**CSYE 7245-Big Data Systems and Intelligence Analysis**

**Analysis & Predict Duration of taxi trip  
Based on Social Streaming & Geospatial data**

Ziwei Fan 001855517

Feb.7th 2017

1. **Abstract:**

The project is to:

(1) Predict the total ride duration of taxi trips in New York City and animate request based on time slot.

(2) Explore how weather data impacts the traffic with csv data & social streaming.

Background:

1. The primary data set is released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. Based on geo-information, the geospatial animation will be visualized by heat map.
2. Temperature, snow-fall or precipitation are quite different monthly, which are more likely to cause influence on traffic: The rainy day, it is more likely a traffic jam and accidents will occurs in downtown area, as well as in snowy day. At the meantime, with the consideration of climate and weather condition, the accident rate may go up.

Hence, in this thesis, the potential association between weather, climate, and the traffic will be explored by weather data, geospatial data and streaming analysis. Besides, according to the analysis, it is more practical to predict for drivers and pedestrian to pay more attention in some areas.

For summary, the several perspectives as below will be described:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Goal** | | | **Realistic scenario** | | |
| **Application** | How practically useful and applicable | **Analysis** | **1.Trip Duration** | | |
| 1. Data Preprocessing 2. Trip duration overview 3. Distribution analysis | 1. Request Animation based on Geospatial data | |
| **2.Weather-Traffic** | | |
| 1. Principle Component Analysis(PCA) 2. Pipeline choosing(TPOT) | | G. Media  Streaming Analysis |
| **Visualization** | Trip Duration Geospatial map  Geospatial Animation on Trip Request | | |
| **Prediction** | **Trip Duration** | | |
| **TPOT**  XGBRegression | | |
| **Measures of validation** | **Score or Comparison** | |  | | |
| **speed** | |  | | |

1. **Introduction:**
2. **Trip duration:**

Based on historical records of cities, it is more likely to visualize the trend and heat map and predict the trip duration in the future. In this segment, the approach or tool I used is:

(1) Distribution analysis on latitude and longitude.

(2) Visualization based on heat map;

(3) Prediction with TPOT (A Python Automated Machine Learning tool that optimizes machine learning pipelines)

(4) Comparison with another method in supervised learning.

1. **Weather effects on Traffic:**

Tremendous range of weather/climate condition may affect the traffic, according to the ‘Road Weather Management program’ [1].

In term of precipitation, the road friction and visibility distance may go down with the degree of wetness of the pavement: Less pavement friction goes with the higher wetness, and lower visibility distance goes with the rainfall. Accordingly, it is more likely that these aspects will affect the vehicle performance or the driver capabilities. Crash risk and the rate of accidents may change according to the chain reaction.

Apart from the references of conditions analytics, what is more is it was recorded by the specific ‘weather-related crash statistics’: In the snow/sleet day, 210,341 crashed in annual average. 4% of vehicle crashed and 739 person killed. [1]

Based on the potential risk from weather condition, it is significant to **explore the association among the traffic status, accident risk and weather condition**:

1. **Analysis Association between accident rate & Weather condition;**
2. **Media Analysis with social streaming;**

Firstly, the raw data could be collected from different departments and be pre-processed, such as weather data from National Weather Service and 2 data sheet from traffic department. Moreover, the text information will be gathered from media, and filter the keyword and analysis based on streaming.

In the next step, according to the result generated by previous steps, prediction could be conducted based on the foundation of algorithms and the result will be greatly helpful to remind the managers of these departments to radiate out the alert to drivers and pedestrians.

1. **Code with Documentation:**
2. **Data Resources[2][3][4]:**

* Weather data in New York City – 2016：<https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016/data>

Description:It contains the weather data collected by the weather station in Central Park between Jan.-Jun. 2016: Min\_Temperature, Max\_Temperature, Average\_Temperature, Precipitation, Snow\_fall, snow\_depth.

* Taxi Trip Duration：

<https://www.kaggle.com/c/nyc-taxi-trip-duration/data>

Description: The data includes individual trip attributes: pickup\_datetime, dropoff\_datetime, pickup\_lat-long etc.

* New York City Taxi with OSRM：

<https://www.kaggle.com/oscarleo/new-york-city-taxi-with-osrm/data>

Description: The data extracted by OSRM involves fastest routes information for each data point: Starting\_Street, End\_Street, total\_distance, total\_TravelTime etc.

The different attributions of 3 data sources will be integrated for analysis more than 1 data source according to specific requirement.

1. **Algorithms:**
   1. **Phase of Analytics:**

* Pearson Correlation Coefficient：

Pearson Correlation Coefficient in this research is to explore the association between different attributes from traffic data and weather data.

[5] Pearson Product-Moment Correlation <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>

‘Pearson Correlation Coefficient is a measure of strength of a linear association between 2 variables and is denoted by r ranged from 0-1: 0 indicates there is on association between the 2 variables. A value greater than 0 indicates a positive association. A value less than 0 indicates a negative association. Basically, it attempts to draw a line of best fit through the data of 2 variables.’

* Visualization with heat map plus image processing-Folium:

The visualized data will be merged with the real map and reflect on the specific map. Each point in this image should be considered as a point represented by lat-long for visualization. Folium is map-based interactive package and it's open source as well. It's based on leaflet js. It enables both the binding of data to a map for choropleth visualizations as well as passing Vincent/Vega visualizations as markers on the map.[13]

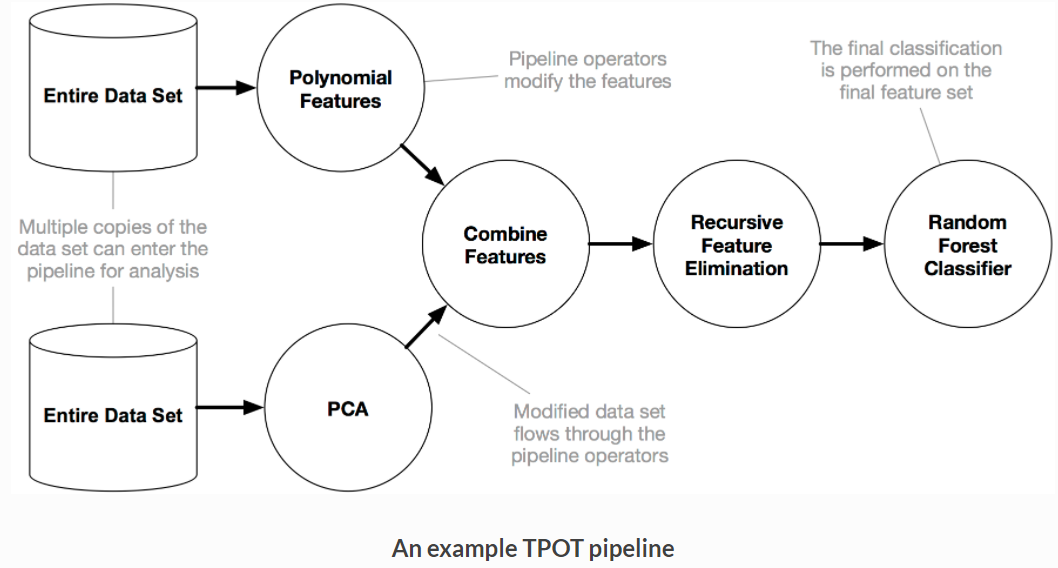
The library has a number of built-in tilesets from OpenStreetMap, Mapbox, and Stamen, and supports custom tilesets with Mapbox or Cloudmade API keys. Folium supports both GeoJSON and TopoJSON overlays, as well as the binding of data to those overlays to create choropleth maps with color-brewer color schemes.

* Principle Component Analysis(PCA):

Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. [12]For a lot of machine learning applications it helps to be able to visualize data. Visualizing 2 or 3 dimensional data is not that challenging. However, even the dataset used in some part is 4 dimensional. PCA can reduce that 4 dimensional data into 2 or 3 dimensions so that you can plot and hopefully understand the data better.

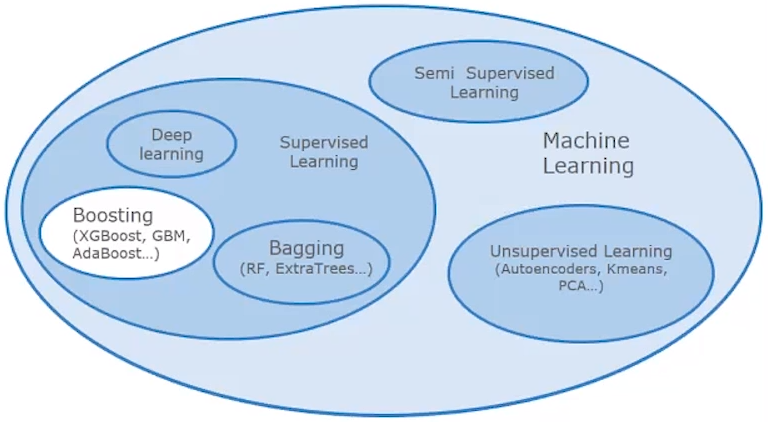
* 1. **Phase of Prediction:**
* TPOT:

TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming. It will automate the most tedious part of machine learning by intelligently exploring thousands of possible pipelines to find the best one for your data. Once TPOT is finished searching (or you get tired of waiting), it provides you with the Python code for the best pipeline it found so you can tinker with the pipeline from there. [6][12][14]



* XGBoost:

An XGBoost (Extreme Gradient Boosting) is an ensemble model that aggregates several trees to provide a more generalizable Machine Learning model. In this program, it might be generated as one of the results of TPOT.



* Algorithm of semantics analysis:

When it comes to Real Time Data Analytics, Spark Streaming provides a single platform to ingest data for fast and live processing in Apache Spark.

* 1. **Validation:**
* Kfoldcrossval Mean Score:

[8] What Is K-Fold Cross Validation? <https://magoosh.com/data-science/2017/12/08/k-fold-cross-validation/>

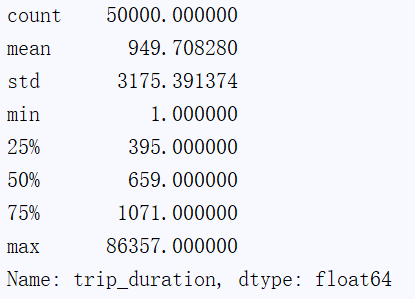
‘In K-Fold Cross Validation, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.’

* Or: Accuracy Score.

[9] 3.3. Model evaluation: quantifying the quality of predictions <http://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score>

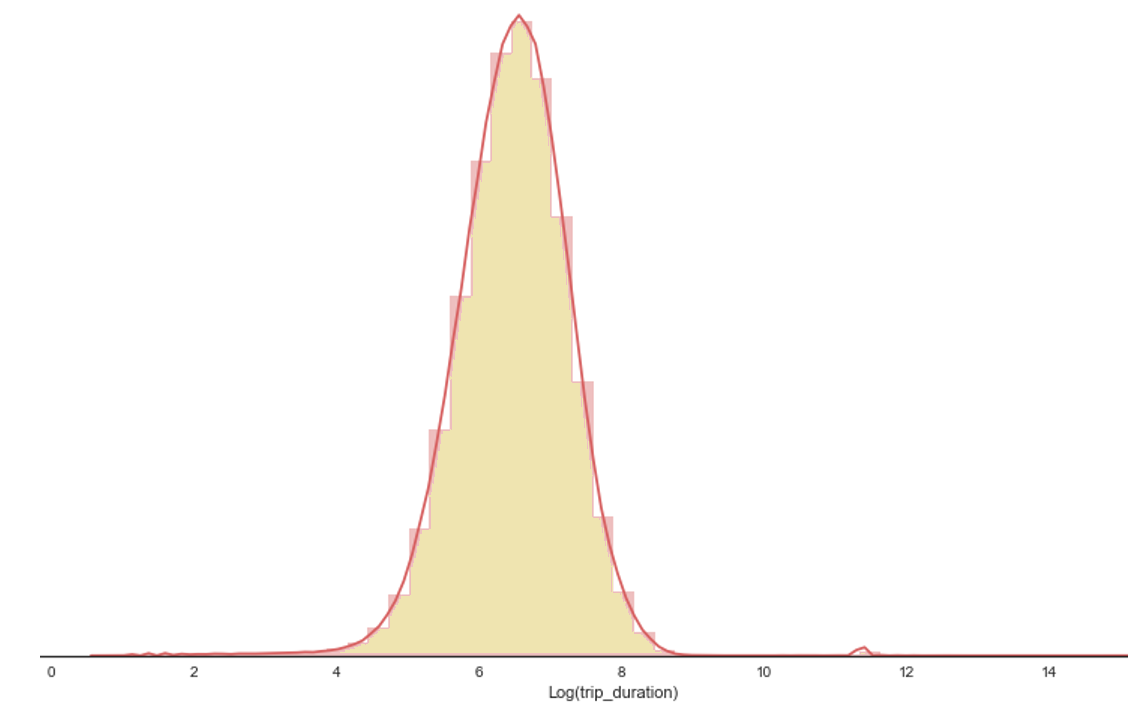
‘Accuracy Score computes the accuracy, either the fraction (default) or the count (normalize=false) of correct predictions. ’

1. **Result & Analytics:**
2. Data Analytics:
   1. Trip duration basic analytics:

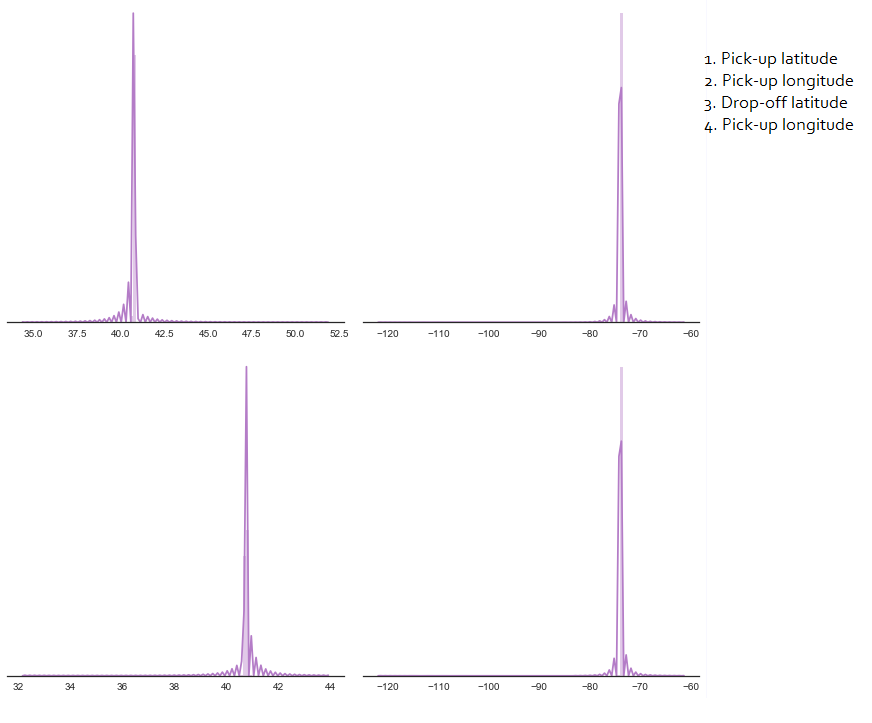


* 1. Distribution analysis:

1. Time based:

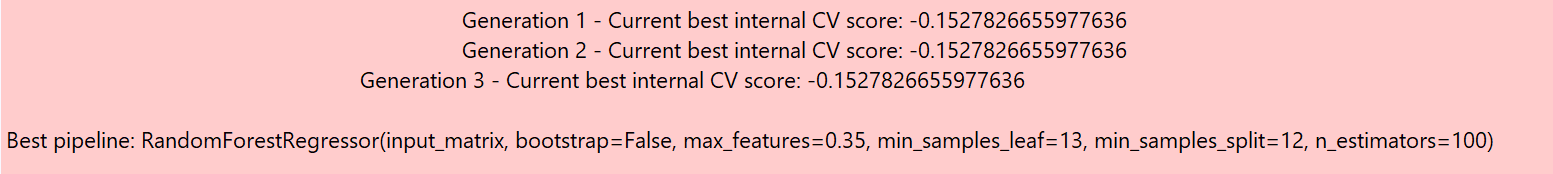


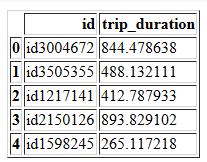
1. Longitude and latitude based:



* 1. Animation on map:
  2. Exploration on trip duration prediction:

TPOT🡪RandomForestRegressor is best!





* 1. Exploration on Weather:

This section will focus on analyzing how these new features behave. The correlation between weather and traffic will be plotted as ‘correlation heatmap. The features like temperature, precipitation etc.

* 1. Streaming analysis detection:

1. **Discussion:**

Q1. How to merge the different cleaned datasets into one dataset without Re-output & Re-read?

Q2. How to choose the best model to train the dataset?

Q3. How to evaluate better?

1. **Reference:**

[1] How Do Weather Events Impact Roads? <https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm>

[2] New York City Taxi Trip Duration: <https://www.kaggle.com/c/nyc-taxi-trip-duration/data>

[3] New York City Taxi with OSRM: <https://www.kaggle.com/oscarleo/new-york-city-taxi-with-osrm>

[4] Weather data in New York City – 2016: <https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016>

[5] Pearson Product-Moment Correlation: <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>

[6] TPOT: http://epistasislab.github.io/tpot/

[7] What Is K-Fold Cross Validation? <https://magoosh.com/data-science/2017/12/08/k-fold-cross-validation/>

[8] 3.3. Model evaluation: quantifying the quality of predictions: <http://scikitlearn.org/stable/modules/model_evaluation.html#accuracy-score>

[9] XGBoost Guidance:

<https://xgboost.readthedocs.io/en/latest/build.html#building-on-windows>

[10]Clustering-Mini Batches K-means:

<http://scikit-learn.org/stable/modules/clustering.html#mini-batch-kmeans>

[11] PCA for Data Visualization:

<https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60>

[12] TPOT: A Python tool for automating data science <http://www.randalolson.com/2016/05/08/tpot-a-python-tool-for-automating-data-science/>

[13] Folium Documentation: <https://folium.readthedocs.io/en/latest/>

[14] TPOT: <http://epistasislab.github.io/tpot/>