NYC Taxi Prediction

Mehmet Bayhoca  
 Department of Computer Science  
 Binghamton University  
 Binghamton, NY  
 mbayhoc1@binghamton.edu

Ozlem Cinar  
 Department of Computer Science  
 Binghamton University  
 Binghamton, NY  
 ocinar1@binghamton.edu

Gizemnur Taskin  
 Department of Computer Science  
 Binghamton University  
 Binghamton, NY  
 gtaskin1@binghamton.edu

ABSTRACT

Our project name is NYC Taxi Prediction. On this project we are focusing on taxi routes especially located in NYC. With little touches (depending on data), our project can be used on any city all over the world. What have we done on project? On project we are focusing on the maximizing the profit for taxi drivers and showing the crowded areas for pick up. Our project has 2 perspectives and we created 2 maps in the end of the project. The first map is for customers to let them see in which area taxi drivers pick customer up and the second map is for drivers to let them see in which areas they can gain more money than others. All in all, this project is done for visual output and in introduction part, we will explain the project more detailed.

KEYWORDS

Taxi, Algorithm, Machine Learning, Dataset, Cluster Analysis, K-Means, Google Maps

1. Introduction

Transportation is one of the especially important area all over world. Especially on crowded city such as NYC, London, Paris and ext. On these crowded cities there are too much demand for taxis and depending on this situation, there are too much taxi drivers. There is big competition among these taxi drivers to gain more money. On this project we are focusing on this area. Our purpose is making taxi drivers gain maximum money during the month and we are using the dataset that is provided by nyc.gov. On our project we are using the pickup locations, drop off locations, distance, tips, fare and etc… To calculate the best locations for taxi drivers for maximum gain. We have used k-means for prediction, random forest for calculating r2\_score, StandardScaler for standardization, and unsupervised learning in k-means to see results during our project. We aimed that we could solve the accuracy and complexity problem with using K-Means since we had massive data to classify our data with K-Nearest Neighborhood method.

Also, we have focused on another area for customers. Sometimes finding a taxi in crowded cities can be difficult. With our project most used areas for pick up is showed in our project by using datasets and can be showed to customer to navigate them crowded areas for picking themselves up. On the next part we will be explain the Machine Learning algorithms that we used on this project to succeed.

2. K-Means Algorithm

On this project we have used the K-Means algorithm to process the datasets. Why we used the K-means algorithm? Actually, we were planning to use KNN on this project, but we have changed our ideas during the project, and we used the K-means. First reason for using K-Means instead of KNN is that K-Means is used on unsupervised learning. At the beginning we were planning to use supervised learning and KNN was fit for supervised learning. But when we changed our approach to unsupervised learning, we decided to use K-Means algorithm. Also, all data have no the same scale, using K-Means performs better when data has no the same scale. While we were using K-Means algorithm, we have used pre-defined external libraries from package ‘sklearn’. cluster import K-Means. K-means is used for clustering data. Due to this reason we have used K-means because our data is too big, and clustering is the best method for our data. K-means is also working good with unlabeled data and we have decided to use k-means.

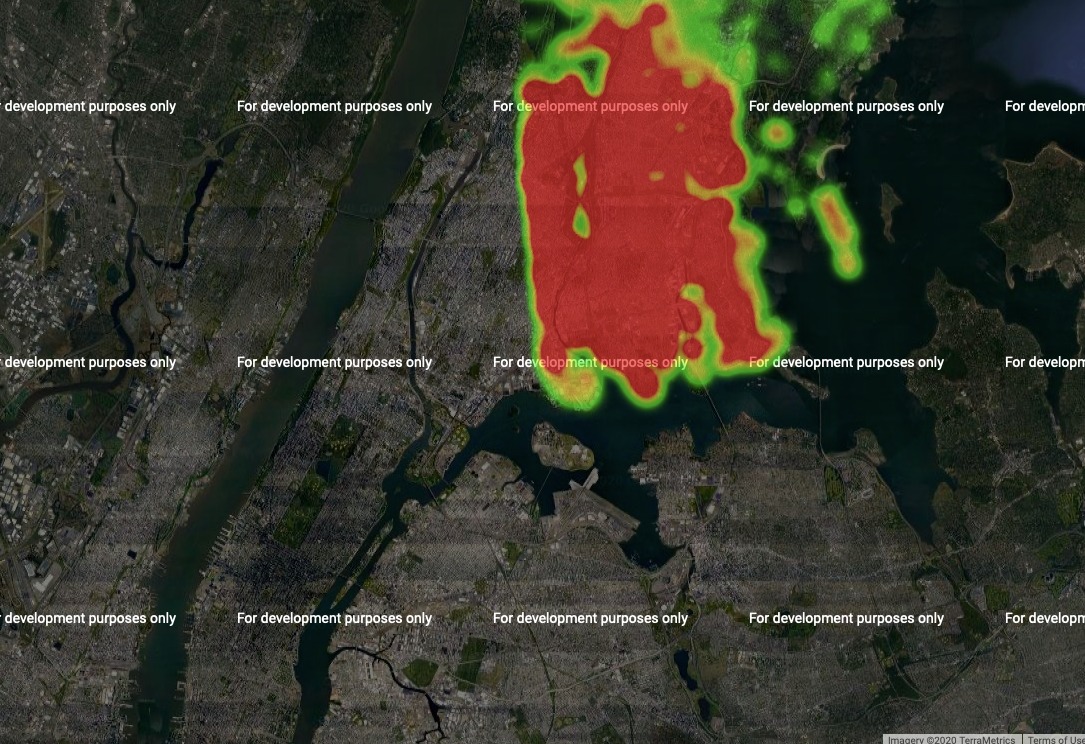
In the code part, we use K-Means twice in lucrative and pickup processes. Therefore, there are two calculations for prediction, standardization, clustering and fitting. After clustering and filtering the input data, we save output data in another new .csv file. In the first part, we use ‘pickup\_longtitude’ and ‘pickup\_latitude’ for calculating pickup intensity on map, while we use ‘pickup\_longtitude’, ‘pickup\_latitude’ and ‘fare\_per\_time’ for calculating more lucrative areas in K-Means process.

3. Random Forest

Random forest is an algorithm that is based on several decision trees. In random forest algorithm, there are many decision trees. With the help of the trees, we are finding best solution for training. As a result, it reduces the probability of overfitting. There are 4 steps for creating this algorithm. First getting values from dataset. The algorithm creates decision trees depending on these values after getting the data. Then, the result is predicted depending on decision trees. However, this algorithm may cause slowness on prediction.

However, we used K-Means in our project for prediction, fitting, and clustering. We tried to apply random forest in the project for choosing the best score, but we failed. We just got the r2\_score for clustered data. All things we did for random forest in command line.

5. Dataset

The dataset we used in this project is selected from [data.cityofnewyork.us](http://data.cityofnewyork.us) website. It is green taxi data aggregated daily, from 04/01/2016 to 05/01/2016. We are pointed the exact latitudes of predicted location. Since the data that have been gathered from taxi drivers and customers, have been change from pickup and drop off location to pick up and drop off id. Since we cannot deal with the ids on the map, we choose the dataset in April 2016 with given search area, New York City. Furthermore, the new and current datasets are available in given websites for future works. The dataset contains all the customers that took a drive on given month. Since that there is big mass of training data, we have struggled to choose best algorithm to classify or clustered the training data.

5. Project Environment

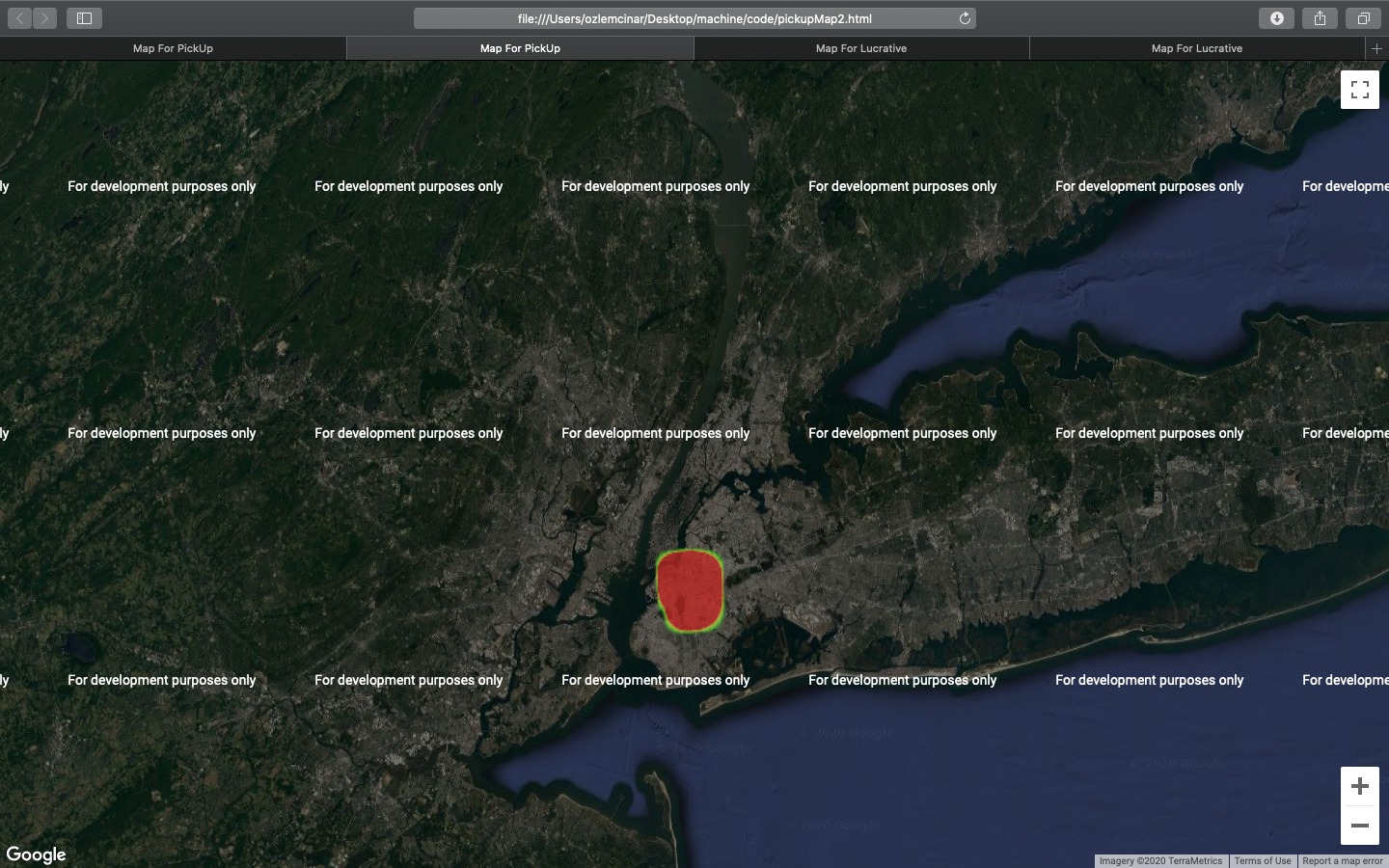
This project is mainly coded in Python language but also there are html files to make project visualized. In python code, there are pre-defined machine learning external libraries in packages and we have used some of them in our project. To run code, user needs to have python3 because some data structures need it. Html codes are used for visualization the predicted points on given New York City map. We use google map API for getting map. The API that we used in the html code is ”[https://maps.googleapis.com/maps/api /js](https://maps.googleapis.com/maps/api%20/js)?key=AIzaSyBnQLjhLNCmdTkdEzl62CrBnf9Sqtjz2L0&libraries=visualization&callback=initMap”. We have two .html files for starting html and ending html. The reason that we have both starting and ending html codes is we are writing values from csv file to points value in html. Then we write txt files into the html file and open it. That makes easier to see predicted points in the map.

There is ‘Makefile’ in the project file to make user compile project easily. When the make file is run by user, command lines will be compiled in terminal and in the end, two Google Maps having titles ‘pickup map’ and ‘lucrative map’ will be displayed.

6. Results

One sample result is as shown Figure 1 where the pickup locations of customers that are taken from our dataset. As seen in the map, red color determines that there is high pickup density in this location.

Figure 1

Yellow means there is low density of pickup latitude. As a result of that, we are predicted that if the driver wants to take customer more frequently, the driver should be in upper Manhattan. And customers will not wait too much since there is a big amount of taxi pickup density in that areas.

The other sample result is as shown in Figure 2 that enlighten the lucrative density for drivers. We are getting the sample data from our dataset same as the pickup density. We are looking pickup lucrative and drop off lucrative column in this dataset. As a result of predicted place that has more lucrative chance depending on probability.

A close up of a tree

Description automatically generated

Figure 2

As a result of the prediction of the project, more lucrative density can be seen in Queen and Brooklyn areas. Therefore, if a driver wants to gain more profit in the New York City, we can give our prediction result which is Queen.

On the other hand, as we mentioned before we were using random forest for train our data and our accuracy was about 61.77%. But we changed our project from supervised to unsupervised. On the other hand, the results can be seen in the google maps. We can compare the maps for accuracy if we enter the test data in addition to training dataset.

Figure 3

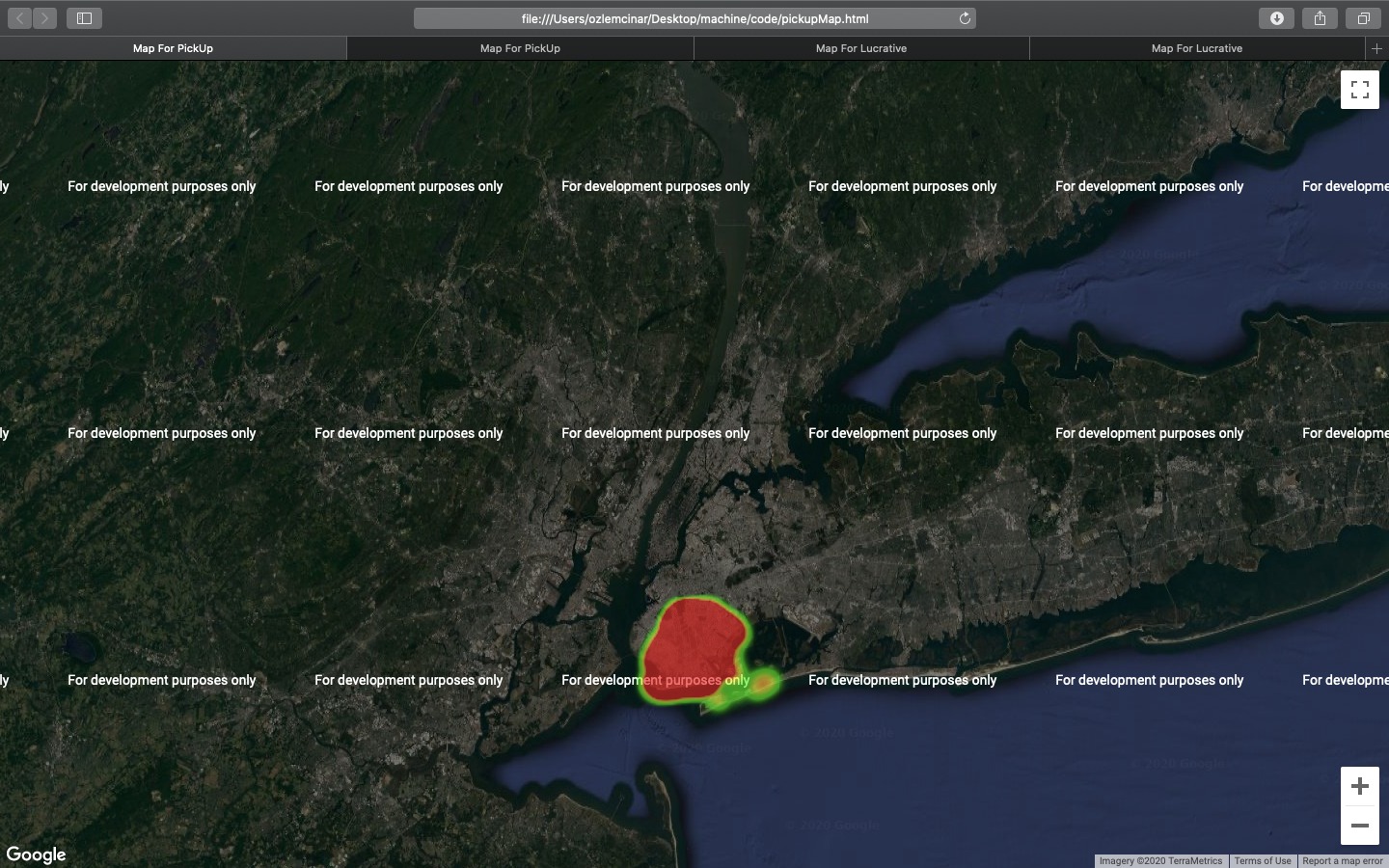


Figure 4

When we compare the two different results of the project according to “fincAccuracy.py” file, Figure 3 is the test data for pickup density and Figure 4 is the training data for pickup density. The similarity is 94% which is shown in Figure 7.

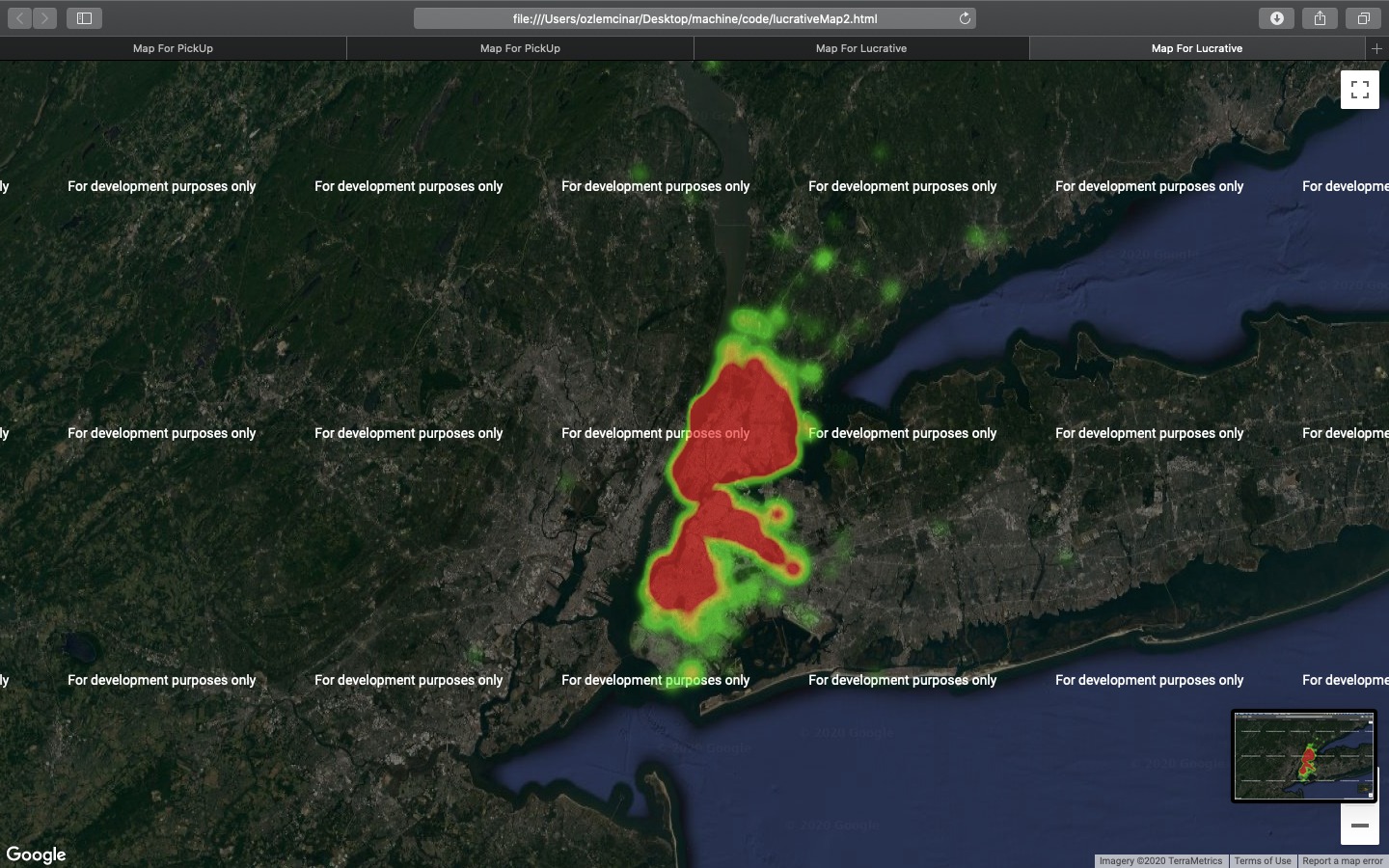


Figure 5

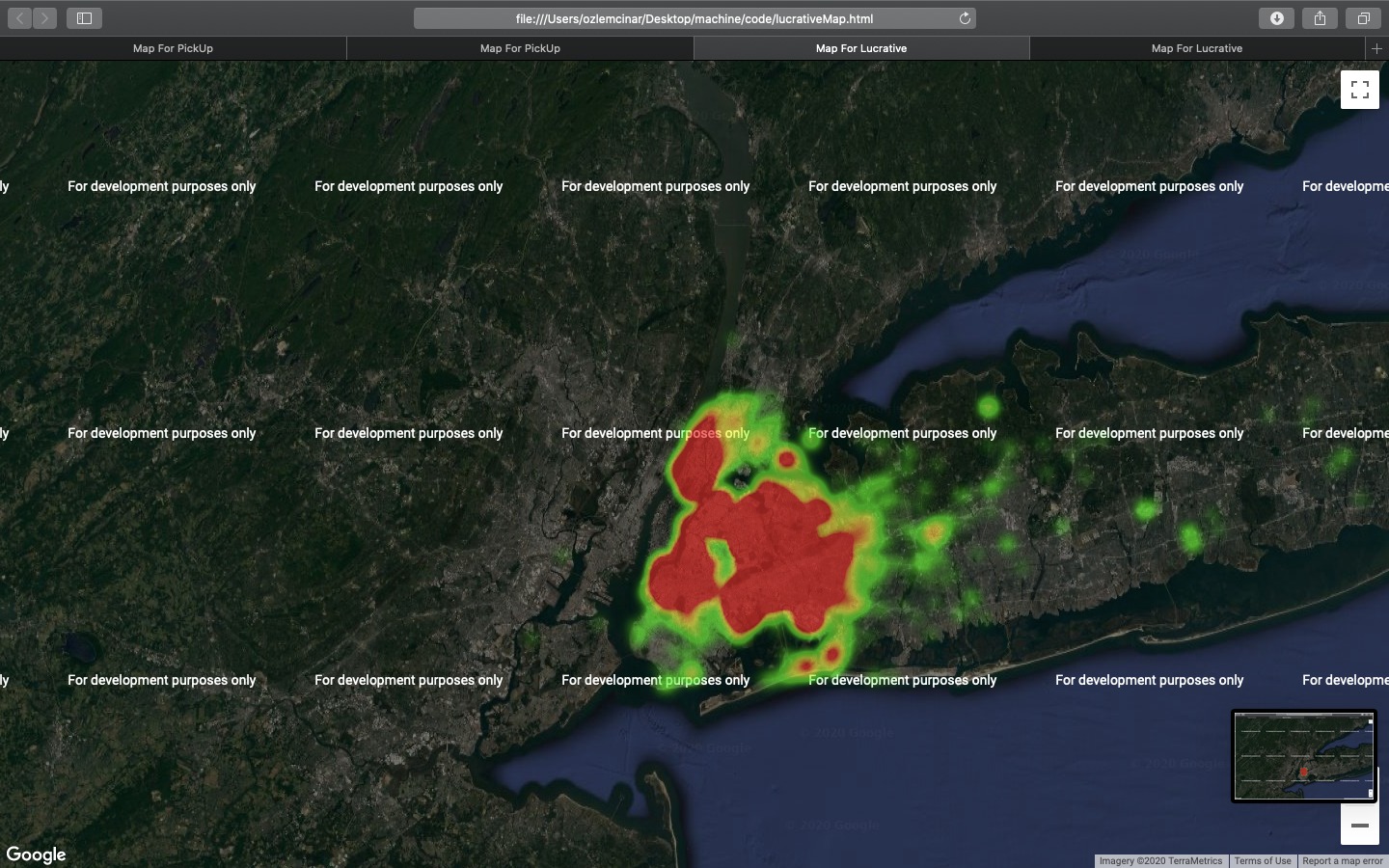


Figure 6

When we compare the two different results of the project, Figure 5 is the test data for lucrative density and Figure 6 is the training data for lucrative density. The similarity is 89% which is shown in Figure 8.

A screenshot of a cell phone

Description automatically generated

Figure 7

A screenshot of a cell phone

Description automatically generated

Figure 8

7. Conclusion

This study base taxi pickup location prediction which analyze the NYC green taxi trip data in April 2016. The project works with gathering data and analyzing it to determine the density of pickup latitude and lucrative. And results are shown in the google maps interface with the helps of html. To sum up, drivers can choose where they should go for individual reasons, to pick up customers without searching it, and gaining more money comparing the other trips, and customers can choose where they should be in for individual reasons to be picked up by drivers with looking the pick-up density areas.

8. References

* [data.cityofnewyork.us](http://data.cityofnewyork.us)
* [https://maps.googleapis.com/maps/api /js](https://maps.googleapis.com/maps/api%20/js)?key=AIzaSyBnQLjhLNCmdTkdEzl62CrBnf9Sqtjz2L0&libraries=visualization&callback=initMap

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00