NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Personalized Marketing Campaign for Upselling Using Predictive Modeling in the Health Insurance Sector

Andreas Melidis

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

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

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

PERSONALIZED MARKETING CAMPAIGN FOR UPSELLING USING PREDICTIVE MODELING IN THE HEALTH INSURANCE SECTOR

by

Andreas Melidis

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

Advisor: Flávio Luís Portas Pinheiro

February 2020

DEDICATION

Dedicated to my beloved family, Anestis, Christina, Daphne and Maya, to my friends and to the interesting people that I have met in my life.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all my Professors that I had the chance to work with and learn from them at NOVA IMS, to Leonardo Vanneschi whose teaching excellence is admirable, to Professor Flávio Pinheiro for his support in producing the final internship report, to all my colleagues, especially to Franklin, Joao, Jorge, Jessica, Susana and Bruno whose support along with their business knowledge, helped me integrate in the company and develop new skills.

I also take this opportunity to express a deep sense of gratitude to Rita Travassos and Bruno Rodrigues who trusted me and were always available when needed.

Last a special thanks to my manager and mentor Magdalena Neate. A person who inspires, dares to take a step beyond and cares. A person who really wants you to shine more than she does.

ABSTRACT

Nowadays, with the oversupply of several different solutions in the private Health Insurance sector and the constantly increasing demand for value for money services from the client's perspective, it becomes clear that Insurance Companies shouldn't only strive for excellence but also engage their client base by offering solutions that are more suitable to their needs.

This project aims, using the power that predictive models can provide, to predict the existing Health Insurance clients who are willing to move in a higher tier product. The case presented above could be described under the term of upselling. The final model will be used for a personalized marketing campaign in one of the most prominent bancassurances in Portugal. At the moment the ongoing upselling campaign, uses only few eligibility criteria.

The outcome of the model has as a goal to assign a probability to each client who is eligible to be contacted for this campaign. The data that were retrieved to train the model, had a buffer period of one week from when the 'event' took place. This is crucial for the business, because there is always the time-to-market parameter which should be taken into consideration in the real world.

The tools that were used for completing this Data Mining project were mostly SAS Enterprise Guide and SAS Enterprise Miner. All the Data Marts that were needed for the particular project, were built and loaded in SAS, so there were no obstacles or connectivity issues. For data visualization and reporting, Microsoft PowerBI was used.

Some of the tables in the Data Marts, are being updated in a daily and other in a monthly basis. Of course, all the historical information is being stored in separate tables, so there is no information loss or discrepancies.

Finally, the methodology that was followed for the implementation of the Data Mining project was a hybrid framework between the SEMMA approach as it is the one that is proposed by SAS Institute to carry out the core tasks of model development and CRISP-DM.

KEYWORDS

Upsell, Marketing Campaign, Insurance, Bancassurance, Predictive Models

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACV Annual Contract Value

AI Artificial Intelligence

CART Classification and Regression Trees

CRISP-DM Cross-Industry Standard Process for Data Mining

ID3 Iterative Dichotomiser 3

IG Information Gain

LoB Line of Business

MCAR Missing Completely at Random

MAR Missing at Random

NMAR Not Missing at Random

NN Neural Networks

NPS Net Promoter Score

SEMMA Sample, Explore, Modify, Model, Assess

4IR Fourth Industrial Revolution

SaaS Software as a Service

1 INTRODUCTION

This report covers the work conducted during the nine months internship in an Insurance Company. The task was to develop a propensity to upsell model for Health Insurance policyholders.

1.1. MODERN'S CUSTOMER CHARACTERISTICS

The digital revolution, also known as the Third Industrial Revolution (Kotler & Keller, 2006), has entirely changed the world of commerce of goods and services. In that process, it created a wide range of new potentialities both to customers and businesses.

If we zoom in to the modern customer, the new characteristics and possibilities that is empowered with, we will notice that a modern customer has:

- Greater consumer buying power as now, with the internet, they can quickly compare similar products from numerous suppliers across the world.
- Easier access to information. Consumers can now be sufficiently informed about products and services.
- Customized products and services. In order for the companies to exceed the needs of the consumers and be more competitive, they transform their products and services in a way to provide flexibility and give greater freedom to customers, design their product based on their needs.
- Various distribution channels and delivery points of the products/services. Nowadays we can state that there are not working hours for the enterprises, when these have an e-shop, as they are open 24 hours per day, 7 days per week. Buyers can access anytime and from everywhere. Moreover, buyers have the freedom to select where and how fast they want the products to be delivered.

It is becoming obvious that this fast-changing world of commerce and at the same time, the rapid advancement of the new technologies mandate companies to adapt into this new environment if they want to survive and make progress.

One of the sectors that is rapidly changing to meet the new needs of the consumers is the Insurance.

1.2 Positioning of Insurance Sector In Maslow's Hierarchy Of Needs

Abraham Maslow, in one of his papers in 1943 (Maslow,1943) with title "A theory of human motivation" posited that people possess a set of unconscious motivation systems that leads them to cover certain needs. These needs could be grouped in the following 'tiers'; starting from the lowest, meaning that it is the first that must be covered. (Figure 1)

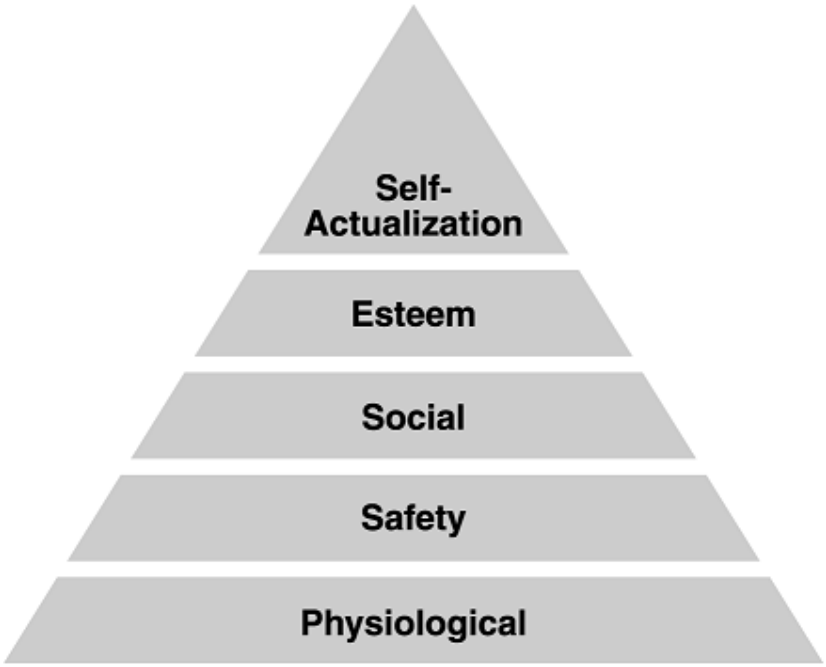

Figure 1– Maslow's Hierarchy of Needs (Image Source:http://www.shyentrepreneur.com/)

One way of obtaining safety and security is by purchasing an insurance policy.

As it becomes clear on one hand we have the Insurance companies knowing that an adequate sense of safety is vital for people if they want to climb in the next tiers of Maslow's pyramid and on the other hand, we have the customers, that now, more than ever are empowered with a vast amount of information and an oversupply of existing companies to choose among.

1.3 Maturity of Advanced Analytics in the Insurance Market

Nowadays, we are traversing the 4th Industrial Revolution(4IR). The IR4 concept was pioneered by Professor Klaus Schwab, founder and chairman of the World Economic Forum

(Schwab,s.d.). 4IR is characterized by the fusion of several technologies and the blurring of lines among the physical, digital and biological spheres.

Due to the high velocity and volume of the new emerging technologies, the need of evaluating the adoption level and the impact in the industries of these technologies was created.

MITRE, is a not-for-profit research corporation, which independently assesses technologies for readiness. Technology Readiness Assessment (TRA), initially developed by NASA, is a metrics-based process that analyzes the maturity of critical technologies.

According to MITRE, assessing the maturity of a particular technology involves determining its readiness for operations across a wide spectrum of environments with a final objective of transitioning it to the user. Application to an acquisition program also includes determining the fitness of a particular technology to meet the customer's requirements and desired outcome for operations.

MITRE's technology maturity s-curve includes the following categories:

1. New or Emerging Technology: Has not reached the tipping point
2. Improving Technology: In the development stage
3. Mature Technology: Second tipping point before the curve turns down
4. Aging Technology: When the downward tail begins

In Accenture's "Broker of the Future" report (Reilly, 2016), they adapted the s-curve to commercial broker technologies.

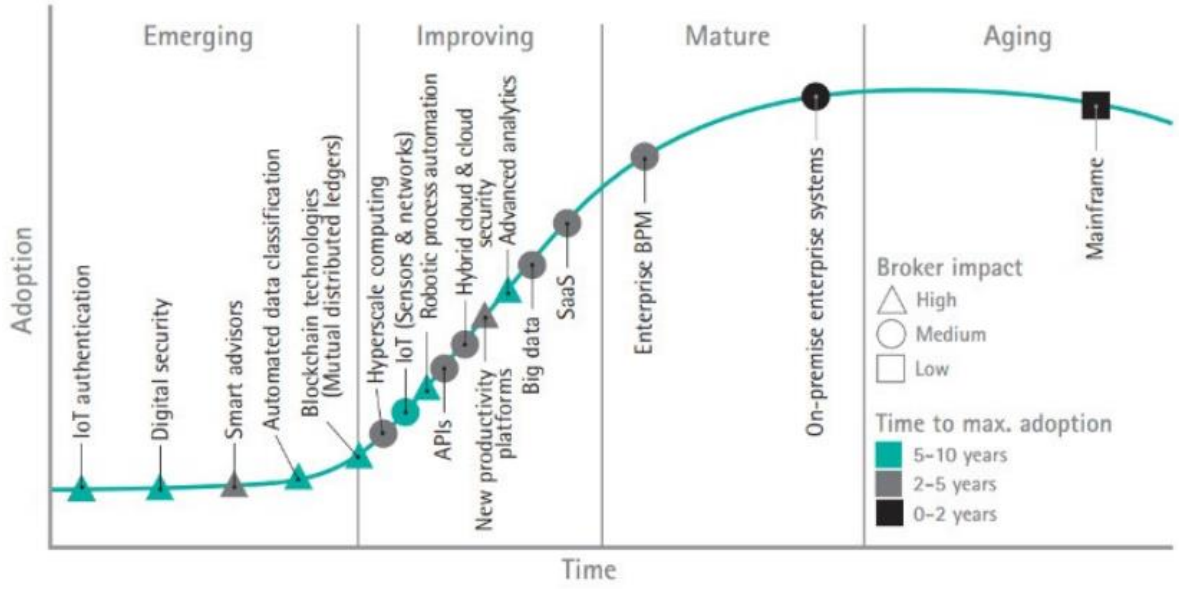

Figure 2– MITRE's technology maturity s-curve (Image Source https://insuranceblog.accenture.com/tech-maturity-s-curve-for-insurers)

It is interesting for the purpose of this report, to focus on the field of Advanced Analytics and the adoption time it takes for the Insurances/Brokers to embrace it, so as the impact that the field of Advanced Analytics has in this sector. As it can be seen (Figure 2), the impact is High, the level of maturity is in the improving stage, approaching the mature state, and the estimated maximum adoption time is between 5 to 10 years.

1.4 The Multiple Benefits of Upselling

As it has been briefly described, the aim of the predictive model was to identify all the existing clients who were more likely to upgrade their product within the company's available products.

The benefits of upselling are manifolds (Figure 3) and diverse for the company and could be seen from several different perspectives. Sometimes these are strictly connected with the extra revenues that the company gets from this action, but sometimes there are side effects of upselling which could be perceived by the company, as even more important than the higher generated revenues.

Similarly, upselling has positive impact for the client who is taking the action and often it is interrelated to the greater relevancy of the possessed product.

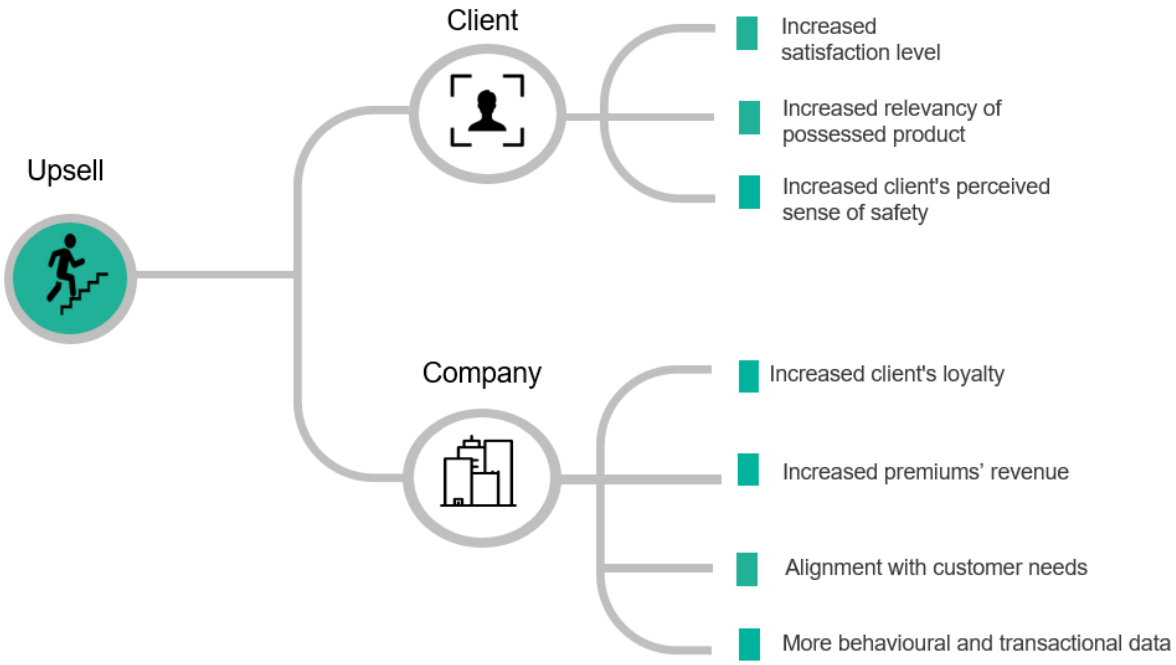

Figure 3– Benefits of Upselling

1.4.1 Business Case – Net Promoter Score

In 2003, Frederick Reichheld introduced in Harvard Business Review the concept of NPS (Reichhald, 2003). He claimed that this single number between 0 to 10, which is the product of several answers of a customer's questionnaire, is adequate for profitable measuring and managing customer loyalty.

- Promoters - people who score from **nine** to **ten**
- Passives - people who score from **seven** to **eight**
- Detractors - people who score from **zero** to **six**

Companies want to improve their revenues –not from the sales of the main service/product- but by offering to the customers enhanced features, subsidiaries to the main.

A business case (Caderholm, 2014) of how upselling could advocate to an enhanced customer experience is presented below. This case refers to JetBlue, an airline company that monitors the customer's experience and satisfaction through NPS.

As it can be seen below (Figure 4) most of the ancillary services of JetBlue had a positive NPS. What is also noticeable is that services with higher NPS are also the most profitable for the company.

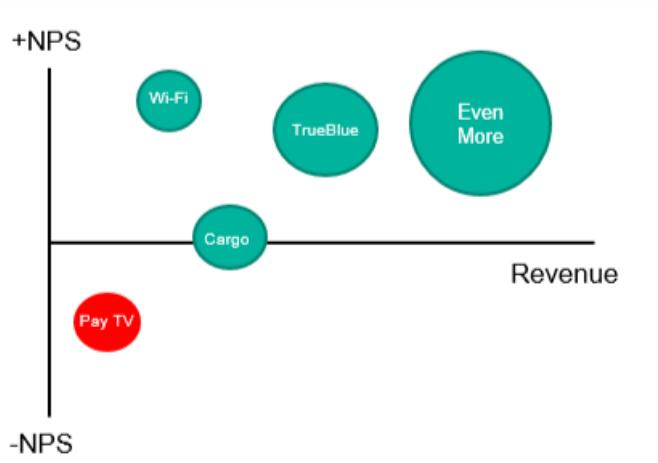

Figure 4– NPS of Upselling

Size of the bubble denotes the relative margin contribution

1.4.2 Business Case – Cost Comparison and Revenue

In the fourth edition of the annual survey that Pacific Crest Securities conducted (Skok, 2015), an investment banking firm focused on Software as a Service companies (SaaS), there was a section, comparing the Customer Acquisition Cost to the Upsell Cost and Renewal Cost.

The survey's responses were answered by senior executive members of the board of 305 SaaS companies.

Some more details of the survey are presented below

- Median of 4MM $ revenues, with 133 companies >5MM $ and 57 >25MM $
- Median of 47 Full-Time employees (range of 2 to 1,200)

- Median customer count of 300; 28% of the recipients of the survey have >1,000 customers

- 70% of participants are located in United States

The responders had to answer the following question: *"How Much Do You Spend for $1 of New Annual Contract Value from a Customer?"*

The main conclusion of the aforementioned question was that the median cost to acquire a new client (1.18$) was 3.2 time higher than the cost for making a customer to upsell (0.28$). The lowest cost is sourcing from the Renewals where the median cost is only 0.13$.

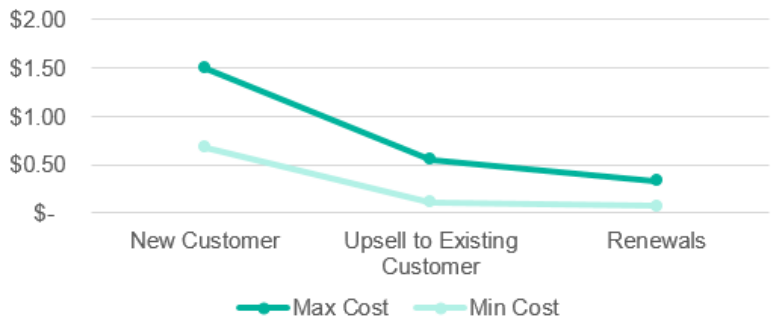

Figure 5– Cost of Upselling VS New Business and Renewal

For the reason of cutting off operational and marketing costs, companies now invest capitals to create efficient ways of identifying clients who are more likely to upsell.

In the same survey, the responders were also asked to answer what percentage of new ACV was coming from upsells to existing customers.

As it is presented in Figure 6, there is an upward trend in the percentage of new ACV that is sourcing from upsells, as the financial size of a company is increasing.

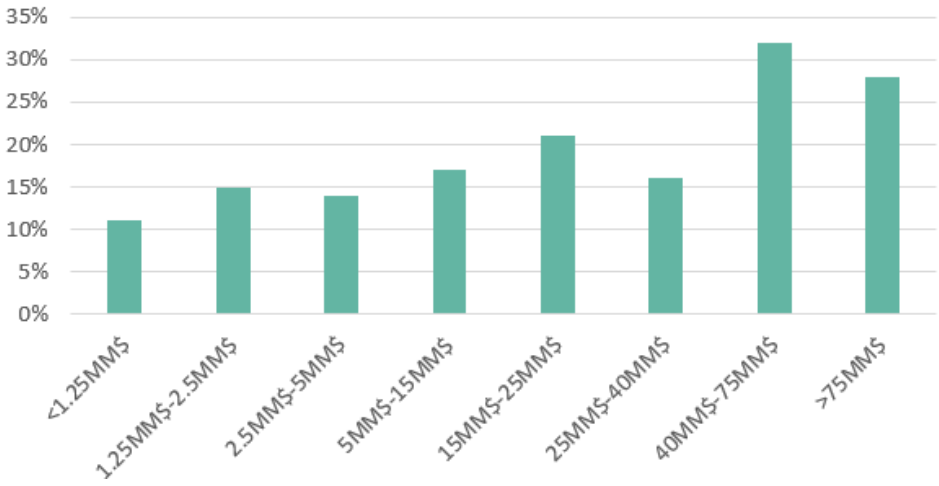

Figure 6– ACV from Upselling

2 THEORETICAL FRAMEWORK

Developing and implementing a predictive model, is only a part in the Data Mining process. In every Data Science project, there are few anterior steps before the modeling phase. Thus it is important to present the main Data Mining methodologies and frameworks as they are described in the literature so as explain the fundamental reasons which led the scientist and businesspeople to create these frameworks/ methodologies.

In addition, the main data preprocessing steps will be presented as well as some of the core and most widely used modelling techniques.

2.1 DATA MINING METHODOLOGIES

The known and well-defined univariate and multivariate statistical methods ceased to be used solely, as they are not sufficient for the large-scale databases and the complexity of the problems that academia and businesses are facing. However, quite frequent are being used as the starting point to prepare more methodologically complex models, by using sophisticated tools and often by using achievements in the field of Artificial Intelligence (AI). In the beginning of the 21st century, the Data Mining techniques are increasingly used, mainly for business and research purposes.

Along with the booming of AI and the several Data Mining techniques, companies have to ensure that the micro and macro-management of Data Mining projects will be executed in a way that will enable all the components that are needed to advocate the efforts to achieve the desirable results. As a result, this led business experts and scientist to create frameworks and methodologies which could facilitate to better manage these kind of projects, by using standardized processes.

The following table (Table 1) is a comparison of selected Data Mining methodologies in three main aspects of knowledge discovery (Rogalewicz & Sika, 2016)

Table 1- Data Mining Methodologies

DM Methodology	Pre-Processing	Main-Processing	Post-Processing
CRISP-DM *CRoss-Industry Standard Process for Data Mining*	Business Understanding Data Understanding Data Preparation	Model	Evaluation, Deployment
KDD *Knowledge Discovery in Database*	Selection, Pre-Processing Transformation	Data Mining	Interpretation and Evaluation
SEMMA *Sampling, Exploration, Modification, Model, Verification*	Sample, Explore, Modify	Model	Assess
VC-DM *Virtuous Cycle of Data Mining*	Identify Transform (Pre-Processing and Main-Processing)		Act, Measure

In this report, CRISP-DM and SEMMA will only be discussed, as a hybrid solution among these two was used for the execution and implementation of the project.

2.1.1 CRISP-DM

A Data Mining methodology, which is broadly used in practice, is known as the CRISP-DM (CRoss-Industry Standard Process for Data Mining).

According to the CRISP-DM concept, the lifecycle of a Data Mining project is consisted by 6 stages (Figure 7).

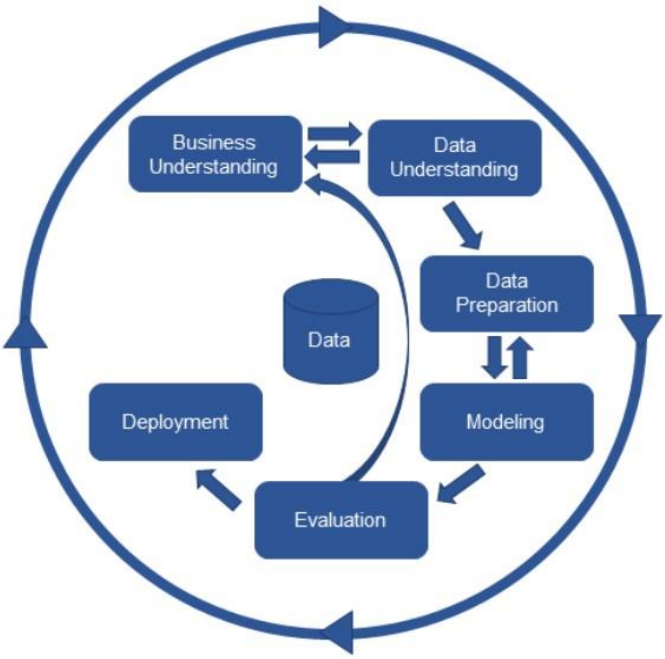

Figure 7– CRISP–DM Methodology (Image Source: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html)

In CRISP-Data Mining, particular attention is paid in comprehending the business context. It is visible by synergic linkages between the first three steps, which can be perceived as the pre-processing stage. The next main- processing stage is the modeling, while the last two advocate in the assessment of obtained results and implementation of results acquired on the basis of modeling.

- **Business Understanding**: Most of the business data mining projects aim to provide a solution to a business problem. Thus it is fundamental to have a clear understanding of company's objectives, the AS-IS situation and the vision in order to facilitate the plan of actions for the specific problem in the next steps.

- **Data Understanding**: This step incorporates the tasks of data understanding and data collection. Emphasis is given to data quality verification so as to a first preliminary data exploration.

- **Data Preparation**: In this stage, data cleaning techniques and data transformation methods are applied to prepare the dataset in a format that can be used as input in the modeling phase. A more thorough data exploration is carried out, providing an opportunity to discover patterns.

- **Modelling**: The modeling stage uses data mining tools to apply algorithms suitable to the task at hand. Depending on the complexity and the time assigned to this task, several sophisticated algorithms can be applied.

- **Evaluation**: Once the modeling phase has been finalized, an evaluation of the constructed models starts by comparing the performance of the models using several evaluation metrics appropriate to the nature of the problem. The outcome of this phase is the selection of the champion model.

- **Deployment**: The final step of CRISP-DM encompasses the communication and deployment of model's outcome. Business stakeholders are being informed about the main insights, difficulties and caveats of the project. In parallel the plan of deployment begins, which usually is a synergy among several teams of the company (IT, Business Automation, Marketing, etc.).

2.1.2 SEMMA

SAS Institute has developed its own, house-in data mining process which incorporates 5 main steps with the acronym SEMMA (SAS Institute Inc, 2017) which stands for *Sample, Explore, Modify, Model, Assess*. According to SAS Institute, this process is applicable across a large variety of Industries and provides these methodologies that enables companies to solve the most complicated business problems such as fraud detection, customer retention, personalized marketing, market segmentation, risk analysis and more. The process has been designed to be executed iteratively.

The steps of the SEMMA approach are being described.

- **Sample** the data by creating one or more data tables. The samples should be sufficiently large so as the incorporated information to be representative to the whole dataset, but small enough to enhance the performance of the process.

- **Explore** the data by searching for relationships, unanticipated or anticipated trends and distribution anomalies in order to gain a deep understanding of the dataset.

- **Modify** the data by creating new features, selecting good predictors, and transforming existing variables to facilitate the model selection process.
- **Model** the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.

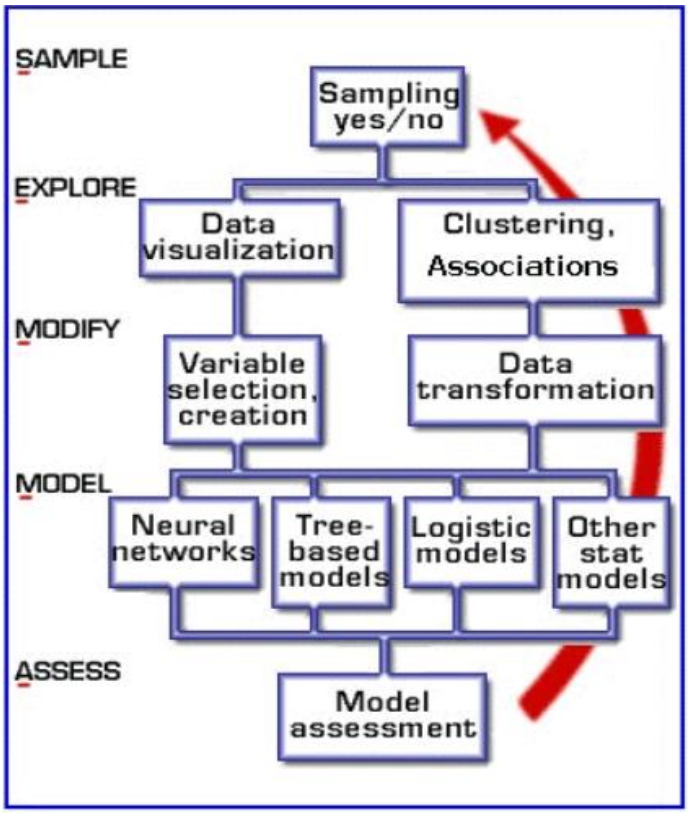

Figure 8– SEMMA Methodology (Image source https://documentation.sas.com/)

2.2 DATA QUALITY

A lot of times, data is far from what is perceived as adequate to be analyzed and used in a Data Mining project. Nowadays, most of data mining techniques can handle some level of imperfection in the data. However, a focus on understanding and improving the data typically improves the quality of the resulting analysis.

Data quality main issues, that usually should be tackled, include the presence of noise, missing or duplicated records and outliers.

According to (Han, Kamber, & Pei, 2017) the quality of the data can be measured taking into account different factors such as *accuracy, completeness, consistency, timeliness, believability and interpretability.*

The reasons why these quality issues exist are manifold as explained below.

- Measurement and data collection errors:
    - The measurement problem results from any measurement process. It is common that the measured value of an object can deviate in an extent from the real value that should have been inserted.
    - The data collection error results while omitting data objects that are inappropriately placed in the dataset.

  Both of these issues could appear either systematically or randomly.

- Incomplete, inaccurate and inconsistent data issues are related to:
    - Data collection instruments that may be faulty. An example for this particular problem is while a human is making the data entry in the system, to accidently insert a wrong value. This error can cause:
        - Inconsistency, as the value will not fall in a predefined set of acceptable values.
        - Noise, as the value may deviate by far, from the median or average estimated value.
    - Missing Values and disguised missing values.
        - Regarding the missing values there are two types of missing values that can be detected. First when it is the intention of the subject not to provide a certain information (i.e. age, place of birth) and second, missing due to inapplicability (i.e. email address, in case that the subject doesn't have an email address).

- Disguised missing values refers to the inaccurate values that intentionally or unintentionally a subject insert to the system (i.e the default value in a questionnaire regarding question about political beliefs is '*I prefer not to say'*) .

- Timeliness refers to the issue of having the data available in the moment it is needed.

  Thus it is crucial to for the business to be aligned with the Data Management Department, regarding the regularity that the systems should be updated.

- Believability affects the level of data trustiness from its users, while interpretability refers to the easiness from the user perspective to comprehend the content of the features.

2.3 Missing Data

An important analysis that precedes the treatment of missing data and should be considered in most of the data mining projects is the nature of the missing values themselves.

The missing data can be categorized according to the relationship between the likelihood of data to be missing and the values of the data, both missing and observed.

What it is tested and examined is the likelihood of data to be missing due to a systematic or a non-systematic reason.

2.3.1 Nature of Missing Data

Missing at Random

First, regarding MAR cases, whether an observation is missing, it is irrelevant and can't be associated with the missingness of the value.

Missing Completely at Random

In this case of missingness, there isn't any relationship between the fact that the data is missing and any values, observed or missing.

It should be emphasized that MCAR can be characterized as a non-systematic behavior, as there is nothing that makes some data more likely to be missing than other.

Not Missing at Random

This is the case where the propensity of a value to be missing, depends on its values. Thus we could describe this type as systematic pattern.

Perhaps, this group of non-responders, feel uncomfortable to answer a specific set of questions. For example, people with low education may have a higher propensity to leave empty a question regarding the level of education.

Another example of Not Missing at Random scenario, which was designed within the scope of this report, is presented in Table 2.

Table 2- Missing Values Example (A1)

ID_Client	No of Cigarettes per day	No of Cardiology Consultations
A	0	2
B	28	12
C		13
D	0	1
E	1	1
F		21

Sometimes it could be intricate to distinguish whether a case refers to Not Missing at Random, Missing at Random or Missing Completely at Random.

Therefore, a good practice is to compare the means of an attribute, segregating by missing and non-missing values associated to the desired to test column. If there is a significant deviation between their mean then this should be perceived as an evidence that missings could characterized as Not at Random Missings.

In the hypothetical scenario that was presented in the Table 2 the following results have been calculated (Table 3) which indicates a pattern for the people who didn't respond as the averages deviate significantly.

Table 3- Missing Values Example (A2)

Responded to the Question	Average No of Cardiology Consultations
Responder	4
Non-Responder (Missing)	17

2.4 OUTLIERS

"An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" (Hawkins, 1980)

The main problems with the outliers are that can inflate the error variance, distort the parametric or/and non-parametric tests and eventually lead to faulty conclusions.

In this section the following topics will be analyzed.

- Which are the reasons that can cause infiltration of outliers in a dataset
- Which are the methods that can be used to detect outliers
- In which cases outliers should be removed

According to Anscombe (Anscombe, 1960) the reasons that could create outliers can be sorted into two main groups. Those that the source is the error in the data and those that arising from the inherent variability of the data.

Outliers from data errors. As it was reported in previous chapter, the quality of the data in the dataset it is extremely important so as the model's results to be trustworthy.

Nonetheless, poor data quality datasets consist a common phenomenon, and it can be the result of a human error during the collection, recording or entry of the data into the system.

These human related errors, can introduce outliers in the dataset. For example, while a user is inserting new tuples of data, may not notice that miskeyed a number that eventually will result to infiltrate an extreme value in to the system.

Outliers from intentional misreporting. Sometimes, it is possible that participants to surveys or to experiments, provide intentionally and consciously inaccurate and faulty information only for sabotaging it (Huck, 2000) or for other, personal interests.

Outliers from sampling errors. An additional reason of the outlier phenomenon could also be due to the sampling method that was selected when balancing the dataset. It is unlike but still possible to happen, that randomly selected observations may incorporate more outlying values than expected.

Outliers as valid natural variation. In this case these data points are legitimately in the dataset.

The detection of outliers is a critical data mining task which among the researchers and scientists has triggered a debate of what constitutes an outlier as well as if an outlier should be removed or not.

Univariate outliers can be detected graphically by using the boxplot (Tukey, 1977).

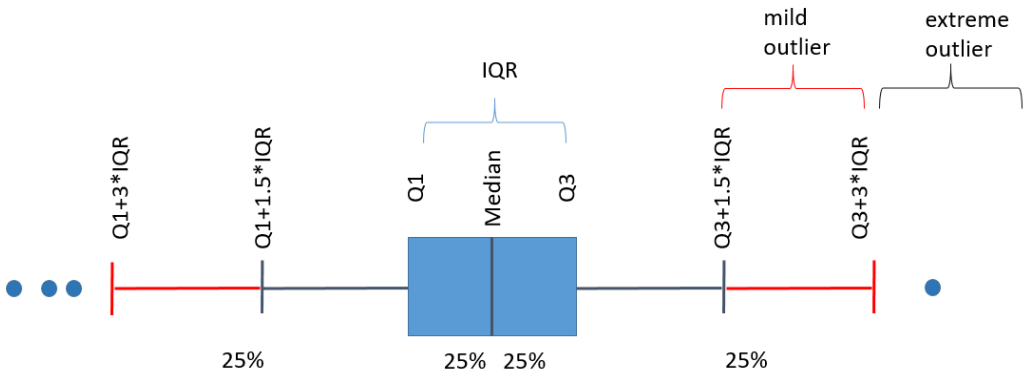

Figure 9– Boxplot

In a boxplot graph, there are two types of outliers which can be distinguished. The minor or mild outliers and the extreme outliers as it is displayed in Figure 9

There are some simple rules of a thumb for detecting outliers, such as z=3 that can be applied and it is relatively effective. However, (Miller, 1991) stated that this method of eliminating

data points which are 3 standard deviations far from the mean could only be effective with specific distributions which are approximating the normal.

With high skewed distributions, skewness could mask the outliers, so this method is not indicated.

Multivariate outlier detection takes into account at least 2 attributes or more. Multivariate outliers cannot be detected by comparing each feature's boxplot, separately.

An example (Figure 10) is given in order to better understand how and why this is happening.

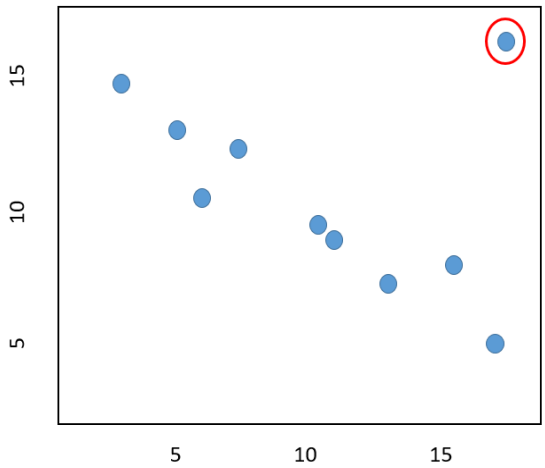

Figure 10– Multivariate Outliers

In the Figure 10 the instance that is located in the upper right corner (circled) can be considered as a multivariate outlier.

However, the same instance in a uni-dimension projection, wouldn't be considered as an outlier, as neither in the X-axis attribute nor in the Y-axis attribute is deviating significantly from the mean.

Multivariate outliers can be detected by using several methods that have been developed such as:

- Statistical based
- Clustering based
- Distance based
- Density based

There is a great deal of debate regarding whether or not an outlier should be removed. According to (Osborne, 2004) the first parameter that should be reviewed is if the outlier is legitimately in the dataset. In case that observations are incorporated illegitimately then the common sense says that they should be removed.

In the opposite scenario, meaning that outliers are legitimately incorporated in the dataset, many opinions have been reported and several reasonings behind them.

First, it has been reported that elimination of outliers, tends to improve the accuracy of the estimators. Moreover, Judd & McClelland (Judd & McClelland, 1989) claim that removal of outliers can provide a more honest estimation of population's parameters.

In case that it is decided to remove the outliers, then all the aforementioned methods can be used to detect and eventually delete them.

In case that deletion of outliers is not desirable for various reasons (e.g. loss of information) then either transformations can be applied to the estimator, especially when the distribution is highly skewed, or truncation.

By using transformations, extreme scores can be kept in the data set, and the relative ranking of scores remains, yet the skew and error variance present in the variable(s) can be reduced (Hamilton, 1992).

When transformation of data is selected as an appropriate solution, then there are two limitations that should be evaluated and considered.

First, transformed variables (e.g. log transformation) change the scale and as a result, it worsens the interpretation between the new transformed variable with the dependent variable.

Furthermore, another problematic issue which confines the broad usage of transformation as a solution to outlier treatment, is that many of the common and widely used transformations do not accept negative values.

Truncation can be seen as an alternative to deletion or transformation. The user recodes the variable which contains outliers and replace the detected outliers with the highest or lowest reasonable value. However, this method is depending a lot to subjective matters as the user

should define which should be the lowest and highest value. Of course statistical methods can be used to define the lower and upper limits.

2.5 Predictive Modeling

"Predictive modeling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or "dependent" variable and various predictor or "independent" variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable"

(Dickey, 2012)

Nowadays, the task of predictive modeling is considered as the enabler of transforming data into knowledge and actionable insights. For many people, and especially people who work closely to business, predictive models are perceived as the process which outcomes to decision-making tool, and consequently allow them to take action based on data-driven conclusions.

Predictive modeling is in the core of a Data Mining project. It aims to construct a concept model that can express the variable that we want to predict (depended variable) as a function of the independent variables. The goal of supervised predictive modeling is to minimize the difference (error) between the predicted and real values.

The representation of the model is consisted by a set of parameters (explanatory variables, operators and/or constants) organized in a structured way. Predictive modeling is the process of training the dataset's instances, by using the input variables (estimators), trying to identify and extract patterns and rules, which sequentially will fit the new unseen data as accurate as possible. The instances that are used to build the model are consequently named as training set.

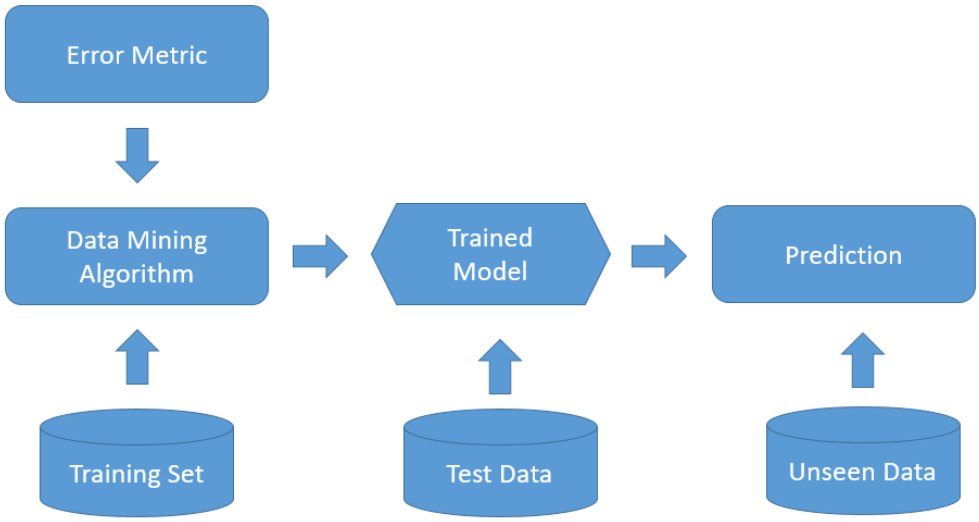

Figure 11– Predictive Modeling Process

3 METHODOLOGY

In this chapter the methodologies that were applied to deploy the project will be discussed in detail.

3.1 DATA MINING PROCESS METHODOLOGY OF THE PROJECT

In this project a hybrid Data Mining methodology which combines the CRISP-DM and SEMMA was used, as CRISP-DM provided the important for the project, step of Business Understanding, and at the same time SEMMA provided a well-defined methodology, designed by SAS Institution.

In Figure 12 it is presented the plan of actions and sub-tasks which were taken in this project.

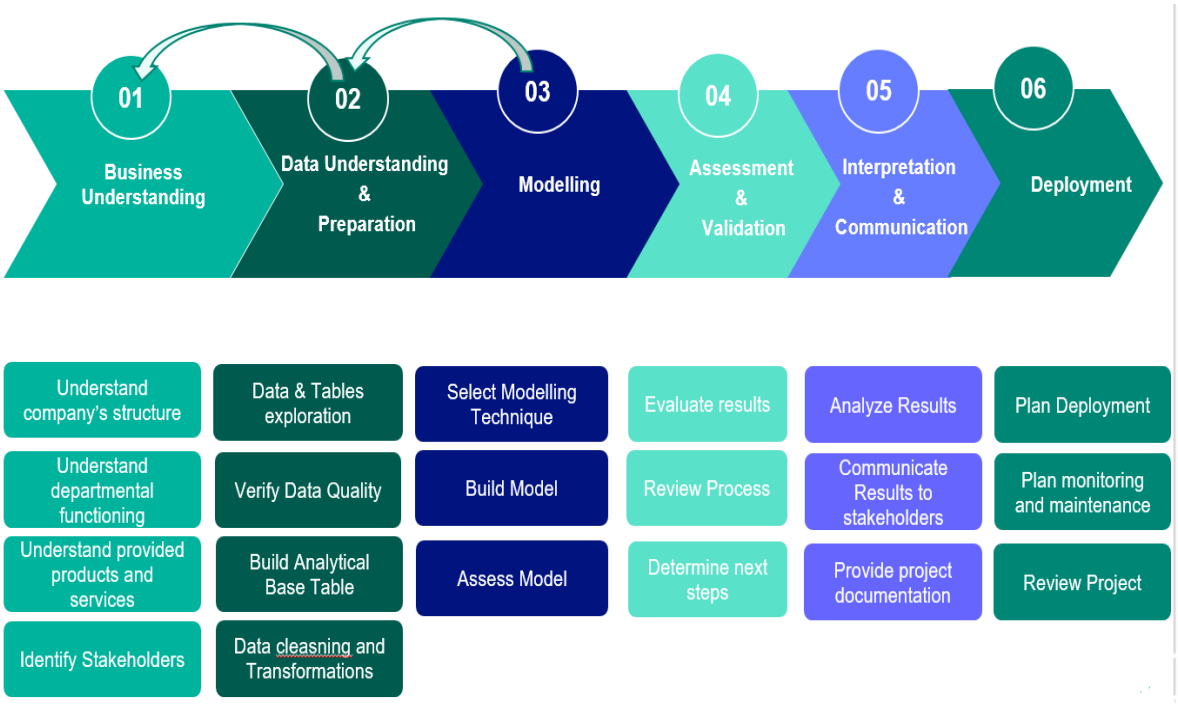

Figure 12– Project's Data Mining Methodology

3.2 PROJECT ROADMAP

As this document is in the context of an Internship, it is worth to present the project's roadmap with all the essential steps that were planned to be executed, the sequence that would be followed as well as the time that had been agreed to be allocated in each one of these steps.

As it can be seen in the following timeline chart (Figure 13), there was a defined project time plan for the implementation of the project. However, in the business context there are endogenous and exogenous factors that don't allow us to always follow it strictly.

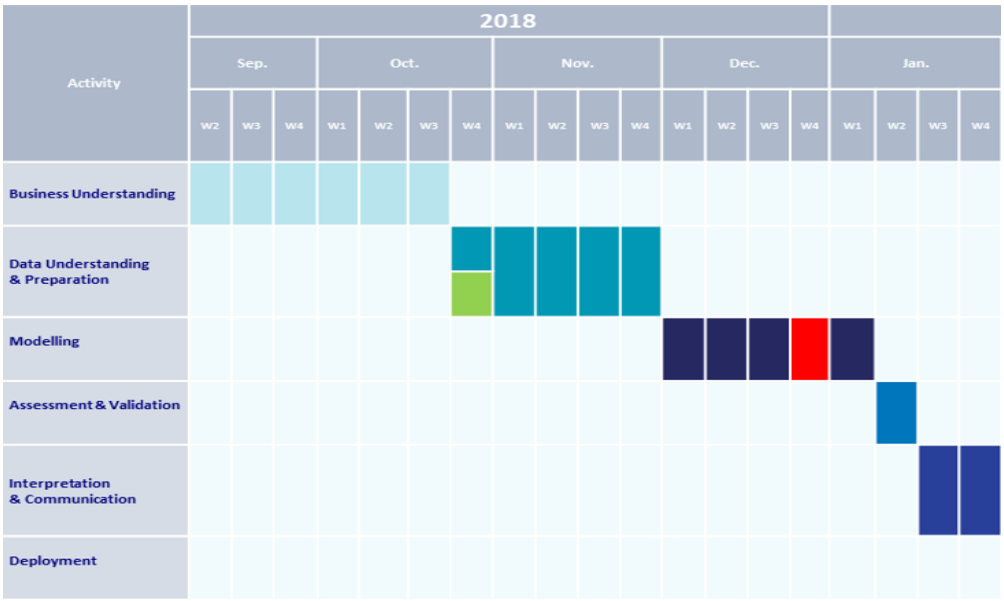

Figure 13– Project's Timeline

Thus sometimes the initial plan has to be adjusted in order to meet the final deadlines.

In the frame of this internship, as it was presented in Figure 13, there was an expected timeline for each one of the phases that were defined in the beginning of the project. However, as it is presented in the Figure 14, the expected timeline didn't match 100% the real timeline.

As the business is an 'alive' entity and deadlines of the final product must be respected, the element of flexibility and adjustability should be present and partial tasks should be executed quicker if needed without sacrificing quality.

There were several explanations why the project faced this mismatch. The main reasons are listed in Table 4.

Table 4- Data Mining Planning

Data Mining phase	Execution Time	Reason
Business Understanding	Delayed	• Lack of domain expertise • Availability of the stakeholders for interviewing
Data Understanding	Delayed	• Lack of domain expertise • Complexity of Data Storage • Technical issues with the software
Data Preparation	Delayed	• Vast amount of data • Exploration of several Data Cleaning and Transformation Methods
Modeling	In advance	• Company's standard processes and knowledge
Assessment	On-time	
Deployment	In advance	• Efficient communication with the IT Dept. , the Marketing Dept. as well as the distribution channels

In Figure 14 it is presented the percentage of the time assigned to each phase versus the actual.

Figure 14– Expected Vs Real Timeline

3.3 Business Understanding

The initiation of the project began with the Business Understanding. This phase focuses on comprehending the objectives and requirements from the perspective of the business, translating these factors into a data mining problem definition and designing a preliminary plan to meet these objectives.

The first step was to understand the available distribution channels of the Health Product Line. In particular, it is necessary to gain an understanding of the existing channels and the sale scenarios that a Health Insurance policy can be issued. Figure 15 displays these channels.

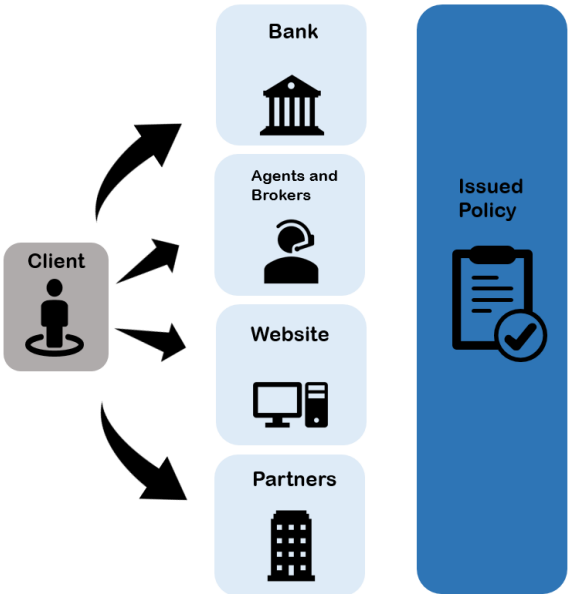

Figure 15– Process of Issuing a policy

Client or prospect client can issue a new policy by any of the following distribution channels; Bank, Agents and Brokers, Website, Partners.

However, there are some significant procedural differences, which affect mainly the stage of underwriting a new policy.

How a policy can be issued across the different channels?

- Bank, Agents and Brokers and Partners: Only by physical presence
- Website: Only by the World Wide Web

A significant proportion of time spent in Business Understanding phase, was specifically regarding the Upsell Model. At this point of the project, it is important that the bigger picture of the expected deliverables, processes that will be considered and outcomes, to clearly be defined, so as misunderstandings in the future to be minimized.

Moreover, additional technical and conceptual issues had been addressed during this stage, such as (1) rules for creating the Target Universe (2) period to be considered for modeling, which will be discussed in later chapters.

As the output of the model would be used in the context of a Marketing campaign the scope of the model should be adjusted in this frame and not be treated isolated.

Thus an AS-IS and TO-BE abstract business process was developed describing the core changes that would occur in the designing, implementation and execution of the marketing campaign, which can be seen in Figure 16.

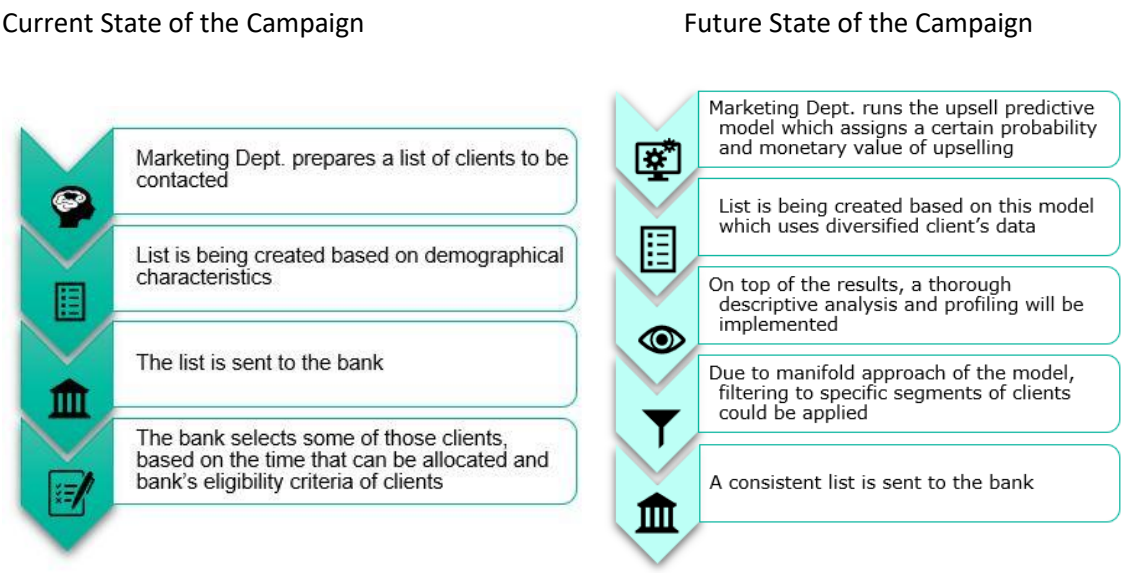

Figure 16– AS–IS Campaign's Implementation VS TO–BE

Moreover, in this step of the Data Mining process the deliverables were discussed and confirmed between the Data Mining Department and the Commercial Performance Department so as the campaign's KPI's which would be used as an evaluation method of the model's performance.

In the following table (Table 5) the deliverables are being displayed

Table 5- Deliverables

Deliverable	Format	User	Description
Predictive Model	SAS Enterprise Miner	Marketing & Data Mining Dept.	The winning model should be extracted in a format that is reusable
Documentation – Main Insights of the model	Word Processor	Stakeholders (Business-oriented)	Interpret the results of the model in a simple way for the Stakeholders
Documentation-Technical Report	Word Processor	Marketing & Data Mining Dept.	A step-by-step technical report
Process of generating Campaign's leads	SAS Enterprise Guide	Marketing & Data Mining Dept.	A process in which scoring and hygiene rules are applied to generate the leads, in a weekly basis

In addition, several interviews with business experts and stakeholders were held in order to listen to the domain experts' opinion regarding the factors that could predict an Upsell in the Health LoB as well as the main triggers.

One of the questions that were called to answer was the following:

"In your personal opinion which are the main characteristics of the clients and main triggers that make them upgrade their Health Insurance policy?"

The answers that were given had a qualitative nature which were summarized in the following groups (Figure 17)

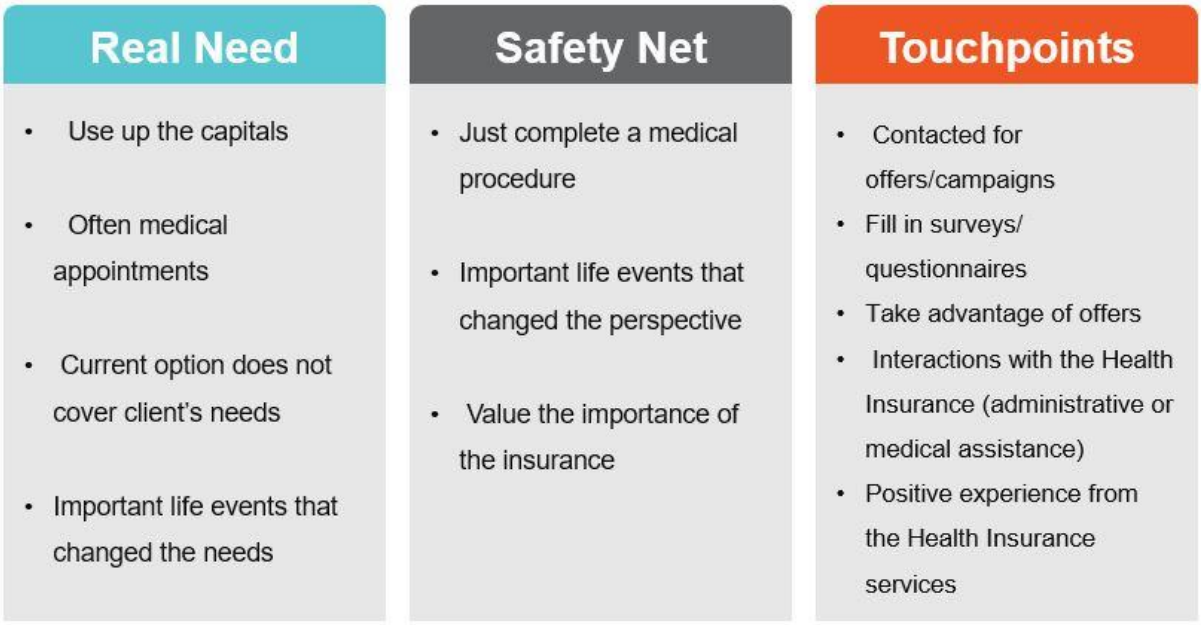

Figure 17– Interviews Insights

By asking these questions, the objective was to decompose the meaning behind the answer, and "synthesize" where needed, variables that would capture the meaning of the answer in a quantitative way.

3.4 Data Understanding and Preparation

Data Understanding step involves taking a closer look at the data available for mining. This is critical in order in later steps to avoid unexpected situations.

This step incorporates accessing and exploring the data by using simple statistics and graphs that visualize them in univariate, bivariate and multivariate way

On the other hand, today's real-world databases are highly susceptible to noisy, missing and inconsistent data due to their typical huge size (Han, Kamber, & Pei, 2017). When the phenomenon of low quality exists, then it is likely to lead in low-quality mining result. There are several techniques that take care of this matter as it will be presented in this chapter

In Figure 18, it is illustrated the Data Understanding and Data Preparation steps and the core tasks that were carried out in order to complete these two main Data Mining processes.

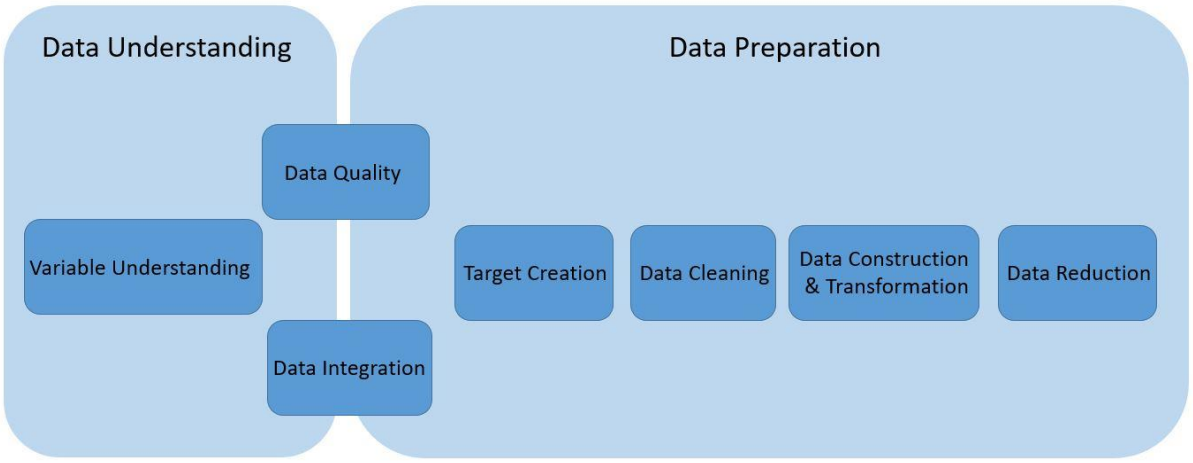

Figure 18– Data Understanding and Preparation Flow

3.5 VARIABLE UNDERSTANDING

First important task that was executed, was the deep understanding of the available variables in the company's data marts, by accessing and exploring the tables in the numerous available data marts. In parallel there was an undergoing process of grouping and checking the availability of specific variables which were indicated from the business experts as potential good predictors.

These group of variables can be listed as following

- Demographical
- Behavioral
- Touchpoints between Insurance and Client
- Product Possession & Usage

It was crucial and at the same time insightful that in the preliminary stage of the Data Mining project, the available data would be described, categorized and analyzed in a meaningful way from various perspectives, so as a high comprehension could be achieved.

Particularly, in this project, first the available to the pre-existing Analytical Base Table variables, were grouped in a meaningful way and the context of these groups was described in

detail so as it would be comparable to capture the degree of match between the business intuition and the existing available variables.

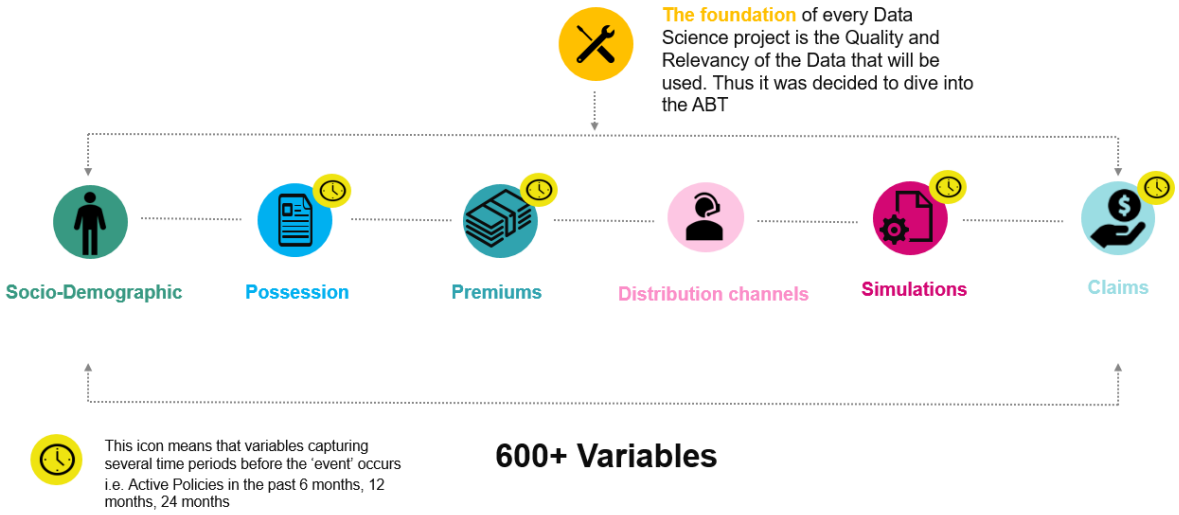

Figure 19– Analytical Base Table

Table 6- Analytical Base Table

Variable Names	Description
Socio-Demographic	
Age Marital Status City Gender etc	This group of variables incorporates basic information about the client such as Age Group, Sex, Profession and Segmentation
Possession	
Ind_has_Auto Ind_1Yr_Issued Ind_has_VidaRisco No_Active_MultiRisco etc	A set of variables which captures the journey of client across Lobs, checks how many active Policies in each LoB the client has, and everything around possession of policies
Premiums	
Premiums_Paid_1Yr	Premium group of attributes, contains

Premiums_Paid_Auto_6Mths Sum_Premiums_Inactive etc	information about the amount paid in each Product Line
Distribution Channels	
Ind_BankClient Ind_PartnerCLient Sum_BankClient_Products etc	A group of variables which captures the real relation between customer and distribution channels, using several ratio variables
Simulations	
Ind_Simulation_Converted Ind Simulation_Not_Converted Ind_Simulations_1Yr etc	Information regarding the Simulations that the client had across the company from all the available channels
Claims	
Ind_Çlaims_1Yr Val_Çlaims_1Yr Val_Claims_Auto_6Mths etc	Information regarding the claims that the client had in the company in all the possessed products

3.6 Data Quality Of The Data Marts

There is a business question and a concern that is raised, which is spread around the Data Experts of big organizations, that refers to the monetary cost of possessing poor data in terms of quality.

Throughout the last few years, big consultancy firms and private research groups have conducted surveys regarding the aforementioned subject.

Below are presented the results of several surveys in a form of infographic

Figure 20– Data Quality in Numbers

Notice that the purpose of data mining is not to prevent the data quality problems, but it focuses on the (Pang-Ning, Steinbach, & Kumar, 2006)

1. detection and correction of the data and
2. use of algorithms which can tolerate poor data quality.

In this project, the quality of the data can be described adequate and therefore the trustiness level was relatively high.

3.7 Data Integration

Large and complex corporations usually tend to also use a data model and storage structure that is complex, which subsequently put in danger the data integrity.

However, nowadays, a significant proportion of enterprises' yearly budget is allocated to Data Warehousing applications. High levels of user satisfaction and return on investments have been reported by using such applications (Graham, Coburn, & Oleson, 1996).

The Insurance company has carefully invested in the top-down approach and developed a centralized Data Warehouse for all the LoB's.

Moreover, dependent data marts have been constructed for every single LoB. Thus it is ensured that the calculated variables across the company are following the same predefined rules eliminating the discrepancies.

In this project, 41 source tables across 4 Data Marts had to be joined, using as a preliminary key the ID of the Health policy concatenated with the NIF of the client, which is unique.

3.8 TARGET DEFINITION AND CREATION

It is crucial for the success of the data mining project to correctly define the Target Universe. By Target Universe, in the specific case of this internship, it is meant all the clients of the Health Insurance, that should be incorporated in the final dataset, before proceeding to the modeling phase.

The target variable was binary, distinguishing the clients who upgraded their Policy from those who didn't, during the selected time period.

3.8.1 Target Universe Creation

More specifically, below are presented the main filters that were applied in the whole universe in order the Target Universe to be created

Individual clients

- Clients within a certain age range
- Clients with an active Health Insurance policy within the given time period
- Clients who possess only one Health Insurance policy
- Clients who possess a policy within a certain set of Options

These filters were applied horizontally to the client base in the selected time period.

Following these hygiene rules applied universally, the target variable could be constructed following the defined rules of what should be considered as an upsell.

In the beginning, business experts had described the upsell action under four different scenarios as presented in Figure 21.

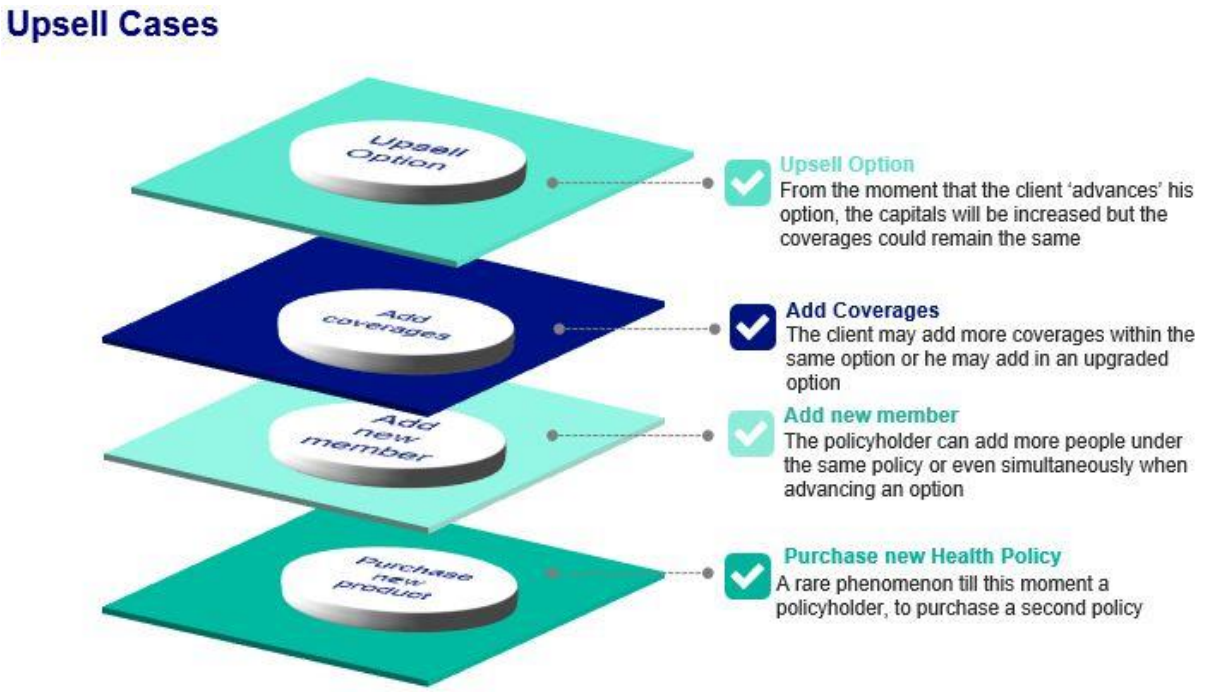

Figure 21– Upsell Cases

In the end, a decision was made by the stakeholders who proposed a single model to be constructed only by considering the propensity to upsell by advancing the Option, as it was believed that the triggers for the remaining types of upsell had different characteristics and the customers different needs.

3.8.2 Target Variable Creation

In the next paragraphs, it will be presented, step-by-step how the target variable was built.

First all the acceptable Options that were in the Target Universe were collected in a single table. It was detected that there were values in the variable which incorporates the names of the possessed by the client Option, that were referring to the same Product but it was spelled differently. This is a case of poor quality of data, which was discussed in the previous chapter (Chapter 3.5) and more precisely the phenomenon of inconsistencies that real world data warehouses face.

An example can be seen in Table 7, where all the listed values refer to the same product, but are spelled differently.

Table 7- Naming Inconsistency

Option Name
Option X
OptionX
Opção X
Option_x

It was required for some modifications to be made, standardizing the values with a unique accepted value, which would be solely used.

Next, it was needed to sort the set of the available options, by tiers and subsequently to assign a numeric value from 1 to 7, as many as the discrete options which were incorporated in the set of products. In order to identify the 'importance' or 'weight' of each option, business knowledge was required.

Table 8- Target Creation

Option_Character	Option_Numeric
Option A	1
Option B	2
…	…
Option G	7

Last step, in the target creation process, was to identify all the clients who altered their option in a given time period. It is important to clarify, that all the policyholders can only upgrade their plan on the renewal date, which occurs once per year.

The process which was selected had as following, presented as pseudocode

1. Get the numeric value of the option for two annuities; One in the annuity that is in the selected time window and the option importance of the annuity prior to this.

2. Subtract the two retrieved values (new_option_importance – old_option_importance)

3. If the result is positive mark it as 1 else 0

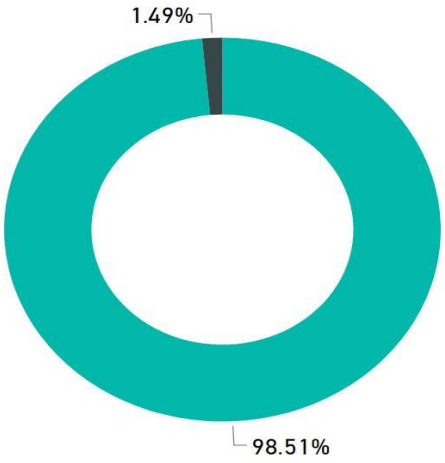

Figure 22– Target Rate

The final dataset was extremely imbalanced, as only 1,49% of the population upsell their product as it can be seen in Figure 22

Finally, an analysis was conducted in order to check whether the phenomenon of seasonality existed in the target universe. In principle it was tested out whether or not spikes of high upsell rates occurred in specific months in the selected time window.

In Figure 23 it is presented the percentage of the total upsell cases, by month.

8.22% of the total upsells took place in January 2017, which was the biggest rate among the rest of the months.

With blue color, the training period is highlighted and with light red color the extra period in which the model was tested.

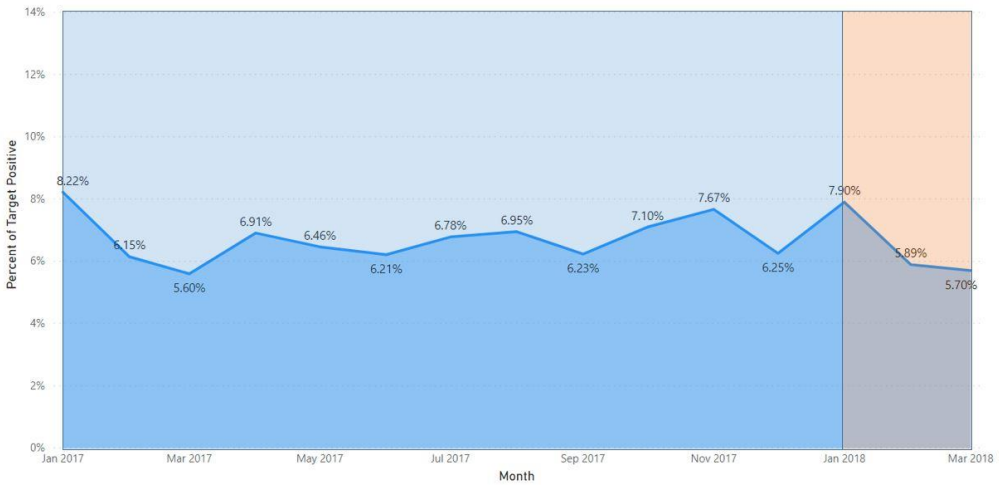

Figure 23– Upsell Rate per Month

As for the Target Universe which was labeled as non-events, a date should be assigned. There were several business approaches of how these dates should be given to the non-events. The two preponderant approaches were the following

- All the 0's to be assigned with the same date which would be in the end of the defined period
- All the 0's to be assigned with a random date

Second solution was selected.

3.9 Data Exploration and Preparation

A preparatory step of modelling is the exploration and preparation of the data. More specifically, during this process, distributions of individual predictors are graphically represented, the degree of missingness is being estimated, unusual values of the independent variables are being exposed as well as the relationship between dependent and independent variables is being revealed.

In Table 9 the major group of activities under the Data Preparation and Exploration umbrella are wrapped and presented.

Table 9- Data Preparation Tasks

Data Preparation Task	What it is tackled	How it is tackled
Data Cleaning	Missing Values	• Delete record • Replace with mean, mode, most frequent class, regressed values
	Duplicated/ Redundant	• Delete record
	Outliers and Noise	• Binning • Filtering with MAD, rare class • Replace extreme values
	Inconsistent	• Business expertise
Data Transformation	Normalization	• Z-Score • Min-Max
	Aggregation	• Aggregated variables
	Feature Construction & Feature Engineering	• Power or Log Transformations • New variables derived from existing
	Generalization	• Drill-Up
	Discretization	• Binning (equal depth, equal width) • Entropy based
Data Reduction	Dimensionality Reduction	• Principal Component Analysis • Heuristic Methods
	Feature Selection	• Chi Square • R square • Cluster-based

Undoubtedly, the more predictors the dataset incorporates, the harder an adequate graphical analysis to be conducted to reveal these relationships.

Thus, and in order for the analyst to rapidly explore the data, there are few guidelines of the most suitable graphs for different exploratory cases, as presented below in Figure 24.

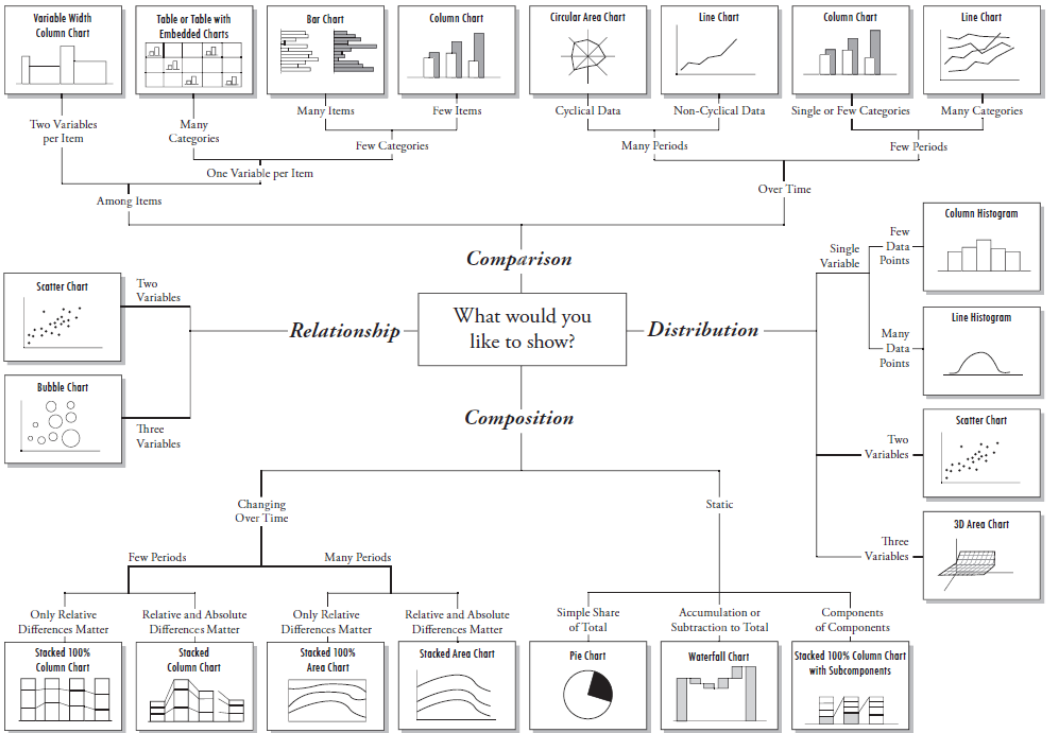

Figure 24– Visualization Table (Source image : https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf)

As examples, in the Figure 25 two variables were plotted to get the first notion of their distribution and visually inspect the skewness and potential outliers

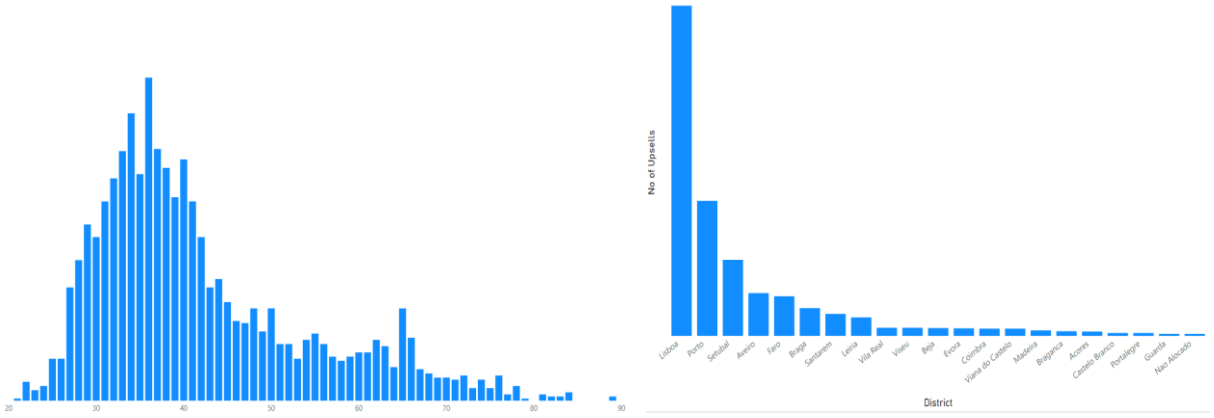

Figure 25– Explanatory Analysis (A)

In the next figure (Figure 26) a line was added, which indicated the upsell rate for each one of the values.

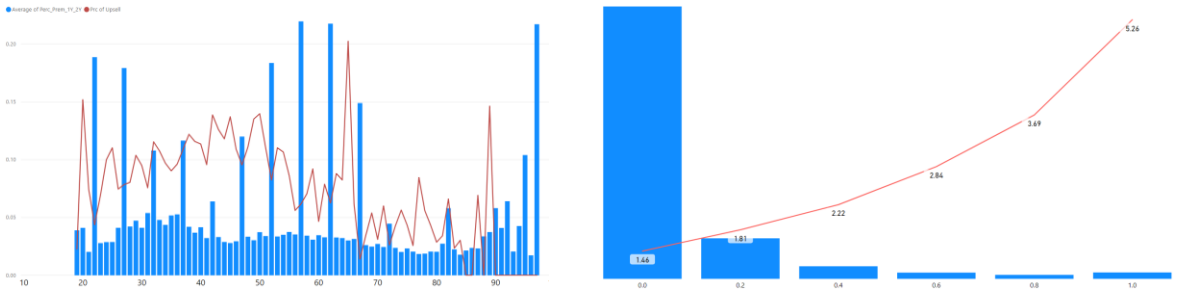

Figure 26– Explanatory Analysis (B)

3.10 Missing Data Treatment

There are several processes which have been developed in order to treat the missing values. However, and before a deep dive into these methods will be presented, it should be reported that there are models that can tolerate incomplete data, thus the treatment of missing data can be skipped.

For example, some of the tree-based models such as CART methodology (Breiman, Friedman, Olshen, & Stone, 1984) incorporate ad-hoc procedures which can tackle internally missing data. More specifically CART methodology uses what is called as surrogate splits whereas, the C5.0 (Quinlan, 1993) handles the missingness in a different way, by using fractional cases.

On the other hand, some of the predictive algorithms cannot tolerate the incomplete data, such as Support Vector Machines and Neural Networks.

Therefore, several ways have been introduced to address this matter which are listed (Han, Kamber, & Pei, 2017)

Deletion: There are two approaches for deletion. First refers to the deletion of all the predictors that incorporate missing values. The second approach encompasses the elimination of the tuples that incorporate missing data.

Fill in manually: This approach, due to the high complexity and the numerous missing data that a real dataset encompasses, quite often is not feasible and besides that, it is characterized as time-consuming, thus it is avoided.

Use a global constant: Another technique that has been proposed and is widely used, is the replacement of all missing values of an attribute, with a constant value such as 'unknown' for categorical features or an extreme numerical value.

Fill in with mean/median: By using this approach, all the missing values of an attribute will be assigned with the same value. As a result of imputing a universal value, could be that the imputed value could deviate by much from the real value, if the latter was known.

Fill in with mean/median of records belonging to the same class: Usually, this approach is a bit more accurate and a more trustworthy estimator can extracted. An example of how this technique can be applied is presented in the Table 10

Table 10- Missing Values Example (B)

ID Client	Operating System	Hours spent on Social Media
A	Windows	7
B	MacOS	3
C	Windows	6
D	Windows	
E	MacOS	2
F	Windows	9

The value in the attribute *'Hours Spent on Social Media'* is missing for the client with ID D, so according to the techniques that have been presented, there is the possibility to impute the missing by the average of this attribute which in the displayed case would be 4,2 hours.

However, if we group by the column '*Operating System*' then, as the missing belongs to the group of '*Windows*', the missing will be replaced by the value of 7,3 which could be perceived as a better proxy.

Fill in with predicted value: The missing value can be predicted by using tree-based rules like decision trees, regressions or other techniques. This method, takes into account other, existing predictors and is more suitable when the type of missing value is defined as NMAR.

User defined: The user assigns a value based on business knowledge or/and experience.

In the context of this project, it was decided to treat the missing values in two different ways taking into account the nature and cause of the missingness. The used methodologies are listed below

- User defined constant value
- Median belonging to the same class

It is important to be highlighted, that the Insurance company follows rigorous procedures which ensure that, at least for the fields that must be filled in, either from the side of the client or from the side of the Company, there will not be missing values.

However, there are some fields, such as the marital status of the person that is being asked in the process of issuing a policy to provide this information, that are optional to be completed. For this reason, it is common to have a great proportion with missing values in such cases.

User defined constant value.

First there were few variables that missingness, was actually meaning absence of an action. Therefore, it was rational to impute 0 as a constant value. An example which was addressed in this project is presented.

It was desirable to retrieve information regarding Health Claims for the Target Universe as it was believed that good predictors could be found.

First, a join which would allow us to have all the records from the Target Universe with the respective Health Claims information for every single record, took place.

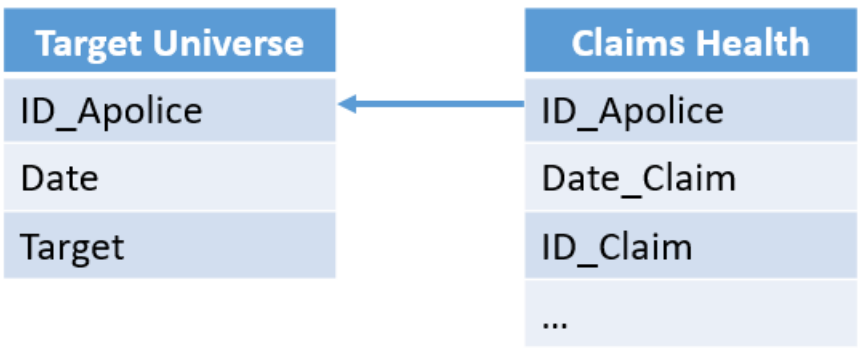

Figure 27– Join Tables

After running the join as displayed in Figure 27, a table similar to the Figure 28 was created.

ID_Apolice	Date	Target	ID_Claim	Date_Claim	Value_Claim
A	15jan2017	0	1adf	5mar2016	100
A	15jan2017	0	2saf	1apr2015	50
B	18feb2017	0	.	.	.
C	4mar2017	1	1asd	1apr2016	3200

Figure 28– Missing Values Example

For clients with missing in value claim, as the ID_Apolice B, the imputed value for the variable 'Value_Claim' was 0.

Median of records belonging to the same group.

The second methodology that was selected, was to fill in some of the attributes with the median of a group.

In this project the grouping variable that was selected, refers to an in-house classification that grouped the clients into 9 different clusters, based on several characteristics.

Finally, it should be mentioned that attributes with more than 25% of missing values, which missingness wasn't due to lack of taking an action, as described in the user defined constant value scenario, was decided to be dropped.

3.11 Outliers Treatment

The last issue that had to be addressed during the phase of data cleaning, was the phenomenon that in the literature review is described under the terms of outlier or influential data points or fringelier (Wainer, 1976).

A univariate outlier detection by creating boxplots was conducted in this project. The first goal was to identify how many of the variables, incorporated clients which could be labeled as outliers.

There was a business decision, based on Company's Intelligence that for some variables where extreme values had been detected, no treatment through exclusion/deletion should be applied as the contained information could be valuable. For example, few clients were identified with exceptional high premiums in some of the Financial products that are available in the Insurance's portfolio. This characteristic could be an estimator of affluence and a good predictor.

Nevertheless, it was decided for some of the attributes with extreme values, including the aforementioned example, the method of truncation to be used. The benefit of truncating the outliers was that there wouldn't be any loss of information and the tendency or orientation of the values would still be present in the variable.

Next, for the rest of the attributes with outliers, it was tested the proportion of clients which would have been excluded, if all the extreme outliers were eliminated from all the variables with extreme values. By doing this, the percent of potential excluded outliers was estimated.

Moreover, another metric that was calculated while testing the volume of univariate outliers, was the following.

It was important to understand if the same clients were labeled 2 or more times as outliers in different variables. Thus the percentage of overlay of the same observations was calculated. This would help to clearly understand if an observation was constantly detected as an outlier.

An example which created for the purposes of this document is being presented in Table 11.

Table 11- Outliers Example

Outliers Attribute A	Outliers Attribute B	Outliers Attribute C
ID A	ID A	ID A
ID B		ID B
ID C	ID C	ID C
ID D		
ID E		
ID F		
	ID G	

- Percent of Outliers appear in just 1 attribute: 58%
- Percent of Outliers appear in just 2 attributes: 14%
- Percent of Outliers appear in more than 3 attributes : 28%

Finally, it was decided, that outliers with overlay of 15 variables to be removed. This decision lessened by 1.35% the Target Universe.

3.12 DATA TRANSFORMATION

There is a broad spectrum of modifications that can be applied to an individual predictor and can serve many functions in quantitative analysis of data (Osborne 2002) which also might improve the utility in the model.

The groups of the available transformations are listed below as described (Han, Kamber, & Pei, 2017)

Smoothing is a technique that can advocate in the process of removing noise from the dataset. Binning and clustering are just two examples of techniques that exist and can be applied. More specifically, binning to remove noise is subdivided into the following methods:

Smoothing by bin means

Smoothing by bin boundaries

Smoothing by bin median

It is common to apply smoothing on data that contain time or sequence effect. An example is presented below where a running 2-point median, is replacing each data point.

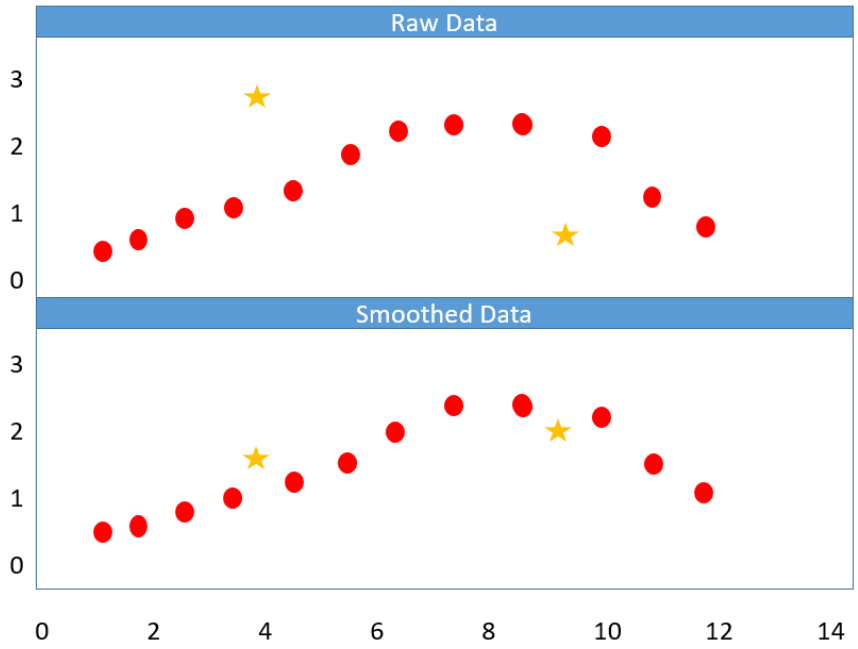

Figure 29– Smoothing by Median

As it can be seen, in the Figure 29 in the Raw Data, there are some instances which deviate significantly from theirs neighbors. By smoothing these anomalies, noise has been almost vanished.

Aggregation is a method which summarizes a group of variables or a group of instances of the same variable into a new one.

An example is given in Table 12 which was created for the purposes of this project.

In the Insurance Sector, the Line of Business of Motor and Multi Risk, can be described under the term of Non-Life Products. Therefore, a new attribute that aggregates the information of the two columns '*No_Policies_Auto'*, '*No_Policies_MultiRisco'* of Table 12, can be summarized into one new column ('No_Policies_NaoVida') as shown in the same table.

Table 12- Data Transformation by Aggregation

ID Client	No_Policies_Auto	No_Policies_MultiRisco	No_Policies_NaoVida
ID A	2	1	3
ID B	5	9	14

Generalization is the process which allows us to drill up and go from an attribute with high specialization, to a more abstract and less informative. In the example shown in the following figure (Figure 30) the most specialized attribute that provides the age of a client is the age (in years) and the most generalized attribute is the age classification; Age classification as well as Age Group, will be explained in the following lines

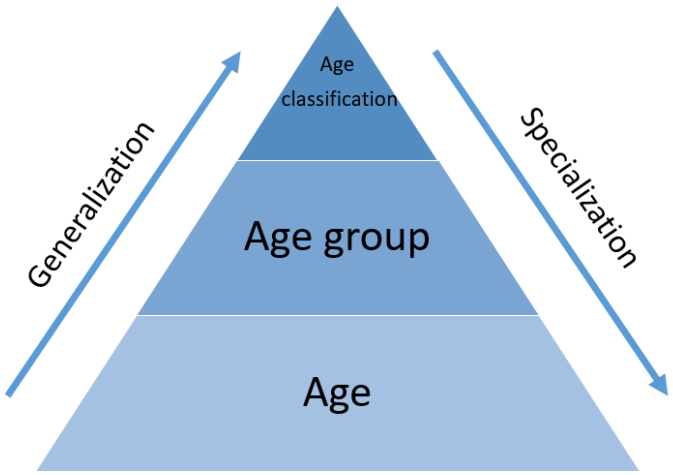

Figure 30– Specialization VS Generalization

In the context of this project, the only attribute that was available regarding the age of the client, was the age in years. Thus, a new set of variables was constructed, in the logic of generalizing this information in two higher levels as it is presented in Table 13.

Table 13- Data Transformation by Generalization

Age	Age Group	Age Classification
0…25	1	Young
26…35	2	
36…45	3	Mature
46…56	4	
56…75	5	Old
76…	6	

Normalization is the process of rescaling an attribute so it falls in a specific, normally small, range.

Feature Construction is a process that discovers missing information between features and augments the space of features by inferring or creating additional features (Matheus, 1991). The new variables that are constructed are the result of existing ones, so no inherently new information is being added. Attribute construction attempts to increase the expressive power of original features (Motoda & Liu)

Several approaches on how to construct new variables exist with data-driven and knowledge-based to be among these. The data driven approach suggests that various operators can be used to construct new variables. On the other hand, knowledge-based construction emphasizes that new features can be the result of the Company's Intelligence.

As it was described previously, several interviews were held aiming exactly to this; to compose new feature's based on domain expertise, if possible. As a result, several knowledge based and data-driven attributes were created.

Below some of them are listed along with the formula which was used:

- Days to Renewal = Date of Renewal – Date of Reference
- Capital Used = Available Capital – Used Capital

3.13 Feature Selection

"The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes" (Han, Kamber, & Pei, 2017)

The role of Feature Selection in a Data Mining project is absolutely critical for several reasons, each one of these, has an effect in different perspectives.

There are many motivations for eradicating irrelevant and redundant features.

First, even though machine learning algorithms have demonstrated a great success into learning patterns that are complex, the level of interpretability of the produced models is sometimes low.

This may be due to the nature of the algorithm. When this is happening, it is referred in the literature as lack of algorithmic transparency. For example, SVM can produce a highly accurate model but to be low in terms of explainability.

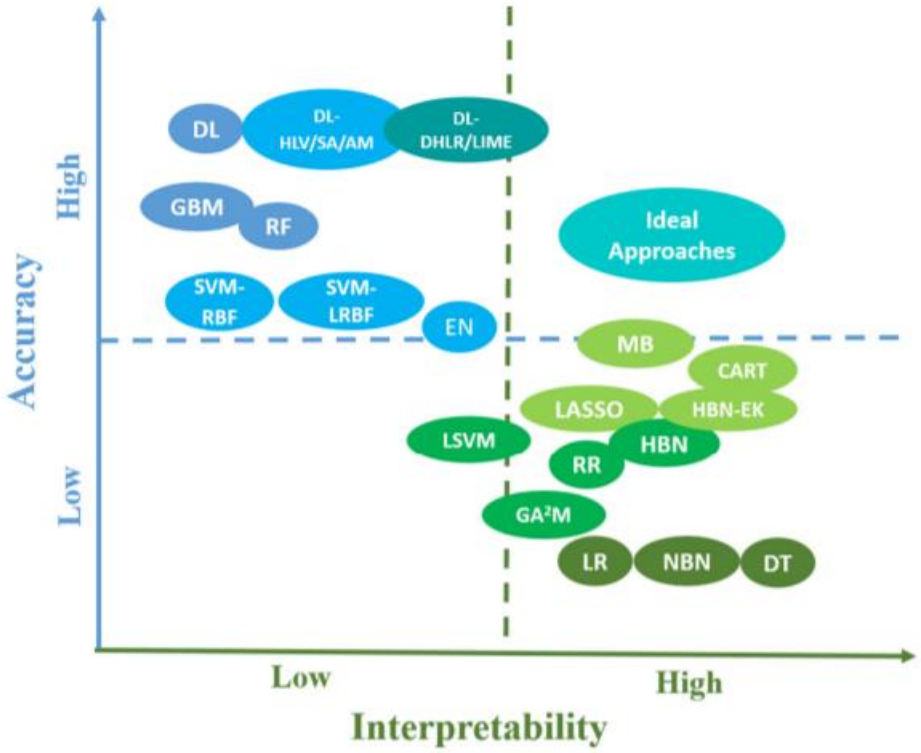

Figure 31– Accuracy VS Interpretability (Source Image https://www.birpublications.org/doi/pdf/10.1259/bjro.20190021)

Abbreviations explained: EN, elastic net; LR, logistic regression; MB, MediBoost; RR, ridge regression; LSVM, linear support vector machine; DT, decision tree; CART, classification and regression tree; DHLR, disentangled hidden layer representation; DL-AM, deep learning with attention mechanisms; DL-HLV, deep learning with combination of handcrafted features and latent variables; GBM, gradient boosting machine; HBN, hierarchical Bayesian network; HBN-EK, hierarchical Bayesian network with expert knowledge; HLV, handcrafted features and latent variables; IP, interpretable; LASSO, least absolute shrinkage and selection operator; LIME, local interpretable model-agnostic explanation;

However, some seemingly high interpretable algorithms like Linear Regression, are also not always easy to be explained due to high dimensionality or due to heavily engineered features.

Moreover, there some models such as Neural Network (NN) and SVM that are sensitive to variables that are irrelevant, and this can cause a drop in the performance of the outcome

Similarly, regression models are sensitive when highly correlated variables are incorporated to the input dataset, due to the phenomenon of multicollinearity.

Last, for some algorithms that are tolerant to high dimensionality and redundancy, it is prudent to exclude them in order to reduce the computational power that will be required and in some cases even lessen the effort of collecting data.

Several methodologies for attribute selection have been developed which can be categorized into two main groups.

Filtering: This is a supervised process which measures the relevance of the variables that are incorporated to the dataset before entering the model. After the filtering is executed, only few variables will be nominated to be an input to the models from this non-iterative process. Most of time filter methods are fast and effective to capture trends between individual predictors and outcome relationship by using measures of statistical significance like p-value. Indicative filter methods are the Chi-Square for categorical variables and Correlation Coefficient.

Wrapper: This method tends to select a better subset of predictors comparing to filter methods, as the process is iterative and repeatedly evaluates a set of variables based on the performance of the model. Wrapper attribute selection can be subdivided in greedy methods like backwards selection where initially all the variables enter the model and recursively are being eliminated or non-greedy methods like Simulated Annealing or Genetic Algorithms which the element of randomness is added in the feature selection process.

In this project, a hybrid method was applied for the step of variable selection which is graphically expressed in Figure32 and will be described in greater detail, in the following paragraphs.

At first, as the software that was used for the Data Mining project has a built-in node for variable selection based on R-squared and Chi-Square tests for both interval and categorical variables, it was selected to run the node and fine tune the parameters (minimum R –square and minimum Chi-square).

Moreover, a Regression node was executed for Variable Selection using the effect of Stepwise. As in the Forward method, Stepwise selection method, begins by default with no

candidate effects in the model and systematically adds effects that are significantly associated with the target. However, and here is the difference with the Forward Selection method, after an effect is added to the model, Stepwise may remove any variable that was already selected and entered the model.

Finally, for the first batch of nodes that were imported to the diagram, a decision tree was selected.

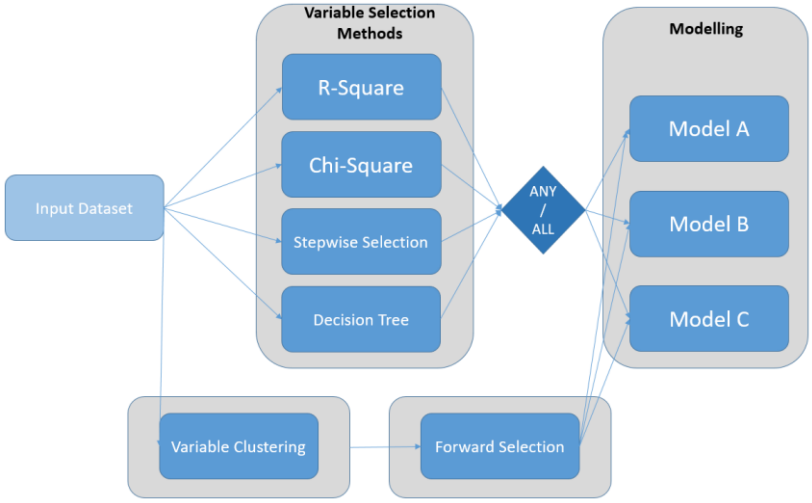

Figure 32– Project's Variable Selection Process

3.14 Predictive Models used in the Project

In this chapter the main algorithms that were used in the modeling phase will be discussed.

3.14.1 Linear Regression

Simple linear regression analysis is a statistical technique that quantifies the relationship between two continuous variables: the dependent variable which we want to estimate and the independent or explanatory variable. Linear Regression draws a line through the data that minimizes the squared error from each point.

Figure 33 illustrates the abstract relationship between *Salary* and *Years of Professional Experience*. The target is to predict the Salary based on the years of professional experience. The graph shows a very linear relationship between *Salary* and *Years of Professional Experience*.

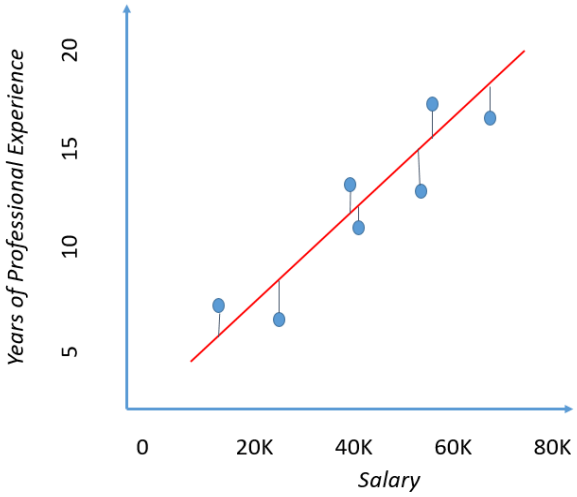

Figure 33– Linear Regression Example

A key metric of how much of the variation can be explained by the independent variable, is the R-squared

The formula to calculate R squared is presented step by step in the following lines

- Calculate the mean of the observed data

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (1)$$

- Estimate the variability of the entire dataset by calculating the Sum of Square Error

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \qquad (2)$$

- Estimate the Regression Sum of Squares

$$SS_{reg} = \sum_i (f_i - \bar{y})^2 \qquad (3)$$

- Estimate the Residual Sum of Squares

$$SS_{reg} = \sum_i (y_i - f_i)^2 \qquad (4)$$

- Estimate the R-squared or coefficient of determination

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (5)$$

3.14.2 Decision Trees

"Decision Tree learning is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete valued functions that is robust to noisy data and capable of learning disjunctive expression"

(Mitchell, 1997)

Decision Trees classify instances by sorting them down the tree, from the root node to some leaf node. The leaf node indicates the final prediction of each record of the dataset. Each node can be seen as an if-then rule which is more intuitive to human readability and understanding.

In Figure 34 an example of the structure of a Decision Tree is presented

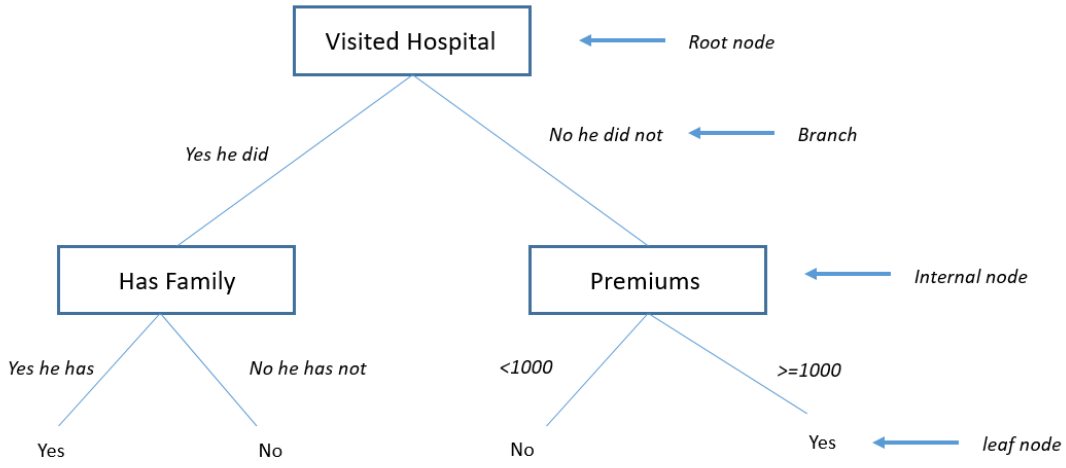

Figure 34– Structure of Decision Tree

Each node (root and internal) of the tree performs a test to a single feature of the input dataset. For example, in Figure 34 the concept is to identify who are these clients of the Health Insurance that are more likely to upsell.

The root node is the attribute if the client visited during the last year a hospital for a medical consultation.

The branch of the root node output to two answers (yes he visited, no he did not visit)

If the value of the root node is positive, then this leads to the first internal node which is the attribute if the client has family members in the insurance policy.

In case that the value of the root node is negative (no he did not visit a hospital for consultation) then another internal node is placed which refers to the continuous attribute of paid premiums of the policy in the last year.

Univariate splitting criteria

Most of times, the discrete splitting functions are univariate. The term of univariate refers to the case that the internal node is using only a single attribute. The algorithm is searching for the best possible split.

There are several metrics for evaluating the goodness of split, based on the origin of the measure, such as:

- Dependence/statistical significance test (e.g. Chi square),
- Information Theory (e.g. Information Gain)

Impurity Based Trees

ID3 algorithm, an indicative example of impurity-based trees, is based on a statistical property called Information Gain.

Information Gain is a quantitative measure which evaluates how well a predictor is able to separate the training records according to their target classification. ID3 algorithm uses this property in order to build the top-down tree, from the root to the last internal node, and selects the best attribute among its candidates recursively, until the tree meets a stopping criteria.

To compute Information Gain, first is needed to estimate Entropy. Entropy measures the homogeneity of examples, that characterize the impurity (or purity) of an arbitrary collection of examples.

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

$E(S) = -\Sigma p_{ici=1} \times log_2(p_i)$

Where p_i is the proportion of S belonging to class i. For a Boolean classification S, the entropy function is computed as:

$E(S) = -[p * log_2(p) + (1-p) * log_2(1-p)]$

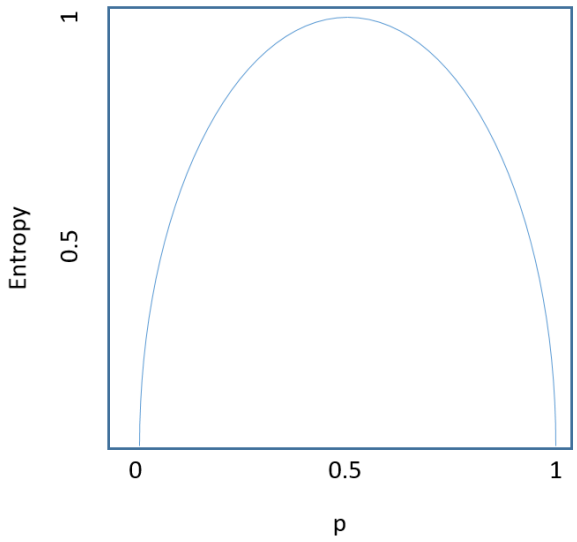

Figure 35– Entropy

Figure 35 shows the variation of the entropy for a Boolean target variable. The maximum (1) is reached when there is no distinction for the target variable, which corresponds to a 50%-50% proportion of event and non-event.

ID3 algorithm aims to find the split that reduces the entropy as much as possible and provides the largest difference in proportion between the target levels.

In the previous example, the class labels were added (**event**-for clients who upsell, **non-event** for clients who did not upsell) and the Entropy was estimated.

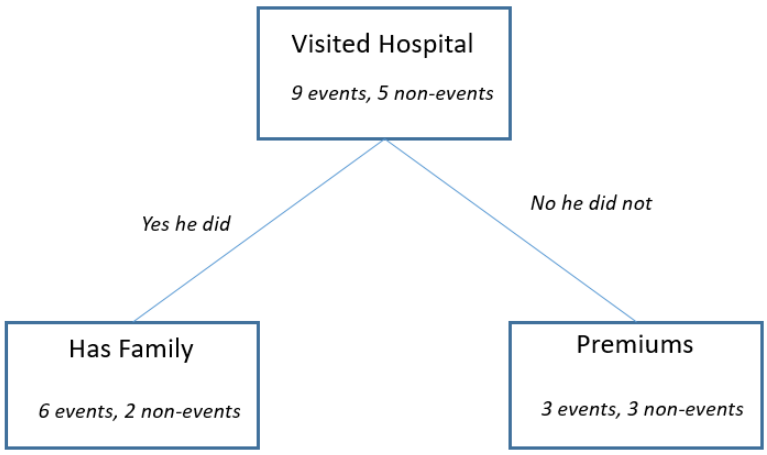

Figure 36– Decision Tree Example

Entropy([9events,5non-events]) = -9/14 log_2 (9/14) -5/14 log_2 (5/14) = 0.94

Information Gain measures the expected reduction in Entropy, and actually defines the effectiveness of an attribute in classifying data (Mitchell, 1997)

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

Gain(S, A) ≡ Entropy(S) - Σ|Sv||S|v ∈$Values(A)$ Entropy(Sv)

Values(A) is the set of all possible values for input A, and Sv is the subset of S for which input A has value v. This measure is computed for each variable. Later, the variable that gives the largest Gain is chosen to split. It is important to notice that the initialization of the algorithm computes the initial entropy of the system by computing the entropy of the target variable.

Continuing the previous example, we calculate the Information Gain.

Values (Visited Hospital) = Yes he did, No he did not

S = [9 events, 5 non-events]

$S_{\text{Yes he did}}$ = [6 events, 2 non events]

$S_{\text{No he did not}}$ = [3 events, 3 non events]

$$\text{Gain(S, Visited Hospital)} = \text{Entropy(S)} - (8/14)\text{Entropy}(S_{\text{yes he did}}) - (6/14)\text{Entropy}(S_{\text{no he did not}})$$

$$= 0.94 - (8/14)*0.811 - (6/14)*1$$

$$= 0.048$$

The stopping criterion is reached when there is no possible increase in Information Gain for a split in a branch or all training examples belong to the same target class. Because Information Gain criteria lack the significance threshold feature of the chi-square criterion, they tend to grow enormous trees.

There are several stopping criteria that can be used, some are inherent from the exhaustive nature of the algorithm meaning that either all training instances belong to the same target class or there is no increase in IG in a potential further split.

However, some hard rules or parameters, can be introduced such as setting the maximum depth of the tree. By setting this parameter it is almost certain that a trade of between Variance and Bias for the training and validation sets will be introduced. In the figure 37 an arbitrary example is given, depicting this case.

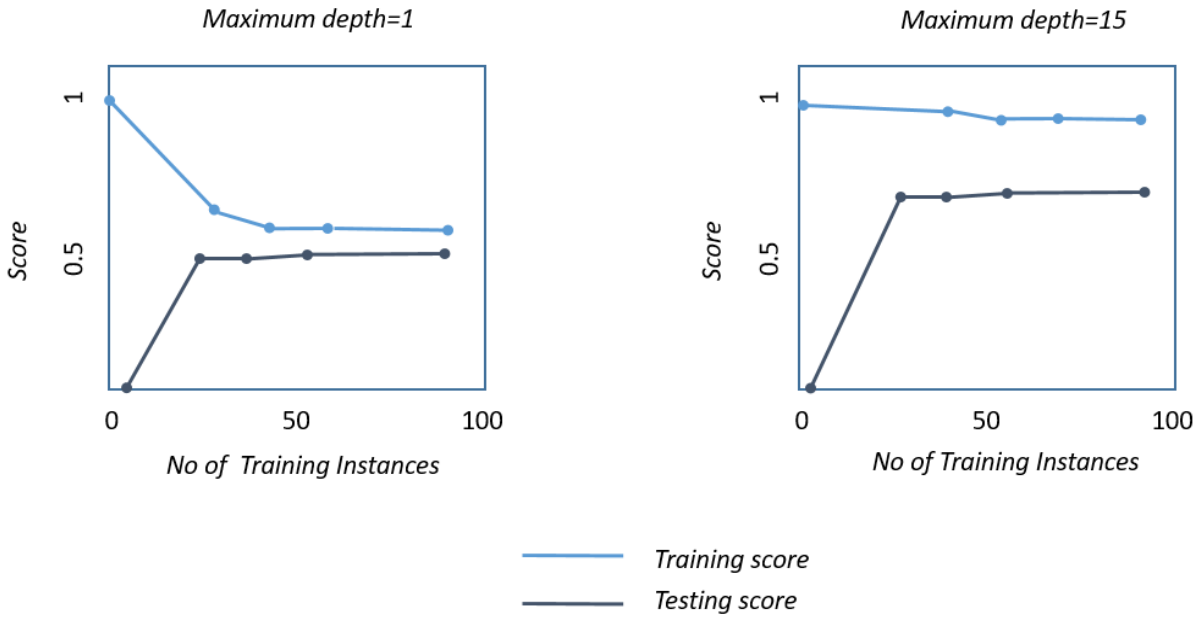

Figure 37– Variance VS Bias

When maximum depth is low we have high bias and the model normally underfits. Adding more training instances is not benefitting the accuracy, instead complexity should be increased if a better fit. However, the variance is relatively low and the test set is performing similarly to the training set.

On the other hand, when maximum depth has been set relatively high, as shown in the right part of Figure 37 then the training set performs well. However, it is clear that the model overfits and the ability of generalization is low.

3.14.3 Gradient Boosting Algorithm

Gradient boosting of regression trees produces competitive, high robust, interpretable procedures for both regression and classification (Friedman, 1999)

Tree boosting creates a series of decision trees which together form a single predictive model. A tree in the series is fit to the residual of the prediction from the earlier trees in the series. The residual is defined in terms of the derivative of a loss function. (SAS Institute Inc, 2017)

It worth to highlight the differences of bagging methods and boosting methods.

Bagging is an ensemble technique in which several *independent* models are built and combine them by conducting some model averaging techniques. (e.g. weighted average, majority vote). All models built simultaneously, in parallel.

Boosting is an ensemble technique in which the models are depended on the previous learner, and sequentially reduce the loss.

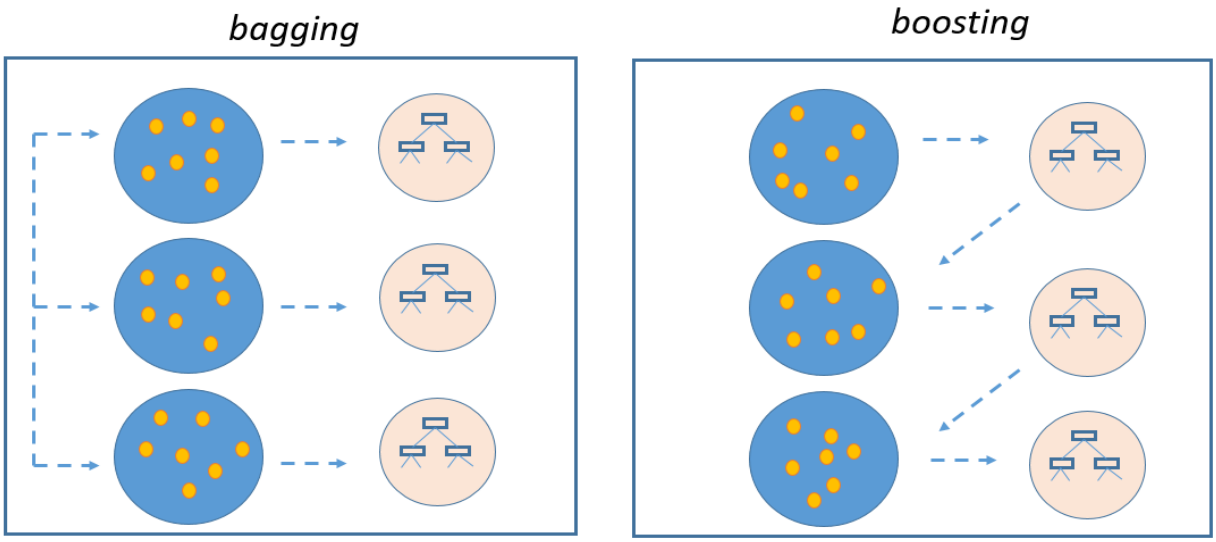

Figure 38– Bagging and Boosting

The logic behind boosting, incorporates the construction of subsequent learners which every time learn from the mistakes of the previous learners. Thus, the probability for an instance to appear in the next model is unequal as only the instances with high errors are promoted to the next learner.

Benefits of Ensemble methods

Using Ensemble models could lead to great advantages (Lantz, 2015):

Better Generalization: Since the output of several is based on a variety of learners no single bias is able to dominate.

Improve the performance on massive or tiny datasets: Many models run into memory or complexity limitations. Then, a possible strategy to overcome this issue is to train several small models than a single full model. Oppositely, in small data sets ensemble models provide a good performance because resampling methods such as bootstrapping are inherently a part of many ensemble designs.

Synthesize data from distinct domains: Since there are no one-size-fits-all learning algorithms, the ensemble's ability to incorporate evidence from multiple types of models with data drawn from different domains.

4 RESULTS

One of the most crucial stages of a Data Science project is the evaluation of the learners. Therefore, several techniques have been introduced in the community for both supervised and unsupervised learning.

The evaluation metrics which will be used to assess the performance of the model, should be aligned with (Solokova & Lapalme, 2009).

- the objective of the model,
- the nature of the target variable (e.g. binary, multi-class, etc)

The focal point of this chapter will be these methods that have been developed to estimate the quality of the predictive model for supervised learning and more specifically for binary classification problems, although some may be applicable to other types of problems.

The evaluation stage usually begins by categorizing the predicted observations in 4 groups with respect to their true values (Sokolova & Lapalme, 2009).

TP (True Positives) Number of elements belonging to C and that are classified as C
TN (True Negatives) Number of elements that do not belong to C and are not classified as C
FP (False Positives) Number of elements that do not belong to C but are classified as C
FN (False Negatives) Number of elements belonging to C and that are not classified as C

Sequentially, these four groups are placed in a 2*2 matrix, known as confusion matrix as shown in Figure 39.

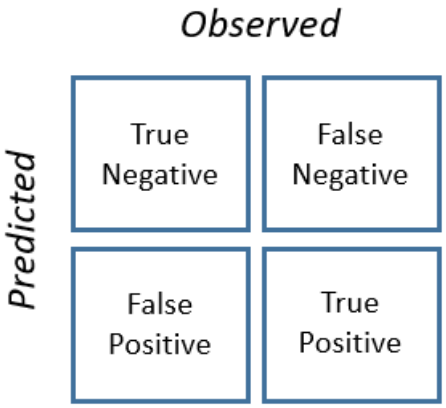

Figure 39– Confusion Matrix

A number of metrics and graphs useful for the analysis can be arranged from the confusion matrix. The following table (Table 14) summarizes the most common.

Table 14- Model's Evaluation Performance Metrics

Metric	Formula
Accuracy	TP + TN) / (TP + TN + FP + FN)
Misclassification Rate	1 – Accuracy
Precision	TP / (TP + FP)
Recall/Sensitivity	TP / (TP + FN)
Specificity	TN / (TN+FP)
False Positive Rate	1 – (TN / (TN+FP))
True Positive Rate	TP (TP/FN)
F-Score	2*((Precision* Specificity) / (Precision + Specificity))

Accuracy is the first indicator which is normally being reviewed but many times can be misleading, especially when dealing with highly imbalanced datasets, thus it is a good practice to accompany it with other metrics such as Lift and AUC.

A simple example of accuracy's limited effectiveness can be seen in a fraud detection case. On the premise that 99% of times, claims in a car insurance are valid, if there is a model that always classifies the instances that are not fraudulent, then it is truthfully to report that the accuracy of the model has a 99% of accuracy.

The Receiver Operating Characteristic (ROC) curve provides a notion of the trade-off between the true positives and false positive rates.

The curve is generated by plotting the True Positive Rate against the False Positive Rate at several threshold values. The performance metric used, is called the Area Under the ROC, which compares how much better the model performs from a random classifier which is the diagonal line as shown in Figure 40

The perfect model has an AUC=1 and a random model, meaning there is absence of discriminant power has an AUC=0.5

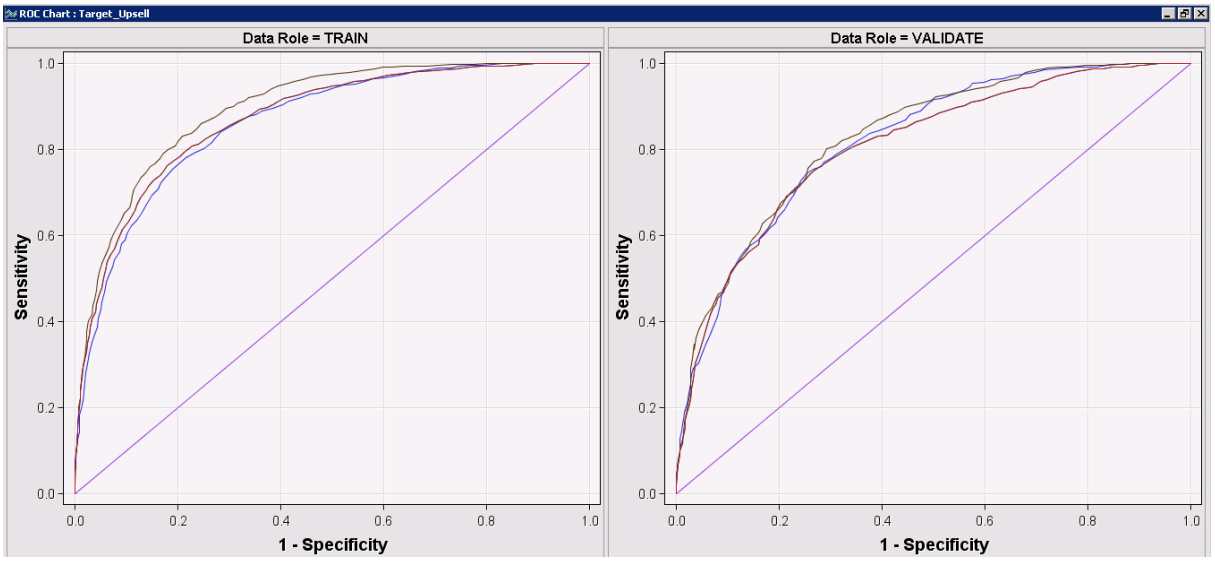

Figure 40–ROC

Another metric that is extensively used, is the Lift. Based on the Lift chart, you can estimate the expected uplift of the model and response rate in any depth as the instances are sorted by posterior probability. By depth it is meant the proportion of the data that will be used for a specific action.

The champion model of this project had a Cumulative Lift of 1.84 for Depth=10% in the validation set (Figure 41). We trained the model in a sample with a proportion 50:50 of events and non-events.

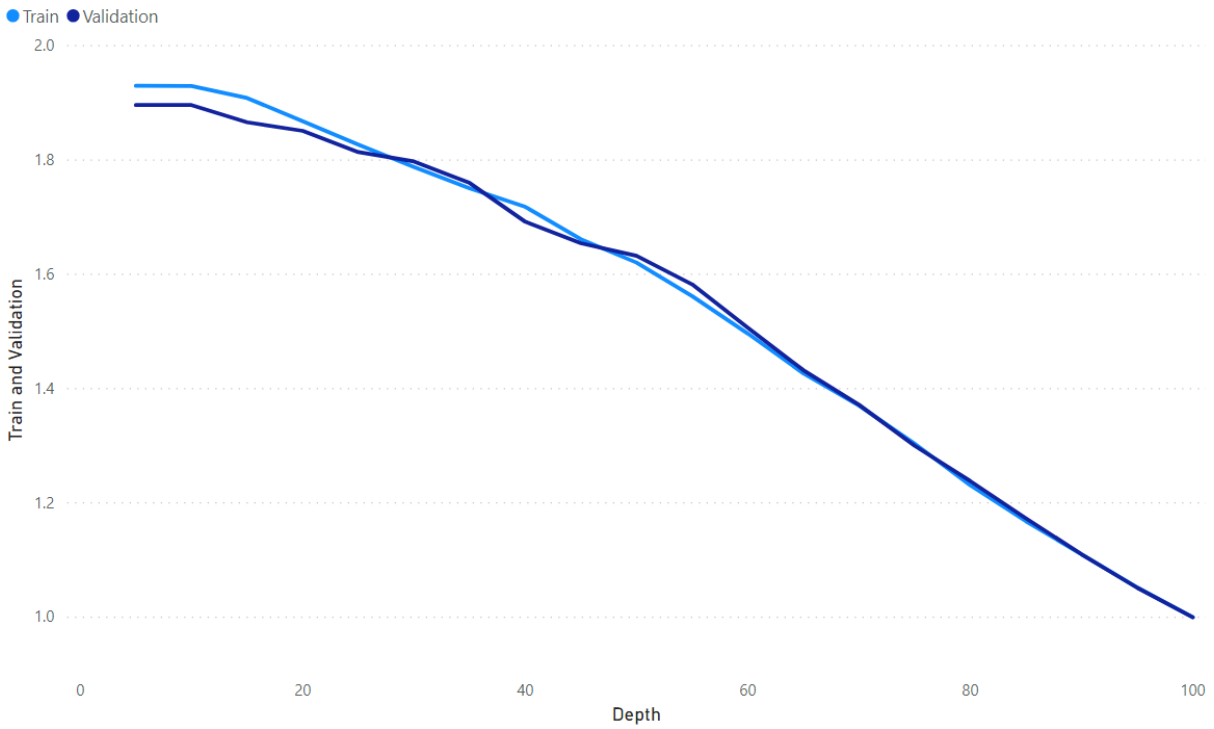

Figure 41– Lift

In Figure 42 it is presented the benchmark when the dataset is balanced.

	Champion Model	Benchmark	
		Random model	Perfect model
ROC Index	0.77	**0.5**	**1**
Cumulative Lift at top decile (10%)	1.81	**1**	**2**

Figure 42–Benchmark of the models

Last, a graph (Figure 43) based on the metric of Lift was created which displays the actual number of clients which were classified as events and non-events for each single depth.

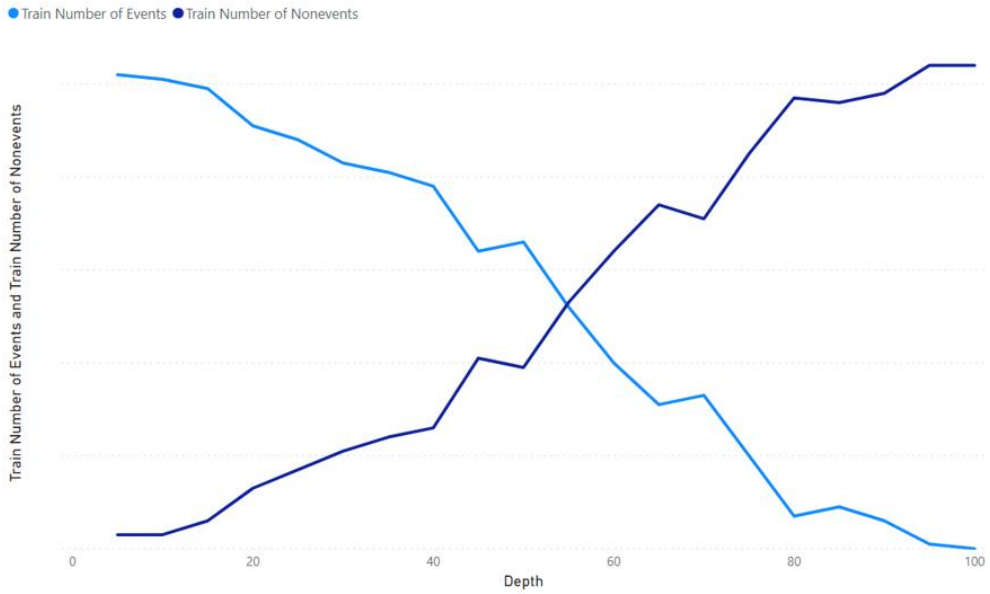

Figure 43– Events VS Non–Events

An important task which often businesspeople ask for, is to interpret the results of the model in a way that the black box of the deployed algorithm is understandable for them, gaining as may insights as possible. In addition, the process of explaining the importance of the variables, can be seen as an additional validation test for the model. Business experts could indicate some spurious variables that the model used, thus it is important to have a good understanding of the final result and predictors.

In this project the scored instances were grouped in deciles and based on the importance of the variables that the champion model used, they were averaged when the variable was numerical and calculating the rate wherever the variables were categorical.

Examples of a numerical and categorical variable are presented below respectively.

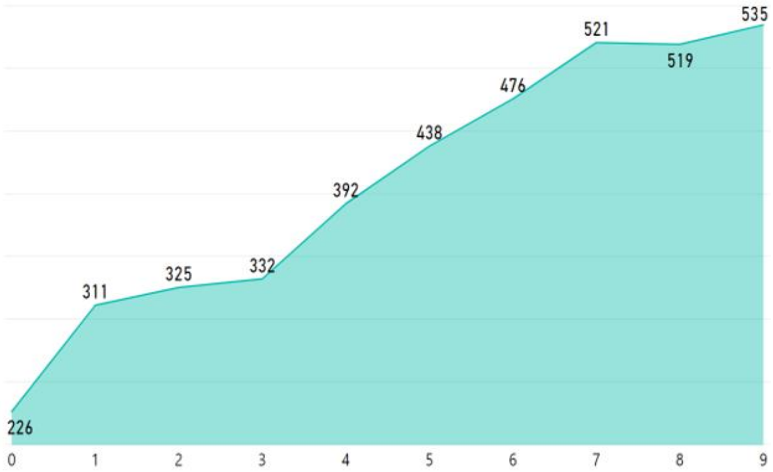

Figure 44– Numeric Variable by Decile

In Figure 44 the numeric Variable A which was used by the model is presented. In the X-axis, the deciles from 0-9 are displayed, sorted in ascending order whereas on the Y-axis the Variable A is displayed. The 9th decile incorporates the highest posterior probabilities. Each decile has the same number of instances. In this example, clients that are classified in the best probability decile (9th), have an average value of 535 while clients in the lowest decile have an average of 226

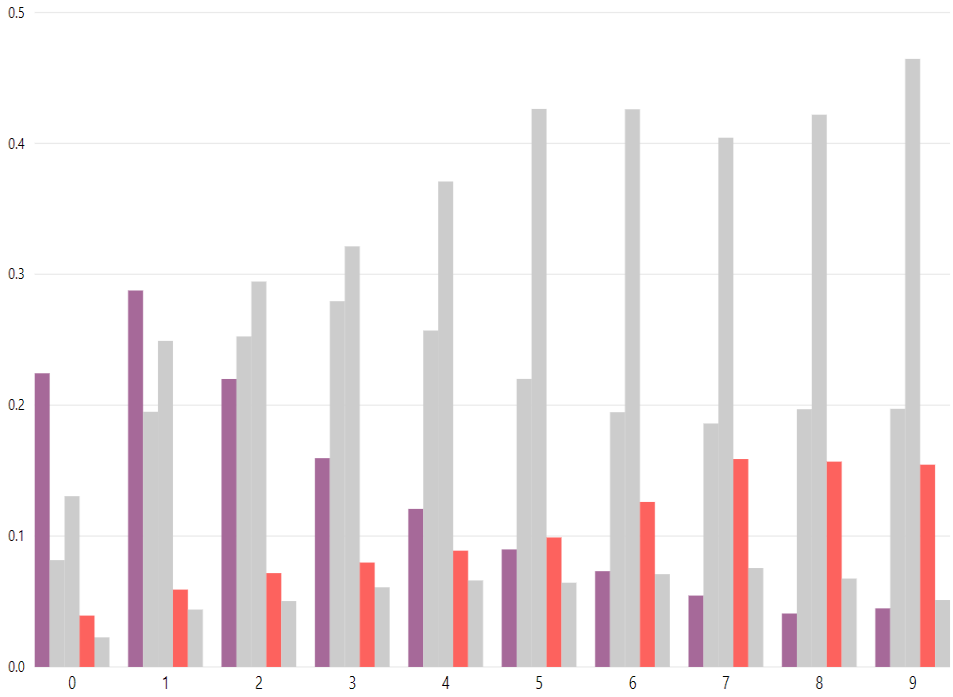

Figure 45– Categorical Variable by Decile

In Figure 45 the categorical variable “Health Insurance product/solution” is presented in the format of ratio. Each decile sums up to 100% (in this graph though some of the products/solutions were excluded in order to ease in the readability).

As it shown, the purple product/solution has a decreasing presence towards the high probability deciles, while the red product/decile has an increasing presence.

5 CONCLUSIONS

The objective of this project, was to construct a model for predicting existing clients of the Health Insurance who are more likely to make an upsell on their policy, following the best academic practices for deploying a Data Mining project along with the business know-how and the pre-existing rules of the Company.

The outcome of the project is now being used in the context of a non-incentive Marketing campaign. The execution of the campaign is through the channel of a call center and the capacity of the campaign varies, depending on the workload of it, as well as the amount of parallel campaigns that are running at the same time.

In the initiation of the project, several interviews were held, which aimed to get a spherical understanding of the Insurance functions together with the important day to day processes. In support of getting to know the business, Human Resources Department, organized a series of talks with managers and key persons from all the Line of Business (Life, Non-Life, Health) providing useful insights.

Subsequently, more exhaustive meetings with the main stakeholders of the project took place, defining the scope and the target of the Health Insurance upsell predictive model.

The construction of the target, the data understanding and the data preprocessing were time consuming tasks but critical for the success of the final outcome.

The exploration of the available datamarts along with the validation process of checking the correctness of the incorporated data, the understanding of the existing variables, the variable transformation and the feature selection, were major tasks and an extra emphasis was given to complete them in the best possible way.

Businesspeople required a report with basic descriptive statistics before the phase of modeling took place, mainly demographic related statistics, such as; age groups with the highest upsell rate as well as geo-related analysis based on the address of the client.

Several algorithms were tested, applying different parametrizations in order the best result to be achieved always in respect of taking care of caveats such as spurious correlations and the phenomenon of overfitting.

The evaluation metrics which were mainly used to assess the performance of the models, were the cumulative lift and the ROC curves.

Once the champion model was selected, a presentation for the stakeholders was prepared in order to communicate the results and explain the key findings.

6 LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

In this project there are few limitations and concerns that should be reported.

First the upsell cases in the Health Insurance were extremely imbalanced as the total number of events, represented roughly 1.5% of the whole universe. No matter how good the model performs, the results cannot be trustful.

Second, as the technique that was selected to deal with the high imbalanced dataset was to under sample the universe, meaning that all the events (1's) would be used, but only a small proportion of the non-events (0's), there was a significant hazard which implicated that important information (for the non-events) would may been extinguished. One of the future improvements that can be tested, is to try out oversampling methods such as SMOTE.

Last, a bias was added in the model, as two categories of clients who upsell were identified and both included in the universe, classified as events. The first category refers to clients who completed an upsell process through the existing ongoing campaign, so it is reasonable to state that there was not any initiative from the client's side but the upsell occurred due to the campaign effectiveness. The second category is consisted by clients who upsell on their own initiative. From the business experts' point of view the trigger for upselling in these two upsell types are clearly contrasting.

A limitation that should be reported, is that in this report important information and insights couldn't be included as the Insurance company has a rigorous regulation rules about data protection.

Regarding the future work, as the model is being used in the frame of a campaign and the channel's capacity is limited, it has already been planned to construct a complementary project that will predict the increased monetary value of a potential upsell, so as to be able to create "Hot Leads". By "Hot Leads" it is meant clients with high probability to upsell and high monetary added value, in case the upsell takes place.

7 BIBLIOGRAPHY

Anscombe, F. J. (1960). *Rejection of Outliers* (Vol. 2). Technometrics.

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

Caderholm, T. (2014, July 30). *Market Realist*. Retrieved 2020, from https://articles2.marketrealist.com/2014/07/jetblues-ancillary-products-performs-terms-net-promoter-score/

Dickey, D. A. (2012). Introduction to Predictive Modeling with Examples. N.Carolina.

Feelders, A. (n.d.). *Handling missing data in trees:surrogate splits or statistical imputation ?*

Friedman, J. H. (1999). *Greedy Function Approximation: A Gradient Boosting Machine.*

Ghani, M. F. (2008). Intelligent heart disease prediction system using data mining techniques.

Grace-Martin, K. (n.d.). Retrieved 2020, from https://www.theanalysisfactor.com/missing-data-mechanism/

Graham, S., Coburn, D., & Oleson, C. (1996). The Foundations of Wisdom: A Study of the Financial.

Grisaffe, D. (2007). QUESTIONS ABOUT THE ULTIMATE QUESTION: CONCEPTUAL CONSIDERATIONS IN EVALUATING REICHHELD'S NET PROMOTER SCORE (NPS). *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 36-53.

Grover, P. (2017). *Medium.com*. Retrieved 2020, from https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

Hamilton, L. C. (1992). *Regressions with graphics: A second course in applied statistics.*

Han, J., Kamber, M., & Pei, J. (2017). *Data Mining: Concepts and Techniques* (3rd ed.). ELSEVIER SCIENCE & TECHNOLOGY.

Hawkins, D. (1980). *Identification of Outliers.* Chapman and Hall.

Hendra, R., & Staum, P. W. (2010). A SAS® Application to Identify and Evaluate Outliers.

Huck, S. W. (2000). *Reading Statistics and research.* New York.

IBM Corporation. (n.d.). *IBM*. (I. Corporation, Editor) Retrieved 2020, from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_data_understanding_phase.htm

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.*

Khun, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models.* Chapman and Hall.

Kotler, P., & Keller, K. L. (2006). *Marketing Management.* New Jersey: Pearson Education Inc.

Lantz, B. (2015). *Machine Learning with R.*

Lipton, Z. (2017). *The Mythos of Model Interpretability.*

Maresca, M. (2019, October 7). *Accenture*. Retrieved 2020, from https://www.accenture.com/us-en/blogs/how-accenture-does-it/predictive-models-how-accenture-now-predicts-the-probability-of-winning-sales

Maslow, A. (1943). A theory of human motivation. *Psychological Review*, 370-396.

Matheus, C. J. (1991). The Need for Constructive Induction. *Eighth International Workshop on Machine Learning*, (pp. 173-177).

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*.

Mitchell, T. M. (1997). *Machine Learning.*

MITRE. (n.d.). Retrieved 2020, from https://www.mitre.org/

Motoda, H., & Liu, H. (n.d.). *Feature Selection, Extraction and Construction.*

Murdoch, J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Interpretable machine learning: definitions,methods, and applications.*

Osborne, J. W. (2004). The Power of Outliers (and Why Researchers Should Always Check for Them).

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Pearson Addison-Wesley.

Quinlan, R. (1993). *C4.5 Programs for Machine Learning.*

Reichhald, F. (2003). The One Number You Need to Grow. *Harvard Business Review*.

Reilly, M. (2016, November 8). *Accenture*. Retrieved 2020, from https://insuranceblog.accenture.com/tech-maturity-s-curve-for-insurers

Rogalewicz, M., & Sika, R. (2016). METHODOLOGIES OF KNOWLEDGE DISCOVERY FROM DATA AND DATA MINING METHODS IN MECHANICAL ENGINEERING. *Management and Production Engineering Review*, 97-108.

Rokach, L., & Maimon, O. (n.d.). Decision Trees. In *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*.

Rust, R., & Huang, M.-H. (2014). The Service Revolution and the Transformation of Marketing Science. *33*(2).

Saluja. (n.d.). Retrieved 2020, from https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281

SAS Institute Inc. (2017). *SAS® Enterprise Miner™ 14.3: Reference Help.* Cary, NC. Retrieved 2020, from https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

Sauro, J. (2014). Retrieved 2020, from https://measuringu.com/missing-data/

Schwab, K. (n.d.). *World Economic Forum*. Retrieved 2019, from https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab

Skok, D. (2015). *ForEntrepneneurs*. Retrieved 2020, from https://www.forentrepreneurs.com/2015-saas-survey-part-1/

Solokova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks.*

Tukey, W. J. (1977). *Exploratory Data Analysis.*

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics*.