NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Predicting Fraud Behaviour in Online Betting

A Data Mining approach

Margarida de Sousa Tedim

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

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

Universidade Nova de Lisboa

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

PREDICTING FRAUD BEHAVIOUR IN ONLINE BETTING

by

Margarida de Sousa Tedim

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Information Analysis and Management.

Advisor / Co Advisor: Prof. Roberto Henriques, PhD

November 2018

DEDICATION

To my Grandmother, for all her love, kindness and dedication.
She'll live in me forever.

ACKNOWLEDGMENTS

First, a special thanks to my parents. To my Mother for always being there, supporting all my pains celebrating all my successes and never leaving my side even when far away from me. To my Father for being my inspiration, my role model and an example during all my life. For trusting in me, encouraging me to go further and always give my best. Thank you.

To my sister, Filipa, for always being there to remember me that I'm beautiful, I'm intelligent and I can do anything that I wish for.

To Ricardo, there are not enough words to thank for the unconditional support, love and dedication during all this project. For being my best friend in all the moments. For always being the calm during the storm. For listen to all my doubts and celebrate even the smallest achievement. For always believing in me even when I didn't. For everything, thank you.

To my advisor, Professor Roberto Henriques, for having followed me through all the project, giving me guidance and support. To my co-advisor Professor Cristina Marreiros for the constant availability and all the initial guidelines.

To all my company and co-workers that made this project possible. In special to José for showing interest in this work even when it was only a draft and for keeping on believing on it during all its development. An even more special thanks to Raquel, for everything she taught me and for always having my back. For having the perfect words in all the situations. Thank you.

To my best friends, Inês, Linda and César. For always being by my side even being all far away from me. For the constant support, helping me and giving me courage to overcome every situation. Thank you.

To all my friends that directly or indirectly help me through this challenge, in special, my oldest friends.

For all the strength and wise advices, and for always being there to listen, a special thanks to Dra. Mafalda.

To my family: Tedim, Sousa, Coutinho and Pereira. Thank you.

ABSTRACT

Fraud isn't a new issue, there are discussions about this matter since the beginning of commerce. With the advance of the Internet this technique gained strain and became a billion-dollar business. There are many different types of online financial fraud: account takeover; identity theft; chargeback; credit card transactions; etc. Online betting is one of the markets where fraud is increasing every day.

In Portugal, the regulation of gambling and online betting was approved in April 2015. In one hand, this legislation made possible the exploration of this business in a controlled and regulated environment, but on the other hand it encouraged customers to develop new ways of fraud. Traditional data analysis used to detect fraud involved different domains such as economics, finance and law. The complexity of these investigations soon became obsolete. Being fraud an adaptive crime, different areas such as Data Mining and Machine Learning were developed to identify and prevent fraud.

The main goal of this Project is to develop a predicting model, using a data mining approach and machine learning methods, able to identify and prevent online financial fraud on the Portuguese Betting Market, a new but already strong business.

KEYWORDS

Online Fraud; Betting Market; Data Mining; Machine Learning; Portugal.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

FF	Financial Fraud
DM	Data Mining
AI	Artificial Intelligence
ML	Machine Learning
SVM	Support Vector Machine
TO	Turnover
GGR	Gross Gambling Revenue
NGR	Net Gambling Revenue
KPI	Key Performance Indicator
WD	Withdrawal
EDA	Exploratory Data Analysis
MAE	Maximum Absolute Error
MEA	Mean Square Error
MLP	Multilayer Perceptron
AML	Anti-Money Laundering
SRIJ	Serviço de Regulação e Inspeção de Jogos (Gambling Regulation and Inspection Service)

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Online Betting became legal in Portugal in April 2015. During 2016, there were made available 2 licenses for Sports Betting and 4 licenses for Casino. Currently 4 more sports betting licenses and 3 more casino licenses are already in operation. Analyzing the *Serviço de Regulação e Inspeção de Jogos* – Gambling Regulation and Inspection Service (SRIJ) Report of the 2nd quarter of 2018 we can conclude that this is a business that is growing every year. The volume of gross revenue increased about 46% when considering the 2nd quarter of 2017 and the 2nd quarter of 2018. Besides this, the number of new registrations also increased in each period. During the analyzed period, and considering all the 8 exploration entities, about 103 thousand players made their registration, more than 40 thousand new registrations when compared with the same quarter in 2017 (SRIJ, 2018).

Having this in mind, we can understand that financial fraud (FF) and match fixing is a matter that is concerning all the institutions with this type of business (Banks, 2012). Being this a field where money is in constant motion and isn't strictly controlled, fraudulent practices become more susceptible to occur. For example, we consider dishonest behaviors when players have more than one account registered in the same website or when they suddenly deposit or withdraw big amounts of money, contrarily to their normal behavior, or when they use a false identity to register. Also, money laundering (AML) is a way of crime that can be combined with fraud and can occur in online betting. In a website, if a player deposits big amounts of money and withdraws it without ever playing it, it can be considered as an AML suspicious situation. This type of cases has already occurred in the overall Betting Market. The principal problem is that the techniques used for controlling and monitoring fraud are mainly manual and outdated. This leads to most fraud cases never being identified or being identified only after they have occurred.

In this context, the development of a model capable of identifying and preventing, on an early stage, fraudulent behaviors using statistical and machine learning (ML) methods has become essential. Data Mining (DM) approaches developed nowadays use frequently Logistic Regression and Artificial Neural Networks models to detect FF (Albashrawi & Lowell, 2016; Sahin & Duman, 2011). Knowing that in Online Betting there are different types of fraud and that this field can suffer changes, other DM supervised logarithms such as: Decision Trees; Support Vector Machine (SVM); Bayesian learning; Discriminant Analysis and Random forests could also be analyzed (Bhowmik, 2008). In another perspective, some authors defend that it is better to use non-supervised algorithms to predict fraud (Xu, Sung, & Liu, 2007). This is an approach that can also be considered having in mind that we are analyzing a very recent business, with just over two years of operation in Portugal, with not such a big fraud database (Kordon, 2010).

The literature that is currently available on this topic is mainly concentrated in explaining and constructing predicting models for financial types of fraud that occur in the banking and health insurance industries. Understanding that Online Betting is a field that is growing every day, generating big amounts of revenue, studying and monitoring fraud has become truly necessary. Furthermore, considering that there is a lack of investigations of this matter in the Portuguese business and also that the techniques used to prevent it are becoming obsolete, the main goal of this project is to create an efficient model capable of predicting dishonest behaviors and diminish the cases of financial fraud.

1.2. STUDY OBJECTIVES

As explained before, the core goal of this project is to implement a predictive model in a Betting Industry in order to improve fraud detection and diminish the cases of dishonest behaviors. The creation of it will involve the traditional five different steps of a data mining process: 1. Collecting the data; 2. Data Preparation; 3. Selecting and Transforming Variables; 4. Processing and evaluating the Model; 5. Test and Validate the Model (Fig.1).

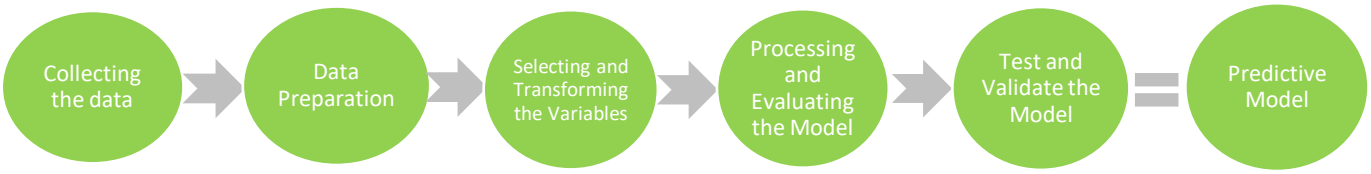


Fig. 1 – Steps to Create a Predictive Model

Having this in mind, the main objectives of this project were settled in order to understand and clarify different topics such as:

- Understand if the risk variables are properly established and coherent with the environment that we are studying;
- Clarify how the company should react when in presence of different types of fraud;
- Conclude on which are the best algorithms to predict fraud;
- Realize whether it is correct to exclude outliers during the modify phase or whether it is better to keep outliers that may contain valuable information considering that we're dealing with potential fraudulent customers;
- Identify which are the indicators that better describe fraudulent behaviors;
- Perceive if it is correct to use the same procedures when dealing with different fraudulent cases.

2. STUDY RELEVANCE AND IMPORTANCE

The project described in this proposal is important, not only for the organization where the model will be implemented but to economic and social areas as well.

Looking first to the organization perspective, this model is relevant, understanding that fraudulent behaviors can lead to large profit losses, state penalizations and overall disturbance of the company. Besides this, knowing that the techniques used until now in this organization are outdated, leading to slow and complex processes, this model would improve the operation of the business and allow the departments involved in this process to have more accurate and faster results. This is a gain having in mind that releasing the human resources of the traditional techniques will allow them to lose less time in identifying fraud cases and concentrate more in understanding and preventing this type of behaviors. Furthermore, knowing that this is the first time that machine learning is implemented on a betting market to predict fraud, the organization will also benefit by being one step ahead of the competition.

Thinking now in an economic and social perspective, this model can also conceive benefits. When dealing with fraud we can realize that we are handling people with criminal intentions, which in our case are associated to a betting market but can also be expanded to other businesses. According to the Kroll Global Fraud Report, “the number of businesses suffering a financial loss as a result of fraud has also increased from 64% in the previous survey period to 69% this year.” (Dajani, 2015) The facility of committing this type of crime, which gains strength with the advancement of the technology, can explain this growth on the fraud cases. In this context, the model proposed by this project would be directly related with diminishing fraud, specifically in the online betting market but having the possibility of reaching other markets too. This is an improvement thinking that we are dealing with criminal behaviors and the faster we stop them the less damage they make to the economy.

3. LITERATURE REVIEW

3.1. DATA MINING

Data Mining is defined as the process of searching useful similar patterns in the data, with the main goal of finding unknown relations that can improve the business. It combines different algorithms that are associated to specialized computational methods that derived from the fields of statistics, artificial intelligence and machine learning. The use of these advanced analytics is what differentiates data mining from the remaining traditional statistics methods (Kotu & Deshpande, 2014).

The standard process of finding similarities and relationships on data is defined by the following phases: (1) data preparation, (2) data mining and (3) data presentation (Fig.2) (F. Chen et al., 2015). Over the years, different methodological frameworks were developed to characterize this process: CRISP-DM, SEMMA, KDD, DMAIC, etc (Shafique & Qaiser, 2014). All these methodologies have similar characteristics that, in a more or less complete way, have as purpose to build a successful data mining solution (Mariscal, Marbán, & Fernández, 2010).

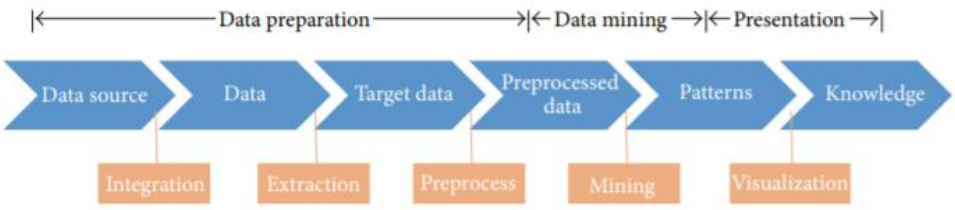


Fig. 2 – The Data Mining overview
(Feng Chen, 2015)

Knowing that in the past decade more data has been created than in the entire previous history and also that the prediction for 2020 is that more than 5,200 gigabytes will be created for every human being, it has become crucial to develop methods capable of extracting and interpreting knowledge as quick as possible (Gantz, Reinsel, & Shadows, 2012). The origin of Data Mining is directly related to this exponential growth of data normally named as Big Data (H. Chen, Chiang, & Storey, 2012). Big data is a term used to describe large and complex data sets that, due to their size and due to being in constant change, have made traditional methods of data processing slow and obsolete (Dean, 2014). Having this in mind, the solution started by implementing models that are based in data, called Data-Driven models. "*The importance of Data Mining arises from the fact that the modern world is a data-driven world.*" (Kantardzic, 2011). These models are focused on computational intelligence and machine learning methods: using a training data set that is representative of all the system's behavior, an algorithm is ran to find unknown but valuable relations between inputs and outputs (Fig. 3) (Solomatine, See, & Abrahart, 2008).

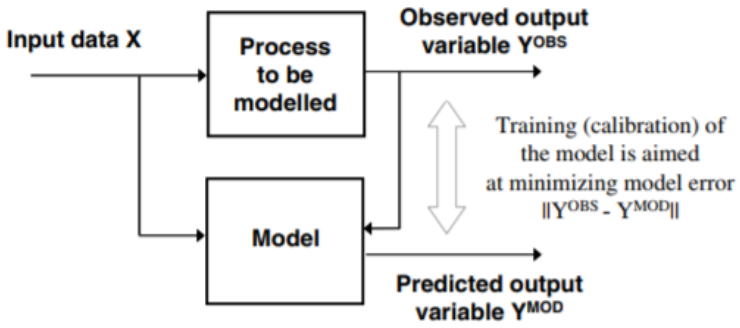


Fig. 3 – General approach to modelling
(Solomatine et al., 2008)

In the past years, data mining has developed different functionalities such as: classification and regression, clustering, association analysis, time series analysis and outlier analysis. In general, all these tasks can be classified into two categories: descriptive and predictive (Kantardzic, 2011). Descriptive modeling has as aim to produce new and valuable information using the available dataset. It finds patterns or tendencies on the data and uses them to help on decision making. On another hand, the goal of predictive modelling is to produce a representative model, based on historical or current facts, that is able to predict unknown or future values (Kotu & Deshpande, 2014). It is necessary that all the training set has input and output attributes. The *target variable*, the variable to be predicted, is given for each observation (Han, Kamber, & Pei, 2012).

All data mining tasks use machine learning algorithms to either describe or predict relevant information from a training set. Logistic regression, neural networks, k-means and decision trees are some examples of these algorithms that enable computers to collect, transform and learn in order to achieve the optimal data mining solution. The recent market changes have forced the optimization of processes, both human and technological. There is an urgent need for decision making to follow the evolution of data. The solution has become to make systems autonomous and intelligent to the extent that they are capable of replacing Man.

There are two principal methods of learning that can be related to data mining: supervised and non-supervised learning. The main difference between them is that the first approach has a target variable or class label and the second approach does not (Dean, 2014). Supervised data mining uses a set of input variables to find and understand the quality of the output variables. It constructs a model based on known input and output variables by generalizing the similarities between them. Then these relations are used to predict for the data set where the output variables are unknown. The more labeled records there are, the better performance the model will have. On the contrary, in non-supervised learning there are no output variables to predict. It is a technique that discovers hidden patterns in unlabeled data (Kononenko & Kukar, 2007).

Unsupervised learning is sometimes associated with describing the data and supervised learning with predicting it. To construct a data mining project, you can either choose one approach or use both. For example: if you have a large and complex dataset you can use clustering, unsupervised learning, to

understand in which group each record belongs. Then you can use regression or classification, supervised learning, on these clusters and predict valuable output variables (Witten, Frank, & Hall, 2011).

To conclude, data mining being directly related with data is a process that can be implemented in many working environments. It is frequently associated with economic businesses but it is not limited to this area. In the last decades, DM has been heavily used in: the medical field, helping to identify the best practices in advance by making records of the patients' diagnoses; political issues, like identification of potential voters or detection of terrorists and criminals; security and quality control (F. Chen et al., 2015). In addition, it has been used in fraud detection by banks, health insurance and telecommunication companies. In this project, we will focus on studying and creating a data mining solution for fraud on the online betting market.

3.2. Data Mining and Fraud

Fraud is an issue that has always been present in the overall business environments but that has gained importance with the development of technology. The Internet of things has brought a plethora of new challenges to explore and it has also triggered new ways of harming companies by creating new techniques of fraud. According to the latest report to the nations of the Association of Certified Fraud Examiners (ACFE), the enterprises in their study estimated a loss of 5% of revenue in a year as a result of fraud. This equates to a loss of more than $6.3 billion, resulting in an average of $2.7 million per case study (ACFE Report, 2016).

The successive growth of the number of fraud cases has led enterprises to focus more on creating processes and building systems able to predict and prevent further fraudulent attacks. In 2014, the EY Forensic Data Analytics Survey reported that 65% of the studied enterprises used spreadsheet tools like Microsoft Excel to do Forensic Data Analytics (FDA) (EY, 2014). Although these tools are important for any FDA program when describing, grouping and filtering, they are not sufficient when identifying and preventing. The same report in 2016 shows that certain developments have been made regarding FDA. Companies are gradually becoming more concerned about reducing the fraud attacks, as the 2016 EY Forensic Data Analytics Survey shows that 74% of the high level executives that answered the survey agree that they need to use better FDA tools in order to improve their current anti-fraud procedures (EY, 2016). Despite this major effort to reduce fraud cases, there are many companies that still do not invest enough resources in this problem either because there is no budget for it or because they have not yet become aware of the losses that this phenomenon can cause.

In general, fraud can be defined as the criminal act of misleading others with the purpose of damaging them or their businesses, in order to get something of worth for their own advantage (Mukherjee, Mukherjee, & Nath, 2016). Normally fraud involves direct legal consequences but in certain cases it can be solved internally. In this project, the term fraud will be referred to in overall, considering cases in which legal consequences can be applied and cases where they are not necessary (Phua, Lee, Smith, & Gayler, 2010).

Nowadays, companies have to be prepared not only to deal with external fraud but also with the possibility of internal fraud (Fig. 4). Firm-level fraud can be committed either by managers, considered high-risk fraud, or by employees, low-risk fraud. Community-level fraud or external fraud can be divided into three main profiles: the average offender; the criminal offender and the organized crime

offender (Phua et al., 2010). These two last profiles described are those that usually concern and affect businesses the most because they usually commit successive fraud attacks in order to find gaps in the detection systems that enable them to profit more and more from companies.

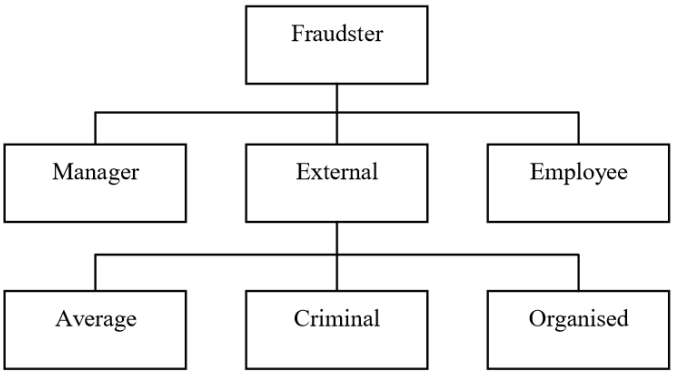


Fig. 4 - Hierarchy chart of fraudsters: firm-level and community-level

(CLIFTON PHUA, 2010)

In order to identify the fraudsters illustrated in Figure 4, usually either internal or external audit controls are made. Although these controls are important, they normally only identify that a transaction was fraudulent after it has already been made. Having this in mind, the solution would be to consolidate these controls with data analytics. This union would return better results when identifying and preventing fraud. The implementation of data mining tools would allow businesses to improve in different levels: (1) improve efficiency by automating the processes; (2) repeatable tests that would allow enterprises to control fraud at any time; (3) wider coverage by using all the database instead of controlling only alarming situations; and (4) early detection by using systems that enable the company to quickly detect a fraud transaction (Mukherjee et al., 2016).

The systems and software which enabled the creation of an antifraud control evolved from different research areas such as artificial intelligence, auditing, database, econometrics, expert systems, algorithms, machine learning, statistics, computing, visualization and others (Phua et al., 2010). It is curious to notice that this new technologies are at the same time creating weakness, which enable fraud to be more accessible to new fraudsters, and producing tools that can be used to provide smarter fraud detection and investigation techniques (Hipgrave, 2013).

In 2014, the ACFE report concluded that one of the most effective tools when reducing fraud losses and fraud scheme duration is proactive data monitoring and analysis. It is very important for antifraud control as it reduces and prevents fraud attacks (Bănărescu, 2015). Regarding the study made in 2014, organizations that used proactive data monitoring/analysis registered a near 60% of reduction in median loss to fraud and faced an approximately 50% reduction in median duration of fraud scheme compared to those that did not use this tool (EY, 2014).

The use of big data intelligence to deal with fraud allows businesses to earlier identify fraud risk and easily find trends and patterns in different types of data, both structured and unstructured, and as a result they can not only solve investigations but also prevent crime (Hipgrave, 2013).

3.3. DATA MINING IN BANKING, MONEY LAUNDERING AND INSURANCE FRAUD

Over the last decades, many articles were published concerning data mining techniques when identifying and preventing fraud. These researches can be categorized into three main financial contexts: internal, banking and insurance (Albashrawi & Lowell, 2016). These businesses are the ones that usually are more sensitive to fraud attacks, probably because they deal with large amounts of money, different technologies and many human situations.

3.3.1. Financial Fraud

Financial fraud disclosed a large amount of cases in the past years and therefore has caused major economic losses. Having this in mind, it has led to many of the papers written about antifraud control being related to this type of crime (West & Bhattacharya, 2016). When referring financial fraud, this includes a set of fraud techniques such as: financial statement fraud; automobile insurance fraud; corporate insurance fraud; money laundering; health insurance fraud; credit card fraud; occupational financial fraud and others. For each of these crimes different data mining methods were used and therefore different algorithms were tested, some performed better in one fraud type and others in another fraud type (West & Bhattacharya, 2015). *"For example, the logistic model can help in detecting financial fraud in automobile insurance, corporate insurance and credit card but it can be considered the best-performing method in the context of corporate insurance fraud".* (Albashrawi & Lowell, 2016)

Regarding fraudulent financial statements, Kotsiantis et al. used different data mining algorithms such as Decision Trees, Neural Networks (NN), Bayesian network, Support Vector Machine (SVM) and nearest neighbor in a sample of 164 Greek firms where 41 committed fraud and 123 didn't. In terms of accuracy, decision trees achieved the best performance managing to correctly classify 91.2% of the total validation set. Besides using individual classifiers they've also tested the database using an ensemble classifier. They've concluded that combining the previous tested algorithms using a stacking variant methodology was the best approach in terms of performance, being able to correctly classify 93.9% of the total validation sets (Kotsiantis, Koumanakos, Tzelepis, & Tampakas, 2006). In 2009, Deng and Mei used an unsupervised learning approach to construct a model able to predict fraudulent financial statements. They've combined both Self-Organization Map (SOM) and K-means clustering algorithms into a clustering model V-KSOM that they applied to a dataset of 100 Chinese firms. As performance measure they have used the Silhouett Index, which is a method that validates the consistency within the clusters. It ranges between -1 and 1 and a high value indicates that the data is well grouped to its own cluster. In the Deng & Mei study, the highest value of Silhouett index was 0.2707 with an accuracy of 89%, proving that unsupervised algorithms can also achieve good results when predicting fraud (Deng & Mei, 2009). Having still in consideration financial statement fraud, in 2008, Liou compared three popular algorithms when building two models: one able to predict business failure (BFP) and other able to detect fraudulent financial reporting (FRD). He tested Neural Networks, logistic regression and decision trees algorithms and concluded that, in both models, logistic regression was the algorithm that exhibits the best results, registering an average accuracy of 99.05% on FRD model and 96.5% on BFP model. These results led this author to conclude that logistic regression is an appropriate methodology for detecting fraudulent reporting and predicting firm failures (Liou, 2008).

Nowadays, financial fraud can also be associated with credit card fraudulent transactions. Yeh & Lien compared six different and popular data mining algorithms in order to study the probability of default of credit card clients. They used a database with 25,000 payments observations where 5,529 were

targeted as cardholders with default payment (fraud). Their main idea was to find the optimal data mining algorithm that could be used to represent the real probability of credit card default. They have compared the quality of the six algorithms, K-nearest neighbour, Logistic Regression, Discriminant analysis, Naïve Bayesian, Neural Networks and Classification trees, using different quality measures: to measure classification accuracy they used lift curves; to estimate the real probability of default they used the Sorting Smoothing Model, by comparing each model with the predicted probability. Neural Networks was the model that performed better in both classification accuracy and predictive accuracy of probability of default, registering a coefficient of determination (R^2) equal to 0.9647 (≅1); a regression intercept of 0.0145 (≅0) and a regression coefficient of 0.9971 (≅1) (Yeh & Lien, 2009). More recently, Dharwa JN & Patel AR proposed a hybrid approach for fraud detection of online financial transactions. They used a database of online shopping transactions to create a model that incorporates data mining techniques, statistics and artificial intelligence in a single platform. This model can be defined as a transaction risk generation model that contains five major components: density-based clustering; linear equation; rules; data warehouse and Bayes theorem. The use of an unsupervised data mining technique is an advantage when thinking that new fraud techniques may also be detected. Flexibility is also guaranteed by the facility of adding new rules or changing old ones to the model. In addition, the Bayes theorem phase allows the model to adapt to changing behaviours of the genuine customer as well as the fraudster (Prof, 2011). In 2014, Olszewski proposed a fraud detection model based on the user accounts visualisation and threshold-type detection. The approach applies the Self-organization map as a technique to visualise the user accounts but using a method of matrices visualization and not the standard vector visualization. After, the fraudulent accounts were identified using the threshold-type binary classification algorithm. This model was tested in three different fraud detection areas: telecommunications; credit card and computer network. In order to measure the quality of the model, ROC curves were employed in the three datasets. They compared the performance of the proposed model with three other reference fraud detection approaches and, in the three examinations, the proposed model registered better accuracy, being able to classify correctly 87.50% in telecommunications and computer network database and 100% in credit card database. These differences on the performances may be justified by the fraudster-specific behaviour - a credit card thief may choose to charge a stolen or cloned card excessively at once but a telecommunication fraudster would be interested in benefiting from the stolen account for a long period (Olszewski, 2014).

3.3.2. Money Laundering

Money laundering can be defined as the process of making illegal income appear legal, by using a legal intermediate such as large investment funds hosted in investment banks. This criminal act is becoming more sophisticated and complex every day. Le Khac & Kechadi developed an efficient solution for anti-money laundering (AML) by constructing a data mining-based approach testing multi algorithms, such as clustering, neural network, genetics algorithm and heuristic algorithm. They applied these techniques together and concluded that their approach would be able to improve the process of detecting ML cases within the investment activities, in particular in terms of running time (Le Khac & Kechadi, 2010). In 2016, Khalaf & El Khamesy tested different neural network types in a bank database with the purpose of creating an AML optimal solution. Although the comparison of the classification results by the different neural networks algorithms showed that the multi-layer perception was the technique that performed better, the Linear Neural Network registered a high performance when training, selecting and testing the data and also recorded the lowest error. For all these reasons, the

authors considered this model as the champion model and agreed that he could be applied in other financial transactions (Khalaf Ahmed Allam El-Din & El Khamesy, 2016).

3.3.3. Insurance Fraud

In what concerns insurance fraud, different studies were already made and can be categorized into two main areas: automobile and health insurance. Viaene, Dedene & Derrig used a database with 1,399 closed personal injury protection (PIP) automobile insurance claims to create a model able to predict fraud. They have explored the explicative capabilities of a Neural Network classifier and reported the findings of applying this classifier on PIP. They have also compared the performance of the NN algorithm with different popular data mining algorithms, such as decision trees and logistic regression (Viaene, Dedene, & Derrig, 2005). In 2014, Rodrigues & Omar also developed a model for antifraud in automobile insurance by using a multi classifier system. They gathered a dataset with 15,421 cases of suspected automobile cases of fraud. Their main objective was to find an algorithm that would allow companies to reduce costs with fraud. Having this in mind, first they created a cost model to fraud detection that allowed them, in the end of the process, to compare the performance of each algorithm. In a second phase, they tested Naïve Bayes, SVM, C4.5 and others individually and also using an ensemble method with average vote function decision. They have concluded that the combination of the classifiers would allow companies to save a higher amount of money concerning fraud than if they were used individually (Rodrigues & Omar, 2014). To conclude on automobile fraud detection, Pinquet et al. applied a bivariate probit model with censoring on a Spanish database with 2,567 suspicious claims. Their main goal was to eliminate the selection bias that is created with traditional auditing policies by using a pure random auditing strategy. This controlled experiment provided an estimated fraud probability for new claims which are not exposed to selection bias. Their results showed that random auditing enables the insurance company to reduce their costs with fraud when compared to the classical audit strategy (Pinquet, Ayuso, & Guill??n, 2007). Healthcare insurance is an area that has also been target of fraudulent attacks. In 2006, Yang & Hwang constructed a model based on process mining able to detect fraud and abuse in healthcare insurance companies. They gathered a dataset with 1,812 medical cases and divided it in fraud and normal cases. First, they used the structure pattern discovery algorithm in order to find patterns in the data, then translated them as features and, finally, they filtered by the feature subset selection algorithm. To evaluate the quality of their detection model, they used specificity and sensitivity measures. Their conclusions were that a structured detection model would be more efficient when detecting fraudulent and abusive cases than a manually constructed detection model (Yang & Hwang, 2006). More recently, Thiprungsri & Vasarhelyi introduced an unsupervised learning model using K-mean algorithm with eight clusters, which was applied in the accounting domain, in particular in the field of audit. Their sample contained 40,080 group life insurance claims of a major US insurance company. After dividing the dataset into eight clusters, they analyzed in detail the ones that were less populated and discovered some unusual characteristics among them. Using cluster-based outlier technique and evaluating which observations had less than 0.6 probabilities of belonging to the cluster, they concluded that 568 claims could be considered as anomalies. The authors consider cluster analysis as a good candidate for fraud and anomaly detection because this unsupervised learning technique surpasses the need to have a structured sample with fraud and non-fraud cases (Thiprungsri & Vasarhelyi, 2011).

3.4. DATA MINING AND ONLINE GAMBLING

In the field of online gambling, data mining techniques are still starting to be explored. Even though this industry is old if you consider it by the physical space (casinos, poker and sports betting on field), it was only legally released online in 1996 in Caribbean and Central America countries (Wood & Williams, 2009). Over the last twenty years, this business has expanded globally and technologically and has allowed customers to play 24 hours a day, seven days a week from home, work or public spaces. This explosive growth created a plethora of new data to explore: wagers; withdrawals; deposits; open rates and even mouse clicks. All this KPI's can be applied in data mining creating a world of possibilities to study (Philander, 2014).

One of the subjects that concern more online gambling companies is to guarantee a responsible gaming. Among other areas, this concept is competent of protecting vulnerable customers; prohibiting underage players and delivering a fair game experience in overall. In 2014, Philander proposed a model able to detect high-risk online gamblers by comparing nine different supervised learning algorithms. Random forest was the algorithm that performed better when classifying likely problem gamblers in training data, but in what concerns hold-out testing samples, neural networks appeared to be the most useful classification method. Despite these results, the main conclusion of the author was that there is a clear need for hold-out testing in this type of data mining research. The available variables in his dataset were considered to be insufficient to reasonably predict high-risk customers (Philander, 2014). Adami et al. used a previous study conducted in live action sports bettors of *bwin platform* and improved it by increasing the set of behavioral markers. Their main goal was to better segment and identify problematic gambles. In order to achieve these better results, they have proposed to add to the original study two indicators that would be able to reveal unsustainable gambling behaviors. This new set of indicators included not only variability of bet size, intensity and frequency of betting and trend of wagered money, but also two important markers: one able to highlight fluctuations between intervals of increasing wager size followed by rapid drops and another able to account the total number of different games played per day by the same gambler. To achieve results, they have divided their dataset into five clusters using k-means algorithm. Although the authors believe they have made an advance on studying gambles behavioral, introducing the possibility of identifying medium-risk customers, they concluded that in order to get even better results physiological studies, using questionnaires, should be integrated in the study (Adami et al., 2013).

In online gambling, more specifically in sports betting, the settlement of the probability of an outcome was always a subject that triggered interest among the bettors. Since the creation of this business, people have questioned how odds are determined. The process beyond the creation of the odds is mainly based in statistics and predictions of what is more likely to happen. Having this in mind a research was developed, using real information made available by different online bookmarkers, to improve the real probability estimation for a given sportive match outcome. The authors used an Adaboost algorithm and constructed a virtual bookmarker following three main steps: first they've explored prospective weak classifiers, then they've drafted them in order to give a weight to their contribution in the end. For further work, the authors concluded that the implementation of a classification algorithm would improve the study, mainly because it would allow to consider as data for the scouting and drafting process both the final results and their statistical confidence (Torre & Malfanti, n.d.).

Another major concern of the online betting companies is to identify and prevent fraudulent attacks. In this area, it's possible to identify several types of crimes that can be committed, two of the most frequent being: match fixing that happens when the fraudster manipulates the result of an event by buying the referee, the coach or even the players; and financial fraud, that can include identity theft, duplication of accounts, money laundering and others. The last Gaming and Gambling cybercrime report made available by the ThreatMetrix in 2017, identified that one in every 20 new accounts and one in every 23 payment transactions is fraudulent (Digital & Network, 2017). Having in mind that fraudulent attacks are directly related to large losses of revenue, data mining solutions are starting to be employed in order to diminish the number of fraud cases in this industry. Until now, the research made on this subject is mainly related with unsupervised learning techniques such as clustering, anomaly and outlier detection.

In what concerns outlier detection and online gambling, in 2008, Manikas presented a model where the main purpose was to detect fraudulent behaviours and addictive gambling by finding and using an optimal method. The database used in this study contains several transactions derived from different online gambling platforms. After doing a detailed explanation of the anomaly detection method, the author chooses to apply a Principal Component Analysis (PCA) base model to the dataset. Having in consideration the dimension of the data and the number of variables existing in the studied dataset, by following this approach the author was able to transform the data into new compressed axes and at the same time keep their variability. The next step was to divide the new data into two different subspaces: normal and anomalous. In order to detect different behaviours in the data, an anomaly threshold was defined using the Q-static method. In order to evaluate the method, two data analyses were considered: time-vertical analysis, in which the datasets capture the activity of all the users within a specific amount of time, and user-based analysis, in which each user is tested for any suspicious behaviour. Both these methods were scanned for outliers. The results of the first analysis were after compared to the results given by the company's traditional method of fraud detection and the outliers encountered in the second analysis were also evaluated. Although the results of the work could not state that the method presented was the most efficient when detecting fraudulent behaviours in this specific type of data, they could prove to be better than the current method applied into the database studied (Manikas, 2008). In 2015, a project was developed in which the ultimate scope was to detect and dissuade money laundering in the gambling industry. The author started by creating, selecting and grouping money laundering indicators by applying statistics into the dataset. In a second phase, he created a profile of the gambling venues and of the beneficiaries in order to better identify anomalies. The profiles were created using pre-defined groups where a mean clustering algorithm was applied in order to identify clusters that are at high risk level. The final phase included the evaluation of the created process by testing potential fraud cases. The precision of the model was settled on 89% and the accuracy on 93% by the test set. The potential cases that registered an anomaly score higher or equal to 90% were considered as fraud. To conclude, the author identified triggers like frequency of gain, amount of small gains, number of gambling venues visited by the player and number of gains as high crucial variables. Outlier detection joined with visualization techniques proved to be extremely suitable for detecting anomalies in the data (Robert, 2015).

Besides the papers presented above concerning data mining and online gambling, in the last years online enterprises specialized only in detecting and preventing fraud and addictive gaming were created. The main goal of this companies is to help online gambling platforms in diminishing the number of fraudulent cases and at the same time reduce the profit losses caused by this matter. Their

techniques of fraud detection are not made available to the public but have as base data mining and machine learning approaches. [1]

Considering that there is not much available literature on fraud detection in online gambling and also that it is mainly concentrated in unsupervised learning algorithms, the project proposed in this report would be a way of starting to compensate the gap that exists in this concern.

[1] https://www.threatmetrix.com/cyber-security-solutions/gaming/
http://www.fraxses.com/applications/

4. METHODOLOGY

Data Mining (DM) analysis is usually conducted by a general process. There are different standard methodologies made available by this approach, three of which are used more frequently: Knowledge Discovery in Databases (KDD), Sample, Explore Modify Model and Acess (SEMMA) and Cross Industry Standard Process for DM (CRISP-DM). Comparing them it is possible to notice that SEMMA and CRISP-DM can be seen as an implementation of KDD (Azevedo & Santos, 2008).

Understanding that in order to achieve the optimal output the DM projects need an ordered structure, in this project we used SEMMA bases to create the predicting model. This methodology was developed by SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model and Assess. The model presented in this paper will mainly be guided by these different steps and will be developed using SAS Enterprise Miner tool. SAS Enterprise miner is a program that uses data mining process to create highly accurate predictive and descriptive models aggregating big amounts of data from across an enterprise.

In order to obtain a data mining process using SAS Enterprise Miner, first you have to create a process flow diagram that is built by dragging nodes from a toolbar, organized by SEMMA categories, and dropping them into a diagram workplace.

Among others, SAS Enterprise Miner is a powerful tool when performing data management, data analysis and reporting. The main benefits of using this program are: it supports the entire data mining process having available a large set of tools; it builds models in a faster way having an easy-to-use approach; the innovative algorithms improve the stability and accuracy of predictions; and it allows the analysts to promote business information and efficiently share results using a single easy-to-interpret framework (SAS, 2016).

4.1. BUSINESS INTRODUCTION

In order to describe the methodology that will be used in the project, first we need to clarify the different variables available in an online betting market. This is a particular business that has its own key performance indicators (KPIs). We will focus on explaining the ones that collect the most important information of the company's business ("Understand your Online Gambling Business with Key Performance Indicators - EveryMatrix," n.d.).

- **Turnover** (TO): describes the overall amount wagered.
- **Gross Gambling Revenue** (GGR) = $TO - Wins\ returned\ to\ players$. It measures the true economic value of gambling.
- **Net Gambling Revenue** (NGR) = $GGR - Tax - Bonus\ Costs$. It represents the true economic value of gambling discounting Taxes and Bonus supported by the company.
- **Margin =** $\frac{GGR}{TO}$**.** Values above zero mean that the company is profiting. Values below zero mean that the players are winning.
- **Average Revenue per User** (ARPU) = $\frac{GGR}{Nº\ Players}$**.** It characterizes the average revenue generated per user in a period of time.
- **Churn Rate**: refers to the proportion of customers that did not place a bet on a certain period of time, they abandoned the website.

These are some of the most important metrics when analyzing a betting environment. To build the predictive model, we will take into consideration both these variables together with others such as: number of deposits/withdraws; amount of deposits/withdraws; total of bets settled; number of registrations, etc.

There are already certain approaches used to perceive if a customer is a risk customer. Using the variables mentioned above we can identify different situations which should then be analyzed with attention and precaution:

- If usually the *Avg. Bet*[2] of a customer is below 10€ and suddenly grows to 1,000€;
- If the volume of transactions, in lifetime, is higher than €4,500 and lower than €5,000 per player;
- When there is an amount of single deposit higher than €5,000 per player;
- If a customer frequently deposits big amounts of money and hardly ever bets it: suspected money laundering;
- When only one customer deposits with different accounts or with different credit cards;
- If there are many large bets in an event that is not much publicized: it tends to happen more in lower leagues;
- If there is an event where there are no bets settled on the day before and two or three hours before the game a large number of bets is settled: this can show that the players were waiting to know the formation of the game;
- If the player doesn't collaborate on completing identity details and financial details.

In order to reach the final goal of creating a Predictive Model for Fraud in Portuguese Online betting different steps were settled:

1. Gather the available data and organize fraud and non-fraud customers;
2. Study the different available literature and understand more about machine learning, data mining algorithms and fraud;
3. Select the best approach to construct the model: supervised or non-supervised models;
4. Choose the existing variables in the business that contribute the most to improve the model and use them to create and transform new variables;
5. Make a 2nd overview to the model:
   5.1 Analyze the descriptive statistics and the visual tools;
   5.2 Decide how to do the partition of the data into: Train, Validate and Test;
   5.3 Realize if all the chosen variables are necessary to the model looking to their variable worth and understanding the correlations among them;
6. Test different data mining predicting algorithms;
7. Make a model comparison and conclude on which is the most accurate model;
8. Evaluate the results using the validation set and implementing the model to new available data.

[2] Avg. Bet = TO/Nº Bets

4.2. SAMPLE

In order to be able to work with the data, first we needed to organize and understand the database structure. The main goal was to identify the data that would be necessary to build the predicting model, verify its quality following imposed standards, and obtain the variables that would be relevant for all the process.

4.2.1. Data

The project database is composed by different data sources and each of them contains different information about the business. The most important ones are described in the table below (Table 1). Customer data source contains all the characteristics of the customer such as: gender; date of birth; first bet date (casino and sports); customer status (active – accounts that are able to bet; frozen – accounts closed for a particular reason; self-excluded – customers who asked not to be able to access the account again); etc. The remaining data sources contain all the customer transaction information such as: number and value of deposits or withdrawals; payment methods; turnover; odds and stakes; game type; etc.

Table 1 - Database

Data Sources	*Description*	*Examples KPI's*
Customer	Main customer information	Age; Gender; Customer Status; Date of registration; (…)
Customer Lifetime	Lifetime KPI's	Last bet date; Balance; Total nº deposits; Total TO; (…)
Sports Market	Main information about Sport Events	League; Event Description; Single/Combo; Branch; (…)
Bet Transactions	All bet changes - Sports and Casino	Open bet date; Game Type; Event; Turnover; (…)
Wallet Transactions	Wallet operations	Operation type; Transaction status; Payment method; Mobile/Web operation; (…)

Having in consideration that each customer can generate millions of billions of transactions and that each of these transactions represent a line on the database, it was necessary to transform the attributes into aggregate, average or temporal variables. The main core was to reconstruct the database in order to obtain only one line for each customer containing all the relevant information.

The database used to build the predictive model contained both earlier identified as fraud customers and regular customers. The first phase of the data preparation involved gathering all the customers that were already identified as fraudsters by the studied enterprise. The company's policy in situations of suspected fraud is to freeze the customer's account, investigate and understand if the suspicions are true or false. When the conclusions are false, it reactivates the customer's account and when they are true, depending on the level of gravity, it keeps the account frozen, contacts the person under investigation and, in last case, reaches the authorities.

The database that was after imported to SAS Enterprise Miner had 429 customers targeted as "Fraud" and 532 customers targeted as "No Fraud". The frozen customers report made available by

the studied company was the base used to collect the fraudulent accounts. Having in consideration that not all the frozen reasons are related to fraud, the report was first cleaned and reorganized so that, in the end, we would only have fraud frozen accounts. The no fraud customers base was randomly selected from the business database, filtering only active customers that already had activity on the platform. In order to differentiate both types of customers, we defined a variable *"Fraud"* which only took the values 0 and 1: 0 for non-fraud customers and 1 for fraud customers. This variable will be used as the target variable, dependent variable, in the model. Taking into consideration that we're developing a supervised learning model, where the main goal is to predict correctly as many observations as possible, the pre-defined target variable contributes to better measures and assessments of the model quality (Dean, 2014).

4.2.2. Variables

The project database is administrated and developed in SQL Management Studio 17 for SQL Server (Database & Solutions, 2015). Having this in consideration, in order to collect all the relevant information about the selected customers, distinct queries were created using the different data sources explained above (Table 1). Considering that we could only have one line for each customer, all the information was transformed in either total, average or maximum/minimum variables.

In order to choose the relevant variables used in the model, different aspects such as wallet operations, betting transactions and customer activity were taken into account. Both the Financial/Validation and Compliance teams were consulted in order to agree in which variables we should use as indicators of fraud behavior. These variables could be grouped into three main categories: demographic indicators; temporal indicators and business indicators (Table 2).

Table 2 - Variables

Customer Indicadores			
Variable	*Description*	*Role*	*Level*
Customer ID	ID User	ID	Nominal
Country	Country of Registration	Reject	Nominal
Gender	Male (M) or Female (F)	Input	Binary
Age	(Current Year - Year of Birth)	Input	Interval
Sports/Casino?	TO Sports > TO Casino - Sports TO Casino > TO Sports – Casino TO Total = 0 – Never Played	Input	Nominal
+ TO Branch	Branch with plus TO Sports	Input	Nominal
+ TO League	League with plus TO Sports	Input	Nominal
+ TO Game	Casino game with plus TO Casino	Input	Nominal
Frozen Date	Account close date	Input	Interval
First Deposit Date	Date of first Deposit	Input	Interval
Level	Low/High or No Fraud	Input	Nominal
Fraud	No Fraud = 0 ; Fraud = 1	Target	Binary
Temporal Indicators			
Variable	*Description*	*Role*	*Level*
Nº days without activity	(Date of Registration – 1st Bet Date)	Input	Interval
Nº Days w/activity	Count of Distinct Bet Dates	Input	Interval
Nº Weeks w/Activity	Count of Distinct Bet Weeks	Input	Interval
Nº Months w/Activity	Count of Distinct Bet Months	Input	Interval

Business Indicators			
Variable	*Description*	*Role*	*Level*
TO Sports	TO with Sports betting	Input	Interval
Total Nº Bets Sports	Count Nº Bets Sports	Input	Interval
Average Bet Sports	(TO Sports/Total Nº Bets Sports)	Input	Interval
GGR Sports	GGR with Sports betting	Input	Interval
Average GGR Sports	(GGR Sports/Total Nº Bets Sports)	Input	Interval
% TO Single	(TO Single Bets/TO Sports)	Input	Interval
% TO Combo	(TO Combo Bets/TO Sports)	Input	Interval
TO Low Leagues[3]	Total TO Low Leagues – Portugal Overview	Input	Interval
Nº Bets Low Leagues	Count Nº Low Leagues - Portugal Overview	Input	Interval
Avg Bet Low Leagues	(TO Low Leagues/Nº Bets Low Leagues)	Input	Interval
GGR Low Leagues	Total GGR Low Leagues - Portugal	Input	Interval
Avg GGR Low Leagues	(GGR Low Leagues/Nº Bets Low Leagues)	Input	Interval
TO Casino	TO with Casino betting	Input	Interval
Nr Bets Casino	Count Nº Bets Casino	Input	Interval
Average Bet Casino	(TO Casino/Total Nº Bets Casino)	Input	Interval
GGR Casino	GGR with Casino betting	Input	Interval
Average GGR Casino	(GGR Casino/Total Nº Bets Casino)	Input	Interval
Nr Distinct Payment Methods	Count Distinct Payment Methods	Input	Interval
Total Approved Deposits	Value of Total Deposits Approved	Input	Interval
Nº Deposits Approved	Nº Total Deposits Approved	Input	Interval
Average Nº Deposits per Day	(Nº Deposits Approved/Nº Days w/ activity)	Input	Interval
Average Deposits Approved	(Value Total Deposits Approved/Nº Total Deposits Approved)	Input	Interval
Total Rejected Deposits	Value of Total Deposits Rejected	Input	Interval
Nº Deposits Rejected	Nº Total Deposits Rejected	Input	Interval
Average Deposits Rejected	(Total Deposits Rejected /Nº Total Deposits Rejected)	Input	Interval
Total Pending Deposits	Value of Total Deposits Pending	Input	Interval
Nº Deposits Pending	Nº Total Deposits Pending	Input	Interval
Average Deposits Pending	(Total Deposits Pending /Nº Total Deposits Pending)	Input	Interval
Total Approved Withdrawals	Value of Total Withdrawals Approved	Input	Interval
Nº Withdrawals Approved	Nº Total Withdrawals Approved	Input	Interval
Average Nr Withdrawal per day	(Nº Withdrawals Approved/Nº Days w/ activity)	Input	Interval
Average Withdrawal Approved	(Total Withdrawals App./Nº Total Withdrawals App.)	Input	Interval
Total Rejected Withdrawals	Value of Total Withdrawals Rejected	Input	Interval
Nº Withdrawals Rejected	Nº Total Withdrawals Rejected	Input	Interval
Average Withdrawal Rejected	(Total Withdrawals Rejected/Nº Total Withdrawals Rejected)	Input	Interval
Total Pending Withdrawals	Value of Total Withdrawals Pending	Input	Interval
Nº Withdrawals Pending	Nº Total Withdrawals Pending	Input	Interval
Average Withdrawal Pending	(Total Withdrawals Pending/Nº Total Withdrawals Pending)	Input	Interval

After having all the database organized, we imported the data into the project and defined roles and levels for each of the selected variables. As referred above, we defined the metric '*Fraud*' as target variable, with a binary level where 1 represents fraud customers and 0 non-fraud customers. Besides

[3] Low Leagues – All the leagues of Portugal excluding 'Primeira Liga' of Soccer.

this, we designated *'Customer ID'* with the ID role and the remaining ones as Input variables. The levels were identified considering if the variables were binary, interval or nominal. We also decided to reject the variable '*Country*' because it only had different values for 4 customers. This difference was caused by a momentary error on the website, in which users could put their nationality on the country field instead of their residence country. In the end, we collected 54 variables, distributed as shown in the figure below (Table 3).

Table 3 – Variables Role and Level Distribution

Role	*Nº Variables*
ID	1
Reject	1
Input	51
Target	1

Level	*Nº Variables*
Nominal	7
Binary	2
Interval	45

4.3. EXPLORE

The second phase of the SEMMA methodology is directly related with making an exploratory data analysis (EDA). The main goal is to characterize the descriptive statistics of the data: identify missing values, determine outliers and understand the value of each variable. The original database is constituted by 51 input variables, where 44 are interval variables and 7 are class variables. Considering this, we analyzed each of these groups of variables separately.

Since we made a predictive modulation, with a supervised learning approach, the results of the target variable were considered for both analysis.

After running the *Stat Explore* node, we noticed that the variables '*Frozen Date'* and *'Level'* had a too high variable worth value, approximately 0.5, when compared to the other variables. This can be explained by the fact that both of them have a particular field for fraud customers and a unique value for non-fraud customers. For example, fraud customers can only have '*low'* and *'high'* as level and non-fraud customers can only have '*no fraud*' as level. Considering this, we decided to reject these variables, with the purpose of being able to increase the performance of the remaining ones.

4.3.1. Interval Variables

Analyzing the Interval variables output (Table 4), the first thing we can conclude is that there are no missing values in the dataset. Looking to the descriptive statistics, another conclusion that we can take is that there might be potential outliers in some variables. The distance between the mean and maximum values supports this conclusion. For example, the variable *"Total_Pending_Deposits"*, for fraud customers, has a mean value approximately 48x bigger than the maximum value and the variable "*Amt_Approved_Deposits*", for fraud customers, has a mean value approximately 201x greater than the maximum value. The presence of outliers in the dataset can be justified by the fact that we're dealing with potential fraudulent customers, who normally have an irregular behavior.

In addition to these conclusions, when analyzing the descriptive statistics, we can also start to understand some variables' behavior with the target variable. The mean value of both casino and sports turnover for fraud customers is approximately 6x greater than the value registered for non-fraud customers. This may show that clients with fraudulent behavior tend to bet higher values than non-fraudulent customers. Besides these, the "Low Leagues" variables behavior also caught our

attention. For fraud customers, the Turnover and the Average Bet registered a maximum value higher than normal and distanced from the mean value. Knowing that we're considering only small leagues, which by themselves are not very appealing to bets, the fact that these values are so high reveal already a certain peculiar behavior.

Table 4 – Statistics of Interval Variables

Target	*Interval Variabels*	*Missing*	*Mean*	*Standard Deviation*	*Median*
0	Avg_Dep__Rejected	0	19	263	5
1	Avg_Dep__Rejected	0	62	378	10
0	Total_Pending_Deposits	0	61	1 947	5
1	Total_Pending_Deposits	0	386	2 893	15
0	Amt_Approved_Deposits	0	473	21 547	55
1	Amt_Approved_Deposits	0	**3 142**	32 139	111
0	Avg_Nr_Dep_per_Day	0	0	1	0
1	Avg_Nr_Dep_per_Day	0	1	1	0
0	Total_Rejected_WD	0	46	3 623	0
1	Total_Rejected_WD	0	446	5 413	0
0	Avg_Nr_WD_per_day	0	0	0	0
1	Avg_Nr_WD_per_day	0	0	0	0
0	TO_Casino	0	312	15 185	0
1	TO_Casino	0	**3 452**	22 270	0
0	Nr_Bets_Casino	0	445	17 687	0
1	Nr_Bets_Casino	0	2 999	26 246	0
0	TO_Sports	0	1 637	76 169	175
1	TO_Sports	0	**8 307**	113 715	209
0	Avg_Bet_Sports	0	7	342	3
1	Avg_Bet_Sports	0	**49**	510	3
0	TO_Low_Leagues	0	64	1 064	5
1	TO_Low_Leagues	0	148	1 568	5
0	Avg_Bet_Low_Leagues	0	4	45	1
1	Avg_Bet_Low_Leagues	0	7	63	1
0	Nr_days_without_activity	0	19	1 958	1
1	Nr_days_without_activity	0	-143	2 929	2
0	Age	0	32	10	30
1	Age	0	32	12	29

Lastly, the variable "*Nr_Days_without_activity*" has a minimum value of *-42,848*. Considering that this variable represents the number of days that passed between the customer registration and the customer first activity on the website, it cannot have negative values and the minimum value should be 0. The customers that registered this value are customers that never had activity on the website. Knowing this, we decided to replace the *-42,848* values by 0 using the SAS Code represented below.

```
Training Code
PROC SQL;
UPDATE &EM_IMPORT_DATA
SET Nr_days_without_activity = 0
Where Nr_days_without_activity<0;
Quit;
```

Fig. 5 – SAS Code Transformation

4.3.2. Class Variables

In what concerns the class variables considered, the first thing to point out is that there are missing values on the variables: "*Favorite_Branch*"; "*Favorite_GameType"* and "*Favorite_League*" (Fig. 7). There are different forms of handling missing data: ignore the data row; replace the missing values with a global constant; replace the missing values with the mean value or median value of the variable or use an algorithm to predict the most probable value (Kaiser, 2014).

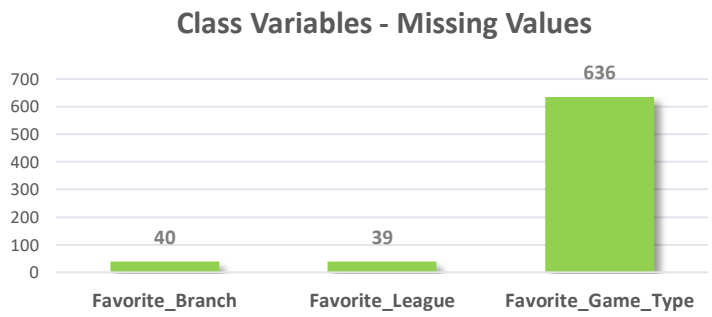


Fig. 6 – Class Variables - Missing Values

Besides this discovery, when observing the class variables' output we could also draw some conclusions through the analysis of the relative frequencies of the variables.

- *Gender*: the majority of the dataset is constituted by Male customers. For both fraud and non-fraud customers this gender represents 83% and 89% respectively (Fig. 8).

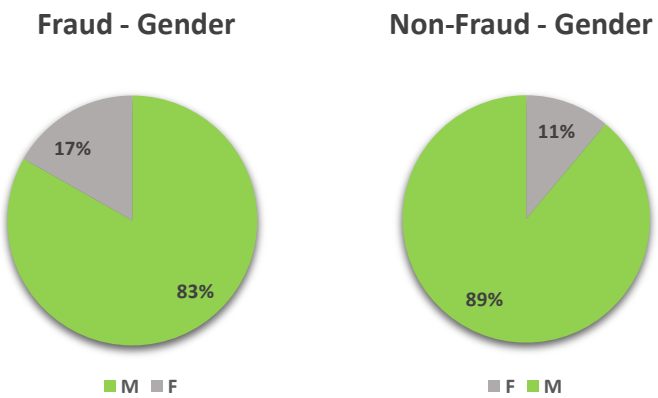


Fig. 7 – *Gender* Frequency Count

- *Sports or Casino*: this variable represents the product where each customer generates more Turnover. For both fraud and non-fraud customers Sports is the product where they place more bets, 77% and 93% respectively. Note that, besides Sports and Casino, for non-fraud customers there is also 18% of customers that never played either of the products. These customers only tried or did deposits/withdrawals which can give information about potential AML situations (Fig. 9).

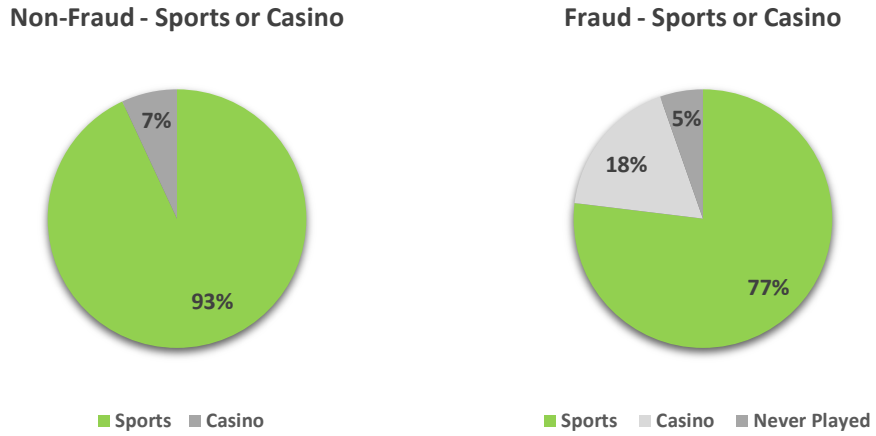


Fig. 8 – *Sports or Casino* Frequency Count

In addition to the analysis of the relative frequencies of the variables, we also analyzed the qui-square test results given for class variables (Table 5). This test shows the association between each class variable and the target variable. Having in consideration that the p-value for all the variables, except *"Favorite_League",* is lower than 5%, we can conclude that these variables will be important when estimating the model. The variable *"Favorite_League"* registered a p-value of 5,93%, which, for being so close to 5%, we'll consider as also contributing positively to the model.

Table 5 – Class Variables - Chi-Square Test

Class Variables	*Chi-Square*	*P-value*
Favorite_League	123,0	0,059
Sports_or_Casino	59,1	<,0001
Favorite_Branch	21,4	0,001
Favorite_Game_Type	11,8	0,008
Gender	6,5	0,011

4.3.3. Data Partition

The data partition is one of the essential steps when developing a predictive model, it assures the quality and efficiency of the model and also prevents overfitting. Overfitting occurs when the learning algorithms adjust too much to the training data and start memorizing it instead of learning all the particularities of it (Khan, 2014). This problem can lead to estimates that wrongly predict unknown data.

In data mining, the samples can be split into three different sets:

- Training set: Sample of data used to train and adjust the model. The bigger, the better the classifier.
- Validation set: Sample of data that controls the training process and the error. The bigger, the better the estimate of the optimal training.
- Test set: Sample of data used to estimate the quality of the final model when predicting unknown data. The bigger, the better the estimate of the algorithm's performance on new data.

By now, there is no optimal data partition defined, it can vary with the sample dimension or the classifier algorithm (Dobbin & Simon, 2011). Considering that our database has approximately 1,000 observations, the partition approach used was 70% for training, 30% for validation and 0% for test. With this splitting, we were able to keep more observations to train the model. The training set had 670 customers, 371 non-frauds and 299 frauds (Fig. 10). The validation set had 291 customers, 161 non-frauds and 130 frauds (Fig. 11).

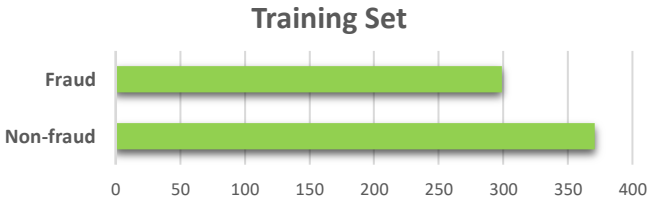


Fig. 9 – Training Set

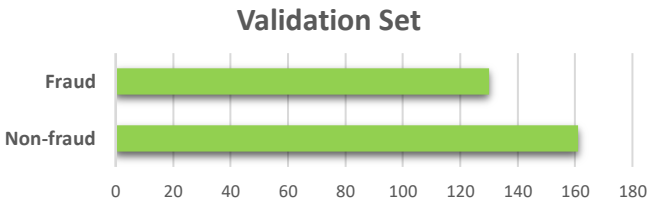


Fig. 10 – Validation Set

4.4. MODIFY

The modify phase exists in order to correct the gaps discovered during the exploration phase. It is the last phase of data preparation before modeling. In this phase, we'll transform, exclude and identify the most valuable variables depending on their specific objective for the study.

4.4.1. Missing Values

Missing values are considered a problem on the dataset, considering that mainly all the standard statistical methods assume that all the analyzed variables have complete information (Soley-bori, 2013). Taking this into account, the presence of absent observations on some variables can reduce the statistical power of the study, decrease the quality of the results, produce biased estimates and lead to baseless conclusions (Kang, 2013).

There are different forms of handling missing data. Two common and similar approaches are the Listwise deletion and the Pairwise deletion. The Listwise deletion method consists in deleting the entire record where there are missing values. The biggest disadvantage of this technique is that, if a big part of the sample has missing values, when excluding it we may be also excluding valuable information that could increase the model's power of prediction. In a less radical perspective, the Pairwise deletion only deletes the values where there is missingness resulting in a lower loss of information. Despite this, there is still a risk of losing precision and induce bias (Vaishnav & Patel, 2015).

Besides these techniques, another way of treating missing values is using imputation methods. These methods consist in substituting each missing value for a reasonable value that is estimated using different approaches. Among others, replacement by the mean or the median, by a constant, using regression approach or an algorithm are some of the most common forms of imputation (Brown & Kros, 2003).

In order to better handle the missing data problem, first of all it is necessary to completely understand the variables and the reasons of missingness. The techniques to apply to treat the missing data will vary with the reasons why data is missing (Soley-bori, 2013).

During the exploration phase, it was already discovered that we've missing values on three class variables of our database: "*Favorite_Branch*"; "*Favorite_League*"; "*Favorite_GameType"* (Fig.7). The reason of this missingness is directly related with the fact that there are some customers that never betted on Sports and, as such, do not have a favorite league or branch. Also, there are some customers that never played casino and, following the logic, do not have a favorite game type. Considering that the missingness on the two first variables represents only 4% of the whole database and on the last one represents 66%, we'll handle both cases differently:

- "*Favorite_GameType":* having in consideration that more than half of the data is missing in this variable and that the variable worth output shows that this variable adds low value to the model, we decided to drop it.
- "*Favorite_Branch*" and "*Favorite_League*": considering that these variables registered good results on the exploration phase and the percentage of missing data is lower than 5%, we decided to replace the missing data by a constant. Knowing the reason of missingness, we replaced all missing data by: *"NoSports"*. Besides this, we noticed while replacing the values that both these variables had 1 observation tagged as 0. Being this a gap on the dataset, we also replaced these values by the "*NoSports"* constant.

4.4.2. Outliers

Outliers can be described as data points that are significantly different from the remaining data, they lie outside the normal region of interest of the input space. Usually, they are also named as abnormalities, deviants or anomalies (Aggarwal, 2013). Understanding the outliers' behavior is considered as essential on a statistical analysis because they can either affect negatively the whole results of the analysis or their behavior can be exactly what the model is looking for.

There are different outlier detection methods that can be divided into parametric and non-parametric methods. The first ones imply that there is a prior knowledge of the underlying data distribution and, on the contrary, the second ones are considered *model-free* (Ben-gal, 2005). Non-parametric methods include data mining methods that are usually based on local distance measures such as clustering techniques. The great advantage of these methods when compared with parametric methods is that they are able to be applied into large datasets.

During the exploration phase, we already identified that some interval variables could have outliers. The presence of these anomalies in the dataset can be justified by the fact that we're dealing with fraudulent customers, who normally have an irregular behavior. Knowing this, we decided to test both *keeping outliers and removing outliers'* approaches excluding only outlier values that were marked as non-fraud customers when building the model. The main goal was not to reject valuable information that fraud outlier observations could bring to the model.

By analyzing the *MultiPlot* node, we detected outliers on twelve different interval variables and excluded them using the manual method – "User Specified" - by establishing new maximum and minimum limits to each variable (Fig.11, Fig.12 e Fig.13).

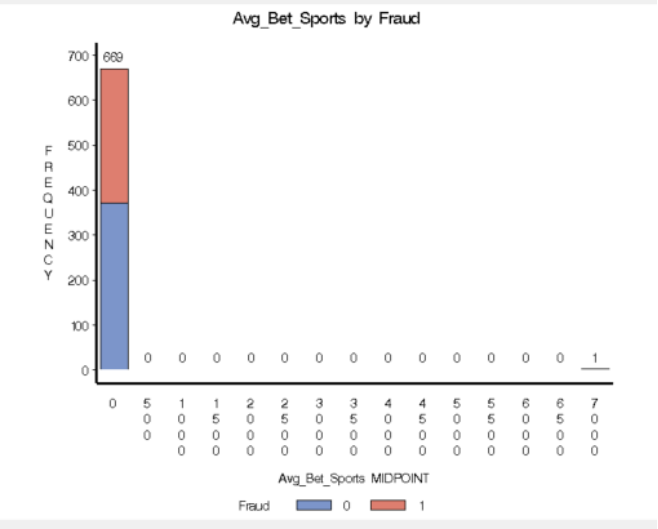


Fig. 11 – MultiPlot Output – *Avg_Bet_Sports*

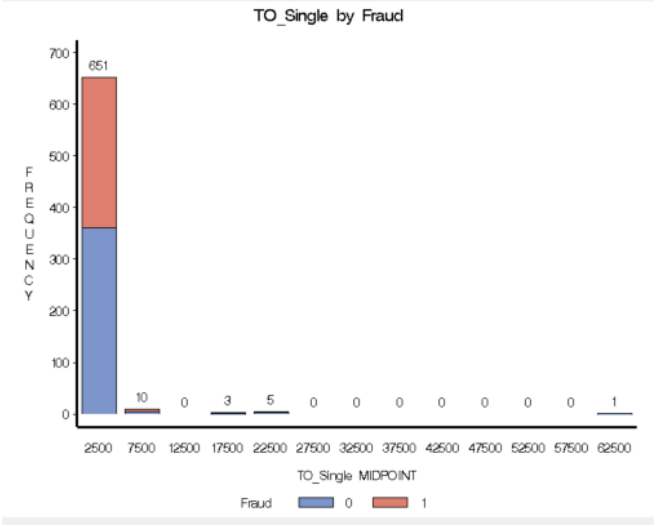


Fig. 12 - MultiPlot Output – *TO_Single*

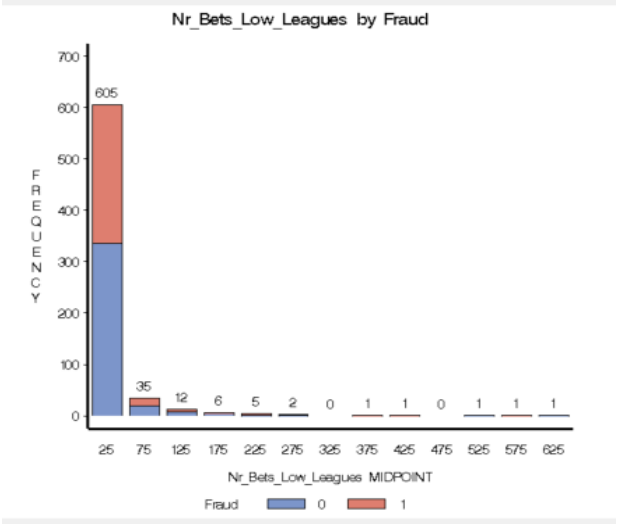


Fig. 13 – MultiPlot Output – *Nr_Bets_Low_leagues*

In the end, 20 observations were excluded from the analysis, representing **2,1%** of the total database. This number of filtered records is adequate in order to avoid bias when building a predictive model (Table 6).

Table 6 – Outliers Filtered Variables

Variable	*Minimum*	*Maximum*
Avg_Bet_Sports	-1	3133
Avg_Dep__Approved	-2	4955
Avg_Dep__Pending	-2	859
Avg_GGR_Casino	-27	10
Avg_GGR_Low_Leagues	-75	112
Avg_GGR_Sports	-36	1108
Nr_Bets_Low_Leagues	-3	574
Nr_Bets_Sports	-34	14902
Nr_Deposits_Approved	-4	810
Nr_Deposits_Rejected	-1	240
TO_Single	-61	12671
TO_Sports	-193	1012923

4.4.3. Dimensionality Reduction

Dimensionality reduction is the study of methods able to reduce the number of attributes describing the model. The main goal is to keep in the sample only variables that are relevant and not redundant, in order to increase the quality of the data and the power of prediction of the data mining model (Dash, n.d.). It is an effective solution to the "Curse of Dimensionality". Bellman (1961) used this expression to state the fact that the sample size needed to estimate a function of several variables to a given degree of accuracy increases exponentially with the number of variables (Verleysen & François, 2005). As such, if the number of variables of input is too large when compared with the available data, it will be difficult to model in a precise way. Dimensionality reduction methods allow us to reduce the complexity of the problem by choosing only variables that add value to its resolution.

The two main concepts that are important to retain are: redundancy and relevance. Relevant variables are considered the ones that have more capacity to describe the target variable. They are the most important ones considering that they are the ones that add more value to the model. Redundant variables are variables that are correlated, explaining the target variable in a similar way and bringing almost the same information to the model.

4.4.3.1. Variable Worth and Correlation Matrix

- Keeping Outliers

In order to identify the most relevant variables, we decided to use the StatExplore output *'Variable Worth'*. It ranks the variables by importance and worth (Fig. 12).

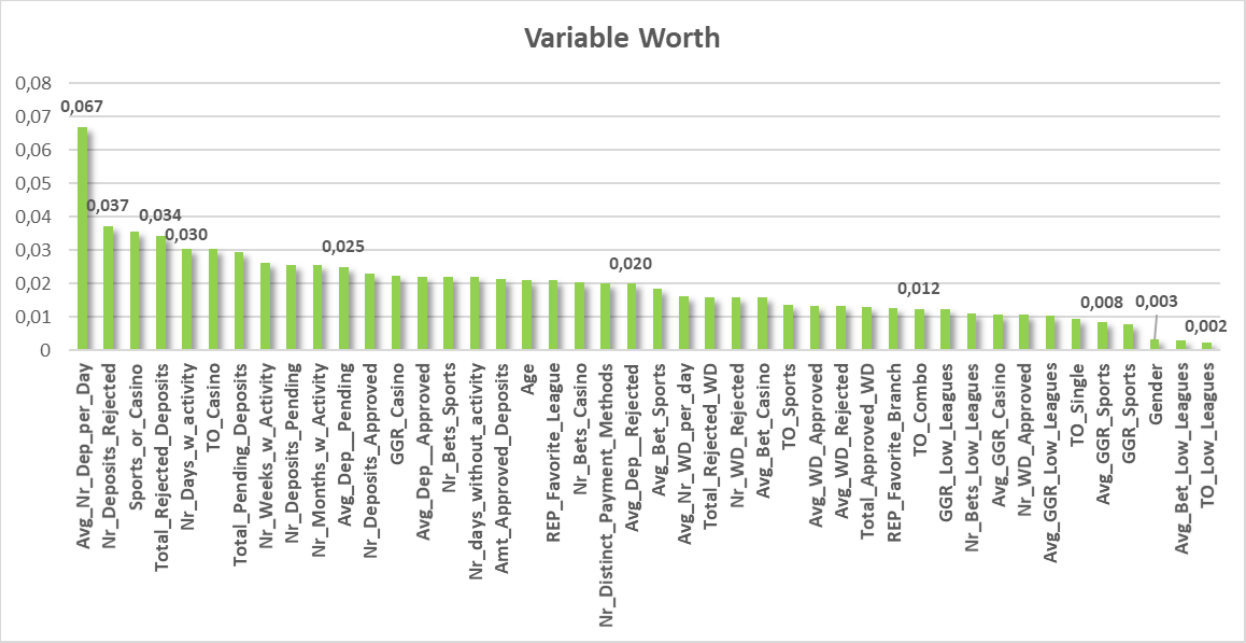


Fig. 14 – Keeping Outliers - Variable Worth

First of all, the variable worth output shows that the value of the variables ranges from 0.002 to 0.067. The variable *"Avg_Nr_Dep_per_Day"* can be considered the one that has the greatest discriminative power of the dependent variable, having registered the value 0.067. Besides this one, the variables *"Nr_Deposits_Rejected", "Sports_or_Casino", "Total_Rejected_Deposits"* and *"Nr_Days_w_activity"* also proved to give a good contribution to the problem resolution by registering a worth higher than 0.030. On the opposite side of the graph, we've *"TO_Low_Leagues"*, *"Avg_Bet_Low_Leagues", "Gender", "GGR_Sports", "Avg_GGR_Sports"* and *"TO_Single"*. All these variables registered a value lower than 0.01, demonstrating to be the less relevant variables and having a low performance when discriminating the dependent variable. Besides this, the variables *"Avg_WD_Pending"*, *"FD_Date"*, *"Nr_Withdrawls_Pending"* and *"Total_Pending_WD"* do not appear on the variable worth output, which can be justified by the fact that these variables do not bring any value to the model. As such, we decided to reject them from the modelling phase.

The redundancy between the variables was analyzed using the correlation matrix. The two most commonly used coefficients to validate the correlation between the variables are: Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient. The Pearson's coefficient holds on the assumptions that the association between the variables must be linear; each of them must have an interval or ratio level and must be normally distributed (Rebekić, Lončarić, Petrović, & Marić, 2015). Considering that our data does not meet all these requirements, we decided to use Spearman's correlation rank coefficient. This coefficient is a non-parametric measure that uses ranks to calculate the correlation between the variables. Instead of using a linear function to associate the

variables, this coefficient describes the relations using a monotonic function. A monotonic function is one that either never increases or never decreases as its independent variable increases (Fig. 13) (Statstutor, n.d.).

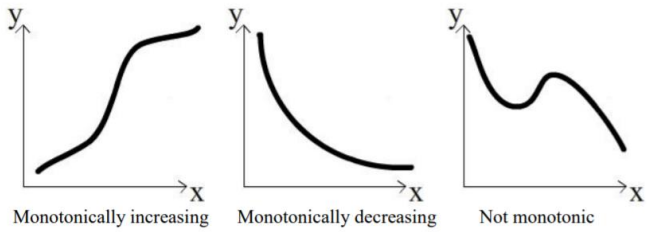


Fig. 15 – Monotonic Function

In order to obtain the Spearman Correlation output (Table 6), we used the SAS Code presented below (Fig.11):

Training Code

```
proc corr data=&EM_IMPORT_DATA.
   nosimple outs=&EM_LIB..&EM_NODEID._USER_CORR_SPEARMAN noprint;
      var %EM_INTERVAL;
run;

%EM_REGISTER(Key=SPEARMAN,Type=DATA);

DATA &EM_USER_SPEARMAN(drop=_Type_);
   Set &EM_LIB..&EM_NODEID._USER_CORR_SPEARMAN;
      If _Type_ ne 'CORR' then delete;
Run;

%EM_REPORT (Key=SPEARMAN, Viewtype=DATA, Autodisplay=Y, Block=Correlation,Description=SPEARMAN);
```

Fig. 16 – SAS Code Spearman Correlation

It was considered that there is redundancy between variables when the absolute correlation coefficient is greater than or equal to 0.8.

The choice of the most valuable variables implies that there is a joint analysis of the two methods presented above. In a first phase, the correlation matrix shows us which variables have a high relation between each other and, in a second phase, the Variable Worth output helps us decide which ones we should keep and which ones we should exclude from the modeling.

Observing the correlation matrix output, the first conclusion we can take is that almost all the variables have a correlation higher than 0.8 with at least one other variable (Annex 9.6). Having this in consideration, we began to analyze the correlations of the variables by those who had more importance on the variable worth output with the objective of rejecting first the less valuable variables. The variable *"Avg_Nr_Dep_per_Day"* had no significant relation with any other variable so we concluded that this variable would be used when modelling. The variable *"Nr_Deposits_Rejected"*

registered a relevant relation with both *"Total_Rejected_Deposits"* (0.97) *and "Avg_Dep__Rejected" (0.82).* Considering that this variable registered a higher worth than the two variables with which it is correlated, we kept the variable *"Nr_Deposits_Rejected"* for modelling and rejected both *"Total_Rejected_Deposits" and "Avg_Dep__Rejected".* Next on the variable worth output comes the variable "Nr_Days_w_activity". This variable showed to have high correlation with six different other variables: *"Nr_Bets_Low_Leagues", "Nr_Bets_Sports", "Nr_Deposits_Approved", "Nr_Months_w_Activity"*, *"Nr_Weeks_w_Activity"* and *"TO_Sports"*. Considering that all these variables had less importance than *"Nr_Days_w_activity",* we decided to exclude them all from modelling. This process was repeated for all variables so that, in the end, we would have kept only those that, at the same time, were not related between each other and had a greater discriminative power of the target variable.

In what concerns class variables, by observing simultaneously the chi-square test results and the variable worth output after all the modifications, we decided to reject both *"REP_Favorite_League"* and *"Gender"* variables. The variable "REP_Favorite_League" recorded a 15% p-value in the chi-square test, which proves that this variable doesn't add any value when estimating the target variable. The rejection of the variable "Gender" can be justified by the fact that this variable was ranked as 42nd in terms of degree of importance, having only a "worth" of 0.003.

Table 7 – Keeping Outliers - Transformed Class Variables – Chi-Square Test

Class Variables	*Chi-Square*	*p-value*
REP_Favorite_League	100	0,150
Sports_or_Casino	48	<,0001
REP_Favorite_Branch	17	0,001
Gender	5	0,034

- Removing Outliers

The same assumptions were used considering now the dataset *removing outliers*. First, we analyzed the variable worth output in order to identify the most valuable variables.

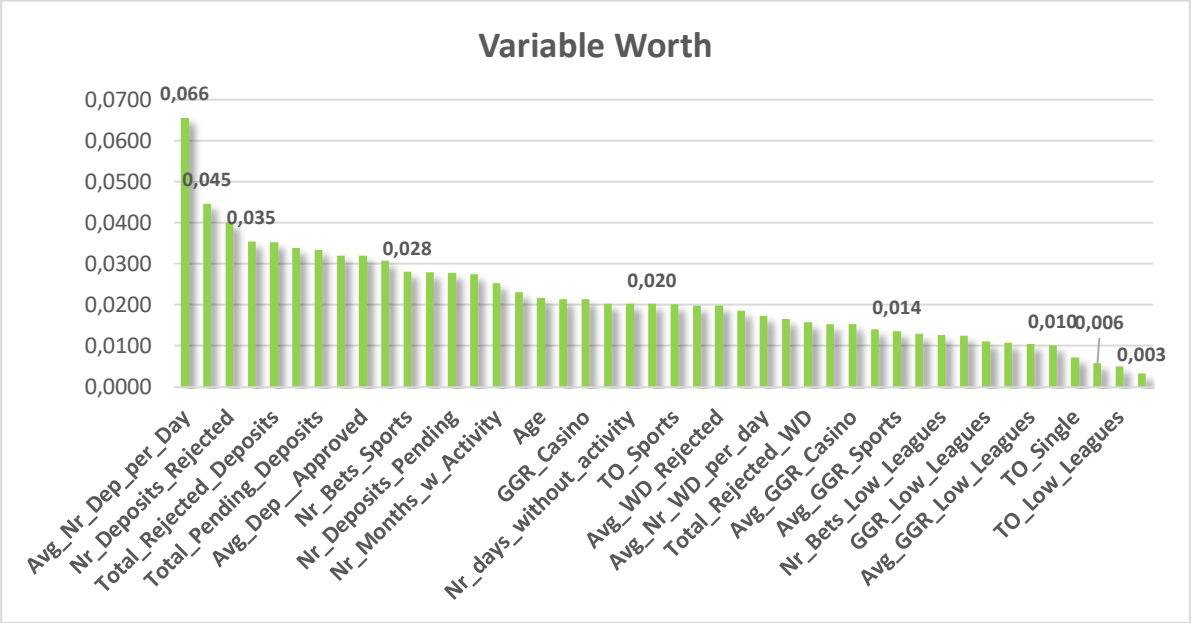


Fig. 17 – Removing Outliers - Variable Worth Output

When compared with the *with outliers approach,* the variable "*Nr_Deposits_Approved*" was the one that registered a higher difference of only 0,02151. It was ranked as 2nd in terms of degree of importance on the *removing outliers approach* and as 12th on the *keeping outliers approach*.

By analyzing the correlations between the variables together with the variable worth output, rejecting the ones that had a correlation higher than 0,8 and starting by the ones that registered higher importance, we decided to keep 28 variables (Table 9).

Table 8 – Removing ouliers - Joint Analysis Correlations and Variable Worth

Variable	*Worth*
Avg_Nr_Dep_per_Day	0,065574
Nr_Deposits_Approved	0,044609
Nr_Deposits_Rejected	0,040131
Sports_or_Casino	0,035381
Total_Pending_Deposits	0,033425
Avg_Dep__Approved	0,031903
TO_Casino	0,030679
Nr_Bets_Sports	0,028007
REP_Favorite_League	0,023049
Age	0,021668
Avg_Bet_Sports	0,021439
GGR_Casino	0,021404
Nr_Distinct_Payment_Methods	0,020322
Nr_days_without_activity	0,020275
Avg_WD_Rejected	0,019758
Avg_WD_Approved	0,018648
GGR_Sports	0,016488
Avg_GGR_Sports	0,013569
REP_Favorite_Branch	0,012477
GGR_Low_Leagues	0,011028
TO_Combo	0,010775
TO_Single	0,007208
Avg_Bet_Low_Leagues	0,005751
Gender	0,003247

As concluded on the *keeping outliers' approach,* by observing simultaneously the chi-square test results and the variable worth output after all the modifications, we decided to reject both *"REP_Favorite_League"* and *"Gender"* variables. The variable "REP_Favorite_League" recorded a 15% p-value in the chi-square test, which proves that this variable doesn't add any value when estimating the target variable. The rejection of the variable "Gender" can be justified by the fact that this variable was ranked last in terms of degree of importance, having only a "worth" of 0.0033.

During this phase, we concluded that we would keep 26 input variables on our model. In data mining, this total number of variables can be classified as too high and may impair the performance of the model. Having this in consideration, we decided to add the Variable Selection node to our dimensionality reduction analysis.

4.4.3.2. Variable Selection

- Keeping Outliers

The joint analysis of the correlation matrix and the variable worth output led us to conclude on keeping 21 input variables: 2 nominal and 19 interval. In order to validate this decision and, at the same time, to make a final dimension reduction of the database, we decided to apply the Variable Selection Node to our variables. The output of this node not only allows us to conclude on which variables to keep and on which to exclude but also illustrates the reason of rejection of each of the variables. In our case, 34 variables were rejected and the reason was *Small R-Square* (Table 8).

Table 9 – Keeping Outliers - Variable Selection Output

Variable Name	*Role*	*Level*	*Reasons for Rejection*
Avg_Nr_Dep_per_Day	Input	Interval	
G_REP_Favorite_Branch	Input	Nominal	
G_REP_Favorite_League	Input	Nominal	
Gender	Input	Binary	
Nr_Bets_Casino	Input	Interval	
Nr_Deposits_Approved	Input	Interval	
Nr_Deposits_Pending	Input	Interval	
Nr_Deposits_Rejected	Input	Interval	
Nr_WD_Rejected	Input	Interval	
Nr_Withdrawls_Pending	Input	Interval	
Nr_days_without_activity	Input	Interval	
Sports_or_Casino	Input	Nominal	
TO_Casino	Input	Interval	
Total_Pending_Deposits	Input	Interval	
Total_Rejected_Deposits	Input	Interval	
Age	Rejected	Interval	Varsel:Small R-square value
Amt_Approved_Deposits	Rejected	Interval	Varsel:Small R-square value
Avg_Bet_Casino	Rejected	Interval	Varsel:Small R-square value
Avg_Bet_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
Avg_Bet_Sports	Rejected	Interval	Varsel:Small R-square value
Avg_Dep__Approved	Rejected	Interval	Varsel:Small R-square value
Avg_Dep__Pending	Rejected	Interval	Varsel:Small R-square value
Avg_Dep__Rejected	Rejected	Interval	Varsel:Small R-square value
Avg_GGR_Casino	Rejected	Interval	Varsel:Small R-square value
Avg_GGR_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
Avg_GGR_Sports	Rejected	Interval	Varsel:Small R-square value
Avg_Nr_WD_per_day	Rejected	Interval	Varsel:Small R-square value
Avg_WD_Approved	Rejected	Interval	Varsel:Small R-square value
Avg_WD_Pending	Rejected	Interval	Varsel:Small R-square value
Avg_WD_Rejected	Rejected	Interval	Varsel:Small R-square value
GGR_Casino	Rejected	Interval	Varsel:Small R-square value
GGR_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
GGR_Sports	Rejected	Interval	Varsel:Small R-square value
Nr_Bets_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
Nr_Bets_Sports	Rejected	Interval	Varsel:Small R-square value

Variable Name	*Role*	*Level*	*Reasons for Rejection*
Nr_Days_w_activity	Rejected	Interval	Varsel:Small R-square value
Nr_Distinct_Payment_Methods	Rejected	Interval	Varsel:Small R-square value
Nr_Months_w_Activity	Rejected	Interval	Varsel:Small R-square value
Nr_WD_Approved	Rejected	Interval	Varsel:Small R-square value
Nr_Weeks_w_Activity	Rejected	Interval	Varsel:Small R-square value
TO_Combo	Rejected	Interval	Varsel:Small R-square value
TO_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
TO_Single	Rejected	Interval	Varsel:Small R-square value
TO_Sports	Rejected	Interval	Varsel:Small R-square value
Total_Approved_WD	Rejected	Interval	Varsel:Small R-square value
Total_Pending_WD	Rejected	Interval	Varsel:Small R-square value
Total_Rejected_WD	Rejected	Interval	Varsel:Small R-square value
REP_Favorite_Branch	Rejected	Nominal	Varsel:Small R-square value, Group variable preferred
REP_Favorite_League	Rejected	Nominal	Varsel:Small R-square value, Group variable preferred

The comparison between the results given by this output and the ones we already reached by analyzing the correlation matrix and the variable worth output led us to exclude 7 variables from the initial selection: *Amt_Approved_Deposits; Avg_Bet_Low_Leagues; Avg_GGR_Sports; GGR_Low_Leagues; Avg_Bet_Sports; TO_Combo and TO_Single.* All these variables proved not to add a greater value to the model and, at the same time, they registered a small R^2 and a small level of importance.

- Removing Outliers

The same procedure described above was applied on the removing outliers model approach. By comparing the variable selection results with the ones already reached with the joint analysis of the correlation matrix and the variable worth output, we decided to exclude 9 variables from the model comparison phase (Table 11).

Table 10 – Removing Outliers – Rejected Variables

Variable	*Worth*	*Reason of Rejection*
Avg_GGR_Sports	0,0136	Varsel:Small R-square value
REP_Favorite_Branch	0,0125	Varsel:Small R-square value, Group variable preferred
GGR_Low_Leagues	0,0110	Varsel:Small R-square value
TO_Combo	0,0108	Varsel:Small R-square value
TO_Single	0,0072	Varsel:Small R-square value
Avg_Bet_Low_Leagues	0,0058	Varsel:Small R-square value
Avg_WD_Pending	0,0000	Varsel:Small R-square value
Nr_Withdrawls_Pending	0,0000	Varsel:Small R-square value
Total_Pending_WD	0,0000	Varsel:Small R-square value

These nine rejected variables, represented on the table above, proved not to bring a high value to the predictive power of the model, either because they had a low importance or because they registered a low R^2.

4.4.3.3. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a useful statistical technique that is commonly used as a data reduction method. Considering its capability of finding patterns in high dimension data, this technique was already applied in different fields such as face recognition and image compression (Smith, 2002). Among others, PCA has the great advantage of allowing the reduction of a large set of variables to a small set of variables without much loss of valuable information and variability of the model.

PCA is a multivariate technique that analyzes a large number of observations, which are described by several correlated quantitative dependent variables, and transforms them into a lower number of uncorrelated variables normally named principal components (Abdi & Williams, 2010). The first principal component explains as much of the variability in the data as possible, and each succeeding component explains as much of the remaining variability as possible.

Having the characteristics described above in consideration, we decided to test PCA in both *keeping and removing outliers model approach.* This decision was directly related with the fact that the initial dataset used in this model had 50 input variables and PCA proved to generate accurate results when in presence of a large dataset. More specifically, this approach was applied in 2008 on a model built to detect fraudulent behaviours and addictive gambling and showed to generate good results when compared with the tools used before (Manikas, 2008).

4.4.4. Metadata

- Keeping Outliers

After analyzing all the variables, correcting missing values and reducing the dimensionality, we used the node Metadata to change the role of the variables that we decided to exclude from the predictive modelling phase.

Table 11 – Keeping Outliers - Metadata

Variable	*Role*	*Level*
Fraud	TARGET	BINARY
CustomerID	ID	NOMINAL
Age	INPUT	INTERVAL
Avg_Dep__Approved	INPUT	INTERVAL
Avg_Nr_Dep_per_Day	INPUT	INTERVAL
Avg_Nr_WD_per_day	INPUT	INTERVAL
GGR_Casino	INPUT	INTERVAL
Nr_Days_w_activity	INPUT	INTERVAL
Nr_Deposits_Rejected	INPUT	INTERVAL
Nr_Distinct_Payment_Methods	INPUT	INTERVAL
Nr_WD_Rejected	INPUT	INTERVAL
Nr_days_without_activity	INPUT	INTERVAL
REP_Favorite_Branch	INPUT	NOMINAL
Sports_or_Casino	INPUT	NOMINAL
TO_Casino	INPUT	INTERVAL
Total_Pending_Deposits	INPUT	INTERVAL

In the end, we kept 14 input variables - 2 nominal and 12 interval - and rejected 36 variables. From this point on, we will only consider these 14 as the ones whose have the greatest discriminative capacity on the *keeping outliers model approach*.

- Removing Outliers

The metadata node was also used on the *removing outliers model approach* in order to apply all the decisions that were made during the dimensionality reduction phase. In this case, we concluded on keeping 16 variables – 1 nominal and 15 interval – and excluding 34 variables (Table 13).

Table 12 – Removing Outliers – Metadata

Variable	*Role*	*Level*
Fraud	Target	BINARY
CustomerID	ID	NOMINAL
Avg_Nr_Dep_per_Day	Input	INTERVAL
Nr_Deposits_Approved	Input	INTERVAL
Nr_Deposits_Rejected	Input	INTERVAL
Sports_or_Casino	Input	NOMINAL
Total_Pending_Deposits	Input	INTERVAL
Avg_Dep__Approved	Input	INTERVAL
TO_Casino	Input	INTERVAL
Nr_Bets_Sports	Input	INTERVAL
Age	Input	INTERVAL
Avg_Bet_Sports	Input	INTERVAL
GGR_Casino	Input	INTERVAL
Nr_Distinct_Payment_Methods	Input	INTERVAL
Nr_days_without_activity	Input	INTERVAL
Avg_WD_Rejected	Input	INTERVAL
Avg_WD_Approved	Input	INTERVAL
GGR_Sports	Input	INTERVAL

There are some differences on the chosen variables from one approach to the other: on the *with outliers approach* we kept the "*Avg_Nr_WD_per_day", "Nr_Days_w_activity", "Nr_WD_Rejected" and "REP_Favorite_Branch"* variables and, on the *without outliers approach,* we kept the *"Nr_Bets_Sports", "Avg_Bet_Sports", "Avg_WD_Rejected", "Avg_WD_Approved", "GGR_Sports" and "Nr_Deposits_Approved"* variables. Having these differences in consideration, it's important to remind that model building is an iterative process, meaning that no single method works best in all cases. Over the next steps, we tested different approaches in order to reach the algorithm that best solves the problem proposed in this report.

4.5. MODEL

The Model phase consists mainly on modelling the previous modified data in order to search automatically for relations between them and reliably predict a desired outcome (Azevedo & Santos, 2008). There are different techniques made available - such as Neural Networks, Regression, Decision Trees and K-means Clustering – and each one of them has particular advantages and performs in a certain way depending on the data mining situations. The final goal is to choose the model that, based on historical data, registers a high accuracy and is realistic when predicting future behaviors of the target variable.

Supervised and Unsupervised learning are the two approaches that can be used when modelling. In this work, we'll only consider supervised learning algorithms. These algorithms have as main goal to accurately predict a specific target value using a subset of data and variables for which this target value is already known.

It's important to point out that there is no simple or completely correct solution when constructing a predicting model. Different data mining algorithms can produce different outcomes and perform on different ways depending on the problem, on the data type and on the approach presented. It is an interactive process that implies that many different techniques are tested.

Having in consideration the literature review, in this study we decided to implement four different algorithms in the model: Logistic Regression, Neural Networks, Decision Trees and Ensemble technique.

4.5.1. Regression

Regression analysis is one of the most commonly used techniques in statistics when in presence of a numeric prediction. It models the relationship between one or more independent/predictor variables and the dependent/target variable. In other words, the regression algorithm estimates the value of the target variable as a function of the predictors for each case in the build data (S. Gupta, 2015). In regression techniques, the independent variable can either be continuous or binary. When in presence of a categorical independent variable with more than two values, in order to apply a regression first it's necessary to convert it to a dummy variable with only two levels (Allan T. Mense, 2017).

There are different regression methods that can be applied in modelling prediction. Two of the most frequently used ones are linear regression and logistic regression. The characteristics that differ them the most are the type of dependent variable of the model and the algorithm approach used.

The simple linear regression model is represented by the mathematical equation: $y = \alpha + \beta X$, where *y* represents the dependent variable, *X* the independent input variables, α the intercept and β the slope of the model (S. Gupta, 2015). One of the particularities of this method is that it requires the dependent variable to be a numeric value, continuous and have a normal distribution. Besides this, it presumes that there is a linear relation between the dependent continuous variable and the remaining independent variables. Least square estimation is the algorithm that is at the basis of the linear regression method. It states that the regression coefficients should be established with the purpose of minimizing the sum of the squared distances of each observed response to its fitted value (Bhalla, 2014).

Logistic Regression is a method that also has its particularities but can be considered as less restricted than the linear regression. In what concerns the target variable, this technique requires it to be dichotomous, having only two levels: 0 (no) or 1 (yes). Logistic regression is based on the Maximum Likelihood Estimation algorithm, which states that regression coefficients should be selected in order to maximize the probability of *y* given X (where y represents the dependent variable and X the independent variables). It can be recognized as an interactive process, considering that the software tests different solutions in order to get the maximum likelihood estimates (Bhalla, 2014).

As already mentioned before, Regression analysis, in particular Logistic Regression, proved to obtain a high level of accuracy when predicting financial fraud and business failure (Liou, 2008). Taking this into consideration, we decided to include this approach on our model comparison.

4.5.2. Neural Networks

Neural Networks (NN) is a powerful machine learning tool that nowadays has been applied in different business problems such as sales forecasting, data validation, fraud prediction and risk management. It is a technique that is inspired on the biological neural networks of the Human brain, composed by a large number of computational units named neurons that connect with each other using layers through an extensive and elaborated communication network. Among other advantages, NN does not impose any statistic restrictions on the independent input variables, has the ability to interpret and model non-linear and complex relationships and also to generalize and predict with high level of accuracy on unseen data.

Over the past years, the number of different types of NN has grown exponentially. Multilayer Perceptron (MLP), Recursive Neural Network (RNN), Convolutional Neural Networks (CNN) and Radial Basis Network (RBF) are some examples of NN approaches. They differ from each other mainly on the learning rules and on the topologies. In this project, we decided to use the MLP neural network with backpropagation algorithm base. This is considered one of the mostly known and frequently used type of NN (Popescu, Balas, Perescu-Popescu, & Mastorakis, 2009). It is an architecture that already proved to be successful in many areas and, specifically, in fraud prediction models and classification (Raghavendra Patidar, 2011) (Yeh & Lien, 2009). Nonlinearity, easy to use, adaptability and robustness are some of the characteristics of the MLP neural networks that allow them to generate such good results.

In order to train the MLP neural network we used the backpropagation algorithm, a technique commonly used with this type of NN. It uses the gradient descent method to identify the minimum of the error function in weight space. The combination of weights which minimizes the error function is considered to be a solution of the learning problem. This technique is also sometimes called backward propagation of errors, because the error is calculated at the output and distributed back through the network layers (Rojas, 1996). Backpropagation algorithm is not only more general than the usual analytical methods, but also has a strong learning capacity. In addition to this, it is easier to use when compared to other machine learning algorithms and it establishes with precision the relations between input and output variables.

The MLP neural networks are constituted by three different layers: input layers; hidden layers and output layers. Input layers receive the data with which the model will be trained and have as goal to keep this data on the network. The hidden layers are responsible for the second phase of the learning process, they define the mechanisms that will be applied during all the system. On the last phase of the NN method, the output layers determine the final solution of all the model. They are all fully connected by a certain weight and resort to a particular activation function. One of the most commonly used activation functions for backpropagation networks is the *sigmoid (*Fig. 16).

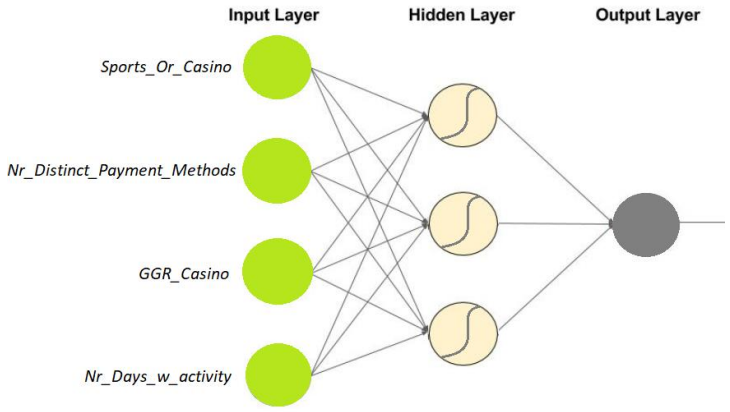


Fig. 18 – MLP Neural Network representation with one Hidden Layer

In order to decide how many hidden layers to apply to the model, commonly different options are tested and, in the end, the selected one would be the one that demonstrates to have the lowest error and the highest accuracy.

4.5.3. Decision Trees

A decision tree is a decision support tool that has the specificity of using a tree-like model to illustrate decisions and their possible consequences. The main objective is to develop a strategy able to reach a particular goal, in this case, maximize the accuracy of the prediction. It has increasingly become a popular algorithm influencing areas such as operations research, in particular decision analysis, machine learning, covering both classification and regression methods, and data mining. This popularity can be justified by the simplicity of the decision trees, not only when considering that they are based in intuitive and easy to understand rules, but also in what concerns data preparation. One of the biggest advantages of this classifier is that it can be interpreted by any person in any language. In addition, missing values, outliers and differences on the scales of the variables do not impair the performance of the algorithm, which results in not needing to do such deep and restricted data preparation. Besides, decision trees do not require any assumptions of linearity of the data, which allows them to be used when in presence of parameters nonlinearly related (Breslow & Aha, 1997).

The tree is represented upside down with its root at the top. The process of developing it involves deciding on which characteristics to choose and what conditions to apply for splitting, along with knowing when to stop (P. Gupta, 2017).

Illustrated on the Fig. 17 is the initial decision tree applied on this model. Note that the variable that is on the top of the tree is the one that adds more value to the model, in this case "*Sports_or_Casino"* variable.

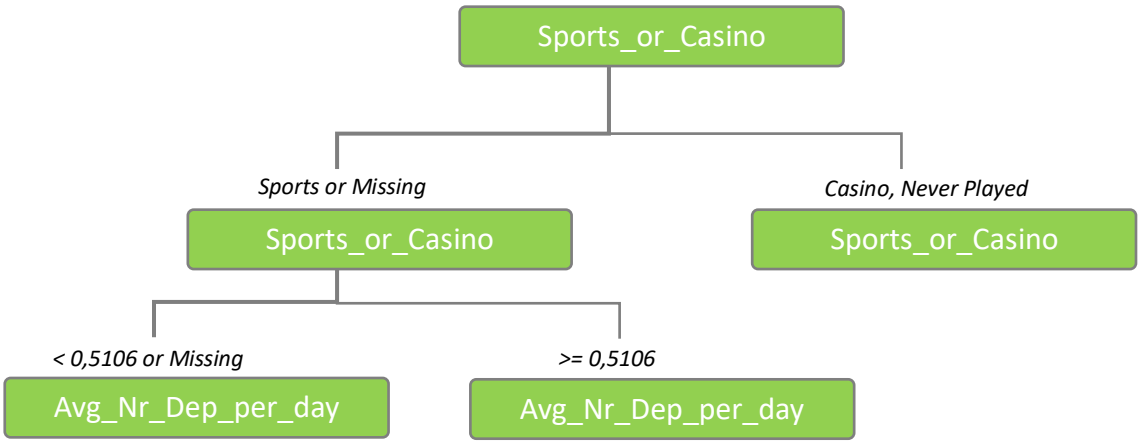


Fig. 19 – Decision Tree model representation

The decision tree process is characterized by the division of the instance space of each internal node into two or more sub-spaces respecting a certain discrete function, in this case, of the different input fraud variables. In the simplest and most frequent case, each step considers a single variable, such that the instance space is partitioned according to the variable's value. In the case of numeric attributes, the condition refers to a range (Rokach & Maimon, 2010). This division is made considering the variables that have the highest capacity of discrimination.

The original scheme grows continuously and downwardly, considering all the meaningful variables and training the algorithm in the best possible way. The learning process stops when there is only one class in the final subsets. The path to this class is considered as the "winning path" that will serve as basis to the required learning.

4.5.4. Ensemble

The Ensemble machine learning methods have a different approach when modelling: instead of concentrating on learning and finding similarities on the data, they are focused on working and improving the performance of the multiple tested algorithms. The main idea is to combine two or more predictive models in order to obtain a potentially more accurate model than the one we would obtain if we considered the models individually (Oza, n.d.).

SAS Enterprise Miner Ensemble node has available three different options to combine the posterior probabilities of the preceding modelling nodes: Average; Maximum and Voting. Average is the one that appears by default and it simple averages the posterior probabilities of the trained models. The maximum function sets the posterior probability as the highest posterior probability of the preceding modelling nodes. Finally, the voting option is used when in presence of a class target variable. It can

be applied in two ways: average and proportion. In the first one, the posterior probabilities are averaged and the most popular class is selected. All the other posterior probabilities are ignored. In the second one, the posterior probability of each class is recalculated by using the proportion of posterior probabilities that predict that class (Maldonado, Dean, Czika, & Haller, 2014).

In both financial and automobile insurance fields, ensemble methods proved to generate accurate results when predicting fraud behavior (Kotsiantis et al., 2006; Rodrigues & Omar, 2014). Having the literature review in consideration, and also that we're working with a class target variable, we decided to include the Ensemble method with average voting to our model comparison.

4.6. ACCESS

The main objective of the last phase of the SEMMA methodology is to evaluate the efficiency and accuracy of the conclusions obtained during the data mining process and estimate how well they perform when solving the project problem (Shafique & Qaiser, 2014). In order to carry out a critical evaluation of the model, to determine the overall reliability of the predicted data it is essential to consider the difference and the deviation between the prevision and the real value. There are different forms of measuring the performance of each algorithm used in the model, such as Relative Operating Characteristic curve (ROC curve) and the area below the curve; mean square error (MSE) and maximum absolute error (MAE) and confusion matrix (accuracy measure and classification error).

4.6.1. Confusion Matrix

The confusion matrix contains information of both real and predicted classifications obtained through a classification system. Commonly, it is designated using a matrix where the columns represent the predicted information and the rows represent the real information (Fig. 18). The entries in the matrix have the following meaning in the context of the study (Santra & Christy, 2012):

- True Positive (TP): total of predicted values that were correctly classified as positive;
- True Negative (TN): total of predicted values that were correctly classified as negative;
- False Positive (FP): total of predicted values that were incorrectly classified as positive;
- False Negative (FN): total of predicted values that were incorrectly classified as negative.

		PREDICTED		
		Positive	Negative	**Total**
REAL	Positive	TP	FN	**TP + FP**
	Negative	FP	TN	**FN + TN**
	Total	**TP + FN**	**FP + TN**	**TP + FP + FN + TN**

Fig. 20 – Correlation Matrix

Using the confusion matrix indicators, it is possible to calculate different measures such as (Sokolova & Lapalme, 2009):

- Accuracy: this metric can be used not only to evaluate the model but also to decide between different models. The higher the accuracy of the model, the better the prediction. When in the presence of a much larger number of negative cases compared to the number of positive cases, it may not be the most adequate performance measure (Bhowmik, 2008). In this case, the accuracy would be high even though the system missed out all the positive cases.

$$Accuracy = \frac{(TN + TP)}{(TN+TP+FN+FP)}$$

- Classification Error (Error): this indicator can be used to evaluate the performance of the classifier and can also help on the decision making of the best model. The lower the error, the better the model.

$$Error = \frac{(FP+FN)}{(TN+TP+FN+FP)}$$

- Recall and Precision: both these measures are able to assess the quality of the model. The precision represents the weight of the TP on the total of positive predictions and the recall describes the sensibility of the model.

$$Precision = \frac{TP}{(TP+FP)} \qquad Recall = \frac{TP}{(TP+FN)}$$

4.6.2. ROC Curve

Another form of evaluating the performance of a classifier is to examine the ROC Curve. It is represented by a graph and it is a technique used for visualizing, organizing and choosing classifiers (Fawcett, 2006). ROC graphs are two-dimensional graphs in which the TP rate or sensitivity (recall) is along the Y-axis and the FP rate or specificity is along the X-axis (Yang & Hwang, 2006). Both these measures are also derived using the classification matrix entries (Kumar & Indrayan, 2011).

$$Sensitivity\ (Recall) = \frac{TP}{(TP+FN)} \qquad Specificity = \frac{FP}{(TP+FP)}$$

The several points of the ROC Curve can give different information about the classifier. On the one hand, the point (0,1) is the point where all the positive and negative cases are correctly classified, it represents the perfect classifier. On the other hand, if the classifier is on the (0,0) point or (1,1) point it means it commits no false positive errors but also gains no true positives on the first case and it commits no true positive errors and also no true negative errors on the second case. In the first case, the strategy would be to never issue positive classification and, on the contrary, in the second case the strategy would be to always issue positive classification (Fawcett, 2006).

Using the ROC Curve, the way of classifying the performance of a model is by examining the area below the curve. This measure ranges between 0.5 and 1. The closer the classifier value is to 1, the better the model performance is.

4.6.3. Lift Chart

Lift Charts are represented by graphs that illustrate the improvement that an algorithm provides when compared with a random guess, it measures the change in terms of a lift score. The x-axis of the chart represents the percentage of the test dataset that is used to compare the predictions and the y-axis represents the percentage of predicted values (Duncan, Guyer, & Rabeler, 2018). Having this in consideration, this is a tool that can also be used to decide between different mining models by comparing the different lift scores of each. In addition, these charts also allow the evaluation of the point at which a model is no longer useful (Vuk, 2006).

This type of visual tool is commonly used on marketing campaigns because, besides helping to decide between different classifiers, it also illustrates the point at which it starts to be less lucrative to target a certain advertising campaign.

The major restriction of the Lift Charts tool is that it cannot be used to measure the accuracy of models that predict continuous numeric variables, since it requires the predicted value to be a discrete value (Duncan et al., 2018).

4.6.4. Mean Square Error (MSE)

In statistics, the Mean Square Error (MSE) of an estimator is used to evaluate the accuracy for continuous variables. It measures the average of the squares of the errors, meaning this it is calculated by the average square difference between the estimated values ($\widehat{xi}$) and the parameter (xi).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (xi - \widehat{xi})^2$$

The MSE ranges between 0 and 1. In order to choose the model with the highest power of prediction, the MSE value should be as close to 0 as possible. The higher the MSE value, the worse the power of prediction of the model.

5. RESULTS AND DISCUSSION

This project had as main goal to create a predictive model able to predict fraud behavior on an online betting market. After completing the modelling phase, testing different assumptions, we used the *Model Comparison Node* in order to assess which model performed best when predicting. All the results' combinations were compared among themselves and also with similar studies.

5.1. KEEPING OUTLIERS

Considering first the *keeping outliers approach,* Table 14 illustrates the combinations that were tested in order to reach the minimum mean square error (MSE) and the maximum ROC Index on the predictive model. It demonstrates both PCA and the Conjugation of Methods of dimensionality reduction together with the train and validation dataset results. For each of these outcomes, four different algorithms were compared: Ensemble; Regression; Decision Tree and Neural Network (NN). NN was evaluated considering a different number of hidden layers and also considering the Multilayer Perceptron Network and the Radial Basis Function Network. Ensemble was tested conjugating all the algorithms and also considering only particular algorithms. In the end, the combination that generated the best results was: NN(5) – Radial; NN(5); NN(4); NN(3); Regression; Decision Tree.

The Ensemble model on the Conjugation of Methods approach was the one that presented the best results, registering, on the validation set, a ROC Index of 0,763 and a Mean Squared Error of 0,193. Besides this model, NN(4) also presented satisfactory results in both PCA and Conjugation of Methods techniques, generating a ROC Index of 0,706 on the first approach and 0,753 on the second one.

Table 13 – Keeping Outliers – Final Results

	KEEPING OUTLIERS					
	PCA			**Conjugation of Methods[4]**		
	Train	*Valid*		*Train*	*Valid*	
	ROC Index	**ROC Index**	**MSE**	**ROC Index**	**ROC Index**	**MSE**
Ensemble	0,803	0,701	0,221	0,794	0,763	0,193
NN (4)	0,801	0,706	0,227	0,776	0,753	0,196
NN (3)	0,783	0,683	0,231	0,772	0,741	0,200
NN (5) - Radial	0,695	0,653	0,233	0,794	0,720	0,210
Regression	0,681	0,625	0,234	0,714	0,709	0,209
NN (2)	0,762	0,639	0,236	0,764	0,703	0,216
NN (5)	0,799	0,673	0,227	0,796	0,697	0,214
NN (1)	0,706	0,635	0,232	0,721	0,678	0,208
Decision Tree	0,708	0,698	0,217	0,713	0,671	0,225

Having in consideration that the Ensemble algorithm on the Conjugation of Methods approach was the one that demonstrated to predict best the study objective, we decided to also analyze the confusion matrix generated for this particular method.

[4] **Conjugation of Methods – dimensionality reduction considering Pearson's Correlation, Variable Worth Output and Variable Selection Node combined.**

Table 14 –Keeping Outliers - Confusion Matrix - Ensemble

		PREDICTED		
		Positive	Negative	**Total**
REAL	Positive	70	24	**94**
	Negative	60	137	**197**
	Total	**130**	**161**	**291**

When analyzing the confusion matrix's output, different evaluation metrics such as accuracy, precision and classification error can be calculated. Ensemble algorithm registered an accuracy of 0,711, a precision of 0,745 and a classification error of 0,289. It wrongly classified only 84 of the 291 observations.

5.2. REMOVING OUTLIERS

The same procedure explained above was followed on the *removing outliers approach* and is illustrated on Table 16. Comparing the PCA and Conjugation of Methods approaches, the second one was the one that generated the best results. Similar to what we concluded on the *keeping outliers approach*, the Ensemble method registered at the same time the highest ROC Index (0,735) and the lowest Mean Squared Error (0,205). These results, although approximate, were still lower than those achieved in the *keeping outliers approach*.

Table 15 – Removing Outliers – Final Results

	WITHOUT OUTLIERS					
	PCA			**Conjugation of Methods**		
	Train	*Valid*		*Train*	*Valid*	
	ROC Index	**ROC Index**	**MSE**	**ROC Index**	**ROC Index**	**MSE**
Ensemble	0,777	0,671	0,234	0,799	0,735	0,205
NN (4)	0,818	0,665	0,239	0,802	0,707	0,220
NN (3)	0,793	0,646	0,249	0,780	0,701	0,219
NN (5) - Radial	0,721	0,633	0,241	0,778	0,717	0,219
Regression	0,708	0,637	0,232	0,644	0,683	0,223
NN (2)	0,781	0,662	0,237	0,744	0,703	0,215
NN (5)	0,781	0,669	0,225	0,806	0,723	0,214
NN (1)	0,726	0,658	0,235	0,732	0,688	0,223
Decision Tree	0,673	0,607	0,242	0,711	0,674	0,224

To conclude, it's also important to notice that the Conjugation of Methods technique conjugated with both the Ensemble or NN(4) algorithms generated the best results on both the *keeping and removing outliers approaches.*

6. CONCLUSION

"In business, as in baseball, the question isn't whether or not you'll jump into analytics. The question is when. Do you want to ride the analytics horse to profitability… or follow it with a shovel" — Rob Neyer.

Online betting is a business that is growing exponentially each year, producing high levels of revenue and generating big amounts of data. There is an urgent need to build analytical procedures and create automatic systems capable of providing knowledge and assisting the development of the enterprises where this field is operating.

The project developed in this study had as main purpose to design a model able to predict fraud behavior on an online betting company, considering different variables, using different analytical assumptions and working with historical data.

The model was created having as base the SEMMA methodology and always having in consideration that, in data mining, there is not only one optimal solution, it is an interactive process. After testing different approaches, it is possible to conclude that the prediction results of the model were good. The algorithm that registered the best results was the Ensemble. It was able to correctly predict 71% of the validation set observations with a precision of 74%.

When comparing with similar studies, Gepp, Wilson, Kumar, & Bhattacharya, in 2012, developed a model of fraud detection in automative insurance that registered an accuracy approximate but still lower than the one obtained on our model, of 69,7% (Gepp, Wilson, Kumar, & Bhattacharya, 2012). Besides, more recently in 2016, an empirical study made on real-life data of financial truncations of an e-commerce organization recorded a precision nearly 1pp lower than the one reached on our study (Fahmi, Hamdy, & Nagati, 2016). Even more related with the field of study, in 2013, Schaidnagel, Petrov, & Laux built a fraud detection model for games merchants that generated weaker results on both Neural Networks and Decision Trees algorithms that the ones generated by our model on the same algorithms (Schaidnagel, Petrov, & Laux, 2013).

In what concerns the different approaches used in this study, we concluded that the model performed better when a conjugation of methods of dimensionality reduction was applied and also when the outliers were not excluded during the data preparation phase. Besides, the Ensemble, Neural Networks algorithm also proved to be able to achieve a good power of prediction of the model.

Regarding the influence of the variables, it is possible to conclude that variables related with the financial profile of the customer, such as average and total of deposits and withdrawals, are the ones that contributed more positively to the predictive power of the model. On the contrary, we could also conclude that some variables brought no value to the model such as "*Gender"* and *"Favorite Game"*.

In conclusion, this work intends to influence and improve the fraud detection processes applied in the online betting industry nowadays and also to be a positive contribute to the researchers in the area of Data Mining and Fraud.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Data mining is defined as the process of exploring large datasets in order to discover unknown patterns, trends, and relationships among the different chosen variables with the purpose of assisting on decision making about business strategies and competitive advantages. It is a methodology that already contributed positively for the improvement of different business fields. In fraud detection, it is a technique that has been growing and improving over time. However, data mining still has some limitations.

The principal limitation found when developing the model presented in this report was the low availability and quality of the data. Data mining is a technique that heavily relies on both these characteristics. Considering that our database only contained data from one year and a half of operation, it didn't had enough time to gather a big number of fraud cases and the ones collected were mainly related with duplication of accounts that can be considered as a less serious type of fraud.

As reference and suggestion for future works, first, it is proposed to consolidate a bigger database with more fraud cases and more different types of fraud.

Besides, it would also be interesting to develop different models focused on the different types of financial fraud already committed on this business: money laundering; chargeback; credit card fraud, etc. By dividing the analysis, new and appealing behaviors could be discovered on the customers.

Another work that could be useful to the online betting area is directly related with fraud in sports events. The influence of the results, referees and even players is an issue that is also heavily present in an online betting company, resulting in large losses of revenue. This type of analysis would allow the exploration in detail of variables more related with the bet and not so much with the customer such as: Live/Pre-match; Date of creation and Market Event Type (Final result; one player to score; which team scores first, among others).

To finish, the study of more developed online betting markets such as the English market would also be very interesting and would bring more knowledge and experience to the Portuguese market.

8. BIBLIOGRAPHY

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.101

ACFE Report. (2016). Report to the Nation on Occupational Fraud & Abuse. *Report to the Nation on Occupational Fraud & Abuse*.

Adami, N., Benini, S., Boschetti, A., Canini, L., Maione, F., & Temporin, M. (2013). Markers of unsustainable gambling for early detection of at-risk online gamblers. *International Gambling Studies*. https://doi.org/10.1080/14459795.2012.754919

Aggarwal, C. C. (2013). *Outlier analysis*. *Outlier Analysis*. https://doi.org/10.1007/978-1-4614-6396-2

Albashrawi, M., & Lowell, M. (2016). Detecting Financial Fraud Using Data Mining Techniques : a. *Journal of Data Science*, *14*(3), 553–570. Retrieved from http://www.jds-online.com/file_download/558/改10-Detecting+Financial+Fraud+Using+Data+Mining+Techniques-JDS_V3.pdf

Allan T. Mense. (2017). Introduction to Regression Techniques, 1–60.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January), 182–185. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Bănărescu, A. (2015). Detecting and Preventing Fraud with Data Analytics. *Procedia Economics and Finance*, *32*, 1827–1836. https://doi.org/10.1016/S2212-5671(15)01485-9

Banks, J. (2012). Online gambling and crime: a sure bet? *The ETHICOMP Journal*. Retrieved from http://shura.shu.ac.uk/6903/

Ben-gal, I. (2005). Outlier Detection. *Data Mining and Knowledge Discovery Handbook*. https://doi.org/10.1007/0-387-25465-x_7

Bhalla, D. (2014). Difference between Linear Regression and Logistic Regression. Retrieved from https://www.listendata.com/2014/11/difference-between-linear-regression.html

Bhowmik, R. (2008). Data Mining Techniques in Fraud Detection. *Proceedings of the Conference on Digital Forensics, Security and Law*, *3*(2), 57–72. Retrieved from http://proceedings.adfsl.org/index.php/CDFSL/article/view/123

Breslow, L. A., & Aha, D. W. (1997). Simplifying decision trees: A survey. *Knowledge Engineering Review*. https://doi.org/10.1017/S0269888997000015

Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management and Data Systems*. https://doi.org/10.1108/02635570310497657

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: Literature review and challenges. *International Journal of Distributed Sensor Networks*. https://doi.org/10.1155/2015/431047

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data To Big Impact. *Mis Quarterly*. https://doi.org/10.1145/2463676.2463712

Dajani, Y. (2015). Global Fraud Report Vulnerabilities on the Rise. *Kroll*, 72–73. Retrieved from

http://fraud.kroll.com/report-archive

Dash, M. (n.d.). Dimensionality Reduction.

Database, E. M. S., & Solutions, M. (2015). *SQL Management Studio for SQL Server User ' s Manual*.

Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. *Wiley & SAS business series*.

Deng, Q., & Mei, G. (2009). Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. In *2009 IEEE International Conference on Granular Computing, GRC 2009*. https://doi.org/10.1109/GRC.2009.5255148

Digital, T., & Network, I. (2017). 2017 Gaming and Gambling Cybercrime Report Global insights from the ThreatMetrix Digital Identity Network.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*. https://doi.org/10.1186/1755-8794-4-31

Duncan, O., Guyer, C., & Rabeler, C. (2018). Lift Chart (Analysis Services - Data Mining). Retrieved from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=sql-server-2017

EY. (2014). Big risks require big data thinking. *Global Forensic Data Analytics Survey 2014*.

EY. (2016). Shifting into high gear: mitigating risks and demonstrating returns Global Forensic Data Analytics Survey 2016, 1–36.

Fahmi, M., Hamdy, A., & Nagati, K. (2016). Data Mining Techniques for Credit Card Fraud Detection : Empirical Study. *Sustainable Vital Technologies in Engineering & Informatics*, (2015), 1–9.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2005.10.010

Gantz, J., Reinsel, D., & Shadows, B. D. (2012). The Digital Universe in 2020. *IDC IView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East."*

Gepp, A., Wilson, J. H., Kumar, K., & Bhattacharya, S. (2012). A Comparative Analysis of Decision Trees Vis a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. *Journal of Data Science*.

Gupta, P. (2017). Decision Trees in Machine Learning. Retrieved from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Gupta, S. (2015). A Regression Modeling Technique on Data Mining. *International Journal of Computer Applications*, *116*, 26.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Hipgrave, S. (2013). Smarter fraud investigations with big data analytics. *Network Security*. https://doi.org/10.1016/S1353-4858(13)70135-1

Kaiser, J. (2014). Dealing with Missing Values in Data. *Journal of Systems Integration*. https://doi.org/10.20470/jsi.v5i1.178

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*. https://doi.org/10.4097/kjae.2013.64.5.402

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition*. *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition*. https://doi.org/10.1002/9781118029145

Khalaf Ahmed Allam El-Din, A., & El Khamesy, N. (2016). Data Mining Techniques for Anti-Money Laundering. *International Journal of Computer Applications*.

Khan, R. Z. (2014). Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (comparative study). *Computer Science,Communication & Instrumentation Devices*, (December 2014), 163–172.

Kononenko, I., & Kukar, M. (2007). *Machine learning and data mining*. *Machine Learning and Data Mining*. https://doi.org/10.1533/9780857099440

Kordon, A. K. (2010). *Applying computational intelligence: How to create value*. *Applying Computational Intelligence: How to Create Value*. https://doi.org/10.1007/978-3-540-69913-2

Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting Fraudulent Financial Statements using Data Mining. In *PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY, VOL 12*.

Kotu, V., & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. https://doi.org/10.1016/C2014-0-00329-2

Kumar, R., & Indrayan, A. (2011). Receiver Operating Characteristic (ROC) Curve for Medical Researchers. *Indian Pediatrics*. https://doi.org/10.1007/s13312-011-0055-4

Le Khac, N. A., & Kechadi, M.-T. (2010). Application of data mining for anti-money laundering detection: A case study. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDMW.2010.66

Liou, F.-M. (2008). Fraudulent financial reporting detection and business failure prediction models: A comparison. *Managerial Auditing Journal*. https://doi.org/10.1108/02686900810890625

Maldonado, M., Dean, J., Czika, W., & Haller, S. (2014). Leveraging Ensemble Models in SAS ® Enterprise Miner ™, 1–15. Retrieved from http://support.sas.com/resources/papers/proceedings11/160-2011.pdf

Manikas, K. (2008). Outlier Detection in Online Gambling, (20008).

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*. https://doi.org/10.1017/S0269888910000032

Mukherjee, S., Mukherjee, T., & Nath, A. (2016). Fraud Analytics Using Data Mining. *International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)*, *3*(4), 1–11.

Olszewski, D. (2014). Fraud detection using self-organizing map visualizing the user profiles. *Knowledge-Based Systems, ScienceDirect*.

Oza, N. C. (n.d.). Ensemble Data Mining Methods. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications.*, *1*, 356–363.

Philander, K. S. (2014). Identifying high-risk online gamblers: A comparison of data mining procedures. *International Gambling Studies*. https://doi.org/10.1080/14459795.2013.841721

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Monash University*. https://doi.org/10.1016/j.chb.2012.01.002

Pinquet, J., Ayuso, M., & Guill??n, M. (2007). Selection bias and auditing policies for insurance claims. *Journal of Risk and Insurance*. https://doi.org/10.1111/j.1539-6975.2007.00219.x

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer Perceptron and Neural Networks. *WSEAS Transactions on Circuits and Systems*.

Prof, G. (2011). A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction. *International Journal of Computer Applications*. https://doi.org/10.5120/1977-2651

Raghavendra Patidar, L. S. (2011). Credit Card Fraud Detection Using Neural Network. *India International Journal of Soft Computing and Engineering*.

Rebekić, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or spearman's correlation coefficient – which one to use ? *Poljoprivreda*. https://doi.org/http://dx.doi.org/10.18047/poljo.21.2.8

Robert, W. (2015). Detecting and Dissuading Money Laundering Through Gambling, (September), 1–72.

Rodrigues, L. A., & Omar, N. (2014). Auto Claim Fraud Detection Using Multi Classifier System, 37–44. https://doi.org/10.5121/csit.2014.4604

Rojas, R. (1996). Neural networks: a systematic introduction. *Neural Networks*. https://doi.org/10.1016/0893-6080(94)90051-5

Rokach, L., & Maimon, O. (2010). Decision Trees. In *Data Mining and Knowledge Discovery Handbook*. https://doi.org/10.1007/0-387-25465-X_9

Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. In *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications* (pp. 315–319). https://doi.org/10.1109/INISTA.2011.5946108

Santra, a. K., & Christy, C. J. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering. *International Journal of Computer Science*.

SAS. (2016). Getting Started with SAS. *Getting Started with SAS Enterprise Miner 14.2*, 80.

Schaidnagel, M., Petrov, I., & Laux, F. (2013). DNA : An Online Algorithm for Credit Card Fraud Detection for Games Merchants, 1–6.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222. Retrieved from http://www.ijisr.issr-journals.org/

Smith, L. I. (2002). A tutorial on Principal Components Analysis Introduction. *Statistics*. https://doi.org/10.1080/03610928808829796

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2009.03.002

Soley-bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *PM931 Directed Study in Health Policy and Management*.

Solomatine, D., See, L. M., & Abrahart, R. J. (2008). Data-Driven Modelling: Concepts, Approaches and Experiences. *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*. https://doi.org/10.1007/978-3-540-79881-1_2

SRIJ, S. de regulação e inspeção dos J. (2018). Relatório 2º Trimestre Atividade do Jogo Online em Portugal.

Statstutor. (n.d.). Spearman's correlation. Retrieved from http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *International Journal of Digital Accounting Research*. https://doi.org/10.4192/1577-8517-v11_4

Torre, G., & Malfanti, F. (n.d.). Data mining for optimal gambling.

Understand your Online Gambling Business with Key Performance Indicators - EveryMatrix. (n.d.). Retrieved from https://everymatrix.com/blog/gambling-business-key-performance-indicators.html

Vaishnav, R. L., & Patel, M. (2015). Analysis of Various Techniques to Handling Missing Value in Dataset. *International Journal of Innovative and Emerging Research in Engineering*, *2*(2), 191–195.

Verleysen, M., & François, D. (2005). The Curse of Dimensionality in Data Mining. *Analysis*. https://doi.org/10.1007/11494669_93

Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2005.04.030

Vuk, M. (2006). ROC Curve , Lift Chart and Calibration Plot. *Metodolos˘ki Zvezki*. https://doi.org/10.1.1.126.7382

West, J., & Bhattacharya, M. (2015). Intelligent financial fraud detection: a comprehensive review. *Computers and Security*. https://doi.org/10.1016/j.cose.2015.09.005

West, J., & Bhattacharya, M. (2016). An investigation on experimental issues in financial fraud mining. In *Proceedings of the 2016 IEEE 11th Conference on Industrial Electronics and Applications, ICIEA 2016*. https://doi.org/10.1109/ICIEA.2016.7603878

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/10.1016/B978-0-12-374856-0.00005-5

Wood, R. T., & Williams, R. J. (2009). *Internet gambling: Prevalence, patterns, problems, and policy options*. *Final Report prepared for the Ontario Problem Gambling Research Centre*.

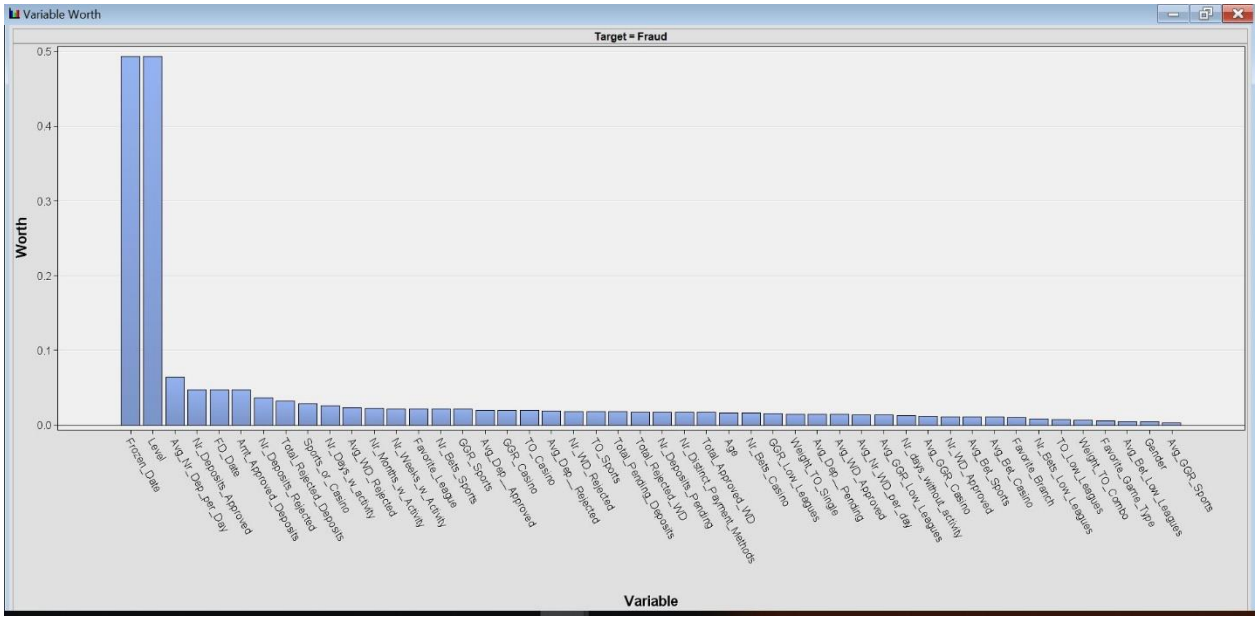
Xu, J., Sung, A. H., & Liu, Q. (2007). Behaviour mining for fraud detection. *Journal of Research and Practice in Information Technology*, *39*(1), 3–18.

Yang, W. S., & Hwang, S. Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2005.09.003

Yeh, I. C., & Lien, C. hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2007.12.020

9. ANNEXES

9.1. VARIABLE WORTH OUTPUT CONSIDERING 'FROZEN DATE' AND 'LEVEL' VARIABLES



9.2. INTERVAL VARIABLES OUTPUT

Target	*Interval Variabels*	*Missing*	*Mean*	*Standard Deviation*	*Median*
0	Total_Rejected_Deposits	0	169	5 367	10
1	Total_Rejected_Deposits	0	1 334	7 914	30
0	Nr_Deposits_Rejected	0	5	22	1
1	Nr_Deposits_Rejected	0	11	26	2
0	Avg_Dep__Rejected	0	19	263	5
1	Avg_Dep__Rejected	0	62	378	10
0	Total_Pending_Deposits	0	61	1 947	5
1	Total_Pending_Deposits	0	386	2 893	15
0	Nr_Deposits_Pending	0	3	11	1
1	Nr_Deposits_Pending	0	6	15	1
0	Avg_Dep__Pending	0	16	317	5
1	Avg_Dep__Pending	0	44	464	5
0	Amt_Approved_Deposits	0	473	21 547	55
1	Amt_Approved_Deposits	0	**3 142**	32 139	111
0	Nr_Deposits_Approved	0	21	69	4
1	Nr_Deposits_Approved	0	32	76	8
0	Avg_Dep__Approved	0	20	770	11
1	Avg_Dep__Approved	0	111	1 151	11
0	Avg_Nr_Dep_per_Day	0	0	1	0
1	Avg_Nr_Dep_per_Day	0	1	1	0
0	Total_Rejected_WD	0	46	3 623	0

1	Total_Rejected_WD	0	446	5 413	0
0	Nr_WD_Rejected	0	0	3	0
1	Nr_WD_Rejected	0	1	4	0
0	Avg_WD_Rejected	0	24	522	0
1	Avg_WD_Rejected	0	73	766	0
0	Total_Pending_WD	0	0	6 182	0
1	Total_Pending_WD	0	494	9 251	0
0	Nr_Withdrawls_Pending	0	0	0	0
1	Nr_Withdrawls_Pending	0	0	0	0
0	Avg_WD_Pending	0	0	6 182	0
1	Avg_WD_Pending	0	494	9 251	0
0	Total_Approved_WD	0	141	6 572	0
1	Total_Approved_WD	0	894	9 753	0
0	Nr_WD_Approved	0	1	5	0
1	Nr_WD_Approved	0	2	6	0
0	Avg_WD_Approved	0	31	731	0
1	Avg_WD_Approved	0	132	1 083	0
0	Avg_Nr_WD_per_day	0	0	0	0
1	Avg_Nr_WD_per_day	0	0	0	0
0	Nr_Distinct_Payment_Methods	0	1	1	1
1	Nr_Distinct_Payment_Methods	0	1	1	1
0	FD_Date	0	42 657	6 943	42 677
1	FD_Date	0	40 298	10 038	42 715
0	TO_Casino	0	312	15 185	0
1	TO_Casino	0	**3 452**	22 270	0
0	GGR_Casino	0	9	643	0
1	GGR_Casino	0	129	956	0
0	Nr_Bets_Casino	0	445	17 687	0
1	Nr_Bets_Casino	0	2 999	26 246	0
0	Avg_Bet_Casino	0	0	3	0
1	Avg_Bet_Casino	0	1	4	0
0	Avg_GGR_Casino	0	0	3	0
1	Avg_GGR_Casino	0	0	2	0
0	TO_Sports	0	1 637	76 169	175
1	TO_Sports	0	**8 307**	113 715	209
0	GGR_Sports	0	344	13 790	50
1	GGR_Sports	0	1 658	20 571	70
0	Nr_Bets_Sports	0	290	1 070	64
1	Nr_Bets_Sports	0	317	1 136	86
0	Avg_Bet_Sports	0	7	342	3
1	Avg_Bet_Sports	0	**49**	510	3
0	Avg_GGR_Sports	0	2	98	1
1	Avg_GGR_Sports	0	-1	147	1
0	TO_Low_Leagues	0	64	1 064	5
1	TO_Low_Leagues	0	148	1 568	5
0	GGR_Low_Leagues	0	7	127	1
1	GGR_Low_Leagues	0	24	170	2
0	Nr_Bets_Low_Leagues	0	21	58	4

1	Nr_Bets_Low_Leagues	0	22	58	6
0	Avg_Bet_Low_Leagues	0	4	45	1
1	Avg_Bet_Low_Leagues	0	7	63	1
0	Avg_GGR_Low_Leagues	0	0	10	0
1	Avg_GGR_Low_Leagues	0	1	10	0
0	TO_Combo	0	4 889	71 652	91
1	TO_Combo	0	961	4 206	101
0	Nr_days_without_activity	0	19	1 958	1
1	Nr_days_without_activity	0	-143	2 929	2
0	Nr_Days_w_activity	0	50	74	21
1	Nr_Days_w_activity	0	45	67	24
0	Nr_Months_w_Activity	0	5	4	4
1	Nr_Months_w_Activity	0	5	4	4
0	Nr_Weeks_w_Activity	0	13	13	8
1	Nr_Weeks_w_Activity	0	12	12	8
0	Age	0	32	10	30
1	Age	0	32	12	29

9.3. CLASS VARIABLES OUTPUT

Target Level	*Class Variables*	*Level*	*Frequency Count*
0	Favorite_Branch		9
1	Favorite_Branch		31
0	Favorite_Branch	0	1
0	Favorite_Branch	BasketBall	12
1	Favorite_Branch	BasketBall	9
0	Favorite_Branch	Handball	1
0	Favorite_Branch	Soccer	500
1	Favorite_Branch	Soccer	377
0	Favorite_Branch	Tennis	9
1	Favorite_Branch	Tennis	12
0	Favorite_Game_Type		372
1	Favorite_Game_Type		264
0	Favorite_Game_Type	Blackjack	24
1	Favorite_Game_Type	Blackjack	17
0	Favorite_Game_Type	Roulette	46
1	Favorite_Game_Type	Roulette	64
0	Favorite_Game_Type	Slots	90
1	Favorite_Game_Type	Slots	84
0	Favorite_League		9
1	Favorite_League		30
0	Favorite_League	0	1
0	Favorite_League	AFC Cup	2
0	Favorite_League	ATP Atlanta Qualifiers	1
0	Favorite_League	ATP Basel	1
1	Favorite_League	ATP Doha	1

1	Favorite_League	ATP Hamburg	1
1	Favorite_League	ATP Madrid	1
1	Favorite_League	ATP Miami	1
0	Favorite_League	ATP Vienna	1
0	Favorite_League	Argentina - Primera Div	4
1	Favorite_League	Argentina - Primera Div	4
1	Favorite_League	Argentina - Superliga	1
0	Favorite_League	Argentina Cup	1
1	Favorite_League	Argentina Cup	1
1	Favorite_League	Australia - League A	1
0	Favorite_League	Australia - NBL	2
1	Favorite_League	Brazil - Paulista	1
0	Favorite_League	Brazil - Serie A	10
1	Favorite_League	Brazil - Serie A	16
0	Favorite_League	Brazil - Serie B	2
1	Favorite_League	Brazil - Serie B	2
0	Favorite_League	Challenger Gimcheon	1
0	Favorite_League	Champions League	50
1	Favorite_League	Champions League	43
0	Favorite_League	Champions League Qualifying	4
1	Favorite_League	Champions League Qualifying	4
0	Favorite_League	Chile - Primera League	1
0	Favorite_League	Chinese - Super League	2
1	Favorite_League	Chinese - Super League	1
0	Favorite_League	Colombia - Primera	3
1	Favorite_League	Copa Libertadores	3
0	Favorite_League	Copa Sudamericana	2
0	Favorite_League	Denmark - Superligaen	1
1	Favorite_League	Denmark Handboldligaen Women	1
0	Favorite_League	England - Championship	2
1	Favorite_League	England - Championship	2
0	Favorite_League	England - Premier League	84
1	Favorite_League	England - Premier League	59
1	Favorite_League	English FA Cup	1
1	Favorite_League	English Football League Cup	2
0	Favorite_League	Euro 2016	1
0	Favorite_League	Euro U21 Championship	1
0	Favorite_League	Eurocup	1
0	Favorite_League	Europa League	13
1	Favorite_League	Europa League	12
0	Favorite_League	Europa League Qualifying	3
1	Favorite_League	Europa League Qualifying	1
0	Favorite_League	FIBA African Championship Women	1
0	Favorite_League	FIFA Confederations Cup	2
1	Favorite_League	FIFA Confederations Cup	2
0	Favorite_League	France - Ligue 1	5

1	Favorite_League	France - Ligue 1	4
0	Favorite_League	France - Ligue 2	1
0	Favorite_League	French Open Men	1
1	Favorite_League	French Open Men	3
0	Favorite_League	French Open Women	1
0	Favorite_League	Friendly	13
1	Favorite_League	Friendly	16
0	Favorite_League	Friendly International	4
1	Favorite_League	Friendly International	8
0	Favorite_League	German DFB Cup	1
1	Favorite_League	German DFB Cup	1
0	Favorite_League	Germany - 1.Bundesliga	15
1	Favorite_League	Germany - 1.Bundesliga	5
1	Favorite_League	Greece Cup	1
1	Favorite_League	ITF China F3	1
0	Favorite_League	ITF Tournaments	4
1	Favorite_League	ITF Tournaments	3
0	Favorite_League	ITF Tournaments, Doubles	1
1	Favorite_League	International Champions Cup	2
1	Favorite_League	Ireland - Premier	1
1	Favorite_League	Israel - Premier League	1
0	Favorite_League	Israel Cup	1
0	Favorite_League	Italy - Serie A	20
1	Favorite_League	Italy - Serie A	17
0	Favorite_League	Italy Cup	1
1	Favorite_League	Italy Cup	1
1	Favorite_League	Japan - J-League 1	1
0	Favorite_League	Japan - J-League 2	1
1	Favorite_League	Japan - J-League 2	1
0	Favorite_League	Lithuania - LKL	1
0	Favorite_League	NBA	40
1	Favorite_League	NBA	41
0	Favorite_League	NFL	1
1	Favorite_League	NHL	2
0	Favorite_League	Netherlands - Eerste Div	2
0	Favorite_League	Netherlands - Eredivisie	2
0	Favorite_League	Netherlands Cup	1
0	Favorite_League	Not Relevant	1
1	Favorite_League	Not Relevant	1
0	Favorite_League	Olympics 2016 - Men	3
0	Favorite_League	Olympics 2016 - Women	1
0	Favorite_League	Portugal - Liga Honra	1
0	Favorite_League	Portugal - Primeira Liga	105
1	Favorite_League	Portugal - Primeira Liga	58
0	Favorite_League	Portugal - Segunda Liga	4
1	Favorite_League	Portugal - Segunda Liga	1

0	Favorite_League	Portugal Cup	2
0	Favorite_League	Portugal League Cup	1
1	Favorite_League	Portugal League Cup	3
0	Favorite_League	Romania - Liga 1	1
0	Favorite_League	Russia - Premier Liga	1
1	Favorite_League	Russia - Premier Liga	1
1	Favorite_League	Russia KHL	1
1	Favorite_League	Slovenia Cup	1
0	Favorite_League	Spain - ACB League	1
0	Favorite_League	Spain - La Liga	38
1	Favorite_League	Spain - La Liga	25
0	Favorite_League	Spain - Segunda	1
0	Favorite_League	Spain Cup	2
1	Favorite_League	Spain Cup	2
1	Favorite_League	Spain Super Cup	1
0	Favorite_League	Turkey - Super Ligi	2
0	Favorite_League	Turkey Cup	2
1	Favorite_League	Turkey Cup	1
1	Favorite_League	UEFA Super Cup	1
0	Favorite_League	UEFA Youth League U19	1
0	Favorite_League	US Open Men	2
1	Favorite_League	US Open Men	1
1	Favorite_League	US Open Women	1
1	Favorite_League	US Open Women Qualifiers	1
0	Favorite_League	USA - MLS	1
1	Favorite_League	USA - MLS	2
1	Favorite_League	USA Open Cup	1
0	Favorite_League	Uruguay - Primera Div	1
0	Favorite_League	WTA Bucharest Qualifiers	1
1	Favorite_League	WTA Dubai	1
0	Favorite_League	World Cup 2018	12
1	Favorite_League	World Cup 2018	13
0	Favorite_League	World Cup Asia Qualifying	1
0	Favorite_League	World Cup Europe Qualify	4
1	Favorite_League	World Cup Europe Qualify	2
0	Favorite_League	World Cup Europe Qualifying	19
1	Favorite_League	World Cup Europe Qualifying	7
1	Favorite_League	World Cup Europe Women Qualifying	1
0	Favorite_League	World Cup South America Qualifying	2
1	Favorite_League	World Cup South America Qualifying	3
0	Gender	F	59
1	Gender	F	72
0	Gender	M	473
1	Gender	M	357
0	Sports_or_Casino	Casino	37
1	Sports_or_Casino	Casino	76

1	Sports_or_Casino	Never Played	23
0	Sports_or_Casino	Sports	495
1	Sports_or_Casino	Sports	330

9.4. VARIABLE WORTH OUTPUT AFTER MODIFY PHASE

Importance	*Variable*	*Worth*
1	Avg_Nr_Dep_per_Day	0,0669
2	Nr_Deposits_Rejected	0,0373
3	Sports_or_Casino	0,0356
4	Total_Rejected_Deposits	0,0343
5	Nr_Days_w_activity	0,0304
6	TO_Casino	0,0304
7	Total_Pending_Deposits	0,0294
8	Nr_Weeks_w_Activity	0,0263
9	Nr_Deposits_Pending	0,0257
10	Nr_Months_w_Activity	0,0254
11	Avg_Dep__Pending	0,0249
12	Nr_Deposits_Approved	0,0231
13	GGR_Casino	0,0224
14	Avg_Dep__Approved	0,0221
15	Nr_Bets_Sports	0,022
16	Nr_days_without_activity	0,0219
17	Amt_Approved_Deposits	0,0215
18	Age	0,0212
19	REP_Favorite_League	0,021
20	Nr_Bets_Casino	0,0203
21	Nr_Distinct_Payment_Methods	0,02
22	Avg_Dep__Rejected	0,02
23	Avg_Bet_Sports	0,0186
24	Avg_Nr_WD_per_day	0,0163
25	Total_Rejected_WD	0,016
26	Nr_WD_Rejected	0,016
27	Avg_Bet_Casino	0,0158
28	TO_Sports	0,0136
29	Avg_WD_Approved	0,0134
30	Avg_WD_Rejected	0,0132
31	Total_Approved_WD	0,0131
32	REP_Favorite_Branch	0,0127
33	TO_Combo	0,0124
34	GGR_Low_Leagues	0,0123
35	Nr_Bets_Low_Leagues	0,011
36	Avg_GGR_Casino	0,0109
37	Nr_WD_Approved	0,0107
38	Avg_GGR_Low_Leagues	0,0104
39	TO_Single	0,0096
40	Avg_GGR_Sports	0,0084
41	GGR_Sports	0,0079

42	Gender	0,0033
43	Avg_Bet_Low_Leagues	0,0031
44	TO_Low_Leagues	0,0024

9.5. Removing Outliers – Variable Selection output

Variable Name	*Role*	*Measurement Level*	*Reasons for Rejection*
Avg_Dep__Pending	Input	Interval	
Avg_Nr_Dep_per_Day	Input	Interval	
Avg_WD_Approved	Input	Interval	
G_REP_Favorite_Branch	Input	Nominal	
G_REP_Favorite_League	Input	Nominal	
Gender	Input	Binary	
Nr_Bets_Casino	Input	Interval	
Nr_Deposits_Approved	Input	Interval	
Nr_Deposits_Pending	Input	Interval	
Nr_Deposits_Rejected	Input	Interval	
Nr_WD_Rejected	Input	Interval	
Nr_days_without_activity	Input	Interval	
Sports_or_Casino	Input	Nominal	
TO_Casino	Input	Interval	
Total_Approved_WD	Input	Interval	
Total_Pending_Deposits	Input	Interval	
Total_Rejected_Deposits	Input	Interval	
Total_Rejected_WD	Input	Interval	
Age	Rejected	Interval	Varsel:Small R-square value
Amt_Approved_Deposits	Rejected	Interval	Varsel:Small R-square value
Avg_Bet_Casino	Rejected	Interval	Varsel:Small R-square value
Avg_Bet_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
Avg_Bet_Sports	Rejected	Interval	Varsel:Small R-square value
Avg_Dep__Approved	Rejected	Interval	Varsel:Small R-square value
Avg_Dep__Rejected	Rejected	Interval	Varsel:Small R-square value
Avg_GGR_Casino	Rejected	Interval	Varsel:Small R-square value
Avg_GGR_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
Avg_GGR_Sports	Rejected	Interval	Varsel:Small R-square value
Avg_Nr_WD_per_day	Rejected	Interval	Varsel:Small R-square value
Avg_WD_Pending	Rejected	Interval	Varsel:Small R-square value
Avg_WD_Rejected	Rejected	Interval	Varsel:Small R-square value
GGR_Casino	Rejected	Interval	Varsel:Small R-square value
GGR_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
GGR_Sports	Rejected	Interval	Varsel:Small R-square value
Nr_Bets_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
Nr_Bets_Sports	Rejected	Interval	Varsel:Small R-square value
Nr_Days_w_activity	Rejected	Interval	Varsel:Small R-square value
Nr_Distinct_Payment_Methods	Rejected	Interval	Varsel:Small R-square value
Nr_Months_w_Activity	Rejected	Interval	Varsel:Small R-square value

Nr_WD_Approved	Rejected	Interval	Varsel:Small R-square value
Nr_Weeks_w_Activity	Rejected	Interval	Varsel:Small R-square value
Nr_Withdrawls_Pending	Rejected	Interval	Varsel:Small R-square value
TO_Combo	Rejected	Interval	Varsel:Small R-square value
TO_Low_Leagues	Rejected	Interval	Varsel:Small R-square value
TO_Single	Rejected	Interval	Varsel:Small R-square value
TO_Sports	Rejected	Interval	Varsel:Small R-square value
Total_Pending_WD	Rejected	Interval	Varsel:Small R-square value
REP_Favorite_Branch	Rejected	Nominal	Varsel:Small R-square value, Group variable preferred
REP_Favorite_League	Rejected	Nominal	Varsel:Small R-square value, Group variable preferred

9.6. KEEPING OUTLIERS – SPEARMAN CORRELATION

Variables	Age	Amt_Approved_Deposits	Avg_Bet_Casino	Avg_Bet_Low_Leagues	Avg_Bet_Sports	Avg_Dep__Approved	Avg_Dep__Pending	Avg_Dep__Rejected	Avg_GGR_Casino	Avg_GGR_Low_Leagues	Avg_GGR_Sports	Avg_Nr_Dep_per_Day	Avg_Nr_WD_per_day	Avg_WD_Approved
Age	1,00	0,05	-0,09	-0,01	-0,04	0,03	0,03	-0,01	-0,02	0,02	-0,02	-0,04	-0,05	-0,04
Amt_Approved_Deposits	0,05	1,00	0,35	0,56	0,44	0,53	0,46	0,56	0,20	0,29	0,24	0,45	0,48	0,51
Avg_Bet_Casino	-0,09	0,35	1,00	0,09	0,06	0,14	0,19	0,22	0,61	0,04	0,00	0,33	0,26	0,27
Avg_Bet_Low_Leagues	-0,01	0,56	0,09	1,00	0,50	0,39	0,28	0,30	0,03	0,49	0,18	0,14	0,30	0,32
Avg_Bet_Sports	-0,04	0,44	0,06	0,50	1,00	0,62	0,19	0,27	0,01	0,24	0,65	0,40	0,28	0,28
Avg_Dep__Approved	0,03	0,53	0,14	0,39	0,62	1,00	0,20	0,38	0,09	0,22	0,44	0,19	0,23	0,24
Avg_Dep__Pending	0,03	0,46	0,19	0,28	0,19	0,20	1,00	0,33	0,09	0,18	0,11	0,21	0,27	0,28
Avg_Dep__Rejected	-0,01	0,56	0,22	0,30	0,27	0,38	0,33	1,00	0,15	0,12	0,13	0,22	0,32	0,34
Avg_GGR_Casino	-0,02	0,20	0,61	0,03	0,01	0,09	0,09	0,15	1,00	-0,01	0,03	0,26	0,07	0,07
Avg_GGR_Low_Leagues	0,02	0,29	0,04	0,49	0,24	0,22	0,18	0,12	-0,01	1,00	0,24	0,15	0,06	0,07
Avg_GGR_Sports	-0,02	0,24	0,00	0,18	0,65	0,44	0,11	0,13	0,03	0,24	1,00	0,48	-0,08	-0,06
Avg_Nr_Dep_per_Day	-0,04	0,45	0,33	0,14	0,40	0,19	0,21	0,22	0,26	0,15	0,48	1,00	0,27	0,27
Avg_Nr_WD_per_day	-0,05	0,48	0,26	0,30	0,28	0,23	0,27	0,32	0,07	0,06	-0,08	0,27	1,00	0,98
Avg_WD_Approved	-0,04	0,51	0,27	0,32	0,28	0,24	0,28	0,34	0,07	0,07	-0,06	0,27	0,98	1,00
Avg_WD_Pending	-0,05	0,08	0,05	0,09	0,10	0,09	-0,01	0,10	0,06	0,10	0,06	0,05	0,08	0,09
Avg_WD_Rejected	-0,10	0,47	0,26	0,33	0,26	0,17	0,31	0,30	0,14	0,10	-0,01	0,28	0,48	0,47
FD_Date	-0,02	-0,12	0,05	-0,15	0,00	0,11	-0,29	0,10	0,12	-0,05	0,07	0,11	-0,08	-0,10
GGR_Casino	-0,02	0,25	0,59	0,04	0,01	0,10	0,13	0,18	0,97	-0,01	0,02	0,30	0,11	0,11
GGR_Low_Leagues	0,06	0,46	0,07	0,44	0,16	0,17	0,26	0,20	-0,01	0,89	0,14	0,16	0,14	0,15
GGR_Sports	0,08	0,81	0,14	0,53	0,44	0,41	0,40	0,41	0,08	0,34	0,48	0,31	0,18	0,21
Nr_Bets_Casino	-0,08	0,37	0,93	0,07	0,01	0,13	0,19	0,23	0,53	0,03	-0,04	0,32	0,29	0,30
Nr_Bets_Low_Leagues	0,09	0,67	0,15	0,57	0,07	0,12	0,34	0,34	0,04	0,29	-0,08	0,04	0,29	0,32
Nr_Bets_Sports	0,10	0,74	0,17	0,52	0,11	0,17	0,38	0,39	0,06	0,20	-0,14	0,03	0,31	0,35
Nr_Days_w_activity	0,09	0,74	0,18	0,48	0,11	0,15	0,38	0,39	0,04	0,18	-0,13	-0,07	0,37	0,40
Nr_Deposits_Approved	0,04	0,92	0,36	0,49	0,27	0,21	0,47	0,49	0,20	0,25	0,10	0,46	0,47	0,50
Nr_Deposits_Pending	0,01	0,49	0,19	0,25	0,10	0,00	0,84	0,29	0,09	0,16	0,03	0,26	0,27	0,29
Nr_Deposits_Rejected	0,02	0,60	0,27	0,27	0,16	0,16	0,37	0,82	0,18	0,11	0,06	0,33	0,33	0,36
Nr_Distinct_Payment_Methods	0,00	0,45	0,16	0,27	0,20	0,19	0,22	0,21	0,11	0,13	0,11	0,27	0,30	0,31
Nr_Months_w_Activity	0,06	0,72	0,24	0,44	0,13	0,13	0,36	0,36	0,07	0,16	-0,06	0,04	0,36	0,39
Nr_WD_Approved	-0,05	0,52	0,28	0,32	0,28	0,22	0,28	0,33	0,08	0,07	-0,06	0,27	0,99	0,98
Nr_WD_Rejected	-0,10	0,46	0,26	0,32	0,25	0,14	0,30	0,29	0,15	0,10	-0,02	0,28	0,48	0,47
Nr_Weeks_w_Activity	0,09	0,74	0,20	0,47	0,13	0,15	0,37	0,37	0,07	0,19	-0,09	-0,01	0,37	0,40
Nr_Withdrawls_Pending	-0,05	0,08	0,05	0,09	0,10	0,09	-0,01	0,10	0,06	0,10	0,06	0,05	0,08	0,09
Nr_days_without_activity	-0,03	0,07	0,07	0,02	0,03	0,06	0,09	0,13	0,01	0,05	0,08	0,10	0,02	0,02
TO_Casino	-0,08	0,38	0,96	0,07	0,03	0,14	0,20	0,25	0,59	0,03	-0,03	0,34	0,30	0,30
TO_Combo	0,00	0,29	0,00	0,24	0,17	0,12	0,18	0,11	-0,04	0,14	0,13	0,06	0,13	0,14
TO_Low_Leagues	0,05	0,73	0,13	0,85	0,33	0,30	0,36	0,38	0,03	0,39	0,05	0,13	0,36	0,39
TO_Single	0,06	0,53	0,06	0,48	0,36	0,30	0,26	0,23	-0,01	0,15	0,02	0,07	0,23	0,25
TO_Sports	0,07	0,86	0,18	0,67	0,52	0,43	0,41	0,47	0,06	0,27	0,13	0,19	0,41	0,44
Total_Approved_WD	-0,05	0,52	0,28	0,32	0,29	0,24	0,29	0,34	0,08	0,08	-0,06	0,27	0,99	0,99
Total_Pending_Deposits	0,03	0,53	0,20	0,30	0,17	0,12	0,94	0,33	0,09	0,19	0,08	0,25	0,30	0,32
Total_Pending_WD	-0,05	0,08	0,05	0,09	0,10	0,09	-0,01	0,10	0,06	0,10	0,06	0,05	0,08	0,09
Total_Rejected_Deposits	0,01	0,63	0,27	0,31	0,22	0,27	0,38	0,91	0,18	0,12	0,10	0,32	0,36	0,38
Total_Rejected_WD	-0,10	0,48	0,26	0,33	0,26	0,16	0,31	0,31	0,15	0,10	-0,01	0,29	0,48	0,47
dataobs	-0,02	-0,27	-0,03	-0,24	-0,13	-0,04	-0,33	0,01	0,07	-0,12	-0,09	-0,09	-0,15	-0,15

Variables	Avg_WD_Pending	Avg_WD_Rejected	FD_Date	GGR_Casino	GGR_Low_Leagues	GGR_Sports	Nr_Bets_Casino	Nr_Bets_Low_Leagues	Nr_Bets_Sports	Nr_Days_w_activity	Nr_Deposits_Approved	Nr_Deposits_Pending	Nr_Deposits_Rejected	Nr_Distinct_Payment_Methods	Nr_Months_w_Activity	Nr_WD_Approved
Age	-0,05	-0,10	-0,02	-0,02	0,06	0,08	-0,08	0,09	0,10	0,09	0,04	0,01	0,02	0,00	0,06	-0,05
Amt_Approved_Deposits	0,08	0,47	-0,12	0,25	0,46	0,81	0,37	0,67	0,74	0,74	0,92	0,49	0,60	0,45	0,72	0,52
Avg_Bet_Casino	0,05	0,26	0,05	0,59	0,07	0,14	0,93	0,15	0,17	0,18	0,36	0,19	0,27	0,16	0,24	0,28
Avg_Bet_Low_Leagues	0,09	0,33	-0,15	0,04	0,44	0,53	0,07	0,57	0,52	0,48	0,49	0,25	0,27	0,27	0,44	0,32
Avg_Bet_Sports	0,10	0,26	0,00	0,01	0,16	0,44	0,01	0,07	0,11	0,11	0,27	0,10	0,16	0,20	0,13	0,28
Avg_Dep__Approved	0,09	0,17	0,11	0,10	0,17	0,41	0,13	0,12	0,17	0,15	0,21	0,00	0,16	0,19	0,13	0,22
Avg_Dep__Pending	-0,01	0,31	-0,29	0,13	0,26	0,40	0,19	0,34	0,38	0,38	0,47	0,84	0,37	0,22	0,36	0,28
Avg_Dep__Rejected	0,10	0,30	0,10	0,18	0,20	0,41	0,23	0,34	0,39	0,39	0,49	0,29	0,82	0,21	0,36	0,33
Avg_GGR_Casino	0,06	0,14	0,12	0,97	-0,01	0,08	0,53	0,04	0,06	0,04	0,20	0,09	0,18	0,11	0,07	0,08
Avg_GGR_Low_Leagues	0,10	0,10	-0,05	-0,01	0,89	0,34	0,03	0,29	0,20	0,18	0,25	0,16	0,11	0,13	0,16	0,07
Avg_GGR_Sports	0,06	-0,01	0,07	0,02	0,14	0,48	-0,04	-0,08	-0,14	-0,13	0,10	0,03	0,06	0,11	-0,06	-0,06
Avg_Nr_Dep_per_Day	0,05	0,28	0,11	0,30	0,16	0,31	0,32	0,04	0,03	-0,07	0,46	0,26	0,33	0,27	0,04	0,27
Avg_Nr_WD_per_day	0,08	0,48	-0,08	0,11	0,14	0,18	0,29	0,29	0,31	0,37	0,47	0,27	0,33	0,30	0,36	0,99
Avg_WD_Approved	0,09	0,47	-0,10	0,11	0,15	0,21	0,30	0,32	0,35	0,40	0,50	0,29	0,36	0,31	0,39	0,98
Avg_WD_Pending	1,00	0,08	0,04	0,06	0,10	0,05	0,04	0,06	0,02	0,05	0,07	-0,01	0,06	0,08	0,03	0,09
Avg_WD_Rejected	0,08	1,00	-0,07	0,18	0,18	0,26	0,29	0,31	0,36	0,36	0,48	0,36	0,42	0,28	0,37	0,48
FD_Date	0,04	-0,07	1,00	0,09	-0,12	-0,18	0,04	-0,21	-0,21	-0,22	-0,16	-0,34	0,07	0,01	-0,25	-0,11
GGR_Casino	0,06	0,18	0,09	1,00	0,02	0,10	0,56	0,07	0,09	0,08	0,25	0,14	0,23	0,13	0,11	0,12
GGR_Low_Leagues	0,10	0,18	-0,12	0,02	1,00	0,50	0,09	0,54	0,45	0,43	0,45	0,32	0,25	0,17	0,37	0,17
GGR_Sports	0,05	0,26	-0,18	0,10	0,50	1,00	0,15	0,67	0,73	0,68	0,76	0,42	0,45	0,32	0,64	0,23
Nr_Bets_Casino	0,04	0,29	0,04	0,56	0,09	0,15	1,00	0,18	0,21	0,22	0,39	0,20	0,30	0,18	0,28	0,31
Nr_Bets_Low_Leagues	0,06	0,31	-0,21	0,07	0,54	0,67	0,18	1,00	0,87	0,84	0,74	0,44	0,45	0,29	0,75	0,33
Nr_Bets_Sports	0,02	0,36	-0,21	0,09	0,45	0,73	0,21	0,87	1,00	0,92	0,80	0,47	0,48	0,34	0,81	0,37
Nr_Days_w_activity	0,05	0,36	-0,22	0,08	0,43	0,68	0,22	0,84	0,92	1,00	0,82	0,47	0,48	0,37	0,90	0,42
Nr_Deposits_Approved	0,07	0,48	-0,16	0,25	0,45	0,76	0,39	0,74	0,80	0,82	1,00	0,56	0,63	0,46	0,79	0,51
Nr_Deposits_Pending	-0,01	0,36	-0,34	0,14	0,32	0,42	0,20	0,44	0,47	0,47	0,56	1,00	0,46	0,27	0,45	0,30
Nr_Deposits_Rejected	0,06	0,42	0,07	0,23	0,25	0,45	0,30	0,45	0,48	0,48	0,63	0,46	1,00	0,32	0,45	0,36
Nr_Distinct_Payment_Methods	0,08	0,28	0,01	0,13	0,17	0,32	0,18	0,29	0,34	0,37	0,46	0,27	0,32	1,00	0,40	0,31
Nr_Months_w_Activity	0,03	0,37	-0,25	0,11	0,37	0,64	0,28	0,75	0,81	0,90	0,79	0,45	0,45	0,40	1,00	0,41
Nr_WD_Approved	0,09	0,48	-0,11	0,12	0,17	0,23	0,31	0,33	0,37	0,42	0,51	0,30	0,36	0,31	0,41	1,00
Nr_WD_Rejected	0,09	0,98	-0,06	0,19	0,19	0,26	0,30	0,32	0,36	0,36	0,48	0,37	0,43	0,29	0,37	0,49
Nr_Weeks_w_Activity	0,05	0,37	-0,24	0,10	0,42	0,67	0,24	0,80	0,87	0,97	0,82	0,46	0,47	0,40	0,96	0,42
Nr_Withdrawls_Pending	0,99	0,08	0,04	0,06	0,10	0,05	0,04	0,06	0,02	0,05	0,07	-0,01	0,06	0,08	0,03	0,09
Nr_days_without_activity	0,09	0,01	0,30	0,00	0,05	0,05	0,05	0,08	0,05	0,05	0,08	0,10	0,20	0,10	0,05	0,02
TO_Casino	0,04	0,29	0,04	0,62	0,08	0,15	0,98	0,17	0,20	0,21	0,40	0,21	0,32	0,18	0,27	0,31
TO_Combo	0,00	0,13	0,02	-0,02	0,19	0,33	0,01	0,28	0,34	0,35	0,31	0,18	0,13	0,17	0,31	0,14
TO_Low_Leagues	0,09	0,38	-0,23	0,06	0,56	0,70	0,15	0,89	0,80	0,76	0,72	0,41	0,43	0,31	0,69	0,40
TO_Single	0,06	0,26	-0,19	0,01	0,27	0,53	0,08	0,48	0,62	0,54	0,50	0,29	0,22	0,20	0,49	0,26
TO_Sports	0,08	0,43	-0,19	0,09	0,45	0,82	0,19	0,75	0,87	0,82	0,81	0,44	0,50	0,37	0,74	0,45
Total_Approved_WD	0,09	0,48	-0,10	0,12	0,17	0,23	0,31	0,33	0,36	0,42	0,51	0,30	0,36	0,31	0,40	0,99
Total_Pending_Deposits	-0,01	0,37	-0,33	0,14	0,32	0,46	0,20	0,43	0,47	0,47	0,56	0,97	0,46	0,28	0,44	0,33
Total_Pending_WD	1,00	0,08	0,04	0,06	0,10	0,05	0,04	0,06	0,02	0,05	0,07	-0,01	0,06	0,08	0,03	0,09
Total_Rejected_Deposits	0,09	0,41	0,08	0,23	0,25	0,47	0,29	0,43	0,47	0,47	0,62	0,42	0,97	0,31	0,44	0,38
Total_Rejected_WD	0,09	0,99	-0,07	0,19	0,18	0,27	0,30	0,32	0,36	0,37	0,48	0,37	0,43	0,29	0,37	0,48
dataobs	-0,01	-0,11	0,74	0,05	-0,18	-0,32	-0,04	-0,31	-0,31	-0,32	-0,31	-0,38	-0,05	-0,15	-0,35	-0,17

Variables	Nr_WD_Rejected	Nr_Weeks_w_Activity	Nr_Withdrawls_Pending	Nr_days_without_activity	TO_Casino	TO_Combo	TO_Low_Leagues	TO_Single	TO_Sports	Total_Approved_WD	Total_Pending_Deposits	Total_Pending_WD	Total_Rejected_Deposits	Total_Rejected_WD	Observation Number
Age	-0,10	0,09	-0,05	-0,03	-0,08	0,00	0,05	0,06	0,07	-0,05	0,03	-0,05	0,01	-0,10	-0,02
Amt_Approved_Deposits	0,46	0,74	0,08	0,07	0,38	0,29	0,73	0,53	0,86	0,52	0,53	0,08	0,63	0,48	-0,27
Avg_Bet_Casino	0,26	0,20	0,05	0,07	0,96	0,00	0,13	0,06	0,18	0,28	0,20	0,05	0,27	0,26	-0,03
Avg_Bet_Low_Leagues	0,32	0,47	0,09	0,02	0,07	0,24	0,85	0,48	0,67	0,32	0,30	0,09	0,31	0,33	-0,24
Avg_Bet_Sports	0,25	0,13	0,10	0,03	0,03	0,17	0,33	0,36	0,52	0,29	0,17	0,10	0,22	0,26	-0,13
Avg_Dep__Approved	0,14	0,15	0,09	0,06	0,14	0,12	0,30	0,30	0,43	0,24	0,12	0,09	0,27	0,16	-0,04
Avg_Dep__Pending	0,30	0,37	-0,01	0,09	0,20	0,18	0,36	0,26	0,41	0,29	0,94	-0,01	0,38	0,31	-0,33
Avg_Dep__Rejected	0,29	0,37	0,10	0,13	0,25	0,11	0,38	0,23	0,47	0,34	0,33	0,10	0,91	0,31	0,01
Avg_GGR_Casino	0,15	0,07	0,06	0,01	0,59	-0,04	0,03	-0,01	0,06	0,08	0,09	0,06	0,18	0,15	0,07
Avg_GGR_Low_Leagues	0,10	0,19	0,10	0,05	0,03	0,14	0,39	0,15	0,27	0,08	0,19	0,10	0,12	0,10	-0,12
Avg_GGR_Sports	-0,02	-0,09	0,06	0,08	-0,03	0,13	0,05	0,02	0,13	-0,06	0,08	0,06	0,10	-0,01	-0,09
Avg_Nr_Dep_per_Day	0,28	-0,01	0,05	0,10	0,34	0,06	0,13	0,07	0,19	0,27	0,25	0,05	0,32	0,29	-0,09
Avg_Nr_WD_per_day	0,48	0,37	0,08	0,02	0,30	0,13	0,36	0,23	0,41	0,99	0,30	0,08	0,36	0,48	-0,15
Avg_WD_Approved	0,47	0,40	0,09	0,02	0,30	0,14	0,39	0,25	0,44	0,99	0,32	0,09	0,38	0,47	-0,15
Avg_WD_Pending	0,09	0,05	0,99	0,09	0,04	0,00	0,09	0,06	0,08	0,09	-0,01	1,00	0,09	0,09	-0,01
Avg_WD_Rejected	0,98	0,37	0,08	0,01	0,29	0,13	0,38	0,26	0,43	0,48	0,37	0,08	0,41	0,99	-0,11
FD_Date	-0,06	-0,24	0,04	0,30	0,04	0,02	-0,23	-0,19	-0,19	-0,10	-0,33	0,04	0,08	-0,07	0,74
GGR_Casino	0,19	0,10	0,06	0,00	0,62	-0,02	0,06	0,01	0,09	0,12	0,14	0,06	0,23	0,19	0,05
GGR_Low_Leagues	0,19	0,42	0,10	0,05	0,08	0,19	0,56	0,27	0,45	0,17	0,32	0,10	0,25	0,18	-0,18
GGR_Sports	0,26	0,67	0,05	0,05	0,15	0,33	0,70	0,53	0,82	0,23	0,46	0,05	0,47	0,27	-0,32
Nr_Bets_Casino	0,30	0,24	0,04	0,05	0,98	0,01	0,15	0,08	0,19	0,31	0,20	0,04	0,29	0,30	-0,04
Nr_Bets_Low_Leagues	0,32	0,80	0,06	0,08	0,17	0,28	0,89	0,48	0,75	0,33	0,43	0,06	0,43	0,32	-0,31
Nr_Bets_Sports	0,36	0,87	0,02	0,05	0,20	0,34	0,80	0,62	0,87	0,36	0,47	0,02	0,47	0,36	-0,31
Nr_Days_w_activity	0,36	0,97	0,05	0,05	0,21	0,35	0,76	0,54	0,82	0,42	0,47	0,05	0,47	0,37	-0,32
Nr_Deposits_Approved	0,48	0,82	0,07	0,08	0,40	0,31	0,72	0,50	0,81	0,51	0,56	0,07	0,62	0,48	-0,31
Nr_Deposits_Pending	0,37	0,46	-0,01	0,10	0,21	0,18	0,41	0,29	0,44	0,30	0,97	-0,01	0,42	0,37	-0,38
Nr_Deposits_Rejected	0,43	0,47	0,06	0,20	0,32	0,13	0,43	0,22	0,50	0,36	0,46	0,06	0,97	0,43	-0,05
Nr_Distinct_Payment_Methods	0,29	0,40	0,08	0,10	0,18	0,17	0,31	0,20	0,37	0,31	0,28	0,08	0,31	0,29	-0,15
Nr_Months_w_Activity	0,37	0,96	0,03	0,05	0,27	0,31	0,69	0,49	0,74	0,40	0,44	0,03	0,44	0,37	-0,35
Nr_WD_Approved	0,49	0,42	0,09	0,02	0,31	0,14	0,40	0,26	0,45	0,99	0,33	0,09	0,38	0,48	-0,17
Nr_WD_Rejected	1,00	0,37	0,09	0,01	0,31	0,13	0,38	0,25	0,42	0,48	0,37	0,09	0,41	0,99	-0,10
Nr_Weeks_w_Activity	0,37	1,00	0,05	0,06	0,24	0,33	0,74	0,52	0,79	0,41	0,46	0,05	0,46	0,37	-0,35
Nr_Withdrawls_Pending	0,09	0,05	1,00	0,09	0,04	0,00	0,09	0,06	0,08	0,09	-0,01	0,99	0,09	0,09	-0,01
Nr_days_without_activity	0,01	0,06	0,09	1,00	0,05	0,12	0,03	-0,02	0,02	0,02	0,09	0,09	0,18	0,02	-0,07
TO_Casino	0,31	0,24	0,04	0,05	1,00	0,01	0,14	0,07	0,19	0,31	0,22	0,04	0,31	0,30	-0,03
TO_Combo	0,13	0,33	0,00	0,12	0,01	1,00	0,28	0,22	0,35	0,14	0,19	0,00	0,13	0,13	-0,10
TO_Low_Leagues	0,38	0,74	0,09	0,03	0,14	0,28	1,00	0,57	0,84	0,40	0,43	0,09	0,44	0,39	-0,32
TO_Single	0,25	0,52	0,06	-0,02	0,07	0,22	0,57	1,00	0,70	0,26	0,31	0,06	0,25	0,26	-0,24
TO_Sports	0,42	0,79	0,08	0,02	0,19	0,35	0,84	0,70	1,00	0,45	0,48	0,08	0,53	0,43	-0,30
Total_Approved_WD	0,48	0,41	0,09	0,02	0,31	0,14	0,40	0,26	0,45	1,00	0,33	0,09	0,39	0,48	-0,16
Total_Pending_Deposits	0,37	0,46	-0,01	0,09	0,22	0,19	0,43	0,31	0,48	0,33	1,00	-0,01	0,44	0,37	-0,38
Total_Pending_WD	0,09	0,05	0,99	0,09	0,04	0,00	0,09	0,06	0,08	0,09	-0,01	1,00	0,09	0,09	-0,01
Total_Rejected_Deposits	0,41	0,46	0,09	0,18	0,31	0,13	0,44	0,25	0,53	0,39	0,44	0,09	1,00	0,42	-0,03
Total_Rejected_WD	0,99	0,37	0,09	0,02	0,30	0,13	0,39	0,26	0,43	0,48	0,37	0,09	0,42	1,00	-0,11
datagbs	-0,10	-0,35	-0,01	-0,07	-0,03	-0,10	-0,32	-0,24	-0,30	-0,16	-0,38	-0,01	-0,03	-0,11	1,00

9.7. SAS ENTERPRISE MINER – FINAL DIAGRAM