Exploratory DataAnalysis (EDA)
on
Telecom Churn Analysis

Sneha Owandkar

Data science trainee,
AlmaBetter, Bangalore

Abstract:

With the rapid development of telecommunication industry, the service providers are inclined more towards expansion of the subscriber base. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more that retaining the existing one. Therefore, by collecting knowledge from the telecom industries can help in predicting the association of the customers as whether or not they will leave the company. The required action needs to be undertaken by the telecom industries in order to initiate the acquisition of their associated customers for making their market value stagnant. Our paper proposes a new framework for the churn prediction model and implements it using the WEKA Data Mining software. The efficiency and the performance of Decision tree and Logistic regression techniques have been compared.

Keywords:

Customer churn, Customer retention, Customer relationship management (CRM), Data mining techniques, Telecom industry.

Introduction:

Data volume has been growing at a tremendous pace over the last two decades due to advancements in information technology. At the same time there has been enormous development in data mining. Many new methods and techniques have been added to process data and gather information. The data gathered from any source is raw data in which the valuable information is hidden. Data mining can be defined as the process of extracting valuable information from data. Data mining techniques have been successfully applied in many different domains. The most difficult problem faced by telecom industry is customer churn. Customer churn models aim to detect customers with a high probability to jump or leave the service provider. A database of customers who might churn allows the company to target those customers and start retention strategies that reduce the percentage of customer churning. Retention of old customers is always the preferable option to the company. Attracting new customers costs almost five to six times more than retaining the old customers. Attracting a new customer includes new recruits of manpower, cost of publicity and discounts. A loyal customer, who has been with a business for quite a long time, tends to generate higher revenues and is less sensitive to competitor prices. Such customers also cost less to keep and in addition, provide valuable word-ofmouth marketing to the business by referring their relatives, friends, and other acquaintances. In telecom Industry, the system is built to provide service to some average number of customers, when the customer number falls below the calculated number. It is considered

as loss to the company [1]. A small step towards retaining an existing customer can lead to a significant increase in revenues and profits. The requirement of retaining customers craves for accurate customer churn prediction models that are both accurate and comprehensible. The Models have to identify customers who are about to churn and their reason for churn to avoid the losses to the telecom industry, a model should be developed to identify the reasons to churn and the improvements required to retain customers.

Problem Statement:

Orange S.A., formerly France Telecom S.A., is a French multinational telecommunications corporation. The Orange Telecom's Churn Dataset, consists of cleaned customer activity data (features), along with a churn label specifying whether a customer cancelled the subscription. Explore and analyse the data to discover key factors responsible for customer churn and come up with ways/recommendations to ensure customer retention.

Customer churn and retention in telecom industry:

Customer churn is a popular measure of lost customers. Telecommunication companies often lose valuable customers and, thus, revenues to the competition. The telecommunication industry has gone through tremendous changes over the last few decades such as addition of new services, technological advancements and increased competition due to deregulation. Customer churn prediction in telecommunication has, thus, become important to industry players in order to protect their loyal customer base, organization growth, and improve its customer relationship management (CRM). Retaining customers with high churn risk is one of the toughest challenges in telecommunication industry today. Due to greater number of service providers as well as more intense competition, customers today have a variety of options to churn. Thus, the telecommunication industry players are waking up to the importance of retaining existing customers as opposed to acquiring new ones. There are many factors that influence customer to churn. Unlike post-paid customers, prepaid customers are not bound by service contracts and they often churn for simplest reasons. Thus, it is quite difficult to predict their churn rate. Another factor is customer loyalty that may be determined by customer service and product quality offered by the service providers. Issues like network coverage issues and reception quality may influence customers to move to the competitor with broader reach and better reception quality. Other factors that increase probability of customers defecting to the competition include slow or inadequate response to complaints and billing errors. Factors such as packaging prices, inadequate features, and older technology may also cause customers to defect to the competition. Customers often compare their providers with others and churn to whoever they feel provides better overall value. A telecommunication company can do just fine if it can take care of existing customers even if it means acquiring no new customers. Globally, the average churn rate among mobile users in telecom industry has been estimated at about 2 percent, which translates to total annual loss of about $100 billion. Kotler estimated that the cost for convincing a regular customer not to churn to the competitor is 16 times less than the cost of searching and establishing contact with a new customer and the cost of attracting new customers is 5 to 6 times more than that for retaining existing ones. Reichheld and Sasser [9] estimated that a service provider can increase profits by between 25 and 85 percent by reducing customer churn rate by 5 percent. This shows the huge impact customer churn rate can have on the business achievement. An analysis of churn rate in different industries shows that it is particularly a major problem in telecommunication industry where it ranges between 20 to 40 percent annually. Technological advancements have helped companies understand that their competitive strategies should ensure high customer retention rates in order to survive in the industry. This especially applies to the telecommunication industry. Thus, significant research activity is now focused on identifying customers with high probability of fleeing to the competition. The deregulation of the telecom industry has increased competition and the situation is only made worse by the fact that customers have more choices than ever.

Just an improvement of 1 percent in customer retention rate could boost company's share price by 5 percent. Poel and Lariviere stated some economical value of customer retention; Successful customer retention means businesses don't have to seek potentially high-risk customers; thus, it can better focus on the needs of existing

customers. Having stored data about long term customers helps companies to understand them well and they become less costly to serve and satisfy. Another economical benefit is that long-term customers are less responsive to competitors' messages. Usually, people tend to share negative experience more than positive ones with friends and relatives. This will create negative perceptions of the company among prospective customers.

Data mining techniques and their applications in customer churn analysis:

The first paragraph under each heading or subheading should be flush left, and subsequent paragraphs should have a five-space indentation. A colon is inserted before an equation is presented, but there is no punctuation following the equation. All equations are numbered and referred to in the text solely by a number enclosed in a round bracket. Ensure that any miscellaneous numbering system you use in your paper cannot be confused designation. In the last few decades there have been significant improvement and changes in the data volumes stored in files, databases, and other repositories. To aid in the decision-making process, it is necessarily vital to come up with powerful techniques of data analysis and interpretation as well as develop tools that can be important in the extraction of interesting hidden patterns and knowledge. Data mining algorithm has the capability of unveiling these patterns and their hidden relationships, and it is an integral component of a complex process that is commonly known as the Knowledge Discovery in Databases (KDD) which explains the steps that must be taken to ensure comprehensive data analysis. According to Shearer [18], CRISP-DM model stands for Cross Industry Standard Process for data mining model. It is mainly for conducting a data mining process, whose life cycle consists of six phases as shown in

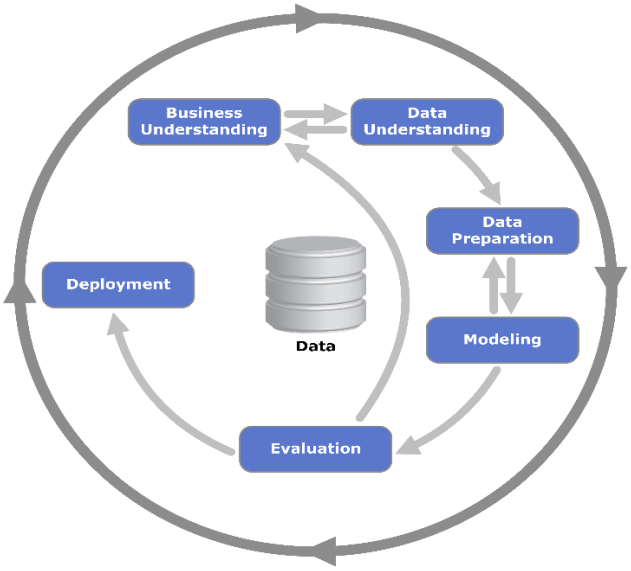

Fig.CRISP-DM model

The first step is to understand the data that serves commercial values. Data preparation entails pre-processing of the raw data containing limited information. This may sometimes involve removal of missing values, quantizing, conversion of categorical variables into numerical. The modelling process involves building a suitable model used to extract the information and also evaluate the information to serve business purposes and accepting the same model after checking for important attributes like performances and accuracy. The final stage involves generation of a report or implementing a repeatable data mining process across the entire firm involved as a deployment and last phase.

Proposed Work:

The study of predicting which persons are going to churn in advance will help the telecommunication industry and the CRM department to identify which persons are going to leave the network. The problem of our work discussed is the classification problem i.e. to classify each subscriber as potential churner or potential non

churner. The framework discussed below is based on the Knowledge Discovery Data (KDD) process [19]. Our framework consists of the following five modules:

Data Acquisition: Acquiring data from the teleset industry is a big task because of the fear of misusing it. The data set for this study acquired from the KDD Cup 2009. It is used to analyse the marketing tendency of customers from the large databases from the French Telecom company Orange [20].

Data Preparation: Since the dataset acquired cannot be applied directly to the churn prediction models, so aggregation of data is required where new variables are added to the existing variables by viewing the periodic usage behaviour of the customers. These variables are very important in predicting the behaviour of customers in advance as they contain critical information used by the prediction models.

Data Pre-processing: Data pre-processing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand. The data gathered from multiple sources first is aggregated and then cleaned as the complete data collected is not suitable for modelling purposes. The records with unique values do not have any significance as they do not contribute much in predictive modelling. Fields with too many null values also need to be discarded.

Data Extraction: The attributes are identified for classifying process. In our work, we have worked with numerical and categorical values.

Decision: The rule set will let the subscribers identify and classify in the different categories of churners and non churners by setting a particular threshold value.

The framework design used in our work is given in Figure 2. We have taken orange dataset which consists of total 18000 attributes.

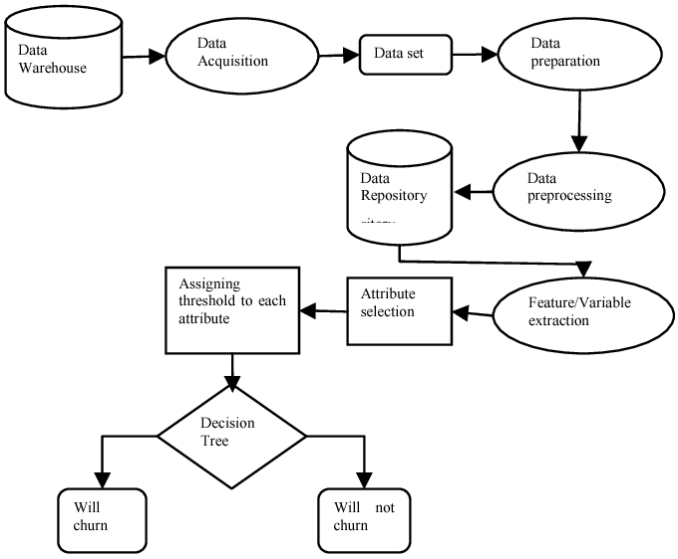

Fig.Churn Prediction Framework

To evaluate a customer's dataset by developing a decision tree the classification is done by altering the tree until a leaf node is attained. When evaluating a customer record a value of churner or non-churner is assigned to its leaf node. Fig 2 presents a simplified decision tree for customer churn prediction in telecom industry.

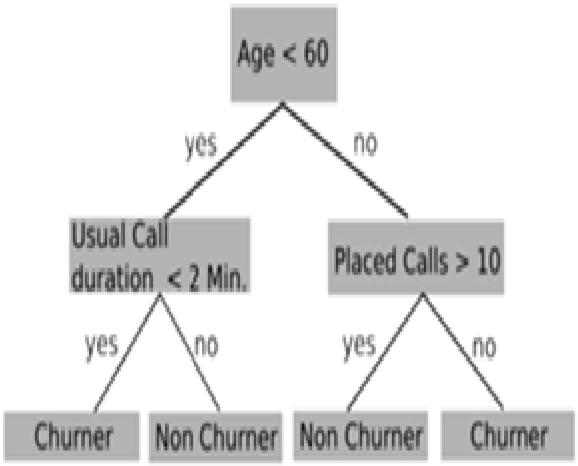

Fig 2 : A simplified churn prediction decision tree

Steps Involved:

1. Understand the data.
2. Univariable study.
3. Multivariate study.
4. Basic cleaning.
5. Test assumptions

Understand the data

Data understanding focuses on the comprehension of the information available in the project. In this step we basically check on the kind of variables provided with the dataset, dtype of the columns, shape of the data frame.

Null values Treatment

Our dataset contains numbers of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

Pandas **isnull()** and **notnull()** methods are used to check and manage NULL values in a data frame The percentage of null values in each variable is found using the following formula.

$$\text{Percentage} = \frac{\text{Number of null values}}{\text{Total number of values}}$$

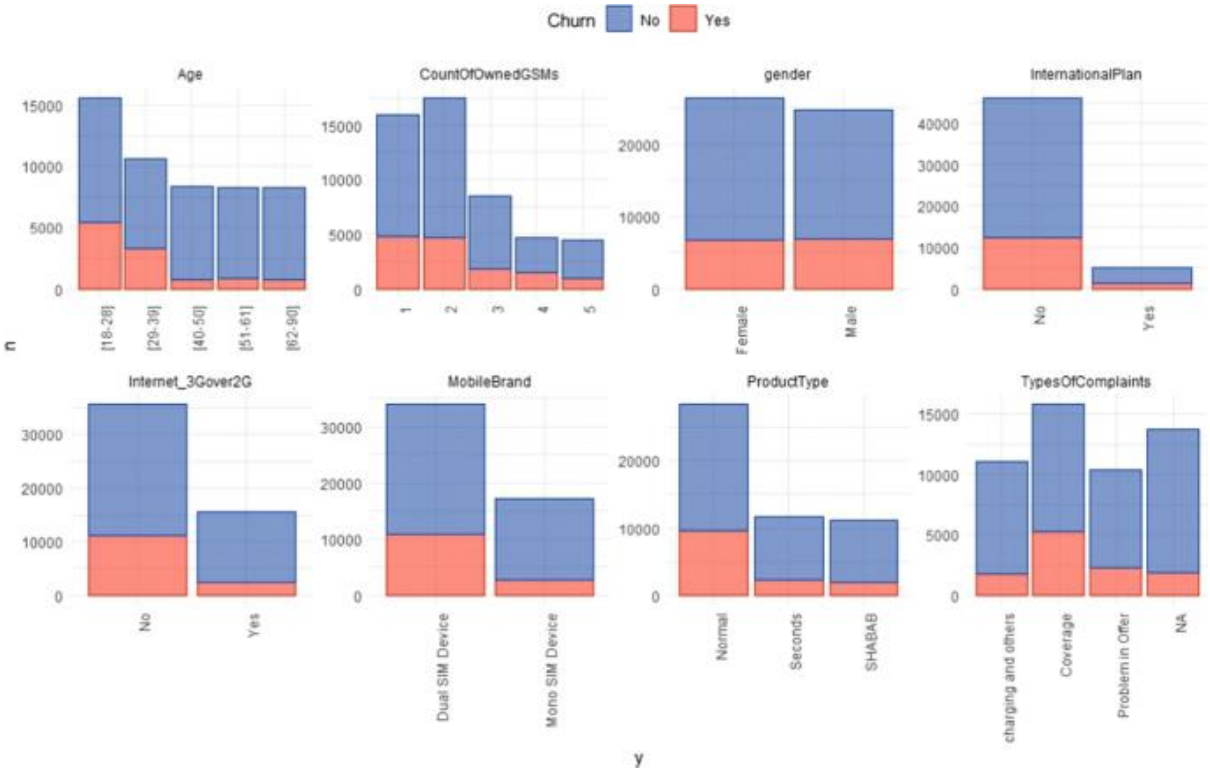

Fig. Customer churn prediction in telecom using machine learning

Conclusion:

There are likely to be tremendous rates of research in data mining and their applications in customer churn, but still, it is an active research field and researchers are searching for more accurate solutions. In this paper we provide a summary of the different data mining methods, and their applications in customer churn prediction. However, from the literature survey it is evident that there has been little research work on covering algorithms and their applications in customer churn, especially when it comes to applying Rules family algorithms in customer churn analysis. Our future work will be applying RULES family techniques on telecom datasets and compare the results with some of the most commonly used techniques in churn prediction as they are very suitable tools for data mining applications.