A Report on the Data Science Process

Aka'sh L. Carver, Chandu Malgari, Sytiva Wheeler, Chandan Chunduru

Computer Science, UNC Greensboro

Abstract

Coronavirus is a continuing worldwide pandemic, which has affected a lot of people including you. Our goal of the project is to develop an analytical framework to study the data coming from United States to understand patterns of COVID-19 effect and spread. This is a consummate report on our technical process and findings.

Keywords: Machine Learning, Data Science, Artificial Intelligence, Big Data, Data Analytics

1. Stage 1: Data Understanding

Preliminary data access and comprehension including reading the data dictionaries and menial data analysis, setting up programming environments. The datasets used were premanaged and we examined the best processes for merging and manipulating the data with python and the numpy and pandas library.

2. Stage 2: Modeling

Modeled data distributions using Poisson distributions and probability mass functions for the USA. Weekly statistics such as kurtosis, mode and mean were taken for the country's new Covid-19 cases and new Covid-19 deaths. This data was then compared to 5 other countries of similar population. Daily new case and death trends for these countries were compared to the USA's daily trends after the data was normalized using the population information available. Both normal and log values were used. The week in which the cases and deaths peaked were found and compared to the other countries chosen previously.

Each member then chose a state to analyze on their own. For these states, weekly statistics similar to the entire US statistics previously found were generated and these data were compared against other states. Weekly statistics of these states new cases and deaths were described, normalized by population, then compared against other states statistics. Then, the state's daily new case and death trends as well as it's top 5 infected counties trends were plotted using normalized data and log normalized values.

Each member, using a state of their choosing, fit a distribution to the new cases caused by Covid-19 in that state. The distributions statistics as well as modality were also described with it, then compared to 5 other states. A Poisson distribution for Covid-19 cases and deaths is modeled for the state and compared to 5 other states. Then, Poisson distribution is modeled for North Carolina counties' Covid-19 new cases and deaths per 100,000 populations. Correlation is performed to identify any patterns between enrichment data variables and Covid-19 and a hypothesis is formed based on that correlation which will be tested in the next stage of the project.

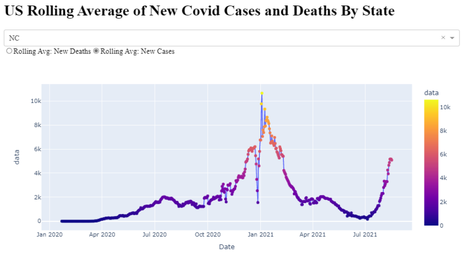
3. Stage 3: Basic Machine Learning and Using Hypothesis Tests

The goal of Stage III is to utilize machine learning and statistical models to predict the trend of COVID-19 cases / deaths. Our team developed linear and nonlinear regression models for the USA Covid-19 new cases and deaths. Using these regression lines, a Root Mean Square Error, or RMSE was calculated while discussing the trade offs of bias/variance error minimization. A trend line is then plotted along with a prediction for the next week's new Covid-19 cases and deaths and the trends are compared to other countries new case and death trends.

Each member chose a state to develop linear and non-linear regression models for and calculated RMSE's with them. Then they identified the counties most at risk with regression models and described their trends. Hospital data, specifically ICU beds, were used to find a states "point of no return" or a point at which they would have no more available hospital beds. Members then used the occupied beds or the amount of Covid Deaths to see if their state was close to this point. Next, the hypotheses created in the previous stage were tested using either Chi-Square tests or two sample and one-sample t-tests. Each member then plotted graphs for each of the aforementioned analyses.

4. Stage 4: Visualization and Dashboard

The final stage's goal was to develop interactive dashboards based on all our previous stage's analysis. We created an interactive graph that showed the trends of Covid-19 cases and deaths while allowing for a linear and log version of the cases to be shown on the graph. The best fitting regression model prediction trend line, both linear and non-linear, are incorporated in the graph. An interactive graph showing the trend line of the moving average was created that allows for selection of different states. A Choropleth map was made showing the USA county level cases and deaths, which is normalized by the overall mean of county populations.



US Rolling Average of New Covid Cases and Deaths By State

NC

○ Rolling Avg: New Deaths ● Rolling Avg: New Cases

Figure 1: US Rolling Averages

References