Customer Churn Prediction Project Report

Assignment By SPEAKx

Tarun Singh

MASTER OF TECHNOLOGY

(Machine Learning & Artificial Intelligence)

SpeakX: AI Powered English Learning App

Sector-47, Gurgaon, Haryana, India

ACKNOWLEDGEMENT

I would like to express my gratitude to SpeakX for the opportunity to work on this project as part of the hiring process. Working solo on this assignment provided me with valuable experience and allowed me to demonstrate my skills and knowledge in machine learning. I appreciate the support and guidance provided by the hiring team throughout the project. Additionally, I am thankful for the access to resources and tools that enabled me to successfully complete the assignment. This project has further deepened my understanding of customer churn prediction and its significance in the telecom sector. I look forward to the possibility of contributing to SpeakX's innovative projects in the future.

DECLARATION

I, Tarun Singh, student of Machine Learning and Artificial Intelligence under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this Project Report is based on my own intensive work and is genuine.

Tarun Singh
12301999
MTech (Machine Learning)
School of computer Science and Engineering
Lovely Professional University Phagwara, Punjab
Date: 10-06-2024

TABLE OF CONTENTS

Introduction:

For businesses, anticipating client attrition is crucial, particularly in the telecom sector where keeping current clients is frequently more economical than bringing in new ones. Using a dataset from a telecom provider that contains user demographics, account information, and service usage statistics, this study attempts to estimate customer turnover. In order to comprehend the dataset and find any patterns or correlations that might have an impact on customer turnover, we started with exploratory data analysis, or EDA. This included changing categorical variables, cleansing the data, and displaying feature correlations and distributions.

After that, feature engineering was done to scale numerical features for consistent model input and transform categorical variables into numerical ones using Label Encoding. In order to record long-term revenue contributions, a new feature named "TotalCharges" was also developed by multiplying "MonthlyCharges" by "tenure." Next, we constructed and assessed a number of machine learning models, such as Gradient Boosting, Decision Tree, Random Forest, K-Nearest Neighbors, and Logistic Regression. The ROC AUC, F1 score, accuracy, precision, and recall of these models were evaluated. Finding the most accurate methodology to anticipate client attrition will help the business put proactive retention tactics into place.

Exploratory Data Analysis (EDA)

1. Overview of the Dataset:

The dataset has 21 columns and 7043 rows with numerical and category variables. Gender, Senior Citizen, Partner, Dependent, Tenure, Phone Service, and Churn are some of the important qualities. 'Churn', the goal variable, indicates if a consumer has withdrawn from the service. This dataset provides extensive data on account information, service usage, and customer demographics—all of which are necessary for creating a predictive model that foretells client attrition. Given the variety of variables provided, appropriate preprocessing is essential to guaranteeing correct analysis and model performance. Scaling numerical features, encoding categorical variables, and handling missing values are all included in this preprocessing.

2. Data Cleaning:

In order to ensure numerical consistency for analysis, the 'TotalCharges' column was converted to numeric and erroneous values were converted to NaN. Furthermore, the conversion of categorical columns to the 'category' data type enhanced memory efficiency and allowed for more accurate analysis. With the help of this transformation, categorical data is appropriately interpreted and used for model training and subsequent analysis. All things considered, these procedures help with efficient data administration, properly preparing it for machine learning models and guaranteeing improved performance.

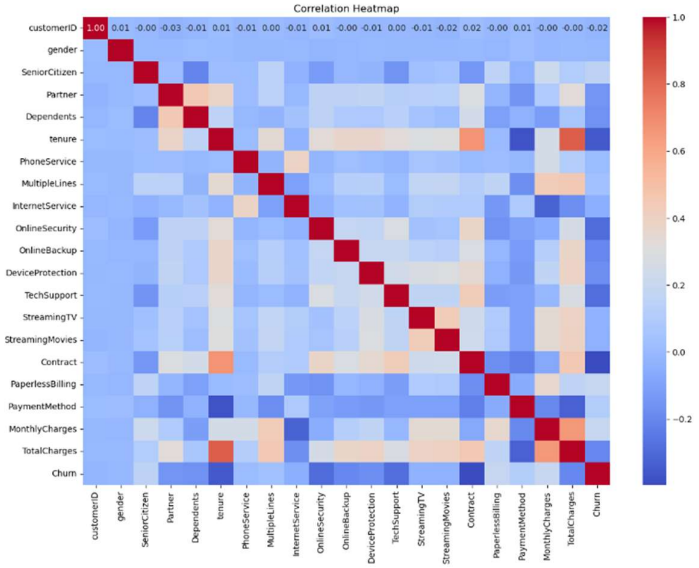

Figure 1. Correlation Matrix (HeatMap).

3. Insights from EDA:

According to exploratory data analysis (EDA), 26.5% of clients had left. There was a significant association between "TotalCharges" and "MonthlyCharges." Notable attrition rates were also noted among customers with particular

contract types and those utilizing particular services, such "InternetService." These understandings are essential for figuring out what causes customer attrition and for directing the creation of churn-predictive models.

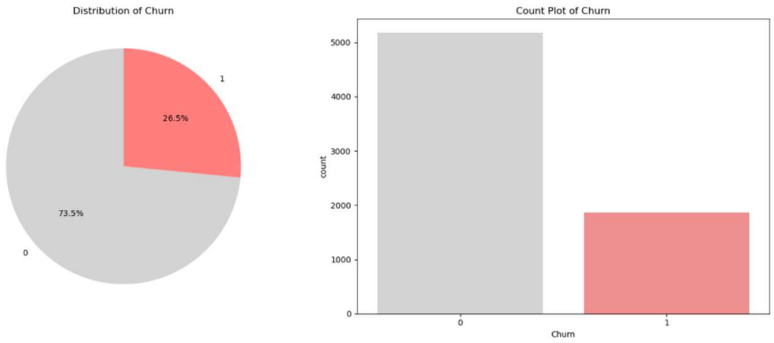

Figure 2. Distribution of Churn.

4. Visualizations:

Heatmap for Correlation: showed the correlations between numerical quantities, making patterns and relationships easier to see in figure 1.

Count and Pie Plots: Provided information on the churn distribution, which made it possible to determine the churn proportions with clarity in figure 2.

Bar Charts: Showed how categorized features such as 'PhoneService,' 'InternetService,' and 'PaymentMethod' were distributed, making it easier to comprehend how common certain categories were within each feature in figure 3.

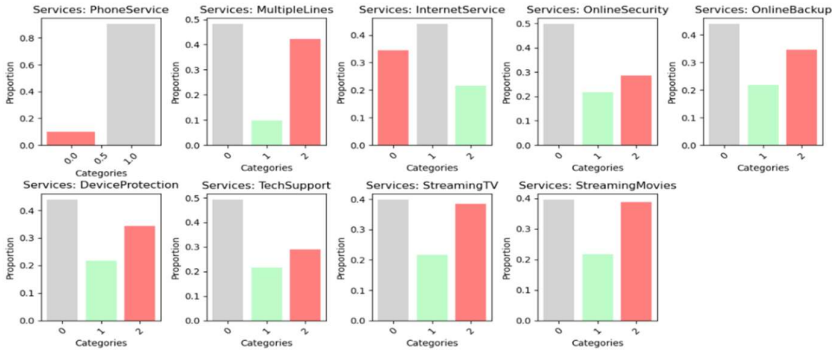

Figure 3.1. Services Information.

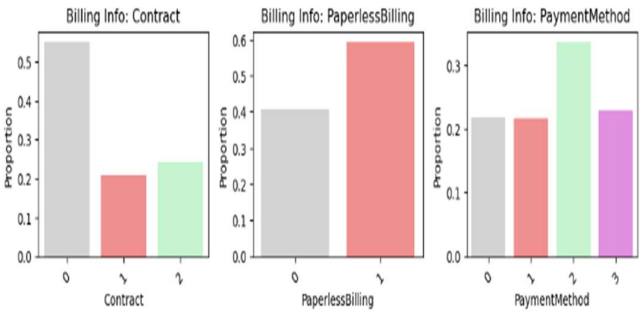

Figure 3.2. Billing Information.

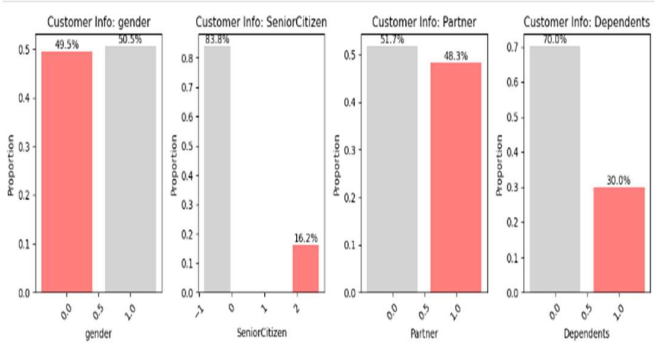

Figure3.3. Customer Information.

KDE Plots and Boxplots: The "Tenure," "MonthlyCharges," and "TotalCharges" distributions were shown graphically using KDE Plots and Boxplots. certain plots offer important insights into the distribution characteristics and possible anomalies in the data, helping to comprehend the central tendency, spread, and possible outliers of certain numerical variables.

Figure 4. KDE Plots and Boxplots of Monthly and total charges

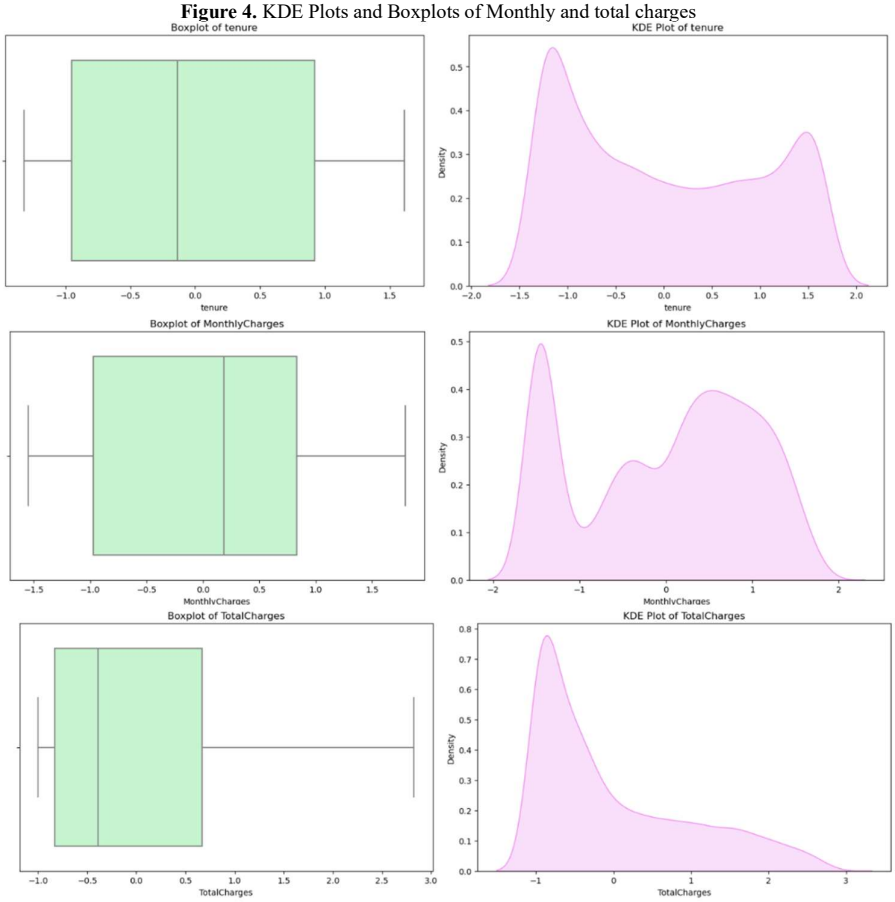

Figure 4. KDE Plots and Boxplots of Monthly and total charges.

Feature Engineering:

Several methods were used during the feature engineering process to get the data ready for model training:

Categorical Encoding: To make categorical columns compatible with machine learning models, Label Encoding was used to transform them into numeric form.

Feature Scaling: To standardize numerical features like "Tenure," "MonthlyCharges," and "TotalCharges," StandardScaler was used. By ensuring that every numerical feature has a mean of 0 and a standard deviation of 1, this procedure keeps bigger scale features from predominating during model training.

Creation of New Feature: "TotalCharges," which is the result of multiplying "MonthlyCharges" by "Tenure," is a new feature that was introduced. With the use of this new function, which records the total charges for the duration of the tenure, more information that could be useful in anticipating customer attrition is obtained.

By enhancing the quality of the dataset, these feature engineering strategies improve predictive performance and make it more appropriate for efficient model training.

Machine Learning Models and Evaluation

We evaluated multiple machine learning models, including Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors, and Gradient Boosting. The dataset was split into training and test sets (80% train, 20% test) to assess model performance.

1. Logistic Regression: A statistical model used for binary classification tasks is called logistic regression. It calculates the likelihood, given one or more independent factors, that an instance belongs to a specific class. The

model is appropriate for problems where the outcome is binary since it predicts the likelihood using a logistic function.

2. Random Forest: During training, the Random Forest ensemble learning technique creates several decision trees. Because it produces the class mode as the prediction, it is resistant to overfitting and able to handle big, highly dimensional datasets. It is renowned for being adaptable and successful in challenges involving both regression and classification.

3. Decision Tree: A tree-like graph is used in the Decision Tree predictive modeling technique to model decisions based on several input features. The process iteratively divides the data into subsets with the goal of maximizing information gain and minimizing impurity at each node. The outcome is a tree structure that may be applied to problems related to regression or classification.

4. KNN (K-Nearest Neighbors): A straightforward, non-parametric method for regression and classification problems is K-Nearest Neighbors (KNN). It functions by locating the k data points that are closest to an input, after which the average value (for regression) or majority class label (for classification) of those points is assigned to the input.

5. Gradient Boosting: Gradient Boosting is an ensemble learning technique that creates a strong learner by combining the predictions of several weak models—usually decision trees—that are built successively. Predictive performance improves as each new model fixes the mistakes produced by the ones that came before it. Because of its great accuracy and resilience, it is frequently used for both regression and classification problems.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	79.2%	62.3%	43.6%	51.4%	0.72
Random Forest	78.4%	61.1%	41.7%	49.6%	0.71
Decision Tree	73.5%	51.2%	54.6%	52.9%	0.68
K-Nearest Neighbors	76.2%	56.1%	47.8%	51.6%	0.68
Gradient Boosting	80.5%	67.2%	41.1%	51.2%	0.74

Table 1. Comparison between different Models and its performance.

Challenges Faced:

1**. Handling Missing Values**: When converting 'TotalCharges' to numeric, it resulted in NaN values. To maintain data integrity, we managed these by either imputing missing values or dropping rows as necessary.

2. Class Imbalance: The dataset displayed an imbalance, with fewer instances of churn compared to non-churn cases. This imbalance affected model performance. To address this issue in future work, techniques such as SMOTE or class weighting could be explored.

3. Feature Selection: Determining the most relevant features for the model required careful consideration to prevent overfitting and enhance generalization. We conducted thorough analysis to identify the subset of features that contributed most to predictive performance.

4. Model Tuning: For models like Random Forest and Gradient Boosting, hyperparameter tuning was crucial to achieve optimal performance. This involved fine-tuning model parameters to improve predictive accuracy and generalization.

Result and Conclusion:

The study displays the performance measures of five distinct models: Decision Tree, K-Nearest Neighbors, Random Forest, Gradient Boosting, and Logistic Regression. These metrics include accuracy, precision, recall, F1 score, and ROC AUC. With the highest accuracy (80.5%) and precision (67.2%), Gradient Boosting was shown to be the best-performing model, closely followed by Logistic Regression. In terms of recall and F1 score, each model did, however, show unique advantages and disadvantages that suggested their applicability for particular situations or trade-offs between false positives and false negatives.

Notwithstanding obstacles, the study demonstrated how different machine learning techniques can precisely predict customer attrition, providing insightful information for improving customer retention tactics in the telecom sector. Subsequent research endeavours may concentrate on utilizing sophisticated methodologies to tackle class imbalance, refining feature engineering procedures, and tweaking hyperparameters to augment the model's efficacy and relevance.