final

chengkan_tao ECO 395 ct32737

```
devtools::install_github(c('rstudio/rmarkdown', 'yihui/tinytex'))
tinytex::install_tinytex()
```

Abstract

Coronavirus Disease 2019 broke out at the end of 2019 and spread all over the world in few months so that millions of people lost their lives. Hence, my final project focuses on COVID-19, especially patients who died as a result of this disease. I plan to find out features of those who died of COVID-19. Here I estimate a tree model, finding binary splitting rules by features in the dataset. Build a logit model to predict the probability that patients can survive from COVID-19. The conclusion is Asian, elder, cormodity, and ICU(severely ill patients) are key words. They have higher probability of death once they get COVID-19

Introduction

I believe COVID-19 and American Presidential Election can be the top topics in 2020, and I am not the only one to set COVID-19 as research object. (Actually, I prefer text mining from Trump's tweets.) In 2020, too many people around the world lost their lives because of coronavirus, so despair can be taken for granted if someone is infected with COVID-19. Those who are hospitalized in ICU can easily get unsure about their own life. The project tries to find out features of the dead, so that patients with different features would get confidence from this result. Of course, it doesn't

mean I don't care about those who have same characteristics. They should face up to the trouble bravely. This project can also act as a warning. Not everyone like prohibition of going out. Someone still goes out without a mouth mask and holds giant public parties. The statistical result of the project would tell which group of people have a relatively high probability of death if they are infected with COVID-19. This group of people should try to reduce the chance of going out and be alerted to symptom. These are reasons why I choose COVID-19.

Method

Data: I download dataset from Centers for Disease Control and Prevention. (Website: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf) The data set I get includes 12 columns (variables) and 25009120 rows (patients). I catch 500000 observations as my data set. I delete all observations including 'NA', 'missing' and 'unknown'.After deleting these observations, I get 14805 observations. These variables includes four timestamp variables, five status and three characteristics of patients. Four timestamps are the earlier of the Clinical Date or the Date Received by CDC, date case first reported to the CDC, date of first positive specimen collection, symptom onset date. Five status are case status (current_status), hospitalization status (hosp_yn), ICU admission status (icu_yn), death status (death_yn), presence of underlying comorbidity or disease (medcond_yn). Three characteristics are age_group, race_ethnicity_combined, and sex. Age group contains 0-9 years, 10-19 years, 20-29years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years, 80years+. Race and ethnicity consists of 'Other', 'Black', 'White', 'Asian', 'Hispanic', 'American_Indian' and 'Native Hawaiian'. Sex contains 'Female', 'Male', 'Other'. Number of observations with other sex is relatively small ,and the variance of it is pretty high. So I didn't take it into my project. There are nine kinds of elements in the age_group, so I set a new variable age. It includes 0-8. 0 match 0-9 years, 1 matched 10-19 years, 2 matches 20-29 years, 3 matches 30-39, 4 matches 40-49 years, 5 matches

50-59 years, 6 matches 60-69 years, 7 matches 70-79 years, 8 matches 80 years+. There are seven kinds of elements in the race_and_ethnicity. I create 6 dummy variables excluding 'Native Hawaiian'. I will use a random subset of this data set with 60000 observations when I build the model because my computer refuses to work with 14805 observations. method: In the data set, all i get are character variables and timestamps. These character variables are only transferred to dummy variables. I will use logit model because my independent variable is death, which is a category variable between 0 and 1. Then calculate the false positive rate and false discovery rate to draw a ROC curve. And I will use classification tree to get splitting rules. In my Rmarkdown, I don't read raw data(2.5G,it's too big)I will show my pre-process in my R. And in my Rmarkdown, I will import processed dataset(small_survey)

Result

There are 20 variables. These are character variables and dummy variables. I will use character variables to build a classification tree and dummy variables to run a logit regression.

```
data1 = small_survey %>%
  group_by(race_ethnicity_combined,sex) %>%
  summarize(race_pct = sum(death_yn == 'Yes')/n())

ggplot(data = data1)+
  geom_col(aes(x=sex,y=race_pct)) +
  facet_wrap(~race_ethnicity_combined)
```

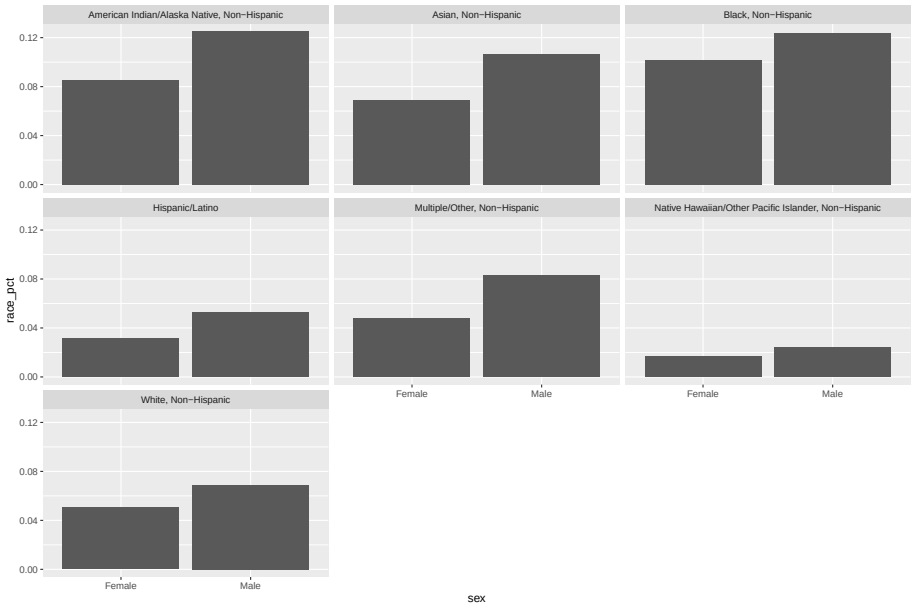


Figure 1 shows the death ratio of the patients which is divided by sex and race. The death ratio is equal to count of the death in the group divided by sum of all members in the group. Through the Figure 1, women's death ratio is always lower than men's ratio, whatever patients' race is. And, we can find death ratio of Asian is pretty high. Male Asian's death ratio is close to 17.5% and female Asian's death ratio is over 10%. Thus, Asian should pay more attention and try to stay at home.

```
data2 = small_survey %>%
  group_by(age_group,sex) %>%
  summarize(age_pct = sum(death_yn == 'Yes')/n())

ggplot(data = data2)+
  geom_col(aes(x=age_group,y=age_pct)) +
  facet_wrap(~sex)
```

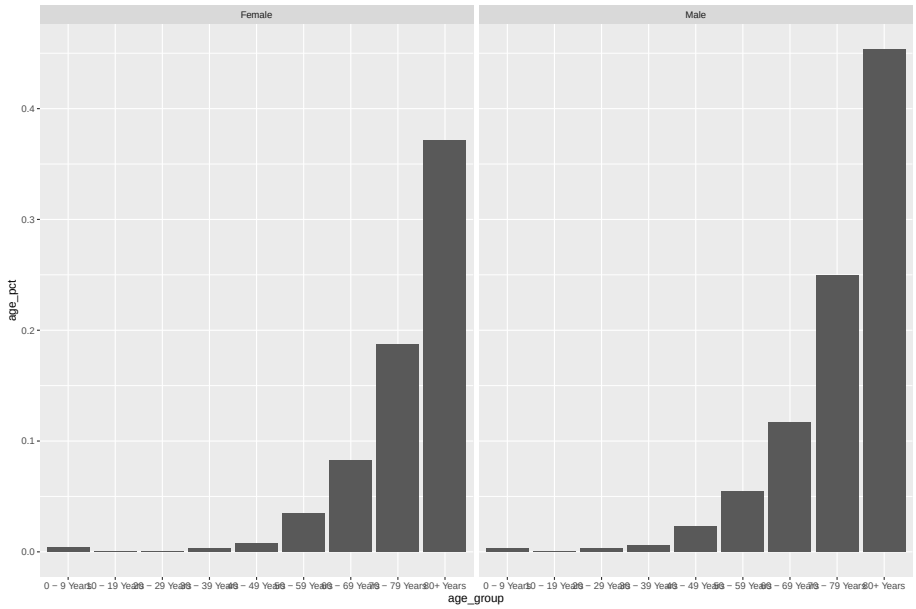
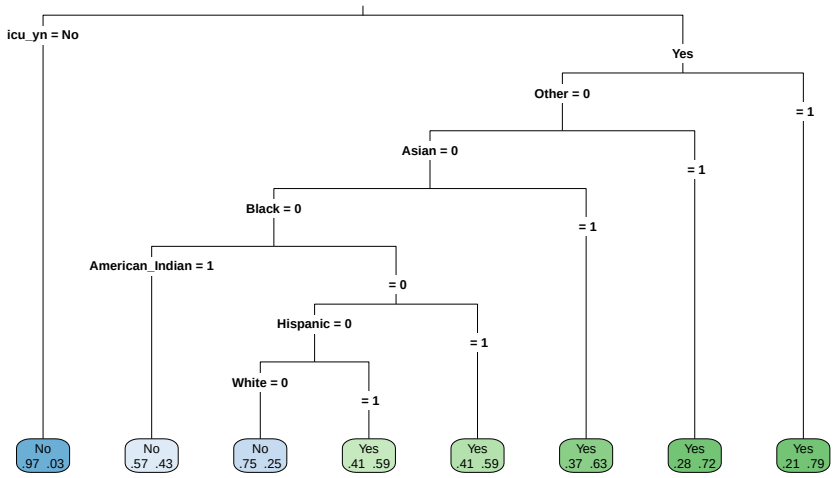


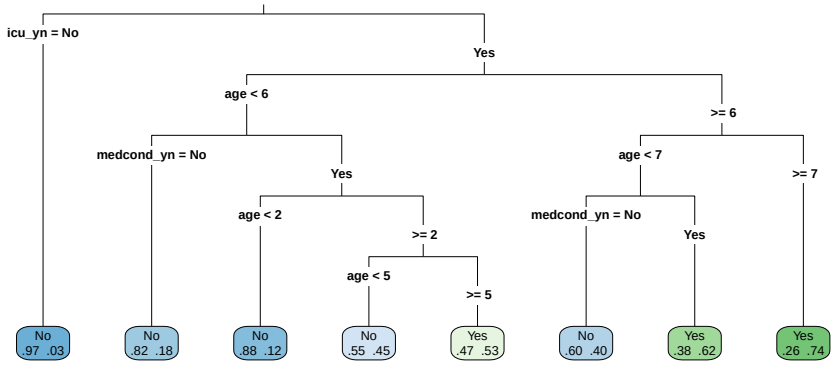
Figure 2 shows a expected trend. Death ratio is high in the 60-69 years group, 70-79 years group and 80 years+ group. Men who are more than 80 years old have a nearlt 50% death ratio when they are infected with COVID-19. The death ratio increases with the age going up. The elder are less likely to survive, facing COVID-19. Similarly, we find that female has less death ratio than male in all age groups.

```
tree = rpart(death_yn~icu_yn+White+Asian+Other+Hispanic+Black+American_Indian,data =sma
rpart.plot(tree, type=3, extra=4)
```



Now, I build a classification tree model. Notice that Yes means death. The tree model includes dummy variables: death_yn, icu_yn, White, Asian, Other, Black, American_Indian. In the picture we find if a patient is hospitalized in ICU, the probability of death is very high. Two final nodes if icu_yn = Yes is no. They are Native_Hawaiian and American_Indian. They are fortunate and maybe their habitat is not heavily affected by COVID-19. But other races are unfortunate when they are in ICU. ICU = Yes means patients are critically ill patients. These kinds of patients would lose their lives at any time. Doctors and Nurses can hardly save their lives. And if patients is not in ICU, their probability of death is much lower. Patients with mild symptoms are more likely to survive from the disease.

```
tree = rpart(death_yn~icu_yn+age+medcond_yn,data =small_survey,control = rpart.control(
rpart.plot(tree, type=3, extra=4)
```



Then the second tree shows that if patients in ICU are more than 50 years old and have comorbidities or other diseases, they are dangerous. Maybe COVID-19 is not the main cause, comorbidities can be also fatal. So people who have chronic diseases or other medical history should be careful. From this tree model, we can also find that young people have higher abilities to cope with diseases because even they who have comorbidities in ICU, the probabiliy of death is lower, but not very low.

Then I will build a logit model to find the effect of age, race and icu on death in the training data set.

```
survey_split = initial_split(small_survey, prop = 0.8)
survey_train = training(survey_split)
survey_test = testing(survey_split)
logit_death = glm(death ~ icu + age + White + Other + Hispanic + Black + Asian + Americ
coef(logit_death)%>%round(2)
```

```
##     (Intercept)             icu             age           White           Other
##          -10.40            3.41            0.99            1.28            1.47
##        Hispanic           Black           Asian American_Indian
```

```
##            1.78            1.84            2.08            1.11
```

First, we should consider the odd. odd = probability of death/(1-probability of death). The changes of variables affect the odd through coefficients. Through the coefficient matrix, there are many dummy variables. For example, patients in ICU are in the same age group. White have a lower probability than Asian. When white is equal to 1, other dummy variables are 0 and it's the same when Asian is equal to 1. So the odd will be divided by e^1.43 and times e^0.74 if it converts from Asian to White. The odd decreases and the probability also decreases. Obviously, age increases from one group to next group, the odd and probability also increase. Icu will increase the probability too.

```
phat_death_test = predict(logit_death,survey_test,type='response')
yhat_death_test = ifelse(phat_death_test>0.5,1,0)



confusion_table_death = table(y=survey_test$death,yhat=yhat_death_test)
confusion_table_death
```

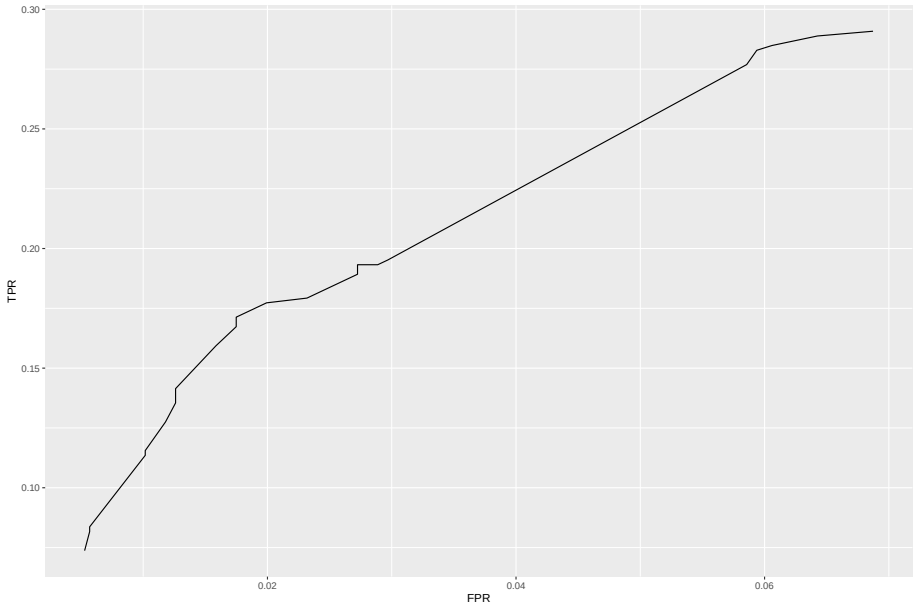
```
##    yhat
## y      0    1
##   0 2743   39
##   1   99   80
```

```
sum(diag(confusion_table_death))/sum(confusion_table_death)
```

```
## [1] 0.9533941
```

```
threshold = seq(0.85, 0.15, by=-0.005)
roc= foreach(k = threshold, .combine='rbind') %do% {
  phat_death_test = predict(logit_death,survey_test,type='response')
  yhat_test = ifelse(phat_death_test > k, 1, 0)
  confusion_matrix = table(y = survey_test$death, yhat = yhat_test)
```

```
  outcome = data.frame(TPR = confusion_matrix[2,2]/sum(survey_test$hosp==1),
                       FPR = confusion_matrix[1,2]/sum(survey_test$hosp==0))
  rbind(outcome)
} %>% as.data.frame()
ggplot(roc) +
  geom_line(aes(x=FPR, y=TPR))
```



This is the ROC curve and the table between y and yhat. The out of sample accuracy is about 95%. Pretty well.

##Conclusion I just want to find what kind of patients face the great danger. From the result we know that Asians have a higher probability of death than other race when infected with COVID-19. Severely ill patients who are in the ICU are very dangerous. The elder should avoid being infected because they have nearly 40%-50% probability of death, especially those have chronic diseases. Both comorbidities and COVID-19 can be killer