Machine Learning Based Churn Prediction

Akash Lal[1], Atharv Dongaonkar[2], Ruturaj Bankar[3], Digvijay Chavhan[4]

[1] *Student, Computer Engineering, MESCOE, Maharashtra, India*
[2] *Student, Computer Engineering, MESCOE, Maharashtra, India*
[3] *Student, Computer Engineering, MESCOE, Maharashtra, India*
[4] *Student, Computer Engineering, MESCOE, Maharashtra, India*

ABSTRACT

Customer turnover is a significant issue and one of the most pressing challenges for big businesses. Companies are working to create methods to forecast prospective customer churn since it directly impacts their revenues, particularly in the telecom industry. As a result, identifying factors contributing to customer turnover is critical to taking the required steps to decrease churn. Our work's essential contribution is developing a churn prediction model that helps telecom carriers estimate which customers are most likely to churn. The model created in this paper employs machine learning methods on a large data platform to provide a novel approach to feature engineering and selection. This research also established churn characteristics that are critical in discovering the fundamental causes of churn to gauge the model's performance. CRM may enhance productivity, offer suitable promotions to a group of potential churn customers based on similar 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 operating characteristics (ROC) area of the suggested churn prediction model are examined. Furthermore, using the rules created by the attribute-selected classifier algorithm gives causes behind the churning of churn clients.

Keyword - *Receiving Operating Characteristics, Deep learning, Convolution Neural Network, churn prediction, Feature selection.*

1. INTRODUCTION

Before subscribing to any of the various Telecom service alternatives available today, consumers undergo 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. Start with the number of users who churned that month to assess your likely monthly churn. Then divide the number of user days in that month by the number of churns per user day to obtain the number of churns per user day. Then multiply the figure by the number of days in the month to get 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 more accurate. Churn prediction systems and sentiment analysis utilise classification and clustering algorithms to characterise churn customers or the reasons for their leave. Because we collect vast amounts of data regularly in the telecom business, mining such data using particular data mining methods is time-consuming, and interpreting predictions using traditional approaches is difficult. Various academics have detailed attempts to minimise churn from massive data sets using both static and dynamic techniques, but such systems still face significant difficulties in identifying churn. Occasionally, such telecommunication data may include churn, making it critical to discover search issues. Customer relationship management must be excellent to successfully locate churn from vast data (CRM).

Using Natural Language Processing (NLP) and machine learning approaches, we suggested churn detection and prediction from a large scale telecommunication data set in this study. The first system is concerned with the strategic NLP process, including data pre-processing, data normalisation, feature extraction, and feature selection. TF-IDF, Stanford NLP, and occurrence correlation approaches have all been offered feature extraction strategies. The whole curriculum was trained and tested using machine learning classification techniques.

2. LITERATURE SURVEY

Telecommunications firms aren't always the most popular among customers, and In the telecom sector, customer loyalty is critical to profitability. People often express dissatisfaction with service providers ' performance, whether it's convoluted billing, spam marketing emails, poor customer support, internet speed, connection, or costly plans. Consequently, it should come as no surprise that telecommunications businesses have a high percentage of client attrition. Customer turnover (attrition) is backbreaking in the telecom business since it manages massive fixed infrastructures that must be compensated by income. Companies frequently prioritize client acquisition, with customer retention being a distant second. However, attracting a new client might cost five times more than keeping an old one. According to a 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 particular service. Most firms might use this data 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 particular client was on the verge of leaving your company, and you could take proactive steps to prevent it?

Customers might cancel for various 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 resulting in consumer dissatisfaction. There is no going back if your organization cannot recognize these signs and take action before the cancel button is pressed; your client 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 valuable tool for gaining important information and training customer churn models. It's all about machine learning when it comes to learning from the past and having crucial knowledge on hand to better future experiences. When it comes to the telecommunications industry, there is a lot of space for growth. The quantity and volume of client data that carriers acquire may help carriers move from reactive to proactive. The development of powerful artificial intelligence and data analytics tools has further aided in using this rich data to combat churn.

Churn prediction employs various techniques, including machine learning and data mining. The decision-tree algorithm is a reliable churn prediction tool [1]. In addition, for churn prediction, a neural network technique [7], data certainty [8], and particle swarm optimization are applied. According to the system, a current collection of software improves the quality of identifying likely churners [2. The roles are classed as a 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. The results are compared to those produced using the C4.5 decision tree, a commonly used approach for classification and prediction. For various reasons, this has resulted in the probability that customers would soon switch to rivals. Improving churn prediction from significant amounts of data using extraction is one strategy they may utilize to achieve this. According to [3,] the formalization of the collecting process's time-window and 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.

Consequently, researchers will be able to significantly minimize 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 they will discontinue their service. Then ask whether they want to renew their membership and provide instructions on how. Customers cannot cancel their subscriptions, although they have a four-week grace period once their membership has expired.

According to [4] They may implement the most effective consumer interaction tactics to increase customer satisfaction levels. In Malaysia's largest telecoms businesses, the researchers used a Multilayer Perceptron (MLP) neural network approach to assess customer churn. They compared the findings to the most used churn prediction methods, including Multiple Regression Analysis and Analyzing Logistic Regression. With the Levenberg Marquardt learning method, the maximum neural network design has 14 input nodes, one hidden node, and one output node (LM). Compared to the most general churn prediction approaches, such as Multiple Regression Analysis and Logistic Regression Analysis, a Multilayer Perceptron (MLP) neural network methodology was used to forecast customer churn at one of Malaysia's largest telecoms businesses.

In system [5] We developed an efficient and descriptive statistical churn model using a Partial Least Square (PLS) technique focusing on highly linked intervals in data sets. According to early findings, the suggested approach produces more reliable results than traditional prediction models and detects essential characteristics to explain churning patterns better. In addition, network administration, overage administration, and problem management procedures are presented and analysed in the context of several essential marketing campaigns.

Burez and Van den Poel [6] Unbalance data sets studies in churn prediction models and compares the performance of random sampling, advanced under-sampling, the Gradient Boosting Method, and Weighted Random Forest. They used metrics to assess the notion (AUC, Lift). The research concludes that the sampling process is superior to the other strategies considered.

Gavril et al. [7] Describes an innovative data mining method to explain the broad dataset type of consumer churn detection. About 3500 consumer details is analyzed based on incoming number as well as outgoing input call and texts. Specific machine learning algorithms were used for training classification and research, respectively. The system's estimated average accuracy is about 90 percent for the entire dataset.

He et al. [8] a prominent Chinese telecoms business created a prediction model based on the Neural Network approach to address the problem of customer churn in a market with around 5.23 million members. The average degree of accuracy was 91.1 %, indicating a high level of predictability.

Idris [9] suggested a genetic engineering solution to modeling AdaBoost-churning telecommunications problems. Two Standard Data Sets verified the series. With a precision of 89%, one from Orange Telecom and the other from cell2cell and 63% for the other one.

Huang et al. [10] The client turnover was investigated using a big data platform. The researchers wanted to demonstrate that big data dramatically increases the cycle of churn prediction depending on the amount, diversity, and velocity of data. 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. The forest algorithm was employed at random by AUC and analysed.

According to [11] Clustered input features are clustered input characteristics that position subscribers in discrete groups using k-means and fuzzy c-means clustering algorithms. These classes are used to build the Adaptive Neuro-Fuzzy Inference System (ANFIS), a prediction model for active churn control. Neuro-fuzzy parallel categorization is the initial phase in the prediction process. FIS then uses the outputs of the Neuro-fuzzy classifier to determine the actions of the churners. To identify inefficiency issues, they might employ success metrics. Customer service network services, operations, and efficiency are linked to churn management measures. The adaptability of GSM numbers is an essential consideration for churner selection.

In System [12] A new collection of applications aimed at improving the detection of prospective churners The characteristics are obtained from call information and client profiles and grouped as a contract, call pattern and the call pattern change description features. The parts are examined using two probabilistic data mining techniques, Nave Bayes and Bayesian Network. The results are compared to those produced using a C4.5 decision tree, which is frequently employed in many classification and prediction applications. These have led to the possibility that consumers may readily migrate to rivals, among other things. Improving churn prediction from significant amounts of data using extraction is one strategy they may utilise to achieve this.
According to [13] Formalization of the time window selection procedure and a literature review. Second, this research examines the rise in churn model consistency by utilizing logistic regression, classification trees, and bagging in conjunction with classification trees to extend the history of customer events from one to seventeen years. As a result, researchers may greatly minimize data-related difficulties in data storage, planning, and research. The amount that customers must pay is determined by the length of the subscription and the promotional value. The newspaper firm sends a letter to inform them that their subscription will expire. Then ask whether they want to renew their membership and provide instructions on how. Customers cannot cancel their subscriptions, although they have a four-week grace period once their membership has expired.

According to [14] the most effective customer retention techniques should be used to effectively reduce customer turnover rates. The research suggests a neural network approach for Multilayer Perceptron (MLP) to predict customer churn in one of Malaysia's leading telecommunications firms. The findings were compared with the most common techniques of churn prediction such as Multiple Regression Analysis and Logistic Regression Analysis. The optimal configuration of the neural network contains 14 input nodes, 1 hidden node and 1 output node with Levenberg Marquardt (LM) learning algorithm. Multilayer Perceptron (MLP) neural network approach to predict client churn in one of the leading telecommunications companies in Malaysia compared to the most common churn prediction techniques, such as Multiple Regression Analysis and Logistic Regression Analysis. In system [15] on Building a predictive churn model that is accurate and concise using a Partial Least Square (PLS) methodology based on highly correlated data sets between variables. A preliminary experiment shows that the model presented provides more accurate performance than traditional models of prediction and identifies key variables to better understand churning behaviors. Additionally, there is a range of basic churn marketing strategies— system management, overage management, and complaint management strategies is presented and discussed.

Burez and Van den Poel [16] studied the problem of unbalance datasets in churn prediction models and compared performance of Random Sampling, Advanced Under-Sampling, Gradient Boosting Model, and Weighted Random Forests. They used (AUC, Lift) metrics to evaluate the model. the result showed that under sampling technique outperformed the other tested techniques.

Gavril et al. [17] presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the number of incoming and outgoing messages and voicemail for each customer. The author applied principal component analysis algorithm ―PCA to reduce data dimensions. Three machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. The author used AUC to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively. The dataset used in this study is small and no missing values existed. He et al. [18] proposed a model for prediction based on the Neural Network algorithm in order to solve the problem of customer churn in a large Chinese telecom company which contains about 5.23 million customers. The prediction accuracy standard was the overall accuracy rate, and reached 91.1%. Idris [19] proposed an approach based on genetic programming with AdaBoost to model the churn problem in telecommunications. The model was tested on two standard data sets. One by Orange Telecom and the other by cell2cell, with 89% accuracy for the cell2cell dataset and 63% for the other one. Huang et al. [20] studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China‘s largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.

3.PROPOSED SYSTEM DESIGN

The goal of the proposed study is to use text analysis and a machine learning classifier to detect churn. During prediction, notice the customer's shifting behaviour pattern. To determine which factors have the most significant impact on churn forecast accuracy. To assess and compute the churn rate month by month and day by day, which helps improve the system's service quality. The proposed research activity will design and build a solution for churn prediction utilizing NLP and machine learning techniques to enhance system accuracy. Then, during projection, we recognize the customer's shifting behaviour pattern. We also assess the factors that most impact the accuracy of churn prediction, and we ultimately review and compute churn rates month by month and day by day, which is beneficial for improving the system's service quality. This paper offers a method for predicting churn from big data sets. The procedure starts with a telecoms synthetic data set that includes imbalanced metadata. To make data preparation, data normalization, feature extraction, and feature selection, as needed. During this execution, they applied several optimization tactics to remove duplicate features that might cause a high error rate. They suggested the system's training and testing execution. After all, steps are completed, the plan describes the categorization accuracy for the entire data set.

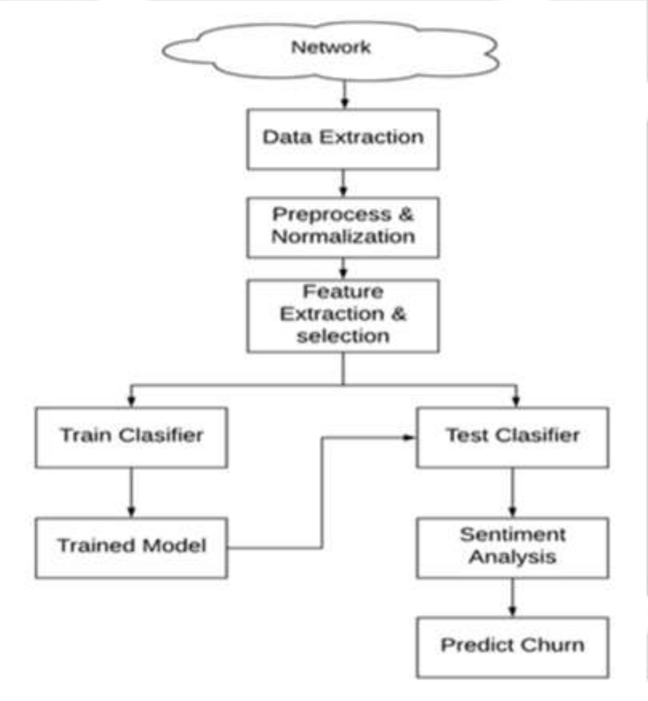

Figure 1: Proposed system design

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, they split the sample data into 70 % for training and 30% for testing. For analysing and improving hyperparameters, we utilised 10-fold cross-validation. We employed engineering tools and a selection technique, and practical function translation. They are 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. Under-sampling 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 developing a supervised method for extracting dimensional categories, choosing appropriate attributes, and minimising duplication by assessing their correlation. The findings reveal that the correlation procedure produces a relatively higher f-score in the weighted frequency of the phrase. In this case, employing weighted word frequency to choose characteristics is crucial. By quantifying the association, the overlap between attributes in a category of aspect is prevented.

1. Data Acquisition: First of all the information for different Telecom Sector Customer based on certain parameters is extracted data.

2. Preprocessing: Then we will apply various preprocessing steps such as lexical analysis, stop word removal, stemming (Porters algorithm), index term selection and data cleaning in order to make our dataset proper.

3. Lexical analysis: Lexical analysis separates the input alphabet into,

4. Word characters (e.g. the letters a-z) and 2) Word separators (e.g space, newline, tab).

5. Stop word removal: Stop word removal refers to the removal of words that occur most frequently in documents.

6. Stemming: Stemming replaces all the variants of a word with a single stem word. Variants include plurals, gerund forms (ing forms), third person suffixes, past tense suffixes, etc.).

7. Data Training: We compile artificial as well as real time using online news data and provide training with any machine learning classifier.

8. Testing with machine learning: We predict online news using any machine learning classifier, weight calculator for real time or synthetic input data accordingly.

9. Analysis: We demonstrate the accuracy of proposed system and evaluate with other existing systems

.

4. CONCLUSIONS

This study focuses on identifying and detecting churn customers from large telecoms data sets and the state-of-the-art analyses of churn prediction systems developed by several studies. Some systems still have issues with linguistic data conversion, which may 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. Suppose a machine-learning algorithm interacts with that approach.

In that case, the complete data set must be tested or confirmed using even sampling strategies that eliminate data imbalance issues and give a trustworthy predicted flow of data. Suppose we deal with the proposed systems with HDFS framework and parallel machine learning algorithm, which will provide better results in low computation cost. In that case, we will implement a proposed plan with various machine learning algorithms. There achieve better accuracy, the input data contains large size and volume if we deal with the proposed systems with HDFS framework and parallel machine learning algorithm, which will provide better results in low computation cost.

5. REFERENCES

[1] Karahoca, Adem, and Dilek Karahoca. "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 Dilek Karahoca. "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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction. Expert Systems with Applications. 2009 Apr 1;36(3):4626-36.

[17] Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data. 2019 Dec;6(1):1-24.

[18] Li L, Jiang P, Xu H, Lin G, Guo D, Wu H. Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China. Environmental Science and Pollution Research. 2019 Jul;26(19):19879-96.

[19] Idris A, Khan A, Lee YS. Genetic programming and adaboosting based churn prediction for telecom. In2012 IEEE international conference on Systems, Man, and Cybernetics (SMC) 2012 Oct 14 (pp. 1328-1332). IEEE.

[20] 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