Project Explanation

Shravan Sundar Ravi

2024-12-08

Customer Churn

In the competitive telecommunications industry, customer churn poses a significant challenge for service providers, impacting revenue and growth potential. This project analyzes customer churn using the Iranian Churn Dataset, which contains 3,150 records collected from an Iranian telecom company's database over 12 months. The dataset includes various attributes such as call failures, SMS frequency, customer complaints, subscription length, and customer value, offering insights into customer behavior. By applying supervised machine learning techniques, this project aims to develop a predictive model that identifies customers at risk of leaving the service. Understanding these predictive factors is crucial for enhancing customer retention strategies, optimizing marketing efforts, and ultimately fostering long-term customer loyalty.

Topic Description

This project examines customer churn in the telecommunications sector using the Iranian Churn Dataset. I am particularly interested in this topic due to the significant impact of customer retention on a company's profitability in a competitive market. The dataset features various attributes such as call failures, complaints, and subscription length, which provide valuable insights into customer behavior. By analyzing these factors, I aim to identify patterns that can inform strategies to enhance customer satisfaction and loyalty, ultimately benefiting both the telecom company and its clients.

Expectations

I expect to uncover key relationships in the dataset, such as a correlation between a higher number of complaints and call failures with increased churn rates, indicating that service quality is crucial for customer retention. Additionally, I hypothesize that longer subscription lengths will be linked to lower churn rates, suggesting that established customers are more loyal. I also anticipate that usage frequency and distinct calls made will reveal insights about customer engagement, with higher usage likely reducing churn. Overall, I aim to identify predictive factors that clarify why customers leave, offering actionable insights for improving retention strategies.

Data Description

Source: The dataset is sourced from the UCI Machine Learning Repository, specifically the Iranian Churn Dataset, Link : https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset

Collection Method: This dataset was randomly collected from an Iranian telecom company's database over a period of 12 months.

Number of Cases: There are a total of 3,150 rows, each representing a customer.

Attribute Descriptions

1. **Call Failure**: The number of times calls attempted by the customer failed, which can indicate service quality issues and potentially influence customer satisfaction.
2. **Complaints**: The total number of complaints registered by the customer, serving as a direct measure of customer dissatisfaction and a potential predictor of churn.
3. **Subscription Length**: The duration of the customer's subscription in months, reflecting customer loyalty and engagement; longer subscriptions may correlate with lower churn rates.
4. **Charge Amount**: The total amount charged to the customer, which could impact their perception of value and satisfaction with the service.
5. **Seconds of Use**: The total time, measured in seconds, that the customer has utilized the service, which may correlate with customer engagement and likelihood of churn.
6. **Frequency of Use**: How often the customer uses the service, indicating their level of engagement and satisfaction.
7. **Frequency of SMS**: The number of SMS messages sent by the customer, which can be an indicator of customer engagement with the service.
8. **Distinct Called Numbers**: The number of different phone numbers the customer has called, providing insight into their social connectivity and usage patterns.
9. **Age Group**: The age category of the customer, which may influence service preferences and churn behavior.
10. **Tariff Plan**: The type of service plan the customer is subscribed to (e.g., prepaid, postpaid), which can affect pricing and churn rates.
11. **Status**: The current status of the customer (active or inactive), which is crucial for determining churn.
12. **Age**: The actual age of the customer, providing demographic insights that may correlate with churn.
13. **Customer Value**: An aggregated metric representing the overall value of the customer to the company, reflecting their potential lifetime revenue.
14. **Churn**: The target variable indicating whether the customer has churned (1) or not (0), which is the main focus of the analysis.

```
import pandas as pd
df = pd.read_csv("Customer Churn.csv")
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
import numpy as np
```

Summary Statistics

```
df.columns = df.columns.str.strip()
```

```
summary_stats = df[['Call  Failure', 'Complains', 'Subscription  Length',
                    'Charge  Amount', 'Seconds of Use', 'Frequency of use',
                    'Frequency of SMS', 'Distinct Called Numbers',
                    'Age Group', 'Tariff Plan', 'Status', 'Age',
                    'Customer Value', 'Churn']].describe()
```

```
print(summary_stats)
```

```
##        Call  Failure    Complains  ...  Customer Value        Churn
## count    3150.000000  3150.000000  ...     3150.000000  3150.000000
## mean        7.627937     0.076508  ...      470.972916     0.157143
## std         7.263886     0.265851  ...      517.015433     0.363993
## min         0.000000     0.000000  ...        0.000000     0.000000
## 25%         1.000000     0.000000  ...      113.801250     0.000000
## 50%         6.000000     0.000000  ...      228.480000     0.000000
## 75%        12.000000     0.000000  ...      788.388750     0.000000
## max        36.000000     1.000000  ...     2165.280000     1.000000
##
## [8 rows x 14 columns]
```

Exploratory Data Analysis

Main Outcome / Target Variable

- **Target Variable**: The main target variable for prediction is **Churn** (1 for churned customers and 0 for retained).
- **Effectiveness**: This variable accurately represents the customer's status, making it suitable for predicting churn likelihood. Its binary nature allows for a straightforward classification model, ideal for exploring factors influencing churn.

No Data Cleaning Was Necessary

New Variables Created

- **Average Monthly Charges**: Calculated by dividing **Charge Amount** by **Subscription Length** to capture monthly spend, potentially indicating value for money perception.

```
#Average Monthly Charge
df['Avg Monthly Charge'] = df['Charge  Amount'] / df['Subscription  Length']
```

- **Complaint Ratio**: Derived by dividing **Complaints** by **Subscription Length** to capture complaint frequency, which may correlate with dissatisfaction.

```
# Complaint Ratio
df['Complaint Ratio'] = df['Complains'] / df['Subscription  Length']
```

Post-Split Procedures

- **Scaling**: Planning to apply scaling (e.g., MinMax or StandardScaler) on the training data for numerical features to improve model stability and accuracy.
- **Feature Engineering and Selection**: Will use training data for feature engineering, including feature importance analysis to fine-tune predictive power.

Excluding Observations

- **Exclusions**: No specific observations are excluded at this point. However, if extreme outliers are discovered in further analysis, they might be flagged or removed for more accurate modeling.

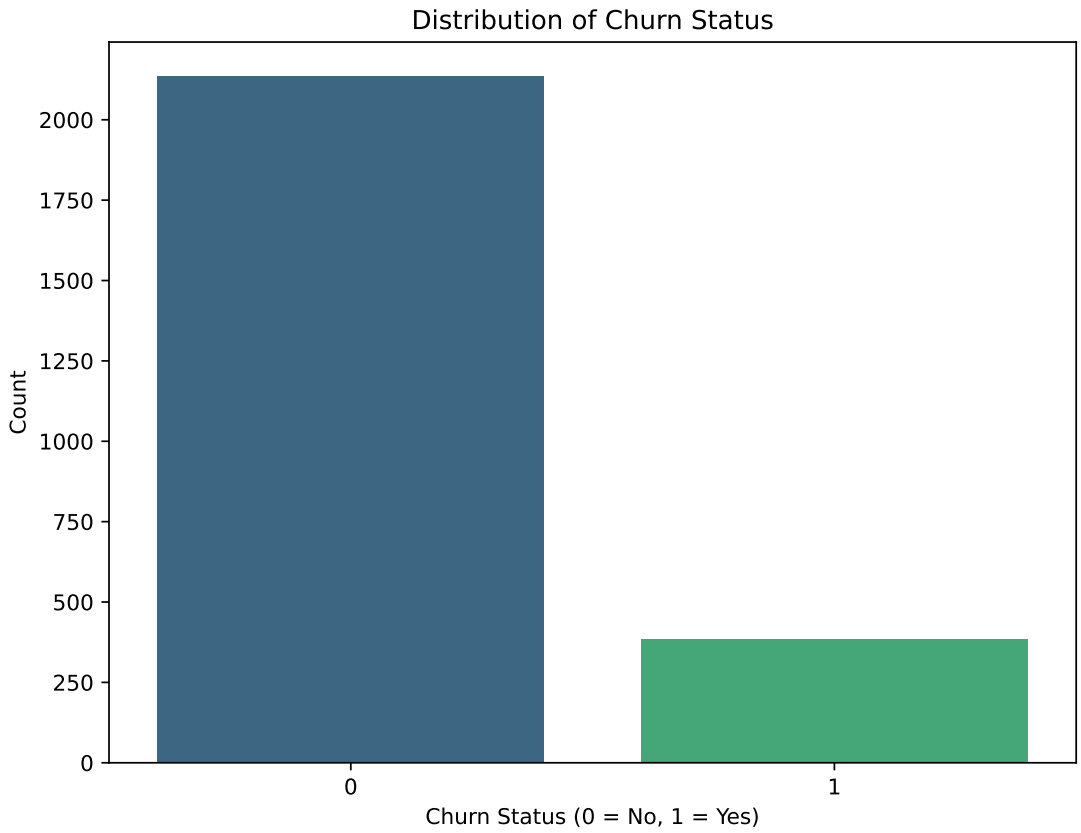
```
# Split data into training and testing sets (80% train, 20% test)
train_df, test_df = train_test_split(df, test_size=0.2, random_state=42)
```

Data Visualizations

Visualization 1: Distribution of Churn Status

- **Type**: Bar Plot
- **Insight**: Visualizing churn vs. non-churn percentages will provide a baseline understanding of class distribution. If there's an imbalance (e.g., significantly more non-churn than churn cases), this may impact model choice and evaluation strategies.

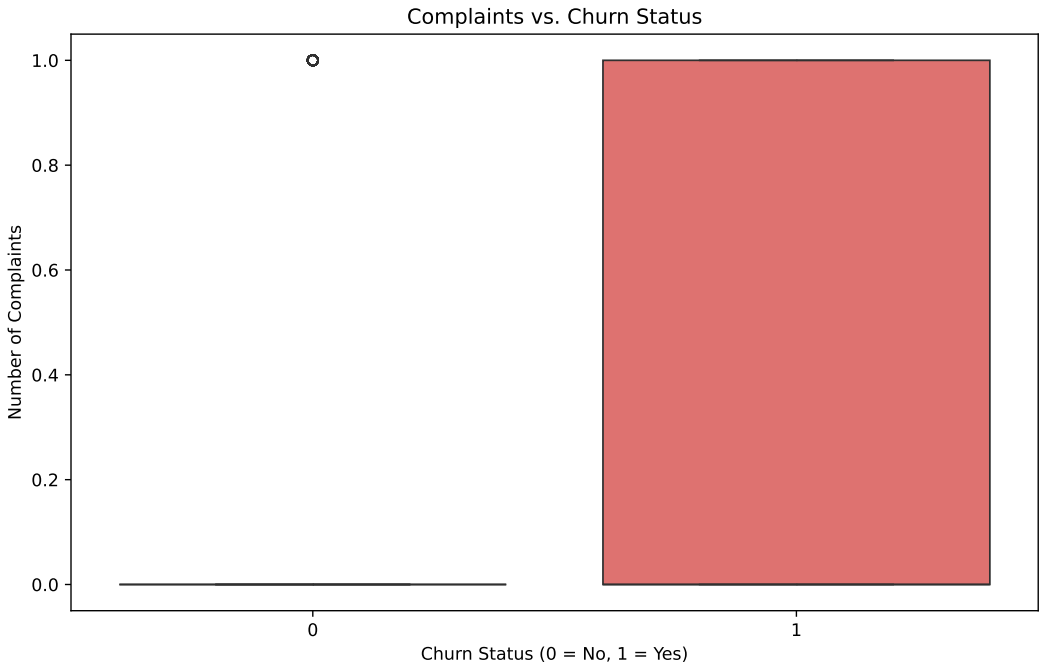
```
# Plot distribution of churn status
plt.figure(figsize=(8, 6))
sns.countplot(data=train_df, x='Churn', hue='Churn', palette='viridis', legend=False)
plt.title('Distribution of Churn Status')
plt.xlabel('Churn Status (0 = No, 1 = Yes)')
plt.ylabel('Count')
plt.show()
```



Visualization 2: Complaints vs. Churn Status

- **Type**: Box Plot
- **Insight**: Comparing complaint frequencies across churn and non-churn groups may reveal if a higher complaint count correlates with higher churn rates. This insight could confirm complaints as a strong predictive feature.

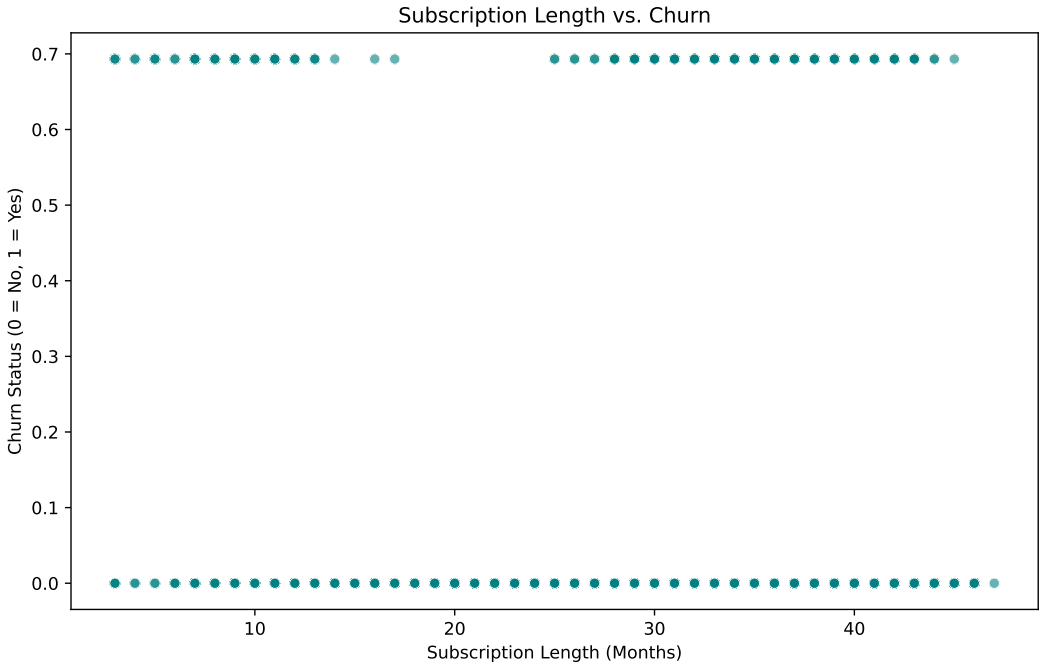
```
# Box plot of Complaints by Churn status
plt.figure(figsize=(10, 6))
sns.boxplot(data=train_df, x='Churn', y='Complains', hue='Churn', palette='magma', legend=False)
plt.title('Complaints vs. Churn Status')
plt.xlabel('Churn Status (0 = No, 1 = Yes)')
plt.ylabel('Number of Complaints')
plt.show()
```



Visualization 3: Relationship between Subscription Length and Churn

- **Type**: Scatter Plot
- **Insight**: This plot can show if longer subscription lengths correlate with lower churn rates, validating whether customer loyalty increases with time and could serve as a retention indicator.

```
# Scatter plot for Subscription Length vs. Churn
train_df['Churn_log'] = np.log1p(train_df['Churn'])
plt.figure(figsize=(10, 6))
sns.scatterplot(data=train_df, x='Subscription  Length', y='Churn_log', alpha=0.6, color="teal")
plt.title('Subscription Length vs. Churn')
plt.xlabel('Subscription Length (Months)')
plt.ylabel('Churn Status (0 = No, 1 = Yes)')
plt.show()
```

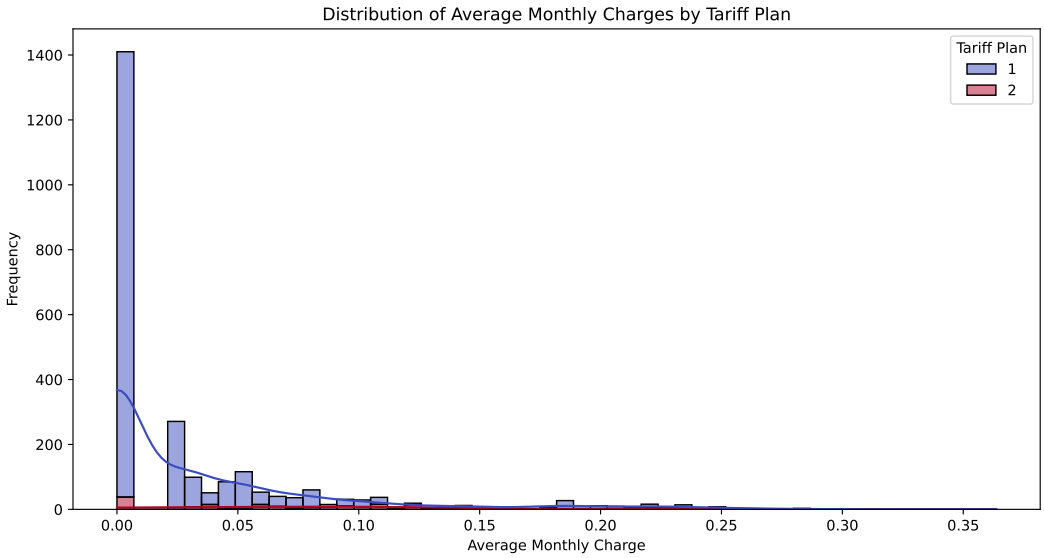


Visualization 4: Average Monthly Charges Distribution by Tariff Plan

- **Type**: Histogram
- **Insight**: Examining monthly charges across different tariff plans could identify which pricing structures lead to higher churn, offering insight into which plans might be less favorable.

```
# Create Avg Monthly Charge variable
train_df['Avg Monthly Charge'] = train_df['Charge  Amount'] / train_df['Subscription  Length']

# Plot histogram of Avg Monthly Charge by Tariff Plan
plt.figure(figsize=(12, 6))
sns.histplot(data=train_df, x='Avg Monthly Charge', hue='Tariff Plan', multiple="stack", palette='coolwa
plt.title('Distribution of Average Monthly Charges by Tariff Plan')
plt.xlabel('Average Monthly Charge')
plt.ylabel('Frequency')
plt.show()
```



Interpretation of the Plot

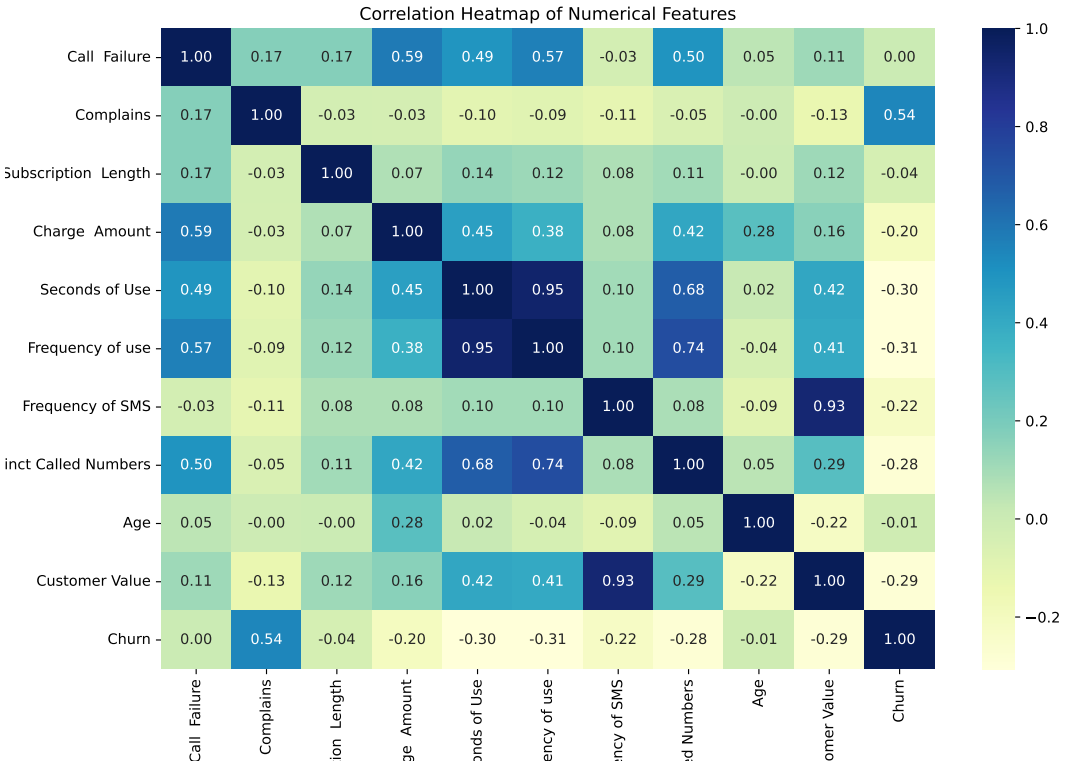
- **Skewed Distribution**: The distribution of average monthly charges is heavily right-skewed, with most customers having low average monthly charges, as shown by the high frequency of low values near 0.
- **Tariff Plan Comparison**:
  - **Tariff Plan 1 (Blue)**: Most of the customers fall under Tariff Plan 1, as evidenced by the higher frequency of charges in this range.
  - **Tariff Plan 2 (Pink)**: There are fewer customers on Tariff Plan 2, with a smaller contribution to each bin, particularly in the lower charge range.
- **KDE Lines**: The KDE lines for each tariff plan illustrate that while both plans generally have low average charges, there are subtle differences in their distributions. The KDE curve helps to smooth out the frequencies, providing an approximate trend of the charges for each tariff plan.

This plot effectively shows the distribution of monthly charges, highlighting that most customers incur relatively low monthly costs, with Tariff Plan 1 being more popular among customers than Tariff Plan 2.

Visualization 5: Correlation between Numerical Features and Churn
Type: Correlation Heatmap
Insight: This heatmap shows how each numerical feature correlates with `Churn`, helping to identify key predictors of churn. For example, a high positive correlation between `Complains` and `Churn` would suggest that customers who file more complaints are more likely to leave. Conversely, a strong negative correlation between `Subscription Length` and `Churn` could indicate that longer-term customers are more loyal and less likely to churn. The heatmap also helps identify any multicollinearity among features, such as between `Seconds of Use` and `Frequency of Use`, which could imply redundancy. This information is essential for feature selection and can improve model accuracy by focusing on the most influential variables for predicting churn.

```
# Calculate correlations for numerical columns
plt.figure(figsize=(12, 8))
correlation_matrix = train_df[['Call  Failure', 'Complains', 'Subscription  Length',
                               'Charge  Amount', 'Seconds of Use', 'Frequency of use',
                               'Frequency of SMS', 'Distinct Called Numbers', 'Age',
                               'Customer Value', 'Churn']].corr()
```

```
# Plot heatmap
sns.heatmap(correlation_matrix, annot=True, cmap="YlGnBu", fmt=".2f")
plt.title("Correlation Heatmap of Numerical Features")
plt.show()
```



The correlation heatmap provides several insights into the relationships between variables in our dataset.Here's a breakdown of the main observations:

- **Complains and Churn:** There is a moderately positive correlation (0.54) between `Complains` and `Churn`, suggesting that customers who file more complaints tend to have a higher likelihood of churning. This indicates that dissatisfaction with service could be a significant predictor of churn.
- **Seconds of Use and Frequency of Use:** These variables show a high positive correlation (0.95), which implies multicollinearity. Since both variables capture customer engagement with the service, you might consider keeping one of these as a feature in the final model to avoid redundancy.
- **Distinct Called Numbers and Frequency of Use:** Another notable correlation (0.74) exists here, implying that customers who use the service frequently tend to call a larger variety of numbers. This could indicate high engagement, which might reduce churn probability.
- **Age and Customer Value:** These variables have a weak negative correlation with `Churn`, which could imply that younger or less valuable customers may be slightly more prone to churn.
- **Other Weak Correlations with Churn:** The variables `Call Failure`, `Subscription Length`, and `Charge Amount`show weaker correlations with `Churn`. This suggests that they might have a limited predictive power individually, but could still provide valuable insights when combined with other variables.

Checking shapes

```
X = train_df.drop(columns=['Churn'])
y = train_df['Churn']

print("Shape of input features (X):", X.shape)
```

```
## Shape of input features (X): (2520, 16)
```

```
print("Shape of output variable (y):", y.shape)
```

```
## Shape of output variable (y): (2520,)
```

Machine Learning Approaches to Explore

- **Logistic Regression**: As a baseline model for binary classification, useful for interpreting coefficients and understanding feature impacts on churn.
- **Decision Trees**: Suitable for identifying key decision points influencing churn, easily visualizable for stakeholder communication.
- **Random Forest**: Offers improved predictive power by averaging multiple decision trees, potentially better at handling imbalanced classes.
- **Support Vector Machines (SVM)**: Good for finding a decision boundary if the data is separable, useful if features show clear groupings between churn and non-churn customers.
- **Gradient Boosting (e.g., XGBoost)**: For enhancing accuracy by correcting weak predictions, ideal if initial models underperform on key metrics like precision and recall.

Visit the git repo for more information : https://github.com/Shravan-Sundar14/DACSS-604-Final-Project