Capstone Project

BAP - R

Project Title: **Predict survival of passengers on the Titanic ship.**

Abstract:

The goal of this project is to analyze the given titanic data and predict the whether a particular passenger survived the sinking of the ship or not. The data was divided into two parts i.e. the training and the test dataset. Models were developed based on the training dataset and applied to the test dataset to find out the accuracy of each model based on the predicted values generated. Based on these values we can determine how good a particular model is for prediction of survival of the passenger.

Submitted By:

Sujay Gokhale

Table Of Contents

Capstone Project – BAP-R

Predict the survival of passengers on the Titanic Ship

Introduction

On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This tragedy shocked the international community and lead to better safety regulations for ships.

Number of observations in the given dataset: 891

The Dataset

Categorical Variables

1) **Survived**: Survival

(0 = No; 1 = Yes)

2) **Pclass**: Passenger Class (1 = 1st, Upper ; 2 = 2nd, Middle; 3 = 3rd, Lower)

3) **Sex**: male, female

4) **Embarked**:

C = Cherbourg;

Q = Queenstown

S = Southampton

Numerical Variables

Age: Passenger Age (In years)

Fare: Passenger Fare (In pounds)

Sibsp: Number of Siblings/Spouses Aboard

```
> summary(mydata)
  PassengerId       Survived          Pclass                                         Name          Sex           Age       
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                  :  1   female:314   Min.   : 0.42  
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :  1   male  :577   1st Qu.:20.12  
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :  1                Median :28.00  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :  1                Mean   :29.70  
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00  
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :  1                Max.   :80.00  
                                                  (Other)                              :885                NA's   :177    
     SibSp           Parch             Ticket          Fare                Cabin       Embarked
 Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00              :687    :  2   
 1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91   B96 B98    :  4   C:168   
 Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45   C23 C25 C27:  4   Q: 77   
 Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20   G6         :  4   S:644   
 3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00   C22 C26    :  3           
 Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33   D          :  3           
                                  (Other) :852                    (Other)    :186           
```

Summary of the data/ Review of Literature

- We can see that there are a total of 891 passengers on board the titanic of which 314 are females and the rest 577 are males.

- The age of the passengers range from a few months old to a maximum of 80 years old. However, all passengers didn't report their age or the data wasn't collected for them so their age is marked as NA. This will be fixed while cleaning the data as such a large chunk of data cannot be ignored or deleted from this dataset.

- Initial impressions of looking at the variables 'Survived' and 'Pclass' seem to be numeric in nature. However, these variables should be of the factor type so we will need to convert this into factor type to ensure correct analysis.

- The names of the passengers travelling seem to be factors so this will need to be converted to the character type.

- One interesting observation regarding the fare is that some of the people travelled for free whereas the highest fare was 512.33, which shows that there was a huge difference in the ticket prices. But it can be possible that the infants were given free tickets whereas the passengers occupying the most luxurious rooms paid higher for their rooms.

- Port of Embarkation -168 people embarked at Cherbourg, 68 at Queenstown and 644 at Southampton. We can also see that for 2 passengers the data is missing.

Transforming Data

Treating missing data

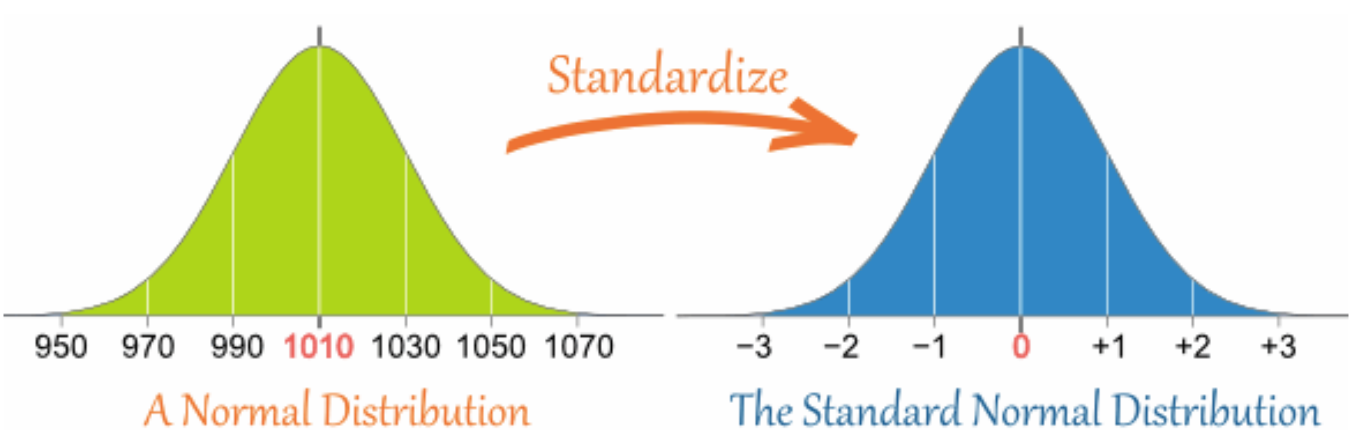
The missing data must be treated to ensure accurate analysis.

mydata$Age[is.na(mydata$Age)] = mean(mydata$Age, na.rm=TRUE)

The Age column contains missing data as we saw in the data summary on the previous page. The above code is used to fix this problem. First we use the [] subset operator to find out which values in the Age column are NA's. Once we have located the NA's we replace them with the mean of the values that are not NA's. This ensures that the new data is the correct representation of the original data and it reduces the errors during analysis.

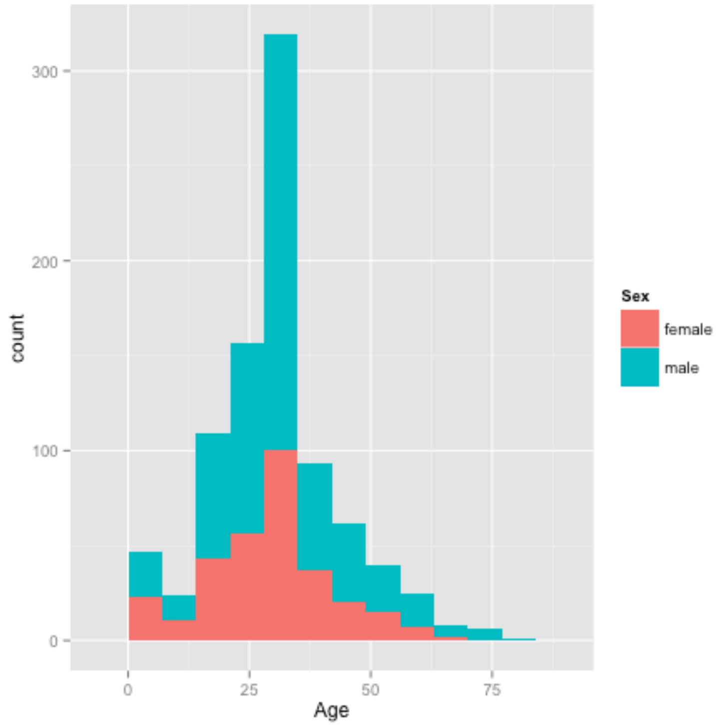
Data Standardization

Data standardization is a process in which data attributes within a data model are organized to increase the cohesion of entity types. In other words, the goal of data standardization is to reduce and even eliminate data redundancy, an important consideration for application developers because it is incredibly difficult to store objects in a database that maintains the same information in several places.



This example shows how standardization works.

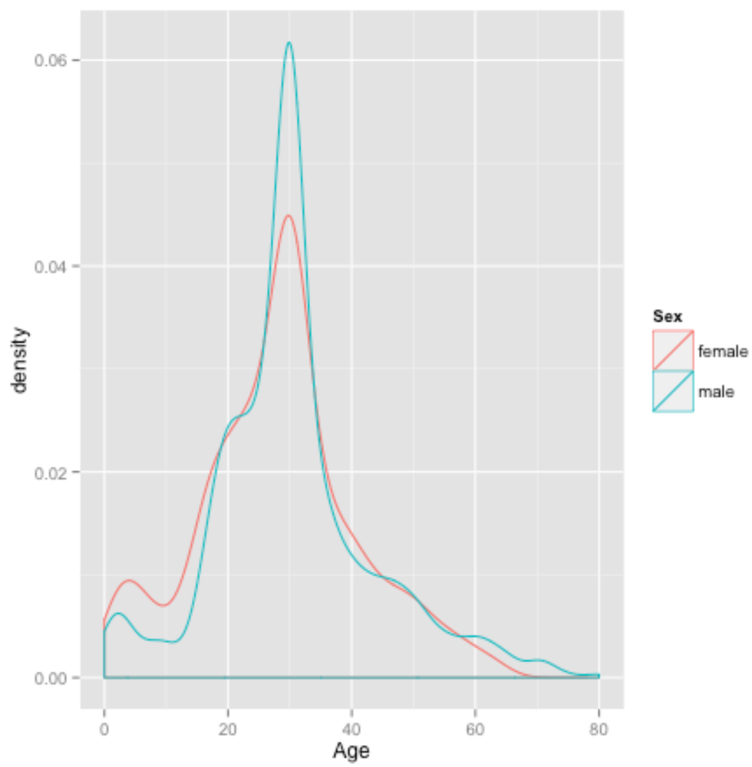
Data visualization

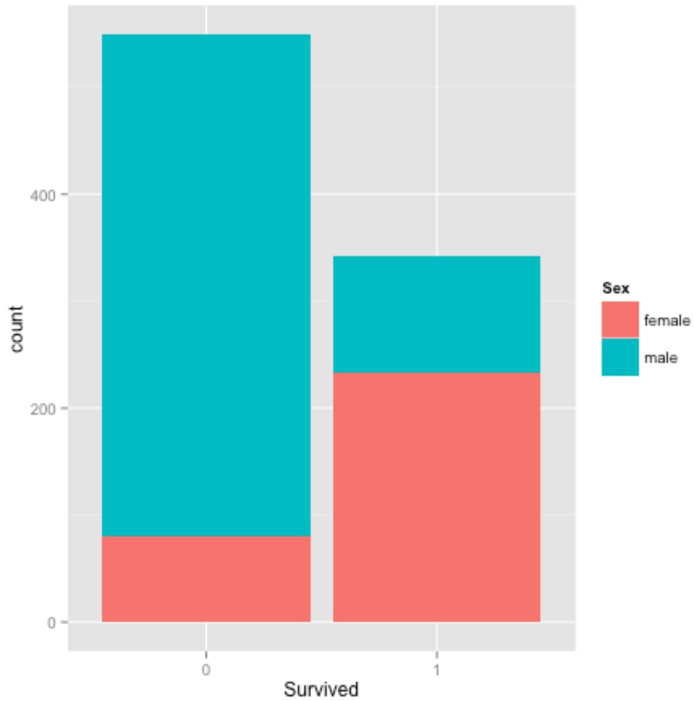


The age distribution of the genders is plotted on the left and we see that this distribution is fairly normal.

However, we notice that there are significantly more number of males as compared to the females sailing on that ill-fated day.

By plotting the line graphs of the same we can make out that the distributions are very similar i.e. both the males and the females constituted a major share between the ages of 20 to 40.

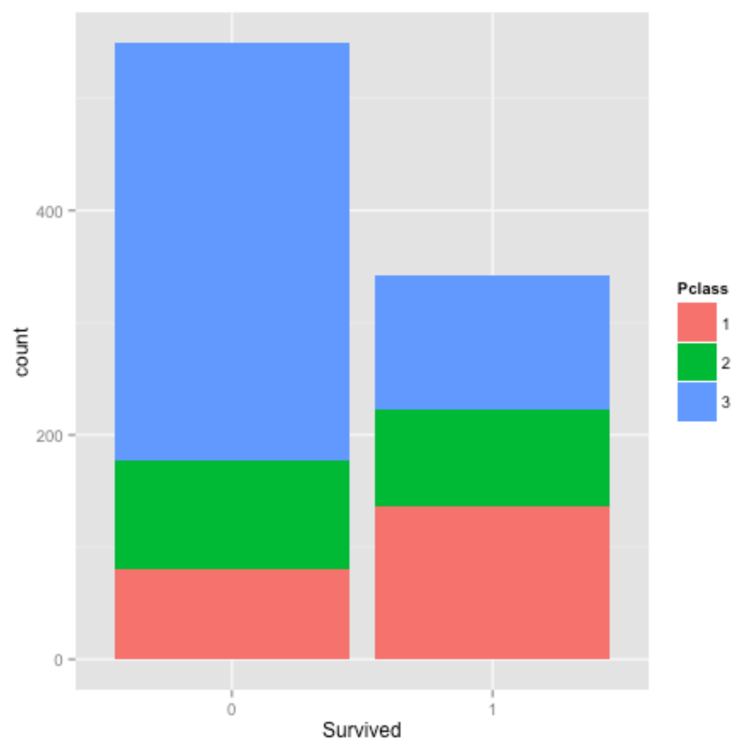


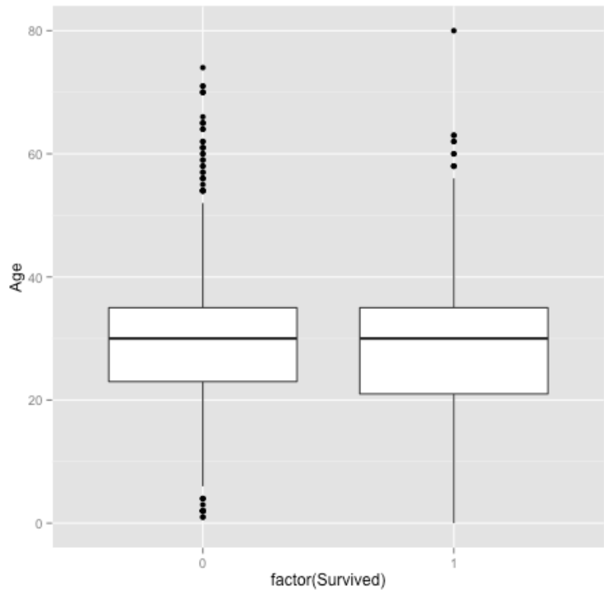


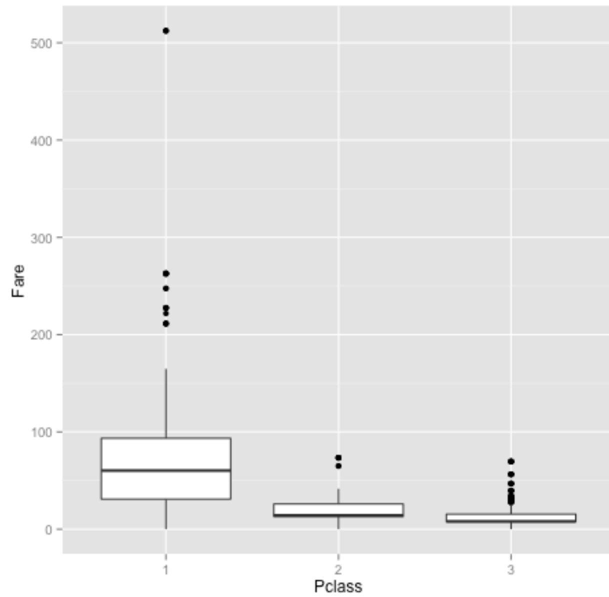
This graph represents the total number of males and females on the ship who survived. By initial examination of this graph one can make out that more number of female passengers survived than males. This could be mainly because of females being given the first preference to board lifeboats over men.

The graph on the right describes the survivors from different classes of the ship. 'Pclass' stands for Passenger Class, 1st Class represents the 'Upper class' and 2nd and 3rd being the middle and lower classes respectively.

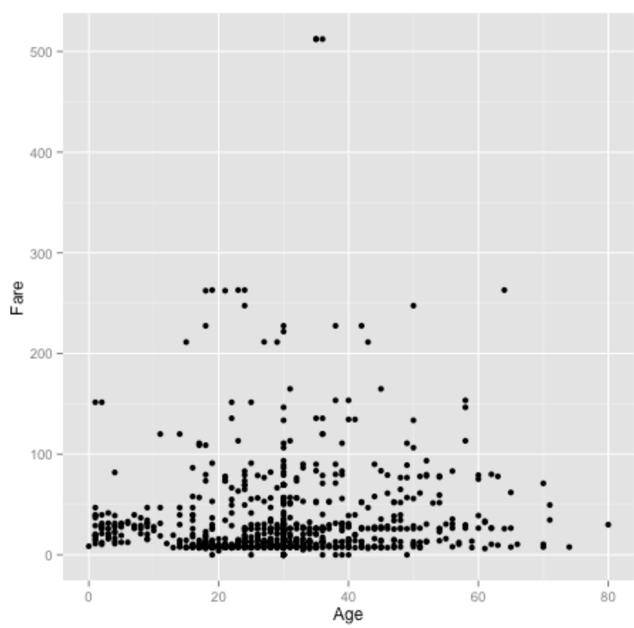
We can see that the passengers from the Upper class were more likely to have survived because they must have been given higher preference over other classes.







The graph on the left shows that age didn't play a major role for a person to have survived. The graph on the right shows that passengers travelling in the 1st or Upper class paid a significantly higher ticket fare than the other classes.



This graph shows us the relationship between the ticket price and the age of the passenger. The graph is very vague and scattered; this shows that the fare is not dependent on the age of the person. Therefore people paid the ticket price based on what class of ticket they purchased.

Steps performed

1. **Installing the necessary packages: -**
    a. caret -
    b. randomForest -
    c. klaR -
    d. plyr -
    e. reshape2 -
    f. ggplot2 –

2. **Fetching the required packages from the library**

3. **Loading the data in R from your working directory**
    a. The data titanicdata.csv is a windows comma separated value (csv) file that contains 12 variables and 891 observations.

4. **Making the necessary conversions to make the data more usable**
    a. Convert the Name variable to character type
    b. Make the Survived variable a factor with two levels i.e. 0 and 1.
    c. Make the Pclass variable a factor with three levels i.e. 1,2 and 3, representing different classes of passengers.

5. **Cleaning the data**
    a. Treating missing data - Assigning the mean of the Ages to NA values
    b. Rounding off Age to the nearest decimal

6. **Developing various plots for exploratory analysis.**

7. **Develop a standardization function that will help to standardize certain non-standardized variables.**

8. **Use the Standardization Function to standardize: -**
    a. Fare
    b. Age

9. **Create a cleaner dataset for analysis**

10. **Creating a Training and Test dataset with 70% being the training data and 30% being the test data.** (Note: Both Training and Test data should wholly represent the original dataset.)

11. Building models

a. **Model 1 –** using *randomForest* function. It implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression. It can also be used in unsupervised mode for assessing proximities among data points.

b. **Model 2 –** using *NaiveBayes* function. Computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule.

c. **Model 3 and 4 –** using *glm* function. glm is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution.

12. Predicting values

a. Predicted values are of factor type. If they come out as numeric as in Model 3 and 4 they need to be converted to factor type. For that purpose the *cut* function is used and the levels are set to match the required values.

13. Creating the actual confusion Matrix based on predicted values.

14. **Interpreting results.**

Results

Model 1

```
> confusionMatrix(p1, data_test$Survived)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 145  28
         1  19  74

               Accuracy : 0.8233
                 95% CI : (0.7721, 0.8672)
    No Information Rate : 0.6165
    P-Value [Acc > NIR] : 1.974e-13

                  Kappa : 0.62
 Mcnemar's Test P-Value : 0.2432

            Sensitivity : 0.8841
            Specificity : 0.7255
         Pos Pred Value : 0.8382
         Neg Pred Value : 0.7957
             Prevalence : 0.6165
         Detection Rate : 0.5451
   Detection Prevalence : 0.6504
      Balanced Accuracy : 0.8048

       'Positive' Class : 0
```

Model 2

```
> confusionMatrix(p2$class, data_test$Survived)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 137  44
         1  27  58

               Accuracy : 0.7331
                 95% CI : (0.6756, 0.7853)
    No Information Rate : 0.6165
    P-Value [Acc > NIR] : 4.134e-05

                  Kappa : 0.4171
 Mcnemar's Test P-Value : 0.05758

            Sensitivity : 0.8354
            Specificity : 0.5686
         Pos Pred Value : 0.7569
         Neg Pred Value : 0.6824
             Prevalence : 0.6165
         Detection Rate : 0.5150
   Detection Prevalence : 0.6805
      Balanced Accuracy : 0.7020

       'Positive' Class : 0
```

Depending on the predicted values from various models we have created a confusion matrix for each model.

Note: -

Sensitivity: is the proportion of actual positive cases that are correctly identified.

Specificity : is the proportion of actual negative cases that are correctly identified.

As we can see that ***Model 1*** is highly accurate i.e. 82.33% and seems to be the best of the models that can be used to predict the whether the passengers survived or not. ***Model 3*** is the next best with 79.32%, followed by ***Model 2*** at 73.31% and the least impressive model in this list is ***Model 4*** with 63.13% accuracy.

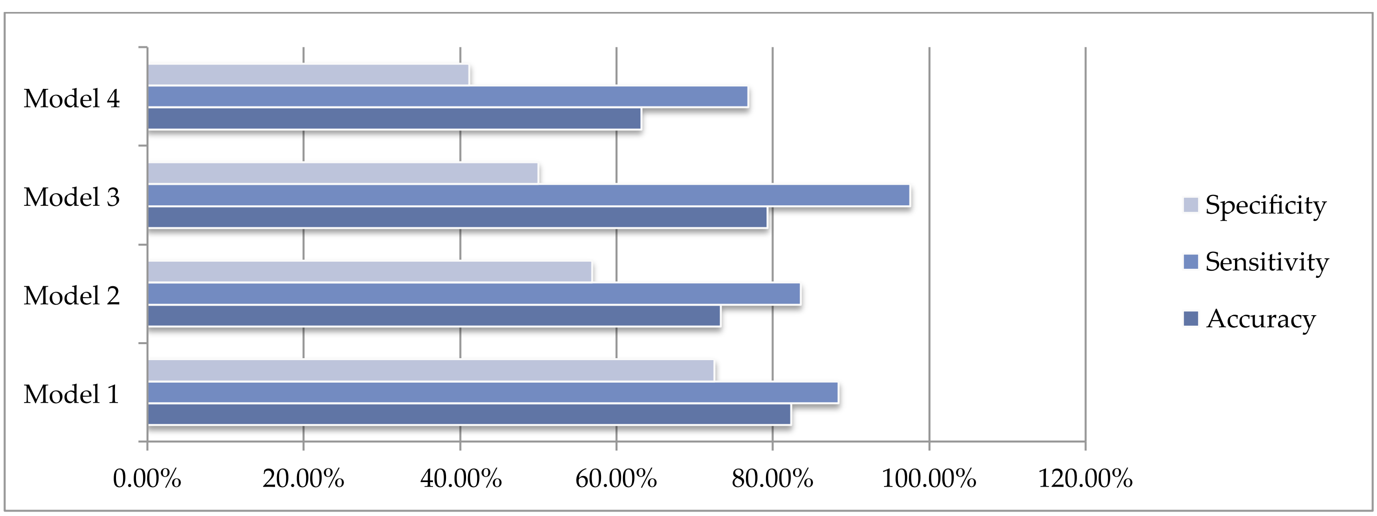
Model 3

```
> confusionMatrix(p3_cut, data_test$Survived)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160  51
         1   4  51

               Accuracy : 0.7932
                 95% CI : (0.7395, 0.8403)
    No Information Rate : 0.6165
    P-Value [Acc > NIR] : 4.694e-10

                  Kappa : 0.521
 Mcnemar's Test P-Value : 5.552e-10

            Sensitivity : 0.9756
            Specificity : 0.5000
         Pos Pred Value : 0.7583
         Neg Pred Value : 0.9273
             Prevalence : 0.6165
         Detection Rate : 0.6015
   Detection Prevalence : 0.7932
      Balanced Accuracy : 0.7378

       'Positive' Class : 0
```

Model 4

```
> confusionMatrix(p4_cut, data_test$Survived)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126  60
         1  38  42

               Accuracy : 0.6316
                 95% CI : (0.5705, 0.6897)
    No Information Rate : 0.6165
    P-Value [Acc > NIR] : 0.33095

                  Kappa : 0.1877
 Mcnemar's Test P-Value : 0.03389

            Sensitivity : 0.7683
            Specificity : 0.4118
         Pos Pred Value : 0.6774
         Neg Pred Value : 0.5250
             Prevalence : 0.6165
         Detection Rate : 0.4737
   Detection Prevalence : 0.6992
      Balanced Accuracy : 0.5900

       'Positive' Class : 0
```

Here is a simple chart showing the comparison of the accuracy of these four models.



Conclusion

One of the reasons that the shipwreck lead to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Considering a hypothetical situation, if we had this data beforehand could we have predicted which passenger is more likely to survive or die? The answer is Yes. Based on the analysis carried out we can say that survival of the passenger can be predicted up to 82.33% accuracy based on the model developed above. However other n' number of models can be developed but in our case we have chosen four models and we have put them to the test.

To conclude I would like to say that R as a programming language for analytics is very powerful and gives immense flexibility to the coder. It helped me to build models far easier than I would have in other languages.

References

1. www.inside-r.org (Information on R packages / code help)

2. www.cran.r-project.org (Information on R packages)

3. en.wikipedia.org (Theoretical Information)

4. images.google.com (Explanatory images)