NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Application of Predicted Models in Debt Management

Developing a Machine Learning Algorithm to Predict Customer Risk at EDP Comercial

Inês Pires Melo

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

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

Universidade Nova de Lisboa

[this page should not be included in the digital version. Its purpose is only for the printed version]

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

APPLICATION OF PREDICTED MODELS IN DEBT MANAGEMENT

by

Inês Pires Melo

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Mauro Castelli

February 2023

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 acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Inês Melo

Lisboa, 2023

ACKNOWLEDGEMENTS

I would like to express my gratitude to my internship team at EDP for their unwavering support and positive spirit throughout my internship, especially to Hélder Oliveira, who was always available to help me and believed in me from the start.

I also cannot thank my family enough, my parents and sister, who have always supported me and been a great pillar, as well as my friends and housemates, Mariana and Sofia, who put up with me every day and have my full heart.

Lastly, I would like to thank my professors whose guidance and mentorship have been instrumental in shaping not only my knowledge but also my character and overall development as an individual.

ABSTRACT

This report is a result of a nine-month internship at EDP Comercial where the main project of research was the application of artificial intelligence tools in the field of debt management. Debt management involves a set of strategies and processes aimed at reducing or eliminating debt and the use of artificial intelligence has shown great potential to optimize these processes and minimize the risk of debt for individuals and organizations.

In terms of monitoring and controlling the creditworthiness and quality of clients, debt management has mainly been responsive and reactive, attempting to recover losses after a client has become delinquent. There is a gap in the knowledge of how to proactively identify at-risk accounts before they fall behind on payments.

To avoid the constant reactive response in the field, it was developed a machine-learning algorithm that predicts the risk of a client becoming in debt by analyzing their scorecard, which measures the quality of a client based on their infringement history.

After preprocessing the data, XGBoost was implemented to a dataset of 3M customers with at least one active contract on EDP, on electricity or gas. Hyperparameter tuning was performed on the model to reach an F1 score of 0.7850 on the training set and 0.7835 on the test set. The results were discussed and based on those, recommendations and improvements were also identified.

KEYWORDS

Debt Management; Machine Learning; Scorecard Prediction; Artificial intelligence

INDEX

1. Introduction 1
   1.1. Company Overview 1
   1.2. The project 2
2. Literature review 3
   2.1. Artificial intelligence (AI) 3
   2.2. Data Mining 3
   2.3. CRISP-DM 4
   2.4. Machine Learning 6
      2.4.1. Decision Trees 6
      2.4.2. Gradient Boosting 8
      2.4.3. XGBoost 8
   2.5. Classification Model Evaluation 10
      2.5.1. Confusion Matrix 10
      2.5.2. Accuracy 11
      2.5.3. Precision 11
      2.5.4. Recall 11
      2.5.5. F1 Score 12
   2.6. Debt management in the energy sector 12
3. Methodology 14
   3.1. Business understanding 14
   3.2. Data Understanding 14
      3.2.1. Data sources 14
      3.2.2. Main Findings 16
      3.2.3. Target analysis 19
   3.3. Data Preparation 20
      3.3.1. Feature engineering 20
      3.3.2. Adding demographic features 20
      3.3.3. Missing values 21
      3.3.4. Outlier detection 21
      3.3.5. Categorical features 21
      3.3.6. Feature scaling 21
      3.3.7. Target distribution 22
      3.3.8. Dealing with correlated features 22

3.3.9. Feature Selection 23
3.4. Model Selection And Evaluation 23
4. Results and discussion 25
4.1. Evaluating first results 25
4.2. Hyperparameter tuning 25
4.3. Discuss final results 28
5. Conclusion 30
6. Limitations and recommendations for future works 31
7. References 32
Appendix 34

LIST OF FIGURES

Figura 1 - Phases of CRISP-DM Process Model Source: (Martinez-Plumed, F. et al., 2000)..... 5
Figura 2- Elements of a Decision tree..... 7
Figura 3 - L1 and L2 Regularization formula 9
Figura 4 - Average expeted annual Consumption by type of consumption 16
Figura 5 - Pie Chart of Types of Consumption 16
Figura 6 - Client distribution of years of consumption 16
Figura 7- Distribution of types of Contracts 17
Figura 8 - Target distribution 19
Figura 9 - F1 Score vs Learning rate 26
Figura 10 - F1 Score vs Number of estimators 27
Figura 11 - Confusion Matrix of Test set 28
Figura 12- Outliers Visualization 35
Figura 13- Ouliers Visualizations part 2 36
Figura 14- Pearson's correlation matrix..... 37
Figura 15 - Spearman's rank correlation matrix 38
Figura 16 - Feature Selection Decision Trees: Gini and Entropy..... 39
Figura 17- Feature Selection using Ridge Classifier..... 40
Figura 18- Feature Selection using Lasso Classifier..... 41

LIST OF TABLES

Tabela 1- Confusion Matrix 10
Tabela 2- Features and Meaning of each one 15
Tabela 3- Descriptive Statistics of the data set 18
Tabela 4- Feature transformations 20
Tabela 5- New Demographic Features and meaning 21
Tabela 6- Strengths and Weaknesses of XGBoost 23
Tabela 7- F1 Score results 25
Tabela 8 - Classification report 28
Tabela 9 - Count of Clients per District 34
Tabela 10 - Metrics grouped by client score 34

LIST OF ABBREVIATIONS AND ACRONYMS

AI: Artificial Intelligence

DM: Data Mining

ML: Machine Learning

1. INTRODUCTION

1.1. COMPANY OVERVIEW

This report is the result of a nine-month internship in the debt and collection management division at EDP Comercial. EDP - Energias de Portugal is a Portuguese electric utilities company that was established in 1976 as a result of the merger of the main companies in the electricity sector in Portugal. It operates in 19 countries, including Brazil, Spain, and the United States, and has over 11,000 employees. According to (Relatório da Qualidade de Serviço, 2021) from EDP - Energias de Portugal, the company's main businesses include the generation, transmission, distribution, and supply of electricity, as well as the supply of gas. In addition, EDP also invests in renewable energy, with a focus on wind, solar, and hydropower. The company has made significant investments in renewable energy, intending to achieve a carbon-neutral business by 2030. Additionally, the group operates in related areas such as engineering, laboratory tests, professional training, energy services, and property management. The EDP Group is composed of various companies operating in the energy sector. Some of the companies that constitute the EDP Group include:

- EDP Distribuição - responsible for the distribution of electricity in Portugal.
- EDP Renováveis - a global leader in the renewable energy sector, focusing on wind and solar power.
- EDP Comercial - responsible for the commercialization of electricity and gas in Portugal and Spain.
- EDP Produção - responsible for electricity production, including thermal and hydroelectric power plants.
- EDP Gás - responsible for the distribution and commercialization of natural gas in Portugal.
- EDP Brasil - responsible for energy generation, distribution, and commercialization in Brazil.
- EDP España - responsible for energy generation, distribution, and commercialization in Spain.
- EDP Labelec - a laboratory and testing services company that provides services to various sectors, including energy, telecommunications, and transportation.

As mentioned, EDP Comercial is a subsidiary of EDP Group that is responsible for the commercialization of electric energy and gas. In a free market with 5.4 million consumers, EDP Comercial serves more than 4 million electric energy consumers (approximately 74% of the market) and over 650000 gas consumers (around 49% of the market) (EDP - Energias de Portugal, 2021). In addition to its core business of selling electricity and gas to residential and commercial customers, EDP Comercial also offers various other services, such as energy efficiency consulting, home automation solutions, and electric vehicle charging infrastructure.

The internship was conducted within the B2C Operational Management division, which focuses on managing the company's business-to-consumer operations. Specifically, in the team responsible for managing debt and collections. The main objectives of this team are to ensure that clients in debt receive proper warnings and to recover the debt owed. The main impact of properly managing debt and collections is it helps to ensure the company's financial stability and helps to minimize losses.

1.2. THE PROJECT

Electric energy is often considered an essential good as it is necessary for powering homes and businesses, as well as for running various industrial processes. Without access to electricity, many aspects of modern life would become difficult or impossible. Additionally, access to electricity is often seen as a basic human right and a key component of economic development, so it is expected that the number of consumers is significantly large. EDP Comercial represents nearly 74% of the free market energy consumers in Portugal (Relatório da Qualidade de Serviço, 2021) and in a free market where the consumer prices are extremely competitive and growing bigger each day, the need for effective debt management is more urgent than ever. The use of data-driven and process mining techniques has become a tool ready to facilitate and improve debt recovery.

The insertion of information management technologies such as data mining and machine learning is becoming extremely useful in the process of decision-making in the sense of analyzing patterns, predicting data, and helping to make decisions based on facts and not just on the experience of the business. According to (So, 2021), debt collection is being revolutionized by artificial intelligence and machine learning. These cutting-edge technologies can analyze vast amounts of data from various sources, uncovering novel insights into delinquency risk and strategies for managing high-risk accounts.

To track and manage the quality of a client in terms of infringements, the team has a scorecard (0-9) associated with each client. A client with a score of 0 is a compliant client that has never been in debt in the past and 9 is the worst quality client possible.

The score is a way to measure the type of reactive response from the team to the client in condition to their score. The proposed theme is to build a machine-learning algorithm that can predict the score of a client before it becomes an in-debt client. The algorithm will be trained on historical data to identify patterns and factors that contribute to clients becoming in debt, and it will use this information to make predictions for new clients. This is a way that makes it possible to have a more preventive response, adverting clients from having debt and focusing more on the most at-risk consumers. This could help the company to target its debt management efforts more effectively, reduce the number of clients who fall into debt, and ultimately improve its bottom line. In the past, debt collection practices mainly involved waiting until a borrower failed to make payments and then taking action to recover the debts. However, with the advent of machine learning, there is now an opportunity to shift this approach to a more proactive one. Machine learning can help to identify accounts that are at risk of defaulting before they actually do, which allows debt collection agencies to take preventive measures to avoid or mitigate potential losses. (So, 2021)

Problem definition: **predicting the risk of a client becoming in debt by analyzing their scorecard, which measures the quality of a client based on their infringement history.**

The following questions will be answered with this report:

- How can we predict the risk of a client?
- Can data analysis give us any insight into the prediction?
- What is the best predictive model for score prediction?
- Can the hyperparameter tuning help the accuracy of the model?

2. LITERATURE REVIEW

2.1. ARTIFICIAL INTELLIGENCE (AI)

Artificial intelligence (AI) is a discipline within computer science and engineering that is devoted to comprehending the principles of intelligent actions and generating technology that mimics these actions. AI research spans a wide spectrum, ranging from designing algorithms that can learn from massive amounts of data (machine learning) to studying human cognitive processes. Certain ones study the replication of intelligent behavior, while others are interested in creating machines that can function independently. (Russell & Norvig, 2010). For many years, businesses have been implementing AI solutions to keep up with the evolving industry landscape. Recent reports from McKinsey suggest that companies are increasingly looking to invest in AI technologies to streamline processes, minimize expenses, and gain a competitive advantage in their respective markets. (Mckinsey, 2021)

In the past, AI was primarily used for automating repetitive tasks, such as data entry and customer service. However, as AI technologies have advanced, companies are now using AI to make strategic decisions and gain insights into their business operations. AI technology can be used to take over and/or complement humans in the execution of monotonous, high-volume jobs, like inputting data and categorizing documents. This allows businesses to use their human resources for more demanding, creative work that necessitates greater thought processes. Furthermore, AI can analyze big amounts of data much more quickly than people can, giving businesses insights into their operations that can aid them in making more informed decisions. (Marr, 2017)

However, as with any new technology, the adoption of AI in business has not been without its challenges. There is a growing concern about the possible effect of AI on employment. Many experts anticipate that AI will result in the replacement of jobs in various industries, which could have significant economic and social implications. This is especially a concern for low-skilled workers who might struggle to find new job opportunities in the AI-powered economy. (Brynjolfsson & Mitchell, 2017)

Despite the challenges, the use of AI in business is expected to continue to grow in the coming years, with many experts predicting that it will become an essential component of business operations. McKinsey Global Institute in 2021 stated that the adoption of artificial intelligence (AI) by some companies is already showing benefits and that it is crucial for other companies to speed up their digital transformations in preparation for the next wave of digital disruption. (Mckinsey, 2021)

2.2. DATA MINING

Data mining is a process that involves using machine learning, statistics, and database technology to identify useful patterns and trends in large, complex, and interrelated data sets. The goal is to extract meaningful information and knowledge that was previously unknown or difficult to discover with standard data analysis methods. Effective data mining requires expertise in both technical areas like machine learning and statistics, as well as in the specific domain of the data being analyzed. (Azuaje, Witten, & E, 2006)

Data mining is a relativity new complex and interdisciplinary field that involves using statistical and machine learning techniques to discover hidden patterns and trends in large datasets. The concept of

exploratory data analysis dates back to the 1960s, but it wasn't until the late 1980s and early 1990s that the term "data mining" was coined to describe the automated process of extracting meaningful information from massive amounts of data. Since then, the field has expanded to include various applications, such as predictive modeling, clustering, and association rule mining, among others. With the emergence of big data, data mining has become more critical than ever, driving innovation and progress in numerous fields. (Han, Kamber, & Pei, 2011)

Data mining is a powerful tool that has found its application in a wide range of fields, including business, engineering, medicine, education, defence, and scientific research. The goal of data mining is to uncover hidden patterns and useful insights from large and complex datasets. In business, data mining has been used to gain insights into customer behaviour, analyze sales trends, and improve marketing strategies. In engineering, it has been applied to improve the quality of products, detect and diagnose faults, and optimize production processes. In medicine, data mining has been used to develop new treatments, identify disease patterns, and create personalized care plans. In education, it has been used to understand student performance and enhance teaching practices. In defence, it has been applied to detect and prevent security threats, and in scientific research, it has been used to analyze complex data and identify patterns in fields such as genomics, astronomy, and climate change. (Tan, Steinbach, & Kumar, 2006)

Some of the commonly used data mining techniques are clustering, classification, regression, association rule mining, and outlier detection. Clustering is the process of grouping similar data points into clusters or segments, based on their similarities and differences. Classification is the process of assigning labels or classes to data points, based on a set of predefined criteria. Regression is the process of predicting numerical values or continuous variables, based on the relationship between the input and output variables. Association rule mining is the process of discovering co-occurring patterns and relationships between different items or variables. Outlier detection is the process of identifying data points that are significantly different from the rest of the data and may require special attention or analysis.

2.3. CRISP-DM

CRISP-DM is a widely used methodology for data mining and analysis, as it provides a structured and systematic approach to the process of knowledge discovery from data.

According to a study by (Azuaje, Witten, & E, 2006), The CRISP-DM methodology offers a complete guide to the entire data mining process and helps ensure that everyone involved in a project understands the process and speaks the same language. The authors also note that the CRISP-DM methodology provides a structure for managing the tasks involved in a data mining project and clarifies the roles and responsibilities of each team member.

In a review of CRISP-DM, (Watson, et al., 2000) states that the methodology emphasizes five main phases: understanding the business problem and the available data, preparing the data for analysis, modeling, evaluating the results, and deploying the models in a real-world context. Verzani also notes

that the CRISP-DM approach is flexible, allowing for modifications to be made to fit the specific needs of a project.

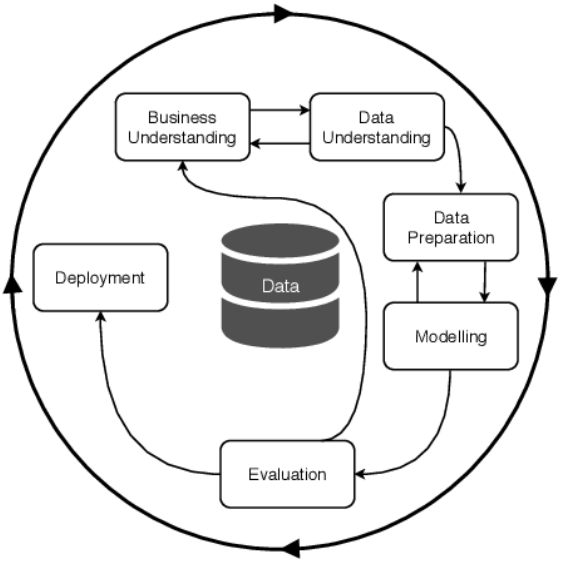

Figura 1 - Phases of CRISP-DM Process Model
Source: (Martinez-Plumed, F. et al., 2000)

As it is shown in the figure, the six steps of the CRISP-DM process are as follows:

Business Understanding: In this step, the objectives of the project are defined and the requirements of the stakeholders are gathered. The goals and expected outcomes of the project are identified, and the feasibility of the project is assessed. It is important to evaluate the business situation to understand what resources are available and what resources will be required. (Schröer, Kruse, & Gómez, 2021).

Data Understanding: the data that is to be used for the project is collected and summarized. Data exploration and visualization are performed to gain an understanding of the structure and contents of the data. (Schröer, Kruse, & Gómez, 2021).

Data Preparation: In this step, the data is cleaned, transformed, and integrated to prepare it for analysis. This may involve removing duplicates, dealing with missing values, and dealing with outliers. To select the appropriate data for analysis, specific criteria need to be established for what data should be included and excluded. If the data is of low quality, cleaning methods can be employed to address the issue. (Schröer, Kruse, & Gómez, 2021).

Modeling: In this step, statistical and machine learning models are applied to the data to uncover patterns and relationships. This step is where the data is mined for insights.

In the data modeling phase, the appropriate modeling technique is chosen and a test case and model are developed. The selection of the data mining technique to be used typically depends on the business problem and the data available. (Schröer, Kruse, & Gómez, 2021).

Evaluation: In this step, the results of the modeling process are evaluated to determine if they meet the goals and expectations of the project. During the evaluation phase, the outcomes are measured against the predetermined business goals. This requires interpreting the results and identifying any necessary follow-up actions. Additionally, it is important to conduct a general review of the entire data mining process. (Schröer, Kruse, & Gómez, 2021).

Deployment: In this final step, the results of the project are deployed and the findings are communicated to the stakeholders. The results can be used to make decisions, solve problems, or support operations.

2.4. MACHINE LEARNING

Machine learning is a subset of artificial intelligence that concentrates on developing algorithms capable of learning from data and making predictions or decisions. These algorithms have found numerous applications, including image recognition, natural language processing, and forecasting customer behaviour. (James, Witten, Hastie, & Tibshirani, 2021).

There are four main types of machine learning:

- Supervised learning: a type of machine learning in which the input data used to train the algorithm includes the correct answers, also known as labels. This type of learning is often used for classification tasks, such as distinguishing between spam and non-spam emails in a spam filter. The algorithm is fed examples of both types of emails, along with their corresponding labels, and uses this information to learn how to accurately classify new, previously unseen emails. (Géron, 2019)
- Unsupervised learning: a type of machine learning that does not use labelled data for training. Instead, it allows the algorithm to find patterns and relationships in the data on its own. In other words, the algorithm must identify similarities and differences in the data without being told what these patterns are. Clustering is a common application of unsupervised learning, where the algorithm groups similar data points together based on their characteristics. (Géron, 2019)
- Semi-Supervised learning: the model is trained using a mixture of labelled and unlabeled data. The goal of semi-supervised learning is to make the most of the labelled data while also taking advantage of the unlabeled data to improve the model's performance.
- Reinforcement learning: the model learns through trial and error. The model receives feedback in the form of rewards or penalties for its actions, and it uses this feedback to make decisions about which actions to take in the future. Reinforcement learning is a technique used in robotics, gaming, and navigation, among other areas, to train a model by allowing it to learn through trial and error without human guidance. (Géron, 2019)

In the energy field, machine learning is increasingly being applied to improve the efficiency and effectiveness of energy production, distribution, and consumption.

The next topics will discuss several machine learning algorithms that will be used or it is important to know the concept of them in this research.

2.4.1. Decision Trees

Decision Trees are a popular and widely used supervised machine learning technique. They are used for both regression and classification problems and they are tree-based models that are easy to interpret and implement. Decision Trees are used to determine a decision or prediction by breaking down a large dataset into smaller and simpler sub-problems until reaching a conclusion.

As stated by (Breiman, Friedman, Olshen, & Stone, 1984), A decision tree is a graphical representation that uses tree-like branches to display different possible outcomes based on the values of various attributes or features. Each branch in the tree represents a different path or choice, and each internal node is a test or decision based on one of the attributes. The terminal nodes, or leaves, represent the final classification or prediction. An example of a graphical representation of a decision tree:

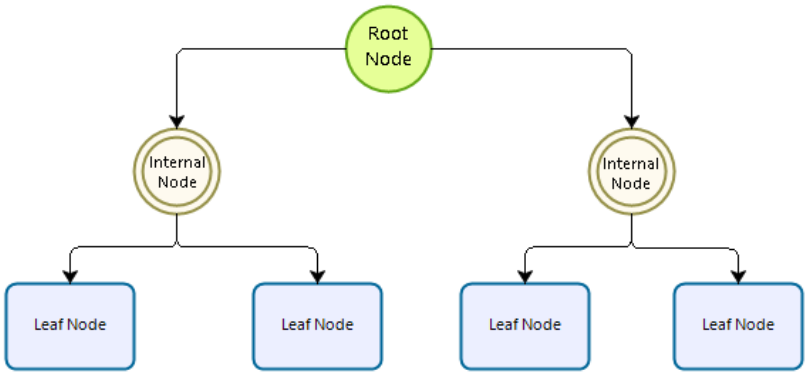

Figura 2- Elements of a Decision tree

Construction of Decision Tree

The learning algorithms for decision trees construct a tree by recursively splitting instances into subsets (James, Witten, Hastie, & Tibshirani, 2021). The splits in the tree are determined based on the feature values, with the objective of maximizing the reduction in impurity in the target variable. The selection of the method for dividing the data at each level of a decision tree affects both the shape of the tree and the effectiveness of the classification. Two popular methods for this are the Gini impurity and information gain. (Hastie, Tibshirani, & Friedman, 2008)

"Gini impurity" is calculated as the sum of the squared probabilities of each class occurring in the node, and ranges from 0 (pure node) to 1 (equal distribution of classes). Suppose we have a dataset D with samples from k different classes. The probability of a sample belonging to a specific class i at a certain node can be represented by p_i. The Gini Impurity of D can be defined as:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

The idea behind “information gain” is that the feature that provides the most information about the target variable, or the most reduction in the uncertainty of the target variable, should be selected as the splitting feature. A feature with a high information gain is considered to be more useful for splitting the data into different classes or categories. In theory, it represents the reduction in entropy achieved by partitioning the data on a particular feature. Entropy, on the other hand, is a measure of the amount of uncertainty or randomness in the data. In decision tree models, it is used to evaluate the purity of a set of examples. The formula for calculating information gain is as follows:

Information Gain = Entropy(parent) - [Weighted Average]*Entropy(children)

Where Entropy(parent) is the entropy of the parent node, and Entropy(children) is the entropy of the child nodes created by splitting the parent node. The weighted average is taken over all child nodes,

with the weight given by the proportion of examples in each child node. Entropy is calculated as follows:

$$E(S) = \sum_{i=1}^{c} p_i \, log_2 p_i$$

Where S represents the sample in which we want to calculate the entropy and p is the proportion of examples in a given class. The entropy is maximum when the data is equally distributed among all classes, and minimum when all the data belongs to a single class.

2.4.2. Gradient Boosting

Gradient boosting is a popular algorithm used in machine learning that can solve various classification and regression problems. It works by combining multiple weak prediction models, often decision trees, into a single strong prediction model, known as an ensemble. This process is done in a way that gradually improves the overall performance of the model. (Friedman, 2001)

In a gradient boosting algorithm, the base learner is typically a decision tree. The algorithm builds an initial model, and then in each subsequent iteration, a new decision tree is added to the ensemble that tries to correct the errors of the previous trees. As described by (Chen & Guestrin, 2016), Gradient boosting is an algorithm that creates a prediction model by adding multiple weak prediction models together in a step-by-step manner. This process enables it to optimize any loss function that can be differentiated.

The main idea behind Gradient Boosting aims to minimize the error of a model by adding new models to the existing ones. This is done in a step-by-step manner, where each new model is added to minimize the loss function at each iteration. The method can be thought of as an approximation to the steepest descent optimization of a general loss function. (Hastie, Tibshirani, & Friedman, 2008).

One of the most popular implementations of gradient boosting is XGBoost (Chen & Guestrin, 2016), which has become a de facto standard for many machine learning tasks. XGBoost provides a highly efficient and scalable implementation of gradient boosting that can handle large datasets and high-dimensional feature spaces.

2.4.3. XGBoost

XGBoost stands for Extreme Gradient Boosting and is an advanced gradient boosting algorithm. XGBoost has been a popular choice for machine learning and data science tasks because of its high level of efficiency, robustness, and accuracy, specially when handling problems that involve large datasets and multi-class classification.

In XGBoost, several decision trees are combined together to make more accurate predictions, which is known as an ensemble learning method. At its core, XGBoost uses decision trees as weak learners, which are combined to form a stronger model. It creates these decision trees iteratively, with each new tree attempting to correct the errors made by the previous tree, this is known as gradient boosting. (Chen & Guestrin, 2016)

XGBoost goes beyond the standard gradient boosting framework by adding several enhancements, including:

Regularization:

To prevent the model from overfitting, the XGBoost algorithm uses regularization techniques such as weight decay, L1 and L2 regularization, and column subsampling.

L1 and L2 are two popular forms of regularization, a technique used in machine learning to reduce overfitting. L1 regularization is also called Lasso regularization and adds the absolute values of the coefficients to the loss function. L2 regularization is also called Ridge regularization and adds the squares of the coefficients to the loss function. (Géron, 2019).

$$L1\ Regularization\ =\ (loss\ function) + \alpha \sum_{j=1}^{p} |b_j|$$

$$L2\ Regularization\ = (loss\ function) + \alpha \sum_{j=1}^{p} b_j^2$$

Figura 3 - L1 and L2 Regularization formula

L1 regularization tends to produce sparse models where many of the coefficients are zero, which can be useful for feature selection, while L2 regularization produces models with smaller but non-zero coefficients. The strength of the regularization penalty is controlled by a hyperparameter, with higher values resulting in more regularization and smaller coefficients.

Adjusting the hyperparameters of L1 and L2 regularization can help find a balance between a model's complexity and its capacity to apply to new data, resulting in improved performance on the test set. (Hastie, Tibshirani, & Friedman, 2008).

Handling missing values:

XGBoost can handle missing values automatically. It can handle missing values in the data by learning how to replace them with a default value. It does this by learning a direction for missing value imputation during the tree construction process. At each node, the algorithm checks if a feature is missing a value and decides on the direction to take based on the gain, which is calculated as the difference between the parent node's loss and the weighted sum of losses in the child nodes. This process helps to improve the performance of the algorithm on data with missing values. (Chen & Guestrin, 2016)

Parallel processing:

XGBoost can take advantage of multiple CPUs or GPUs to train and make predictions faster. XGBoost is known for its ability to efficiently use available computational resources, and one of its main strengths is its ability to parallelize the construction and evaluation of decision trees. This allows XGBoost to take advantage of multi-core processors and distributed computing environments for faster processing. (Chen & Guestrin, 2016)

Weighted instances:

XGBoost allows you to assign weights to instances, which can be useful for imbalanced datasets. Weighting instances is a technique that can be used to balance the positive and negative weights of imbalanced datasets. By assigning weights to the instances, the algorithm can focus on the underrepresented class and improve its performance on the test set. (Chen & Guestrin, 2016)

The final model in XGBoost is a combination of all the decision trees created during training. Each tree outputs a prediction, and these predictions are combined using weighted averaging. The weights assigned to each tree depend on its performance during training.

2.5. CLASSIFICATION MODEL EVALUATION

Classification metrics are used to assess the accuracy and performance of a classification model. These metrics provide a way to measure how well the model can correctly classify instances into different classes based on the input features. Evaluation of classification models is essential to build a reliable and effective model. Many of these metrics are based on the Confusion Matrix, which summarizes the actual and predicted classification outcomes, making it a useful tool to analyze the model's performance. (Grandini, Bagli, & Visani, 2020).

2.5.1. Confusion Matrix

A confusion matrix is a technique used to evaluate the performance of a classification model on a dataset where the outcomes are already known. It offers a means to measure the accuracy of the model and is particularly useful when dealing with imbalanced categories. (Géron, 2019)

A confusion matrix displays the actual classes of the data along one axis and the predicted classes along the other. Each cell of the matrix shows the count of instances that were classified as belonging to a particular class. The cells along the diagonal correspond to the correctly classified instances, while the cells off the diagonal correspond to the instances that were classified incorrectly.

		Actual values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TP

Tabela 1- Confusion Matrix

In other words, is composed of four main components: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). According to (Géron, 2019), True positives refer to the instances where the model correctly identified the positive class, while false positives are the instances where the model incorrectly identified a negative instance as positive. True negatives refer to the instances where the model correctly identified the negative class, and false negatives are the instances where the model incorrectly identified a positive instance as negative.

2.5.2. Accuracy

Accuracy is a widely used metric for evaluating classification models that measures the number of correct predictions divided by the total number of predictions. It provides an overall idea of the model's performance by computing the proportion of true positives and true negatives in the predicted data. (Hossin & M.N, 2015)

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

While accuracy is a useful metric for balanced datasets, it can be misleading in the case of imbalanced datasets where the number of instances in each class is different. In such cases, a model that always predicts the majority class will have high accuracy, but will not necessarily perform well in predicting the minority class.

2.5.3. Precision

Precision is a classification metric that quantifies the accuracy of positive predictions made by a model. It is defined by dividing the number of true positives by the sum of true positives and false positives. In other words, precision measures the proportion of true positive predictions out of all the positive predictions made by the model. (Géron, 2019).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

A high precision means that the model is predicting correctly a high proportion of positive cases. Precision is commonly used to evaluate the performance of a model in situations where the cost of a false positive is high, i.e., when the model falsely classifies a negative instance as a positive class. In such cases, the focus is on minimizing the number of false positives, and the precision metric is used to measure the proportion of true positives among the predicted positives.

2.5.4. Recall

Recall (otherwise known as sensitivity or true positive rate) is a measure of the classifier's capability to correctly identify positive occurrences. It is calculated by dividing the number of true positive predictions by the sum of true positives and false negatives, representing the proportion of actual positive instances that are correctly identified by the classifier. (Géron, 2019).

This means that recall is a measure of how well the classifier is able to correctly identify instances of the positive class (i.e., true positives) among all instances that actually belong to the positive class (true positives + false negatives).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

In situations where the number of positive instances is significantly lower than the number of negative instances, imbalanced datasets for example, recall is often considered a more useful metric than accuracy. This is because recall measures the ability of the model to correctly identify important instances, even if they are rare, while accuracy only measures the overall number of correct predictions. (Fawcett, 2006).

2.5.5. F1 Score

According to (Géron, 2019), The F1 score is a metric that takes into account both precision and recall to provide an overall evaluation of the performance of a classification model.

$$F1 = 2\,\frac{precision \times recall}{precision + recall}$$

It is calculated as the harmonic mean of precision and recall, and is useful in situations where both metrics are important. This metric is particularly helpful in scenarios where there is an imbalance between the number of positive and negative instances or when the cost of false positives and false negatives is not the same.

The F1 score ranges from 0 to 1, with 1 being the best possible score. A high F1 score indicates good performance in both precision and recall.

2.6. DEBT MANAGEMENT IN THE ENERGY SECTOR

Debt management B2C in the energy sector refers to the strategies and practices that companies in the energy industry use to manage their debts and financial obligations towards their customers (business-to-consumer, or B2C).

Debt management is a critical aspect of credit risk management for utilities and is essential for maintaining financial stability and minimising default risk (Esgalhado, Higginson, Jacques, Matecsa, & Selandari, 2019). The report highlights the importance of effective debt collection practices, regular communication with customers, and the use of advanced technologies to automate and streamline debt management processes.

One of the several goals of customer debt management in the energy sector is to provide education and resources to help customers better understand their energy usage and costs. This can include providing information on energy efficiency and conservation on their invoices, as well as assistance with identifying and accessing financial assistance programs.

Customer debt management in the energy sector requires also working with customers to develop payment plans or other solutions to manage and reduce their debt. This means negotiating extended payment plans, connecting customers with financial assistance programs, and even providing temporary energy assistance to help clients keep their lights on during difficult financial times. Energy companies may also offer energy efficiency programs, on-bill financing and other solutions to help customers reduce their energy usage and costs. The paper from McKinsey also notes that utilities can benefit from using digital technologies and automation to enhance their debt management practices, including credit scoring, collections, and restructuring. This approach can help reduce operating costs and increase the amount of debt collected. Furthermore, digital technologies can provide significant improvements in customer experience throughout the debt collection process. Some utilities have developed a comprehensive digital platform for delinquent customers, covering debt overview, payment support, renegotiation, and customer support. (Esgalhado, Higginson, Jacques, Matecsa, & Selandari, 2019)

Overall, debt management in the energy sector is a complex and dynamic field that requires a comprehensive understanding of financial risk and consumer behavior. Companies must continuously

innovate and adapt their debt management strategies to ensure financial stability and minimize default risk.

3. METHODOLOGY

3.1. BUSINESS UNDERSTANDING

The first step needed is to gather all the important information about the client data available on the EDP database. The way to achieve that success is by gaining an in-depth understanding of the business and its operations. For that, CRISP-DM was the methodology used.

It is clear that the team has a few main obstacles to overcome in order to improve the effectiveness of debt recovery: get an insight into who is most likely to become in debt. And why? And after knowing that, how can we improve that?

Objective: build a machine learning classification model that identifies patterns that can help predict which customers are most likely to default on their loans.

Scope:

- Identify the main variables that correlate with a customer being a defaulter of payments
- Test different models and their parameters to choose the one with the best accuracy
- Learn and assess patterns for each class
- Apply the model and help identify and mitigate potential financial risks

3.2. DATA UNDERSTANDING

3.2.1. Data sources

It was collect all the important information about the client data available on the EDP database, which means all the variables corresponding to the information of the contract(s) and consumption of the clients. Unfornetlly, as EDP can not have access to information such as purchasing power, annual income or other information that leads us the assume the probability of a person becoming in debt, it was collected as much information available as possible including their demographic information, consumption patterns, billing information, and payment history. The collection of the data was based on SAS tables connected to the database servers of the company and performed on SAS Enterprise Guide.

The dataset is composed of all active clients of the company, meaning, the ones that had at least one active contract on electricity or gas. That makes a total of 3M people. The features used were:

Column name	Meaning
NOME_PRIMEIRO	Name of the client
COD_CAE	Type of consumption: Domestic or other
COD_CATEGORIA_CLIENTE	Customer category: Person or organization
COD_TIPO_CLIENTE	Client type: Domestic or non-domestic
COD_CLASSE_CONTA	Account Class: Small business or residential
COD_VALOR_PRESENTE	Level of promotional campaigns gifts
FLG_EDP_ONLINE	if it uses the App or not

Contratos Ativos	Number of active contracts
Contrato ativo mais Antigo	the seniority of the oldest active contract (days)
MEAN_of_FLG_TARIFA_SOCIAL	Percentage of accounts that have social help
MEAN_of_QTD_POTENCIA_INST	Mean of the installed Power in all places of consumption
MEAN_of_QTD_CONSUMO_ANUAL_ESP	Mean of the expected annual consumption in all places of consumption
MAX_of_QTD_POTENCIA_INST	Highest Power installed
Contratos Encerrados	Number of closed contracts associated with the client
Antiguidade Cliente	days of client seniority
Flg_Tel	if it has a phone registered
Flg_Email	if it has an email registered
Distrito	Main District of the client
MEAN_of_FLG_FATURA_ELETRONICA	Percentage of accounts with electronic invoices
MEAN_of_FLG_DEB_DIRETO	Percentage of accounts that pays with direct debit
MEAN_of_Pack01	if it has the base pack
Postal code	Postal code
Tipo Contratos	A- energy only B- gas only C- Services Only D- Energy and gas
Variedade Serviços	how many different services did the client request
Quantidade de Serviços	how many times did the client request services
Multibanco	Percentage of accounts that pays with an atm
Dia PGMT	the average day of the month the client pays the bill
Tempo até pagamento	days of delay between the due date and the payment date
Clientes antigos	how many different clients shared the same place/s of consumption of the current client
avg_score_clienteant	The average score of past clients
locais de consumo antigos	Number of places of consumption of the client
Score	Client score

Tabela 2- Features and Meaning of each one

3.2.2. Main Findings

An analysis of the data was conducted, and the results revealed some key insights that are particularly relevant to the business. The following are the main observations that were drawn from this analysis.

Based on the data, it appears that the majority of the clients (89.36%) are domestic consumers. This is not surprising given that the field is B2C, meaning that the primary customers are individuals and households rather than organizations. This high percentage of domestic consumers indicates that the business is primarily focused on managing the energy and/or gas services provided to individual homes, rather than to large organizations or commercial properties. Given these observations, it was explored if there were any notable differences in the consumption patterns of domestic vs. non-domestic consumers.

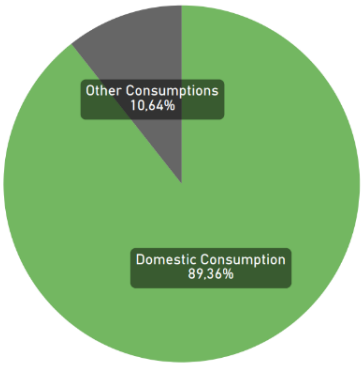

Figura 5 - Pie Chart of Types of Consumption

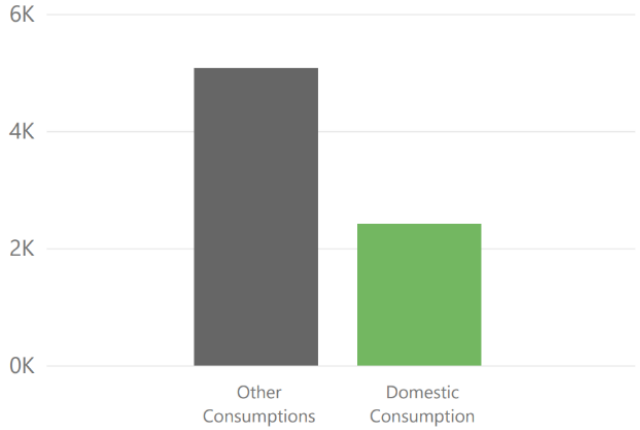

Figura 4 - Average expeted annual Consumption by type of consumption

The data also shows that the majority of consumers have been customers for 5-10 years. This could indicate that these clients have established a long-term relationship with the company and may be less likely to switch to a different provider.

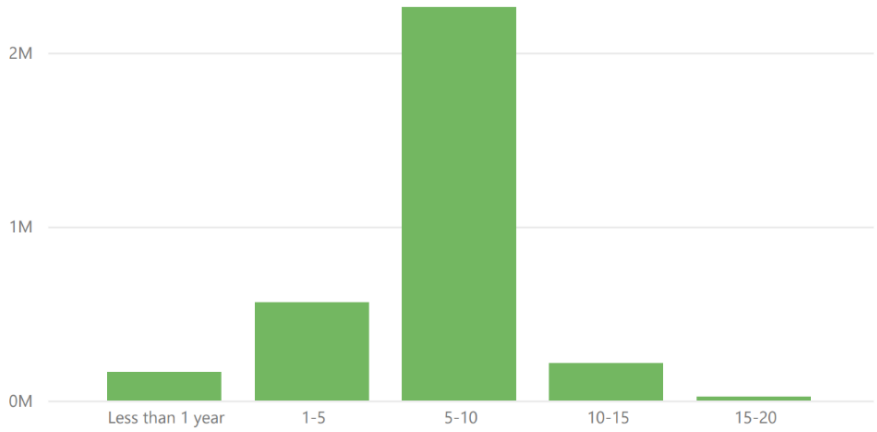

Figura 6 - Client distribution of years of consumption

Since the sample of customers used is only those who have an associated contract, i.e. gas or energy, it is normal that there is no representation of class C (services only). It seems the majority of clients (over 80%) have an electric-only contract with the company and there is a small number of clients with a gas-only contract (less than 1% of the total). The remaining clients (just under 20%) have both electric and gas contracts.

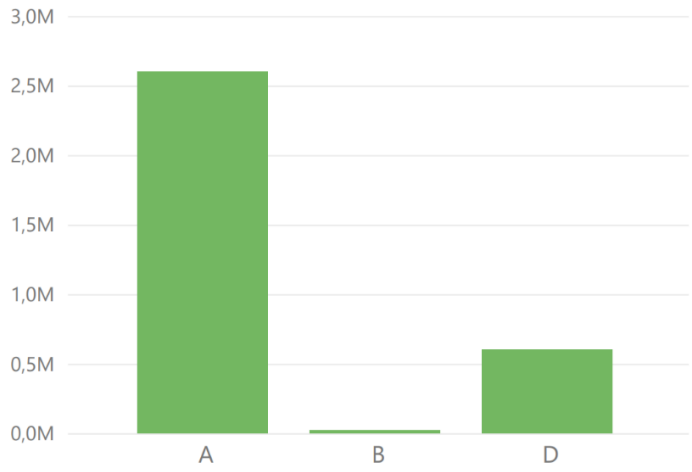

Figura 7- Distribution of types of Contracts

To further the analysis of the data it was performed descriptive statistics on the dataset, obtaining summary statistics such as the mean, standard deviation, minimum and maximum values, and quartiles.

Feature	count	mean	std	min	25%	50%	75%	max
Contratos Ativos	3225590	1,41	0,85	1	1	1	2	134
Contrato ativo mais Antigo	3225590	2155,75	1133,26	0	1256	2403	2960	5708
MEAN_of_FLG_TARIFA_SOCIAL	3225590	0,14	0,35	0	0	0	0	1
MEAN_of_QTD_POTENCIA_INST	3225590	8,17	5,3	0	6,9	6,9	7,76	999,9
MEAN_of_QTD_CONSUMO_ANUAL_ESP	3225590	2700,09	3835,5	-20196	996	1920	3228	936630
MAX_of_QTD_POTENCIA_INST	3225590	9,28	5,82	0	6,9	6,9	10,4	999,9
Contratos Encerrados	1137341	2,05	2,38	1	1	1	2	513
Antiguidade Cliente	3225590	2538,63	1106,29	3	1955	2709	3236	5708
MEAN_of_FLG_FATURA_ELETRONICA	3225590	0,48	0,5	0	0	0	1	1
MEAN_of_FLG_DEB_DIRETO	3225590	0,61	0,49	0	0	1	1	1
MEAN_of_Pack01	3225590	0,94	0,24	0	1	1	1	1
Variedade Serviços	1014783	1,81	1,27	1	1	1	2	23
Quantidade de Serviços	1014783	34,41	30,13	1	10	29	52	2332
Multibanco	3225590	0,31	0,46	0	0	0	1	1
Dia PGMT	3154275	16,03	8,66	1	9	16	23	31
Tempo até pagamento	3154275	-2,73	16,41	-38	-5	-5	-2	1554

Clientes antigos	789234	1,58	1,11	1	1	1	2	140
avg_score_clienteant	789231	1,1	1,46	0	0	0,5	2	9
locais de consumo antigos	1137340	1,66	1,7	1	1	1	2	272

Tabela 3- Descriptive Statistics of the data set

This helped to gain an initial understanding of the distribution of the data and identify any potential outliers or anomalies. The output of the descriptive statistics was also helpful in selecting appropriate transformation techniques for the data. By conducting descriptive statistics, a baseline was established for the data and provided a foundation for further analysis and modeling. Based on the statistical description of the dataset, some observations can be made:

About the contracts: The average number of active contracts is 1.41 indicating that there are some customers with multiple active contracts. The mean of the oldest active contract is 2155.75 days (almost 6 years), with a standard deviation of 1133.26 days. The minimum value is 0, indicating that there may be some new customers in the dataset. The average number of closed contracts is 2.05, with a standard deviation of 2.38 but the maximum number of closed contracts is 513, indicating that some customers have closed a large number of contracts.

About the properties of the contracts: The mean value of the FLG_TARIFA_SOCIAL variable is 0.14, indicating that only a small percentage of customers are eligible for a social tariff. The mean value of QTD_POTENCIA_INST is 8.17, with a standard deviation of 5.3 but the maximum value is 999.9, which could be an outlier. The minimum value of QTD_CONSUMO_ANUAL_ESP is -20196, which seems to be an invalid value and may need further investigation. The mean value of Pack01 is 0.94, indicating that most customers have subscribed to the first package.

About services: The average variety of services used by customers is 1.81, with a standard deviation of 1.27. This indicates that some customers use multiple services. The average number of services used by customers is 34.41, with a standard deviation of 30.13 and the maximum number of services used is 2332, indicating that some customers use a large number of services.

About payments: The average value of Multibanco is 0.31, indicating that a small percentage of customers use this payment method. The mean value of FLG_DEB_DIRETO is 0.61, indicating that most customers have chosen direct debit as a payment method. The mean value of FLG_FATURA_ELETRONICA is 0.48, indicating that almost half of the customers have opted for electronic invoicing. The mean value of Dia PGMT is 16.03, indicating that customers tend to pay their bills around the middle of the month. The mean value of Tempo até pagamento is -2.73, indicating that on average, customers pay their bills 2.73 days before the due date.

About the places of consumption: The average number of “Clientes antigos” is 1.58, with a standard deviation of 1.11. The maximum number of old customers is 140, indicating that some locations have previously been with customers connected to the company. The mean value of avg_score_clienteant is 1.1, with a standard deviation of 1.46, indicating that some of the previous customers with the location of the current customer have a score above 0 and have been in debt with the company. The average number of old consumption locations is 1.66, with a standard deviation of 1.7 and the maximum number of old consumption locations is 272, indicating that some customers have closed contracts on more than one previous location.

3.2.3. Target analysis

The target being analyzed and worked on is based on the customer risk scorecard of the company. The scorecard has the following consideration factors:

- If it has active contracts
- If it has electronic invoices active
- If it has direct debit activated
- The monthly and annual average bill
- The number of different debt warnings made, such as shut-off notices, payments reminders, power cuts, and complaint letters
- The percentage of client debt and lapsed debt
- The number of:
    - Fully paid payment agreements,
    - Cancelled payment agreements
    - Active payment agreements
- The amount of debt associated with the payment agreements

Based on these rules, the more defaults, the higher the score of the client.

After consulting with the business area, it was decided to divide the score into four categories: 0, 1-3, 4-6, and 7-9. This decision was made due to the fact that proportionally speaking, there are not many people different from the score zero, meaning that it is already a significant difference in terms of business if someone has a score above zero. The categories made are grouped scores that are not so different in terms of business but are different from each other (between categories). For example, a person with a score of zero is a lot different than one with a score of one, but a person with a score of one is not so different from a person with a score of 2 or 3.

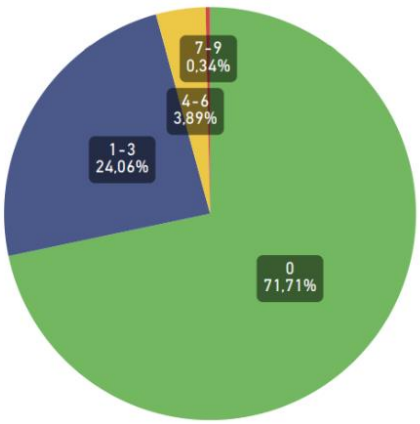

Figura 8 - Target distribution

3.3. DATA PREPARATION

Data transformation is a major step in the data preparation process that is indispensable for building accurate and effective models. The main goal is to convert the raw data into a format that can be easily understood and analyzed by a machine learning algorithm.

3.3.1. Feature engineering

Feature engineering is a crucial step in the preprocessing of a dataset prior to training a machine learning model. After deleting duplicated rows, the changes made to the columns were:

Feature	Transformation
Contratos Encerrados	Fill the missing values with zero, because a missing value on closed contracts means no closed contracts
Variedade Serviços	Fill the missing values with zero, because a missing value on services variety means no services were requested
Quantidade de Serviços	Fill the missing values with zero, because a missing value on services quantity means no services were requested
NOME _PRIMEIRO	The name of the client was used to determine the gender, a new column named *gen* was created and the name was dropped
Antiguidade Cliente	Transformed from days to years
Contrato ativo mais Antigo	Transformed from days to years
Clientes antigos	Fill the missing values with zero, because a missing value on former clients means no former clients are associated with the current
avg_score_clienteant	Fill the missing values with zero, the same logic as above
locais de consumo antigos	Fill the missing values with zero, the same logic as above
Score	As explained in section 3.2.3, the score was divided into four classes (0, 1-3, 4-6, 7-9)

Tabela 4- Feature transformations

3.3.2. Adding demographic features

To better predict the risk of debt, it was necessary to add demographic information to the dataset. The company only had basic information about its clients, which was not ideal for accurately predicting their risk of debt. Information such as age, number of people in the household, annual income, etc. was missing.

To solve this problem, some variables were created to help get a notion of the financial state of the consumer location. This corresponds to the location of the consumer's residence, which is publicly known, such as the local population, average purchasing power, average geographical age, etc.

These demographic features might help to provide a more complete picture of the client and can be useful in making more accurate predictions about their risk of debt. The features added to the data were:

Column	Meaning
0-14	number of people in the range of age from 0 to 14 years old
15-24	number of people in the range of age from 15 to 24 years old
25-64	number of people in the range of age from 25 to 64 years old
65 e mais	number of people in the range of age 65+
Homens	number of men
Mulheres	number of women
Poder de compra per capita	Per capita purchasing power

Tabela 5- New Demographic Features and meaning

3.3.3. Missing values

The best approach to dealing with missing values depends on the specific dataset, the amount of missing data, and the overall goals of the analysis. A small number of missing values were identified, with a maximum of 3% of the total and only on categorical features. To address this issue, the chosen method was to fill these missing values with the mode, which means filling with the most frequent value in the respective feature column. This method was chosen because it is a straightforward way to handle missing values, especially for small amounts of missing data.

3.3.4. Outlier detection

Identifying and removing outliers is a crucial step in data preprocessing that should be performed prior to applying any statistical or machine learning techniques to the data. Outliers can have a major impact on the outcome of analyses. Therefore, it is important to handle outliers before proceeding with data analysis to ensure accurate and reliable results. (Enderlein, 1987).

After a combination of visualization through box plots, knowledge of the business and the Interquartile Range (IQR), the difference between the 75th and 25th percentile of the data, meaning the observations that are less than Q1 - 1.5 * IQR or greater than Q3 + 1.5 * IQR are considered outliers, 2.63% of the data were considered outliers and eliminated.

3.3.5. Categorical features

To represent categorical variables, such as: COD_CAE, COD_CATEGORIA_CLIENTE, COD_TIPO_CLIENTE, COD_CLASSE_CONTA, COD_VALOR_PRESENTE, FLG_EDP_ONLINE, MEAN_of_FLG_TARIFA_SOCIAL, Flg_Tel, Flg_Email, Distrito, MEAN_of_FLG_FATURA_ELETRONICA, MEAN_of_FLG_DEB_DIRETO, Tipo Contratos, Multibanco, PONTOS DÍVIDA, gen, it was used the common technique of dummy variables. This will help improve the performance of the model, as they capture the relationship between the original categorical variable and the target variable.

3.3.6. Feature scaling

Feature scaling is an important preprocessing step in many machine learning algorithms. It refers to the process of transforming the variables or features in the dataset to a common scale. This helps to ensure that each variable or feature contributes equally to the model, avoiding bias and improving the performance of the algorithm. As noted by (Géron, 2019), machine learning algorithms tend to work more effectively or reach optimal results more quickly when the features have been standardized to be on similar scales. Without feature scaling, variables or features with larger values can dominate the distances, leading to suboptimal performance. Therefore, the dataset was normalized using the

method of min-max scaling. Min-max scaling, also known as normalization, scales the values to a range between 0 and 1.

3.3.7. Target distribution

Analysing the target distribution, it's clear it's an imbalanced target variable, which could impact the accuracy of the predictions. Having an unbalanced dataset can harm the accuracy of the predictive model in several ways. This is because when the target classes are imbalanced, the model will learn to predict the majority class more often, leading to a high accuracy score but with a low precision and recall score. This means that the model will not perform well in identifying the minority class, leading to biased results. Additionally, an unbalanced dataset can lead to overfitting on the majority class, making it difficult for the model to generalize to unseen data. Therefore after considering different techniques to handle this issue, it was decided to stratify the data when splitting it into a training and testing set, that involves distributing the target variable evenly among the two sets, so that the proportion of target categories in both sets is similar to the proportion of target categories in the entire dataset. This was done to ensure that the training set and testing set are representatives of the entire dataset and that there is a balance of the target categories in both sets.

Another technique considered was undersampling. It is a technique that involves reducing the size of the majority class to balance the target variable. Although it is an effective way to handle imbalanced datasets, it can also lead to a loss of important information and may result in a biased model. This was the case since it was found that undersampling resulted in a significant reduction in the size of the dataset, which in turn impacted the performance of the predictive model.

Therefore, it was decided not to use undersampling and instead use stratified sampling, as discussed earlier, to ensure a balanced distribution of target variables in the training and testing sets. This approach helped to prevent the over-representation or under-representation of any particular target category, which could have resulted in biased models or inaccurate model performance evaluation.

3.3.8. Dealing with correlated features

Correlation analysis is a valuable technique in the process of selecting relevant features for a target variable, as it can help identify predictors that may have a strong relationship with the target. (Kelleher, Mac, Aoife, & Arcy, 2015). In the context of data analysis, correlation analysis helps identify features that are highly correlated between themselves and the target variable. This information can then be used to make informed decisions about which features to include or exclude from a predictive model.

The correlation was calculated using various methods, including Pearson's correlation coefficient and Spearman's rank correlation coefficient. These methods measure the linear relationship between two variables, with values ranging from -1 (strong negative correlation) to 1 (strong positive correlation). Features that are highly correlated with each other (multicollinearity) or with the target (high correlation) - correlation coefficient of more than 0.8 with each other and more than 0.7 with the target- were removed during correlation analysis to avoid unstable and irrelevant solutions in the model since high correlations between the features and target can result in overfitting or redundant information.

3.3.9. Feature Selection

Preparing data for training models often involves feature selection, which is crucial for improving model performance and reducing computational time, especially when dealing with large datasets. By reducing the dimensionality of the data, feature selection helps to identify the relevant predictors and avoid the problem of having too many potential predictors, which can make the modeling process more difficult and time-consuming (Guyon & Elisseeff, 2003). Therefore, using predictive models for feature selection can help in identifying the most important features that have a strong impact on the target variable. This leads to a more efficient and effective model, as well as increased interpretability. The algorithms experimented to select the features for the model were:

- Embedded methods such as Decision trees, ridge classifier and lasso classifier.
- Filter-Based Methods such as Correlation
- Wrapper methods such as RFE (recursive feature elimination), were used for the features selected by the methods above.

3.4. Model Selection And Evaluation

After the data is fully prepared and explored, it is ready for model selection. The first step made was to split the data into train and test, with the train representing 70% and the remaining 30% the test part. To ensure all classes are being represented while the data is being trained by the model, it was applied 10-fold cross validation as well.

Since we are dealing with a multiclassification problem, a high number of features, and the presence of both categorical and numerical features, the following models were considered for training: Random Forest Classifier, Decision Trees Classifier, Neural Networks and XGBoost.

Given this research is facing a very large size dataset, which contains millions of observations and numerous features, I need a model that could handle such large and complex data while still producing accurate predictions. The main obstacle here in the selection and implementation of the model is computational resources. Considering that, and the balance between the weaknesses and strengths of the model, I chose to use XGBoost primarily because of its efficiency and scalability with computational resources.

Strengths:	**Weaknesses:**
High predictive accuracy	Computationally intensive, especially when dealing with large datasets or complex models.
Flexibility: can be used for a wide range of tasks, including feature selection, missing value imputation, and anomaly detection.	Sensitivity to hyperparameters, and tuning these hyperparameters can be time-consuming and computationally expensive.
Scalability: highly scalable and can handle large datasets with millions of features and millions of samples.	Interpretability: XGBoost models can be difficult to interpret
Parallel processing, which can speed up training times.	Potential for overfitting
Regularization, which can help prevent overfitting and improve generalization performance.	

Tabela 6- Strengths and Weaknesses of XGBoost

When dealing with imbalanced datasets, accuracy may not be a reliable metric as it only measures the proportion of correct predictions. Instead, it is essential to use an evaluation metric that considers both precision and recall to assess the model's overall performance in a balanced manner. This approach is necessary because traditional metrics may not account for the false positive and false negative rates, which can significantly impact the effectiveness of the model.

The F1 score is a popular metric used in classification tasks to measure the balance between precision and recall. It is particularly useful when the dataset is imbalanced, meaning that one class has a much larger number of samples than the others, which is the case.

Precision and recall are two important metrics in classification. Precision measures how many of the predicted positive samples are actually positive, while recall measures how many of the actual positive samples are correctly predicted. In other words, precision focuses on the accuracy of positive predictions, while recall focuses on the ability to find all positive samples.

The F1 score is the harmonic mean of precision and recall. It is calculated as follows:

$$F1 = 2\,\frac{precision \times recall}{precision + recall}$$

The F1 score ranges from 0 to 1, with 1 being the best possible score. A high F1 score indicates good performance in both precision and recall. By choosing the F1 score, I am able to evaluate the trade-off between precision and recall and get a single metric that summarizes the overall performance of my model. This is particularly important when the cost of false positives and false negatives is not equal, as it allows me to optimize the model for the particular application at hand.

4. RESULTS AND DISCUSSION

This chapter presents the results of applying the model and discusses the implications of the findings. It starts by providing an overview of the selected model followed by the hyperparameter tuning process. Then, it reports the final performance of the model with the selected evaluation metrics and compared it to baseline models. Finally, the limitations are discussed and the potential future directions of this study are mentioned.

4.1. EVALUATING FIRST RESULTS

The model was first applied with the default parameters, using different sets of variables constructed during feature selection. It generated the following results:

	F1 Score	
Features	**Train**	**Test**
Decision Trees	0.7836	0.7800
Correlations	0.7804	0.7772
Ridge Regression	0.7816	0.7788
Lasso Regression	0.7611	0.7590
RFE	0.7873	0.7839

Tabela 7- F1 Score results

The results show that different feature selection methods can have a significant impact on the performance of the XGBoost model. In this case, the RFE (Recursive Feature Elimination) method resulted in the highest F1 scores on both the training and testing datasets. This indicates that the selected features by this method were the most relevant and informative for the classification task. These features were the following: Contratos Ativos, Contrato ativo mais Antigo, MEAN_of_QTD_CONSUMO_ANUAL_ESP, Contratos Encerrados, Antiguidade Cliente, MEAN_of_Pack01, Variedade Serviços, Tempo até pagamento, avg_score_clienteant, COD_CAE_other, COD_CATEGORIA_CLIENTE_2, COD_VALOR_PRESENTE_A, COD_VALOR_PRESENTE_B, COD_VALOR_PRESENTE_M, COD_VALOR_PRESENTE_MA, COD_VALOR_PRESENTE_MB, MEAN_of_FLG_TARIFA_SOCIAL_1, Flg_Tel_1, Flg_Email_1, Distrito_BEJA, Distrito_COIMBRA, Distrito_SANTAREM, Distrito_VISEU, MEAN_of_FLG_DEB_DIRETO_1, Tipo Contratos_D, Multibanco_1.

On the other hand, the Decision Trees and Correlations methods produced F1 scores that were lower than RFE, but still relatively high. This suggests that the features selected by these methods also have some predictive power, but may not be as informative as those selected by RFE. The Ridge and Lasso Regression methods produced the lowest F1 scores, indicating that the selected features may not be as relevant or informative for the classification task as those selected by the other methods.

4.2. HYPERPARAMETER TUNING

The choice of hyperparameters can significantly affect the performance of a model, and hyperparameter tuning is the process of finding the best combination of hyperparameters that results in the best model performance and possibly less overfitting.

Finding the best hyperparameters is crucial for creating an accurate machine learning algorithm. However, there is no universal rule for selecting hyperparameters, as different values can have a significant impact on the model's predictive ability. Using cross-validation, we can estimate the test error for different hyperparameter values, and select the best one. Yet, with a large number of possible values for each hyperparameter, finding the best set of hyperparameters can be a daunting task. To address this, techniques like grid search, randomized search, and Bayesian optimization can be used to navigate the vast hyperparameter space and find the best combination of values. (James, Witten, Hastie, & Tibshirani, 2021)

Since the most well-performed set of features were the RFE, it is the one that is going to be used for hyperparameter tuning. The first step made to adjust parameters was to analyse the learning rate and the number of estimators of the model and how affects its performance.

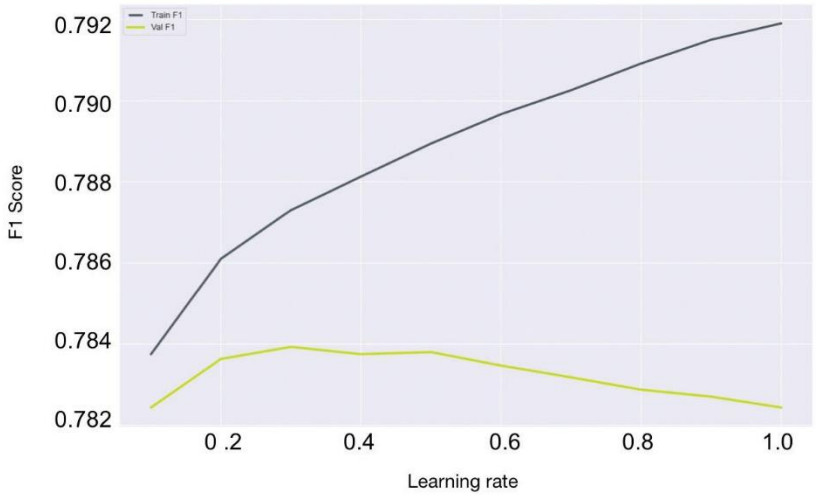

Figura 9 - F1 Score vs Learning rate

To prevent overfitting, it is used the learning rate, also called shrinkage, which reduces the weights of each feature during boosting iterations proportionally to their current weights. This helps achieve better generalization performance, but it may take longer to converge. A smaller learning rate is usually paired with a higher number of trees to compensate for the slower convergence. (XGBoost Documentation — xgboost 1.7.4 documentation)

This means that a smaller learning rate may result in longer training times since the model takes smaller steps to optimize the objective function. On the other hand, a larger learning rate may result in faster training times, but the model may converge to a suboptimal solution and may require additional steps to improve the performance, and clearly, that is shown in the figure. Therefore, the choice of the learning rate should be balanced between the success rate and performance time based on the available computing resources, meaning the main focus of looking at the figure is to find the optimal trade-off.

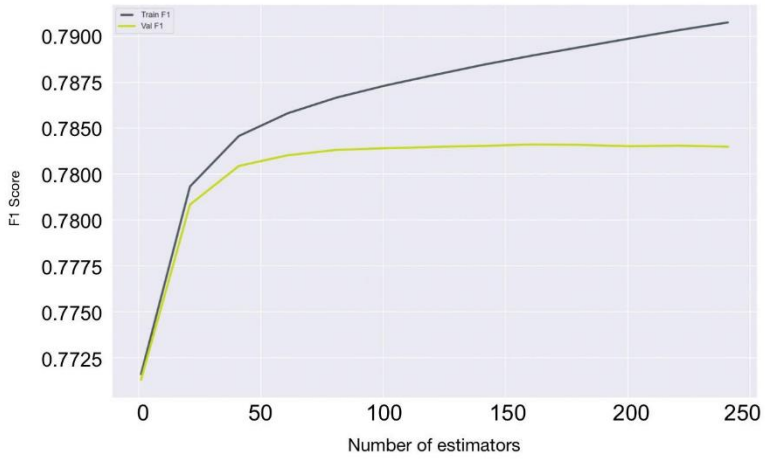

Figura 10 - F1 Score vs Number of estimators

The number of estimators is a hyperparameter in the XGBoost algorithm that refers to the number of decision trees to be used in the model. We can see in the figure that increasing the number of estimators can improve the performance of the model, but can also significantly increase the training time and memory requirements. In general, increasing the number of estimators beyond a certain point may lead to diminishing returns in performance improvement, while significantly increasing the time and computational resources required to train the model.

While choosing the method to use to explore the hyperparameters, it was considered grid search and randomized search. Given the large dataset, the big set of different parameters to explore, and the limited computational resources, it may not be feasible to use grid search, which exhaustively searches over all possible combinations of hyperparameters, as it may take a very long time to complete. Therefore, alternative methods such as randomized search, which randomly samples a subset of hyperparameters, may be more practical and efficient in this scenario. According to (Bergstra, Ca, & Ca, 2012), they found that random search can achieve the same level of performance for hyperparameter tuning as grid search, but with fewer attempts and less time spent on computation. This is because the search space for hyperparameters is complex and includes both continuous and discrete parameters, making the grid too coarse to capture all the details. The study also found that although random search may produce performance estimates with high variance, increasing the number of attempts can compensate for this without significantly increasing computational costs compared to grid search.

Therefore, it was decided to use randomized search instead of grid search to tune the hyperparameters of the model. After running 100 iterations for the search, it was found a set of parameters that were able to reduce the overfitting of the model, even though the score was not improved. That leaves us with the final result of 0.7850 on the training set and 0.7835 on the test set.

4.3. DISCUSS FINAL RESULTS

This chapter is dedicated to interpreting the results and the predictions of the model. In the case of a classification model, it is important to evaluate the performance of the model in a meaningful way, taking into account the specific requirements of the task at hand. To do this, I have generated a classification report that presents precision, recall, f1-score, and support for each class in the dataset. This provides a detailed overview of the model's performance in each class, allowing us to evaluate its strengths and weaknesses. By considering these metrics, we can gain insights into how well the model is able to identify different types of instances and we can make informed decisions about how to improve its performance.

	Precision	**Recall**	**F1-Score**	**Support**
0	0,84	0,94	0,89	651070
[1-3]	0,65	0,46	0,54	218689
[4-6]	0,64	0,32	0,43	34352
[7-9]	0,57	0,09	0,15	2828
accuracy			0,8	906939
macro avg	0,47	0,45	0,5	906939
weighted avg	0,78	0,8	0,78	906939

Tabela 8 - Classification report

Looking at the results, we can see that the model achieved high precision (0.84) and recall (0.94) for class 0, showing that it performed well in identifying instances of this class. However, for classes 1-3 and 4-6, the precision and recall are lower, indicating that the model did not perform as well to correctly identify these classes. Class 7-9 has the lowest precision, recall, and F1-score, indicating that the model struggles the most in identifying instances of this class.

The macro-averaged F1-score is 0.50, indicating that the performance of the model is moderate, however, the macro-average F1 score treats all classes equally and calculates the F1 score for each class separately, and then takes the unweighted average of those scores. This means that the contribution of each class to the final score is the same, regardless of the number of samples in that class. Therefore, it is more appropriate to use metrics that consider the class distribution, such as the weighted average or micro-average F1 score. The weighted average F1 score calculates the average F1 score, weighted by the number of samples in each class, giving more weight to the performance of the classes with more samples.

As seen before we can observe the weighted average F1-score is 0.78, indicating the overall performance of the model is good but suggesting that the model performed better on the larger

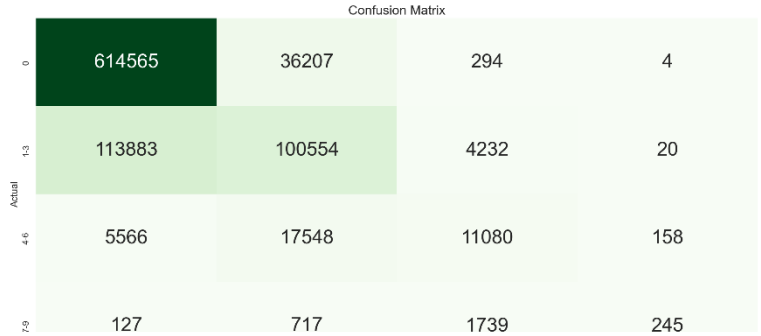

Figura 11 - Confusion Matrix of Test set

classes and struggled with the smaller ones. To further analyse the strengths and weaknesses of the model it was constructed a confusion matrix.

After constructing and analyzing the confusion matrix, the main observations that were drawn are:

As seen before it is clear that class 0 has a bigger sample and it is correct to assume that the model performed better in this class by looking at the classification report. We can further confirm this by looking at the confusion matrix as well since it has the highest number of true positives and the lowest number of false positives and false negatives. We can also see that the class most mistaken by class 0 is class 1-3. The same happens for class 1-3 in the sense that is most mistaken by class 0. Class 1-3 has a large number of false positives, which means that many samples are incorrectly predicted to belong to class 1-3 when they actually belong to another class. On the other hand, class 4-6 has a large number of false negatives, which means that many samples that belong to class 4-6 are incorrectly predicted to belong to another class. In class 7-9 we can also conclude that is often mistaken by class 4-6. Also from class 7-9, we can see that has the lowest number of true positives, indicating that it is the most difficult class to predict accurately. Overall, we can observe from the confusion matrix that it is common for a class to be misclassified as its neighbouring class. This pattern can be expected, as there may not be a significant difference in the features between these neighbouring classes. Therefore, a classifier may have a higher chance of making false positive or false negative errors when distinguishing between these classes.

5. CONCLUSION

This report was based on the main project developed during the internship and it explored the use of machine learning algorithms for predicting the risk of clients in terms of debt score at EDP Comercial. The project addresses the urgent need for effective debt management in a competitive and growing market, where data-driven techniques can significantly improve debt recovery. To achieve that demand it was created a predictive model that is able to identify patterns in clients based on their scorecard, which measures the quality of a client based on their infringement history. The data set was constituted of client contracts information, services required, payment history and information on the places of consumption. Before applying the model, the data preparation and feature engineering process were crucial to the performance of the model, including adding features to help the model succeed. After training the model using XGBoost classifier, was conducted hyperparameter tuning to help reduce the overfitting of the model and the model ended up achieving an overall performance of 0.78. It was concluded that one of its main weaknesses is correctly identifying the lower classes, but it has good performance in predicting the bigger classes.

Although more work is needed to optimize the model's performance and expand the sample size of the data set and add more information on the client's financial profile, this project shows promising results and provides a foundation for future work on debt recovery management.

By being proposed this project the majority of this internship gave me exposure to the practical application of data-driven techniques, specifically the theoretical and practical knowledge acquired during the Master's program. This was achieved by applying data analysis, process mining, and machine learning to customer data and therefore it was possible to develop a predictive model that can identify at-risk accounts before they fall behind on payments, allowing the company to respond preventively and proactively. The collection of the data was made using SAS Guide but all the rest was continued on Jupyter Notebook, the main system used during the master in several courses. During the internship, I also had the opportunity to gain a wealth of new knowledge and insights, particularly in the areas of business operations and management of a large portfolio of clients in the energy industry. Through this experience, I was able to explore new algorithms and software tools, which helped me to better understand and apply the concepts learned during my Master's program.

Overall, the internship provided a valuable and relevant experience that enhanced my understanding of the energy industry and solidified my knowledge of the subjects already studied in the master’s program.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

In the process of developing a model to predict client debt, I encountered a few limitations that should be addressed in future works. Firstly, the size of the dataset is very large and building a better model that can handle this amount of data requires significant computational resources. That was the main obstacle in this project and for further exploration, it is important to level up the computation resources and search for more efficient and scalable algorithms that can handle such a large dataset.

Secondly, the imbalance in the number of samples across classes, particularly for the lower sampled classes, may have had a negative impact on the model's performance. One way to fix this issue in the future is by collecting more data for the underrepresented classes. This can lead to a better representation of the different classes in the dataset, which can improve the model's ability to accurately predict the risk of default.

Lastly, while we collected as much information about the clients as possible, one big limitation regarding the client information is there may be additional variables that are not currently included in the dataset that could improve the model's performance. Future works can explore the inclusion of additional features related to the client's purchasing power, annual income, and other related factors. This can lead to a more comprehensive model that can better capture the complexities of the client's financial situation.

In summary, I believe this model has a high potential for improving the way debt management is handled. Despite the limitations presented, the developed model is able to assist the operations team in predicting clients who are at higher risk of default and acting more proactively in debt recovery. By leveraging the available data, the model provides a useful tool for identifying clients who require more attention and resources in debt recovery efforts. Although there is room for improvement in the model's accuracy and ability to handle the imbalanced dataset, it still has the potential to reduce the operational costs associated with debt collection and improve the overall efficiency of the recovery process. As such, future efforts could focus on addressing the limitations and enhancing the model's performance to better serve the needs of the operations team. In the end is imperative that the model can handle large datasets, address the class imbalance issue, and include more features that can improve the model's performance.

7. REFERENCES

Aurélien Géron. (2019, 6). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION.*

Azuaje, F., Witten, I., & E, F. (2006, 1). *Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques.*

Bergstra, J., Ca, J., & Ca, Y. (2012). *Random Search for Hyper-Parameter Optimization Yoshua Bengio.* http://scikit-learn.sourceforge.net.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.*

Brynjolfsson, E., & Mitchell, T. (2017, 12). *What can machine learning do? Workforce implications.* American Association for the Advancement of Science. https://doi.org/10.1126/science.aap8062

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* http://dx.doi.org/10.1145/2939672.2939785

Delen, D., Sharda, R., & Turban, E. (2013, 12). *Business Intelligence: A Managerial Perspective on Analytics.*

EDP - Energias de Portugal. (2021). *Relatório da Qualidade de Serviço.* https://www.edp.pt/pdf/relatorio_qualidade_servico.pdf

Enderlein, G. (1987). *Hawkins, D. M.: Identification of Outliers. Chapman and Hall, London – New York 1980, 188 S., £ 14, 50.*

Esgalhado, B., Higginson, M., Jacques, F., Matecsa, M., & Selandari, F. (2019). *Getting a grip on bad debt: Practical steps to help utilities boost their resilience.*

Fawcett, T. (2006, 6). *An introduction to ROC analysis.* North-Holland.

Friedman, J. (2001, 10). *Greedy function approximation: A gradient boosting machine.* Institute of Mathematical Statistics.

Grandini, M., Bagli, E., & Visani, G. (2020, 8). *Metrics for Multi-Class Classification: an Overview.* AN OVERVIEW. https://arxiv.org/abs/2008.05756v1

Guyon, I., & Elisseeff, A. (2003). *An Introduction to Variable and Feature Selection.*

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems).*

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction.*

Hossin, M., & M.N, S. (2015, 3). *A Review on Evaluation Metrics for Data Classification Evaluations.*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition.*

Kelleher, J., Mac, B., Aoife, N., & Arcy, D. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, workflows, and case studies.*

Marr, B. (2017, 7). *The Biggest Challenges Facing Artificial Intelligence (AI) In Business And Society.*

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., . . . Flach, P. (2000, 8). *Crisp-dm: towards a standard process modell for data mining.* IEEE Computer Society.

Mckinsey. (2021, 12). *The state of AI in 2021.*

Russell, S., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach.*

Schröer, C., Kruse, F., & Gómez, J. (2021, 1). *A Systematic Literature Review on Applying CRISP-DM Process Model.* Elsevier.

So, K. (2021, 1). *How AI Is Modernizing The Collections Process.*

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining Instructor's Solution Manual.*

Watson, H., Grecich, D., Shearer, C., Herdlein, S., Fong, J., Wong, H., & Fong, A. (2000). *Statement of Purpose The CRISP-DM Model: The New Blueprint for Data Mining* . Retrieved from www.dw-institute.com/journal.htm

XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/

APPENDIX

District	Nº of clients
	44 863
AVEIRO	204 129
BEJA	47 131
BRAGA	269 780
BRAGANCA	52 840
CASTELO BRANCO	72 222
COIMBRA	135 946
EVORA	46 900
FARO	182 553
GUARDA	59 248
ILHA DA GRACIOSA	5
ILHA DA MADEIRA	277
ILHA DAS FLORES	1
ILHA DE PORTO SANTO	4
ILHA DE SANTA MARIA	3
ILHA DE SAO JORGE	13
ILHA DE SAO MIGUEL	148
ILHA DO FAIAL	17
ILHA DO PICO	11
ILHA TERCEIRA	44
LEIRIA	157 827
LISBOA	675 414
PORTALEGRE	35 133
PORTO	542 329
SANTAREM	141 679
SETUBAL	262 242
VIANA DO CASTELO	92 949
VILA REAL	75 294
VISEU	126 588

Tabela 9 - Count of Clients per District

SCORE	Average of Contratos Encerrados	Average of Contratos Ativos	Average of Quantidade de Serviços
0	1,9	1,39	31,25
1	2,17	1,46	37,55
2	2,23	1,46	40,49
3	2,25	1,49	41,04
4	2,31	1,45	41,73
5	2,88	1,62	42,01
6	2,72	1,48	42,9
7	3,14	1,53	45,12
8	3,48	1,44	47,49
9	4,1	1,32	43,53

Tabela 10 - Metrics grouped by client score

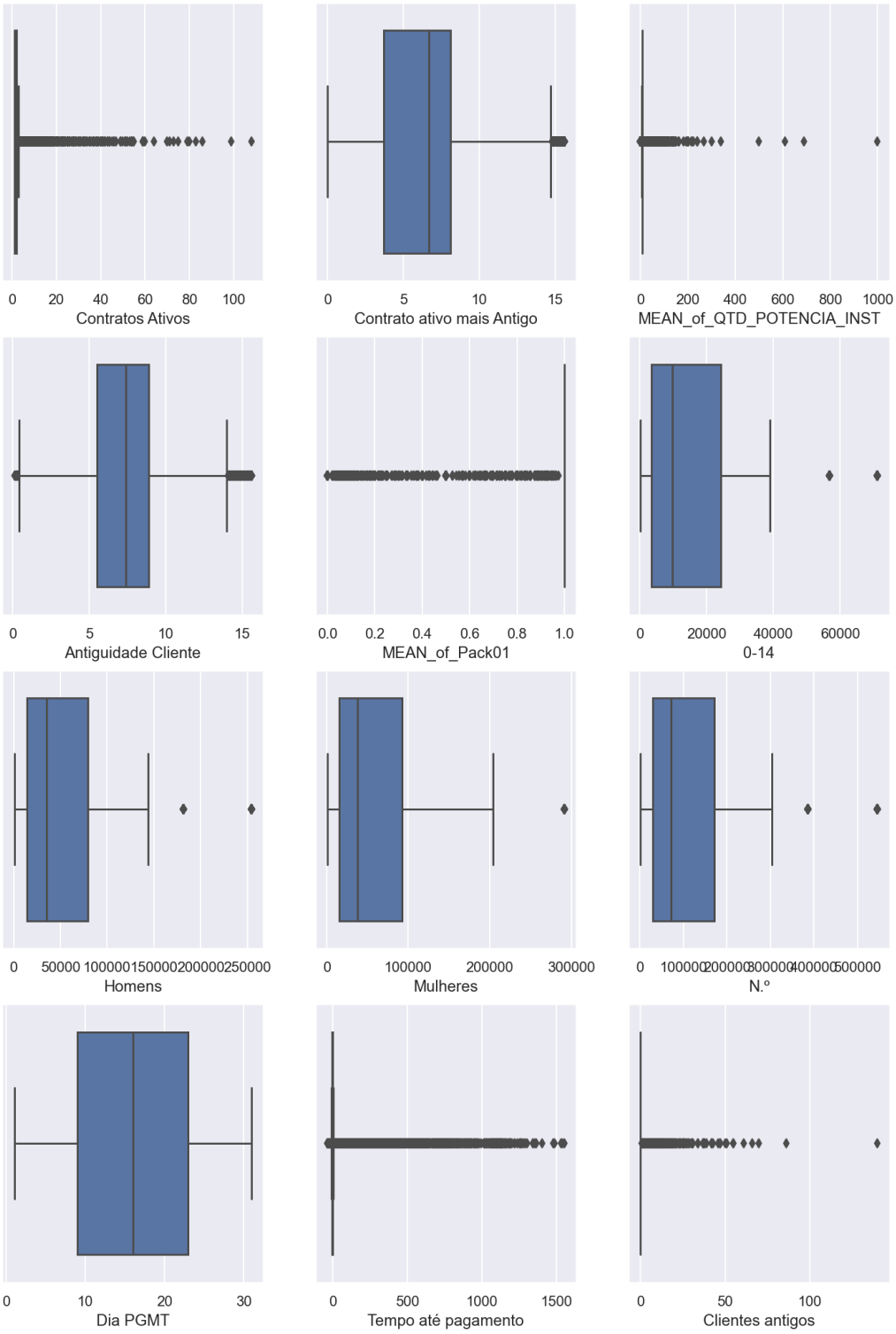

Figura 12- Outliers Visualization

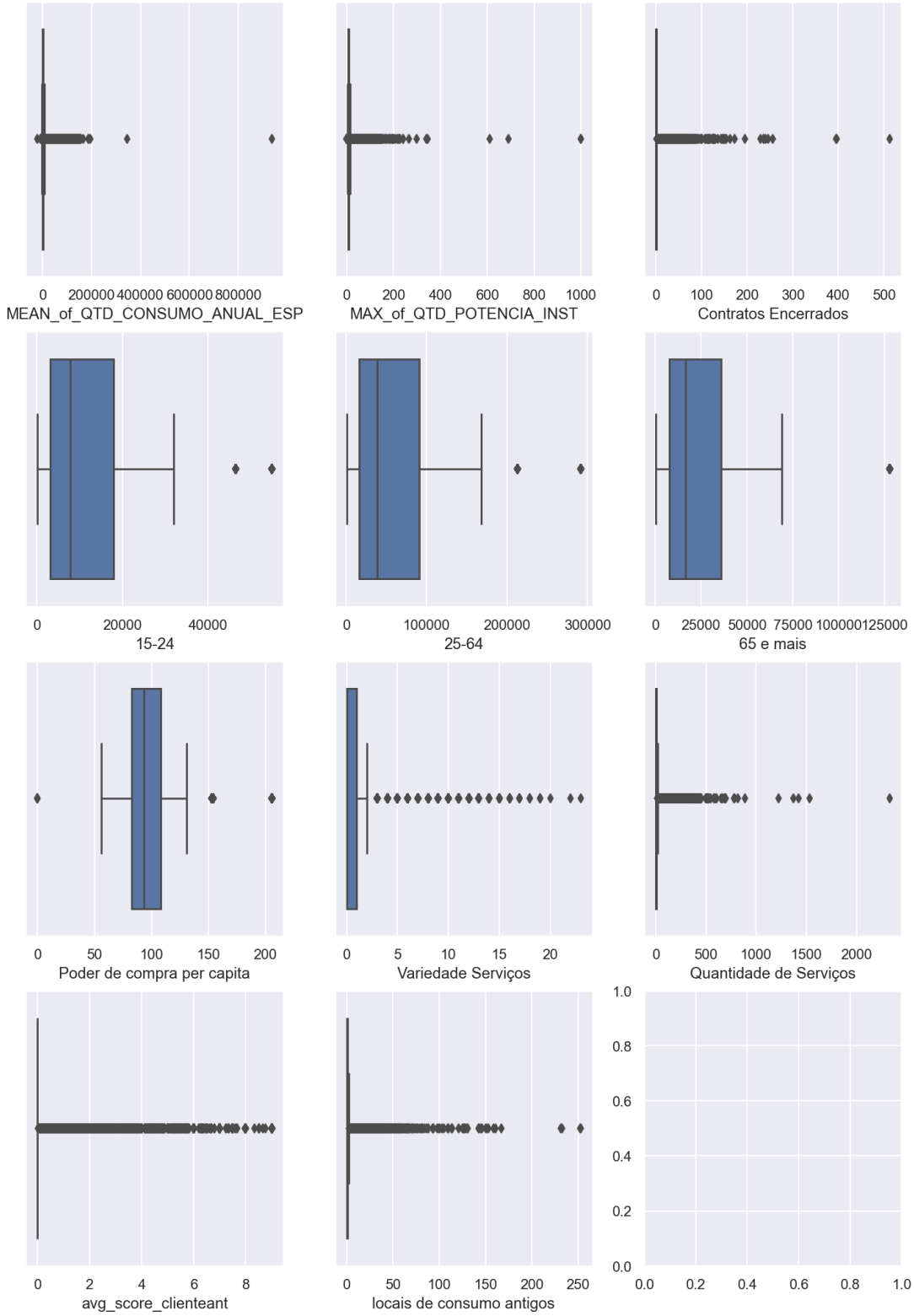

Figura 13- Ouliers Visualizations part 2

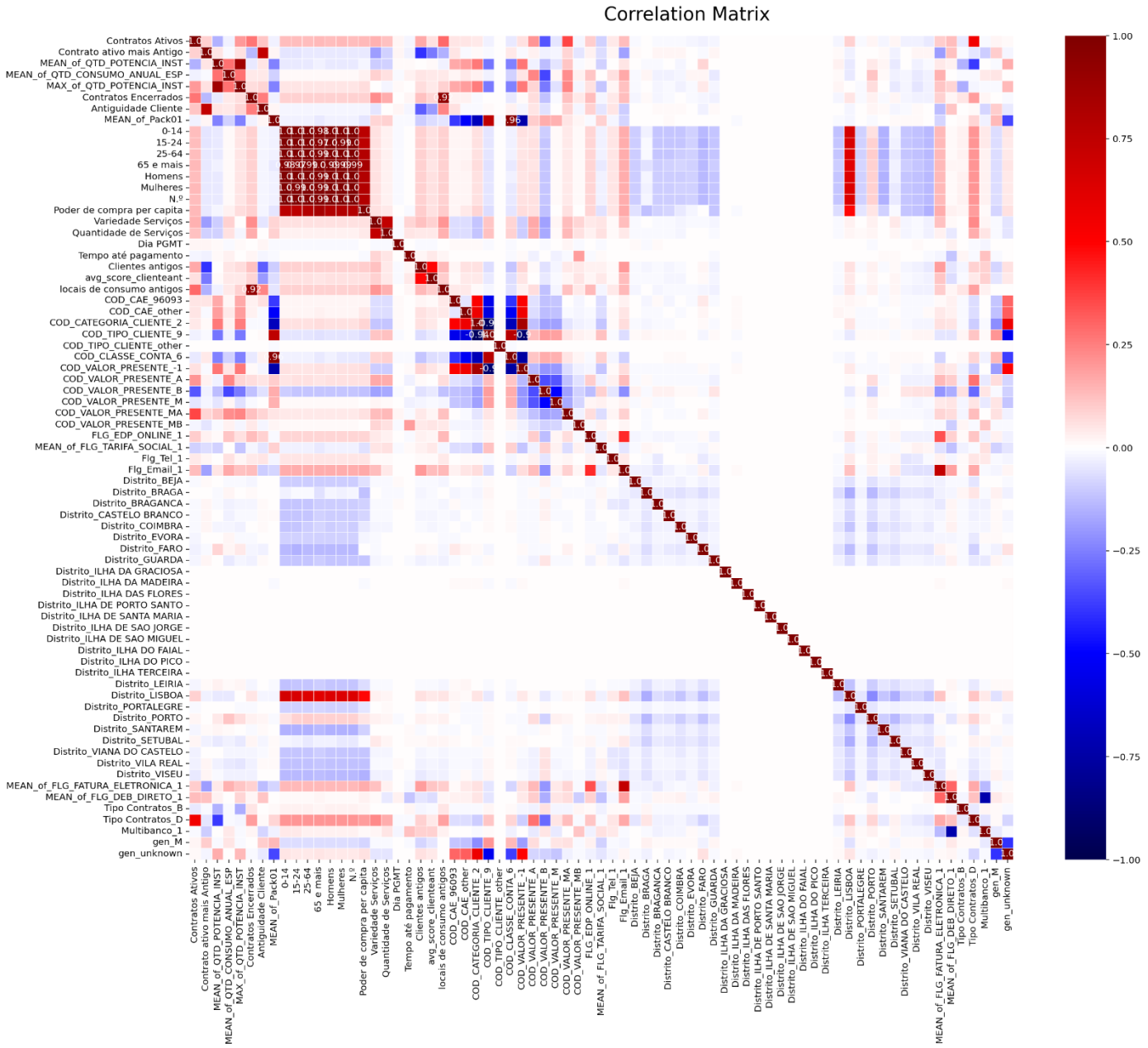

Figura 14- Pearson's correlation matrix

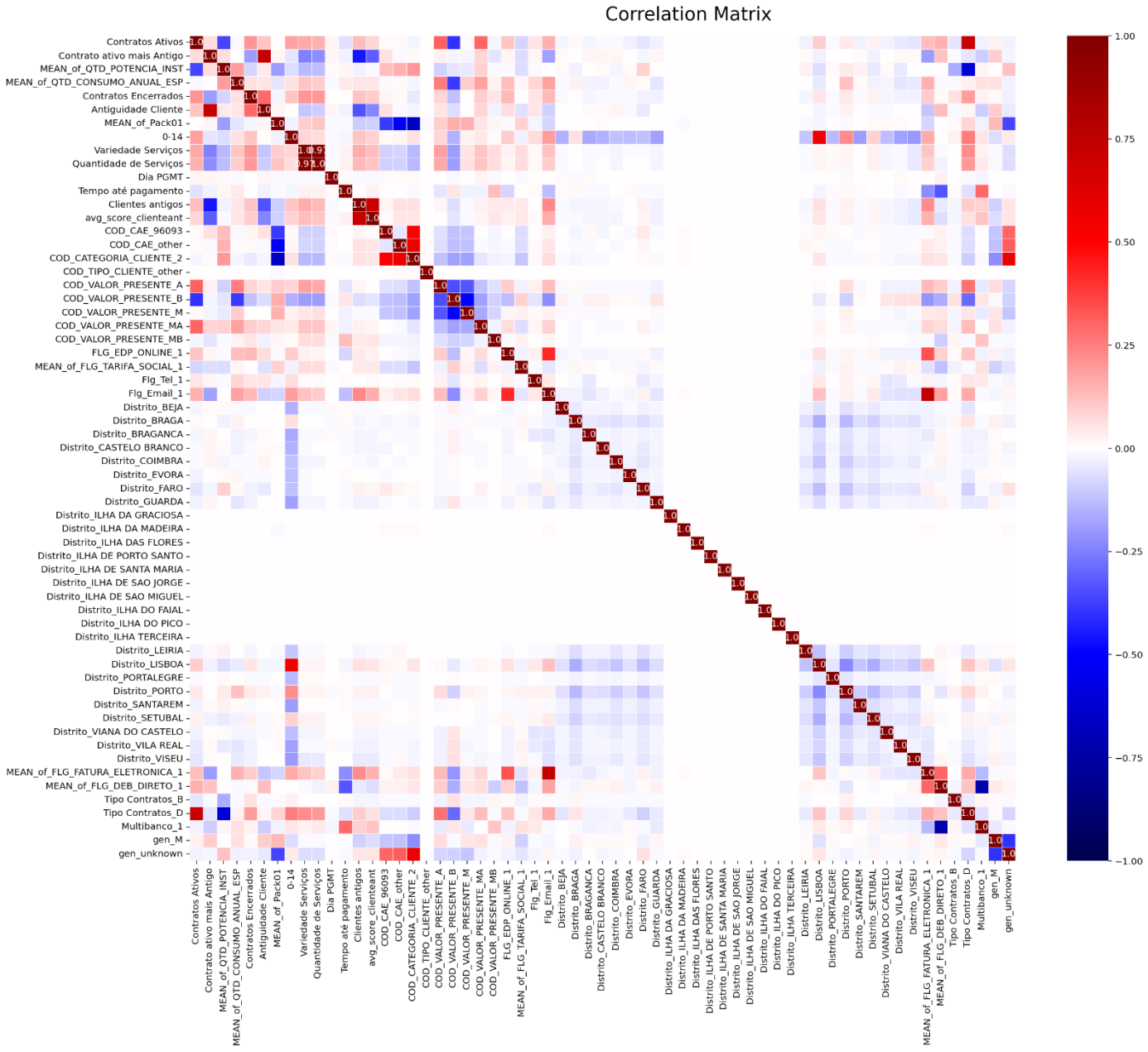

Figura 15 - Spearman's rank correlation matrix

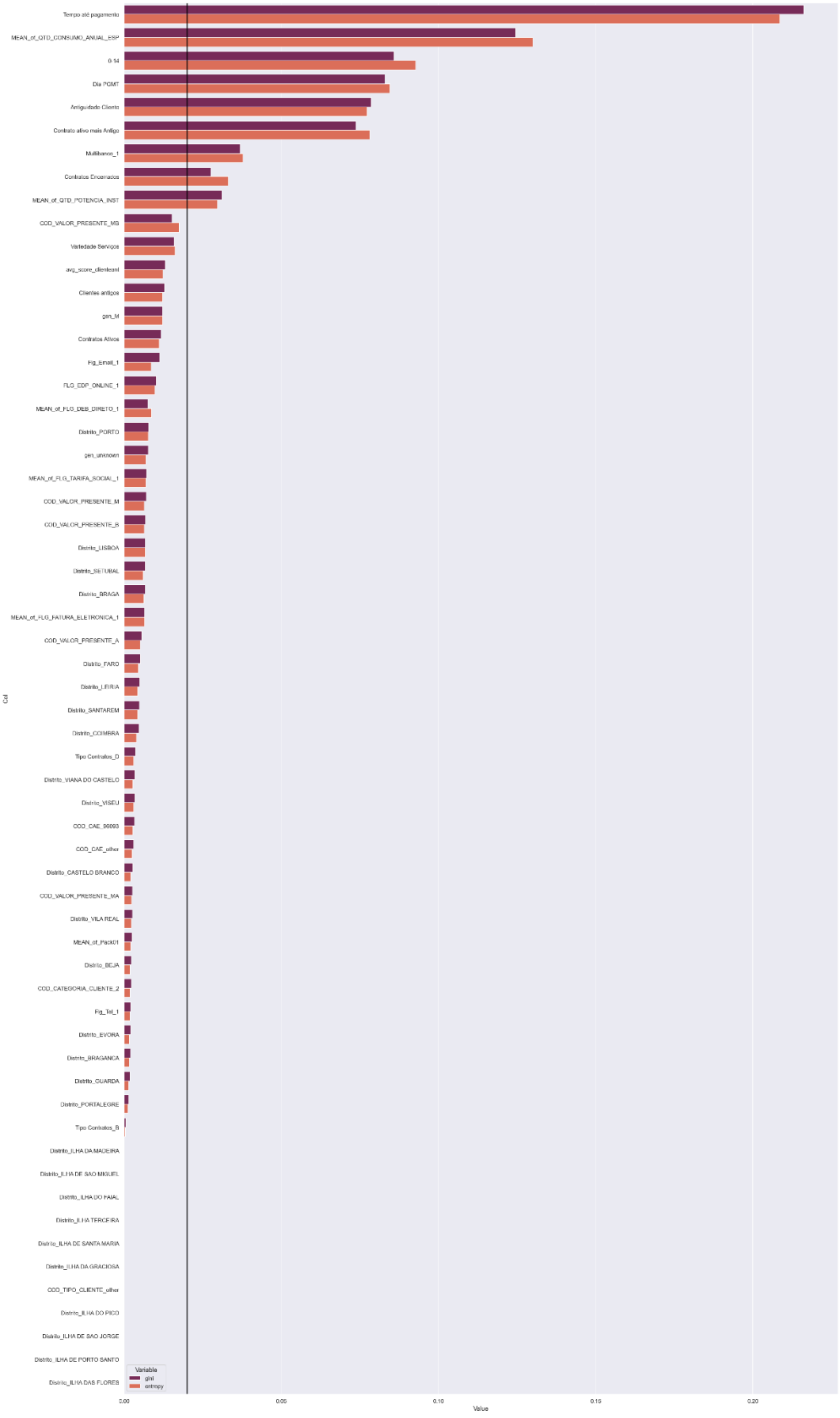

Figura 16 - Feature Selection Decision Trees: Gini and Entropy

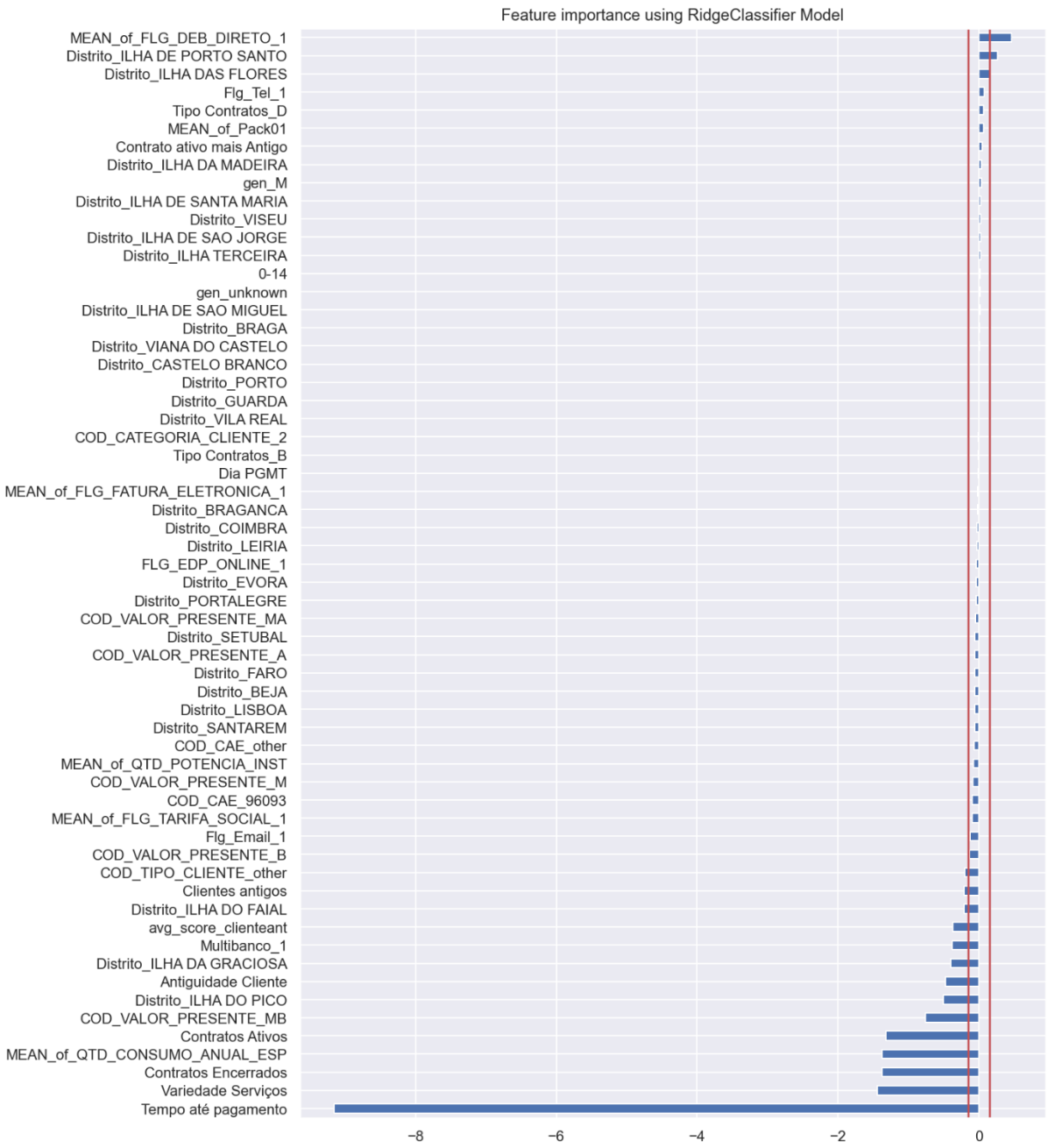

Figura 17- Feature Selection using Ridge Classifier

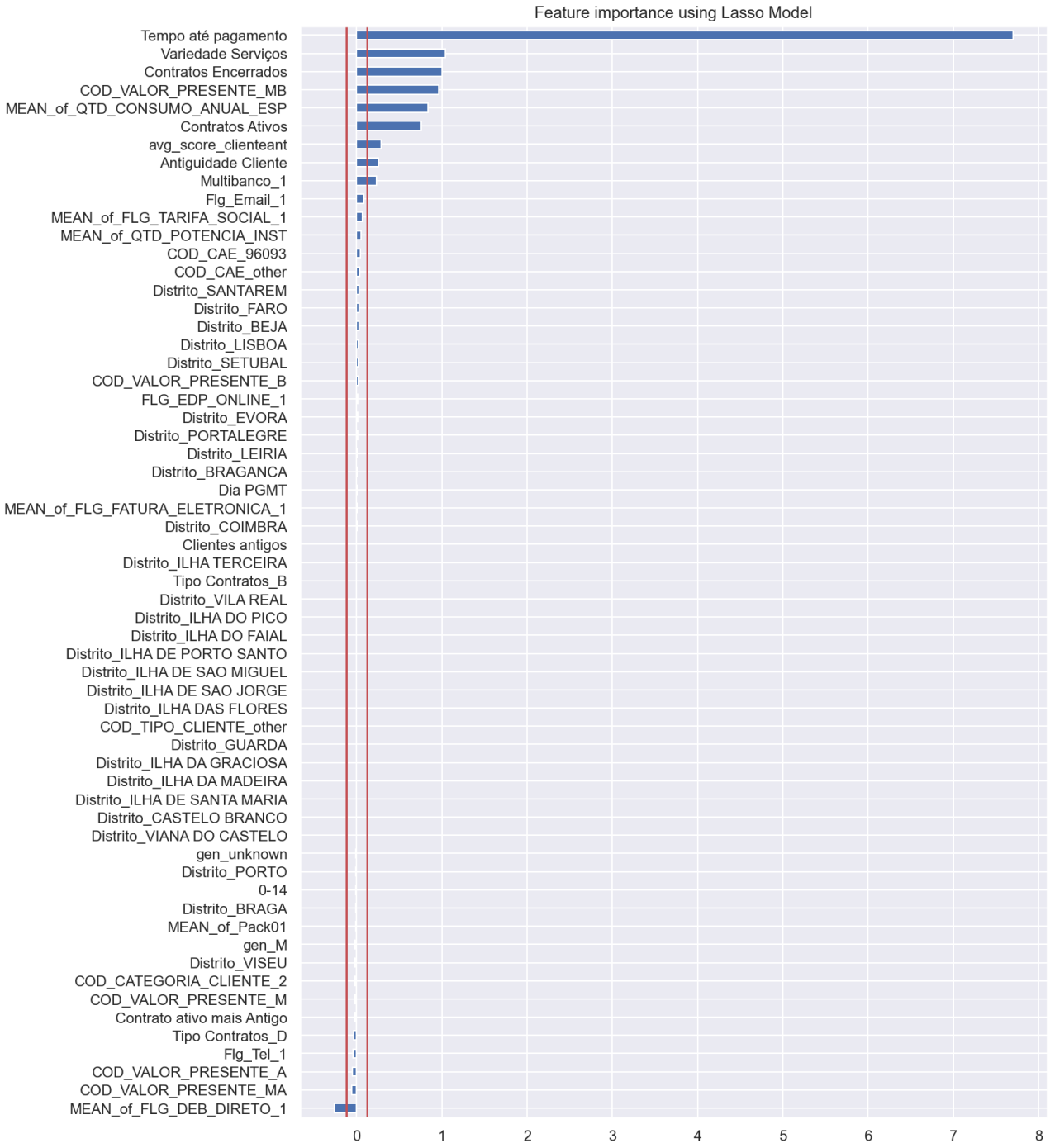

Figura 18- Feature Selection using Lasso Classifier

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

Universidade Nova de Lisboa