679 Final Exam

Group 10 - Zhaosheng Xie, Jinzhe Zhang,
Haoqi Wang, Zhitian Liu

Abstract

The Surveillance, Epidemiology, and End Results (SEER) Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS).The SEER research data include SEER incidence and population data associated by age, sex, race, year of diagnosis, and geographic areas (including SEER registry and county). For this project, we focus on analyzing the head and neck cancers data set.

Background

After we looked through the related materials and had some explorations about the data, we decided to focus on finding out what kinds of factors that affect the doctor's diagnosis when they give the patient surgery suggestions. In other words, what we are interested in doing in this project is figuring out what kind of patients are recommended to do the surgery.

In this project, after data processing and EDA, we built the Logistic regression model, multilevel logistic regression model and random forest model to explore the relationship among variables from the transformed SEER data. What's more, the

classifier model can also be used to forecast whether a doctor will recommend a patient for surgery.

Data Processing

After reading the transformed SEER data set in R, we can see that there are 25 variables and 23291 observations in our data set. There's no missing values in the data.

```
## [1] 0

## [1] "Surgery performed"

## [2] "Recommended, unknown if performed"

## [3] "Not recommended"

## [4] "Recommended but not performed, patient refused"

## [5] "Recommended but not performed, unknown reason"

## [6] "Not performed, patient died prior to recommended surgery"

## [7] "Not recommended, contraindicated due to other cond; autopsy onl
y (1973-2002)"
## [8] "Unknown; death certificate; or autopsy only (2003+)"
```

Surgery decision	recommend
Surgery performed	1 (recommend)
Recommended, unknown if performed	
Recommended but not performed, patient refused	
Recommended but not performed, unknown reason	
Not performed, patient died prior to recommended surgery	0 (not recommend or unknown)
Not recommended	
Not recommended, contraindicated due to other cond; autopsy only (1973-2002)	
Unknown; death certificate; or autopsy only (2003+)	

Based on the transformed data from SEER, to start our analysis, we first create a

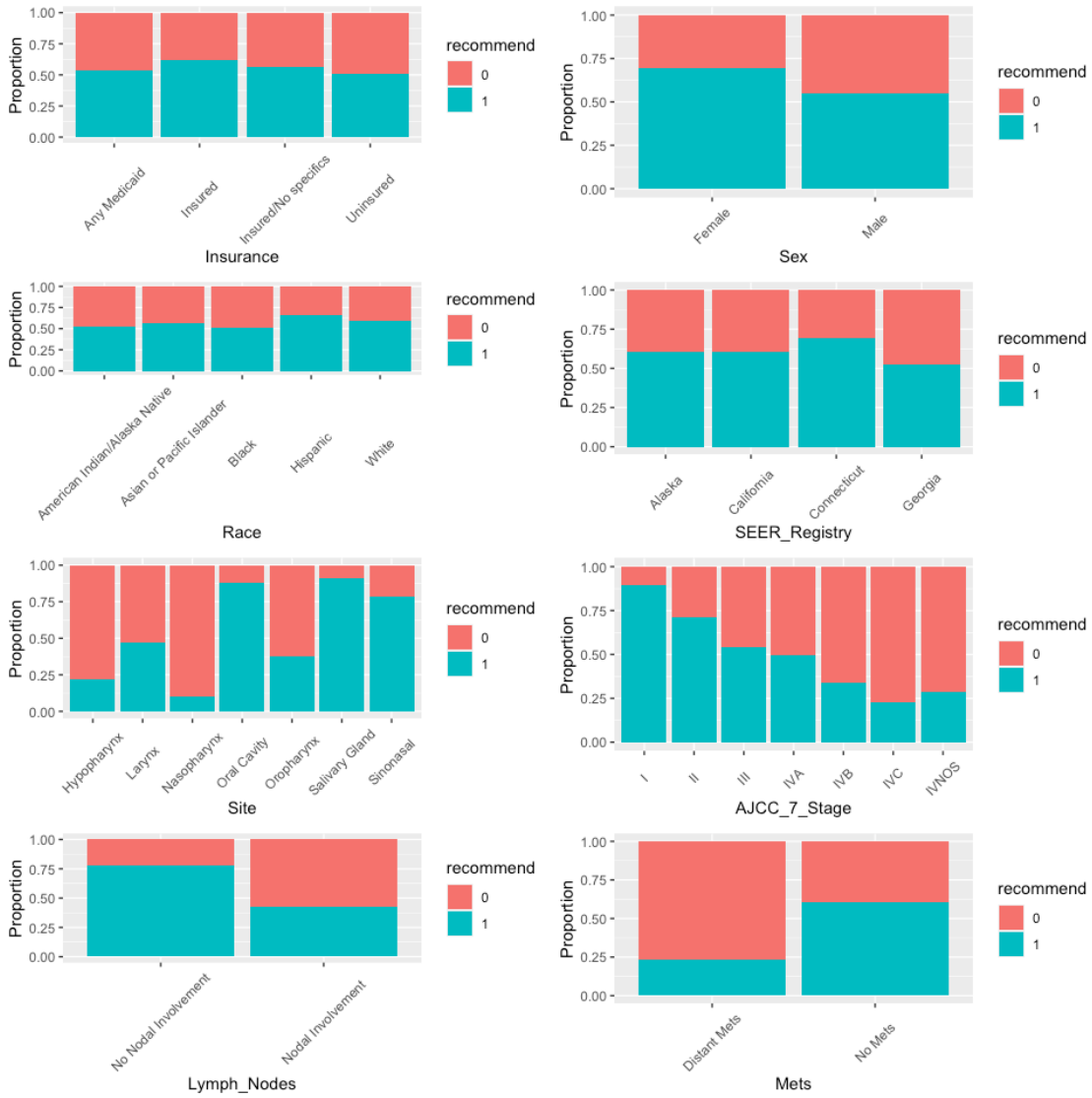
response variable that indicates whether or not the patient is recommended to perform the surgery according to the doctors. So we create the variable- “recommend” by extracting the information from “surgery decision” column.

The goal of our analysis in this project is to figure out what kind of patients are recommended to do the surgery. So what we need is only the demographic information of the patients and the information about the cancer they have. We don’t need to know （don’t want to know） whether the patient has died and whether they have undergone chemotherapy or radiotherapy. Hence we have to exclude the “Cause of death”, “Chemotherapy”, “Radiation”, “Surgery_Decision”, “Surgery_Performed” from the data set. The “Study ID” also does not need to include in our dataset.

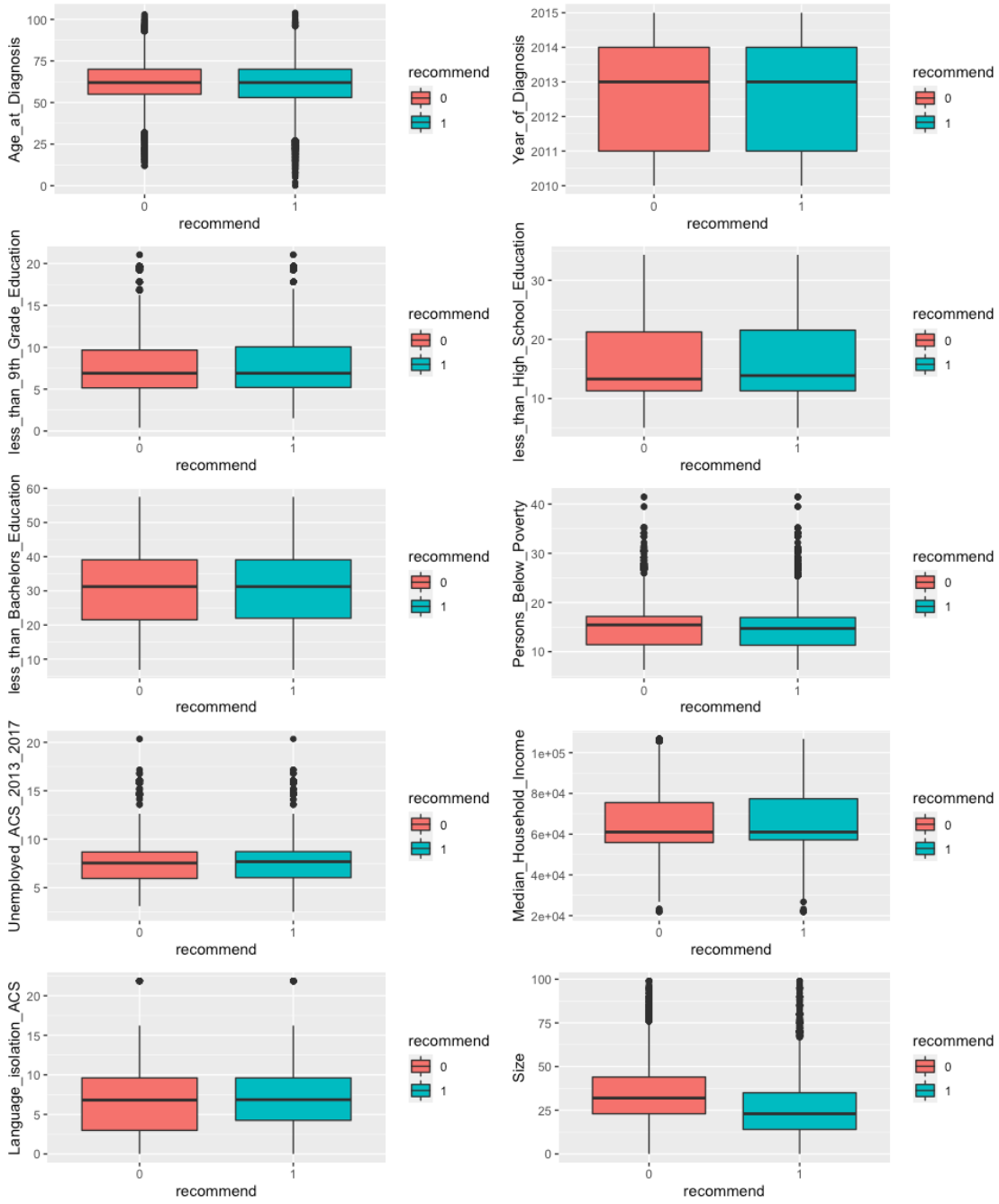
After excluding the variables we don’t need, we have the data set with below variables and ready to perform some EDA.

EDA

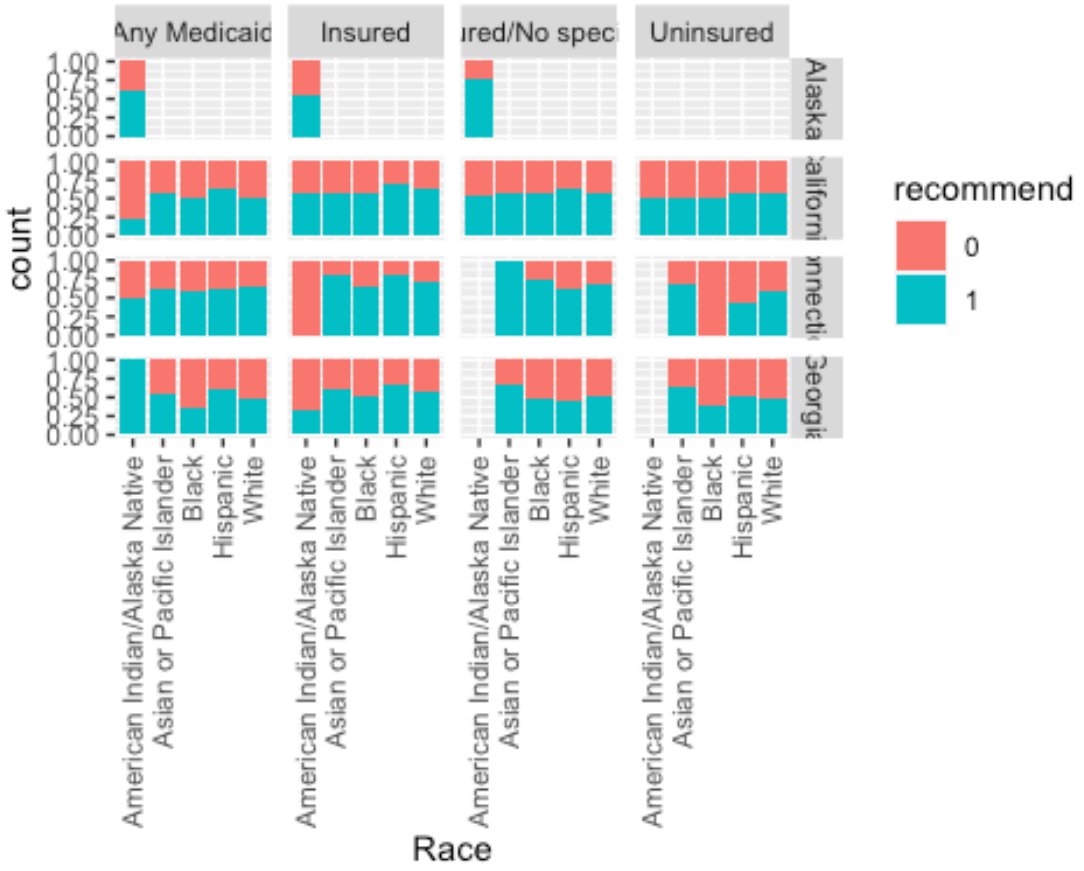
In the EDA section, we firstly used the ggplot2 package to generate some bar plots and box plots to check whether each variable is related to the doctor’s recommendation. Then we performed more advanced plots and visualize the correlation between some features to see if there are highly correlated variables.



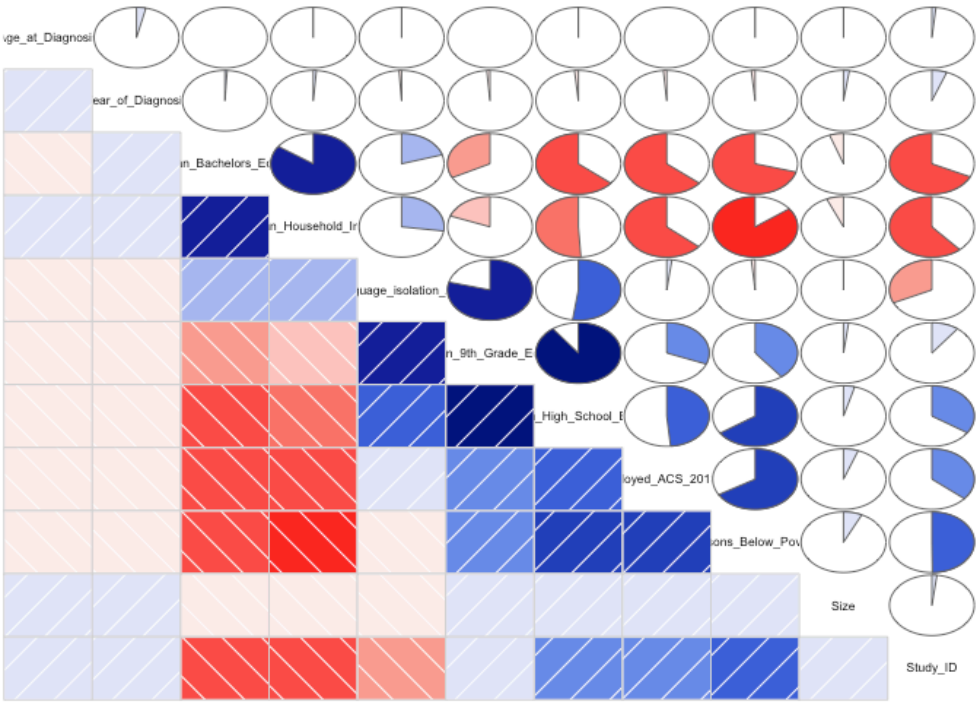
Firstly, we want to find the relationship between each categorical variable and recommendation. From this figure, we can see 'Sex','Site','AJCC.7.Stage','lymph.nodes' and 'mets' have the significant impact on the recommendation. But 'Race' 'Insurance','SEER.Registry' have the slight impact on whether doctor recommend the patient to do the surgery.



Similarly, we also tried to find the relationship between the continuous variable and recommendation. As the figure shows, ‘size’,‘language_isolation’ has bigger influence on the recommendation.



Then we combined three categorical variables randomly, we tried to find the interaction by this method when we did the model. As the figure shows, the proportion of the recommend of the person who are white in different 'REER_Registry' vary. So we consider this in our model when we need to add the interactions into the model.



When we have done the eda part, we found that some relationships among some variables. For example, the boxplots show that the distribution of the recommendation for the “less_than_High_School_Education” and “less_than_Bachelors_Education” are generally same. And then we think that we can combine the highly correlated colunms.

Model

In the modeling part, we use logistic regression and random forest to explore the important features for doctors to decide whether or not to recommend the patients to perform a surgery.

Logistic Regression

hypothesis test

Logistic regression model is a very popular modeling method when the outcome is binary. Before we fit a logistic regression model, we firstly apply hypothesis testing between different features versus“recommend”.

For the continuous variables, we performed two sample t-tests to determine whether the relationship between these variables and “recommend” is statistically

significant. For the categorical data. We performed the Chi-square test and Fisher exact test for the same purpose.

```
## # A tibble: 10 x 11
##    variables .y.   group1 group2    n1    n2 statistic     df
  p
##    <chr>     <chr> <chr>  <chr>  <int> <int>     <dbl>  <dbl>     <d
bl>
##  1 Age_at_D… value 0      1       9502 13789     6.98  22272. 3.03e-
 12
##  2 Language… value 0      1       9502 13789    -6.04  20040. 1.59e-
  9
##  3 less_tha… value 0      1       9502 13789    -3.80  20486. 1.43e-
  4
##  4 less_tha… value 0      1       9502 13789    -0.754 19524. 4.51e-
  1
##  5 less_tha… value 0      1       9502 13789    -1.12  20154. 2.63e-
  1
##  6 Median_H… value 0      1       9502 13789    -3.40  19769. 6.81e-
  4
##  7 Persons_… value 0      1       9502 13789     4.33  19833. 1.47e-
  5
##  8 Size      value 0      1       9502 13789    37.8   20341. 7.23e-
303
##  9 Unemploy… value 0      1       9502 13789    -0.654 19698. 5.13e-
  1
## 10 Year_of_… value 0      1       9502 13789     2.75  20477. 5.99e-
  3
## # … with 2 more variables: p.adj <dbl>, p.adj.signif <chr>

## # A tibble: 23 x 4
##    variables                              p     p.adj p.adj.signif
##    <chr>                              <dbl>     <dbl> <chr>
##  1 Age_at_Diagnosis                3.03e- 12 1.52e- 11 ****
##  2 Language_isolation_ACS          1.59e-  9 5.30e-  9 ****
##  3 less_than_9th_Grade_Education   1.43e-  4 2.86e-  4 ***
##  4 less_than_Bachelors_Education   4.51e-  1 5.01e-  1 ns
##  5 less_than_High_School_Education 2.63e-  1 3.29e-  1 ns
##  6 Median_Household_Income         6.81e-  4 1.14e-  3 **
##  7 Persons_Below_Poverty           1.47e-  5 3.67e-  5 ****
##  8 Size                            7.23e-303 7.23e-302 ****
##  9 Unemployed_ACS_2013_2017        5.13e-  1 5.13e-  1 ns
## 10 Year_of_Diagnosis               5.99e-  3 8.56e-  3 **
## # … with 13 more rows
```

We create a dataframe which includes the hypothesis test result for each variable, as we can see from the above output, only three variables give us no significant result. They are “Unemployed_ACS_2013_2017”, “less_than_High_School_Education”, “less_than_Bachelors_Education” . For these three variables, there doesn’t exist a

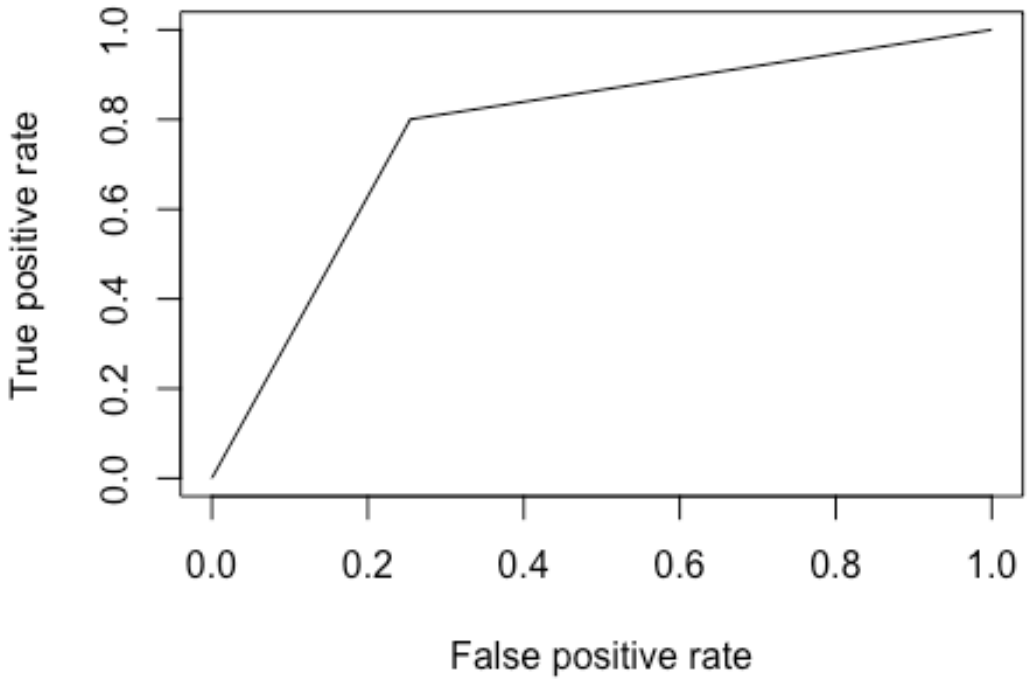
clear relationship with recommend intention. So we won't include these three features into the logistic regression for now.

Baseline Model

We firstly centralized the continuous variables in the data set. After randomly splitting the data into train and test sets, we fit a baseline logistic regression with all other variables left. From the summary output, we can see that there are several variables that give a low p-value indicating these variables are a meaningful addition to the model. It is to say, these features are more important to decide the outcome by logistic regression.

The AIC value for this basic logistic regression model is 10593. After this, we predict the test data set using the baseline model and get the accuracy 0.7785, also calculate the AUC value is 0.773.

```
## [1] 10593.07

##
## FALSE  TRUE
##  2579  9066

## [1] 0.7785316
```

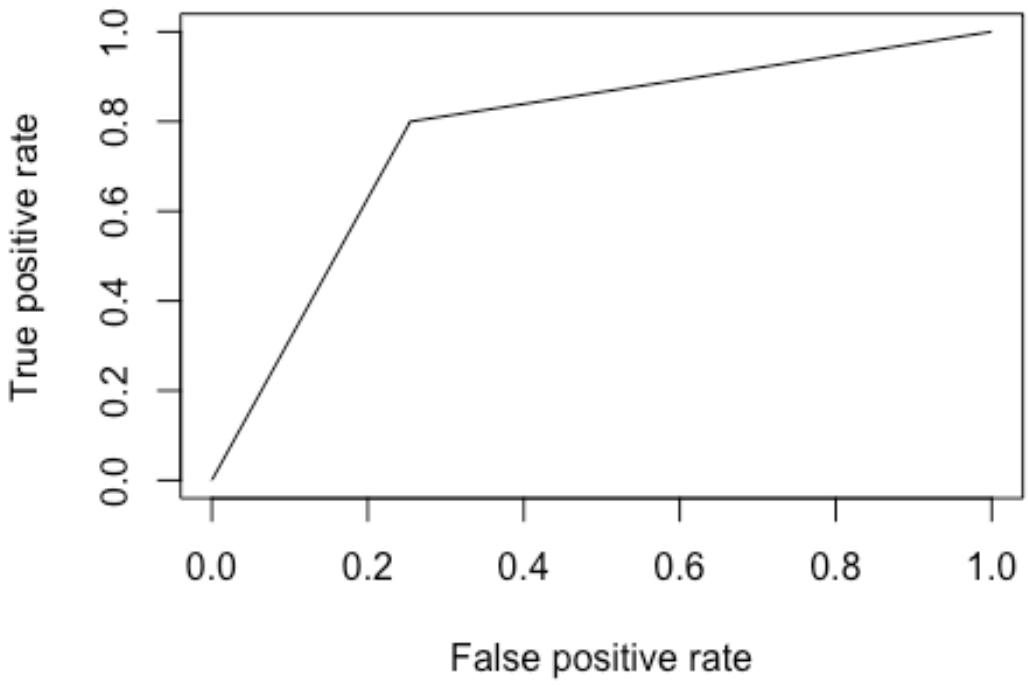


```
## [1] 0.7729084
```

Improved Logistic Regression Model

Now let's remove all insignificant independent variables from the prediction model, In this model we will remove "Sex" and "Mets" and "Race" , also the intercept. The AIC decreased to 10589 and the predicted accuracy and the AUC value remain steady.

```
## [1] 10589.2

##
## FALSE  TRUE
##  2582  9063

## [1] 0.7782739
```



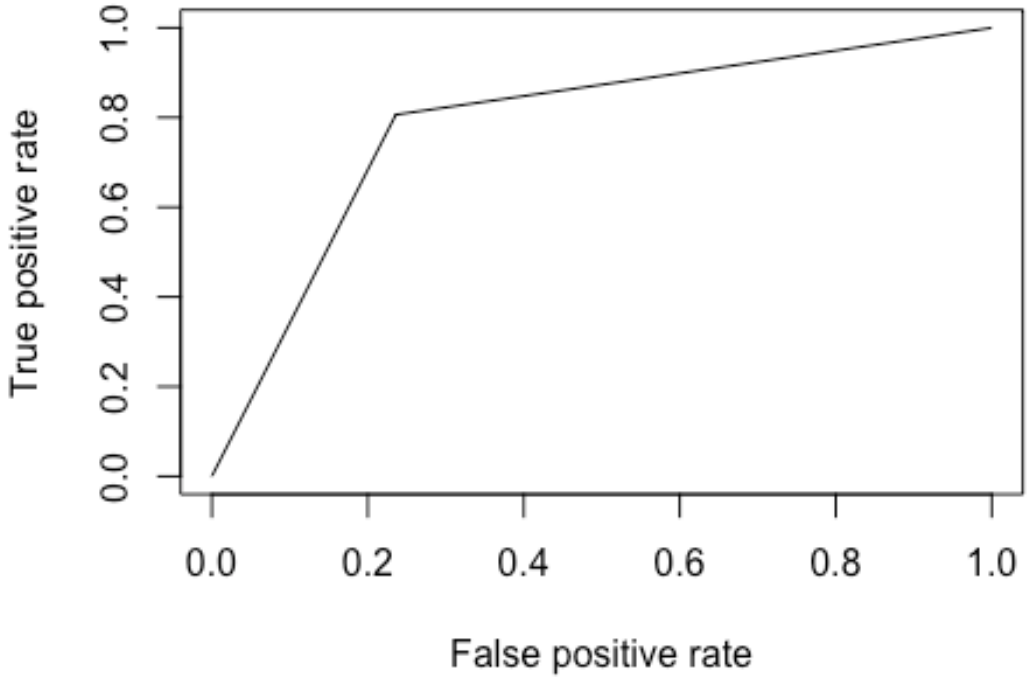

```
## [1] 0.772731
```

Final model: add interaction

From the EDA part,we notice some of the variables in the transformed data may be highly correlated. So based on our knowledge, The AJCC stage of a cancer patient is decided by it's tumor size (T stage) and the Lymph Nodes(N stage), the Site of the

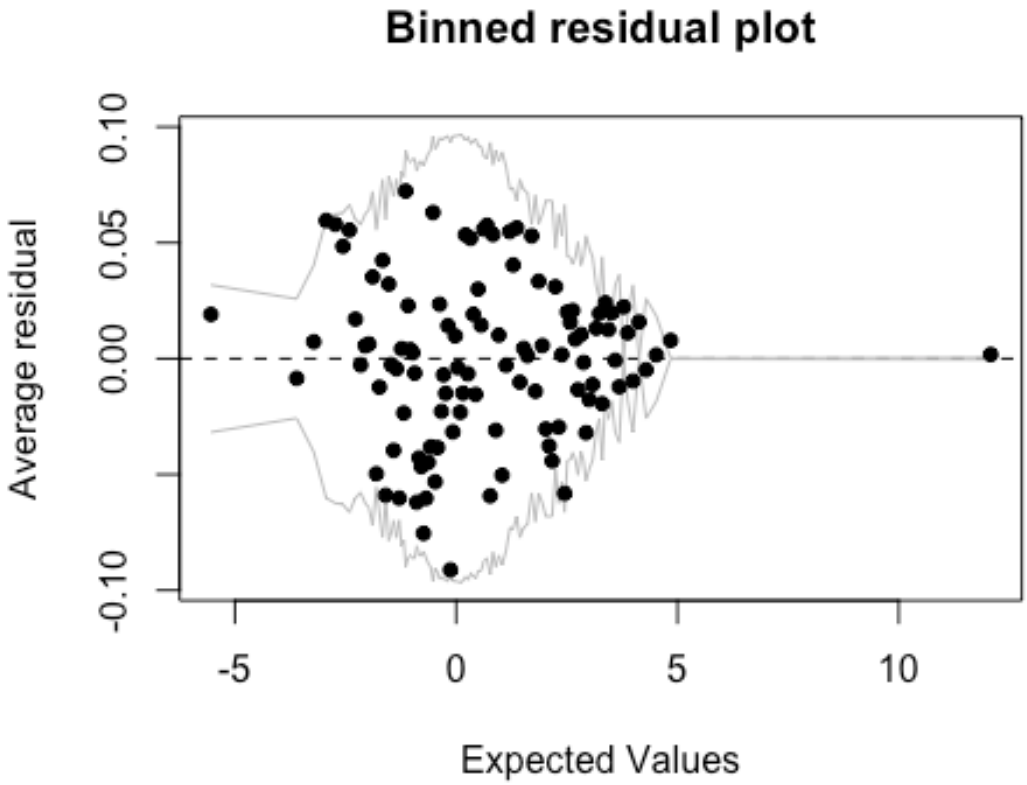
cancer he has, also the distant metastasis (M stage). So for these variables, we may need to add interaction terms in the model. Also, the poverty of the patient may have an interaction effect with his educational level, unemployment status, income and so on. We may need to also consider these into our model.

In our final model, we add these interaction terms and from the result, we can see that the model can produce an accuracy of around 80% and the AIC score improved to 10308, the AUC value increased to 0.785, which is an acceptable level for a regression model.



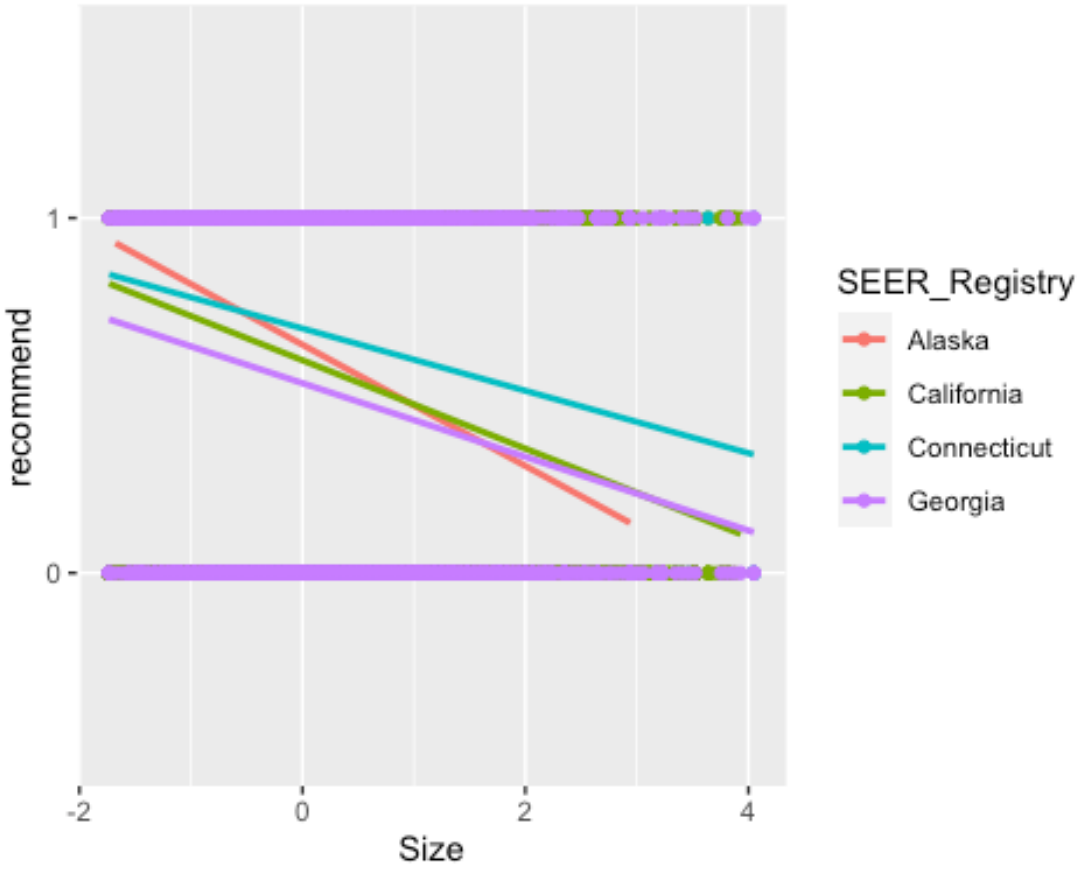

```
## [1] 0.7852195
```

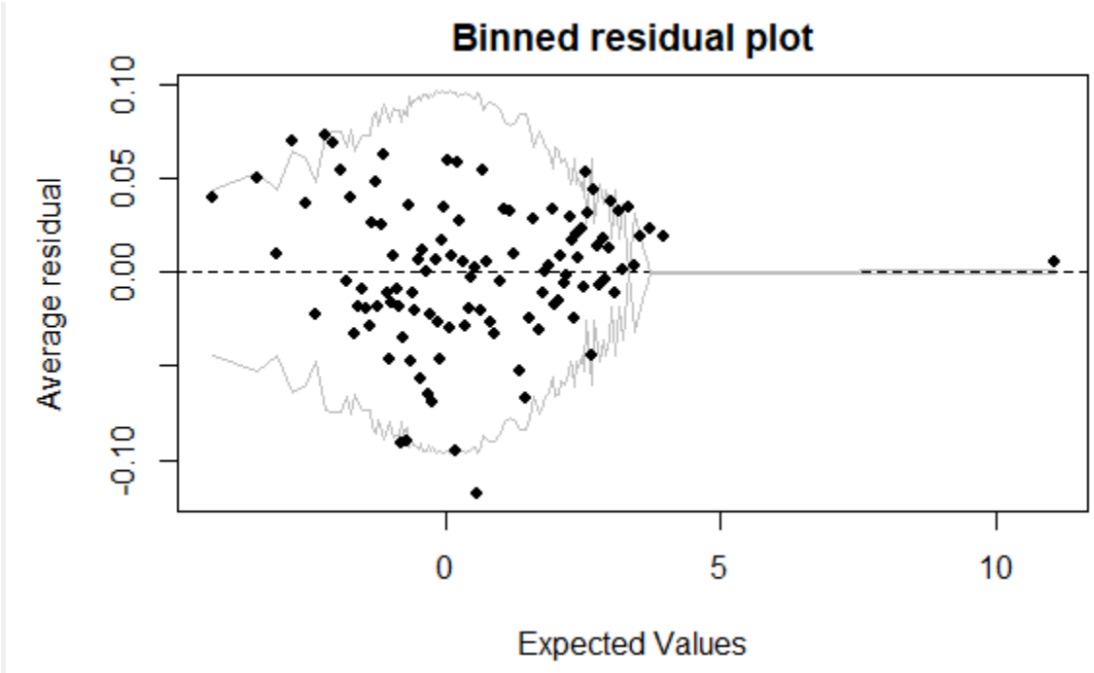
We also checked the binned residual plot for this model. The residual plot shows a fairly random pattern. This random pattern indicates that the model provides a decent fit to the data.



Multilevel Logistic Regression

After getting a logistic regression model, we think that there is still space to improve the model. Since the response variable is binary, and we want to add the random effect to make the model better, so we choose the multilevel logistic regression. We considered that different regions may have different diagnosis standards among the doctors. So “SEER_Registry” might be thought of as random effect. And we also draw the figure to double check our suspectation.





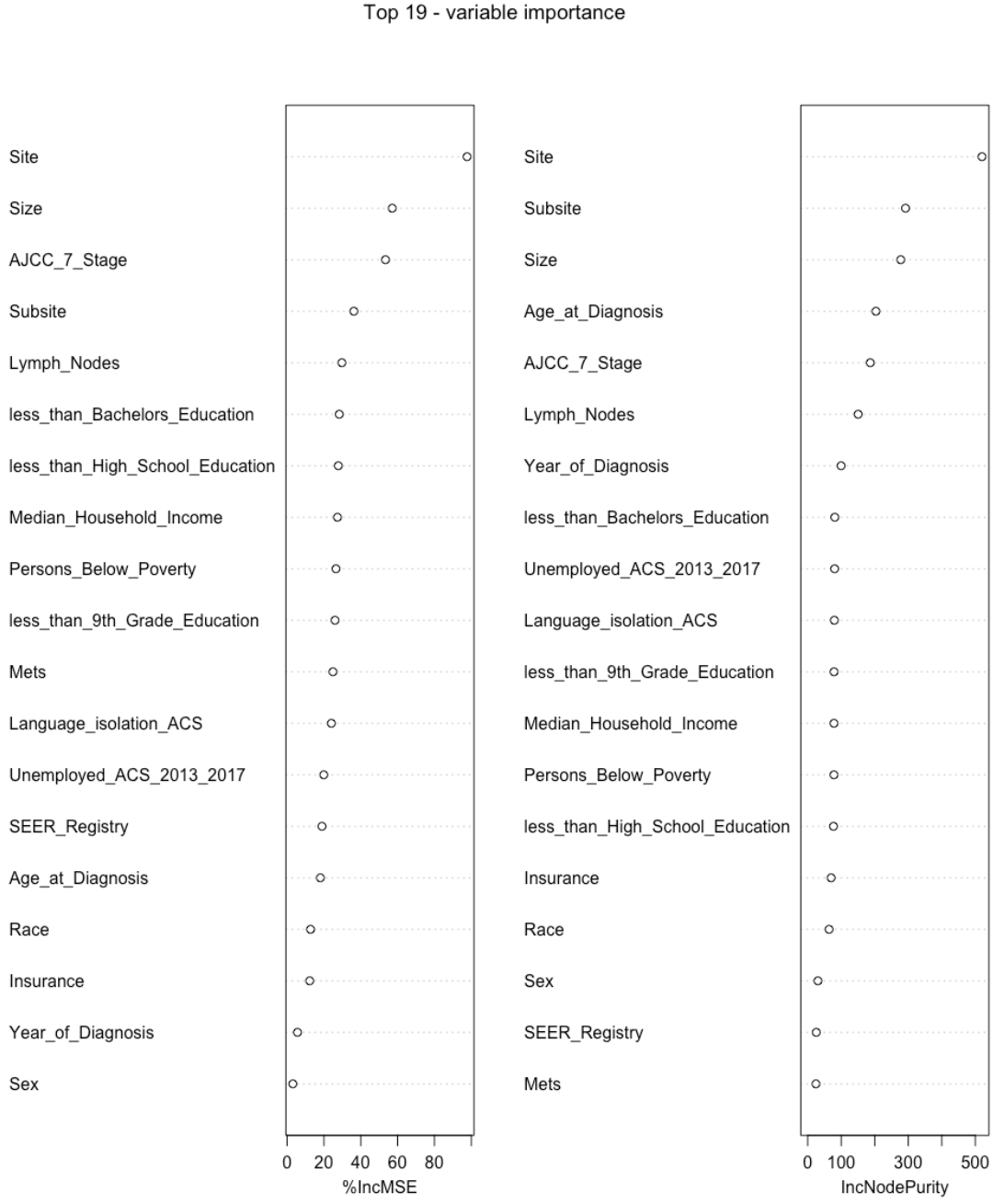
This model spends too much time to get the result, we move this to appendix. While the result of this model is much better than the logistic baseline model, so if we have more time, we can improve this later!

Random Forest

First we processed our data and chose some features that are actually associated with our response. The dataset was divided into 2 datasets: 50% training dataset and 50% testing dataset.

Then we built a 19-factors model using a training dataset as our baseline model. After this, we used the importance() function to get important features. The least important variable will be removed. Repeat this process. So the next model will be one factor less than the last model. Finally, we calculated accuracy and chose a 15-factor model which removes “Mets”, “SEER_Registry”, “Sex” and “Race” features. The accuracy of this 15-factor model is around 78.72%.

```
##                                    %IncMSE IncNodePurity
## Sex                               3.105416      30.38465
## Year_of_Diagnosis                 5.638290      99.49508
## Age_at_Diagnosis                 18.048628     203.42714
## Race                             12.713995      63.75763
## Insurance                        12.262365      69.75289
## SEER_Registry                    18.931404      25.47751
## less_than_9th_Grade_Education    26.013867      78.40259
## less_than_High_School_Education  27.799035      76.75917
## less_than_Bachelors_Education    28.311866      80.77016
## Persons_Below_Poverty            26.523549      78.22006
## Unemployed_ACS_2013_2017         19.900721      80.19872
## Median_Household_Income          27.337777      78.24645
## Language_isolation_ACS           24.035459      79.32309
## Site                             97.737714     519.93128
## Subsite                          36.244346     291.80823
## AJCC_7_Stage                     53.424174     186.62869
## Size                             57.097228     277.51790
## Lymph_Nodes                      29.738338     150.41953
## Mets                             24.882845      24.45553
```



There are two ways to evaluate the importance of features: Mean decrease accuracy and Mean decrease gini: Mean decrease accuracy describes the degree to which the prediction accuracy of random forest is reduced when a variable is changed into a random number. The greater the value, the greater the importance of the variable. Mean decrease gini calculated the heterogeneity influence of each variable on the observed values of each node in the classification tree through the Gini index. The higher the value, the greater the importance of the variable.

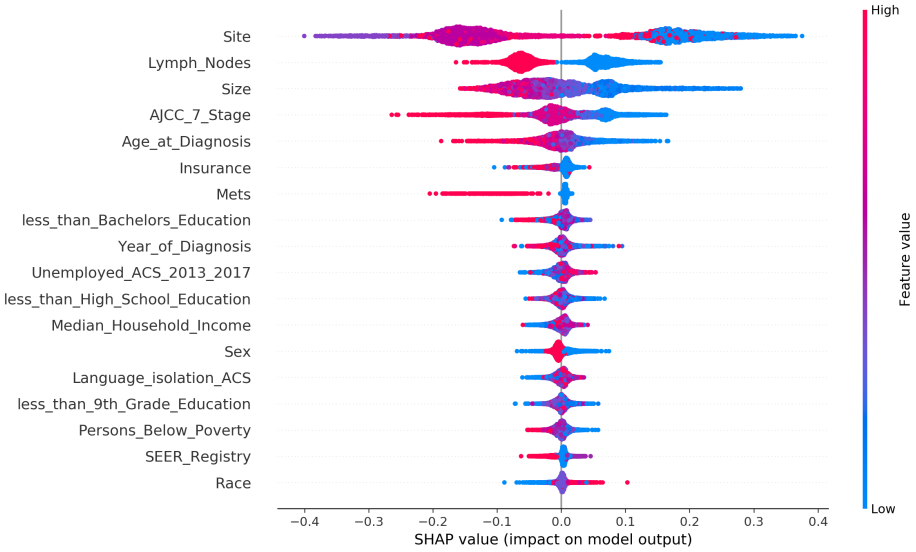
From the Mean decrease accuracy plot, we can see the top 5 most important variables are “Site”, “Size”, “AJCC_7_Stage”, “Subsite” and “Median_Household_Income”. The two least important characteristics are “Year_of_Diagnosis” and “Sex”. This also makes sense.
From the Mean decrease gini plot, we can see the top 5 most important variables are “Site”, “Subsite”, “Size”, “Age_at_Diagnosis” and “AJCC_7_Stage”. The two least important characteristics are “Mets” and “SEER_Registry”.

By comparing the results of these two plots we find they are generally similar, but there are some small differences. For example, in the left plot, “Year_of_Diagnosis” is the least important variable but in the right picture, this is the 7th important feature.

In this confusion matrix, column is real situation and row is prediction. The calculation method of accuracy is to divide the sum of the diagonals by the sum of all values.

SHAP, which stands for SHapley Additive exPlanations, is probably the state of the art in Machine Learning explainability. This algorithm was first published in 2017 by Lundberg and Lee (here is the original paper) and it is a brilliant way to reverse-engineer the output of any predictive algorithm.

In a nutshell, SHAP values are used whenever we have a complex model (could be a gradient boosting, a neural network, or anything that takes some features as input and produces some predictions as output) and we want to understand what decisions the model is making. For calculation, we use Python to calculate the marginal contribution of a feature when it is added to the model, and then take the mean value, namely the SHAP baseline value of a certain feature, into account the different marginal contributions of this feature in all feature sequences.



As shown above, our group used the Shap package in the Python environment to consider and include the effect of the interaction between each feature. The color of the feature tends to be red if the value of the feature is higher while it tends to be blue if the value of feature is lower. For example, when the value of Lymph_Nodes is “No Nodal Involvement”, it will have a negative effect on recommending surgery while the value is “Nodal Involvement” it will have the same but positive effect as “No Nodal Involvement”. However,both effects are not always the same among those factors, such as Mets and Size, the effect of high value in one side has much more influence on recommending surgery decisions than the other side. In other words, it means that the doctor does not tend to recommend patients to do surgery with distant Mets while cancer with no Mets will not impact much on recommending surgery decisions.

Conclusion

To conclude our analysis, both the logistic regression model and the random forest model can give us similar important factors which contribute to the “recommend” variable. To be specific, “the AJCC stage of cancer”, also “Site” of their cancer, the “Size” of the tumor, and the “lymph nodes”, these features related to the cancer situation are the most important factors the doctors need to consider to decide whether or not a patient needs surgery. Besides, the demographic information of the patients also matters, especially the age of diagnosis and the insurance status are very important features according to the models. What’s more, the features like educational level, the unemployment status and household income, poverty situation…..together, these variables also include some information for the doctors to take into consideration when trying to give surgery suggestions.

Discussion & Limitation

For the modeling part, logistic regression can give us quick and accurate predictions. We can filter the features during the modeling selection process, it's also easy to interpret. However, there is a lot of characteristic variables and interaction terms. The coefficient output seems to be too lengthy. It is difficult to find out the specific effect of each variable on the outcome. For the multilevel model, we just simply add some random effects based on the baseline logistic regression model, indeed it can give a better performance, however, fitting this model needs to spend a very long time, it's hard to do the model selection so we drop this idea for now.

Random Forests can be used for both classification and regression tasks. It also works well with both categorical and numerical data without re-scaling or transformation of variables. However, Random Forests are not easily interpretable. They provide feature importance but it does not provide complete visibility into the coefficients as linear regression. In this project, the categorical features in the Shap value of the tree model can only be converted into numerical ones, which we need to improve.

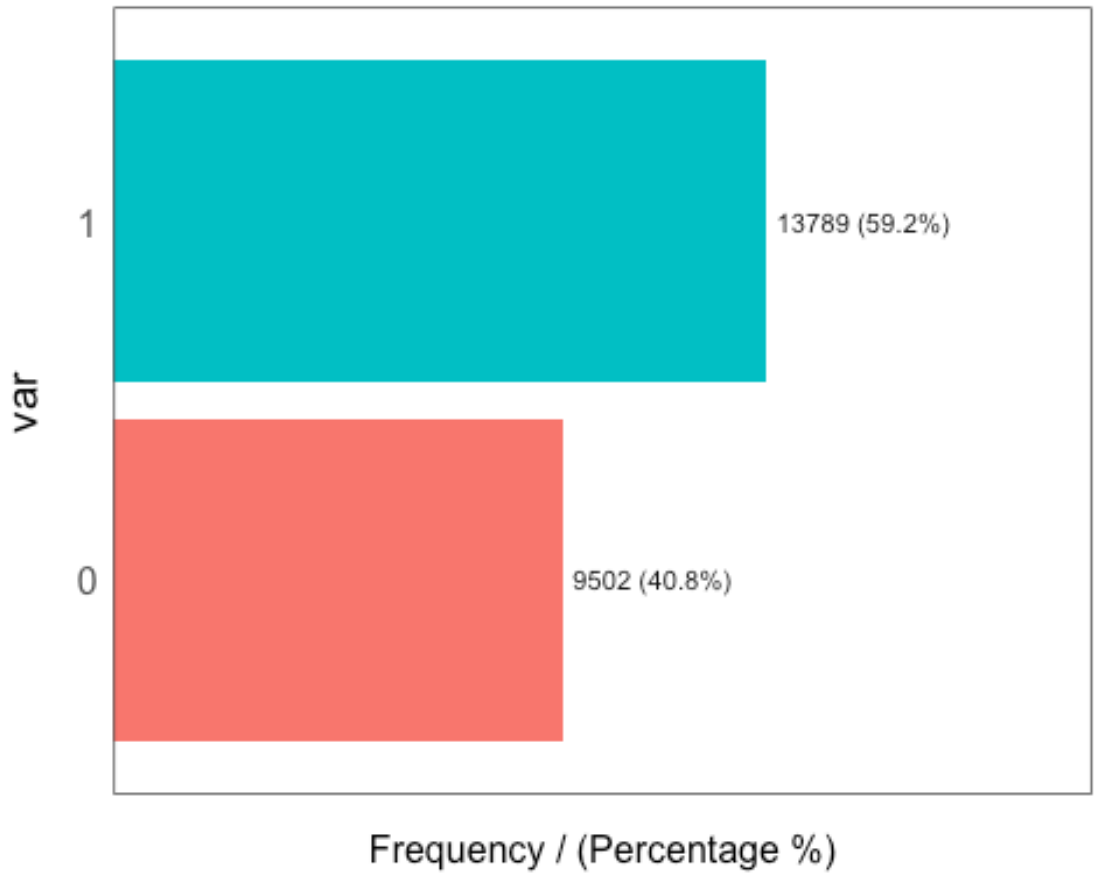
There are also some other limitations to the data we used. Firstly, all the patients in the model were diagnosed from 2010 to 2015, it's a little outdated, the doctor's diagnostic criteria after 2015 may be changed. Secondly, For some variables, there are some biases since some of the categories in specific variables (like race and SEER registry) have too few observations. These biases may influence our modeling results.

Our next step for this project will be to look into more details in adding random effects on the logistic regression model and also try other algorithms. In addition, we need to find a better explanation of the Shap value in the random forest model.

Appendix

The proportion of different

```
freq(logisdata$recommend)
```




```
##   var frequency percentage cumulative_perc
## 1   1     13789       59.2            59.2
## 2   0      9502       40.8           100.0
```

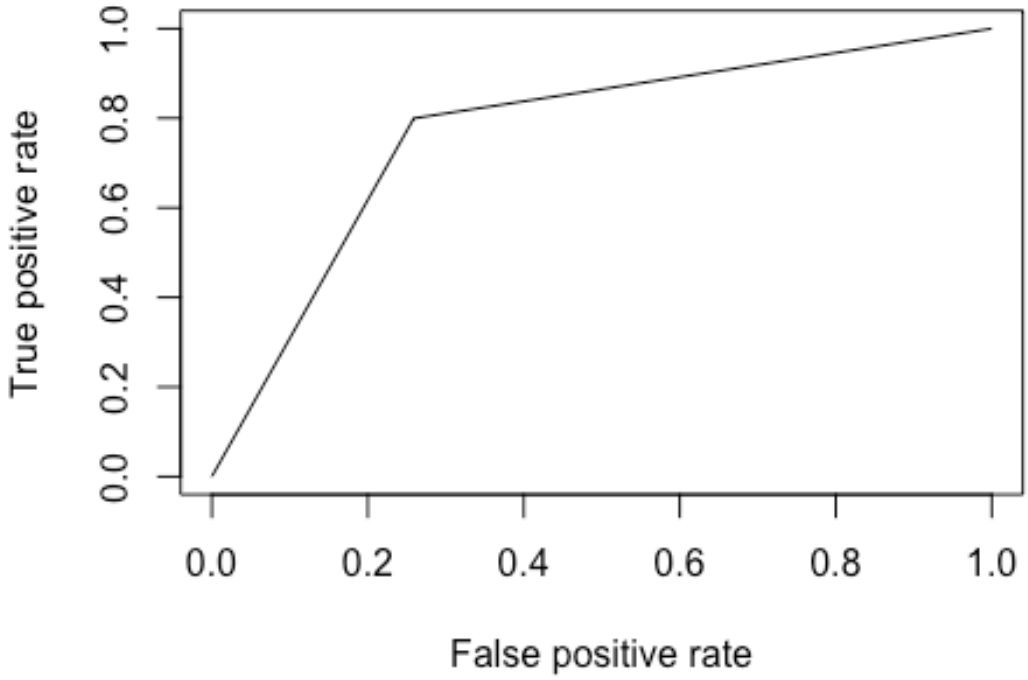
```
names(logisdata)[names(logisdata) == "% Language isolation ACS 2013-2017 (households)"] <- "Language_isolation"
  logisdata$Language_isolation <- ifelse(logisdata$Language_isolation<4,"Best",
                                         ifelse(logisdata$Language_isolation>=4&logisdata$Language_isolation<8,"Good",
                                                ifelse(logisdata$Language_isolation>=8,"Bad",3)))


logisdata$Persons_Below_Poverty <- ifelse(logisdata$Persons_Below_Poverty< 10,"rich",ifelse(logisdata$Persons_Below_Poverty>=10&logisdata$Persons_Below_Poverty<20,"average",
                                                ifelse(logisdata$Persons_Below_Poverty>=20&logisdata$Persons_Below_Poverty<30,"poor",ifelse(logisdata$Persons_Below_Poverty>=30&logisdata$Persons_Below_Poverty<50,"very poor",4))))
```

```
  sum(ifelse(logisdata$Persons_Below_Poverty=="very poor" & logisdata$R
ace=="Asian or Pacific Islander",1,0))

## [1] 0

  logisdata$Persons_Below_Poverty=as.factor(logisdata$Persons_Below_Pov
erty)
    ggplot(logisdata,aes(x=Insurance,..count..))+
    geom_bar(aes(fill=recommend),position="fill")+
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = .5))
+
    facet_grid(Mets~ logisdata$Lymph_Nodes)
```

Multilevel Logistics Regression Model

```
fit_multi<-glmer(factor(recommend)~Age_at_Diagnosis+factor(Mets)+factor
(Insurance)+factor(Sex)+
                 factor(Race)+factor(Site)+factor(Subsite)+less_than_Hig
h_School_Education+less_than_9th_Grade_Education+less_than_Bachelors_Ed
ucation+Unemployed_ACS_2013_2017+(1 + Size|SEER_Registry), data = logis
data.train,              family=binomial(),control=glmerControl(optimize
r = "Nelder_Mead",optCtrl=list(maxfun=100000)))

## fixed-effect model matrix is rank deficient so dropping 6 columns /
coefficients
```

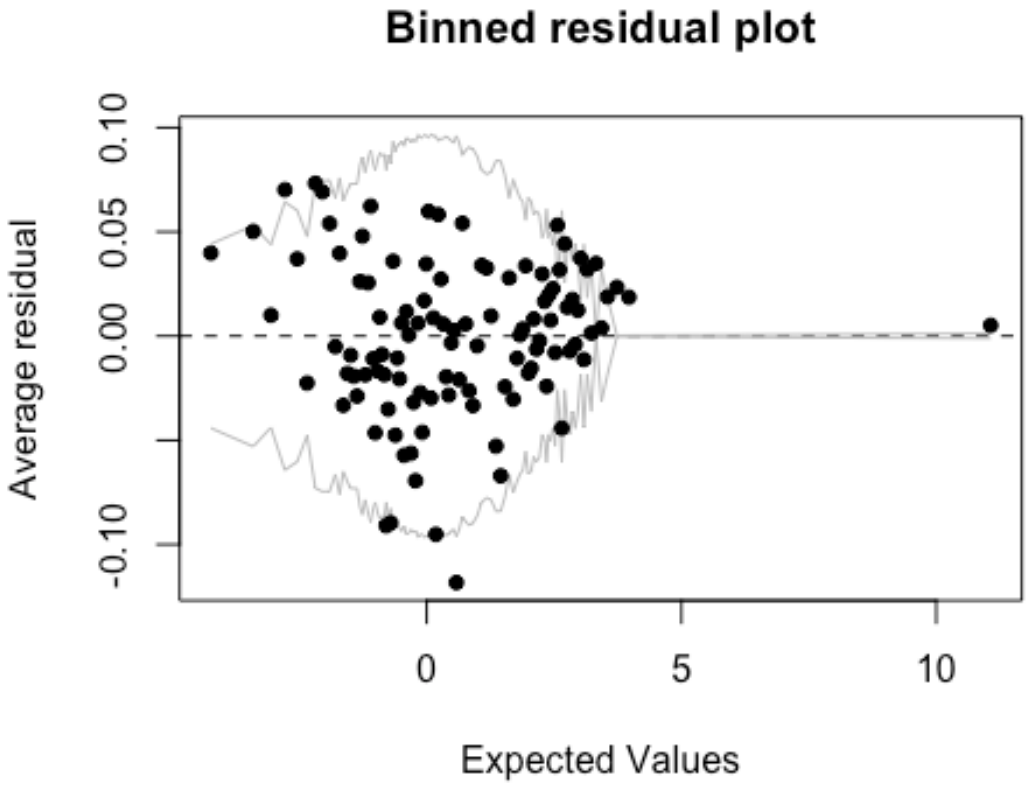
```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$ch
eckConv, :
## Model failed to converge with max|grad| = 0.014787 (tol = 0.002, com
ponent 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$ch
eckConv, : Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?

summary(fit_multi)


#AIC 10848
#BIC 11488
#ACC 0.783
logisdata.test$predicted.reco = predict(fit_multi, newdata=logisdata.te
st, type="response")
logisdata.test$predicted.reco=ifelse(logisdata.test$predicted.reco>0.5,
1,0)

pred <- prediction(logisdata.test$predicted.reco, logisdata.test$recomm
end)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf)
```

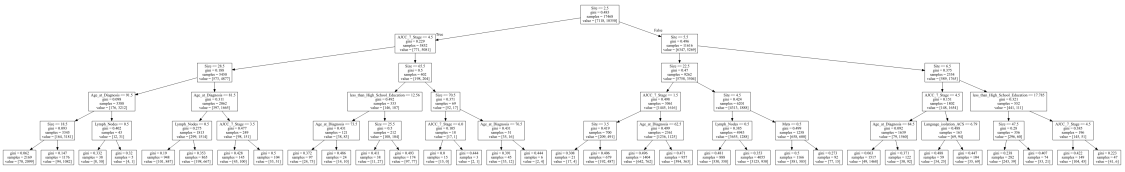


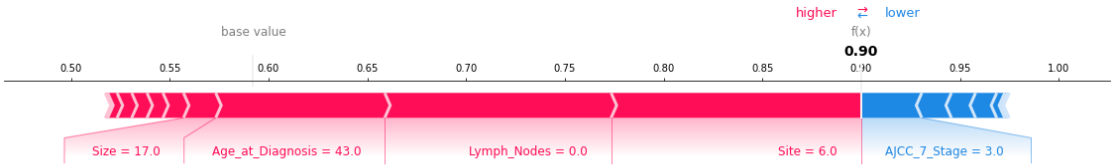

```
auc <- performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
#acceptable 0.785
auc
```

```
## [1] 0.7703225
```

```
binnedplot(predict(fit_multi), resid(fit_multi,type="response"))
```


Binned residual plot
Average residual
0.10
0.05
0.00
-0.10
0
5
10
Expected Values




higher
lower
base value
f(x)
0.90
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Size = 17.0
Age_at_Diagnosis = 43.0
Lymph_Nodes = 0.0
Site = 6.0
AJCC_7_Stage = 3.0