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

**Prediction on Traffic Situation based on Synthetic Analysis**

Ziwei Fan 001855517

fan.ziw@husky.neu.edu

**Abstract**

The time spent on road primarily depends on multiple perspectives: Regions, weather, accidents etc. For taxi or Uber driver, **how to avoid traffic congestion and achieve benefit maximization?** In details:

1. How could the drivers keep track of the traffic situation by visualization?
2. How could drivers acquire the distribution information about requests number?

For these questions, the focus of paper is to:

1. **Visualize the traffic and requests number by geospatial information.**
2. **Explore how weather data impacts the traffic with data & explore if it could be validated by sentiment analysis.**
3. **Predict the total ride duration of taxi trips in New York based on TPOT.**

Weather data, geospatial data, and traffic data, have been utilized to be visualized according to the different requirements. For prediction in this case, the best pipeline generated by TPOT is DecisionTreeRegressor. Additionally, the correlation between traffic and weather is different from the common sense. It is just a weak negative correlation -0.385 by sentiment analysis and similar with the result of Pearson Correlation.

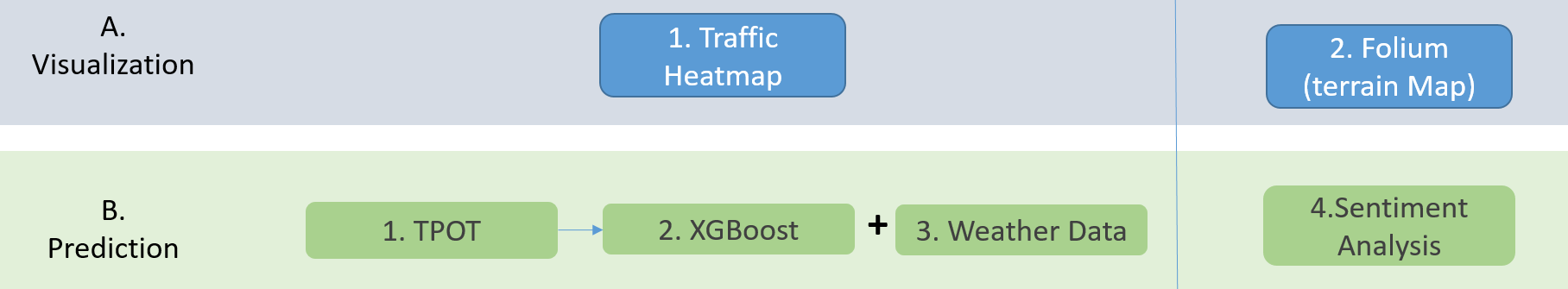
**Introduction**

Longer trip duration and traffic congestion are terrible issues for traffic system and taxi drivers. For leveraging the traffic resources and avoid traffic jam, the synthetic traffic monitoring and prediction are necessary.

For this goal, diverse sorts of data in the system should be considered:

1. In terms of weather, it is shown that tremendous range of weather/climate condition may (not must) affect the traffic, according to the ‘Road Weather Management program’. [1]
2. For drivers, the heatmap of hour changing will directly reflect the hot areas of traffic. Additionally, the prediction on duration trip will be more accurate to get to know what the hot area will be and show the hot areas predicted for the drivers.
3. Besides, they could keep track which regions have the most number of requests and they could receive the traffic information from operation center, so that they could balance and make a decision to achieve benefit maximization.

In the implementation, the basic architecture is as below:



**Data source**

1. [**New York City Taxi Trip Duration**](https://www.kaggle.com/c/nyc-taxi-trip-duration)**:**

 The data was originally published by the NYC Taxi and Limousine Commission (TLC)[6].It contains pickup time, geo-coordinates, number of passengers, and several other variables.

1. **Taxi with OSRM:**

OSRM(Open source routing machine)is a Modern C++ routing engine for shortest paths in road networks. Get in Touch. Features. Flexible import of OpenStreetMap data.[7] This dataset is used to extract information about the fastest routes for each data point.

It contains data about Starting & End Street, total distance, number of steps, distance of per step etc.

1. **New York Weather Data:**

Weather data collected from the National Weather Service. It contains the first six months of 2016, for a weather station in central park. It contains for each day the minimum temperature, maximum temperature, average temperature, precipitation, new snow fall, and current snow depth.

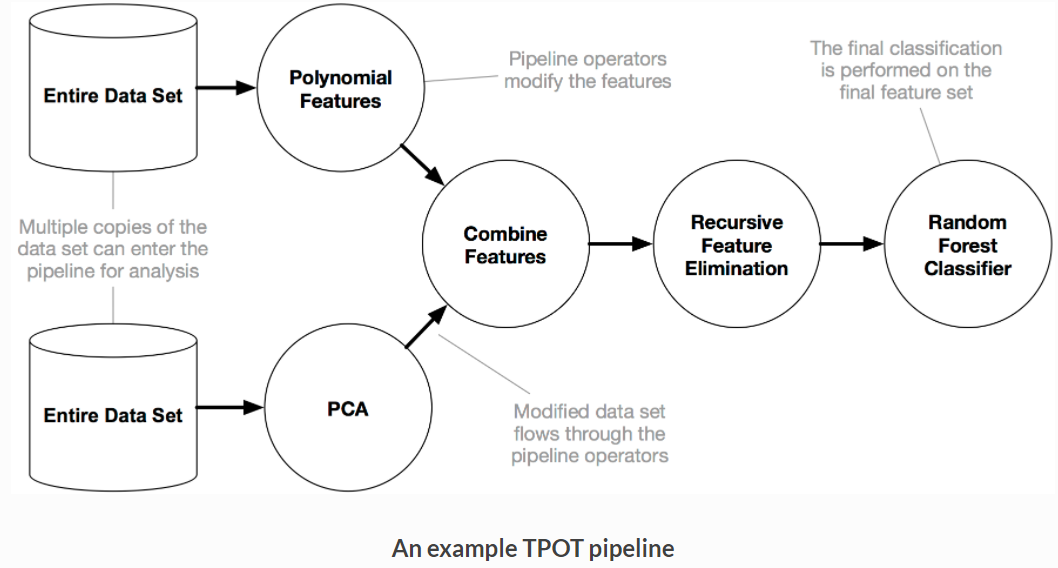
**Method**

Prior to the training, some work for pro-pressing has been completed. 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. The approaches will be explained in more details:

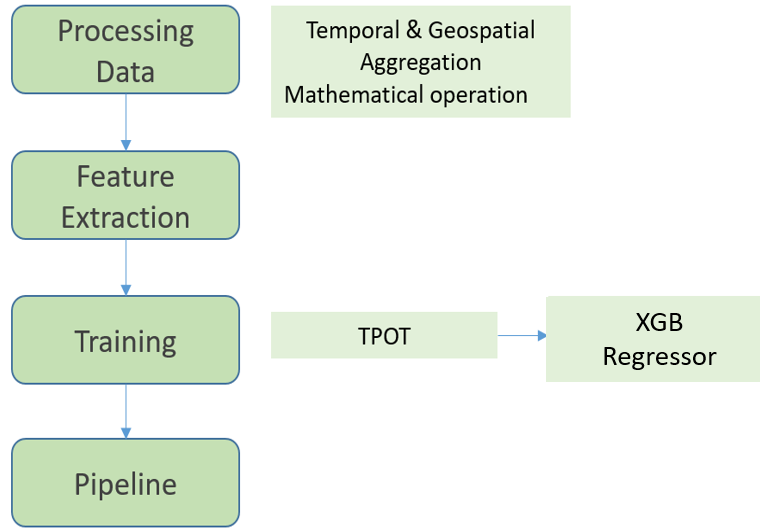
1. **TPOT:**

TPOT package is used to handle the cross validation and hyper-parameters and predict the trip duration. [3][4]

It 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 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.

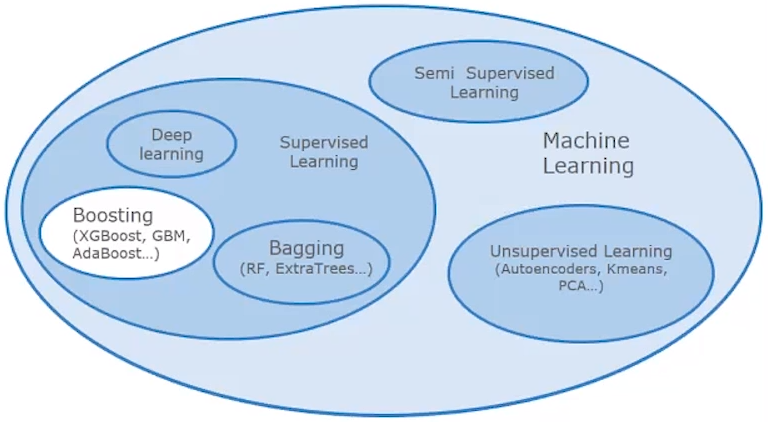


In this case, I will apply TPOT by this progress:

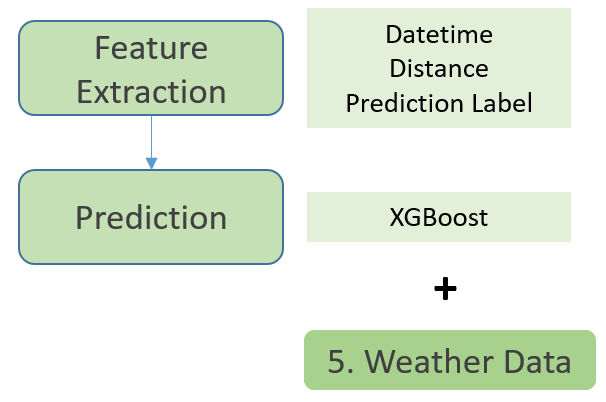


1. **XBGoost:**

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. However, in this case, another model-GradientBoostingRegressor has been generated after TPOT processing. Thus, XGBoost could be applied for further validation or mixed with weather data for exploring correlation.



In order to further predict with weather data, and validate the result generated by TPOT.

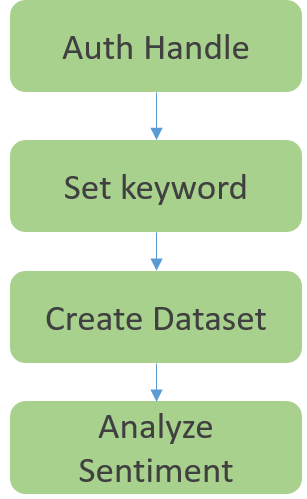


1. **Sentiment analysis:**

Sentiment Analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker. [2] With the different keyword, it is likely that we could explore the attitude correlation between traffic and weather.

As one of the approaches to validate the correlation explored previously, sentiment analysis has been applied based on twitters. In this validation, the attitudes towards weather and traffic have been checked routinely and compared with the correlation results calculated previously: If the results are similar.

It will focus on the twitter dataframe and explore if there is any correlation between weather and traffic. This is an approach to validate the Pearson correlation calculated by algorithm.

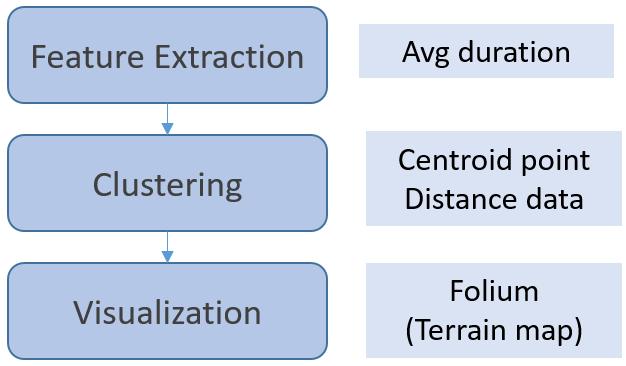


1. **Visualization by heatmap & Folium:**

For real terrain visualization, Folium is map-based interactive package and it is open source as well. It’s based on leaflet and enables both the binding of data to a map for choropleth visualizations, as well as passing Vincent/Vega visualizations as markers on the map.[5]

In the driving, the visualized and colorful map with request information could navigate the driver to the region with more request.

I will apply it into scenario of the request number: Folium is applied for generating terrain map in order to reflect the number of request more directly.

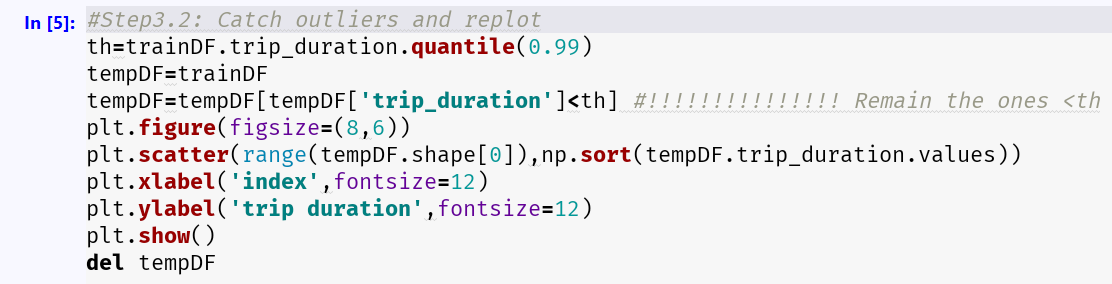


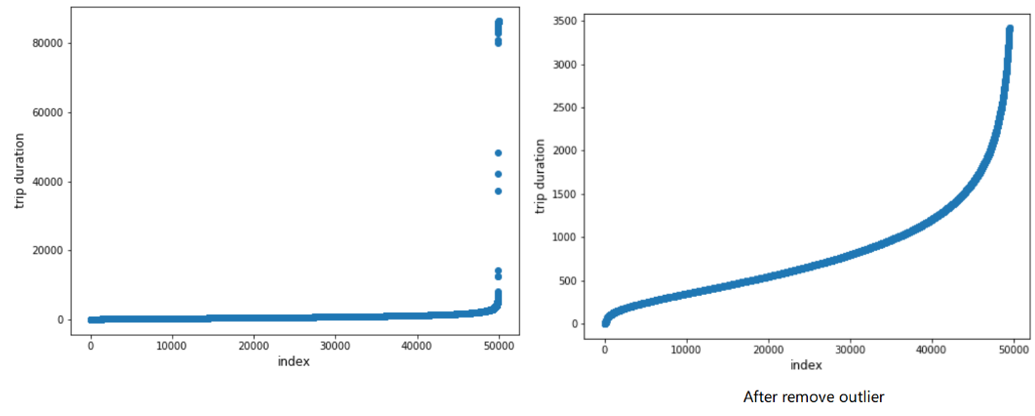
**Result and validation**

Based on the methods mentioned briefly, the visualization could be presented directly according to the data frame after feature extraction and model training. Additionally, a model has been generated by TPOT. The prediction on trip duration has been conducted with RMSE. In details:

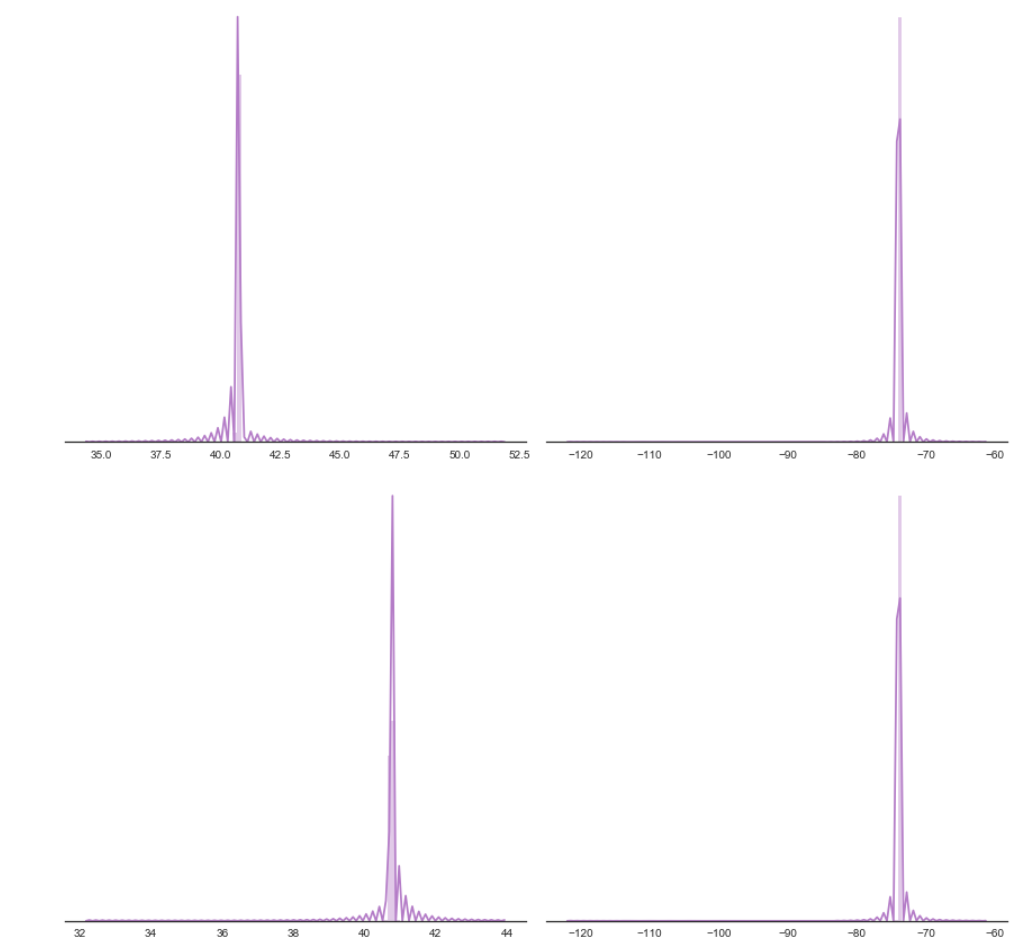
1. **Pro-processing and check the data:**

The new plot has been smoothened after removing the null data, outliers. In this case, we just remain the ones within the range of threshold (0.99).



****

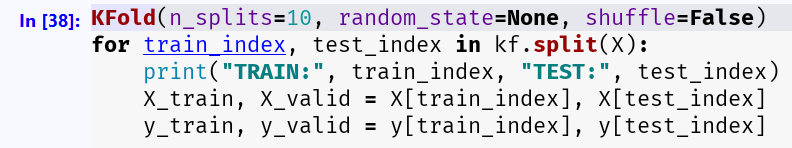
Then, let us check the lat-long distributions are then used them to have a heat map kind of view of given lat-longs in next steps. It is obvious that we get the findings that most duration centered around these ranges: Latitude should be between 40.6 to 40.9 and Longitude should be between -74.05 to -73.70. Then we will focus on these range more closely.

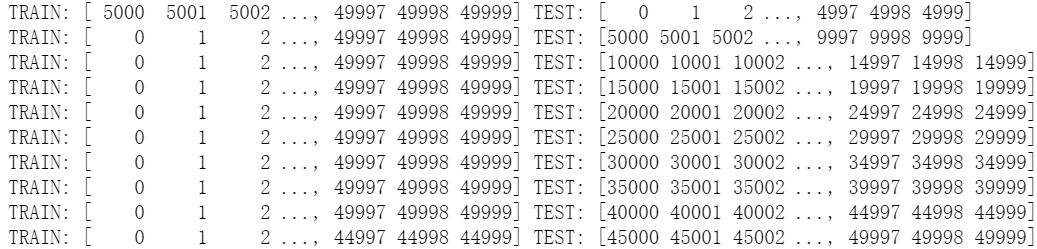


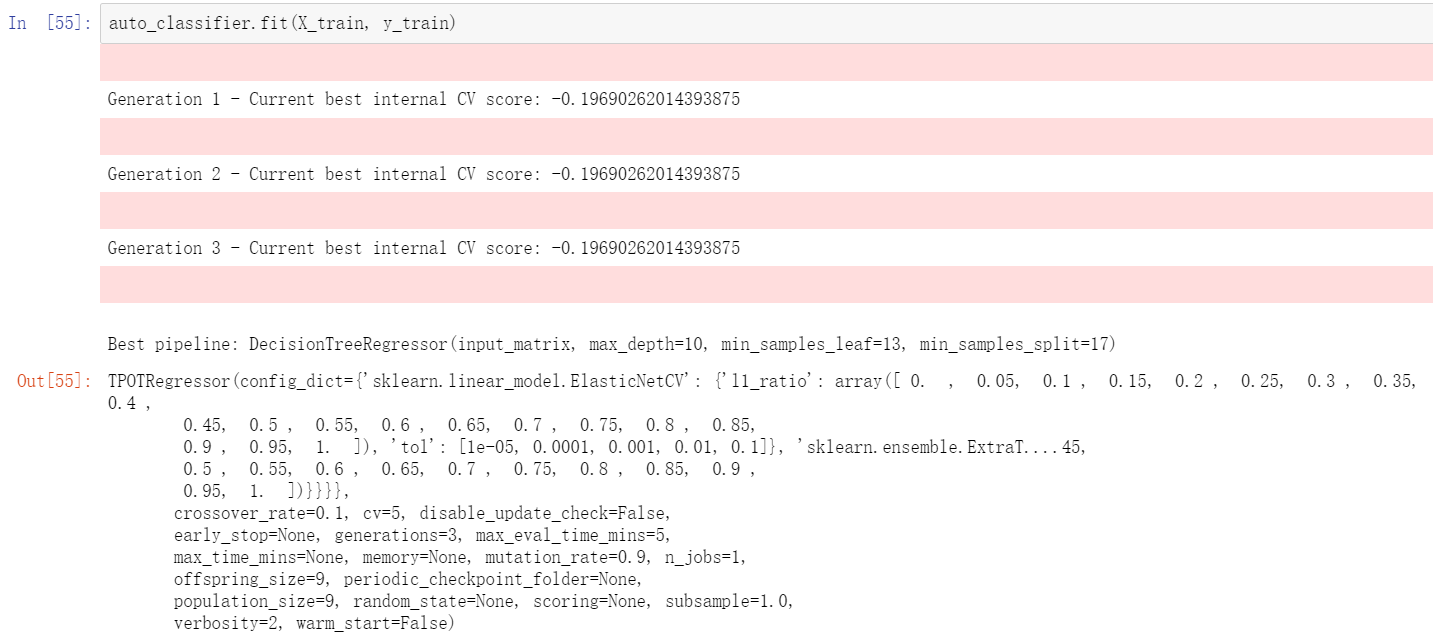
1. **Pipeline:**

After importing TPOTRegressor and KFold, the data has been trained by TPOTRegressor with n\_splits=10.

The best result of TPOT is Decision Tree Regressor with proper parameters.

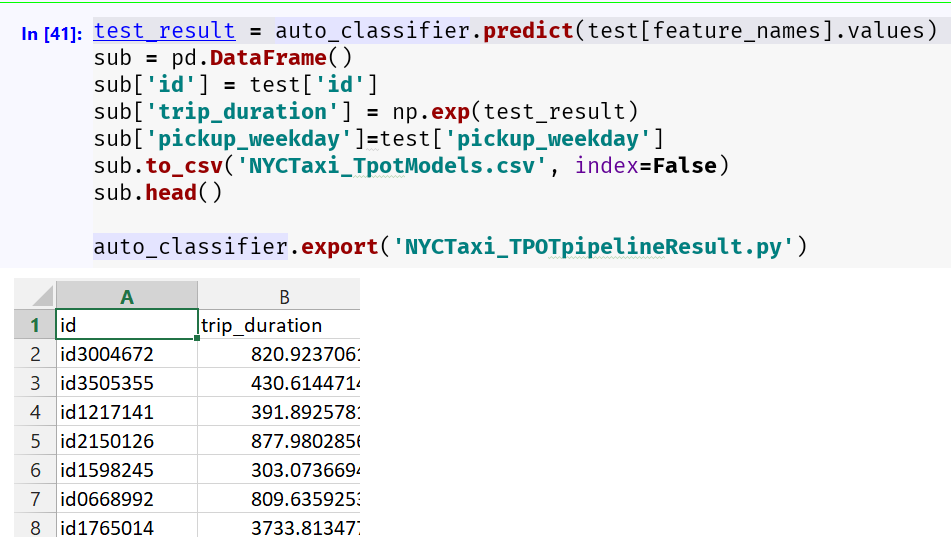






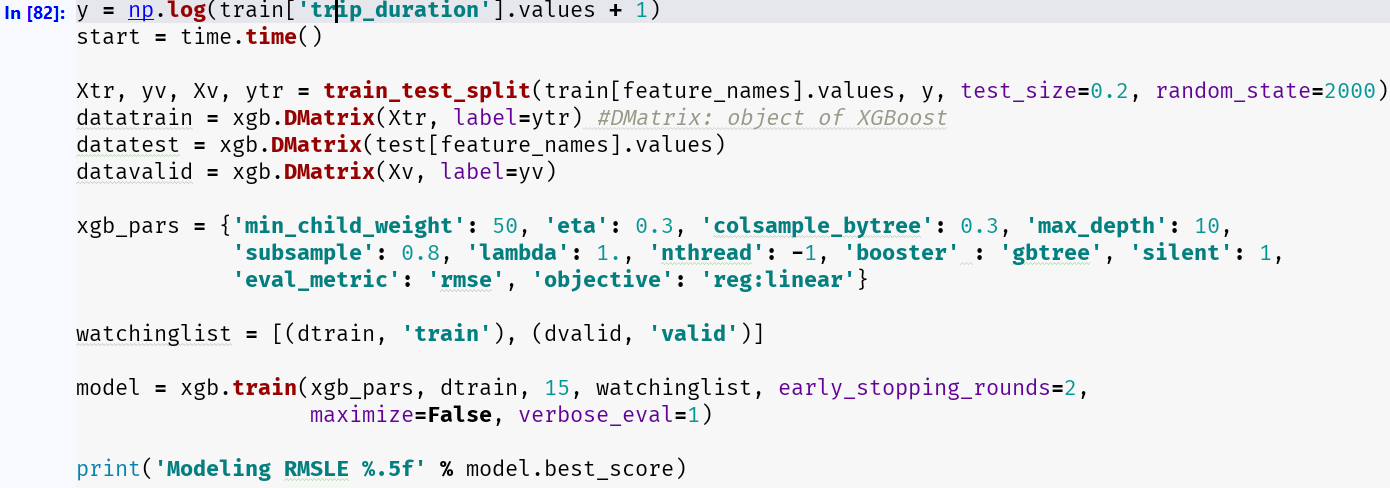
**3. Predict by the ‘best’ pipeline:**

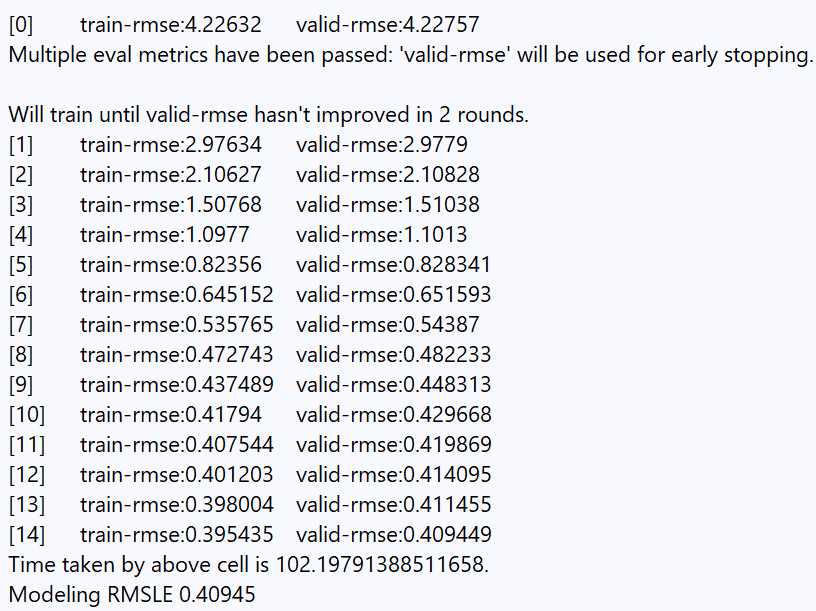
After generating the best pipeline, now we could predict the trip duration with it:



1. **Train and validate by XGBoosting:**

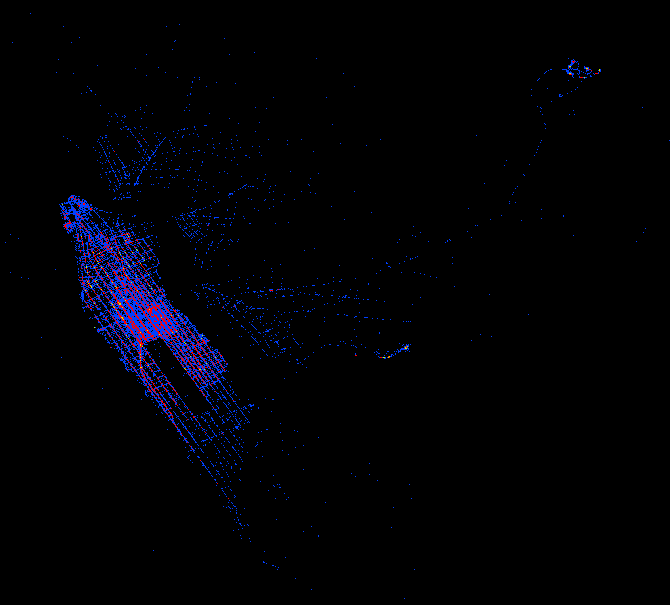
It will train until valid RMSE has not improved in 2 rounds. Finally, the modeling RMSLE is 0.40945. (Root Mean Squared Error (RMSE) and Root Mean Squared Logarithmic Error (RMSLE) both are the techniques to find out the difference between the values predicted by machine learning model and the actual values.)





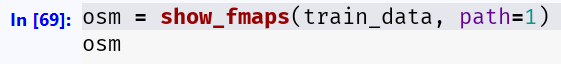
1. **Visualization by Traffic Heatmap:**

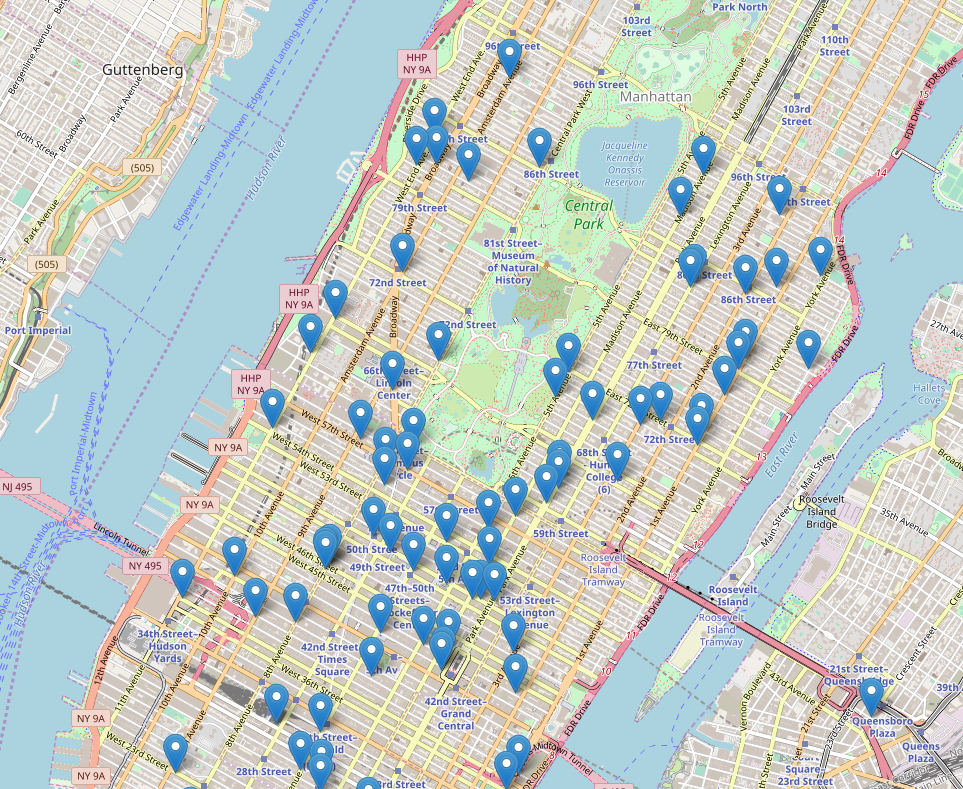
The heatmap will present the request according to different hour. The picture has taken a summary of lat-long and their count, and assign a different color for different count range.



1. **Visualization by Terrain map Folium:**

Before generating map and adding the pick up and drop coordinates, firstly, the function has been defined to calculate summary of given list of clusters.

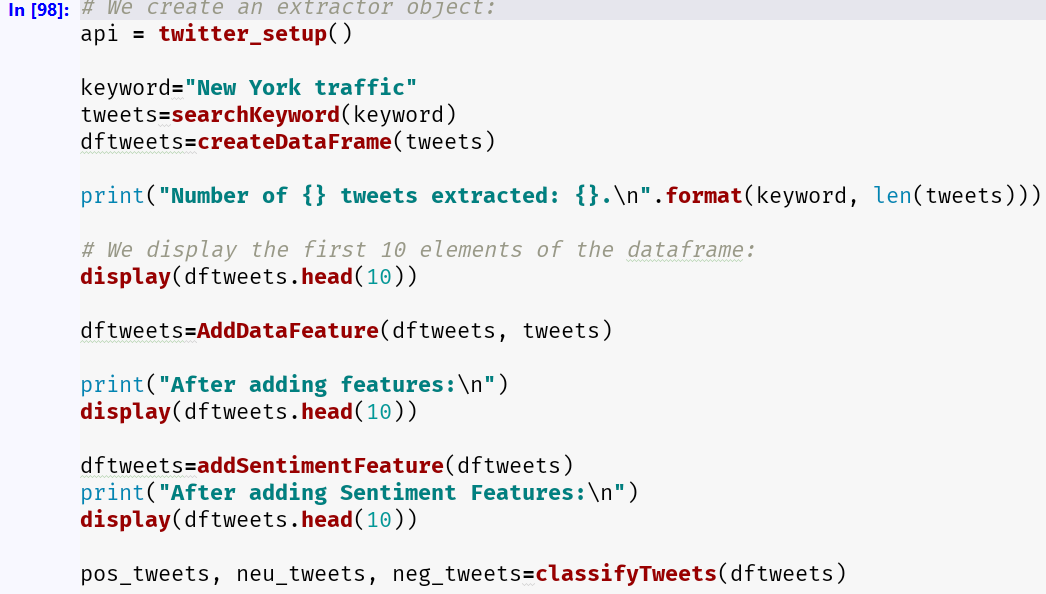


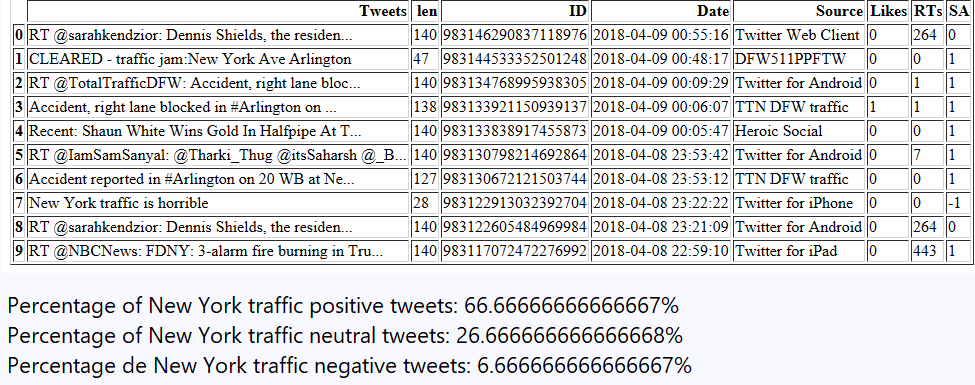


1. **Validation by Sentiment Analysis:**

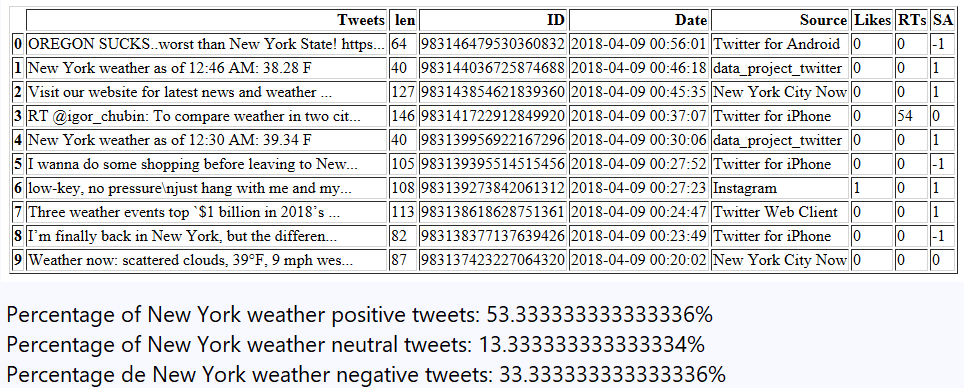
Let us check if the sentiment of twitters has some consistency with the result generated by Pearson Correlation.

After handling authorization, defining the keyword, adding tweets data into dataframe, sentiment analysis could be applied by importing TextBlob.

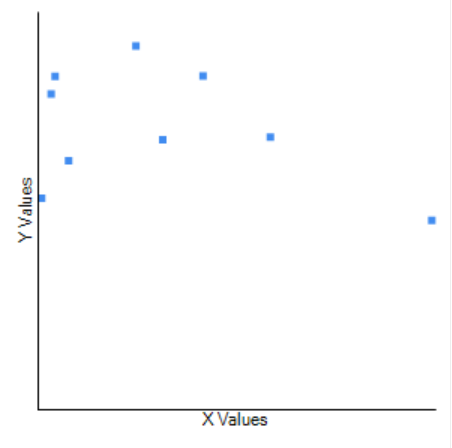




From this result, we find out the percentage of traffic positive tweets is 66.6% and percentage of negative tweets is 6.6%;



After 10 iterations of sentiment analysis, the correlation could be calculated as the chart. Pearson Correlation Coefficient is -0.385. That means there is a weak negative correlation between weather and traffic: It is likely that if the sentiment of weather is positive, the traffic will be a little worse. Vice versa. However, the correlation is not strong as imagine.



**Conclusion**

1. The trip duration could be predicted according to history records. The best model is DecisionTreeRegressor and current best internal CV score is -0.19.
2. The weather roughly affected the traffic, but not too much as common sense: It is used to be known that the traffic will getting worse with a worse weather. However, the traffic may be affected by diverse aspects of weather: perhaps the measures have been taken in advance, the numbers of vehicles has decreased due to severe weather. Thus, the weather could affect the traffic with -0.385 of Pearson correlation.
3. The geospatial heatmap could be plotted in visualized map based on features extracted and merged.

**Future work**

1. Further sentiment experience: In the sentiment analysis, just the keyword with broader and neutral meaning have been filtered and searched, like ‘weather’, ‘traffic’, ‘temperature’ etc. How about negative and positive keyword and will it affect the result? Like ‘accidents’, ’crash’, ’gale’, ‘sunshine’ etc.
2. Validation by more models: The model generated by TPOT may quite different. One time it was XGBoostRegressor. It changed to GradientBoostingRegressor and sometimes it become to DecisionTree. What is the reason?

**References**

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

[2] Sentiment Analysis: https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/

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

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

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

[6] NYC TLC Trip data:

http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml

[7] Open source routing machine: <http://project-osrm.org/>

[8] sklearn.tree.DecisionTreeRegressor Tutorial: <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>