BIOSTAT 625 Final Project

Li Liu, Lingxuan Kong, Yichu Wang, Yubo Shao

Introduction

Starting from the end of 2019, COVID-19, a worldwide pandemic, has changed everyone's life forever. Nearly every country is suffering from this pandemic, and, as time goes by, many researchers have developed and updated COVID-19 related data in the United States, we use the data on Kaggle [1] in this project, which is the 3,142 counties of the United States spanning a diverse range of social, economic, health, and weather conditions. County-level data on health, socioeconomics, and weather can help address one of the primary tasks of the uncover challenge, which is to identify which populations are at the greatest risk for COVID19. The data is a combination of data from the New York Times, CDC, and so on, and the time span of the data is from 2020-01-21 to 2020-12-04. When observing the data, we have found several interesting results, such as the relationship between the incidence and mortality rates and two parties in the United States.

In this project, our contribution is twofold. Firstly, plots and other forms of visualization data can impress the audience most, instead of pure numbers, so we try to do data visualization to produce plots, but in a dynamic and faster way. Secondly, as an individual, if we want to select a place to live, under this pandemic, which place might be the best one to move to? We tried to draw inference between environmental factors and some major concerned variables. In summary, we want to complete two tasks in this project: Inference and Prediction. The goal of doing inference is to find significant predictors for our interested outcomes, while for prediction, we want to see whether we can use machine learning to make a prediction on an unknown dataset (test set).

The following sections are organized as follows: Section 2 is data preprocessing and visualization, which shows the basic procedures of our data preprocessing and preliminary visualization results. Section 3 and 4 are summaries of model inference and prediction results. Section 5 is the conclusions and lessons we learned from this group project. Section 6 is the acknowledgments.

Data Preprocessing and Visualization

Details for Data Cleaning

Since the size of our original dataset is 1.38GB, cleaning data requires some tricks. For example, `fread` function in `data.table` package is used in the whole procedure, and it will perform better on computers with many CPU cores. The whole procedure of data cleaning is shown as below:

1. In order to show the prevalence of COVID-19 across the United States, we calculated the state-level incidence, mortality, and fatality rates at different time points based on the county-level number of cases and deaths and the baseline population of each county.

2. We removed Hawaii, Alaska, Virgin Islands, Puerto Rico, Northern Mariana Islands, and only keep the remaining 48 states and Washington, D.C... We also marked the selected states as blue and red states

based on the results of the 2020 presidential election. If a county has no statistics at any point in time, we treat its cases and deaths as 0 if we find the county has not started getting its statistics yet, and if it ends its statistics recording earlier than the selected time point, we will use the reported data of the most recent date before the selected time point.

3. Since calculating the three rates for each state at different dates is actually a repetitive task, we take advantage of `parSapply` and `foreach` function in `Parallel` package and combine it with our own function, which significantly improves the efficiency of our code.

4. According to the purpose of our study, we generated multiple different data sets based on the original data set. Since our original data type is longitudinal, we extracted the data for each different county on the last day of the dataset as our forecast dataset. We further deleted variables with more than 10% missing values, and Joplin, Missouri was also deleted because it had more than 90% missing data. The final dataset we used for our predictive models contained 53 features and 3105 different counties (including the counties of the same name in different states).

5. Finally, we calculated the average morbidity, mortality, and fatality rates of these counties, and introduce an indicator variable for these 3 rates respectively to each of these counties to indicate whether it was higher than the average rate. We used the binary indicator variables as the outcome of our predictive model.

Visualization

In order to reflect the COVID-19 pandemic trends in Republican and Democratic-dominated states at different times, four-time points are selected as the sample for display, which is March 31, 2020, June 30, 2020, September 30, 2020 and December 4, 2020 (December 4, 2020 is the maximum time contained in the dataset). Additionally, in order to show the increase in the prevalence of COVID-19 in red and blue states over time, we calculate the overall incidence rate, overall mortality rate, and overall fatality rate of all red states and all blue states on each day contained in the whole data set, and do a data visualization by line charts. See Figure 1 to 4 for details.

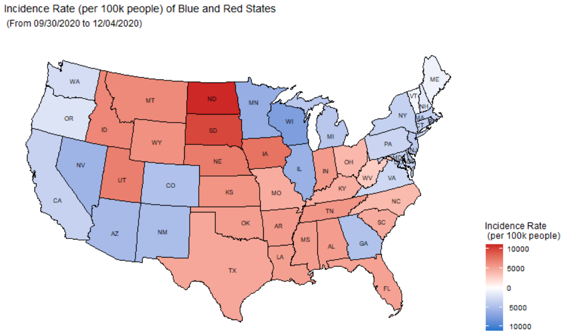

Figure 1: In the map, the intensity of the color reflects the severity of the incidence rate. The fourth quarter map of the incidence rate is used as an example

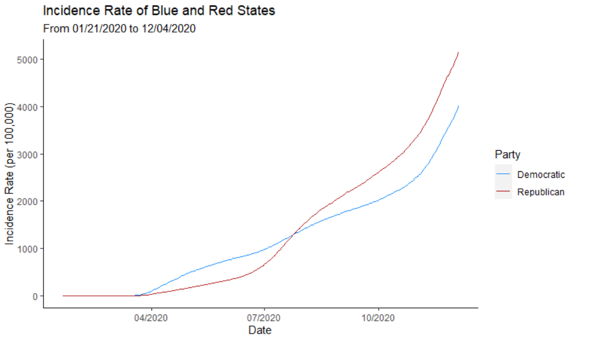

Figure 2: In the incidence rate line chart, the rate of the Democratic Party is relatively large than the Republican Party in the almost first half of the year. Then the Republican Party overtakes the Democratic Party

Model Inference

In this part, we want to use Logistic Regression to find the most significant predictors in our settings.

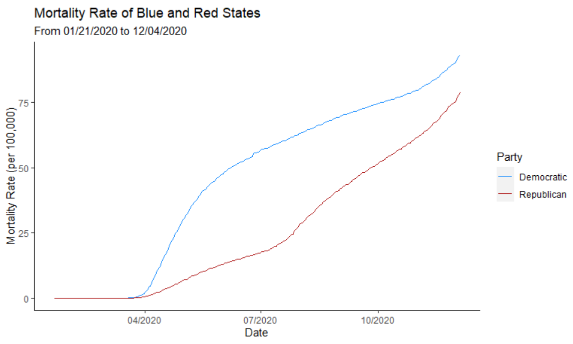

Figure 3: In the mortality rate line chart, the mortality rate of the Democratic Party is higher than the Republic Party for the whole year

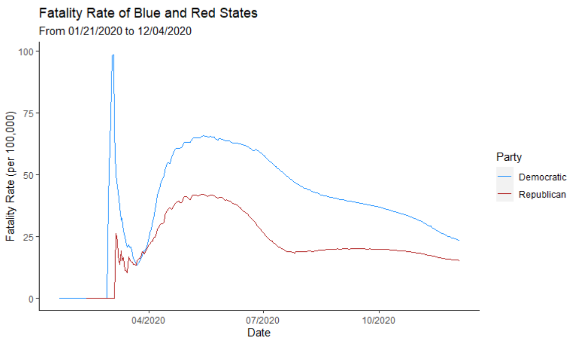

Figure 4: In the fatality rate line chart, the Democratic Party has a similar trend with the Republican Party and it is always higher than the latter

We have in total 53 variables in this dataset, in order to do a forward selection more efficiently, we set up a new workflow and make a more flexible function. As the first step, we fix 53 models with single predictors and get p-values. Those variables with less than 0.05 p-value will be considered as having higher potential to be strongly connected with the outcome we are interested in. Compared to normal forward selection procedure, this step saved much computing time by avoiding fitting a lot of models using variables with low potential to be significant in the final model. In our case, we have 35 variables left after screening, 33% variables were dropped. Then 10 most important variables are selected as predictors of the outcome we interested in. The results are shown in Table 1.

Variable Name	p-value
percent_single_parent_households_CHR	7.41×10^{-38}
percent_low_birthweight	3.63×10^{-36}
percent_fair_or_poor_health	5.42×10^{-23}
violent_crime_rate	5.28×10^{-8}
preventable_hospitalization_rate	2.03×10^{-12}
mental_health_provider_rate	5.75×10^{-6}
percent_severe_housing_problems	4.57×10^{-5}
severe_housing_cost_burden	1.75×10^{-4}
average_daily_pm2.5	1.01×10^{-3}
injury_death_rate	3.64×10^{-3}

Table 1: Top 10 significant variables

Note that this novel procedure produces a lot of flexibility. To begin with, the forward selection function is based on two criteria, AIC and least p-value. We can choose one preferred criteria. Meanwhile, the stopping criteria is also flexible, we set two options in this function. The first stop criteria is the natural one, if no more variables can be included into the model, then the selecting process stops. Another stop criteria is the final number of predictors included in the model is set as a parameter.

Model Prediction

Besides inference, the prediction part is also our interest. Machine learning algorithms are the most common methods used for this objective [3]. In this part, two models are used for prediction, which are Logistic Regression and Random Forest. We also consider 2 other models for this part, which are Support Vector Machine (SVM) and Deep Neural Network (DNN). However, they are not used for prediction because SVM requires all features to be numerical data. In this dataset many variables are categorical data, so even if we can get a prediction result, its theoretical meaning does not exist. As for DNN, since the final dataset is not too big (only thousands of observations), DNN will not get a better result in common sense. For saving time and GPU resources, this method is not taken into consideration, either.

The brief introductions and corresponding results are listed below in two sub-sections. Please note that the data is split into 2 parts, where 85% of the data is training data while 15% of it is test data. The prediction result is obtained from the test data.

Logistic Regression

Logistic Regression is a regression model with the formula

$$\text{logit}(p) = \ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = \beta^T x \tag{1}$$

where k is the number of predictors, $\beta = (\beta_0, \ldots, \beta_k)^T$ and $x = (1, x_1, x_2, \ldots, x_k)^T$. To fit this model, Maximum Likelihood Estimation (MLE) is considered. Note that by mathematical transformation, we have

$$P(y|x;\beta) = h_\beta(x)^y (1 - h_\beta(x))^{1-y} \tag{2}$$

where $h_\beta(x) = \frac{1}{1+e^{-\beta^T x}}$ for each data point (x, y). So the likelihood function can be written as

$$L(\beta) = \prod_i P(y_i|x_i;\beta) = \prod_i h_\beta(x_i)^{y_i} (1 - h_\beta(x_i))^{1-y_i} \tag{3}$$

Then the model can be solved by log transformation and numerical optimization methods such as Stochastic Gradient Descent can be used for solving the objective problem since it has no closed-form solution. We ignore this part since `glmnet` in R can help us tackle these details.

Random Forest

Random Forest is another machine learning method exploiting Ensemble Learning principles. In classification mode, it trains many decision trees and uses voting to provide a final decision, where for each decision tree, they will generate decisions by enumerating conditions related to the decision on the trees.

Three important properties of it support us to adopt this method. Firstly, ensemble learning can make several weak classifiers collaborate for a strong classifier. Secondly, Random Forest does not have overfitting problems [2]. Lastly, since each tree is trained independently, parallel computing techniques (`registerDoParallel(·)`) play an important role in accelerating the training of Random Forest.

Results and Comparison

Here are the results for Logistic Regression and Random Forest. The outcome is the binary form of fatality rate.

true pred	No	Yes
No	335	57
Yes	70	158

Table 2: Logistic Regression

true pred	No	Yes
No	352	135
Yes	53	80

Table 3: Random Forest

	Acc	AUC
LR	82.1%	80.66%
RF	75.16%	72.63%

Table 4: Comparison

Surprisingly, Logistic Regression behaves much better than Random Forest.

Conclusion

1. Although we thought that parties can be one significant feature that affects our interested outcomes, this is not the case by experiments, which shows that media and superificial visuailization result may confuse people a lot in interpreting results.
2. Even if Random Forest has been applied broadly in real datasets, in this project it behaves as an inferior baseline, which proves the "no universal" theorem [4], saying there is no universal machine learning algorithm that can get the best result on all datasets.
3. How to deal with categorical data is worth considered, besides transforming categorical features into factors, some embedding methods such as one-hot encoding are also worth considered, especially when using machine learning algorithms. This part is one of our future works for this project.

Discussion: An Open Question

There is a problem arising when completing the inference and prediction of the models. As mentioned before, the outcome variables are binary forms of mortality, fatality, and incidence rate. When we put the continuous (original) version of these three variables as predictors, we can find a near 100% accuracy in the test set. However, these three variables are not the top 10 significant variables in Logistic Regression. It is so strange to find that “mortality rate is not a significant predictor of mortality rate”.

Here is one possible explanation for this phenomenon: since decision trees in Random Forest will select a value x as boundary value and then divide the value as a classification boundary (for example, $X \leq x$ will be assigned 0 and for $X > x$, 1.). When the boundary is chosen as the average value of these three rates, this feature can be transformed just the same as the three outcomes, which will lead to a near 100% accuracy since the feature is nearly the same as the outcome. But since this procedure is missing in Logistic Regression and this transformation eliminates the statistical information of the three values, it is reasonable to say these three predictors will never be top significant features for the outcomes.

Acknowledgements

Yubo Shao completes the data cleaning work perfectly, making the data really easy to handle for Li Liu (working for model prediction) and Lingxuan Kong (working for model inference). The visualization graphs are drawn by Yichu Wang, who also vividly shows the trend of COVID-19 progress by an animation shown on the slides. The open question is found by Li Liu and Lingxuan Kong. The report is collaborated by all 4 team members and we are really grateful and excited to work as a team to complete this interesting final project.

References

[1] Data source. https://www.kaggle.com/johnjdavisiv/us-counties-weather-health-hospitals-covid19-data?scriptVersionId=48607875.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[4] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.