NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Predicting prepayment in home loans:

Modelling full and partial prepayment in the Portuguese banking sector using machine learning methods

Joana Maria Estalagem Lopes

Dissertation presented as a partial requirement for obtaining a Master's degree in Information Management

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

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

PREDICTING PREPAYMENT IN HOME LOANS:

MODELLING FULL AND PARTIAL PREPAYMENT IN THE PORTUGUESE BANKING SECTOR USING MACHINE LEARNING METHODS

by

Joana Maria Estalagem Lopes

Dissertation presented as a partial requirement for obtaining a Master's degree in Information Management, Specialization in Knowledge Management and Business Intelligence

Advisor: Prof. Dr. Jorge Miguel Ventura Bravo

November 2021

ACKNOWLEDGEMENTS

To Professor Jorge Miguel Ventura Bravo, for his support, shared knowledge, support in the dissertation's changes, and endless emails in this pandemic reality.

To the Bank that provided the data for this dissertation, in particular to the model development team for the support given, for the availability presented to have fortnightly meetings, unlimited calls, emails, and messages.

To my employer for the help in finding the necessary data and support provided throughout this challenging year, in particular to Alexandra Ferreira for her mentorship and guidance.

To my family and friends for the support, mentoring, concern, and emotional help they provided me throughout this dissertation, the masters and my life. To my parents, who always supported me, even when worried about my (lack of) sleeping hours. To my sister for the unconditional support and advice she has always given me. To Sara Maciel and Miriam Lopes for the friendly and understanding words for my absences. And last but not least, to my boyfriend, for his help, support, companionship, presence, and understanding for my absences and fatigue.

RESUMO

Existe um pré-pagamento quando ocorre um reembolso antecipado de um empréstimo por parte do tomador, i.e., o tomador paga mais que o montante contratual acordado. Tal pode ocorrer como parte do principal em dívida (reembolso parcial) ou o valor total do principal em dívida (reembolso total). Do ponto de vista de um banco, o estudo do reembolso antecipado - seja total ou parcial - é importante, pois resulta numa mudança nos fluxos de caixa calendarizados. Em particular, há uma diminuição nos fluxos de caixa futuros resultantes de um evento futuro desconhecido.

Assim, o principal objetivo deste estudo é a modelação dos eventos de pré-pagamento no crédito à habitação de um grande banco português, através de uma abordagem de *machine learning*, avaliando o seu desempenho através da utilização de técnicas como a Area Under the Receiver Operating Characteristic Curve (ROC), o *gain or lift* e Kolmogorov-Smirnov. Tal permite o estudo do fenómeno das amortizações antecipadas (ou pré-pagamentos) no mercado Português, utilizando dados reais, e através de modelos de *machine learning*.

Uma vez que foram utilizados dados reais, a primeira parte deste estudo prendeu-se com o pré-processamento dos dados, de modo a garantir que os modelos não incluíam ruído e problemas de qualidade de dados. A segunda parte prendeu-se com a computação dos modelos de *machine learning*, testando modelos de *artificial neural network* e *random forest*, com a comparação da *performance* destes através de métricas como o ROC, *gain or lift* e Kolmogorov-Smirnov.

Os resultados obtidos revelam que os modelos de pré-pagamento total e parcial apresentam bom desempenho nas três métricas de desempenho analisadas. Ambos os modelos apresentam resultados positivos e demonstram que os modelos apresentam bons resultados preditivos e capacidade discriminatória, sendo o modelo de amortização parcial superior ao modelo de amortização total, com uma diferença que, embora não muito grande, merece destaque.

Este estudo é particularmente relevante dada a sua análise num banco português, e a aplicação de modelos de *machine learning* na modelação de pré-pagamento, para os quais os estudos são escassos. Por outro lado, têm recentemente ocorrido esforços (por parte do banco onde o estudo se encontra incluído) para a atualização dos modelos tradicionais atualmente em vigor.

KEYWORDS

Amortizações antecipadas; Pré-pagamento; Crédito Habitação; Modelos de *machine learning*; Florestas Aleatórias; Redes Neuronais.

ABSTRACT

There is a loan prepayment when there is an early repayment of a loan from the borrower, i.e. the borrower pays more than the contractual amount due. The repayment may be part of the outstanding principal (partial repayment) or the total principal outstanding (full repayment). From a Bank's perspective, the study of early repayment – be it full or partial – is relevant as they result in a change in the schedule cash flows. In particular, there is a decrease in the future cash flows resulting from an unknown future event.

Hence, the primary purpose of this study is the modelling of the prepayment events in the mortgage loans of a large Portuguese bank, through a machine learning approach, assessing its performance through the use of techniques such as the Area Under the Receiver Operating Characteristic Curve (ROC), the Gain or Lift, and Kolmogorov-Smirnov statistic. This allows for the test of the prepayment phenomena in the Portuguese reality, using real Bank data, and through the use of machine learning models.

As there was a use of real-life data, the first part of this study implied the pre-processing of the data, to ensure that the noise and data quality problems were not part of the models. The second stage implied the computation of the machine learning models, which occurred through the testing of Artificial Neural Network and Random Forest models, with the comparison of its performance using the ROC, Gain or Lift and Kolmogorov-Smirnov statistic.

The results obtained reveal that both the total and partial prepayment models perform well in all the three performance metrics analysed. Both models present positive results and demonstrate that the models have good predictive results and discriminatory capacity. The partial repayment model is superior to the full repayment model, with a difference that is worthy of mention although not very large.

This study is particularly relevant given its analysis in a Portuguese bank and the application of machine learning models in modelling prepayment, for which studies are scarce. Furthermore, there have been occurring efforts (in the bank where this study is framed) to update the traditional models currently in force.

KEYWORDS

Pre-payment; Early repayment; Mortgage Loans; Machine Learning; Random Forest; Neural Network.

INDEX

1. Introduction 1
    1.1. Problem Identification.. 4
2. Literature Review 6
        2.1.1. Prepayment in Loans 6
        2.1.2. Prepayment in the Portuguese market 8
        2.1.3. Machine learning models 9
    2.2. Information Considered 14
    2.3. Performance assessment 14
3. Methodology 16
    3.1. Data Pre-Processing 20
        3.1.1. Data Cleaning 21
        3.1.2. Data Transformation 26
        3.1.3. Data Reduction... 30
    3.2. Development of Models.. 32
    3.3. Performance Assessment 33
        3.3.1. Area Under the Curve 33
        3.3.2. Gain or Lift 34
        3.3.3. Kolmogorov-Smirnov 35
4. Results and Discussion 36
    4.1. History of Prepayments: Comparison between the bank and the financial system 36
    4.2. Modelling Prepayment... 37
        4.2.1. Full Prepayment 38
        4.2.2. Partial Prepayment.. 40
    4.3. Further Discussion.. 42
        4.3.1. Comparison between full and partial prepayment.................... 42
        4.3.2. Comparison between model and profiling 46
5. Conclusions... 48
    5.1. Limitations and Recommendations for Future Work 49
6. Bibliography... 50
7. Appendix... 58
    Appendix 1. Methodological Steps and Software 59
    Appendix 2. Dataset variables 60
    Appendix 3. Data Description – Histogram, Bar Chart and Box Plot.................. 65

Appendix 4. Data Pre-Processing – Post-Windsorizing and Impute node 95
Appendix 5. Generalization – Purpose of Loan 103
Appendix 6. Generalization – Marital Status 106
Appendix 7. Generalization – Profession 107
Appendix 8. Pearson Correlation 133
Appendix 9. Stepwise Regression Variable Selection – Full repayment............. 134
Appendix 10. Stepwise Regression Variable Selection – Partial repayment........ 136
Appendix 11. LASSO Regression Variable Selection – Full repayment 138
Appendix 12. LASSO Regression Variable Selection – Partial repayment............ 140

FIGURE INDEX

Figure 1 – Evolution of households' patrimony. Source: BPStat (Banco de Portugal, 2021d, 2021e) 2

Figure 2 – Evolution of credit to individuals in the national market. Source: BPStat (Banco de Portugal, 2021b, 2021a, 2021c) 3

Figure 3 – Evolution of RWA, by risk type. Source: EBA reporting (European Banking Authority, 2019) 4

Figure 4 – Amount of prepayment for total and partial repayments (bar chart) and number of total and partial repayments (line chart). Source: Report on Monitoring of Retail Banking Markets from Banco de Portugal, data aggregated by the author (Banco de Portugal, 2019) 9

Figure 5 – Generic schema of an MLP network (Desai et al., 1996) 11

Figure 6 – Schema of a decision tree (Anderson, 2007)....13

Figure 7 – Representation of the methodology followed, adapted by the author from the literature. Source: Handhika et al., 2019; Lessmann et al., 2015; Munkhdalai et al., 2019; Xia et al., 2017. 16

Figure 8 – Reduction of records through data reduction. 22

Figure 9 – Generic schema of an MLP network (Desai et al., 1996) 33

Figure 10 – Depiction of ROC curve and respective AUC (Narkhede, 2018) 34

Figure 11 – Cumulative Lift chart example. Source: (SAS, 2021e) 35

Figure 12 – Amount of prepayment for total and partial prepayment (bar chart) and number of total and partial prepayment (line chart), comparison between the financial system and Bank. Source: Report on Monitoring of Retail Banking Markets from Banco de Portugal, data aggregated by the author, and Bank's internal data (Banco de Portugal, 2019). 36

Figure 13 – ROC chart for the full prepayment models 39

Figure 14 – Cumulative lift chart for the full prepayment models 39

Figure 15 – ROC chart for the partial prepayment models 41

Figure 16 – Cumulative lift chart for the partial prepayment models 42

Figure 17 – Methodological steps and software of the steps performed 59

Figure 18 – Histogram and box plot of the year of construction 65

Figure 19 – Histogram and box plot of the municipality 66

Figure 20 – Histogram and box plot of the opening date of the loan 67

Figure 21 – Histogram and box plot of the district 68

Figure 22 – Histogram and box plot of the loan purpose 68

Figure 23 – Histogram and box plot of the LTV of the current property evaluation 69
Figure 24 – Histogram and box plot of the LTV of the original property evaluation 70
Figure 25 – Histogram and box plot of the early repaid amount 71
Figure 26 – Histogram and box plot of the financed amount 71
Figure 27 – Histogram and box plot of the residual amount..72
Figure 28 – Histogram and box plot of the number of instalments paid 73
Figure 29 – Histogram and box plot of the loan term 74
Figure 30 – Histogram and box plot of the residual term 74
Figure 31 – Histogram and box plot of the total partial early repayments............................75
Figure 32 – Histogram and box plot of the total amount repaid 76
Figure 33 – Histogram and box plot of the interest rate 76
Figure 34 – Histogram and box plot of the spread rate 77
Figure 35 – Histogram and box plot of the contract end date 77
Figure 36 – Histogram and box plot of the date of birth...78
Figure 37 – Histogram and box plot of the marital status...79
Figure 38 – Histogram and box plot of the level of education 80
Figure 39 – Histogram and box plot of the profession..81
Figure 40 – Histogram and box plot of the yearly income 81
Figure 41 – Histogram and box plot of the scoring 82
Figure 42 – Histogram and box plot of the payment incident indicator 83
Figure 43 – Histogram and box plot of the check inhibition indicator..................................84
Figure 44 – Histogram and box plot of the number of days past due 84
Figure 45 – Histogram and box plot of the monthly instalment in the financial system.........85
Figure 46 – Histogram and box plot of the monthly instalment in the bank 85
Figure 47 – Histogram and box plot of the number of products in the financial system 86
Figure 48 – Histogram and box plot of the number of products in the bank.........................87
Figure 49 – Histogram and box plot of the percentage of credit card usage 88
Figure 50 – Histogram and box plot of the balance in sight deposits, 6 months 88
Figure 51 – Histogram and box plot of the balance in sight deposits, 12 months 89
Figure 52 – Histogram and box plot of the balance in term deposits, 6 months 89
Figure 53 – Histogram and box plot of the balance in term deposits, 12 months 90
Figure 54 – Histogram and box plot of the debtors in the national financial system 90
Figure 55 – Histogram and box plot of the real operations in the national financial system ..91
Figure 56 – Histogram and box plot of the real operations in the bank 91
Figure 57 – Histogram and box plot of the potential operations in the national financial system 92

Figure 58 – Histogram and box plot of the potential operations in the bank 93
Figure 59 – Histogram and box plot of the amount of potential credit in the national financial system 93
Figure 60 – Histogram and box plot of the amount of real credit in the national financial system 93
Figure 61 – Histogram and box plot of the amount of potential credit in the bank 94
Figure 62 – Histogram and box plot of the amount of real credit in the bank........................94
Figure 63 – Comparison of the loan purpose before and after conversion 97
Figure 64 – Comparison of the financed amount before and after impute and outlier smoothing...97
Figure 65 – Comparison of the current LTV before and after impute and outlier smoothing .98
Figure 66 – Comparison of the origination LTV before and after impute and outlier smoothing 98
Figure 67 – Comparison of the number of paid instalments before and after impute 98
Figure 68 – Comparison of the interest rate before and after impute and outlier smoothing99
Figure 69 – Comparison of the spread rate before and after impute and outlier smoothing .99
Figure 70 – Comparison of the marital status before and after conversion 99
Figure 71 – Comparison of the age before and after impute 100
Figure 72 – Comparison of the profession before and after conversion 100
Figure 73 – Comparison of the yearly income before and after impute and outlier smoothing 100
Figure 74 – Comparison of the scoring before and after impute and outlier smoothing......101
Figure 75 – Comparison of the check inhibition before and after impute and outlier smoothing 101
Figure 76 – Comparison of the monthly instalments in the financial system before and after impute and outlier smoothing 102
Figure 77 – Pearson correlation matrix 133

TABLE INDEX

Table 1 – Macroeconomic variables added, and the respective source 18
Table 2 – Variables added to the dataset 19
Table 3 – Example of the calculation performed 19
Table 4 – Descriptive statistics for data assessment (Vidal & Barbon, 2019) 20
Table 5 – Strategies for working with missing values. Source: (Vidal & Barbon, 2019) 21
Table 6 – Statistical descriptions of numerical variables 24
Table 7 – Statistical descriptions of the categorical variables 25
Table 8 – Aggregation categories in the loan purposes. Source: Author aggregation 28
Table 9 – Loan purposes which were eliminated..29
Table 10 – Aggregation categories in the marital status. Source: Bank's internal aggregation 29
Table 11 – Aggregation of categories in the loan professions. Source: author aggregation based on *Classificação Portuguesa das Profissões – Grande Grupo* by INE (Instituto Nacional de Estatística, 2011)..30
Table 12 – Filter methods according to variables' types (Brownlee, 2019; Kaushik, 2016) 30
Table 13 – AUC thresholds...34
Table 14 – Percentage of the targets in the dataset..37
Table 15 - Results of the performance assessment metrics for the full prepayment, highlighted the best model for the metric 38
Table 16 - Results of the performance assessment metrics for the partial prepayment, highlighted the best model for the metric...41
Table 17 – Variables considered in the full prepayment model 44
Table 18 – Variables considered in the partial prepayment model.......................................46
Table 19 – Analysis of variables category between the full and partial models 46
Table 20 - Analysis of variables category between the full and partial models 47
Table 21 – Variables considered in the dataset 64
Table 22 - Statistical descriptions of numerical variables 96
Table 23 - Statistical descriptions of categorical variables..97
Table 24 – Original categories and mapping to aggregated categories in loan purpose.......105
Table 25 - Original categories and mapping to aggregated categories in marital status 106
Table 26 – Original categories and mapping to aggregated categories in the profession variable 132
Table 27 – Variables selected using the Stepwise Regression in the full repayment............135
Table 28 – Variables selected using the Stepwise Regression in the partial repayment.......137

Table 29 – Variables selected using the LASSO Regression in the full repayment 139
Table 30 – Variables selected using the LASSO Regression in the partial repayment 141

LIST OF ACRONYMS AND ABBREVIATIONS

ANN	Artificial Neural Networks
AUC	Area Under the ROC Curve
BME	Bayesian Model Ensemble
BS	Brier Score
CC	Consistency Check
CPR	Conditional Prepayment Rate
DSTI	Debt service to income
ECB	European Central Bank
ECL	Expected Credit Loss
EU	European Union
FPR	False Positive Rate
FS	Financial System
GIGO	Garbage-in, garbage-out
IV	Information Value
KS	Kolmogorov-Smirnov
LTV	Loan-to-value
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLP	Multilayer perceptron
RF	Random Forest
ROC	Receiver Operating Characteristic
RWA	Risk weighted assets
TPR	True Positive Rate
US	United States of America
USA	United States of America

1. INTRODUCTION

Housing wealth is the most important asset Portuguese families hold in their portfolios, representing around 60% of the individuals' net worth in 2020 (Figure 1). Although there is a decrease in the housing weight in the total financial assets, there is an overall upward trend, decreasing in crisis years, namely in the 2010s financial crisis. The primary driver of families' net worth is housing, with a dramatic change in the last 50 years – with an increase in homeownership, as opposed to a majority of tenants; this can be seen in the chart with the overall increase in the net assets (Banco de Portugal, 2021d, 2021e; Xerez, Pereira, & Cardoso, 2019; Xerez, Rodrigues, Lima, & Cardoso, 2019).

According to the analysis of the censuses from 1970 and 2011, the percentage of tenants and homeowners changed significantly. While in 1970 the ratio was balanced, in 2011, 73% of families were homeowners and 27% were tenants. This change is explained by the public policies to encourage the acquisition or construction of own housing, the development of a subsidized credit regime and housing savings accounts. However, a shift in this trend with the 2010's sovereign debt crisis, with a contraction in home loans and an increase in the rental market, must be noted. (Xerez, Pereira, et al., 2019; Xerez, Rodrigues, et al., 2019)

Furthermore, this tendency is enhanced by article 65 of the Portuguese Constitution, which states the right to housing and, in particular, homeownership. In 1976, the resolution of the Council of Ministers defined the access to the purchase of own homes by the families as an elementary principle of the housing policy. These policies and Portugal's entry into the European Union (EU) led to the increase in homeownership by Portuguese families (*Constituição Da República Portuguesa*, 1976; *Resolução Do Conselho de Ministros, Nº67, 1ª Série*, 1976; Xerez, Pereira, et al., 2019; Xerez, Rodrigues, et al., 2019).

In addition to the legal and financing components, cultural aspects must also be considered, particularly in southern European countries. Here, homeownership is associated with wealth, freedom, housing satisfaction, and a safety net in older years in the face of reduced income in retirement (Elsinga & Hoekstra, 2005; Xerez, Pereira, et al., 2019).

This trend may also be observed in Figure 1, which demonstrates an increase in Portuguese families' patrimony. The patrimony associated with housing represents the majority, which demonstrates the study findings and the tendency for homeownership in the Portuguese market and, subsequently, for mortgage credits to aid this homeownership (Banco de Portugal, 2021d, 2021e).

Building up housing wealth through homeownership and mortgage repayment is also by far the main way European households set aside for old age (Household Finance and Consumption Survey, ECB 2016). In the Euro area countries, the household's wealth (excluding pension wealth, the present value of all future expected pension benefits) is primarily held in the form of real assets, which represent 82.2% of total assets owned by households with the remaining assets being financial. The largest component of real assets is the household main residence, representing 60.2% of total real assets, followed by other real estate property (Bravo et al., 2019).

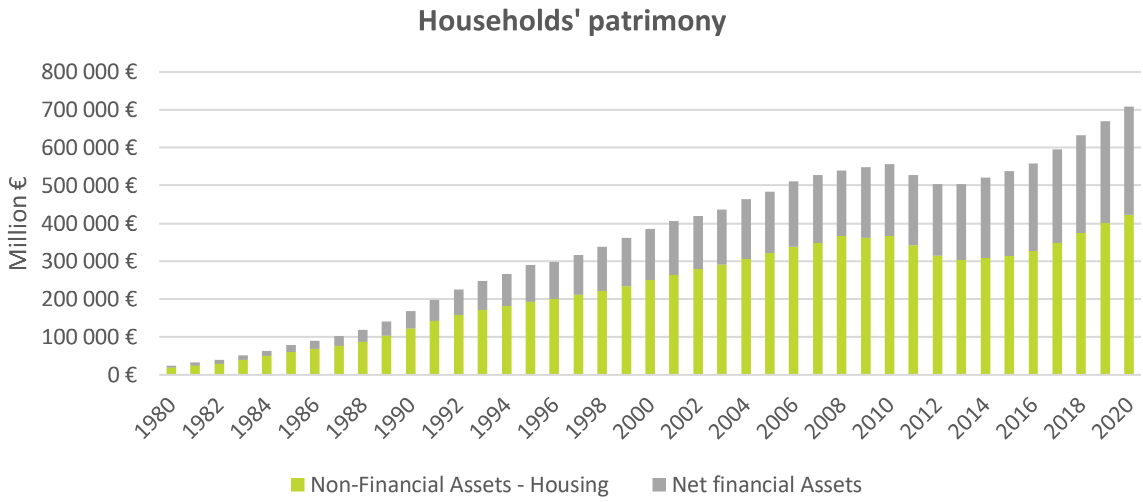

Figure 1 – Evolution of households' patrimony[1]. Source: BPStat (Banco de Portugal, 2021d, 2021e)

This increase in homeownership is accompanied by the increase in mortgage loans, shown in Figure 2, representing 77% of the total loans to individuals in June 2021. As with the individual's net worth, in years of crisis, such as at the beginning of the decade, the credit granted tends to decrease, with mortgage loans decreasing sharply in this period (Banco de Portugal, 2021b, 2021a, 2021c).

Thus, given the representativeness of mortgage loans for individuals, there are incentives to apply strategies to reduce the monthly installments. One of these is the early repayment, or prepayment, which occurs when there is an early repayment of a loan from the borrower, i.e., the borrower pays more than the contractual amount due. This may be part of the outstanding principal (partial repayment) or the total principal outstanding (full repayment) (Banco de Portugal, 2021a; Jacobs et al., 2005; LaCour-Little, 2008).

[1] Here, net financial assets are given by the total financial assets, net of liabilities.

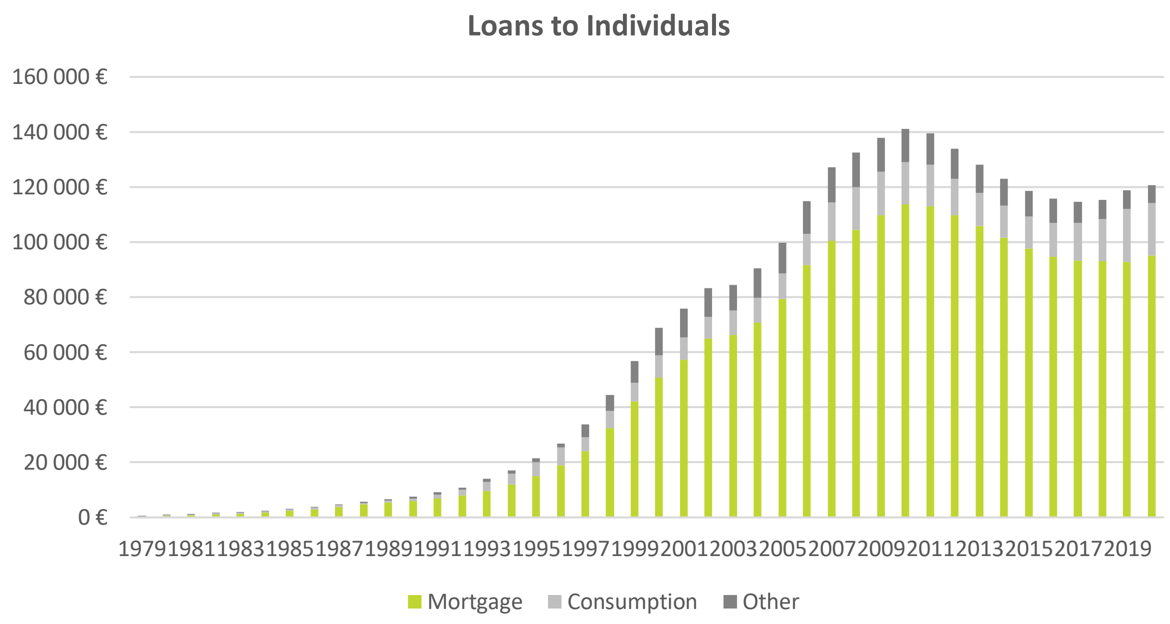

Figure 2 – Evolution of credit to individuals in the national market. Source: BPStat (Banco de Portugal, 2021b, 2021a, 2021c)

On the financial institutions' side, in the last years and in the wake of the pandemic, there have been low and negative interest rates, which have been decreasing banks' profitability, eroding banks' net interest margins. Net interest margins are mainly comprised of structural elements (such as high-quality liquid assets, demanded to fulfil regulatory requirements such as the liquidity coverage ratio) and the margin on assets and liabilities (more linked to client business, such as the ability to reprice deposits, and excess liquidity from deposits). This has been leading to an increase in the risk-taking by banks as a strategy to counter this trend, namely through expanding mortgage lending and consumer credit at weaker terms. In particular, the early repayment, or repayment, presents a risk for the banks as it reduces the future cash-flows and, consequentially, its liquidity (Albertazzi et al., 2020; Bohn et al., 2020).

The European Central Bank (ECB) states that their economic, political, and debt sustainability are the main risk drives for European banks. Credit risk is encompassed in these three main risks, which is, in a simplified way, the potential of a customer failing to fulfil its contractual obligations with the financial institution – generally speaking, failure to meet the contractual payments. Therefore, one of the focuses of financial institutions relates to credit risk management, whose main objective is to maximize the rate of return, adjusted to the institution's risk, i.e. maintaining risk exposure within acceptable parameters as per the institution's risk policy and the regulatory constraints. In addition, liquidity risk is closely monitored by the regulators and is defined by measuring the risk of bank's incurring in losses from the inability to fulfil its payment obligations (Basel Committee on Banking Supervision, 2000; European Central Bank, 2020).

As can be seen in Figure 3, credit risk encompasses the majority of the European banks' risk-weighted assets (RWA). Additionally, home mortgages represent a significant percentage of banks' outstanding debt and the large majority of the loans to individuals, as shown by Figure 2. (Banco de Portugal, 2021b, 2021a, 2021c; Li, 2014)

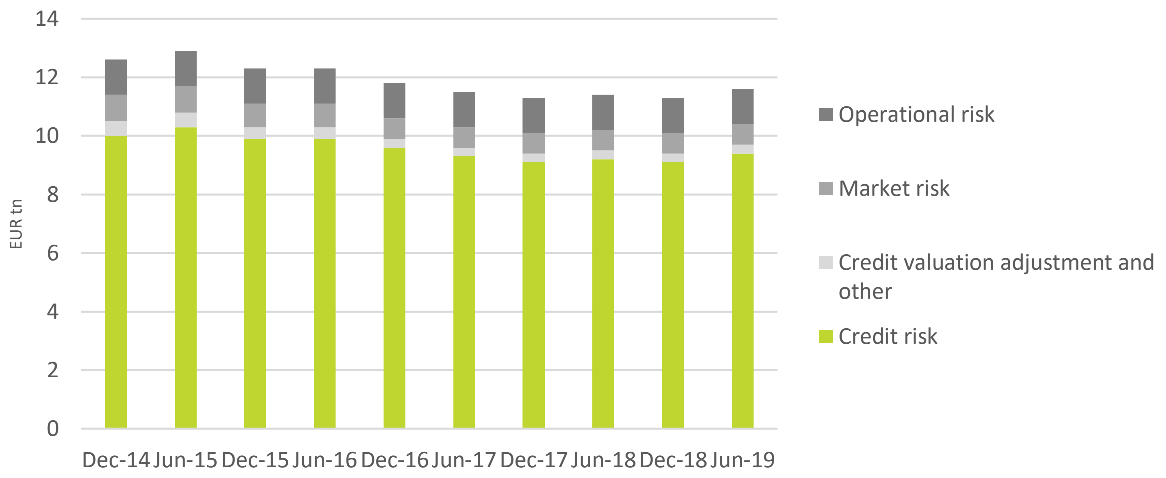

Figure 3 – Evolution of RWA, by risk type. Source: EBA reporting (European Banking Authority, 2019)

1.1. Problem Identification

Prepayments, the payment of an amount that exceeds the scheduled amortization of the loan, have usually been predicted through the use of option-based pricing theory and time-series analysis, being the majority of the studies found. Studies using machine learning approaches are found in more recent papers, though they are scarce. Furthermore, most of the studies focus on the American mortgage market, whose reality is significantly different from the Portuguese.

In the United States of America (US or USA), mortgages are highly sensitive to long-term interest rate changes, being prepayment costs priced into the interest rate. Differently, in Portugal there are lump-sum prepayment penalties, induced by statutory requirements, and charged to the clients for early repayment. This results in a significant difference between the two markets, as in Portugal homeowners bear a part of the prepayment risk, with banks holding residual risk[2] (Deloitte, 2019; European Central Bank, 2004; Fang & Munneke, 2021; Sirignano, Sadhwani, & Giesecke, 2018).

As such, with this study, a different set of mortgage markets will be analyzed compared to most studies – a sizeable Portuguese banking institution – whose reality, as mentioned, differs from the majority of the papers. Furthermore, it will model these events through machine learning models, exploring the works of the most recent studies, particularly Random Forest (RF) and Artificial Neural Network (ANN) models. The models to be estimated will use historical information on credit behaviour to determine which characteristics will allow for the prediction of the customer behaviour during the contract's lifetime, that is, whether they will tend to comply with the terms of the contract or pay earlier than scheduled. (Mester, 1997)

Hence, the primary purpose of this paper is to model prepayment in a large Portuguese bank, having as specific objectives the (i) implementation of machine learning models for these events and assess its classification performance through the use of techniques such as the Area Under the Receiver

[2] Since the funding of sources are tipically by retail deposits and other retail instruments. However, there is a tendency to increase the share of market funding, through mortgage covered bonds and mortgage-backed securities.

Operating Characteristic Curve (AUC), the Gain or Lift, and Kolmogorov-Smirnov statistic (Lessmann et al., 2015), (ii) the use of data from a bank on the Portuguese market, where there is not a significant number of studies, (iii) the comparison of the variables selected in the models when analyzed in a profiling and estimation perspective[3], and (iv) the presentation of the results to the model development team in the bank and compare the performance between the models currently used and the resulting models and variables used.

The study is particularly relevant given both its analysis in Portuguese banking and the application of machine learning models in modelling prepayment, for which studies are scarce internationally. Furthermore, there have recently been internal projects in the bank to update the traditional models, currently in force, where this study was framed.

Throughout the development of this study, there were fortnightly meetings with the model development team in the bank, currently testing Decision Tree (DT) models, with a discussion of the intermediate results, data sources, and transformations. The study's final output has the capacity of being leveraged for variable selection and testing of additional models and data transformations, as decision trees provide a degree of transparency and suitability superior to the random forest and artificial neural network models.

[3] This approach will compare the variables selected by the models before and after the prepayment, i.e. when the model is a predictive model versus when it is a profiling model.

2. LITERATURE REVIEW

2.1.1. Prepayment in Loans

A mortgage is prepaid when there is an early repayment, or prepayment, of a loan from the borrower, i.e. the borrower pays more than the contractual amount due. This may be part of the outstanding principal (partial repayment) or the total principal outstanding (full repayment). The amount of partial repayment is defined by the client and results in a reduction of the value of monthly installments, whereas a reduction in the contract term results in a change of the contract (Banco de Portugal, 2021a; Jacobs et al., 2005; LaCour-Little, 2008).

From a Bank's perspective, the study of early repayment – be it full or partial – is important as they result in a change in the schedule cash flows. In particular, there is a decrease in the future cash flows resulting from an unknown future event. Meis (2015) refers that it impacts financial institutions in two perspectives: (i) in the asset and liability management of the bank and (ii) because prepayments lead to interest rate risk[4] (Jacobs et al., 2005; Kishimoto & Kim, 2014; Meis, 2015).

Various studies have tried to understand the clients – mortgagors – motivations for early repayment, i.e. the deviation from the contractual installments. This deviation from the agreed repayment schedule may be due to changes in economic conditions (usually an improvement in these conditions, which encourage borrowers to reduce the amount due in the loan), or due to a transfer of the loan to another credit institution (with Banco de Portugal stating that the latter is the major cause for full early repayment in Portugal and caused by, for example, divorce or change of job location).

The literature suggests two approaches to predict when customers will repay: (i) an optimal prepayment approach and (ii) exogenous prepayment rules. The first assumes that the option to prepay is exercised when the mortgage value exceeds the outstanding debt plus transaction costs if any (ignoring individual factors of the borrowers). Taking mortgage default as a (put) option, early literature used the Black and Scholes (1973) pioneered contingent claims framework. Using this approach, the key drivers of default were home values and interest rates (Gerardi et al., 2013; Chamboko & Bravo, 2020). The latter relates the observed prepayment (and default) with a set of explanatory variables (such as scoring, income, loan-to-value ratio, loan age, interest rates, housing prices, and refinancing incentives.

Studies identify refinancing incentives as the most crucial factor influencing prepayment, i.e. prepayment event when the mortgage rate is below the contractual rate. These exogenous models overcome the shortcomings of the optimal model, which cannot explain behaviour and borrowers not prepaying when it is optimal, as consumers have other non-financial motivations to prepay (Banco de Portugal, 2021a; Charlier & van Bussel, 2001; Goodarzi et al., 1998; Jacobs et al., 2005; LaCour-Little, 2008; Meis, 2015; Saito, 2018; Sirignano et al., 2018).

This study considers prepayment through the exogenous approach, which does not assume that the borrowers will always follow rational reasoning. The primary determinants considered in these types

[4] Though, as it will be shown ahead, the prepayment risk is not as relevant in the Portuguese banking market.

of models are (Borovkova, 2017; Charlier & van Bussel, 2001; Dickinson & Heuson, 1994; Jacobs et al., 2005):

i. **Refinancing incentive** – which has been considered the most important element, is the price mechanism, such as a drop in the market rates. Variables considered include tax considerations, burnout effect, and the media effect;

ii. **Housing turnover and seasoning** – which depends on mobility rate of the labour force, seasonal factors (such as the month of the year, where prepayment tends to be higher in December and lower in January and February), and seasoning curve (i.e. the relation between the loan age and the prepayment rate, which yields an S-shaped curve);

iii. **Macroeconomic factors** – which relates to the housing market and demographic factors;

iv. **Microeconomic or loan-specific factors** – which relate to factors such as age and type of mortgage, type of house, loan-to-value (LTV) ratio, socio-economic status, and marital status.

These exogenous models, where it is considered that clients do not always make the optimal financial decisions, use models such as Bayesian models, Proportional Hazard Model, and Logistic Regression, most frequently the latter two. In the Proportional Hazard model, the hazard rate is the probability of the event – in this case, prepayment – occurring in the next month, given the mortgage has not been prepaid before. Here, the baseline hazard is the "typical" prepayment profile. Despite its advantages (such as the estimation of results even in incomplete datasets, its interpretability, flexibility to include factors in the model, and the production of superior results in comparison to Logistic Regression models), it is not widely used by financial institutions due to its complexity and lack of experience from finance professionals (as this model derives from the medical sciences).

In the Logistic Regression model, the dependent variable is the prepayment event, and the independent variables are given by macroeconomic factors and loan-specific variables. Financial institutions widely use this model (e.g. in credit scoring), and it is, as such, also a popular model for prepayment. However, Borovkova states that it lacks the flexibility and interpretability of the Proportional Hazard Models and produces inferior results (Borovkova, 2017; Charlier & van Bussel, 2001; Goodarzi et al., 1998; Green & Shoven, 1986; Kau et al., 1990; LaCour-Little, 2008; Schwartz & Torous, 1992; Chamboko & Bravo, 2016, 2019a,b).

More recent and scarce studies have modeled prepayment (and default probability) using machine learning and deep learning models. This approach overcomes the limitation of using a pre-specified form, usually a linear one (namely in variable interactions which are a significant component of the nonlinear effects), and are being developed here to support traditional models currently in force on financial institutions, as they allow for identification of complex relations between input and target variables. Two types of machine learning models have been used to model prepayment – Neural Networks and Random Forests (Borovkova, 2017; Deloitte, 2019; Sirignano et al., 2018). Classifiers using statistical or operational research methods and other machine learning methods such as genetic algorithms, homogenous ensembles, and heterogeneous ensembles have also been investigated (Ashofteh & Bravo, 2019, 2021a,b)

The studies of prepayment risk and modelling arise mainly from the USA, considering the specificities of the mortgage market in this country – with the majority of the loans from government-sponsored enterprises (from the Federal Home Loan Mortgage Corp.: Fannie-Mae and Freddie-Mae, and Government National Mortgage Association: Ginnie Mae). While these do not originate loans, they originate a secondary market for home loans. These studies, arising from the USA, usually model prepayment through option-pricing methodologies for mortgages subject to prepayment risk (Emmons, 2008; European Central Bank, 2004).

Hence it should be noted the differences between the majority of Portuguese mortgages – the origin of the dataset – and the US mortgages, the focus of most studies on prepayment risk. The handling of prepayment in US mortgages is, as mentioned before, handled by US mortgages agencies (Fannie Mae and Freddie Mac). Here, the prepayment risk is hedged in fixed income and swap markets, and thus an unexpected change in the prepayments can generate volatility in bond markets. Hence, prepayment activity is highly sensitive to long-term interest rate changes, as such prepayment costs are priced into the interest rate.

This is not the case for the Portuguese mortgage loans, where statutory requirements induce lump-sum prepayment penalties, which are charged to the clients for early repayment. This results in significant differences between the two markets, as in Portugal, the homeowners bear a part of the prepayment risk, with banks holding the residual risk[5]. This is the case for the majority of the European market, except for the Danish mortgage market, with long-term fixed-rate mortgage loans with embedded options of a penalty-free prepayment, such as the US (European Central Bank, 2004). Given these differences, the refinancing incentives are not as significant in this study.

2.1.2. Prepayment in the Portuguese market

As previously mentioned, it is important to note the difference between prepayment in Portuguese mortgage loans and US mortgages. In the studies made on the US mortgage systems, the prepayment activity is highly sensitive to long-term interest rate changes, as such prepayment costs are priced into the interest rate. Whereas in Portugal there are lump-sum prepayment penalties induced by statutory requirements and charged to the clients for early repayment. This results in a significant difference between the two markets, as in Portugal the homeowners bear a part of the prepayment risk, with banks holding the residual risk. This is the case for the majority of the European market, except for the Danish mortgage market, with long-term fixed-rate mortgage loans with embedded options of a penalty-free prepayment, such as the US (European Central Bank, 2004).

Hence, in the euro area, housing loans remain to a large extent on banks' balance sheets as they are mainly financed via bank deposits or, to some extent, via the issuance of covered bonds, i.e., banks tend to support a more cautious behaviour of lenders concerning the loans originated, and there is not a significant presence in secondary markets, given the supervisory goal to keep the financial institution's risk and balance sheet transparent (European Central Bank, 2009).

The similarities between most of the previous studies and the model to be defined relate to the management difficulties, as clients tend not to follow rational option exercising strategies. However, the risk is less concentrated than in the US, given that widespread mortgage prepayment penalty fees

[5] Since the funding of sources are typically by retail deposits and other retail instruments. However, there is a tendency to increase the share of market funding, through mortgage covered bonds and mortgage-backed securities.

apply, and retail deposits provide the funding of mortgages in Europe. Hence, European mortgage prepayment risk is mainly faced by both the originating bank and homeowners (European Central Bank, 2004).

In the particular case of the Portuguese market, there is a penalty for the mortgagor (i.e. the client) for exercising the prepayment, in particular, the following maximums are defined for commissions to be charged (Banco de Portugal, 2021a; Goodarzi et al., 1998; Mercer Oliver Wyman, 2003):

- Contracts with variable interest rate: equivalent to 0.5% of the repaid capital;
- Fixed interest rate contracts: equivalent to 2% of the repaid capital.

2.1.2.1. Historical prepayments in Portugal

Banco de Portugal publishes a yearly "Report on Monitoring of Retail Banking Markets" where, among others, shares the data on the amount and number of prepayments in Portuguese banks.

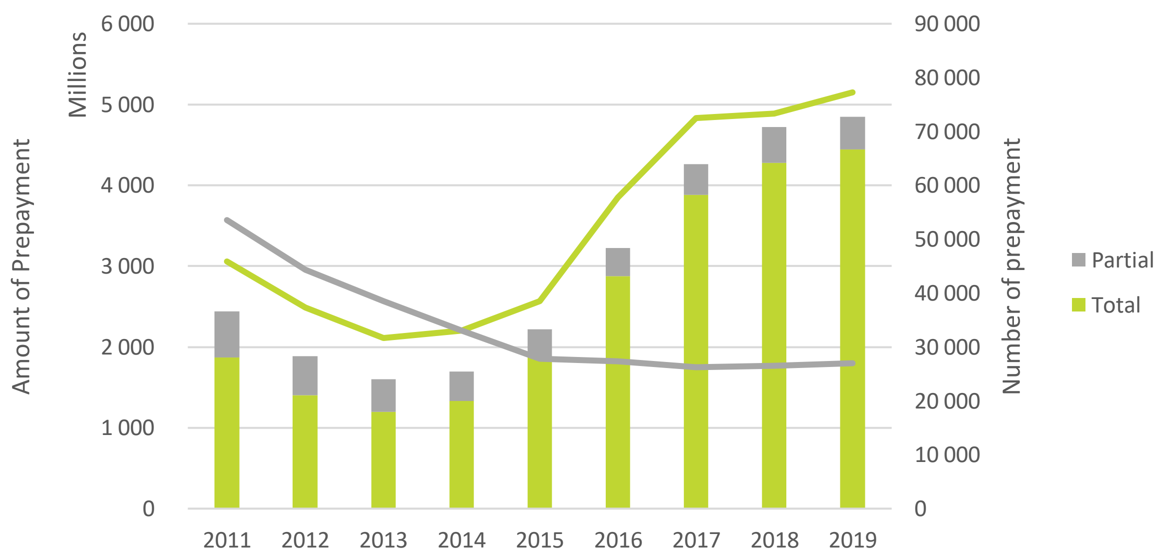

Figure 4 – Amount of prepayment for total and partial repayments (bar chart) and number of total and partial repayments (line chart). Source: Report on Monitoring of Retail Banking Markets from Banco de Portugal, data aggregated by the author (Banco de Portugal, 2019)

These reports, depicted in Figure 4, show a decrease in both the prepayment amount and number of prepayments during the financial crisis, with a significant increase in the number of total prepayments and an overall increase in the prepayment amount in economic recovery times. There has been a significant decrease in the amount and number of partial prepayments in the last ten years, stabilizing at around 26,000 per year since 2015 (Banco de Portugal, 2012-2019).

2.1.3. Machine learning models

To perform this study, machine learning models will be used, which overcome the limitation of a pre-specified form, usually a linear one, especially variable interactions, which are a significant component of the nonlinear effects. Two types of machine learning models have been used to model prepayment – neural networks and random forests (Deloitte, 2019; Sirignano et al., 2018; Sousa et al., 2013).

The growing application of machine learning (ML) is associated with increased computing power and a reduction in the investment needed to use this growing capacity. Hao (2018), of the Massachusetts Institute of Technology (MIT), defines ML algorithms as the use of statistical methods to find patterns in large amounts of data, namely numerical data, text, images, and interactions (such as social media interactions and clicks). Stanford University defines ML as the scientific method that allows computers to function without being explicitly programmed in its online course. SAS describes ML as the data analysis method that automates the construction of analytical models, being a branch of artificial intelligence whose rationale is that systems can learn from the data, identify patterns and make decisions that minimize human intervention, where its iterative nature allows models to adapt when exposed to new data. IBM adds to the previous definitions that, when data is ingested in the model over time, the models learn from it and increase its accuracy. In short, the rationale that governs ML is simple – find a pattern and apply it (Hao, 2018; IBM Cloud Education, 2020; SAS, 2020; Stanford University, 2020).

There are, generically, three methods for the learning of ML models (Hao, 2018; IBM Cloud Education, 2020):

i. **Supervised learning**: where the data is categorized, and it is through this categorization that the algorithm knows what to look for, and the model is created. This is the learning used in the prepayment models where customers are categorized by whether the event (prepayment) occurred, where the algorithm will try to find patterns and similarities between the clients who prepaid;

ii. **Unsupervised learning**: where the data is not categorized, and the algorithm aims to discover patterns in the data, grouping them according to their similarities. This type of learning can be used to categorize types of target customers for the launch products or the application of discounts;

iii. **Reinforced learning**: where the algorithm learns by trial and error to achieve a clear objective, being reinforced (or penalized) as its behaviour helps or delays reaching the objective. This type of learning is the foundation for Google's AlphaGo.

IBM defines the methodology for developing any ML model or application in four steps: (IBM Cloud Education, 2020)

i. **Select and prepare training data**: the records will be divided between training data – which will be used to train the model – and test data – which will be used to select the model and test its performance. The training data is equivalent to a representative set of data, which the model will ingest to solve the problem. The data must be pre-processed, that is, analysed for the existence of outliers, biases, distribution of variables, and randomized.

ii. **Choice of an algorithm**: the type of algorithm will depend on the type of data (with or without categorization), the amount of data, and the type of problem to solve.

iii. **Algorithm training**: the training occurs as an iterative process that involves presenting the records and variables to the model and comparing the results obtained from those that should have been obtained. With the analysis of this comparison, the model will adjust the parameters and will run again.

iv. **Use and improvement of the model**: the final step involves the use of the model in new data, allowing it to improve its precision and effectiveness over time

2.1.3.1. Artificial Neural Network

Artificial neural networks (ANN) were designed to mimic how the human brain works, consisting of a series of interconnected nodes representing neurons that are usually structured in layers. The nets can thus be described as networks of computational elements that respond to inputs and learn to adapt to the environment and data. These learn from the data introduced in the model and distinguish the relationships between credit customers and their prepayment rate, determining which characteristics are more important for this prediction. The most used network structure is the multilayer perceptron (MLP or backpropagation), as they allow for non-linear data, being theoretically capable of modelling any decision process (Anderson, 2007; William Edward Henley, 1995; Mcclelland & Rumelhart, 1986; Mester, 1997; West, 2000).

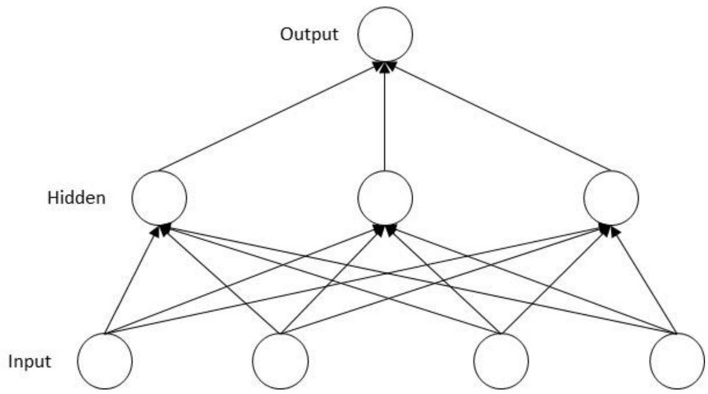

Figure 5 – Generic schema of an MLP network (Desai et al., 1996)

In a schema as presented in Figure 5, there are three layers of nodes: an input layer that propagates information to the "hidden" layer, which receives a weighted sum of the inputs and calculates the output value through an appropriate transformation (as a sigmoid and hyperbolic tangent function) and the output layer, which receives the calculated values. The weight calculation between the nodes in an MLP network uses the backpropagation rule, which minimizes the difference between the expected output values and the actual values (Desai et al., 1996; William Edward Henley, 1995; Mcclelland & Rumelhart, 1986; West, 2000).

The input nodes correspond to the variables used in the model, and the output nodes correspond to the customer's value. The value of the model outputs, for the k-record can be expressed, thus, according to the input values (X) and the weights of the network (w) (West, 2000):

$$Y_k = \sum_{h=1}^{2} w_{kj} \left(g \left(\sum_{i=1}^{2} w_{ji} X_i \right) + w_{jb} \right) + w_{ib} \, , k = 1,2, \tag{1}$$

where i is given by the input neurons, j the neurons of the hidden layers and b equals the skew values and $g(.)$ is the transfer function (e.g. hyperbolic function) (West, 2000).

ANN models allow for increased flexibility compared to more traditional statistical models. They have no assumptions regarding the distribution of variables or the functional form of the relationship between them and allow for highly nonlinear relationships (Altman et al., 1994; Fractal Whitepaper, 2003; Lessmann et al., 2015; Mester, 1997; Munkhdalai et al., 2019; Sirignano et al., 2018).

The main advantages of ANN models are their flexibility, as previously mentioned, without assuming the distribution of variables and allowing for the existence of highly nonlinear relationships; their capacity to process large volumes of data, without suffering the disadvantages related to sparse data; the "hidden" layer allows the data to contain complex nonlinear relationships and dependencies between variables, allowing ANNs to recognize these interactions and relationships between them; the parallel nature of the model may be more appropriate for complex and multidimensional data, and the output layer allows having multiple nodes (Anderson, 2007; Henley, 1995; Mcclelland & Rumelhart, 1986; Sirignano et al., 2018).

The main disadvantages of these models are: (i) their "black box" nature, which does not allow to describe the contributions of the different characteristics to the classification rule, which can result in data that overfits; (ii) they use a lot of data and computational power, requiring many iterations until the final model, meaning a high processing time for the training phase; (iii) they are expensive to implement and maintain and there is a possibility of obtaining illogical results.

Given their opaque nature, ANNs are not suitable for environments where the logic behind decisions must be understood. A potential solution to increase transparency is the use of numerical summaries, which make it possible to understand the relative importance of each variable (Altman et al., 1994; Anderson, 2007; Fractal Whitepaper, 2003; William Edward Henley, 1995; Mcclelland & Rumelhart, 1986).

2.1.3.2. Random Forest

A decision tree consists of a series of sequential nodes – the tree trunks – which divide subsets of the dataset based on the values of the characteristic under analysis; and leaves that specify the predicted class (or probability of belonging to the class). The tree's construction follows the rationale that each segregation, given by the nodes, will increase the purity of the descending nodes (compared to the parent node). Generically, there are two components in the construction of decision trees (William Edward Henley, 1995):

i. Selection of rules to segregate nodes – growth phase: which includes the breakdown of the source leaf into a series of nodes;

ii. Selection of when to declare the node as terminal – pruning phase: a tree is developed until the records on a specific leaf are below an established threshold or no longer possible to expand further. Thus, the pruning phase means the replacement of decision nodes by leaves.

Below is an example of a simplified decision tree:

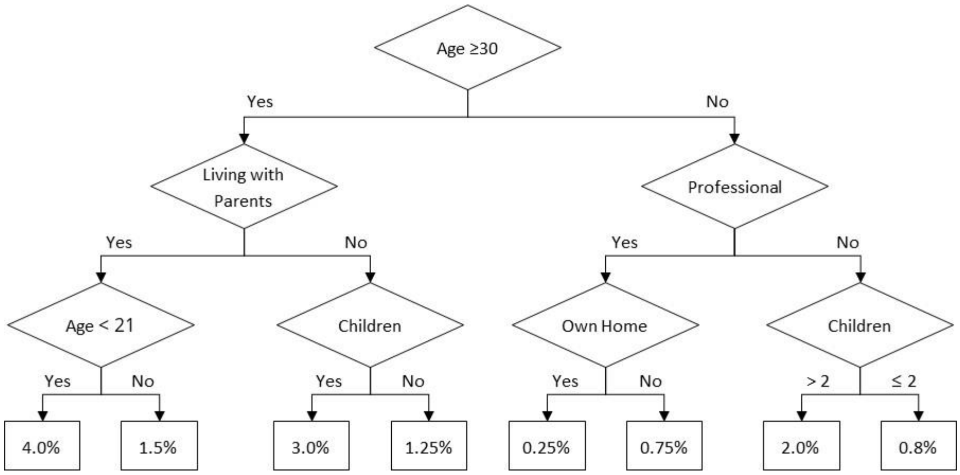

Figure 6 – Schema of a decision tree (Anderson, 2007)

One of the problems with decision trees is lower precision (compared to other ML models) and generalization difficulties, weaknesses that can be overcome using multiple trees. Hence appear the random forest models, which consist of a series of classifiers with the structure of decision trees, allowing each tree to "vote" on the most popular class. The various decision trees develop a "forest", which allows the forecast accuracy to be significantly increased.

These tree-based homogeneous ensemble algorithms (as the same algorithm, decision trees, is always used) can be divided into two families: boosting and bagging algorithms. Boosting algorithms, of which Adaboost (Adaptative Boosting) is one of the most famous, combine multiple weak learners into one giving more weight to the worst models. Here, the trees are grown through successive reweighting of the training data, increasing the weights of the records that were poorly classified in the previous iterations. This method of forming a model by readjusting earlier, weaker models, giving greater weight to poorly classified records is known as "boosting".

Bagging models use bootstrap aggregations (bagging), where trees are trained independently on bootstrap samples (sampling with replacement) of the same size as the training data and average the individual predictions (Breiman, 2001; Freund & Schapire, 1996; Ho, 1995; Larkin & McManus, 2018; Lessmann et al., 2015; Mishina, Tsuchiya, & Fujiyoshi, 2014; SAS, 2021b; Wyner, Mease, & Bleich, 2017).

Bayesian Model Ensemble (BME) or Averaging is an alternative ensemble learning classifier aiming at finding a composite model that best approximates the actual data generation process (known historical data) and its multiple sources of risk. The BME composite model design is set to be superior to the individual candidate models because, first, it explicitly addresses model uncertainty. Second, because each model's shortcomings are ideally compensated within a statistically (data) driven optimal combination. Third, because conditioning the statistical inference on a set of statistical models minimizes the individual model-based biases and produces more realistic confidence intervals. This in turn improves the out-of-sample forecasting precision and provides a more accurate representation of forecast uncertainty for decision-making (Bravo et al., 2021; Bravo & Ayuso, 2020, 2021; Bravo, 2019, 2021).

The main advantages of random forests are their precision, robustness to outliers, and noise in the data; it calculates internal estimates of the error, strength, correlation, and importance of variables. It is simple and easily configurable since it has few parameters that must be parameterized. However, tend to have lower performance when irrelevant variables are included in the model, and the interpretability is also reduced compared to single tree models (Breiman, 2001; Friedman, 2001; Nikulski, 2020).

2.2. INFORMATION CONSIDERED

The models will need to be trained using data that will allow for recognizing patterns and trends in the data. Typically data may be used at the mortgage portfolio level or individual loan data, as in this study (Charlier & van Bussel, 2001; Jacobs et al., 2005).

For the purposes considered in this study, research has mainly used the following information (Shunqin Chen et al., 2021; Li, 2014; Liang, Jin, & Wang, 2019; Louzis, Vouldis, & Metaxas, 2010; Saito, 2018; Sousa et al., 2013):

i. **Loan characteristics**: loan amount, loan age, the amount for the monthly payment (installment rate), interest rate, homeownership status/category for the loan request, term of the loan, loan-to-value (both at origination and monthly, this is calculated as the loan amount divided by the underlying property value), the value of collateral security, type of collateral security, location of the property, number of days overdue, and an indicator of prepayment;

ii. **Client personal characteristics**: age, sex, marital status, number of dependents, district of address, educational qualification, monthly income, occupation/work status, and employment history;

iii. **Client credit history**: which includes information on the number and amount of open credits, client history with financial institutions (checking account, the average balance in checking account, loans outstanding, loans defaulted, number of days with delay in payments, collateral/guarantee), and the client's financial capacity (total assets of the borrower, gross income of the borrower, gross income of the household, monthly costs of household, debt to income ratio);

iv. **Macroeconomic variables**, which affect aggregated behaviour: evolution of house pricing, unemployment rate, GDP and GDP monthly variation, divorce rate, the month of the year, consumer confidence, minimum wage variation.

2.3. PERFORMANCE ASSESSMENT

To estimate the performance of the models, the measures used can be distinguished into three main categories (Brownlee, 2018; Grebenar, 2018; Lessmann et al., 2015; Narkhede, 2018):

i. **Measures that assess the discriminatory capacity of the model**, such as the Area Under the Receiver Operating Characteristic Curve (AUC) and the Gini Index. The ROC (Receiver Operating Characteristic) curve represents the probability curve, so the AUC represents the degree of discrimination of the model, i.e., how much the model can distinguish between classes. This indicator always takes values between 0 and 1, and the higher the better;

ii. **Measures that assess the accuracy of the model's probabilistic forecasts**, such as the Brier Score. The Brier Score (BS) calculates the mean square error between the predicted probabilities and the expected values. It always takes values between 0 and 1, and the smaller the better. The study used for this benchmark (Lessmann et al.) focuses on credit scoring scorecards where the probability is particularly relevant, which is not as relevant in our study;

iii. **Measures that assess the correctness of the predicted categories**, such as the classification error. The misclassification calculates the ratio between the wrongly classified records (both false positives and false negatives) and the total number of records. It always takes values between 0 and 1, and the smaller the better.

According to Lessmann et al. (2015), it should be considered more than one metric to measure performance, as it was considered that, even being popular methods, only jointly do they allow for the evaluation of various models' precision angles. These will be the AUC, which can be replaced by the H-measure, the Gain or Lift – which measures the effectiveness of a classification model – and the Kolmogorov-Smirnov – which measures the degree of separation between the negative and positive distributions (Anderson, 2007; Lessmann et al., 2015; Siddiqi, 2012).

3. METHODOLOGY

To create a prepayment model, it is essential the availability of a large dataset with a significant history and high quality of loan prepayments (Sousa et al., 2013). To guarantee high-quality data, the work to be performed will be divided into three phases and will begin with data pre-processing and variable selection, where it is decided what variables should be used for the models. Moreover, it will be proceeded by the development of the models and their performance assessment. These are iterative and cyclical phases, and, as such, after the first phase of testing the models, it will be performed further data pre-processing to refine and improve the results.

This process can be summarized by the diagram below, whose framework was adapted from the literature, with a diagram of the steps performed and the software where they were performed in Appendix 1 (Handhika et al., 2019; Lessmann et al., 2015; Munkhdalai et al., 2019; Xia et al., 2017):

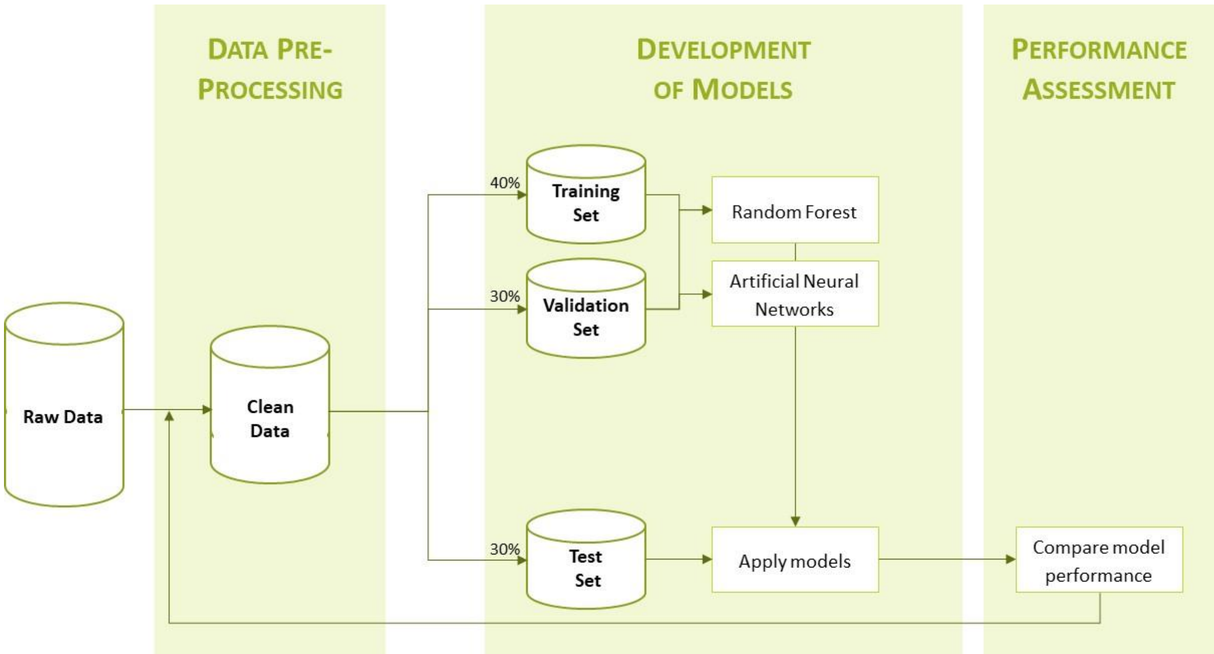

Figure 7 – Representation of the methodology followed, adapted by the author from the literature.
Source: Handhika et al., 2019; Lessmann et al., 2015; Munkhdalai et al., 2019; Xia et al., 2017.

The dataset to be used for these models is comprised of monthly observations in mortgage loans in a large Portuguese bank gathered from January of 2011 to June 2020. It is comprised of 69 variables, where 48 are inputs from the bank, and the remaining were added by the author and detailed in this chapter. The complete list of variables is detailed in **Appendix 2**.

Given the adverse shock in the payment behaviour caused by the Covid-19 pandemic, and consequent state support for home loans, such as moratoria[6], the data from 2020 was excluded from the sample, and considered the timeframe from January 2011 to December 2019. In addition, and to guarantee a minimum amount of granularity, contracts need to have, at least, two years' term (and, as such, the

[6] Moratoria, in the context of home loans, resulted in a suspension of the payments, and arose in the context of the pandemic outburst as a way to protect consumers and their permanent home mortgages. (*Decreto-Lei n.º 10-J/2020*, 2020)

minimum contract end date was January 2013, and the maximum opening date was December 2017). With this initial selection, the data comprised 90.165.830 records, from 1.029.040 contracts, with the oldest contract from 1970.

As shown in the literature review, the following information is mainly used for these models:

i. **Loan characteristics**: loan amount, loan age, the amount for the monthly payment (installment rate), interest rate, homeownership status/category for the loan request, term of the loan, loan-to-value (both at origination and monthly, this is calculated as the loan amount divided by the underlying property value), the value of collateral security, type of collateral security, location of the property, number of days overdue, an indicator of default and an indicator of prepayment;

ii. **Client personal characteristics**: age, sex, marital status, educational qualification, monthly income, occupation/work status, debt to income ratio;

iii. **Client credit history**: number and amount of open credits;

iv. **Macroeconomic variables**, which affect aggregated behaviour: evolution of house pricing, unemployment rate, GDP and GDP monthly variation, divorce rate, the month of the year, consumer confidence, minimum wage variation.

The dataset's variables, detailed in **Appendix 2,** were, therefore, grouped in the following categories, resulting from an adaptation from the list above:

i. **Loan characteristics** – includes data on the contract in question, which includes the total amount owed, and loan characteristics;

ii. **Client** – with the client personal characteristics, which includes financial, demographic, and employment indicators;

iii. **Behaviour in the bank and financial system** – with the client credit history, which includes the client history with financial institutions;

iv. **Point in time** – includes the month and year of observation;

v. **Macroeconomy –** includes the variables defined in Table 1.

As referred above, and according to the studies by World Bank and Dastile et al. (2020), macroeconomic variables were added, as they are important in the study of the client's behaviour and their response to the economy and its changes. Thus, the following macroeconomic variables were be added to the dataset, according to the ones used by Bellotti and Cook (Bellotti & Crook, 2009; Dastile et al., 2020; The World Bank Group, 2019):

Name	Description	Source	Periodicity
ED_LICENC_TVH	The number of licensed buildings, year-on-year change. I.e. authorization granted by the City Councils under specific legislation for the execution of Works (new constructions,	Statistics Portugal (INE)	Monthly

Name	Description	Source	Periodicity
	extensions, transformations, restorations, and demolitions of buildings).		
ENDIV_PART_TVH	Indebtedness of families and non-profit institutions serving families in Portugal, year-on-year change.	Bank of Portugal (Bpstat)	Monthly
GRAU_POUP_PART_TVH	Degree of household savings, year-on-year change.	Statistics Portugal (INE)	Monthly
IND_COINC_TVH	Coincident indicators for private consumption, year-on-year change. It seeks to capture the underlying evolution of the year-on-year variation in private consumption.	Bank of Portugal (Bpstat)	Monthly
IND_PRECOS_HAB_TVH	Housing price index, which measures the evolution of housing prices in the residential market in the national territory, year-on-year change.	Statistics Portugal (INE)	Quarterly
IND_SENT_ECO_TVH	Economic sentiment indicator, year-on-year change. This short-term indicator allows the monitoring of the evolution of the economic environment and anticipating the evolution of the main macroeconomic aggregates for Portugal.	Bank of Portugal (Bpstat)	Monthly
N_FOGOS_CONST_TVH	Number of licensed dwellings in new buildings for family housing, year-on-year change.	Statistics Portugal (INE)	Monthly
PERSP_SIT_EC	Outlook on the country's economic situation over the next 12 months, year-on-year change.	Statistics Portugal (INE)	Monthly
PIB	GDP at market prices, year-on-year change.	Bank of Portugal (Bpstat)	Quarterly
TAXA_INFLACAO_TVH	Harmonized consumer price index, year-on-year change.	Bank of Portugal (Bpstat)	Monthly
TAXA_JURO_DP_TVH	Interest rate in term deposits (< 1 year, private individuals), year-on-year change.	Bank of Portugal (Bpstat)	Monthly
TAXA_JURO_HAB_TVH	Interest rate in mortgage loans (private individuals), year-on-year change.	Bank of Portugal (Bpstat)	Monthly
TX_DESEMPREGO_TVH	Unemployment rate of the active population aged between 15 and 74 years, year-on-year change.	Statistics Portugal (INE)	Monthly
TX_DIVORCIO_TVH	Number of marriages dissolved by divorce, year-on-year change.	Statistics Portugal (INE)	Yearly

Table 1 – Macroeconomic variables added, and the respective source

As mentioned above, the initial datasets had monthly information for each of the contracts. To account for the monthly and yearly information, there were three modelling options: (i) treat the data as panel data and apply fixed or random effects (Allison & Christakis, 2017; Park, 2011; Williams, 2018), (ii) treat the data as longitudinal data (Shuo Chen et al., 2014; Shuo Chen & Bowman, 2011; Jing et al., 2011) and (iii) embed the information from the previous months (and years) in the variables used for modelling (Deloitte, 2019). For this study, and given the flexibility it allows, the third option will be used, i.e. incorporating past information in the variables. As such, and to account for the history of the contract, the following variables were added:

Name	Description	Formula
TOTAL_AMORT_PARCIAL	Total partial early repayments. It is a calculated variable based on the target variable, which indicates the existence of early repayments.	Count of the partial repayments, until the time ID of the observation.
TOTAL_MONTANTE_AMORT	Total amount repaid. It is a calculated variable based on the amount prepaid.	Sum of the partial repayment amount, until the time ID of the observation.

Table 2 – Variables added to the dataset

These variables were added in a way to guarantee that no future information was considered, i.e., follows the rationale shown in the example below:

ID	Date	Target Amort Partial	Amount Amort	Total Amort Partial	Total Amount Amort
1	201001	0	0	0	0
1	201002	0	0	0	0
1	201003	0	0	0	0
1	201004	1	150	0	0
1	201005	0	0	1	150
1	201006	0	0	1	150
1	201007	1	300	1	150
1	201008	0	0	2	450
2	202001	1	500	0	0
2	202002	0	0	1	500
2	202003	0	0	1	500
2	202004	0	0	1	500

Table 3 – Example of the calculation performed

To perform a first explanatory phase, the data will be analyzed in two steps: (Grebenar, 2018)

i. **Univariate analysis**: in which the variables' discriminatory power will be analyzed, using descriptive statistics, such as the following:

Statistics [per variable]	Why?
Number of unique values	Variations in the population
Number of missing values	Correction of the data collected
Mean, median, and mode	Characterization of the average record
Histograms	How the population is distributed
Five highest and lowest values	Possible outliers/values likely to be errors

Table 4 – Descriptive statistics for data assessment (Vidal & Barbon, 2019)

ii. **Multivariate analysis**: in which combinations of variables will be analysed for their correlations, as highly correlated variables may generate collinearity issues in the models.

In our dataset, comprised of 69 variables, this initial exploratory phase was performed by analysing the histograms and box plots, which confirmed the existence of outliers. In particular, variables regarding the behavior in the bank and financial system (e.g. amount of credit, number of operations, and financial products) and income demonstrate a series of extreme values that may bias the models' results.

Furthermore, as shown in the following chapter, the variables are typically not highly correlated, and there are a series of variables with a high percentage of missing values. The histograms, bar charts, and boxplots can be found in **Appendix 3**.

3.1. Data Pre-Processing

As shown in Figure 7, data needs first to be processed and cleaned; thus, the first phase of the figure is data pre-processing. Pre-processing will allow for more accurate results, improving the consistency of the results and smoothing data, aiding its interpretation and use. This pre-work, where data gets transformed or encoded, is used to allow machines to parse it and prevent results known as GIGO easily – garbage in, garbage out, i.e. if the data has much noise and is incorrect, the results the model produces will not be good and will be untrustworthy (Gavrilova & Bolgurtseva, 2020; Pandey, 2019).

In broad terms, data pre-processing encompasses three steps (Gavrilova & Bolgurtseva, 2020; Jain, 2019; Vidal & Barbon, 2019):

› **Data cleaning** – this step implies performing a data quality assessment, identifying irrelevant, missing, and noisy data and phenomena such as outliers. After the conclusion of this step, it is expected that the data set is clear and complete.

› **Data transformation** – this step implies the generation of a different representation of the data, which may improve the model's predictive power.

› **Data reduction** – this step implies the analysis of variables relations to decrease the number of variables used in the models, which enables more accurate results (when the correct variables are selected) and more efficient models (when fewer variables are considered).

3.1.1. Data Cleaning

As stated previously, data cleaning implies a prior data quality assessment. The data cleaning wishes to analyse if the values within the same variables are consistent (an example of inconsistency would be the same variable including 'female' and 'woman' to identify the female gender), the presence of outliers (which are extreme results that may or not arise from errors in the data set), and missing values for the variables. This assessment is crucial to identify which records must be cleaned from the data, allowing for an improvement in the accuracy and consistency of the results (Gavrilova & Bolgurtseva, 2020).

Data cleaning usually involves a set of analyses (Gavrilova & Bolgurtseva, 2020; Jain, 2019; Pandey, 2019; Vidal & Barbon, 2019):

I. Missing data: when the variables have missing records, these may be sparse or plenty.
If the missing records are sparse, and even many missing for the same client, that record is usually eliminated. If the variable has a lot of missing values, it is usually eliminated from the data set. Where there are not many missing values, and where these were not deemed significant, these may be replaced through interpolation (either through the absolute mean/median/mode or the mean/median/mode of the k-nearest neighbours). According to Vidal & Barbon, below are presented four strategies to work with missing records:

Strategy	When?	Pros	Cons
Remove rows	› Large dataset › Few missing values There needs to be a complete dataset	Uses complete records and does not make assumptions	Data loss
Replace with unique value (e.g. '99999999')	› A high percentage of missing values in various variables Missing values may be possible for variables in the future	Preserves data and can accommodate missing values in the future	May introduce bias if the reason why there are missing values does not persist in the future
Replace with the average value	› Limited data › Few missing values Missing values are not possible for variables in the future	Preserves data	Assumes past missing values are no different from average records
Replace with a predicted value	› Limited data Few missing values	Preserves data and may be more realistic than using average values	Increases complexity of the model

Table 5 – Strategies for working with missing values. Source: (Vidal & Barbon, 2019)

II. Noisy data: this is usually due to data entry errors and faulty data collection, and it may manifest as duplicate/semi-duplicate records and inconsistent variables. It can be handled through binning (aggregate data in smaller segments of the same data and applying dataset preparation for each of them), regression (smoothing the data to fit a regression function), and clustering (aggregating similar groups in a cluster).

As described above, the initial dataset had a considerable dimension – around 90 million records – resulting from the joining of several tables and the use of historical information, resulting in some data quality issues.

The process of data cleaning involved the following stages, which led to an iterative reduction of records[7]:

I. The tables were crossed using an inner join, i.e. it was considered that the following fields were essential and could not, therefore, have any missing values: target variables (both partial and full repayment), prepayment amount, contract end and start date, loan term, the residual amount of the loan.
II. To reduce noisy data, it was defined a series of consistency checks (CC). When these conditions did not hold, the contract was eliminated:
    a. If the number of installments paid reduces, the residual amount must also reduce;
    b. If there is one full prepayment, the future residual amount must be null;
    c. There can only be one full prepayment per contract;
    d. The residual amount cannot increase in two sequential months;
    e. The financed amount cannot increase in two sequential months;
    f. The financed amount must always be less than or equal to the residual amount;
    g. The age of the client must lie between 18 and 80;
    h. The number of days past due must be less than or equal to 365 days;
    i. The value of monthly installments in the bank must be less than the amount of credit (liabilities) of the client in the bank.

These data cleaning stages led to the following reduction of records, using the numbers above:

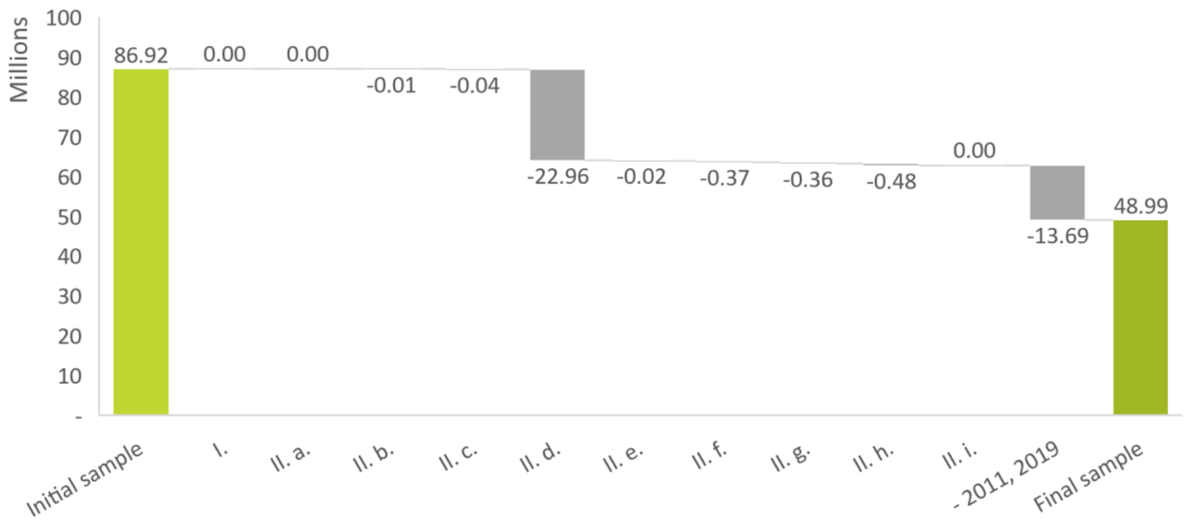

Figure 8 – Reduction of records through data reduction.

[7] Note that the elimination of records involved eliminating every record of the contract. I.e. if the contract had a data quality issue in one specific month, every record from that contract is eliminated.

The most significant reduction is the increase of the residual amount in consequential months, even though this may happen through credit reinforcement, these records were disregarded, for the purpose of this exercise. The second highest reduction resulted from the elimination of the observations from 2011 and 2019, in order to reduce the size of the dataset. The 37.93 million records which were eliminated implied a reduction of 243,087 contracts.

After this initial assessment, the numerical variables descriptive statistics were the following:

Variables	# Missing Values	%	Mean	Maximum	Minimum
ANO_CONSTRUCAO	2 172 884	52.8%	1 811	9 999	1
DATA_ABERTURA	-	0.0%	13-02-1943	28-12-1957	25-05-1920
DT_NASCIMENTO	425 687	10.4%	10-08-1906	29-07-1939	-10193
ED_LICENC_TVH	-	0.0%	-2.13%	29.87%	-25.06%
ENDIV_PART_TVH	607 333	14.8%	-2.62%	0.13%	-4.03%
GRAU_POUP_PART_TVH	-	0.0%	46.14%	230.77%	-86.15%
IDADE	425 687	10.4%	49	80	19
IND_COINC_TVH	-	0.0%	0.08%	2.90%	-6.40%
IND_PRECOS_HAB_TVH	-	0.0%	2.57%	12.24%	-8.17%
IND_SENT_ECO_TVH	-	0.0%	3.06%	27.39%	-18.65%
INIB_CHEQUE	425 678	10.4%	0	1	0
LTV_ATUAL	1 102 664	26.8%	1	206	0
LTV_ORIG	1 105 286	26.9%	1	206	0
M_PRS_MENS_BANCA	495 351	12.0%	670	970 534	0
M_PRS_MENS_BANK	497 743	12.1%	640	970 534	0
MONTANTE_AMORT	-	0.0%	72	1 221 613	0
MONTANTE_FINANCIADO	303	0.0%	61 564	2 585 000	125
MONTANTE_RESIDUAL	-	0.0%	33 129	2 585 000	0
N_DIAS_ATRASO	943 850	23.0%	0	296	0
N_FOGOS_CONST_TVH	-	0.0%	1.95%	84.98%	-48.06%
N_OPER_BANCA_POT	-	0.0%	1	13	-
N_OPER_BANCA_REAIS	-	0.0%	2	27	-
N_OPER_BANK_POT	-	0.0%	0	8	-
N_OPER_BANK_REAIS	-	0.0%	1	10	-
N_PREST_PAGAS	943 850	23.0%	135	449	-1
N_PRODUTOS_BANCA	495 351	12.0%	4	93	0

Variables	# Missing Values	%	Mean	Maximum	Minimum
N_PRODUTOS_BANK	497 743	12.1%	2	33	1
PERC_UTILIZA	1 792 319	43.6%	0	66	0
PERSP_SIT_EC	-	0.0%	-18.82%	10.00%	-59.80%
PIB	-	0.0%	0.50%	3.60%	-3.60%
PRAZO	-	0.0%	348	720	24
PRAZO_RESIDUAL	-	0.0%	205	652	0
RENDIMENTO	596 670	14.5%	40 021 988	20 084 001 000 000	0
RESP_BANCA_POT	42	0.0%	75 206	374 293 858	-
RESP_BANCA_REAIS	498 858	12.1%	105 368	374 293 858	-
RESP_BANK_POT	512	0.0%	111 572	316 797 858	-
RESP_BANK_REAIS	495 754	12.1%	152 087	316 797 858	-
SALDO_DO_06M	143 142	3.5%	8 357	19 031 830	- 167 708
SALDO_DO_12M	143 142	3.5%	8 081	10 923 831	- 167 676
SALDO_DP_06M	2 254 455	54.8%	39 658	17 214 821	-
SALDO_DP_12M	2 254 455	54.8%	38 882	14 747 105	-
SCORING	1 012 650	24.6%	5	10	1
T_JURO	303	0.0%	2	28	-
T_SPREAD	303	0.0%	1	17	0
TAXA_INFLACAO_TVH	-	0.0%	0.93%	3.30%	-0.40%
TAXA_JURO_DP_TVH	-	0.0%	-27.60%	38.89%	-55.83%
TAXA_JURO_HAB_TVH	-	0.0%	-6.01%	59.18%	-29.91%
TOT_DEVEDORES_BANCA	499 008	12.1%	2	24	1
TOTAL_AMORT_PARCIAL	-	0.0%	0	6	-
TOTAL_MONTANTE_AMORT	-	0.00	242	1 221 613	-
TX_DESEMPREGO_TVH	-	0.0%	-4.35%	21.77%	-22.12%
Z_FIM_CTTO	-	0.0%	08-02-1972	17-12-2007	31-12-1952

Table 6 – Statistical descriptions of numerical variables

For the descriptive variables, the number of unique, missing values and the missing percentage was the following:

Variables	# Missing Values	%	# Unique Values	Mode
C_Postal	2 172 884	52.8%	67 524	2840000 (Arrentela)
Concelho	2 763 885	67.2%	308	Lisboa
Distrito	2 763 885	67.2%	29	Lisboa
ESTADO_CIVIL	46	0.0%	12	4 (Married)
HAB_PROF	430 886	10.5%	9	2 (High School)
FINALIDADE	303	0.0%	118	1180 (Permanent home purchase)
IND_CREDITO	0	0.0%	6	1 (Regular credit)
PROFISSAO	430 886	10.5%	645	232005 (Teacher)

Table 7 – Statistical descriptions of the categorical variables

We can observe that there is a high percentage of missing values in the age of the propriety, municipality, district, percentage of use of credit cards, and balance in term loans. These can be segmented into two typologies:

- Variables in which the records are missing, and there are no justifications besides data quality issues. Here, as stated above, the variables will be removed from the dataset. This is the case for the age of the propriety, municipality, and district, which are all variables that are given by the client (usually at the stage of the deed of the housing credit agreement), which may explain the data quality problems;
- Variables where missing means that there is no amount in that product. This is the case for potential credit operations and balance in term loans, as clients may not have this kind of limits or products. These records will be kept in the sample and substituted by 0.

For the remainder of the variables, whose missing values percentage was less than 50% (no more than 30%), the missing values will be imputed based on a predicted value, as per Table 5, through SAS Enterprise Miner's Impute Node. The imputation occurred for the variables[8] using as imputation method the Tree. Here, the other dataset observations are used to impute the missing values, where the variable in question is used as a target (e.g. when imputing values for the "SCORING", the scoring will be considered the target). When the variables display a non-normal distribution, this method allows for more accurate results, with less bias to the median or average (SAS, 2021c; Vidal & Barbon, 2019).

[8] Variables with use of Impute node: amount of real credit in the bank and financial system, balance in sight deposits, check inhibition indicator, client age, debt-service rate in the bank and financial system (variable added and explained in the next chapter), financed amount, indebtedness of families in Portugal, interest rate, LTV (both current and original), monthly instalments in the bank and financial system, number of days past due, number of debtors, number of instalments paid, number of products in the bank and financial system, number of real operations in the bank and financial system, percentage of term elapsed (variable added and explained in the next chapter), scoring, spread rate, yearly income of the client.

For the categorical variables, whose '-1' represents the unknown class, the missing values were replaced by -1.

Finally, and to remove the extreme values and reduce sample variability, it was applied windsorizing to the numerical variables with the most extreme outliers[9]. It consisted in replacing the lower extremes by the first percentile, and higher extremes with the ninety-nine percentile. Windsorizing assumes that these percentiles result in more plausible values and alleviates the bias by substituting with more attenuated values (Ghosh & Vogt, 2012; Grebenar, 2018).

The histograms and bar charts before and after this change can be found in **Appendix 4**, for the variables where there was a change, detailed in footnotes 8 and 9.

3.1.2. Data Transformation

This step implies transforming the data to be suitable for the models and the creation of new variables. These new variables will be based on the existing ones, which may have more predictive power and will assist the analysis to be performed in the third step (i.e. data reduction) and can also be a requirement for some of the methodologies used.

Data transformation usually occurs using the following transformations (Gavrilova & Bolgurtseva, 2020; Jain, 2019; Vidal & Barbon, 2019):

I. Scaling of variables: scaling the data values in a specified range (usually from -1 to 1, or 0 to 1), through techniques such as decimal scaling or Z-score, which is useful for algorithms based on distance metrics.

II. Attribute selection: construction of new attributes from the original set, which may mean developing financial stability ratios, for example.

III. Discretization: attributing interval levels or conceptual levels (e.g. positive/negative or young/ adult/senior) to numerical attributes for improving efficiency.

IV. Conversion: converting categorical data to numeric, attributing numbers to the categories, so that it can be widely used for the models.

V. Generalization: converting very granular data to a higher level (e.g. home address generalized to town or country).

VI. Linearization: transforming the variable so that its relationship with the target variable is linear.

[9] Variables with windsorizing: amount of real credit in the bank and financial system, balance in sight and term deposits, financed amount, interest rate, LTV (both current and original), monthly instalments in the bank and financial system, number of debtors, number of products in the bank and financial system, number of real and potential operations in the bank and financial system, percentage of credit card usage (only for the observations above the 99th percentile), residual amount (only for the observations above the 99th percentile), spread rate, yearly income of the client.

In this study, the models used – artificial neural networks and random forests – do not require scaling of variables, conversion, or linearization.

It was, however, performed attribute selection and generalization. The first consisted of creating five additional variables, and the latter of the conversion of highly granular categorical variables to a more aggregated level.

Thus, it was created the following variables:

- **Debt-service ratio, in the bank**: a measure of the proportion of the client's monthly installment in the bank in monthly income. This was calculated using the following formula:

$Tx_{esforco_{bank}} = \frac{M\_PRS\_MENS\_BANK}{RENDIMENTO/12}$, as the income represents the yearly income. (2)

- **Debt-service ratio, in the financial system**: a measure of the proportion of the client's monthly installment in the financial system in monthly income. This was calculated using the following formula:

$Tx_{esforco_{bank}} = \frac{M\_PRS\_MENS\_BANCA}{RENDIMENTO/12}$, as the income represents the yearly income. (3)

- **Percentage of residual term elapsed**: measure the proportion of the term in the contract that has elapsed. This was calculated using the following formula:

$$Perc\_prazo = \frac{PRAZO\_RESIDUAL}{PRAZO} \quad (4)$$

- **Total of partial repayments**: total partial early repayments. Results from the count of the partial repayments, until the time ID of the observation, whose calculation approach was detailed in chapter 3.

- **Total amount in early repayments**: total amount repaid. Results from the sum of the partial repayment amount, until the time ID of the observation, whose calculation approach was detailed in chapter 3.

The generalization was performed for the location of the property, for the purpose of the loan, for marital status, and for the client job:

- **Postal code**: it was aggregated by county and district, based on the mapping by *CTT - Correios de Portugal, S.A.*. This aggregation resulted in a reduction from 67.524 unique values, to 308 counties and 29 districts. (CTT, 2021)

- **Purpose of loan**: it was aggregated based on major groups, defined by the author, based on the description of the loan's purposes. This aggregation reduced the 118 unique values (and 117 unique numeric values) to 11, plus a twelfth class of missing. The complete mapping is shown in Appendix 5, with the 11 classes presented below:

Loan purpose - aggregation
Acquisition permanent home

Loan purpose - aggregation
Acquisition secondary home
Acquisition property home
Acquisition other home
Acquisition land / construction
Works
Installation of prefabricated homes
Investments in real estate
Acquisition garage / others
Credit restructuring
Acquisition of goods

Table 8 – Aggregation categories in the loan purposes. Source: Author aggregation

The analysis of the loan purposes led to the elimination of contracts which had one the loan purposes below, as they were deemed as unrelated to the mortgage contracts:

Description
Auto Ligeiro Peso Bruto 2.500
Auto Ligeiro Peso Bruto 2.500
Moto Novo
Caravana
Formacao Profissional
Outras Desp Educacao/Formacao
Despesas Com Saude
Ferias/ Viagens/ Lazer
Festas Familiares
Automovel Novo
Curso Superior
Curso Especializado/Executivo
Mestrado-Portugal
Doutoramento-Portugal

Description
Pos-Graduacao
Mba
Saude - Cirurgia Estetica
Saude - Medicina Dentaria
Licenciatura-Portugal
Automovel Usado
Outros Cursos-Portugal

Table 9 – Loan purposes which were eliminated

- **Marital Status**: it was aggregated based on major groups defined in the bank's internal system. This aggregation reduced 12 unique values to 4, plus a twelfth class of missing. The complete mapping is shown in, with the 4 classes presented below:

Marital Status - aggregation
Single
Married/De facto Union
Separated / Divorced
Widower

Table 10 – Aggregation categories in the marital status. Source: Bank's internal aggregation

- **Profession**: it was aggregated based on the highest group presented by INE in *Classificação Portuguesa das Profissões – Grande Grupo*. The author performed this mapping based on job descriptions and the chapter "*Estrutura*" of the document. This aggregation reduced the 645 unique values (and 563 unique numeric values) to 11, plus a twelfth class of missing. The complete mapping is shown in Appendix **7**, with the 11 classes presented below (Instituto Nacional de Estatística, 2011):

Profession - aggregation
Armed Forces Professions
Representatives of the legislative power and executive bodies, directors and executive managers
Specialists in intellectual and scientific activities
Intermediate level technicians and professions
Administrative staff
Personal, safety and security services workers and vendors

Profession - aggregation
Farmers and skilled workers in agriculture, fishing and forestry
Skilled workers in industry, construction and crafts
Plant and machine operators and assembly workers
Unskilled workers
Student

Table 11 – Aggregation of categories in the loan professions. Source: author aggregation based on *Classificação Portuguesa das Profissões – Grande Grupo* by INE (Instituto Nacional de Estatística, 2011)

Lastly, the categorical variables already had numerical attributes and hence did not require conversion.

3.1.3. Data Reduction

Data reduction through feature selection carries as main benefits the reduction in processing time and storage requirements, allowing for a better data understanding and visualisation. Reducing the variables in the dataset diminishes the risk of the curse of dimensionality, having as the primary goal improving the models' performance (Guyon & Elisseeff, 2003).

Generally speaking, feature selection methods may be divided into three main categories (Brownlee, 2019; *Feature selection in machine learning*, 2013; Guyon & Elisseeff, 2003; Kaushik, 2016):

I. Filter methods: these evaluate the features against a proxy and encompass a pre-processing step, and are, therefore, independent of the choice of predictor and model. They are computationally efficient and statistically robust against overfitting. The specific methods used are divided according to the type of variable and are presented in the table below:

Input \ Output	Numerical	Categorical
Numerical	› Pearson's Correlation › Spearman's › T-test	› Linear Discriminant Analysis › ANOVA › Kendall's
Categorical	› ANOVA › Kendall's	› Chi-squared › Mutual information

Table 12 – Filter methods according to variables' types (Brownlee, 2019; Kaushik, 2016)

II. Wrapper methods: these use a subset of features, where they train the model and, based on the inferences from this model, decide to add or remove features from the subset. As these evaluate the models in the same metric as they are being optimized, wrapper methods tend to generate the highest accuracy and give the best results. However, with every feature added or eliminated, the process must restart, becoming time intensive and computationally expensive. Two methods used are forward selection and recursive feature elimination. In forward selection, the models commence with no features, which are added iteratively and assessed if they improve, or not, the performance of the model (the order used follows variable importance). In backward elimination, the model begins with every feature, which are progressively eliminated, following the least significant order.

III. <u>Embedded methods</u>: results of combining the previous methods through the use of algorithms that have built-in feature selection methods. These methods guide their search by estimating changes in the objective function, with two of the most popular being LASSO and RIDGE regression, which have penalization functions to reduce overfitting. LASSO's penalty is equivalent to the absolute value of the magnitude of coefficients, whereas RIDGE's penalty is equivalent to the square of the magnitude of coefficients.

In this data set, the leading data reduction step was the passage from monthly to annual data, which was performed by selecting the observations from January of each year. This reduction allowed for a significant improvement in computational performance. Reducing the dataset from 48.99 million records to 4.06 million records. This position of information was mainly driven by the operational efficiency it allowed, and, retrieving the information from January, it can be assessed the contracts' characteristics before the event, i.e., prior to the prepayment. This allows for an analysis of the characteristics that mainly help to explain the prepayment events.

With regards to variable selection, it was first analyzed the correlation of the variables, based on Table 12. As both the numerical and categorical variables had numerical values, only Pearson's Correlation needed to be performed, based on SAS's `PROC CORR`, which generated the matrix shown in **Appendix 8.** In this matrix, it can be seen that there is no significant correlation with the target variables: the highest correlated variable with the full repayment is the amount of previous repaid capital (with a correlation of 0.10) and in the partial repayment is the number of previous repayments (with a correlation of 0.13). In terms of the explanatory variables:

› There are strong correlations between the year and the macroeconomic variables, as expected, and between the macroeconomic variables themselves;

› There is, logically, a strong negative correlation between the date of opening of the loan and the number of installments paid.

› The income and age are positively correlated, i.e., the older the client, the more income it makes (with a correlation of 0.29).

› Both LTVs are negatively correlated with the financed amount (which means that the higher the LTV, the riskier to the bank and, thus, the lesser the financed amount), with a correlation of -0.20 for current LTV and -0.23 for origination LTV.

› The monthly installments are, naturally, positively correlated to the overall amount of liabilities in the bank and the financial system (with correlations of around 0.70).

After performing the filter methods, stepwise regression was performed to analyze which variables can be added to the model in terms of their value. The stepwise regression was computed through the use of the `Regression Node` in SAS Enterprise Miner, and selected the following variables:

› The **full repayment model** selected 39 variables, which are detailed in **Appendix 9** ordered by importance. This presents as additional variables compared to the partial repayment model, the macroeconomic variables of the number of licensed buildings, the economic sentiment indicator and the divorce rate, and the opening date of the contract, the check inhibition indicator, the number of days past due, the number of financial products in the bank, the yearly income, the amount of real liabilities in the financial system and bank, the amount of potential liabilities in the financial system, the debt-service ratio in the financial system, the number of debtors in the financial system and the number of potential operations in the bank.

- The **partial repayment model** selected 30 variables, which are detailed in Appendix 10, ordered by importance. This presents as additional variables compared to the full repayment model, the macroeconomic variables of the coincident indicators for private consumption and the interest rate in term deposits, and the monthly installments in the financial system, the debt-service ratio in the bank, the number of financial products in the financial system and the number of potential operations in the financial system.

Finally, to simplify this auxiliary process, between the LASSO and RIDGE regression, the LASSO regression was performed as it involves a more simplified data handling process. Hence, the LASSO Regression was computed using the `PROC GLMSELECT` in SAS, using ten-fold cross-validation and selection equal to "LAR", i.e. "Least Angle Regression". The variables selected were the following (SAS, 2021d; Ulloa, 2017):

- The full repayment model selected 45 variables, which are detailed in Appendix 11 ordered by importance. This presents as additional variables compared to the partial repayment model, the macroeconomic variables of the inflation rate, the indebtedness of families, the interest rate in term deposits and the economic sentiment indicator, and the amount of real credit in the financial system, the age, the number of days past due, the number of financial products in the financial system, the balance in term deposits, 12 months, the number of potential operations in the financial system, the monthly instalments in the financial system and the residual loan term.

- The partial repayment model selected 40 variables, which are detailed in Appendix 12 ordered by importance. This presents as additional variables compared to the full repayment model, the macroeconomic variables of the unemployment rate, the housing price index, the number of licensed buildings and the GDP, and the origination LTV, the number of real operations in the bank, the amount of potential credit in the bank and balance in sight deposits, 12 months.

3.2. Development of Models

The models will be computed using SAS Enterprise Miner. The nodes used will be detailed in the corresponding chapters. As depicted in Figure 7, the dataset will be segregated between training, validation, and test set in the commonly used ratio of "40%-30%-30%" (Baesens et al., 2003; Lessmann et al., 2015):

3.2.1.1. Artificial Neural Networks

As stated above, ANN models are designed to mimic the human brain through a series of interconnected nodes. The schema presented below presents the three layers of nodes: an input layer that propagates information to the "hidden" layer, which receives a weighted sum of the inputs and calculates the output value through an appropriate transformation – examples of transformation functions might be sigmoid and hyperbolic tangent function – and the output layer. The calculation of the weight between the nodes, in an MLP network, is done using the backpropagation rule, which aims to minimize the difference between the expected value of the output values and the actual values (Desai et al., 1996; William Edward Henley, 1995; Mcclelland & Rumelhart, 1986; West, 2000).

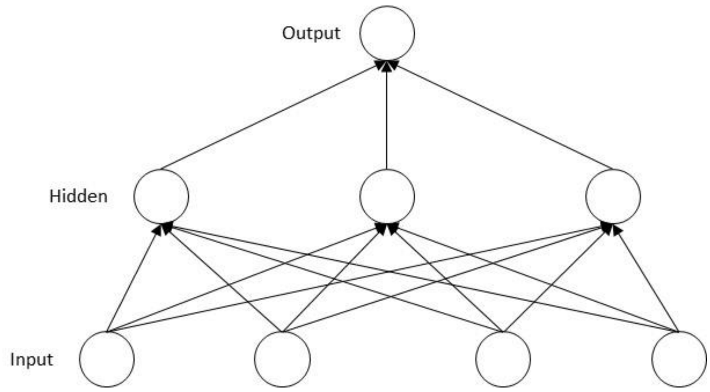

Figure 9 – Generic schema of an MLP network (Desai et al., 1996)

The ANN will be implemented in SAS Enterprise Miner using the Neural Network node, where the multilayer perceptron architecture will be used; and the AutoNeural, which assesses different network configurations and selects the most appropriate to capture the relationship between the dataset and target (SAS, 2021f; Zhao, 2018).

3.2.1.2. Random Forest

As previously mentioned, random forests are tree-based ensemble algorithms and will be used, in this study, both boosting and bagging algorithms. Boosting algorithms, of which Adaboost is a particular model, combine multiple weak learners into one giving more weight to the worst models. Bagging models use bootstrap aggregations (bagging), where trees are trained independently on bootstrap samples (sampling with replacement) of the same size as the training data and average the individual predictions (Larkin & McManus, 2018; Mishina et al., 2014; SAS, 2021b).

The random forest models will be implemented in SAS Enterprise Miner. The boosting algorithm will be implemented using the Gradient Boosting node, where the weak learners are aggregated using gradient descent, a first-order iterative optimization algorithm to find a loss function (Friedman, 2001, 2002; SAS, 2021a). The bagging algorithm will be implemented by using the HP Forest node (Nord & Keeley, 2016; SAS, 2021b).

3.3. PERFORMANCE ASSESSMENT

As mentioned in 2. Literature Review, according to Grebenar (2018) and Lessmann et al. (2015), the performance assessment should use as performance accuracy measurements the AUC, the Gain or Lift, and the Kolmogorov-Smirnov.

As in the previous chapters, these will be implemented using SAS Enterprise Miner, throughout the following approaches, segregated per each of the measures.

3.3.1. Area Under the Curve

The AUC measures the percentage of results which are the ROC curve – which yields a graphical plot of the percentage of 'bads' rejected (True Positive Rate - TPR) versus the percentage of 'goods' rejected

(False Positive Rate - FPR), and is as good as closer to one (Brownlee, 2018; Lessmann et al., 2015; Narkhede, 2018; Vidal & Barbon, 2019). For this exercise, the percentage of "bads" rejected (TPR) are clients that prepaid and the model identified these clients as having a prepayment; and the percentage of "goods" rejected (FPR) are the clients that did not prepay, although the model identified these clients as having a prepayment.

Mandrekar defines the following thresholds for the AUC (Mandrekar, 2010):

AUC Value	Discriminatory Ability
≤ 0.5	› No discrimination ability
]0.5; 0.7]	› Nonacceptable results
]0.7; 0.8]	› Acceptable results
]0.8; 0.9]	› Excellent results
> 0.9	› Outstanding results

Table 13 – AUC thresholds

This method will be computed using the Model Comparison node, which generates the ROC curve and corresponding AUC value. This approach allows for a generation of a plot depicting the curve and the value of the index, for as many models as are being tested. Below is shown an example of the chart generated by this operation:

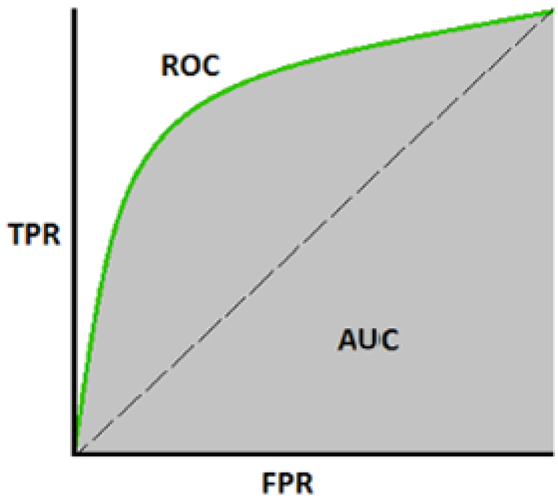

Figure 10 – Depiction of ROC curve and respective AUC (Narkhede, 2018)

3.3.2. Gain or Lift

The gains or lift chart measures the effectiveness of the model, calculated as the ratio between the results obtained using the model, and without it.

The gain chart table divides the range into deciles. A model with no predictive power would be expected to predict around 10% of events in each of 10 deciles. A good discriminating model would predict a higher number of events in the top decile, from which the response will decline monotonically (SAS, 2021e).

This will be computed using the Model Comparison Node of SAS Enterprise Miner, which calculates the cumulative lift using the following formula (SAS, 2021e):

$$\begin{aligned}&Cumulative\ Lift \\ &= cumulative\ ratio\ of\ \%\ Captured\ Response\ within\ decile\ to\ the\ baseline\ \%\ response\end{aligned} \tag{5}$$

This generates a chart similar to the one below, where the better models will have the higher curve, as the chart displays, representing the advantage in using the predictive model compared to a naïve model (i.e. a model at random, whose cumulative lift is 1). The Decision Tree model captures the event 3.5 times better than the random model in the example below.

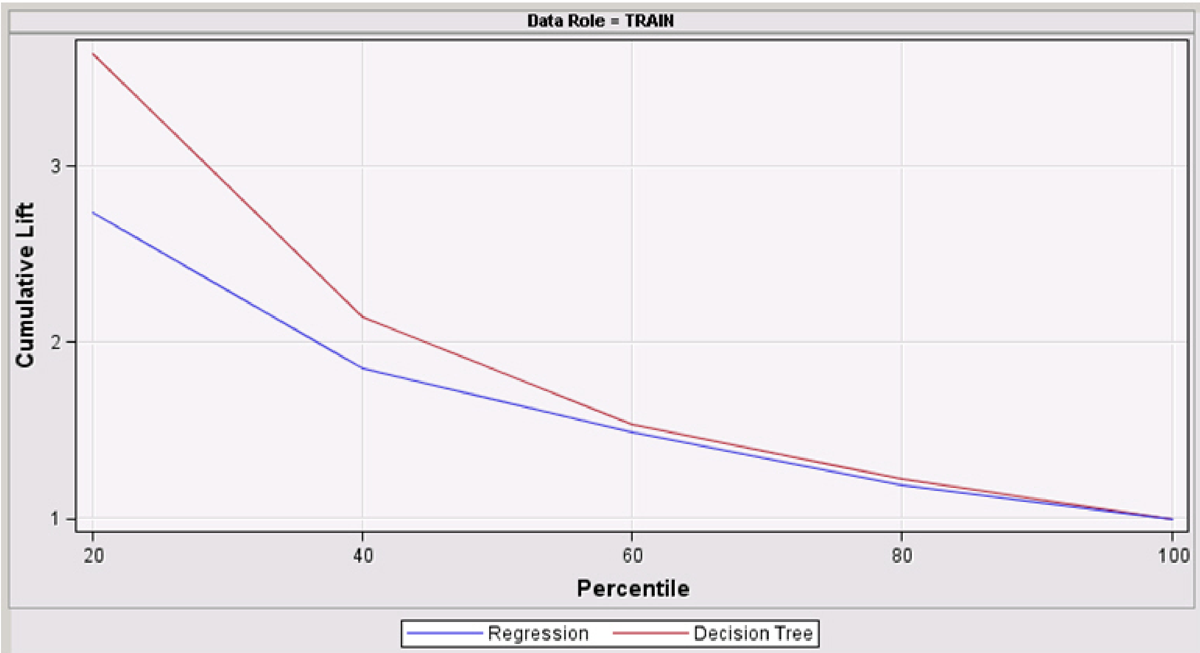

Figure 11 – Cumulative Lift chart example. Source: (SAS, 2021e)

3.3.3. Kolmogorov-Smirnov

The Kolmogorov-Smirnov (KS) measures the degree of separation between positive and negative distributions. A value below 20% indicates a questionable model, whereas above 70% means it is probably a 'too good to be true' model (Anderson, 2007).

The formula is given by:

$$D_{KS} = Max\left\{F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)\right\} \tag{6}$$

where $F(Y_i)$ is the theoretical cumulative distribution of the distribution being tested, i is the point in analysis, and N is the sample size.

This will be computed using the Model Comparison node of SAS Enterprise Miner, which calculates the KS.

4. RESULTS AND DISCUSSION

For the modelling of prepayments on mortgage loans, based on machine learning approaches, as described in the previous chapters, two models will be carried out: one model aims at modelling full repayment (when the customer amortises the entire outstanding balance, i.e. settles its debt), and the other aims at modelling partial repayments (when the customer amortises only part of the outstanding balance, being higher than the contracted amount of the scheduled amortisation).

The chapter is, thus, divided into an analysis of the bank's data and its relationship with the history of pre-payments in the financial system in Portugal, an analysis of its comparability, and the incidence of the targets on the dataset. After this preliminary analysis of the dataset considered in the study, the model results are presented, and, finally, additional and comparative analyses of the model results obtained are presented.

4.1. History of Prepayments: Comparison between the bank and the financial system

As mentioned in chapter 2.1.2.1, Banco de Portugal publishes a yearly "Report on Monitoring of Retail Banking Markets" where, among others, shares the data on the amount and number of prepayments in Portuguese banks. The figure below reflects a comparison of the behaviour between the financial system (equivalent to Figure 4), shown in the left chart, and the bank's data, shown in the right chart.

Here, it can be seen that there is a comparable behaviour between the Bank and the Financial System, with an inversion on the tendency of the number of prepayments, with partial being the majority before 2015, with around 66% in 2012 in the Bank and 54% in the financial system. Furthermore, the prepayment amount also demonstrates a similar tendency, with a decrease after the crisis and until 2013, with a sharp increase of around 52% in the Bank and 45% in the financial system.

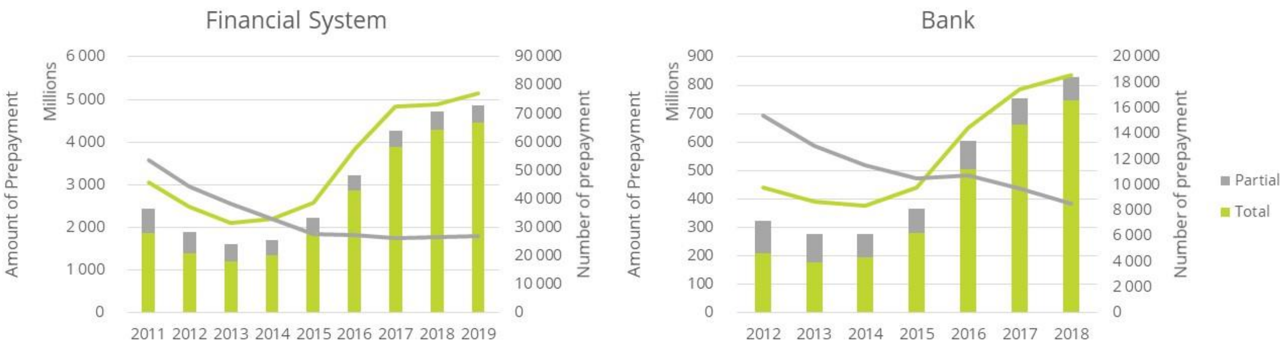

Figure 12 – Amount of prepayment for total and partial prepayment (bar chart) and number of total and partial prepayment (line chart), comparison between the financial system and Bank. Source: Report on Monitoring of Retail Banking Markets from Banco de Portugal, data aggregated by the author, and Bank's internal data (Banco de Portugal, 2019).

It must also be mentioned that the individual analysis per Bank is impacted not only by macroeconomic factors, but also by its own commercial activity and market competition. This materialises, for example, in the Bank having a pro-active retention policy or repurchasing of credit, this may impact how the Bank interacts with its clients and, as thus, the prepayments.

These prepayments, being full or partial, are rare events in the Bank, with the percentage between observations and observations with positive events shown in the table below:

Target	Total of records	Number of prepayments	Percentage of positive cases
Full repayment	4 055 416	86 887	2.11%
Partial repayment	4 055 416	54 369	1.32%

Table 14 – Percentage of the targets in the dataset

4.2. Modelling Prepayment

As stated, two sets of models will be computed under two different targets: artificial neural networks and random forests for full and partial prepayments.

The two targets will be computed, and explained, separately, given that they typically have distinct causes. Full repayments are typically associated with phenomena that are more difficult to predict, such as a change of bank (with the total transfer of the residual amount to the other bank), a change of job or divorce (because they may imply a change in the mortgagors' home), and, more generally, the purchase of a new house (with the payment of the current loan and the opening of a new one). Partial repayments are phenomena that may be more associated with the behaviour of mortgagors and their financial situation.

As detailed in the previous chapters, the two sets of models will be artificial neural networks and random forests, with different parameter combinations being trained, to test which combination provides the best fit to the data and gives the best performance. Thus, the results are shown for the following models:

- **Artificial neural networks**:
  - With multilayer perceptron architecture, are tested models without variable selection with 3, 5 and 8 hidden units, with variable selection through stepwise regression with 3 hidden units, and variable selection through chapter 3.1.3 with 3 hidden units.
  - With autoneural node, i.e. with the model selecting the architecture to be used, are tested models without variable selection with 3 and 8 hidden units, with variable selection through stepwise regression with 3 hidden units, and variable selection through chapter 3.1.3 with 3 hidden units.
- **Random forest**:
  - With the bagging approach, are tested models with a maximum depth of 50 splitting rules and 30 trees, without variable selection, with variable selection through stepwise regression, and variable selection through chapter 3.1.3 with 3 hidden units.
  - With the boosting approach, are tested models with 70 iterations without variable selection, with variable selection through stepwise regression, and variable selection through chapter 3.1.3 with 3 hidden units.

4.2.1. Full Prepayment

The fifteen typologies of models, described above, are applied to the data after the pre-processing chapter 3.1, in short, noise reduction, outliers, missing values and aggregation of categories, and after the dataset is split into training, validation and test set in a ratio of 40-30-30. The training set will be used to, as the name implies, train the model, the validation set will be used to select the model, and the test set will be used to assess the model performance, being an unbiased sample as the model has never been in contact with this data. Lastly, and as described previously, the models will be assessed through the AUC, the cumulative lift and the Kolmogorov-Smirnov statistics.

The table below summarizes these results for both the train and test set, highlighting the model that performs best in each metric in green.

	Train			Test		
Methodology	**AUC**	**Cum. Lift**	**Kolmogorov -Smirnov**	**AUC**	**Cum. Lift**	**Kolmogorov -Smirnov**
ARTIFICIAL NEURAL NETWORK						
MLP - with stepwise - 3 hidden units	0.80	4.19	0.43	0.80	4.16	0.43
MLP - with LASSO - 3 hidden units	0.76	3.45	0.38	0.76	3.48	0.38
MLP - no variable selection - 8 hidden units	0.78	3.72	0.40	0.77	3.72	0.39
MLP - no variable selection - 5 hidden units	0.76	3.45	0.37	0.76	3.44	0.38
MLP - no variable selection - 3 hidden units	0.76	3.36	0.37	0.76	3.36	0.37
Auto Neural - with stepwise - 3 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
Auto Neural - with LASSO - 3 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
Auto Neural - no variable selection - 3 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
Auto Neural - no variable selection - 8 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
RANDOM FOREST						
Bagging - with stepwise	0.86	5.48	0.53	0.82	4.68	0.46
Bagging - selection with LASSO	0.86	5.61	0.54	0.83	4.81	0.47
Bagging - no variable selection	0.86	5.50	0.52	0.82	4.72	0.46
Gradient Boosting - with stepwise	0.50	1.00	0.00	0.50	1.00	0.00
Gradient Boosting - selection with LASSO	0.50	1.00	0.00	0.50	1.00	0.00
Gradient Boosting - no variable selection	0.50	1.00	0.00	0.50	1.00	0.00

Table 15 - Results of the performance assessment metrics for the full prepayment, highlighted the best model for the metric

A graphical depiction of the AUC – the ROC curve, is shown below:

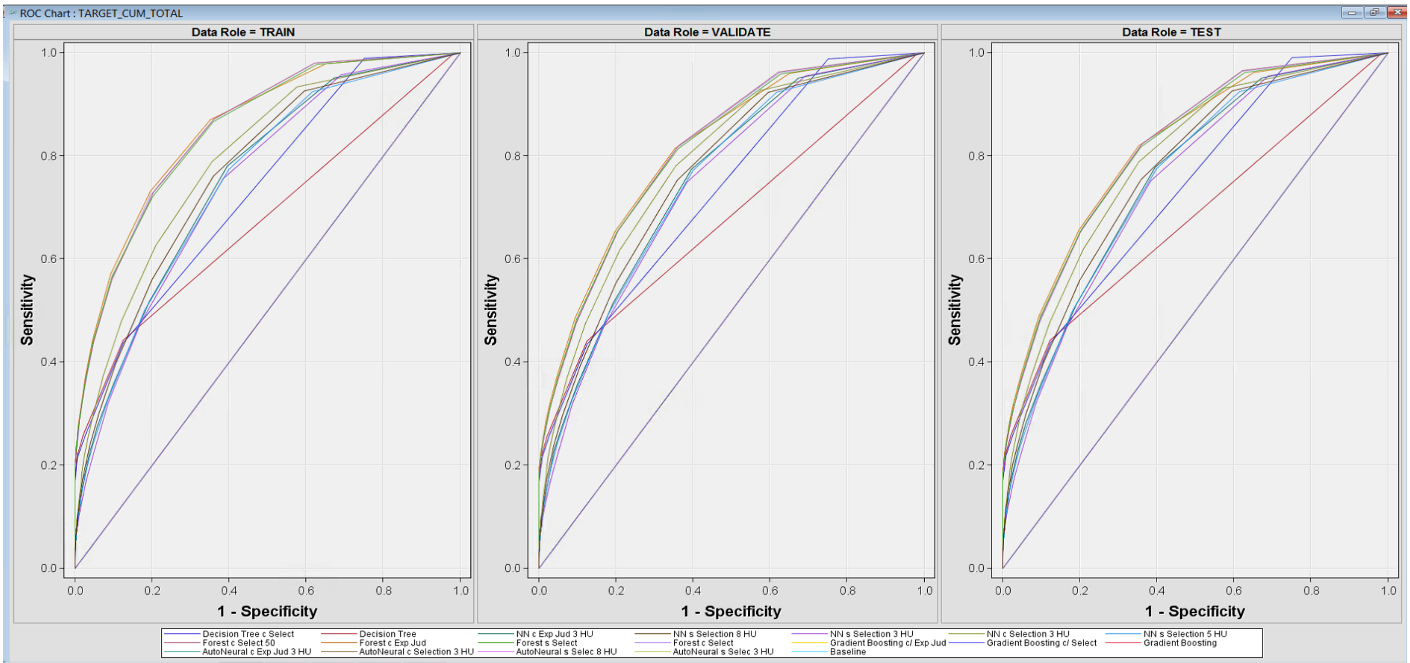

Figure 13 – ROC chart for the full prepayment models

And a graphical depiction of the cumulative lift is shown below:

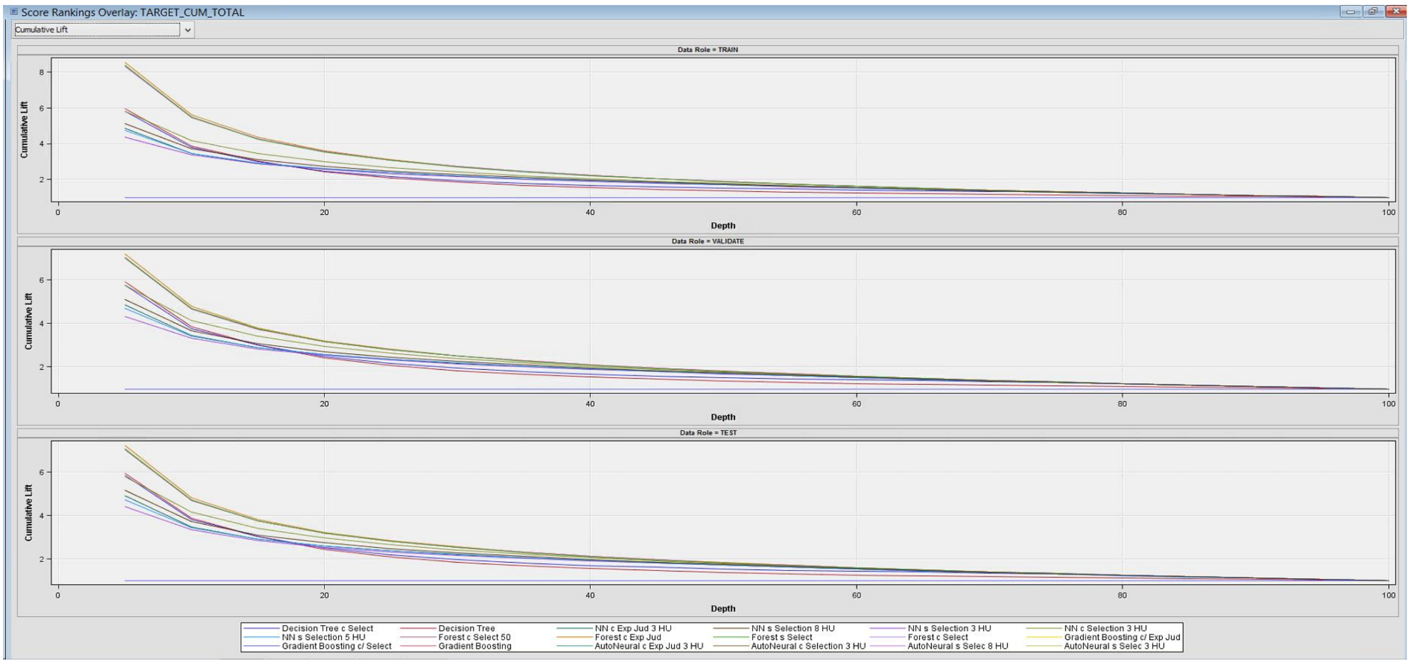

Figure 14 – Cumulative lift chart for the full prepayment models

These results show that the best model for the full prepayment target is the random forest, using the bagging approach, i.e. HP Forest node, with variable selection through LASSO regression. This model achieves a ROC of 0.86 in the training data set and 0.83 in the test data set. These results show that the tree models perform better than the ANN models, whose performance, measured through the ROC, is higher than 0.75 but less than 0.80. However, in the analysis of the performance deterioration between the training and test set, the ANN models show a stabler model, with a decrease of around 0.01 (with the overall best model, in the random forest, with a decrease of 0.04). In the cumulative lift and KS, the performance is also superior in this model, with a more significant difference to the other models, as can be depicted in the values of the table, and graphically, with the two lines referring to the boosting models separated from the rest.

It shall also be noted that the models using the autoneural and gradient boosting node do not achieve acceptable results, with random or naïve models.

4.2.2. Partial Prepayment

The fifteen typologies of models, described above, are applied to the data after the pre-processing chapter 3.1, in short, noise reduction, outliers, missing values and aggregation of categories, and after the dataset is split into training, validation and test set in a ratio of 40-30-30. The training set will be used to, as the name implies, train the model, the validation set will be used to select the model, and the test set will be used to assess the model performance, being an unbiased sample as the model has never been in contact with this data. Lastly, and as described previously, the models will be assessed through the AUC, the cumulative lift and the Kolmogorov-Smirnov statistics.

The table below summarizes these results for both the train and test set, highlighting the model that performs best in each metric in green.

	Train			Test		
Methodology	**AUC**	**Cum. Lift**	**Kolmogorov -Smirnov**	**AUC**	**Cum. Lift**	**Kolmogorov -Smirnov**
ARTIFICIAL NEURAL NETWORK						
MLP - with stepwise - 3 hidden units	0.89	6.63	0.60	0.89	6.54	0.59
MLP - with LASSO - 3 hidden units	0.89	6.62	0.60	0.88	6.56	0.59
MLP - no variable selection - 8 hidden units	0.88	6.32	0.58	0.88	6.26	0.57
MLP - no variable selection - 5 hidden units	0.87	6.28	0.57	0.87	6.24	0.56
MLP - no variable selection - 3 hidden units	0.89	6.61	0.60	0.88	6.55	0.59
Auto Neural - with stepwise - 3 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
Auto Neural - with LASSO - 3 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
Auto Neural - no variable selection - 3 hidden units	0.87	6.21	0.57	0.87	6.17	0.56
Auto Neural - no variable selection - 8 hidden units	0.50	1.00	0.00	0.50	1.00	0.00
RANDOM FOREST						
Bagging - with stepwise	0.92	7.44	0.68	0.90	6.72	0.60
Bagging - selection with LASSO	0.93	7.60	0.69	0.90	6.75	0.61

	Train			Test		
Methodology	**AUC**	**Cum. Lift**	**Kolmogorov -Smirnov**	**AUC**	**Cum. Lift**	**Kolmogorov -Smirnov**
Bagging - no variable selection	0.93	7.56	0.69	0.89	6.59	0.60
Gradient Boosting - with stepwise	0.50	1.00	0.00	0.50	1.00	0.00
Gradient Boosting - selection with LASSO	0.50	1.00	0.00	0.50	1.00	0.00
Gradient Boosting - no variable selection	0.50	1.00	0.00	0.50	1.00	0.00

Table 16 - Results of the performance assessment metrics for the partial prepayment, highlighted the best model for the metric

A graphical depiction of the AUC – the ROC curve, is shown below:

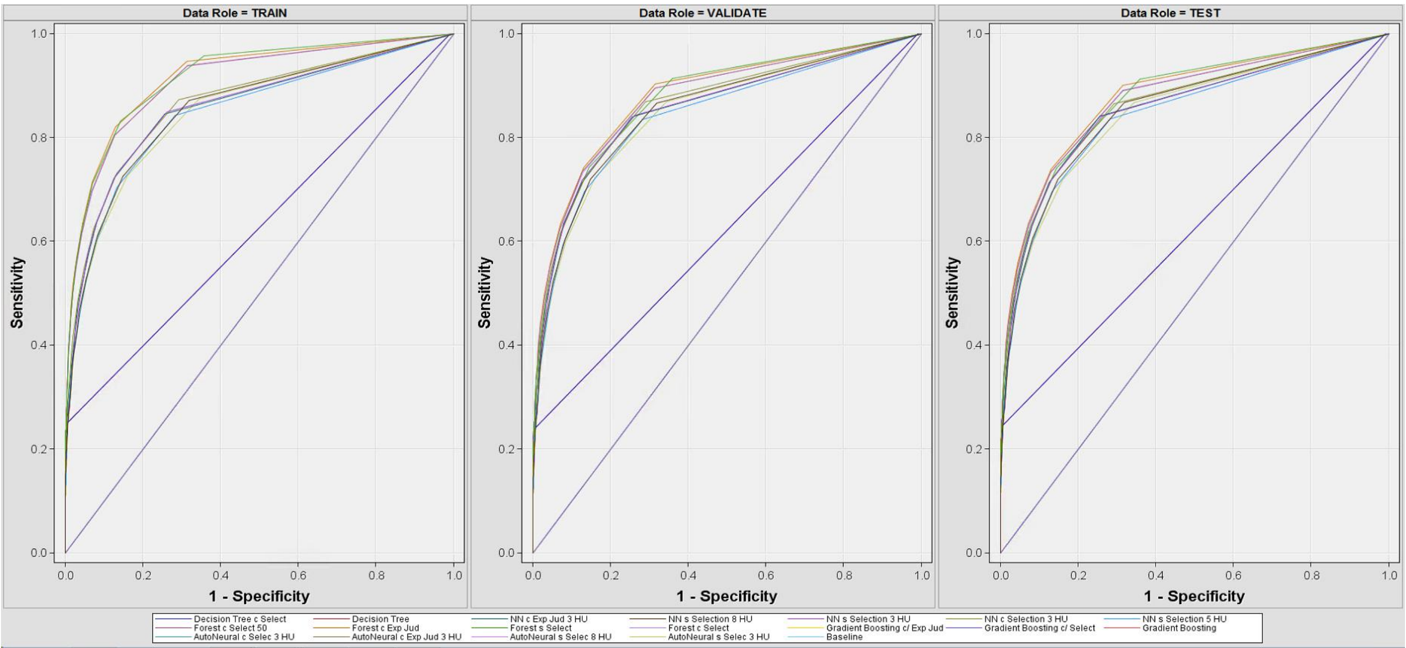

Figure 15 – ROC chart for the partial prepayment models

And a graphical depiction of the cumulative lift is shown below:

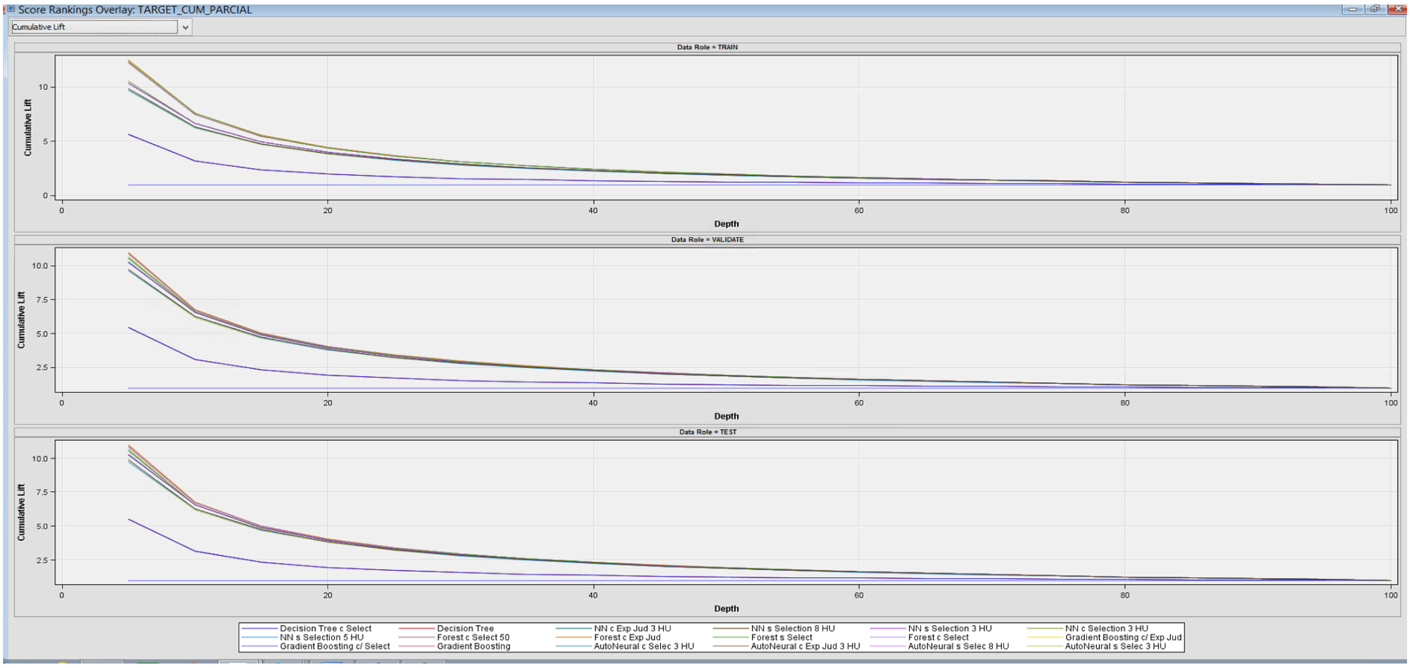

Figure 16 – Cumulative lift chart for the partial prepayment models

These results show that the best model for the partial prepayment target is also the random forest, using the bagging approach, i.e. node HP Forest, with variable selection through LASSO regression. This model achieves a ROC of 0.93 in the training data set and 0.90 in the test data set. These results show that the tree models perform better than the ANN models, whose performance, measured through the ROC, is around 0.88. However, in the analysis of the performance deterioration between the training and test set, the ANN models show a stable model, not showing significant decreases (with the overall best model, in the random forest, with a decrease of 0.03). In the cumulative lift and KS, the performance is also superior in this model.

It shall also be noted that the models using the autoneural (except for the autoneural without variable selection and three hidden units) and gradient boosting node do not achieve acceptable results, with random or naïve models.

4.3. Further Discussion

After obtaining models that meet the thresholds for performance and demonstrate predictive power, additional analyses were carried out, allowing further insights into the dataset and leveraging the machine learning models as a tool for better knowledge of the data and consequential phenomena present.

This first analysis compares the variables selected in the full and partial prepayment model, to analyse the differences between the two events and gain more knowledge of them. The second analysis involved comparing the models obtained from two different perspectives by analysing the data before the prepayment, knowing that it had occurred, and after.

4.3.1. Comparison between full and partial prepayment

A macro conclusion arises from the performance analysis of the two models: as expected, the partial prepayment model performs better than the full prepayment model. This can be explained, as indicated above, by the events that lead to full amortizations (e.g., change of bank, job displacement or divorce). However, to obtain a less empirical analysis of this phenomenon, this study analysed the most used variables in the models. This analysis, followed by averaging of parameter estimates is part of Bayesian model averaging methods, which can be further research, as noted in chapter 5. Conclusions.

This analysis is possible as both models with better performance are random forest models, where it is possible to analyse the number of splitting rules in which the variable participates.

As such, the tables below show the variables considered in each model, with the number of splitting rules per variable, ordered from highest (variables with greater distinctive power) to the lowest.

	# Splitting rules
Variable	**Full**
MONTANTE_RESIDUAL	5664
TOTAL_MONTANTE_AMORT	4653
PRAZO	4368
PERC_PRAZO	4093
N_PREST_PAGAS	4035
PRAZO_RESIDUAL	3921
LTV_ATUAL	3767
DATA_ABERTURA	3674
T_JURO	3584
FINALIDADE	3567
LTV_ORIG	3507
IND_CREDITO	3157
IDADE	2985
M_PRS_MENS_BANK	2963
MONTANTE_FINANCIADO	2887
PROFISSAO	2886
M_PRS_MENS_banca	2773
SALDO_DO_06M	2675
RENDIMENTO	2612
T_SPREAD	2590
RESP_BANK_REAIS	2553
SALDO_DO_12M	2287
RESP_BANCA_REAIS	2214

	# Splitting rules
Variable	Full
scoring	2110
N_DIAS_ATRASO	1974
TX_ESFORCO_BANCA	1902
SALDO_DP_06M	1771
SALDO_DP_12M	1714
ESTADO_CIVIL	1587
RESP_BANCA_POT	1540
TAXA_JURO_DP_TVH	1480
TX_DIVORCIO_TVH	1462
ANO	1372
tot_devedores_banca	1356
TOTAL_AMORT_PARCIAL	1335
TAXA_INFLACAO_TVH	1323
IND_SENT_ECO_TVH	1312
n_produtos_banca	1274
N_OPER_BANCA_REAIS	1234
N_PRODUTOS_BANK	1195
ENDIV_PART_TVH	1135
N_OPER_BANCA_POT	1069
N_OPER_BANK_REAIS	1020
Perc_utiliza	868
N_OPER_BANK_POT	791
INIB_CHEQUE	75

Table 17 – Variables considered in the full prepayment model

	# Splitting rules
Variable	Partial
MONTANTE_FINANCIADO	5635
LTV_ATUAL	4336
LTV_ORIG	3856
MONTANTE_RESIDUAL	3794
M_PRS_MENS_BANK	3647

	# Splitting rules
Variable	Partial
N_PREST_PAGAS	3512
DATA_ABERTURA	3353
PERC_PRAZO	3193
TOTAL_MONTANTE_AMORT	3030
PRAZO	2621
SALDO_DP_06M	2564
scoring	2500
FINALIDADE	2481
N_OPER_BANCA_REAIS	2446
SALDO_DO_06M	2278
RESP_BANK_REAIS	2233
Z_FIM_CTTO	2207
SALDO_DO_12M	1990
PROFISSAO	1984
RENDIMENTO	1898
TOTAL_AMORT_PARCIAL	1724
T_SPREAD	1703
TX_ESFORCO_BANCA	1687
tot_devedores_banca	1651
IND_CREDITO	1560
T_JURO	1551
RESP_BANK_POT	1493
ESTADO_CIVIL	1380
RESP_BANCA_POT	1284
N_PRODUTOS_BANK	1196
PIB	1118
TX_DESEMPREGO_TVH	1108
ANO	981
IND_PRECOS_HAB_TVH	972
ED_LICENC_TVH	914
N_OPER_BANK_REAIS	867
Perc_utiliza	837

	# Splitting rules
Variable	Partial
N_OPER_BANK_POT	620

Table 18 – Variables considered in the partial prepayment model

These tables show that the models consider a similar number of variables (the full model considers 46 variables, whereas the partial considers 38), with 32 variables in common. As such, there is a similarity between the variables present in both models. As stated in Table 19, the major differences are related to the use of additional macroeconomic variables and more variables regarding the client's behaviour in the bank and financial system in the full prepayment model.

Variable Category	# Full	# Partial
Loan characteristics	14	14
Client	5	4
Behaviour in Bank and Financial System	21	15
Macroeconomy	5	4
Point in time	1	1

Table 19 – Analysis of variables category between the full and partial models

In particular, the full prepayment model considers as additional variables (not considered in the partial prepayment) the residual term of the loan, the client's age, the amount of monthly instalments in the financial system, the real liabilities in the financial system, the number of days past due, the balance in term deposits, 12 months, the number of financial products in the financial system, the number of potential operations in the financial system and the check inhibition indicators and, as macroeconomic variables the interest rate in term deposits, the divorce rate, the inflation rate, the economic sentiment indicator and the indebtedness of families.

The partial prepayment model considers as additional variables (not considered in the full prepayment) the date of contract ending, the potential liabilities in the bank and, as macroeconomic variables, the GDP, the unemployment rate, the housing price index and the number of licensed buildings.

4.3.2. Comparison between model and profiling

The models built were based, as previously mentioned, on annual observations, which are positioned in January of each year, the target being an indicator of whether there was a prepayment (total or partial) in that year. It allows the analysis of the customer's conditions at the beginning of the year, before the event.

The results obtained (and mostly the variables used in the model) will be compared with an auxiliary model that was built, which is based on annual observations positioned in December, i.e., it allows for the analysis of customer conditions after the prepayment event.

This analysis allows a comparison between a predictive model and a profiling model and provides a set of insights that can be further explored, regarding which variables diverge the most after the prepayment event.

This analysis shows that the profiling models utilize a more reduced number of variables, in particular in the macroeconomic variables.

	Predictive		Profilling	
Variable	**Full**	**Partial**	**Full**	**Partial**
Loan characteristics	14	14	10	9
Client	5	4	3	4
Behaviour in Bank and Financial System	21	15	15	12
Macroeconomy	5	4	1	2

Table 20 - Analysis of variables category between the full and partial models

5. CONCLUSIONS

The study's primary purpose was to model prepayment events in a large Portuguese bank using machine learning models (in particular random forest and artificial neural network) as the studies in both the Portuguese market and through the use of machine learning models were scarce.

The results obtained reveal that both the total and partial prepayment models perform well. For this analysis, three distinct performance metrics were used - AUC, cumulative lift and Kolmogorov-Smirnov statistics. The model with the best performance to model full prepayments obtained a ROC of 0.83, with an excellent discriminatory ability as per Mandrekar thresholds, with a cumulative lift of 4.81, well above 1 (the naïve, or random model), and KS of 0.47, being the questionable limit of 0.20, in contrast, a KS of 0.70 being a model with results too good to be true. The model with the best performance to model partial prepayments obtained a ROC of 0.90, with an excellent discriminatory ability as per Mandrekar thresholds, with a cumulative lift of 6.75, well above 1 (the naïve, or random model), and KS of 0.61, being the questionable limit of 0.20, in contrast, a KS of 0.70 being a model with results too good to be true. (Anderson, 2007; Mandrekar, 2010)

The three metrics analysed allow for three distinct conclusions to be inferred:

- Both models present positive results and demonstrate that the model has good predictive results and discriminatory capacity;
- The partial repayment model is superior to the full repayment model, with a difference which, although not very large, is worthy of mention;
- Finally, the best models are the best in all metrics; there is consistency in the metrics when selecting the best model.

The analysis of the most relevant variables in the models, possible by the use of random forest models, allows for the analysis in two dimensions:

i. **Comparison between full and partial prepayment**: the models consider a similar number of variables (the full model considers 46 variables, whereas the partial considers 38), with 32 variables in common. As such, there is a similarity between the variables present in both models. The major differences relate to the use of additional macroeconomic variables and more variables in the behaviour in the financial system and bank in the full prepayment model.
ii. **Comparison between model and profiling**: the profiling models utilize a lower number of variables, in particular, less macroeconomic variables.

The models obtained, which use machine learning models with a more opaque nature, were compared, in meetings with the Bank, with the models that the modelling team was developing, which are decision trees models. The models presented in this study present a superior performance. However, it should be mentioned that they use a larger number of variables and, given their nature, present restrictions regarding usability in a banking context due to their reduced transparency, which will be described in the following chapter.

5.1. Limitations and Recommendations for Future Work

The approach followed in this model, being machine learning models, entails the inherent limitations, emphasising that this model applies to the financial system, which is subject to intense regulation. The World Bank and EBA released papers on the usability of machine learning, where they focus on the main limitations on the usage of machine learning in the financial system: (European Banking Authority, 2020; The World Bank Group, 2019)

- **Explanation and interpretability**: a model is considered explainable if humans can understand how a result is reached, on what grounds that result is based, or what justifies the result. This implies not only an explanation and interpretation of the results but also transparency on the processes inherent to the data, processing, algorithms and training methods.

- **Traceability and auditability**: a model is considered traceable and auditable if every step and criteria can be traced throughout the modelling process, allowing its replication by third parties and oversight.

- **Bias**: a model may inadvertently make biased decisions, which discriminate against a group of clients. This may occur due to the data used, selection bias, and with the propagation of historical social bias, for example, when a class is less represented in the training set, the model will learn from few examples and will not be able to generalize correctly.

Even though the selected models allow for an analysis of the variables used and their importance, they do not allow for an analysis of the rules inherent to the decision. This results in a reduction of transparency, interpretability and auditability by the regulators, which would be even worse had the best model be of the artificial neural network family, where there is an added layer of complexity. Furthermore, given the use of historical data, biases may be inherent, however, their scope is reduced by the available variables (i.e. the dataset does not have gender, race, region or belief variables). In addition, for the models to be implemented in the financial system, the training dataset must be tested against different timeframes, to ensure its representativeness.

Thus, given these limitations in machine learning models in the financial system context, the suggested future work concerns deepening the comparative analysis between profiling and prediction models to extract insights from the machine learning models. For example, it can be further studied the variables considered in each of the models, their individual impact on model performance. These analyses can then be leveraged into the models currently in force, which meet the regulatory requirements.

In addition, and as referred above, the comparison between the profiling and prediction models, followed by averaging the parameter estimates, being part of Bayesian model averaging methods could give further insights on the data.

6. BIBLIOGRAPHY

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, *35*(3), 1275–1292. https://doi.org/10.1016/j.eswa.2007.08.030

Allison, P. D., & Christakis, N. A. (2017). *Fixed-Effects Methods for the Analysis of Nonrepeated Events*.

Altman, E., Marco, G., & Varetto, F. (1994). Corporate Distress Diagnosis: Comparisons using linear discriminant analysis and neural networks (The Italian Experience). *Journal of Banking and Finance*, *18*, 505–529.

Anderson, R. (2007). *The Credit Scoring Toolkit*. (O. U. Press, Ed.).

Ashofteh, A. & Bravo, J. M. (2019). A non-parametric based computationally efficient approach for credit scoring. Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao 2019 [CAPSI 2019 - 19th Conference of the Portuguese Association for Information Systems, Proceedings. 4]. https://aisel.aisnet.org/capsi2019/4.

Ashofteh, A., & Bravo, J. M. (2021a). A Conservative Approach for Online Credit Scoring. Expert Systems With Applications, Volume 176, p. 1-16, 114835. https://doi.org/10.1016/j.eswa.2021.114835

Ashofteh, A. & Bravo, J. M. (2021b). Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code]. https://doi.org/10.24433/CO.1963899.v1. Associated Publication: "A Conservative Approach for Online Credit Scoring", Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2021.114835

Assembleia Constituinte. Constituição da República Portuguesa (1976).

Assembleia da República. Decreto-Lei n.º 10-J/2020 (2020).

Baesens, B., Gestel, T. Van, Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring, *5682*. https://doi.org/10.1057/palgrave.jors.2601545

Banco de Portugal. (2019). *Relatório de Acompanhamento dos Mercados Bancários de Retalho*.

Banco de Portugal. (2021a). Como reembolsar e transferir. Retrieved May 25, 2021, from https://clientebancario.bportugal.pt/pt-pt/como-reembolsar-e-transferir

Banco de Portugal. (2021b). Empréstimos-Particulares-PRT-Consumo-M€ (OIFM). Retrieved August 1, 2021, from https://bpstat.bportugal.pt/serie/12557382

Banco de Portugal. (2021c). Empréstimos-Particulares-PRT-Habitação-M€ (OIFM). Retrieved August 1, 2021, from https://bpstat.bportugal.pt/serie/12557378

Banco de Portugal. (2021d). Empréstimos-Particulares-PRT-Outros fins-M€ (OIFM). Retrieved August 1, 2021, from https://bpstat.bportugal.pt/serie/12556761

Banco de Portugal. (2021e). Património dos particulares - Património financeiro líquido. Retrieved August 1, 2021, from https://bpstat.bportugal.pt/serie/12561030

Banco de Portugal. (2021f). Património dos particulares - Património não financeiro - Habitação. Retrieved August 1, 2021, from https://bpstat.bportugal.pt/serie/12561031

Basel Committee on Banking Supervision. (2000). *Principles for the Management of Credit Risk*. *Bank for International Settlements*. https://doi.org/10.1002/14651858.CD012104

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*(12), 1699–1707. https://doi.org/10.1057/jors.2008.130

Black, Fischer, and Myron Scholes (1973). The pricing of options and corporate liabilities. Journal of Political Economy 81: 637–54.

Bohn, A., Plantefève, O., Poppensieker, T., & Schneider, S. (2020). *How banks can ease the pain of negative interest rates*.

Borovkova, S. (2017). *A note on prepayment modelling for residential mortgages*.

Bravo, J. M. (2020). Longevity-Linked Life Annuities: A Bayesian Model Ensemble Pricing Approach. CAPSI 2020 Proceedings, 29. https://aisel.aisnet.org/capsi2020/29.

Bravo, J. M. (2021). Pricing participating longevity-linked life annuities: A Bayesian Model Ensemble approach. European Actuarial Journal. https://doi.org/10.1007/s13385-021-00279-w

Bravo, J. M., Ayuso, M. (2020). Mortality and life expectancy forecasts using bayesian model combinations: An application to the portuguese population. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, E40, 128–144. https://doi.org/10.17013/risti.40.128–145.

Bravo, J. M., Ayuso, M. (2021). Forecasting the retirement age: A Bayesian Model Ensemble Approach. Advances in Intelligent Systems and Computing, Volume 1365 AIST, 123–135 [2021 World Conference on Information Systems and Technologies, WorldCIST 2021] Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_12.

Bravo, J. M., Ayuso, M., & Holzmann, R., (2019). Making use of Home Equity: The Potential of Housing Wealth to Enhance Retirement Security. IZA DP Series No. 12656, September, IZA Institute of Labour Economics, Germany. Available at https://ideas.repec.org/p/iza/izadps/dp12656.html

Bravo, J. M., Ayuso, M., Holzmann, R., Palmer, E. (2021). Addressing the Life Expectancy Gap in Pension Policy. Insurance: Mathematics and Economics, 99, 200-221. https://doi.org/10.1016/j.insmatheco.2021.03.025.

Breiman, L. (2001). Random Forests. *Kluwer Academic Publishers*, 5–32.

Brownlee, J. (2018). A Gentle Introduction to Probability Scoring Methods in Python. Retrieved August 17, 2020, from https://machinelearningmastery.com/how-to-score-probability-predictions-in-python/#:~:text=with sample code).-,Brier Score,error in the probability forecasts.&text=The Brier score can be,() function in scikit-learn.

Brownlee, J. (2019). How to Choose a Feature Selection Method For Machine Learning. Retrieved November 15, 2020, from https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. Risk Management, 18(4), 264–287. https://doi.org/10.1057/s41283-016-0006-4

Chamboko, R., & Bravo, J. M. (2019a). Frailty correlated default on retail consumer loans in Zimbabwe. International Journal of Applied Decision Sciences, 12(3), 257–270.

https://doi.org/10.1504/IJADS.2019.100436

Chamboko, R., & Bravo, J. M. V. (2019b). Modelling and forecasting recurrent recovery events on consumer loans. International Journal of Applied Decision Sciences, 12(3), 271–287. https://doi.org/10.1504/IJADS.2019.100440

Chamboko, R., & Bravo, J. M. (2020). A Multi-State Approach to Modelling Intermediate Events and Multiple Mortgage Loan Outcomes. Risks, 8(2), 64. https://doi.org/10.3390/risks8020064Charlier, E., & van Bussel, A. (2001). *Prepayment Behavior of Dutch Mortgagors: An Empirical Analysis*.

Chatterjee, S., & Barcun, S. (1970). A Nonparametric Approach to Credit Screening. *Journal of the American Statistical Association*, *65*(329), 150–154.

Chen, Shunqin, Guo, Z., & Zhao, X. (2021). Predicting mortgage early delinquency with machine learning methods. *European Journal of Operational Research*, *290*(1), 358–372. https://doi.org/10.1016/j.ejor.2020.07.058

Chen, Shuo, & Bowman, F. D. (2011). A Novel Support Vector Classifier for Longitudinal Highdimensional Data and Its Application to Neuroimaging Data. *Statistical Analysis and Data Mining*, *4*(6), 604–611. https://doi.org/10.1002/sam.10141.A

Chen, Shuo, Grant, E., Wu, T. T., & Bowman, F. D. (2014). Some recent statistical learning methods for longitudinal high-dimensional data. *WIREs Computational Statistics*, *6*(1), 10–18. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4181610/

Coffman, Y. C. (1986). The Proper Role of Tree Analysis in Forecasting the Risk Behavior of Borrower. *Management Decision Systems*.

Conselho de Ministros. Resolução do Conselho de Ministros, nº67, 1ª série (1976).

CTT. (2021). Base de Dados Códigos Postais Portugal. Retrieved April 3, 2021, from https://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring : A systematic literature survey. *Applied Soft Computing Journal*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Deloitte. (2019). How to Improve Prepayment Modelling - Exploring the Added Value of Machine Learning. Retrieved January 14, 2021, from https://www2.deloitte.com/nl/nl/pages/risk/articles/how-to-improve-prepayment-modelling.html

Desai, V. S., Crook, J. N., & Overstreet, G. A. J. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *2217*(95).

Dickinson, A., & Heuson, A. J. (1994). Mortgage Prepayments: Past and Present. *Journal of Real Estate Literature*, *2*(1), 11–33.

Eddy, Y. L., & Muhammad, E. N. B. A. (2017). Credit Scoring Models: Techniques and Issues. *Journal of Advanced Research in Business and Management Studies*, *7*(2), 29–41.

Elsinga, M., & Hoekstra, J. (2005). Home Ownership and Housing Satisfaction. *Journal of Housing and the Built Environment*, *20*(4), 401–424.

Emmons, W. R. (2008). The Past, Present and Future of the U.S. Mortgage Market. Retrieved June 4, 2021, from https://www.stlouisfed.org/publications/central-banker/summer-2008/the-past-present-and-future-of-the-us-mortgage-market

ECB (2016). The Household Finance and Consumption Survey: results from the second wave. ECB Statistics Paper No 18, December − Household Finance and Consumption Network.

European Banking Authority. (2019). *Risk Assessment of the European Banking System*. https://doi.org/10.2853/61391

European Banking Authority. (2020). EBA Report on Big Data and Advanced Analytics, (January), 1–60.

European Central Bank. (2004). *Financial Stability Review*.

European Central Bank. (2009). *ECB Monthly Bulletin*.

European Central Bank. (2020). *ECB Banking Supervision: Risk assessment for 2020*.

Fang, L., & Munneke, H. J. (2021). A spatial analysis of borrowers' mortgage termination decision – A nonparametric approach. *Regional Science and Urban Economics*, *86*(103595). https://doi.org/10.1016/j.regsciurbeco.2020.103595

Feature selection in machine learning. (2013). Retrieved November 15, 2020, from https://algorithmia.com/blog/feature-selection-in-machine-learning

Fractal Whitepaper. (2003). Comparative Analysis of Classification Techniques.

Freund, Y., & Schapire, R. E. (1996). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Gavrilova, Y., & Bolgurtseva, O. (2020). What Is Data Preprocessing in ML? Retrieved October 30, 2020, from https://serokell.io/blog/data-preprocessing

Gerardi, K., Herkenho, K., Ohanian, L. E. & Willen, P. (2013). Unemployment, Negative Equity and Strategic Default. Working Papers Series. Atlanta: Federal Reserve Bank of Atlanta.

Ghosh, D., & Vogt, A. (2012). Outliers: An Evaluation of Methodologies. In *Joint Statistical Meetings* (pp. 3455–3460).

Goodarzi, A., Kohavi, R., Harmon, R., & Senkut, A. (1998). *Loan Prepayment Modeling*.

Grebenar, T. (2018). *Behavioural Model of Assessment of Probability of Default and the Rating of Non-Financial Corporations*.

Green, J., & Shoven, J. B. (1986). The Effects of Interest Rates on Mortgage Prepayments. *Journal of Money, Credit and Banking*, *18*(1), 41–59.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Handhika, T., Fahrurozi, A., Zen, R. I. M., Lestari, D. P., Sari, I., & Murni. (2019). Modified Average of the Base-Level Models in the Hill-ClimbingBagged Ensemble Selection Algorithm for Credit Scoring. *Procedia Computer Science*, *157*, 229–237. https://doi.org/10.1016/j.procs.2019.08.162

Hao, K. (2018). What is machine learning? Retrieved August 24, 2020, from https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

Henley, W . E ., & Hand, D. . J. . (1996). A k-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk. *Journal of the Royal Statistical Society*, *45*(1), 77–95.

Henley, William Edward. (1995). Statistical aspects of credit scoring.

Ho, T. K. (1995). Random Decision Forests.

IBM Cloud Education. (2020). Machine Learning. Retrieved August 24, 2020, from https://www.ibm.com/cloud/learn/machine-learning

Instituto Nacional de Estatística. (2011). *Classificação Portuguesa das Profissões 2010*. *Lisboa: INE*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Classificação+Portuguesa+das+Profissões+2010#0

Jacobs, J. P. A. M., Koning, R. H., & Sterken, E. (2005). *Modelling Prepayment Risk*.

Jain, D. (2019). Data Preprocessing in Data Mining. Retrieved October 30, 2020, from https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Jing, B., Liu, H., & Liu, M. (2011). *Deep Modeling of Longitudinal Medical Data*.

Kau, J. B., Keenan, D. C., Muller, W. J., & Epperson, J. E. (1990). Pricing Commercial Mortgages and Their Mortgage-Backed Securities. *Journal of Real Estate Finance and Economics*, *3*, 333–356.

Kaushik, S. (2016). Introduction to Feature Selection methods with an example (or how to select the right variables?). Retrieved November 15, 2020, from https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Kishimoto, N., & Kim, Y. (2014). Prepayment behaviors of Japanese residential mortgages. *Japan & The World Economy*, *30*, 1–9. https://doi.org/10.1016/j.japwor.2013.12.002

LaCour-Little, M. (2008). Mortgage Termination Risk: A Review of the Recent Literature. *Journal of Real Estate Literature*, *16*(3), 297–326.

Larkin, T. K., & McManus, D. J. (2018). *Social Media, Anonymity, and Fraud : HP Forest Node in SAS ® Enterprise Miner ™*.

LC, T. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*, 149.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, M. (2014). *Residential Mortgage Probability of Default Models and Methods*.

Liang, Y., Jin, X., & Wang, Z. (2019). *Loanliness : Predicting Loan Repayment Ability by Using Machine Learning Methods*.

Louzis, D. P., Vouldis, A. T., & Metaxas, V. L. (2010). *Macroeconomic and bank-specific determinants of non-performing loans in Greece: a comparative study of mortgage, business and consumer loan portfolios*.

Makowski, P. (1985). Credit scoring branches out. *Credit World*, *75*, 30–37.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research*, *136*(1), 190–211. https://doi.org/10.1016/S0377-2217(01)00052-2

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega: International Journal of Management Science*, *31*(2), 83–96. https://doi.org/10.1016/S0305-0483(03)00016-1

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Mangasarian, O. L. (1965). Linear and Nonlinear Separation of Patterns by Linear Programming. *Operations Research*, *13*(3), 444–452.

Mcclelland, J. L., & Rumelhart, D. E. (1986). *Parallel Distributed Processing Explorations in the Microstructure of Cognition Volume 1 : Foundations*. *The MIT Press*.

Meis, J. (2015). *Modelling prepayment risk in residential mortgages*. ERASMUS UNIVERSITY ROTTERDAM.

Mercer Oliver Wyman. (2003). *Study on the Financial Integration of European Mortgage Markets*.

Mester, L. J. (1997). What is the Point of Credit Scoring ? *Business Review (Federal Reserve Bank of Philadelphia)*.

Mishina, Y., Tsuchiya, M., & Fujiyoshi, H. (2014). Boosted random forest. *Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, *2*, 594–598. https://doi.org/10.5220/0004739005940598

Mpofu, T. P., & Mukosera, M. (2014). Credit Scoring Techniques : A Survey. *International Journal of Science and Research*, *3*(8).

Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability (Switzerland)*, *11*(3), 1–23. https://doi.org/10.3390/su11030699

Narkhede, S. (2018). Understanding AUC - ROC Curve. Retrieved August 17, 2020, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nikulski, J. (2020). The Ultimate Guide to AdaBoost, random forests and XGBoost. Retrieved August 28, 2020, from https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f

Nord, C., & Keeley, J. (2016). An Introduction to the HPFOREST Procedure and its Options, 1–12. Retrieved from https://www.mwsug.org/proceedings/2016/AA/MWSUG-2016-AA20.pdf

Pandey, P. (2019). Data Preprocessing : Concepts. Retrieved October 30, 2020, from https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

Park, H. M. (2011). *Practical Guides To Panel Data Modeling : A Step by Step*.

Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, *112*(2), 310–321. https://doi.org/10.1016/S0377-2217(97)00398-6

Saito, T. (2018). *Mortgage Prepayment Rate Estimation with Machine Learning*.

SAS. (2020). Machine Learning - What it is and why it matters. Retrieved August 24, 2020, from https://www.sas.com/en_us/insights/analytics/machine-learning.html

SAS. (2021a). SAS Documentation - Gradient Boosting Node. Retrieved June 19, 2021, from https://documentation.sas.com/doc/en/emref/14.3/n0t6j7sk2xn3mon1e7ulvypjppew.htm

SAS. (2021b). SAS Documentation - HP Forest Node. Retrieved July 6, 2021, from https://documentation.sas.com/doc/en/emref/14.3/p1uhmtoprigyvkn147i1tw9e2ax0.htm

SAS. (2021c). SAS Documentation - Impute Node. Retrieved July 2, 2021, from https://documentation.sas.com/doc/en/emref/14.3/p1lmg1hmgbpz5cn1cxhmqfucnxlj.htm#n1iyh1abfi52p4n1177i21jtqswu.

SAS. (2021d). SAS Documentation - LASSO. Retrieved June 17, 2021, from https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_glmselect_details10.htm

SAS. (2021e). SAS Documentation - Model Comparison Node. Retrieved from https://documentation.sas.com/doc/en/emref/14.3/p01jgc9rmzsg37n1lfncp67t0unm.htm

SAS. (2021f). SAS Documentation - Neural Network Node: Reference. Retrieved July 6, 2021, from https://documentation.sas.com/doc/en/emref/14.3/p0zbgj1tu3h1uhn1x6regixbdg7v.htm

Schwartz, E. S., & Torous, W. N. (1992). Prepayment, Default, and the Valuation of Mortgage Pass-Through Securities. *The Journal of Business*, *65*(2), 221–239.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*. (J. W. & Sons, Ed.).

Sirignano, J. A., Sadhwani, A., & Giesecke, K. (2018). Deep Learning for Mortgage Risk. *SSRN Electronic Journal*.

Smalz, R., & Conrad, M. (1994). Combining evolution with credit apportionment: A new learning algorithm for neural nets. *Neural Networks*, *7*(2), 341–351.

Sousa, M. R., Gama, J., & Brandão, E. (2013). *Introducing Time-Changing Economics into Credit Scoring*.

Stanford University. (2020). Machine Learning Course Syllabus. Retrieved August 24, 2020, from https://www.coursera.org/learn/machine-learning#about

The World Bank Group. (2019). *Credit Scoring Approaches Guidelines*.

Ulloa, P. (2017). MACHINE LEARNING: Running A LASSO Regression in SAS. Retrieved June 17, 2021, from https://www.linkedin.com/pulse/machine-learning-running-lasso-regression-sas-paul-ulloa-mba

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*.

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking and Finance*, *22*(10–11), 1421–1439. https://doi.org/10.1016/S0378-4266(98)00059-4

Vidal, M. F., & Barbon, F. (2019). *Credit Scoring in Financial Inclusion*.

Wang, Y., Wang, S., & Lai, K. K. (2005). A New Fuzzy Support Vector Machine to Evaluate Credit Risk, *13*(6), 820–831.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*.

Williams, R. (2018). *Panel Data 3 : Conditional Logit / Fixed Effects Logit Models*.

Wyner, A. J., Mease, D., & Bleich, J. (2017). Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. *Journal of Machine Learning Research*, *18*, 1–33.

Xerez, R., Pereira, E., & Cardoso, F. D. (2019). *Habitação Própria em Portugal numa Perspetiva*.

Xerez, R., Rodrigues, P. G., Lima, J. de M., & Cardoso, F. D. (2019). Shifting from a homeowner society to a rental market? Over a decade of housing policy in Portugal, 2007-2017. In *Housing Policy and Tenure Types in the 21st Century - A southern european perspective*. https://doi.org/10.1007/s10901-005-9023-4

Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274–13283. https://doi.org/10.1016/j.eswa.2011.04.147

Zhao, K. (2018). *Predictive Modeling Using Artificial Neural Networks in SAS® Enterprise Miner*.

Zhou, X., Zhang, D., & Jiang, Y. (2008). A new credit scoring method based on rough sets and decision tree. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5012 LNAI*, 1081–1089. https://doi.org/10.1007/978-3-540-68125-0_117

7. APPENDIX

Appendix 1. METHODOLOGICAL STEPS AND SOFTWARE

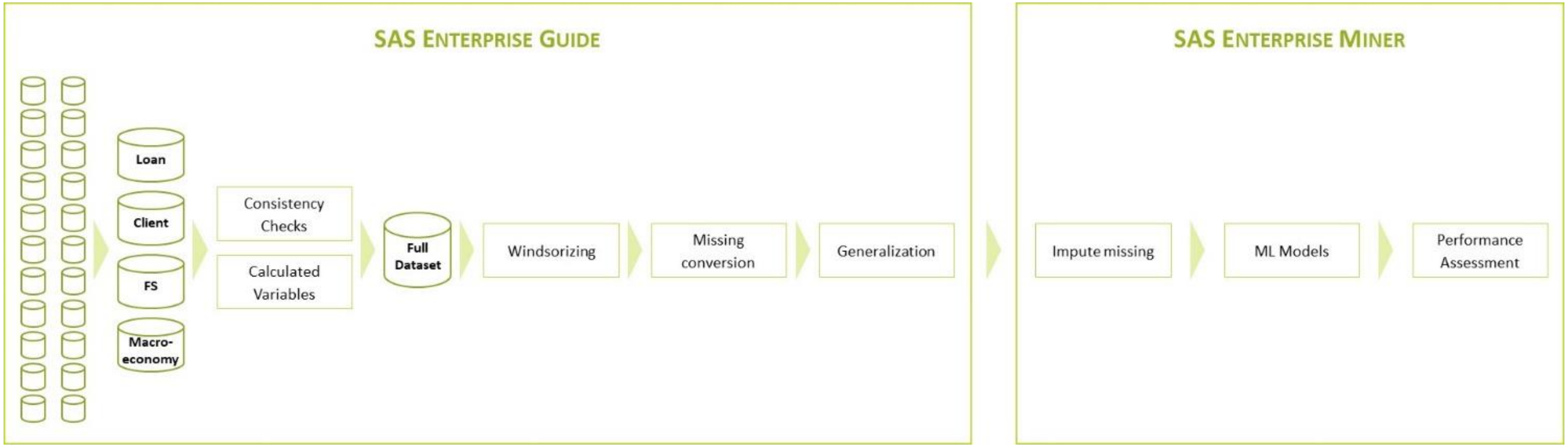

Figure 17 – Methodological steps and software of the steps performed

Appendix 2. DATASET VARIABLES

The dataset is comprised of the following variables:

Variable Category	Name	Description	Source	Periodicity
Operation	ANO_CONSTRUCAO	Year of construction of the property associated with the mortgage loan.	Bank	Monthly
	ANTIGUIDADE_IMOVEL	Age of the property. Calculated variable, based on the year of construction, i.e. difference between the year of the data and the year of construction of the property.	Calculated	Monthly
	C_POSTAL	Postal code of the property, with seven digits.	Bank	Monthly
	CONCELHO	Municipality of the property, extracted from CTT's data, based on the property's postal code.	CTT	Monthly
	DATA_ABERTURA	Opening date of the loan.	Bank	Origination
	DISTRITO	District of the property, extracted from CTT's data, based on the property's postal code.	CTT	Monthly
	FINALIDADE	Purpose of the loan.	Bank	Monthly
	LTV_ATUAL	Loan-to-value according to the current value of the property.	Bank	Monthly
	LTV_ORIG	Loan-to-value according to the value of the property at the origination date.	Bank	Monthly
	MONTANTE_AMORT	Amount early repaid in the loan.	Bank	Monthly
	MONTANTE_FINANCIADO	Total amount financed in the loan.	Bank	Monthly
	MONTANTE_RESIDUAL	Residual amount of the loan, i.e. difference between the amount financed and the total already amortized by the customer.	Bank	Monthly
	N_PREST_PAGAS	Number of instalments paid.	Bank	Monthly
	PERC_PRAZO	Percentage of residual term elapsed. Measured as the percentage of the term in the contract that has elapsed.	Calculated	Monthly
	PRAZO	Loan term.	Bank	Monthly
	PRAZO_RESIDUAL	Residual loan term, i.e. difference between the loan term and the period already elapsed.	Bank	Monthly
	TARGET_AMORT_PARCIAL	Target variable, which indicates if the operation had a partial early repayment.	Bank	Monthly
	TARGET_AMORT_TOTAL	Target variable, which indicates if the operation had a total early repayment.	Bank	Monthly
	TOTAL_AMORT_PARCIAL	Total partial early repayments. This is a calculated variable, based on the target	Calculated	Monthly

Variable Category	Name	Description	Source	Periodicity
		variable, which indicates the existence of early repayments.		
	T_JURO	Interest rate of the loan.	Bank	Monthly
	T_SPREAD	Spread rate of the loan.	Bank	Monthly
	Z_FIM_CTTO	Contract end date.	Bank	Origination
Client	DT_NASCIMENTO	Date of birth of the client.	Bank	Origination
	ESTADO_CIVIL	Marital status of the client. This variable has the following list of values: - Unknown - Single - Married with common-law marriage - Married with separation of property - Married in communion of acquired regime - Married in dotal regime - De facto union - Judicially separated from persons and assets - Divorced - Widower - Married - Judicially separated from property	Bank	Origination
	HAB_PROF	Qualification / level of education of the client. This variable has the following list of values: - Primary education - High school - Bachelor degree - Master degree - Doctorate - No studies - Superior professional technical courses - Unknown	Bank	Origination
	IDADE	Age of the client. Calculated variable, based on the client's date of birth, i.e. difference between the year of the information and the client's date of birth.	Calculated	Yearly
	PROFISSAO	Client profession.	Bank	Origination
	RENDIMENTO	Yearly income of the client.	Bank	Origination
	SCORING	Monthly client notation.	Bank	Monthly

Variable Category	Name	Description	Source	Periodicity
Behaviour in Bank and Financial System	IND_CREDITO	Payment incident indicator. This variable has the following list of values: - Regular credit - Other indications, as long as delay in payment is ≤ 30 days - Delays in payment > 30 days - Restructured due to financial difficulties - Default	Bank	Monthly
	INIB_CHEQUE	Check inhibition indicator.	Bank	Monthly
	M_PRS_MENS_BANCA	Amount of monthly instalments in the national financial system.	Bank	Monthly
	M_PRS_MENS_BANK	Amount of monthly instalment in the bank.	Bank	Monthly
	N_DIAS_ATRASO	Number of days overdue.	Bank	Monthly
	N_OPER_BANCA_REAIS	Number of operations in the national financial system. These operations include effective credit in a regular situation, overdue credit, credit written-off to assets, renegotiated credit, credit overdue in judicial litigation and credit written-off to assets in judicial litigation.	Bank	Monthly
	N_OPER_BANK_REAIS	Number of operations in the bank. These operations include effective credit in a regular situation, overdue credit, credit written-off to assets, renegotiated credit, credit overdue in judicial litigation and credit written-off to assets in judicial litigation.	Bank	Monthly
	N_OPER_BANCA_POT	Number of operations in the national financial system. These operations include potential credit.	Bank	Monthly
	N_OPER_BANK_POT	Number of operations in the bank. These operations include potential credit.	Bank	Monthly
	N_PRODUTOS_BANCA	Number of financial products in the national financial system.	Bank	Monthly
	N_PRODUTOS_BANK	Number of financial products in the bank.	Bank	Monthly
	PERC_UTILIZA	Percentage of use of credit cards.	Bank	Monthly
	RESP_BANCA_POT	Total amount of credit (liabilities) of the client in the national financial system. These operations include effective credit in a regular situation, overdue credit, credit written-off to assets, renegotiated credit, credit overdue in judicial litigation and credit written-off to assets in judicial litigation.	Bank	Monthly

Variable Category	Name	Description	Source	Periodicity
	RESP_BANK_POT	Total amount of credit (liabilities) of the client in the bank. These operations include effective credit in a regular situation, overdue credit, credit written-off to assets, renegotiated credit, credit overdue in judicial litigation and credit written-off to assets in judicial litigation.	Bank	Monthly
	RESP_BANCA_REAIS	Total amount of credit (liabilities) of the client in the national financial system. These operations include potential credit.	Bank	Monthly
	RESP_BANK_REAIS	Total amount of credit (liabilities) of the client in the bank. These operations include potential credit.	Bank	Monthly
	SALDO_DO_06M	Total balance in sight deposits, six months.	Bank	Monthly
	SALDO_DO_12M	Total balance in sight deposits, twelve months.	Bank	Monthly
	SALDO_DP_06M	Total balance in term deposits, six months.	Bank	Monthly
	SALDO_DP_12M	Total balance in term deposits, twelve months.	Bank	Monthly
	TOT_DEVEDORES_BANCA	Number of debtors in the national financial system, associated with the customer.	Bank	Monthly
	TX_ESFORCO_BANCA	Debt-service ratio, in the financial system. Measure of the percentage of the client's monthly instalment in the financial system in the income.	Calculated	Monthly
	TX_ESFORCO_BANK	Debt-service ratio, in the bank. Measure of the percentage of the client's monthly instalment in the bank in the income.	Calculated	Monthly
Point in time	MÊS	Month of the observation.	Bank	-
	ANO	Year of the observation.	Bank	-
Macroeconomy	ED_LICENC_TVH	Number of licensed buildings, year-on-year change. I.e. authorization granted by the City Councils under specific legislation, for the execution of Works (new constructions, extensions, transformations, restorations and demolitions of buildings).	Statistics Portugal (INE)	Monthly
	ENDIV_PART_TVH	Indebtedness of families and non-profit institutions serving families in Portugal, year-on-year change.	Bank of Portugal (Bpstat)	Monthly
	GRAU_POUP_PART_TVH	Degree of household savings, year-on-year change.	Statistics Portugal (INE)	Monthly
	IND_COINC_TVH	Coincident indicators for private consumption, year-on-year change. This seeks to capture the underlying evolution of the year-on-year variation in private consumption.	Bank of Portugal (Bpstat)	Monthly

Variable Category	Name	Description	Source	Periodicity
	IND_PRECOS_HAB_TVH	Housing price index, which measures the evolution of housing prices in the residential market in the national territory, year-on-year change.	Statistics Portugal (INE)	Quarterly
	IND_SENT_ECO_TVH	Economic sentiment indicator, year-on-year change. This short-term indicator allows the monitoring of the evolution of the economic environment and anticipating the evolution of the main macroeconomic aggregates for Portugal.	Bank of Portugal (Bpstat)	Monthly
	N_FOGOS_CONST_TVH	Number of licensed dwellings in new buildings for family housing, year-on-year change.	Statistics Portugal (INE)	Monthly
	PERSP_SIT_EC	Outlook on the country's economic situation over the next 12 months, year-on-year change.	Statistics Portugal (INE)	Monthly
	PIB	GDP at market prices, year-on-year change.	Bank of Portugal (Bpstat)	Quarterly
	TX_DIVORCIO_TVH	Number of marriages dissolved by divorce, year-on-year change.	Statistics Portugal (INE)	Yearly
	TAXA_INFLACAO_TVH	Harmonized consumer price index, year-on-year change.	Bank of Portugal (Bpstat)	Monthly
	TAXA_JURO_DP_TVH	Interest rate in term deposits (< 1 year, private individuals), year-on-year change.	Bank of Portugal (Bpstat)	Monthly
	TAXA_JURO_HAB_TVH	Interest rate in mortgage loans (private individuals), year-on-year change.	Bank of Portugal (Bpstat)	Monthly
	TX_DESEMPREGO_TVH	Unemployment rate of the active population aged between 15 and 74 years, year-on-year change.	Statistics Portugal (INE)	Monthly

Table 21 – Variables considered in the dataset

Appendix 3. DATA DESCRIPTION – HISTOGRAM, BAR CHART AND BOX PLOT

The charts performed can be found below.

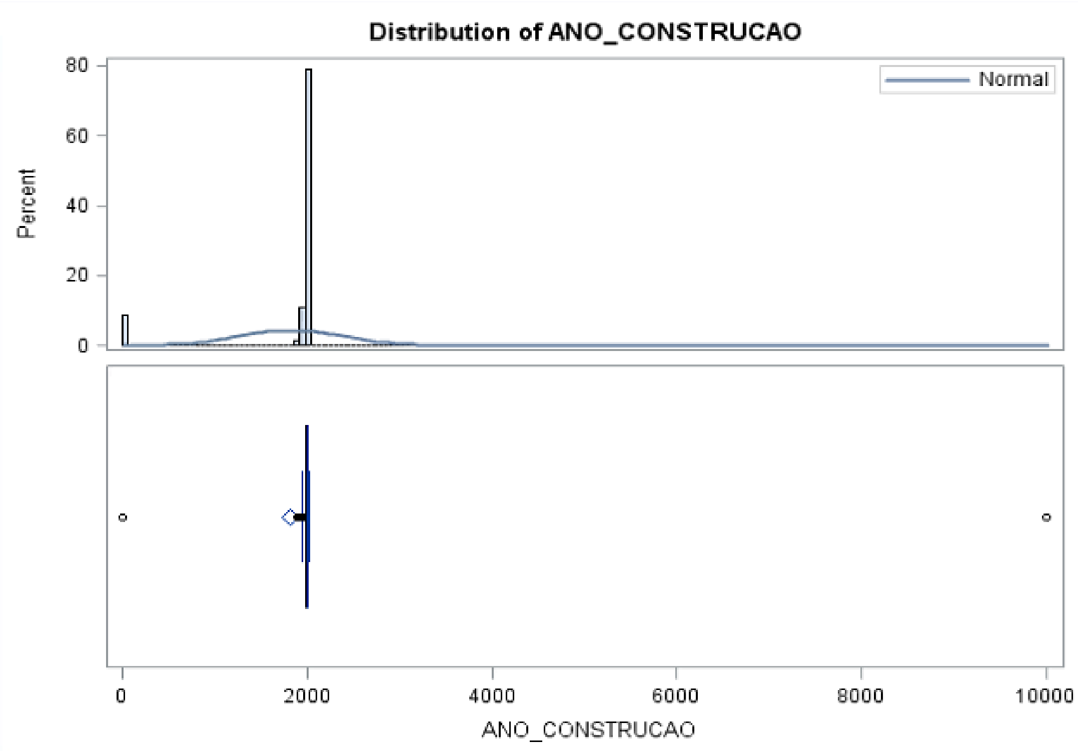

Figure 18 – Histogram and box plot of the year of construction

As can be seen by the chart above, the year of construction of the property shows two extreme values: 0 and 9999, which entail data quality errors.

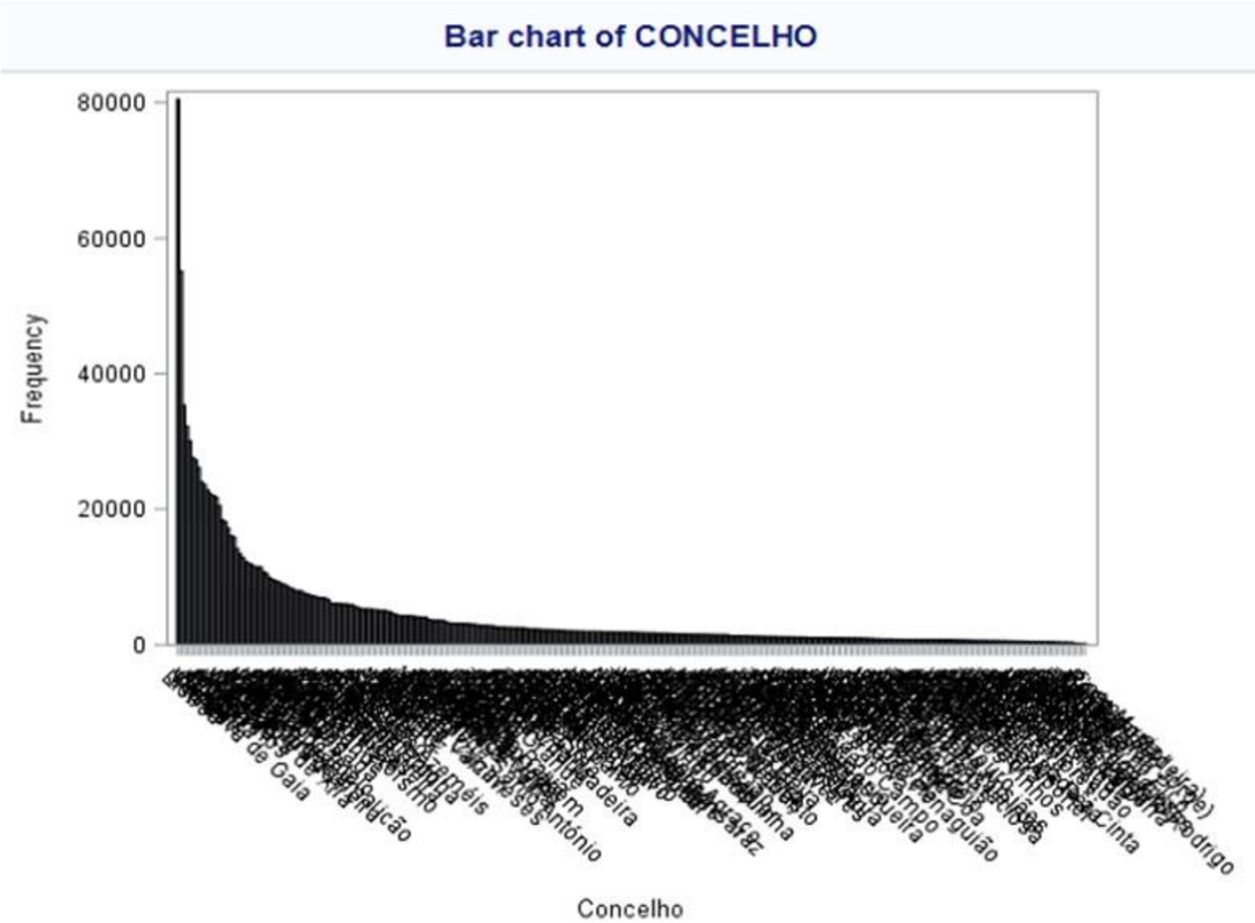

Figure 19 – Histogram and box plot of the municipality

As can be seen by the chart above, the municipality displays many unique values, with a left tail being predominant, i.e. with the majority of observations in a few classes.

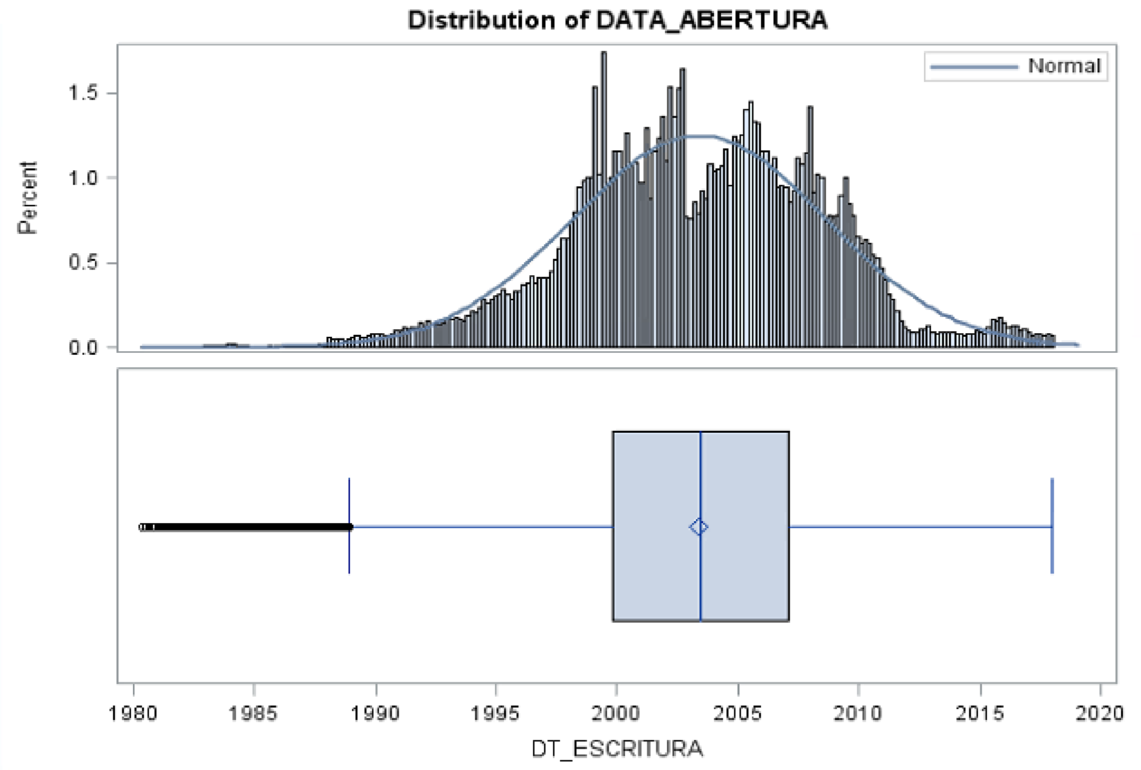

Figure 20 – Histogram and box plot of the opening date of the loan

As can be seen by the chart above, the opening date shows a distribution similar to the normal distribution, with a slight skew towards the right.

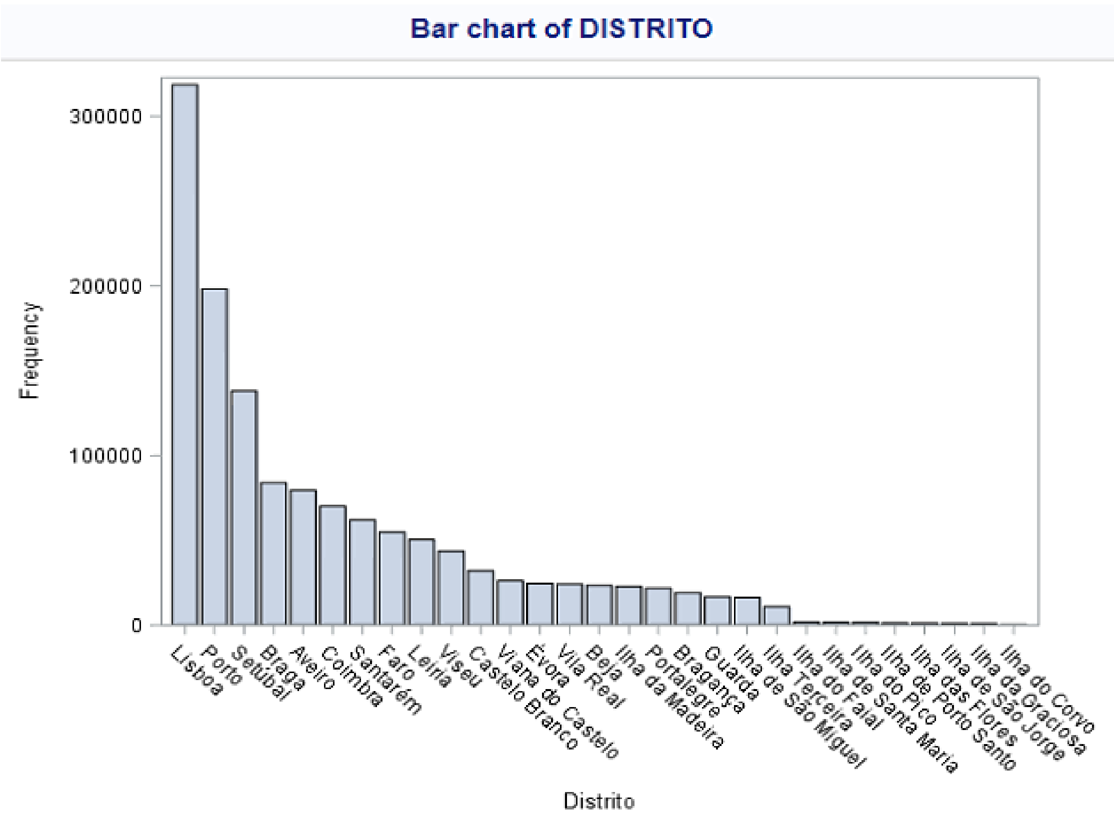

Figure 21 – Histogram and box plot of the district

As can be seen by the chart above, and as expected, the district has a lower level of granularity than the municipality, with the majority of loans from Lisbon and Oporto, the two main cities in Portugal.

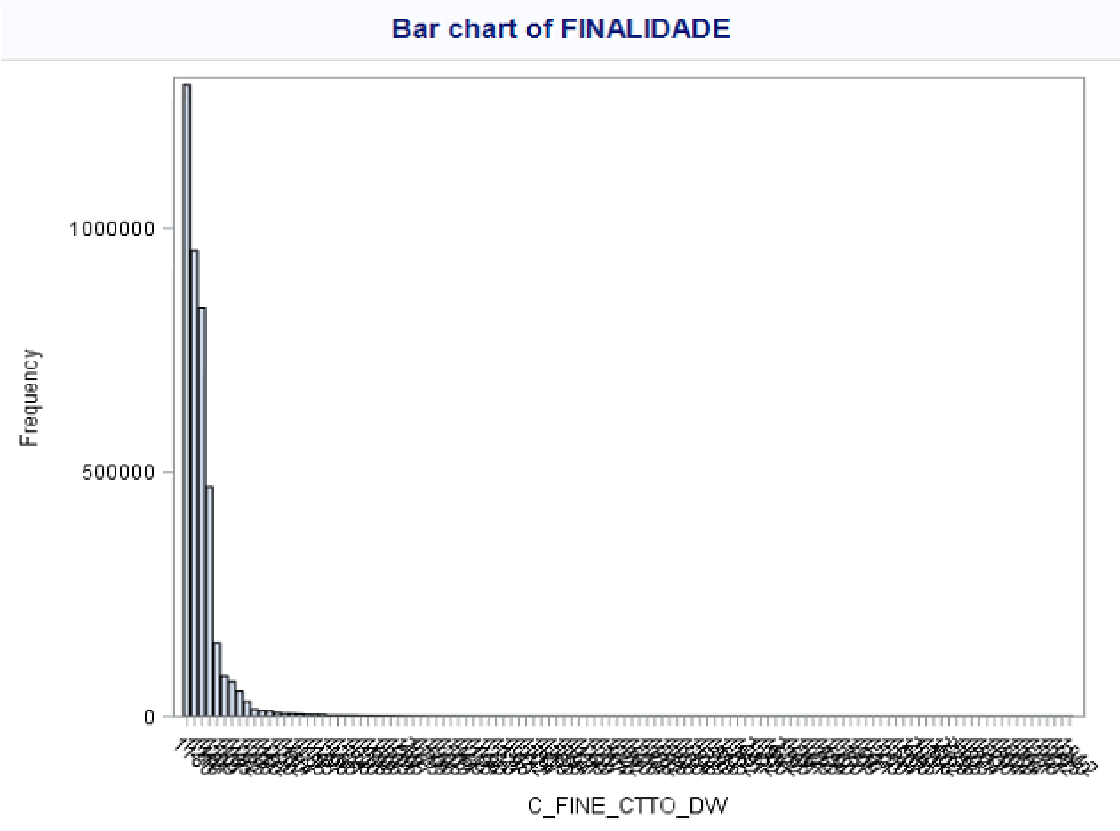

Figure 22 – Histogram and box plot of the loan purpose

As can be seen by the chart above, the purpose of the loan displays many unique values, with the majority of observations in 5 classes.

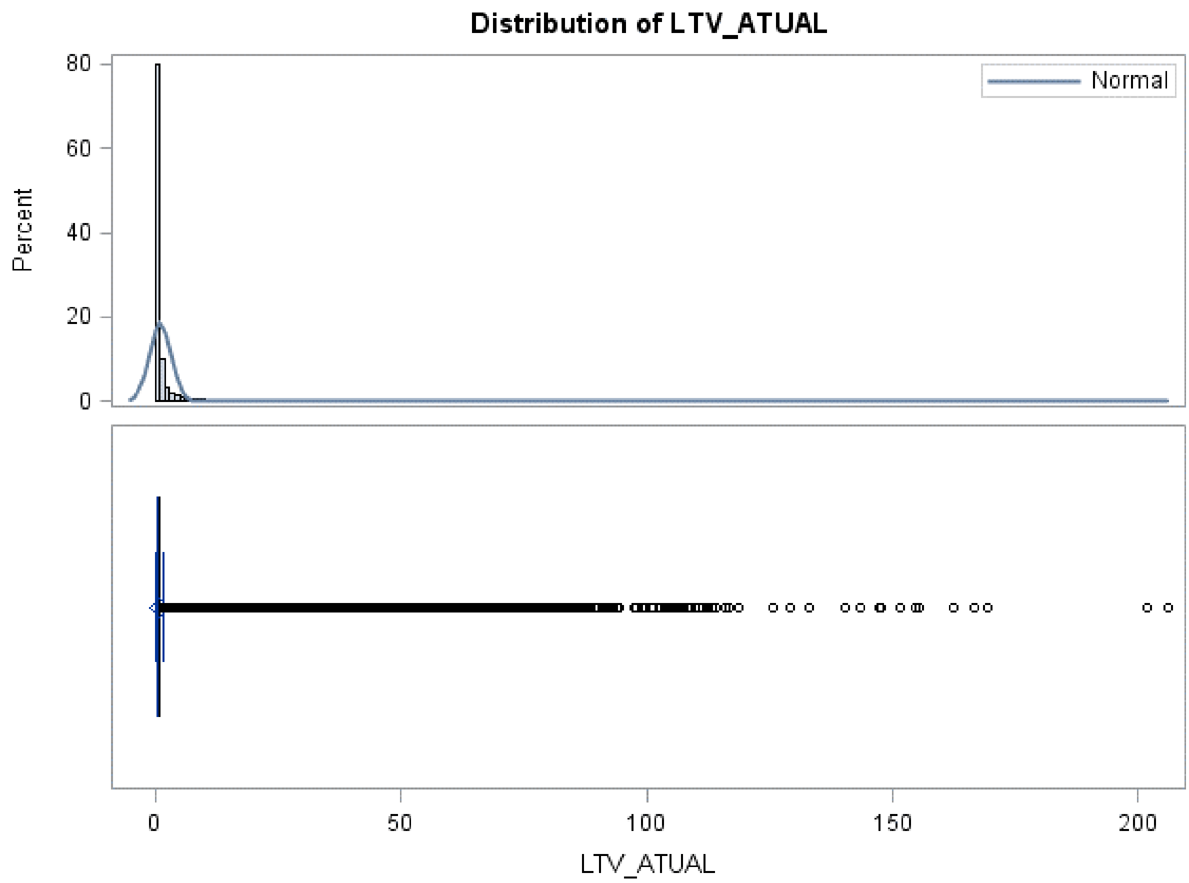

Figure 23 – Histogram and box plot of the LTV of the current property evaluation

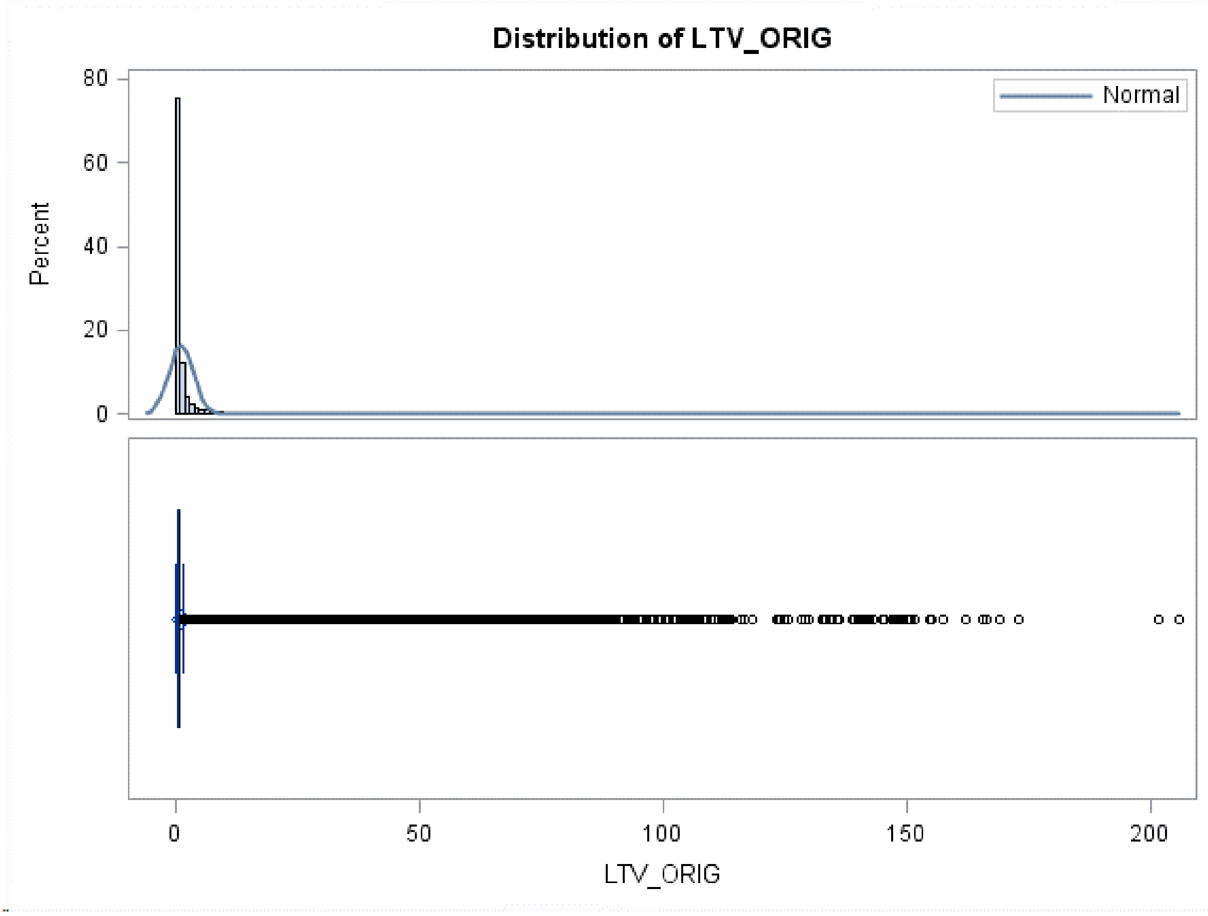

Figure 24 – Histogram and box plot of the LTV of the original property evaluation

Both LTV show a left-skewed distribution, being severely impacted by the outliers.

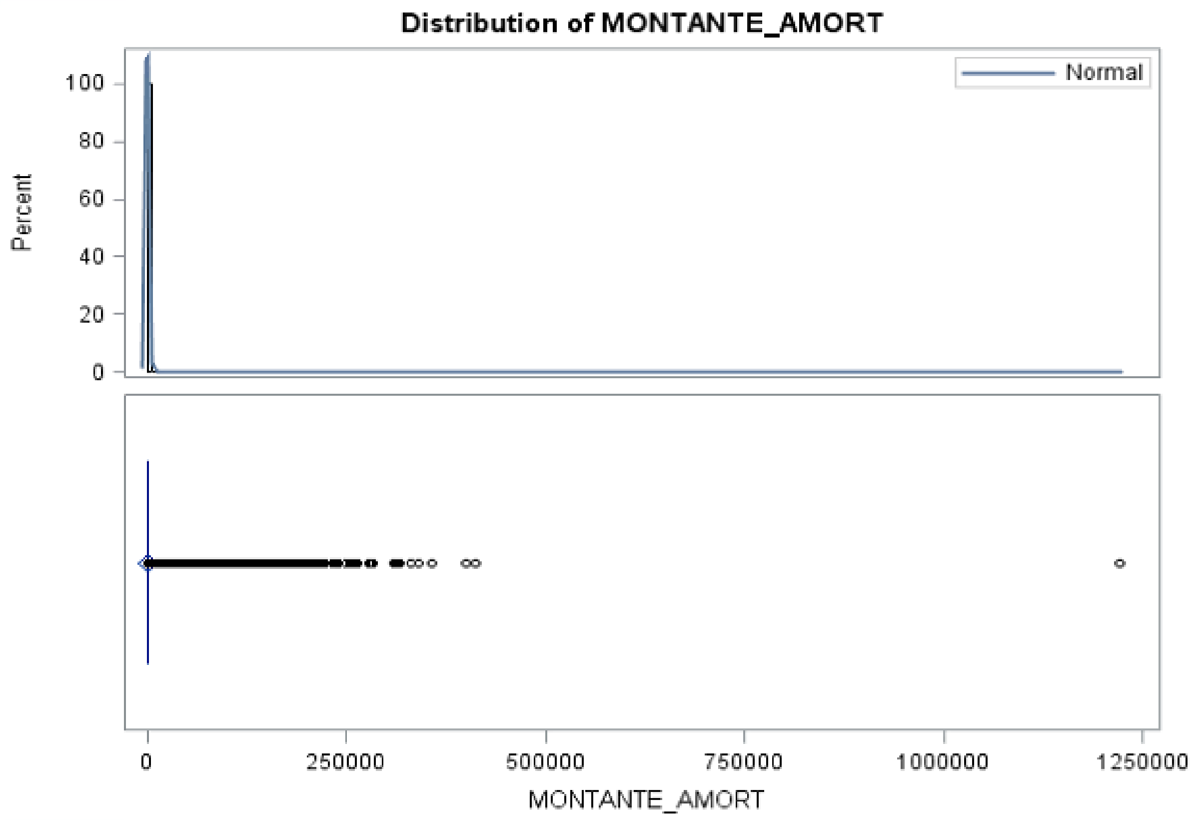

Figure 25 – Histogram and box plot of the early repaid amount

The early repaid amount shows a left-skewed distribution, being severely impacted by the outliers.

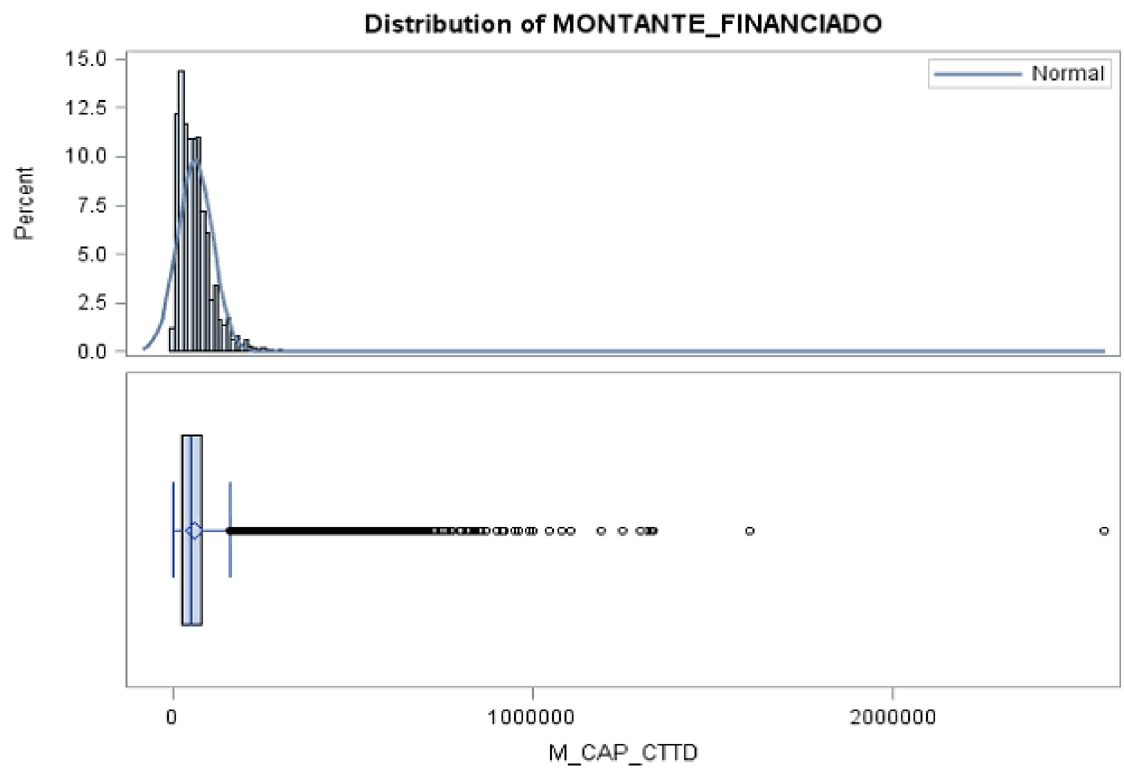

Figure 26 – Histogram and box plot of the financed amount

As can be seen by the chart above, the amount financed in the loan shows a left-skewed distribution, which is severely impacted by the outliers.

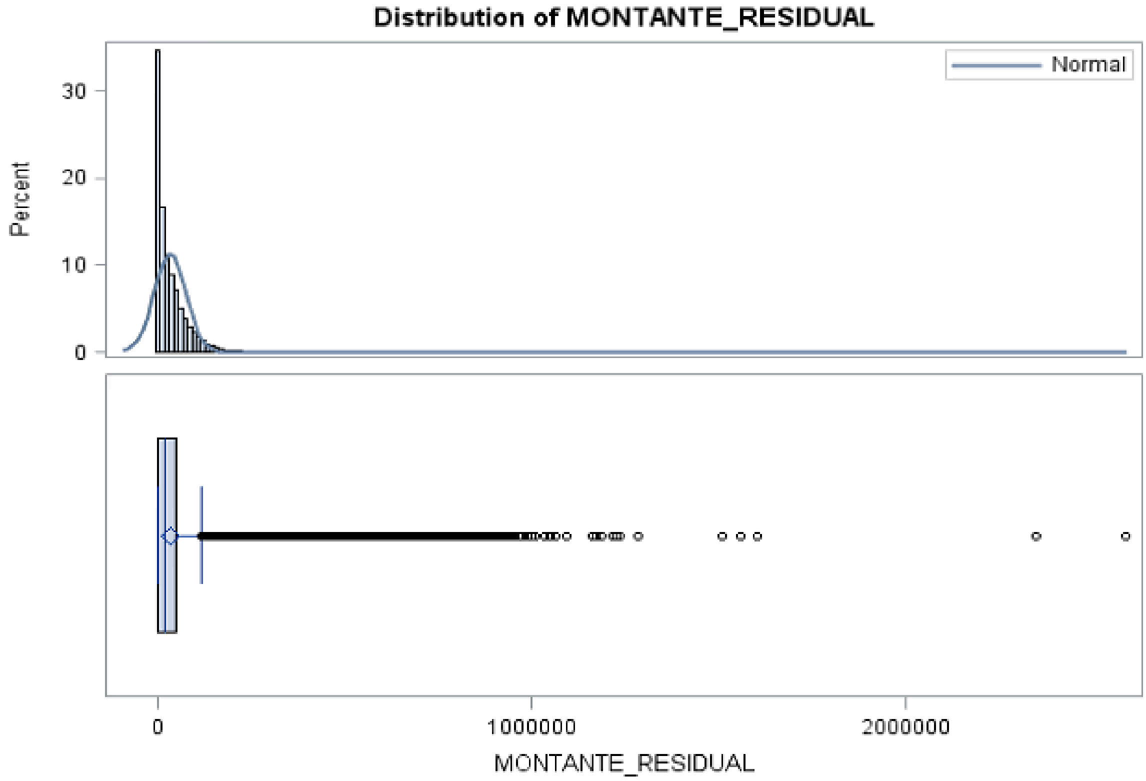

Figure 27 – Histogram and box plot of the residual amount

As with the amount financed, the residual amount shows a left-skewed distribution, which is severely impacted by the outliers.

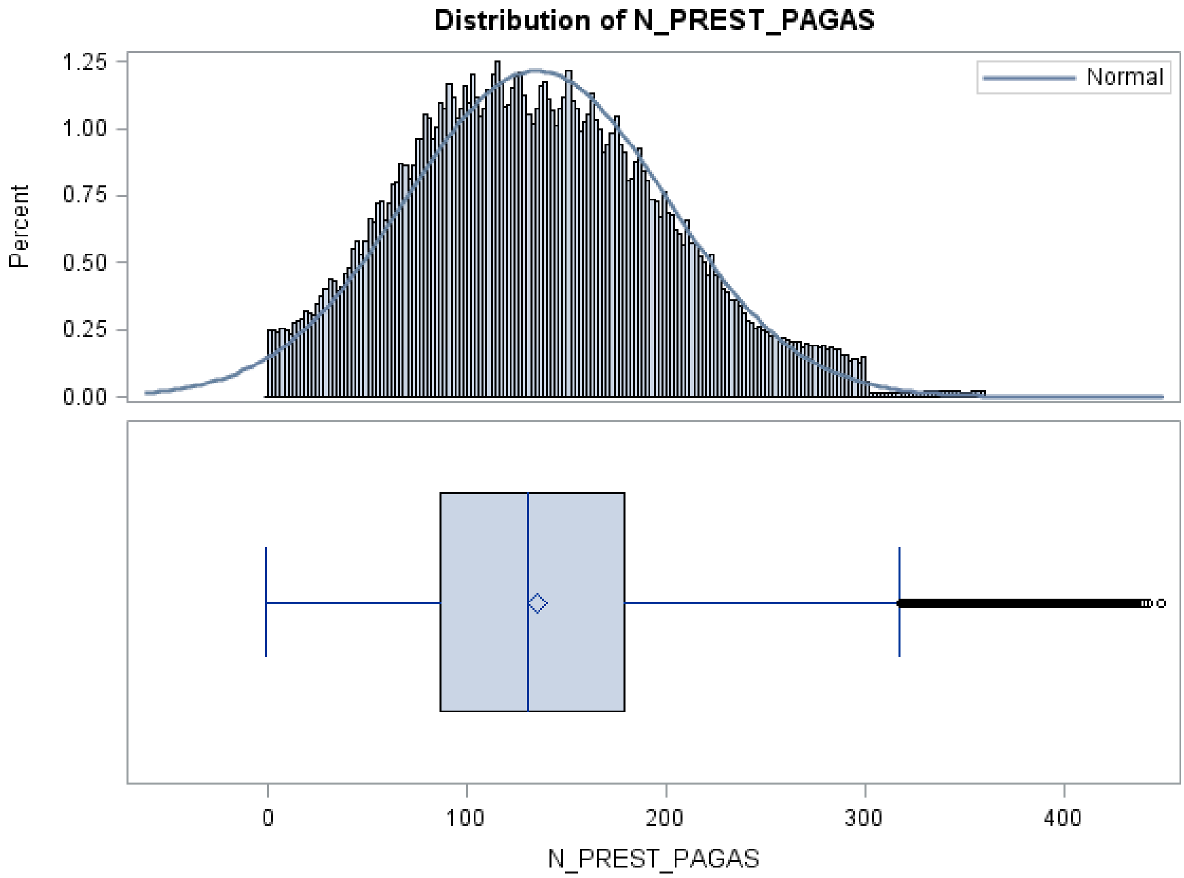

Figure 28 – Histogram and box plot of the number of instalments paid

As can be seen by the chart above, there is a predominance of the right tail, being affected by contracts with a significant number of instalments paid.

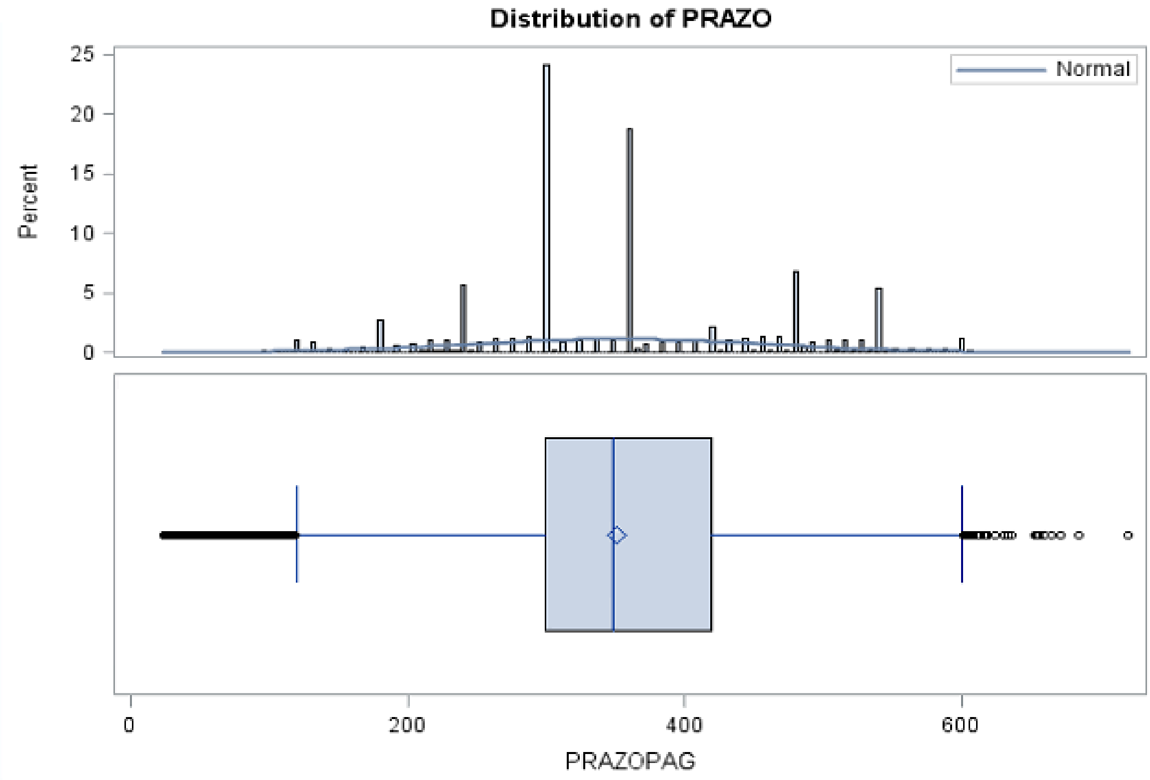

Figure 29 – Histogram and box plot of the loan term

As can be seen by the chart above, there is a predominance of some loan terms (mainly 25 and 30 years).

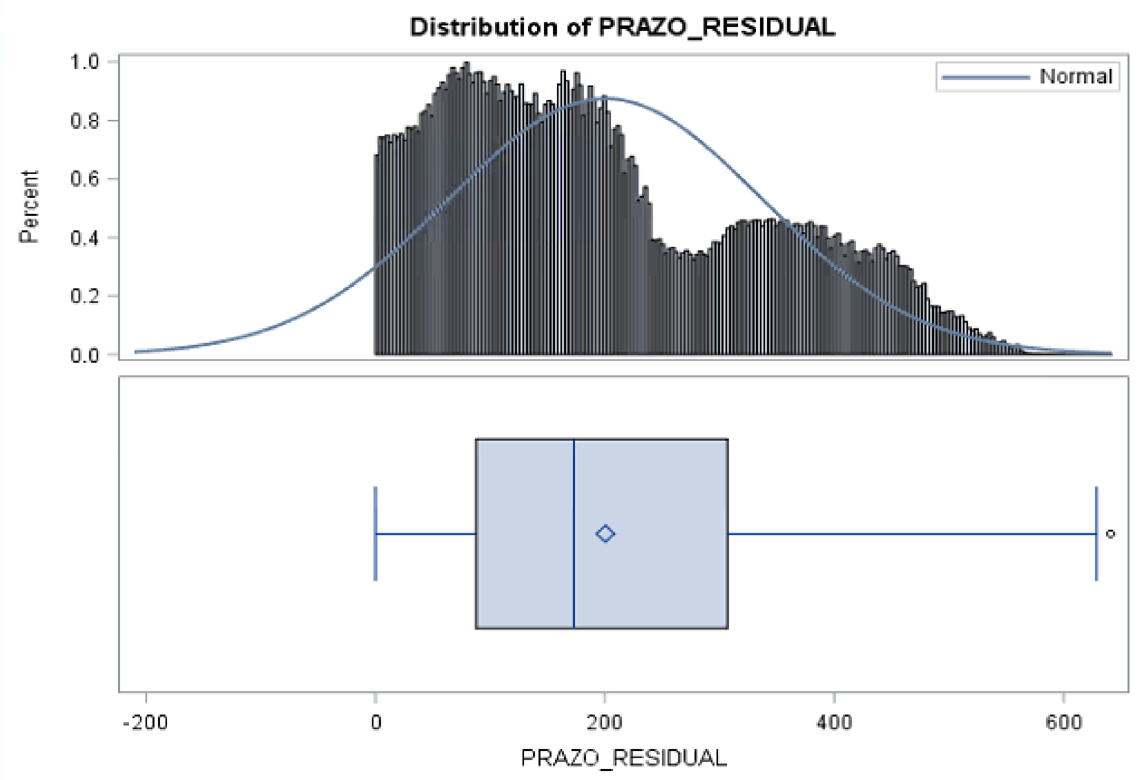

Figure 30 – Histogram and box plot of the residual term

As can be seen by the chart above, the residual loan term demonstrates the usual cadence in loan reduction over time.

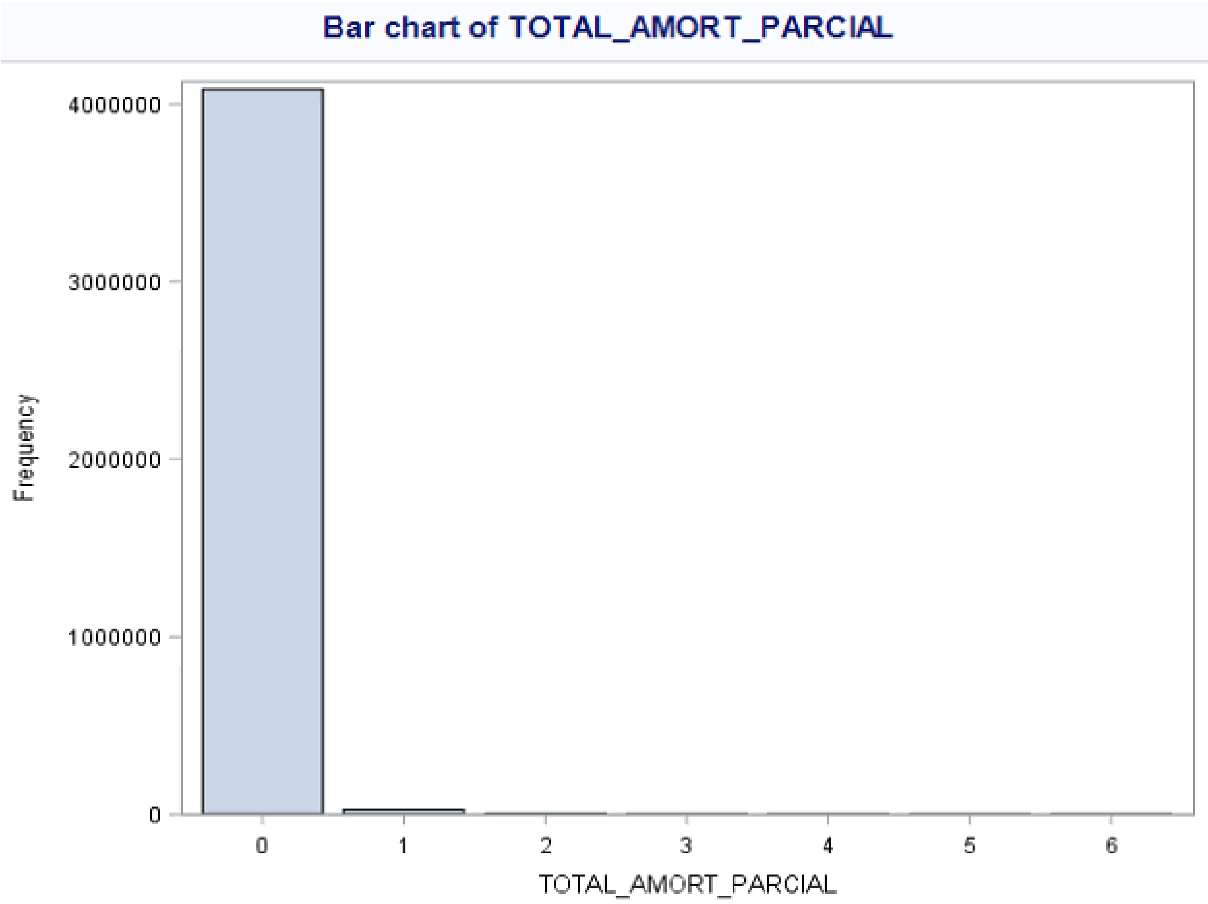

Figure 31 – Histogram and box plot of the total partial early repayments

As can be seen by the chart above, there is a high predominance of contracts without any partial amortization. The majority of contracts with partial repayments have no more than 2.

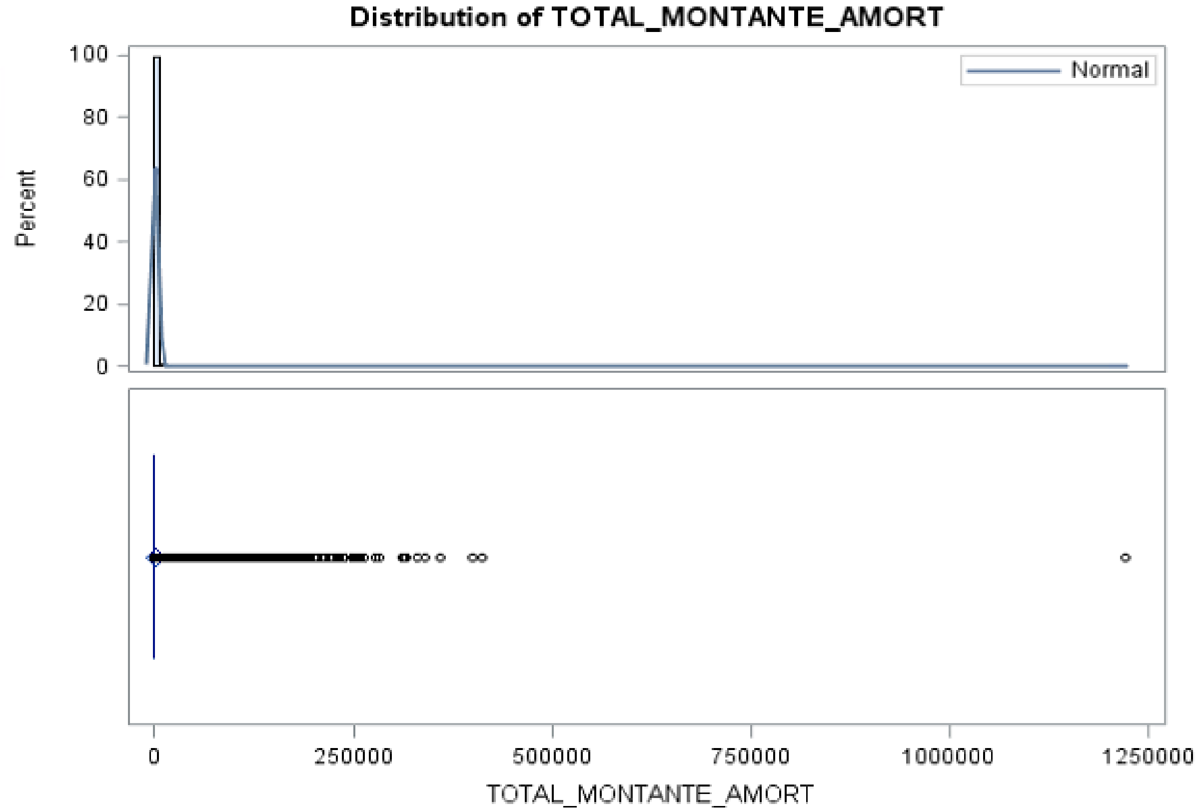

Figure 32 – Histogram and box plot of the total amount repaid

As can be seen by the chart above, and in line with the number of partial prepayments, there is a concentration on the left side of the chart.

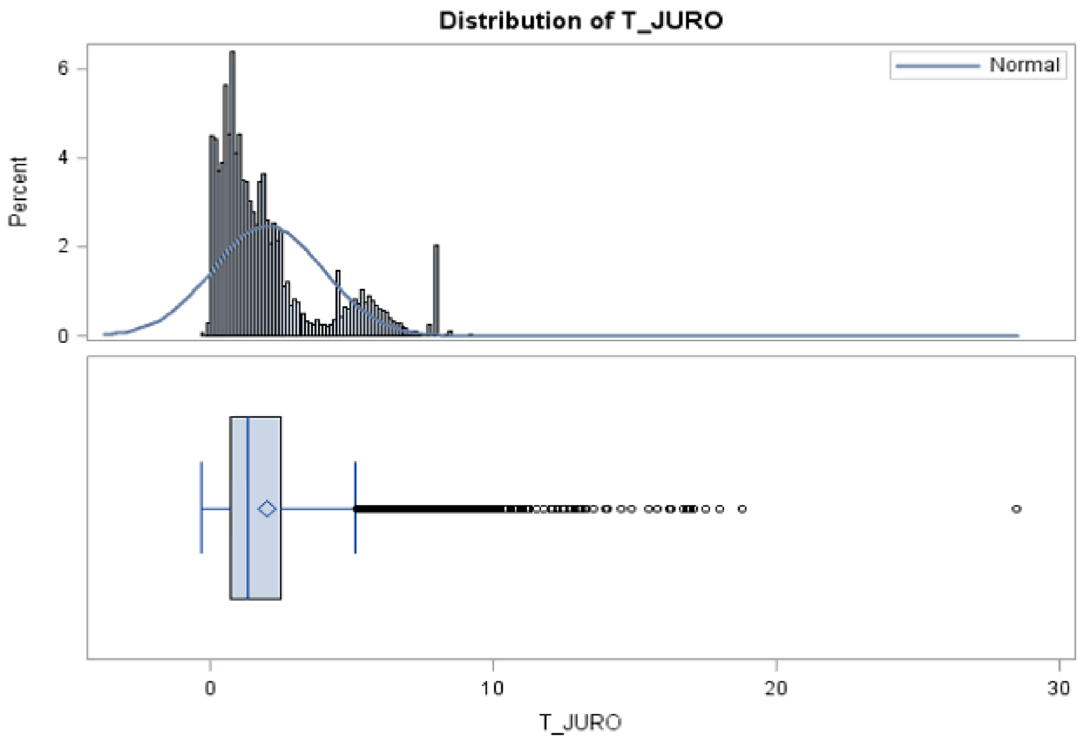

Figure 33 – Histogram and box plot of the interest rate

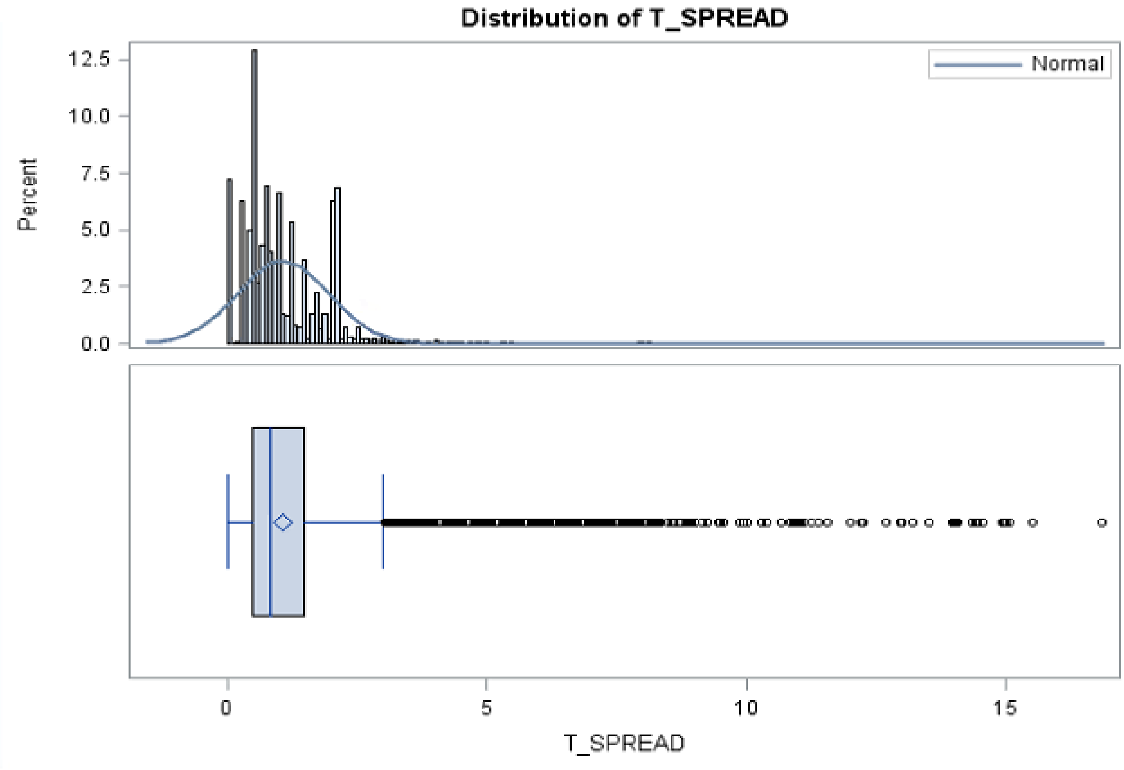

Figure 34 – Histogram and box plot of the spread rate

Both the interest and spread rate show a left-skewed distribution, which is severely impacted by the outliers.

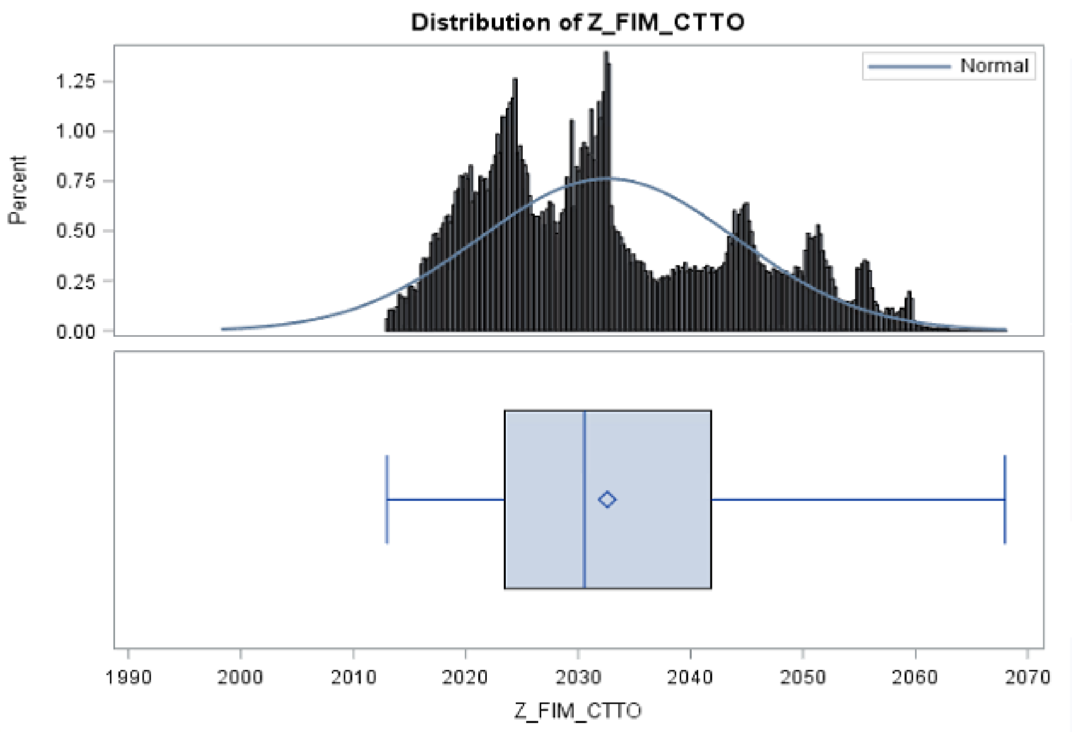

Figure 35 – Histogram and box plot of the contract end date

As can be seen by the chart above, the contract end date demonstrates a normal distribution without the presence of outliers.

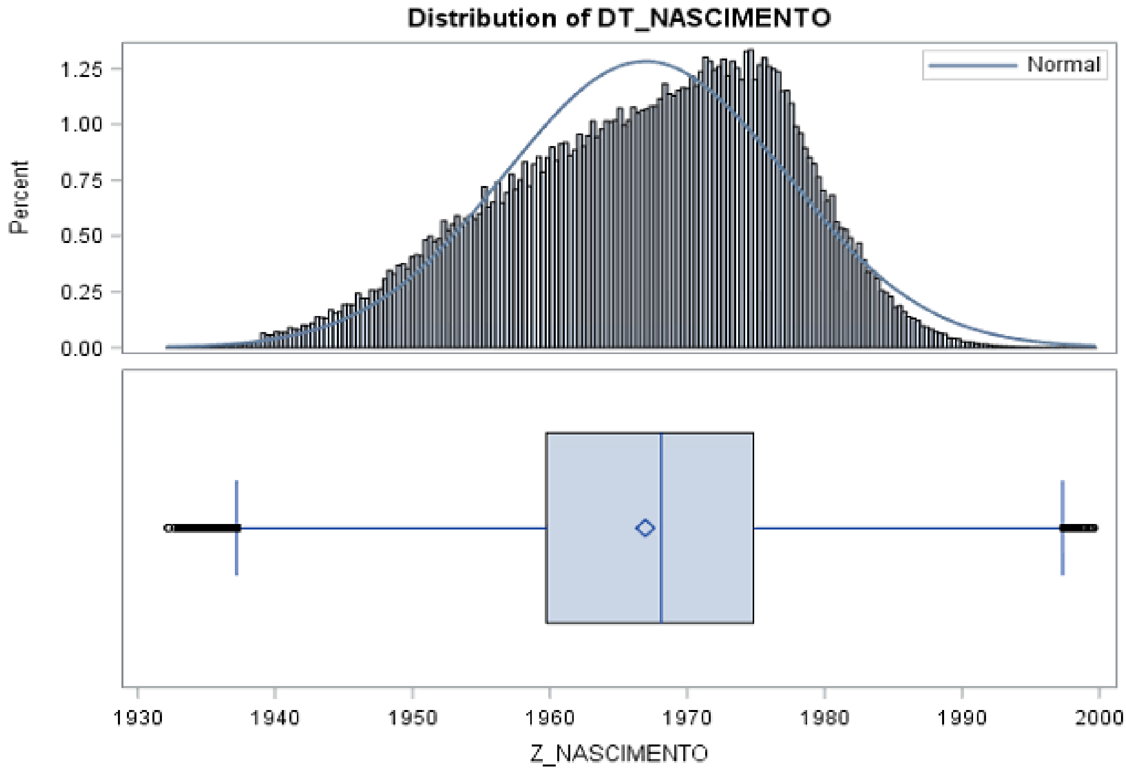

Figure 36 – Histogram and box plot of the date of birth

As can be seen by the chart above, the date of birth demonstrates a normal distribution with few outliers.

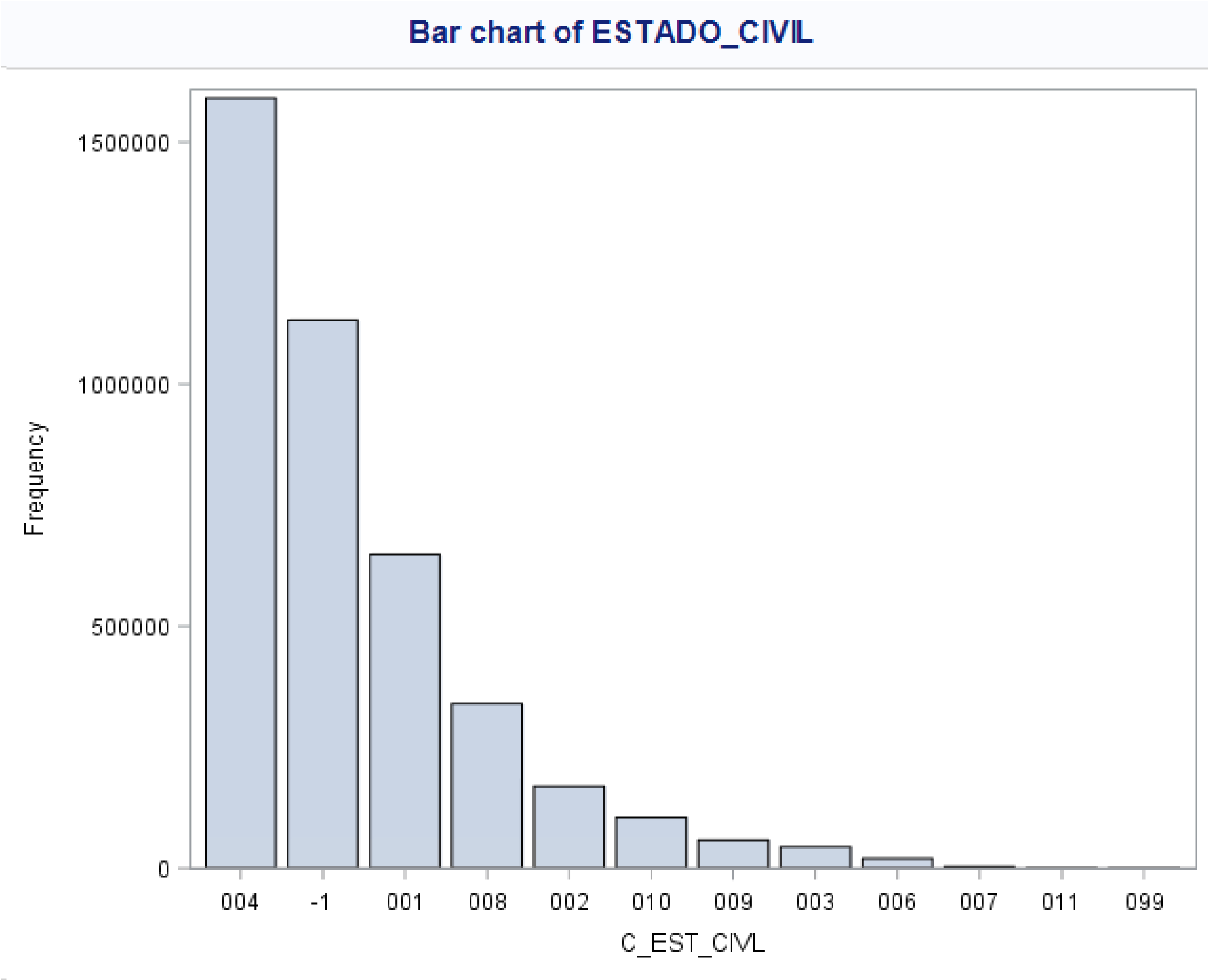

Figure 37 – Histogram and box plot of the marital status

As can be seen by the chart above, the majority of clients are married. The variable holds the following values:

-1, 0 and 099. Unknown
1. Single
2. Married with common-law marriage
3. Married with separation of property
4. Married in communion of acquired regime
5. Married in dotal regime
6. De facto union
7. Judicially separated from persons and assets
8. Divorced
9. Widower
10. Married
11. Judicially separated from property

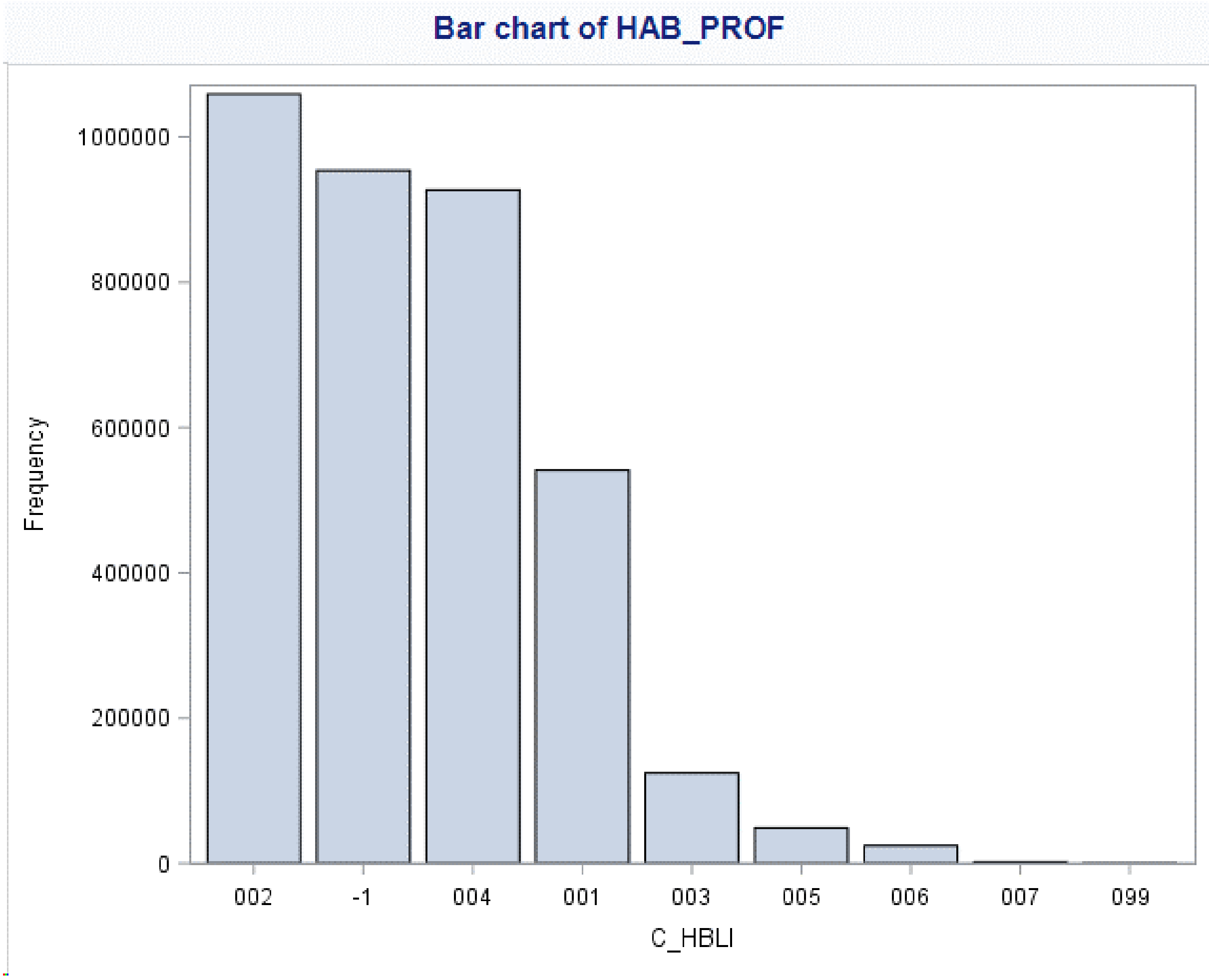

Figure 38 – Histogram and box plot of the level of education

As can be seen by the chart above, most clients have finished high school followed by a bachelor's degree. Furthermore, this variable, HAB_PROF, holds the following values:

- -1. Unknown
- 1. Primary education
- 2. High school
- 3. and 4. Bachelor degree
- 5. Master degree
- 6. Doctorate
- 7. No studies
- 98. Superior professional technical courses
- 99. Other

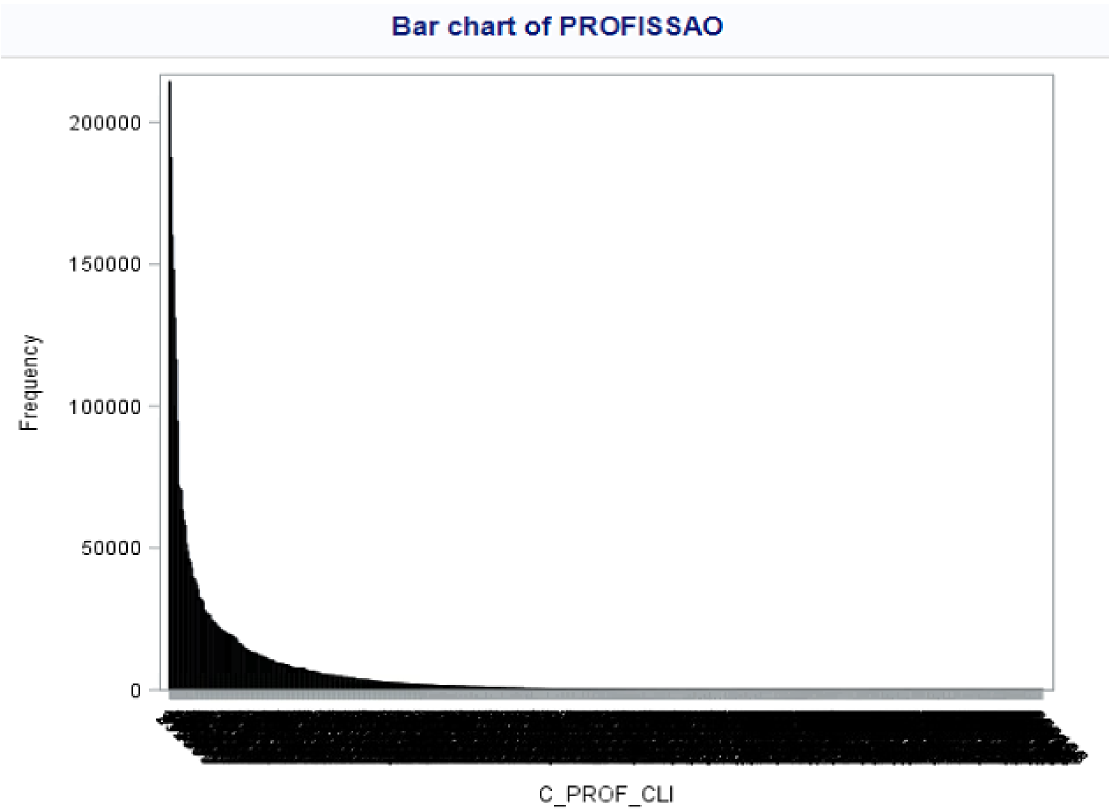

Figure 39 – Histogram and box plot of the profession

As can be seen by the chart above, the profession displays a significant amount of unique values (563 unique ones), with the majority of observations in a few classes.

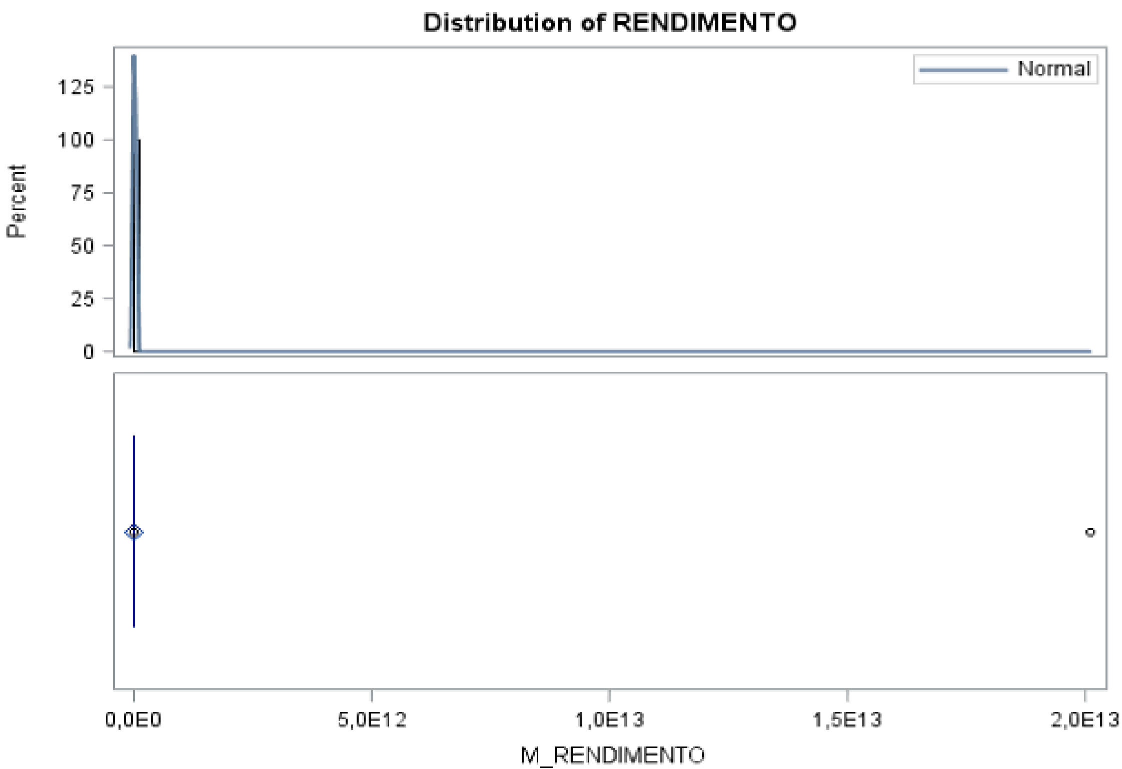

Figure 40 – Histogram and box plot of the yearly income

The client’s income distribution is severely impacted by the outlier with 20.084.001.000.000 €, a data quality issue.

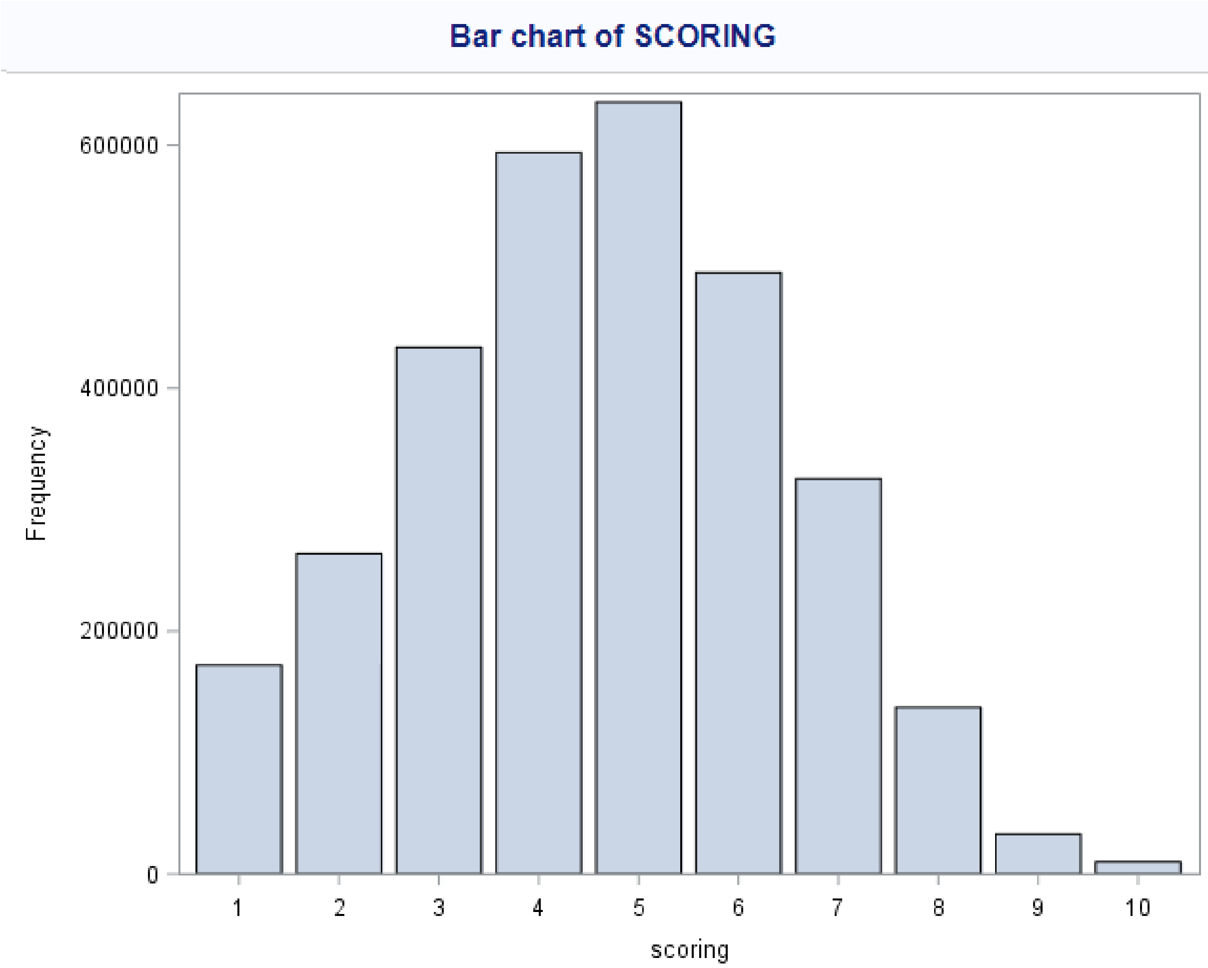

Figure 41 – Histogram and box plot of the scoring

The client’s scoring shows a normal distribution between the lower scoring levels (indicating the “better” clients) and higher scoring levels.

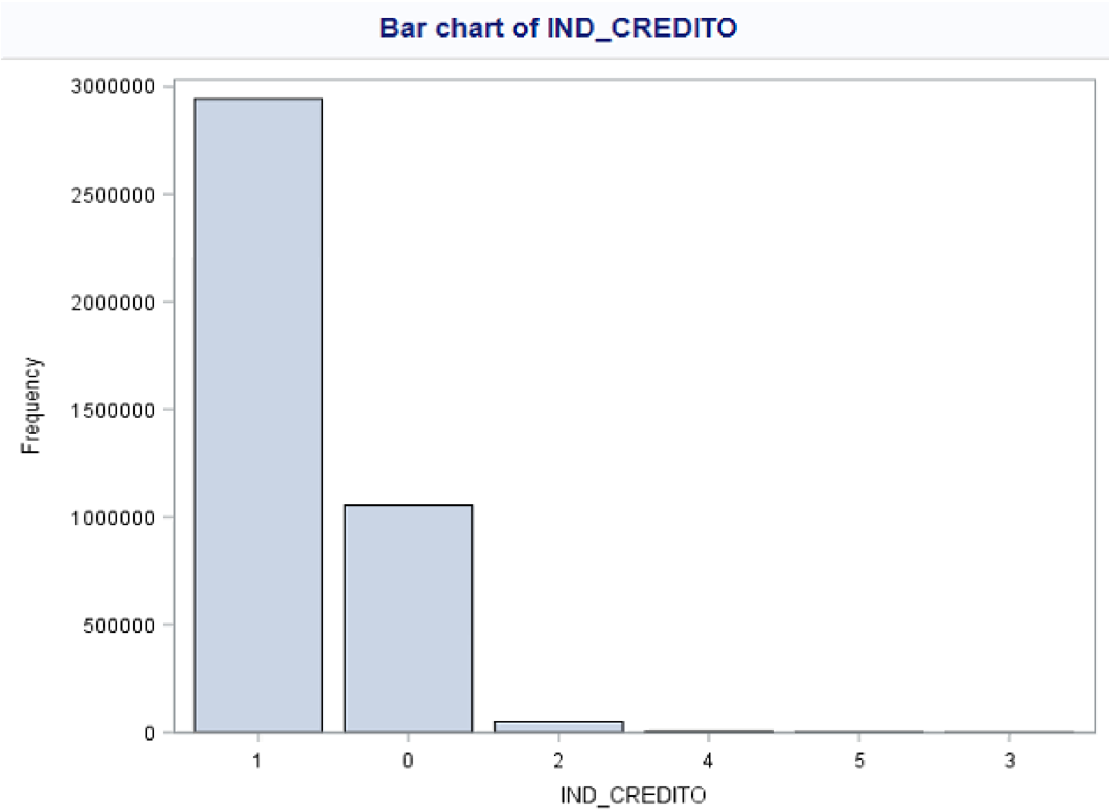

Figure 42 – Histogram and box plot of the payment incident indicator

As can be seen by the chart above, the majority of clients have a regular payment indicator. Furthermore, this variable, IND_CREDITO, holds the following values:

0. Missing value;
1. Regular credit;
2. Other indications, as long as delay in payment is ≤ 30 days;
3. Delays in payment > 30 days;
4. Restructured due to financial difficulties;
5. Default.

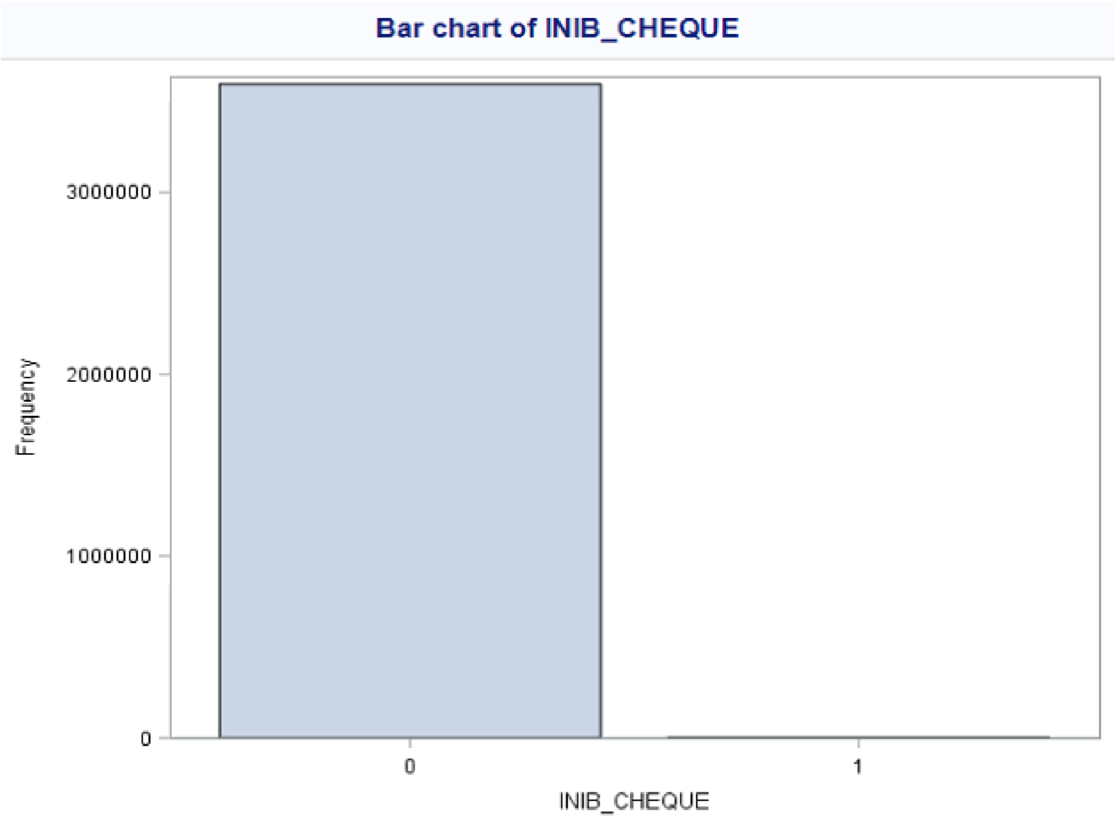

Figure 43 – Histogram and box plot of the check inhibition indicator

As can be seen by the chart above, most clients do not have a check inhibition.

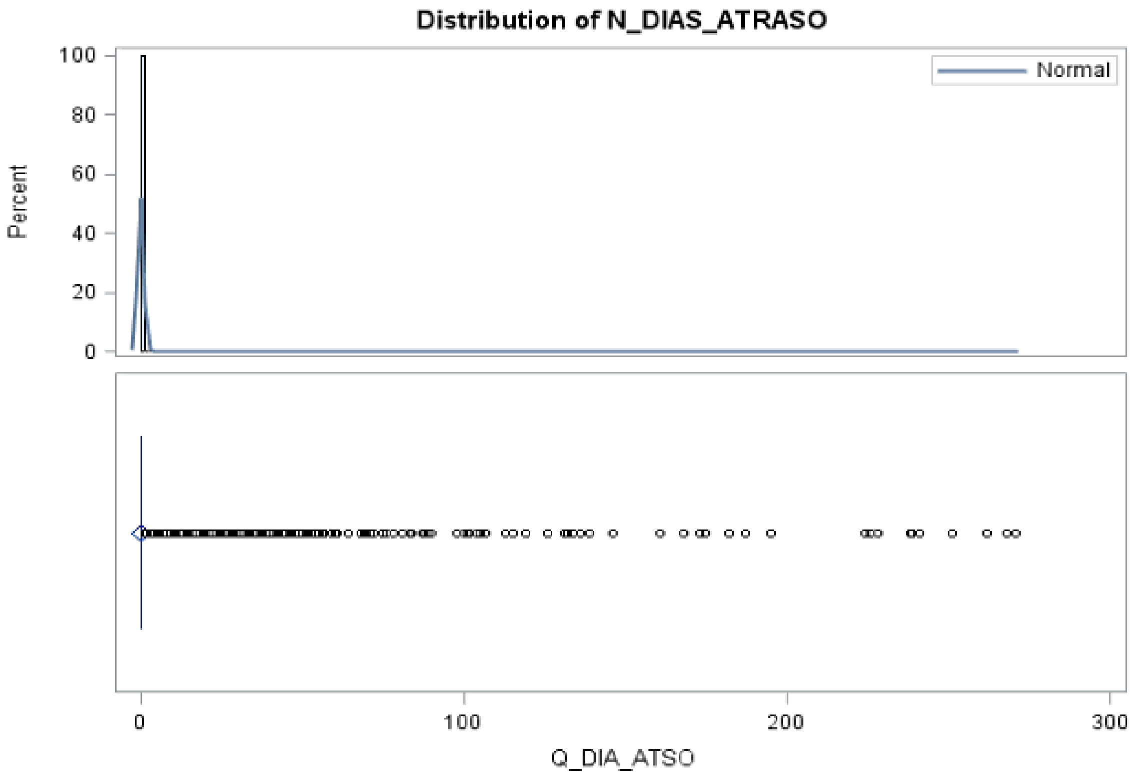

Figure 44 – Histogram and box plot of the number of days past due

As can be seen by the chart above, the large majority of clients does not have any days overdue.

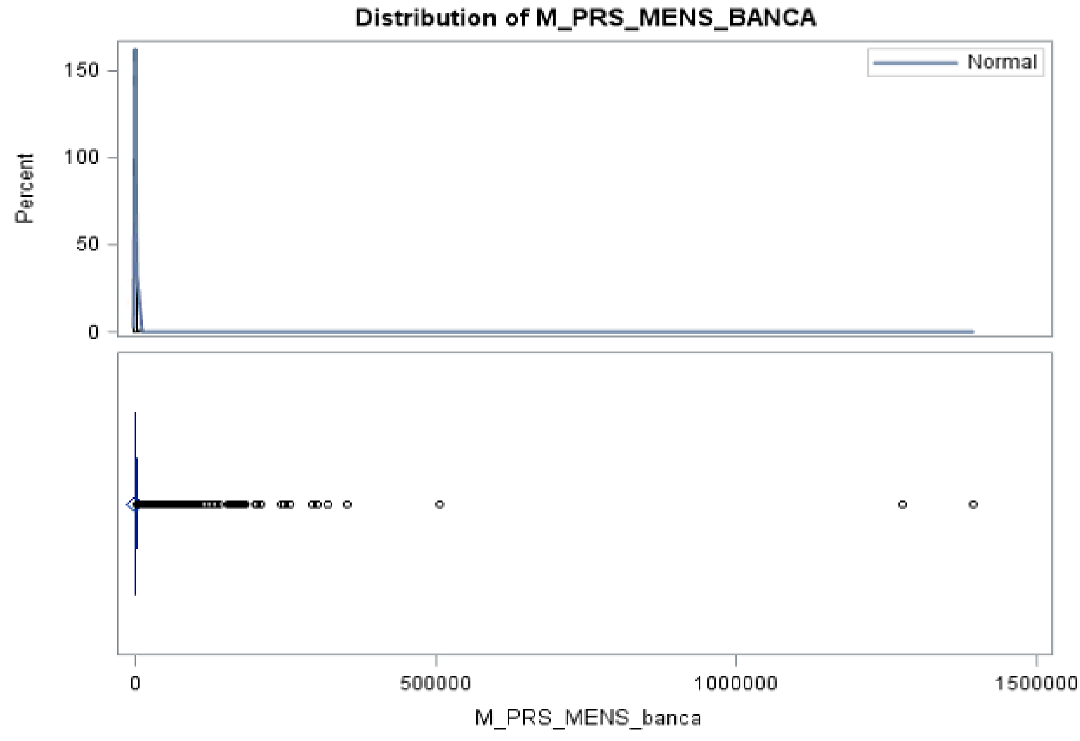

Figure 45 – Histogram and box plot of the monthly instalment in the financial system

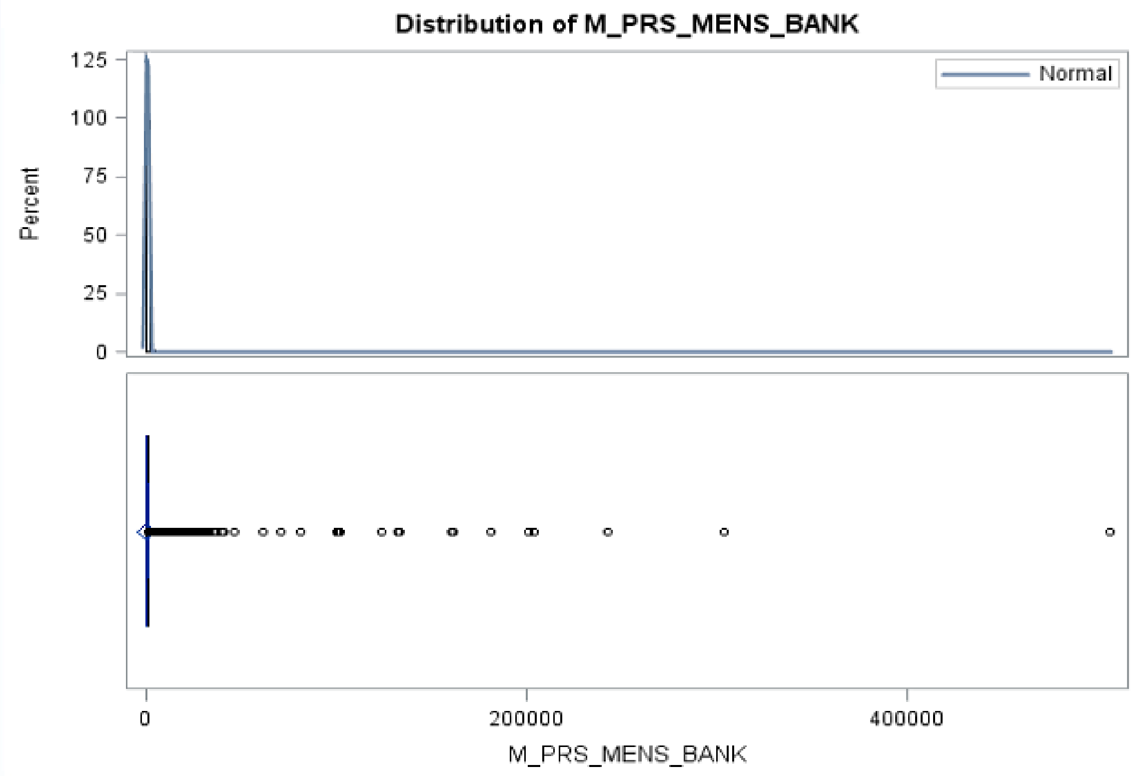

Figure 46 – Histogram and box plot of the monthly instalment in the bank

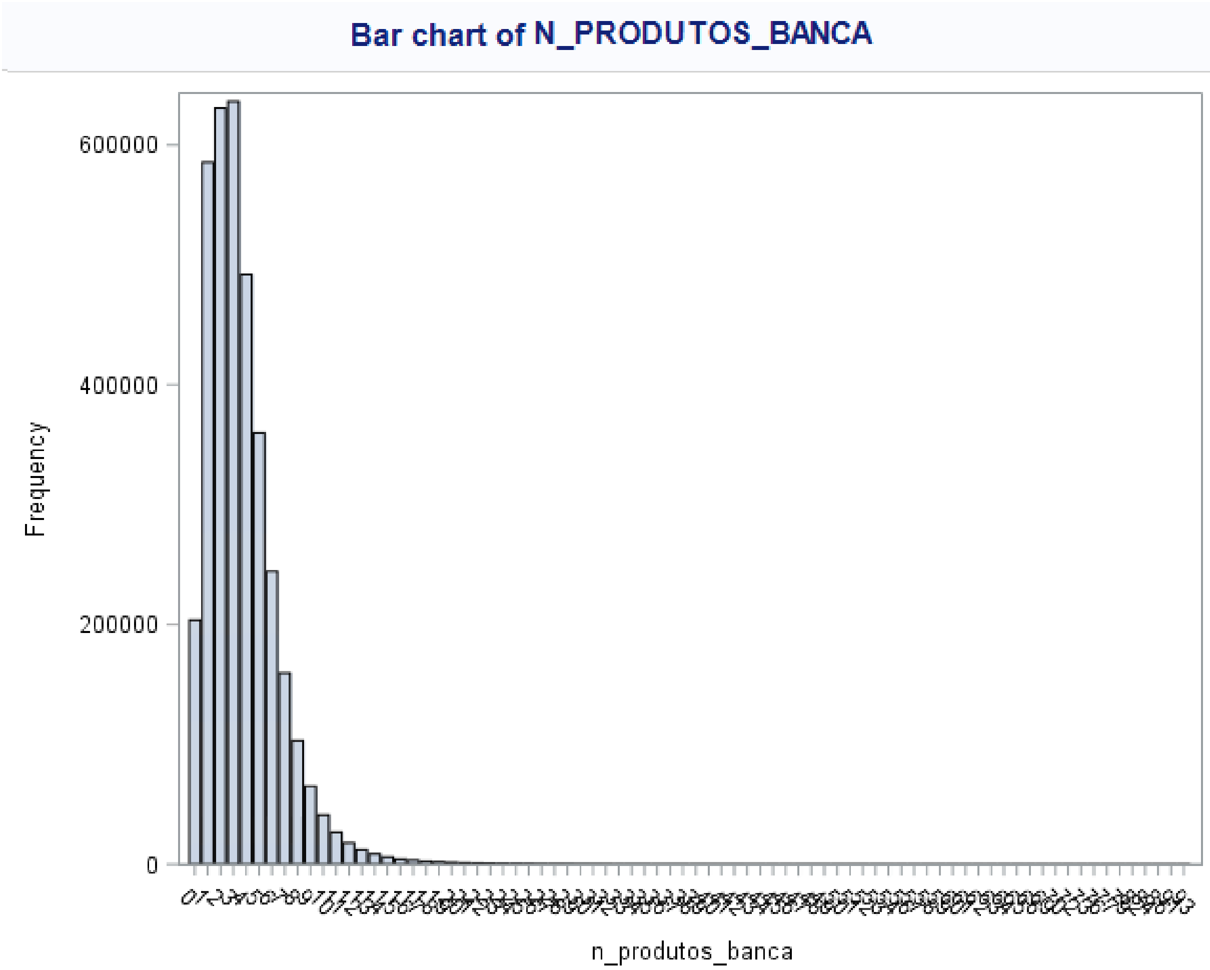

Figure 47 – Histogram and box plot of the number of products in the financial system

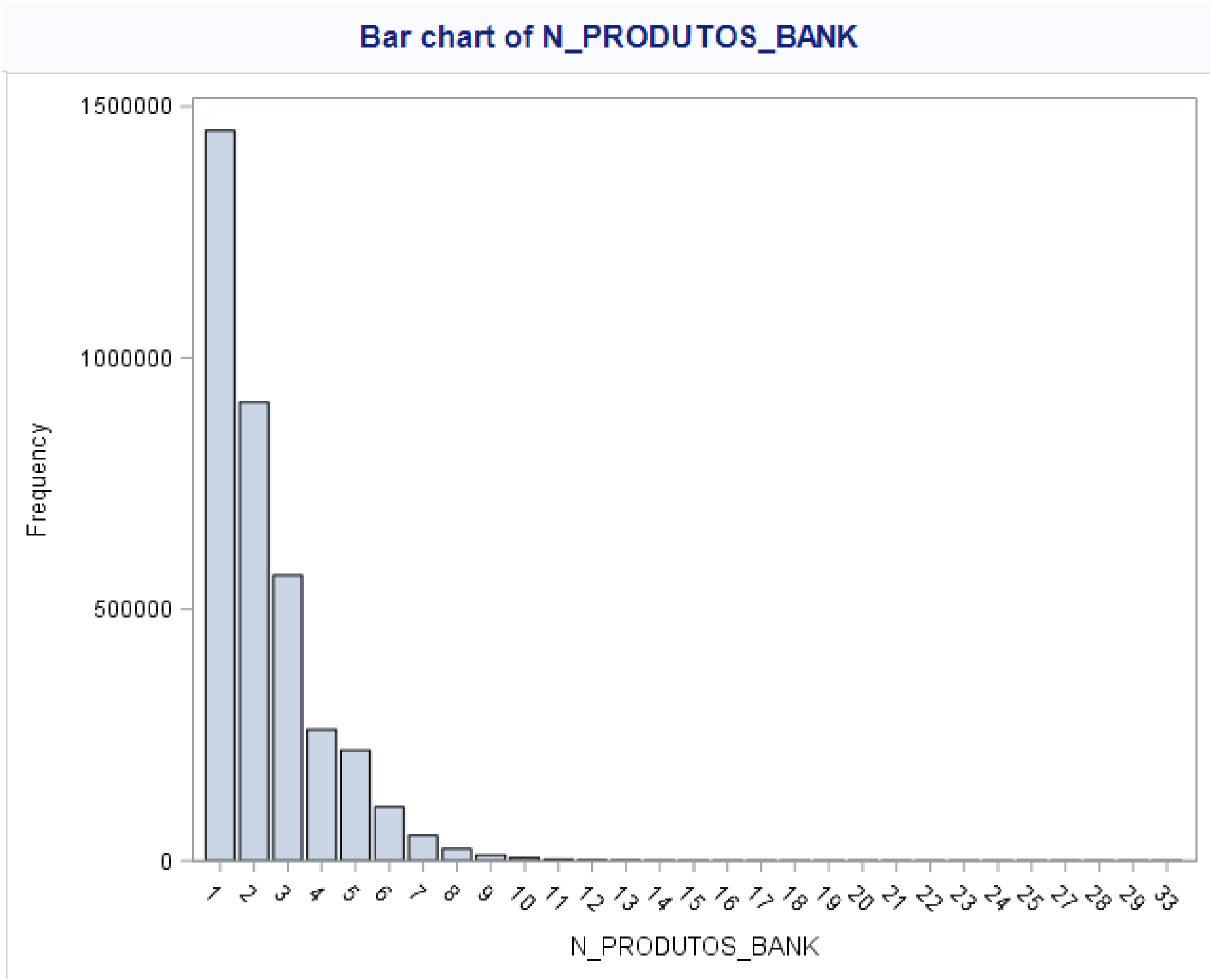

Figure 48 – Histogram and box plot of the number of products in the bank

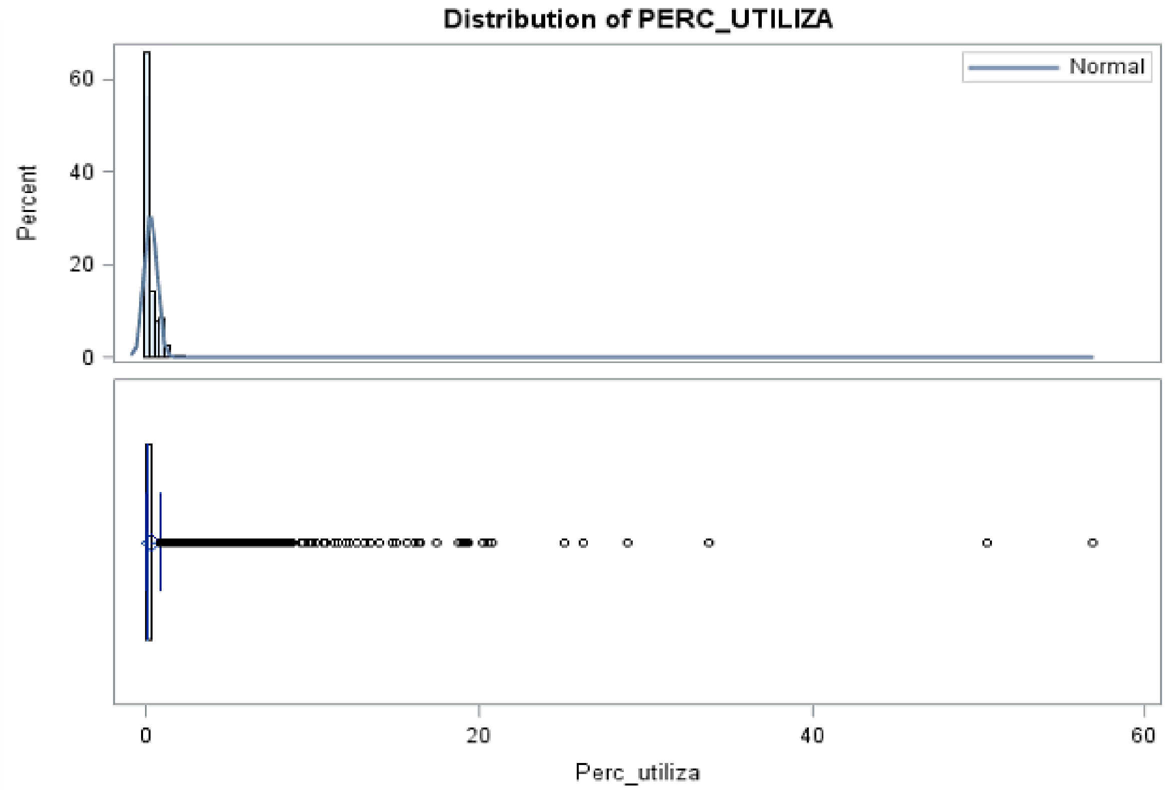

Figure 49 – Histogram and box plot of the percentage of credit card usage

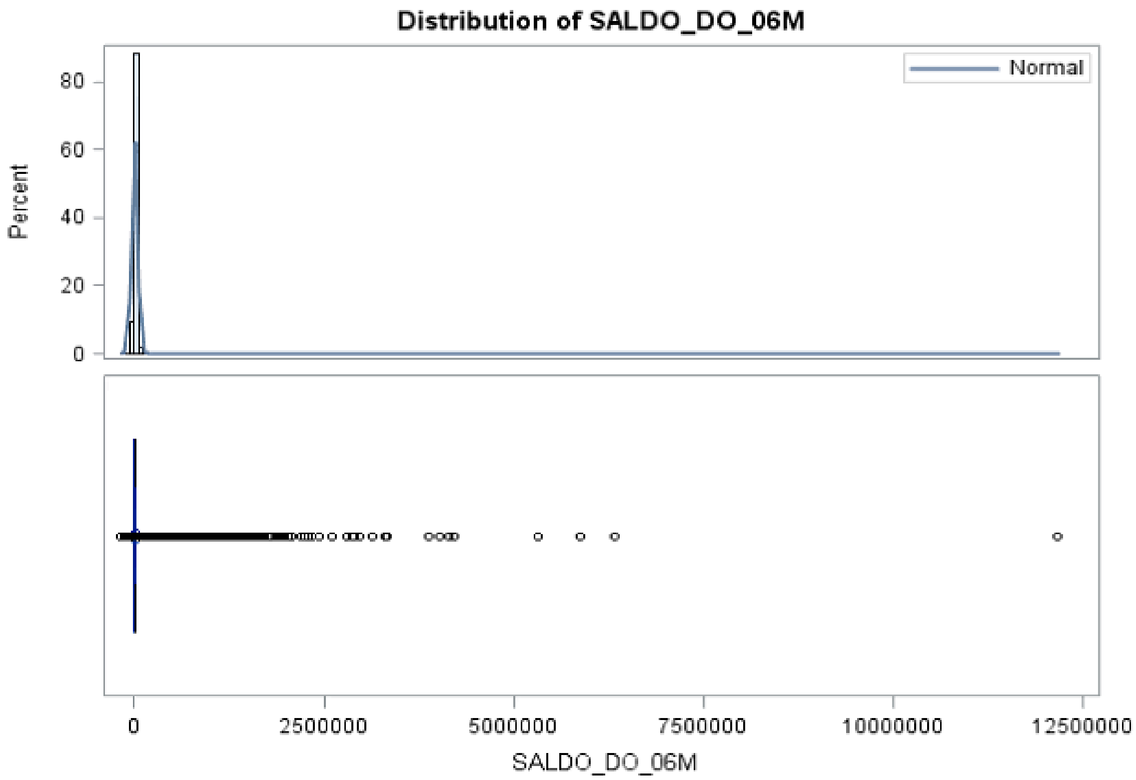

Figure 50 – Histogram and box plot of the balance in sight deposits, 6 months

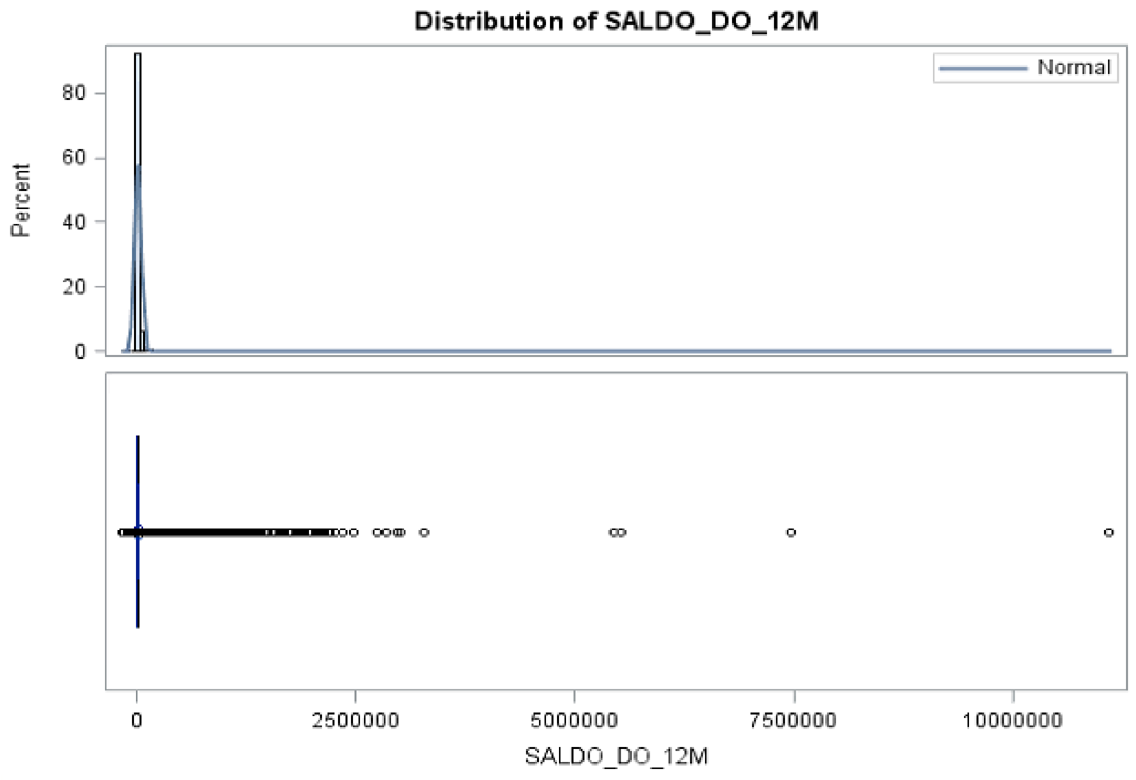

Figure 51 – Histogram and box plot of the balance in sight deposits, 12 months

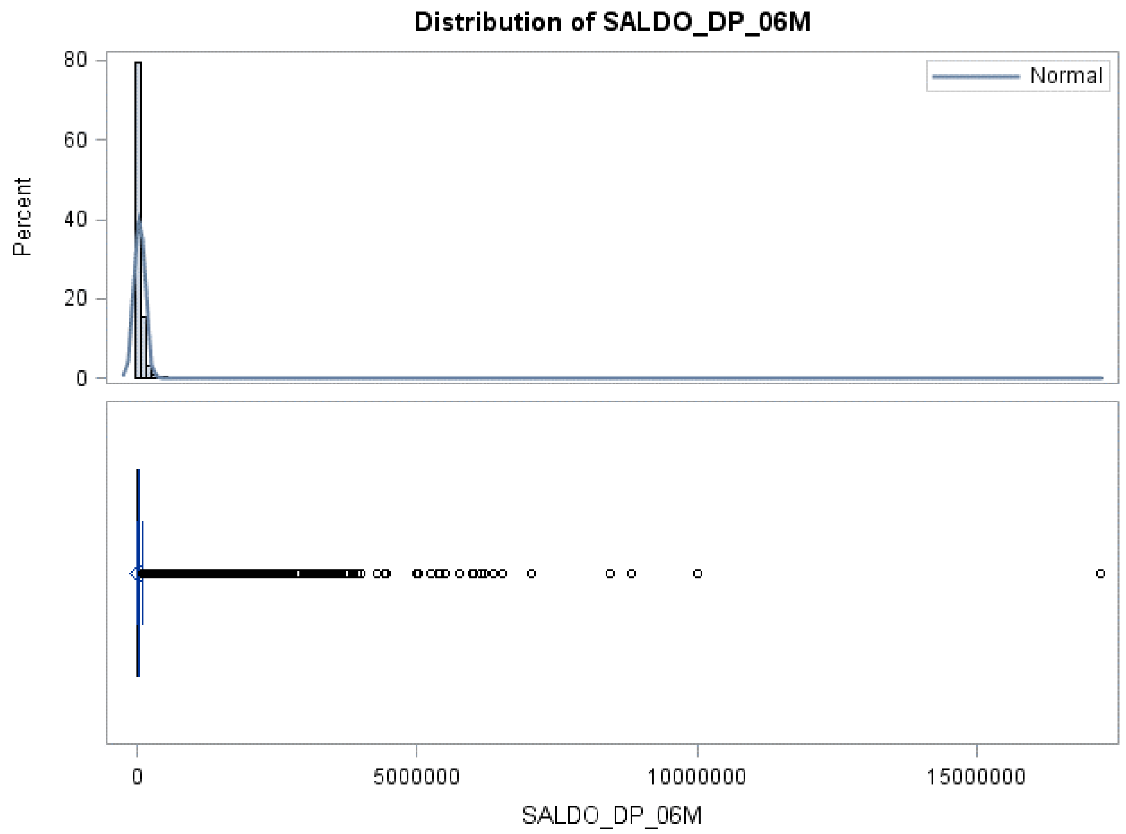

Figure 52 – Histogram and box plot of the balance in term deposits, 6 months

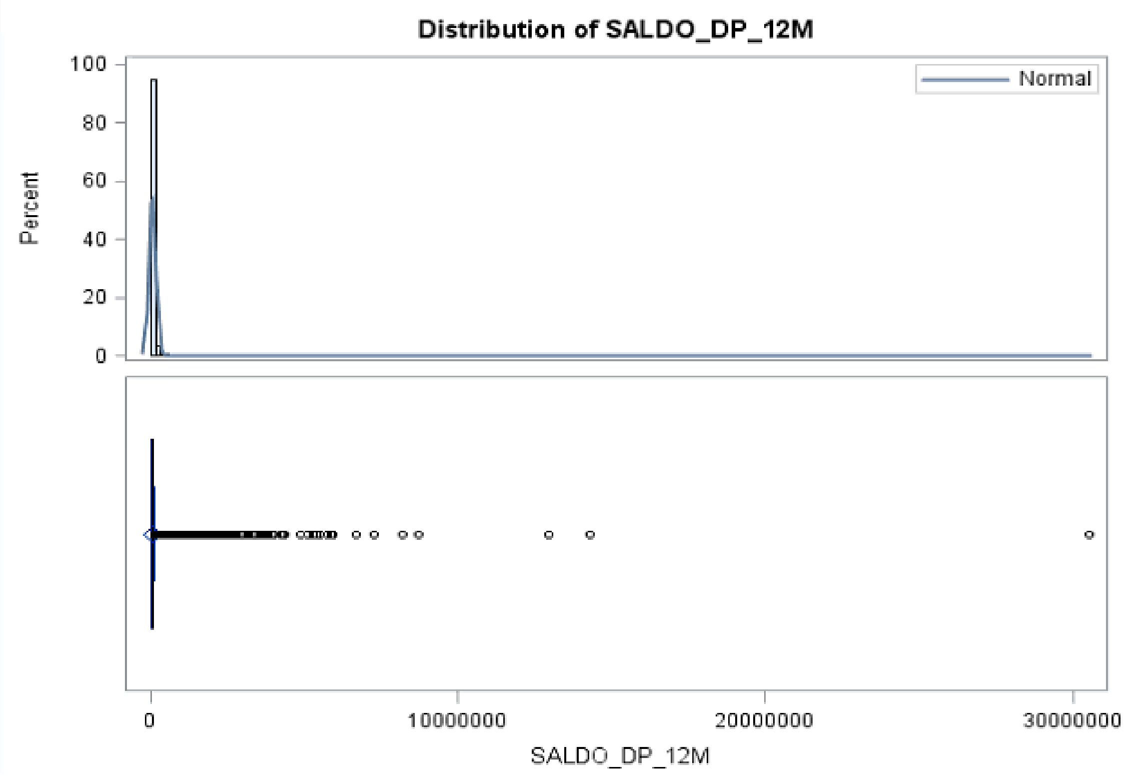

Figure 53 – Histogram and box plot of the balance in term deposits, 12 months

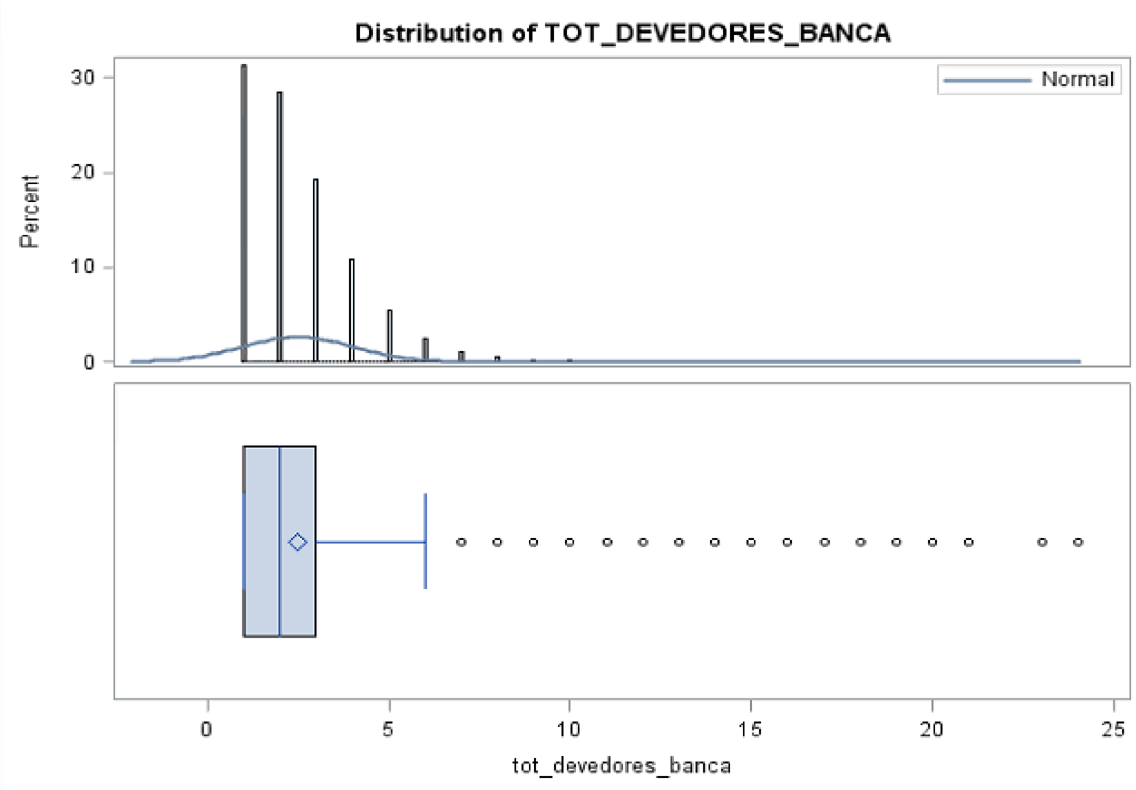

Figure 54 – Histogram and box plot of the debtors in the national financial system

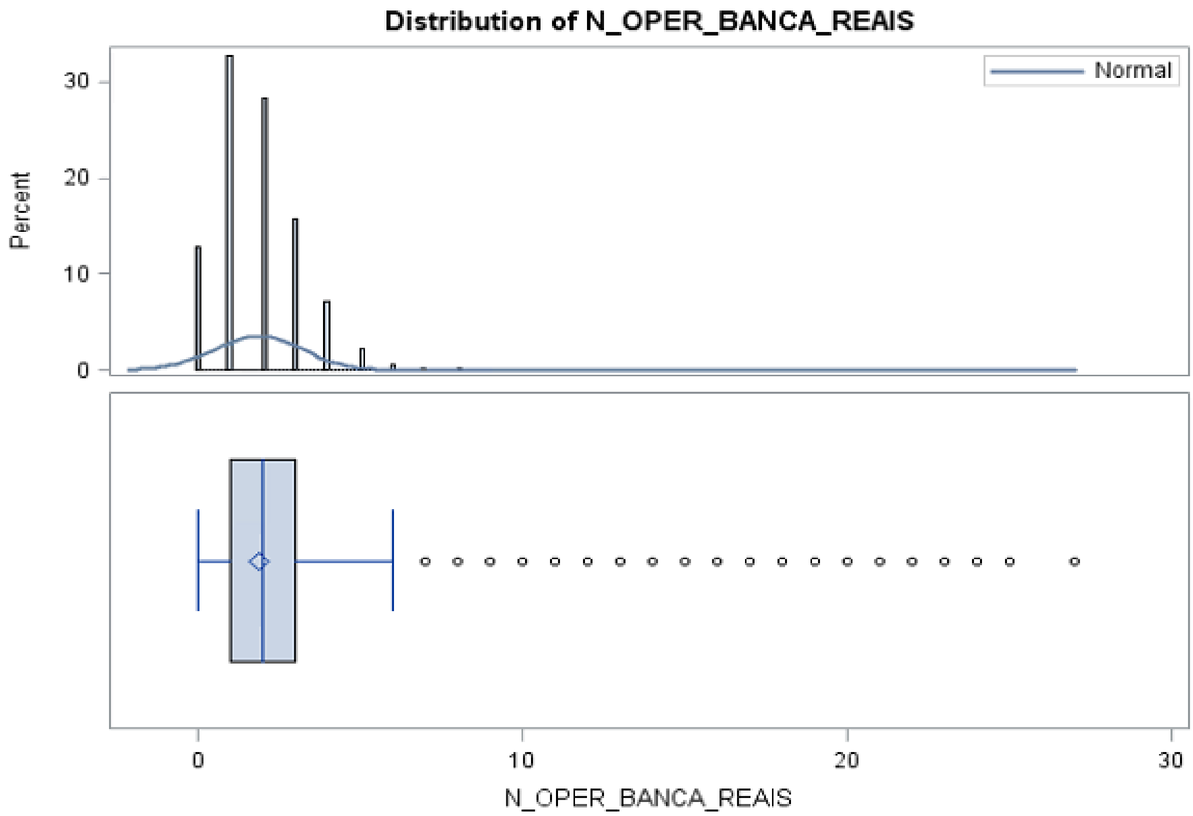

Figure 55 – Histogram and box plot of the real operations in the national financial system

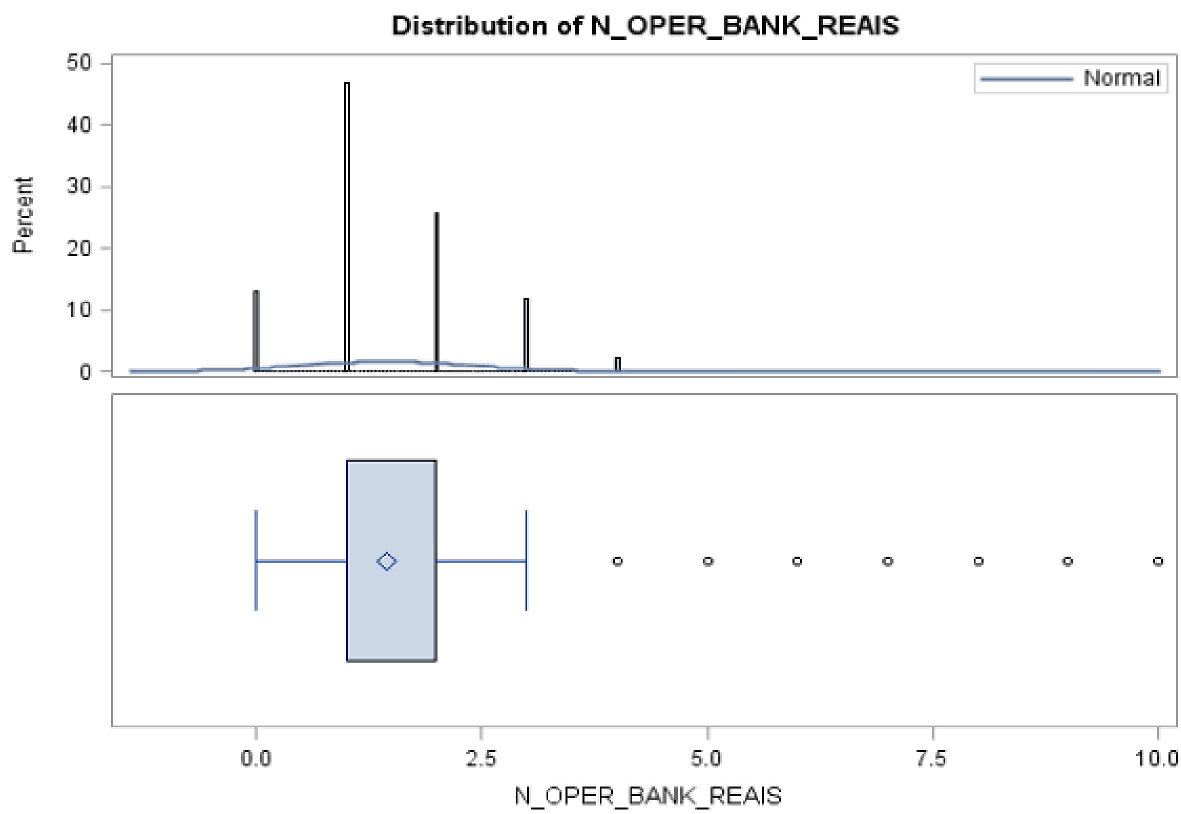

Figure 56 – Histogram and box plot of the real operations in the bank

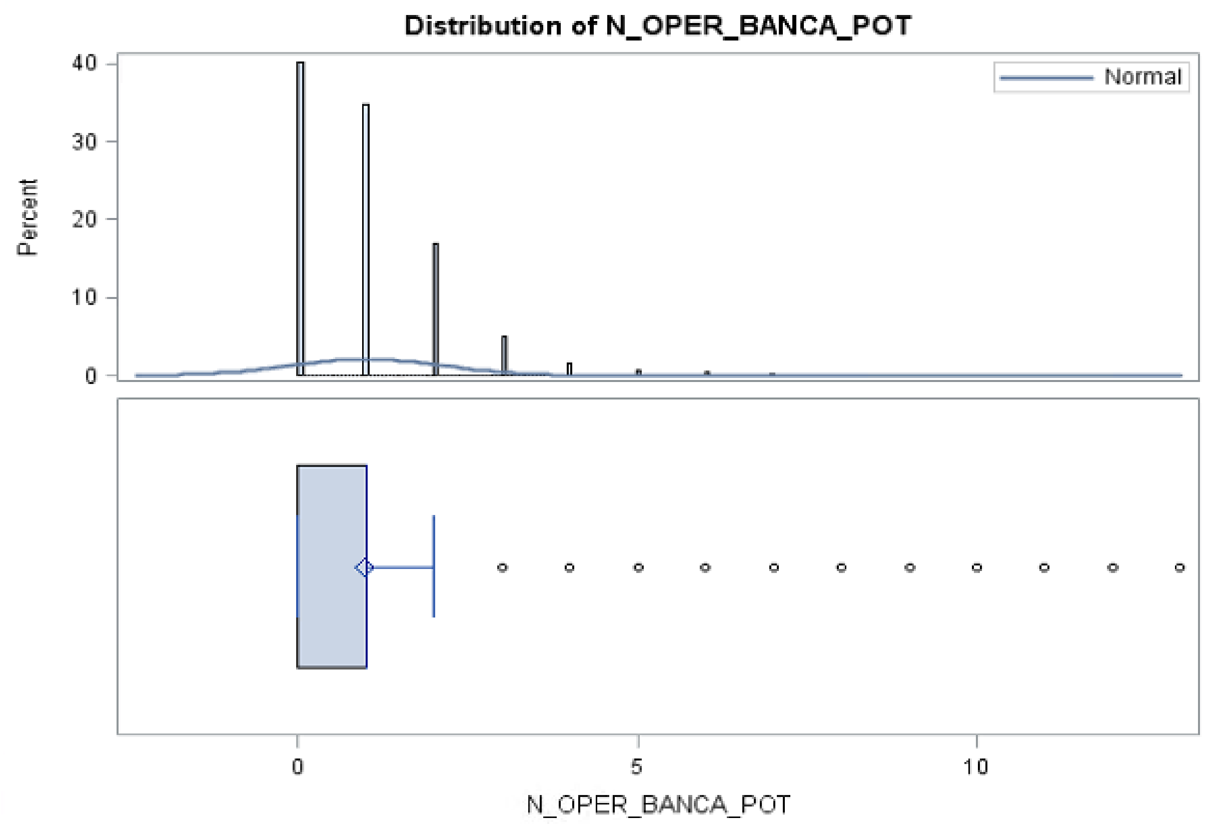

Figure 57 – Histogram and box plot of the potential operations in the national financial system

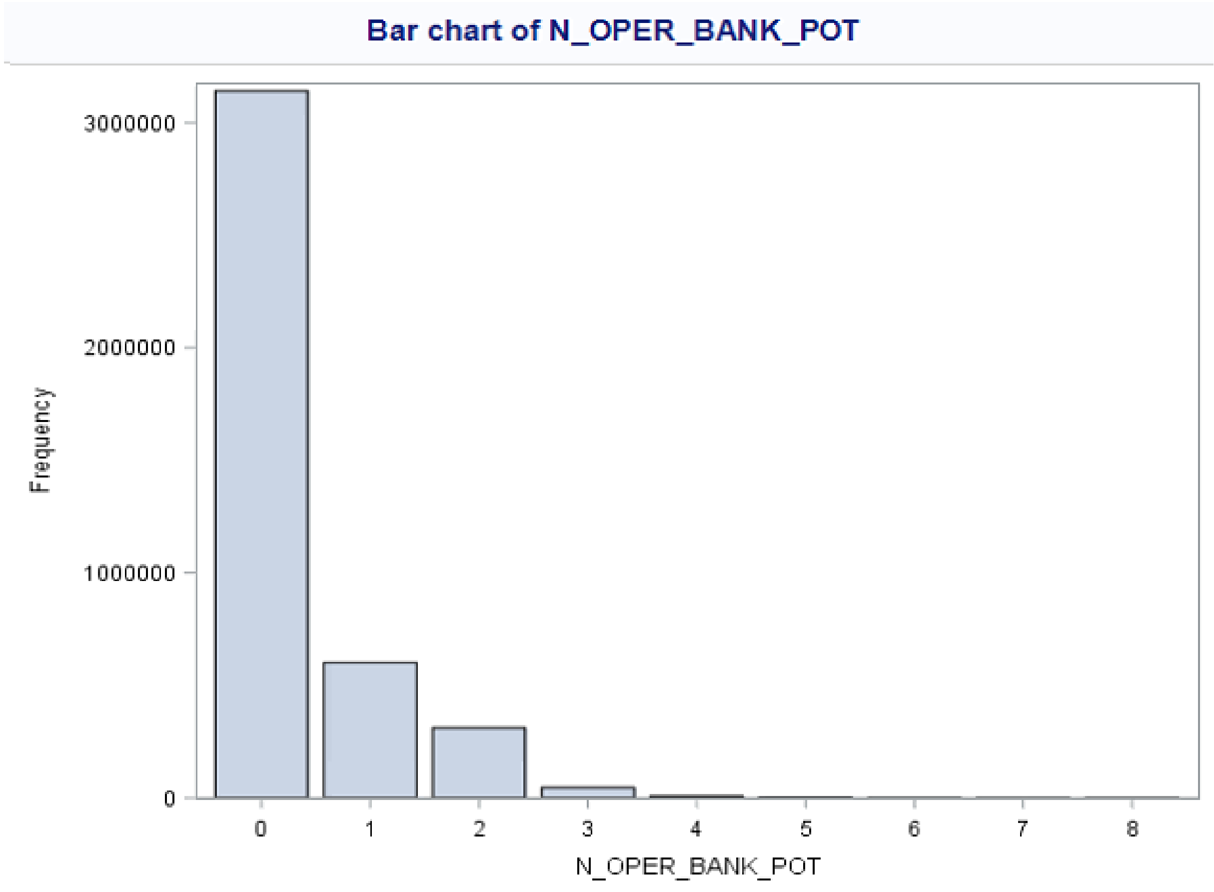

Figure 58 – Histogram and box plot of the potential operations in the bank

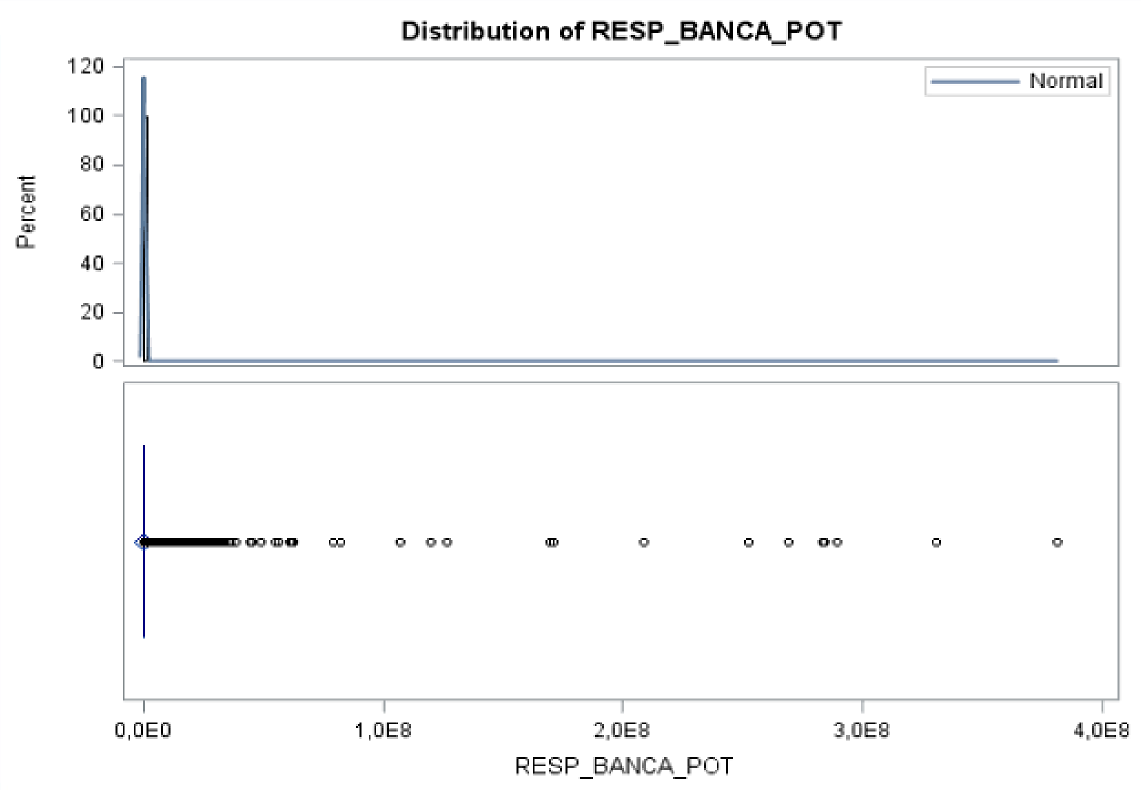

Figure 59 – Histogram and box plot of the amount of potential credit in the national financial system

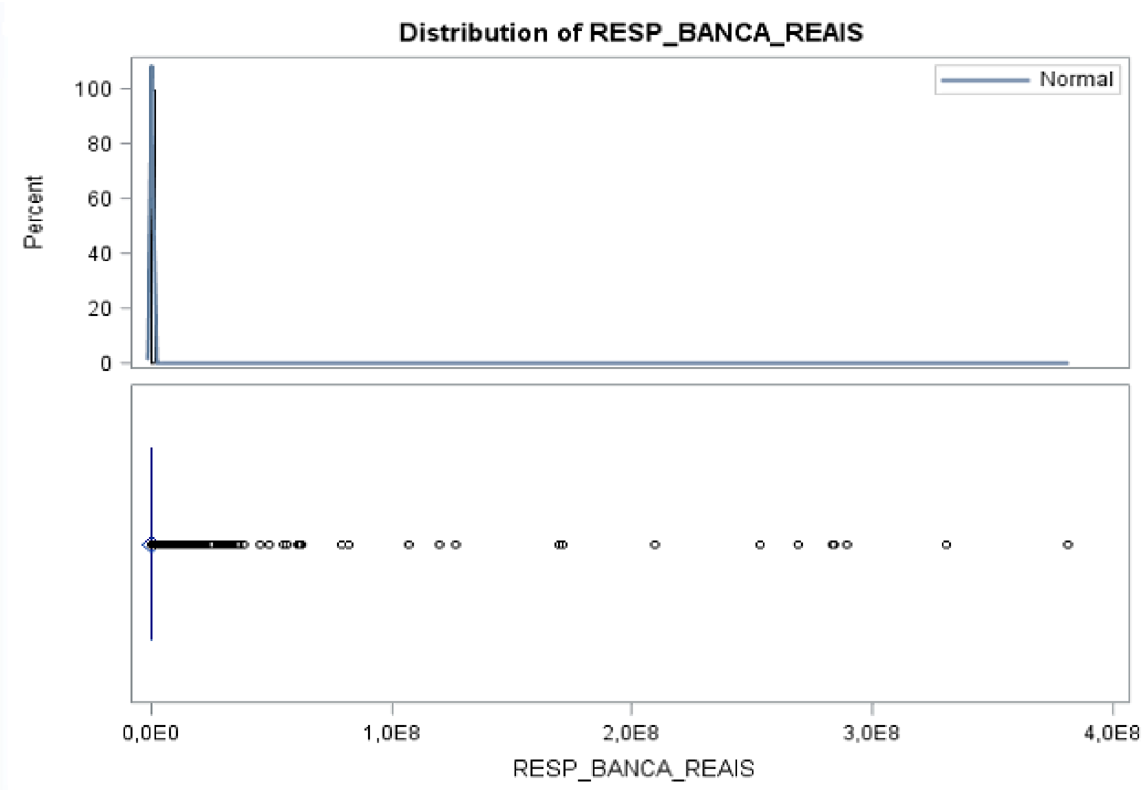

Figure 60 – Histogram and box plot of the amount of real credit in the national financial system

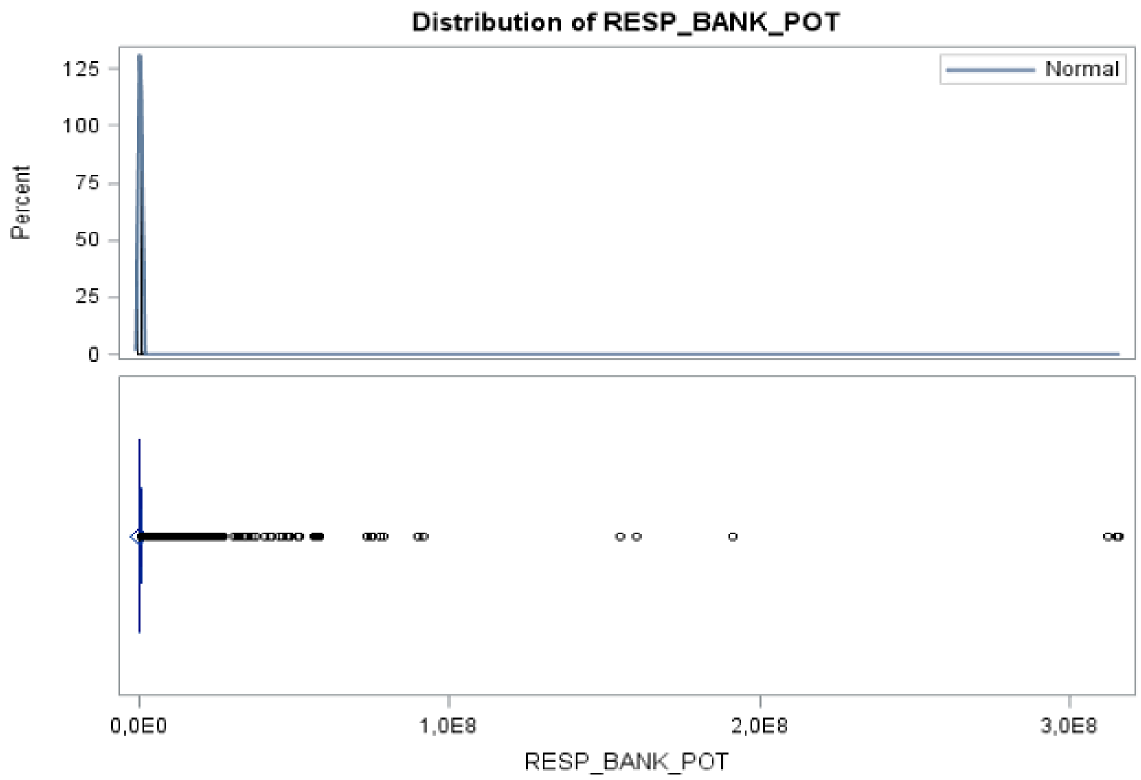

Figure 61 – Histogram and box plot of the amount of potential credit in the bank

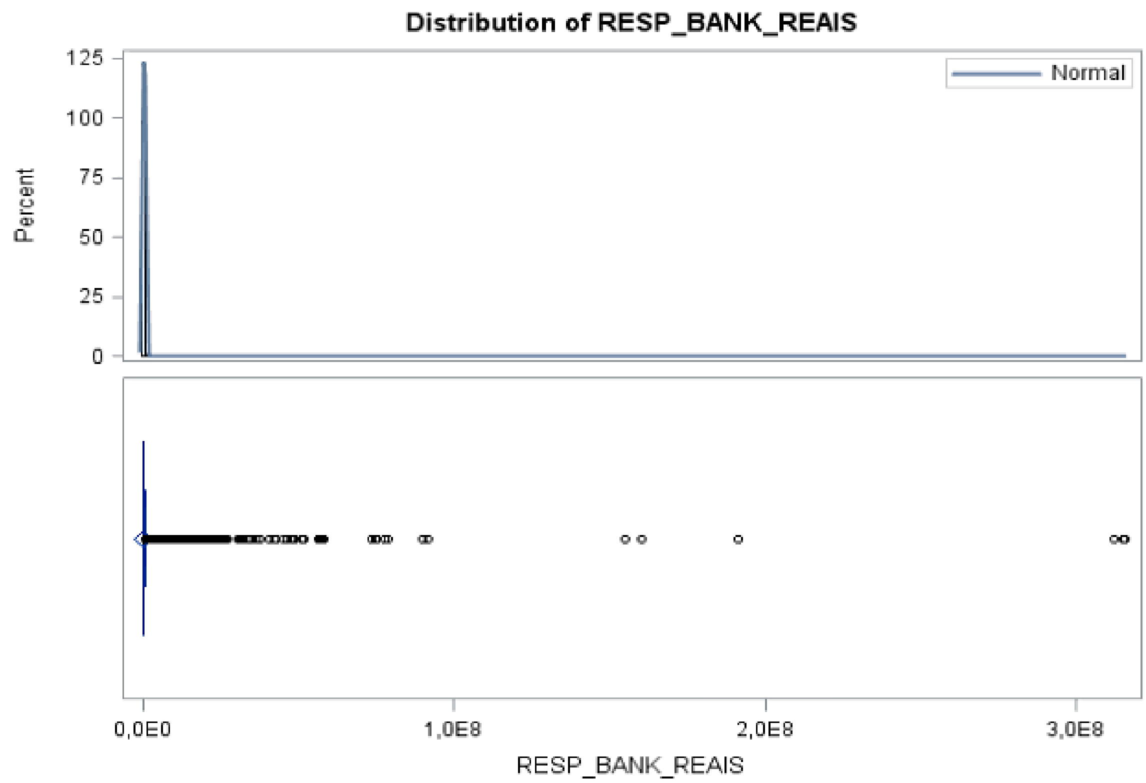

Figure 62 – Histogram and box plot of the amount of real credit in the bank

The variables shown above are all highly affected by the outliers presented.

Appendix 4. DATA PRE-PROCESSING – POST-WINDSORIZING AND IMPUTE NODE

After performing the Data Cleaning, as per chapter 3.1.1, i.e. after smoothing the outliers, using the windsorizing method, and imputing the missing values, the descriptive statistics, histograms, and bar graphs are as follows:

Numerical variables:

Variables	# Missing Values	%	Mean	Maximum	Minimum
DATA_ABERTURA	-	0.0%	12-02-1943	28-12-1957	25-05-1920
ED_LICENC_TVH	-	0.0%	-2.13%	29.87%	-25.06%
ENDIV_PART_TVH	-	0.0%	-2.21%	0.13%	-4.03%
GRAU_POUP_PART_TVH	-	0.0%	46.15%	230.77%	-86.15%
IDADE	-	0.0%	49	80	19
IND_COINC_TVH	-	0.0%	0.08%	2.90%	-6.40%
IND_PRECOS_HAB_TVH	-	0.0%	2.57%	12.24%	-8.17%
IND_SENT_ECO_TVH	-	0.0%	3.06%	27.39%	-18.65%
INIB_CHEQUE	-	0.0%	0	1	0
LTV_ATUAL	-	0.0%	1	10	0.0238262
LTV_ORIG	-	0.0%	1	11	0.024542
M_PRS_MENS_banca	-	0.0%	616	2 944	0
M_PRS_MENS_BANK	-	0.0%	601	2 862	0
MONTANTE_AMORT	-	0.0%	72	1 221 613	0
MONTANTE_FINANCIADO	-	0.0%	60 860	225 000	5 000
MONTANTE_RESIDUAL	-	0.0%	32 575	182 323	0
N_DIAS_ATRASO	-	0.0%	0	296	0
N_FOGOS_CONST_TVH	-	0.0%	1.96%	84.98%	-48.06%
N_OPER_BANCA_POT	-	0.0%	1	5	-
N_OPER_BANCA_REAIS	-	0.0%	2	6	1
N_OPER_BANK_POT	-	0.0%	0	3	-
N_OPER_BANK_REAIS	-	0.0%	2	4	1
N_PREST_PAGAS	-	0.0%	137	449	-1
N_PRODUTOS_BANCA	-	0.0%	4	14	1
N_PRODUTOS_BANK	-	0.0%	2	8	1
PERC_PRAZO	-	0.0%	1	1	-
PERC_UTILIZA	-	0.0%	0	2	0

Variables	# Missing Values	%	Mean	Maximum	Minimum
PERSP_SIT_EC	-	0.0%	-18.81%	10.00%	-59.80%
PIB	-	0.0%	0.50%	3.60%	-3.60%
PRAZO	-	0.0%	348	720	24
PRAZO_RESIDUAL	-	0.0%	206	652	0
RENDIMENTO	-	0.0%	18 250	98 764	0
RESP_BANCA_POT	-	0.0%	108 868	788 230	-
RESP_BANCA_REAIS	-	0.0%	140 728	840 376	244
RESP_BANK_POT	-	0.0%	102 435	788 230	-
RESP_BANK_REAIS	-	0.0%	125 314	840 376	244
SALDO_DO_06M	-	0.0%	7 127	98 307	- 2 738
SALDO_DO_12M	-	0.0%	6 977	92 672	- 2 566
SALDO_DP_06M	-	0.0%	16 441	383 000	-
SALDO_DP_12M	-	0.0%	16 116	374 568	-
SCORING	-	0.0%	5	10	1
T_JURO	-	0.0%	2	8	-
T_SPREAD	-	0.0%	1	4	0
TAXA_INFLACAO_TVH	-	0.0%	0.93%	3.30%	-0.40%
TAXA_JURO_DP_TVH	-	0.0%	-27.60%	38.89%	-55.83%
TAXA_JURO_HAB_TVH	-	0.0%	-6.01%	59.18%	-29.91%
tot_devedores_banca	-	0.0%	2	7	1
TOTAL_AMORT_PARCIAL	-	0.0%	0	6	-
TX_DESEMPREGO_TVH	-	0.0%	-4.35%	21.77%	-22.12%
Z_FIM_CTTO	-	0.0%	11-02-1972	17-12-2007	31-12-1952

Table 22 - Statistical descriptions of numerical variables

Categorical variables:

Variables	# Missing Values	%	# Unique Values	Mode
ESTADO_CIVIL	-	0.0%	5	2 (Married/De facto Union)
PROFISSAO	-	0.0%	12	3 (Specialists in intellectual and scientific activities)
FINALIDADE	-	0.0%	12	1 (Acquisition permanent home)

Variables	# Missing Values	%	# Unique Values	Mode
IND_CREDITO	-	0.00	6	1 (Regular credit)

Table 23 - Statistical descriptions of categorical variables

<u>Comparison of before and after histograms and bar charts</u>. Note that are only shown the variables where there is a change in the values, i.e. where it was performed some kind of data transformation:

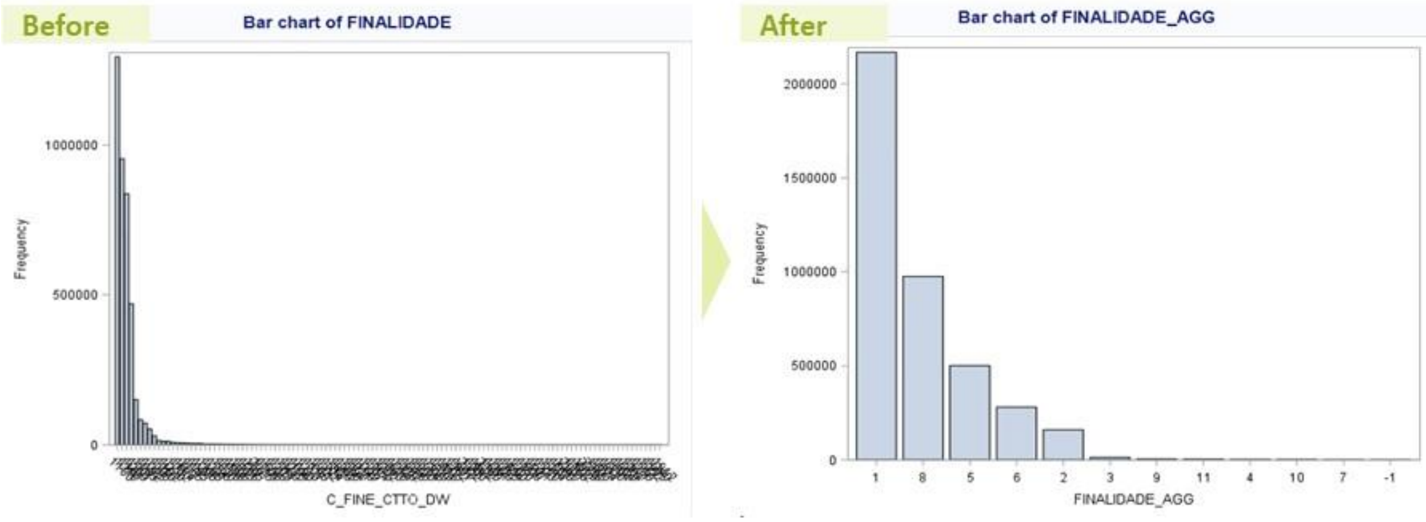

Figure 63 – Comparison of the loan purpose before and after conversion

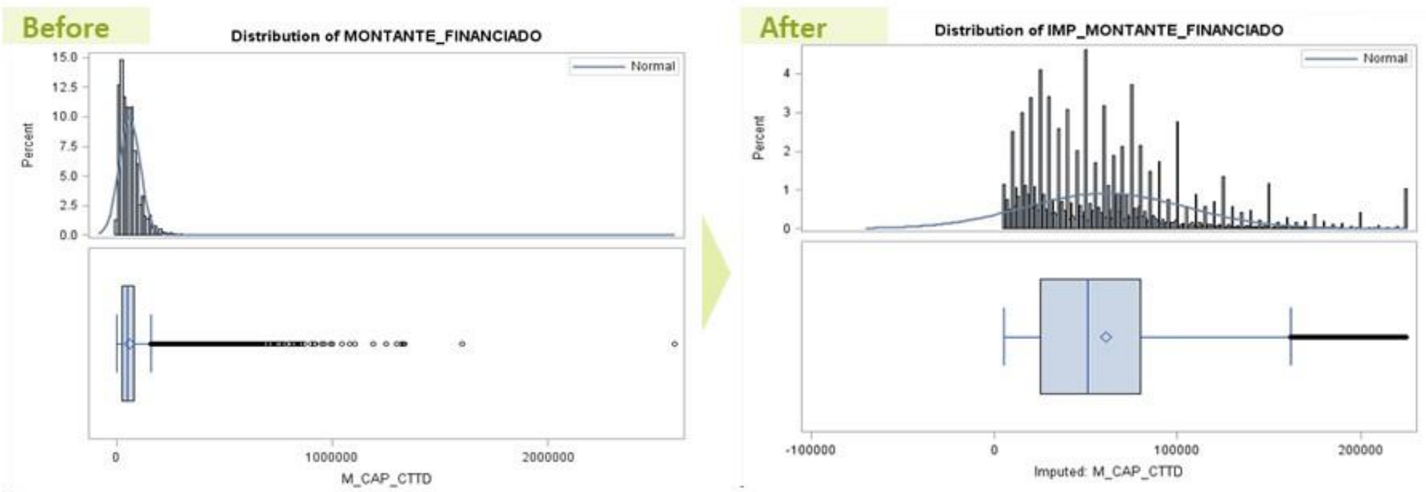

Figure 64 – Comparison of the financed amount before and after impute and outlier smoothing

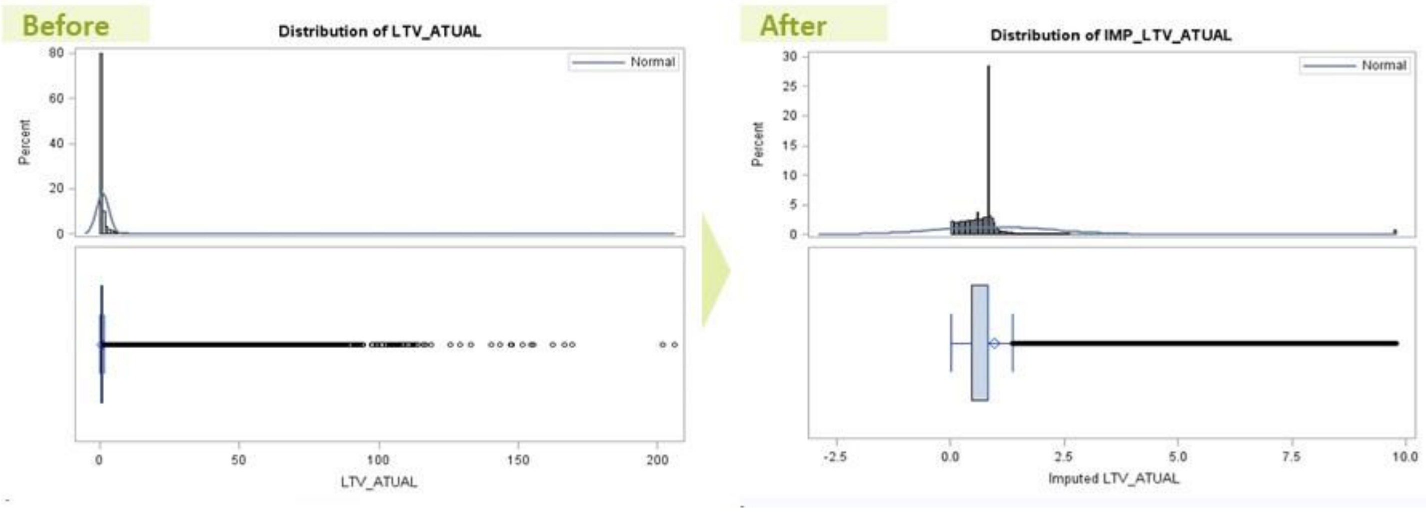

Figure 65 – Comparison of the current LTV before and after impute and outlier smoothing

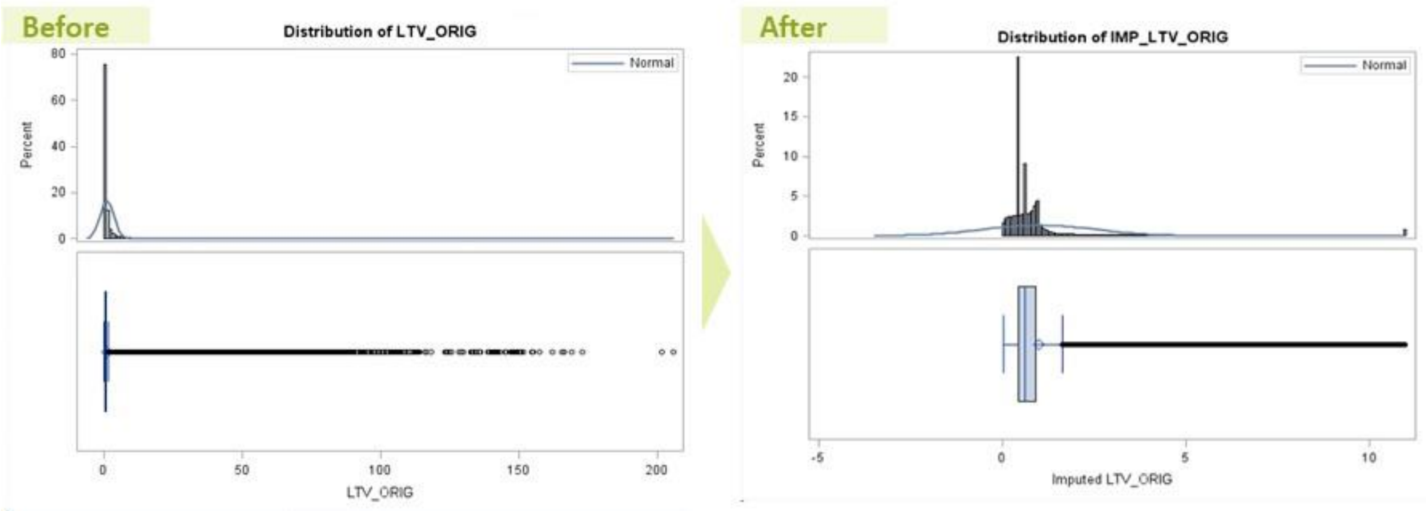

Figure 66 – Comparison of the origination LTV before and after impute and outlier smoothing

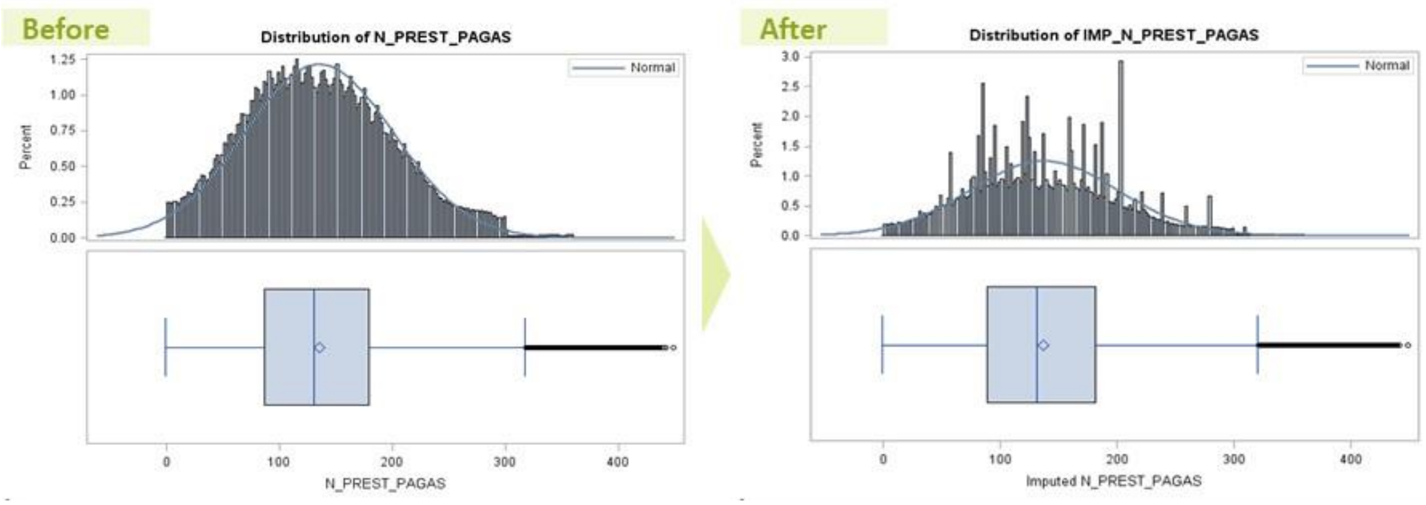

Figure 67 – Comparison of the number of paid instalments before and after impute

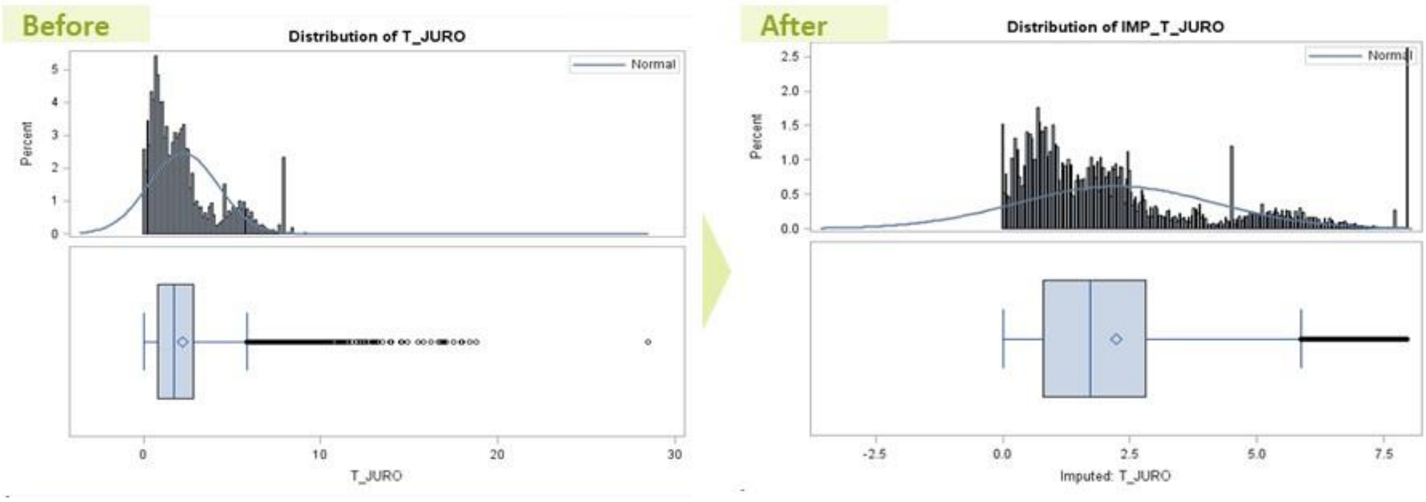

Figure 68 – Comparison of the interest rate before and after impute and outlier smoothing

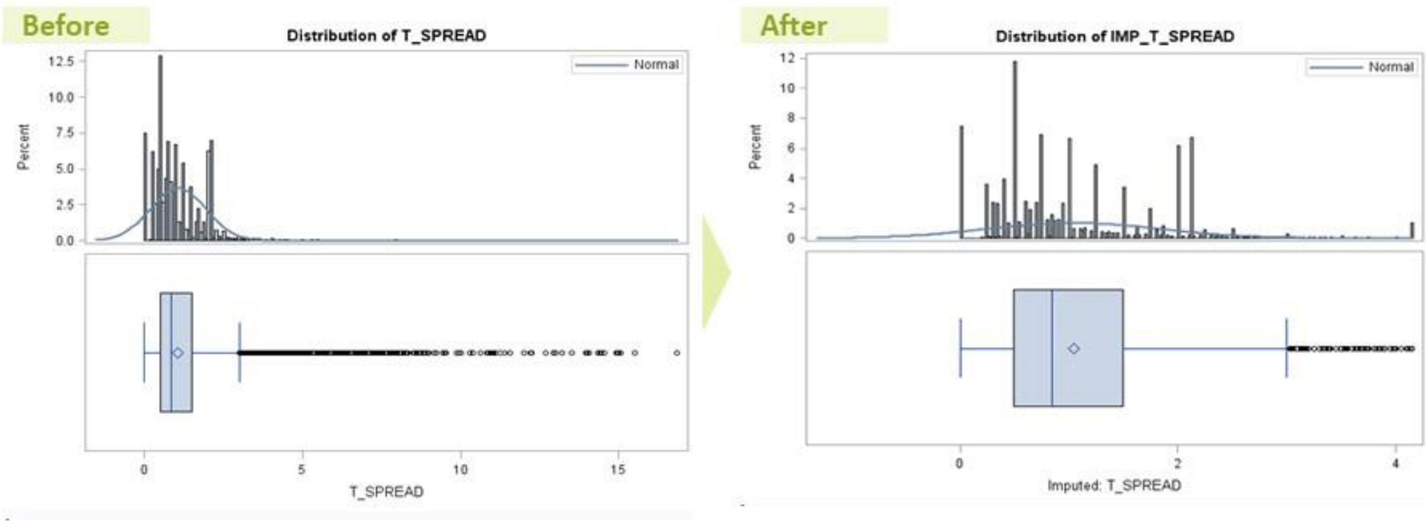

Figure 69 – Comparison of the spread rate before and after impute and outlier smoothing

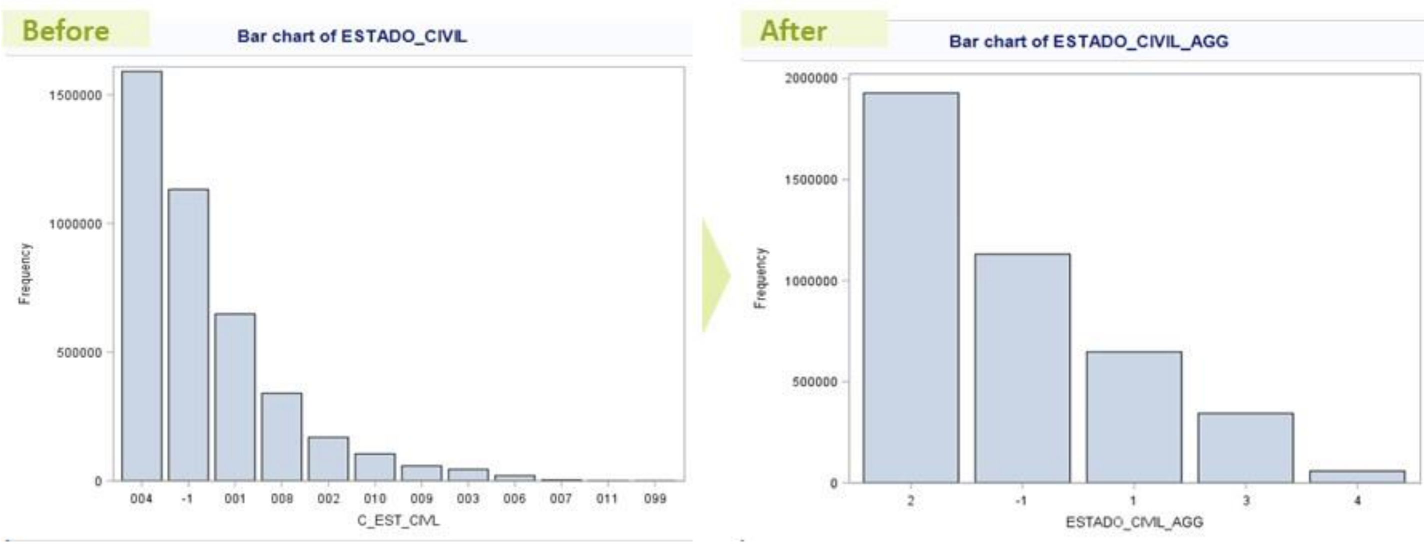

Figure 70 – Comparison of the marital status before and after conversion

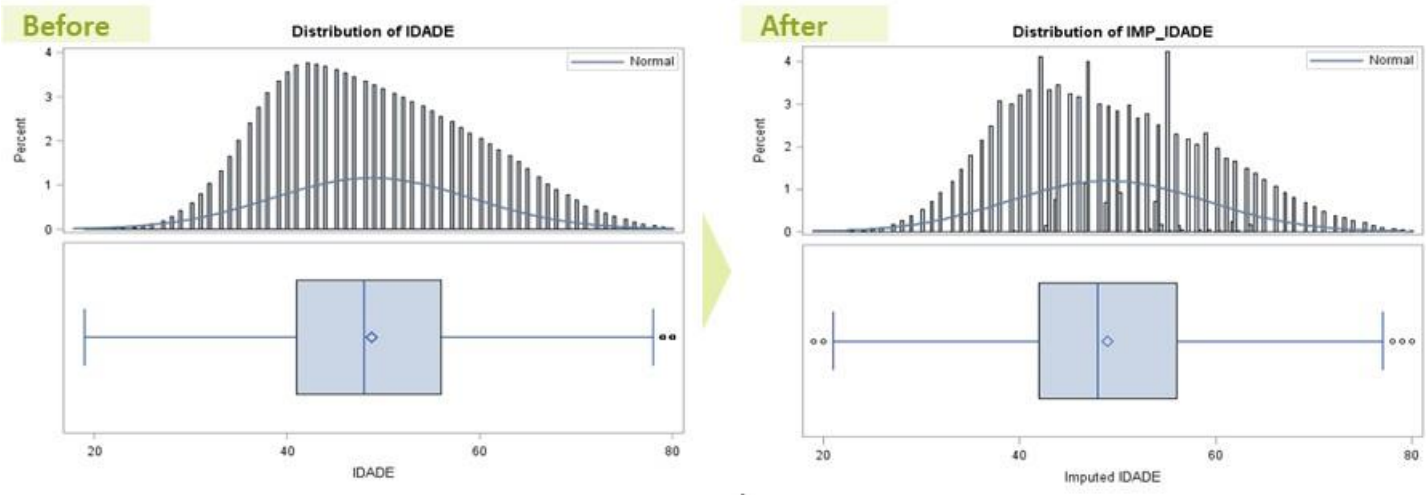

Figure 71 – Comparison of the age before and after impute

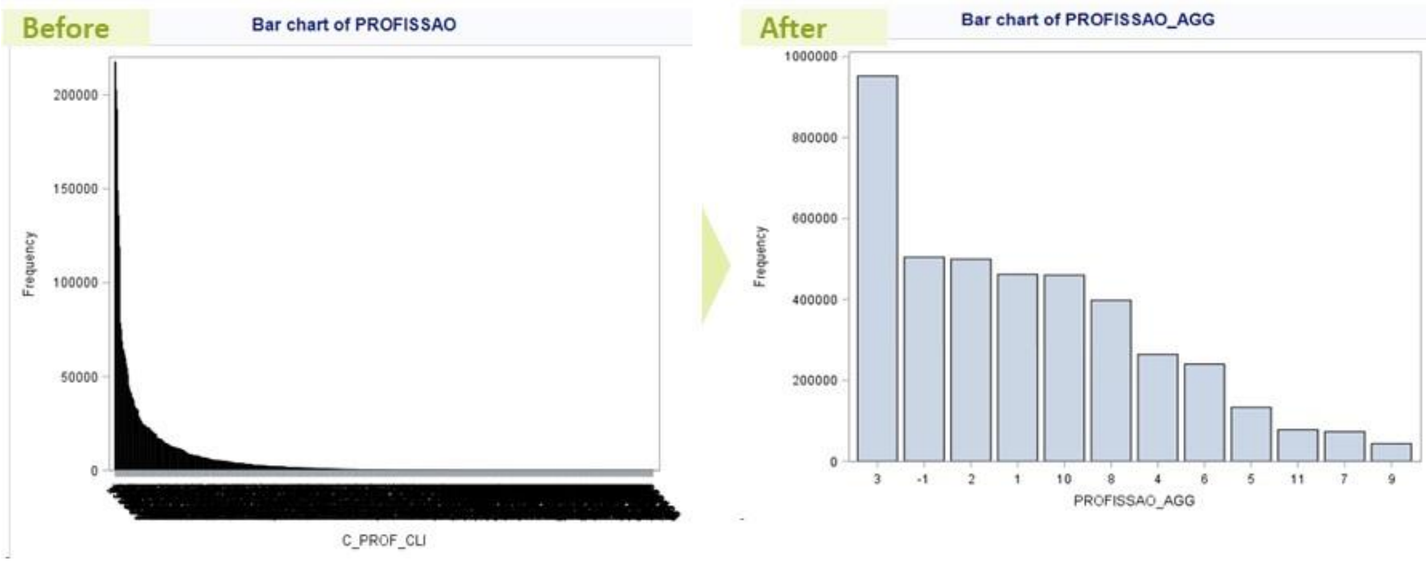

Figure 72 – Comparison of the profession before and after conversion

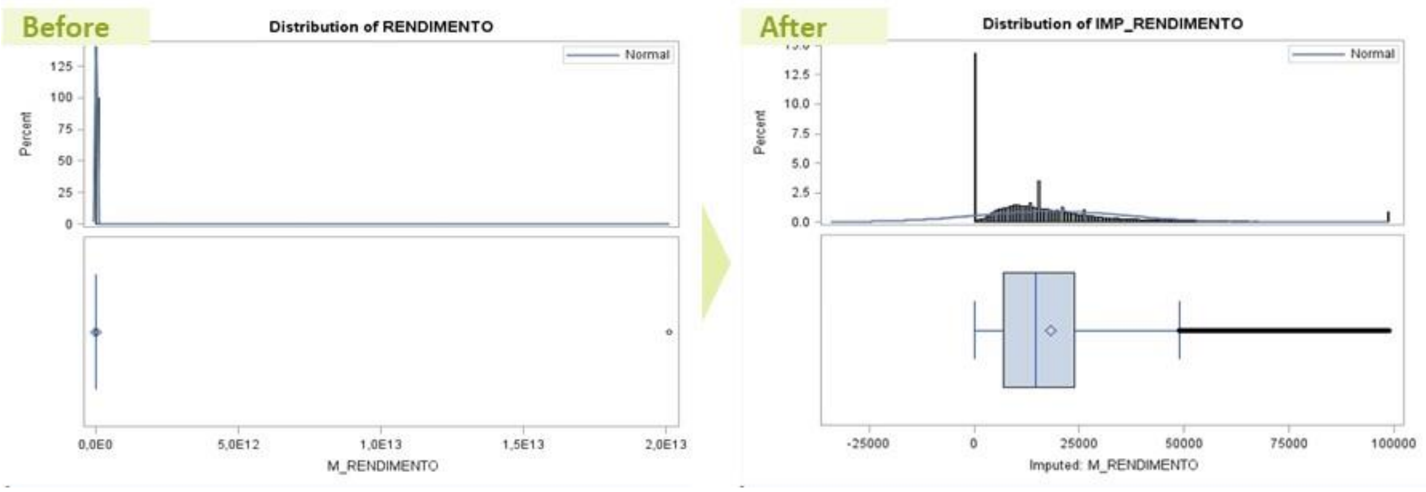

Figure 73 – Comparison of the yearly income before and after impute and outlier smoothing

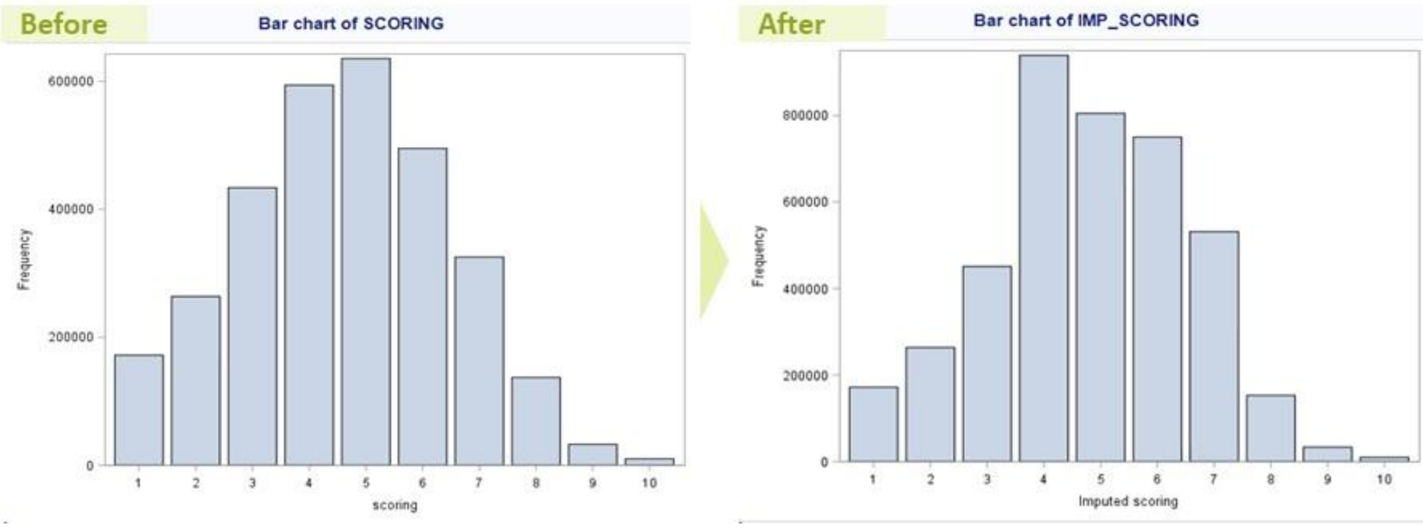

Figure 74 – Comparison of the scoring before and after impute and outlier smoothing

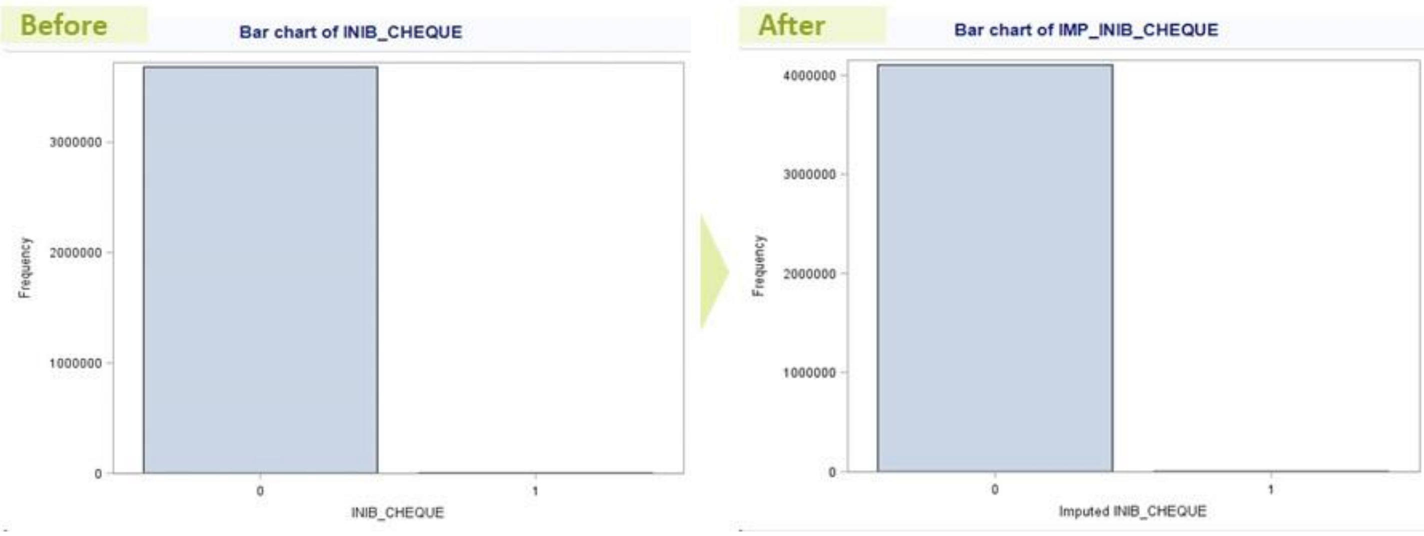

Figure 75 – Comparison of the check inhibition before and after impute and outlier smoothing

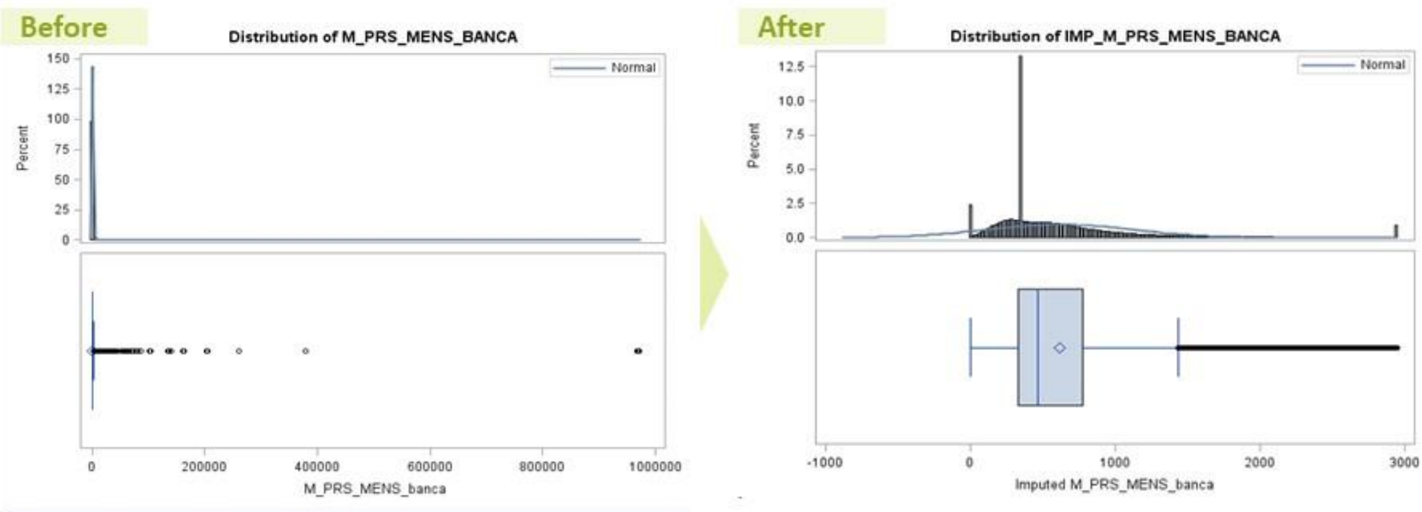

Figure 76 – Comparison of the monthly instalments in the financial system before and after impute and outlier smoothing

Appendix 5. GENERALIZATION – PURPOSE OF LOAN

Loan purpose – original	Loan purpose – aggregation
AQUISICAO DE HABITACAO PERMANENTE - NOVA	Acquisition permanent home
AQUISICAO DE HABITACAO PERMANENTE - USADA	Acquisition permanent home
AQUISICAO DE HABITACAO SECUNDARIA - NOVA	Acquisition secondary home
AQUISICAO DE HABITACAO SECUNDARIA - USADA	Acquisition secondary home
AQUISICAO DE HABITACAO RENDIMENTO - NOVA	Acquisition secondary home
AQUISICAO DE HABITACAO RENDIMENTO - USADA	Acquisition property home
AQUISICAO DE IMOVEL PARA RENDIMENTO	Acquisition property home
AQUISICAO DE IMOVEL P- SERVICO	Acquisition property home
AQUISICAO OUTRAS FINALIDADES	Acquisition other home
AQUISICAO TERRENOS CONSTRUCAO	Acquisition land / construction
OBRAS NA HABITACAO PERMANENTE	Works
OBRAS EM IMOVEL P/ RENDIMENTO	Works
OBRAS DE CONSERVACAO ORDINARIA NA HABITACAO PERMANENTE	Works
OBRAS DE CONSERVACAO ORDINARIA NA HABITACAO SECUNDARIA	Works
OBRAS DE CONSERVACAO ORDINARIA NA HABITACAO RENDIMENTO	Works
OBRAS DE CONSERVACAO EXTRAORDINARIA NA HABITACAO PERMANENTE	Works
OBRAS DE CONSERVACAO EXTRAORDINARIA NA HABITACAO SECUNDARIA	Works
OBRAS DE CONSERVACAO EXTRAORDINARIA NA HABITACAO RENDIMENTO	Works
OBRAS DE BENEFICIACAO NA HABITACAO PERMANENTE	Works
OBRAS DE BENEFICIACAO NA HABITACAO SECUNDARIA	Works
OBRAS DE BENEFICIACAO NA HABITACAO RENDIMENTO	Works
OBRAS EM IMOVEL P- RENDIMENTO	Works
OBRAS EM IMOVEL P- SERVICO PRO	Works
OBRAS OUTRAS FINALIDADES	Works
OBRAS POR INQUILINOS	Works
CONSTRUCAO DE HABITACAO PERMANENTE	Acquisition land / construction
CONSTRUCAO DE HABITACAO SECUNDARIA	Acquisition land / construction

Loan purpose – original	Loan purpose – aggregation
CONSTRUCAO DE HABITACAO RENDIMENTO	Acquisition land / construction
CONSTRUCAO DE IMOVEL P- RENDIM	Acquisition land / construction
CONSTRUCAO DE IMOVEL P- SERVIC	Acquisition land / construction
CONSTRUCAO OUTRAS FINALIDADES	Acquisition land / construction
INSTALACAO DE CASAS PRE-FABRICADAS - HABITACAO PERMANENTE	Installation of prefabricated homes
INSTALACAO DE CASAS PRE-FABRICADAS - HABITACAO SECUNDARIA	Installation of prefabricated homes
INSTALACAO DE CASAS PRE-FABRICADAS - HABITACAO RENDIMENTO	Installation of prefabricated homes
INVESTIMENTO NAO ESPECIFICADO EM IMOBILIARIO	Investments in real estate
INVESTIMENTO EM IMOVEIS PARA H	Investments in real estate
INVESTIMENTO EM IMOVEIS PARA SERVICO	Investments in real estate
INVEST IMOV ARREND NAO HABIT	Investments in real estate
CREDIOBRAS HABITACAO PERMAN	Works
CREDIOBRAS HABITACAO SECUND	Works
CREDIOBRAS - HABITACAO RENDIME	Works
OUTROS BENS DE CONSUMO	Works
AQUISICAO DE GARAGEM	Acquisition garage / others
AQUISICAO PRODUTO NAO BANCARIO	Acquisition garage / others
COMERCIO E SERVICOS	Acquisition garage / others
CONSTRUCAO	Works
PARTICULARES	Acquisition garage / others
REESTRUTURACAO DE CREDITO	Credit restructuring
OUTRAS APLICACOES FINANCEIRAS	Investments in real estate
ELECTRODOMESTICOS / MOBILIARIO / DECORACAO	Acquisition of goods
AQUISICAO DE TERRENO	Acquisition land / construction
OBRAS DE REABILITACAO URBANA	Works
AQUISICAO FRACCAO USADA NAO HIPOTECADA	Acquisition garage / others
AQUISICAO MORADIA NOVA FIN.OUT.I.CRED	Acquisition secondary home
AQUISICAO IMOVEIS OUTROS	Acquisition garage / others

Loan purpose – original	Loan purpose – aggregation
OBRAS CONSERVACAO ORDINARIA DA FRACCAO	Works
HABITACAO PROPRIA PERMANENTE	Acquisition permanent home
AQUISICAO HABITACAO PROPRIA NAO PERMANENTE	Acquisition secondary home
BENEFICIACAO HABITACAO PROPRIA NAO PERMANENTE	Acquisition secondary home
CONSTRUCAO - OUTRAS	Works
COMPLEMENTO AQUISICAO HABITACAO PROPRIA PERMANENTE	Acquisition of goods
COMPLEMENTO OBRAS HPP	Works
COMPLEMENTO CONSTRUCAO HPP	Acquisition land / construction
CREDITO HABITACAO PARA APOIO A DESEMPREGADOS - D.L.103/09	Acquisition permanent home
REESTRUTURACAO DE CREDITOS NO GRUPO	Credit restructuring
CREDITO HABITACAO - CH IMOVEIS ENTIDADE PUBLICA - HPP	Acquisition property home
CREDITO HABITACAO - CH IMOVEIS ENTIDADE PUBLICA - HSEC	Acquisition property home
AQUISICAO IMOVEIS PARA FINS TURISTICOS	Acquisition property home
AQUISICAO IMOVEL PARA VENDA	Acquisition property home

Table 24 – Original categories and mapping to aggregated categories in loan purpose

Appendix 6. GENERALIZATION – MARITAL STATUS

Marital Status – original	Marital Status – aggregation
DESCONHECIDO	Unknown
DESCONHECIDO	Unknown
SOLTEIRO	Single
CASADO EM REGIME DE COMUNHAO GERAL DE BENS	Married/De facto Union
CASADO EM REGIME DE SEPARACAO DE BENS	Married/De facto Union
CASADO EM REGIME DE COMUNHAO DE ADQUIRIDOS	Married/De facto Union
CASADO-REGIME DOTAL	Married/De facto Union
UNIDO DE FACTO	Married/De facto Union
SEPARADO JUDICIALMENTE DE PESSOAS E BENS	Separated / Divorced
DIVORCIADO	Separated / Divorced
VIUVO	Widower
CASADO	Married/De facto Union
SEPARADO JUDICIALMENTE DE BENS	Separated / Divorced
DESCONHECIDO	Unknown

Table 25 - Original categories and mapping to aggregated categories in marital status

Appendix 7. GENERALIZATION – PROFESSION

Profession – original	Profession – aggregation
FUNCIONARIO PUBLICO	Administrative staff
OFIC.OUT.PROF.DAS FORCAS SVC SEGUR.C/FUNC.CMD DIR.OU CHEFIA	Personal, safety and security services workers and vendors
OUTROS ESPECIALISTAS EM ENGENHARIA (EXCEPTO ELECTROTECNOLOGI	Specialists in intellectual and scientific activities
PROFISSIONAL PARAMEDICO	Specialists in intellectual and scientific activities
TECNICO DOS SERVICOS DE SAUDE COMUNITARIA	Intermediate level technicians and professions
ASSISTENTE DE MEDICOS	Intermediate level technicians and professions
AGENTES DE CREDITO E EMPRESTIMOS	Specialists in intellectual and scientific activities
AGENTE DE SERVICOS DE LICENCIAMENTO	Specialists in intellectual and scientific activities
ESCRIVAO E SIMILARES	Administrative staff
PESSOAL DE COMPANHIA E AJUDANTES DE QUARTO	Unskilled workers
CONDUTOR DE MOTOCICLOS	Plant and machine operators and assembly workers
CONDUTOR DE VEICULOS ACCIONADOS A MAO OU AO PE	Plant and machine operators and assembly workers
TRABALHADOR DA RECOLHA DE RESIDUOS	Unskilled workers
OFICIAL DE MARINHA	Armed Forces Professions
OFICIAL DE ADMINISTRACAO NAVAL	Armed Forces Professions
OFICIAL ENGENHEIRO NAVAL	Armed Forces Professions
OFICIAL DE INFANTARIA	Armed Forces Professions
OFICIAL DE ARTILHARIA	Armed Forces Professions
OFICIAL DE CAVALARIA	Armed Forces Professions
OFICIAL DE ENGENHARIA MILITAR	Armed Forces Professions
OFICIAL DE MATERIAL MILITAR (EXERCITO)	Armed Forces Professions
OFICIAL DE ADMINISTRACAO MILITAR (EXERCITO)	Armed Forces Professions
OUTROS OFICIAIS DO EXERCITO	Armed Forces Professions
OFICIAL PILOTO AVIADOR	Armed Forces Professions
OFICIAL DA AREA DE OPERACOES AEREAS	Armed Forces Professions
OFICIAL DA FORCA AEREA DA AREA DE MANUTENCAO DE SISTEMAS DE	Armed Forces Professions
OUTROS OFICIAIS DA FORCA AEREA	Armed Forces Professions

Profession – original	Profession – aggregation
SARG.COMUNICACOES (MARINHA)	Armed Forces Professions
SARG.OPERACOES (MARINHA)	Armed Forces Professions
SARG.MANOBRA SVC (MARINHA)	Armed Forces Professions
SARG.TEC.ARMAMENTO (MARINHA)	Armed Forces Professions
OUT.SARG.MARINHA EQUIP/DOS	Armed Forces Professions
SARGENTO DE INFANTARIA	Armed Forces Professions
SARGENTO DE ARTILHARIA	Armed Forces Professions
SARG.MAT.MILITAR (EXERCITO)	Armed Forces Professions
OUTROS SARGENTOS DO EXERCITO	Armed Forces Professions
SARGENTO DA AREA DE OPERACOES AEREAS	Armed Forces Professions
SARG.FORCA AEREA AREA MANUT.DE SIST.DE ARMAS	Armed Forces Professions
SARGENTO DE POLICIA AEREA	Armed Forces Professions
OUT.SARG.FORCA AEREA	Armed Forces Professions
PRACA COMUNICACOES (MARINHA)	Armed Forces Professions
PRACA FUZILEIRO	Armed Forces Professions
PRACA DE OPERACOES (MARINHA)	Armed Forces Professions
OUT.PRACAS MARINHA EQUIP/DOS	Armed Forces Professions
PRACA DE INFANTARIA	Armed Forces Professions
PRACA TRANSMISSOES (EXERCITO)	Armed Forces Professions
PRACA MAT.MILITAR (EXERCITO)	Armed Forces Professions
OUTRAS PRACAS DO EXERCITO	Armed Forces Professions
PRACA AREA OPERACOES AEREAS	Armed Forces Professions
OUTRAS PRACAS DA FORCA AEREA	Armed Forces Professions
DIR.PROD.NA AGRIC.	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.DAS IND.CONSTRUCAO ENG.CIVIL	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERENTE DO COMERCIO A RETALHO	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERENTE DO COMERCIO POR GROSSO	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERENTE HOTEIS SIMILARES	Representatives of the legislative power and executive bodies, directors and executive managers

Profession – original	Profession – aggregation
DIR.GERENTE RESTAURACAO (RESTAURANTES SIMILARES)	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.DOS SVC DAS TEC.INF.COMUNICACAO (TIC)	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DE TRANSPORTES	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.ARMAZENAGEM DIST.RELACIONADOS	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.SUC.DE BANCOS SVC FINANCEIROS SEGUROS	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DOS SERVICOS DE EDUCACAO	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DOS SERVICOS DE SAUDE	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.DOS SVC CUIDADOS A PESSOAS IDOSAS	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.DOS SVC APOIO SOCIAL	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.BIBLIOTECAS ARQ.MUSEUS GALERIAS ARTE MONUMENTOS NACIONAI	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.OUT.SVC ESPECIALIZADOS N.E.	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERENTE DOS CENTROS DESPORTIVOS RECREATIVOS CULTURAIS	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.ESTRATEGIA PLANEAMENTO	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR FINANCEIRO	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DE RECURSOS HUMANOS	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DE MARKETING	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DE PUBLICIDADE	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DE RELACOES PUBLICAS	Representatives of the legislative power and executive bodies, directors and executive managers
DIRECTOR DE COMPRAS	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.INVEST.DESENVOLVIMENTO	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.DAS IND.TRANSFORMADORAS	Representatives of the legislative power and executive bodies, directors and executive managers
FLORICULTOR	Farmers and skilled workers in agriculture, fishing and forestry
AGRIC.TRAB.QUALIF.CULT.AGRICOLAS MISTAS	Farmers and skilled workers in agriculture, fishing and forestry
PROD.TRAB.QUALIF.NA PROD.BOVINOS	Farmers and skilled workers in agriculture, fishing and forestry
AGRIC.TRAB.QUALIF.AGRIC.PROD.ANIMAL COMBINADAS ORIENTADOS P/	Farmers and skilled workers in agriculture, fishing and forestry
DIR.DAS IND.EXTRACTIVAS	Representatives of the legislative power and executive bodies, directors and executive managers
GEOFISICO	Specialists in intellectual and scientific activities

Profession – original	Profession – aggregation
ESTATICISTA E DEMOGRAFO	Specialists in intellectual and scientific activities
PROGRAMADOR DE SOFTWARE	Specialists in intellectual and scientific activities
PROGRAMADOR WEB E DE MULTIMEDIA	Specialists in intellectual and scientific activities
ADM.ESPEC.CONCEPCAO BASE DADOS	Specialists in intellectual and scientific activities
ADMINISTRADOR DE SISTEMAS	Specialists in intellectual and scientific activities
OUT.ANALISTAS PROGRAMADORES SOFTWARE APLICACOES	Specialists in intellectual and scientific activities
ESPEC.REDES INFORMATICAS	Specialists in intellectual and scientific activities
ARQUITECTO PAISAGISTA	Specialists in intellectual and scientific activities
ENGENHEIRO DE CONSTRUCAO DE EDIFICIOS	Specialists in intellectual and scientific activities
ENGENHEIRO ELECTRONICO	Specialists in intellectual and scientific activities
ENGENHEIRO DE TELECOMUNICACOES	Specialists in intellectual and scientific activities
ENGENHEIRO INDUSTRIAL PROD.	Specialists in intellectual and scientific activities
ENGENHEIRO DO AMBIENTE	Specialists in intellectual and scientific activities
OUT.ENG.RELACIONADOS C/MINAS METALURGIA	Specialists in intellectual and scientific activities
ESPEC.PROTECCAO DO AMBIENTE	Specialists in intellectual and scientific activities
CONSULTOR ACTIVIDADES DAS PESCAS	Specialists in intellectual and scientific activities
MEDICO DE ESPECIALIDADES MEDICAS	Specialists in intellectual and scientific activities
MEDICO DE ESPECIALIDADES CIRURGICAS	Specialists in intellectual and scientific activities
MEDICO ESTOMATOLOGISTA	Specialists in intellectual and scientific activities
MEDICO MEDICINA GERAL FAMILIAR	Specialists in intellectual and scientific activities
ENFERMEIRO ESPEC.EM ENFERMAGEM MEDICO-CIRURGICA	Specialists in intellectual and scientific activities
ENFERMEIRO ESPEC.EM REABILITACAO	Specialists in intellectual and scientific activities
ENFERMEIRO ESPEC.EM ENFERMAGEM COMUNITARIA	Specialists in intellectual and scientific activities
ENFERMEIRO ESPEC.EM SAUDE MATERNA OBSTETRICA	Specialists in intellectual and scientific activities
ENFERMEIRO ESPEC.EM SAUDE INFANTIL PEDIATRICA	Specialists in intellectual and scientific activities
ENFERMEIRO ESPEC.EM SAUDE MENTAL PSIQUIATRICA	Specialists in intellectual and scientific activities
AUXILIAR DE ENFERMAGEM	Intermediate level technicians and professions
PROFESSOR DOS ENSINOS, TECNOLOGICO, ARTISTICO E PROFISSIONAL	Specialists in intellectual and scientific activities

Profession – original	Profession – aggregation
OUTROS PROFESSORES DE LINGUAS	Specialists in intellectual and scientific activities
OUTROS PROFESSORES DE ARTE	Specialists in intellectual and scientific activities
ESPEC.EM FORMACAO DESENVOLVIMENTO RECURSOS HUMANOS	Specialists in intellectual and scientific activities
ESPEC.EM HIGIENE SAUDE AMBIENTAL LABORAL	Specialists in intellectual and scientific activities
CONSULTOR FINANCEIRO INVESTIMENTOS	Specialists in intellectual and scientific activities
ANALISTA FINANCEIRO	Specialists in intellectual and scientific activities
ESPEC.EM POLITICAS ADM.	Specialists in intellectual and scientific activities
OUT.ESPEC.EM ASSUNTOS JURIDICOS N.E.	Specialists in intellectual and scientific activities
BIBLIOTECARIOS OUT.ESPEC.INF.RELACIONADOS	Specialists in intellectual and scientific activities
ESPEC.EM CIENCIAS POLITICAS	Specialists in intellectual and scientific activities
INTERPRETE E OUTROS LINGUISTAS	Specialists in intellectual and scientific activities
PINTOR DE ARTE	Specialists in intellectual and scientific activities
DESIGNER PRODUTO INDUSTRIAL OU EQUIPAMENTO	Specialists in intellectual and scientific activities
DESIGNER INTERIORES ESPACOS OU AMBIENTES	Specialists in intellectual and scientific activities
BAILARINO	Specialists in intellectual and scientific activities
ACTOR	Specialists in intellectual and scientific activities
REALIZADOR DE CINEMA E TEATRO	Specialists in intellectual and scientific activities
DIR.FOTOGRAFIA SOM MONTADOR RELACIONADOS	Specialists in intellectual and scientific activities
TECNICO DAS CIENCIAS FISICAS	Intermediate level technicians and professions
TECNICO DE TELECOMUNICACOES	Intermediate level technicians and professions
TEC.CONTROLO INSTALACOES INDUSTRIA QUIMICA	Intermediate level technicians and professions
TECNICO DE QUIMICA INDUSTRIAL	Intermediate level technicians and professions
TEC.EM REDES SIST.DE COMPUTADORES	Intermediate level technicians and professions
TECNICO DA WEB	Intermediate level technicians and professions
TEC.GRAVACAO AUDIOVISUAL	Intermediate level technicians and professions
TECNICO DE EMISSOES DE RADIO	Intermediate level technicians and professions
TEC.EMISSOES TELEVISAO	Intermediate level technicians and professions
TECNICO DE CARDIOPNEUMOGRAFIA	Intermediate level technicians and professions

Profession – original	Profession – aggregation
TECNICO DE MEDICINA NUCLEAR	Intermediate level technicians and professions
OUT.TEC.EQUIPAMENTO DIAGNOSTICO TERAPEUTICO	Intermediate level technicians and professions
TECNICO DE RADIOLOGIA	Intermediate level technicians and professions
TECNICO DE RADIOTERAPIA	Intermediate level technicians and professions
OFICIAL MAQUINISTA DE NAVIOS	Intermediate level technicians and professions
CALIBRADOR VERIFICADOR PROD.(EXCEPTO ALIMENTOS BEBIDAS)	Skilled workers in industry, construction and crafts
OUT.ENC.INDUSTRIA TRANSFORMADORA	Intermediate level technicians and professions
TEC.INSPECCAO VEIC.	Intermediate level technicians and professions
TEC.METALURGIA BASE INDUSTRIA EXTRACTIVA	Intermediate level technicians and professions
TECNICO DE ANALISES CLINICAS	Intermediate level technicians and professions
TEC.ANATOMIA PATOLOGICA CITOLOGICA TANATOLOGICA	Intermediate level technicians and professions
TEC.DAS CIENCIAS VIDA (EXCEPTO CIENCIAS MEDICAS)	Intermediate level technicians and professions
TECNICO AGRICOLA	Intermediate level technicians and professions
TECNICO DA PRODUCAO ANIMAL	Intermediate level technicians and professions
TEC.FLORESTAL (INCLUI CINEGETICO)	Intermediate level technicians and professions
TECNICO DE OPTICA OCULAR	Intermediate level technicians and professions
TEC.ASSISTENTE FISIOTERAPIA SIMILARES	Intermediate level technicians and professions
OUT.PROF.NIVEL INTERMEDIO SAUDE N.E.	Intermediate level technicians and professions
TERAPEUTA OCUPACIONAL	Specialists in intellectual and scientific activities
TERAPEUTA DA FALA	Specialists in intellectual and scientific activities
AUDIOLOGISTA	Specialists in intellectual and scientific activities
OUT.PROF.SAUDE DIVERSOS N.E.	Intermediate level technicians and professions
ACUPUNCTOR	Specialists in intellectual and scientific activities
PROF.NIVEL INTERMEDIO MEDICINA TRADICIONAL COMPLEMENTAR	Intermediate level technicians and professions
INSTRUTORES MONITORES ACTIVIDADE FISICA RECREACAO	Intermediate level technicians and professions
AGENTE IMOBILIARIO GESTOR PROPRIEDADES	Intermediate level technicians and professions
ORGANIZADOR CONFERENCIAS EVENTOS	Intermediate level technicians and professions
OUT.ESPEC.EM VND MAT.TEC.MEDICO (EXCEPTO TIC)	Intermediate level technicians and professions

Profession – original	Profession – aggregation
TECNICO DE COMPRAS	Intermediate level technicians and professions
CORRETOR COMERCIAL	Intermediate level technicians and professions
CHEFE DE ESCRITORIO	Intermediate level technicians and professions
SECRETARIO ADMINISTRATIVO EXEC.	Intermediate level technicians and professions
SECRETARIO DA AREA JURIDICA	Intermediate level technicians and professions
SECRETAIRE MEDICAL	Intermediate level technicians and professions
OUT.TEC.DAS CIENCIAS FISICAS ENG.N.E.	Intermediate level technicians and professions
TECNICOS DE GALERIAS, BIBLIOTECAS, ARQUIVOS E MUSEUS	Intermediate level technicians and professions
LOCUTOR APRESENTADOR RADIO TELEVISAO OUT.MEIOS COMUNICACAO	Specialists in intellectual and scientific activities
JOGADOR PROFISSIONAL FUTEBOL	Intermediate level technicians and professions
CICLISTA PROFISSIONAL	Intermediate level technicians and professions
TREINADOR DE DESPORTOS	Intermediate level technicians and professions
ARBITRO (JUIZ) DE DESPORTOS	Intermediate level technicians and professions
INSTRUTOR DE DESPORTOS	Intermediate level technicians and professions
OPER.CONTABILIDADE ESCRITURACAO COMERCIAL	Administrative staff
OPER.DADOS PROCESSAMENTO PAGAMENTOS	Administrative staff
EMPREGADO SVC APOIO A PROD.	Administrative staff
CONTROLADOR TRANSPORTES TERRESTRES MERCADORIAS	Administrative staff
EMPREGADO CONTROLO DOS SVC TRANSPORTES AEREOS MARITIMOS	Administrative staff
EMPREGADO DE BIBLIOTECA	Administrative staff
TEC.REGISTOS MEDICOS INF.SOBRE SAUDE	Intermediate level technicians and professions
EMPREGADO SVC PESSOAL	Administrative staff
OUTRO PESSOAL APOIO TIPO ADMINISTRATIVO N.E.	Administrative staff
FISCAL ENC.PORTAGEM	Intermediate level technicians and professions
LEITOR DE CONTADORES	Unskilled workers
RECEPCIONISTA EXCEPTO HOTEL	Administrative staff
RECEPCIONISTA DE HOTEL	Administrative staff
OUTRO PESSOAL RECEPCAO INF.A CLIENTES	Administrative staff

Profession – original	Profession – aggregation
EMPREGADO DOS CENTROS DE CHAMADAS	Administrative staff
CHEFE DE COZINHA	Intermediate level technicians and professions
ENC.LIMPEZA TRAB.DOMESTICOS EM ESCRITORIOS HOTEIS OUT.ESTABE	Unskilled workers
GOVERNANTE DOMESTICO	Personal, safety and security services workers and vendors
AJUDANTE DE COZINHA	Unskilled workers
PREP/DOR REFEICOES RAPIDAS	Unskilled workers
ASSISTENTE VND ALIMENTOS AO BALCAO	Personal, safety and security services workers and vendors
AUXILIAR DE PROFESSOR	Personal, safety and security services workers and vendors
AUXILIAR DE SAUDE	Personal, safety and security services workers and vendors
AJUDANTE FAMILIAR	Personal, safety and security services workers and vendors
PRESTADOR CUIDADOS A ANIMAIS	Personal, safety and security services workers and vendors
MASSAGISTA DE ESTETICA	Personal, safety and security services workers and vendors
TEC.NIVEL INTERMEDIO APOIO SOCIAL	Intermediate level technicians and professions
OUT.TRAB.DOS SVC PESSOAIS N.E.	Intermediate level technicians and professions
ADIVINHADOR E SIMILARES	Personal, safety and security services workers and vendors
AGENTE POLICIA SEGUR.PUB.	Personal, safety and security services workers and vendors
AGENTE DE POLICIA MARITIMA	Personal, safety and security services workers and vendors
AGENTE DE POLICIA MUNICIPAL	Personal, safety and security services workers and vendors
SARG.GUARDA NACIONAL REPUB.NA	Personal, safety and security services workers and vendors
GUARDAS GUARDA NACIONAL REPUB.NA	Personal, safety and security services workers and vendors
GUARDA DOS SERVICOS PRISIONAIS	Personal, safety and security services workers and vendors
REPOSITOR PROD.EM PRATELEIRAS	Unskilled workers
DEMONSTRADOR	Personal, safety and security services workers and vendors
ENC.LOJA (ESTABELECIMENTO)	Personal, safety and security services workers and vendors
VENDEDOR EM LOJA (ESTABELECIMENTO)	Personal, safety and security services workers and vendors
ASSISTENTE ESTACAO SVC AO CONDUTOR	Personal, safety and security services workers and vendors
TRAB.NAO QUALIF.FLORICULTURA HORTICULTURA	Unskilled workers

Profession – original	Profession – aggregation
OUT.PROD.ES TRAB.QUALIF.S CRIACAO ANIMAL	Farmers and skilled workers in agriculture, fishing and forestry
MOTOSSERRISTA	Farmers and skilled workers in agriculture, fishing and forestry
SAPADOR FLORESTAL	Farmers and skilled workers in agriculture, fishing and forestry
MESTRE CONTRAMESTRE ARRAIS PESCA MARITIMA COSTEIRA	Farmers and skilled workers in agriculture, fishing and forestry
OUT.TRAB.QUALIF.S PESCA MARITIMA COSTEIRA	Farmers and skilled workers in agriculture, fishing and forestry
PESCADOR DE AGUAS INTERIORES	Farmers and skilled workers in agriculture, fishing and forestry
MERGULHADOR	Skilled workers in industry, construction and crafts
MESTRE CONTRAMESTRE ARRAIS PESCA MARITIMA DO LARGO	Farmers and skilled workers in agriculture, fishing and forestry
OUT.TRAB.QUALIF.S PESCA MARITIMA DO LARGO	Farmers and skilled workers in agriculture, fishing and forestry
PESCADOR E MARINHEIRO PESCADOR, DE PESCA MARITIMA DO LARGO	Farmers and skilled workers in agriculture, fishing and forestry
AGRICULTOR DE SUBSISTENCIA	Farmers and skilled workers in agriculture, fishing and forestry
CRIADOR ANIMAIS SUBSISTENCIA	Farmers and skilled workers in agriculture, fishing and forestry
AGRIC.CRIADOR ANIMAIS PROD.COMBINADA SUBSISTENCIA	Farmers and skilled workers in agriculture, fishing and forestry
MINEIRO	Plant and machine operators and assembly workers
ENC.INDUSTRIA EXTRACTIVA	Intermediate level technicians and professions
OUT.TRAB.QUALIF.S PEDRA SIMILARES	Skilled workers in industry, construction and crafts
OUT.OPER.ES INSTALACOES FIXAS MAQ.DIVERSAS N.E	Plant and machine operators and assembly workers
ASSENTADOR DE REFRACTARIOS	Skilled workers in industry, construction and crafts
ARMADOR DE FERRO	Skilled workers in industry, construction and crafts
OUT.TRAB.QUALIF.S EM BETAO ARMADO SIMILARES	Skilled workers in industry, construction and crafts
MONTADOR ALVENARIAS PRE-ESFORCADOS	Skilled workers in industry, construction and crafts
ENCARREGADO DA CONSTRUCAO	Intermediate level technicians and professions
CARPINTEIRO LIMPOS TOSCO	Skilled workers in industry, construction and crafts
CARPINTEIRO NAVAL	Skilled workers in industry, construction and crafts
CONSTRUTOR CASAS RUDIMENTARES	Skilled workers in industry, construction and crafts
COLOCADOR TELHADOS COBERTURAS	Skilled workers in industry, construction and crafts
LADRILHADOR	Skilled workers in industry, construction and crafts
ASSENTADOR TACOS AFAGADOR MADEIRA	Skilled workers in industry, construction and crafts

Profession – original	Profession – aggregation
MONTADOR DE TUBAGENS	Skilled workers in industry, construction and crafts
ELECTRICISTA CONSTRUCOES SIMILARES	Skilled workers in industry, construction and crafts
COLOCADOR PAPEL PAREDE PINTOR DECORADOR SIMILARES	Skilled workers in industry, construction and crafts
PINTOR A PISTOLA SUPERFICIES	Skilled workers in industry, construction and crafts
ENC.DAS IND.METALURGICAS BASE FAB.PROD.METALICOS	Intermediate level technicians and professions
SERRALHEIRO MOLDES CUNHOS CORTANTES SIMILARES	Skilled workers in industry, construction and crafts
REG.OPER.MAQ.-FERRA/AS CMD NUMERICO COMPUTORIZADO P/ TRAB.ME	Skilled workers in industry, construction and crafts
RECTIFICADOR RODAS POLIDOR AFIADOR METAIS	Skilled workers in industry, construction and crafts
REP/DOR BICICLETAS SIMILARES	Skilled workers in industry, construction and crafts
OPER.MAQ.P/ CORTE SOLDADURA ISOLAMENTO FAB.ENROLAMENTO CABLA	Plant and machine operators and assembly workers
OUT.OURIVES TRAB.DIAMANTES INDUSTRIAIS	Skilled workers in industry, construction and crafts
OLEIRO	Skilled workers in industry, construction and crafts
MODELADOR FORMISTA CERAMICA	Skilled workers in industry, construction and crafts
OPER.INSTALACOES P/ O FAB.PROD.CERAMICOS	Skilled workers in industry, construction and crafts
ENC.DAS IND.TRANSF.MINERAIS NAO METALICOS	Intermediate level technicians and professions
CORTADOR DE VIDRO	Skilled workers in industry, construction and crafts
TRABALHADOR DE VIDRO DE OPTICA	Skilled workers in industry, construction and crafts
TRAB.OUT.OFICIOS DIVERSOS N.E.	Skilled workers in industry, construction and crafts
OPER.INSTALACOES P/ O FAB.VIDRO	Skilled workers in industry, construction and crafts
ARTESAO DE ARTIGOS EM MADEIRA	Skilled workers in industry, construction and crafts
ARTESAO RENDAS BORDADOS TAPECARIAS MANUAIS	Skilled workers in industry, construction and crafts
OUT.TRAB.MANUAIS ARTIGOS TEXTEIS COURO MAT.SIMILARES	Skilled workers in industry, construction and crafts
ENC.DAS IND.PASTA PAPEL IMPRESSAO SIMILARES	Intermediate level technicians and professions
OUT.SUPERVISORES PESSOAL ADMINISTRATIVO	Intermediate level technicians and professions
ENCADERNADOR	Skilled workers in industry, construction and crafts
OUT.TRAB.RELACIONADOS C/O ACABAMENTO IMPRESSAO	Skilled workers in industry, construction and crafts
MATADOR DE ANIMAIS	Skilled workers in industry, construction and crafts
CORTADOR DE CARNE	Skilled workers in industry, construction and crafts

Profession – original	Profession – aggregation
PREP/DOR CONSERVADOR PEIXE	Skilled workers in industry, construction and crafts
ENC.DAS IND.ALIM.DAS BEBIDAS	Intermediate level technicians and professions
OPER.MAQ.PROD.PADARIA PASTELARIA CONFEITARIA MASSAS ALIMENTI	Plant and machine operators and assembly workers
PASTELEIRO	Skilled workers in industry, construction and crafts
CONSERVEIRO FRUTAS LEGUMES SIMILARES	Skilled workers in industry, construction and crafts
PROVADORES CLASSIFICADORES ALIMENTOS BEBIDAS	Skilled workers in industry, construction and crafts
TRAB.DO TRATA/0 MADEIRA	Skilled workers in industry, construction and crafts
TRAB.DO TRATA/0 CORTICA	Skilled workers in industry, construction and crafts
ENC.DAS IND.MADEIRA CORTICA	Intermediate level technicians and professions
ENC.DAS IND.TEXTEIS DO VESTUARIO CALCADO CURTUMES	Intermediate level technicians and professions
TRAB.COSTURA SIMILARES	Skilled workers in industry, construction and crafts
OUT.TRAB.SIMILARES A ESTOFADOR	Skilled workers in industry, construction and crafts
CURTIDOR DE PELES	Skilled workers in industry, construction and crafts
OPER.INSTALACOES PROCESSAMENTO ROCHAS	Plant and machine operators and assembly workers
OPER.INSTALACOES PROCESSAMENTO MINERIOS	Plant and machine operators and assembly workers
PERFURADOR POCOS SONDADOR SIMILARES	Plant and machine operators and assembly workers
TEC.CONTROLO INSTALACOES PROD.METAIS	Intermediate level technicians and professions
OPER.INSTALACOES FORNOS PRIMEIRA TRANSF.METAIS	Plant and machine operators and assembly workers
OPER.INSTALACOES FORNOS SEGUNDA FUSAO VAZADORES LAMINADORES	Plant and machine operators and assembly workers
OPER.INSTALACOES TRATA/0 TERMICO METAIS	Plant and machine operators and assembly workers
OPER.INSTALACOES MAQ.P/ TRATA/0 TERMICO PROD.QUIMICOS	Plant and machine operators and assembly workers
OPER.INSTALACOES MAQ.P/ FILTRAGEM SEP/CAO QUIMICA	Plant and machine operators and assembly workers
OPER.INSTALACOES MAQ.P/ REACCAO VERIFICACAO PROD.QUIMICOS	Plant and machine operators and assembly workers
TEC.OPERACAO INSTALACOES REF.PET.GAS NATURAL	Intermediate level technicians and professions
OPER.INSTALACOES MAQ.P/ PET.GAS	Plant and machine operators and assembly workers
OPER.MAQ.A VAPOR CALDEIRAS	Plant and machine operators and assembly workers
TEC.OPERACAO INSTALACOES TRATA/0 AGUA	Intermediate level technicians and professions
OPER.MAQ.P/ TRAB.O CIMENTO	Plant and machine operators and assembly workers

Profession – original	Profession – aggregation
OPER.MAQ.P/ TRAB.OUT.MINERAIS	Plant and machine operators and assembly workers
OPER.MAQ.P/ TRAB.A PEDRA	Plant and machine operators and assembly workers
OPER.MAQ.P/ FAB.MOLAS P/ ESTOFOS COLCHOES VEIC.AUTO.OU OUT.F	Plant and machine operators and assembly workers
OPER.MAQ.P/ FAB.PROD.ARAME	Plant and machine operators and assembly workers
OPER.INSTALACOES MAQ.P/ OUT.TRATA/0S QUIMICOS	Plant and machine operators and assembly workers
OPER.MAQ.REVESTIMENTO METALIZACAO ACABAMENTO METAIS	Plant and machine operators and assembly workers
OPER.MAQ.EQUIP.P/ TRAB.MADEIRA	Plant and machine operators and assembly workers
OPER.MAQ.EQUIP.P/ TRAB.CORTICA	Plant and machine operators and assembly workers
OPER.MAQ.P/ O FAB.PROD.PAPEL	Plant and machine operators and assembly workers
OPER.MAQ.TECER TRICOTAR	Plant and machine operators and assembly workers
OPER.MAQ.BRANQUEAR TINGIR LIMPAR TECIDOS OUT.TEXTEIS	Plant and machine operators and assembly workers
OPER.MAQ.LAVANDARIA	Plant and machine operators and assembly workers
OPER.MAQ.P/ PREP/R PELES C/PELO COURO	Plant and machine operators and assembly workers
OUT.OPER.ES MAQ.P/ O FAB.PROD.TEXTEIS PELE C/PELO COURO	Plant and machine operators and assembly workers
OPER.MAQ.FAB.PROD.LACTEOS	Plant and machine operators and assembly workers
OPER.MAQ.TRATA/0 FRUTOS LEGUMES FAB.AZEITE OLEOS ALIM.MARGAR	Plant and machine operators and assembly workers
MONTADOR MAQUINARIA MECANICA	Plant and machine operators and assembly workers
MONTADOR EQUIP.ELECTRICOS ELECTRONICOS	Plant and machine operators and assembly workers
OUT.TRAB.MONTAGEM	Plant and machine operators and assembly workers
OPER.MAQ.EMBALAR ENCHER ROTULAR	Plant and machine operators and assembly workers
GUARDA-FREIOS AGULHEIRO AGENTE MANOBRAS CAMINHOS-DE-FERRO	Plant and machine operators and assembly workers
MOTORISTA DE TAXIS	Plant and machine operators and assembly workers
MOTORISTA DE AUTOCARROS	Plant and machine operators and assembly workers
GUARDA-FREIO DE ELECTRICO	Plant and machine operators and assembly workers
MOTORISTA VEIC.PESADOS MERCADORIAS	Plant and machine operators and assembly workers
OPER.MAQ.AGRICOLAS FLORESTAIS MOVEIS	Plant and machine operators and assembly workers
OPER.MAQ.ESCAVACAO TERRAPLENAGEM SIMILARES	Plant and machine operators and assembly workers
OPER.GRUAS GUINDASTES SIMILARES	Plant and machine operators and assembly workers

Profession – original	Profession – aggregation
OPERADOR DE EMPILHADORES	Plant and machine operators and assembly workers
VENDEDOR AMBULANTE (EXCEPTO ALIMENTOS)	Unskilled workers
VENDEDOR CENTROS CONTACTO	Unskilled workers
VENDEDOR AO DOMICILIO	Unskilled workers
PORTEIRO DE EDIFICIOS	Personal, safety and security services workers and vendors
LAVADOR DE VEICULOS	Unskilled workers
OUTRO TRAB.LIMPEZA MANUAL	Unskilled workers
DISTRIBUIDOR MERCADORIAS SIMILARES	Unskilled workers
PORTEIRO DE HOTELARIA	Personal, safety and security services workers and vendors
OUT.PROFISSOES ELEMENTARES DIVERSAS N.E.	Unskilled workers
EMPREGADO LAVABOS SIMILARES	Unskilled workers
OUT.TRAB.POLIVALENTES	Unskilled workers
TRAB.NAO QUALIF.AGRIC.(EXCLUI HORTICULTURA FLORICULTURA)	Unskilled workers
TRAB.NAO QUALIF.PROD.ANIMAL	Unskilled workers
TRAB.NAO QUALIF.AGRIC.PROD.ANIMAL COMBINADAS	Unskilled workers
TRAB.NAO QUALIF.FLORESTA	Unskilled workers
TRAB.NAO QUALIF.PESCA	Unskilled workers
TRAB.NAO QUALIF.AQUICULTURA	Unskilled workers
TRAB.NAO QUALIF.DAS MINAS	Unskilled workers
TRAB.NAO QUALIF.DAS PEDREIRAS	Unskilled workers
TRAB.NAO QUALIF.CONSTRUCAO EDIFICIOS	Unskilled workers
TRAB.TRIAGEM RESIDUOS	Unskilled workers
EMBALADOR MANUAL INDUSTRIA TRANSFORMADORA	Unskilled workers
TANOEIRO EMBUTIDOR OUT.SIMILARES A MARCENEIRO	Farmers and skilled workers in agriculture, fishing and forestry
SUPERVISOR CARGAS DESCARGAS	Intermediate level technicians and professions
SEM PROFISSAO	Unknown
DIRIGENTE SUPERIOR ADM.PUB.	Representatives of the legislative power and executive bodies, directors and executive managers
DIRIGENTE SUPERIOR ADM.PUB.	Representatives of the legislative power and executive bodies, directors and executive managers

Profession – original	Profession – aggregation
DIRIGENTE SUPERIOR ADM.PUB.	Representatives of the legislative power and executive bodies, directors and executive managers
DIRIGENTE ORGANIZACOES INTERESSE ESPECIAL	Representatives of the legislative power and executive bodies, directors and executive managers
OUT.DIR.SVC NEGOCIOS ADM.	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERAL GESTOR EXEC.EMPRESAS	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERAL GESTOR EXEC.EMPRESAS	Representatives of the legislative power and executive bodies, directors and executive managers
PRODUTOR DE CINEMA E TEATRO	Specialists in intellectual and scientific activities
PRODUTOR DE CINEMA E TEATRO	Specialists in intellectual and scientific activities
PROD.REALIZADOR TELEVISAO RADIO	Specialists in intellectual and scientific activities
ANALISTA EM GESTAO ORGANIZACAO	Specialists in intellectual and scientific activities
OUTROS AGENTES DE NEGOCIO	Intermediate level technicians and professions
OUTROS AGENTES DE NEGOCIOS	Intermediate level technicians and professions
DIR.GERENTE OUT.SVC N.E.	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERAL GESTOR EXEC.EMPRESAS	Representatives of the legislative power and executive bodies, directors and executive managers
DIR.GERAL GESTOR EXEC.EMPRESAS	Representatives of the legislative power and executive bodies, directors and executive managers
COMERCIANTE LOJA (ESTABELECIMENTO)	Personal, safety and security services workers and vendors
OUTROS OFICIAIS DO EXERCITO	Armed Forces Professions
OUTROS OFICIAIS DA MARINHA E EQUIPARADOS	Armed Forces Professions
ARQUITETO DE EDIFICIOS	Specialists in intellectual and scientific activities
FISICO	Specialists in intellectual and scientific activities
OUT.TEC.DAS CIENCIAS FISICAS ENG.N.E.	Intermediate level technicians and professions
FISICO	Specialists in intellectual and scientific activities
ASTRONOMO	Specialists in intellectual and scientific activities
METEOROLOGISTA	Specialists in intellectual and scientific activities
QUIMICO	Specialists in intellectual and scientific activities
GEOLOGO	Specialists in intellectual and scientific activities
OCEANOGRAFO	Specialists in intellectual and scientific activities
MATEMATICO	Specialists in intellectual and scientific activities
ACTUARIO	Specialists in intellectual and scientific activities

Profession – original	Profession – aggregation
ENGENHEIRO SISTEMAS(INFORMATICA)	Specialists in intellectual and scientific activities
ANALISTA SISTEMAS(INFORMATICA	Specialists in intellectual and scientific activities
ARQUITECTOS,ENGENHEIROS E ESPECIALISTAS SIMILARES	Specialists in intellectual and scientific activities
ARQUITECTO	Specialists in intellectual and scientific activities
URBANISTA	Specialists in intellectual and scientific activities
ENGENHEIRO CIVIL	Specialists in intellectual and scientific activities
ENGENHEIRO ELECTROTECNICO	Specialists in intellectual and scientific activities
ENGENHEIRO MECANICO	Specialists in intellectual and scientific activities
ENGENHEIRO NAVAL	Specialists in intellectual and scientific activities
ENGENHEIRO QUIMICO	Specialists in intellectual and scientific activities
ENGENHEIRO DE MINAS	Specialists in intellectual and scientific activities
ENGENHEIRO METALURGICO	Specialists in intellectual and scientific activities
CARTOGRAFO E AGRIMENSOR	Specialists in intellectual and scientific activities
CARTOGRAFO E AGRIMENSOR	Specialists in intellectual and scientific activities
BIOLOGO	Specialists in intellectual and scientific activities
BIOLOGO	Specialists in intellectual and scientific activities
MEDICO ESPECIALIDADES TECNICAS	Specialists in intellectual and scientific activities
FARMACOLOGISTA OUT.ESPEC.RELACIONADOS	Specialists in intellectual and scientific activities
FARMACOLOGISTA OUT.ESPEC.RELACIONADOS	Specialists in intellectual and scientific activities
MEDICO ESPECIALIDADES TECNICAS	Specialists in intellectual and scientific activities
FARMACOLOGISTA OUT.ESPEC.RELACIONADOS	Specialists in intellectual and scientific activities
ENGENHEIRO AGRONOMO	Specialists in intellectual and scientific activities
ENGENHEIRO FLORESTAL	Specialists in intellectual and scientific activities
ENGENHEIRO INDUSTRIAL PROD.	Specialists in intellectual and scientific activities
MEDICO MEDICINA GERAL FAMILIAR	Specialists in intellectual and scientific activities
MEDICO DENTISTA	Specialists in intellectual and scientific activities
VETERINARIO	Specialists in intellectual and scientific activities
FARMACEUTICO	Specialists in intellectual and scientific activities

Profession – original	Profession – aggregation
ENFERMEIRO DE CUIDADOS GERAIS	Specialists in intellectual and scientific activities
PROFESSOR DOS ENSINOS UNIVERSITARIO SUPERIOR	Specialists in intellectual and scientific activities
PROFESSOR DOS ENSINOS BASICO (2º 3º CICLOS) SECUNDARIO	Specialists in intellectual and scientific activities
ESPEC.EM METODOS ENSINO	Specialists in intellectual and scientific activities
ESPEC.DO TRAB.SOCIAL	Specialists in intellectual and scientific activities
ESPEC.EM METODOS ENSINO	Specialists in intellectual and scientific activities
FORMADOR EM TEC.INF.	Specialists in intellectual and scientific activities
OUTROS PROFESSORES DE MUSICA	Specialists in intellectual and scientific activities
OUT.ESPEC.DO ENSINO N.E.	Specialists in intellectual and scientific activities
CONTABILISTA AUDITOR REVISOR OFIC.CONTAS SIMILARES	Specialists in intellectual and scientific activities
ESPEC.EM RECURSOS HUMANOS	Specialists in intellectual and scientific activities
ESPEC.EM RELACOES PUB.S	Specialists in intellectual and scientific activities
ESPEC.EM PUBLICIDADE MARKETING	Specialists in intellectual and scientific activities
ADVOGADO	Specialists in intellectual and scientific activities
ADVOGADO	Specialists in intellectual and scientific activities
MAGISTRADO (JUDICIAL DO MINISTERIO PUBLICO)	Specialists in intellectual and scientific activities
MAGISTRADO (JUDICIAL DO MINISTERIO PUBLICO)	Specialists in intellectual and scientific activities
CONSERVADOR DOS REGISTOS CIVIL AUTOMOVEL COMERCIAL PREDIAL	Specialists in intellectual and scientific activities
NOTARIO	Specialists in intellectual and scientific activities
ARQUIVISTA	Specialists in intellectual and scientific activities
CURADOR DE MUSEUS	Specialists in intellectual and scientific activities
ECONOMISTA	Specialists in intellectual and scientific activities
CONTABILISTA AUDITOR REVISOR OFIC.CONTAS SIMILARES	Specialists in intellectual and scientific activities
SOCIOLOGO	Specialists in intellectual and scientific activities
SOCIOLOGO	Specialists in intellectual and scientific activities
ANTROPOLOGO E SIMILARES	Specialists in intellectual and scientific activities
ARQUEOLOGO	Specialists in intellectual and scientific activities
GEOGRAFO	Specialists in intellectual and scientific activities

Profession – original	Profession – aggregation
HISTORIADOR	Specialists in intellectual and scientific activities
FILOLOGO	Specialists in intellectual and scientific activities
FILOLOGO	Specialists in intellectual and scientific activities
TRADUTOR	Specialists in intellectual and scientific activities
PSICOLOGO	Specialists in intellectual and scientific activities
ASSISTANT SOCIAL	Specialists in intellectual and scientific activities
AUTOR E ESCRITOR	Specialists in intellectual and scientific activities
JORNALISTA	Specialists in intellectual and scientific activities
ESCULTOR	Specialists in intellectual and scientific activities
DESIGNER GRAFICO OU COMUNICACAO MULTIMEDIA	Specialists in intellectual and scientific activities
OUT.ARTISTAS ARTES VISUAIS	Specialists in intellectual and scientific activities
COMPOSITOR	Specialists in intellectual and scientific activities
MUSICO	Specialists in intellectual and scientific activities
MUSICO	Specialists in intellectual and scientific activities
CANTOR	Specialists in intellectual and scientific activities
COREOGRAFO	Specialists in intellectual and scientific activities
ATOR	Specialists in intellectual and scientific activities
MINISTRE DU CULTE/MEMBRE ORDRE RELIGIEUSE | MINISTRO DE CULTO	Specialists in intellectual and scientific activities
TEC.NIVEL INTERMEDIO ESTATISTICA MATEMATICA SIMILARES	Intermediate level technicians and professions
TECNICO CIENCIAS FISICAS	Intermediate level technicians and professions
TECNICO DAS CIENCIAS QUIMICAS	Intermediate level technicians and professions
TECNICO DE ENGENHARIA CIVIL	Intermediate level technicians and professions
TECNICO DE ELECTRICIDADE	Intermediate level technicians and professions
TECNICO DE ELECTRONICA	Intermediate level technicians and professions
TECNICO DE ELECTRONICA	Intermediate level technicians and professions
TEC.MANUT.E REP/CAO MOTORES AVIAO	Intermediate level technicians and professions
TECNICO DE ELECTRONICA	Intermediate level technicians and professions
INSPECTORES TEC.SAUDE DO TRAB.AMBIENTE	Intermediate level technicians and professions

Profession – original	Profession – aggregation
TECNICO DE GAS	Intermediate level technicians and professions
DESENHADORES E TECNICOS AFINS	Intermediate level technicians and professions
TOPOGRAFO E SIMILARES	Specialists in intellectual and scientific activities
CARTOGRAFO E AGRIMENSOR	Specialists in intellectual and scientific activities
CARTOGRAFO E AGRIMENSOR	Specialists in intellectual and scientific activities
CARTOGRAFO E AGRIMENSOR	Specialists in intellectual and scientific activities
OUT.TEC.DAS CIENCIAS FISICAS ENG.N.E.	Intermediate level technicians and professions
OUT.TEC.DAS CIENCIAS FISICAS ENG.N.E.	Intermediate level technicians and professions
INSPECTORES TEC.SAUDE DO TRAB.AMBIENTE	Intermediate level technicians and professions
PROGRAMADOR DE APLICACOES	Specialists in intellectual and scientific activities
TEC.APOIO AOS UTILIZADORES DAS TEC.INF.COMUNICACAO (TIC)	Intermediate level technicians and professions
TEC.OPER.DAS TEC.INF.COMUNICACAO (TIC)	Intermediate level technicians and professions
OUT.TEC.CONTROLO PROCESSOS INDUSTRIAIS	Intermediate level technicians and professions
TEC.GRAVACAO AUDIOVISUAL	Intermediate level technicians and professions
FOTOGRAFO	Intermediate level technicians and professions
TEC.SIST.DE COMUNICACOES VIA RADIO	Intermediate level technicians and professions
TEC.SIST.DE COMUNICACOES VIA RADIO	Intermediate level technicians and professions
OFIC.CONVES PILOTO NAVIOS	Intermediate level technicians and professions
OFIC.CONVES PILOTO NAVIOS	Intermediate level technicians and professions
PILOTO DE AERONAVES	Intermediate level technicians and professions
CONTROLADOR DE TRAFEGO AEREO	Intermediate level technicians and professions
TEC.SEGUR.SIST.ELECTRONICOS AERONAUTICOS	Intermediate level technicians and professions
TECNICO DE ENGENHARIA CIVIL	Intermediate level technicians and professions
TECNICO DE ENGENHARIA CIVIL	Intermediate level technicians and professions
INSPECTORES TEC.SAUDE DO TRAB.AMBIENTE	Intermediate level technicians and professions
INSPECTORES TEC.SAUDE DO TRAB.AMBIENTE	Intermediate level technicians and professions
INSPECTORES TEC.SAUDE DO TRAB.AMBIENTE	Intermediate level technicians and professions
INSPECTORES TEC.SAUDE DO TRAB.AMBIENTE	Intermediate level technicians and professions

Profession – original	Profession – aggregation
OUT.AGENTES NIVEL INTERMEDIO ADM.PUB.P/ APLIC.LEI SIMILARES	Intermediate level technicians and professions
OUT.TEC.INSPECTORES MECANICA	Intermediate level technicians and professions
ESPEC.EM HIGIENE SAUDE AMBIENTAL LABORAL	Specialists in intellectual and scientific activities
ESPEC.EM HIGIENE SAUDE AMBIENTAL LABORAL	Specialists in intellectual and scientific activities
DIETISTA E NUTRICIONISTA	Specialists in intellectual and scientific activities
OPTOMETRISTA OPTICO OFTALMICO	Specialists in intellectual and scientific activities
TERAPEUTA ASSISTENTE DENTARIO	Specialists in intellectual and scientific activities
TEC.PROTESES MEDICAS DENTARIAS	Intermediate level technicians and professions
FISIOTERAPEUTA	Specialists in intellectual and scientific activities
TEC.ASSISTENTE VETERINARIOS	Intermediate level technicians and professions
TEC.ASSISTENTES FARM.	Intermediate level technicians and professions
PARTEIRA	Intermediate level technicians and professions
OUT.ESPEC.EM MEDICINA TRADICIONAL ALTERNATIVA	Specialists in intellectual and scientific activities
PROFESSOR DO ENSINO BASICO (1º CICLO)	Specialists in intellectual and scientific activities
EDUCADOR DE INFANCIA	Specialists in intellectual and scientific activities
PROFESSOR DO ENSINO ESPECIAL	Specialists in intellectual and scientific activities
PILOTO DE AERONAVES	Intermediate level technicians and professions
INSTRUTOR DE CONDUCAO	Personal, safety and security services workers and vendors
CORRETOR BOLSA CAMBISTA SIMILARES	Intermediate level technicians and professions
CORRETOR BOLSA CAMBISTA SIMILARES	Intermediate level technicians and professions
CORRETOR BOLSA CAMBISTA SIMILARES	Intermediate level technicians and professions
AGENTE DE SEGUROS	Intermediate level technicians and professions
EMPREGADO DAS AGENCIAS VIAGENS	Administrative staff
DIRECTOR DE VENDAS	Representatives of the legislative power and executive bodies, directors and executive managers
DELEGADO DE INFORMACAO MEDICA	Specialists in intellectual and scientific activities
REPRESENTANTE COMERCIAL	Intermediate level technicians and professions
AVALIADOR IMOVEIS SEGUROS OUT.BENS	Intermediate level technicians and professions
DESPACHANTE TRANSITARIO SIMILARES	Intermediate level technicians and professions

Profession – original	Profession – aggregation
DESPACHANTE TRANSITARIO SIMILARES	Intermediate level technicians and professions
DESPACHANTE TRANSITARIO SIMILARES	Intermediate level technicians and professions
TECNICO DA AREA DO EMPREGO	Intermediate level technicians and professions
OUTROS AGENTES DE NEGOCIOS	Intermediate level technicians and professions
REPRESENTANTE COMERCIAL	Intermediate level technicians and professions
TEC.NIVEL INTERMEDIO DOS SVC JURIDICOS RELACIONADOS	Intermediate level technicians and professions
SOLICITADOR	Specialists in intellectual and scientific activities
OUT.TEC.ADMINISTRATIVOS CONTABILIDADE	Intermediate level technicians and professions
TESOUREIRO	Intermediate level technicians and professions
TEC.NIVEL INTERMEDIO ESTATISTICA MATEMATICA SIMILARES	Intermediate level technicians and professions
INSPECTOR ALFANDEGA FRONTEIRA	Intermediate level technicians and professions
INSPECTOR ALFANDEGA FRONTEIRA	Intermediate level technicians and professions
AGENTE ADM.TRIBUTARIA	Intermediate level technicians and professions
AGENTE SVC SEGUR.SOCIAL	Intermediate level technicians and professions
INSPECTOR DETECTIVE POLICIA	Intermediate level technicians and professions
OUT.ARTISTAS INTERPRETES CRIATIVOS DAS ARTES DO ESPECTACULO	Specialists in intellectual and scientific activities
DESIGNER DE TEXTEIS E MODA	Specialists in intellectual and scientific activities
DECORADOR	Intermediate level technicians and professions
JORNALISTA	Specialists in intellectual and scientific activities
CANTOR	Specialists in intellectual and scientific activities
OUT.ARTISTAS INTERPRETES CRIATIVOS DAS ARTES DO ESPECTACULO	Specialists in intellectual and scientific activities
OUT.ATLETAS DESPORTISTAS COMPETICAO	Intermediate level technicians and professions
TOUREIRO CAVALEIRO TAUROMAQUICO OUT.PROF.SIMILARES	Intermediate level technicians and professions
OUTROS SARGENTOS DO EXERCITO	Armed Forces Professions
EMPREGADO ESCRITORIO EM GERAL	Administrative staff
DACTILOGRAFO OPER.PROCESSAMENTO TEXTO	Administrative staff
OPERADOR DE REGISTO DE DADOS	Administrative staff
TECNICO DE SECRETARIADO	Administrative staff

Profession – original	Profession – aggregation
CAIXA BANCARIO E SIMILAR	Administrative staff
OPER.DOS SVC ESTATISTICA FINANCEIROS SEGUROS	Administrative staff
EMPREGADO DE ARMAZEM	Administrative staff
OUT.TEC.NIVEL INTERMEDIO DAS ACTIVIDADES CULTURAIS ARTISTICA	Intermediate level technicians and professions
CONTROLADOR TRANSPORTES TERRESTRES PASSAGEIROS	Administrative staff
CARTEIRO E SIMILARES	Administrative staff
CHEFE DE ESTACAO DE CORREIOS	Intermediate level technicians and professions
BILHETEIRO	Personal, safety and security services workers and vendors
OPERADOR DE CAIXA	Personal, safety and security services workers and vendors
EMPREGADO BANCA NOS CASINOS OUT.EMPREGADOS APOSTAS	Personal, safety and security services workers and vendors
PENHORISTA E PRESTAMISTA	Administrative staff
COBRADOR (AGENTE COBRANCA E LEITURA)	Administrative staff
PESSOAL INF.ADMINISTRATIVA	Administrative staff
OPERADOR DE CENTRAL TELEFONICA	Administrative staff
OPERADOR DE CENTRAL TELEFONICA	Administrative staff
REPRES.DO PODER LEGISL.ORGAOS EXEC.S	Representatives of the legislative power and executive bodies, directors and executive managers
ASSISTENTES VIAGEM COMISSARIOS	Personal, safety and security services workers and vendors
FISCAL COBRADOR TRANSPORTES PUBLICOS	Personal, safety and security services workers and vendors
GUIA INTERPRETE	Personal, safety and security services workers and vendors
EMPREGADO DE APROVISIONAMENTO	Administrative staff
EMPREGADO DE ARMAZEM	Administrative staff
COZINHEIRO	Personal, safety and security services workers and vendors
EMPREGADO DE MESA	Personal, safety and security services workers and vendors
EMPREGADO DE BAR	Personal, safety and security services workers and vendors
TRIPULACAO CONVES NAVIOS SIMILARES	Plant and machine operators and assembly workers
AUXILIAR CUIDADOS CRIANCAS	Personal, safety and security services workers and vendors
OUT.TRAB.DOS CUIDADOS PESSOAIS SIMILARES NOS SVC SAUDE	Personal, safety and security services workers and vendors

Profession – original	Profession – aggregation
OUT.TRAB.DOS CUIDADOS PESSOAIS SIMILARES NOS SVC SAUDE	Personal, safety and security services workers and vendors
CABELEIREIRO E BARBEIRO	Personal, safety and security services workers and vendors
ESTETICISTA	Personal, safety and security services workers and vendors
MANICURA,PEDICURA E CALISTA	Personal, safety and security services workers and vendors
AGENTE FUNERARIO	Personal, safety and security services workers and vendors
OUT.TEC.NIVEL INTERMEDIO DAS ACTIVIDADES CULTURAIS ARTISTICA	Intermediate level technicians and professions
DISC JOCKEY	Specialists in intellectual and scientific activities
ASTROLOGO	Personal, safety and security services workers and vendors
BOMBEIRO	Personal, safety and security services workers and vendors
OUTROS AGENTES DE POLICIA	Personal, safety and security services workers and vendors
OUTRO PESSOAL DOS SVC PROTECCAO SEGUR.	Personal, safety and security services workers and vendors
PESSOAL DE AMBULANCIAS	Intermediate level technicians and professions
MANEQUIM E OUTROS MODELOS	Personal, safety and security services workers and vendors
OUT.TRAB.RELACIONADOS C/VNDS N.E.	Personal, safety and security services workers and vendors
OPERADOR DE CAIXA	Personal, safety and security services workers and vendors
AGRIC.TRAB.QUALIF.CEREAIS OUT.CULT.EXTENSIVAS	Farmers and skilled workers in agriculture, fishing and forestry
AGRIC.TRAB.QUALIF.CULT.ARVORES ARBUSTOS	Farmers and skilled workers in agriculture, fishing and forestry
AGRIC.TRAB.QUALIF.HORTICULTURA	Farmers and skilled workers in agriculture, fishing and forestry
TRAB.QUALIF.JARDINAGEM	Farmers and skilled workers in agriculture, fishing and forestry
PROD.TRAB.QUALIF.NA PROD.OUT.ANIMAIS CARNE	Farmers and skilled workers in agriculture, fishing and forestry
APICULTOR TRAB.QUALIF.APICULTURA	Farmers and skilled workers in agriculture, fishing and forestry
APICULTOR TRAB.QUALIF.APICULTURA	Farmers and skilled workers in agriculture, fishing and forestry
OUT.TRAB.QUALIF.S FLORESTA SIMILARES	Farmers and skilled workers in agriculture, fishing and forestry
AQUICULTOR (AQUACULTOR) TRAB.QUALIF.AQUICULTURA AGUAS INTERI	Farmers and skilled workers in agriculture, fishing and forestry
PESCADOR MARINHEIRO PESCADOR PESCA MARITIMA COSTEIRA	Farmers and skilled workers in agriculture, fishing and forestry
TRABALHADOR DAS PEDREIRAS	Plant and machine operators and assembly workers
CANTEIRO	Skilled workers in industry, construction and crafts

Profession – original	Profession – aggregation
TRAB.NAO QUALIF.CONSTRUCAO EDIFICIOS	Unskilled workers
PEDREIRO	Skilled workers in industry, construction and crafts
CALCETEIRO	Skilled workers in industry, construction and crafts
CIMENTEIRO/ARMADOR DE FERRO	Skilled workers in industry, construction and crafts
OUT.CARPINTEIROS SIMILARES	Skilled workers in industry, construction and crafts
MONTADOR DE ANDAIMES	Skilled workers in industry, construction and crafts
OUT.TRAB.QUALIF.S CONSTRUCAO ESTRUTURAS BASICAS SIMILARES N.	Skilled workers in industry, construction and crafts
OUT.ASSENTADORES REVESTIMENTOS	Skilled workers in industry, construction and crafts
ESTUCADOR	Skilled workers in industry, construction and crafts
TRAB.QUALIF.EM ISOLAMENTOS ACUSTICOS TERMICOS	Skilled workers in industry, construction and crafts
VIDRACEIRO	Skilled workers in industry, construction and crafts
CANALIZADOR	Skilled workers in industry, construction and crafts
PINTOR DE CONSTRUCOES	Skilled workers in industry, construction and crafts
PINTOR-DECORADOR VIDRO CERAMICA OUT.MAT.	Skilled workers in industry, construction and crafts
LIMPADOR CHAMINES OUT.ESTRUTURAS EDIFICIOS	Skilled workers in industry, construction and crafts
SOLDADOR	Skilled workers in industry, construction and crafts
FUNILEIRO E CALDEIREIRO	Skilled workers in industry, construction and crafts
BATE-CHAPA VEIC.AUTO.	Skilled workers in industry, construction and crafts
OUTRO PREP/DOR MONTADOR ESTRUTURAS METALICAS	Skilled workers in industry, construction and crafts
SERRALHEIRO CIVIL	Skilled workers in industry, construction and crafts
ARMADOR MONTADOR CABOS METALICOS	Skilled workers in industry, construction and crafts
MERGULHADOR	Skilled workers in industry, construction and crafts
FORJADOR E FERREIRO	Skilled workers in industry, construction and crafts
OPER.PRENSA FORJAR ESTAMPADOR SIMILARES	Skilled workers in industry, construction and crafts
FORJADOR E FERREIRO	Skilled workers in industry, construction and crafts
REG.OPER.MAQ.-FERRA/AS CONVENCIONAIS P/ TRAB.METAIS	Skilled workers in industry, construction and crafts
MECANICO REP/DOR VEIC.AUTO.	Skilled workers in industry, construction and crafts
TEC.MANUT.E REP/CAO MOTORES AVIAO	Skilled workers in industry, construction and crafts

Profession – original	Profession – aggregation
MECANICO REP/DOR MAQ.AGRICOLAS INDUSTRIAIS	Skilled workers in industry, construction and crafts
ELECTROMECANICO ELECTRICISTA OUT.INSTALADORES MAQ.EQUIP.ELEC	Skilled workers in industry, construction and crafts
MECANICO REP/DOR EQUIP.ELECTRONICOS	Skilled workers in industry, construction and crafts
INSTALADOR REP/DOR TEC.INF.COMUNICACAO	Skilled workers in industry, construction and crafts
INSTALADOR REP/DOR LINHAS ELECTRICAS	Skilled workers in industry, construction and crafts
TRAB.QUALIF.DO FAB.REP/CAO INSTRUMENTOS PRECISAO	Skilled workers in industry, construction and crafts
TRAB.QUALIF.DO FAB.AFINACAO INSTRUMENTOS MUSICAIS	Skilled workers in industry, construction and crafts
JOALHEIRO	Skilled workers in industry, construction and crafts
OUTROS OLEIROS E SIMILARES	Skilled workers in industry, construction and crafts
POLIDOR ACABADOR ARTIGOS VIDRO	Skilled workers in industry, construction and crafts
LAPIDADOR GRAVADOR VIDRO CERAMICA OUT.MAT.	Skilled workers in industry, construction and crafts
PINTOR-DECORADOR VIDRO CERAMICA OUT.MAT.	Skilled workers in industry, construction and crafts
OUT.TRAB.QUALIF.S DO FAB.INSTRUMENTOS PRECISAO ARTESAOS SIMI	Skilled workers in industry, construction and crafts
OPERADOR DE PRE-IMPRESSAO	Skilled workers in industry, construction and crafts
OUT.PREP/DORES CARNE PEIXE SIMILARES	Skilled workers in industry, construction and crafts
PADEIRO	Skilled workers in industry, construction and crafts
TRAB.DO FAB.PROD.LACTEOS	Skilled workers in industry, construction and crafts
MARCENEIRO	Skilled workers in industry, construction and crafts
ALFAIATE E COSTUREIRO	Skilled workers in industry, construction and crafts
CHAPELEIRO	Skilled workers in industry, construction and crafts
OPER.MAQ.COSTURA	Skilled workers in industry, construction and crafts
BORDADOR	Skilled workers in industry, construction and crafts
VENDEDOR EM QUIOSQUE EM MERCADOS	Personal, safety and security services workers and vendors
ESTOFADOR	Skilled workers in industry, construction and crafts
PREPARADOR E ACABADOR DE PELES	Skilled workers in industry, construction and crafts
SAPATEIRO	Skilled workers in industry, construction and crafts
TRAB.FAB.FOGUETES (FOGUETEIRO)	Skilled workers in industry, construction and crafts
OPERADOR DE FUNDICAO	Skilled workers in industry, construction and crafts

Profession – original	Profession – aggregation
OPER.INSTALACOES P/ O TRAB.MADEIRA CORTICA	Skilled workers in industry, construction and crafts
OPER.INSTALACOES P/ O FAB.PASTA PAPEL PAPEL	Skilled workers in industry, construction and crafts
OPER.INSTALACOES MAQ.P/ MOAGEM SUBSTANCIAS QUI.	Skilled workers in industry, construction and crafts
TEC.OPERACAO INSTALACOES PROD.ENERGIA	Intermediate level technicians and professions
TEC.OPERACAO INCINERADORES	Intermediate level technicians and professions
ENC.DAS IND.REF.DO PET.QUI.PROD.FARM.TRANSF.MAT.PLASTICAS BO	Intermediate level technicians and professions
OPER.MAQ.P/ O FAB.PROD.FOTOGRAFICOS	Plant and machine operators and assembly workers
OPER.MAQ.P/ O FAB.PROD.BORRACHA	Plant and machine operators and assembly workers
OPER.MAQ.P/ O FAB.PROD.MAT.PLASTICAS	Plant and machine operators and assembly workers
OUTROS OPERADORES DE IMPRESSAO	Plant and machine operators and assembly workers
OPER.MAQ.P/ PREP/R FIAR BOBINAR FIBRAS TEXTEIS	Plant and machine operators and assembly workers
OPER.MAQ.FAB.CALCADO SIMILARES	Plant and machine operators and assembly workers
OPER.MAQ.PREP/CAO CARNE PEIXE	Plant and machine operators and assembly workers
OPER.MAQ.MOAGEM CEREAIS TRANSF.ARROZ FABRICACAO RACOES	Plant and machine operators and assembly workers
OPER.MAQ.P/ PREP/CAO CHA CAFE CACAU	Plant and machine operators and assembly workers
OPER.MAQ.P/ PREP/CAO VINHOS OUT.BEBIDAS	Plant and machine operators and assembly workers
OPER.MAQ.P/ O FAB.DO TABACO	Plant and machine operators and assembly workers
MAQUINISTA DE LOCOMOTIVAS	Plant and machine operators and assembly workers
MOTORISTA AUTO.LIGEIROS CARRINHAS	Plant and machine operators and assembly workers
TRIPULACAO CONVES NAVIOS SIMILARES	Plant and machine operators and assembly workers
TRIPULACAO CONVES NAVIOS SIMILARES	Plant and machine operators and assembly workers
OUT.TRAB.POLIVALENTES	Unskilled workers
VENDEDOR EM QUIOSQUE EM MERCADOS	Personal, safety and security services workers and vendors
PRESTADOR DE SERVICOS NA RUA	Unskilled workers
TRAB.LIMPEZA EM CASAS PARTICULARES	Unskilled workers
TRAB.LIMPEZA EM ESCRITORIOS HOTEIS OUT.ESTABELECIMENTOS	Unskilled workers
TRAB.LIMPEZA EM ESCRITORIOS HOTEIS OUT.ESTABELECIMENTOS	Unskilled workers
LAVADEIRO E ENGOMADOR DE ROUPA	Unskilled workers

Profession – original	Profession – aggregation
MEMBRO ORDEM RELIGIOSA TEC.APOIO RELIGIOSO	Intermediate level technicians and professions
COLOCADOR ANUNCIOS (MONTADOR ANUNCIOS)	Unskilled workers
AUXILIAR APOIO ADMINISTRATIVO (CONTINUO)	Unskilled workers
ESTAFETA	Unskilled workers
BAGAGEIRO	Unskilled workers
SEGUR.(VIGILANTE PRIVADO) OUT.PORTEIROS SIMILARES	Personal, safety and security services workers and vendors
SEGUR.(VIGILANTE PRIVADO) OUT.PORTEIROS SIMILARES	Personal, safety and security services workers and vendors
OUTRO PESSOAL DOS SVC PROTECCAO SEGUR.	Personal, safety and security services workers and vendors
CANTONEIRO DE LIMPEZA	Unskilled workers
COVEIRO	Unskilled workers
AQUICULTOR (AQUACULTOR) TRAB.QUALIF.AQUICULTURA AGUAS MARITI	Farmers and skilled workers in agriculture, fishing and forestry
TRAB.NAO QUALIF.ENG.CIVIL	Unskilled workers
OUT.TRAB.NAO QUALIF.S INDUSTRIA TRANSFORMADORA	Unskilled workers
CONDUTOR VEIC.TRACCAO ANIMAL	Unskilled workers
CARREGADORES DESCARREGADORES NAO QUALIF.S MERCADORIAS	Unskilled workers
ASSISTENTE ESTACAO SVC AO CONDUTOR	Personal, safety and security services workers and vendors
DOMESTICA/DONA DE CASA	Unknown
ESTUDANTE	Student
SEM PROFISSAO	Unknown

Table 26 – Original categories and mapping to aggregated categories in the profession variable

Appendix 8. PEARSON CORRELATION

In the matrix shown below, the greener the value the closer to 1, i.e. positive correlation; the redder the value, the closer to -1, i.e. negative correlation.

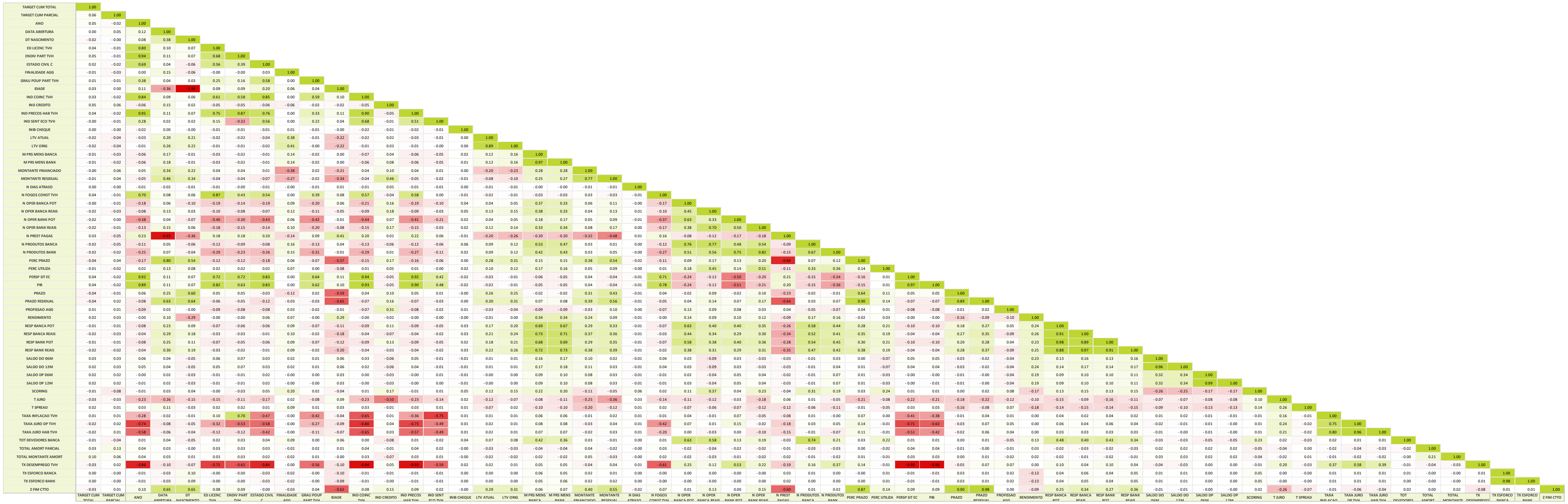

Figure 77 – Pearson correlation matrix

Appendix 9. STEPWISE REGRESSION VARIABLE SELECTION – FULL REPAYMENT

Variables selected
DATA_ABERTURA
ED_LICENC_TVH
ESTADO_CIVIL_AGG
FINALIDADE_AGG
IDADE
INIB_CHEQUE
LTV_ATUAL
LTV_ORIG
MONTANTE_FINANCIADO
M_PRS_MENS_BANK
N_DIAS_ATRASO
N_PREST_PAGAS
N_PRODUTOS_BANK
RENDIMENTO
RESP_BANCA_REAIS
RESP_BANK_REAIS
SALDO_DO_06M
SALDO_DO_12M
TX_ESFORCO_BANCA
T_JURO
T_SPREAD
scoring
tot_devedores_banca
IND_CREDITO
IND_SENT_ECO_TVH

Variables selected
MONTANTE_RESIDUAL
N_OPER_BANCA_REAIS
N_OPER_BANK_POT
N_OPER_BANK_REAIS
PERC_PRAZO
PROFISSAO_AGG
Perc_utiliza
RESP_BANCA_POT
RESP_BANK_POT
SALDO_DP_06M
TOTAL_AMORT_PARCIAL
TOTAL_MONTANTE_AMORT
TX_DIVORCIO_TVH
Z_FIM_CTTO

Table 27 – Variables selected using the Stepwise Regression in the full repayment

Appendix 10. STEPWISE REGRESSION VARIABLE SELECTION – PARTIAL REPAYMENT

Variables selected
ESTADO_CIVIL_AGG
FINALIDADE_AGG
LTV_ATUAL
LTV_ORIG
MONTANTE_FINANCIADO
M_PRS_MENS_BANK
M_PRS_MENS_banca
N_PREST_PAGAS
SALDO_DO_06M
SALDO_DO_12M
TX_ESFORCO_BANK
T_JURO
T_SPREAD
n_produtos_banca
scoring
IND_COINC_TVH
IND_CREDITO
MONTANTE_RESIDUAL
N_OPER_BANCA_POT
N_OPER_BANCA_REAIS
N_OPER_BANK_REAIS
PERC_PRAZO
PROFISSAO_AGG
Perc_utiliza
RESP_BANK_POT
SALDO_DP_06M
TAXA_JURO_DP_TVH

Variables selected
TOTAL_AMORT_PARCIAL
TOTAL_MONTANTE_AMORT
Z_FIM_CTTO

Table 28 – Variables selected using the Stepwise Regression in the partial repayment

Appendix 11. LASSO REGRESSION VARIABLE SELECTION – FULL REPAYMENT

Variables selected
ANO
PERC_PRAZO
TOTAL_AMORT_PARCIAL
PRAZO
TAXA_INFLACAO_TVH
SALDO_DO_06M
ENDIV_PART_TVH
TOTAL_MONTANTE_AMORT
T_SPREAD
DATA_ABERTURA
SALDO_DP_06M
T_JURO
MONTANTE_RESIDUAL
RENDIMENTO
TAXA_JURO_DP_TVH
N_OPER_BANCA_REAIS
PROFISSAO_AGG
FINALIDADE_AGG
TX_ESFORCO_BANCA
RESP_BANCA_REAIS
IDADE
M_PRS_MENS_BANK
IND_SENT_ECO_TVH
MONTANTE_FINANCIADO
N_DIAS_ATRASO
N_OPER_BANK_POT
IND_CREDITO

Variables selected
INIB_CHEQUE
n_produtos_banca
LTV_ATUAL
Perc_utiliza
SALDO_DP_12M
ESTADO_CIVIL_AGG
N_OPER_BANCA_POT
RESP_BANK_REAIS
N_PRODUTOS_BANK
M_PRS_MENS_banca
scoring
tot_devedores_banca
RESP_BANCA_POT
N_PREST_PAGAS
SALDO_DO_12M
LTV_ORIG
N_OPER_BANK_REAIS
PRAZO_RESIDUAL

Table 29 – Variables selected using the LASSO Regression in the full repayment

Appendix 12. LASSO Regression Variable Selection – Partial repayment

Variables selected
TOTAL_AMORT_PARCIAL
TOTAL_MONTANTE_AMORT
scoring
MONTANTE_FINANCIADO
N_PREST_PAGAS
N_OPER_BANCA_REAIS
LTV_ATUAL
PRAZO
RESP_BANK_REAIS
M_PRS_MENS_BANK
T_SPREAD
RENDIMENTO
MONTANTE_RESIDUAL
T_JURO
tot_devedores_banca
SALDO_DP_06M
LTV_ORIG
SALDO_DO_06M
PERC_PRAZO
FINALIDADE_AGG
PIB
TX_DESEMPREGO_TVH
ESTADO_CIVIL_AGG
IND_PRECOS_HAB_TVH
ANO
PROFISSAO_AGG
TX_ESFORCO_BANCA

Variables selected
ED_LICENC_TVH
IND_CREDITO
N_OPER_BANK_POT
Perc_utiliza
N_PRODUTOS_BANK
N_OPER_BANK_REAIS
DATA_ABERTURA
RESP_BANCA_POT
RESP_BANK_POT
INIB_CHEQUE
SALDO_DO_12M

Table 30 – Variables selected using the LASSO Regression in the partial repayment