

Introduction

- Efficient transportation systems are the backbone of urban economies.
- Increasing population density leads to traffic congestion and fluctuating ride demands.
- NYC heavily relies on its diverse modes of transit, including taxis subways and rideshares.

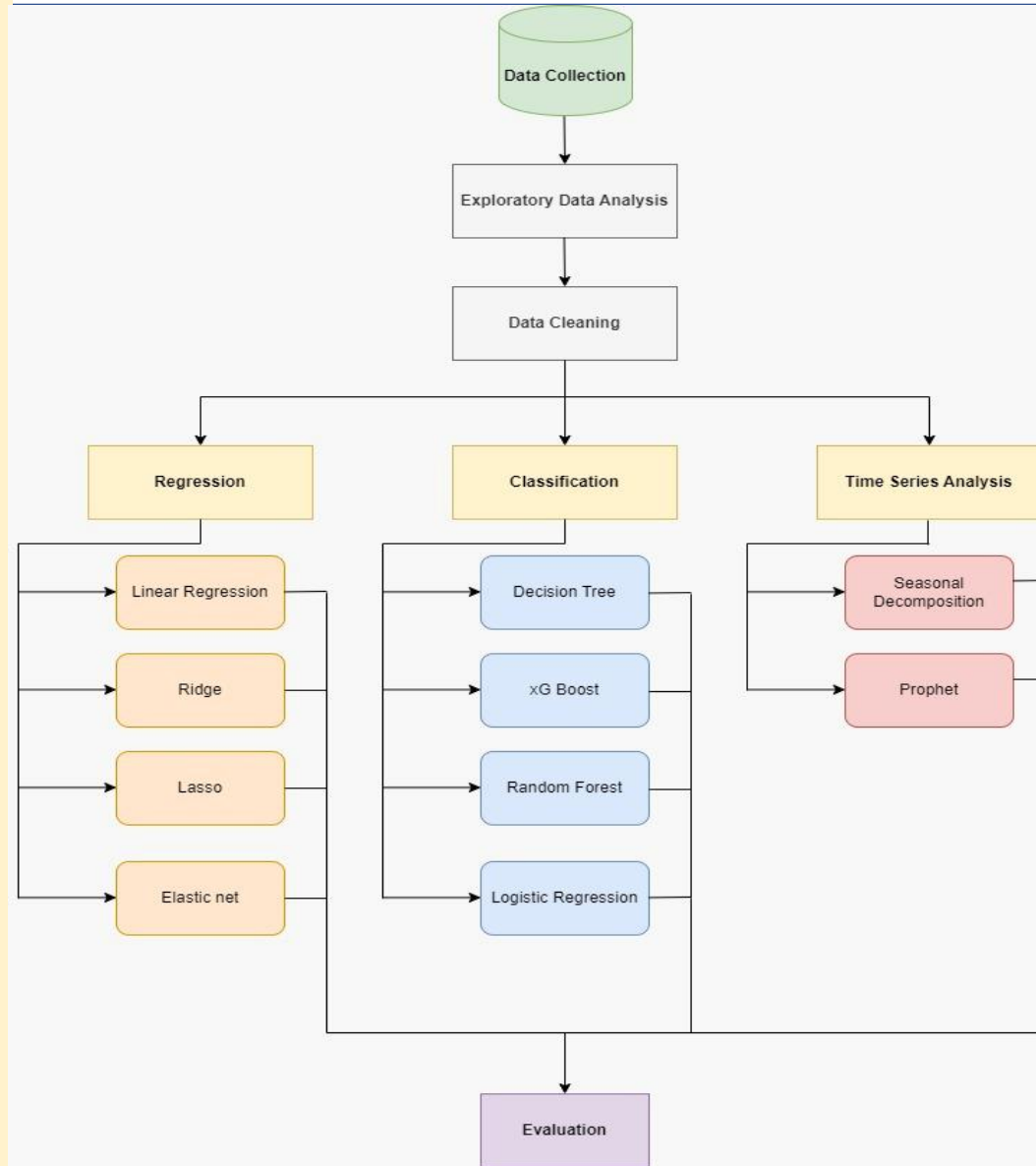
Motivation

1. Understanding how transportation operates can improve urban mobility and reduce inefficiencies.
2. Predicting fares dynamics, transportation patterns and passenger's behavior can help optimize resource allocation
3. Insights from this study can support sustainable urban planning and smart transit systems.

Questions?

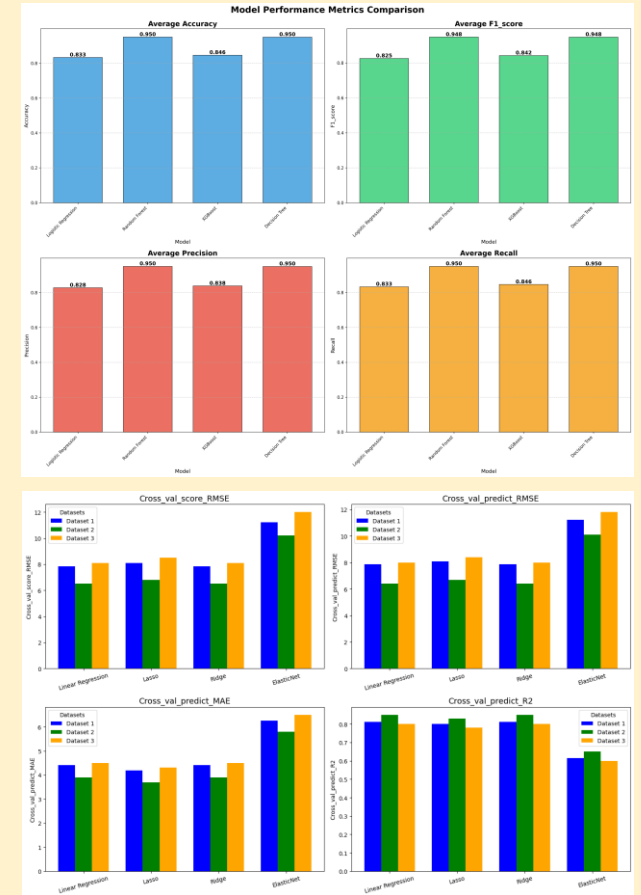
- Can we predict fair amount based on other features like trip distance, passenger count, etc?
- Can we predict payment type based on other features?
- Can we forecast future demand and revenue based on past data?

Predictive Modeling and Machine learning Insights from NYC Taxi Data



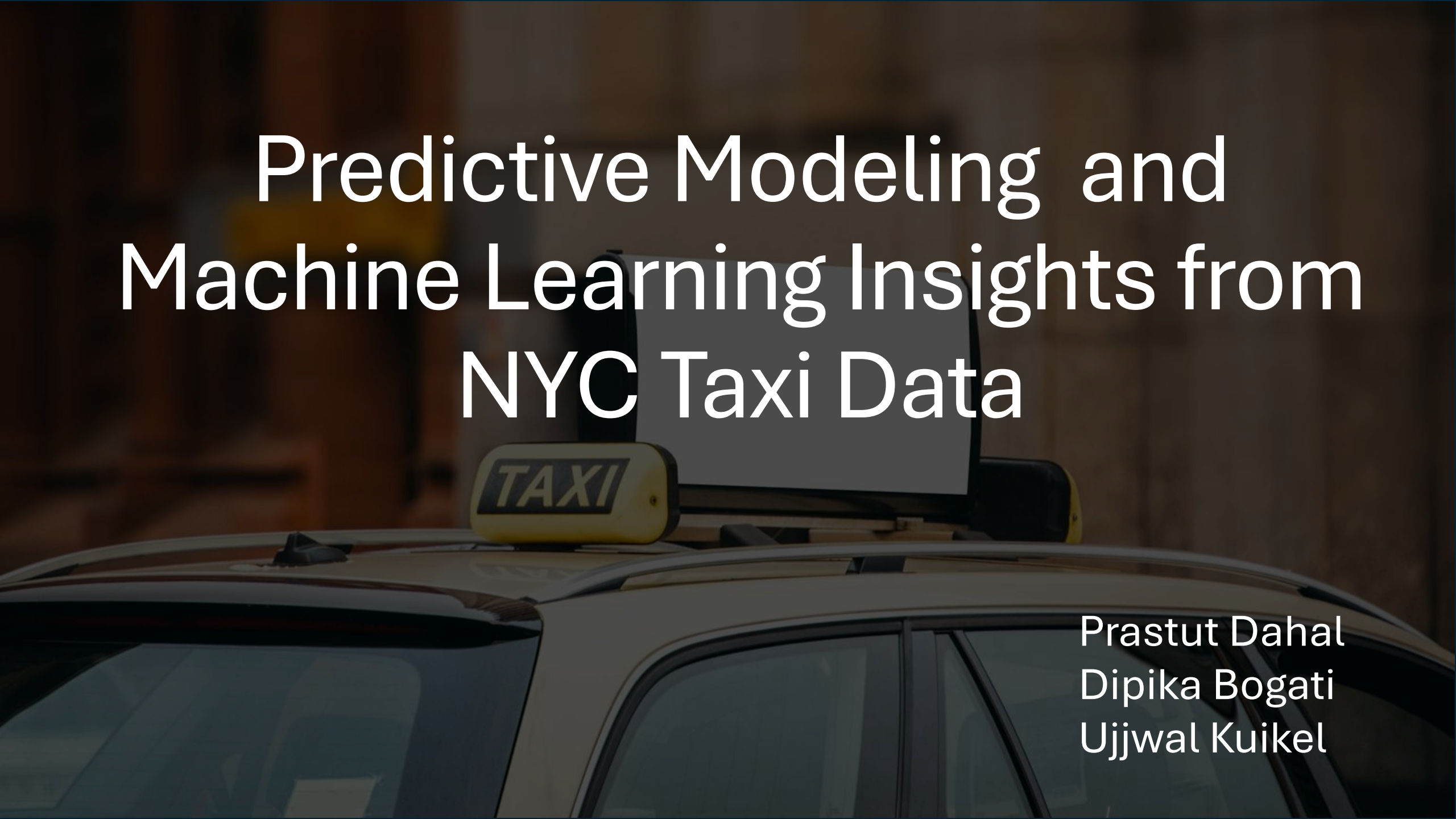
Results

- Ridge seems to be the most consistent performer on regression across all datasets, particularly in terms of balancing RMSE, MAE, and R^2 .
- Random Forest and Decision Tree consistently performed the best across all metrics, scoring 0.950 in accuracy, precision, and recall, and 0.948 in F1 score.



Prastut Dahal,
Ujjwal Kuikel,
Dipika Bogati

Predictive Modeling and Machine Learning Insights from NYC Taxi Data

The background of the slide is a photograph of a taxi, likely a yellow cab, with a 'TAXI' sign on its roof. The image is darkened and serves as a backdrop for the title text.

Prastut Dahal
Dipika Bogati
Ujjwal Kuikel

Introduction



Efficient transportation systems are the backbone of urban economies.



Understanding how transportation operates can improve the mobility and reduce inefficiencies



Predicting fares dynamics, transportation patterns and passenger's behavior can help optimize resource allocation



Insights from this study can support sustainable urban planning and smart transit systems.

Questions we answered



Can we predict fair amount based on other features like trip distance, passenger count, etc?

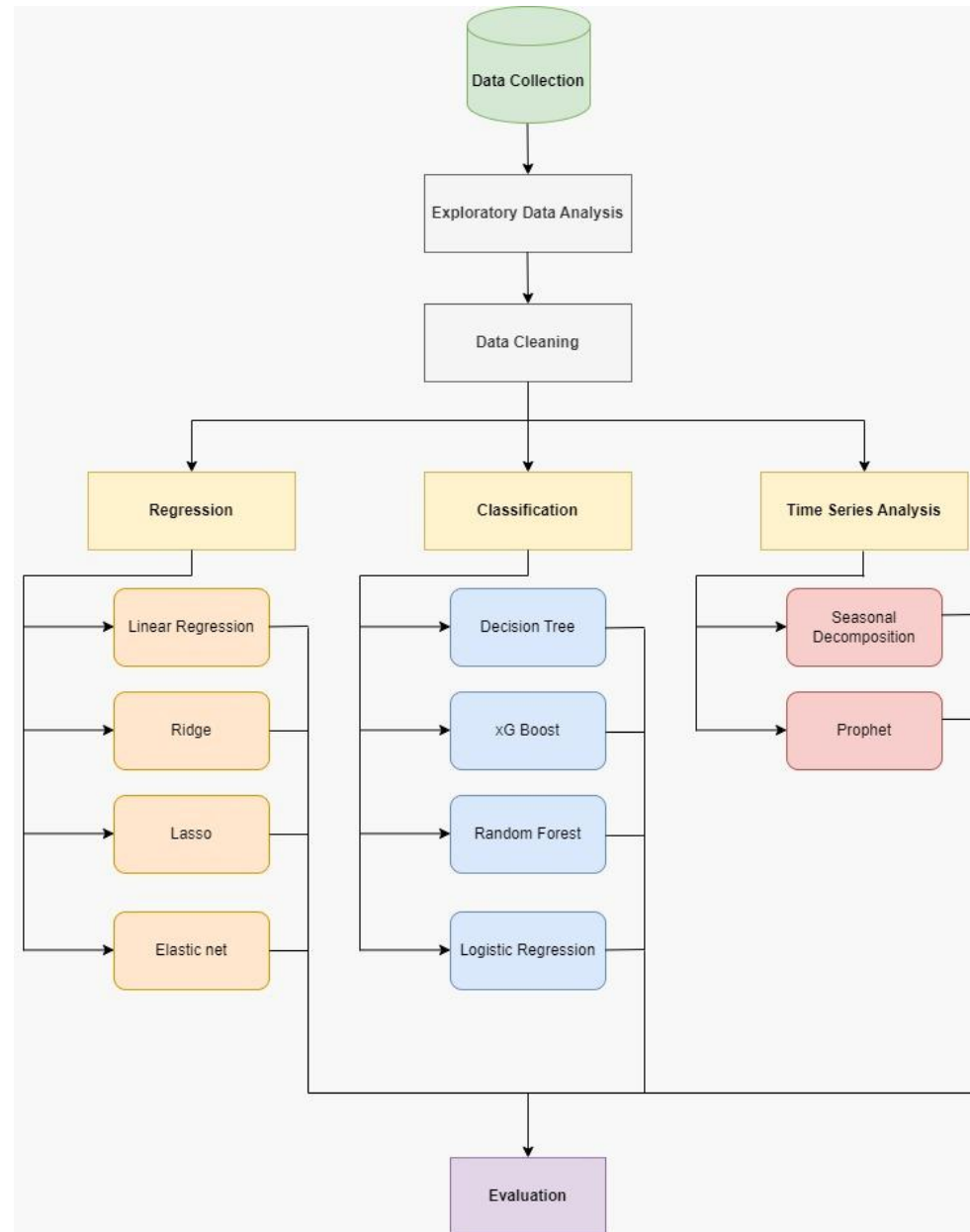


Can we predict payment type based on other features?



Can we forecast future demand and revenue based on past data?

Methodology



Dataset

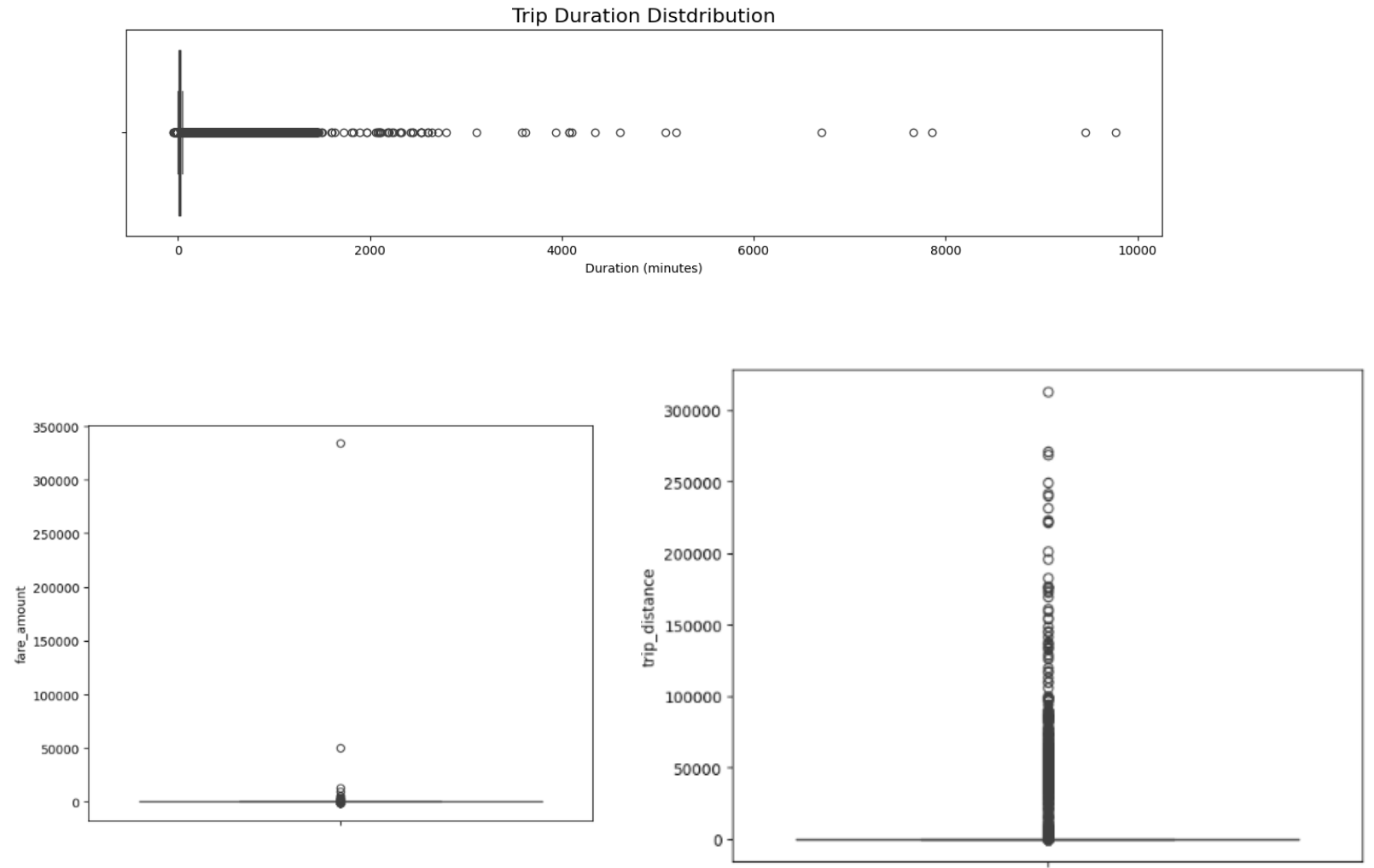
Source:

- NYC Taxi & Limousine Commission (TLC) Yellow Taxi Trip Records
- Open dataset with millions of trip records from NYC taxis

Key Features:

- **Trip Details:** Pickup and drop-off locations, timestamps, trip distance
- **Fare Information:** Total fare, tip amount, payment type
- **Passenger Data:** Number of passengers per trip
- **Time-Related Attributes:** Pickup time, dropoff time

EDA



Data Cleaning



Handling Missing Data:

Checked for missing values and imputed or removed rows/columns as needed.



Feature Engineering:

Created new features (e.g., trip_duration, hour_of_day, day_of_week) based on raw data.



Encoded Categorical Variables:

Used one-hot encoding or label encoding for categorical data.



Handled Duplicates:

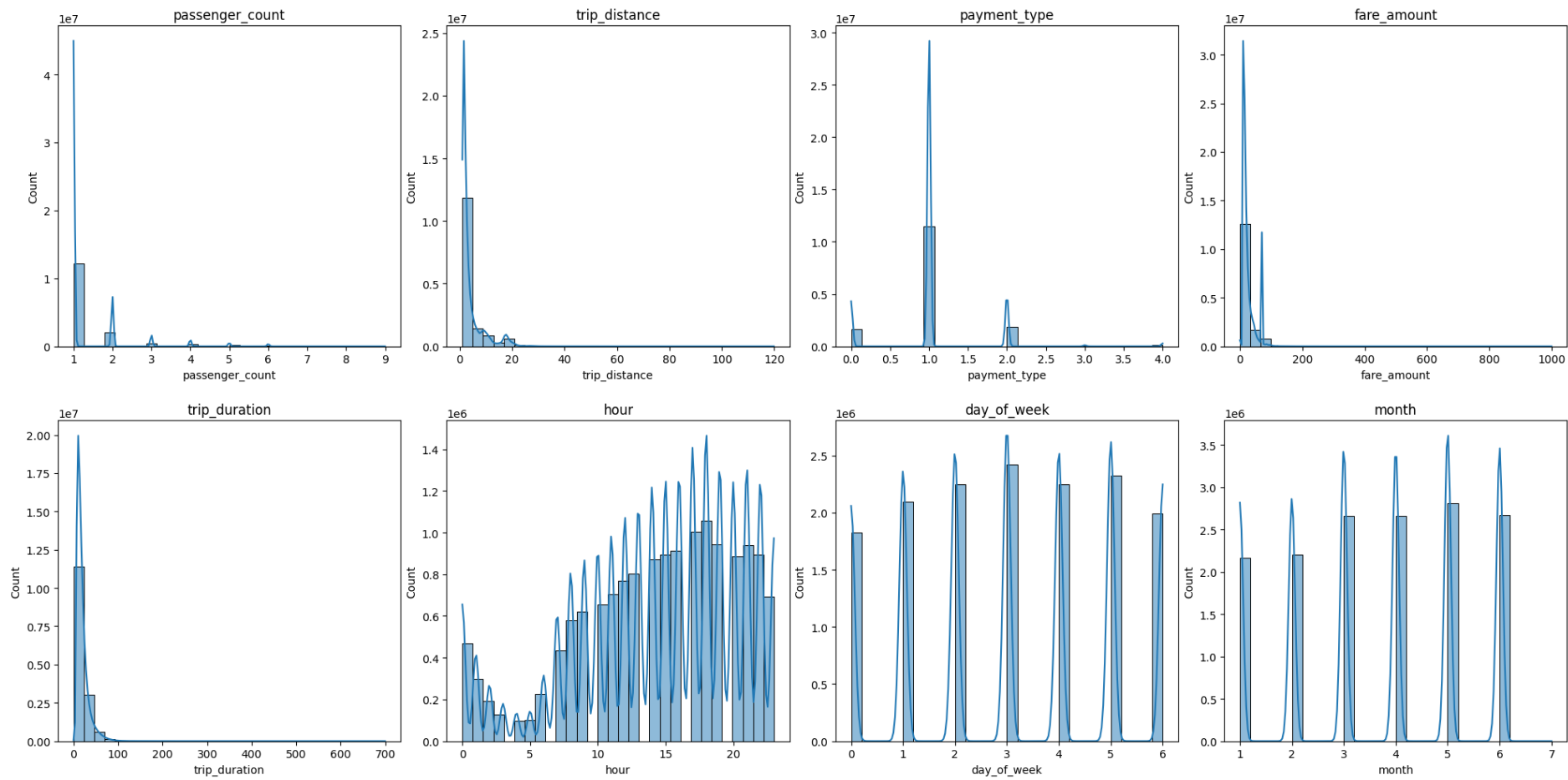
Checked for and removed duplicate rows to avoid skewing results.



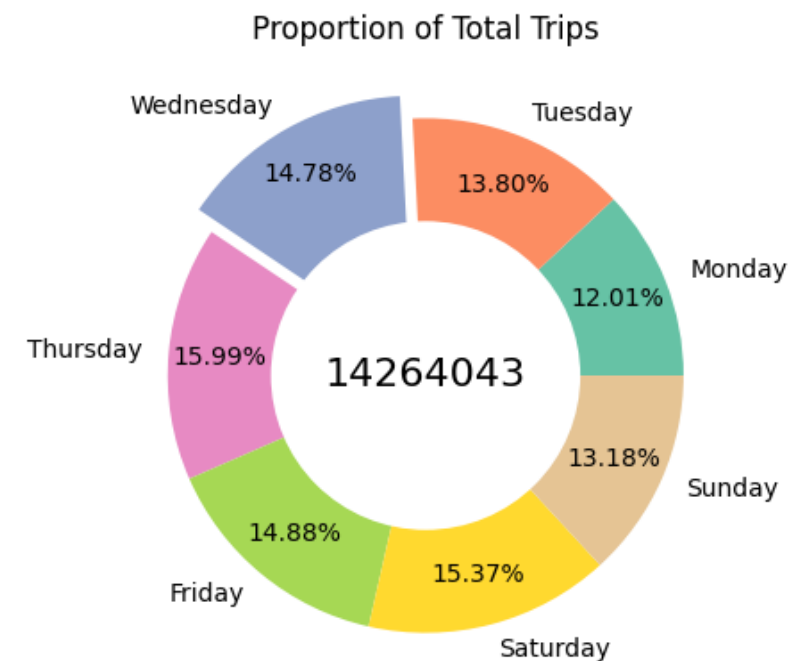
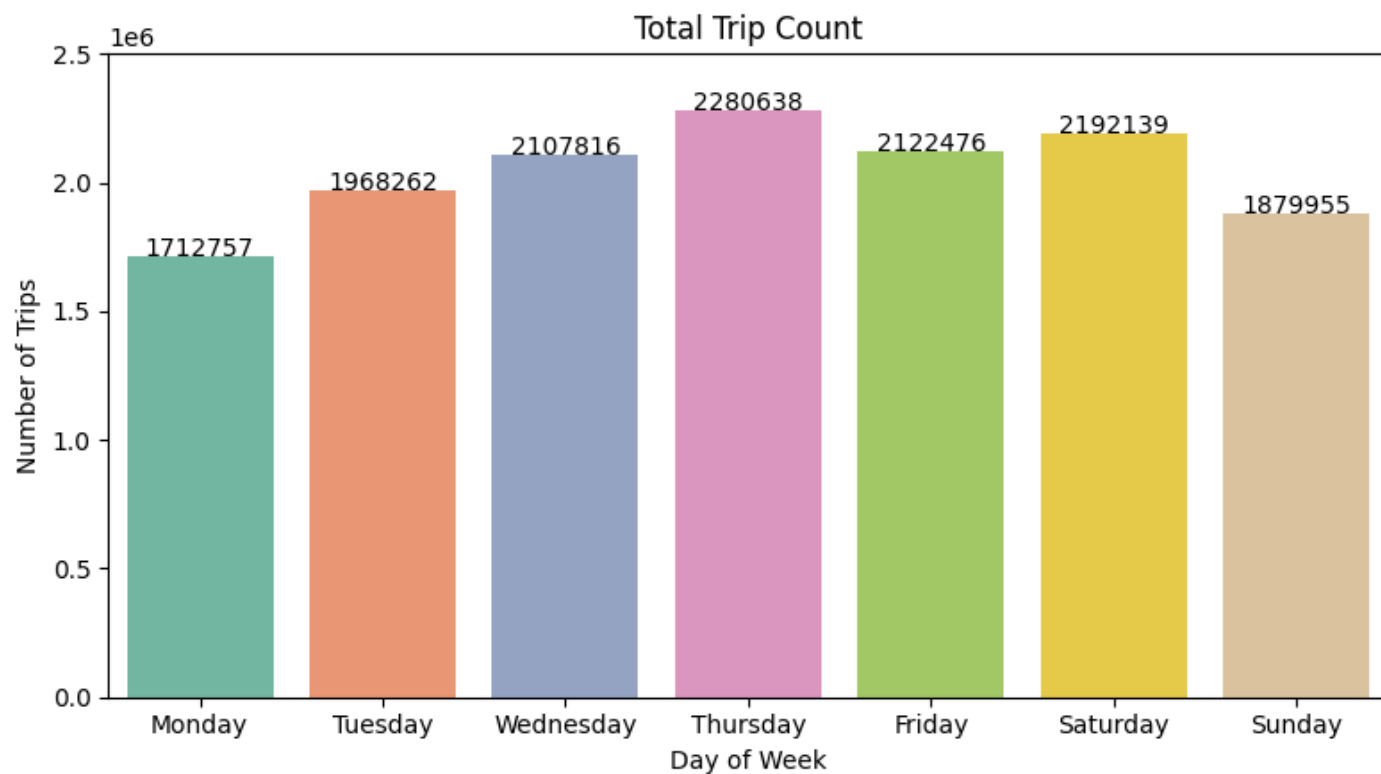
Outlier Detection and Removal:

Identified and removed outliers in numerical features (e.g., extreme trip fares, durations).

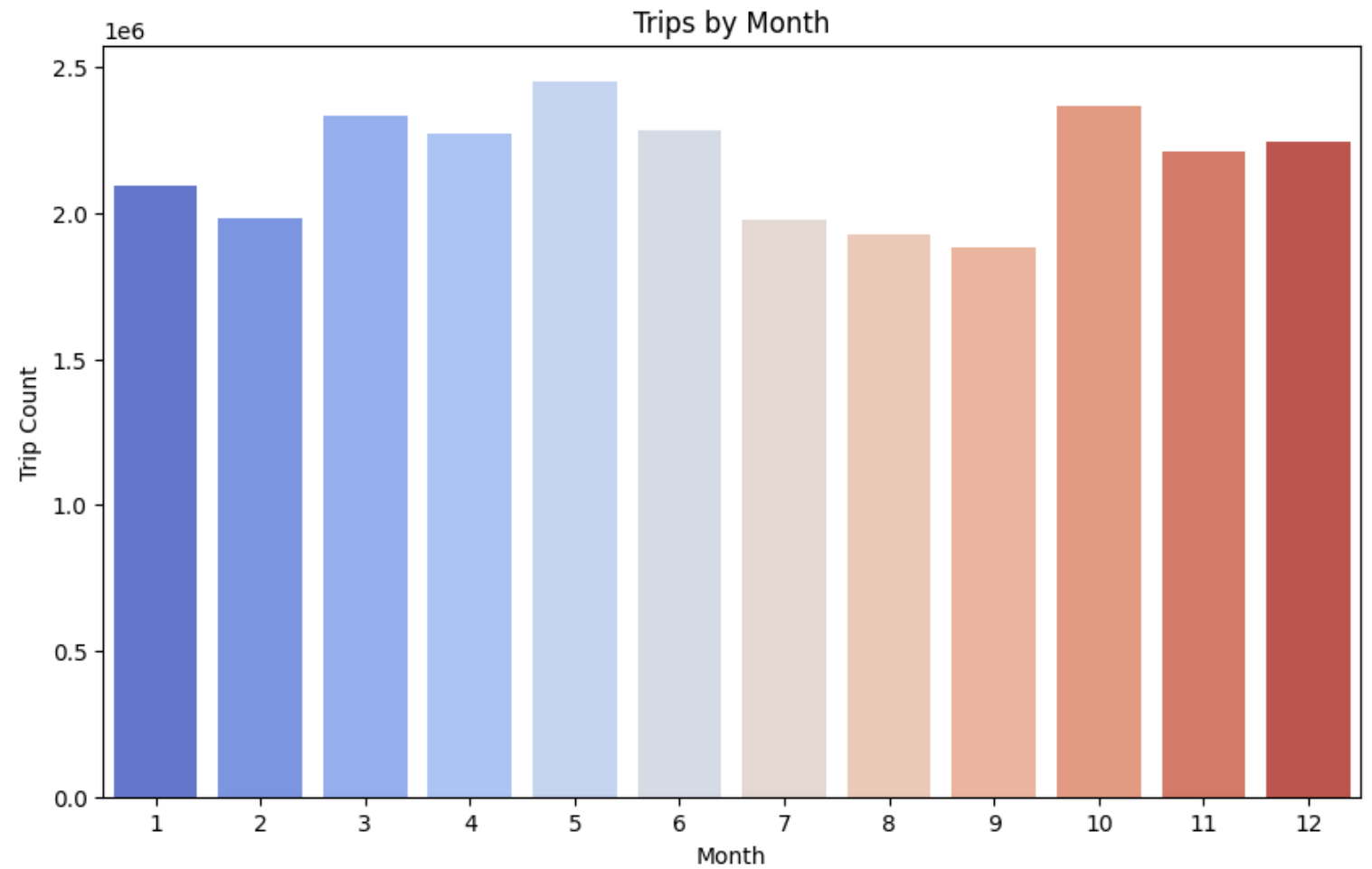
EDA



Trip Distribution on Day of Week



Trips By Month



Regression

Linear Regression:

Fits a straight line that best represents the linear relationship between features and target.

Lasso Regression:

Uses L1 regularization to perform feature selection by driving some coefficients to zero.

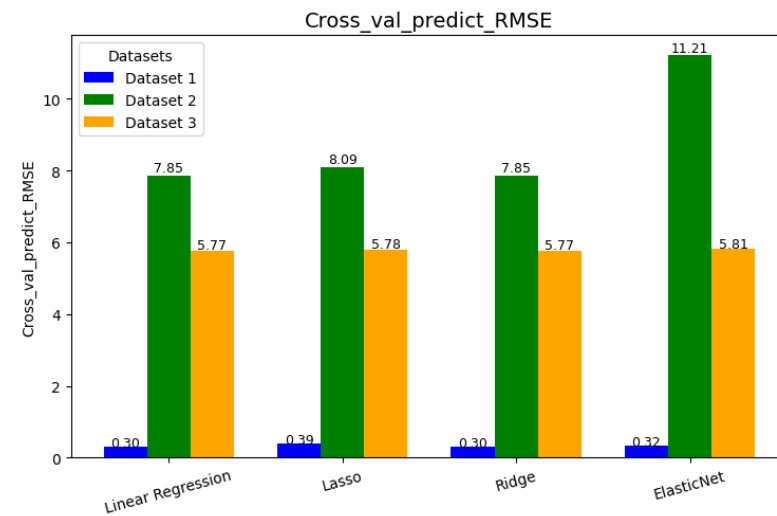
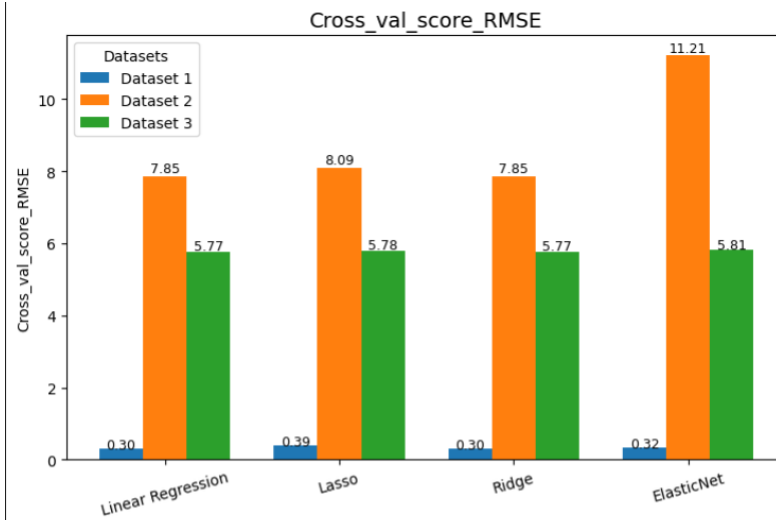
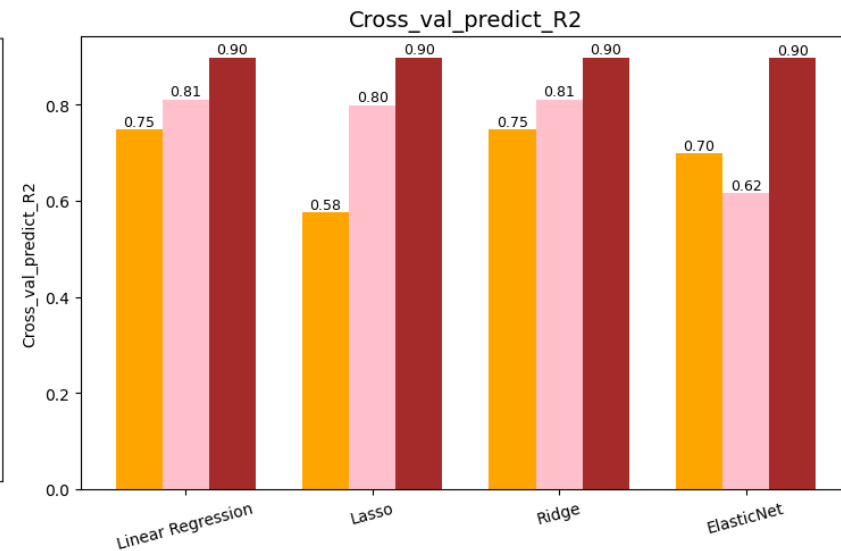
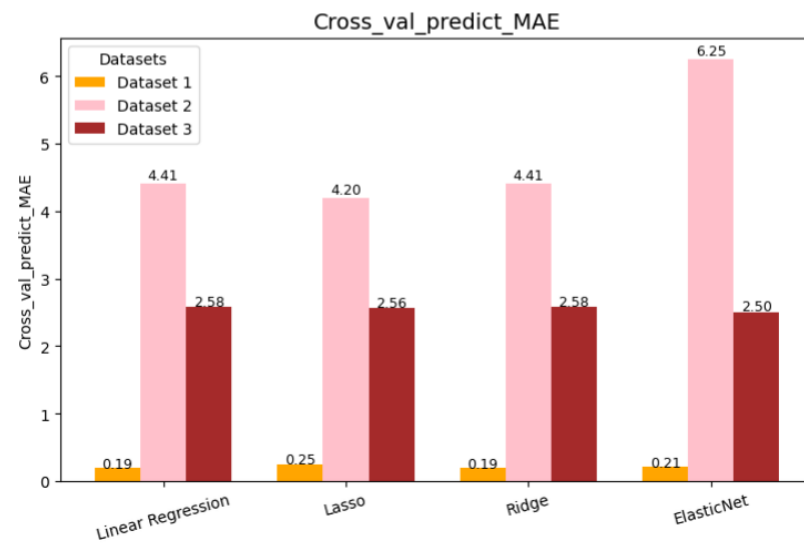
Ridge Regression:

Applies L2 regularization to prevent overfitting by shrinking coefficients.

Elastic Net:

Combines L1 and L2 regularization for balanced feature selection and coefficient shrinkage.

Evaluation

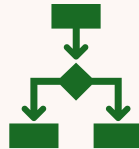


Classification



Logistic Regression:

Assumes there is linear relationship between features and target values, using a sigmoid function to transform linear combinations into probabilities.



Decision Tree:

Captures both linear and non-linear relationships by hierarchically splitting data based on feature importance, creating a tree-like model of decisions.



Random Forest:

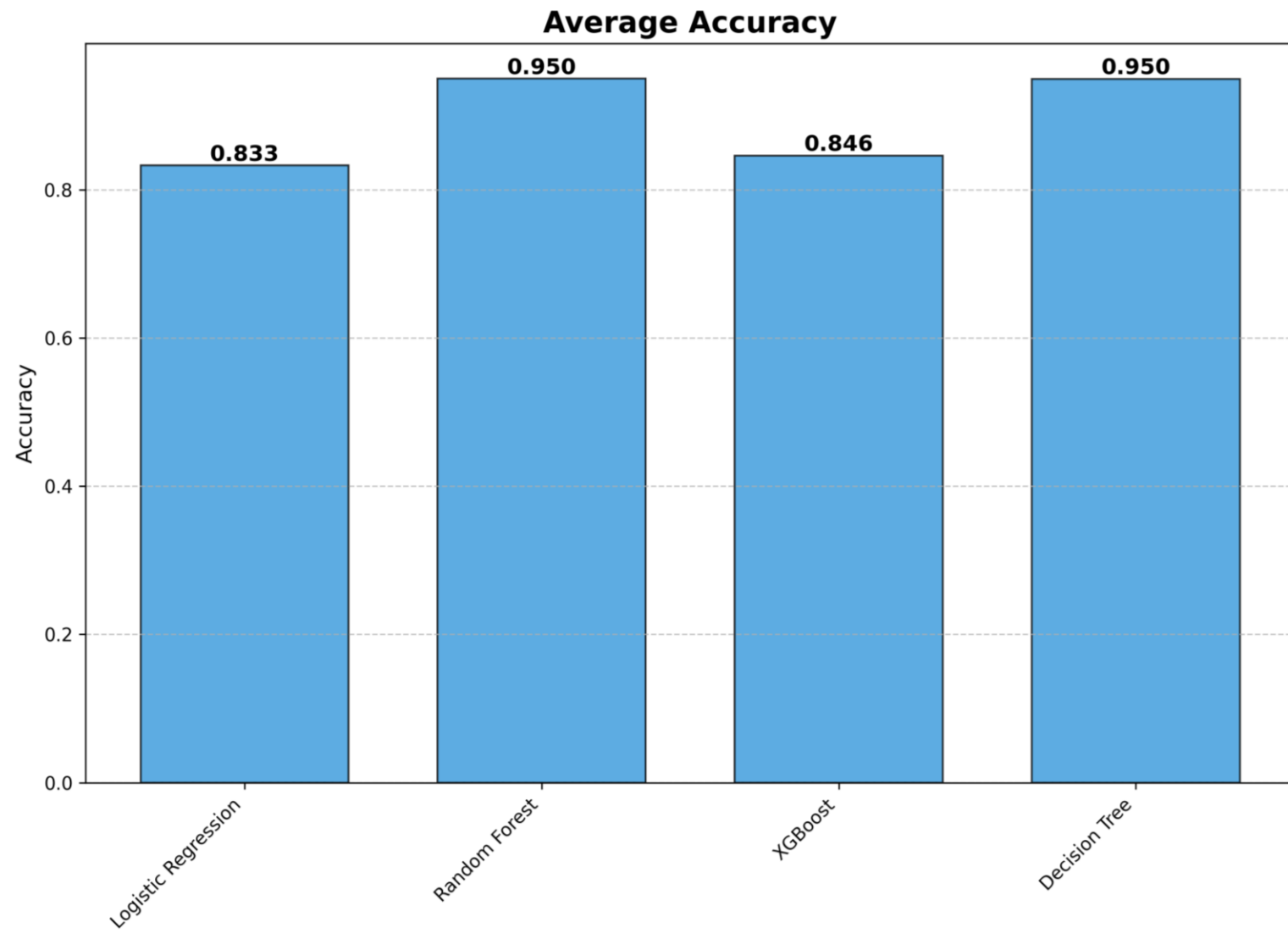
Combines multiple decision trees to make better predictions by using bootstrap aggregation and reducing overfitting through tree diversity.



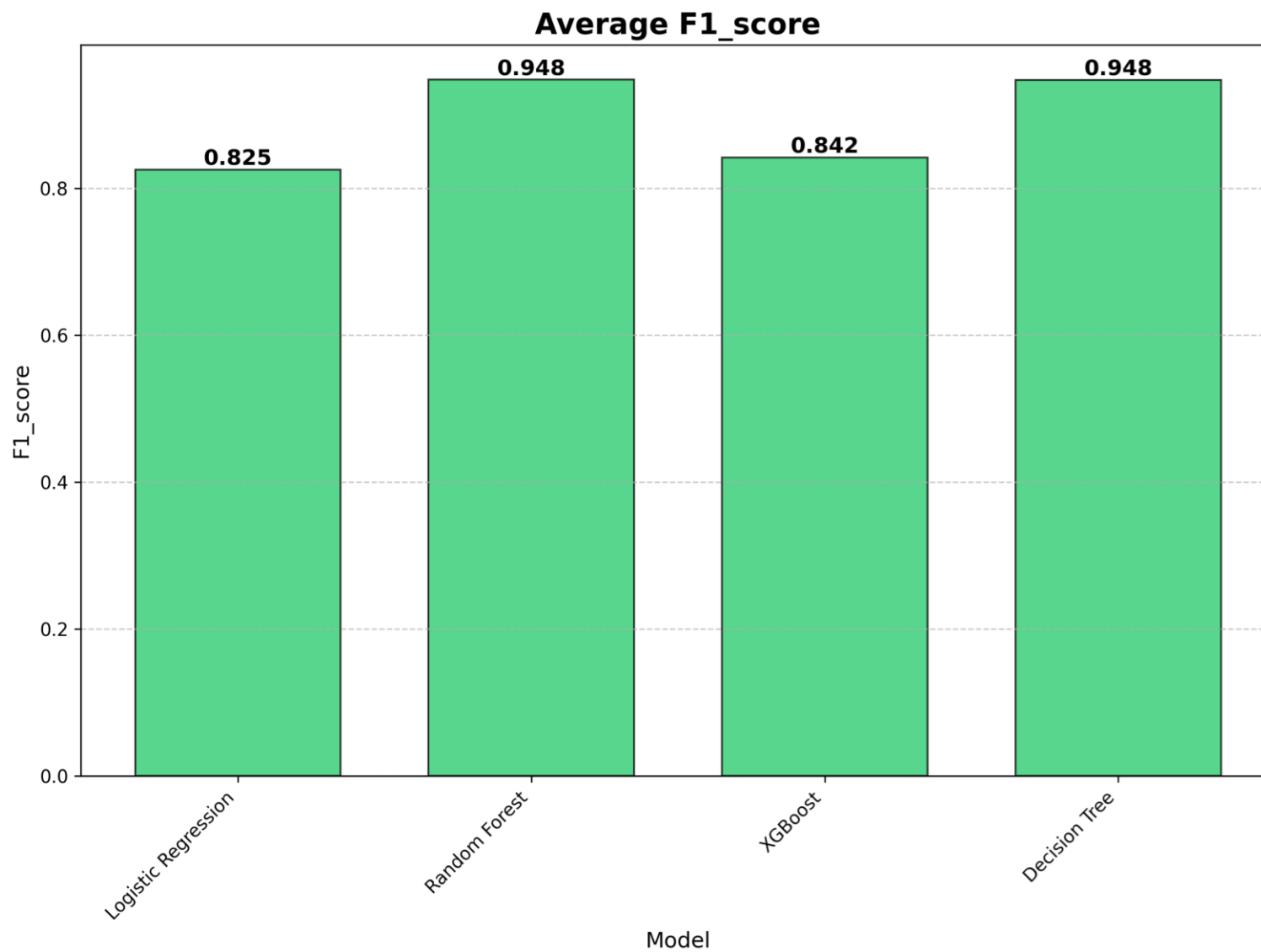
XGBoost:

Unlike random forest, it creates one decision tree at a time, with each tree strategically designed to fix the mistakes from the previous tree using gradient descent.

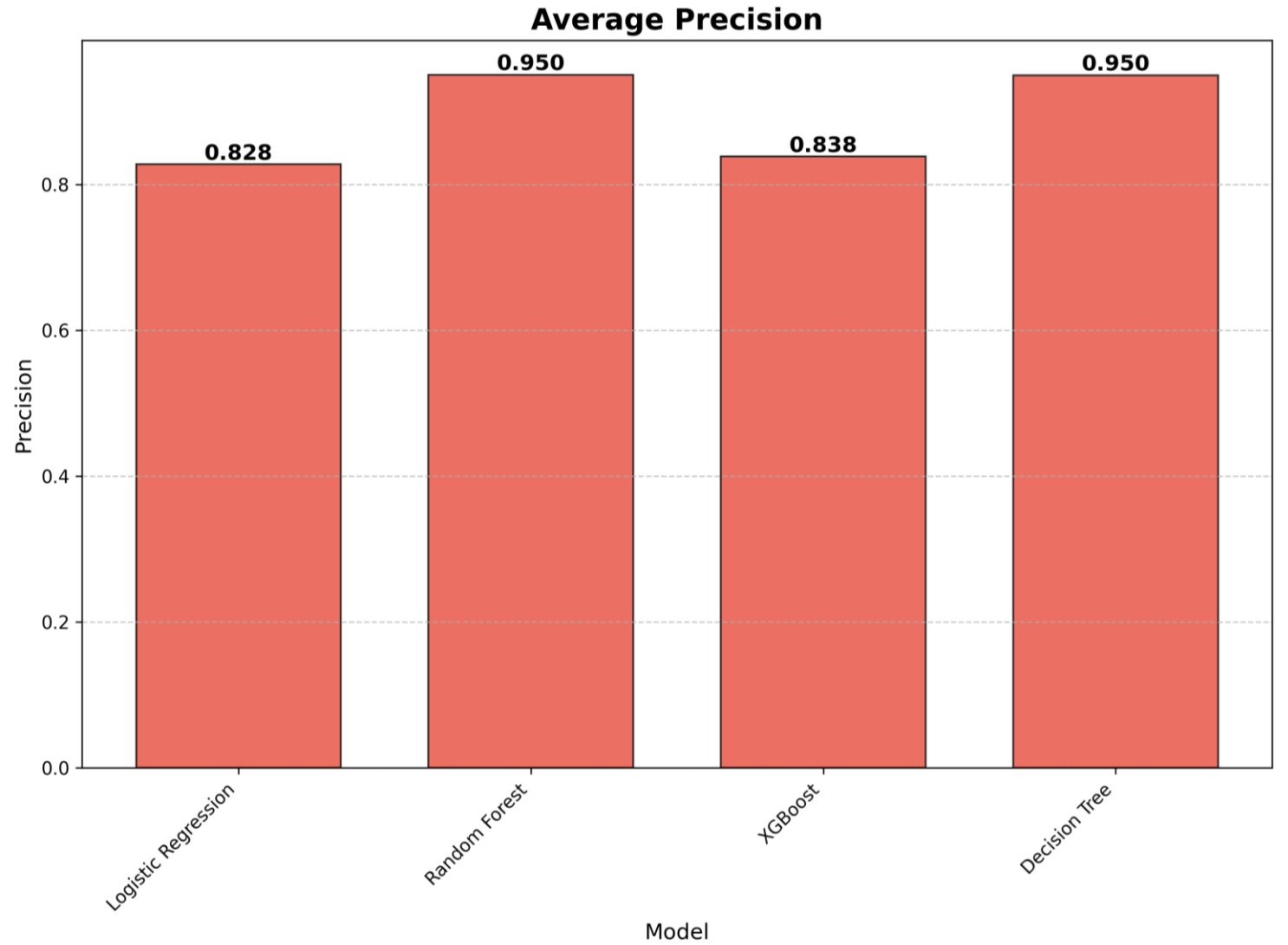
Average Accuracy



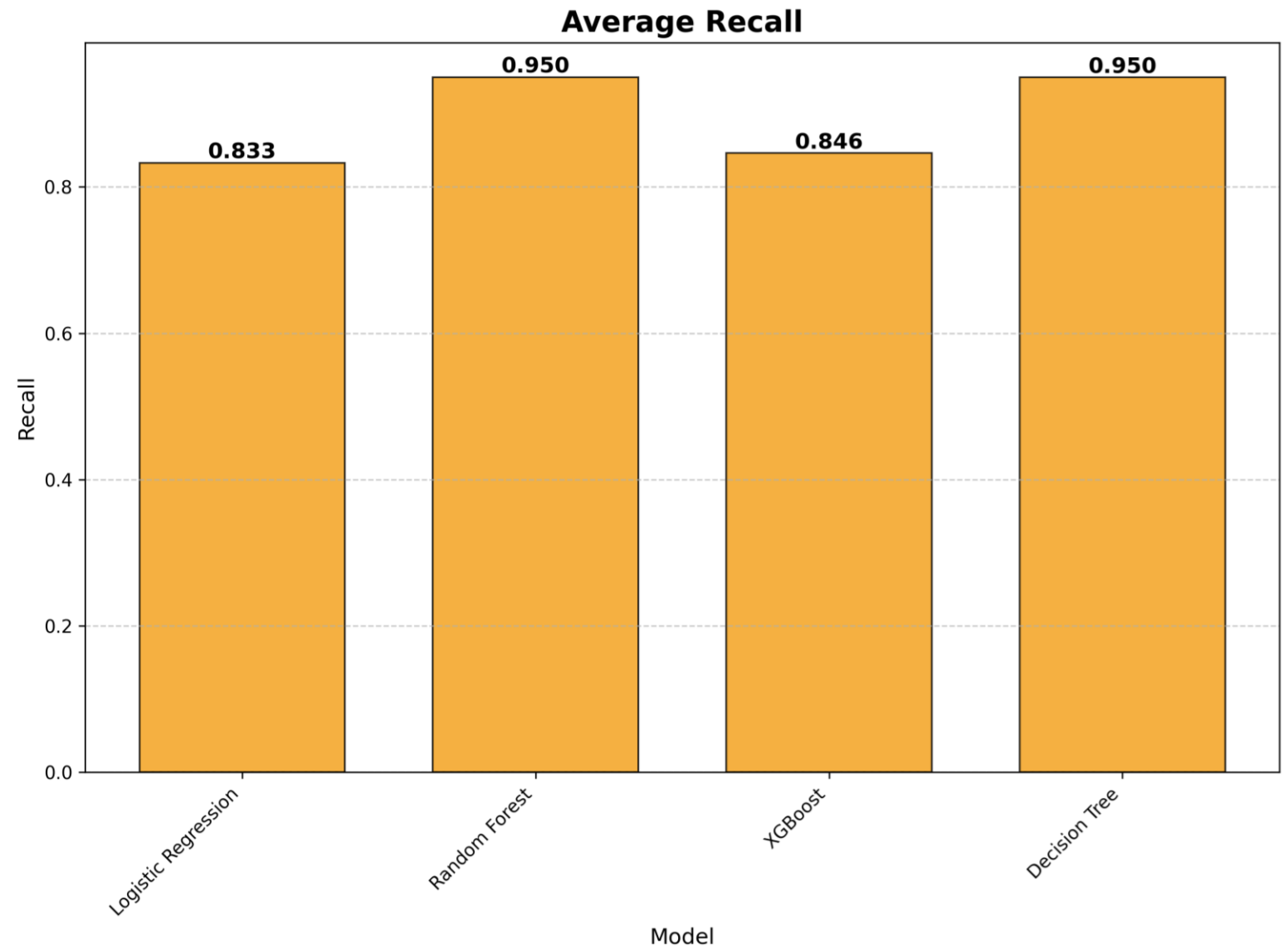
F1_Score



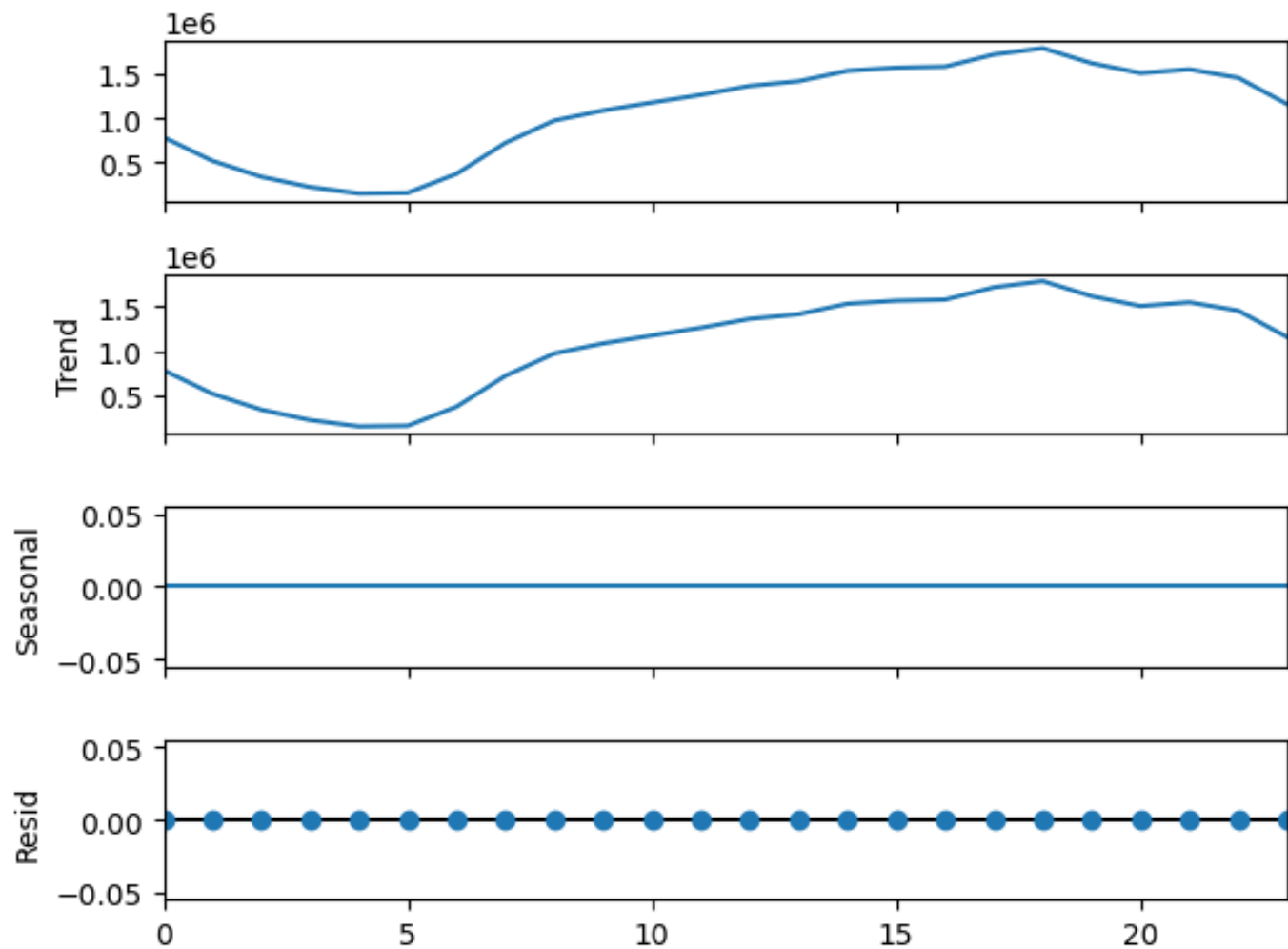
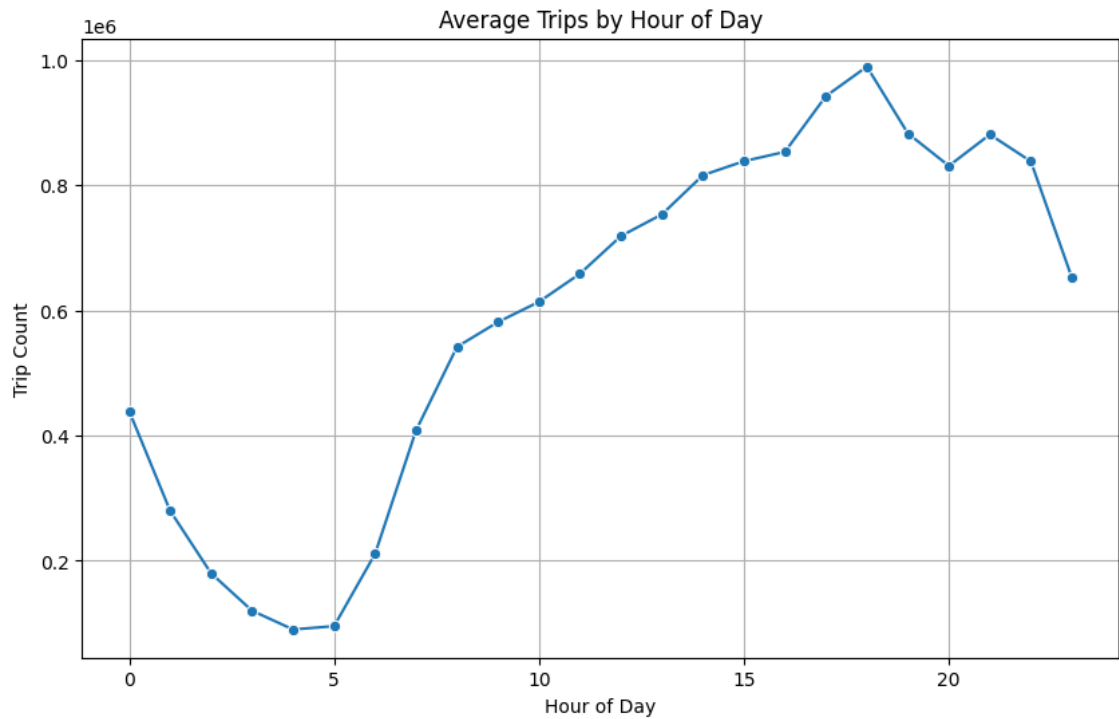
Precision



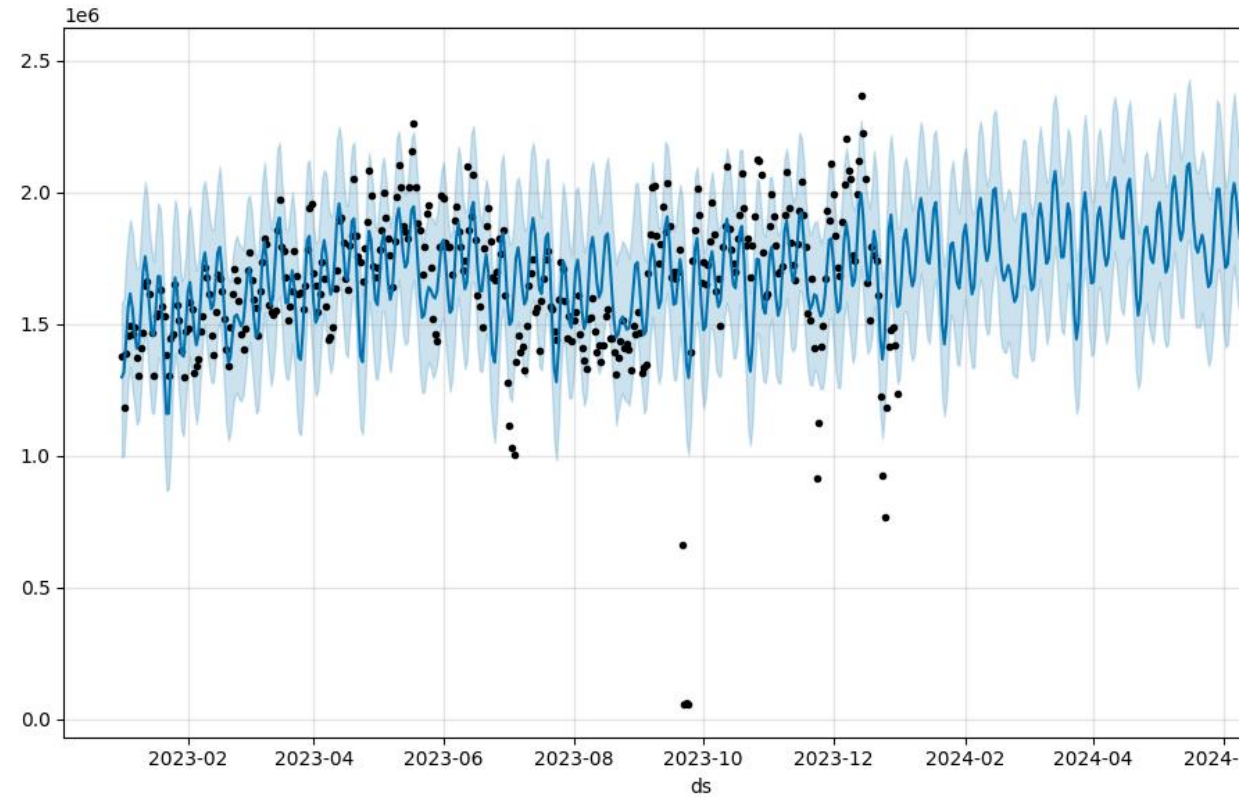
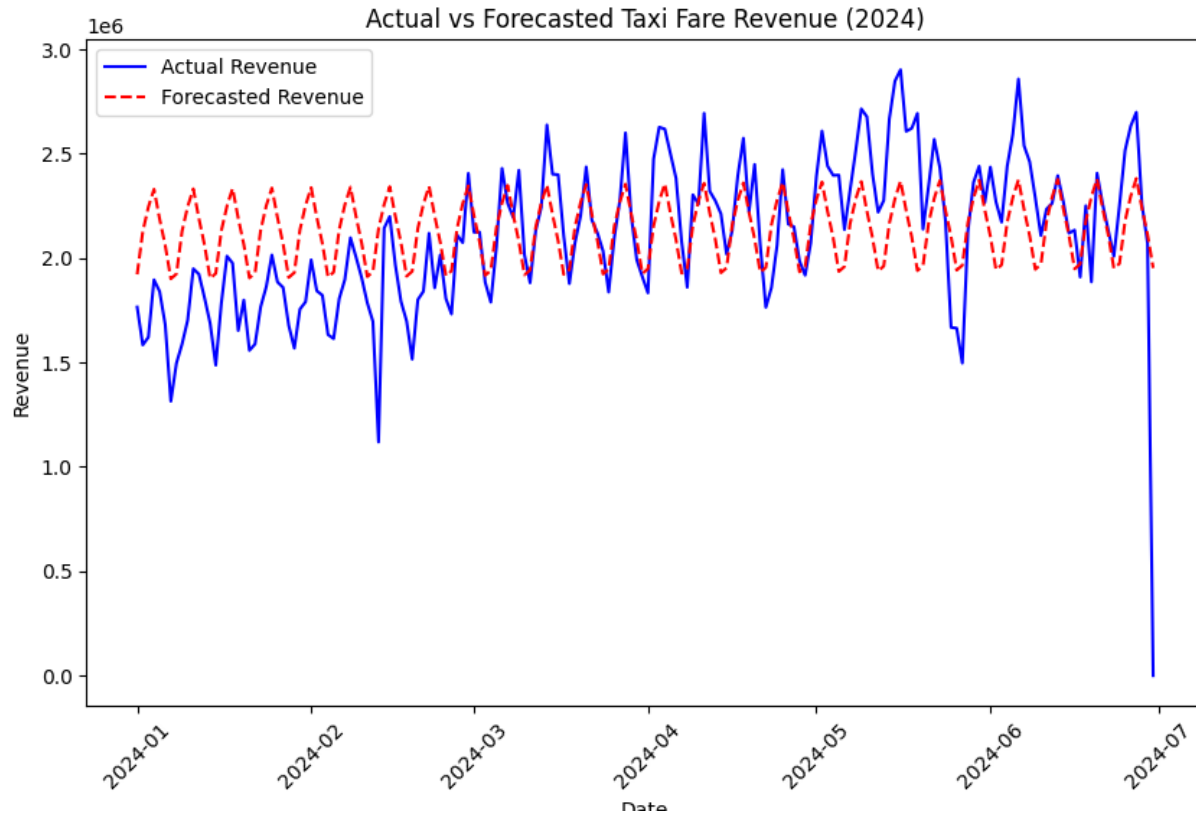
Recall



Hourly Demand

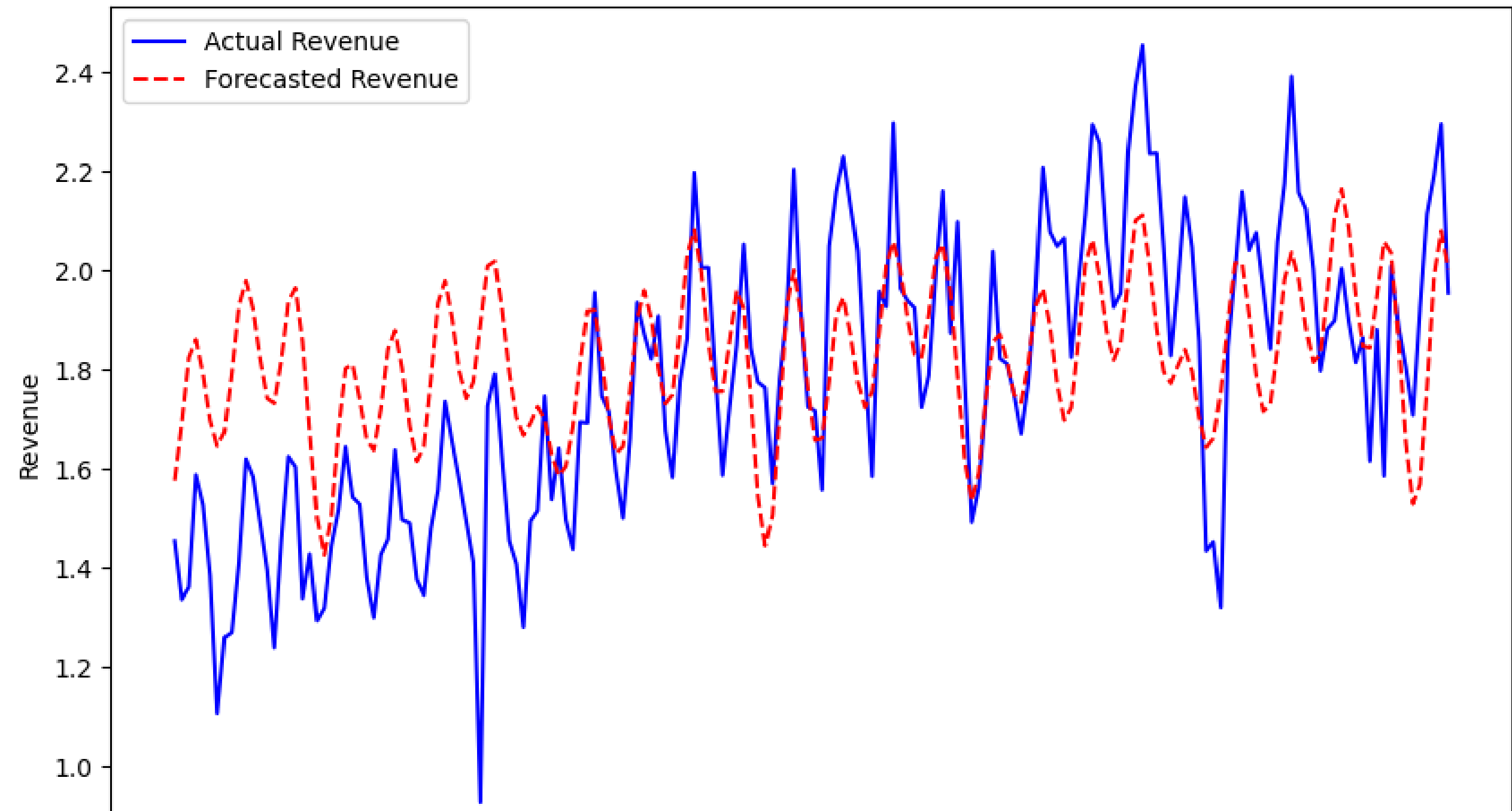


Actual vs Forecasted Fare Revenue

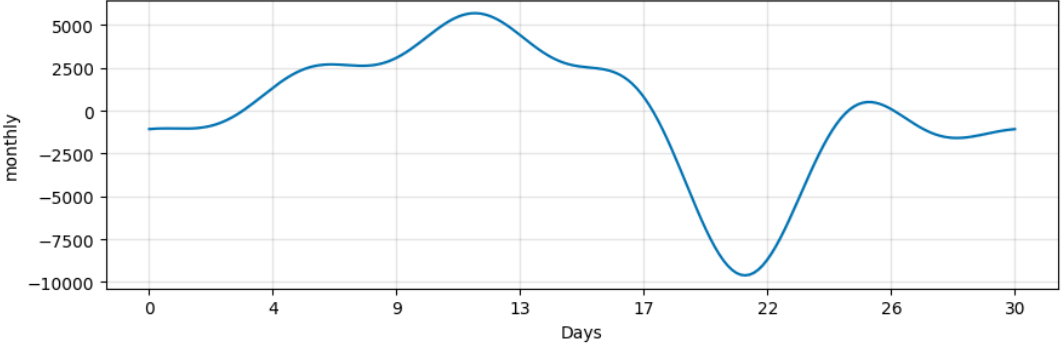
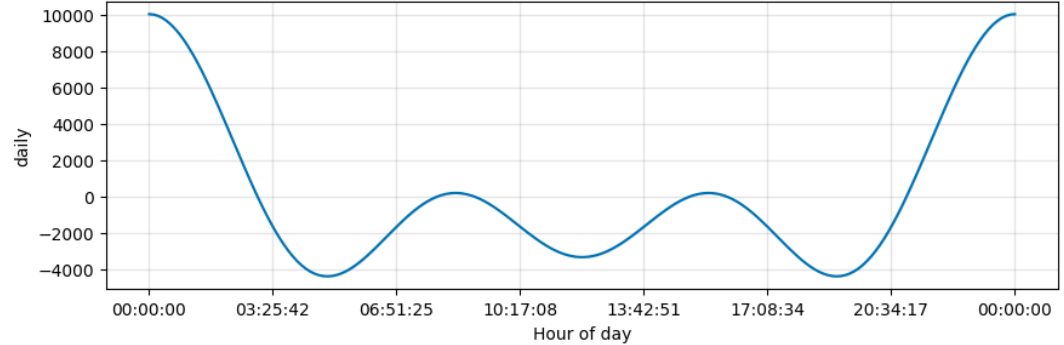
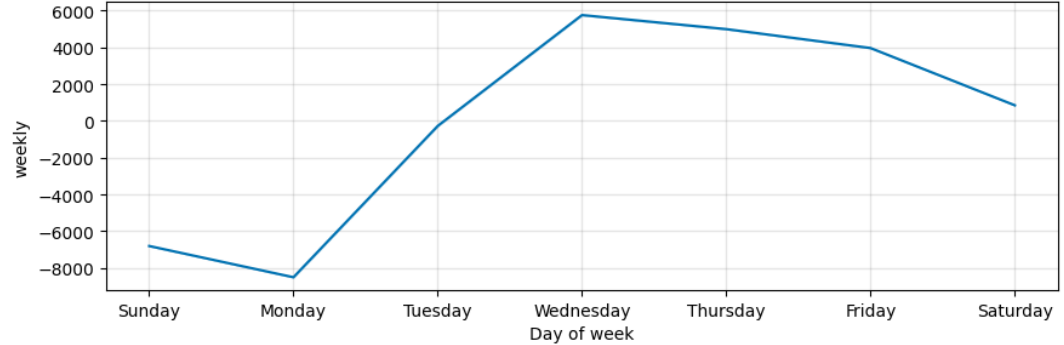
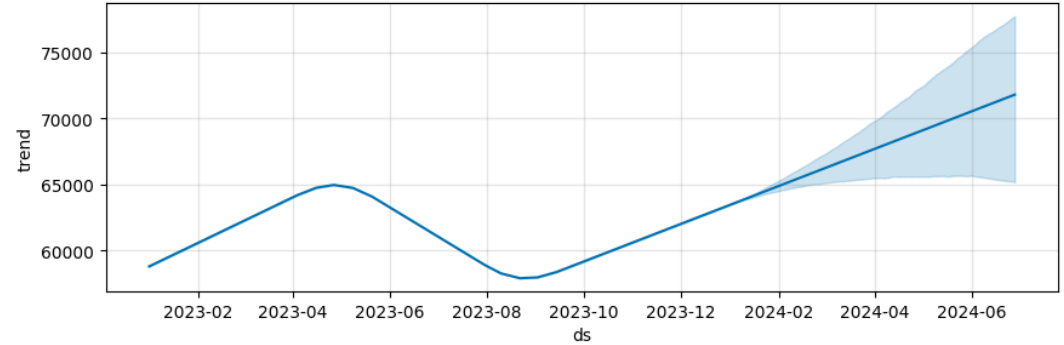
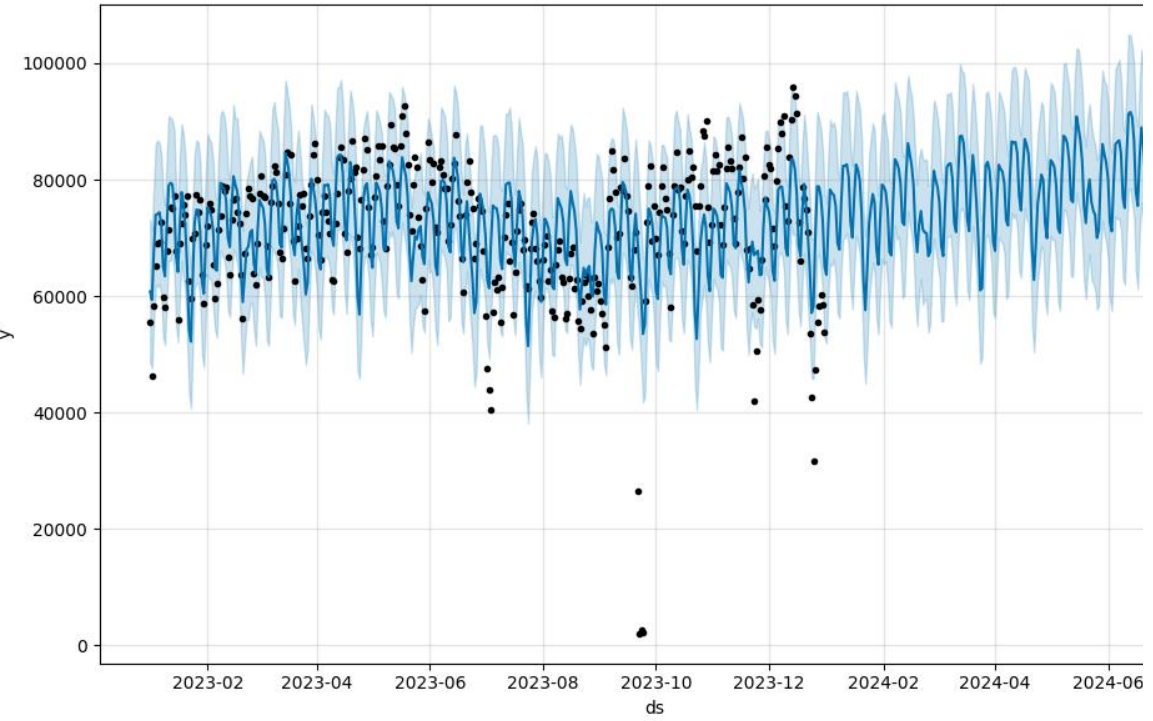


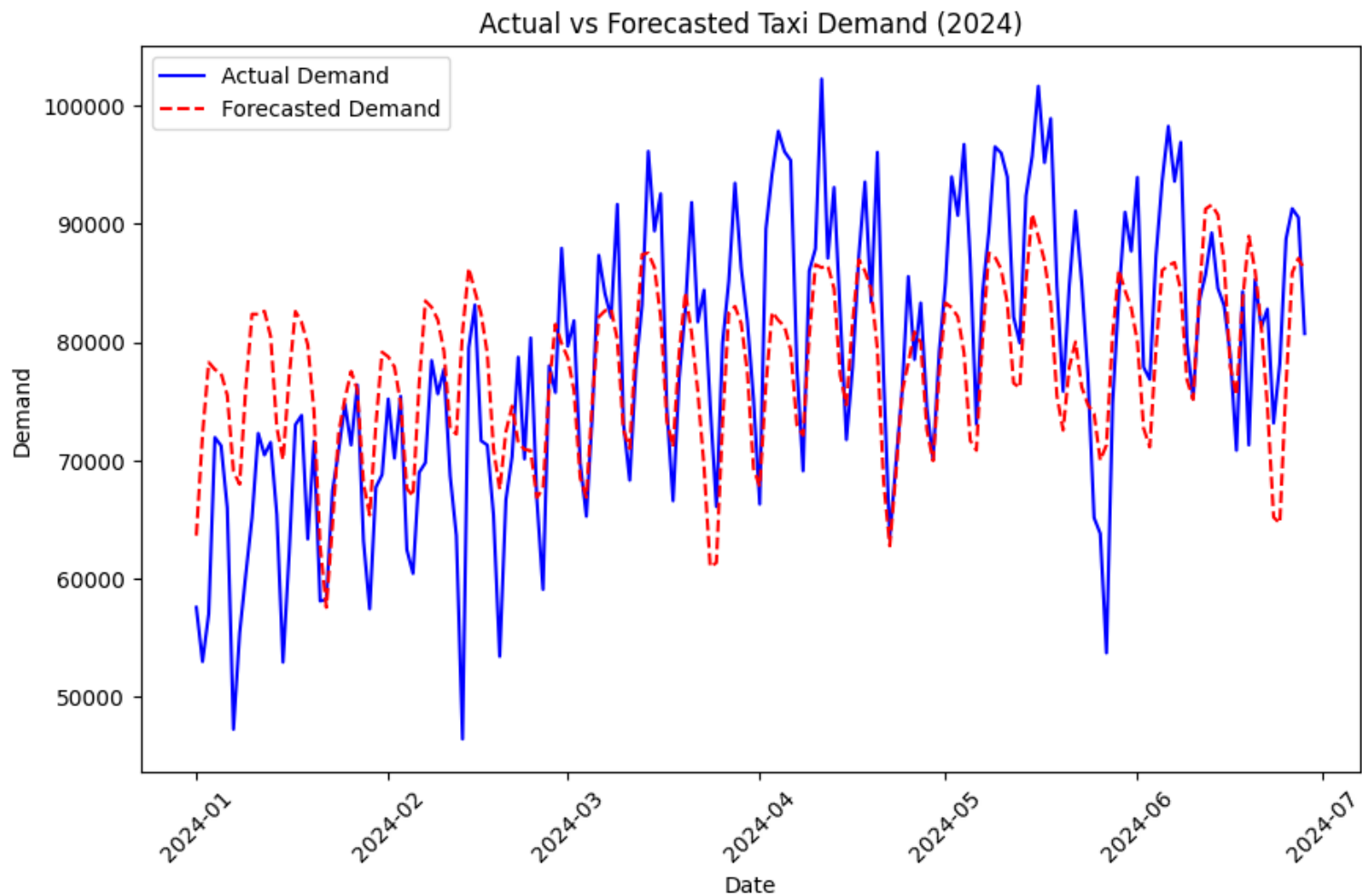
Actual vs Forecasted Taxi Fare Revenue (2024)

1e6



Taxi Demand





MAE: 6892.943946538229
RMSE: 8770.956216919876
MAPE: 0.09277875136923078%

Actual vs Forecasted Taxi Demand

Conclusion

- To conclude, we effectively predicted fare amount, payment types and also future revenue and demand.
- Predictions help enhance efficiency and customer experience.
- Results validate the potential of regression and classification models in transportation data analysis.
- Future work can include exploring more advanced models and deep learning for higher accuracy.
- Also, Develop real-time prediction models to handle dynamic changes in demand and improve decision-making.