2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS) | 979-8-3503-4616-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACCESS57397.2023.10199700

Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry

Kiran Deep Singh
Chitkara University Institute of Engineering and Technology,
Chitkara University,
Rajpura, Punjab, India
kdkirandeep@gmail.com

Prabh Deep Singh
Department of Computer Science and Engineering, Graphic Era Deemed to be University.
Dehradun, India
ssingh.prabhdeep@gmail.com

Ankit Bansal
Chitkara University Institute of Engineering and Technology,
Chitkara University,
Rajpura, Punjab, India
erankitbansal@gmail.com

Gaganpreet Kaur
Chitkara University Institute of Engineering and Technology
Chitkara University,
Rajpura, Punjab, India
kaur.gaganpreet@chitkara.edu.in

Vikas Khullar
Department of Computer Science and Engineering,
Graphic Era Deemed to be University
Dehradun, India.
vikas.kullar@gmail.com

Vikas Tripathi
Department of Computer Science and Engineering,
Graphic Era Deemed to be University.
Dehradun, India
vikastripathi.be@gmail.com

***Abstract*— The telecommunications business is one of the key industries with a higher risk of revenue loss owing to client turnover and environmental impact. Thus, efficient and effective churn management includes targeted marketing campaigns, special promotions, or other incentives to keep the customer engaged in technological progress. There are a lot of machine learning algorithms available now, but very few of them can effectively take into account the asymmetrical structure of the telecommunications dataset. The efficiency of machine learning algorithms may also vary depending on how closely they approximate the real-world telecommunications data rather than the publicly available dataset. As a result, the researchers used various predictive models, including XGBoost, for this dataset. The accuracy achieved on the native dataset is 82.80%. Results show the effectiveness of the predictive model with great technological capabilities.**

Keywords—Churn Prediction; Telecommunication's Customers; Machine Learning, XGBoost, Exploratory Data Analysis.

I. INTRODUCTION

Telecom churn prediction is an essential aspect of customer relationship management for telecommunications companies. By identifying customers at risk of leaving, companies can take steps to retain them before they leave. Additionally, by monitoring the effectiveness of retention strategies and making adjustments as needed, telecommunications companies can ensure that they are effectively retaining customers and maximizing the value of their customer base [1].

The process of telecom churn prediction begins with collecting data on customer behavior. This data can include information on call patterns, usage patterns, demographics, and other relevant information. One of the most important factors in telecom churn prediction is identifying early warning signs of potential churn. For example, a customer who has not made a call in several months or has had multiple negative interactions with customer service may be at a higher risk of leaving. By identifying these early warning signs, a company can take steps to retain customers before they leave.

There are a few different methods that telecommunications companies use to predict churn. These algorithms can analyze the data and identify patterns associated with customers at risk of leaving. Another common method is to use statistical models, such as logistic regression or survival analysis. These models can analyze the data and estimate the probability of a customer leaving [2].

Once the data has been analyzed, telecommunications companies can use this information to develop strategies to retain customers. The targeted marketing campaigns, special promotions, or other incentives to keep the customer engaged. It is also important to note that customer retention strategies should be tailored to the specific needs of each customer [3]. Another important aspect of telecom churn prediction is measuring the effectiveness of retention strategies. The tracking metrics include customer retention rate, customer lifetime value, and customer satisfaction. By monitoring these metrics, telecommunications companies can determine the effectiveness of their retention strategies and make adjustments as needed.

Exploratory Data Analysis, referred to as EDA, enables analysts to gain insights into the data and identify patterns, trends, and outliers [4]. In the context of the telecommunication industry, EDA can be used to analyze customer data such as demographic information, usage patterns, contact information, payment history, and customer service interactions [5], [6]. Analyzing this data can help understand customer behavior and identify factors associated with a higher churn risk.

EDA typically begins with data cleaning and preparation, including removing missing or duplicate data and ensuring that the data is in the correct format [7], [8]. Telecommunications datasets used for churn prediction typically include a variety of properties or features that can be used to identify patterns and trends associated with customers at risk of leaving. Some common properties or features found in these datasets include:

979-8-3503-4616-9/23/$31.00 ©2023 IEEE

1. Demographic information: This can include the age, gender, income, and education level of the customer [9].
2. Usage information: This can include call history, text message history, data usage, and bill amount.
3. Contact information: This can include contract start and end date, contract length, and contract type (e.g., prepaid or postpaid).
4. Payment information: This can include payment history and payment method.
5. Device information: This can include the type of device the customer uses and the date the device was purchased.
6. Customer service interactions can include information such as the number of customer service interactions and the nature of those interactions.
7. Network information: This can include information such as the coverage area of the customer, the strength of the signal, and the availability of certain features (e.g., 4G or 5G).
8. Other external factors: This can include information such as weather, economic conditions, and events that could affect customer behavior.

This paper's structure continues: Section II discusses related work. Section III describes the general methodology, while Section IV covers the experiment and dataset. Section V covers the result and discussion. Section VI concludes the findings.

II. Related Work

Recently, the area of telecom churn prediction, with several studies conducted to understand the factors contributing to churn and develop effective prediction models.

Stephani et al. [10] examined performance indicators across six algorithms on the same dataset and found that Random Forest achieved the highest accuracy (89 percent). SVM using various kernel functions has been applied to a dataset of 3333 subscribers and 21 variables by Ionut Brandusoiu et al.[11] . Meanwhile, Hemlata Jain et al. [10] conducted prediction analysis using Logistic Regression and Logit Boost independently on the same dataset. They found that Logistic Regression yielded a higher accuracy (85.2385%) than Logit Boost. Sabbeh [12] compared ten different Machine Learning algorithms on the same dataset, finding that Random Forest and AdaBoost achieved the highest accuracy after undergoing 10-fold cross-validation of the dataset.

Different independent machine learning approaches have been tested on a publically available dataset with 5000 samples, with XGBoost coming out on top with an accuracy of 95% and an F- score of 80% [13], [14]. They suggested a method to cluster datasets with comparable properties based on span and then trained the clusters with only SVM, Logistic Regression, and Decision Tree for the combined model. Three classifiers' performances were compared on each cluster, and the best metrics were pooled. The authors' cumulative model accuracy using this method is 99%. J. Pamina et al. [15] used a different publicly available dataset with 7043 subscribers and implemented KNN, XGBoost, and Random Forest in their research. By employing XGBoost on this dataset, they achieved an accuracy of 79.8 percent. K-Means Clustering and XGBoost were applied to a public dataset, including 7043 samples in the study by Pan Tang [15]. Accuracy was improved from 79% with XGBoost alone to 81% with the K-Means clustering technique applied to the training set [16].

The author used a different publicly available dataset with 7043 subscribers in their research by employing XGBoost on this dataset to achieve an accuracy of 79.8 percent. Different independent machine learning approaches have been tested on a publically available dataset with 5000 samples, with XGBoost coming out on top with an accuracy of 95% and an F- score of 80%. They suggested a method to cluster datasets with comparable properties based on span and then trained the clusters with only Support Vector Machine (SVM), Logistic Regression, and Decision Tree for the combined model. Three classifiers' performances were compared on each cluster, and the best metrics were pooled. The authors' cumulative model accuracy using this method is 99%. K-Means Clustering and XGBoost were applied to a public dataset, including 7043 samples in the study by Pan Tang [17].Accuracy was improved from 79% with XGBoost alone to 81% with the K-Means clustering technique applied to the training set.

III. Methodology

The methodology used for Exploratory Data Analysis (EDA) and customer churn prediction in the telecommunications industry is shown in Fig 1, which typically involves several steps.

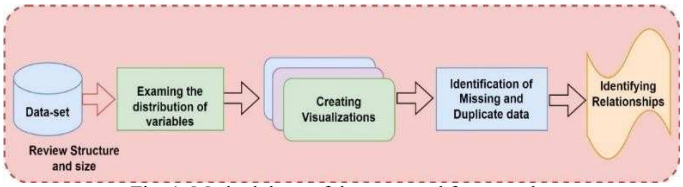

Fig. 1. Methodology of the proposed framework.

Data Collection: The first step in the methodology is to collect relevant data on customer demographics, usage patterns, contract information, payment history, and customer service interactions. This data is typically collected from the company's databases and customer relationship management (CRM) systems [18], [19].

Data Cleaning and Preparation: Ensuring that the data is in the correct format. It also includes dealing with any outliers or missing values.

Data Analysis: After the data is cleaned, various techniques can be used for Exploratory Data Analysis (EDA), such as data visualization, statistical analysis, and correlation analysis. These techniques are used to understand the distribution of the data and identify patterns, trends, and outliers.

Modeling: Once the data is analyzed, modeling techniques such as machine learning, statistical modeling, and survival analysis can be used for customer churn prediction. These techniques are trained on historical data to identify patterns associated with customers at risk of churn.

Evaluation: Based on the evaluation results, the model can be fine-tuned to improve its performance.

In this study, the data set is taken related to the telecom churn prediction for analysis, which is available for free download through the archival data platform Kaggle. The 20 predictor variables in this dataset are mostly concerned with customers' typical behaviors.

XGBoost is an implementation of the gradient boosting algorithm, an ensemble method that combines multiple decision trees to improve the model's performance.

This algorithm is known for handling large datasets and high-dimensional feature spaces. It can also deal with missing values and handle categorical variables effectively. XGBoost is a powerful algorithm that can be used for binary and multi-class classification [20]. It is an efficient algorithm with several important features that make it stand out from other gradient-boosting algorithms. One of the main features of XGBoost is its ability to handle missing data, a common problem in real-world datasets. Another important feature is its ability to handle categorical variables often present in datasets. It also includes several regularization techniques that help to prevent overfitting, which is a common problem in machine learning models. It includes L1 and L2 regularization, which penalize models with large weights, and a technique called "shrinkage," which slows down the learning rate as the model progresses. Moreover, it also includes a feature called "early stopping," which allows the algorithm to stop training when the model's performance stops improving and helps to prevent overfitting.

The XGBoost algorithm uses a specific decision tree called a gradient-boosted tree. A gradient-boosted tree is a decision tree that is fit using gradient boosting. Gradient boosting is an iterative algorithm that fits decision trees to the data by minimizing a loss function.

The loss function used in the XGBoost algorithm is the sum of the negative log-likelihoods of the observations, which is a common loss function for classification problems.

IV. Experiment and Dataset

The data set is investigated to have a deeper comprehension of the patterns present in the data and possibly formulate some hypotheses. First, the distribution of each variable is examined and moved to slice and dice data to identify any intriguing trends.

Demographics: First, we need to understand the customers in terms of their gender, age range, partners, and whether or not they have dependents. Only 16 percent of clients are considered to be senior folks. Therefore, the majority of our younger clients are represented in the data. Roughly half of the customers are in committed relationships, but only thirty percent have children or other dependents. In addition, as was to be predicted, most customers who do not have a spouse do not have any dependents. This percentage is 80 percent. There is no variation in the distribution of customers with or without dependents and partners based on gender, as was discovered when comparing the percentage of customers with and without dependents and partners by gender. Fig 2 shows the Boxplot for Churn vs. tenure, and the total No of customers per their tenure is shown in Fig 3. In addition, there is no distinction based on genderfor senior citizen status.

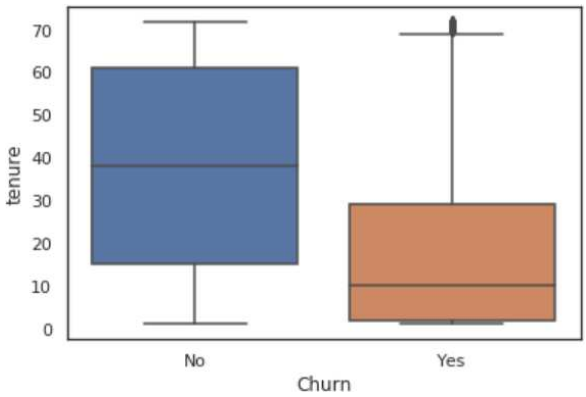

Fig. 2. Boxplot for Churn vs. tenure.

Details about Customer Accounts:

Tenure: Looking at the histogram, it is clear that many consumers have just been with the telecom business for one month, while quite a few have been with the company for almost 72 months. Fig 4 presents the total number of customers by their contract type. It may be because various clients have various contracts. Because of this, the client's ability to remain with or depart from the telecom provider may be more or less straightforward, depending on the contract terms they entered.

Contracts: Let's break down the number of customers by contract type to comprehend the preceding graph better. Fig 5 and Fig 6 show the Churn distribution as total and monthly, respectively.

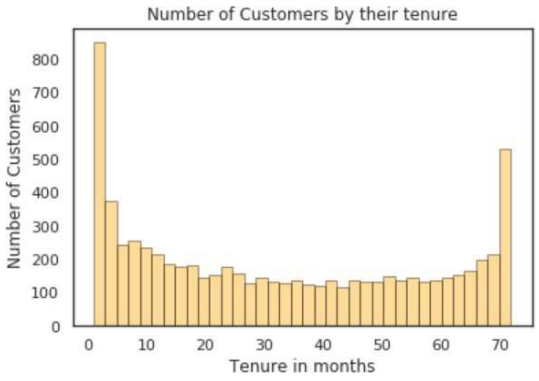

Fig. 3. Total No of customers as per their tenure.

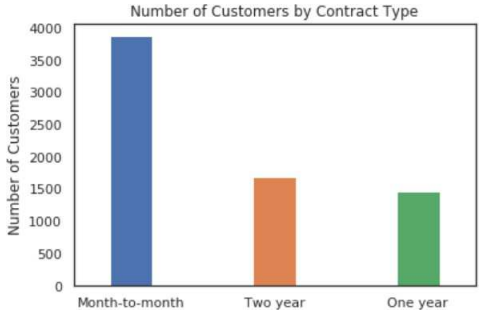

Fig. 4. Total number of customers by the type of theircontract.

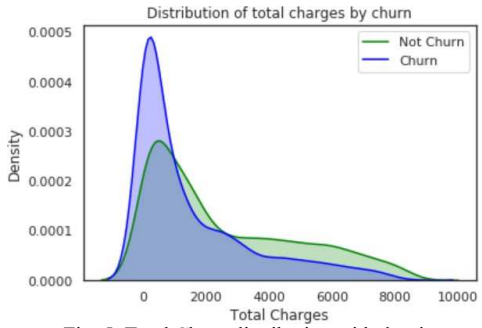

Fig. 5. Total Churn distribution with density.

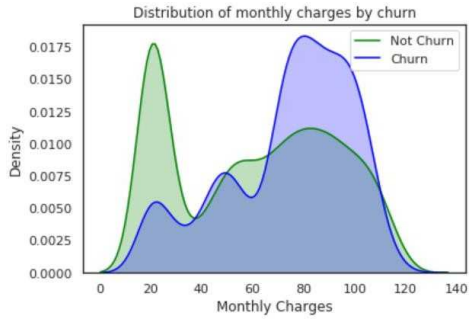

Fig. 6. Monthly Churn distribution with density.

After completing the preceding EDA, various predictive models like Random Forest, Adaptive Boosting (ADA Boost), SVM, Logistic Regression, and XG Boost are created and compared to some models. In logistic regression, it is essential to scale the variables to all fall within the range of 0 to 1. It assisted in increasing the accuracy from 79.7% to 80.0%. Additionally, the relevance of variables is consistent with the Random Forest algorithm.

TABLE 1: ACCURACY SCORE OF PREDICTIVE MODELS.

S. No.	Predictive Model	Accuracy score
1	Random Forest	0.808
2	ADA Boost	0.815
3	SVM	0.820
4	Logistic Regression,	0.807
5	XG Boost	0.828

V. Discussions

According to the random forest algorithm, the most significant variables for predicting churn are the monthly contract, tenure, and total costs. The random forest results are quite comparable to those of logistic regression and consistent with what we had anticipated from our EDA. SVM improves the accuracy to as much as 82%. For a more accurate prediction, however, we must examine the real positive and true negative rates in greater detail, including the Area Under the Curve (AUC). Although XG Boost is a paradigm for slow learning based on the concept of Boosting, it boosts the accuracy of test results to over 82%, as shown in Table 1. XG Boost is the superior technique over all others.

VI. Conclusion

This research analyzed XGBoost's performance in predicting customer turnover in the telecommunications sector. The significance of predictive model algorithms cannot be firmly stated since machine learning techniques directly depend on the nature of the dataset, and research studies have been done exclusively on the datasets available. As a result, this study has used both external and internal datasets to evaluate XGBoost's efficacy in churn prediction. In the future, the generalizability of the proposed approach can be improved; researchers can look at sampling strategies and text mining. To further investigate customer behavior in churn prediction.

REFERENCES

[1] S. Tiwari, S. Kumar, and K. Guleria, “Outbreak trends of CoronaVirus (COVID-19) in India: A Prediction,” Disaster Medicine and Public Health Preparedness, pp. 1–9, Apr. 2020.

[2] S. Ramirez and I. Lizarazo, “Detecting and tracking mesoscale precipitating objects using machine learning algorithms,” International Journal of Remote Sensing, vol. 38, no. 18, pp. 5045–5068, May 2017.

[3] S. Syed-Abdul, S. Malwade, A. A. Nursetyo, M. Sood, M. Bhatia, D. Barsasella, M. F. Liu, C. C. Chang, K. Srinivasan, M. Raja, and Y. C. J. Li, "Virtual reality among the elderly: A usefulness and acceptance study from Taiwan," BMC Geriatr, vol. 19, no. 1, 2019.

[4] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” BMC Medical Informatics and Decision Making, vol. 11, no. 1, Jul. 2011.

[5] Kumar Gourav and A. Kaur, “Computation Offloading Scheme Classification Using Cloud-Edge Computing for Internet of Vehicles (IoV),” pp. 459–485, Sep. 2022.

[6] R. Sharma, V. Kukreja, Rajesh Kumar Kaushal, A. Bansal, and A. Kaur, “Rice Leaf blight Disease detection using multi-classification deep learning model,” Oct. 2022.

[7] V. Mittal, D. Gangodkar, and B. Pant, "Deep Graph-Long Short-Term Memory: A Deep Learning Based Approach for Text Classification," Wirel Pers Commun, vol. 119, no. 3, pp. 2287–2301, Aug. 2021.
[8] P. Matta, B. Pant, and U. K. Tiwari, "DDITA: A naive security model for IoT resource security," Advances in Intelligent Systems and Computing, vol. 670, pp. 199–209, 2019.
[9] S. K. Sood and K. D. Singh, "An Optical-Fog assisted EEG-based virtual reality framework for enhancing E-learning through educational games," Computer Applications in Engineering Education, vol. 26, no. 5, pp. 1565–1576, 2018.
[10] A. Kaur, G. Singh, V. Kukreja, S. Sharma, S. Singh, and B. Yoon, "Adaptation of IoT with Blockchain in Food Supply Chain Management: An Analysis-Based Review in Development, Benefits and Potential Applications," Sensors, vol. 22, no. 21, p. 8174, Oct. 2022.
[11] I. Brandusoiu, G. T.- Margin, and undefined 2013, "Churn prediction in the telecommunications sector using support vector machines," academia.edu.
[12] S. F., "Machine-Learning Techniques for Customer Retention: A Comparative Study," International Journal of Advanced Computer Science and Applications, vol. 9, no. 2, 2018..
[13] S. K. Sood and K. D. Singh, "Optical fog-assisted smart learning framework to enhance students' employability in engineering education," Computer Applications in Engineering Education, vol. 27, no. 5, pp. 1030–1042, 2019.
[14] P. D. Singh, R. Kaur, K. D. Singh, and G. Dhiman, "A novel ensemble-based classifier for detecting the COVID-19 disease for infected patients," Information Systems Frontiers, vol. 23, pp. 1385–1401, 2021.
[15] J. Pamina, B. Raja, "An effective hybrid learning system for telecommunication churn prediction," Expert Systems with Applications, vol. 40, no. 14, pp. 5635–5647, 2013.
[16] S. Kaur, K. D. Singh, P. Singh, and R. Kaur, "Ensemble model to predict credit card fraud detection using random forest and generative adversarial networks," in Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2, 2021, pp. 87–97.
[17] M. S. Gaur and B. Pant, "Trusted and secure clustering in mobile pervasive environment," Human-centric Computing and Information Sciences, vol. 5, no. 1, Oct. 2015..
[18] P. Kaur and K. Singh, "Detection of Heart Diseases using Machine Learning and Data Mining," Int J Comput Appl, vol. 178, no. 31, pp. 975–8887, 2019.
[19] K. D. Singh, P. Singh, and S. S. Kang, "Ensembled-based credit card fraud detection in online transactions," in AIP Conference Proceedings, 2022, vol. 2555, no. 1, p. 50009.
[20] P. Kaur, P. Singh, and K. Singh, "Air pollution detection using modified traingular mutation based particle swarm optimization," Int. J. Eng. Technol., vol. 6, pp. 2005–2015, 2019.

