Prediction of Taxi Fares in New York City

Pradyumna YM
PES1201700986
CSE dept., PES University

Anush V Kini
PES1201701646
CSE dept., PES University

Punit Pranesh Koujalgi
PES1201701502
CSE dept., PES University

***Abstract*—Every month, about 1.5 million taxi trips are completed in the city of New York. With the rise of taxi aggregation services like Uber and Lyft who have established a monopoly in this business in the recent times, taxi fares have become unpredictable due to various schemes like surge pricing. We plan to predict the taxi fares given the time and pickup and drop-off coordinates. The yellow cabs of New York City have partnered with Google to solve this task of predicting taxi fares, which has published a large dataset on kaggle through a competition. We approach this problem by using techniques to improve the data and come up with a model that can predict taxi fares accurately and efficiently using this published data in order to avoid the hassles of data collection. We feel that this will help the yellow cabs regulate their prices for a given trip according to inputs from our model. This also allows us to optimise the travel costs for passengers by predicting the start time for the least cost from point A to point B.**

***Index Terms*—Regression, Fare Optimisation, Machine Learning, Data Analytics**

I. Introduction

Recent years have seen the rise of app based taxi services like Uber and Lyft. This sudden growth in competition has left traditional cab services at a severe disadvantage. Further, most app based taxi services use the concept of 'surge pricing' to inflate the service rates according to supply and demand. This has led to unpredictable taxi fares putting the consumer at a severe disadvantage. Predicting the fare of a taxi ride can help passengers determine a favourable time to begin their journey while also assisting drivers to choose more profitable rides. Taking this idea a step further, a system that would predict the optimal time for a ride with least fare be a useful tool for passengers in suggesting the best time to take as cab ride.

In this paper, we report on the various techniques we have used to handle the dataset. In section 2, we briefly look at some of the related works. In section 3, we explain the preprocessing methods that we have used to clean our dataset. We also take a look at the feature engineered columns we have added to our dataset. In section 4, we explore some of the baseline models as a comparative study. In section 5, we describe the improvements we plan to make in the future to our model and data. In section 6, we summarise on the important aspects of our report.

II. Related Work

The taxi fare problem has mostly been approached as a regression problem. Various techniques have been modeled to predict the fare with minimal error.

Rishabh Upadhyay et al. [1] have classified the taxi fare amount into 5 different categories, (low, normal, med, high, extra high). Further, they have feature engineered columns holiday, time, weekend, distance from airport, city centre and popular tourist places. They have approached this problem with 2 different methods. Firstly, a 5 layer DNN, with ReLU units, and softmax output layer. Secondly, They have used a stacked classifier approach, where there are a set of base classifiers trained on the training data. Next, there is a classifier trained on these outputs of the base classifiers and the test data, which will be used to classify the taxi fares. Though the validation results of the neural network are promising, they fail to generalise to the test set. Further, no comparative study has been performed on similar approaches to the stacked classifier like AdaBoost classifier. Further, the classification approach used is of little use in the real world.

Xinwu Qian et al. [2] have used an approximate dynamic programming approach to model the underlying semi-Markov process. In their paper, they focus particularly on the time of pickup as a feature, which is actually an important factor in deciding the taxi fares. Also, the method discussed in this paper is computationally economical, considering the fact that no large ML models are trained.

Hai Yang et al. [3], focus on the linear distance-fare model used in Hong Kong and it's downfalls. They have proposed to model the taxi fares in Hong Kong as a nonlinear profitability-based taxi service model, which factors in the initial flag-fall charge of taxis and distance of the taxi ride.

K. Tziridis et al. [4] have investigated into the important features of influence for the prediction of Airfares,where they concluded that the day of the week and holidays are very important.They have approached the problem as both a regression problem and as a classification problem. With a limited data of about 1800 flights, they were able to achieve good results using bagging regression trees.

Frank Ivis et al. [5] have compared various techniques that can be employed to compute geographical 2-dimensional and spherical distances. Further, they have discussed methods of handling latitude and longitude data., which is pivotal in solving our regression problem.

Nitin R Chopde et al. [6], have used the Haversine formula as a reliable approximation to the distance between two GPS coordinates in their work to find the shortest path between two Coordinates in Google Maps, using Djikstra's algorithm.

Christophoros Antoniades et. al [7] have used various models such as linear regression with forward selection, and higher order terms of attributes, Lasso Regression, and Random Forests. They propose to transform the GPS coordinates in such a way that the streets of the city of New York are parallel to the x and y axes, as this might lead to a better split of the data in case of Random Forest Regressors. They have obtained impressive results through these models.

A. Our Approach

Most of the above approaches have focused mainly on predicting the optimal fares prices for the driver-passenger dynamic with the primary aim of aiding the taxi drivers. We propose to first perform a comparative study of the state of the art machine learning models as mentioned above literature. Further, we will then optimize these models to suit our problem statement of predicting the fare given information only available prior to the trip. We then plan to supplement these models by adding a system which determines the optimal time to embark on a trip. This system will aid passengers to minimise their fare amount. This solves the problems that are faced by the consumers due to the uncertain pricing schemes used by Taxi aggregators, while it can also be used by yellow cabs to fairly approximate their fares.

III. Data

We chose a dataset on NYC taxi fare prices had 7 columns and 55 million rows, due to which we loaded the data in chunks of 5 million rows, and cleaned, preprocessed the data, performed feature engineering and then concatenated these chunks after changing the datatypes of some columns to reduce the memory required. The descriptions of the attributes can be found here:

1) key: A unique string identifying each row.
2) pickup_datetime: Timestamp of when the taxi ride began.
3) pickup_longitude: The longitude coordinate of where the taxi ride began.
4) pickup_latitude: The latitude coordinate of where the taxi ride began.
5) dropoff_longitude: The longitude coordinate of where the taxi ride ended.
6) dropoff_latitude: The latitude coordinate of where the taxi ride ended
7) passenger_count: The number of passengers in the taxi ride.

After carefully examining the dataset, we decided to add more features to the dataset as the ones already present were too crude tor models to make sense of. So, we decided to add the following features:

- distance: a Haversine distance between the pickup and drop-off coordinates.
- weekday: an integer representing the day of the week.
- time: time in minutes since 12:00 AM that day.
- holiday: boolean value indicating whether it is a public holiday or not.
- year: the year during which the taxi ride was taken.

IV. Experiments

To start off, we implemented a simple Linear Regression model using the OLS approach, using all of the features described above. We were able to obtain an RMS error rate of about 5.38, which is a very promising start. We have also tried out Ridge and Lasso regressions, which have not produced RMS errors of 6.16 and 9.4 respectively. These results on linear regression models are without transformations performed on any of the columns in the dataset. We plan to obtain better results as they need to be optimised through cross validation for the shrinkage factor. We also tried using Random Forest Regressors, which also gave us good results, with an RMSE rate of about 4.47 on the test set.

TABLE I
A tabulation of the baseline results

Model	RMSE error
Linear Regression (OLS)	5.38
Ridge Regression	6.16
Lasso Regression	9.4
Random Forest	4.47

V. Future Works

We plan to work on tuning the parameters for the already established baseline models to tweak their performances further. We then plan to look into our dataset once again in order to add features if any are required. Later, we plan to look at various state of the art ML models that can be used for this regression problem. Once we come up with the best possible model for this task, we aim to be able to recommend passengers the optimal cost and start time based on the source and destination of the trip.

VI. Conclusion

We explore the current state of the art models in fare prediction in this work. After careful analysis of the dataset through visualization, we have feature engineered columns and modified the dataset. Further,we have built a few baseline models which have generated promising results. We plan to improve these baseline models, and then build model to beat the baseline and achieve state of the art results.

REFERENCES

[1] R. Upadhyay and S. Lui, "Taxi fare rate classification using deep networks," 09 2017.

[2] X. Qian and S. V. Ukkusuri, "Time-of-day pricing in taxi markets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1610–1622, June 2017.

[3] H. Yang, C. Fung, K. Wong, and S. Wang, "Nonlinear pricing of taxi services," *Transportation Research Part A: Policy and Practice*, vol. 44, pp. 337–348, 06 2010.

[4] K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 1036–1039.

[5] F. J. Ivis, "Calculating geographic distance: Concepts and methods," 2006.

[6] M. Nichat, "Landmark based shortest path detection by using a* algorithm and haversine formula," 04 2013.

[7] C. Antoniades, D. Fadavi, and A. F. Amon, "Fare and duration prediction : A study of new york city taxi rides."