Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4321 - 4326
Received 25 April 2021; Accepted 08 May 2021.

Churn Analysis in Telecommunication Industry using Machine Learning Techniques

Vibhor Shah, Computer Science & Engineering, SRM Institute of Science and Technology, Kattankulathur,

Deepak Harbola, Computer Science & Engineering, SRM Institute of Science and Technology, Kattankulathur,

Dr. S. Thenmalar, Asst. Prof., Computer Science & Engineering SRM Institute of Science Technology, Kattankulathur

ABSTRACT

Customers are the most important pillar of an organizations success hence every company emphasize on giving satisfaction of its services to the customers. As the telecommunication and the information technology sector is growing the number of companies are increasing hence there is a stiff competition in the market. The increasing rivalry among firms has compelled the companies to take the issue of churn seriously in present years. Churn happens when the customer leaves the organization for another company due to advantages like good services, less prices which the other company provides. Since the cost of acquiring new customers is higher than cost to retain the current customers hence the churn analysis becomes an important part to study and get predictions of the possible and potential customers who can churn in the future. The agenda of this work is to predict customer churn using methodologies of machine learning with data analysis. In this work the use of decision tree, k-nearest neighbour, support vector machine is done. Kaggle website is used for dataset purposes.

Keywords—
churn customer, big data, machine learning, decision trees, logistic regression, knn, svm

I. INTRODUCTION

Due to rapid growth in the market of telecommunication sector and information technology the competition between companies is increasing. As the customers can choose among many different operators, companies try to give more time to retain their current clients rather than acquiring new ones. As customers are major source of income for an organization hence predicting customer churn is an important part for the existence of telecom company. Therefore, better tools must be used to provide better insight on predicting churn. This prediction can help the companies to execute better marketing strategies and manage their networks the churn of customers is believed to be a threat to the growth of the company and studies (Umayaparvathi and Iyakutti 2016) have shown that machine learning methodologies are great to predict the churn.

II. STATE OF THE ART (LITERATURE SURVEY)

Many methods and machine learning techniques are being applied to predict and form analytical data on churn reduction. The following studies tells about the existing work and problems to overcome.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4321 - 4326
Received 25 April 2021; Accepted 08 May 2021.

a) Study conducted by Tanneedi tells us that customer churn is a major hurdle for telecom sector. It tells that big data techniques with machine learning are good to find churn. Techniques such as Decision Trees have been used. The result of this study tells us that as volume and quality of service improves, the churn rate decreases.
b) Study of Huang, (2015) lays emphasis on 3Vs which are volume, variety and velocity. This means that by using huge amount of training data from large number of features and increasing velocity of data processing the prediction can be done effectively. This technique gave the accuracy rate of 0.96 of churners from list.
c)Study of Almana, (2014) pointed that prepaid customers are more likely to churn than postpaid customers and emphasized on this factor to predict churn. As prepaid customers are not covered in contracts hence more chances of churning.
d)Arifin and Samopa (2018) stressed on 3 variables - voice, data usage and service and concluded that if company want lesser churn rate than focus on these variables is of utmost importance
e) Alwin (2018) told that churning can be decreased by studying the history of the user analytically. Logic regression model can be of great help to do precise calculations
f) Saini (2017) studied on factors like cost of using another operator is less hence churning. Technique known as exhaustive CHAID used to predict customers who can churn in future.
g) Sjarif (2019) laid focus on the importance for a company to have model which uses KNN algorithm as this results in improving the accuracy form 80% to more than 95%.

III.PRPOSED WORK

Every organization wants less of its customers to churn hence the analysis. Therefore, this study aims to reduce churning by predicting the probability of which customers are more likely to churn. As obtaining new customers is costly hence retaining them is the best idea. The following figure tells us steps to follow for system proposed:

1. Collecting Data - Data of telecommunication industry is used and which will be appropriate for the analysis is used.

2.Data Pre-Processing - This has 3 procedures: first, data cleaning which is done by replacing missing data with null values. Second. feature extraction which is done by setting the target variable and doing its correlation with other variables. Third, transformation of data where string values are converted into numerical values so as to use in machine learning models.

3.Data Preparation - This is done by converting categorical to numeric data, using machine learing techniques for scaling. Identifying the target variable and other data exploratory analysis is also done.

4.Data Estimation/Prediction - The final outcome will tell how much probable is the customer likely to churn. Final output will show the list of customersids with the probability of the churning.

5. Data Visualization - Various python libraries are used for the same like matplotlib, seaborn, pandas_profiling and plotly. The relation between variables is showed using scatter plot, pie chart, bar chart.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4321 - 4326
Received 25 April 2021; Accepted 08 May 2021.

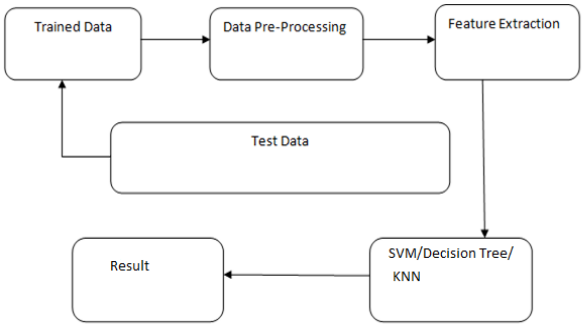

Fig.1System ArchitectureDiagram

IV. IMPLEMENTATION

A. Data Set Analysis - The dataset used is fetched from Kaggle website. First exploratory analysis done using pandas_profiling. Then string values are converted to numeric so as the machine learning model can use them like no in churn is assigned 0 whereas yes is assigned as 1. Also functions are used to replace spaces with null values. Also datatype of charges column is changed to float. Then using pie chart and other plots the correlation between different variables is established.

First rows

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	Multiple
0	7590-VHVEG	Female	0	Yes	No	1	No	No phor
1	5575-GNVDE	Male	0	No	No	34	Yes	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No
3	7795-CFOCW	Male	0	No	No	45	No	No phor
4	9237-HQITU	Female	0	No	No	2	Yes	No
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes
7	6713-OKOMC	Female	0	No	No	10	No	No phor
8	7892-POOKP	Female	0	Yes	No	28	Yes	Yes
9	6388-TABGU	Male	0	No	Yes	62	Yes	No

Last rows

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	Mul
7033	9767-FFLEM	Male	0	No	No	38	Yes	No
7034	0639-TSIQW	Female	0	No	No	67	Yes	Yes

Fig.2 - Dataset

B. Algorithms Used – 1. Logistic Regression - It is model which is used for statistical purposes. It uses logistic function. In this the combination of input value takes place linearly with the help of weight or coefficients to estimate output. The ouput values is binary(0 or 1).

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4321 - 4326
Received 25 April 2021; Accepted 08 May 2021.

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

Fig. 3 - Logistic regression equation

Here the predicted output is y, b0 is intercept term, for single input x the coefficient is b1.

2. K-Nearest Neighbor (KNN) - In this supervised machine learning algorithm labeled data is classified into different labels and regression is applied on them. It makes judgement of input data by taking help by storing the same data. K value is decided and data points are labelled based on its distance with K. Below Manhattan formula can be used to define different distances.

$$D = \sum_{i=1}^{n} |x_i - y_i|$$

Fig 4 Manhattan Distance

3. SVM (Support Vector Machine) - This regression technique divides data points by creating a boundary line and divides them into different classes. This line is Hyperplane. S for support here means that it takes decision on new data by taking help/support of extreme data points. The classes are differentiated by using these extreme points for e.g. a bus and a train can be assumed as data points with dissimilar features treated as extreme points. Now, if we want to identify whether the new vehicle is bus or train than we have to consider the dissimilar feature which are the extreme data points.

4. Random Forest - Uses supervised machine learning method, for both regression and classification problems. It uses method of combining many classifiers to find solution to complex problem called ensemble learning and imporve efficiency of model. It predicts the datasets class by combining multiple trees and hence gives better output as some decision trees could give wrong output. To get stronger random forest classifier we need to make two assumptions: First, the variable should have some real values to get accurate results rather than guessing. Second, there should be low correlations for each tree predictions.

C. Training & Testing – At first the data is visualized using various python libraries like matplotlib, seaborn, plotly and graph is plotted between different variables as an example below scatter plot between tenure and churn rate.

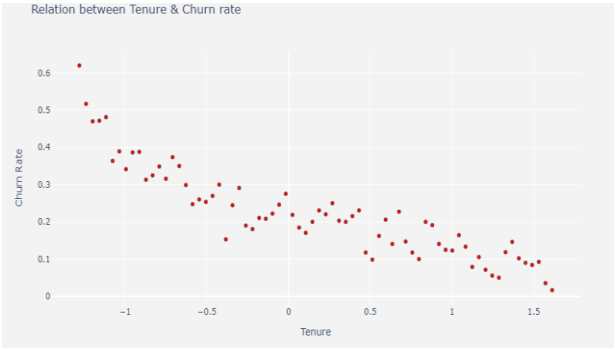

Fig. 5 Tenure & Churn Rate

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4321 - 4326
Received 25 April 2021; Accepted 08 May 2021.

The training set consists of 70% of whole dataset whereas 30% is the testing set. Dataset contains data in numerical and string format which is understandable by human beings but cannot be understood by machine learning models hence it is important that categorical data is converted into numeric format using hot or label encoding so that implementation can be done and we can get accurate results as ML algorithms work better on data which is labeled. Random state is kept as 50 and all four algorithms are applied knn, svm, random forest and logistic regression. At last a confusion matrix is created so as to compare their accuracy with each other. Also scaling is used so as to provide better correlation between variables of interest. The correlation is also derived using Pearson's correlation coefficient (r). Its range is between -1(total negative correlation) and +1(total positive correlation).

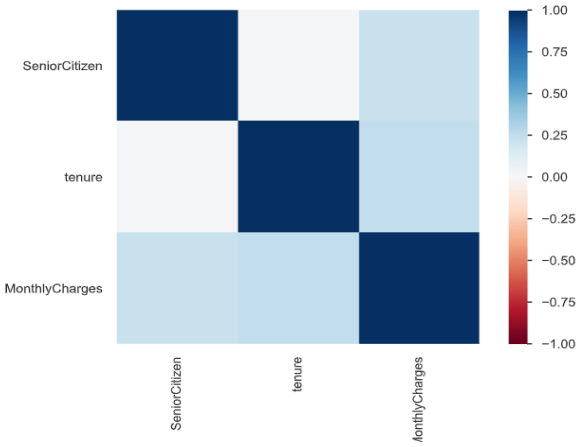

Fig.6Pearson'scorrelationanalysis

V. RESULTSDISCUSSION

First the score of each algorithm is calculated which tells us which one of the algorithm performs better than the rest in the prediction of churn. This comparison made between algorithms tells us the best one to use in churn prediction. The final score is rounded off to 2 decimal places. Using method sort we have sorted the scores in descending order and assigned the index using set_index accordingly. Also after the usage of confusion matrix and using predict_proba method the probability of churn is calculated. From the figure below we can see that logistic regression has the highest score hence will provide with the best customer churn prediction.

Out[28]:

	Score	Model
0	81.14	Logistic Regression
1	80.66	Support Vector Machine
2	79.38	Random Forest
3	76.87	K-Nearest Neighbor
4	73.27	Decision Tree

Fig. 7 Accuracy of MLalgorithms

VI. CONCLUSION

One of the biggest problems of telecom industry is customer churn. So it becomes important for the organization to know the possible customers going to churn so that there will be minimal losses. Hence to maintain a customer base which is loyal to the organization is beneficial as the costs associated with maintaining it is less. The increasing rivalry among firms have made the prices to go down hence the increase in churn, so the need of this

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 4321 - 4326
Received 25 April 2021; Accepted 08 May 2021.

prediction is much more necessary. Big data techniques applied with machine learning have made the prediction and analysis of churn much more easier. In predicting churn ml approaches are critical. The report has three main objectives. One is to look at customer churn. Second, how useful is prediction model of churn in telecom sector. Third, to compare algorithms which will provide better prediction to reduce churn rates. Techniques like logistic regression, KNN, svm have proved to of great help and will also be in the future for a business to analyze its churn rate.

REFERENCES

[1]Tanneedi, N.N.P.P. (2016). Customer Churn Prediction Using Big Data Analytics. Master Thesis, Blekinge Institute of Technology.
[2]Huang, F, Zhu, M, Yuan, K, Deng, E.O. (2015). Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. pp .607–618.
[3]Almana, A.M., Aksoy, M.S., Alzahrani, R. (2014). A Survey on Data Mining Techniques in Customer Churn Analysis for Telecom Industry. Int. Journal of Engineering Research and Applications, 4(5), pp. 165-171.
[4]Alwin, P.K.D.N.M., Kumara, B.T.G.S., Hapuarachchi, H.A.C.S. (2018). Customer Churn Analysis and Prediction in Telecommunication for Decision Making. 2018 International Conference on Business Innovation (ICOBI), pp. 40-45.
[5]Azeem, M, Usman, M. (2018). A fuzzy based churn prediction and retention model for prepaid customers in telecom industry. International Journal of Computational Intelligence Systems, 11(1), pp. 66 – 78.
[6]Saini, N, Monika, S, Garg, K. (2017). Churn Prediction in Telecommunication Industry using Decision Tree. International Journal of Engineering Research & Technology (IJERT), 6(4), pp. 439-443.
[7]Sjarif, N.N.A., Yusof, M.R.Y., Wong, D.H., Ya'akob, Ibrahim, R, & Osman, M.Z.(2019). A Customer Churn Prediction using Pearson Correlation Functionand K Nearest Neighbor Algorithm for Telecommunication Industry. Int. J. Advance Soft Compu. Appl, 11(2), pp. 46-59.
[8]Yabas, U, Chankya, H.C. (2013). Churn prediction in subscriber management for mobile and wireless communications services. IEEE Publications.
[9]Madan, M, Dave, M, Nijhawan, V.K. (2015). A Review on: Data Mining for Telecom Customer Churn Management. International Journal of Advanced Research in Computer Science and Software Engineering, 5(9), pp. 813-817.
[10]Gupta, M, Gandhi, A.B., Gupta, S.C. (2018). Machine Learning as Intelligent tool for Churn Prediction in Telecommunication Industry. International Journal of Computer Applications, 181(10), pp. 16-22.
[11]Azeem, M, Usman, M. (2018). A fuzzy based churn prediction and retention model for prepaid customers in telecom industry. International Journal of Computational Intelligence Systems, 11(1), pp. 66 – 78.
[12]Mamčenko, J, Gasimov, J. (2014). Customer Churn Prediction in Mobile Operator Using Combined Model. In Proceedings of the 16th International Conference on Enterprise Information Systems, pp. 233-240.