How Weather Influence Lyft Price

Yuhan Wang, Yi Gu, Tianhao Gao, Yige Sun

ABSTRACT

Lyft has changed the face of taxi ridership, making it more convenient and comfortable for riders. However, most consumers are unclear about what they are charged for besides distance. It is a challenging task to understand how Lyft set their rates. This gets more complicated with changes in weather. In this project, we attempt to estimate weather's influence on Lyft's price in Boston. We include different weather features to this analysis to see how it impacts the changes in the price rates.

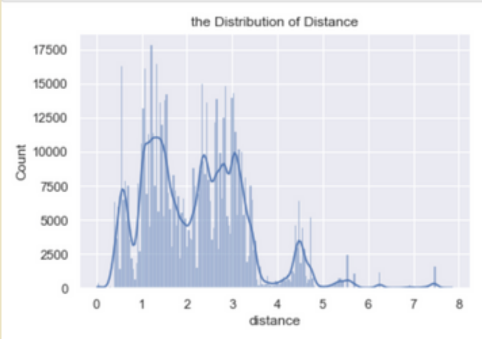
INTRODUCTION

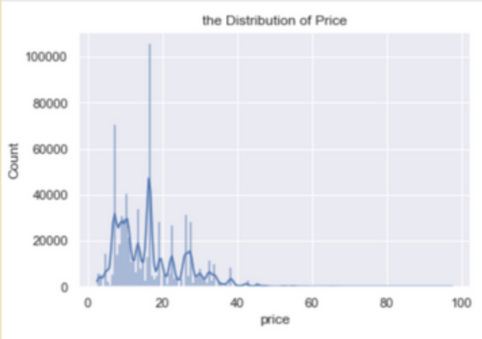
Now when people want to go to some places which are a little bit far, Lyft is a good transportation choice. You do not need to worry about the parking location, and you can have a good rest on the trip. In Boston, there were 42.2 million rides started from here in 2019. However, the weather in Boston often changes, which may influence price rates of Lyft. For example, on snowy days, the price of Lyft rides may become higher. This makes our group feel curious about the relationship between Lyft price and the weather in Boston. In this research project, we will the related data frames from Lyft's official website to analyze different weathers' influence on Lyft price rates.

The motivation for our topic comes from daily life. Lyft, a famous ride-hailing service providing applications, is commonly used when planning a trip. The slight change in its pricing might greatly affect our decision on trip mode. From our perspectives, the price of a trip might be related to the weather condition, since available drivers might be limited on rainy days and snowy days due to limited supply versus tremendous demand. However, other weather factors might affect total pricing, since, in the rainy seasons, people might not be willing to travel far, resulting in a relatively low total fee. All in all, understanding the relationship between weather and pricing could help us judge whether we should travel by Lyft in certain weather conditions.

METHODOLOGY

The data chosen for this project is simulating Lyft rides using real prices. These real prices were determined by considering if someone were to take a ride on Lyft it would cost that price. Lyft does not make its data public. This cab ride data was collected by using Uber & Lyft API queries and corresponding weather conditions. The data was collected for a singular week in the month of November 2018.
We needed to process the data before we could use it because it was not uniform. In our EDA process, the rain column had null values which were supposed to mean it did not rain. They were changed to 0 so that they could be put on the visualization as quantitative data. The prices also had certain missing values. We removed the rows which had missing values for prices because that was the most important factor. We also show the distribution of price and distance:





Then we divided the distance by price and save it as the column in our dataset, so it would have higher relationship with weather features. Finally, the time values were in the form of time stamps so they were converted to date times. All of this processing was done in python.

Two datasets are involved in this project: weather and cab. In order to build the model, we combined two datasets based on the timestamp and the location where weather and cab ridership are shared in common.
We trained and tested on a 70-30 training/testing split of our dataset. For the evaluation, we performed grid search cross-validation on every type of model, tuning the hyperparameter to find the best model. The mean-squared error would be calculated by comparing the predicted results from the test set with the split result test set among all algorithms. The one with the minimum value would be selected.

After discussion, we decided to explore 4 algorithms: **Linear Regression, Random Forest Regression, Decision Tree Regression, and the K-nearest neighbors algorithm**.

Linear Regression: we think that maybe there's a linear relationship between our data since we predict that worse weather will lead to a higher price.

Random forest: It fits many decision trees onto the data, averaging them to produce an output. It can generalize data using its built-in function to perform feature selection, which makes the prediction accurate.
Decision Tree Regression: The Decision tree would both investigate the linear and non-linear relationship, which might be great for this project.
K-nearest neighbors algorithm: To identify a suitable value of k (number of neighbors), instead of iterating the value by the loop, we implement Grid Search Cross Validation, aiming to tune the hyperparameter for the model which has the best performance. The accuracy of the model would be determined by the mean squared errors between the predicted set and the split test set

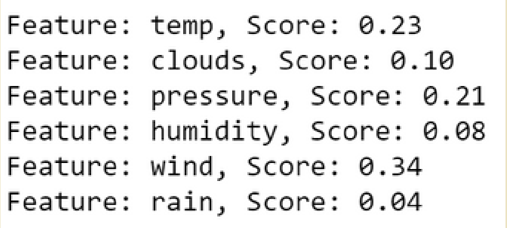
RESULTS AND EVALUATION

Linear Regression:
We first tried Linear Regression. The metric used to evaluate is MSE. The MSE between the predicted set and the split test set is **6.4**, which indicates that the model is not accurate enough.
No hyperparameter could be tuned in linear regression.

K-Nearest neighbors Algorithm:
We then tried the KNN model, where **n_neighbors = 1000**. The MSE of this model is **6.38**. Then, we use GridSearchCV to tune the hyperparameter n_neigbors in the range 1 to 1000, and 100 as step size. The optimal MSE we get is **6.13** where **n_neighbor = 100**.

Decision Tree Regression:
We got the MSE as **5.69**, which is better than the KNN model and Linear Regression. No hyperparameter needed to be tuned.

Random Forest Regression:
To get the best accuracy, we use *SelectFromModel* to evaluate the importance of each features:



```
Feature: temp, Score: 0.23
Feature: clouds, Score: 0.10
Feature: pressure, Score: 0.21
Feature: humidity, Score: 0.08
Feature: wind, Score: 0.34
Feature: rain, Score: 0.04
```

We can find that **temp**, **clouds**, **pressure**, and **wind** have more importance.
So we just chose these four features to train this model.

Still using Grid Search CV to tune the hyperparameter of this model using n_estimators in the range **[1, 1000]** and **200** as step size, we got the best parameter as **600** and the MSE as **5.65**. This MSE is better than the above two models but is still not very good. Our group guesses this may be because the weather reason does not **have such a big influence** as other reasons(like locations), and we may **not have enough data**, and **the dataset might be unbalanced.** If we have more data maybe it will be better.

IMPACTS

Our models show the relationship between different weather features and the price/distance of Lyft. Since it is specifically based on Lyft, we think that maybe Lyft company may have interest on these models. If they think that the weather features really influence on the price/distance, they can modify the price to get better profit. Not only Lyft, other taxi companies may also can get some information from the models, as the weather influence on Lyft may be similar on other taxis. Also, Lyft customers in Boston may also want to know about the relationship between the price/distance and the weather, so they can consider about going out by Lyft in different weathers.

CONCLUSION

In this project, we use four machine learning models to predict Lyft's price rate based on weather features. By tuning hyperparameters through Grid Search Cross Validation, Random Forest Regression could achieve the best performance with **MSE = 5.65**. Also, based on the feature selection in KNN regression, the **wind** is the most important weather feature, rather than rain, which is commonly believed to be the most important one.
Thus, we could not conclude that it is an accurate model. The possible reason would be that weather features might be hard to quantify and be evaluated. Also, the dataset we used just record **one week's** ridership, which is unbalanced.
For future research, more data is required to generate the model, including more varieties of weather features, like snowing days.

REFERENCE

Uber & Lyft Cab prices https://www.kaggle.com/datasets/ravi72munde/uber-lyft-cab-prices
Prophet Facebook: Automatic Forecasting Procedure https://github.com/facebook/prophet
Uber and Lyft averaged 96,000 rides a day in Boston in 2017: Report https://boston.curbed.com/2018/5/2/17310438/uber-and-lyft-boston-rides