Capstone Proposal

NYC Taxi Fare Prediction with Deep Learning

Domain Background:

Majority of residents of New York city rely on public transportation for commute. Only 22% of residents own vehicle as compared to national average of 91% Americans who at least own one vehicle. Taxi companies in New York city are fourth largest provider of public transportation providing over 1.2 million rides to city dwellers and visitors. Therefore, taxis are lifeline of transportation in New York city area than the rest of America.

Recently, it has been noticed that traditional taxi companies are struggling to stay in the business due to stiff competition from ride hailing apps. Accurate fare pricing could benefit the profitability of the taxi owners. Moreover, new ride hailing apps also need to calculate the price of the ride accurately in order to maintain the profitability for struggling taxi business in New York city. Ride hailing apps have 65% more trips than taxi trips. Ride hailing through apps (Lyft, Uber etc.) has soared to 15 million rides and taxi rides are fewer than 10 million. Ride hailing app popularity increased taxi usage mainly due to cheaper pricing and easy availability.

Motivation of this project is to familiarize with current mapping and pricing techniques such as openstreetmap for plotting data. Another part of the motivation includes use the machine learning techniques such as deep learning, for drawing meaningful results and make informed decisions from city data. This has a huge implications in near future as more and more cities are opening their datasets.

This is a very recent competition on Kaggle ended just 3-4 weeks ago. It offers very interesting information in terms geospatial information and price prediction challenges.

Estimating the fare just based on linear regression of distance between two points results in large errors, though it is simplest way to predict it. This is shown by the creator of the competition in this Kernel.

Problem Statement

Objective of the project is to predict the fare accurately, given taxi pick up and drop off locations, on a given date with given number of passengers. Root mean square error is calculated and serve as a metric to check the accuracy of the model.

Dataset and Inputs

In this project, dataset in the Kaggle competition, New York City Taxi Fare Prediction, used to predict the fare amount (including tolls) for the ride, given the pickup and drop off location. This dataset most likely originated from the NYC Taxi & Limousine Commision (TLC) dataset for yellow and green taxis.

This is a feature rich large dataset provided the following details and contains 55 million lines.

ID in dataset

- key - Unique string identifying each row in both the training and test sets. Comprised of pickup_datetime plus a unique integer.

Features in dataset

- 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 feature

- fare_amount - `float` dollar amount of the cost of the taxi ride. This value is present for the training dataset and predicted by the machine learning model using test dataset.

Solution Statement

Solution will include the building deep neural network using Keras and Tensorflow for this multidimensional data. Model takes cleaned data of date, time, pick up, drop off location and number of passenger data. Model then predicts the fare for these inputs with highest possible accuracy.

Benchmark Model

Multivariable regression model will serve as a benchmark model. It will provide a solution with simple approach towards fare prediction problem.

Evaluation Metric

Root mean squared (RMSE) error is used as the evaluation of the model. This is typically done for regression model. RMSE measures difference between prediction by the machine learning model and actual value.

Outline of Project Design

Following things will be performed with dataset to make final predictions of taxi fare with a deep learning model

1. Data cleaning to remove data points with
   a. Negative fares, very large fares, same start and end location, null or zero entries
2. Feature engineering and analysis
   a. Convert the location information of pickup and dropoff to haverisine distance
   b. Create timestamp information column to identify the day, month, year and hour of the ride to consider effect of patterns due to time as well as weather changes.
   c. Time of the ride also sets the peak hour and late night features for higher pricing which are finally hot encoded.
   d. Identify specific rides between airport and dropoff to Manhattan area for considering the fare rules of NY Taxi and Limousine Commission.
3. Data exploration
   a. Plot the new features against fare price to identify any patterns
4. Data splitting
   a. Dataset is huge so it is possible to split the dataset into training, validation and test data set.
5. Model development with deep learning
   a. Deep learning model will be built with Keras and Tensorflow.
   b. This is a large dataset so stochastic batch gradient descent with mean squared error as loss function will be most suitable for optimization of neural network.
   c. Since this is a regression, most probably less than 10 deep layers will be suitable.
6. Evaluation of the model
   a. Evaluation of the model will be done against the test dataset to find the RMSE.
   b. Speed and resources used for calculation.

References:

1. https://www.kaggle.com/c/new-york-city-taxi-fare-prediction
2. https://datasmart.ash.harvard.edu/news/article/case-study-new-york-city-taxis-596
3. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
4. http://www.aei.org/publication/chart-of-the-day-creative-destruction-the-uber-effect-and-the-slow-death-of-the-nyc-yellow-taxi/
5. https://datasmart.ash.harvard.edu/news/article/analytics-city-government