Customers Churn Management: A Machine Learning Perspective – at Vodafone Panafon S.A. –

Thesis & Final Intership Report

B*Sc* Thesis
Chryssoula-Maria Nampouri
Department of Management Science and Technology
Athens University of Economics and Business
Athens, Greece
t8150096@aueb.gr

Academic Advisor: Prof. Emmanoyil Zaxariadis
Company Supervisor: Georgina Bilali
Athens, February 2020

Abstract

Within a saturated market, customer acquisition no longer ensures sustainable revenue. The paradigm has moved towards Customer Retention. Churn Management is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues, companies have realized that waiting for a cancellation request to act, is not an efficient strategy. Instead, they should take into account the risk of churn and step in before customers make that decision. The most promising way is through Machine Learning, considering the Customer Churn problem as a predictive classification problem of churners and non-churners.

This thesis aims to provide a literature review on Machine Learning approach, the widespread methods and techniques for its application to classification problems, with an emphasis on the Churn Management. More specifically, the research questions that this thesis attempts to answer are the following:

1. *What are the main Machine Learning approaches for classification problems?*
2. *How Machine Learning powered churn analysis for modern day businesses?*

In addition, a particular methodology of Churn Model is proposed and implemented in the context of an internship. The research was based on a data set containing 175,513 customers, of whom 13% were churners. Several models implemented in order to predict those churners. However, the best results were obtained by applying a Voting Classifier of LightGBM and Random Forest.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Prof. Emmanoyil Zahariadis, and to my supervisor, Georgina Bilali, for their constant guidance, advice and support during my internship, and throughout the implementation of the particular thesis. Their formidable knowledge, advice and motivation have contributed not only to the fulfilment of my objectives in the context of my internship, but also to the wider advancement of my skills and personality.

— Athens, February 2020

Contents

List of Figures

List of Tables

Part I: Thesis

Title:

Customers Churn Management: A Machine Learning Perspective

Chapter 1

Introduction

In the dynamic and competitive market environment, where customers are free to choose from plenty of providers, even a bad experience and the customer may quit.

Customer Relationship Management (CRM) is a comprehensive strategy for building, managing and strengthening loyal and long-lasting customer relationships [Xia and Jin, 2008]. It is broadly acknowledged and extensively applied to different fields, e.g., telecommunications, banking and insurance, retail market, etc. Within the context of CRM, it is a common knowledge that the longer the customer stays with a company, the longer the profit can be made out of them. Hence, retaining its existing customer or preventing them from leaving or switching service providers is one of the key areas in CRM.

Losing customers not only leads to a direct loss of revenue, but also leads to an increased need of attracting new customers, which includes advertisement, promotion, offers, effort to know customers' needs and time to build sustainable relationships. [Athanassopoulos, 2000]. According to surveys[1] [2], the cost of acquisition of a new customer is estimated to be ranging from $ 300 to $ 600, and it costs roughly five times as much to sign on a new customer as to retain an existing one [Mozer et al., 2000] . Yet, in some cases, it may be 20 times more expensive [Vafeiadis et al., 2015].

Therefore, the ability to identify customers at risk of churn, while there is still enough time for action, gives a huge competitive advantage to every manager in order to take corresponding actions early and re-engage them. To predict the latent churn customers, several studies have focused on *Machine Learning*.

[1] https://evolvepg.com/About/Whats-Evolving/ArticleID/28//A-Customer-Saved-Is-Worth-a-Customer-Earned%E2%80%A6Times-5

[2] https://www.forbes.com/sites/alexlawrence/2012/11/01/\/five-customer-retention-tips-for-entrepreneurs/#1da2c82e5e8d

Machine Learning is the science that is "concerned with the question of how to construct computer programs that automatically improve with experience" (Mitchell, 1997) [3]. The basic premise of Machine Learning is to build algorithms that can receive input data and through statistical analysis discover a pattern to them and finally predict an output that characterize them. There are several applications for Machine Learning (ML), the most significant of which is Predictive Data Mining that explores large amounts of data - also known as "Big Data".

Big Data is no fad. The world is growing at an exponential rate and so is the size of the data collected across the globe. Data is becoming more meaningful and contextually relevant, breaking new grounds for Machine Learning (ML), in particular for deep learning (DL) and artificial intelligence (AI), moving them out of research labs into production [Jordan and Mitchell, 2015]. The problem has shifted from collecting massive amounts of data to understanding it-turning it into knowledge, conclusions, and actions. Multiple research disciplines, from cognitive sciences to biology, finance, physics, and social sciences, as well as many companies believe that data-driven and "intelligent" solutions are necessary to solve many of their key problems [Kersting, 2018]. A recent report from the McKinsey Global Institute asserts that Big Data and Machine Learning (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation [Manyika et al., 2011].

Based on the type of data given and the purpose of the problem, four main approaches of ML systems are recognised; **supervised learning**, **unsupervised learning**, **semi-supervised learning** and **reinforcement learning** [Emerson et al., 2019, Mohri et al., 2018]:

1. **Supervised Learning:** A set of labelled examples is received as input data (training set) and predictions are made for all unseen data (test set). The algorithm infers a function linking each set of inputs with the expected or labelled output in the process of training.

2. **Unsupervised Learning**: In contrast to supervised learning, unsupervised learning systems handle unlabelled data and therefore try to find structures within the data by creating classes on their own.

3. **Semi-Supervised Learning:** A Hybrid or Semi-Supervised Learning system, combines supervised and unsupervised learning, using both labelled and unlabelled data to train models. This is useful, when there is limited data or the process of labelling data could introduce biases.

4. **Reinforcement Learning:** A Reinforcement Learning algorithm or agent learns by interacting with its environment. The agent receives rewards for performing correctly and penalties for performing incorrectly. It is a type of dynamic programming that learns by trial and error, based on feedback from past experiences. Like unsupervised learning, it does not require labelled data.

[3] https://www.frontiersin.org/articles/10.3389/fdata.2018.00006/full#h2

Since the proposed task is a classification problem, concerning the possible classes of customers (churners and non-churners), this thesis will focus on supervised learning methods and more specifically on classification problems, in which the output of instances admits only discrete, unordered values. Particularly, I will attempt to make a literature review on some of the most well established and popular Machine Learning algorithms for classification problems, taking into consideration reliability, efficiency and popularity in the research community. The final goal is to suggest the ways that Machine Learning can be applied to the Customer Churn Management. Insights will be given into the crucial factors that can drive a successful retention campaign and how these variables can be tracked and effectively managed through Machine Learning. More specifically, the research questions that this thesis attempts to answer are the following:

1. *What are the main Machine Learning approaches for classification problems?*

2. *How Machine Learning powered churn analysis for modern day businesses?*

The rest of the thesis is structured as follows. **Chapter 2** is dedicated to the background of Supervised Machine Learning. A literature review is provided to get acquainted with the most major topics. **Chapter 3** introduces Customer Churn Management review. The main objective is to fully understand the problem of Customer Churn and determine the factors that Machine Learning seems to be an amazingly promising solution. Last but not least, **Chapter 4** aims to summarize and discuss the findings of the bibliographic overview and list the limitations as well.

All the above literature review is a theoretical overview of Machine Learning technique and its application in Customer Churn Management and is implemented into practise in Part II of this artefact in the context of an internship. The result is a step-by-step implementation and evaluation of a Churn Model using Machine Learning Algorithms.

Chapter 2

Background of Supervised Machine Learning

A major difference between humans and computers has been for a long time that human beings tend to automatically improve their way of tackling a problem. Humans learn from previous mistakes and try to solve them by correcting them or looking for new approaches to address the problem. Traditional computer programs do not look at the outcome of their tasks and are therefore unable to improve their behavior. The field of Machine Learning addresses this exact problem and involves the creation of computer programs that are able to learn and therefore improve their performances by gathering more data and experience [Lucket and Schaefer-Kehnert, 2016].

This section gives an overview of the relevant theoretical foundations and answers the following questions:

1. *What are the most promising algorithms in Machine Learning? When to use each algorithm?*
   This is answered in Section 2.1.

2. *How to measure the performance of a Machine Learning system?*
   This is answered in Section 2.2.

3. *What are the main issues of Machine Learning systems?*
   This is answered in Section 2.3.

To answer these questions, the first section deals with the most promising algorithms for supervised learning, with emphasis on classification problems. Afterwards, the next section is dedicated to the evaluation of a Machine Learning system, presenting some of the most accurate evaluation metrics commonly used for such problems. Finally, the chapter is closed by addressing some possible issues of Machine Learning implementation.

2.1 Machine Learning Techniques in Classification Problems

2.1.1 Logistic Regression

Logistic Regression (LR) is the most widely used probabilistic statistical classification model in many fields for binary data (0-1 response) prediction. It takes linear combination of features - i.e.,

$$z = w_o + w_1 x_1 + w_2 x_2 + ... + w_n x_n$$

and applies non-linear function to it. The most commonly used non-linear function for binary classification problems is sigmoid $\sigma(z)$ and the output is ginen by the following equation:

$$\sigma(z) = \frac{1}{1 - e^{-z}} \tag{2.1}$$

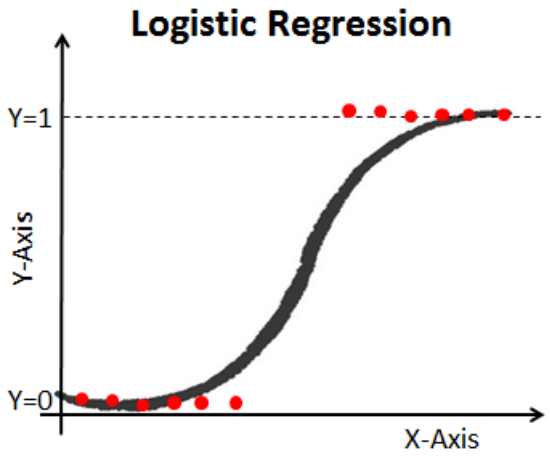


Figure 2-1: Logistic Regression for Binary Classification [Navlani, 2020c].

Logistic regression provides lots of ways to regularize your model and you do not have to worry as much about your features being correlated, like you do in Naive Bayes. It has been widely applied due to its simplicity, efficiency, easy interpretation and usage of limited computational resources. What is more, logistic regression allows easy model updating using stochastic gradient descent, unlike Decision Trees [sec: 2.1.4] or SVMs [sec: 2.1.3] [Sharmistha, 2019, Harlalka, 2018].

Mostly LR performs well when the relationship in the data is approximately linear. However, it performs poorly if complex non-linear relationships exist between the variables. In addition, it requires more statistical assumptions before being applied than other techniques. Also, the prediction rate gets affected, if there are missing data in the data set [Vafeiadis et al., 2015].

2.1.2 Instance-Based Algorithms

Instance-based learning algorithms are lazy-learning algorithms (Mitchell 1997), as they delay the induction or generalization process until classification is performed. Lazy-learning algorithms require less computation time during the training phase than eager-learning algorithms (such as decision trees, neural and Bayes nets) but more computation time during the classification process. One of the most straightforward instance-based learning algorithms is the nearest neighbour algorithm.

k-Nearest Neighbour (kNN) is based on the principle that the instances within a data set will generally exist in close proximity to other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbours. The kNN locates the k nearest instances to the query instance and determines its class by identifying the single most frequent class label.

Consider the Figure 3-2 below. If k is set to three, then the query instance will be assigned to the green class. On the other hand, if k is set to seven, then the instance will be assigned to the red class.

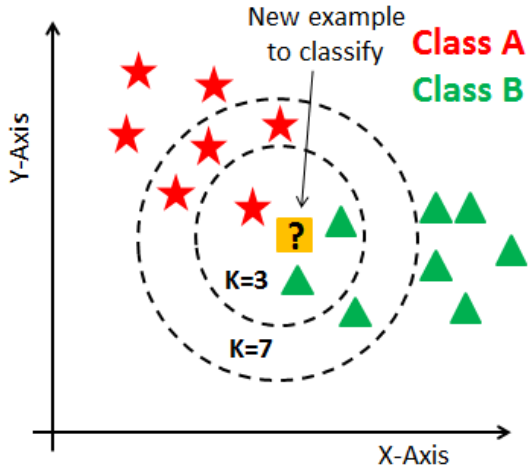


Figure 2-2: k-Nearest Neighbours Classifier[Navlani, 2020a].

The choice of k affects the performance of the kNN algorithm. Consider the following reasons why a k-nearest neighbour classifier might incorrectly classify a query instance:

- When noise is present in the locality of the query instance, the noisy instance(s) win the majority vote, resulting in the incorrect class being predicted. A larger k could solve this problem.
- When the region defining the class, or fragment of the class, is so small that instances belonging to the class that surrounds the fragment win the majority vote. A smaller k could solve this problem.

The power of kNN has been demonstrated in a number of real domains, but there are some reservations about the usefulness of kNN, such as: (i) they have large storage requirements, (ii) they are sensitive to the choice of the similarity function that is used to compare instances, (iii) they lack a principled way to choose k, except through cross-validation or similar, computationally-expensive technique [Guo et al., 2003].

2.1.3 Support Vector Machines

Support Vector Machines (SVMs) is a Machine Learning technique based on structural risk minimization. The basic approach to classify the data, starts by trying to create a function that splits the data points into the corresponding labels with (a) the least possible amount of errors or (b) with the largest possible margin. This is due to the fact that larger empty areas next to the splitting function result in fewer errors, because the labels are better distinguished from one another.

In many instances, a data set may be very well separable by multiple functions without any errors. Therefore, the margin around a separating function is being used as an additional parameter to evaluate the quality of the separation. Margin is defined as the minimum distance between the decision boundary and every data point. SVMs choose that function that maximizes the margin [Bishop, 1995]. The Figure 3-1 demonstrates this exact rule. In this case the separation with the black line is the better one, since it distinguishes the two classes in a more precise manner.

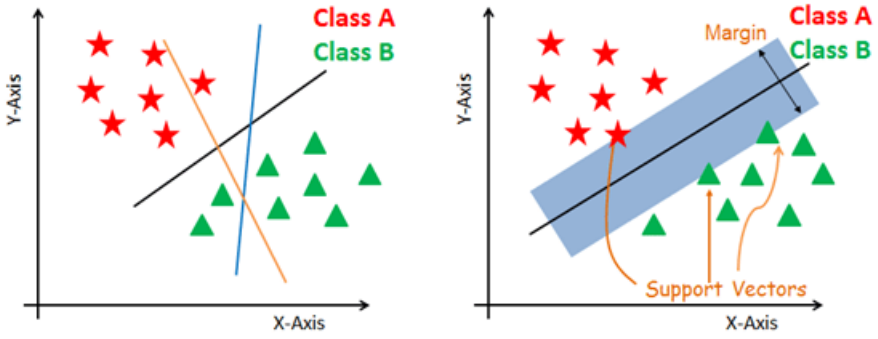


Figure 2-3: Visualization of a Support Vector Machine splitting a data set into two classes, by using three different linear separations resulting in differently sized margins around the splitting functions [Navlani, 2020b].

2.1.4 Decision Trees

Decision Trees (DT) are models that describe the conditional distribution of label y given x features (predictors). More specifically, are tree-shaped structures representing sets of decisions capable to generate classification rules for a specific dataset, or as Berry and Linoff noted "a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules". In these tree structures, leaves (terminal nodes or external nodes) represent class labels and branches (internal nodes) represent conjunctions of features that lead to those class labels. More descriptive names for such tree models are Classification Trees if the response is discrete or Regression Trees if the response is continuous.

The CART model subdivides the predictor space [$x = (x_1, x_2, ..., x_k)$] as follows. Each internal node has an associated splitting rule that uses a predictor to assign observations to either its left or right child nodes. The terminal nodes thus identify a partition of the observation space according to the subdivision defined by the splitting rules. For quantitative predictors, the splitting rule is based on a split value s and assigns observations for which $\{x_i \leq s\}$ or $\{x_i > s\}$ to the left or the right child node. For qualitative predictors, the splitting rule is based on a category subset C, and assigns observations for which $\{x_i \in C\}$ or $\{x_i \notin C\}$ to the left or the right child node. [Bayesian CART Model Search]. The feature that best divides the training data would be the root node of the tree. There are numerous methods for finding the feature that best divides the training data but a majority of studies have concluded that there is no single best method [Kotsiantis et al., 2006].

One of the most useful characteristics of decision trees is their comprehensibility. People can easily understand why a decision tree classifies an instance as belonging to a specific class. CART is, also, flexible in practice in the sense that it can easily model non-linear or non-smooth relationships. It has the ability of interpreting interactions among predictors. It also has great interpretability due to its binary structure and it is often effective in high-dimensional data. However, CART has several drawbacks such as it tends to overfit the data. In addition, since one big tree is grown, it is hard to account for additive effects. Another disadvantage is that it does not support online learning, so you have to rebuild your tree when new examples come on [Harlalka, 2018, Vafeiadis et al., 2015].

2.1.5 Artificial Neural Networks

Artificial Neural Networks (ANN) is a very popular approach to address complex problems with high dimensional and uncorrelated data. ANNs have become a key technology in the development of ML. They were first proposed over 75 years ago, inspired by the workings of the human brain. The brain receives the stimulus from the outside world, does the processing on the input, and then generates the output. As the task gets complicated, multiple neurons form a complex network, passing information among themselves. An Artificial Neural Network tries to mimic a similar behavior.

ANN is a complex non-linear function obtained by hierarchically synthesizing very simple non-linear functions, such as the logistic/sigmoid function. Such a simple non linear function has little learning capacity. It is essentially Logistic Regression, where we can only learn linear decision boundaries. In neural networks we combine multiple simple functions, so that overall construct very complex non-linear functions. **So, the increased flexibility comes from the combination!** In the terminology of neural networks the non-linear function $h(\cdot)$ is called activation function.

A neural network in the simplest case consists of three layers:

- ***Input Layer:*** Input data vector
- ***Hidden Layer:*** An intermediate layer
- ***Output Layer:*** The output function values

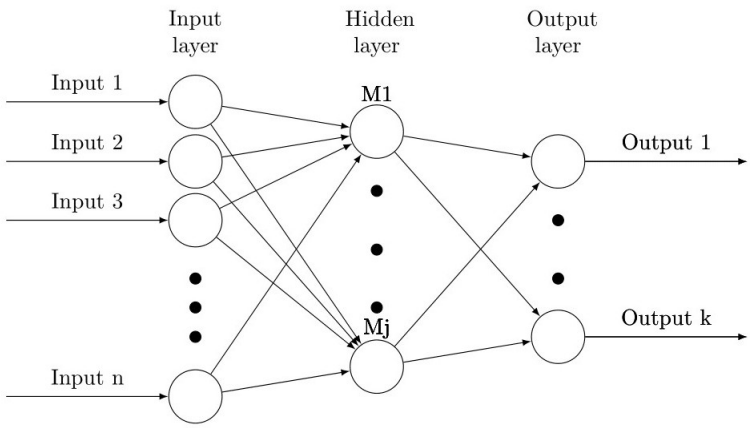


Figure 2-4: Example of an Artificial Neural Network with one Hidden Layer

In the above case the data flow is the following: The input data is fed to the input layer, the neurons (nodes) perform a linear transformation on this input using weights and biases. Post that, an activation function $h(\cdot)$ is applied on the above result (until this step, it is exactly a Logistic Regression process). Finally, the output from the activation function moves to the next hidden layer and the same process is repeated, until the output layer, in which the output is converted into probabilities. This forward movement of information is known as the forward propagation.

And what if the output generated is far away from the actual value? Using the output from the forward propagation, error is calculated. Based on this error value, the weights and biases of the interconnections are readjusted. This process is known as back-propagation. After the update of the weights, the forward propagation starts from the beginning, until there is convergence.

Obviously, the most crusial part of neural networks are the activation functions that transform linear combination of features into non-linear. Until now, only sigmoid has been discussed, but there are several activation functions for the hidden layers [Gupta, 2017,Serengil, 2017,Walia, 2017], some of the most popular are presented below:

- ***Linear Function***:

$$h(z) = z \tag{2.2}$$

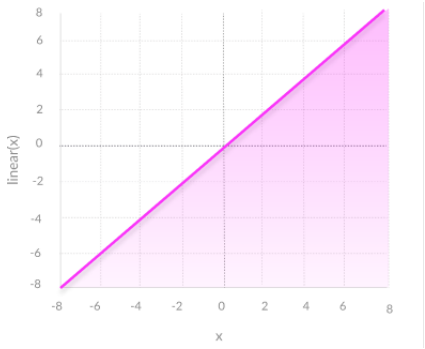


Figure 2-5: Linear Activation Function

– Suitable only for the output layer in case of regression.

- ***Hyperbolic Tangent Function (tanh)***:

$$\mathrm{h(z)} = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{2.3}$$

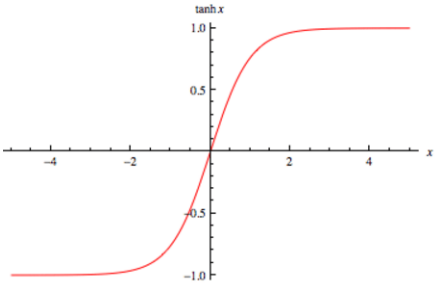


Figure 2-6: Hyperbolic Tangent Activation Function

– Takes values between -1 and 1; takes both negative and positive values; strictly inceasing.

- ***ReLU Function***:

$$h(z) = max(0, z) \tag{2.4}$$

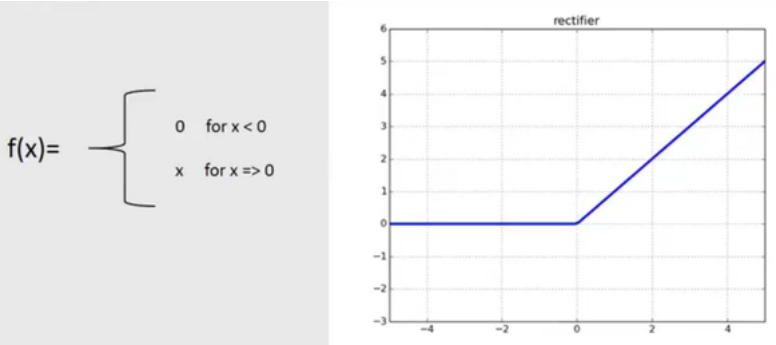


Figure 2-7: ReLU Function

– Takes values from 0 up to plus infinity; strictly inceasing.

- ***SoftPlus Function***:

$$h(z) = \ln(1 + e^z) \tag{2.5}$$

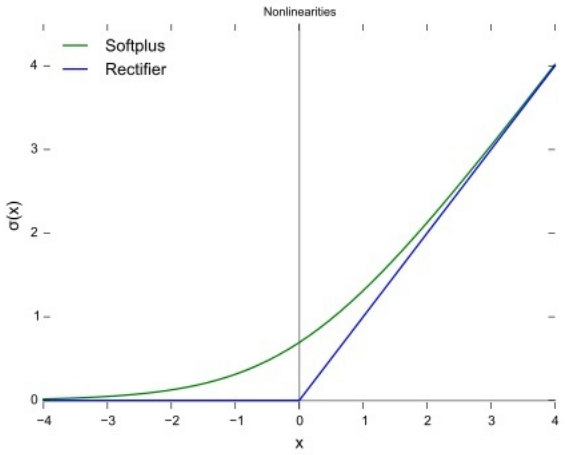


Figure 2-8: SoftPlus Activation Function

– Takes values from 0 up to plus infinity;

ReLU and Softplus are largely similar, except near 0 where the softplus is enticingly smooth and differentiable. It is much easier and efficient to compute ReLU than softplus function which has $\ln(\cdot)$ and $exp(\cdot)$ in its formulation.

But what kind of activation functions do we need for the output layer?

The functional form of the final activation function for each output depends on what kind of output data we wish to model/predict. If in the k^{th} output we wish to predict real values (regression), then the corresponding function will be linear. If instead in the k^{th} output we wish to predict binary {0, 1} values, then we could use the sigmoid. If we are interested in classification with multiple classes we can use the ***Softmax:***

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \qquad for\ j = 1, 2, ..., K \tag{2.6}$$

– where z is the linear combination of features, as we have already seen in Logistic Regression and k are the possible classes. So, the above equation calculates the probability of a data point belonging to each class label.

Furthermore, Artificial Neural Networks can be divided into three main categories:

- ***Multilayer Perceptrons (MLPs):*** All networks that do not gain feedback from the network itself. This means that the input data flows in one direction, from the input nodes through 0 to n hidden nodes to the output nodes. There is no information given backwards to readapt the system. MLPs are suitable for classification prediction problems, where inputs are assigned a class or label. They are also, suitable for regression prediction problems, where a real-valued quantity is predicted given a set of inputs. Data is often provided in a tabular format, such as you would see in a CSV file or a spreadsheet.

- ***Convolutional Neural Networks (CNNs):*** CNNs, were designed to map image data to an output variable. They have proven so effective that they are the go-to method for any type of prediction problem involving image data as an input. The benefit of using CNNs is their ability to develop an internal representation of a two-dimensional image. This allows the model to learn position and scale in variant structures in the data, which is important when working with images. Generally, CNNs work well with any kind of data that have spatial relationships.

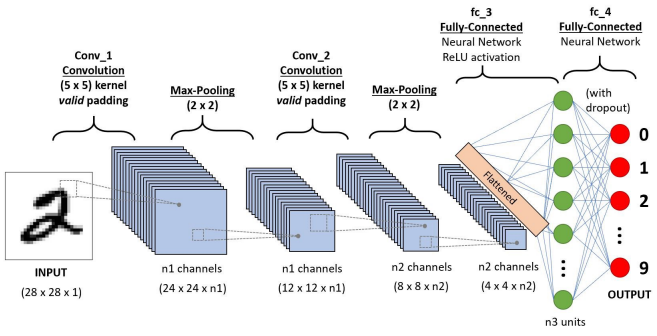


Figure 2-9: Example of Convolutional Neural Network

- ***Recurrent Neural Networks (RNNs):*** All networks that contain a feedback option and therefore are able to reuse data from later stages, for the learning process, in earlier stages. More specifically all RNNs have feedback loops in the recurrent layer. This lets them maintain information in "memory" over time. RNNs have received the most success when working with sequences of words and paragraphs, generally called Natural Language Processing (NLP). This includes both sequences of text and sequences of spoken language represented as time series. They are also, used as generative models that require a sequence output, not only with text, but on applications such as generating handwriting.

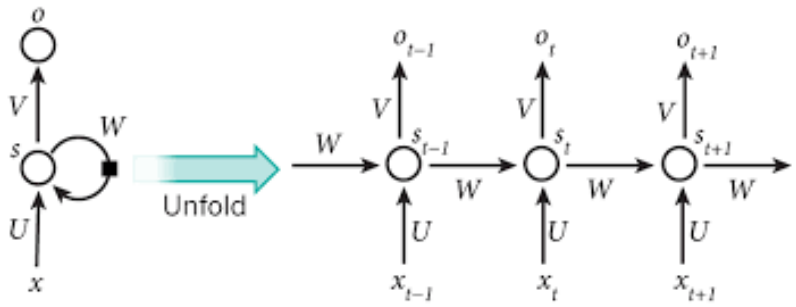


Figure 2-10: Example of Recurrent Neural Network

RNNs can be very difficult to train for problems that require learning long-term temporal dependencies because of vanishing and noisy gradients [SuperDataScience, 2018, Hochreiter, 1998]. In such cases, the **Long Short-Term Memory, or LSTM Network** is perhaps the most successful RNN because it overcomes the problems of training a Recurrent Network and in turn has been used on a wide range of applications. LSTM Networks use special units in addition to standard units. These units include a "memory cell" that can maintain information in memory for long periods of time. A set of gates is used to control when information enters the memory, when it is output, and when it is forgotten. This architecture let them learn longer-term dependencies.

To sum up, MLP network is suggested for classification and regression problems that use tabular data sets, CNN works well with images and RNN-LSTM have received the most success when working with text and speech data [Brownlee, 2018].

2.1.6 Ensemble Learning

Ensemble methods are meta-algorithms that combine several Machine Learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). Most ensemble methods use a single base learning algorithm to produce homogeneous base learners, i.e. learners of the same type, leading to homogeneous ensembles [Smolyakov, 2017]. There are also some methods that use heterogeneous learners, i.e. learners of different types, leading to heterogeneous ensembles. In order for ensemble methods to be more accurate than any of its individual members, the base learners have to be as accurate as possible and as diverse as possible.

In this subsection, we will look at a few basic ensemble techniques, as well as more advanced ones:

Basic Ensemble Techniques:

- The **max voting method** is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a "vote". The predictions which we get from the majority of the models are used as the final prediction.

- Similar to the max voting technique, **in averaging method** we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.

- **Weighted average** is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction.

Bagging methods combine the results of multiple models (for instance, all decision trees) to get a generalized result. If all models are created on the same set of data, then there is a high chance of giving the same result since they are getting the same input. To overcome this problem, bootstrapping technique is used. Bootstrapping is a sampling technique in which we create subsets of observations from the original dataset, with replacement. Bagging (or Bootstrap Aggregating) technique uses these subsets (bags) to get a fair idea of the distribution (complete set).

One of the most popular bagging algorithms is **Random Forest (RF)**. The base estimators in random forest are decision trees. RF randomly selects a set of features which are used to decide the best split at each node of the decision tree. Looking at it step-by-step, this is what a random forest model does:

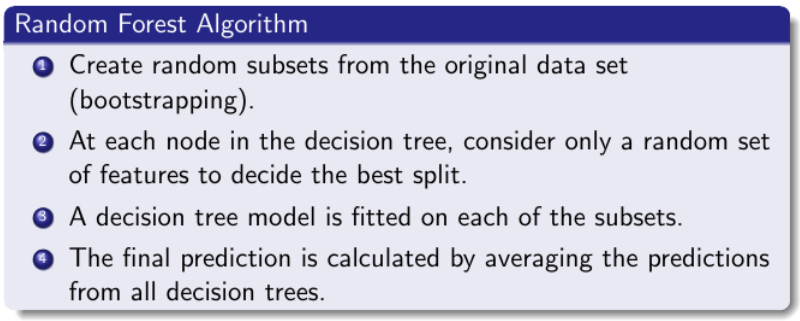


Figure 2-11: Random Forest Algorithm

To sum up, Random forest randomly selects data points and features, and builds multiple trees (Forest).

Boosting method is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. The way boosting works is described in the below steps:

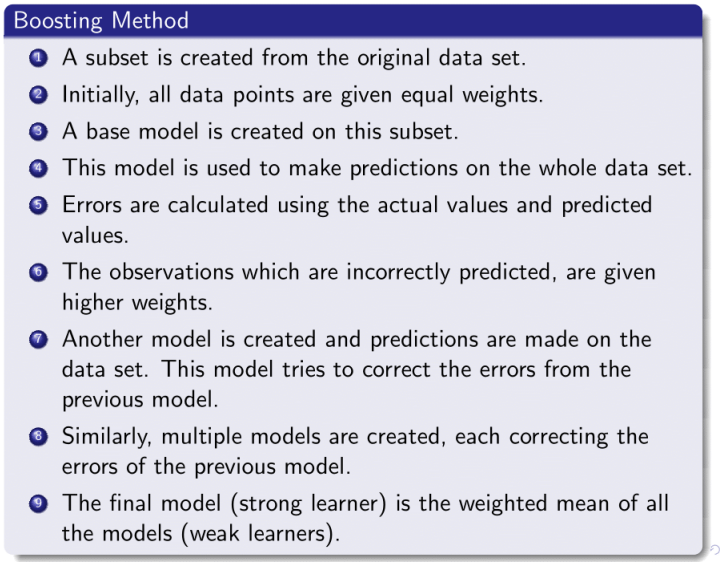


Figure 2-12: Boosting Method

Thus, the boosting algorithm combines a number of weak learners to form a strong learner. The individual models would not perform well on the entire dataset, but they work well for some part of the data set. Thus, each model actually boosts the performance of the ensemble. Two of the most establised boosting methods are **XGBoost** and **LightGBM**:

XGBoost stands for eXtreme Gradient Boosting. XGBoost is a scalable and accurate implementation of gradient boosting machines based on decision trees and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit

of memory and hardware resources for tree boosting algorithms. The scalability of XGBoost is due to several important systems and algorithmic optimizations. These innovations include:

- An in-built capability to handle missing values.
- The power of parallel processing makes learning faster and enables quicker model exploration (XGBoost uses multiple CPU cores to execute the model).
- It follows an effective Tree Pruning technique. A GBM would stop splitting a node when it encounters a negative loss in the split. Thus it is more of a greedy algorithm. XGBoost on the other hand make splits up to the max_depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain.
- More importantly, XGBoost exploits out-of-core computation and enables data scientists to process hundred millions of examples on a desktop.

XGBoost is an algorithm that has been widely recognized in a number of Machine Learning and data mining challenges. Take the challenges hosted by the Machine Learning competition site Kaggle for example. Among the 29 challenge winning solutions published at Kaggle's blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with neural networks in ensembles [Chen and Guestrin, 2016].

LightGBM is based on decision trees, as well as XGBoost, yet it follows a different strategy. Whereas XGBoost uses decision trees to split on a variable and explore different cuts at the variable (the level-wise tree growth strategy), LightGBM concentrates on a split and goes on splitting from there in order to achieve a better fitting (this is leaf-wise tree growth strategy). This allows LightGBM to reach first and fast a good fit of the data, and to generate alternative solutions compared to XGBoost (which is good, if you expect to blend, i.e. average, the two solutions together in order to reduce the variance of the estimated). Algorithmically talking, figuring out as a graph the structures of cuts operated by a decision tree, LightGBM peruses a depth-first search (DFS) [Boschetti and Massaron, 2018].

LightGBM is prefixed as "Light" cause of its high speed. LightGBM can handle large size of data and takes lower memory to run. It has more complex trees due to the leaf-wise strategy leading to a higher accuracy in prediction but also to a higher risk of overfitting; therefore, it is particularly ineffective with small data sets. LGBM can also leverage parallelization and GPU usage and thus data scientists are widely using LGBM for data science application development. Last but not least, it is memory parsimonious because it does not store and handles continuous variables as they are, but it turns them into discrete bins of values (histogram-based algorithms).

2.2 Evaluation of Machine Learning

Another very important part in Machine Learning, is the problem of how a computer program notices which of its results were appropriate and which contained mistakes. Model Evaluation is an integral part of the model development process. It helps to find the best model that represents the data and how well the chosen model will work in the future. Normally, there are two methods of evaluating models in data science, *Hold-Out* and *kfold Cross-Validation* [Khurana, 2019]:

Hold-Out is done by splitting the data set into three subsets, training set, validation set and test set. Training set is a subset of the whole data set used to train the predictive models. Evaluating model's performance with the data used for the training is not acceptable in data science, because it can easily generate overoptimistic and overfitted models. Thus, an independent subset of the data set, called validation set is used to assess the performance of these models by calculating some performance indicators. Validation set provides a test platform for fine tuning model's hyperparameters and selection of the best-performing model. Furthermore, it can be used for regularization by early stopping: stop training when the error on the validation data set increases, as this is a sign of overfitting [Prechelt, 2012]. Finally, the test set or unseen examples is a subset of the data set to assess the likely future performance of a model. If a model fits to the training set much better than it fits the test set, overfitting is probably again the cause. Not all Machine Learning algorithms need a validation set. In that case, the model uses two-thirds for training and the other third for estimating performance.

On the other hand, **k-fold Cross-Validation** is used only when a limited amount of data are available, in order to achieve an unbiased estimate of the model performance. In k-fold cross-validation, the data set is divided into k subsets of equal size. Then, models are trained k times, each time leaving out one of the subsets from training and use it as the test set. If k equals the sample size, this is called "leave-one-out" and if k is set to two then, it is a single train/test split.

The choice of evaluation metrics that are used in the validation and test set depends on the given Machine Learning task (such as classification, regression, ranking, clustering, topic modeling, among others). Unfortunately, it is not a topic that has, generally, been given much thought in the fields of Machine Learning and Data Mining. More often than not, common off-the-shelf metrics are applied without much attention being paid to their meaning. In this section we will review some of the most important indicators used in classification problems and give an intuitive idea of what could go wrong with them.

Confusion Matrix is suggested as one of the best approaches to illustrate the performance of Machine Learning programs. Particularly, in binary classification problems is a four cell contingency table that distinguishes between true positive, false positive, true negative and false negative predictions [Powers, 2011].

		Actual Values: p	Actual Values: n	
		p	**n**	**total**
Predicted Values	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Table 2.1: Confusion Matrix

The **accuracy value** is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Number of Total Predictions}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

Accuracy does not distinguish between the types of errors it makes (False Positive versus False Negatives). While this is acceptable only if there are equal number of samples belonging to each class (i.e., if it is a balanced data set). For example, consider that there are 93% samples of class A and 7% samples of class B in the training set. Then the model can easily get 93% training accuracy by simply predicting every training sample belonging to class A (***Benchmark Method***). When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the test accuracy would drop down to 60%. Classification accuracy is great, but gives the false sense of achieving high accuracy.

The real problem arises, when the cost of misclassification of the minor class samples are very high, like in the case of a medical classification problem whose goal is to discriminate cancerous (positive class) from non cancerous patients (negative class). If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests [Japkowicz, 2006]. In such cases, the **recall** value in conjuction with **precision** value are more effective, because they overcome this exact problem of benchmark.

The **recall value** or *sensitivity* or *True Positive Rate* (TPR) is the proportion of true positive cases that are correctly predicted positive.

$$recall = \frac{TP}{TP + FN} \tag{2.8}$$

Conversely, the **precision value** or *confidence* denotes the proportion of predicted positive cases that are correctly true positives.

$$precision = \frac{TP}{TP + FP} \tag{2.9}$$

Precision assesses to what extent the classifier was correct in classifying examples as positives, while recall assesses to what extent all the examples that needed to be classified as positive were so [Japkowicz, 2006]. Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. So, depends on the type of problem if one should go for high precision or high recall.

For example, for rare cancer data modeling, anything that does not account for false-negatives is a crime. Recall is a better measure than precision. For recommendations, false-negatives is less of a concern. Precision is better here. A website may have thousands of free customers registrations every week. The call center team wants to call them all, but it is impossible due to limited time. So, they have to select those with good chances to buy. It does not matter if they call someone that is not going to buy (so precision is not so important), but it is very important that all of them with high temperature are always in the target list, so they do not go without buying. That means that the predictive model needs to have a high recall, no matter the precision. However, in most cases precision and recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75).

F-Measure or *F1-score* aims to combine the statements of recall and precision by using the harmonic mean between the two:

$$F - Measure = 2 \times \frac{\text{precision } \times \text{ recall}}{\text{precision } + \text{ recall}} \tag{2.10}$$

Accuracy is used when the true positives and true negatives are more important, while F-Measure is used when the false negatives and false positives are crucial [Huilgol, 2019].

What is more, recall, precision and F-Measure focus only on the positive examples and predictions and this is a great advantage over accuracy. What they do not do, however– which, incidentally, accuracy does–, is capture any information about how well the model handles the negative cases [Powers, 2011, Japkowicz, 2006]. For instance, in the medical domain be able to recognize that a truly healthy patient is, indeed, healthy.

The **Receiver Operating Characteristic Curve (ROC)** is a probability curve and the **Area under ROC Curve (AUC)** represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting zeros as zeros and ones as ones. The measures it uses are the true positive rate (TPR) or recall and the false positive rate (FPR).

$$\textbf{specificity} = \frac{TN}{TN + FP} \tag{2.11}$$

$$\mathbf{FPR} = 1 - specificity = \frac{FP}{TN + FP} \tag{2.12}$$

ROC Analysis plots the false positive rate on the x-axis of a graph and the recall on the y-axis, where:

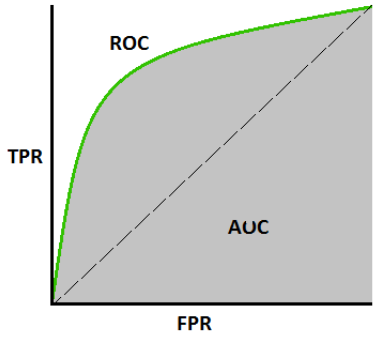


Table 2.2: AUC-ROC Curve [Narkhede, 2018]

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting zeros as ones and ones as zeros. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

So, let's interpret above statements:

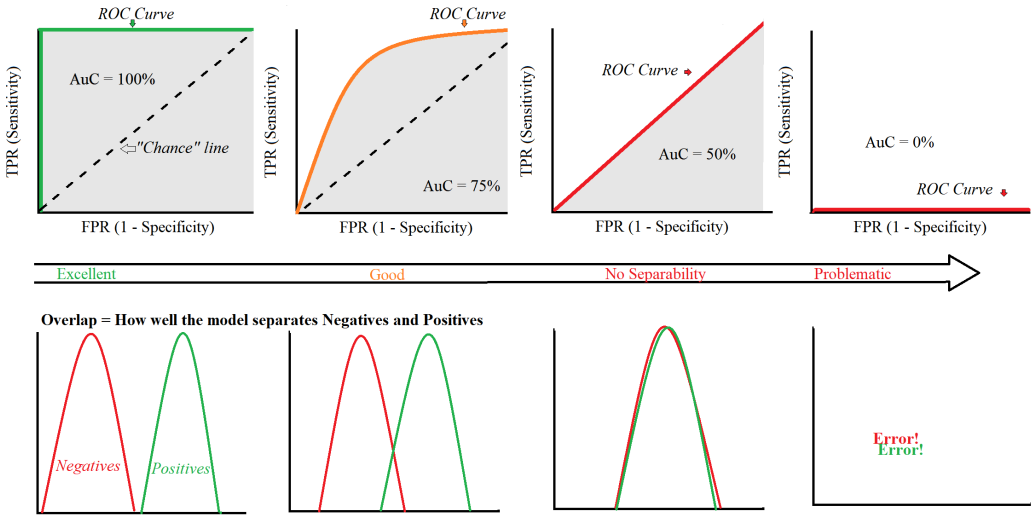


Table 2.3: AUC-ROC Curve [Glen, 2019]

In multi-class model, we can plot N number of AUC-ROC Curves for N number classes using One vs All methodology. So for example, if you have three classes named X, Y and Z, you will have one ROC for X classified against Y and Z, another ROC for Y classified against X and Z, and a third one of Z classified against Y and X.

2.3 General Issues of Machine Learning Techniques

The most important part in Machine Learning is the selection of relevant data. The data-driven decisions are only as good as the quality of data you gather. Even the most complex model with the state-of-the-art techniques will have low performance, if the data are not relevant and valid. If a requisite expert is available, then s/he could suggest which fields (attributes, features) are the most informative. If not, then the simplest method is that of "brute-force", which means measuring everything available in the hope that the right (informative, relevant) features can be isolated [Kotsiantis et al., 2006]. However, a data set collected by the "brute-force" method is not directly suitable for induction. It contains in most cases noise and missing feature values, and therefore requires significant pre-processing [Zhang et al., 2003]. Data preparation can be more time consuming than data mining, and can present equal, if not more, challenges than data mining.

So, what can be wrong with data? There is a hierarchy of problems that are often encountered in data preparation and pre-processing:

Data Entry Errors: These errors should be checked for by the data handling software, ideally at the point of input, so that they can be re-entered. They are generally straightforward, such as coming across negative prices, when positive ones are expected. If correct values cannot be entered, the problem is converted into missing value category, by simply removing the data.

Anomaly Detection (Outliers): Detecting outliers is of major importance for almost any quantitative discipline. Most Machine Learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results. The use of visualization can often be a powerful tool. Visualization is particularly good at picking out "bad values" that occur in a regular pattern. Many surveys also, have introduced contemporary techniques for anomaly detection. These surveys focus on analytical methods, unsupervised anomaly detection techniques – distances or densities are used to give an estimation of what is normal and what is an outlier and deep learning – when anomalies are known and labeled correctly [Erfani et al., 2016, Goldstein and Uchida, 2016].

Missing Values Imputation: Missing values is an unavoidable problem in dealing with most real world data sources. Generally, there are some important factors to be taken into account when processing unknown feature values. One of the most important ones is the source of "unknown-ness": (i) a value is missing because it was forgotten or lost; (ii) a certain feature is not applicable for a given instance (e.g., it does not exist for a given instance); (iii) for a given observation, the designer of a training set does not care about the value of a certain feature [Kotsiantis et al., 2006]. Depending on the circumstances, researchers have a number of methods to choose from to handle missing data. On the one hand, there are the commonly used methods like, deletion method or imputation with the mean/median value to numeric features or the mode to objects [Acuña and Rodriguez, 2004].

On the other hand, more complex methods (ML) have been proposed to treat missing values [Acuña and Rodriguez, 2004, Ragel and Crémilleux, 1998] like, kNN classifier, accociation rules etc. There are also, lots of easy to use packages that offers several of the above techniques in order to handle missing values automatically [Moritz and Bartz-Beielstein, 2017].

The Curse of Dimensionality: The curse of dimensionality refers to the phenomena that occur when classifying, organizing, and analyzing high dimensional data that does not occur in low dimensional spaces, specifically the issue of data sparsity and "closeness" of data. On the one hand, ML excels at analyzing data with many dimensions. On the other hand, as we add more dimensions we also increase the processing power we need to analyze the data, and we also increase the amount of training data required to make meaningful models – each feature increases the data set requirement exponentially. To overcome the issue of the curse of dimensionality, *Dimensionality Reduction* and *Feature Selection* is used to reduce the feature space [Van Der Maaten et al., 2009]. PCA (Principal Component Analysis) is one of the most traditional tools used for dimension reduction. It transforms the data into the most informative space, thereby allowing the use of lesser dimensions which are almost as informative as the original data. Meanwhile, feature selection is the process of identifying and removing as many irrelevant and redundant features as possible (Yu and Liu 2004). This reduces the dimensionality of the data and enables data mining algorithms to operate faster and more effectively.

Imbalanced Data Set: Imbalanced data sets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed by two classes: The majority (negative) class and the minority (positive) class. These type of sets suppose a new challenging problem for Data Mining, since standard classification algorithms usually consider a balanced training set and this supposes a bias towards the majority class. Various real-world classification tasks, such as medical diagnosis, text categorization and fraud detection suffer from this phenomenon [Ertekin et al., 2007]. To handle imbalanced data sets, there a lot of techniques such as, over-sampling – increases the number of minority class members in the training set, down-sampling – reduces the number of majority samples, use of ensemble learning techniques – convert an imbalanced data set into multiple balanced ones and then build a number of classifiers on these multiple data with a specific classification algorithm [Sun et al., 2015].

In this work, Machine Learning methods will be used in the context of Customer Churn Management problem, which will be explained in *chapter 3*.

Chapter 3

Customer Churn Management

"Customers opt for a product or a service for a particular period, which can be rather short – say, a month. Thus, a customer stays open for more interesting or advantageous offers. Plus, each time their current commitment ends, customers have a chance to reconsider and choose not to continue with the company. Of course, some natural churn is inevitable, and the figure differs from industry to industry. But having a higher churn figure than that is a definite sign that a business is doing something wrong."

— Alex Bekker, the Head of Data Analytics Department at ScienceSoft

Customer retention is a major problem and one of the most important concerns for large companies. As markets have become increasingly saturated, companies have attempted to identify ways in which to improve customer loyalty, satisfaction, and retention. The marketing approach within many organizations has gradually evolved from product-centric to customer-centric. This approach is supported by modern database technologies, which enable companies to obtain the knowledge of who the customers are, what they have purchased, when they purchased it, and predictions on behavior. Academics have generated a large body of research that addresses part of that challenge with a particular focus on predicting **customer churn**.

This section will focus on the relevant theoretical foundations used to understand, analyze and manage churn by answering the following questions:

1. *What is customer churn, and why does it matter to a business?*
   This is answered in Section 3.1.

2. *Why customers churn?*
   This is answered in Section 3.2.

3. *How companies deal with churners?*
   This is answered in Section 3.2.

4. *How Machine Learning powered churn analysis for modern day businesses?*
   This is answered in Section 3.4.

5. *Why Machine Learning is the predominant method for retention?*
   This is answered in Section 3.5.

To answer these questions, the first section defines the Customer Churn Problem and its importance to businesses. Afterwards, the main factors that lead to customer churn are briefly presented. The next section describes the basic customer retention techniques that companies use to follow. After that, we suggest the way Machine Learning can be applied to customer retention and churn analysis. The section is closed by addresing the points, in which ML is the predominant method.

3.1 Introduction to Customer Churn

Customer churn, also known as customer attrition, customer turnover, or customer defection is a business term used to describe lost of clients or customers. Many businesses with large customer bases, particularly subscriber-based businesses (like telecommunication companies, banks and retail companies) monitor and manage their churn numbers very closely. Some of the most valuable measurements of customer churn are:

- The total number of lost customers within a specific time frame.
- The percentage of lost customers within a specific time frame.
- The recurring business value that is lost.
- The percentage of the recurring business value that is lost.

The metric tracked is typically known as the "churn rate" and is expressed as a percentage [Klepac et al., 2014]. Basic calculation to express churn rate is relatively straightforward: number of customers that defected over the period divided by the total number of customers during the period:

$$\textbf{Churn Rate (\%)} = \frac{\text{Number of Churned Customers}}{\text{Total Number of Customers}} \tag{3.1}$$

Customer attrition is an important issue for any company, and it is especially important in mature industries where the initial period of exponential growth has been left behind [Klepac et al., 2014]. This is due to the fact that new customers typically churn at a higher rate than customers that have stuck around for a bit. So, if a company is growing and mainly acquires new customers, its churn rate will skew higher than it really is and will not be representative. However, customer churn is an agonizing reality that affects all businesses at some point. Not even the largest or most successful companies are spared from customers' defection. According to Lincoln Murphy an acceptable churn

rate is in the $5 - 7\%$ range **annualy**, depending upon whether you measure customers or revenue [Murphy,].

Lots of studies has shown that is more profitable to keep and satisfy existing clients than to constantly attract new ones. As we have already discussed, the cost of acquition of a new customer is $6 - 7$ times more than retain an existing one. Furthermore, according to marketing metrics the probability of selling to an existing customer is $60 - 70\%$, while the probability of selling to a new prospect is $5 - 20\%$, which is reasonable due to the customer relationship and loyalty. And besides, it is not by chance that 80% of a company's future revenue will come from just 20% of its existing customers. [evolve].

Taking into consideration the statistics mentioned above, the first rule of any business is to retain customers and build a loyal relationship with them, and thereby avoid customer acquisition costs. In fact, only 18% of companies have a greater focus on retention vs. 44% that focus on customer acquisition. However, to be efficient, companies have to find a way to balance between these two activities. It is true that only 40% of companies and 30% of agencies have an equal focus on acquisition and retention [Saleh,].

3.2 Causes of Customer Churn

The first step for lasting and sustainable business growth is to understand what has caused previously loyal customers and users to abandon ship and maybe find a new provider instead. As a result, there is a need of categorized churners, based on their churn action, in order to handle them severally in a more targeted and effective way. Lots of studies has shown that churning customers can be divided into three main groups which are **voluntary churners**, **non-voluntary churners** and **silent churners** [Farhaoui and Moussaid, 2019, Klepac et al., 2014]:

Non-voluntary churners are the easiest to identify, as they are the customers who have had their service withdrawn by the company. There are several reasons why a company could revoke a customer's service, most often abuse of service and non-payment of service.

Voluntary churn is more difficult to determine, because this type of churn occurs when a customer makes a concious decision to terminate his service with the provider. Voluntary churn can be sub-divided into two main categories, **incidental churn** and **deliberate churn**:

- Examples of **incidental churn** include changes in the customer's financial circumstances, so that the customer can no loger afford the service, or a move to a different geographical location where the company's service is anavailable. Incidental churn usually explains only a small percetage of a company's voluntary churn.

- **Deliberate churn** is the problem that most churn management solutions try to battle. This type of churn occurs when a customer decides to move his custom to a competing company. Reasons that could lead to a customer's deliberate churn include technology-based reasons, when a customer discovers that a competitor is offering the latest products, while their existing supplier cannot provide them the same product/service. Economical reasons include finding the product at a better price from a competing company. Examples of other reasons for deliberate churn include quality factors such as poor coverage, or possibly bad experiences with call centers etc.

"71% of consumers have ended their relationship with a company due to poor customer service. "

— KISSMetrics

"Price is not the main reason for customer churn, it is actually due to the overall poor quality of customer service."

— Accenture Global Customer Satisfaction Report (2008)

"A customer is 4 times more likely to defect to a competitor if the problem is service-related than price- or product-related."

— Bain & Company

Silent churners are those that left the company without any prior reason that the company can make use for reducing the future churners [Farhaoui and Moussaid, 2019].

"For every customer complaint, there are 26 other unhappy customers who have remained silent."

— Lee Resource

"96% of unhappy customers don't complain, however 91% of those will simply leave and never come back."

— Financial Training services

Different factors lead to different kinds of churn – each requiring a specific approach. Improving customer satisfaction reduces cancellations that result in voluntary churn, while using decline management techniques minimizes payment declines that lead to involuntary churn. Churn analysis would help to identify the cause of this churn, opening up opportunities to implement effective retention strategies.

3.3 Customer Retention Techniques

Generally, there are two basic approaches to manage customer churn, **untargeted approaches** and **targeted approaches** [Neslin et al., 2006]. Untargeted approaches rely on superior product and mass advertising to increase brand loyalty and retain customers.

On the other hand, targeted approaches rely on identifying customers who are likely to churn, and then either providing them with a direct incentive or a customized service plan to stay. Targeted approaches can be sub-divided into two main categories: the **proactive campaigns** and the **reactive campaigns**:

With the reactive campaigns, the firm waits until the customer contacts the firm to cancel his or her account. Customers do not churn until the end of their subscription period arrives and they do not renew, because they have already paid up until the end of their subscription period. So, if they have only canceled, there is still a chance to win them back before their subscription end [Tate, 2020], typically with a financial incentive, e.g., a rebate to stay. With the proactive campaigns, the firm takes into account the risk and tries to identify *in advance* customers who are likely to churn at some later date. Then the firm targets these customers with special programs or incentives to forestall the customer from churning.

At this point, it is worth noting that, by the definition of the moment of churn – i.e. the moment the subscription ends and renewal does not happen, it is not actually possible for new signups to churn in their first month, and this does away with the issue of how new customer growth can distort the churn calculation. New signups should not be included in churn or the total number of customers for their first month.

Firms typically implement multiple campaigns. The reactive and proactive programs need to be coordinated and finally, all these efforts need to be integrated with the firm's marketing strategy. To develop and evaluate a single retention campaign, the firm needs to: First, identify customers who are at risk of not being retained. Second, diagnose why each customer is at risk – *section 3.2*. Third, decide what kind of churners the particular campaign aims to target. Next, decide when to target these customers and with what incentive and/or action. Finally, implement the campaign and evaluate it. These steps are applicable to both proactive and reactive campaigns [Ascarza et al., 2018].

However, some difference between the two campaigns are worth noting. On the one hand, reactive campaigns are simpler because the firm does not need to identify who is at risk – the customer who calls to cancel self-identifies. Besides, once they have request to cancel in themselves, it seems more straightforward to find out their underlying pain points and renegotiate their contract successfully. So, "rescue rates" can readily be calculated to evaluate the program, and subsequent behavior can be monitored.

On the other hand, proactive campaigns tend to be more challenging because not all customers can be rescued, and customers learn that informing the firm about their intention to churn can be copiously rewarded by the firm, endangering the long-run sustainability of reactive churn management. Hence, the incentive offered to win back the customer in a reactive campaign is typically high value, relative to a proactive campaign because the firm is certain that the customer will churn [Ascarza et al., 2018].

Obviously, targeted proactive programs have potential advantages of lower incentive costs (since the incentive may not have to be as high as when the customer has to be "bribed" at the last minute not to leave) and not training customers to negotiate for better deals under the threat of churning [Neslin et al., 2006]. However, they are more challenging starting from the basic task of identifying who is at risk of churn and thus, they can be very wasteful if churn predictions are inaccurate, because in that case firms are wasting incentive money on customers who would have stayed anyway. As a result, state-of-the-art analytics are required to balance the costs of false positives (targeting a customer who had no intention to leave) and false negatives (failing to identify a customer who was truly at risk). The goal then is to predict customer churn as accurately as possible.

The next chapter will focus on how to predict possible churners for targeted proactive campaigns using Machine Learning techniques.

3.4 Proactive Customer Retention using Machine Learning

As we have already mentioned, companies have been mostly using traditional reactive churn techniques based on simple statistics models. They have been listing high-propensity churners and addressing their needs through special concierge services. But, as the market becomes more competitive over time, businesses have realized that this approach for churning customers is not a sustainable process and they need a holistic proactive churn management strategy thats takes into account the risk, the level and the cost of the retention, as well as the causes of such behavior. A large amount of research suggest Machine Learning to be that strategy.

ScienceSoft's Alex Bekker stresses the importance of Machine Learning for proactive churn management: "As to identifying potential churners, Machine Learning algorithms can do a great job here. They reveal some shared behavior patterns of those customers who have already left the company. Then, ML algorithms check the behavior of current customers against such patterns and signal if they discover potential churners" [1].

So, the main question Machine Learning aims to answer in churn management is: *who is at risk?* or *what is the probability of a customer to churn?* This entails using a predictive model to identify customers at risk of not being retained or in general of generating lower retention metrics. In the

[1] https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html

simplest case, the dependent variable could be 0-1 (churner/non-churner). Then, the predictive churn model would be a straightforward Machine Learning classification tool: look at the user activity from the past and check who is active after a certain time and who is not and create a model that probabilistically identifies the steps and stages when a customer (or segment) is leaving your service or product [Kundu, 2018].

Some possible categories of churners are:

- Closure of an Account
- Non-Renewal of a Contract or Service Agreement
- Use Another Service Provider
- Defaulter

In that case, the predictive churn model would be a multi-class classification model due to the different categories of churners. Generally, the different types of churners that will be predicted using the churn model are based on the current retention campaign and the types of customers the company aims to target.

Having a predictive churn model gives the awareness and quantifiable metrics to fight against in retention efforts. This gives the ability to pattern habits of customers who leave, and step in before they make that decision. Without this tool, businesses would be acting on broad assumptions, not a data-driven model that reflects how customers really act. Without a strong understanding of customers and their behavior, it is hard to keep them satisfy and therefore, retain them.

The key element of a churn predictive model is the **selection of relevant data**. The more relevant data you gather, the more accurate the model will be. Therefore, the first phase involves the identification of data sources that will help to determine the behavior of the customers or broaden company's knowledge about them. In this case the proper sources of data are the ones that assess the triggers that caused past clients to ultimately leave the company. These include well-researched predictors like customer satisfaction, usage behavior, switching costs, customer characteristics, customer complaints and marketing efforts, as well as more recently explored factors such as emotions and social connectivity [Ascarza et al., 2018, Kundu, 2018]:

- **Customer Characteristics:** Demographics (zip code, gender, age, occupation, annual income, geographical location etc.), psychographic segment, customer tenure, etc.
- **Usage Behavior:** Products, usage level, coupon usage, number of products, etc.
- **Customer Satisfaction:** Emotion in e-mails, store visits/online, customer service calls, usage trends, etc.

- **Marketing:** Mail responders, response to direct mail, open newsletters, response to click-to-call, previous marketing campaigns, acquisition method, acquisition channel, etc.
- **Complaints:** Complaint resolutions, complaint priority, frequency of complaints, etc.
- **Switching Costs:** Pricing plan, add-on services, ease of switching, etc.
- **Purchase History:** Frequency of purchase, date of last purchase, value of purchases, balances/store credit, etc.
- **Social Connectivity:** Neighbor churn, social network connections, social embeddedness, neighbor/connections usage

Social connectivity factors can predict churn. In the context of telecommunications, it has been shown that high "social embeddedness", – the extent to which the customer is connected to other customers within the network, is negatively correlated with churn [Ascarza et al., 2018]. Furthermore, the behavior of a customer's connections also affects his own retention. For example, a customer is more likely to churn from a service/company if his contacts or friends within the company churn.

Another important question about relevant data sources is whether churn prediction can be improved using ultra-fine-grained "big data". These are actions consumers take such as visiting a web page, visiting a specific location, "linking" something on social media, etc. These kind of data are valuable for a company, while they can provide much knowledge about the customers' sentiments associated with the company and mostly about the customers' relationships with the competitors. However, the downside of data from social media monitoring is that they often have very high dimensionality and extreme sparsity. Thus far, there is no direct evidence of this sort of data improving retention prediction.

Depending on the number of data sources, this phase may last from three weeks to up to three months [Detko, 2019] and it is a very important stage because well-defined business conditions will later translate into accuracy in metrics. However, even if all the above data seems to be relevant for the churn model, there is a difference between determining the best predictors of churn and understanding why the customer is at risk of churning [Ascarza et al., 2018]. For example, demographic variables might predict churn, but these variables rarely cause customers to leave the company. As a result, it is of crusial importance to proceed to a data analysis, that will find out the most representative variables of each type of churners.

So, once sufficient data have been collected for the churn model, the next step is the **data analysis phase**. During this phase, it is essential to understand the data set and the features that characterize the customers. Visualization of variables' distribution would help, as well as a preliminary statistical analysis to determine potential variables for modeling. The main objective,

however, is to detect some important independent variables that have positive or negative correlation with the dependent one.

After the data analysis phase, the **data preparation phase** follows. It is said that data scientists spend 80% of their workday preparing data [Detko, 2019]. As it was discussed in the *section 2.3*, properly formatted and validated data improves data quality and protects applications from potential landmines such as null values, unexpected duplicates, incorrect indexing, and incompatible formats.

Once the data have been chosen, prepared, and cleaned, modeling can begin. This stage involves creating a predictive model, which is fed with some training data and tries to identify patterns in them. Then, it can process additional testing data to make predictions on. It is always good to start with a simple base model, so you can get a base line to measure performance against. This stage also, includes *Model Evaluation*, *Performance Monitoring* and *Hyperparameters Fine-Tuning*.

Model Evaluation: While training a model is a key step, how the model generalizes on unseen data is an equally important aspect that should be considered in every machine learning pipeline. Could the model be merely memorizing the data it is fed with, and therefore unable to make good predictions on future samples, or samples that it has not seen before? Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

Performance Monitoring: The most suitable perfomance indicators must be carefully chosen in order to evaluate model's performance correct. In case of customer churn, accuracy and F-measure are not the most representative metrics, since data sets that refers to customers and past churners are mostly (or should be) imbalanced. Bearing in mind the churn rate for a sustainable company, customers that stay should be proportionally many more than those who churn. Consequently, accuracy may be very high, just because all non-churners have been correctly predicted, which has no meaning for churn management. Thus, the metrics that are mainly suggested are the confusion matrix, the recall and the precision, since they give a detailed overview of the misclassified data.

Meanwhile, customers databases commonly consist of thousands of entries. So, if a business team wants to run a retention campaign, they are able to contact only a small portion of them, i.e. the portion with the highest probabilities to churn. Thus, they do not care to predict them all, while they will not be able to contact them all. Instead they want, that portion of customers that will finally target, no matter how small it is, at least be correctly predicted and do not include customers that would have stayed anyway. So, to be more precise the metric that seems to count more in churn prediction is the precision.

3.5 Advantages of Machine Learning

Adopting Machine Learning for churn prediction has several advantages over traditional business rules:

1. Machine Learning generally relies on finding patterns and relationships in large amounts of data. The rules discovered by the Machine Learning model are guaranteed to be supported by evidence instead of intuition.

2. Unlike humans, who are limited by the number of variables/factors they can account for when crafting their business rules, ML algorithms can process and extract patterns from many variables, which results in more complex and comprehensive rules.

3. Assuming there is high quality data available, Machine Learning is able to learn highly accurate rules in a much shorter time compared to a human, who usually needs a significant amount of experience and domain knowledge to devise accurate rules, i.e., the ROI tends to be higher for Machine Learning.

4. A fourth advantage is that Machine Learning is able to timely detect concept drift and adapt the rules accordingly, thus it is more adaptive to changes, i.e., if the accuracy of the churning predictions start to degrade over time, Machine Learning quickly detects it and adapts the rules to the new scenario, ensuring the prediction of churners is reliable for the business.

In short, companies nowadays have an ever-growing data pool, but slow processing of information. With Machine Learning, they are now able to analyze huge and complex data, while delivering accurate results at a faster rate. Through this change, they can avoid potential unknown risks and find great opportunities.

Chapter 4

Conclusions

4.1 Summarization of Findings

Although variations of Machine Learning have long been around, the discipline has developed rapidly in recent years. Many factors have combined to derive this development. Increased computer power has made real time processing feasible for many complex tasks, increased connectivity has driven innovation and automation in the delivery of traditional tasks and services, the potential to extract useful information from the vast amounts of data generated via the internet (Big Data) has led to novel analytical methods. Alongside this, the development of easy to use programming languages, such as Python and R, and Machine Learning focused frameworks such as TensorFlow, has contributed to the wide investigation of Machine Learning applications in industry. The success of these applications is driving Machine Learning commercial research into further applications. Businesses increasingly adopt data-driven culture and Machine Learning techniques to optimize most of their internal processes.

From the research I conducted, I noticed that within a saturated market, customer acquisition no longer ensures sustainable revenue. The paradigm has moved towards customer retention. The current scenario, full of challenges for every subscriber-based provider, states churn management as the door to revenue growth. Different factors lead to different kinds of churners, each of those requires a special treatment. Businesses have now, realized that waiting for a cancellation request to act, is not an efficient strategy. Instead, they should take into account the risk of churn and try to identify those customers in advance. The most promising way is through Machine Learning, considering the Customer Churn Problem as a predictive classification problem of churners and non-churners. Then the companies will be able to pattern the habits of customers who leave, and step in before they make that decision.

4.2 Limitations

As discussed above, there is much hype surrounding Big Data. Firms are constantly exhorted to set strategies in place to collect and analyze Big Data, and warned about the potential negative consequences of not doing so. However, the Wall Street Journal has suggested that although companies sit on a treasure trove of customer data, for the most part they do not know how to use it [1]. Machine Learning may have been established in the research community for NLPs, Computer Vision and Speech-to-Text problems, however in industry and mainly in respect to Human Behaviour problems is still imature.

Furthermore, due to the strict regulations that GDPR has set in Europe, there is a growing pressure on organizations to better protect their sensitive customer and operational data. Customers should give their consent before companies store their personal data and be assured for the purpose of their usage. Moreover, they are able to ask for their deletion and the deletion of metadata as well, whenever they want to. As a result, whatever analysis is based on these kind of data and especially when the results will be used by the company, then an extra attention should be paid. In parallel, companies when storing or transmitting personal data (metadata), should use an encryption policy and ensure that this encryption solution meets the current standards. Due to this policy, even the company itself has not always the ability to identify the person to whom the encryption relates.

[1] https://pdfs.semanticscholar.org/7fee/58f6c5616d4324c13be9c61db35bd54b419c.pdf

Part II: Internship

Company:

Vodafone-Panafon Hellenic Telecommunications Company S.A

Position:

Data Scientist Intern

Chapter 1

Introduction

During the final semester of my studies at the Athens University of Economics and Business, Greece, from November to February 2019-20, I implemented my internship at Vodafone-Panafon SA.[1] Throughout this three-month intership, I had the chance to work with professionals in the area of Data Science and improve my skills in the domain.

1.1 Company

Vodafone Group is a multinational telecommunications company. Its registered office is located in Newbury, Berkshire, England and its global headquarters is based in London, England. It predominantly operates services in the regions of Asia, Africa, Europe, and Oceania.

Vodafone is a leader in the technology communications through mobile, fixed, broadband and TV. It has an extensive experience in connectivity, convergence and the Internet of Things, as well as championing mobile financial services and digital transformation in emerging markets. Since making the first mobile call in the UK on 1 January 1985, Vodafone has grown into an international business and one of the most valuable brands in the world. It has mobile operations in 24 countries, partner with mobile networks in 42 more, and provide fixed broadband in 19 markets. As of 30 September 2019, Vodafone Group had approximately 625 million mobile customers, 27 million fixed broadband customers and 22 million TV customers, including all of the customers in Vodafone's joint ventures and associates.

More specifically, Vodafone offers a range of communications service to both consumers and businesses in multiple regions. For European Consumers provides a range of mobile services, including calls, text and data whether customers are at home or travelling abroad. On the other hand, Vodafone identifies mainly three categories of Businesses:

[1] https://www.vodafone.gr/

- **SoHo**, which means Small office–Home office, and describes the Businesses that own from 1–9 telephone lines – e.g. family businesses.
- **SME**, which means Small-Medium Enterprises and consists of Businesses that own from 10–50 telephone lines – e.g. a law firm.
- **Major Accounts**, that are Enterprises and Organizations that own more than 50 telephone lines – e.g. an advisory company or a Ministry.

To the above businesses, Vodafone delivers mobile, fixed and a suite of converged communications services to support the growing needs of business customers, ranging from small home offices to large multi-national companies. Also, it offers a diverse range of Internet of Things (IoT) services to business, including managed connectivity, automotive and insurance services, smart metering and health solutions. Cloud & Security portfolio includes both public and private cloud services, as well as cloud-based applications and products for securing networks and devices. Last but not least, Vodafone sells capacity on its global submarine and terrestrial cable systems. The services include international voice, IP transit and messaging.

Vodafone believes that the opportunities and promise of a better digital future should be accessible to all. Through the technology, Vodafone will work to bridge the divides that exist and help people to contribute equally and fully to society. The goal is to create a *Gigabit Society* characterised by hyper-connection, very high speed, very low latency, ultra-dense coverage and very secure exchange of data. This will create new social and economic benefits through the IoT and new digital developments.

Particularly, Vodafone is committed to investing in the network infrastructure and coverage to deliver a high-quality service that allows individuals and businesses to connect confidently anywhere and at any time. By 2025, with the Gigabit networks, citizens will access an ever-growing range of services in real-time and businesses can develop new products and services to meet the needs of future generations. The IoT will create more efficient, safer and smarter transport; and mobile financial services will reduce poverty and enable access to essential services like healthcare and education. These services enable incredible innovation and technologies to be developed to help make peoples' lives easier, healthier, smarter and more fulfilling.

1.2 Intership Goals

Intership is absolutely the best way for someone to harness the skill, knowledge, and theoretical practice he/she learnt in the university. You can acquire endless amounts of education in your life, however, that knowledge does not always translate to the working life. Personally, through my internship at Vodafone company I aimed to accomplish a variety of goals.

First of all, I wanted to advance my technical skills. The main goal was to experience a real-world machine learning project in a large multinational company. I set the aim to participate in the process of handling and preprocessing large amounts of data and implement complex machine learning algorithms to get the best possible results.

Apart from technical knowledge, this internship constituted my first work experience, out of my comfort zone. I was finally in a real work environment, working for a project with a real impact for the company, under real working conditions in fixed work shifts, deadlines, being in constant collaboration with other employees and supervised for my progress. Better communication, collaboration, adaptability and flexibility are some examples of the skills that I developed.

1.3 Goal of the Report

Through the implementation of my final report, I attempt to summarize and evaluate all the knowledge, experience, and overall impact of my internship. This report provides a thorough description of my three-month activities and evolution at Vodafone. Also, inspired by the project I worked on, I present a synopsis of the methodology that was used for the implementation of a predictive Churn Model using Machine Learning techniques.

Chapter 2

Department and Role

2.1 Organizational Structure

Vodafone is a multinational company and thus, has a complex hierarchical organization structure. Its corporate structure consists of various departments that contribute to the company's overall mission and goals. Apart from the common departments including Finance, Human Resources, Legal & External Affairs, Vodafone in Greece consists of four more main departments; the Consumer Business Unit which handles all operations of Retail Customers (consumers), the Enterprise Business Unit which supports the Businesses, the Commercial Operations Department which is responsible for the digital era of operations and the Technology Depatment.

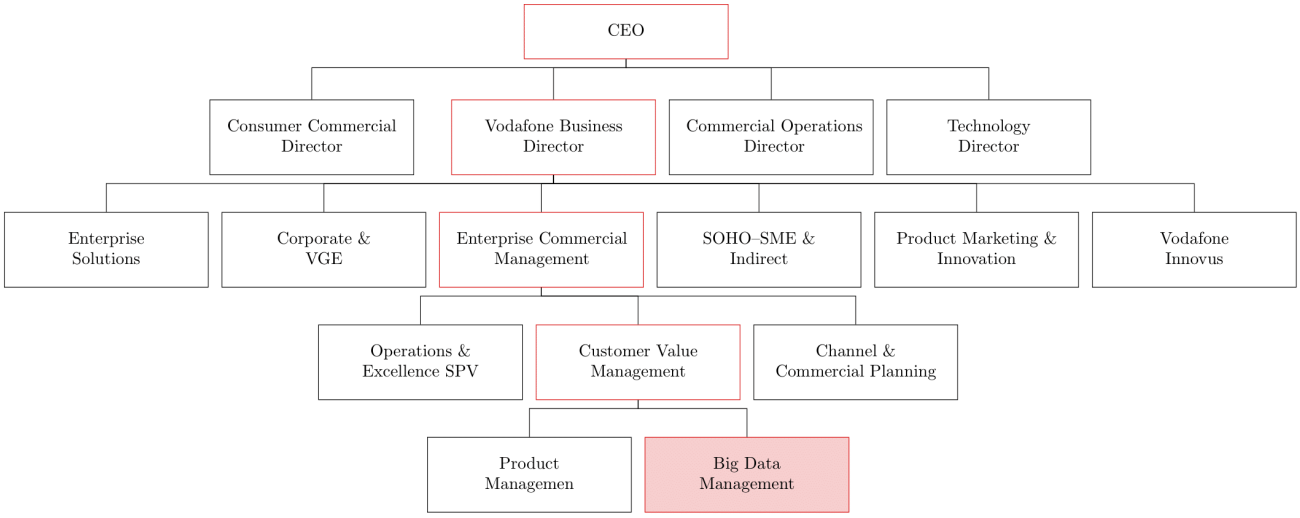


Figure 2-1: Vodafone's Organizational Chart.

2.2 My Role

During my intership, I was assigned to the Customer Value Management team (CVM) of the Enterprise Business Unit (EBU). The supervisor of the department is Dimitrios Gratsias, who leads two sub-teams: The Product Management team and the Big Data team. The Product Management team consists of eight employees and is responsible for several products/services campaigns related to Enterprises. For instance, retention campaigns, migration campaigns, product-mix campaigns, VDSL upgrades, renewals, churners, etc.

The position I was assigned to, was in the sub-team responsible for Big Data and Data Science projects for Vodafones' business partners. The Data Science team oriented only toward Enterprises is a newcomer team for Vodafone, which until now was consisted of one person: Georgina Bilali, the Big Data Scientist. Moreover, one year ago, there was not even a person dealing with big data for Enterprises in the whole Vodafone in Greece. The only related department was the one responsible for consumers. However, Vodafone Group has realized that its of crucial importance to enhance Enterprises support in Greece as well, by utilization of new technologies, big data and machine learning capabilties.

The aim of this newcomer team is to support the whole CVM and Product Managers by providing them with usefull insights and solutions for their campaigns. Particularly, while vodafone's customer database consists of thousands of companies, if CVM wants to run a campaign, they are able to contact only a small portion of them. In this case, my team is responsible for finding those customers that satisfy the requirements of the specific campaign and are more likely to successfully respond.

During the period of my intership, Data Science team was specifically requested to built a predictive model that identifies the customers that are in risk of churn. The purpose was to help retention campaigns target the correct population, act in advance and reduce churn rate. This project involved first the implementation of a Churn Model using Machine Learning and then the creation of the targeted list, with the customers in the top performing quartile of churn, for the next retention campaign.

2.2.1 Required Skills

In order to fulfil the goals of the project, three main skills were required. The first skill involved basic familiarity with research conduction. Since, I had to explore the best practices and state-of-the-art approaches in churn prediction, I had to be familiar with research performance, particularly digital libraries where research is maintained, and how to properly study and evaluate research.

The second skill required, for the implementation of the Churn Model, was experience with Python. I had already sufficient knowledge of Python from University courses, and thus this project allowed me to build and expand on my previous knowledge by exploring new libraries and better programming practices.

The third skill involved experience with Machine Learning. During my studies, I had the chance to get familiarized with plenty of Machine Learning techniques both algorithmically and programmatically and work on several Machine Learning projects as well. The challenge in that case, however, was the handling of a large-scale project, with large amounts of data and complex relationships.

Additional skills that were also beneficial to my internship was collaboration, ease of communication, flexibility and problem solving. Four years at the department of Management Science and Technology have equipped us with competencies to work professionally in a large company such as Vodafone. Although we are not experts in Artificial Intelligence and Machine Learning techniques, I believe our university studies have assisted in building a mindset of continuous learning. Additionally, working on multiple group projects at University has contributed to the development of several soft skills, such as communication within teams, professionalism, diligence and attention to detail, which are some important factors for success cultivated through hard work.

2.2.2 Expected Results

The expected results of my internship were the successful implementation of the Churn Model and its correct utilization from the business units. Apart from these, it was my personal intention to satisfy the expectations of my supervisor at Vodafone, and also to manage favourable collaboration with my colleagues.

Chapter 3

Activities during the Intership

3.1 Activities

My first week in Vodafone was like an adaption period. Waiting for my laptop to be set-up, it was a good chance to be informed about the CVM responsibilities and generally about Vodafone's activities. Thankfully, all my colleagues were eager to help me and show me their work in detail, making me feel a real part of the team. Each of them is responsible for a different product/service, so every day I was learning something new. For example, I devoted one day to the colleague responsible for VDSL upgrade, another day to someone responsible for the customers of CYTA, that was taken over by Vodafone, another day to the person responsible for the migration campaigns, etc. These days were so useful to get to know each other and mostly to become familiar with the systems and the practices that Vodafone uses.

Furthermore, since I had already been informed about the project that would be involved, during my spare time I seized the opportunity to study related published research with the most commonly used approaches for the Churn Model. Academics and Businesses have generated a large body of research about that topic, so I had to explore the most efficient techniques and get familiarized with the problem business-wise, as well. From the second week, my laptop was ready and I was able to start the development of the project. By combining all the suggestions and knowledge I gathered from the research I performed, I developed in collaboration with my supervisor and two more interns the approach that we would follow for the accomplishment of the project.

Apart from this project, during my internship I was often requested to provide some analytics reports on various subjects. For instance, Vodafone in the past was cooperating with Mikel in order to provide offers to its customers, – i.e. Mikel code 1+1 coffee. Now, Vodafone is cooperating with Gregorys instead of Mikel. So, I was requested to provide a report summarizing the gain of this new cooperation. Indicatively, the questions I had to answer were:

- *How many customers have stopped asking for code?*
- *How many customers that were not asking for code for Mikel, ask now for Gregorys?*
- *Are revenues increasing or decreasing?*

All the above information are saved in the Vodafone's database, in which I had not access. Thankfully, a colleague of mine was eager to help me and export the list with the required information from the database. Then, I was able to proceed to the analysis based on this list. Another time, I was requested to provide a report analyzing the performance of revenues comes from the enterprises in the last three years month by month. To accomplish this task, I was provided again with the list of revenues and then I used Python to analyze the data. In addition, when a colleague of the Product team was in need for assistance with his tasks, for instance with the monthly update of his presentations, if I was available, I would always help.

Apart from the main project and the ad-hoc tasks, in the last month of my intership I was involved in two more ongoing projects, the *Geolocation Data Project* and the *Customer Satisfaction Project*:

- **Geolocation Data Project** is a new project for Vodafone and was assigned to the Data Science team of the EBU. The main business-wise purpose for Vodafone is to take advantage of the antennas and the signals that customers' mobiles constantly broadcast. In this way, Vodafone will have millions of geolocation data, which can be used, for example to visualize traffic and solve several case studies for its business partners. So, the final goal is to convert all this information to products usefull for businesses. This project was initially running in Spain and now is in an early stage in Greece. For better understanding of the systems used in this project, my team has weekly skype-calls (hands-over) with the corresponding team in Spain.
- **Customer Satisfation Project** is another ongoing project that was assigned to the Data Science team. Every time a customer is being served by Vodafone, he receives a message to vote his service, write comments and tell if his problem finally solved. All this data are saved in a platform, called *Medallia* which provides very usefull insights about customers' satisfaction and customers' complaints. Vodafone aims to analyze those data, proceed with a sentiment analysis and make use of them in several aspects of its activities, – e.g. in retention practices, in the churn model, etc.

In parallel, employees have access to Vodafone University. Vodafone University is an internal learning platform where every Vodafone employee can search from a variety of educational materials and create an individual career learning curriculum that ideally suits to his interests and job specialization. Technology topics, digital courses, hard & soft skills trainings and many more, can

all be found and explored on this platform. In fact, I observed an intense interest of my supervisor to educate herself by spending time on this platform. Following her example, I tried to adopt the practice and utilize my free time to develop skills and learn from her.

Last but not least, it is mandatory for all the EBU employees of Vodafone to participate in the *Customers Day Event* every Thursday from 09:00 to 13:00. This event is a day that businesses work to get to know their customers better. The options we had were the following:

- **Call Center:** You can go to a call center of Vodafone and listen to live calls to customers. The employees in the call center may call customers to inform them about their products/services, to suggest them new services or upgrades, to solve possible issues, etc.
- **Vodafone Shop:** The second option was to visit a Vodafone shop and see in real-time the customer care and the customers' issues.
- **Business Agent:** The third option was to follow a 'Business Agent' around in visits to SoHo and SME companies, that use another service provider, in order to inform them about Vodafone's products.
- **Account Manager:** The last option was to follow an Account Manager around in SME and Major Enterprises of Vodafone. Account Managers are personal agents of Vodafone that interact with specific companies. Providing Enterprises with their own advisor to be informed and serviced, Vodafone achieves long standing and stable business relationships with them.

Customers days are essential for Business teams, especially when a campaign is running. In this case, they want to monitor its progress, ensure that they will reach their monthly goals and list the customers' demands. For this reason, Business teams organize twice a month some meetings with regard to customers days, in order to discuss their observations and suggest solutions.

3.2 Main Project

As mentioned before, I had the chance to participate in a large-scale project related to Machine Learning. More specifically, from the first week of my internship my supervisor assigned me the project of the end-to-end Churn Model development. For this development, I had to collaborate with her and two other interns. Throughout the internship, I worked on various tasks such as data preparation, anomaly detection, models implementation, fine-tuning of hyperparameters and model evaluation. For the development of the model Anaconda Distribution [1] and Python 3.6.5 [2] were used. More details about the project will be provided in the next chapter of this report.

[1] https://www.anaconda.com/distribution/
[2] https://www.python.org/downloads/release/python-365/

Chapter 4

Detailed Description and Results

4.1 Data Set Description

The model was focused only on SoHo companies, which have from 1 to 9 telephone lines. SME and Major Accounts require special treatment and interface only with personal agents of Vodafone, as we have already mentioned. Thus, they are not included in retention campaigns.

Generally, the ways Vodafone identifies its Business customers are two: The unique *AFM Number* of each company and the *MSISDNs*, which are the unique codes of the telephone lines of the company. For SoHo customers that have multiple telephone lines, each AFM Number contains several MSISDN codes. The particular model uses as unique feature the MSISDNs and not the general AFM Number, while a company may switch service provider for only some of its lines and not necessarily for all. Moreover, the Churn Model was requested to be implemented, as part of the mobile retention campaigns. As a result, the MSISDNs that were used for the model, were those that correspond to mobile telephone lines. Furthermore, Vodafone identifies five types of customers (four types of churners):

- **No Disc:** No Disc means No Disconnection and includes the MSISDNs that renew their contract.
- **Port Out:** Port Out means Portability Out and includes the MSIDSNs that switch service provider.
- **Disc:** Disc means Disconnection and includes the MSISDNs that completely disconnect their telephone line.
- **XPOST:** XPost includes the MSISDNs that continue with Vodafone, but change their number from Postpaid to Prepaid.
- **Bad Dept:** Bad dept includes the MSISDNs that do not pay for their service.

So, the main objective of the specific Churn Model was to identify those mobile-lines owners that are more likely to port out, those that will disconnect their mobile line, those that will change their mobile number from postpaid to prepaid and those that will be the "Bad depts". From now on, with the term "churner", we will consider anyone that belongs to one of above categories.

The first step was to built a data set with customers that sometime in the past renewed their contract with Vodafone and with customers that had been characterized as churners. As mentioned before, as interns we did not have access to the database of Vodafone, so this data set was built by our supervisor Georgina, who then make it available to us. More specifically, for the initial data set seven months were used, while churners from only one month were too few in comparison with the non-churners. So, the seven monthly data sets were merged into one and the initial size of the data set was 1,759,199 mobile telephone lines or MSISDNs (rows) and 967 features (columns). However, to built a more balanced data set, non-churners were kept from **only one month** and from the rest of months were kept only the churners examples.

The initial data set consisted of 967 features. All these features were used to characterize each MSISDN and included information related to customers' personal data, usage data, payment data, issues of customers, etc. The target variable was set as a multivalued variable, 0 for non-churners, 1 for Port out, 2 for Bad Dept, 3 for Disconnections, and 4 for XPost and thus, we had to deal with a multi-class classification problem.

Initially, we proceed to some business-wise deletions of rows. For instance, we dropped from the data set the MSISDNs that were in the *Welcome Phase*. As established in the literature review in the first part of the report, new signups should not be included in churn prediction, while the end of their subscription does not arrive soon. The same applies to those who renewed their contract in the last two months, so they were dropped as well. Apart from them, we also, dropped the owners that belong to special categories, such as the military ones, who have specific Tariff Plans and services. Last but not least, it was essential to check for the MSISDNs uniqueness and validity because the aggregation of data from multiple sources may lead to conflicts.

After these modications, the new size of the data set was 175,513 rows and 967 features. From the total number of customers, only 6,4% were the Ports out, 4% the XPost, 1,5% the Disconnections and 1,1% the Bad Dept. The rest 87% were the No Disc customers. The figure below depicts the distribution of the churners categories (class label) in the data set:

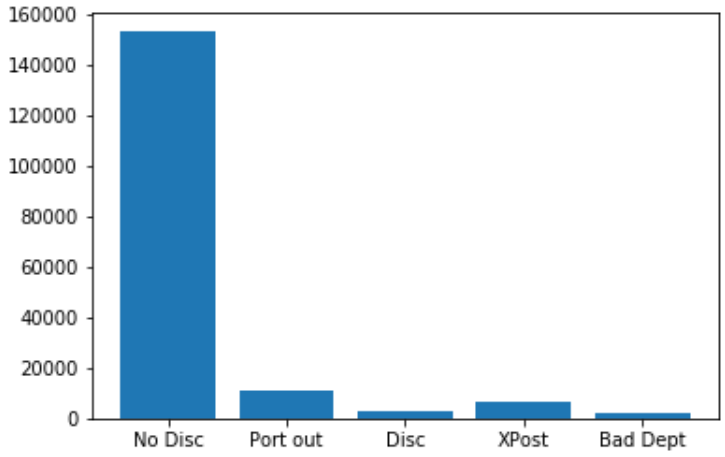


Figure 4-1: Distrubution of Churners.

As it emerges, the total number of customers that would be used for the train of the Churn Model was 175,513, of which only 13% were churners and actually splitted in four other categories. So, we had to create a model that could have identified correctly that 13% of customers, using an extremely imbalanced input.

Before we continue with the methodology that we followed for the data preprocessing step, it is worth to mention that before that, we made an in-depth analysis on the data set, for its better understanding. As mentioned in the therotical part of the report, the key element in the implementation of a model is the effective handling of the data set. Thus, in this step we delved into the features, trying to understand their mean and use, check for their distribution, their correlations and proceed to several visualizations. During this phase and due to the large number of variables, our supervisor, who is an expert in the Vodafone's customers database, thankfully suggested to us which features are the most informative in order to focus on them. The particular part of the project is unfortunately not publicly available, while it contains sensitive data about the company.

4.2 Data Preprocessing

The next step for the development of the Churn Model was the Data Preprocessing. At this stage, we mainly used Scikit-Learn library [1] provided by Anaconda. First of all, we started by dropping some markedly unsuitable features, in order to decrease the size of the data set as much as possible:

- **Drop Features with High Percentage of Missing Values:** The features that have over 85% missing values were dropped, while they did not have enough information to contribute

[1] https://scikit-learn.org/stable/

to the model. Only two features have in total more than 85% missing values, so 965 features remained. The specific percentage was chosen after several trials.

- **Drop Features having only One Value:** Following the same logic, we drop the features that had only one value to all the instances. In total there were 16 features in the data set having only one value and thus, 949 remained.

- **Drop Low-Variance Features:** Low-variance features, are those that in the majority have the same value. However, because the data set was imbalanced and non-churners hold the 95% of the data set, it would not be representative to find the low-variance features in the entire data set only. For example, the data set may contain a feature that has the same value for non-churners, but it is differentiated in the other classes. If we had dropped that feature, then we would have "lost" a feature that possibly distinguish the classes perfectly. Contrarily, the ideal case is to find features that have the same value in non-churners and distinct ones to the rest of the classes. Thus, we decided to find first the features with over 95% same value in the entire data set and then to delve into each class separately. The threshold percetage of the same value for each class was set to 90%. Using these thresholds, 226 features deleted and 723 remained.

- **Drop Highly-Correlated Features:** One more attempt to drop features and decrease the size of the data set was through the highly-correlated features, while they contain almost the same information and do not affect the model. The only case, in which correlated features are useful is when they are also correlated to the target and punctuate their existence in the model. The threshold for dropping these features was set to 95%. As a result, 135 features were dropped and 588 remained.

- **Drop Business-Wise Features:** Last but not least, we proceed to some business-wise deletions, after the analysis we had made and 582 features finally remained.

So, the size of the final data set was 175,513 rows and 582 features. After, concluding with the features that would be used in the model, the next step was the transformation of them, where it was needed. First of all, we had to deal with the existing missing values. Missing values of numerical variables were replaced with mean values of class label, missing values of objects were replaced with zero value, in order to be identified as a seperated possible value and flag variables were imputed with the mode of class label. Then, we proceed to the transformation of the features based on their values' type:

- **Numerical Features:** Numerical features, integers and floats, were rescaled using some methods of the sklearn. During this process, with tested both standardization and normal-

ization methods to check which one was fitting better in our data and we finally ended up in standardization.

- **Object Features:** As for the object features, we had to convert them into model-understandable numerical data, while most of the Machine Learning models cannot recognize any other type expect from numbers. For this convertion, we had two options: The Label-Encoder technique and the One-hot-Encoding technique. LabelEncoder can turn for example [dog,cat,dog,mouse,cat] into [1,2,1,3,2], but then the imposed ordinality means that the average of dog and mouse is cat. On the other hand, One-Hot-Encoding has the advantage that the result is binary rather than ordinal and that everything sits in an orthogonal vector space. However, even the one-hot seems to be a better choice, we finaly chose the label-encoder technique, due to the fact that our features had high cardinality and the feature space blew up, causing the Curse of Dimensionality problem.

Apart from the above techniques, we tried and others that did not perform adequately well. For instance, in order to deal with the imbalanced distributions of the examples across the known classes, we used several resampling methods, like over-sampling and down-sampling of the sklearn library, but the results were not improved and thus, they were not implemented at the end. Moreover, we attempted to implement Dimensionality Reduction, using *Principal Components Analysis (PCA)* method as well as *SelectKBest* method, which means select features according to the k highest scores, of the sklearn library as well. Both of these methods, reduced the models' performance, so they were not included in the process.

Lastly, we tried to detect outliers in the data set in order to be excluded from the training phase of the model. The ideal result of that detection would be to identify outliers mainly in the churners classes. Then, the instances of these classes would be reduced, but they would also become cleansed and distinguishable. Thus, it would be easier for the model to learn their pattern and then any mistake in these categories would be like an outlier. However, we faced two basic problems with this approach: The fist problem was that the most outliers were belonging in the majority class of non-churners and the second one was that the samples of the minority classes were already too few, so we could not reduce them a lot.

4.3 Models Implementation

The next step in the process, was the implementation of the models. To train the models we used the 80% of the data set as the training set and the rest 20% of the data set as the test set. The training data set comprised of 140,410 customers, of whom 6,4% were the Port out, 1,1% were the Bad Dept, 1,5% the Disc and 4% the XPost. The test set comprised of 35,103 customers, of whom 6,4% were the Port out, 1,1% were the Bad Dept, 1,5% the Disc and 4% the XPost.

The methods used in the model were LightGBM [2], Random Forest [3] and Sequential Neural Network Model [4] and a model was created for each of the chosen technologies. Each of the models was constructed using the same training data set.

Unfortunately, while these models were run locally, we did not have the computational power to support Cross-Validation technique for the rotation of the data. For the same reason, we could not implement very complex models, as they require a careful parameter tuning and our laptop could not afford methods like Grid Search, that scan the data to configure the optimal parameters. As a result, we had to run the models for different parameters mannualy. Certainly, we tried and other methods such as XGBoost, Decision Trees that did not perfom well and we tried to implement the SVM Classifier, but it was interrupted due to its endless execution time.

In order to evaluate our models more precisely, we also, loaded a validation data set that contained a new single month. The purpose was to make predictions on that month, as if it was for a real monthly retention campaign. The validation data set comprised of 171,318 customers, of whom 1,2% were Port out, 0,15% Bad Dept, 0,3% Disconnections and 0,6% XPost. We followed the same process as for the initial data set, in order to be in the same format and with the same features and then after we have trained the models, we tested them on that.

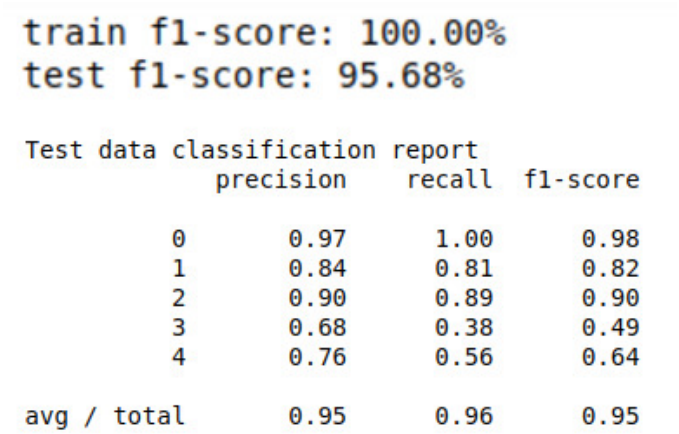
	Training Set	**Test Set**	**Validation Set**
No Disc	87%	87%	97,7%
Port out	6,4%	6,4%	1,2%
Bad Dept	1,1%	1,1%	0,15%
Disc	1,5%	1,5%	0,3%
XPost	4%	4%	0,6%
	140,410	35,103	171,318

The model that we finally concluded after several trails, was a Voting Classifier of LightGBM and Random Forest. We used Random Forest Classifier, because in the validation data set achieved better results in the non-churners category and we combined it with LightGBM that performed well on the rest of categories. The code for the implementation of the Churn Model is unfortunately not publicly available. However, I will present some of the results of the model in order to taste of its performance. The metrics that were used for the evaluation of the models were the confusion matrix and the classification report that summarizes the recall and the precision of each class. However, the confusion matrix will not be presented, while contains analytical information of the missclassified data.

[2] https://lightgbm.readthedocs.io/en/latest/
[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[4] https://keras.io/getting-started/sequential-model-guide/

4.4 Results

The results on the Test Data set are briefly presenting in the bellow figure:



```
train f1-score: 100.00%
test f1-score: 95.68%

Test data classification report
             precision    recall  f1-score

          0       0.97      1.00      0.98
          1       0.84      0.81      0.82
          2       0.90      0.89      0.90
          3       0.68      0.38      0.49
          4       0.76      0.56      0.64

avg / total       0.95      0.96      0.95
```

Figure 4-2: Voting Classifier of LightGBM & Random Forest on Test Data Set.

As we can see, the specific Churn Model predicts very good the customers that Port out (class:1), the customers that are Bad Dept (class:2) and so-so the XPost (class:4). Its lowest performance results from the Discnonnection Customers (class:3), which was expected to some degree due to their undefined nature. For instance, the may be customers that left to abroad and disconnect their line (on another occasion they may have renewed their contract) or customers that cannot afford the service anymore and prefer to disconnect their line, instead of not paying, or customers that want to change service provider, but they want to change their number as well.

In any case, the results are quite adequately, while Vodafone is not able to contact all these customers and will target a portion from each category anyway. So, it does not matter if they cannot predict them all correctly. The important metric, as we have already mention is the precision. According to the above classification report, even for a Disconnection campaign (that Disconnections do not perfom as well as the other categories), if they target 100 customers, they will find 68, which is a good ratio.

As for the results on the Validation Data Set, the performance of the model was in lower levels. More specifically, it assigned many customers belonging to non-churners category into the other categories. As a result the precision of the churners was reduced, while the recall increased. This tendency in the performance of models in the validation set is a usual phenomenon in Machine Learning, while they have to do with a totally new data set that corresponds to real conditions. Remember that we trained a model using churners from seven months and we test it on data set from a single month.

However, even in this case the model will probably bring the hoped-for results. Vodafone does not aim to reduce its churn rate by 100%. Even a 5% decrease in churn rate, will rapidly increase its revenues. So, we splitted the results into quartiles and by using a threshold in the prediction rate for each category, we finally targeted the top quartile that belongs to. The precision for the top 20% of each class label was around 85%. The threshold will be set-up based on the number of churners a particular campaign aims to "save". The fewer churners a campaign aims to save, the higher the precision rate will be.

4.5 Project Tools

For the development of the above project, the following tools were used:

Environment	Anaconda
Programming Language	Python 3.6.5
Editor	Jupyter Notebook
Data set Analysis	Pandas
	Statistics
	Seaborn
	Matplotlib
Data Preprocessing	NumPy
	SKlearn
Models Implementation	Sklearn
	LightGBM
	Keras
Evaluation Metrics	Sklearn

Table 4.1: Project Tools.

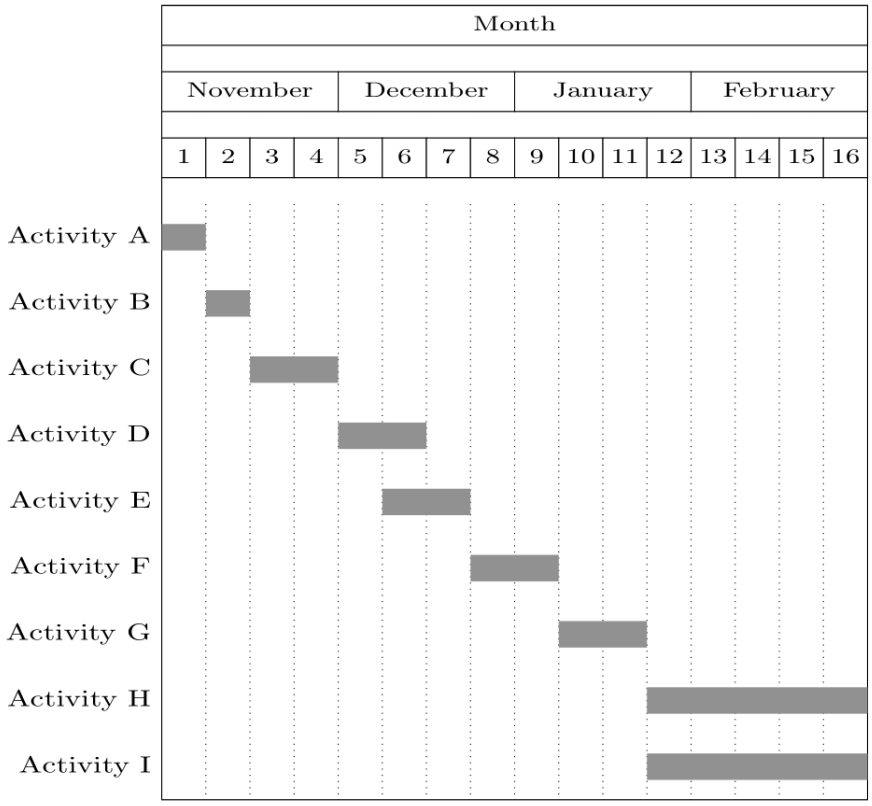
Chapter 5

Time Management

My internship at Vodafone started on *November 18th, 2019* and was completed on *February 17th, 2020*. The time duration of all activities is depicted in the following timetable:

Task	**Duration**
Research Conduction	1 week
Data Set Analysis	1 week
Data Preparation	2 weeks
Models Implementation	2 week
Model Validation	2 weeks
Model Evaluation	2 week
Churn Model Presentation	2 week
Medallia Project	5 weeks - Until now
Geolocation Data Project	5 weeks - Until now

Table 5.1: Tasks During the Internship.



Activity A: Reaserch Conduction
Activity B: Data Set Analysis
Activity C: Data Preparation
Activity D: Models Implementation
Activity E: Models Validation
Activity F: Model Evaluation
Activity G: Churn Model Presentation
Activity H: Medallia Project
Activity I: Geolocatio Data Project

Figure 5-1: Gantt Chart (Weekly View).

Chapter 6

Skills

The following knowledge and skills that were developed in the context of the University courses were utilised and further improved during my internship at Vodafone:

Skill	Related Methods	Example
Project Management	Gantt chart	Planning of activities/trainings, better management of deadlines
Machine Learning Algorithms	Logistic Regression, Decision Trees, Random Forest, LightGBM, Neural Networks	
Programming Languages	Python	Implementation of the Churn Model in Python
Machine Learning Libraries	NumPy, Pandas, Sklearn, Keras	
Database Management Systems	SQL	
Statistics		Distributions of variables, correlations of variables, probabilities
Analytics Tools	Excel, Medallia	Customer Satisfaction Analysis
Visualization Tools	Power BI, QGIS	Viewing, editing, analysis of geospatial data & exporting graphical maps

Table 6.1: Skills.

Chapter 7

Conclusion

Working at Vodafone was a pleasant experience. The headquarters are located in the suburb of Chalandri, Athens and host the majority of the employees. The interiors of the buildings are *Google-like* designed in a modern and playful style with the use of wallpapers, colorful furnitures, table-soccers, ping-pong tables and analytics dashboards. The offices follow an open-space layout, where big tables allow teams to collaborate by sitting nearby and changing easily workspace. People from all hierarchy levels work at the same workplace. Each employee is equipped with a personal computer given thus the freedom to work at different places. Additionally, there are several glass walled conference rooms on each workspace, which allow teams to discuss for their progress, have a call with a partner, make a presentation or celebrate a team's achievement. It is certainty a work environment that motivates employees to work harder and be creative.

Apart from that, team of CVM was truly amazing. All the colleagues were always willing to provide me with assistance, guidance and explain me their work in detail. Their cooperation and generally their relationships are admirable, anything like someone would expect to see in a large multinational company. They make me feel comfortable and part of their team from the first moment. In parallel, my supervisor provided me with a variety of knowledge, skills and best practices. Her commitment, passion and diligence on programming was a continuous motivation for me to work harder. This internship was a big opportunity for me to learn about myself, apply the knowledge acquired from my studies and improve my technical skills, as well my soft skills.

Finally, the most significant experience I acquired through my internship was the continuous sense of responsibility. Despite having previously worked on multiple and diverse University projects, it was the particular project I implemented at Vodafone that transferred me the great sense of responsibility due to the fact that, for the first time, a project I worked on would be evaluated in production. Knowing that your work will be assessed not only by your supervisor but also by clients, and have an actual impact on them, provides a different perspective on your project.

References

[Acuña and Rodriguez, 2004] Acuña, E. and Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In Banks, D., McMorris, F. R., Arabie, P., and Gaul, W., editors, *Classification, Clustering, and Data Mining Applications*, pages 639–647, Berlin, Heidelberg. Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-17103-1_60.

[Ascarza et al., 2018] Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., Hardie, B. G., Lemmens, A., Libai, B., Neal, D., et al. (2018). In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5(1-2):65–81. https://www.hbs.edu/faculty/Publication%20Files/ascarza_et_al_cns_17_e08d63cf-0b65-4526-9d23-b0b09dcee9b9_538a6ea6-a480-4841-b9f0-a87be24989ba.pdf.

[Athanassopoulos, 2000] Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research*, 47:191–207. https://doi.org/10.1016/S0148-2963(98)00060-5.

[Bishop, 1995] Bishop, C. M. (1995). *Pattern Recognition and Machine Learning*. Oxford University Press.

[Boschetti and Massaron, 2018] Boschetti, A. and Massaron, L. (2018). *Python Data Science Essentials: A practitioner's guide covering essential data science principles, tools, and techniques, 3rd Edition*. Packt Publishing Ltd. https://books.google.gr/books?id=cP1wDwAAQBAJ.

[Brownlee, 2018] Brownlee, J. (2018). When to use mlp, cnn, and rnn neural networks. https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. https://dl.acm.org/doi/pdf/10.1145/2939672.2939785?download=true.

[Detko, 2019] Detko, D. (2019). Customer churn analysis: How to retain customers using machine learning. Available at Predica Group, https://www.predicagroup.com/blog/customer-churn-analysis/. [Online; accessed Feb 16, 2020].

[Emerson et al., 2019] Emerson, S., Kennedy, R., O'Shea, L., and O'Brien, J. (2019). Trends and applications of machine learning in quantitative finance. *8th International Conference on Economics and Finance Research (ICEFR 2019)*. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3397005.

[Erfani et al., 2016] Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121 – 134. http://www.sciencedirect.com/science/article/pii/S0031320316300267.

[Ertekin et al., 2007] Ertekin, S., Huang, J., Bottou, L., and Giles, L. (2007). Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, page 127âĂŞ136, New York, USA. Association for Computing Machinery.

[Farhaoui and Moussaid, 2019] Farhaoui, Y. and Moussaid, L. (2019). *Intelligent Big Data Analysis to Design Smart Predictor for Customer Churn in Telecommunication Industry*. Studies in Big Data. Springer International Publishing. https://books.google.gr/books?id=hz0GwAEACAAJ.

[Glen, 2019] Glen, S. (2019). Roc curve explained in one picture. Available at Data Science Central: https://www.datasciencecentral.com/profiles/blogs/roc-curve-explained-in-one-picture. [Online; accessed Feb 10, 2020].

[Goldstein and Uchida, 2016] Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4):1–31. https://journals.plos.org/plosone/article%3Fid%3D10.1371/journal.pone.0152173.

[Guo et al., 2003] Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In Meersman, R., Tari, Z., and Schmidt, D. C., editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 986–996, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Gupta, 2017] Gupta, D. (2017). Fundamentals of deep learning – activation functions and when to use them? https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/. [Online; accessed Feb 7, 2020].

[Harlalka, 2018] Harlalka, R. (2018). Choosing the right machine learning algorithm. https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f.

[Hochreiter, 1998] Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

[Huilgol, 2019] Huilgol, P. (2019). Accuracy vs. f1-score. Available at Medium: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2. [Online; accessed Feb 9, 2020].

[Japkowicz, 2006] Japkowicz, N. (2006). Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning*, pages 6–11. https://www.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-003.pdf.

[Jordan and Mitchell, 2015] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260. https://science.sciencemag.org/content/349/6245/255.

[Kersting, 2018] Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*. https://doi.org/10.3389/fdata.2018.00006.

[Khurana, 2019] Khurana, Y. (2019). Difference between model validation and model evaluation? https://medium.com/yogesh-khuranas-blogs/difference-between-model-validation-and-model-evaluation-1a931d908240. [Online; accessed Feb 18, 2020].

[Klepac et al., 2014] Klepac, G., Kopal, R., and Mršić, L. (2014). Churn problem in everyday business. In Fagerberg, J., Mowery, D. C., and Nelson, R. R., editors, *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*, Research essentials, chapter 1, pages 1–25. IGI Global. https://books.google.gr/books?id=oaR_BAAAQBAJ&pg=PA1&dq=voluntary+and+nonvoluntary+churners&hl=el&source=gbs_toc_r&cad=4#v=onepage&q=voluntary%20and%20nonvoluntary%20churners&f=false.

[Kotsiantis et al., 2006] Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190. https://doi.org/10.1007/s10462-007-9052-3.

[Kundu, 2018] Kundu, A. (2018). Machine learning powered churn analysis for modern day business leaders. https://towardsdatascience.com/machine-learning-powered-churn-analysis-for-modern-day-business-leaders-ad2177e1cb0d. [Online; accessed Feb 16, 2020].

[Lucket and Schaefer-Kehnert, 2016] Lucket, M. and Schaefer-Kehnert, M. (2016). Using machine learning methods for evaluating the quality of technical documents. Master's thesis, Linnaeus University. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A920202&dswid=5594.

[Manyika et al., 2011] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. (2011). Big data: The next frontier for innovation, competition and productivity. Report, McKinsey. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx.

[Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. The MIT Press. https://pdfs.semanticscholar.org/e923/9469aba4bccf3e36d1c27894721e8dbefc44.pdf.

[Moritz and Bartz-Beielstein, 2017] Moritz, S. and Bartz-Beielstein, T. (2017). imputets: time series missing value imputation in r. *The R Journal*, 9(1):207–218. http://cran.seoul.go.kr/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf.

[Mozer et al., 2000] Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., and Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, 11(3):690–696. https://www.researchgate.net/publication/3302790_Predicting_subscriber_dissatisfaction_and_improving_retention_in_the_wireless_telecommunications_industry.

[Murphy,] Murphy, L. Saas churn rate: What's acceptable? Available at Sixteen Ventures: https://sixteenventures.com/saas-churn-rate. [Online; accessed Feb 14, 2020].

[Narkhede, 2018] Narkhede, S. (2018). Understanding auc - roc curve. Available at Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Online; accessed Feb 10, 2020].

[Navlani, 2020a] Navlani, A. (2018 [accessed February 9, 2020]a). Knn classification using scikit-learn.

[Navlani, 2020b] Navlani, A. (2019 [accessed February 9, 2020]b). Support vector machines with scikit-learn.

[Navlani, 2020c] Navlani, A. (2019 [accessed February 9, 2020]c). Understanding logistic regression in python.

[Neslin et al., 2006] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn

models. *Journal of marketing research*, 43(2):204–211. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.489.5495&rep=rep1&type=pdf.

[Powers, 2011] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. https://dspace2.flinders.edu.au/xmlui/bitstream/handle/2328/27165/Powers%20Evaluation.pdf?sequence=1&isAllowed=y.

[Prechelt, 2012] Prechelt, L. (2012). Early stopping — but when? In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 53–67, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Ragel and Crémilleux, 1998] Ragel, A. and Crémilleux, B. (1998). Treatment of missing values for association rules. In Wu, X., Kotagiri, R., and Korb, K. B., editors, *Research and Development in Knowledge Discovery and Data Mining*, pages 258–270, Berlin, Heidelberg. Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/3-540-64383-4_22.

[Saleh,] Saleh, K. Customer acquisition vs.retention costs – statistics and trends. https://www.invespcro.com/blog/customer-acquisition-retention/. [Online; accessed Feb 15, 2020].

[Serengil, 2017] Serengil, S. I. (2017). Softplus as a neural networks activation function. https://sefiks.com/2017/08/11/softplus-as-a-neural-networks-activation-function/.

[Sharmistha, 2019] Sharmistha, C. (2019). A comprehensive study of linear vs logistic regression to refresh the basics. https://towardsdatascience.com/a-comprehensive-study-of-linear-vs-logistic-regression-to-refresh-the-basics-7e526c1d3ebe. [Online; accessed Jan 27, 2020].

[Smolyakov, 2017] Smolyakov, V. (2017). Ensemble learning to improve machine learning results. Available at Medium: https://blog.statsbot.co/ensemble-learning-d1dcd548e936. [Online; accessed Feb 11, 2020].

[Sun et al., 2015] Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., and Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, 48(5):1623 – 1637.

[SuperDataScience, 2018] SuperDataScience (2018). Recurrent neural networks (rnn)-the vanishing gradient problem. https://www.superdatascience.com/blogs/recurrent-neural-networks-rnn-the-vanishing-gradient-problem. [Online; accessed Feb 9, 2020].

[Tate, 2020] Tate, A. (2020). How to calculate customer churn rate (the best saas churn formula). https://www.profitwell.com/blog/the-complete-saas-guide-to-calculating-churn-rate-and-keeping-it-simple. [Online; accessed Feb 14, 2020].

[Vafeiadis et al., 2015] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., and Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9. https://doi.org/10.1016/j.simpat.2015.03.003.

[Van Der Maaten et al., 2009] Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13. https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf.

[Walia, 2017] Walia, A. S. (2017). Activation functions and it's types-which is better? Available at Towards Data Science: https://towardsdatascience.com/activation-functions-and-its-types-which-is-better-a9a5310cc8f.

[Xia and Jin, 2008] Xia, G. E. and Jin, W. D. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering-Theory & Practice*, 28(1):71–77. https://www.sciencedirect.com/science/article/pii/S187486510960003X.

[Zhang et al., 2003] Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17:375–381. https://www.researchgate.net/publication/220355854_Data_Preparation_for_Data_Mining.