Universitat Politecnica de Catalunya (UPC)

Barcelona, Spain

Master Thesis

Prediction of Customers' Churn in Telecommunication Industry

Author: Lorisa Perdoci

Thesis Advisor: Luis Jofre-Roca

Contents

GLOSSARY

Some important indexes in the telco industry are:

Term	Definition
ARPA	(Average Revenue per Account) = Total Revenue / Number of Accounts
ARPU	(Average Revenue per Unit) = Total Revenue / Number of MSISDN
Bundle	Ways that telecom companies package (bundle) their services
Churn Rate	No. of Customers Churning / (Closing Base-Opening Base)
KPI	Key Performance Indicator
Margin of Postpaid Customers	Total Postpaid Customers / Total Customers
MSISDN	Mobile number
New Activations	New customers for a company
Port-In	Customers who come to an Operator from another Operator
Port-Out	Customers who leave the Operator to find another
Postpaid	Customers with fixed term contract
Prepaid	Customers without a contract who pay in advance for the service provided by the operator
Tenure	Time of customer in the network

ABSTRACT

In the developed world, mobile markets have reached saturation on subscriber penetration and connections growth. The challenge for operators has evolved from attracting new customers to retaining existing ones. Various components have an impact on churn. Therefore, it is very important to understand the behavior of the customers, encourage them in spending more and then predicting the future by preventing their attrition. As the industry is evolving, the biggest challenge for operators is to engage with consumers and retain their loyalty by delivering more competitive and innovative value-added services.

While understanding consumer needs remains essential to improve customer retention, other emerging tariffs and services are likely to carry a long-term impact on churn (including national, international and roaming bundles tariffs and mobile services). The churn might be voluntary in cases they want to leave the network they actually are using, or involuntary churn in case of unpaid bills. The methodology used to do the right evaluations in order to achieve strong results in this field is very large and varied. The scope of this thesis is to identify and analyze different appropriate models that can help the data analysts to find the churners in Telecommunication industry.

In this thesis we are going to discuss on two important topics in telecommunication markets and their respective predictive models, which tend to understand the customer behavior towards different competitors: market share in telecommunication industry and customer churn.

1. INTRODUCTION

The growing demand and consumption of data services is leading the scientist community in analyzing and finding new models that better fit to those data. Mobile telephony is highly used by consumers and enterprises as well. As traffic increases, customer and network data increases as well. Furthermore, traffic carried on telco networks is becoming increasingly rich, with voice being supplanted by data services such as streaming audio and video. LTE deployments and associated marketing campaigns have boosted the consumption of data services. This increased data traffic comes from a variety of sources (smartphones, TVs, tablets, and laptops) and multiple channels (social media, web chat, email, and voice calls).

Machine-to-machine (M2M) communications have also increased the data service usage as customers use the network to control devices in places like the home or their car. Meanwhile, the telco's companies also have operational data such as billing, network, location data, and call detail records, presented in a structured format, which is typically contained in SQL databases. Semi structured and unstructured data such as call logs, social media messages, text messages, emails, customer feedback documents, system logs, and sensor data is present. The rate of data growth will exceed the capacity of Telco's existing data warehouses.

To support internal decision-making processes, these new and varying datasets need to be ingested by storage platforms and processed using analytic tools to gain insights. The insights will inform how to drive increased ARPU (Average Revenue per Unit), predict and reduce customer churn, deliver improved customer experience through personalized services, and limit the operational costs associated with network management (through network design, planning, and optimization).

So, the main focus of this thesis is to understand the telecom customer behaviour towards different competitors while studying the market share in telecommunication industry and customer churn. This will be done while exploring different machine learning algorithms like LSTM and XGBOOST. The major aim of this study is to identify churners so that the retention strategies could be targeted upon them and the company may flourish by maximizing its overall revenue.

Methodology includes the main concepts that are needed in order to work out this thesis. The scope is to be useful when we use the methodology to the applied statistics. The second chapter will give a detailed analysis of the models chosen to cover the three topics mentioned above. A comparison of the chosen models and others taken in consideration during this thesis is done.

Chapter 3 is the core of this thesis, where the analysis and the results of the different models are included. We are going to divide this section into two sub sections corresponding to our two areas of research. For each of them it will be explained in detail the identification of variables, a descriptive analysis, the implementation of the chosen machine learning model and the results

Finally conclusions and future work is discussed. Many different adaptations, tests, and experiments have been left for the future due to the need to focus on more conceptual approaches (i.e. the experiments with real data are usually very time consuming, requiring even days to finish a single run). Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or simply curiosity.

2. METHODOLOGY

This section considers the main pillars of the master's thesis in order to deal with the research of the identification and analysis of the methodologies used to estimate the customer attrition in the industry of the telecommunication. Descriptive Statistics is the very first discipline that will help us to evaluate the data we have in our possession. Indicators are the tools that can measure the behaviour of a phenomenon that is considered representative for the analysis and are used to monitor or evaluate the degree of success or adequacy of the activities implemented. A strong point to be focused on is the information system, without which every single calculation would be too complex to calculate. Also it is important a deep knowledge in the field of Mathematical Analysis and Marketing. The second one helps us formulate the needs of any business. Below we are going through the datasets and respective machine learning models for each of the main topics of this thesis.

2.1. METHODOLOGY GENERAL CONCEPTS

So far, we have done an evaluation of all the knowledge we have gained in order to advance in our research. That is why the study is focused in standard disciplines combined with the applied statistics. As we have reaffirmed above in Telecommunications industry it is very important to understand the behaviour of the customers, encourage them in spending more and then predicting their future by preventing their attrition.

The methodology used to do the right evaluations in order to achieve strong results in this field is very large and varied. Descriptive Statistics is the very first discipline that will help us evaluate the data we have in our possession. Graphical analysis is the main representation of the analytics. Inferential statistics will help examining the entire population instead of a sample. The sampling techniques allow us optimizing the sample extraction criteria so that we can obtain the same information from the sample, which would have been obtained by having the entire collective. Sample is a piece of data taken from the whole data which is continuous in the time domain. Therefore, that is the reason why sampling techniques is one of the most important disciplines that

we will refer in this paper. Sampling is defined as “The process of measuring the instantaneous values of continuous-time signal in a discrete form.”

Methods of collecting, summarizing, analyzing, and interpreting variable numerical data are summarized in the chapter Statistical Methods. They can be contrasted with deterministic methods, which are appropriate where observations are exactly reproducible or are assumed so. KPIs or Key Performance Indicators are the tools that can measure the behavior of a phenomenon that is considered representative for the analysis and are used to monitor or evaluate the degree of success or adequacy of the activities implemented. A strong point to be focused in is the information system, without which every single calculation would be too complex to calculate. Considerable should be the knowledge in the field of Mathematical Analysis and Marketing. Marketing knowledge helps us formulate the needs of any business.

2.1.1. Descriptive and inferential statistics

Descriptive statistic is the discipline in which the methodologies used by a tester are studied to collect, represent and process the observed data for analysing a certain phenomenon. [1]

Data may be quantitative or qualitative. The data is sorted in statistical tables. Graphic representations: Bar graph, pie chart, histogram. Examples of some graphical representations can be seen in Figure 2.1. The analytical representation of the theoretical distributions is to find an interpolating mathematical function that adequately represents an observed statistical phenomenon. Its purpose is to get a general view of the data and the distributions of the variables by diagrams, tables, and basic statistics, such as mean and standard deviation. The descriptive analysis is a necessary part of the research and is always conducted before doing any statistical tests or more complicated modelling. This part presents some common techniques for descriptive data analysis.

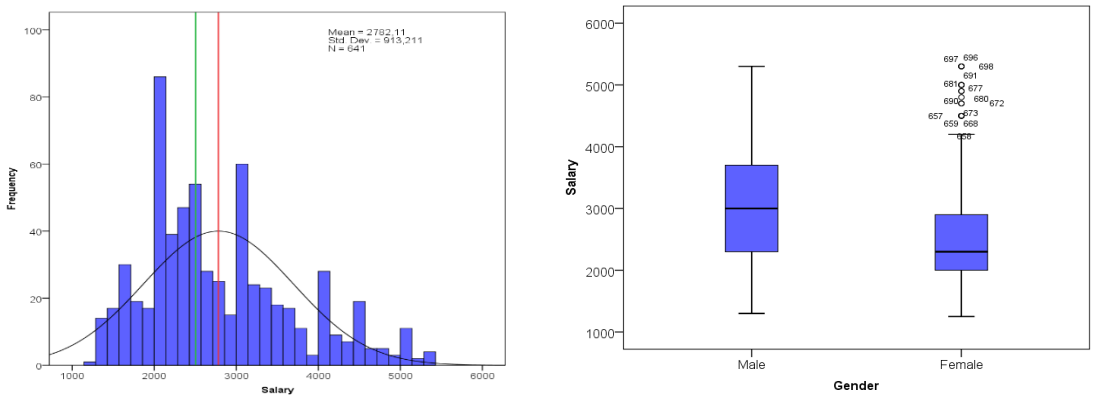

Figure 2.1. Example of a graph plot and box plot

Regression: examines the linear relation between one or more explanatory variables (or independent ones) and a criterion (or dependent) variable: $Y = \alpha + \beta X$. By correlation is meant a relationship between two statistical variables such that each value of the first variable matches with "certain regularity" the value of the second one. The degree of correlation between two variables is expressed by the so-called correlation indices which assume values between - 1 and + 1. We refer to descriptive statistics when addressing a direct population survey, instead of statistical inference when starting from examining a sample to have information about the entire population. Addressing a statistical inference problem should refer to a Model. The three main techniques used in the inference are theory of estimation, use of confidence intervals and test theory. In probability, it is considered a phenomenon that can be observed solely from the point of view of the possibility or not of its occurrence, regardless of its nature. Between two extremes, known as event and event impossible, there are more or less likely events.

A random variable X is a numeric variable whose measured value may change by repeating the same measurement experiment. X can be a continuous or discrete variable. The maximum information that can be given to questions of the type: what is the probability that in a future measurement the value of X is between a and b, where a <b?

These probabilities identify the distribution of the random variable. In the field of statistical inflection, two schools of thought are distinguished, linked to different concepts, or interpretations, of the meaning of probability: Classical / frequentist inferences and Bayesian inferences. A probability distribution is a mathematical model that links the values of a variable to the probability that these values can be observed. Formally, probability distributions are expressed by a

mathematical law called probability density function (indicated by f(x)) or probability function (indicated by p(x)) respectively for continuous or discrete destruction. Discrete distribution examples are Binomial, Poisson etc. Instead of continuous distribution are the Normal, Exponential, and Weibull.

2.1.2. Sampling techniques

Statistical data may come from the following sources: censuses, sample surveys, and data processing collected within administrative procedures, that is, administrative source data. Statistical surveys - performed on a sample or total population - are developed according to a process that, starting with the definition of detection targets, collects and processes the data and concludes with the analysis and dissemination of the results. The sampling techniques allow optimizing the sample extraction criteria so that we can obtain the same information from the sample, which would have been obtained by having the entire collective. In this way one can get the same information, with costs, however, significantly lower and, often very important, with extremely quick times (e.g. electoral projections).

By sample is meant that group of elementary units, a particular subset of the population, identified in it so as to allow, with a definite risk of error, the generalization of analysis results to the whole population [2]. A sampling plan defines a method by which you select items that are part of the sample. A first major distinction to note is that of probabilistic samples and non-probabilistic samples. The probabilistic sample selection methods can be different. They are distinguished by: simple random sample, stratified sample, cluster sample, systematic sample, multi-stage sample, etc. Depending on how the sample units are selected.
The parameters of interest in the population are usually the average and the total. The estimators that are often used are those of Hansen-Horvitz and Horvitz-Thompson. Correctness and efficiency evaluate the bounty of the estimators used.

2.1.3. Statistical methods for business

The analysis of historical series includes a series of statistical methods to investigate the historical series, determine the process underlying it, and make predictions. According to the traditional approach, it is assumed that the process has a deterministic part, which allows it to be dissociated in trendy, cyclical and/or seasonal components, and that the difference between the theoretical data

of the deterministic model and the observed data is attributable to a residual random component [3]. According to the modern approach, however, it is assumed that the process described has been generated from a stochastic process described by a parametric probabilistic model. The groups should be unit assemblies on the one hand as homogeneous as possible and on the other as separate as possible. This suggests introducing distance indices so as to clarify the notion of proximity and homogeneity. Cluster Analysis (CA) consists of a set of statistical techniques to identify groups of units similar to a set of characters taken into account, and according to a specific criterion. The objective that we set ourselves is basically that of bringing together heterogeneous units into more subordinate and mutually exhaustive subsets. The statistical units are, in other words, subdivided into a number of groups depending on their level of "resemblance" from the values that one Series of preset variables assumes in each unit.

The joint measurement analysis and a multivariate analysis technique that takes into account consumer preferences in the choice of goods and services. Through Conjoin Analysis you can check: the degree of relevance to each level or mode of each feature, and the importance each individual attributes to a feature of a product or service [4].

Performance Plan - The Plan's definition process has followed some logical phases: defining the history, the current and the identity of the organization; Analysis of the external and internal context; Definition of strategic objectives and strategies; Definition of operational objectives and operational plans; the development of programming and control systems and the improvement actions to be promoted.

The index is a ratio between two numbers and is intended to compare two entities. The indicator is the sum of one or more factors such as GDP (Gross Domestic Product), which will affect the final result from the sum of the individual factors. Indicators are tools that can show (measure) the behavior of a phenomenon that is considered representative for the analysis and are used to monitor or evaluate the degree of success or adequacy of the activities implemented.

2.1.4. Bayesian Inference

The Bayesian approach might have a very important role in Telco industry due to computational reasons. Epistemological reasons and pragmatic reasons are also considerable reasons that guide us in our study. As explained by Brunero Liseo in “Introduzione alla statistica Bayesiana” (Liseo,

2008), from an epistemological point of view the reasons for using this method are based on a simple and direct inductive reasoning method, according to the information available on a certain set of phenomena, in a certain moment of life that wants to calculate the probability of future events or, more generally of events for which it is not known whether they are verified or not [5].

Bayesian logic is consistent with very logical basis and free of risk counterexamples, always waiting for innovation when it is used the method of induction, and it is necessary to produce statements of probabilistic nature of events that we do not know if it will happen or less. Pragmatic reasons are related to the need to take in consideration the extra-experimental information of the problem that need to be solved [6]. That refers to the Bayesian setting. In telecommunication, for example, when assessing the probability that a customer might leave the network due to the reduction of some particular offer those that are the a-priori probabilities (extra-experimental info) and are nothing else but the information on the specific offer we need to include in our problem solving.

Also very useful in this sector is to have the information at a level of disaggregation sufficiently high. This need goes under the name of "small area estimation" that refers to the difficulty of producing information for areas of which we do not have access to the sample. So, estimating the possibility to churn for a single customer that belongs to a sample of a company for which we do not have data, might be possible using the Bayesian method. So, an intrinsic characteristic of the Bayesian method is precisely that of being able to assume, in a simple and naturalway, different levels of association between the units of the sample, allowing the phenomenon of "borrowing strength" which allows the production of estimates sufficiently stable for those areas with no sample data.

The Bayesian Method gives the possibility to integrate, using Bayes theorem, all the information generated by the statistical experiment with the "a priori" data. Monte Carlo methods, based or not on the properties of Markov chains, gives the possibility to generate a sample of whatever dimensions, independent identically distributed by the distribution a posteriori of the parameters we are interested in. That's why in a very large contest the Bayesian approach permits the flexibility of a model which is very difficult to be achieved through classical methods.

Figure 2.2. shows clearly the concept of borrowing strength (Vaart, 2013) [7]. In high dimensions, as we are considering telco data, the potential gain is large. A-priori knowledge should make this gain even bigger.

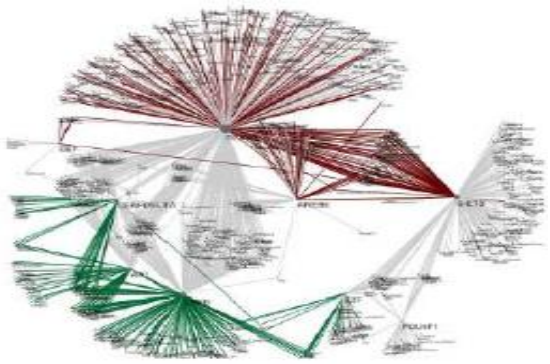

Figure 2.2. Borrowing strength

For each of the topics to be treated in this master thesis (market share in telecommunication industry and customer churn) we have chosen two respective machine learning models.

2.2. MARKET SHARE PREDICTION

Market share is the percent of total sales in a sector generated by a particular company. It is calculated by taking the company's sales over the period and dividing it by the total sales of the sector over the same period. Telecommunication industry is a fast pacing one, so having a look continuously at the market share and trying to forecast the near future helps them respond to the changes in market share using different campaigns to attract their customers to be able to retain them, but also get new customers over time.

The market share is being studied on a 1-year dataset with daily data.

2.2.1. Dataset

Market share seems like a pretty straightforward problem. But the challenge for every company, especially in the telecommunication industry, is how to tell exactly how the competitor is doing

with their sales. So, we know exactly the new customers of the company we might be working for, but not those of the competitors.

That is why we are going for a different approach. Let's suppose a given market has 3 different telecom providers: operator1, operator2 and operator3. I am doing this research with data provided from operator1 telecom provider. What I have in my database are incoming calls from operator2 and operator3 users towards an operator1 customer, and calls from operator1 towards operator2 and operator3. So, we can build a database like the one shown in Table 1.

Date (daily)	Number of calls (operator 1 to operator 2)	Number of calls (operator 1 to operator 3)	Number of calls (operator 2 to operator 1)	Number of calls (operator 3 to operator 1)
03/01/2021	2526	832	2611	503
04/01/2021	2657	745	2682	675
05/01/2021	2816	750	3149	631

Table 2.1. Dataset sample of Market Share

What we are interested though would be to track calls from numbers which haven't appeared previously in our database within that specific operator. This way we are tracking at once the movement of customers from operator1 to operator2 or operator3 (also known in telecommunication industry as port-in/port-out) and new customers as well.

We have collected one year of daily data which was pre-processed to turn the above information in market share percentage of all three operators. The sample of this data can be seen in Table 2.1.

2.2.2. Machine learning Model Identification

In the market share machine learning problem we have to deal with a time series one.

In practice a suitable model is fitted to a given time series and the corresponding parameters are estimated using the known data values. The procedure of fitting a time series to a proper model is termed as Time Series Analysis [8]. It comprises methods that attempt to understand the nature of the series and is often useful for future forecasting and simulation.

In time series forecasting, past observations are collected and analysed to develop a suitable mathematical model which captures the underlying data generating process for the series [9, 10].

The future events are then predicted using the model. This approach is particularly useful when there is not much knowledge about the statistical pattern followed by the successive observations

or when there is a lack of a satisfactory explanatory model. Time series forecasting has important applications in various fields. Often valuable strategic decisions and precautionary measures are taken based on the forecast results. Thus making a good forecast, i.e. fitting an adequate model to a time series is very important. Over the past several decades many efforts have been made by researchers for the development and improvement of suitable time series forecasting models.

To solve this time series problem the below machine learning models were taken into consideration:

- ***ARIMA (AutoRegressive Integrated Moving Average) Time Series Prediction***

  ARIMA, is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. [11]

  Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modelled with ARMA models.

  Formally it is said that a time series yt follows an ARIMA model (p, q) if it satisfies the relation:

  $y_t = a_0 + a_1 y_{t-1} + ... + a_p y_t - p + \varepsilon_t$,

  where

  $\varepsilon_t = u_t + b_1 u_t - 1 + ... + b_q u_t - q$

  it is called a moving average process of order q or MA (q) (MA stands for Moving Average) and where the errors u_t are a white noise, that is a succession of uncorrelated random variables with zero mean and finite variance. The ARMA model can be considered as a way to approximate the autocovariances of y_t. The reason is that any time series y_t with finite covariance can be written as an AR (autoregressive, model) or as an MA with uncorrelated errors, although the AR or MA models may require an infinite order.

  A stationary process can be written in the form of a moving average. This result, known as Wold's representation theorem, is one of the fundamental results underlying the analysis of stationary time series. In some cases the autocovariances can be better approximated using an ARMA (p, q) model with small p and q rather than a pure AR model with only a few lags.

From a practical point of view, however, ARMA model estimation is more difficult than AR model estimation, and ARMA models are more difficult to extend to the case of additional regressors than AR models.

In general, once the order (p, q) has been chosen, the parameters of an ARMA model (p, q) can be estimated e.g. through the maximum likelihood estimator. As for the autoregressive model, the choice of the order of the model must meet the opposing needs of a good adaptation to the data and of parsimony in the number of parameters to be estimated. If the data show the presence of non-stationarity, it is sometimes possible to remove this non-stationarity through the transformation into prime differences, $yt - yt - 1$. The ARMA model (p, q) applied to the data thus transformed is called ARIMA model (Autoregressive Integrated Moving Average) with parameters (p, 1, q). The transformation of the data into first differences can be applied $d \geq 0$ times, thus obtaining the ARIMA model (p, d, q). In particular, the ARIMA model (p, 0, q) coincides with the ARMA model (p, q). A simple example of an ARIMA model is given by the random walk $yt = yt - 1 + ut$, which is an ARIMA (0, 1, 0).

- ***Support Vector Machines (SVM)***

Recently, forecasting of future observations based on time series data has received great attention in many fields of research. Several techniques have been developed to address this issue in order to predict the future behaviour of a particular phenomenon.

The application of SVM for time series forecasting is relatively new. The initial results shown in this research have clearly demonstrated the potential of this approach in predicting time series data. Future works can be centered on investigating the performance of SVM with different kernel functions and optimal hyper parameters of SVM forecasting model, which has the potential to improve the accuracy of the forecast.

Compared to other neural network regressors, SVM has three distinct characteristics when it is used to estimate the regression function. First, SVM estimates the regression using a set of linear functions that are defined in a high-dimensional feature space. Second, SVM carries out the regression estimation by risk minimization, where the risk is measured using Vapnik's insensitive loss function. Third, SVM implements the SRM principle which minimizes the risk function consisting of the empirical error and a regularized term.

From the implementation point of view, training SVM is equivalent to solving the linearly constrained quadratic programming problem with the number of variables twice as that of the number of training data points. The sequential minimal optimization (SMO) algorithm extended by Scholkopf and Smola [12], [13] is very effective in training SVM for solving the regression estimation problem.

Support vector machines (SVM) and artificial neural networks (ANN) are alternative methods that can be used for forecasting in nonlinear time series and can overcome the problems of nonlinearity and non stationarity.

- ***LSTM (Long Short Term Memory) Networks***

  LSTM cells are used in recurrent neural networks that learn to predict the future from sequences of variable lengths. Note that recurrent neural networks work with any kind of sequential data [14].

  LSTM neural networks are capable of solving numerous tasks that are not solvable by previous learning algorithms like RNNs (Recurrent Neural Networks). Long-term temporal dependencies can be captured effectively by LSTM, without suffering much optimization hurdles. This is used to address the high-end problems.

  LSTM networks are indeed an improvement over RNNs as they can achieve whatever RNNs might achieve with much better finesse. As intimidating as it can be, LSTMs do provide better results and are truly a big step in Deep Learning. With more such technologies coming up, you can expect to get more accurate predictions and have a better understanding of what choices to make.

  The main idea behind LSTM cells is to learn the important parts of the sequence seen so far and forget the less important ones.

  This is important for telecommunication industry data which might be affected by seasonal changes. This is one of the main reasons why this model is chosen for our scenario.

2.3. CUSTOMER'S CHURN

Churn prediction is probably one of the most important applications of data science in the commercial sector. The thing which makes it popular is that its effects are more tangible to comprehend and it plays a major factor in the overall profits earned by the business.

Churn is defined in business terms as ‘when a client cancels a subscription to a service they have been using’, in our case the subscription of telecommunication provider. So, Churn Prediction is essentially predicting which clients are most likely to cancel a subscription i.e ‘leave a company’ based on their usage of the service.

2.3.1. Dataset

Studying and trying to predict customer’s churn is one of the main problems that companies in overall, but especially telecommunication companies try to solve. There are many factors which impact on one’s decision to join or leave a company, but only a few can be predictable.

Now let’s have a look how a customer’s independent behaviour and characteristics can lead to predict if he might be a possible churner. We are taking into consideration the attributes defined below:

- customer’s last 9 months activity: 0 indicates that the customer hasn’t performed any outgoing or incoming activity during a month; 1 indicates that the customer has performed at least one outgoing or incoming activity (call/message) during a month
- port out flag (0/1 field) indicates whether the customer performed a port out or not in the past
- deactivation flag (0/1 field) indicates whether the customer has ever been deactivated before
- revenue of last month shows the amount of money a customer has spent on his last month (month of observation in the end)
- total revenue of last 3 months shows the amount of money a customer has spent three last months
- maximum recharge of last 3 months
- bundle purchase last 9 months activity
- tenure of the customer in months shows how long has a person been a customer of this operator
- activation date shows when did the customer join the company for the first time (customer might have performed a portout and then portin again in between)
- location shows where is the customer registered (which city of the country)
- age shows the age of the customer

- month is the month of observation

The data has been pre-processed and a sample of it can be seen in Table 2.2.

MSISDN	9M ACTIVIT	PORT OUT FLA	DEACT FLA	REV 1	REV 3	MAX RECH 3	BND PURCH 9	TENURE MONT	ACTIVATION DA1	LOCATION	AGE	MONTH
35568xxxx866	011011111	0	0	1000	2500	1500	011	2	01/12/2021	TIRANA	25	202202
35568xxxx273	111111111	0	0	1500	3000	1500	111	5	05/09/2021	DURRES	28	202202
35568xxxx682	111001100	1	0	0	1200	1200	100	25	16/04/2020	TIRANA	32	202202

Table 2. Dataset sample of Customer Churn

By having this information on the customers, we can use a supervised machine learning model to predict who are the customers who might want to port out from our telecom operator.

2.3.2. Machine learning Model Identification

Customer's churn prediction is a problem that telecommunication companies have been trying to solve for a long time now and using different approaches.

A common task that appears in different machine learning applications is to build a non-parametric regression or classification model from the data. When designing a model in domain-specific areas, one strategy is to build a model from theory and adjust its parameters based on the observed data. Unfortunately, in most real-life situations such models are not available. In most situations, even initial expert-driven guesses about the potential relationships between input variables are not available to the researcher. The lack of a model can be circumvented if one applies non-parametric machine learning techniques like neural networks, support vector machines, or any other algorithm at one's own discretion, to build a model directly from the data [15]. These models are built in the supervised manner, which means that the data with the desired target variables has to be prepared beforehand.

Based on the data we have been able to collect, there have been a couple of machine learning algorithms taken into consideration before deciding on XG BOOST (Extreme Gradient Boosting):

- ***Gradient Boosting Machine Tree***

  Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak

learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.

In gradient boosting machines, or simply, GBMs, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic squared-error loss, the learning procedure would result in consecutive error-fitting.

In general, the choice of the loss function is up to the researcher, with both a rich variety of loss functions derived so far and with the possibility of implementing one's own task-specific loss.

This high flexibility makes the GBMs highly customizable to any particular data-driven task. It introduces a lot of freedom into the model design thus making the choice of the most appropriate loss function a matter of trial and error. However, boosting algorithms are relatively simple to implement, which allows one to experiment with different model designs.

Moreover the GBMs have shown considerable success in not only practical applications, but also in various machine-learning and data-mining challenges [16] (Bissacco et al., 2007; Hutchinson et al., 2011; Pittman and Brown, 2011; Johnson and Zhang, 2012).

To design a particular GBM for a given task, one has to provide the choices of functional parameters $\Psi(y, f)$ and $h(x, \theta)$. In other words, one has to specify what one is actually going to optimize, and afterwards, to choose the form of the function, which will be used in building the solution. It is clear that these choices would greatly affect the GBM model properties. The GBM framework provides the practitioner with such design flexibility.

- ***Decision Trees***

  'Decision tree' is a collective name for two different machine learning methods: a regression tree and a classification tree. A regression tree is used for numerical target variables. The churn problem requires a classification tree approach, which can have

categorical or binary dependent variables. A modern and common-used abbreviation for decision tree is CART(classification and regression tree).

The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes).

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value.

Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Internal nodes are represented as circles, whereas leaves are denoted as triangles.

Note that this decision tree incorporates both nominal and numeric attributes. Given this classifier, the analyst can predict the response of a potential customer (by sorting it down the tree), and understand the behavioural characteristics of the entire potential customers population regarding churn. [17]

Each node is labelled with the attribute it tests, and its branches are labelled with its corresponding values.

- ***XG BOOST***

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

It trains faster especially on larger datasets. [18]

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

XGBoost is going to be explained in more details when we will be talking about the Customer Churn as it is the chosen algorithm to solve that machine learning problem.

3. IMPLEMENTATION AND ANALYSIS

We are going to divide this section into two sub sections corresponding to our two areas of research. For each of them it will be explained in detail the identification of variables, a descriptive analysis, the implementation of the chosen machine learning model and the results.

3.1. MARKET SHARE FORECASTING

3.1.1. Identification of variables and data pre-processing

There are various internal and external factors to be taken into consideration when you want to predict how the market share is going to be in the upcoming days. A company might have just released a huge campaign that you know of, but what you do not know is what is planning the competitor. Therefore, you might be expecting a raise in the market share with this big move, and you do get many new customers from the competitor. However, suddenly your competitor does another big campaign, and they get some other customers from you. The market share might change or not, but you cannot know what your competitor's strategy is for the near future.

Since we are working on daily data and we need to forecast daily market share of the next month, we need to work with data we are certain of and rely on the past patterns. Therefore, we need to identify what have been our new activations and competitor's, but also the port-in/port-outs from these operators in the market (the movement between operators). Again, we have the information on our end, but do not have access to the information of the competitors. Nevertheless, there is one way to identify what we need to know by the database, which contains information on calls activity.

Every telecom provider stores information of incoming and outgoing calls. Each call record has a mobile number who initiated the call, and another one who received and answered the call, also the duration of the call.

But we need a couple of other transformation to identify the new activations and port-outs. Note that we don't need to differentiate between the two perse. We need to identify these among other customers. How we can do this is by checking which numbers haven't appeared previously in our

database within that specific operator. This way we are tracking at once the port-in/port-outs and new customers as well.

Finally, we need to translate this information in percent of the market share.

3.1.2. Descriptive analysis

Descriptive statistic is a set of techniques used to describe the basic characteristics of the data collected in the study. They provide a simple synthesis of the sample and the collected measures. Along with simple graphic analysis, it constitutes the initial starting point for any analysis of quantitative data. In our case we will go through our dataset over time and try to identify the outliers.

First, let us have a look at the change of the market share over time. We are studying three operators and the historical view of the market share for each of them as is shown in Figure 3.1 with different colours.

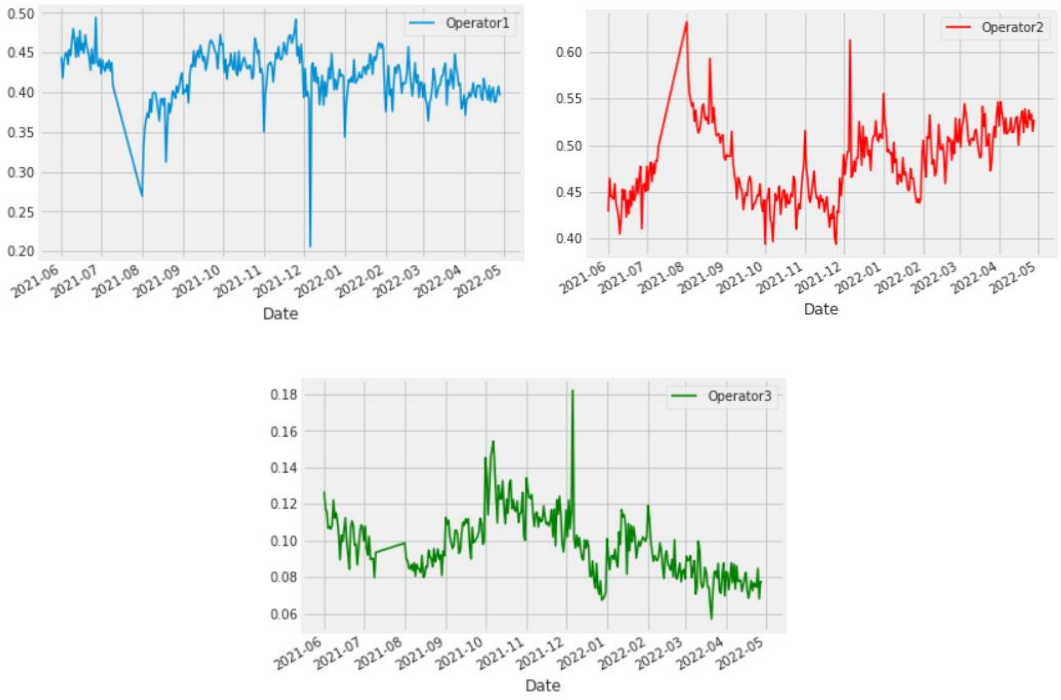

Figure 3.1. Market Share representation in different graphs of Operator1 in blue, Operator2 in red, and Operator3 in green

Maybe having a look at the market share change over time at the same graph might give us some better insights.

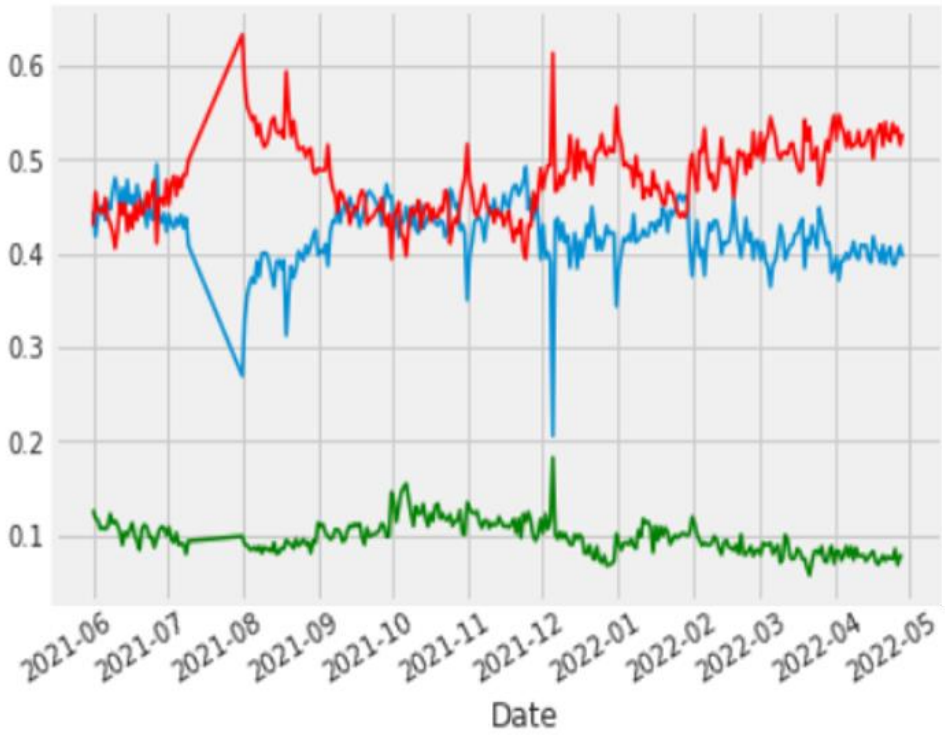

Figure 3.2. Market Share representation in same graph of Operator1 in blue, Operator2 in red, and Operator3 in green

As it may be seen from Figure 3.2. it appears that the big players in the industry of telecommunication in this specific market are Operator1 (in blue) and Operator2 (in red). There also seem to be some correlation between the two because the graphs of the two are almost complementary and symmetrical while the green one of Operator3 is almost linear and stable.

Talking about noticing a correlation between different operators in the historical data, let us go ahead and get the daily return for all the operators and compare the daily percentage return of each two of them to check how correlated they are. Daily return on a stock is used to measure the day to day performance of stocks, it is the price of stocks at today's closure compared to the price of the same stock at the day-before closure. [20] In our case we are not comparing price stocks but new activations.

Seaborn and pandas (python libraries) make it very easy to repeat this comparison analysis for every possible combination of stocks in our technology stock ticker list. We can use sns.pairplot()

[21] to automatically create this plot. We can simply call pairplot on our DataFrame for an automatic visual analysis of all the comparisons.

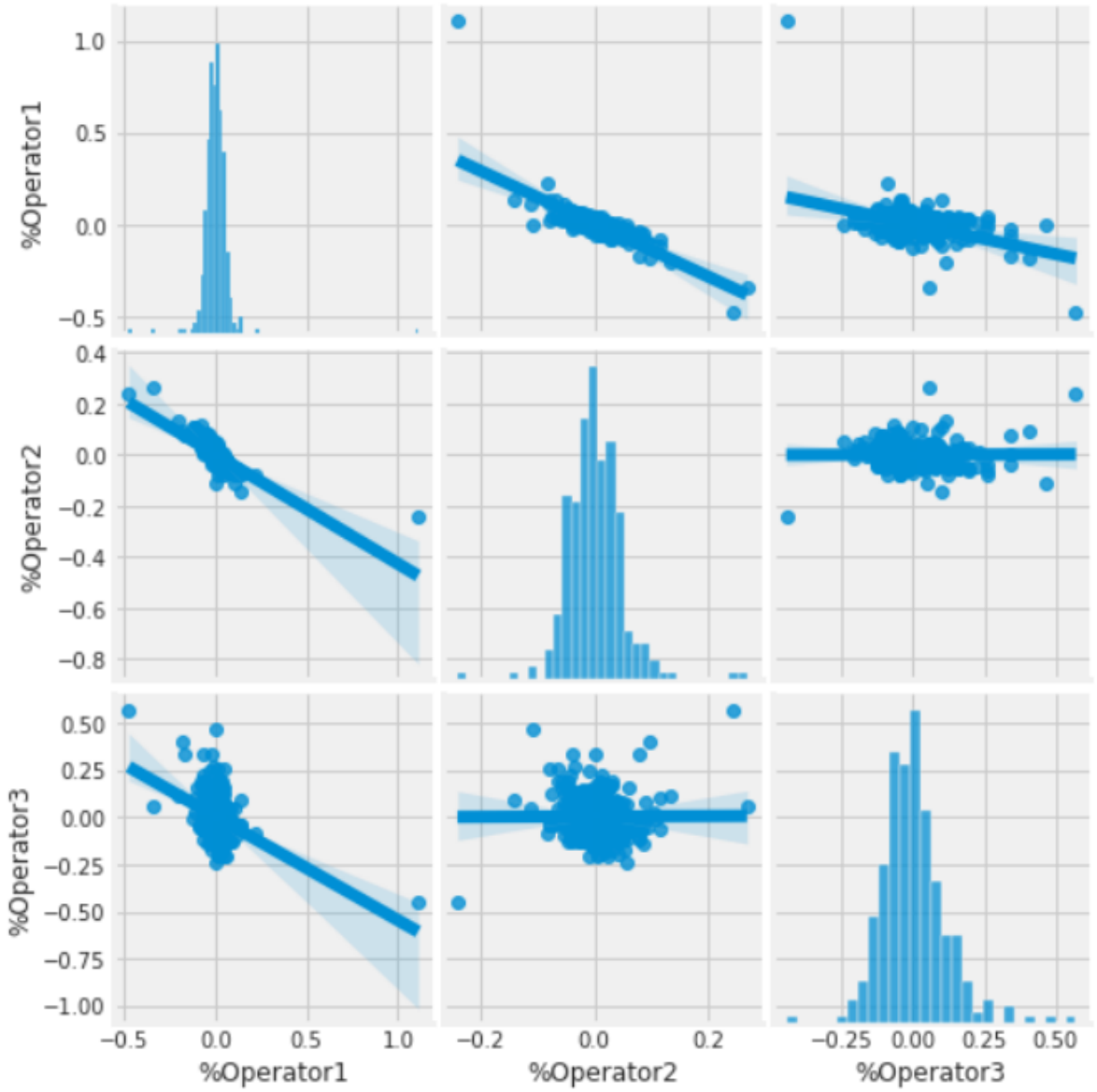

Figure 3.3. Relationship on daily return between three operators using sns.pairplot().

In Figure 3.3. we can see all the relationships on daily returns between all the stocks. A quick glance confirms the correlation between Operator1 and Operator2. It is a negative correlation that usually depicts when one is going down, the other is going up. In the beginning of this research, we started with a strong statement saying that the operators nowadays are fighting to retain their customers rather than getting new ones. That is because the market is already saturated and we

know the real struggle is retaining customers from the competitors. This can be confirmed by the correlation between the two big players of this telecommunication market. It is obvious from the graphs we have seen so far that Operator1 and Operator2 are continuously getting each other's customers.

While just calling sns.pairplot() is really simple we can also use sns.PairGrid() for full control of the figure.

Finally, we also do a correlation plot (Figure 3.4.), to get actual numerical values for the correlation between the stocks' daily return values. By comparing the closing market shares, we see an interesting negative relation between Operator1 and Operator2. We confirm once again the negative strong relation between Operator1 and Operator2. While on the other hand we can see there is no correlation whatsoever between Operator2 and Operator3, but there is a somewhat tighter negative relation between Operator1 and Operator3. Now we have already noticed that Operator1 and Operator2 are the two big players in this telecommunication market. With Operator3 being the smallest one in the market, it is natural that it can aim to get market share from Operator2 which is not the biggest, but not the smallest either. That is why we can notice a stronger correlation of Operator3 with Operator1 instead of Operator2.

Figure 3.4. Correlation plot for the daily returns.

3.1.3. LSTM implementation

LSTM stands for Long short-term memory. LSTM cells are used in recurrent neural networks that learn to predict the future from sequences of variable lengths. Note that recurrent neural networks work with any kind of sequential data.

The main idea behind LSTM cells is to learn the important parts of the sequence seen so far and forget the less important ones. [22]

This is important for telecommunication industry data which might be affected by seasonal changes. As we have seen from the descriptive analysis performed in our dataset in the above section. This is one of the main reasons why this model is chosen for our scenario.

How does LSTM work?

Long short-term memory (LSTM) units (or blocks) are a building unit for layers of a recurrent neural network (RNN). A RNN composed of LSTM units is often called an LSTM network [23]. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for "remembering" values over arbitrary time intervals; hence the word "memory" in LSTM. Each of the three gates can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM; hence the denotation "gate". There are connections between these gates and the cell.

The expression long short-term refers to the fact that LSTM is a model for the short-term memory which can last for a long period of time. An LSTM is well-suited to classify, process and predict time series given time lags of unknown size and duration between important events. LSTMs were developed to deal with the exploding and vanishing gradient problem when training traditional RNNs.

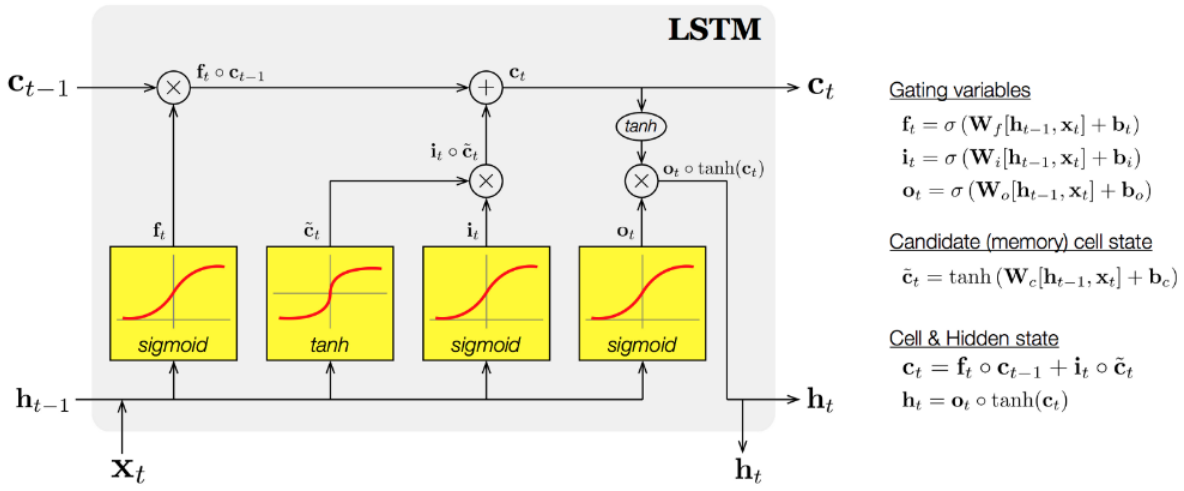

Figure 3.5. LSTM logical functioning.

Components of LSTMs:

So, the LSTM cell contains the following components

- Forget Gate “f” (a neural network with sigmoid)
- Candidate layer “C"(a NN with Tanh)
- Input Gate “I” (a NN with sigmoid)
- Output Gate “O”(a NN with sigmoid)
- Hidden state “H” (a vector)
- Memory state “C” (a vector)
- Inputs to the LSTM cell at any step are Xt (current input) , Ht-1 (previous hidden state) and Ct-1 (previous memory state).
- Outputs from the LSTM cell are Ht (current hidden state) and Ct (current memory state)

Working of gates in LSTMs

First, LSTM cell takes the previous memory state Ct-1 and does element wise multiplication with forget gate (f) to decide the present memory state Ct. If forget gate value is 0 then previous memory state is completely forgotten else f forget gate value is 1 then previous memory state is completely passed to the cell (Remember f gate gives values between 0 and 1).

3.1.4. Results

Let us now try to predict the market share of the telecommunication market between the three operators we have been observing in this study.

We start by observing each of the operators individually. First, we create a dataframe with the dataset we are going to work with. In figure 3.6. a historical view of Operator 1 can be seen for the last 1 year.

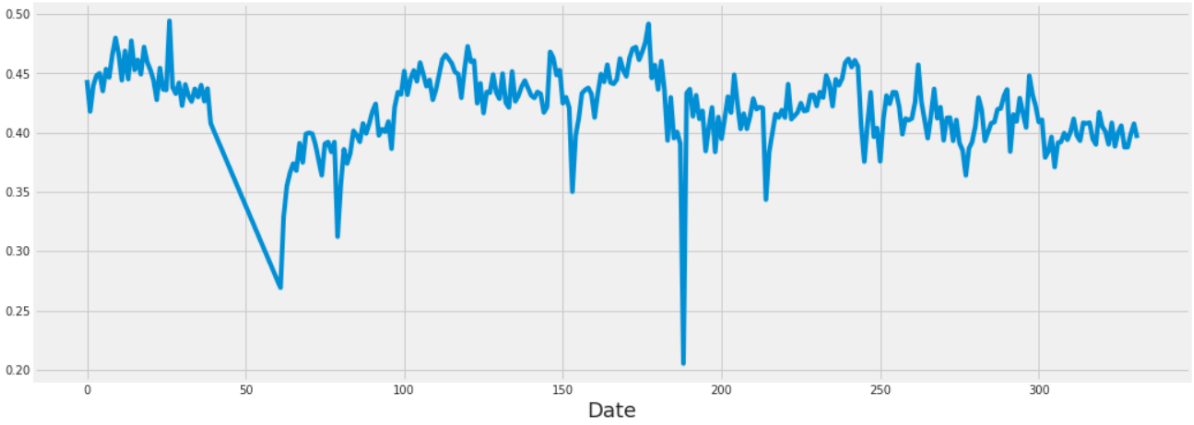

Figure 3.6. Operator1 data to feed the LSTM model.

Then we go on with dividing the dataset into training and testing subsets. Training data will be having 80% of the whole dataset. The other 20% will be used for testing.

We are then going to build the LSTM model as per figure 3.7.

```
# Build the LSTM model
model = Sequential()
model.add(LSTM(128, return_sequences=True, input_shape= (x_train.shape[1], 1)))
model.add(LSTM(64, return_sequences=False))
model.add(Dense(25))
model.add(Dense(1))

# Compile the model
model.compile(optimizer='adam', loss='mean_squared_error')

# Train the model
model.fit(x_train, y_train, batch_size=1, epochs=20)
```

Figure 3.7. Python code to build the LSTM model

We are using the root mean squared error (RMSE) for which the obtained value for the Operator1 is 0.024. Root-Mean-Square-Error or RMSE is one of the most popular measures to estimate the accuracy of our forecasting model's predicted values versus the actual or observed values while training the regression models or time series models [24]. Taking the square root of the average squared errors has some interesting implications for RMSE. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.

Finally, we plot the predicted data alongside the actual values and the result can be seen in figure 3.8. While in Table 3.1. We compare side by side the ground truth and predicted values.

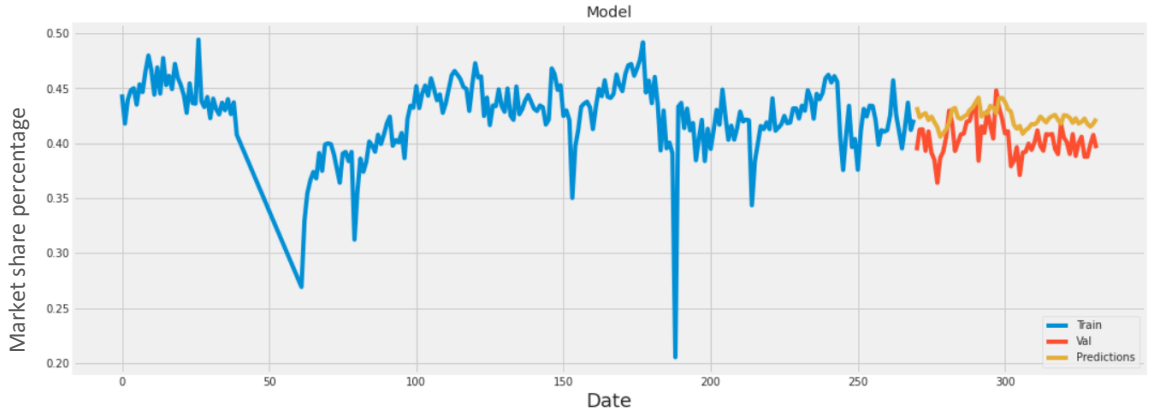

Figure 3.8. Visualization of prediction of Operator1 on top of ground truth-values

	Operator1	Predictions
270	0.393548	0.432957
271	0.412463	0.423551
272	0.412721	0.425952
273	0.393177	0.427319
274	0.410754	0.421080
...	...	...
327	0.387791	0.422442
328	0.387816	0.417553
329	0.399054	0.415207
330	0.407603	0.417964
331	0.395494	0.422350

Table 3.1. Predictions vs Ground Truth for Operator1 side by side

Now let us do the above steps and analysis again for the other two operators. Figure 3.10. shows the historical view of Operator2 over the last year.

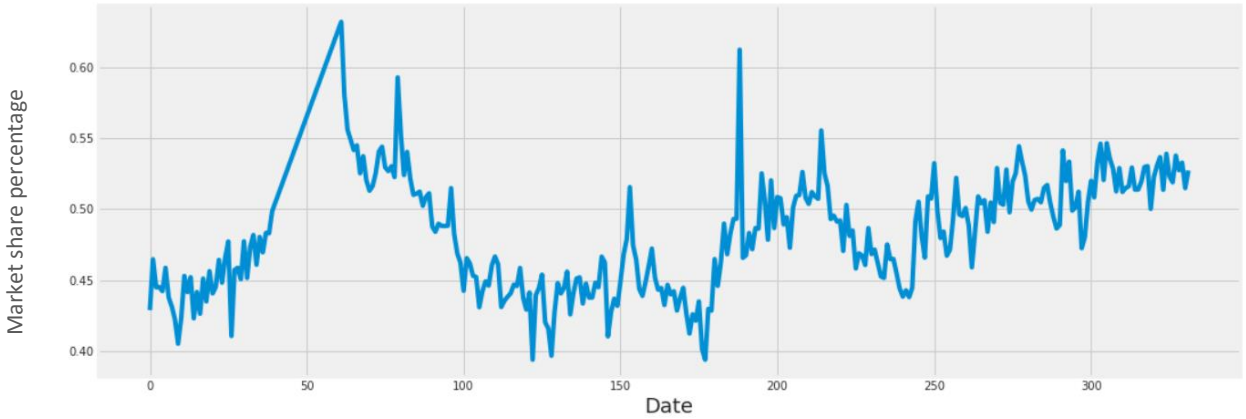

Figure 3.10. Operator2 data to feed the LSTM model.

The RMSE obtained value for the Operator2 is 0.02.

In Figure 3.11. a visualization of the prediction of next 60 days of data can be shown on top of ground truth-values. It is interesting to notice how the line in orange (which are the predicted values) follow to red line trajectory quite closely. The numerical results can then be seen side to side on Table 3.2.

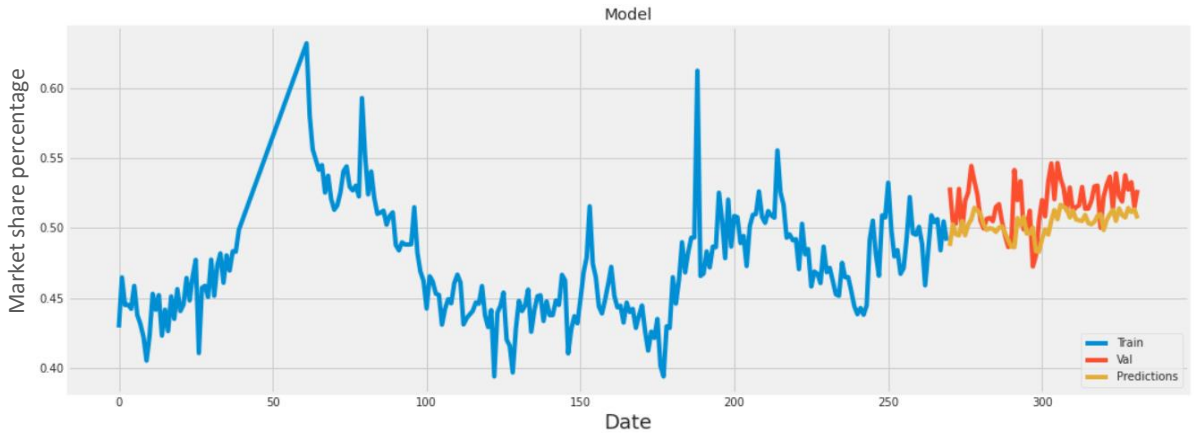

Figure 3.11. Visualization of prediction of Operator2 on top of ground truth-values

	Operator2	Predictions
270	0.529032	0.487114
271	0.504451	0.502624
272	0.503180	0.495475
273	0.527964	0.494760
274	0.497783	0.504958
...	...	...
327	0.537791	0.507570
328	0.527489	0.514272
329	0.532861	0.511472
330	0.514918	0.513394
331	0.527325	0.506789

Table 3.2. Predictions vs Ground Truth for Operator2 side by side

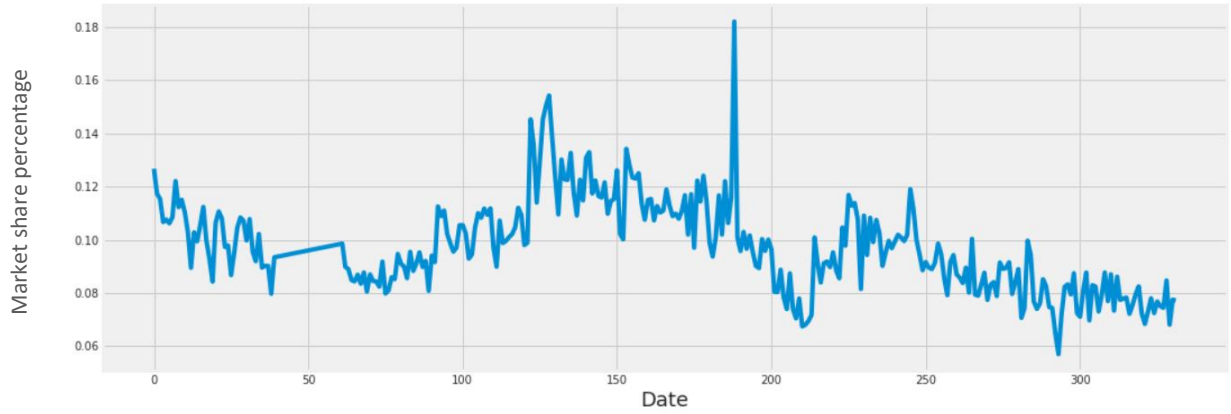

Figure 3.13. Operator3 data to feed the LSTM model.

The RMSE obtained value for the Operator3 is 0.01.

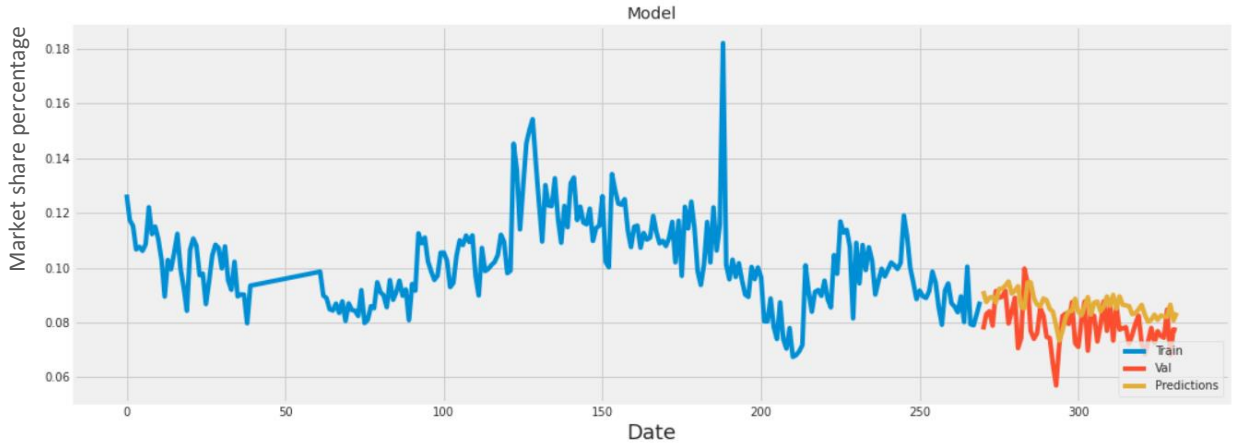

Figure 3.14. Visualization of prediction of Operator3 on top of ground truth-values

	Operator3	Predictions
270	0.077419	0.091485
271	0.083086	0.087491
272	0.084099	0.089039
273	0.078859	0.089401
274	0.091463	0.087336
...	...	...
327	0.074419	0.082222
328	0.084695	0.082044
329	0.068085	0.086479
330	0.077478	0.080702
331	0.077181	0.083607

Table 3.3. Predictions vs Ground Truth for Operator2 side by side

3.2. CUSTOMER CHURN PREDICTION

Churn prediction is probably one of the most important applications of data science in the commercial sector. What makes it popular is that its effects are more tangible to comprehend and it plays a major factor in the overall profits earned by the business.

Churn is defined in business terms as 'when a client cancels a subscription to a service they have been using', in our case the subscription of telecommunication provider. So, Churn Prediction is essentially predicting which clients are most likely to cancel a subscription i.e 'leave a company' based on their usage of the service [25].

3.2.1. Identification of variables and data pre-processing

Let's go back again to the above explanation of churn for a little bit: *"So, Churn Prediction is essentially predicting which clients are most likely to cancel a subscription i.e 'leave a company' based on their usage of the service".* It mentions the usage of service as a key indicator to lead us to the possibility for a customer to churn or not. So, let's stick to this and identify some variables which might help us with it.

The most intuitive thing we might need to consider is whether the customer is performing any sort of incoming or outgoing activity at all or not and for how long. That is how the variable 9M activity was created. It takes into consideration the last 9 months and for each month we have displayed a 0 or 1 to show lack or presence of incoming/outgoing activity of any type. The reason is pretty straightforward: if a customer hasn't been performing any type of activity for a couple of months now, he most probably has already left.

The same logic applies to recharges/bundles a customer has purchased for the last 9 months. A time windows of 9 months is being used because of the fast pacing industry telecommunication is.

Something else we can consider is the past history of this customer with port outs and deactivations. So, if a customer has already performed a port out in the past, and then came back, there might be a chance that he moves again, or it's all the opposite, if they come back it means they already made up their mind. Either case, this is a variable which definitely helps to train an algorithm on customer churn.

Since we are already here studying the past history of the customer, a well-known attribute which is usually studied in churn propensity models is the tenure of the customer. It is very intuitive why, customers who have being staying the longest with a company are considered loyal and don't tend to change companies. Other than that in almost every company there are special campaigns being made from time to time for loyal customers so they benefit from the company too much for leaving for another company.

Another variable might be the activation date of the customer in this telecommunication provider which doesn't necessarily correlate with the tenure of the customer. It has to do with the fact whether the customer has some time in the past performed a port out or a deactivation.

We talked about recharges/bundles of the last 9 months being a variable, which will for sure have a significant weight on this prediction model. However it is also important how much revenue is this customer generating with his recharges i.e. how much a customer is able to spend with this telecom provider. Having said this we can identify some other variables like revenue of the most recent month, total revenue of last 3 months, max recharge of last 3 months.

Also it is natural that in this kind of analysis we also take into consideration some personal data like the age or demographic data like the location of the customer.

A sample of the used dataset is shown at Table 3.4

MSISDN	9M ACTIVIT	PORT OUT FLA	DEACT FLA	REV 1	REV 3	MAX RECH 3	BND PURCH 9	TENURE MONT	ACTIVATION DA1	LOCATION	AGE	MONTH
35568xxxx866	011011111	0	0	1000	2500	1500	011	2	01/12/2021	TIRANA	25	202202
35568xxxx273	111111111	0	0	1500	3000	1500	111	5	05/09/2021	DURRES	28	202202
35568xxxx682	111001100	1	0	0	1200	1200	100	25	16/04/2020	TIRANA	32	202202

Table 3.4. Sample data of Customer Churn

3.2.2. Descriptive analysis

First let us have a look at the correlation of Churn with other variables at Figure 3.17.

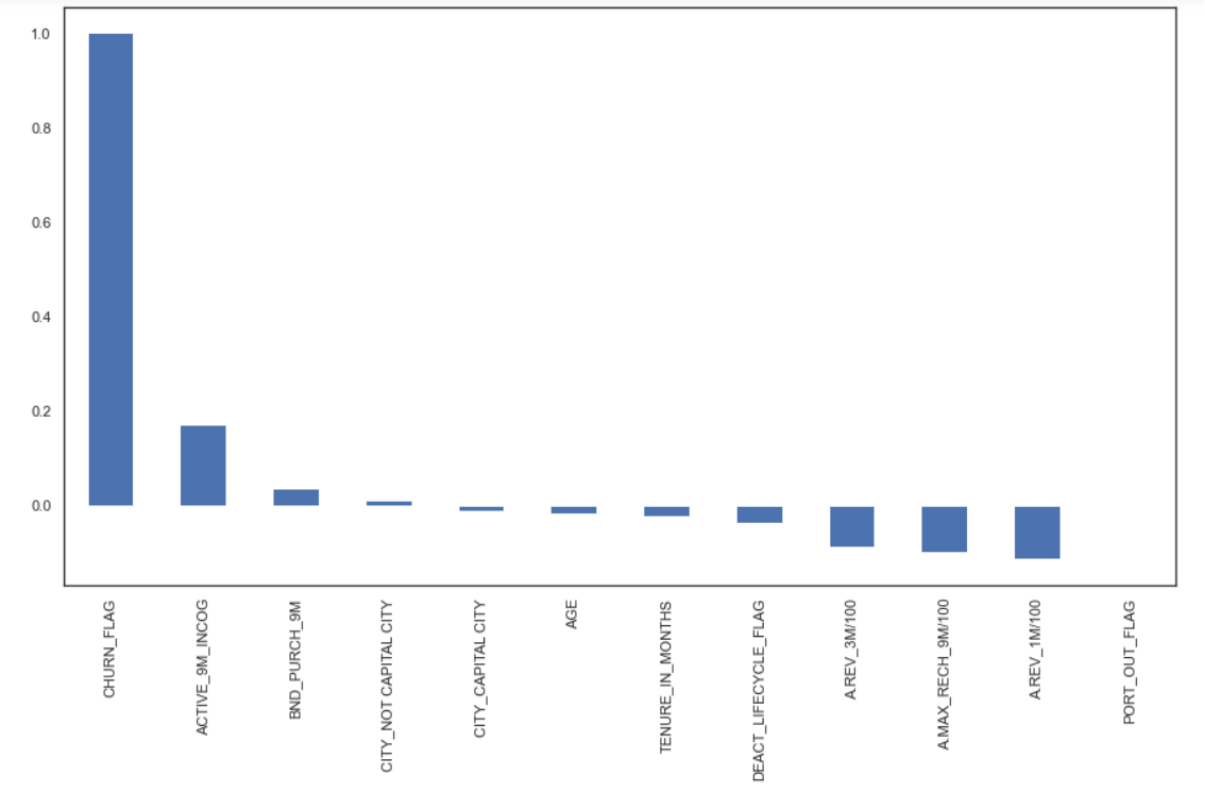

Figure 3.17. Correlation of Churn with other variables

Last 9 months activity, bundle purchase of last 9 months and location seem to be positively correlated with churn. While, tenure, age, revenue of last months and last 3 months and max recharge of last 9 months seem to be negatively correlated with churn. In different markets recharges have different significance. For markets like the Albanian one which is dominated by

the prepaid bundles rather than post-paid contracts, studying recharges plays an important role in studying churn predictions.

We will explore the patterns for the above correlations below before we delve into modelling and identifying the important variables.

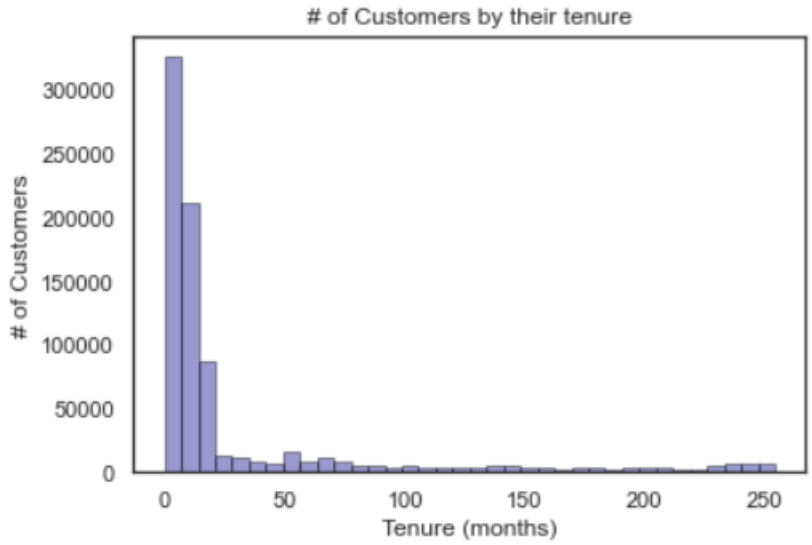

Figure 3.18. Number of Customers by their tenure

- ***Tenure:*** After looking at the histogram of figure 3.18. we can see that a lot of customers have been with the telecom company for just a month, while the majority have a tenure of less than 2 years. Before that, the number of customers is almost constant, and these seem to be the loyal customers of the operator. This could be potentially because the operator is following a more aggressive strategy toward getting new customers rather than retaining the existing ones.

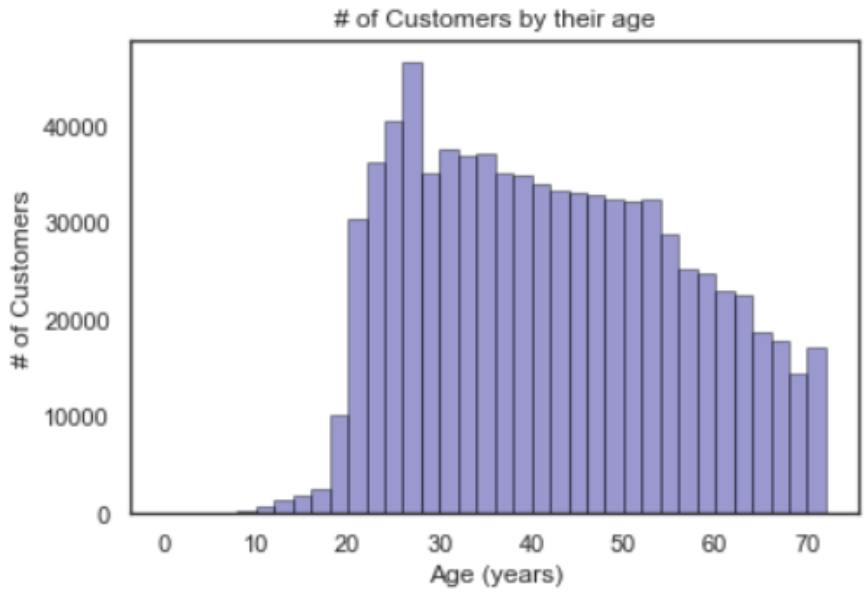

Figure 3.19. Number of Customers by their age

- ***Age:*** As can be depicted from the above histogram in Figure 3.19 customers vary in age from around 20 to around 70. The age group with the highest number of customers seem to be mid twenties.

Now let's take a look at our predictor variable Churn and understand its interaction with other important variables as was found out in the correlation plot.

Let's first look at the churn rate in our data in figure 3.20.

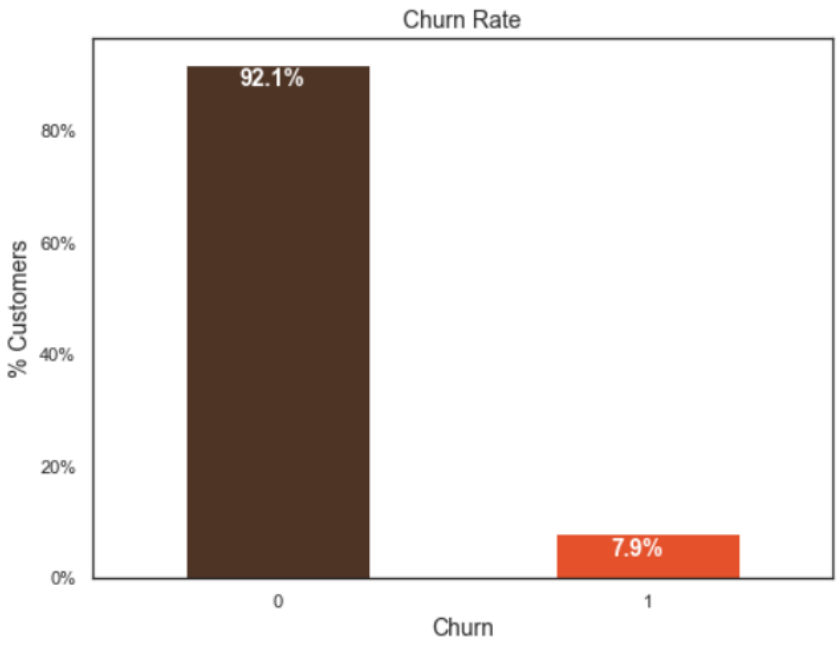

Figure 3.20. Churn rate in our sample data

In our data, 92% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skeweness could lead to a lot of false negatives. We will see in the modelling section on how to avoid skewness in the data.

Lets now explore the churn rate by tenure, seniority, contract type, monthly charges and total charges to see how it varies by these variables.

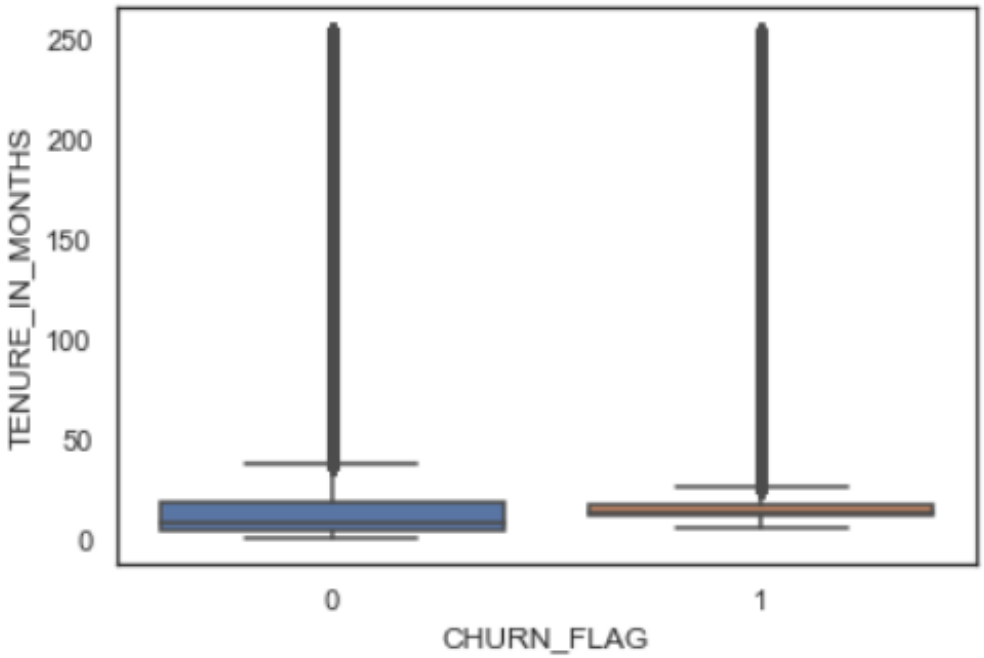

Figure 3.21. Churn vs Tenure

- ***Churn vs Tenure:*** As we can see form the above plot in figure 3.21, the customers who do not churn, they tend to stay for a longer tenure with the telecom company
- ***Churn by last month money spent:*** Customers who haven't spent any money in their mobile number last month tend to churn more than the rest. This is shown in figure 3.22.

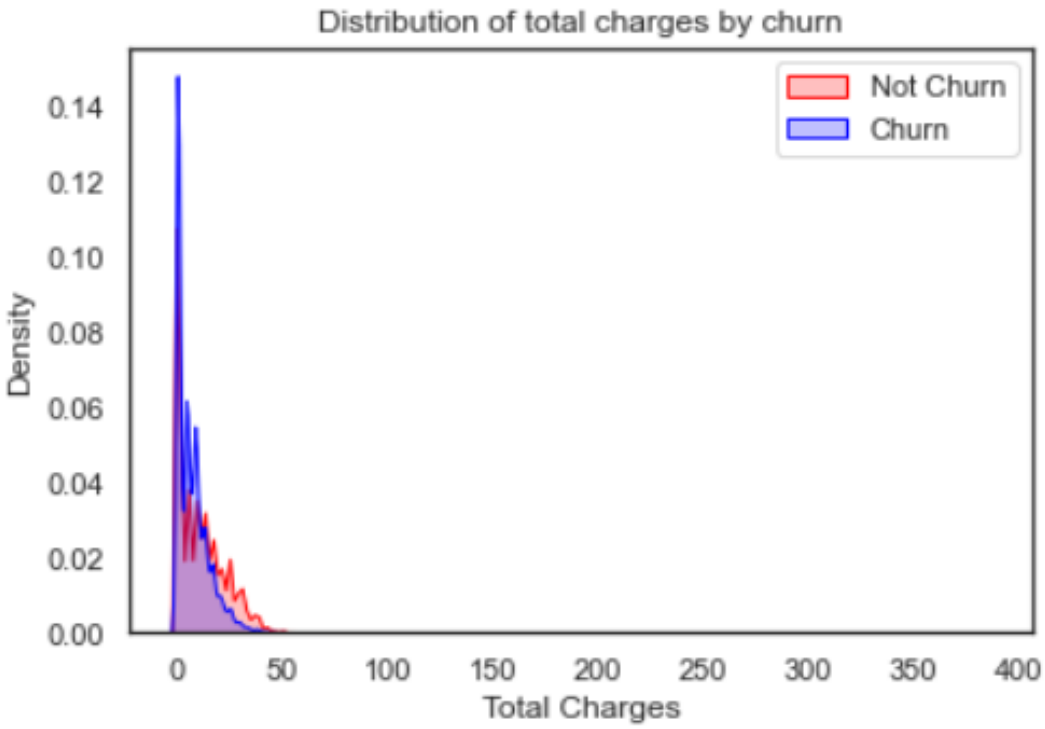

Figure 3.22. Distribution of last month recharges by churn

3.2.3. XGBOOST algorithm implementation

XGBoost [26] (Extreme Gradient Boosting) algorithm is a boosting algorithm for classifying regression tree models [27] which is coming from the gradient lifting decision tree. XGBoost is used for customer churn prediction in this paper. A general flow of the Boosting algorithm based on the classification regression tree is shown in Figure 3.23. First, we need to learn a tree from the sample to obtain the first estimation result Y1, and the second tree is learned with Y based on the difference between the real label and the predictive label in the previous step. By analogy, the algorithm error can be effectively reduced.

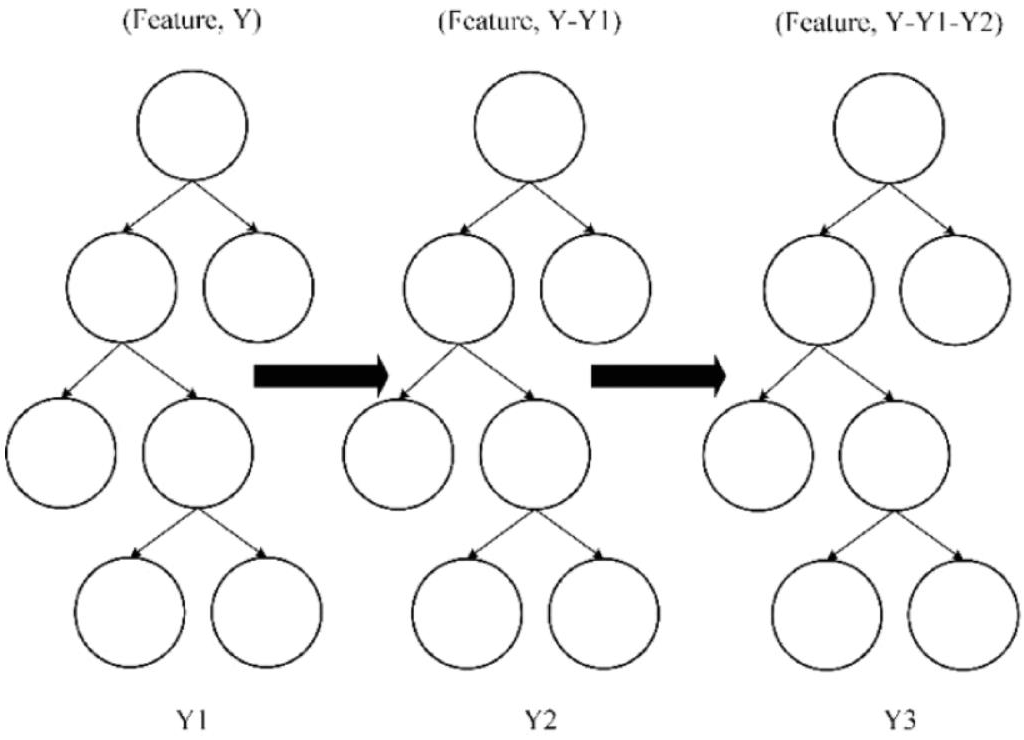

Figure 3.23. Boosting Algorithm flow

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

System Optimization:

1. Parallelization: XGBoost approaches the process of sequential tree building using parallelized implementation. This is possible due to the interchangeable nature of loops used for building base learners; the outer loop that enumerates the leaf nodes of a tree, and the second inner loop that calculates the features. This nesting of loops limits parallelization because without completing the inner loop (more computationally demanding of the two), the outer loop cannot be started. Therefore, to improve run time, the order of loops is interchanged using initialization through a global scan of all instances and sorting using parallel threads. This switch improves algorithmic performance by offsetting any parallelization overheads in computation.

2. Tree Pruning: The stopping criterion for tree splitting within GBM framework is greedy in nature and depends on the negative loss criterion at the point of split. XGBoost uses 'max_depth' parameter as specified instead of criterion first, and starts pruning trees backward. This 'depth-first' approach improves computational performance significantly.
3. Hardware Optimization: This algorithm has been designed to make efficient use of hardware resources. This is accomplished by cache awareness by allocating internal buffers in each thread to store gradient statistics. Further enhancements such as 'out-of-core' computing optimize available disk space while handling big data-frames that do not fit into memory.

Algorithmic Enhancements:

1. Regularization: It penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting.
2. Sparsity Awareness: XGBoost naturally admits sparse features for inputs by automatically 'learning' best missing value depending on training loss and handles different types of sparsity patterns in the data more efficiently.
3. Weighted Quantile Sketch: XGBoost employs the distributed weighted Quantile Sketch algorithm to effectively find the optimal split points among weighted datasets.
4. Cross-validation: The algorithm comes with built-in cross-validation method at each iteration, taking away the need to explicitly program this search and to specify the exact number of boosting iterations required in a single run.

Expressions (1-3) below give the calculation flow of gradient boosting training. The following expressions calculate the target of the n-th tree model. The first behavior inthe model defines a regularization term, which can reduce overfitting to improve the generalization ability of the model. The second behavior is the first three terms from Taylor's formula, which contains constant terms, first and second derivatives. And the first three terms represent the original very well and not complexity. From the expressions we can see that one of the advantages of XGBoost is that it's accurate to the second derivative.

Among them, the objective function of each round is calculated by the below expressions , and an ft is chosen to minimize our objective function, that is, the error between the predicted result and the actual result is reduced after adding *ft* . Where l is the error function and Ω is a regularization term, the error function tries to fit the training data as much as possible, and the regularization term encourages a simpler model. When the model is simple, the randomness of the results of the finite data fitting is relatively small, which is not easy to over-fitting, which makes the prediction of the final model more stable.

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) + constant \tag{1}$$

When the error function l is not a square error, the first three terms of Taylor expansion are used to approximate the original objective function as the below formula.

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right) + constant \tag{2}$$

where gi is the first derivative of the error function and hi is the second derivative of the error function.

$$\begin{aligned} g_i &= \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \\ h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \end{aligned} \tag{3}$$

Then, by removing the constant term, namely, the difference between the real value and the predicted value of the previous round. The objective function only depends on the first and second derivatives of the error function of each data point.

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right) \tag{4}$$

Based on the realization of XGBoost, the algorithm first ranks the eigenvalues, because the tree model needs to determine the best segmentation points, and then stores them in a number of blocks. This structure is reused in later iterations, which has greatly reduced the computational complexity. In addition, the information gain of each feature needs to be calculated in the process of node splitting, so the calculation of information gain can be parallelized by using this data structure. [27]

3.2.4. Results

After going through the above Descriptive analysis we will develop the predictive model XGBoost. Deciding the tree performance can have various different meanings. The amount or extent where the results of the classifier can be interpreted is classified as a measure call Interpretability. This is a measurement where it can be very hard to assess different classifiers based on it since it is subjective. As mentioned earlier, interpretability of decision trees can be easy until some point; however, it is inevitable that it might become very hard to interpret if the tree becomes complex. In some cases, the performance is measured by speed, sometimes by the size of the grown tree and in most cases it is measured by accuracy. Speed is very desiderable in the telco due to the dinamics of this industry.

Accuracy Based metrics are various measures that show the performance of classifiers on rating systems or percentages. Accuracy based metrics have dominated the evaluation methods and techniques since they give the most realistic and easily calculable results. Some of them are accuracy (recognition rate), error rate, recall, specificity and precision. Robustness is how reliable or correct predictions a classifier makes when it encounters noisy data or data with missing values. And of course this is one of the most important metrics that is used to evaluate a decition tree.

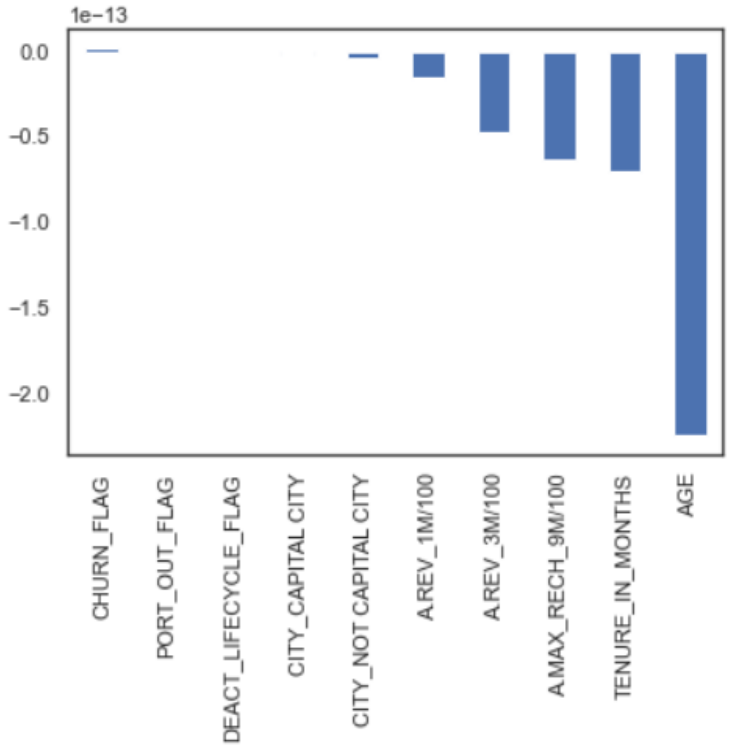

Figure 3.24. Correlation of Churn with other variables

After implementing the XGBoost Algorithm, we were able to get an accuracy score of 0.982.

We can see that some variables have a negative relation to our predicted variable (Churn). Negative relation means that likeliness of churn decreases with that variable. Let us summarize some of the interesting features below:

As we saw in our exploratory data analysis, having a bigger tenure in months reduces chances of churn. Tenure along with the Age have the most negative relation with Churn as predicted by XGBoost.

Generating a high last month revenue also has a negative relationship with Churn. The average revenue of last 3 months has an even higher negative impact on Churn as it means the customer is more consistent in spending money with this operator.

4. COUNCLUSIONS AND FUTURE WORK

One of the most important steps in Telecommunications industry is to understand the behavior of the customers, encourage them in spending more and then predicting their future by preventing their attrition. The churn might be voluntary in cases they want to leave the network they actually are using, or involuntary churn in case of unpaid bills. The methodology used to do the right evaluations in order to achieve strong results in this field is very large and varied.

In this thesis we addressed two problems: Customer Churn in telecommunication industry and Market Share forecasting using machine learning algorithms. The sampling method in this case is a probabilistic one, simple randomization.

A discussion on different machine learning models for each of our topics of interest has been provided. In particular we proposed three machine learning models for market share forecasting: ARIMA, Support Vector Machines and LSTM and we decided to go with the latter one because the main idea behind LSTM cells is to learn the important parts of the sequence seen so far and forget the less important ones. This is important for telecommunication industry data, which might be affected by seasonal changes.

Again, for the other topic of interest Customer Churn, three machine learning algorithms were proposed: Gradient Boosting Machine, Decision Trees and XGBoost. For this machine learning problem, XGBoost was chosen as the most appropriate algorithm that covers our needs. XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

From an experimental point of view, the obtained results for both machine learning problems were very good ones. The LSTM model for the Market Share had an RMSE of 0.02446 for Operator1, 0.020848 for Operator2 and 0. 010315 for Operator3. These are very good indicators for time

series forecasting. The XGBoost model for Customer Churn had an accuracy of 0.98266, which is again a really good indicator for this kind of machine learning problem.

Many different adaptations, tests, and experiments have been left for the future work to be pursued from this point (i.e. the experiments with real data are usually very time consuming, requiring even days to finish a single run). Additional work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or simply curiosity.

BIBLIOGRAPHY

[1] Salvan, A., & Pace, L. (1997). *Principles of Statistical Inference.* Singapore: World Scientific.

[2] Han, J. K., M., & Pei,, J. (2011). *Data Mining: Concepts and Techniques.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[3] Azzalini, A. (2001). *Inferenza Statistica: una Presentazione basata sul Concetto di Verosimiglianza.* Milano: Springer-Italia.

[4] H.Kutner, M., Nachtsheim, C. J., Neter , J., & Li, W. (Fifth Edition). *Applied Linear Statistical Models.* Irwin : McGraw-Hill.

[5] Liseo, B. (2008). *Introduzione alla statistica Bayesiana.* Dispensa.

[6] Zhou, Y., Zhang, T., & Chen, Z. (2006). Applying Bayesian Approach to Decision Tree. *International Conference on Intelligent Computing ICIC,* 290-295.

[7] Vaart, A. v. (2013). *Bayesian statistics and the borrowing of strength in high-dimensional data analysis.* Retrieved from https://www.knaw.nl/shared/resources/actueel/bestanden/20130910_big_data_science_presentatie_aad_van_der_vaart.pdf

[8] K.W. Hipel, A.I. McLeod (1994). *"Time Series Modelling of Water Resources and Environmental Systems",* Amsterdam, Elsevier 1994.

[9] G.P. Zhang, (2007). *"A neural network ensemble method with jittered training data for time series forecasting",* Information Sciences 177 (2007), pages: 5329–5346.

[10] G.P. Zhang, (2003). *"Time series forecasting using a hybrid ARIMA and neural network model",.* Neurocomputing 50 (2003), pages: 159–175.

[11] H. Park (1999). *"Forecasting Three-Month Treasury Bills Using ARIMA and GARCH Models",* Econ 930, Department of Economics, Kansas State University, 1999.

.

[12] A. J. Smola and B. Schölkopf, (1998). *"A Tutorial on Support Vector Regression,"*. Royal Holloway College, London, U.K., NeuroCOLT Tech. Rep. TR 1998-030.

[13] S. K. Shevade, S. S. Keerthi. *Improvements to SMO Algorithm for Regression.* Dept. of Mechanical and Production Engineering, National University of Singapore

[14] F. Karim, S. Majumdar, H. Darabi (2019). *Insights into lstm fully convolutional networks for time series classification.* IEEE Access, 7 (2019), pp. 67718-67725

[15] Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. *Springer Science & Business Media*.

[16] Bissacco, A., Yang, M.-H., Soatto, S., (2007). *Fast human pose estimation using appearance and motion via multi-dimensional boosting regression.* Conference on Computer Vision and Pattern Recognition, 2007, IEEE, pp. 1–8.

[17] Saini, N., Monika, & Garg, K. (2017). Churn Prediction in Telecommunication Industry using Decision Tree. *International Journal of Engineering Research & Technology (IJERT),* 439-443.

[18] T. Chen and C. Guestrin, (2016). *"XGBoost: A scalable tree boosting system",.* Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining

[19] R. Pugliesea, S. Regondia, R. Marini (2021). *Machine learning-based approach: global trends, research directions, and regulatory standpoints.* Data Science and Management Volume 4.

[20] S.E. Schaeffer, S.V.R. Sanchez (2020). *Forecasting client retention — a machine-learning approach.* J. Retailing Consum. Serv., 52 (Jan.) (2020).

[21] Seaborn Python library *https://seaborn.pydata.org/*

[22] J. Brownlee (2016). *Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras*

[23] F. A. Gers, J. Schmidhuber and F. Cummins (2000). *"Learning to Forget: Continual Prediction with LSTM".* Neural Computation, vol. 12

[24] Chail, T., Draxler, R., R. (2015). *Root mean square error (RMSE) or mean absolute error (MAE)*

[25] Qureshi, S. A., Rehman, A. S., Qamar, A. M., & Aatif , K. (2013). Telecommunication Subscribers' Churn. *IEEE,* 131-136.

[26] Lu, N., Hua, L., & Lu, J. (2014). A Customer Churn Prediction Model in Telecom Industry Using Boosting. *IEEE*, 1659-1665.

[27] W. XingFen, Y. Xiangbin and M. Yangchun. (2018). *"Research on user consumption behavior prediction based on improved XGBoost algorithm"*. Proc. IEEE Int. Conf. Big Data (Big Data),