Ride-Hailing Apps Pricing Analysis and Prediction

EAS 503: Final Report
Submitted by: Group 28

Motivation:

Ride-hailing apps are widely used to hire vehicles and the driver as a convenient mode of commute. Uber & Lyft are the most well-known ride-hailing apps. The app users simply need to enter their pickup location and the destination, confirm their booking by making the payment. Although the source and destination are unchanged or the distance of commute for different locations remains the same, customers most might need to pay different price for same travel distance.

Hence, we want to understand the contributing factors/features from the dataset that can help us determine the posterior price.

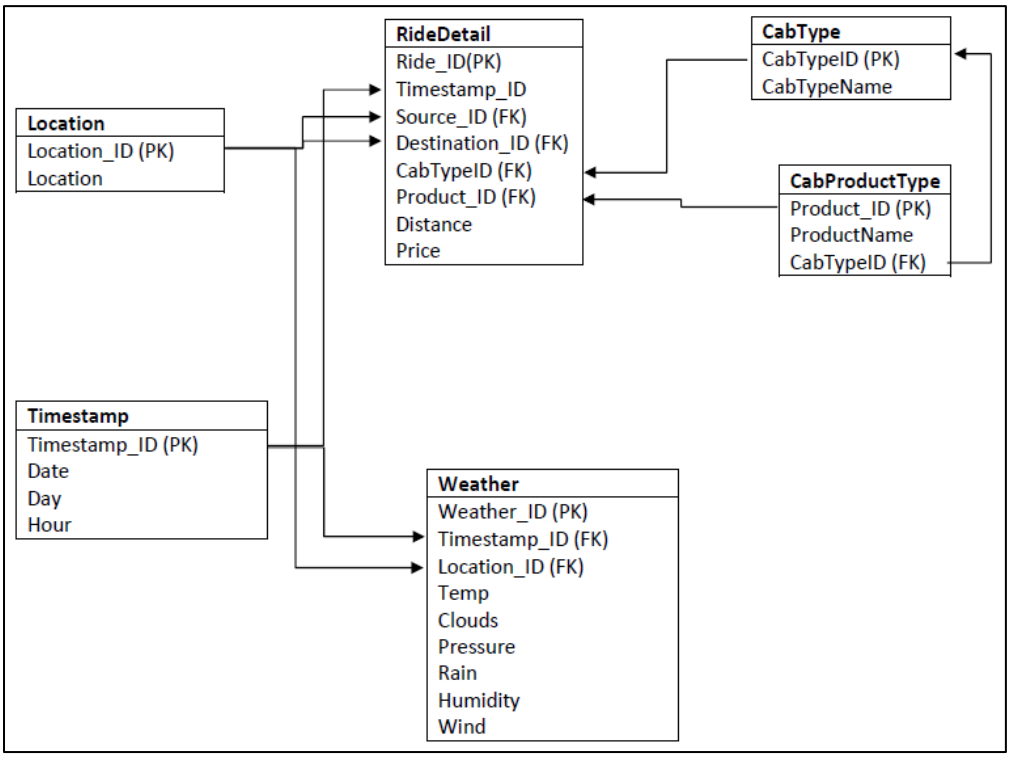
Data Description:

Dataset: https://buffalo.box.com/s/afevpzhs9uy1igfpym7z0u3yg9cvvi0a
(Alternative Link): https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices

The dataset includes two different CSV files, namely 'cab_rides.csv' (693,071 rows, 10 columns) consisting of the pricing data and 'weather.csv' (6276 rows, 8 columns) consisting of weather conditions. These data are based on the usage of Uber & Lyft in Boston-based locations.

Approach and Analysis

- Creation and maintenance of relational database: A relational database was created using the two CSV files. A normalized database with 5 tables (*RideDetail, Location, CabType, CabProductType, Weather, Timestamp*) was created.



- Exploratory Data Analysis:

The data split for *'Uber'* to *'Lyft'* is 52-48. The data for the available cab type is evenly split among the available cab types. As the data was evenly split there was no need for resampling of data.

We analyzed the data to check for geographical patterns, the number of rides booked by location. There was no significant difference in the number of bookings in terms of location. Similarly, there were no significant differences in the number of bookings with other features.

From the above analysis, we clearly understand the data is evenly distributed. As the desired variable of analysis is 'price', we had performed basic analysis on price and features using visualization. One key observation is, for higher price points the range of ride distance is less than that of lower price points. We have analyzed timely patterns for cost of the ride at hour of the day and day of the week. We have found a significant increase in price for bookings on Sundays and Mondays. On Sundays the significant increase in bookings was from 7 am to 10 am and the peaks on the later hours of the day for Mondays. Form the analysis of price across the ride class, we seen price for luxury class ride has higher mean and range compared to regular and shared class.

Predictive Modeling:

Most of the features has shown a nonlinear or no relation to the price on visual analysis. To best predict the price using the data, a nonlinear model is best suited. We have used an Artificial Neural Network model, as they can accommodate for the nonlinearity and having multiple hyper parameters that can be adjusted to fine tune the model to fit the data.

For the model, missing values and outliers in the data are handled with minimal data loss. To improve the performance of the model and limit the drastic growth of outputs from the nodes, we have normalized. We have encoded all the categorical features using one hot encoding. As the neural network models tends to overfit the data, to test the generalizability of the model and accuracy on unseen data we have randomly split the data to train and test. A neural network model with four hidden layers. The RMSE on train set is 1.97 and the test set is 1.99. The below is the summary of the model architecture.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 128)               6912
_________________________________________________________________
dense_1 (Dense)              (None, 64)                8256
_________________________________________________________________
dense_2 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_3 (Dense)              (None, 8)                 264
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 9
=================================================================
Total params: 17,521
Trainable params: 17,521
Non-trainable params: 0
_________________________________________________________________
```

CONCLUSION:

Thus, using the dataset obtained, we created a normalized database, performed some exploratory data analysis and built a model using neural network. With the built model, the price of the ride can be predicted with an average error of $2.

As a future enhancement to the project, the applications of the project can be used only by the individual users whereby they can compare their commuting price for Uber & Lyft, but can also be used to impose price regulations to prevent unnecessary price surges.