Ensemble Regression Models for Short-term Prediction of Confirmed COVID-19 Cases

Raushan Raj, Anand Seetharam, and Arti Ramesh
Department of Computer Science, SUNY Binghamton
{rrausha1, aseethar, artir}@binghamton.edu

Abstract

Accurately predicting the number of new COVID-19 cases is critical to understanding and controlling the spread of the disease as well as effectively managing scarce resources (e.g., hospital beds, ventilators). To this end, we design a regression based ensemble learning model comprising of Linear regression, Ridge, Lasso, ARIMA, and SVR that takes the previous 14 days' data into account to predict the number of new COVID-19 cases in the short-term. The ensemble model outputs the best performance by taking into account the performance of all the models. We consider data from top 50 countries around the world that have the highest number of confirmed cases between January 21, 2020 and April 30, 2020. Our results in terms of relative percentage error show that the ensemble method provides superior prediction performance for a vast majority of these countries with less than 10% error for 5 countries and less than 40% error for 27 countries.

1 Introduction

COVID-19 is a major global pandemic that has impacted the lives of people around the world. In spite of severe lockdowns in countries around the world to curb its spread, more than 4 million people around the world have tested positive for the virus by May 15, 2020. As the virus spreads unabated, a large of number of individuals continue to get infected globally every day. For example, in USA, starting from a handful of cases in early March, the number of confirmed cases has exceeded 1.4 million by May 15, 2020. Making accurate short-term predictions of the number of COVID-19 cases is critical for upgrading scarce resources such as hospital beds and ventilators as well as procuring vital medicines, particularly in developing countries. Additionally, as countries start to reopen, accurate short-term predictions are necessary to quickly identify new cluster of cases and take appropriate measures.

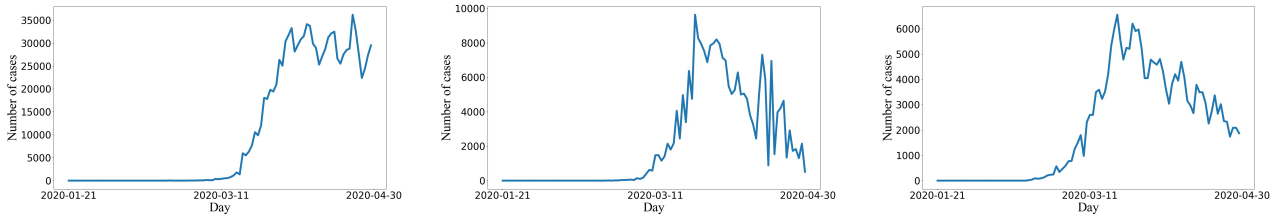
Therefore, in this paper, our goal is to develop a regression-based ensemble model comprising of Linear regression, Ridge, Lasso, ARIMA, and SVR to predict the number of COVID-19 cases in the short-term future using the number of confirmed cases in the past 14 days. The ensemble model selects the best performing model among the above-mentioned ones for the particular dataset in consideration. We consider the data from 50 countries around the world that have the highest number of confirmed cases between January 21, 2020 and April 30, 2020 and execute our ensemble model [jhu, 2020]. We note that our effort builds on and is complementary to existing efforts in this direction [onl, 2020].

We observe from our experiments that a single model (i.e., Linear regression, Ridge, Lasso, ARIMA, and SVR) does not always provide the best performance, with model performance dependent on the country in consideration. This necessitates the use of ensemble learning models that automatically select the best among various models. We evaluate the performance of the model in terms of the relative percentage error (RPE) and investigate the 1-day and the average over 3-day performance. We observe that the ensemble model provides good prediction performance for a vast majority of these countries with less than 10% error for 5 countries and less than 40% error for 27 countries.

Our approach is simple and relies solely on the past data to predict the future. We hypothesize that the difference in prediction performance for the various countries is due to the following reasons—*i)* the difference in testing capacity of the different countries, *ii)* the severity of the lockdown and the social distancing measures, and *iii)* the veracity of the number of cases being reported by some countries. For example, under pressure from the international community, China recently revised its actual number of confirmed cases [chi, 2020].

2 Data & Problem Statement

We use the data collected and distributed by Johns Hopkins COVID-19 Github data repository [jhu, 2020], which provides an overview of COVID-19 cases (confirmed, deaths, and recovered) for countries around the world. The data on the site is updated daily. For the purpose of this study, we select the top 50 countries with the highest number of COVID-19 confirmed cases. Figure 1 shows the daily number of confirmed cases for the top 3 countries. We omit China from the list of countries studied because their numbers were drastically modified recently after pressure from the international community [chi, 2020].



(a) USA (b) Spain (c) Italy

Figure 1: Cases per day for USA, Spain and Italy - the three countries with the maximum number of confirmed cases

The confirmed COVID-19 case prediction problem can be cast as a time-series prediction problem where we consider data for the past n time steps (i.e., x_1, x_2 ..... x_n) and predict k time steps into the future (i.e., $\hat{y}_1$, $\hat{y}_2$, ... $\hat{y}_k$). Note that y_1, y_2, ... y_k are the actual output values. Statistical regression models are ideally suited for such time-series prediction problems and therefore, in this work, we develop a predictive modeling system using an ensemble of statistical regression models to predict future confirmed COVID-19 cases based on past confirmed cases.

3 Ensemble Regression Models

In this section, we provide an overview of our COVID-19 confirmed cases prediction system. It takes data collected and distributed by Johns Hopkins COVID-19 Github data repository [jhu, 2020] as input and outputs the predicted number of confirmed cases in the future. It comprises of two main components: *i)* the data pre-processing component that pre-processes the data, and *ii)* the prediction component consisting of five different models (i.e., Linear regression, Ridge, Lasso, ARIMA, and SVR) that takes the pre-processed data to generate the predictions. The ensemble layer then selects the best model for the particular dataset under consideration. Figure 2 show the architecture of our prediction architecture.

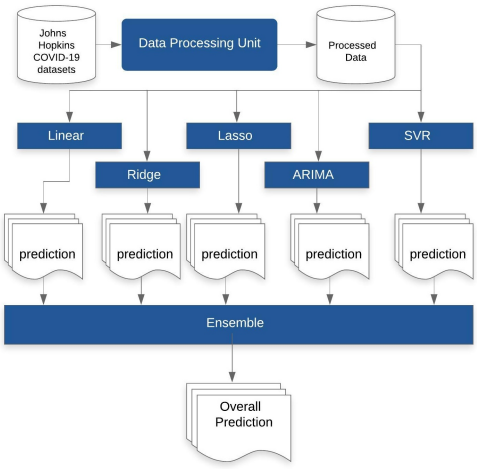


Figure 2: Ensemble model for prediction

We next provide a brief overview of each of the five statistical regression models.

Linear Regression: Linear regression fits a straight line that best models the data.

Ridge: Often when there are fewer observations for training, the underlying cost function can overfit the training observations and Ridge and Lasso regression variants are used to regularize the cost function. In both these regression variants, the procedure involves adding a penalty term to limit the value of the regression coefficients. Ridge regression, also known as L2 regularization, adds an L2 penalty which equals the square of the magnitude of the regression coefficients. This causes all the coefficients to be altered by a factor to still have a non-zero value, but can shrink the coefficients.

Lasso: Lasso regression, also known as L1 regularization, adds an L1 penalty which equals the absolute value of the coefficients. The inclusion of this penalty can possibly induce sparseness in the model driving some regression coefficients to zero.

ARIMA: This statistical model uses a combination of auto-regression, moving average and differencing to model the data.

SVR: Support Vector Regression extends the principles of Support Vector Machines (SVMs) to regression. Similar to SVMs, SVR establishes a decision boundary around the model and maximizes the number of points that falls within this decision boundary.

Our **ENSEMBLE** model then combines the predictions from all these different models to output the best prediction for the dataset under consideration.

To enable the different regression models to make accurate predictions, we generate sequences containing n days as input and k days as output. We adopt a sliding window approach that moves the window one day at a time to cover the entire time series data, generating a sequence of length $n + k$ days each time the window slides. We observe from our experiments that values of n = 14 and k = 3 provide overall good prediction performance.

4 Performance Evaluation

In this section, we first present the metrics used for evaluation and then present the experimental results. Our code for the project is available here [cod, 2020].

For a time series prediction model, the usual metrics used for evaluation are the root mean squared error (RMSE), the mean absolute error (MAE), and the relative percentage er-

ror (RPE). The RMSE and MAE are not appropriate metrics for the COVID-19 case prediction problem because the number of cases have been increasing rapidly in all countries during the time period considered in this study. For example, if the number of cases increases from 100 to 120, the MAE is 20, whereas if it increases from 1000 to 1200 the MAE is 200. The percentage error in prediction in both cases is 20%. Therefore, we consider the relative percentage error (RPE) in our study which is defined as,

$$RPE_{day_j} = \frac{100 * (\sum_{i=1}^{h} \frac{|\hat{y}_{ij} - y_{ij}|}{y_{ij}})}{h} \quad (1)$$

where y_{ij} is the i^{th} test sample for the j^{th} day, $\hat{y}_{ij}$ is the predicted value of y_{ij}, and h is the number of test samples.

We observe from our experiments that for some countries during the initial few days of the spread of the virus, the number of cases is low which throws off the prediction performance of the model. Therefore, we run the models considering 95%, 90%, and 85% of the data for each country and investigate the prediction performance. For example at 95%, if the total number of cases for the entire time period of the study is 100, we consider at least 95 cases or more, leaving out data entirely from days in the beginning. The reason behind this is that for most countries the number of cases during the first few days is small and sometimes even 0. This makes it challenging to predict for those days because if the actual number of cases for a particular day is 0 and our model predicts 1, the percentage error is ∞.

Table 1 shows the average over 3-days RPE for the various countries. The table reports the RPE as well as the percentage of data points used for generating the prediction. We report results for the countries for which the RPE is below 40% in Tabke 1. We observe from the table that 5 countries including USA have less than 10% error in prediction, while 27 countries have less than 40% error in prediction. We also investigate the 1-day prediction performance of the ensemble model and observe that the overall quality of the predictions are better than the average over 3-day predictions. Additionally, we observe that Linear regression, Ridge, and Lasso provide similar prediction performance. This is expected because of the underlying similarity of these models and the relatively simple nature of the data.

4.1 Discussion of Results

We discuss some interesting observations about the evolution of the number of COVID-19 cases in the different countries to give us a better understanding of the prediction performance. We observe that while our model provides good prediction performance for USA and Italy, the countries with the first and third highest infection rates, respectively, it does not provide good performance (around 50%) for Spain, the country with the second highest infection rate. This can be attributed to the high variation in the number of cases from the middle to the end of April (Figure 1b). Similarly, we observe from the data that the number of cases in UAE is continuously increasing. This is the primary reason for the superior performance of our ensemble model. In fact, we observe that if we leave the first 20% data, the RPE for UAE is only 2.7. While the government there is taking preventive action, the lockdowns have not been too severe, resulting in the number of cases increasing steadily. UAE went with night lockdown (i.e.,10 hour lockdown) and then imposed a complete lockdown in early April. However, the restrictions were eased soon and they reverted back to 10 hour lockdowns. India implemented a nationwide lockdown since mid March and parts of the country are still in lockdown. But due to the high population density, it has seen a steady increase in the number of new cases, thus resulting in the model providing good prediction performance. Looking at the absolute numbers, we believe that the number of people infected without the lockdown would have been significantly higher compared to the current number of cases.

We observe from our experiments that a number of countries including Spain, Australia, Norway, Greece, Pakistan, Thailand, and others have prediction performance worse than 40%. A closer look at these countries shows that this set consists of both developed and developing countries. Economically speaking, these countries are also spread across the spectrum. While it is unclear as to why these countries have worse performance in comparison to others, we hypothesize that the difference in prediction performance for the various countries is due to the following reasons—*i)* the difference in testing capability of the different countries, *ii)* the severity of the lockdown and the social distancing measures, and *iii)* the veracity of the number of cases being reported by some countries.

For example, Australia saw its peak in mid March. The strict measures taken by its government to prevent the spread of COVID-19 resulted in a significant decrease in the number of new cases by April and thus our predictive model gave higher RPE for those days. A similar situation is seen when we study the data for Japan, Norway, and Germany. All these countries reached their peak by the end of March or early April and strict lockdown measures resulted in a drastic decline in the number of new cases. Once again, as the rate of decrease was high and could not be accurately captured by the number of cases in the last 14 days', the performance of our predictions for those days was less accurate, leading to overall higher RPE. In comparison to these countries, Sweden asked its people to self isolate, but did not enforce any the nation-wide lockdown. As a consequence, COVID-19 has spread significantly and the number of cases still continue to increase. Due to this combination of self-isolation and open economy, we can see oscillations in the number of cases, making it a difficult prediction task.

Developing countries such as Pakistan and South Africa have all resorted to severe lockdowns, but the number of cases are still increasing on average. One reason could be the inability to maintain social distancing while performing everyday activities in these countries because of their socio-economic situation. Additionally, we see sudden variations in the number of confirmed cases with some days having significantly larger number of confirmed cases than others. This could be the result of skewed testing or reporting, but such drastic variations result in our model providing overall poor prediction performance for these countries.

Country	Total Cases	Linear	Ridge	Lasso	ARIMA	SVR	Ensemble	Percentage of Data Used
Russia	106498	8.72	8.57	8.71	8.7	5.15	**5.15**	90.49
Saudi Arabia	22753	14.21	14.09	14.19	7.57	19.44	**7.57**	90.42
India	34863	8.07	8.03	8.06	9.87	15.54	**8.03**	86.29
United Arab Emirates	12481	9.35	9.35	9.38	18.08	9.57	**9.35**	85.59
US	1069424	15.72	15.67	15.72	9.84	16.26	**9.84**	95.9
Iran	94640	20.96	20.83	20.94	13.51	17.74	**13.51**	90.49
United Kingdom	172481	18.54	18.49	18.54	15.38	18.42	**15.38**	91.45
Canada	54457	21.48	21.43	21.46	16.65	25.64	**16.65**	91.4
Italy	205463	18.6	18.57	18.6	22.47	20.79	**18.57**	95.06
Chile	16023	19.37	19.35	19.35	21.45	28.2	**19.35**	95.34
Poland	12877	24.2	24.13	24.11	19.78	35.48	**19.78**	95.08
Colombia	6507	21.2	21.17	21.09	19.82	27.73	**19.82**	86.08
Mexico	19224	20.14	20.1	20.13	24.76	29.96	**20.1**	95.59
Brazil	87187	20.62	20.68	20.63	34.77	20.46	**20.46**	86.05
Argentina	4428	32.34	32.28	32.06	22.94	46.32	**22.94**	86.7
Romania	12240	25.69	25.65	25.63	23.03	33.74	**23.03**	95.29
Turkey	120204	23.96	23.9	23.96	24.74	23.15	**23.15**	95.26
Serbia	9009	48.18	48.01	47.98	24.13	55.0	**24.13**	87.0
Indonesia	10118	25.2	25.17	25.17	25.27	34.85	**25.17**	95.55
Philippines	8488	36.2	36.13	36.09	25.21	49.63	**25.21**	90.54
Singapore	16169	76.16	75.96	76.09	34.06	25.44	**25.44**	85.78
Sweden	21092	32.8	32.74	32.76	26.42	42.55	**26.42**	95.15
Denmark	9356	32.31	32.22	32.22	27.99	34.97	**27.99**	95.25
Finland	4995	46.81	46.7	46.49	32.25	68.22	**32.25**	85.99
Israel	15946	32.99	33.06	33.1	70.63	76.68	**32.99**	85.14
Netherlands	39512	33.49	33.49	33.49	39.53	36.85	**33.49**	95.67
Belgium	48519	34.94	34.93	34.94	38.78	38.0	**34.93**	95.35
Malaysia	6002	35.18	35.08	35.03	50.29	64.3	**35.03**	96.03
Peru	36976	35.16	35.18	35.16	47.74	43.22	**35.16**	95.28
Dominican Republic	6972	43.53	43.47	43.45	38.5	60.58	**38.5**	95.52
Panama	6532	46.32	46.24	46.17	38.7	53.99	**38.7**	95.21
Germany	163009	38.98	38.95	38.99	59.62	46.69	**38.95**	95.54

Table 1: Three day ahead average RPE comparison across various countries

5 Related Work

We note that our effort builds on and is complementary to existing efforts in this direction [onl, 2020]. Due to the recent nature of the problem, there is limited peer-reviewed research in this space and most work is in the form of interactive websites. For example, the model in [cov, 2020c] estimates the social distancing using geolocation data from mobile phones and uses it for forecasting. Similarly, the model in [cov, 2020a] takes individual state-by-state re-openings in the United States into account and investigates their impact on the number of infections. The COVID-19 Simulator assumes that contact rates will increase by 20% after stay-at-home orders are lifted and uses that for making projections [cov, 2020b]. In comparison to existing work, we perform short-term prediction of the number of confirmed COVID-19 cases in countries around to world and not just in the United States.

6 Conclusion and Future Work

In this paper, we designed an ensemble regression model comprising of Linear, Ridge, Lasso, ARIMA, and SVR for the predicting the number of COVID-19 cases, and demonstrated that it provides good prediction performance for a number of different countries. Our designed model is computationally efficient, requires minimal supervision to execute, and can be readily adopted by countries that lack the computational infrastructure and/or expertise. In future, we plan to design an interactive website where people can easily view the predictions for the COVID-19 confirmed cases. Additionally, we also plan to extend our models and make them work at the regional/county levels to assist with local governance. As more data becomes available, we also plan to investigate the applicability of models such as Random Forest Regressor and Gaussian Conditional Random Fields.

References

[chi, 2020] China revises death toll by 50%. https://www.bbc.com/news/world-asia-china-52321529, 2020.

[cod, 2020] Code for project. https://bitbucket.org/rrausha1/covid-19-case-prediction/src/master/, 2020.

[cov, 2020a] Covid-19 projections using seir. https://covid19-projections.com/about/, 2020.

[cov, 2020b] Covid-19 simulator. https://www.covid19sim.org/team, 2020.

[cov, 2020c] The university of texas covid-19 modeling consortium. https://covid-19.tacc.utexas.edu/projections/, 2020.

[jhu, 2020] Johns hopkins COVID-19 github data repository. https://github.com/CSSEGISandData/COVID-19, 2020.

[onl, 2020] Center for disease control and prevention (cdc) covid-19 forecasts, 2020. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html.