Forecasting Short-term Future COVID19 Cases Based on Historical Data

Tristan Chen, Jessica Nguyen, Hera Chan

Abstract and Introduction

- At the start of the COVID-19 outbreak, our underprepared knowledge of the virus led to millions of deaths. Two years have passed since the outbreak, and now with more data and resources available, we aim to forecast future patterns of the pandemic in order to be more readily prepared for the future.
- Our objective is to build and train time series forecasting models to predict the number of cases in the short term future
- If successful, models such as these will allow us to prepare adequate resources and supplies in advance, possibly decreasing the number of COVID-19 cases and deaths
- It is challenging to accurately predict the spread of COVID-19 due to the amount of factors that play a part in the pandemic. In addition, because the pandemic is so widespread, there are bound to be some inconsistencies in the data.

Data Collection and Preparation

- Our data was collected from the COVIDCast Epidata API, which provides data on COVID-19 from multiple different sources
- Only data from 02/22/2020 to 09/27/2020 for some counties in California were collected
- For our ground truth, we collected the daily proportion of new cases per 100,000 people
- The 5 features that we collected were:
  - Estimated % of doctor visits about COVID Symptoms
  - Estimated % of doctor visits with confirmed COVID cases
  - Estimated number of new hospital admissions with COVID associated diagnoses based on medical records
  - Estimated number of new hospital admissions with COVID associated diagnoses based on claims
  - Daily proportion of new cases per 100,000 people
- We then dropped the rows with missing data

References

[1] Delphi Research Group. (2020). *Data sources and signals*. Delphi Epidata API. Retrieved December 4, 2021, from https://cmu-delphi.github.io/delphi-epidata/api/covidcast_signals.html

[2] Sher, V. (2021, March 24). *Time series modeling using Scikit, Pandas, and Numpy*. Medium. Retrieved December 4, 2021, from https://towardsdatascience.com/time-series-modeling-using-scikit-pandas-and-numpy-682e3b8db8d1

Model training and Evaluation

Feature Importance (Figure 1)

- Calculated feature importance by performing grid search cross validation on the random forest model
- The features outpatient_covid (t-2), confirmed_cases_prop (t-1), and outpatient_covid (t-1) were the 3 most important features

Random Forest (Figure 2, Figure 3)

- Performed cross validation on our random forest model using GridSearchCV and TimeSeriesSplit to find the best hyperparameters
- Fitted the model using these parameters on 90% of our dataset as the training and validation set

SVR (Figure 4)

- Found best parameters using similar method to the random forest model
- Fitted the SVR model on these parameters, using the same split previously used

Model Evaluation (Figure 5)

- Training, validation, and testing performances were measured using mean squared error
- The random forest outperforms the SVR model

Conclusion

- Our random forest and SVR model were able to predict the general trend of the model, however the models had trouble with predicting the spikes
- However, our models predict better for some counties, and worse for others
- In addition, our interpretable model unexpectedly outperformed our complex model.
- Building the models were not difficult, but preparing a consistent dataset that the models could accurately train on was challenging. We had to drop counties with missing values to increase accuracy of results.

Future Work

- Try fitting data using other models, such as regression or neural networks
- Create models for forecasting COVID-19 cases for different counties, states, and time ranges

Visualization

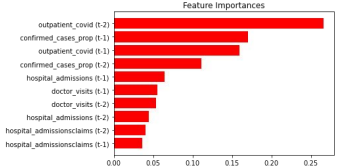

Figure 1

Visualization (Cont.)

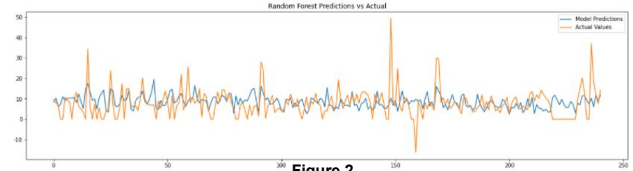

Figure 2

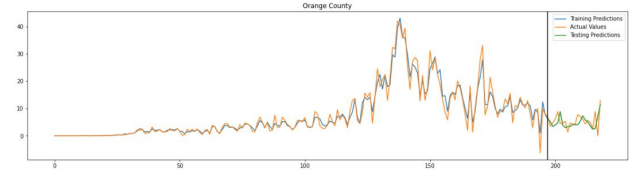

Figure 3

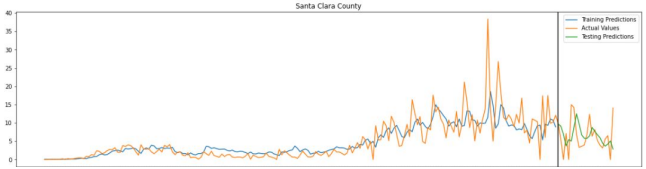

Figure 4

	Train Score	Val Score	Test Score
Random Forest	3.757635	75.708820	48.578106
SVR	18.209154	121.085624	65.281242

Figure 5

Coordination

- Members primarily used Discord and Facebook Messenger to communicate.
- Jessica and Tristan collected and prepared the Data. Jessica and Hera worked on the Random Forest model while Tristan worked on the Support Vector Machine model.
- All members worked together for the poster and summary.