Uber / Lyft Price Prediction

By Anthony Schams & Phoebe Wong

Data Description

- Kaggle data on rideshare services Uber and Lyft, combined with weather data
- Gathered from end of November 2018 - Mid December 2018
- Collected by making API calls to Uber, Lyft, and a weather API
- Limited to Boston Area
- 635,242 observations after cleaning

Feature Engineering

- One hot encoded categorical features: origins/destinations (neighborhoods), and types of ride (UberX, Uber Black, Lyft, etc.)
- Make additional time-related features with epoch time: day of the week, AM/PM, rush hours (7-9am; 5-9pm), weekends nights(PM on Fri, Sat, Sun). Also added feature that identify days with sporting events (Patriots, Celtics, Bruins).
- Log transformed all numerical variables (price, distance, temperature, humidity, wind)
- Total 55 features

Linear regression for predicting price

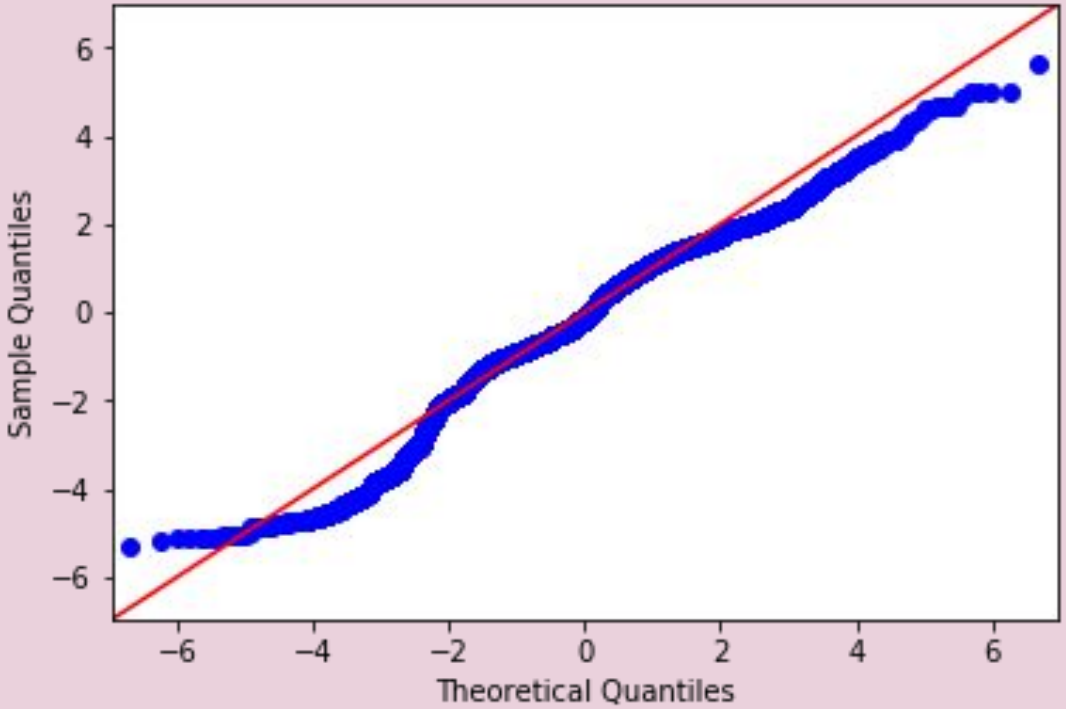
Preliminary exploration:

- OLS Model including all features had Adjusted R-squared of 0.933
- But QQ plot shows non-normal residuals

Model selection using LassoCV & RidgeCV in sklearn:

- 80/20 train test split
- Tried to address possible issues: bimodal distributions; outliers; but all have non-normal residuals.
- Best model w normal residuals includes features chosen with random forest.
    - Adjusted R-Sq: 0.666 ; 9 features selected using Lasso

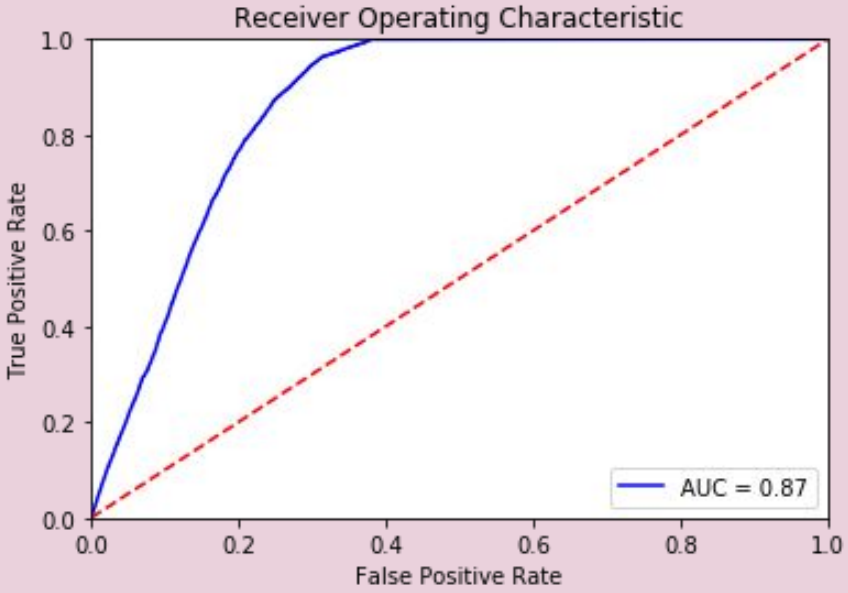
Linear Regression Plots

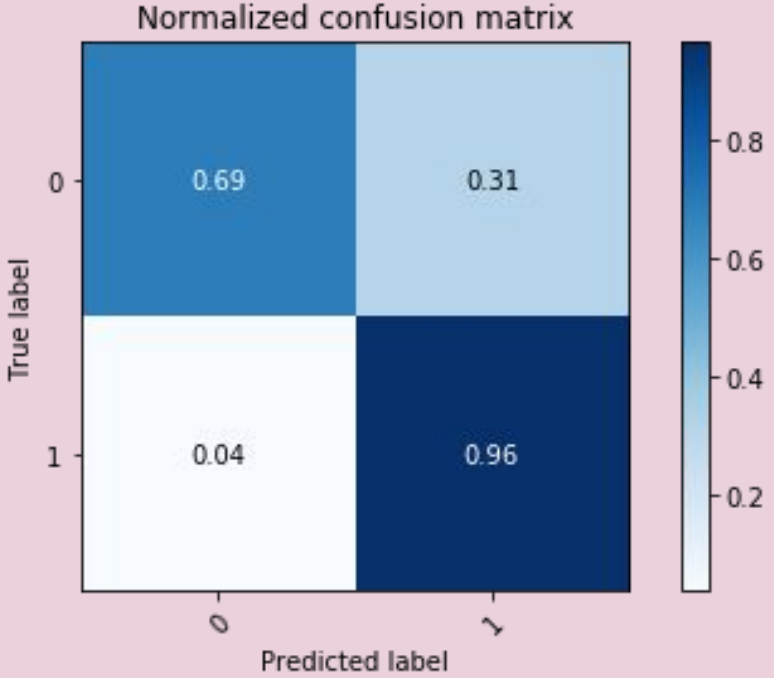


Logistic regression for predicting surge multiplier

- Encoded surge multiplier to 1 when multiplier >1; 0 otherwise
- Use logistic regression cross validation (LogisticRegressionCV) in sklearn
- Class imbalance: ~ 3% positive
- Resampling to address class imbalance:
  - Subsample for speed
  - Same 80/20 train test split
  - Separated to two dataframes: target = 0 (n = 491,486) & 1 (n = 16,707)
  - Random sampling 16.7k obs from target = 0 dataframe, append to target = 1, for 10 times and average the scores
  - L1 and L2 regularization yielded very similar results with this method

Logistic Regression Plots





Comparing Price Over Time

Uber consistently cheaper than Lyft

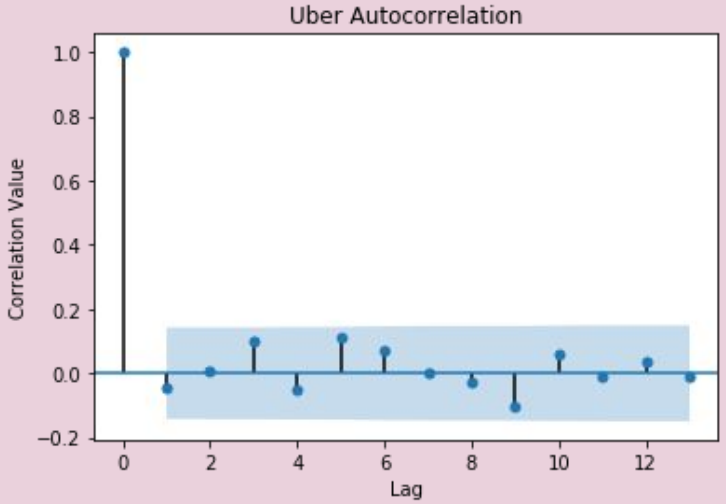
Is there Time Series Trend between the two companies?

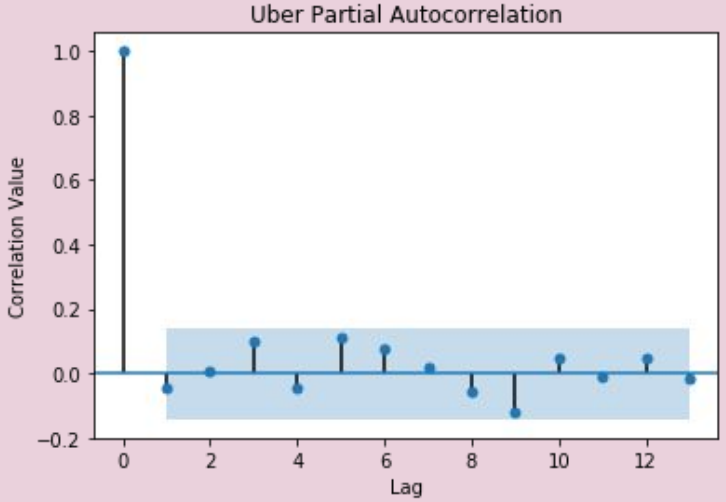
Uber Prices Time Series

No Trend

No Autocorrelation

No Partial Autocorrelation





Lyft Prices Time-Series

No Trend

No Autocorrelation

No Partial Autocorrelation