CAPSTONE PROPOSAL

Domain Background

New York City is one of the most populous cities in the United States. For an overly crowded place like this, traffic has been a major concern for over several years now. The transportation system here is a network of complex infrastructure that has one of the largest subway systems in the world, an aerial tramway, a vehicular tunnel etc. Taxis account for a significant contribution towards the daily commute of people residing in NYC. There was a survey conducted to know the number of taxi trips per day and it turned out to be a whopping number of around 850,000 trips per day through taxis. This number accounts to approximately 1/10th of the total population in New York City.

Problem Statement

This project aims to predict the estimated fare for a taxi trip in New York City for the given pickup and drop off locations. The inference could be achieved through supervised regression machine learning task. There are several factors that could influence the estimated fare and some of these include the point of pickup, destination, peak hours, high booking requests at the respective location, low availability of drivers, weather, number of passengers, time of booking request etc.

The above problem statement is from Kaggle https://www.kaggle.com/c/new-york-city-taxi-fare-prediction#description.

Datasets and Inputs

The aim of this project is to predict the estimated taxi fare. Hence, our target algorithm is fare_amount. Remaining columns are the input features which are subset of some of the reasons mentioned in problem statement as they influence the fare_amount.

Features

- pickup_datetime - timestamp value indicating when the taxi ride started.
- pickup_longitude - float for longitude coordinate of where the taxi ride started.
- pickup_latitude - float for latitude coordinate of where the taxi ride started.
- dropoff_longitude - float for longitude coordinate of where the taxi ride ended.
- dropoff_latitude - float for latitude coordinate of where the taxi ride ended.
- passenger_count - integer indicating the number of passengers in the taxi ride.

Target

- fare_amount - float dollar amount of the cost of taxi ride. This value is only in the training set; this is what we are predicting in the test set and it is required in our submission CSV.

Solution Statement

This is a clear supervised machine learning regression task. To tackle this problem I am going to try some of the below regression models:

- Linear Regressor
- Random Forest Regressor
- One of Adaboost and gradboosting regressors
- Deep Learning (if required)

Grid search will be used for finetuning the parameters for boosting models.

Benchmark Model

The author of this dataset in Kaggle calculated the basic estimate based on just the distance between the two points which resulted in an RMSE of $5-$8.

Evaluation Metrics

There are many ways to estimate how good a regression predictive model is. Some of the common evaluation metrics are

- Root mean squared error or RMSE
- Mean Absolute Error
- R2 score

Since the benchmark model for evaluation in Kaggle used RMSE, same will be used for evaluating the goodness of our prediction.

Project Design

- Data Preprocessing & cleaning – Visualize different features by looking at min, max, median and histogram. Will check to see if there are any outliers and do the data cleaning to remove noise.
- Model Selection- Experiment with different regression models mentioned earlier.
- Model Tuning- use GridSearchCV to finetune the hyperparameters and optimize the model.

References

- http://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/
- https://www.google.com/publicdata/explore?ds=kf7tgg1uo9ude_&met_y=population&hl=en&dl=en
- https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/