NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

SYSTEM FOR FRAUD DETECTION: CUSTOMER SEGMENTATION AND PREDICTIVE ANALYSIS

VIA-VERDE PORTUGAL

Carolina Sottomayor Moser Machado

Project Work presented as partial requirement for obtaining the master's degree in Information Management

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

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

CUSTOMER SEGMENTATION AND PREDICTIVE ANALYSIS OF VIA VERDE

by

Carolina Sottomayor Moser Machado

Project Work presented as partial requirement for obtaining the master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Advisor: *Prof.* Rui Alexandre Henriques Gonçalves

November 2019

ACKNOLEDGMENTS

I would like to express my gratitude to all those which have supported, helped, commented or otherwise assisted me throughout the process.

First, I would like to thank my supervisor Professor Rui Gonçalves who helped me throughout this process with his guidance, knowledge, and advices.

Additionally, I would like to thank the University Nova Information Management School and all professors that thought me during this master's degree as their courses have contributed with knowledge presented in this work project.

Furthermore, given the consulting nature of this project, I appreciate all contributions received by Via Verde Portugal that provided information and data regarding the toll industry. This includes a special thank you to Eng.ª Margarida Cordeiro and Susana Ferraz.

Finally, I appreciate the support and the information sharing that I felt from my Brisa colleagues and most importantly, I thank my friends and family for listening to me and encouraging me to do my best. None of this could have happened without all of them.

ABSTRACT

Brisa - the largest Portuguese motorway operator in Portugal - focus on these main factors: the digital world, the affirmation of individual choice, the transport integration, the collaborative business models and the sharing economy. These are some of the aspects that make Brisa a benchmark of the motorway sector at an European level.

Consistent with this vision, Via Verde was created in 1991 and has been growing since then. Via Verde's mission is to make customers lives more practical. It offers greater speed and convenience to its customers, through faster ways of payment. In the end of 2018, Via Verde had 3.4 million of active identifiers, "the result of customer growth with 261 thousand new net subscribers compared to 2017", said Brisa's mobility services collection company. [2019]

Based on the relevant concepts of information, knowledge and intelligence associated with Data Mining, this paper seeks to identify and describe the practical applicability of this information technology in current and relevant areas of the company activity.

The initial phase of the study began with a literature review focused on the techniques and approaches behind the predictive analysis. This was followed by the data collection and validation process. The data collection phase is of great importance in the elaboration of any scientific research and all the care with this phase aims to guarantee the quality of the information. The data used in this dissertation was mainly provided by the company in study – Via Verde Portugal - and is a sample of 406 Million transactions for the year of 2017. The final stage consisted of a study caused by the processing of data and the development of the most appropriate model.

KEYWORDS

Beacon Technology; Payment Risk; Customer Segmentation; Fraud; Predictive Model; Neural Networks; Decision Trees; Regression;

INDEX

1. INTRODUCTION 1

1.1. Background and Theoretical Framework 1

1.2. Company History 1

1.3. Problem Identification 3

1.4. Study Objectives 4

2. LITERATURE REVIEW 6

2.1. Operational Risk 6

2.2. Fraud 9

2.3. Payment Fraud 13

2.4. Analytical Methods for Fraud Detection in Payment 13

2.4.1. Data Mining Processes 15

2.4.2. Predictive Models 18

2.4.2.1. Defining the Target 19

2.4.2.2. Logistic Regression 20

2.4.2.3. Decision Trees 21

2.3.2.3.1. Decision Trees Representation 21

2.3.2.3.2. Growing a Decision Tree 22

2.4.2.4. Artificial Neural Networks 26

2.4.2.5. Ensemble Models 29

3. METHODOLOGY 32

3.1. Research Methodology 32

3.2. Data Collection Process 32

3.3. Modelling 33

3.3.1. Sample 33

3.3.1.1. Variables 33

3.3.1.2. Sampling Techniques 34

3.3.1.3. Normalization 36

3.3.2. Explore 37

3.3.3. Modify 41

3.3.3.1. Target Variable - Binary 42

3.3.3.2. Dimensionality Reduction 42

3.3.3.3. Outliers 42

3.3.3.4. Data Partition 43

3.3.4. Model and Assess 44
3.3.4.1. Error Rate 45
3.3.4.2. Accuracy 46
3.3.4.3. Other Measures 46
3.3.4.4. ROC Curve and AUC (Area Under Curve) 47

4. RESULTS AND DISCUSSION 48

5. CONCLUSIONS 54

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS 58

7. BIBLIOGRAPHY 59

8. ANNEXES 64

LIST OF FIGURES

Figure 1 - "From the infrastructures era to the Mobility era" (Source: Author elaboration).. 2
Figure 2 - Evolution over the years of 2016 and 2017 (Source: Author elaboration, using data from 2016 and 2017 from Via Verde Portugal)..... 4
Figure 3 - Reasoning Process (Source: Author elaboration) 5
Figure 4: The "4-Cause Definition" of Operational Risk According to Basel (Source: Guidelines on Operational Risk Management, Oesterreichische Nationalbank (OeNB), 2006) 7
Figure 5 - Fraud Triangle Model (Source: Source: Wells, J. T., 2005. Principles of fraud examination. Hoboken, New York: John Wiley and Sons) 10
Figure 6 - Top Merchants Affected by Fraud Transactions (Source: ONLINE PAYMENT FRAUD WHITEPAPER, 2015) 13
Figure 7 - CRISP-DM Life Cycle (Source: Adapted from R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining")..... 16
Figure 8 – SEMMA (Source: Adapted from A. I. R. L. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview") 17
Figure 9 - Data Triangle methodology (Source: Author elaboration, using information from Reeve A. (2013). "Managing Data in motion: Data Integration. Best Practice Techniques and Technologies". 152-154) 19
Figure 10 - Sigmoid Function..... 21
Figure 11 - Illustration of the Decision Tree (Source: SAS Institute)..... 22
Figure 12 - Entropy of a Binary Variable 23
Figure 13 - *CHAID* structure - Decision Tree example (Source: Author elaboration) 25
Figure 14 - *CART* structure - Decision Tree example (Source: Author elaboration) 25
Figure 15 - Artificial Neural Network Representation 27
Figure 16 - Backpropagation Algorithm 27
Figure 17 - *Step* Activation Funtion..... 28
Figure 18 - *Sigmoid* Activation Funtion 28
Figure 19 – *Tanh* Activation Function 29
Figure 20 - *ReLU* Activation Function..... 29
Figure 21 - Bagging Method (Source: Author elaboration, using information from Bulhmann, P. (2012). "Bagging, Boosting and Ensemble Methods")..... 30
Figure 22 – Boosting Method (Source: Author elaboration, using information from Bulhmann, P. (2012). "Bagging, Boosting and Ensemble Methods")..... 31
Figure 23 - Stacking Method 31
Figure 24 - Workflow Diagram 32

Figure 25 - Analytical Model used in the study 33
Figure 26 - Distribution of the dataset according to the dependent variable (Debt) 34
Figure 27 – SMOTE (Source: Made by Author) 35
Figure 28 - SMOTE steps (Source: Made by Author) 36
Figure 29 - Multi-Class Dependent Variable (Source: Made by the author, PowerBI) 37
Figure 30 - Binary Dependent Variable (Source: Made by the author, PowerBI) 37
Figure 31 – Influence of Transaction Payment Method Variable (Source: Made by the author, PowerBI) 38
Figure 32 - Transaction Payment Method by Target Variable (Source: Made by the author, PowerBI) 38
Figure 33 - Influence of the Date Variable by Day of the Week (Source: Made by the author, PowerBI) 39
Figure 34 - Influence of the Date Variable by Day of the Month (Source: Made by the author, PowerBI) 39
Figure 35 - Influence of the Tariff Class (Source: Made by the author, PowerBI) 39
Figure 36 - Influence of the Rescission Date Variable (Source: Made by the author, PowerBI) 40
Figure 37 - Influence of the Type of Client Variable (Source: Made by the author, PowerBI) 40
Figure 38 – Number and Value of Observations per Client (Source: Made by the author, PowerBI) 41
Figure 39 - Binary Target Variable Code 42
Figure 40 - Partitions used in the Predictive Model 43
Figure 41 – ROC Curve (Source: Made by the author) 47
Figure 42 – Predictive Models ROC Curve (Source: SAS Miner) 49
Figure 43 – SAS Segment Size Profile (Source: SAS Miner) 52
Figure 44 – SAS Costumer Segment Profile (Source: SAS Miner) 52

LIST OF TABLES

Table 1 - Number of Via Verde enrollments (Source: Author elaboration, using data from 2013 to 2017 from Via Verde Portugal) 2
Table 2 - The Seven Operational Risk Event Types Projected by Basel II (Source: Bielski, 2003; BdP, 2010 e BCBS, 2004) 8
Table 3 - CRISP-DM & SEMMA (Source: Adapted from SAS) 18
Table 4 - Comparison of CHAID vs CART algorithm (Source: Author elaboration, , using information from Lu, Y. (2015). "Decision tree methods: applications for classification and prediction", Shanghai Archives of Psychiatry, 130-133) 26
Table 5 - Data Partition 43
Table 6 – Confusion Matrix 45
Table 7 – Statistic Comparison [Fit Statistic: _AUR_] 49
Table 8 – Statistic Comparison [Fit Statistic: GAIN and LIFTC] 50
Table 9 – Ensemble 10 Confusion Matrix 51
Table 10 – Neural 10 Confusion Matrix 51
Table 11 – Predictive model error rate for the binary dependent variable 55
Table 12 – Neural Network Success Percentage used in previous studies 55

LIST OF ABBREVIATIONS AND ACRONYMS

BCBS	Basel Committee on Banking Supervision
CART	Classification and Regression Trees
CHAID	Chi-Square Automatic Interaction Detection
DT	Decision Trees
MLP	Multi-Layer Perceptron
VVP	Via Verde Portugal

1. INTRODUCTION

This first chapter consists of a brief contextualization of the business, the background and the problem description. The main objectives are also presented below.

1.1. BACKGROUND AND THEORETICAL FRAMEWORK

To survive in today's competitive market, a company needs adequate technology to innovate the services and products it offers. Industrial Revolution in the mid-eighteenth century made a profound impact on the productive process, which reached both the economic and social levels. Since then, technology has been a part of the evolution of the companies. Every day companies are forced to adopt new technologies, which is necessitated by the society that is going through rapid changes. The challenge that big data is posing will continue to be one of the most exciting opportunities for the next years.

In the very world where the boundaries between industries, technologies and regulatory bodies are becoming increasingly indistinct, fraudsters are looking for soft targets to attack beyond their traditional ways. In PwC's 2018 Global Economic Crime and Fraud Survey, only 49% of global organizations admitted to having fallen victim to fraud and economic crime. However, Didier Lavion, the Principal of the Global Economic Crime and Fraud Survey Leader, says that that number should be much higher – about 51%. (PWC, 2018)

Naturally, the use of big data has become indispensable for the prevention of fraud. Existing technologies are now enough to analyse the data generated today. However, the future is still an unknown environment, and it is important for companies to recognise the benefits that these new technologies will bring to their businesses and should be encouraged to exploit these techniques. It is possible to gain a comprehensive look at all the channels connected by the clients that will be seen in a unique way, regardless of the connection. (Accenture Consulting, 2018)

A greater amount of information will be generated to outline the fraudster's profile and treat the customers differently, and new information will be added to the fraud prevention process. Furthermore, the speed of identifying these activities will increase, generating financial benefits for the companies and their customers. (Accenture Consulting, 2018)

1.2. COMPANY HISTORY

Brisa Portugal Highways (Brisa Autoestradas de Portugal) was founded in 1972, leading the market and establishing a recognized operational model for road infrastructures.

Due to the rapid evolution of technology, companies of all kinds face an increasingly competitive environment and more globally dispersed competition (Porter, 1990). To satisfy customer requirements while competing with other companies, it is important to create and develop an efficient flow that ensures the production of first-class products (Wann-Yih et al., 2004). These characteristics define a new era, in which the concept of mobility has taken on a broader significance whilst it has also introduced new challenges. Via Verde was born in this new era, in 1991, by relying on an innovative electronic payment system: Beacon technology.

Initially, Via Verde was created to prevent delays on toll roads and bridges by collecting tolls without requiring cars to stop. A tag is attached to the vehicle's windshield that transmits information about the vehicle, and the toll amount is debited directly from the driver's bank account. This payment involves no administrative costs and a fully integrated, cross-bank network.

For example, using the toll roads is the fastest way to get around Portugal: a drive from Lisbon to Aveiro, for example, would take just under 3 hours on the toll roads but around 5 and a half hours on the non-toll roads[1]

Currently, with recent updates to its system, Via Verde offers other services in addition to the collection of motorway tolls, such as ex-SCUTS, electronic fuel payment, and car park fee payment.

Over the years, Via Verde has become part of the daily routine of road users. It began with the development of a set of services under the Via Verde brand that use a collaborative ecosystem logic—functionalities and services that are complementary to each other, some operated directly by Via Verde, others integrated through a platform such as Via Verde Planner.

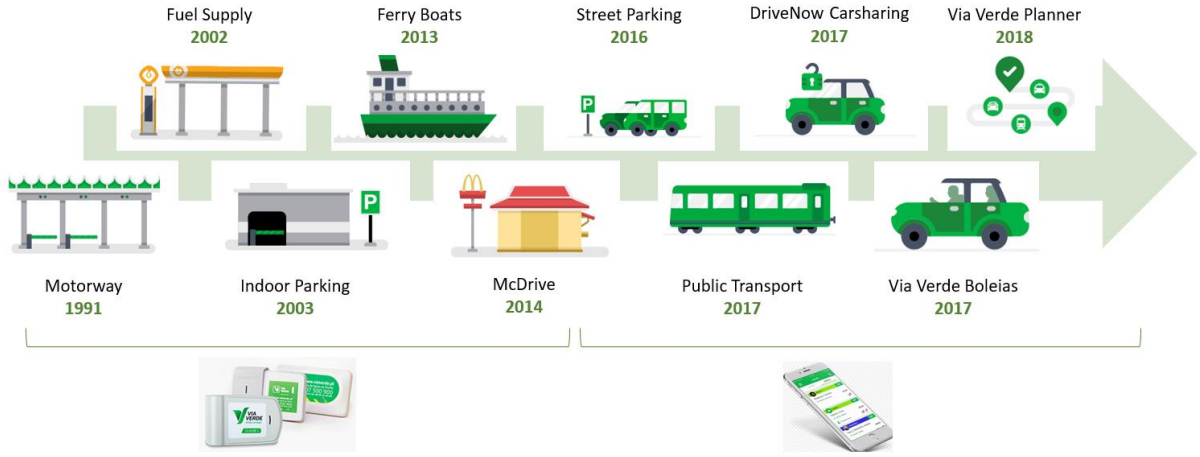


Figure 1 - "From the infrastructures era to the Mobility era"
(Source: Author elaboration)

These developments have had a significant impact on Via Verde's financial performance in recent years.

Year	Number of Enrollments
2013	96.4k
2014	107k
2015	128.5k
2016	167.2k
2017	174.9k

Table 1 - Number of Via Verde enrollments (Source: Author elaboration, using data from 2013 to 2017 from Via Verde Portugal)

[1] https://www.viamichelin.pt/

However, along with company growth has come the risk of non-payment by customers. It is important to mention that, of Via Verde's more than 759.700 clients, 55.700 clients have rescinded the contract—some forced by payment failure, the last phase of an infringement procedure, others by option.

When an infringement occurs is necessary to investigate what is the situation and how often it occurs. There are two main paths when analyzing infringement:

- When relates to McDrive, Gas Supply, Indoor Car Parking, Ferries, the service will be blocked, and no further use will be possible until the client regulates the situation.
- If is regarding the highway tolls the situation changes, since there is no physical way to stop vehicle to pass (e.g. toll barrier). This happens when there is a problem of registration of the transaction (registration of the vehicle that does not coincide with the registration in the contract, non-payment of transactions, alteration of the Multibanco card associated with the contract, or other reason) is triggered a yellow signal to warn of this problem.

In these cases, the holder of the Via Verde subscription contract may receive various notices (letters, mobile messages, e-mails, according to the data available in the contract) and, if the problem is not solved, Via Verde may terminate the contract (the identifier being invalid).

The infractions can occur for several reasons. The most common are the following situations:

- No identifier associated with the vehicle or with an invalid identifier (canceled / terminated). This situation occurs when, for example, a vehicle that does not have an active contract with Via Verde Portugal, SA, uses the toll road for vehicles equipped with a Via Verde identifier.
- Identifier with payment problems. To enroll in the Via Verde payment system, implies the existence of a valid debit card, to ensure the payment of debits resulting from the uses of the service. If the debit card associated with the identifier is invalid (expiry / renewal / cancellation or other reason, information provided by SIBS, which reflects the communication received from the bank issuing the ATM card), it causes non-payment of transactions.
- Ticket issued manually. An invoice issued for a passage made on a Via Manual barrier, provides that payment is to be made to the concessionaire within 8 days. After this deadline, if the payment is not verified, the process of administrative infraction is initiated.
- Violation of a manual barrier. This situation occurs whenever the payment of any toll rate is not made using the manual way.

1.3. PROBLEM IDENTIFICATION

Launched in 1991, this system developed by Brisa, known as the automatic toll collection system, was quickly expanded to other features, having exceeded 300 million transactions in 2014, and was considered a great value-added service for the customer. The widespread implementation of electronic toll systems offers great advantages in terms of efficiency, convenience, and safety, not only for payment but also for ease of travelling. However, there are some risks associated with this business, including the risk of non-payment. This study aims to construct a predictive model to prevent non-paying clients.

An infringement occurs when a vehicle passes through the Via Verde lane without having an active contract or, when it does have an active contract, it is invalid for some reason. In the event the Via

Verde contract is valid, non-payment can occur in cases where it is impossible for Via Verde to collect tickets.

The evolution of the infringements during the years 2016 and 2017 can be observed in the following graph:

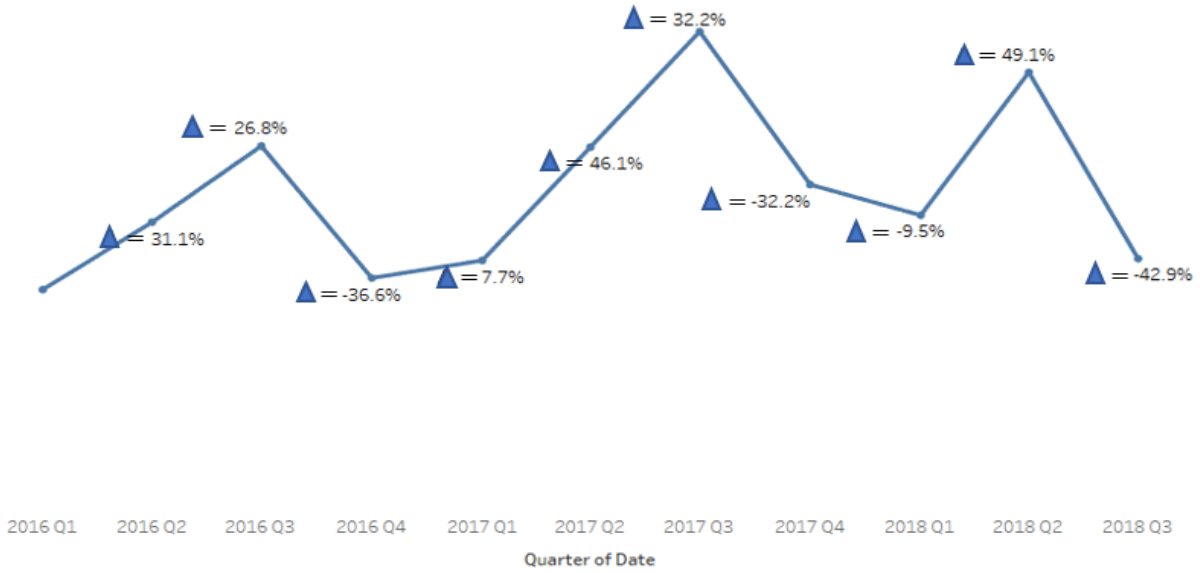


Figure 2 - Evolution over the years of 2016 and 2017 (Source: Author elaboration, using data from 2016 and 2017 from Via Verde Portugal)

There are certain variables that contributes to this grow - knowing which are the most relevant and weightier in the process can add value to the company. At first, we need to distinguish data from information. Although linked, these concepts do not have the same meaning. Data is all information and initially does not bring any value to the company, as it has not yet been filtered or interpreted. They are given with potential. Information is a set of data that has been categorized and useful information can be extracted. The important thing in this process is to be able to distinguish what can be used with relevance and what has no value whatsoever. For instance, the data provided only had the 4 first numbers of the Postal Code. This data, by itself, doesn't bring an insight to our study, however, when combining with the Location (provided by CTT Data Base) gives us the chance to see where the most part of costumers and the routes used are.

As seen, Data analysis is an important step in any decision-making process within a company. But, in order for such information to be relevant and to ensure positive returns for the business, it is essential for the organization to ensure data quality. This means consulting the origin of the facts, verifying the composition of the elements, evaluating the consistency of the available information, among other procedures. During the study this was one of the main struggles, a lot of variables had outliers and non-sense information. This is due to the fact that the main client information is provided by the costumer and might not be the most accurate or can be misspelled.

1.4. Study Objectives

In an attempt to respond to the problems related to the grow of the road toll non-payment, a general goal was set, and several specific objectives defined. This master's degree final project

consists on a payment fraud study analysis of the most known Beacon technology in Portugal – Via Verde.

The intent is to explore the design of a customer segmentation process, and which variables to consider. To achieve this purpose, the following questions will be answered:

1. Which variables should be included in customer segmentation?
2. How can a customer segmentation process be designed when looking to how often do they purchase and how much do they spend (Monetary Value - Identify the variables that most contribute to the success/ debt situation)?
3. How can we predict/ anticipate future frauds? (Apply predictive methods - MLP neural networks, regression, or decision tree)

A predictive model will be developed with the main objective of predicting whether a client will be in debt to Via Verde in the future based on the same variables as customer segmentation (e.g., variables associated with the possibility of losses associated with the non-fulfilment of a client's contract). The process will be developed through qualitative and quantitative research based on a literature review and data provided by the company. Only variables with high influence on the case study will be implemented in this project.

The diagram below describes the reasoning process:

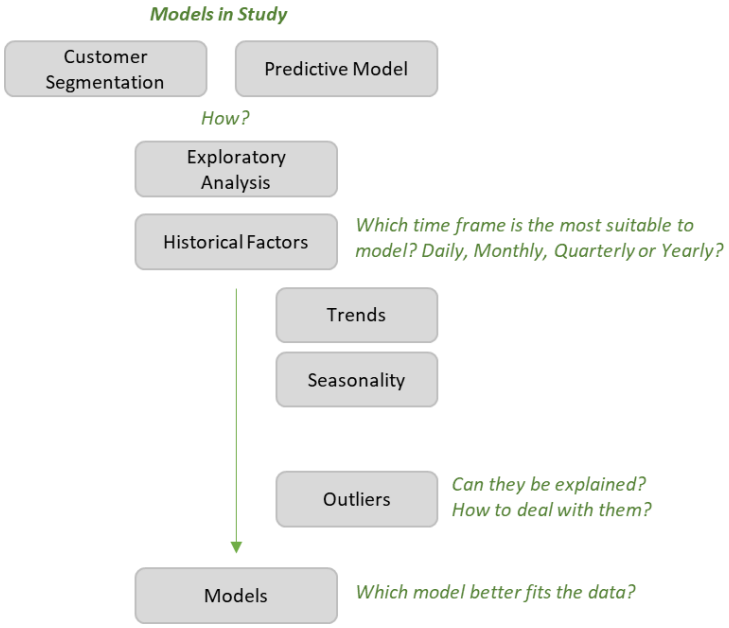


Figure 3 - Reasoning Process (Source: Author elaboration)

2. LITERATURE REVIEW

Risk management is a preponderant factor for the survival of any business. Solomon et al. (2000: 449), incorporates in the concept of risk all types of risks (financial and non-financial) that companies face and, considers, that risk can be understood as uncertainty as to the amount of results associated with both earning potential and loss exposure.

The type of risks can be distinguished according to their nature:

- Financial risk: when risk is directly related to the monetary assets and liabilities of the institution;
- Non-financial risk: when risk arises from circumstances (social, political or economic phenomena) or internal (human resources, technologies, procedures and others) to the institution;
- Other risks: specific risk whose negative impact results in an imbalance for the entire financial system, whether at the level of the country or the world

In this study we will develop the concept one of the Non-Financial Risk - Operational Risk, due to analysis failure, processing operations, internal and external fraud and insufficient or inadequate human resources [BdP (Bank of Portugal): Notice No. 5/2008, Article 11].

Operational Risk and Payment Fraud are key concepts for detecting fraudulent behavior. These will be developed in the following subchapters. In addition, identifying and exposing methods and fraud detection algorithms (namely, data mining and networks) will clarify and classify the various existing methods.

2.1. OPERATIONAL RISK

By virtue of its nature, the operational risk is inseparable from a company's business as it is related to all business activities. It is the most prevalent specific risk; all measures to control and mitigate its effects depend on the specific profile of an organisation. It is considered to be a cultural risk, since its approach and treatment practices particularly affect day-to-day business (Christl & Pribil, 2006).

One of the first definitions of operational risk was provided by an internationally recognised entity in 1993, which is allied to the uncertainty pertaining to the losses resulting from inadequate systems and controls, human error and management failures (Group of Thirty[2], quoted by Magalhães in 2012). Since then, the said concept has been developing, with several researchers and official entities presenting their own definitions.

In the context of the definition issued by the Basel Committee, Brink (2002), Chernobai, Rachev and Fabozzi (2007) defined that operational risk consists on the loss resulting from four dimensions – people (inadequate or failed internal processes), systems, processes and external events:

[2] Consultative Group on International Economic and Monetary Affairs, Inc.
The Group of Thirty aims to deepen the understanding of international economic and financial issues and explore the international repercussions of decisions taken in the public and private sectors. (Source: http://www.group30.org/)

- People – These can occur due to several facts, such as, lack of knowledge of the institution's products, commercial pressure to achieve objective, segregation of functions, misunderstanding, omission, distraction or negligence of employees or third parties engaged in and fraudulent behavior (tampering with controls, intentional non-compliance with standards, leakage of privileged information, misappropriation of information, misinformation). (Mestchian, 2003)

- Systems – The risks that can arise from this dimension include system failures caused by degradation problems, quality and data integrity, inadequate, application-related risks, hardware failures, and storage and data recovery, among others. (Mestchian, 2003)

- Processes – Refers to inefficiencies in the organization's business processes. It includes the processes of value movement, such as sales and marketing, product development and customer support, as well as all back office processes, such as information, human resources and operations processing. The occurrence of non-observance of operational norms and limits results in the lack of functioning of committees, non-compliance with credit, undue custody of confidential documents, non-implementation of controls, lack of compliance, lack of monitoring / conciliation and others. (Mestchian, 2003)
  The implementation of internal control procedures aims to avoid errors and risks, however, they may be incorrectly drawn and executed, becoming a source of risk, since they will be inefficient (Davis, 2005).

- External Events – Losses that may occur due to external events, such as, criminal acts, natural disasters, terrorism, money laundering, external fraud and information leakage by outsourcing companies. (Brink (2002).

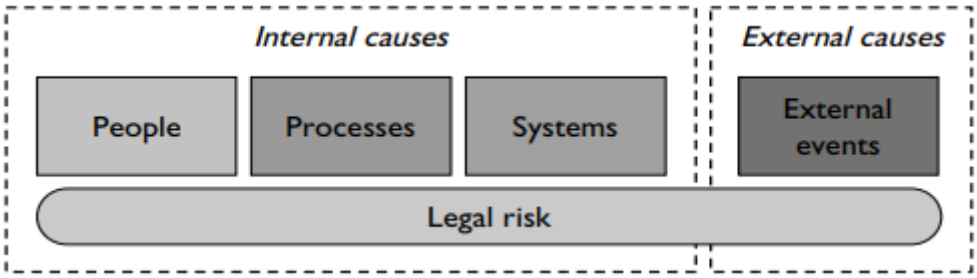


Figure 4: The "4-Cause Definition" of Operational Risk According to Basel
(Source: Guidelines on Operational Risk Management, Oesterreichische Nationalbank (OeNB), 2006)

Given the BCBS' definition, certain criticisms were made in this regard. Wahler (2002) argues that the aforementioned risk originates from internal and external sources: (i) change: external and internal causes; (ii) complexity: in products, processes and technology; (iii) complacency: inefficient management of the business and its risk, unlike the credit and market risk events that are influenced by the institution's transactions and business partners. Furthermore, according to Hadjiemmanuil (2003), the definition provided by the committee is "opaque" and "open" because it does not specify the factors that constitute operational risk or its relation to other forms of risk. Tattam (2011) refers to the BCBS' definition as "narrow" since it only mentions the risk loss, with no reference to the possibility concerning opportunity, positive consequence or monetary gain. Chaudhury (2010) points out that this risk is specific to the company and to certain operations and, unlike the market, the

credit, the interest rate and the foreign exchange risks, a higher level of operational risk exposure is generally not rewarded by higher expected return.

For Davis (2005), the difficulty of an operational risk is revealed when comparing it with other types of risks. At the credit risk level, the amount of loans made and to whom, with an approximation of the probability of non-compliance through the analysis of tenderers and with support for scoring models, are known at the outset. In the operational risk, these models have no capacity to predict risk or the degree of risk exposure (CHORAFAS, 2003). The fact that a process has never had an operational failure cannot be regarded, as it being exempt from potential risks. According to Akkizidis and Bouchereau (2006), the risk should only be accepted when the benefits outweigh the costs. Operational risks can only be eliminated if the institution no longer exists.

Herghiligiu and Cocris (2014) report that the types of operational risks are represented by categories/risk classes. These include internal fraud, external fraud, legal and liability losses, non-compliance with regulations, processing errors, information security breaches, inadequate business practice, disaster recovery, business continuity, and physical security failures. (Table 2)

Types of Operational Risk Event	**Definition**
Internal fraud	Acts of fraud committed internally in an organization go against its interest. Losses arising from acts intentionally committed to fraud, assets misappropriation or legislation, regulations or business policies circumventing.
External fraud	Losses arising from acts intentionally intended to commit fraud, misappropriation of assets or circumvention of legislation by a third part. Theft, check fraud, and breaching the system security like hacking or acquiring unauthorized information are the frequently encountered practices under external fraud.
Practices on employment and safety at the workplace	Losses arising from acts that are not in accordance with labor, health or safety legislation or agreements, as well as the payment of personal injury or acts related to differentiation/discrimination.
Customers, products and business practices	Losses arising from the intentional or negligent breach of a professional obligation in relation to specific customers (including fiduciary and fitness requirements) or the nature or design of a product.
Damage to physical assets	Losses arising from damage or loss caused to physical assets by natural disasters or other events (such as rapid and unexpected changes in climatic conditions)
Disturbance of business activities and system failures	Losses due to disturbance of business activities or system failures (hardware or software)
Execution, delivery and management of processes	Losses due to Failure in delivery, transaction or process management, as well as in relationships with commercial counterparts and vendors.

Table 2 - The Seven Operational Risk Event Types Projected by Basel II
(Source: Bielski, 2003; BdP, 2010 e BCBS, 2004)

One of the main operational risk challenges lies in developing a management approach that helps the top management to define the different categories of operational risk to be considered in each of the business lines. Under new regulation rules, each institution is allowed to adopt its own definition of operational risk. This individual definition is subject to certain requirements. It should provide a clear understanding of what an operational risk is, consider the material risks that the business faces, and include the main causes of operating losses (Walsh, 2003). It is essential to have an appropriate risk

management environment that identifies, assess, monitor, control and mitigates the risk, to establish contingency plans and to establish operational risk management policies and carry out their regular evaluation. The supervisor is responsible for (i) carrying out periodic inspections covering this risk, (ii) ensuring that banks establish and evaluate such policies and (iii) documenting and internally disseminating the processes and controlling the operational risk; identifying and evaluating the operational risk inherent in all products, activities, processes and systems, defining a minimum of loss and developing processes to periodically monitor the risk profile and exposure to significant losses, including reporting to the top management body and the regulator, are some of the measures presented by the Basel Committee which emphasize the development of an adequate risk management environment (BCBS, 2001). Goncalves (2011) adds that the Operational Risk Management based on an effective process/ system leads to reduction of losses and operational costs, market and investors image improvement, and also reinforces the level of satisfaction of employees, shareholders, institution customers and the financial market.

Fraud is a specific part of operational risk, being either internal or external. Prevention and detection of this risk are the motivation of this work. In the next section we explain its details and debate its importance.

2.2. FRAUD

In 1950, Donald R. Cressey, Ph.D., interviewed 250 felons over a period of five months. The felons were selected based on the following two behavioural criteria: (1) the person should have accepted a position, job or role of trust, and (2) they must have violated that trust.

Cressey concluded that the three factors that were always present when respondents reported the breach of the trust were:

1. Realized they had a financial problem that was not likely to be shared with another person in their environment or conviviality;
2. Had knowledge or awareness that this problem could be solved secretly for breach of the position of financial trust received, and
3. Were able to rationalize their own conduct, so that they allowed them to adjust their conception of themselves as trustworthy.

For these reasons, the author concluded that at the base of fraudulent behaviour lies a human process defined by a combination of the three essential factors mentioned above, namely, a non-shareable financial problem, an opportunity to commit a breach of trust and rationalization on the part of the offender, so that they would have a 'clear conscience' even after committing the fraud. (Cressey, 1950, p. 738-743)

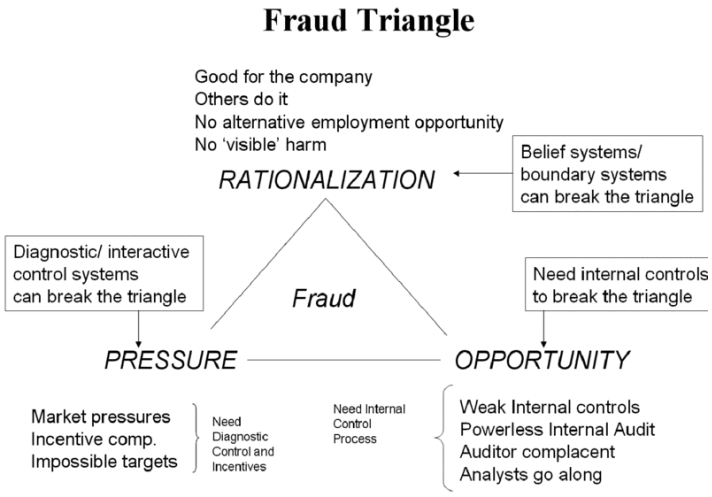


Figure 5 - Fraud Triangle Model (Source: Source: Wells, J. T., 2005. Principles of fraud examination. Hoboken, New York: John Wiley and Sons)

The pressure on the individual corresponds to the financial problem not shared by the individual, such as the fear of losing their current occupation, inability to reach or maintain a certain standard of living, and other personal problems. The fact that the individual had the opportunity, the necessary knowledge, and the information to commit the fraud are failures on part of corporate governance, and so is the lack of opportunity to resolve the unshared problem. Rationalization constitutes the process of defining the act as justifiable and/or acceptable.

The concept of fraud has developed over time and has become increasingly exhaustive. The word 'fraud' has a Latin origin and originally means 'harm done to someone'. Picket in 2000 (cited by Moura & Silva, 2004. p. 550) defined fraud as any behaviour by which a person has an intention to gain an advantage over another person. This definition has become more and more comprehensive, yet it is not yet complete (Moura & Silva, 2004) because it is found in the most diverse forms and across different domains (Pimenta & Afonso, 2012).

According to Wells, 2007 (as cited in Pimenta 2009), 'in the broadest sense, fraud can include any crime for profit, using as main *modus operandus* the achievement'. However, according to Pimenta (2009), the achievement does not directly imply fraud because, to be considered fraud, there must be damages, which are typically monetary in nature.

There are different types of Fraud:

- Corporate Service Fraud
    - Payment Fraud - This type of fraud involves falsely creating or diverting payments. Examples include creating false bank account records that allow fraudulent payments to be made. Other examples include generating fake payments, making fraudulent payments to yourself, intercepting and changing payee details, or intend not to pay.
    - False Accounting Fraud - This type of fraud involves changing the way company accounts are presented so that they do not reflect the actual value or financial

activities of the company. This fraud commonly includes overstating assets and / or understanding liabilities.
    - Acquisition Fraud - Engages in a third-party procurement process and covers the acquisition of goods, services and construction projects. Acquisition fraud usually involves collusion to perpetrate fraud that encompasses bidding irregularities, bid manipulation, or payment requests - often for goods (and sometimes services) that have not been delivered or are below what was specified as the request.
    - Asset Exploitation and Information Fraud - Examples of fraudulent activity reported in this category include sick leave staff who work elsewhere, abuse of flexible working hours, misuse of company time, and misrepresentation or misrepresentation
    - Travel and subsistence, payment and other licenses Fraud - Fraud in this area involves activities such as completing fraudulent payment requests or creating false payroll records. Examples of fraud include unclaimed travel claims, false claims of customer entertainment, exaggerated claims, forged signatures authorizing payment, etc
    - Receiving Fraud - This type of fraud involves using the organization's assets for unofficial purposes and / or providing information to outsiders for personal gain. This excludes direct theft from insiders, such as stealing stationery or other physical assets. (Bologna e Lindquist, 1995)
- Institutional Investment Fraud
    - Pyramid or Ponzi Scheme Fraud - These well-known types of fraud involve an unsustainable business model, in which investments to other later investors are used to pay off previous investors, giving the impression that early participants investments dramatically increase in value in a short space of time. of time. These types of scams often appear at the beginning of a recession when investors want to withdraw their money from the scheme, leading to their sudden collapse and exposure. (Reena Aggarwal, May Hu and Jingjing Yang, 2015)
- Business Trading Fraud
    - Long and Short Business Fraud - This type of fraud occurs when a seemingly legitimate business is established with the intention of defrauding its suppliers and customers. This can happen after the company develops a good reputation and credit history or when the apparent business has been in operation for only a few months (short term fraud, usually Internet related). (Cassim M, 1987)
- General Business Fraud
    - Bankruptcy-Related Insolvency and Fraud - Insolvency fraud occurs when a company is dealing in a fraudulent manner and usually occurs before the company's early insolvency. Directors (or shell directors) often set the phoenix immediately before or after the insolvency of the first company, in order to obtain assets from the first company and to avoid paying its debts at the same time. However, there are several provisions of the Insolvency Law that allow liquidators and / or creditors to take action against those individuals who personally try to take shelter behind the corporate veil of the company. (Richard Perkoff, 2019)

In Portugal, according to Pimenta (2009), the quantification of fraud becomes quite complex due to various institutional, cultural, and cognitive factors. Nonetheless, it is possible to obtain, with a low probability of error, a set of values to estimate that fraud represents between 1.5% and 2% of Portugal's GDP. This being the case, it is important to note the different types of fraud: internal fraud (by an employee of an organisation), external fraud (from an entity outside of an organisation), interpersonal fraud (by one or more individuals against another), and business fraud (from one organisation to another organisation or individual) (Soares, 2008). In this study, we will focus on external fraud.

2.3. PAYMENT FRAUD

The amount of electronic transactions has significantly increased over the last few years, mainly due to the spread of electronic commerce technology. This resulted in an increase of fraud cases, resulting in billions of dollars losses every year, worldwide. (Brandao G., Caldeira E., Pereira A., 2014)

Payment fraud is any type of false or illicit transaction. The perpetrator deprives the casualty of assets, personal property, interest or sensitive information through the Internet. (ACFE, 2010)

According to a study from Juniper Research[3], in 2015, the average value of a fraudulent transaction is significantly higher than the average of that same legitimate transaction. The most affected categories were airlines with 46 % of fraudulent transactions, followed by money transfer with 16%:

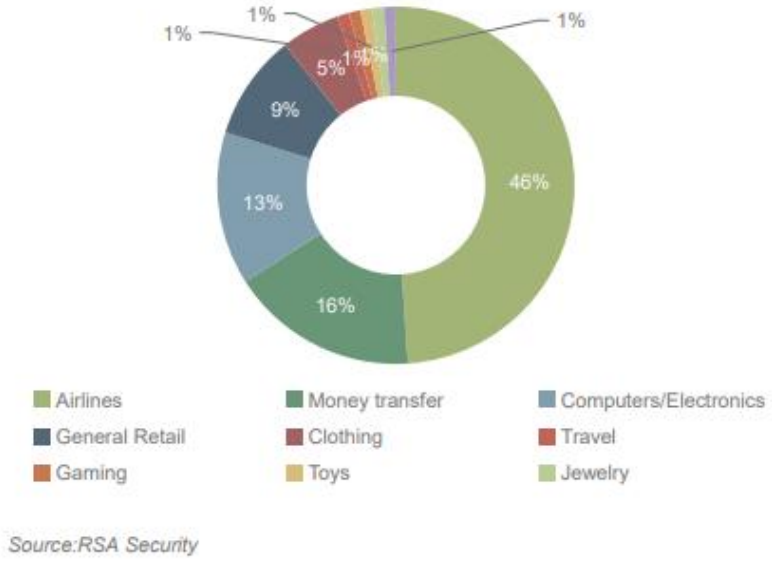


Figure 6 - Top Merchants Affected by Fraud Transactions
(Source: ONLINE PAYMENT FRAUD WHITEPAPER, 2015)

As mentioned in the previous chapter Fraudsters use an assortment of strategies leading to substantial losses. Fraudsters use exposed data in a decentralized, worldwide production process translating stolen information into false installments. (Richard J. Sullivan, 2014)

2.4. ANALYTICAL METHODS FOR FRAUD DETECTION IN PAYMENT

Fraud prevention consists of ensuring fraud does not occur prior to finalising the payment terms of a transaction. Fraud detection comes into place when prevention measures fail to prevent fraud and consists of identifying fraud as soon as possible after it occurs. However, both fraud prevention and detection activities are in constant evolution. This is due to the fact that, whenever a new fraud prevention or detection method is implemented, criminals revise their strategies and try new

[3] Juniper Research Group is a consulting and analytics firm. It is a global leader for industry research with mobile, online & digital market research specialists, providing market intelligence, consulting, data and forecasting.

methods of getting around the fraud prevention or detection measures. As a result, new methods for fraud detection are elaborated and the cycle repeats itself (Bolton, R.J., Hand, D.J., 2002).

There are several methods capable of detecting fraud: business rules, auditing, social networks, and statistical and data mining models.

- The use of business rules for fraud detection is very important, as they serve to represent the user's requirements and internal process conditions that are essential to keeping up with the growth of companies, including making software useful and up-to-date (Wan-Kadir & Loucopoulos, 2004).
- The main aim of the audit is not to detect fraud, but rather to express an opinion on the veracity of the financial statements. However, in the execution of the work, errors or frauds may be detected. This approach is very costly and time-consuming, being only applied to small samples (Copeland et al., 2012; Francisco, 2014; Schiller, 2006).
- Social Networks consist in the implementation of networks connecting suspicious or fraudulent entities by modelling connections between entities in claims (Baesens et al., 2015; Jans, Van Der Werf, Lybaert, & Vanhoof, 2011)
- A wide range of data mining techniques is used in fraud detection. So far, the logistic regression model (which measures the relation between a categorical variable – fraud or not fraud – with the other independent variables) is the most used model in detecting financial fraud (Albashrawi, 2016).

The choice of these methods must be made accordingly to the set of characteristics that each company presents with the needs and particularities of the business. Most failures in fighting fraud are due to errors in the detection phase, which increases the probability of propagation. Fraud detection methods allow diagnosing fraudulent activities whereas prevention aims to avoid fraud. (Baesens et al., 2015; Hartley, 2016)

The present study will be focus on the use of data mining for fraud detection.

The use of data mining for fraud detection requires the existence of a well-defined problem, based on procedural models, and cannot be solved with query and reporting tools (Lavrac et al., 2004). This technique aims to find unknown facts with a statistical basis that are triggered from the data (Elkan, 2001) in an efficient way.

Today, the data discovery and decision-making phenomenon is more than a trend. Data mining refers to extracting knowledge from a large set of observed data in order to discover the unsuspected relationships and hidden patterns in data and presenting it to users in an innovative, understandable, and useful way (Adeniyi, Wei, & Yongquan, 2016).

Data mining is also considered the "process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules" (Berry & Linoff, 1997, p. 5). It uses machine learning, artificial intelligence and statistical technologies to extract information from a data set and transforming it into an understandable structure that could never be found through manual analysis alone.

Berry and Linoff (2010) later had cause to regret the 1997 reference to "automatic and semi-automatic means", feeling that it misled the role of data exploration and analysis.

In its first years, data mining was mainly used for unsupervised learning, involving identifying clusters and associations, without making any assumptions about the structure of the data (Stephens and Tamayo, 2003). Nowadays, the main goal has changed dramatically. A researcher will now examine large data sets to discover patterns as before, but will also uncover new information and predict failure points and outcomes for the future.

A few concise definitions of Data Mining are presented below.

- "DM is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" – Fayyad et. all (1996)
- "[Data Mining is] the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different (…) resources, to reveal otherwise hidden meanings" - Hearst (1999)
- "[Data Mining is] "Extracting useful information from large data sets." - Hand et al. (2001)
- "Data Mining is the process of automatically discovering useful information in large data repositories" – Tan, Steinbach & Kumar (2006)
- "[Data Mining is] the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques." - Gartner Group, information technology research firm.

2.4.1. Data Mining Processes

In order to have a clear analysis of the applied techniques it is essential to have an overview of the Data Mining Process.

There are several methodologies proposed by authors, however the most used methodologies are SEMMA (Sample, Explore, Modify, Model, Assess) and CRISP-DM (CROSS Industry Standard Process for Data Mining). (Santos & Azevedo, 2005) They will be described in the following subsections.

CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a process for performing data mining, with the goal of having a consistent procedure, repeatability and objectivity. The data mining process provides a view of the project lifecycle, containing the project phases, their tasks and the relationships between these tasks. (R. Wirth and J. Hipp, 2000)

The CRISP-DM methodology is described in terms of hierarchical process and consists of six crucial steps, which are described and illustrated below (Figure X). The six phases were thought so that they could be applied in any business area. [CRISP-DM, 1999]

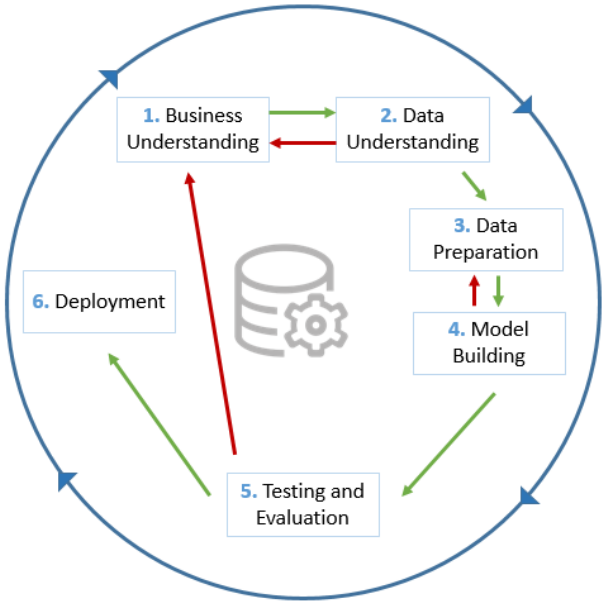


Figure 7 - CRISP-DM Life Cycle
(Source: Adapted from R. Wirth and J. Hipp,
"CRISP-DM: Towards a standard process model for data mining")

1. Business Understanding – There should be an analysis of the context for the demands of the project. More specifically, what are the scenarios (or not) for the project to start in the best possible way?
2. Data Understanding – The characteristics and limitations of the databases, their history, their composition and their type should be understood, and it should be determined if the data are sufficient to understand the proposed problem.
3. Data Preparation – There are numerous techniques and technologies that can be used when working with data. Some of the many tasks include inserting missing treatment, converting different types of data according to need and understanding whether the data are categorical and continuous and if they should be standardized or not. In this step, the construction of variables is also fundamental to the success of any model.
4. Model Building – At this point, techniques will be used that are more adherent to the objective of the project, be it a prediction, classification, grouping or regression.
5. Testing and Evaluation – In this step, a rigorous assessment of the results is performed so that there is confidence in the project before it is delivered. If the project objective has not been reached, it may be necessary to go back to the first step.
6. Deployment – In this step, the project is finished. It is the least technical, but not the least important, stage of the data mining process. Here the result is delivered to the customer in the form of a report or a system deployment for real-time data access.

SEMMA

The SEMMA is a methodology developed by SAS Institute, linked to SAS enterprise miner. It is a cyclic scheme, as can be seen in the following representation.

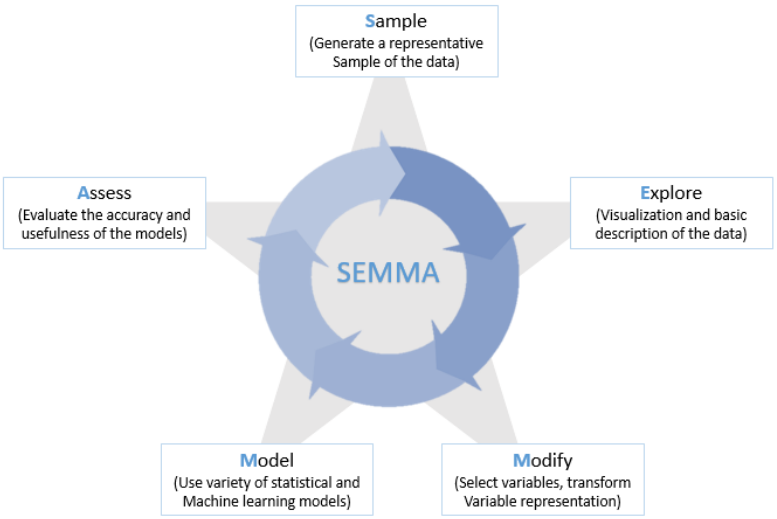


Figure 8 – SEMMA
(Source: Adapted from A. I. R. L. Azevedo and M. F. Santos,
"KDD, SEMMA and CRISP-DM: a parallel overview")

The acronym SEMMA is described as:

- Sample - Sample of representative population data, is normally partitioned into training, validation and test sets;
- Explore - helps redefine the entire process of knowledge discovery by search for trends and anomalies in the data through statistical techniques is the processing of data;
- Modify - allows you to select and transform the variables in view of the type of model used, is based on the exploratory phase to manipulate the information. (to convert nominal variables in numerical);
- Model - through data mining models, search combinations in information that better predicts the expected result with the model;
- Access (evaluation) - evaluates the results obtained by measuring the performance of the data mining process allowing the optimization of the results by the adjustment of the model.

The SEMMA model allows the adjustment of the data according to several regression or classification simultaneously, and the model with the best performance in terms of the mean squared error will be selected to predict the data.

CRISP-DM vs SEMMA

As seen, both methodologies give an overview of the development of a predictive model, structuring the data process in phases that are interrelated, converting the knowledge discovery development in an interactive process.

In a first analysis it can be deduced that in terms of processes for developing a Data Mining project the methodology CRISP-DM is more complete than SEMMA, by incorporating the Understanding phases of the Business and Implementation. However, a more detailed analysis integrates Business Understanding into the Sample phase of the SEMMA methodology that it is not possible to constitute a coherent and solid sample without genuine of all aspects presented. With regard to the evaluation phase of the SEMMA methodology if it is considered that the obtained knowledge is applied it is assumed that the phase of Implementation (present in the CRISP-DM methodology) is also present (Azevedo& Santos, 2008).

Tasks	CRISP - DM	SEMMA
Project Initiation	Business Understanding	-
Data Access	Data Understanding	Sample and Explore
Data Transformation	Data Preparation	Modify
Model Building	Modelling	Model
Project evaluation	Evaluation	Assessment
Project Finalization	Deployment	-

Table 3 - CRISP-DM & SEMMA
(Source: Adapted from SAS)

Based on these differences, it was decided to follow the CRISP-DM methodology with the SEMMA approach during this work, in particular by adapting itself to the project in question and also for being one of the methodologies most used in this type of research work (Predictive Models). One of the main reasons was the fact that SEMMA methodology is focused on SAS Enterprise Miner software and on model development.

2.4.2. Predictive Models

"An economist is an expert who will know tomorrow why the things he predicted yesterday didn't happen."
—Earl Wilson

Predictive analytics consists of the techniques, tools, and technologies used to find models that can anticipate outcomes with a significant probability of accuracy. Prediction brings insight into the unknown. Accurate predictions can transform businesses by empowering their decision-support systems.

Prediction is a process that consists of three distinct but related parts: capture, predict, and act. 'Capture' refers to capturing, or collecting, relevant data, while 'predict' refers to using various techniques, including data mining, text mining, and statistical analysis, to anticipate an outcome. Finally, 'act' refers to deploying models into operational processes to optimise decisions at the right time to change outcomes for the better.

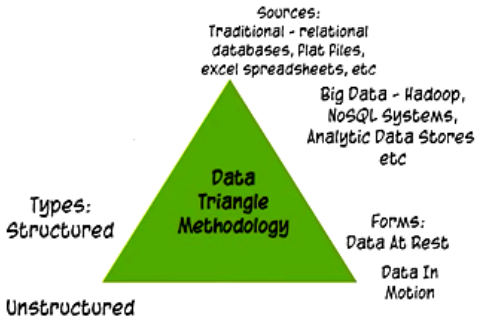


Figure 9 - Data Triangle methodology
(Source: Author elaboration, using information from Reeve A. (2013). "Managing Data in motion: Data Integration. Best Practice Techniques and Technologies". 152-154)

Predictive Analytics is often express in many ways. A few concise definitions of predictive modelling are presented below.

> "Predictive analytics is the process of discovering interesting and meaningful patterns in data."
> (Dean Abbott, 2014, Applied Predictive Analytics. Principles and techniques for the professional data analyst)

> "Predictive modelling (also known as supervised prediction or supervised learning) starts with a training data set. The observations in a training data set are known as training cases (also called training examples, instances, or records). The variables are called inputs (also known as predictors, features, explanatory variables, or independent variables) and targets (also known as response, outcome, or dependent variable). For a given case, the inputs reflect your state of knowledge before measuring the target"
> (Christie et al., 2011, Applied Analytics Using SAS Enterprise Miner)

> "Predictive Analytics helps connect data to effective action by drawing reliable conclusions about current conditions and future events"
> (Gareth Herschel, Research director of Gartner Group)

There are several aspects of the model building process that are worthy of further discussion. In the following subsections, we will describe the relevance of defining the target in a model and the three most common predictive modelling approaches.

2.4.2.1. Defining the Target

Defining and measuring the target variable to be predicted by the model is the first step in a data mining project. The target variable should have all the information that we want to predict – the outcome.

In the present work the target variable will be the "Debt" - case where the client has been in debt or is in debt. The approaches taken were:

- Binary Target- It takes the value 1 if the transaction was fully paid, and 0 otherwise. To notice that the binary variable can either be numeric, 0 and 1, or character, Y and N.

- Multi-Class Target - it combines the dependent variables in study into a discrete variable composed by 4 distinct classes

These two methods will be further described later, in the methodology chapter.

2.4.2.2. Logistic Regression

Logistic Regression is a type of regression normally used when the dependent variable is dichotomous or binary. It allows the estimation of the associated probability of occurrence of a given event in a set of exploratory variables. After estimating the probability of an instance, the classification of it as event or non-event can be made.

Logistic regression ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular machine learning method. The dependent variable follows the Bernoulli[4] distribution having an unknown probability.

In any regression model the right quantity is the mean value of the response variable given the independent variable value. This quantity is called conditional mean value and can be expressed as E[Y|X] where Y represents the response variable and X is the explanatory variable:

$$E[X|Y = x] = \pi(x) = \frac{e^{\beta 0+\beta 1x}}{1 + e^{\beta 0+\beta 1x}}$$

Since the association between the target variable and the inputs is not a linear function, a linkage function denominated *logit* is used to establish the relationship between the dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. This transformation is defined as:

$$logit\big(\pi(x)\big) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$logit\big(\pi(x)\big) = \frac{\frac{e^{\beta 0+\beta 1x}}{1 + e^{\beta 0+\beta 1x}}}{1 - \frac{e^{\beta 0+\beta 1x}}{1 + e^{\beta 0+\beta 1x}}} = \frac{\frac{e^{\beta 0+\beta 1x}}{1 + e^{\beta 0+\beta 1x}}}{\frac{1}{1 + e^{\beta 0+\beta 1x}}}$$

$$logit\big(\pi(x)\big) = \ln\big(e^{\beta 0+\beta 1x}\big) = \beta 0 + \beta 1x$$

This transformation takes on special importance because the model with this transformation has several properties of the linear regression model:

- The logit function is linear in the parameters;
- It can be continuous;

Their values can assume $\mathbb{R}$

The final value is interpreted as the estimated posterior probability - In order to associate predicted values to these probabilities the sigmoid function is used. The function maps any real value into another value between 0 and 1.

[4] The Bernoulli distribution is a special case of the Binomial distribution where n=1 – just one trial.

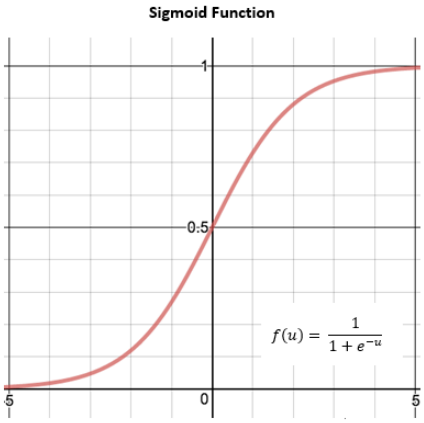


Figure 10 - Sigmoid Function

The goal of logistic regression is to correctly predict the category of the outcome for individual cases - any value of x above 0.5 will be classified as Class 1, otherwise will be classified as Class 0. The regression coefficients for the logistic regression are calculated using maximum likelihood estimation or MLE. This means that the choice of an adequate model is based on the significance of the coefficients associated with the input variable.

2.4.2.3. Decision Trees

A decision tree is a map of possible outcomes of a variety of related choices. It allows an individual or organization to compare possible actions based on their costs, probabilities, and benefits.

Nowadays, Decision Trees is one of the most popular predictive algorithms due to their structure and interpretability. They are a simple, but powerful form of multiple variable analysis. They provide unique capabilities to complement and substitute for:

- Traditional statistical forms of analysis
- Data mining tools and techniques
- Multidimensional forms of reporting and analysis found in the field of business intelligence

2.3.2.3.1. Decision Trees Representation

A decision tree usually begins with a single node, which then, splits into possible outcomes. Each of these results leads to additional nodes, which will again, split into other possibilities.

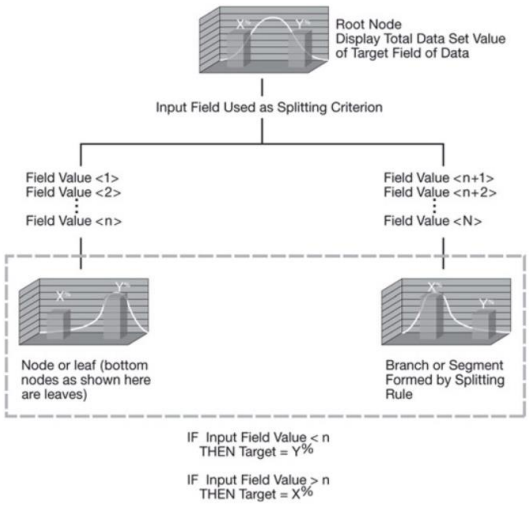


Figure 11 - Illustration of the Decision Tree (Source: SAS Institute)

There are three types of nodes: probability nodes, decision nodes, and end nodes (also called leaves). The probability node, represented by a circle, shows the probabilities of a certain result. A decision node, represented by a square, shows the decision to be made, and an end node shows the final result of a decision path. All nodes, including the bottom leaf nodes, have mutually exclusive assignment rules. In consequence, observations from the parent data set can be found in one node only.

2.3.2.3.2. Growing a Decision Tree

Decision tree growing is done by creating a decision tree from a data set. Splits are selected, and class labels are assigned to leaves when no further splits are required or possible. To measure the goodness of a split different functions can be used, the most known are *Entropy* and *Chi-Square,* both approaches are available in SAS Enterprise Miner.

Entropy

Entropy characterizes the (im)purity of the data: in a dataset, it is a measure of the lack of homogeneity of the input data when comparing with is actual classification. For example, the entropy is maximal (equal to 1) when the data set is heterogeneous (Mitchell, 1997).

The entropy function *E* of a collection *S* in a *c* class classification is defined as:

$$\text{Entropy(S)} = \sum_{i=1}^{c} -p_i \ \log_2 p_i$$

Where p_i is the proportion of *S* belonging to class i.

Some of the most useful features of Entropy are:

- Maximum Entropy $\log_2 c$ if $\rho_i = \rho_j \ \forall$ i ≠ j
- Entropy(S) = 0 if $\exists\ i$ such as $\rho_i = 1$
- By hypothesis, 0. $\log_2 0 = 0$

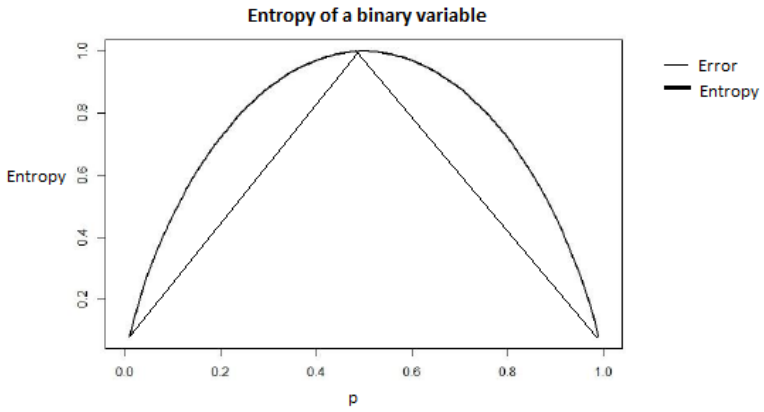


Figure 12 - Entropy of a Binary Variable

Figure 8 shows the variation of the entropy for a binary target variable versus the error. The maximum is reached when there is a 50/50 proportion of event and non-event. The aim of the algorithm is to find the split that minimizes the entropy, which provides the largest difference in proportion between the target levels.

The information gain of an attribute A from a data sub-set S gives the expected entropy reduction caused by the partitioning of the examples according to an input variable. Let P(A) be the set of values that A can assume; Let x be an element of this set and let s_x be the subset of S formed by the data in which A = x; the entropy obtained by partitioning S as a function of attribute A is given by:

$$Entropy(A) = \sum_{x \epsilon P(A)} \frac{|s_x|}{|S|} \, Entropy\ (s_x)$$

The Information *Gain* relative to a collection *S* and an input *A* is defined as:

$$Gain(S, A) = Entropy(S) - Entropy(A)$$

The construction of a decision tree has three objectives: to decrease the entropy (the randomness of the objective variable), to be consistent with the data set and to have the smallest number of nodes.

Gini Index

The Gini index, developed by Conrado Gini in 1912, measures the degree of heterogeneity of the data. Therefore, it can be used to measure the impurity of a node. This index in a given node is given by:

$$Gini\ Index = 1 - \sum_{i=1}^{c} {P_i}^2$$

Where P_i is the relative frequency of each class in each node, and c is the number of classes.

When this index is equal to zero, the node is pure. On the other hand, when it approaches the value one, the node is impure (increases the number of classes evenly distributed on this node).

In situations where Gini criterion is used in binary partitioning trees, it tends to isolate the records that represent the most frequent class in a branch. When using entropy, the number of records in each branch is balanced.

Both Impurity measures have different techniques to split the node in decision tree based models. However, most of the times, they are quite consistent with each other, the performance of a model won't change with the use of Gini Index or Entropy.

Algorithms

There are several known algorithms that implement a decision tree. There is no precise way to determine the best algorithm. Their performance may vary depending on the volume of data and the situation in which they are being used.

One of the first algorithms developed was the ID3. His concept of creation used the idea of inference systems and machine learning concepts. In a short time, other algorithms have also appeared: C4.5, CART (Classification and Regression Trees), CHAID (Chi Square Automatic Interaction Detection) and others. In this present subsection the CHAID will be described.

CHAID (Chi-Square Automatic Interaction Detection)

In CHAID algorithm, input groupings are formed by combining values in the input if their relationships with the target are similar. Values are indistinguishable from a statistical point of view if the pairwise differences between two values relative to the target are not statistically significant. By selecting the input variable with the lowest significant p-value, the algorithm is intrinsically selecting the variable that has the stronger relationship with the target variable at each step (Ritschard, 2010).

Like other decision trees algorithm, CHAID advantages are that is a highly visual algorithm, easy to interpret because it uses multi-way slip[5] by default. It need rather large samples sizes to work effectively since with small samples sizes the respondent groups can quickly become too small for reliable analysis. One important advantage of CHAID over alternatives is that the multiple regression is non- parametric.

The CHAID algorithm can produce more than two branches at any level of the tree. The first branch of the tree is created when the independent variable, which has greater interaction with the dependent variable, is selected. Each node has homogeneous values according to selected variables. The process is performed on all independent variables to find the best number of classes.

[5] Multi-way split: use as many partitions as distinct value

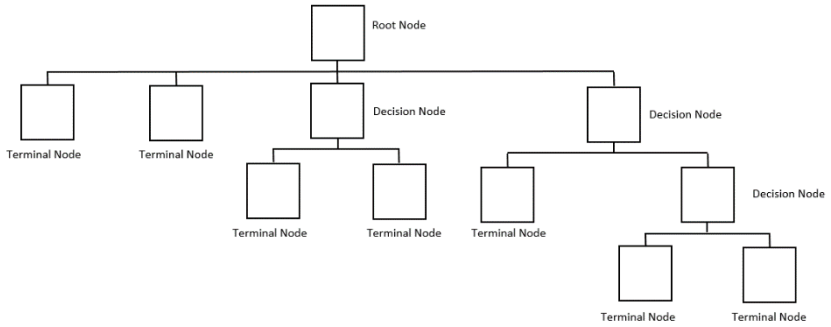


Figure 13 - *CHAID* structure - Decision Tree example (Source: Author elaboration)

CART (Classification and Regression Trees)

The trees constructed by the CART algorithm are indicated for non-linear problems, achieving satisfactory results for both numerical and categorical variables. The growth of the tree is binary - each node has two branches/sub-trees, so that the values of the dependent variable are more homogeneous than the previous division. Within a tree, there are many simpler sub-trees, so the tree obtained has the possibility of being pruned once the process is finished, as shown in the figure 10.

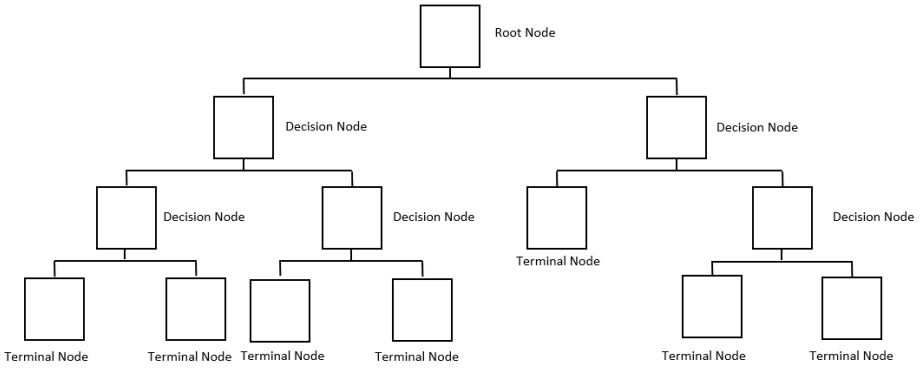


Figure 14 - *CART* structure - Decision Tree example (Source: Author elaboration)

Through the process, instead of determining when a node is terminal or not, it continues to provide tree growth until it is no longer possible to do so, for example, when reaching a minimum number of data in the sample. After all the terminal nodes have been found, the tree is defined as maximal, that is, the tree of maximum size

After finding the maximal tree, some subtrees, obtained by pruning some branches of this tree, are observed by testing the error rates and the best of them is chosen.

The following table shows a brief comparison between the two algorithms – CHAID and CART:

	CHAID	CART
Advantages	Uses dependent variable of categorical type; Doesn't need to follow pre-set parameters; No pruning treatment.	Uses dependent variable of any type; Doesn't need to follow pre-set parameters; Generates binary trees.
Disadvantages	Requires large amounts of data for satisfactory results.	The generated tree has many levels.
Measure used to select input variable	Chi-square	Gini index for nominal targets and variance reduction for interval targets.

Table 4 - Comparison of CHAID vs CART algorithm
(Source: Author elaboration, , using information from Lu, Y. (2015). "Decision tree methods: applications for classification and prediction", Shanghai Archives of Psychiatry, 130-133)

2.4.2.4. Artificial Neural Networks

The human brain is considered the most fascinating existing carbon-based processor, consisting of approximately 10 billion neurons. All the functions and movements of the organism are related to the functioning of these small cells. The neurons are connected to each other through synapses, and together they form a large network, called Neural Network.

Artificial neural networks consist of a method of solving artificial intelligence problems, constructing a system that has circuits that simulate the human brain, including its behavior - learning, making mistakes and making discoveries. More than that, they are computational techniques that present a model inspired by the neural structure of intelligent organisms and that acquire knowledge through experience.

There are many different types of artificial neural networks, differing in their learning rules and topologies. In this case, the one used is the Multi-Layer Perceptron (MLP), one of the most prestigious and successful architectures used in predictive and classification problems.

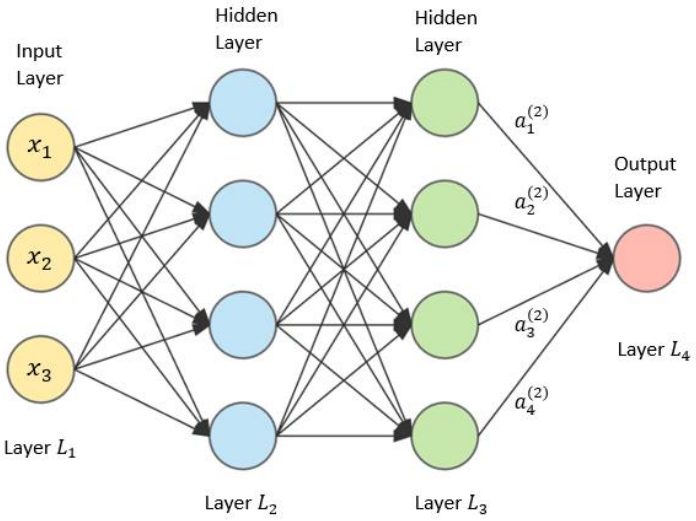


Figure 15 - Artificial Neural Network Representation

The type of neural networks shown has three layers. The first layer, called Input Layer, is characterized by receiving the data which the model will be trained and stores it in the network; the second layer, called Hidden Layers, is where most processing is done through weighted connections; and the third layer, called Output Layer, is where the final result is completed and displayed. The layers are linked through synapses with the exception of the first layer that is where the whole process begins.

MLP uses a supervised learning technique called backpropagation algorithm for training, characterized by decreasing the gradient throughout the network, responsible for minimizing the mean square error of the model output.

The process begins with the initial random distribution of the associated weights. Each variable of the initial layer is trained and synthesized by the hidden layers and the values obtained in the output layer are later compared with the real ones, ending the process with the calculation of the error value that the network obtained. This error is then gradually reduced, altering and adjusting the weight of synapses.

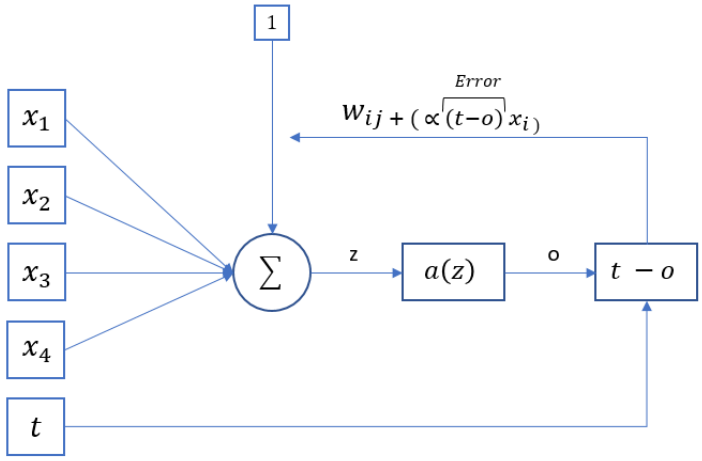


Figure 16 - Backpropagation Algorithm

The function a(z) is called activation function. This function is the tool that enables the information pass through the network. It allows small changes in the weights and bias to only cause a small change in the output. This part of the process is crucial for the algorithm since is where the network learns – it decides if the information that the neuron is receiving is relevant to the information provided or should be ignored.

There are several types of activation functions and this is an active research area, as Artificial Intelligence grows.

Binary Step Function

One of the most commonly used activation functions is the ”threshold” function.

$$f(s) = \begin{cases} 1, & s > 0 \\ -1, & otherwise \end{cases}$$

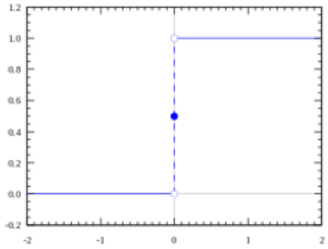


Figure 17 - *Step* Activation Funtion

This activation function is particularly useful when we want to solve classification problems, it will conclude whether or not the neuron should be activated.

Sigmoid Function

A common activation function is the sigmoid activation function because of its properties such as non-linear, monotonically increasing and easily differentiable. The function ranges from 0 to 1 having a shape format S.

$$f(s) = \frac{1}{1 + e^{-s}}$$

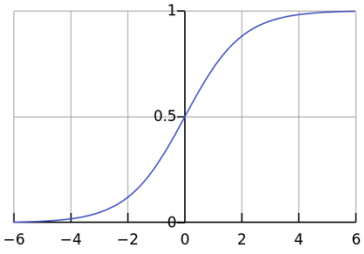


Figure 18 - *Sigmoid* Activation Funtion

One of the problems associated to this function is when the gradients become very small. This means that the gradient is approaching zero and the network is not really learning.

Hyperbolic tangent – Tanh

The Tanh activation function works similarly to the sigmoid function, but symmetrical to the origin – bounded from -1 to 1.

$$f(s) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

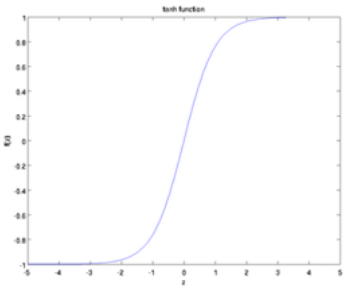


Figure 19 – *Tanh* Activation Function

Both Sigmoid and Tanh functions are sometimes avoided due to the Vanishing Gradient problem - when more and more Hidden layers are added to the model, making the learning speed of the next hidden layers in the model getting faster and faster. The Training process will then take longer and the Prediction Accuracy of the Model will decrease.

ReLU

The ReLU function is the rectified linear unit:

$$f(s) = \max(0, s)$$

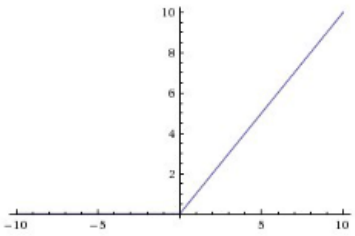


Figure 20 - *ReLU* Activation Function

The main advantage of using the ReLU function over other activation functions is that it does not activate all neurons at the same time - if the input is negative, it will be converted to zero and the neuron will not be activated. This means that at the same time, only a few neurons are activated, making the network sparse, efficient and easy for computing.

2.4.2.5. Ensemble Models

The Ensemble models consist of the combination of previous models to obtain more accurate predictions than can be gained from any of the individual models. By combining predictions from several models, limitations in individual models may be avoided, resulting in a higher overall accuracy (Low Error), higher consistency (avoids overfitting) and reduces bias and variance errors.

The principle is simple and depends mainly on two factors: the precision and the diversity of the models that make up the ensemble. The accuracy of the models should be quite similar, but the way they are predicted must be different. It is fundamental to notice that the ensemble model can be more accurate than the individual models only if the individual models disagree with one another.

There are two main approaches for model ensembling:

- Use similar classifiers and combine them together (for example Bagging and Boosting)
- Combine different classifiers (using model stacking)
-

Bootstrap Aggregation (Bagging)

Multiple models of the same learning algorithm trained with subsets of dataset randomly picked from the training dataset.

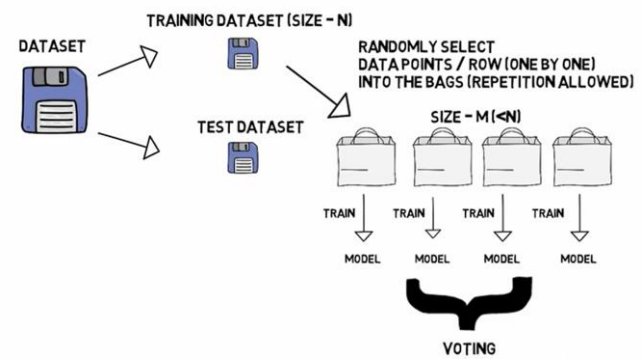


Figure 21 - Bagging Method
(Source: Author elaboration, using information from Bulhmann, P. (2012). “Bagging, Boosting and Ensemble Methods”)

Boosting

Boosting is a little variation of Bagging, where we give more emphasis on selecting points which give wrong predictions in order to improve accuracy.

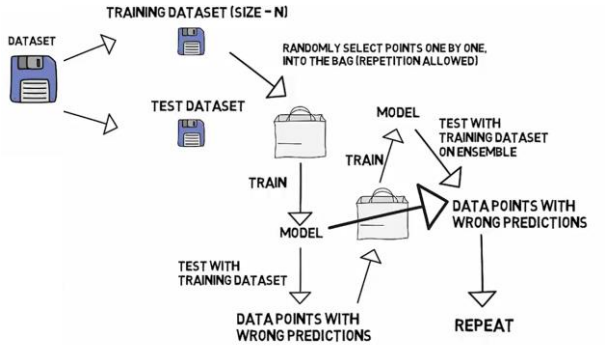


Figure 22 – Boosting Method
(Source: Author elaboration, using information from Bulhmann, P. (2012). "Bagging, Boosting and Ensemble Methods")

Stacking

Model stacking is an efficient ensemble method in which the predictions generated by using various machine learning algorithms are used as inputs in a second learning level algorithm – combination of classification or regression models via meta-classifier or meta-regressor. This second-level algorithm is trained to optimally combine the model predictions to form a new set of predictions.

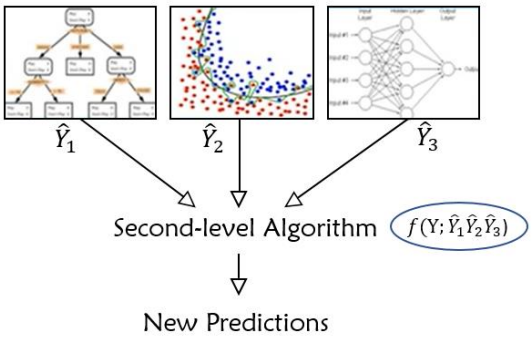


Figure 23 - Stacking Method
(Source: Author elaboration, using information from Bulhmann, P. (2012). "Bagging, Boosting and Ensemble Methods")

3. METHODOLOGY

3.1. RESEARCH METHODOLOGY

Having the right methodology is the key to acquire and organize the dataset and to be able to achieve results. The approach will be in the following order: (A) the preparation of the dataset for a statistical study, (B) detail of the value creation metric, timeframe and explanatory variables and ultimately, (C) create the processes and methods for the Costumer Segmentation and Predictive Analysis. In general terms, the workflow can be described in the following diagram:

Figure 24 - Workflow Diagram

The scientific method emphasizes the choice of systematic procedures, so that a certain situation under study is described and explained. Your choice should be based on two basic criteria: the nature of the objective in which it is applied and whether the objective is in view of the study (Fachin, 2001).

3.2. DATA COLLECTION PROCESS

The data collection process can be difficult and time-consuming due to security, privacy and cost issues. In average, the time allocated to the data preparation process is around 80%. In some cases, this value can reach 85% which is common in case where the data is stored in many warehouses. (Guo & Li, 2008)

In this case, the data collection came from two different sources - Company Database, Via Verde Portugal, and Centraldedados [6], which were later processed and complemented with each other in the Microsoft Office Excel 2016 tool.

In SAS Enterprise Miner, the imported data was graphically and structurally analysed, consequently reducing the dimensionality of the variables integrating the model, which was complemented with the treatment of outliers and the use of different partitions. Three predictive tools were used - neuronal networks, decision tree and regression - that originated several models with different error rates obtained through the evaluation of the model.

The data used in this dissertation was mainly provided by the company in study – Via Verde Portugal - and is a sample of all 406 Million transactions for the year of 2017[7]. The data provided contains data from toll transactions, McDonalds transactions, PharmaDrive transactions and Interior and exterior Parking. The data provided allowed not only the necessary analyses (relationships addresses, costumer types and usage frequency) but also to add variables to these analyses to present a more complete and comprehensive study.

[6] Public database with a complete list of postal codes in Portugal, including the islands.
[7] RELATÓRIO INTEGRADO 2017 - BRISA

The database used in this model includes more than 1.000.000 transactions during the year of 2017. After studying the data provided and the business model it was decided to analyse the data on a daily basis. This conclusion comes from the fact that majority of customers uses the same routes/highways during the week, mainly to go to work, and usually uses a different route on the weekend.

It is also important to note that, as secondary data will be used, the quality of the database will be carefully analysed as it may compromise the reliability and robustness of the system.

3.3. MODELLING

The Analytical Model used in this project is SEMMA (Sample, Explore, Modify, Model, and Assess), a Data Mining process developed by SAS Institute, responsible for creating the software used - SAS Enterprise Miner 14.1.

SEMMA is characterized by five fundamental steps and is responsible for better performance, simplifying and streamlining all the steps required in the predictive model.

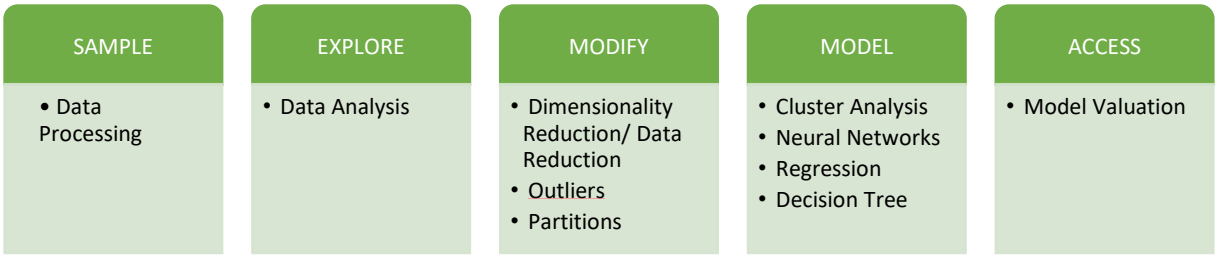


Figure 25 - Analytical Model used in the study

3.3.1. Sample

The Sample stage is characterized by selecting the data set for modelling. Includes the entire process performed to form the database used in the study – the variables used, the normalization process to which they were exposed and, finally, the partitions that constitutes the model.

3.3.1.1. Variables

As shown in the table below, the model consists of 33 nominal variables, 2 Ordinal variables, 7 Time ID, 1 binary variable and 2 interval variables. Some of the most relevant variables are:

- Contract Modality – can be one of four options: Purchase, Rental, Prepaid or Old;
- Postal Code
- Payment Method – Credit, SIBS, Direct Debit, Prepaid, Rent a Car, Interoperability
- Rescission date
- Market Type – can be either Tools, Parking or McDrive
- Payment Status – Pending, Confirmed, Error, Waiting for confirmation, sent to recovery, no payment

An extend description of the variables can be found in Annex 1.

The dependent variable on this study, "Debt", will be treated in a different way, with the objective to get a more precise result: an ordinal and binary type.

3.3.1.2. Sampling Techniques

The problem of unbalanced classes is a relatively recent one, which arose when machine learning evolved from its embryonic state, in which it was purely scientific, to being 78 an applied technology, widely used in business, industry and scientific research. Although this problem has been identified for some time, it has only been the subject of research for about 10 years. Its importance grew when it became apparent that unbalanced databases harmed the accuracy of models generated by classifying algorithms [Chawla, Japkowicz & Kotcz, 2004].

In the classification domain, a dataset is said to be unbalanced when there are many more cases of some classes than others [Chawla, Japkowicz & Kotcz, 2004]. Classes with little exposure are called rare classes [Weiss, 2004]. Another type of imbalance that has aroused the interest of the community is an imbalance within a class (intraclass), in which the distribution of data is very large, leading to the occurrence of rare cases [Han, Wang & Mao, 2005].

Chawla, Japkowicz & Kotcz (2004) divide the proposed solutions for class imbalances into three subareas:

1. Methods of data sampling, aimed at reducing the imbalance in databases;

2. Data cleansing methods, which serve to better define the decision space learned by the classifiers; and

3. Other methods that are important for solving the problem of unbalanced classes, but that are only presented for information and will not be detailed here because they are not the focus of this research.

In this study, the dataset in question is unbalanced. Of the total observations (1.048.150), approximately 98.78% are transactions paid at first time (=0) and approximately 1.22% were reported as late (=1) as shown in Figure 28.

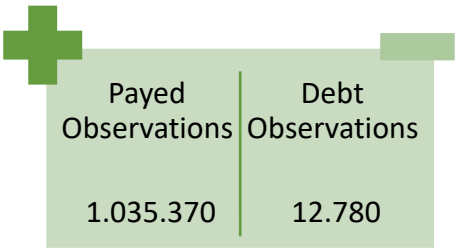


Figure 26 - Distribution of the dataset according to the dependent variable (Debt)

It can be said that for each observation with a historical debt transaction, there are 80 with no issue, making the existence of a debt transaction rare, i.e., a rare class or, more generally, an imbalanced class (Weiss, 2004). This is what happens most of the time in datasets with real data, where the

"normal" class is predominant and only a small percentage involves observations of the class of interest (Chawla, 2009; Chawla et al., 2002).

Sampling methods aim to change the training data distribution in order to increase the accuracy of the models. This is achieved by eliminating observations on the majority class (referred to undersampling in the literature) or by replication of the minority class (called oversampling). According to Jo and Japkowicz (2004), there is no guarantee that the original distribution of training data is the most appropriate for the construction of classifiers. The sampling methods are used to change the distribution of the data in such a way that it is possible to generate better classifiers for them.

The simple oversampling techniques are widely criticized by the scientific community, since many of them only replicate existing positive cases. Merely replicating existing cases of the minority class actually increases the classifier's visibility for this class. However, there is the undesired effect of overfitted models, that is, very specific models for these replicated cases, thus impairing their generalization power to the class of interest.

An oversampling technique, SMOTE (Synthetic Minority Over-sampling Technique) was introduced by Chawla, Bowyer, Hall, & Kegelmeyer (2002) as a solution to this situation.

Figure 27 – SMOTE
(Source: Made by Author)

Chawla et al. (2002) developed a different method of doing oversampling of the minority class, which consists in the generation of synthetic cases (artificial cases) for the class of interest from the existing cases. These new cases will be generated in the vicinity of each case of the minority class, in order to grow the decision region and thus, increase the generalization power of the classifiers generated for this data.

Visually, in the sample space of the data set, these new synthetic cases will be randomly interpolated along the straight-line segment linking each case of the minority class to one of its nearest k randomly chosen neighbors.

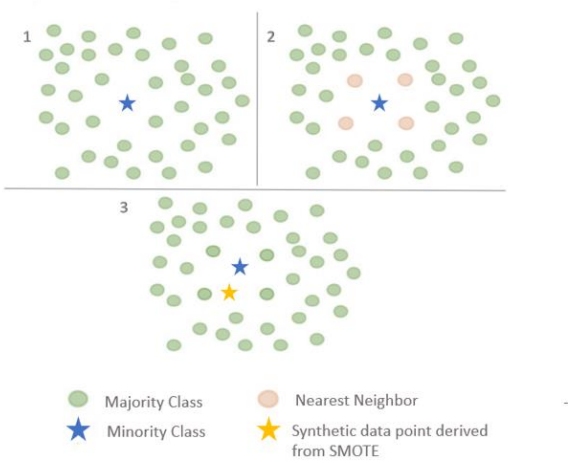


Figure 28 - SMOTE steps
(Source: Made by Author)

In SAS Enterprise Miner, one way to bias the classification of a rare event is to over-sample the rare event. To solve the unbalanced data in the study the "Level Based" option in the Sampling Node was used (Stratified property section).

3.3.1.3. Normalization

Normalization aims to improve the accuracy and efficacy of the records in order to achieve more real and correct results. Expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect or "weight." Normalizing the data attempts to give all attributes an equal weight.

There are several methods used to carry out a good standardization, the most known are the *min-max normalization, Decimal Scaling* and z-score normalization:

$$zscore = \frac{x - \mu}{\delta} \qquad Decimal\ Scoring = \frac{x}{10^i} \qquad Min\ Max = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where

x – Value
μ - Average
δ – Standard Deviation
J is the smallest integer such that Max(|DecimalScoring|) <1

Data normalization means transforming all variables into a specific range. Since in this case they are in the same measure there is no need to standardize the interval variables – price and discount. Therefore, no normalization was applied in the pre-processing phase.

3.3.2. Explore

Exploration helps refine the discovery process. The second stage of the process consists in the exploration of the data, by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

As mentioned in an earlier stage, in order to maximize and improve future forecasting, the dependent variable, debt, will be presented in two different ways.

The following Figure (Figure 30) represents the multi-class dependent variable after its categorization within the sample used in this model. As can be seen, the largest debt class is 3 (Transactions with due payment) representing about 83.4% of the total debt sample, while the less popular is class 1 (Prevention Process associated), with around 1.21%.

Figure 29 - Multi-Class Dependent Variable
(Source: Made by the author, PowerBI)

In its binary nature, the dependent variable will be divided into Payed Transaction and Debt Transaction. This classification is attributed through a mathematical calculation that relates the various types of payment fraud mentioned before. These two types of Target Variable will be described in the following two subchapters.

Figure 30 - Binary Dependent Variable
(Source: Made by the author, PowerBI)

It is possible to verify by observing the previous graph, that the selected Via Verde Transactions sample is represented by a significant frequency of paid transactions, which, following the criteria described above, has about 64% of the total transactions included in this study. As mentioned before, the data is not well balanced in his original form, and for that reason, the transactions in debt only account for, approximately, 1.2% of the total transactions in the database.

It can also be seen that the dependent variable depends on the transaction payment method chosen by the client. It can be seen through Figure 32 that the most popular/relevant payment method within the total sample is 20 - SIBS. It is important to mention that, even though the most popular

payment method is SIBS, the class with a higher percentage of debtors are 90 and 200 – Transactions to be paid in person, via mail or shop.

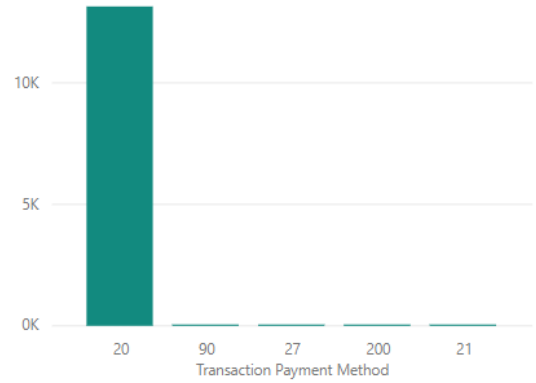


Figure 31 – Influence of Transaction Payment Method Variable
(Source: Made by the author, PowerBI)

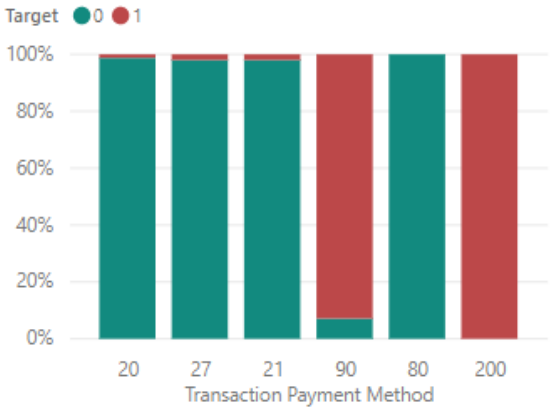
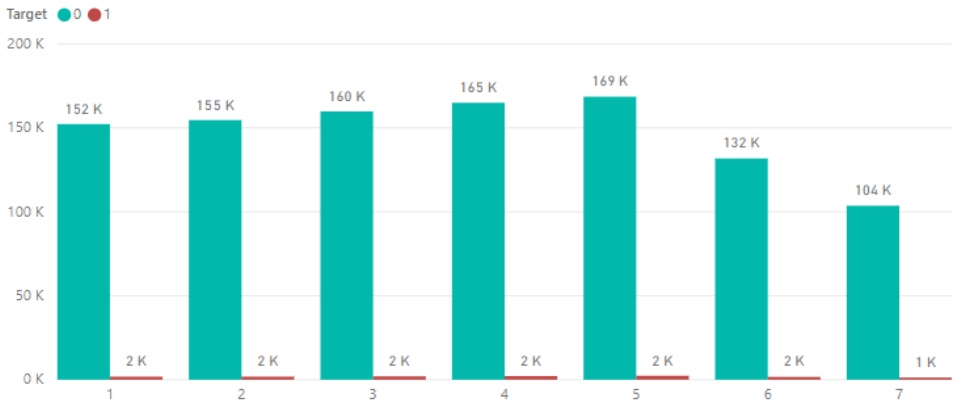


Figure 32 - Transaction Payment Method by Target Variable
(Source: Made by the author, PowerBI)

When looking at the days of the week is possible to see that the ratio between observation paid and in debt is constant among all days. What is curious is the fact that the number of transactions decreases throughout the month. As mentioned before, a higher frequency doesn't translate in a higher contribution to the model.



1 – Monday
2 –Tuesday
3 – Wednesday
4 – Thursday
5 – Friday
6 – Saturday
7 - Sunday

Figure 33 - Influence of the Date Variable by Day of the Week
(Source: Made by the author, PowerBI)

Figure 34 - Influence of the Date Variable by Day of the Month
(Source: Made by the author, PowerBI)

Another variable that has a big impact on the Dependent Variable is the Tariff Class – the classes of the vehicle. As observed in the next figure the vehicles with Class 1 are the most frequent – Motorcycles and vehicles with height measured vertically as from first axis below 1.10m.

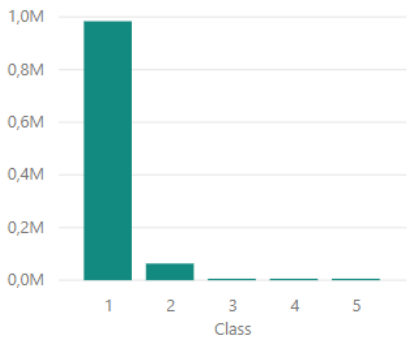


Figure 35 - Influence of the Tariff Class
(Source: Made by the author, PowerBI)

From the Total Debt Observations, around 5% already cancelled the contract with Via Verde. This Variable, Rescission Date, doesn't have a big impact on the study since it could have been voluntary or forced by the company.

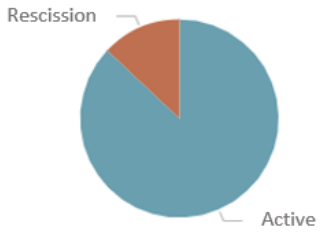


Figure 36 - Influence of the Rescission Date Variable
(Source: Made by the author, PowerBI)

It is interesting to see that the costumers with higher number of contracts are the Particulars/Single Contract with more that 70% of the Total. It is also possible to observe that 10% of the Sample doesn't have an associated type of client. For analysis purposes this label was kept, since it contributes as much as the remaining two labels with regards to the number of transactions and respective amount (Figure 39)

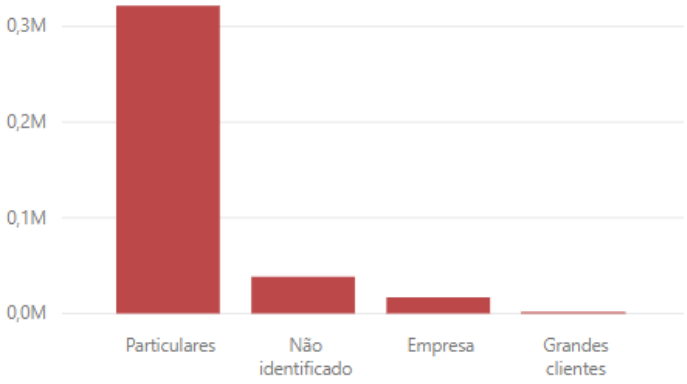


Figure 37 - Influence of the Type of Client Variable
(Source: Made by the author, PowerBI)

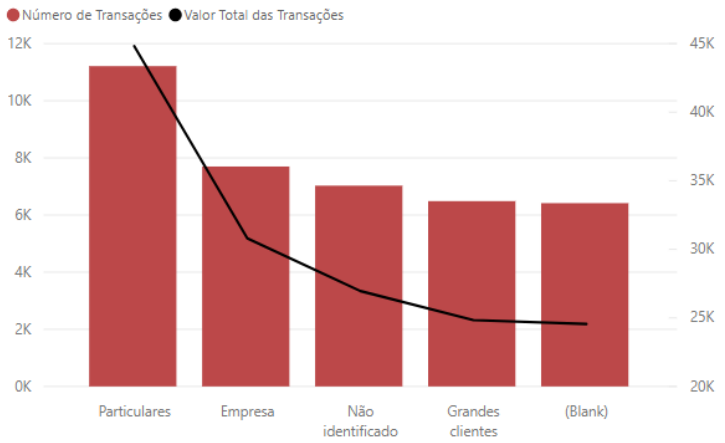


Figure 38 – Number and Value of Observations per Client
(Source: Made by the author, PowerBI)

For the comparative analysis of variables, we used bivariate graphs, due to their large usefulness in statistical analysis, which allows us to determine the empirical relationship between any variable in study and the dependent variable, so it is possible to verify how explanatory is for the model.

Following its simplicity and clarity in this study, the dependent variable was the fundamental pillar in the following analysis. Knowing if a particular variable is an influencing factor for a debt observation is an essential milestone and one of the objectives of this study. For that, we only analyzed those that greater relief and have the more importance present for the dependent variable and consequently for the model.

3.3.3. Modify

This stage is used to create, select, manipulate and transform data in need, being that, as already mentioned, the dependent variable will take the binary form.

This stage is used to create, select, manipulate and transform data in need, based on the discoveries in the exploration phase. The need to look for outliers and to reduce the number of variables also appears in this phase, with the goal to narrow them down to the most significant one.

Two new variables were created in order to have a better contribute to the model:

- Contract Status

  IF ISNULL([Rescission Date]) THEN "Contrato Activo" ELSE "Contrato Rescindido" END

- Number of Identifiers/Beacons per costumer

  DISTINCTCOUNT(tag_id)

3.3.3.1. Target Variable - Binary

The SAS code, shown below, combines the dependent variables "Payment Status", “Discount reason” e “Prevention Processes” into a discrete binary variable.

The process executed in its binary transformation was based on the relation between the transaction payment status, the number of times that the transaction was refused by the bank and if the transaction was part of a prevention process. In case of any of the previous actions happened it considered to be a payment failure, represented by the value "1", and if otherwise by "0".

```
IF (PAYMENT STATUS=1 OR PAYMENT STATUS=5 OR PAYMENT STATUS=10) OR (Discount
reason>0) OR (Prevention Processes>0)

THEN Debt =1

ELSE Debt =0

RUN;
```

Figure 39 - Binary Target Variable Code

3.3.3.2. Dimensionality Reduction

To avoid the complexity of the model, a reduction of the variables under study was performed. The less relevant and less important variables for the dependent variable (DepVar) were discarded in order to get a more synthesized and less complex analysis. This process will be discussed in the next chapter.

3.3.3.3. Outliers

An outlier is an observation that appears to be quite distanced from the others, which could have a great impact on the interpretation of results. Outliers significantly affect the process of estimating statistics resulting in overestimated or underestimated values.

These values significantly modify statistical measures which may cause entropy in the analysis and can lead to interpretation errors. To resolve the variables with outliers it is necessary to choose one of two possible solutions, remove the record or remove the variable.

The missing values and the outliers were analyzed during the data collection phase. There are several mathematical methods for identifying outliers, such as mean absolute deviation (MAD), Extreme Percentiles, Modal Center and Standard Deviation from the Mean. However, these were not considered, since a first visual analysis was enough to remove all significant outliers.

Regarding the treatment of missing values, the models that contain neural networks tend to ignore observations that contain missing values, thereby reducing the size of the training data weakens its predictive power. In order to avoid this scenario, it is appropriate to fill these values.

The different approaches for handling missing values and outliers can drastically change the results of the study. The detection of the missing values and the outliers was done visually through the Filter node. After a visual analysis of the interval variables, the following approach was taken:

- Contract Modality – 39% of the observations were blank. The decision taken was to drop the variable since the Contract Type Description gives the same idea and is full filled.
- Postal Code 4 digits – to remove the records that are blank (1 record) and records that represent a date (1 record).
- Tariff Class – 0.14% of the observations were blank. The decision taken was to fill the blank with the mode, the most frequently occurring value found.

It is important to mention that, even though there is a big range of values in the interval variables, *Price* and *Dis*count, they were not considered outliers, since they are represented in the same measure – the variety depends on the type/size of the client.

3.3.3.4. Data Partition

The last process of this stage consisted of partitioning the sample into training, where the training is performed, validation set, used in the performance evaluation, and the test sets, which is used to evaluate the performance of the trained network when in operation. It is important to understand that different partitions of the data will result in different conclusions.

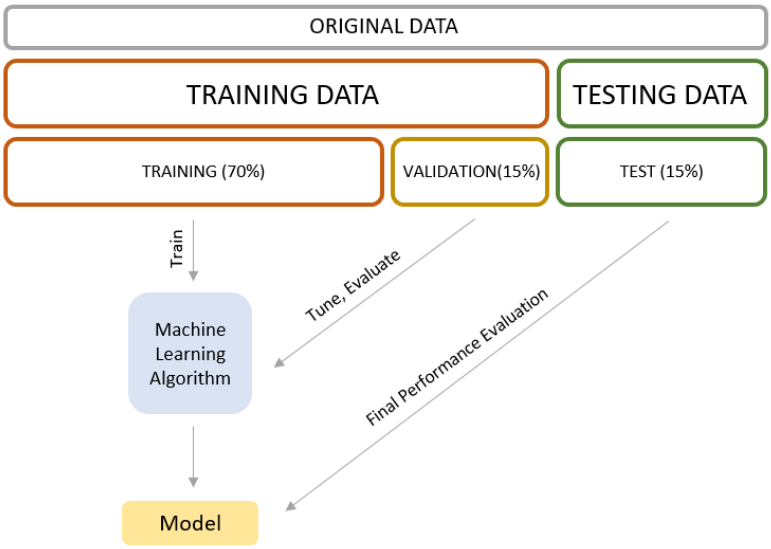


Figure 40 - Partitions used in the Predictive Model

The distribution of the sample was defined to be 70% training, 15% validation and 15% testing data. The following table displays the result of the data partition:

Level / Data Role	0	1	Totals
Train	724.731	9.271	734.003
Validation	155.366	1.921	157.287
Test	155.298	1.988	157.286

Table 5 - Data Partition

The method selected for the data partition was the *"Stratified".* This method tends to offer better results since it maintains the proportion in relation to the target variable in the training data, validation data and test data. The large training data is intended to avoid over-learning of the model.

3.3.4. Model and Assess

The model in this study is characterized by supervised learning - neural networks, decision trees and regression - which is performed by its greater precision. (Lee, Booth and Alam, 2005) In practical terms, it receives a set of input data in its function and consists of deciphering the style bars between them. The main thing is to get the most truthful and concise result in the behavioral specifications of future actions, based on historical observations.

There are a large number of algorithms used in Data Mining. However, depending on the problem and the type of data, some are more adequate than others, which does not mean that they solve the problem more precisely and quickly.

SAS allows us to create a predictive model in Enterprise Miner that is able to adjust the data according to various regression or classification. Simultaneously, the model with the best performance in terms of the mean square error, is selected by the score node to make the forecast of the data. This task is time-consuming, and it requires an excellent knowledge of each algorithm and the domain (Koblar, 2012).

For Bação (2006: 41-42), all methods of discovery of knowledge suffer from the tendency to adapt too much to the data set that is used to train [overlearning]. Only with the training data there is no way to whether the relationships found can be generalized to the entire population in study or if they only occur in the training set and therefore have no value for generalize to the population.

The first algorithm to be study was the Logistic Regression, a parametric model for further comparison to the decision trees. Stepwise was the Selection Model property chosen in the Model Selection Subgroup. This specification makes SAS Enterprise Miner to use stepwise variable selection to build the logistic regression model. A procedure to select or exclude variables from a model that is based on an algorithm that checks the importance of the variables, including or excluding them from the model, based on a decision rule. The importance of the variable is defined in terms of a measure of statistical significance of the coefficient associated with the variable for the model. This statistic depends on the assumptions of the model.

As mentioned earlier in the study, another popular and easy-to-understand algorithm used was the Decision Trees. In this step, SAS Enterprise Miner automatically train the full decision tree and consequently prune the tree to an optimal size by select split rules at each step to maximize the split decision logworth[8]. In order to get more accuracy in the model a Maximum Depth of 6, 10 and 12 was tested. This means that SAS will train the tree that includes up to 6, 10 and 12 generation of the root node.

The Gradient Boosting node was used in order to generate a set of decision trees to form a single predictive model. Gradient Boosting Algorithm uses several weak algorithms to create a more

[8] Split decision logworth is a statistic that measures the effectiveness of a particular split decision at differentiating values of the target variable (SAS Enterprise Miner)

powerful precise algorithm. Instead of using a single estimator, having multiples will create a more stable and robust algorithm. Like decision trees, boosting makes no assumptions about the distribution of the data.

Regarding the Neural Network Model there were two possible approaches that could be taken:

- Neural Network Node
- AutoNeural Node

Since the structure of the model in construction is known, the Neural Network Node was used. This node can accommodate a wider variety of nonlinear relationships between a set of predictors and a target variable than the logistic regression. The AutoNeural Node is used to find the best network configuration that best describes the relationship in a data set for further network training.

It is essential to measure the error associated with any forecast. Be aware of the difference and deviation of the forecast against the real value is essential not only to carry out a critical evaluation of the model, but also to determine the overall accuracy of the forecast data and to know if they are within reasonable limits. For this, there are metrics responsible for the evaluation of the predictive tool used, such as classification error and quadratic mean error. This will be further discussed in the next sub chapters.

3.3.4.1. Error Rate

According to Kate (Kate et al. 2004), to evaluate the results of the predictive model, involves measuring the degree of uncertainty associated with forecasts.

The confusion matrix makes it easy to see the number of correct ratings and the number of predicted classifications for each class of a given set of examples, according to the classifier under analysis. It becomes a useful tool to analyze the quality of the classifier in recognizing examples of different categories (Han and Kamber, 2006).

The Confusion Matrix is used to evaluate the quality of the predictive model, giving the proportion of false positives and false negatives of the model.

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	**TP**	FN
	negatives	FP	**TN**

Table 6 – Confusion Matrix

TP, True Positives, refers to the number of examples of the 'positive' category correctly predicted as a 'positive' category; FN, False Negatives, represents the number of examples from the 'positive' category incorrectly predicted as a 'negative' category. FP, False Positive, refers to the number of

examples of the 'negative' category incorrectly predicted as 'positive' category and TN, True Negatives, represents number of correctly predicted category negative examples as a negative category.

The Error Rate (ERR) is a predictive error used only in discrete variables (binary, multiclass), which uses the confusion matrix as support to its calculation.

$$ERR = \frac{FP + FN}{TP + FP + TN + FN} = \frac{FP + FN}{P + N}$$

The best error rate is 0.0, whereas the worst is 1.0.

3.3.4.2. Accuracy

Accuracy (ACC) is the number of correctly predicted data points out of all the data points. It is calculated as the total number of the correct predictions divided by the total number of observations:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{P + N}$$

The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by *1 – ERR*.

3.3.4.3. Other Measures

Often, accuracy is used along with precision and recall, which are other metrics that use various ratios of true/false positives/negatives. Together, these metrics provide a detailed look at how the algorithm is classifying data points.

Precision, also called as Positive Predictive Value, is the metric that calculates positive predictive values, and Recall (REC), also called as Sensitivity (SN) or True Positive Rate (TPR), is the model's sensitivity:

$$PREC / PPV = \frac{TP}{TP + FP} \qquad SN / REC / TPR = \frac{TP}{TP + FN}$$

The best sensitivity is 1.0, whereas the worst is 0.0. Same interpretation is made for Recall.

Faceli et al. (2011) considers precision as an accuracy measure of the classifier and recall as a measure of its completeness. The analysis of these two measures separately is not usually done. However, the weighted harmonic mean of these two measures created the F-measure, also called as F-Score. This metric combines precision and recall to bring a unique number that indicates the overall quality of your model and works well even with data sets that have disproportionate classes:

$$F\ Measure = 2 * \frac{REC * PREC}{REC + PREC}$$

The F Measure is best if there is some sort of balance between Precision (PREC) and Recall (REC) in the system.

3.3.4.4. ROC Curve and AUC (Area Under Curve)

Generally, sensitivity and specificity are difficult to reconcile characteristics, that is, it is complicated to increase the sensitivity and specificity of a test at the same time. ROC (receiver operator characteristic curve) curves are a way of representing the normally antagonistic relationship between the sensitivity and specificity of a quantitative diagnostic test over a continuum of cutoff point values.

In geometric terms, this curve is represented by a Cartesian graph that combines the rate of false positives (*1 - Specificity*) with the rate of true positives (*Sensitivity*):

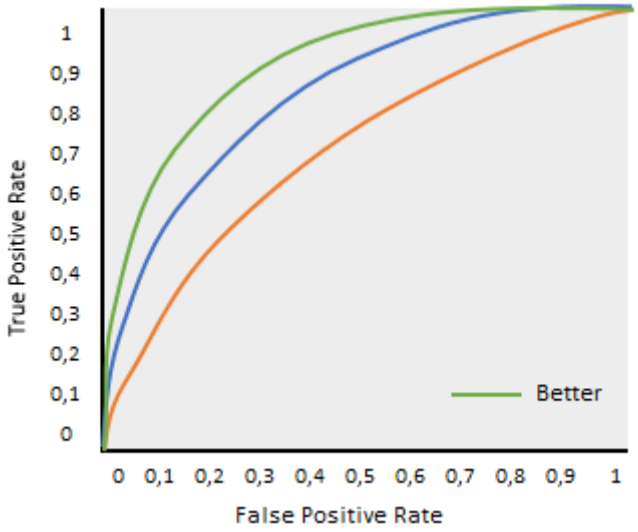


Figure 41 – ROC Curve (Source: Made by the author)

The accuracy of the statistical test is measured by the area below the ROC curve, which according to its value, which will be between 0.5 and 1, indicates how good the model under study is

4. RESULTS AND DISCUSSION

The purpose of this study was to predict whether or not a client will be in debt to Via Verde in the future, based on a set of previously defined data.

In order to evaluate the best model, several models were trained and compared in the model comparison node, SAS Miner:

- Artificial Neural Network for hidden layers 1 through 7
- Decision Tree
- Logistic Regression

This node provides criteria derived from several sources: classification measures as ROC charts and corresponding area under the curve and classification rates, data mining measures as lift and gain measures and statistical measures as Gini statistics.

From the results obtained, the ROC (receiver operating characteristic) curve shows that, for the Training Role, *the:*

- *Ensemble 10 - ensemble (5 to 4)* node, with neural network with 4 and 5 hidden layers;
- *Ensemble 4 - ensemble (4 to 7)*, with neural network with 4, 5, 6 and 7 hidden layers, and;
- *Ensemble 7 - ensemble (3 to 4)*, with neural network with 3 and 4 hidden layers,

are the best performing models since they are the ones that have the curve most adjusted to the upper left corner as shown in Figure 44.

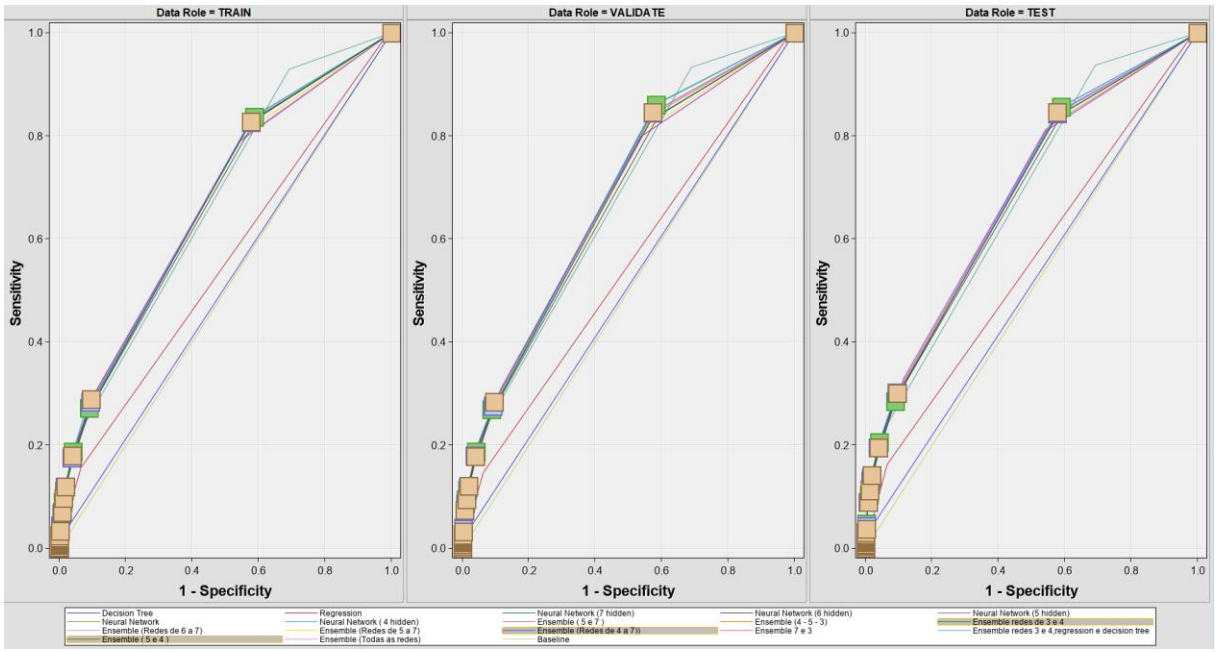


Figure 42 – Predictive Models ROC Curve
(Source: SAS Miner)

In addition to the previous point, the models with greater area below the ROC curve (Area under Curve or ROC Index Statistic) were identified:

Model Node	Model Description	Train: Roc Index	Valid: Roc Index	Test: Roc Index
Tree	Decision Tree	0,508	0,508	0,511
Ensmbl10	**Ensemble (5 e 4)**	**0,695**	**0,702**	**0,708**
Neural3	Neural Network (5 hidden)	0,69	0,697	0,705
Ensmbl4	**Ensemble (Redes de 4 a 7))**	**0,695**	**0,702**	**0,71**
Ensmbl8	Ensemble (4 - 5 - 3)	0,694	0,701	0,709
Ensmbl	Ensemble (Todas as redes)	0,694	0,702	0,71
Ensmbl7	**Ensemble redes de 3 e 4**	**0,695**	**0,701**	**0,711**
Ensmbl5	Ensemble (Redes de 5 a 7)	0,693	0,701	0,709
Ensmbl9	Ensemble (5 e 7)	0,69	0,697	0,705
Ensmbl2	Ensemble redes 3 e 4,regression e Decision Tree	0,691	0,697	0,707
Ensmbl6	Ensemble (Redes de 6 a 7)	0,693	0,7	0,71
Ensmbl3	Ensemble 7 e 3	0,684	0,691	0,7
Neural2	Neural Network	0,684	0,691	0,7
Neural5	Neural Network (7 hidden)	0,684	0,691	0,7
Neural4	Neural Network (6 hidden)	0,692	0,699	0,707
Reg	Regression	0,549	0,546	0,552
Neural10	Neural Network (4 hidden)	0,694	0,701	0,708

Table 7 – Statistic Comparison [Fit Statistic: _AUR_]

As mentioned before the AUC is the area enclosed by the ROC curve. A perfect classifier has AUC = 1 and a completely random classifier has AUC = 0.5. We can see that the best model(s) has an AUC of 0.7. This means that there is a 70% chance that the model will be able to distinguish between positive class and negative class. It is important to mention that AUC is not directly comparable to

accuracy. Some models may be poorly calibrated but still achieve a good AUC score. This happens because its relative ordering is correct (e.g. predicting [0.05, 0.95, 0.07, 0.09] is the same as predicting [0.61, 0.80, 0.65, 0.70]). Therefore, it is incorrect to interpret that the model is "70% accurate". (Tom Fawcett, 2006)

It is clear that the same nodes depicted in the graphic have exhibited a better AUC; however, there are three models that are very close to those mentioned above, which should not be excluded from analysis.

The Gini Index also shows Ensemble 10 [0.39] and Ensemble 4 [0.39] as being the most favorable, followed by Ensemble 8, Ensemble 10, Ensemble 5 and Neural 10. It is also important to mention that this Index is equivalent to the AUC measuring the area between the cumulative response curve and a 45-degree angle.

Taking the Average Square Error into consideration, we can see that the best fit is Ensemble 4, followed by Neural 10 and Ensemble 10. The best fit is determined by the smallest means squared error. However, there was a difference of less than 0.01 between the models; therefore means squared error calculations were not used as a chosen criteria.

Regarding the Data Mining Measures, the gains chart or lift chart is generally used to choose among competing models:

Model Node	Model Description	Train: Gain	Valid: Gain	Train: Cumulative Lift	Valid: Cumulative Lift
Tree	Decision Tree	13,42519059	14,69568953	1,13425191	1,146956895
Ensmbl10	**Ensemble (5 e 4)**	**189,7651573**	**186,8823188**	**2,897651583**	**2,868823187**
Neural3	Neural Network (5 hidden)	183,1313848	179,4148254	2,831313857	2,794148252
Ensmbl4	**Ensemble (Redes de 4 a 7))**	**188,0697229**	**186,112096**	**2,880697238**	**2,861120959**
Ensmbl8	Ensemble (4 - 5 - 3)	187,6956301	186,112096	2,876956311	2,861120959
Ensmbl	Ensemble (Todas as redes)	184,9871149	183,2365975	2,849871158	2,832365974
Ensmbl7	Ensemble redes de 3 e 4	183,0667857	185,3932214	2,830667867	2,853932213
Ensmbl5	Ensemble (Redes de 5 a 7)	186,2966068	183,2365975	2,862966077	2,832365974
Ensmbl9	Ensemble (5 e 7)	181,624053	183,2365975	2,816240539	2,832365974
Ensmbl2	Ensemble redes 3 e 4,regression e decision tree	178,0512467	177,1829166	2,780512476	2,771829165
Ensmbl6	Ensemble (Redes de 6 a 7)	181,7700306	178,2044752	2,817700315	2,782044751
Ensmbl3	Ensemble 7 e 3	175,0798668	181,0799737	2,750798677	2,810799736
Neural2	Neural Network	175,0798668	181,0799737	2,750798677	2,810799736
Neural5	Neural Network (7 hidden)	175,0798668	181,0799737	2,750798677	2,810799736
Neural4	Neural Network (6 hidden)	180,6274263	169,5779799	2,806274272	2,695779798
Reg	Regression	88,06262245	79,85119952	1,880626231	1,798511994
Neural10	**Neural Network (4 hidden)**	**190,1435694**	**184,6743468**	**2,901435703**	**2,846743466**

Table 8 – Statistic Comparison [Fit Statistic: GAIN and LIFTC]

The gain shows how much better the model is performing than a random model. Other aspects taken into consideration includes:

- Response rate—Ensemble 10 and Neural 10 performed better performance at around 2.3%;
    - The model with the greatest percentage response
- Captured Response—Ensemble 10 and Neural 10, at 9.1%
    - The model with the largest recorded response values using the decile range specified in the Selection Depth property.
- Misclassification Rate—Neural 10 and Ensemble 10 had a better performance at 2.44%;
    - The model with the lowest misclassification rate.

It is clear that Ensemble 10 and Neural 10 are the best predictive models. Analyzing the confusion matrix of these two models, it can be concluded that both models had a favorable outcome, with an error rate of around 1.25%.

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	97	3
	negatives	6397	507282

Table 9 – Ensemble 10
Confusion Matrix

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	98	1
	negatives	6390	507290

Table 10 – Neural 10
Confusion Matrix

Positive Predictive Power (Precision) = 0.970	Positive Predictive Power (Precision) = 0.989
Negative Predictive Power = 0.987	Negative Predictive Power = 0.987
Sensitivity = 0.015	Sensitivity = 0.015
Specificity = 0.999	Specificity = 0.999
Accuracy = 0.987	Accuracy = 0.987

In order to understand the customer profile a closer look to the profile segmentation was taken. Customer segmentation is critical for companies of all types and industries, as it helps to find the right audience profile to work with. Therefore, there has to be a differentiation in the form of treatment.

Through the SAS Node Costumer profile, it was able to group groups of people with the same characteristics, preferences, desires and similar tastes. This is usually done to treat them equally so that there may be a closer approximation of the consumer (s) to the product / service.

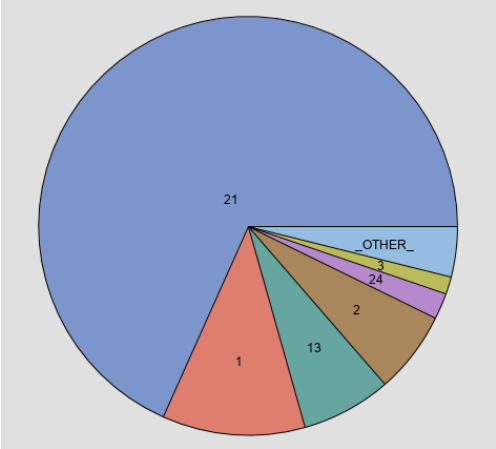


Figure 43 – SAS Segment Size Profile
(Source: SAS Miner)

The previous graph shows that Segment 21 contains more than half of the costumers, containing 68,3% of the population. One of the most interesting aspect is to understand the smaller segments, since they might reveal interesting information about the data (Segment 24, 3 and _OTHER_).

Looking at the worth of a variable is visible that the largest segment, Segment 21, resulted partly from the Price Range, Tag modality, Client Type, Tariff Class and City:

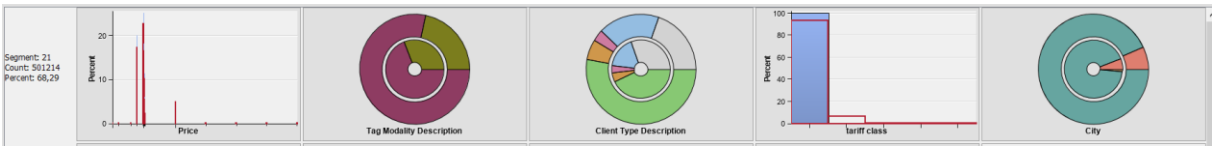


Figure 44 – SAS Costumer Segment Profile
(Source: SAS Miner)

The most frequent client is a client that mostly used tolls under 8€, has a purchased tag (78,4%), is a particular client (with 53%), has a Class 1 vehicle and is from Lisbon (93.3%).

Regarding the construction of the model, the results revealed that the variables Tag Modality, City and Tariff class are the factors that most contribute to increase the intention to use Via Verde and consequently, combining with Transaction Payment method, the ones that most contribute to commit fraud. It was also noted that the Type of Client (e.g. Particular, Company, etc) is an influencing factor in the fraudulent activity.

Looking at the days of the week, it can be verified that on Sunday the fraudulent behavior decreases to almost half. This confirms one of the first fact mentioned early in the study – the transactions routes vary from weekday and weekend. Majority of particular clients uses the same toll roads/ highways to go to work. It was also possible to seen that the frequency of the use of the toll roads decreases throughout the month. The last few days of the month registered a decrease of above 50% on the number of transactions.

According to table 7, there are 12 models with similar result when analyzing the data, therefore, all twelve models should be careful investigated. From the author's point of view these models may not be representative since the data analyze is a small sample of the entire year (around 0.25% of the 2017 transactions). Another reason for this conclusion is the fact that 99% of the transactions are done by costumers from Lisboa, and nowadays, the clients are spread throughout the country.

Within the analyzed data, the proportion of transactions that revealed being fraudulent is small. This might have two reasons: either the sample is not representative of the reality, or within the entire universe, the costumers that act fraudulent are not significant.

In conclusion, Fraud is a phenomenon whose definition is dependent on its context and is difficult to measure, yet it is of great importance to organizations from both an ethical and economic perspective. Despite the inherent complexity of the phenomenon and the inevitable incompleteness of the models, it is necessary to look for tools to understand and mitigate fraud. The literature examined showed advances in defining occupational fraud in the broad sense, considering various forms of organizational misconduct, not only limited to unlawful acts (those provided for by law), but encompassing internal actions and against the interests of the organization.

5. CONCLUSIONS

The main objective of this project was the study of a payment fraud analysis through a model capable of predicting the payment success of one of the most known Beacon technology in Portugal – Via Verde, using specific variables and historical data.

Using combinations of keywords/ expressions with the terms (data mining, debt, customer segmentation, predictive analysis) and using the Google Scholar and NOVA Discovery applications, it was possible to verify that the specific literature available that addresses the relationship between methods of predictive analysis, data mining and credit scoring for non-financial entities is still not very significant in Portugal.

This project demonstrated an approach to developing a predictive model. Two main data mining processes (CRISP-DM and SEMMA) were presented and related to a predictive model development to establish guidelines for the practical part. Then, the algorithms applied during the practical phase and their evaluation were reviewed based on the literature.

One of the key factors of this study was the data-understanding phase, where the data sources and the nature of the data was presented, cleanse and transformed. During this stage, a descriptive analysis of the data was conducted with the aim of analyzing the distribution of the variables and investigate the existence of irregularities such as missing values, outliers, and redundancy. An important decision was also made during this phase: the definition of the target variable. After developing the model, the Binary Target Variable was the most relevant for the model. The final step of this phase was the aggregation of all input variables into one Analytical Base Table. *[1. Which variables should be included in customer segmentation?]*

For the customer segmentation, majority of the records correspond to a specific profile, Client Type "Particular" with a purchased tag, with Class 1 vehicle from Lisbon. It was visible that these clients have a high use within Via Verde Tools, however with a low cost per purchase (majority under 8€ per purchase) *[2. How can a customer segmentation process be designed when looking to how often do they purchase and how much do they spend (Identify the variables that most contribute to the success/ debt situation)?]*

Although data understanding and data preparation took a large portion of the process, the focus of the project was on the predictive modelling techniques. Several configurations of logistic regression models, decision trees, artificial neural networks, and ensemble models were created and evaluated. The final model was selected based on various metrics and its performance on test data was also analyzed, confirming that the model contributes to the improvement of the payment system. *[3. How can we predict/ anticipate future frauds?]*

Table 11 presents the final classification errors for each of the final chosen models and their predictive tools:

Predictive Model	Error Rate	Target Variable
Ensemble 10 - Neural Network with 4 and 5 hidden layers	0.0126138	Binary
Neural 10 - Neural Network with 4 hidden layers	0.0125960	Binary

Table 11 – Predictive model error rate for the binary dependent variable

Regarding the methodologies used, the neural networks provided a best predictive result, presenting a clear homogeneity in any of the developed models. This result was already presented in previous studies, where the highest success percentages and most noticeable data were obtained exactly from this type of Model:

Authors	% Sucess
Kaur & Nidhi, 2013	93.3%
Ghiassi et al., 2015	94.1%
Riwinoto, M.T., Zega, & Irlanda, 2015	58.1%
Rhee & Zulkernine, 2016	88.8%

Table 12 – Neural Network Success Percentage used in previous studies

On a variable side, it was possible to conclude that some of the variables provided are not relevant for the study and therefore, were excluded. On the other hand, there is a group of very explanatory variables that contributed to the model predictive success, such as, Transaction Payment Method, Contract Modality and Client Type.

It is important to keep in mind that the network does not allow the identification of the different fraud typologies as it only allows verifying which costumer committed fraud or not. Throughout the development of the models, we can verify that these are not a sufficiently conclusive tool to say that a particular costumer will or will not commit fraud. In order to conclude this study the three main objective questions were answered:

1. Which variables should be included in customer segmentation?

   The results revealed that variables like Tag Modality, City, Client Type and Tariff Class are the factors that most contributes for the characterization of the customer segmentation. These variables combining with the Transaction Payment Method are the ones that most contributes to analyze their attitudes and behavior. We have also verified that the Transaction Date is a key factor for the analysis since there is a routine change from weekday to weekend – a high decrease of users when comparing both.

2.	How can a customer segmentation process be designed when looking to how often do they purchase and how much do they spend?	It is visible that there are two types of costumers - the ones that do daily journeys, and the ones that do sporadic journeys. The costumers that do daily journeys generally use short and less expensive routes, usually to go to work, school, etc. These costumers also do long journeys but two/three times a year. These costumers have a high use of the product but spend less amount. The second type of client are the one that do not have a routine and does not uses the product on a daily basis, but when it uses the product is usually to do a long journey, and therefore the cost is higher (Longer the journey, higher the cost).
3.	How can we predict/ anticipate future frauds?	Fraud prevention is not a static process. There is no starting or finishing point. Fraud prevention is a continuous cycle that involves monitoring, detection, decision-making, incident management, and learning. Despite the fact that the developed model did not have a conclusive result, the company focus on preventing future payment fraud can probably be done by developing a Predictive Model using the same methodology studied. The Company in study, and all the others, should strive to constantly learn from the fraud incidents and incorporate these findings into the detection process. Nowadays, the biggest challenge companies' face is to quickly identify and determine fraudulent attacks while improving the costumer experience.

In this type of companies, fraud is a dynamic process that is constantly in evolution. It is design in a way that is developing on a daily basis with new fraudulent schemes emerging. It is crucial to identify and prevent fraud to minimize the losses of companies with this crime. However, as mentioned throughout this study, we could also verify that, although the use of predictive models is a useful and beneficial tool for the companies to try to prevent these types of situations, it should be a complement to other detection methodologies and fraud prevention.

Human behavior, from the perspective of organizational studies and sociology, cannot be considered in isolation, but needs the interactionist perspective to analyze the micro-level (individual), the meso-level (culture) and the macro-level (organization) in order to better understand the causes and antecedents of fraud (VAUGHAN, 2007). If individuals commit fraud, how varied they may be, there is

an encouraging vision: not only can we prevent, in order to increase the effectiveness of procedures for dealing with fraud, but there is also the possibility of relative prediction as to agent formation.

The company in study is aware of this situation and has been working on the subject. In order to mitigate the toll nonpayment Via Verde launched "Pagamento de Portagens" portal which allows the settlement of toll debts by domestic or foreign customers before they reach the Tax Authority, avoiding the payment of fines.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The realization of this dissertation was subject to some restrictions that had an impact on the study.

The first limitation found was the fact that the data used as the basis for this case study was small and probably not fully representative of the population (96% were from Lisbon, for example). The access to data from different sectors would not only allow to compare the different type of population but also check which one is the most fraudulent, if there are entities operating in the same way.

With the results of the predictive models, it would be possible, when new clients subscribe Via Verde would be possible to classify and compare them with fraudulent clients, allowing the company to anticipation possible fraudulent behaviors. So, if the new client belongs to the cluster of fraudulent profiles it could be stated that the probability of committing fraud was quite high.

The second limitation that can be applied to almost all Predictive models is the constant acceleration adoption. Not only can traditional problems be but also problems that were never thought before.

It would also be interesting to develop the clusters identified in the previous chapter. Once the clusters are created, the application of predictive models, such as the use of decision trees, would allow us to identify common characteristics that fraudulent clients have. Knowing fraudulent clients better would make possible to characterize them and to understand more accurately and completely what characteristics can be considered risky.

7. BIBLIOGRAPHY

Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst

Accenture Consulting (2018). Winning at the point of attack

ACFE (2010). Report to the Nation on Occupational Fraud and Abuse, http://www.acfe.com/rttn. Accessed: 6/10/2010

ACFE (2016). Report to the Nations on Occupational Fraud and Abuse, 2016 Global Fraud Study

Adeniyi, D. A., Wei, Z., and Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, 12(1), 90–108.

Azevedo, A. I. R. L. and Santos, M. F. (2008). "KDD, SEMMA and CRISP-DM: a parallel overview," IADS-DM.

Akkizidis, I. S. B. V. (2006). Guide to optimal operational risk & Basel II. Taylor & Francis Group, LLC.

Alvarez, G. and Petrovic, S. (2003). A new taxonomy of web attacks suitable for efficient encoding. Computers & Security, 22(5):435–449.

Bação, F. (2009). Introdução ao Data Mining, Apontamentos Mestrado. Lisboa: NOVA IMS.

Basheer, I., and Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. Journal of Microbiological Methods, 43(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

Baesens, B., Vlasselaer, V. Van, and Verbeke, W. (2015). Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. (WILEY, Ed.).

Bhalla, D. (2014). Difference between linear regression and logistic regression. Retrieved from http://www.listendata.com/2014/11/difference-between-linear-regression.html. (Accessed in: 10/05/2018)

Brandao, G., Caldeira, E. and Pereira, A. (2014). Fraud Analysis and Prevention in e-Commerce Transactions https://www.researchgate.net/publication/287299598_Fraud_Analysis_and_Prevention_in_e-Commerce_Transactions (Accessed in: 14/12/2018)

Buecker A., Morelli T. and Shearer C. (2010). “IBM SPSS predictive analytics: Optimizing decisions at the point of impact”

Barry de Ville and Neville P. (2013). “Decision Trees for Analytics Using SAS Enterprise Miner”

Bulhmann, P. (2012). “Bagging, Boosting and Ensemble Methods” (Article)

Cassim, M. (1987). “Fraudulent and ‘Reckless’ Trading and Section 424 of the Companies Act” SALJ LJ 162

Chaudhury, M. (2010). A review of the key issues in operational risk capital modeling. The Journal of Operational Risk 5(3), 37–66.

Craven, M. W., and Shavlik, J. W. (1997). Using neural networks for data mining. Future Generation Computer Systems, 13(2–3), 211–229. https://doi.org/10.1016/S0167-739X(97)00022-8 (Accessed in: 10/05/2018)

Cumby, J. A., and Barnes, J. G. (1995). "Strategic investment in service quality: protecting profitable customer relationships", in Swartz, T.A., Bowen, D.E., & Brown, S.W. Advances in Services Marketing and Management, JAI Press, Greenwich, Connecticut, 4, 229-248.

Cressey, D. R. (1950). The criminal violation of financial trust. American Sociological Review 15 738–743.

Copeland, L., Edberg, D., Panorska, A. K., and Wendel, J. (2012). Applying Business Intelligence Concepts to Medicaid Claim Fraud Detection. Journal of Information Systems Applied Research, 5(1).

CRISP-DM: CROSS industry standard process for data mining (1999). http://www.crisp-dm.org/, (Accessed in: 05/12/2018).

Davis, E. (2005). “Loss Data Collection and Modelling.” Pp. 1–2 in Operational Risk: Practical Approaches to Implementation, ed. E. Davis. London: Risk Books.

Delen D. (2018). “Real-World Data Mining: Applied Business Analytics and Decision Making”

Faceli, K., Lorena, A. C., Gama, J. and Carvalho, A. C. P. L. F. (2011). Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. LTC. Rio de Janeiro.

Fachin, O. (2001). Fundamentos de metodologia. São Paulo: Saraiva.

Gennari, J. H., Langley, P., and Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40(1–3), 11–61. https://doi.org/10.1016/0004-3702(89)90046-5 (Accessed in: 10/05/2018)

Golmohammadi, K. and Zaiane, O.R. (2012). Data mining applications for fraud detection in securities market. In: 2012 European intelligence and security informatics conference (EISIC), IEEE. pp 107–114

Gonçalves, R. (2001). Sistema de informação para gestão de Risco Operacional em instituições financeiras.

Guo, T. a O., and Li, G. (2008). Neural Data Mining for Credit Card Fraud Detection. In Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming (pp. 12–15).

Jack, G. and Robert, J. (1995). Fraud Auditing and Forensic Accounting: New Tools and Techniques by Bologna, 2nd edition

Hadjiemmanuil, C. (2003). “Legal Risk and Fraud: Capital Charges, Control and Insurance.” Pp. 74–100 in Operational Risk: Regulation, Analysis and Management, ed. C. Alexander. London: Prentice Hall-Financial Times. Operational Risk: A Survey.

Han, J., Kamber, M., and Pei, J. (2012). Data Mining: Concepts and techniques (3rd ed.).

Hashem, I. A., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 98-115.

Jonsson, P., and Mattsson, S. A. (2005). Logistik – Läran Om effektiva materialflöden. Lund: Studentlitteratur AB, ISBN 978-91-44-04182-7.

Larose, D. T. and Larose, C.D. (2015). Data Mining and Predictive Analytics

Lee, K., Booth, D., and Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. Expert Systems with Applications, 29(1), 1–16. https://doi.org/10.1016/j.eswa.2005.01.004 (Accessed in:20/05/2018)

Liu, Y. and Schumann, M. (2005). Data mining feature selection for credit scoring models. J Oper Res Soc 56(9):1099–1108

Lu, Y. (2015). “Decision tree methods: applications for classification and prediction”, Shanghai Archives of Psychiatry, 130-133

Luell, J. (2005). Analytical fraud detection. Master’s thesis, University of Zurich (Accessed in: 18/12/2018)

Maia, A. J., de Sousa, B. and Pimenta, C. (2017). Fraude em Portugal - Causas e Contextos

Mainon, O., and Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook, 2nd Edition. Outlier detection.

Maklan, S., Peppard, J. and Klaus, P. (2015). Show me the money: improving our understanding of how organizations generate return from technology-led marketing change. Eur J Mark 49(3/4):561–595

Mestchian, P. (2003). “Operational Risk Management: The Solution is in the Problem.” Pp. 3–14 in Advances in Operational Risk: Firm-wide Issues for Financial Institutions. London: Risk Books.

Moro, S., Cortez, P., and Rita, P. (2016). A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features https://link.springer.com/article/10.1007/s00521-015-2157-8 (Accessed in:15/08/2018)

Moura, H. da S., and Silva, A. C. R. da. (2004). Auditoria de Fraude: Instrumentos na Prevenção de Fraudes Contra as Empresas.

Neelamegham, R., and Chintagunta, P. (1999). A Bayesian Model to Forecast New Product Performance

Oesterreichische Nationalbank (OeNB). (2006). Guidelines on Operational Risk Management

Pimenta, C. (2009). Esboço de Quantificação da Fraude em Portugal.

Pimenta, C., and Afonso, Ó. (2012). Notes on the Epistemology of Fraud.

Porter, M. (1990). The competitive advantage of nations. New York: Free Press, ISBN: 9780684841472.

Reinartz, W., Kumar, V. (2000). "On the profitability of long-life customers in a non-contractual setting: an empirical investigation and implications for marketing". Journal of Marketing, 64(4), 17-35.

Reeve A. (2013). “Managing Data in motion: Data Integration. Best Practice Techniques and Technologies”. 152-154

Reena, A., May, H., and Jingjing, Y. (2015). “The Journal of Portfolio Management Special China Issue 2015”, 41 (5) 92-109.
https://doi.org/10.3905/jpm.2015.41.5.092

Perkoff, R. (2010). Commercial fraud and insolvency

Santos, M. F. and Azevedo, C. S., (2005). “Data Mining - Descoberta de Conhecimento em Bases de Dados”, FCA

SAS Institute Inc. (2009). Getting Started with SAS Enterprise Miner 6.1

SAS Institute Inc. (2003). Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach, Second Edition.

Siegel, E. (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Wley & Sons, Inc.

Soares, M. (2008). Contributo do Data Mining na Detecção e Prevenção de Fraude.

Stephens, S. and Tamayo, P. (2003). “Supervised and unsupervised data mining techniques for life sciences.” Curr Drug Disc.

Sullivan, J. R. (2014). Controlling Security Risk and Fraud in Payment Systems. Pp 49-62

Tattam, D. (2011). A Short Guide to Operational Risk. Gower Publishing, Ltd, 3-26.

The Royal Society. (2017). Machine Learning: The power and Promise of computers that learn by example.

Vaughan, D. (2007). Beyond Macro- and Micro-Levels of Analysis, Organizations, and the Cultural Fix. In: PONTELL, H.N.; GEIS, G.L. (Eds). International Handbook of WhiteCollar and Corporate Crime. Cap. 1, p. 3-23.

Wahler, B. (2002). Process-Managing Operational Risk – Developing a Concept for Adapting Process Management to the Needs of Operational Risk in the Basel II Framework. http://papers.ssrn.com/sol3/papers.cfm?abstract id=674221.

Wann-Yih W., Chwan-Yi C., Ya-Jung W., and Hui-Ju T., (2004). "The influencing factors of commitment and business integration on supply chain management", Industrial Management & Data Systems, 104(4), 322 – 333.

Wan-Kadir, W. M. N., and Loucopoulos, P. (2004). Relating Evolving Business Rules to Software Design.

Wirth, R., and Hipp, J. (2000). "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, pp. 29-39.

8. ANNEXES

Anex 1

Variable	*Role*	*Level*	*Observed Values*	*Average*	*Content*	*Table*
Client_Id	*ID*	*Nominal*			Client ID	VV_Client
DepVar "Debt"	*Target*	*Binary*			Dependent Varibale	
Client_Type_Description	*Input*	*Nominal*	Particulares, Empresa Grandes Clientes, Não Identificado		Client type description	VV_Client
Contract_Id	*ID*	*Nominal*			Contract ID	VV_Contract
Contract_Modality_Id	*ID*	*Nominal*	10, 20, 26, 30		Contract Modality	VV_Contract
Postal_Code_4_Digits	*Input*	*Nominal*			Postal Code present in the contract	VV_Contract
District	*Input*	*Nominal*			District associated to the Postal Code	District_CTT
Country	*Input*	*Nominal*			Country present in the contract	VV_Contract
Contract_Rescision_Date	*Input*	*Time ID*			Contract Rescision Date	VV_Contract
Contract_Type_Description	*Input*	*Nominal*	Credito, Isento, Pre-Pago, Rent-a-Car, SIBS/DD		Payment Method	VV_Contract
Service_Type	*Input*	*Nominal*	Parques, Portagens		Service Type	VV_Net_Element
Net_Element	*Input*	*Nominal*			Highway Entry Lane	VV_Net_Element
AE	*Input*	*Nominal*			Highway Initials	VV_Net_Element
AE Name	*Input*	*Nominal*			Highway Description	VV_Net_Element
Operator	*Input*	*Nominal*			Highway Operator	VV_Net_Element
City	*Input*	*Nominal*			Operator City	VV_Net_Element

Country	*Input*	*Nominal*	Portugal		Operator Country	VV_Net_Element
Tag_Id	*ID*	*Nominal*			Tag ID	VV_Tag
Tariff_Class	*Input*	*Ordinal*	1-5		Tariff Class	VV_Tag
Tag_Rescission_date	*Input*	*Time ID*			Identifier Rescision Date	VV_Tag
Tag_Modality_Description	*Input*	*Nominal*	Compra, Aluguer, Pre-Pago CTT, Facturação Repartida		Identifier Modality	VV_Tag
Transaction_Id	*ID*	*Nominal*			Transaction ID	VV_Transactions
Transaction_Type	*Input*	*Nominal*	10		Transaction Type	VV_Transactions
Market_Id	*ID*	*Nominal*	1, 2, 150, 160		Market ID	VV_Transactions
Payer	*Input*	*Nominal*			Customer ID	VV_Transactions
AVC	*Input*	*Ordinal*	0-5		Tariff Class taken from the toll, where 0 is an error in the reading	VV_Transactions
Format_Id	*ID*	*Nominal*	S		Flag for toll transactions, parking in parks and fuel station ("S");	VV_Transactions
Usage_Datetime	*Input*	*Time ID*			Transaction date	VV_Transactions
Product_Id	*ID*	*Nominal*	1 - 2		Product ID	VV_Transactions
Service_Id	*ID*	*Nominal*	1 - 2		Service ID	VV_Transactions
Num_Operators	*Input*	*Nominal*			Number of Operators present in the transaction	VV_Transactions
Entry_Date	*Input*	*Time ID*			Entry Toll Date	VV_Transactions
Entry_Time	*Input*	*Time ID*			Entry Toll Time	VV_Transactions
Entry_Lane	*Input*	*Nominal*			Entry Toll Lane	VV_Transactions

OPID	*Input*	*Nominal*			Entry Toll Operator ID	VV_Transactions
Exit_Date	*Input*	*Time ID*			Exit Toll Date	VV_Transactions
Exit_time	*Input*	*Time ID*			Exit Toll Time	VV_Transactions
Exit_Lane	*Input*	*Nominal*			Exit Toll Lane	VV_Transactions
OPID1	*Input*	*Nominal*			Exit Toll Operator ID	VV_Transactions
Price	*Input*	*Interval ar*			Entry Toll Price	VV_Transactions
Discount	*Input*	*Interval ar*			Discount Toll Price	VV_Transactions
Transaction_Payment_Method	*Input*	*Nominal*			Transaction Payment Method	VV_Transactions
Payment Status	*Input*	*Nominal*	1,4,5,7,9,10,13,30		Transaction Payment Status	VV_Transactions
Discount_reason	*Input*	*Nominal*	0-36		Number of payment refusals	VV_Transactions
Prevention_num	*Input*	*Nominal*	0-18		Number of times that the client was part of a Prevention Process	VV_Transactions