National University of Computer and Emerging Sciences

Project Report

Data Science

Covid-19 Detection and Forecast

Authors
K173755 Noman Anjum
K173795 Muhmmad Mustafa Manga

Supervisor
Dr. Atif Tahir

May 30, 2021

0.1 Research Goal

Coronavirus disease 2019 (COVID-19), also known as the coronavirus, or COVID, is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease has since spread worldwide, leading to an ongoing devastating pandemic.[1]

Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. Symptoms may begin one to fourteen days after exposure to the virus.[1]

Our research has two basic aims:

- Detection/Classification Of COVID 19 with the help of potential symptoms.
- Time series forecasting for number of cases of COVID 19.

With the help of correct data, we can make use of symptoms to classify if the subject is COVID Positive Or Negative. This can help in effectively screening out people for further testings and save testing equipment, which are costly and not abundantly available. Cost of COVID test in Pakistan is around Rupees 6500.[2]

Further with the help of data we can use time series forecasting algorithms to get an estimate of number of cases and make proper plannings for future. Time series forecasting has another important application; with the help of symptom's count we can predict number of cases of COVID 19. This approach can help us getting an estimate of number of COVID cases if we have proper data about symptoms, like number of reported cases of headache, short of breath and cough etc. Collecting data about the symptoms is relatively easy, and hence it can help us get an estimate for COVID cases without actually having COVID tests.

0.2 Retrieving Data

Primary data we needed for our research should essentially have samples containing different symptoms and a feature defining whether sample was diagnosed with COVID 19 or Not. To get this data we crawled and searched a number of Governmental Data Websites of different countries, including [3],[4] and [5]. However most of the websites only contained data specifying number of cases during different months between January 2020 and March 2021 along with distributions of ages and number of casualties and number of cases reported. This data which is no doubt very important, could help us train our forecasting algorithms however it couldn't help us in training our machine learning models.

Our extensive research for data collection lead us to the Website of Israel, which is run by Government Of Israel. This website has huge data regarding COVID 19, including data containing symptoms for an individual, their age and whether the individuals was diagnosed with COVID or NOT. This dataset[6] also contains the date on which the test for the COVID was conducted, this date could be helpful for us to perform time series analysis and forecasting. As this dataset got matched with our requirements, we decided to select this data with our further process.

0.3 Data Preparation

As we analyzed the dataset after downloading it, we came to know that data has thousands of missing values. Further, a alot of features and values in feature are in Hebrew language as dataset was obtained from Israel's Government website(National Language of Israel is Hebrew). In order to translate dataset in English we used Google translate(an online translating service provided by Google Inc). Since the features and variables in Hebrew were finite, we did manual translation of those values instead of using any API for the translation. Hebrew feature names and variables were then replaced by English translation using a python Library Pandas.

Second stage for data preparation was to handle missing values, Our analysis revealed that data was very huge and and samples with missing values despite being in large quantity was only 16% of original data. Hence we decided to remove samples with null values after getting insights and visualizations from the data.

Finally for time series analysis we converted our "test_date" column to pandas DateTime format and made it the index of our dataset, which is the basic requirement in order to aggregate data on Monthly/Weekly/Yearly basis.

0.4 Data Exploration

Exploration of Data provided us with some interesting and very useful insights over the COVID cases in Israel between January 2020-March 2021. Firstly we came to know about number of missing values in the data and it's ratio with the actual data which was about 16%. Secondly data contained the additional information about person's gender and whether the person was above 60 years old or not. Insights of the symptoms of COVID lead us to conclude that there were a huge number of samples with cough, fever and sore throat, however they were not diagnosed as COVID positive. Major symptoms leading to COVID were sore throat, shortness of breath and headache. It was found that the ratio of Male and Female Diagnosed with COVID was equal.Time series analysis of COVID cases signifies that trend of symptoms follows the trend in Positive COVID cases which means that there is a solid relation of COVID and theoretical symptoms defined by health WHO and other health sectors.

0.5 Model Building

Selection of Model is an important part for classification task, as different data distribution needs different learning algorithms for effective learning of features for the prediction. We used two different approaches in order to identify best Machine Learning Algorithm, which can learn patterns in our dataset effectively and come up with a good accuracy for predictions.

Multiple Models and Ensembling

We initially selected following Machine Learning Algorithms For Our Project.

- Naive Bayes Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Multi Layer Perceptron(with 32,64 nodes sequence)

Along with these models we used Voting Classifier which is an Ensembling Technique. Ensembling classifier works on "Wisdom of Crowd" approach, and chooses a label for a sample based on majority votes.

Kfold Cross Validation

In "Cross Validation" a model is trained using K-1 of the folds as training data, the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute a performance measure such as accuracy). We used a 3 fold validation on our dataset, so that our models see all of the data and we can avoid underfitting of the model.

With 3 Fold Cross Validation the best algorithm was found to be "Decision Tree Classifier" and "Ensemling Voting Classifier" with an accuracy of aroud 91%. Since Voting classifier was using Decision Tree along with Multi-Layer Perceptron Model which is computationally expensive on training, we decided to use "Decision Tree Classifier" as Classification Model and dropped the Voting Classifier.

Time Series Analysis

Our second goal of project was to forecast number of COVID cases on monthly bases. For this we transformed our data. "test_date" column from dataset was casted to date_time format. This feature was then used as index of the dataset, and dataset was aggregated on Monhtly bases using this index. With the help of aggregated data we performed Univaritate and Multi Variate Time series analysis and forecasting using ModelTime Package build in R language.

Forecasting Algorithms

- ARIMA with and without XGBoost
- xponential Smoothing (ets)
- Prophet
- MARS/EARTH
- Linear Regression (Parsnip)
- Prophet With XGBoost
- RandomForest
- GLMNET

Forecasting Algorithms Evaluation

Above mentioned forecasting Algorithms were run on Univariate and Multivariate data. For Univariate forecasting we used monthly sum of COVID cases and Date with the formula NoOfCases = test_date. Through this univariate forecasting we can get an estimate of Number of COVID cases in Future, learning the trend in dataset. This estimate can help us forming strategies to cope up with any further wave of COVID. This analysis can also help understanding trend of COVID increase with different seasons and weather conditions.

Multivariate Time series forecasting was performed with the help of symptoms count. For this the sum on symptom was calculated on monthly aggregated data. Hence above mentioned models were trained on Multivariate Data with formula NoOfCases = test_date+CoughCount+SoreThroatCount +FeverCount+HeadAcheCount+ShortnessOfBreathCount. Primary purpose of multivariate forecasting is to get an estimate of number of COVID cases with the help of symptoms count.

Out of These eight forecasting algorithms we selected the one with least MAPE(mean average precision error), For Univariate Forecasting, Prophet With XGBoost was the algorithm with least MAPE and for MultiVariate Forecasting GLMNET was the one with least MAPE . We used these two models for future forecasting.

0.6 Presentation & Automation

For the purpose of Automation and Presentation Of Project we decided to develop a web based system. Details of Our web based system is as follow:

System Description

We developed web based application for the purpose of Automation of our project. This web application will allow the dynamic retraining of model as we get more data. Hence the Graphical User Interface of the system provides an option to upload latest dataset which is in correct format and then retrain our models to capture new details from the dataset. Our application also provides a portal where a person can provide his health details(Boolean values for different symptoms), his age and gender and get a result whether he is potentially caught by the novel COVID.

Technologies

- We used a python's Fast API framework for the development of our backend. FastAPI is a modern, fast (high-performance), web framework for building APIs with Python 3.6+ based on standard Python type hints.

- Javascript framework React.js has been used as a frontend technology. React is a very popular Javascript framework for frontend, It is blazingly fast and dynamic due to it's client side rendering feature. It provides a very smooth and responsive User Interface for a pleasing user experience.

0.7 APPENDIX

References and Bibliography

[1] https://en.wikipedia.org/wiki/COVID-19

[2] https://www.shifa.com.pk/pcr/

[3]https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

[4] https://covid19.who.int/table

[5] https://ourworldindata.org/coronavirus/country /pakistan

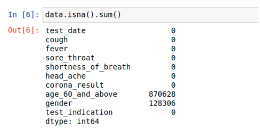
[6] https://data.gov.il/dataset/covid-19

Algorithms

Naïve Bayes : is a simple learning algorithm that utilizes Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy.
Decision tree : analysis involves making a tree-shaped diagram to chart out a course of action or a statistical probability analysis. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.
Random forest : is a supervised learning algorithm. The ”forest” it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.
Multilayer perceptron : is a class of feedforward artificial neural network. The term MLP is used ambiguously, sometimes loosely to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons.
ARIMA : In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average model is a generalization of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series.
Exponential smoothing : is a rule of thumb technique for smoothing time series data using the exponential window function. Whereas in the simple moving average the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time.
Prophet : is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data.
GLMNET : is a package that fits generalized linear and similar models via penalized maximum likelihood.It fits linear, logistic and multinomial, poisson, and Cox regression models. It can also fit multi-response linear regression, generalized linear models for custom families, and relaxed lasso regression models.

Reference Images

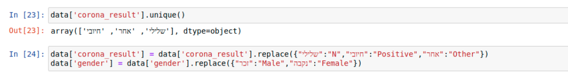
Data Preparation



```
In [6]: data.isna().sum()

Out[6]: test_date                  0
        cough                      0
        fever                      0
        sore_throat                0
        shortness_of_breath        0
        head_ache                  0
        corona_result              0
        age_60_and_above      870628
        gender                128306
        test_indication            0
        dtype: int64
```

Figure 1: Number of Missing/Null values within each column.



```
In [23]: data['corona_result'].unique()

Out[23]: array(['שלילי', 'אחר', 'חיובי'], dtype=object)

In [24]: data['corona_result'] = data['corona_result'].replace({"שלילי":"N","חיובי":"Positive","אחר":"Other"})
         data['gender'] = data['gender'].replace({"זכר":"Male","נקבה":"Female"})
```

Figure 2: Replacing Hebrew keywords with English Translation.

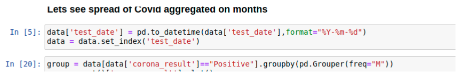


Figure 3: Aggregating Data and making Date Column as Index.

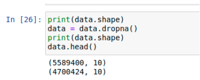


Figure 4: Number of samples before and after dropping Missing/Null values.

Visualizations and Insights Of Data

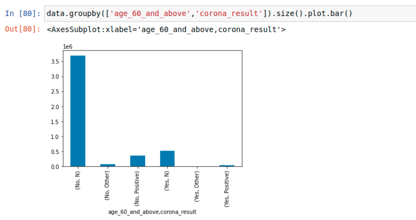


Figure 5: Plot specifying Age Factor Against COVID diagnosed.

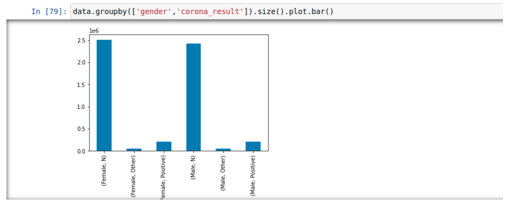


Figure 6: Plot specifying Gender Factor Against COVID diagnosed.

Figure 7: Count for symptoms for Positive and Negative Diagnosed Samples.

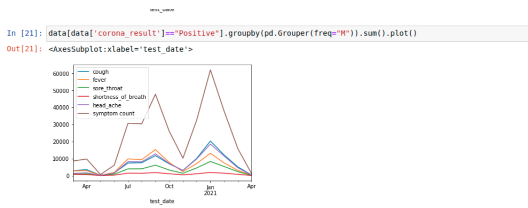


Figure 8: Trend between symptom count and Number of Positive cases.

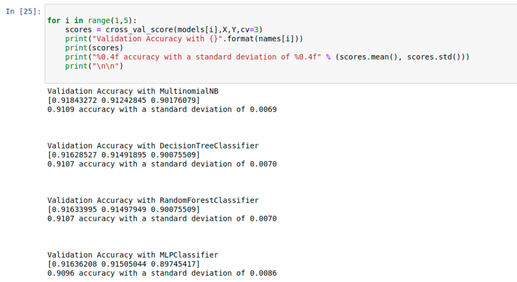


Figure 9: Mean accuracy with Cross Validation on Multinomial Naive Bayes/Decision Tree and Random Forest Classifier.

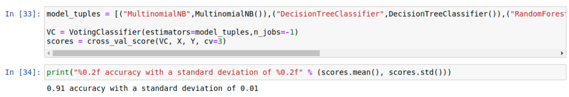


Figure 10: Cross Validation With 3 Folds On Voting Classifier.

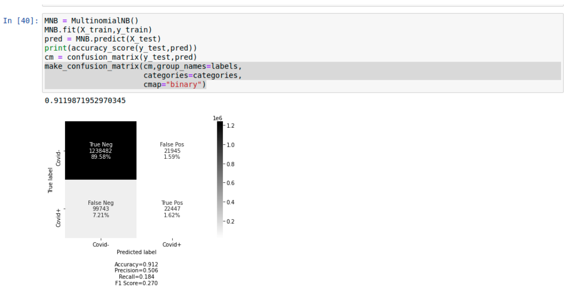


Figure 11: Testing Multinomial Naive bayes on test data.

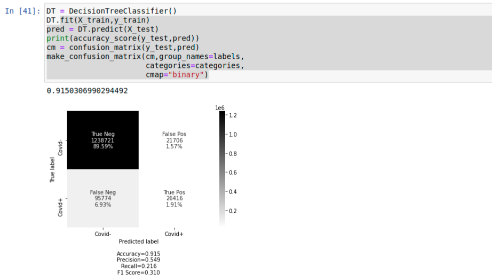


Figure 12: Mean accuracy with Cross Validation on Decision Tree.

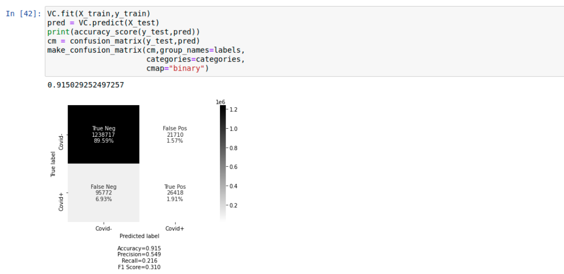


Figure 13: Mean accuracy with Cross Validation on Voting Classifier.

Time Series Model Training/Testing Results

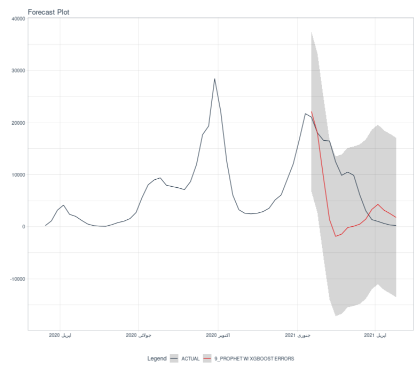


Figure 14: Training And testing result on Univariate Time Series Model.

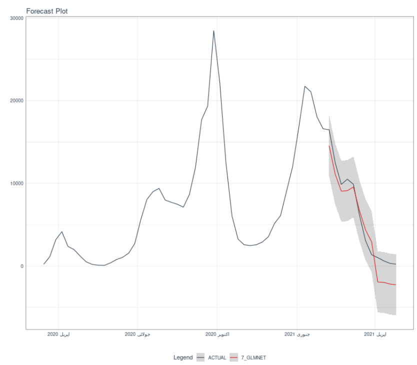


Figure 15: Training And testing result on Multi-Variate Time Series Model.

Figure 16: 6 month Forecast of Number Of COVID cases via UniVariate Time Series Model.

	A	B	C	D	E	F	G	H
36	35	NA	ACTUAL	actual	2020-11-08	2484	NA	NA
37	36	NA	ACTUAL	actual	2020-11-15	2584	NA	NA
38	37	NA	ACTUAL	actual	2020-11-22	2911	NA	NA
39	38	NA	ACTUAL	actual	2020-11-29	3559	NA	NA
40	39	NA	ACTUAL	actual	2020-12-06	5159	NA	NA
41	40	NA	ACTUAL	actual	2020-12-13	6110	NA	NA
42	41	NA	ACTUAL	actual	2020-12-20	9089	NA	NA
43	42	NA	ACTUAL	actual	2020-12-27	12079	NA	NA
44	43	NA	ACTUAL	actual	2021-01-03	16731	NA	NA
45	44	NA	ACTUAL	actual	2021-01-10	21738	NA	NA
46	45	NA	ACTUAL	actual	2021-01-17	21059	NA	NA
47	46	NA	ACTUAL	actual	2021-01-24	18035	NA	NA
48	47	NA	ACTUAL	actual	2021-01-31	16599	NA	NA
49	48	NA	ACTUAL	actual	2021-02-07	16479	NA	NA
50	49	NA	ACTUAL	actual	2021-02-14	12487	NA	NA
51	50	NA	ACTUAL	actual	2021-02-21	9884	NA	NA
52	51	NA	ACTUAL	actual	2021-02-28	10508	NA	NA
53	52	NA	ACTUAL	actual	2021-03-07	9889	NA	NA
54	53	NA	ACTUAL	actual	2021-03-14	6131	NA	NA
55	54	NA	ACTUAL	actual	2021-03-21	3063	NA	NA
56	55	NA	ACTUAL	actual	2021-03-28	1372	NA	NA
57	56	NA	ACTUAL	actual	2021-04-04	1026	NA	NA
58	57	NA	ACTUAL	actual	2021-04-11	639	NA	NA
59	58	NA	ACTUAL	actual	2021-04-18	338	NA	NA
60	59	NA	ACTUAL	actual	2021-04-25	242	NA	NA
61	60	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-05-02	-323.14900418488	-15640.410143267	14994.112134897
62	61	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-05-09	-765.12758862736	-16082.38872771	14552.133550455
63	62	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-05-16	-1062.5742673635	-16379.835406446	14254.686871719
64	63	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-05-23	-916.31143822911	-16233.572577311	14400.949700853
65	64	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-05-30	-145.31580015036	-15462.576939233	15171.945338932
66	65	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-06-06	-164.34027070763	-15481.60140979	15152.920868375
67	66	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-06-13	-126.11106498736	-15443.37220407	15191.150074095
68	67	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-06-20	214.62506183985	-15102.636077242	15531.886200922
69	68	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-06-27	2065.9717288115	-13251.289410271	17383.232867894
70	69	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-07-04	3994.3959054342	-11322.865233648	19311.657044516
71	70	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-07-11	6661.238798816	-8656.0223402662	21978.499937898
72	71	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-07-18	8148.4863334498	-7168.7748056324	23465.747472532
73	72	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-07-25	8679.3638971318	-6637.8972419504	23996.625036214
74	73	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-08-01	7246.4117733362	-8070.849365746	22563.672912418
75	74	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-08-08	7008.1855571087	-8309.0755819734	22325.446696191
76	75	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-08-15	6570.5681454313	-8746.6929936509	21887.829284513
77	76	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-08-22	6561.9431187978	-8755.3180202843	21879.20425788
78	77	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-08-29	7521.8103940518	-7795.4507450304	22839.071533134
79	78	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-09-05	10042.269821997	-5274.9913170851	25359.530961079
80	79	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-09-12	15250.169179748	-67.091959333944	30567.43031883
81	80	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-09-19	22216.796011486	6899.5348724039	37534.057150568
82	81	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-09-26	25640.436645165	10323.175506083	40957.697784247
83	82	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-10-03	22212.504720617	6895.2435815343	37529.765859699
84	83	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-10-10	14690.205892681	-627.05524640166	30007.467031763
85	84	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-10-17	6718.4085596429	-8598.8525794393	22035.669698725
86	85	9	PROPHET W/ XGBOOST ERRORS	prediction	2021-10-24	2007.2347693198	-13310.026369762	17324.495908402

Figure 17: 6 month Forecast of Number Of COVID cases via UniVariate Time Series Model.

	A	B	C	D	E
1	cough	fever	sore_throat	shortness_of_breath	head_ache
2	2911	1955	546	194	1096
3	1358	1061	447	167	883
4	893	654	302	108	599
5	785	573	282	93	579
6	669	610	273	79	538
7	2511	2014	1023	266	2095
8	2368	1955	973	274	1951
9	2299	1944	981	263	1900
10	3030	2629	1260	344	2696
11	4374	3992	1981	506	4060
12	5380	5033	2377	722	4955
13	4709	4624	2051	675	4451
14	4567	4504	2117	647	4330
15					
16					

Figure 18: 6 month Test Data used to get Forecast via MultiVariate Time Series Model.

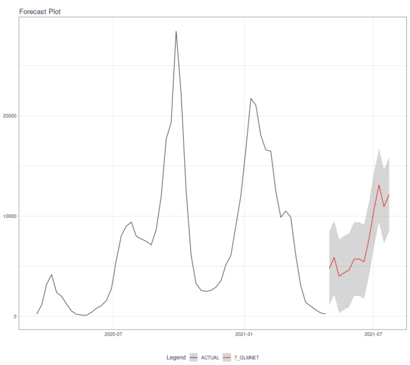
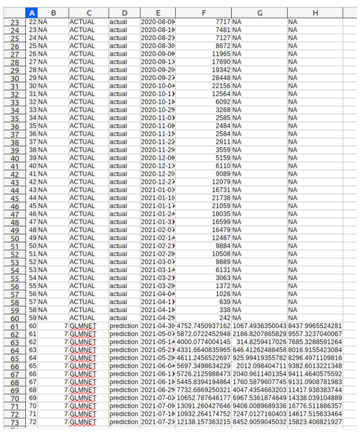


Figure 19: 6 month Forecast of Number Of COVID cases via MultiVariate Time Series Model.



	A	B	C	D	E	F	G	H
23	22	NA	ACTUAL	actual	2020-08-09	7717	NA	NA
24	23	NA	ACTUAL	actual	2020-08-16	7481	NA	NA
25	24	NA	ACTUAL	actual	2020-08-23	7127	NA	NA
26	25	NA	ACTUAL	actual	2020-08-30	8672	NA	NA
27	26	NA	ACTUAL	actual	2020-09-06	11965	NA	NA
28	27	NA	ACTUAL	actual	2020-09-13	17690	NA	NA
29	28	NA	ACTUAL	actual	2020-09-20	19342	NA	NA
30	29	NA	ACTUAL	actual	2020-09-27	28448	NA	NA
31	30	NA	ACTUAL	actual	2020-10-04	22156	NA	NA
32	31	NA	ACTUAL	actual	2020-10-11	12564	NA	NA
33	32	NA	ACTUAL	actual	2020-10-18	6092	NA	NA
34	33	NA	ACTUAL	actual	2020-10-25	3268	NA	NA
35	34	NA	ACTUAL	actual	2020-11-01	2585	NA	NA
36	35	NA	ACTUAL	actual	2020-11-08	2484	NA	NA
37	36	NA	ACTUAL	actual	2020-11-15	2584	NA	NA
38	37	NA	ACTUAL	actual	2020-11-22	2911	NA	NA
39	38	NA	ACTUAL	actual	2020-11-29	3559	NA	NA
40	39	NA	ACTUAL	actual	2020-12-06	5159	NA	NA
41	40	NA	ACTUAL	actual	2020-12-13	6110	NA	NA
42	41	NA	ACTUAL	actual	2020-12-20	9089	NA	NA
43	42	NA	ACTUAL	actual	2020-12-27	12079	NA	NA
44	43	NA	ACTUAL	actual	2021-01-03	16731	NA	NA
45	44	NA	ACTUAL	actual	2021-01-10	21738	NA	NA
46	45	NA	ACTUAL	actual	2021-01-17	21059	NA	NA
47	46	NA	ACTUAL	actual	2021-01-24	18035	NA	NA
48	47	NA	ACTUAL	actual	2021-01-31	16599	NA	NA
49	48	NA	ACTUAL	actual	2021-02-07	16479	NA	NA
50	49	NA	ACTUAL	actual	2021-02-14	12487	NA	NA
51	50	NA	ACTUAL	actual	2021-02-21	9884	NA	NA
52	51	NA	ACTUAL	actual	2021-02-28	10508	NA	NA
53	52	NA	ACTUAL	actual	2021-03-07	9889	NA	NA
54	53	NA	ACTUAL	actual	2021-03-14	6131	NA	NA
55	54	NA	ACTUAL	actual	2021-03-21	3063	NA	NA
56	55	NA	ACTUAL	actual	2021-03-28	1372	NA	NA
57	56	NA	ACTUAL	actual	2021-04-04	1026	NA	NA
58	57	NA	ACTUAL	actual	2021-04-11	639	NA	NA
59	58	NA	ACTUAL	actual	2021-04-18	338	NA	NA
60	59	NA	ACTUAL	actual	2021-04-25	242	NA	NA
61	60	7	GLMNET	prediction	2021-04-30	4752.7450937162	1067.4936350043	8437.9965524281
62	61	7	GLMNET	prediction	2021-05-07	5872.0722452948	2186.8207865829	9557.3237040067
63	62	7	GLMNET	prediction	2021-05-14	4000.0774004145	314.8259417026	7685.3288591264
64	63	7	GLMNET	prediction	2021-05-21	4331.6640835965	646.41262488458	8016.9155423084
65	64	7	GLMNET	prediction	2021-05-28	4611.2456522697	925.99419355782	8296.4971109816
66	65	7	GLMNET	prediction	2021-06-04	5697.3498634229	2012.098404711	9382.6013221348
67	66	7	GLMNET	prediction	2021-06-11	5726.2125988473	2040.9611401354	9411.4640575592
68	67	7	GLMNET	prediction	2021-06-18	5445.8394194864	1760.5879607745	9131.0908781983
69	68	7	GLMNET	prediction	2021-06-25	7732.6869250321	4047.4354663203	11417.938383744
70	69	7	GLMNET	prediction	2021-07-02	10652.787646177	6967.5361874649	14338.039104889
71	70	7	GLMNET	prediction	2021-07-09	13091.260427646	9406.0089689336	16776.511886357
72	71	7	GLMNET	prediction	2021-07-16	10932.264174752	7247.0127160403	14617.515633464
73	72	7	GLMNET	prediction	2021-07-23	12138.157363215	8452.9059045032	15823.408821927

Figure 20: 6 month Forecast of Number Of COVID cases via MultiVariate Time Series Model.