Cheng Zhu 3034947319 cheng_zhu@berkeley.edu

Machine Learning Models for the Prediction of US County's Monthly Covid Death Numbers

Author: Cheng Zhu

1. Abstract

This project aims to apply machine learning models to predict the monthly death number in US counties. Nearly one hundred features on the county level from three datasets are considered for the accurate prediction of the death number in every county in March 2022. After thoughtful feature engineering, regularization, and cross-validation, an optimal linear regression model is proposed for the task. Root mean squared error (RMSE) and R-squared value are used to evaluate the models. The prediction can help local officials to be prepared for the severity of the upcoming covid trend and help them make the best opening decisions for their county.

2. Introduction

Since the COVID-19 pandemic started in early 2020, intensive data has been collected by various research and government agencies to assist policymakers in making the right decisions for the public. One of the critical decisions is maintaining the balance between economic opening and the restrictions on lockdown. The first element of this tradeoff is highly related to every citizen's daily life. People simply cannot handle the severe consequences of shutting down economic activities forever and pray that the virus magically disappears.[1] On the other hand, enforcing some travel and opening restrictions is crucial in slowing down the spread and saving lives. In the US, the local officials of each county have great autonomy in making their policies. In order to make the optimal decision, people need to be able to predict the severity of the pandemic in the future so that the correct policies can be made. In this project, I would like to apply machine learning models to predict each county's monthly death numbers. Specifically, I focused on building a model using previous monthly death and confirmed case numbers, and county's demographic and economic data to predict each county's COVID-19 death count of March 2022 accurately.

Current literature has been majorly focused on using numerical models to predict growth in COVID-19 death cases at a country or state level.[2,3] Although they can offer some insight into the general trend of the pandemic, county officials need more local data to make the right decisions. In fact, the pandemic hit different parts of the US at different time frames. For example, New York County was hit head-on, and the death cases were extremely high during the early half of 2020, while more inland counties might experience a gentler impact overall. As a result, people need a better understanding of each county's future COVID-19 development using their own unique data.

3. Description of Data

In this project, the data are collected by the Data 100 staff and the author of this project. Two datasets collected by the staff are "time_series_covid19_deaths_US.csv" and "time_series_covid19_confirmed_US", which have daily COVID-19 death and confirmed case numbers on the county level, respectively. In addition, the dataset I collected is the "county_complete.csv",[4] which includes a wide range of demographic and economic data on the county level. For instance, the age and race distribution, the GDP per capita, the poverty rate, and

people's educational background of the county. There are nearly a hundred features in the dataset I collected, but I only hand-picked about 10 features that I am most interested in and believe may impact the death amount of March 2022.

4. Description of Methods

4.1. Train-Test Splitting

After cleaning and merging the datasets, a working dataframe is achieved. It contains monthly confirmed and death numbers on the county level from Feb/2020 to Mar/2022. Also, it has demographic and economic data for each county. A detailed description of the dataframe can be found in the analysis notebook. The shape of the dataframe is 2386 rows and 68 columns. Firstly, a train-test split was done to randomly select 2000 rows as the training set and the rest 386 rows as the test set (holdout set). The trend of COVID-19 deaths in the United States is shown in Figure 1. There are multiple waves in the death number as illustrated in the figure. Also, to look at the trend on the county level, Baldwin, Alabama is randomly selected (Figure 2). It is shown that Baldwin was hit more severely in the fourth wave (delta variant surge), while the country, in general, was hit more severely in the third wave. This again highlights the necessity to analyze and predict the death numbers on a county level.

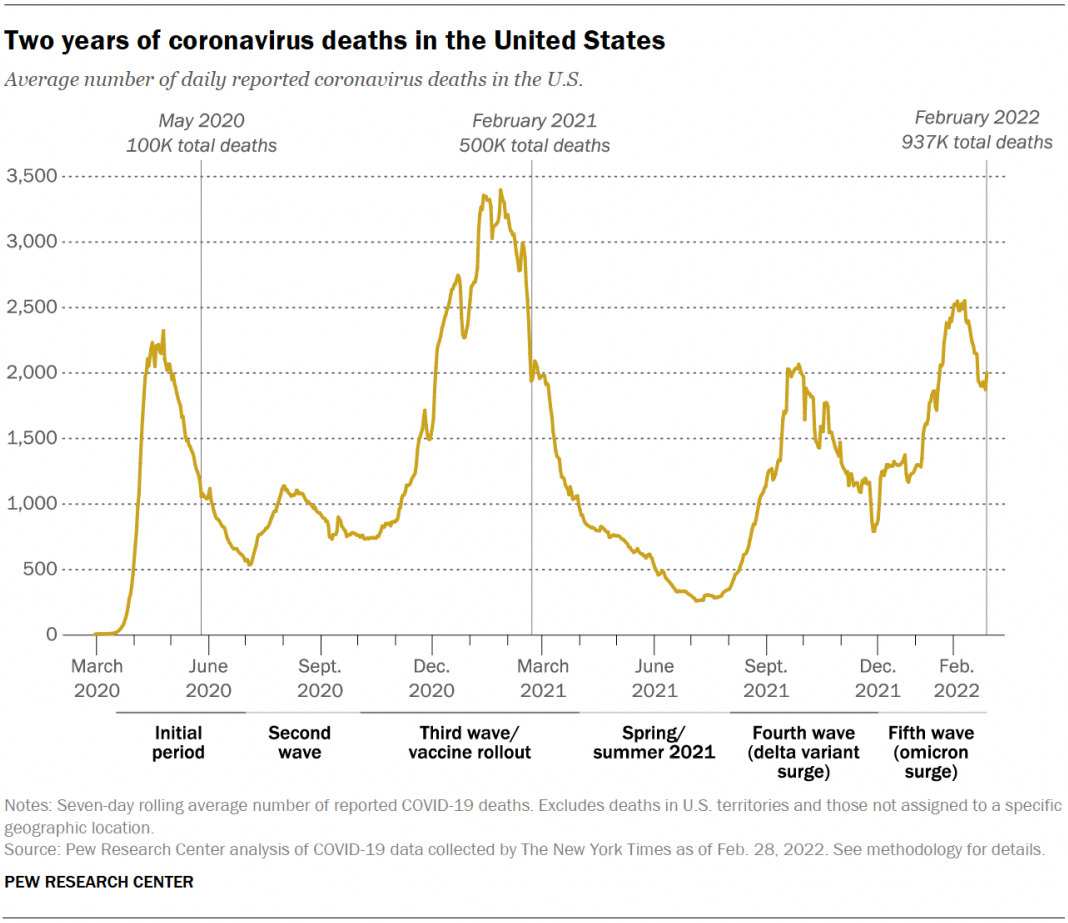

Figure 1. Two years of coronavirus deaths in the United States.[5]

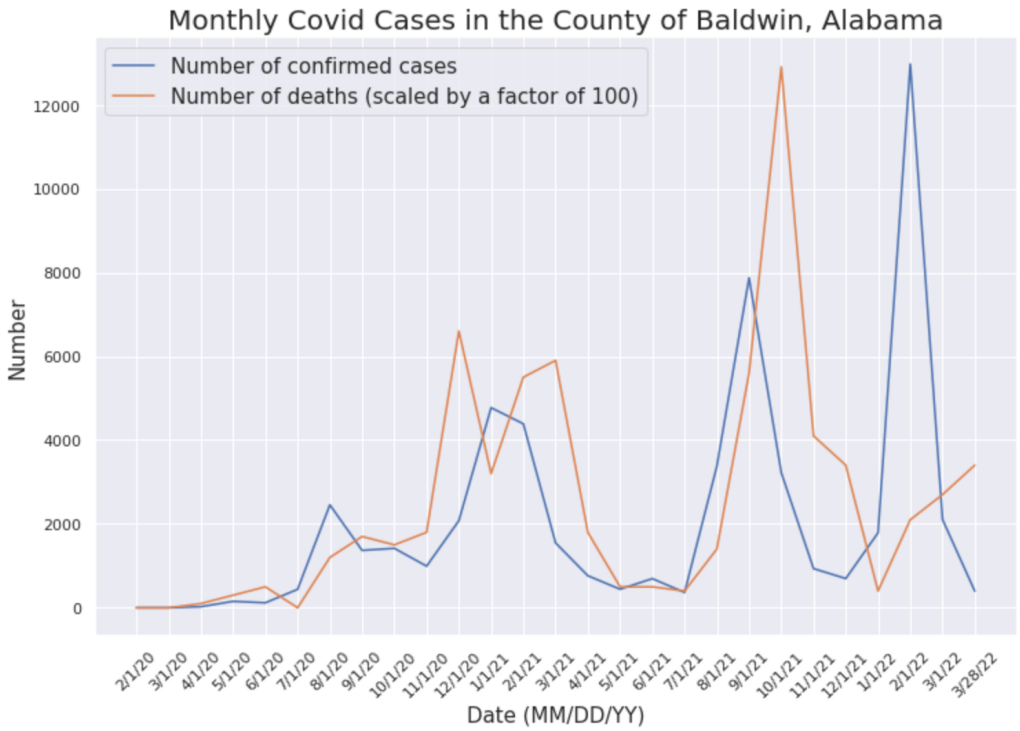

Figure 2. Two years of coronavirus confirmed cases and deaths in Baldwin, Alabama.

4.2.Exploratory Data Analysis

Understanding the correlation among the features and the correlation between each feature and the target column (March 2022 death number) is very important for the later modeling process, so I used Seaborn heatmap and matplotlib line plots to visualize these correlations of our training set. Firstly, the correlation between the monthly death numbers is shown in Figure 3. It is clearly shown that the first few months of the pandemic, namely Feb/2020 to June/2020, are very poorly correlated to the numbers of later months, while nearly all the later months are correlated with each other. As a preliminary feature screening, all the death numbers of later months can be selected, but the overfitting issue will likely be inevitable, which will be resolved in later sections. Similarly, the correlation between monthly confirmed cases and March/2022 death number (Figure 4.) also shows that the death number is highly correlated to months after June/2020, with a few exceptions such as May, June, and December/2021. These three months are the "valley" period, where fewer deaths occurred. In the current period (March/2022 in this study's case), people are facing another surge in the death numbers caused by the omicron variant, so the data from the "valley" period may not be very meaningful for our prediction.

Having a sense of the relationship between the March/2022 death number and the demographic and economic data is also essential for utilizing the most helpful features. As shown in Figure 5, only the feature of "pop_2019", which represents a county's population in 2019, is highly correlated to the death number of March/2022. Other factors, such as age and race distribution, income per capita, household people density, people's educational background, etc., are not highly correlated to the death number. This finding is very interesting and is the opposite of my expectations. According to a previous study in July 2021 regarding the correlation of COVID-19 death rate and demographic factors,[6] the death rate in African American and Hispanic

communities are much higher than in other communities. One possible explanation for my finding is that the coronavirus has changed a lot since the early stage of the pandemic. It has become much more contagious and less lethal so that, at least in the US, human factors do not matter that much in determining the death number. However, this finding should be investigated more carefully and rigorously, but it is beyond the scope and objective of the current project.

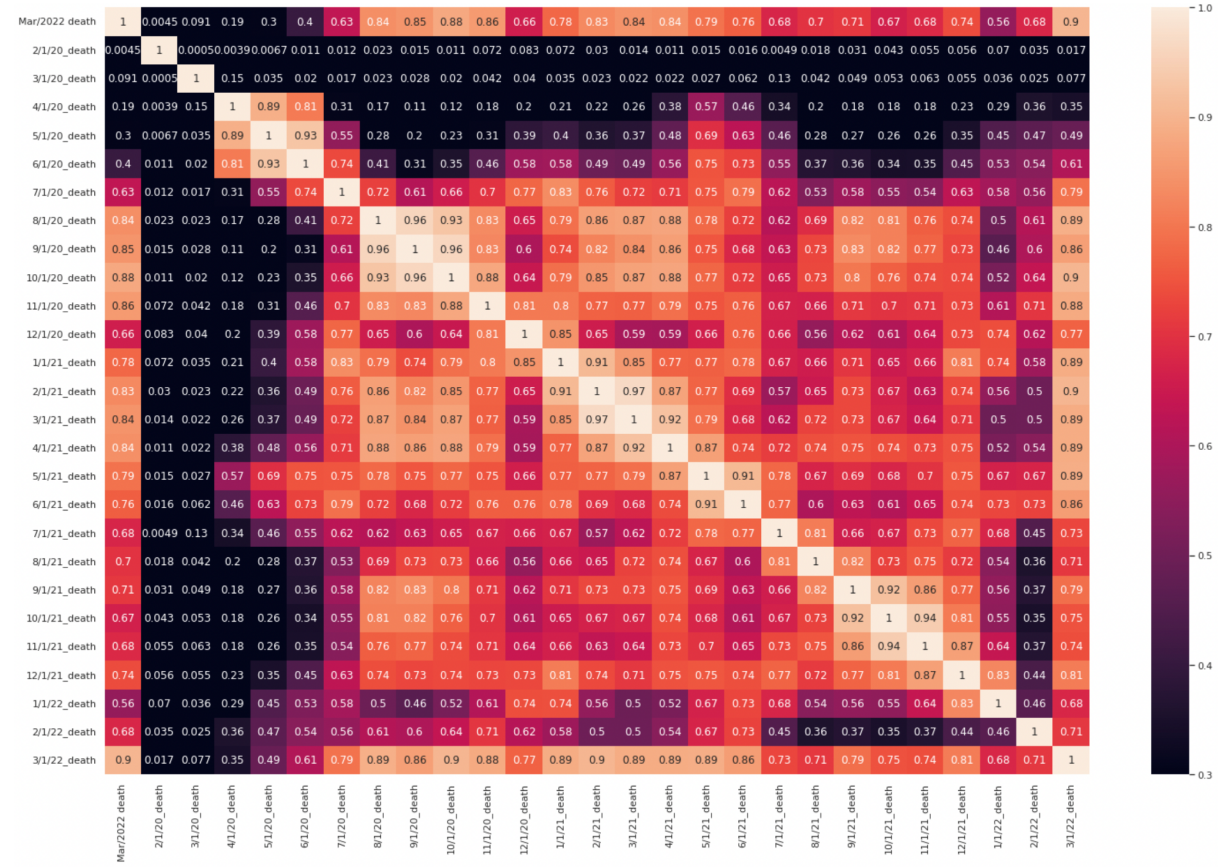

Figure 3. The correlation among monthly death numbers.

Cheng Zhu 3034947319 cheng_zhu@berkeley.edu

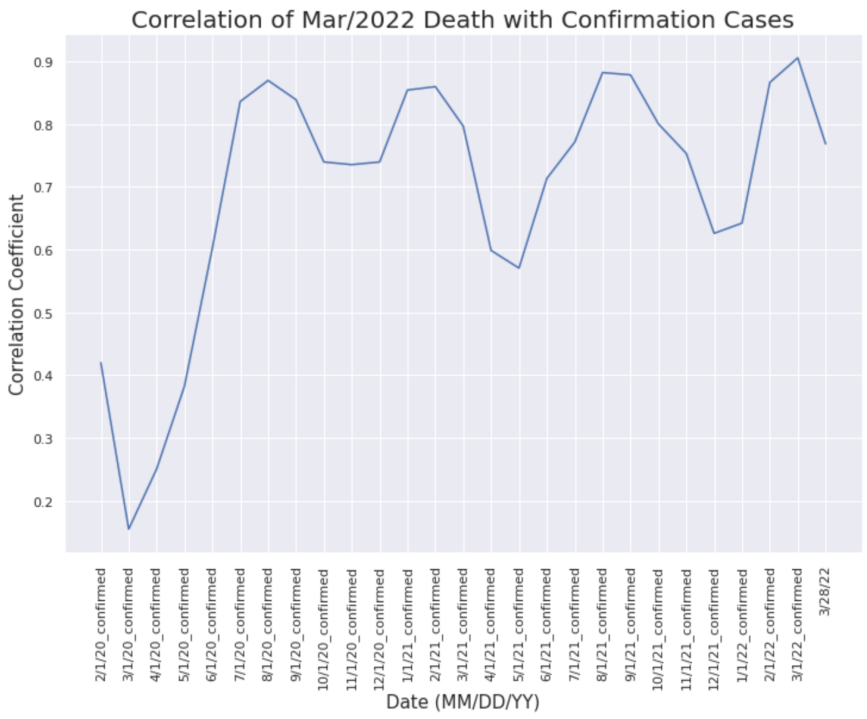

Figure 4. The correlation between monthly confirmed cases and March/2022 death number.

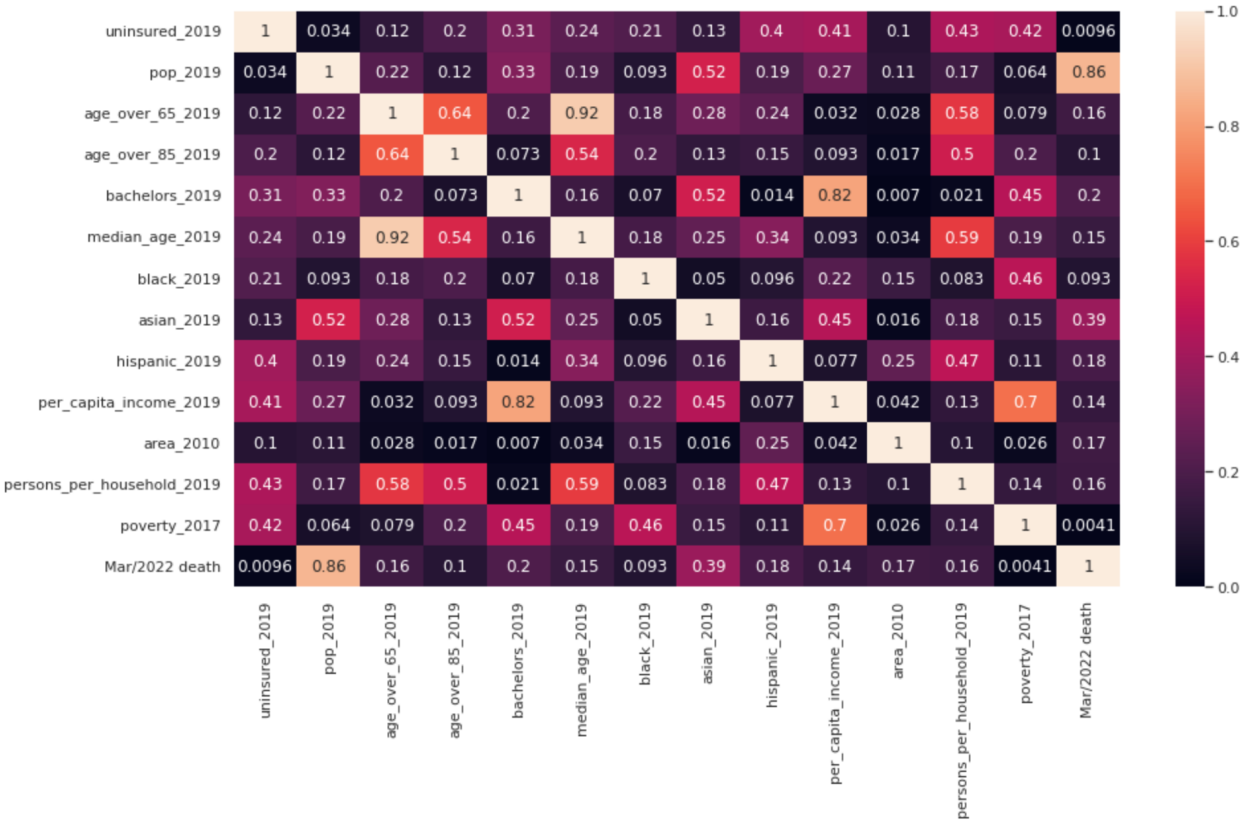

Figure 5. Correlation of the demographic and economic data with the March/2022 death number.

Cheng Zhu 3034947319 cheng_zhu@berkeley.edu

4.3. Data Modeling and Inferences

Linear regression models were used to predict the death number of each county in March 2022 because, according to previous analysis, that number is highly correlated to many features, such as the death/confirmed case numbers of various months and the population. First, all the selected features are applied to fit the training set, and the test set results are shown in Figure 6. It seems that most numbers are below 10. Deleting some outliers may help with the modeling. The root mean squared error (RMSE) is used to evaluate the model. The testing set RMSE is 16.04. The following efforts will be focused on how to reduce this number.

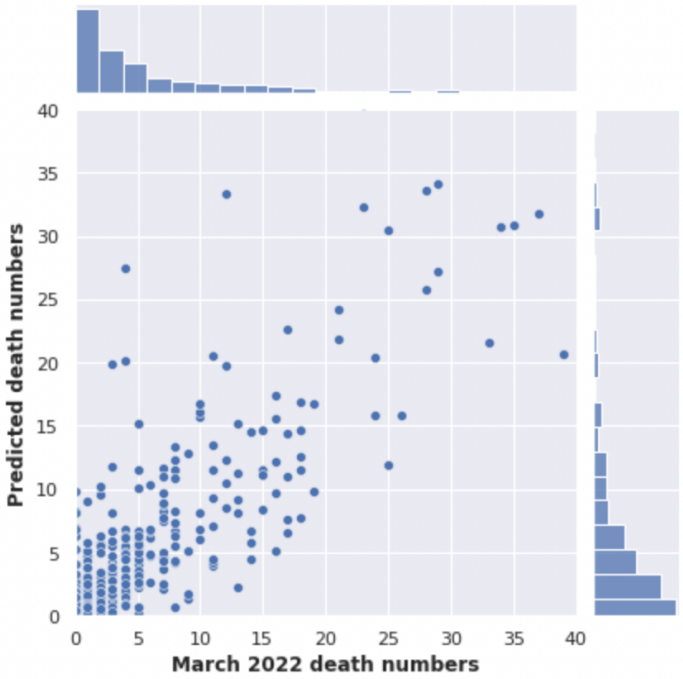

Figure 6. The predicted and real March 2022 death numbers of the test (holdout) set using all the features.

One improvement is only to select some features to train the model. The below section focuses on showing the RMSE of the model with different numbers of features. Figure 7 demonstrates that selecting more features will always decrease the training error, but the holdout error may increase if we include more features. Also, from this figure, it is noticeable that when some features are added, the holdout error will increase. I consider those features not useful, so I only selected the features that can help decrease the holdout error. Now the holdout RMSE is decreased to 11.46. The features I selected are the death numbers from May/2021 to Feb/2022, and the confirmed case numbers of Aug/2020, Dec/2020, and from June/2021 to Mar/2022. In my opinion, the major rationale behind this selection is that the coronavirus has been constantly evolving, and it has become more contagious and less lethal. Hence, the data from the period before the delta-variant wave is not that useful in predicting the current death numbers.

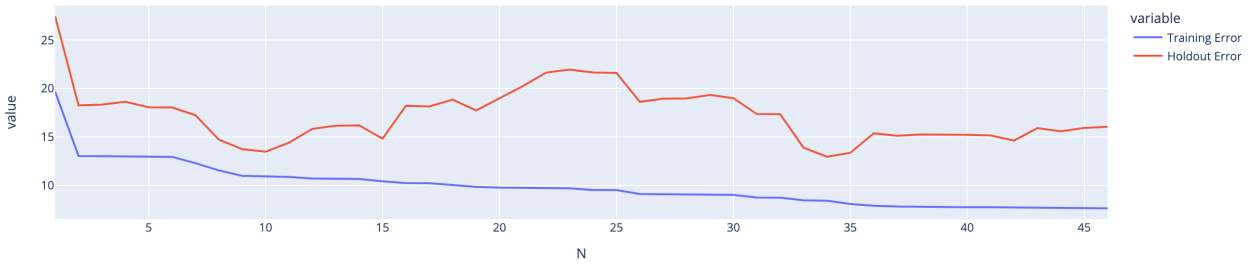

Figure 7. The training and holdout error for different numbers of features.

Although some more useful features are selected, it has been shown in section 4.2. that the features are highly correlated. As a result, applying ridge regression will be helpful to further decrease the testing error since it is a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear independent variables.

In order to properly regularize a model, the features should be at the same scale. To achieve this, a standard scaler is created to ensure every column has zero mean and a standard deviation of 1. The error of the training and test sets are shown in Figure 8 with different regularization strength (alpha). The best alpha is the one that minimizes the holdout error. It is about 100.

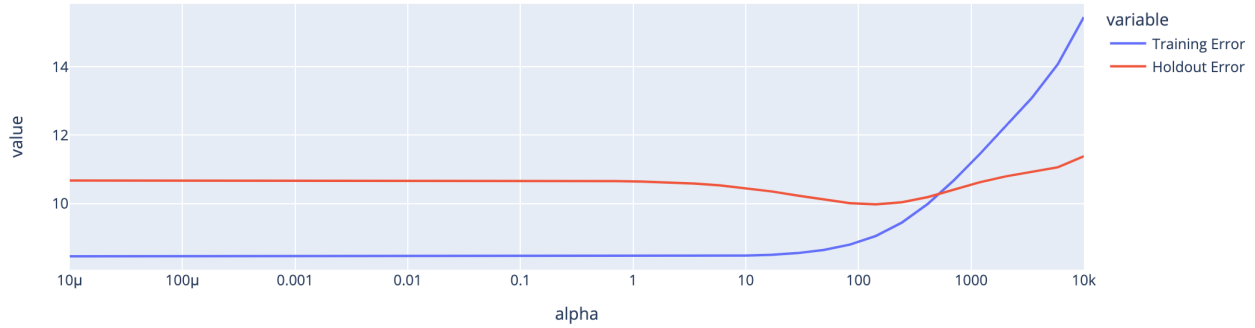

Figure 8. The error of the training and test sets with different regularization strength.

Another approach is K-fold cross validation (Figure 9), which allows us to use more data for training instead of setting aside some specifically for hyperparameter selection. In this project, 5-fold cross validation is used to further confirm the optimal alpha value for regularization. The result is shown in Figure 10. The cross-validation error shows a similar dependency on alpha relative to the holdout error. It also indicates that the optimal alpha is around 100.

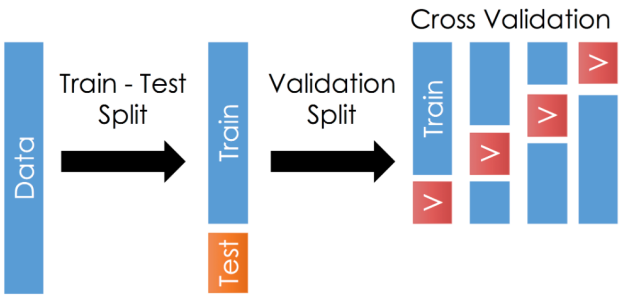

Figure 9. The illustration of the cross-validation method.

Cheng Zhu 3034947319 cheng_zhu@berkeley.edu

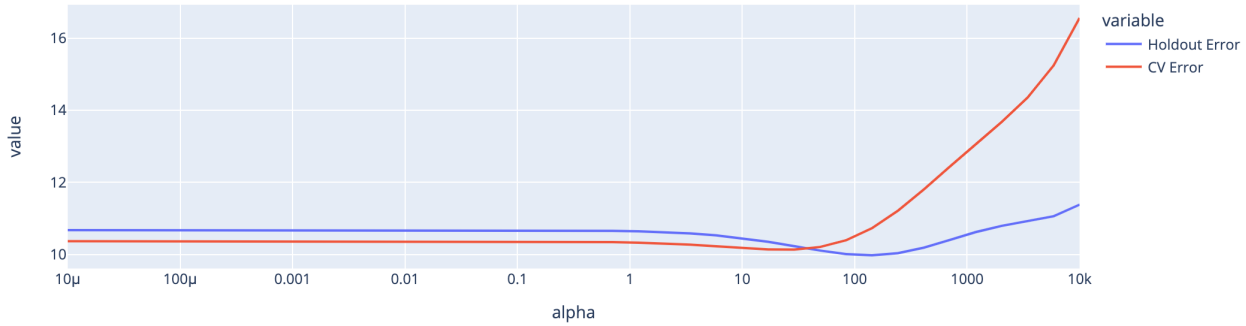

Figure 10. The cross-validation (CV) error and test error with different regularization strength.

5. Discussion

When using a ridge regression model with alpha equals 100, the holdout RMSE is reduced to 9.99. The holdout RMSE of various models are summarized in table 1. The predicted and real March 2022 death numbers of the test (holdout) set using selected features and L2 regularization are shown in figure 11. The R-squared value of the optimal model is 80.60754785739461%.

Models	with all the features	with selected features	with regularization
Holdout RMSE	16.044361953579422	11.461781283768994	9.99195572688303

Table 1. Comparison of the RMSE of different models.

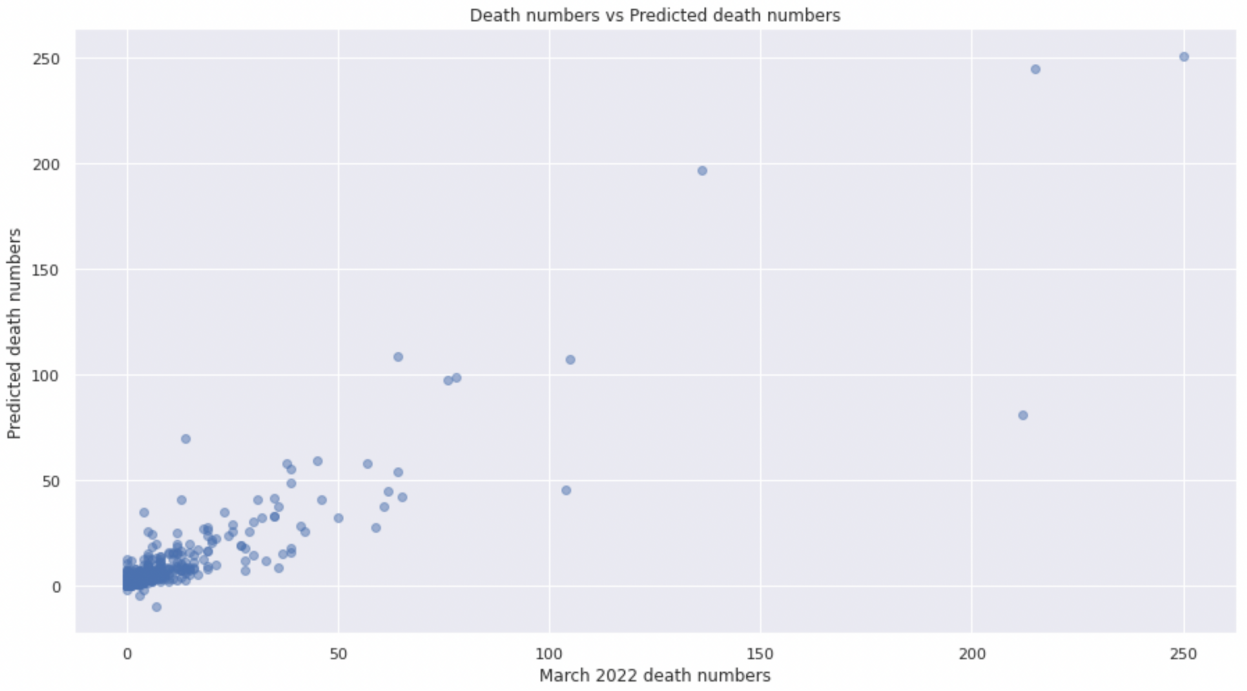

Figure 11. The predicted and real March 2022 death numbers of the test (holdout) set using selected features and L2 regularization.

One limitation in this model is that many rows in the original three datasets are not valid. The three datasets have 3342, 3342, and 3142 rows, respectively. However, when the data is cleaned, there are only 2386 rows left. One row represents one county in every dataset, so almost one-third of the county in the US is not considered in this project. Future efforts should focus on searching for more complete datasets.

There are two interesting findings in this project. The first is that demographic factors, such as age and race distribution, income per capita, etc., are not highly correlated to counties' March 2022 death number. This finding is the opposite of my expectations. I think it is worthwhile to confirm that this is really the case with deeper analysis in the future. The second is that the monthly death and confirmed case numbers after roughly June 2021 are the most useful in predicting the death number for March 2022, which may indicate that the data of the delta variant surge is most closely related to the current situation. In the future, we can explore the reasons behind this phenomenon.

6. Conclusion

Predicting the COVID-19 death number in a county in the upcoming month is a very important task. If a model can best predict the numbers, then local officials can make the right policies beforehand to be prepared for the future wave. This task can be achieved with the help of linear regression models. The death number of March 2022 is closely correlated to the death and confirmed case numbers in the past, particularly the ones after June 2021. The holdout RMSE of the model can be reduced to below 10 by feature engineering and regularization, and the R-squared value is about 80%.

7. References

[1] W. Guan; N. Zhong. Strategies for Reopening in the Forthcoming COVID-19 Era in China. *Natl. Sci. Rev.* **2022**, 9, 3, nwac054.
[2] R. Nair, et al. Predicting the Death Rate Around the World Due to COVID-19 Using Regression Analysis. *Int. J. Swarm Intell. Res.* **2022**, 13, 2.
[3] S. Roy, P. Ghosh. Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. *PLoS ONE* **2020**, 15, 10, e0241165.
[4] https://www.openintro.org/data/?data=county_complete
[5] https://www.pewresearch.org/politics/2022/03/03/the-changing-political-geography-of-covid-19-over-the-last-two-years/
[6] M. Karmakar, et al. Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US. *JAMA Netw. Open* **2021**, 4, 1, e2036462.