Classification of Titanic Passenger Data and Chances of Surviving the Disaster

Machine Learning Kaggle Competition

DATA621 FINAL PROJECT

Abdellah AitElmouden | Gabriel Abreu | Jered Ataky | Patrick Maloney

Introduction

The goal of the project was to predict the survival of passengers based off a set of data. We used Kaggle competition "Titanic: Machine Learning from Disaster" (see https://www.kaggle.com/c/titanic/data) to retrieve necessary data and evaluate accuracy of our predictions. The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with the outcome (whether or not a passenger survived). We used this set to build our model to generate predictions for the test set.

Modeling Plan

- Data Exploration: summary statistics and simple visualizations were created to search for relationships between the variables.

- Data Preparation: null values were imputed and new features were engineered.

- Logistic Regression Modeling: A binomial logistic regression model was used as an initial comparison for the following model that was simplified using stepwise regression.

Modeling Plan

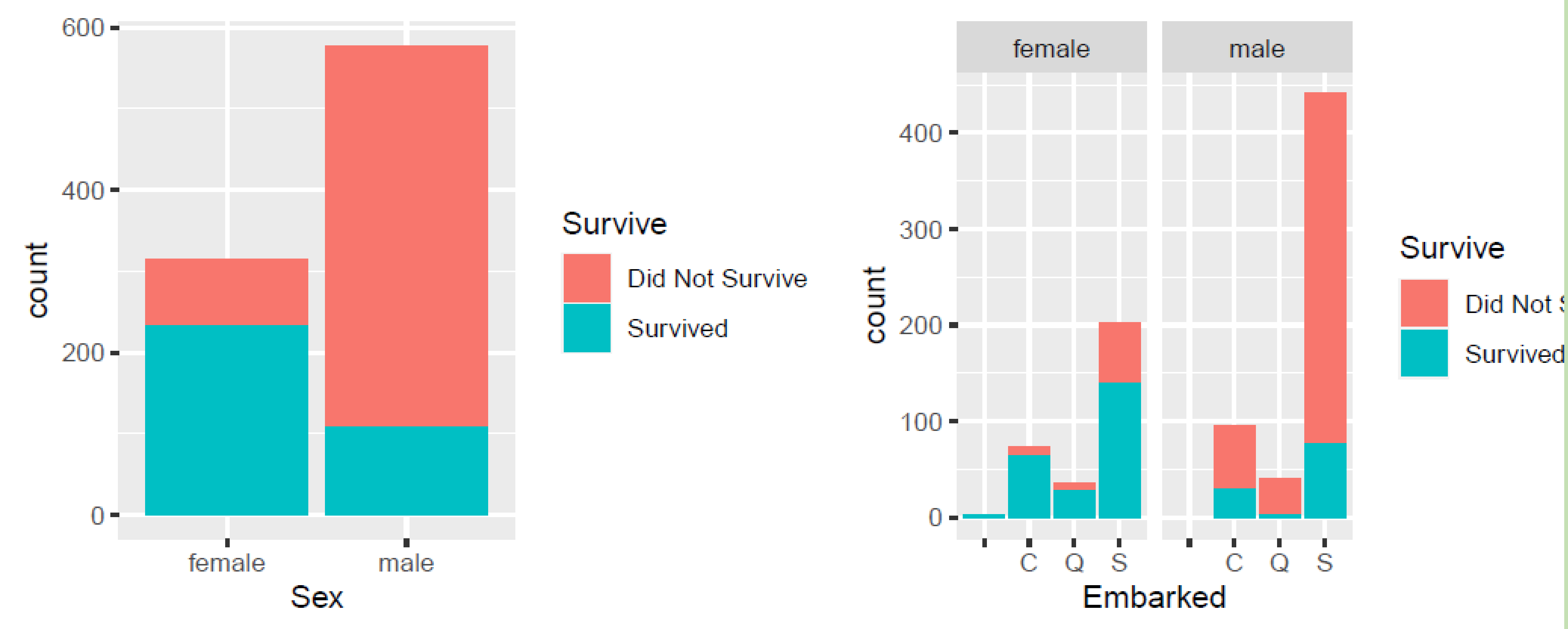
- Random Forest Modeling: two different random forest functions were used from different packages to be sure we had the best version.

- Evaluation: the test data was then cleaned and run through the models and their performance was evaluated. Additional tweaks to the models were made in attempt to improve performance.

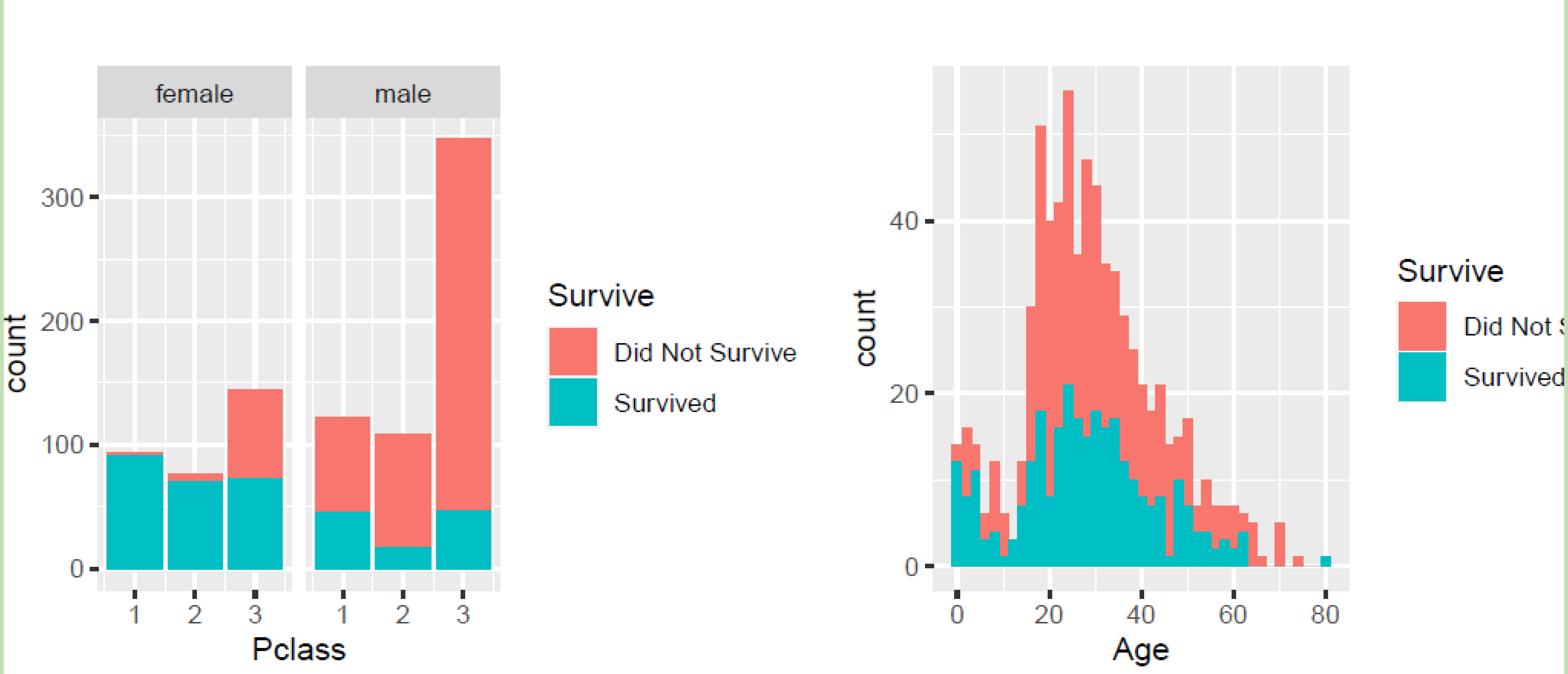
Variable Descriptions

```
| Variable | Desctiption                                                          |
|----------|----------------------------------------------------------------------|
| survival | Survival (0 = No; 1 = Yes)                                           |
| pclass   | Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)                          |
| name     | Name                                                                 |
| sex      | Sex                                                                  |
| age      | Age                                                                  |
| sibsp    | Number of Siblings/Spouses Aboard                                    |
| parch    | Number of Parents/Children Aboard                                    |
| ticket   | Ticket Number                                                        |
| fare     | Passenger Fare                                                       |
| cabin    | Cabin                                                                |
| embarked | Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton) |
```

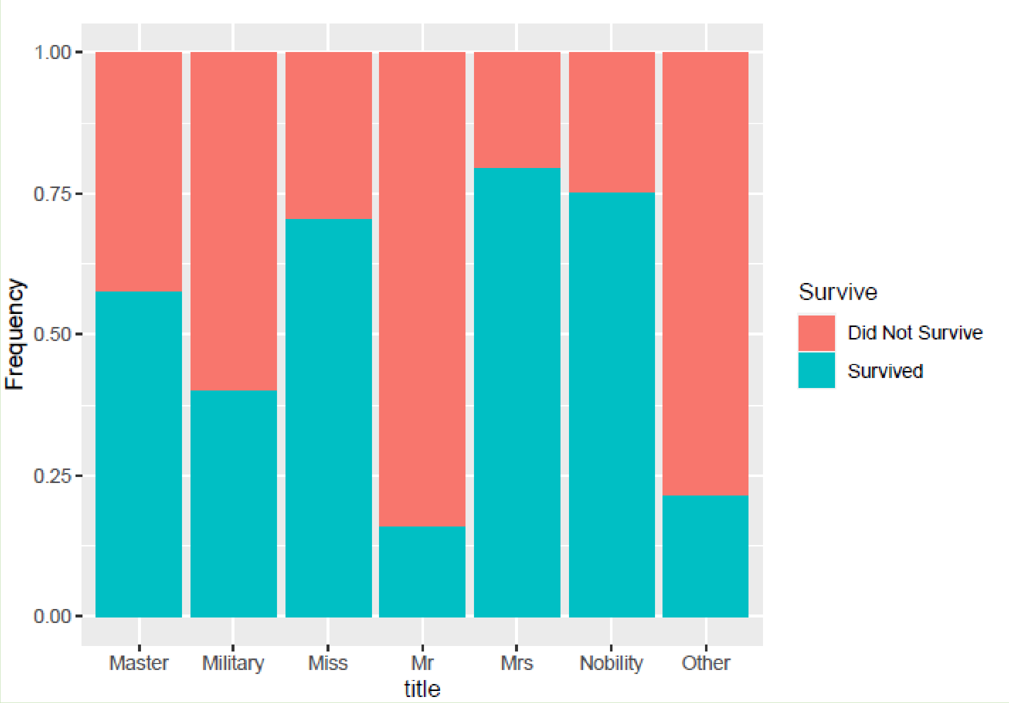
Data Exploration



Data Exploration

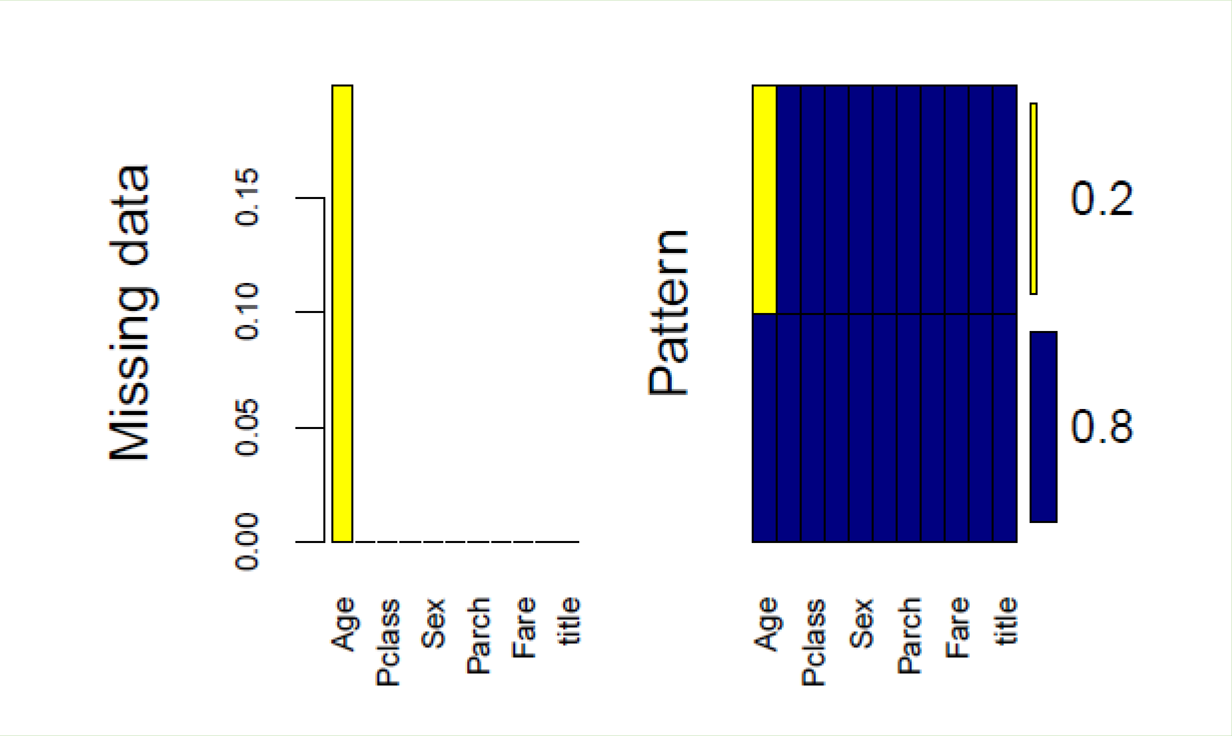


Data Exploration



Based on the visuals, it seemed gender and class had an effect on a passenger's probability of surviving. We then looked at the titles associated with the passenger's name.

Data Preparation



There are three variables with missing or empty values based on our exploration of the data and visualizations: Embark, Cabin, and Age. Only passengers 62 and 830 are missing their embark ports. We randomly assigned them a value of “C”. The column Cabin had too many missing values to impute or fill, so we dropped the Cabin column from the training data set.

Modeling

After exploring the patterns and creating new features, now I will build statistical models to predict the fate of the passengers in the test data set.

Three machine learning methods are used in this project:

- Binomial with logit link function (w/ Imputed data)
- Stepwise regression
- Random Forest

Model 1: Binomial with logit link function

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + title + Embarked + Fare +
##     Age + SibSp + Parch, family = binomial(link = "logit"), data = train3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4068  -0.5433  -0.3760   0.5370   2.5761
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    20.721916 481.641207   0.043  0.96568
## Pclass         -1.140275   0.161793  -7.048 1.82e-12 ***
## Sexmale       -15.092293 481.640636  -0.031  0.97500
## titleMilitary  -2.779350   1.133695  -2.452  0.01422 *
## titleMiss     -15.629763 481.640884  -0.032  0.97411
## titleMr        -3.405782   0.548291  -6.212 5.24e-10 ***
## titleMrs      -14.742217 481.640937  -0.031  0.97558
## titleNobility  -2.728641   1.529793  -1.784  0.07448 .
## titleOther     -3.995306   0.979950  -4.077 4.56e-05 ***
## EmbarkedQ      -0.103967   0.397173  -0.262  0.79350
## EmbarkedS      -0.414950   0.248222  -1.672  0.09459 .
## Fare            0.003291   0.002608   1.262  0.20689
## Age            -0.031319   0.009698  -3.229  0.00124 **
## SibSp          -0.569119   0.127482  -4.464 8.03e-06 ***
## Parch          -0.365867   0.136334  -2.684  0.00728 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
```

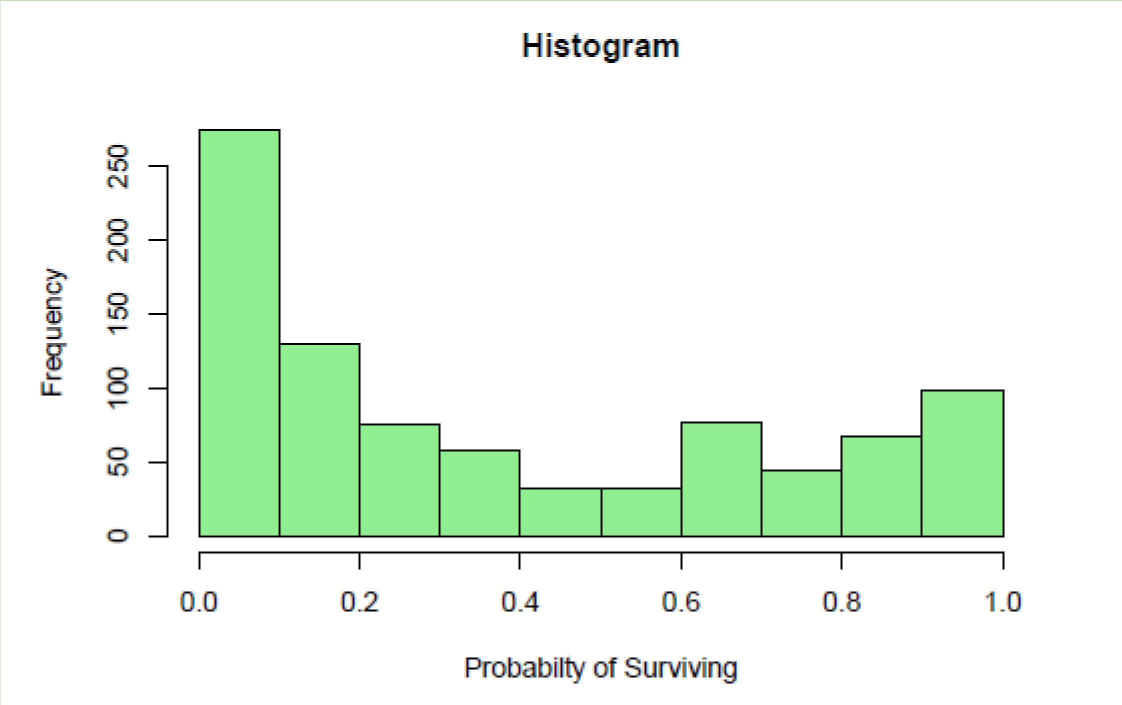
This model is used when the data have binary outcomes. We fit a generalized linear model (binomial with logit link function) with Pclass, Sex, title, Embarked, Fare, Age, SibSp and Parch, as predictors of the number of survived passenger. we can see that Pclass, Sex, Age and SibSp are all significant variables.

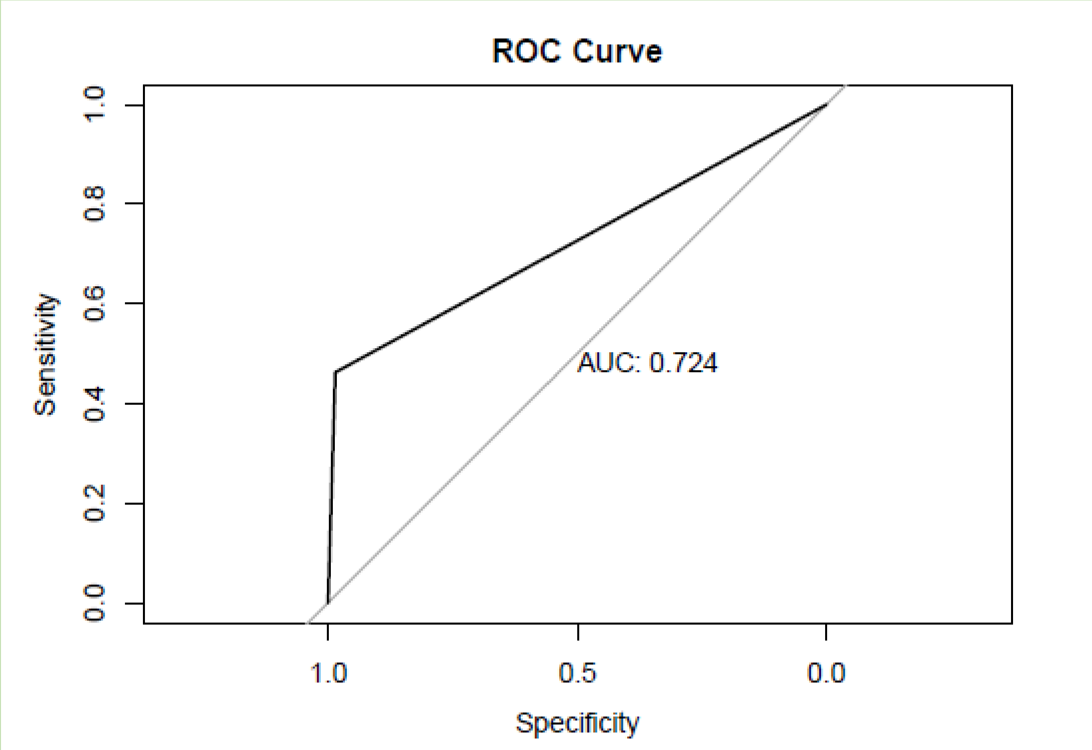
Model 2: Stepwise

```
## Start:  AIC=752.12
## Survived ~ Pclass + Sex + title + Embarked + Fare + Age + SibSp +
##     Parch
##
##            Df Deviance    AIC
## - Embarked  2   725.41 751.41
## - Fare      1   723.92 751.92
## <none>          722.12 752.12
## - Sex       1   726.54 754.54
## - Parch     1   729.88 757.88
## - Age       1   733.03 761.03
## - SibSp     1   747.11 775.11
## - title     6   780.25 798.25
## - Pclass    1   772.73 800.73
##
## Step:  AIC=751.41
## Survived ~ Pclass + Sex + title + Fare + Age + SibSp + Parch
##
##          Df Deviance    AIC
## <none>        725.41 751.41
## - Fare    1   728.49 752.49
## - Sex     1   729.60 753.60
## - Parch   1   733.84 757.84
## - Age     1   736.90 760.90
## - SibSp   1   753.19 777.19
## - title   6   783.55 797.55
## - Pclass  1   777.98 801.98
```

```
Call:
glm(formula = Survived ~ Pclass + Sex + title + Fare + Age +
    SibSp + Parch, family = binomial(link = "logit"), data = train3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4747  -0.5493  -0.3854   0.5225   2.6670

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    20.321420 481.372829   0.042 0.966327
Pclass         -1.132427   0.157823  -7.175 7.21e-13 ***
Sexmale       -14.986038 481.372309  -0.031 0.975164
titleMilitary  -2.813948   1.128833  -2.493 0.012674 *
titleMiss     -15.516876 481.372559  -0.032 0.974285
titleMr        -3.444979   0.545960  -6.310 2.79e-10 ***
titleMrs      -14.666556 481.372612  -0.030 0.975694
titleNobility  -2.640764   1.543909  -1.710 0.087185 .
titleOther     -3.920256   0.965141  -4.062 4.87e-05 ***
Fare            0.004198   0.002594   1.618 0.105608
Age            -0.031856   0.009633  -3.307 0.000943 ***
SibSp          -0.591944   0.126686  -4.673 2.97e-06 ***
Parch          -0.378039   0.135425  -2.791 0.005247 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  725.41  on 878  degrees of freedom
AIC: 751.41
```

Model 2: Stepwise



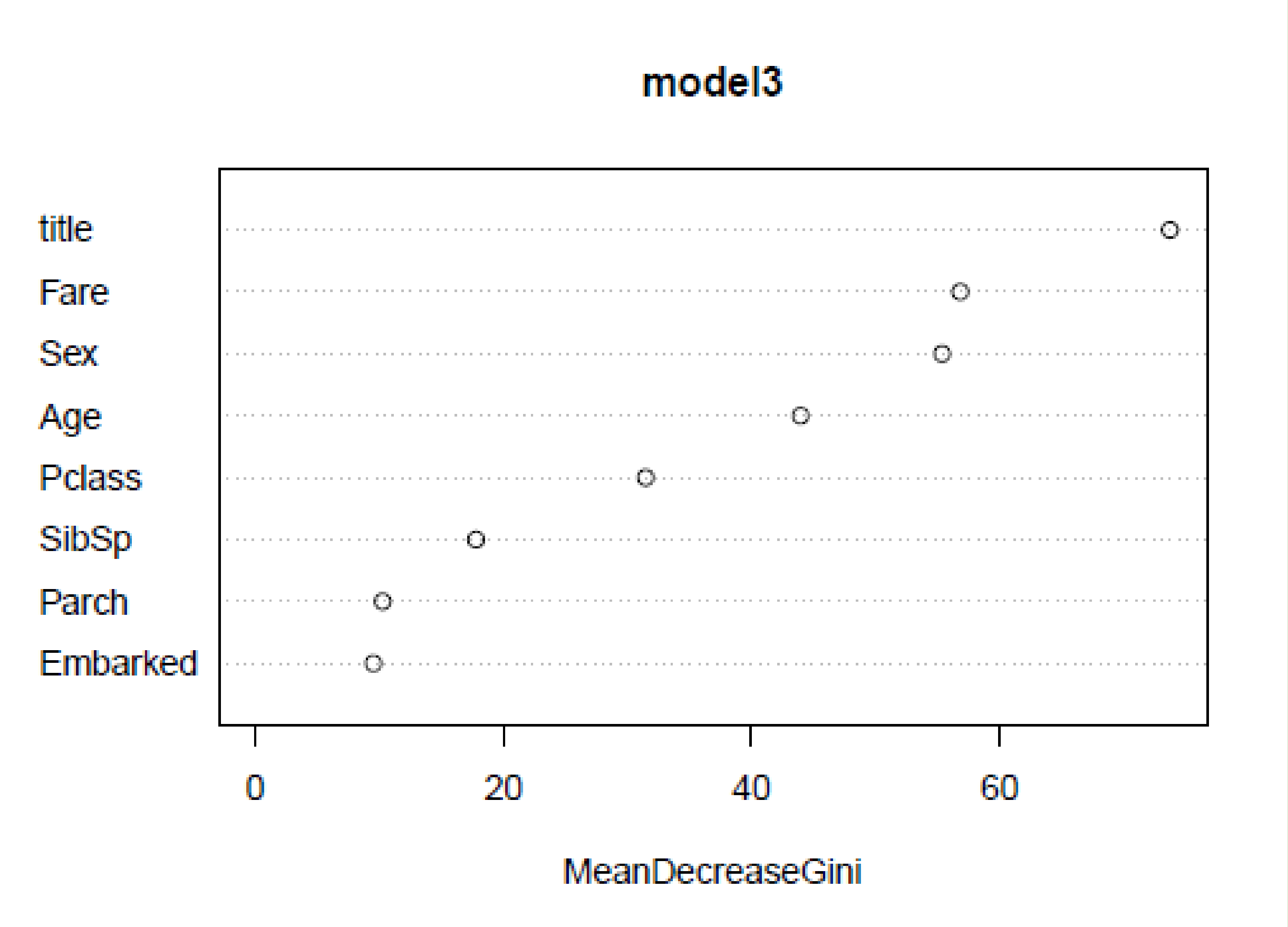


Model 3: Random Forest

Random Forest is our favorite machine learning algorithm so far. We will use the randomForest function from the randomForest package.

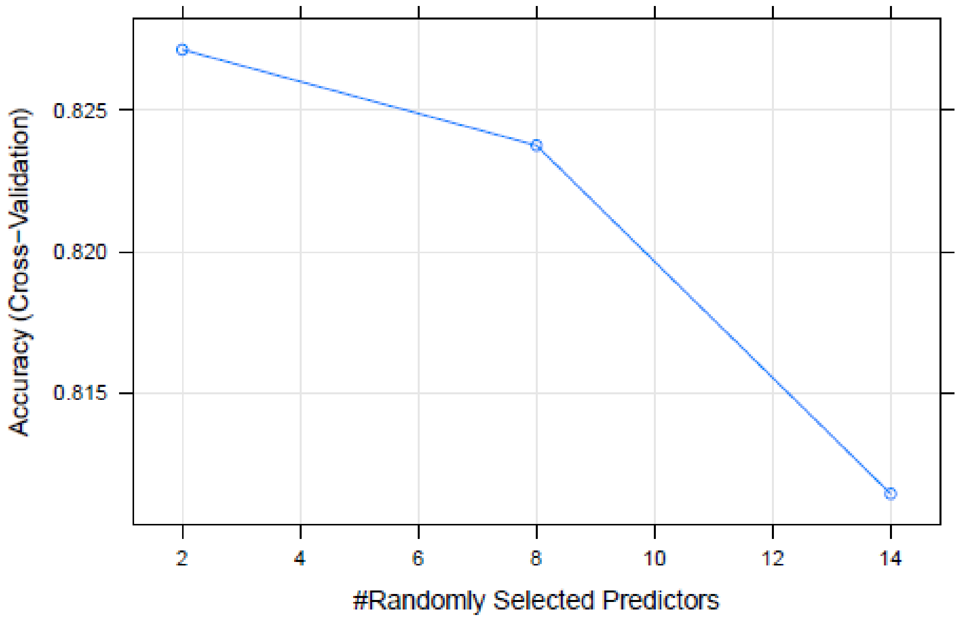
```
Random Forest

891 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 801, 802, 802, 802, 802, 803, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8238662  0.6207001
   8    0.8361880  0.6488591
  14    0.8170480  0.6110314

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 8.
```

Model 3: Random Forest

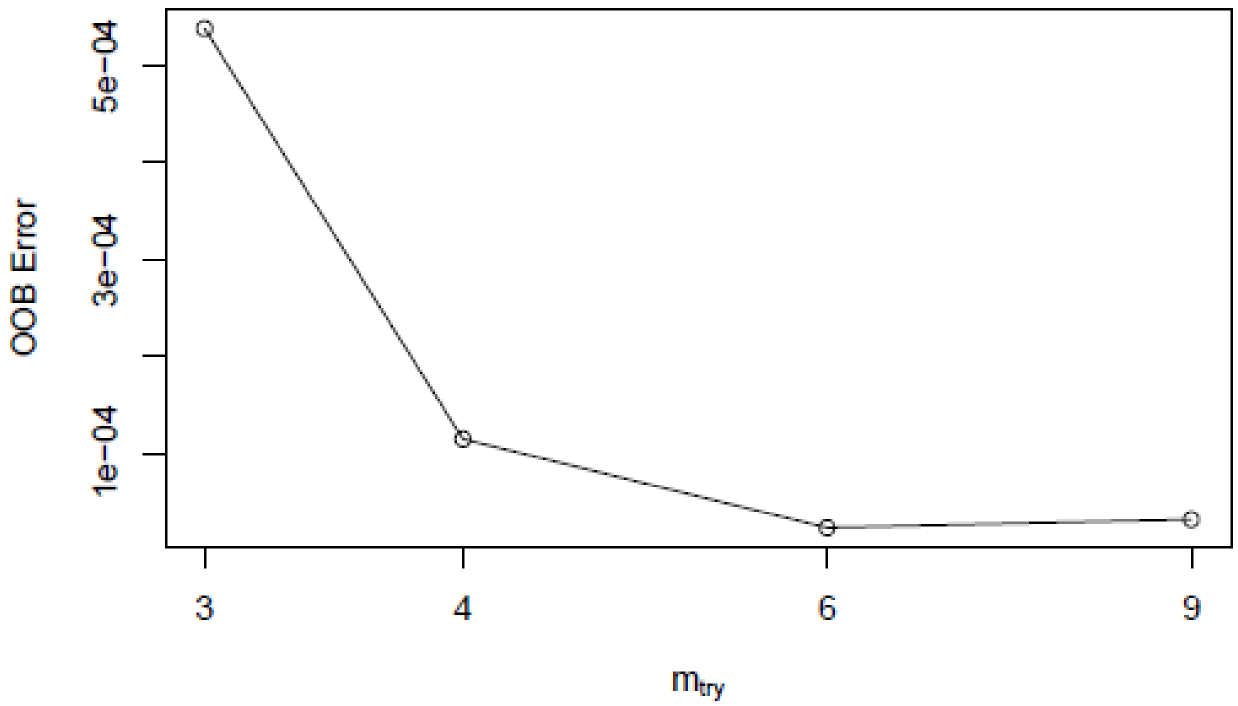


The plot shows the importance of the variables judged by the mean decrease accuracy. A variable is considered the most important if the accuracy of the model without it decreased the most compared to the full model.

Model 3: Random Forest




```
## -3.704311 0.01
## 0.7988898 0.01
## -0.3660311 0.01
```



Model Selection

The model with the most accurate result is model 2. Tweaking the thresholds changed results on Kaggle but changing the threshold to 0.75 seemed to be optimal (produced score of 0.77511), while the random forest models produced scores of 0.75358.

The models might see greater accuracy testing different methods of imputation. The age column saw the
greatest amount of missing values, focusing on creating accurate age values will most likely improve the
models.

Summary

In this project, We practiced:

- Exploratory data analysis with tidyversse, ggplot2, and rpart..etc
- We learned Several machine learning algorithms, and modeling with caret, randomForest, and other packages.
- Feature Engineering techniques.
- We used Rstudio and all the skills and methodologies we learned during this semester.