US Covid-19 Prediction with L1 Regularized Linear Regression

Yue Wang

Department of Data Science, Brown University

DATA 1030: Hands-on Data Science

Professor Andras Zsom

Dec 07, 2021

Introduction

Covid-19 was first diagnosed and reported in December 2019, and surprisingly, continued to affect the world for almost two years with no signs of extinction so far. The prediction of Covid-19 cases seems to be an important way to measure the severity of it and to determine which protocols should be practiced and to what extent to contain it worldwide.

This project aims to predict the number of future daily Covid-19 cases by the previous daily counts of cases in the US with regression models. Data were maintained and renewed by Tableau, with two data sources. The US data was retrieved from New York Times along with other 1,929 countries, and the rest of the world's data was from JHU CSSE Global Timeseries. The dataset has 13 columns which represent the cumulative number of people with positive Covid cases, county name, province/ state name, report date, continent name, data source name, the number of daily new deaths, county FIPs number, country alpha-3 code, country short name, country alpha-2 code, daily new positive cases, and cumulative deaths, ordered as the dataset. Each row provides the above info per day per county. There are 2,435,986 entries from 219 countries in total and this project mainly concentrates on the 2,186,968 rows representing the US only, which is still a large proportion of the data due to the lack of information on county or province in most of the other countries.

Data, code, and figures can be found at the GitHub Repository: https://github.com/YueWangpl/covid_pred.

Exploratory Data Analysis

Figure 1 shows the overall trend of the daily cases varying over time. It appears that there could be a seasonal correlation where cases increase slowly during winter and more rapidly during summer. The histogram of case count in Fig. 2 also confirms such assumption with a sign of multi-modal distribution, at least bi-modal. Whereas more data over at least 3 years would be needed to draw the conclusion. The first and only negative count appeared on June 4th, 2021. According to the data source, the New York Times explained the negative values stating that health officials frequently remove or reassign cases and deaths after receiving new information, resulting in small decreases in total state and county tallies. Common reasons include removing duplications or cases and deaths that turn out to involve people who live in other jurisdictions. Occasionally, jurisdictions report larger decreases after changing case or death definitions.

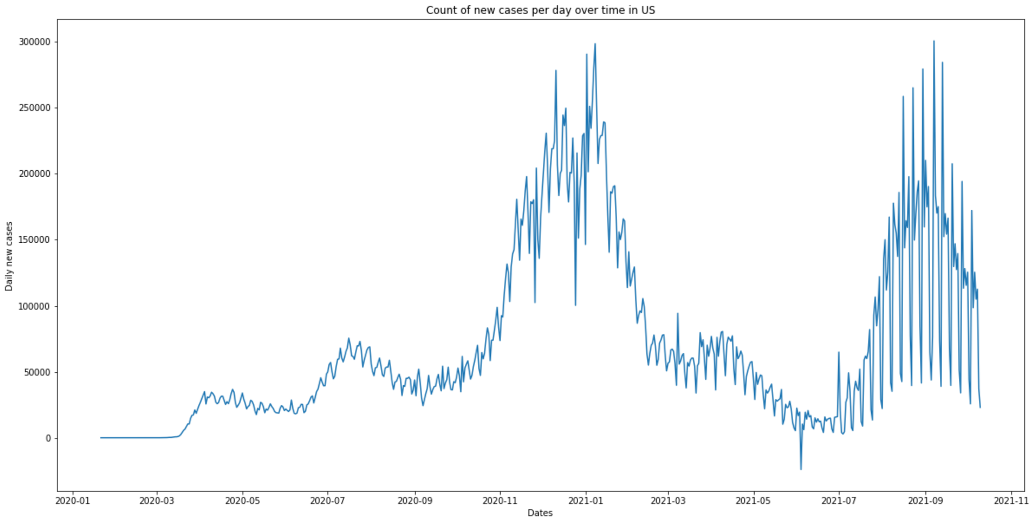


Fig. 1 US Daily Count of Positive Cases vs Time

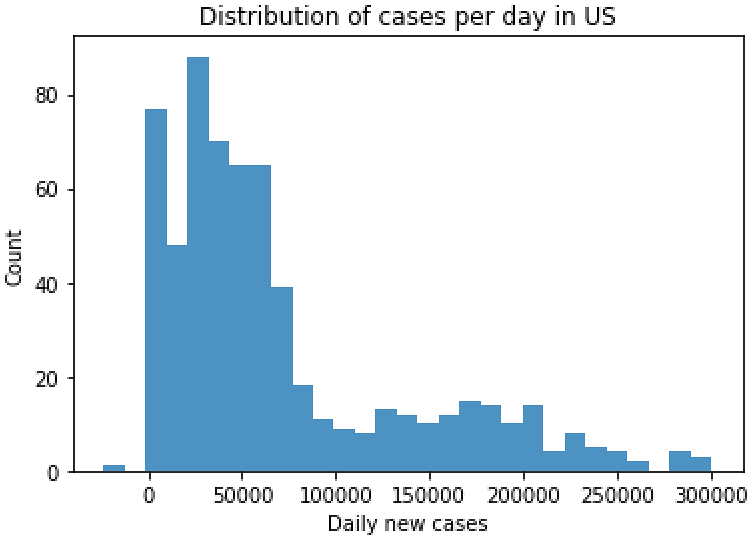


Fig. 2 US Daily Count of Positive Cases Histogram

Figure 3 displays the average daily reports among different states in the US. California with the most population also reports the most daily cases on average. Whereas, Arizona falls behind it while being the 14th most populated state and the 33rd state ranking by the density of population (both from Census 2020). Fig. 4 presents the violin plot of daily cases grouped by different states. Symmetrical log scale was applied on both positive and negative sections of the y-axis to improve the readability of the graph. Higher variances are commonly exhibited in states with the most daily reported cases, alongside some outliers, such as New York.

Figure 5 exhibits a positive correlation between the number of daily deaths and the number of daily cases via a scatter plot.

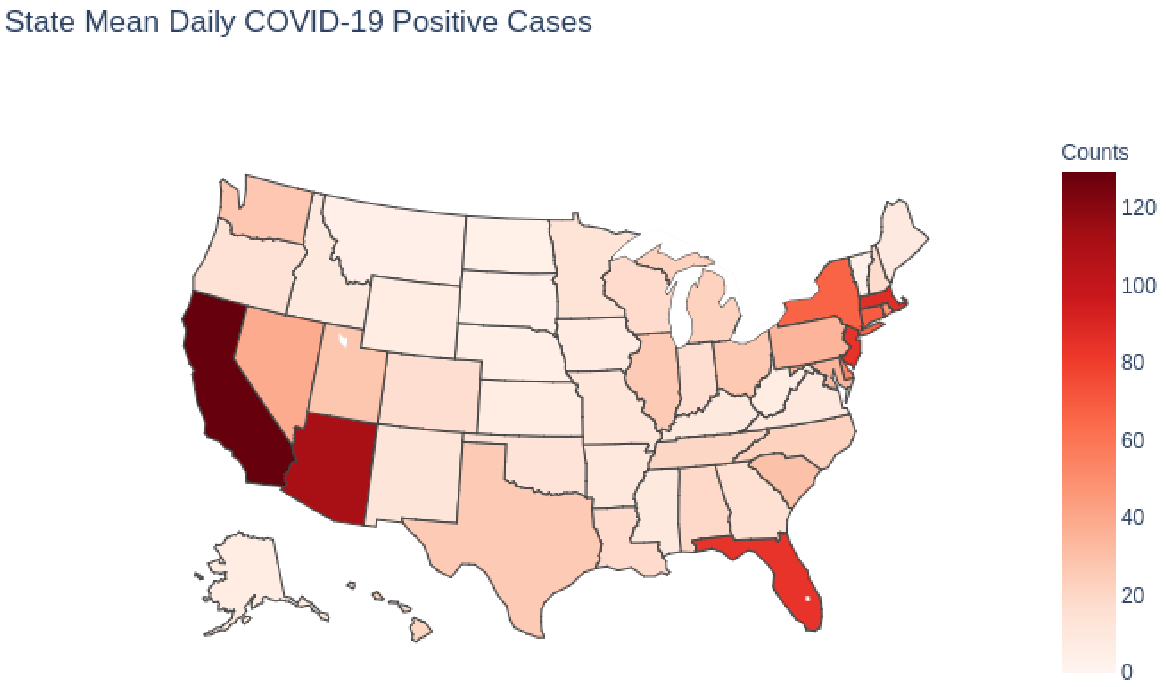


Fig. 3 US Daily Count of Positive Cases by State

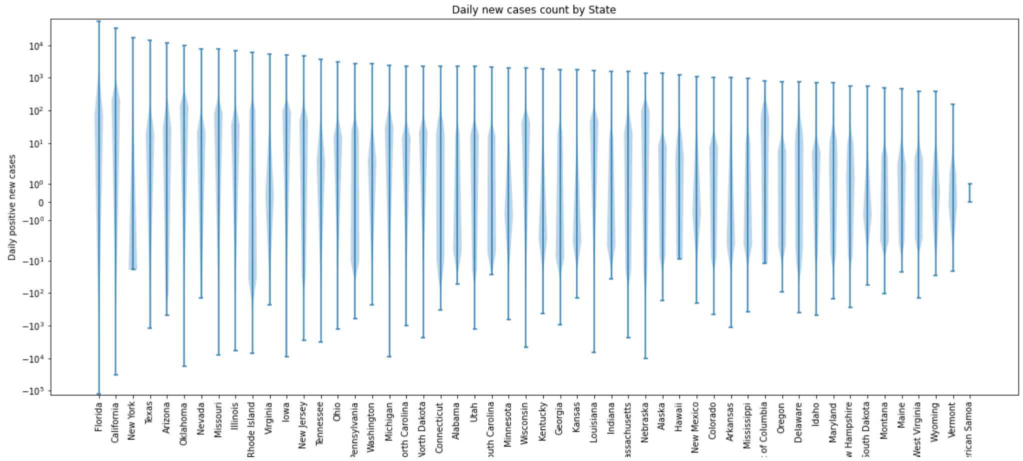


Fig. 4 US Daily Count of Positive Cases by State Violin Plot

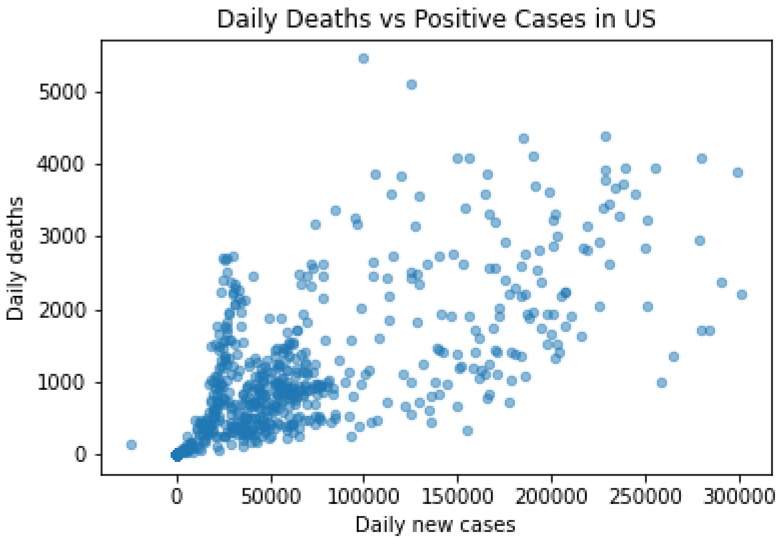


Fig. 5 US Daily Deaths vs Daily Count of Positive Cases

Methods

Data Preprocessing

The data being used has a time-series structure, so it is not i.i.d. However, 35 days were designed to be the look-back period, so 35 columns were created with the recorded number of new positive cases on the previous i days, i∈[1, 35] to be used as features in order to predict the target variable. The first 35 rows were dropped as a consequence of the inability to retrieve the Covid counts of all 35 days earlier than them. After feature engineering, this new model-ready data frame, with 35 feature variables and 1 target variable, should now be i.i.d. since each row does not rely on other values besides its own fields. Additionally, the target variable is the count of daily cases of the US instead of states or counties, so this new data frame has no group structure either. Therefore, it's safe to perform a train-test-split. However, after comparing the R2 scores of simple Train Test Split and Time Series Split, the R2 score with Time Series Split appears to be 0.72 which is lower than two standard deviations from the mean, 0.89, of Train Test Split, because the daily case count was increased dramatically after November 2020 according to Fig. 1, and predicting that partition of data has no practical meaning for our task. Therefore, Time Series Split was determined to be the splitting method. 60% data was used as the training set, 20% as the validation set, and the last 20% was designed to be the testing set.

All columns of the model-ready data frame are continuous. As the last step of preprocessing, a standard scaler was fit on the features of the training set and transformed on the feature variables of all three sets.

Machine Learning Pipeline

Four ML algorithms were tested including linear regression, linear regression with L1 and/ or L2 regularization, random forest, XGBoost. Validation scores are calculated by 3 time-series split. As shown in Fig. 6, the first three splits represent the cross-validation split, and the last split represents the split of the test set.

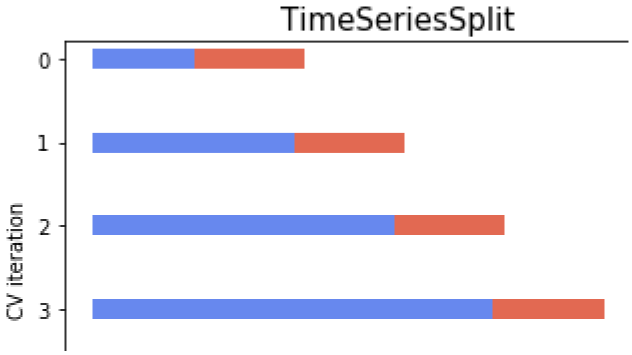


Fig. 6 Time Series Split

MAE and MSE were considered but ultimately, R2 score was selected to be the metric due to the large scale of y values. Tuned parameters are displayed in Table. 1.

Table. 1. ML algorithms and Parameters

Model	Parameter Grid	Best Parameters	Validation Score	Test Score
Baseline	N/A	N/A	0	-0.0008
LinReg	N/A	N/A	0.81	0.56
Lasso	**alpha ∈ [1e-6, 1e8]**	**alpha': 112.20184543019607**	**0.83**	**0.58**
Ridge	alpha ∈ [1e-6, 1e12]	alpha': 5.623413251903491	0.83	0.57
Elastic Net	alpha ∈ [1e-6, 1e8] l1_ratio ∈ [0, 1]	alpha': 0.03548133892335753, 'l1_ratio': 0.45	0.84	0.56

Random Forest	n_estimators ∈ {3, 10, 30, 70, 80, 90, 100, 120, 150, 300, 500}
	max_depth ∈ {1, 3, 5, 7, 8, 9, 10, 30}
		max_depth': 9,
		'n_estimators': 120
			0.39	0.07

XGBoost	learning_rate ∈ [1e-6, 1] n_estimators ∈ [1, 500] max_depth ∈ [0, 100]	learning_rate': 0.03162277660168379, 'max_depth': 25, 'n_estimators': 500	0.45	0.19

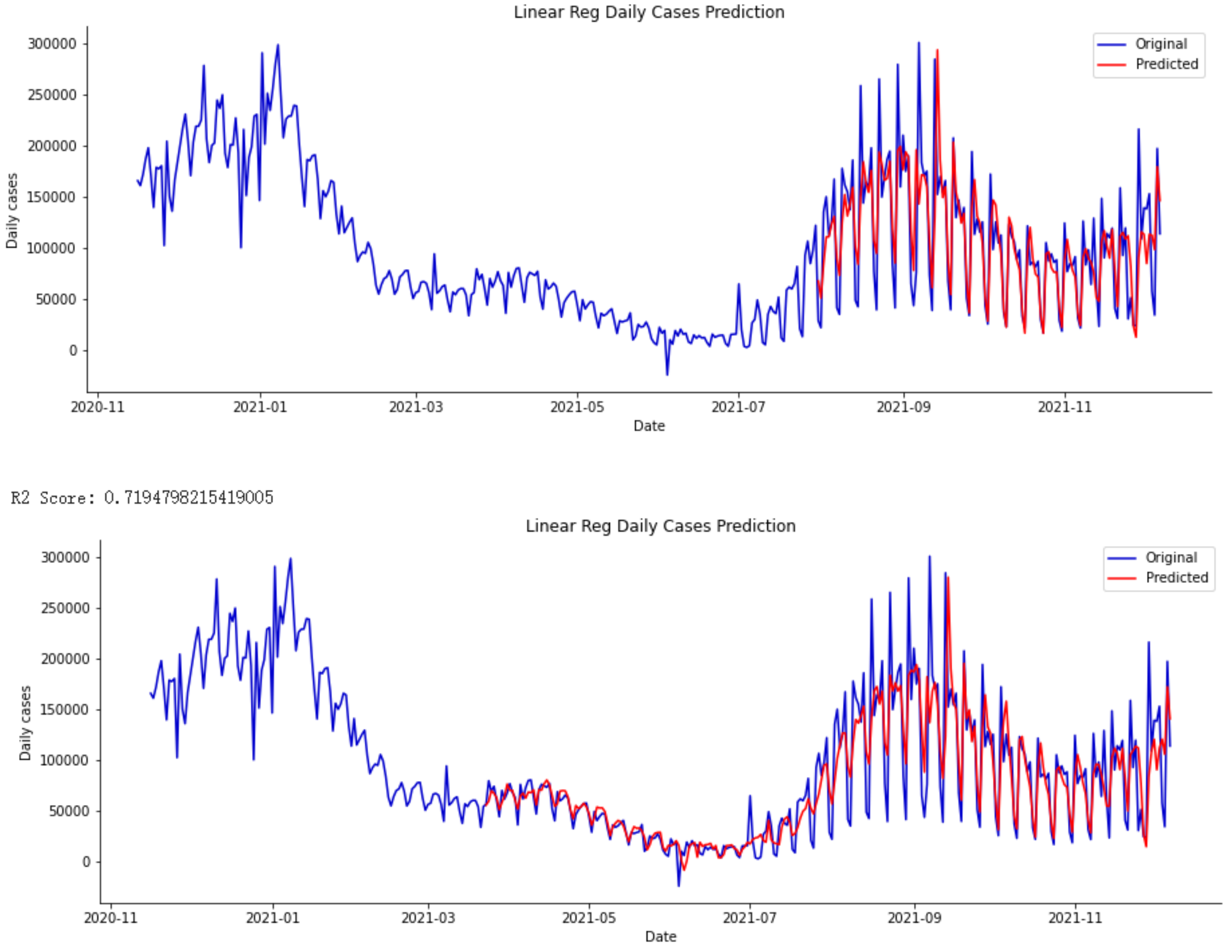
Therefore, a linear regression model with L1 regularization with alpha = 112.20 was decided to be the final model.

Results

The baseline model of a regression task would be to predict future values with the average number of the current dataset, and by definition, the R2 score of such calculation is 0. And by experiment, the R2 score of the baseline model is -0.0008. Since the dataset was split by time series split, the standard deviation of the R2 score was not able to be calculated by switching the random state of the split.

As mentioned above, the ML algorithm was determined to be the Lasso regression model with alpha = 112.20, which produced the validation score to be 0.83 with a standard deviation of 0.039, and the best test score to be 0.58 among other regression models.

If the test size is increased and retrain the model with the best set of parameters, the R2 score can increase to 0.72 because of the more spiky behavior in the recent data compared to spring this year.



```
R2 Score: 0.7194798215419005
```

Fig. 7 Model Prediction *Top*: 20% data as the test set. *Bottom*: retrained with test size = 40%.

Feature importances are calculated by coefficients, feature permutation score, and by SHAP values. We can observe a periodic pattern that the feature importances do not decrease monotonically, but they decrease with oscillation.

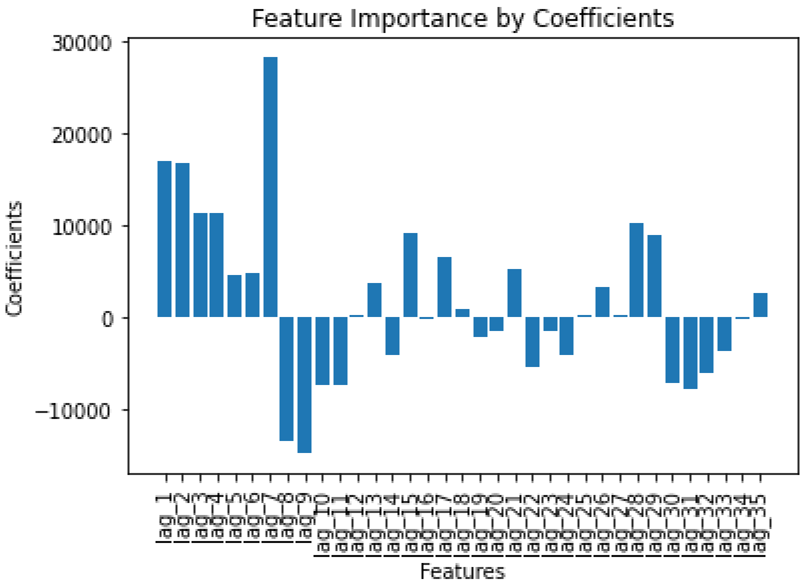


Fig. 8 Feature importances by Coefficients

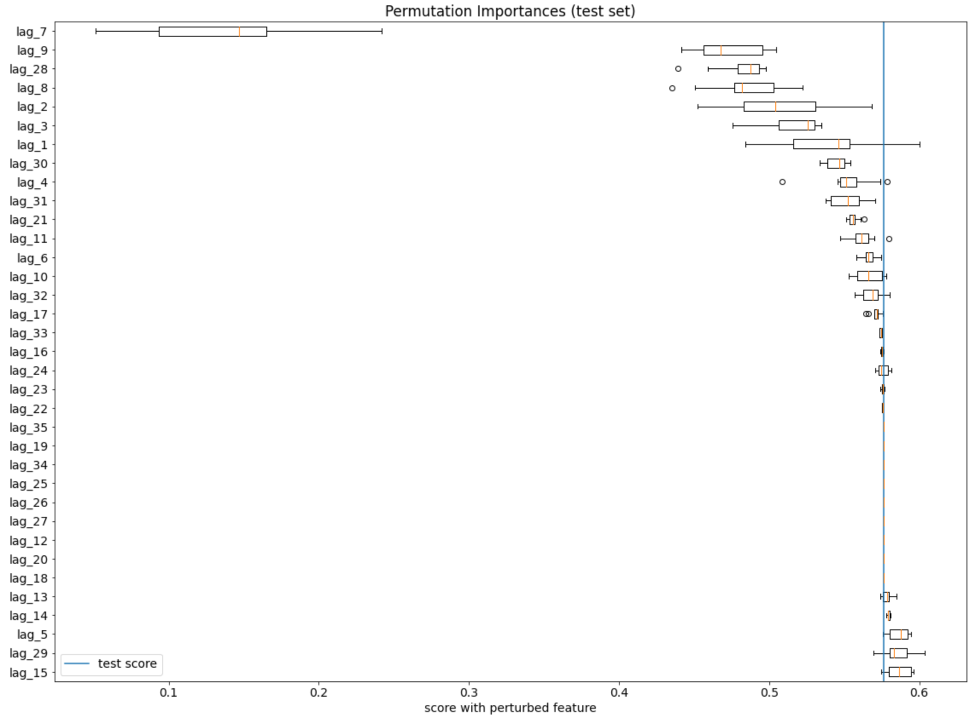


Fig. 9 Feature importances by Permutation

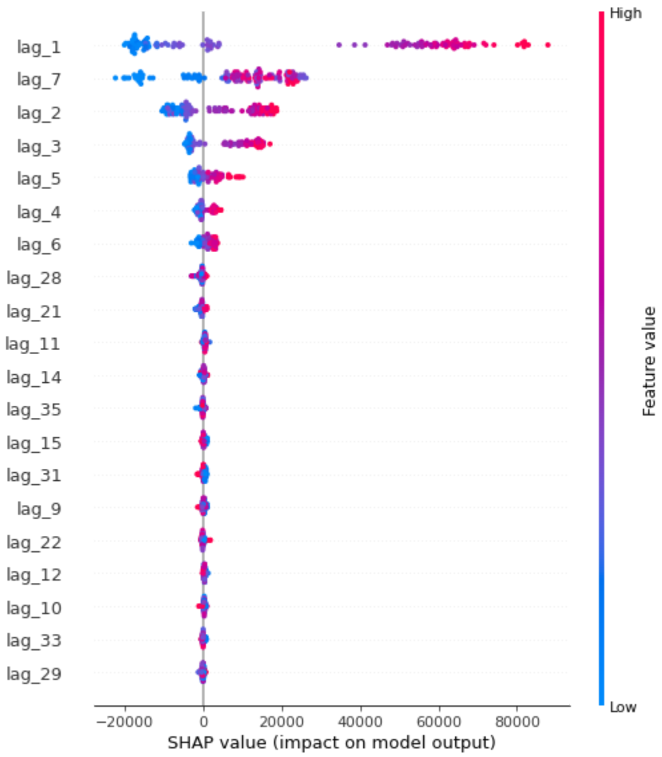


Fig. 10 Feature importances by SHAP

Local feature importances are also measured by SHAP as shown in Fig. 11. A periodic pattern can also be observed.

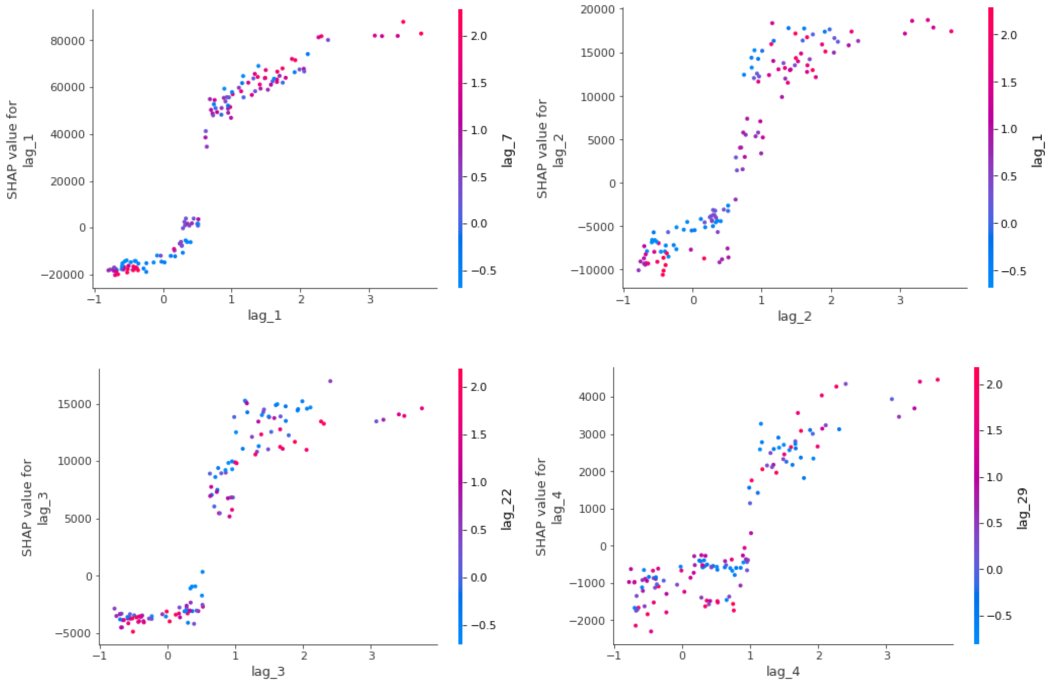


Fig. 11 Local feature importances by SHAP

It’s surprising that the feature of lag_7, representing the case count 7 days before the prediction is the most important feature. Considering that the recorded data might not be on time by the health officials, it is smart for the New York Times to report the weekly average rather than daily counts as described on their website. However, it also turns out that the least important feature is lag_15 which is the count of 15 days earlier than the report date.

Outlook

Only domestic data were used by my method, and data from other countries were completely ignored. A better model could be built with additional features such as the daily count of other countries that are geographically close to the US or countries without a travel ban by the US. In terms of algorithms, deep learning models with LSTM layers may also be used to predict future points.

References

Tableau. (2020, October 12). *Covid-19 (coronavirus) Data Resource Hub.* Tableau. https://www.tableau.com/covid-19-coronavirus-data-resources

Liebeskind, M. (2020, April 22). *5 Machine Learning Techniques for Sales Forecasting*. Towards Data Science. https://towardsdatascience.com/5-machine-learning-techniques-for-sales-forecasting-598e4984b109

The New York Times. (2021, December 07). *Track Corona Virus Cases in Places Important to You*. The New York Times. https://www.nytimes.com/interactive/2021/us/covid-cases-deaths-tracker.html