2020-10-8

Using Deep Neural Network to find the most effective method for cab driver in New York.

Zitao Jiang
987857

Introduction

Deep Neural Network is the cutting-edge algorithm for machine learning now. It has better performance in general than other traditional machine learning algorithm such as Support vector machine or Decision tree.

Figure 1

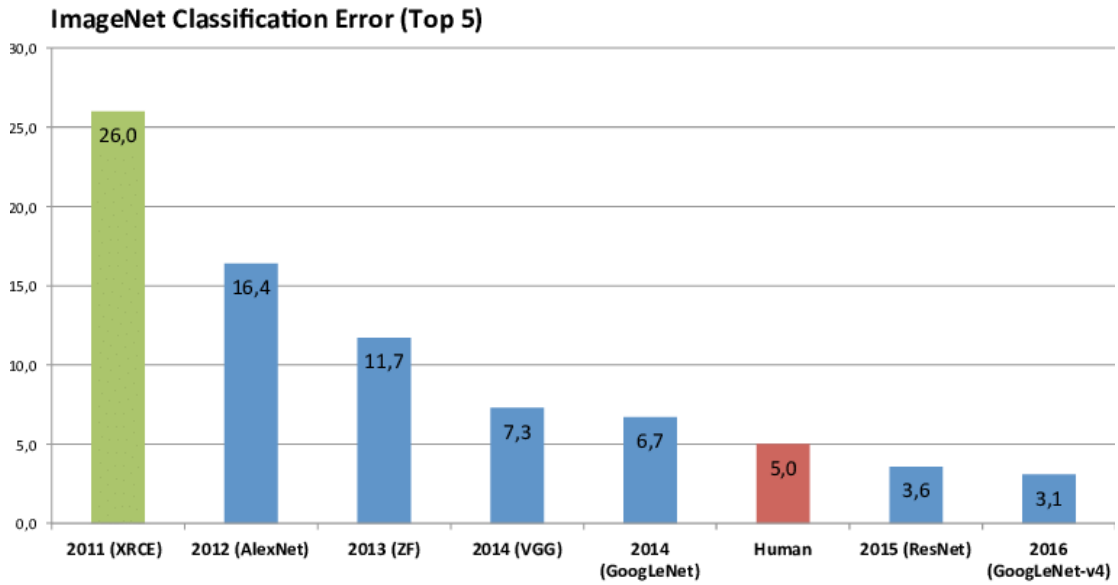


Figure 1 shows that since 2012 deep neural network has much better performance than SVM.

In most hard board game human ever created GO, Alphago which based on Deep Neural Network beats top human player in 2016. The Deep Neural Network also dominance many video games now such as StarCraft2 and Dota2. In this research, deep neural network tries to find most effective method for cab driver in New York city. More specifically, try to find which combination of attributes in order to maximize the total amount in each trip and to predict total amount in each trip. The potential stakeholder could be normal cab driver in New York. They can use the most effective method to increase their profit. The cab company such as uber and lyft could use this report finding to refine their business model and use better cab distribution strategy to increase their profit.

Data and Attribute Selection

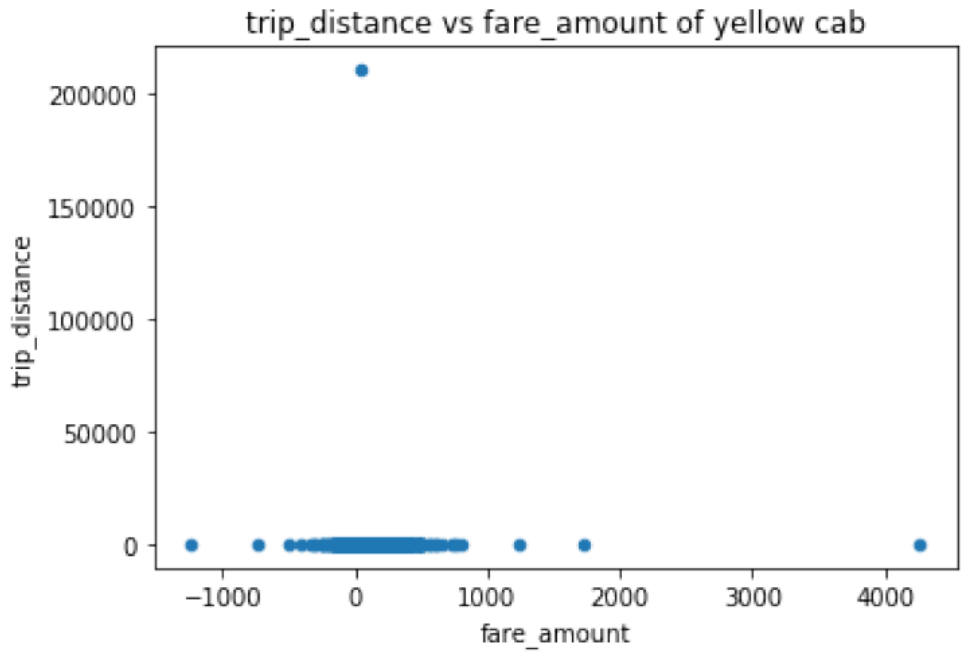
The data set used is Jan 2020 yellow cab data set from New York city government. Because it's the newest dataset since covid-19 hit New York, because covid-19 is not a normal situation in New York city. This factor needs to be separated in this report. And in the old data set, New York government records exact GPS location, although it can boost model accuracy. But for the ethical and privacy reason, in this report location ID is been used instead of GPS location. Because with combination of other data, individual's exact location and private information could be leaked. It also has 10^6 orders of magnitude instances which is large enough to make the model converge. The tpep_pickup_datetime, trip_distance, PULocationID and total_amount in yellow cab data set is been choosen to analysis. And daily temperature weather data from https://www.wunderground.com/ is been used as an external dataset. Because weather is strong factor for customers choose to used cab or public transport.

Pre-processing and Cleansing

First transfer csv file to feather file in order to save computation time and resource.

Similar to the assignment 1 report, first plot trip_distance and fare_amount to see any outlier in the dataset.

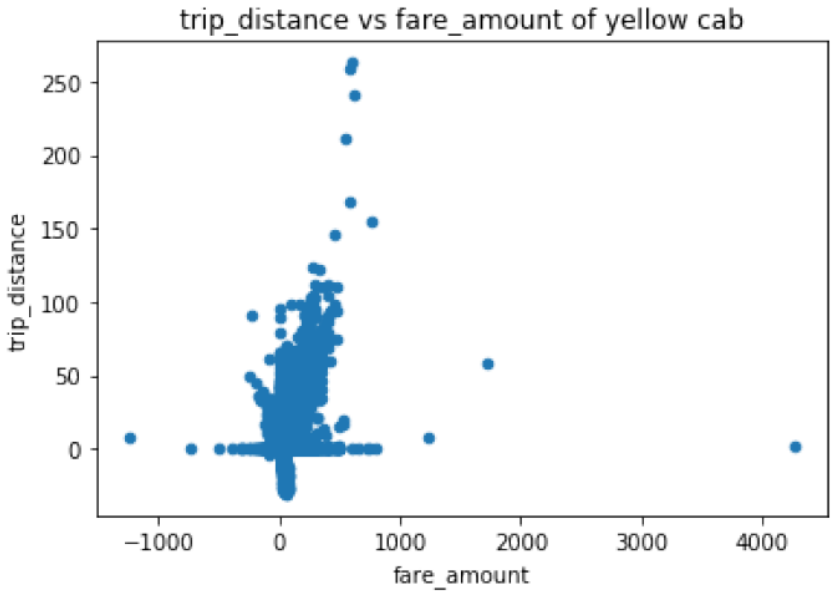
Figure 2



From figure2 there is an instance has over 200,000 mile and very small fare_amount, it might cause by sensor failure. It been treated as outlier and been discard from dataset

Plot new graph

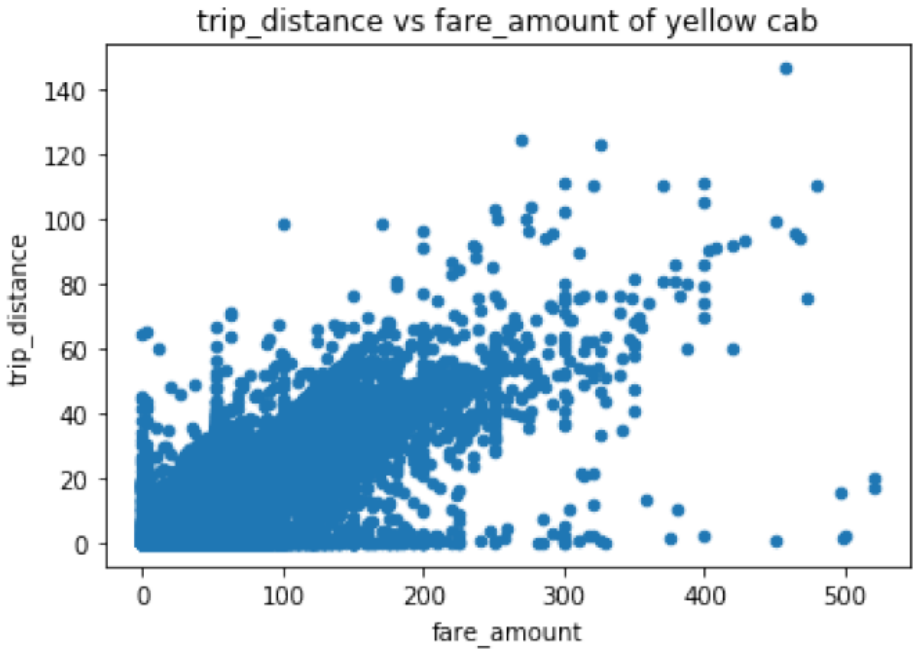
Figure 3



From figure 3, because it's unreasonable has negative fare_amount and trip_distance. Therefore treated as outlier and discard from dataset. Also the data point which has over 750 fare_amount and over 150 miles trip distance. They seem far away from main cluster. They will been treated as outlier and discard from dataset.

Plot final plot

Figure 4



From figure 4, there is rough linear relationship between fare_amount and trip_distance. All data point seems reasonable.

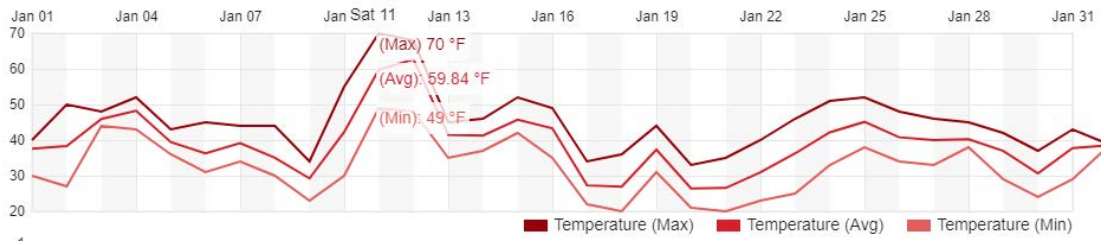
There is not any missing data in tpep_pickup_datetime, trip_distance, PULocationID and total_amount in yellow cab dataset. Also there is not any missing value in weather dataset. Therefore missing value do not need to be considered in this report.

For datatime feature attribute change string to timestamp datatype.

Finally merge weather dataset with yellow cab dataset.

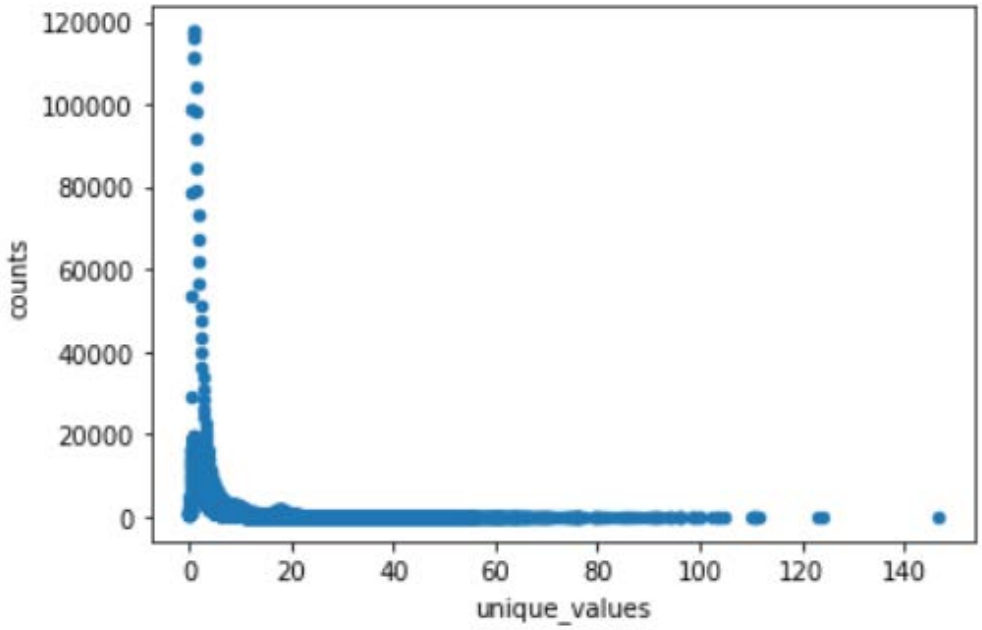
Descriptive analysis

Summery of weather data:



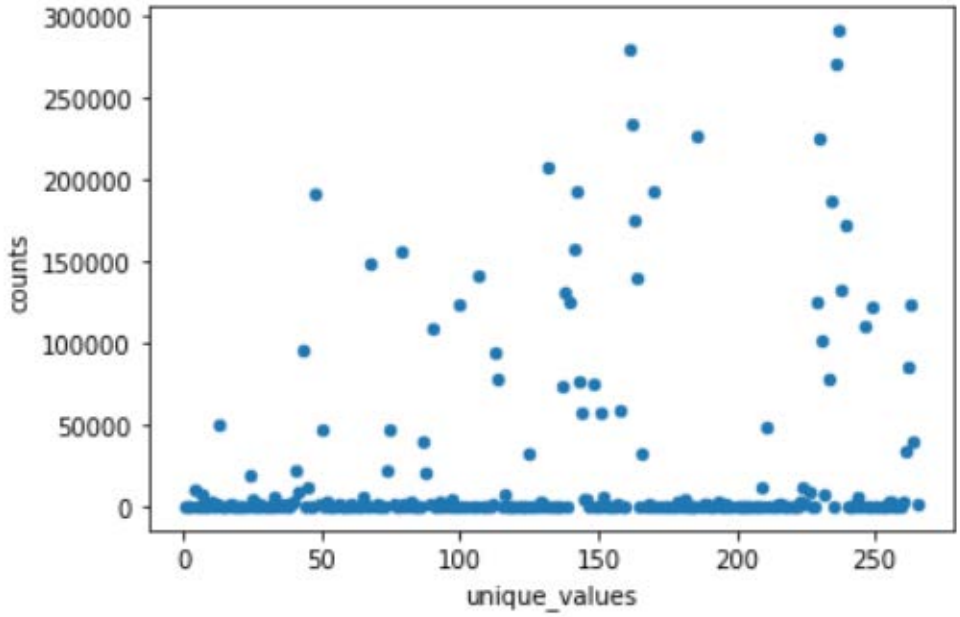
Average temperature between 30 and 50 F, a little hot between 11 Jan and 13 Jan.

Summery of trip distance:



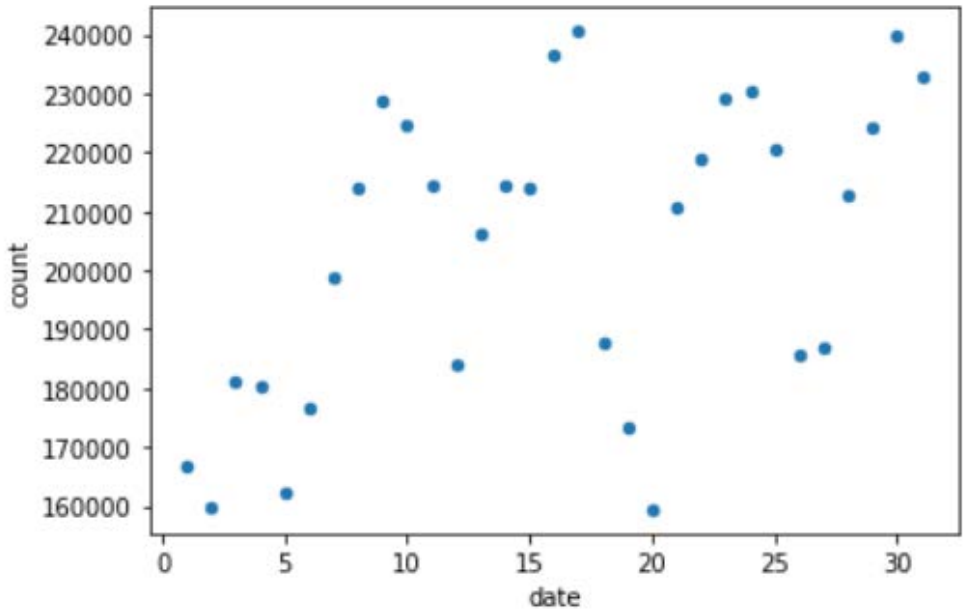
Most trip distance is below 10 miles

Summery of PULocationID



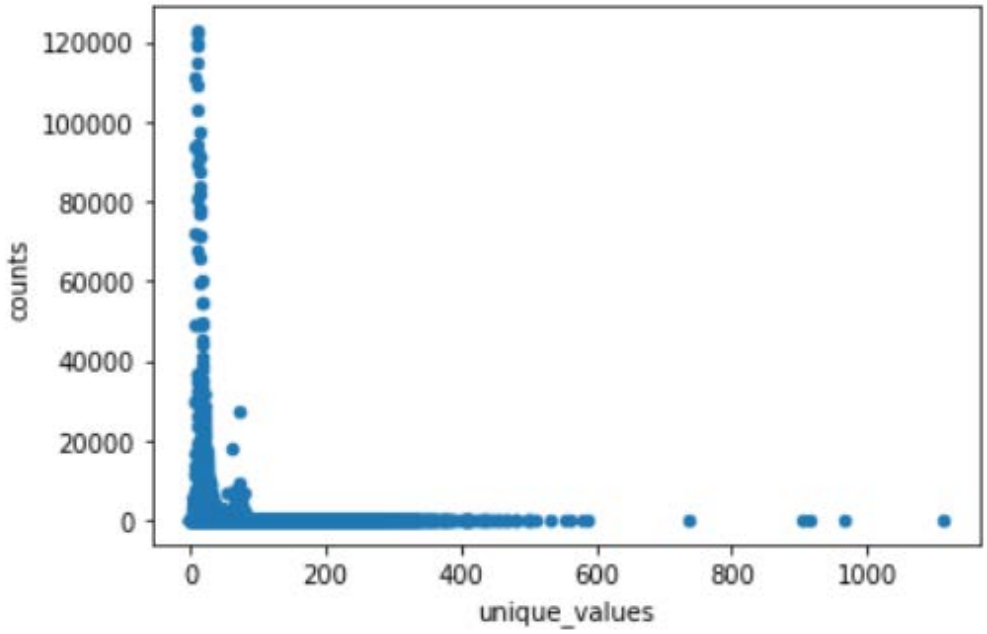
Most location do not many pick up, their counts are below 50000. Location ID 237,161,236 has the most pick up and way above other location.

Summery of tpep_pickup_datetime:



Datetime seem do not have any pattern.

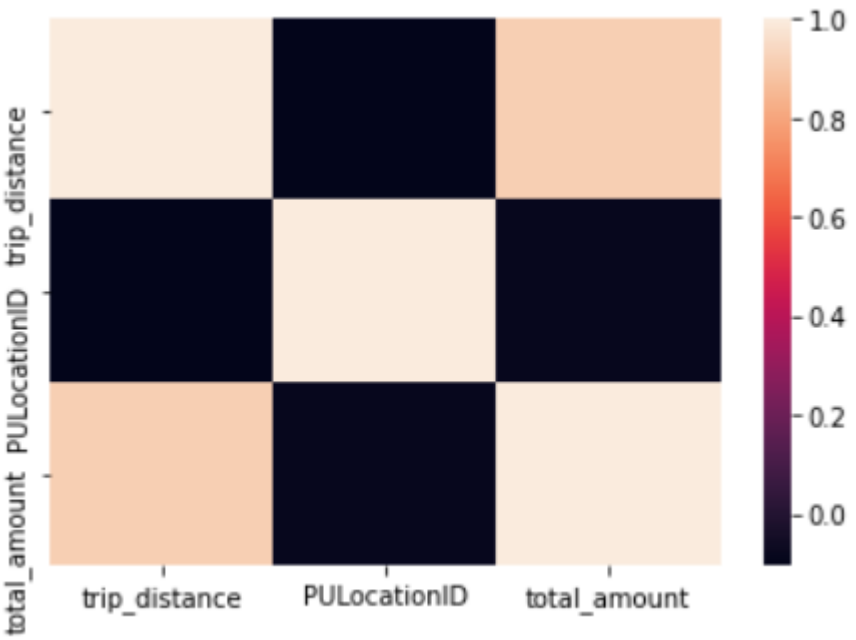
Summery of total_amount:



Most trip cost below 50 dollars, trip cost around 100 also has a little raise. It might cause by trips from or to airport.

Pairwise relationship:

Datetime and weather is independent from taxi dataset. Therefore, exclude them for this part.



Trip_distance and total amount has a strong correlation, PUlocation seem has small correlation with other two attributes.

Modelling

The statistical model used in this report is deep neural network. It was first introduced in 1949 by Donald Hebb. In Hebbian learning, neuron wire together if they fire together.[1]

Math equation for on layer:

$$y_i = f\left(\left[\sum_j w_j x_{ij}\right] + b\right) = f(\mathbf{w} \cdot \mathbf{x_i} + b)$$

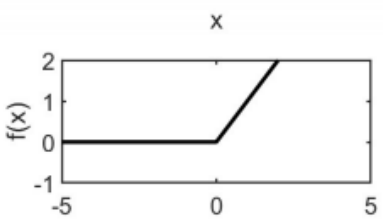
Input x, weight w, bias b, output y. Hyper-parameter: activation function f.

Neural network does not have many assumptions. It can do classification(output layer contains multiple neurons) or regression(output neuron has identity activation function). The neural network with multiple layers and non-linear activation function can approximate any continuous function on R^n(in simple word, it can learn anything unlike other traditional method)

The choice of activation function is RELU:

- rectified linear unit (ReLU):

$$f(x) = \max(0, x)$$



In the practice, RELU converge much faster the logistic function. It saved computational power and time. It was inspired by biological neural network where neurons do not activation unless it passes the threshold. It was first introduced by Hahnloser in 2000 with biological motivations and mathematical justifications.[2]

Because deep neural network is a really powerful algorithm and in theory it can find any pattern in the dataset. In practice, it has very high accuracy too. In order to find pattern in yellow cab dataset, deep neural network with RELU is a good choice.

Split train/test to 80/20, cost function is mean square error. It's regression model

Fit training data in to the model.

```
Model performance: 0.99
```

After tunning parameter(number of hidden nodes and layers, activation function) and try to find optimal value. The model performance seem does not change. Therefore stick with original parameter.

Analysis of results

Final model after refinement with RELU activation function. The final model has millions of weights and bias. Larger weight means strong connection between two neurons (they more likely fire together). In theory, last layer tries to find any pattern in previous layer.

The final model can be used for cab company in specific time which cab geo distribution will gain most profits. The cab driver can use this model to go to specific location in specific time in order to gain most total_amount in each trip.

Reference

1. Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley and Sons. ISBN 9780471367277.
2. Hahnloser, R.; Sarpeshkar, R.; Mahowald, M. A.; Douglas, R. J.; Seung, H. S. (2000). "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit". Nature. 405 (6789): 947–951. Bibcode:2000Natur.405..947H. doi:10.1038/35016072. PMID 10879535. S2CID 4399014.