Churn Prediction in SaaS CAE Industry

João Paulo Sousa Morais

Master's Dissertation

FEUP Supervisor: Prof. Vera Miguéis
Company Supervisor: Sam Prabhu

Integrated Master in Industrial Engineering and Management

June 2021

Abstract

Customer retention is a critical factor for success in Software-as-a-Service companies. Making retention efforts is a priority with this business model, which is why the topic of Churn is relevant in today's world. Understanding who has behaviors similar to a customer that churn, and so is likely to follow the same faith, is a must for companies who want to minimize their retention costs by increasing their efficiency. Once companies are capable of properly predicting risky customers, they have the knowledge of where to invest their time and resources to change that trend, instead of spreading these resources with all the accounts.

Churn is considered to happen when a customer makes the decision not to continue developing businesses with a product or service provider. In SimScale, with a subscription-based business model established, a customer is considered churned when it decides to not renew the subscription, which is why the efforts of retaining this customer need to be done before the subscription end period arrives so that retention activities can be put in place.

Aiming to develop a predictive model, that SimScale could use to analyze their customer behaviors, this project followed a CRISP-DM methodology to better structure its developments. A study on the data collected by the company was developed and after data cleaning and preparation steps, the dataset was arranged to be used as input for machine learning classification techniques.

Different algorithms were tested and Artificial Neural Networks was considered the best to anticipate the churn event. The other algorithms included in the study were the Random Forest, Logistic Regression, Decision Tree, Support Vector Machines, and Gradient Boosting, with some of them showing potential signs of capability under different conditions. Within the modeling process, feature selection strategies were deployed and also the SMOTE, for data balancing purposes since the original dataset is unbalanced. Hyperparameter tuning was performed with nested cross-validation and grid search. To accelerate the grid search method, the first iterations included a randomized function that tested a smaller amount of combinations of parameters, as an initial indication for the surroundings of the best combination achievable.

The proposed model will provide to the SimScale Customer Success Team a tool to complement their identification toolbox of potential churners during the following yearly quarters.

Resumo

A retenção de clientes é um fator decisivo para o sucesso de empresas de software como serviço. As tarefas de retenção são uma prioridade neste modelo de negócio e, por isso, a temática *Churn* é relevante no mundo atual. Identificar os clientes que têm comportamentos semelhantes aos de clientes que previamente cancelaram ou não renovaram e que, como tal, provavelmente irão seguir o mesmo desfecho, é uma obrigação para empresas que desejam minimizar os seus custos de retenção aumentando sua eficiência. Uma vez capazes de prever adequadamente os clientes de risco, as empresas sabem onde investir o seu tempo e recursos para alterar essa tendência, e assim evitar distribuir esses recursos em clientes estáveis.

Churn ocorre quando um cliente toma a decisão de não dar continuidade aos negócios com um fornecedor de produto ou serviços. Na SimScale, cujo modelo de negócios é baseado em subscrições, um cliente é considerado como *churned* quando decide não renovar a subscrição. Razão pela qual os esforços para reter esse cliente precisam de ser feitos antes que o período de término da subscripção seja atingido para que as atividades de retenção possam ser colocadas em prática.

Com o objetivo de desenvolver um modelo preditivo que a SimScale possa utilizar para analisar o comportamento dos seus clientes, este projeto seguiu uma metodologia CRISP-DM para estruturar os seus desenvolvimentos. Foi desenvolvida uma análise sobre os dados recolhidos pela empresa e, após as etapas de limpeza e preparação dos mesmos, estes foram organizados para puderem servir de objeto de estudo para as técnicas de classificação de *machine learning*.

Vários algoritmos foram testados e a *Artificial Neural Network* foi considerada a melhor técnica para antecipar o evento de *churn*. Os restantes algoritmos incluídos no estudo são *Random Forest*, *Logistical Regression*, *Decision Tree*, *Support Vector Machine* e *Gradient Boosting*, com alguns deles a apresentarem sinais de capacidade em diferentes cenários. No decorrer do processo de modelagem, estratégias de seleção de recursos foram incluídas, tal como a técnica SMOTE, com o objetivo de balancear os dados, uma vez que o conjunto de dados original é caraterizado por um desequilibrio no total de clientes *churned*. O ajuste de parâmetros foi realizado com *nested cross-validation* e *grid search*. Para agilizar o método de *grid search*, as primeiras iterações incluíram uma função aleatória que testou uma quantidade menor de combinações de parâmetros, como uma indicação inicial da vizinhança da melhor combinação possivel de parametros.

O modelo proposto é uma ferramenta de identificação de potenciais *churners*, preparado para ser utilizado pela equipa de Customer Success da SimScale.

Acknowledgements

My first word of appreciation goes to Prof. Vera Miguéis for the support, guidance and availability shown during the entire project.

I would like to thank the SimScale Customer Success Team, in particular to Sam Prabhu, for trusting me with this challenge and supporting me during the last six months.

I would like to thank my friends and family for all the fun and great moments shared. They gave me the necessary energy to move forward with the project and the challenges that arouse from it.

Lastly, I would like to thank my great parents and sister for not only the support and encouragement given during this project but more importantly for always attempting to challenge me to do better and be better.

"Without data you're just another person with an opinion."

W. Edwards Deming

Contents

Acronyms and Symbols

AHP Analytical Hierarchy Process

ANN Artificial Neural Network

AUC Area Under the Curve

CAE Computer Assisted Engineering

CNN Convolutional Neural Network

CP Churn Period

CRISP-DM Cross Industry Standard Process for Data Mining

CRM Customer Relationship Management

CS Customer Success

CSM Customer Success Manager

DM Data Mining

DT Decision Tree

EDA Exploratory data analysis

FN False Negative

FP False Positive

GB Gradient Boosting

HR Human Resources

IT Information Technology

KDD Knowledge Discovery in Databases

LSTM Long Short Term Memory

LT Latency Period

MAR Missing at Random

MCAR Missing Completely at Random

MNAR Missing Not at Random

OP Observation Period

RF Random Forest

RFM Recency, Frequency and Monetary

ROC Receiver Operator Characteristic

SEMMA Sample, Explore, Modify, Model and Assess

SMOTE Synthetic Minority Over-sampling Technique

SVM Support Vector Machines

TN True Negative

TP Target Period

TP True Positive

List of Figures

List of Tables

Chapter 1

Introduction

The focus of this chapter is contextualizing and presenting some of the problems that exist in the SaaS business model, and in particular, why predicting customer churn can help solve some of those problems.

1.1 Project Motivation

In today's world, Software-as-a-Service based companies need to focus on retaining customers to be financially viable. In this business model, the clients pay a periodical amount to have access to a certain service during a certain period. This recurring revenue system generates a need to sustain clients for a certain period, in order to overcome the initial costs the SaaS companies have to acquire clients. The initial costs can be related to marketing and advertising, salaries of the human force, directly and indirectly, involved with the acquisition of the client, or even the costs involved in the training needed for the client to use the service successfully. The clients need to stay with the company, on average, for more than twelve months to enable the generation of profit (Ge et al., 2017). When a customer decides to churn before the break-even period is achieved, the companies accumulate financial losses. It is also relevant to note that a significant factor in favor of putting efforts in retaining customers is that the costs of acquiring a new customer are higher than the costs of retaining them (Judy Strauss and Raymond Frost, 2014).

The awareness of retention rates and, more in specific, identifying which customers are more likely to churn is critical to a company. Being capable of properly identifying which customers to target and focus on, to change their probability of churning, can be crucial.

This project is a consequence of the SimScale interest in understanding the variables explaining the churn event of past customers and will help the Customer Success Team identifying the clients more likely to churn and to actively implement customer-oriented strategies focusing on those clients and variables in an attempt to change the tendency they are following and possibly extending the subscription period of this clients.

1.2 SimScale and Its Processes

SimScale is the world first production-ready SaaS application for engineering simulation. Founded in 2012 by five friends, David Heiny, Vincenz Dölle, Johannes Probst, Alex Fischer, and Anatol Dammer, started as a consulting company and then, once the founders identified the opportunity and power that cloud services could offer to this industry, moved to a software company with the development of the SimScale Platform. From the beginning, SimScale has been expanding its customer base through various markets, having now clients in all continents of the world. This expansion is creating a need to prioritize actions in the Customer Success(CS) team, which is responsible for nurturing and supporting the clients in their journey with the product the company offers.

Currently, the customers are segmented into categories considering their likelihood of churning. These categories are periodically and manually assigned by the CS team, which represents a huge workload and involves being in constant contact with all customers, which is not always possible. To help with the evaluation of the customer involvement with the service, before this project, some metrics were developed to measure the client's usage, engagement, and relationship with the support team. Based on these scores, the team selects which proactive actions are needed and which customers should be reached to understand possible problems or shift negative tendencies being captured by these metrics.

With time, it was clear that although these metrics were used daily and improved the day-to-day planning of the team, they were not a true indicator of the propensity to churn on the final period of the subscription. In fact, churn could only be anticipated very close to the end date agreed between both parts, when typically there was a contact to clarify this question.

1.3 Objectives

The main goal of this project is the proper identification of the customers that will churn, so that they can be reengaged and churn can be avoided.

To achieve this goal, the expected outcome of this project is a predictive model, capable of identifying customers likely to churn. Moreover, this study aims to explore whether or not the use of existing aggregated metrics like Usage Score, Engagement Score, and Relationship Score have a positive influence on the accuracy of the churn prediction.

This project includes a learning phase where multiple techniques are tested to find the most suitable solution for the context and purpose of the project.

Once SimScale has a proper predictive model, the intention is to integrate it into the existing CRM tools used in the company, so that it can help the CSM's in their day-to-day planning. The project is aligned with the company strategy of allowing the business to scale without compromising the quality of the support given by the team and should also positively impact the retention rates. Once the team can identify the proper clients to reach, the efficiency will improve and they

will be more capable to deal with the expected increase in the number of active clients without a significant increase in the team size.

1.4 Dissertation structure

Following the Introductory chapter, chapter 1 the document presents four more chapters.

The chapter 2, **Literature Review**, compiles a background overview on the relevant topics of the project. Based on previous related work, this chapter will provide various clarifications on the techniques, concepts, algorithms, and methodologies used during this project.

Chapter 3, **Implementation**, presents the process applied in order to achieve the objectives presented in the introductory section. The study followed the established standard methodology CRISP-DM. As such, this chapter follows its guidelines, which includes sections concerning business understanding, data understanding, data preparation, and modeling stage. In the current chapter, the deployment phase is also approached.

Chapter 4, **Experiments and results**, develops the second part of the CRISP-DM modeling data phase. It also comprehends the data evaluation phase, with respect to the decision of whether the developed models meet the business success criteria or not.

To finalize, chapter 5, **Conclusions**, builds, an overall assessment of the developed work. Some possible next steps are also proposed in this chapter.

Chapter 2

Literature Review

Customer churn and Customer Relationship Management in a broader sense, as mentioned before, are relevant concerns for firms in various contexts and in particular for Software-as-a-Service companies, operating on recurring revenue systems, where customer retention is critical for business growth and healthy economic performance (Ge et al., 2017).

With the evolution of Artificial Intelligence and Data Minding over the years and the increasing awareness of its applicability helping on the prediction of churn phenomenons, there have been multiple studies performed on different markets and business models. This chapter intends to present an overview of what studies have been published around this field and, with this, present some relevant techniques and theoretical concepts relevant when discussing this scientific area of knowledge.

2.1 Customer Relationship Management

The CRM concept emerged from the IT world in the '90s. Although it is regularly used to describe technology-based customer solutions like sales force automation's (Mohan and Deshmukh, 2013), and it is common to find it in the academic community used mutually with the term "relationship marketing", CRM is typically used when speaking about technology solutions. It is described as "information-enabled relationship marketing" by Ryals and Payne (2001). CRM capabilities in organizational processes are a reflection of the companies skills and knowledge to "identify attractive customers and prospects, initiate and maintain relationships with attractive customers, and leverage these relationships into customer level profits" (Morgan et al., 2009). Some CRM analytical tools have been developed to help companies with these tasks and normally these applications are capable of analyzing datasets made from the information provided by the customers or collected from the customers usage, by the companies, to give answers to some important questions firms have on today's world (Lazarov et al., 2007).

2.1.1 Customer Retention

CRM techniques are applied with various objectives. Kracklauer et al. (2004) proposed the idea of CRM having impact in four dimensions of the customer journey with the company. The diagram bellow illustrates these dimensions:

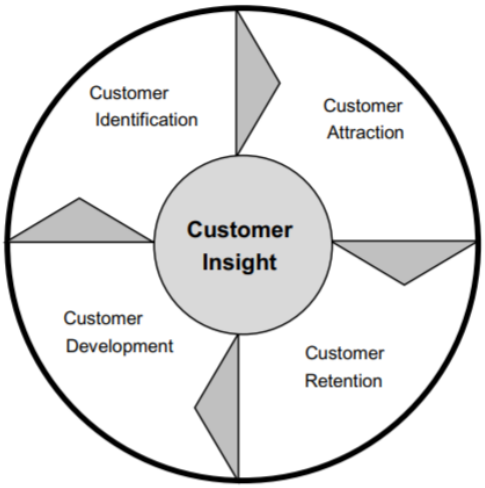


Figure 2.1: Customer Relationship Management Dimensions. (Kracklauer et al., 2004)

The insights provided from CRM techniques can help identify and attract potential customers, develop stronger relations with customers, and retain current customers.

For SaaS business model companies, customer retention is critical for achieving economic success. The recurring revenue system that characterizes this business model reinforces the importance of keeping the existing clients, since they need to remain subscribing the service for a certain period, to be a profitable client to the service provider. Only after that period, the initial cost incurred to acquire the customer is compensated by the subscription periodical value (Ge et al., 2017).

The economic value associated with customer retention is recognized in literature and is explained by Van den Poel and Larivière (2004), based on the following arguments:

1. Successful customer retention decreases the need for seeking new and potentially risky customers, which allows companies to increase focus on nurturing their existing customers and building stronger relationships.

2. Long-term customers buy more

3. Positive word-of-mouth from existing customers typically results in new referrals for the company.

4. Servicing costs tend to reduce with time due to greater knowledge of the existing customer.

5. Stable customers are less sensitive to competitive marketing activities.

6. Customer loss creates opportunity costs because of sales reduction

7. Attracting new customers is five to six times more expensive than customer retention.

Customer retention is a common goal of the use of machine learning techniques to support decision-making within companies (Ngai et al., 2009). Increasing customer retention rates can be done on multiple levels. Companies can develop segmented marketing campaigns where machine learning can have an influence, helping in the segmentation of customer base.

2.1.2 Churn Management

Analyzing and managing the event named churn starts by understanding what defines churn. Generally, churn can be defined as a "marketing-related term characterizing a consumer who is going from one company to another" (Glady et al., 2009). However, in different contexts, it has some distinctive aspects. As an example, in telecommunications, the event of churning happens when a customer cancels the subscription with a mobile service supplier (Wei and Chiu, 2002) and in the gaming industry, it is considered to happen when a player stops playing for a certain time (Rothmeier et al., 2021). Studies focused on churn in a certain industry or specific field should formalize the moment when churn happens to avoid misunderstandings.

Even though defining churn is not a trivial process, it is possible to define three types of churn (Lazarov et al., 2007):

- **Deliberate churn** happens when a customer decides to cancel or not renew a contract with a certain provider to start relationships with a new provider and can be explained by lack of satisfaction with the service or lack of loyalty with the service, which by itself can be explained by multiple reasons, such as not having competitive price plans.

- **Incidental churn** happens when a customer leaves a supplier without the aim of switching to a different one, and can be explained by changes in the circumstances, financial or geographical, for example, resulting in no further need of the service.

- **Non-voluntary churn** happens when a provider decides to discontinue the contract with the customer, which can be justified with abuse or non-payment of the service.

Since Non-voluntary churn is under the companies control, there is no need to predict this event. Moreover, Incidental churn is only responsible for a small part of the company's churn. A large part of the solutions for churn management focus on Deliberate churn (Hadden et al., 2007).

There are two basic paths for managing customer churn: untargeted and targeted approaches. Untargeted approaches, where the main focus is on providing superior product and mass advertising. This approach ultimately tends to increase brand loyalty and retain customers. Targeted approaches focus on the identification of customers likely to churn and preventing it from happening. These approaches can be classified as Reactive or Proactive. The first intends to fight churn only after the customer shows the intention to cancel the contract, and the second intends to

predict churn and try to avoid it before the intention is communicated by the customer (Jahromi, 2009).

Recommendations in the literature are that companies should continuously search for customer dissatisfaction motives and identify the risk in advance, since the chances of preventing churn are higher when using proactive approaches (Valtola, 2019). Companies are investing resources and time, implementing machine learning techniques that can help to identify the customers who require efforts to be kept.

2.2 Churn Prediction

Nowadays, machine learning and machine learning have evolved to a level that has revolutionized churn management processes. Currently, using machine learning models to predict churn is common and widely proliferated across the academic community, but also on the professional one.

2.2.1 Churn definition process

In the process of modeling churn, one of the initial stages is defining the periods under study and what event or description characterizes a churner. The three main periods recognized in churn prediction studies are the observation, latency, and target period (Veloso, 2013):

- **Observation Period:** Past customer behavior data is collected with the aim of training the prediction model. An increase in the amount of data obtained in this stage enhances the accuracy of the model but can also decrease the computational performance, which suggests that multi-data sizes should be tested, if possible, in order to find the right balance.
- **Latency Period:** In this stage, comprehended between the observation period and the beginning of the target period, strategies are implemented to retain customers that were identified as potential churners. The period length depends on the time needed to implement the retention actions.
- **Target Period:** Stage when customers are classified as being churners or not accordingly to what is used to define churn in the context under analysis.

As mentioned in chapter 2.1.2, the concept of churn can be adapted to various industry types and business models. Its definition requires a good understanding of the business and its day-to-day operations (Ahn et al., 2020). The process of defining churn is not always clear and there are a few strategies that help on this characterization (Veloso, 2013):

- **Static:** Definition based on a fixed established period, when a customer is considered to be churned if during a defined period does not make a commercial interaction with the company.
- **Dynamic:** Definition various for each client. Chiang et al. (2003) presented two different examples for this concept, within the network banking service industry. In a first approach,

a customer with no transactions for a period longer than the average interval of time between transactions in the known past. In the second approach, churn is considered to happen when the time between transactions is longer than the longest time interval without transactions known.

- **Partial:** Churn happens when a client reduces the use of a company service to a competitor company. As an example, Oliveira (2012) developed an evaluation based on the expenses of the customer with the goal of determining whether there was a three-month period in which a customer made less than 40% of the purchases of the previous quarter. Identifying partial churn is important for churn management since partial churn can turn into total abandonment of the relationship between the company and the client later on (Buckinx and den Poel, 2005).

Multiple machine learning techniques can be and are applied in literature, to churn prediction related problems. The next section explores what has already been done and conclude what should be done on this field.

2.2.2 Machine learning techniques and explanatory variables

2.2.2.1 Knowledge Discovery in Databases

Machine learning is an established process in different industries and businesses due to its ability to use data to make analyses, predict trends and find patterns. To properly deal with loads of data there is a need to use adequate structured techniques, being one of the firsts and most commonly used ones the KDD. KDD is "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" Fayyad et al. (1996). The authors place machine learning as a step of the KDD process, and as a means to extract patterns from data.

The KDD process consists of five steps (Azevedo and Santos, 2008):

1. **Selection:** this stage consists on creating a data-set where discovery can happen;

2. **Pre-processing:** this stage consists on the target data cleaning and pre-processing in order to obtain consistent data;

3. **Transformation:** this stage consists on the transformation of the data using dimensionality reduction or transformation methods;

4. **Data mining:** this stage consists on identifying interesting patterns according to the machine learning objective;

5. **Interpretation/Evaluation:** this stage consists on the interpretation and evaluation of the identified patterns.

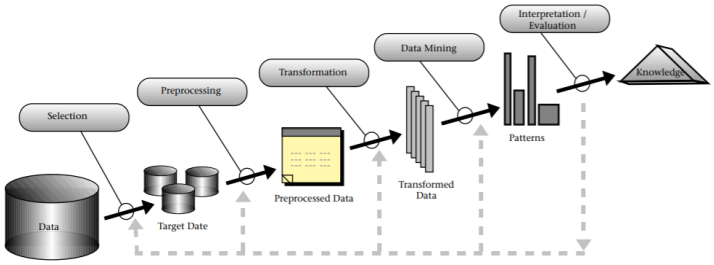


Figure 2.2: An Overview of the Steps That Compose the KDD Process. (Fayyad et al., 1996)

2.2.2.2 Feature related techniques

Once the problem definition is understood the next step is analyzing the data and identifying the predictors, for example, the explanatory variables. Feature engineering is by definition the process of identifying and extracting valuable features to act as inputs for machine learning models (Ferreira, 2019). Kim et al. (2017) tells us that, if executed properly, this feature engineering process improves the prediction performance and enables the discovery of important insights. It is referenced by some authors as the most important task on machine learning projects (Domingos, 2012).

Normally, study groups include the features that intuitively provide the most useful information, but this intuition in professional problems is based on feelings or beliefs from the team involved in the development of the model, which is not always easy to prove or explain. With the goal of getting more value from untreated data, some other teams follow representation schemes that are already structured and work as guidelines for this process. RFM (Recency, Frequency, and Monetary) is a common strategy used for customer segmentation and has a focus on customer behavior. As an example of an RFM application, Sheikh et al. (2019) based its segmentation work for a fin-tech Industry dataset on the RFM model, where recency was measured by the interval length between the last transaction and the end of the period of studies, the frequency with the number of transactions which occurred in the period, and monetary by the aggregated value of the transactions performed during the studied period.

Including irrelevant features in an analysis or study can negatively impact the model performance, which explains why it is critical to properly select the features to include. As mentioned before, the selection can be done manually and based on feeling, but there are various techniques developed for this purpose.

Saeys et al. (2007) explored different techniques of feature selection and aggregated them into three categories based on the principle behind them: filter, wrapper, and embedded.

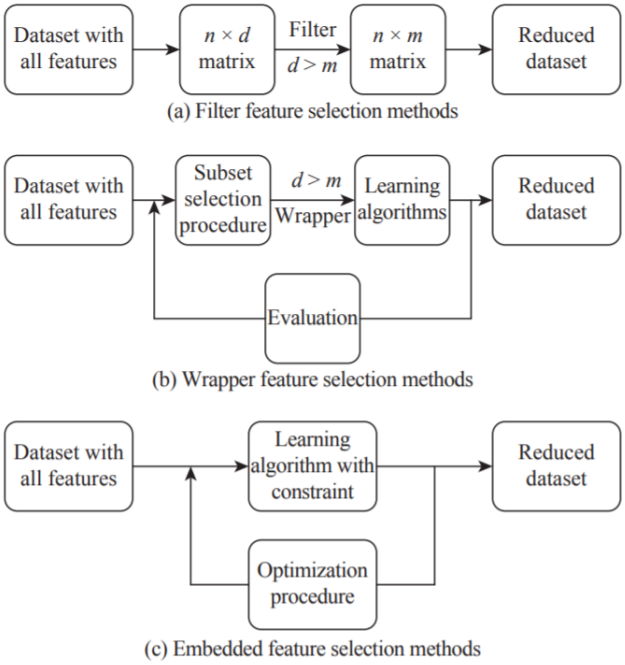


Figure 2.3: Flowchart of filter, wrapper, and embedded feature selection methods. (Liu et al., 2019)

Filter techniques, illustrated in figure 2.3 a), define a ranking over the list of available features by identifying what features are more useful for classification. Approaches in this category assess the relevance of each feature by analyzing the intrinsic properties of the data without any interaction with the classifier and between them, thereby ignoring feature dependencies which may lead to worse classification performance (Beniwal et al., 2012). While trying to overcome the issue of ignoring feature dependencies, various multivariate filter techniques were developed and can partially incorporate some of these dependencies.

Different from filter techniques, wrapper methods, illustrated in figure 2.3 b) include the model hypothesis search in the feature selection process. machine learning algorithms are used to infer the performance of a given set of features and an iterative searching process, considering the possible combinations of feature subsets, is conducted (Beniwal et al., 2012). The wrapper approach typically is more time-consuming and more computationally expensive, since it involves an evaluation of each attribute subset under consideration in the search (Ahmad and Dey, 2005). Commonly used examples of wrapper techniques are forward selection and backward elimination. Forward selection uses an empty set of features as a starting point and from the original set of features, the best is added to the empty set. Every subsequent iteration, the best one of the remaining features is added to the set. Backward elimination, in opposition, uses as a starting point the full set of features and, at every iteration, removes the worst one. Combining both methods by selecting the best attribute and removing the worst from the remaining ones is a possibility to reduce the computing time from the individual methods (Han et al., 2011).

Lastly, the embedded methods, illustrated in figure 2.3 c), in which the search for optimal subsets of features is built into the classifier construction, are seen as a search in the combined space of feature subsets and hypotheses (Beniwal et al., 2012). Embedded methods, like the commonly used Logistic Regression, have the advantage of interacting with the classification model and still being less computationally intensive than wrapper methods.

2.2.2.3 Prediction Algorithms

Prediction problems can be differentiated based on the target variable type the problems focus on predicting. Regression problems, aim to develop a function that maps a set of records to numerical data. Classification problems have the same objective but the target variable is categorical (Sarchana and Kelangovan, 2014).

In churn prediction problems, the goal is to classify a customer as being churner or non-churner and typically are developed as classification problems. Verbeke et al. (2011) defended that for predicting churn, three key aspects must be analyzed, which are: accuracy, comprehensibility, and justifiability. Accuracy in churn prediction models is the percentage of instances correctly determined as churners or as non-churners and it indicates the model's predictive capabilities. Comprehensibility and justifiability are not focused on predicting, but more on understanding the reasons for churn and what can the company do to avoid it. The majority of the studies spend most of the attention on the accuracy of the models, but by guarantying comprehensibility and justifiability, companies are more capable of developing effective and successful retention strategies (Verbeke et al., 2011).

There is not an agreement on which machine learning technique is more suitable for churn prediction problems in general. A large number of studies regarding churn prediction implement different algorithms on datasets and then establish comparisons between them. While trying to predict churn of employees in the human resources business, for example, Bandyopadhyay and Jadhav (2021) implemented different algorithms, which are Random Forest, Naive Bayes, and Support Vector Machine, and concluded that for the dataset explored, the Random Forest technique is the most accurate.

Random Forest is an ensemble algorithm, such as Gradient Boosting. Ensemble algorithms combine different classifiers to obtain better performance than the one obtained with each classifier alone (Polikar, 2012). In particular, Random Forest is an ensemble of decision trees (DT) and a decision tree is a diagram with nodes linked by edges, where nodes are features and edges are classification rules. The final nodes of the diagrams are named leaves and represent an outcome/prediction. DT is used as an individual classifier as well, but when integrated within the RF technique, randomness is added to the model and in the process of splitting a node, it looks for the feature among a random subset of features with the best input for that stage of the tree (Tan et al., 2005). The other ensemble algorithm mentioned above, Gradient Boosting (GB), as the name indicates belongs to the group of boosting methods. These methods focus on converting weak learners into strong ones and GB does this by weighting observations and giving more weight to the ones harder to classify. A loss function is used to assign the weights and present how

the model's coefficients correlate with the provided data. By doing this, the observations badly classified will have a better chance of being used in the next iterations of the model and being better interpreted by the new data used (Ferreira, 2019).

One of the most used techniques in churn prediction problems is the Logistic Regression since it is easy to implement, good at dealing with non-linearity, and presents solid results (Miguéis et al., 2012). Features values are combined with coefficient values, calculated using training data, to predict an output result, which can then, with the logistic function, translate to a value between 0 and 1 which represents the probability of churn.

Khodabandehlou and Zivari Rahman (2017) develop a comparative analysis of Support Vector Machines, Decision Trees, and Artificial Neural Network, with a dataset provided from the grocery retailer industry. It is suggested and concluded from this study that not only variables provided by the RFM method should be included in the model, since by mixing the RFM variables with the variables number of purchased items, the number of returned items, the discount, the distribution time and prize, the predictive capabilities improved. From the predictive techniques tested in this study, the one with higher accuracy was Artificial neural network (ANN).

ANNs are inspired in the way human brain processes information, gathering their knowledge from the data provided and detecting the patterns and relationships in data. An ANN is made from a large number of artificial neurons connected with different weights, creating a neural structure with organised in layers. The neural network behavior is determined by the transfer weights between neurons, by the learning rule implemented, and by the architecture itself (Agatonovic-Kustrin and Beresford, 2000).

Apart from techniques already mentioned above, Kim et al. (2017) explored two deep learning algorithms: Convolutional Neural Network and Long Short Term Memory Pointing. Since deep learning algorithms have provided improvements in performance on different machine learning studies, the authors applied it in a churn related problem, but concluded that no benefit came from these methods and also defended that feature selection and a proper definition of churn for every particular application are more relevant for fighting churn and developing models for it.

2.2.2.4 Evaluation criteria

Once a model is prepared, it needs to be evaluated using proper metrics so that its performance can be measure and, if intended, it can be compared to different models.

The metrics used are highly dependent on the object under evaluation. Commonly used evaluation metrics in churn-related studies are Accuracy, Sensitivity, and Area Under the Curve are obtained from a confusion matrix (Fawcett, 2006). It is a concept designed to summarize the predictive answers of machine learning algorithms.

In machine learning classification problems, as mentioned above, accuracy is frequently analyzed. It calculates the percentage of correct predictions on the test data:

$$Accuracy = \frac{Numberofcorrectpredictions}{Totalnumberpredictions} = \frac{TP+TN}{TP+TN+FN+FP} \tag{2.1}$$

Although it is commonly used, it is important to know that accuracy can be misleading when the dataset is imbalanced, which in churn problems means that the number of churns and non-churns on the dataset is largely different. Such differences are expected in churn-related problems which indicate that accuracy should not be used alone to measure the performance of a churn prediction model (Weiss, 2004).

Other metrics that allow for conclusions to be taken still in the presence of unbalanced data scenarios are available and the already mentioned Area Under the ROC curve (AUC) or other metrics such as F1-score, serve this purpose:

- **F1-score** is the harmonic mean of two simpler metrics, recall and precision. It is an attempt to capture what those two measures capture simultaneously, which can be difficult since normally an improvement in precision is linked to a decrease in recall. Individually, Recall in the churn context presents the proportion of actual churners that were correctly classified and precision calculates the proportion of predicted churners that are correctly classified.

$$Reccall = \frac{TP}{Actualresults} = \frac{TP}{TP+FN} \tag{2.2}$$

$$Precision = \frac{TP}{Predictedresults} = \frac{TP}{TP+FP} \tag{2.3}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

- **Area Under the Curve** and Receiver Operator Characteristic are connected with the curve from the first method being the second. The ROC line, exemplified in figure 2.4 is determined using the TP rate (sensitivity) against the FP rate (specificity), and using customized different probability intervals that determine if a certain value is a positive or negative outcome. AUC assumes values from 0 to 1, where 1 represents a perfectly capable model and 0.5 could be obtained by randomly guessing if a customer will churn or not.

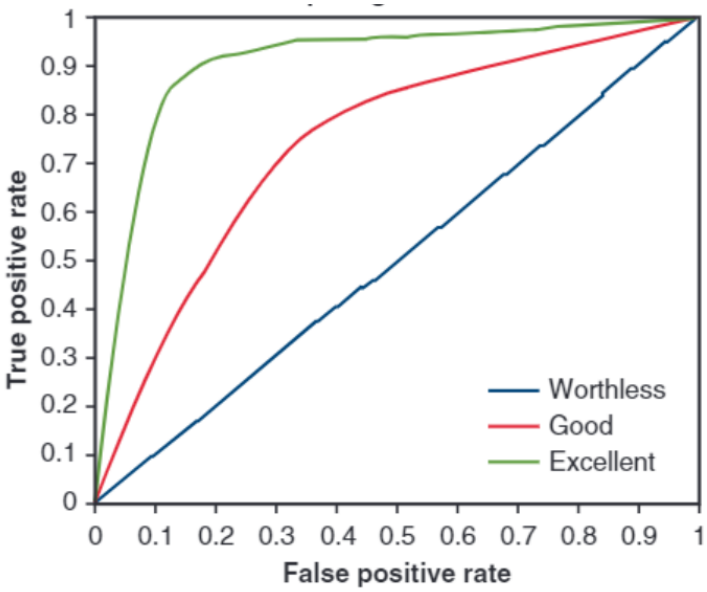


Figure 2.4: ROC Curve Exemplification. (Ferraris, 2019)

From a business perspective, and in particular for churn prediction modeling, the costs of a bad classification of a non-churner are lower than from an improper classification of a churner. After the identification of future churners it is expected that extra efforts are put into action to change the current path of those customers. A non-churner, classified as churner, affect by these retention measures will only be more engaged with the product and will not have major negative consequences apart from possible extra unnecessary expenses taken in the extra efforts to retain these customers. Although a miss-classification of a true churner can result in losing that customer and the respective future cash incomes that it represents. The different impacts of improper classifications tell that, in the churn context prediction models, sensitivity should be prioritized but never neglecting specificity.

2.3 Final impressions

Literature does not provide clear references on which techniques should be used for churn prediction problems in general. There are several alternatives that should be explored to identify the best technique for the particular dataset available.

Every churn problem has it owns particularities, starting on the basic definition of what churn is in that particular context, which makes it difficult to generalize guidelines on how to approach these problems. Studies around churn predictive models become relevant for analysis since later in time they can provide a good base for better conclusions to be taken in this study field.

Chapter 3

Methodology

The amount of data available is increasing and the profits for using it are following the same trend.

In chapter 2.2.2.1, the KDD model to attack machine learning problems is presented. Other methodologies used in machine learning projects, such as CRISP-DM and SEMMA are widely used in academic studies and in industry. When compared, Azevedo and Santos (2008) states that the last two can be seen as an implementation of the KDD.

CRISP-DM incorporates steps preceding and following the KDD stages: Business understanding and Deployment, which in the present study, add value and can be critical for its success (Azevedo and Santos, 2008). The model in total consists of six phases (Shafique and Qaiser, 2014), which will serve as divisions on the present chapter and will help the study achieve the proposed goals presented in chapter 1.3:

1. **Business understanding:** focus on the understanding of the project objectives and requirements from a business perspective; conversion into a machine learning problem; design of a preliminary strategy to achieve those objectives; definition of the business success criteria;

2. **Data understanding:** data collection, familiarization with the data, exploratory data analysis; confirmation of the data quality;

3. **Data preparation:** preparation of the final data-set, from the raw data, which includes tasks such data cleaning, feature selection, creation of new variables, and others;

4. **Modeling:** selection and application of several modeling techniques;

5. **Evaluation:** comparison and evaluation of the models prepared;

6. **Deployment:** implementation of the model developed on the actual business context and confirming its effects on it.

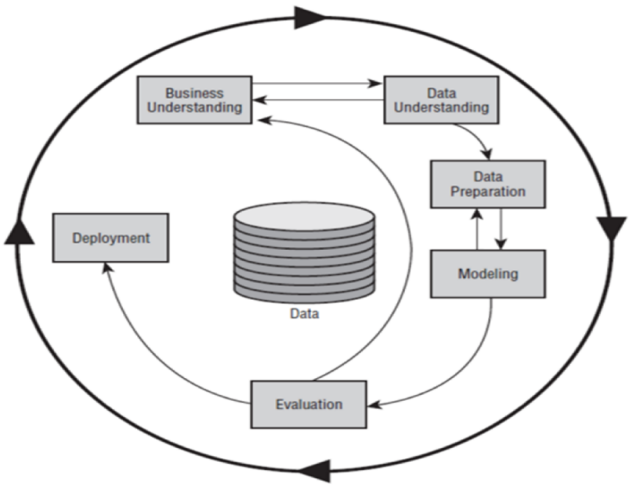


Figure 3.1: CRISP-DM Process Model. (Shafique and Qaiser, 2014)

Solving a problem creates different new ones and questions arise from it. This methodology embraces that fact and helps to deal with the lessons taken from past projects, incorporating them into new ones, as represented in figure 3.1.

This chapter, based on the CRISP-DM method, explains how this project will be developed and implemented.

3.1 Business understanding

The company is facing considerable growth in the customer base, in terms of financial value and number of users. The growth is explained by an increase in the number of new users, followed by new deals successfully finalized or expansion contracts from current customers, and by making sure the retaining rates from existing customers remain high.

Realizing customer retention is critical for the successful growth, due to the reasons explained in chapter 2.1.1, the company used as case study has putting efforts to improve the existing processes used to identify risky customers and ultimately improve the efficiency of tasks related to customer retention, also named churn fighting tasks.

Nowadays, the Customer Success Team is responsible for following the developments of every paying user of the platform. This means nurturing customers during the process and the license period, identifying any negative signs that might indicate future churn. When potential churners are spotted, the team is expected to evaluate the case, and ultimately attempt to change the expected outcome for those customers by following retention action plans developed internally. Every quarter of the year, a list of clients with ending contracts on the following quarter is generated and these customers are associated with a category that describes the likelihood of them to churn. The available classifications are explained in table 3.1:

Table 3.1: Retention status classification.

Category	Description
Likely Yes	The customer is expect to renew
Likely No	The customer is expect to churn

The classification is manually done by the team, based on the judgment of the Customer Success Managers, their interactions with the customers, and the analysis of daily calculated aggregated measures explained in detail in chapter 3.2.2.

With the already mentioned enlargement of the number of customers to nurture and support, it is more important than before to improve the efficiency of the identification and prioritization of customers to target. Only this way, the team can provide the same quality of support and preserve the positive retention rates, without also scaling the costs involved in these processes. The introduction of a churn prediction model can help obtain this goal since the manual task performed today could be eliminated and help the team recognize earlier which customers are likely to churn, even without necessarily even communicating with them. Such models can also provide knowledge to the company since they help understand what is justifying customer abandonment.

The study will take into consideration the entire list of clients from the company. Currently, the customers are segmented accordingly to the financial value paid by the client, and a subjective evaluation of what these value could be in the future with possible expansion deals. Intuitively the company is more interested in retaining larger customers due to the future cash flows that those clients represent. However, considering that these clients will be receiving more attention from the team due to its high demands, the automatic prediction can have bigger importance to capture what is happening in lower value accounts that tend to be more autonomous and harder for the team to follow. Such factor explains why the entire list of current and past clients, independently of its characteristics, was considered.

The ultimate goal of this study is the development of a model to be applied by the Customer Success Team, which will allow the team to better identify the customers who will not renew in the following quarter. With this objective, the event of churn is formally defined using the Observation Period (OP), Latency Period (LP) and Churn Period (CP), with the three lasting four months individually. Customers who during the OP did not declare the intention to not renew and that later in the CP present that intention or stop paying for the service are considered churners. More information on the data available for this study will be given in chapter 3.2.2, but due to limitations on this area, the OP considered for the modeling part was shorted to the day the clients informed the intention to churn, which explains why the results from the study need to be carefully evaluated and its outcome can be considered a proof of concept to be replicated later when more adequate data is available. More information on these topic will be provided later, on chapter 4 and chapter 5. A representation of the periods used and the ones intended for later use are illustrated in figure 3.2 and 3.3.

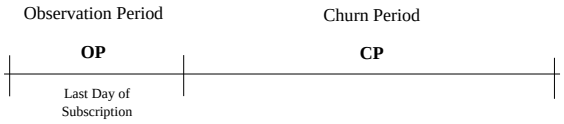


Figure 3.2: Churn definition periods used.

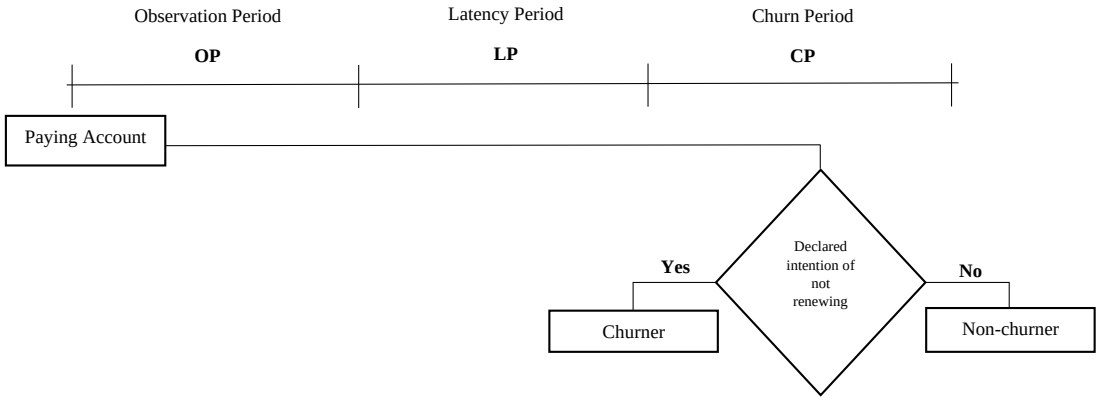


Figure 3.3: Churn definition periods intended for the deployment stage.

3.2 Data understanding

This second phase of CRISP-DM process, as mentioned above, focuses on collecting data, confirming its quality, and exploring it to reveal hidden information.

3.2.1 Data collection

SimScale saves information about their clients, their usage, and their interactions. Initially, a study on what variables were available and could add value to the model was performed. Once that was clear, a report with all the existing accounts in the CRM tool was generated with all the previously selected variables included in a single table.

On this initial report, collected on day 11/06/2021, there was a total of 1387 accounts. However, there was a need to filter the results to only include customer that paid a minimum of one invoice, have a clear definition of active or churned customer, have the Churn Scores registered (explained with detail in chapter 3.2.2), properly tracked and without missing values. The existence of accounts without this information is explained by the process followed by the existing sales team. During the process of looking for potential clients, and when a lead shows interest in the product, an account with its properties is created and left without these details until the contract is signed and the onboarding stage starts. It happens that the some deals never move over this stage and the accounts are left with those empty fields on the CRM dataset. With these criteria

list applied, the data set was left with 435 accounts and there is a guarantee that all the data being tracked is meaningful and, ultimately, relevant for modeling.

3.2.2 Initial data description

After collecting the data, it is important to understand the variables used in the project. To properly develop any analysis and infer any conclusion, there is a need to recognize the OP and CP under which the data was obtained. For this study, as mentioned in chapter 3.1, the OP and CP was neglected to last data available for each client, due to a lack of data availability to approach the problem in another way. For Churned customers, the data used was regarding the last day each account was considered as Non-churner, and for active customers the data regards the last day before the predictive model was created.

The initial data set is structured by different types of variables. Some variables are simple geographical characteristics of each account, such as *Billing Country*, others focus on financial aspects like *Total Won Amount*. Some variables focus on characterizing the relationship with the client and its usage, like *Depth of Relationship* and *Last Platform Job*, respectively. All the variables included in the data set are listed and explained in table A.1.

Three particularly interesting variables included are *Usage Risk*, *Engagement Risk*, and *Relationship Risk*. These are aggregated measures used daily by the Customer Support Team to evaluate the account usage, the relationship developed by the Support Team and the End users on the client's side, and the engagement of the decision-maker with SimScale. On tables 3.2, 3.3 and 3.4 the variables used to calculate each individual scores are detailed and also the weights considered for all the accounts.

Table 3.2: Usage Score calculations.

Usage Score	Variables	Weight
	Days since last job	50%
	Computing quota remaining (%)	20%
	Perceived satisfaction	20%
	Number of support cases	10%

Table 3.3: Engagement Score calculations.

Engagement Score	Variables	Weight
	Days since last Business Review	40%
	Days since last contact	30%
	Sentiment / Mood	20%
	Days since last approach	10%

Table 3.4: Relationship Score calculations.

Relationship Score	Variables	Weight
	Depth of Relationship	30%
	Total number of overdue invoice days	20%
	Tenure Length	20%
	Number of platform licenses	20%
	Success Story	10%

Ultimately these three scores are aggregated in a single one name ChurnScore, also integrated into the initial data set. In table 3.5, the ChurnScore calculation is deconstructed and the weights of each score used to calculate the general one are shown.

Table 3.5: Churn Score calculations.

Churn Score	Weight
Usage Score	50%
Engagement Score	25%
Relationship Score	25%

3.2.3 Exploratory Data Analysis

Before preparing the data it is important to analyze in detail the available data, its trends, and relations between variables since, for example, some data analysis techniques are limited on the type of variables that they are able to process. Exploratory data analysis (EDA) is a step from the Data understanding stage, where the main goal is to identify patterns, outliers, collinearities and correlations Schuh et al. (2019). EDA techniques also have an impact on facilitating the variables description and summarization Akrong Hesse and Ofosu (2018).

In this section, the variables listed in section 3.2.2 are also carefully analyzed with the objective of revealing quality data problems, that with an early identification can be treated easily, and will result in a smother data preparation phase.

Myatt (2007) proposes that an insightful initial categorization of each variable, based on the type of values and the measurement scale used to compare them, is needed to properly understand the data. Categorizing the variables allows making future choices on how to visualize data or even how to read it. There are two types of variables in the dataset: Numerical and Categorical.

The list of variables explored and their classification is presented in table 3.6.

Table 3.6: Initial dataset features.

Variable	Variable type	Level of measurement
Billing Country	Categorical	Nominal
ChurnScore (Default)	Numerical	Interval
ChurnScore Engagement Risk	Numerical	Interval
ChurnScore Relationship Risk	Numerical	Interval
ChurnScore Usage Risk	Numerical	Interval
Depth of Relationship	Numerical	Ratio
Last Business Review Date	Categorical	Ordinal
Next Business Review Date	Categorical	Ordinal
Last Platform Job Date(Day)	Numerical	Ratio
License Count	Numerical	Ratio
Number of Support Cases	Numerical	Ratio
Perceived Satisfaction	Numerical	Ratio
Potential	Categorical	Nominal
Sentiment / Mood	Numerical	Ratio
Success Story / Case Study Status	Categorical	Nominal
Tenure (Days)	Numerical	Interval
Total Won Amount	Numerical	Interval
Usage Frequency	Categorical	Ordinal
Valid Computing Quota Amount Assigned	Numerical	Interval
Valid Computing Quota Amount Used	Numerical	Interval

With the variable types distinguished, during the next part of the current section, the characterization will be different for numerical and categorical variables.

Regarding qualitative variables, and following Komorowski et al. (2016), a simple univariate EDA method for categorical variables is conducted . This analysis is shown in the table 3.7.

Table 3.7: Summary of initial categorical variables value and frequency.

Variable	Value	Frequency
Billing Country	Australia	2.88%
	Austria	0.96%
	Barbados	0.24%
	Belgium	0.72%
	Brazil	0.72%
	Canada	5.52%
	Chile	0.48%
	China	0.24%
	Chinese Taipei	0.48%
	Colombia	0.24%

Variable	Value	Frequency
	Croatia	0.72%
	Denmark	1.92%
	Faroe Islands	0.24%
	Finland	1.68%
	France	2.88%
	Germany	10.79%
	Hong Kong	0.24%
	India	1.20%
	Ireland	0.96%
	Israel	1.92%
	Italy	2.88%
	Japan	1.20%
	Korea, Republic of	0.24%
	Lebanon	0.24%
	Malaysia	0.24%
	Mexico	0.24%
	Morocco	0.24%
	Netherlands	3.12%
	New Zealand	0.24%
	Norway	2.64%
	Oman	0.24%
	Peru	0.48%
	Poland	0.48%
	Qatar	0.72%
	Russian Federation	0.24%
	Singapore	0.48%
	Slovakia	0.24%
	Slovenia	0.24%
	South Africa	0.48%
	Spain	0.48%
	Sweden	1.20%
	Switzerland	4.08%
	Thailand	0.96%
	United Arab Emirates	0.48%
	United Kingdom	13.43%
	United States	29.50%
Last Business Review Date	Date type variables	
	...	
Next Business Review Date	Date type variables	

Variable	Value	Frequency
	...	
Potential	High Value	10.79%
	Low Value	58.27%
	Medium Value	30.94%
Success Story / Case Study Status	Confirmed	4.76%
	Low Chance	2.38%
	Published	47.62%
	Rejected	13.10%
	Selected	9.52%
	Suggested	16.67%
	Work in Progress	5.95%
Usage Frequency	Daily	4.56%
	Inactive	74.10%
	Occasionally	11.03%
	Weekly	10.31%

From the frequency values listed in table 3.7, and regarding the *Billing Country* variable, it is possible to conclude that a considerable amount of countries included in the initial data set have a residual presence. In section 3.2.3.1, a transformation to this residual values will be defined.

From table 3.8, some characteristics of the numerical variables became clear. Negative values were not expected, however they exist in the *Last Platform Job Date(Day)* and *Tenure(Days)*. Different ranges between variables are detected, i.e analysing the max and min values of *Depth of Relation* and *Valid Computing Quota*. Both these characteristics will be approached in section 3.3.

For quantitative variables, and following Komorowski et al. (2016), several descriptive measures were computed. Table 3.8 collects, for each variable, the mean, minimum, median, maximum and 1st and 3rd quartiles.

Table 3.8: Summary of initial numerical variables.

Variable	Mean	Min	1st Quartile	Median	3rd Quartile	Max
ChurnScore (Default)	59.1	4.0	47.0	61.0	73.0	90.0
ChurnScore Engagement Risk	60.2	0.0	50.0	60.0	80.0	90.0
ChurnScore Relationship Risk	61.5	0.0	50.0	60.0	80.0	100.0
ChurnScore Usage Risk	53.2	0.0	30.0	50.0	70.0	90.0
Depth of Relationship	5.9	1.0	4.8	6.0	7.0	10.0
Last Platform Job Date(Day)	101.8	-181.0	6.0	42.0	133.0	1204.00
License Count	1.3	0.0	1.0	1.0	1.0	34.0
Number of Support Cases	2.1	0.0	0.0	1.0	3.0	26.0
Perceived Satisfaction	5.9	0.0	5.0	6.0	7.0	10.0
Sentiment / Mood	5.6	1.0	4.0	6.0	7.0	10.0
Tenure (Days)	537.5	-44.0	212.0	368.0	730.0	2314.0
Total Won Amount	8429.8	0.0	4500.0	6239.9	9000.2	153187.0
Valid Computing Quota Amount Assigned	24479.6	0.0	8000.0	12000.0	20000.0	1666012.0
Valid Computing Quota Amount Used	7677.0	0.0	281.5	2317.5	6943.2	306243.0

3.2.3.1 Variables explanation and analysis

Based on the data capture in table 3.7, table 3.8, and graphical techniques explained during this section, some conclusions were taken. Graphical EDA methods summarize data in a diagrammatic/pictorial way. Both univariate and multivariate methods are shown in this section.

1. Starting by analyzing the existence of missing values, it is evident from figure 3.4 that the variables *Depth of Relationship*, *Last Business Review Date*, *Next Business Review Date*, *Perceived Satisfaction*, *Sentiment / Mood*, *Success Story / Case Study Status* and *Valid Computing Quota Amount Used*do not have data for all the data samples. Dealing with missing values involves understanding the reasons that justify their existence and this process will be developed in section 3.3.

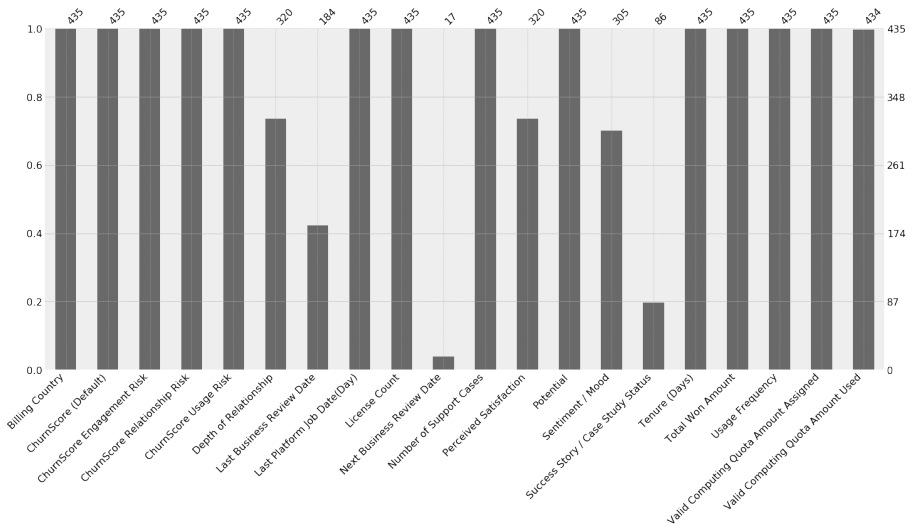


Figure 3.4: Summary of non missing values.

2. The large majority of the clients are considered Low Potential, which is understandable since deals with higher Potential are generally more complex to identify since they involve extra steps for understanding expansion potential. It is also verified in figure 3.6 that the majority of the accounts have similar amounts of *Total Won Amount*, although the higher values in this field are encountered in High-value accounts.

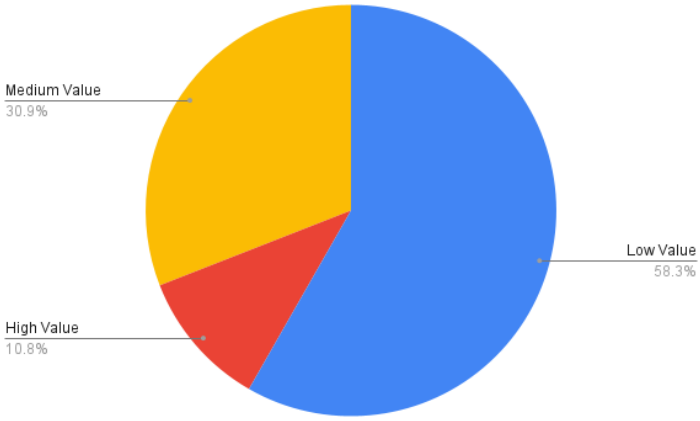


Figure 3.5: Number of clients proportion of each *Potential*.

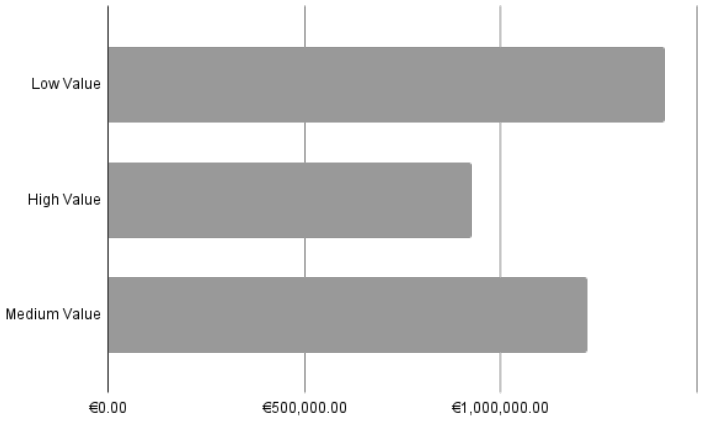


Figure 3.6: Total Won Amount distribution for each Potential type.

3. The existence of negative values on the variables *Last Platform job Date(Day)* and *Tenure(Days)* indicates a need for adjustments since these values should not exist. They measure the elapsed time between the last platform job executed by the client and the length of the relationship, respectively, with the day when the data was extracted or for Churned clients, i.e. the last day of the subscription.

4. Looking into the *Billing Country* variable available in the initial dataset, there are 47 countries listed. Since a large number of these countries have a residual frequency within the

dataset, only five of them (United States, United Kingdom, Switzerland, Germany, and Canada) collect more than 60% of the accounts.

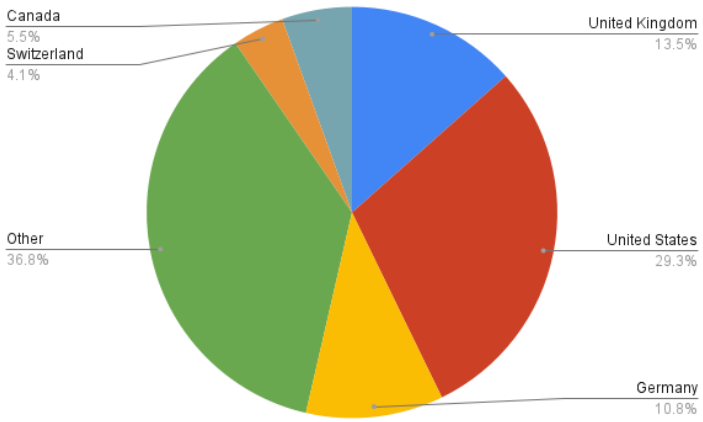


Figure 3.7: Countries distribution.

3.2.4 Data quality

From the previous section, it becomes clear that the initial dataset has some problems needed fixing. Finding solutions for these problems is critical for the success of the data preparation and modeling stage.

There is not a single solution to improve data quality because it is highly dependent on the specific dataset, which is why a solid data understanding can significantly influence the quality of our data (Pipino et al., 2002).

From the figure 3.4, and more clearly on figure 3.9, it is evident that some variables have a high number of missing values that in the modeling stage can not be there. Other questions to approach in the data preparation section are the existence of the negative values, and in particular on the *Billing Country* variable the existence of a high number of low frequent answers that can be aggregated.

Table 3.9: Missing Values count and frequency.

Variable	Number of missing values	Frequency
Billing Country	0	0%
ChurnScore (Default)	0	0%
ChurnScore Engagement Risk	0	0%
ChurnScore Relationship Risk	0	0%
ChurnScore Usage Risk	0	0%
Depth of Relationship	115	26.43%
Last Business Review Date	251	57.70%
Last Platform Job Date(Day)	0	0%
License Count	0	0%
Next Business Review Date	418	96.01%
Number of Support Cases	0	0%
Perceived Satisfaction	115	26.43%
Potential	0	0%
Sentiment / Mood	130	29.88%
Success Story / Case Study Status	349	80.23%
Tenure (Days)	0	0%
Total Won Amount	0	0%
Usage Frequency	0	0%
Valid Computing Quota Amount Assigned	0	0%
Valid Computing Quota Amount Used	1	0.23%

3.3 Data preparation

After the Data Understanding section, the insights obtained are translated into improvements in the dataset during the current section. Preparing data can include different tasks: data cleaning, recording, selection and production of training and testing datasets, or elements may be merged or aggregated in this step (Nisbet et al., 2017).

3.3.1 Data cleaning

The data cleaning process is used to identify unreasonable or inaccurate data and then improve its quality through corrections or elimination of errors and omissions (Natarajan et al., 2010). Following the same order used to approach the identified problems in section 3.2.4, in the current section missing values are the first targeted problem.

3.3.1.1 Missing Values

Missing values can be explained by multiple reasons and it is important to understand that data is not always collected with the goal of being used for machine learning ends. A simple misunderstanding of the team for manual input variables, or a shift in the importance of a certain variable that becomes relevant only after a certain period, can justify the lack of some values. Problems related to incomplete datasets are common since the majority of the existing modeling techniques require complete datasets which are hard to find in the industry Osborne (2016).

Missing data represents different problems, such as a reduction of statistical power, creation of bias in the estimation of parameters, reduction in the representativeness of the samples, and it makes the task of analyzing the study results harder.

In 1976, Rubin described three popular missing data mechanisms, based on the assumptions previously mentioned, describing the relation of the variables and the chances of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). This classification is based on the relation between the reason of missing values and the other variables.

The missing values from the current dataset are presented in table 3.9, and have different justifications:

- *Depth of Relationship* - This variable is a manual input provided by the CS Team, and indicates how strong is the relation between the users and the Account Owner. Null variables mean that this information was not provided and this can be explained by a lack of knowledge to fill the field or a simple forgetting to provide this information. Since there is no pattern on why the data is missing, it is assumed as an MCAR. Since these values represent a significant percentage of the information (26.43% of the total accounts collected), the samples are maintained and the null values are replaced with the statistical mode of the remaining samples.

- *Last Business Review Date* - This variable represents the date of the Last Business meeting. It is considered an MNAR and when a value is missing, it means no business meeting happened with such client. This variable does not need to be filled since, as explored in section 3.3.2, it will be transformed into a new one that does not have the missing value problem.

- Next Business Review Date - This variable represents the date of the Next Business meeting. It is considered an MNAR and when a value is missing, it means no business meetings are scheduled with such client. Therefore, there is a reason for them to be missing, related to the variable itself. Similar to the previous variable, the problem will be solved in section 3.3.2.

- *Perceived Satisfaction* - This variable is a manual input provided by the CS Team, and indicates how the team believes the client's satisfaction is. Null variables, similar to what happens in the *Depth of Relationship* variable, means that this information was not provided

and this can be explained by a lack of knowledge to fill the field or a simple forgetting to provide this information. Since these values represent a significant percentage of the information (26.43% of the total accounts collected), the samples are maintained and the null values are replaced with the statistical mode of the remaining samples.

- *Sentiment / Mood* - This variable is a manual input provided by the CS Team, and indicates how the team believes the client's overall satisfaction with the subscription. Null variables, similar to what happens in the *Depth of Relationship* and *Perceived Satisfaction* variable, means that this information was not provided and this can be explained by a lack of knowledge to fill the field or a simple forgetting to provide this information. Since these values represent a significant percentage of the information (29.88% of the total accounts collected), the samples are maintained and the null values are replaced with the statistical mode of the remaining samples.

- *Success Story / Case Study Status* - After analyzing this specific variable, it was verified that the field is only completed when there is the intention to develop these initiatives with the customers. These values are MNAR and null values represent a new category named "No intention" and the problem is solved.

- *Valid Computing Quota Amount Used* - This variable counts the number of core hours/-computing quota already spent by the customer. It is not expected to have missing values, since it should have a starting value of zero and increase with the computing time used. It is considered an MCAR. Since these values do not represent a significant percentage of the information (0.23% of the total accounts collected), the samples are removed.

3.3.1.2 Inconsistencies

As mentioned previously, the dataset has some inconsistencies detected during the data quality stage.

Some variables that are expected to only carry positive values are presenting some rows with negative ones. Visible in the Minimum statistical field, shown in table 3.8. The variables *Last Platform Job Date(Days)* and *Tenure(Days)* have these problem and both are justified by an error in the organization of data. To fix and prevent further errors, samples under these conditions were eliminated.

3.3.2 Data construction

Initially the data set comprises two variables named *Next Business Review Date* and *Last Business Review Date*, that register the date from the last and the next scheduled Business meeting for each customer, respectively. The features mentioned in section 3.3.1, have missing values, that are explained by having customers where simply this meeting did not exist in the past and is not scheduled for the future. During this section two variables are developed to better caption these

information and avoid the missing values problem. A summary of the features developed in this section is available in table 3.10.

A variable is created to replace the *Next Business Review Date*. Assuming that the date of the Next business review is booked by the CS team, and this precise date does not have any additional information apart from being a proof of openness and interest from the clients side, to discuss business alignments and product development's. The new variable will contemplate two possible values, "No next Br" and "Yes next Br", and they are defined based on the existence of a Next Business Review date.

The *Last Business Review date* variable, as it is initially, is considered a categorical variable and does not allow to establish conclusions on the elapsed time since the meeting happened and the present. By creating a new variable that analyzes this date and categorizes the time period by the number of yearly quarters in between, or in cases where no meeting was arranged, it is defined as "No", the relevant information is clearer and easier to access.

Table 3.10: Summary of new variables added.

Variable	Description	Calculation Rule
Has Last Business Review Date	Number of yearly quarters since the last Business Review meeting was performed.	Categorization of the elapsed time between the last BR date and the present day.
Has Next Business Review Date	Whether a next Business Review meeting is scheduled or not.	Categorization of existing value of next BR date.

3.3.3 Data formatting

At the end of the Data preparation stage, data is expected to be ready to use in the modeling stage. Considering machine learning techniques generally require numeric data, and the project dataset has categorical variables, not in this format, there is a need to convert the categorical variables into numerical ones. For this purpose, the Scikit-learn library for Python was used, similarly to the rest of the data modeling stage procedures applied. It offers different options to make the conversion, namely One-Hot Encoder, whose effects are exemplified in table 3.8.

The One-Hot Encoder is selected, and it converts the categorical variables by creating a new binary variable for each category, with a value of 0 or 1 if the original value corresponds to that category or not, respectively. The technique is not perfect and its major drawbacks are the larger number of features that result from it and not being the optimal choice when using tree-based models, such as DT or RF. For splitting algorithm's like tree-based ones, the binary variables created are considered independent from each other, which makes them unlikely to be early choices on the tree building, with the continues variables gaining relevance.

id	Value
1	A
2	B
3	C
4	D
5	E
6	B

(a) Without encoding.

id	A	B	C	D	E
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	0	1	0	0	0

(b) One-Hot Encoding transformation.

Figure 3.8: Example of Encoding techniques transformation effects.

Another important question to approach for improving the dataset quality is the different ranges present in the variables. Some machine learning techniques are sensitive to these differences with variables with larger ranges having more influence in the models. Scaling techniques have been developed and force the variables to have similar ranges.The commonly used Standardization technique, mandatory for some machine learning techniques, is applied recurring to the Scikit-learn library, mentioned above, which includes the StandardScaler function. It converts the variables in order to have a mean equal to 0 and a standard deviation value equal to 1.

The proposed changes are assembled with the Pipeline function available in the Python library that was used in the project and the resulting dataset is appropriate to be used in the modeling stage. Although we could use this dataset as it is, further improvements are executed in the following two sections.

3.3.4 Data selection

Understanding which variables should be used as inputs for the algorithms can positively impact the efficiency and performance of future stages of the project. Feature selection techniques have as outputs more efficient datasets, less overfitting probabilities, and reduction of computational time (Beniwal et al., 2012).

To discover high correlated variables, the correlation values are estimated for all the combinations of variables in the current dataset, already including the features mentioned during the current chapter. Features with high correlation carry redundant information that can be eliminated to reduce the data size and feature count, ultimately helping to identify patterns and trends easier (Shardlow, 2016). Correlation calculation techniques take into consideration the type of variables under analysis. The available methods have different scales and should not be compared. Three possible combinations of variables are found in the project dataset and different methods were selected for this task, according to that:

- **Categorical variable / Categorical variable:** For a combination between two categorical variables, the analysis will be based on the Chi-squared and Cramer's V tests. For the

chi-squared statistics, the null hypothesis is rejected when the p-value is below a reference value of 5%, normally, and when rejected, dependency between variables is detected. Table A.2 summarizes for the combinations available, the results obtained with this test and it is possible to identify some dependencies. One limitation of the Chi-squared test is not allowing to evaluate how strong the relation between the the variables is, and while trying to understand that, Cramer's V is used since it allows to obtain this measure and takes in consideration the dataset size, which is seen as a limitation of the Chi-squared test as well. For this analysis the Scipy Python library was adopted. The output value from the Cramer's V is within the range of [0, 1], with a higher values indicating stronger relations. In table A.3 the calculated values for the Cramer's V test are summarized. Crewson (2016) proposes that for values until 0.1, association between variables is limited or non-existence, and between 0.1 and 0.3 there is a low association. For most combinations of variables in the study dataset the test values are within these two categories which is not significant enough to use as an argument to consider removing variables from the dataset. Although, between *Potential* and *Billing Country*, and between *Has Last Business Review Date* and *Has Next Business Review Date*, the values are in a category where relation is considered moderate, within 0.3 and 0.5, which is explained by the nature of the variables.

- **Categorical variable / Numerical variable:** For a combination between a categorical and numerical variable, the analysis will be based on the Kruskal-Wallis H-test. ANOVA test is not adequate since the dataset is not normally distributed and the standard deviations of the variables are different, which are requirements for the use of this technique (Cardinal and Aitken, 2005). Kruskal test is used to evaluate if, between two or more groups of a categorical variable, there are significant differences on a numerical one. The higher the H value obtained for each combination, the stronger evidence is to support the null hypothesis, which states the difference between some of the groups is statistically significant. The values obtained are illustrated in the table A.4, and the relations that stand out, due to the fact that high H values are obtained for *Usage Frequency* and *ChurnScore Usage Risk*. This is understandable since both are metrics related to the usage of the clients and focus on measuring the frequency and quality of the usage.

- **Numerical variable / Numerical variable:** For a combination of two numerical variables, the analysis will be based on the Pearson correlation coefficient. The method evaluates the linear dependence of variables with a classification between the range [-1, 1]. Positive values mean that features follow the same rising or downtrends. For example, if feature A increases, then feature B will also increase, and the same goes for decreasing trends. Negative values mean that if features follow opposite trends. The correlation between the numerical variables can be found in table A.5. Extreme values of 1 and -1 represent high correlation, and has expected the higher values calculated are between the ChurnScores variables and others that, has explained in section 3.2.2, are used to calculate these scores.

From the analyzes developed, some choices could be done regarding the removal of some correlated variables.

The variables used to calculate the aggregated Scores, detailed in section 3.2.2, present a considerable correlation with these scores. In chapter 4, one experiment will test whether or not the use of these variables has a positive or negative impact on the predictive capabilities of different models.

Although categorical variables show significant statistical correlation, since the dataset size and the number of variables are not large, and the metrics focus on different points of the accounts life cycle, the study will remain considering all these variables.

Other feature selection technique analysed and tested during the project was the filter technique based on the variables importance score provided by the random forest technique, in resemblance to other studies mentioned in the literature review. These scores allow some interpretation of what variables are more related to the dependent variable. Variables with low importance's were removed from the dataset, until the total number of variables was reduced to half the original size.

3.3.5 Data sampling

The number of churners in the data set under analysis only represents 17.3% of the total number of accounts listed (72 churners for 417 accounts). This is defined as an imbalanced class and needs to be corrected in order to be used in predictive algorithms, since they operate under the assumption that data is equally distributed and are biased to classify all samples as the larger class Fernández et al. (2018). A typical strategy to deal with unbalanced datasets is re-sampling, which based on the original dataset rebuilds it into an improved one. This can be achieved with under-sampling or over-sampling techniques, with the difference being the principle behind them. Under-sampling focuses on reducing the majority class by removing some of its samples, and Over-Sampling focuses on adding samples to the minority class.

The technique used in this project is the Synthetic Minority Over-sampling Technique (SMOTE), recurring to the python imbalanced-learn library. It generates artificial samples, based on already existing samples of the minority class, and adds them to the original dataset. Farquad et al. (2014) developed different tests, using different percentages of over/under-sampled data, a combination of both and SMOTE, and SMOTE achieved the best results. Although over-sampling can be the cause of overfitting, it was selected, due to the small dataset available that with underfitting would even be more restricted, and relevant information would probably be lost in the process.

In chapter 4, the effects of SMOTE will be tested in comparison to results with unbalanced data.

3.4 Modeling

Section 3.3 of the current chapter has the main objective of preparing the data to a point that it is adequate for the modeling stage of this machine learning study. In this section of the chapter, the modeling techniques adopted are presented.

3.4.1 Modeling technique

The first step in modeling stage is splitting the data into two groups, which are named training set and test set. The training set is the part of the original dataset used to train the model and the test set is used to evaluate its performance. A proportion of 60:40, respectively, is used for this project. During the split, it is important to force both groups to be representative of the original data set, with the ratio of churners and non-churners remaining the same.

As mentioned in the Literature Review chapter, there are a large number of algorithms that can be used for machine learning studies. Moreover, there is not a study that identifies the best algorithms for churn prediction problems. Thus, following other studies mentioned in chapter 2, the Random Forest, Logistic Regression, Decision Tree, SVM, Gradient Boosting, and ANN were used in this project.

These techniques are controlled by hyperparameters that need to be adjusted for better performance achievements Feurer et al. (2019). Hyperparameter tuning is performed with the help of the Scikit-learn Python library functions, as follows: Initially, a randomized search is used to have a first impression of the range of the most favorable hyperparameters, and then Grid search, to better tune on the neighborhood of the previous discovered values. Grid search requires a list of values for the hyperparameters under the tuning process and tries all combinations possible within those lists, saving the best combination from those. Each combination was evaluated using the area under the ROC curve (AUC).

The algorithms tested and the respective hyperparameters tuned during this study are presented in table 3.11.

Table 3.11: Hyperparameters optimized.

Algorithm	Hyperparameters
Random Forest	max_features n_estimators
Logistic Regression	C
Decision Tree	max_depth max_features min_samples_split
SVM	C gamma
Gradient Boosting	max_depth max_features min_samples_split
ANN	epochs

The already mentioned grid search technique is combined with the nested cross-validation to avoid using the same data in the tuning and evaluation process, which can result in overfitting (Cawley and Talbot, 2010). This combination aggregates two loops, with each one using the cross-validation method. The k-fold cross-validation splits the training set into k equal folds, holds one for validation, and uses the others as training data sets. The performance is measured by averaging the scores from each iteration of the loop.

The inner loop focuses on hyperparameter tuning, and the grid search method is integrated into this section, while the outer loop focuses on the optimal parameters chosen by the inner one. For the current study, a k value of 4 is used.

3.5 Evaluation

Based on the information presented in section 2.2.2.4, multiple metrics are used to evaluate the performance of machine learning algorithms.

Confusion matrices are used to summarize the performance of the predictive models tested in chapter 4.

		Actual	
		Positive	**Negative**
Predicted by model	**Positive**	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Figure 3.9: Confusion matrix layout. (Arisholm et al., 2010)

The four categories have different meanings for different contexts. For churn problems they should be understood as:

- **True Positive (TP) -** churners properly classified
- **False Positive (FP) -** non-churners incorrectly classified as churners
- **False Negative (FN) -**churners incorrectly classified as non-churners
- **True Negatives (TN) -** non-churners properly classified

For the churn context, the metrics used are: accuracy, precision, recall, F1-score, ROC AUC. The valuable insights extracted from each of these measures are detailed in section 2.2.2.3.

3.6 Deployment phase

CRISP-DM last stage is the Deployment stage, where the discoveries from the studies are actually implemented and the plan for this implementation is developed. The Customer Success Team will

use the output of this project to define customers likely to churn in the following yearly quarter. The study provides a proof of concept of what techniques and methodologies should the team follow when retraining the model with more accurate data. Guidelines for what and how those variables should be collected for proper integration are also provided at the end of the study. A script is arranged to properly treat the data, following the recommendations explained in section 3.3.

Chapter 4

Experiments and results

During this chapter, the results obtained in the modeling stage of the CRISP-DM model are summarized. Four experiments were performed, following the procedure described in section 3.4, with differences between them so that a conclusion about their effects can be established.

The first experiment will work as a benchmark for comparisons with the other three since it is the simplest one, where only hyperparameters are tuned and no feature selection or data sampling techniques are implemented. The second experiment applies the data sampling technique and allows for comparisons and confirmation of the necessities of these methods for more comprehensive results. The third experiment, adds the features selection process, based on the importance's calculated with random forest technique. The last and fourth one, the aggregated measures, explained in section 3.2.2, are manually removed from the dataset and the conditions from the second experiment are replicated. Hyperparameter tuning is a common procedure for all experiments and the best combinations found are exhibited for all experiments. For the ANN model, no hyperparameter optimization was performed and the number of runs considered was constant.

4.1 Experiment 1

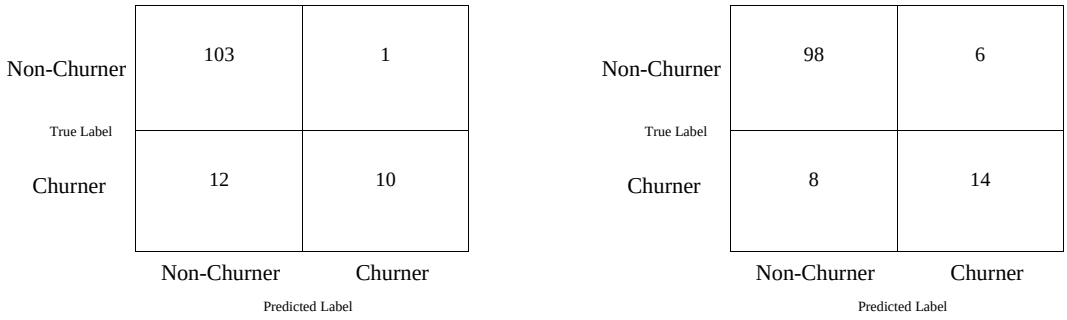
The first experiment is developed without the use of data sampling or feature selection techniques. A hyperparameter tuning study is implemented and the best combination of parameters for each technique is shown in table 4.1. Confusion matrices are calculated for each algorithm and the respective evaluation metrics are summarized in table 4.2.

Table 4.1: Best set of hyperparameters found - Experiment 1.

Algorithm	Hyperparameters	Best parameters set discovered
Random Forest	max_features	20
	n_estimators	1400
Logistic Regression	C	1
Decision Tree	max_depth	41
	max_features	8
	min_samples_split	9
SVM	C	979
	gamma	0.0001
Gradient Boosting	max_depth	98
	max_features	8
	min_samples_split	20
ANN	epochs	50

Table 4.2: Evaluation Metrics - Experiment 1.

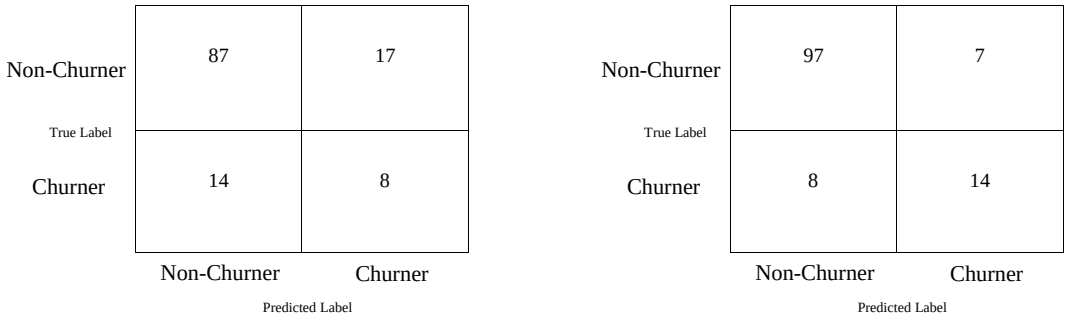
Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0,90	0,91	0,45	0,61	0,72
Logistic Regression	0,89	0,70	0,64	0,67	0,79
Decision Tree	0,75	0,32	0,36	0,34	0,60
SVM	0,88	0,66	0,64	0,65	0,78
Gradient Boosting	0,93	0,93	0,63	0,76	0,81
ANN	0,87	0,63	0,63	0,63	0,78



(a) Random Forest

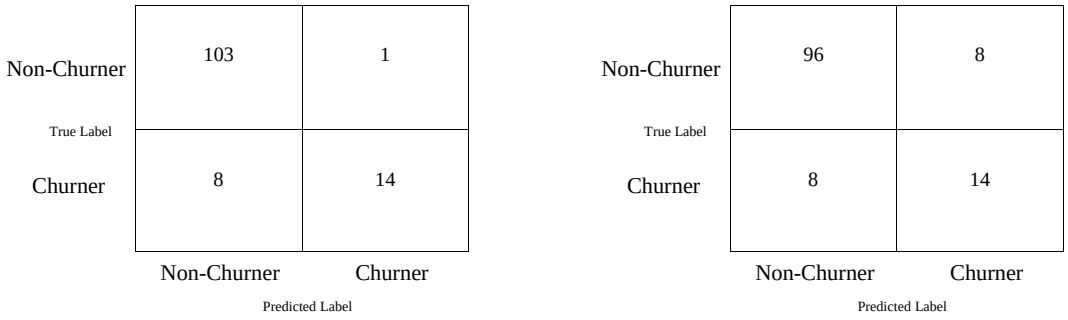
(b) Logistic Regression

Figure 4.1: Confusion Matrix Experiment 1- Random Forest and Logistic Regression.



(a) Decision Tree (b) SVM

Figure 4.2: Confusion Matrix Experiment 1- Decision Tree and SVM.



(a) Gradient Boosting (b) ANN

Figure 4.3: Confusion Matrix Experiment 1- Gradient Boosting and ANN.

Analysis

Analyzing table 4.1, it is not clear that a certain algorithm over-performed the others. When looking at the metric used as evaluation criteria during the modeling process, i.e. AUC, the best performer is the Gradient Boosting with a value of 81%, while the others show similar AUCs excepting the Decision Tree technique that reached only 60%, which is not positive. Considering a miss classification of a true churner is the most costly to the company, the metric recall has significant relevance in the Churn context. With this in mind, it can be seen in the table 4.2 that apart from the Decision Tree, the algorithms shared similar recall values, around 64%. Although precision and accuracy should be used as decision criteria due to its limitations described in section 2.2.2.4, the Gradient Boosting performed slightly better as well, which places a good indication on this algorithm for the next experiments.

The next experiment will allow testing the impact of data sampling techniques in an unbalanced dataset like the one available for the project.

4.2 Experiment 2

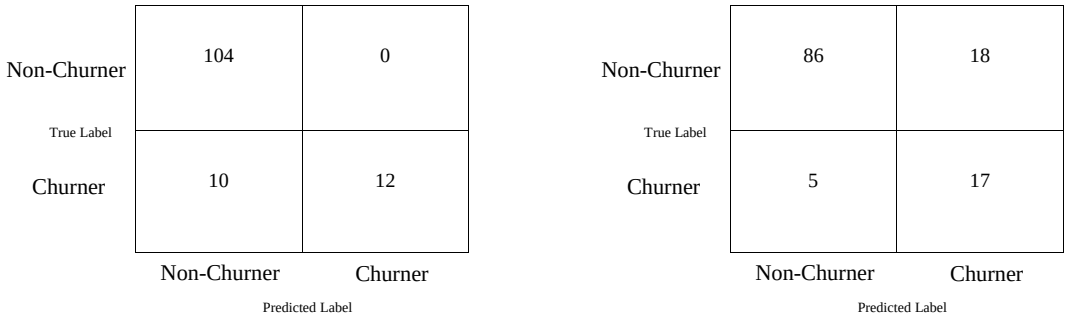
Experiment 2 contemplates the SMOTE sampling technique, described in section 3.3.5. The results of the models are summarized and presented with the same approach as in Experiment 1.

Table 4.3: Best set of hyperparameters found - Experiment 2.

Algorithm	Hyperparameters	Best parameters set discovered
Random Forest	max_features	20
	n_estimators	356
Logistic Regression	C	4
Decision Tree	max_depth	47
	max_features	10
	min_samples_split	9
SVM	C	109
	gamma	0,1
Gradient Boosting	max_depth	66
	max_features	4
	min_samples_split	10
ANN	epochs	50

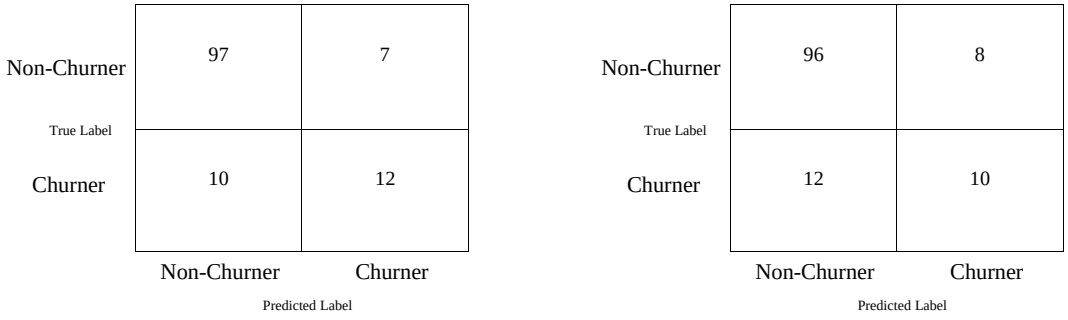
Table 4.4: Evaluation Metrics - Experiment 2.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0,92	1,00	0,54	0,70	0,77
Logistic Regression	0,82	0,49	0,77	0,60	0,80
Decision Tree	0,86	0,63	0,54	0,58	0,74
SVM	0,84	0,55	0,45	0,50	0,84
Gradient Boosting	0,87	0,65	0,59	0,61	0,76
ANN	0,82	0,50	0,81	0,62	0,82



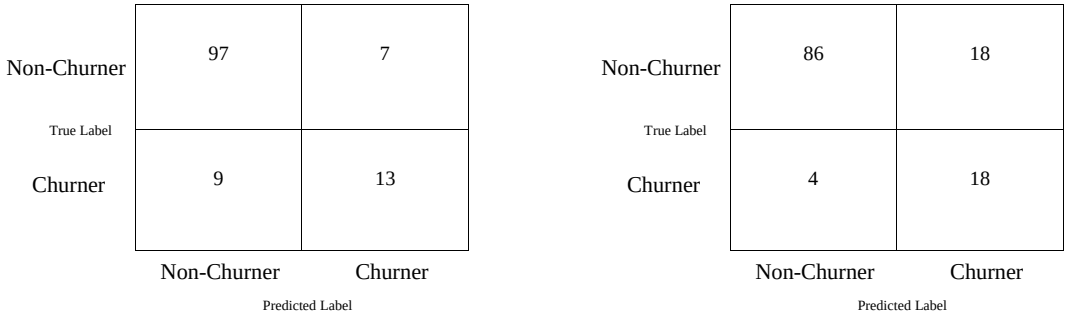
(a) Random Forest (b) Logistic Regression

Figure 4.4: Confusion Matrix Experiment 2- Random Forest and Logistic Regression.



(a) Decision Tree (b) SVM

Figure 4.5: Confusion Matrix Experiment 2- Decision Tree and SVM.



(a) Gradient Boosting (b) ANN

Figure 4.6: Confusion Matrix Experiment 2- Gradient Boosting and ANN.

Analysis

The data sampling, as expected, resulted in changes in the overall performance of the algorithms. Since now the dataset is balanced, accuracy becomes more significant for analysis purposes. It is relevant to point out that Random Forest achieved a very good 92% accuracy score, followed by the Gradient Boosting with 82%. However, both are more prepared to properly identity non-churners instead of churners since the recall values of both do not outperform the rest of the algorithms. Random Forest was capable of identifying all the non-churners, as shown in figure 4.4 a), but 10 real churners were not identified, which is a significant number, considering the total number of churners in the test sample. The best performers on the identification of churners in experiment 2 were the ANN and the Logistic Regression, with the first obtaining a recall value of 81%, the highest between the two.

Considering the miss-classification of the true churner is more costly than the other classification errors, the more promising algorithms in this experiment are the ANN and the Logistic Regression, but if one had to be chosen the ANN would move forward since it outperformed on the recall metric an all the other metrics.

Gradient Boosting, which showed positive results experiment 1, did not outperform the other algorithms in experiment 2.

4.3 Experiment 3

Experiment 3 includes another technique that aims to improve the feature selection process. Using the importance of each variable for the random forest algorithm in the last experiment, a group of the most important half of the existing variables available initially is selected and used for this experiment, with the other half being ignored. The selected variables had importance's higher than 0.7% and are listed in table 4.5.

Table 4.5: Random Forest importance's for selected features - Experiment 3.

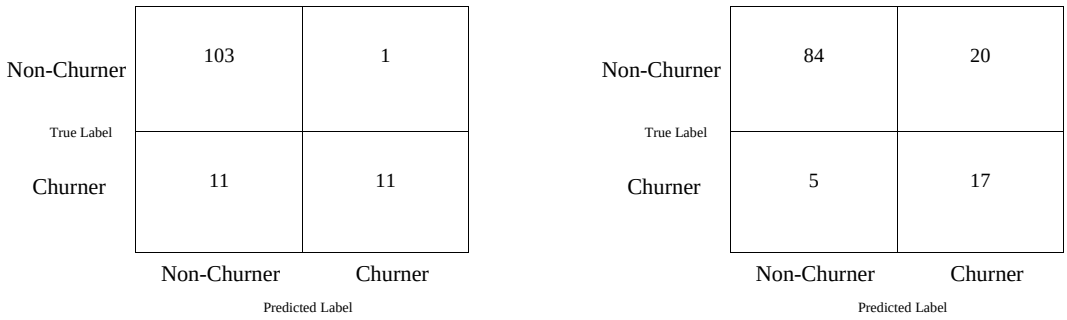
Variable Id	Variable Name	Importance
2	ChurnScore Relationship Risk	0.20003
3	ChurnScore Usage Risk	0.18709
5	Last Platform Job Date(Day)	0.08516
1	ChurnScore Engagement Risk	0.07733
37	Yes next BR	0.07501
10	Tenure (Days)	0.05749
11	Total Won Amount	0.04847
0	ChurnScore (Default)	0.03742
24	Daily	0.03305
13	Valid Computing Quota Amount Used	0.03039
12	Valid Computing Quota Amount Assigned	0.02019
8	Perceived Satisfaction	0.01554
7	Number of Support Cases	0.01457
9	Sentiment / Mood	0.01280
6	License Count	0.01277
39	long peridod	0.01081
16	Medium Value	0.00939
21	United Kingdom	0.00938
4	Depth of Relationship	0.00899
30	No intention	0.00710

Table 4.6: Best set of hyperparameters found - Experiment 3.

Algorithm	Hyperparameters	Best parameters set discovered
Random Forest	max_features	20
	n_estimators	570
Logistic Regression	C	1
Decision Tree	max_depth	29
	max_features	11
	min_samples_split	9
SVM	C	100
	gamma	0,1
Gradient Boosting	max_depth	46
	max_features	4
	min_samples_split	5
ANN	epochs	50

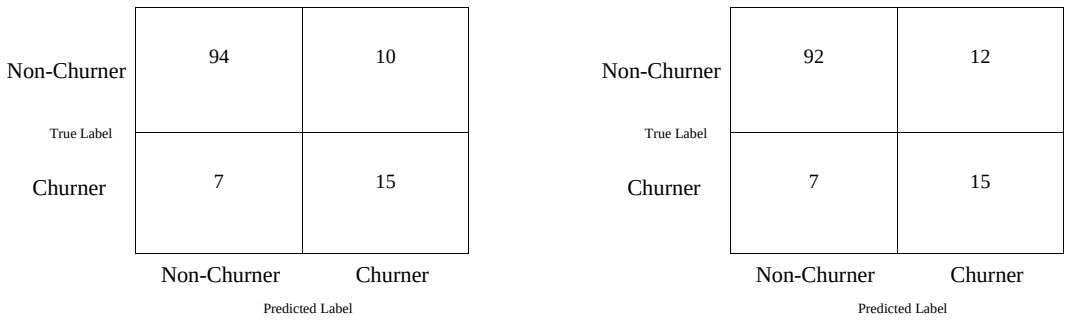
Table 4.7: Evaluation Metrics - Experiment 3.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0,90	0,92	0,50	0,65	0,75
Logistic Regression	0,80	0,46	0,77	0,58	0,79
Decision Tree	0,86	0,60	0,68	0,64	0,79
SVM	0,85	0,55	0,68	0,61	0,85
Gradient Boosting	0,89	0,75	0,54	0,63	0,75
ANN	0,80	0,44	0,54	0,49	0,70



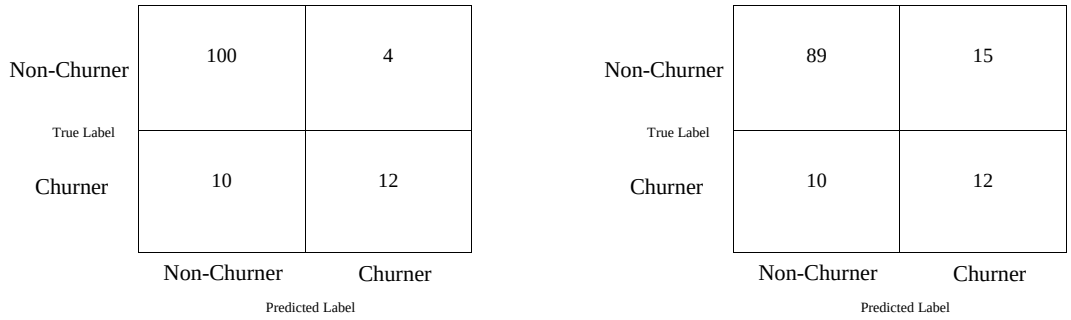
(a) Random Forest (b) Logistic Regression

Figure 4.7: Confusion Matrix Experiment 3- Random Forest and Logistic Regression.



(a) Decision Tree (b) SVM

Figure 4.8: Confusion Matrix Experiment 3 - Decision Tree and SVM.



(a) Gradient Boosting (b) ANN

Figure 4.9: Confusion Matrix Experiment 3 - Gradient Boosting and ANN.

Analysis

Feature selection, which reduces the number of variables included in the model to half, had different impacts on the algorithms.

Again there is not a clear best choice within the tested models. Random Forest presented the best accuracy score with a high capability of identifying non-churners. In fact, only one was incorrectly predicted, as shown in the confusion matrix of figure 4.7 a), but with a mediocre capability on the identification of true churners, with only half of the existing ones properly classified. The opposite happens for the Logistic regression, where the classification of churners is the best within the tested models, with a recall score of 77%.

4.4 Experiment 4

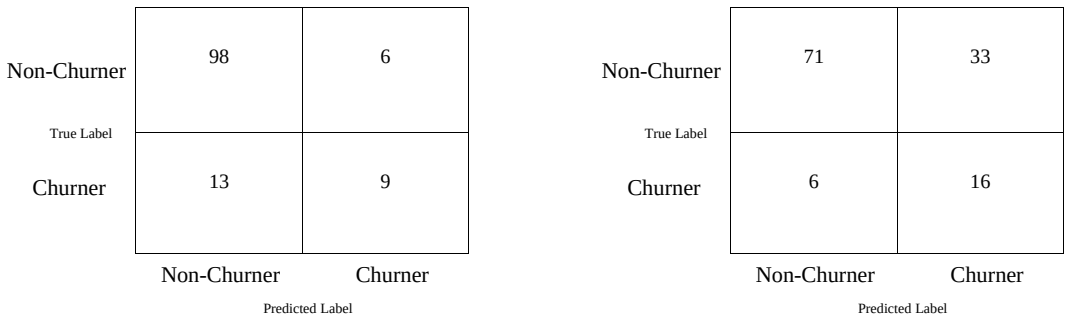
Experiment 4, uses as starting point the experiment 2 dataset, and removes the aggregated measures ChurnScore (Default), ChurnScore Engagement Risk, ChurnScore Relationship Risk and ChurnScore Usage Risk, to infer if these descriptive measures used daily in the CS team have a positive impact on the predictive process.

Table 4.8: Best set of hyperparameters found - Experiment 4.

Algorithm	Hyperparameters	Best parameters set discovered
Random Forest	max_features	20
	n_estimators	519
Logistic Regression	C	97
Decision Tree	max_depth	10
	max_features	5
	min_samples_split	8
SVM	C	100
	gamma	0.1
Gradient Boosting	max_depth	46
	max_features	4
	min_samples_split	5
ANN	epochs	50

Table 4.9: Evaluation Metrics - Experiment 4.

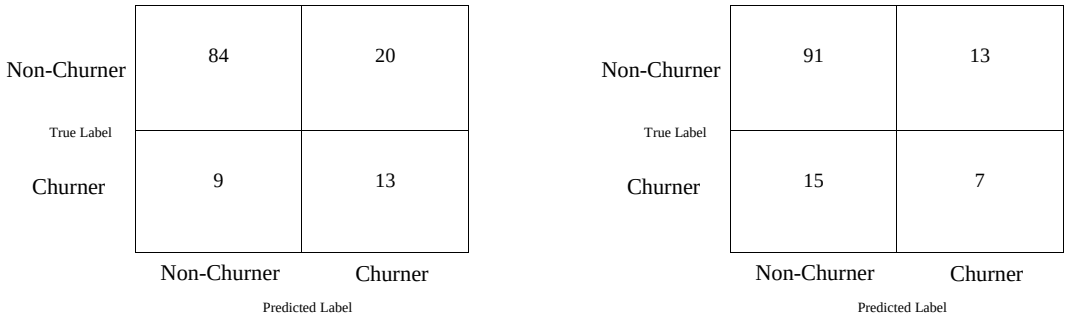
Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0,85	0,60	0,41	0,49	0,68
Logistic Regression	0,69	0,33	0,73	0,45	0,69
Decision Tree	0,77	0,39	0,59	0,47	0,70
SVM	0,78	0,35	0,32	0,33	0,60
Gradient Boosting	0,89	0,75	0,54	0,63	0,75
ANN	0,86	0,58	0,82	0,68	0,85



(a) Random Forest

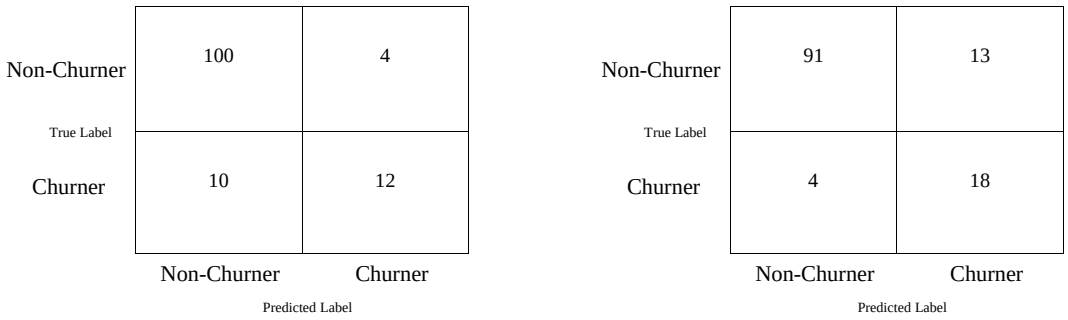
(b) Logistic Regression

Figure 4.10: Confusion Matrix Experiment 4 - Random Forest and Logistic Regression.



(a) Decision Tree (b) SVM

Figure 4.11: Confusion Matrix Experiment 4 - Decision Tree and SVM.



(a) Gradient Boosting (b) ANN

Figure 4.12: Confusion Matrix Experiment 4- Gradient Boosting and ANN.

Analysis

Removing the aggregated Scores from the dataset, similarly to the feature selection technique, provided different effects to the algorithms.

Again there is not a clear best choice within the tested models. Gradient Boosting presented the best accuracy score with a low capability of identifying non-churners which might invalidate the use of this technique. The next best technique when looking at accuracy is the ANN, also being the best performer in identifying churners. Again this comes with less assertiveness in non-churners identification, where Gradient Boosting and Random Forest are the best ones.

4.5 Results Summary

The evaluation phase, explained in section 3.5 is a step of the CRISP-DM method. In the previous section, the tested models were evaluated and compared with statistical measures. It is required

to add another level of judgment, not focused on machine learning, but one which integrates the business perspective where these models will be integrated later on, while in the deployment phase.

As explained in section 1.3, the end goal of the project is helping the SimScale Customer Success Team understand which customers should be the focus of retention campaigns due to their likelihood to churn. With a practical view, a predictive model capable of classifying churners, avoiding miss classifying non-churners as churners, and churners as non-churners, is what is desired. These miss classifications have different consequences and should be analyzed accordingly. This explains why during the analysis of the results, there has been a focus on recall, a metric that captures this relation.

At the moment to take decisions based on the outcome of the model, customers wrongly targeted as churners, will be approached and enter on retention campaigns, which will result in the loss of working time and on a ineffective use of resources. On the other hand, customers classified as non-churners that are actually at risk to churn will not be the target, which is the real threat and ultimately what the project is trying to avoid.

A first conclusion that can be developed based on the comparison's between experiments 1 and 2, is that the capability of predicting churners improved in most of the algorithms, which was expected when providing to the algorithm training process more samples of churners it could analyze and collect information from. With the experiment 2 conditions, ANN is the technique with the lowest number of true churners being classified as non-churners, which is why it could be said that this is the best performer for the project proposes. However, it is relevant to mention the lack of capability for identifying non-churners when compared with the Random Forest results. The low comprehensibility the ANN technique offers can be negative for future studies on identifying critical variables related to churners.

The feature selection strategies used in experiments 3 and 4 have not provided a clear tendency of improvement for the algorithms tested. However, in experiment 4, the ANN technique resulted in a better capability of non-churners identification, while remaining with high accuracy for churners.

In summary, the Artificial Neural Network model is considered the most efficient if aggregated Scores are decided to be ignored. Considering ANN does not provide comprehensibility capabilities, the inclusion of aggregated scores does not pay off in the eventuality of trying to explain churn with them. Thus it is suggested to not include them when on the deployment stage.

Chapter 5

Conclusions

The project's end goal was to develop a proof of concept of a model capable of predicting churn for SimScale customers. The followed process selected to organize and structure the project work, i.e. CRISP-DM, revealed to be a great choice for this project since the sequence of stages gave significant insights for the following steps of the project development. The deployment phase is still under development for the reasons explained in the future work section.

A dense analysis of the available data was developed so that the decision of including a variable in the dataset had a clear justification. Experimental attempts to identify the best models and techniques to be applied to the data set were performed and evaluated accordingly to success measures. The evaluation metrics were carefully selected and adjusted to the project context, allowing to decide whether the models generated met or not the success criteria. The Analytical Neural Network technique has a better performance when balancing its performance on different levels. Random Forest and Logistic Regression that performed well for non-churners identification are recommended to be tested in future work since their outcomes can change when more data characterizing churners is available.

Even though the project was helpful to get an understanding of the performance of different techniques under different conditions, some challenges and other improvements were identified. The difficulties are detailed in section 5.1 and the proposals for future works to complement this study are detailed in section 5.2.

5.1 Challenges

Once the problem and goals were defined, there was a need to define the project plan. It needed to contemplate the techniques to use and test for the exploratory analysis, feature selection, and even which classifying techniques should be studied. Since machine learning and more in particular prediction problems had already been largely developed and studied, there was a significant amount of strategies and alternatives that could be used, and this resulted in a large pool of options that needed to be filtered so that an organized study could be developed with the resources available.

During the initial stage, when defining the data set to analyze, some variables required treatment, and others needed to be simply ignored, since they are stored daily and each day some of the variables are over-written, not allowing to obtain time-dependent information. This is a significant limitation since the observation period is largely affected. The above-mentioned problem resulted in consequent delays of the modeling stage due to new problems being discovered, and more insights about the variables, being obtained from the CS team.

Another limitation that complement the previously mentioned ones, was the lack of experience planning a project in such area of study. Miss expectations on necessary planning and computing time limited the number of techniques tested.

5.2 Future Work

When developing the literature review section 2.2.2.3, on the topics explored during this project, it was clear that a possible path for further studies is exploring the benefits of the comprehensibility of the models for the churn problem. Understanding why customers cancel subscriptions or decide to not renew seems an interesting and profitable knowledge for companies to have, which is why comprehensibility could be valued even if accuracy is reduced. This relation, of course, is also a potential target of analysis. The best performer technique in this project is considered to be the ANN, which due to its nature is not interpretable. Techniques like Active Learning-Based Approach and AntMiner+, are existing options that enhance the comprehensibility and could be tested (Kamalakannan and Mayilvahanan, 2018).

Hyperparameter tuning is performed for the majority of the algorithms tested, however the ANN parameters due to its nature, should be studied with the aim of achieving a better outcome for the predictive model. Different architectures could also be explored in future studies with the available data.

Regarding feature selection, some techniques explored in the literature review, section 2.2.2.2, such as Forward selection or Backward elimination, could be tested. Both were considered for the current study but due to the computational time needed, were not developed. However, they could be interesting and add predictive value.

One proposal that is not expected to improve directly the predictive model, but is more focused on the aggregated scores, explained in section 3.2.2, is the differentiation of the calculations of these scores according to their potential category. While in the data understanding stage it became clear, in internal debating with the Customer Success Team, that the potential was crucial to determine the expectations of what defines success in terms of usage, relationship, and engagement for each client. Currently, the weights do not have the type of client in consideration, as shown in table 3.5. To better capture this effect, an Analytical Hierarchy Process was developed to determine different weights for each churn score metric for each account Potential. The results are summarized in table 5.1 and were implemented in parallel to the existing ones. A future discussion about its impact on the day-to-day usage of the CS team should be performed to determine if the

segmentation work performs as expected and if these weights should replace the current ones and so be integrated into the predictive model training as well.

Account Potential	Score	Current Weight	Proposed Weight
High Value	Usage Score	50,00%	35,91%
	Relationship Score	25,00%	56,44%
	Engagement Score	25,00%	7,65%
Medium Value	Usage Score	50,00%	36,61%
	Relationship Score	25,00%	53,21%
	Engagement Score	25,00%	10,18%
Low Value	Usage Score	50,00%	68,77%
	Relationship Score	25,00%	23,44%
	Engagement Score	25,00%	7,78%

Figure 5.1: ChurnScore Weight calculation resultant of a AHP process.

As mentioned in section 3.1, there are limitations on the availability of data that allows for observation and target periods to be fixed. The last proposal for future work is the preparation of dataset storage that collects data, starting on the current quarter of the year until the second next quarter. Using these data, further studies can be developed with more representative data. Such was not feasible during this study due to time limitations.

Bibliography

Agatonovic-Kustrin, S. and Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, 22(5):717–727.

Ahmad, A. and Dey, L. (2005). A feature selection technique for classificatory analysis. *Pattern Recognition Letters*, 26(1):43–56.

Ahn, J., Hwang, J., Kim, D., Choi, H., and Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. *IEEE Access*, 8:220816–220839.

Akrong Hesse, C. and Ofosu, J. (2018). *STATISTICAL METHODS FOR THE SOCIAL SCIENCES*. Akrong Publications Limited.

Arisholm, E., Briand, L. C., and Johannessen, E. B. (2010). A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. *Journal of Systems and Software*, 83(1):2–17.

Azevedo, A. and Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview. In *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008*, pages 182–185.

Bandyopadhyay, N. and Jadhav, A. (2021). Churn Prediction of Employees Using Machine Learning Techniques. *Tehnički glasnik*, 15(1):51–59.

Beniwal, S., Jambheshwar, G., and Arora, J. (2012). Classification and feature selection techniques in data mining Discovering Overlapping Community Structure in Networks through Co-clustering View project Classification and Feature Selection Techniques in Data Mining. *International Journal of Engineering Research and Technology*, 1(6).

Buckinx, W. and den Poel, D. V. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, 164(1):252–268.

Cardinal, R. and Aitken, M. (2005). *ANOVA for the Behavioral Sciences Researcher*. Psychology Press.

Cawley, G. C. and Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Technical report, Journal of Machine Learning Research.

Chiang, D. A., Wang, Y. F., Lee, S. L., and Lin, C. J. (2003). Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications*, 25(3):293–302.

Crewson, P. (2016). *Applied Statistics*. Online ebook.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the acm*.

Farquad, M. A., Ravi, V., and Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing Journal*, 19:31–40.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *AAAI Press / The MIT Press*.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing.

Ferraris, V. A. (2019). Commentary: Should we rely on receiver operating characteristic curves? From submarines to medical tests, the answer is a definite maybe!

Ferreira, I. (2019). Churn Prediction in Digital Marketing. *Faculty of Engineering of University of Porto*.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2019). *Automated Machine Learning Methods, Systems, Challenges*. Springer US.

Ge, Y., He, S., Xiong, J., and Brown, D. E. (2017). Customer churn analysis for a software-as-a-service company. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*, pages 106–111.

Glady, N., Baesens, B., and Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 197(1):402–411.

Hadden, J., Tiwari, A., Roy, R., and Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers and Operations Research*, 34(10):2902–2917.

Han, J., Kamber, M., and Pei, J. (2011). Third Edition : Data Mining Concepts and Techniques. *Journal of Chemical Information and Modeling*, 53(9):1689–1699.

Jahromi, A. T. (2009). *Predicting Customer Churn in Telecommunications Service Providers*. PhD thesis, Luleá University of Technology.

Judy Strauss and Raymond Frost (2014). *E-Marketing*. Pearson Education Limited.

Kamalakannan, T. and Mayilvahanan, P. (2018). Efficient Customer Churn Prediction Model Using Support Vector Machine with Particle Swarm Optimization. *International Journal of Pure and Applied Mathematics*, 119(10):247–254.

Khodabandehlou, S. and Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, 19(1/2):65–93.

Kim, S., Choi, D., Lee, E., and Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. *Plos One*.

Komorowski, M., Marshall, D. C., Salciccioli, J. D., and Crutain, Y. (2016). *Exploratory Data Analysis*, pages 185–203. Springer International Publishing, Cham.

Kracklauer, A., Mills, D., and Seifert, D. (2004). Customer management as the origin of collaborative customer relationship management. *Collaborative Customer Relationship Management - Taking CRM to the Next Level*, pages 3–6.

Lazarov, V., München, T. U., Capota, M., and München, T. U. (2007). Churn prediction. *Business Analytics Course. TUM Computer Science*.

Liu, H., Zhou, M., and Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3):703–715.

Miguéis, V. L., Van den Poel, D., Camanho, A. S., and Falcão e Cunha, J. (2012). Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification*, 6(4):337–353.

Mohan, A. and Deshmukh, A. K. (2013). Conceptualization and Development of a Supply Chain-Customer Relationship Management (SC2R-M) Synergy Mode. *Journal of Supply Chain Management Systems*, 2(3):9–25.

Morgan, N. A., Vorhies, D. W., and Mason, C. H. (2009). Research notes and commentaries market orientation: Marketing capabilities, and firm performance. *Strategic Management Journal*, 30(8):909–920.

Myatt, G. J. (2007). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. John Wiley & Sons, Inc., New Jersey.

Natarajan, K., Li, J., and Koronios, A. (2010). Data mining techniques for data cleaning. In *Engineering Asset Lifecycle Management - Proceedings of the 4th World Congress on Engineering Asset Management, WCEAM 2009*, pages 796–804. Springer, London.

Ngai, E. W., Xiu, L., and Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2 PART 2):2592–2602.

Nisbet, R., Miner, G., and Yale, K. (2017). *Handbook of statistical analysis and data mining applications*. Elsevier Inc.

Oliveira, V. L. M. (2012). *Management in Retailing Supported by Data Mining Techniques*. PhD thesis, Faculty of Engineering of University of Porto.

Osborne, J. W. (2016). *Regression & Linear Modeling Best Practices and Modern Methods*. Clemson University.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4):211–218.

Polikar, R. (2012). Ensemble Learning. In *Ensemble Machine Learning*, pages 1–34. Springer US.

Rothmeier, K., Pflanzl, N., Hullmann, J. A., and Preuss, M. (2021). Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game. *IEEE Transactions on Games*, 13(1):78–88.

Ryals, L. and Payne, A. (2001). Customer relationship management in financial services: towards information-enabled relationship marketing. *Journal of Strategic Marketing*, 9:3–27.

Saeys, Y., Aki Inza, I., and Larrañ Aga, P. (2007). A review of feature selection techniques in bioinformatics. *BIOINFORMATICS REVIEW*, 23:2507–2517.

Sarchana and Kelangovan (2014). Survey of Classification Techniques in Data Mining Want more papers like this? Survey of Classification Techniques in Data Mining. *International Journal of Computer Science and Mobile Applications*, 2:65–71.

Schuh, G., Prote, J. P., Molitor, M., Sauermann, F., and Schmitz, S. (2019). Databased learning of influencing factors in order specific transition times. In *Procedia Manufacturing*, volume 31, pages 356–362. Elsevier B.V.

Shafique, U. and Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12(1):217–222.

Shardlow, M. (2016). An Analysis of Feature Selection Techniques. Technical report, The University of Manchester.

Sheikh, A., Ghanbarpour, T., and Gholamiangonabadi, D. (2019). A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *Journal of Business-to-Business Marketing*, 26(2):197–207.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, us ed edition.

Valtola, V. (2019). *A Study of Customer Retention*. PhD thesis, Arcada.

Van den Poel, D. and Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 157(1):196–217.

Veloso, J. (2013). *Um Modelo para Previsão de Churn na Área do Retalho*. PhD thesis, Universidade do Minho.

Verbeke, W., Martens, D., Mues, C., and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3):2354–2364.

Wei, C. P. and Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications*, 23(2):103–112.

Weiss, G. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations*, 6:7–19.

Appendix A

Appendix

A.1 Initial variables

Table A.1: Initial dataset features description.

Variable	Description
Billing Country	Country of the client
ChurnScore (Default)	Indicator of likelihood to churn
ChurnScore Engagement Risk	Indicator of engagement
ChurnScore Relationship Risk	Indicator of relationship
ChurnScore Usage Risk	Indicator of usage
Depth of Relationship	Quality of the relationship between client and support team
Last Business Review Date	Date from the last business review meeting
Next Business Review Date	Date from the last business review meeting scheduled
Last Platform Job Date(Day)	Elapsed days since last job was executed in the platform
License Count	Number of licenses registered per account
Number of Support Cases	Number of support tickets requested per account
Perceived Satisfaction	Account owner felling of clients satisfaction on the platform
Potential	Account potential for commercial expansion
Sentiment / Mood	Account owner felling of clients satisfaction with the service
Success Story / Case Study Status	Stage of Success Story proposal
Tenure (Days)	Length of the relationship
Total Won Amount	Financial amount won from the client
Usage Frequency	Frequency of the client usage for the platform
Valid Computing Quota Amount Assigned	Valid amount of simulation credits initially available for the client
Valid Computing Quota Amount Used	Valid amount of simulation credits used by the client

A.2 Correlation between categorical variables

Chi-Squared Analysis

Table A.2: Chi-Squared analysis results.

	Potential	Success Story / Case Study Status	Usage Frequency	Billing Country	Has Next Business Review Date	Has Last Business Review Date
Potential	-	Independent (fail to reject H0)	Dependent (reject H0)	Dependent (reject H0)	Dependent (reject H0)	Dependent (reject H0)
Success Story / Case Study Status	-	-	Independent (fail to reject H0)	Dependent (reject H0)	Independent (fail to reject H0)	Independent (fail to reject H0)
Usage Frequency	-	-	-	Independent (fail to reject H0)	Independent (fail to reject H0)	Independent (fail to reject H0)
Billing Country	-	-	-	-	Dependent (reject H0)	Independent (fail to reject H0)
Has Next Business Review Date	-	-	-	-	-	Dependent (reject H0)
Has Last Business Review Date	-	-	-	-	-	-

Cramer's V Analysis

Table A.3: Cramer's V Analysis Results.

	Potential	Success Story / Case Study Status	Usage Frequency	Billing Country	Has Next Business Review Date	Has Last Business Review Date
Potential	-	0	0.1564821932	0.3598943337	0.1560886199	0.1552587309
Success Story / Case Study Status	-	-	0.0774274772	0.09795549443	0	0.07286911637
Usage Frequency	-	-	-	0	0.0613616112	0.05834954381
Billing Country	-	-	-	-	0.1291341038	0.05004737657
Has Next Business Review Date	-	-	-	-	-	0.3829353371
Has Last Business Review Date	-	-	-	-	-	-

A.3 Correlation between categorical and numerical variables

Kruskal Analysis

Table A.4: Kruskal Analysis Results.

	ChurnScore (Default)	ChurnScore Engagement Risk	ChurnScore Relationship Risk	ChurnScore Usage Risk	Depth of Relationship	Last Platform Job Date(Day)	License Count	Number of Support Cases	Perceived Satisfaction	Sentiment / Mood	Tenure (Days)	Total Won Amount	Valid Computing Quota Amount Assigned	Valid Computing Quota Amount Used
Potential	79.69253027	44.97160722	53.6586901	53.84159317	27.19030561	19.83136113	27.46203757	25.76241752	4.949049245	17.59355741	7.120776649	106.6014832	27.4347339	37.26629285
Success Story / Case Study Status'	9.858412679	6.050179984	9.206041187	13.41105987	3.195686624	11.80376625	2.451108803	6.76882485	6.640976701	6.628857073	46.75738171	6.434728697	4.716015095	18.0977858
Usage Frequency	57.52457757	36.16481904	11.76390096	94.63903168	6.135338386	129.8298444	10.57749268	6.130287853	5.194023396	20.34049368	1.474333773	26.39905877	25.4785792	37.25818779
Billing Country	47.06650851	7.158622627	37.63781732	38.31579445	9.340251102	8.772525648	7.332815193	5.186798421	9.302812208	32.95443401	5.056620575	23.2683926	5.756490215	16.70766959
Has Next Business Review Date	11.98395639	29.73740518	0.937408212 4	6.665513084	1.624189616	7.642472158	1.800118195	1.520202115	0.477888470 3	1.174804169	10.07957251	4.571165444	0.225306137 2	0.281349603 1
Has Last Business Review Date	73.16009631	123.0418996	28.57839572	46.07157969	25.54022036	19.8036307	7.964499361	25.21128854	7.491785066	18.37064287	13.58286948	31.68974302	8.782782643	6.241431309

A.4 Correlation between numerical variables

Pearson Analysis

Table A.5: Pearson Analysis Results.

	ChurnScore (Default)	ChurnScore Engagement Risk	ChurnScore Relationship Risk	ChurnScore Usage Risk	Depth of Relationship	Last Platform Job Date(Day)	License Count	Number of Support Cases	Perceived Satisfaction	Sentiment / Mood	Tenure (Days)	Total Won Amount	Valid Computing Quota Amount Assigned	Valid Computing Quota Amount Used
ChurnScore (Default)	1.000000	0.562158	0.852428	0.852235	-0.281173	0.395293	-0.104773	-0.391242	-0.267789	-0.146487	-0.127419	-0.267415	-0.143716	-0.384283
ChurnScore Engagement Risk	-	1.000000	0.258724	0.368468	-0.188783	0.254003	-0.179125	-0.153899	-0.131915	-0.164429	0.237336	-0.223226	-0.005704	-0.118584
ChurnScore Relationship Risk	-	-	1.000000	0.556271	-0.315354	0.169690	-0.077893	-0.385931	-0.178227	-0.106238	-0.224142	-0.249526	-0.186095	-0.410543
ChurnScore Usage Risk	-	-	-	1.000000	-0.125223	0.543718	-0.045896	-0.320672	-0.297457	-0.092173	-0.161426	-0.162192	-0.089793	-0.282176
Depth of Relationship	-	-	-	-	1.000000	-0.177717	0.056930	0.189563	0.529996	0.643806	-0.114862	0.172388	0.144295	0.193976
Last Platform Job Date(Day)	-	-	-	-	-	1.000000	-0.055201	-0.178218	-0.227215	-0.163997	0.126202	-0.072323	-0.062474	-0.123709
License Count	-	-	-	-	-	-	1.000000	0.063615	-0.055242	0.058710	0.014355	0.662994	0.072301	0.034563
Number of Support Cases	-	-	-	-	-	-	-	1.000000	0.107699	0.120863	0.130360	0.170192	0.251570	0.345149
Perceived Satisfaction	-	-	-	-	-	-	-	-	1.000000	0.545297	-0.015487	-0.002312	0.035114	0.132004
Sentiment / Mood	-	-	-	-	-	-	-	-	-	1.000000	-0.152958	0.146987	0.073257	0.131995
Tenure (Days)	-	-	-	-	-	-	-	-	-	-	1.000000	0.024793	0.104912	0.062865
Total Won Amount	-	-	-	-	-	-	-	-	-	-	-	1.000000	0.232464	0.195545
Valid Computing Quota Amount Assigned	-	-	-	-	-	-	-	-	-	-	-	-	1.000000	0.398698
Valid Computing Quota Amount Used	-	-	-	-	-	-	-	-	-	-	-	-	-	1.000000