Machine Learning Engineer Nanodegree

Capstone Proposal 03/18/2019

New York City Taxi Fare Prediction

Domain Background

New York Yellow Taxis do not offer a flat rate other than to certain airports. Fare charges are based on a meter counting time and distance. The meters are generally very accurate.

Base charges:

- Base fare – $2.50
- New York State tax surcharge – $0.50
- Additional charge from 4 pm to 8 pm on weekdays – $1.00
- Additional charge from 8 pm to 6 am everyday – $0.50

Once the meter is running:

- When the taxi is moving, the charge is $0.50 per one-fifth of a mile with speed being 6 miles per hour or more
- If the taxi is sitting still or must slow down due to traffic, the charge is $0.50 for 2 minutes of time stopped or travelling below 6 miles per hour

Airport fares:

- Flat fares from JFK to Manhattan – $52
- Fare to Newark Airport – Metered fare in addition to extra $17.50 surcharge, and also tolls for a tunnel

Therefore, it is always safe to have a good prediction of the taxi fare in New York.

Problem Statement

Given a training set of 55 million Taxi trips in New York since 2009 in the training data and 9914 records in the testing data, the goal is to predict the fare of a taxi trip with the help of information such as pick-up and drop-off locations, the pick-up date and time, and number of passengers travelling.

Datasets and Inputs

File Descriptions

- **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) for which the **fare_amount** needs to be predicted

- **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.

Data fields

ID

- **key** – Unique string (comprised of **pickup_datetime** plus a unique integer) identifying each row in both the training and test sets. This is useful to simulate a submission file while doing cross-validation within the training set.

Features

Feature Name	Feature Description	Feature Data Type
pickup_datetime	Value indicating when the taxi ride started	timestamp
pickup_longitude	Longitude coordinate of where the taxi ride started	float
pickup_latitude	Latitude coordinate of where the taxi ride started	float
dropoff_longitude	Longitude coordinate of where the taxi ride ended	float
dropoff_latitude	Latitude coordinate of where the taxi ride ended	float
passenger_count	Number of passengers in the taxi ride	integer

Target

- **fare_amount** – float dollar amount of the cost of the taxi ride. This value is only in the training set and needs to be predicted for the test set.

Solution Statement

Analysis of the factors that will affect the cost of a taxi trip

1. **Trip distance** – The more is the distance to be travelled, the higher is the fare and vice-versa.

2. **Time of travel** – The fare may vary based on different hours during the day or based on peak hours.

3. **Days of the week** – Fare may vary based on whether we are travelling on weekdays or on weekends.

4. **Weather conditions** – The availability of taxis may vary based on weather conditions. So, the fare will vary based on the number of taxis available.

5. **Trip to/from airport** – Trips to or from the airport are generally fixed.

6. **Pick-up or drop-off neighbourhood** – Fare may vary based on the location

7. **Availability of taxi** – Higher is the number of taxis available, lower will be the price and vice-versa.

Exploratory Data Analysis

1. **Distribution of fare amount**

2. **Distribution of Geographical Features** –
    a. Analysis of fare amount based on latitudes and longitudes

    b. Analysis of fare amount based on whether pick-up or drop-off location is related to airports in New York

    c. Variation of fare amount based on neighbourhood.

    d. With some research, we came to know that New York City encompasses five county-level administrative divisions called boroughs: The Bronx, Brooklyn, Manhattan, Queens, and Staten Island. So, further analysis can be done based on whether the pick-up or drop-off location is in any of the regions mentioned above.

3. **Distribution of Trip Distance** –
    a. Analysis of fare amount based on trip distance. For the calculation of trip distance based on pick-up coordinates and drop-off coordinates, we will use the **Haversine Distance** formula.

Haversine formula determines the ***great-circle distance*** between two points on a sphere given their longitudes and latitudes.

4. **Distribution of pick-up date and time** –
    a. Analysis of fare amount based on date, time of the day, day of the week, weekdays, weekends, months, years

5. **Distribution of passenger count** –
    a. Analysis of fare amount based on number of passengers.

    b. Further analysis can be done from number of passengers along with pick-up date and time

Models to be evaluated

1. **Linear Regression** (as base model)
2. **Random Forest**
3. **LightGBM** (boosting tree-based algorithm)

Benchmark Model

For this problem, we will use a **Linear Regression** model as the benchmark model.

Evaluation Metrics

We will evaluate the proposed models based on Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Project Design

The project will be implemented in three steps:

1. Exploratory Data Analysis
2. Data pre-processing
3. Cleaning up data
4. Implementation of base model
5. Implementation of proposed models
6. Evaluation of proposed models based on proposed evaluation metrics