COVID-19 Short-Term Prediction Using Socio-Demographic and COVID-19 Statistical Data

Case Study Applied on New York Counties

# Summary

In this study, we use machine learning to correlate between socio-demographic data and the spread of COVID-19 within a community. Finding such correlation will be critical to predict the expected number of positive COVID-19 cases. Given that number of cases is highly dependent on the number of testing, we use in this study statistical COVID-19 data which include current infection rate, current infection density, etc. Predicting the expected number of COVID-19 within a specific area will help officials and the health community to better prepare for upcoming outbreak.

The aim of this study is use socio-demographic data along with COVID-19 statistical data from the previous 3 days to predict the next day number of new COVID-19 cases.

# Attributes:

|  |  |
| --- | --- |
| Attribute | Description |
| Density | Population Density |
| Education | Percentage of population 25 years and older having a high school degree or higher |
| Unemployment | Unemployment rate |
| Sex Ratio | Ratio of Male to 100 Female |
| Age Median | Median of population age |
| Public Transportation Commute | Percentage of workers that use public transportation as a way to commute (excluding taxicab) |
| Infection Rate | Current COVID infection rate (number of cases/population) |
| Infection Density | Current COVID infection density (number of cases/area) |
| Average Increase | Average daily percentage of COVID cases increase (calculated from the previous 4 days) |
| Covid1 | Number of COVID cases on day 1 |
| Covid2 | Number of COVID cases on day 2 |
| Covid3 | Number of COVID cases on day 3 |
| Covid4 | Number of COVID cases on day 4 |
| Covid5 | Number of New COVID cases on day 5 (To predict) |

# Limitation

This version is intended to show that it is possible to correlate the socio-demographic traits with the spread of COVID-19. In addition, this serves as an initial proof that machine learning can be used to predict the future numbers of new COVID-19 cases. We cannot draw strong conclusions from the results of this model given the limitation of data used. Positive numbers of COVID cases depends on the number of COVID tests being performed. In addition, timeseries forecasting could be implemented when having more timeseries data.

Hence, this version of the project is intended to show a proof of concept of the use of machine learning and socio-demographic data to predict infection rate/number of covid-19. To enhance this project furthermore, additional effort needed to:

* Gather Data for all US counties
* Extract Testing Rate per county
* Extract more daily COVID number of cases (and predict next-day, next-2-day and next-3-day cases)
* Extract Hospitalization Rate of COVID-19 patients to predict the expected future cases in need for hospitalization)

# Method

We aim to use machine learning to predict the next-day number of new COVID-19 cases. This is a regression problem.

## Normalization

Min-Max Normalization will be done on the following attributes:

* Density
* Infection Density
* Infection Rate

## Technique

As a proof of concept, we will test as a start two regression techniques that are able to give some insights even when having small number of data points:

* Logistic Regression
* Support Vector Machine (SVR)

We will perform 10-folds cross validation to make sure that every point will have the chance to appear in training and testing.

## Performance Evaluation

To evaluate our model, we plan to ……..

# Data

As a test case, we will apply our methodology on New York counties. Note that Bronx, Kings, New York county, Queens and Richmond are represented by New York City. Hence, we have a total of 58 case.

Socio-Demographic data is collected from the Census Bureau (latest available version)

COVID-19 Daily cases are scribed from the Department of Health of the State of New York (<https://coronavirus.health.ny.gov/county-county-breakdown-positive-cases>)

# Results