Kevin Qualls
DATA 1030
Sep. 30, 2019

DATA 1030 Proposal

Problem Description

For this project, I would like to examine data on Uber and Lyft prices, and see if prices can be predicted based on distance, source of pick-up, day of week, and time of day. It would be interesting to see if the prices my model predicts are in line with those of Uber and Lyft. The data collected was from rides in the Boston, Massachusetts area November 25 – Dec 18, 2018.

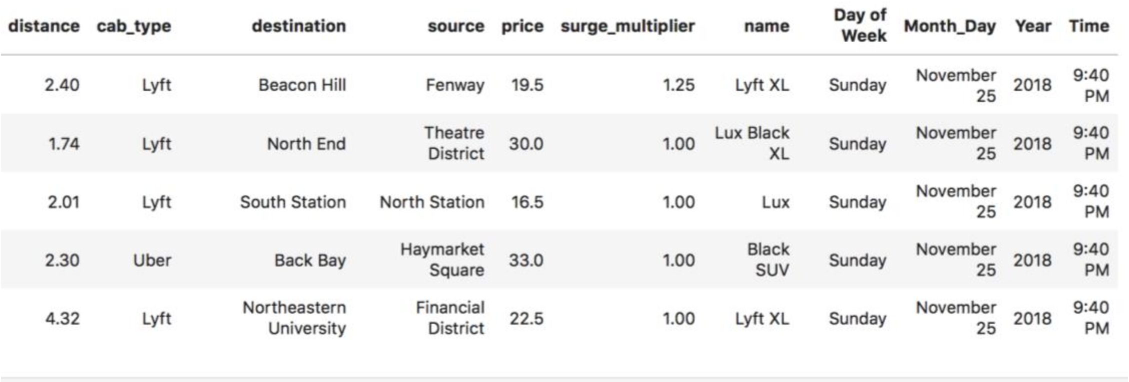
Requirements

- The target variable is price of Uber and Lyft.
- This would be a regression problem, since price is a numerical feature.
- The information gathered from this machine learning model will prove interesting/important, since users can strategically determine when to purchase a Lyft or Uber at minimal cost, based on distance, day of week, source of pick up, and time of day.

Original Data Set

- *cab_rides.csv* || 693,071 rows || 10 features
    - Feature 1: distance: distance traveled (numeric) || Scaler Encoder
    - Feature 2: cab_type: lyft or uber (categorical) || OneHot Encoder
    - Feature 3: time_stamp: time of day (categorical) || Ordinal
    - Feature 4: destination: place person is going to (categorical) || OneHot Encoder
    - Feature 5: source: place person is leaving from (categorical) || OneHot Encoder
    - Feature 6: price: price of ride (numeric)
    - Feature 7: surge multiplier: multiplier for price (numeric) || Scaler Encoder
    - Feature 8: ID: id for passenger (categorical) || OneHot Encoder
    - Feature 9: Product ID: type of lyft or uber (categorical) || OneHot Encoder
    - Feature 10: name: similar to Feature 9 (categorical) || OneHot Encoder

Originally there were 10 features, but I decided to split up Feature 3 in Excel into 4 columns: day of week, month and day, year, and time. I also deleted Features 8 and 9, as they were unnecessary. Below is a final version of the table, with 11 features and 693,071 rows:



distance	cab_type	destination	source	price	surge_multiplier	name	Day of Week	Month_Day	Year	Time
2.40	Lyft	Beacon Hill	Fenway	19.5	1.25	Lyft XL	Sunday	November 25	2018	9:40 PM
1.74	Lyft	North End	Theatre District	30.0	1.00	Lux Black XL	Sunday	November 25	2018	9:40 PM
2.01	Lyft	South Station	North Station	16.5	1.00	Lux	Sunday	November 25	2018	9:40 PM
2.30	Uber	Back Bay	Haymarket Square	33.0	1.00	Black SUV	Sunday	November 25	2018	9:40 PM
4.32	Lyft	Northeastern University	Financial District	22.5	1.00	Lyft XL	Sunday	November 25	2018	9:40 PM

Reference: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices/downloads/uber-lyft-cab-prices.zip/4

There was also a second dataset for this problem, called *weather.csv*, however, I found this dataset unnecessary for my analysis, so I decided to just focus on *cab_rides.csv*.

New Data Set

- *cab_rides.csv* || 693,071 rows || 11 features
    - Feature 1: distance: distance traveled (numeric) || Scaler Encoder
    - Feature 2: cab_type: lyft or uber (categorical) || OneHot Encoder
    - Feature 3: destination: place person is going to (categorical) || OneHot Encoder
    - Feature 4: source: place person is leaving from (categorical) || OneHot Encoder
    - Feature 5: price: price of ride (numeric) || Target Variable, therefore no encoder
    - Feature 6: surge multiplier: multiplier for price (numeric) || Scaler Encoder
    - Feature 7: name: type of ride (categorical) || OneHot Encoder
    - Feature 8: Day Of Week (categorical) || OneHot Encoder
    - Feature 9: Month_Day (categorical) || Ordinal Encoder
    - Feature 10: Year (categorical) || Ordinal Encoder
    - Feature 11: Time (categorical) || Ordinal Encoder

Summary of Previous Work

Kernel #1: Starter: Uber & Lyft Ride prices : Random Forrest

This project merged the weather and cab_rides dataset, and then split it into a training and test data set. It seems it uses a model to predict prices, and this model has a 91.86% accuracy rate. A confusion matrix was also made to predict the surge multiplier. At the end, it was calculated the accuracy of the classifier that was made is 76.25%.

Kernel #2: Starter: Uber & Lyft Cab prices d40ffa30-6

This project only focused on exploratory data analysis, with no prediction models.

Preprocess the Data

I preprocessed the following features with a Scaler Encoder, since they were numeric values, without a minimum or maximum: “distance”, “surge multiplier”.

I preprocessed the following features with a OneHot Encoder, since they were categorical and could not be ordered: “cab_type”, “destination”, “source”, “name”, “Day of Week”.

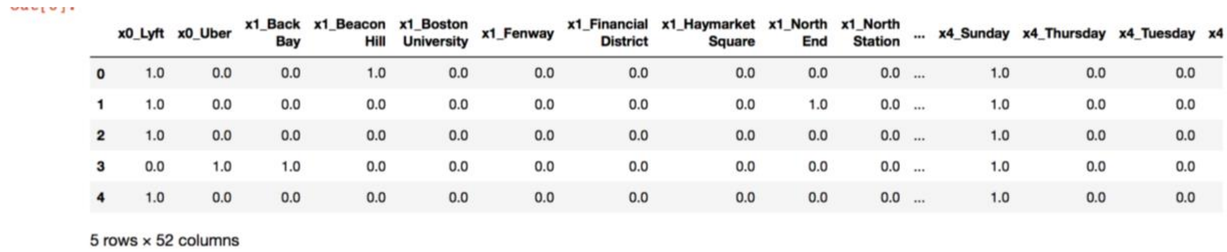
I preprocessed the following features with an Ordinal Encoder, since they were categorical and could be ordered: “Month_Day”, “Year”, “Time”.

Lastly, I did not transform the feature, “price”, since it is the predictor variable.

Reference: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices/downloads/uber-lyft-cab-prices.zip/4

Kevin Qualls
DATA 1030
Sep. 30, 2019

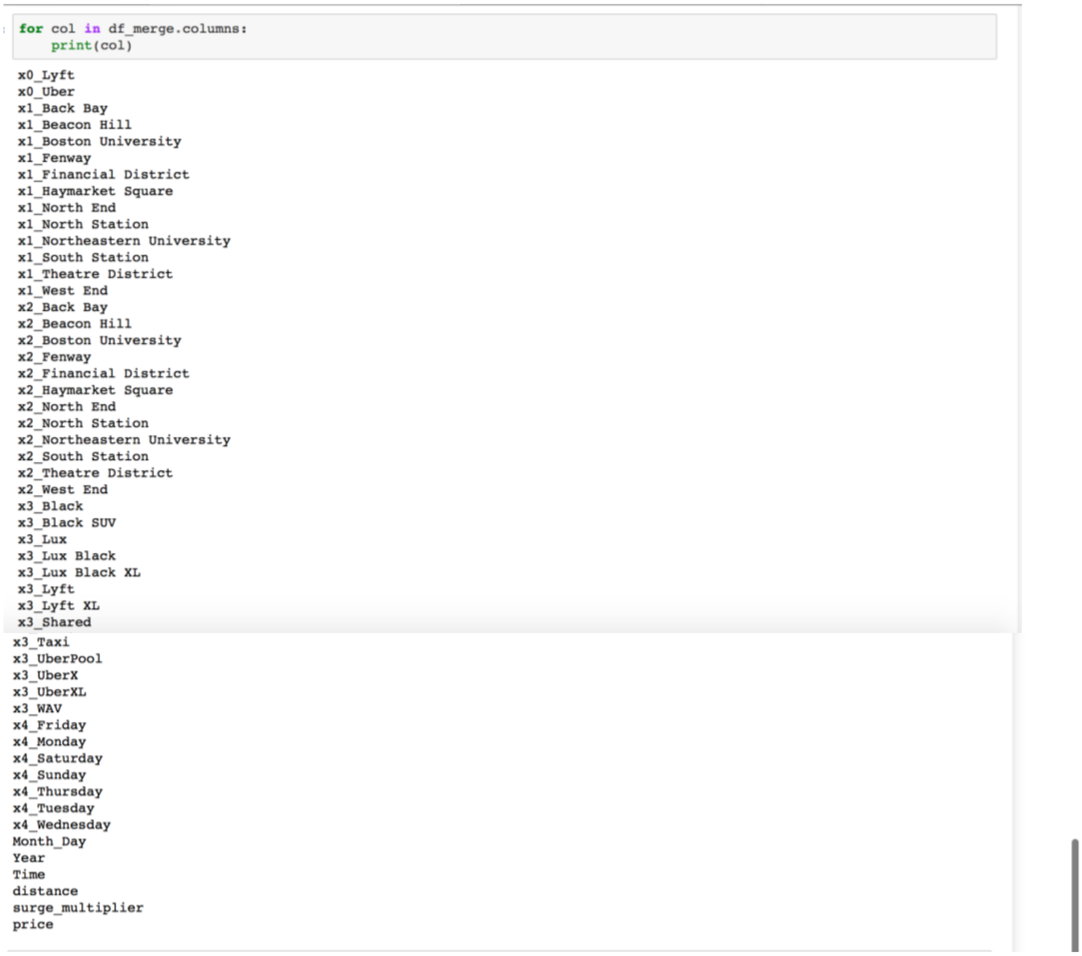
The final preprocessed dataset has 5 rows and 52 features, shown below:



	x0_Lyft	x0_Uber	x1_Back Bay	x1_Beacon Hill	x1_Boston University	x1_Fenway	x1_Financial District	x1_Haymarket Square	x1_North End	x1_North Station	...	x4_Sunday	x4_Thursday	x4_Tuesday	x4
0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	
1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	1.0	0.0	0.0	
2	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	
3	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	
4	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	

5 rows × 52 columns

Here is a list of all 52 features as well:



```
for col in df_merge.columns:
    print(col)
```

```
x0_Lyft
x0_Uber
x1_Back Bay
x1_Beacon Hill
x1_Boston University
x1_Fenway
x1_Financial District
x1_Haymarket Square
x1_North End
x1_North Station
x1_Northeastern University
x1_South Station
x1_Theatre District
x1_West End
x2_Back Bay
x2_Beacon Hill
x2_Boston University
x2_Fenway
x2_Financial District
x2_Haymarket Square
x2_North End
x2_North Station
x2_Northeastern University
x2_South Station
x2_Theatre District
x2_West End
x3_Black
x3_Black SUV
x3_Lux
x3_Lux Black
x3_Lux Black XL
x3_Lyft
x3_Lyft XL
x3_Shared
x3_Taxi
x3_UberPool
x3_UberX
x3_UberXL
x3_WAV
x4_Friday
x4_Monday
x4_Saturday
x4_Sunday
x4_Thursday
x4_Tuesday
x4_Wednesday
Month_Day
Year
Time
distance
surge_multiplier
price
```

Reference: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices/downloads/uber-lyft-cab-prices.zip/4