THE BATTLE OF NEIGHBOURHOODS

By,
Akash Sarkar

BACKGROUND

Imagine the following scenario:

- You like to plan ahead and always review your options and make your choices about where you will visit and eat up front before you travel.
- You are flying to Chicago for a Data Science Conference.
- You arrive in Chicago the day the conference starts but you've managed to convince your boss to delay your return by a few days giving you time to explore.
- But you know no one in Chicago to show you around to all the top sites and to bring you to the best restaurants.
- Also the last time you went to a conference you were mugged and had you passport. money and credit cards stolen so you're now nervous of going somewhere without first researching the venue and the surrounding area.
- The conference is next week and you don't have time to do all the research you'd like.

INTRODUCTION:

When driven by venue and location data from FourSquare, backed up with open source crime data, it is possible to present the cautious and nervous traveler with a list of attractions to visit supplemented with a graphics showing the occurrence of crime in the region of the venue.

A high level approach is as follows:

- The travelers decides on a city location [in this case Chicago]
- The ForeSquare website is scrapped for the top venues in the city
- From this list of top venues the list is augmented with additional geographical data
- Using this additional geographical data the top nearby restaurants are selects
- The historical crime within a predetermined distance of all venues are obtained
- A map is presented to the to the traveler showing the selected venues and crime statistics of the area.
- The future probability of a crime happening near or around the selected top sites is also presented to the user

WHO IS THIS SOLUTION TARGETED?

This solution is targeted at the cautious traveler. They want to see all the main sites of a city that they have never visited before but at the same time, for whatever reasons unknown,

Some examples of envisioned users include:

- A single white female traveler
- An elderly traveler that has had previous back experiences when travelling

There are many data science aspect of this project including:

- Data Acquisition
- Data Cleansing
- Data Analysis
- Machine Learning
- Prediction

Now that the conference is over the Data Scientist can explore Chicago and feel much safer.

DATA SECTION:

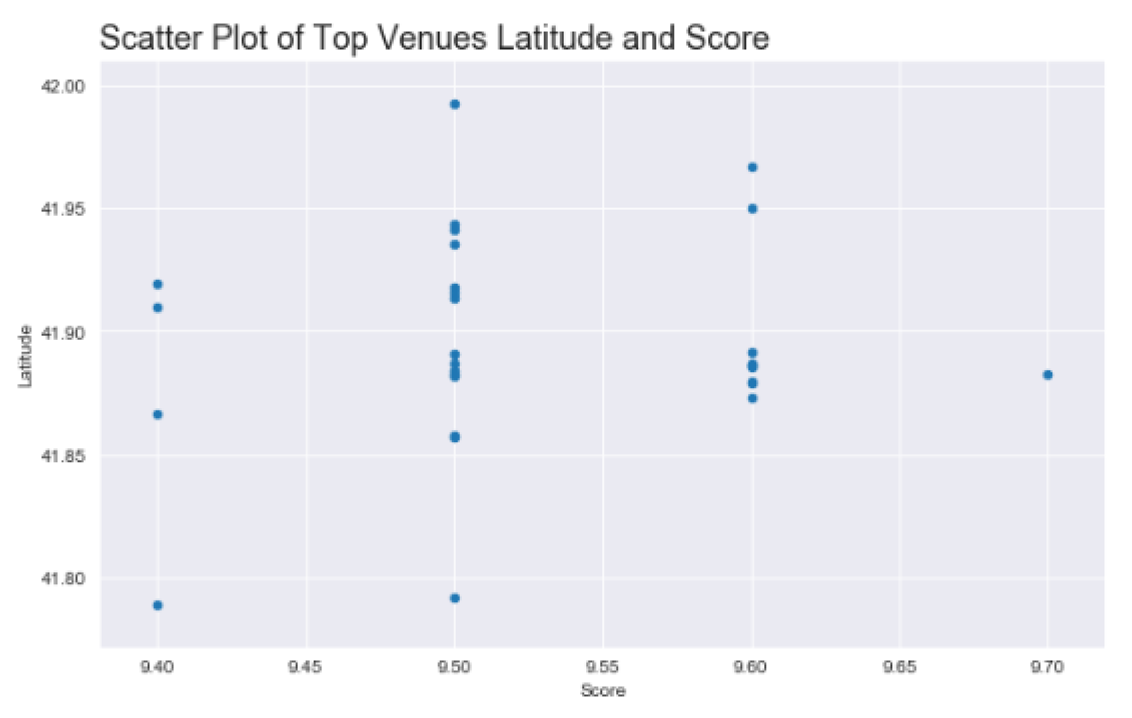
Given the size of the project and for simplicity only the following scenario will be addressed:

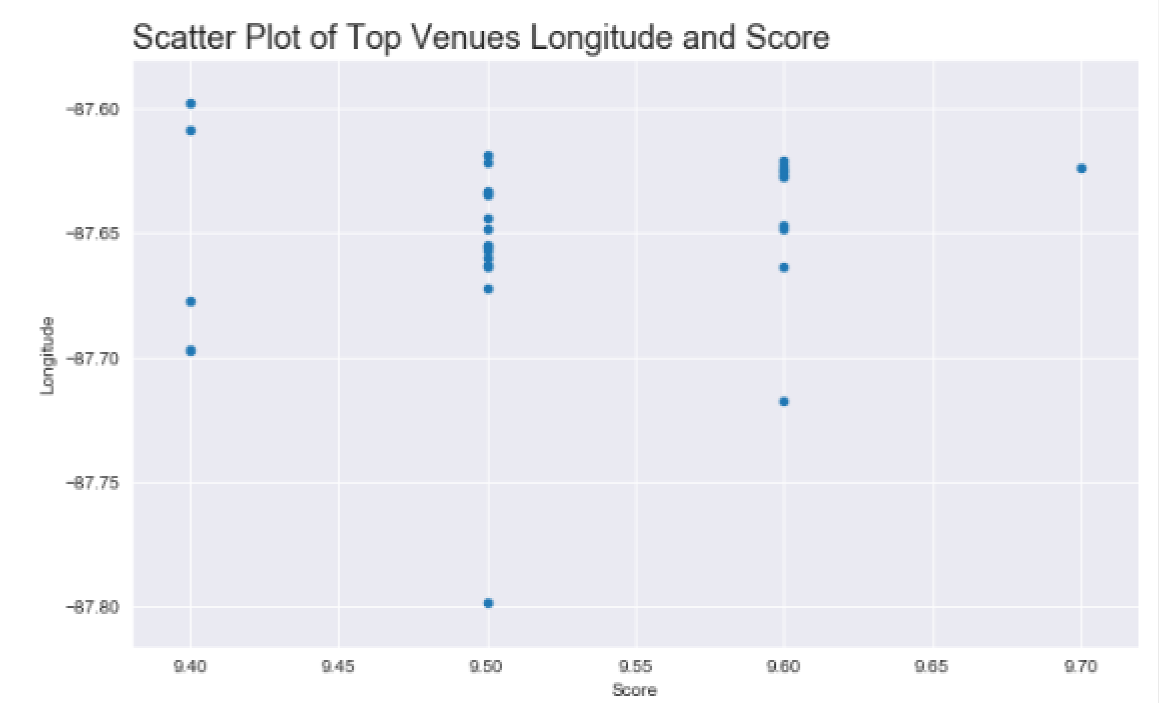
- Query the FourSqaure website for the top sites in Chicago
  - Go to www.foursquare.com, enter the city of your choice and select Top Picks from *I'm Looking For* selection field.
- Use the FourSquare API to get supplemental geographical data about the top sites
  - Using the id field extracted from the HTML it is then possible to get further supplemental geographical details about each of the top sites from FourSquare using the API call
- Use the FourSquare API to get top restaurant recommendations closest to each of the top site
- Use open source Chicago Crime data to provide the user with additional crime data

METHODOLOGY:

Exploratory Data Analysis

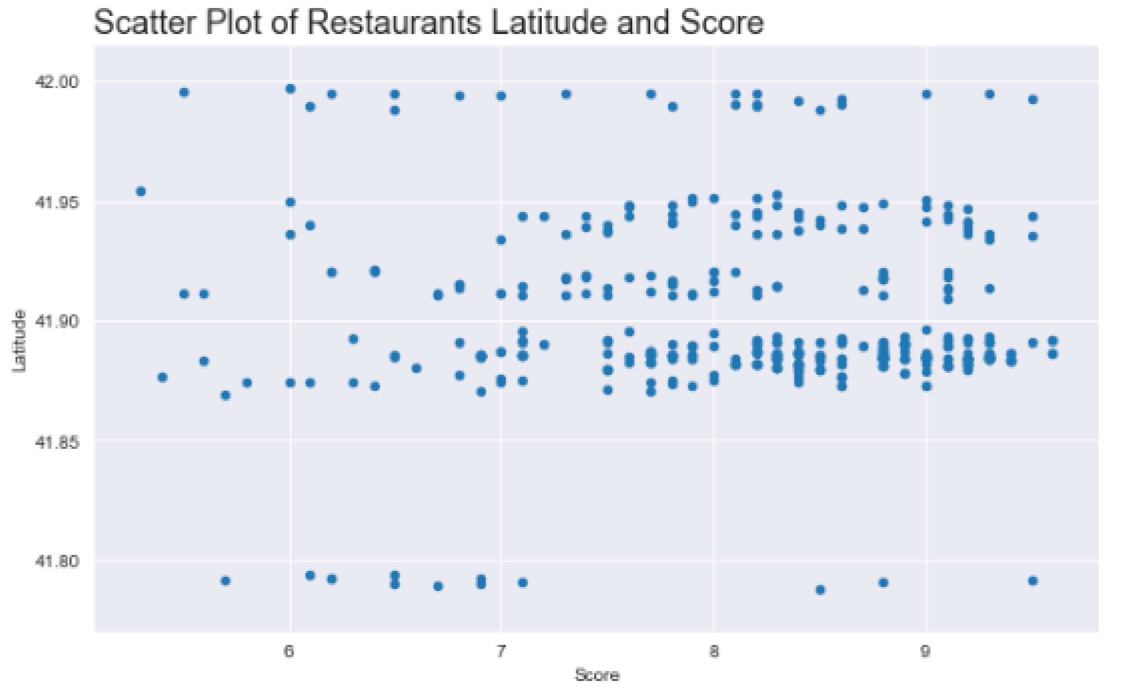
The first round of exploratory analysis was to examine the Top Venues and Restaurants Dataframes to determine if there was any correlation between variables.

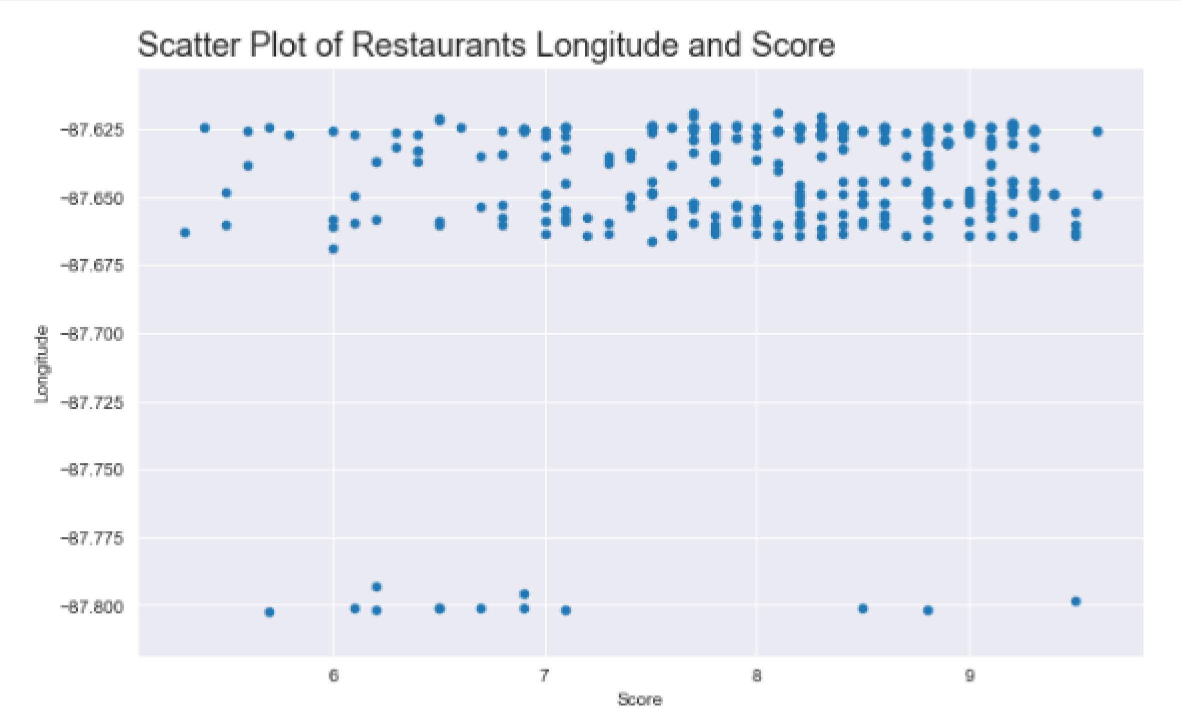




METHODOLOGY (CONTINUED)

Although nothing obvious to would appear that the top venues are centered arounf the -87.65 Longitude, the Restaurant data was examined next.





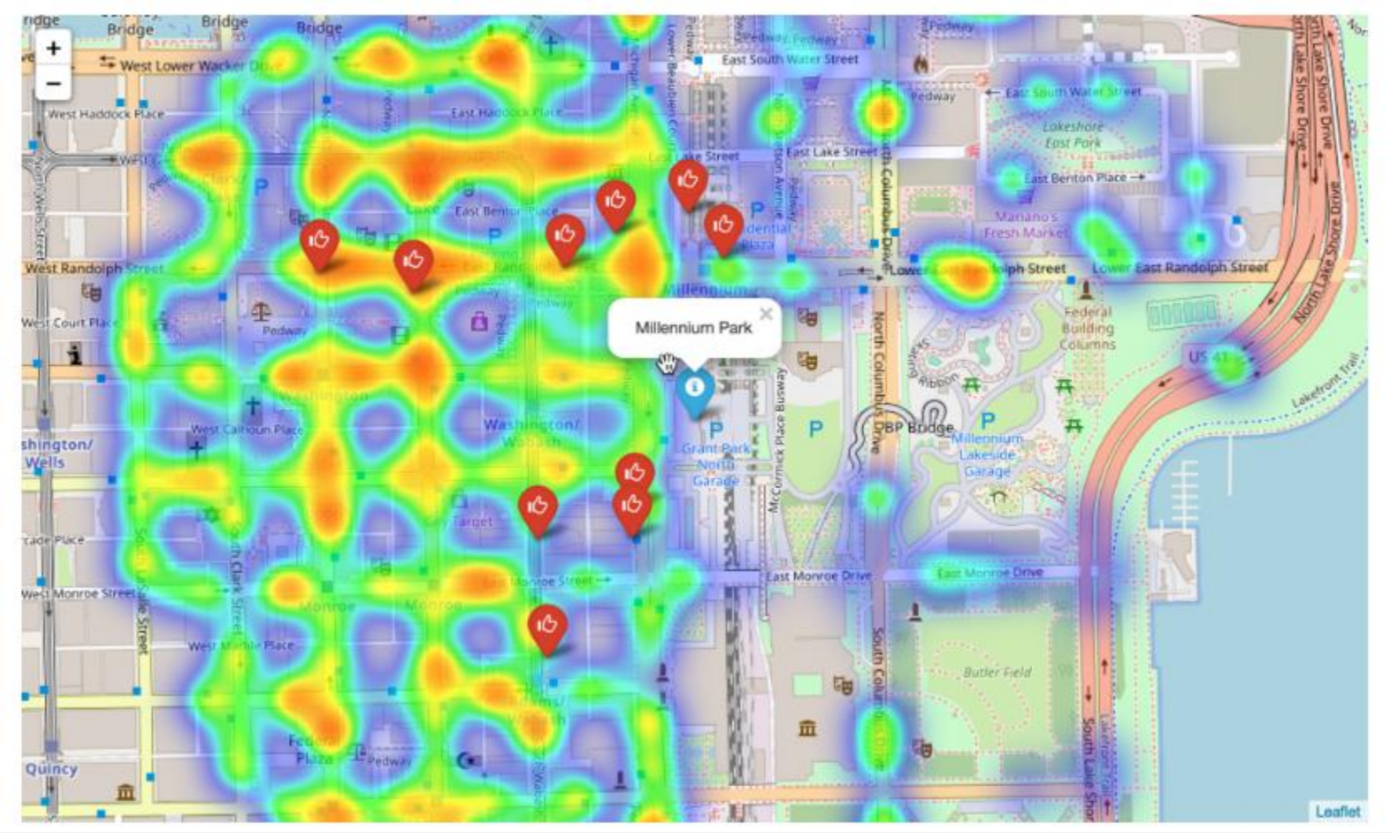
FURTHER VISUALISATION

In this section a preview of the type of data that will be displayed to a user of the proposed solution is shown.

- For each of the Top 10 Venues:
- All crimes within 750 meters of the venue are added to a dataframe
- All restaurants associated with the venue are added to a dataframe
- A folium Map is created centered on the venue
- A heatmap of the crimes in the area are overlaid
- the venue is marked on the map
- The top 10 scored restaurants are marked on the map

HEATMAP OF MILLENNIUM PARK

- The location of the attraction and the 10 top rated venues are clearly shown.
- The Top Venue is shown using a blue marker, the restaurants are shown using a red marker
- Also shown is the heatmap of cimes within 750 meters over the course of the entire previous year



MODELLING

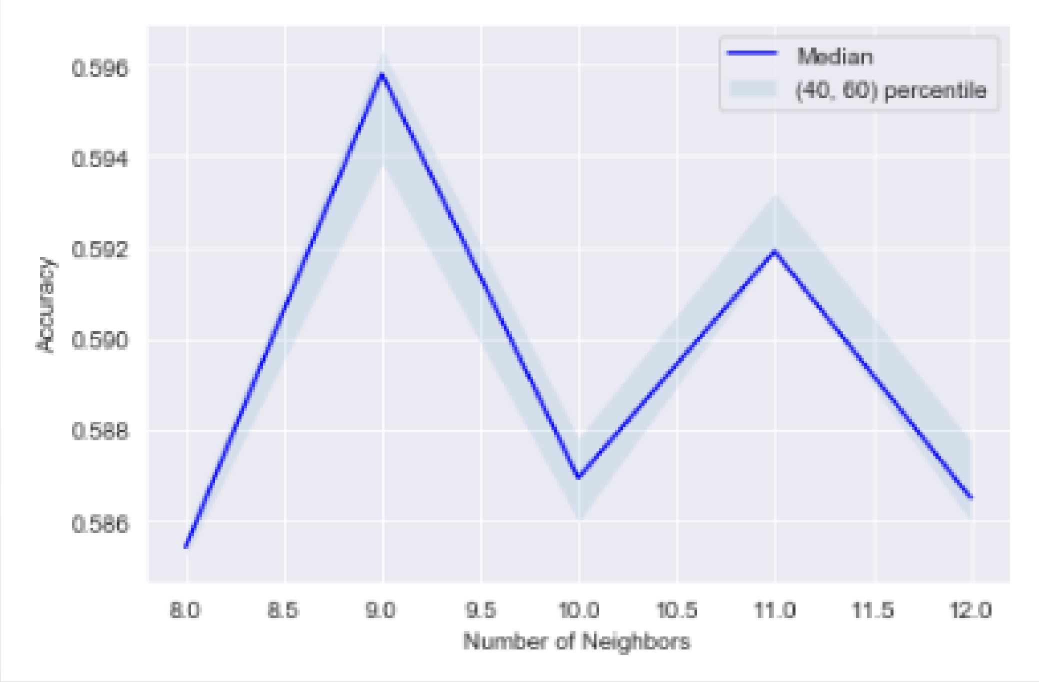
Five model type were then chosen to be evaluated:

- K Nearest Neighbours
- Decision Trees
- Logestic Regression
- Naive Bayes
- Decision Forest using a Random Forest

However, Logistic Regression and Naive Bayes models did not return any models with an accuracy greater that 0.61.

K NEAREST NEIGHBOR (KNN)

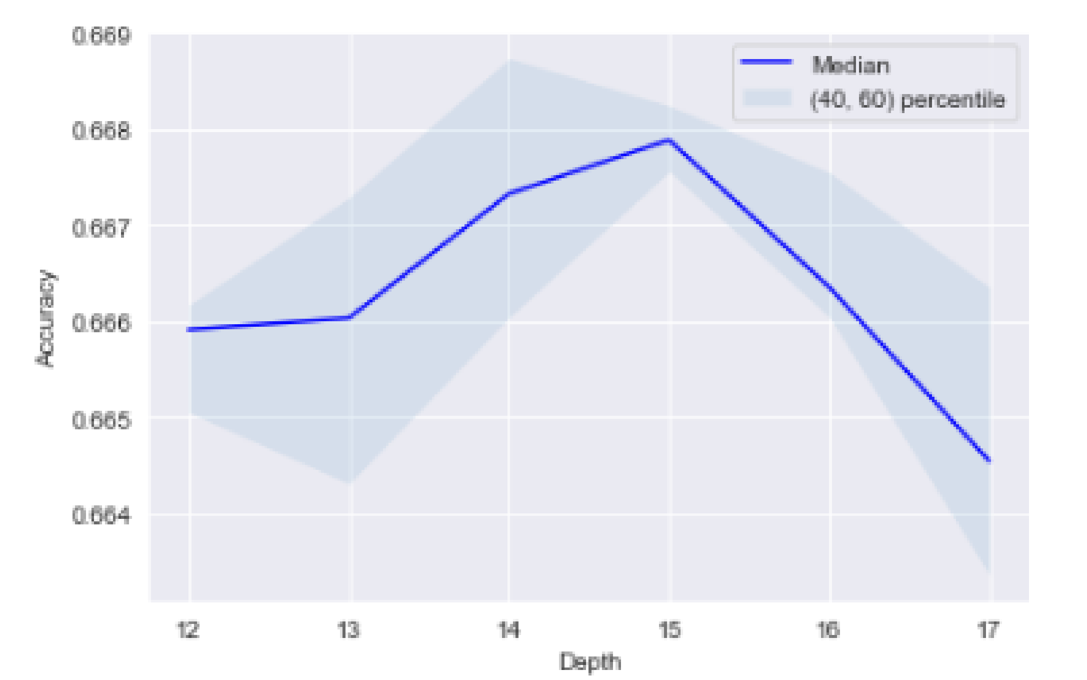
KNN Model was quick to execute and through the process of evaluation it was discovered the K = 9 gave the best results K Nearest Neighbors



KNN was not particularly fast taking approximately 10 minutes per model.

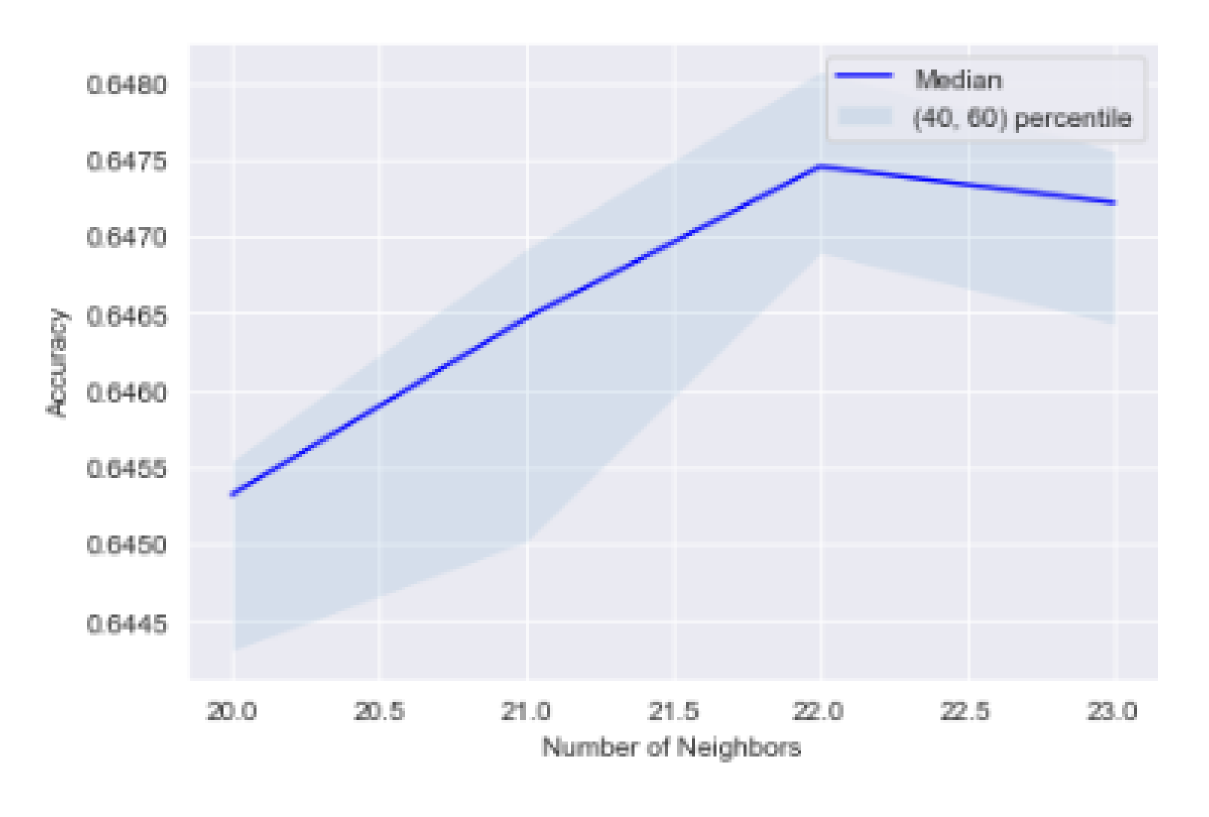
DECISION TREE

The Decision Tree model was particularly fast taking only 10 seconds per model. This meant that it was easy to try multiple different parameters. A tree depth of 15 gave the best model performance:



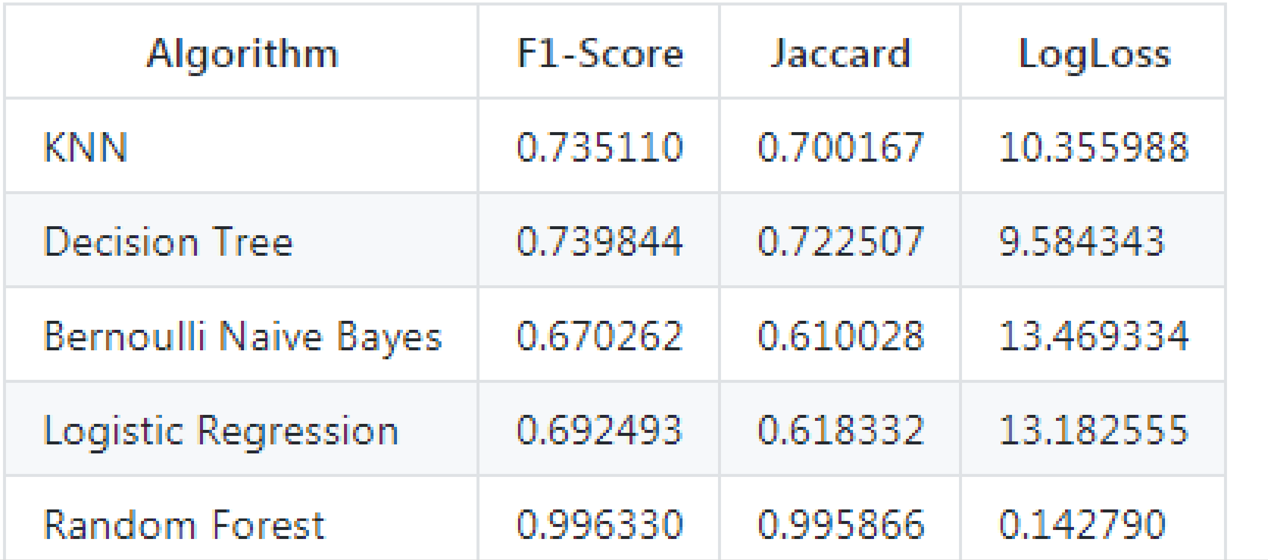
DECISION FOREST USING A RANDOM FOREST

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.



BEST MODEL

- Using the the crime data for the top two occuring crimes each of the top performing models where further evaluated to to determine which model performed the best using F1-Score, Jaccard Score and Log Loss.
- Randon forest was determined to be the best model.

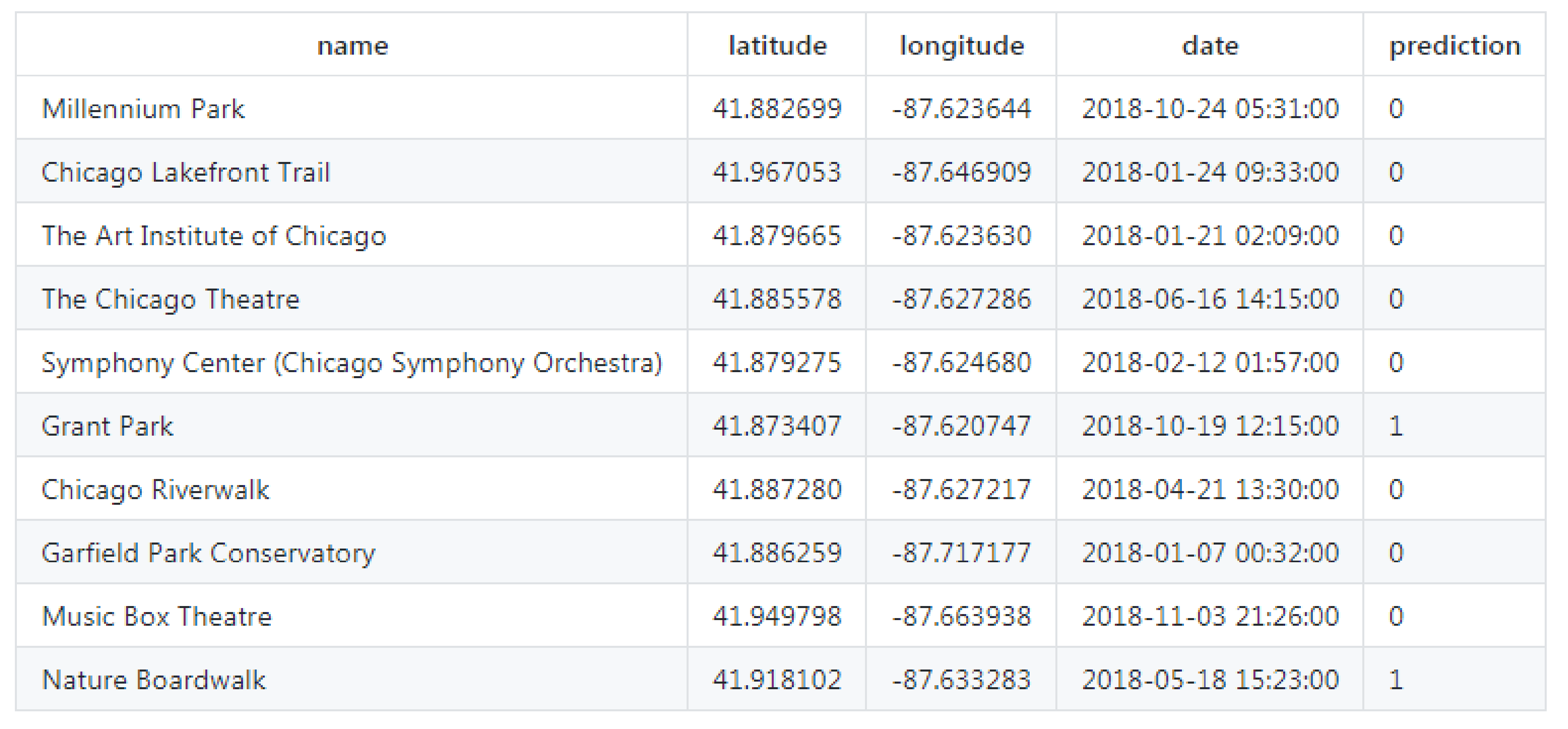


Algorithm	F1-Score	Jaccard	LogLoss
KNN	0.735110	0.700167	10.355988
Decision Tree	0.739844	0.722507	9.584343
Bernoulli Naive Bayes	0.670262	0.610028	13.469334
Logistic Regression	0.692493	0.618332	13.182555
Random Forest	0.996330	0.995866	0.142790

RESULTS AND PREDICTION

Final 10 venues were identified.

Of the top ten venues 8 were identified as potentially dangerous to visit and 2 were deems safe. As there is no data to compare the predictions against the best way we will visualise the data again.

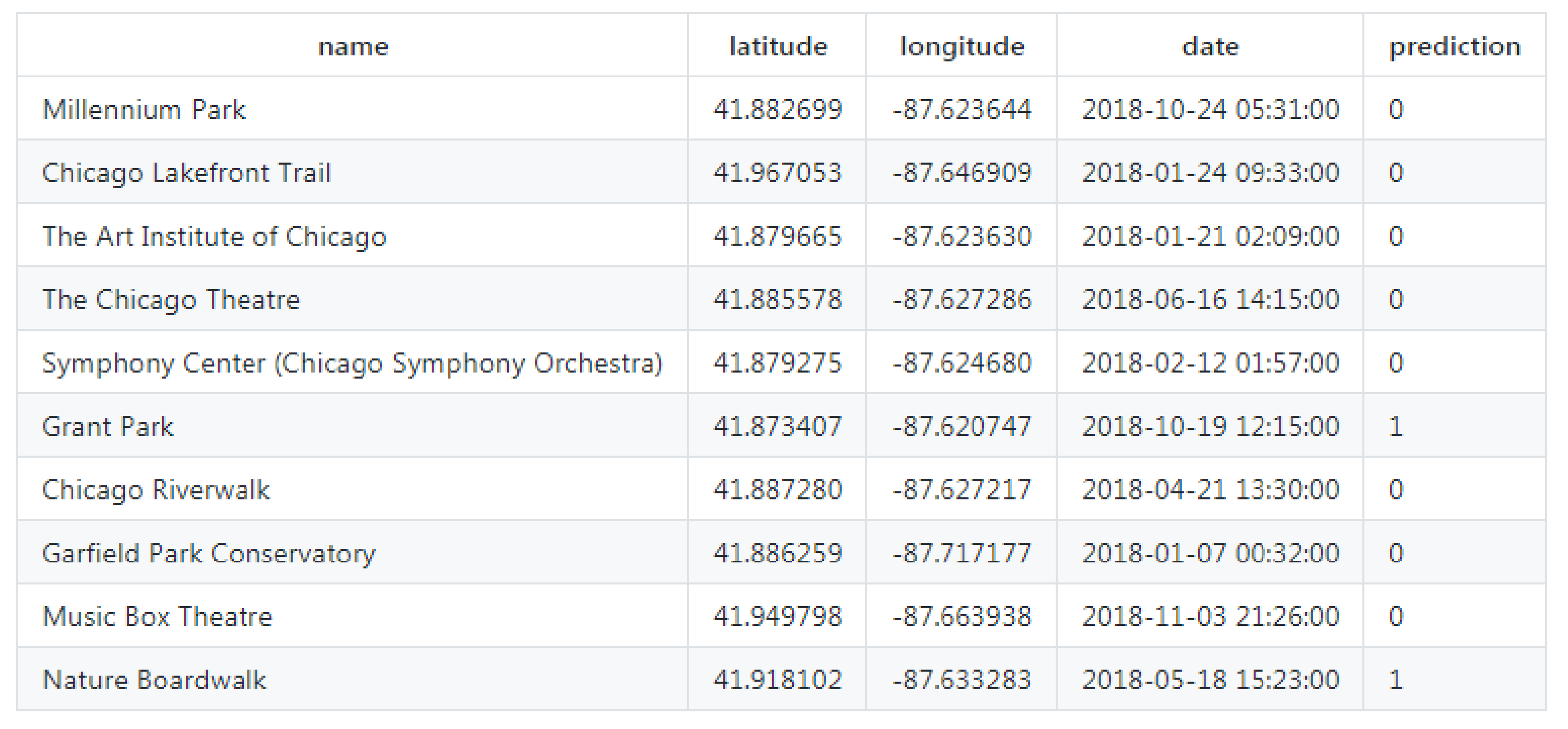


name	latitude	longitude	date	prediction
Millennium Park	41.882699	-87.623644	2018-10-24 05:31:00	0
Chicago Lakefront Trail	41.967053	-87.646909	2018-01-24 09:33:00	0
The Art Institute of Chicago	41.879665	-87.623630	2018-01-21 02:09:00	0
The Chicago Theatre	41.885578	-87.627286	2018-06-16 14:15:00	0
Symphony Center (Chicago Symphony Orchestra)	41.879275	-87.624680	2018-02-12 01:57:00	0
Grant Park	41.873407	-87.620747	2018-10-19 12:15:00	1
Chicago Riverwalk	41.887280	-87.627217	2018-04-21 13:30:00	0
Garfield Park Conservatory	41.886259	-87.717177	2018-01-07 00:32:00	0
Music Box Theatre	41.949798	-87.663938	2018-11-03 21:26:00	0
Nature Boardwalk	41.918102	-87.633283	2018-05-18 15:23:00	1

RESULTS AND PREDICTION

Final 10 venues were identified.

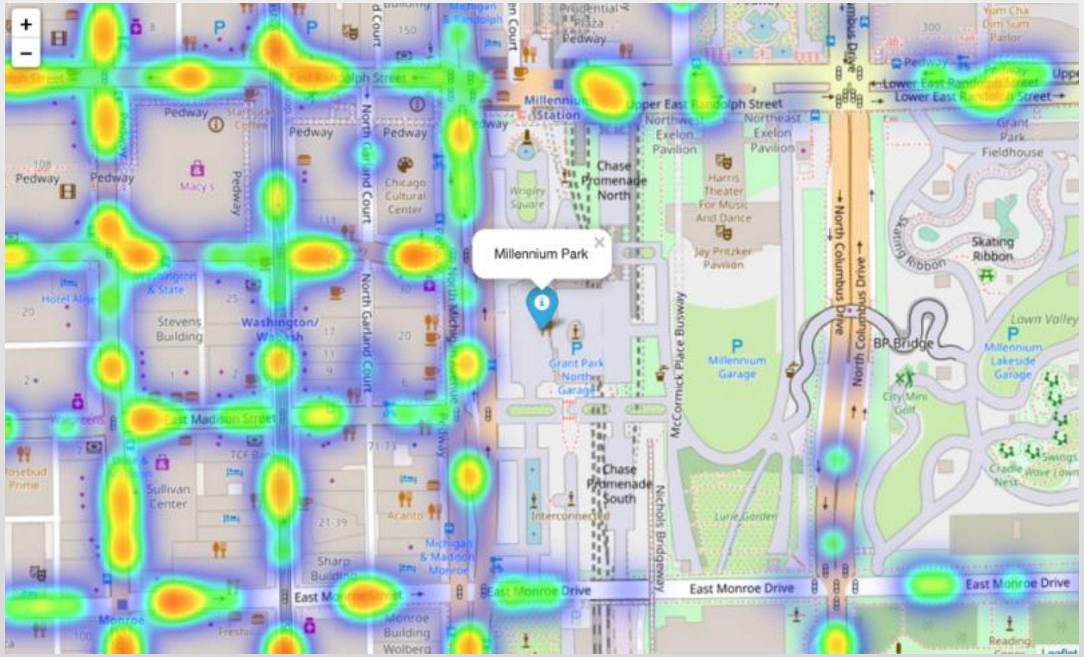
Of the top ten venues 8 were identified as potentially dangerous to visit and 2 were deems safe. As there is no data to compare the predictions against the best way we will visualise the data again.



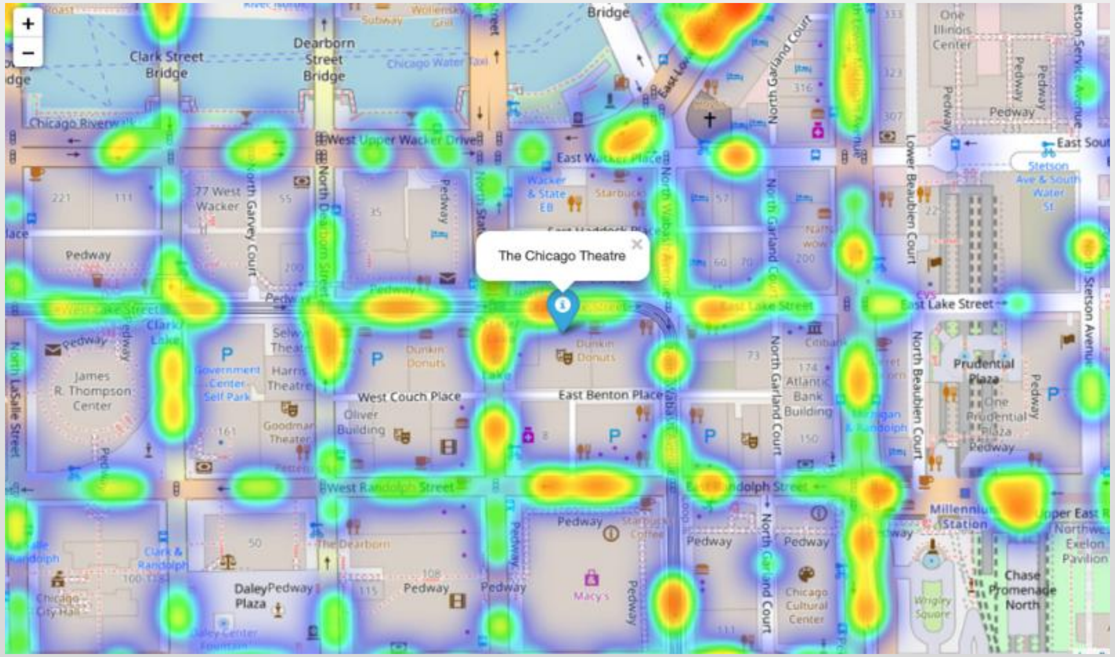
name	latitude	longitude	date	prediction
Millennium Park	41.882699	-87.623644	2018-10-24 05:31:00	0
Chicago Lakefront Trail	41.967053	-87.646909	2018-01-24 09:33:00	0
The Art Institute of Chicago	41.879665	-87.623630	2018-01-21 02:09:00	0
The Chicago Theatre	41.885578	-87.627286	2018-06-16 14:15:00	0
Symphony Center (Chicago Symphony Orchestra)	41.879275	-87.624680	2018-02-12 01:57:00	0
Grant Park	41.873407	-87.620747	2018-10-19 12:15:00	1
Chicago Riverwalk	41.887280	-87.627217	2018-04-21 13:30:00	0
Garfield Park Conservatory	41.886259	-87.717177	2018-01-07 00:32:00	0
Music Box Theatre	41.949798	-87.663938	2018-11-03 21:26:00	0
Nature Boardwalk	41.918102	-87.633283	2018-05-18 15:23:00	1

VISUALIZATIONS OF PREDICTIONS

These two images are of Millennium Park and of The Chicago Theatre. Both of these venues were identified as likely to be susceptible to crime.



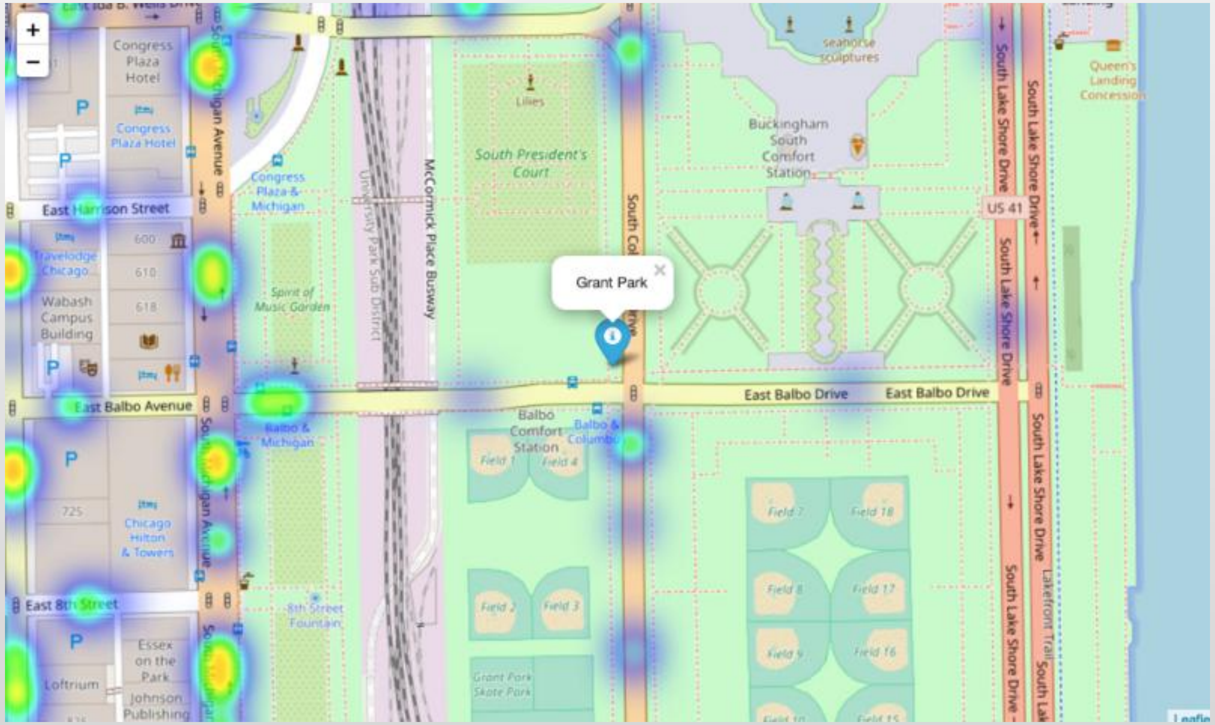
a) Millennium Park



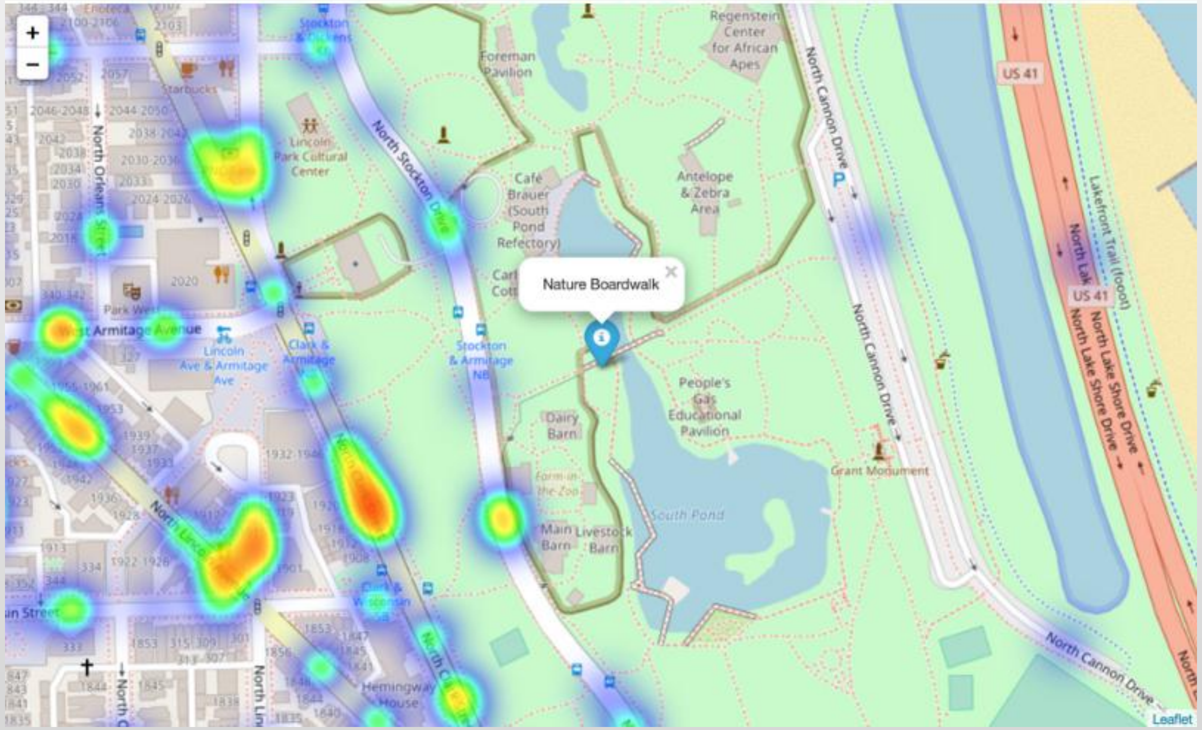
b) Chicago Theatre

VISUALIZATIONS OF PREDICTIONS

These images are from Grant Hill and Nature Boardwalk. Although both show signs of criminal activity, both have far less than Millennium Park and The Chicago Theatre.



a) Grant Hill



b) Nature Boardwalk

CONCLUSIONS AND DISCUSSIONS

- Although all of the goals of this project were met there is definitely room for further improvement and development as noted below. However, the goals of the project were met and, with some more work, could easily be devleoped into a fully phledged application that could support the cautious traveller in an unknown location.
- Of the contributing data the Chicago Crime data is the one where more data would be good to have. Also not every city in the world makes this data freely available so that is a drawback.
- FourSquare proved to be a good source of data but frustrating at times. Despite having a Developer account I regularly exceeded my hourly limit locking me out for the day. This is why Pickle was used to store the captured data.

FURTHER DEVELOPMENT

The following are suggestions how this project could be further developed:

- Best time to visit each venue
- Suggestions for morning, afternoon, evening and night time
- Daily itineraries
- Route planning and transportation
- Time lapse of the crime in the area of the venue
- Favorite dining preferences could be used to choose the restaurants