# **Adv. Machine Learning**

# **Telecom Customer Churn**

Project Summary

**Group 18**

Pragya Sharma (51) sec-02

Swathi Vegiraju (56) sec -01

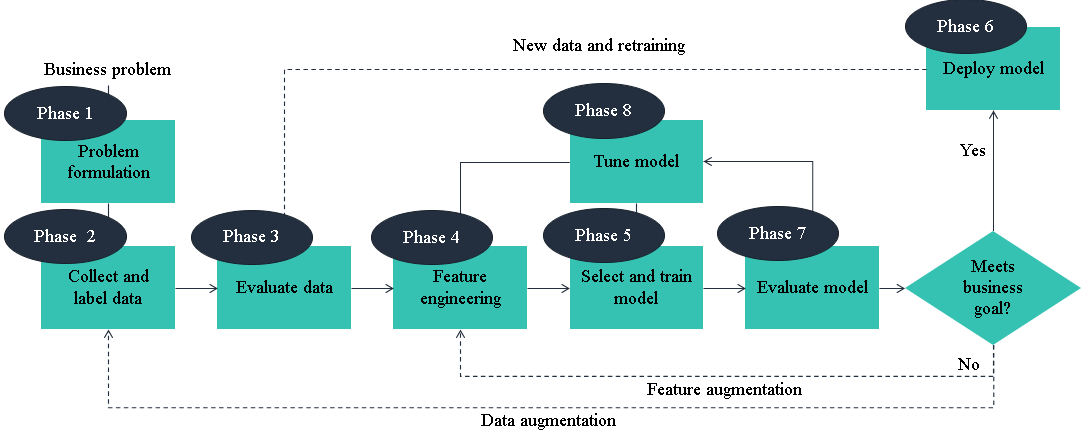
Rahul Gamara (10) sec-02

**Introduction:**

In today's hyper-competitive business landscape, maintaining customer satisfaction is paramount for success. One of the biggest challenges companies face is customer churn, the phenomenon where customers discontinue using their services. This report focuses on analyzing the "Telco Customer Churn" dataset to help companies predict which customers are at high risk of churn, particularly within the telecommunications industry. Customer churn is of utmost importance as each lost customer directly translates to lost revenue for a company. By understanding the factors influencing churn, businesses can implement strategies to mitigate it, thereby safeguarding their revenue streams and fostering long-term customer relationships. Leveraging a machine learning pipeline for analysis enhances the predictive capabilities, enabling proactive measures to retain valuable customers.

**Methodology:**

This project integrates the machine learning pipeline learned in our course, covering data preprocessing, feature engineering, hyperparameter tuning, model selection, and evaluation. After cleaning and transforming the "Telco Customer Churn" dataset and enhancing features, hyperparameter tuning optimizes algorithm performance. We selected logistic regression, random forest, and XGBoost for churn prediction. Evaluation includes confusion matrices and popular performance metrics like F1 score, precision, recall, and accuracy. By leveraging these techniques, businesses gain a robust tool to predict and prevent customer churn, maximizing revenue and fostering customer loyalty.



**About Dataset:**

The dataset used in this project is collected from the [**kaggle website**](https://www.kaggle.com/code/jasonyu001/churn-rate-prediction-log-rf-xgb#3.Predicting), the dataset contains **7044 Instances/Rows** and **21 features/columns** as described below:

**Customer ID**: Unique identifier for each customer

**Gender**: Gender of the customer (e.g., Male, Female)

**Senior Citizen**: Indicates if the customer is a senior citizen (1: Yes, 0: No)

**Partner**: Indicates if the customer has a partner (Yes, No)

**Dependents**: Indicates if the customer has dependents (Yes, No)

**Tenure**: Number of months the customer has stayed with the company

**Phone Service**: Indicates if the customer has phone service (Yes, No)

**Multiple Lines**: Indicates if the customer has multiple lines (Yes, No, No phone service)

**Internet Service**: Type of internet service subscribed by the customer (DSL, Fiber optic, No)

**Online Security**: Indicates if the customer has online security service (Yes, No, No internet service)

**Online Backup**: Indicates if the customer has online backup service (Yes, No, No internet service)

**Device Protection**: Indicates if the customer has device protection service (Yes, No, No internet service)

**Tech Support**: Indicates if the customer has tech support service (Yes, No, No internet service)

**Streaming TV**: Indicates if the customer has streaming TV service (Yes, No, No internet service)

**Streaming Movies**: Indicates if the customer has streaming movie service (Yes, No, No internet service)

**Contract**: Type of contract the customer has (Month-to-month, One year, Two year)

**Paperless Billing**: Indicates if the customer uses paperless billing (Yes, No)

**Payment Method**: Payment method used by the customer (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic))

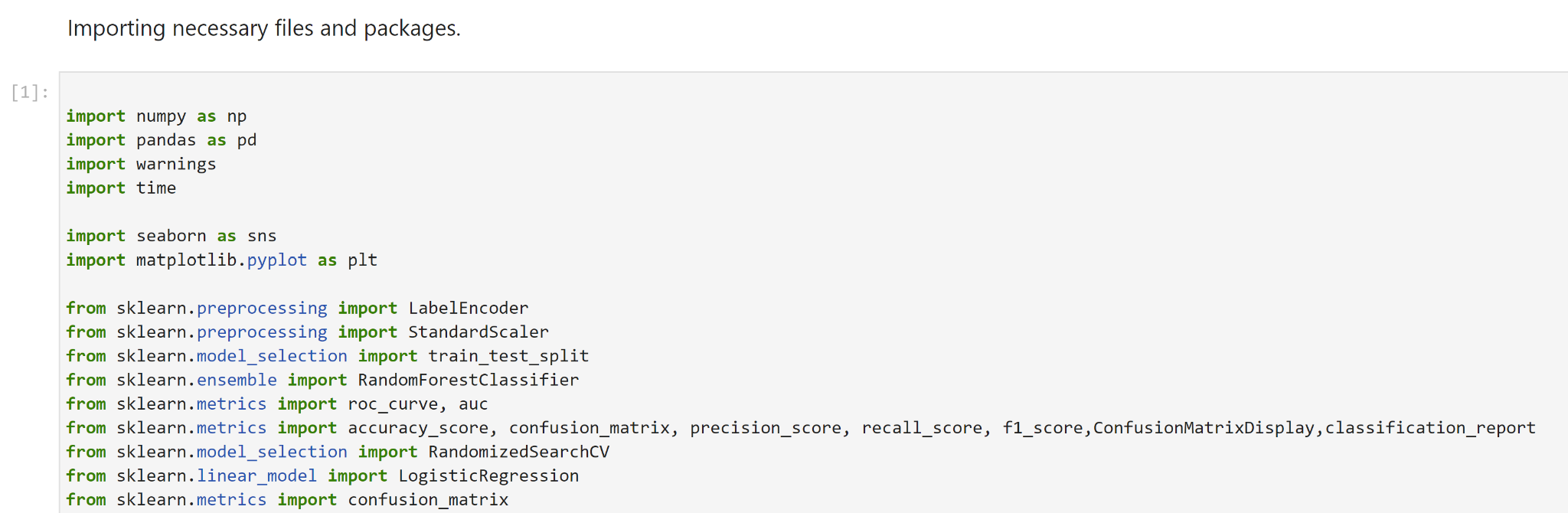
**Monthly Charges**: Monthly charges billed to the customer

**Total Charges**: Total charges billed to the customer

**Churn**: Indicates if the customer churned (Yes, No)

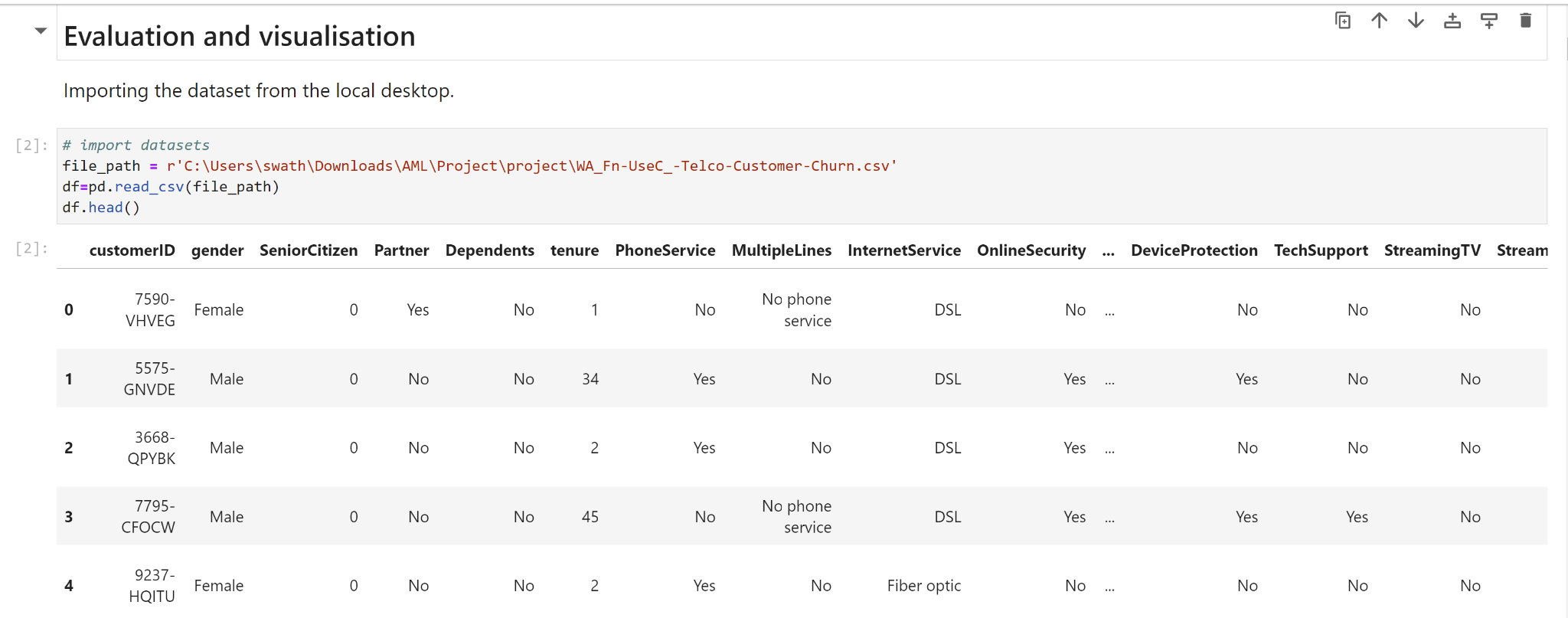
**Data Cleaning & Preprocessing:**

Importing the required libraries.

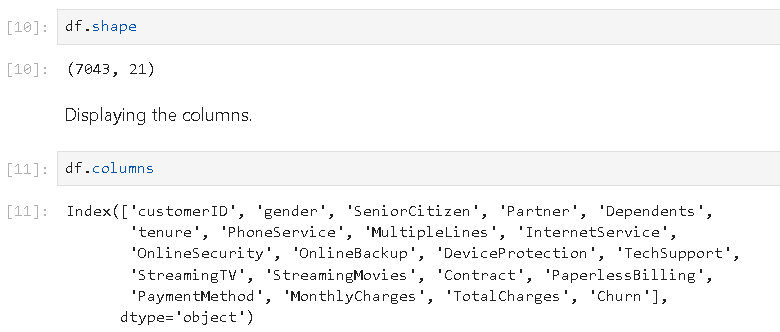


**Evaluation and Visualization:**

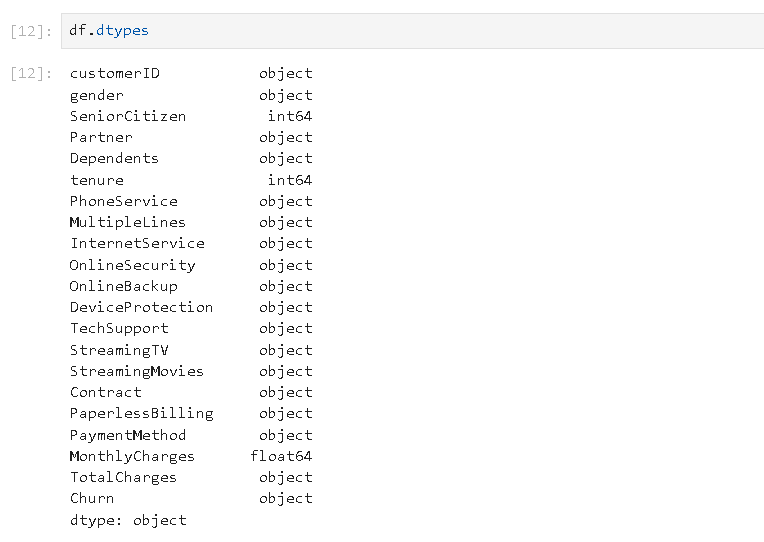
Reading the csv file from the local desktop.



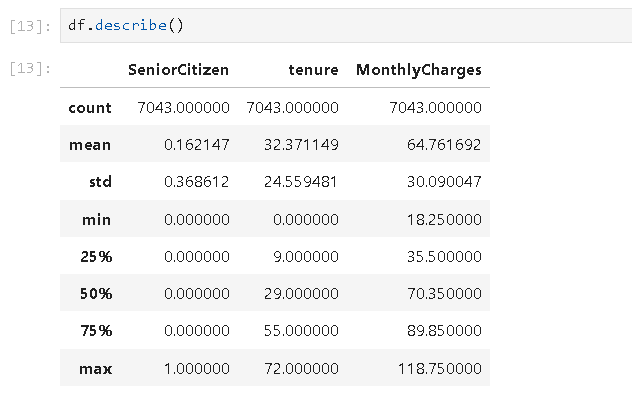
The dataset comprises 7043 rows and 21 columns. The columns are *customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup,DeviceProtection,TechSupport, StreamingTV,StreamingMovies, Contract,PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn*

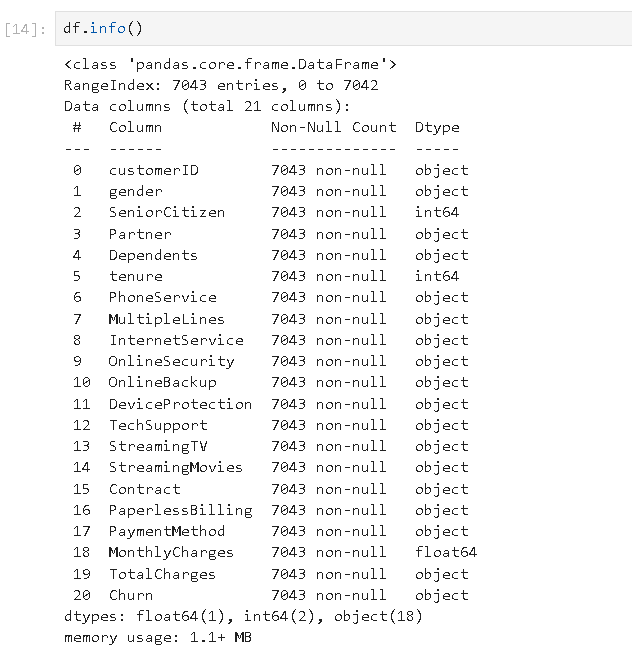


Displaying the datatypes of each column. *SeniorCitizen* and *tenure* is of int64 data type while *MonthlyCharges* is of float64 data type. Rest all are object data type.



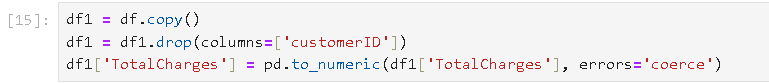
Next, we go through the descriptive statistics of the data set. This function shows the count, mean, standard deviation, minimum and maximum value and the three quartiles for the quantitative columns.

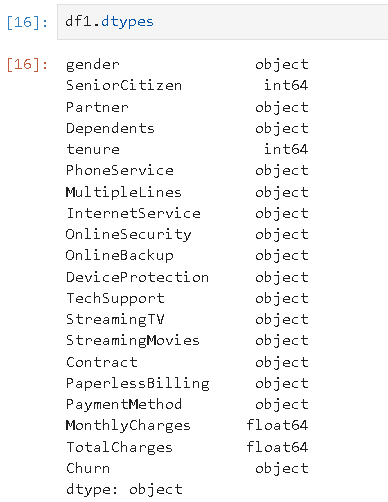




We can see above that the columns '*SeniorCitizen*', '*tenure*', '*MonthlyCharges*' and '*TotalCharges*' are the only numerical columns. However, the *TotalCharges* column's data type is object. As a result, we need to change its datatype from object to float.

Also, the column *customerID* has no such significance in our analysis hence we decide to drop it.





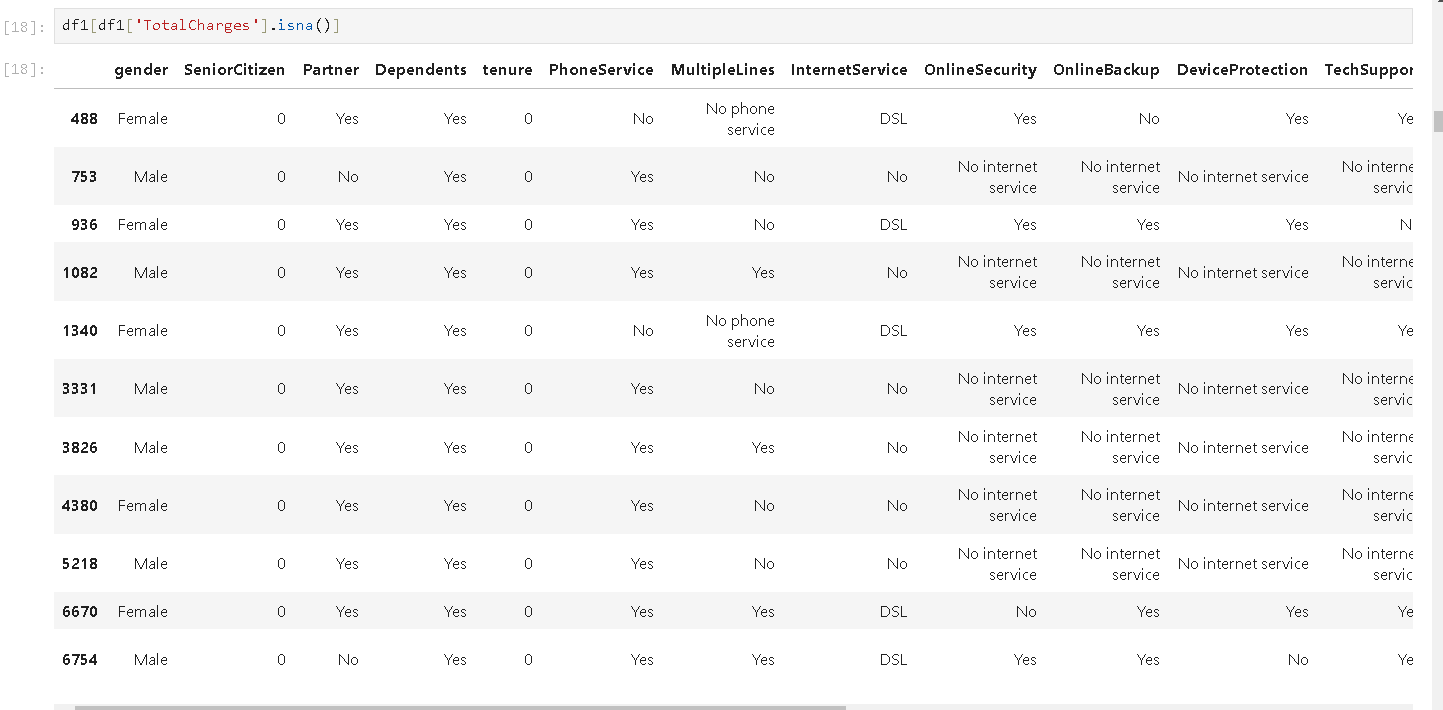
We can see the changes clearly that the *TotalCharges* is now a float datatype column and *customerID* is no longer a feature in the dataset.

**Checking for missing values:**

Checking for null values if any. There were 11 missing values in *TotalCharges*.

****

We can notice that *tenure* has all values 0. Adding to that, all 0 tenure rows are also with null TotalCharges, and with no Churn. It is very likely that these are new customers who have just started using the service within the past month and have not been charged yet, hence there is no way to determine their churn status.

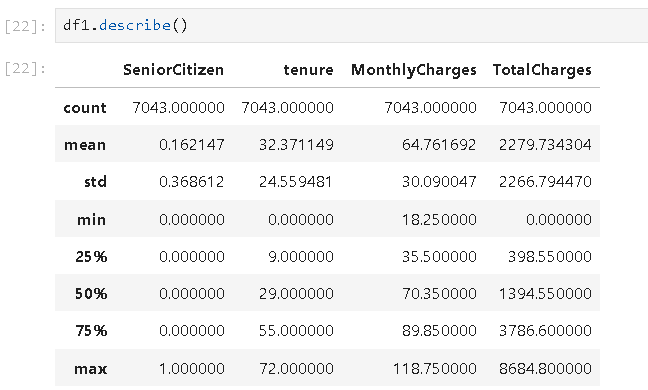




Since there are only 11 rows, we will just impute the null values with 0.

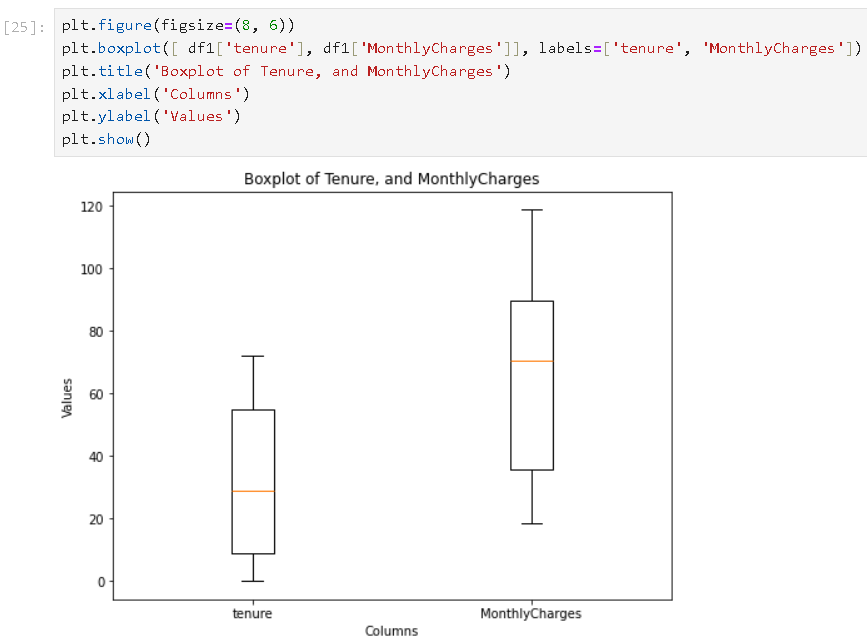


We can see the count value for *TotalCharges* go from 7032 to 7043, which confirms the successful imputation.



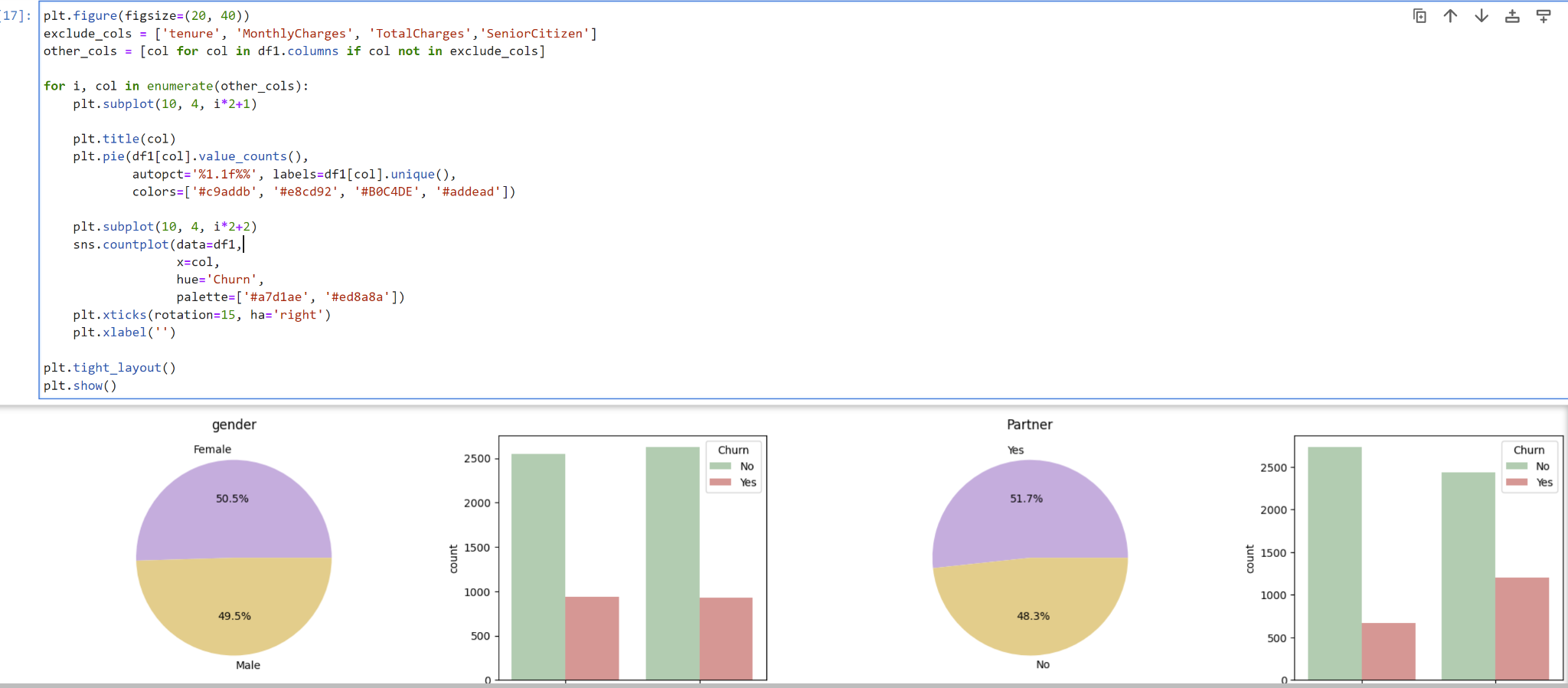
**Checking for outliers:**

Plotting a box plot to check if the dataset has any outliers or not. Here we can see there are no outliers.



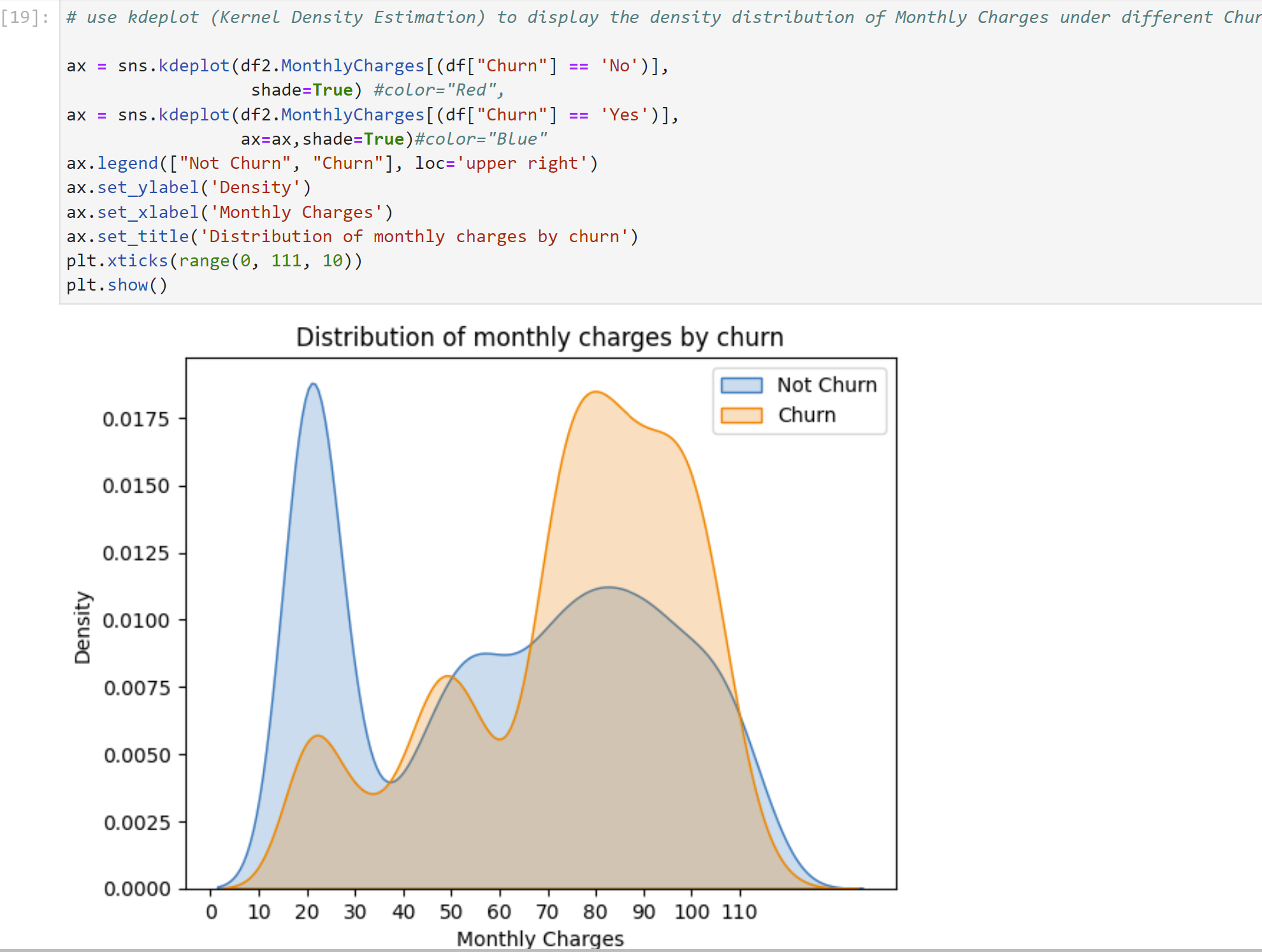
**Exploratory Data Analysis:**

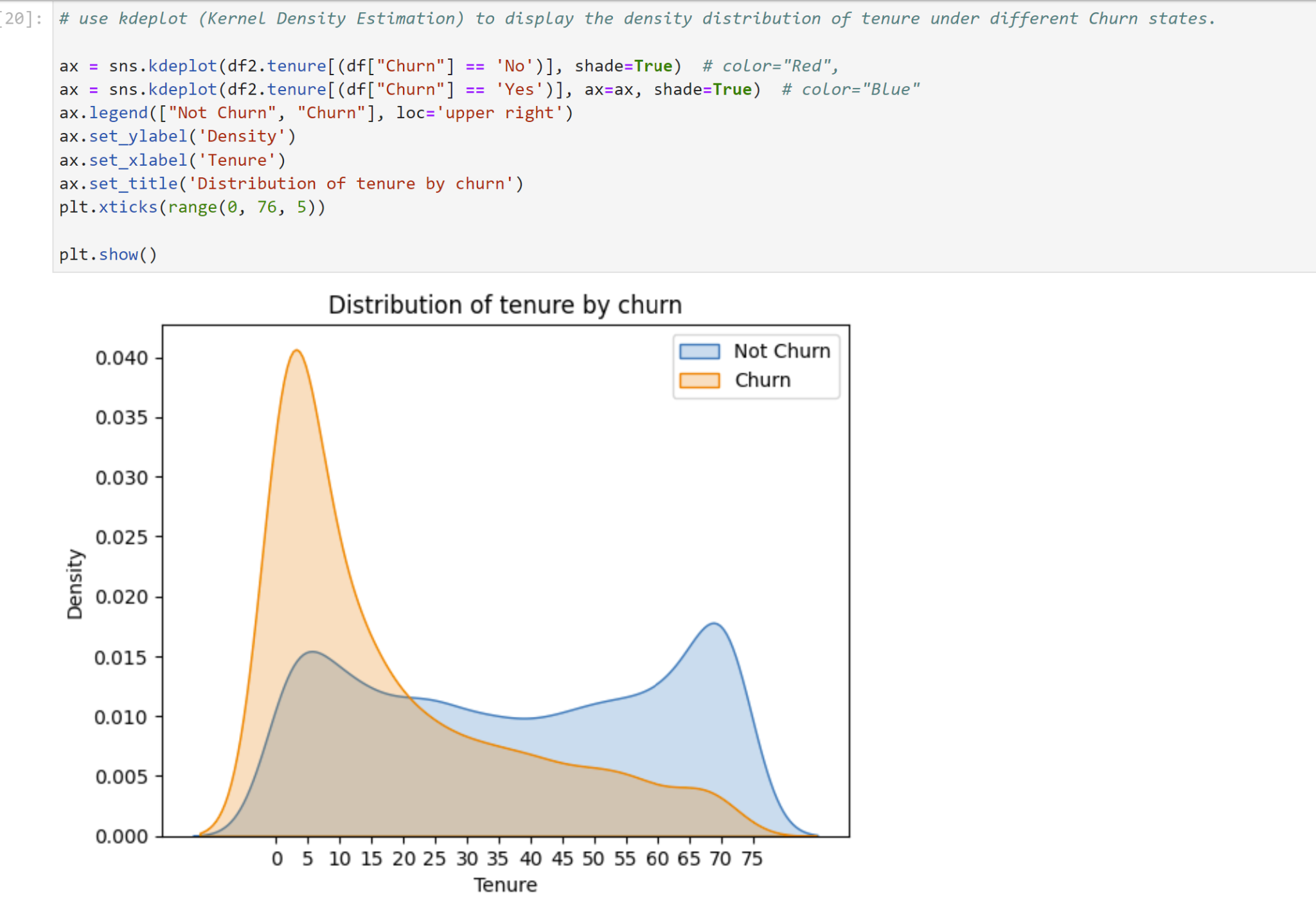
We will now Generate a series of countplots to visualize the relationship between selected categorical variables and the churn status of customers.



Insights: gender - gender does not affect the client's decision Partner & Dependents - clients in relationships, as well as clients with children, are less likely to refuse services.

Understanding the relation between Numerical variables with respect to Target variable





Insights: we can say that as the Price increases customers are mostly likely to churn and the customers who stay for short period of time are more likely to churn.

**Feature Encoding:**

With the help of LabelEncoder() we are converting all the categorical variables into numerical variables.



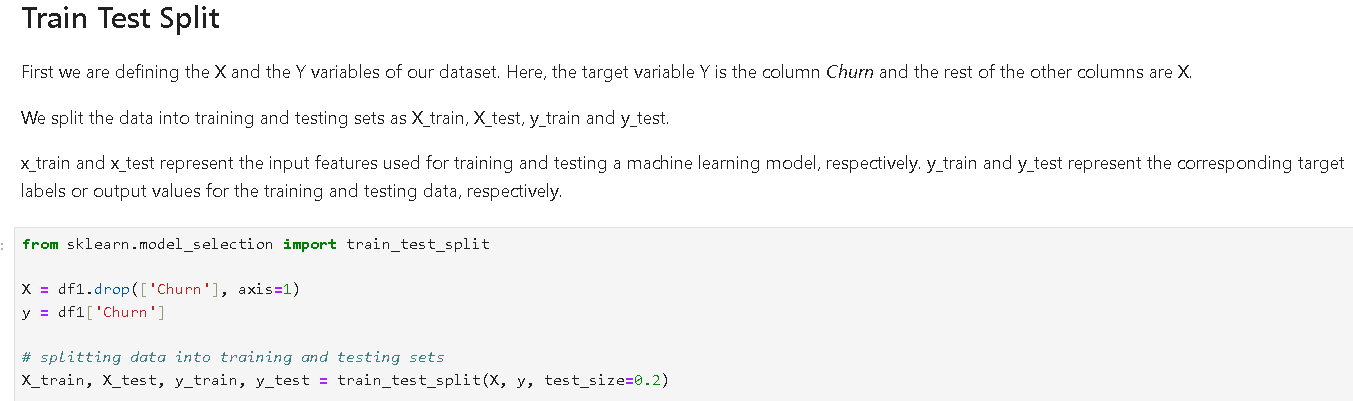
## **Feature Scaling:**

With the help of StandardScaler() we are scaling the data because few columns are not in the range with respect to other columns such as tenure, monthlycharges and totalcharges

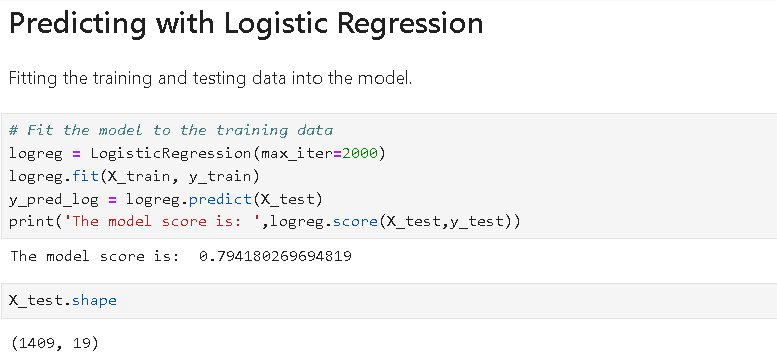


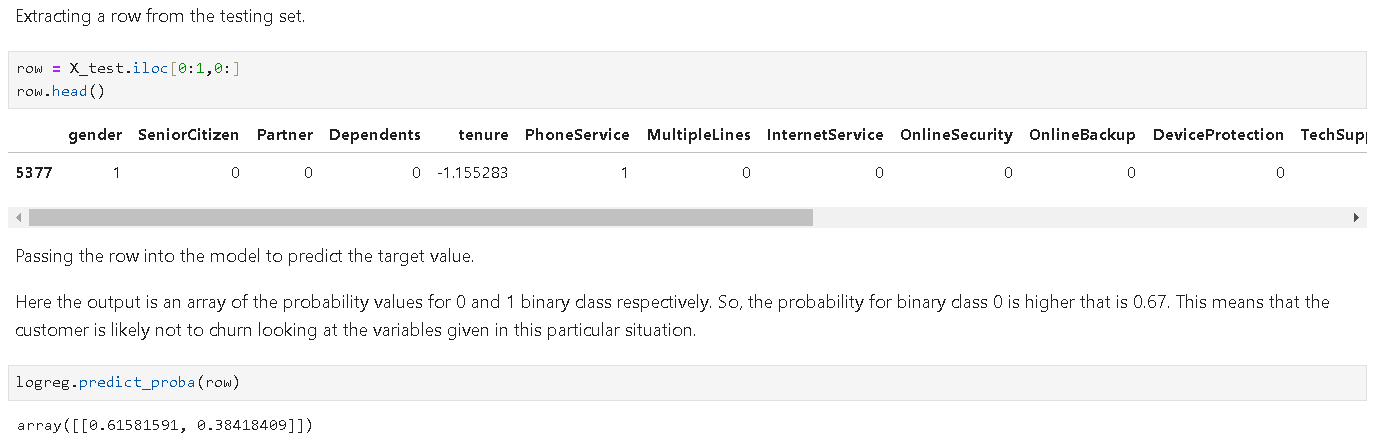
**Train Test Split:**

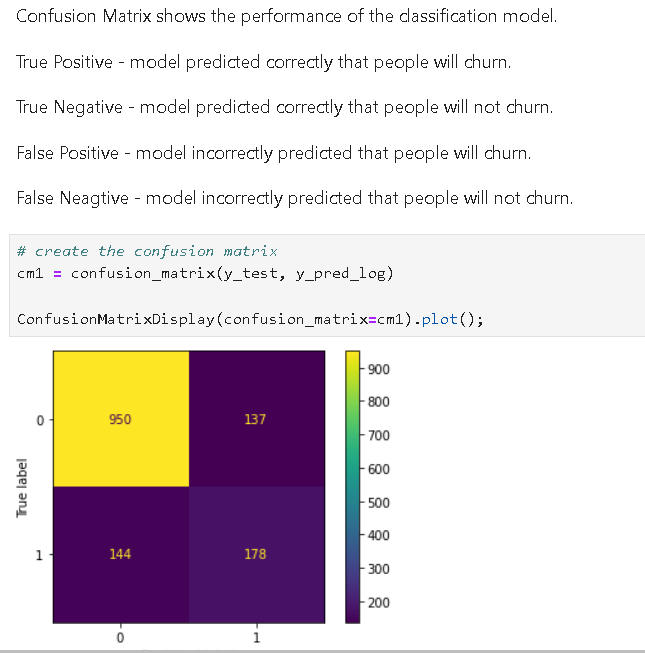
First, we will define the x and y variables. Then we will split the dataset into training and testing sets.

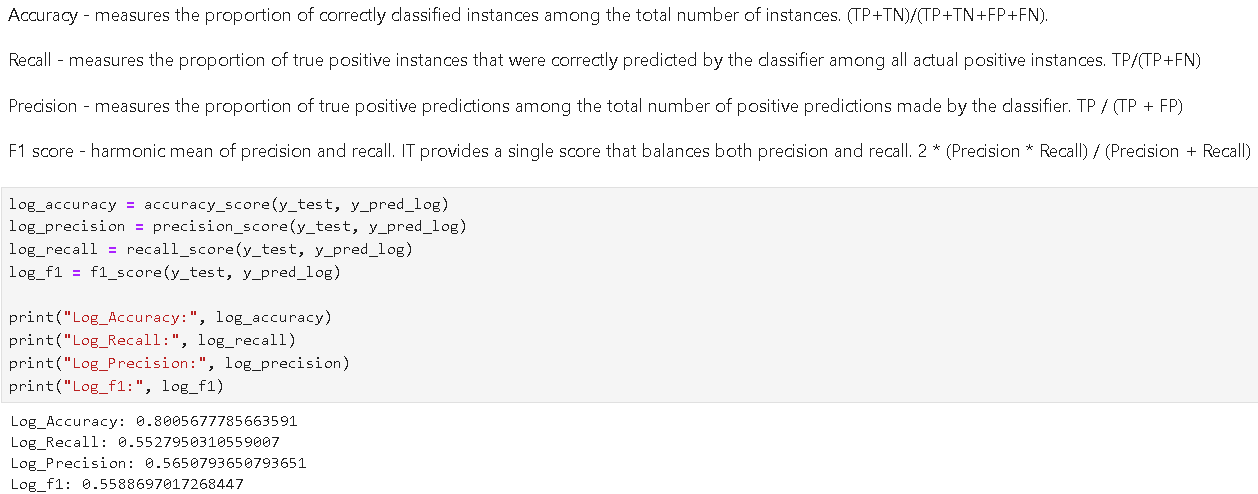
****

After splitting the data into training and testing sets, we applied the three models on the data one by one. The models we picked for analysis are Logistic Regression, Random Forest and XGBoost.

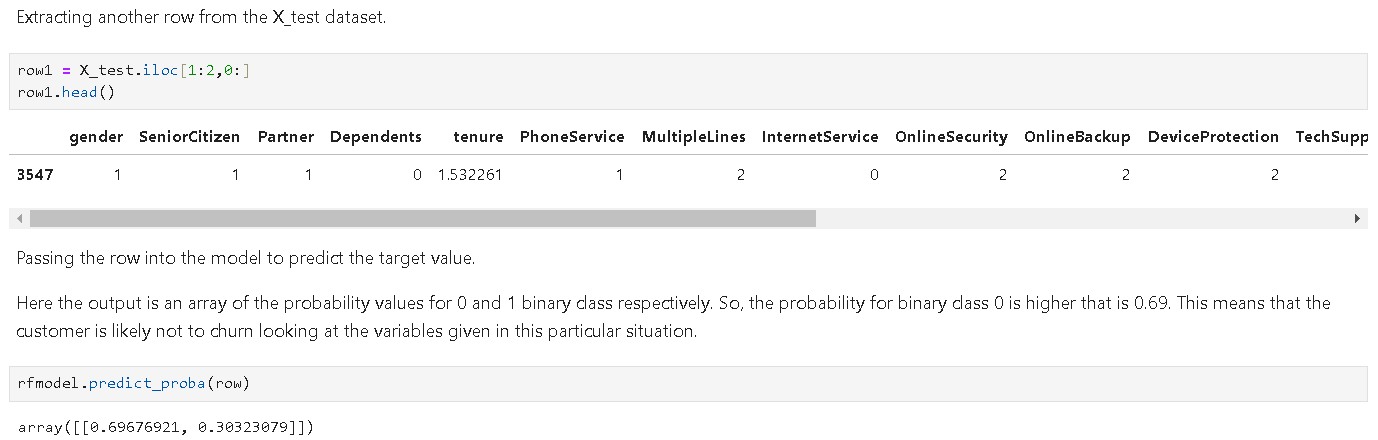
****

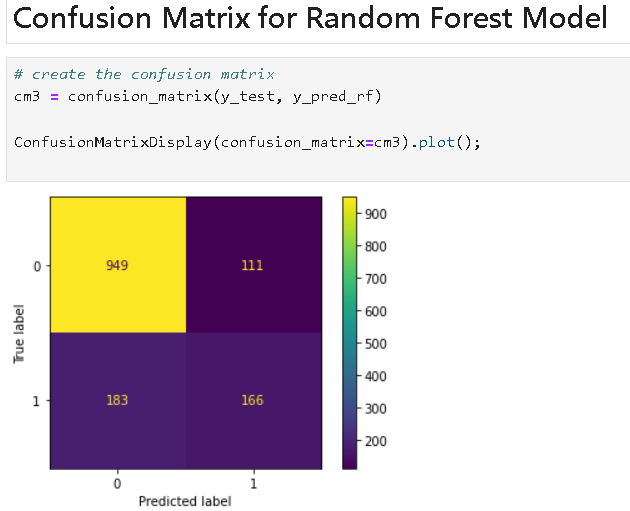
****

****

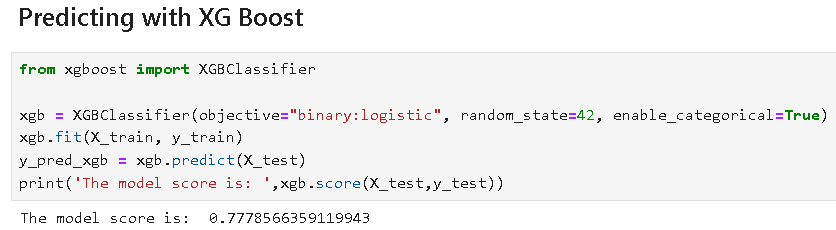
****

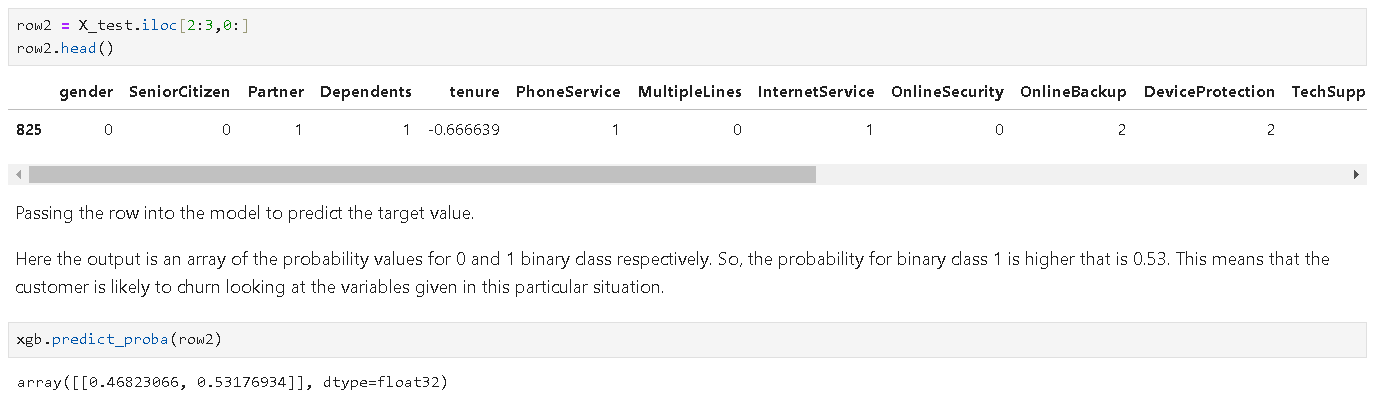
****

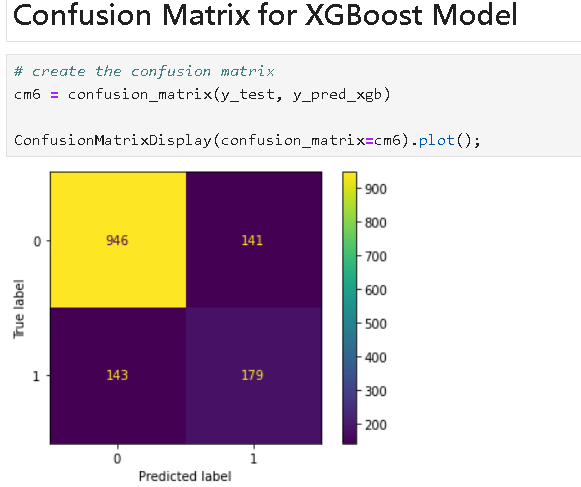
****

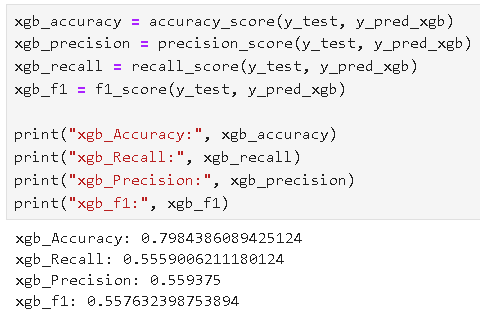
****

****

****

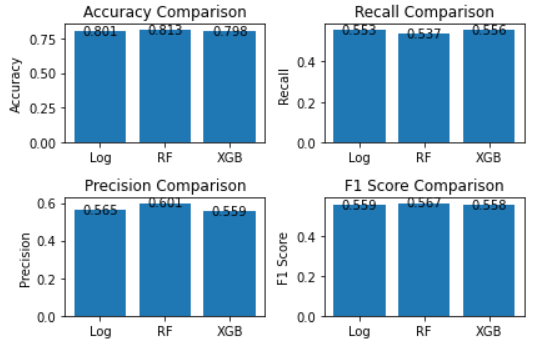
****

****

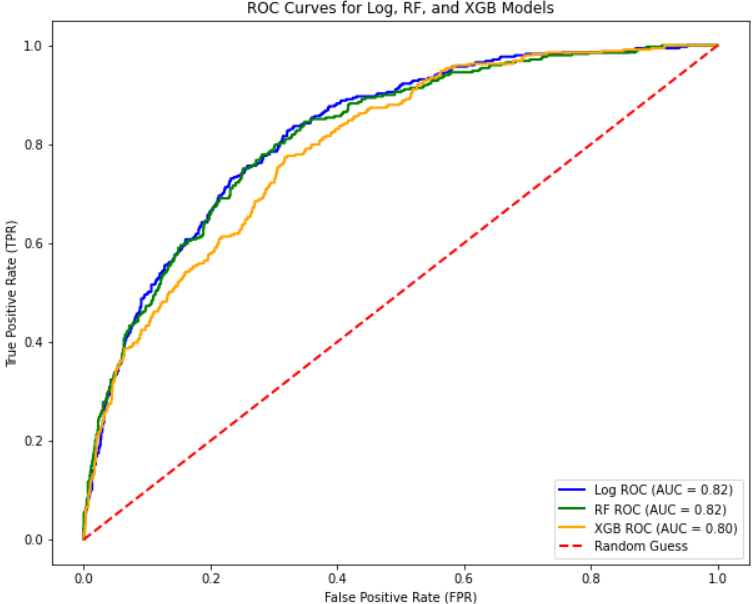
****

Below is a comparative graph showing the results from all the three models and corresponding ROC curve.

XGBoost excels with the highest accuracy (0.798) and recall (0.798), indicating it correctly identifies most instances and avoids false negatives. Random Forest boasts the best precision (0.601), meaning its positive predictions are often true, but it misses some actual positives compared to XGBoost.



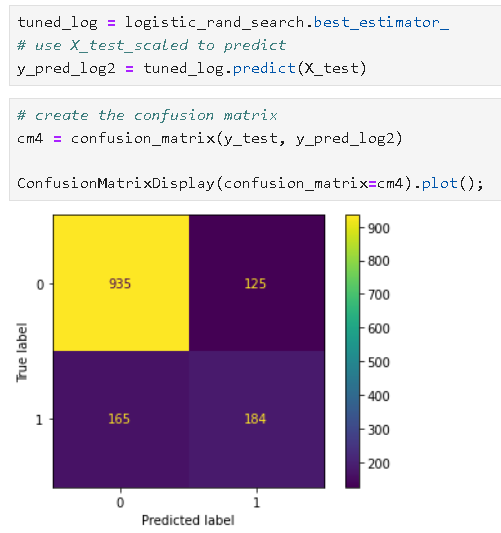
In the output below, the XGB model has the highest AUC (Area Under the Curve), at 0.82, which indicates it performs better at distinguishing between positive and negative classes compared to the other two models. Although both Log and RF models have an AUC of 0.80, the curve for the XGB model suggests it achieves a better balance between True Positive Rate (TPR) and False Positive Rate (FPR) across all thresholds.

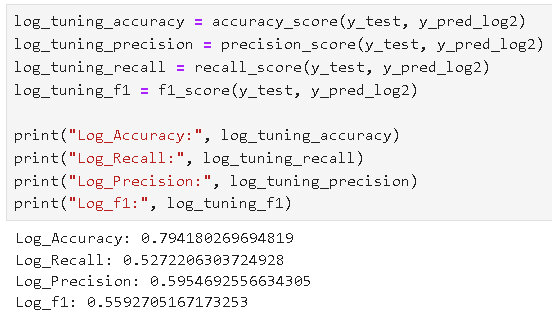


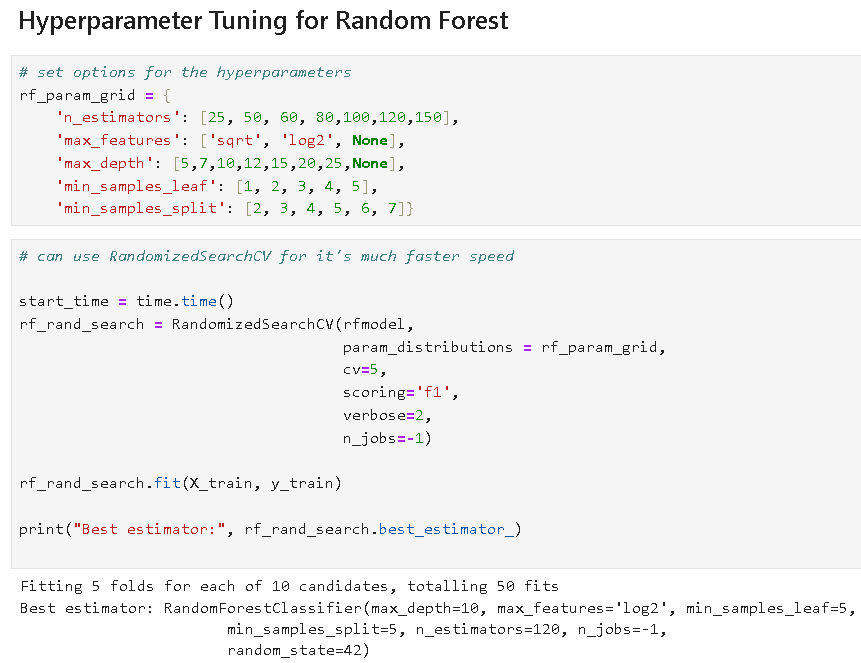
**Hyperparameter Tuning:**

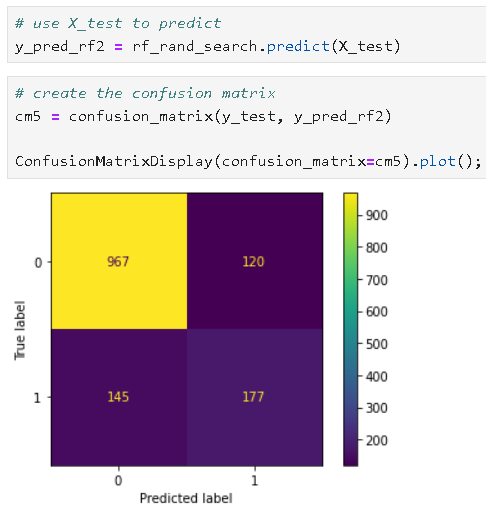
The code below tunes hyperparameters (settings) for a logistic regression model. It uses RandomizedSearchCV to try various combinations from defined ranges (param\_distributions) and pick the best based on a scoring metric (scoring). The code creates a RandomizedSearchCV object with the model, hyperparameter options, validation folds (cv), and number of parallel jobs (n\_jobs) to find the optimal settings for the model.

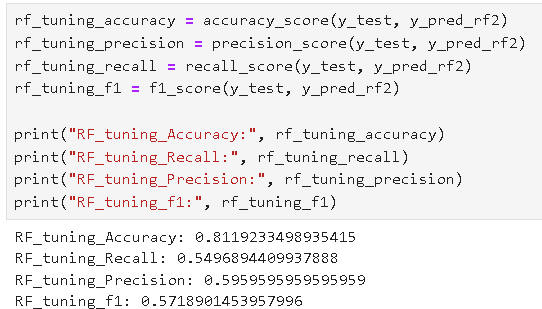
****

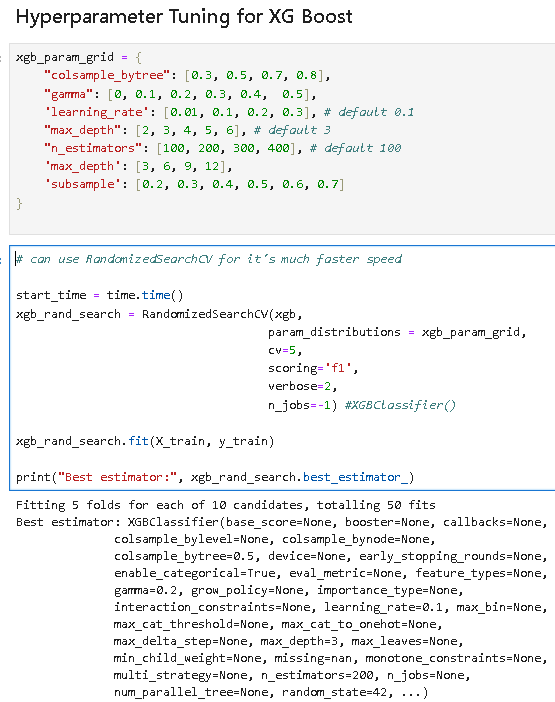
****

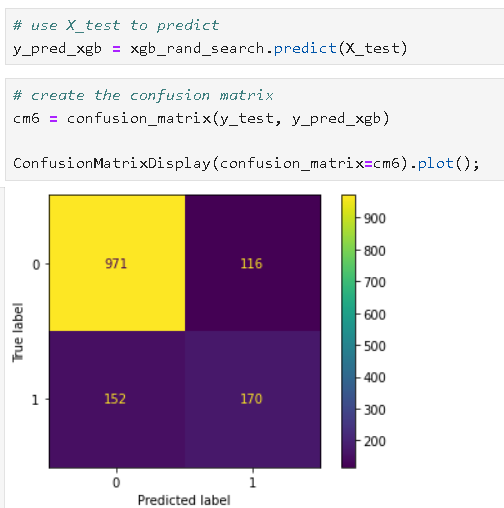
****

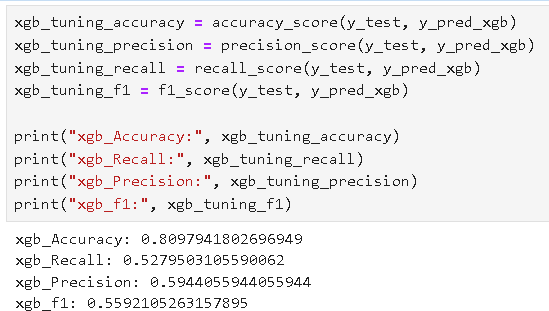
****

****

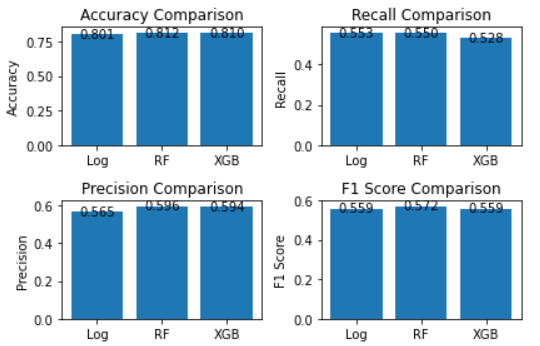
****

****

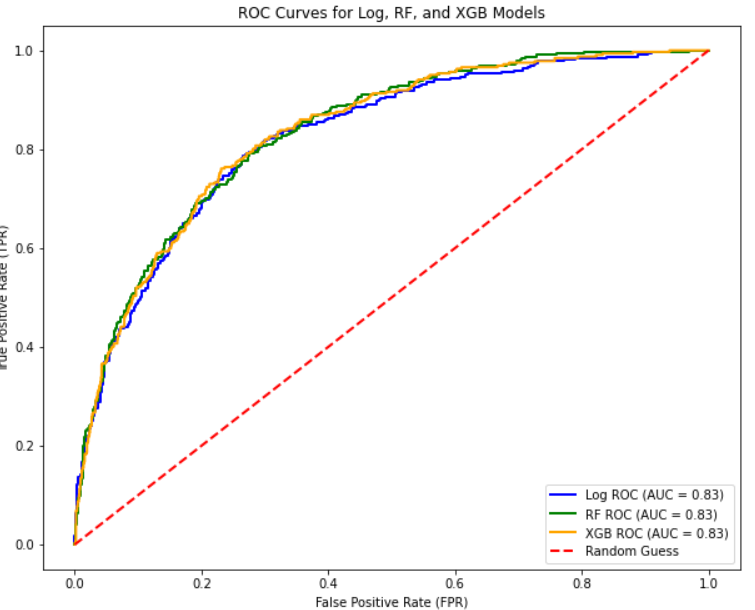
****

****

The results compare three models (Logistic Regression, Random Forest, XGBoost) for classification. XGBoost shines with the highest accuracy (0.798) and a balanced F1 score (0.687), indicating it excels at correctly identifying cases while avoiding false positives and negatives. Random Forest boasts the best precision (0.601) but misses some true positives.

****

The ROC curves show similar performance (AUC = 0.83) for all three models (Logistic Regression, Random Forest, XGBoost) in distinguishing positive from negative cases. While AUC provides a good overall measure, the actual curves can reveal differences. Here, a closer look at the curves themself is necessary to identify any potential variations in the balance between True Positive Rate (TPR) and False Positive Rate (FPR) across different classification thresholds.



**Conclusion:**

In conclusion, the telecom customer churn project effectively analyzes customer data to predict and prevent churn. By employing machine learning techniques like Logistic Regression, Random Forest, and XGBoost, the project identifies key factors influencing churn and provides actionable insights for businesses. Random forest emerges as the top-performing model, showcasing its efficacy in accurately predicting churn and enabling companies to implement targeted retention strategies. Overall, the project equips businesses with valuable tools to safeguard revenue and foster long-term customer relationships in the competitive telecom industry.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***