Machine Learning Engineer Nanodegree
New York City Taxi Trip Duration

Cosa Santos

June 23, 2019

1 Domain Background

The chosen project for the Capstone Project is the New York City Taxi Duration Trip competition from Kaggle. The challenge is to build a model that predicts the total ride duration of taxi trips in New York City.

1.0.1 2. Problem Statement

By considering historical information such as pick up and drop off date, hour and geo-localization, and trip duration regarding taxi trips, the objective is to predict the duration of each taxi trip given in a specific test set, which does not present neither drop off time, nor trip duration. The prediction will be performed through a supervised learning regressor to be defined given the data structure.

1.0.2 3. Datasets and Inputs

The challenge provides two data sets. * train.csv - the training set, which contains 1.458.644 observations * test.csv - the testing set, which contains 625.134 observations

The training set contains eleven features:

- `id` - a unique identifier for each trip
- `vendor_id` - a code indicating the provider associated with the trip record
- `pickup_datetime` - date and time when the meter was engaged
- `dropoff_datetime` - date and time when the meter was disengaged
- `passenger_count` - the number of passengers in the vehicle (driver entered value)
- `pickup_longitude` - the longitude where the meter was engaged
- `pickup_latitude` - the latitude where the meter was engaged
- `dropoff_longitude` - the longitude where the meter was disengaged
- `dropoff_latitude` - the latitude where the meter was disengaged
- `store_and_fwd_flag` - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- `trip_duration` - duration of the trip in seconds

The testing set doesn't have neither the `trip_duration`, target variable, nor the `dropoff_datetime`. So,the model should be trained on the training set, and predict the trip_duration for the testing set observations. The predictions should be submitted to the Kaggle's challenge for calculating the final score.

The variables pickup_latitude, pickup_longitude, dropoff_latitude and dropoff_longitude will be used to calculate the trip distance between the pickup and dropoff locations. Also, they are important and will be analysed because where in NY the each trip starts and ends influences greatly the velocity in which the taxi can travel.

The variables pickup_datetime are important, and also will be analysed profoundly, because the moment on time in which the trips occur also influences greatly the the taxi velocity.

The remaining variables will be analysed in order to conclude about their influence on the target variable.

The data sets provided by the Kaggle challenge were made avaiable in Big Query on Google Cloud Platform, originally published by NYC Taxi and Limousine Commission (TLC).

2 Solution Statement

The solution will be provided by a regression method. The optimization will be performed after a preprocessing of the data and through minimization of the root mean square logarithm error (RMSLE), cross validation for detecting over and/or underfitting, and grid-search for fine tunning hyperparameters.

3 Benchmark Result

The benchmark result is the winning score of the Kaggle's challenge when the winning team submitted the provided testing set. The challenge's metric is the RMSLE, and the winning score is 0.28976.

4 Evaluation Metrics

The evaluation metric is the one from the Kaggle's challenge, RMSLE, in order to make the score comparable to the benchmark result. The metric is quite similar to the root mean squared error, however, considers the squared difference of the logarithm values of prediction ($\hat{y}$) and true values(y).

$$\text{RMSLE}(y, \hat{y}) = \sqrt{\frac{1}{N}\sum_k \log(1 + y_k) - \log(1 + \hat{y}_k)}$$

5 Project Design

The project will be segregated in four major parts: data preprocessing, model selection, benchmark comparison and conclusion.

5.1 Data Preprocessing

Data preprocessing will be segregated into two sections: Generals and Specifics.

5.1.1 Generals

In this section, data quality will be considered and performed.

- Are there any missing values?
- The features are formatted on a convenient way?
- Is there any unnecessary information?

Along with that, from the features `pickup_latitude`, `pickup_longitude`, `dropoff_latitude` and `dropoff_longitude`, the feature `distance` will be calculated.

5.1.2 Specifics

In this section, the data will be explored statistically and visually, with the objective of understanding the features influences, detecting and dropping outliers and finding transformations that will make the training and prediction better, as rescalling or *one-hot-encoding*, for example. This analysis will be divided in four subsections: Distance anaysis, Geo-localization analysis, Time related analysis and the remaining features analysis.

For the analysis, the velocity will be calculated, `distance` over `trip_duration` as the trip duration may depend on traffic, that depends greatly on date and time.

5.2 Model Selection

This parte will train the models and consider the RMSLE scoring for choosing the best one. As the testing set provided doesn't have the target feature, `trip_duration`, the training set will be splitted into training and testing set for the model selection.

5.2.1 Initial Model Evaluation

For the initial model evaluation, a set of different kind regressors will be trained in different data set sizes, using cross validation, in order to verify underfitting and overfitting between regressors. No tunning of hyperparameters will be performed.

5.2.2 Model Tunning

The best model from the previous step will have its hyperparameters tunned through gridsearch in order to select the best hyperparameters.

5.2.3 Final Validation

Once chosen the best model, it must pass through a set of final validations in order to check its robustness.

5.3 Benchmark Comparison

Once the model has been selected and trained on the whole provided training set, the predictions over the provided testint set will be submitted to the private leaderboard on Kaggle. The result will be comparised with the bechmark, which is `0.28976`.

5.4 Conclusion

The last section it is the conclusion with a reflextion about the data, the model and a proposal possible improvements.