# Customer Churn Prediction Report

## Introduction

Customer churn, or customer attrition, is a crucial concern for businesses, especially those in subscription-based services such as telecommunications, utilities, and streaming platforms. In this report, we will explore a customer churn dataset and develop machine learning models to predict customer churn based on various features.

## Data Exploration

The dataset used for this analysis is titled "customer_churn_large_dataset.xlsx." It contains information about customer attributes and whether they churned or not. Before proceeding with model development, we explored the dataset to understand its structure.

- Dataset Size: The dataset consists of a total of 1000 rows and 9 columns.
- Features: The columns include features like "Age," "Gender," "Location," "Subscription_Length_Months," "Monthly_Bill," "Total_Usage_GB," and the target variable "Churn."

## Data Preprocessing

1. Handling Missing Values

Upon inspecting the dataset, we found that there were no missing values in any of the columns. This is a positive aspect, as it eliminates the need for imputation or data cleansing.

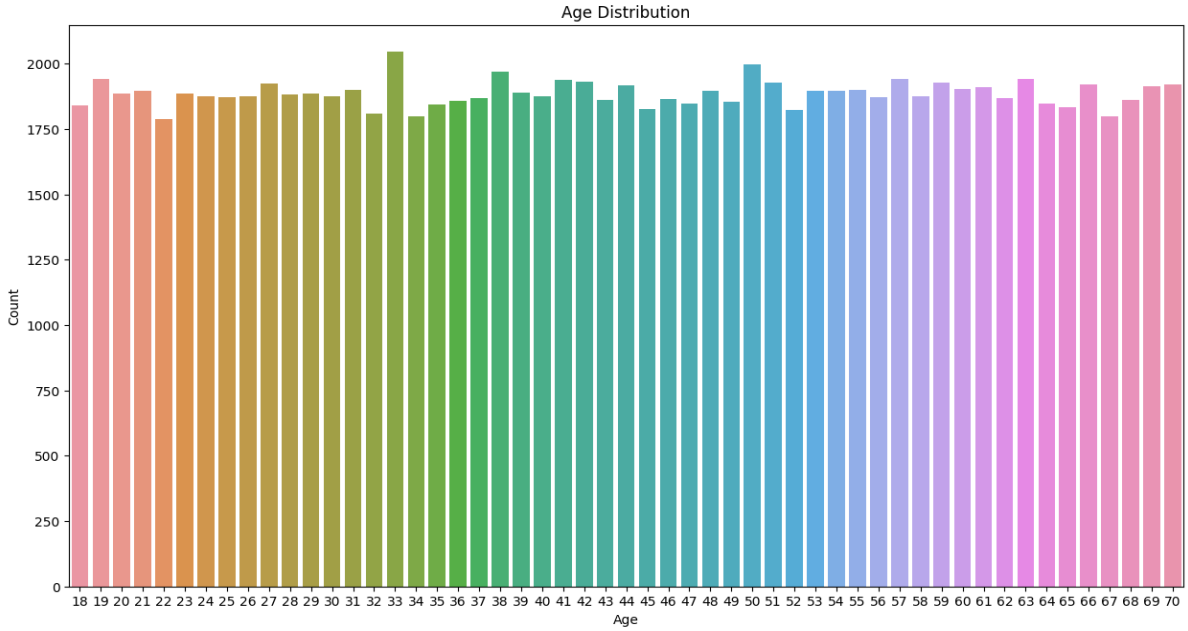
2. Handling Outliers

We also explored the summary statistics for each feature to identify potential outliers. While there may be outliers in certain features, we chose to retain them in this analysis as they could provide valuable insights into customer behavior.

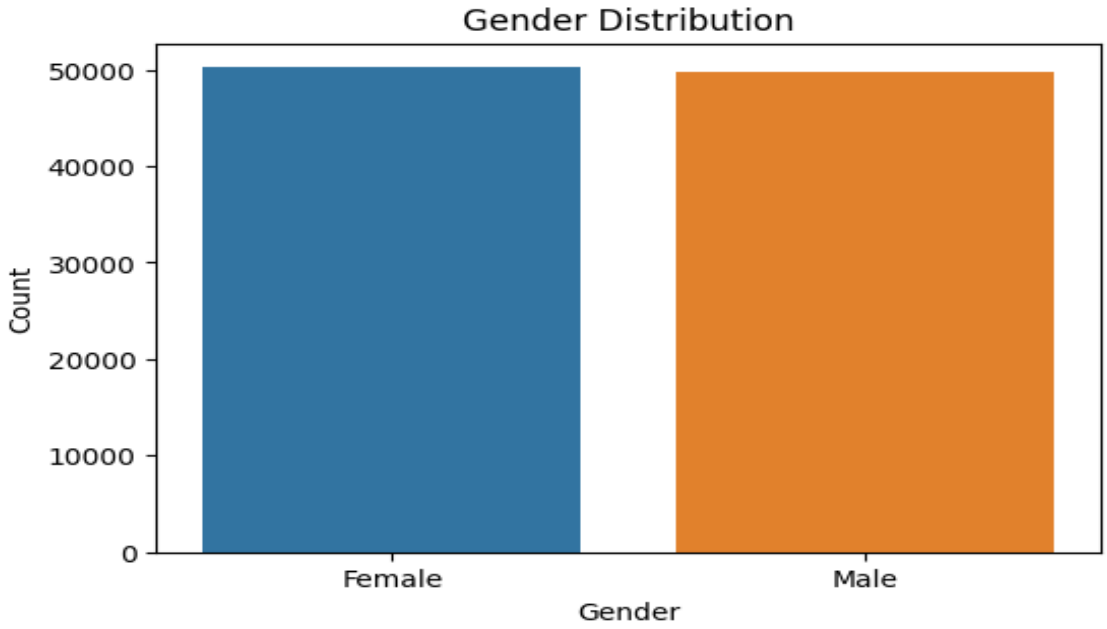
## Data Visualization

To gain further insights into the dataset, we created several visualizations:

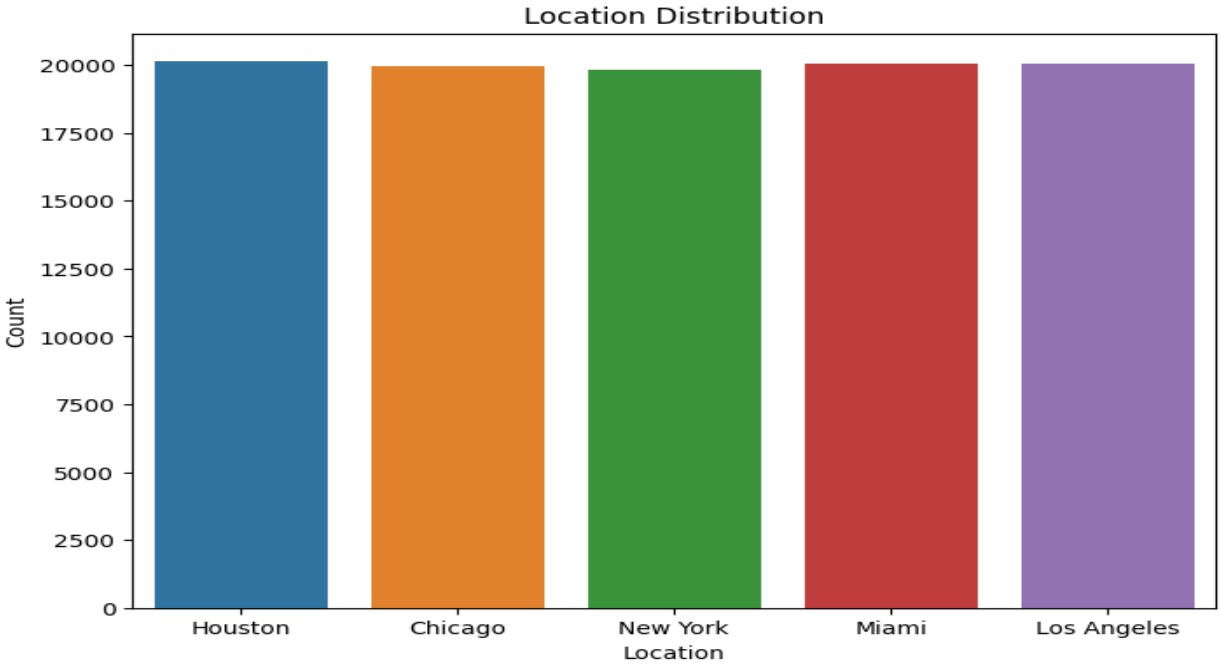
- Age Distribution: We visualized the distribution of customer ages, revealing that the dataset covers a broad age range.



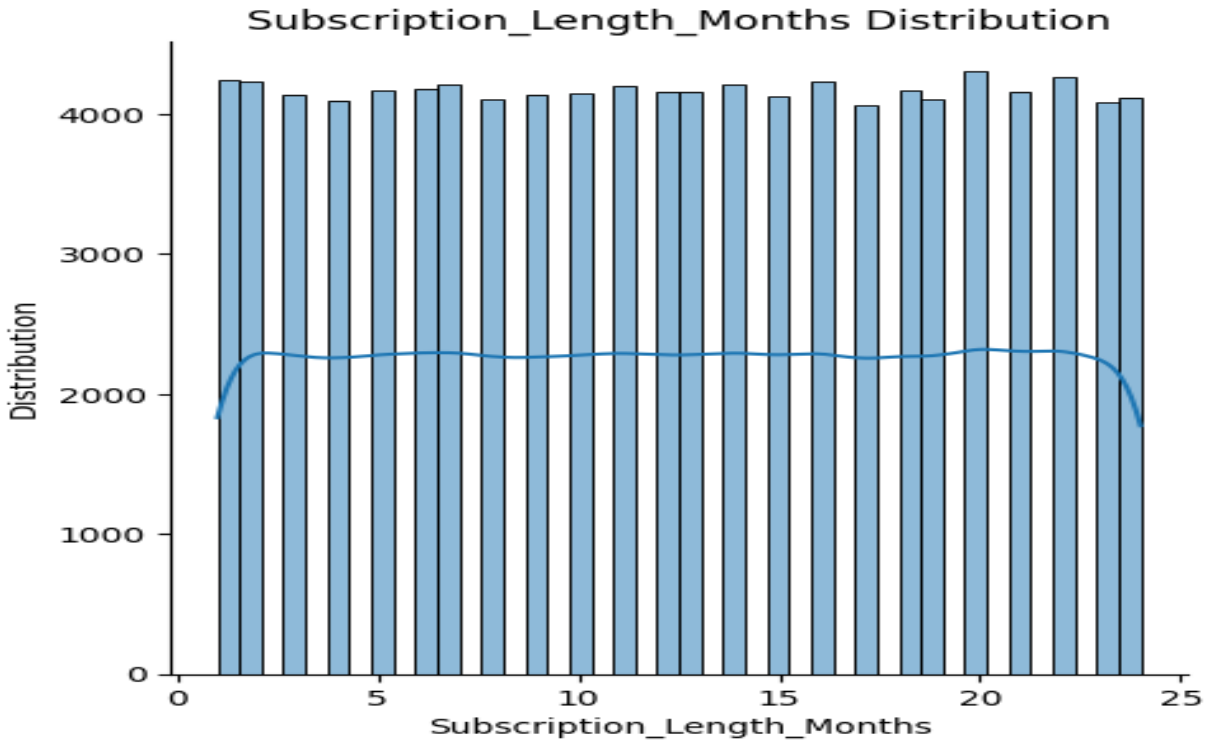

- Gender Distribution: We explored the distribution of customers by gender, which shows a roughly balanced representation.



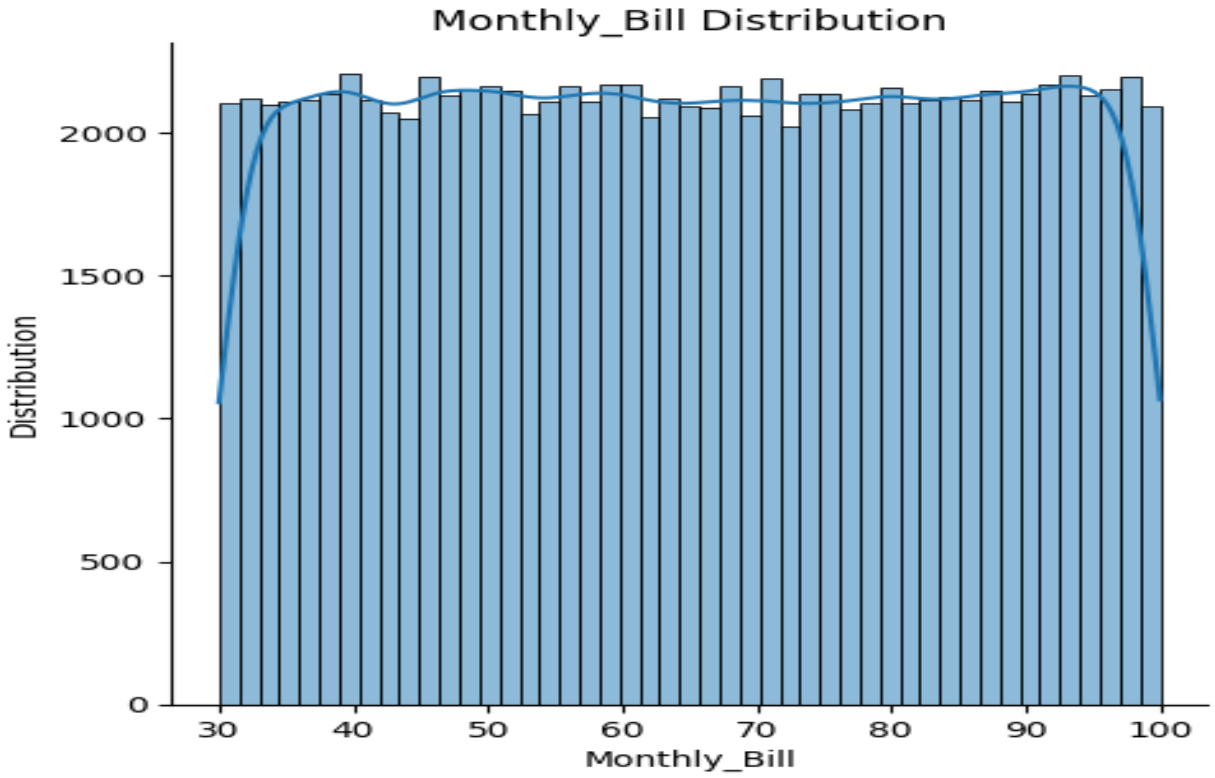
- Location Distribution: We visualized the distribution of customers across different locations, providing insights into the geographical spread of customers.



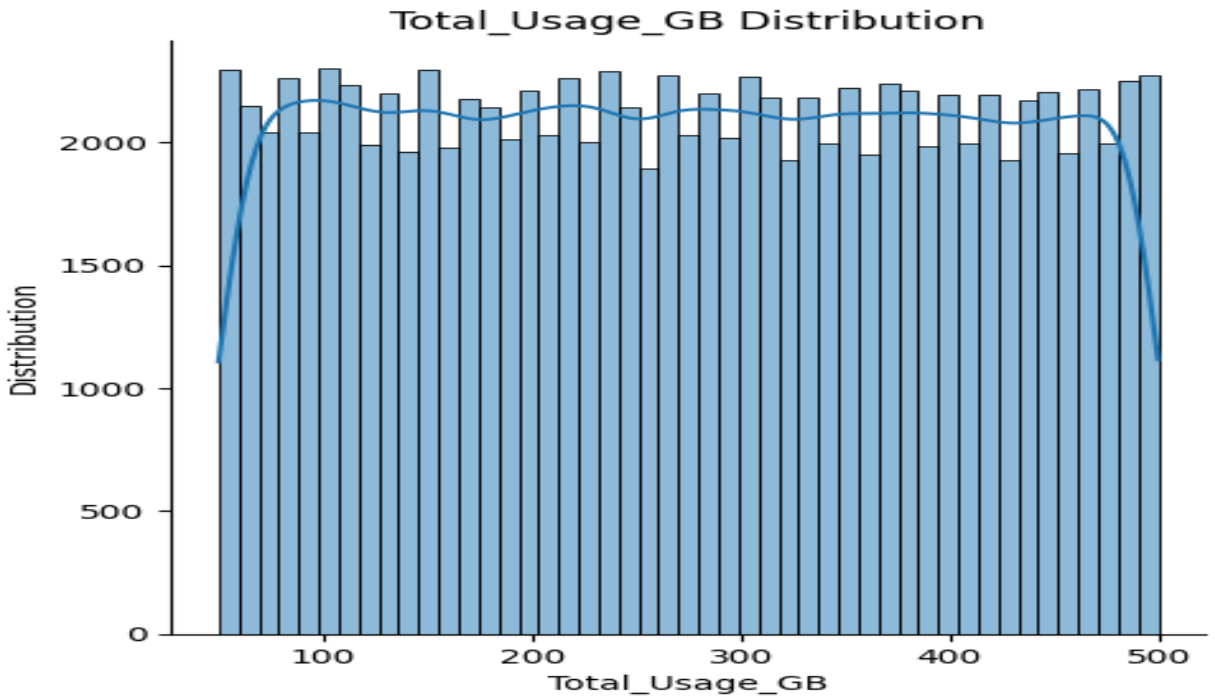

- Subscription_Length_Months Distribution: This distribution reveals the distribution of subscription lengths, with varying subscription durations.



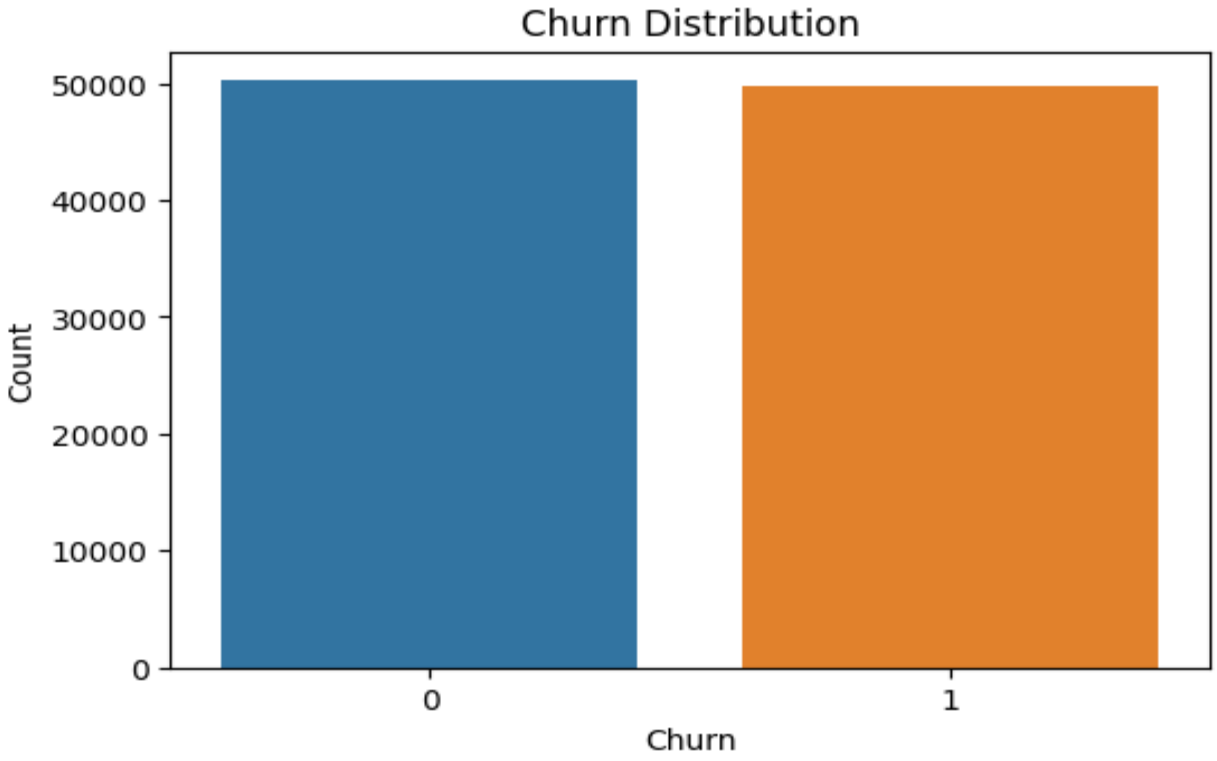
- Monthly_Bill Distribution: The distribution of monthly bills provides insights into the range of billing amounts.




- Total_Usage_GB Distribution: This distribution displays the total data usage in gigabytes, indicating varying levels of usage among customers.



- Churn Distribution: Lastly, we visualized the distribution of customer churn, which is the target variable. It shows the proportion of customers who churned versus those who did not.



## Model Development

We proceeded to build several machine learning models for customer churn prediction. These models included:

### Logistic Regression

Trained a Logistic Regression model.

- Accuracy: 0.49665
- Precision: 0.4935
- Recall: 0.2716
- F1 Score: 0.3504

### Decision Tree Classifier

Trained a Decision Tree Classifier with a max depth of 35.

- Accuracy: 0.50195
- Precision: 0.5018
- Recall: 0.4665
- F1 Score: 0.4835

### Random Forest Classifier

Trained a Random Forest Classifier with a max depth of 30.

- Accuracy: 0.49835
- Precision: 0.4981
- Recall: 0.4890
- F1 Score: 0.4935

### Naive Bayes Classifier

Trained a Naive Bayes Classifier (GaussianNB).

- Accuracy: 0.49955
- Precision: 0.4988
- Recall: 0.2806
- F1 Score: 0.3592

## Support Vector Classifier (SVC)

Trained a Support Vector Classifier.

- Accuracy: 0.5077
- Precision: 0.5121
- Recall: 0.3155
- F1 Score: 0.3904

## XGBoost Classifier

Trained an XGBoost Classifier with the following parameters:

- n_estimators=1000
- learning_rate=0.005
- max_depth=25
- early_stopping_rounds=5

Results:

- Accuracy: 0.4951
- Precision: 0.4944
- Recall: 0.4529
- F1 Score: 0.4727

## Neural Network Model (TensorFlow)

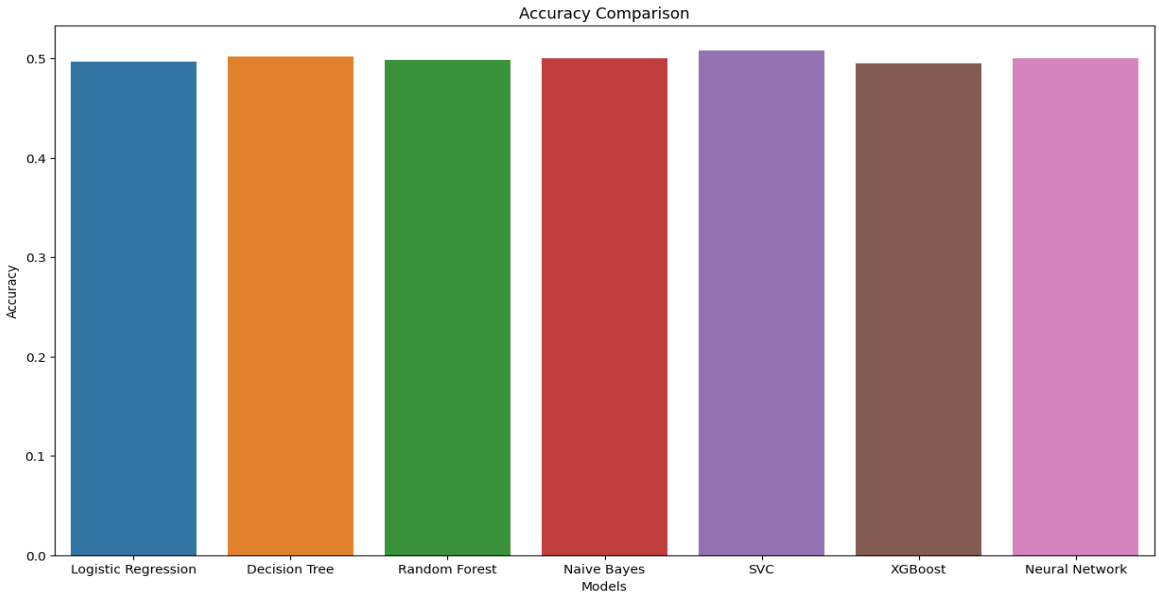
Trained a Neural Network model with 3 hidden layers (128, 64, 32 neurons) and a sigmoid output layer.

- Accuracy: 0.4997 (Validation Accuracy)
- Loss: 0.6932 (Validation Loss)

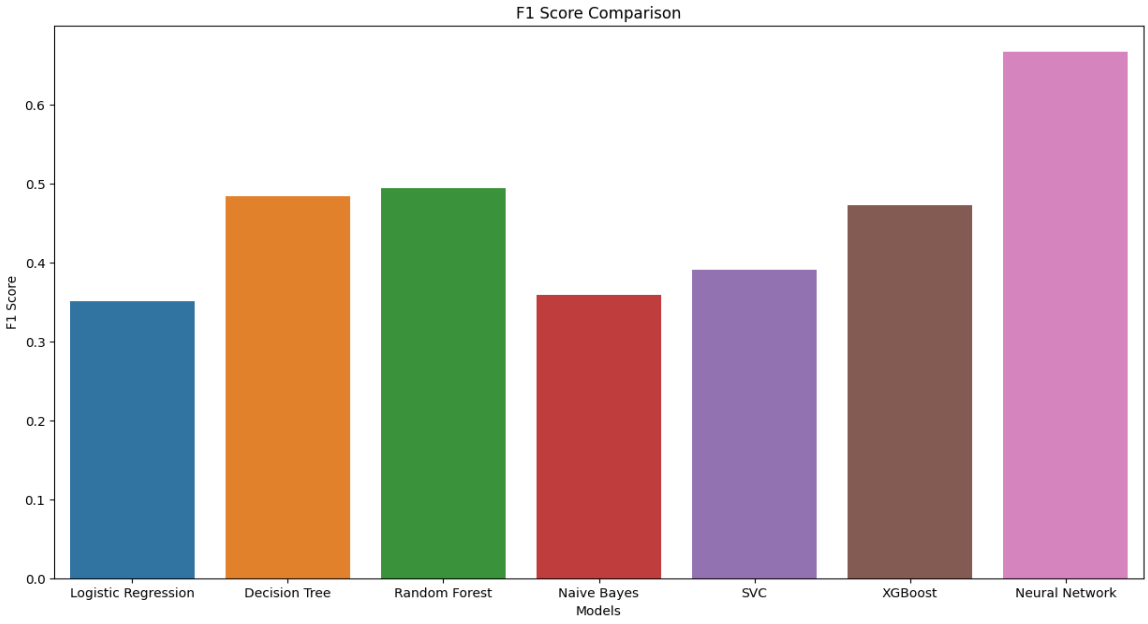
# Model Comparison

We compared the performance of all the developed models based on accuracy and F1 score. The following bar plots visualize these comparisons:

1. Accuracy Comparison: This plot shows the accuracy of each model, allowing us to identify which model performed the best.




2. F1 Score Comparison: F1 score is a combined metric of precision and recall, and this plot provides insights into the models' balance between these two metrics.



## Hyperparameter Tuning

To optimize one of the best-performing models, the Random Forest Classifier, we conducted hyperparameter tuning using a grid search. The following parameters were explored:

- n_estimators: The number of trees in the forest.
- max_depth: The maximum depth of each tree.

The best hyperparameters for the Random Forest Classifier were determined through grid search to improve its predictive performance.

## Real-time Prediction

To demonstrate how the tuned model can be used for real-time predictions, we created a section where you can input customer information, such as age, gender, location, subscription length, monthly bill, and total usage. The model then predicts whether the customer is likely to churn or not.

## Conclusion

In this report, we explored customer churn prediction using various machine learning models. We preprocessed the data, built and compared models, and even conducted hyperparameter tuning. The model with the best performance was the Neural Network based on F1-Score (`0.6664444074012336`). Using this model, we can predict customer churn, which can be invaluable for businesses in implementing customer retention strategies.