MA 679 Final Report

Group 7

2021/5/4

Introduction

After having some initial exploration of SEER data, we decided to see whether there is a bias between the doctor's decision and the treatment guideline in terms of each observation's race and gender. The treatment guideline is based on the TNM cancer stage, we only have information about the AJCC cancer stage, which means we cannot match all cancer types with their guidelines. Therefore, we only chose cancer on Salivary Gland to do further exploration since its treatment matches well with the data.

Data Processing

In the data processing step, we first selected all the Salivary Gland Cancer. In the 'NCCN Guidelines' file, there is no treatment guideline information for stage 'IVC' and 'IVNOS' of cancer Salivary Gland. We removed observations of the 'IVNOS' stage since there are only 6 observations and this amount is trivial compared with the whole dataset. As for the 'IVC' stage, according to the information provided in 'head-and-neck.pdf' page 87, there is no preferred treatment, and the treatment should be individualized based on patient characteristics. Thus, we removed all observations of the 'IVC' stage because we could not decide whether the given therapy follows the guideline or not. Other than these two stages, based on the `NCCN Guidelines`, we found that only stage 'IVB' has a different preferred number one therapy from others, its therapy1 is Radiation while other stages are recommended to have Surgery. Depending on this information, we create a binary column to indicate whether each individual is given treatment following the guideline or not, in this case, '0' means the treatment does not follow the guideline, '1' means the treatment follows the guideline. Meanwhile, we replace all the blank space and symbols in column names for convenience.

Also, we created some new variables based on the EDA: 1. Insurance2: This is a categorical variable with levels 0,1,2. We find that people in the "Insured" class have the lowest rate of being given treatment not following the guideline, while people in the "Uninsured" class have the highest rate of that. "1" indicates the "Insured" class, "2" indicates the "Uninsured" class, and "0" indicates all others.

2. Subsites2: This is a categorical variable with two levels 0 and 1. We find that in the subsite class "C08.1-Sublingual gland", all the observations are given treatment following the guideline. "1" indicates an individual is in this class, "0" indicates all others.

The specific plots and other results from EDA will be shown in the next part.

Because we decided to use machine learning, we also separate the data into training data and test data and guarantee that the proportion of 0 and 1 responses in both two samples are similar.

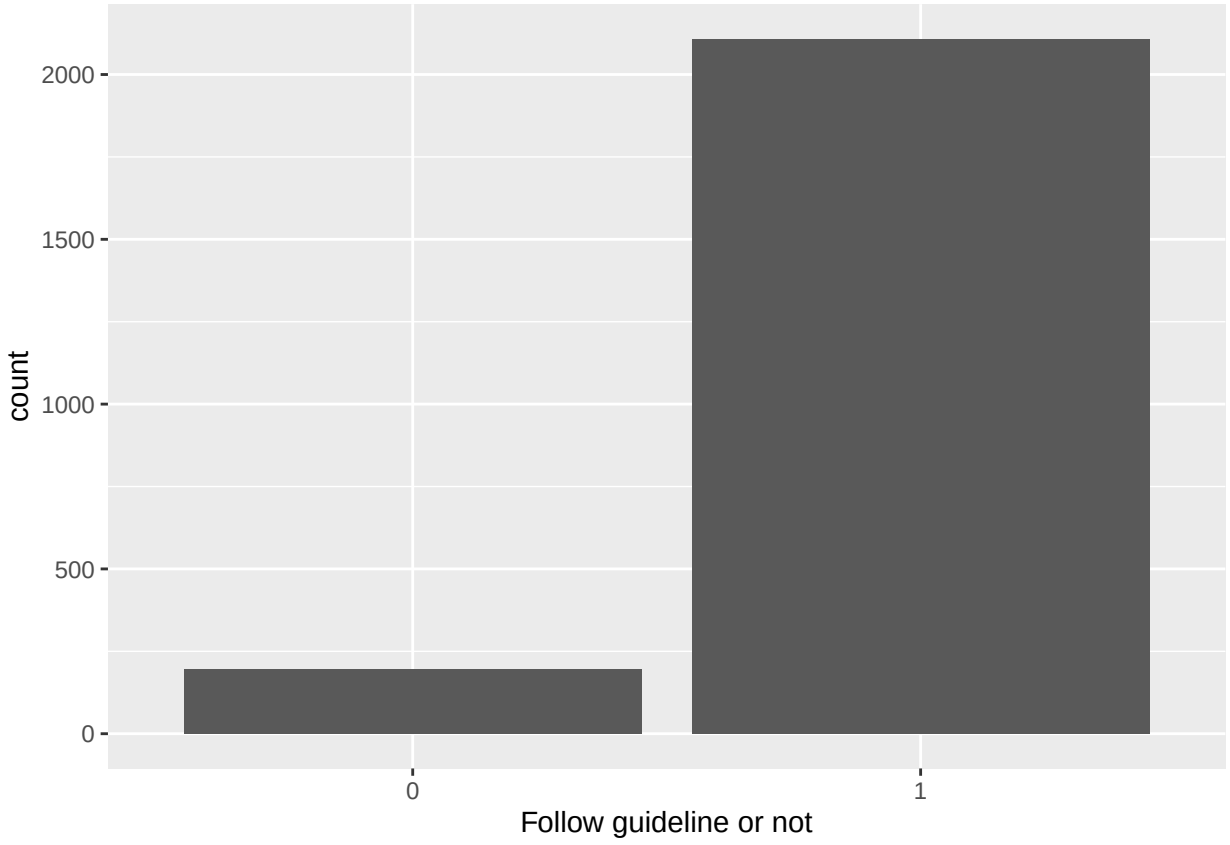
EDA

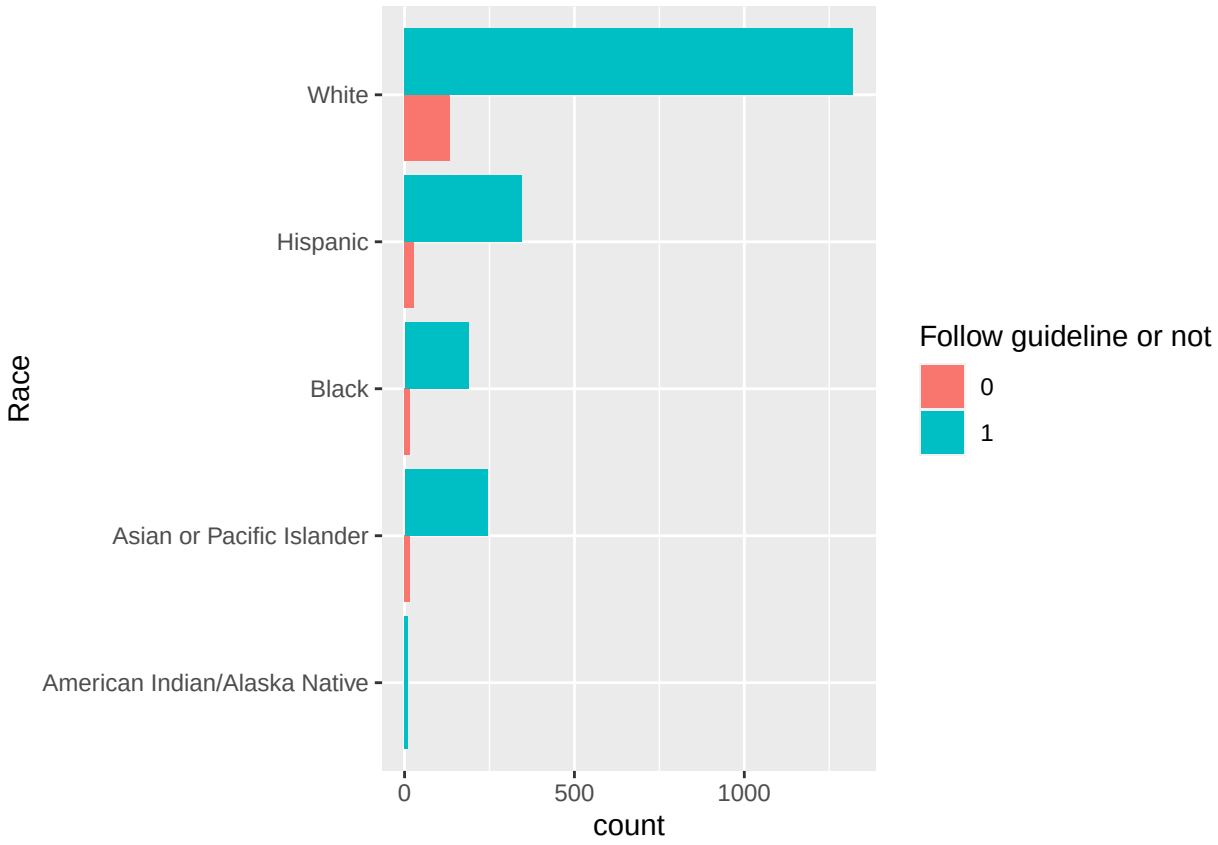
From the bar plots, we observed that this data is imbalanced in terms of Race, Gender, AJCC stage, and Insurance type, etc.

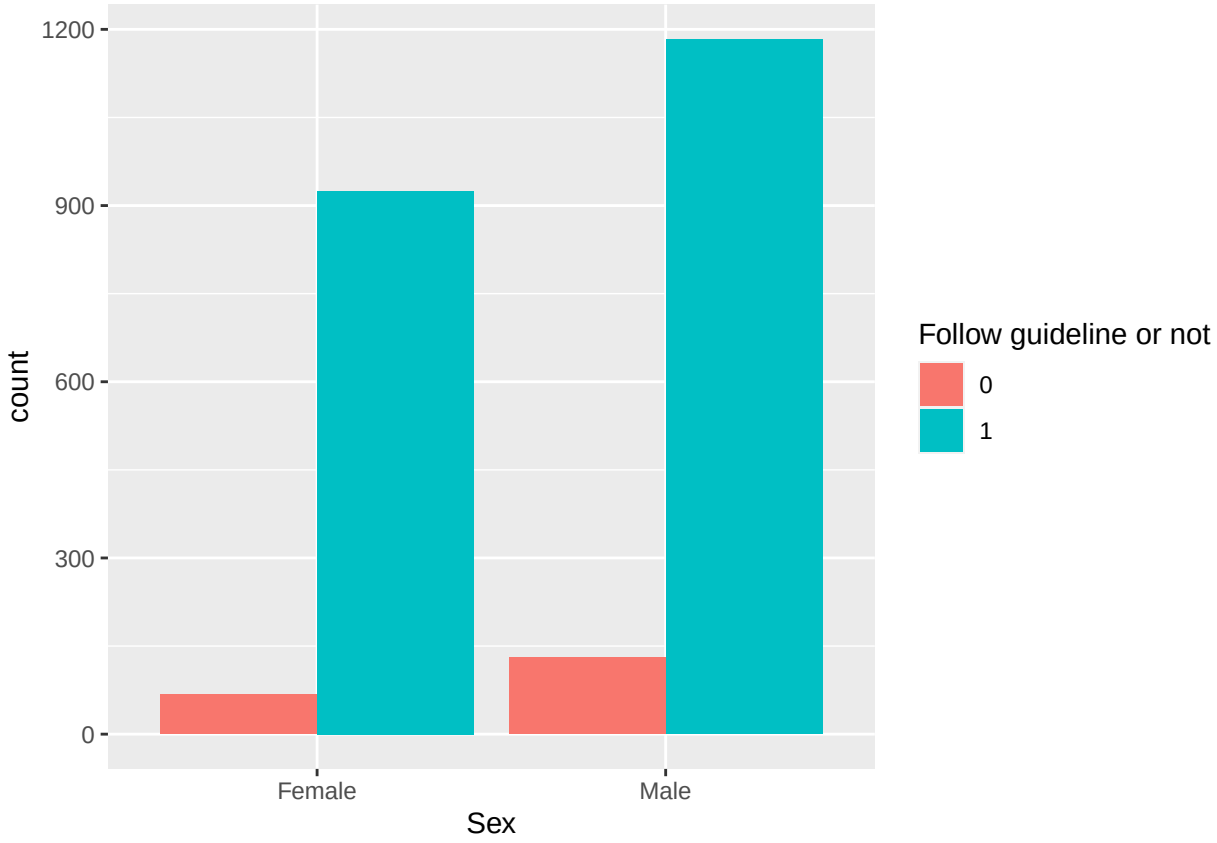
Besides this, it's notable that there are only 197 observations in class 0 while 2108 observations in class 1, so we also have extremely imbalanced classes in our outcome. (plots)

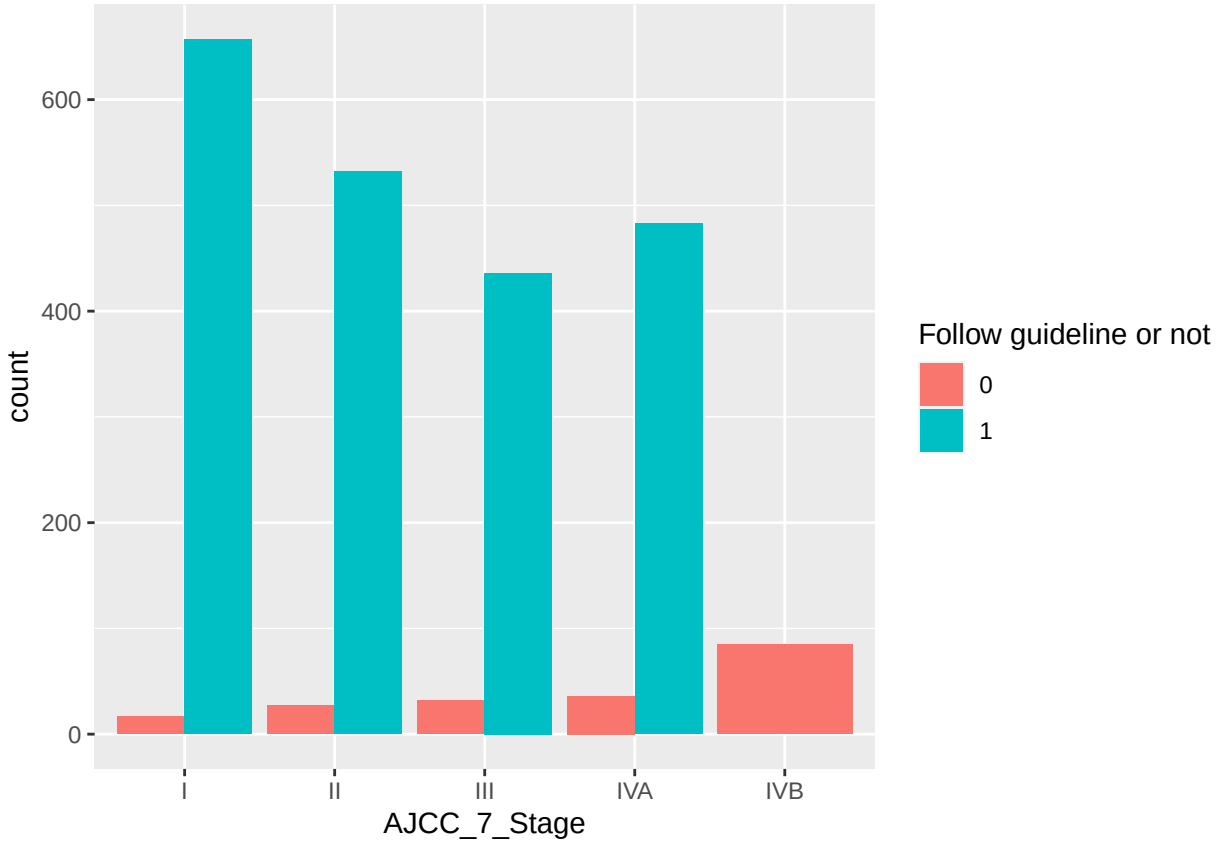
From the correlation matrix, we also noticed that, variables like education, including 9th grade, High school and bachelor are highly correlated. So we exclude them from modeling.

Other plots from our EDA are in the Appendix section.




Race
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
0
500
1000
count
Follow guideline or not
0
1


1200
900
600
300
0
count
Female
Male
Sex
Follow guideline or not
0
1



Modeling

We tried the logistic regression and multilevel logistic regression as our baselines, and then we moved to the classification tree, SVM, and random forest to see if they could return better results.

Since we have very imbalanced classes, we chose not to use the accuracy from the confusion matrix as a standard to compare the model performances because it is misleading in this case. For example, the model could classify half of the 0s wrong, but 99% of 1s correct, but since we have only 197 zeros but 2108 ones in the data frame, the accuracy from the confusion matrix would still be super high. However, we hope our model can perform equally well for each class. What we decided to use for comparing models are:

1. the confusion matrix result, how many 0s and 1s are misclassified
2. AUC

We used 0.5 as the threshold and classified a probability greater than 0.5 as class1 and others as class0. The AUC results from logistic regression and multilevel regression are 0.72 and 0.754. Random forest with stratified sampling returns an AUC of about 0.8. For all other models, the accuracy of predicting class1 is higher than 90%, but the accuracy of predicting class0 is only about 50% while that of the random forest model with stratified sampling is over 70%.

Based on those two criteria, we decided to use the random forest model for the following analysis.

Random Forest

We first fitted a normal randomForest model and tuned the hyperparameters mtry and ntree. We got maximum AUC at 0.725 for mtry = 7 and ntree = 600.

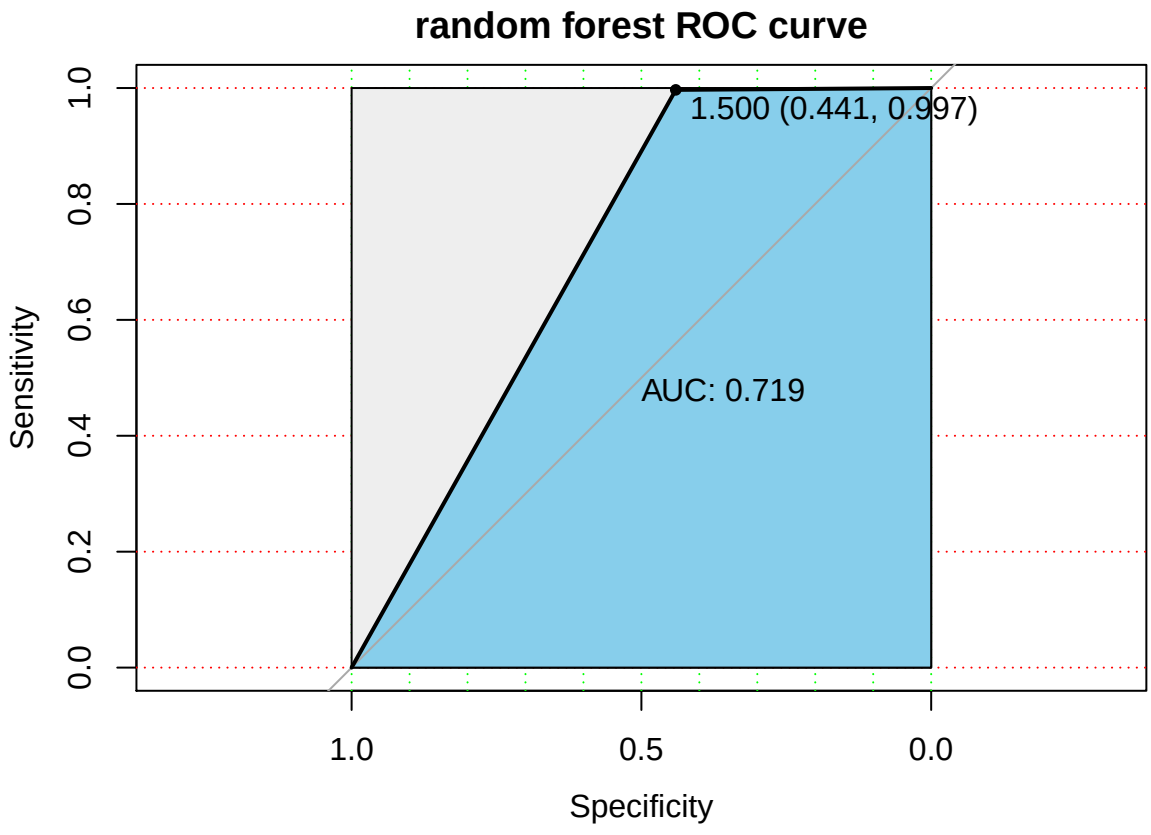
Next, to deal with the imbalanced classes problem, we tried the stratified sampling method in randomForest with different ratios of class 0s and class 1s. We tuned the partition of the sample size and used 2:3 (60 for class0 and 90 for class1) which returned the maximum AUC. We also tuned mtry and ntree, and the values that returned the largest AUC are mtry = 5 and ntree = 3000.

We drew the variable importance plot using the mean decrease Gini. In both models, AJCC_7_Stage, Age_at_Diagnosis, and Year_of_Diagnosis contribute most to whether the patient followed the guidelines. For the bias that we are considering about, it seems that race, gender, and insurance tend to have a small influence.

```
##
## pred_forest   0   1
##           0  26   2
##           1  33 630
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

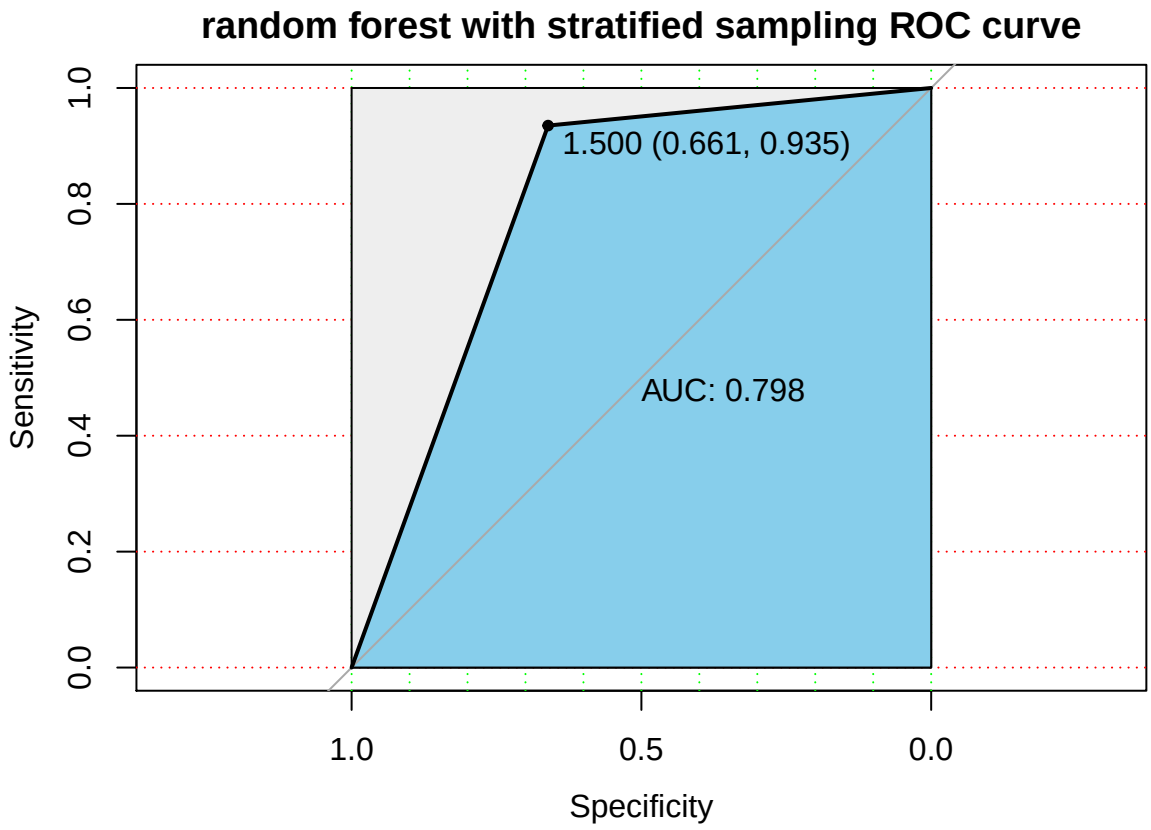


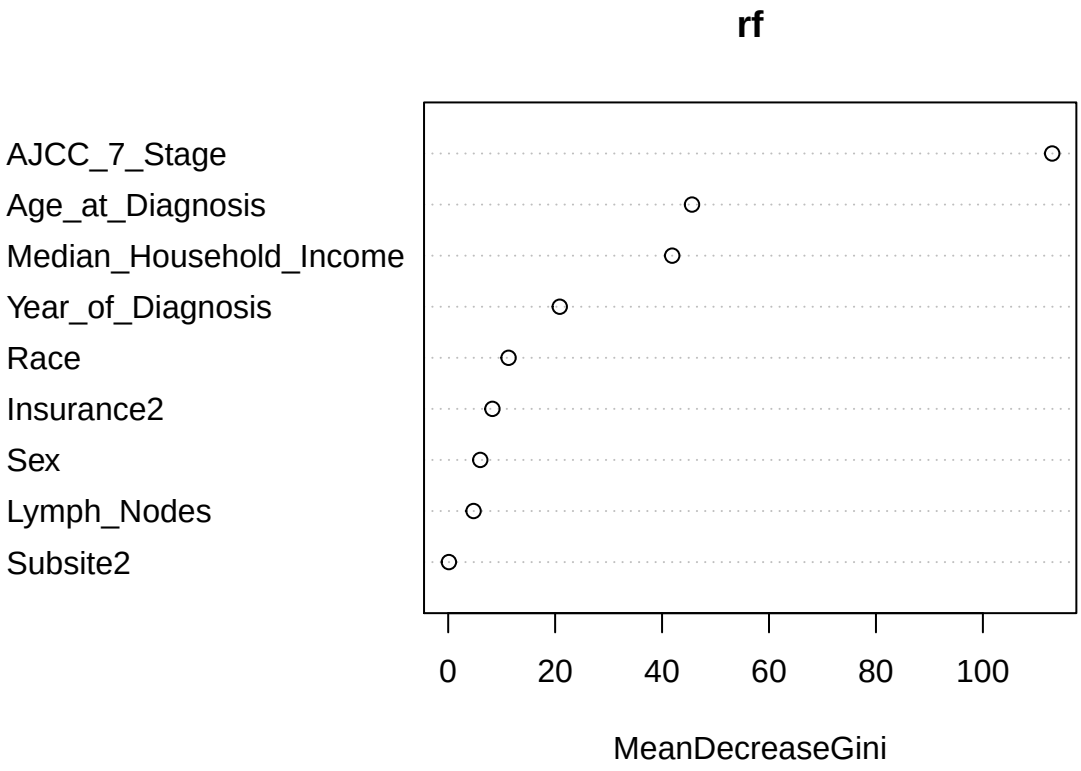

```
##
```

```
## pred_forest2   0   1
##             0  39  41
##             1  20 591
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```





Result

To see if there is bias, we compared two models where one of them contains Race and Gender as predictors and the other does not. If we can find an obvious difference in the prediction, then there might be some bias. We still used the confusion matrix result and AUC as standards to compare the model results.

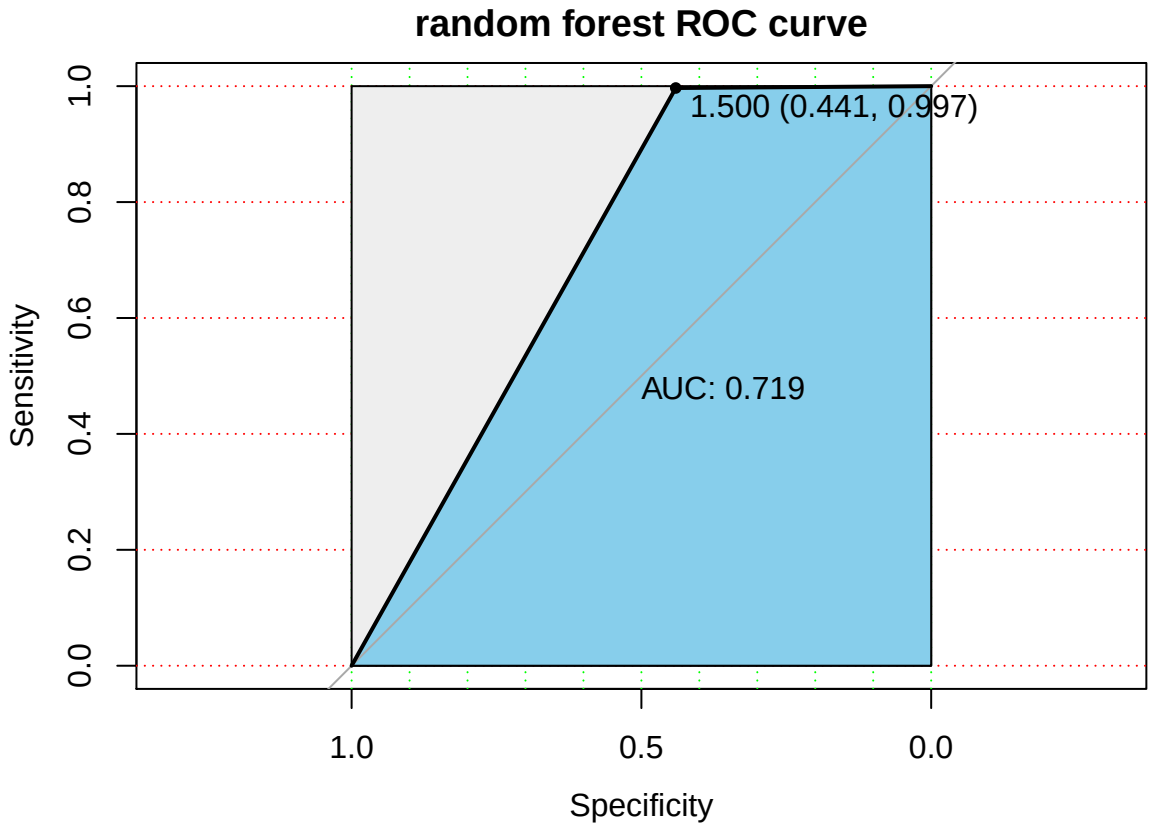
For the random forest model without stratified sampling method, we found the model without Race and Gender as predictors had a slightly worse prediction, to be specific, this exclusion resulted in the model misclassifying about 4 more observations in class1 to class0, and this causes the AUC decreased around 0.5%.

For the random forest model with the stratified sampling method, the result is very similar. The model without Race and Gender as predictors misclassified 9 more observations in class1 to class0 compared with the model with Race and Gender, and its AUC is about 2% lower.

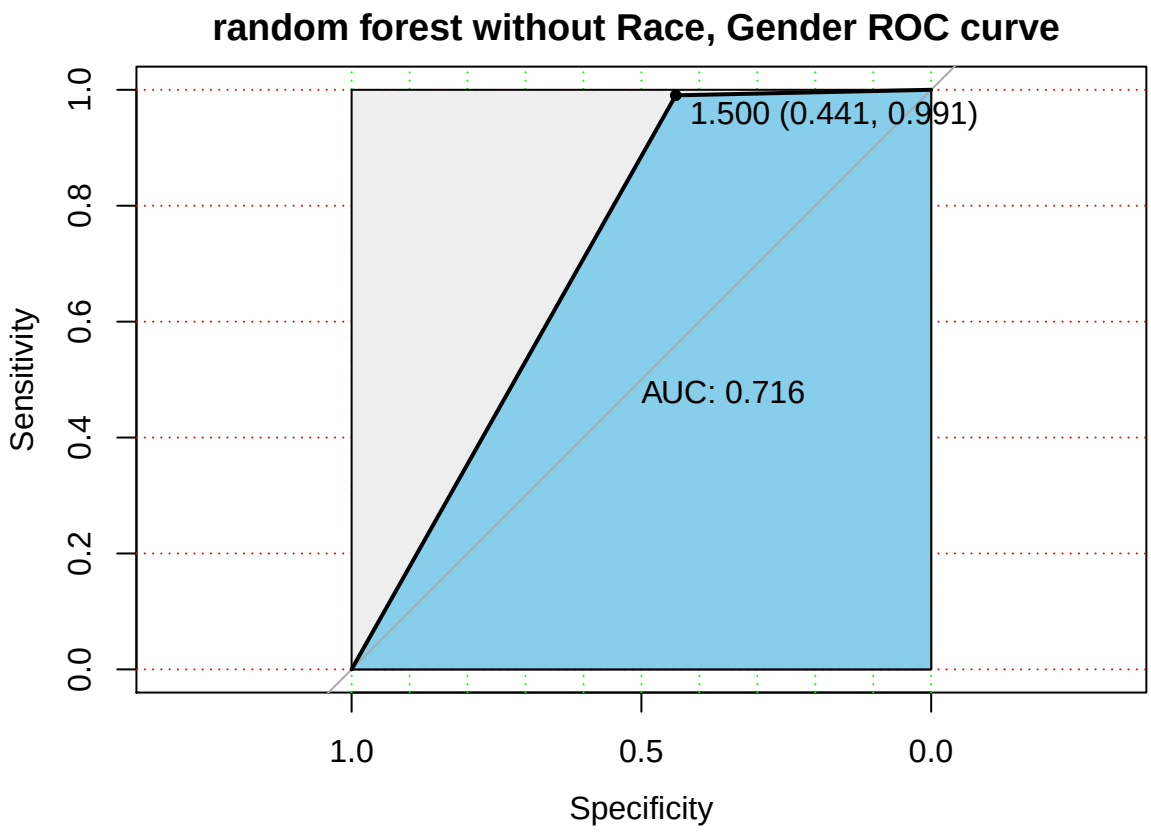
From what we got from the results, since there was no obvious difference in the prediction, we cannot conclude the existence of bias in either race or gender.

```
##
## Attaching package: 'Metrics'

## The following object is masked from 'package:pROC':
##
##     auc

## The following objects are masked from 'package:caret':
```

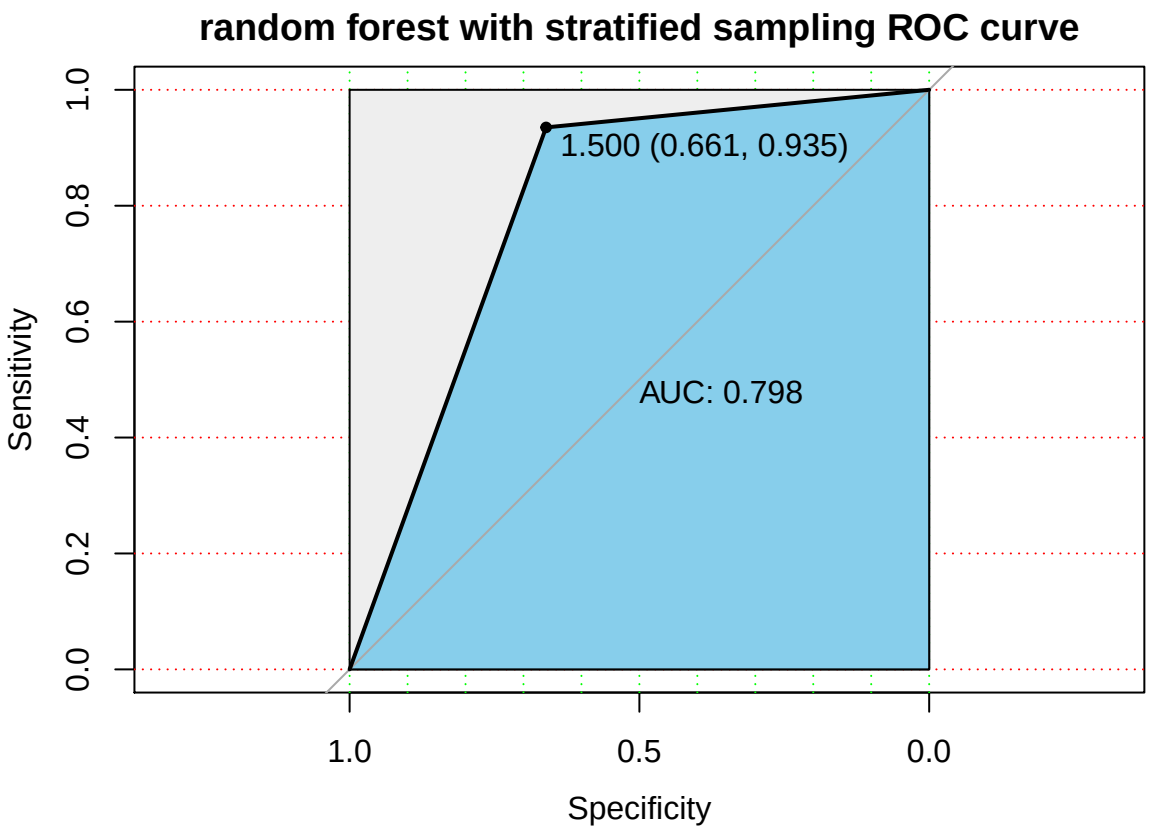
```
##
##     precision, recall

##
## pred_forest   0   1
##           0  26   2
##           1  33 630

##
## pred_forest1_2   0   1
##               0  26   6
##               1  33 626

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



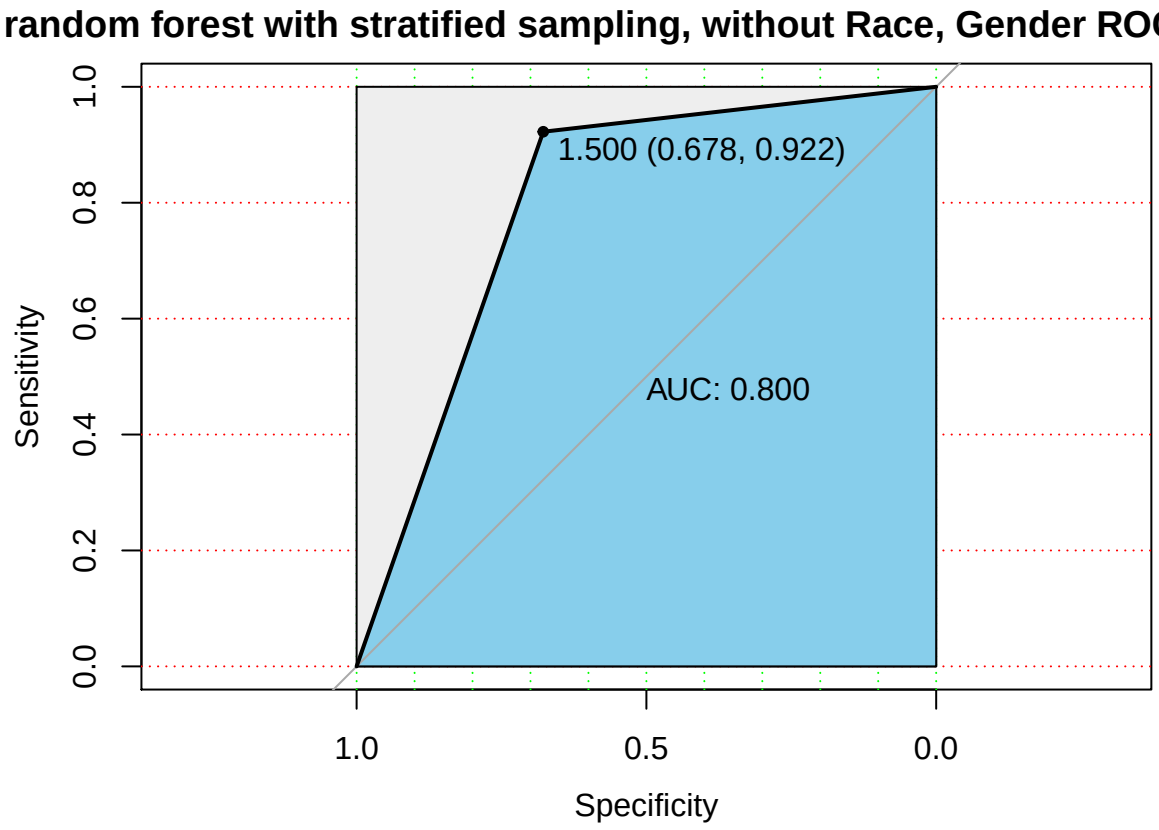

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
##
## pred_forest2   0   1
##            0  39  41
##            1  20 591

##
## pred_forest2_2   0   1
##              0  40  49
##              1  19 583

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = test$follow, predictor = as.numeric(pred_forest2))
##
## Data: as.numeric(pred_forest2) in 59 controls (test$follow 0) < 632 cases (test$follow 1).
## Area under the curve: 0.7981
```

```
##
## Call:
## roc.default(response = test$follow, predictor = as.numeric(pred_forest2_2))
##
## Data: as.numeric(pred_forest2_2) in 59 controls (test$follow 0) < 632 cases (test$follow 1).
## Area under the curve: 0.8002
```

Discussion

Because we do not have information like TNM stage, so we cannot match treatment guidelines with all head and neck cancer. Therefore, we only choose one site here, and the conclusions we made are all based on that certain site. Once we get to the TNM stage, we can do that on all the data, and it will give us a broader view of bias.

We do our best to fit a model performing well on both class 1 and class 0, as you can see the model is not perfect due to the limitation of data imbalance. This is a question that often can be seen in machine learning since in the real world those imbalance things always happen. Our model here is only a reference to give a

sense of what and how this bias will happen, and due to the accuracy of the model, we cannot assert that it must or it must not have bias.

Appendix