Machine Learning from Disaster

Kaggle Titanic Competition

Roshane Taylor,
April Sapp,
Michael Puerto,
Reshma Jayaram,
Jazarai Sturdivant

Our goal

Data from 891 people on board the Titanic will be partitioned into 70% train and 30% test datasets. Both of these datasets contain the same variables, but for different passengers. Splitting the data into parts is common practice in predictive analytics and it tests for over/under fitting models by introducing new data that was not used to create the model to test its prediction accuracy. For prediction, we are less worried about the format of our predictive formulas. Instead, we are focused on the performance of our predictive formulas. We will be using AUC and ROC to compare our models.

Viewing the Data

train <- read.csv("train.csv")

view(train)
head(train)
tail(train)

summary(train)
names(train)

```
> summary(train)
  PassengerId
 Min.   :  1.0
 1st Qu.:223.5
 Median :446.0
 Mean   :446.0
 3rd Qu.:668.5
 Max.   :891.0

    Survived
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.3838
 3rd Qu.:1.0000
 Max.   :1.0000

     Pclass
 Min.   :1.000
 1st Qu.:2.000
 Median :3.000
 Mean   :2.309
 3rd Qu.:3.000
 Max.   :3.000
```

```
                                    Name
Abbing, Mr. Anthony                   :  1
Abbott, Mr. Rossmore Edward           :  1
Abbott, Mrs. Stanton (Rosa Hunt)      :  1
Abelson, Mr. Samuel                   :  1
Abelson, Mrs. Samuel (Hannah Wizosky):  1
Adahl, Mr. Mauritz Nils Martin        :  1
(Other)                               :885
    Sex           Age
female:314   Min.   : 0.42
male  :577   1st Qu.:20.12
             Median :28.00
             Mean   :29.70
             3rd Qu.:38.00
             Max.   :80.00
             NA's   :177
    SibSp
Min.   :0.000
1st Qu.:0.000
Median :0.000
Mean   :0.523
3rd Qu.:1.000
Max.   :8.000
```

```
         Cabin      Embarked
            :687     :  2
B96 B98     :  4    C:168
C23 C25 C27:  4     Q: 77
G6          :  4    S:644
C22 C26     :  3
D           :  3
(Other)     :186
```

```
     Parch
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.3816
3rd Qu.:0.0000
Max.   :6.0000

      Ticket
1601     :  7
347082   :  7
CA. 2343:  7
3101295  :  6
347088   :  6
CA 2144  :  6
(Other)  :852
      Fare
Min.   :  0.00
1st Qu.:  7.91
Median : 14.45
Mean   : 32.20
3rd Qu.: 31.00
Max.   :512.33
```

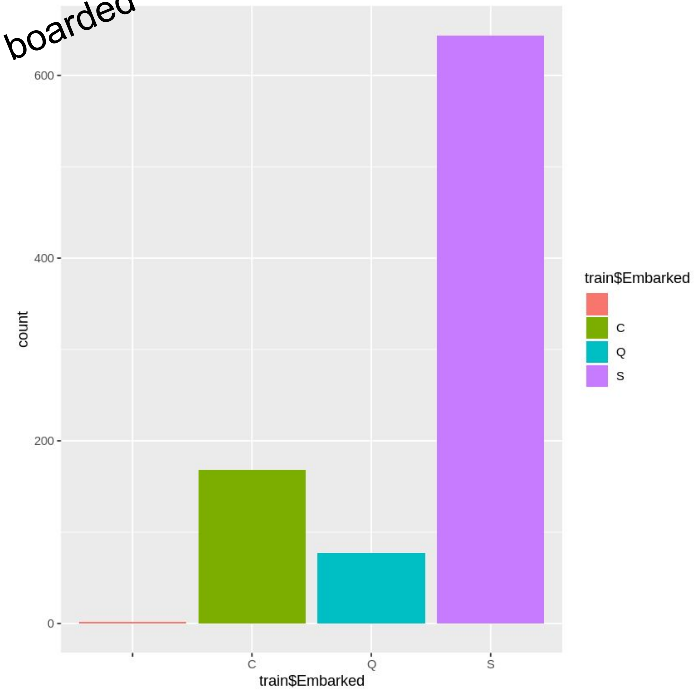
Viewing the Data

```
> head(train)
  PassengerId Survived Pclass
1           1        0      3
2           2        1      1
3           3        1      3
4           4        1      1
5           5        0      3
6           6        0      3
                                                 Name    Sex
1                             Braund, Mr. Owen Harris   male
2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
3                              Heikkinen, Miss. Laina female
4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female
5                            Allen, Mr. William Henry   male
6                                    Moran, Mr. James   male
  Age SibSp Parch           Ticket    Fare Cabin Embarked
1  22     1     0        A/5 21171  7.2500              S
2  38     1     0         PC 17599 71.2833   C85        C
3  26     0     0 STON/O2. 3101282  7.9250              S
4  35     1     0           113803 53.1000  C123        S
5  35     0     0           373450  8.0500              S
6  NA     0     0           330877  8.4583              Q
```

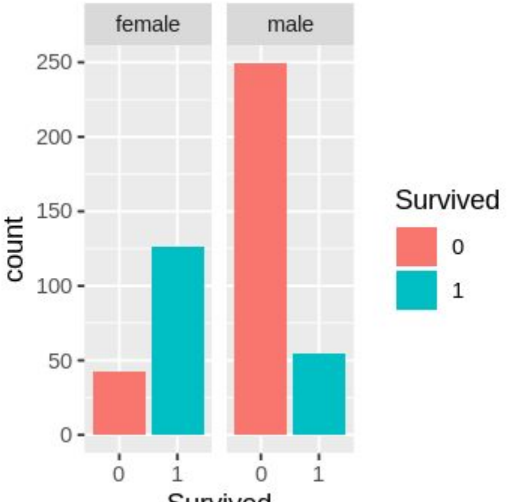
```
> names(train)
 [1] "PassengerId"
 [2] "Survived"
 [3] "Pclass"
 [4] "Name"
 [5] "Sex"
 [6] "Age"
 [7] "SibSp"
 [8] "Parch"
 [9] "Ticket"
[10] "Fare"
[11] "Cabin"
[12] "Embarked"
```

Understanding the data.

How many people boarded where?



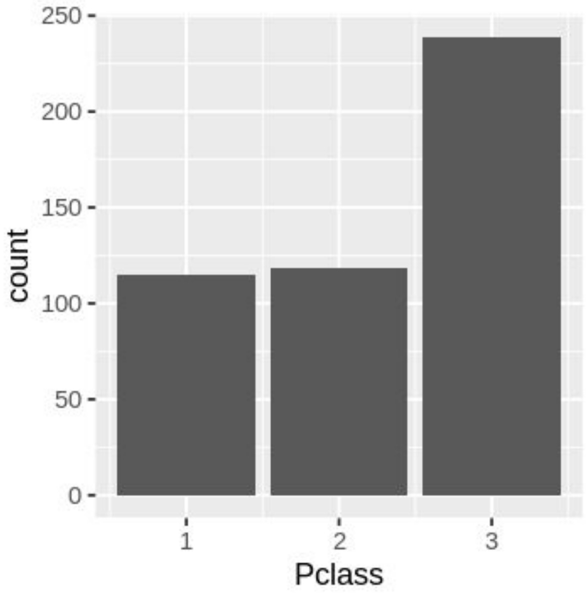
Do males survive more/less than females?



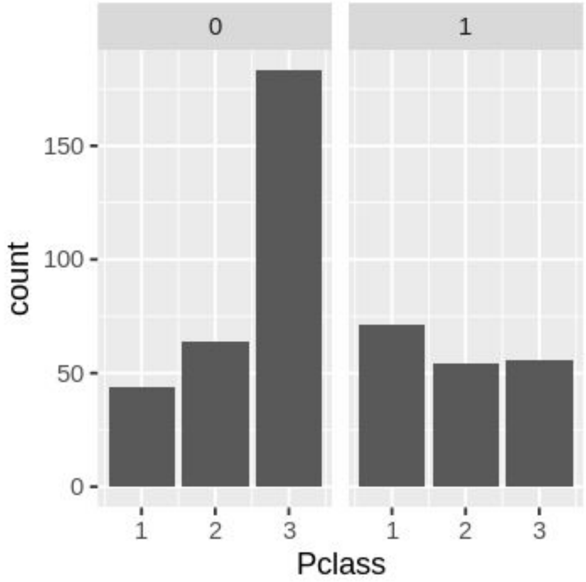
We should look at the proportions!

Understanding the data

How many people belong to each class?

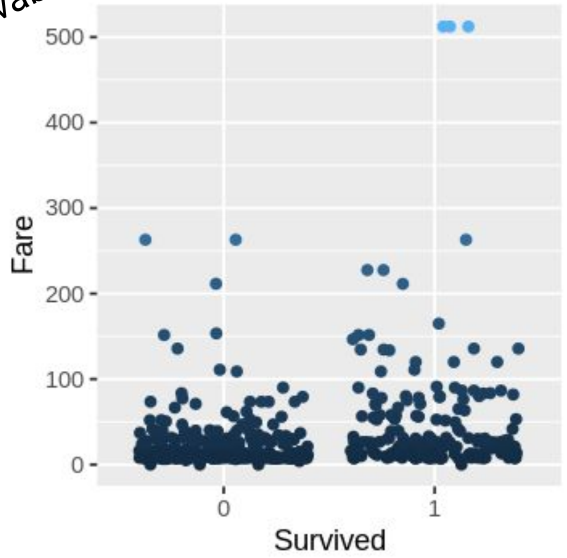


Is class related to survivability?

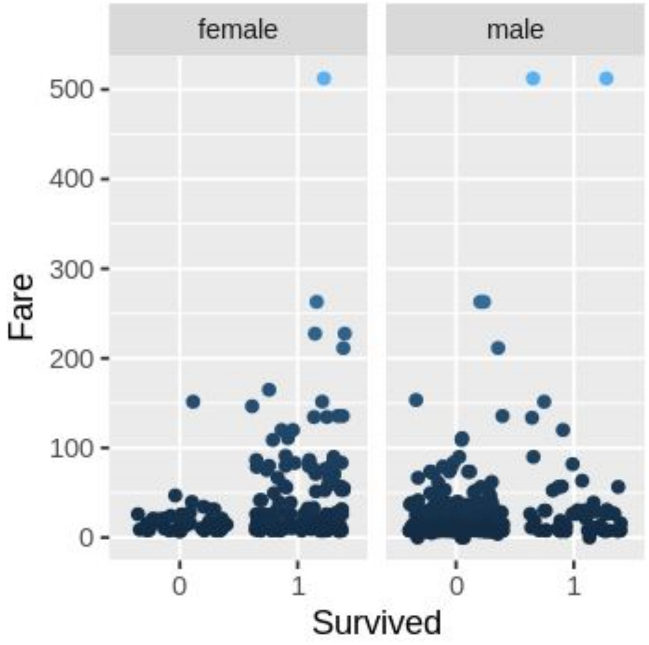


Understanding the data

Does fare appear to increase chances of survivability?



Is this different for males and females?



Selecting Variables

Excluded Variables

- **Name** : This variable contains the names of passengers and the peoples that were on the boat, duplicated information with SibSp and Parch
- **Ticket** : Ticket numbers provide no information relevant to survival
- **Embarked**: Port of boarding correlates with fare
- **Cabin**: a lot of missing values, relationship with fare

Included Variables

- **PassengerID:** Easily reference passengers
- Survived: able to identify survival
- **Sex:** complete data
- Age: Speaks to physical condition
- **SibSp:** Groups could be a factor
- Parch: Groups could be a factor
- **Fare:** Location of cabin and port information
- **Pclass:** Passenger Class information

Steps

Using dplyr, we can select which variables to keep. Also, we can use these functions to delete observations with NA.

```r
test <- test %>%
select(-c(Cabin, Ticket, Name, Embarked))
train <- train %>%
select(-c(Cabin, Ticket, Name, Embarked))
```

```r
train <- na.omit(train)

```

Data Cleaning: Missing Values

We believe that missing values that are currently in the dataset are missing at random, and that the survivability of persons on board are independent from one another.

For example: knowing that a person on board is 14 years old and has 3 siblings onboard doesn't tell me about the survivability about some other person onboard.

Therefore, we will only keep the complete cases to help build our model.

```
1 train %>%
2 summarise_all(funs(sum(is.na(.))))
```

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
0	0	0	0	111	0	0	0

Data Cleaning the missing test data for submission

The test data that was provided by kaggel had some missing data… We dealt with them in these ways.

Missing data in Fare variable: Thankfully there was only one missing value, though we would like to have better imputation methods, we just used the average of Fare.

Missing data in Age variable: Unfortunately, there were much more missing values in this variable. We could do 1 of 2 things: Get rid of the variable in our models, or impute the values. We do both in this case, finding the best model out of it’s class to predict survival. Then dropped Age from the model OR impute Age in the test data.

We just used the average of age to impute the missing values in this score too.

Furthermore

We will be splitting our training data down further to compare how the model uses “new” data.

We decided to further split the training data because the survived column/data does not exist in the test dataset, and we would like to see how the model would perform with data not use to train the model.

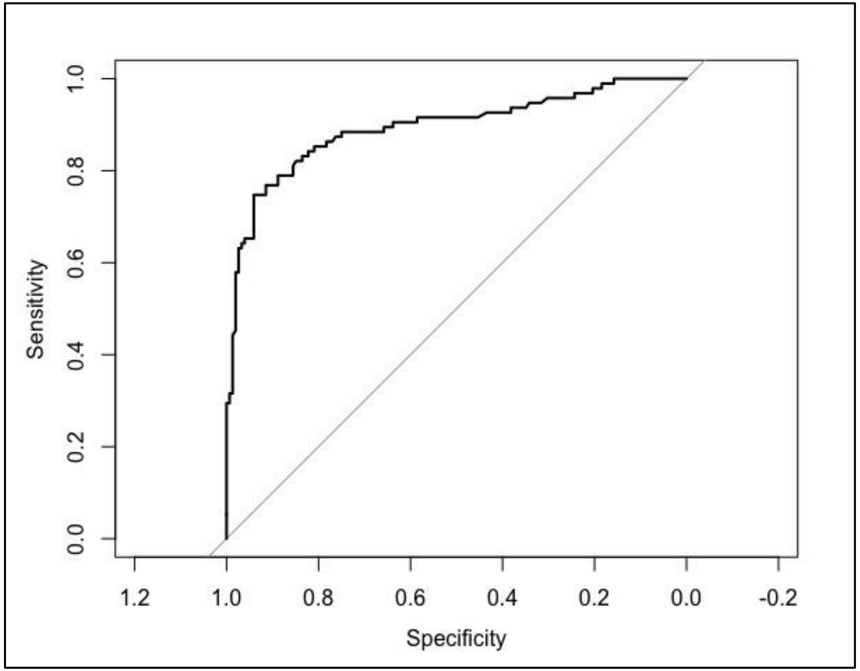
Trying Several Logistic Models

```
Survived ~ Pclass + Sex + Age + SibSp +
Parch + Fare

Area under the curve: 0.8036

Survived ~ Pclass + Sex + Age

Area under the curve: 0.8069

Survived ~ Pclass*Sex*Age*SibSp + Parch

Area under the curve: 0.8134
```

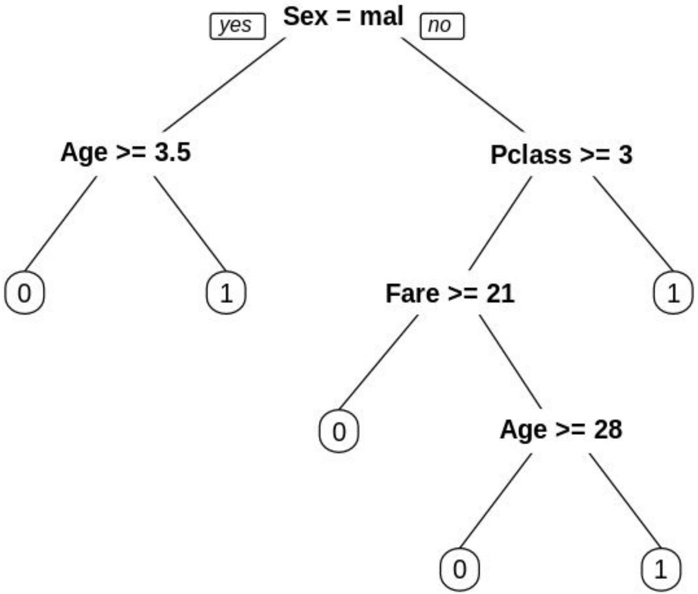
```
Survived ~ Fare

Area under the curve: 0.7211

Survived ~ Pclass + Sex + Age + SibSp
+ Parch + Fare

Area under the curve: 0.8036
```

Selecting the Best Logistic Model

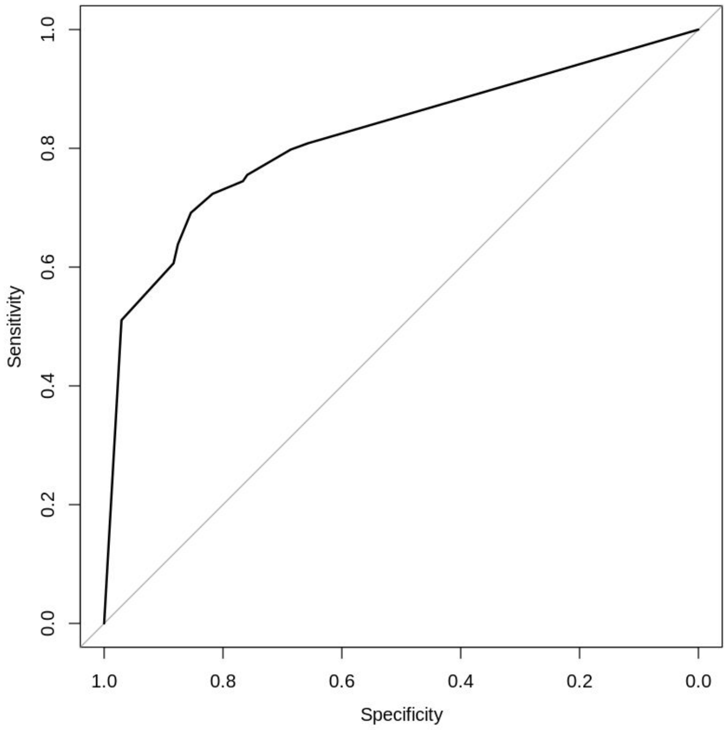
```
Survived ~
Pclass*Sex*Age*SibSp*Parch,

 Area under the curve: 0.8325
```



Decision tree

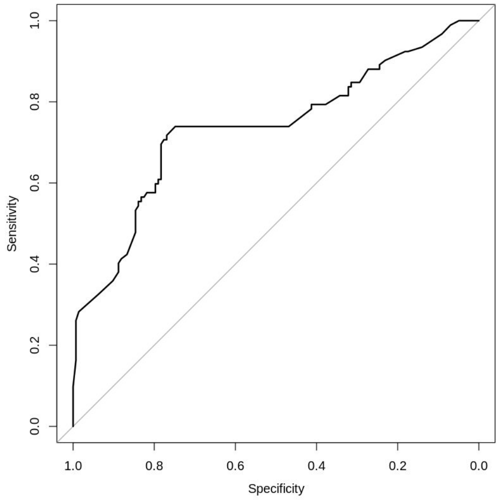


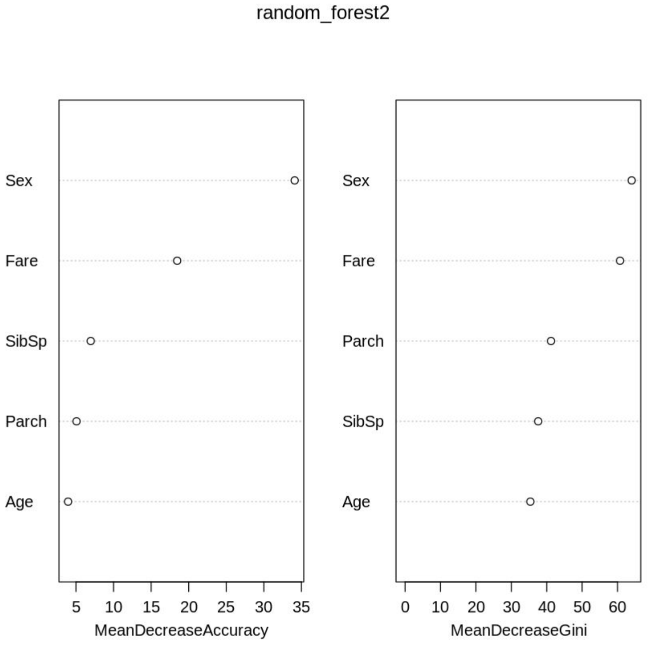
Area under the curve : 0.82



Random Forest

This random forest model has an AUC of 0.71

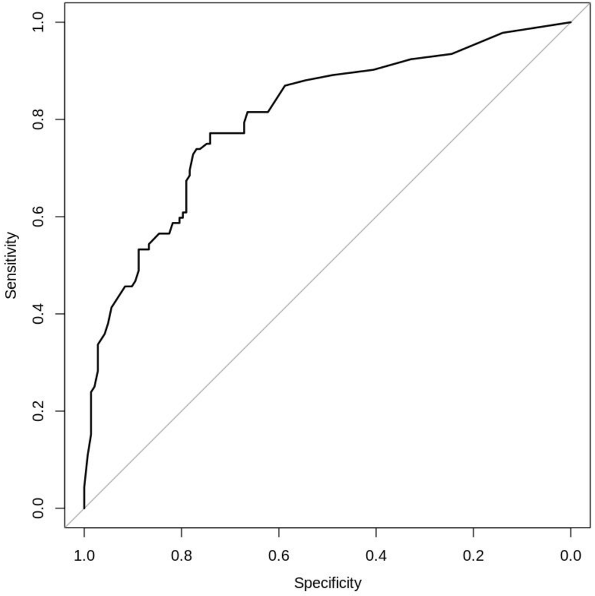




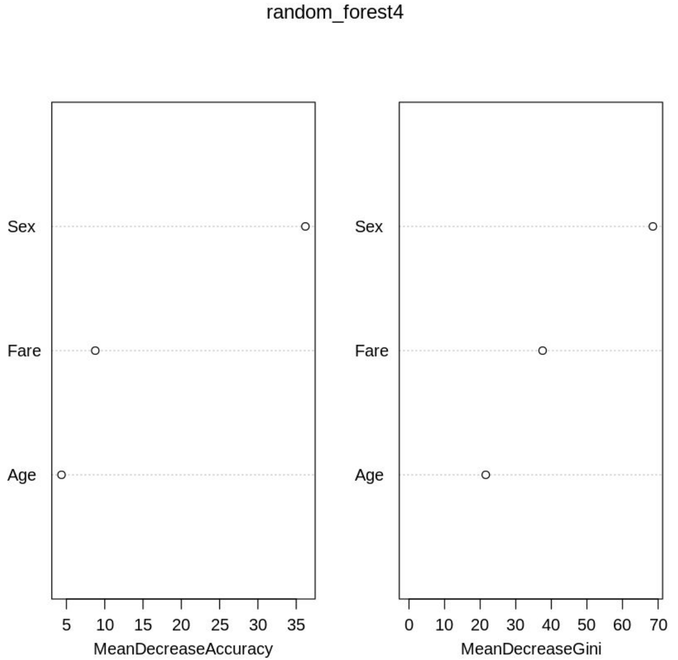

```
train_ROSE <- ROSE(Survived ~ ., data = train)$data

random_forest2 <- randomForest(Survived ~  Sex + Age + SibSp + Parch + Fare,
                               data = train_ROSE, ntree=100, importance = TRUE)
```

Simpler Random Forest

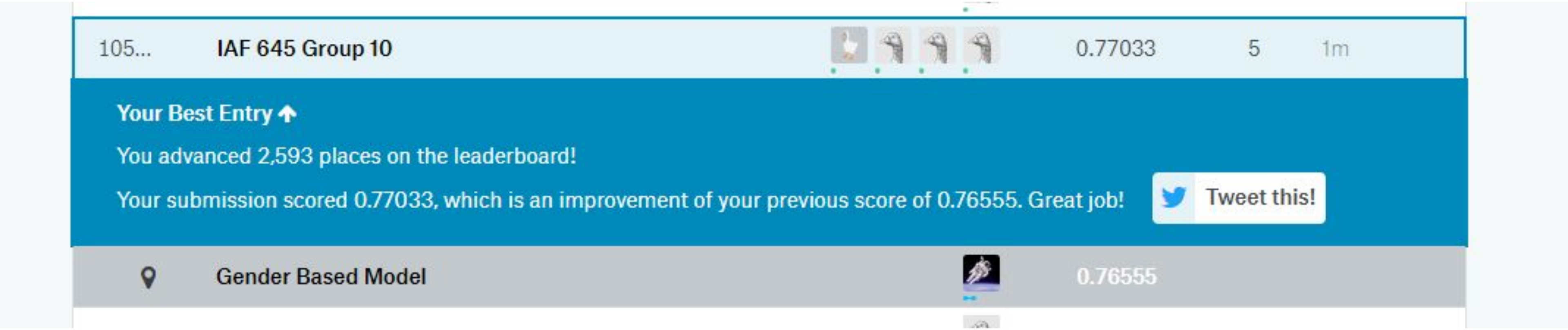
```
random_forest4 <- randomForest (factor(Survived) ~ Sex + Age +
Fare, data = train, ntree=100, importance = TRUE)
```



This random forest has an AUC of 0.80



Which Model is the Best?



Our best model was the logistic regression. In this particular entry, we removed Age as predictor variable from the model rather than impute age. We compared this to other models where age was imputed. These will be featured in the next slide.

Spoiler: this model which did not include and impute age, did better than the model that did include and impute age :)

glm_without_age.csv 3 minutes ago by mowgli glm without age `Pclass*Sex*SibSp*Parch`	0.77033	☑
glm_with_age.csv 3 minutes ago by mowgli glm with age `Pclass*Sex*Age*SibSp*Parch`	0.76555	☐
randomF_simple_with_age.csv 4 minutes ago by mowgli 3 most important variables Randomforest `Sex + Age + Fare`	0.76076	☐
randomF_simple_without_age.csv just now by mowgli rf simple without age `Sex + Fare`	0.76555	☐
randomF_all_without_age.csv 5 minutes ago by mowgli randomForest all variables without age `Sex + SibSp + Parch + Fare`	0.73205	☐
randomF_all_with_age.csv 12 minutes ago by mowgli `Survived ~ Sex + Age +` randomForest using all variables with a `SibSp + Parch + Fare` means	0.76076	☐

It's worth noting that the random forest model that only contains 2 variables does pretty well compared to other, more complex models.

Who Will Survive using our Models?

1. We plan to identify and categorize at risk individuals using our predictive models
   a. Ideally we would like to be able to explain why our predictor variables lead to target variable
2. Big Picture: Can we reduce number of casualties in future cruise ship accidents?

The end :)

Mowgli- Exploratory Data Analysis

Reshma- Logistic Regression

Roshane- Decision Tree

Jazarai and April- Random Forest