Linear Statistical Analysis Project Report
Fall 2020

MOHAMMED SHOEBUDDIN HABEEB
002521737

Predicting the Odds of Survival aboard the Titanic using Generalized Linear Model

INTRODUCTION:

The sinking of the Titanic is one of the most infamous shipwrecks in history. This resulted in the death of 1502 out of 2224 passengers and crew. We observe that there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. I would like to build a Generalized Linear Model - Logistic Regression model to predict the probability that someone lived on the titanic using passenger data (i.e. name, age, gender, socio-economic class, etc.).

In this paper, after all the assumptions are met, we build a Generalized Linear model, i.e. Logistic Regression and interpret the estimated coefficients in terms of the predicted survival odds. Create confidence intervals for the coefficient estimates and the coefficient standard errors. Now I would like to estimate the proportion of incorrectly predicted outcomes with the correctly predicted outcomes using Receiver Operating Characteristic curve to see a graphical display of the predictive ability of the model

DATA:

The Titanic datasets are obtained from Kaggle (https://www.kaggle.com/c/titanic/data). there are originally three datasets form the website which autometically splitted into three portions; for one being the trained data include the survival outcome (binary 0: died or 1: survived value), another one being the tested data without the survival outcome, and a third one only contained the actual survival outcome of passengers listed in the tested data.

This study used the datasets to make prediction on the survival outcome of passengers in the tested data with a model built from the trained dataset. Since this is a binary outcome prediction, the logistic regression analysis will be used to model.

EXPLORING AND DATA PREPARATION:

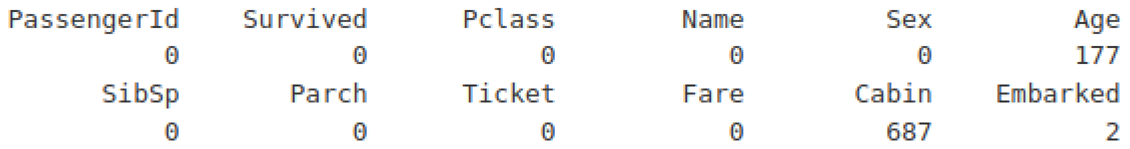
Since the datasets are given seperately as trained and tested data, they will be kept as it is. The thing that needed to be done is to merge the actual survival outcome of passengers from tested data with other information in that dataset. The column of survival outcome (dependent variable) is merged with the rest of the independent variables/features of the passegers from the tested dataset by passengerId. The trained dataset contains 891 observations (passenger information) and 12 features (information of passengers), and the tested dataset contains 418 observations.

The Pcalss (passenger class) feature are converted into factor in both dataset because the value of 1,2,3 should not be treated as numerical but as category levels for analysis later.

While the Kaggle website describes information of the dataset, and mentioned that 'age' that are less then 1 is recorded as fractional numbers. The observations with age value less than 1 is then assess here out of curiosity for both dataset, and it looks like the fractions indicates the months-old of infants out of 12 months.

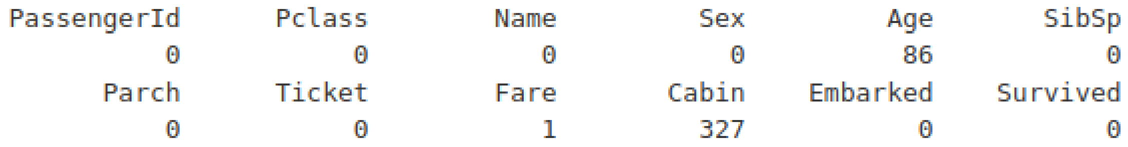
I use sapply() function to count the number of observations with each feature that contains 'NA'. There are many missing age and Cabine values in both dataset, while theere are 2 values missing in the Embarked feature in the trained data, and 1 value missing in the Fare feature.

Train Set :



```
PassengerId    Survived      Pclass        Name         Sex         Age
          0           0           0           0           0         177
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0         687           2
```

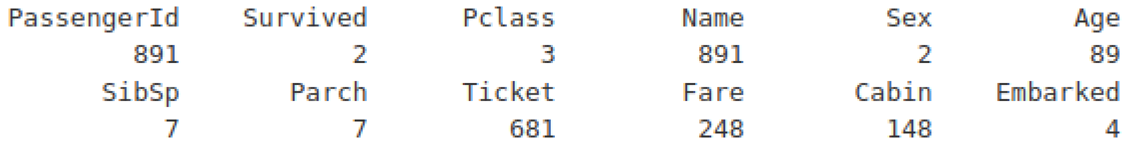
Test Set :



```
PassengerId      Pclass        Name         Sex         Age       SibSp
          0           0           0           0          86           0
      Parch      Ticket        Fare       Cabin    Embarked    Survived
          0           0           1         327           0           0
```

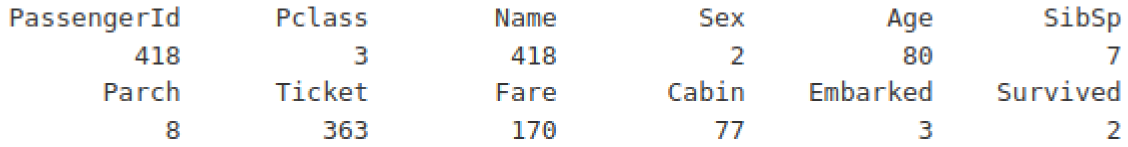
Similarly, the number of unique observations per column is revealed below.

Train



```
PassengerId    Survived      Pclass        Name         Sex         Age
        891           2           3         891           2          89
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          7           7         681         248         148           4
```

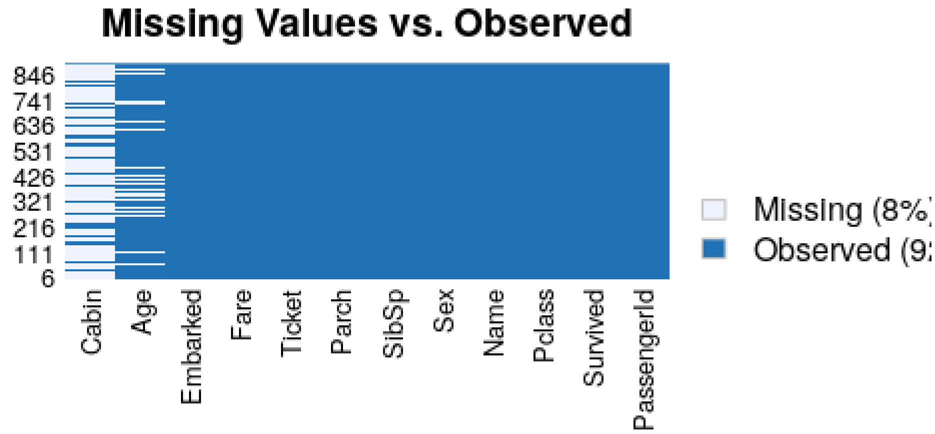
Test :



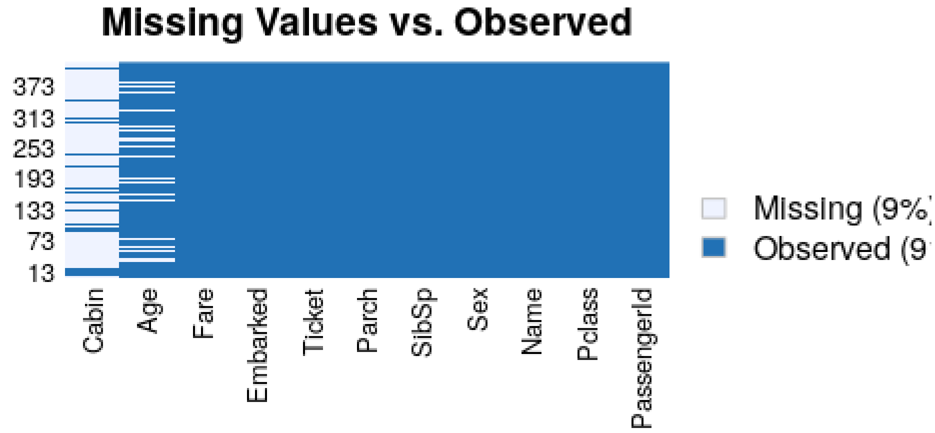
```
PassengerId      Pclass        Name         Sex         Age       SibSp
        418           3         418           2          80           7
      Parch      Ticket        Fare       Cabin    Embarked    Survived
          8         363         170          77           3           2
```

Using the missmap() function under the Amelia package, the visualization of the amount of missing and observed values per features is observed below. Most information in the Cabin and Age features are missing in both datasets.

Missing values in train set



Missing Values in Testing set



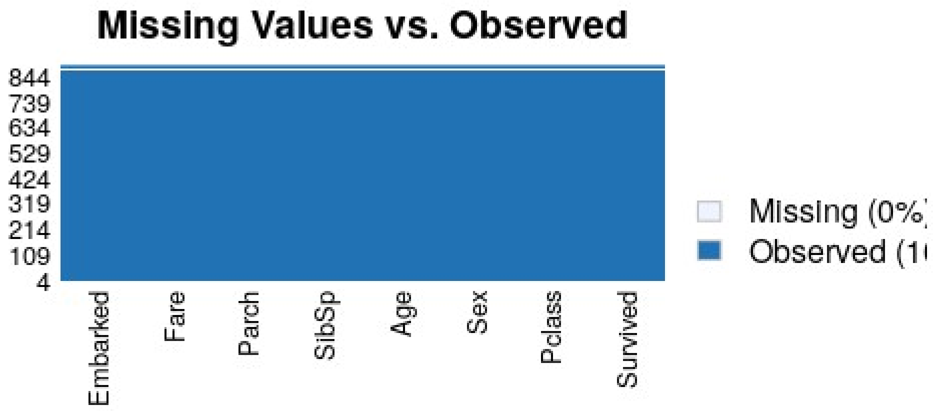
Features that contains many missing values, such as ‘Cabin’, and features that are assumed to be insignificant to predict survival, such as ‘Ticket’ and ‘Name,’ will be excluded. The features that are left for analysis are: Sex, Age, Pclass, Sibsp, Parch, Fare, Embarked and Survived.

To deal with the misssing value in Age feature, there are several way such as using mean, median or mode to fill in the missing value. Here the mean age of the Titanic population is used to applied on the age column that are missing in both datasets.

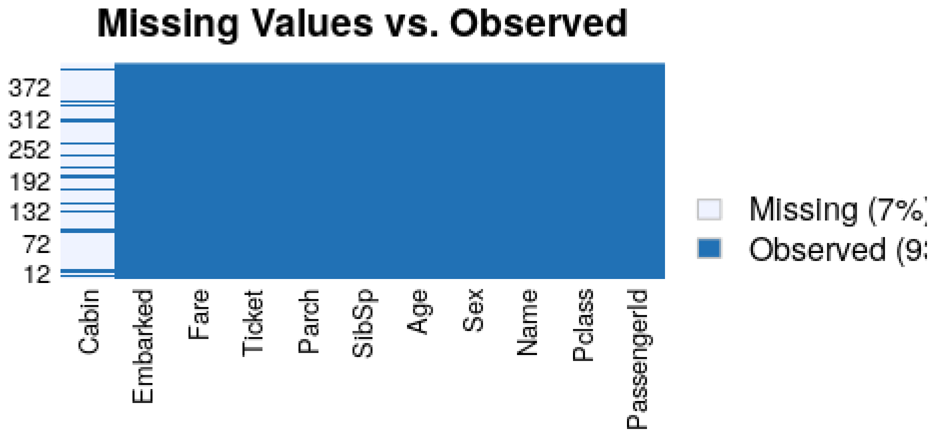
Since there’s only two missing observations in Embarked’ and one in ‘Fare’ features, those observations are removed assuming that should not lose too much information compared with the rest of the datasets.

Once modified these datasets are shown below with much improvement and missing values and we use missmap() again to to see this issue is resolved.

Training Dataset



Testing Dataset



Model Training for Data

Using the generalized linear model, glm() function, make a logistic regression analysis using 'Survived' feature as outcome, with the rest of features in the training dataset as independent predictors. Specified binomial(link = 'logit') in the family argument will analyze the data using logistic regression.

The output of the logistic regression object, reg.model, shows that catergorical features, 'Pcalss' and 'Sex', and the numerical features 'Age' and 'Sibsp' are significant features for predicitng survival outcome at alpha = 0.05 level. Th rest of the model coefficients suggests insignificant contribution in survival prediction. For example, increase one unit in age will decrease the log odd of survival by 0.039; being a male will decrease the log odd of survival by 2.7 compared to female; and being in class2 will decrease the log odd of survival by 0.92, being in class3 will decrease the log odd of survival by 2.15.

We try to find the summary of the model

We also try to understand the significant features and variables

```
> summary(reg.model)

Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = titanic.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6451  -0.5914  -0.4239   0.6237   2.4430

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.288732   0.565036   9.360  < 2e-16 ***
Pclass      -1.100128   0.143525  -7.665 1.79e-14 ***
Sexmale     -2.718565   0.200795 -13.539  < 2e-16 ***
Age         -0.039949   0.007854  -5.086 3.65e-07 ***
SibSp       -0.325914   0.109434  -2.978   0.0029 **
Parch       -0.093022   0.118728  -0.783   0.4333
Fare         0.001916   0.002376   0.806   0.4200
EmbarkedQ   -0.030989   0.381989  -0.081   0.9353
EmbarkedS   -0.419628   0.236813  -1.772   0.0764 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.82  on 888  degrees of freedom
Residual deviance:  784.12  on 880  degrees of freedom
AIC: 802.12

Number of Fisher Scoring iterations: 5
```

The 95% confident interval for each predictor’s coefficient in the logistic regression model is also shown below.

```
> confint(reg.model)
Waiting for profiling to be done...
                   2.5 %       97.5 %
(Intercept)  4.201129910  6.419738896
Pclass      -1.384525784 -0.820671511
Sexmale     -3.121432392 -2.333342196
Age         -0.055637481 -0.024807709
SibSp       -0.551160082 -0.121903094
Parch       -0.332056669  0.136817379
Fare        -0.002494639  0.007069559
EmbarkedQ   -0.783935885  0.715600907
EmbarkedS   -0.883498379  0.046082606
```

We see that Parch, Embarked and Fare are insignificant so we remove them to fit for stepwise calibration. Using stepwise procedure with interactions and then removing the insignificant terms, we obtain the following model:

```
glm(formula = Survived ~ Sex + Pclass + Age + SibSp + Sex:Pclass + 
    Pclass:SibSp + Pclass:Age + Sex:Age, family = "binomial", 
    data = titanic.train)
```

```
> summary(reg.model)

Call:
glm(formula = Survived ~ Sex + Pclass + Age + SibSp + Sex:Pclass + 
    Pclass:SibSp + Pclass:Age + Sex:Age, family = "binomial", 
    data = titanic.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.9058  -0.6813  -0.4192   0.4915   2.6186  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)     4.33125    1.32466   3.270 0.001077 ** 
Sexmale        -3.96515    1.14054  -3.477 0.000508 ***
Pclass         -0.99595    0.45766  -2.176 0.029542 *  
Age             0.04442    0.03161   1.405 0.159948    
SibSp           0.78122    0.46681   1.674 0.094227 .  
Sexmale:Pclass  1.00206    0.34081   2.940 0.003280 ** 
Pclass:SibSp   -0.43359    0.17019  -2.548 0.010845 *  
Pclass:Age     -0.02699    0.01088  -2.480 0.013127 *  
Sexmale:Age    -0.04746    0.01938  -2.449 0.014311 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.82  on 888  degrees of freedom
Residual deviance:  753.88  on 880  degrees of freedom
AIC: 771.88

Number of Fisher Scoring iterations: 6
```

For an easier interpretation, we can transform these values into odd’s ratios

```
Coefficients:
   (Intercept)        Sexmale         Pclass            Age          SibSp  Sexmale:Pclass    Pclass:SibSp  
Pclass:Age     Sexmale:Age  
       4.33125       -3.96515       -0.99595        0.04442        0.78122        1.00206 -0.43359  
-0.02699       -0.04746  
```

Linear contrast can be done using the wald.test() function in the aod package for the model object. For example, the first one is comparing the ‘Age’ feature coefficients from the model, which is the all three level including the reference level. The p-value = 0.1 indicates statistical insignificant difference for the levels of the ‘Age’ feature. The second one is comparing the statistical significance of ‘Pclass’ and its p-value indicates that the passenger classes are significantly difference.

Goodness of Fit:

The ANOVA table is created by adding the terms of the model sequentially.
Since the residual deviance of the model decreases with each added predictor variable along with the fact that the p-values are significant, there is evidence that our fitted model is a good fit.
Cooks distances for the data are created, yet none of them are significantly large. This indicates that there are no influential points.
We can also perform Wald Tests on each of the predictors to check and see if they are needed in the model.

```
Wald test:
----------

Chi-squared test:
X2 = 4.5, df = 2, P(> X2) = 0.1
```

```
> wald.test(b = coef(reg.model), Sigma = vcov(reg.model), Terms = 3:4)
Wald test:
----------

Chi-squared test:
X2 = 21.2, df = 2, P(> X2) = 2.5e-05
```

Moreover, exponentiate the model coefficients can look at the result and interpret its meaning at a different angle. Below are table of the “odd ratio” value for each predictor coefficient relative to the survival and their respective 95% confident interval odd ratio value.

95% confidence intervals for the odds ratios are as follows

```
                     OR       2.5 %       97.5 %
(Intercept)    76.0394519 6.560069786 1204.1012821
Sexmale         0.0189651 0.001747379    0.1571769
Pclass          0.3693706 0.144553814    0.8753543
Age             1.0454221 0.982575886    1.1124213
SibSp           2.1841248 0.877675525    5.5414461
Sexmale:Pclass  2.7238913 1.450106016    5.5725297
Pclass:SibSp    0.6481791 0.461272849    0.9022472
Pclass:Age      0.9733752 0.952731706    0.9943152
Sexmale:Age     0.9536474 0.917866878    0.9904616
```

The anova() function for the model object allows to see the null and residuals deviances. The difference between these two deviances shows how well the model is performing against the null deviance. The residuals deviance column allows to see the drop of deviance value by additional respective predictor term added. The table shows that adding Pcalssm Sex, age and Sibsp has significantly reduce the residuals deviances while the other terms aren't anymore at alpha = 0.05 level.

```
> anova(reg.model, test = 'Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: Survived

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           888    1182.82
Sex           1  266.212       887     916.61 < 2.2e-16 ***
Pclass        1   89.782       886     826.82 < 2.2e-16 ***
Age           1   22.121       885     804.70  2.56e-06 ***
SibSp         1   14.434       884     790.27 0.0001452 ***
Sex:Pclass    1   22.511       883     767.76  2.09e-06 ***
Pclass:SibSp  1    5.530       882     762.23 0.0186959 *
Pclass:Age    1    2.317       881     759.91 0.1279832
Sex:Age       1    6.031       880     753.88 0.0140554 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Furthermore, using teh pR2() function in the pscl package allows to see a linear regression R-square value equivelent, which is the McFadden R-square index. This is equivelently saying that the logsitic regresion model has well explained 36% of variation in the survival prediction.

Since the residual deviance of the model decreases with each added predictor variable along with the fact that the p-values are significant, there is evidence that our fitted model is a good fit. Cooks distances for the data are created, yet none of them are significantly large. This indicates that there are no influential points.

Like the results before, these p-values indicate that each of the predictor variables are significant in predicting the odds that a passenger on the Titanic survives or does not survive.
Lastly, we can use the Hosmer-Lemeshow Goodness of Fit Test to determine model adequacy.

```
> hoslem.test(reg.model$y,fitted(reg.model),g=10)

	Hosmer and Lemeshow goodness of fit (GOF) test
```

```
data:  reg.model$y, fitted(reg.model)
X-squared = 27.264, df = 8, p-value = 0.0006365
```

From this test we see that p-value is .0006365 so we can say that there is strong evidence that our model is a good fit and adequate for predicting survival based on Sex, Class, Pclass and SibSp

McFadden R2 value between 0.2 and 0.4 is considered good. Therefore, since our McFadden R2 is .3626 we can say that the model selected is an excellent fit for predicting survival.

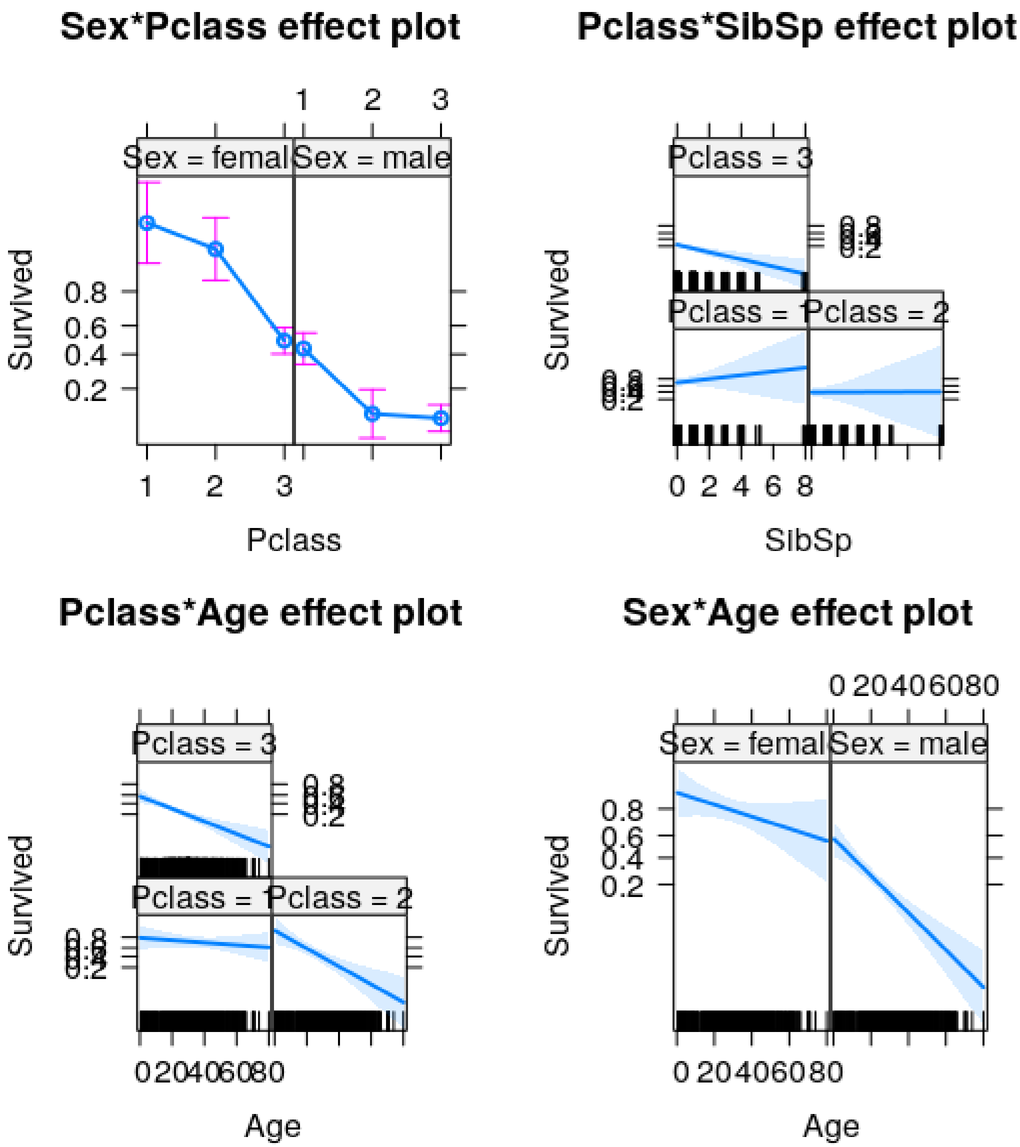
```
> pR2(reg.model)
fitting null model for pseudo-r2
              llh       llhNull          G2      McFadden         r2ML         r2CU
-376.9406259 -591.4088880  428.9365242    0.3626396    0.3827575    0.5202942
```

Collinearity :

After assessing the goodness of fit of the logistic model, we will check to see if there is any collinearity between the predictor variables. We will check this using variance inflation factors.

```
> vif(reg.model)
              Sex      Pclass         Age       SibSp   Sex:Pclass Pclass:SibSp   Pclass:Age      Sex:Age
      35.62102     17.22230     18.97842     19.39245     21.40169     20.58712     12.59078
13.33962
```

Effect :

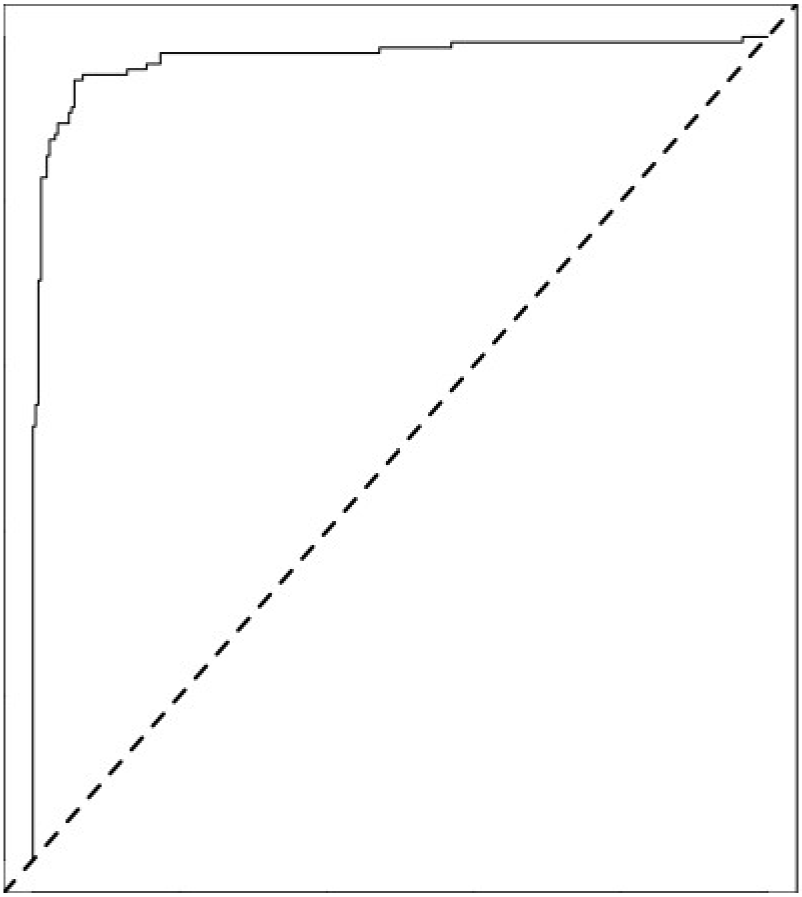


Model Evaluation :

We evaluate the model on prediction set to get the following accuracy

```
> print(paste('Accuracy', 1 - misClassifiError))
[1] "Accuracy 0.937649880095923"
```

The Receiver Operating Characteristic (ROC) curve is plotted below for false positive rate (FPR) in the x-axis vs. the true positive rate (TPR) in the y-axis. It shows the detection of true positive while avoiding the false positive. This is the same as measuring the unspecificity (1 - specificity) in x-axis, against the sensitivity in y-axis. This ROC curve in particular shows that its very closed to the perfect classifier meaning that its better at identifying the positive values. An index for that is the AUC (area under the curve) of this ROC, which is 0.975 for this case



CONCLUSION:

The initial model proposed was not representative of the data given which was shown in the initial modeling. Once we Stepwise formulate we get a modified model which has better results and gives convincing representation of predictions with an accuracy of 94%.

However, even from the logistic regression model, we can easily see that the Titanic survival outcome is highly depended on several predictors, such as sex, age and passenger class. In particular, female are more likely to survived than male while keeping other predictors conditions constant, older people are less likely to survived while keeping other predictors conditions constant; and lastly, people from a lower class are less likely to survived keeping other predictors conditions constant.