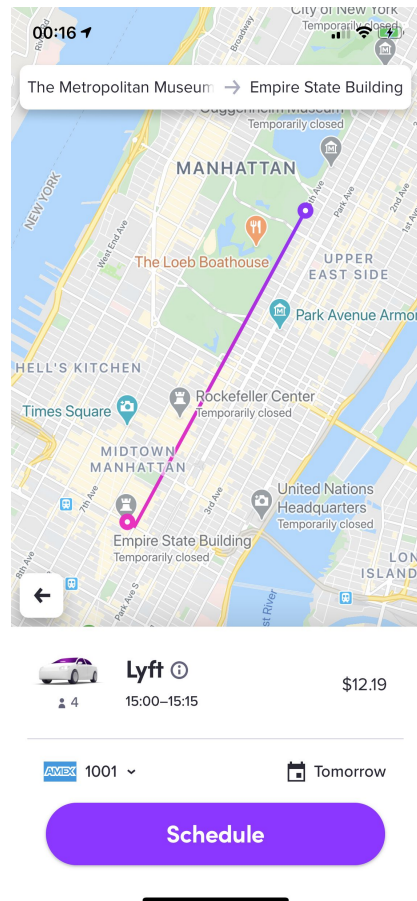
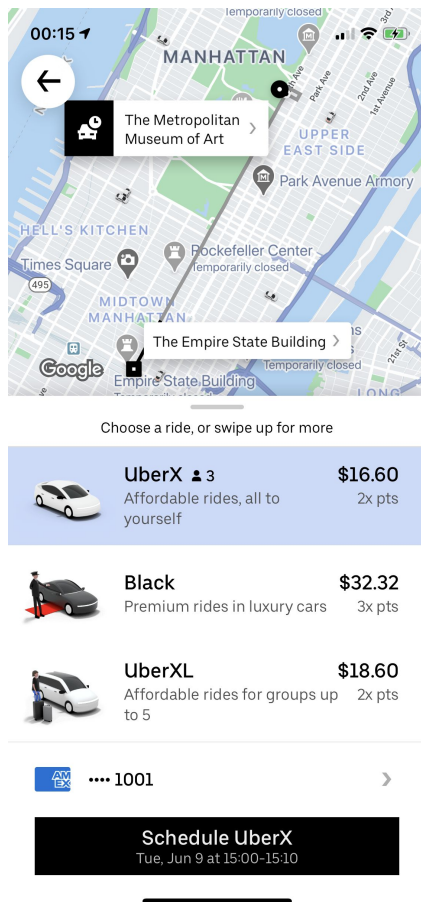


NYC Taxi Price Estimator

Know before you go. No math, no surprises.

Motivation & Problem

- Problem:
 - We usually do not know cost for a taxi ride ahead of time
 - Unable to compare taxi fare with price estimates provided by ride-hailing apps to plan ahead
- Scenario:
 - My friend and I are visiting NYC, and we plan to visit The Metropolitan Museum of Art in the morning and go to Empire State Building at 3pm tomorrow. Should we take a taxi, Lyft, or Uber?



Data

- Source:
 - Kaggle Competition: [New York City Taxi Fare Prediction](#)
- Dataset:
 - train.csv: input features and target for taxi trips in NYC from 2009 to 2015 (~ 55M rows)
- Features:
 - pickup_datetime - timestamp value indicating when the taxi ride started
 - pickup_longitude - pickup longitude coordinate
 - pickup_latitude - pickup latitude coordinate
 - dropoff_longitude - dropoff longitude coordinate
 - dropoff_latitude - dropoff latitude coordinate
 - passenger_count - the number of passengers in the taxi ride
- Target:
 - fare_amount - dollar amount of the cost of the taxi ride
- Subset Data:
 - Filtered data by a specific year (configurable)
 - Default: the most recent year - 2015
- Feature Engineering:
 - Extracted from pickup_datetime
 - Day of Week
 - Hour
 - Dropped pickup_datetime
- Train Test Split:
 - Performed stratified sampling on hour and day of the week for training and test sets
 - Number of total observations and test set ratio are configurable
 - Default # of observations: 50,000
 - Default test set ratio: 0.3

Model & Performance

- Model: Random Forests

- Reasons:
 - Convenient feature importance for inference and easy to interpret
 - Robust to outliers in predictor space
 - Computationally scalable: build trees in parallel
 - Automatically discards irrelevant predictors
- Used default parameters

- Baseline Performance

- Basic estimate based on just the distance between the pickup and dropoff points
- RMSE of \$5-\$8 depending on models

- Performance on Test Set:

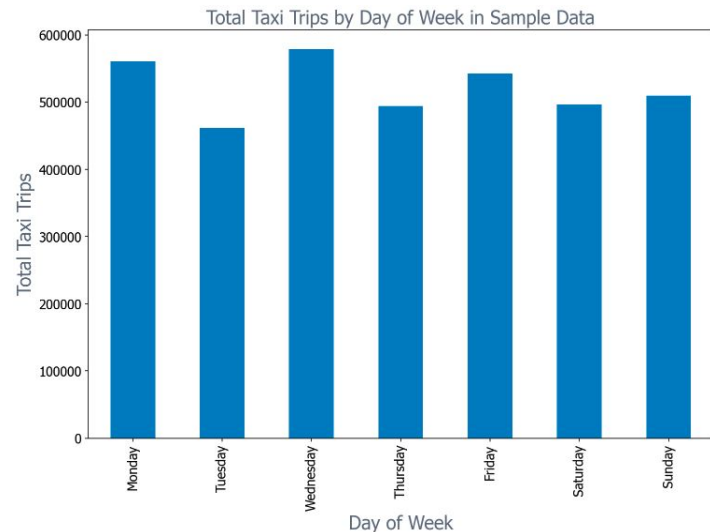
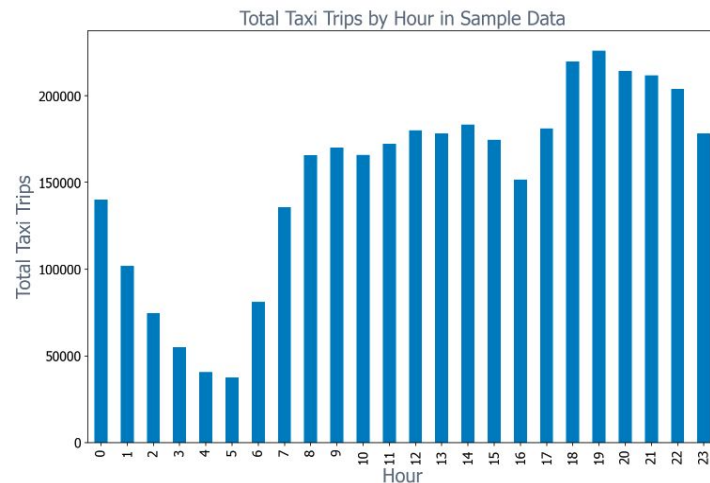
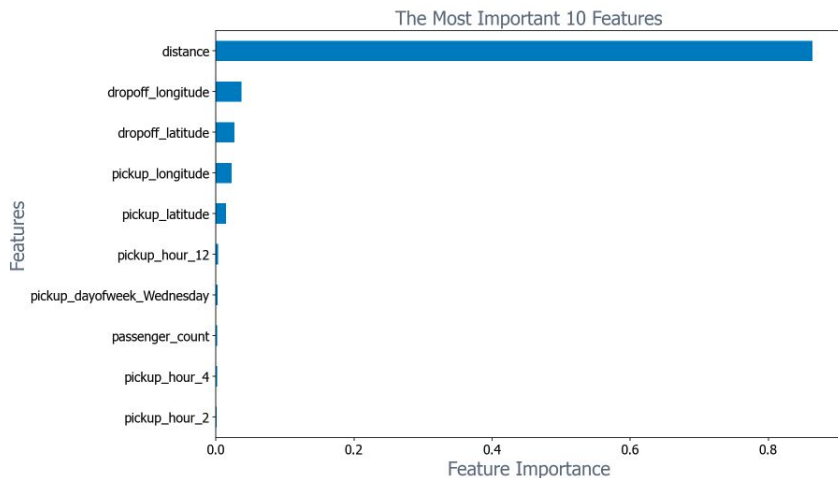
- **RMSE: \$3.852**
- MAE: \$2.033
- R-squared: 0.894

- Success Criteria in Project Charter:

- Achieved ML performance metric criterion
 - RMSE has to be less than \$4 for the app to go live
- Unable to evaluate business success metric:
 - the app used to estimate 100 trips per day

Insights

- Distance and locations are the most important features
- Pickup hours seem to be more important than day of week, probably because the difference in traffic volume by hour > difference in traffic volume by day of week
- Findings agree with taxi fare rules: Taxi fare = \$2.50 initial charge. Plus 50 cents per 1/5 mile when traveling above 12mph or per 60 seconds in slow traffic or when the vehicle is stopped



Thank You!

Lirong Ma
www.github.com/lirongm
