A REPORT ON TITANIC DATASET USING LOGISTIC REGRESSION

Factors Deciding Survival Of Passengers

NOVEMBER 20, 2019

NITIN AGGARWAL - 100328731
JINISH KANPARA - 100330202
MAYANK BANSAL- 100328735

Table of Contents

About The Dataset:

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 342 out of 891 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “What sort of people were more likely to survive?” using passenger data that have following variables:

Sr. No.	Variable Name	Description Of variable	Type of Variable
1	Passenger ID	Unique Id for each passenger	Categorical
2	Survived	Whether Passenger Survived (1) or Died (0)	Categorical
3	Pclass	Socio-Economic Classes	Categorical
4	Name	Name of Passenger	Categorical
5	Sex	Gender	Categorical
6	Age	Age of Passenger	Quantitative
7	SibSp	Siblings or spouse with passenger	Categorical
8	Parch	Parents or Children with passenger	Categorical

9	Ticket	Ticket Number	Categorical
10	Fare	Price of Ticket	Quantitative
11	Cabin	Cabin allotted to passenger	Categorical
12	Embarked	Place for boarding (Q-Queenstown, S- Southampton, C- Cherbourg)	Categorical

Fig 1: Table of variables

As our response variable 'Survived' is dichotomous in nature, so we are performing logistic regression in R language on this dataset. Firstly, we have loaded our dataset into the R and done a preliminary analysis to know our dataset and we found some of the missing values in the variable Age. Then we filled the values in the Age column by the median of Age relative to our categorical variable Pclass so that it will be as close to actual values.

Secondly, after analyzing we merged 2 columns SibSp and Parch to create a new categorical variable named 'Alone' with two categories 'Yes' and 'No'. We have done so as probability of survival increases if passenger is with someone.

Thirdly, we have found median fare according to Pclass to find out correct levels of Pclass as we want to give maximum weight to that Pclass which has maximum Fare. We found out that Pclass 1 has maximum Fare and Pclass – 3 has the lowest fare for which we have recoded the levels of Pclass by interchanging the levels of 1 and 3 in the dataset, by treating it as an ordinal variable. After cleaning the data, we dropped the columns passenger id, ticket, names because these variables have no significance in determining whether a person will survive or not and further, we dropped Sibsp and Parch from the dataset because we have made a calculated variable 'Alone' by merging these 2 variables. At last, we are left with 5 independent variables and one response variable -Survived in our dataset.

Data Cleaning And Manipulation:

After loading the dataset, in our preliminary analysis, we found 177 missing values in Age column. We filled the missing values of Age by taking the median of the remaining

```
> #Checking total umber of null values in our dataset
> colSums(is.na(titanic))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket
          0           0           0           0           0           0           0           0           0
       Fare       Cabin    Embarked
          0           0           0
> #Checking median of Age according to Each Pclass
> titanic %>%
+     group_by(Pclass) %>%
+     summarize(median_age = median(Age,na.rm = TRUE))
# A tibble: 3 x 2
  Pclass median_age
   <int>      <dbl>
1      1         37
2      2         29
3      3         24
>
> #Filling missing values of Age with median value of age according to the Pclass
>
> titanic[c("Age","Pclass")]<- titanic[c("Age","Pclass")] %>%
+     mutate(Age = ifelse(is.na(Age) & Pclass==1,37 ,Age)) %>%
+     mutate(Age = ifelse(is.na(Age) & Pclass==2,29 ,Age)) %>%
+     mutate(Age = ifelse(is.na(Age) & Pclass==3,24 ,Age))
> colSums(is.na(titanic))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket
          0           0           0           0           0           0           0           0           0
       Fare       Cabin    Embarked
          0           0           0
```

Age column grouped according to the Pclass (screenshot attached below).

Fig 2: Filling age null values by taking median of Age grouped according to Pclass

After filling null values in the Age column, we looked at our dataset more closely and found that SibSp (Siblings/Spouses) and Parch (Parents/children) columns can be merged into one to make it more interpretable. We derived a new categorical column namely, Alone with two levels, Yes and No.

```
> #Creating a new calculated variable group to know how many people were actually present alongwith that person
>
> titanic$Alone<-  titanic$SibSp + titanic$Parch
> titanic <- titanic %>% mutate(Alone= ifelse(Alone==0,"Yes", "No"))
> titanic$Alone <- factor(titanic$Alone, levels = c("No","Yes"))
> head(titanic$Alone)
[1] No  No  Yes No  Yes Yes
Levels: No Yes
```

Fig 3: Creating Calculated variable Alone from SibSp and Parch

In our preliminary analysis, we found that there are three classes in the variable Pclass with levels 1, 2 and 3. According to the description of the variable, the Pclass should be treated as ordinal data because each level signifies the Passenger's class. For eg. Business class > Premium Economy class > Economy class. To prove this, we calculated the median fare of each class and concluded that Pclass 1 is the most expensive followed by Pclass 2 and then 3. Therefore, we recoded the Pclass variable by assigning the value of 3 to 1 and vice versa as can be seen in the screenshot below.

```
> #median fare according to each Pclass
>
> titanic %>%
+     group_by(Pclass) %>%
+     summarize(median_fare = median(Fare, na.rm = TRUE))
# A tibble: 3 x 2
  Pclass median_fare
   <int>       <dbl>
1      1       60.3
2      2       14.2
3      3        8.05
>
> #Treating Pclass into ordinal data and assigning the values 3 to pclass 1 as its weightage(meadian fare) is more
 and 1 to Pclass 3 as its median fare is less.
>
> titanic$Pclass<- as.character(titanic$Pclass)
> titanic$Pclass<- as.numeric(recode(titanic$Pclass, "'1'= '3';'2'='2'; '3'='1'"))
> head(titanic)
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch
1           1        0      1                                     Braund, Mr. Owen Harris   male  22     1     0
2           2        1      3 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3           3        1      1                                      Heikkinen, Miss. Laina female  26     0     0
4           4        1      3        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5           5        0      1                                    Allen, Mr. William Henry   male  35     0     0
6           6        0      1                                            Moran, Mr. James   male  24     0     0
            Ticket    Fare Cabin Embarked Alone
1        A/5 21171  7.2500              S    No
2         PC 17599 71.2833   C85        C    No
3 STON/O2. 3101282  7.9250              S   Yes
4           113803 53.1000  C123        S    No
5           373450  8.0500              S   Yes
6           330877  8.4583              Q   Yes
>
```

Fig 4: Making Pclass a Quantitative Variable by giving Weightage according to Status

Lastly, we dropped the unwanted columns from our dataset such as PassengerId, Name, SibSp, Parch, Ticket, Cabin. We have choses=n to drop these columns because of the following reasons:

1. **PassengerId:** It does not affect if a person will survive or not.
2. **Name:** The survival of the person will also is independent of the name of the passenger.

3. **SibSp and Parch:** We have derived a new column, Alone which is a combination of these two columns. Therefore, dropping these columns will prevent us from multicollinearity.
4. **Ticket:** Ticket number is not an important criterion to decide if that person would survive or not.
5. **Cabin:** Even the Cabin number does not define a person's survival.

After dropping these columns, we are left with Pclass, Age, Fare, Sex, Embarked and Alone columns.

```
> #Dropping unwanted columns
>
> titanic<- titanic[,!names(titanic)%in%c("PassengerId","Ticket","Name","SibSp", "Parch","Cabin")]
> str(titanic)
'data.frame':   891 obs. of  7 variables:
 $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass  : num  1 3 1 3 1 1 3 1 1 2 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age     : num  22 38 26 35 35 24 54 2 27 14 ...
 $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
 $ Alone   : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 2 1 1 1 ...
> head(titanic)
  Survived Pclass    Sex Age    Fare Embarked Alone
1        0      1   male  22  7.2500        S    No
2        1      3 female  38 71.2833        C    No
3        1      1 female  26  7.9250        S   Yes
4        1      3 female  35 53.1000        S    No
5        0      1   male  35  8.0500        S   Yes
6        0      1   male  24  8.4583        Q   Yes
```

Fig 5: Deleting Non-Significant Variables

Significance Test:

1. Variance Test
2. 2 Sample T-Test
3. Chi-square Test

Variance Test:

1. **Age and Survived:** To check the association between Quantitative Variables (Age) and categorical response variable (Survived), first, I performed variance test to check whether the variances of two groups (Survived or Dead) are equal or not. Upon testing, I found the positive result for Age variable.
    - **Null Hypothesis:** Variance of two groups (Survived or Dead) according to age is equal
    - **Alternate Hypothesis:** Variance of two groups (Survived or Dead) according to age is not equal.

```
> #Variance and 2 sample t-test of Age
> #Variances are same for both the groups
>
> a<-subset(titanic, Survived==0)$Age
> b<-subset(titanic, Survived==1)$Age
> var.test(a,b)

        F test to compare two variances

data:  a and b
F = 0.83995, num df = 548, denom df = 341, p-value = 0.07097
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6921009 1.0149908
sample estimates:
ratio of variances
         0.8399524
```

Fig 6: Performing variance test on Age vs Survived

P-value=0.07 > level of significance=0.05 that is why we fail to reject the null hypothesis and conclude that the variance of 2 groups according to age are equal.

2. **Fare and Survived:** To check the association between Quantitative Variables (Fare) and categorical response variable (Survived), first, I performed variance test to check whether the variances of two groups (Survived or Dead) are equal or not. Upon testing, I found the negative result for the Fare variable.
    - **Null Hypothesis:** Variance of two groups (Survived or Dead) according to fare is equal
    - **Alternate Hypothesis:** Variance of two groups (Survived or Dead) according to fare is not equal.

```
> a<-subset(titanic, Survived==0)$Fare
> b<-subset(titanic, Survived==1)$Fare
> var.test(a,b)

        F test to compare two variances

data:  a and b
F = 0.22214, num df = 548, denom df = 341, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1830368 0.2684300
sample estimates:
ratio of variances
         0.2221384
```

Fig 7: Performing variance test on fare vs Survived

P-value=2.2e-16 < level of significance=0.05 that is why we can reject the null hypothesis and conclude that the variance of 2 groups according to fare is not equal.

2 Sample T-test:

1. **Age and Survived** We have performed this test to check whether Quantitative variable (Age) influences the response variable (Survived or Dead).

    - **Null Hypothesis:** Means of two groups (Survived or Dead) according to age is equal.

- **Alternate Hypothesis:** Means of two groups (Survived or Dead) according to age is not equal.

P-value=0.1587 > level of significance=0.05 that is why we fail to reject the null hypothesis and conclude that the mean of 2 groups according to age is equal. Thus, we conclude that age does not influence the response variable.

```
> #Two sample t test says that the mean age of two groups are same
>
> t.test(a,b,var.equal = T)

        Two Sample t-test

data:  a and b
t = 1.4105, df = 889, p-value = 0.1587
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5034568  3.0758976
sample estimates:
mean of x mean of y
 29.56011  28.27389
```

Fig 8: Performing 2 sample t-test on Age vs Survived

2. **Fare and Survived** We have performed this test to check whether Quantitative variable (Fare) influences the response variable (Survived or Dead).

    - **Null Hypothesis:** Means of two groups (Survived or Dead) according to fare is equal.
    - **Alternate Hypothesis:** Means of two groups (Survived or Dead) according to fare is not equal.

P-value=2.699e-11 < level of significance=0.05 that is why we can reject the null hypothesis and conclude that the mean of 2 groups according for fare is not equal. Thus, we conclude that age does affect the response variable.

```
> t.test(a,b,var.equal = F)

        Welch Two Sample t-test

data:  a and b
t = -6.8391, df = 436.7, p-value = 2.699e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -33.82912 -18.72592
sample estimates:
mean of x mean of y
 22.11789  48.39541
```

Fig 9: Performing 2 sample t-test on fare vs Survived

Chi-square Test:

1. Pclass and Survived**:** We performed a chi-square test to check the association between Categorical Variables (Pclass) and the categorical response variable (Survived or Dead).
    - **Null Hypothesis:** Pclass and Survived variable are independent of each other in the population.
    - **Alternate Hypothesis:** Pclass and Survived variable are not independent of each other in the population.

```
> p.table<- table(titanic$Survived, titanic$Pclass)
> chisq.test(p.table)

        Pearson's Chi-squared test

data:  p.table
X-squared = 102.89, df = 2, p-value < 2.2e-16
```

Fig 10: Chi Square test on Pclass vs Survived

P-value=2.2e-16 < level of significance=0.05 that is why we reject the null hypothesis and conclude that variables Pclass and Survived are not independent of each other in the population means there is an association between both the variables.

2. Sex and Survived: We performed a chi-square test to check the association between Categorical Variable (Sex) and the categorical response variable (Survived or Dead).
    - **Null Hypothesis:** Sex and Survived variable are independent of each other in the population.
    - **Alternate Hypothesis:** Sex and Survived variable are not independent of each other in the population.

```
> sex.table<- table(titanic$Survived,titanic$Sex)
> sex.table

    female male
  0     81  468
  1    233  109
> chisq.test(sex.table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  sex.table
X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Fig 11: Chi Square test on Sex vs Survived

P-value=2.2e-16 < level of significance=0.05 that is why we reject the null hypothesis and conclude that variables Sex and Survived are not independent of each other in the population means there is an association between both the variables.

3. Alone and Survived: We performed a chi-square test to check the association between Categorical Variable (Alone) and the categorical response variable (Survived or Dead).
    - **Null Hypothesis:** Alone and Survived variable are independent of each other in the population.

- **Alternate Hypothesis:** Alone and Survived variable are not independent of each other in the population.

```
> aln.table<- table(titanic$Survived, titanic$Alone)
> aln.table

     No Yes
  0 175 374
  1 179 163
> chisq.test(aln.table)

        Pearson's Chi-squared test with Yates' continuity correction

data:  aln.table
X-squared = 36.001, df = 1, p-value = 1.973e-09
```

Fig 12: Chi Square test on Alone vs Survived

P-value=1.973e-09 < level of significance=0.05 that is why we reject the null hypothesis and conclude that variable Alone and Survived are not independent of each other in the population means there is an association between both the variables.

4. Embarked and Survived**:** We performed a chi-square test to check the association between Categorical Variable (Embarked) and the categorical response variable (Survived or Dead).
   - **Null Hypothesis:** Embarked and Survived variable are independent of each other in the population.
   - **Alternate Hypothesis:** Embarked and Survived variable are not independent of each other in the population.

```
> p.table<- table(titanic$Survived, titanic$Embarked)
> chisq.test(p.table)

        Pearson's Chi-squared test

data:  p.table
X-squared = 29.671, df = 3, p-value = 1.619e-06
```

Fig 13: Chi Square test on Embarked vs Survived

P-value=1.619e-06 < level of significance=0.05 that is why we reject the null hypothesis and conclude that variable Embarked and Survived are not independent of each other in the population means there is an association between both the variables.

Multicollinearity:

We have conducted the multicollinearity test to check whether independent variables are highly correlated with each other or not and we found that none of the variables are highly correlated with one another. GVIF^(1/(2*DF)) of each variable is <3.16. So, we can conclude that there is no multicollinearity between the variables.

```
> #-----------------------------MULTICOLLINEARITY TEST---------------------------------------
> #Since GVIF^(1/(2*Df)) is less than 3.16, we can say that there is no multicolinearity in our dataset
>
> vif(glm(Survived~., family = binomial(link = "logit"), data= titanic))
             GVIF Df GVIF^(1/(2*Df))
Pclass   1.936397  1        1.391545
Sex      1.192774  1        1.092142
Age      1.417721  1        1.190681
Fare     1.449538  1        1.203968
Embarked 1.174629  3        1.027188
Alone    1.222377  1        1.105612
```

Fig 14: Checking multicollinearity of variables in the model

Correlation:

Correlation is performed to check the strength of the relation between two quantitative variables. In our dataset, we are left with two quantitative variables, Age and Fare. We are checking the relation between them.

```
> pairs.panels(titanic[,!names(titanic)%in%c("Survived", "Pclass", "Sex","Alone")])
```

Fig 15: Command to graphically see the correlation between the variables

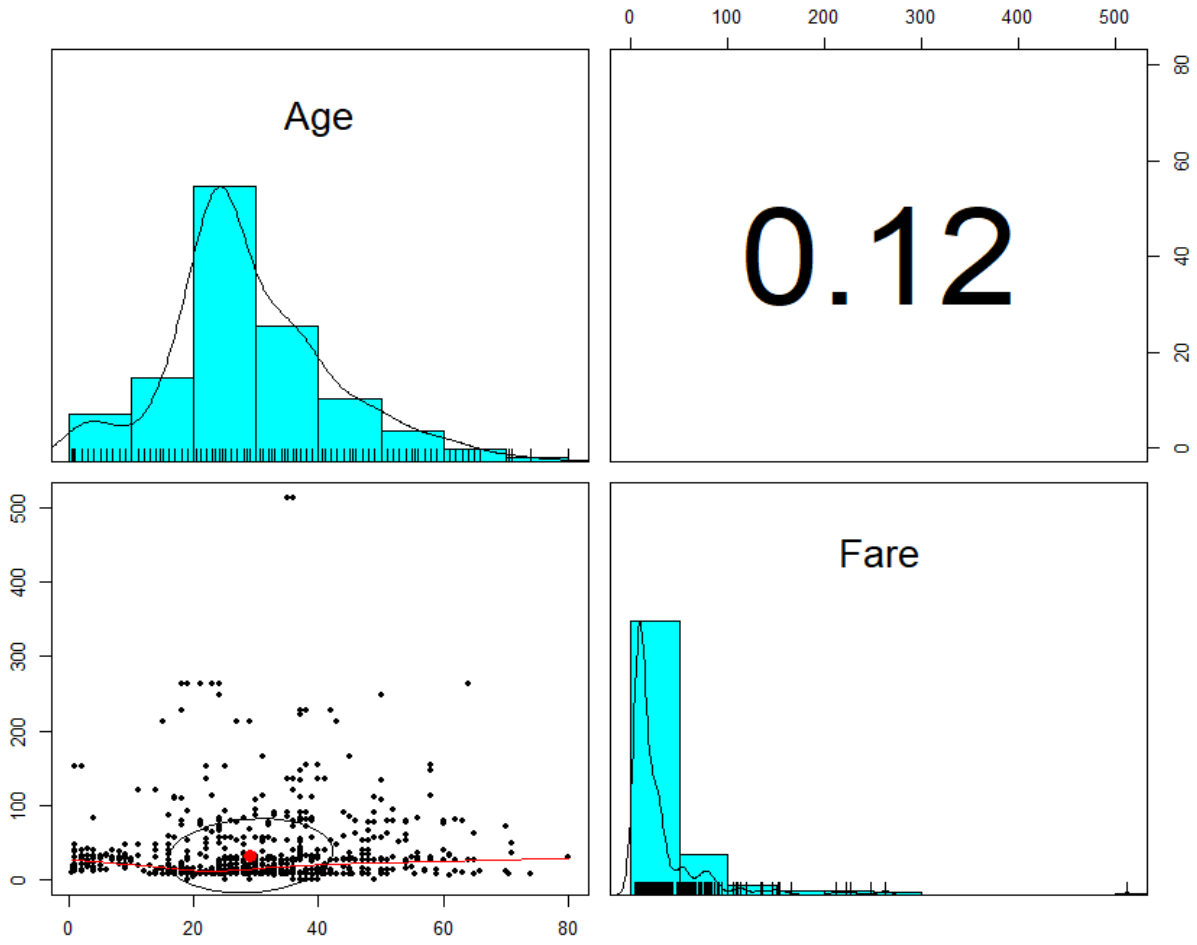

Fig 16: Checking correlation between quantitative variables

Interaction Terms:

To find out the interaction terms we have plotted interaction plot between sex and Pclass. The interaction plot shows that the lines are not parallel so we conclude that two variables are interacting with each other and could show significance in our model.

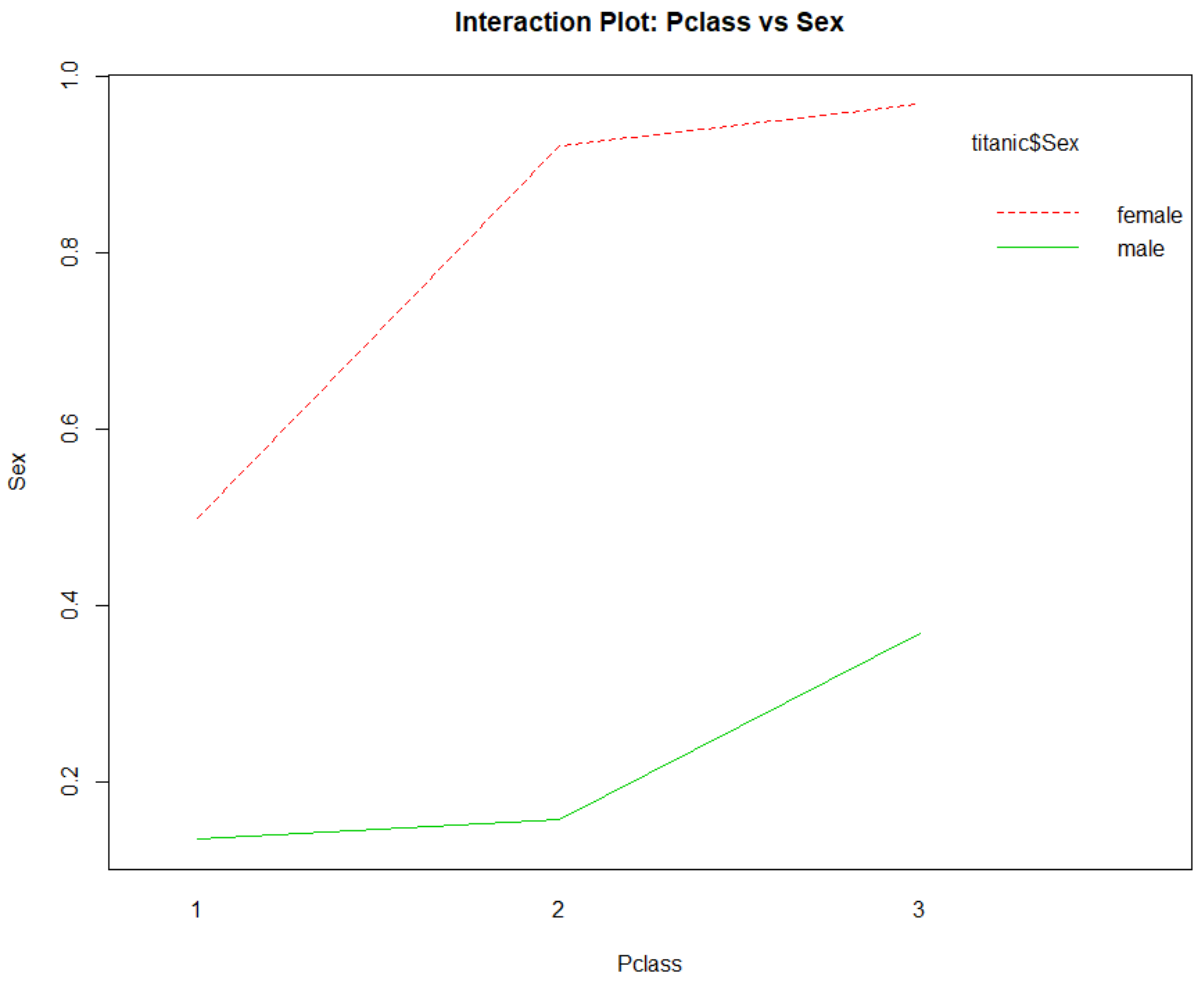

Fig 17: Interaction plot between Pclass and Sex. The lines are not parallel which shows interaction between the two variables

Splitting The Dataset:

In the next step, we decided to split our dataset into 2 sets (train and test) by taking a split ratio of 0.7 randomly. We did so because after training our model, we need to have a dataset to see how well our model performs on the unseen observations. It is a kind of validating our model.

```
> #--------------------------------Splitting the dataset-------------------------------------
>
> set.seed(100)  #It is used so that each time the dataset we get after splitting is the same
> sample_size<- sample.split(titanic$Survived, SplitRatio = 7/10) #Splitting the dataset into 70/30 ratio
> train<-subset(titanic, sample_size==T)
> test<-subset(titanic, sample_size==F)
> nrow(train)
[1] 623
> nrow(test)
[1] 268
```

Fig 18: Splitting the dataset

Model Building and Variable Selection:

We decided to build 2 models:

1. Model 1 (without interaction term): We have modelled Survived as a function of Age, Sex, Pclass, Fare and Alone.

```
> model1<- glm(Survived~Pclass+Sex+Age+Fare+Alone, family = binomial(link = "logit"), data=train)
> summary(model1)

Call:
glm(formula = Survived ~ Pclass + Sex + Age + Fare + Alone, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6978  -0.6972  -0.4381   0.6915   2.5113

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.045871   0.287623  -0.159    0.873
Pclass       1.306803   0.172040   7.596 3.06e-14 ***
Sexmale     -2.474373   0.228752 -10.817  < 2e-16 ***
Age         -0.042860   0.009027  -4.748 2.05e-06 ***
Fare        -0.001165   0.002771  -0.420    0.674
AloneYes     0.041860   0.227402   0.184    0.854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 829.60  on 622  degrees of freedom
Residual deviance: 579.94  on 617  degrees of freedom
AIC: 591.94

Number of Fisher Scoring iterations: 5
```

Fig 19: Model1 without interaction

2. Model 2 (with interaction term): We have taken an interaction term (Pclass*Sex) while modeling Survived as a function of Age, Sex, Pclass, Fare, Alone and Pclass*Sex.

```
> model2.int<- glm(Survived~Pclass+Sex+Age+Fare+Alone+Pclass*Sex, family = binomial(link = "logit"), data=train)
> summary(model2.int)

Call:
glm(formula = Survived ~ Pclass + Sex + Age + Fare + Alone +
    Pclass * Sex, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3215  -0.6920  -0.4890   0.4897   2.3972

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.528748   0.489743  -3.122 0.001799 **
Pclass          2.448944   0.378013   6.478 9.27e-11 ***
Sexmale        -0.263406   0.578707  -0.455 0.648992
Age            -0.044325   0.009451  -4.690 2.73e-06 ***
Fare           -0.001434   0.003141  -0.457 0.647865
AloneYes        0.008481   0.236525   0.036 0.971395
Pclass:Sexmale -1.474296   0.389870  -3.782 0.000156 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 829.60  on 622  degrees of freedom
Residual deviance: 560.88  on 616  degrees of freedom
AIC: 574.88

Number of Fisher Scoring iterations: 6
```

Fig 20: Model2 with interaction

After that, we have used backward elimination method to take only the significant variables for modeling.

Backward Elimination:

In R**,** backward elimination is performed using minimum AIC criteria. It takes AIC of Null model as a base cut off in the first iteration and removes the variable with the AIC less than the NULL model’s AIC in the second iteration and then in the next step it takes AIC of removed variable as the cut off AIC to discard another variable whose AIC is less than the new cut off and it will perform the iterations again and again until all the variables have AIC greater than the cut off AIC, which is our refined set of variables.

We have applied backward selection procedure to both our models and then built new models accordingly.

```
> step(model1, direction = "backward" )
Start:  AIC=580.13
Survived ~ Pclass + Sex + Age + Fare + Alone

         Df Deviance    AIC
- Alone   1   568.14 578.14
- Fare    1   568.71 578.71
<none>        568.13 580.13
- Age     1   583.49 593.49
- Pclass  1   628.45 638.45
- Sex     1   718.05 728.05

Step:  AIC=578.14
Survived ~ Pclass + Sex + Age + Fare

         Df Deviance    AIC
- Fare    1   568.85 576.85
<none>        568.14 578.14
- Age     1   583.87 591.87
- Pclass  1   628.68 636.68
- Sex     1   724.69 732.69

Step:  AIC=576.85
Survived ~ Pclass + Sex + Age

         Df Deviance    AIC
<none>        568.85 576.85
- Age     1   583.93 589.93
- Pclass  1   651.67 657.67
- Sex     1   727.68 733.68

Call:  glm(formula = Survived ~ Pclass + Sex + Age, family = binomial(link = "logit"),
    data = train)

Coefficients:
(Intercept)       Pclass      Sexmale          Age
   -0.09683      1.24262     -2.52667     -0.03430

Degrees of Freedom: 622 Total (i.e. Null);  619 Residual
Null Deviance:      829.6
Residual Deviance: 568.9        AIC: 576.9
```

Fig 21: Performing backward elimination on model 1 to select variables

Refined Models:

train.model1: **-**After performing backward elimination on model 1, we have modeled Survived as a function of Pclass, Sex ,and Age.

```
> train.model1<- glm(Survived~ Pclass+Age+Sex, family = binomial("logit"), data=train)
> summary(train.model1)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7255  -0.7003  -0.4367   0.6967   2.5071

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02268    0.27756  -0.082    0.935
Pclass       1.26530    0.14956   8.460  < 2e-16 ***
Age         -0.04192    0.00872  -4.807 1.53e-06 ***
Sexmale     -2.45485    0.22110 -11.103  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 829.6  on 622  degrees of freedom
Residual deviance: 580.2  on 619  degrees of freedom
AIC: 588.2

Number of Fisher Scoring iterations: 5
```

Fig 22: Refined model1 after backward elimination

train.model2: **-**After performing backward elimination on model 2, we have modeled Survived as a function of Pclass, Sex, Age and Pclass*Age, according to the results of backward procedure.

```
> train.model2<- glm(Survived~Pclass+Age+Sex+Pclass*Sex, family=binomial(link="logit"), data=train)
> summary(train.model2)

Call:
glm(formula = Survived ~ Pclass + Age + Sex + Pclass * Sex, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3563  -0.6932  -0.4897   0.4950   2.3919

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.527750   0.487981  -3.131 0.001744 **
Pclass         2.414548   0.373302   6.468 9.93e-11 ***
Age           -0.043581   0.009044  -4.819 1.44e-06 ***
Sexmale       -0.244131   0.575526  -0.424 0.671428
Pclass:Sexmale -1.483048   0.391841  -3.785 0.000154 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 829.60  on 622  degrees of freedom
Residual deviance: 561.12  on 618  degrees of freedom
AIC: 571.12

Number of Fisher Scoring iterations: 6
```

Fig 23: Refined model2 after backward elimination

Model Selection Using Backward Elimination In SAS:

With a motivation of getting a different model from SAS so that we can compare the two models, one from R and another from SAS, we put our model containing all the variables in SAS.

In SAS, backward elimination works by eliminating the variables with the highest p-value in each step. For example, in the very first step it will remove the variable that has a greater p-value than your pre-defined cut-off value, say 0.05. In the next step it will again fit a model containing all the variables excluding the selected variable in the first step. It will again check for the variable with a p-value greater than 0.05 and if it finds such variable, it will again remove that variable in the next step, and it will continue until it finds a model with all the variables having p-value less than 0.05.

Coincidentally, after fitting our model in SAS using backward elimination procedure, we found a model that is similar to the model we obtained by performing backward elimination in R. Therefore, we decided to go with the two models previously used in R for further analysis.

SAS Result After Performing Backward Elimination:

```
proc import datafile="H:\Data\WINXP\Desktop\titanic_cleaned.csv"
out=mydata dbms=csv replace;
getnames=yes;
run;

proc print data=mydata;
run;

proc logistic dat=mydata;
    class Sex(param=ref ref=First) Embarked(param=ref ref=First) Alone(param=ref ref=First);
    model Survived(Event="1")=
        Pclass Sex Age Fare Embarked Alone
        /selection=backward slstay=0.05 hierarchy=single details;
run;
```

Fig 24: SAS code for model fitting

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	891
Number of Observations Used	891

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	549
2	1	342

Probability modeled is Survived='1'.

Backward Elimination Procedure

Class Level Information			
Class	Value	Design Variables	
Sex	female	0	
	male	1	
Embarked	C	0	0
	Q	1	0
	S	0	1
Alone	No	0	
	Yes	1	

Step 0. The following effects were entered:

Fig 25: SAS output

Intercept Pclass Sex Age Fare Embarked Alone

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1188.655	812.639
SC	1193.447	850.978
-2 Log L	1186.655	796.639

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	390.0164	7	<.0001
Score	346.9003	7	<.0001
Wald	237.9093	7	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	72.0132	<.0001
Sex	1	176.2953	<.0001
Age	1	21.9002	<.0001
Fare	1	0.0068	0.9345
Embarked	2	5.6369	0.0597
Alone	1	0.2360	0.6271

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3275	0.3374	0.9425	0.3316
Pclass		1	1.2387	0.1460	72.0132	<.0001
Sex	male	1	-2.6108	0.1966	176.2953	<.0001
Age		1	-0.0369	0.00788	21.9002	<.0001
Fare		1	-0.00018	0.00216	0.0068	0.9345
Embarked	Q	1	-0.0824	0.3737	0.0486	0.8256
Embarked	S	1	-0.4945	0.2335	4.4853	0.0342

Fig 26: SAS output(cont.)

Alone	Yes	1	0.0959	0.1974	0.2360	0.6271

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pclass	3.451	2.592	4.594
Sex male vs female	0.073	0.050	0.108
Age	0.964	0.949	0.979
Fare	1.000	0.996	1.004
Embarked Q vs C	0.921	0.443	1.916
Embarked S vs C	0.610	0.386	0.964
Alone Yes vs No	1.101	0.748	1.621

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.2	Somers' D	0.704
Percent Discordant	14.8	Gamma	0.704
Percent Tied	0.0	Tau-a	0.333
Pairs	187758	c	0.852

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	72.0132	<.0001
Sex	1	176.2953	<.0001
Age	1	21.9002	<.0001
Fare	1	0.0068	0.9345
Embarked	2	5.6369	0.0597
Alone	1	0.2360	0.6271

Step 1. Effect Fare is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
		Intercept and

Fig 27: SAS output(cont.)/ Variable Fare is removed

Criterion	Intercept Only	Covariates
AIC	1188.655	810.645
SC	1193.447	844.192
-2 Log L	1186.655	796.645

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	390.0096	6	<.0001
Score	346.8944	6	<.0001
Wald	237.8766	6	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	87.4090	<.0001
Sex	1	176.2938	<.0001
Age	1	21.9190	<.0001
Embarked	2	5.7229	0.0572
Alone	1	0.2711	0.6026

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3252	0.3361	0.9359	0.3333
Pclass		1	1.2336	0.1319	87.4090	<.0001
Sex	male	1	-2.6104	0.1966	176.2938	<.0001
Age		1	-0.0368	0.00787	21.9190	<.0001
Embarked	Q	1	-0.0806	0.3731	0.0466	0.8290
Embarked	S	1	-0.4913	0.2303	4.5495	0.0329
Alone	Yes	1	0.0998	0.1917	0.2711	0.6026

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pclass	3.433	2.651	4.447
Sex male vs female	0.074	0.050	0.108
Age	0.964	0.949	0.979
Embarked Q vs C	0.923	0.444	1.917

Fig 28: SAS output(cont.)

Embarked S vs C	0.612	0.390	0.961
Alone Yes vs No	1.105	0.759	1.609

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.1	Somers' D	0.704
Percent Discordant	14.6	Gamma	0.706
Percent Tied	0.3	Tau-a	0.334
Pairs	187758	c	0.852

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
0.0068	1	0.9345

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	87.4090	<.0001
Sex	1	176.2938	<.0001
Age	1	21.9190	<.0001
Embarked	2	5.7229	0.0572
Alone	1	0.2711	0.6026

Step 2. Effect Alone is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1188.655	808.918
SC	1193.447	837.672
-2 Log L	1186.655	796.918

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq

Fig 29: SAS output(cont.) / Variable Alone is removed

Likelihood Ratio	389.7372	5	<.0001
Score	346.7170	5	<.0001
Wald	237.7985	5	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	87.6941	<.0001
Sex	1	188.6279	<.0001
Age	1	21.9634	<.0001
Embarked	2	5.7369	0.0568

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3472	0.3335	1.0837	0.2979
Pclass		1	1.2244	0.1307	87.6941	<.0001
Sex	male	1	-2.5816	0.1880	188.6279	<.0001
Age		1	-0.0360	0.00768	21.9634	<.0001
Embarked	Q	1	-0.0526	0.3683	0.0204	0.8863
Embarked	S	1	-0.4834	0.2298	4.4232	0.0355

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pclass	3.402	2.633	4.396
Sex male vs female	0.076	0.052	0.109
Age	0.965	0.950	0.979
Embarked Q vs C	0.949	0.461	1.953
Embarked S vs C	0.617	0.393	0.968

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.0	Somers' D	0.704
Percent Discordant	14.6	Gamma	0.707
Percent Tied	0.4	Tau-a	0.333
Pairs	187758	c	0.852

Fig 30: SAS output(cont.)

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
0.2779	2	0.8703

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	87.6941	<.0001
Sex	1	188.6279	<.0001
Age	1	21.9634	<.0001
Embarked	2	5.7369	0.0568

Step 3. Effect Embarked is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1188.655	810.640
SC	1193.447	829.810
-2 Log L	1186.655	802.640

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	384.0148	3	<.0001
Score	342.9157	3	<.0001
Wald	236.1923	3	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	97.7595	<.0001
Sex	1	194.3886	<.0001
Age	1	22.8996	<.0001

Fig 31: SAS output(cont.) / Embarked is removed

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.00231	0.2423	0.0001	0.9924
Pclass		1	1.2411	0.1255	97.7595	<.0001
Sex	male	1	-2.6067	0.1870	194.3886	<.0001
Age		1	-0.0366	0.00765	22.8996	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pclass	3.459	2.705	4.424
Sex male vs female	0.074	0.051	0.106
Age	0.964	0.950	0.979

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.2	Somers' D	0.694
Percent Discordant	14.8	Gamma	0.701
Percent Tied	1.0	Tau-a	0.329
Pairs	187758	c	0.847

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
6.0476	4	0.1956

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Pclass	1	97.7595	<.0001
Sex	1	194.3886	<.0001
Age	1	22.8996	<.0001

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq

1	Fare	1	5	0.0068	0.9345
2	Alone	1	4	0.2711	0.6026
3	Embarked	2	3	5.7369	0.0568

Fig 32: SAS output(cont.)

The final model we obtained from SAS is:

Survived = 0.00231 + 1.2411 (Pclass) - 2.6067(Sex:male) – 0.0366(Age), which is the same as obtained in the R output (shown below).

```
> train.model1<- glm(Survived~ Pclass+Age+Sex, family = binomial("logit"), data=titanic)
> summary(train.model1)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7122  -0.6525  -0.4339   0.6413   2.4738

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.002313   0.242346   0.010    0.992
Pclass       1.241094   0.125523   9.887  < 2e-16 ***
Age         -0.036631   0.007655  -4.785 1.71e-06 ***
Sexmale     -2.606730   0.186965 -13.942  < 2e-16 ***
---
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:  802.64  on 887  degrees of freedom
AIC: 810.64

Number of Fisher Scoring iterations: 5
```

Fig 33: R output for the chosen model

Likelihood Ratio Test

After choosing the two refined models, we performed a likelihood ratio test to check the goodness of fit for each model and to select the best model among these two models.

The reduced model is our first refined model with the lesser number of variables (model without interaction term) and our full model is the model with a greater number of variables i.e. the model with an interaction term

- **Null Hypothesis:** Reduced model is appropriate (Survived ~ Pclass+ Age+ Sex)
- **Alternate Hypothesis:** Full model is appropriate (Survived ~ Pclass+ Age+ Sex+ Pclass*Sex)

```
> #--------------------------------Likelihood-Ratio Test--------------------------------
> #Its result shows that the full model (model with interaction term) is appropriate
> lrtest(train.model1,train.model2)
Likelihood ratio test

Model 1: Survived ~ Pclass + Age + Sex
Model 2: Survived ~ Pclass + Age + Sex + Pclass * Sex
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -290.10
2   5 -280.56  1 19.081  1.253e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Fig 34: Likelihood ratio test

After performing likelihood ratio test, we analyzed that the p-value (1.253e-05) of the test is much smaller than our level of significance’s value of 0.05, thus, we reject null hypothesis and conclude that our full model is appropriate.

Wald Test:

The Wald test is used to check whether a parameter is significant for a given model. Wald test is also used to test whether a parameter will explain the variability in a model or not. We have taken the second model to perform Wald’s test.

- **Null Hypothesis:** All parameters are zero except intercept.
- **Alternate Hypothesis:** At least one of the parameters is not zero.

```
> Anova(train.model2)
Analysis of Deviance Table (Type II tests)

Response: Survived
           LR Chisq Df Pr(>Chisq)
Pclass       86.075  1  < 2.2e-16 ***
Age          25.604  1  4.192e-07 ***
Sex         150.033  1  < 2.2e-16 ***
Pclass:Sex   19.081  1  1.253e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Fig 35: Wald Test

After performing the Wald test, we analyzed that deviance for each variable is greater than 3.84 (level of significance=0.05 and degree of freedom = 1). Therefore, we can reject the null hypothesis and conclude that at least one of the parameters is not zero.

Classification Report:

A Classification report is used to measure the quality of predictions from a classification algorithm. Since our dataset is unbalanced, means the number of survived observations is less than the number of dead people, therefore it is not a good criterion to check the efficiency of our model.

We have still made a classification report for both the models. We have taken the cut off value as 0.38 by taking the proportion of survival upon total.

Cut Off Value: 0.38

```
> predictions1<- predict(train.model1, test, type="response")
> predictions2<- predict(train.model2, test, type="response")
> cutoff<- mean(train$Survived)
> cutoff
[1] 0.3836276
```

Fig 36: Proportion

Predictions:

For predictions, we made a function so that it assigns value as 1 whose predicted value is greater than the cut off value of 0.38 and the rest as 0.

Predictions1 tells the predicted values for model 1 and predictions2 does the same for model 2.

```
> #----------Function To Assign Survival Category Based On Probabilit Cutoff----
----
>
> for (i in (1:length(predictions1))) {
+   if (predictions1[i]>cutoff){
+     predictions1[i]<- 1}else{
+       predictions1[i]<-0
+     }}
>
> for (i in (1:length(predictions2))) {
+   if (predictions2[i]>cutoff){
+     predictions2[i]<- 1}else{
+       predictions2[i]<-0
+     }}
> head(predictions1)
 2  4 10 12 14 23
 1  1  1  1  0  1
> head(predictions2)
 2  4 10 12 14 23
 1  1  1  1  0  1
```

Fig 37: Filling Survival Values based on Cut-Off

Classification Report for Model 1:

```
> class_rpt.model1<-xtabs(~predictions1+test$Survived) #for my test dataset
> class_rpt.model1
            test$Survived
predictions1   0   1
           0 139  20
           1  26  83
>
```

Fig 38: Classification Report for Model 1

Classification Report for Model 2**:**

```
> class_rpt.model2<-xtabs(~predictions2+test$Survived) #for my test dataset
> class_rpt.model2
            test$Survived
predictions2   0   1
           0 149  26
           1  16  77
```

Fig 39: Classification Report for Model 2

Sensitivity, Specificity and Accuracy:

The **sensitivity** of a test (also called the true positive rate) is defined as the proportion of survived people who will have a positive result.

The **specificity** of a test (also called the true negative rate) is defined as the proportion of dead people who will have a negative result.

The **accuracy** of a model defines the number of correct predictions that is made by our model.

Model 1:

```
> sensitivity1<- 83/(83+20)
> specificity1<- 139/(139+26)
> accuracy1<- (139+83)/(139+83+20+26)
> sensitivity1 #80.5%
[1] 0.8058252
> specificity1 #84.2%
[1] 0.8424242
> accuracy1    #82.8%
[1] 0.8283582
```

Fig 40: Sensitivity, Specificity and accuracy for Model 1

Model 2:

```
> senstivity2 <- 77/(77+26)
> specificity2 <- 149/(149+16)
> accuracy2<- (149+77)/(149+26+77+16)
> senstivity2  #74.7%
[1] 0.7475728
> specificity2 #90.3%
[1] 0.9030303
> accuracy2    #84.3%
[1] 0.8432836
```

Fig 41: Sensitivity, Specificity and accuracy of Model 2

ROC curve:

The receiver operating characteristic (ROC) curve, which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR) as the x coordinate, is an effective method of evaluating the performance of diagnostic tests. We have plotted the ROC curve for both our models and find out the area under curve (AUC) for both the plots.

According to the results of ROC AUC, our model 2 is better than model 1 as it is covering an area of around 84.6% which is greater than the area covered by model 1 which is 83.9%.

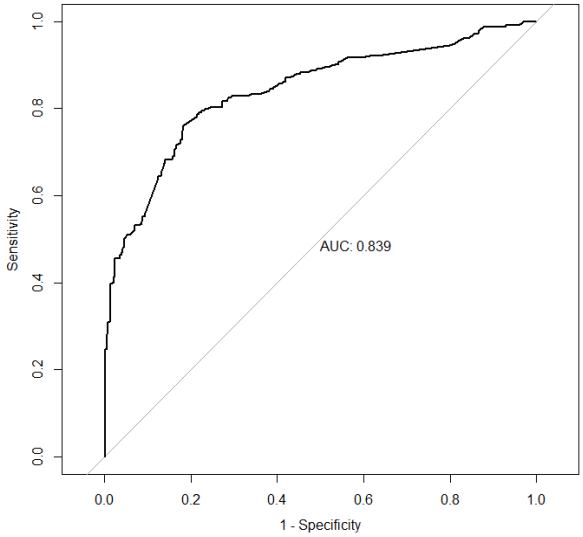

Fig 42: ROC curve for Model 1

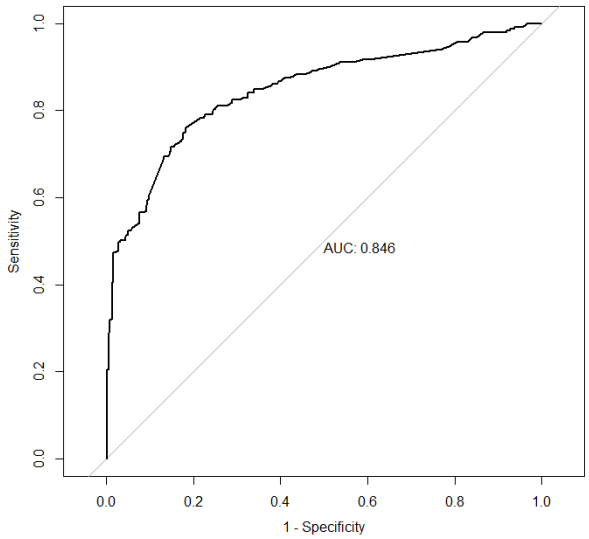

Fig 43: ROC curve for Model 2

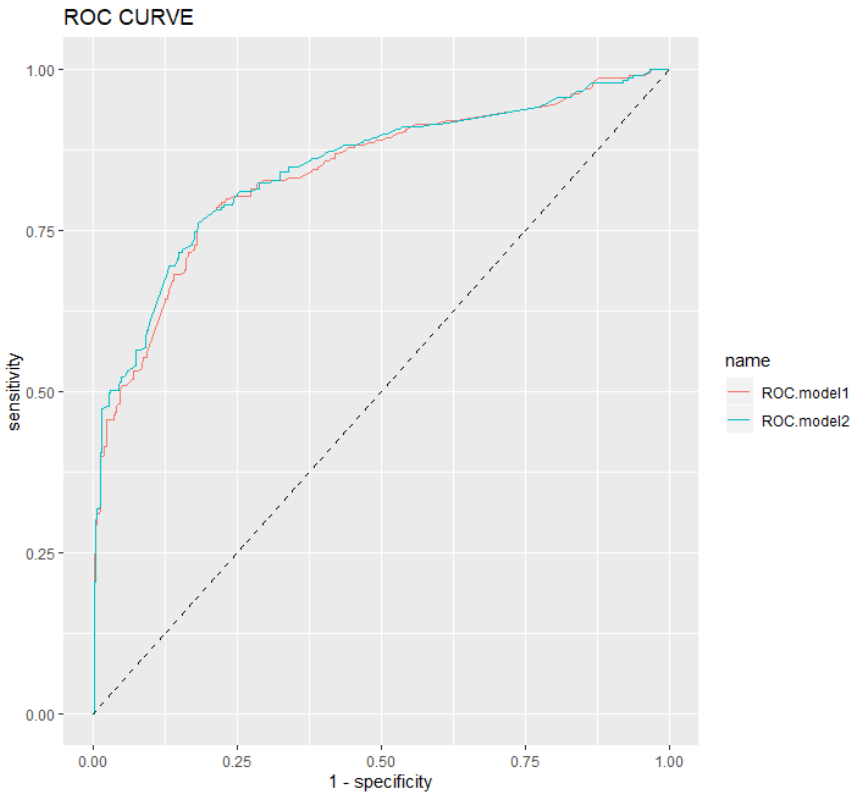

Fig 44: ROC Curve for Model 1 and Model 2

Hosmer Lemeshow Test:

Our data is ungrouped; therefore, the Hosmer Lemeshow test is appropriate to judge the model's goodness of fit.

Null Hypothesis: The model is a good fit.

Alternate Hypothesis: The model is not a good fit.

Our results show the p-value of 0.624 which is greater than 0.05 (at 95% level of significance), therefore, we failed to reject the null hypothesis and conclude that our model is a good fit.

```
> #-----------------------------Hosmer Lemeshow test---------------------
> #The results shows that the fitted model does not show lack of fit
>
> hl<-hoslem.test(train$Survived, fitted(train.model2))
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  train$Survived, fitted(train.model2)
X-squared = 6.2075, df = 8, p-value = 0.624
```

Fig 45: Hosmer Lemeshow Test

Final Model:

After performing several tests, we have chosen our final model as follows:

Survived = -1.53 + 2.41 (Pclass) - 0.04 (Age) - 0.24 (Sexmale) - 1.48 (Pclass*Sexmale)

Interpretation Of The Coefficient Estimates:

1. **Age:** With every one year increase in Age, the estimated odds of survival are multiplied by a factor of e^(-0.04), which is equal to 0.96, keeping all other variables constant.

2. **Pclass*Sexmale:** The effect of Pclass on the log odds of survival depends on the levels of Sex.
3. **Pclass And Sex:** As these variables are involved in interaction, we can't interpret them.

Sigmoidal Curve of The Fitted Model:

We have plotted the sigmoidal curve of the predicted probability of each observation of the test dataset against each observation number. The plot is shown below:

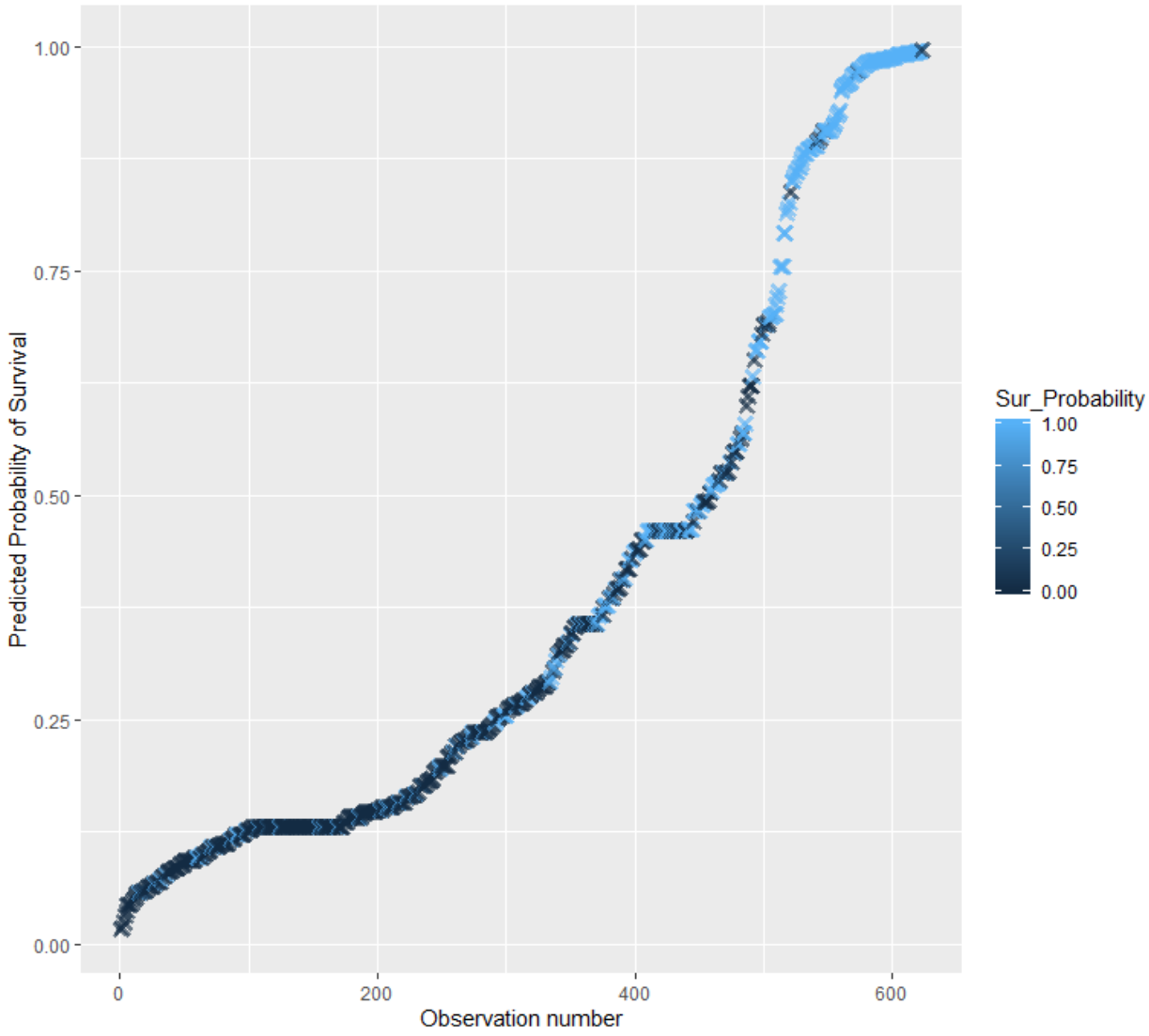

Fig 46: Sigmoidal Curve of The Fitted Model