RESEARCH ARTICLE OPEN ACCESS

Customer Churn Prediction in Telecom SectorUsing Machine Learning Techniques

Auti Divya.S, Deshmukh Rajeshwari.B, Dumbre Komal.G
Computer Enginnering DepartementJCOE, Kuran, SPPU- Pune

Dr.A.A.Khatri
Associate Professor Computer Engineering Department JCOE, Kuran SPPU - Pune

ABSTRACT
Now-a-days churn is a one of the biggest problem in the telecom industry. Companies can understand the customers who leave the services using advanced machine learning (ML) technology. Data is important in each and every sector, which makes churn prediction the biggest challenge for telecom sector. Churn prediction is also called as attrition. There are many reasons why the customers are leaving the company or services such as canceling their subscriptions due to poor services, leaving the company due to network issue, poor connectivity, etc. Customer churn prediction can be found out by using various machine learning algorithms. Previously used algorithms such as Logistic Regression, Decision Tree, etc. These algorithms does not provide high accuracy. There are other supervised learning algorithms which can provide better accuracy and are efficient than previously used algorithms. The main purpose of this research is to stop the customer from getting churn and not only predicting the customer churn but finding the reasons behind customer problems.
Keywords— Customer Attrition, KNN, Random Forest, SVM, Churn, No Churn.

I. INTRODUCTION

With the rapid development of telecommunication industry, to meet the need of surviving in the competitive environment of telecom sector, the retention of existing customers has become a huge challenge. Maintaining existing customers is far easier than making new ones. Churn is a vast field to be work upon its not limited to an organization or in telecommunication industries. [1]

To reduce churn rate try to find out factors that increase customer churn is important to take necessary actions. The main aim of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely to churn.

In customer churn prediction one can analyzing the past behavior of customers and accordingly finding the real reason behind the churn, then predicting whether churn will happen in future by customer. On the basis of customer feedback the churn model will give the proper output related to churn and no churn. [2]

Customers can very easy switch service from one organization to another for better service quality or price rate. Telecom sector are convinced that recruiting new customers is far more expensive and hard than existing customers. [3]

II. REVIEW OF DIFFERENT CHURN PREDICTION SYSTEMS

1) The churn prediction is the process of collecting the customer feedback about the product and tries to retain the customer within their organization. In this system they have used algorithms like decision tree, Logistic regression. The data they got was mostly balanced and categorical data than they begin with data cleaning, pre-processing, removing unwanted columns, feature selection, label encoding. Then they used logistic regression and found features with highest weight assigned leading to cause of churn. The advantage was logistic regression gives higher accuracy than other algorithm. One of the limitation was that, for retention of customer churn recommendation system can be used.

2) In this paper 2 classification model of supervised learning as the data used is labelled data which comes under supervised learning category of Machine Learning. They have used 2 algorithms, which are KNN and Logistic Regression. KNN and Logistic regression has been applied on the clean dataset using python as a platform as both of the algorithms are under the supervised machine learning so a confusion matrix is preferred to be performed in order to find models performance. In that they have used the different parameters. It has been found that KNN is better than logistic regression in terms of accuracy. They does not consider deep learning methods which will increase accuracy.

3) This work aims to predict customer churn in commercial bank as early as possible using efficient in Machine Learning method such as KNN, SVM, Decision Tree and RF Classifier. A better result is achieved by using RFC together with oversampling. The performance of classifiers varies when using different feature selection methods. Results shows that DT and RFC accuracy increases after oversampling. The model was in the form that can use minimal information and give maximum throughput for prediction. The study only use small amount of data.

4) The Churn prediction from the ustomers become the important aspect in the telecommunication in order to retain the customers with organizations and to have an increase in their revenue. This increases the need for modeling a churn prediction system that is not only accurate but also includes the comprehensibility and justifiability. The LDT (Lower Distance Zone) and UDT (Upper Distance Zone) is used. The dataset is iterated for these algorithms each, and takes the best of the iterated samples. They had used 4 different sets of datasets. They consider the best ones efficiency as the resultant value. LDT and UDT has low efficiency so to overcome this drawback they had used RFC and SVM, gives better accuracy. Different techniques are used to achieve an efficiency of 100%.

III. PROPOSE SYSTEM

In this system, we use various algorithms like Random Forest, Support vector machine and k nearest neighbour to find accurate values and which helps us to predict the churn of the customer. We are implementing the model by having a dataset that is trained and tested, which have maximum correct values. Machine Learning trains a model on known data so that it can predict future outcomes. Sentiment Analysis is the process of detecting text to determine if the emotional tone of the message is positive, negative or neutral. It works on textual on the basis of customer feedback.

A. K-Nearest Neighbor Algorithm:

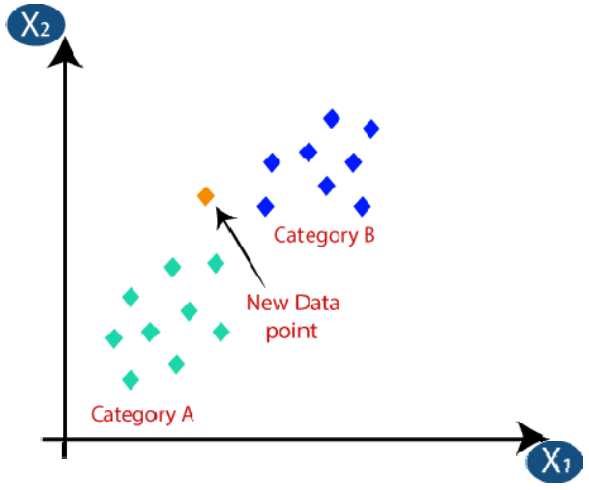


Fig 1: K-Nearest Neighbor

K-Nearest Neighbour is one of the easiest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm find out the similarity between the new data points and available data points and put the new data points into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarities. This means when new data comes then it can be easily classified into a well suite category by using KNN algorithm.

KNN is also called as lazy learner algorithm because it does not learn from the training set instantly instead it stores the dataset and at the time of classification, it performs an action on the dataset. A supervised machine learning algorithm s one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data. In other words, it classifies a new data point based on similarity.

B. Support Vector Machine:

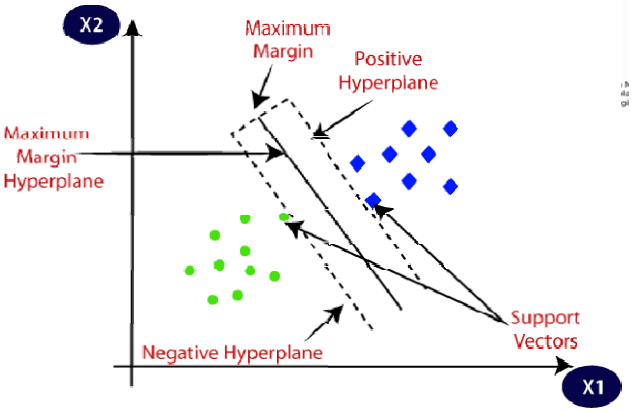


Fig 2: Support Vector Machine

SVM which means support vector machine is one of the most popular supervised learning algorithm in machine learning. The main goal of SVM is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the category in the future this best decision is called a hyperplane.

SVM chooses the major points that help in creating the hyperplane. Margin is nothing but the distance between two bounding hyperplanes. Points which are closer to the hyperplane are known as the support vector points, because these points are consider in the result of algorithm and not others. There are some of the kernels in SVM. Kernels are mathematical function. Kernel is the method of utilizing a linear classifier to solve a non-linear problem.

SVM is different from other algorithms because the decision boundary expands the distance from the nearest data points of all the classes.

C. Random Forest Classifier:

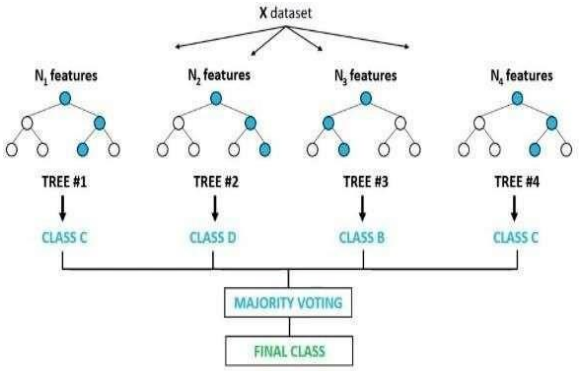


Fig 3: Random forest classifier

A Random Forest Algorithm is a supervised machine learning algorithm and is used for Classification and Regression problems in Machine Learning. Random Forest gather the output of multiple decision trees to reach a final result on basis of majority of number decision trees. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.

Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. IN the random forest algorithm the more the number of trees it give better results. Similarly, and also the higher accuracy It is based on the concept of ensemble learning

There are a lot of benefits to using Random Forest Algorithm, it reduces the risk of over fitting and the required training time. Random Forest algorithm runs efficiently in large databases and produces highly accurate predictions.

IV. SYSTEM ARCHITECTURE

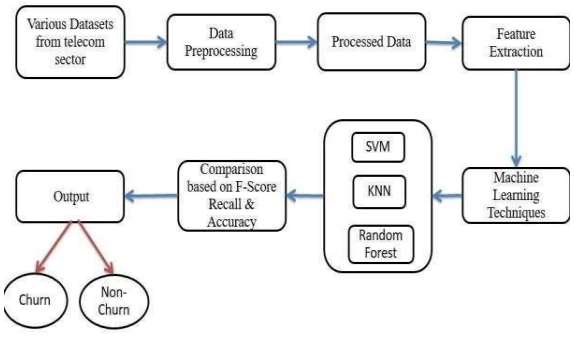


Fig 4: System Architecture

1. Dataset:-

The data set is the important aspect for everything. Dataset is a collection of data. A dataset is the data required in machine learning. Here in this system we are taking dataset of different companies under telecom sector such as airtel, idea, jio, BSNL, etc. For the purpose of predicting churn in telecom sector.

2. Data Preprocessing:-

Data preprocessing is the process in which the data is cleaned and organized. The dataset is preprocessed to check missing values, noisy data, symbolic data or null values before executing of the algorithm.

Data have been collected from different sources, so there is different type of format 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 should be free from noisy data, null values, an incorrect size.

3. Processed data:-

After data preprocessing the next step is the processed data. Once the data is preprocessed which means all the unwanted values are remove which not necessary in execution of algorithms. The data is ready for getting processed.

4. Feature Extraction:-

Feature Extraction is a crucial step. It is the process in which the raw data is converted into numerical data. The dataset consists of many features out of which Choosing the needed features, which helps to improve Performance measurement of the model.While remaining features will have less importance.

5. Implementing different machine learning techniques:-

Implement various algorithms is the next step in the system architecture. The techniques which are implement are support vector, random forest and k nearest neighbor machine learning algorithms. Comparing all the algorithm and then finding out which algorithm is giving better accuracy.

6. Comparison:-

Comparison of various machine learning evaluation metric that measures a model's accuracy. That is f-score, accuracy and recall. With the help of confusion matrix accuracy is obtained.

7. Output:-

Now the prediction of all the algorithms. The algorithm which has highest Accuracy will give the accurate result or prediction regarding to churn or no churn.

V. RESULT

We performed several experiments on the proposed churn Model using machine learning algorithms on the dataset. InFig.6, we can see that results obtained while performing.
The experiment using the Random Forest algorithm has the highest accuracy. Random Forest (RF) is a useful algorithm that suite for classification and can handle nonlinear data very efficiently. Random Forest performance more accurate result and accuracy as compared to the other techniques

	precision	recall	f1-score	support
No churn	0.85	1.00	0.92	847
call churn	0.00	0.00	0.00	13
churn	0.00	0.00	0.00	124
internet churn	0.00	0.00	0.00	9
message churn	0.00	0.00	0.00	6
accuracy			0.85	999
macro avg	0.17	0.20	0.18	999
weighted avg	0.72	0.85	0.78	999

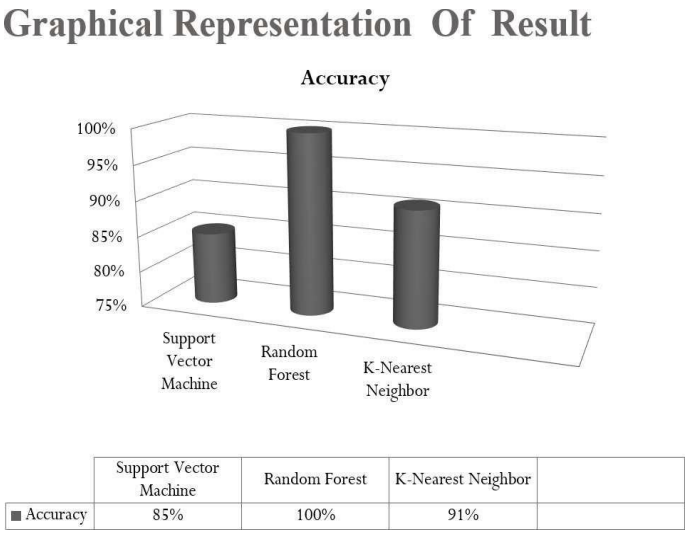
Fig5: Confusion matrix of Support Vector Machine

	precision	recall	f1-score	support
No churn	1.00	1.00	1.00	847
call churn	1.00	1.00	1.00	13
churn	1.00	1.00	1.00	124
internet churn	1.00	1.00	1.00	9
message churn	1.00	1.00	1.00	6
accuracy			1.00	999
macro avg	1.00	1.00	1.00	999
weighted avg	1.00	1.00	1.00	999

Fig 6: Confusion matrix of Random Forest

	precision	recall	f1-score	support
No churn	0.92	1.00	0.96	847
call churn	0.44	0.31	0.36	13
churn	0.90	0.50	0.64	124
internet churn	0.75	0.33	0.46	9
message churn	0.00	0.00	0.00	6
accuracy			0.91	999
macro avg	0.60	0.43	0.48	999
weighted avg	0.90	0.91	0.90	999

Fig7: Confusion matrix of K-Nearest Neighbor



	Support Vector Machine	Random Forest	K-Nearest Neighbor	
Accuracy	85%	100%	91%	

VI. CONCLUSION

Churn prediction is an important issue in telecom sector as reducing churn can help in gain profit. With the help of machine learning techniques one can achieve this and stop customer from getting churn or leaving the service. The Machine Learning method can give better accuracy than existing traditional ML models. The churn prediction is done on the basis of customer's review which classifies the customer into churn and non-churn. To solve the problem related to churn prediction there is a need to identify customer churn chances as fast as possible. And try to find out why customers are dissatisfied with the service reason behind and then try to fulfill the requirement of the customers.

The importance of this type of model in the telecom market is to help companies make more profit. Hence, we have built a model that can predict the churn of customers in Telecom Company.

VII. FUTURE SCOPE

This type of research in the telecom sector is to help companies retain customers and hence retaining the overall profile and further helping to generate more profit. In future there will be opportunities to refine the prediction by adding different features like geographical area, costing, bandwidth (5G-6G), etc. With evolving algorithms, once can find out newer and better text based algorithms which will help in improving accuracy of the prediction system. One can also consider other huge dataset for training the system, which will predict accurate output.

VIII. REFERENCES

[1] Kulkarni, A., Patil, A., Patil, M., & Bhoite, S. (2019). Customer Churn Analysis and Prediction. International Journal of Computer Applications Technology and Research, 8. Retrieved from https://ijcat.com/archieve/volume8/issue9/ijcatr080 91005.pdf

[2] Bhatanagar, M., & Srivastava, D. (2019). A Robust Model for Churn Prediction using Supervised Machine Learning. doi:https://doi.org/10.1109/IACC48062.2019.8971494

[3] Rahaman, M., & Kumar, V. (2020). Machine Learning Based Customer Churn Prediction In Banking. doi:https://doi.org/10.1109/ICECA49313.2020.929 7529

[4] Geetha, V., Punitha, A., Nandhini, A., Nandhani, T., Shakila, S., & Sushmitha, R. (2020). Customer Churn Prediction In Telecommunication Industry UsingRandom Forest Classifier. doi:https://doi.org/10.1109/ICSCAN49426.2020.92 62288

[5] Erik Nesset, Ola Bergem, (2021). Building chain loyalty in grocery retailing by means of loyalty programs – A study of 'the Norwegiancase

[6] Nusrat Parvin, Sayaka Zaman, Samia Amin (2021) Journal of Retailing and Consumer Services

[7] Mishra, A., & Reddy, U. (2017). A comparative study of customer churn prediction in telecom industry using ensamblelearning classifier. doi:10.1109/ICICI.2017.8365230

[8] Mariya Hendriksen , Pim Nauts (2020), Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers, IEEE Conference

[9] Auti, D., Deshmukh, R., Dumbre, K., Khatri, A. (2022). A Review On Customer Churn Prediction Using Different Algorithms. doi:20.18001.GSJ.2022.V9I11.22.40326