***Prediction Of Customer Churn In Telecom Industry***

***Abstract— The main task of customer churn prediction is to estimate subscribers who may want to leave from a company and provide solutions to prevent possible churnes. In recent years, estimating churners before they leave has became valuable in the environment of increased competition among companies. The research in this paper was done to estimate churners for companies in the telecommunication industry showing how prediction efficacy is increased by balancing the data with down sampling and classifying by the rotation forest method. The performance level of these techniques are compared with Antminer and C4.5 decision tree. The comparisons are done by using the dataset taken from American Telecommunication Company and accuracy, sensitivity and specificity are used for the performance criteria.***

***Keywords–Customer Churn Prediction; Data Mining; Machine Learning; Telecommunication; Logistic Regression.***

***I. INTRODUCTION***

*Customer churn has became highly important for companies because of increasing competition among companies, increased importance of marketing strategies and conscious behavior of costumers in the recent years. Customers can easily trend toward alternative services. Companies must develop various strategies to prevent these possible trends, depending on the services they provide. During the estimation of possible churns, data from the previous churns might be used. An efficient churn predictive model benefits companies in many ways. Early identification of customers likely to leave may help to build cost effective ways in marketing strategies. Customer retention campaigns might be limited to selected customers but it should cover most of the customer. Incorrect predictions could result in a company losing profits because of the discounts offered to continuous subscribers. Therefore, the right predictions of the churn customers has became highly important for the companies. The prominent role that the telecommunication sector has come to occupy worldwide makes it all the more important to develop prediction mechanisms along the lines of churn prediction. Few statistics show the importance of the customer retains in this sector. One of the remarkable studies shows that 1% increase in the customer retain campaigns may result in the 5% increase in the overall values of the companies [1]. In wireless network telecommunication industry, the monthly rate of customer churn is 2.2% and the annual rate of customer churn is 27% [2]. The yearly cost of customer churn is 4 billion dollars in Europe and America, and it is 10 billion dollars in the entire world [2]. We may suppose that 1.5 million customers would stay in the same company by increasing the correct prediction at the rate of 1%. This may yield to 54 million dollars benefit for the companies annually [3]. In the literature, many researches have been conducted to increase the prediction rates of costumer churns in the telecommunication industry. The scope of this researches covers creating new models, developing existing models, combining of existing models, attribute derivation and outlier analysis techniques. Tsai and Lu used two different hybrid models to develop a customer churn prediction model. The developed hybrid model is a combination of two artificial neural networks and the second hybrid model is a combination of self organizing maps and artificial neural networks. First models are used for data reduction and second models are used for actual classifier [5]. Kechadi and Buckley used attribute derivation process to increase the correct prediction rate [2]. Bayesian Belief Network method is tried in a study which is conducted by Kisioglu and Topcu [1]. Verbeke et al. increased the accuracy by using two different rules extraction method. This methods were AntMiner+ and ALBA [6]. Bock and Poel used two different rotation based ensemble classifiers. These are Rotation Forest and Adaboosts [7]. Yeshwanth et al. suggested a new hybrid model that combines C4.5 decision tree and genetic programming [8]. Zhao et al. used one class support vector machine to increase the performance [3]. Ghorbani et al. created a new hybrid model by combining neural network, tree models and fuzzy modeling [9]. Ant-Miner+ algorithm is working by using the ‘divide and conquer’ technique. Firstly, it starts with all of the training data. Then it creates the best rule which includes a subset of training data and then the best rule is added to the list of previously discovered rules. After that the samples which are covered by this rules are removed from the training data and everything starts again with the reduced training data-set. This iteration continues until when there is only a few remaining samples in training data. At this stage, a default rule is created which covers the remaining samples. Logistic Regression method is a new generation Regression algorithm. It bases to creating subsets by using principal component analysis method as a feature extraction technique [4]. In this research, it has been observed that Logistic Regression method gives better results than antminer+ method which is used by Verbeke. To make comparison, the same data-set is used with Verbeke’s research and same evaluation criteria, such as accuracy, sensitivity and specificity ratios are examined. It is accepted that; supposing a customer that will leave as would not leave and losing him is much more important than giving unnecessary promotions to customers who will not leave as would leave. For these reason sensitivity is seems to be more important than specificity. The rest of the paper is organized in this following manner: Section 2 explains the Logistic Regression. Section 3 presents the data, data processing and evaluation criteria. In section 4, results of rotation forest, Ant-Miner+ and C4.5 methods are compared. Finally, our conclusion is offered.*

*II. LOGISTIC REGRESSION*

***Churn Prediction***

*In terms of telecommunication the customers leaving the current company and moving into another is called as churn, and in the present scenario considering the surge in the number of churn customers the industry is trying its best to retain the profitable customers and this is named as churn management.*

*The objective of the paper is to classify the possible customers that might churn*

*In many areas statistical analysis is used to predict the customers that might churn.*

*The outcomes of churn analysis are the below*

* *Improved retention*
* *Propensity modelling*
* *Prioritized marketing*
* *Increased customer value*

*The types of churn can be classified as*

1. *Voluntary churn*
2. *Involuntary churn*

*When the industry decides to eliminate the customer then it becomes an involuntary churn. The reason behind this might be due to fraud activities, nonpayment of bills and people who would not use the services. When the customer decides to leave the company then it becomes a voluntary churn and such kind of churn happens in a abrupt and unexpected manner. So it is very important to identify the customers that might churn and avoid the same, this can be achieved by doing a churn analysis.*

***Churn Managemet***

*Since acquiring new customers is challenging it is very important to retain the current customers.*

*Churn can be reduced by analyzing the past history of the potential customers systematically. Large data is maintained about the customers and on performing a proper analysis on the same it is possible to predict the probable customers that might churn. The information that is available can be analyzed in different ways and thereby provide various ways for the operators to envisage the churning and evade the same.*

***Data collection***

*For analysis the data that is available in the telecom dataset has been used and prediction has been done for the same.*

***Data preparation***

*Before the data can be analyzed we have to clean the data and keep it ready so that the desired results can be derived from it. Data has be clean so that the redundancy and errors can be removed because having such data will lead to incorrect results as well.*

*In this paper a Churn Analysis has been applied on Telecom data, here the agenda is to know the possible customers that might churn from the service provider. Python programing is used for the same this will help give a statistical computing for the data available, here backward logistic regression is been used to achieve the same. The end result would give us the probability of churn for each customer.*

*Here to do churn analysis Logistic regression is been used, Logistic regression is a statistical method here the resultant variable is categorical, rather than continuous. Logistic regression limits the prediction to be in the interval of zero and one.*

*In this paper we are using Logistic Regression, this involves taking the variables as features and calculating the churn rate of the customer based on the mentioned features. The logistic regression is easier algorithm for checking out the churn details.*

*The dataset that is used has 22 variables available. These are related to Gender, customer\_id, PhoneService etc. The dataset has over 7043 customer related information available.*

*After applying Logistic regression the approach inserted the new cleaned data into a new file called as the “prediction”. This new column will give the probability of the customers that might churn from the telecom provider, here the data that is used is of 2016 and we know the customers that are churned, we are trying to use Logistic regression model here and thereby come to conclusion is the model is accurate and based on the accuracy then decide if the model has to be considered for future predictions or not.*

***Prediction***

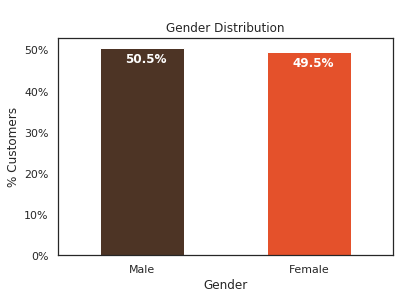
*The business is interested in the final product and it is very important to represent your result in a “graphical representation” such a way that it is understandable and the result helps business make the needed predictions which in turn brings profit. There are many tools that help achieve the same for example, Tableau, Power BI, qlikview etc.*

***Data Visulation Tools***

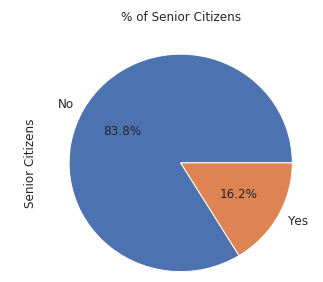
*The best way to get your message across is to use visualization tools, by representing data visually it is possible to uncover the surprising patterns and the patterns that would go unnoticed if we took the stats alone.*

*Here in this approach, data is already cleaned and the result is populated in a file called “Prediction” which will be used to visually show how the data appears and the impact.*

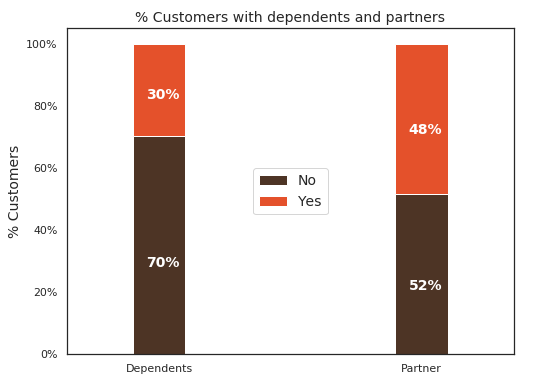
***Gender classification:***

**

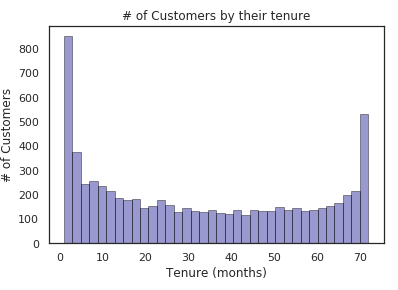
***Senior citizen classification:***

** ***By classifying the data into senior citizens and non senior citizens we can able to calculate the churn rate is more among which group of people and provide offers or any attractive measures can be taken to decrease the churn rate***

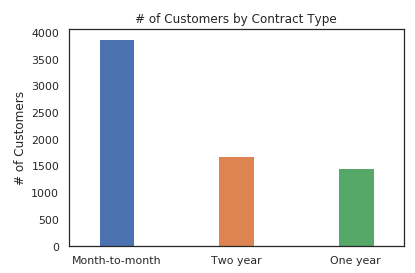
***Partner and dependent status***

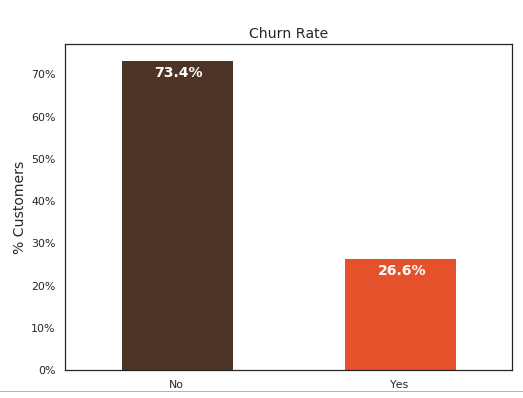
*****About 50% of the customers have a partner, while only 30% of the total customers have dependents***

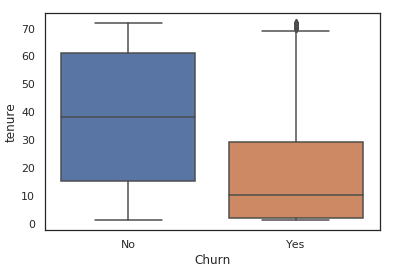
***Based on the customer tenure:***

*****if the customer’s are from a long period of time there is less churn and the probability of churning for new customers is high***

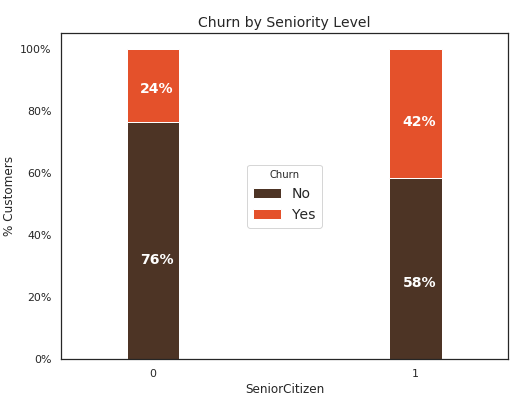
***Contracts:***

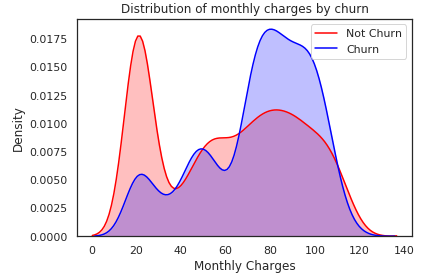
** ***Churn Rate:***

*****This bar plot is used for predicting the churn rate and the interactions with other important variable   
Churn Rate vs Tenure:***

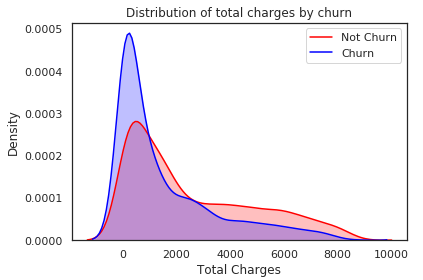
*****As we can see from the plot, the customers who don’t churn, they tend to stay for a longer time in telecom company***

***Churn Rate vs Seniority:***

*****senior citizens have almost double the churn rate than younger population  
Churn By Monthly Charges:***

*Higher % of customers churn where the monthly charges are high*

***Churn By Total Charges:***

*****it seems that there is higher churn where the total charges are lower***

***Results***

*Telecommunication industry always suffers from a very high churn rates when one industry offers a better plan than the previous there is a high possibility of the customer churning from the present due to a better plan in such a scenario it is very difficult to avoid losses but through prediction we can keep it to a minimal level. In this paper the method used is Logistic Regression (backward logistic regression) and this helps to identity the probable churn customers and then make the necessary business decisions. Using a decision tree would give a more appropriate result, by using logistic regression the result achieved is 80.02% accurate.*

***Future Scope***

*The future scope of this project would be to get an accuracy of 100% and for this we can make use of decision tree and in particular using a hybrid classification techniques to point out existing suggestion between churn prediction and customer lifetime value. The result and the accuracy can be bettered if we use more variables in the data. The dynamic nature of the industry has ensured that data mining has been increasingly significant and the industry is totally relying on the results that the data would help predict.*

***Reference:***

1. *Business Intelligence and Insurance, White Paper, Wipro Technologies,Bangalore,2001*
2. *Telcom dataset from kaggle(*[*https://www.kaggle.com/puja19/telcom-customer-churn*](https://www.kaggle.com/puja19/telcom-customer-churn)*)*
3. [*https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/*](https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/)
4. *Rehman, Ali Mustafa Qamar, Aatif Kamal*
5. *Wei, I. Chiu, Turning telecommunication call details to churn prediction: a data mining Approach expert System with applications.*