Advanced Research in Electrical and Electronic Engineering
p-ISSN: 2349-5804; e-ISSN: 2349-5812 Volume 6, Issue 1 January-March, 2019, pp. 16-19
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Analytics in Telecom Industry

Vatsala Vatsyayana[1] and Mohd Khurram Shakir[2]

[1]Lal Bahadur Shastri Institute of Management, Delhi
[2]Lal Bahadur Shastri Institute of management, Delhi
E-mail: [1]vatsala_vatsyayana@lbsim.ac.in, [2]mohd_khurramshakir@lbsim.ac.in

Abstract—*Analytics in Telecom industry is basically a technique of business intelligence which is specially applied and designed to satisfy the complex needs of telecommunication organizations. Also, Telecom analytics helps in decreasing operational costs and maximizing the profits by reducing fraud, increasing sales and improving risk management. The telecom industry is observing a cluster of competition and an era of hyper consumerization. The most immediate and significant impact is the increase in customer churn which is the measure of the number of customers moving out of a particular service over a repetitive period. Due to this the telecom industry operators are investing more efforts in retaining and maintaining a customer base. The charges of services to those customers remains same, leading to stagnant revenues but other factors which are leading to churn are to be improved. Thus, the operators are focusing broadly on three business areas that are fulfilment, revenue and a real world example is that Airtel is losing it's postpaid customers at regular intervals nowadays. For this we have collected certain data which after analysing gave us the precise idea of the factors which are affecting the customer base of the company. The result of this interpretation is that the churn rate depends on the "CALL DROP" and "PAYMENT" factors.*

Keywords*: Churn, Telecom, Telecommunications, Consumerization, Loyalty.*

Introduction:

Data Analytics is a very impressive field of study used as a necessity in almost every organization. This is an interdisciplinary field of study based on the collection processes and systems to extract the knowledge, information and insights of data in different forms and converting the data into living actions, irrespective of the fact that the data is unstructured, semi-structured or structured. This is quite same as to gain information from the set of huge database of consisting of different types of information. Nowadays the biggest problem that the telecom industry is facing today is the churn. Churn rate is defined as the rate at which the customers switch from one service to another .This churn rate depends upon various factors which influence a customer to change from one network to another.

Analytics in Telecom Industry

Classification Techniques:

Predicting future churners is of utmost importance to the telecom industry. Various classification methods are deployed for the predictions. Thus it becomes essential to identify the correct method for this purpose.

Some classification techniques useful for this purpose are:

Famous Techniques gaining momentum by Telecommunication industries to retain customers:

Upselling:

Upselling is defined as a strategy to sell a more advanced and expensive version of a product that the customer is buying or he/she already has that product plus adding extra features as well as add-ons to the same.

Upselling is important because primarily it's easy money. Besides that, when used properly, upselling also allows the industry to come close to the customers and create additional product awareness. Upsells are mostly small purchases that the buyer doesn't have to these upsell think a lot about. The advantage is those are extremely profitable for salesperson and eventually are also important and profitable for the company.

Cross-selling:

Cross-selling is the concept of selling products that vary with each other but should be related to the product the customer is buying or he/she already has that already bought. For example, a customer is buying a Television set and the salesperson offers that person a PS, then this condition is called a cross-sell.

Random Forest Method:

Random forest was originally a method of combining several CART style decision trees using the process of bagging .The early development of random forest was influenced by the method of random subspace of Ho in 1998, the work of Geman & Amit in 1997 and the method of random split selection from Dietterich in 2000. Some core ideas of random forest also matches the early work of Carter & Kwokt in 1988 which is based on the concept of ensembles of decision trees. Random forest models are created by combining the analysis of several trees and each of the tree is being trained separately.

There are three main assumptions that are needed to be made while constructing a random tree. These are:

(1)Splitting of the leaves.

(2)Predictor type to be used in each leaf

(3) Injecting randomness into the trees.

Explanation of (1): For the purpose of splitting the leaf, a bunch of candidate splits are gathered or created and building a criteria to choose between them. A more useful approach is to choose that candidate split which is responsible for optimizing a purity function over the leaves.

Explanation (2): The most useful choice for predictors of each leaf is to make the use of the average response on the training points which are falling in that particular leaf.

Explanation (3): Injecting and applying the randomness into the construction of tree can be done in many ways. In this case, the thresholds are to be chosen randomly or by optimization of some or all of the data in the leaf or it can be done randomly.

Findings:

Finding 1: According to the research, it can be seen that the selection of variables is simple in case of Random Forest than in Logistic Regression. The reason behind this is that, Random Forest are easy to apply when the data is large as compared to Logistic Regression.[by Raphael Couronné Philipp Probst and Anne-Laure Boulesteix on 27 June 2018]

Finding 2: The overall results were observed for about 240 datasets and in those Random Forest has shown better acuaracy in about 70% of the cases than logistic regression which has shown better results in about 30% of the cases. As a whole our result indicate the increase in the use of Random Forest method having default parameter values as a standard method. [Raphael Couronné, Philipp Probst, Anne-Laure Boulesteix in 2017]

Finding 3: In variety of cases, applying data by using Random Forest over fitting gets reduced. [by Gerard Biau in 2012]

Logistic Regression:

Logistic regression analysis is one of the most widely used and preferred regression methods that are used in implementation of models which are having binary dependent variables. Logistic regression also defines the relationship between the binary result variable and independent variables comprising of both continuous and discrete variables.

There are 3 types of Logistic regression

(1)**Binary Logistic Regression** (BLOGREG) Analysis: It is the logistic regression analysis that is to be made with the help of dependent variables which includes binary results .

(2)**Ordinal Logistic Regression** (OLOGREG) Analysis: This as the name suggests is a method implemented when the resultant variable is ordinal. In case of coding ordinal scaled data the results should have a structure of natural order such as didn't like/like/like much etc.

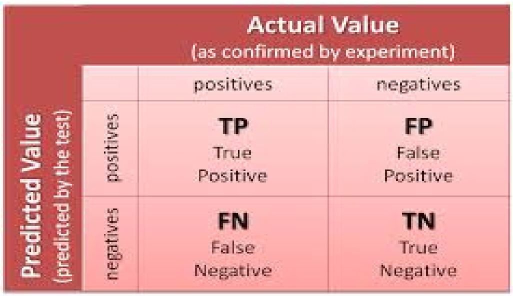
(3)**Nominal Logistic Regression** (NLOGREG) Analysis: It is the method which is implemented when the result variable is nominal. In case of coding the values the categories do not have to be in order.

Findings:

Finding 1: The study suggested that Logistic Regression gives better results than Random Forest technique when we consider the overall correct classification rate. Various strategies were taken for handling the missing values in case of logistic regression as well as Random Forest. [Ming Geng BMed,1992]

In logistic regression the missing values for every variable are grouped in a separate category having the label as "unknown", but in case of Random Forests, the advanced approach is adapted to automatically handle missing values. It is observed that this advanced approach is not effective when data values are not missing at random because the overall classification rate obtained from this advanced approach is very close to that obtained from the simple method.

Finding 2: Also in the research it was observed that the Support Vector Machines requires two parameters to be chosen for predictions those are kernel and slack variable, So when this data is applied to the Logistic Regression, it shows good results but when applied to Random Forest technique the data haven't shown good and accurate results.[by Tomas Pranckevicius, Virginijus Marcinkevicius in 2017]



		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

Explanation: In the above figure we have seen that there are 4 quadrants:

(1) 1st quadrant is "TP"- It means that whatever probability have been predicted in this quadrant actually exists.

(2) 2nd quadrant is "FP"- It means that whatever probability have been predicted does not exists or matches with actual probabilities.

(3) 3rd quadrant is "FN"-It means that probability prediction states that the criteria doesn't exist but in reality it do exist.

(4)4th quadrant is "TN"- It means that whatever predicted probability states that there is no presence of the criteria but it reality we do have presence of the criteria.

Accuracy: It is as the name suggests defines or determines the accuracy of the model.

Accuracy = TP + TN/ TOTAL

Misclassification Rate: It is the rate which tells that overall how often the model is wrong.

The formula for Misclassification Rate is:

Misclassification = FP + FN/TOTAL

Sensitivity: It is defined as how many times it was actually "yes" when we predicted the model as "yes". Sensitivity = TN/ACTUAL NO

Specificity: It is defined as how many times it was actually "no" when we predicted the model as "no". Formula = TN/ACTUAL NO.

Precision: It is defined as that when the model predicts "yes", so how often it is a actual "yes". Formula = TP/PREDICTED YES

Prevalence: It is defined as how often the "yes" condition actually occurring in out model.

Formula = ACTUAL YES/TOTAL

Decision Tree Technique:

Decision Tree is a graphical representation having branches to show various outcomes and results of the decision.

Findings:

Findings 1: The biggest problem with the results obtained by applying Logistic Regression is that they are difficult to interpret. This is the reason why decision trees are preferred to Logistic Regression.[Decision Trees Are Usually Better Than Logistic Regression by Tim Bock]

Finding 2: A research paper was read on the prediction through Decision tree and it was found out that while predicting churn rate of a data set the level of accuracy of Logistic Regression and Decision Tree is almost same (Logistic Regression-80.65% and Decision Tree-80.63%).So this denotes that Decision Tree also gave almost same level of accuracy and is easy to interpret than Logistic Regression.[Decision Trees Are Usually Better Than Logistic Regression by Tim Bock]

Finding 3: It was noticed that Decision Tree models are preferred to Logistic Regression models because they explain the factors of the model in a much more informative way than Logistic Regression models.[By Rajesh S.Brid 2018]

Conclusion:

Telecom industries have large data sets and to find out better results mostly Decision Tree technique is preferred to Logistic Regression because the models obtained from Logistic Regression are complex to interpret. It is also observed that in an experiment done on various types of data sets Random Forest models have shown much better results than models of Logistic Regression. But when we consider the overall correctness of the classification rate Logistic Regression models tends to have more correctness than models of Random Forest.

If the data has continuous variables then Logistic Regression yields better results but overall if the data is having independent variables as categorical then Random Forest's results are much better. Thereby it can be safely concluded that no technique is better than the other, it depends on which type of data or conditions a researcher is considering while creating the model. In the research process, it could be seen that features which are mostly responsible for the increasing churn rate in telecommunication industries are Call drop and Payment. This has been concluded because, they are the two factors which are taken as independent in most researches.

References:

[1] 1.Customer Churn Analysis: Using Logistic Regression to Predict At-Risk Customers by Sunil Kappal Dec 14,18. https://dzone.com/articles/customer-churn-analysis-using-logistic-regression

[2] A comparison of Logistic Regression to Random Forests for exploring differences in risk factors associated with stage at diagnosis between black and white colon cancer patients, 1992 http://d-scholarship.pitt.edu/7034/1/realfinalplus_ETD2006.pdf

[3] Random forest versus logistic regression: a large-scale benchmark experiment Raphael Couronné, Philipp Probst, Anne-Laure Boulesteix in 2017, https://epub.ub.uni-muenchen.de/39955/1/TR.pdf

[4] Decision Trees Are Usually Better Than Logistic Regression by Tim Bock https://www.displayr.com/decision-trees-are-usually-better-than-logistic-regression/

[5] Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies by Adnan Idris and Asifullah khan,https://www.researchgate.net/publication/256918723_Churn_prediction_in_telecom_using_Random_Forest_and_PSO_based_data_balancing_in_combination_with_various_feature_selection_strategies

[6] Random forest versus logistic regression: a large-scale benchmark experiment by Raphael Couronné Philipp Probst and Anne-Laure Boulesteix on 27 June 2018, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5

[7] Analysis of a Random Forests Model by Gerard Biau in 2012. , http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

[8] Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification by Tomas Pranckevicius, Virginijus Marcinkevicius in 2017.
https://www.researchgate.net/publication/318056374_Comparison_of_Naive_Bayes_Random_Forest_Decision_Tree_Support_Vector_Machines_and_Logistic_Regression_Classifiers_for_Text_Reviews_Classification
Decision Trees - A simple way to visualize a decision – Medium Medium by Rajesh S.Brid 2018
https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

[10] Churn Analysis in Telecommunication using Logistic Regression by Helen Treasa and Rupali Wagh in 2017
http://www.computerscijournal.org/vol10no1/churn-analysis-in-telecommunication-using-logistic-regression/
Analysis of Churn Prediction: A Case Study on Telecommunication Services in Macedonia
Aleksandar J. Petkovski, Biljana L. Risteska Stojkoska, Kire V. Trivodaliev, and Slobodan A. Kalajdziski
Analysis of Churn Prediction: A Case Study on Telecommunication Services in Macedonia
Aleksandar J. Petkovski, Biljana L. Risteska Stojkoska, Kire V. Trivodaliev, and Slobodan A. Kalajdziski

[11] Analysis of churn prediction: A case study on Telecommunication of services in Macedonia by Aleksander J. Petkovski, Biljana L.Risteska Stojkoska, Kire V.Trivodaliev and Slobodan A. Kalajdziski
https://www.researchgate.net/publication/312573014_Analysis_of_churn_prediction_A_case_study_on_telecommunication_services_in_Macedonia

[12] Comparison Between SVM and Logistic Regression: which one is better? By Diego Alejandro Salazar, Jorge Evan Velez, Juan Carlos Salazar in 2012,
https://www.researchgate.net/publication/260773845_Comparison_between_SVM_and_Logistic_Regression_Which_One_is_Better_to_Discriminate

[13] Customer churn prediction using improved balanced random forests
https://www.researchgate.net/publication/222929949_Customer churn_prediction_using_improved_balanced_random_forests.