Project Stage III Report

Contributors: Amantii Samson, Alejandro Penaloza,
Peter Yuan, Ritu Joshi and Quinn Tjin-a-soe

Goals

The goal of Stage II is to utilize machine learning and statistical models to predict the trend of COVID-19 cases / deaths.

Task 1:

- Team:
    - Develop Linear and Nonlinear (polynomial) regression models for predicting cases and deaths in the US.
        - Start your data from the first day of infections in the US. X-Axis, number of days since the first case, Y-Axis number of new cases and deaths.
        - Aim to predict 1 week in advance. Use older data to validate your models. Use Root Mean Square Error (RMSE) to see the evaluation.
        - Describe the trends as compared to other countries.
- Member:
    - Utilize Linear and Nonlinear (polynomial) regression models to compare trends for a single state and its counties. Start your data from the first day of infections. X-Axis, number of days since the first case, Y-Axis number of new cases and deaths. Calculate error using RMSE.
    - Identify which counties are most at risk. Model for top 5 counties with cases within a state and observe their trends.
    - Utilize the hospital data to calculate the point of no return for a state. Use percentage occupancy / utilization to see which states are close and what their trend looks like.
    - Utilize decision tree, random forest, and ARIMA (https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/) based models to compare and contrast the performance of each. X-variable is the previous weeks data, Y-variable is current week data. So you are predicting one week in advance. Calculate RMSE error. This is just based on the number of cases.
      Ritu Joshi - A regression model was created for each machine learning algorithm. Out of those ARIMA models performed well with least RMSE.
      Peter Yuan - A regression model was created for each machine learning algorithm. I use all the data except the last two weeks' data to test. The RMSE for polynomial regression , random forest, and decision tree are close, around 17000, and the decision tree is little better than the other two models. The RMSE for ARIMA is around 21000.

- Use 5 different variables from the enrichment data to predict the spread rate (cases and deaths) of COVID-19 in a county. Compare Random Forest and Decision Trees (RMSE error).
    - For example, percentage of population in a certain age-group, socio-economic status, public transportation, work from home, etc.
    - Show the relative importance of variables and explain why.

Ritu Joshi - Hospital Beds data merged with the COVID-19 dataset was used to perform further operations. In this task, the complete dataset was grouped by county and aggregated. Next, the whole dataset was normalized by population, text variables were ignored. 5 features were extracted which became the X variable for training and testing. "Conf_diff: variable was used as an independent variable and used for both training and testing.

After normalizing and dropping all the unnecessary variables, the dataset is split into 70% training and 30% testing datasets. Decision Tree regressor and Random forest regressors were used for modeling the data. RMSE was calculated for both Confirmed cases and deaths and the results are presented below:

RMSE for confirmed cases - Random Forest: 97.222668, Decision Tree: 132.287868

RMSE for deaths - Random Forest: 8.303236, RMSE for deaths - Decision Tree: 10.123922

Peter Yuan- I have selected 5 variables as following:

1. DP03_0001E Estimate!!EMPLOYMENT STATUS!!Population 16 years and over
2. DP03_0028E Estimate!!OCCUPATION!!Civilian employed population 16 years and over!!Service occupations
3. DP03_0080E Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS)!!Families!! 35,000*to* 49,999

4. DP03_0084E Estimate!!INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS)!!Families!! 150,000*to* 199,999
5. DP03_0098E Estimate!!HEALTH INSURANCE COVERAGE!!Civilian noninstitutionalized population!!With health insurance coverage!!With public coverage

Since the y variable is also the proportion of case and death. So the RMSE is a float number. I have builded 4 models, 2 random forest trees and 2 decision trees to predict the case and death separately. The RMSE of random forest is lower than the decision tree , so in my case, the random forest fits better.

In the meantime, I use method feature_importances_ to test the importance of each variable. According to the result, the feature 1 is relatively important to case number, the feature 2 is relatively important to death number it does make sense, because no matter if you are rich or not, it won't decide if you get infected, but if you work in public then you get a higher chance to meet the infected person.

On the other hand, if you got infected. Then the income will decide how you get treated, so this is important to the death.

Amantii Samson-

Using the Hospital Beds enrichment dataset, I grouped the dataset using the states to organize all the data and then I selected the five variables that I would be using to predict the spread of cases and deaths within a state. The five variables that I selected to predict the spread of cases and deaths were: NUM_LICENSED_BEDS, NUM_STAFFED_BEDS, NUM_ICU_BEDS, BED_UTILIZATION, and AVG_VENTILATOR_USAGE.

These variables appeared to me to be the most effective means of analyzing the hospital beds dataset in relation to each state. The varying values within each variable, along with the importance of each factor in a real hospital, provide a more robust analysis of the dataset for each state.

Alejandro Penaloza

The 5 variables I used were from the ACS Sex and Age Dataset and they were as follows:

1. **Estimate!!Total!!Total population**
2. **Estimate!!Male!!Total population**
3. **Estimate!!Female!!Total population**
4. **Estimate!!Total!!Total population!!AGE!!15 to 19 years**
5. **Estimate!!Total!!Total population!!AGE!!20 to 24 years**

I chose these 5 because they're all relevant within each other. When you look to analyze the covid spread rate in a specific county, you want to know the total population in that county, and when getting more into specifics you'd want to know which chunk of this population is female, male, age 15-19, and age 20-24. It's good to know the rates at which they spread within these specific sex and age groups!