**New York City neighbourhoods and their susceptibility to COVID-19**

**Introduction**

The coronavirus disease 2019 (COVID-19) is an infectious disease that has resulted in a global pandemic that has affected the lives of millions of people around the world. While the impact of this disease has been felt everywhere, some neighbourhoods have been hit harder than others. In this project, I investigated how the number and types of venues in neighbourhoods in New York City (NYC) affects the number of cases in that neighbourhood. This investigation will be important to many stakeholders like the government, hospitals and ordinary people. With the results of this study, the government will be able to identify the types of neighbourhoods which require more attention with regards to the pandemic, while hospitals in such neighbourhoods can prepare themselves for a large influx of patients and ordinary people living in such neighbourhoods can take the necessary precautions to avoid exposure.

**Data**

I used three main sources of data for this project. Firstly, I used the data from the New York City government[[1]](#footnote-1) , to obtain the number of COVID cases in each zip code. I also used publicly available data to obtain the latitude and longitude for each zip code[[2]](#footnote-2)[.](https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/)

Finally, I used the Foursquare API to obtain the data for the venues in each zip code neighbourhood using the latitude and longitude. In particular, I will collect the number of venues in each category like restaurants, parks etc. This will be used to identify what types of venues increase the risk of a neighbourhood having a large number of cases.

**Methodology**

I began by assigning the data points to two classes based on whether the covid case rate (defined as the number of cases in a zip code area divided by its population) is above or below the median rate. This allowed me to produce 2 classes of equal size. Using this I could produce a map of New York City with each postal code area marked as low (green) or high (red).

Map

Description automatically generated

I then chose three categories for the venues, namely, stores, restaurants and entertainment and found the proportion of each category in each of these postal codes. For example, in the case of restaurants, I included any venue type with the word ‘restaurant’ present as well as any cafes. I made sure that at least 10% of the venues in each postal code were captured in these categories so that the data used makes sense in this context. I also included the population of each postal code region as a feature. The assumption I made here is that each postal code region in NYC has roughly the same area and so the population is a good estimate of the population density.

I then built a decision tree classifier, a gradient boosted classifier and logistic regression classifier based on this data. For the 2 decision tree based classifiers, feature scaling is unnecessary as they are invariant under monotonic transformations of the data. For the logistic regression classifier, I scaled the data using the standard procedure of finding the number of standard deviations away from the mean of each point.

**Results**

Having trained all three classifiers discussed above, I cross-validated them using the five-fold cross validation procedure. I obtained the F1 score of each classifier as listed below:

|  |  |
| --- | --- |
| **Classifier** | **F1 score** |
| Decision tree | 0.68 |
| Gradient boosted decision tree | 0.65 |
| Logistic Regression | 0.59 |

It can be seen that all three classifiers have obtained F1 scores of greater than 0.5 indicating that they have had some success in classifying the data points into the correct classes. The highest success was achieved by the simple decision tree classifier with an F1 score of 0.68 closely followed by the boosted decision tree. This could be because the size of my data set was quite small and only 4 features were used, thus limiting the success of the gradient boosting and logistic regression algorithms.

The decision tree that was obtained can be seen below:**Diagram

Description automatically generated with medium confidence**

**Discussion**

One can get some interesting findings from the above tree. Firstly, we see that, while both classes are initially balanced, there is a significant split at the root node based on whether the entertainment venues comprise of less than or greater than 6.8% of the venues considered. This shows that even a relatively small proportion of entertainment venues can lead to an increase in covid cases. The largest class labelled as ‘high’ contains 38.8% of the initial data and has the properties that the entertainment proportion is larger than 6.8% and the store proportion is between 1.5 and 24.6%. This implies that this class contains a a large number of restaurants and entertainment venues suggesting that these factors have a significant impact on the covid case rate.

Now, examining the largest leaf node with the label ‘low’, we see that it contains the data points with entertainment venues with a proportion of less than 6.8% and stores with a proportion of less than 55.1%. This implies that the restaurant proportion is greater than 38.1%. Hence, we see that restaurant proportion is not as important when determining covid case rate when compared to the entertainment proportion.

This analysis can be extended in many ways to obtain more accurate and reliable results. Firstly, the current study only looked at areas in New York City. By extending this analysis to the rest of the country and even the rest of the world, we would obtain more data points, which should make the classifiers more reliable. Also, by doing this, one may also investigate the differences in these results for different countries. Furthermore, one could include many more features such as the demographic details of each area as well as various other venue categories. This will allow for a more detailed analysis of the factors that affect the COVID case rate.

**Conclusion**

From the analysis in this investigation, we may conclude that entertainment venues pose a significant risk for COVID transmission. As was found using the decision tree classifier, the areas with the proportion of entertainment venues being more than even 7% have a significantly larger probability of having a high case rate. This shows that the government should place a larger focus on such areas when combatting this pandemic.

1. <https://www1.nyc.gov/site/doh/covid/covid-19-data.page> [↑](#footnote-ref-1)
2. <https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/> [↑](#footnote-ref-2)