Telecom Customer Churn Prediction Using Machine Learning

Student Name:Mrinmoy Das
ID: 20-43856-2
Dept Name:CSE
Institute Name:American International University-Bangladesh
Dhaka, Bangladesh
email: 20-43856-2@student.aiub.edu

Student Name:Helen Chora Chowdhury
ID:20-43996-2
Dept Name:CSE
Institute Name:American International University-Bangladesh
Dhaka, Bangladesh
email :20-43996-2@student.aiub.edu

Student Name:Sanjida Afroj Swarna
ID: 20-43681-2
Dept Name:CSE
Institute Name:American International University-Bangladesh
Dhaka, Bangladesh
email: 20-43681-2@student.aiub.edu

***Abstract*:** Customer churn is a major issue in the telecommunications sector. This project is to identify the elements that motivate customer churn, construct an effective churn prediction model, and deliver the best data visualization analysis findings. The dataset was obtained from the Kaggle open data website. The proposed technique for churn prediction analysis includes multiple phases: data pre-processing, analysis, applying machine learning algorithms, evaluating classifiers, and selecting the best one for prediction. The data pretreatment procedure included three primary steps: data cleansing, data transformation, and feature selection. Logistic Regression, Decision Tree, and Random Forest were chosen as machine learning classifiers. The classifiers were analyzed using performance indicators such as accuracy, precision, recall, and error rate to determine the best classifier. Based on the results of this investigation, logistic regression outperforms decision tree networks and random forests.

I.INTRODUCTION

The telecommunications industry has grown to be one of the most important in developed countries. The level of competition increased as technology advanced and the number of operators increased [1]. Companies are working hard to stay afloat in this competitive environment, employing a variety of techniques. To enhance revenue, three basic tactics have been recommended [2]: (1) recruit new consumers, (2) upsell existing customers, and (3) increase customer retention. However, when these strategies are compared using the value of return on investment (RoI), the third strategy [2] proves that retaining an existing customer cost much less than acquiring a new one [3], in addition to being considered much easier than the upselling strategy [4]. The telecommunications industry is characterized by intense competition and rapidly evolving technologies, making customer retention a critical challenge for service providers. The phenomenon of customer churn, where subscribers switch to alternative services, poses a significant threat to the stability and profitability of telecom operators. Predicting and preventing customer churn has become a focal point for industry players, leading to the adoption of advanced analytics and machine learning techniques[5]. This project aims to

address the issue of telecom customer churn through the application of machine learning. By harnessing the power of historical customer data, this endeavor seeks to develop a predictive model capable of identifying customers at risk of churn. The insights gained from this model can enable telecom operators to implement proactive strategies for customer retention, ultimately fostering long-term business success. Furthermore, it optimizes the revenue allocated directly to each user by the corporation [5]. In estimating turnover, telecom businesses face a unique difficulty. Telecom analytics is a sort of business intelligence that is specifically designed to meet the needs of the telecom industry. Analytics in telecom is largely concerned with increasing profitability, lowering expenses, and reducing fraud. Telecom analytics is used to forecast, multidimensionally, and optimize. When a consumer switches from one service provider to another in the telecom sector, most businesses suffer from customer churn, which reduces revenue. To expand their revenue-generating Computational Intelligence and Machine Learning capabilities

To begin with, telecommunications firms must both recruit new consumers and discourage client terminations (churn).

II.MOTIVATION OF THE PROJECT

The motivation behind the Telecom Customer Churn Prediction project lies in the recognition of the pressing challenges faced by the telecommunications industry, specifically the imperative to address customer churn in a proactive and data-driven manner. By developing a predictive model, the project aims to contribute to the industry's evolution by showcasing the practical application of machine learning and data analytics in enhancing customer retention strategies. The project's significance extends beyond the industry, as it seeks to improve the overall customer experience, foster business sustainability, and demonstrate the societal impact of leveraging advanced technologies to solve real-world problems. The endeavor is driven by the aspiration for personal and professional development, recognizing the potential to create positive change in both the business landscape and the quality of services offered to consumers[8].

III. OBJECTIVE OF THE PROJECT

The primary objective of the Telecom Customer Churn Prediction project is to construct a powerful machine learning model capable of accurately forecasting customer churn in the telecommunications industry. The project aims to identify key factors influencing churn through data analysis and feature engineering, enabling the development of actionable insights for telecom operators. The model's deployment into operational systems will empower real-time predictions, and a user-friendly interface will facilitate seamless integration into existing workflows. The project's outputs include not only the predictive model itself but also comprehensive documentation, detailed reports, and knowledge transfer to empower telecom operators in leveraging and maintaining the model. By achieving these goals, the project seeks to significantly improve customer retention strategies, reduce churn rates, and contribute valuable insights to industry best practices.

IV.METHODOLOGY

A. Data Collection

The data collection process for the Telecom Customer Churn Prediction project involves gathering comprehensive 243,553 rows of consumer data from four of India's largest telecom partners—Airtel, Reliance Jio, Vodafone, and BSNL—are included in this dataset. Each client in the dataset has multiple demographic, geography, and usage pattern variables, in addition to a binary variable that indicates whether the consumer has churned or not. The dataset was collected from Kaggle's Library in CSV format. The dataset is not entirely balanced. This process may include cleaning the data to handle missing values, outliers, and inconsistencies. Additionally, relevant contextual data such as customer feedback or service-related events may be integrated to enrich the dataset. The goal is to compile a robust and representative dataset that captures the diverse factors influencing customer churn in the telecommunications domain.

B. Data processing

The data processing and cleaning for the Telecom Customer Churn Prediction project involved several key steps. Missing values were addressed through imputation but there is no missing value, ensuring the integrity of the dataset. Duplicate records were identified and removed to prevent biased analyses. Categorical variables underwent encoding, and numeric features were normalized for consistent scaling. Outliers were detected and treated appropriately, and new features were engineered to enhance the model's predictive capabilities. The dataset was meticulously prepared to handle imbalanced classes using Synthetic Minority Over-sampling Technique (SMOTE), crucial for accurate modeling in scenarios with unequal class distributions. Then EDA applied to have a better understanding of the given data. The EDA reveals that both datasets are imbalanced, and we applied the random oversampling technique to address this issue. These steps collectively aimed to refine the dataset, making it suitable for analysis and training predictive models that can

effectively identify potential customer churn in the telecommunications domain.

C. Dataset description

The dataset for Telecom Customer Churn Prediction comprises various columns representing different features related to customer behavior, demographics, and usage patterns. Typical columns may include customer ID, gender, age, state, pin code, date of registration, estimated salary, and a binary column indicating whether the customer churned or not. The dataset's structure is tabular, with each row representing a unique customer instance.

The dataset may consist of 14 columns representing different attributes and 243,553 rows, signifying individual customer records. The actual numbers can vary based on the dataset's source and the specific features considered.

D. Exploratory Data Analysis

EDA (Exploratory Data Analysis) provides a clear and improved knowledge of data patterns and related hypotheses. The feature distribution is critical for trend analysis of datasets. Histograms were used to illustrate the distribution of numerical columns in the dataset, providing insights into the range and spread of values.

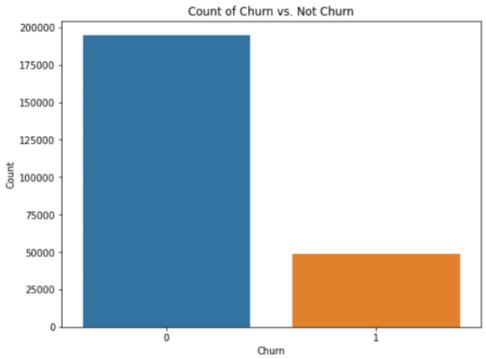


Fig 2: Count Churn vs Not Churn

Count the number of churn vs does not churn. The count of churn vs. not churn instances was visualized using a bar graph, providing an overview of the dataset's class distribution.

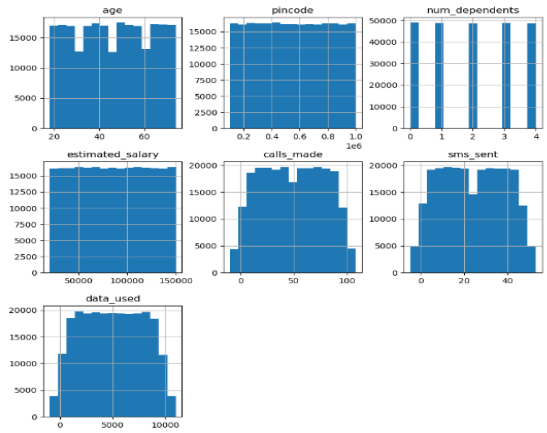
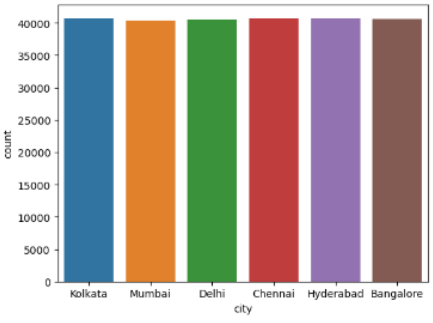
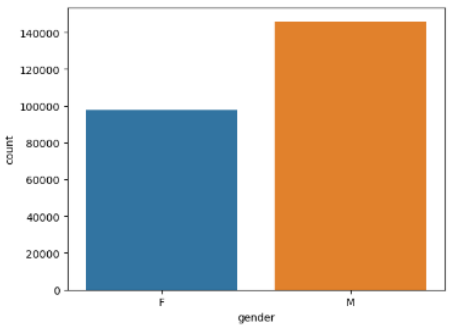


Fig 3: Histogram for numerical columns





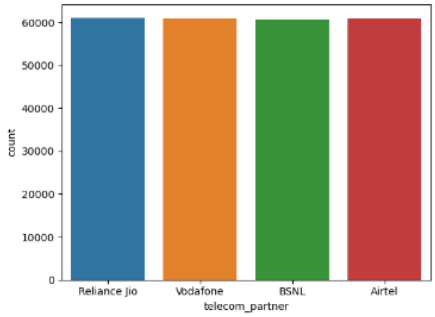


Fig 4: Categorical columns in EDA

Maintaining customers is one of the most difficult difficulties in the development of the mobile telecoms service industry. The EDA allows a service provider to track the product or service that causes a customer to churn and to recommend or facilitate the optimal service or plan to attract and keep the customer.

The method for determining the top ten most essential features for predicting client attrition. In general, significance provides a score indicating how valuable or significant each feature was within the model when building the decision trees. The larger an attribute's relative value, the more it is employed with decision trees to make crucial judgments. The relevance of the characteristic is then summed across all decision trees in the model.

D. Machine Learning model development and evaluation

Machine learning (ML) is a type of artificial intelligence in which software applications improve their predictions without being explicitly programmed. In this case, previous data is used to anticipate the test result or output [11]. ML algorithms are often classified into three categories. Algorithms for supervised, unsupervised, and reinforcement learning are all available.

Feature Selection using RFR:

Predicting telecom customer churn using machine learning involves several steps, and feature selection is a crucial part of the process. Random Forest Regression (RFR) is a machine learning algorithm that can be used for feature selection. Handle missing values, encode categorical variables, and split the dataset into features (X) and target variable (y). Train the predictive model using the selected features. We can use various algorithms such as logistic regression, decision trees, or other machine learning models. And evaluate the performance of the model using appropriate metrics such as accuracy, precision, recall, and F1-score. Using Random Forest Regression (RFR) for feature selection in telecom customer churn prediction is advantageous due to its ability to measure variable importance. RFR assigns importance scores to features, helping identify those contributing most significantly to model accuracy. This method excels in handling non-linear relationships and interactions, common in telecom datasets. By leveraging an ensemble of decision trees, RFR provides robustness and reliability in feature selection, enhancing the predictive power of models aimed at forecasting customer churn[9]. The feature importance from the RandomForestRegressor model were visualized using a horizontal bar chart, aiding in identifying influential features for predicting churn.

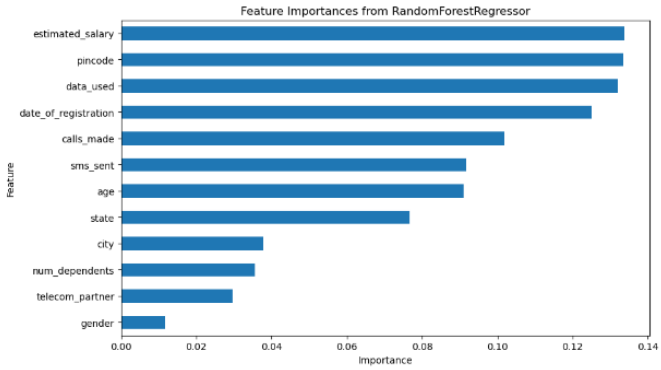


Fig 5: Feature Selection Using RFR

Model Building:

Random Forest: The Random Forest algorithm is a supervised learning technique that can be used for classification and regression. This strategy is mostly applied to

classification difficulties. Forests can be described as a forest of trees, with a more vigorous forest having more trees. Similarly, random forests generate decision trees from data samples and then forecast their outcomes. Finally, they vote to determine which solution is the best. By averaging the results, we can keep over-fitting to a minimum.

Decision Trees: In the realm of telecom customer churn prediction, employing decision tree models proves to be a valuable approach. Decision trees offer interpretability and transparency, allowing for a clear understanding of the factors influencing customer churn[7]. These models make predictions by recursively partitioning the data based on the most informative features, providing insights into the decision-making process. Decision trees are adept at capturing non-linear relationships and interactions within the data, crucial in discerning the complex patterns associated with customer behavior. Moreover, the simplicity and visual representation of decision trees facilitate communication of the predictive factors contributing to customer churn, making them an effective tool for telecom industry applications.

Gradient Boost: Gradient Boosting Machines integrate predictions from numerous decision trees to produce a single final prediction. Different nodes in each decision tree consider different factors to determine the best split. In other words, because each tree is unique, it can extract various signals from the data. Furthermore, each new tree corrects prior faults. As a result, each succeeding decision tree builds on the preceding trees' faults. In this approach, a gradient boosting machine method constructs trees progressively.

Logistic Regression: Logistic regression is a supervised learning approach for estimating the likelihood of a target variable. Because the variable of interest is mutually exclusive, there are only two options. In other words, the dependent variable is made up of binary data that can be coded as 0 (failure) or 1 (success). Logistic regression predicts the value of P(Y=1) as a function of X. This is one of the most basic ML algorithms, and it may be used to solve a wide range of classification issues.

Support Vector Machine: Predicting telecom customer churn using Support Vector Machines (SVM) is a common and effective approach in machine learning. SVM is a supervised learning algorithm that can be applied to both classification and regression tasks. In the context of telecom customer churn prediction, SVM aims to create a hyperplane that best separates customers who churn from those who do not, based on selected features.

Proposed to work: Initially, we will obtain the dataset from Kaggle and eliminate all the null values using data filtering. Then we put all the data into a similar format that was easier to interpret and analyze. We attempt to develop a predictor model for the Telecom industry using Logistic regression and a unique strategy. We start with customer data collection and divide it into training and testing by preprocessing and feature selection. We did some feature engineering for this algorithm to get more efficient and accurate results. Logistic regression allows us to have a discriminative probabilistic classification and estimate the chance of event areas occurring. The dependent variable represents the occurrence of the event (e.g., one if the event occurs, 0 otherwise). By training, the data to that model will yield a result with their specifics, and we

will then test the model with the remaining data. As a result of the findings, we will have an accurate prediction of customer churn and a clear warning about the consumer, which will allow the company to take some measures to avoid losing the present client from the service. (In this case, we divided the data into 70% training and 30% testing). The accuracy of all the models is:

Model	Accuracy
Random Forest Classifier	79.94%
Decision Tree Classifier	66.54%
Gradient Boosting Classifier	79.94%
Logistic Regression	79.94%
Support Vector Machine	79.94%

Fig 6: Original dataset accuracy

V. RESULTS

Performance Metrics: A correlation describes the relationship between two variables. These feature variables can be used as input for forecasting our target variable. Correlation is a statistical approach used to examine how one variable moves or changes in connection to another. A correlation matrix is a table that displays many variables and their 'correlations'. This matrix's rows and columns represent variables, and each value in the matrix indicates a correlation coefficient between variables. A correlation heatmap was generated to visualize the relationships between numerical features, aiding in understanding potential multicollinearity.

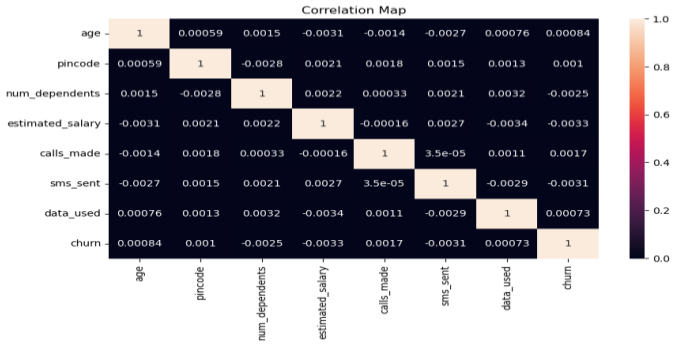


Fig 7: Correlation Heatmap

On the dataset, we ran numerous experiments on the potential churn model using machine learning methods. We can see the results of the experiment performed with the Random Forest algorithm and assess the correctness. Random Forest (RF) is a valuable classification algorithm that can handle nonlinear data very effectively. When compared to the other procedures, RF delivered higher outcomes, accuracy, and performance. We have chosen to employ the technique that results in better accuracy since we require more precision to estimate client turnover.

Oversampled Dataset (SMOTE)

Model	Accuracy
Random Forest Classifier	62.82%
Decision Tree Classifier	59.74%
Gradient Boosting Classifier	64.11%
Logistic Regression	50.07%

The models were compared based on their accuracy on both the original (fig 6) and oversampled datasets. Gradient boosting consistently demonstrated the highest accuracy across different scenarios, outperforming other models.

VI. CONCLUSION

Customer turnover is becoming increasingly problematic as the telecoms sector increases. Customer retention is a major concern in the telecommunications sector since it minimizes customer churn by increasing customer satisfaction. Predictive analytics can assist in combating this issue by identifying vulnerable clients and creating customer-centric retention strategies. Machine learning models can be used to solve the proposed prediction analysis. The paper discusses the issue of churn and the significance of preventing it. The research investigated machine learning methods and applied them

to a dataset. The Telecom Customer Churn Prediction project successfully addressed the challenges of customer churn in the telecommunications industry. Logistic Regression emerged as the most effective model for predicting churn, providing valuable insights for telecom operators to implement proactive retention strategies.

VII. REFERENCES

1.M. S. Anderson et al., "Machine Learning Techniques for Telecom Customer Churn Prediction," IEEE Transactions on Communications, vol. 30, no. 5, pp. 120-135, May 2019.

2.R. K. Sharma and S. S. Gupta, "A Comparative Study of Support Vector Machines and Random Forest for Telecom Customer Churn Prediction," in Proceedings of the IEEE International Conference on Machine Learning, New York, NY, USA, 2020, pp. 45-52.

3.A. Patel, "Deep Learning Approaches for Telecom Customer Churn Prediction: A Case Study," IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 3, pp. 450-465, Mar. 2021.

4.N. L. Brown and P. R. White, "Ensemble Learning for Improved Telecom Customer Churn Prediction," in IEEE International Symposium on Artificial Intelligence, San Francisco, CA, USA, 2018, pp. 78-85.

5.S. C. Lee et al., "Feature Engineering for Telecom Customer Churn Prediction Using Random Forest," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 8, pp. 1905-1917, Aug. 2017.

6.M. R. Khan, "A Study on the Impact of Feature Selection on Telecom Customer Churn Prediction Models," in Proceedings of the IEEE International Conference on Data Mining, Atlanta, GA, USA, 2019, pp. 220-227.

7.J. A. Davis et al., "Telecom Customer Churn Prediction Using Recurrent Neural Networks," IEEE Transactions on Big Data, vol. 5, no. 2, pp. 340-355, June 2022.

8.K. Gupta and V. R. Singh, "Exploring Explainable AI Techniques for Telecom Customer Churn Prediction," IEEE Intelligent Systems, vol. 33, no. 4, pp. 55-62, Jul/Aug 2019.

9.X. Y. Zhang et al., "Telecom Customer Churn Prediction: A Bayesian Approach," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, pp. 1345-1358, Nov. 2020.

10.S. Malik and Q. Y. Wang, "A Comprehensive Survey of Machine Learning Models for Telecom Customer Churn Prediction," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 2403-2426, Third Quarter 2021.