Discovering Customer Behavioural Patterns from Financial Transactions

Ayrton Senna Azzopardi

Supervised by Dr Joel Azzopardi

Department of Artificial Intelligence

Faculty of ICT

University of Malta

May, 2021

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in A.I..

WWW.UM.EDU.MT

The research work disclosed in this publication is funded by Entropay Services through sponsorship of the author to attend the M. Sc. AI.

First edition, Monday 31st May, 2021

Dedicated to all my loved ones.

Acknowledgements

I dedicate this dissertation to my family, for their unconditional love and continuous support. They are the ones who instilled in me sound values and principles, without which I would have never reached such point in my life.

I would also like to thank my supervisor, Dr. Joel Azzopardi for his guidance throughout the past years. The completion of this dissertation would not have been possible without his continuous feedback and contribution.

Abstract

In an era where there is a significant need of converting huge amounts of data into useful and valuable information, the task of data mining saw a huge increase in importance, popularity and applicability. Data mining refers to the process of examining large volumes of data so as to discover hidden patterns, relationships and other insights conveyed in the data. It allows companies to process their data, while aiming to produce new growth opportunities in order to outperform their competition. Nowadays, many companies and businesses are resorting to data mining techniques, to save costs and time, as well as to understand customers' needs and market conditions. Analysing such data is beneficial for any company, leading to better informed business decisions, higher profits and more contented clients.

In this dissertation, we present a data mining study that is applied on millions of transactional records that were collected for a number of years, by a leading virtual credit card company based in Malta. In this study, 2 machine learning techniques, namely Artificial Neural Networks (ANNs) and Gradient Boosting (GBM), are analysed to identify the best modelling framework that predicts the churning behaviour of this company's customers. Apart from helping the marketing department of this firm by providing a model that predicts customers that are willing to churn, in this study we contribute to literature by analysing and identifying the minimum amount of customer activity needed, to predict churn. In addition, we also analyse the "cold start" problem by performing a time-series experiment based on the hardly any data available at the beginning of the customer purchase history.

When evaluating our system, we have seen that the Gradient Boosting Model (GBM) predictive model is more suitable than an Artificial Neural Network (ANN) in our Customer Relationship Management (CRM) problem, since it is capable of identifying the majority (70%) of churners correctly whilst the constructed ANN is more capable of classifying the non-churners. Furthermore, we have shown that reducing the observation window size does not reflect in huge performance loss, thus giving the ability of leveraging prediction performance with the amount of data observed. Finally, we have seen that the demographic information present in our dataset is not effective to predict churn behaviour of customers.

Contents

List of Figures

List of Tables

List of Abbreviations

1

Introduction

With the advancements in web technologies, online shopping as well as online gambling have rapidly increased in popularity in the past years. In fact, worldwide retail e-commerce sales are showing logarithmic growth and forecasted to have an increase of around $670 billion in the year of 2020 [1], whilst the online gambling market is forecasted to grow by $48.6 billion between 2017 and 2024 [2]. The ability of making a financial transaction instantly and from across the world is one of the sole reasons that e-commerce and gambling websites grew in such a rapid pace. However, when making an online transaction, a user's financial data is being exposed on the internet. This has been a major concern to some users. In fact, they are resorting to methods that are able to protect their financial information.

A popular approach is making use of virtual credit cards. A virtual credit card is essentially a randomly generated number designed to keep one's traditional credit card information secure whilst shopping online. Initially, funds are deposited into an account linked with this virtual credit card number, and then the actual virtual credit card is used when performing transactions online. Since funds are deposited through a secure platform provided by the virtual credit card company, the actual bank account details of a user are not used or else observed in any way when making a transaction on the web. Moreover, the actual bank will not have any knowledge whatsoever how clients are spending their money. This is extremely beneficial to online gamers as certain banks are rejecting attempts of using credit cards at gambling websites. As for this reason, many customers and gamblers across the globe are making use of virtual credit cards as a way to protect their financial information.

[1] Sept 2020: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

[2] Sept 2020: https://www.statista.com/statistics/270728/market-volume-of-online-gaming-worldwide/

Therefore with the increase in popularity of e-commerce and gambling websites, millions of users are now registering with virtual credit card companies so as to keep their traditional bank details protected whilst online. Consequently, virtual credit card companies are dealing with huge amounts of incoming data in the form of new users, creation of credit cards, deposit of funds and mostly online transactions. In fact, these companies are highly concerned with 'Big Data' since their data is large in volume and is increasing exponentially. Such data needs to be analysed and processed, in order to extract meaningful knowledge about the different users and any trends that might be beneficial for the respective company.

Data mining refers to the process of examining large volumes of data so as to discover hidden patterns, relationships and other insights conveyed in the data (Bilal Zorić, 2016; Cil et al., 2018; Keramati et al., 2016). This task allows companies to process their data, while aiming to produce new growth opportunities in order to outperform their competition. Nowadays, many companies and businesses are resorting to data mining techniques, to save costs and time, as well as to understand customers' needs and market conditions. Analysing such data is beneficial for any company, leading to better informed business decisions, higher profits and more contented clients.

With regards to virtual credit card companies, the task of data mining can be performed to predict a client's lifetime value. This involves forecasting how long a client will keep using the company's service based on one's activity. This information is beneficial to identify which clients are the best, which needs to be retained and which are willing to switch to a competitor. Moreover, data mining results can also be used to enhance the marketing sector of a company. Knowing a customer's behaviour and preferences, is valuable when trying to send the right message to the right user at the right time.

1.1 | Motivation

Despite being highly required and beneficial, there are still challenges that need to be addressed when attempting to discover behavioural patterns from big financial data.

One of the main challenges that is encountered by researchers contributing to the financial domain, is the sensitive nature of such financial data and its implications on scientific research. According to Martens et al. (2016), studies on data processing and analysis for financial businesses, highly depend on close collaborations with the industry Cil et al. (2018); Safinejad et al. (2018). For this reason, any company data together with the information resulting from the research rarely get shared with the scientific community, due to its sensitive nature. Thus, applicability and comparison of methods proposed by other researchers is very limited and restricted. In light of such issue, evaluating any proposed methods is also challenging due to the lack of gold standard datasets in this research domain.

Furthermore, a challenge that researchers come up against when contributing to CRM, is the well-known "cold start" problem. The task of modelling customers' behavioural trends and preferences to predict their possible future actions, becomes rather challenging when companies attempt to manage recently registered customers. This is due to the fact that in the early stages, very little customer-company interaction would have been observed for a relatively new customer. For firms in the financial domain such as virtual credit card companies, this means that only a small amount of financial transactions are available when modelling new customers. In fact, Padilla and Ascarza (2020) state that "companies have difficulties leveraging existing data when they attempt to make inferences about customers at the beginning of their relationship".

1.2 | Proposed Solution

Having described how financial companies are having trouble retrieving any marketing value from their increasingly growing data, and how the task of solving such problem has its own challenges, we now propose our solution to this research problem. In this dissertation, we present a data mining study that is applied on millions of transactional records that were collected for a number of years, by a leading virtual credit card company based in Malta.

In this study, a number of machine learning techniques are analysed to identify the best modelling framework that predicts the churning behaviour of this company's customers. Apart from giving value to the marketing department of the contributing firm by providing a model that predicts whether a customer is churning or not, in this study we contribute to literature by analysing and identifying the minimum amount of customer activity needed, to predict churn. In addition, we also analyse the "cold start" problem by performing a time-series experiment based on the hardly any data available at the beginning of the customer purchase history.

1.3 | Aims and Objectives

The aim of this dissertation is to research machine learning techniques and develop a modelling framework that is capable of predicting whether a customer is churning based on one's virtual credit card transactions, which framework can then be utilised by the co-operating company to enhance its marketing strategies in attempting to retain customers. The aforementioned aim will be fulfilled by achieving the following research objectives:

- Extract a number of dynamic features from the provided raw financial transactions, so as to find the most effective feature set when predicting customer churn.
- Build a machine learning setup that is effective in modelling and predicting customer churn using the extracted features.
- Determine the minimum amount of customer activity needed, in order to effectively predict whether a customer is churning or not.
- Determine whether demographic features together with any initial financial transactions can be used to predict whether a relatively new customer will continue to use the company's services or else churns.

Despite the fact that there are quite a few researchers to have studied customer churn in the financial domain, the concept of predicting churn using dynamic features solely extracted from raw financial transactions, as presented in this dissertation, is still quite a novel approach.

Dynamic features are those for which their values change over time sequentially, for example the amounts spent in each day. On the other hand, static features simply contain a fixed value, such as the total amount of purchases or the total amount spent in purchases.

Traditional customer churn prediction solutions focus solely on domain specific and static features. Furthermore, their predictive models are generally based on features representing customer socio-demographic information, customer satisfaction reviews, service usage aggregation and product or account type ownership amongst others (Kaya et al., 2018).

Latest research works have shown that dynamic behavioural predictors tend to be more effective in representing customer behaviour for churn prediction (Hsu et al., 2019; Kaya et al., 2018; Leung and Chung, 2020). In fact, the first of our objectives is to extract meaningful dynamic features from a customer's virtual credit card usage history, and validate their usefulness with regards to customer churn prediction.

Furthermore, our second objective is to construct a machine learning setup that is able to model the churn behaviour of customers and predict churn activities. Once again, in Section 3.1.1 we review a number of works that employ different machine learning techniques in their studies while tackling the task of customer churn prediction in the finance sector.

However the generated framework with the best results in such studies will be specific to the dataset that is used in that particular experiment. It does not entail that the same framework or technique will obtain the best performance results when applied on a different dataset. For this reason, our second objective still adds value to the research as it evaluates and finds the best performing technique on the data provided in this dissertation.

Subsequently, our third objective involves the experimentation and determination of the optimal observation window size i.e. the minimum amount of customer activity needed to predict with good performance, whether a customer is churning or not. From all the similar systems found in literature, only Leung and Chung (2020) attempts to test different observation window sizes. However, this experiment was solely an evaluation of just 2 observation periods differing in length (4 months vs 6 months). We fill this research gap by performing a more extensive, time-series experiment in order to determine the minimum amount of customer spending history needed.

Finally, in our fourth objective we tackle the "cold start" problem within the task of churn prediction in the financial domain. This objective is mainly motivated by the fact that since financial companies or any other institutions, contain barely any observed data for newly registered customers, the latter are generally excluded from further analysis so as to not disrupt the study. In this dissertation, we fill this gap by applying another time-series experiment using the only data available at a specific time.

1.4 | Contributions

Through the objectives mentioned in Section 1.3, we will be contributing to the fields of Artificial Intelligence (AI) and Fintech since we further explore the novel approach of utilising dynamic features in customer churn modelling, and also because these features are extracted solely from raw financial transactions.

In addition, we will also be filling the research gap surrounding the identification of the appropriate observation window size. As previously mentioned, we will be performing and evaluating a time-series experiment to determine the minimum sufficient amount of customer purchase history required to predict customer churn behaviour.

1.5 | Document Structure

The rest of the dissertation is structured in the following chapters. In Chapter 2 we present a brief background on the concept of machine learning and also on the techniques utilised in this study in order to build the machine learning setup. Chapter 3 reviews the approaches taken by similar systems found in literature. In Chapter 4, we discuss and justify the design decisions taken and also explain the implementation details of this study. In Chapter 5, we evaluate our objectives set in Section 1.3, taking into consideration how other researches evaluated similar systems in literature. Finally, Chapter 6 concludes this dissertation outlining our contributions to literature, and how this study can benefit from future work.

2

Background

In this chapter, we present a brief background on the concept of machine learning and also on the techniques utilised in this dissertation so as to build our machine learning setup. Initially, we discuss the field of machine learning and outline its benefits and applicability in the current era. The different categories of machine learning techniques are also described hereunder. We then give an overview of two machine learning techniques utilised in this dissertation, namely gradient boosting and artificial neural networks.

2.1 | Machine Learning

Machine Learning is a field of AI that refers to the notion of having a computer program that learns from any given data and also adapts to new observations, without any human intervention. The concept of machine learning comprises of various algorithmic tools that without being explicitly programmed, are capable of automatically learning from data observations, gain knowledge from experience and enhance their learning behaviour when observing new data. (Badillo et al., 2020; Holzinger, 2019)

More briefly, machine learning algorithms focus on analysing and utilising the given data to learn a generalised concept by themselves. In fact, the process of learning begins with observing the given data or training examples and then infer patterns to be able to make predictions and better decisions on new observations in the future. In other words, machine learning algorithms can be seen as inducing general functions or concepts from specific training examples.

In an era where 'Big Data' is a major concern for numerous industries and companies, the field of machine learning saw a great increase in popularity as well as scientific research (Badillo et al., 2020). The fact that certain machine learning algorithms are able to process large amounts of data and infer generic concepts instantly, has offered data analysts new opportunities in how their data can be automatically examined and analysed without any human intervention. In fact, machine learning is being widely applied in health care, financial services, retail, automotive and transportation industries amongst others, as well as in certain government agencies.

Machine learning techniques are generally classified into three major categories, namely:

- **Supervised Learning** - In supervised learning, algorithms are trained on labelled data, meaning the provided data is associated with the true answer. Thus the task of such algorithms is to learn how to predict the answer for future observations based on the concept that has been inferred from the training set of observations. In other words, the purpose of supervised algorithms is to learn the mapping between the input variables and the output answer. Furthermore, supervised machine learning algorithms are mainly used to solve the following two problems:
  - **Classification** - For classification problems, supervised algorithms are used to predict a label. Thus, the output answer is a class.
  - **Regression** - For regression problems, supervised algorithms are used to predict a quantity. Thus, the output answer is numerical.
- **Unsupervised Learning** - On the contrary to supervised techniques, in unsupervised learning, data is unlabelled and the output result is yet unknown. Therefore, the task of unsupervised algorithms is to investigate, infer and identify any hidden patterns or structures in the provided dataset. Generally, such algorithms group or link unsorted information according to similarities, patterns and differences without any prior knowledge. In fact, unsupervised machine learning algorithms are generally used to perform the following tasks:
  - **Clustering** - Clustering refers to the process of grouping similar items together into a single group or cluster, such that items in the same cluster are more similar to each other than to those in other clusters.
  - **Association** - Association refers to the process of identifying links of interest between the different items in the data, where such associations are generally in the form of rules.
- **Reinforcement Learning** - In reinforcement learning, algorithms are used to maximise some form of cumulative reward. More precisely, algorithms perform different actions in search of the ideal behaviour towards solving a specific problem. In

order to find this ideal behaviour, each action taken returns a reward or error that is utilised by the algorithm as feedback. In comparison to other types of learning, reinforcement learning does not require labelled data as it focuses on exploring the goal result by capitalizing on current knowledge.

(Badillo et al., 2020; Holzinger, 2019)

2.2 | Gradient Boosting

Gradient boosting is a supervised machine learning technique, developed by Friedman (2002) and that can be utilised to solve both classification and regression problems. Boosting in machine learning terms, refers to a technique that is capable of tuning weak predictive models so as to become better predictors. This is mainly done by generating a single predictive model as an ensemble of numerous weak predictive models - these are generally decision trees. A weak predictive model is one that is slightly better than a random guesser. Therefore, a decision tree performing with an accuracy better than 50% is a fine example of a weak learner.

The gradient boosting technique can be viewed as a "stage-wise additive model" as it builds the ensemble model by adding a weak predictive model every time, leaving the existing weak learners in the model untouched (Friedman, 2001). At each step, this technique analyses the current ensemble and exploits the misclassification error of the current model to try to reduce it. For that reason, gradient boosting technique can be described as an optimization problem where the goal is to reduce the loss of the model, by adding further weak predictive models in a gradient descent like approach.

To summarise the above, gradient boosting algorithms are based on three elements, mainly:

- **Loss Function** - This refers to the function that needs to be optimised by the algorithm when adding weak learners to the ensemble. The gradient boosting framework allows for any differentiable loss function to be used, so it merely depends on the type of problem being solved. Squared error functions can be used for regression whilst logarithmic losses can be used for classification.

- **Weak Learner** - Weak learners refer to the individual predictive models that are added to the ensemble that make up the final model. As aforementioned, decision trees are used as weak learners in gradient boosting. This is mainly because of the fact that they are capable of creating splits in the data, facilitating the addition of their outcomes and correction of predictions.

- **Additive Model** - This refers to the way a decision tree is chosen and added to the ensemble. As discussed above, trees are added to minimize the misclassification error of the model. This is done by identifying the appropriate parameters of the tree that reduces the loss function. This approach is known as “functional gradient descent”.

(Friedman, 2001, 2002)

Gradient boosting is a greedy machine learning technique. For that reason, it requires a number of regularisation methods to restrict the fitting process and prevent “over-fitting” the dataset. The following are 4 types of improvements that are typically used to regularise the model:

- **Shrinkage** - This refers to the learning rate of the algorithm, which weights the results of each decision tree in the ensemble so as to slow down the learning process. Slowing down the learning process, means that more decision trees will be required in the model, thus leaving space for future trees to improve the performance of the model.
- **Random Sampling** - This refers to randomly selecting a subset of the data whenever the algorithm adds a new tree to the model. Only the selected sample is then utilised when parameterising the decision tree to be added. This type of boosting is called “stochastic gradient boosting”.
- **Tree Constraints** - These constraints refer to limitations performed on the individual decision trees. These are essential to ensure that the weak learners of the model, remain weak. The number of trees in the model, the depth of each tree, the size of each tree and the minimum number of observations required to consider a split, are some of the restrictions that can be implemented on the construction of decision trees.
- **Penalised Learning** - Apart from limiting the structure of the trees, other penalisation methods can be applied on the weighted values of the leaves. L1 and L2 regularization of weights are amongst the popular regularization functions used to smoothen the learnt weights.

(Friedman, 2001, 2002)

2.3 | Artificial Neural Networks

ANNs are computational algorithms inspired from the biological nervous system of the human body, or rather how the human brain processes information. Similar to gradient boosting techniques described previously in 2.2, ANNs also form part of the family of supervised machine learning algorithms and thus can be applied to both classification and regression problems. These algorithms are intended to artificially replicate the behaviour of the biological systems found in the human brain known as "neurons". The biological nervous system in the human brain constitutes of millions of neurons, interconnected by synapses. Neurons are capable of receiving sensory input through dendrites, process it and then send electronic and chemical signals towards other neurons through synapses. In view of this, such system can be viewed as a network of neurons and synapses, capable of processing information.

Inspired from such concept, an ANN is an information processing algorithm that simulates how the human brain processes information. In fact, an ANN is made up of a collection of interconnected processing units known as "artificial neurons", which in turn work together simultaneously to process information and generate meaningful results. Each connection in an ANN, allows the transmission of signals between the corresponding two neurons. Similar to a biological neuron, the purpose of an artificial neuron is to receive a signal, process it and then signal other neurons that are directly connected to it. (Kröse et al., 1993; Mehlig, 2019)

It is worth noting, that a signal between two artificial neurons is typically a real number, and a connection or edge between such two neurons is assigned a weight that modifies the strength of the signal at that connection. With this in mind, the output of an artificial neuron is merely a weighted linear combination of its inputs. However to introduce non-linearity in ANNs so as to solve complex problems, an activation function is applied on the output of the neuron. This function outlines whether a neuron should be activated i.e. transmit the signal, or not based on the computed output. (Kröse et al., 1993; Mehlig, 2019)

In addition, neurons in an ANN are segmented into different layers. There are three main layers in an ANN, each having a different purpose and its own transformation of inputs. These are as follows:

- **Input Layer** - The purpose of the input layer is to sequentially feed in observations from the given dataset into the network. The number of artificial neurons in this layer is equal to the number of explanatory variables of the data. Moreover, any input received by this layer, is left unchanged and is forwarded to the succeeding layer.

- **Hidden Layer** - The purpose of the hidden layer is to perform transformations on the inputs received from the input layer, with the aim of learning a generalised concept. There can be more than one hidden layer in an ANN. Every neuron in a hidden layer transforms the received inputs by applying the activation function described above, and then transmits a signal accordingly to other connected neurons in the following layer.

- **Output Layer** - The purpose of the output layer is solely to convert signals coming from neurons in the last hidden layer, into an output variable depending on the concept being learnt.

In an ANN, the learnt concept is represented in the actual weights of the individual connections between neurons. For this reason, the learning process involves continuous weight updating. This is done by computing a cost function that calculates the difference between the predicted output and the actual value of the current data observation. The computed result is then sent back through the network, so as to adjust the weights for the next observation. The goal is to reduce the cost function after each training example. This concept is known as "back-propagation" and is constantly performed on the network until the cost function is minimized to a satisfiable value. (Kröse et al., 1993; Mehlig, 2019)

Even though ANNs are nowhere near the capabilities of the human brain, they are still considered as one of the basic fundamentals of artificial intelligence. Having nonlinearity capabilities, saw ANNs be successful in solving complex problems such as time-series predictions, pattern recognitions, data processing and robotics amongst others.

2.4 | Summary

In this chapter we gave a brief background about machine learning and the machine learning techniques explored in this dissertation. In view of this, we initially explained the concept of machine learning and how it can be of great benefit to data analysts dealing with huge amounts of data. Furthermore, we also outlined the three different approaches of machine learning techniques. Subsequently, we discussed the framework of gradient boosting and also listed some of the elements required by a gradient boosting machine learning algorithm. Finally, we gave a brief overview of artificial neural networks. We explained the inspiration behind the technique, the structure of the neural network and also the learning process in general.

3

Literature Review

In this chapter we give an overview of the related work found in literature with regards to our proposed solution and the concepts applied in this study.

In this dissertation, we present a study on how data mining can be applied on large financial data, whilst aiming to extract meaningful information for the co-operating virtual credit card company so as to make better informed decisions. Moreover, we construct a machine learning model that predicts customer churn behaviour from virtual credit card transactions.

For this reason, we start this chapter by giving an overview of the work done till now on customer churn prediction, focusing mainly on systems and approaches applied in the banking industry or else on financial data.

In practice, constructing a machine learning model or a machine learning framework does not only entail choosing a machine learning technique and applying it on the data provided. Generally, researchers are provided with raw and unstructured data, for which a number of valid features require to be identified and extracted. Furthermore, prior to the process of feature extraction, a conscious decision is commonly made on how much data is observed and utilised by the algorithm for every customer.

With this in mind, we also give an overview of the decisions taken by academics during their research when tackling and performing the extraction of churn related features, the modelling of their machine learning approach and finally the choice of the observation window size, summarising these different tasks in sections 3.1.2, 3.1.3 and 3.1.4 respectively.

3.1 | Customer Churn

In an era where there is a significant need of converting huge amounts of data into useful and valuable information (Cil et al., 2018), the task of data mining saw a huge increase in importance, popularity and applicability. Data mining refers to the process of examining large volumes of data so as to discover hidden patterns, relationships and other insights conveyed in the data (Bilal Zorić, 2016; Cil et al., 2018; Keramati et al., 2016).

It allows companies to process their data, while aiming to produce new growth opportunities in order to outperform their competition. As a result, the task of data mining has been studied and applied in numerous fields "such as marketing, finance, banking, health care, customer relationship management and organizational learning" amongst others (Cil et al., 2018).

Undoubtedly, the financial and banking industry has been evolving quite substantially in recent years. This mainly depends on the clear change in customer's expectations and choices, together with new emerging technologies as well as the need for crucial availability of financial services. Consequently, existing companies and institutions in such industry are facing extreme competition, mainly due to the interfering changes done by not only direct competitors but also by new entrants to the industry and other start-up companies that are providing innovative financial solutions. (Shirazi and Mohammadi, 2019)

For this reason, Shirazi and Mohammadi (2019) state that in order to maintain a competitive edge over other solutions so as to stay within the customers' financial path, analysing and addressing customer attraction and customer retention has become one of the major priorities when it comes to strategic planning within the financial industry.

Furthermore to such claim, marketing departments in different industries are focusing more on managing the churn behavior of customers, meaning identifying which customers are willing to end their relationship with the company, rather than investing in strategies in an attempt to acquire new customers (Bilal Zorić, 2016; Farquad et al., 2014; Kaya et al., 2018; Keramati et al., 2016; Kim et al., 2005; Leung and Chung, 2020; Rosa, 2019; Safinejad et al., 2018; Shirazi and Mohammadi, 2019; Szmydt, 2018).

This is mainly due to the fact that keeping existing customers is far less costly than finding and attaining new customers (Bilal Zorić, 2016; Kaya et al., 2018; Keramati et al., 2016; Kim et al., 2005; Leung and Chung, 2020; Rosa, 2019; Safinejad et al., 2018; Shirazi and Mohammadi, 2019; Szmydt, 2018).

Consequently, the process of churn prediction has had its fair share of interest from both the business sector as well as academic researchers. Hereafter, we will go through some of the work found in literature with regards to predicting the churn behaviour of customers in the financial sector.

3.1.1 | Similar Systems

Kim et al. (2005) are a few of the first academics that started researching and performing customer churn prediction within the financial sector. In fact, Kim et al. (2005) claim to have introduced the machine learning technique of Support Vector Machines (SVMs) to the analysis of customer churn behaviour. According to Kim et al. (2005), SVM modelling was still emerging as a new technique. Hence, one can assume that during those years, SVMs were not applied to machine learning tasks as much as they are being applied nowadays.

In view of this, they evaluate the effectiveness of an SVM in modelling the churn behavior of credit card customers. They apply this analysis on a dataset provided by a credit card company in Korea, containing demographic and credit card usage information for 9,210 customers in which 50% of these maintained their credit card during April 1997 and October 2000 whilst the remaining customers churned during the same period. Furthermore, Kim et al. (2005) only used a 3-month period when observing the credit card usage of customers as they believe that such period is adequate to understand the behaviour of a customer.

After tuning the parameters of the SVM i.e. "the upper bound C and the bandwidth of the kernel function", they compare the constructed model against a three-layer Back-Propagation Neural Network (BPN). In their experiments, the SVM model obtained better prediction results, outperforming the BPN acting as the baseline model. In addition, Kim et al. (2005) state that the process of parameter tuning is a vital step within a machine learning task as different parameter values drastically change the prediction performance.

Similarly, Farquad et al. (2014) also investigate the applicability of SVMs towards customer churn prediction. However, Farquad et al. (2014) go a step further in their work as they construct a hybrid model, where apart from being able to predict whether a customer is churning or not, it is capable of extracting a number of informative rules on the customers, using the SVM as the underlying model. Their approach can be viewed in 3 phases.

Initially, Farquad et al. (2014) reduce the number of features used when predicting churn, through a recursive feature elimination process. This process trains a linear SVM so as to compute a rank for each feature. At each step, the feature associated with the lowest rank is removed. Hereafter, Farquad et al. (2014) extract the support vectors computed after training the SVM with the reduced feature set so as to predict the churn behaviour of customers. These support vectors together with the predicted values are used to construct a new dataset to be utilised in the final phase.Eventually, Farquad et al. (2014) implement a Naive-Bayes Tree (NBTree) to purposely generate meaningful rules giving more insights about the churn behaviour of customers.

Farquad et al. (2014) employ this hybrid model on a credit card dataset provided by a Latin American bank that experienced a substantial amount of churns within their credit card clientele. The dataset comprises of 14,814 customers, however is highly imbalanced as only 7% of customers are churners. For this reason, Farquad et al. (2014) experiment with various sub-sampling techniques including under-sampling and over-sampling as well as the Synthetic Minority Oversampling TEchnique (SMOTE). Originally, the dataset contains both socio-demographic and behavioural data, yet the feature selection process only maintains 6 behavioural features that are observed within the 3 months prior to the labelling period.

Correspondingly, there are other works in literature that follow a similar approach to the one applied in Farquad et al. (2014), where researchers aim at generating a number of informative rules when analysing customer churn. The generated rules tend to group customers into different segments according to common behaviour. Generally, such information is extremely beneficial to companies and organisations when they are planning new marketing strategies with regards to customer retention.

In view of this, as part of their customer churn prediction model, Keramati et al. (2016) aim to outline common characteristics of those customers that are predicted to churn. To identify any hidden behavioural patterns, Keramati et al. (2016) employ a Decision Tree (DT) model as their customer churn prediction model. This is mainly because, by nature, DTs generate clear and significant "if-then" rules, allowing Keramati et al. (2016) to fulfill their aim.

Keramati et al. (2016) make use of a bank's sample dataset containing 4,383 customers that are enrolled within the bank's electronic banking services. The dataset is highly imbalanced with merely 1.5% of customers labelled as churners. For this reason, Keramati et al. (2016) perform random sampling with replacement by utilising a bootstrap sampling module from RapidMiner [1] data mining tool. Keramati et al. (2016) state that the data is limited to only "customer dissatisfaction, level of service usage and customer-related" information. After pre-processing this data, forward selection and backward elimination methods were both performed and evaluated in order to select the best possible feature set before constructing the decision tree and generate the

[1]Sept 2020: https://rapidminer.com/glossary/data-mining-tools/

desired meaningful rules.

Similarly, Cil et al. (2018) also utilise DTs in their quest to discover meaningful knowledge from their dataset. Cil et al. (2018) are provided with a dataset consisting of socio-demographic information together with nearly 4,000,000 investment fund transactions of around 65,525 customers of a specific bank. The investment funds transaction data are pre-processed by summing up the amounts for each month and calculate a percentage change form the original principal amount of the customer. These percentage changes are then converted from numeric to categorical values through the use of clumping. During this process, the number of clusters or categories to use is decided in close collaboration with the business depending on studies that were previously performed. It is worth mentioning that for these features, Cil et al. (2018) decided to observe up to 6 months prior to the closure or inactivity of a customer's account from the customer's entire investment fund transaction history.

Subsequently, Cil et al. (2018) perform 2 types of analysis. Initially, they model a DT on the computed investment fund transaction order data so as to determine the transactional patterns of customers that closed off their account with the bank. The learnt patterns in the form of DT rules are then utilised on future customers to predict those that potentially are willing to churn. Hereafter, Cil et al. (2018) model another DT, this time using the socio-demographic features, so as to determine the common socio-demographic characteristics of customers that end their relationship with the bank. J4.8, PART, JRip, Naive Bayes and OneR classification algorithms are used for the first analysis, whilst ID3 decision tree is used for the latter analysis when discovering socio-demographic rules for the bank's marketing department.

Shirazi and Mohammadi (2019) are another pair of academics that employ DTs as the classification model for their customer churn prediction task. In addition, Shirazi and Mohammadi (2019) apply big data analytics specifically towards customer churn prediction in the retiree segment. According to the "Personal and Commercial Banking" department of their "target bank", the customer retiree segment was identified as the highest strategic priority. In fact, the dataset utilised in this study comprises of all mass and non-mass affluent customers that are between 50 to 71 years old having at least a single active account and are not yet retired at the beginning of the observation window (i.e. from November 2011 till September 2015). This entails that apart from the churning event, Shirazi and Mohammadi (2019) also analyse the retirement event within the customers of their "target bank".

Their big data analysis can be divided in 2-fold. Initially, Shirazi and Mohammadi (2019) construct a DT using the Classification and Regression Tree (CRT) method as their customer churn prediction model. This modelling phase utilises both structured and unstructured data, including customer socio-demographics, information on the enrolled bank products of a customer, "pixel data" that depicts the customer's online tracks on various websites, Interactive Voice Response (IVR) logs, phone conversations and finally the actual financial data. Subsequently, Shirazi and Mohammadi (2019) employ a General Linear Model (GLM) analysis, verifying a number of hypotheses regarding the behaviour of churning. In this final phase, a series of T-Tests and ANOVA statistical and correlation tests are performed on the mentioned hypotheses. As a result of their study, Shirazi and Mohammadi (2019) claim that the behaviour of customers significantly correlates with their churn event and also that performing targeted marketing and "offering the right product at the right time" reduces the rate of churn.

Contrary to some of the works already reviewed in this section, some of the studies found in literature do not focus on acquiring meaningful rules or else determine common characteristics of churning customers, but rather focus on predicting whether a customer is churning or not as efficiently as possible. As a matter of fact, recent studies have widely adopted the standard ANN machine learning technique when tackling the problem of customer churn prediction within the financial sector.

Bilal Zorić (2016) proposes a neural network based framework that is capable of predicting the likelihood of churn for customers of a small Croatian bank. The dataset used in this study contains merely socio-demographic information and levels of service usage for 1,866 customers. Bilal Zorić (2016) designed and constructed an ANN with three hidden layers consisting of 8, 4 and 2 neurons, by making use of Alyuda NeuroIntelligence [2] software package. Their motivation involves the fact that customers that are registered within multiple bank products, tend to not close off his relationship with the bank and thus work should be focused on customers with a few products or accounts. In fact, after examining the features of some of the customers and the predicted likelihood of churn, Bilal Zorić (2016) concludes that the bank should start addressing the needs of students and young customers including favourable interests, student loans and online services, as they are more likely to churn.

Similar to Bilal Zorić (2016), Safinejad et al. (2018) also employ the non-linear ANN technique in their study to predict future churn rate of customers from their financial transactions. Safinejad et al. (2018) make use of a dataset that was retreived from a financial institution and contains raw financial transactions of more than 4,500 customers recorded between 2009 and 2011. The 3-year observation window is divided into seasonal intervals and for each interval (12 in total), Recency (R), Frequency (F), Monetary (M) and Length (L) variables are calculated as features.

[2] Sept 2020: https://www.alyuda.com/product/neural-networks-software

Furthermore, Safinejad et al. (2018) describe a "fuzzy dynamic model" that can be split into 3 phases. Firstly, a weighted-RFML model is utilised to cluster customers and identify the segment representing the most valuable customers. Since the 4 factors differ in importance, the weights of this model are computed based on experts' opinions acquired through a questionnaire. Secondly, a fuzzy rule-based model that takes as inputs the L, F and M variables and outputs a 3-mode (low, medium, high) churn rate value, is developed. This is applied on all the 12 seasonal intervals of all customers. Thirdly, Safinejad et al. (2018) models the prediction of future churn rate using 2 models - ARIMA as a linear machine learning model and ANN as a non-linear model. The employed ANN is a two-layer feed forward network with 10 hidden neurons, constructed using MathWorks' Neural Network Time-Series tool [3]. Safinejad et al. (2018) conclude that the ANN model outperformed the linear model claiming to have identified a suitable model for customer churn prediction together with an appropriate definition of churn in the finance sector.

Rosa (2019) is another researcher that utilises the ANN machine learning technique when predicting customer attrition. Furthermore, Rosa (2019) proposes a data-driven machine learning framework that is capable of tackling and predicting customer churn behaviour in one of the major and most well-known retail banks in Portugal. This study is provided with a sample dataset containing behavioural information of more than 90,000 customers out of which only 1,588 are labelled as churners. In this study, both the observation and the labelling window are chosen to be 6 month periods. According to Rosa (2019), students make up of nearly 25% of churned customers during 2017. Due to their barely any involvement with the bank, thus making the task of customer churn prediction even more difficult, Rosa (2019) decided to exclude students within this study.

The behavioural information that is fed into the modelling technique, involves merely socio-demographic features and binary variables representing what actions were or were not taken by the customer during the observation window. As previously mentioned, Rosa (2019) develops a customer churn predictive model through the use of ANNs. In fact, Rosa (2019) constructs and evaluates 4 different neural networks, all with differing number of hidden layers. Hereafter, the best performing model is validated on a new dataset of nearly 135,000 customers whose data was observed 6 months later than the original dataset. Despite obtaining an accuracy score of 84%, the model only managed to predict nearly 25% of the churned customers.

Most of the customer churn prediction systems reviewed above, focus solely on do-

[3]Sept 2020: https://www.mathworks.com/help/deeplearning/ref/ntstool.html

main specific and static features. Some of these features that are extracted and used in such traditional attempts, generally represent product or account type ownership, service usage aggregation and socio-demographic information. Dynamic behavioural patterns in a customer's financial transactions, are rarely considered.

In fact, Kaya et al. (2018) try to fill this research niche by exploring the spatio-temporal patterns and choice behaviour of customers from their financial transactions and determine whether such behaviour relates to the customer churn event. First of all, Kaya et al. (2018) perform this study on a dataset provided by a major financial organisation in an OECD country, consisting of "demographic information, credit card transactions, money transfers and electronic fund transfers" of over 100,000 customers. Secondly, Kaya et al. (2018) extract novel features based on the spatio-temporal and choice patterns of customers, hidden within the financial transactions.

The spatio-temporal features that are namely *diversity*, *loyalty* and *regularity* measure how varied or constant customers are within their purchase behaviour with regards to time and location perspective. On the other hand, the financial choice patterns outline how customers disperse their spending with regards to merchants, purchase categories and locations of merchants. Kaya et al. (2018) also introduce the *entropy of choice* behavioural feature which represents the entropy or variety of customers' choices with regards to products and merchants. Hereafter, Kaya et al. (2018) employ Random Forests (RFs) classification technique so as to predict the churn behaviour of customers. Kaya et al. (2018) claim that churn activities can be effectively predicted using dynamic behavioural patterns and furthermore, using domain-independent variables specifically those that are based on the spatio-temporal patterns in human activities.

Similar to Kaya et al. (2018), Leung and Chung (2020) also propose a dynamic classification framework that aims to capture the actual customer behavioural patterns required for effective churn prediction. This study is applied on a dataset comprising of over 32,000 customers of a well-known retail bank in Florida, USA that offers various savings and loan products. For every customer, this dataset contains static predictors, such as the traditional demographic and product ownership variables, and also account activity predictors including aggregation of financial transactions and service usage. In view of this, a trend factor is computed for each account activity predictor, aiming to capture the trends of account activities within the observation period.

Leung and Chung (2020) experiment with both the observation window (4 months vs 6 months) and the labelling window (2 months vs 3 months). After computing the trend factors for each different period, Leung and Chung (2020) evaluates 3 different supervised machine learning models, namely Logistic Regression (LR), RF and GBM. Leung and Chung (2020) concludes that with 6 months of data, the models obtained better accuracy than with 4 months of data. On the other hand, accuracy decreases rapidly as the prediction window is extended. Furthermore, RF and GBM outperformed the LR prediction model. Finally it is worth mentioning that to overcome the issue of having an imbalanced dataset, Leung and Chung (2020) make use of multiple observations of the same customers from different type periods. The authors state that this is a limitation of their framework as this entails a potential lack of independence between the training instances.

Deep learning models are known to be able to capture high-level representations from the huge amounts of customer data being created at an increasing rate due to the increasing amount of financial activities (Hsu et al., 2019). Following on the previous studies, dynamic behavioural predictors have shown to be more effective in representing customer behaviour for churn prediction.

In view of this, Hsu et al. (2019) develop a Recurrent Neural Network (RNN) feature extractor with GRU. The aim is to better model the time dependencies found within a customer's credit card spending history, as a result of which, a number of dynamic features are then extracted for customer churn prediction. Consequently, Hsu et al. (2019) propose an innovative approach by combining this strong dynamic feature extraction from RNN with a RF. This enhanced RNN-RF model is therefore capable of combining dynamic and static features allowing for better performance when predicting credit card customer churn.

Hsu et al. (2019) evaluate their innovative approach on an open dataset from UCI Machine Learning Repository [4]. This dataset consists of 30,000 instances with 23 features each, separated into 5 static socio-demographic features and 18 dynamic features describing monthly the customer's service usage in a 6 month period. Finally, Hsu et al. (2019) conclude that the RNN-RF predictive model outperformed other benchmark models and also stated that the model performed better with more training instances.

[4]Sept 2020: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

3.1.2 | Churn Related Features

In this section, we give an overview of the churn related features utilised by the prediction systems related to our proposed solution that have just been reviewed in Section 3.1.1. As previously mentioned, the task of extracting features from the raw dataset is a crucial step in the whole machine learning process. This is because the performance of the constructed machine learning model highly depends on the predictors it is given. In view of this, in Table 3.1 we list the different feature types used by similar systems found in literature.

Year	Researchers	Features
2005	Kim, S. et al.	Demographic and static service usage information
2014	Farquad, M. A. H. et al.	Socio-demographic and dynamic behavioural information
2016	Keramati, A. et al.	Socio-demographic, customer dissatisfaction and static service usage information
2016	Bilal Zorić, A.	Socio-demographic and static service usage information
2018	Cil, F. et al.	Socio-demographic and dynamic behavioural information
2018	Kaya, E. et al.	Dynamic behavioural features with spatio-temporal and financial choice information
2018	Safinejad, F. et al.	Dynamic behavioural information
2019	Hsu, T.-C. et al.	Demographic and dynamic behavioural information
2019	Rosa, N. B. C.	Socio-demographic and static service usage information
2019	Shirazi, F. et al.	Socio-demographic and unstructured behavioural information
2020	Leung, H. C. et al.	Static service usage and dynamic behavioural information

Table 3.1: Summary of the different churn related features extracted and used by related systems.

3.1.3 | Modelling Approaches

In this section, we give an overview of the different modelling approaches employed by the prediction systems related to our proposed solution, reviewed in Section 3.1.1. As one might have noticed, the machine learning models applied for customer churn prediction differ by time due to emerging new technologies and also depending on the features that were extracted or that were already available. In view of this, in Table 3.2 we list the different modelling approaches taken by similar systems found in literature.

Year	Researchers	Machine Learning Methods Used
2005	Kim, S. et al.	SVM
2014	Farquad, M. A. H. et al.	SVM
2016	Keramati, A. et al.	DT
2016	Bilal Zorić, A.	ANN
2018	Cil, F. et al.	DT
2018	Kaya, E. et al.	RF
2018	Safinejad, F. et al.	ANN
2019	Hsu, T.-C. et al.	Combined RNN and RF
2019	Rosa, N. B. C.	ANN
2019	Shirazi, F. et al.	DT
2020	Leung, H. C. et al.	LR, RF and GBM

Table 3.2: Summary of the different modelling approaches constructed and employed by related systems.

3.1.4 | Data Observation

In this section, we give an overview of the different data used by similar systems and also outline the different choices taken regarding the size of the observation window. The related systems review given in Section 3.1.1, outlines how academics in this field are dependent on the cooperation of financial firms and institutions in terms of data. One can appreciate the fact that there is a lack of gold standard datasets in this research domain due to the sensitivity of financial data. Furthermore, the size of the observation window differs between systems. In view of this, in Table 3.3 we list the different data used and also the different choices taken on the size of the observation window by similar systems found in literature.

Year	Researchers	Real Data Sample	Observation Period
2005	Kim, S. et al.	Credit card data from a Korean credit card company	3 months
2014	Farquad, M. A. H. et al.	Credit card data from a Latin American bank	3 months
2016	Keramati, A. et al.	E-banking service data from a bank	2 years
2016	Bilal Zorić, A.	Banking services data from a Croatian bank	N/A
2018	Cil, F. et al.	Investment funds transactional data from a bank	6 months
2018	Kaya, E. et al.	Financial transactional data from a financial organisation	9/12 months
2018	Safinejad, F. et al.	Raw financial transactional data from a financial institution	3 years
2019	Hsu, T.-C. et al.	Open dataset from UCI Machine Learning Repository [5]	6 months
2019	Rosa, N. B. C.	Banking services data from a Portuguese bank	6 months
2019	Shirazi, F. et al.	Data from multiple sources on the retiree segment of a bank	N/A
2020	Leung, H. C. et al.	Banking services usage data from a bank in Florida, USA	4/6 months

Table 3.3: Summary of the different data and observation window sizes utilised in related systems.

3.2 | Summary

In this chapter we discussed the work found in literature that is related to our proposed solution. We first gave a brief background on data mining and the problem of customer churn prediction. Here, we also mentioned how the process of customer churn retention has become one of the major strategic priorities within the financial industry. Subsequently, we discussed the different approaches and decisions performed by similar churn prediction systems. Finally, we summarised how academics tackled the extraction of churn related features, the modelling of their machine learning approach and also the choice of the observation window size.

4

Design & Implementation

After reviewing related systems found in literature, in this chapter we shift our focus towards our proposed solution, outlining its design choices together with all the implementation details. In fact, we first give a brief overview of the whole solution and how its different phases are aligned with the objectives set in Section 1.3. Furthermore we also discuss all our design choices, justifying our decisions using the literature discussed in Chapter 3.

Subsequently, an analysis of the provided dataset is given together with a discussion of any pre-processing performed on the data. Hereafter, we describe in more detail the implementation aspect of our solution, focusing mainly on the development done to achieve the set objectives.

4.1 | Design Choices

The main objectives of our solution as discussed in Section 1.3, are i) to extract dynamic features from raw financial transactions in order to capture the customer behavioural patterns for churn prediction, ii) to construct a machine learning setup that is capable of predicting customer churn, iii) to determine the minimum amount of customer activity needed prior to churning and iv) to determine whether demographic information combined with any initial financial transactions can be used to predict churning behaviour for relatively new customers.

Initially, the framework pre-processes the raw financial transactions to filter out any missing or redundant data. In the pre-processing stage, the financial transactions are also segmented into several time periods as a preparatory work for future analysis. Hereafter, numerous features are generated from the financial transaction records so as to extract and represent any behavioural patterns hidden within the data. Furthermore, customers are also classified and labelled according to our customer churn definition. Subsequently, the extracted features and the generated customer churn labels are fed into a machine learning technique in order to model the churn behaviour of customers and be able to predict churn activities.

Consequently, after determining the best set of features and the best performing predictive model, an experiment is performed where different observation window sizes are evaluated to determine the minimum amount of customer activity required prior to the churn event. Finally, a time-series experiment is performed to determine whether demographic information combined with any spending information that is available at the beginning of the customer's relationship with the company, can be used to predict whether new customers are willing to continue using the company's service.

In the following sections, we discuss the decisions made in each of the above mentioned tasks, justifying our choices with the use of literature.

4.1.1 | Extracted Features

In this study we extract different types of features aimed at representing the customer behaviour required for churn prediction.

In our review of literature, we have seen how Bilal Zorić (2016); Cil et al. (2018); Farquad et al. (2014); Hsu et al. (2019); Keramati et al. (2016); Kim et al. (2005); Leung and Chung (2020,?); Rosa (2019); Shirazi and Mohammadi (2019) all make use of socio-demographic information within their churn predictive systems. For this reason, we followed the traditional approach and decided to extract and make use of any customer demographics found in the provided dataset as well.

Furthermore, apart from socio-demographic features, Bilal Zorić (2016); Keramati et al. (2016); Kim et al. (2005); Leung and Chung (2020); Rosa (2019) also make use of static predictors representing the service usage of customers within the observation time period. With this in mind, we compute a number of statistical features that aggregate different aspects of the purchase history of a customer within a particular observation window.

Latest research has shown that dynamic behavioural features tend to be more effective in representing customer behaviour for churn prediction (Hsu et al., 2019; Kaya et al., 2018; Leung and Chung, 2020). For this reason, some of the statistics that are to be computed on the entire observation window, will also be applied for each month period in the window, creating a new set of macro-average monthly statistics. Such feature extraction process is similar to the one performed by Leung and Chung (2020). However Leung and Chung (2020) then compute a trend factor value based on the computed dynamic features, prior to inputting the features into the predictive model. At this stage, we followed the approach taken by Kaya et al. (2018) where we input the dynamic features directly in the predictive model.

Apart from the monthly statistical features, we also compute a feature vector containing the amounts spent by the customer on each day of the observation window. The motivation behind this set of features is to allow the predictive model to train on variables that resemble the raw financial transactions as much as possible in an attempt to not loose any information regarding the dynamic behaviour of customers within the observation period during the feature extraction process.

Finally, inspired by the innovative idea of having features representing the choice behaviour of customer with regards to merchants, purchase categories and locations of merchants Kaya et al. (2018), we generate a merchant vector containing the number of purchases done within the observation window towards each Merchant Category Code (MCC).

As can be seen, in this study we make use of different types of features including demographic, static and dynamic predictors in order to predict the churn behaviour of customers in the financial industry. The full list of features is shown in Table 4.1. These are to be evaluated in Section 5.3 so as to select only those features that actually capture the behaviour of customers.

4.1.2 | Machine Learning Techniques

Since our focus was never on acquiring meaningful rules or else determining common characteristics of churning customers, but rather on predicting whether a customer is churning or not as efficiently as possible, we followed the approach taken by Bilal Zorić (2016); Rosa (2019); Safinejad et al. (2018), that is employing a Neural Network as the customer churn predictive model. In Section 2.3, we described how ANNs are intended to artificially replicate the behaviour of the biological systems found in the human brain. In fact, together with other researchers contributing to the field of churn prediction, we believe that this non-linear predictive model is a suitable contender in modelling the churn behaviour.

Furthermore, we also employ a Gradient Boosting Model (GBM) as the customer churn predictive model. The fact that GBM is capable of tuning weak predictive models so as to become better predictors by generating a single predictive model as an ensemble of numerous weak ones, inspired us to make use of such technique in our quest to predict customer churn. In addition, Leung and Chung (2020) stating that GBM performed better than a LR and was also on par with RF, gave us further motivation in constructing such predictive model.

4.1.3 | Data Observation and Labelling

Most of the studies reviewed in Section 3.1.1, did not opt for an observation window exceeding 6 months. In fact, most of the systems we reviewed employ an observation window size varying between 3 to 6 months. As a result, in Section 5.4 we perform an experiment where we train our predictive model on varying observation window sizes starting from 1 month worth of data up to 6 months. The results of such experiment will give us an inclination of the minimum amount of customer activity needed to predict churn.

It is worth mentioning that for the evaluation of the extracted features and also that for the constructed predictive models, the observation window size is taken to be 3 months. We have chosen an observation window size of 3 months since it is the smallest window size that was used by the similar systems reviewed in Section 3.1.4.

On the other hand, when we are observing the financial transactions in the period following the observation window so as to determine the churn label for customers, we decided to only consider the succeeding month. This is mainly because according to Leung and Chung (2020), prediction accuracy decreases instantly as the prediction window increases. Furthermore, companies and marketing departments would find it more beneficial if they can predict what activity is expected in the coming month. In addition, we do not employ any fuzzy logic in our customer churn definition, meaning that a customer can either be labelled as "Churned" or not "Not Churned". In fact, our customer churn definition is quite straightforward - if a customer has at least 1 transaction in the labelling window then the label is "Not Churned" else "Churned".

4.1.4 | The Cold Start Problem

Padilla and Ascarza (2020) overcome the "cold start" problem within Machine Learning solutions for CRM, by developing a probabilistic modelling framework that given the behavioural patterns observed at the start of the customer's experience, is able to simulate and predict with a probability future actions of this customer. However Padilla and Ascarza (2020) state that their probabilistic model is still dependent on a number of variables that are required upon customer registration, making it challenging when applying it to empirical scenarios.

For this reason, we believe that this problem requires a research on its own. Thus, we opted to perform a simple time-series experiment so as to examine other simpler possibilities to overcome the issue of having barely any data to use when predicting newly registered users. In this time-series experiment, we start predicting churn using initially the demographic features of customers and then gradually add the amounts spent by the customer each day. This experiment is performed on only the first month of the customer's history since then we will overlap with the previous experiment.

4.2 | Implementation Details

In this section, we describe any relevant details regarding the implementation of the different aspects of our proposed solution. Initially, we start by describing the dataset used in this study and describe any pre-processing steps done. Hereafter, we describe the extracted features and subsequently, explain the implementation details of our machine learning classifiers. All implemented code is developed in Python programming language.

4.2.1 | Data Description and Pre-Processing

In this study, we examine a financial dataset provided by a leading virtual credit card company based in Malta. This dataset consists of 2,996,700 financial transactions that were performed by 266,459 different customers between 1st January 2016 and 10th January 2018. It is worth mentioning that the provided transactions were not all made in the same currency. For consistency purposes, we convert the amounts of all transactions from their original currency to their equivalent Euro amount.

As previously mentioned, these financial transactions are segmented into several time periods as preparatory work for future analysis. In fact, the first step of pre-preprocessing involves splitting the original dataset into multiple subsets where each subset corresponds to a single month between January 2016 and January 2018. Consequently, from each subset, we build a dictionary where each distinct user is associated to an array of transactions performed within that corresponding month.

Hereafter, these monthly dictionaries are utilised to generate observation windows of any size prior to feature extraction. In addition, such dictionaries are also used when labelling customers since the labelling window is taken to be a single month.

In every dataset, there can be outliers or rather data that can be filtered out so as not to disrupt the learning performance of the predictive model. With this in mind, prior to filtering out customers that we think might hinder the learning process of the machine learning model, we define and analyse 4 different user activity levels so as to make a better conscious decision. The different user activity definitions are as follows:

1. Active in all months of the observation window
2. Active in at least the first and last month of the observation window
3. Active in at least the last month of the observation window
4. Active in the majority of the months of the observation window

In Figure 4.1, we showcase how churn percentages within customers change with different observation window sizes. On the other hand in Figure 4.2, we showcase the amount of eligible customers in each different observation window. Based on these 2 distributions, we decide to filter customers based on the second activity definition, entailing that any evaluation or experiments done in this study, will be applied only on customers that are active in at least the first and last month of the observation window. Intuitively, such definition ensures that any churn prediction is performed on customers that are currently active and have not churned already.

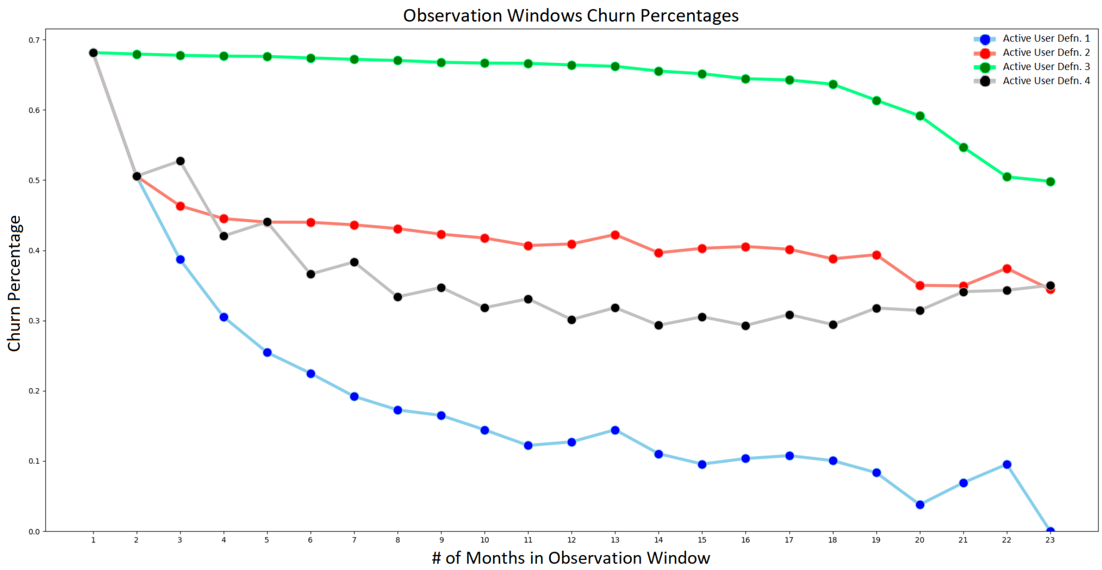

Figure 4.1: This graph depicts how the churn percentage amongst customers changes with increasing observation window size and how this differs amongst the different user activity definitions.

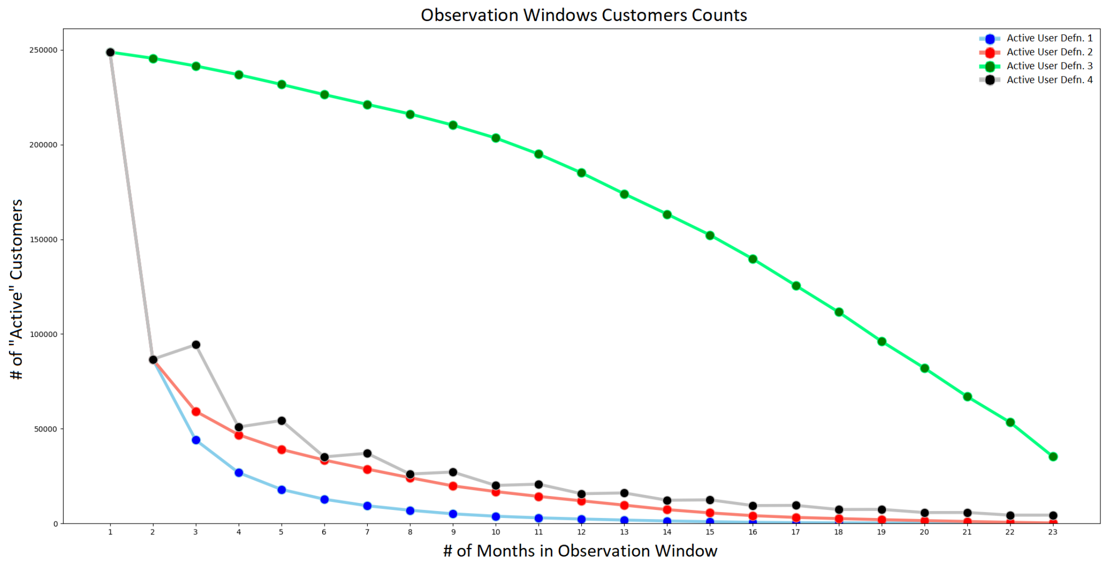

Figure 4.2: This graph depicts how the amount of customers observed changes when the observation window increases and how this differs amongst the different user activity definitions.

4.2.2 | Extracted Features

In this study we extract different types of features aimed at representing the customer behaviour required for churn prediction. In Section 4.1.1 we justified our choice of extracting and making use of demographic information, "global" statistics that are relative to the entire observation window, dynamic monthly statistics for each month in the observation window, a vector containing daily purchase amounts and another vector containing the number of purchases done towards each MCC. In Table 4.1 we give the full list of features specifying whether the feature is categorical or numerical. One-Hot Encoding is performed on all categorical features.

Feature Category	Feature	Type
Demographic	User Age	Numeric
	User Age Group	Categorical
	User Country Code	Categorical
	User Currency Code	Categorical
Global Statistics	Average Number of Days Between Purchases	Numeric
	Median Number of Days Between Purchases	Numeric
	Number of Days Between First and Last Purchases	Numeric
	Number of Purchases on each Week Day	Numeric
	Week Day with Most Purchases	Categorical
	Total Number of Purchases	Numeric
	Total Amount Spent in Purchases	Numeric
	[User History] Average Number of Purchases Per Day	Numeric
	[User History] Average Number of Purchases Per Week	Numeric
	[User History] Average Number of Purchases Per Month	Numeric
	[User History] Average Amount Spent Per Day	Numeric
	[User History] Average Amount Spent Per Week	Numeric
	[User History] Average Amount Spent Per Month	Numeric
	[Observation Window] Average Number of Purchases Per Day	Numeric
	[Observation Window] Average Number of Purchases Per Week	Numeric
	[Observation Window] Average Number of Purchases Per Month	Numeric
	[Observation Window] Average Amount Spent Per Day	Numeric
	[Observation Window] Average Amount Spent Per Week	Numeric
	[Observation Window] Average Amount Spent Per Month	Numeric
Monthly Statistics	Average Number of Days Between Purchases	Numeric
	Median Number of Days Between Purchases	Numeric
	Number of Days Between First and Last Purchases	Numeric
	Number of Purchases on each Week Day	Numeric
	Week Day with Most Purchases	Categorical
	Total Number of Purchases	Numeric
	Total Amount Spent in Purchases	Numeric
Daily Purchases Amounts Vector	Amount Spent in each Day	Numeric
Merchant Category Code Counts Vector	Number of Purchases towards each Merchant Category Code	Numeric

Table 4.1: Detailed list of all the extracted features.

4.2.3 | Machine Learning Techniques

In Section 4.1.2, we explained and justified why we chose to construct an Artificial Neural Network (ANN) model and a Gradient Boosting (GBM) model as our customer churn classifiers. Both classifiers are implemented using readily available libraries. We decided to implement our ANN using Keras[1] which is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. On the other hand, we decided to construct our GBM using XGBoost[2]. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

In order to instantiate either one of these models, we need to choose preset values for a number of hyper-parameters. In Section 5.2, we come up with a range of possible values for each parameter and then perform a test to determine the best parameter setup.

4.2.4 | The Cold Start Problem

In Section 4.1.4, we discussed that a time-series experiment is to be performed where we initially try to predict the churning behaviour of newly registered customers using their demographic information. Then, as new data starts coming in for these customers, we add daily purchase amounts to the feature set so as to perform the customer churn prediction once again, measuring the performance each time.

For this experiment, we do not apply any filtering to select the newly registered customers. Thus, for all the customers found in the data (more than 260,000), the appropriate features have been extracted to be used by the predictive model.

As aforementioned, this time-series experiment only observes the first month of the customers' history. The analysis is done on a weekly basis, meaning that the churn prediction is performed every week.

As a result, the first month of the customer's history is split into 4 weeks to create 4 observation windows of a week in size. Hereafter for each weekly observation window, we extract the daily purchases amount vector i.e. a feature vector of length 7 comprising of the purchase amounts for every day in that week.

Initially, we attempt to predict customer churn using only the demographic information of the user. Then after observing a week of customer's history, the purchases amount vector for that week is appended to the current feature set before attempting to predict churn one again.

In Section 5.5, we output and discuss the results of this simple experiment.

[1] Sept 2020: https://github.com/keras-team/keras
[2] Sept 2020: https://github.com/dmlc/xgboost/tree/master/python-package

4.3 | Summary

In this chapter, we gave an overview of our proposed solution. We started the chapter by discussing the decisions and design choices we took in order to fulfill our objectives, making use of the reviewed literature to justify our choices. Subsequently, we described the financial dataset that was provided by the virtual credit card company cooperating in this study. Hereafter, the complete list of extracted features is provided. We then discussed briefly how the machine learning techniques examined in this study, are implemented and finally we explain in detail the time-series experiment to be performed as an investigation of the "cold start" problem within Machine Learning.

5

Evaluation

In this chapter, we present the evaluation conducted on the proposed system in relation to the objectives mentioned in Section 1.3. Essentially, the evaluation of our system is divided into 3 parts: evaluation of the features extracted from the raw financial transactions and of the model employed to predict customer churn behaviour, evaluation of the varying observation window sizes to determine the minimum amount of customer history required for churn prediction and finally the evaluation of the usage of demographics together with any initial behavioural observations when predicting churn for relatively new customers. We first discuss how similar systems have been evaluated in literature, and using this knowledge we construct an evaluation plan. For each objective of our solution, we then describe the experiments conducted and discuss the results obtained together with any observations and conclusions we deduced.

5.1 | Similar Systems

As already mentioned in the motivation of this dissertation (Section 1.1), academics and researches working on CRM in the financial sector, require close collaboration with the business side, in order to acquire substantial amount of financial data that can be processed and analysed. In addition, any company data together with the information resulting from research rarely get shared with the scientific community, due to its sensitive nature. For this reason, there is a lack of gold standard datasets in this research domain, making the comparison of methods proposed by other researchers extremely challenging. In fact, all the related systems reviewed in Section 3.1.1, do not compare their findings against those obtained in other research work, but solely evaluate their own proposed system on a dedicated test set and discuss the results in terms of various performance metrics.

Apart from the traditional Accuracy score, the performance of most classification models is measured using the Area Under the Receiver Operating Characteristics Curve (AUROC) metric. This metric provides an aggregate measure of performance across all possible classification thresholds. This metric has been used to evaluate the systems described in Hsu et al. (2019); Kaya et al. (2018); Keramati et al. (2016); Rosa (2019). Furthermore, generally researchers tabulate the comparison of the predicted and actual values through a confusion matrix, as in Cil et al. (2018) and Rosa (2019). In addition, the cell values of the confusion matrix can be utilised to compute other metrics such as True Positive Rate (Sensitivity), True Negative Rate (Specificity), False Positive Rate, False Negative Rate, Precision and Recall, that can be used to infer more conclusions on the performance of the classification model.

In our evaluation, we will make use of the AUROC score as the main performance metric, however we will compute other metrics based on the confusion matrix in our discussion of results. Furthermore, we follow the general approach of splitting the data into training and testing sets using the 90:10 split.

5.2 | Parameter Tuning of the Machine Learning Models

In order to instantiate our classifiers so as to evaluate the objectives of the proposed solution in the subsequent sections, parameter tuning is required for both machine learning models. As aforementioned, both the ANN and GBM classification models require a set of hyper-parameters that need to be carefully selected since they have an affect on the learning capabilities of the models and thus can hinder their prediction performance.

The best set of parameters are different for each research problem and are determined mostly using an empirical process. For this reason, a Grid Search is implemented. A Grid Search can be considered as an optimization problem. Its aim is to optimize the set of parameters in relation to a performance metric. This form of search tries all possible combinations of parameters (within a pre-specified range) to determine the best performing set of parameters.

Apart from the range of values for each hyper-parameter of the classification models, we also need to select the type of features to be used and the observation window size. As previously mentioned in 4.1.3, the observation window size is taken to be 3 months until we conduct the experiment in Section 5.4. On the other hand, for the parameter tuning process, we will use a combination of demographics, global and monthly statistics to have a mixture of static and dynamic features.

In addition, the range of potential parameter values for both the ANN and GBM classification models, together with the best performing parameter setup resulting from the Grid Search, are shown in Tables 5.1 and 5.2 respectively.

Hyper-Parameter	**Possible Parameter Values**	**Selected Value**
Hidden Layers	[1, 2, 3]	3
Neurons in Hidden Layer	[50, 150, 250, 500, 750, 1000]	500
Dropout Rate	[0, 0.5]	0.5
Activation Function	[linear, relu, tanh, softmax, sigmoid]	sigmoid
Optimizer Function	[sgd, rmsprop, adadelta, nadam, adamax]	rmsprop

Table 5.1: List of the different parameters to be tuned for ANN, with their potential values and their resulting best performing value.

Hyper-Parameter	**Possible Parameter Values**	**Selected Value**
Learning Rate	[0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]	0.1
Maximum Depth of Tree	[6, 10, 14, 18]	18
Subsampling Ratio	[0.5, 0.75, 1]	1
Column Subsampling Ratio	[0.5, 0.75, 1]	0.5
L1 Regularisation	[0, 5, 10]	5
L2 Regularisation	[1, 5, 10]	5

Table 5.2: List of the different parameters to be tuned for GBM, with their potential values and their resulting best performing value.

5.3 | Extracted Features and Machine Learning Techniques Used

In order to determine the most effective features in capturing customer behavioural patterns and the best performing modelling technique in customer churn prediction, a greedy search is employed. We compute all the combinations of the different feature categories and feed the computed feature set into our two predictive models. We measure the prediction performance of each model using the AUROC metric, and tabulate the results for our ANN model in Table 5.3 and those obtained by our GBM model in Table 5.4.

Feature Category					AUROC
Demographics	Global Statistics	Monthly Statistics	Merchant Vector	Daily Vector	
X					0.5
	X				0.5741
		X			0.5432
			X		0.5319
X	X				0.5761
X		X			0.532
X			X		0.5249
	X	X			0.5765
	X		X		0.5806
		X	X		0.5314
X	X	X			0.5888
X	X		X		0.5826
X		X	X		0.5712
	X	X	X		0.5758
X	X	X	X		0.5947
				X	0.5591
X				X	0.5529
	X			X	0.5627
		X		X	0.5517
			X	X	0.5729
X	X			X	**0.6159**
X		X		X	0.5803
X			X	X	0.5778
	X	X		X	0.5873
	X		X	X	0.5947
		X	X	X	0.5699
X	X	X		X	0.5846
X	X		X	X	0.5804
X		X	X	X	0.5652
	X	X	X	X	0.5974
X	X	X	X	X	0.5824

Table 5.3: Evaluation results of the extracted features and of the ANN predictive model.

The ANN classification model obtained its highest AUROC score when trained on demographic information, global statistics aggregating the entire 3-month observation window and the vector comprising of the daily purchase amounts, obtaining a score of 0.62. With such performance, we can conclude that the implemented ANN model has some form of ability in distinguishing between the 2 classes.

Feature Category					AUROC
Demographics	Global Statistics	Monthly Statistics	Merchant Vector	Daily Vector	
X					0.5332
	X				0.6508
		X			0.6387
			X		0.6347
X	X				0.6579
X		X			0.6448
X			X		0.6354
	X	X			0.6563
	X		X		0.6643
		X	X		0.6561
X	X	X			0.6602
X	X		X		0.6586
X		X	X		0.6523
	X	X	X		0.6691
X	X	X	X		0.6634
				X	0.6847
X				X	0.6788
	X			X	0.6825
		X		X	0.683
			X	X	0.6812
X	X			X	0.6845
X		X		X	0.69
X			X	X	0.6786
	X	X		X	0.6864
	X		X	X	0.6907
		X	X	X	0.6883
X	X	X		X	0.6872
X	X		X	X	0.6909
X		X	X	X	0.6895
	X	X	X	X	**0.6927**
X	X	X	X	X	0.6892

Table 5.4: Evaluation results of the extracted features and of the GBM predictive model.

On the other hand, the GBM classification model performed at its best when trained on global statistics acquired from the entire observation window, monthly statistics obtained for each month in the window, vector consisting of the number of purchases done towards the different MCCs and finally vector comprising of the daily purchase amounts. In this scenario, the GBM obtained an AUROC score of 0.69, implying that the constructed GBM distinguishes the 2 classes way better than the implemented ANN model. Furthermore, it can be noticed that the GBM improved its score by a few percentages when more features are observed.

In addition, for the best performing ANN and the best performing GBM, we compute other performance metrics using traditional values from the confusion matrix so as to have a better understanding of the models' prediction performances. In fact, we compute the Sensitivity, Specificity, False Positive Rate, False Negative Rate and the Precision metric. The confusion matrices and the obtained measurements are shown in Tables 5.5, 5.6, 5.7 and 5.8 for the ANN and GBM respectively.

		Actual	
		Churn	**Non-Churn**
Predicted	**Churn**	1118.0	780.6
	Non-Churn	1479.0	2227.4

Table 5.5: Confusion Matrix of the ANN predictive model.

Metric	**Score**
Sensitivity	0.4305
Specificity	0.7405
False Positive Rate	0.2595
False Negative Rate	0.5695
Precision	0.6211

Table 5.6: Evaluation results of the ANN predictive model based on the Confusion Matrix.

		Actual	
		Churn	**Non-Churn**
Predicted	**Churn**	1815.0	943.0
	Non-Churn	782.0	2065.0

Table 5.7: Confusion Matrix of the GBM predictive model.

From these metrics, one can conclude that the GBM with 70% sensitivity, is more capable of predicting positive cases i.e. "Churners", whilst on the other hand ANN with 74% specificity, is more capable of identifying negative cases i.e. "Non Churners". Furthermore, GBM is quite consistent and manages to incorrectly classify both positive and negative cases around 30% of the time. On the other hand, ANN incorrectly classifies positive cases as negative around 57% of the time. With regards to how many predicted

Metric	Score
Sensitivity	0.6989
Specificity	0.6865
False Positive Rate	0.3135
False Negative Rate	0.3011
Precision	0.6581

Table 5.8: Evaluation results of the GBM predictive model based on the Confusion Matrix.

positive cases were actually correct, the GBM edges the ANN model with a 3.7% better precision. To conclude this section, it is fair to say that the implemented GBM is more suitable to predict the churn behaviour of customers.

In view of this, the constructed GBM model and all the extracted features bar customer demographics, will be utilised in the remaining experiments.

5.4 | Varying Observation Window Sizes

In this section, we examine different observation window sizes and check how the prediction performance of the classification model changes in return, so as to determine the minimum amount of customer purchase history required and still predict churn with the same performance. The results of such experiment are shown in Table 5.9.

Number of Months	AUROC Score
1	0.6674
2	0.6883
3	0.6927
4	0.6865
5	0.6801
6	0.6850

Table 5.9: Evaluation results of the different observation window sizes.

Despite varying the observation window size from 1 month up to 6 months, the performance of the predictive model does not change much however. The 3-month observation window remained with the best performance metric score, with the other window sizes not managing to cap that. Despite obtaining the lowest AUROC score (0.67), the 1-month observation window is approximately only 2.5% off the top. It can be said that decreasing the amount of purchase history observed does not reflect a huge loss in predictive performance.

These results hint that the current machine learning framework might be over complicated for the problem and the data we are trying to model in this experiment. This behaviour, known as a "high variance" situation, tends to lead to over-fitting. In this experiment, as we were increasing the observation window size, we were adding more features to the vector comprising of the daily purchase amounts. As a result, the model was getting complex with more features, without improving its performance. This problem can be tackled by applying data-dimensionality reduction techniques including Principal Component Analysis (PCA) or feature elimination techniques.

5.5 | The Cold Start Problem

In this section, we conduct a time-series experiment where we examine how effective the classification model is when predicting churn on new customers. The results of such experiment are shown in Table 5.10.

Number of Weeks	AUROC Score
0	0.5026
1	0.5746
2	0.6175
3	0.6524
4	0.6828

Table 5.10: Evaluation results of the usage of demographics combined with the initial purchase observations for churn prediction on new users.

Results show that knowing just the age, country and currency information of a customer, is not enough to be able to predict whether a newly registered user is willing to continue using the company's services or rather stops and defaults. Predicting customer churn with only demographic data is as effective as tossing a coin. It is worth noting that with a few weeks of purchase data, the prediction performance increases quite rapidly, reaching the levels of having a 3-month observation of purchases. Both this experiment and the one preceding it, have shown that a month's worth of data is still quite sufficient to predict whether a customer is defaulting in the next month. We can conclude that by only observing the purchase data of the current month, we can infer churn predictions for the following month.

5.6 | Summary

In this chapter, we described how the proposed solution of this dissertation was evaluated. We first gave an overview of the evaluation approach, generally seen in various related works found in literature. Thereafter, we discussed the experiments performed on each one of our objectives and the obtained results. We have seen that the GBM predictive model is more suitable than an ANN in our CRM problem, since it is capable of identifying the majority (70%) of churners correctly whilst the constructed ANN is more capable of classifying the non-churners. Furthermore, we have shown that reducing the observation window size does not reflect in huge performance loss, thus giving the ability of leveraging prediction performance with the amount of data observed. Finally, we have seen that the demographic information present in our dataset is nowhere near effective to predict churn behaviour of customers.

6

Conclusion

In this dissertation, we presented a data mining study that was applied on millions of financial transactions collected for a number of years, by a leading virtual credit card company based in Malta. In this study, the Artificial Neural Network (ANN) and the Gradient Boosting Model (GBM) machine learning techniques were both analysed so as to identify the best framework that accurately models the customer churn behaviour of customers. Furthermore, quite a number of different features have been extracted from the raw financial transactions. These include demographic features, "global" statistics that are relative to the entire observation window, dynamic monthly statistics for each month in the observation window, a vector containing daily purchase amounts and another vector containing the number of purchases done towards each MCC. Evaluation experiments were done to determine i) the feature set that best captures customer behaviour for churn prediction, and ii) the classification model that is effective in predicting the churn behaviour of customers. In view of this, the GBM classifier applied on all the extracted features mentioned above except for demographic data, resulted in the best machine learning framework of this study, obtaining an AUROC score of 0.6927. In addition, we also observed that our learning framework is capable of correctly identifying 70% of "Churners", potentially making it a suitable solution in CRM.

In this dissertation, we also investigated different sizes of the observation window by experimenting with 1 month to 6 month time-periods. Results show that decreasing the observation window to a month's length does not extensively affect the predictive performance of the classifier giving the ability of leveraging prediction accuracy with the amount of data observed. In fact, we conclude that by observing the purchase behaviour of customers in the current month, one can infer churn predictions for the following month. In addition, we attempted to investigate and possibly overcome the "cold start" problem within machine learning, by performing a time-series experiment starting from the initial customer acquisition until a month of purchase history is observed. In this experiment, we start predicting customer churn using only demographic information and in time, combine any new purchase data. This experiment showed that for the current dataset, predicting churn behaviour using only customer demographics is similar to tossing a coin.

6.1 | Achieved Aims and Objectives

The 4 objectives outlined in Section 1.3 have all been fulfilled within this study.

In this dissertation, we have extracted 2 categories of dynamic features, namely the macro-average monthly statistics that aggregates purchase behavioral information per month of the observation window, and the daily purchase amounts vector. After evaluating all of the extracted features through a greedy search, we have proven that dynamic behavioural predictors tend to be more effective in representing customer behaviour for churn prediction (Hsu et al., 2019; Kaya et al., 2018; Leung and Chung, 2020). This is because our best performing predictive setup was applied and hence require both of our novel dynamic feature categories.

Furthermore, in this dissertation we have constructed a machine learning framework that is capable of predicting customer churn behaviour from raw financial transactions. As already mentioned, this framework is capable of correctly identifying 70% of "Churners". Although in research one can find related works that predicts churn with a better performance score than our 0.6927 AUROC score, the sensitivity of our framework should not go unnoticed considering we are predicting churn behaviour of over 50,000 different customers.

Moreover, when evaluating our solution, we experimented with different time-periods for the observation window. As already mentioned, results were surely promising, showcasing that a decrease in the observation window barely affects the prediction performance of our framework. In fact, we state that a month's worth of data is enough to predict whether customers are defaulting in the following month. Despite of this, we still recommend to attempt this experiment again and apply data-dimensionality reduction techniques to prevent cases of over-fitting or generating too complex models with a large number of features.

Finally, despite failing to overcome the "cold start" problem within a machine learning setup, we succeeded in attempting to examine such issue on the dataset we were provided. We have seen how the knowledge of just the age, country and currency information of a customer, is not anywhere sufficient enough to be able to predict whether a newly registered customer is going to default or not in the coming month or so. An approach that can be attempted to overcome this problem, is to directly prompt the user for information upon registration. The information given by the actual customer can then be fed and used within the predictive models.

Nevertheless, as mentioned in Section 1.4, by achieving the set objectives we contributed to literature by exploring the novel approach of extracting dynamic features from raw financial transactions to model customer churn behaviour and also by identifying the appropriate and sufficient observation window size through a computed experiment.

6.2 | Limitations

In this study, we have worked with a dataset of financial transactions spanning just over 2 years. Ideally, we would have had a longer amount of data coverage to be able to observe more customers and be able to identify and model different behavioural patterns.

Furthermore, experiments involving machine learning techniques took a considerable amount of time due to the large amount of processing power they require. In the future, we might possibly be opting for a more powerful architecture that is hosted on the cloud. In the past years, Azure, AWS as well as Google Cloud have significantly invested in both AI and machine learning, and are now offering various cloud services that can be applicable to our line of work.

6.3 | Future Work

In some of the related work reviewed in this dissertation, we have seen how certain machine learning frameworks are capable of extracting meaningful rules which tend to group customers into different segments according to common behaviour. Such information is beneficial to the marketing department and managerial personnel of companies as they have knowledge on the actual characteristics of customers that are willing to churn. The work performed in this dissertation can be further improved by augmenting the constructed framework to a tree-based model in order to extract meaningful behavioural rules.

Furthermore, after addressing the problem of customer churn prediction, it now makes sense to tackle the problem of predicting the next purchases of customers. The approaches performed in collaborative recommendation systems can be adopted and tweaked to our purpose.

Hereafter, our work can be enriched by providing a dashboard that makes use of the results obtained in our study, which in turn illustrates the different customer segments as well as the modelled churn and purchase behvaiour of customers. In addition, graphical plots on the data are also to be included in this dashboard to show different behavioural trends of customers. For instance, transactions can be aggregated according to merchant to identify those merchants that a particular customer transacts to the most. The day and time that a customer usually transacts at can also be observed through the appropriate graphical visualisations. Moreover, knowing the country where a transaction is originating can also generate interesting observations. Finally, this can be linked with real-world events happening at the same time of the analysis, to derive any consequences that may affect a customer's spending patterns.

6.4 | Final Remarks

In this dissertation we have constructed a machine learning framework that given millions of raw financial transactions, it is capable of modelling customer behaviour and predict churn activities. From the evaluation conducted in this dissertation, we have shown that although there is room for improvement, our system succeeds in achieving the objectives set for this study.

References

Badillo, Solveig, Banfai, Balazs, Birzele, Fabian, Davydov, Iakov I, Hutchinson, Lucy, Kam-Thong, Tony, Siebourg-Polster, Juliane, Steiert, Bernhard, and Zhang, Jitao David. An introduction to machine learning. *Clinical Pharmacology & Therapeutics*, 107(4):871–885, 2020.

Bilal Zorić, Alisa. Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex Systems: INDECS*, 14(2):116–124, 2016.

Cil, Fatih, Cetinyokus, Tahsin, and Gokcen, Hadi. Knowledge discovery on investment fund transaction histories and socio-demographic characteristics for customer churn. *International Journal of Intelligent Systems and Applications in Engineering*, 6(4):262–270, 2018.

Farquad, Mohammed Abdul Haque, Ravi, Vadlamani, and Raju, S Bapi. Churn prediction using comprehensible support vector machine: An analytical crm application. *Applied Soft Computing*, 19:31–40, 2014.

Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Friedman, Jerome H. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

Holzinger, Andreas. Introduction to machine learning & knowledge extraction (make). *Machine learning and knowledge extraction*, 1(1):1–20, 2019.

Hsu, Te-Cheng, Liou, Shing-Tzuo, Wang, Yun-Ping, Huang, Yung-Shun, et al. Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1572–1576. IEEE, 2019.

Kaya, Erdem, Dong, Xiaowen, Suhara, Yoshihiko, Balcisoy, Selim, Bozkaya, Burcin, et al. Behavioral attributes and financial churn prediction. *EPJ Data Science*, 7(1):41, 2018.

Keramati, Abbas, Ghaneei, Hajar, and Mirmohammadi, Seyed Mohammad. Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 2(1):10, 2016.

Kim, Sun, Shin, Kyung-shik, and Park, Kyungdo. An application of support vector machines for customer churn analysis: Credit card case. In *International Conference on Natural Computation*, pages 636–647. Springer, 2005.

Kröse, Ben, Krose, Ben, van der Smagt, Patrick, and Smagt, Patrick. An introduction to neural networks. 1993.

Leung, Hoiyin Christina and Chung, Wingyan. A dynamic classification approach to churn prediction in banking industry. 2020.

Martens, David, Provost, Foster, Clark, Jessica, and de Fortuny, Enric Junqué. Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly*, 40(4), 2016.

Mehlig, Bernhard. Artificial neural networks. *arXiv preprint arXiv:1901.05639*, 2019.

Padilla, Nicolas and Ascarza, Eva. Overcoming the cold start problem of crm using a probabilistic machine learning approach. 2020.

Rosa, Nelson Belém da Costa. *Gauging and foreseeing customer churn in the banking industry: a neural network approach*. PhD thesis, 2019.

Safinejad, Fatemeh, Noughabi, Elham Akhond Zadeh, and Far, Behrouz H. A fuzzy dynamic model for customer churn prediction in retail banking industry. In *Applications of Data Management and Analysis*, pages 85–101. Springer, 2018.

Shirazi, Farid and Mohammadi, Mahbobeh. A big data analytics model for customer churn prediction in the retiree segment. *International Journal of Information Management*, 48:238–253, 2019.

Szmydt, Marcin. Predicting customer churn in electronic banking. In *International Conference on Business Information Systems*, pages 687–696. Springer, 2018.