Machine Learning Capstone

New York City Taxi Trip Duration

Domain Background: This is a Kaggle competition. It is based on the taxi service in the New York city. The competition dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this playground competition. Based on individual trip attributes, participants should predict the duration of each trip in the test set. Prediction of trip duration is very important they can provide insights on traffic pattens, road blockage, With ride sharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to their estimated fare and ride duration, since the competing apps provide these metrics upfront. This is academic paper where machine learning was applied to this type of problem.

Ref: http://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJuniorNewYorkCityCabPricing-report.pdf

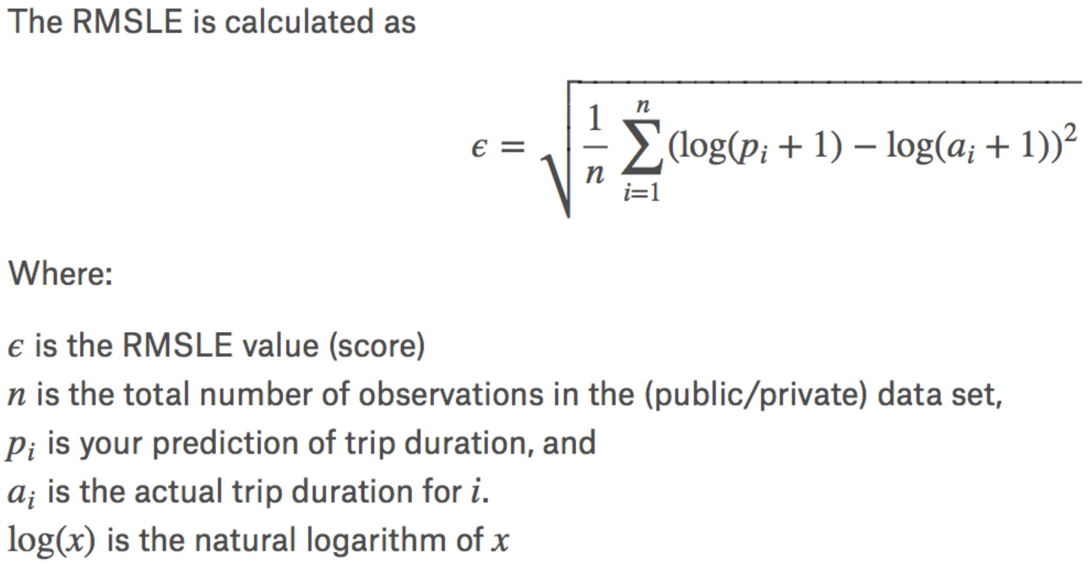
Problem Statement: The competition is about regression building a model to predict the total trip duration based on given data about past taxi trip duration data. Explore the given data and predict which features are effecting the trip duration most so that they can be used in real cases every day. I treat this as a regression problem.

Datasets and inputs: The dataset for the competition can be found in kaggle platform. The features that are present in the data set are *id, vendor_id, pickup_datetime, dropoff_datetime, passenger_count, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, store_and_fwd_flag, trip_duration*. Here *trip_duration* feature is the target variable. These are very important features one can guess for predicting the trip duration, but there are more important features that are absent like traffic flow density variation with time and route, weather data … which can be helpful in making accurate predictions. So more data can be collected or created and added to the given data if the aim is to generate the best model. There are 1458644 train examples with 11 features in the training data I want to split and use this training data for training and cross validation purposes. *Ref for data:* **https://www.kaggle.com/c/nyc-taxi-trip-duration/data**.

Solution Statement: The solution I chose for this problem is to create an XGboost model to predict the trip duration using the data given and then fine tune the model to get a better model. The model can be applicable to the data set because the data suitable for it we can Encode any categorical variables and use them in model. Any metric useful for regression task can be used like RMSE - Root mean squared error, RMLSE - Root mean squared logarithmic error .. etc. By using the trained XGB model we can produce predictions for any number of future cases. I want to use Principle component analysis and clustering techniques later in order to create an improved model.

Benchmark Model: There are many great kernels in kaggle platform for this competition which can serve as for a evaluation model. I want to use this very good kaggle kernel by Beluga as evaluation or reference: **https://www.kaggle.com/gaborfodor/from-eda-to-the-top-lb-0-367**, The model presented is very sophisticated and is on top of the leader board. This model helps me in evaluating my performance to check what features need to be added to the data to improve the performance how to better scale feature and tune model. The kernel also helps me figure out where I my model is going wrong So I want to use this kernel as my reference.

Evaluation Metric: As suggested by kaggle I want to use RMLSE as error metric it is a vey simple evaluation metric and works well even if there are very small and large values for target variable. This metric suits well for this competition



The RMSLE is calculated as

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

Where:

ϵ is the RMSLE value (score)
n is the total number of observations in the (public/private) data set,
p_i is your prediction of trip duration, and
a_i is the actual trip duration for i.
$\log(x)$ is the natural logarithm of x

Project Design:

- **Loading data:** I will start by loading data into a pandas data frame.

- **Handling missing values:** Find if there are any missing values for each feature and replacing them with appropriate values.
- **Observe data:** Next I will observe data statics and related summary.
- **Curating data variables into useful formats:** Features like pickup_datetime need to be split in to useful and interpretable forms by adding day, month, year, hour features to data.
- **Visualising data:** Plot some visualisations to see any outliers, trends and patterns in data. Plots include box plots to find outliers, scatter plots to find patterns in data.
- **Removing outliers:** Removing extreme outliers so that the model wont get effected of them.
- **Adding relevant features for the problem:** Adding features like distance is useful in this context because trip duration generally depends on distance also.
- **Encoding Categorical Variables:** Because models can only interpret numerical data.
- **Building Models:** I chose to build a liner regression model and an XGBoost model for this problem.
- **Tune Parameters:** There are many parameters to tune for XGBoost model, so I want to use grid_search_cv to tune those parameters.
- **Comparing the results with Benchmark model:** Finally I want to compare the performance of my model with bench mark model and do necessary modifications to increase performance of my model.