2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC) | 979-8-3503-2614-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/R10-HTC57504.2023.10461822

Predicting Customer Churn In Telecom Industry: A Machine Learning Approach For Improving Customer Retention

Abhikumar Patel
Department of Information and Communication Technology, SOT
Pandit Deendayal Energy University
Gandhinagar, Gujarat, India
abhi.pict19@sot.pdpu.ac.in

Amit G Kumar
Department of Electronics and Communication Engineering, SOT
Pandit Deendayal Energy University
Gandhinagar, Gujarat, India
amit.Gkumar@sot.pdpu.ac.in

***Abstract*— “In this research paper, we aim to analyze the factors that contribute to customer churn in the telecom industry and develop effective retention strategies to reduce churn rates and improve customer satisfaction. Ourfindings can help telecom companies develop effective retention strategies to reduce customer churn and improve customer satisfaction while maintaining affordable pricing for their services. The machine learning model we developedcan be used as a tool for telecom companies to predict and prevent customer churn, leading to improved revenue and profitability. In addition to investigating the factors leading to customer churn, we propose a novel machine learning model that leverages advanced predictive analytics to identify at-risk customers and personalize retention interventions, further enhancing its applicability for telecom companies seeking to mitigate churn and foster long-term customer loyalty. Our research provides actionable insights and a comprehensive approach towards sustainable growth in the highly competitive telecom industry.”**

Keywords— Supervised Machine Learning, Customer churn prediction, Telecom industry, Customer Retention, XGBoost Classifier

I. INTRODUCTION

The use of machine learning algorithms has shown promising results in predicting customer churn in various industries, including the telecom industry. In this research paper, we focus on using supervised machine learning algorithms to predict customer churn in the telecom industry. Our objective is to develop a predictive model that can identify customers at risk of churn and provide insights for the telecom industry to develop effective retention strategies. Through this research, we aim to contribute to the growing body of knowledge on customer churn prediction in the telecom industry. Our goal is to provide valuable insights to help telecom companies develop data-driven retention strategies to improve customer satisfaction, reduce churn rates, and ultimately enhance their competitiveness in the market.

II. RELATED WORK

Several studies have been conducted in the field of customer churn analysis, with varying techniques and data sources. In their study, A. K. Malik used a decision tree algorithm to identify the key factors influencing customer churn in the telecommunications industry [3]. Similarly, D. Vetrithangam used a random forest algorithm to predict customer churn in the same industry [6]. In another study, Jajam, N. utilized a support vector machine algorithm to predict customer churn in the mobile telecommunications industry, achieving an accuracy rate of over 80% [5].

Additionally, a study conducted by P. Bhuse applied clustering techniques to customer churn data in the telecommunications industry to identify customer segments at risk of churn [7]. In contrast, K. D. Singh utilized logistic regression to analyze the effect of customer satisfaction on customer churn in the telecommunications industry [8]. In addition to the studies mentioned above, several other studies have explored different machine-learning algorithms and techniques for predicting customer churn in the telecommunications industry. For example, A. Mishra and U. S. Reddy (2017) used a hybrid model that combined decision trees and logistic regression to predict customer churn in a Chinese telecommunications company. The model achieved high accuracy and outperformed other algorithms, including support vector machines and random forests. Similarly, Jajam, N. (2023) applied a gradient-boosting algorithm to a dataset from a US- based telecommunications company and found that the most significant predictors of churn were customer tenure, contract type, and monthly charges.

Despite these promising results, the accuracy and effectiveness of churn prediction models can be affected by

979-8-3503-2614-7/23/$31.00 ©2023 IEEE

several factors, including data quality, feature selection, and model overfitting.

III. Methodology

The goal is to gain insights into the characteristics of the customers who have churned and identify any patterns or factors that contribute to churn. Once these insights have been obtained, the next step is to develop a model that can accurately predict churn based on the available data.

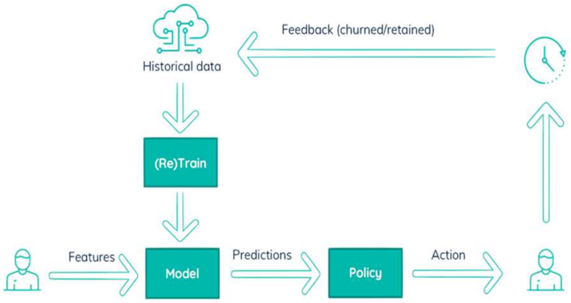

Fig. 3.1. Complete layout of the proposed research work

A. Dataset Exploration:

The Telecom Churn Dataset[2] contains information about customers of a telecom company, including attributes such as state, account length, area code, whether the customer has an international plan or voicemail plan, and various measures of the customer's usage, such as the number of voicemail messages, total minutes and charges for calls made during the day, evening, and night, and the total number of customer service calls made. The dataset consists of 2666 rows and 20 columns. This could give a better understanding of any skewness or outliers in the data.

B. Pre-processing Data:

The method used for imputing missing values (mean or median) could be explained and justified. Additionally, any outliers that were detected during the exploration step could be handled, either by removing them or by applying a transformation to the data.

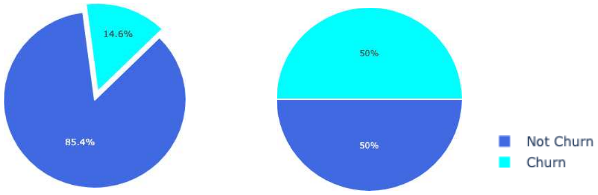

Fig. 3.2.[A] Churn Count after SMOTE. [B] Churn Count Before SMOTE.

SMOTE, which stands for Synthetic Minority Over-sampling Technique, is a popular method used to address the class imbalance problem in machine learning. This technique involves creating synthetic samples of the minority class by interpolating between existing samples in that class. SMOTE has been shown to improve the performance of models in predicting the minority class, especially when combined with other techniques such as under-sampling the majority class. Overall, SMOTE is a useful tool in addressing class imbalance and can lead to more accurate and reliable predictions.

In addition to checking churn rates for different area codes, the demographics of customers in each area code could be analyzed to identify any patterns that may be contributing to churn. This could help in understanding the factors that contribute to customer churn.

C. Data Visualization:

The correlation matrix provides valuable insights into the relationship between various features and the churn rate of customers. From the matrix, Customers with the International Plan churn more frequently, as we have noticed. This information can be used to investigate the reasons behind the high churn rate in this area and implement measures to address them. 74%, which is more than 50% of the average churn rate. In the context of customer churn in telecom, the correlation matrix plot helps to identify the relationships between various features such as call duration, number of calls made, account length, and others.

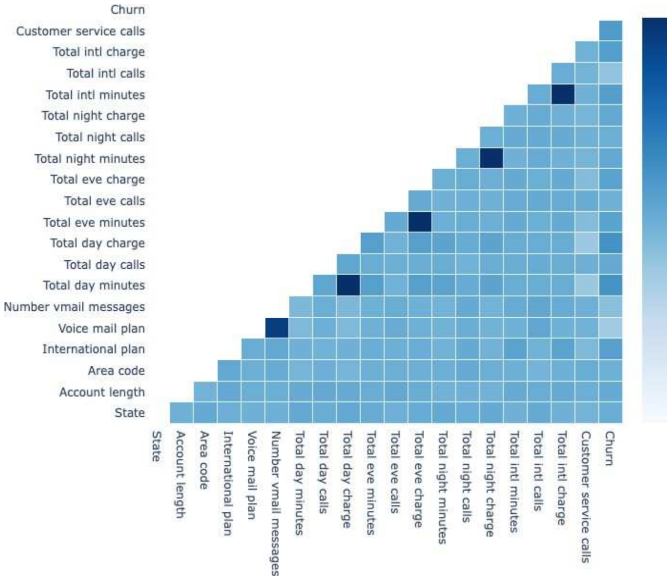

Fig. 3.3. Correlation Matrix

This plot can help to determine the strength and direction of the correlation between these features and the target variable, which is customer churn. By analyzing the correlation matrix plot, we can gain insights into which features are most strongly correlated with customer churn, which can be used to develop more accurate predictive models.

D. Model Building:

The specific machine learning algorithms used for building the models could be listed, along with their hyper parameters and any feature engineering techniques used. The metrics used for evaluating model performance, such as accuracy or F1 score, could also be explained in more detail.

IV. Results And Discussions

We selected a diverse set of machine learning algorithms to build our customer churn prediction model. We will use the Bernoulli Naive Bayes algorithm as a baseline, as it is a simple yet powerful algorithm for text classification tasks. Additionally, we will use the GaussianNaive Bayes algorithm, Support Vector Classifier, K-Nearest Neighbours with both Euclidean and Chebyshev distances, Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier.

A. Bernoulli Naive Bayes algorithm:

Bernoulli Naive Bayes is a probabilistic algorithm that applies the Bayes theorem with an assumption of independencebetween the features. It is suitable for binary and Boolean features, making it a good fit for the telecom churn dataset withits mostly binary and Boolean features. In the case of thetelecom industry, it can be used to classify customers into churnor non-churn groups based on binary features such as whether they have an international plan or not.

B. Gaussian Naive Bayes algorithm:

Gaussian Naive Bayes is a probabilistic algorithm that assumes the input features to be normally distributed. This algorithm is useful for classification problems where the features are continuous. In the context of our research, Gaussian NB can be used to classify customers as churners or non-churners based on their telecom service usage patterns.

C. Logistic Regression (baseline) algorithm:

Logistic Regression is a popular algorithm used for binary classification tasks. In our research, we have used Logistic Regression as a baseline algorithm to evaluate the performance of other algorithms. In the context of customer churn, the algorithm estimates the probability ofa customer churning or not churning based on the relevant features such as international plan, customer service calls,and call minutes.

D. Support Vector Machine:

The support Vector Machine (SVM) algorithm is a powerful and widely used machine-learning technique for classification tasks. It works by finding the best hyperplane that separates the data into different classes. SVM has been shown to perform well in various domains, including customer churn analysis in the telecom industry. In our study, we used SVM as one of the classification algorithms and achieved an accuracy of 74%.

E. K-Nearest Neighbours (Chebyshev):

KNN with the Chebyshev distance metric is also used as a classification algorithm for customer churn analysis. The Chebyshev distance metric is a way of measuring the distance between two points in space, and it takes into account the maximum difference in each dimension.

F. Decision Tree Classifier:

The decision tree classifier is a popular machine-learning algorithm that can handle both categorical and numerical data. It works by splitting the data into subsets based on the most significant variable, intending to create a decision tree that can accurately classify future data points.

G. K-Nearest Neighbours (Euclidean):

The K-Nearest Neighbours (KNN) classifier is a non-parametric algorithm used for classification and regression tasks. In our study, we used the Euclidean distance metric to calculate the distances between the samples in the feature space.

H. Random Forest Classifier (RFC):

Random Forest Classifier (RFC) is a popular ensemble learning algorithm that builds multiple decision trees and combines their predictions to obtain better accuracy and reduce overfitting. In our research, we have used RFC as one of the classification models for predicting customer churn in the telecom industry.

I. XGBoost Classifier:

XGBoost achieved the highest accuracy of 94%among all the algorithms we used. This indicates that it canbe an effective tool for predicting customer churn in the telecom industry. Its ability to handle missing data and its fast training speed make it an attractive choice for large datasets.

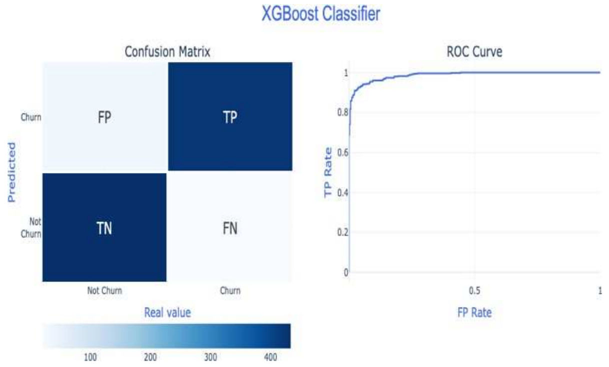

Fig. 4.1. Confusion Matrix and ROC Curve for Support Vector Machine

Another way to use the churn prediction model for customer retention is by identifying the customers who are at high risk of churning and taking proactive measures to retain them. A rate-per-threshold chart is a useful tool to evaluate the performance of a binary classification model at different decision thresholds. From the rate-per-threshold chart for the XGBoost classifier, it can be seen that the model performs well across a wide range of decision thresholds. 5, the model achieves a TPR of 0.

Table 1. The Results Of Obtained Experiments For Used Machine Learning Classification Models

Model	Acc	Precision	Recall	F1 Score	Area Under Curve
Bernoulli NB	0.64	0.64	0.64	0.64	0.64
Gaussian NB	0.73	0.71	0.8	0.73	0.75
LR (baseline)	0.74	0.73	0.76	0.74	0.75
SVM	0.74	0.73	0.77	0.74	0.75
K - NN (Chebyshev)	0.83	0.77	0.94	0.83	0.85
DT Classifier	0.84	0.9	0.77	0.84	0.83
K - NN (Euclidean)	0.87	0.82	0.97	0.87	0.89
RF Classifier	0.89	0.94	0.83	0.89	0.88
XGBoost Classifier	0.94	0.96	0.93	0.94	0.94

The precision score measures the proportion of correctly identified churn cases out of all predicted churn cases. A high precision score indicates that the model has correctly identified a high proportion of churn cases. The recall score, on the other hand, measures the proportion of correctly identified churn cases out of all actual churn cases. A high recall score indicates that the model has identified a high proportion of actual churn cases. In summary, the XGBoost Classifier is an effective model for predicting customer churn, and it can be used in conjunction with proactive customer retention strategies to reduce the churn rate and improve customer satisfaction.

VI. CONCLUSION

Our analysis suggests that the telecom industry should focus on improving its international plan services and addressing customer concerns after the fourth service call to reduce customer churn. Furthermore, companies should also aim to improve customer engagement and satisfaction during peak hours to reduce the likelihood of churning. The XGBoost classifier can be used as a reliable model to predict customer churn, and companies can use its predictions to identify high-risk customers and take proactive measures to retain them.

REFERENCE

[1] A. Zaky, S. Ouf and M. Roushdy, "Predicting Banking Customer Churn based on Artificial Neural Network," 2022 5th International Conference on Computing and Informatics (ICCI), New Cairo, Cairo, Egypt, 2022, pp. 132-139, doi: 10.1109/ICCI54321.2022.9756072.

[2] Mnassri, B. (2019) *Telecom churn dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets (Accessed: 02 February 2023).

[3] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[4] A. Larasati, D. Ramadhanti, Y. W. Chen and A. Muid, "Optimizing Deep Learning ANN Model to Predict Customer Churn," 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang, Indonesia, 2021, pp. 1-5, doi: 10.1109/ICEEIE52663.2021.9616714.

[5] Jajam, N. (2023) 'Arithmetic optimization with ensemble deep learning SBLSTM-RNN-IGSA model for customer churn prediction', *IEEE Access*, 11, pp. 93111–93128. doi:10.1109/access.2023.3304669.

[6] A. Raj and D. Vetrithangam, "Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 139-144, doi: 10.1109/CISES58720.2023.10183530.

[7] P. Bhuse, A. Gandhi, P. Meswani, R. Muni and N. Katre, "Machine Learning Based Telecom-Customer Churn Prediction," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 1297-1301, doi: 10.1109/ICISS49785.2020.9315951.

[8] K. D. Singh, P. Deep Singh, A. Bansal, G. Kaur, V. Khullar and V. Tripathi, "Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry," 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Kalady, Ernakulam, India, 2023, pp. 197-201, doi: 10.1109/ACCESS57397.2023.10199700.

[9] A. Mishra and U. S. Reddy, "A Novel Approach for Churn Prediction Using Deep Learning," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, 2017, pp. 1-4, doi: 10.1109/ICCIC.2017.8524551.