Predicting Infection of COVID-19 in California

Kankshat Patel, Andy Nguyen, Nishant Yadav, Johnny Yu
Fall 2021 Data Science Capstone

1. Abstract

Effective screening of COVID-19 enables a quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed and aim to assist medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. The primary goal of the research is to develop an early prediction tool for the dissemination of new coronaviruses known as the COVID-19. The established machine-learning approach trained the model on past COVID-19 data to predict COVID-19 test results with high accuracy. Our framework can be used, among other considerations, to prioritize testing for COVID-19 when testing resources are limited.

2. Methodology

2.1 Dataset

Data was extracted from the covidcast API between the dates of March 1, 2020 and March 1, 2021, as the dates encapsulated many peaks of COVID-19 cases throughout the state of California. The utilized data focused on confirmed COVID-19 cases rather than possible exposures, such as positive COVID-19 cases, hospital admission due to COVID-19 related symptoms, and outpatient doctor visits with a confirmed COVID-19 diagnosis.

2.2 Data Preparation

Rather than omitting observations with missing values, imputation was utilized. The mean for each feature for each day replaced missing values for the corresponding date. Furthermore, the data was shifted per county such that the values for the previous days were appended to each observation in sequential order. As shown in fig 1, one observation for the date of 2020-03-04 contained values for 2020-03-04 as well as 2020-03-05.

time_value	change_health	hospital_admit	doc_visits	ground_truth	ground_truth+1	change_health-1	hospital_admit-1	doc_visits-1
2020-03-02	0.013988	0.119067	0.000000	0.0	1.0	0.011162	0.119646	0.000000
2020-03-03	0.012357	0.119366	0.000000	1.0	0.0	0.013988	0.119067	0.000000
2020-03-04	0.011438	0.119776	0.019267	0.0	0.0	0.012357	0.119366	0.000000
2020-03-05	0.010639	0.119895	0.016927	0.0	0.0	0.011438	0.119776	0.019267

Figure 1

An interesting caveat was the confirmed number of cases per day per county. To accurately predict the number of cases in the future and make our model valuable in any sense, the confirmed number of cases for the present day as well as the future day was a required feature to obtain tangible results, as depicted in fig 2 below. Following the split of data into training and validation, the data was fit into a decision tree regressor model and evaluated on the testing data.

geo_value	time_value	change_health	hospital_admit	doc_visits	ground_truth	ground_truth+1	change_health-1	hospital_admit-1	doc_visits-1
6001	2020-06-10	0.085265	1.013521	2.383065	48.0	86.0	0.102985	1.094606	2.540969
6001	2020-06-11	0.070310	0.879963	2.287956	86.0	97.0	0.085265	1.013521	2.383065
6001	2020-06-12	0.083679	0.744354	2.288105	97.0	54.0	0.070310	0.879963	2.287956
6001	2020-06-13	0.070867	0.749456	2.330506	54.0	50.0	0.083679	0.744354	2.288105
6001	2020-06-14	0.095950	0.730906	2.460149	50.0	53.0	0.070867	0.749456	2.330506
6001	2020-06-15	0.103072	0.847577	3.259097	53.0	108.0	0.095950	0.730906	2.460149

Figure 2

2.3 Model setup

2.3.1 Decision Tree

Decision trees, a specific type of machine learning, are based on covariates to create a model for predicting outcomes. Currently, artificial intelligence, including decision tree modelling, is being used in the COVID-19 pandemic for early detection and diagnosis, monitoring treatment, tracing contacts, developing drugs and vaccines, predicting cases and fatalities and even identifying the most vulnerable groups.

The decision tree model generated by the cross-validation training data yielded resulting performance metrics such as R-squared and mean squared error were extracted from the cross-validation results. Our cross validation was run with python's SciKitLearn function TimeSeriesSplit as our data samples were at fixed time intervals, rendering the traditional train_test_split function useless as our data must not be randomly shuffled.

Tuning the depth parameter of our tree within the ranges of 2 to 17 and running a TimeSeriesSplit cross validation on that set yielded an optimal depth of 5, as shown below in fig 3. Running and testing our model on a tree with depth equal to five results in an r-squared score of 0.88. We see the predicted number of COVID cases versus the true number of COVID cases in fig 4 below.

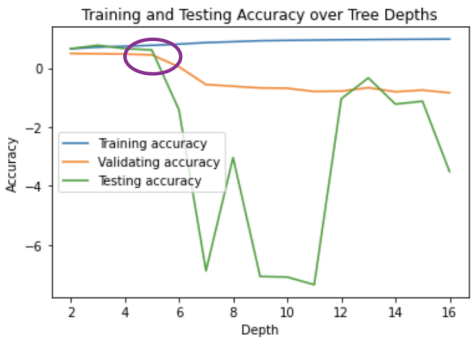

Figure 3

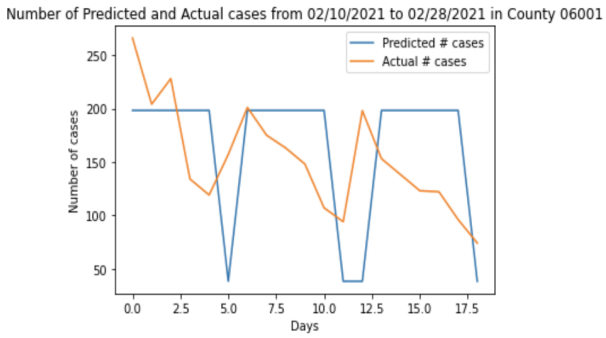

Figure 4

2.3.3 Support Vector Regression

The SVR is a type of ML managed algorithm for regression classification. SVR depends on a variety of statistical functions as a non parametric technician. The set of the kernel function converts data input into the form you like. In order to overcome regression problems using a linear function,

SVR maps the vector(x) input(s) in the n-dimensional space called the function space(z) when dealing with non-linear regression problems.

After running TimeSeriesSplit on SVR to determine the ideal kernel, we found the rbf kernel, C value of 100, and gamma value of 0.1, yielded the highest r-squared value of 0.37. Fig 5 below outlines the predicted number of cases in Alameda county throughout one month.

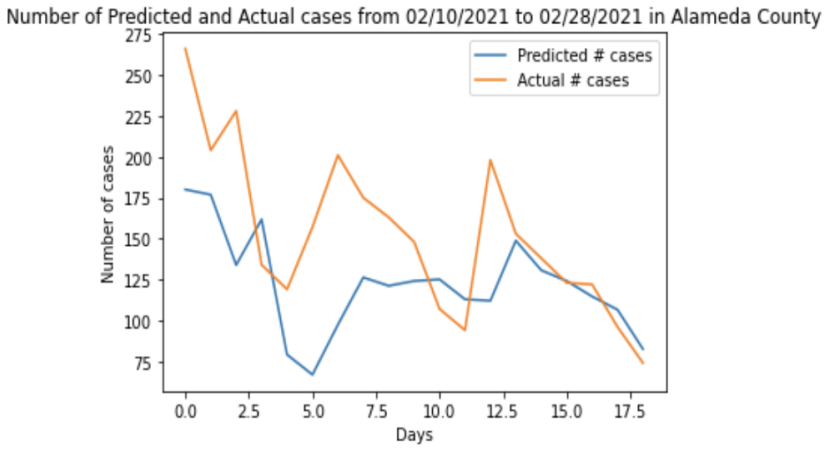

Figure 5

3. Results

3.1 Findings

In this project, the possibility of a COVID-19 outbreak was suggested as a global ML-based predictive method. The framework analyzes data sets containing real data from the past day and uses machine learning algorithms to make forecasts about future days.

The decision tree model was effective in predicting the number of COVID-19 cases in the upcoming days and struggled to follow the spikes of COVID-19 cases. Given a prediction period of 17.5 days, the decision tree model has a large MSE as the predicted number of COVID-19 cases follow the general trend of the true number of COVID-19 cases, yet differ significantly from the actual values. The simple explanation is that the model does not do well at predicting the exact number of COVID-19 cases but adequately predicts the general trend of cases. Furthermore, from fig 4 we see the model does well to predict the spikes in the number of COVID-19 cases which is a significant finding as the goal of the model is to predict the trend and spread of COVID-19 before it happens.

Compared to the SVR, the decision tree more accurately predicts the number of COVID-19 cases; however, the SVR model predicts the spikes in COVID-19 cases more precisely than our decision tree model, as shown in fig 5. Moreover, the SVR model solely predicts the spikes of COVID-19 cases before they happen, not the general trend of cases.

3.1 Conclusion

The correctness of the model could be increased by introducing related attributes like several hospitals, the immune system of the infected person, age of the patient, gender of the patient, steps

taken to combat the proliferation of the virus, and so on to make it completely informative. We can extend this model to a nationwide level by simply incorporating more counties into our covidcast fetching portion to predict COVID-19 cases for all regions within the United States.

The model gives results on the basis of data developed by information given by health agencies. Thus, forecasting may not be 100% accurate, but it can surely be used as a corrective measure. For future work further enhancement can be done by combining new factors and algorithms with deep learning to get more accurate results. Using machine-learning techniques within a shared database could generate predictive insights, showing the patterns in communities that precede outbreaks and helping dictate where and when lockdowns and social distancing orders should be implemented.