COVID-19 Forecast

Atherv Gole, Natasha Leodjaja, Nathan Roll, and Romina Fareghbal

Abstract & Motivation

The COVID-19 pandemic has led to a greater need for time based modeling of infectious disease. To study the rate of change in COVID-19 cases in California and the factors that are involved in determining the number of cases, there are a number of methods that can be used. The analysis focus in this project is a comparison between Simple Linear Regression and a more complex Support Vector Regression using a non-linear RBF kernel. Using the signal data collected from COVIDcast, the models were cross validated by breaking the time series data into segments five times with increasingly larger training sizes while maintaining chronological order. To evaluate performance of the models, mean squared errors and goodness of fit were compared as well as the testing predictions and ground truth COVID incidence values. It seems that between the two models, the linear regression model provides a compromise between model performance, training and prediction time, and interpretability.

Technical Details

Our five-feature labeled dataset (available here) was synthesized from the COVIDcast API. Missing data was handled through replacement, based on previously recorded data per county. An 85/10/5 train/test/validation split was performed prior to training the OLS model and the 5-step decision tree classifier. All preparation, modelling, and analysis were conducted in python 3.7 with the sklearn module. Visualizations were generated using matplotlib and plotly.

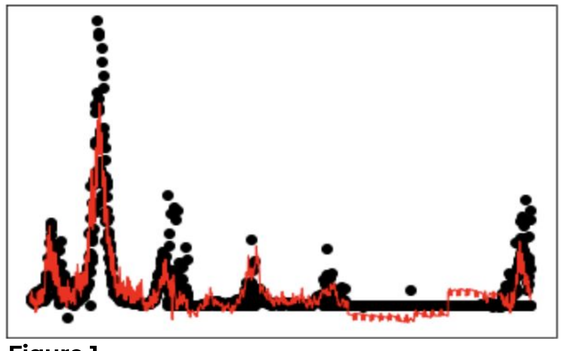


Figure 1

Research Questions

- Given the past data, can we build predictive models that forecast COVID cases?
- Which features are relevant to the prediction and how should that affect our policies?

Methods

Linear Regression
Feature Importance
Support Vector Machine (non linear)

Figure 2

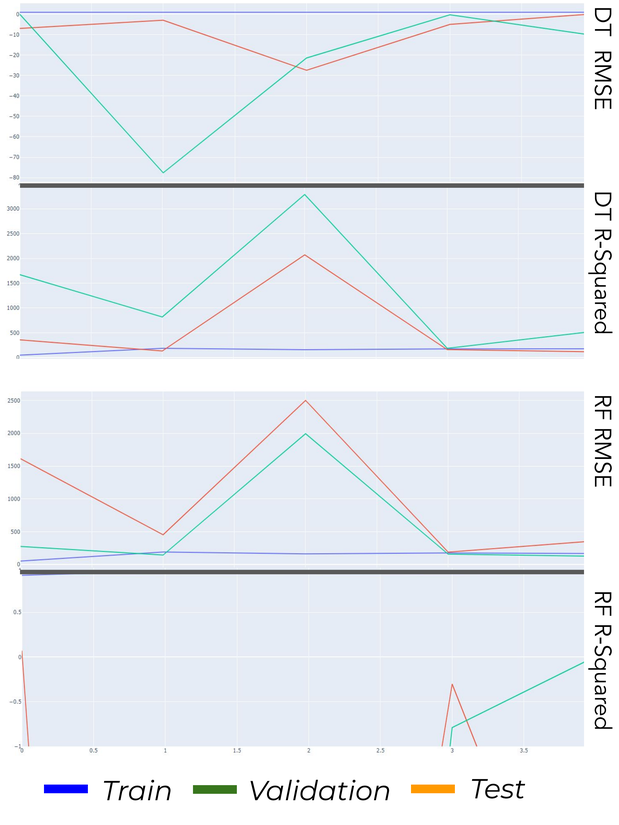
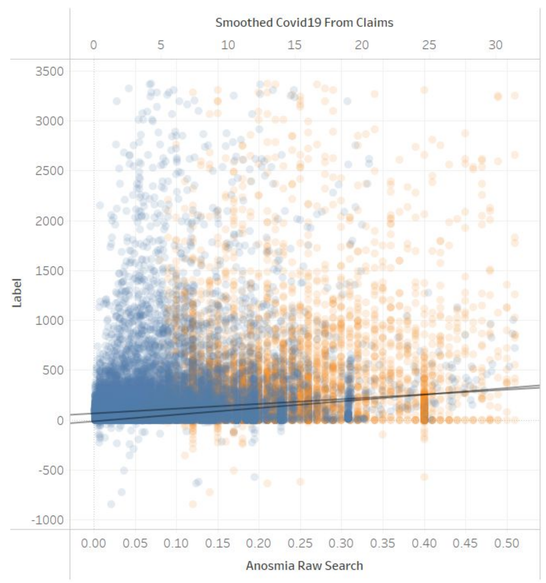


Figure 3



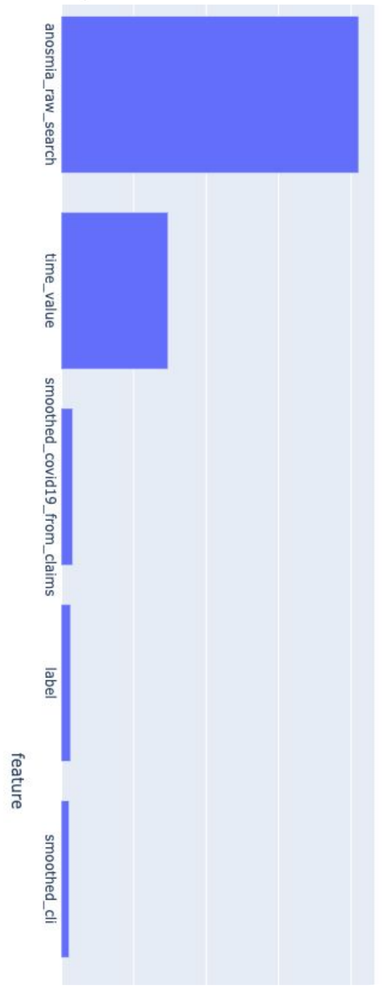
Conclusions

- The performance of both models is fairly similar but the peaks in predictive accuracy are at different portions of the cross validation time series splits.
- The mean squared error for the linear regression model is higher than the MSE for the SVM, which can be expected due to the lower flexibility of the model, resulting in higher error for portions of the data that are closer to polynomial trends than linear.
- In this case, the simpler model seemed to provide a better overall fit. This is a valid case since we are dealing with effects over time following a general downward trend.

Future Work

Potential future improvements to the model could involve feature reduction through AIC/BIC or similar criterion for non-linear models. A polynomial regression model might have also provided a better fit especially as rates of increase and decrease changed throughout time.

Figure 4



Coordination

1. Atherv: covidcast, data preparation, linear regression, support vector machine, poster, and summary.
2. Natasha: covidcast, feature importance, and poster.
3. Nathan: data preparation, support vector machine, and poster.
4. Romina: support vector machine and poster

References

[1] Brownlee, Jason. "Feature Selection for Time Series Forecasting with Python." *Machine Learning Mastery*, 15 Sept. 2020, https://machinelearningmastery.com/feature-selectio-time-series-forecasting-python/.

[2] Harmon, Clay. "Covid-19 Time-Series Analysis with Pandas and Python: The Grim RIFR." *Medium*, The Startup, 10 Dec. 2020, https://medium.com/swlh/covid-19-time-series-analysis-with-pandas-and-python-the-grim-rifr-b078e6a327ca