Furthermore, it also analyzed when it is convenient to use alternative data and how it impacts the performance of credit scoring models, whether applied to individuals or companies. In real-world applications, our work impacts both lenders and borrowers. The lenders benefit from our study's results because it allows them to optimize their risk management process. On the one hand, changing the observation window from 12 to six months reduces management and data storage costs.

This reduction also increases the population of behavior models by reducing the loan age constraint. Additionally, it allows lenders to know for how long they should use social information to improve their credit scoring models. This result also contributes to the efficiency of their processes. Borrowers benefit from models with smaller observation windows since the negative events they have had in the past are forgotten by the credit scoring models more quickly. In this way, the borrowers can receive an accurate creditworthiness assessment faster. Additionally, showing the value of the social-interaction data allows the borrowers to complement their credit information when they do not have it. Therefore, it increases the possibilities of financial inclusion, especially for those cases with little or no credit information.

Our study suggests numerous lines of investigation. First, we would like to extend our research period; our 24-month dataset only allowed us to gain insights from the first 12 months of the borrowers' behavior. Based on our results, it is feasible that the impact of social interaction networks stabilizes after 12 months in the business scoring problem. A more extended observation period would also allow us to study mortgages that have a slower evolution than consumer and commercial credits. Second, we would like to understand what happens in other domains, either using other types of networks or studying microcredits or peer-to-peer lending. Finally, in this work, we studied creditworthiness assessment performance dynamics through multiple independent models trained with time-dependent datasets; we would like to design a framework that inherently handles time dependency.

Acknowledgments

This work would not have been accomplished without the financial support of CONICYT-PFCHA / DOCTORADO BECAS CHILE / 2019-21190345. The second author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery Grant RGPIN-2020-07114]. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. The last author thanks the partial support of FEDER funds through the MINECO project TIN2017-85827-P and the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 777720.

Chapter 4

Assessment of creditworthiness models privacy-preserving training with synthetic data[4]

Abstract

Credit scoring models are the primary instrument used by financial institutions to manage credit risk. The scarcity of research on behavioral scoring is due to the difficult data access. Financial institutions have to maintain the privacy and security of borrowers' information refrain them from collaborating in research initiatives. In this work, we present a methodology that allows us to evaluate the performance of models trained with synthetic data when they are applied to real-world data. Our results show that synthetic data quality is increasingly poor when the number of attributes increases. However, creditworthiness assessment models trained with synthetic data show a reduction of 3% of AUC and 6% of KS when compared with models trained with real data. These results have a significant impact since they encourage credit risk investigation from synthetic data, making it possible to maintain borrowers' privacy and to address problems that until now have been hampered by the availability of information.

Keywords: credit scoring; synthetic data; generative adversarial networks; variational autoencoders

4.1. Introduction

For decades financial institutions have used mathematical models to determine borrowers' creditworthiness and consequently manage credit risk. The main objective of these models is to characterize each borrower with the probability of not complying with their contractual obligations (The Basel Committee on Banking Supervision, 2000), avoiding to give loans to applicants that will not be able to pay them back. Despite all the years of research on credit scoring, there is still little done on behavioral scoring models, which are the credit scoring models used on those clients who have already been granted a loan, because it requires large

[4] The following is a copy of the paper presented at the Hybrid Artificial Intelligent Systems Conference. Please cite this paper as follows: Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña (2022). Assessment of Creditworthiness Models Privacy-Preserving Training with Synthetic Data. In: , et al. Hybrid Artificial Intelligent Systems. HAIS 2022. Lecture Notes in Computer Science(), vol 13469. Springer, Cham.

volumes of data and a relevant historical depth (Goh & Lee, 2019; Kennedy et al., 2013). In addition, financial institutions are often reluctant to collaborate in this type of investigation due to concerns about data security and personal privacy. Until now, the use of synthetic data in credit scoring is mainly restricted to balancing the minority class in classification problems using the traditional SMOTE (Gicić & Subasi, 2019), variational autoencoders (Wan et al., 2017), and lately generative adversarial networks (Fiore et al., 2019; Lei et al., 2020; Ngwenduna & Mbuvha, 2021). In these studies, synthetic records of the minority class are generated, and the original data set is augmented. In this paper, we present a framework that allows us to train a model on synthetic data and then apply it to real-world data. We also analyze if the model copes with data drift by applying both models to real-world data representing the same problem but obtaining the dataset one year later. The main findings of our work are:

- It is possible to train a model on synthetic data that achieves good performance in real situations.
- As the number of features increases, the synthesized data quality gets worse.
- There is a performance cost for working in a privacy-preserving environment. This cost corresponds to a loss of predictive power of approximately 3% if measured in AUC and 6% in KS.

4.2. Related Work

4.2.1. Credit Scoring

Credit scoring aims to manage credit risk, defined as the potential for a borrower to default on established contractual obligations (The Basel Committee on Banking Supervision, 2000). These models intensively use borrower data, demographic information, payment behavior, and even alternative data sources such as social networks (Muñoz-Cancino et al., 2021; Óskarsdóttir et al., 2019), psychometrics (Djeundje et al., 2021), and geolocation (Simumba et al., 2021).

4.2.2. Generative models for synthetic data generation

Generative models are a subset of machine learning models whose main objective is to learn the real-data distribution and then to generate consistent samples from the learned distribution. Working with synthetic data allows addressing problems where real-data is expensive to obtain, where a large dataset is needed to train a model, or where the real-data is sensitive or cannot be shared (Torres, 2018). For years, statistical methods were the most used ones to estimate the real-world data joint distribution. In this group, Gaussian Mixture Models are the most utilized for this task when there are fewer continuous variables. At the same time, Bayesian Networks are commonly used for discrete variables. The main problem of these methods is dealing with datasets containing numerical and categorical variables. They also present problems when the continuous variables have more than one mode and the categorical variables present small categories (Xu, 2020). During the last years, deep learning models have gained popularity to generate synthetic data due to their performance and because they allow us to deal with the problems mentioned above. The generative adversarial networks and the variational autoencoders stand out within these models.

4.2.2.1. Generative Adversarial Networks

Generative adversarial networks are a deep learning framework based on a game theory scenario where a generator network $\mathcal{G}(\cdot)$ must compete with a discriminator network $\mathcal{D}(\cdot)$. The generator network produces samples of synthetic data that attempt to emulate real data. In contrast, the discriminator network aims to differentiate between real examples from the training dataset and synthetic samples obtained from the generator (Goodfellow, Bengio, & Courville, 2016). Its most basic form, vanilla GAN, $\mathcal{G}(\cdot)$ maps a vector z from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to a vector $\hat{x}$ in the data domain X. While $\mathcal{D}(\cdot)$ outputs a probability that indicates whether $\hat{x}$ is a real training samples or a fake sample drawn from $\mathcal{G}(\cdot)$ (Xu, 2020). The generator $\mathcal{G}(\cdot)$ and the discriminator $\mathcal{D}(\cdot)$ are alternatively optimized to train a GAN. Vanilla GANs have two main problems, representing unbalanced categorical features and expressing numerical features having multiple modes. To solve this, Xu et al. (2019) (Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019) present a conditional generator (CTGAN) that samples records from each category according to the log-frequency; this way, the generator can explore all discrete values. Moreover, the multimodal distributions are handled using kernel density estimation to assess the number of modes in each numerical feature.

4.2.2.2. Variational autoencoders

Autoencoders (AE) are an unsupervised machine learning method that enables two main objectives: low-dimensional representation and synthetic data generation. Variational Autoencoder (Kingma & Welling, 2013) interpret the latent space produced by the encoder as a probability distribution modeling the training samples as independent random variables, assuming the posterior distribution defined by the encoder $q_\theta(z|x)$ and generative distribution $p_\phi(x|z)$ defined by the decoder. To accomplish that the encoder produces two vectors as output, one of means and the other of standard deviations, which are the parameters to be optimized in the model. Xu et al. (2019) (Xu et al., 2019) present TVAE, a variational autoencoder adaption for tabular data, using the same pre-processing as in CTGAN and the evidence lower bound (ELBO) loss.

4.3. Methodology and Experimental Design

4.3.1. Dataset

In this work, we use a dataset provided by a financial institution already used for research on credit scoring (Muñoz-Cancino et al., 2021; Muñoz-Cancino et al., 2022). This dataset includes each borrower financial information and social interactions features over two periods: January 2018 and January 2019; each dataset contains 500,000 individuals. Each borrower is labeled based on their payment behavior in the following 12-month observation period. Each borrower in the 2018 dataset is labeled as a defaulter if it was more than 90 days past due between February 2018 and January 2019 and is labeled as a non-defaulter if it was not more than 90 days past due. Borrowers from the Jan-2019 dataset are similarly tagged. This dataset contains three feature subsets: X_{Fin} corresponds to the borrower's financial information, X_{Degree} corresponds to the number of connections the borrower has in the social interaction network, and X_{SocInt} are the features extracted from the social interactions.

4.3.2. Synthetic data generation

A step to privacy-preserving credit scoring model building is to generate a synthetic dataset that mimics real-world behavior. In order to accomplish this, we compare the performance of two state-of-the-art synthetic data generators, CTGAN and TVAE, defined in Sect. 4.2. The first experiment (**S01**) only compares these methods using borrowers' features X_{Fin}. The objective of this stage is to find a method to generate synthetic data from real data, and it is not part of this study to find the best way to generate them. Despite not generating an exhaustive search for the best hyper-parameters, we will test two different architectures (Arch) for each synthesizer. Arch A is the default configuration for both methods. In the case of CTGAN, Arch B set up the generator with two linear residual layers and the discriminator with two linear layers, both of size (64, 64). In the case of TVAE Arch B, set hidden layers of (64, 64) for both the encoder and the decoder. Then, in experiment **S02**, we train a new synthesizer using the best architecture from **S01**. This experiment uses the borrowers' features X_{Fin} and exclusively one feature from the network data, the node degree X_{Degree}. We only include node degree because its feature enables us to reconstruct an entire network using the random graphs generators. Finally, in experiment **S03**, the borrowers and social interaction features (X_{Fin}+X_{Degree}+X_{SocInt}) are used to train a synthesizer. This experiment corresponds to the traditional approach to generating synthetic data from a dataset using social interaction features.

4.3.3. Borrower's creditworthiness assessment

The objective of this stage is to have a framework that allows us to estimate the borrower's creditworthiness from a feature set. This modeling framework is based on previous investigations (Muñoz-Cancino et al., 2021; Muñoz-Cancino et al., 2022). This stage begins by discarding attributes with low or null predictive power and selecting uncorrelated attributes. The correlation-based selection method begins by selecting the attribute with the highest predictive power. It then discards the possible selections if the correlation exceeds a threshold ρ. This step is repeated until no attributes are left to select. To ensure the model generalization capability, we work under a K-fold cross-validation scheme; in this way, the feature selection and the model training use K-1 folds, and the evaluation is carried out with the remaining fold. Additionally, we use two holdout datasets, one generated with information from the same year as the training dataset but not contained. The second contains information from one year later. Both the results of the validation fold and the holdout dataset are stored to use a t-test later to compare different models (Flach, 2012, Ch. 12).

4.3.4. Evaluation Metrics

In this section, we describe a set of metrics that will help us to evaluate the performance of the synthetic data generators and the classification models used for creditworthiness assessment. The area under the curve **(AUC)** is a performance measure used to evaluate classification models (Bradley, 1997). The AUC is an overall measure of performance that can be interpreted as the average of the true positive rate for all possible values of the false positive rate. A higher AUC indicates a higher overall performance of the classification model (Ho, Mo, & Chan-Hee, 2004). Another classification performance measure is the **F-measure**. This metric is calculated as the harmonic mean between precision and recall. It is beneficial for dichotomous outputs and when there is no preference between maximizing the model's precision or recall (Hripcsak & Rothschild, 2005). Kolmogorov-Smirnov (KS) statistic measures the distance separating two cumulative distributions (Hodges, 1958). The KS statistic ranges between 0 and 1 and is defined as $D = \max_x |F_1(x) - F_2(x)|$, where F_1 and F_2 are two cumulative distributions. In the case of creditworthiness assessment, we are interested in the difference between the cumulative distributions of defaulters and non-defaulters, and a higher D indicates a higher discriminatory power. However, in the case of synthetic data generation, we are interested in the real data distribution and the synthetic data distribution being as similar as possible; in this way, a lower D indicates a better synthetic data generation. In order for all the acceptance criteria to be the same, we define the **KSTest** as $1 - D$; in this way, a higher KSTest indicates a better synthetic data generator. In the synthetic data generation problem, the KS is only valid to measure the performance for continuous features; to handle categorical features, we will use the chi-square test (CS). The CS is a famous test to assess the independence of two events (McHugh, 2013). We will call **CSTest** to the resulting p-value for this test. Therefore a small value indicates we can reject the null hypothesis that synthetic data has the real data distribution. In the synthetic data generation problem, we want to maximize the CSTest.

4.3.5. Experimental setup

The parameters of the univariate selection are set at $KS_{min} = 0.01$ and $AUC_{min} = 0.53$, i.e., we discard feature with a univarite performance lower than KS_{min} or AUC_{min}. In the multivariate selection process, we set $\rho = 0.7$ in the process to avoid high correlated features (Akoglu, 2018). The N-Fold Cross-Validation stage is carried out considering $N = 10$, and in each iteration, the results of regularized logistic regression and gradient boosting (Friedman, 2001) models are displayed.

4.4. Results and Discussion

In this section, we present the results of our methodology. We start with the implementation details. Then, we compare the synthesizers, and finally, we analyze the creditworthiness assessment performance of the models trained using synthetic data.

4.4.1. Implementation Details

In this work, we used the Python implementations of Networkx v2.6.3 (Hagberg et al., 2008) and Synthetic Data Vault (SDV) v5.0.0 (Patki, Wedge, & Veeramachaneni, 2016) for networks statistics and synthetic data generation, respectively. To conduct the experiments, we used a laptop with 8 CPU cores Intel i7 and 32 GB of RAM.

4.4.2. Synthetic Data Generation Performance

The first objective is to analyze the performance of the methods to generate synthetic data presented above, CTGAN and TVAE. Table 4.1 shows the results obtained. The features Synthesizer training features corresponds to the training feature set, while Arq indicates the network architecture defined in Sect. 4.3.2. The experiment S01 consisted in comparing both synthesizer using two different architectures. It is observed that a reduction in the number of layers reduces the execution times considerably in both cases, being TVAE, the one that presented the fastest execution times. KSTest show us the performance to synthesize continuous features, where TVAE achieves better performance than CTGAN. The difference between TVAE architectures is almost negligible when evaluate continuous features performance. The performance to synthesize categorical features is measured using CSTest. In this case, TVAE obtained higher performance again, the differences between architectures is slightly higher to architecture A. Another popular approach to measuring the synthesizer performance is training a classifier to distinguish between real and synthetic data. The column Logistic Detection in Table 1 shows the result after training a logistic regression model; the value displayed corresponds to the complementary F-measure. In this way, values closer to 1 indicate that the classifier cannot distinguish between real and synthetic data, and values closer to 0 mean the classifier efficiently detects synthetic data. It can be seen that TVAE achieve the best performance, but this performance decreases as we include more features to the synthesizer.

Tabla 4.1: Synthetic data generators performance

Experiment	Synthesizer training features	Synthesizer	Arch	Exec Time (m)	CSTest	KSTest	Logistic Detection
S01	X_{Fin}	CTGAN	A	410	0.836	0.864	0.697
			B	260	0.861	0.846	0.749
		TVAE	A	230	0.962	0.868	0.803
			B	130	0.952	0.861	0.756
S02	$X_{Fin} + X_{Degree}$	TVAE	B	140	0.935	0.836	0.644
S03	$X_{Fin} + X_{Degree} + X_{SocInt}$	TVAE	A	400	0.924	0.809	0.539
S03	$X_{Fin} + X_{Degree} + X_{SocInt}$	TVAE	B	320	0.907	0.825	0.542
S03	$X_{Fin} + X_{Degree} + X_{SocInt}$	TVAE	B	465	0.930	0.819	0.513

4.4.3. Creditworthiness assessment performance on real data

This section establishes a comparison line for the performance of the models trained with synthetic data. In order to establish this comparison, we first trained classifiers using real-world data and tested their performance using the holdout datasets previously defined. Table 4.2 shows the results of training models according to the methodology described in 4.3.3. The performance is measured using AUC and KS on the two holdout datasets; the 10-folds mean and its standard deviation are shown for each statistic. For each feature set, we trained two classifiers, logistic regression and gradient boosting. The results show that gradient boosting obtains better results compared to logistic regression. More details of this comparison are shown in Table 4.3, where we quantify the higher predictive power of gradient boosting.

Tabla 4.2: Creditworthiness assessment performance for models trained on real data

Classifier training features	Classifier	Holdout 2018		Holdout 2019	
		AUC	KS	AUC	KS
X_{Fin}	GB	0.88 ± 0.001	0.59 ± 0.002	0.82 ± 0.001	0.50 ± 0.002
X_{Fin}	LR	0.87 ± 0.001	0.58 ± 0.001	0.82 ± 0.001	0.50 ± 0.002
$X_{Fin} + X_{Degree} + X_{SocInt}$	GB	0.88 ± 0.001	0.59 ± 0.002	0.82 ± 0.001	0.50 ± 0.002
$X_{Fin} + X_{Degree} + X_{SocInt}$	LR	0.87 ± 0.001	0.58 ± 0.002	0.83 ± 0.001	0.50 ± 0.002
$X_{Degree} + X_{SocInt}$	GB	0.61 ± 0.002	0.17 ± 0.002	0.62 ± 0.001	0.18 ± 0.002
$X_{Degree} + X_{SocInt}$	LR	0.60 ± 0.001	0.17 ± 0.002	0.61 ± 0.001	0.18 ± 0.002

Based on the results presented above, we will select gradient boosting for the comparisons against the models trained on synthetic data that we will present in the next section.

Tabla 4.3: Gradient boosting and logistic regression comparison on real data (holdout 2018)

Classifier training features	AUC diff (%)	KS diff (%)	AUC diff p-value	KS diff p-value
X_{Fin}	0.70%	1.65%	0.000	0.000
$X_{Fin} + X_{Degree} + X_{SocInt}$	0.84%	1.91%	0.000	0.000
$X_{Degree} + X_{SocInt}$	1.65%	2.36%	0.000	0.000

4.4.4. Creditworthiness assessment performance on synthetic data

This section aims to know how the performance of a creditworthiness assessment model (the classifier) behaves when trained on synthetic data and applied to real-world data. Table 4.4 shows the performance indicators on real-world data. Considering all synthesizers are trained with at least the feature set X_{Fin}, the results of training the classifier with X_{Fin} are also displayed for all synthesizers. It is observed that regardless of the synthesizer, training the classifier incorporating at least feature set X_{Fin} produces similar performances in 2018 except in **S02**. However, when we analyze how much the model degrades, the model trained with synthetic X_{Fin} from synthesizer **S01** is the one that suffers a minor discrimination power loss. It can be explained in part that a better synthesizer manages to capture better the proper relationship between the borrower features and the default.

Tabla 4.4: Creditworthiness assessment performance on real data for model trained on synthetic data

Synthesizer Experiment	Classifier training features	holdout 2018		holdout 2019	
		AUC	KS	AUC	KS
S01	X_{Fin}	0.85 ± 0.003	0.53 ± 0.002	0.82 ± 0.002	0.48 ± 0.002
S02	X_{Fin}	0.82 ± 0.001	0.51 ± 0.001	0.80 ± 0.001	0.46 ± 0.002
S03	X_{Fin}	0.85 ± 0.002	0.55 ± 0.002	0.80 ± 0.002	0.46 ± 0.002
S03	$X_{Fin} + X_{Degree} + X_{SocInt}$	0.85 ± 0.002	0.56 ± 0.003	0.80 ± 0.002	0.47 ± 0.003
S03	$X_{Degree} + X_{SocInt}$	0.60 ± 0.002	0.16 ± 0.002	0.61 ± 0.003	0.18 ± 0.002

The comparison of performance obtained by the models trained with synthetic data against the models trained on real-world data is presented in Table 4.5. We can understand this comparison as the cost of using synthetic data, and it corresponds to the loss of predictive power to preserve the borrower's privacy. We can observe that in the best cases, this decrease in predictive power is approximately 3% and 6% when we measure the performance in AUC and KS, respectively.

Tabla 4.5: Comparison between models trained using synthetic data and models trained on real data. ** denotes when the difference is statistically significant using 0.05 as the p-value threshold, while * uses 0.1.

Synthesizer Experiment	Classifier training features	holdout 2018		holdout 2019	
		AUC diff	KS diff	AUC diff	KS diff
S01	X_{Fin}	-3.59%**	-10.09%**	-0.86%**	-3.92%**
S02	X_{Fin}	-6.24%**	-13.24%**	-3.32%**	-6.48%**
S03	X_{Fin}	-2.81%**	-6.01%**	-3.21%**	-6.70%**
S03	$X_{Fin} + X_{Degree} + X_{SocInt}$	-3.12%**	-5.68%**	-2.54%**	-4.73%**
S03	$X_{Degree} + X_{SocInt}$	-1.85%**	-4.31%**	-0.69%**	1.10%*

4.5. Conclusions

This work aimed to use synthetic data to train creditworthiness assessment models. We used a massive dataset of 1 million individuals and trained state-of-the-art synthesizer methods to obtain synthetic data and achieve this goal. Then, we presented a training framework that allows us to analyze trained models with synthetic data and observe their performance on real-world data. In addition, we observed their performance one year after being trained to see how susceptible they are to data drift. Our results show that lower quality synthetic data is obtained as we increase the number of attributes in the synthesizer. Despite this, it is possible to use these data to train models that obtain good results in real-world scenarios, with only a reduction in the predictive power of approximately 3% and 6% when we measure the performance in AUC and KS, respectively. These findings are of great relevance since they allow us to train accurate creditworthiness models. At the same time, we keep borrowers' privacy and encourage financial institutions to strengthen ties with academia and foster collaboration and research in credit scoring without the privacy and security restrictions.

4.6. Future Work

Our future work will delve into how to synthesize social interactions' information in the form of graphs and not as added attributes to the training dataset since, as we show, this deteriorates the quality of the synthetic data.

Acknowledgements

This work would not have been accomplished without the financial support of CONICYT-PFCHA / DOCTORADO BECAS CHILE / 2019-21190345. The second author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery Grant RGPIN-2020-07114]. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. The last author thanks the support of MICIN UNDER project PID2020-116346GB-I00.

Chapter 5

Conclusions

This doctoral thesis has produced valuable insights and advancements to enhance the creditworthiness assessment. We have identified valuable knowledge impacting the predictive performance of credit scoring models by blending different network representation learning methodologies and using synthetic data for training models. Our first study's conclusions emphasize the importance of incorporating social-interaction data into credit scoring, especially for thin-file borrowers. We have improved creditworthiness assessment performance by combining various graph representation learning approaches, such as hand-crafted feature engineering and Graph Neural Networks. This methodology enhances the predictive power of credit scoring models and contributes to financial inclusion by providing a proper credit risk assessment for unbanked applicants.

Our second study explored the dynamics of creditworthiness assessment performance based on credit history, loan repayment behavior, and social network data. Our findings indicate that incorporating social interaction features adds significant value, particularly in the early stages of the lending process. Additionally, as more credit history becomes available, creditworthiness assessment performance increases at a decreasing rate. This effect was observed up to six months from the loan granting when it stabilized. This finding is meaningful because it reduces the temporal extent of the datasets necessary to train and research behavioral credit scoring models. These insights allow for more efficient risk management, expand the possibilities of financial inclusion, and leverage second-chance banking, providing an opportunity for borrowers with a negative credit history but a recent good credit behavior to regain access to the financial system at an accelerated pace.

Lastly, the third study introduces the use of synthetic data in training creditworthiness assessment models. Our results demonstrate that these models can perform well in real-world scenarios while preserving borrower privacy. This approach opens up opportunities for collaboration between financial institutions and academia, encouraging research and innovation in credit scoring without compromising privacy and security.

These conclusions highlight the potential for enhancing creditworthiness assessment models through integrating social-interaction data, understanding the dynamics of credit behavior, and leveraging synthetic data for training purposes. By addressing these key aspects, we can further refine credit risk management procedures, promote financial inclusion, and facilitate meaningful collaborations between academia and industry to pursue more accurate and robust credit scoring methodologies.

Bibliography

Akoglu, H. (2018, 08). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. doi: 10.1016/j.tjem.2018.08.001

Anderson, R. A. (2022). *Credit intelligence and modelling: Many paths through the forest of credit rating and scoring.* Oxford University Press.

Apostolik, R., Donohue, C., & Went, P. (2009). *Foundations of banking risk: an overview of banking, banking risks, and risk-based banking regulation* (Vol. 507). John Wiley & Sons Incorporated.

Arsov, N., & Mirceva, G. (2019). Network embedding: An overview. *arXiv preprint arXiv:1911.11726*.

Aziz, S., & Dowling, M. (2019). Machine learning and ai for risk management. In *Disrupting finance: Fintech and strategy in the 21st century* (pp. 33–50). Cham: Springer International Publishing. doi: 10.1007/978-3-030-02330-0_3

Baidoo, E. (2020). *A credit analysis of the unbanked and underbanked: an argument for alternative data* (PhD dissertation). Analytics and Data Science Institute, Kennesaw State University.

Banu, I. M. (2013). The impact of credit on economic growth in the global crisis context. *Procedia Economics and Finance*, *6*, 25-30. (International Economic Conference of Sibiu 2013 Post Crisis Economy: Challenges and Opportunities, IECS 2013) doi: https://doi.org/10.1016/S2212-5671(13)00109-3

Bhalla, D. (2016). Observation and performance window [Computer software manual]. (Retrieved from https://www.listendata.com/2016/08/observation-and-performance-window.html. Accessed July 10, 2022)

Biron, M., & Bravo, C. (2014). On the discriminative power of credit scoring systems trained on independent samples. In M. Spiliopoulou, L. Schmidt-Thieme, & R. Janning (Eds.), *Data analysis, machine learning and knowledge discovery* (pp. 247–254). Cham: Springer International Publishing.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Bravo, C., Thomas, L. C., & Weber, R. (2015). Improving credit scoring by differentiating defaulter behaviour. *The Journal of the Operational Research Society*, *66*(5), 771–781.

Bravo, C., & Óskarsdóttir, M. (2020). Evolution of credit risk using a personalized pagerank algorithm for multilayer networks. *arXiv preprint arXiv:2005.12418*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brown, K., & Moles, P. (2014). Credit risk management. *K. Brown & P. Moles, Credit Risk Management*, *16*.

Carta, S., Ferreira, A., Reforgiato Recupero, D., & Saia, R. (2021). Credit scoring by leveraging an ensemble stochastic criterion in a transformed feature space. *Progress in Artificial Intelligence*, *10*(4), 417–432.

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* (p. 785–794). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2939672.2939785

Cnudde, S. D., Moeyersoms, J., Stankova, M., Tobback, E., Javaly, V., & Martens, D. (2019). What does your facebook profile reveal about your creditworthiness? using alternative data for microfinance. *Journal of the Operational Research Society*, *70*(3), 353-363. doi: 10.1080/01605682.2018.1434402

Cusmano, L. (2018). SME and entrepreneurship financing: The role of credit guarantee schemes and mutual guarantee societies in supporting finance for small and medium-sized enterprises. *OECD SME and Entrepreneurship Papers, No. 1*. doi: 10.1787/35b8fece-en

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems* (pp. 3844–3852).

Diallo, B., & Al-Titi, O. (2017). Local growth and access to credit: Theory and evidence. *Journal of Macroeconomics*, *54*, 410-423. (Banking in Macroeconomic Theory and Policy) doi: https://doi.org/10.1016/j.jmacro.2017.07.005

Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, *163*, 113766. doi: https://doi.org/10.1016/j.eswa.2020.113766

Fang, F., & Chen, Y. (2019). A new approach for credit scoring by directly maximizing the kolmogorov–smirnov statistic. *Computational Statistics & Data Analysis*, *133*, 180–194.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34.

Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448-455.

Flach, P. A. (2012). *Machine learning - the art and science of algorithms that make sense of data*. Cambridge University Press.

Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: Lessons from peer-to-peer lending. *International Journal of Industrial Organization*, *51*, 185 - 222. doi: https://doi.org/10.1016/j.ijindorg.2016.09.002

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine.

Annals of statistics, 1189–1232.

García, V., Marqués, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*, *47*, 88-101. doi: https://doi.org/10.1016/j.inffus.2018.07.004

Gicić, A., & Subasi, A. (2019). Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. *Expert Systems*, *36*(2), e12363.

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020). Network based credit risk models. *Quality Engineering*, *32*(2), 199-211. doi: 10.1080/08982112.2019.1655159

Goel, A., & Rastogi, S. (2021). Credit scoring of small and medium enterprises: a behavioural approach. *Journal of Entrepreneurship in Emerging Economies*.

Goh, R. Y., & Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, *2019*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 855–864). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2939672.2939754

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research.* doi: https://doi.org/10.1016/j.ejor.2021.03.006

Hagberg, A., Swart, P., & SChult, D. (2008). Exploring network structure, dynamics, and function using networkx. In *In proceedings of the 7th python in science conference (scipy* (pp. 11–15).

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Ho, P. S., Mo, G. J., & Chan-Hee, J. (2004). Receiver operating characteristic (roc) curve: Practical review for radiologists. *kjr*, *5*(1), 11-18.

Hodges, J. (1958). The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, *3*(5), 469–486.

Hripcsak, G., & Rothschild, A. S. (2005, 05). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296-298.

Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, *27*(4), 623-633. doi: https://doi.org/10.1016/j.eswa.2004.06.007

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847-856. doi: https://doi.org/10.1016/j.eswa.2006.07.007

Hurley, M., & Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, *18*(1), 5.

Kao, L.-J., Lin, F., & Yu, C. Y. (2021). Bayesian behavior scoring model. *Journal of Data Science*, *11*(3), 433-450. doi: 10.6339/JDS.201307_11(3).0004

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012, 12). Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, *6*(4). doi: 10.1145/2382577.2382579

Kennedy, K., Mac Namee, B., Delany, S., O'Sullivan, M., & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications*, *40*(4), 1372-1380. doi: https://doi.org/10.1016/j.eswa.2012.08.052

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N., & Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N., & Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Kleinberg, J. M. (1999, 9). Authoritative sources in a hyperlinked environment. *J. ACM*, *46*(5), 604–632. doi: 10.1145/324133.324140

Kozeny, V. (2015). Genetic algorithms for credit scoring: Alternative fitness function performance comparison. *Expert Systems with Applications*, *42*(6), 2998-3004. doi: https://doi.org/10.1016/j.eswa.2014.11.028

Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., & Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring. *Decision Support Systems*, *120*, 106-117. doi: https://doi.org/10.1016/j.dss.2019.03.011

Kyeong, S., Kim, D., & Shin, J. (2022). Can system log data enhance the performance of credit scoring?-evidence from an internet bank in korea. *Sustainability*, *14*(1). doi: 10.3390/su14010130

Lei, K., Xie, Y., Zhong, S., Dai, J., Yang, M., & Shen, Y. (2020). Generative adversarial fusion network for class imbalance credit scoring. *Neural Computing and Applications*, *32*(12), 8451–8462.

Leskovec, J., & Sosič, R. (2016). Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(1), 1.

Letizia, E., & Lillo, F. (2019). Corporate payments networks and credit risk rating. *EPJ Data Science*, *8*(1), 21.

Liu, Y. (2001). New issues in credit scoring application. *Work report No*, *16*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4765–4774). Red Hook, NY, USA: Curran Associates Inc.

Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, *65*, 465-470. doi: https://doi.org/10.1016/j.engappai.2016.12.002

Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research*, *261*(2), 656 - 665. doi: https://doi.org/10.1016/j.ejor.2017.02.037

Mashanovich, N. (2017). Credit scoring: Part 2 – credit scorecard modelling methodology [Computer software manual]. (Retrieved from https://www.worldprogramming.com/blog/datascience/credit__scoring__pt2/". Accessed July 10, 2022)

Masyutin, A. (2015). Credit scoring based on social network data. *Business Informatics*, *3*(33), 15-23.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, *23*(2), 143–149.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, 113986. doi: https://doi.org/10.1016/j.eswa.2020.113986

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2022). Assessment of creditworthiness models privacy-preserving training with synthetic data. In *Hybrid artificial intelligent systems* (pp. 375–384). Cham: Springer International Publishing.

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023a). On the combination of graph data for assessing thin-file borrowers' creditworthiness. *Expert Systems with Applications*, *213*, 118809. doi: https://doi.org/10.1016/j.eswa.2022.118809

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023b). On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance. *Expert Systems with Applications*, *218*, 119599. doi: https://doi.org/10.1016/j.eswa.2023.119599

Muñoz-Cancino, R., Rios, S. A., Goic, M., & Graña, M. (2021). Non-intrusive assessment of covid-19 lockdown follow-up and impact using credit card information: Case study in chile. *International Journal of Environmental Research and Public Health*, *18*(11). doi: 10.3390/ijerph18115507

Muñoz-Cancino, R., Ríos, S. A., & Graña, M. (2023). Clustering cities over features extracted from multiple virtual sensors measuring micro-level activity patterns allows one to discriminate large-scale city characteristics. *Sensors*, *23*(11). doi: 10.3390/s23115165

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. (2017). Learning feature engineering for classification. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 2529–2535). doi: 10.24963/ijcai.2017/352

Neto, R., Jorge Adeodato, P., & Carolina Salgado, A. (2017). A framework for data transformation in credit behavioral scoring applications based on model driven development. *Expert Systems with Applications*, *72*, 293-305. doi: https://doi.org/10.1016/j.eswa.2016.10.059

Ngwenduna, K. S., & Mbuvha, R. (2021). Alleviating class imbalance in actuarial applications using generative adversarial networks. *Risks*, *9*(3).

Nikolaidis, D., Doumpos, M., & Zopounidis, C. (2017). Exploring population drift on consumer credit behavioral scoring. In *Operational research in business and economics* (pp. 145–165). Springer.

Niu, B., Ren, J., & Li, X. (2019, 12). Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information*, *10*(12), 397. doi: 10.3390/info10120397

Ntwiga, D. B. (2016). *Social network analysis for credit risk modeling* (Unpublished doctoral dissertation). University of Nairobi.

Óskarsdóttir, M., Bravo, C., Vanathien, J., & Baesens, B. (2018a, 11). Credit scoring for good: enhancing financial inclusion with smartphone-based microlending. In *Proceedings of the thirty ninth international conference on information systems.* San Francisco, California, USA.

Óskarsdóttir, M., Bravo, C., Vanathien, J., & Baesens, B. (2018b, 7). Social network analytics in micro-lending. In *29th european conference on operational research (08/07/18 - 11/07/18).* Valencia, Spain.

Óskarsdóttir, M., & Bravo, C. (2021). Multilayer network analysis for improved credit risk prediction. *Omega*, *105*, 102520. doi: https://doi.org/10.1016/j.omega.2021.102520

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, *74*, 26 - 39. doi: https://doi.org/10.1016/j.asoc.2018.10.004

Óskarsdóttir, M., Bravo, C., Verbeke, W., Baesens, B., & Vanthienen, J. (2018). *Effects of network architecture on model performance when predicting churn in telco.*

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, *85*, 204 - 220. doi: https://doi.org/10.1016/j.eswa.2017.05.028

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, *201*(2), 490-499. doi: https://doi.org/10.1016/j.ejor.2009.03.008

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*, 8026–8037.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016, Oct). The synthetic data vault. In *2016 ieee international conference on data science and advanced analytics (dsaa)* (p. 399-410).

Putra, S. G. P., Joshi, B., Redi, J., & Bozzon, A. (2020). A credit scoring model for smes based on social media data. In M. Bielikova, T. Mikkonen, & C. Pautasso (Eds.), *Web engineering* (pp. 113–129). Cham: Springer International Publishing.

Rabecca, H., Atmaja, N. D., & Safitri, S. (2018). Psychometric credit scoring in indonesia microfinance industry: A case study in pt amartha mikro fintek. In *The 3rd international conference on management in emerging markets (icmem 2018)* (pp. 620–631). Bali, Indonesia.

Radović, O., Marinković, S., & Radojičić, J. (2021). Credit scoring with an ensemble deep learning classification methods–comparison with traditional methods. *Facta Universitatis, Series: Economics and Organization*, 029–043.

Rajan, R. G., & Zingales, L. (1996, September). *Financial dependence and growth* (Working Paper No. 5758). National Bureau of Economic Research. doi: 10.3386/w5758

Rathi, S., Verma, J. P., Jain, R., Nayyar, A., & Thakur, N. (2022). Psychometric profiling of individuals using twitter profiles: A psychological natural language processing based approach. *Concurrency and Computation: Practice and Experience*, e7029. doi: 10.1002/cpe.7029

Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M. A., & Bravo, C. (2021). Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications. *Expert Systems with Applications*, *169*, 114486. doi: https://doi.org/10.1016/j.eswa.2020.114486

Roa, L., Rodríguez-Rey, A., Correa-Bahnsen, A., & Valencia, C. (2021). *Supporting financial inclusion with graph machine learning and super-app alternative data.*

Romero, D. M., Uzzi, B., & Kleinberg, J. (2019). Social networks under stress: Specialized team roles and their communication structure. *ACM Transactions on the Web (TWEB)*, *13*(1), 1–24.

Ruiz, S., Gomes, P., Rodrigues, L., & Gama, J. (2017). Credit scoring in microfinance using non-traditional data. In E. Oliveira, J. Gama, Z. Vale, & H. Lopes Cardoso (Eds.), *Progress in artificial intelligence* (pp. 447–458). Cham: Springer International Publishing.

Shumovskaia, V., Fedyanin, K., Sukharev, I., Berestnev, D., & Panov, M. (2020). *Linking bank clients using graph neural networks powered by rich transactional data.*

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). John Wiley & Sons.

Simumba, N., Okami, S., Kodaka, A., & Kohtake, N. (2021). Spatiotemporal integration of mobile, satellite, and public geospatial data for enhanced credit scoring. *Symmetry*, *13*(4). doi: 10.3390/sym13040575

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, *295*(2), 758-771. doi: https://doi.org/10.1016/j.ejor.2021.03.008

Sukharev, I., Shumovskaia, V., Fedyanin, K., Panov, M., & Berestnev, D. (2020). *Ews-gcn: Edge weight-shared graph convolutional network for transactional banking data.*

Tan, T., & Phan, T. Q. (2018). Social media-driven credit scoring: The predictive value of social structures. *Available at SSRN 3217885*.

The Basel Committee on Banking Supervision. (2000, 09). Principles for the management of credit risk. *Basel Committee Publications*, *75*.

The Global Financial Index. (2022). *The global findex database 2021: Financial inclusion, digital payments, and resilience in the age of covid-19.* (Retrieved from https://openknowledge.worldbank.org/bitstream/handle/10986/37578/9781464818974.pdf. Accessed July 3, 2022)

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications.* SIAM.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149-172. doi: https://doi.org/10.1016/S0169-2070(00)00034-0

Torres, D. G. (2018). *Generation of synthetic data with generative adversarial networks*

(Unpublished doctoral dissertation). Ph. D. Thesis, Royal Institute of Technology, Stockholm, Sweden, 26 November.

Van, L. T.-H., Vo, A. T., Nguyen, N. T., & Vo, D. H. (2021). Financial inclusion and economic growth: An international evidence. *Emerging Markets Finance and Trade*, *57*(1), 239-263. doi: 10.1080/1540496X.2019.1697672

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505 - 513. doi: https://doi.org/10.1016/j.ejor.2014.04.001

Vlasselaer, V. V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38 - 48. doi: https://doi.org/10.1016/j.dss.2015.04.013

Wan, Z., Zhang, Y., & He, H. (2017). Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 ieee symposium series on computational intelligence (ssci)* (p. 1-7).

Wei, Y., Yildirim, P., Van den Bulte, C., & Dellarocas, C. (2016). Credit scoring with social network data. *Marketing Science*, *35*(2), 234-258. doi: 10.1287/mksc.2015.0949

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11), 1131-1152. doi: https://doi.org/10.1016/S0305-0548(99)00149-5

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29–39).

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21. doi: 10.1109/TNNLS.2020.2978386

Xu, L. (2020). *Synthesizing tabular data using conditional gan* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *CoRR*, *abs/1907.00503*.

Zeng, G., & Zeng, E. (2019). On the three-way equivalence of auc in credit scoring with tied scores. *Communications in Statistics-Theory and Methods*, *48*(7), 1635–1650.

Zhang, M., & Chen, Y. (2018). Link prediction based on graph neural networks. In *Advances in neural information processing systems* (pp. 5165–5175).