# Telecom Customer Churn Analysis

# Contents

Project Overview

Data Cleaning

Exploratory Data Analysis

Feature Engineering

Data Preprocessing

Machine Learning Models

# Project Team Members

Parthiv Patel

Prachiti Jadhav

Sukhman Legha

Janki Joshi

Aishwarya

# Part 1:
# Project Overview

# Project Overview

## Technologies

python™ jupyter NumPy

seaborn hvPlot matplotlib

pandas statsmodels scikit learn

## Problem Statement

The goal of this analysis is to understand and predict customer churn for a telecom company. Customer churn, also known as customer attrition, refers to when a customer stops doing business with a company or stops using its services. In the context of a telecom company, this could mean a customer discontinuing their phone or internet service.

- To collect information on more than 7000+ distinct customer behaviour towards a telecom company
- To clean, prepare, organize and handle the missing data, outliers,incorrect record and so on
- Prepare a statistical report to analyse the trend in the customer retention for telecom businesses
- Propose several research questions and use visualization techniques to identify the researched questions

## Data Resources

- Kaggle - https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction

## Attributes

7000+ Customer Dataset

Monthly Charges

Online Security

Churn Rate

# Part 2:
# Data Cleaning

# Data Cleaning

## Missing Values Analysis

The presence of missing values in the dataset was identified using the isnull().sum() method. The generated output showed that there are no missing values in any of the columns.

## Handling Non-numeric Values in 'TotalCharges' Column

During further exploration, it was discovered that the 'TotalCharges' column contained 11 non-numeric values. To address this issue, the column was converted to numeric type using pd.to_numeric() with errors set to 'coerce' to treat errors as NaN. Rows with missing values in the 'TotalCharges' column are filtered out and dropped from the dataset.

## Dropping Unnecessary Columns

To streamline the dataset and remove unnecessary information,  'customerID' column was dropped as it is not required for the analysis. The resulting DataFrame now serves as the clean and preprocessed dataset for further analysis.

# Part 3:<br>Exploratory Data Analysis

# Exploratory Data Analysis

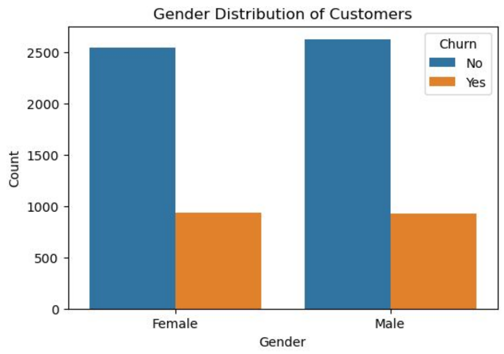
1. **Definition of EDA:**
   - EDA is a critical first step in the data analysis process.
   - It involves examining raw data to understand its structure and components.

2. **Objectives of EDA:**
   - To discover patterns and relationships within the data.
   - To identify anomalies or outliers that may need special treatment.
   - To test assumptions about the data's distribution and statistical properties.

3. **Techniques Used in EDA:**
   - Visualization tools like histograms, box plots, scatter plots, and heat maps.
   - Statistical summaries and techniques, such as mean, median, mode, variance, and correlation analysis.

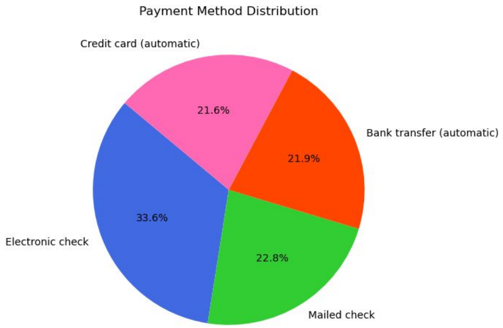
# Exploratory Data Analysis

## Gender Distribution of Customers



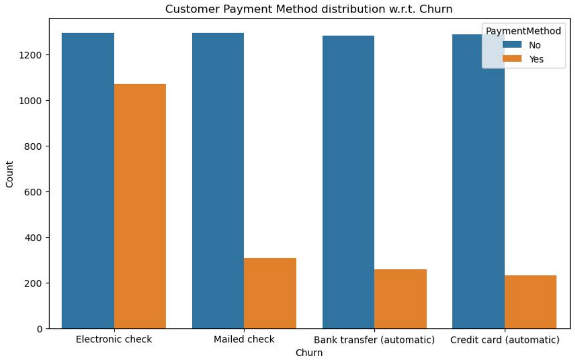
- The count of female customers who have not churned is over 2500, while those who have churned is approximately 1000.
- The count of male customers who have not churned is over 2500, while those who have churned is approximately 1000.
- Overall, the chart suggests that the number of customers who have not churned is substantially higher than those who have, for both genders.

## Payment Method Distribution



- "Electronic check" accounts for the largest portion, at 33.6%.
- "Bank transfer (automatic)" represents 21.9%.
- "Credit card (automatic)" is close behind at 21.6%.
- "Mailed check" makes up 22.8%.
- The chart visually conveys that electronic checks are the most common payment method among the customers, while the other three methods are used relatively evenly.

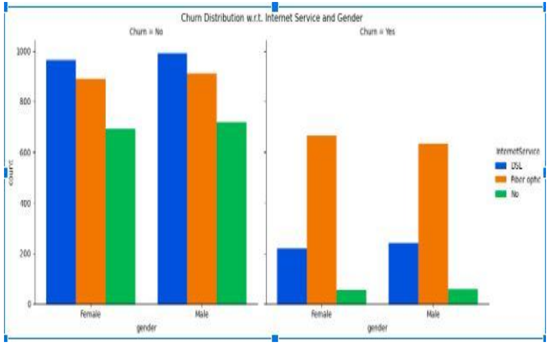
## Customer Payment Method Distribution



- "Electronic check" shows a high count of churned customers compared to the other payment methods.
- For "Mailed check" and "Bank transfer (automatic)", the count of customers who have not churned is higher than those who have churned.
- "Credit card (automatic)" has the lowest count of churned customers and the highest count of customers who have not churned among the payment methods presented.

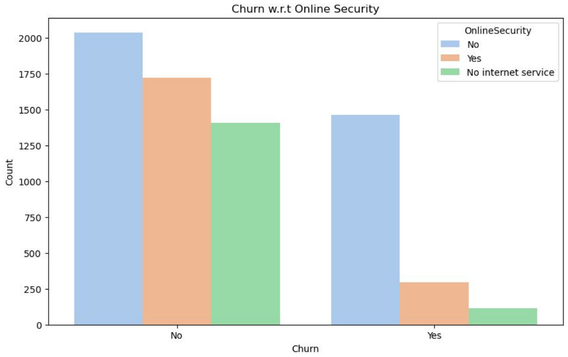
# Exploratory Data Analysis

## Internet Service and Gender Distribution



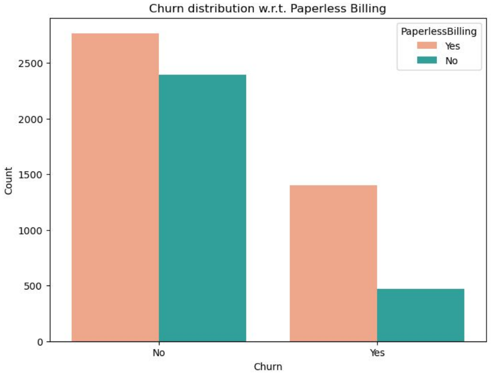
- In the "No Churn" section, both female and male customers predominantly use DSL and Fiber optic services, with fewer customers having no internet service.
- Among the customers who have churned, the trend is similar, but with a drastic reduction in numbers. The majority of both female and male customers who have churned were using Fiber optic service.
- The count of individuals  who have churned and had no internet service is notably small, compared to other services.

## Churn w.r.t Online Security



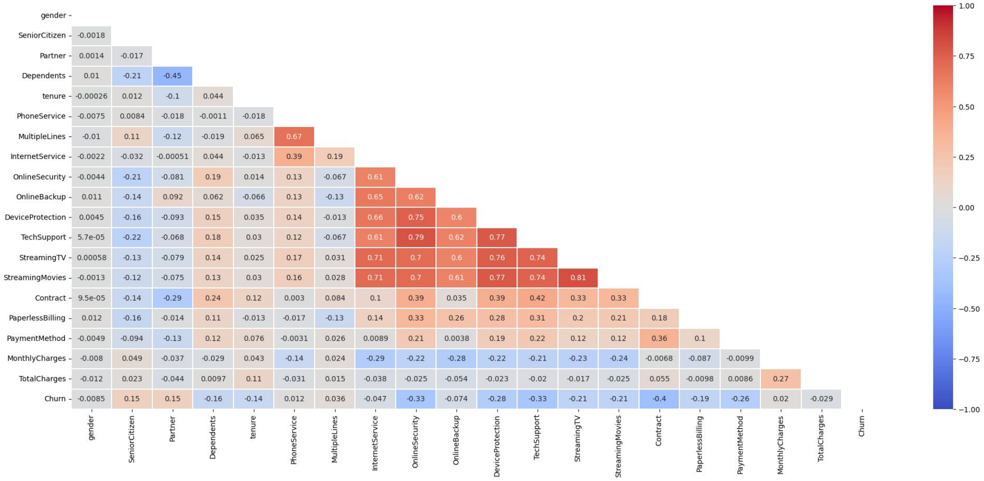
- A substantial number of customers who have not churned do not have online security. This number is significantly higher than those who have online security and those with no internet service.
- Among the customers who have churned, the majority also did not have online security, but the count is much less than those who have not churned.
- Very few customers with no internet service have churned.The data suggests that not having online security may be associated with higher churn, regardless of the churn status. However, customers with no internet service show the least likelihood of churning.

## Churn w.r.t Paperless Billing



- It can be observed that more number of customers that have not churned prefer paperless billing i.e receiving their bills online versus receiving a physical copy of the bill in mail
- The same also holds true for the majority of people who have churned, although the difference between people preferring paperless bills is significantly more than those who want to receive a hard copy of the bill.
- In general, it can be concluded that in either cases, paperless billing is a popular option which contributes towards saving paper and reduces the possibility of identity theft caused due to mail tampering.

# Exploratory Data Analysis



**Some notable correlations are:**

- "MultipleLines" and "PhoneService" show a strong positive correlation, which is highlighted in red.
- "Dependents" and "Partner" show a strong negative correlation, highlighted in blue.
- "StreamingTV" and "StreamingMovies" also have a high positive correlation with each other.
- "Churn" seems to have a moderate negative correlation with "tenure," "OnlineSecurity," "TechSupport," "DeviceProtection," "StreamingTV," and "StreamingMovies," which might suggest that customers with longer tenure or those who use these services are less likely to churn.

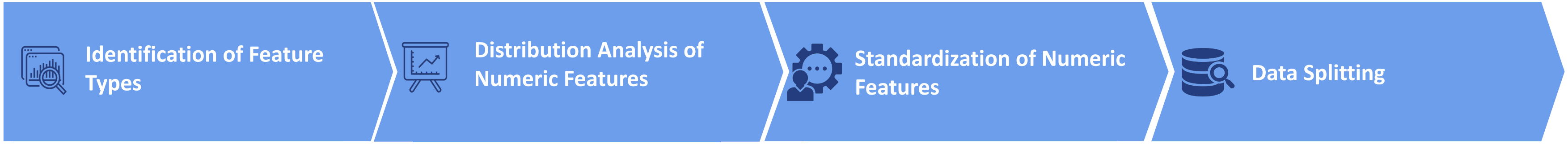
# Part 4:
# Feature Engineering

# Feature Engineering

1. **Feature Engineering Overview:**
   - Enhancing data features for improved machine learning model performance.
2. **Key Techniques:**
   - Deriving new features (e.g., 'age' from 'date of birth').
   - Normalizing data (e.g., logarithmic transformations).
   - Binning and encoding categorical variables.
3. **Encoding Categorical Variables:**
   - Transforming non-numeric categories into machine-readable formats.
   - Example: `encode_categorical_to_int` function for integer encoding.
4. **Importance:**
   - Crucial for model compatibility and predictive accuracy.
   - Balances informative feature creation with model complexity.

# Part 5:
# Data Preprocessing

# Data Preprocessing



### Identification of Feature Types

The dataset was analyzed to distinguish between numeric (num_cols) and categorical (cat_cols_ohe) features. This distinction is vital for applying suitable preprocessing techniques to different data types.

### Distribution Analysis of Numeric Features

Distribution plots were created for numeric features - 'tenure', 'Monthly Charges', and 'Total Charges'.
These plots are instrumental in understanding the data distribution, identifying patterns, and detecting any potential outliers or skewness.

### Standardization of Numeric Features

The numeric columns were standardized using the StandardScaler.
This step normalizes the data, ensuring that features with larger scales do not dominate the model's learning process.

### Data Splitting

The dataset was divided into features (X) and the target variable (y), where 'Churn' is the target.
Subsequently, the data was split into training and testing sets, with a 70-30% ratio. The stratify parameter was used to maintain a consistent distribution of the target variable across these sets.

# Part 6: Machine Learning Models

SORRY, KID, OUR MACHINE LEARNING CRM WITH PREDICTIVE ANALYTICS SAYS YOU'RE GETTING COAL THIS YEAR.
TOM FISHBURNE

# K-Nearest Neighbors Algorithm

The K-Nearest Neighbors (K-NN) algorithm is a simple, yet powerful machine learning technique used for classification. K-NN works by finding the 'k' closest training examples in the feature space to a given test point. The output is then determined based on these 'k' nearest neighbors.

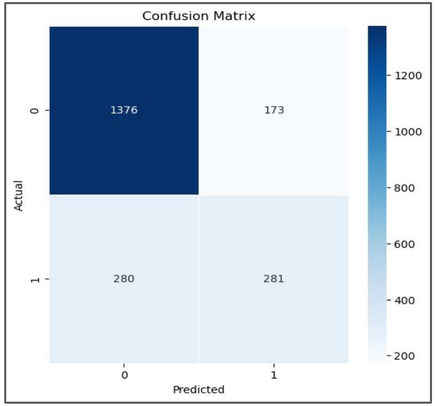
## Model Performance & Evaluation

### Accuracy Matrix

The KNN model, with k set to 16, achieved an accuracy of **78.53%** on the test dataset. This metric reflects the overall proportion of correctly predicted instances (both true positives and true negatives) in relation to all predictions made.

### Confusion Matrix

- True Negatives (TN): 1376, indicating non-churned customers correctly classified.
- True Positives (TP): 281, indicating churned customers correctly classified.
- False Positives (FP): 173, indicating non-churned customers incorrectly classified as churned.
- False Negatives (FN): 280, indicating churned customers incorrectly classified as non-churned.



# Neural Network Model

A neural network consists of layers of interconnected nodes (neurons). Each node is a simple processing unit that performs a basic computation. The basic structure includes an input layer, one or more hidden layers, and an output layer.

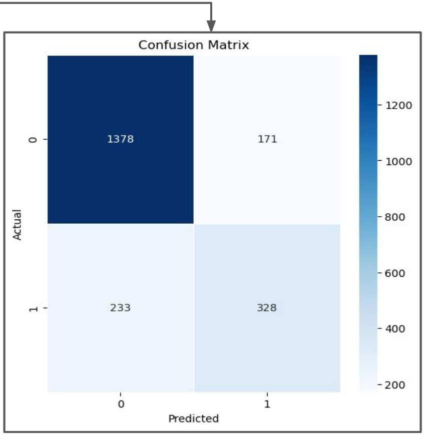
## Model Performance & Evaluation

### Accuracy Matrix

Post-training, the model achieved an accuracy of **80.85%** on the test set, indicating a high level of predictive performance for this binary classification task.

### Confusion Matrix

- True Negatives (TN): 1378, where the model correctly identified non-churned customers.
- True Positives (TP): 328, where the model accurately predicted churned customers.
- False Positives (FP): 171, where non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 233, where churned customers were incorrectly predicted as non-churned..



# Decision Trees

**Decision Trees is a predictive modeling algorithm employed in classification tasks. It assumes the conditional independence of features and is widely used for scenarios where interpretable decision rules are valuable. For Telecom Customer Churn prediction, Decision Trees are utilized to create a model that can effectively classify customers as churned or non-churned.**

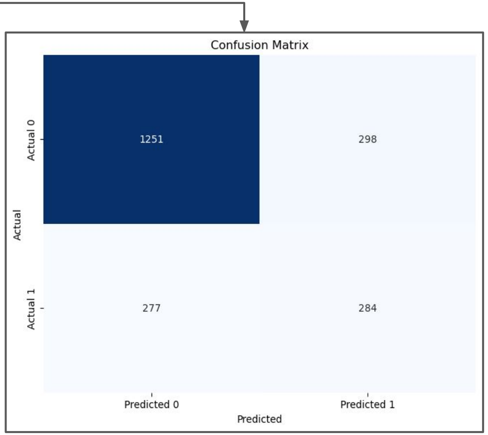
## Model Performance & Evaluation

### Accuracy Matrix

Decision Trees model achieved an accuracy of **72.75%** on the test set. This indicates the proportion of correctly classified instances, showcasing the model's effectiveness in distinguishing between customers who churned and those who did not.

### Confusion Matrix

- True Negatives (TN): 1251, where the model correctly identified non-churned customers.
- True Positives (TP): 284, where the model accurately predicted churned customers.
- False Positives (FP): 298,where  non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 277, where churned customers were incorrectly predicted as non-churned.



# Stochastic Gradient Descent Classifier

The Stochastic Gradient Descent (SGD) Classifier is a machine learning algorithm used for classification tasks. It optimizes the model parameters by updating them iteratively based on small, random subsets of training data, making it well-suited for large datasets. Stochastic Gradient Descent Classifier (SGD) model is employed for predicting Telecom Customer Churn.

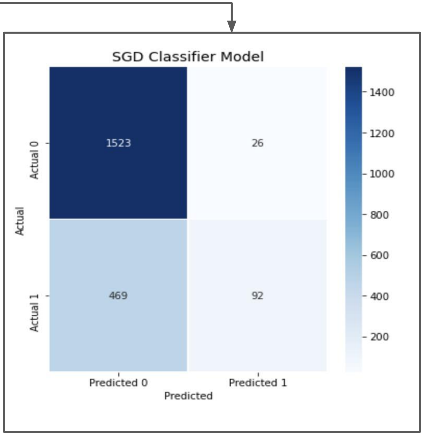
## Model Performance & Evaluation

### Accuracy Matrix

The Stochastic Gradient Descent Classifier model achieved an accuracy of **75.59%** on the test set. This indicates the proportion of correctly classified instances, showcasing the model's effectiveness in distinguishing between customers who churned and those who did not.

### Confusion Matrix

- True Negatives (TN): 1523, where the model correctly identified non-churned customers.
- True Positives (TP): 92, where the model accurately predicted churned customers.
- False Positives (FP): 26,where  non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 469, where churned customers were incorrectly predicted as non-churned.



# Logistic Regression

**The Logistic Regression model is a statistical method used for binary classification. It is a popular algorithm due to its simplicity, interpretability and efficiency. It provides the probability of the event occurring which is why it was found highly efficient for predicting Telecom Customer Churn.**

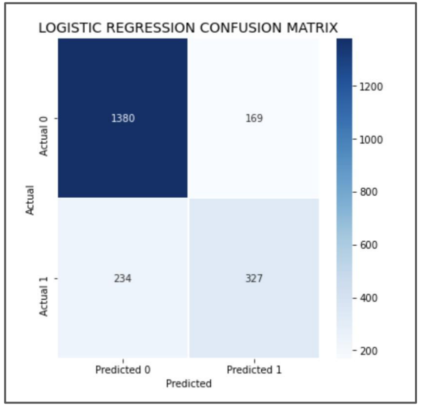
## Model Performance & Evaluation

### Accuracy Matrix

The Logistic Regression  model achieved an accuracy of **80.90%** on the test set. This indicates the proportion of correctly classified instances, showcasing the model's effectiveness in distinguishing between customers who churned and those who did not.

### Confusion Matrix

- True Negatives (TN): 1380, where the model correctly identified non-churned customers.
- True Positives (TP): 327, where the model accurately predicted churned customers.
- False Positives (FP): 169,where  non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 234, where churned customers were incorrectly predicted as non-churned.



# Support Vector Machine

**The Support Vector Machine is a fast and dependable classification algorithm that performs very well with a limited size data to analyze. Support Vector Machine model is employed for predicting Telecom Customer Churn.**

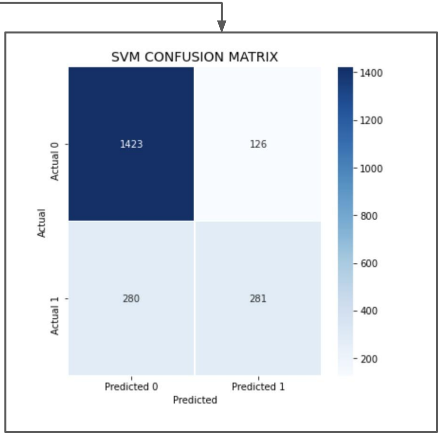
## Model Performance & Evaluation

### Accuracy Matrix

The Support Vector Machine model achieved an accuracy of **80.75%** on the test set. This indicates the proportion of correctly classified instances, showcasing the model's effectiveness in distinguishing between customers who churned and those who did not.

### Confusion Matrix

- True Negatives (TN): 1423, where the model correctly identified non-churned customers.
- True Positives (TP): 281, where the model accurately predicted churned customers.
- False Positives (FP): 126,where  non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 280, where churned customers were incorrectly predicted as non-churned.



# Naives Bayes Algorithm

Naive Bayes is a probabilistic algorithm used for classification tasks. It's particularly well-suited for scenarios where there are multiple features describing each instance, and the assumption of conditional independence among features holds reasonably well. Naive Bayes (NB) model is employed for predicting Telecom Customer Churn.

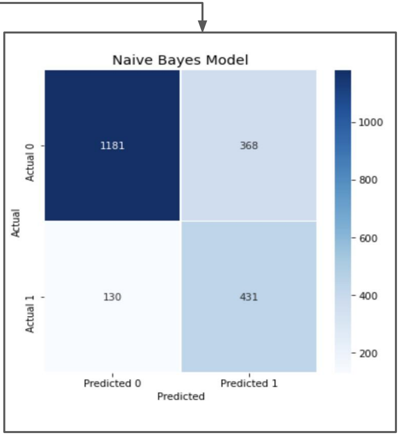
## Model Performance & Evaluation

### Accuracy Matrix

Naive Bayes model achieved an accuracy of **76.40%** on the test set. This indicates the proportion of correctly classified instances, showcasing the model's effectiveness in distinguishing between customers who churned and those who did not.

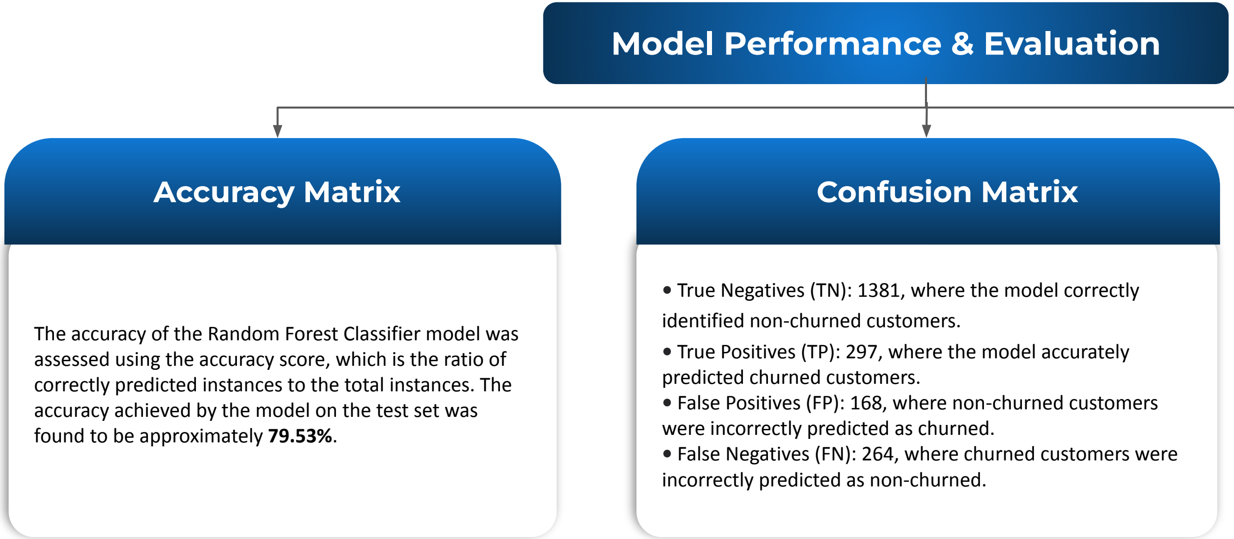
### Confusion Matrix

- True Negatives (TN): 1181, where the model correctly identified non-churned customers.
- True Positives (TP): 431, where the model accurately predicted churned customers.
- False Positives (FP): 368,where  non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 130, where churned customers were incorrectly predicted as non-churned.



# Random Forest

**Random Forest model operates by constructing a large number of decision trees during training and outputting the class that is the mode of the classes output by individual trees. Here, random forest classifier was employed for the predictive modeling of customer churn.**



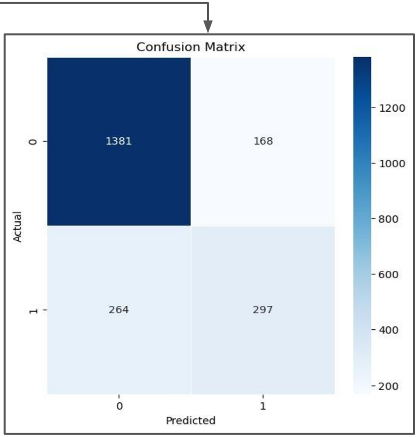
## Model Performance & Evaluation

### Accuracy Matrix

The accuracy of the Random Forest Classifier model was assessed using the accuracy score, which is the ratio of correctly predicted instances to the total instances. The accuracy achieved by the model on the test set was found to be approximately **79.53%**.

### Confusion Matrix

- True Negatives (TN): 1381, where the model correctly identified non-churned customers.
- True Positives (TP): 297, where the model accurately predicted churned customers.
- False Positives (FP): 168, where non-churned customers were incorrectly predicted as churned.
- False Negatives (FN): 264, where churned customers were incorrectly predicted as non-churned.



# Key Takeaways

**This analysis aimed to predict customer churn for a major telecom company using detailed customer data and machine learning models.**

Based on the findings, several meaningful insights around churn drivers and customer segments were found:

- The exploratory data analysis revealed associations between certain customer attributes and increased likelihood of churn.
    - Customers with higher monthly charges and month-to-month contracts most likely to churn
    - Fiber optic users and those without online security more prone to quit
    - Longer-tenured customers much less likely to churn

- The Logistic Regression model achieved the highest test accuracy of 80.90%.
- Other techniques like neural networks (accuracy 80.85%), SVM(accuracy 80.75%) and random forests (with accuracy of 79.53%) also performed reasonably well in maximizing key metrics.

# Recommendations

**Given these results, some practical applications would be:**

- Marketing team designs targeted promotions and bundles for high churn risk segments to persuade them to stay
- Customer service team focuses on quickly resolving issues and personalized engagement for vulnerable segments
- Product team upgrades fiber optics infrastructure to improve capabilities and mitigate service dissatisfaction
- Retention efforts emphasize newer customers who are more likely to churn
- Competitive and flexible pricing models retain price-sensitive customers paying higher monthly charges

# It’s The End Of Our Presentation!