Bike Share System in SF Bay Area

Author: Owen Hsu

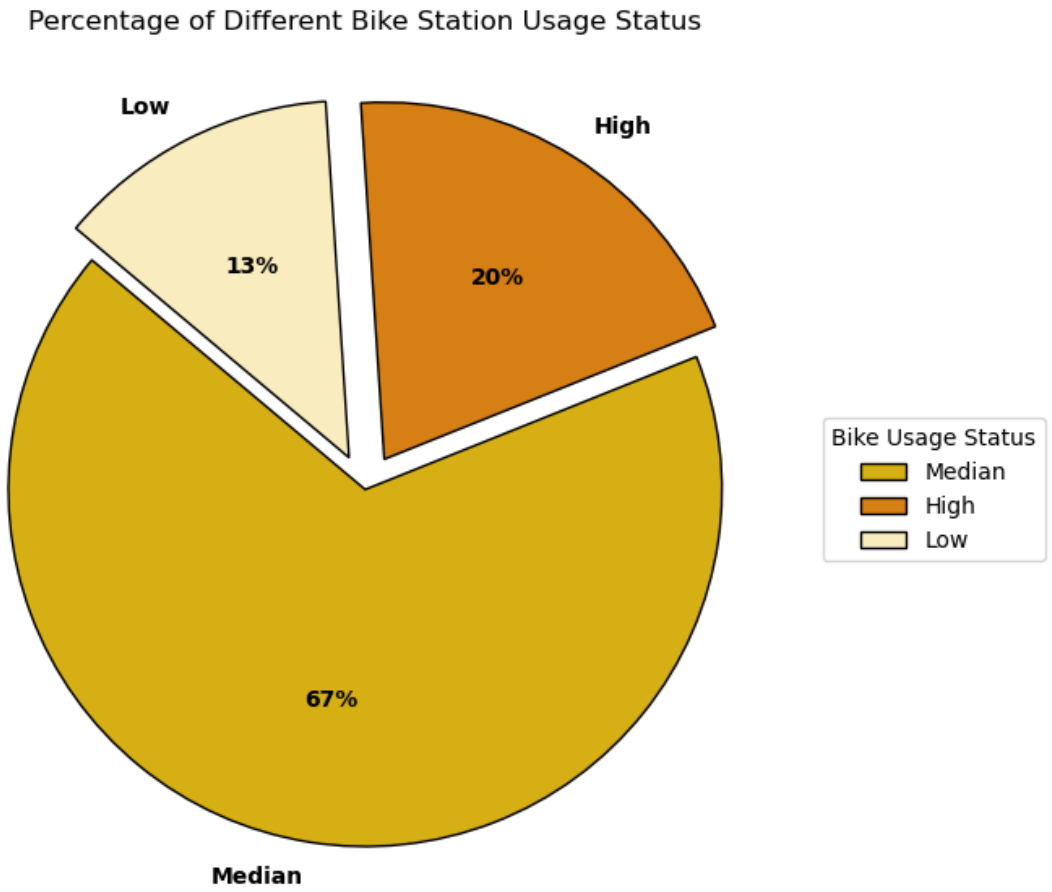
Project Overview

- **The Problem Area:**
  - How might we enhance the efficiency and accessibility of the bike share system in the SF Bay Area?
- **Approach**
  - Building a machine learning model to predict usage patterns, so that we can optimize bike availability through strategic redistribution, ensuring they are accessible where and when they are most needed.
- **Potential Impacts**
  - Better user experience and increased usage by 10%
  - Cost savings on transportation by 5%
  - Reduction in CO2 emissions by 5%

Dataset Overview

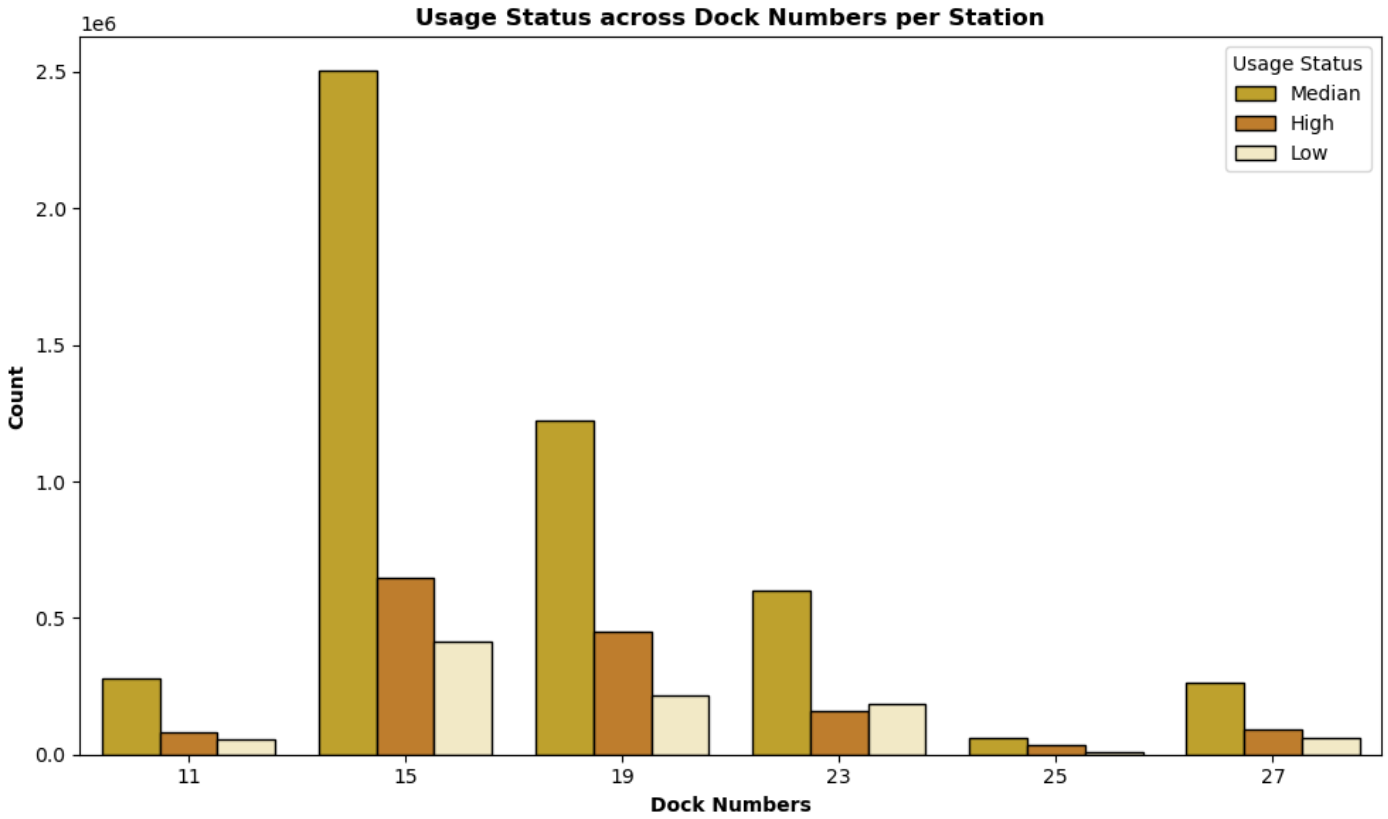
- **Original Datasets:**
  - Station, Status, and Weather datasets
  - 71,984,434 rows; 7 columns, 4 columns, and 24 columns respectively
- **Merged, Filtered, and Cleaned Dataset:**
  - 7,337,194 rows; 15 columns (13 numerical and 2 datetime)
  - Station ID, Date/Time, Available Bikes, Available Docks, Total Docks, and Weather Info

EDA



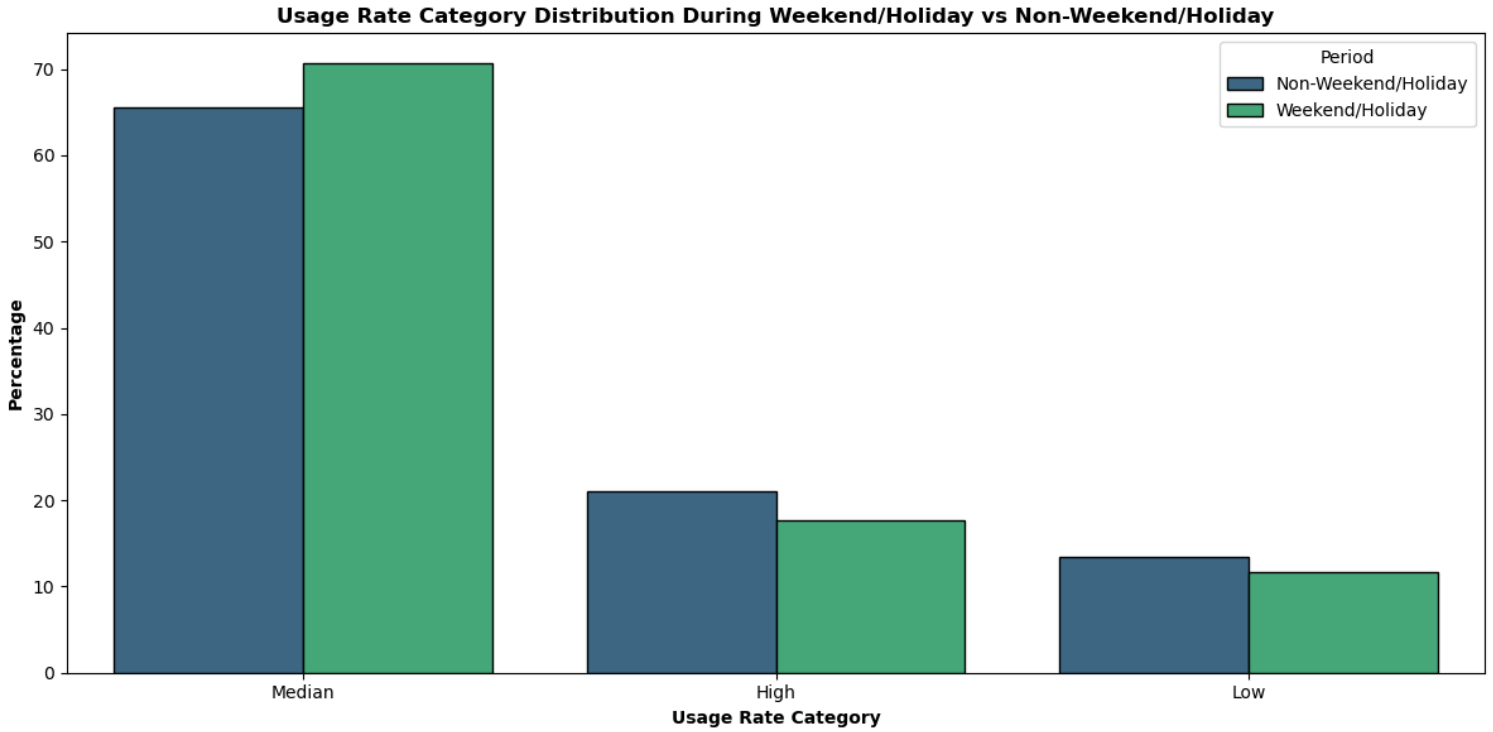
- **Target Variable - Bike Usage Rate**
  - Median: 67%
  - High: 20%
  - Low: 13%

EDA



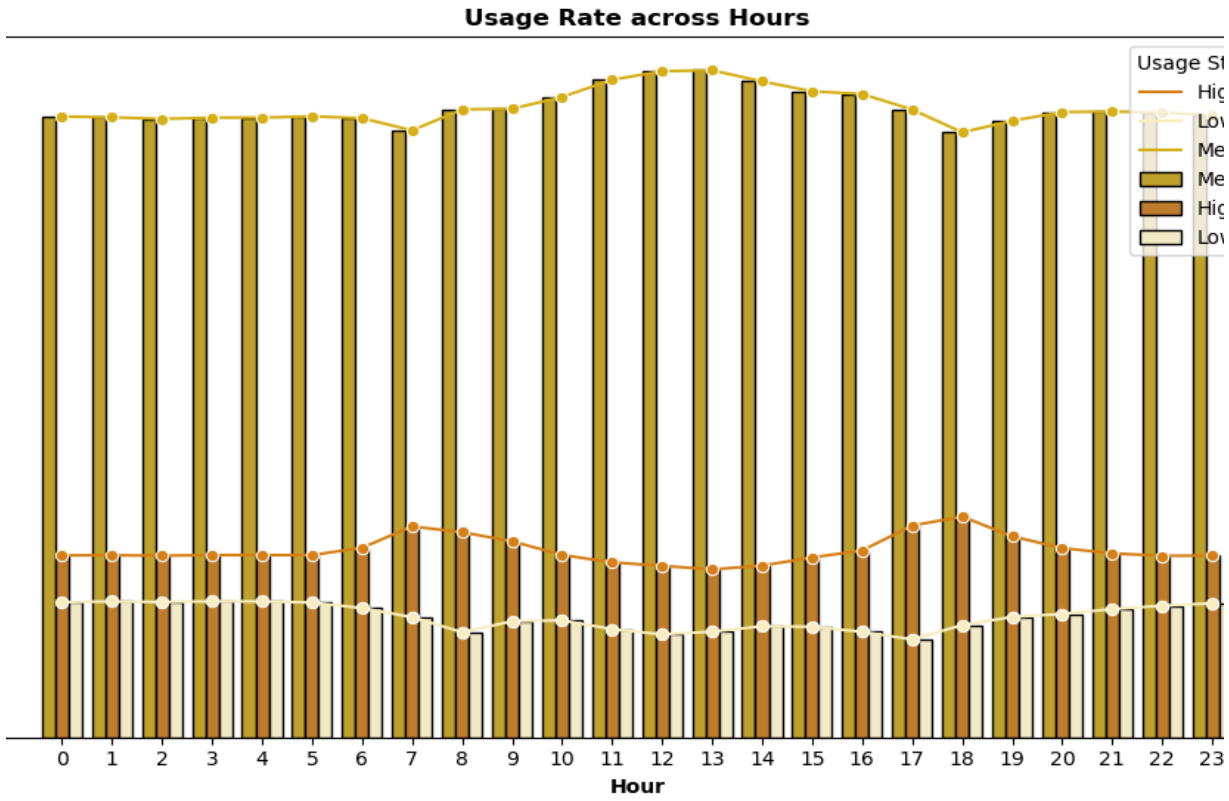
- The lower the dock number of the station, the more likely it is to have either a higher or lower usage rate.
- This is expected, as smaller bike stations' usage rates are more sensitive to the available bike and dock numbers.

EDA



- High usage rates:
  - 21% to 17%
- Low usage rates:
  - 13% to 11%
- Median usage rates:
  - 65% to 70%

EDA

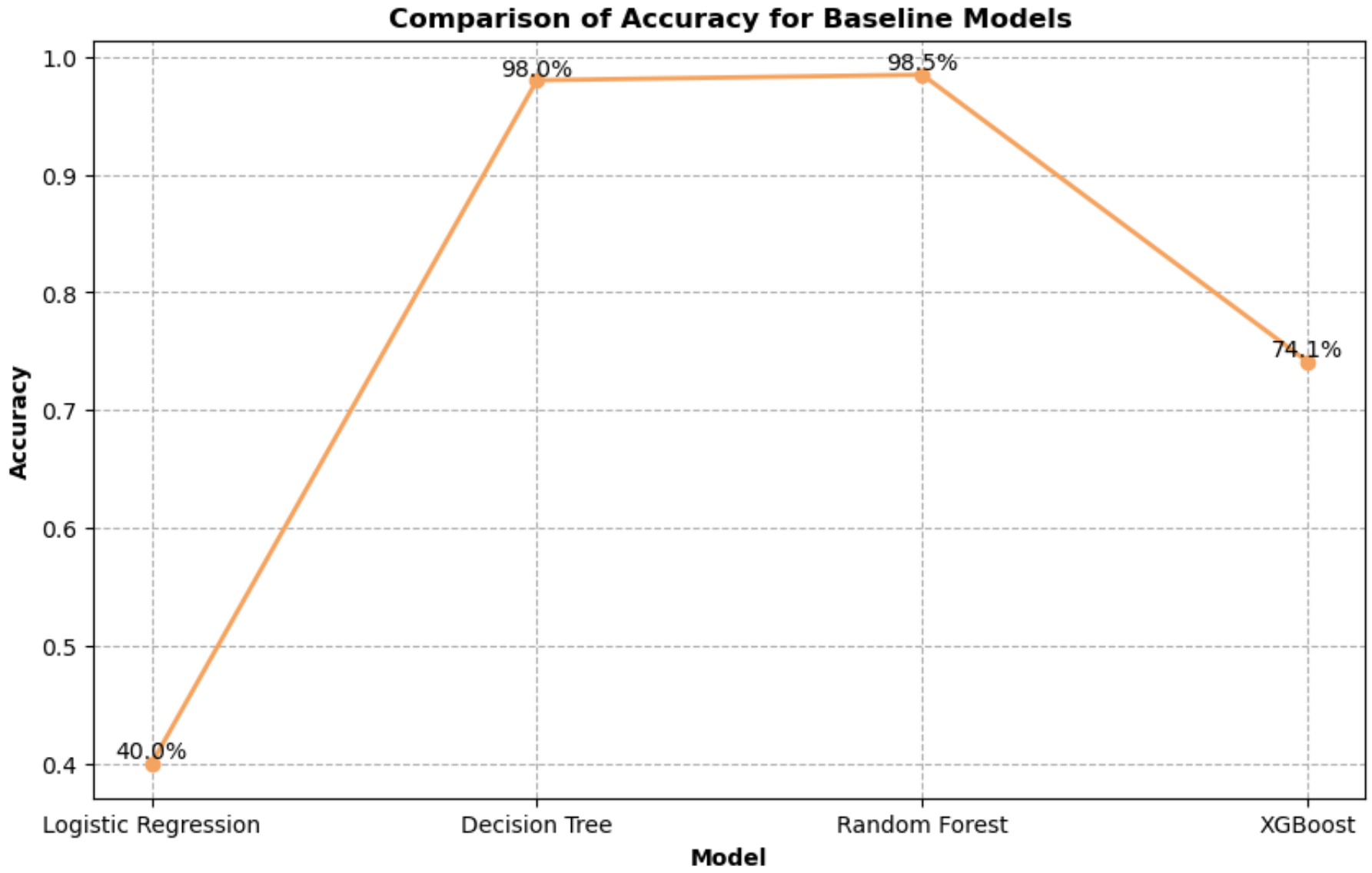


- 7 - 9 AM and 5 - 7 PM have the highest high bike usage rates during the day.
- 8 AM and 5PM have the lowest low bike usage rates during the day.
- This could be commuter influence.

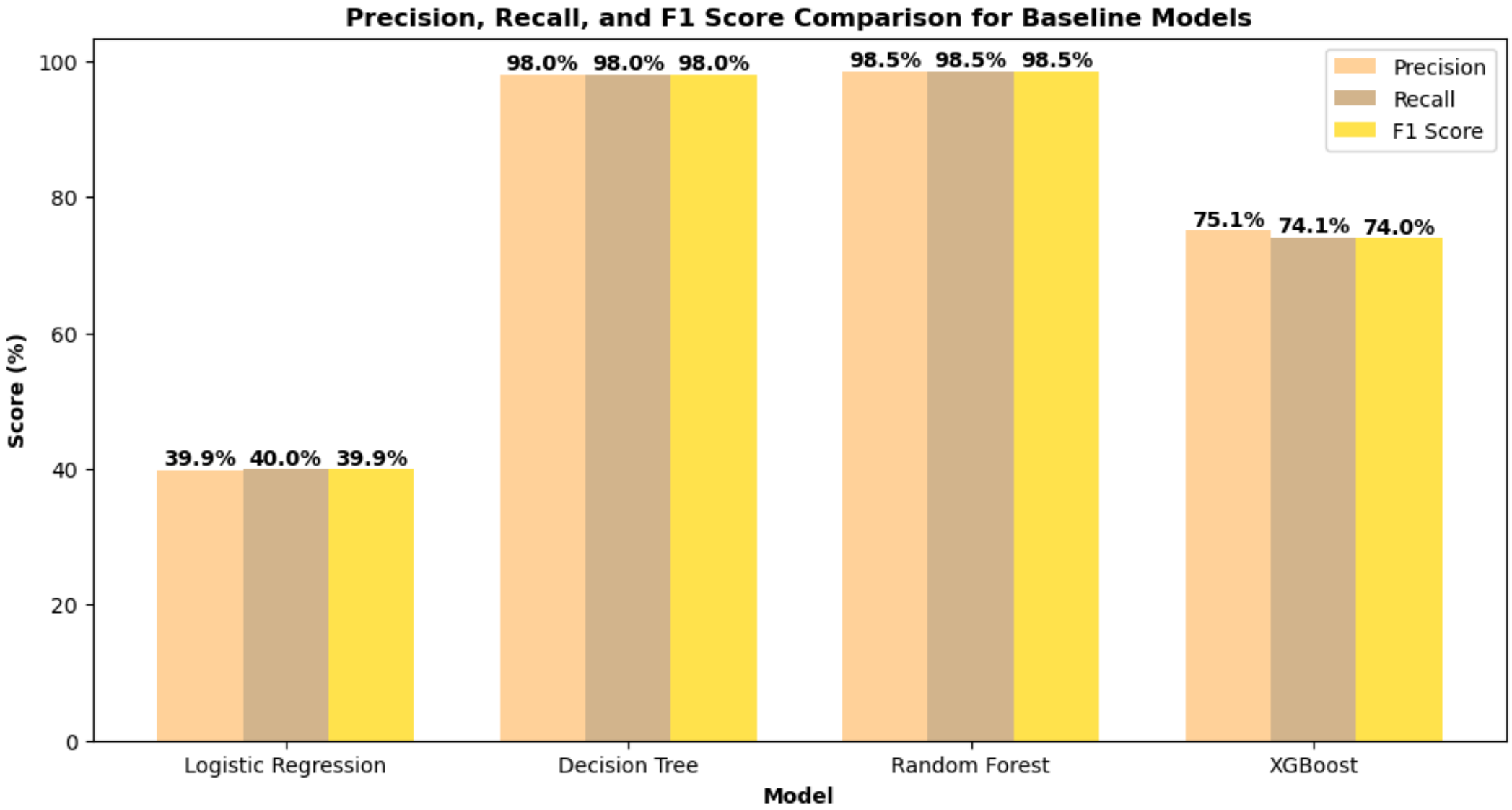
Summary of Baseline Models

Model	Train Score (%)	Test Score (%)	Elapsed Time (seconds)
Logistic Regression	39.92%	39.97%	20.41
Decision Tree	100.00%	98.02%	238.01
Random Forest	99.99%	98.48%	8055.33
XGBoost	74.12%	74.10%	174.61

Summary of Baseline Models



Summary of Baseline Models



Next steps for advanced modeling

- **Hyperparameter Tuning**
  - Grid search
  - Cross validation
- **Fit the models with the best parameters**
  - Compare baseline models vs. tuned models
  - Compare ROC AUC curve for each tuned model
- **Model Selection**
  - Model evaluation

Thank you