NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Application of Predictive Models for Bank Loans

Estimating a Reference Price to Improve the Loan's Negotiated Amount

Francisco João Mourato Calha

Internship Report

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Application of Predictive Models for Bank Loans

Estimating a Reference Price to Improve the Loan's Negotiated Amount

by

Francisco João Mourato Calha

Internship Report presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Prof. Fernando José Ferreira Lucas Bação, PhD, Universidade Nova de Lisboa

February, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisboa, 28 February 2024]

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to my team at Caixa Geral de Depósitos. Each member played a crucial role throughout my internship and development as a data scientist, always available to help me. A special acknowledgment goes to Luana and Teresa, who entirely supported and believed in me and were always there for encouragement.

Additionally, I would like to extend my thanks to Professor Fernando Bação for his guidance and patience during this period.

Lastly, I cannot forget to express my appreciation to my partner, Sara. Her unconditional support and patience were invaluable. I am also grateful to my family for always standing by me, especially my grandmother, who was always a source of motivation.

ABSTRACT

The banking sector has been significantly impacted by rapid technological advancements, prompting financial institutions to explore ways to adapt to these changes. This project resulted from an internship at Caixa Geral de Depósitos, where the primary objective was to assist the commercial chain in negotiating a specific type of loan with company clients. This was achieved by developing a predictive model that estimates a reference price for the loan's negotiated amount based on the historical operations of similar clients. The price must address the client's risk and differ based on the type of guarantee provided. After analyzing the data, it was found that two separate models would need to be developed due to the presence of a group of products with business limitations. The project explored various algorithms, including Extreme Gradient Boosting, Random Forest, Gradient Boosting, CatBoost, and Support Vector Regression. Extreme Gradient Boosting was chosen as the optimal algorithm for both product groups, achieving an R^2 of 0.844, an MAE of 18.26, and an RMSE of 33.78 in the normal credit product group, and an R^2 of 0.68, an MAE of 18.49, and an RMSE of 33.43 in the group of products with business restrictions.

KEYWORDS

Bank Loan; Client Risk; Estimating Reference Price; Extreme Gradient Boosting; Machine Learning

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

Bp	Basis Points
CGD	Caixa Geral de Depósitos
CI/CD	Continuous Integration and Continuous Deployment
CRISP-DM	Cross Industry Standard Process for Data Mining
DT	Decision Tree
ETL	Extract, Transform and Load
GB	Gradient Boosting
LGD	Loss Given Default
LinearSVR	Linear Support Vector Regression
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
PARE	Preço Ajustado ao Risco da Empresa
RF	Random Forest
RMSE	Root Mean Squared Error
SHAP	Shapley Additive Explanations
SVR	Support Vector Regressor
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

1.1. COMPANY OVERVIEW

Established over a century ago, Caixa Geral de Depósitos (CGD) is a Portuguese state-owned financial institution and one of the largest banks in Portugal. Founded in Lisbon in 1876, CGD provides access to a variety of high-quality financial products and services (Relatorio e Contas - CGD, 2022).

The CGD Group, apart from CGD, is also composed of various subsidiaries and affiliated companies, which contribute to the diversification of services provided and the brand's expansion at international levels. While CGD is the core bank that offers traditional commercial banking services, the subsidiaries and affiliates end up operating in various sectors of activity, such as the asset management ensured by Caixa Gestão de Ativos and CGD Pensões. Moreover, CGD Group's investment banking and venture capital platform contributes to support the economic growth, innovation, and entrepreneurship in Portugal and abroad. Although CGD mainly operates in Portugal, the group holds a presence across Europe and on other continents, including Africa, Asia, and America (*CGD Group Worldwide*, n.d.).

The year of 2022 proved to be a challenging year for the global economy. Still affected by the worldwide pandemic, the world economy also had to face the social and economic impacts originated from the war in Europe. These events contributed to an increase in volatility in the financial markets. Nevertheless, CGD has demonstrated the resilience needed to adapt to this sequence of exceptional events and actively aimed to pursue growth, innovation, and transformation to address future sustainability challenges. The institution's growth was notably recognized, and in the same year, several international studies highlighted its reputation in the eyes of consumers, even earning recognition from the UK Banker magazine, part of the Financial Times group, which ranked CGD as the top bank in Portugal in the TOP 1000 World Banks publication (Relatorio e Contas - CGD, 2022).

This internship was conducted within the Advanced Analytics division, which has the focus on obtaining machine learning models to support other departments of the institution in the decision-making process and promote technological progress and innovation across the different departments.

1.2. THE PROJECT

The ability to lend credit cautiously plays a fundamental role in the business activity and success of a bank or other financial institution, as well as in the overall economic growth. This way, companies can finance themselves with capital they may not have immediately available to support their business activities and investments. On the other hand, the bank is increasing its profitability through interest, fees, and commissions applied to the transaction. However, this will only be possible if the client can repay the borrowed amount plus these additional charges. Therefore, it is crucial for the bank to manage the credit risk associated with any loan to reduce potential financial losses and safeguard the bank's financial health.

Therefore, the client's profile becomes a valuable asset when providing credit. This profile is available for the manager to consult and includes the client's details for both identification and assessment. After assessing the client, and in case of approval, a negotiation phase takes place to discuss some additional charges and determine the spread rate to be applied. The all-in price is the resultant sum of these negotiable factors, and it falls to the manager to negotiate this price with the client in order to maximize the loan's profitability for the bank. To assist in these negotiations, the manager also has access to the Preço Ajustado ao Risco da Empresa (PARE), representing the minimum recommended price for the manager to propose. This price already addresses the risk of the company as well as other factors. This indicator is crucial in the negotiation phase but only sets a lower boundary for the operation price. So, when negotiating the all-in price, there may be cases where the client would be willing to pay a bit more for the operation. However, without having a higher reference value more adjusted to that client, there may be a lack of confidence in negotiating a higher price, driven by the concern of potentially losing the client.

By obtaining a reference value for the all-in price of a client based on the history of operations carried out by clients with similar characteristics, the manager has a new indicator that can be decisive during the negotiations. It is important to note that the proposed reference price will always be a recommendation based on similar clients without imposing an upper limit. In the end, the goal here will always be to assist the manager in decision-making, improving his confidence and giving him more control over the negotiations, allowing the bank to maximize the profitability of the loan.

Problem definition: **obtaining a reference value for the all-in price for each client based on the historical operations performed by similar customers.**

The report will address the following questions:

Can a Machine Learning model help improve the profitability of a loan?

Can the final results confirm the business requirements expected by the business partner?

2. LITERATURE REVIEW

2.1. FINANCING CONCEPT

Financing is crucial for both new and existing companies, supporting their business activities, investments, and expansion initiatives by providing necessary funds. According to Barbosa (2016), while some companies essentially finance themselves internally, utilizing funds generated from their own operations, profits, or retained earnings, others primarily rely on external funds obtained from sources outside the organization. The author even highlights in the study that companies often rely on bank loans as a primary source of external funding.

2.1.1. Bank Loans

In the banking industry, banks and other financial institutions have several products to offer to the customers, and one of their primary sources of revenue comes from loans (adapted from Majumdar et al., 2022). The term loan typically refers to providing credit by the bank to the customer, with the expectation of future repayment, including not only the borrowed amount but also additional interest or financial charges applied.

The interest rate corresponds to the sum of two factors: the base rate, also designated as the reference rate, and the spread.

$$Interest\ Rate = Base\ Rate + Spread$$

While the base rate corresponds to the interest rate at which central banks grant loans to commercial banks on international market, the spread represents the bank's profit margin on the interest rate for granting that loan (adapted from Kaczmarczyk, 2019). Regarding the charges applied, these can be fees or commissions. While some of them are mandatory, others are optional, and their value depends on the negotiation.

Additionally, these credit products address the customers' specific needs, and they can even be grouped into families based on common characteristics, such as the type of loan, the purpose of borrowing, repayment terms, or other factors.

Overall, the success or failure of the bank's financial performance is strongly related to the loan repayment at the agreed-upon term. When customers fulfill their responsibilities, the bank realizes a profit. However, if the customer defaults, the bank incurs a loss, which can severely impact its financial health.

2.1.2. Credit Risk Impact on Loans

Excessive unsuccessful loans can severely damage the bank's financial stability. They can even affect the stability of the entire financial system, as evidenced in past financial crises, such as the international crisis of 2008, where excessive exposure to high-risk credit loans contributed to the collapse of several financial institutions (adapted from Shi et al., 2022) .

According to Jafar Hamid & Ahmed (2016), credit risk can be defined as the probability of “the loan won’t be return back at time or at all”. Subsequently, the interest rate and the charges applied can reflect the risk that the creditor assumes by financing a company.

When it comes to assessing the default risk for companies, the balance sheet, also designated as a statement of financial position, is a valuable tool. This statement provides essential information about the company's financial health, liquidity, and solvency. By examining this information, creditors can determine whether a company can meet its obligations and evaluate its overall financial performance.

Additionally, by providing guarantees, clients can mitigate the risk associated with the operation, which can facilitate credit access and lead to lower interest rates. When a loan is secured by guarantee, not only does it reduce the lender's risk, as the bank can claim the guarantee to recover part or all of the loan amount in case of default, but it also provides an incentive for the client to repay the loan (adapted from Barro, 1976).

2.1.3. Credit Support Lines

In order to provide financial support to companies as a response to economic crises and emergency situations, or to stimulate specific activity sectors, governments, financial institutions, and other organizations develop programs and initiatives to promote credit lines with several benefits. For instance, in May 2023, CGD announced the creation of seven new credit lines for small and medium-sized enterprises to stimulate innovation and investment in specific fields, promoting sustainable transformation, and supporting the creation and competition of new businesses, among others (*Caixa-FEI Apoio Micro e PME*, n.d.).

Additionally, these support lines often include favorable credit conditions, such as lower interest rates, increased flexibility in loan repayment terms and an associated guarantee covering a portion of the financing to promote economic growth. Moreover, they can be limited on the financing amount and on the spread to be applied.

However, not all companies are commonly eligible, as each support line is intended for a specific event, with appropriate restrictions based on factors such as activity sector, business volume, and company size.

2.2. Related Work

The banking industry has not thoroughly explored the idea of obtaining a reference price for the negotiated amount of a loan. However, CGD has already investigated this subject in a similar approach that estimates the combined amount paid by company clients in the spread and additional charges for a different type of loan.

The project faced some challenges regarding the amount of available data to train a machine learning (ML) model. As this project was explored during the COVID-19 pandemic, which

affected the business economy, it only used data from the year in question and the previous year. Only client information was used to develop the model, and the all-in price of each client was obtained by aggregating the different all-in prices of the contracted operations by each one. Moreover, it was found necessary to obtain two distinct models. One to address the normal credit products associated with that type of loan and the other to deal with credit support lines that were trained separately. It was observed that the dataset of credit support lines had more records (distinct clients) than the dataset addressing the normal credit products.

Furthermore, in this project, the decision was to remove only extreme outliers from the target with no further removal techniques explored due to the lack of data. Regarding the process of handling categorical variables, the technique explored was Ordinal Encoding, as the categorical variables had a large number of unique values.

Some ML algorithms were tested, including tree-based models, such as Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and CatBoost, as well as other linear models. One linear model considered was Linear Support Vector Regression (LinearSVR), which is limited to a linear kernel. Regardless of the dataset, the ensemble algorithms outperformed the DT and linear models. For the dataset related to the normal credit products, Random Forest was the model that achieved the best performance with an MAE of 38.70, an RMSE of 56.87, and an R^2 of 0.149. For the dataset related to credit support lines, CatBoost was the selected model with an MAE of 21.12, an RMSE of 29.47, and an R^2 of 0.145. These values, as well as the ones presented later in the proposed report, were multiplied by the same factor to enable a better comparison across the projects. The lower errors obtained in the credit support lines dataset when compared to the other dataset can be justified by the fact that this dataset had more records available, and these lines of credit are typically more limited in terms of all-in prices. These models presented some limitations on handling higher values of all-in price, which typically correspond to clients with higher risk associated.

So, in a posterior development, the inclusion of the client's financial rating, which assesses the company's risk of default and incorporates various financial elements from the company's balance sheet, was even explored. Upon the introduction of this variable, the models recognized it as an important factor, helping them to generalize better. GB outperformed the other tested algorithms in both datasets, demonstrating an overall improvement in the evaluation metrics when compared to the early developments. For the dataset related to normal credit products, the achieved MAE value was 40.53, the RMSE value was 55.94, and the R^2 was 0.419. On the other hand, for credit support lines, the MAE was 20.21, the RMSE was 26.32, and the R^2 was 0.26. Additionally, this inclusion proved to be beneficial for the prediction of higher target values.

The positive results of the previous project were the motivation behind the initiation of this reported project. Additionally, it was noticed that Extreme Gradient Boosting (XGBoost) was not explored, and some studies now encourage the exploration of this algorithm in regression

tasks. One of them was a comparative study conducted by Grinsztajn et al. (2022), which explores the performance of several machine learning models, including XGBoost, RF, and GB, comparing their performance across a large number of datasets and different random search iterations. The author even states that the great performance of these algorithms in tabular data is attributed to the fact that they are ensembles, and the base weak learner is the DT. Furthermore, the study highlighted the overall excellent performance of XGBoost and its ability to outperform the other models explored in regression tasks. However, a separate study conducted by S et al. (2016) compared the performance of XGBoost and GB in different regression datasets and found evidence that XGBoost is not always the most accurate.

3. METHODOLOGY

The team's common methodology for approaching and completing this kind of ML project follows a hybrid approach explored by Baijens et al. (2020), where the Scrum methodology is integrated into the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology in the context of data science. The author states that "although CRISP-DM was intended to be an iterative model, evidence suggests that it has been used in a rather waterfall-like approach" and recommends incorporating "Agile practices alongside a waterfall approach during a data science project", emphasizing the compatibility and benefits gained by combining these methodologies. Indeed, in a business environment, it is quite common that insights gained during the process or changes in the requirements have an impact on the subsequent phases. So, this hybrid approach benefits essentially from the solid and well-defined structure of CRISP-DM for planning and executing the data mining process and, simultaneously, it leverages the adaptability and flexibility introduced by the Scrum methodology, allowing for changes in the project requirements and promoting a frequent communication with the stakeholders.

According to Shearer (2000), CRISP-DM divides the data mining project's lifecycle into six distinct phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, as shown in Figure 1.

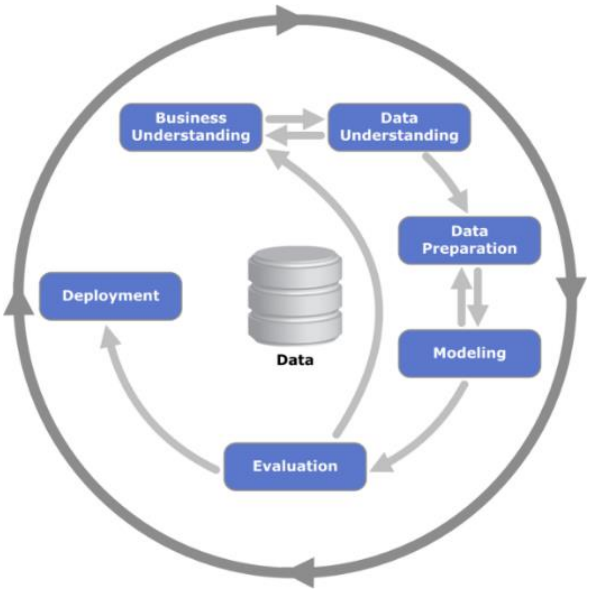

Figure 1 - Phases of CRISP-DM process

Source: (Swamynathan, 2019)

Moreover, it is important to understand that in a company like CGD, with an enormous database and an ongoing need to develop ML models to support various departments of the bank, the data understanding phase and data preparation phase can be challenging and very time-consuming. Consequently, the team has developed its own framework to optimize data processing and preparation tasks by constructing specialized tools to help in these processes. So, any new data source intended to be integrated into the team's information domains to support the machine learning models is processed by Shyness, a tool specifically developed by the team for data pre-processing tasks, ready to read the Extract, Transform and Load (ETL) rules, provided by the user, planned to be applied to that source. This process is illustrated in Figure 2.

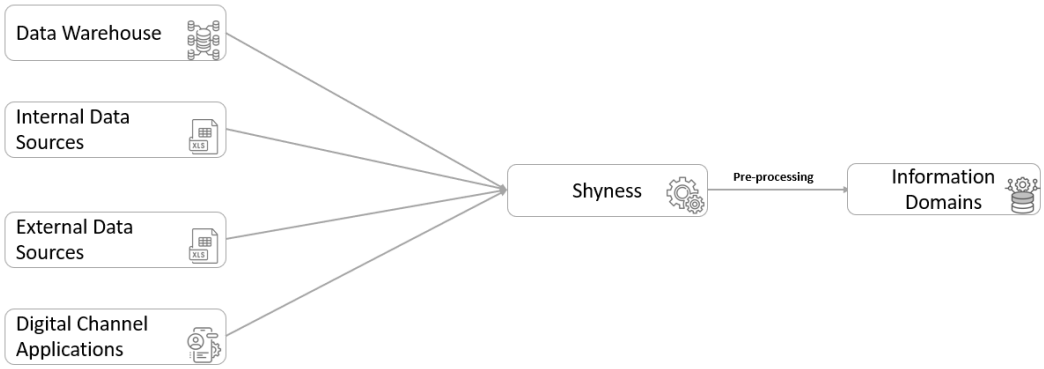

Figure 2 - Pre-processing framework using Shyness.

The data preparation is ensured by Shad, another tool developed by the team, which can read and execute the desired data preparation rules to be applied directly to the information domains or to other data preparations already stored by Shad. It is also capable of automatizing and storing the data preparations to be used in future developments of other ML projects.

Additionally, the modeling phase is also supported by another developed tool, designed by Grouper, where the user has access to simulate and set the space of parameters for different algorithms. For each model, a range of evaluation metrics are calculated, and together with a bunch of visualizations, make it easier to analyze the performance of each model and compare them.

In the following subchapters, a brief description of each phase of CRISP-DM will be presented, along with its implementation within the context of the project and utilizing the tools described above to achieve the proposed objective.

3.1. BUSINESS UNDERSTANDING

In this initial phase, the focus lies on understanding the project from a business perspective and identifying the stakeholders' needs, requirements, and expected outcomes. Moreover, it

is essentially to evaluate the availability of the resources to accomplish the project (adapted from Wirth & Hipp, 2020).

As mentioned earlier, when negotiating a loan for a credit product with a client, the manager can negotiate the spread rate and some additional charges applied to mitigate the risk associated with the operation even more and to retrieve better profitability from the loan. The all-in price is the resultant sum of both negotiable factors. During the negotiations, the manager already has access to the PARE price, which is the minimum recommended price to apply. However, relying solely on this price may be insufficient to grant confidence to the manager in proposing a higher all-in price with the fear of potentially losing the client.

The objective of this project is to obtain a reference all-in price for each client belonging to the *Empresas* and *Grandes Empresas* segments, which are the medium and large-sized company clients, to assist the manager during the negotiation of the negotiable factors to be applied for a specific credit product family. The reference price for each client will be based on operations performed by similar clients and must address specific business requirements. In other words, it is supposed to vary depending on the type of guarantee provided and the client's risk.

Furthermore, Figure 3 illustrates the initial approach designed for this project. However, its feasibility must be analyzed based on the available data.

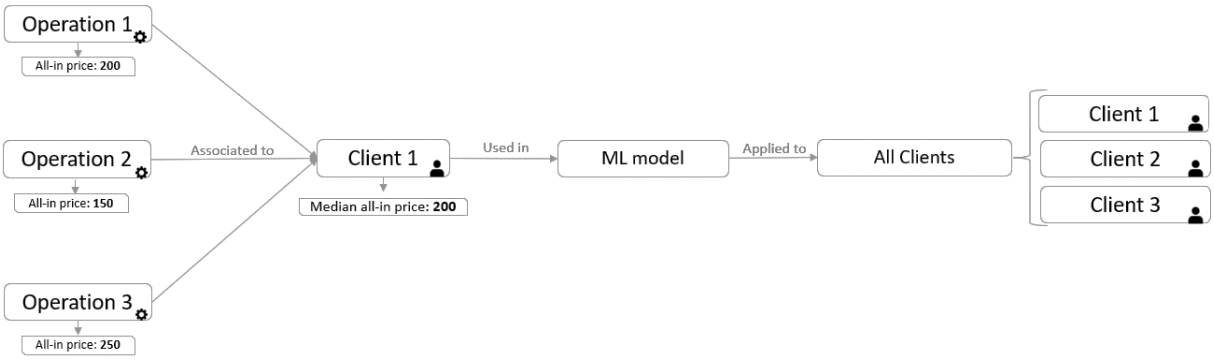

Figure 3 - ML model at client level framework

3.2. DATA UNDERSTANDING

This phase starts by collecting the necessary data. Then, the dataset is explored to become familiarized with it. It involves examining the data, identifying potential data quality issues, performing data visualizations to unveil potentially hidden patterns, and even analyzing pre-established hypotheses (adapted from Wirth & Hipp, 2020).

3.2.1. Data Collection

First, it will be necessary to obtain the target. So, the collected data can be classified into two categories: data required to address this task and data collected to train the ML model. Regarding the first point, all the information about the contracted operations must be collected. This information is already present in the team's information domains, and it contains all the information about operations that were already evaluated and approved by CGD, including information about other contracted products. The product family, the operation's term, or the fees and charges applied are examples of features present in this source.

Regarding the rest of the information needed to train the ML model, it is imperative to gather only the personal information strictly necessary about the client, not only financial but also demographic. For this purpose, there was also available a source of prepared information by the team not only with domestic information about the client collected by the company but also with public information coming from DUN & Bradstreet. This financial information encompasses the balance sheet items, from the ones more general such as total active or total passive, to items more specific as the net income of the period or the amount of intangible assets. Besides this, some evaluation business metrics and ratios were also available, computed by the team across previous projects, to appraise the company's activity. The company's age, the company's activity sector, the company's district, the total number of employees, the year of business start, and the client's segmentation are some examples of demographic features to consider.

Furthermore, the client's financial rating is stored in a different information domain, from which the rating ESG was also retrieved, which evaluates the company in terms of sustainability.

Finally, it is also necessary to gather information about each guarantee associated with the operation, retrieving its type, coverage percentage, and rank in case of default. Besides these, the Loss Given Default (LGD), which is a risk metric used to estimate the potential loss for the organization in case the borrower defaults, considering the given guarantee, was the last information needed to gather. However, this was the only one not present in the team's information domains and had to be processed with Shyness.

3.2.2. Acquisition of the Target

For the construction of the target, it will be necessary to have the information about the contracted operations already processed and stored in a team's information domain. All operations not related to the credit product family of interest or in the block stage have been filtered out.

For this credit product family, only a set of commissions must be included in the all-in price as additional charges to be considered. So, the calculus of all-in price can be given by the following formula:

$$all\text{-in } price = spread + \sum commissions$$

The following steps were to select only the commissions to be considered in this case, distinguish whether a commission is periodic or not, and standardize the claim's periodicity of the commission and the operation's term in years.

Regarding the commission value, the manager has to consult reference values whether it is a fixed amount or a fixed rate. Based on these values, negotiations determine the final commission value to be applied. This information also needs to be standardized in order to obtain the total amount paid in commissions. For this process, an internal reference file with specific formulas was used to address each case, considering the claim's periodicity, the operation's term, and the type of commission value to obtain the commission amount in basis points (bp). Subsequently, the total amount paid in commission for each operation was obtained by summing the individual amounts.

Finally, to calculate the all-in price in basis points, the spread rate column was multiplied by 100 to be also converted into basis points and then summed with the total charge amount.

3.2.3. Data Exploration

In this phase, it was necessary to initially analyze the customer and operation universes to gain insights into the preparations needed for the final dataset, as well as to assess the impact of certain variables on describing the target.

3.2.3.1. Universe Exploration

Exploring the universe in terms of both clients and operations is a crucial initial step in our data exploration process. Initially, the focus lay on examining the contracted operations in the current year and the previous year as well. A significant difference was identified between the total number of contracted operations and the total number of distinct clients. The latter was significantly lower, making it very challenging to explore the development of an ML model, as can be seen in Figure 4.

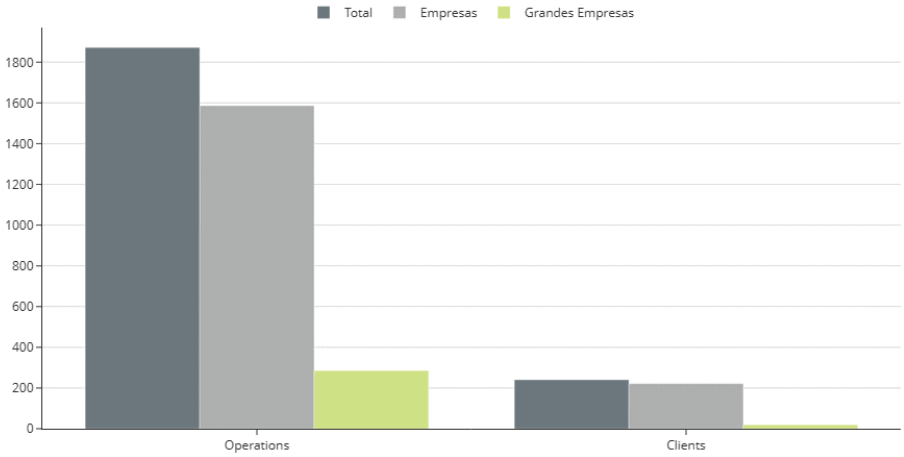

Figure 4 - Comparison of the number of operations and distinct clients

As an attempt to mitigate this restriction, a longer temporal scale was explored. However, this project was performed in a post-pandemic period, which may have had an impact on product demand and acquisition behavior over the previous years. So, it was decided to assess the feasibility of using data from the last three years in addition to the current year for model development. By comparing the target's distribution in each year, it was possible to observe similarities in the distributions, as illustrated in Figure 5, supporting the inclusion of data from 2020 and 2021. Consequently, it was chosen to retain these data modifications. It is important to note that, for reporting purposes, the target values were multiplied by a factor.

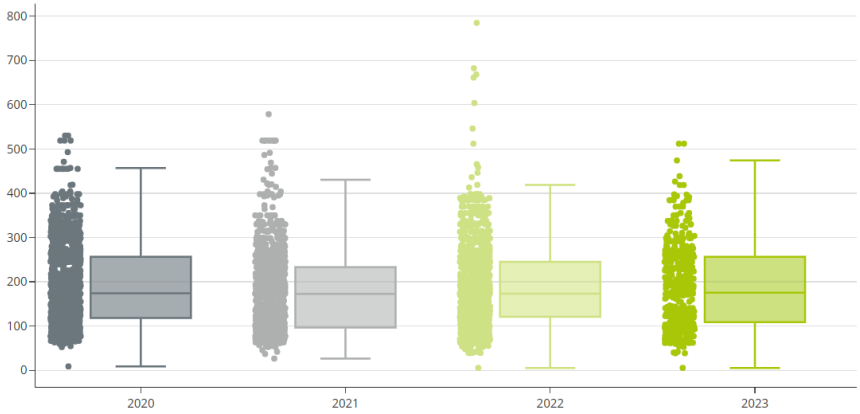

Figure 5 - Target distribution over the years

While the total number of distinct operations significantly increased, the total number of distinct clients did not show a corresponding improvement. To address this limitation, one last approach was explored to expand the pool of distinct clients available for model training. This involved incorporating another client segment, specifically *Negócios*, which includes small-sized enterprises. However, this approach was not favorably viewed, as it is generally expected that clients in this segment might be associated with higher all-in prices compared to those in

Grandes Empresas or even *Empresas* and their characteristics might significantly differ from larger companies.

Given the challenges in acquiring a substantial number of distinct customers for modeling, it was decided to explore the model's development at the operational level. This change will have an impact on the framework used to obtain the model. Instead of aggregating the all-in prices of all operations performed by a client and assigning the value to the client for training and evaluating the ML model, will be used the individual operations themselves, without combining them, for both training and evaluation, as illustrated in Figure 6.

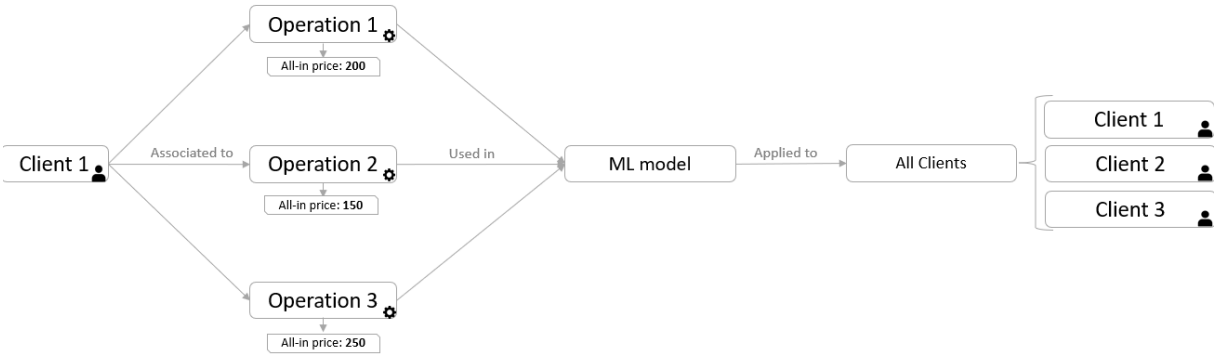

Figure 6 - ML model at operation level framework

As the objective remains the same, it will only be used to predict the target, mainly the client information variables associated with the operation. So, the final features of the model will always be related to the client, enabling its application to all clients at the end. This approach not only significantly increases the number of records available for modeling but also introduces variability in the target. Records with identical input features can be assigned to different all-in prices. An auxiliary example of the dataset proposed can be consulted in Appendix A. Moreover, this approach eliminates the apparent necessity of incorporating the *Negócios* segment. As a result, it was decided to stick with this approach and consider only the contracted operations associated with clients from the *Empresas* and *Grandes Empresas* segments.

Furthermore, with the business partner's domain knowledge of the products within this credit family, it was possible to identify a group of products related to credit support lines, and as already mentioned, they have various acquisition limitations, resulting in all-in prices typically lower than usual. Subsequently, the impact of each product group on the all-in price was examined. Consider this group composed of credit support lines as A-Restricted and the group with the rest of the products as A-Normal. Figure 7 highlights the differences in all-in prices between these two groups, making it necessary to address each group separately to obtain a better model's generalization for each one.

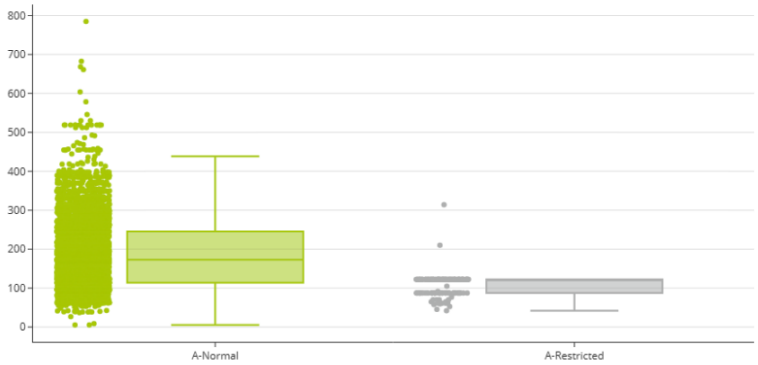

Figure 7 - Target distribution comparison for both groups

Furthermore, the total number of operations associated with each group was explored, and the distribution of the target variable across the different years for group A-Restricted was promptly analyzed. Regarding the all-in price distributions over the years, these remain identical. However, the results highlighted a clear limitation in producing a model solely for A-Restricted, as this group comprises only around two hundred operations.

To address this limitation on A-Restricted, the business partner proposed the inclusion of operations from a different financial product family. Both product families have similar characteristics, and the all-in price may be identical. Additionally, one advantage of this inclusion is that a group of products with identical limitations can also be identified, and designed by B-Restricted. Therefore, the all-in price distributions for both restricted groups were explored to better understand the impact of this approach. Figure 8 indicates that the two groups are not entirely similar, with the majority of B-Restricted operations falling within the 60 bp to 180 bp range for all-in prices. On the other hand, A-Restricted operations are more inclined to have all-in prices of either 90 bp or 125 bp, exhibiting less variability. So, the inclusion of B-Restrict can bring additional variability to the all-in prices of operations associated with A-Restricted, expanding the predictions for this group beyond the typical 90 bp or 125 bp values.

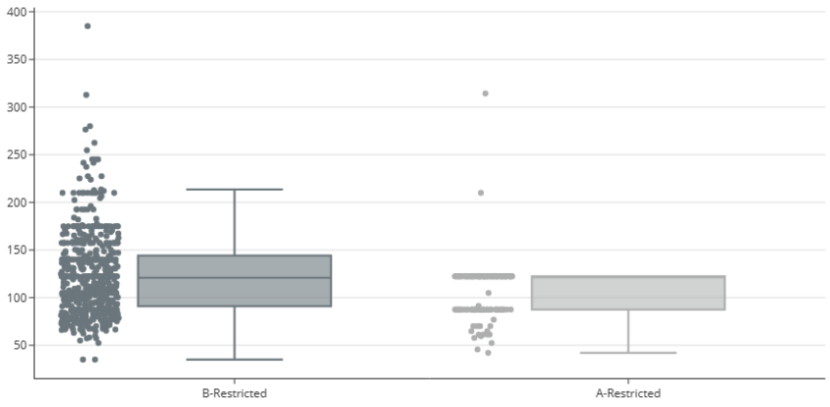

Figure 8 - Target distribution comparison between both restricted groups

It was observed that B-Restricted had nearly three times as many operations as A-Restricted, considering only *Empresas* and *Grandes Empresas*. This discrepancy might lead to a potential challenge in the model's generalization for A-Restrict associated with this type of loan. To address this imbalance, it was decided to incorporate operations from the *Negócios* segment. Moreover, credit support lines are typically launched to support mainly micro, small, or some medium-sized enterprises, and each credit support line is subject to specific constraints, which can complicate the process of generalizing an ML model. Including *Negócios* not only ensures a more balanced representation of operations from each product family but also introduces some variability into the data and prepares the model to handle different credit support lines. For instance, in Figure 9, it is possible to observe two records in *Empresas* with higher all-in values compared to the rest of the data could be considered outliers when analyzing solely this segment. Nevertheless, when including *Negócios*, there is a much greater representation of those values, and they might correspond to the same credit support line.

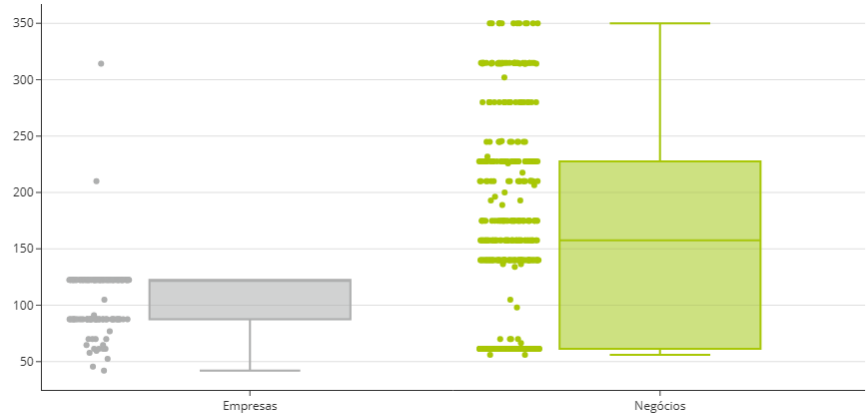

Figure 9 – Target distribution comparison for each client segment within A-Restrict

Overall, instead of producing a single model to address the main project objective, two ML models will be obtained separately. One is to address the products from A-Normal and will only use contracted operations from clients belonging to the *Empresas* and *Grandes Empresas* segments. The other will address the products from A-Restrict, which, due to the low number of operations associated, includes operations from a similar financial product family, B-Restrict. Moreover, for A-Restricted, operations from the three different client segments will be considered, while for B-Restricted, only operations associated with clients from *Empresas* and *Grandes Empresas* will be considered. In the future, consider the Normal group as the one related to A-Normal and the Restricted group as the group which combines A-Restricted and B-Restricted. In the end, the Normal group contained 4437 distinct operations, while the Restricted group contained 1225.

3.2.3.2. Main Findings

A deeper exploration of the data was then conducted, assessing the data quality associated with each group and performing some relevant analyses in the project's context.

Although most of the data comes from information domains and prepared data sources stored by the team, which already ensure a high level of data quality, inconsistencies, and missing values were spotted. Some guarantees associated with the operations had a cover percentage of zero, which is not possible. Regarding the missing information, some columns had missing values that must be filled in, while others, as they came from prepared data sources, were already filled in and must also be analyzed, as shown in Table 1.

Table 1 - Missing information

Variable	Variable's Description	Percentage of missing information (Normal dataset)	Percentage of missing information (Restricted dataset)	Missing information already filled	Filled value
c_notz_ratg	Financial Rating	3.25	6.37	No	-
c_ratg_esg_sust	Sustainable Rating	3.18	5.22	No	-
abt_characterization_districto_emp	District (internal source)	3.81	11.27	No	-
abt_characterization_company_age	Company age	0	3.76	Yes	0
dun_distrito	District (external source)	0.11	2.29	Yes	SEM INFORMACAO
cae_sector_section_description	Activity Sector	0.11	2.29	Yes	sem_sector
dun_ano_inicio_negocio	Started Business Year	0.11	2.29	Yes	9999
dun_ano_balanco	Financial Year	0.11	2.29	Yes	9999
dun_empregados	Number of employees	0.11	2.29	Yes	-1

Also, columns related to the client's balance sheet or other related financial information already had missing values filled with zero, and their total percentage of missing information was analyzed. For the Normal group, a total of 64 variables were found, with more than 50 percent of the information missing. On the other hand, for the Restricted group, 67 variables were identified as possessing more than 50 percent of missing information.

Moreover, both clients' ratings, ESG and financial, were represented as text variables when they should be numerical with the integer values ranging, for the client's financial rating between 1 and 15 and for the ESG rating between 1 and 3.

Furthermore, it was explored the total number of operations with at least one guarantee associated. For the Normal group, the analysis showed a predominance of operations with guarantees and, considering only these, on average, they were associated with approximately 2 different guarantees. On the other hand, for the Restricted group, was observed a more balanced dataset in terms of operations having or not any guarantee associated, and each operation with at least one guarantee has an average of 2.82 guarantees associated.

Afterward, it's difficult to perform a deeper analysis for each variable present on both datasets due to the presence of a large number of variables. So, in cooperation with the business partner was identified a set of possible meaningful variables that might help describe the target and were worth analyzing.

The client's financial rating was one of the variables indicated that might impact the all-in price. By analyzing the distribution of the all-in price for each rating, it could be concluded that, for the Normal group, in general, as the client's rating increases, the all-in price tends also to increase, as shown in Figure 10. However, this relation was not so evident within close rating levels, mainly within lower and even higher levels. Moreover, it was analyzed that operations without client rating, described by 'nan', had, commonly, lower all-in prices assigned.

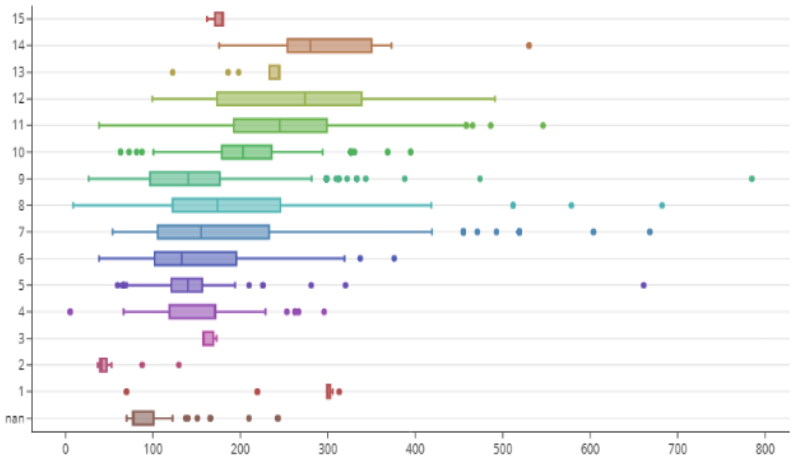

Figure 10 - All-in price distribution over the client's financial rating for the Normal group

Also, aggregating all the rating levels in three classes, where class [1-5] comprehends levels between 1 and 5, class [6-10] between 6 and 10, and class [11-15], it was observed that the

classes [1-5] and [11-15], which are the ones with the lowest and higher levels of ratings, respectively, were less represented in the dataset, as presented in Figure 11.

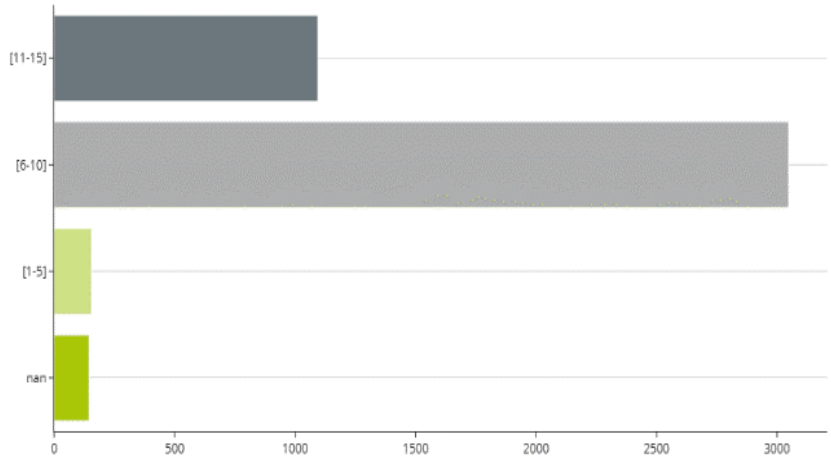

Figure 11 - Number of operations per rating class for Normal group

For the Restricted group, it was also concluded that there is a tendency to increase the all-in price as the client's rating increases and this tendency was also not so clear within lower and higher levels of rating, as shown in Figure 12.

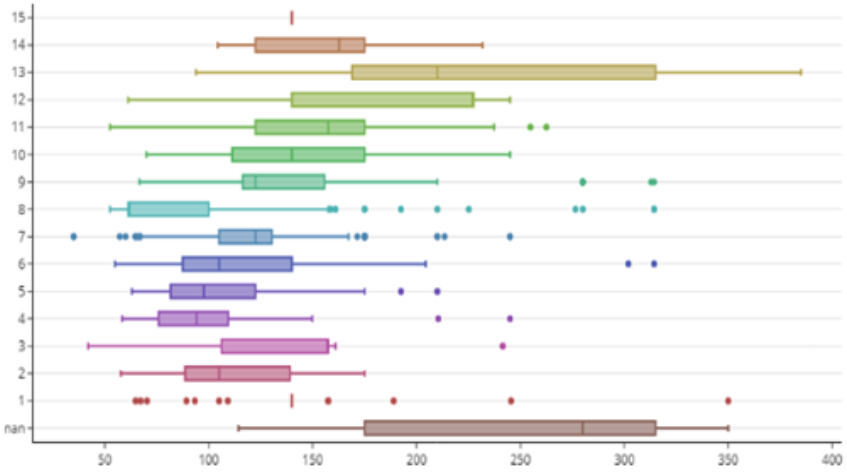

Figure 12 - All-in price distribution over the client's financial rating for Restricted group

Moreover, the classes [1-5] and [11-15] were more similar in terms of the total number of operations, as illustrated in Figure 13. Additionally, operations without a client's rating typically had higher all-in prices.

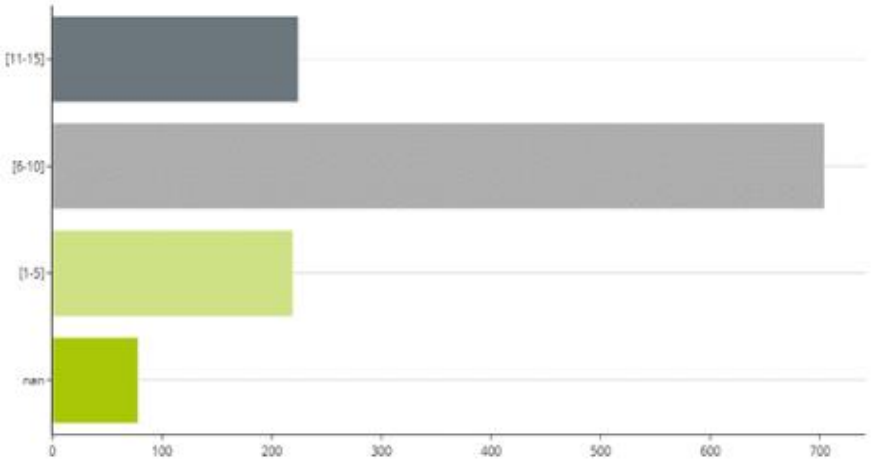

Figure 13 - Number of operations per rating class for Restricted group

The business activity sector of the client was also expected to impact the all-in price as some sectors might be associated with higher risk. A strong relationship between this variable and the target was not identified for the Normal group. The fact that this variable is especially dominated by two sectors, with similar distributions of all-in prices, makes it more challenging to evaluate this hypothesis. The sector actividades_de_informacao_e_de_comunicacao is usually associated with higher all-in prices. These results are presented in Figure 14 and Figure 15.

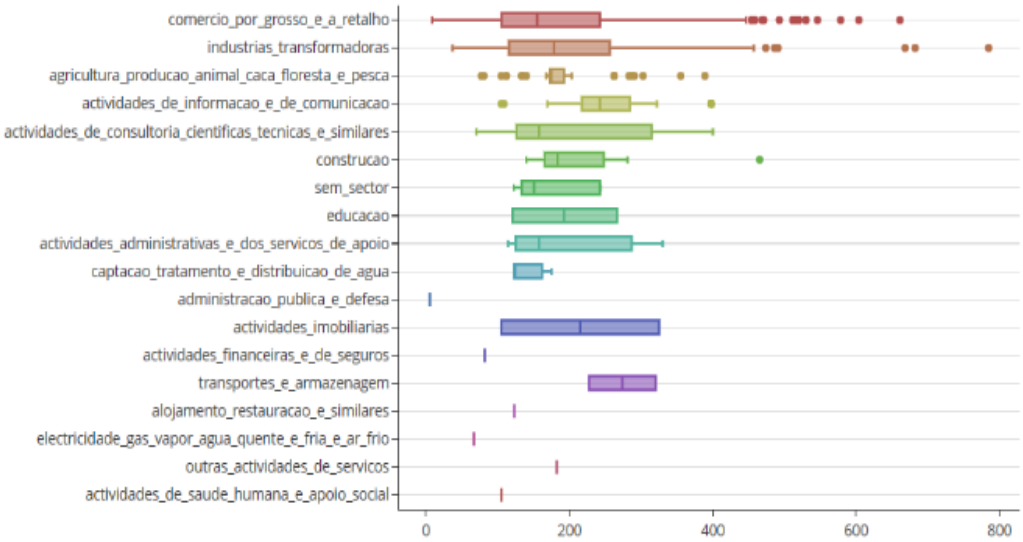

Figure 14 - All-in price distribution over the business activity sector for Normal group

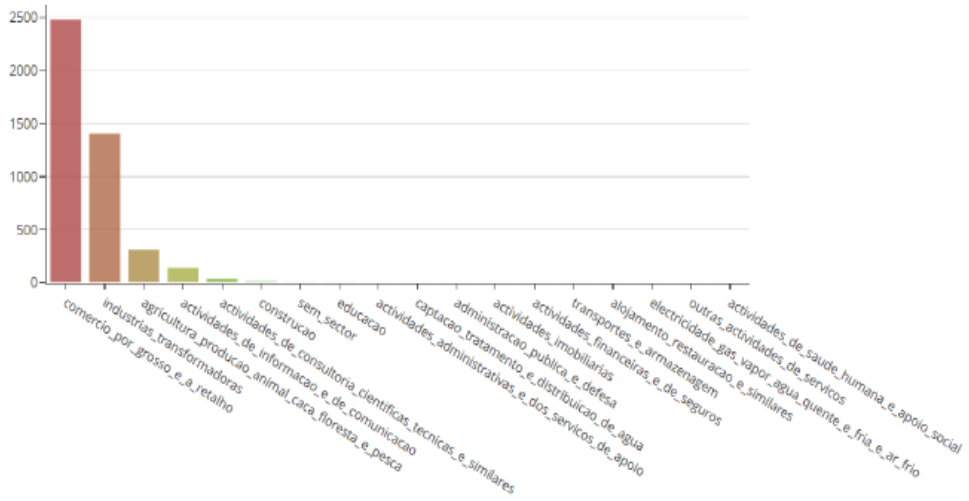

Figure 15 - Number of operations over the business activity sector for Normal group

For the Restricted group, the most predominant sector differs from the Normal group. Additionally, the three most represented sectors show similarities in terms of all-in prices, as shown in Figure 16, making it challenging to discern the impact of this variable on the target. Generally, the sectors agricultura_producao_animal_caca_floresta_e_pesca and administracao_publica_e_defesa are typically associated with higher all-in prices, as well as clients without sector (sem_sector). However, evaluating the impact of this variable on other sectors is difficult due to their lower representation, as presented in Figure 17.

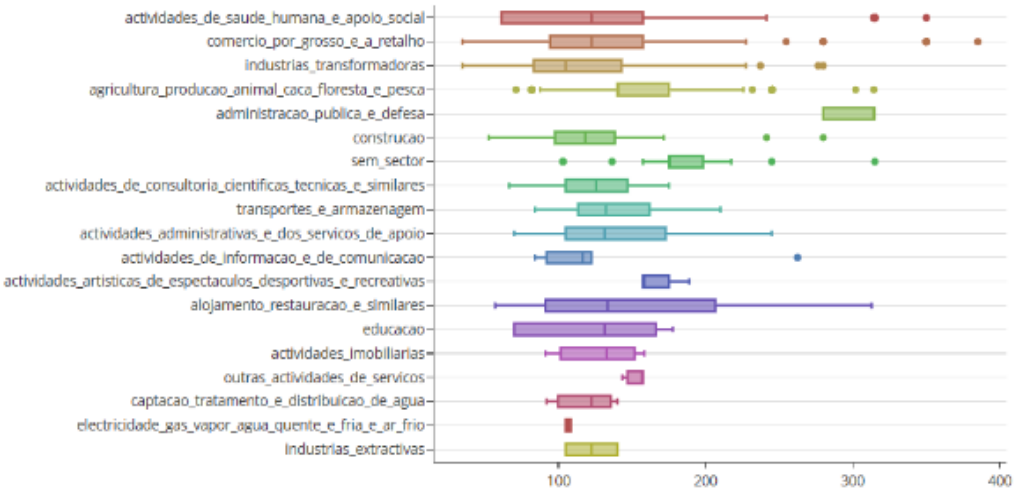

Figure 16 -All-in price distribution over the business activity sector for Restricted group

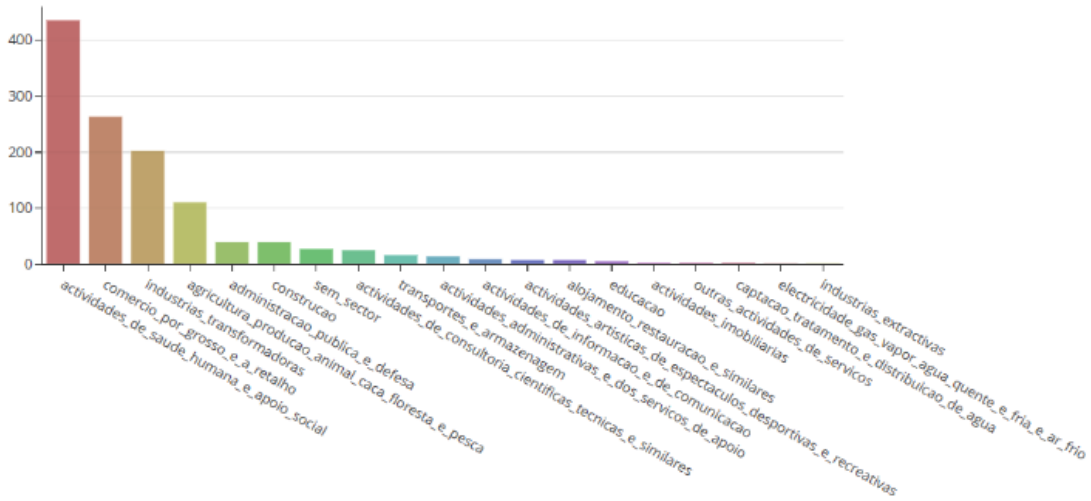

Figure 17 - Number of operations over the business activity sector for Restricted group

3.3. DATA PREPARATION

In this phase, the focus lies on integrating the data and performing the necessary transformations to ensure a dataset of high quality and relevance for subsequent phases.

As it was identified, there is a need to obtain two distinct models, each one addressing a different dataset. So, rerunning all the process whenever it is necessary to test the impact of certain preparations can be challenging. Thus, this process was split into two phases. In one initial phase, all the necessary and common preparations were applied for both datasets and, in principle, do not need to be rerun so often. In a second phase, specific preparations for each group were implemented, already addressing all the data transformations performed before. All this process was conducted using the Shad tool.

3.3.1. Common Preparations

As most of the data used comes from information domains recurrently needed for other projects, most of it already ensures good data quality. However, it is still necessary to perform some actions to further enhance it.

3.3.1.1. Removing Inconsistencies

Every guarantee with a cover percentage of 0 was identified and subsequently removed from the dataset.

3.3.1.2. Feature Engineering

By calculating the difference between the current year and the year when the business activity started (dun_ano_inicio_negocio), it was obtained the company's age (dun_company_age) in years, which is a more interpretable metric and will be useful for missing values imputation.

Then, the information about the client's district is present both in domestic information, which is collected by CGD (abt_characterization_districto_emp), and DUN & Bradstreet information (dun_distrito), which is external. So, it was decided to create a new variable designed by abt_district, which combines both columns by completing the missing information in abt_characterization_districto_emp with the corresponding information present in dun_distrito, which comes from DUN & Bradstreet.

The previous strategy was also employed for the company's age, preserving the domestic information from abt_characterization_company_age and completing the missing information with the respective values of dun_company_age. This variable was designed by company_age.

Also, an auxiliary indicator has been created to specify whether an operation is related to normal credit lines, credit support lines, or even the financial product family included in the Restricted group.

3.3.1.3. Handling Missing Data

At first, the missing values in the guarantee's type code and its description were set to '000' and 'Sem Garantia', respectively. In LGD and guarantee's rank, missing values were assigned to high values. LGD was set to its maximum value of 0.45, while the guarantee's rank was set to 999. Additionally, missing values in the cover percentage of the guarantee were set to 0. It is crucial to fill in these variables correctly, as they will be vital for aggregating the information later.

It is important to keep the missing values identified on the financial rating of the client (c_notz_ratg) as missing. One main reason is the fact that this rating has a valid date. So, this variable can include several ratings, ranging from lower to higher risk, that were not yet valid, and inputting them can lead to misinformation when training the model. Therefore, it was filled with -1, and the same approach was applied to the ESG rating (c_ratg_esg_sust).

Missing values on abt_district and company_age still needed to be filled in cases where the information was missing for both columns used to create these variables. The decision was to fill the missing values in abt_district with 'Sem Informacao' and company_age with 0.

3.3.1.4. Data Cleaning

Once each operation can be associated with multiple guarantees, there were duplicated operations, where the only differences were in the guarantee's information. So, in order to retrieve only one guarantee per operation, priority rules were defined with the help of the business partner and promptly applied. These rules involved the LGD of the guarantee, its rank, and its cover percentage on the operation. Upon this data preparation, it was necessary to take a step back and explore the data again to better understand the sampling of the different types of guarantees present. The results are presented in Appendix B.

Then, the data type of each rating, financial and ESG, was rectified and converted to integer instead of string and all labels from categorical variables suffered a sequence of text preparations to ensure normalization.

3.3.2. Modeling Preparations

The following step was to obtain two distinct datasets, one addressing operations from the Normal group, where all operations not associated with this group were filtered out, and the other dataset only composed of operations associated with the Restricted group. With the data properly separated among the different datasets, several specific data preparations that depend on each group were explored.

3.3.2.1. Handling Outliers

As observed before, during Data Understanding subchapter, there are, in fact, high values in the target, but none can be considered as a clear outlier. It can also be genuine data points carrying important information but with low representativity and, in this case, there are no extreme values. Moreover, keeping records with a high target's value can be beneficial since providing a prediction slightly higher than the actual value can help increase the negotiable margin. So, the decision was to not remove any record.

3.3.2.2. Categorical Features

For the process of converting categorical labels into numerical representations, the Ordinal Encoder was applied due to its simplicity and space efficiency. This encoding assigns a unique integer value to each label, which makes it more efficient than the One-Hot Encoder, which utilizes the common dummy variables technique to represent the categorical. However, the Ordinal Encoder "implies an order to the variable that may not actually exist" (Potdar et al., 2017). As some categorical variables present in both datasets have a higher number of labels, using the One-Hot Encoder will end up generating a very sparse dataset.

3.3.2.3. Feature Scaling

Although some ML algorithms are not sensitive to the different scales across the input features, such as tree-based algorithms, distance-based algorithms can be seriously impacted by it, for instance, SVR (adapted from Géron, 2019). Features with larger scales may dominate the distance calculations, and for these algorithms to interpret all the features in the same scale, feature scaling must be applied. In this case, for both datasets, the chosen algorithm was the MinMaxScaler to normalize each feature between zero and one.

3.3.2.4. Feature Selection

Reducing the dimensionality of the data can bring several benefits for the training phase because not only will it require less computation, but it will also reduce the time spent on this phase. Moreover, as the number of features increases, the sparsity of the data and the computation complexity also increase, which may lead to overfitting. This relation is known as the curse of dimensionality, and the solution lies in reducing the number of features (adapted from Guyon & Elisseeff, 2003).

Consequently, not only all the previously identified variables related to the client's balance sheet or other financial information, with more than 50 percent of the information missing, were removed, but also variables too specific about the client's financial information. For instance, were only retrieved the aggregated items in the information coming from the balance sheet. In addition, this approach can further facilitate the model's acceptance, as the final features selected by the model might be more intuitive.

To select the final features, it was explored the feature importance provided by some ML models as a feature selection technique. As both datasets are composed of numerical and categorical features, and ensembles usually provide better accuracy than weak learners, GB and RF were the explored algorithms to perform this task. Usually, the initial approach is to select either 10, 15, or even 20 features at the beginning, depending on the size of the training set. Then, the number of selected features can be altered to a more suitable value according to the cumulative importance of the features and the business knowledge.

Furthermore, it was decided to analyze the top fifteen most important futures considered by both algorithms for the Normal group and the top ten most important futures for the Restricted group.

For the Normal group, both algorithms identified the client's financial rating as one of the most crucial variables in describing the target, as illustrated in Figure 18 and Figure 19. However, neither GB nor RF identified the guarantee's type as one of the most important features.

Feature	Feature Importance
info_sir_pnc_tot	0.272356
c_notz_ratg	0.188607
metric_autonomia_financeira	0.133322
metric_ebitda_vs_volume_negocios	0.063383
metric_financiamentos_obtidos	0.041612
metric_financiamentos_obtidos_vs_ebitda	0.036737
company_age	0.036714
dun_empregados	0.033007
info_sir_pc_tot	0.031735
info_sir_ac_tot	0.022995
dun_financiamentos_obtidos_nao_corr	0.022589
info_sir_anc_tot	0.02045
metric_rentabilidade_liquida_vendas	0.016469
metric_rendibilidade_capitais_proprios	0.016137
dun_emprestimos_genericos_nao_corr	0.014166

Figure 18 - Top 15 Feature Importance GB

Feature	Feature Importance
c_notz_ratg	0.139309
info_sir_pnc_tot	0.094045
metric_autonomia_financeira	0.088159
dun_financiamentos_obtidos_nao_corr	0.084704
info_sir_tot_act	0.065491
info_sir_ebitda	0.058515
dun_emprestimos_genericos_nao_corr	0.049864
info_sir_cp_and_pass_tot	0.046481
metric_rentabilidade_liquida_vendas	0.040712
info_sir_ac_tot	0.035178
metric_financiamentos_obtidos	0.02906
dun_empregados	0.026257
metric_ebitda_vs_volume_negocios	0.025989
info_sir_pass_tot	0.025848
metric_financiamentos_obtidos_vs_activo	0.023078

Figure 19 - Top 15 Feature Importance RF

Furthermore, it was possible to conclude through Figure 20 and Figure 21, that not only did GB retrieve more cumulative importance with the 15 features compared to RF, but it was also possible to maintain a higher value of cumulative importance using this algorithm while reducing the number of features.

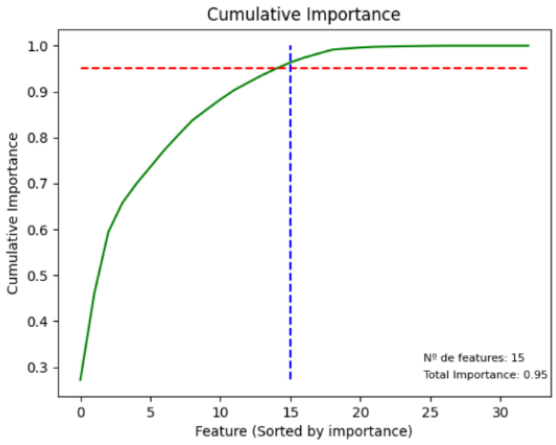

Figure 20 - GB cumulative feature importance

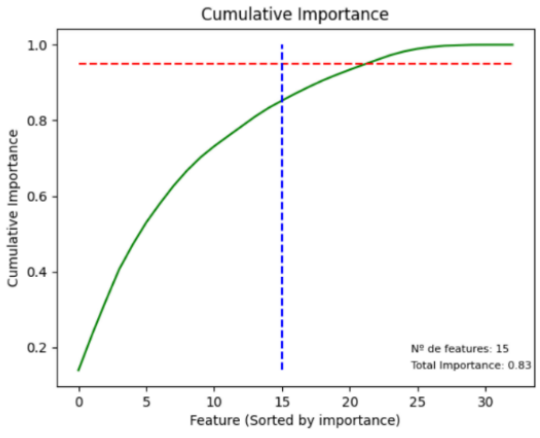

Figure 21 - RF cumulative feature importance

So, the decision was to proceed with GB and further explore the top 7 most important features, which still offer a favorable trade-off between the number of features and cumulative importance, resulting in a total importance of 0.77.

Moreover, it is crucial to analyze the presence of highly correlated features, as they may provide redundant information to the model. To evaluate the correlation within the set of seven features, the Pearson correlation coefficient was employed. This coefficient assesses the linear relationship between two variables, ranging from -1 to 1. Values closer to 1 or -1 signify a high linear correlation, while values closer to 0 indicate a low correlation (adapted from Swamynathan, 2019). Since only one pair exhibited a high correlation, as shown in Figure 22, the decision was to remove the feature metric_financiamentos_obtidos since info_sir_pnc_tot was considered the most important feature.

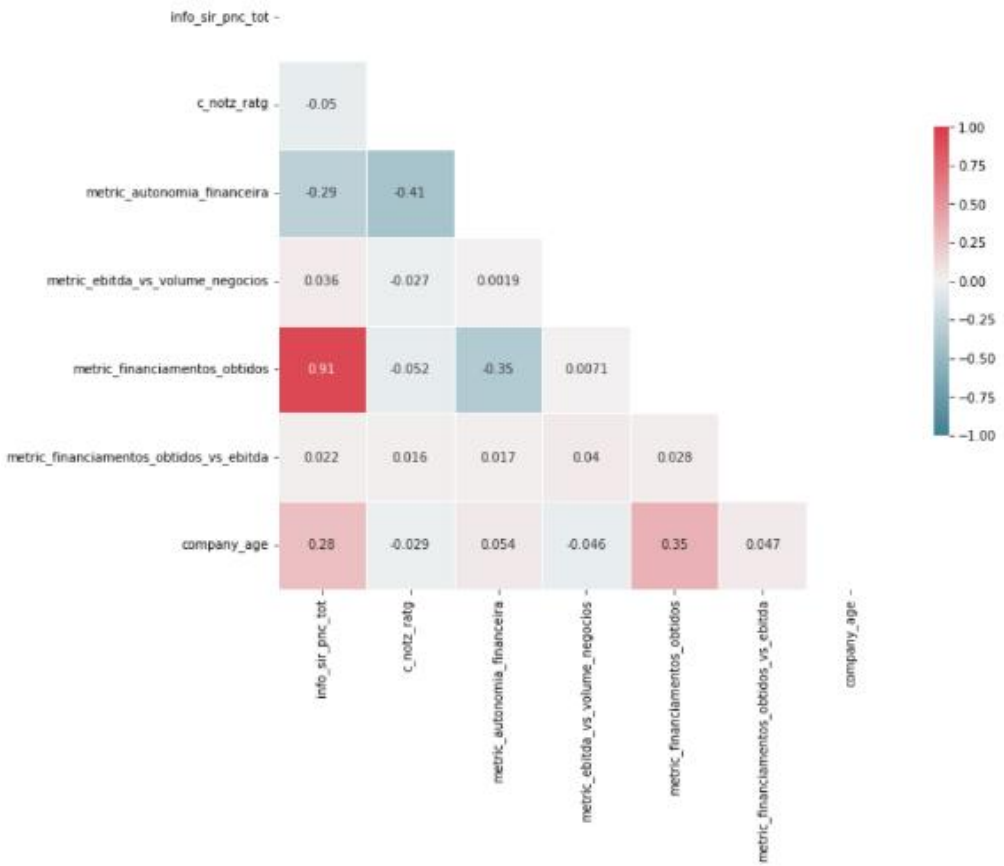

Figure 22 - Correlation matrix for Normal group

Furthermore, since the guarantee's type was not present in this set of features and its inclusion is crucial to ensure the model's applicability, it had to be incorporated.

A similar process was applied for the Restricted group, beginning with the analysis of the top 10 most important features identified by each algorithm.

Once again, and for the same reasons, with the help of Figure 23, Figure 24, Figure 25, and Figure 26, GB was selected as the method for feature selection. It was decided to analyze the top 7 features further, resulting in a total cumulative importance of 0.86. Furthermore, it was concluded that the feature info_sir_volume_negocios should be removed due to being highly correlated with info_sir_cp_tot, as seen in Figure 27. Subsequently, the guarantee's type was also included.

Feature	Feature Importance
info_sir_cp_tot	0.371484
info_sir_volume_negocios	0.149308
company_age	0.091108
c_notz_ratg	0.079976
metric_netdebt_vs_ebitda	0.074738
dun_empregados	0.048938
metric_autonomia_financeira	0.040379
dun_situacao_liquida	0.02429
cae_sector_section_description	0.019784
i_tipo_prod	0.018343

Figure 23 - Top 10 Features Importance GB

Feature	Feature Importance
info_sir_cp_tot	0.110119
dun_situacao_liquida	0.099542
dun_vendas	0.092789
info_sir_volume_negocios	0.074616
c_notz_ratg	0.059865
info_sir_cp_and_pass_tot	0.05825
info_sir_tot_act	0.056394
metric_autonomia_financeira	0.056382
info_sir_ebitda	0.044495
info_sir_pc_tot	0.037076

Figure 24 - Top 10 Features Importance RF

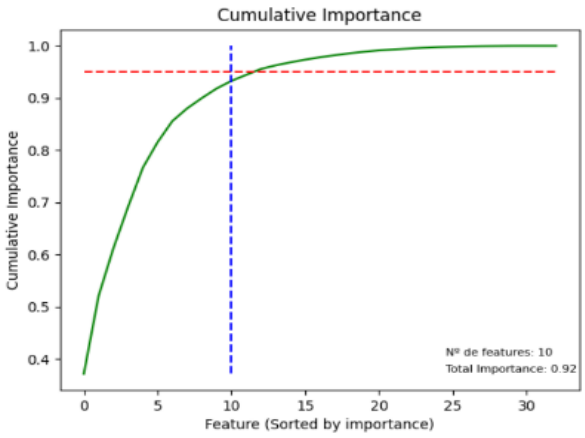

Figure 25 - GB cumulative feature importance (Restricted group)

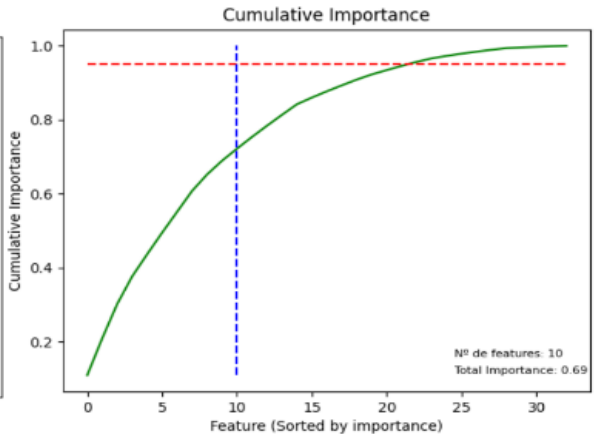

Figure 26 - RF cumulative feature importance (Restricted group)

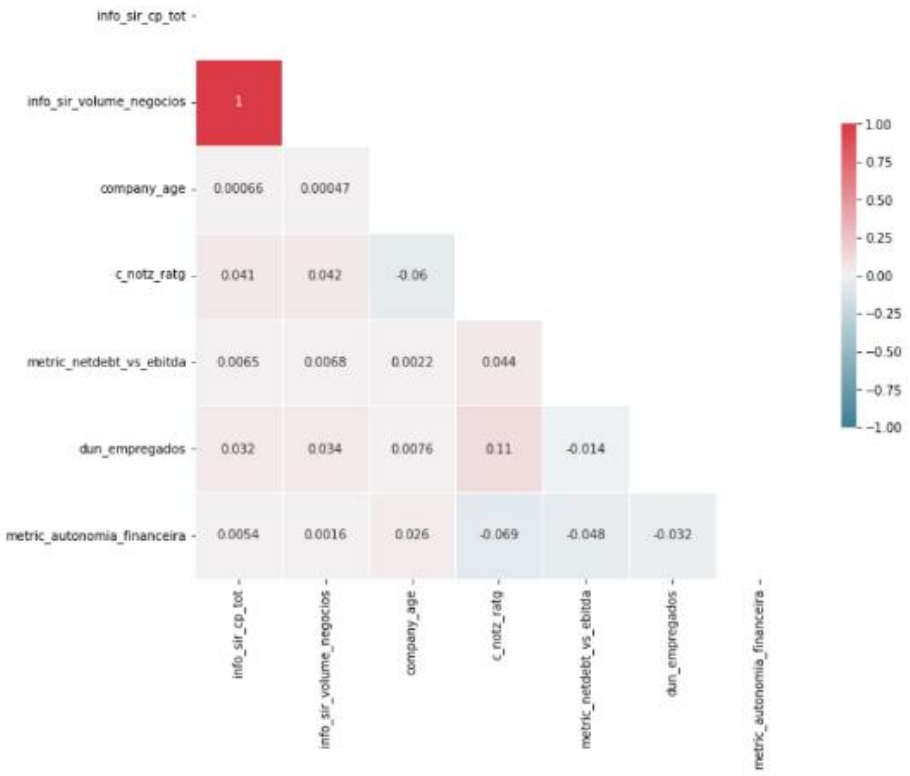

Figure 27 - Correlation matrix for Restricted group

3.4. MODELING

This phase is dedicated to testing different predictive models, defining the test design to be applied, as well as selecting the appropriate modeling techniques and algorithms intended to be used (adapted from Shearer, 2000).

3.4.1. Test Design

With the data already prepared for each required group, the following step consisted of dividing each dataset into training and testing sets. For the Normal group, 70% of the data was allocated for training and the remaining 30% for testing purposes. On the other hand, for the Restricted group, considering a lower amount of data, the decision was to split with a ratio of 80-20.

Afterwards, to prevent the risk of overfitting and to obtain a more robust evaluation of each model's performance, k-fold cross-validation technique was applied for both groups, with k equal to 5. It consists of dividing the training dataset into five subsets without replacement, using four of them for training and the remaining one for validation. This process is performed five times, with each subset serving as a validation set only once, where the results are

averaged at the end to obtain a model's performance evaluation less sensitive to the randomness in a particular split (adapted from Swamynathan, 2019).

3.4.2. Models

As stated before, this project aimed to obtain a reference value for the all-in price for each client based on the historical operations performed by similar customers. To address this regression problem, some ML models were employed, and is important to understand their concepts within the scope of this research.

In order to understand the algorithms chosen in this project, it is necessary to first have a better understanding of what a DT is.

This algorithm starts by establishing the splitting criterion, which determines the quality of the splits at each node by evaluating the impurity or error within. Next, the feature and the feature's value that minimize the splitting criterion are selected. This process is then repeated across the child nodes, which were created as a result of the split until a specific stopping criterion is met or when no more possible splits can be performed. Each terminating node, also designated as a leaf node, represents the final prediction, and each non-leaf node specifies a test to be conducted on a descriptive feature. Additionally, branches represent the different paths or decisions the model can take at each node (adapted from Kelleher et al., 2020).

Furthermore, this algorithm offers robustness against outliers as well as robustness against the addition of insignificant input variables for the model, also known as internal feature selection (adapted from Friedman, 2001).

However, DT can become too deep which makes them prone to overfitting, which is the phenomenon when "the model fits the data too well, capturing all the noises" (Swamynathan, 2019). So, techniques like pruning the tree, which involves selecting a stopping criterion to limit the tree's growth, or using ensemble methods, like RF, can be explored to mitigate this issue.

RF is a well-known example of bagging, where multiple weak learners, in this case, DT, are combined to build a stronger model called an ensemble. Additionally, each tree is trained independently and in parallel with a random subset of observations and features. In the end, the final prediction is often determined by averaging the predictions from the multiple individual decision trees (adapted from Kelleher et al., 2020; Swamynathan, 2019).

According to Ali et al. (2012), by combining multiple DTs, RF not only keeps the benefits of this algorithm but also obtains more robust results. Also, by introducing randomness in the selection of the training samples and the features to be considered, RF overcomes the overfitting problems with no need to prune the trees.

Furthermore, one major advantage of this algorithm is the fact that it can assign importance scores to each feature based on the mean reduction of impurity brought by the feature on every computed tree. The higher the score, the more influential the feature (adapted from RandomForestRegressor Documentation, n.d.).

Similar to RF, GB is also an ensemble that uses DT as base models, inheriting this algorithm's strengths. The main difference compared to RF is that GB builds trees by sequentially adding new models to the ensemble to make later ones "specialized in areas where earlier models struggled with" (Kelleher et al., 2020).

In general, this algorithm starts with a single weak learner predicting the training data. Then, its main objective is to reduce the loss function by iteratively adding a new base learner to the ensemble to fit the negative gradient of the loss function. As a result, subsequent models are trained to correct the errors made by earlier models rather than predicting the original target values as the initial model did (adapted from Swamynathan, 2019).

The success and effectiveness of the GB framework have inspired various implementations and extensions of the algorithm, with CatBoost being one of them. It was specially designed to handle categorical features efficiently during training. By computing statistics and even feature combinations, CatBoost can retrieve more information about the categorical features and, even, better capture the interaction between them without the need of preprocessing to be applied beforehand (adapted from Dorogush et al., 2018).

According to Dorogush et al. (2018), the main cause of overfitting on classical boosting implementation depends on the fact that "the gradients employed at each step are derived from the same set of data points on which the current model was constructed on". So, CatBoost generates different permutations to provide several training sets for individual trees.

As stated before in the Related Work, RF, GB, and CatBoost achieved good performance when predicting the negotiated amount of a loan, and as such, these models were chosen.

Additionally, XGBoost has become one of the most popular approaches in the ML community and for that reason, it was also selected (adapted from Chen & Guestrin, 2016).

XGBoost algorithm is a powerful algorithm described by Swamynathan (2019) as a more regularized version of GB. It emerged to address and optimize certain aspects of the traditional GB, offering efficiency, scalability, and performance improvements. Therefore, XGBoost was also selected.

Contrary to the traditional GB that stops when encountering a negative loss, XGBoost splits the tree up to a specified maximum depth regardless of the loss value in the split (adapted from Swamynathan, 2019).

It controls the model's complexity by relying on the regularization term, which, according to Chen & Guestrin (2016), helps avoid overfitting. Moreover, the usage of subsampling columns

is also described by authors as a feature capable of not only reducing the running time but also slightly increasing the model's performance and helping prevent overfitting.

Additionally, Chen & Guestrin (2016) also emphasize that the scalability offered by XGBoost is a key contributor to the effectiveness of the algorithm. The implementation of parallel processing and distributed computing allows the construction and evaluation of the trees to occur simultaneously and across multiple machines.

Finally, Support Vector Regressor (SVR) was chosen for its versatility as it allows the exploration of multiple kernels to handle also non-linear problems (adapted from Basak et al., 2007).

This algorithm identifies which are the most influential data points that help maximize the minimal margin between the regression curve and these points, also designed as support vectors (adapted from G. Li et al., 2019).

Essentially, SVR training is formulated as a convex optimization problem, which ensures that the global optimum solution exists and will be found, rather than a local optimum (adapted from Ghanbari & Goldani, 2021). However, it is sensitive to possible outliers and noisy data that can seriously impact the performance of this model (adapted from Sabzekar & Hasheminejad, 2021).

Before training the different models, it is important to understand that no rule or formula can a priori determine the best combination of hyperparameters for each model without testing them. Therefore, what is done is to define a set of hyperparameters that will be combined and tested to seek to optimize the model's performance. This process of seeking optimization for the hyperparameters is called hyperparameter tuning, and it might be a very time-consuming and resource-intensive process. However, applying this technique is expected to contribute not only to achieving a better performance in various models but also to avoiding underfitting or overfitting.

For this project, the hyperparameter tuning technique employed was Random Search with the incorporation of cross-validation to obtain a more robust estimation of the model's performance. It is well-known for its efficiency in testing a broad range of hyperparameters. This efficiency is largely attributed to its random selection of combinations to test, avoiding an exhaustive exploration of all possible combinations (adapted from Swamynathan, 2019).

Moreover, to help the process of understanding the selection of the final features, Shapley Additive Explanations (SHAP) values were analyzed. It has become a very popular tool in the ML field, as it provides a better explanation for complex models, allowing to better understand the contribution of the features and increasing trust over the final selected features (adapted from Saluja et al., 2021).

3.5. EVALUATION

During this phase, the selected models are evaluated based on success criteria and the business goals. Afterward, it is important to assess whether the model aligns with the business requirements and discuss possible reasons in case of failure (adapted from Shearer, 2000).

In order to assess the performance of a regression model, there are several metrics that are commonly used, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination.

Regarding MAE, as the name implies, it measures the average absolute difference between the predicted values and the actual observed values, quantifying the average error magnitude. The value of this metric ranges between 0 and positive infinite, with 0 meaning a perfect regression. It is computed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where n is the number of observations in the dataset, y_i corresponds to the actual value of the target variable for observation i and $\hat{y}_i$ to the correspondent predicted value.

In its turn, MSE computes the mean of the squared errors, which implies that its value is not on the same scale as the dependent variable but squared. As a result, MSE is more sensitive to outliers than MAE. So, it's suitable for cases when outliers are present and need to be detected once it penalizes more large errors. In order to bring the MSE to the scale of the dependent variable, the squared root is applied, and RMSE emerges (adapted from Chicco et al., 2021):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Similar to MAE and MSE, the RMSE value also ranges between 0 and positive infinitive, with 0 representing the perfect fit.

Willmott & Matsuura (2005) have suggested that RMSE is an ambiguous metric of average error, indicating that MAE should be used instead. Based on this study, years later, Chai & Draxler (2014) not only refute the main idea presented by Willmott & Matsuura (2005) of avoiding RMSE but also reinforces the need to evaluate a regression model by combining several metrics rather than focus solely on one. Each metric has its own limitations, and combining multiple metrics can provide a more comprehensive evaluation of the model's performance.

Regarding the coefficient of determination, also known as R^2, according to Chicco et al. (2021), it can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The author further highlighted the interpretation challenges of other commonly used metrics, such as MAE, MSE, and RMSE, when compared to R^2 for assessing the model's performance. As these other metrics evaluate the quality of fit in terms of the errors obtained, their values alone fail to communicate the quality of the regression performance.

Moreover, R^2 is superiorly limited by 1, indicating that the model is able to explain all of the variability in the dependent variable. Even not having a lower boundary, a value of 0 indicates that the model does not explain any of the variability in the dependent variable, and negative values of this metric indicate “a worse fit than the average line” (Chicco et al., 2021).

Therefore, there is a clear limitation when accessing the performance of a model with R^2, which is the fact that it does not directly provide any information about the magnitude of the errors. Li (2017) also recognizes the interpretability limitations of MSE, RMSE, and MAE on accessing the model’s performance but also demonstrates that R^2 can be a biased, misleading, and insufficient measure to predict the model’s accuracy.

Therefore, MAE, RMSE, and R^2 were used to evaluate each model when applied to the test set.

In order to evaluate the models' performance at the client level, two additional datasets were considered. One related only to contracted operations from normal credit lines of the credit family proposed, and the other to address the ones associated with credit support lines. Afterward, the operations were aggregated by client and guarantee’s type, using the corresponding median of the all-in prices. The preference for the median over the mean lies in the fact that it is a more robust measure when outliers are present.

Additionally, it was decided to extend the temporal scale to the latest possible date. This decision was made to include the maximum number of unique clients from the *Empresas* and *Grandes Empresas* segments with contracted operations. Considering testing only in new clients is very limited since the number of new clients in that period was not enough to draw conclusions.

Moreover, it is important to note that customer information may change over time, and it will be necessary to assess whether the model can effectively handle these changes. This involved analyzing not only how far the predictions are from the actual values but also how effectively they can differentiate the suggested reference price based on the customer’s financial rating and the type of guarantee provided.

Finally, each model must be applied to the entire universe of clients in the *Empresas* and *Grandes Empresas* segments, evaluating whether the models' results meet the business requirements.

3.6. DEPLOYMENT

In the final phase of the project, the results obtained are shared with the stakeholders, and the plan for maintaining and monitoring them is discussed, especially if they are intended to be integrated into the day-to-day business operations (adapted from Shearer, 2000).

After obtaining the final models and the results aligned with the business partners' vision, both models were moved to the production stage, and their results were ready to be integrated into the company's framework.

Each model maintenance is ensured by an individual Continuous Integration and Continuous Delivery (CI/CD) pipeline developed to replicate the process of achieving, testing, and deploying the model results periodically in an automated manner.

Changes in data pose a common challenge for models in production, and for that reason, monitoring these models is crucial. However, the ML community is still in the early stages of understanding essential metrics and developing trigger alerts for detecting data deviations from normal behavior (adapted from Paleyes et al., 2022). Consequently, currently, the monitoring of regression models lies in the feedback of everyone involved, and it is crucial to closely monitor the models' performance in the real-world scenario to retrain each model in case of need.

4. RESULTS AND DISCUSSION

The objective of this project is to obtain a reference value for the all-in price for each client based on the historical operations performed by similar customers. This was accomplished by following the workflow illustrated in Figure 28.

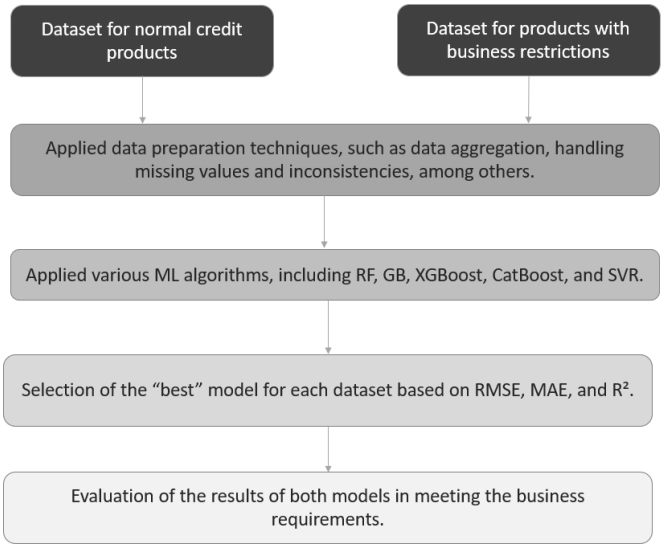

Figure 28 - Workflow diagram

In this chapter, it will be presented the results obtained from testing the different selected models on both datasets. Subsequently, the chosen final model for each group was evaluated at the client level, focusing solely on normal credit lines and credit support lines within the proposed credit family for the *Empresas* and *Grandes Empresas* segments. Finally, both models were applied to all CGD clients from the previous segments and the results were evaluated and discussed.

4.1. FIRST RESULTS EVALUATION

For the Normal group, the results achieved by the different selected models using the top 15 features considered by GB are present in Table 2, as well as the results obtained before and after including the guarantee's type to better understand the impact of this variable. The results of MAE and RMSE were multiplied by the same factor applied earlier during the Universe Exploration subchapter to enable a better comparison of these values in the context of the problem. Additionally, it can be observed that, in general, the inclusion of this variable did not significantly impact the performance of the models, except for SVR, which demonstrated a significant decline. As mentioned earlier, one advantage of DT is its robustness against the inclusion of non-important variables. The ensemble method used inherited this advantage from DT, which may help justify these results.

Table 2 - Models' results for Normal group

		MAE	RMSE	R^2
Top 15 Features considered by GB	RF	31.54	51.59	0.636
	GB	18.71	34.84	0.834
	CatBoost	64.10	80.43	0.113
	XGBoost	18.69	34.65	0.836
	SVR	66.35	87.99	0.264
6 Features	RF	33.52	53.68	0.605
	GB	18.80	34.52	0.837
	CatBoost	64.26	80.57	0.109
	XGBoost	18.18	34.13	0.841
	SVR	56.97	79.87	0.290
7 Features including guarantee's type	RF	33.43	53.44	0.609
	GB	18.72	34.52	0.837
	CatBoost	64.30	80.64	0.108
	XGBoost	18.26	33.78	0.844
	SVR	67.79	90.79	0.288

So, given that including the guarantee's type is crucial to ensure the model's applicability, XGBoost was chosen, as it obtained better results than the other models when considering this variable. Achieving an MAE of 18.26, which is relatively low considering the context of the problem, with the median of all-in prices around 175 bp. In addition, the RMSE value of 33.78, nearly twice the MAE value, suggests the presence of some records in the test set where the errors are significantly large. One main reason for this discrepancy could be the fact that the data contained several records with high target values, as observed in Figure 7 for A-Normal. Moreover, the R^2 value of 0.844 suggests that a high proportion of the variability in the data was explained using this set of features. From SHAP values visualization, presented in Figure 29, it was even possible to conclude that high values of the client's financial rating (c_notz_ratg) lead to higher values in the target. At the same time, the guarantee's type (d_fami_prd_cta_gar) cannot discriminate the target as effectively.

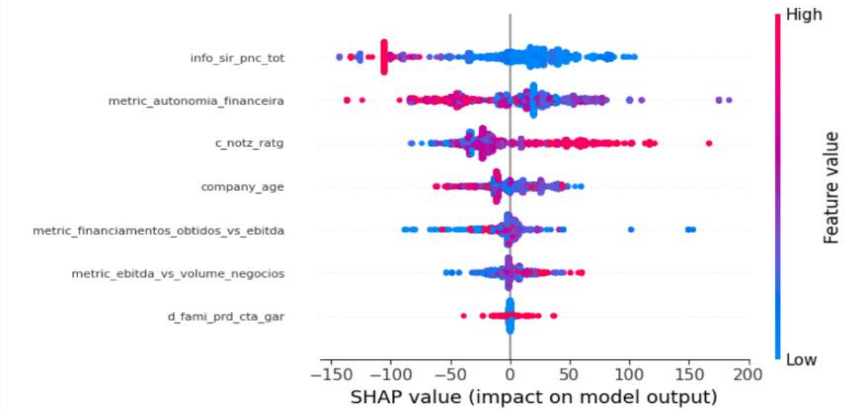

Figure 29 - SHAP values visualization for Normal group

For the Restricted group, the different models' results are presented in Table 3. It is possible to observe that the inclusion of the guarantee's type slightly improved the overall performance of the models, suggesting that this variable may provide new helpful information to describe the target that was not being considered before. RF and XGBoost achieved good results, and the overall performance did not vary by much, with RF achieving better MAE than XGBoost, but in the other metrics, XGBoost showed higher values.

Table 3 - Models' results for the Restricted group

		MAE	RMSE	R^2
Top 10 Features considered by GB	RF	18.14	33.42	0.680
	GB	18.19	33.46	0.683
	CatBoost	41.90	54.41	0.151
	XGBoost	18.52	33.26	0.679
	SVR	37.72	53.04	0.193
6 Features	RF	18.23	33.61	0.675
	GB	20.35	33.28	0.682
	CatBoost	42.15	54.58	0.146
	XGBoost	19.19	33.88	0.667
	SVR	37.63	49.56	0.287
7 Features including guarantee's type	RF	18.25	33.55	0.677
	GB	19.50	33.20	0.679
	CatBoost	42.01	54.43	0.151
	XGBoost	18.49	33.43	0.680
	SVR	36.60	49.40	0.291

When analyzing the SHAP values visualizations produced by both models, as presented in Figure 30 and Figure 31, it was possible to observe that XGBoost offers a better understanding of the contribution of the client's rating, indicating that higher values of this feature lead to higher values of all-in price. This is a crucial pattern to capture, and RF shows more difficulty handling it.

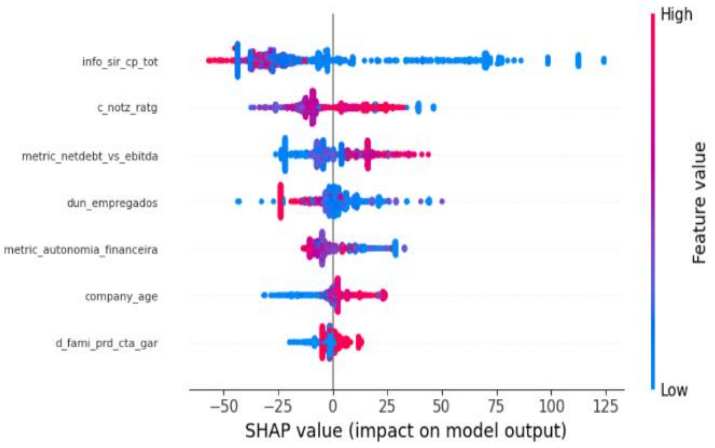

Figure 30 - SHAP values visualization of XGBoost

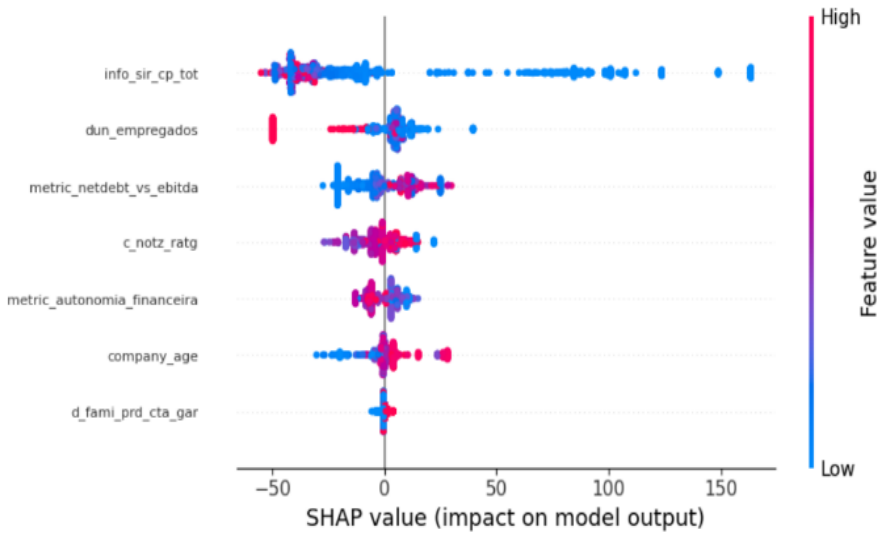

Figure 31 - SHAP values visualization of RF

So, the decision was to proceed with XGBoost, which achieved MAE, RMSE, and R^2 values of 18.49, 33.43, and 0.68, respectively. While the MAE value is similar to that achieved in the Normal group, its impact on the Restricted group is different because the all-in values in this group are typically concentrated at lower values. Figure 8 and Figure 9 had already highlighted the discrepancy among the different subgroups within the Restricted group, which helps justify why the performance metrics are not better.

In the following subchapter, the chosen models will be evaluated by applying them to the dataset created at the client level. The aim is to gain a better understanding of how well these models perform in meeting the business requirements.

4.2. Model Evaluation at Client Level

Firstly, the results for normal credit lines were analyzed by applying the model developed for the Normal group. The analysis started with a comparison between the distributions of the client's all-in prices and the predicted ones across the proposed client segments, as presented in Figure 32 and Figure 33.

Similar to the project described in Related Work, this model also tends to slightly increase the all-in prices, in general. Regarding the predictions obtained for clients in the *Grandes Empresas* segment, they are significantly influenced by clients in the *Empresas* segment due to the substantial difference in the number of operations associated with each segment during the model training.

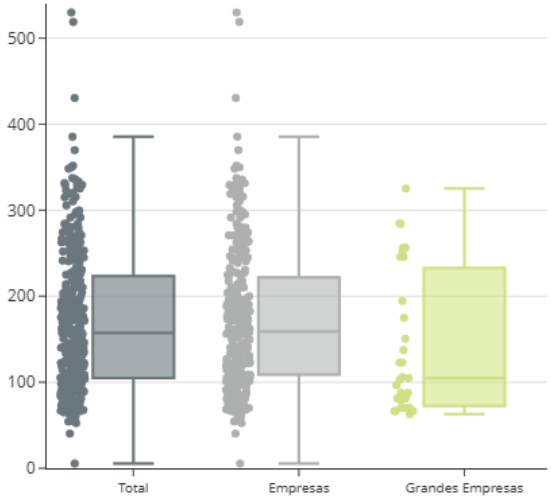

Figure 32 - Actual values of all-in price over the client segment for normal credit lines

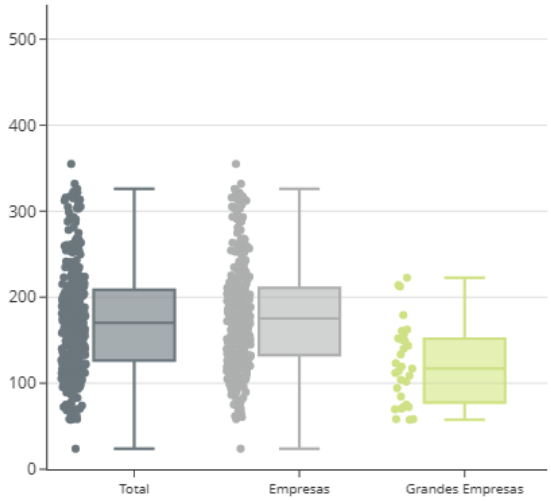

Figure 33 - Predicted values of all-in price over the client segment for normal credit lines

Subsequently, the predictions were analyzed across client ratings and compared with the actual values, as illustrated in Figure 34 and Figure 35. It was observed that the model effectively captures the expected pattern in the rating by increasing the all-in price as the client's rating increases.

Rating 1 stands out with a deviation between the predictions and actual values that should be further analyzed since these clients are typically reliable customers, and the model increased the recommended price. However, it indicates that most of the operations used for training associated with this rating had considerably high values of the all-in price. A deeper analysis concluded that these operations belonged to clients whose rating was subsequently changed to a higher value. Despite this, the predictions are not significantly far from the actual values.

Moreover, clients associated with higher ratings typically have a higher probability of default, and there is a lack of data for clients in these ratings as they might become unprofitable for the bank. In general, for rating 15, the model even tends to increase the prices, but for rating 14 and 13, the model tends to suggest a decrease in the all-in prices. Despite the lower variability in intermediary rating levels, the model is able to better discriminate the suggested all-in price for these ratings.

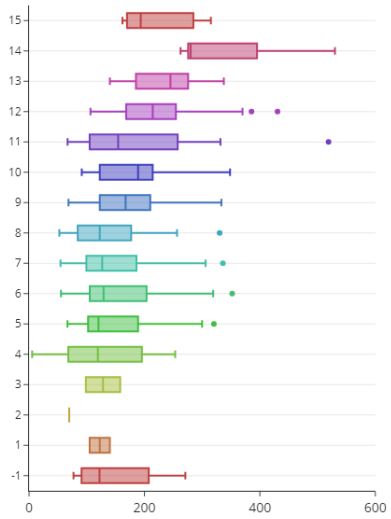

Figure 34 - Actual values of all-in price over the rating for normal credit lines

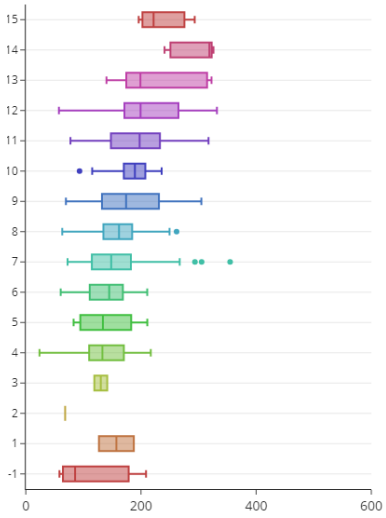

Figure 35 - Predicted values of all-in price over the rating for normal credit lines

Overall, the model is able to suggest higher all-in prices compared to the actual values, indicating clients who might be receptive to paying more for the loan. On the other hand, it also suggests lower prices and several factors can explain that. One reason may be the lack of distinct clients used to train the model, while another factor could be the manager's role in negotiating prices. In some cases, the manager could negotiate a higher all-in price with a client despite a historical tendency towards lower prices for similar ones.

When conducting a similar analysis for the distinct types of guarantees, interpreting the results becomes more challenging, primarily because this variable is predominantly dominated by one label, as observed in Figure 50, and the variations observed in the predictions across different types might be attributed to the influence of other variables in the model. The analysis is shown in Figure 36 and Figure 37.

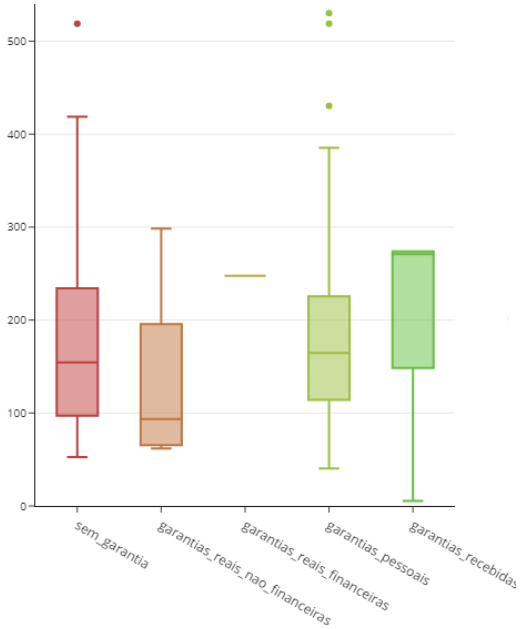

Figure 36 - Actual values of all-in price over the type of guarantee for normal credit lines

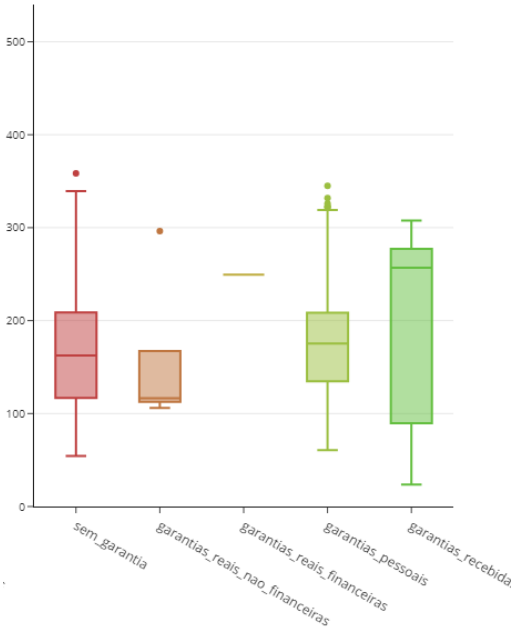

Figure 37 - Predicted values of all-in price over the type of guarantee for normal credit lines

For the credit support lines, a similar sequence of analyses was conducted. A clear tendency to increase the all-in prices can be observed in Figure 38 and Figure 39, which might be explained by the discrepancy in the all-in prices among the various subgroups within the Restricted group.

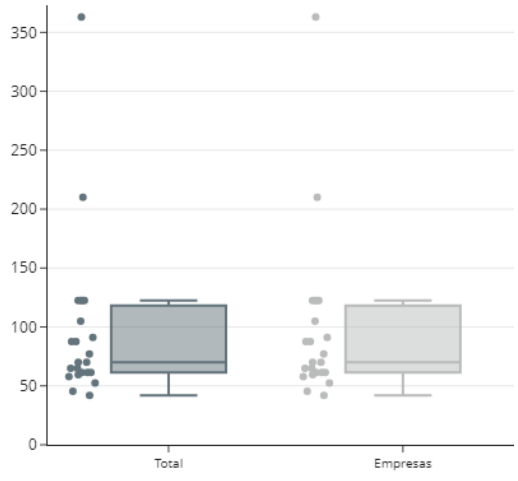

Figure 38 - Actual values of all-in price over the client segment for credit support lines

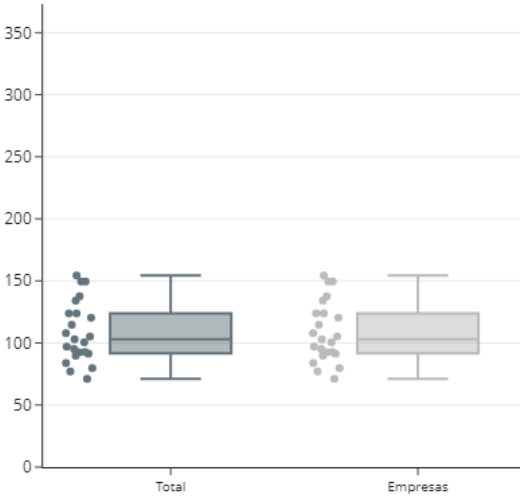

Figure 39 - Predicted values of all-in price over the client segment for credit support lines

Regarding the client's rating, although a clear relationship between this variable and the all-in price is not evident in the actual values, the predicted values are more aligned with the client's rating, suggesting an increase in the all-in price as the rating increases.

This analysis is shown in Figure 40 and Figure 41. The main reason behind this result might arise from the decisions made throughout the process of obtaining the dataset to develop the model. Figure 12 had already highlighted this relationship between the client's rating and all-in price, and even though the operations from credit support lines of *Empresas* and *Grandes Empresas* alone cannot fully describe this relation, the other subgroups within the Restricted group helped the model understand this pattern.

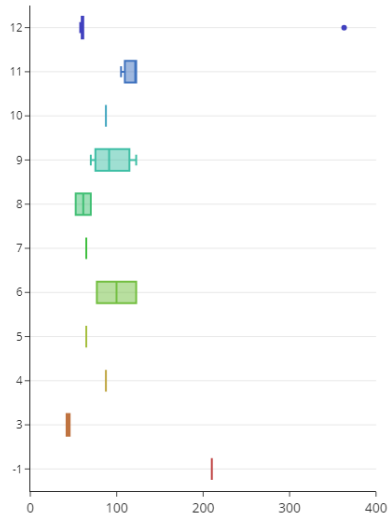

Figure 40 - Actual values of all-in price over the rating for credit support lines

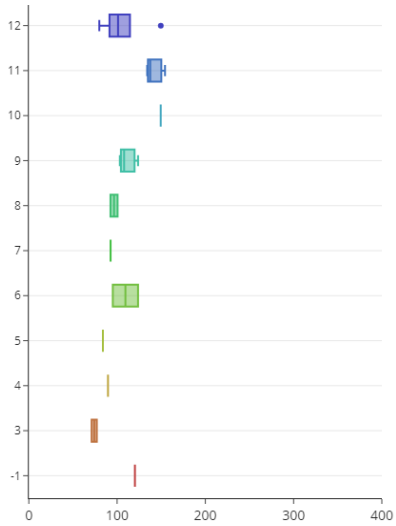

Figure 41 - Predicted values of all-in price over the rating for credit support lines

In general, the model recognizes that clients with a higher risk should be associated with higher prices, except for rating 12, where the predictions are higher but not higher than rating 11. This discrepancy is acceptable, as credit support lines are more limited in the negotiated all-in price. Suggesting values for this rating higher than the prices of rating 11 might lead to predictions far away from the actual prices possible to negotiate with the client.

Concerning the impact of the type of guarantees for these credit support lines, it is not as crucial to analyze since they are supposed to be typically associated with a particular type (garantias_pessoais), and operations without guarantee associated (sem_garantia) or from a different type might be rare or inconsistencies. Nevertheless, the results are presented in Figure 42 and Figure 43.

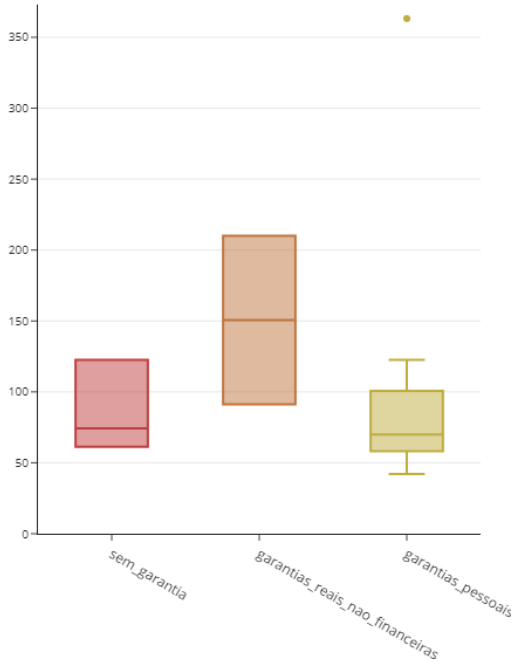

Figure 42 - Actual values of all-in price over the type of guarantee

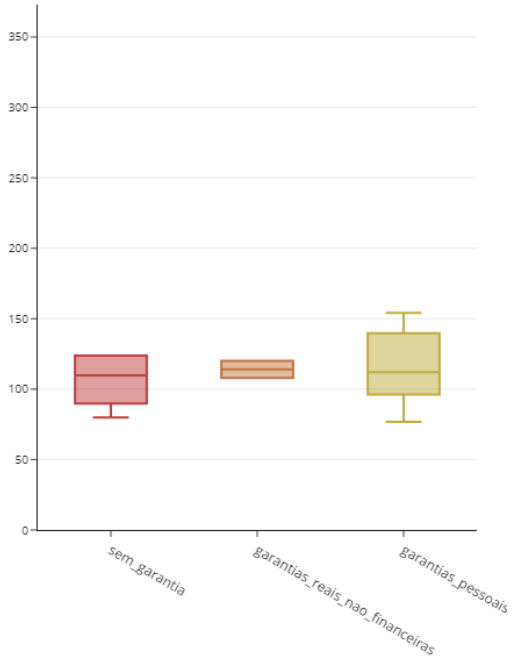

Figure 43 - Predicted values of all-in price over the type of guarantee

4.3. Obtaining the Final Results

After analyzing the model results, each model was applied to all CGD clients in the *Empresas* and *Grandes Empresas* segments, considering the possibility of the client providing each type of guarantee. This universe of clients is significantly larger than the previous, and it is crucial to analyze them as they represent the real-world outcomes.

For clients from *Empresas*, the predictions of credit support lines are generally slightly lower than the predictions for normal credit lines, as illustrated in Figure 44 and Figure 45. Moreover, for normal credit lines, the model suggests prices typically lower for *Grandes Empresas*, possibly because these companies are usually more financially stable. However, for credit support lines, the same does not happen, and the most probable reason for that is the fact that larger companies usually are not eligible for these lines of credit.

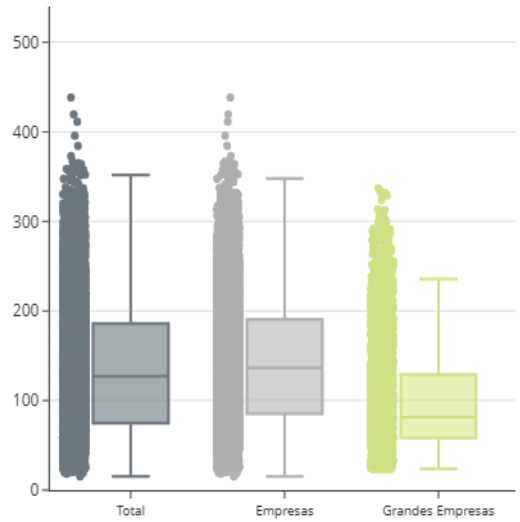

Figure 44 - Final predictions of all-in price over client segment for normal credit lines

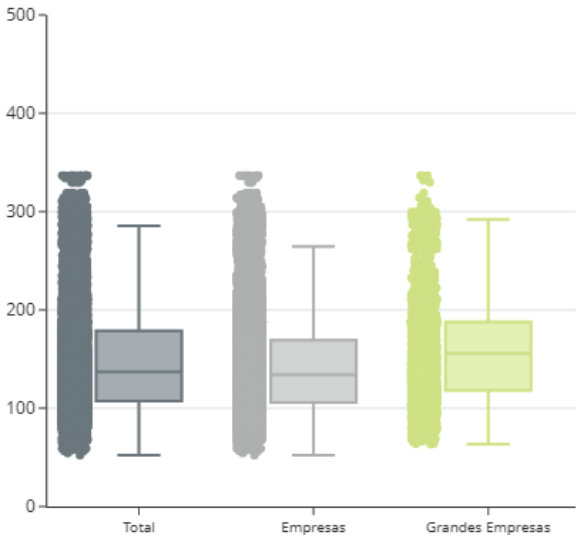

Figure 45 - Final predictions of all-in price over client segment for credit support lines

In terms of client ratings, the results presented in Figure 46 and Figure 47 still demonstrate an increase in the all-in price as the rating ascends. While normal credit lines show a higher variation in the all-in price based on the rating, credit support lines exhibit a more gradual increase. Moreover, in the collected data, some levels of rating were missing for credit support lines associated with clients from *Empresas* and *Grandes Empresas*, and the presence of the other subgroups within the Restricted group might have facilitated a better prediction of these ratings.

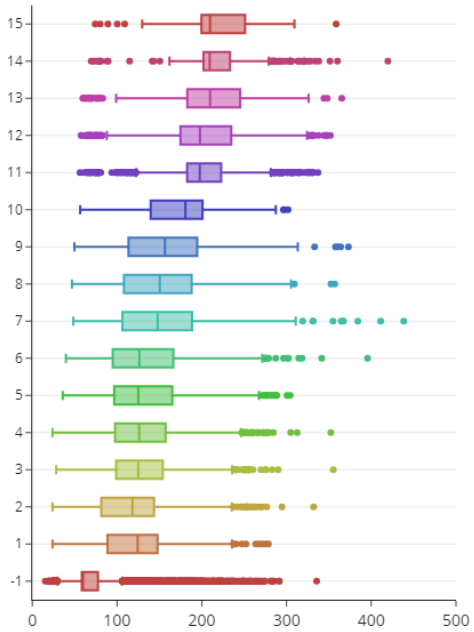

Figure 46 - Final predictions of all-in price over the rating for normal credit lines

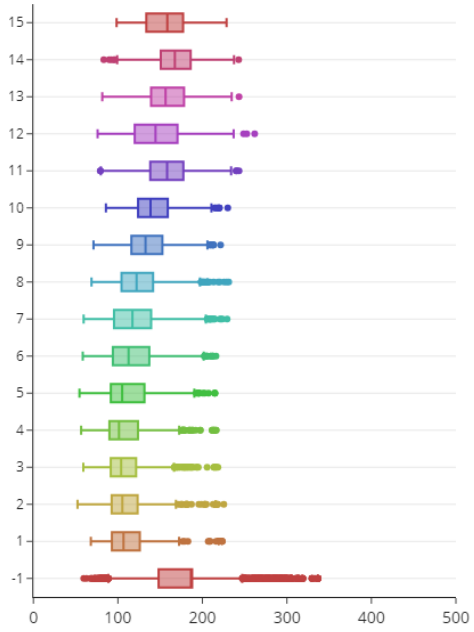

Figure 47 - Final predictions of all-in price over the rating for credit support lines

Regarding the guarantee's type, as already mentioned, its impact is not so significant on credit support lines since they are usually associated with one specific guarantee's type. Even though the model is able to differentiate the price slightly. However, for normal credit lines, the predictions are quite similar among themselves, indicating that the model is not able to better discriminate the all-in price, whether the client provides a guarantee or not, nor among the different types. These results are presented in Figure 48 and Figure 49.

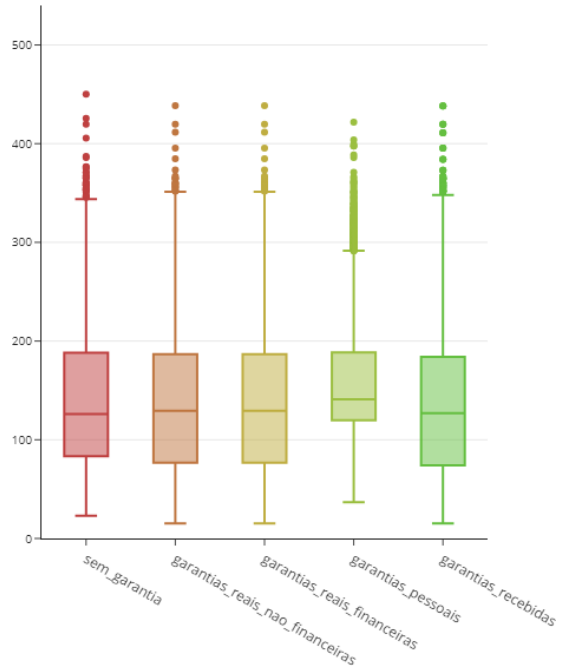

Figure 48 - Final predictions of all-in price over the type of guarantee for normal credit lines

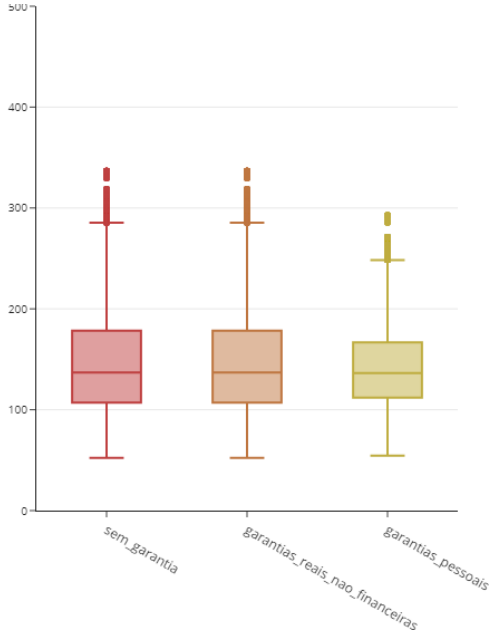

Figure 49 - Final predictions of all-in price over the type of guarantee for credit support lines

Even though there is insufficient sampling from the other types, it could be reasonable to expect that the fact of the client did not provide a single guarantee would increase the all-in price. However, the results did not align with the expectations, and several reasons may explain them beyond the difference in data sampling with and without providing a guarantee.

For instance, the client's characteristics may not imply the presence of a guarantee, as the client might be deemed capable of repaying the financing. Moreover, Meyer & Nagarajan (1996) suggest that the presence of credit guarantees can trigger a learning process for banks where they realize that borrowers who benefit from the guarantee are not as risky and unprofitable as initially anticipated. As a result of this learning, banks may become more willing to provide loans to these borrowers in the future, even without the need to provide a guarantee.

As a result, the model predictions were subsequently adjusted following internal official documents to mimic the difference that was supposed to be observed based on the type of guarantee.

5. CONCLUSIONS AND FUTURE WORKS

This report arises from the main project developed during my internship at CGD. The project aims to support the commercial chain by providing a reference value for the all-in price negotiated with clients from the *Empresas* or *Grandes Empresas* segment for a specific type of loan. To accomplish this objective, a predictive model was developed, and it was important that it could reflect the impact of the client's risk and the type of guarantee provided by the client. So, all the necessary data was collected both to construct the target and to obtain the ML model. By analyzing the data, it was identified not only that an adjustment to the initial approach was needed but also that two ML models should be produced separately due to the presence of credit support lines that had special acquisition requirements. Before testing different ML models, a sequence of data preparations was employed. Afterward, XGBoost was considered the best model for both groups and posterior analyses were conducted to assess whether the models' results aligned with the business requirements. Finally, the maintenance and monitoring of the models was discussed.

Almost the entire process was conducted using tools developed by the team to address common needs across ML projects, such as data preprocessing, data preparation, and modeling. Data analyses were performed using Jupyter Notebook.

Certainly, the main expected contribution of this project is the increase of the profitability for the loan of credit lines within this credit family for clients from *Empresas* and *Grandes Empresas*. Both models were capable of recommending all-in prices higher than the actual prices negotiated, suggesting clients that would probably be receptive to agree to pay a bit more to secure the loan. Moreover, when negotiating the all-in price with the client, the manager will have a new indicator based on the operations carried out by similar clients as a reference, which might improve his confidence throughout the negotiations. Therefore, since the model obtained for credit support lines also addresses the segment of *Negócios* and another financial product family, its application can even be explored for those groups.

In addition to increasing the profitability of this type of loan, it was also expected that the models obtained would be capable of recognizing the client risk as one of the most important variables and differentiating the price based on the type of guarantee provided by the client. These business requirements were highly important to ensure the applicability of the model, and the data analysis process conducted proved to be crucial in assessing their feasibility.

Regarding the client's risk, both models reflected the expected outcome, demonstrating that this variable has indeed an impact on the predictions and is recognized by the models as one of the most important features. However, the significance of the guarantee's type in describing the target is less pronounced. This limitation has minimal implications for credit support lines, but for normal credit lines, it would be expected that the type of guarantee could influence the suggested all-in price. However, analyses highlighted a predominance of

a type of guarantee in the data, with insufficient sampling for the remaining types, which can help justify the model results.

By developing this report, the internship becomes a more complete experience, allowing me not only to apply the theoretical and practical knowledge acquired during the master's program but also to explore and develop new capabilities. Indeed, a notable advantage of this experience in CGD was the opportunity to apply ML in a real-world scenario, comprehending and analyzing the data from a business perspective and exploring new ML techniques, models, frameworks, and tools, which helped me to better comprehend the concepts learned during the master's program. Moreover, I became more familiar with concepts from the banking sector, gaining a deeper understanding of the impact of certain decisions.

Therefore, this project faced several limitations, some of them already explored and partially mitigated to accomplish the objective. By addressing them in future works, more robust models might be achieved.

One major limitation of this project lies in the fact that the application of an ML model dependent on the operation's features would imply predicting in real-time or almost real-time, which is not currently possible. This approach certainly would be more complete, enabling to also understand the impact of operation characteristics over the predictions.

Also, several data limitations were faced throughout this report. One of them was the lower number of distinct clients, which invalidated the approach of obtaining the model at the client level. Increasing this number would also contribute to increasing the data variability and reducing the risk of biasing the model toward a specific set of input values.

Additionally, the proportion of operations related to credit support lines from *Empresas* and *Grandes Empresas* was the lowest within the Restricted group. Achieving a higher representation of this kind of operation would help to obtain a better model's generalization for this group.

Moreover, the presence of categories with low representation, both in the client's financial rating and in the type of guarantee, ends up complicating the model's generalization for these categories.

Regarding the team's tools, although they facilitate the process of ML by allowing a quick and simple application of several ML techniques, not all techniques present in common ML libraries are available. For instance, Isolation Forest and Local Outlier Factor, which are two common and effective algorithms used for anomalies detection, are currently being implemented in the team's framework and, in future works, certainly, will be explored the impact of using these algorithms for outliers' detection and removal.

Also, by developing a more sophisticated monitoring for both models, a better understanding of the models' performance can be obtained, whether to identify the strengths of each model or to detect cases where they are struggling.

Unfortunately, the explorations were limited due to time and computational resource constraints. It is recommended to conduct a more in-depth exploration of both feature exclusion and selection, as well as to test a broader range of machine learning algorithms.

Additionally, it is advisable to experiment with different numbers of folds in cross-validation and analyze their impact on the models' predictions. Exploring different techniques for hyperparameter tuning, such as GridSearch, which exhaustively tests all hyperparameter combinations, might lead to better results.

In summary, despite the limitations presented, both developed models hold great potential to support the decision-making process and increase the profitability of these products for the company. They can also serve as a solid foundation for the development of similar projects addressing other banking products. In the coming stages, refining these models based on the insights gained will likely enhance their effectiveness and expand their applicability within the company. However, there is room for improvement in the models' overall performance. Consequently, future efforts could focus on addressing the limitations presented, refining the models, and optimizing them further to align with specific needs.

BIBLIOGRAPHICAL REFERENCES

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5). https://ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf

Baijens, J., Helms, R., & Iren, D. (2020). Applying Scrum in Data Science Projects. *2020 IEEE 22nd Conference on Business Informatics (CBI)*, 30–38. https://doi.org/10.1109/CBI49978.2020.00011

Barbosa, L., & Pinho, L. S. (2016). Estrutura de financiamento das empresas. *Banco de Portugal*. https://www.bportugal.pt/sites/default/files/anexos/papers/re201601_p.pdf

Barro, R. J. (1976). The Loan Market, Collateral, and Rates of Interest. *Journal of Money, Credit and Banking*, *8*(4), 439. https://doi.org/10.2307/1991690

Basak, D., Pal, S., & Chandra Patranabis, D. (2007). *Support Vector Regression*. Neural Information Processing – Letters and Reviews. https://www.researchgate.net/profile/Srimanta-Pal/publication/228537532_Support_Vector_Regression/links/00b7d51c2a0d86d3d9000000/Support-Vector-Regression.pdf

Caixa-FEI Apoio Micro e PME. (n.d.). Retrieved January 27, 2024, from https://www.cgd.pt/Institucional/Sala-de-Imprensa/2023/Pages/Caixa-FEI-Apoio-Micro-e-PME.aspx

CGD Group Worldwide. (n.d.). Retrieved December 18, 2023, from https://www.cgd.pt/English/International-Activity-CGD/Pages/CGD-Group-Worldwide.aspx

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient boosting with categorical features support* (arXiv:1810.11363). arXiv. http://arxiv.org/abs/1810.11363

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Second edition). O'Reilly Media, Inc.

Ghanbari, M., & Goldani, M. (2021). *Support Vector Regression Parameters Optimization using Golden Sine Algorithm and its application in stock market* (arXiv:2103.11459). arXiv. http://arxiv.org/abs/2103.11459

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?* https://doi.org/10.48550/ARXIV.2207.08815

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Jafar Hamid, A., & Ahmed, T. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. *Machine Learning and Applications: An International Journal*, *3*(1), 1–9. https://doi.org/10.5121/mlaij.2016.3101

Kaczmarczyk, K. (2019). Methods for calculating loan profitability for a bank. *Central European Review of Economics & Finance*, *29*(1), 5–22. https://doi.org/10.24136/ceref.2019.001

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (Second edition). The MIT Press.

Li, G., Yang, J., Wu, C., & Ma, Q. (2019). *Maximal Margin Distribution Support Vector Regression with coupled Constraints-based Convex Optimization* (arXiv:1905.01620). arXiv. http://arxiv.org/abs/1905.01620

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? *PLOS ONE*, *12*(8), e0183250. https://doi.org/10.1371/journal.pone.0183250

Majumdar, A., Banerjee, R., Ghosh, R., Ghosh, S., Bhattacharjee, R., Pallob, S., Ghosal, A., Dhar, P., & Banerjee, N. (2022). Loan Prediction by using Machine Learning. *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY*, *9*(1). https://ijirt.org/master/publishedpaper/IJIRT155556_PAPER.pdf

Meyer, R. L., & Nagarajan, G. (1996). *EVALUATING CREDIT GUARANTEE PROGRAMS IN DEVELOPING COUNTRIES*. https://core.ac.uk/download/pdf/159565519.pdf

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Computing Surveys*, *55*(6), 1–29. https://doi.org/10.1145/3533378

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4), 7–9. https://doi.org/10.5120/ijca2017915495

RandomForestRegressor Documentation. (n.d.). Scikit-Learn. Retrieved January 6, 2024, from https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Relatorio e Contas—CGD. (2022). https://www.cgd.pt/Investor-Relations/Informacao-Financeira/CGD/Relatorios-Contas/2022/Documents/Relatorio-Contas-CGD-2022.pdf

S, R., Uzir, N., R, S., & Banerjee, S. (2016). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. *International Journal of Control Theory and Applications*. https://www.researchgate.net/publication/318132203_Experimenting_XGBoost_Algorithm_for_Prediction_and_Classification_of_Different_Datasets

Sabzekar, M., & Hasheminejad, S. M. H. (2021). Robust regression using support vector regressions. *Chaos, Solitons & Fractals*, *144*, 110738. https://doi.org/10.1016/j.chaos.2021.110738

Saluja, R., Malhi, A., Knapič, S., Främling, K., & Cavdar, C. (2021). *Towards a Rigorous Evaluation of Explainability for Multivariate Time Series* (arXiv:2104.04075). arXiv. http://arxiv.org/abs/2104.04075

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Open Journal of Social Sciences*, *5*(4), 13–22.

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Swamynathan, M. (2019). *Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python*. Apress. https://doi.org/10.1007/978-1-4842-4947-5

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Wirth, R., & Hipp, J. (2020). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.* https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

APPENDIX A

Table 4 - Example of the dataset at operation level

Operation ID	Client ID	District	Activity Sector	Business Volume	Financial Rating	All-in Price
1	1	Lisboa	Sector 1	X	7	**100**
2	1	Lisboa	Sector 1	X	7	**230**
3	1	Lisboa	Sector 1	X	7	**175**
4	2	Porto	Sector 6	Y	8	**350**
5	3	Porto	Sector 3	Z	6	**270**
6	3	Porto	Sector 3	Z	6	**160**

APPENDIX B

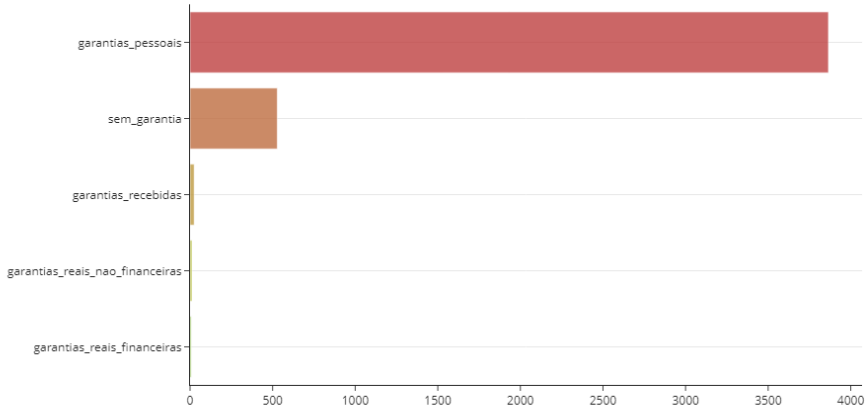

Figure 50 - Number of operations over guarantee's type for Normal group

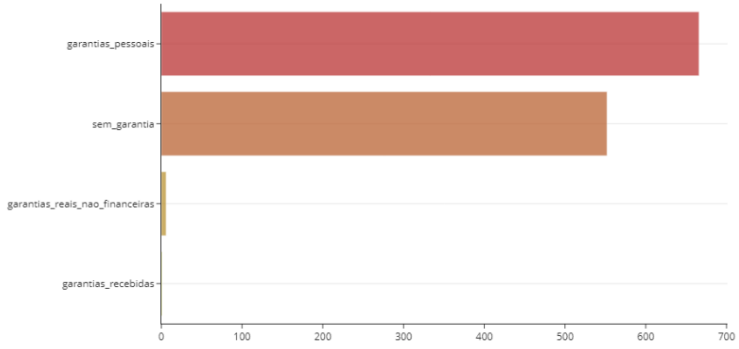

Figure 51 - Number of operations over guarantee's type for Restricted group

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa