Machine Learning Engineer Nanodegree, Udacity

Capstone Proposal

By Gagan Saini

New York City Taxi Fare Prediction

Contents

Domain Background

With the surge in app-based taxi providers, it has become very important to predict the taxi fare for a trip in advance. The user can then decide whether he wants to avail the ride or not. Also, as there are number of providers, a user can compare the fare of different providers for a trip and then choose the best one.

One simple way to estimate the fare is by using the distance of the ride and then calculate the fare based on the number of kms travelled and estimated price per km. But the price prediction in real world is more complex than that. Various other factors may affect the pricing like toll taxes, number of passengers, pickup and drop-off points, time of the day the ride is taken, advance booking etc. Some other variable factors are also considered by the taxi providers these days. For example, the prices will be high in case of high demand which is usually called surge pricing, the route taken by for the ride etc.

Fare prediction in advance directly affects the business because that is the only parameter based on which the customer will decide to go with the taxi provider or not. So, it is very important to predict the fare as much accurately as possible.

There are many academic research papers trying to solve the problem of fare & time prediction. The time taken can also directly affect the total fare of a ride. In [1], they have applied Support Vector Regression (SVR) to predict the time taken for a ride in advance. In [2], Neural Networks (SSNN) are used for time prediction. In [3], they have used deep neural networks for the fare prediction.

Problem Statement

To predict the fare amount for a taxi ride in New York City given the longitude and latitude coordinates of the pickup and drop-off locations, date & time of the pickup and number of passengers.

This is clearly a regression problem. Regression is used to predict the value of a dependent variable based on a number of independent variables.

In this problem, the fare to be predicted is the dependent variable and the features like pickup and drop off locations, number of passengers, and datetime of pickup are the dependent variables. We may deduce new feature out of the given features like distance travelled, special pickup and drop-off locations as few pickup and drop-off locations like airport may have higher fares than others. More details about the data are presented in the next section.

Dataset & Inputs

The dataset is provided by Kaggle [4]. It contains the following three files:

- train.csv - Input features and target fare_amount values for the training set (about 55M rows).
- test.csv - Input features for the test set (about 10K rows).
- sample_submission.csv - a sample submission file in the correct format (columns key and fare_amount). This file 'predicts' fare_amount to be $11.35 for all rows, which is the mean fare_amount from the training set.

As the test.csv file does not have the corresponding value of the target variable. We can not use this data to verify our model. So, I will use the data from the train.csv file only by splitting the data into train, validation and test sets.

Features

Column Name	Data type	Description
pickup_datetime	timestamp	The date and time when the ride started.
pickup_longitude	float	Longitude coordinate of where the taxi ride started.
pickup_latitude	float	Latitude coordinate of where the taxi ride started.
dropoff_longitude	float	Longitude coordinate of where the taxi ride ended.
dropoff_latitude	Float	Latitude coordinate of where the taxi ride ended.
passenger_count	integer	Number of passengers in the taxi ride

Target Variable

Column Name	Data type	Data Range	Description
fare_amount	float	- $300 to $93963.36	Total cost of the taxi ride in dollars.

Solution Statement

The problem can be solved by using regression techniques. At this point I am not sure which one will be best suited for the problem in hand. So, I will try a handful of regression techniques and will choose the one which works best in terms of predicting the fare. I will try Linear Regression, Decision tree regressor, SVM and Ensemble methods.

Benchmark Model

A simple Linear regression model will be used as a benchmark model. One such model [5] is provided on the Kaggle in the problem description of this completion. This model has an RMSE of 5.74184 on Kaggle. I will rerun this model on my trimmed representative data set and compare the performance of this model with the other models implemented by me.

Evaluation Metrics

RMSE (Root mean Squared Error) is used as an evaluation metrics in the original Kaggle Competition. So, I will also use the RMSE for the evaluation of my models.

RMSE measures the difference between the predictions of a model, and the corresponding ground truth [6]. It is a measure of the spread of the y values about the predicted y values.

RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where, y_i is the truth value and $\hat{y}_i$ is the predicted value.

A large RMSE means a large average error. So smaller values of RMSE are desirable.

Project Design

I will use the following plan for solving this problem:

1. As the data is huge (~55M rows), I will choose trim down the data set into a reasonably small representative data set.
2. As part of data visualization, I will try to get some insights into data by plotting some graphs.
3. As part of data preprocessing, I will remove duplicate data entries and the entries with missing data.
4. Train different models like Decision Trees, SVM, Ensemble methods on the data with the given features or some new derived features like distance.
5. Test the model on test set and evaluate each model's performance by calculating the RMSE.
6. Tune the hyperparameters of the models to get minimum RMSE.
7. Compare the model with the lowest RMSE with the performance of the benchmark model.

References

[1] Wu, Chun-Hsin, Jan-Ming Ho, and Der-Tsai Lee. "Travel-time prediction with support vector regression." IEEE transactions on intelligent transportation systems 5.4 (2004): 276-281

[2] Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. "Accurate freeway travel time prediction with state-space neural networks under missing data." Transportation Research Part C: Emerging Technologies 13.5 (2005): 347-369.

[3] Rishabh Upadhyay, Simon Lui, “Taxi Fare Rate Classification Using Deep Networks” Singapore University of Technology and Design (Sep 2017).

[4] https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data

[5] https://www.kaggle.com/dster/nyc-taxi-fare-starter-kernel-simple-linear-model

[6] https://www.kaggle.com/c/new-york-city-taxi-fare-prediction#evaluation