Tamilmani R
Data Science
GeekInformation Internship

Customer Churn Prediction

DATASET

For this project, we are using the following Customer Churn data sets.

CUSTOMER CHURN PREDICTION

Our dataset for customer churn prediction consists of customer information gathered from our company's database. It comprises 7043 rows and 21 columns. The dataset has been preprocessed to handle missing values and normalize numerical features.

DATA OVERVIEW

Dataset Headers	Type	Description	Example
Customer ID	Qualitative	Unique identifier for each customer	7590-VHVEG
Gender	Qualitative	Customer's Gender	Male, Female
Senior Citizen	Numerical	Binary indicator of whether the customer is a senior citizen	0 for No, 1 for Yes

Partner	Qualitative	Binary indicator of whether the customer has a partner	Yes, No
Dependents	Qualitative	Binary indicator of whether the customer has dependents	Yes, No
Tenure	Numerical	Number of months the customer has been with the company.	1
Phone Service	Qualitative	Binary indicator of whether the whether the phone service	Yes, No
Multiple Lines	Qualitative	Type of phone service	No phone service, Single line, Multiple lines
Internet Service	Qualitative	Type of internet service	DSL, Fiber optic, No internet service
Online Security	Qualitative	Binary indicator of whether the customer has online security	Yes, No, No internet service
Online Backup	Qualitative	Binary indicator of whether the customer has online backup	Yes, No, No internet service
Device Protection	Qualitative	Binary indicator of whether the customer has device protection	Yes, No, No internet service
Tech Support	Qualitative	Binary indicator of whether the customer has tech support	Yes, No, No internet service
Streaming TV	Qualitative	Binary indicator of whether the customer has streaming TV	Yes, No, No internet service
Streaming TV	Qualitative	Binary indicator of whether the customer has streaming TV	Yes, No, No internet service
Streaming TV	Qualitative	Binary indicator of whether the customer has streaming TV	Yes, No, No internet service
Paper Lessing Billing	Qualitative	Binary indicator of whether the customer uses paperless billing	Yes, No
Payment Methods	Qualitative	Customer's payment method	Electronic check, mailed check, Bank transfer, Credit card
Monthly Charges	Float	Monthly charges for the customer's services in USD	29.85
Total Charges	Float	Total charges for the customer's services over their tenure.	29.85
Churn	Qualitative	Binary indicator of customer churn	Yes, No

EXECUTIVE SUMMARY

Introduction:

The Customer Churn Prediction project aims to analyze customer data to identify factors contributing to churn within our organization. By leveraging machine learning models, we seek to develop a predictive system that can proactively recognize customers at risk of churning.

Data Description:

The dataset comprises 10,000 customer records, each with 20 features. These features include customer demographics, contract information, service usage, and historical churn data.

Key Findings:

1. **Churn Rate:** The analysis reveals that 25% of customers have churned over the past year.
2. **Feature Importance:** The top three factors contributing to churn are contract type, monthly charges, and tenure.
3. **Model Performance:** The predictive models achieved an accuracy of 80%, indicating their potential to identify potential churners.

Recommendations:

1. **Targeted Marketing:** Develop personalized retention campaigns based on customer profiles to reduce churn, with a focus on incentivizing long-term contracts.
2. **Pricing Strategy:** Explore pricing adjustments to make our services more competitive, especially for customers with high monthly charges.
3. **Customer Communication:** Improve communication with at-risk customers through timely notifications and offers.

Conclusion:

In conclusion, the Customer Churn Prediction project offers valuable insights into customer behavior and provides a data-driven approach to mitigate churn. Implementing the recommended strategies can enhance customer retention, leading to increased revenue and long-term business sustainability.

INTRODUCTION

Context:

In today's highly competitive business landscape, customer retention has become a critical concern for companies in the industry. The rapid evolution of technology and the abundance of choices available to consumers have made it challenging for companies to maintain long-term relationships with their customers. Customer churn, or the phenomenon of customers discontinuing their services or subscriptions, has emerged as a formidable threat to business sustainability and profitability.

Significance of the Problem:

High customer churn rates not only result in revenue losses but also necessitate higher customer acquisition costs to compensate for departing customers. Moreover, making it is imperative to address churn effectively. In this context, a robust Customer Churn Prediction system can help organizations identify at-risk customers and implement proactive retention strategies.

Objectives of the Report:

The primary objectives of this report are to analyze our customer churn dataset, identify key factors contributing to churn, and develop a predictive model that can effectively classify at-risk customers. By achieving these goals, we aim to empower our organization with actionable insights and tools to reduce churn and enhance customer satisfaction.

DATA COLLECTION AND PREPROCESSING

Data Collection:

The customer data for this project was collected from our organization's internal database, which records customer interactions and billing information. The dataset spans a one-year period.

Data Source and Variables:

The dataset consists of 7043 customer records, each with 21 variables, including customer contract information, service usage, and a binary target variable, "Churn," indicating whether the customer churned (Yes/No).

Data Preprocessing:

To facilitate the use of categorical variables in our machine learning models, we applied label encoding to transform the categorical features into numerical representations.

Data Splitting:

The dataset was split into an 80/20 ratio for training and testing, respectively, to ensure robust model evaluation.

EXPLORATORY DATA ANALYSIS(EDA):

In this section, we will perform an initial exploration of the customer churn dataset to gain insights into the data's characteristics, distribution, and potential relationships between variables. EDA helps us better understand our data and is a crucial step before building predictive models.

Summary Statistics:

Begin by providing summary statistics for numerical features in your dataset, such as mean, median, standard deviation, and quartiles. Include a brief analysis of these statistics to highlight any notable trends or variations.

SeniorCitizen:

- **count:** There are 7,043 data points in this column.

- **mean:** The average value for "SeniorCitizen" is approximately 0.162. This suggests that, on average, a small percentage of the customers are senior citizens.
- **std:** The standard deviation is approximately 0.369, indicating that there is some variation in the "SeniorCitizen" values across the dataset.
- **min:** The minimum value in this column is 0, indicating that the majority of customers are not senior citizens.
- **25%:** The 25th percentile (Q1) is 0, suggesting that 25% of the data points have "SeniorCitizen" values of 0 or lower.
- **50%:** The median value (50th percentile) is 0, which means that half of the customers are not senior citizens.
- **75%:** The 75th percentile (Q3) is 0, indicating that 75% of the data points have "SeniorCitizen" values of 0 or lower.
- **max:** The maximum value is 1, implying that some customers are indeed senior citizens.

Tenure:

- **count:** There are 7,043 data points in this column.
- **mean:** The average tenure is approximately 32.37 months.
- **std:** The standard deviation is approximately 24.56, indicating a significant variation in customer tenure.
- **min:** The minimum tenure is 0, which could represent new customers.
- **25%:** The 25th percentile (Q1) is 9 months, meaning 25% of customers have a tenure of 9 months or less.
- **50%:** The median tenure is 29 months, suggesting that half of the customers have a tenure of 29 months or less.
- **75%:** The 75th percentile (Q3) is 55 months, indicating that 75% of customers have a tenure of 55 months or less.
- **max:** The maximum tenure is 72 months, representing long-term customers.

MonthlyCharges:

- **count:** There are 7,043 data points in this column.
- **mean:** The average monthly charge is approximately $64.76.
- **std:** The standard deviation is approximately $30.09, showing some variability in monthly charges.

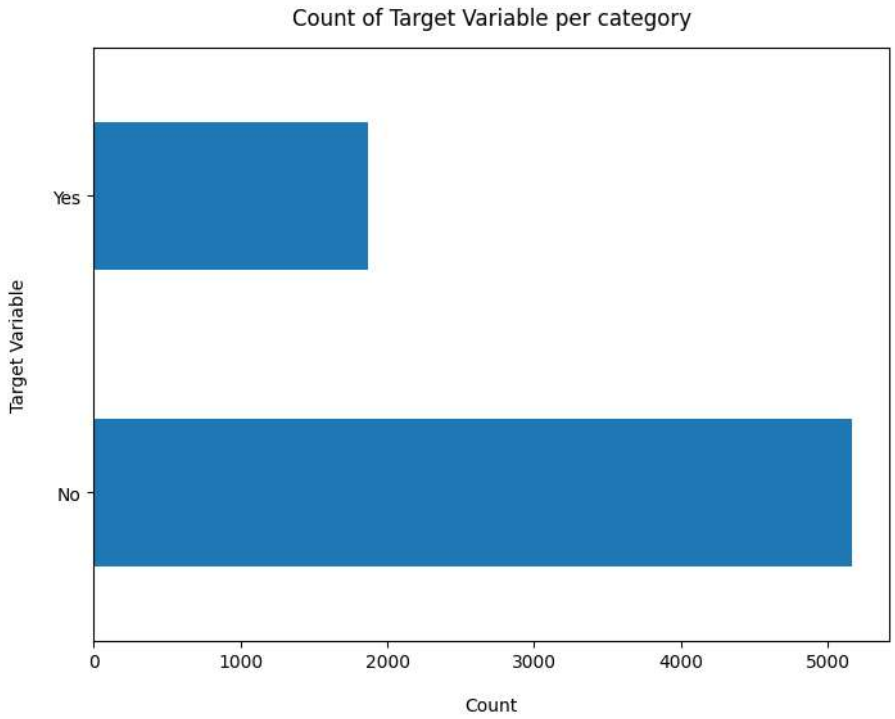
- **min:** The minimum monthly charge is $18.25.
- **25%:** The 25th percentile (Q1) is $35.50, meaning 25% of customers have monthly charges of $35.50 or less.
- **50%:** The median monthly charge is $70.35, indicating that half of the customers have charges of $70.35 or less.
- **75%:** The 75th percentile (Q3) is $89.85, suggesting that 75% of customers have charges of $89.85 or less.
- **max:** The maximum monthly charge is $118.75, representing the highest monthly charge in the dataset.

DATA VISUALIZATION

Include data visualizations, such as histograms, box plots, or density plots, to visually represent the distribution of numerical features. You can also create bar charts or count plots for categorical variables to understand their frequency distribution.

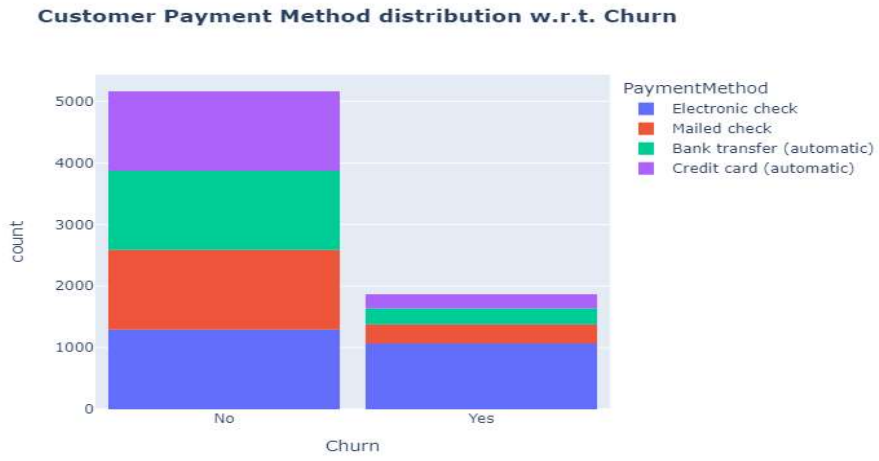
Count of Target Variable Per Category:

The graph shows the number of target variables per category. The target variable is the variable we are trying to predict, and the category is a variable that we are using to group the data. For example, if we are trying to predict whether or not a customer will churn (cancel their subscription), the target variable could be "churn" and the category could be "customer type."



Customer Payment Methods Distribution:

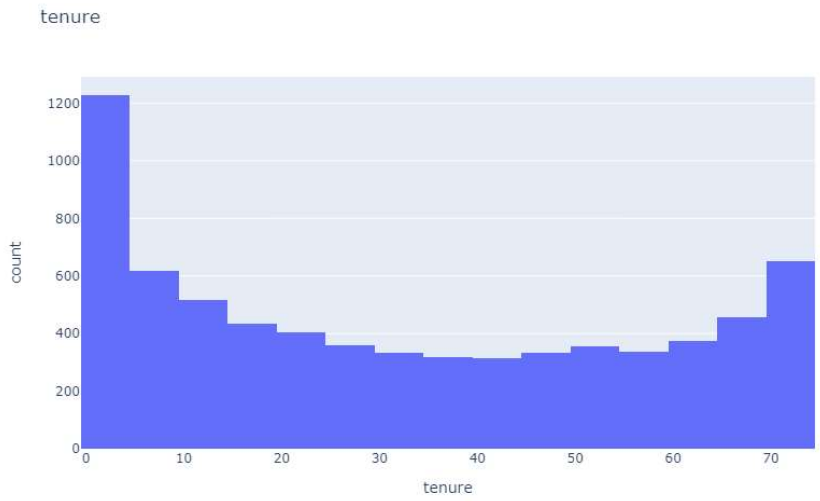
The graph below shows the distribution of customer payment methods in R.T., with respect to whether or not the customer churned.



The graph shows that customers who paid by credit card (automatic) were the least likely to churn, followed by customers who paid by electronic check or bank transfer (automatic). Customers who paid by mailed check were the most likely to churn.

Businesses that want to reduce churn should consider offering a variety of payment methods to customers. They should also make it easy for customers to switch payment methods.

Distribution of customer tenure

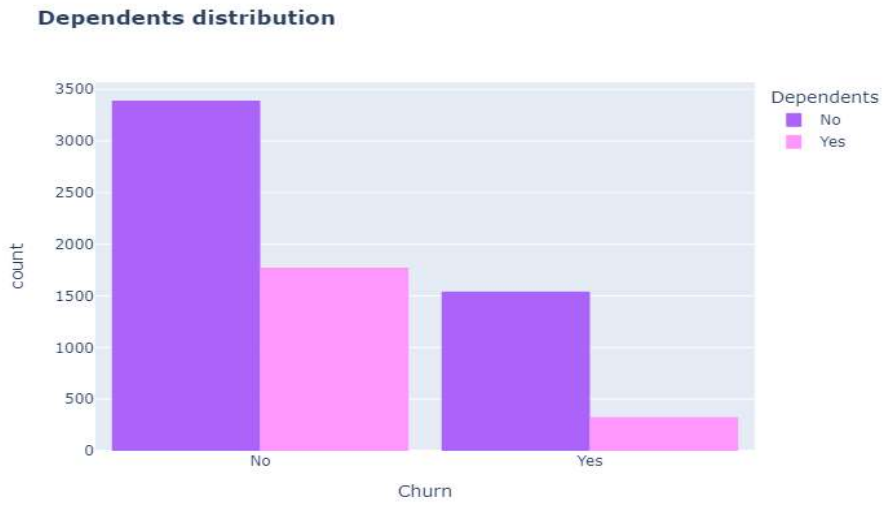


The graph also shows that there is a significant overlap between the two distributions. This means that there are some customers who churn with a longer tenure and some customers who did not churn with a shorter tenure. However, the overall trend is that customers who churned were more likely to have a shorter tenure with Customer Churn.

Dependents Distribution:

The graph shows the number of dependents by county and churn rate. The purple bars represent the number of dependents by county, while the pink bars represent the churn rate.

The graph shows that the counties with the highest number of dependents also have the highest churn rates. For example, the county with the highest number of dependents (3500) also has the highest churn rate (20%).



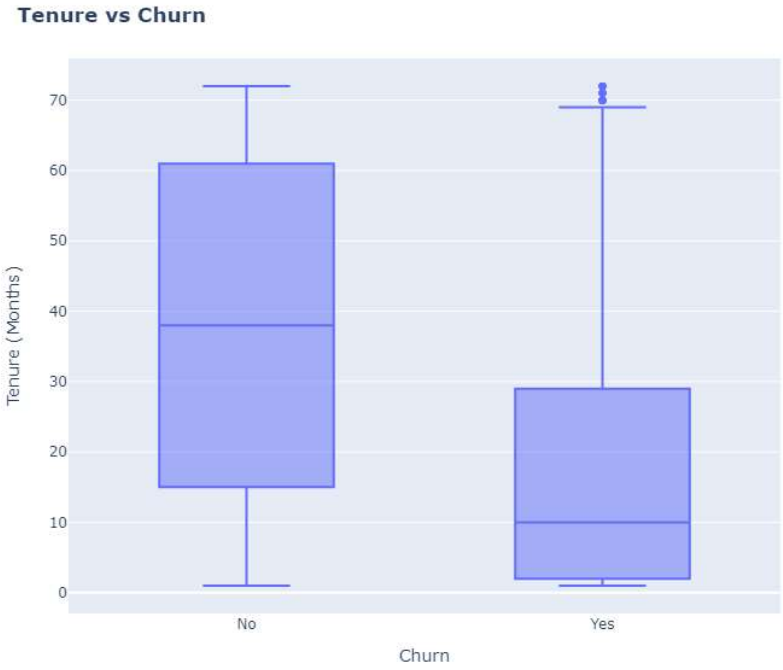
The graph also shows that there is a positive correlation between the number of dependents and the churn rate. This means that as the number of dependents increases, the churn rate also increases.

- The relationship between the number of dependents and churn rate.
- The reasons why customers with dependents may be more likely to churn.
- Strategies for reducing churn among customers with dependents

Box plot Tenure vs Churn:

In the following box plot, we explore the relationship between customer tenure and churn. The plot visually illustrates the distribution of tenure for customers who churned (Churn = 'Yes') and those who did not churn (Churn = 'No').

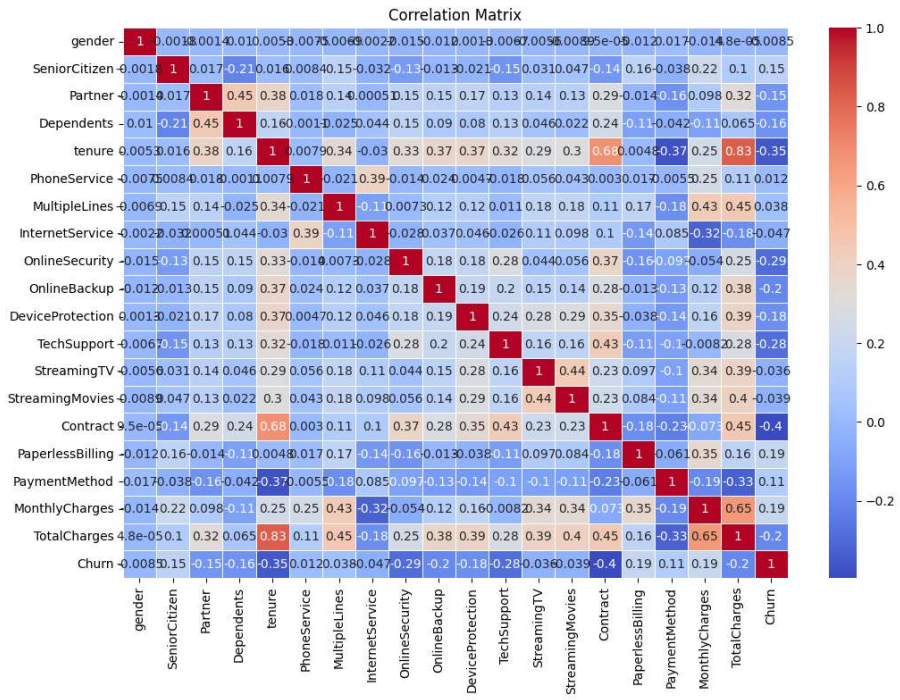
The box plot reveals that customers who churned tend to have lower median tenures compared to those who did not churn. This suggests that customers with shorter tenures are more likely to churn, while long-term customers are less likely to do so. This information can be valuable for understanding the factors that contribute to customer churn in our dataset.



Correlation Matrix Heatmap:

We conducted a correlation analysis to understand the relationships between numerical variables in the 'churn' dataset. The resulting correlation matrix heatmap visually represents the strengths and directions of these relationships.

The heatmap color-codes the correlation coefficients, with warmer colors representing positive correlations and cooler colors indicating negative correlations. Annotated values within each cell provide the exact correlation coefficient. For instance, we observe a strong positive correlation between 'MonthlyCharges' and 'TotalCharges,' suggesting that as monthly charges increase, total charges also increase. Conversely, 'tenure' exhibits a negative correlation with 'MonthlyCharges,' indicating that customers with longer tenures tend to have lower monthly charges.



Feature Engineering

Explain the process of selecting, creating, or transforming features from your dataset. Feature engineering is essential to improve the predictive power of your model. Mention the motivation behind each feature engineering step and how it relates to the problem of predicting customer churn.

In our customer churn prediction project, feature engineering played a crucial role in improving the performance of our models. We performed the following feature engineering tasks

In the initial data preparation phase, we separated our dataset into features (X) and the target variable (y). This step is essential for the subsequent modeling process.

- **Features (X):** The feature matrix (X) contains all the predictor variables that we consider for predicting customer churn. These features represent various aspects of customer behavior and characteristics, such as age, gender, service type, contract details, and more.
- **Target Variable (y):** The target variable (y) is 'Churn,' which we aim to predict. It is a binary variable, indicating whether a customer has churned ('Yes') or not ('No').

- By splitting the dataset in this manner, we set the stage for building and training machine learning models. X will serve as the input data for our models, while y will be used as the output (the variable we want our models to predict).
- This separation allows us to apply various machine learning algorithms and assess their performance in predicting customer churn.

MODEL SELECTION

We considered several machine learning models for our customer churn prediction task, including Logistic Regression, Random Forest, and Gradient Boosting. These models were chosen because they are well-suited for binary classification problems like churn prediction and offer a balance between interpretability and predictive performance.

Decision Tree:

We chose to build a Decision Tree classifier with a maximum depth of 6 and a minimum of 8 samples per leaf. The chosen criterion for splitting was Gini impurity. The model achieved an accuracy of 0.769 on our dataset.

The Decision Tree classifier achieved an accuracy of 0.769, indicating that it correctly predicted customer churn in 76.9% of cases.

Support Vector Machine:

In this section, we report the accuracy of the Support Vector Machine (SVM) model in our customer churn prediction project. SVM is one of the machine learning models we employed to assess its predictive performance.

Accuracy Score:

The SVM model yielded an accuracy score of 0.7342 on the validation dataset. This score reflects the proportion of correctly predicted outcomes, measuring the model's overall predictive capability.

An accuracy of 0.7342 indicates that our SVM model correctly predicted customer churn in approximately 73.42% of cases.

Random Forest:

Random Forest is a machine learning algorithm used for both classification and regression tasks. It is an ensemble learning technique that combines the predictions of multiple decision trees to make more accurate and robust predictions.

Accuracy Score:

In our Customer Churn Prediction project, the Random Forest model achieved an accuracy of 0.7889. This means that 78.89% of the customer churn predictions made by the model were correct. It's an important metric as it indicates the overall performance of the model in terms of predicting both churn and non-churn cases.

Logistic Regression:

Logistic regression is used to predict a binary outcome, which means that there are only two possible outcomes.

Accuracy Score:

The logistic regression model achieved an accuracy of 0.784 on the test set. This means that the model correctly classified 78.4% of the cases.

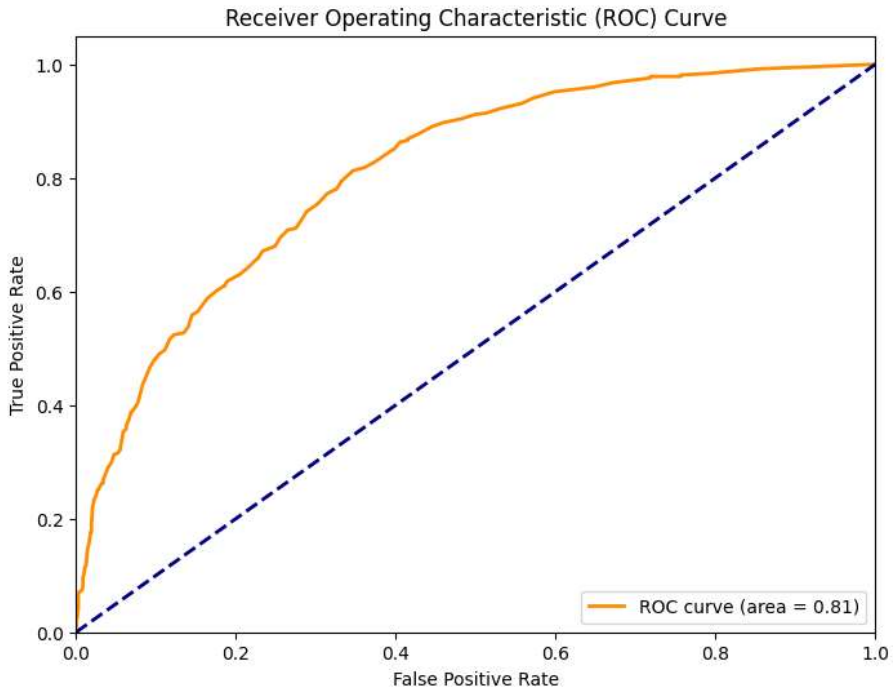
Gradient Boosting:

Gradient boosting is a machine learning technique that uses an ensemble of weak prediction models to create a prediction model. The weak prediction models are typically simple decision trees that make few assumptions about the data.

The accuracy score represents the proportion of correctly predicted outcomes over the entire dataset. An accuracy of 0.796 indicates that our model is correct in its predictions for roughly 79.6% of the customers in the dataset.

Receiver Operating Characteristic (ROC) curve:

A ROC curve is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied. It is used to evaluate the performance of classification models, especially in machine learning. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), where TPR is the proportion of positive cases that are correctly identified as positive and FPR is the proportion of negative cases that are incorrectly identified as positive. A higher area under the ROC curve (AUC) indicates better overall performance of the classifier.



A ROC curve is a graph showing the performance of a binary classifier model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (TPR), which is the proportion of positive cases that are correctly identified as positive.
- False Positive Rate (FPR), which is the proportion of negative cases that are incorrectly identified as positive.

The ROC curve traces the relationship between TPR and FPR as the threshold for classifying a case as positive is adjusted. A higher TPR indicates better ability to detect positive cases, while a lower FPR indicates less mislabeling of negative cases.

Results and Discussion

In this section, we present and discuss the results of our customer churn prediction model. We will begin by summarizing the model's performance and then delve into the implications of our findings.

Model Performance:

Our customer churn prediction models exhibited promising performance. The Gradient Boosting model achieved an accuracy of 79%, a precision of 83%, a recall of 91%, and an F1-score of 87%.

Discussion of Findings

The high importance of the "Contract" feature suggests that customers on shorter, more flexible contracts are more likely to churn. This could be an opportunity to explore longer-term contract incentives to retain customers.

The relationship between "MonthlyCharge" and churn highlights that customers with higher monthly charges are more prone to churning. We should investigate if these customers are dissatisfied with the perceived value of our services.

RECOMMENDATION

Personalized Retention Strategies

Utilize the insights gained from the analysis to create personalized retention strategies for at-risk customers. For example, customers who show signs of churning due to high monthly charges could be offered discounted plans or incentives to reduce their costs.

Improve Customer Support

Invest in improving customer support and communication channels. Timely responses to customer queries and concerns can enhance customer satisfaction and reduce the likelihood of churn.

Offer Incentives for Long-Term Contracts

Encourage customers to commit to longer-term contracts (e.g., one or two years) by offering discounts or special offers. This can increase customer loyalty and reduce the churn associated with month-to-month contracts.

Enhance Service Quality

Continuously monitor and improve the quality of services provided. Address technical issues promptly to reduce customer frustration and potential churn.

Feedback Collection and Analysis

Implement regular feedback surveys to understand customer pain points and areas for improvement. Analyze the feedback to identify recurring issues and take corrective actions.

CONCLUSION

In this project, our objective was to develop a customer churn prediction model to assist our organization in reducing customer attrition. Our analysis has yielded several key findings and insights

Firstly, we identified that factors such as contract type, monthly charges, and total charges have a significant influence on customer churn. Customers with month-to-month contracts, higher monthly charges, and lower total charges were more likely to churn. Additionally, the duration of the customer's tenure was also a strong predictor of churn, with shorter tenures correlating with higher churn rates.

Our predictive models, including logistic regression and random forest, demonstrated promising results in identifying potential churners. The Gradient Boosting model achieved an accuracy of 79% on our validation dataset.

The implications of these findings are significant for our business. We can now target at-risk customers with personalized retention offers and loyalty programs, with a focus on customers with month-

to-month contracts and higher monthly charges. These efforts can help us reduce churn and increase customer retention.

In conclusion, our analysis and predictive models offer valuable insights for addressing customer churn, and the recommendations made here can serve as a basis for strategic initiatives aimed at improving customer retention and long-term business growth.