IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907237 |

Analysis of Customer Churn for Quality-of-Service Parameters using Machine Learning

Mrs. Archana Paike, Mrs. Sayali Ambavane

Department of Computer Engineering, Government Polytechnic, Pune, Maharashtra, India

Department of Computer Engineering, Government Polytechnic, Pune, Maharashtra, India

ABSTRACT: Churn prediction system that classifies churn consumers and the causes for their churning of telecom customers using classification and clustering algorithms. 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 minimise churn from huge 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 in order to successfully identify churn from vast data (CRM). Using Natural Language Processing (NLP) and machine learning approaches, we presented churn identification and prediction from large-scale telecommunication data sets in this study. The first system is concerned with the strategic NLP process, which includes data preparation, data normalisation, feature extraction, and feature selection. TF-IDF, Stanford NLP, and occurrence correlation approaches have all been offered as feature extraction strategies. The whole curriculum was trained and tested using machine learning classification techniques. Finally, the experiment analysis explains how to assess the proposed system's performance and compare it to certain current systems.

KEYWORDS: Natural language processing, churn prediction, machine learning, telecom industry, customer relationship management

I. INTRODUCTION

Writing comments tends to churn more regularly in today's computer environment, but voice mail plan users tend to churn less frequently. Customers who have made four or more customer service calls churn at the same rate as customers who have made four or more calls. Using several machine learning methodologies, we determine the average churn rate during model training and assess it for testing. Predicting accurate turnover is crucial for maximising an organization's revenues, as we highlighted in our research. The rest of the paper is laid out as follows: Section 2 provides a summary of recent research, section 3 outlines suggested study, system overview, and datasets description, and section 4 provides observations. Contribution to research in Section 5 Section 7 finishes the study with section 8 future works, which deals with the implementation of churn prediction systems.

1.1 Background

Clustering algorithms are clustered input functions that use k-means and fuzzy c-means to place subscribers in independent, separate groups, according to [1.] The Adaptive Neuro Fuzzy Inference Framework (ANFIS) is used to build a prediction model for effective churn control using these groupings. The parallel categorization of Neuro soft is the initial step toward prediction. The outputs of Neuro fuzzy classifiers are then used as feedback by FIS to determine the churners' behaviour. In order to discover inefficiencies, progress measures might be employed. The facilities, methods, and performance of the customer support network are all markers of churn reduction. The versatility of GSM numbers is an important factor in churner selection.

A current set of software in System [2] to improve the quality of identifying likely 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. For a variety of reasons, this has resulted in the probability that customers would soon switch to rivals. 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,] the 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

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907237 |

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 have 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 (MLP) neural network approach to assess customer churn. The findings were compared 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, 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 (MLP) neural network methodology was used to forecast customer churn at one of Malaysia's largest telecoms businesses.

Using a Partial Least Square (PLS) technique, system [5] focused 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 management, overage administration, and problem handling procedures are presented and analysed in the context of a few basic marketing campaigns.

[6] Burez and Van den Poel 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. Metrics were used to assess the notion (AUC, Lift). The research concludes that the sampling process is superior than the other strategies considered.

[7] Gavril et al. The article describes a novel data mining approach for explaining the many types of customer churn detection datasets. On the basis of incoming numbers as well as outbound input calls and texts, around 3500 customer information are examined. For training categorization and study, certain machine learning algorithms were applied. 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 business with roughly 5.23 million members. The average degree of precision 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. The forest algorithm was employed at random by AUC and analysed.

Makhtar et al. [11] advocated using rough set theory as a statistical definition of churn in telecom. The Rough Set classification method has beaten the other algorithms, as described in this article.

The issue of imbalanced data sets, where churned customer groups are below active customer levels, has been addressed in several studies, making churn estimate a major worry. Amin et al.[12] evaluated six different oversampling strategies to the problem of telecom churn forecasting. Other algorithms (MTDF and rules development based on genetic algorithms) outperformed the others, according to the data. amazing algorithms for screening.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907237 |

II. LITERATURE SURVEY

We have surveyed several recent trends in this field and tabulated the techniques, datasets used and research gap in Table 1

Table1.Brief overview of survey

No	Technique	Dataset	Extracted Features	Research Gap
1	x-Means clustering algorithms and Neuro Fuzzy algorithm [1]	GSM operation data, 24,900 customers 22 attributes Turkey dataset	Some value-added services and some values added services	System reflects good accuracy on structured dataset only.
2	Naïve Bayes, Decision Tree[2]	European operator 106,405 customers 112 attributes	Contract, usage pattern patterns, and calls pattern	High error rate to detect actual churn due to redundant features.
3	Neural network, Regression [3]	Unknown 129,892 customers 113 attributes	Demographic, Value added, usage pattern	Heterogeneous dataset tedious to handle in similar patterns environments.
4	Neural network, Regression [4]	Unknown, 169 customers 10 attributes	Demographic, Billing data, usage pattern, customer relationship	High space complexity generate in each layer
5	Stepwise variable selection partial least squares [5]	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral information, Customer care and demographics	Redundant features should be generating high error rate.
6	Artificial Neural Network [6]	ML Dataset of UCI 2,427 user's information with 20 attributes	Demographics, Usage pattern, Value added services	It works only define statically parameters.
7	Binomial logistic regression model [7]	Iranian telco operator 3150 customers 15 attributes	Demographic, call usage pattern, customer care service	Language influence should be generate irrelevant features vector.
8	Generalized additive models (GAM) [8]	Belgian 134, 120 customers 27 attributes	Demographic Usage patter, bill and payment	High error rate during unknown text prediction.
9	Logistic regression Decision tree [9]	Polish mobile operator 122098 customers 1381 attributes	Demographic, call data records, customer care services	Its works only synthetic data only and high data reduction rate.
10	Decision tree as well as machine learning [10] algorithms has used.	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral data, of customer care and feature information	Behaviors information generate the churn possibility sometime it generate false ratio.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907237 |

III. PROPOSED WORK

In this paper, we offer a method for predicting churn from big data sets. The system starts with a telecoms synthetic data set that includes some imbalance meta data. To do data preparation, data normalisation, feature extraction, and feature selection, as needed. During this execution, several optimization tactics were applied to remove duplicate features that might cause a high error rate during the execution. The proposed system execution for training and testing is shown in Figure 1. After all steps are completed, the system describes the categorization accuracy for the full data set.

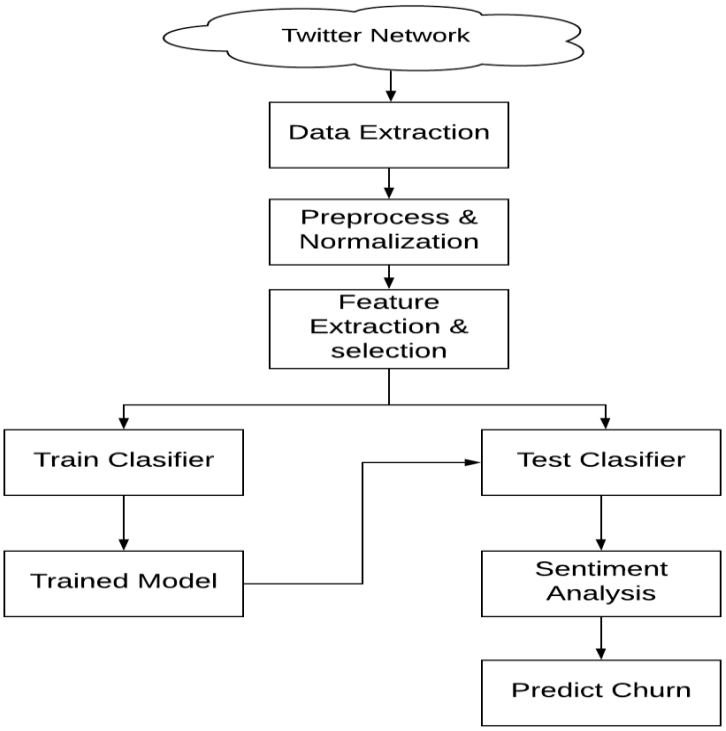

Fig.2 Proposed system overview

1.2 System overview

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. We employed engineering tools, as well as a selection technique and effective function translation. 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 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 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 very crucial. By quantifying the relationship between characteristics in a category of aspect, the overlap between features in that category is avoided..

1.3 Datasets used

We utilised a telecom industry dataset from Kaggle.com to forecast churn consumers since it includes data from both churn and non-churn clients. It has around 21 characteristics and 7043 rows with the churn class labelled as yes or no. The class label is the last property with a number value such as 1 or 2.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907237 |

IV. OBSERVATIONS

- The rule generation provides better classification accuracy than other classification techniques which is define in [12].
- System [7] provide accuracy for churners as well as non-churners model around 99.10% and for BN algorithm it should be around, 99.55% as well as MLP, and 99.0 0% for SVM respectively.
- Hybrid method has used for churn prediction which generates around 90% classification accuracy in [1]
- Neural Network has used for classification as well as accuracy prediction in [4] which provides around 91.28% accuracy on large dataset

The below analysis is the system classification graph. The graphs display how system classify the overall inputs into classification various instances. The proposed system is implemented with RNN combination, which gives all results with satisfactory level. For performance evaluation 5000 instances given for training and 1500 reviews given for testing with different cross validation. Here system compares the proposed results with two different existing systems.

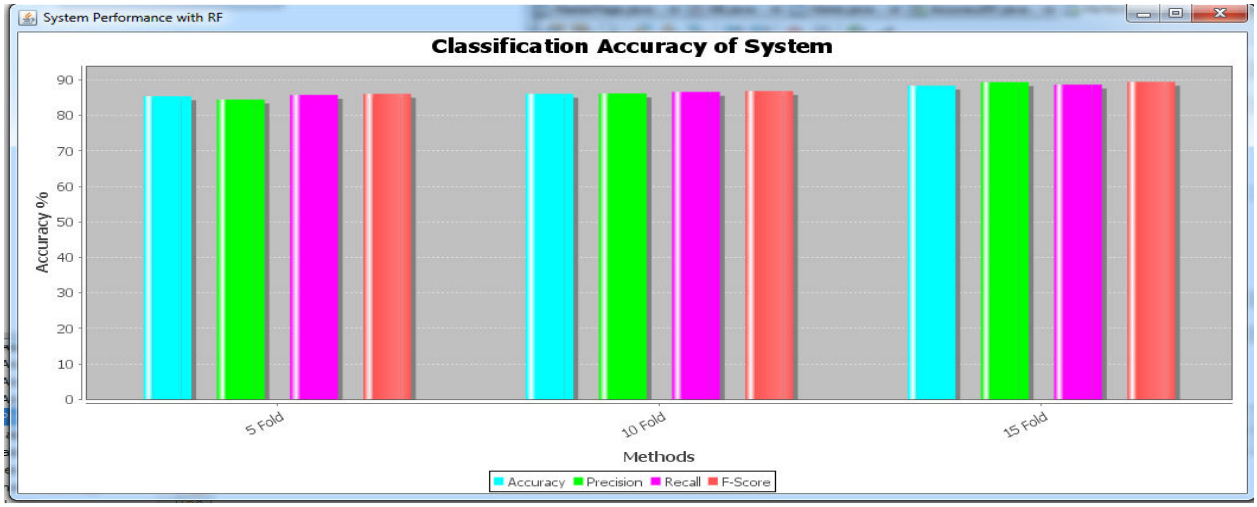

Figure 3 : Classification accuracy of RF

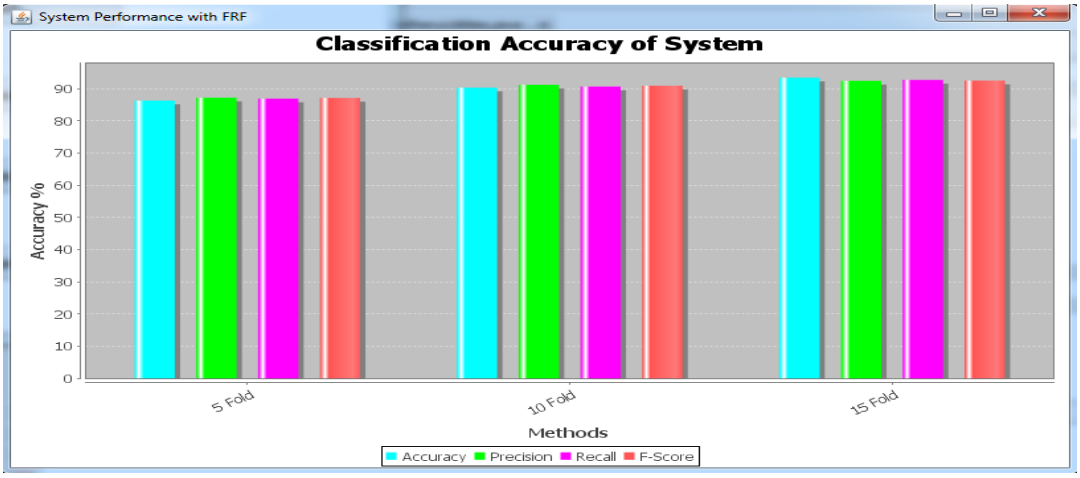

Figure 4 : Classification accuracy of FRF

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907237 |

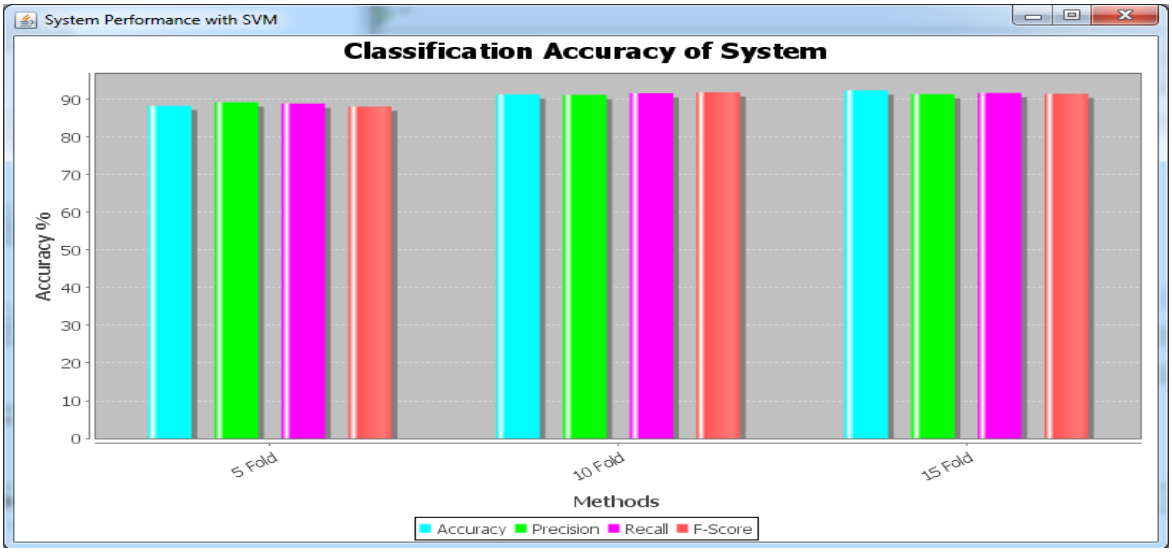

Figure 5 : Classification accuracy of SVM

V. CONCLUSION

The goal of this study is to identify and detect churn consumers from huge telecommunication data sets. The state of the art discusses churn prediction systems produced by diverse studies. Many systems are still plagued by linguistic data conversion difficulties, which may result in a high rate of errors during execution. Many academics have advocated combining Natural Language Processing (NLP) approaches with various machine learning algorithms to achieve high accuracy when input is organised. When dealing with such a system, any machine learning algorithm must analyse or verify the complete data set using an even sampling strategy, which eliminates data imbalance and ensures a constant data flow for prediction.To implement a proposed system with large heterogeneous dataset in Hadoop distribution File System (HDFS) will be future interesting task of this system,

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 Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.

12 Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details