NYC Taxi Fare Prediction Analysis

Joe Mendoza

Phase 5, 2025

Business Understanding

- For this project we are going to develop machine learning models to predict Taxi fares.

- Accurate fare predictions to optimize strategies and improved customer service

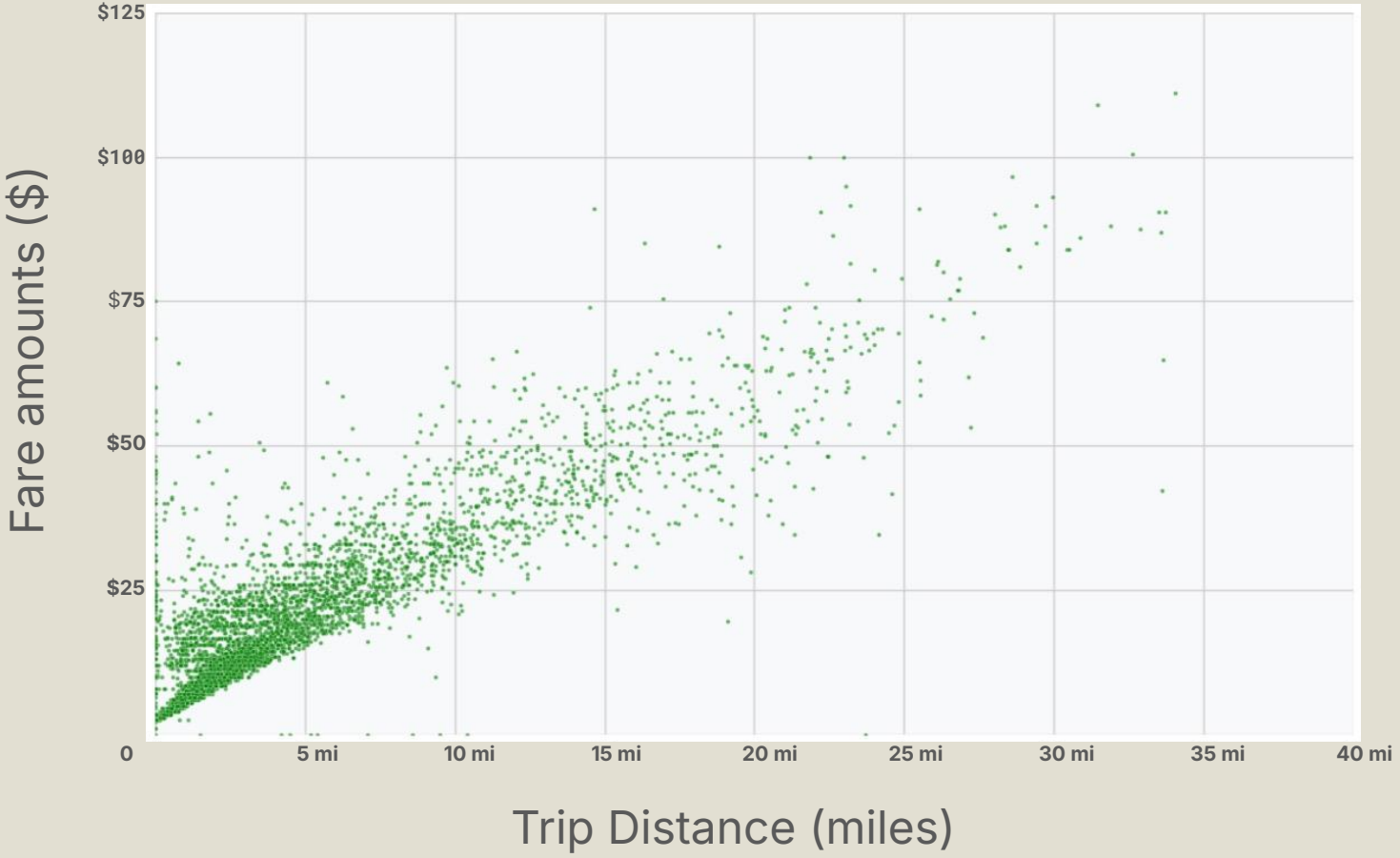
Data Understanding

- We are using the “Taxi Trip Data NYC” data set from Kyle
- Data set as over 83,000 rows and 20 columns
- For this data set spans from 12/14/20 - 08/02/21
- Dataset it will be cleaned and set up for analysis

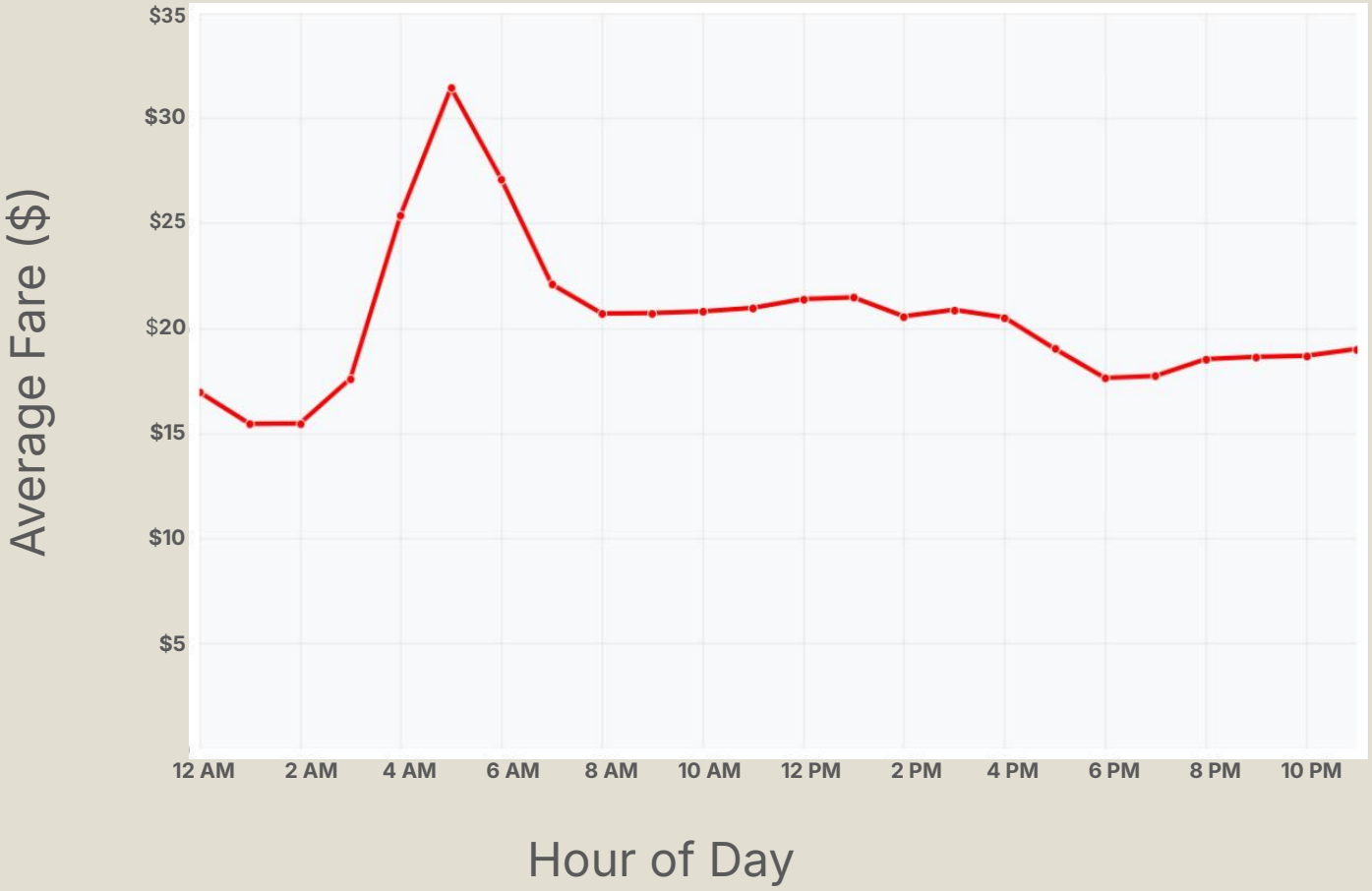
Exploratory Data Analysis

- Creating graphs showing four different analysis.

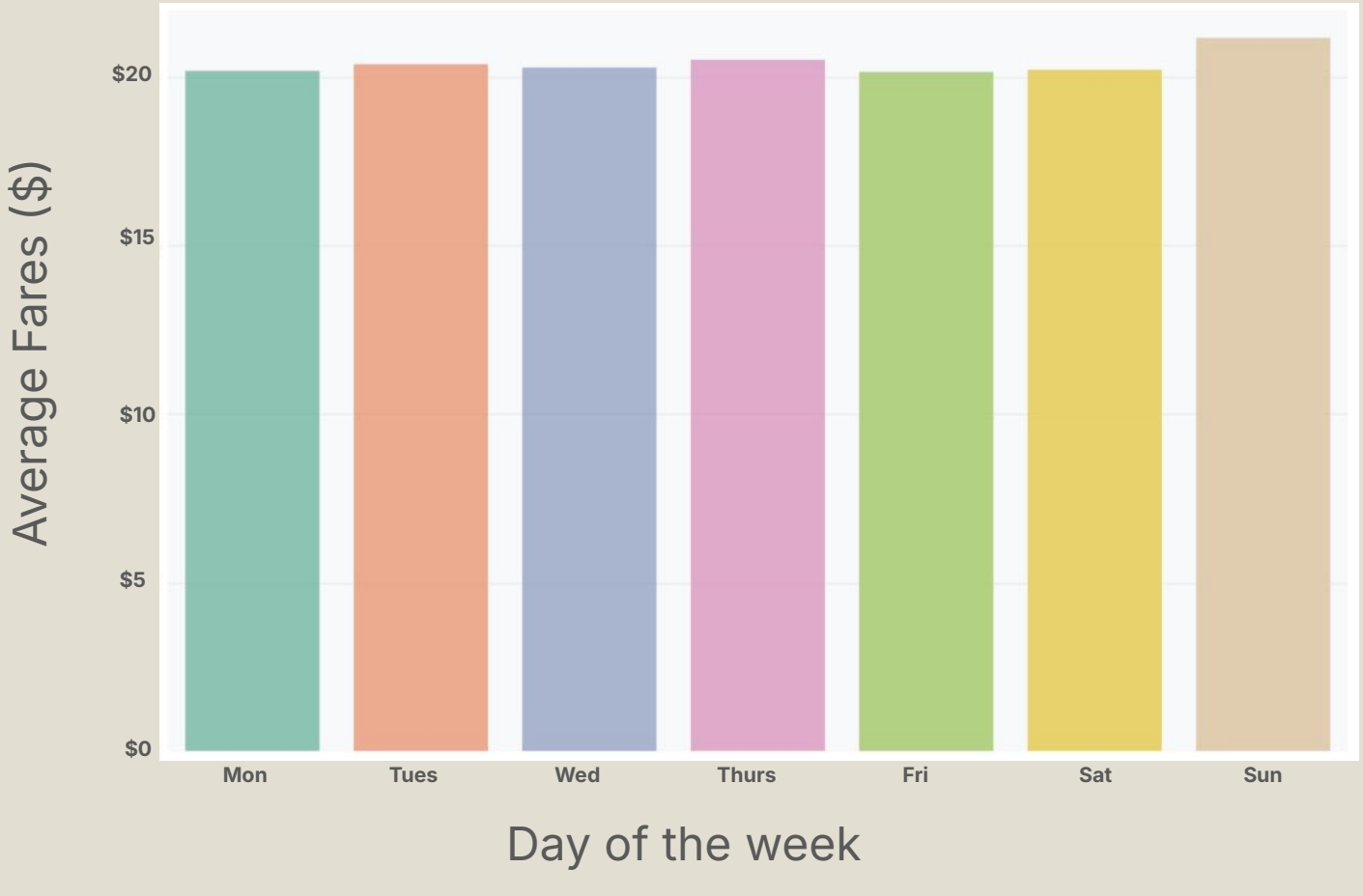
Fare vs Trip Distance



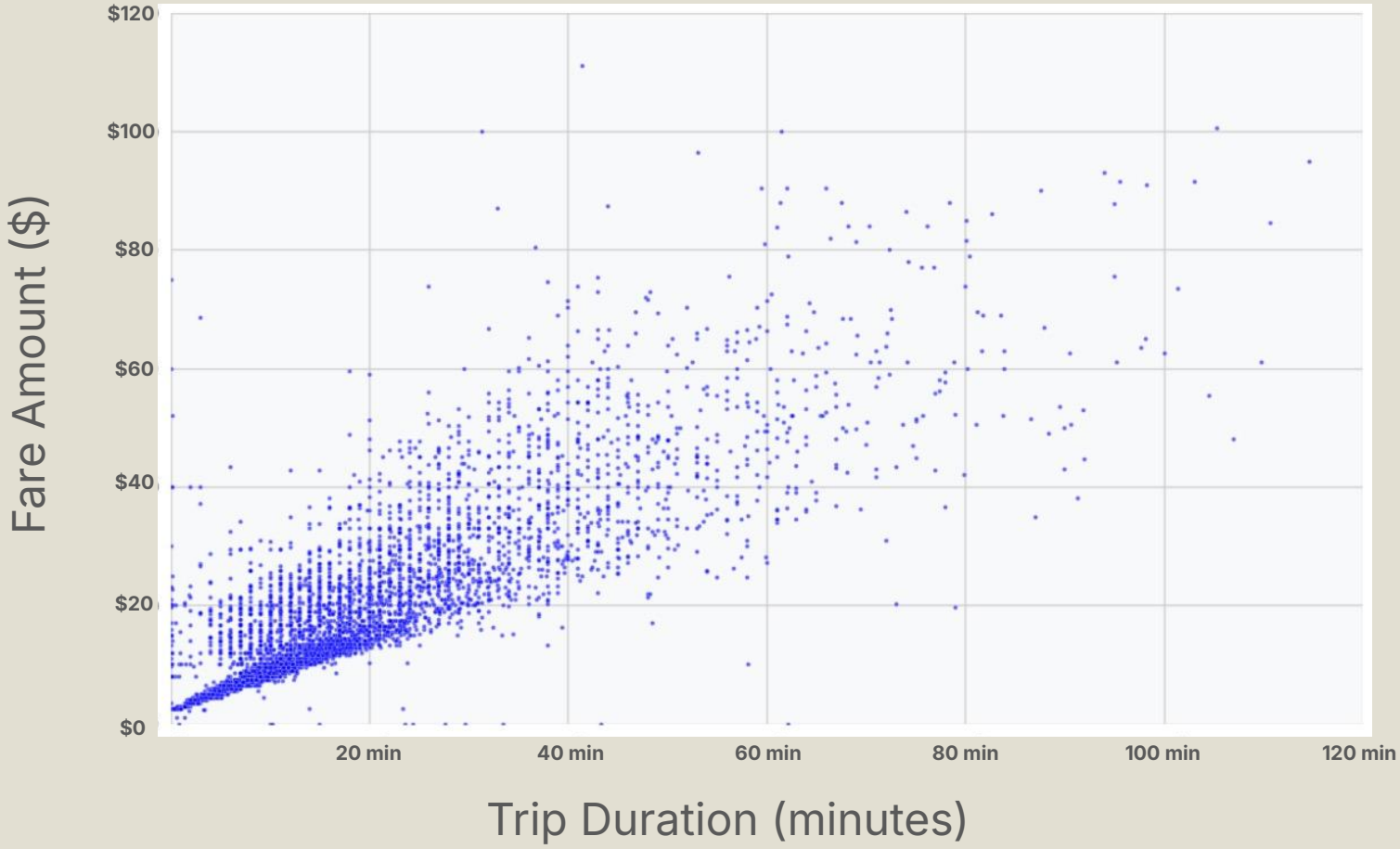
Fare by Hour of Day



Fare by Day of the Week



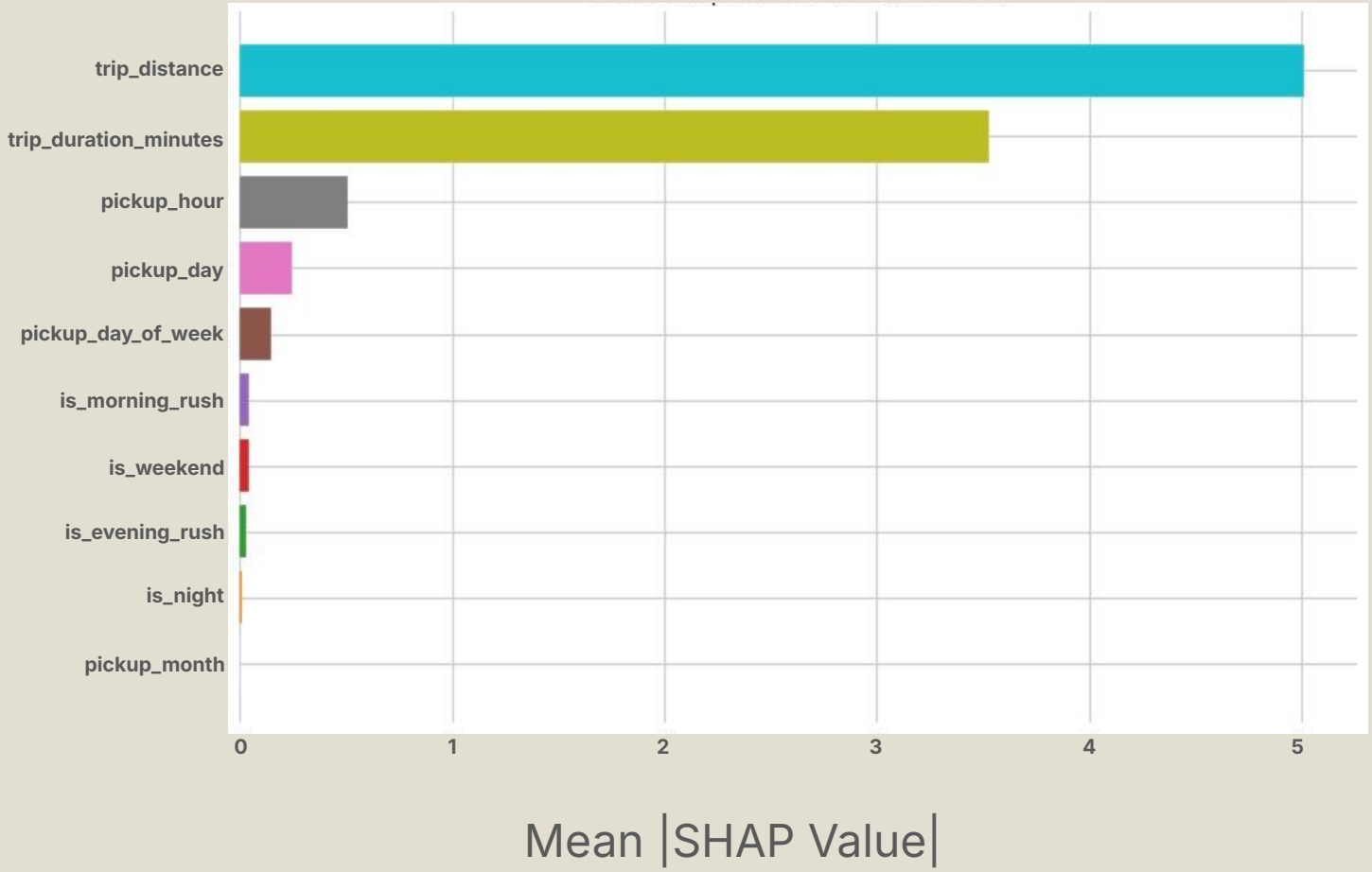
Trip Duration vs Fare



Machine Learning Analysis

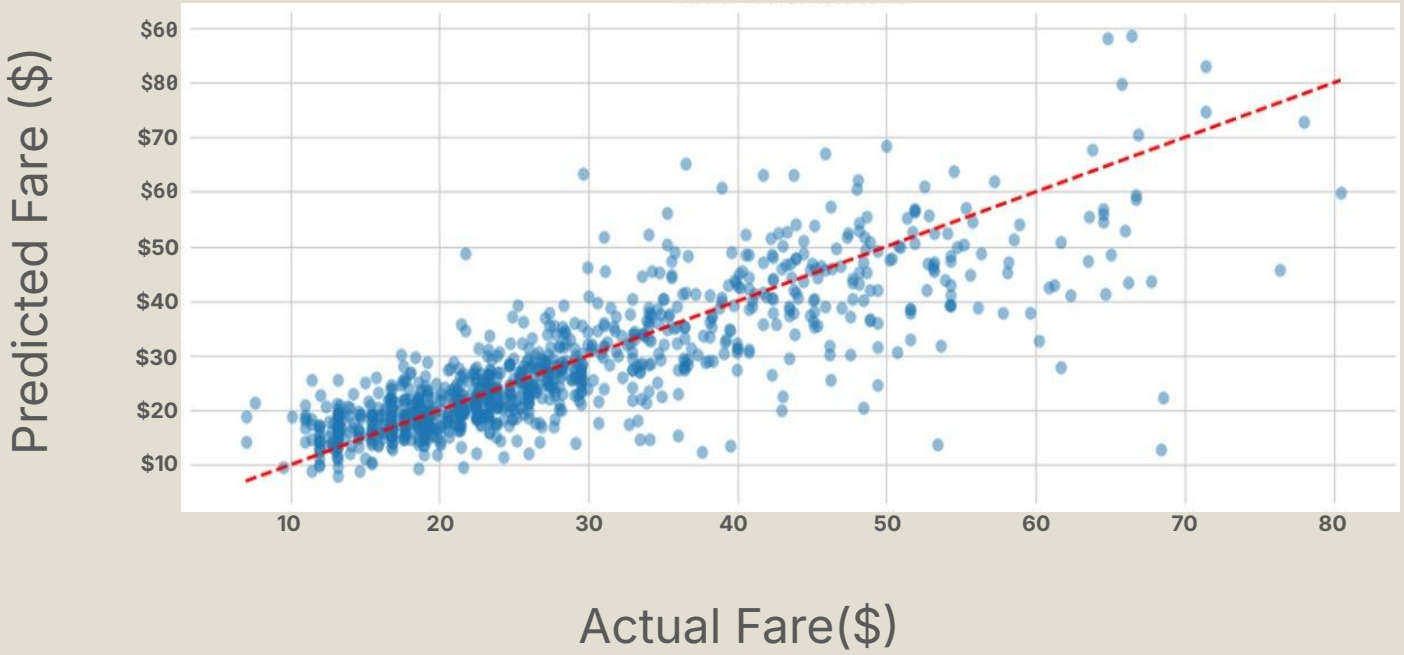
- Train all models
- Evaluate all models performance

Feature Importance-Random Forest



Prediction Analysis-Random Forest

Actual vs Predicted Fares



Conclusions

- Predicted average fair is $27.76
- Weekend trips were predicted at $24.81
- Short trips were predicted at $21.14
- Long trips were predicted at $40.85
- Fare column and the key features were distance, duration, and time patterns.
- Random Forest was our best performing model
- On average predictions were off by $ 5
- Pandemic consideration

Limitations

- Data set is only from 12/14/2020 to 08/01/2021
- New York City market only
- No seasonal variations
- Geographic location limitations
- Traffic patterns not captured in dataset
- Airport service charges not captured either

Recommendations

- Integrate NYC TLC zone data
- Develop real time data pipeline for life predictions
- Develop multi city prediction capabilities
- Advance analytics to implement recommendations

Next Steps

- Implement continuous model monitoring
- Automated retraining pipeline
- Fallback models
- Data validation practices

Thanks!

Email: joe_mendoza@icloud.com

Github: @dbvimpec