PSTAT 131 Project Final Report

Michael Kwok

5/31/2020

I. Introduction

This report will first explore the data in the Titanic dataset provided by the Department of Biostatistics of Vanderbilt University. The response variable is whether or not a passenger survived (1 for yes, 0 for no), and the covariates provide information about the Titanic passengers such as sex, age, and passenger class. The goal is to answer the research question: which covariates, if any, are most relevant in predicting whether or not a passenger would survive, and why?

II. Data

a) Contents

There are 1309 rows and 11 columns in the dataset. 38.197% of the passengers survived. Besides survival, we are given the following information about the passengers: passenger class, name, sex, age, number of siblings/spouse on board, number of parents/children on board, ticket number, fare amount, cabin number, and port of embarkation.

The numerical predictors are age, number of siblings/spouse, number of parents/children, and fare. The remaining predictors are categorical.

The following table gives the first 10 rows of the dataset.

```
## # A tibble: 10 x 11
##    Pclass Survived Name  Sex     Age SibSp Parch Ticket  Fare Cabin
##     <int>    <int> <fct> <fct> <dbl> <int> <int> <fct>  <dbl> <fct>
##  1      1        1 Alle~ fema~ 29        0     0 24160  211.  B5
##  2      1        1 Alli~ male   0.92     1     2 113781 152.  C22 ~
##  3      1        0 Alli~ fema~  2        1     2 113781 152.  C22 ~
##  4      1        0 Alli~ male  30        1     2 113781 152.  C22 ~
##  5      1        0 Alli~ fema~ 25        1     2 113781 152.  C22 ~
##  6      1        1 Ande~ male  48        0     0 19952   26.6 E12
##  7      1        1 Andr~ fema~ 63        1     0 13502   78.0 D7
##  8      1        0 Andr~ male  39        0     0 112050   0   A36
##  9      1        1 Appl~ fema~ 53        2     0 11769   51.5 C101
## 10      1        0 Arta~ male  71        0     0 PC 17~  49.5 <NA>
## # ... with 1 more variable: Embarked <fct>
```

b) Graphs

Below are visual representations of the predictors. Shaded in dark gray is the death rate.

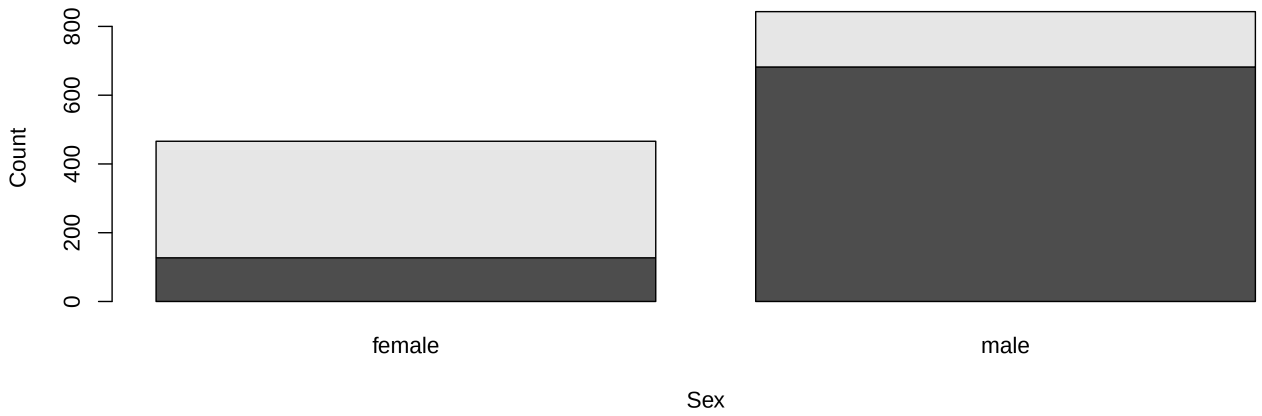


Figure 1: Frequency of Sex

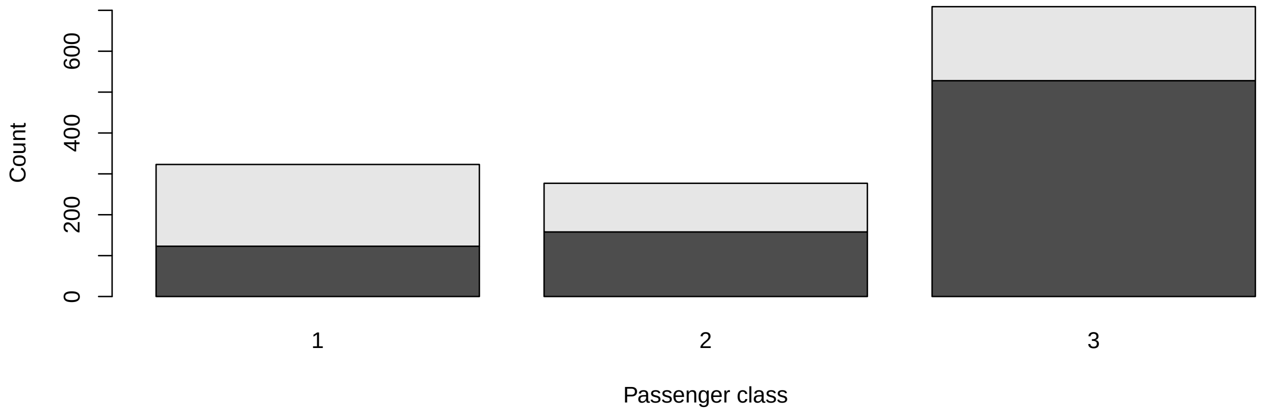


Figure 2: Frequency of Passenger Class

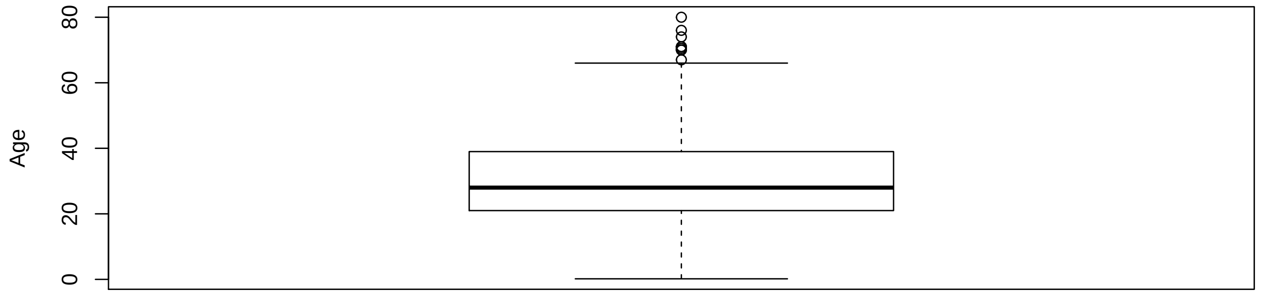


Figure 3: Distribution of Age

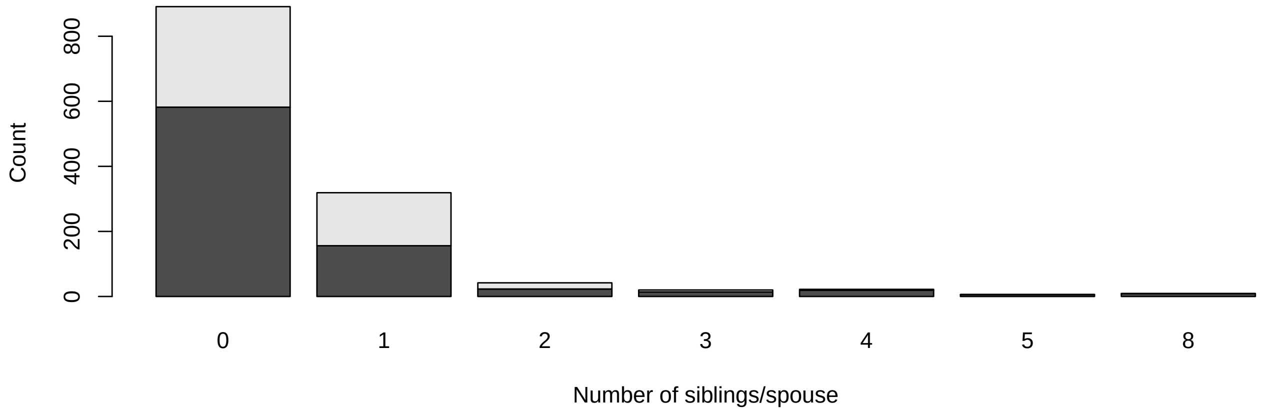


Figure 4: Frequency of Siblings/Spouse

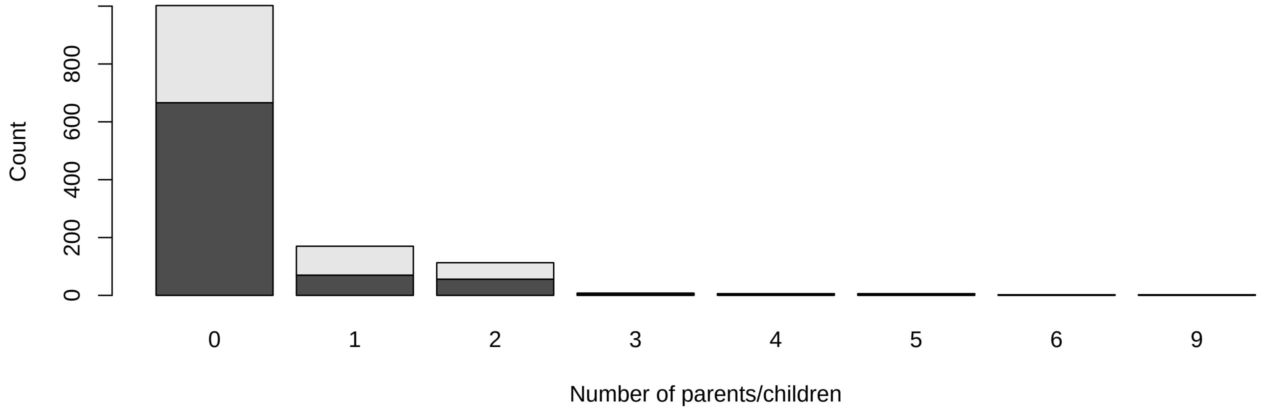


Figure 5: Frequency of Parents/Children

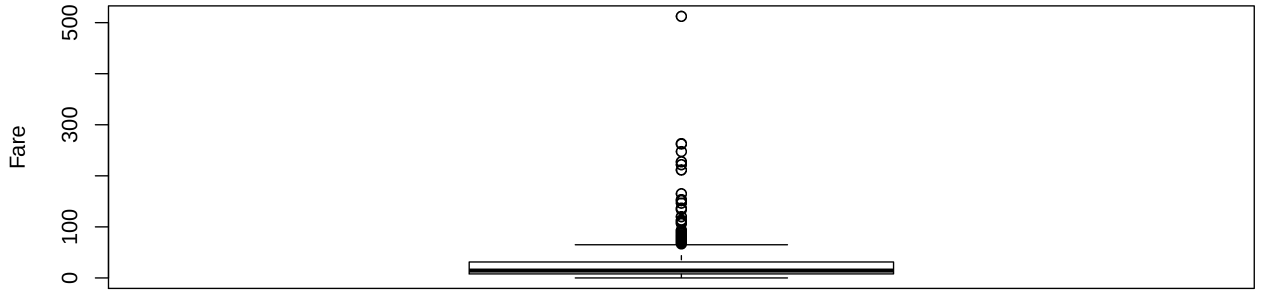


Figure 6: Distribution of Fare

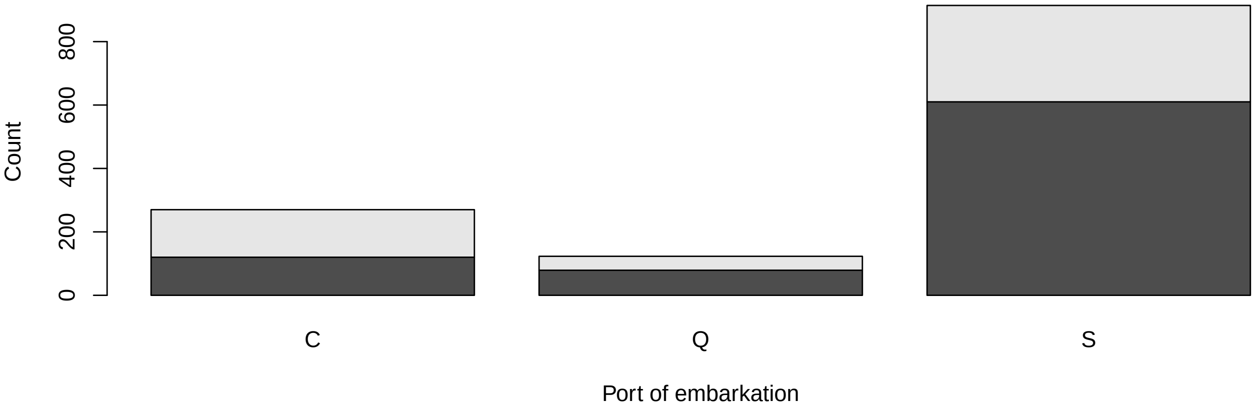


Figure 7: Frequency of Port of Embarkation

c) Graph interpretations

Figure 1 shows that the there were nearly twice as many males as there were females, and most of the females survived while most of the males perished. Figure 2 shows that the majority of the passengers belonged in class 3, followed next by class 1 and then class 2; most of class 3 perished, about half of class 2 perished, and less than half of class 1 perished. Figure 3 shows that most of the passengers were between ages of 20 and 40, with the average being around 30. Figure 4 shows that the majority of passengers had either 0 or 1 siblings/spouses on board with them. Figure 5 shows that the majority of passengers had either 0, 1, or 2 parents/children on board with them. Figure 6 shows that the majority of fares were less than 100, but there were some outliers above 100. Figure 7 shows that the majority of passengers embarked from Southampton, followed next by Cherbourg and then Queenstown. Most of the passengers from Southampton and Queenstown perished while a little less than half of those from Cherbourg perished.

d) Missing data

77.464% of the Cabin variable, 20.092% of the Age variable, and 0.153% of the Embarked variable are missing.

III. Methods

a) Variable removal and transformation

The following predictors will be removed: Cabin due to its lack of data, Name because that is not a feasible categorical predictor, and Ticket because it is not consistently numeric or categorical, so it would be difficult to use. After this removal, there will be 7 predictors left.

Next, any missing values will be replaced by the median (for numerical values) or mode (categorical values) for that predictor. Namely, the median of age 28 and the mode port of embarkation Southampton will replace all NA values.

Lastly, the Survived response variable will be converted into Yes/No.

b) Splitting the data

The dataset will then be split into 45% training set, 20% validation set, and 35% hold-out test set. This is equivalent to 589, 262, and 458 observations, respectively.

c) Classifiers

The goal is to classify the response variable Survived (Yes or No) based on the predictors. Three models will be fitted to the training data: a classification tree, a pruned tree, and a random forest. See Appendix for the code that builds the models.

d) Model selection

The final model will be based on the validation error rates.

e) Details on fitting the models

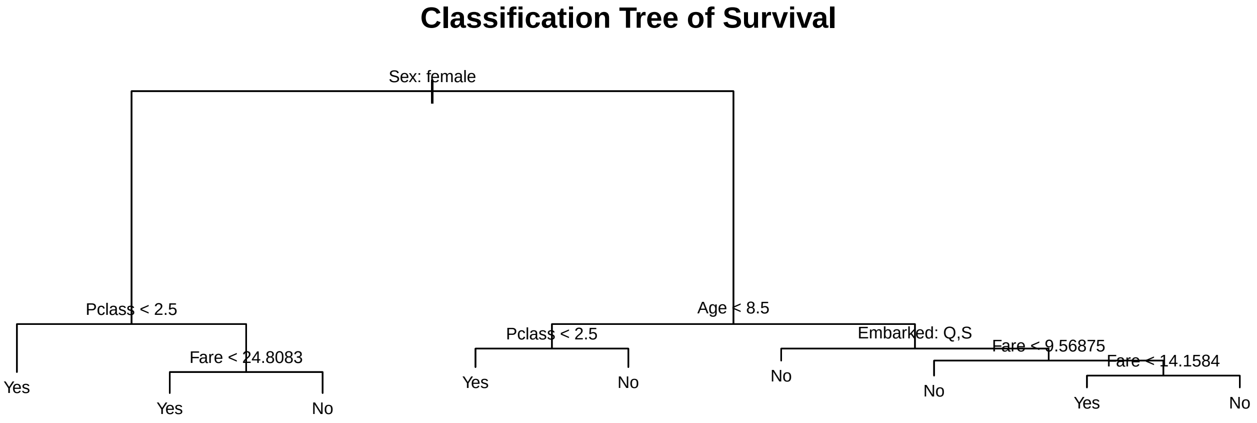
The classification tree is built on all the predictors using the tree() function from the tree library.

The pruned tree's optimal tree size is found using the cv.tree() function, which finds that the optimal number of terminal nodes is 9. The pruned tree is then built using all the predictors with the prune.misclass() function.

The random forest model is built using all predictors with the randomForest() function from the randomForest library.

IV. Model Building

a) Classification Tree

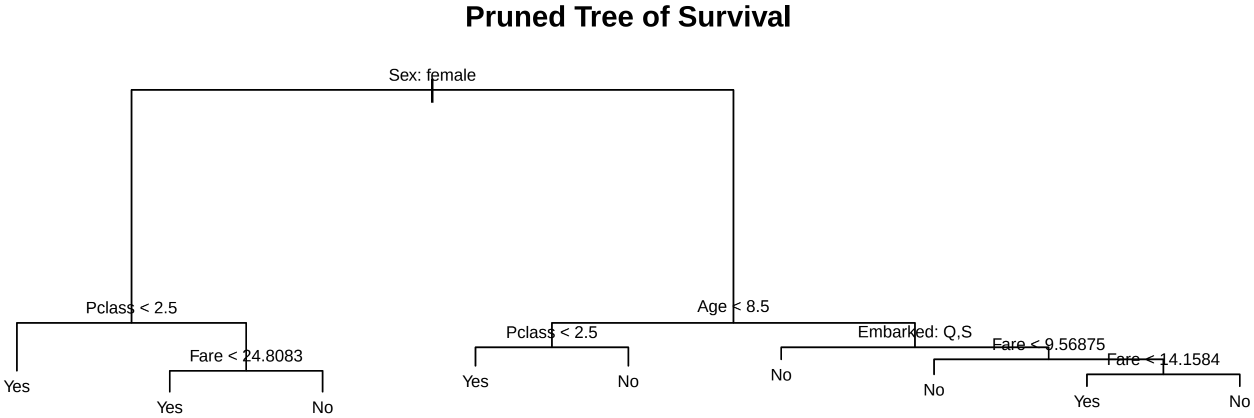


The classification tree splits on 5 of the predictors: Sex, Pclass, Age, Fare, and Embarked.

The tree's first split is on Sex. If the passenger was female and either in Pclass 1 or 2, then they are predicted to survive. If they were in Pclass 3, they'd only be predicted to survive if their fare was less than 24.81. If a male's age was less than 9, then they are predicted to survive only if they were in Pclass 1 or 2. If their age was 9 or greater, then they are predicted to survive only if they embarked from Cherbourg and their fare was between 9.57 and 14.15.

The training error of the model is 0.165 while the validation error is 0.218.

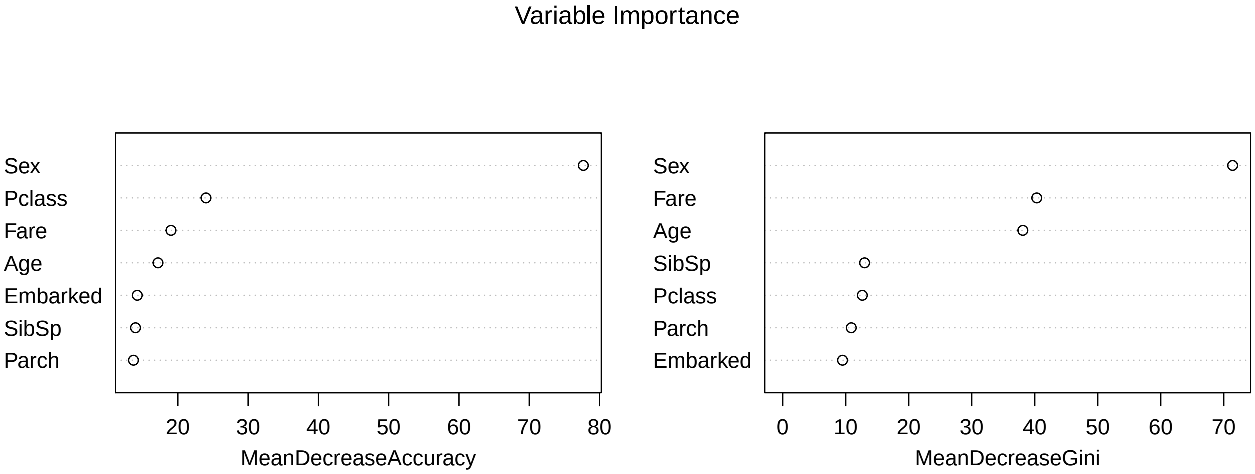
b) Pruned Tree



The pruned tree turns out to be exactly the same as the original tree.

It has 9 terminal nodes and 106 cross-validation errors. Just like the original tree, the training error of the model is 0.165 while the validation error is 0.218.

c) Random Forest



The random forest finds that Sex, Age, and Fare are all among the top 4 most important variables for both the mean decrease in accuracy and gini index metrics. For the mean decrease in accuracy, the fourth variable of the top 4 variables is Pclass; for the gini metric, it's SibSp.

The out-of-bag estimate of the error rate is 0.192. The training error of the model is 0.117 while the validation error is 0.191. Both the training error and validation error are better than those of the previous two models.

d) Selecting the final model

Out of all three models, the random forest has the best training and validation error rates, so it is selected to be the final model.

V. Conclusions

The final random forest model has a performance of 0.194 on the hold-out test set.

From our findings, we can conclude that women and children were generally the priority passengers to get onto lifeboats. Furthermore, having more siblings/spouses on board would make it more likely that a passenger would perish since they'd spend precious time gathering their family members together. Lastly, it's possible that the the distance from cabins to lifeboats was related to fare and passenger class, making certain passengers more likely to survive.

Potential research directions include further exploring the relationships among predictor variables, such as how fare interacts with passenger class and how sex interacts with age.

Appendix

```r
set.seed(2020)

# tree model
tree.fit <- tree(Survived~., data=train)
```

```r
set.seed(2020)

cv <- cv.tree(tree.fit, FUN=prune.misclass, K=10)
best.cv <- cv$size[which.min(cv$dev)]

# build model
pt.fit <- prune.misclass(tree.fit, best=best.cv)
```

```r
set.seed(2020)

# random forest model
rf.fit = randomForest(Survived~., data=train, importance=TRUE)
```

References

The dataset was obtained from http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets