PSTAT 197A Final Project

By Andy Nguyen, Kankshat Patel, Johnny Yu, and Nishant Yadav

Abstract & Introduction

Goal: To build a predictive model using past data that would forecast the future COVID 19 cases.

Use Case: Given recent events and the emergence of new variants, modeling the expected spread of COVID19 would be helpful as we try to end this pandemic.

Challenges: Distinguishing which features are useful, dealing with null values in the dataset.

Data Collection and Preparation

1. Collection
   a. Utilized data relevant to confirmed covid cases
   b. Strayed from data with many NA/missing values
   c. Mobility data helpful for long term (7 Days+)
2. Prep
   a. Impute NA/missing values with daily average value across all counties
   b. Shifted data by one day to prep for time-series forecasting model
3. Processing
   a. Set date to represent split of training and testing data
   b. Split data such that data remained sequential rather than random sample

Model Training and Evaluation

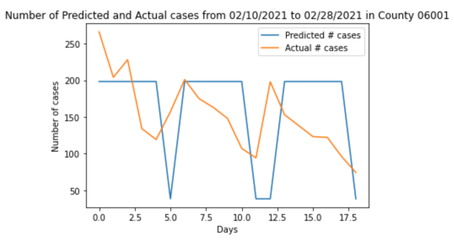

Figure 1: Predicted vs. Actual COVID cases with LR model

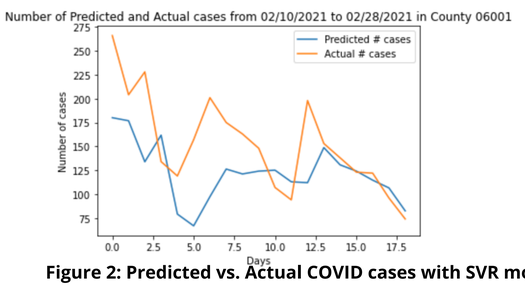

Figure 2: Predicted vs. Actual COVID cases with SVR model

Both Figures 1 and 2 represent a plot of our predicted COVID cases vs. actual COVID cases for a specific county. We can see that for the same county, the simple model does a much better job at predicting the actual number of COVID cases as the predicted plot mirrors the true plot a lot better than the SVR model.

Conclusion/Next Steps

Findings:

- Linear Regression performed better than Decision Tree Regressor, a 92% score vs. a 88% score.
- Both simple models performed better than the complex SVR model which resulted in a 36% score.
- We attribute the discrepancy between the simpler and complex values to the lack of overfitting in the simpler model.

Next Steps:

- Impute data with more specifics
- Find additional time-series relationships of past and present variables
- Increase amount of data
- Deeper dive into more models

References

https://towardsdatascience.com/time-series-modeling-using-scikit-pandas-and-numpy-682e3b8db8d1

http://www.kasimte.com/2020/02/09/linear-regression-from-time-series-data-using-scikit-learn.html

https://towardsdatascience.com/how-to-model-time-series-data-with-linear-regression-cd94d1d901c0

Coordination

We split the data preparation as well as the data processing portion equally and worked equal amounts on the poster and summary