Project Report: Customer Churn Analysis and Prediction

Mukand Krishna

8/18/23

1. Introduction:

Customer churn refers to the phenomenon where customers stop using a company's products or services. It is a crucial concern for businesses, as losing customers can lead to decreased revenue and growth. In this report, we analyze customer churn in a telecom company and build predictive models to identify factors that contribute to churn. Our goal is to develop accurate models that can predict whether a customer is likely to churn or not, based on various features.

2. Objectives:

The objectives of this analysis are:

- Data Exploration and Visualization: We begin by exploring the dataset, understanding its columns, and visualizing patterns that might indicate factors influencing churn.
- Data Preprocessing: We clean and preprocess the data, handling missing values and converting categorical variables into a suitable format for analysis.
- Model Selection and Evaluation: We employ various machine learning models, such as Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Ada Boost, to predict customer churn. We evaluate these models using accuracy, recall, and precision metrics.
- Interpretation and Analysis: We interpret the results from the models, analyze their strengths and weaknesses, and discuss the significance of different features in predicting churn.

i. Data Cleaning:

Mean imputation is a technique used to fill in missing values in the 'TotalCharges' column of the dataset. Mean imputation involves replacing missing values with the mean (average) value of the available data in the same column. This approach assumes that the missing values are missing at random and that the mean value is a reasonable estimate for the missing entries.

Filling the missing values in TotalCharges column with the mean of TotalCharges values.

```
1 data.fillna(data["TotalCharges"].mean())
```

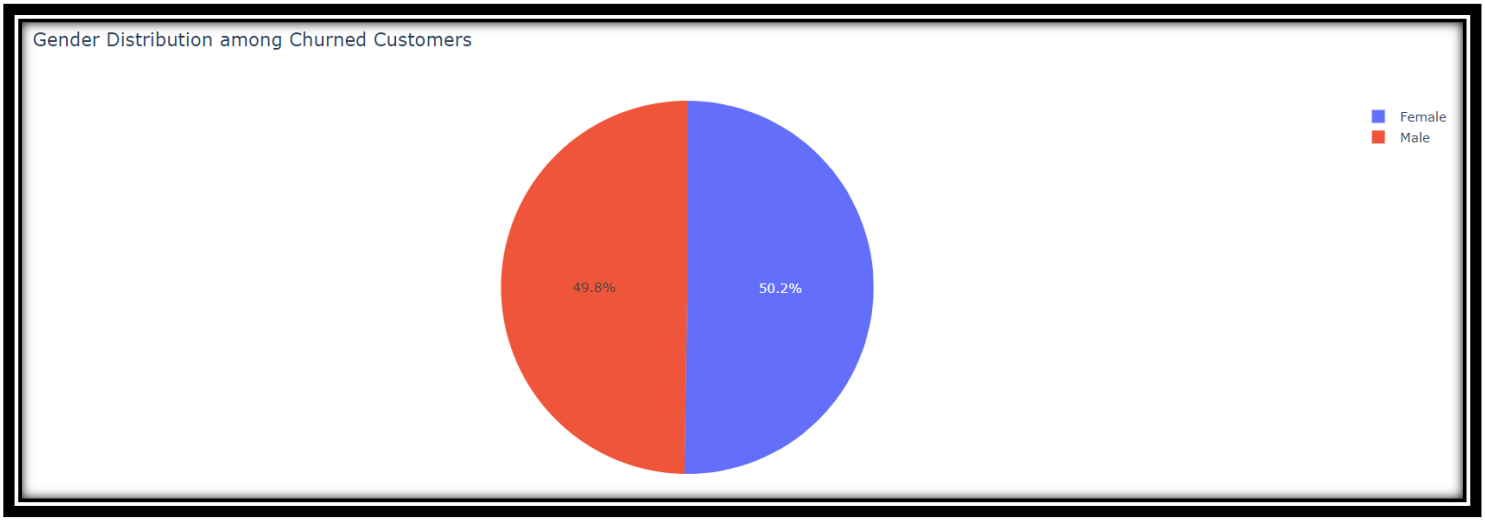
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	Mult
0	7590-VHVEG	Female	0	Yes	No	1	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	
3	7795-CFOCW	Male	0	No	No	45	No	
4	9237-HQITU	Female	0	No	No	2	Yes	
...	...	...	...	...	...	...	...	

```
1 data.isnull().sum()
```

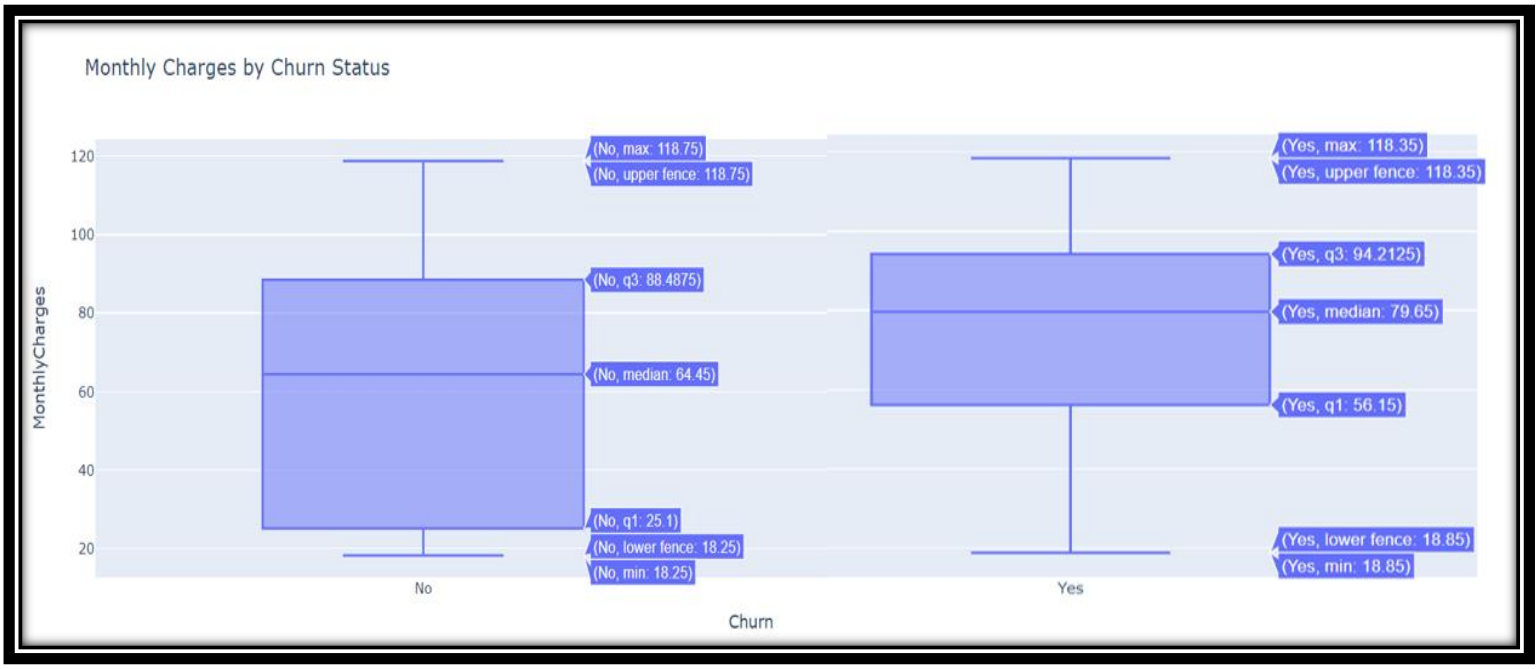
```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV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

ii. Data Exploration and Visualization:

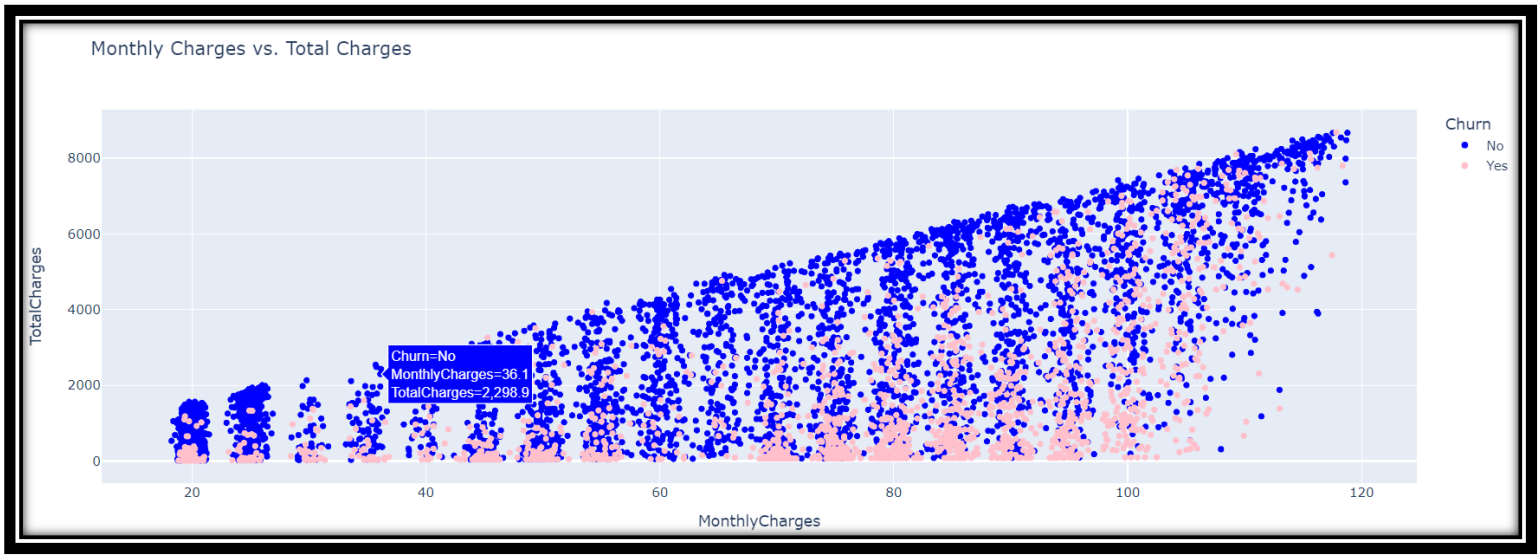
Pie Chart of Gender Distribution for Churned

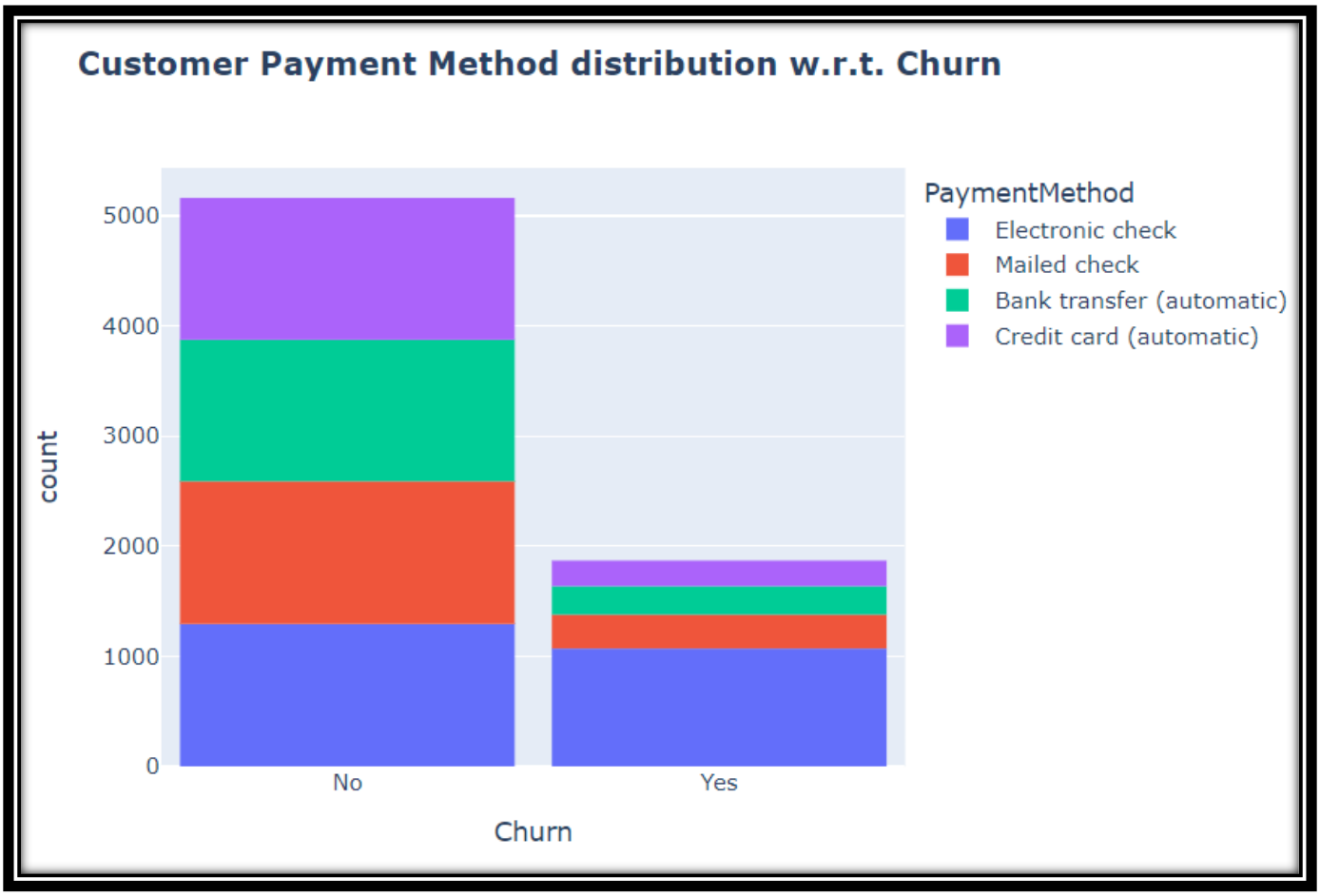


Box Plot of Monthly charges by Churn

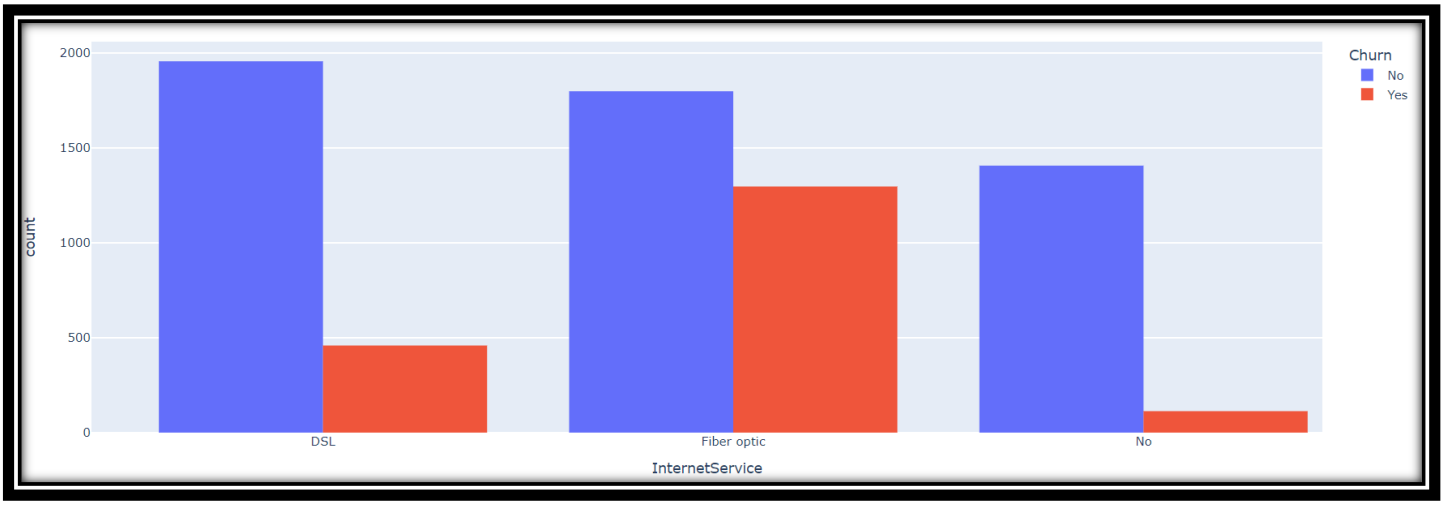


Scatter plot of Monthly vs Total Charges

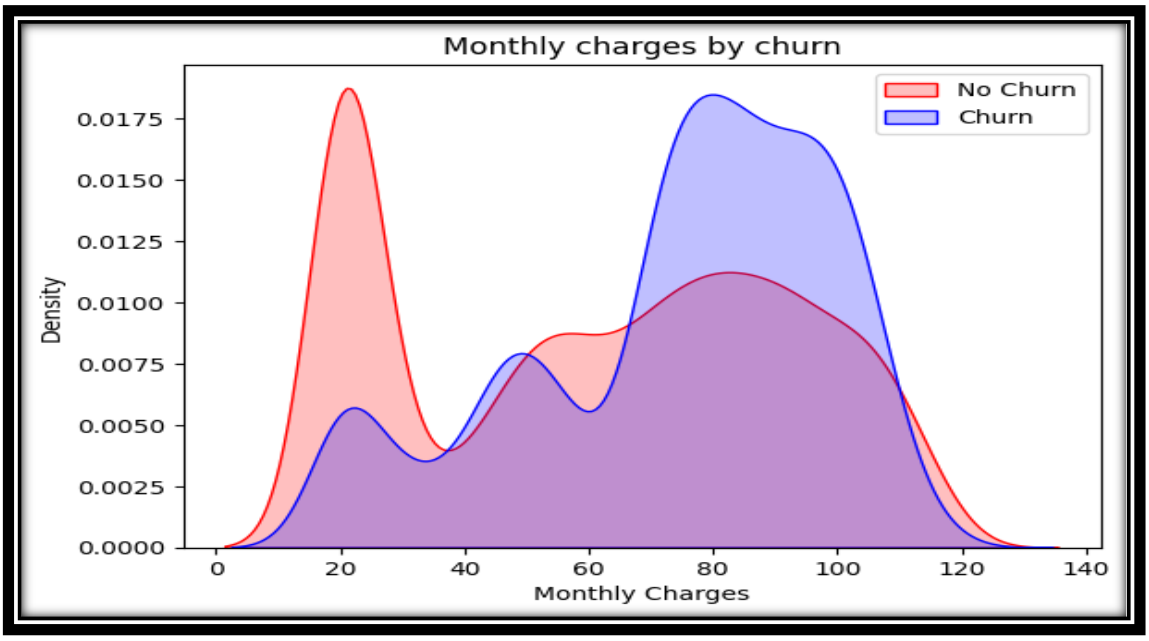


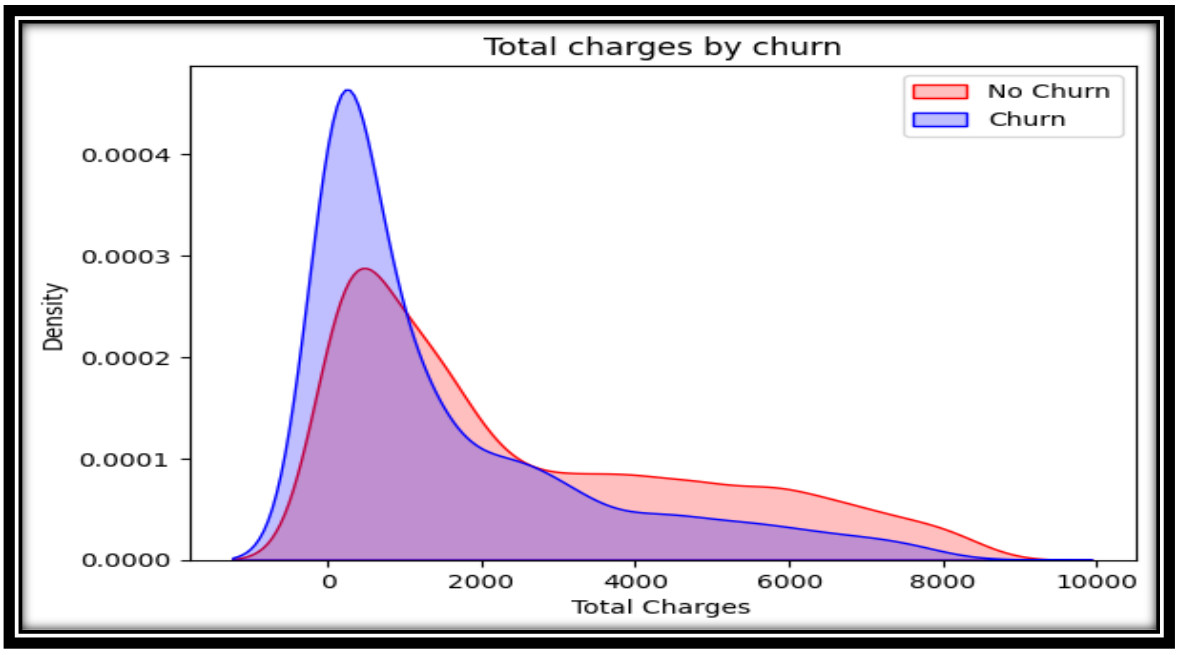


Bar plot of Internet Service

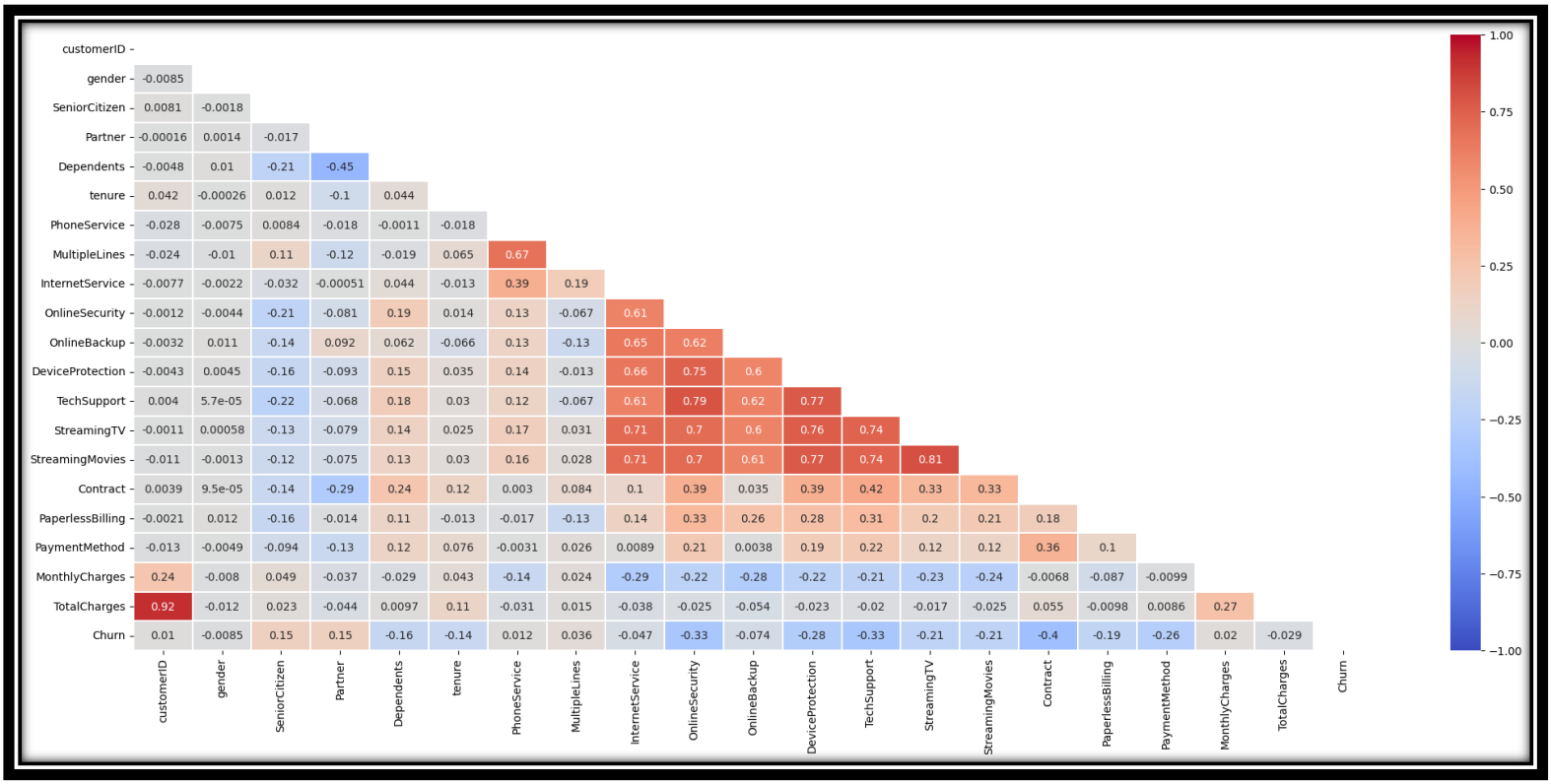


KDE Plots of Monthly and Total Charges





Determining how one variable changes in relation to another variable



Machine learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Ada Boost, **to predict churn**.

Evaluating each model's accuracy, recall, and precision to determine its performance in predicting churn.

```
Model: Support Vector Machine
Accuracy: 0.8078561287269286

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.92      0.87      1539
           1       0.70      0.50      0.59       574

    accuracy                           0.81      2113
   macro avg       0.77      0.71      0.73      2113
weighted avg       0.80      0.81      0.80      2113

================================================
```

```
Model: Logistic Regression
Accuracy: 0.8121154756270705

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.90      0.87      1539
           1       0.69      0.57      0.62       574

    accuracy                           0.81      2113
   macro avg       0.77      0.74      0.75      2113
weighted avg       0.80      0.81      0.81      2113

================================================
```

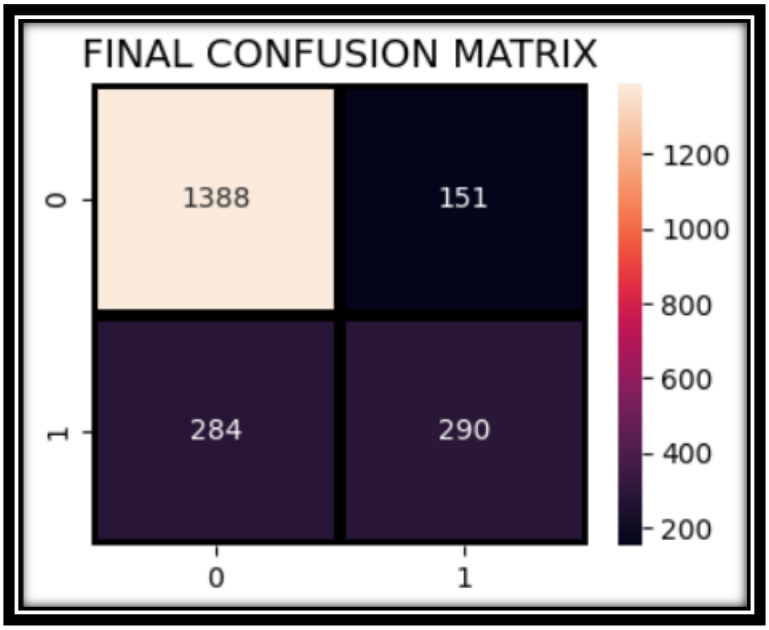
Confusion matrix is a tool used to evaluate the performance of predictive models for customer churn. This helps to understand how well the model's predictions align with the actual outcomes.

Consists of four values: **True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).**
TP: instances where the model correctly predicts churned customers.

TN: represents correct predictions of non-churned customers.

FP: instances where the model incorrectly predicts churn when it shouldn't.

FN: represents where the model fails to predict churn when it should.



There are total **1388+151=1539 actual non-churn values** and the algorithm predicts 1388 of them as non-churn and 151 of them as churn.

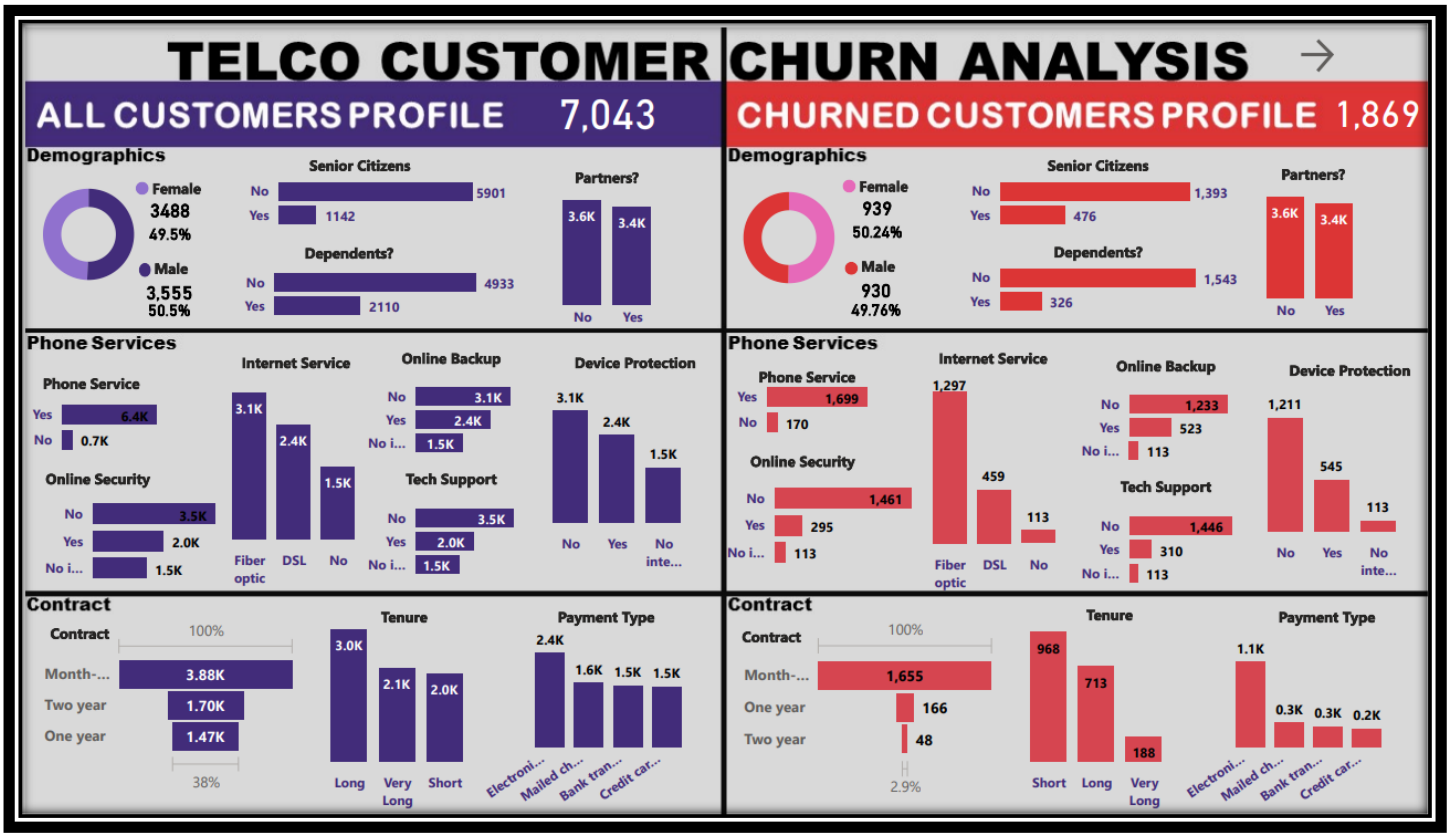
While there are **284+290=574 actual churn values** and the algorithm predicts 284 of them as non-churn values and 290 of them as churn values.

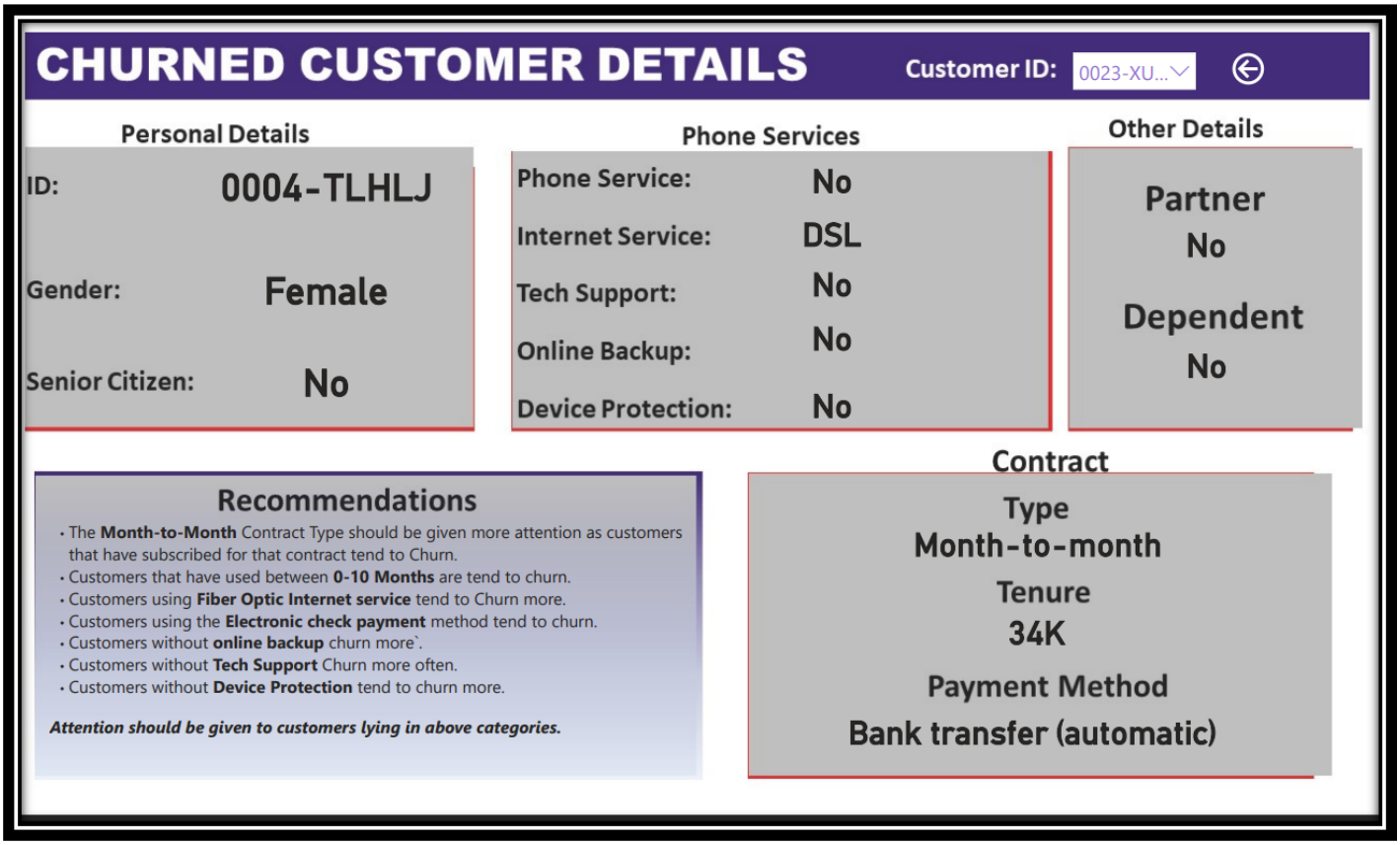
3. Results and Analysis:

- Gender distribution among churned customers shows a pattern where a larger proportion of female churn compared to male.
- Customers using electronic check as their payment method are more likely to churn.
- Higher Monthly Charges, lower tenure, and lower Total Charges contribute to higher churn rates.
- Among the models, Logistic Regression, Gradient Boosting, and Support Vector Machine achieve the highest accuracy and precision in predicting churn.

4. Power Bi Dashboard:

The dashboard includes interactive visualizations that allow users to explore the data and gain insights.





5. Conclusion:

In this project, we successfully explored and visualized customer churn patterns in a telecom company's dataset. By building and evaluating various machine learning models, we were able to predict customer churn with a significant degree of accuracy and precision. The insights from this analysis can help businesses understand factors contributing to churn and make informed decisions to retain valuable customers.

The **key takeaways** from our analysis are the identification of features strongly associated with churn and the recommendation to focus on reducing Monthly Charges and increasing customer tenure to minimize churn. We advise exploring ways to improve the electronic check payment process to reduce churn rates among customers.