Machine Learning Models and Alternative Data in Credit Scoring: Statistical and Financial impact

Luisa Fernanda Roa Ballen

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Industrial Engineering

of

Universidad de los Andes.

Department of Industrial Engineering

Universidad de los Andes

June 16, 2021

Acknowledgements

I would first like to thank my advisor, Dr. Carlos Valencia, who guided me throughout this project. Your insightful feedback and valuable discussions were crucial for the development of this work. I would also like to extend my gratitude to Dr. Alejandro Correa-Bahnsen for allowing me to work on this project, for his support and helpful advice. Finally, I would like to thank my family for their support and encouragement during this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Super-Apps worldwide have revolutionized the on-demand delivery market by creating digital platforms that allow finding anything a consumer might need, from purchase prepared foods and groceries to airplane and event tickets. In addition, these platforms have created ecosystems in which millions of transactions are transacted daily, specifically given that by 2020 the number of smartphone users were 3.5 billion [1] that is almost half of the worldwide population. Hence, consumers have highly adopted Super-Apps as these centralize the offer of services and products with a clear and intuitive interface [2].

Furthermore, some of these companies have developed electronic payment platforms that offer payment services, bank transfers, cash withdrawals, phone recharges, and transfers among users. The largest and best-known Super-Apps worldwide are WeChat and Alipay, China's internet giants Tencent and Alibaba. WeChat emerged as a messaging application, while Alipay was founded as the payment department of Taobao. Now, these portals have more than 40 and 90 functionalities, respectively, ranging from financial services to transportation, food delivery, e-commerce, and multimedia content [3].

Although the revolution of these platforms began in Asia, the expansion of these business models to different regions has been seen in recent years. Specifically, in Latin America, Supper-Apss like Mercado Libre, Movile, and Rappi has created a complete ecosystem within its application, which offers everything from grocery purchases, bets, and games to money transfers and cash withdrawals. The good reception of these platforms is partly due to the versatility they give their users to find different services in one place and specifically the financial facilities. With each new service or functionality that Super-Apps bring to the market, especially financial ones, these increasingly constitute themselves as financial technology companies (fintech) [4].

Being part of the financial market represents great challenges, such as creating disruptive services to the innovative evaluation of users to assign the right financial products. However, Super-Apps have a competitive advantage over traditional banking to face these challenges. Furthermore, since Super-Apps have the data generated by users on their platforms and at the same time access to transactional data common to banking, supper applications can better understand the needs, profiles, and interactions of users who want to access alternative financial services. In this way, this new data known as alternative data [5] is essential to provide information on the financial behavior of unbanked consumers and to strengthen knowledge of individuals that already are in the financial system.

For financial companies, be they traditional or fintech entities, evaluating credit risk is a fundamental task because based on this assessment, they decide whether to make a loan or not. According to the World Bank in 2006 [6], credit scoring was defined as a statistical technique that combines several financial characteristics to form a single score to assess a borrower's creditworthiness. It stands out in this definition that a financial characterization must be used for the credit evaluation and although currently the bank recognizes that there is a proliferation in the use of machine learning and that the alternative data is part of the variables to consider [7], large credit bureaus and financial institutions rely primarily on past credit behavior and financial information. The latter excludes from the evaluation system those young individuals who want to start a financial life or those who have never agreed for various reasons. Especially in recent years, research into creating more inclusive and fair credit valuation systems has seen tremendous growth, and the use of alternative data has shown strong credit scoring [8]. For example, Alibaba has consumer spend behavior data of over 420 million customers, which has been used to build its proprietary Sesame credit score. In comparison, National Credit Bureau, run by People's Bank of China has data on 300 million people.

Currently, one of the most popular all-in-one applications in Latin America is expanding to offer more services and part of the company's efforts are focused on creating financial services. Hence, exploring new data and creating credit models become essential for the company to assess users accurately. The need for this research arises after the results obtained in [9], in which a first approximation to a credit score model with the Super-App transactional variables was made, finding that the alternative data contribute significantly to the statistical and economic performance of the model. However, since the alternative data to be explored is abundant, this research aims to make a detailed analysis not only of transactional data but also of the networks that are formed between users within the Super-App, whether by sharing a credit card or device or location, by having p2p transactions between them and by sharing the first digits in a credit card known as BIN. In addition, it is intended to explore different techniques that allow improving the quality of data, such as feature selection. Likewise, the aim is to study different machine learning models to benchmark statistical and financial performance.

Preliminary results of this work were presented in [10], accepted for publication at the Intelligent Systems Conference (IntelliSys) 2021.

1.0.1 Justification

The problem that this research is intended to attack has a financial, regulatory, and social impact. First, for the Super-App, the research proposes a solution that allows evaluating the creditworthiness of the users to access the new financial services that the company offers. Herein lies the importance, at an organizational level, of having models with good performance since, for each credit assigned to a bad payer or the non-access to credit of a good payer, the company must assume the costs that this implies. In addition, it is important to consider that this research gives the Super-App a competitive advantage because although different Fintech companies have begun to explore alternative data, the companie can be a pioneer with some variables and gain more significant participation in the alternative financial market.

Likewise, the regulatory impact of this research poses challenges for regulators since it is necessary to ensure that decisions based on machine learning models with alternative data are as accurate as possible but also as unbiased, transparent, and fair as possible. In this way, it is necessary to mitigate the potential risks of these new approaches so that once the information is generalized, it leads to greater banking penetration.

Finally, the social implications lie in the financial inclusion of people who do not have banking services. Access to financial services allows people, especially the most vulnerable, to have the capacity and tools to manage their money. This inclusion directly impacts individuals and their families as it can drive economic growth through financing.

1.0.2 Research propose

Alternative data and machine learning models have proven to be features and tools that allow people to be financially evaluated more accurately. Furthermore, they have allowed the bankarization of people who are practically invisible to the traditional financial system. Hence, the current research has three main objectives:

1. Develop credit score machine learning models in order to provide a credit assessment for the granting of credits and financial products.

2. Identify and generate variables with alternative data that allow to have a competitive advantage over traditional financial institutions and Fintech competitors.

3. Analyze and quantify the statistical and financial impact of the proposed models and selected variables for the company.

Chapter 2

Literature Review

2.0.1 Traditional credit scoring

The granting of credits is one of the main axes of banking institutions and consists of granting money loans with the commitment that the borrower will gradually return these resources in the future. The granting of credit is a critical and fundamental phase in risk management since candidates must be studied and evaluated in order to determine their financial status and payment capacity to guarantee the generation of a healthy portfolio. Financial institutions use credit rating systems to measure the risk associated with each loan application [11] and assign a credit score to clients that reflects the approximate, individual likelihood of default.

The estimation of credit scores is evaluated through a predictive model and generally, for consumer credit score, the variables considered include age, marital status, income, credit bureau, and historical financial behavior within the same entity [12]. Specifically, the bureau score is usually acquired by banking entities to strengthen their internal risk assessment systems. This metric is calculated only upon the centralized information held at a Credit Reference Agency (Credit Bureau) as Equifax, Experian, and TransUnion using the Fair Isaac and Company (FICO) method. This method was introduced in 1989 and scores within a range of 300 - 850 points, and this consists of weighing the payment history, use of credit, length of credit history, requests for new credits, and types of credits used [13].

Formally, a credit score is a statistical model that allows the estimation of the probability $\hat{p}_i = P(y_i = 1|\mathbf{x}_i)$ of a customer i defaulting upon a contracted debt.

Additionally, since the objective of credit scoring is to estimate a classifier c_i to decide whether or not to grant a loan to a customer i, a threshold t is defined such that if $\hat{p}_i < t$, then the loan is granted, that is, $c_i(t) = 0$, and denied otherwise, that is, $c_i(t) = 1$ [14]. Although in the literature there are different studies with more than 40 classification models, the standard in the industry is still the logistic regression [15] because it has the advantage of being a completely interpretable model and highly predictive with highly correlated variables with the default as the credit history. Authors such as Sohn et al. [16], and Vanderheyden and Priestley [17] present models with an acceptable performance but restricted to banked people since the variables are mostly from the financial history.

2.0.2 Alternative credit scoring

Traditional credit score models only consider those who have a history in the financial sector; people who have never accessed any financial service are invisible to the traditional financial sector. For this reason, in the last decade, the use of alternative data has grown significantly to strengthen the assessment of banked people and provide a credit evaluation for the unbanked. According to Aitken [18], the alternative data can be defined as non-financial information that constitutes the unbanked as a category of knowledge and offers a particular type of financing. It also classifies the data into three main groups non-financial payments, individual behavior, and data ingestion, these capture information on payment ability, individual behavior, and consumption behavior. For instance, Djeundje et al. [19] investigated the use of demographic variables, use of email, and psychometric variables for assigning credit scores in a lending platform. The authors found that accuracy was high enough to assess individuals when considering such alternative data when conventional credit history data was unavailable.

The use of alternative data along with machine learning techniques has been shown to have superior results to conventional models with traditional data. First, Berg et al. [20] show how the digital footprint left by the interaction of individuals with a website improves the prediction of default for banked people and gives the opportunity to effectively evaluate users without prior scores. Moreover, Zhang et al. [21] found that by merging traditional information with information from the social networks of users of a P2P network, traditional credit scores are exceeded. Likewise, Giudici et al. [22] proposed augment traditional credit scoring methods using alternative data similarity networks in a peer-to-peer network finding that centrality measures from the networks improve accuracy and the model explainability.

Efforts to find the best way to evaluate people financially have also led different authors to explore the impact of Deep learning algorithms. For instance, Han et al. [23] makes use of people's geographic footprints to first understand the credit characteristics of the regions through graphic convolutional networks and then evaluate the credit score using the trajectories of each individual. The results show that this approach can increase the accuracy of the credit evaluation compared to the reference methods. Likewise, Zhu et al. [24] proposes a model that combines the Relief algorithm for features selection with convolutional neural networks, finding that the proposed methodology exceeds the performance of models such as logistic regression.

Furthermore, in recent literature, there is a new source of information provided by graphs. This last form of representing data allow to identify behaviors and relationships between entities in a system, and several authors have demonstrated the contribution to credit risk assessment. First, Óskarsdóttir et al. [25] presents a credit score model in which it complements financial information with data from the call network and the study of the spread of default in it. In this way, the results show that the use of network effects, even on its own, could lead to predictive results as accurate as those provided by traditional information. The same authors, in the work [26], study correlated default (probability of multiple defaults occurring) in a multilayer network by using a personalized PageRank algorithm that ranks nodes with respect to a source of influence and can measure the effect of default across networks. With network theory, they find that propose networks provide valuable insights for credit risk propagation hence for credit assessment. Also, Westland et al. [27] investigated how by using a network based information of communications and travels of borrowers, peer-to-peer credit scoring can improve. Finding that graph topology is an essential predictor of loan profitability and risk assessment as good borrowers tend to be more connected to the rest of the graph than are borrowers who are likely to default.

Given the value that the graphs have generated for credit assessment, different studies have used different methodologies directly on the graphs to have more robust predictions that take advantage of the graph directly. For example, Lin et al. [28] propose an industrial-scale distributed network representation framework in which, based on different user interactions within Alipay platform, define unsupervised and supervised network representation for default prediction. The unsupervised network representation methods capture global information from the graph, while the supervised methods local information. Furthermore, Liang et al. [29] proposes a new learning approach to define credit risk and credit limits based on the Multi-view-aware Mixture-of-Experts network, which in one of its components uses a graph neural network to capture information from the users social network to improve performance and giving greater interpretability to the credit score prediction.

Chapter 3

Preliminaries

3.0.1 Graph theory

A graph is defined as $G = (V,E)$ where V is a set of nodes or vertices and E is a set of edges connecting pairs of nodes. When a direction is defined between a pair of nodes, that is $(V_1,V_2) \neq (V_2,V_1)$, then the graph can be defined as a directed graph where the edges indicate a one-way relationship, otherwise it is defined as an undirected graph and $(V_1,V_2) = (V_2,V_1)$. Graphs can be represented by a list of vertices $V = \{V_1,V_2,V_3,V_4\}$ and edges $E = \{(V_1,V_2),(V_2,V_3),(V_2,V_4),(V_4,V_1)\}$ or by means of a squared adjacency matrix, $A_G = \{a_{ij}\}$, which contains a row and a column for each vertex. In this matrix, if the vertex V_i is adjacent to the vertex V_j then A_{ij} takes a value greater than zero. The values within the adjacency matrix can be understood as the weights of the edges. The diagonal of the adjacency matrix is 0 unless the graph contains self-loops, and the adjacency matrix is symmetric if the graph is undirected.

The nodes represent objects or entities within the graph; hence, different nodes can be contained in the same graph. When there is only one type of node, it is known as a monopartite graphs, when it has two or more they are called bipartite or multipartite graphs. A bipartite graph $G = (V_1 \cup V_2, E)$ has two disjoint subsets of vertices $\{V_1(G),V_2(G)\}$ and the edges only join vertices of different subsets, no edges are joining two elements of V_1 or V_2. The adjacency matrix for a bipartite graph has a dimension of $M \times M$ where $M = |V_1| + |V_2|$ and its a block matrix where $V_1 \times V_1$ and $V_2 \times V_2$ are 0.

For graph analysis, there are two groups of measures and algorithms, graph-level and vertex-level measures. In the first group, measures refer to the global properties of the graph. In contrast, the second group measures indicate properties of the vertices, the scope of this research is focused on the last group and specifically in degree, eigenvector centrality, PageRank, and Louvain community measures.

Degree

The degree of a vertex $d(v_i)$ is the total number of edges connected to the vertex i. In a directed graph, the degree is defined as the sum of the in-degree and out-degree, which are the number of inward edges going into v_i and outward edges originating v_i. Thus, the degree captures the measure of connectivity of a vertex as well as its centrality since the higher the degree, the more central the vertex is.

$$d_i = \sum_j a_{ij} \forall i \in V$$

Eigenvector centrality

The eigenvector centrality measures the importance of vertices for the connectivity of the graph considering the importance of the neighbors [30]. For a vertex i, eigenvector centralitry is define as $x_i = \frac{1}{\lambda} \sum_k a_{k,i} x_k$ where λ is a constant. In vector notation the above expression can be rewritten as the eigenvector equation $\mathbf{Ax} = \lambda \mathbf{x}$, the vector $\mathbf{x}$ is an eigenvector of A corresponding to λ.

Pagerank

Pagerank algorithm is another centrality measure, initially created to rank web pages in search engines, that captures the influence of a vertex in the graph [31]. It is computed by randomly traversing the graph and counting the frequency of reaching each vertex. The rank of vertex i is define as

$$r_i = (1-d) + d \sum_{v_j \in B_i} \frac{r_j}{l_j}$$

where d is a constant between 0 and 1 known as damping factor, v_j is a vertex in the neighbor linking to v_i, l_j and r_i are the outgoing edges and the rank of v_j

respectively. When writing the systems of equations in a matrix form, we have that

$$\mathbf{r} = (\mathbf{I} - d\mathbf{H})^{-1}(1-d)\mathbf{I}$$

and, since the algorithm is iterative, the final Pagerank scores are obtained through the power iteration method, that follows the next scheme

1. Initialize he ranking vector $\mathbf{r}^0$ as $\frac{1}{|V|}$ for all of the components.

2. Iterate $\mathbf{r}^{t+1} = \mathbf{M} \cdot \mathbf{r}^t$ where $\mathbf{M}$ is the stochastic adjacency matrix.

3. Stop iterations when $\mathbf{r}^t$ converge, $|\mathbf{r}^{t+1} - \mathbf{r}^t| < \varepsilon$

Louvain

In a graph, communities can be formed, and these are usually identified since the connectivity between the inside vertices is higher than the rest of the nodes in the graph. The Louvain method for community detection is an algorithm that seeks to optimize the modularity of the graph [32]. Modularity is defined as

$$Q = \sum_{i=1}^{c} (e_{ii} - a_i^2)$$

where e_{ii} is the percentage of edges in the module i and a_i is the percentage of edges with at least one end in module i.

$$e_{ii} = |\{(u,v) : u \in V_i, v \in V_i, (u,v) \in E)\}|/|E|$$

$$a_i = |\{(u,v) : u \in V_i, (u,v) \in E)\}|/|E|$$

High modularity means more edges within the module than expected by chance; thus, the quality of the given division communities is high. In order to maximize modularity, the Louvain method has two phases. First, each vertex i is extracted from its community and inserted into the community of each neighbor j of i. For each of the neighboring communities of i to which it was inserted, the change in its modularity is calculated [formula 1], and i is inserted in the one with the highest value. This process is repeated for each vertex until there is no change in the modularity of any community. The next stage of the Louvain Method is to define a new network where the vertices are the communities defined in the previous phase and the relationships within the communities are understood as self-loops, and the relationships between communities is a weighted of the relationships of the original nodes.

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

Shortest Path

Path algorithms are one of the most studied in graphs, these methods seek to find the shortest or longest path between a pair of nodes $\min dist(u,v)$. There are different algorithms that, according to the need, traverse the graph and consider its characteristics differently.

Local Neighbor Feature Aggregation

The aggregation of neighbors features starts from exploring with graph queries the characteristics of the neighborhood of the nodes. The i-th neighborhood $N_i(v)$ of a node v is the set of nodes adjacent to v that are at a distance i. The concept of graph queries is the simplest graph analytics process in which it is sought to identify patterns or characteristics of the graph. Specifically, the aggregation of features consists of, starting from a node v_j and its neighborhood $N_i(v_j)$, adding the characteristics of interest of the neighbors to capture information of the local context of the node.

3.0.2 Graph Neural Networks

Graph Neural Networks (GNN) is a class of deep learning methods designed to make inferences about graphs. These methods are part of the machine learning branch, Geometry Deep Learning (GDL), which seeks to generalize deep neural networks to non-euclidean data. Traditional neural networks work for structured data such as images with a defined structure, however, for complex structures with graphs or manifolds, the proposed models are not generalizable. Next, an introduction to the graph neural networks is presented, followed by different variations proposed in the literature for classification problems.

As defined above, a graph is an arrangement of nodes V with relationships between them represented by an adjacency matrix A. In addition, each node i can contain a set of features x_i and a label l_i so the definition of a graph is also defined as $G = (V, A, X, L)$ where X is an $n \times |V|$ matrix of features where n is the number of features and $|V|$ the number of vertices and L is a vector of labels of length $|V|$. It should be noted that the vector can have a dimension c, where $c < |V|$, in the case that not all nodes have a label assigned and the nature of the problem is semi-supervised.

The basic concepts behind graph neural networks are neighbor aggregation and message passaging. The representation (embedding) of a node is defined by the aggregation of its features and those of its neighbors.

3.0.2.1 Graph Convolutional Neural Networks

Graph convolutional networks, defined by Kipf et al. in [33], are multi-layer feed-forward neural networks that propagate and transform node features across a graph, its propagation rule is given by

$$X^{(k+1)} = \sigma\left(D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}X^{(k)}W^{(k)}\right),$$

where $W^{(k)}$ is a trainable weight matrix in the k-th layer, $\sigma(\cdot)$ is an activation function, $\tilde{A} = A + I$, and $X^{(k)}$ is the k-th layer node representation where $X^{(0)} = X$.

3.0.2.2 GraphSage

GraphSage is a non-spectral approach that computes node representations in an inductive manner. It was introduced by Hamilton et al. in [34], as an alternative to many of the transductive methods in the literature. Graphsage considers the neighborhood of each node, and then performs a specific aggregator over it, the result then fed into a recurrent neural network in order to propagate information between different layers of the model. This can be expressed as

$$\begin{aligned} x_{\mathcal{N}_i}^{(k+1)} &= \text{aggregate}\left(\{x_j^{(l)} : j \in \mathcal{N}_i\}\right) \\ \tilde{x}_i^{(k+1)} &= \sigma\left(W \cdot \text{concat}(x_i^{(k)}, x_{\mathcal{N}_i}^{(k+1)})\right) \\ x_i^{(k+1)} &= \text{norm}(\tilde{x}_i^{(k+1)}). \end{aligned}$$

In practice the aggregator is taken to be the average, and the normalizer to be normalization with respet to the l^2 norm.

3.0.2.3 TAGCN

Topology Adaptive Graph Convolutional Networks (TAGCN) were introduced by Du et al. in [35], these networks are a vertex domain approach to the convolutional neural network problem on graphs where the graph convolution operation is defined in terms of a shift-invariant filter, this filter can be expressed as as a multiplication by polynomials of the graph adjacency matrix. More explicitly,

$$x^{(k+1)} = \sigma\left(\sum_{c=1}^{C_k}\sum_{j=0}^{K} g_{c,j}^{(k)}\left(D^{\frac{-1}{2}}AD^{\frac{1}{2}}\right)^j x_c^{(k)}\right),$$

where σ is a ReLU activation function, C_k is the dimension of the k-th layer, K is the degree of the filter as a polynomial of the adjacency matrix, and the $g_{c,j}^{(k)}$ are the coefficients of said polynomial. The reader should notice that $x_c^{(k)}$ is the vector corresponding to the entries of c-th feature of each vertex in the k-th layer of the network.

3.0.3 Supervised learning model

In supervised learning, there is a set of P predictor variables $[x_{1i}, x_{2i}, \ldots, x_{1P}]$ and a label y_i for each observation i, the objective of supervised algorithms is to find the function $\hat{f}(x)$ that best approximates $y = f(x)$ by minimizing the loss function $L(y, f)$. There are many supervised models, both linear and non-linear, used for regression and classification problems. However, in recent years the Extreme Gradient Boosting or XGBoost algorithm has proven to be a super algorithm with good performance in different problems and competitions [36]. The general idea of the algorithm is based on a sequential assembly of decision trees (CART) to learn from the result of the previous trees and correct the error produced by them.

Ensemble is a technique used in machine learning and consists of adding weak learners so that their predictive power is combined. Two very well-known ensemble learning methods are bagging and boosting. The second is the one used by XGBoost and, therefore the reason for part of its name. In boosting, trees are built sequentially so that each subsequent tree aims to reduce errors from the previous trees. Therefore, an initial model $\hat{f}(x)$ trained to predict y is defined, which has the residuals $r = y - \hat{f}(x)$ and the following model $\hat{f}^b(x)$ is fit over this residuals. The updated $\hat{f}(x)$ model is defined as $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$ where λ is a parameter known as learning rate. The residuals are also updated as $r_i \leftarrow r_i + \lambda \hat{f}^b(x_i)$ and the final boosted model is

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

In addition, the high predictive power of XGBoost is leveraged by the use of the Gradient Tree Boosting algorithm, because although this algorithm also uses boosting to build a model, the residuals or pseudo-residuals can be obtained by

$$r_i m = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i))}\right]$$

Thus, trees $\hat{f}^{*b}(x)$ are fitted for the residuals of each iteration and an optimal coefficient γ is also defined to get a final boosted tree model $\hat{f}^b(x) = \hat{f}^{b-1}(x) + \gamma \hat{f}^{*b}(x)$.

3.0.4 Financial evaluation

The models that are usually used to make a credit evaluation do not consider the costs associated with a wrong prediction, both the cost of accepting a customer who is a defaulter and that of rejecting a good payer customer are considered equal. Correa Bahnsen et al. [37] presents a cost-sensitive approach for credit risk assessment where the cost of a wrong prediction is not constant among the observations and is evaluated as presented in Table 3.1.

The cost of correct predictions is zero while the cost of a false negative C_{FN_i} is

Table 3.1: Credit scoring example-dependent cost matrix

	Actual Positive $y_i = 1$	Actual Negative $y_i = 0$
Predicted Positive $c_i = 1$	$C_{TP_i} = 0$	$C_{FP_i} = r_i + C^a_{FP}$
Predicted Negative $c_i = 0$	$C_{FN_i} = Cl_i \cdot L_{gd}$	$C_{TN_i} = 0$

the loss over the credit line define as the credit line Cl_i times the loss given default L_{dg} and the cost of a false positive C_{FP_i} is the sum of the profit r_i when declining a good payer plus the cost of giving the credit to another costumer C^a_{FP}. The profit generated by a consumer i is defined as $r_i = PV(A(Cl_i, int_{r_i}, l_i), int_{cf}, l_i) - Cl_i$, where the first term are the financial institution gains and expenses with lending rate int_{ri}, cost of funds int_{cf} and monthly payments A represented in present value by means of PV as below.

$$A(Cl_i, int_{r_i}, l_i) = Cl_i \frac{int_{r_i}(1+int_{r_i})^{l_i}}{(1+int_{r_i})^{l_i} - 1},$$

$$PV(A, int_{cf}, l_i) = \frac{A}{int_{cf}} \left(1 - \frac{1}{(1+int_{cf})^{l_i}}\right)$$

For the term C^a_{FP}, the assumption is that since the credit is not granted to the client i then it will be granted to an alternative client who will have an average line of credit $\overline{r}$ and an average profit $\overline{u}$ that are weighted by the probability of payment pi_0 and default pi_1 respectively.

$$C^a_{FP} = -\overline{r} \cdot \pi_0 + \overline{Cl} \cdot L_{gd} \cdot \pi_1$$

Once the cost-matrix is defined, the cost improvement can be expressed as the cost savings as

$$Savings = \frac{Cost_l - Cost}{Cost_l},$$

where $Cost$ is calculated as the sum of individual costs matrices for each customer

$$Cost = \sum (1-c_i) * y_i * C_{FN_i} + (1-y_i) * c_i * C_{FP_i},$$

and $Cost_l$ is the cost of the cost-less class [38] overall results of a model.

Chapter 4

Methodology

In this section, we aim to describe the data used in this research as well as the methodology for the construction of tabular and graph-based features. Moreover, we present the proposed graphs and their characteristics.

4.0.1 Data

The data for the development of this research was provided by a Latin American Super-App that, as mentioned above, is granting new financial products such as credit cards. Since the Super-App does not yet have its own data on the defaulted payments of its users, the labels of whether a user is defaulted or not were provided by a financial partner. The sample of users considered consists of 38,342 observations of the Super-App's active users with financial products with the partner; the default rate is 12.9%.

Now, the data provided by the Super-App can be classified into two groups: transactional data and data from interactions or relationships. From this data, all the predictor variables that hoped to model financial aptitude were constructed. From the first group, 149 features were built, of which only 41 were considered at the end due to feature selection, these variables are entirely behavioral and transactional. From the information of the second group, five different networks are built, and different variables that could indicate a user's creditworthiness were obtained based on graphs. In Figure 4.1 a diagram of the data and the construction of features is presented, in the following subsections each of these is detailed.

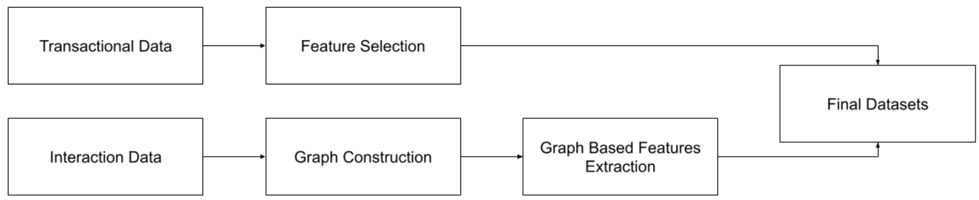

Figure 4.1: Data processing

4.0.2 Tabular features

From the transactional data of the users of interest, 141 variables are constructed that can be divided into demographic, transactional, and behavioral variables.

Demographic: variables that allow making a socioeconomic profile of the user are considered, in this way to be able to have a proxy for variables that capture the financial capacity from assets such as cell phone brand, place of residence, number, and score of registered cards, number of addresses for order delivery, internal segmentation, among others. The total of constructed variables are.

Transactional: variables that capture consumption in verticals of food delivery, groceries, liquor, pharmacy, and e-commerce. From these variables, it is sought to describe the consumption patterns and preferences of the users. Some of the variables are the number of total orders, amount spent, use of discounts, organic orders, canceled orders, orders canceled due to payment error, fraudulent orders, among others.

Behavioral: variables that describe the user's application usage behavior such as frequency of purchase, recency, most used vertical, most used payment type, weekend or night purchases, churn probability, among others. These variables seek to identify behaviors that differentiate bad payers from good ones.

Starting from the 141 defined tabular variables, it was decided to make feature selection to use in the models only those that capture the most predictive information to reduce the computational cost of the modeling and even improve the performance of the models. To this end, we use three different methodologies to find highly informative variables and then select those variables that have been significant or selected by at least two of the three methods, as shown in figure 4.2. Finally, drop high correlated variables is decided, use a Filter method and use Boruta Algorithm [39].

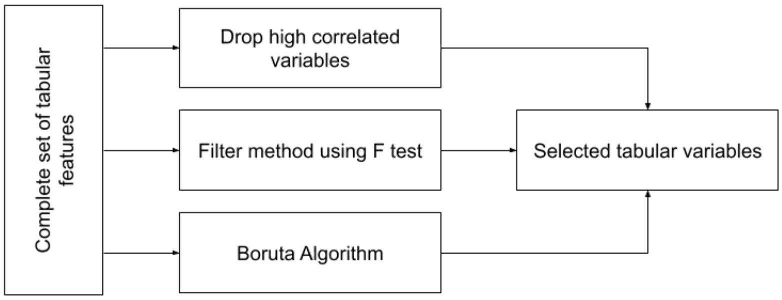

Figure 4.2: Tabular feature selection

Using the first methodology, the features that had a correlation greater than 0.9 are dropped, so no variables have redundant information. Moreover, the statistical measure selected to score the variables by a filter method is the F-score, with this the 50 variables with the highest score are selected. Finally, using Boruta, an algorithm in which it randomly reproduces values of the different predictor variables and then compares them with the original variables to measure the importance of the variable. Only the variables of greater importance than the random ones are considered important, in total there are 78. After identifying the variables selected by each method, we proceed to select only those that at least two methods considered necessary.

4.0.3 Graph construction

With the aim of making the most of all the information generated by the Super-App, six graphs are constructed from the interaction data, in which only one type of relationship is considered in 5 and all relationships in the last one. In the construction of the graphs, not only the relationships between the 38 thousand users were considered, but also the connections to users who did not have the defaulter label. As graphs had to capture the complete environment of the users to have the best approximation. Below is the detail of each graph

1. **P2P**: The graph of peer-to-peer transactions in the virtual wallet platform of

the Super-App. This graph is a directed graph on the set of users, where for each edge its source is the user who initiated the transfer, and its direction points towards the user who receives the transfer. When considering all of the interactions of 38,342 users, the network ends up having 88,270 vertices and 214,637 edges.

2. **BIN**: In hopes to better understand users with similar socioeconomic status, we decided to construct a graph that allowed us to capture information from users with similar financial products. We construct a graph that relates users to the BINs, where we only kept those users and BINs inside the population of interest and those users that used any of said BINs at least once in October 2020. The graph has as nodes the BINs and the users, and the graph's edges are directed edges from users to BINs. The graph has 901,366 users, 9,096 BINs, and 1,646,201 edges.

3. **GEO**: To capture users with similar sociodemographic data, we decided to construct a graph whose vertices consisted of geohashes and users, and whose edges connected users with the geohashes from the delivery addresses used inside of the App. As with the BIN-graph, we then restricted the graph to those users and geohashes related to our population of interest together with the users who had been in said geohashes at least once in October 2020. This graph has 276,260 users y 34,224 geohashes and 1,104,142 edges.

4. **CC**: In order to capture users with similar financial background, we considered a graph that consists of users and credit cards, where the edges connect a user with one of their registered credit cards. Inside of a Super-App, it is not uncommon for different users to have registered the same credit cards, in order to consider the financial behavior of those users that may share payment methods with the population given to us by the Super-App ally, we decided to inductively grow our network. The graph ended up having 136,009 users, 385,014 credit cards and 634,870 edges.

5. **DV**: We construct a graph whose vertices consist of users and devices, and

where the edges relate users with their registered devices. The graph was constructed in a similar way to the CC-graph. This graph has 247,844 users, 385,014 devices, and 707,948 edges.

6. **MIX**: In a final graph, in order to combine all the information of the users, a network that considers all the relationships and nodes previously stated is defined. That is, a graph with a total of 2 million nodes from the five types of nodes (users, bines, credit cards, geohash, and devices) and 4.1 million relationships resulting from relationships by P2P transfers, have a device, have a credit card, share geohash and bines in common.

A summary of structure details of these graphs is presented in Table 4.1. Also, Figure 4.3 presents a small portion of the graphs in this order P2P transfers, credit cards, devices, BINs, geohash and MIX. Users are represented by pink (defaulters) and orange nodes (non-defaulters), credit cards, devices, BINs, and geohash are shown in red, green, blue, and purple, respectively.

Table 4.1: Graphs structure details

Graph	Nodes	Edges
P2P	88,270	214,637
Credit Card	576,042	634,870
Devices	632,858	707,948
Bines	910,431	1,646,201
Geohash	310,484	1,104,142
Mix	2,064,245	4,193,862

4.0.4 Graph based features

From these networks, we computed what we call *graph-features*, which include centrality, community, and neighborhood variables, for each of the users the population of interest:

1. *Total degree*: Total number of connections of a node.

2. *Eigenvector centrality*: Score for each node based on the importance of its neighbors [30].

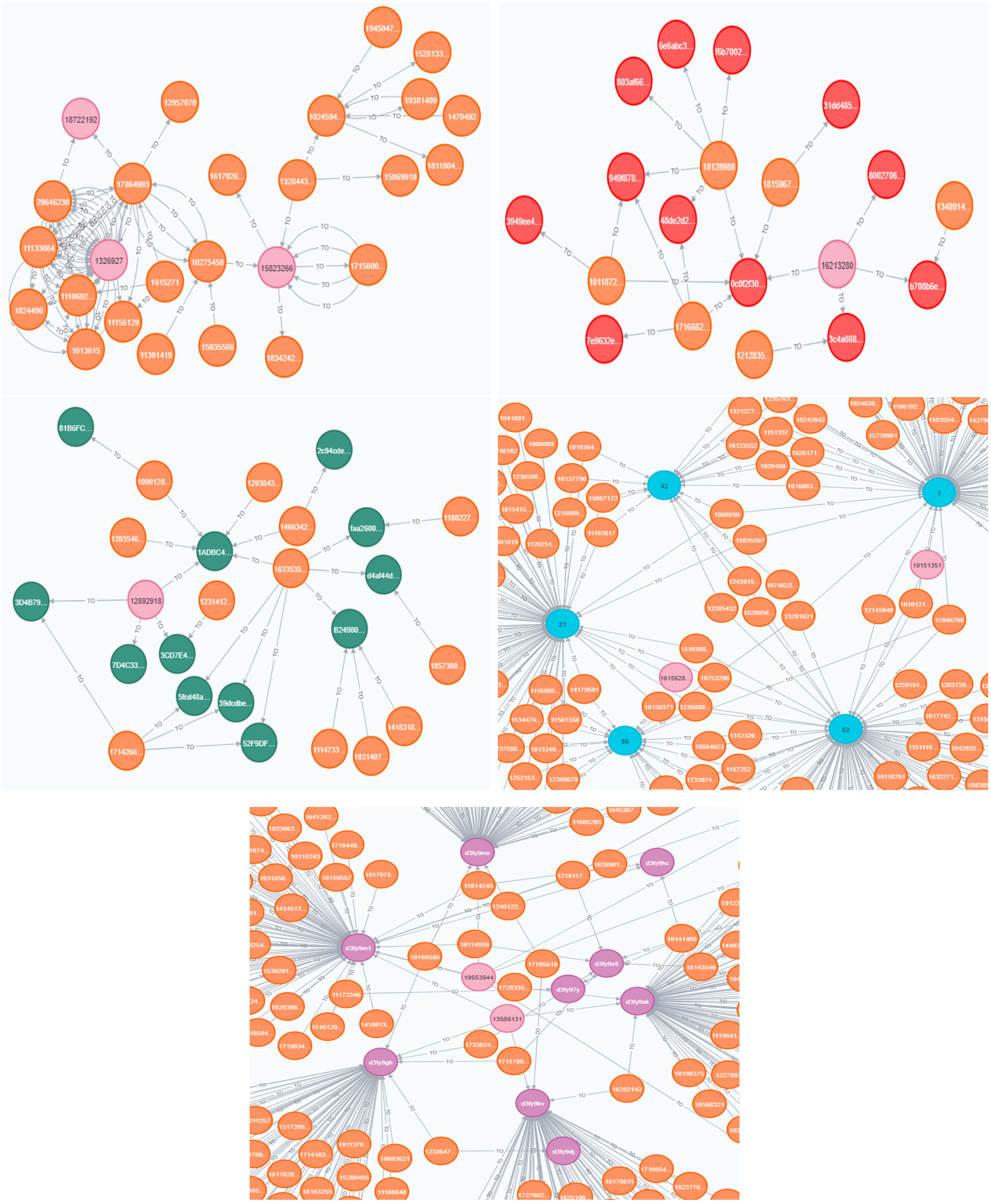

Figure 4.3: Proposed graphs examples

3. *Pagerank*: Score that measures the importance of a node by randomly traversing the graph and counting incoming and outgoing edges each node weighted by their importance [31].

4. *Louvain*: It is a method for extracting communities from a graph and identifying which nodes belong to them [32].

5. *Shortest path*: Algorithm that finds the shortest path between a pair of nodes. In this case, the shortest path from the user nodes to any defaulter node was calculated.

6. *Average neighbors features*: In this category, for each user we compute the averages of certain variables of the Super-App over their neighborhood. For instance, in the P2P network for user A that is connected to users B and C, an example of one type of neighborhood variable of A is the average of the orders cancelled with credit card of users B and C. It is considered important to obtain information from the neighborhood as it allows to capture information on the possible behavior of the user from their peers. The variables selected are: orders canceled due to payment error, orders paid with credit card, maximum credit card score, and if the users is prime.

Chapter 5

Experiments

This section presents the different experiments that are divided into tabular and GNN models. First, a traditional machine learning approach is taken, and a classification model is used to predict whether a user will default. In the second, the graph neural networks framework is used to run models directly on the graphs for node classification.

5.0.1 Tabular model

For the tabular methods, we used an XGBoost [36] with five randomized bootstraps, trained with 70% of the data and tested with the remaining 30%. First, a *tabular based model* is defined in which only the 41 tabular features of the Super-App are considered, which will allow to see the statistical and financial gains of the variables based on graphs. Then, define a *graph based tabular model* where only the features absorbed in graphs will be considered, this is repeated six times, one for each defined graph. Finally, a *hybrid tabular model* is proposed in which the mix of tabular and graph-based features is considered for each graph.

Each of these models is evaluated for its performance with the ROC AUC score and KS score. On the one hand, the AUC is one of the most used measures in classification problems as this evaluate the discriminating capacity of a model by measuring the area under the curve by the tradeoff of true and false positives at different thresholds. Likewise, the AUC has the advantage that it is not highly affected by imbalanced data such as accuracy. On the other hand, the KS statistic measures the maximum difference between two cumulative distributions (true positive rate and

Table 5.1: Parameters to estimate the financial savings

Parameter	Value
Interest rate (int_r)	40%
Cost of funds (int_{cf})	10%
Loss given default (L_{gd})	75%

false positive rate), that is how able is the model to discriminate between events and non events. Furthermore, to see whether there is a substantial difference in model performance when considering graph-based features, statistical tests are performed. Finally, the models are also evaluated with the cost sensitive method presented in section 3, the parameters considered are presented in Table 5.1.

5.0.2 GNN model

For the experiments based on the neural network graph framework, four different types of semi-supervised models in the P2P, credit cards, and devices graphs considering only numerical variables of the Super-App were used. The graph neural networks used are developed for homogeneous graphs, therefore, for the credit card and device networks, a preprocessing of folding is done so that the networks are only with user nodes and a relationship represents whether they share a card or device. Figure 5.1 specifies the folding process, wherein the heterogeneous network on the left, the pink nodes represent users and the blue nodes represent any other entity like credit cards or devices, the relationships between them are understood as user A has entity A. Once the network is folded, it is a homogeneous graph as it only has one type of node and an edge exists if two nodes share an entity. Moreover, Table 5.2 specifies the sizes of the new homogeneous networks.

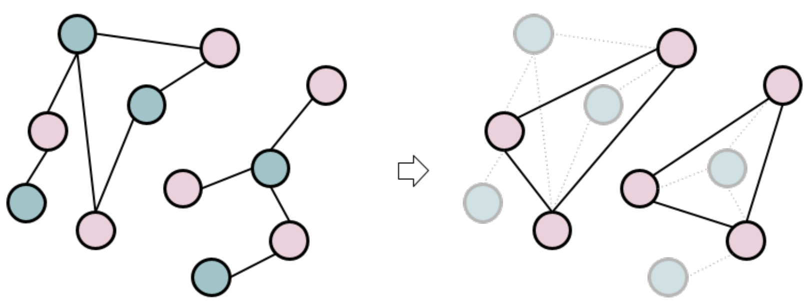

Figure 5.1: Graph folding

Table 5.2: Homogenous graphs structure details

Graph	Nodes	Edges
Credit Card	136,009	1,364,150
Devices	247,844	42,494,824

The growth of the number of edges in the homogeneous graph is due to the nature of some credit card and devices nodes with a high degree. These nodes are considered super nodes because they have a large number of connections, and when folding the graph, all possible combinations between their immediate neighbors are considered. For example, if a device node has 10 connected users, then the number of connections goes from 10 in a heterogeneous graph to 45 in a homogeneous graph. These nodes were not limited or eliminated as it is considered that they can capture dynamics and essential information.

Furthermore, each experiment uses the same GNN structure, except for the type of layer and the additional parameters that each needs. In general, all GNNs have two layers: a hidden layer with a ReLU activation feature and 16 neurons and an output layer with a size of 2. We train our models with respect to a cross-entropy loss function, the learning rate is 0.02 and are trained with 200 epochs. Given that the data is imbalanced, we decided to weight the cross-entropy loss so that we could improve the precision of our models.

Chapter 6

Results

In this chapter we present the experimental results of each of the proposed models.

6.0.1 Tabular Model

The average performance obtained from the tabular base model is 0.684 in AUC and 0.280 in KS. Regarding the results obtained in the graph-based tabular model, in general, it is identified that when considering only variables of the graph, the performance does not have a better performance than the base model. Figure 6.1 presents AUC results, only the BIN network seems to capture enough information, such as transactional variables with an average AUC equal to that of the base model. On the other hand, the results obtained with the KS metric are similar as can be seen in Figure 6.2, where the BIN graph seems to give the best performance compared to the other graphs with a KS statistic of 0.287. The fact that the BIN graph of a similar performance to the base model indicates the predictive power of the graphs, and this may be explained by the fact that the BINs incorporate financial information and internal segmentation of traditional financial entities.

The results obtained for the hybrid models are presented in Figure 6.3 and Figure 6.4. Overall, the proposed graphs improve the performance of the tabular base model, particularly the BIN graph and the MIX graph allow better prediction of defaulters in terms of AUC and KS statistic. First, the gain when considering variables based on the BIN graph is 0.037 in AUC and 0.052 in KS compared to the tabular based model. The latter indicates that although the tabular variables consider information based on the credit cards registered by the user as a card level, the number of

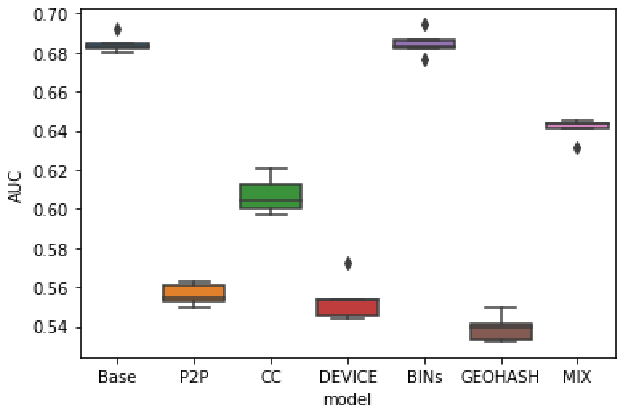

Figure 6.1: AUC performance by graph based tabular model

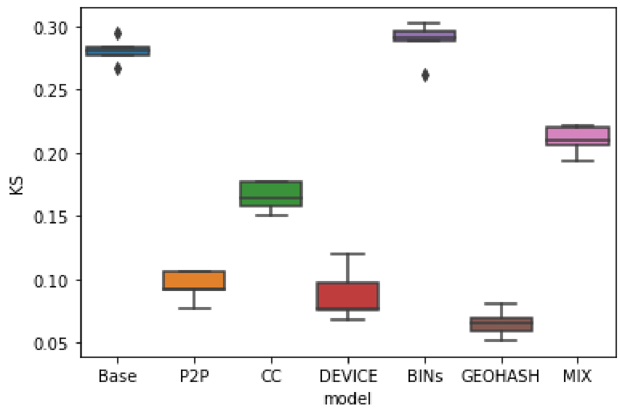

Figure 6.2: KS performance by graph based tabular model

cards, and card scores, the graph manages to capture new and important information to evaluate creditworthiness. Moreover, although the MIX graph considers the connections of bins, this has a lower performance than the model that only considers BIN graph features, which could be explained by the fact that when adding relationships that are not informative, the different graph algorithms or graph queries capture fewer signals of bad or good payers. Nevertheless, the MIX graph improves in 0.023 and 0.036 concerning the tabular base model in AUC and KS statistics.

In order to identify those graphs that provide a statistically significant gain concerning the tabular based model, Mann-Whitney non-parametric mean tests are performed. Table 6.1 shows the p-values of the Mann Whitney test for both AUC and KS statistics. Overall, the models that consider variables based on the bin graphs and the MIX graph generate a statistically dignifying gain in the two metrics.

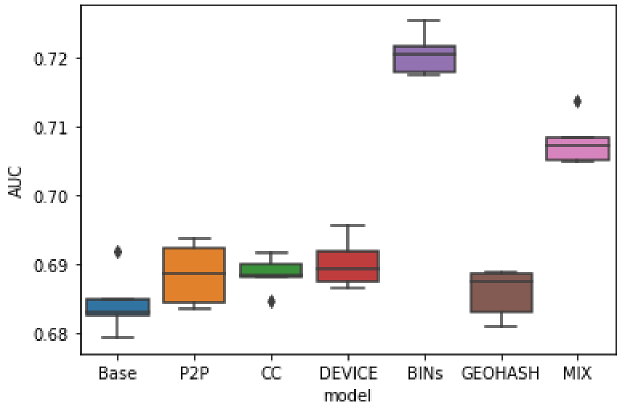

Figure 6.3: AUC performance by hybrid tabular model

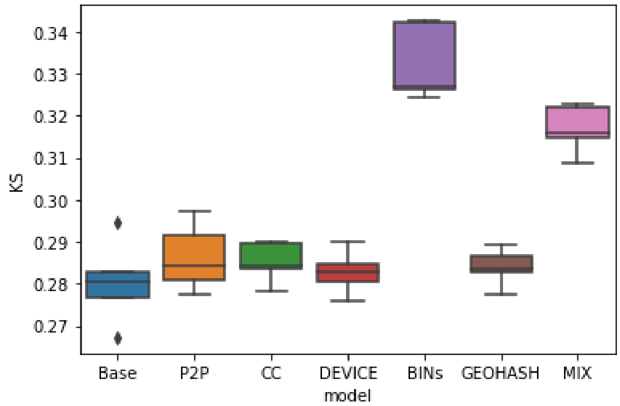

Figure 6.4: KS performance by hybrid tabular model

Moreover, the devices graph also gives a significant gain of AUC, although in the previous box plot it is not clear, the performance of this hybrid model is 0.690.

Graph	AUC	KS
P2P	0.0718	0.1481
CC	0.1050	0.1481
DEVICE	**0.0473**	0.3371
BINS	**0.0060**	**0.0060**
GEOHASH	0.2654	0.2016
MIX	**0.0060**	**0.0060**

Table 6.1: Mann Whitney test P-Value for performance metrics Hybrid Model

Furthermore, a model with a good statistical performance does not necessarily imply that it also has a good financial performance. Figure 6.5 shows the savings obtained by [37] example-dependent cost-sensitive methodology, as can be seen all the graphs except that of P2P transactions exceed the average savings of the based tabular model. Although the best financial improvements are given again by the BINS and MIX graphs, in this case a considerable increase is also seen in the models that consider the credit cards, devices, and geohash graphs. Thus, it can be argued from a financial point of view that all graphs except P2P are options that offer financial benefits and some statistics.

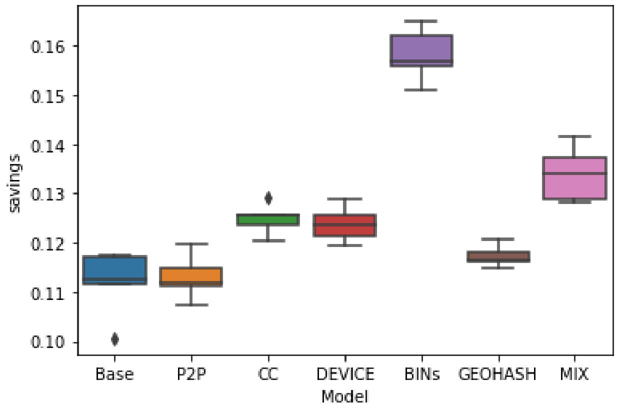

Figure 6.5: Financial savings

Since we used a black-box model, SHapley Additive exPlanations [40] are used to identify the feature importance and the effect of each of the variables in credit risk assessment. Figures 6.6 and 6.7 present the SHAP important features of the tabular hybrid model with the BINs and Mix graphs. In this plot, features are sorted on the y-axis according to their relevance, and the x-axis measures the impact on the model's output according to the value taken by the variable, red a high value and blue a low value. Thus, SHAP allows a local interpretation of how the probability of default increases or decreases from the variable's value.

Regarding the hybrid model with features of the BIN graph, the five most influential variables in model two are from the graph. First, the PageRank centrality variable captures the most information on defaults and implies that the higher centrality, the greater the probability of default. This centrality measure indicates that the importance within the graph given by the connections of users with BINs are indicators of non-payment and that the BINS contain information on the segmentation and information that traditional entities have. Moreover, the centrality variable calculated by eigenvector centrality, the fourth most important variable, also seems to capture a non-linearity with the probability of default since it is observed that low values are concentrated in the center of default the x-axis and high values in the extremes. Furthermore, the transactional variables with the most significant influence are related to the use and characteristics of registered credit cards. Orders with payment errors, for instance, is the second variable and is interpreted as the more errors in payments, the greater the probability of default, this being an early indicator of lack of liquidity. Likewise, variables such as the maximum score of the credit card, the ownership score of a card (whether it is own or not), and the percentage of orders paid with a credit card have similar behavior, as they take high values less probability of non-payment; this is can be interpreted as experience with financial products and their use.

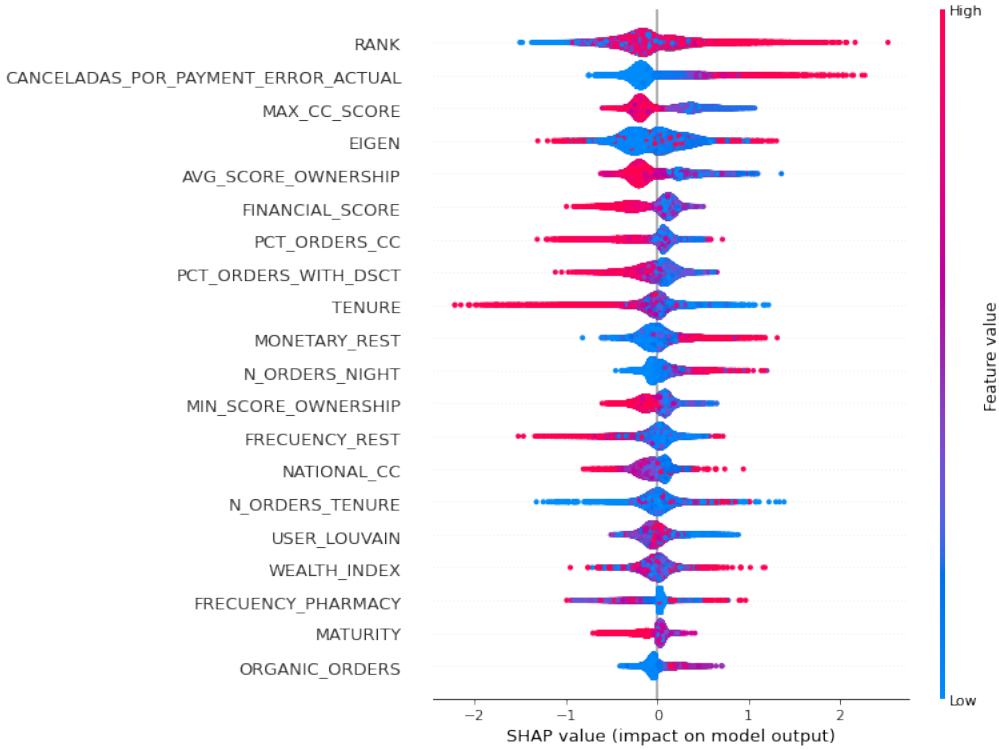

Figure 6.6: Feature importance tabular hybrid model BIN graph

Finally, overall there is similar importance in the hybrid tabular model with features of the MIX graph (Figure 6.7). Again the most important tabular variables are related to the registered credit cards and their use. Regarding the variables based on the graph, the centrality measured by eigenvector centrality is the most important, indicating that the greater the centrality, the greater the probability of default. In addition, the variable *U2_AVG_PCT_ORDERS_CC* that indicates the average of the orders paid with the credit card of the neighbors takes relevance and implies that as the neighborhood uses its card more, the lower my probability of default. Hence, as the level of banking in the neighborhood, so users in a context where there is greater use of the card can also replicate the behavior.

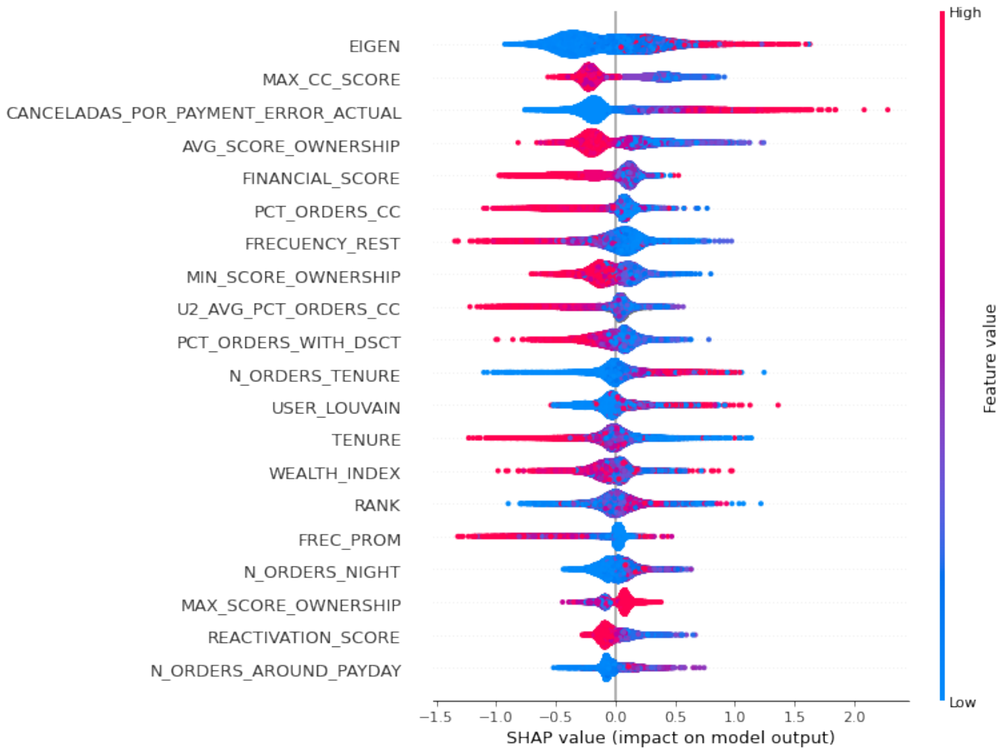

Figure 6.7: Feature importance tabular hybrid model MIX graph

6.0.2 GNN Model

The performance of each GNN are shown in Table 6.2. Overall, the GNNs that achieve the best performance in terms of AUC regardless of the network used are GraphSAGE and GCN, also the network with the highest predictive value when using GNN models for credit risk assessment is the P2P network. GraphSAGE gives the best performance of the P2P graph with an AUC of 0.67, which indicates that the aggregation of neighbor features is valuable as better representations can be generated to identify good and bad payers. For the credit card graph, the best AUC is 0.637 obtained through a GCN followed by GraphSAGE, the difference in AUC between these two models is not significant however, a marked difference between the weights of true negatives (sensitivity) and true positives (Specificity) is evident, where graphSAGE has greater power to identify bad payers and good payers GCN. The same is true for the network of devices in which this time the GCN gives the highest rate of sensitivity and GrapSAGE of specificity.

Table 6.2: GNNs performance

Graph	GNN	AUC	Sensitivity	Specificity
P2P	GraphSAGE	**0.6753**	0.649	0.701
	GCN	0.6571	0.677	0.636
	TAGCN	0.6441	0.600	0.687
Credit Cards	GraphSAGE	0.6319	0.641	0.622
	GCN	**0.6373**	0.577	0.696
	TAGCN	0.6234	0.636	0.610
Devices	GraphSAGE	**0.6169**	0.575	0.658
	GCN	0.6137	0.624	0.603

Chapter 7

Discussion

The results obtained from the tabular models demonstrate the additional value generated by the alternative data for credit evaluation, specifically those of centrality based on graphs and tabular variables related to the use of credit cards. This contribution becomes relevant for individuals with little or no financial record as considering models with comprehensive information and from different sources promotes financial inclusion by providing services to social groups traditionally neglected by traditional entities. These platforms then leverage economic growth and efficiency in the allocation of resources by facilitating access to credit. Furthermore, this alternative data allows them to complement their financial profile for individuals already banked, giving rise to a more robust creditworthiness assessment.

Another implication of the results obtained is the improvement of the operational efficiency of the Super-Apps since the variables based on graphs improve not only the statistical performance but also improve the economic performance. In this way, companies can maintain their appetite for risk and improve the costs of misclassification errors by selecting a credit scoring model that fits their costs and dynamics. Therefore, Super-Apps has a competitive advantage as it improves the offer of financial services and encourages product innovation for new population sectors traditionally excluded. The above supports the idea that smart data, machine learning open new growth paths for financial services [41].

On the other hand, super-apps have the opportunity to redefine credit risk since they are not limited to variables commonly used in credit scoring models. Henceforth, Super-Apps can define risk within the platform based on new interactions, behavior, and transactionality, creating profiles or scores that differ between defaulters and non-defaulters. Thus, determining solvency only with information and data from the platform generates value and advantage since it is possible to evaluate an individual independent of the banking history and attract more users due to its versatility and capacity for inclusion. Likewise, this allows Super-Apps to take advantage of their platforms to create a gamified score in which users can interact to improve their scores and adhere to the platform.

Finally, with the results, the idea of encouraging regulators to consider alternative Super-App data in credit evaluation models is supported likewise, that they promote regulations that include the use of alternative data for the alignment of these platforms and the information they imply. In turn, Super-Apps must design a compliance management program that fully complies with consumer protection laws and regulations. Robust compliance management includes adequate testing, monitoring, and controls to ensure that the issues are understood and addressed consumer protection risks [42].

.

Chapter 8

Conclusions

In this paper, the statistical and financial impact of alternative data is studied, particularly graphs leverage by a Super-App, for creditworthiness assessment through machine learning models. For this purpose, six different networks are presented based on different interactions within a Super-Apps, from which graph-based features are extracted. Additionally, we build tabular variables based on the behavior and transactionality of the users in the application. Then, feature selection is made, and two types of experiments are proposed, one using a binary classification model XGBoost and GNNS to take advantage of the structure of the graphs. Overall, it is shown that the alternative data generated provides useful information to strengthen the credit evaluation, in particular the data extracted from a graph of bins to capture insightful information about the payment trend in tabular models for the interpretation the SHAP additive values methodology is used. In addition, it is identified that the networks of BINs, device cards, and MIX generate savings when evaluating the model with a cost-sensitive approach, so it is encouraged to include the costs of business operation for a performance aligned with the business strategy. On the other hand, GNNs show lower performance, possibly due to the size of the dataset, given that these algorithms are 'data hungry' and the unbalanced nature of the problem. Finally, the role of Super-Apps in credit systems and the financial inclusion that they can provide is discussed, along with the possibility of redefinition of credit risk and regulatory implications.

Bibliography

[1] Ash Turnesr. 1 billion more phones than people in the world. Accessed 1/04/21.

[2] Betsy Atkins. Are super apps the future? Accessed 1/04/21.

[3] Asian Insights Office . Super apps in financial services: Business models and opportunities. Technical Report sector briefing 81, DBS Group Research, September 2019.

[4] T. Sun. *Balancing Innovation and Risks in Digital Financial Inclusion-Experiences of Ant Financial Services Group*, chapter 2, pages 37–43. 08 2017.

[5] Naeem Siddiqi. *Intelligent Credit Scoring*. John Wiley & Sons, Inc., January 2017.

[6] Charles B Wendel and Matthew Harvey. Sme credit scoring: key initiatives, opportunities, and issues. Technical report, The World Bank, 2006.

[7] The World Bank Group. Credit scoring approaches guidelines. Technical report, The World Bank Group, 2019.

[8] World Artificial Intelligence Conference. Lufax CTO Mao Jinliang: AI is reshaping the wealth management industry, Aug 2019. Accessed 2020-04-02.

[9] Luisa Roa, Alejandro Correa-Bahnsen, Gabriel Suarez, Fernando Cortés-Tejada, María A. Luque, and Cristián Bravo. Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications, 2020.

[10] Luisa Roa, Andrés Rodríguez-Rey, Alejandro Correa-Bahnsen, and Carlos Valencia. Supporting financial inclusion with graph machine learning and super-app alternative data. *arXiv preprint arXiv:2102.09974*, 2021.

[11] Rosa Puertas Medina and Maria Luisa MartÃ Selva. AnÃ¡lisis del credit scoring . *Revista de AdministraÃ§Ãde Empresas*, 53:303 – 315, 06 2013.

[12] Seyed Mahdi Sadatrasoul, Mohammad Gholamian, Mohammad Siami, and Zeynab Hajimohammadi. Credit scoring in banks and financial institutions via data mining techniques: A literature review. *Journal of AI and Data Mining*, 1:119–129, 04 2013.

[13] Experian. Credit score basics. tips for unlocking your credit potential, 2019. Accessed 14/09/20.

[14] Lyn Thomas, Jonathan Crook, and David Edelman. *Credit Scoring and its Applications*. SIAM, USA, second edition edition, 2017.

[15] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124 – 136, 2015.

[16] So Young Sohn, Dong Ha Kim, and Jin Hee Yoon. Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing*, 43:150 – 158, 2016.

[17] Bob Vanderheyden and Jennifer Priestley. Logistic ensemble models. *arXiv preprint arXiv:1806.04555*, 2018.

[18] Rob Aitken. 'all data is credit data': Constituting the unbanked. *Competition & Change*, 21(4):274—300, 2017.

[19] Viani B Djeundje, Jonathan Crook, Raffaella Calabrese, and Mona Hamid. Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163:113766, 2021.

[20] Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies*, Accepted for publication, 2019.

[21] Yuejin Zhang, Hengyue Jia, Yunfei Diao, Mo Hai, and Haifeng Li. Research on credit scoring by fusing social media information in online peer-to-peer lending. *Procedia Computer Science*, 91:168 – 174, 2016. Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management (ITQM 2016).

[22] Paolo Giudici, Branka Hadji-Misheva, and Alessandro Spelta. Network based credit risk models. *Quality Engineering*, 32(2):199–211, 2020.

[23] Xiao Han, Ruiqing Ding, Leye Wang, and Hailiang Huang. Creditprint: Credit investigation via geographic footprints by deep learning. *arXiv preprint arXiv:1910.08734*, 2019.

[24] B. Zhu, W. Yang, H. Wang, and Y. Yuan. A hybrid deep learning model for consumer credit scoring. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 205–208, 2018.

[25] María Óskarsdóttir, Cristián Bravo, Carlos Sarraute, Jan Vanthienen, and Bart Baesens. 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, 2019.

[26] Cristián Bravo and María Óskarsdóttir. Evolution of credit risk using a personalized pagerank algorithm for multilayer networks. *arXiv preprint arXiv:2005.12418*, 2020.

[27] J Christopher Westland, Tuan Q Phan, and Tianhui Tan. Private information, credit risk and graph structure in p2p lending networks. *arXiv preprint arXiv:1802.10000*, 2018.

[28] Jianbin Lin, Zhiqiang Zhang, Jun Zhou, Xiaolong Li, Jingli Fang, Yanming Fang, Quan Yu, and Yuan Qi. Netdp: An industrial-scale distributed network representation framework for default prediction in ant credit pay. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1960–1965. IEEE, 2018.

[29] Ting Liang, Guanxiong Zeng, Qiwei Zhong, Jianfeng Chi, Jinghua Feng, Xiang Ao, and Jiayu Tang. Credit risk and limits forecasting in e-commerce consumer lending service via multi-view-aware mixture-of-experts nets. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 229–237, 2021.

[30] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[31] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[32] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[33] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[34] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[35] Jian Du, Shanghang Zhang, Guanhang Wu, José MF Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017.

[36] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[37] Alejandro Correa Bahnsen, Djamila Aouada, and Bjorn Ottersten. Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. In *2014 13th International Conference on Machine Learning and Applications*, pages 263–269, Detroit, USA, 2014. IEEE.

[38] Alejandro Correa Bahnsen, Djamila Aouada, and Bjorn Ottersten. Example-Dependent Cost-Sensitive Decision Trees. *Expert Systems with Applications*, 42(19):6609–6619, 2015.

[39] Miron B Kursa, Witold R Rudnicki, et al. Feature selection with the boruta package. *J Stat Softw*, 36(11):1–13, 2010.

[40] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[41] J Heggestuen. The mobile payments report: Forecasts, user trends, and the companies vying to dominate mobile payments. Accessed 26/03/21.

[42] Sezgi Guler-Fuechec. Alternative data in america: Policy and concerns. Accessed 18/05/21.