Name: Favour Ihechim Jacob

Project Title: Customer Churn Classification for a Telecom Company

i. **Data Analysis Report:** In the course of the project, an exploratory analysis was conducted on the data (https://www.kaggle.com/datasets/blastchar/telco-customer-churn) and the following were the outcome:

a) Data distribution: the dataset for XYZ Telco firm comprised of 21 columns and 7043 rows (7,042 customers, 1 row represent the customer metadata). The dataset housed information of her customers and their attributes such as CustomerID, Gender, Partner, Dependents, SeniorCitizens etc. The following observations were made from the dataset:

i. The gender distribution of XYZ's customer base is nearly balanced with a total of 50.5% as Male and 49.5% as female (this trend plays out again for customers who churned by gender)
ii. Nearly 26% of customers who churned are elderly (SeniorCitizens) others are not.
iii. Nearly 36% of customers who churned have partners others do not.
iv. Nearly 18% of customers who churned have dependents others do not.
v. All 11 of null values representing 11 customers did not churn but for a comprehensive analysis, the records were dropped.
vi. 1,869 customers did churn whereas 5,163 customers were retained.

b) Data types: Overall, the data explored is a mix of numerical (integer and float data types) and categorical data (objects). The dataset has three(3) data types as follows:

Float64 (1 column)
Int64 (2 columns)
Object (18 columns)

NB: The **TotalCharge** column was converted from an object to numeric data type as money is usually numeric and not otherwise – in real world scenario.

c) Missing values: The dataset was quite clean; there were no duplicates/a redundant record although, there were 11 nulls/missing values in the **TotalCharge** column that corresponded to 11 customers. Interestingly, all customers within this category did not churn. For the sake of regularizing the data, affected rows were dropped.

d) Outliers: based on the result from the scatterplot plotted and other metrics explored, there was no observation that posed as an outliers - all observations were clustered and had normal distances from each other.

ii. **<u>Data Visualization Report:</u>** following the completion of the data analysis phase, the analysed data was plotted in several forms (scatter plots, box plots, pie charts, histograms) to gain insights from the data. The several plots and visuals gave answers to various pertinent questions asked in the course of the project. Questions like, but not limited to:

a) How many customers have churned and what is the churn rate?
b) What is the gender distribution like?
c) What are the demographics of the customers who have churned?
d) Did gender affect the rate of churn?
e) Are there any difference in churn rate between customers with partners and those without? How about dependents?
f) Is there any particular service associated with churn?
g) Are customers on long term contract less likely to churn?

iii. **<u>Classification Model Report:</u>** following the analysis and the visualization phases where the problem was well understood, two models were developed:

i. Logistics Regression Model
ii. Support Vector Machine (SVM) Model).

The performance of both models was compared with the latter outperforming the former slightly in terms of the metrics used for comparison – **accuracy score, precision score, recall score and the F1 score** (kindly see comparison in attached jupyter file). In both cases, the dataset was split into the train and test subsets with the train set, set to 80% of the entire dataset.

iv. **<u>Final Report:</u>**

a. **Problem Statement:** The Telco Customer Churn is a focused customer retention program targeted at predicting customer behaviours to retain customer.

b. **Dataset Description:** The dataset consist of customer's data for XYZ Telco firm with each row representing a unique customer; the columns represent the various customers' attributes such as a unique identifier (CustomerID), sex (Gender) etc. In total, there are 21 columns and 7,043 rows.
The dataset includes information about:

i. Customers who left within the last month – the column is called Churn.
ii. Services that each customer has signed up for – Phone, MultipleLines, Internet, OnlineSecurity, OnlineBackup, DataProtection, TechSupport, StreamingTV and Movies

iii. Customer account information – how long they have been a customer, Contract, Payment Method. PaperlessBilling, MonthlyCharges and TotalCharges.
iv. Demographic information about cusstomers – Gender, AgeRange and if they have Partners and Dependents.

Conclusion

From the insight gained from the data, the following was deduced:

i. Gender does not affect the churn
ii. Most young and single customers (with no partners and dependents) happen to churn meaning they experience some level of dissatisfaction.
iii. Customers without certain services (tech support, robust storage, online security, data protection) are more likely to churn.
iv. Most customers that churned use electric check.
v. Long term contract customers are less likely to churn compared to the others.

Recommendation

XYZ firm needs to come up with favourable policies that will improve her efficiency and minimize the rate of churn along the lines of the highlighted conclusions.