International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 3, May-June 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Random Forest Classifier Algorithm for Churn Using Customer Review

Ms.Kavita Babasaheb Khatal.(ME comp) , Prof. M. A.Wakchaure (*Ph.D.*)

Computer Department, Amrutvahini College of Engineering, Sangmner. Dist. Ahmednagar

khatalkavita3@gmail.com

--************************----------------------------------

Abstract:

Customer attrition is a significant issue and one of the most pressing challenges for big businesses. Companies are attempting to create methods to forecast possible customer churn because of the direct impact on revenues, particularly in the telecom business. Finding the variables that cause customer turnover is critical in order to take the required steps to decrease churn. Our key contribution is the development of a churn predictions model that helps telecom providers estimate which customers will be able to churn. The model created in this paper employs computational methods on a large data platform to provide a novel approach to feature engineering and selection. This study also defined churn characteristics that are critical in discovering the fundamental causes of churn in order to gauge the model's performance. CRM may boost productivity, offer suitable promotions to a group of potential churn subscribers based on previous behaviour patterns, and vastly improve the company's marketing efforts by identifying the main churn drivers from customer data. The accuracy, precision, recall, f-measure, and receiving performance parameters (ROC) area of the suggested churn forecasting model are all examined. It also gives aspects that contribute to customer attrition via the rules provided by the essential element classification model.

Keywords **—Receiving Operating Characteristics, Deep learning, Convolution Neural Network, churn prediction, Feature Selection**

--************************----------------------------------

I. INTRODUCTION

When connecting to any of the several Telecom service alternatives available today, consumers go through a lengthy decision-making process. Telecom companies' services aren't very distinctive, and number portability is popular. Customer loyalty is a Problem. As a result, it's becoming more critical for telecoms firms to detect circumstances that cause consumers to unsubscribe and take preventative efforts to keep them. Begin by calculating the number many users who churned during that month cheval cheval. The number of continues to pump per user day is then defined as the total number of user days for that month. Then multiply the figure by the number of days in the month to obtain the monthly churn rate. According to studies undertaken over the last several years, data mining approaches are more successful in anticipating customer attrition. Developing an effective churn prediction model is a time-consuming process that includes everything from identifying relevant predictor variables (features) from a vast amount of accessible customer data to selecting an appropriate predictive data mining approach for the feature set. In addition to the network data that they create, telecom industries gather a vast number of client-related data such as consumer profiling, calling patterns, and democratic data. It is possible to categorise a customer's attitude of going away or not going away based on their history of calling behaviour and behaviour. According to studies conducted over the last decade, data mining approaches are more successful in forecasting turnover. Churn prediction strategies that use predictive modelling are also thought to be more accurate. Churn prediction technologies and sentiment analysis use classification and clustering methods to categorise churn consumers and the reasons for their departure. Because we collect vast amounts of data on a regular basis in the telecom business, mining such data using particular data mining methods is a time-consuming operation, and interpreting predictions using traditional approaches is difficult. Various academics have detailed attempts to reduce churn from huge data sets using both static and dynamic techniques, however such systems still have significant challenges with genuine churn detection. Such telecommunication data may sometimes include churn, making it critical to discover search issues. Customer relationship management must be excellent in order to successfully identify churn from vast data (CRM).

Using Natural Language Processing (NLP) and machine learning approaches, we suggested churn detection and predictions from large-scale telecommunications data sets in this study. The first system is concerned with the strategic NLP process, which includes data from before the, data normalisation, feature extraction, and feature engineering. TF-IDF, Stanford NLP, and occurrence correlation approaches have all been offered as feature extraction strategies. The whole curriculum was trained and tested using statistical and machine learning techniques.

II. LITERATURE SURVEY

Telecommunications firms are not always the most popular among customers. In the telecom sector, customer loyalty is critical to profitability. Whether it's convoluted billing, spam marketing emails, poor customer support, internet speed, connection, or costly plans, people often express dissatisfaction with service providers' performance. As a consequence, it should come as no surprise that telecommunications businesses have a high percentage of client attrition. Customer turnover (attrition) is especially difficult in this market since telecom carriers operate enormous fixed infrastructures that must be compensated by income. Customer acquisition is frequently prioritised by businesses, with customer retention coming in second. However, attracting a new client might cost five times as much as keeping an old one. According to study conducted by Bain & Company, increasing client retention rates by 5% may improve earnings by 25% to 95%. Customer attrition, often known as churn, is a measure that reflects customers who discontinue doing business with a firm or a certain service. Most firms might use this data to attempt to identify the causes of high churn rates and develop reactive action plans to address those causes. But what if you knew ahead of time that a certain client was on the verge of leaving your company, and you could take proactive steps to prevent it? Customers might cancel for a variety of reasons, including poor service quality, customer service delays, pricing changes, new rivals joining the market, and so on. Typically, there is no one cause, but rather a chain of events that led to client dissatisfaction. If your organisation is unable to recognise these signs and take action before the cancel button is pressed, there is no going back; your consumer has already left. However, you still have something important in the form of data. Your consumer provided plenty of hints as to where you fell short. It may be a useful tool for gaining important information and training customer churn models. Machine learning is all about learning from the past and having crucial knowledge on hand to better future experiences. There is a lot of space for growth in the telecoms sector. Carriers' vast amounts of client data may help them move from a reactive to a proactive posture. The development of powerful artificial intelligence and data analytics tools has further aided in the efficient use of this rich data to combat churn. Churn prediction employs a variety of techniques, including machine learning and data mining. The decision-tree algorithm is a dependable churn prediction approach [1]. For churn prediction, we also apply a neural network technique [7], data certainty [8], and particle swarm optimization. According to system [2,] a current collection of software is being developed to improve the standard for recognising potential churners. The roles are classed as deal, request pattern, and call pattern adjustments overview functions and are retrieved from request information and client accounts. The properties are assessed using two probabilistic data mining techniques, Nave Bayes and Bayesian Network, and the results are compared to those produced using the C4.5 decision tree, a commonly used approach for classification and prediction. These have led to the potential that customers may easily switch to rivals, among other things. Improve churn prediction from big amounts of data using extraction in the near future is one strategy that may be utilised to achieve this. According to [3,] formalisation of the collecting process's time-window, as well as a literature study. Second, this study examines the growth in churn model accuracy by extending the length of customer events from one to seventeen years using logistic regression, classification trees, and bagging combined with classification trees. As a consequence, researchers will be able to significantly minimise data-related demands such as data collecting, preparation, and analysis. The cost of a subscription is determined by the duration and promotional nature of the subscription. The newspaper industry is sending them a letter informing them that their service will be discontinued. Then ask whether they want to renew their membership and provide instructions on how to do so. Customers cannot cancel their subscriptions, although they do enjoy a four-week grace period once their membership has expired.

According to [4,] the most effective consumer interaction tactics may be employed to effectively increase customer satisfaction. In one of Malaysia's largest telecoms businesses, the researchers used a Multilayer Perceptron neural network approach to assess customer churn. The findings were compared to the most used churn prediction methods, including Multiple Regression Analysis and Analysing Logistic Regression. With the learning algorithm Levenberg Marquardt, the largest neural network design has 14 input nodes, 1 hidden node, and 1 output node (LM). When compared to the most prevalent churn prediction approaches, such as Multiple Regression Analysis and Logistic Regression Analysis, a Multilayer Perceptron neural network approach was used to forecast customer churn at one of Malaysia's largest telecoms businesses. Using a Partial Least Square (PLS) technique, system [5] focuses on highly linked intervals in data sets to create an efficient and descriptive statistical churn model. According to early findings, the suggested approach produces more trustworthy results than traditional prediction models and detects essential characteristics to better explain churning patterns. In addition, network administration, overage administration, and problem management procedures are presented and analysed in the context of several basic marketing campaigns. Burez and Van den Poel [6] compare the performance of random sampling, Advanced Under-Sampling, Gradient Boosting Method, and Weighted Random Forest in churn prediction models. Metrics were used to assess the notion (AUC, Lift). The research concludes that the sampling process is superior than the other strategies considered. Gavril et al. [7] provide a unique data mining strategy for explaining customer churn detection over a wide dataset type. On the basis of incoming and outgoing input calls and messages, around 3500 customer information are examined. For training categorization and research, certain machine learning algorithms were utilised. For the total dataset, the system's estimated average accuracy is about 90%. He et al. [8] constructed a prediction model based on the Neural Network approach to solve the problem of customer churn for a large Chinese telecoms firm with roughly 5.23 million members. The average degree of accuracy was 91.1 percent, indicating a high level of predictability.

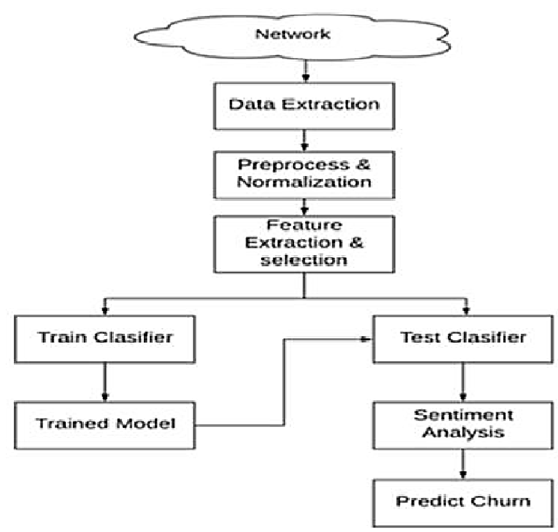
To mimic AdaBoost-churning telecommunications difficulties, Idris [9] proposed a genetic engineering technique. The series was confirmed by two Standard Data Sets. One from Orange Telecom and the other from cell2cell, both with an accuracy of 89 percent, and 63 percent for the other. On the big data platform, Huang et al. [10] investigated customer turnover. The researchers wanted to demonstrate that depending on the amount, diversity, and velocity of data, big data dramatically increases the cycle of churn prediction. Data from

China's largest telecoms company's Project Support and Business Support Department was intended to be stored in a large data repository for fracture engineering. AUC evaluated using the forest algorithm at random. According to [11], clustered input characteristics are used using kmeans and fuzzy c-means clustering algorithms to divide subscribers into distinct groups. These classes are used to build the Adaptive Neuro Fuzzy Inference System (, which is a prediction model for active churn control. Neuro fuzzy parallel categorization is the initial phase in the prediction process. The results of the Neuro fuzzy classifier are then used by FIS to determine churner actions. Success measurements may be used to identify inefficient issues. Customer service network services, operations, and efficiency are all linked to churn management indicators. The adaptability of GSM numbers is an important factor in churner selection. A new collection of applications has been added to System [12] to enhance the detection of probable churners. The characteristics are obtained from call records and client profiles and are grouped as contract, call pattern, and call pattern changes description features. The characteristics are examined using two probabilistic data mining techniques, Nave Bayes and Bayesian Network, and the results are compared to those produced using a C4.5 decision tree, which is frequently used in many classification and prediction applications. These have led to the possibility that consumers may readily migrate to rivals, among other things. Improve churn prediction from big amounts of data using extraction in the near future is one strategy that may be utilised to achieve this.

III. PROPOSED METHODOLOGY

The proposed study activity will use text analysis, natural language processing, and machine learning to detect churn. During prediction, detect the customer's shifting behaviour pattern. To determine which factors have the greatest impact on churn forecast accuracy. To review and compute the churn rate month by month and day by day, which is helpful for improving the system's service quality. To improve system effectiveness, the proposed research activity will design and build a solution for churn prediction utilising NLP and machine learning based. Then, during prediction, we recognise the customer's shifting behaviour pattern. We also examine the factors that most affect churn prediction accuracy, and we ultimately review and compute churn rate every month and day, which is beneficial for improving system customer experience. We suggested churn prediction using big scale data in this study. The system starts with a telecoms synthetic data set that includes certain imbalance Meta data. To do data preparation, data normalisation, feature extraction, and feature selection, as needed. During this execution, optimization procedures were applied to remove duplicate features that might cause a high error rate. The suggested system's training and validation execution. After all steps are completed, the system describes the categorization accuracy for the full data set.

A. Architecture



The goal of this kind of study in the telecommunications sector is to assist firms in increasing their profits. Forecasting turnover has become one of the most significant sources of revenue for telecom firms. As a result, the goal of this study was to develop a system for the Telecom Company that could forecast client attrition. AUC values will be high for such prediction models. To analyse and build the model, the sample data was split into 70 percent for training and 30 percent for testing. For analysing and improving hyper parameters, we utilised 10-fold cross-validation. Engineering tools, effective function translation, and a selection strategy were applied. Making the user interface machine learning algorithms friendly. Another issue was discovered: the data was unbalanced. Customers' turnover accounts for just approximately 5% of the entries. Undersampling or tree methods that are not impacted by this issue have been used to fix a problem. Our various classifiers can be more effective in identifying churn in vast data and offering accurate predictions. This study contributes to the development of a supervised method for extracting dimensional categories, choosing appropriate attributes, and minimising duplication by assessing correlation between them. The findings reveal that the correlation method produces a relatively higher f-score in the weighted frequency of the phrase. In this case, employing weighted word frequency to choose characteristics is very crucial. The relationship between characteristics in a category of aspect is measured to eliminate overlap.

1. Data collection: First, information for various Telecom Sector customers is retrieved using particular characteristics.
2. Pre-processing: Next, we'll do lexical analysis, stop word removal, stemming (Porters method), index term selection, and data cleaning to ensure that our dataset is complete.
3. Lexical analysis: Lexical analysis divides the input alphabet into two categories: 1) word characters (a-z) and 2) word separators (e.g space, newline, tab).
4. Stop word removal: Stop word removal is the process of removing terms from papers that appear repeatedly.
5. Stemming: Stemming is the process of replacing all of a word's variations with a single stem term. Plurals, gerund forms (ing forms), third - person omniscient suffixes, past tense suffixes, and other variations exist.
6. Data Training: We aggregate fake and real-time news data from the internet and train any machine learning classifier.
7. Machine learning testing: We use any computational modelling classifier or weight calculator for real-time or synthetic input data to forecast online news.
8. Evaluation: We showed that the proposed system's effectiveness and compare it to other current systems.

B. Algorithm

1) *Bagging Classifier:*

Input: inp 1…..n all input parameters, Desired Threshold Th.
Output: Executed for output as lable.
Step 1 : Read all records from database (R into DB)
Step 2: Parts [] ← Split(R)
Step 3:

$$\mathrm{CVal} = \sum_{k=0}^{n} \mathbf{Parts[k]}$$

Step 4: check (Cval with Respective threshold)
Step 5: T← get current state with timestamp
Step 6 : if(T.time > Defined Time)
Read all measure of for TP and FN
Else continue. Tot++
Step 7: calculate score = (TP *100 / Tot)
Step 8 : if (score >= Th)
Generate event
end for

2) *Decision tree classifier*:

Input : Selected feature of all test instances D i….n , Training database policies {T 1 ………….T n }
Output: No. of probable classified trees with weight and label.
Step 1: Read (D into D[i])
V←Extract features (D)
Step 2: N←Count_Features(D)
Step 3: for each(c into TrainDB)
Step 4: Nc[i] ==Ext_Features(c)
Step 5: select relevant features of w= {Nc[i], N}
Step 6: Statement (w>t)
Step 7: Return Tree Insatnce { Nc[i], N, w, label}

3) Knearest Neighbour classifier

Input: Training Rules Tr[], Test Instances Ts[], Threshold T.
Output: Weight w{0-1}
Step 1 : Read each test instance from (TsInstnace from Ts)
Step 2 : Read each test instance from (TsInstnace from Ts)
Step 3 : TsIns $= \sum_{k=0}^{n}\{Ak \ldots An\}$
Step 4 : Read each train instance from (TrInstnace from Tr)
Step 5 : $TrIns = \sum_{j=0}^{n}\{Aj \ldots \ldots Am\}$
Step 6 : $w = WeightCalc(TsIns, TrIns)$
Step 7 : if (w >= T)
Step 8 : Forward feed layer to input layer for output OutLayer[] ← {Tsf,w}
Step 9 : optimized feed layer weight, Cweigt ← OutLayer [0]
Step 10 : Return Cweight

Input: Train_DatasetF TrF[], Test_DatesetF TsF[], Threshold T.
Output: Classified label
Step 1: Read R {All attributes} from current parameters.
Step 2: Map with train features with each sample.
Step 3: Calculate distance of train DB with same evidences

$$distance = \sum_{k=0}^{n}(TrF, TsF)$$

Step 4 : evaluate distance> threshold
Step 5: Return the predicted label

IV. RESULT AND DISCUSSIONS

The system categorization graph is shown below. The graphs show how the system categorises the aggregate inputs into separate cases. The suggested system uses an updated RF combination that produces good results in all areas. 5000 instances were supplied for training while 1500 reviews were given for assessment with various classification model and performance assessment. The projected outcomes are compared to two distinct current systems in this system.

Table 1: Comparative analysis of various classification algorithms

No	Method	Acc	Pre	Rec	F-1
1	URF	0.95	0.95	1	0.97
2	DT	0.89	0.92	0.96	094
3	Bagging Classifier	0.94	0.96	0.94	0.97
4	Knearestneighbors	0.81	0.86	0.92	0.89

The comparative study of multiple classification methods for the recommended time series forecasting module is shown in Table 1. KNeighbors has the lowest accuracy, but URF classification has the best accuracy with a 95 percent cross validation rate.

Figure 2 below shows a similar set of data.

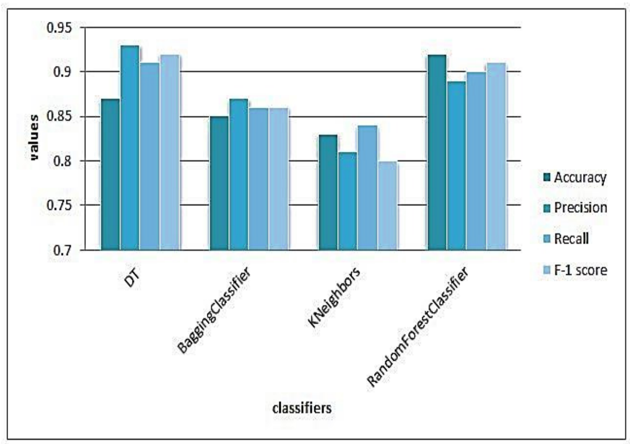


Fig. 2. Time required (in milliseconds) for complete transaction with different records

V. CONCLUSIONS

This study aimed at identifying and detecting churn customers from large telecoms data sets, and the state-of-the-art analyses churn prediction systems developed by various studies. Some systems still have issues with linguistic data conversion, which might result in a high rate of errors during execution. Many academics have proposed combining Natural Language Processing (NLP) approaches with different machine learning algorithms in the hopes of achieving excellent data structuring results. If a training algorithm interacts with that approach, the full data set must be tested or confirmed using even sampling strategies that eliminate data imbalance issues and give a trustworthy predicted flow of data. If we deal with the proposed systems with HDFS framework and parallel machine learning algorithm, which will provide better results in low computation cost, in the future direction to implement a developed methodology with various algorithms to achieve better accuracy, as well as the input signal contains large size and volume.

REFERENCES

[1] Karahoca, Adem, and DilekKarahoca. "GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system." Expert Systems with Applications 38.3 (2011): 1814-1822.

[2] Kirui, Clement, et al. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining." International Journal of Computer Science Issues (IJCSI) 10.2 Part 1 (2013): 165.

[3] Ballings, Michel, and Dirk Van den Poel. "Customer event history for churn prediction: How long is long enough?." Expert Systems with Applications 39.18 (2012): 13517-13522.

[4] Ismail, Mohammad Ridwan, et al. "A multi-layer perceptron approach for customer churn prediction." International Journal of Multimedia and Ubiquitous Engineering 10.7 (2015): 213-222.

[5] Lee, Hyeseon, et al. "Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model." Decision Support Systems 52.1 (2011): 207-216.

[6] Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009; 36(3):4626–36.

[7] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.

[8] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.

[9] Idris A, Khan A, Lee YS. Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). 2012. p. 1328–32.

[10] Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international Conference on management of data. 2015. p .607–18

[11] Karahoca, Adem, and DilekKarahoca. "GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system." Expert Systems with Applications 38.3 (2011): 1814-1822.

[12] V.Geetha, A. Punitha, A.Nandhini, T. Nandhini, S.Shakila, R.Sushmitha. "Customer Churn Prediction in Telecommunication Industry Using Random Forest Classifier." International Journal of Computer Applications 64.5 (2020): 39-42.