**Predicting the Number of Covid-19 Cases Based on Neighborhood Venues**

Robin Zhao

**1. Introduction**

**1.1 Background**

During 2020, the number of COVID-19 cases has surged across the state of New York. After strict COVID-19 safety guideline enforcements, the infection rate in NYC gradually decreased, and New York city was starting to open up. The reopening of the New York City has four phases, and in around July 2020, New York was entering phase three of the reopening, where outdoor dining, local gyms, shopping malls and other entertainment facilities were reopened. However, with these facilities opening, the COVID-19 infection rate came up simultaneously. In January 2021, New York City’s restaurant’s indoor dining were forced to close again because of the surging virus. There is clearly a positive correlation between the reopening businesses and the virus infection rate. With different infection rates in different neighborhoods of New York City, it would be interesting to examine how different venues in a neighborhood affects that neighborhood’s infection rate as well as if number of venues can effectively predict the neighborhood’s infection rate.

**1.2 Problem**

Different venues in a neighborhood can affect the overall neighborhood’s infection rate. This paper is to examine the relationships between venues and the neighborhood’s infection rate and try to effectively predict the neighborhood’s infection rate based on the venue lists.

**1.3 Interest**

The government would be very interested in an accurate prediction of COVID-19 infection rates in different neighborhoods in order to come up with strategies to combat it. Moreover, individuals would be very interested too because it can give them some precautions in traveling into another possibly highly infected neighborhood.

**2. Data Acquisition**

One dataset is found on NYC Health website in the Percent Positive and Test Rate of Molecular Testing by ZIP code section. This dataset is able to give us the information of the neighborhood name, its ZIP code, and its infection rate. Another dataset is found on opendatasoft in the US Zip code Latitude and Longitude section. This second dataset is able to give us the coordinates for each neighborhood by its ZIP code, and hence we can us Foursquare to acquire the venue lists for each NY neighborhood by its latitude and longitude. Finally, we can conduct our analysis on the venue lists and the infection rates.

**3. Methodology**

After cleaning the data, a data frame consists of neighborhood name, infection rate, neighborhood latitude and neighborhood longitude is created. After requesting nearby venue list of each neighborhood, a categorical dataset is created, where for each neighborhood, its nearby venues are all marked “1” under each corresponding venue name columns. Grouping by the neighborhood, a new final dataset shows the frequency of each venue in each neighborhood, and then our final dataset is ready to be analyzed.

First, K-means clustering machine learning method is applied to the dataset, putting all neighborhoods into 5 clusters, and then folium package is used to visualize the labeled neighborhoods in a map:

Map

Description automatically generated

Next, a bar graph is drawn using the dataset, showing how infection rates in neighborhoods comparing to each other. Neighborhoods are sorted by the infection rate in descending order:

Chart, bar chart

Description automatically generated

This is useful because the graph can give a visualization of which neighborhoods are the most dangerous place to go into, and that can give the government some guidance of how to cope with the virus.

Then, another column is created, giving each neighborhood’s infection rate a label of low, medium, and high level. After splitting the dataset into training and testing sets, the K-NN classification method is applied, with x being the total frequency of restaurants, stores, and gyms in each neighborhood, y being the level of infection rates, and k clusters that can give the best accuracy. After our model is training, it is used to predict the x test set. Then decision tree, support vector machine classification are applied in the same way.

**4. Results**

In the map of clustered neighborhoods, the colored dots indicate the New York neighborhoods are being divided into 4 clusters, which roughly corresponds to Manhattan, Queens, Brooklyn, and Williamsburg locations. This is not surprising, as the density of venues in each Borough is roughly the same and different than others. For example, venues in Manhattan are generally more dense than other Boroughs. In the final machine learning methods, accuracy scores for three different machine learning classification techniques (K-NN, Decision Tree, Support Vector Machine) are indicated in the following table:

Graphical user interface, text, application

Description automatically generated with medium confidence

Since the accuracy score for Support Vector Machine classification is the highest, we can choose this method to predict the infection rate based on one neighborhood’s venue list.

**5. Discussion**

The accuracy score of Support Vector Machine classification is around 0.77, which indicates that the model we used to predict the infection rate is relatively accurate. This means that given any neighborhood in the New York state, even if there is not enough COVID-19 testing to determine an infection rate, we can still be able to deduce a relatively accurate infection rate by only using the venues inside the neighborhood. This finding can greatly help the government in combating the virus in many ways. For example, when issuing vaccines to a remote area when there is not enough COVID-19 testing data, the government can determine how to allocate the vaccine by predicting that neighborhood’s infection rate based the facilities inside the neighborhood. This can greatly improve the vaccine distribution efficiency as well as the vaccine’s effectiveness. However, since different states have different lockdown policies, the above model can only be used to predict the areas in New York because it is trained only by the New York data. But this modelling technique is easily transferable to training the dataset in other locations. Finally, because of not perfect accuracy in predicting the infection rate, this model can only be used for reference.

**6. Conclusion**

We have used location data from Foursquare, Percent Positive and Test Rate of Molecular Testing by ZIP code data from NYC Health website, and US Zip code Latitude and Longitude data to predict the infection rate in each neighborhood based on the frequency of different types of venues. The accuracy of our prediction model is 0.77, which is sufficient to use for reference when examining COVID-19 spreading patterns or making government decisions to combat the virus. Some future works would be incorporating data from every state in the US to develop a mode that is more accurate, exploring other venues that can potentially affect the infection rate, and researching new ways to teach people about the danger of gathering venues and to avoid them.