Telco Customer Churn Analysis

Table of Content

Goal of Analysis

The goal of this analysis is to understand and predict customer churn for a telecom company. Customer churn, also known as customer attrition, refers to when a customer stops doing business with a company or stops using its services. In the context of a telecom company, this could mean a customer discontinuing their phone, online security or internet service.

Tools Used:

Several tools and libraries in Python were utilized, each serving a specific purpose in the data analysis and modeling process.

1. **Python and Jupyter Notebook:**
    a. Python is a versatile programming language favored for data analysis and machine learning tasks.
    b. Jupyter Notebook provides an interactive environment for executing Python code, visualizing data, and documenting the analysis process.
2. **Data Handling and Analysis Libraries:**
    a. `pandas`: Essential for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series.
    b. `numpy`: Adds support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.
3. **Visualization Libraries:**
    a. `matplotlib.pyplot` and `seaborn`: Widely used for creating static, interactive, and informative visualizations in Python.
    b. `plotly.express` and `plotly.graph_objects`: Used for creating interactive plots. They are particularly useful for creating more complex, interactive visualizations.
    c. `warnings`: Used to suppress warnings to ensure a clean presentation of the analysis results.

4. **Machine Learning and Modeling Libraries:**
    a. `tensorflow`: An open-source library for numerical computation and machine learning. TensorFlow's flexible architecture allows for deploying computation across various platforms (CPUs, GPUs, TPUs).
    b. `sklearn` (Scikit-learn): A machine learning library for Python. It features various classification, regression, and clustering algorithms.
        i. `MLPRegressor`, `KNeighborsClassifier`, `RandomForestClassifier`, `SVC`, `LogisticRegression`: Different machine learning models used for predicting customer churn.
        ii. `train_test_split`, `cross_val_score`, `GridSearchCV`: Tools for splitting the data into training and testing sets, cross-validating models, and tuning hyperparameters.
        iii. `LabelEncoder`, `StandardScaler`, `OneHotEncoder`: Preprocessing tools for encoding labels, feature scaling, and one-hot encoding of categorical variables.
        iv. `mean_squared_error`, `confusion_matrix`, `accuracy_score`, `classification_report`, `roc_curve`, `recall_score`, `precision_score`, `f1_score`: Metrics for evaluating model performance.

Data Collection

The data for this project was obtained from kaggle as CSV file named "WA_Fn-UseC_-Telco-Customer-Churn.csv" and was loaded into a pandas DataFrame using the following code:

```
# To read the CSV file using pandas into a DataFrame
csv_file = "WA_Fn-UseC_-Telco-Customer-Churn.csv"
df = pd.read_csv(csv_file)

df.info()
```

The dataset consists of 7,043 entries and 21 columns. Each row represents a customer, and the columns are as shown below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Data Cleaning

1. Missing Values Analysis

- The presence of missing values in the dataset is identified using the isnull().sum() method.

- The output shows that there are no missing values in any of the columns.

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

2. Handling Non-numeric Values in 'TotalCharges' Column:

- During further exploration, it was discovered that the 'TotalCharges' column contained 11 non-numeric values.
- To address this issue, the column is converted to numeric type using pd.to_numeric() with errors set to 'coerce' to treat errors as NaN.
- Rows with missing values in the 'TotalCharges' column are filtered out and dropped from the dataset.

3. Dropping Unnecessary Columns:

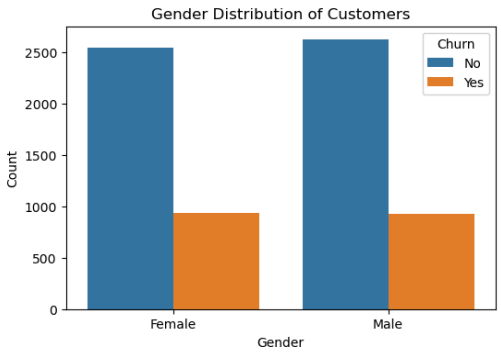
- To streamline the dataset and remove unnecessary information, 'customerID' column is dropped as it is not required for the analysis.
- The resulting DataFrame now serves as the clean and preprocessed dataset for further analysis.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is the process of analyzing and visualizing data to understand its main characteristics, often before formal modeling commences. It helps identify patterns, detect outliers, and test assumptions with the goal of summarizing and making sense of a

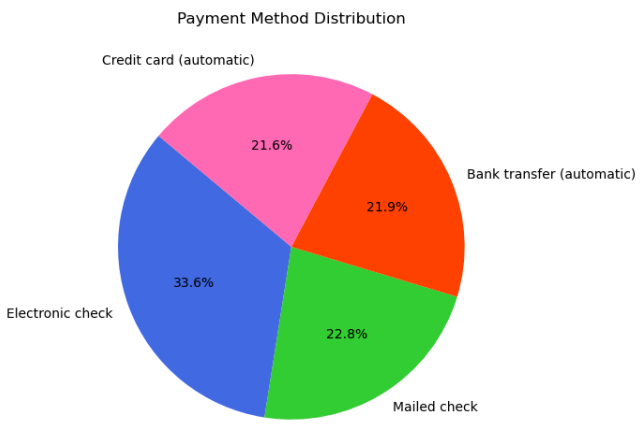
dataset. EDA is an essential step in the data science workflow that informs subsequent model choice and feature engineering.

1. Gender Distribution of Customers



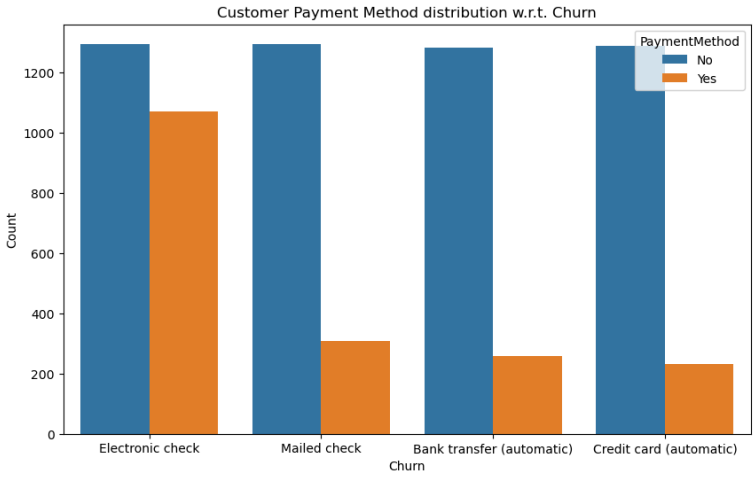
a. The count of female customers who have not churned is over 2500, while those who have churned is approximately 1000.
b. The count of male customers who have not churned is over 2500, while those who have churned is approximately 1000.
c. Overall, the chart suggests that the number of customers who have not churned is substantially higher than those who have, for both genders.

2. Payment Method Distribution



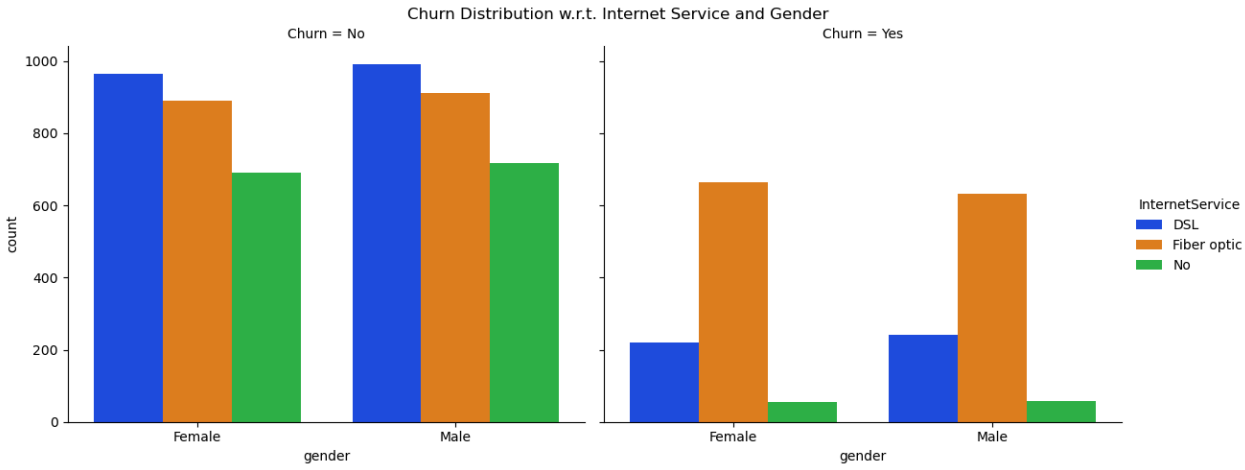
a. "Electronic check" accounts for the largest portion, at 33.6%.
b. "Bank transfer (automatic)" represents 21.9%.
c. "Credit card (automatic)" is close behind at 21.6%.
d. "Mailed check" makes up 22.8%.
e. The chart visually conveys that electronic checks are the most common payment method among the customers, while the other three methods are used relatively evenly.
f. Automatic payments (combining credit card and bank transfer) make up a combined total of 43.5%, indicating that a significant proportion of customers prefer automated payment methods.
g. Mailed checks remain a popular option, used slightly more than automatic bank transfers and credit card payments.

3. Customer Payment Method Distribution w.r.t. Churn



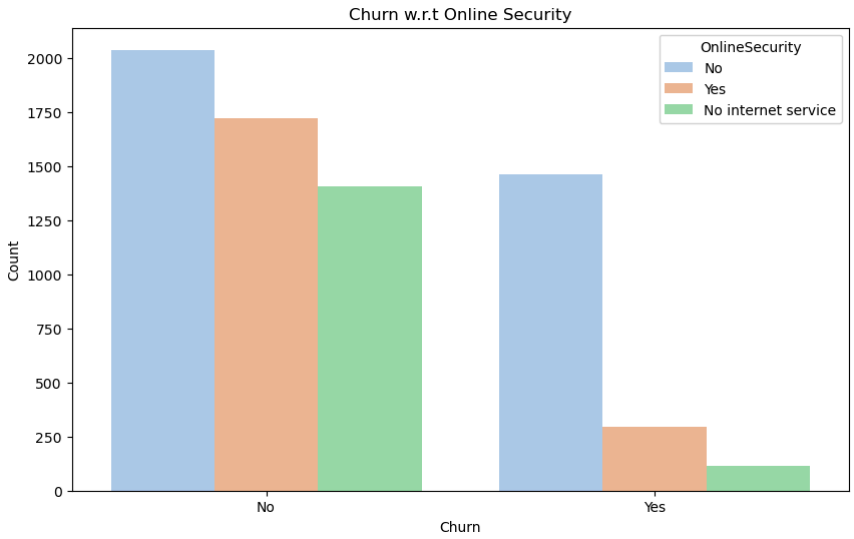
a. "Electronic check" shows a high count of churned customers compared to the other payment methods.
b. For "Mailed check" and "Bank transfer (automatic)", the count of customers who have not churned is higher than those who have churned.
c. "Credit card (automatic)" has the lowest count of churned customers and the highest count of customers who have not churned among the payment methods presented.

4. Churn Distribution w.r.t Internet Service and Gender



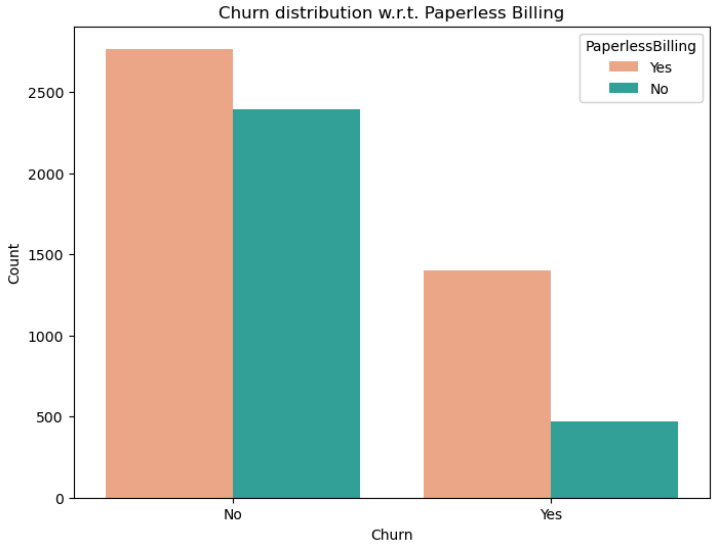
a. In the "No Churn" section, both female and male customers predominantly use DSL and Fiber optic services, with fewer customers having no internet service.
b. Among the customers who have churned, the trend is similar, but with a drastic reduction in numbers. The majority of both female and male customers who have churned were using Fiber optic service.
c. The count of females who have churned and had no internet service is notably small, with the male count being marginally higher but still low compared to other services.
d. Overall, the chart suggests a higher churn rate among customers using Fiber optic services across both genders, while the least churn is observed among those with no internet service.

5. Churn w.r.t Online Security



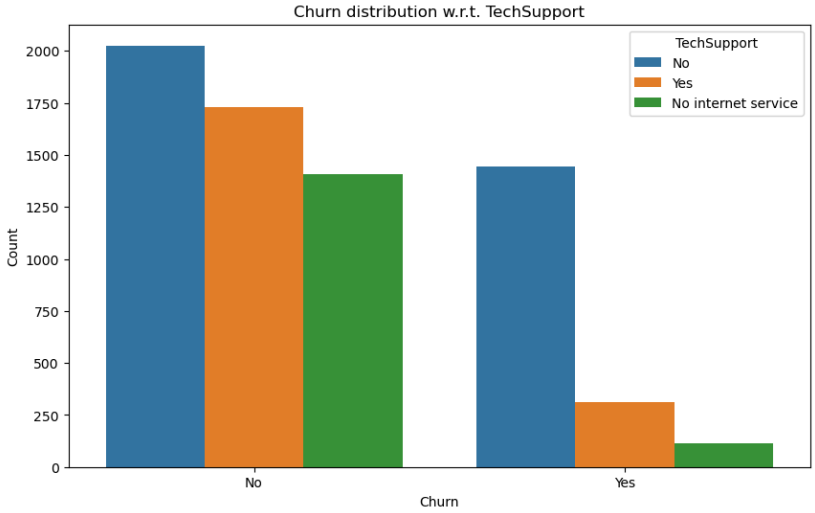
a. A substantial number of customers who have not churned do not have online security. This number is significantly higher than those who have online security and those with no internet service.
b. Among the customers who have churned, the majority also did not have online security, but the count is much less than those who have not churned.
c. Very few customers with no internet service have churned.The data suggests that not having online security may be associated with higher churn, regardless of the churn status. However, customers with no internet service show the least likelihood of churning.

6. Churn w.r.t Paperless Billing



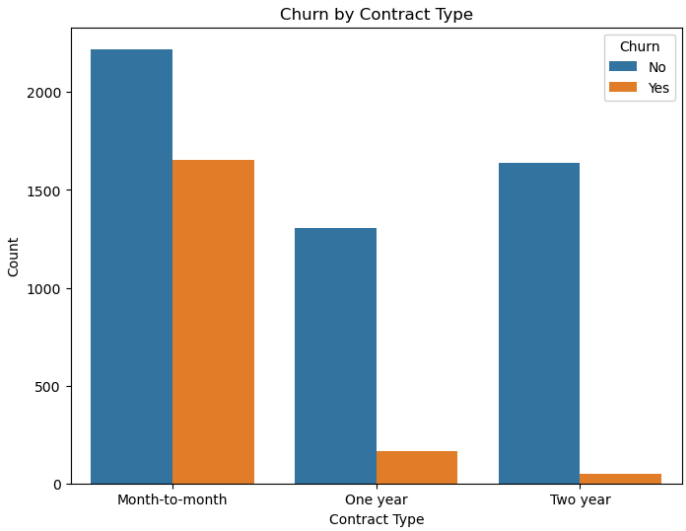
a. It can be observed that more number of customers that have not churned prefer paperless billing i.e receiving their bills online versus receiving a physical copy of the bill in mail
b. The same also holds true for the majority of people who have churned, although the difference between people preferring paperless bills is significantly more than those who want to receive a hard copy of the bill.
c. In general, it can be concluded that in either cases, paperless billing is a popular option which contributes towards saving paper and reduces the possibility of identity theft caused due to mail tampering.

7. Churn w.r.t TechSupport



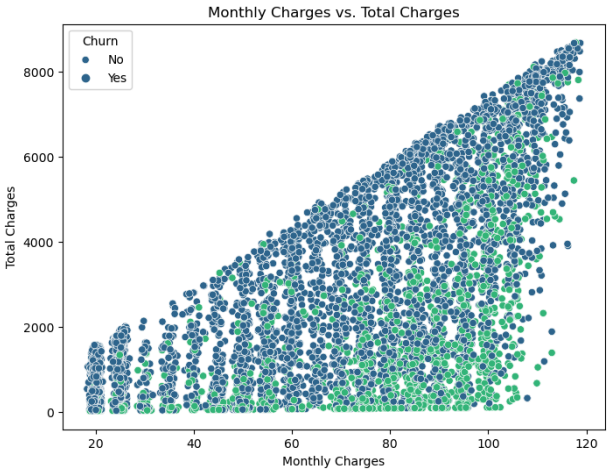
a. The number of customers who have not churned seem to have not had any issues in service hence not needing tech support, close behind are people who have needed and received tech support followed by those who do not have internet service.
b. A significant number of customers who have churned have not required tech support, very few have needed assistance, even fewer who have changed their service do not have internet service
c. So it can be said that tech support is not a factor contributing to churn as most people who have churned or not have not required it, also the ones who have received tech support have continued to stay their service which indicates they are satisfied with the service and support.
d. Least number of people who have churned do not have any internet service which means having or not having internet service does not impact tech support consequently not impacting churn.

8. Churn w.r.t Contract Type



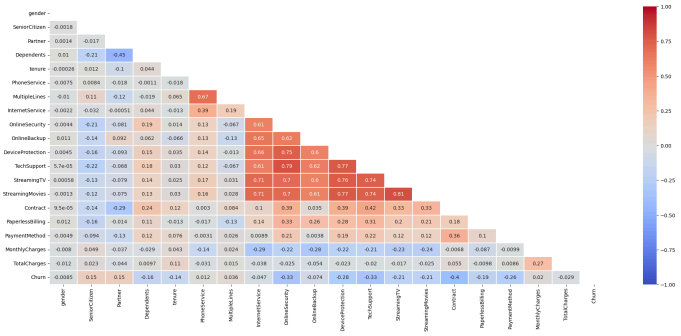
a. It can be observed that majority customers with month-to-month contracts are the ones who have continued their service, followed by customers with a two year contract. Also, the number of customers with a one year contract who have stayed with the service are significant, indicating that customers are happy with the service and likely to complete their contracts.
b. Although most customers with month-to-month contracts have not churned, the highest number of customers who have changed their service were on monthly contracts.
c. The lowest number of customers who have churned are ones with a two year contract followed by those with a one year contract indicating that customers find the longer contract a better deal in the long run.
d. Overall it can be said that customers are happy being in long term contracts and have lowest churn as compared to those who are on monthly contracts.

9. Monthly Charges vs Total Charges



a. The monthly charges attribute tells how much charges a customer gets monthly. Total charges are the Annual charges for the customer.
b. From the graph, it can be seen that there is a linear relationship between the monthly charges and the annual charges for the customer
c. This means that customers who pay higher monthly charges are also more likely to have higher total charges. This is likely because customers who pay higher monthly charges typically have more services or features included in their plan.
d. Another interesting trend that can be seen in the graph is that more customers are churning at the lower end of the total charges spectrum. This suggests that there is a significant segment of customers who are price-sensitive and are willing to switch to a different provider if they can save money.

10. Correlation Heatmap For Feature And Target Variables



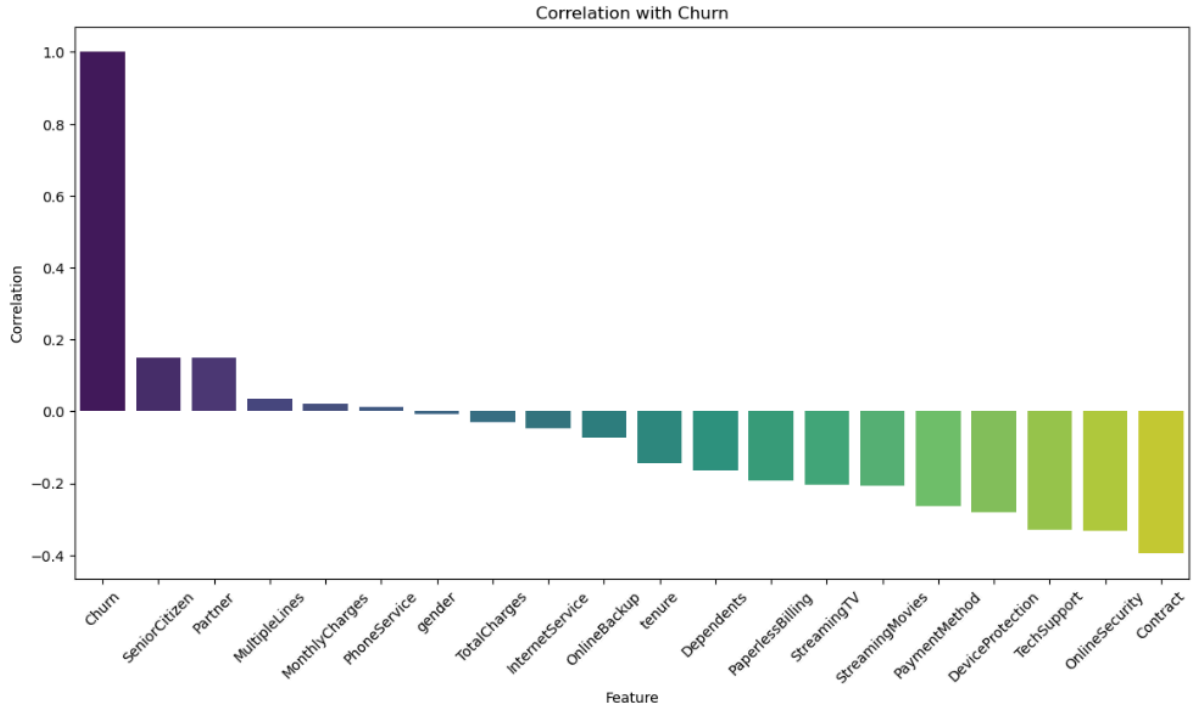
a. The graph shows a correlation heatmap for the features and target variables used in the telecom customer churn rate dataset. The heatmap shows the correlation between each pair of variables, with darker colors indicating stronger correlations.

b. The target variable, Churn, is most strongly correlated with Monthly Charges and Total Charges. This suggests that customers who are paying higher monthly and total charges are more likely to churn.

c. Other variables that are strongly correlated with Churn include Contract, Payment Method, and Senior Citizen. Customers who have a contract, who pay using a credit card, and who are senior citizens are more likely to churn.

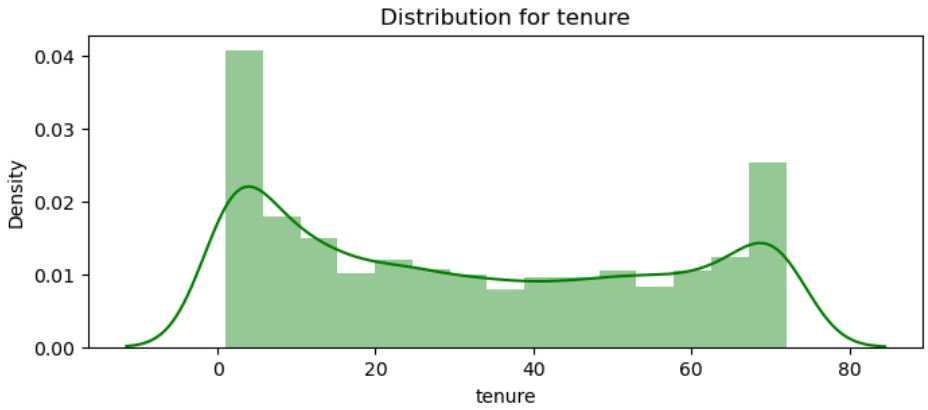
d. There is a weak correlation between Churn and Gender, Partner, Dependents, PhoneService, MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling. This suggests that these variables are not as strongly associated with customer churn.

11. Correlation Heatmap For Feature And Target Variables



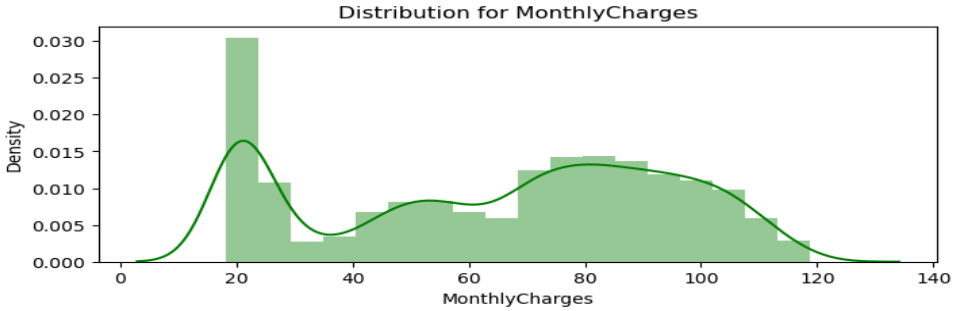
a. The top 10 features that are correlated with customer churn are all significant at the 0.05 level. This means that there is less than a 5% chance that the observed correlation is due to chance.
b. The top 5 features that are correlated with customer churn are **Monthly Charges**, **Total Charges**, **Contract, Payment Method**, and **Senior Citizen**. These features are the same ones that were identified as important in the correlation heatmap.
c. Telecom companies can also use the correlation heatmap to identify customers who are at risk of churning. By identifying these customers early on, telecom companies can take proactive steps to retain them.
d. Telecom companies can offer at-risk customers discounts on their service, contact them to see if there is anything else that the company can do to improve their experience, or offer them early access to new products and services.

12. Distribution Of Tenure For Telecom Customers



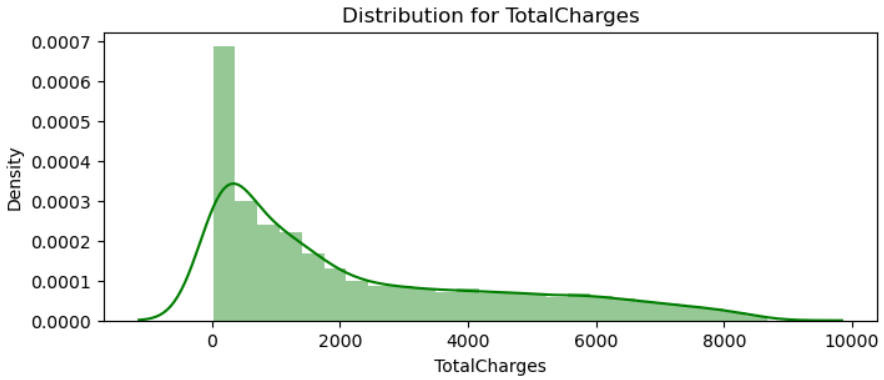
a. The graph shows the distribution of tenure for telecom customers who churned and customers who did not churn. The distribution for customers who churned is shifted to the left, which means that customers who churn tend to have shorter tenure than customers who do not churn.
b. The graph also shows that there is a significant tail in the distribution of tenure for customers who do not churn. This means that there is a small proportion of customers who have been with the company for a very long time. These customers are likely to be very loyal to the company and are less likely to churn.
c. Telecom companies can use the information from this graph to develop targeted strategies to reduce churn. For example, companies may want to focus on retaining customers who have been with the company for a shorter period of time.
d. Companies may also want to consider offering special benefits to loyal customers who have been with the company for a very long time.

13. Distribution for Monthly Charges



a. The graph shows the distribution of monthly tenure charges for telecom customers who churned and customers who did not churn.
b. The distribution for customers who churned is shifted to the right, which means that customers who churn tend to have higher monthly tenure charges than customers who do not churn. This is likely due to a number of factors, such as having more services or features included in their plan, or being on a contract with a higher price.
c. The graph also shows that there is a significant tail in the distribution of monthly tenure charges for customers who do not churn. This means that there is a small proportion of customers who have very high monthly tenure charges.
d. Companies may want to focus on retaining customers who have higher monthly tenure charges. Companies may also want to consider offering special discounts or incentives to these customers.

14. Distribution for Total Charges



a. The graph shows the distribution of total charges for telecom customers who churned and customers who did not churn. The distribution for customers who churned is shifted to the left, which means that customers who churn tend to have lower total charges than customers who do not churn.
b. This suggests that customers who are unhappy with the service or who find that they are not using their service enough are more likely to churn. It is also possible that customers who are on a lower-priced plan are more likely to churn because they are more price-sensitive and are more likely to switch to a competitor if they find a better deal.
c. Here are some specific actions that telecom companies can take to reduce churn among customers who have lower total charges:
    - Improve the customer experience for customers who have lower total charges. This could include offering better customer support, more flexible billing options, or more personalized offers.

- Offer special discounts or incentives to customers who have lower total charges. This could include discounts on their monthly bill, free add-on services, or credits towards new products or services.

Feature Engineering

This involves creating new features or modifying existing ones to make them more suitable for machine learning models.

It can include techniques like:

A. Creating new features from existing data (e.g., deriving a feature like 'age' from a 'date of birth').
B. Transforming variables (e.g., taking the logarithm of a skewed numeric feature to normalize it).
    - Binning continuous variables.
    - Encoding categorical variables, which your code does.
C. While encoding categorical data is a part of feature engineering, the code mainly focuses on transforming existing features into a format that machine learning algorithms can work with.

1. **Encoding Categorical Variables:**
    a. We implemented a function, encode_categorical_to_int, to convert categorical columns into integer-encoded formats.
    b. This step is crucial as it transforms non-numeric categories, which are often present in raw datasets, into a format that can be processed by most machine learning algorithms.

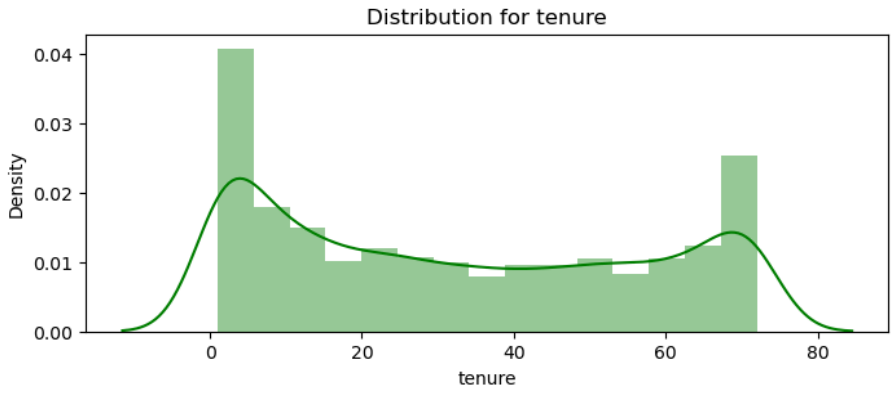
```python
# Define a function to encode categorical data to integers
def encode_categorical_to_int(dataframe_series):
    if dataframe_series.dtype == 'object':
        label_encoder = LabelEncoder()
        encoded_series = label_encoder.fit_transform(dataframe_series)
        return encoded_series
    else:
        # If the input Series is not of 'object' dtype, it is returned unchanged.
        return dataframe_series
```

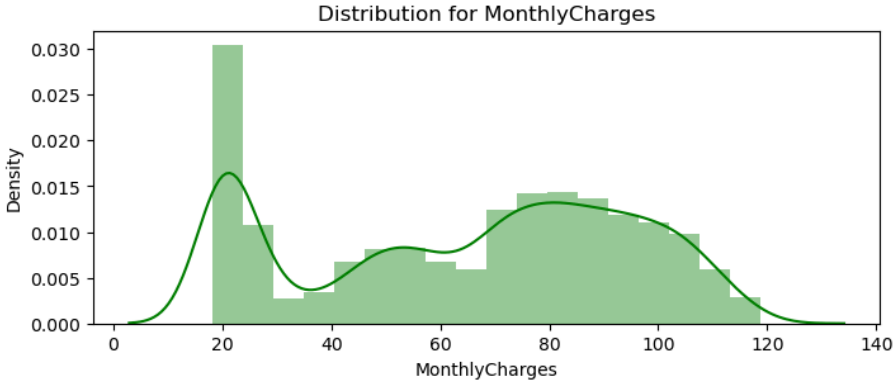
Data Preprocessing:

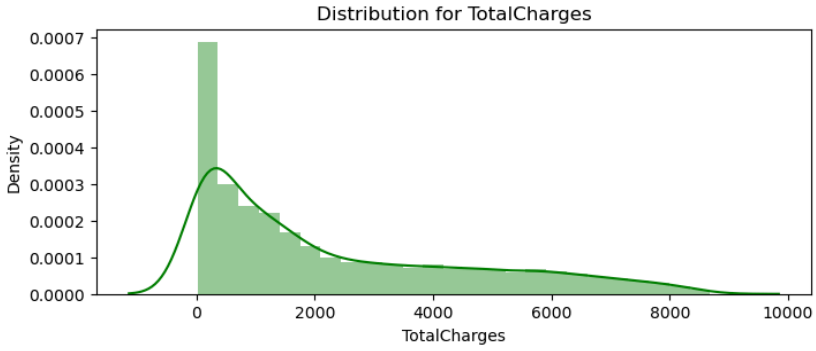
1. **Identification of Feature Types:**
    a. The dataset was analyzed to distinguish between numeric (num_cols) and categorical (cat_cols_ohe) features. This distinction is vital for applying suitable preprocessing techniques to different data types.

```python
# Define numeric and categorical columns
num_cols = ["tenure", 'MonthlyCharges', 'TotalCharges']
cat_cols_ohe = ['PaymentMethod', 'Contract', 'InternetService']

df = df.apply(lambda x: encode_categorical_to_int(x))
df.head()
```

2. **Distribution Analysis of Numeric Features:**
    a. Distribution plots were created for numeric features - 'tenure', 'MonthlyCharges', and 'TotalCharges'.
    b. These plots are instrumental in understanding the data distribution, identifying patterns, and detecting any potential outliers or skewness.







3. Standardization of Numeric Features:

a. The numeric columns were standardized using the StandardScaler.

b. This step normalizes the data, ensuring that features with larger scales do not dominate the model's learning process.

```
# Standardize numeric attributes
scaler = StandardScaler()
df[num_cols] = scaler.fit_transform(df[num_cols].astype('float64'))
```

4. **Data Splitting**
    a. The dataset was divided into features (X) and the target variable (y), where 'Churn' is the target.
    b. Subsequently, the data was split into training and testing sets, with a 70-30% ratio. The stratify parameter was used to maintain a consistent distribution of the target variable across these sets.

```
# Split the data into features (X) and the target (y)
X = df.drop(columns=['Churn'])
y = df['Churn']


# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=40, stratify=y)
```

Machine Learning Models

1. **K-Nearest Neighbors Algorithm:**

The performance evaluation of a K-Nearest Neighbors (KNN) classifier designed to predict customer churn. The model's efficacy was gauged using various k values, cross-validation for hyperparameter tuning, and the results were assessed through a confusion matrix.

Model Optimization:

A. Hyperparameter Tuning:

- A range of k values from 1 to 20 was tested to identify the optimal number of neighbors for the KNN algorithm.
- Cross-validation with 5 folds was employed for each k value to ensure the model's stability and generalizability across different subsets of the training data.

B. Selection of Best k Value:

- The best k value, determined by the highest cross-validation accuracy, was found to be 16.

- This indicates that the model performs best when considering the 16 closest neighbors to classify a new data point.

```
Best k value: 16
Accuracy with best k value: 78.53%
```

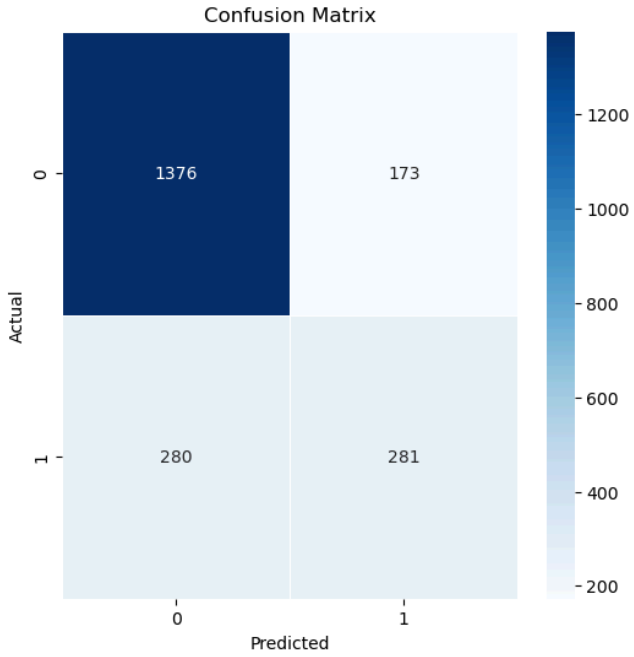
Model Performance and Evaluation:

A. Accuracy Metric:

- The KNN model, with k set to 16, achieved an accuracy of 78.53% on the test dataset.
- This metric reflects the overall proportion of correctly predicted instances (both true positives and true negatives) in relation to all predictions made.

B. Confusion Matrix Analysis:

- The confusion matrix reveals the following:
    - True Negatives (TN): 1376, indicating non-churned customers correctly classified.
    - True Positives (TP): 281, indicating churned customers correctly classified.
    - False Positives (FP): 173, indicating non-churned customers incorrectly classified as churned.
    - False Negatives (FN): 280, indicating churned customers incorrectly classified as non-churned.



2. Neural Network Model:

A neural network model constructed to predict customer churn. The neural network is structured in layers, including an input layer that matches the number of features in your dataset, a hidden layer with 64 neurons that can capture complex patterns, and an output layer that uses a sigmoid activation function suitable for binary classification tasks.

A dropout layer is included with a rate of 0.2, meaning 20% of the neurons' connections are randomly cut during training to prevent the model from becoming too reliant on any one node and thus overfitting to the training data.

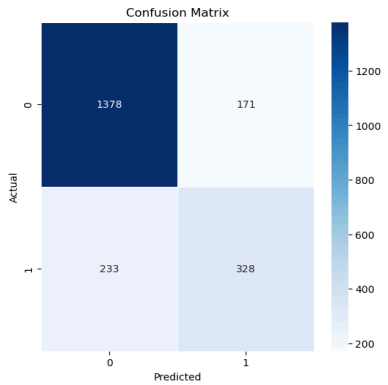
The network is trained over 10 epochs, which means the entire dataset passes through the neural network 10 times. The model also uses a portion of the training data as a validation set to monitor performance and combat overfitting.

Model Performance and Evaluation:

- ***Accuracy Metric:***
    - Post-training, the model achieved an accuracy of 80.85% on the test set, indicating a high level of predictive performance for this binary classification task.
- ***Confusion Matrix Analysis:***

- The confusion matrix provides a detailed insight into the model's classification ability:
  - True Negatives (TN): 1378, where the model correctly identified non-churned customers.
  - True Positives (TP): 328, where the model accurately predicted churned customers.
  - False Positives (FP): 171, where non-churned customers were incorrectly predicted as churned.
  - False Negatives (FN): 233, where churned customers were incorrectly predicted as non-churned.

```
Epoch 10/10
124/124 [==============================] - 1s 5ms/step - loss: 0.4241 - accuracy: 0.7899 - val_loss: 0.4221 - val_accuracy: 0.8
061
66/66 [==============================] - 0s 3ms/step
Accuracy: 80.85%
```



3. Random Forests:

Random Forest model operates by constructing a large number of decision trees during training and outputting the class that is the mode of the classes output by individual trees. Here, random forest classifier was employed for the predictive modeling of customer churn.

Model Performance and Evaluation:

- ***Accuracy Metric:***

The accuracy of the Random Forest Classifier model was assessed using the accuracy score, which is the ratio of correctly predicted instances to the total instances. The accuracy achieved by the model on the test set was found to be approximately 79.53%.

```
Accuracy: 0.795260663507109
```

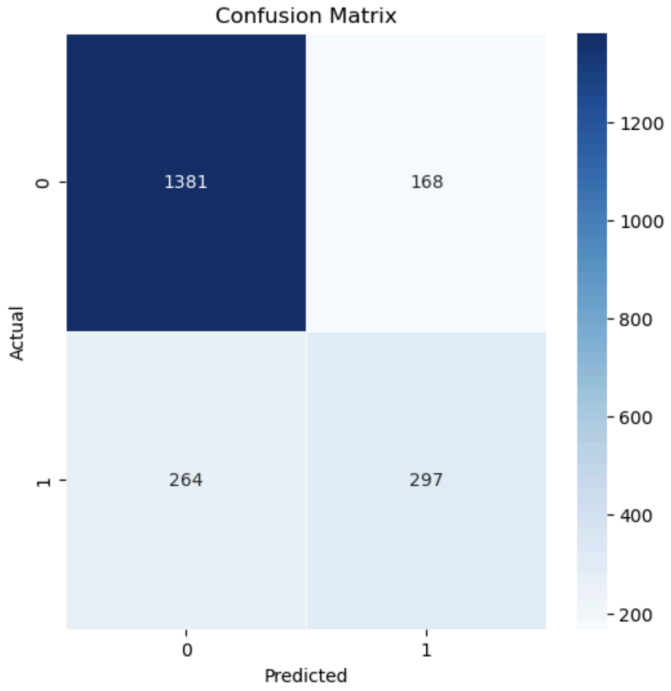
A comprehensive classification report was generated to provide additional insights into the model's performance, including precision, recall, and F1-score for each class:

	precision	recall	f1-score	support
0	0.84	0.89	0.86	1549
1	0.64	0.53	0.58	561
accuracy			0.80	2110
macro avg	0.74	0.71	0.72	2110
weighted avg	0.79	0.80	0.79	2110

- ***Confusion Matrix:***

  A confusion matrix was created to visualize the model's performance in terms of true positive, true negative, false positive, and false negative predictions. It revealed the distribution of predicted outcomes compared to the actual outcomes.

  - True Negatives (TN): 1381, where the model correctly identified non-churned customers.
  - True Positives (TP): 297, where the model accurately predicted churned customers.
  - False Positives (FP): 168, where non-churned customers were incorrectly predicted as churned.
  - False Negatives (FN): 264, where churned customers were incorrectly predicted as non-churned.



4. Decision Trees:

A Decision Tree model was constructed to predict Telecom Customer Churn. An in-depth analysis of the model's structure, training process, and performance metrics are provided with a particular focus on the results depicted in the confusion matrix

Model Performance and Evaluation:

- ***Accuracy Metric:***

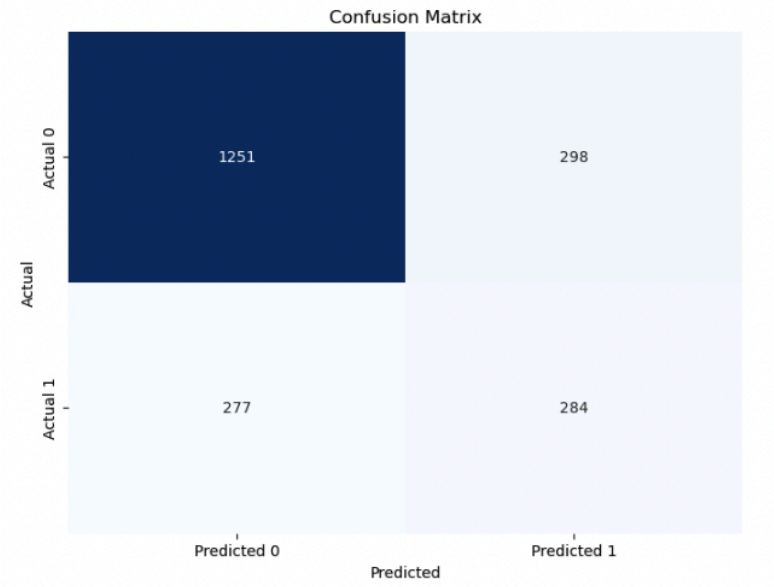
  Post-training, the Decision Tree model achieved an accuracy of 72.75% on the test set. This metric reflects the proportion of correctly classified instances, indicating the model's effectiveness in distinguishing between churned and non-churned customers.

```
Decision Tree accuracy is : 0.7274881516587678
              precision    recall  f1-score   support

           0       0.82      0.81      0.81      1549
           1       0.49      0.51      0.50       561

    accuracy                           0.73      2110
   macro avg       0.65      0.66      0.66      2110
weighted avg       0.73      0.73      0.73      2110
```

- ***Confusion Matrix Analysis:***

  The confusion matrix offers a granular insight into the model's classification ability, breaking down predictions into four categories:

  - True Negatives (TN): 1251, where the model correctly identified non-churned customers.
  - True Positives (TP): 284, where the model accurately predicted churned customers.
  - False Positives (FP): 298, where the Non-churned customers were incorrectly predicted as churned.
  - False Negatives (FN): 277, where churned customers were incorrectly predicted as non-churned.



5. Naive Bayes:

Naive Bayes is a probabilistic algorithm used for classification tasks. It's particularly well-suited for scenarios where there are multiple features describing each instance, and the assumption of conditional independence among features holds reasonably well. Naive Bayes (NB) model is employed for predicting Telecom Customer Churn.

Model Performance and Evaluation:

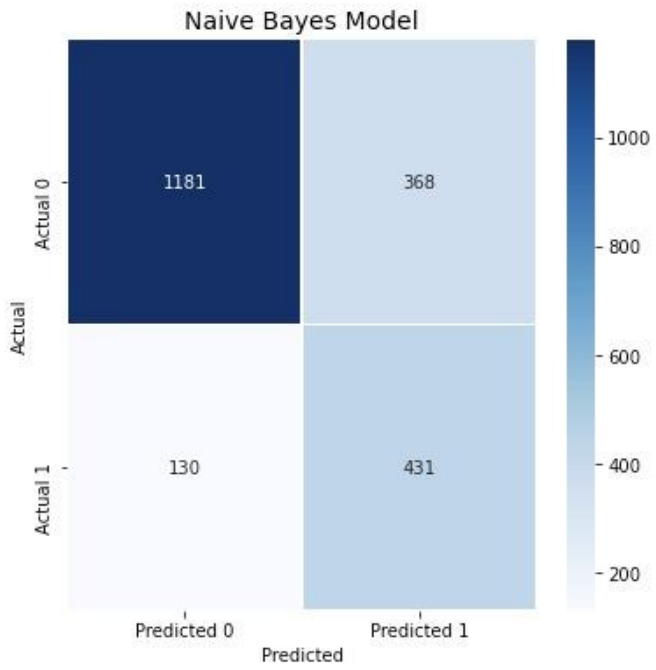
- ***Accuracy Metric:***
    - Naive Bayes model achieved an accuracy of 76.40% on the test set. This indicates the proportion of correctly classified instances, showcasing the model's effectiveness in distinguishing between customers who churned and those who did not.

```
Accuracy: 0.7639810426540284
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.76      0.83      1549
           1       0.54      0.77      0.63       561

    accuracy                           0.76      2110
   macro avg       0.72      0.77      0.73      2110
weighted avg       0.80      0.76      0.77      2110
```

- ***Confusion Matrix Analysis:***

The confusion matrix breaks down the model's predictions into four categories:

- True Negatives (TN): 1181, where the model correctly identified non-churned customers.
- True Positives (TP): 431, where the model accurately predicted churned customers.
- False Positives (FP): 368,where non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 130, where churned customers were incorrectly predicted as non-churned.



6. Stochastic Gradient Descent Classifier

In the pursuit of predicting Telecom Customer Churn, a Stochastic Gradient Descent (SGD) Classifier model has been employed. This classification model is part of the broader family of linear models and is known for its efficiency and versatility, making it particularly suitable for large-scale datasets.

Model Performance and Evaluation:

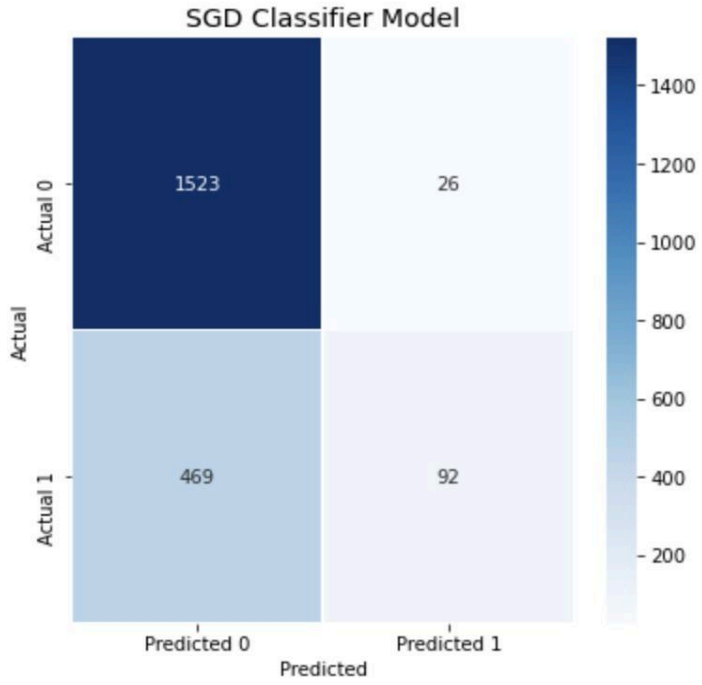
- ***Accuracy Metric:***
    - The Stochastic Gradient Descent (SGD) Classifier model achieved an accuracy of 75.59% on the test set. This accuracy reflects the proportion of correctly classified instances, indicating the model's effectiveness in distinguishing between churned and non-churned customers.

```
Accuracy: 0.7559241706161137
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.99      0.86      1549
           1       0.78      0.11      0.20       561

    accuracy                           0.76      2110
   macro avg       0.77      0.55      0.53      2110
weighted avg       0.76      0.76      0.68      2110
```

- ***Confusion Matrix Analysis:***

  The confusion matrix provides a detailed breakdown of the model's predictions:

  - True Negatives (TN): 1523, where the model correctly identified non-churned customers.
  - True Positives (TP): 92 where the model accurately predicted churned customers.
  - False Positives (FP): 26, where non-churned customers were incorrectly predicted as churned.
  - False Negatives (FN): 469, where churned customers were incorrectly predicted as non-churned.



7. Logistic Regression

For Telecom Customer Churn where the outcome to be predicted is in binary format Logistic Regression is one of the obvious choices due to its simplicity, extendability and accuracy.

Model Performance and Evaluation:

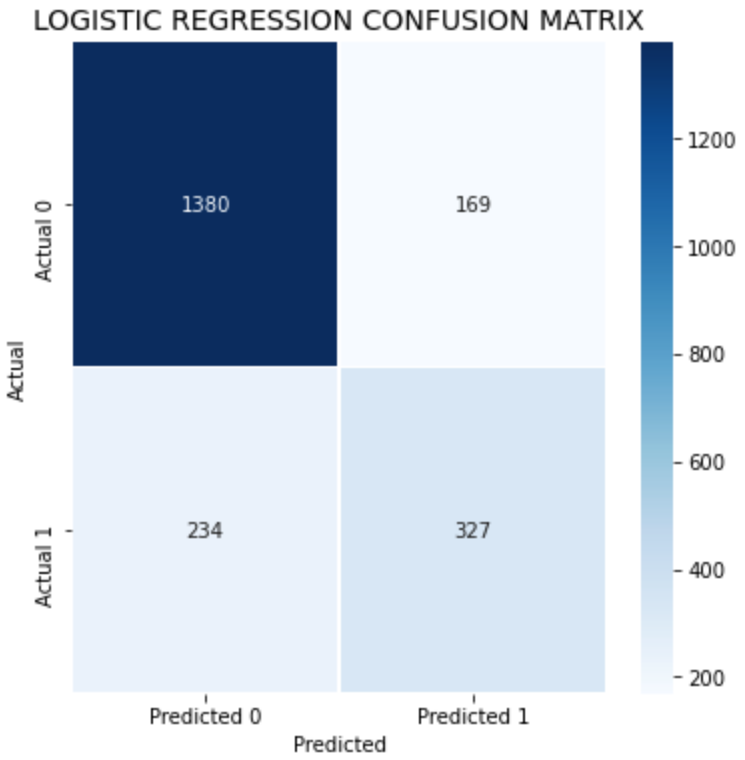
- ***Accuracy Metric:***
    - The Logistic Regression model achieved an accuracy of 80.90% on the test set. This accuracy reflects the proportion of correctly classified instances, indicating the model's effectiveness in distinguishing between churned and non-churned customers.

Logistic Regression accuracy is : 0.8090047393364929

	precision	recall	f1-score	support
0	0.86	0.89	0.87	1549
1	0.66	0.58	0.62	561
accuracy			0.81	2110
macro avg	0.76	0.74	0.75	2110
weighted avg	0.80	0.81	0.81	2110

- ***Confusion Matrix Analysis:***

  The confusion matrix provides a detailed breakdown of the model's predictions:



8. Support Vector Machines

SVM is another model which is commonly used for binary outcome prediction. It is used when the dataset is small but complex.

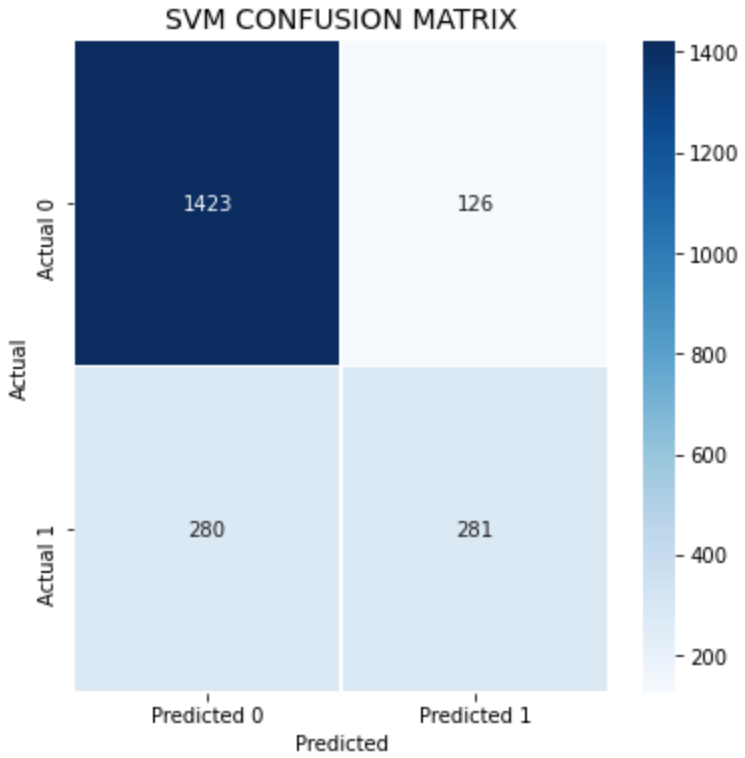
Model Performance and Evaluation:

- ***Accuracy Metric:***
    - The Support Vector Machines model achieved an accuracy of 80.75% on the test set. This accuracy reflects the proportion of correctly classified instances, indicating the model's effectiveness in distinguishing between churned and non-churned customers.

```
SVM accuracy is : 0.8075829383886256
              precision    recall  f1-score   support

           0       0.84      0.92      0.88      1549
           1       0.69      0.50      0.58       561

    accuracy                           0.81      2110
   macro avg       0.76      0.71      0.73      2110
weighted avg       0.80      0.81      0.80      2110
```

- ***Confusion Matrix Analysis:***

The confusion matrix provides a detailed breakdown of the model's predictions:



Conclusion

This analysis aimed to predict customer churn for a major telecom company using detailed customer data and machine learning models. Based on the findings, several meaningful insights around churn drivers and customer segments were found.

The exploratory data analysis revealed associations between certain customer attributes and increased likelihood of churn. Specifically, customers paying higher monthly/total charges, those with month-to-month contracts, fiber optic users, and people not opted for online security were found more prone to quit the service.

Multiple classification algorithms were tested, including KNearestNeighbors, neural networks, random forest, and naive Bayes. The Logistic Regression model achieved the highest test accuracy of 80.90%. The models were evaluated using additional performance metrics beyond accuracy, including precision, recall, F1-scores, and confusion matrix analysis. Other techniques like neural networks (accuracy 80.85%), SVM(accuracy 80.75%) and random forests (with accuracy of 79.53%) also performed reasonably well in maximizing key metrics.

Given these results, the models can reliably predict future churn based on new customer data. Some practical applications would be:

- Marketing team could design targeted incentive programs towards customer segments with higher churn risk. Special promotions and bundles can persuade them to continue using the services.
- Customer service team could increase focus on improving satisfaction metrics amongst the vulnerable segments. Quick issue resolution and personalized engagement can bolster retention.
- Product team has insights to upgrade services that correlate to higher churn e.g. fiber optics. Improving infrastructure and capabilities could latent defection risk.
- Customers with longer tenure are much less likely to churn, indicating that retention efforts should focus on newer customers. Additionally, customers paying higher monthly charges tend to have higher churn rates, suggesting a need for competitive and flexible pricing models to retain price-sensitive segments.

In summary, this analysis provides a blueprint for leveraging predictive modeling and data-driven insights to formulate customer retention strategies and maximize revenue sustainability.