A Research about The Dynamic Pricing System on A Region Specified Sample Combining Deep Neural Network and Random Forest Model

Mu-Huan Lee, Yaguang Liu, Ziyao Ding, Youhui Yang

Abstract— **Deep Learning have become one of the most popular research area these days. In this paper we are going to find out the potential of using a deep neural network to predict Lyft's prices. We will also compare the performance of our DNN model with traditional machine learning models.**

I. INTRODUCTION

Uber and Lyft's prices are not constant like public transport, nor does they have a price calculating formula like taxi. We are interested in what factors are affecting the prices. If we know what is affecting the price, can we predict the price given the information? Deep learning is one of the most popular research area these days. We want to know if a deep neural network can do a good job at predicting Uber and Lyft's prices. We will build multiple models based on different algorithms, and compare those models with the DNN model.

II. RELATED WORKS AND DISCUSSION

With many possible factors to these pricing models, many researches has been done to this topic. Many methods are used in these researches, which provided us with a better acknowledgement to this topic. We will discuss some of them bellow.

A. Demands and Weather

[1] and [2] both agree that demands should be a critical vector to decide the price in the system. Weather, in common sense, would largely impact demand of riding services. However, no further data or result is available in these papers showing the importance of whether. So we decide to test the effect of whether data on the data set we are using.

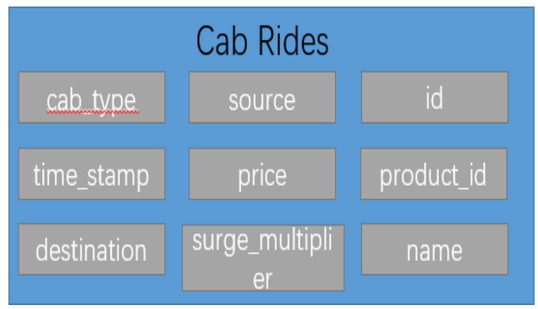
B. Models Estimation

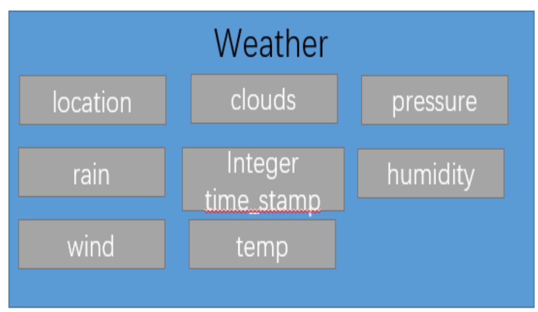
[3] has a overall estimation of machine learning models' performance in predicting the price for RoD (Ride on Demand) service. The result shows that decision tree and random forest models have an outstanding performance in this problem on the data set provided by DiDi in an undisclosed city of China. It indicates us that feature selection could play a more important rule in this question. However, that raises our curiosity about how would DNN model perform in the data collected from Boston area, and how well would it do when we combine it with the features selected out of the random forest model.

III. DATA SET

All our data were accessed trough [4]. According to the description of the data set, the Cab Rides data are collected from Uber and Lyft for a week in November 2018, for an interval of 5 minutes. While the Weather data are collected ever hour.

Attributes of the data set are as bellows.





IV. MODELS AND EXPECTATION

A. LASSO and Linear Regression

Linear regression is a linear approach. The way it works is to model the relationship between a scalar response and one or more explanatory variables. Linear regression tries to construct the relationship between two variables by fitting a linear equation with observed data. For example, a person might want to relate the heights of individuals to their weights or waistline using a linear regression model. It is one of the most poplar models ever. Lasso regression is a type of linear regression and it uses shrinkage, which means data will go towards a center point, like the variation, mean, etc. The lasso procedure encourages easy models that do not have many parameters. This model fits well when you want to automate some model selection, for example, variable selection and parameter elimination.

B. Random Forest

After tuning the parameters, we found that random forest regressor performs the best. This means our price prediction problem might be a feature selection problem. So we decided to see what feature plays the most important role.
below is the list of the importance ranking of the features.

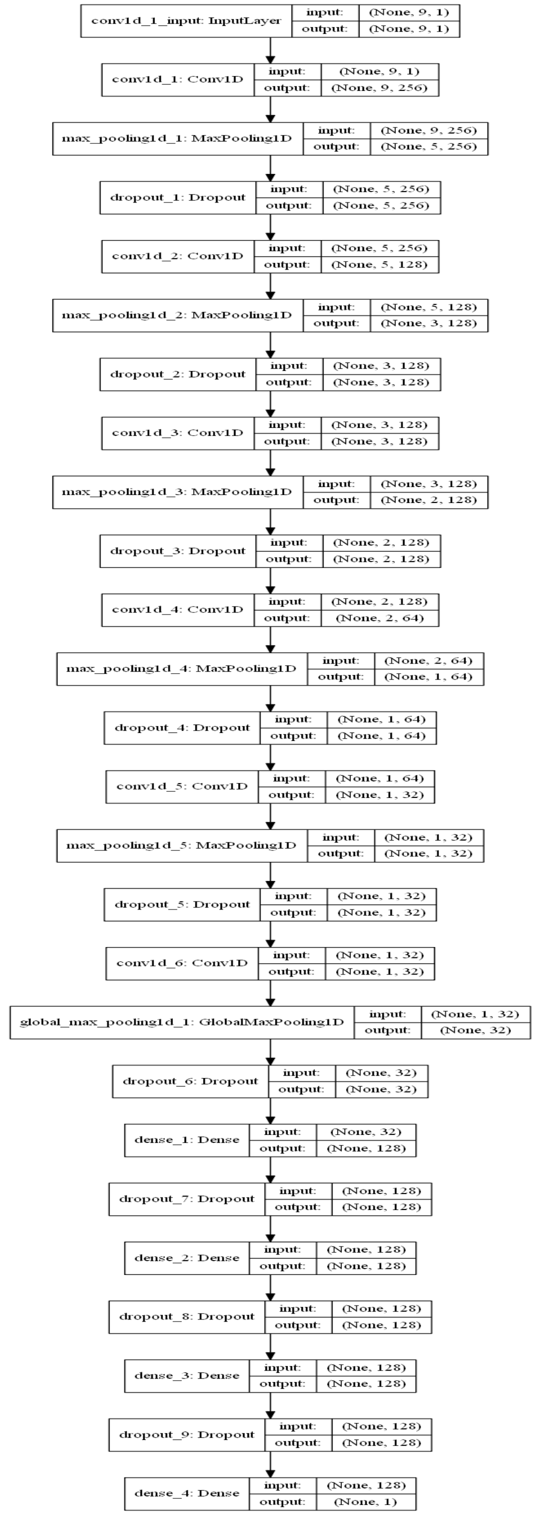
	importance
distance	0.663544
pressure	0.094210
day	0.067150
hour	0.047708
temp	0.043696
wind	0.031848
humidity	0.021651
clouds	0.018874
rain	0.011318

From the result, we can see that distance is far more important than the other factors in this model. However, it is surprising to see that pressure is having a such high importance to the prices, and raining has a really low score. It is possibly caused by the data collection. The data is collected among a week. In such a short period of time, the weather could be stable and thus the randomness of the importance of the weather factors would be understandable.

C. Deep Neural Network

DNN(deep neural network) is a artificial neural network with multiple hidden layers. There is a tradeoff about the depth of neural network. Deeper structure can provide more flexibility. It increases the non-linearity of the approximated function. But it also increases the risk of overfitting. Without activation function, neural network converts a no-linear problem into higher dimension to make it solvable. Activation function is a transfer function that transfer the input signal of a unit to a proper output signal. The using of activation function provide non-linear transformation to neural network. It makes neural network capable for more complex problems.To reduce the risk of overfitting, dropout is used as a regularizing approach to improve the generalization of DNN model.

In our neural network, we built 6 1D(1 dimensional) convolution layers with 6 1D max pooling layers and 4 dense layers. For each hidden layer, we use ReLu(Rectified Linear Units) as our activation function. We choose ReLu because it is idempotent. And ReLu requires less computational resources compare to other activation functions, for example, sigmoid and tanh. We also use dropout to reduce the risk of overfitting. For the optimizer, we use SGD(stochastic gradient descent) with momentum. Momentum can help the model's gradients change in the right direction.

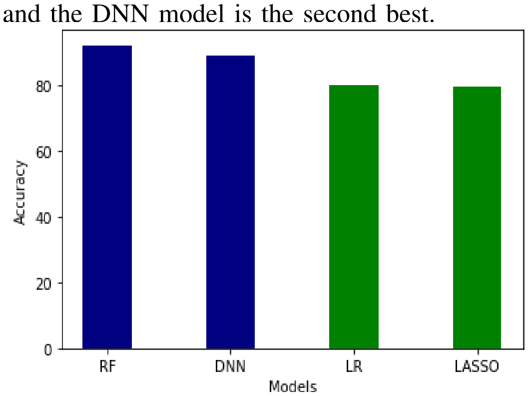


V. METRIX AND RESULTS

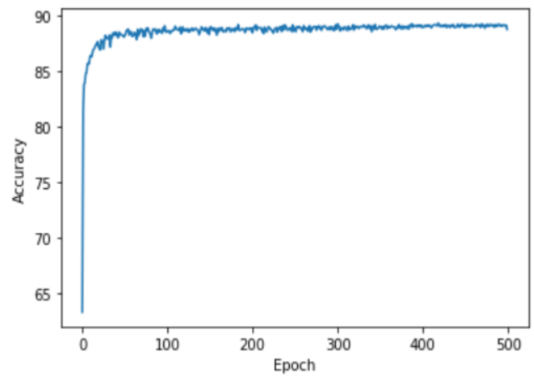
We choose to use MAPE (Mean Absolute Percentage Error) as our evaluation method for the following reasons:

- The MAPE metric has the advantage of being scale-independent and easily interpretable, with its percentage form. Furthermore, as the price multiplier is always positive, we can avoid getting undefined values in MAPE.
- Because of the special properties of price multipliers, other metrics such as MAE, MSE or RMSE can be directly represented from the prediction accuracy (i.e., in what percentage we have a absolute difference being 0, 0.1, 0.2...,0.6 between the predicted price multiplier and the ground truth), while MAPE cannot. This is because the price multiplier only takes discrete values such as 1.0, 1.1, ..., 1.6 (and we round the predicted price multiplier to these values), and so the difference between the predicted price multiplier and the ground truth also takes discrete values. As a result, metrics such as MAE/MSE/RMSE can be calculated from the prediction accuracy directly.
- Using MAPE to measure the relative prediction error is a common practice in evaluating forecast accuracy on, for example, human mobility pattern, taxi demand prediction and so on. Moreover, the baseline predictors we use in our evaluation also use the MAPE metric. To compare our results with baselines, we also use MAPE so that it can give a sense as to how our prediction model performs.

The figure bellow is the accuracy between all the model we have built. We can see that random forest performs the best, and the DNN model is the second best.



We think what causes this result is because some features are much more important than the other features, and random forest is good at finding these important features. The accuracy of our DNN model is quite similar to the random forest model. These two models outperforms lasso model and linear regression model. The figure bellow is the accuracy recorded during each training epoch.



VI. LIMITATION AND POSSIBLE FURTHER RESEARCH

A. Limitation

This research is only for the Uber and Lyft price in the great Boston Area. It is restricted by both the location and the companies.

The data is collected in the interval of a week in November to December, 2018. Hence, the features could be largely biased by the climate of that week.

Ride on Demand services uses price multiplier to judge the prices of trips during rush hours or between popular destinations. However, the data set only has 20975 inputs whose multiplier is not 1, given the total size is 693071. The small sample of rides with multiplier other than 1 could also influence the performance of the deep learning model. It would be likely that the model would rate the importance of the multiplier lower.

B. Further Research

Future work includes try different types of normalization on our data and improve the structure of neural network. The data collection could also be improved. More data should be collected in different period of the whole year, and under different weather situations, so that the model could capture more information from the weather factors.

VII. CONCLUSIONS

This paper investigates the performances of random forest, deep neural network, lasso regression and linear regression on the prediction of Lyft price in Boston area. The features includes distance, time, and weather. The result shows that random forest gets the best performance among the algorithms that we tried. The deep neural network's performance is very close to the performance of random forest. Lasso and linear regression have similar accuracy score.

REFERENCES

[1] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, and D. M. Chiu, "Fine-grained Dynamic Price Prediction in Ride-on-demand Services: Models and Evaluations," Mobile Networks and Applications, 2019.

[2] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, D. Zhang, and D. M. Chiu, "A Simple but Quantifiable Approach to Dynamic Price Prediction in Ride-on-demand Services Leveraging Multi-source Urban Data," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 3, pp. 1–24, 2018.

[3] Ismail Saadi, Melvin Wong, Bilal Farooq, Jacques Teller, Mario Cools, An Investigation into Machine Learning Approaches for Forecasting Spatio-temporal Demand in Ride-hailing Service.
[4] RaviMunde, Uber Lyft Cab prices Cab and Weather dataset to predict cab prices against weather, Accessed on: Oct. 10, 2019. [Online], Available: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices