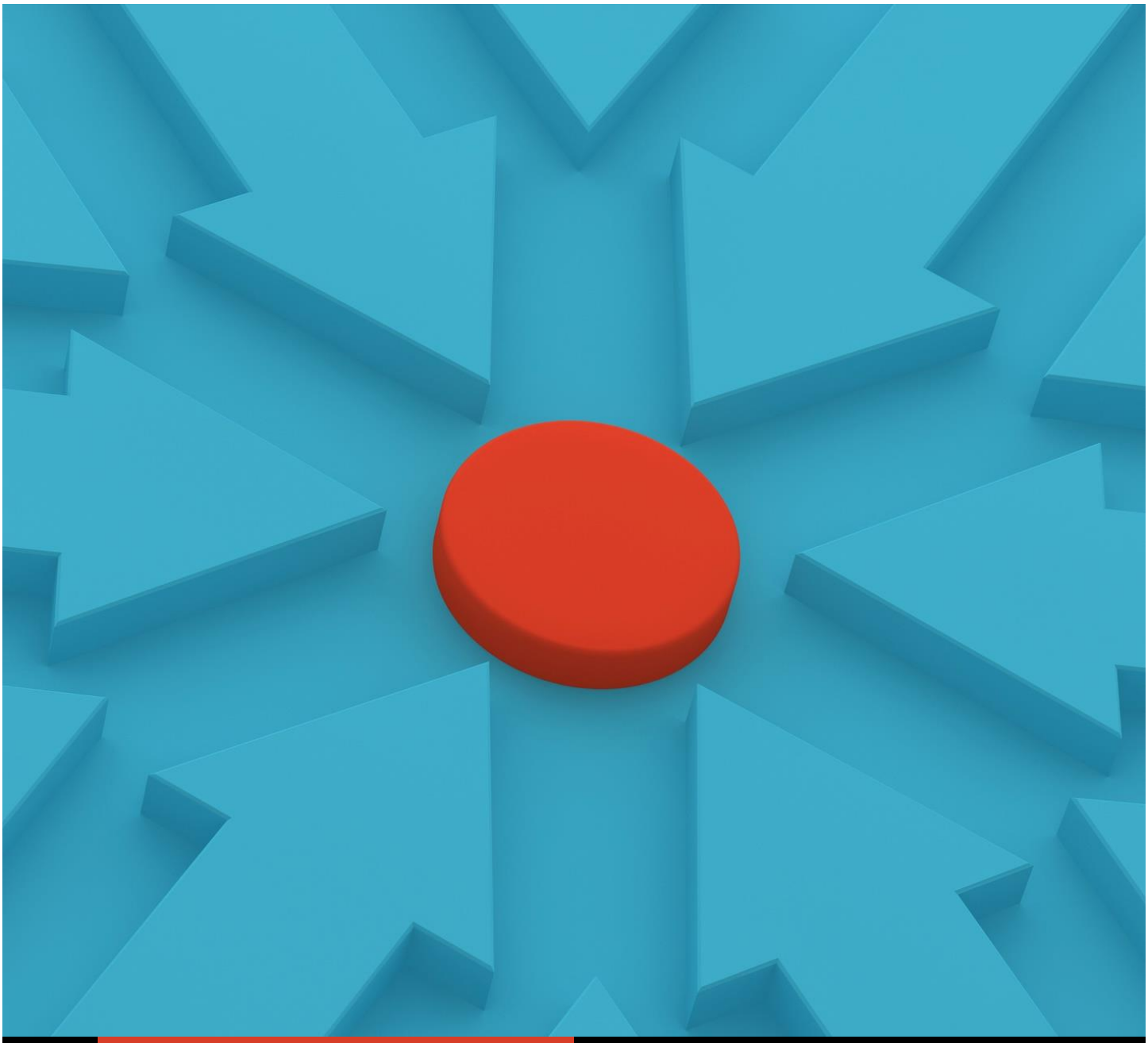
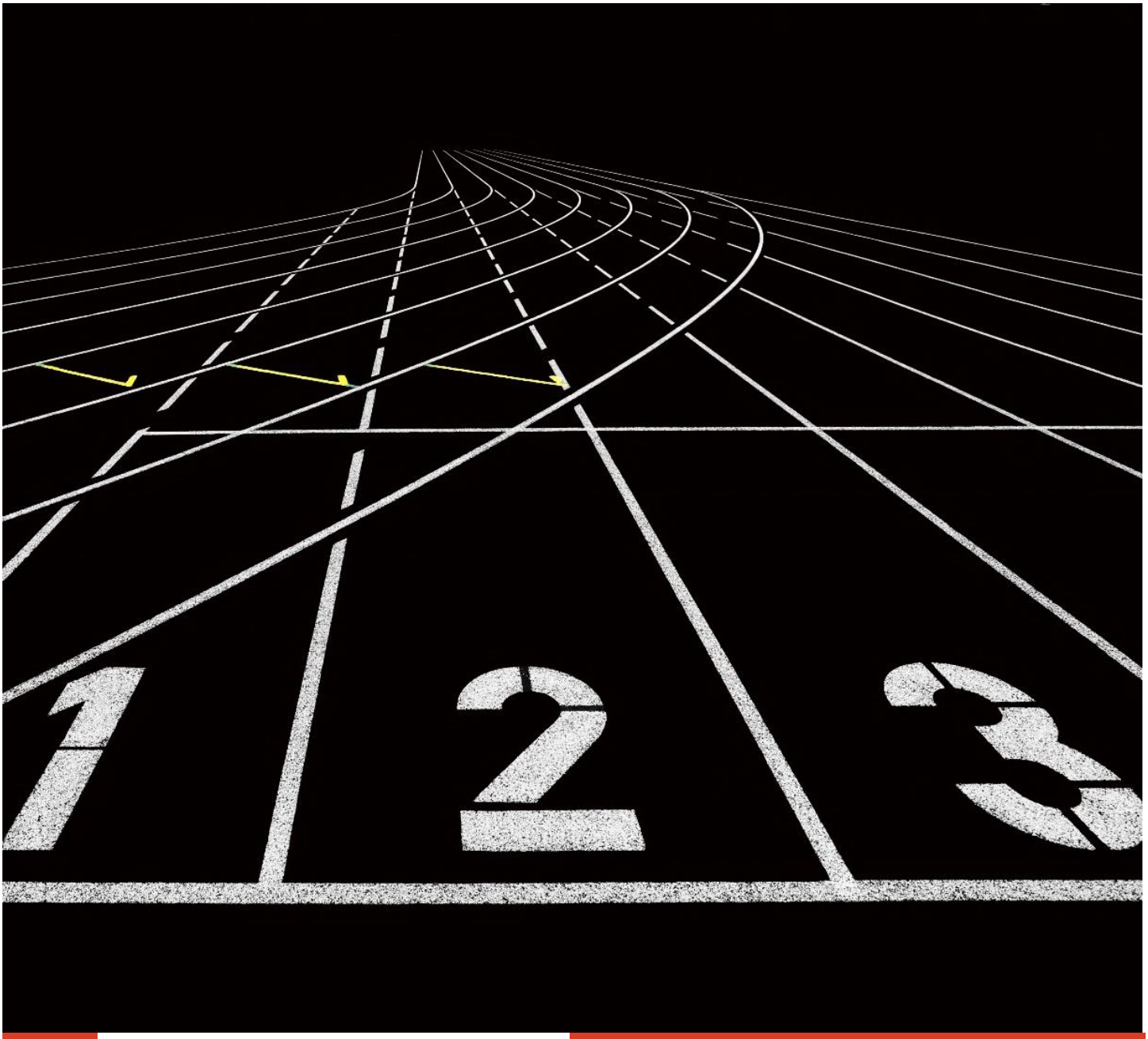


REPORT

PREDICTING CUSTOMER CHURN

MIHIR SRIVASTAVA



INTRODUCTION

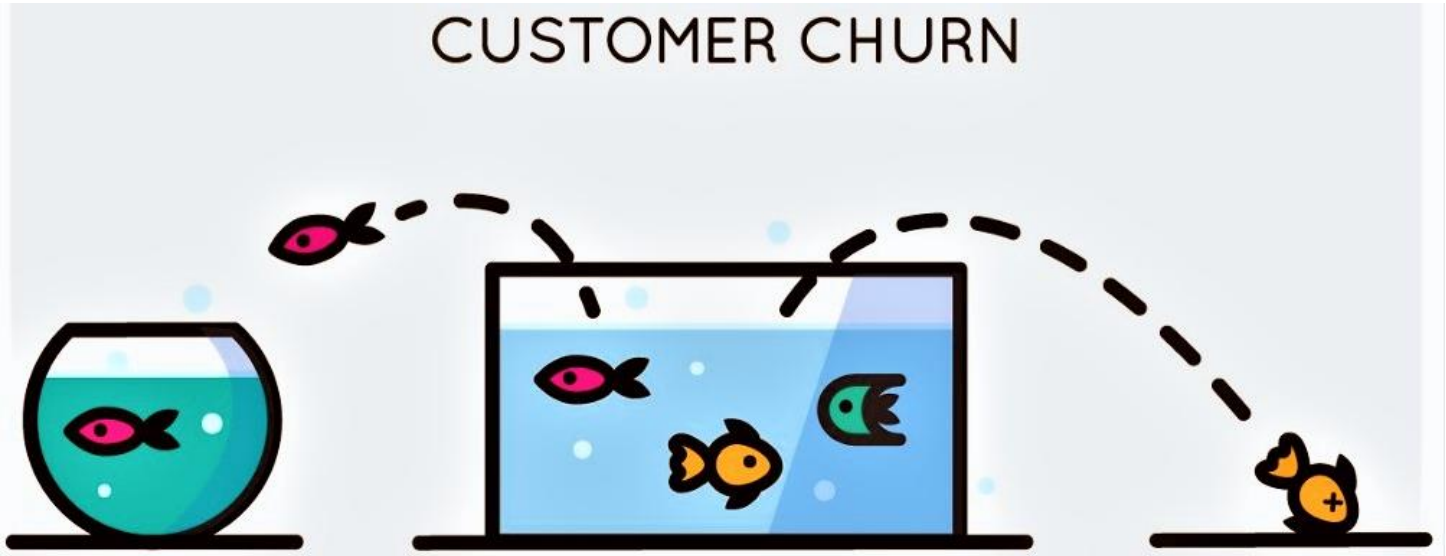
Customer churn, or the loss of customers, is a critical challenge faced by telecommunications companies.

Understanding the factors that contribute to churn can help companies proactively address customer retention strategies. In this report, we aim to predict customer churn in a telecommunications company using python and machine learning models.

DATASET

The dataset contains information about customers, including demographics, services subscribed to, contract details, and churn status. Here are some key features:

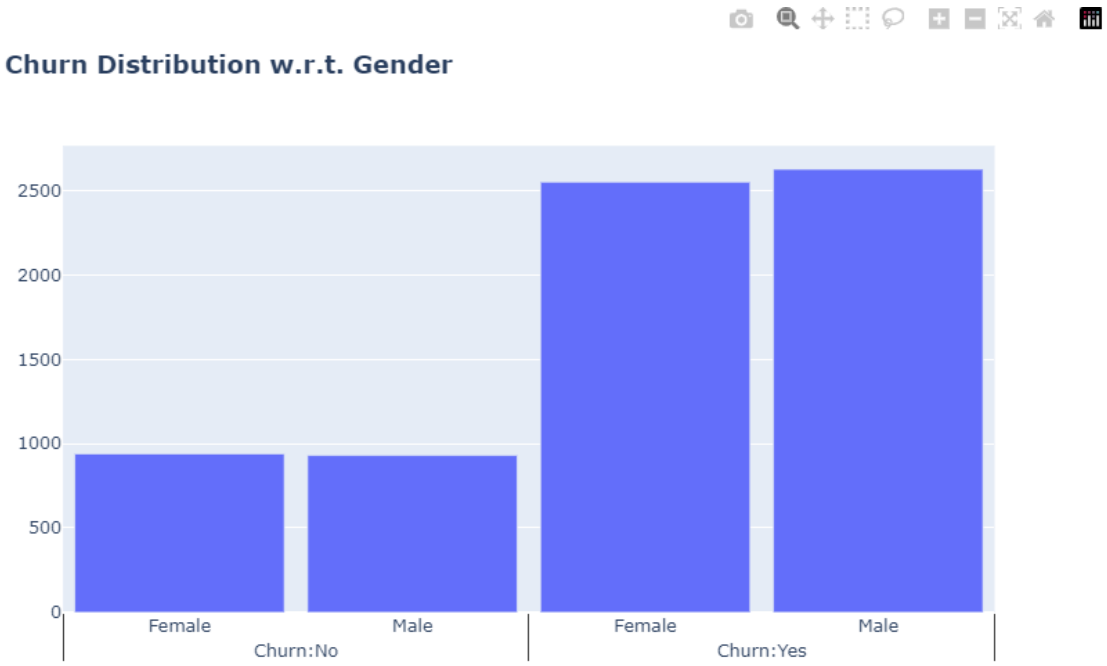
- customerID: Unique identifier for each customer
- gender: Gender of the customer
- SeniorCitizen: Whether the customer is a senior citizen or not (1 for yes, 0 for no)
- Partner: Whether the customer has a partner or not
- Dependents: Whether the customer has dependents or not
- tenure: Number of months the customer has stayed with the company
- PhoneService: Whether the customer has phone service or not
- MultipleLines: Whether the customer has multiple lines or not
- InternetService: Type of internet service subscribed to by the customer
- OnlineSecurity: Whether the customer has online security or not
- OnlineBackup: Whether the customer has online backup or not
- DeviceProtection: Whether the customer has device protection or not
- TechSupport: Whether the customer has tech support or not
- StreamingTV: Whether the customer has streaming TV or not
- StreamingMovies: Whether the customer has streaming movies or not
- Contract: Type of contract subscribed to by the customer
- PaperlessBilling: Whether the customer has paperless billing or not
- PaymentMethod: Payment method used by the customer
- MonthlyCharges: Monthly charges for the services subscribed to
- TotalCharges: Total charges incurred by the customer
- Churn: Whether the customer churned or not (target variable, 1 for yes, 0 for no)



EXPLORATORY DATA ANALYSIS (EDA):

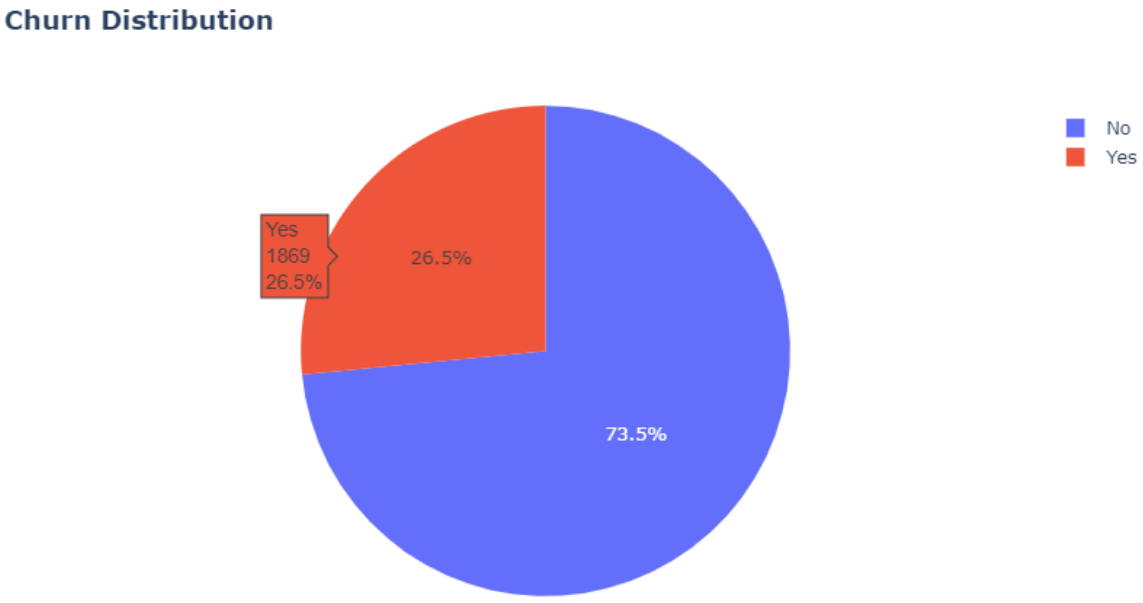
Churn Distribution by Gender:

The churn distribution is similar for both genders, with no significant difference observed.



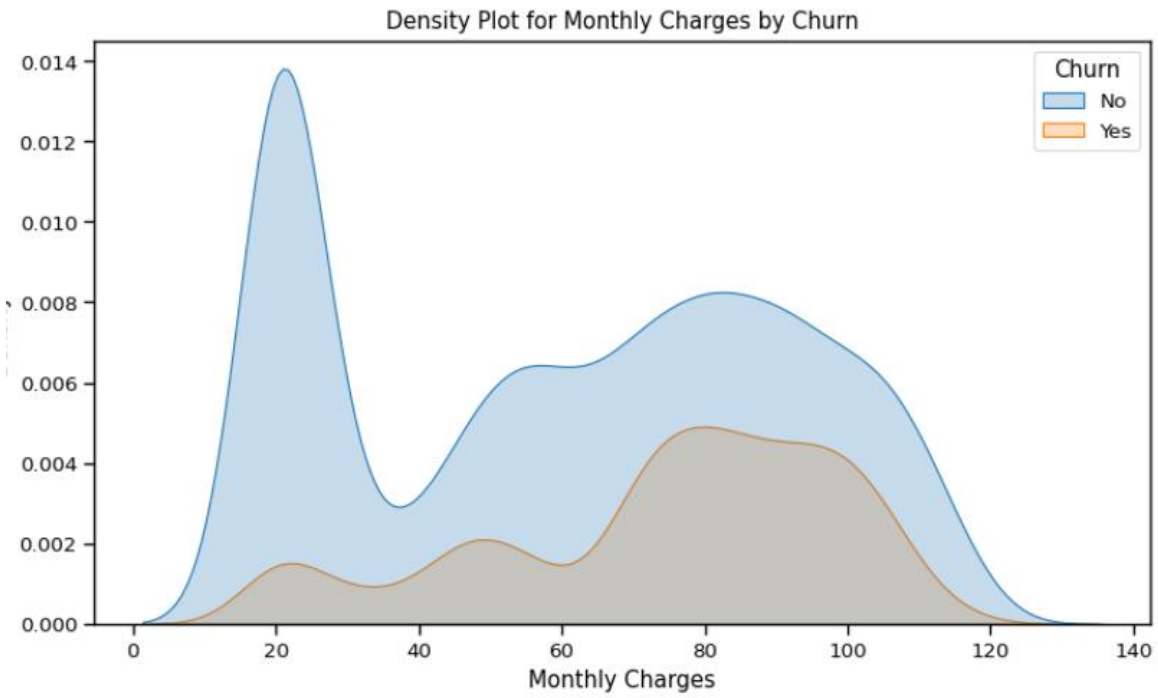
Churn Distribution Percentage-wise:

Approximately 26% of customers have churned, indicating a considerable churn rate
.



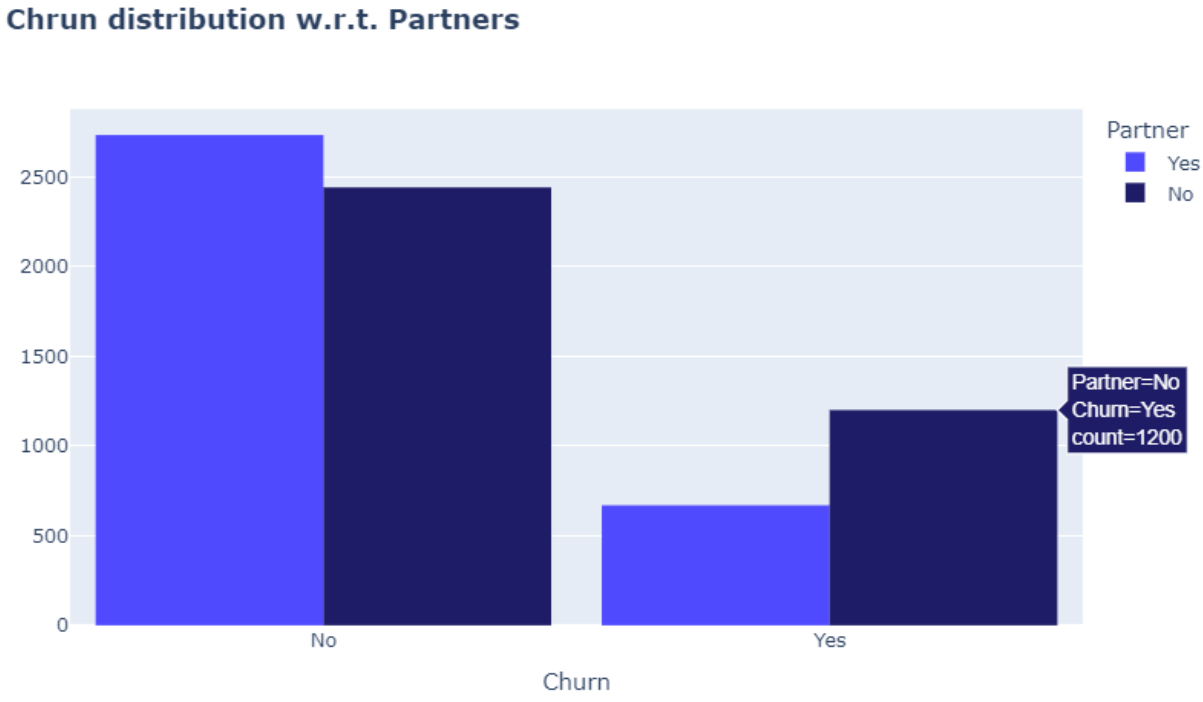
Relationship between Churn and Monthly Charges:

Customers with higher monthly charges are more likely to churn, indicating a potential correlation between charges and churn.



Churn Distribution with Respect to Partners:

Customers without partners have a slightly higher churn rate compared to those with partners.



And many more plots are in notebook.

FEATURE ENGINEERING

Creating Average Charges as a New Feature:

Feature engineering involves creating new features or modifying existing ones to improve the performance of machine learning models. In the context of the telecommunications dataset provided, feature engineering can help to better capture the relationships between the features and the target variable (Churn).

A new feature, 'AverageCharges' is created by dividing 'TotalCharges' by 'tenure'. This represents the average monthly charges incurred by the customer.

```
Index(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
       'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',
       'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn',
       'AverageCharges'],
      dtype='object')
```

```
data.AverageCharges
```

```
0        29.850000
1        55.573529
2        54.075000
3        40.905556
4        75.825000
           ...
7038     82.937500
7039    102.262500
7040     31.495455
7041     76.650000
7042    103.704545
Name: AverageCharges, Length: 7032, dtype: float64
```

MODELING

Random Forest Model:

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

- Random Forest is a powerful ensemble learning technique that can handle non-linear relationships and interactions between features.
- It is well-suited for classification tasks like predicting customer churn.

Logistic Regression Model:

Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is one of the simplest types of neural networks and provides the probabilities of a binary outcome.

- Logistic Regression is a classic linear model used for binary classification tasks.
- It's interpretable and provides insights into the importance of each feature.

Evaluation Results:

- Both models achieved decent performance in predicting customer churn.
- Random Forest model: Accuracy = 0.781, Precision = 0.623, Recall = 0.465, F1 Score = 0.530
- Logistic Regression model: Accuracy = 0.788, Precision = 0.623, Recall = 0.513 F1 Score = 0.563

```
Logistic Regression Model:
Accuracy: 0.788
Precision: 0.623
Recall: 0.513
F1-Score: 0.563


Random Forest Model:
Accuracy: 0.781
Precision: 0.617
Recall: 0.465
F1-Score: 0.530
```

CONCLUSION

Project Overview

The goal of this project was to predict customer churn in a telecommunications company using machine learning models. The dataset provided contained various customer attributes, including demographic information, service usage, and account information. I applied feature engineering, exploratory data analysis (EDA), and built two machine learning models: Logistic Regression and Random Forest.

Exploratory Data Analysis (EDA) Findings

1. **Churn Distribution**:
    - The dataset showed an imbalance between churned and non-churned customers, with more customers not churning.
2. **Churn vs. Gender**:
    - The analysis indicated no significant difference in churn rates between male and female customers.
3. **Churn vs. Monthly Charges**:
    - Customers with higher monthly charges showed a higher likelihood of churning.
4. **Churn vs. Contract Type**:
    - Customers with month-to-month contracts had higher churn rates compared to those with one or two-year contracts.
5. **Churn vs. Partner**:
    - Customers without a partner had higher churn rates.

Feature Engineering

1. **Conversion of Categorical Variables**:
    - I converted categorical features such as gender, InternetService, Contract, and PaymentMethod to numeric values using one-hot encoding.
2. **Creation of New Features**:
    - A new feature, AverageCharges, was created by dividing TotalCharges by tenure.
3. **Handling Missing Values**:
    - Missing values in TotalCharges were handled by converting non-numeric entries to NaN and subsequently dropping rows with NaN values in AverageCharges.