Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms

IJERT Journal

International Journal of Engineering Research & Technology (IJERT)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms

V. Kavitha
Assistant Professor (Ad-hoc),
Department of Computer Science and Engineering;
JNTUA College of Engineering; Pulivendula, Andhra Pradesh, India.

S. V Mohan Kumar
Department of Computer Science and Engineering;
JNTUA College of Engineering; Pulivendula, Andhra Pradesh, India.

G. Hemanth Kumar
Department of Computer Science and Engineering;
JNTUA College of Engineering; Pulivendula, Andhra Pradesh, India.

M. Harish
Department of Computer Science and Engineering;
JNTUA College of Engineering; Pulivendula, Andhra Pradesh, India.

***Abstract*:- In the Telecommunication Industry, customer churn detection is one of the most important research topics that the company has to deal with retaining on-hand customers. Churn means the loss of customers due to exiting offers of the competitors or maybe due to network issues. In these types of situations, the customer may tend to cancel the subscription to a service. Churn rate has a substantial impact on the lifetime value of the customer because it affects the future revenue of the company and also the length of service. Due to a direct effect on the income of the industry, the companies are looking for a model that can predict customer churn. The model developed in this work uses machine learning techniques. By using machine learning algorithms, we can predict the customers who are likely to cancel the subscription.Using this, we can offer them better services and reduce the churn rate. These models help telecom services to make them profitable. In this model, we used a Decision Tree, Random Forest, and XGBoost.**

Keywords: Telecom churn, Xgboost(Extreme Gradient Boosting) Classification algorithms, Decision Trees, Random Forest. .

I. INTRODUCTION

The telecommunications sector has displayed one of the central industries in developed countries. Service companies like these suffer, particularly from the loss of valuable customers due to competitors known as customer churn. The scientific progress and the growing number of operators increased the level of opposition. Companies are pulling hard to survive in this aggressive market, depending on complicated strategies. The customer churn causes a considerable loss of telecom services and becomes a severe problem. Three main approaches have been introduced to generate more profits to get new customers, upsell the current customers, and increase the holding period of customers. However, comparing these strategies using the value of return on investment (RoI) of each into account has shown that the third approach is the most successful strategy, proves that maintaining an existing customer costs much lower than getting a new one, in extension to being held much easier than the upselling tactics. To implement the third strategy, companies have to reduce the potential of customer's churn, known as "the customer movement from one provider to another." Customers' churn is a significant concern in service sectors with great aggressive services. On the other hand, foretelling the customers who are expected to leave the company will serve a potentially big-hearted extra revenue source if it is given in the early phase. Many types of research confirmed that machine learning technology is highly efficient in predicting this situation. This method is applied learning from past data.

A. *Existing System*

Customer churn prediction has been performed using various techniques, including data mining, machine learning, and hybrid technologies.These techniques enable and support companies in identifying, predicting, and retaining churn customers. They also help industries in CRM and decision making. Most of them used decision trees in common as it is one of the recognized methods to find out the customer churn, but it is not appropriate for complex problems [1]. But the study shows that reducing the data improves the accuracy of the decision tree [2].In some cases, data mining algorithms are used for customer prediction and historical analysis. The techniques of regression trees were discussed with other commonly used data mining methods like decision trees, rule-based learning, and neural networks [3].

B. *Proposed System*

In this system, we use various algorithms like Random Forest, XGBoost & Logistic Regression to find accurate values and which helps us to predict the churn of the customer. Here we implement the model by having a dataset that is trained and tested, which makes us have maximum correct values. Fig.1 shows the proposed model for churn prediction and describes its steps. In the Initial step, data preprocessing is performed in which we do filtering data and convert data into a similar form, and then we make feature selection.

In the further step prediction and classification is done using the algorithms like Random Forest, XGBoost, Logistic Regression(LR). Training and testing the model with the data set, we observe the behavior of the customer and analyze them. In the final step, we do analysis based on the results obtained and predict the customer churn.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

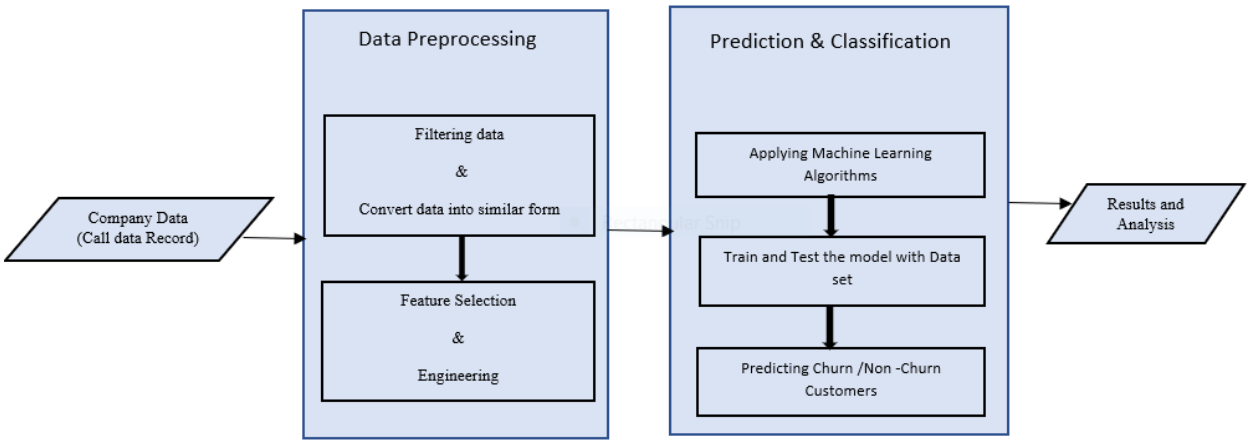

Fig 1. Proposed Model for Customer Churn Prediction

II. METHODOLOGY

A. *Data Set*

As we know, the data set is the starting point for everything; it should have full-fledged data to make the machine learn about the problem. Datasets can be generated or developed from the scrap information available on the internet. Some issues we have to create a dataset that makes sense that tells how to respond based on real-time inputs for the problem datasets can be gathered from the internet every day. A dataset is a collection of data. Most commonly, a data set has contents of a single database table, or a single statistical data matrix, where every column of the table describes a particular variable, and each row matches a given member of the data set in question. The data set lists the values of the variables, such as height, the weight of an object, for each member of the data set. Each value is recognized as a datum. As we know, the data set is the starting point for this process.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	...	Yes	
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	...	No	
7	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	...	No	
8	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	...	Yes	
9	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	...	No	

Z10 rows × 21 columns

B. *Data Preprocessing*

Data set is a collection of feathers and N number of rows. Many values are in different formats. In a dataset, they may be duplicate values or null values that may lead to some loss inaccuracy, and there may be dependent.

Data have been collected from different sources, so there use a different type of format to notate a single value like gender someone represents M/F or Male/Female. The machine can understand only 0 and 1, so an image will be in 3-dimension data should be reduced to a 2-dimension format like data show to free from noisy data, null values, an incorrect size. Data cleaning can be performed by panda's tabular data and OpenCV for images.

1. Data Filtering and Noise Removal

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.

2. Feature selection & Engineering

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.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

C. *Prediction & Classification*

Many techniques have been proposed for customer churn prediction in the telecommunication industry. In these three modeling techniques are used as predictors for the churn prediction. These techniques are outlined as :

1) Random Forest

We use Random Forest to predict whether the customer is going to cancel his subscription. Random Forest uses Decision trees for classifying whether the customer is going to cancel his subscription. The random forest consists of a large number of decision trees. A decision tree points to a specific class. A class with more number of votes will be the classifier for a particular customer. Decision trees are sensitive to the data they are trained in. To avoid this, we use Bagging. Bagging is a kind of process where we take a random sample from the dataset for training decision trees.

2) Logistic Regression

By using logistic regression, we can predict the probability of a churn i.e., the likelihood of a customer to cancel the subscription. Logistic regression is a supervised learning algorithm used for classification. In Logistic regression, we set a threshold; based on the limit, and only the classification is made using logistic regression. The threshold value is variable, and it is dependent on the classification problem itself.

3) XGBoost

XGBoost is the abbreviation for eXtreme Gradient Boosting. The primary purpose of using XGBoost is due to its execution speed, and it's model performance. XGBoost uses ensemble learning methods; i.e., it uses a combination of different algorithms and produces output as a single model. XGBoost supports parallel and distributed computing while offering efficient memory usage.

III. PROPOSED WORK

Initially, we will get the dataset from Kaggle, and by data filtering, we removed all the null values. Then we converted all the data into a similar form, which more natural to understand and analyze. By using Logistic regression and having a different approach, we try to implement a predictor model for the Telecom company. Here we have a customer data set, and by preprocessing and feature selection, we divide the data set for training and testing. For this algorithm, we have made some feature engineering to have more efficient and accurate results using that algorithm.

Logistic regression helps us to have a discriminative probabilistic classification and can estimate the probability of occurring event places. The dependent variable presents the event occurrence (e.g., it will be one if the event takes place, 0 otherwise). By training, the data to that model will get a result having their details, and then we will test the model with the remaining amount of data. Therefore we will get an accuracy based on the findings by which we can predict the customer churn and can a clear warning about the customer, and this can help the company to take some measures which will help not to lose the existing customer from the service.(Here we divided data into 80% - training,20%-testing).

By getting both the results, we will try to fix the y1 train data and y2 test data to the model fit to make the model learn from the historical data. In this, the epochs are used to make the model learn the same data repeated times. By using the CNN model, we visualized the data, and by that, we can know the model accuracy of the resulted data, and then we can have a prediction of the churn.(Fig.6,Fig.7)

Similarly, we used the other two techniques to know which will provide us more accurate results. In the Random Forest, we used the same dataset, and by applying the technique, we trained the model and tested it out to get the results in the confusion matrix, which will show us the obtained output, and we can notice the accuracy (fig.3). The result obtained from the XGBoost model is shown in (fig.4), where we can observe the accuracy obtained by using that technique.

IV. RESULT AND ANALYSIS

We performed several experiments on the proposed churn model using machine learning algorithms on the dataset. In Fig.2, we can observe the results obtained while performing the experiment using the Random Forest algorithm and can check the accuracy. Random Forest(RF) is a useful algorithm that suite for classification and can handle nonlinear data very efficiently.RF produced better results and better accuracy and performance compared to the other techniques. As we should need better accuracy to predict the customer churn, we prefer to use the technique, which results in better accuracy. Similarly, we can observe the results obtained when using the Logistic regression technique(Fig.4) and XGBoost (Fig.5). Finally, we visualized the data using the CNN model. In Fig.6, Fig.7 we can observe the visualized data.

	precision	recall	f1-score	support
0	0.62	0.52	0.56	440
1	0.85	0.89	0.87	1321
accuracy			0.80	1761
macro avg	0.73	0.70	0.72	1761
weighted avg	0.79	0.80	0.79	1761

Fig 3.Confusion matrix of Random Forest.

	precision	recall	f1-score	support
0	0.57	0.56	0.56	440
1	0.85	0.86	0.86	1321
accuracy			0.79	1761
macro avg	0.71	0.71	0.71	1761
weighted avg	0.78	0.79	0.78	1761

Fig 4.Confusion matrix of Logistic regression.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

	precision	recall	f1-score	support
0	0.58	0.50	0.54	440
1	0.84	0.88	0.86	1321
accuracy			0.78	1761
macro avg	0.71	0.69	0.70	1761
weighted avg	0.78	0.78	0.78	1761

Fig 5.Confusion matrix of XGBoost.

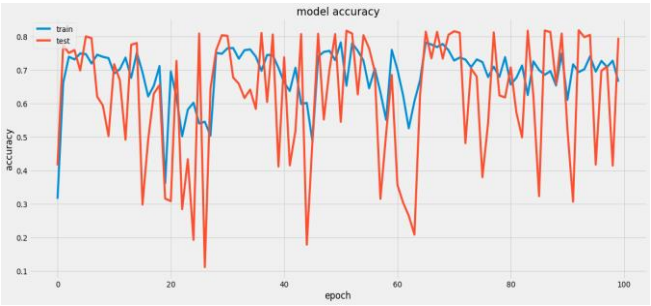

Fig.6.Visualizing CNN model val_acc and accuracy.

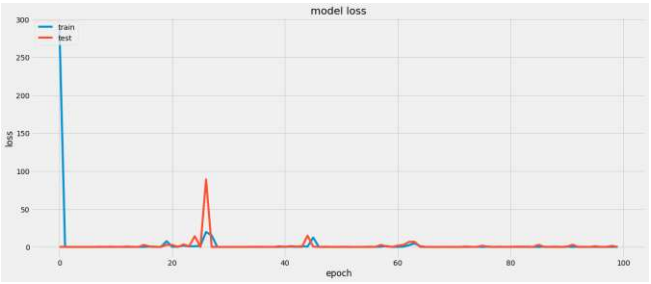

Fig.7.Visualizing CNN model val_loss and loss.

V. CONCLUSION

The importance of churn prediction will help many companies, mainly in telecom industries, to have a profitable income and achieve good revenue. Customer churn prediction is the major issue in the Telecom Industry, and due to this, companies are trying to keep the existing ones from leaving rather than acquiring a new customer. Three tree-based algorithms were chosen because of their applicability and diversity in this type of application. By using Random Forest, XGBoost, and Logistic regression, we will get more accuracy comparing other algorithms. Here we are using the dataset of some customers about their service plan and checking the values of them and have a precise prediction, which will help to identify the customers who are going to migrate to other company services. By this, the 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.Random Forest produced better accuracy among the various methods.

In the coming days, we will further research on lazy learning approaches to have better customer churn prediction.To know the changing behavior of the customers, the study can be extended by using Artificial Intelligence techniques for trend analysis and customer prediction.

REFERENCES

[1] V. Lazarov and M. Capota, ‘‘Churn prediction,’’ Bus. Anal. Course, TUM Comput. Sci, Technische Univ. München, Tech. Rep., 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.7201&rep=rep1&type=pdf.

[2] Vadakattu, B. Panda, S. Narayan, and H. Godhia, ‘‘Enterprise subscription churn prediction,’’ in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 1317–1321.

[3] 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

[4] A Novel Approach for Churn Prediction Using Deep Learning https://www.researchgate.net/publication/328819998_A_Novel_Appoach_for_Churn_Prediction_Using_Deep_Learning

[5] S. A Survey on Customer Churn Prediction using Machine Learning Techniques.https://www.researchgate.net/publication/310757545_A_Survey_on_Customer_Churn_Prediction_using_Machine_Learning_Techniques

[6] Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors https://arxiv.org/pdf/1703.03869.pdf

[7] Automated Feature Selection and Churn Prediction usingDeepLearning Models. https://www.irjet.net/archives/V4/i3/IRJET-V4I3422.pdf

[8] Effectual Predicting Telecom Customer Churn using Deep Neural Network.https://www.ijeat.org/wp-content/uploads/papers/v8i5/D6745048419.pdf

[9] Customer Churn Prediction in Telecommunication with Rotation Forest Methodhttps://www.researchgate.net/publication/282981765_Customer_churn_prediction_in_telecommunication

[10] Customer churn prediction in telecom using machinelearninghttps://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6

AUTHORS PROFILE

V.KAVITHA, ASSISTANT PROFESSOR (AD-HOC), DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING; JNTUA COLLEGE OF ENGINEERING; PULIVENDULA, ANDHRA PRADESH, INDIA.

G.Hemanth kumar, Department of Computer Science and Engineering; JNTUA College of Engineering; Pulivendula, Andhra Pradesh, India.

S.V MOHAN KUMAR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING; JNTUA COLLEGE OF ENGINEERING; PULIVENDULA, ANDHRA PRADESH, INDIA

M.HARISH, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING; JNTUA COLLEGE OF ENGINEERING; PULIVENDULA, ANDHRA PRADESH, INDIA