Predicting the Price of an Uber/Lyft ride

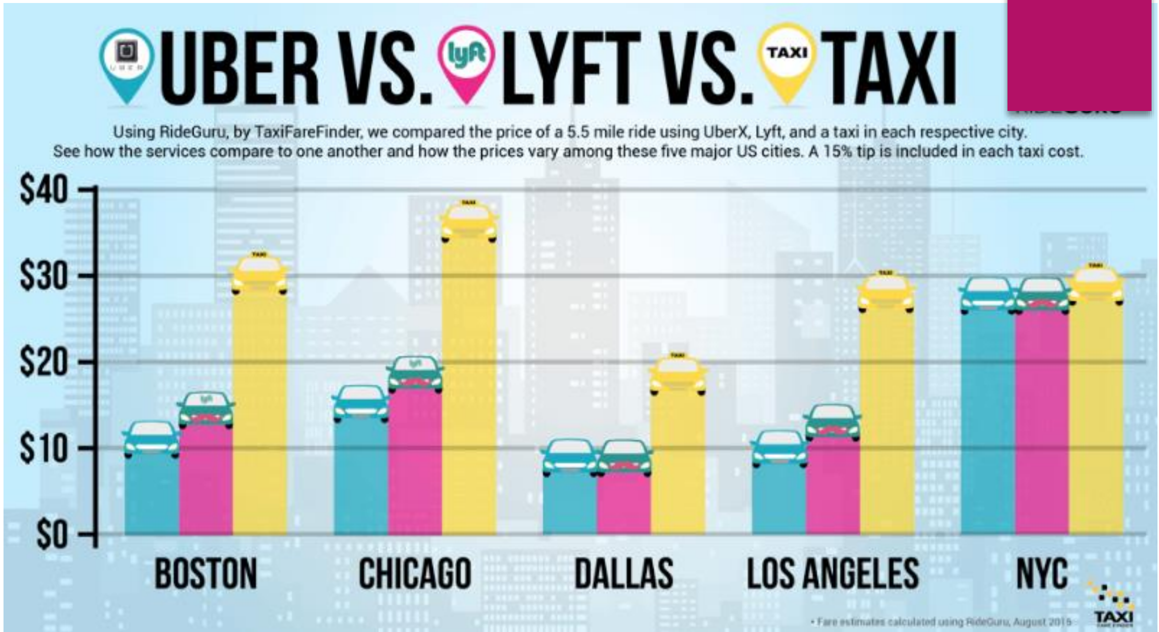
BY DEVESH GOKALGANDHI

THANKS TO SPRINGBOARD MENTOR: NADAV RINDLER

The Problem

- Uber and Lyft dominate the ride-service industry, with over 100 million users across both applications.
- Rides, offered by these applications, are a much cheaper alternative to standard yellow taxis in major cities.

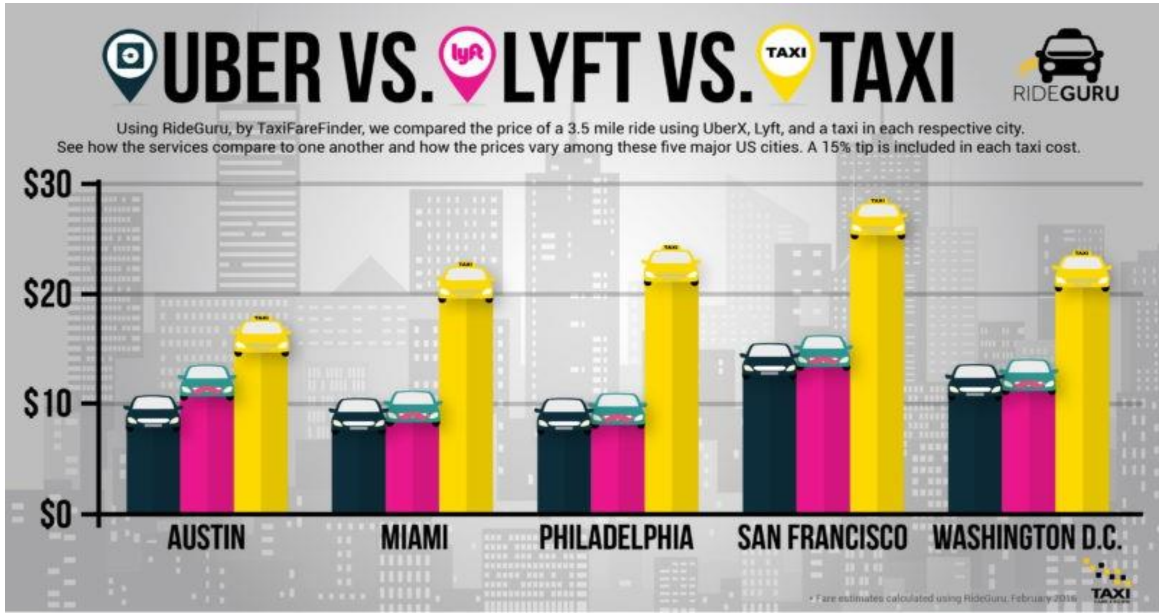
What factors influence the price of Uber/Lyft rides?



UBER VS. LYFT VS. TAXI

Using RideGuru, by TaxiFareFinder, we compared the price of a 5.5 mile ride using UberX, Lyft, and a taxi in each respective city.
See how the services compare to one another and how the prices vary among these five major US cities. A 15% tip is included in each taxi cost.

• Fare estimates calculated using RideGuru, August 2015



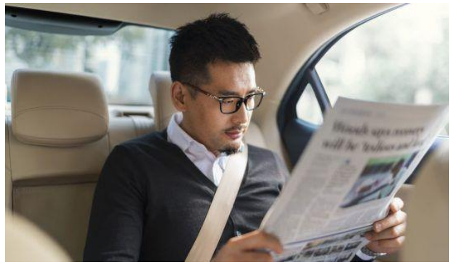
UBER VS. LYFT VS. TAXI

Using RideGuru, by TaxiFareFinder, we compared the price of a 3.5 mile ride using UberX, Lyft, and a taxi in each respective city.
See how the services compare to one another and how the prices vary among these five major US cities. A 15% tip is included in each taxi cost.

• Fare estimates calculated using RideGuru, February 2016

Who might care?

Riders



Competing Ride-Service Applications

Gett®

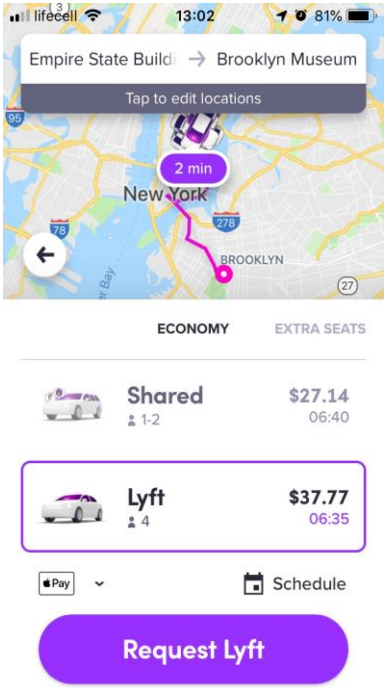
curb

zTrip®

BRIDJ

What factors might influence price?

- Ride Factors
  - Distance
  - Surge-Multiplier
  - Ride-Type
  - Pickup/Dropoff Location
  - Pickup Time of Day
  - Pickup Day of Week



lifecell 13:02 81%

Empire State Build → Brooklyn Museum

Tap to edit locations

ECONOMY EXTRA SEATS

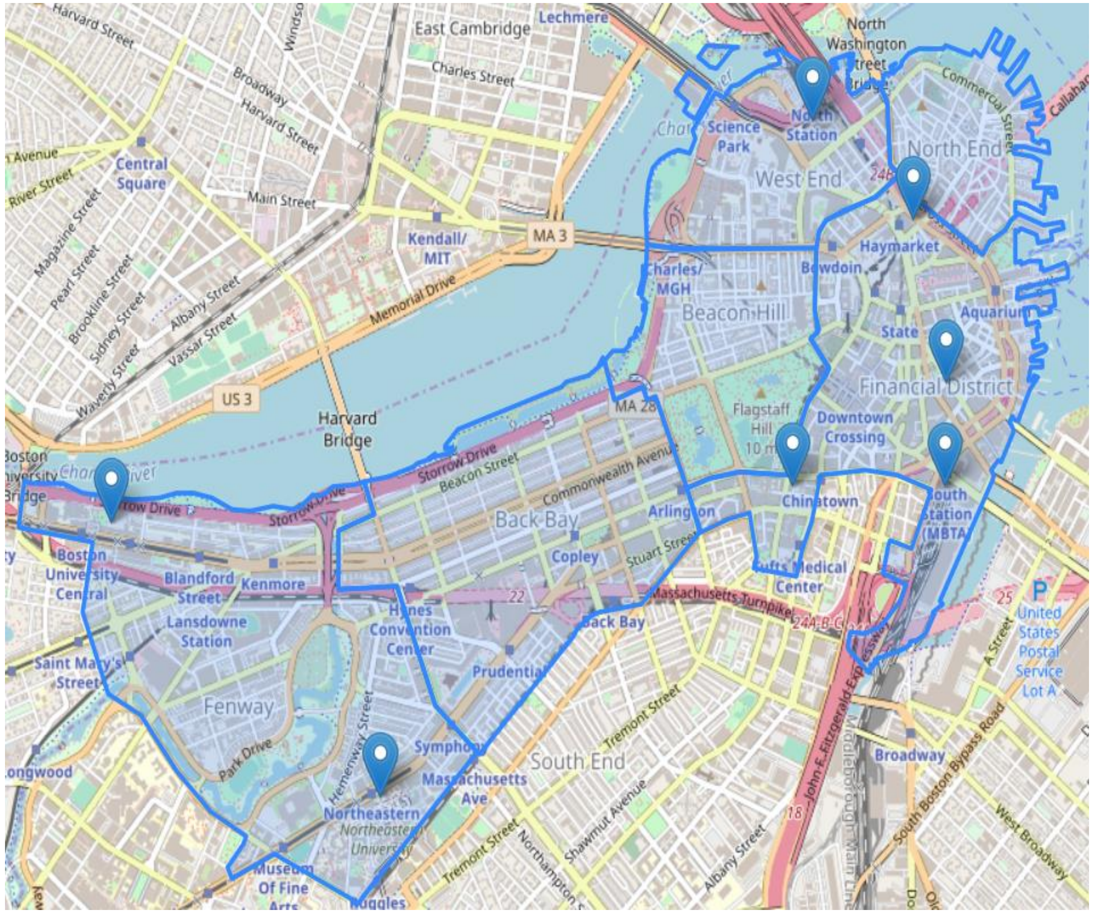
Shared 1-2 $27.14 06:40

Lyft 4 $37.77 06:35

- Weather Factors
  - Temperature
  - Rain
  - Humidity
  - Pressure
  - Wind
  - Clouds

Data Information

- Ride pickup and dropoff occurring within 6 major Boston neighborhoods: North End, West End, Back Bay, Beacon Hill, Fenway, and Downtown.
- Within these areas: Boston University and Northeastern University (in Fenway), North Station (in North End), Haymarket Square, Financial District, South Station, and Theatre District (in Downtown)



Data Information

Ride Data Specifics

- Acquired through real time API queries
- Ranges from Nov 26 – Dec 18, 2018
- File Format: csv
- Over 600000 ride instances (approx. 50000 for each of12 ride-types). Each row is a unique ride instance from either Uber or Lyft

Weather Data Specifics

- Acquired through wunderground.com
- Ranges from Nov 26 – Dec 18, 2018
- File Format: csv
- Each weather instance describes the measurement of each weather feature (temp, rain, wind, etc.)at specific time

Merging

Weather instance merged by time-stamp, occurring within 1 hour of pickup, on ride pickup location

Sport Occurrence

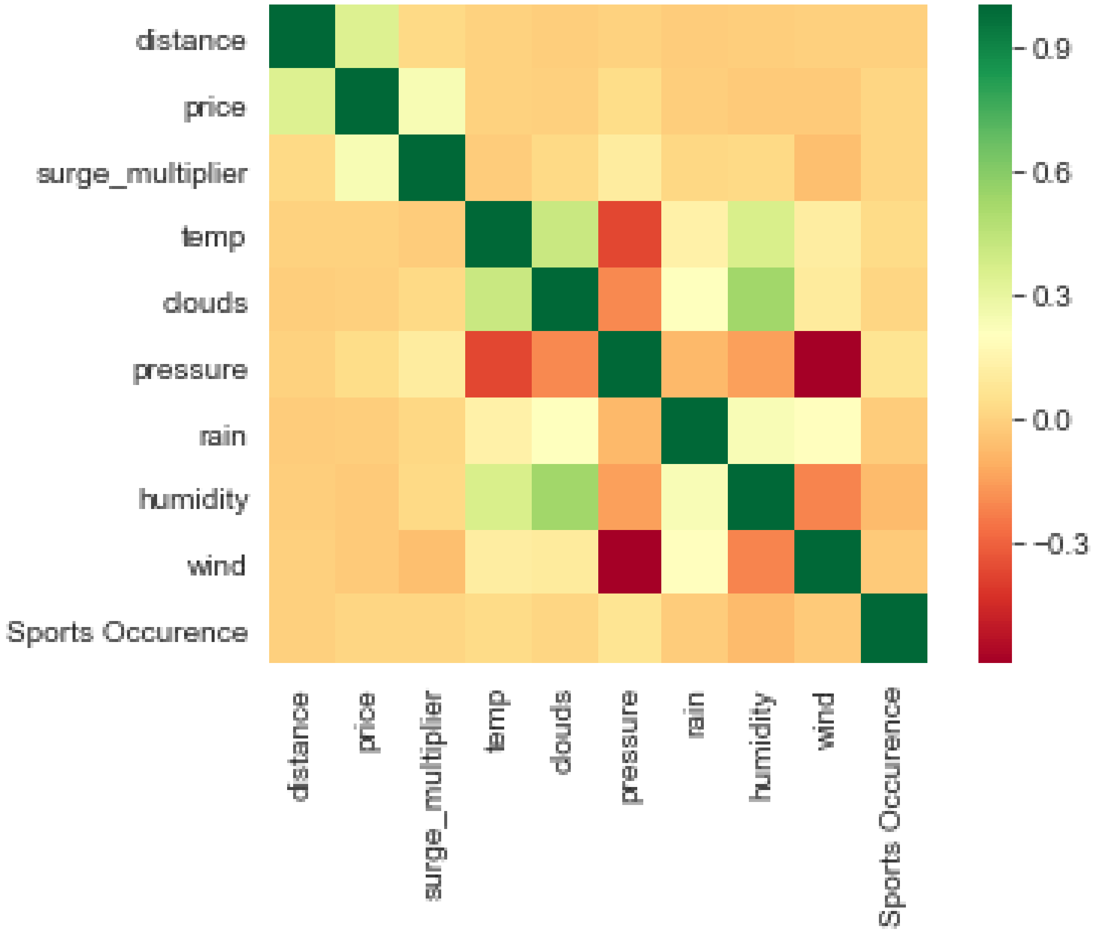
- Dummy variable column added to dataset representing the occurrence of a Celtics game or Bruins game.
- 1 if game occurred within 2 hours of ride pickup or ride drop-off, 0 otherwise.
- Only for rides going to and coming from North Station (where TD Garden is located).

Data Exploration

HTTPS://GITHUB.COM/DGOKALGA/SPRINGBOARD-DATA-SCIENCE/BLOB/MASTER/CAPSTONE-1/DATASTORYTELLING.IPYNB

1. VALUE PER DISTANCE
2. TEMPORAL FACTORS
3. SPORTS OCCURRENCE
4. WEATHER FACTORS

What quantitative features correlate with ride price?



Distance and surge-multiplier positively correlate with the ride price.

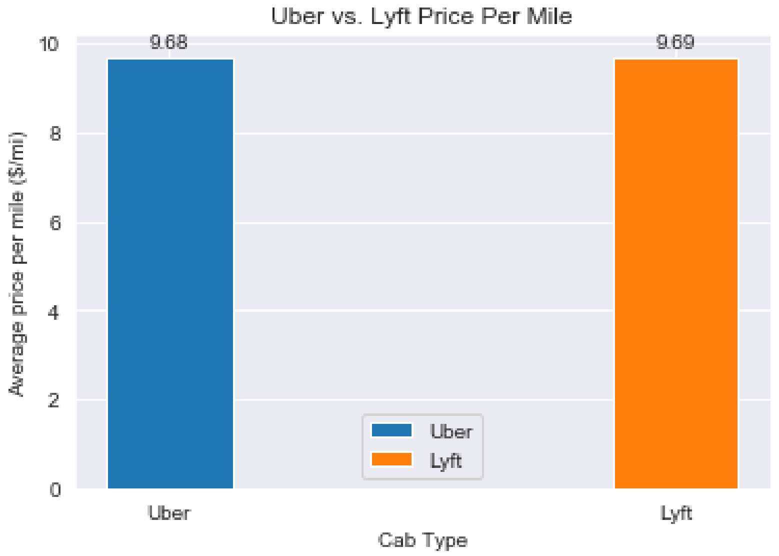
Some correlation between weather features.

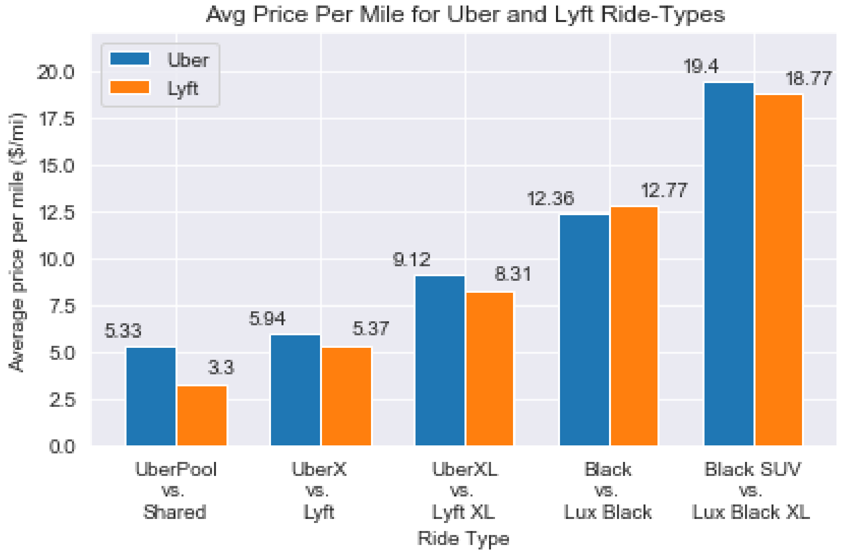
Value Per Distance

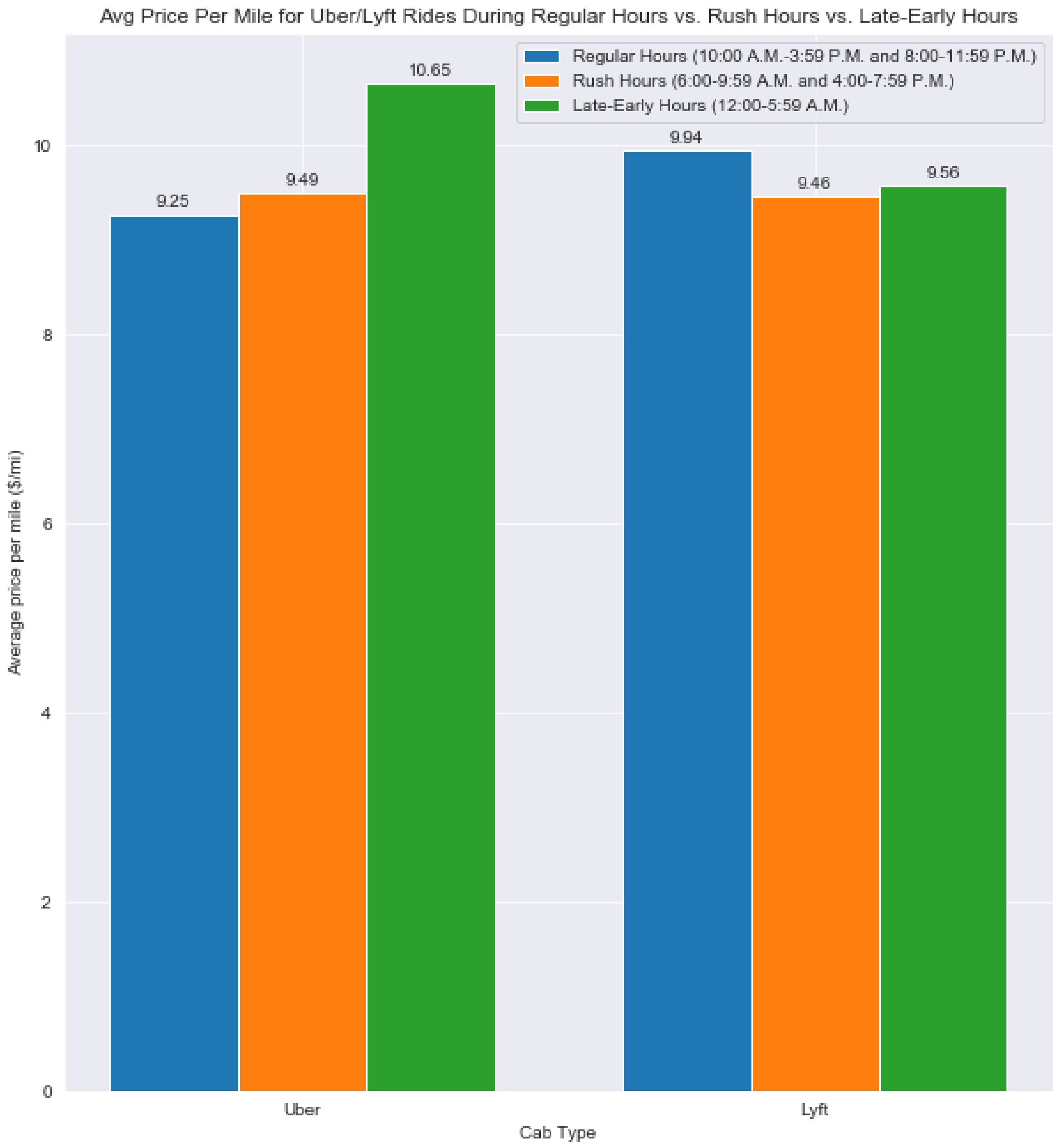
Looking at the value of a ride per distance for both Uber and Lyft ride-service applications, on average, both cost relatively the same per mile.

When comparing similar ride-types between Uber and Lyft, four of the five Uber ride-types: UberPool, UberX, UberXL, Black SUV are the more expensive option per mile, on average, compared to their Lyft counterpart.

Lux Black, on average, was found to be the more expensive option per mile than its Uber counterpart, Uber Black.



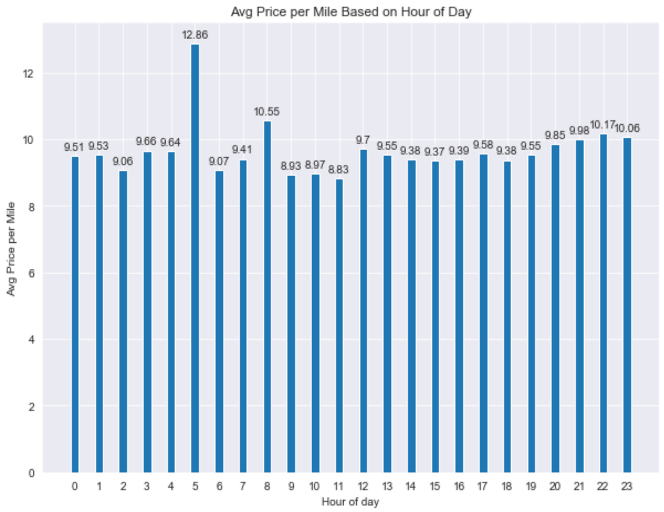


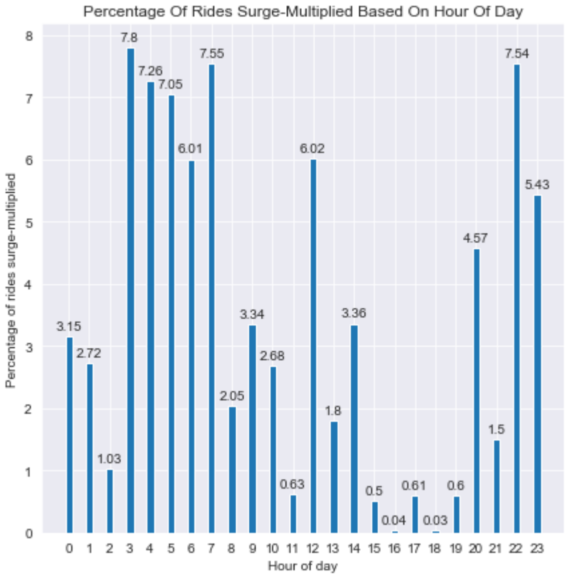


Prices Higher During Late-Early Hours for Uber and Regular Hours for Lyft

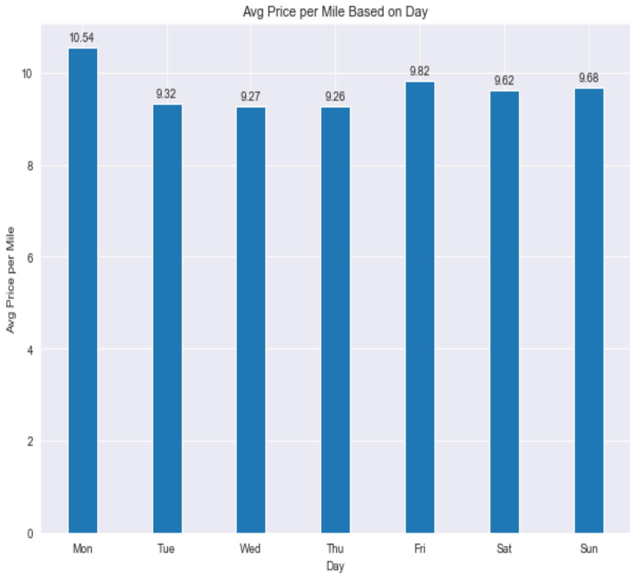
Prices Higher During Late-Early Hours, and Weekends (except Mon)

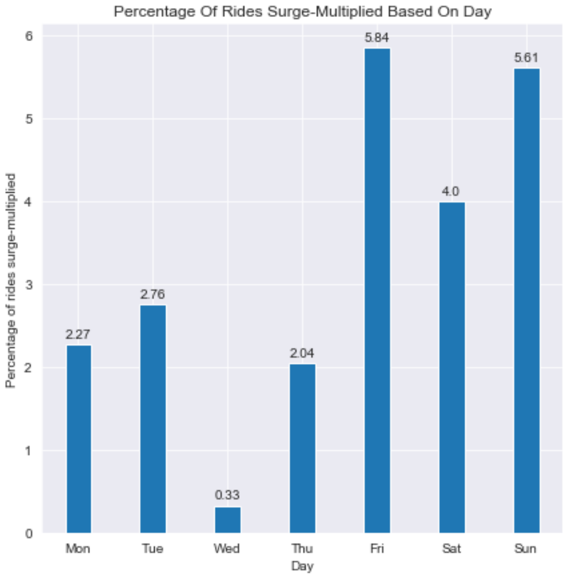
Hour of Day


Avg Price per Mile Based on Hour of Day


Percentage Of Rides Surge-Multiplied Based On Hour Of Day

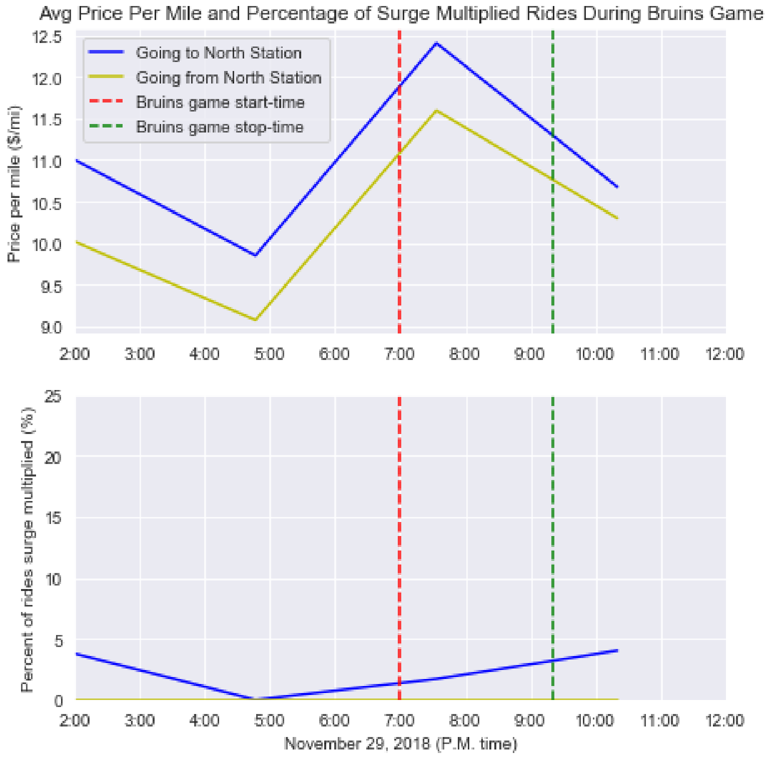
Day of Week


Avg Price per Mile Based on Day

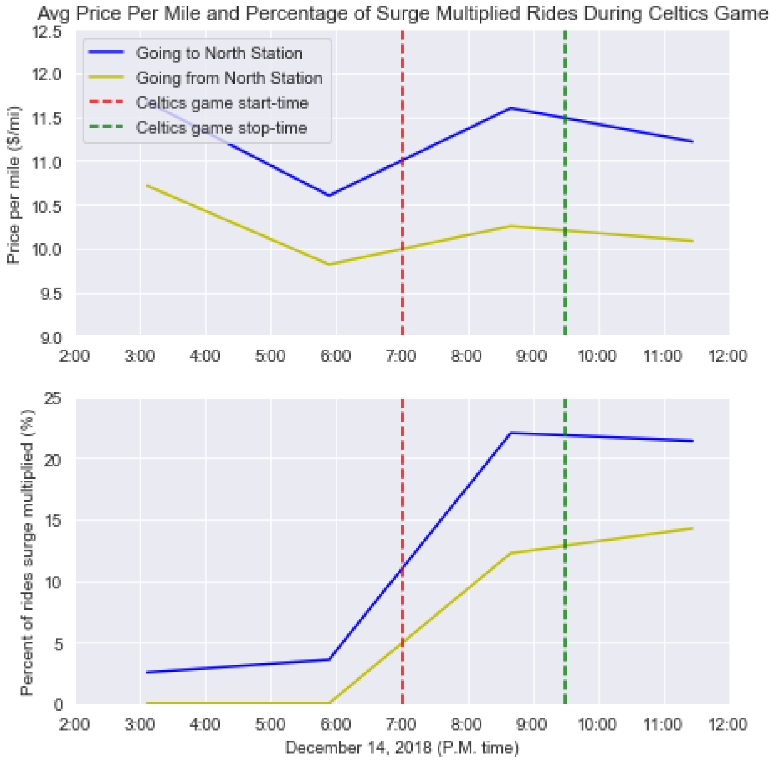

Percentage Of Rides Surge-Multiplied Based On Day

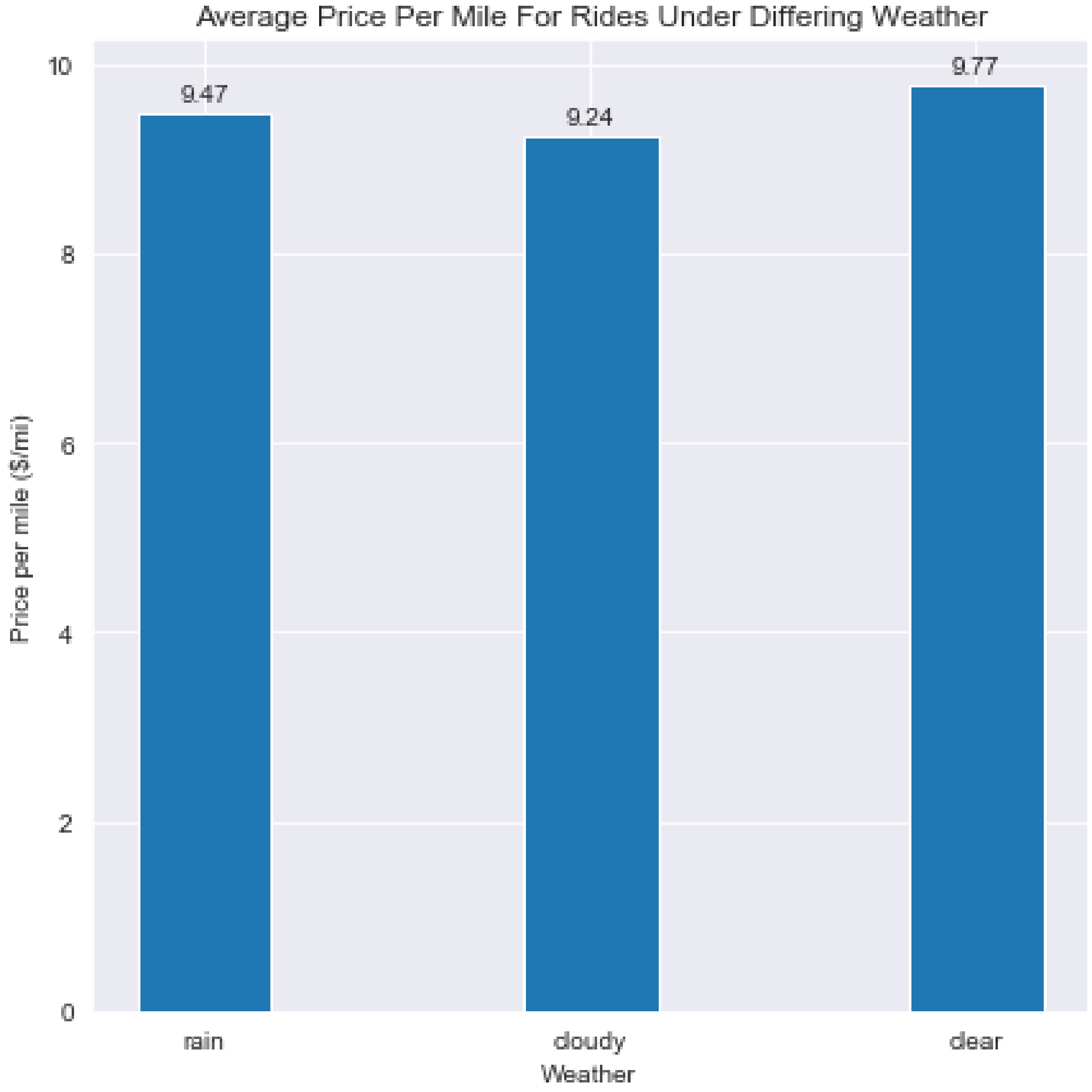
Prices Higher and Surged During and After the Game

Bruins Game


Avg Price Per Mile and Percentage of Surge Multiplied Rides During Bruins Game

Celtics Game


Avg Price Per Mile and Percentage of Surge Multiplied Rides During Celtics Game



Prices cheaper under rainy and cloudy weather

Regression Modeling

HTTPS://GITHUB.COM/DGOKALGA/SPRINGBOARD-DATA-SCIENCE/BLOB/MASTER/CAPSTONE-1/UBERLYFT_REGRESSION.IPYNB

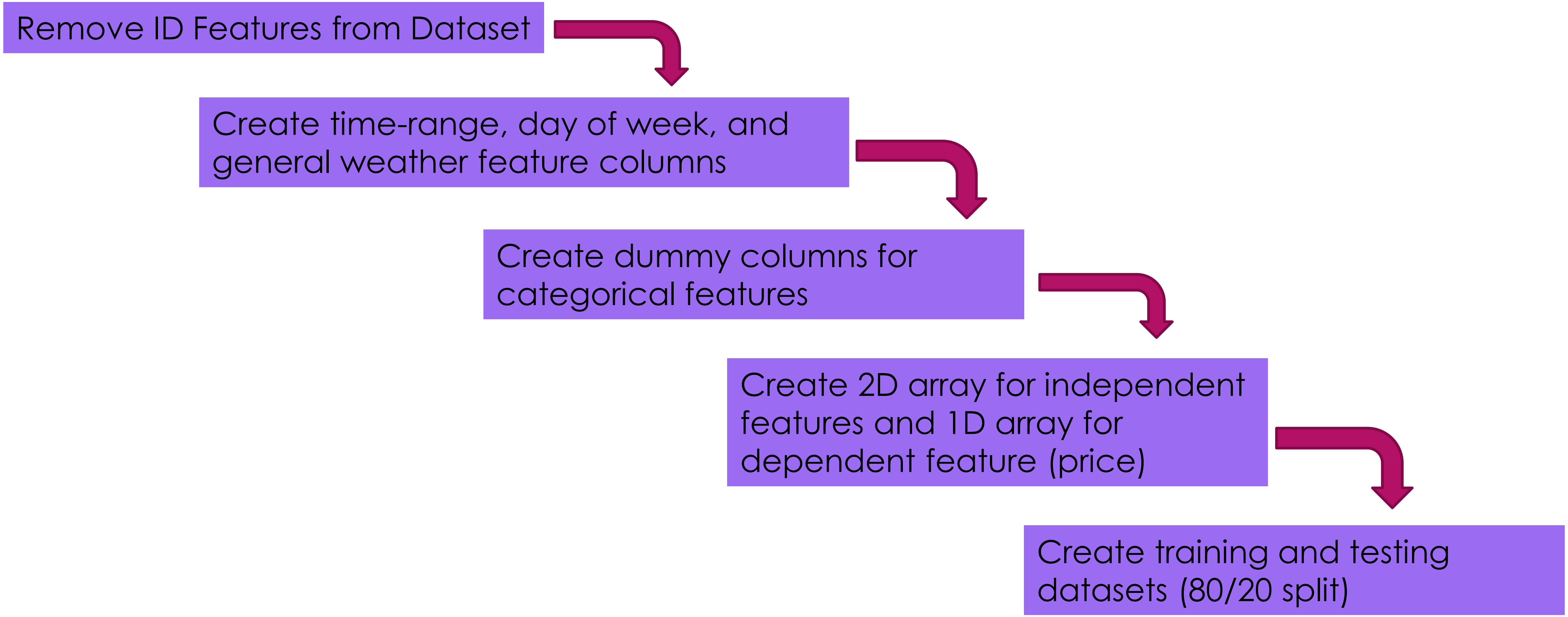
REGRESSION MODELS IMPLEMENTED (USING SCIKIT-LEARN AND STATSMODEL):

1. MULTIVARIATE LINEAR REGRESSION
2. LASSO REGRESSION
3. RIDGE REGRESSION

METRICS CALCULATED:

1. R-SQUARED
2. MEAN ABSOLUTE ERROR (MAE)
3. MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)
4. ROOT MEAN SQUARED ERROR (RMSE)

Modeling Pre-Processing



Remove ID Features from Dataset

Create time-range, day of week, and general weather feature columns

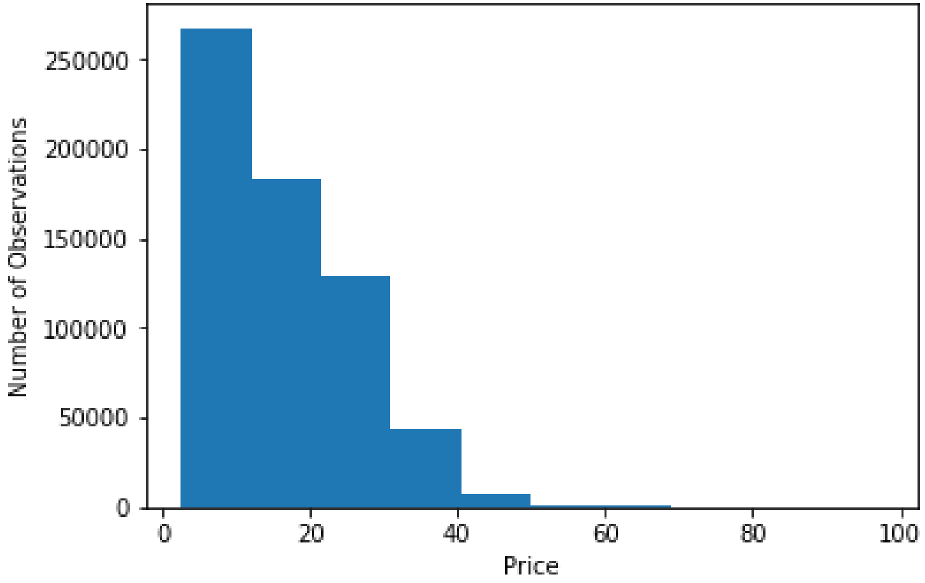
Create dummy columns for categorical features

Create 2D array for independent features and 1D array for dependent feature (price)

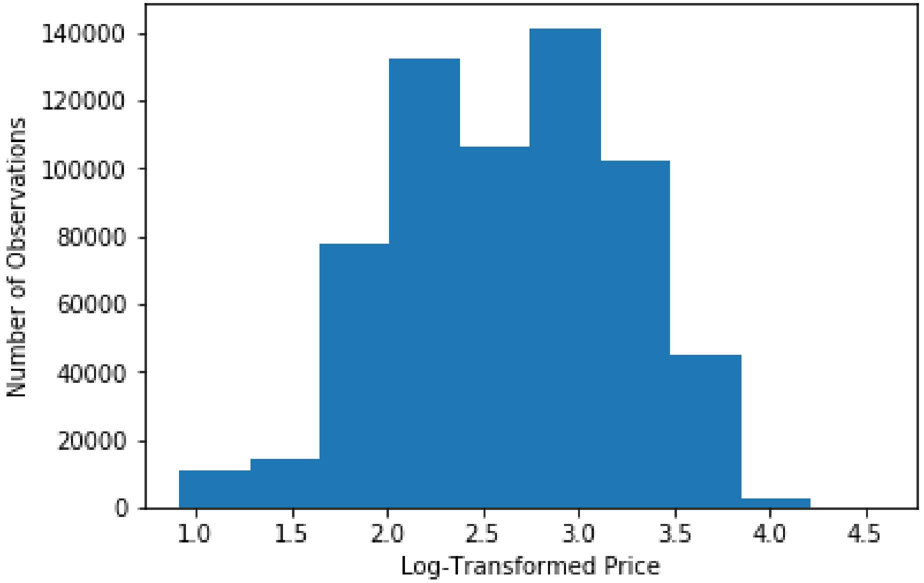
Create training and testing datasets (80/20 split)

Distribution of the Target Variable

- Distribution of Price

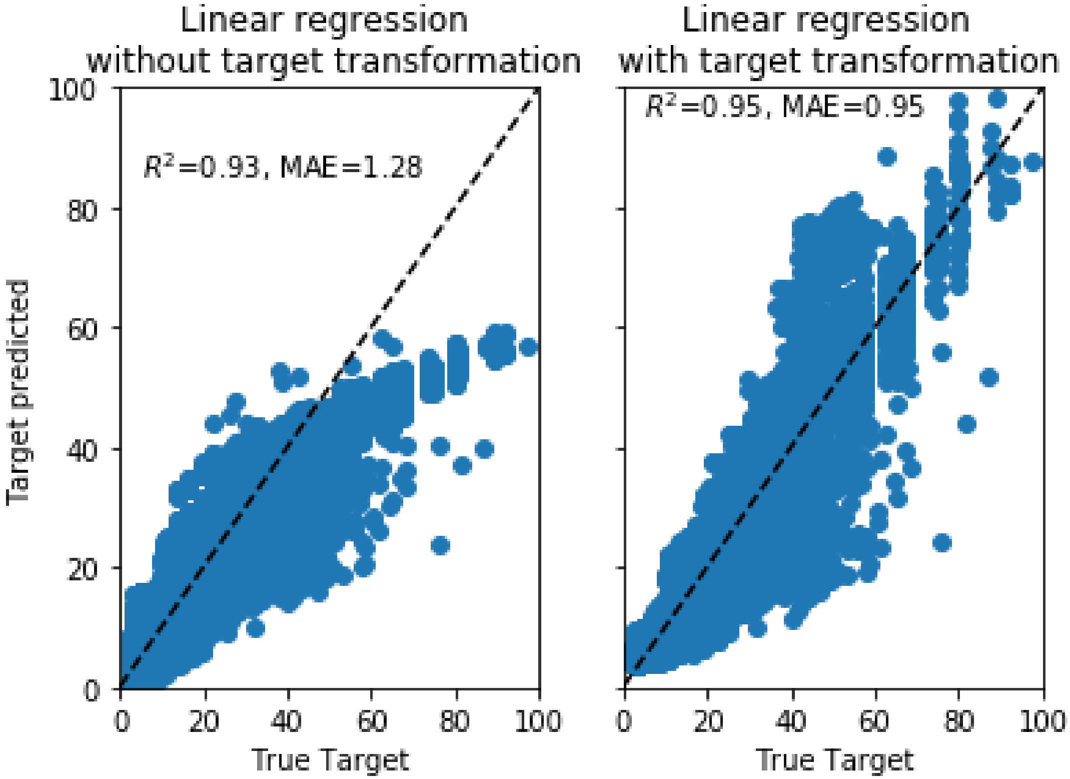


- Distribution of Log-Transformed Price



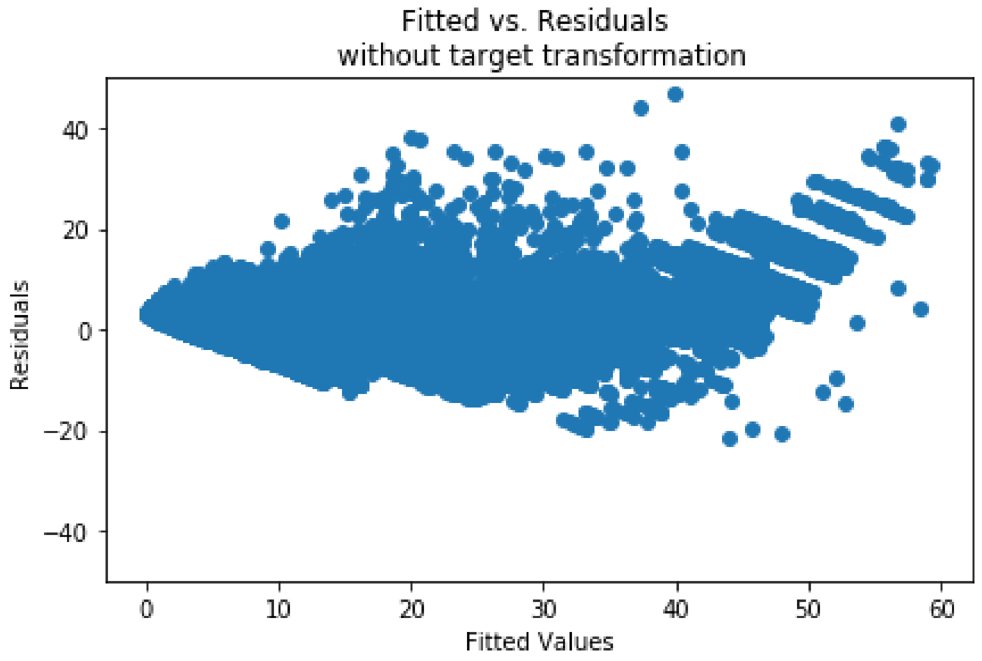
Multivariate Linear Regression

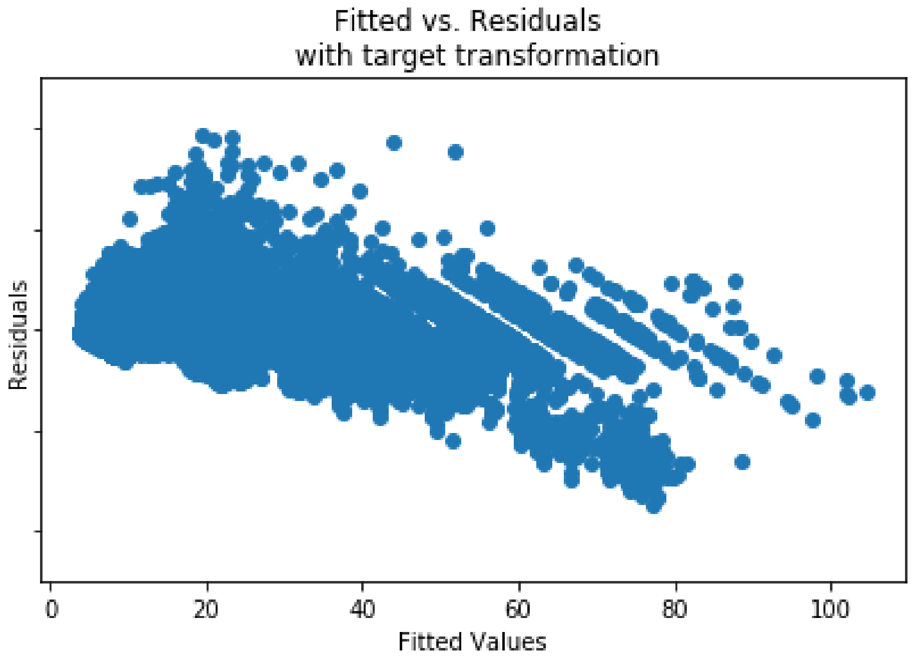
Regression performed on training data, for both non-transformed and transformed price



Multivariate Linear Regression (cont.)

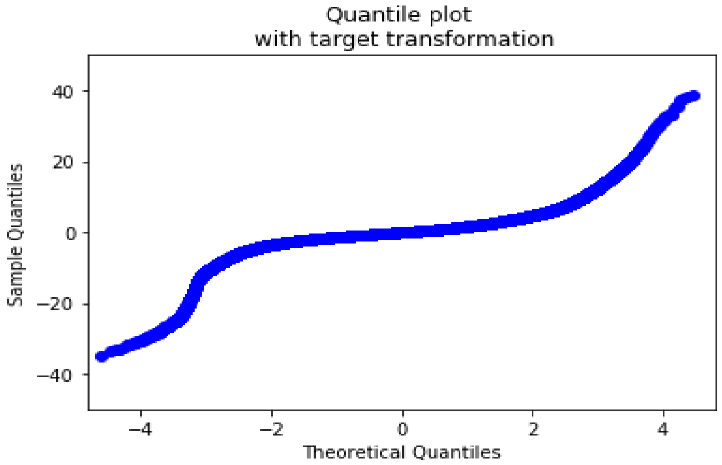
Transformed target variable model capture higher prices well



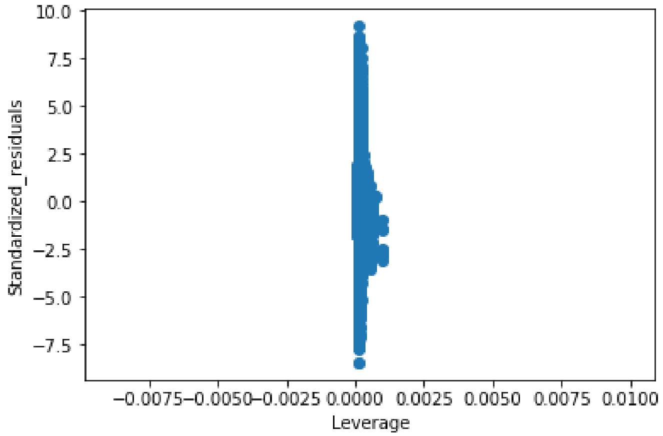


Multivariate Linear Regression (cont.)

Residuals are normally distributed with heavy tails



No evidence of highly influential points, as high leveraged points have small standardized residuals



Lasso and Ridge Regression

Lasso Regression

- GridSearch and Cross Validation to determine best alpha. Price is log-transformed
- Best alpha value of 0, same as linear regression
- How model evaluates test data with lowering alpha values:

alpha	R-Squared	MAE	MAPE	RMSE
.01	.845	2.44	15.83	3.68
.001	.944	1.44	9.85	2.22
,0005	.945	1.41	9.64	2.20
0	.945	1.40	9,54	2.19

Ridge Regression

- GridSearch and Cross Validation to determine best alpha. Price is log-transformed
- Best alpha, with 5-fold cross validation: 2.778
- How the model evaluate test data:

Alpha	R-Squared	MAE	MAPE	RMSE
2.778	.945	1.40	9.54	2.19

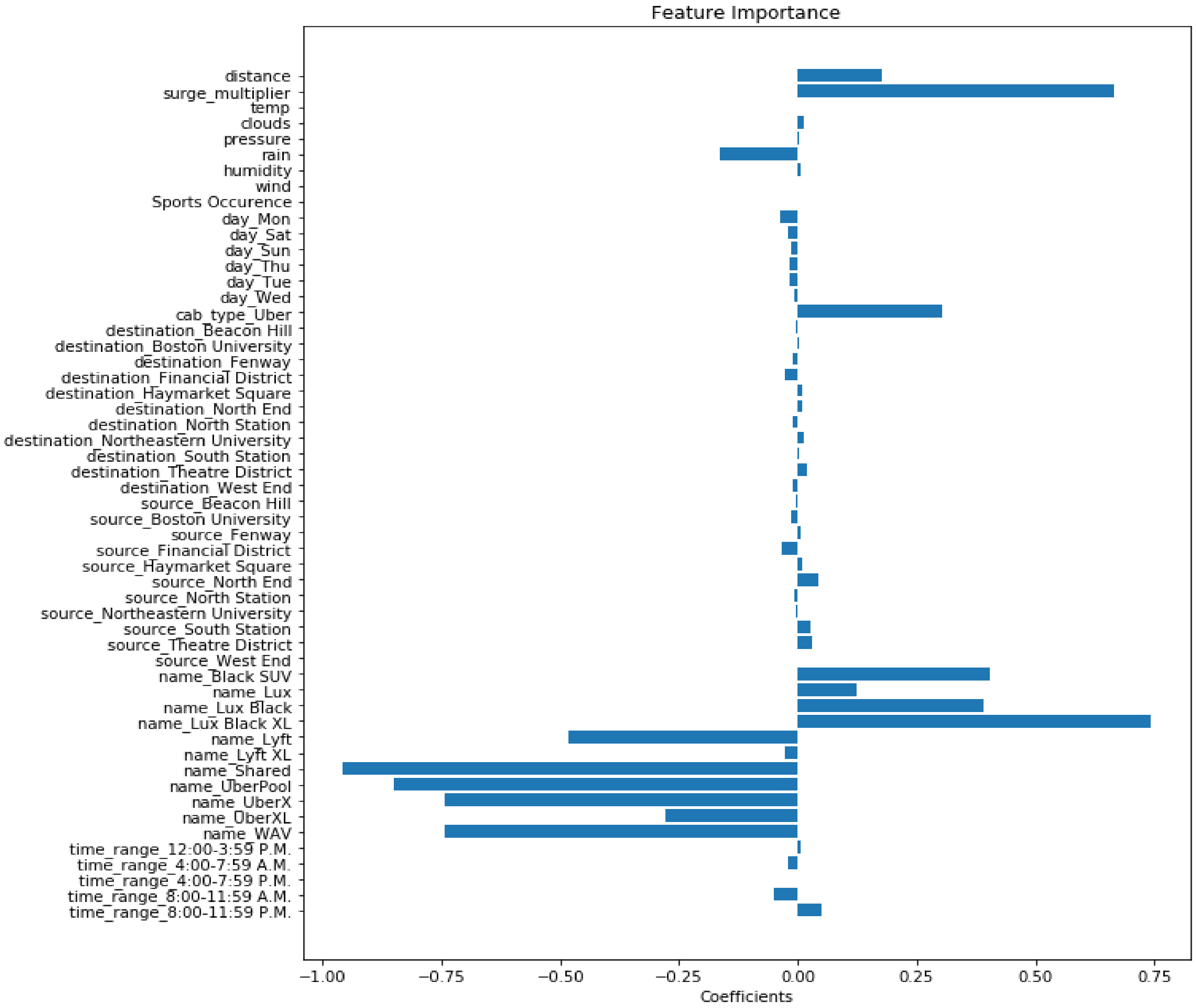
Model Comparisons on Test Data

When predicting for the training data or test data, the inverse function is applied to the predicted value for real valued prices on log transformed models.

Log-Linear Regression Model is the best model, as seen from its evaluation on test data, and simplicity compared to Lasso and Ridge regression.

Model	R-Squared	MAP	MAPE	RMSE
Linear Regression	.931	$1.71	12.95%	2.47
Log-Linear Regression	.945	$1.40	9.54%	2.19
*Lasso Regression	.945	$1.40	9.54%	2.19
Ridge Regression	.945	$1.40	9.54%	2.19

*Lasso regression with alpha = 0, which is the same as linear regression



Feature Importance

Coefficients do not have a direct relationship with the dependent variable, price, because of the log-linear relationship

Feature Importance (cont.)

- Top 5 features:
  1. Ride-Type
  2. Surge Multiplier
  3. Distance
  4. Cab-Type
  5. Rain

- Two features, humidity and Sports Occurrence, have p-values > 0.05, not statistically significant!
- These two features are removed from the dataset, and log-linear regression is re-ran.
- Model did not improve significantly.

Model	R-Squared	MAE	MAPE	RMSE
No features removed model	.945	1.40	9.54	2.19
Features removed model	.945	1.40	9.54	2.19

Recommendations to Clients

- Riders, when considering shared rides, solo rides, rides with up to 6 people, or premium black care service rides with up to 6 people, should choose Lyft over Uber.
- Otherwise, choose Uber for premium black car service rides for up to 4 people.

- Riders should expect prices to be cheaper under rainy weather, during daytime hours, or on most of the weekdays except Monday.
- More rides have a chance to be surged at late-early hours in the day, and generally on weekends.

- Competing ride service applications should use the model and offer discounted rides less than a $1.40 less than the predicted price to drive demand.
- These competitors should also lower the surge multiplication of rides during late-early hours or weekends to provide a less expensive alternative to Uber and Lyft

Limitations and Future Improvements

- We assume all rides are independent, even though there may be some time-correlations.
- We assume Uber and Lyft assign a price to a ride with the same algorithm.
- Used only rides from Boston, and only occurring from a span of 2 weeks.
- Missing Uber Taxi ride-type.

- To improve the model, we want to implement other regression models, such as regression trees, random forest, and gradient boosting, to capture some off the non-linear relationships.
- We want to include other major cities ride data to lessen bias towards Boston.
- Include data over the period of years and determine whether price trends change over time.
- Include other features such as traffic data, latitude and longitude of ride-pickup and drop-off, or other transportation features to improve the model.

Conclusion

- Log-Linear Regression Model best evaluates unseen ride data, with an error of $1.40, or a percentage error of 9.5%, within the actual ride price.
- 4 of the 5 Uber ride-types: UberPool, UberX, Uber XL, and Uber Black SUV, on average, are more expensive per mile than their Lyft counterparts. Lyft Black, on average, is more expensive per mile than Uber Black.
- Rides, on average, are cheaper under rainy and cloudy weather, than clear weather.
- All features, except humidity and Sports Occurrence, are used from both the ride and weather data.
- With more complex models, the evaluation of unseen data can be improved for the future.