IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8692

CUSTOMER CHURN PREDICTION

Senthilnayaki B [1], Swetha M [2], Nivedha D [3]

Teaching Fellow, Information Science and Technology, CEG. Anna University, Chennai, Tamil Nadu[1]

UG – Information Science and Technology, CEG. Anna University, Chennai, Tamil Nadu[2]

UG – Information Science and Technology, CEG. Anna University, Chennai, Tamil Nadu[3]

Abstract Customer churn analysis and prediction in telecom sector is an issue now a days because it's very important for telecommunication industries to analyze behaviors of various customer to predict which customers are about to leave the subscription from telecom company. So machine learning techniques and algorithm plays an important role for companies in today's commercial conditions because gaining a new customer's cost is more than retaining the existing ones. This project focuses on various machine learning techniques for predicting customer churn through which we can build the classification models such as Logistic Regression, Random Forest and lazy learning and also compare the performance of these models.

Keywords— churn , machine learning , Logistic regression , Random Forest , K-nearest-neighbors

1. INTRODUCTION

The customer who cease a product or service for a given period is referred as churner. In a telecommunication company, the individual who has opted service from a firm is referred to as Churn. The individual who probably intends to depart from the firm in near future was predicted by the churn model. Many industries build a model like a churn as a common application for data mining technique. Mobile telephone organizations present across the globe are almost on the verge of building their own churn model. Furthermore, to retain the customers, churn results can be efficiently utilized for various other goals.Churn Management approach is actually the first step in building a model. In general, the project needs a churn model in the best way instead of taking a single method which has the best lift. So here we have built an automated application as a default for a long run.In this digital era, the client of one company may also be a consumer of one or more telecommunication firms. Some of us may use different carriers based on the distance and some others may use different carriers based on the different plans they offer. While performing the analysis using machine learning customer experience tends to provide valuable insights. Some people will change their service providers from time to time. Increase or decrease in the calling rate will also depend on different job responsibilities. Based on the availability of the data various situations may reflect**.**

2. LITERATURE SURVEY

Irfan Ullah et al., [6] identified churn factors that are essential in determining the root causes of churn. By knowing the significant churn factors from customers' data, Customer Relationship Management (CRM) can improve productivity, recommend relevant promotions to the group of likely churn customers based on similar behavior patterns, and excessively improve marketing.

Kavitha V et al., [7] used a Decision Tree, Random Forest, and XGBoost to predict the customers who are likely to cancel the subscription which can offer them better services and reduce the churn rate.By preprocessing and feature selection ,the data set for training and testing. For the above mentioned algorithm, it is necessary to do some feature engineering to have more efficient and accurate results.

Krishna Sai and Sasikala [8] implemented an EDA using Visualization, statistical tests for feature selection and Data mining methods for predicting the likely churners by utilizing a Logistic Regression Model.Here dataset has been analysed by using the data visualization techniques before entering into the modeling process.

3. SYSTEM DESIGN

It is very crucial to make the data useful because unwanted or null values can cause unsatisfactory results or may lead to producing less accurate results. In the data set, there are a lot of incorrect values and missing values. We analyzed the whole dataset and listed out only the useful features. The listing of features can result in better accuracy and contains only valuable features as to come up with the specific information like the owner, place of registration, address.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8692

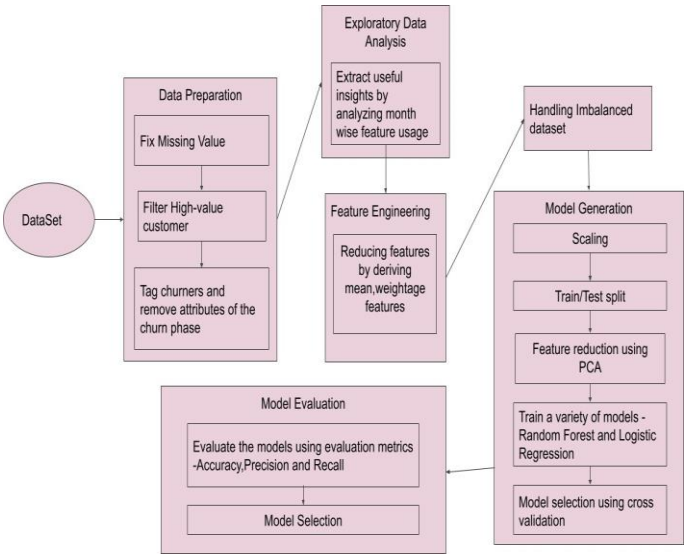

Figure 3.1: Architectural Design for Customer Churn Prediction

Feature selection is a crucial step for selecting the required elements from the data set based on the knowledge. The dataset used here consists of many features out of which we chose the needed features, which enable us to improve performance measurement and are useful for decision-making purposes while remaining will have less importance. The performance of classification increases if the dataset is having only valuable variables and which are highly predictable. Thus having only significant features and reducing the number of irrelevant attributes increases the performance of classification. Many techniques have been proposed for customer churn prediction in the telecommunication industry. Here by using logistic regression, Random Forest and KNN we can predict the probability of a churn i.e., the likelihood of a customer to cancel the subscription and we can evaluate the models using performance metrics like accuracy , precision and recall score.

4. IMPLEMENTATION

Load the dataset and print the first 5 records of the dataframe to check the loaded dataset. Here mobile_number is the unique id column for each customer. Columns are segregated by months of june(6), july(7), august(8), september(9) for the year 2014. It has about a lack of customer records and 226 columns. In order to filter the high value customer records, derived the column of average recharge amount of June and July month(the good phase), take only the records that is more than the 70th percentile of the average recharge amount .Drop the remaining records which is not required and print the count of rows and columns of newly filtered data.

Step 1 Get the average of sixth month recharge amount and seventh month recharge amount of all customers
Step 2: : Get data greater than 70 percentile of average recharge amount.
Step 3: Drop unwanted column
Step 4: : *var*1 *Add* 9*th month call f eatures and data f eatures D* Deriving churn column.
Step 5: if *var*1 = *False* **then**
Step 6:fill churn column value with 1(churn = 1)
Step 7: else
Step 8: fill churn column value with 0(non-churn = 0)
Step 9: **end if**
Step 10: Do column split based on month
Step 11: Drop all ninth month features

4.1. HANDLING MISSING VALUE

In order to fix the missing value in dataset check for the count of missing values in the dataset and list the columns with the missing values. Then pass the dataframe to get_cols_split helper function and get the column categories and pass the month's column list to get_cols_sub_split helper function and get the columns sub-categories.here fb_user and night_pack_user columns are of nominal type 0 and 1. Since missing values could be of another type, imputing them as 2. Missing values for some set of columns seem to be as data not available. So imputing them with 0.Few date columns have some missing values. But let's leave that as is for now and will use that later point in time.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8692

Step 1: for jun to aug **do**.
Step 2: Gemonth's incoming, outgoing calls usage and recharge columns using *get col split* procedure
Step 3: end for
Step 4: Fill missing values of fb user and night pack user month columns with 2
Step 5 :Fill missing values of recharge columns with 0.
Step 6: Fill missing values of call usage columns with 0

4.2 EXPLORATORY DATA ANALYSIS

Due to data imbalance churn rate is low in the overall dataset. In order to fix it analysis is performed on certain important features column like age on network(AON), incoming calls usage, outgoing calls usage, operator wise calls usage, recharge amount, recharge count, average revenue per user and 2G and 3G. These columns seem to have outliers at the top percentile which is treated using outliers treatment. The outlier treatment is to cap the outliers at the 99th percentile for the above mentioned features column which derives some mandatory features. Remove the columns of date to perform sample logistic regression on the available data.

.**Step 1:** : **for** *f eatures = AON, ic, og, odu, rech amt, rech count, arpu...* **do**.
Step 2: for jun to aug **do**.
Step 3: Merge column with churn column and plot columns with churn label
Step 4:end for
Step 5:end for
Step 6:for not end of column list **do**
Step 7: Cap the values with the 99th percentile and 1 percentile
Step 8: Apply that percentile to data
Step 9:end for
Step 10: Remove the outliers for specific columns like roam og mou 8,arpu 7,loc og mou 8,loc ic mou 7,std og mou 7

4.3 MODEL GENERATION

Now that the class is well balanced it is splited into train(70%) and test (30%) dataset. Apply PCA on the training dataset for dimensionality reduction and feature selection.Draw the screeplot forthe PCA components and pick the right number of PC components to build the model and chose 60 PCA components for model building using the following PCA algorithm.

Step 1: : Consider a Data with n-dimensions .
Step 2: : Subtract the mean - from each of the data dimensions.
Step 3: Calculate the covariance matrix
.**Step 4:** Calculate the eigen values and eigenvectors of the covariance matrix
Step 5: Reduce dimensionality and form feature vector
Step 6: : Deriving the new data FinalData = RowFeatureVector x RowZeroMeanData
Build models like Random Forest , KNN and LogisticRegression .Stratified k-fold cross validation method is usedto select the best model by estimating their performance

4.4 MODEL EVALUATION

Evaluating the models using appropriate evaluation metricslike accuracy, precision and recall that is more important to identify churners than the non-churners accurately. Draw ROC graph, confusion matrix and classification report for Random Forest model which is predited as the best. Also obtain the number of correctly predicted and wrongly predicted records by random forest model

5. EXPERIMENTAL RESULT

Stratified k-fold cross validation technique is applied to select the best model by estimating their performance metrics. Figure 5.1 shows that Random Forest is the best model with cross validation score of 96.3% where as KNN and Logistic Regression has cross validation score of 88.8% and 81.72% respectively.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8692

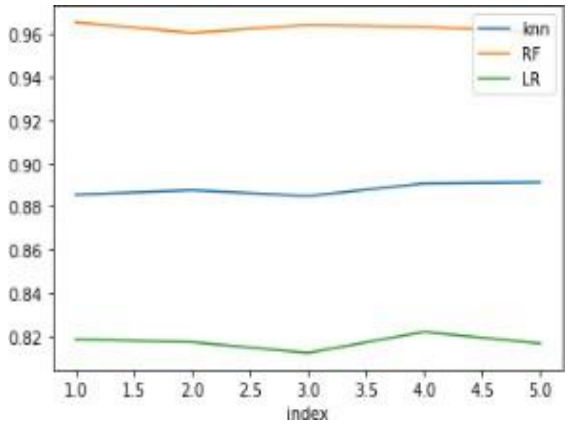

Figure 5.1: Analysis using Cross Validation

Evaluate the Random Forest model based on theirperformance metrics like accuracy, precision and recall that is more important to identify churners than the non-churners accurately. Figure 5.2 shows the results obtained while performing the experiment using the Random Forest algorithm and can check the accuracy. This clearly shows that Random Forest performs better with non-linear data than other machine learning models.

Figure 5.2: Classification Report

Classification Report :

	precision	recall	f1-score	support
0	0.97	0.94	0.96	8241
1	0.94	0.97	0.96	8169
accuracy			0.96	16410
macro avg	0.96	0.96	0.96	16410
weighted avg	0.96	0.96	0.96	16410

TN = 7755, FP = 486, FN = 208, TP = 7961

Figure 5.3 shows the confusion matrix of the Random Forest model which clearly depicts the correct and incorrect counts of both churn and non-churn. Here the correctly predicted churners and non-churners counts 7961 and 7755 respectively. Non-Churners who are wrongly predicted as churners counts 208 and churners who are wrongly predicted as non-churners counts 486 .

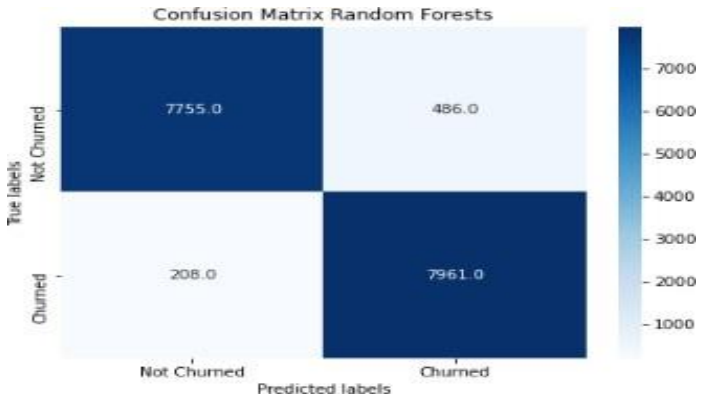

Figure 5.3: Confusion Matrix Of Random Forest Model

6. CONCLUSION

In Telecom Company can have a clear view and can provide them some exiting offers to stay in that service. The obtained results show that our proposed churn model produced better results and performed better by using machine learning techniques.In upcoming time it is necessary to reduce further more features in order to obtain better accuracy and introducing some more machine learning models for better performance.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 6, June 2021

DOI: 10.17148/IARJSET.2021.8692

REFERENCES

1. Abhishek and Ratnesh ,"Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques", In the proceedings of 2017 *International Conference on Advanced Computation and Telecommunication,* Bhopal, India, 2017.
2. Abinash and Srinivasulu U ,"Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco", In the proceedings of *2017 International Conference on Inventive Computing and Informatics* , Coimbatore, India, pp. 721-725, 2017.
3. Trupti S. Gaikwad; Snehal A. Jadhav; Ruta R. Vaidya; Snehal H. Kulkarni. "Machine learning amalgamation of Mathematics, Statistics and Electronics". *International Research Journal on Advanced Science Hub*, 2, 7, 2020, 100-108. doi: 10.47392/irjash.2020.72
4. Alae and El Hassane , "A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers", In the proceedings of *Intelligent Systems and Computer Vision* , Fez, Morocco, 2017.
5. Salini Suresh; Suneetha V; Niharika Sinha; Sabyasachi Prusty; Sriranga H.A. "Machine Learning: An Intuitive Approach In Healthcare". *International Research Journal on Advanced Science Hub*, 2, 7, 2020, 67-74. doi: 10.47392/irjash.2020.67
6. Anuj and Prabin ,"A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services", *International Journal of Computer Applications* ,Volume 27, No.11, pp. 26-31, 2011.
7. Balasubramanian M, and Selvarani M ,"Churn Prediction in Mobile Telecom System using Data Mining Techniques ", *International Journal of Scientific and Research Publications*, Volume 4, Issue 4, pp. 1-5, 2014.
8. Irfan Ullah , Basit Raza, Ahmad Kamran Malik , Muhamad Imran , Saif Ul Islam and Sung Won Kim., "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", In the proceedings of *IEEE Access*, vol. 07, no. 2169-3536, pp. 60134 - 60149, 2019.
9. Varshaa a; Vinitha V; Usha Nandhini D; Yogeshwaran R; Soundharya B M. "Artificial intelligence and its applications- A Review". *International Research Journal on Advanced Science Hub*, 1, 2, 2020, 1-4. doi: 10.47392/irjash.2019.11
10. Kavitha V, Hemanth G , Mohan S.V and Harish M , "Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms", *International Journal of Engineering Research & Technology(2278-0181),* Vol. 9, Issue 05, pp. 181-184, 2020.
11. Kiran and Surbhi , "Customer Churn Analysis in Telecom", *Industry International Conference for Reliablity*,Noida , India , 2015.
12. Krishna B.N, and Sasikala ,"Predictive Analysis and Modeling of Customer Churn in Telecom using Machine Learning Technique ,"In the proceedings of *International Conference on Trends in Electronics and Informatics* , Tirunelveli, India, pp. 6-11, 2019. .
13. Rahul J and Usharani T ,"Churn Prediction in Telecommunication Using Data Mining Technology", *International Journal of Advanced Computer Science and Applications*, Vol. 2, No.2, pp. 17-19, 2013.
14. Roshin Reji , Rohit Zacharias , Sebin Antony and Merlin Mary James, "Churn Prediction in Telecom sector using Machine Learning*", International Journal of Information Systems and Computer Sciences*, Vol. 8, No.2, pp. 832–937, 2019
15. Sato T, Huang B.Q , Huang Y, Kechadi M.T and Buckley B , "Using PCA to Predict Customer Churn in Telecommunication Dataset", *International conference on Advanced data mining and applications*, Vol. 2,pp. 26-27, 2010.
16. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, ''Customer churn prediction in telecommunication industry using data certainty,'' J. Bus. Res., vol. 94, pp. 290–301, Jan. 2019.
17. S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, ''Telecommunication subscribers' churn prediction model using machine learning,'' in Proc. 8th Int. Conf. Digit. Inf. Manage., Sep. 2013, pp. 131–136.
18. T. Jahromi, M. Moeini, I. Akbari, and A. Akbarzadeh, ''A dual-step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers,'' J. Innov. Sustainab., vol. 1, no. 2, pp. 2179–3565