Machine Learning Engineer Nanodegree

Capstone Proposal

Wenzhe(Emma) Ding

September 12th, 2017

Proposal

Domain Background

There are currently over 13,000 licensed taxicabs and over 50,000 taxicab drivers providing transportation for passengers in New York City via street hails(http://www.nyc.gov/html/tlc/html/industry/yellow_taxi.shtml). The New York City Taxi & Limousine Commission (TLC) has released data with detailed information of each taxi trip from January 2016 through July 2016. To understand trip patterns and answer more specific questions such as what is the rush hours of NYC taxi trips, what features are relevant to predict taxi trip duration, what are most popular taxi regions at different time of the day and etc. we need to investigate such data sets. In fact, many data scientists have done analysis using such public taxi data, some are listed below.

- http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/ used an older NYC taxi data (January 2009 to June 2015) to analyze travelling patterns. It also presented findings of how Uber has changed the landscape for taxis.
- https://www.kaggle.com/headsortails/nyc-taxi-eda-update-the-fast-the-curious aimed at predicting the duration of taxi rides in NYC. It integrated NYC weather data and fastest routes data to study their impact on the taxi trip duration.
- https://www.kaggle.com/gaborfodor/from-eda-to-the-top-lb-0-367 focused on the feature extraction and aimed at finding best possible feature set for grandient boost tree algorithm.

In addition to these efforts, reasearchers have done extensive study to learn transportation mode using machine learning techniques. A few research papers are showned below:

- http://ieeexplore.ieee.org/document/7063936/?reload=true compared the performance of several methods including K-nearest neighbor, support vector machines, and tree-based models that comprise a single decision tree, bagging, and random forest methods.
- http://ieeexplore.ieee.org/abstract/document/7366148/ applied principal component analysis and semi-supervised Gaussian mixture models to classify different transportation modes.
- a deep learning method is used to identify different transportation modes of smartphone users in http://ieeexplore.ieee.org/abstract/document/8006227/.

Furthermore, being a transportation engineering major, I am especially interested in studying people's travelling behavior by investigating public transportation data.

Problem Statement

The New York City Taxi & Limousine Commission hosts a kaggle competition (https://www.kaggle.com/c/nyc-taxi-trip-duration) to challenge data scientists to build a model that predicts the total ride duration of taxi trips in New York City, which also serves as the major goal of this project. Available features in the data include pickup time, geo-coordinates, number of passengers, and several others. Root Mean Squared Logarithmic Error is used to compare the performances among models.

Datasets and Inputs

Both the train data and test data are provided by the competition, which is also under the folder "data" in this repo and title as "train.csv" and "test.csv" respectively. The time range of both data set is from January 2016 to July 2016. The training set contains 1458644 trip records and the testing set contains 625134 trip records. Detailed explanation of data fields are shown below, which is also available at the "Data" section of the competition (https://www.kaggle.com/c/nyc-taxi-trip-duration/data).

- 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

Basically, the data points from "train.csv" would be split into 80% training set and 20% validation set and we will use only training set for model training. Validation set will be used for parameter selection and to avoid overfitting of the model being built. Because there is no labels on test data, we will upload the model predictions to the kaggle competition and obtain the scores for the test data from the LeaderBoard (https://www.kaggle.com/c/nyc-taxi-trip-duration/leaderboard).

Solution Statement

The first step is exploratory data analysis (EDA) and both univariate and bivariate analysis will be conduncted to study data features. Possible techniques include histogram plot, time series analysis, etc. Understanding the difference between train and test set is also necessary becasue if there is huge differenc between them then feature extracted from train set would not apply to test set.

There are many possible methods to predict trip duration and one of them is gradient boosted trees regressor using XGBoost (a package in python http://xgboost.readthedocs.io/en/latest/), which is one of ensemble methods to deal with supervised learning problems. The main principle behind this method is that a group of “weak learners” can come together to form a “strong learner”. They typically less prone to overfitting and make the model more robust,unlikely to be influenced by small changes in the training data. The input of this algorithm is a set of numerical features and the output is a number, which is also the prediction of the algorithm. We will implement two methods other than gradient boosted trees regressor, which are decision tree regression and linear regression to compare the performances among them.

Benchmark Model

A benchmark model is decision tree regression model, which is widely used for supervised learning problems. It uses a tree like structure to specify a series of conditions that are tested to determine the value for a sample. We will obtain the best-performed decision tree using grid search.

Evaluation Metrics

The evaluation metric for this project is Root Mean Squared Logarithmic Error (RMSLE), which is defined here (https://www.kaggle.com/wiki/RootMeanSquaredLogarithmicError). RMSLE penalizes an under-predicted estimate greater than an over-predicted estimate.

Project Design

Firstly, we will conduct comprehensive EDA before building models for prediction as the insights can be valuable for both our model building as well as the community. A few analysis of the data needs to be done in this process:

- check whether there are missing values in the features of interest
- remove outliers: we could use z-scores, box plots, scatter plots to screen outliers. We will try using inter quartile range (IQR) to detect outliers. If this is not a profer method, we could use linear regression to fit to a data set, then rank the errors, and throw out the top 10% and refit iteratively as a test for stability.
- feature engineering: after EDA, we could extract useful features that have impact on the target variable and these features could include in model training as well.

Afterwards, we will implement algorithms which will includes split the training data into train-validation set and train XGBregressor. We could obtain feature importance from the results of XGBregressor, which can use as a measurement to filter out useless features. Finally, we'll compared the performance of the best performed decision tree model (obtained by grid search) and gradient boosted tree model. Since there's no labels in the test set, performance will be compared based by the rank on the kaggle competition LeaderBoard.
