**Project Outputs**

Data frame = SURdf (There in the CODE file)

Head(first 6 rows and cols data)

A screenshot of a computer screen

Description automatically generated

Dimensions of data (No of ROWS,COLS in Data)

A white background with black text

Description automatically generated

Structure of Data frame (shows COLUMN names and Data types)

A screenshot of a computer

Description automatically generated

Summary (For all COLUMNS, gives All statistical details like Mean,Mode,IQR(Inter Quartile Range),Median)

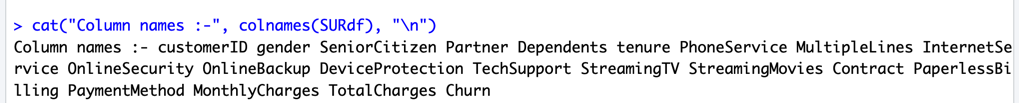
A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Displaying all COLUMN names



Data Type of each COLUMN

A close-up of a computer screen

Description automatically generated

Rough / Indirect Check of Missing Values

A graph with text overlay

Description automatically generated

Dropping the “customerID” column

A screenshot of a computer screen

Description automatically generated

* On deep analysis, we can find some indirect missingness in our data (which can be in form of blankspaces). Let's see that!

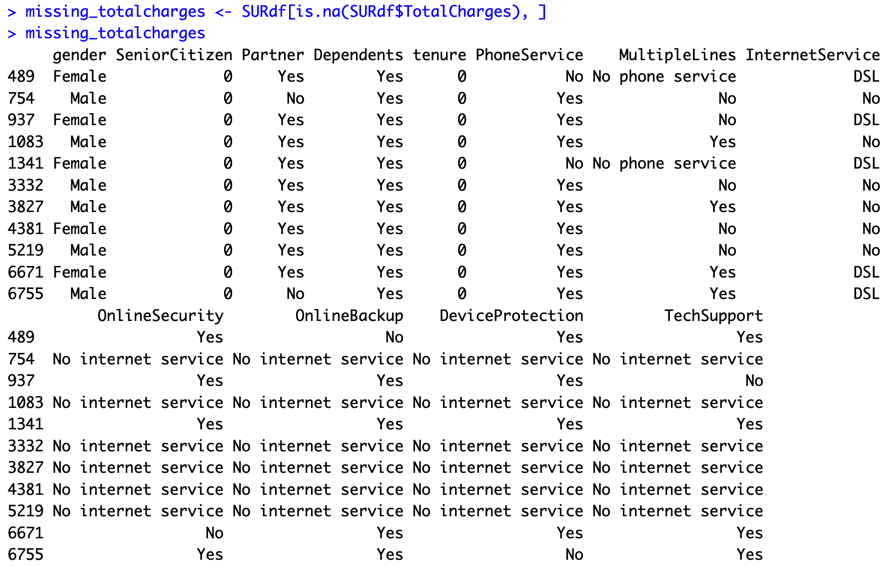
Checking missing values in each column

A screenshot of a computer

Description automatically generated

* Here we see that the TotalCharges has 11 missing values. Let's check this data.

Filtering rows where ‘TotalCharges’ is NA and Viewing rows with missing ‘TotalCharges’



A screenshot of a computer screen

Description automatically generated

* It can also be noted that the Tenure column is 0 for these entries even though the MonthlyCharges column is not empty. Let's see if there are any other 0 values in the tenure column.

Get the indices of rows where 'tenure' is equal to 0

A close up of numbers

Description automatically generated

* There are no additional missing values in the Tenure column. Let's delete the rows with missing values in Tenure columns since there are only 11 rows and deleting them will not affect the data.

Drop rows where 'tenure' is equal to 0

A close-up of a math problem

Description automatically generated

To solve the problem of missing values in TotalCharges column, I decided to fill it with the mean of TotalCharges values.

A screenshot of a computer screen

Description automatically generated

Count the number of missing values (NA) in each column

A close-up of several words

Description automatically generated

Map values in the 'SeniorCitizen' column: 0 -> "No", 1 -> "Yes"

A screenshot of a computer screen

Description automatically generated

Summarize the 'InternetService' column

A close-up of words

Description automatically generated

Summarize the specified numerical columns

A screenshot of a number of numbers

Description automatically generated

Data Visualization

PLOT – 1 A and B

A pie chart with text on it

Description automatically generated A pie chart with text on it

Description automatically generated

* 26.6 % of customers switched to another firm.
* Customers are 49.5 % female and 50.5 % male.

Gender count where Churn is NO

A white background with blue text

Description automatically generated

Gender count where churn is YES

A white background with blue text

Description automatically generated

A diagram of a pie chart

Description automatically generated

* There is negligible difference in customer percentage/ count who chnaged the service provider. Both genders behaved in similar fashion when it comes to migrating to another service provider/firm.

A graph of different colored bars

Description automatically generated

* About 75% of customer with Month-to-Month Contract opted to move out as compared to 13% of customers with One Year Contract and 3% with Two Year Contract.

A graph of a customer payment method

Description automatically generated

* Major customers who moved out were having Electronic Check as Payment Method.
* Customers who opted for Credit-Card automatic transfer or Bank Automatic Transfer and Mailed Check as Payment Method were less likely to move out.

Get unique values from the 'InternetService' column

A close-up of a sign

Description automatically generated

Filter rows where gender is "Male" and count combinations of InternetService and Churn

A screenshot of a computer

Description automatically generated

Filter rows where gender is "Male" and count combinations of InternetService and Churn

A screenshot of a computer program

Description automatically generated

A graph of different colored bars

Description automatically generated

* A lot of customers choose the Fiber optic service and it's also evident that the customers who use Fiber optic have high churn rate, this might suggest a dissatisfaction with this type of internet service.
* Customers having DSL service are majority in number and have less churn rate compared to Fibre optic service.

A purple and white lines

Description automatically generated with medium confidence

* Customers without dependents are more likely to churn

A chart with orange and green lines

Description automatically generated

* Customers that doesn't have partners are more likely to churn

A screenshot of a graph

Description automatically generated

* Most customers churn in the absence of online security,

A red and yellow lines

Description automatically generated

* Customers with Paperless Billing are most likely to churn.

A screenshot of a graph

Description automatically generated

* Customers with no TechSupport are most likely to migrate to another service provider.

A green screen shot of a green screen

Description automatically generated

* Very small fraction of customers don't have a phone service and out of that, 1/3rd Customers are more likely to churn.

A graph of a graph showing the amount of charge

Description automatically generated

* Customers with higher Monthly Charges are also more likely to churn

A graph showing the distribution of charge

Description automatically generated

 **High Total Charges customers are less likely to churn**, likely due to longer tenure or satisfaction with services.

 **Churn is more frequent among customers with lower Total Charges,** highlighting a potential area to focus on improving customer retention strategies for new or low-value customers.

A graph of a diagram

Description automatically generated with medium confidence

* New customers are more likely to churn

A graph of numbers and a graph

Description automatically generated with medium confidence

Strong correlations between tenure, MonthlyCharges, and TotalCharges suggest that longer-tenure customers contribute significantly to revenue.

Lower tenure and shorter Contract types are the strongest predictors of churn, emphasizing the need to focus on retaining short-term and new customers.

Encouraging longer contracts and addressing dissatisfaction among customers with short tenure or month-to-month plans can help reduce churn.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer code

Description automatically generated

A graph showing a distribution of the amount of charge

Description automatically generated with medium confidence

A graph with green lines

Description automatically generatedA graph showing the distribution of electricity

Description automatically generated

Since the numerical features are distributed over different value ranges, I will use standard scalar to scale them down to the same range.

A graph with a line

Description automatically generatedA graph showing a distribution curve

Description automatically generatedA graph with a line

Description automatically generated

A computer screen shot of a computer program

Description automatically generated

A person holding a sign

Description automatically generated with medium confidence

A close up of a logo

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a graph

Description automatically generated

**Machine Learning Model Evaluations and Predictions**

KNN Algorithm

A close-up of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

SVM Algorithm

A computer code with blue text

Description automatically generated

A screenshot of a computer

Description automatically generated

RANDOM FOREST Algorithm

A computer screen shot of a program

Description automatically generated

A close-up of blue text

Description automatically generated

A screenshot of a computer

Description automatically generated

Random forest Confusion Matrix

A blue squares with white text

Description automatically generated

Random Forest ROC Curve

A graph with a line drawn on it

Description automatically generated

Logistic Regression Algorithm

A computer code with blue text

Description automatically generated

The dataset is **imbalanced**, with approximately 74% of the samples belonging to class 0 and only 26% to class 1.

A computer screen shot of a program

Description automatically generated

A screenshot of a computer code

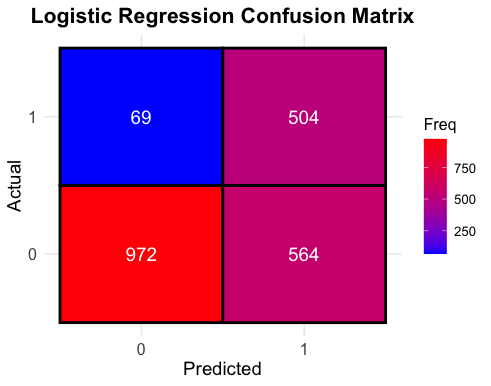
Description automatically generated

Logistic Regression CLASSIFICATION REPORT

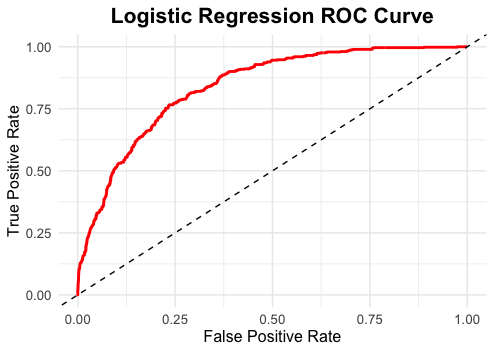
A screenshot of a computer

Description automatically generated

Logistic Regression CONFUSION MATRIX



Logistic Regression – ROC Curve

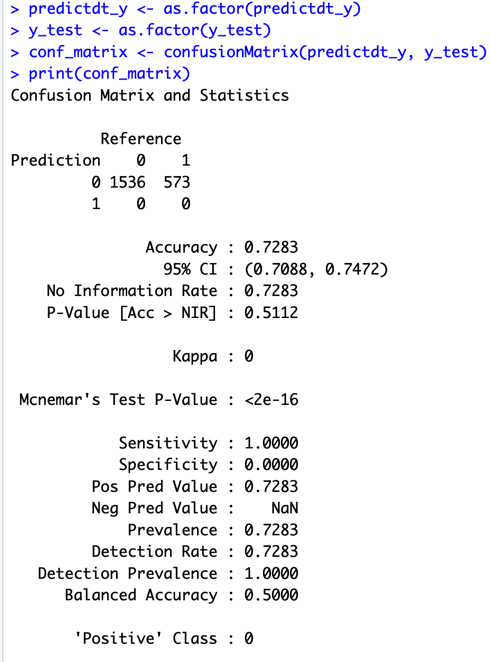


DECISION TREE

A computer code with blue text

Description automatically generated

Decision Tree - CLASSIFICATION REPORT



XG BOOST CLASSIFIER

A close-up of a code

Description automatically generated

XGBOOST Classification Report

A screenshot of a computer screen

Description automatically generated

XGBOOST Confusion Matrix

A diagram of a confusion matrix

Description automatically generated