Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

Churn Predictions using NLP and ML

Prof. Dr. Monika Rokade[2], Abhijit Shingarwade[1], Prof. Dr. Sunil Khatal[3]

[1]Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, India

monikarokade4@gmail.com

[2]Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, India

abhijit.shingarwade@gmail.com

[3]HOD, Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, India

khatalsunils88@gmail.com

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract:

The churn prediction system identifies customers at risk of attrition and analyzes the factors contributing to telecom customer turnover by leveraging classification and clustering algorithms. The telecom industry typically gathers vast amounts of data, which can make the application of certain data mining techniques cumbersome and interpreting predictions with standard methods challenging. Numerous studies have focused on minimizing churn in large datasets. However, these systems continue to face significant challenges in effectively detecting churn. In some cases, telecommunications data might already contain churn signals, underscoring the need for precise search and detection methods. Efficient customer relationship management is critical for accurately identifying turnover within extensive datasets. In this study, we demonstrated churn detection and prediction by utilizing comprehensive telecom datasets with natural language processing (NLP) using machine learning methods. The main system emphasizes a planned NLP approach, incorporating feature extraction, feature selection, data normalization, and data preparation. We implemented feature extraction approaches using NLP methodologies. We trained and assessed the system as a whole using hybrid machine learning methods for classification. The experimental analysis highlights the methodology for evaluating the performance of the proposed system, comparing it with existing approaches.

Keywords: CRM, Machine learning, and natural language processing (NLP).

I. INTRODUCTION

Recent research has shown that data mining approaches are superior at forecasting client turnover. Developing an effective churn prediction model is crucial and requires significant effort. From the large amount of customer data that is available, you need to pick out the right predictor variables. Telecommunications companies gather extensive consumer-related data, including customer profiles, calling habits, and demographic information, alongside the network data they produce. Created on the client's historical calling behavior, it is possible to identify their propensity to disengage or remain engaged. Research directed over the previous decade indicates that data mining approaches are

increasingly successful in forecasting turnover. The predictive modeling approaches used in churn prediction are regarded as more precise. In the telecommunications sector, the daily generation of substantial data makes the use of particular data mining techniques laborious, while the interpretation of predictions derived from conventional methods becomes challenging. Effectively identifying turnover from extensive data enhances customer relationship management (CRM). The first system covers the strategic NLP process, which includes feature extraction, feature selection, data normalization, and data pre-processing. We use machine learning classification algorithms to teach and assess the whole curriculum.

II. LITERATURE SURVEY

In [1] Clustered input features refer to characteristics that group subscribers into distinct clusters using algorithms like k-means and fuzzy c-means clustering. The first step in the prediction process involves neuro-fuzzy parallel classification. Following this, the FIS utilizes the outputs from the neuro-fuzzy classifier to assess churner behaviors. This helps identify potential inefficiencies, which may be measured using success metrics. Additionally, the flexibility of GSM numbers plays a critical role in selecting churners.

In System [2] double probabilistic data mining methods, Naive Bayes and Bayesian Networks, are employed to analyze these features. The outcomes are then related to persons produced by a C4.5 decision tree, a widely used tool in various classifications in addition prediction tasks. This approach suggests that consumers might easily switch to competitors, among other factors. One potential strategy to improve churn prediction is to utilize data extraction techniques to analyze large datasets more effectively.

According to [3] a formalization of the time window selection process, followed by a comprehensive literature review. The approach helps researchers significantly reduce data-related challenges in areas like storage, planning, and analysis. A newspaper company sends out notifications to inform customers of their subscription's expiration, asking if they wish to renew and providing guidance on how to do so.

According to [4] successfully diminish client attrition; it is essential to use the most effective retention techniques. For Malaysia's leading telecom firms, the study suggests predicting customer attrition with the Multilayer Perceptron (MLP) approach. They contrasted the outcomes with those of reputable churn prediction methods, such as logistic regression analysis in addition to multiple regression analysis. The ideal neural network topology, as determined by the Levenberg-Marquardt (LM) learning algorithm, has one output node, one hidden node, and fourteen input nodes.

According to [5] they employ a Partial Least Squares (PLS) approach utilizing highly correlated dataset intervals to create a precise and efficient predictive churn model. Initial experiments show that this model surpasses current prediction models, shedding light on key factors that help in understanding churn behavior more effectively.

In [6] the tests show how well different approaches, such as weighted random forests, gradient boosting models, random sampling, and advanced undersampling, work in churn prediction models that use datasets that aren't balanced. We evaluated the model's performance using measures such as AUC and

lift. The findings indicated that the undersampling technique produced superior performance relative to the other strategies.

According to [7] they propose an advanced data mining methodology for identifying client attrition within an extensive dataset. The investigation was performed on data from more than 3,500 users, concentrating on the number of incoming and departing calls and texts. They developed and evaluated the classification system using several machine learning methodologies. The system's anticipated accuracy throughout the whole dataset is about 90%.

In [8] a predictive model was created using a neural network algorithm to tackle the customer attrition issue at a major China telecom firm with over 5.23 million clients. The model attained an overall accuracy of 91.1%, indicating a significant degree of prediction proficiency.

According to [9] proposed using genetic programming to model the churn challenges encountered by telecommunications firms in relation to AdaBoost. Two standard datasets, Orange Telecom and Cell2Cell, confirmed the sequence with accuracies of 89% and 63%, respectively.

In [10] they conducted the analysis of client attrition using a big data display place. The assistants aimed to validate that extensive data may significantly enhance churn forecast accuracy, contingent upon the volume of data available and the speed of its processing. In this investigation, we assessed the random forest approach using AUC.

III. PROPOSED SYSTEM DESIGN

This paper presents an approach for predicting churn using large datasets. The process begins with a synthetic telecom dataset containing imbalanced metadata. It involves steps such as data preparation, normalization, feature extraction, and selection. We implemented various optimization strategies during the process to eliminate redundant features that could potentially lead to increased error rates. Figure 1 illustrates the training and testing phases of the proposed system. The system reports the classification accuracy for the entire dataset once all steps are complete.

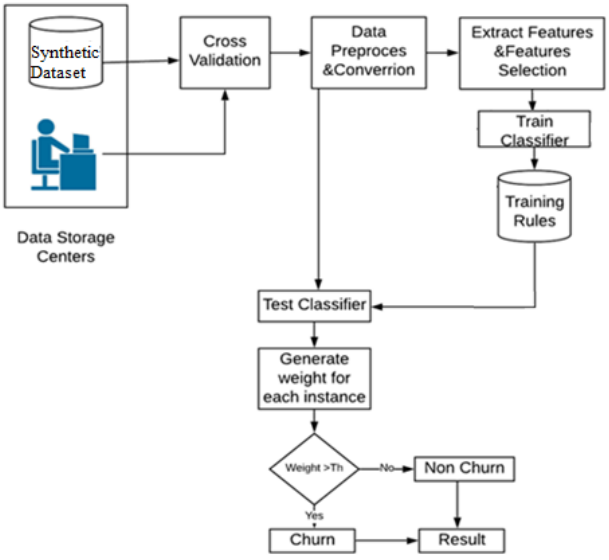

Figure 1: Proposed System Design

Implement Module

Data Pre-Processing:

Data Pre-processing – Before working on data it's necessary to pre-process it. Data may contain missing values that leads to poor result to avoid this, pre-processing is necessary. Filtering and noise removal: In filtering process unwanted and unnecessary features are removed and only keep useful features. In noise removal, null values, space, missing characters are removed.

Feature Extraction:

Feature extractions remove the unwanted data from the data set and keep only accurate and complete data that should be processed.

Data Training: We gather both synthetic and real-time churn data and utilize it to train machine learning classifiers effectively.

Testing with machine learning: We use machine learning classifiers and a weight calculation mechanism that are designed to work well with both real-time and fake input data to predict churn data.

Analysis: We showcase the accuracy of the proposed system in addition compare its performance against other existing systems.

IV. RESULT AND ANALYSIS

We carried out the implementation in an open-source Java environment. Intel(R) Core(TM) i3-2328M CPU @ 2.20GHz 2.20 GHz the system's Java 3-tier analytics platform. We used the Telecom Sector dataset to classify emails as spam or non-spam. We conducted an experimental analysis using an ensemble machine learning approach to validate the results.

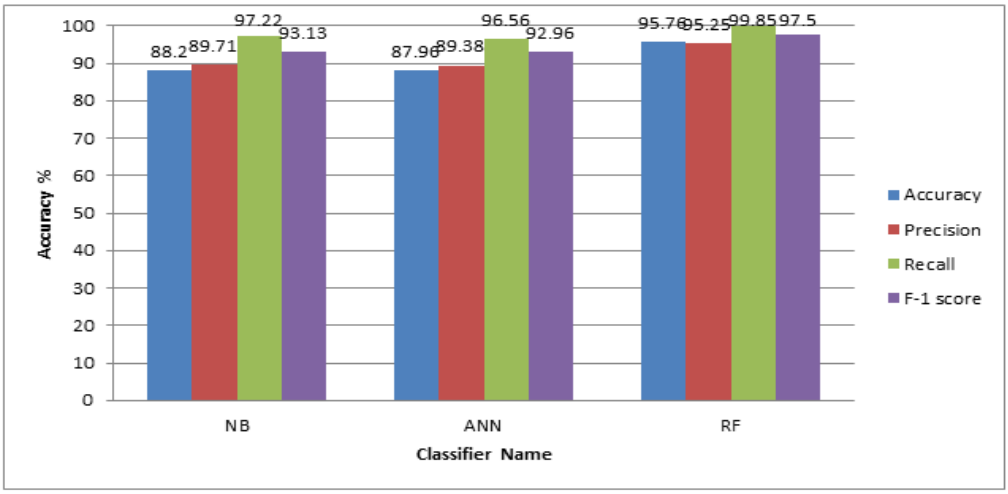

Figure 2: Accuracy of system analysis

Figure 2 depicts the categorization accuracy of the proposed system in comparison to many leading methodologies. The paper demonstrates the dataset's ability to identify churn and non-churn situations using several deep learning and machine learning classifiers. The suggested model detects whether or not churn data is malicious with a high accuracy rate of up to 95.40%.

V. CONCLUSION

This research mainly aims to discover and recognize churned customers among extensive telecom databases. The discourse emphasizes cutting-edge churn prediction technologies created by many academics. Nevertheless, several systems continue to have difficulties in digesting linguistic input, resulting in elevated error rates during execution. To resolve these challenges, several academics have suggested the amalgamation of natural language processing (NLP) methodologies with machine learning methods. This combination may improve data organization and overall efficiency. When using machine learning algorithms with these methodologies, it is crucial to ensure the accuracy of the whole dataset using even sampling techniques. This will mitigate issues related to data imbalance and ensure the reliability of the predictive data flow.

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.