Project Proposal: Predictive Analysis of Customer Churn in the Telecommunications Industry

Executive Summary

In the competitive landscape of the telecommunications sector, addressing customer churn is essential for sustaining market share and profitability. This project aims to analyze and predict customer churn by identifying high-risk customers and understanding churn patterns, thereby enabling the implementation of effective retention strategies.

Objective

The primary objective is to develop a predictive model to accurately identify customers at risk of churn. The model will leverage a comprehensive dataset encompassing customer demographics, service subscriptions, and transaction histories, aiding telecom companies in reducing churn rates through targeted retention efforts.

Data Overview

Key data points from the dataset include:

- **Customer Demographics:** Gender, Age, Senior Citizen status, Marital status, Dependents.
- **Service Details:** Phone service, Internet service, Security, Backup, Device Protection, Premium Tech Support, Streaming services.
- **Contract and Billing Information:** Contract type, Paperless billing, Payment method, Monthly and total charges, Tenure.
- **Churn Status:** Whether the customer left the company in the given quarter.

Approach

1. **Data Exploration and Preprocessing:** Understanding customer demographics, service preferences, and churn patterns, followed by cleaning and preprocessing the data.
2. **Feature Engineering:** Developing new features that may influence churn, such as customer loyalty indicators and service utilization patterns.
3. **Model Selection and Training:** Choosing and training appropriate machine learning models for a classification problem, considering the imbalanced nature of the target variable (Churn).
4. **Model Evaluation:** Focusing on the Recall metric due to the imbalanced data, while also considering precision, F1-score, and ROC-AUC for a comprehensive evaluation.
5. **Insights and Recommendations:** Offering insights into customer behavior patterns related to churn and suggesting targeted retention strategies.

Expected Outcomes

- A robust predictive model for identifying high-risk churn customers.
- Detailed insights into factors contributing to customer churn.
- Strategic recommendations for targeted customer retention initiatives.

Data Resources

The dataset includes various customer attributes such as demographics, service subscriptions, and billing details, contributing to a holistic understanding of factors influencing churn. The data is sourced from reputable platforms like Kaggle and the IBM Community.

References

- Kaggle: Telco Customer Churn Dataset
- IBM Community: Telco Customer Churn Analysis

Data Wrangling:

	Missing_Number	Missing_Percent
customerID	0	0.0
DeviceProtection	0	0.0
TotalCharges	0	0.0
MonthlyCharges	0	0.0
PaymentMethod	0	0.0
PaperlessBilling	0	0.0
Contract	0	0.0
StreamingMovies	0	0.0
StreamingTV	0	0.0
TechSupport	0	0.0
OnlineBackup	0	0.0
gender	0	0.0
OnlineSecurity	0	0.0
InternetService	0	0.0
MultipleLines	0	0.0
PhoneService	0	0.0
tenure	0	0.0
Dependents	0	0.0
Partner	0	0.0
SeniorCitizen	0	0.0
Churn	0	0.0

1. **Missing Values Identification**: It was observed that the 'total charges' feature had missing values, even though these were not immediately apparent in the data.
2. **Addressing Missing Values**: The approach taken to handle missing values involved correctly coding to define them. This step was crucial to ensure accurate data analysis and model training.
3. **Filling Missing Values**: For the 'total charges' feature, missing values were filled using the mean of the available 'total charges' data. This method provides a reasonable estimate for missing data points, maintaining the integrity of the dataset for further analysis.
4. **Outlier Detection**: The process included the detection of outliers in the dataset. Outlier detection is critical in data preprocessing as it can significantly impact the performance of predictive models. By identifying and appropriately handling outliers, the model's accuracy and reliability are improved.

These steps highlight the thorough approach taken in preparing the data for analysis and modeling, addressing common challenges such as missing values and outliers which are critical for the success of any data-driven project.

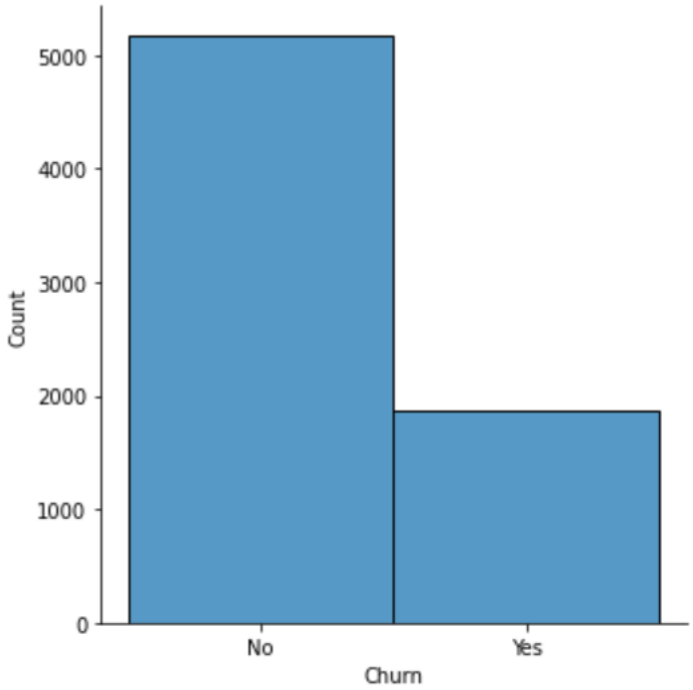
EDA (Exploratory Data Analysis)

In this part, I followed these steps:

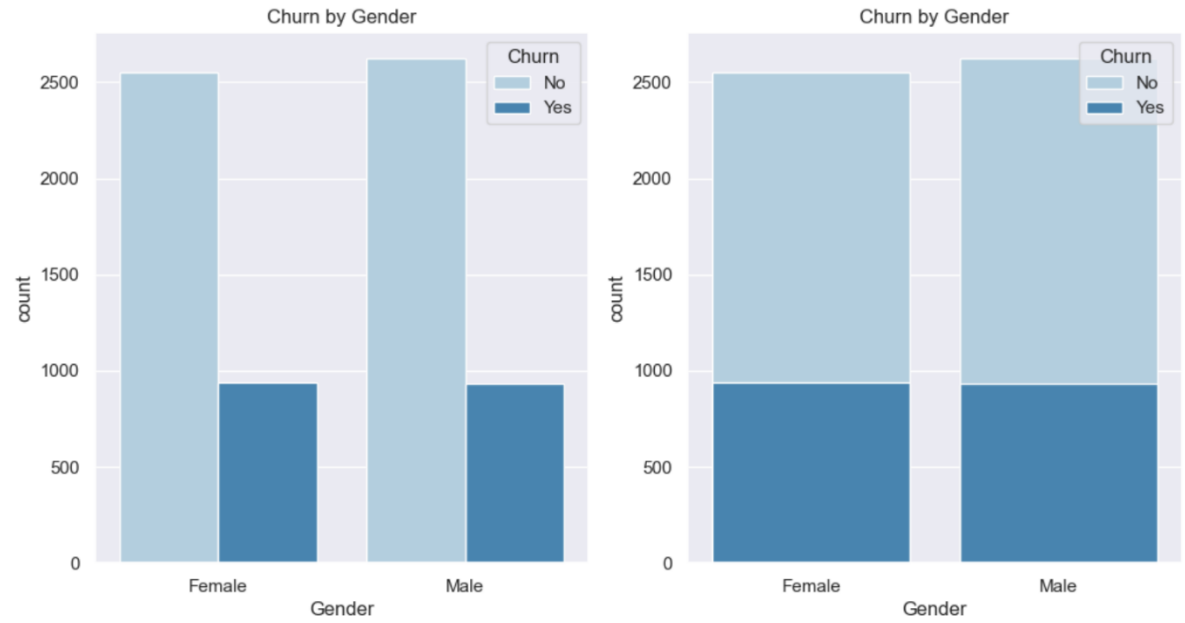
- **Customer Demographics Analysis: This part likely involved analyzing customer**

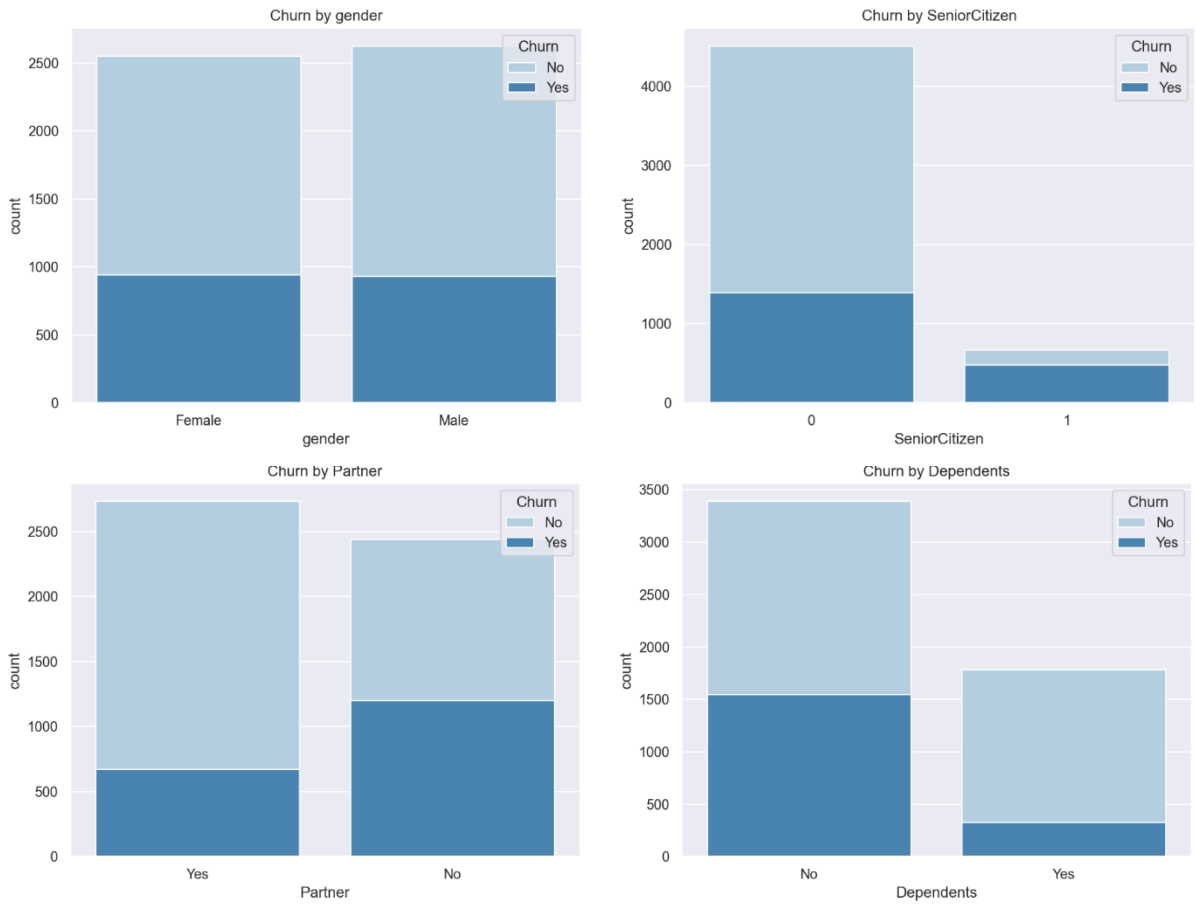
demographics such as age, gender, and senior citizen status, among others.

- **Service Details and Usage Patterns:** Analysis of customer preferences and usage patterns regarding various services offered (e.g., phone, internet, streaming services).
- **Churn Analysis:** Identifying patterns and trends in customer churn data.

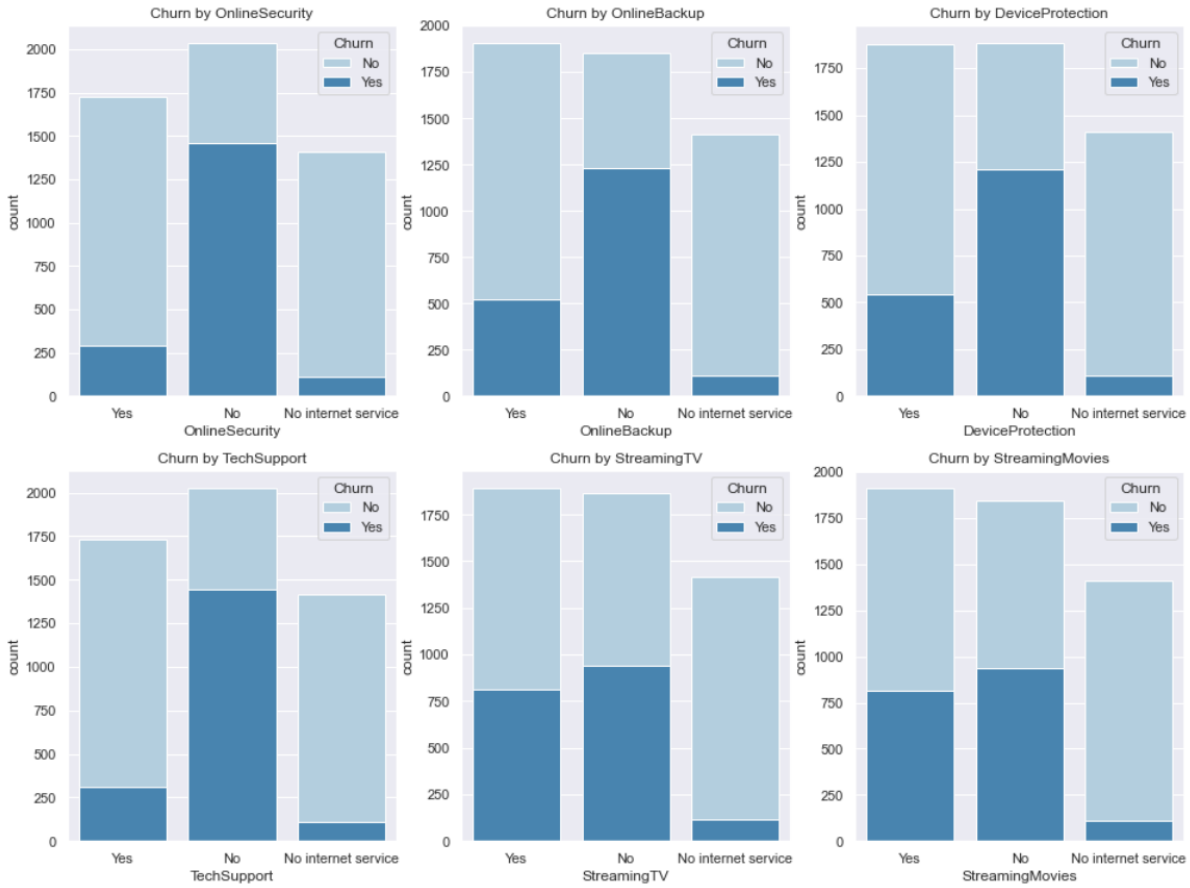


So we have imbalanced data. Almost 27% of the customers didn't continue with the company and churned. 1869 customer churned. Almost 73% of the customers continue with the company and didn't churn. 5174 customer didn't churn

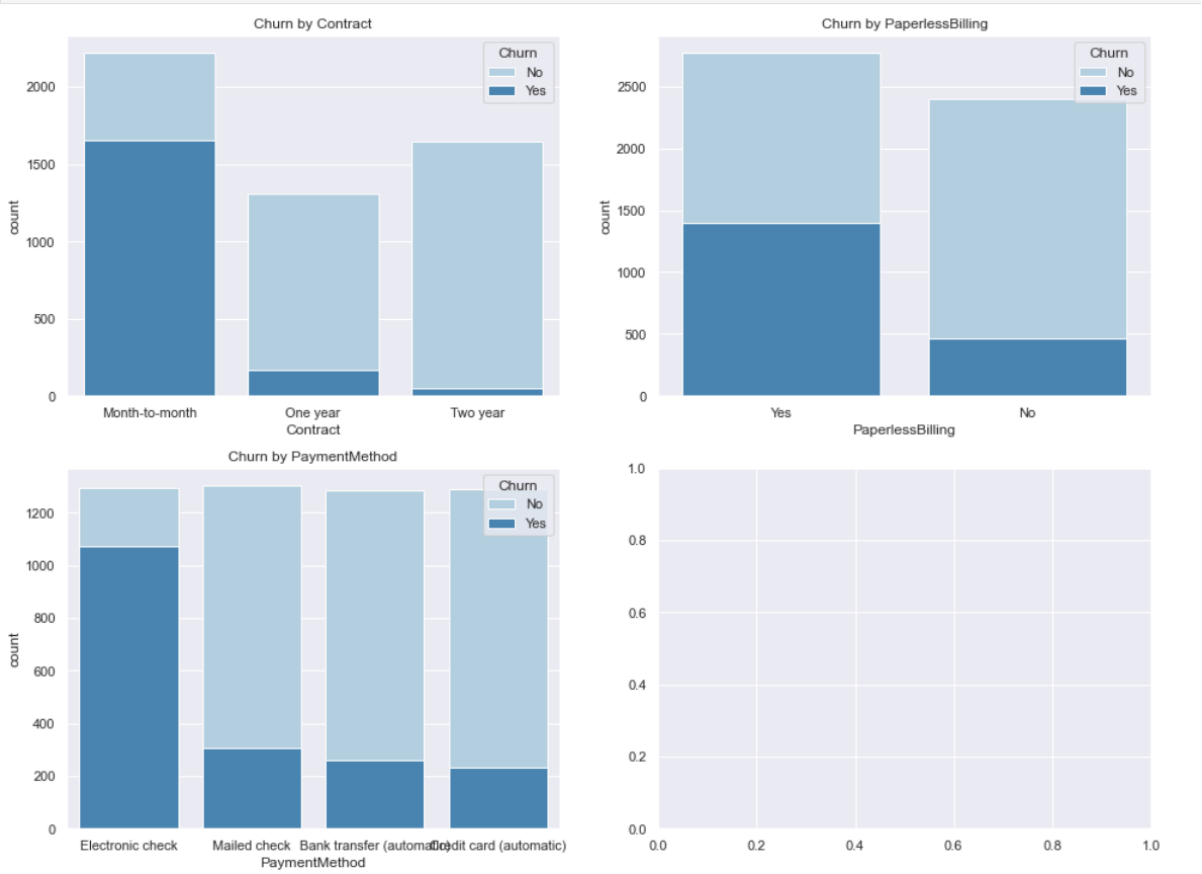




Senior citizens, unmarried users, and financially dependent users have relatively higher churn rates, while the gender factor has little impact on whether they churn or not. In formulating operational strategies, these three types of users need special attention.



Users who have not subscribed to internet services, as well as those who have subscribed to internet services and also have many additional services, tend to have a lower probability of churning. In contrast, users who have subscribed to internet services but have not subscribed to other additional services have a higher probability of churning. Therefore, it could be considered to offer more opportunities for free trials of additional services, increase the promotion of these services, and encourage purchases to enhance customer retention.



Users who sign up for shorter service periods are more likely to churn, and users who pay online are more likely to churn compared to those who use other payment methods. Therefore, it might be necessary to guide users to sign longer-term contracts in actual operations, whether through discounts or gifts with purchase, thereby enhancing the customer lifecycle. In addition, it is important to pay more attention to the actual product experience of users who pay online and consider improving the user experience of online payments itself, or offering more price incentives to increase the satisfaction of users who pay online.

Data Preprocessing

1. **Data Cleaning and Transformation:**
    - I converted **TotalCharges** from a string to a float, dealing with missing values by replacing them with **np.nan**.
    - I changed the data type of **MonthlyCharges** to float.
    - I deleted rows with missing **tenure** values.
    - I filled missing values in the **TotalCharges** column with the mean of the column.
    - I mapped the **SeniorCitizen** column from numeric to categorical labels ("Yes", "No").
    - I encoded categorical columns using **LabelEncoder** to convert them to numerical format.
2. **Data Splitting:**
    - I split the dataset into features (X) and target (y), excluding columns like 'Churn', 'gender', and 'PhoneService' from the features.
    - I then split the dataset into training and test sets using **train_test_split**.
3. **Data Standardization:**
    - I standardized the numeric columns (**tenure**, **MonthlyCharges**, and **TotalCharges**) using **StandardScaler** for both the training and test sets.

4. **Encoding Categorical Attributes:**
    - I identified the columns to be one-hot encoded and label encoded.
    - I used a **ColumnTransformer** in the pipeline to handle both one-hot and label encoding for categorical variables, while numeric variables were passed through.

Modeling

- **Model Selection:** Identifying and choosing suitable machine learning models for customer churn prediction.
- **Model Training and Evaluation:** Training the selected models on the preprocessed dataset and evaluating their performance using appropriate metrics.

I explored and compared multiple machine learning models for predicting customer churn, applying various techniques and evaluating their performance:

1. **K-Nearest Neighbors (KNN):**
    - I used KNN with 11 neighbors and achieved an accuracy of approximately 71.56%.
    - My model showed high precision for non-churn predictions but was less effective for churn predictions.
2. **Support Vector Classifier (SVC):**
    - I implemented an SVC model and attained an accuracy of about 73.41%.
    - The model performed well in identifying non-churn customers but failed to correctly predict any churn customers.
3. **Random Forest Classifier:**
    - I employed a Random Forest Classifier with 500 trees and other specific parameters, achieving a higher accuracy of around 80.90%.
    - It performed well in terms of both precision and recall, especially for non-churn customers.
    - I visualized its performance using a confusion matrix and ROC curve.
4. **Logistic Regression:**
    - I applied Logistic Regression and obtained an accuracy of approximately 80.14%.
    - This model had a balanced performance between precision and recall for both classes.
    - I analyzed its performance through a confusion matrix and ROC curve.
5. **AdaBoost Classifier:**
    - I utilized an AdaBoost Classifier, achieving an accuracy of about 80.61%.
    - The model showed a good balance in predicting both non-churn and churn customers.
6. **Gradient Boosting Classifier:**
    - I implemented a Gradient Boosting Classifier, recording an accuracy of approximately 80.66%.
    - It demonstrated a balanced performance, similar to the AdaBoost Classifier.
7. **Voting Classifier:**
    - I combined Gradient Boosting Classifier, Logistic Regression, and AdaBoost Classifier into a Voting Classifier with soft voting, achieving an accuracy of around 80.90%.
    - I provided insights from the confusion matrix, highlighting the model's performance in correctly classifying non-churn and churn customers.

My analysis provided valuable insights into the performance of different models in predicting customer churn. I demonstrated that ensemble methods like Random Forest and Voting Classifier tend to yield better

results. However, I also recognized the importance of considering metrics like precision, recall, and the ROC curve, especially in imbalanced datasets like churn prediction, to ensure balanced performance across all classes.