ECO 395 Project: COVID-19 Death Rate Prediction

Xiaohan Sun/ Liyuan Zhang/ Evelyn Cheng

5/12/2021

Abstract

The pandemic of COVID-19 now is becoming the world's greatest threat, which is a problem we all need to fight together. A prediction of COVID-19 death rate can help both the government and hospital to plan and mobilize social resources reasonably and rapidly. In this project, we have created 5 models to predict the death rate of COVID-19 base on the data from the Centers for Disease Control and Prevention. ROC curve, precision and accuracy rates are used to select a best model. The result shows that tree model and logit model have a better performance than others.

Introduction

COVID-19 has now become a global pandemic, spreading rapidly among individuals in most countries in the world, and therefore has become the world's greatest threat. As of May 11, 2021, a total of 33,515,804 cases have been diagnosed in the United States, with a cure rate of 79.1% and a fatality rate of 1.8%. According to data released by Johns Hopkins University, the cumulative number of deaths from the COVID-19 in the United States exceeds 580,000; data released by the US Centers for Disease Control and Prevention shows that there are close to 580,000 deaths.

In the COVID-19 epidemic raging around the world, data analysis and data visualization have once again released energy that cannot be underestimated. From infection tracking, case tracing and time sorting at the beginning of the epidemic, to forecasting of the epidemic, new topics and discourses about the spread of the virus flooded in.

The forecast of COVID-19 case growth and mortality is critical to the decisions of political leaders, businesses and individuals responding to this pandemic. COVID-19 is an emerging disease, and initially there is no historical data to guide scientists to predict its impact on the country over time. The ability to predict the progress of this pandemic is essential, and is aimed at fighting this pandemic and controlling its spread. We considered 5 different models, and we used these models to predict the number of death cases and rate in order to provide reference values for people in need. The goal of our project is first to find the best predictive model in daily situations, and secondly to utilize these models to prepare more for the health care system.

This project uses health data from February 2020 when the epidemic broke out to April 2021. It also integrates race, gender, age, state of residence, exposure to the environment, and access to health care and resources to predict Covid-related deaths and death rates for all parts of the United States.

Data and Methods

Data Preprocessing

Our primary source of data is Data.CDC.gov, which is one of the major operating components of the Department of Health and Human Services and is recognized as the nation's premiere health promotion, prevention, and preparedness agency. The data including 522743 rows and 10 variables. Below table is about the detail of each variable.

Variable	Description
case_month	Date associated with disease or specimen collection
res_state	State of residence
age_group	Age group [0-17 years;18-49 years;50-64 years; 65+ years]
sex	Male and Female
race	American Indian/Alaska Native; Asian; Black; Multiple/Other; Native Hawaiian/Other Pacific Islander; White
exposure_yn	In the 14 days prior to illness onset, did the patient have any of the following known exposures: domestic travel,international travel, cruise ship or vessel travel as a passenger or crew member, workplace, airport/airplane, adult congregate living facility (nursing, assisted living, or long-term care facility), school/university/childcare center, correctional facility, community event/mass gathering, animal with confirmed or suspected COVID-19, other exposure, contact with a known COVID-19 case?
symptom_status	The symptom status of this person
hosp_yn	Was the patient hospitalized?
icu_yn	Was the patient admitted to an intensive care unit (ICU)?
death_yn	Did the patient die as a result of this illness?

Since the sample sizes of American Indian/Alaska Native, Multiple/Other, Native Hawaiian/Other Pacific Islander in `race` are not enough, we merged these categories into a new category named "Other". Also, for the `res_state`, we merged the categories with small sample size into a new category named "Other".

Before fit models, we need to pick up and engineer the features we used in the models. From the table above, most features are categorical variables, thus we need to preprocess these features first. Here are two methods we use:

1. Encode categorical variables as dummy variables: instead of using "one-hot encoding", we choose to assign numbers manually to encode categorical variables. For example, we encode `sex` by assign 1 to "Male" and 0 to "Female", but they both in the one column `sex`.

2. Factorize categorical variables: we factorize all the categorical variables in order to do classification.

Both methods are applied to `res_state`, `age_group`, `sex`, `race`, `exposure_yn`, `symptom_status`, `hosp_yn`, `icu_yn`, `death_yn`.

Models

In this project, we used 50% of data set to train 5 models and the rest of data set to test the performance of these models in order to find a best one to predict the COVID-19 death rate. Here are the five models:

1. Linear model: in this model, we use dummy variables (`res_state`, `age_group`, `sex`, `race`, `exposure_yn`, `symptom_status`, `hosp_yn`, `icu_yn`) to predict COVID-19 death rate.

2. logit model: in this model, we use factorized categorical variables to do classification. (features we used are same as linear model.)

3. Naive Bayes model: in this model, we use factorized categorical variable `death_yn` for a class label, dummy variables (`res_state`, `age_group`, `sex`, `race`, `exposure_yn`, `symptom_status`, `hosp_yn`, `icu_yn`) as a set of features to predict COVID-19 death rate.

4. Tree model: in this model, the "stop points" are (1) if the split improves the deviance by a factor of 0.0002 (0.02%), (2) it has at least 30 observations for controlling tree growth.

5. Random Forest model: in this model, we use factorized categorical variables to do classification. (features we used are same as linear model.)

Since these variables we used are all categorical variables, we decide to use out-of-sample accuracy, precision and ROC curve to evaluate the performance of these models.

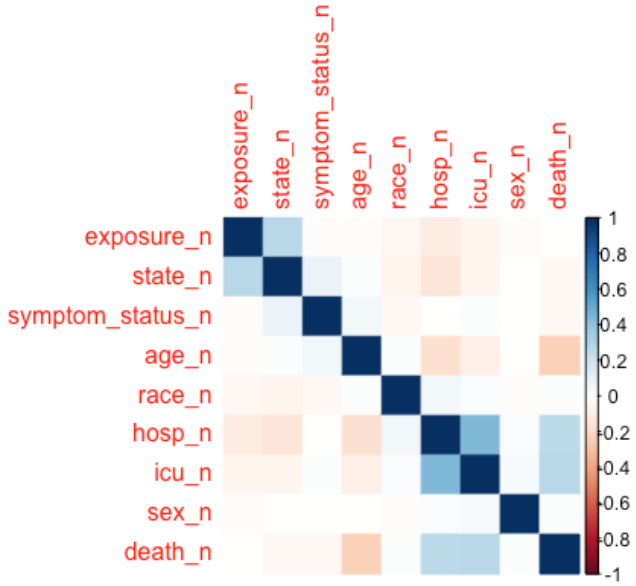
1. Accuracy rate: correct predictive rate
2. Precision: positive predictive value = 1 - FPR
3. ROC curve : false positive rate (FPR), (The true positive rate) TPR

Results

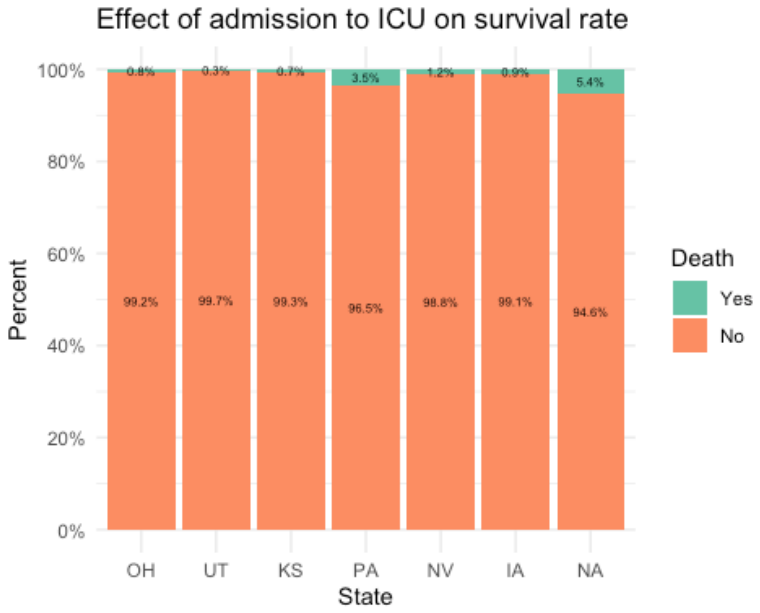
In order to explore the relationship between variables more intuitively, we use data visualization, a graphical representation of the potential relationship between variables.

exploration of data

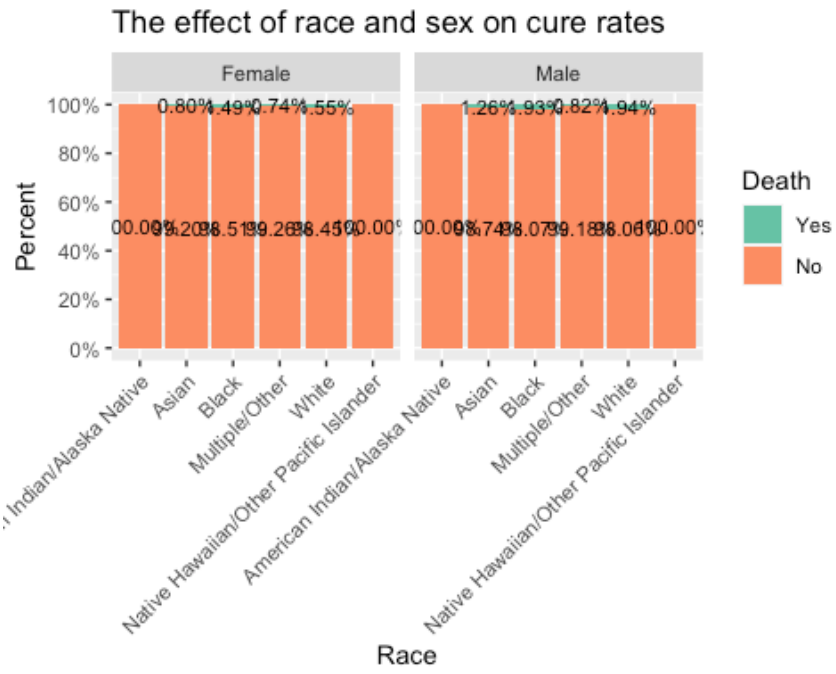
First of all, let's look at the general relationship between features.



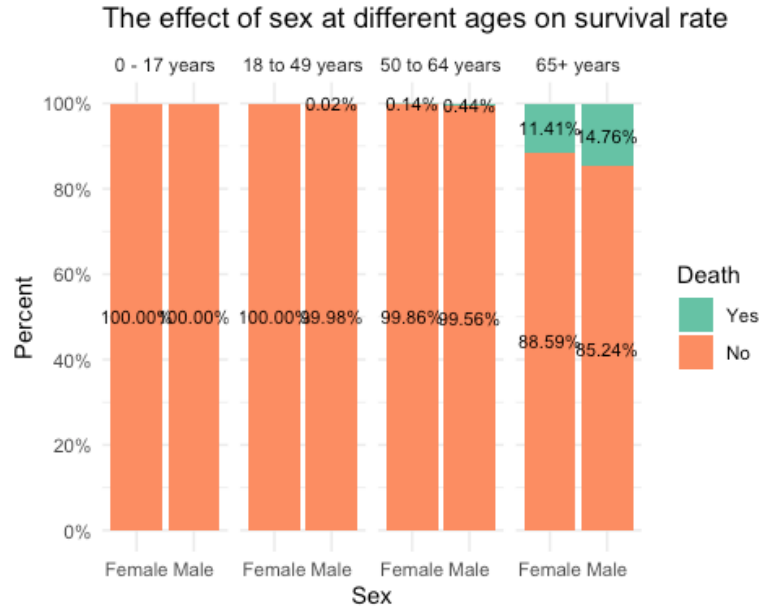
Second, we mapped the survival rates of the populations in the different states. From the chart, we find that Utah has the lowest survival rate,which is 0.8% . Because there are many gaps in the data statistics, we only calculated the mortality rates of the 6 states with large sample size, and classified the data of the rest states as Other.



Third, We looked at the effect of race and sex on the cure rate for COVID-19, and we found that white men had the lowest cure rate, at 98.06% . American Indian/Alaska Native and Hawaiian/Other Pacific Islander have the highest cure rates of 100% . We found that the effect of race on the cure rate of Covid-19 was not significant, with a difference of only 1.94% between the highest and lowest. So we think this is an unsignificant variable

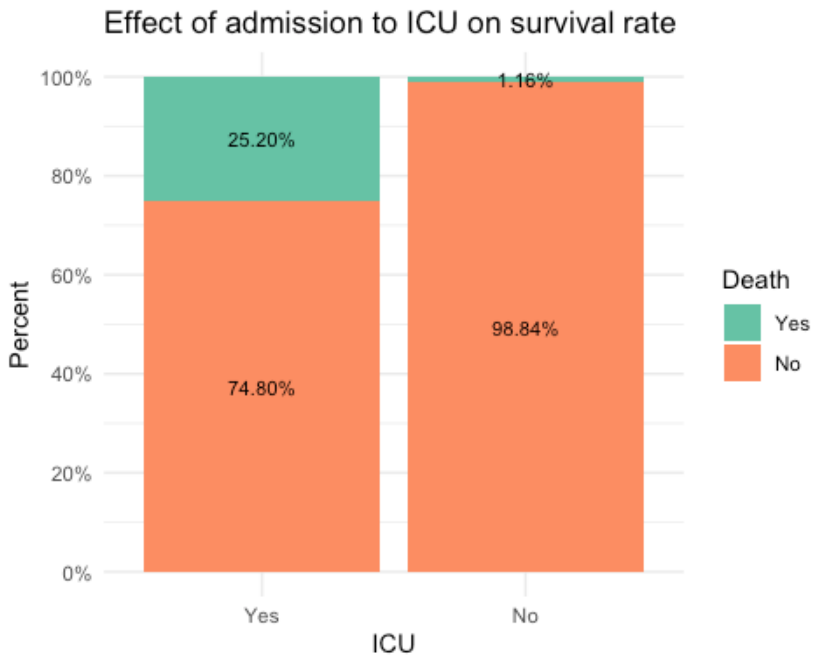


Below is the table about the effect of sex at different ages on survival rate of Covid-19. As can be seen clearly in the diagram, the cure rate decreases as the patient ages incease. In the under-17 group, all COVID-19 patients were cured, while in the over-65 group, only about 88% were cured. We can also find that the cure rate is lower in Male than in Female. The lowest cure rate was found in male over 65, who had a cure rate of only 87.31%. Thus, we can draw a conclusion that, sex and age group is an import variable to impact the cure rate of Covid-19.



To find out the relation between the patient whether the patient admitted to an intensive care unit (ICU) and the Covid-19 cure rate, we draw out the table about the Effect of

admission to ICU on survival rate. For the table below, we can find that patients who admitted to the ICU have a much lower cure rate(73.69%) than those who do not(98.88%). This is a reasonalbe and common sense phenomenon, because the symptoms of patients entering the ICU will be more serious. We therefore used the ICU as an important factor in determining the rate of cure.



Fit Models

Linear Model

Confusion Matrix:

```
##      y_hat_lm
## y_lm       0      1
##   No  253725   3468
##   Yes   2905   1273
```

Accuracy:

```
## [1] 0.975617
```

Precision:

```
## [1] 0.2685088
```

Logit Model

Confusion Matrix:

```
##        y_hat_logit
## y_logit      0      1
```

```
##      No  256564    629
##      Yes   3495    683
```

Accuracy:

```
## [1] 0.9842217
```

Precision:

```
## [1] 0.5205793
```

Naive Bayes Model

Confusion Matrix:

```
##       yhat_test_nb
## y_test      0      1
##    No  255536   1657
##    Yes   2920   1258
```

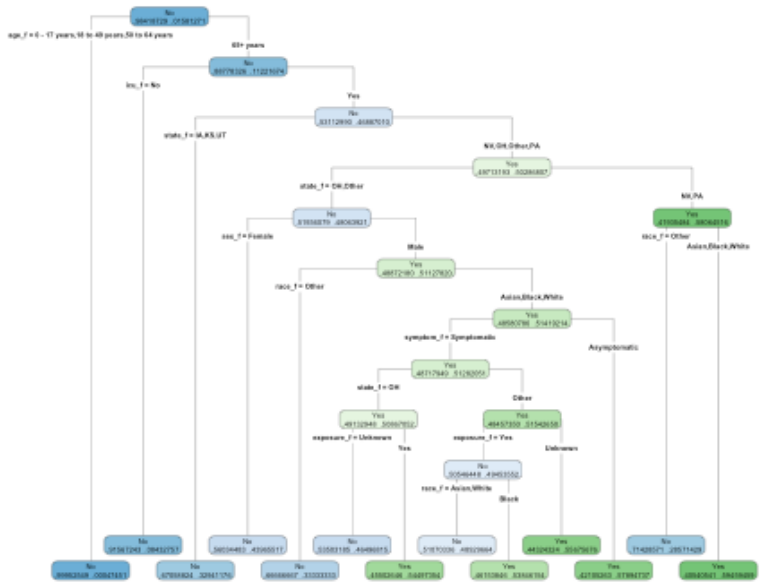
Accuracy:

```
## [1] 0.9824885
```

Precision:

```
## [1] 0.4315609
```

Tree Model



Confusion Matrix:

```
##        y_hat_tree
## y_tree      0      1
##    No  255830   1363
##    Yes   2974   1204
```

Accuracy:

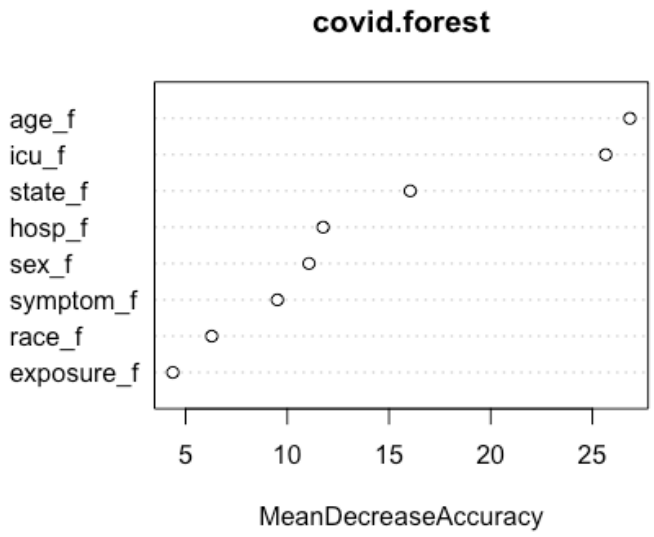
```
## [1] 0.9834067
```

Precision:

```
## [1] 0.46903
```

Random Forest Model

The graph below is variable importance plot, from the graph, we know that the `res_state`, `age_group` and `icu_yn` are the most important features in predicting death rate, `sex`, `race`, `hosp_yn` are less important features.



Confusion Matrix:

```
##        y_hat_tree
## y_tree      0      1
##    No  256901    292
##    Yes   3832    346
```

Accuracy:

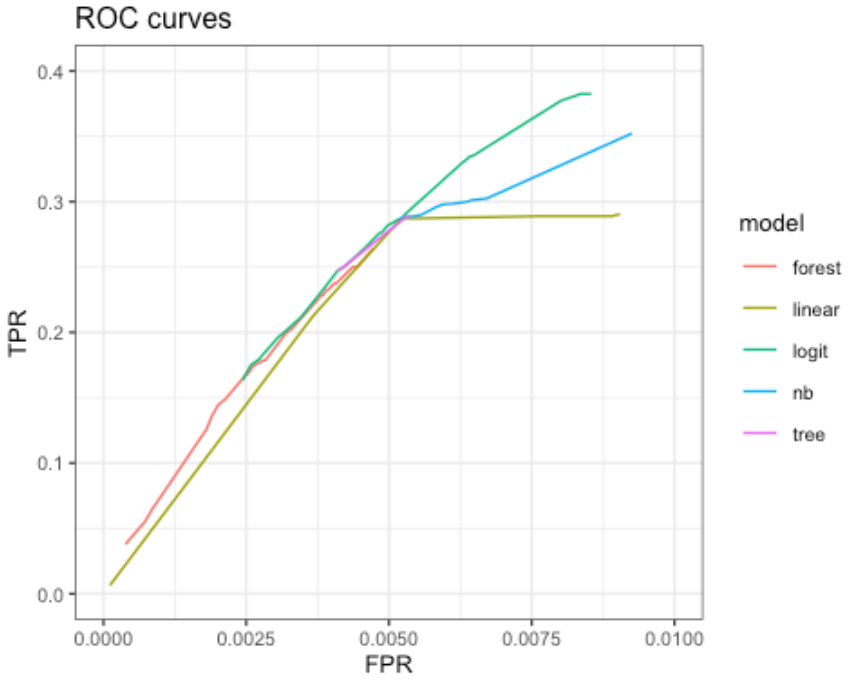
```
## [1] 0.9842217
```

Precision:

```
## [1] 0.5423197
```

ROC Curve

Since a ROC curve that is more "up and to the left" represents better performance, from the graph below, we know that logit model and tree model have a better performance than others.



Conclusion

For model accuracy, there is no significant difference among the five models. For model precision, we find that random forest model > logit model > tree model > naive bayes model > linear model. And for ROC curve, we notice that tree model, logit model and random forest model are better. Thus, tree model and random forest model, in this project, can predict the COVID-19 death better.

There are also many limitations in our model. For instance, in the raw data, the number of records that people death is much smaller than the number of records that people not death, which hurts the performance of models.

In the real life, a good predictive model for COVID-19 death rate can be useful to many area.

For government, the policy maker can make an appropriate plan in advance in order to avoid social panic and prevent further spread of the disease. For example, if there are probably many people death by the disease, the government could make a plan for disposal human remains which can be infection source for people.

For hospitals, by using the predictive model, they can plan the hospital beds properly. Also, for patients with a high probability of death, hospitals can prepare in advance and pay high attention to these people.