**Predicting Customer Churn in a Telecommunications Company**

Name: - Raunak Raj

Registration Number: - 12002617

Course: - Integrated B. Tech M. Tech Computer Science and Engineering

College: - Lovely Professional University

**Objective: -** The primary objective of this project is to develop a predictive model that can identify customers at risk of churning, enabling the company to take proactive measures to retain them.

**About Data Set: -** The dataset, sourced from Kaggle, comprises 21 columns and 7,044 entries. Key features, including gender, senior citizen status, partner status, dependent status, tenure, and churn, will be utilized for both data visualization and the development of a predictive model.

**Data Pre-processing: -**

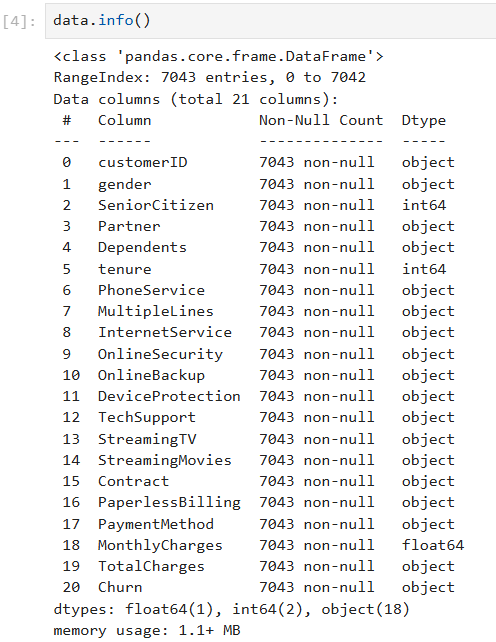
1. Firstly, the data.info() command was utilized to determine the column names and their respective data types. The results of this command are illustrated in Figure 1.

Figure 1: - Result of data.info() command

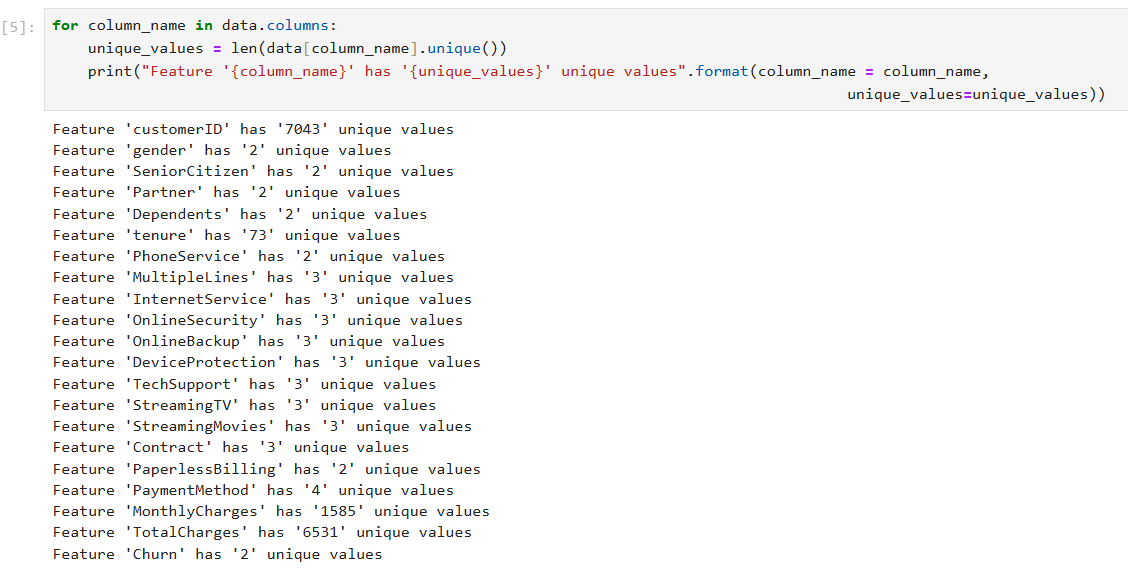
1. Subsequently, the unique values for each feature were examined using the unique() command, with the output displayed in Figure 2.

Figure 2: - Unique Values Associated with Each Features

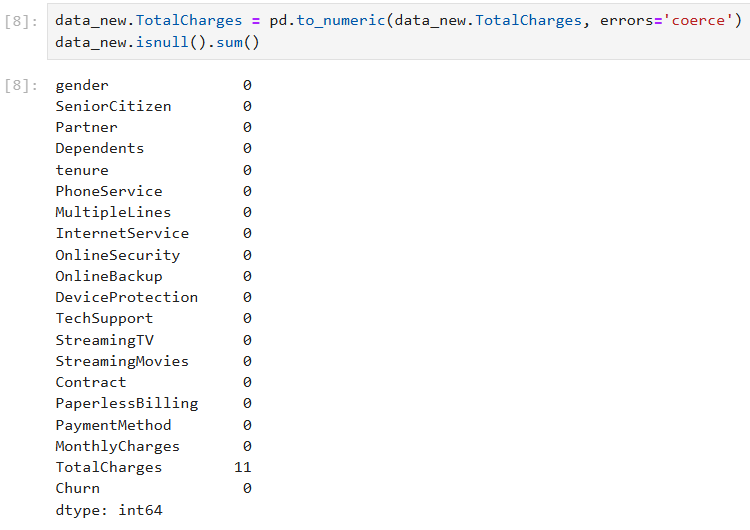
1. Next, the presence of null values in each feature was assessed using the isnull().sum() command. As depicted in Figure 3, the "TotalCharges" feature contained 11 null values. These 11 null entries were then removed from the dataset using the dropna() command.

Figure 3: - Null Values Associated with Each Features

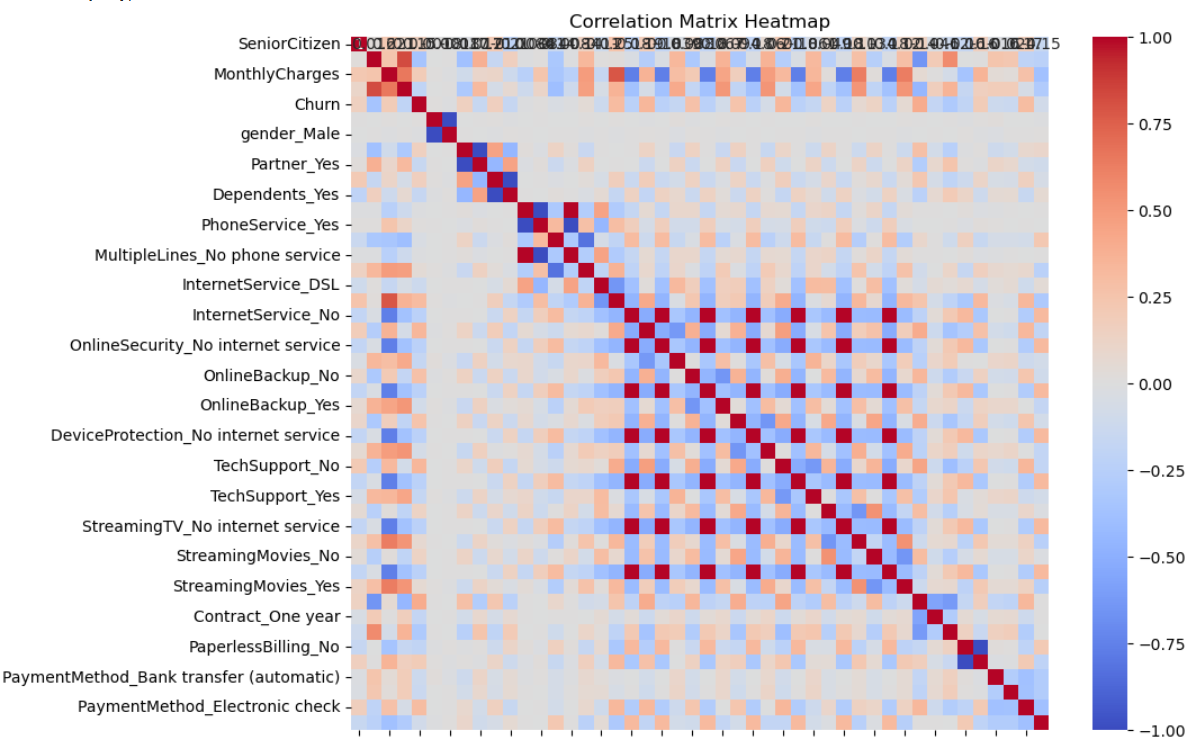
**Correlation Matrix: -** A correlation matrix was generated using the corr() method, focusing on the "Churn" column. Subsequently, a heatmap representing the correlation matrix was created utilizing the matplotlib.pyplot library. The heatmap is depicted in Figure 4.

Figure 4: - Correlation Matrix Heatmap

**DATA VISUALIZATION**

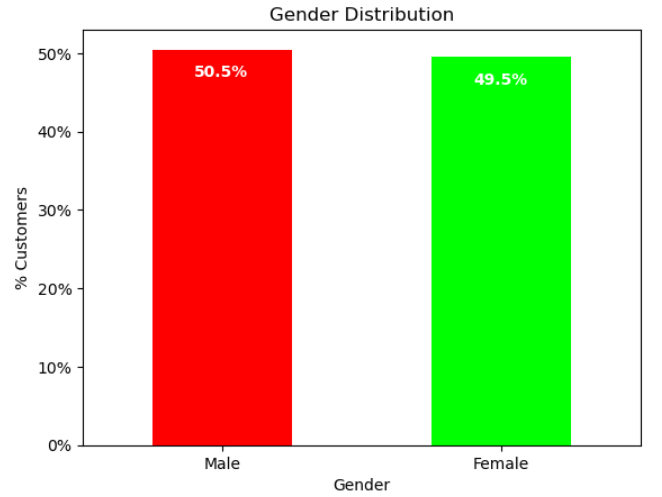
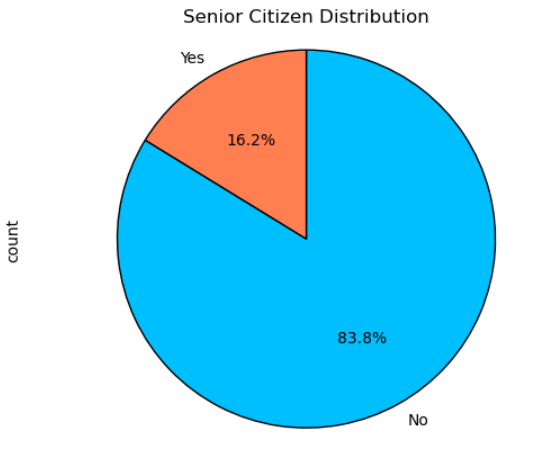
* **GENDER DISTRIBUTION: -** The analysis of the dataset revealed that 50.5% of the employees in the organization are male, while 49.5% are female. The analysis was done with the help of Bar Chart (Figure 5).

Figure 5: - Gender Distribution

* **SENIOR CITIZEN DISTRIBUTION: -** The analysis of the dataset revealed that 16% of the employees in the organization falls are senior citizens, while 83.8% are young people. The analysis was done with the help of Pie Chart (Figure 6).

Figure 6: - Senior Citizen Distribution

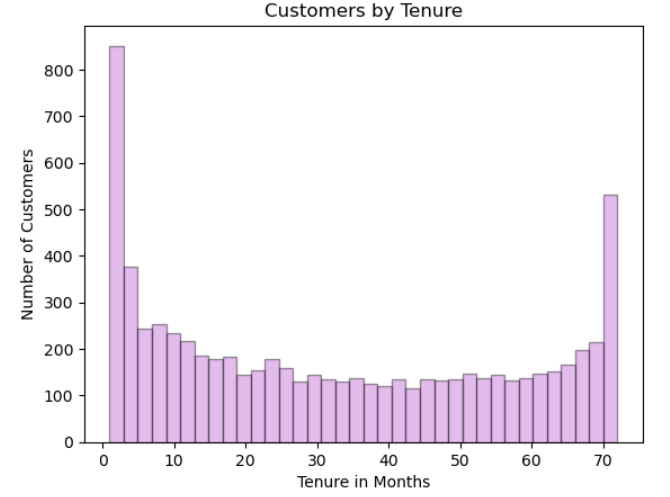
* **CUSTOMER BY TENURE: -** Upon examining the histogram (Figure 7), it is evident that a significant number of customers have been with the telecom company for just one month, while a considerable number have remained for approximately 72 months.

Figure 7: - Customers by Tenure

* **CUSTOMER BY CONTRACT: -** The graph (Figure 8) indicates that the majority of customers are on month-to-month contracts. Additionally, there are an equal number of customers with 1-year and 2-year contracts.

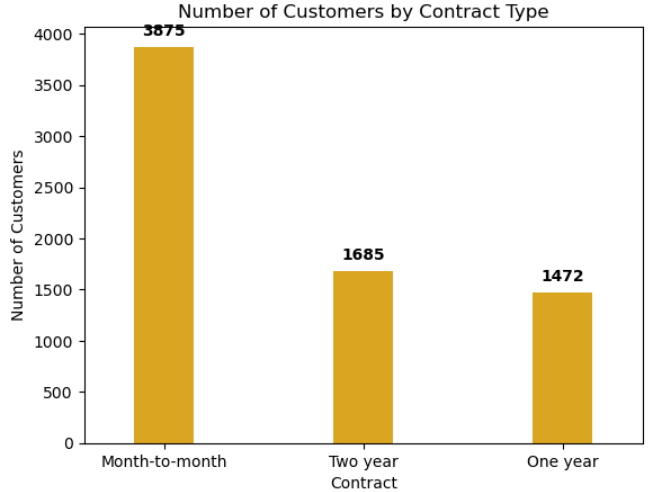


Figure 8: - Customers by Contract Type

* **CHURN RATE: -** The graph (Figure 9) indicates that the Churn Rate. According to the figure 73.4% customers don’t churn.

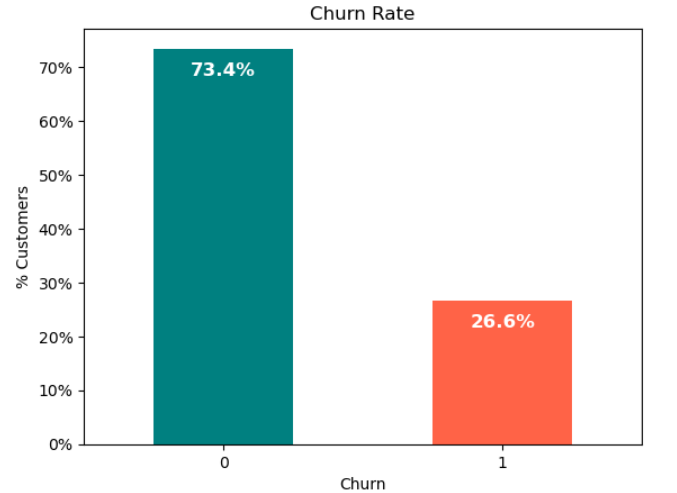


Figure 9: - Churn Rate

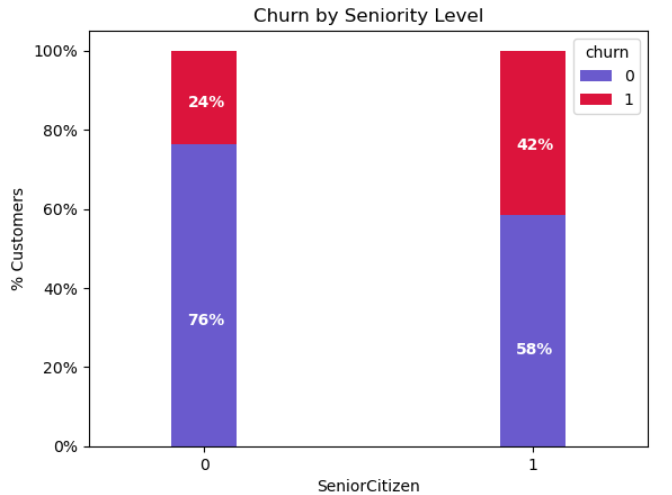
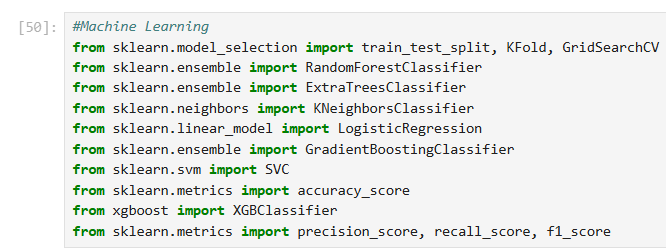
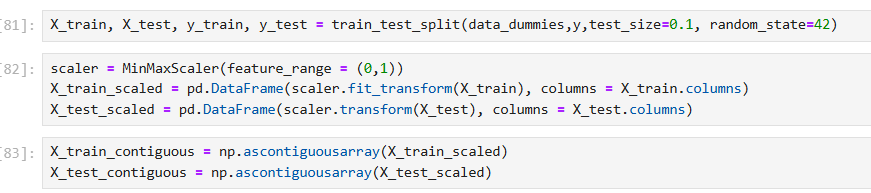
* **CHURN RATE BY SENIORITY: -** The graph (Figure 10) indicates that the Churn Rate by seniority. Senior citizens exhibit nearly twice the churn rate compared to the younger population.

Figure 10: - Churn Rate by Seniority

**MACHINE LEARNING AND PREDICTION MODEL**



Figure 11: - Importing Libraries

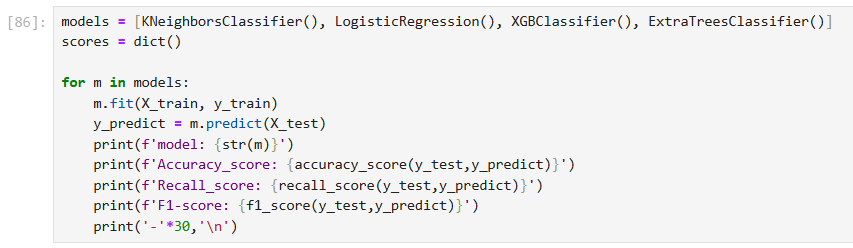
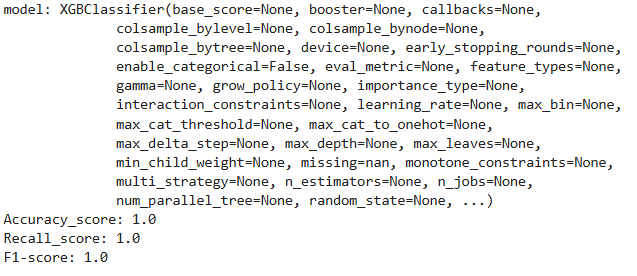
Figure 12: - Splitting The Data into Training and Testing Sets

Figure 13: - ML Algorithms

* Four machine learning algorithms were applied to the dataset: K-Nearest Neighbors (KNN), Logistic Regression, XGBoost Classifier (XGBClassifier), and Extra Trees Classifier (ExtraTreesClassifier). The accuracy, recall score, and F1 score of these models are presented in Figures 14, 15, and 16, respectively.

Figure 14: - Results of KNN and Logistic Regression



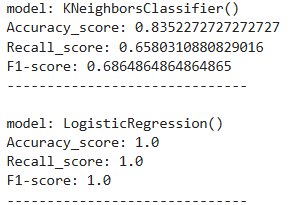


Figure 15: - XGB Classifier Results

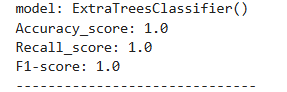


Figure 16: - Extra Trees Classifiers

* **CHALLENGES FACED: -**

1. Data Quality: - Ensuring that the dataset is clean, relevant, should not contain missing values, outliers and noise. This can compromise the model’s performance. From Figure 3 we can clearly tell that “TOTAL CHARGES” features had 11 null values.
2. FEATURE SELECTION: - Creating the model with most relevant features was complex and time consuming. “Customer ID” feature/column was discarded initially as it was not correlated with the “CHURN” hence it could have led to model not performing optimal.
3. ML ALGORITHMS: - Four machine learning algorithms were applied to the dataset: K-Nearest Neighbors (KNN), Logistic Regression, XGBoost Classifier (XGBClassifier), and Extra Trees Classifier (ExtraTreesClassifier). The accuracy, recall score, and F1 score of these models are presented in Figures 14, 15, and 16, respectively. Notably, only the KNN algorithm yielded good values, while the implementation of the other algorithms faced difficulties.