Titanic Analysis : Machine Learning

Ankita Pal

7/28/2020

Titanic Data :-

RMS Titanic was a British passenger liner operated by the White Star Line that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after striking an iceberg during her maiden voyage from Southampton to New York City. Of the estimated 2,224 passengers and crew aboard, more than 1,500 died. Here we have a data about the passengers information along with whether they survived or not.
The data has the following variables described below:
1. PassengerId - ID of the Passengers
2. Survived - Passenger Survival Indicator (Categorical ; 1 = Survived, 0 = Not Survived)
3. Pclass - Passenger Class (Categorical)
4. Name - Passenger Name
5. Sex - Sex of Passengers (Categorical)
6. Age - Age of Passengers (Continuous)
7. SibSp - Number of Siblings/Spouses Aboard (Continuous)
8. Parch - Number of Parents/Children Aboard (Continuous)
9. Ticket - Ticket Number
10. Fare - Passenger Fare
11. Cabin - Passenger Cabin
12. Embarked - Port of Embarkation (Categorical ; C = Cherbourg, Q = Queenstown, S = Southampton)

Objective :-

The main objectives of the analysis of Titanic Dataset are:
1. How does features depend upon chance of Survival?
2. Predicting the Survival using preferred features in Q1.
3. Predicting the Survival in the entire ship and finding the proportion of survived passengers.

Loading the Required Packages :-

```
library(titanic)
library(dplyr)
library(caret)
library(ggplot2)
library(rpart)
library(rpart.plot)
```

```
library(randomForest)
library(tibble)
```

Loading the Titanic Data :-

The data can be loaded using the given `zip` file which contains two `.csv` files. But R contains `titanic` package which can be used to load the data.

```
dat1 <- titanic_train
dat2 <- titanic_test
head(dat1)
```

Table 1: Titanic Data

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

Data Processing :-

The data processing step contains checking for missing values and filling those missing values, changing the structure of different variables according to their use and so on.

```
data <- bind_rows(dat1,dat2)
# Finding missing values
colSums(is.na(data))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
```

```
colSums(data == "")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
```

There are quite a number of missing values in Age(263/1309) and Embarked. The 418 missing values in Survived are from the `titanic_test` data, so not required filling them up.
Finally cleaning and processing the data.

```
data_clean <- data %>%
  mutate(Survived = factor(Survived),
         Pclass = factor(Pclass),
         Age = ifelse(is.na(Age),median(Age,na.rm = TRUE),Age),
         Embarked = ifelse(Embarked == "",'S',Embarked)%>%factor(),
         FamilySize = SibSp + Parch + 1) %>%
  select(Survived,Sex,Pclass,Age,SibSp,Parch,FamilySize,Fare,Embarked)
```

On rechecking whether there are any missing values we find that there are not left missing values in `data_clean`.

Analysis the Data :-

Analysis for Objective 1

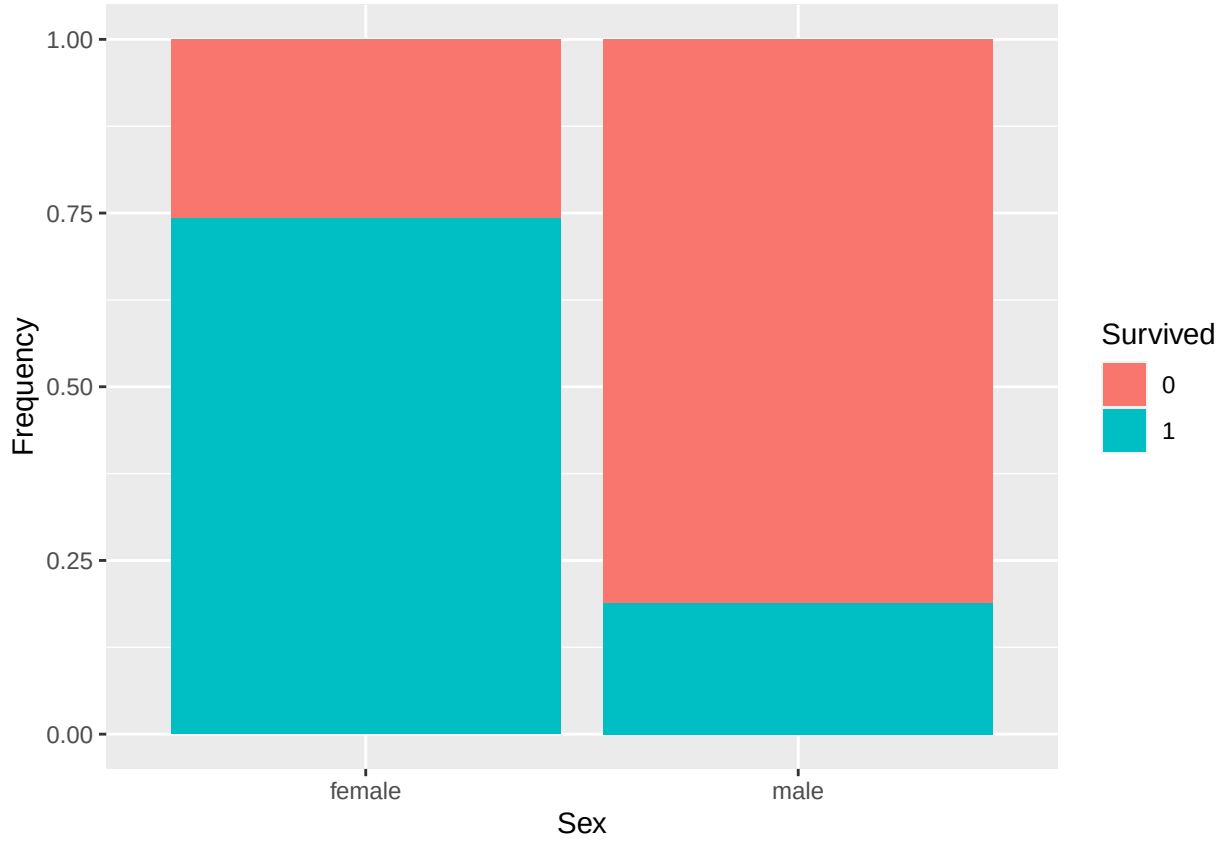
The train data has the Survived information, so here we will use only train data set.

```
train_set <- data_clean[1:dim(titanic_train)[1],]
```

The features that we will be concerned with are Sex, Pclass, Age, SibSp, Parch, FamilySize, Fare and Embarked. So, firstly checking all the features with the Survived information, i,e., how well each of the features are related to the chance of Survival.

1. **Sex as a function of Survival**

```
ggplot(train_set,aes(x = Sex,fill = Survived)) +
  geom_bar(position = "fill") + ylab("Frequency")
```




```
train_set %>% group_by(Sex) %>% summarize(Survived = mean(Survived == 1)*100)
```

```
## # A tibble: 2 x 2
##   Sex    Survived
##   <chr>     <dbl>
## 1 female     74.2
## 2 male       18.9
```

So, if the Sex = Female, then chance of Survival was greater about 74%.

2. **Age as a function of Survival**

```
ggplot(train_set,aes(x = Age,fill = Survived)) +
  geom_histogram(binwidth = 3,position = "fill") + ylab("Frequency")
```

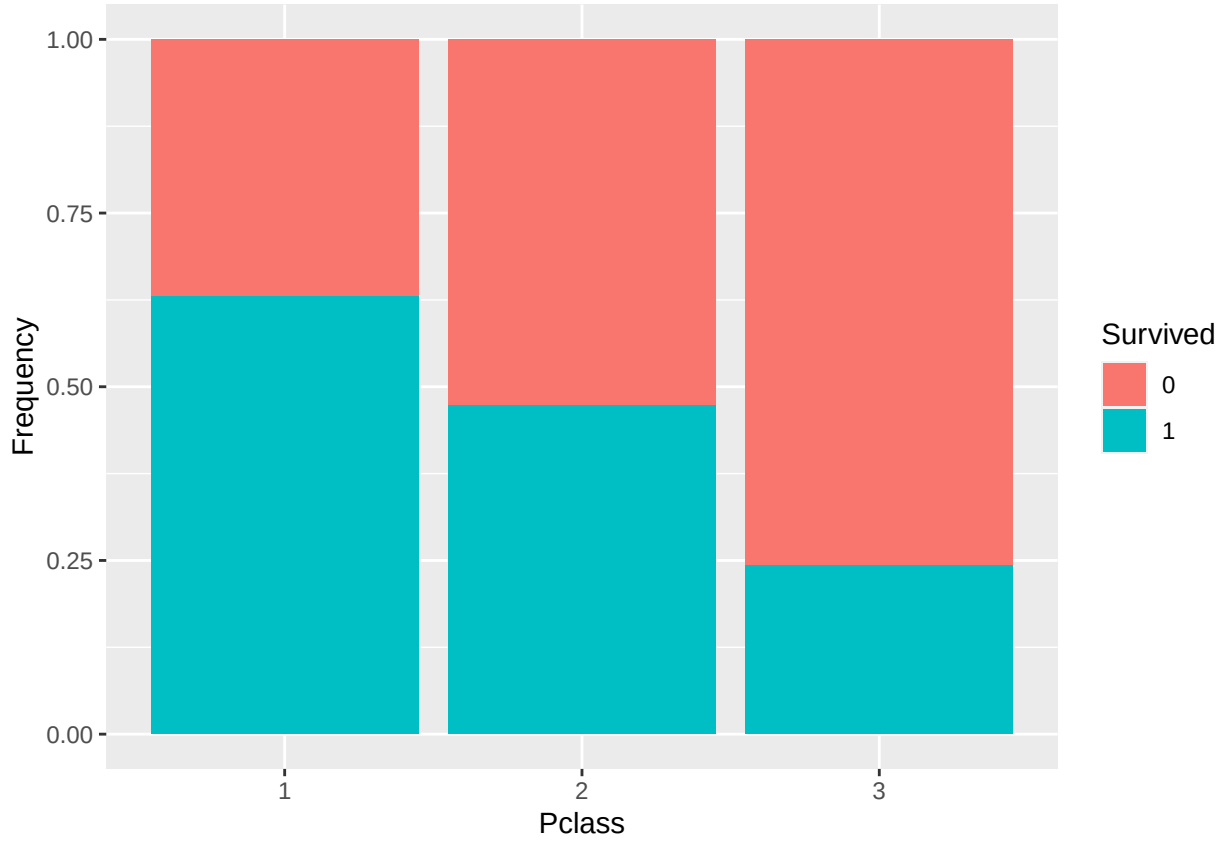

```
train_set %>% group_by(Age) %>% summarize(Survived = mean(Survived == 1)*100)
```

```
## # A tibble: 88 x 2
##      Age Survived
##    <dbl>    <dbl>
##  1  0.42    100
##  2  0.67    100
##  3  0.75    100
##  4  0.83    100
##  5  0.92    100
##  6  1        71.4
##  7  2        30
##  8  3        83.3
##  9  4        70
## 10  5       100
## # ... with 78 more rows
```

So, children less than 15y/o & old people >=80 has higher chance of survival.

3. **Pclass as a function of Survival**

```
ggplot(train_set,aes(x = Pclass,fill = Survived)) +
  geom_bar(position = "fill") + ylab("Frequency")
```



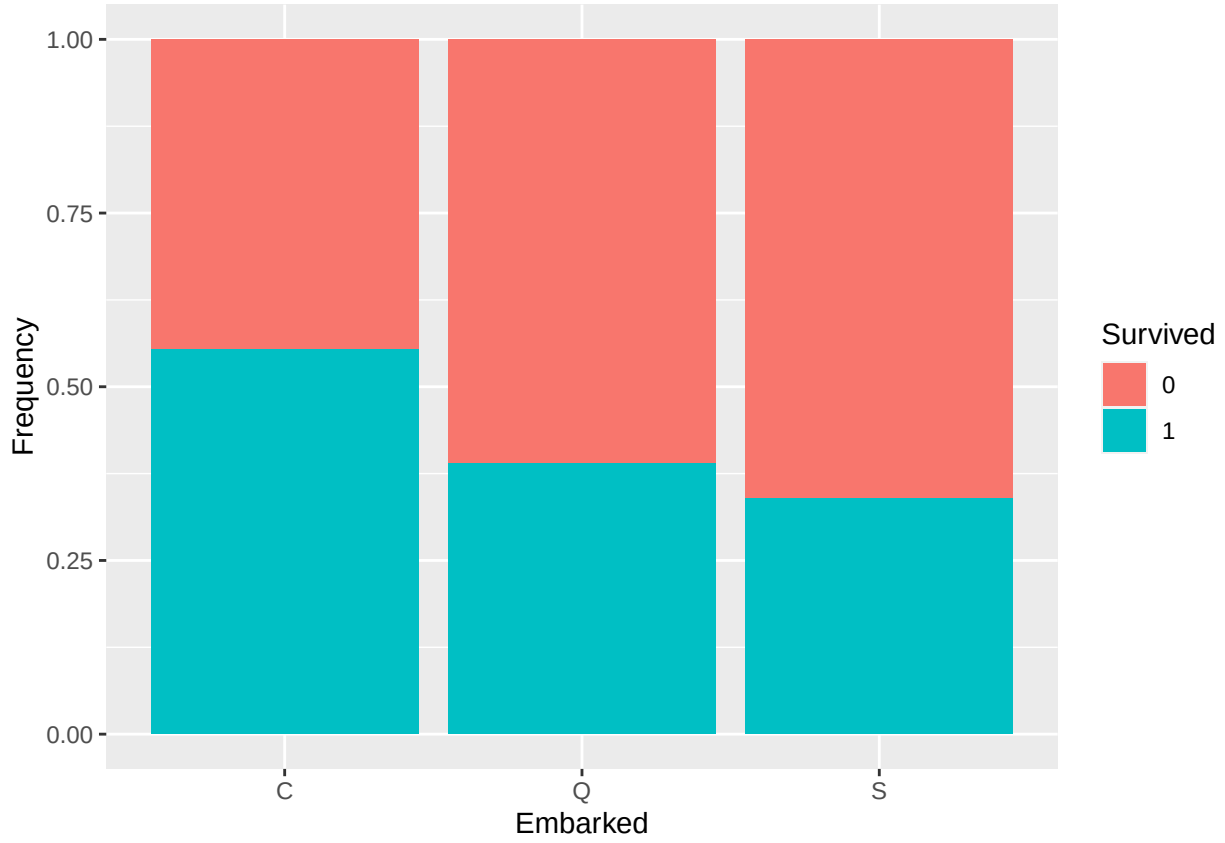

```
train_set %>% group_by(Pclass) %>% summarize(Survived = mean(Survived == 1)*100)
```

```
## # A tibble: 3 x 2
##   Pclass Survived
##   <fct>     <dbl>
## 1 1          63.0
## 2 2          47.3
## 3 3          24.2
```

So, if Passenger is from 1st class, chance of survival is greater, i.e., 63%.

4. **Embarked as a function of Survival**

```
ggplot(train_set,aes(x = Embarked,fill = Survived)) +
  geom_bar(position = "fill") + ylab("Frequency")
```



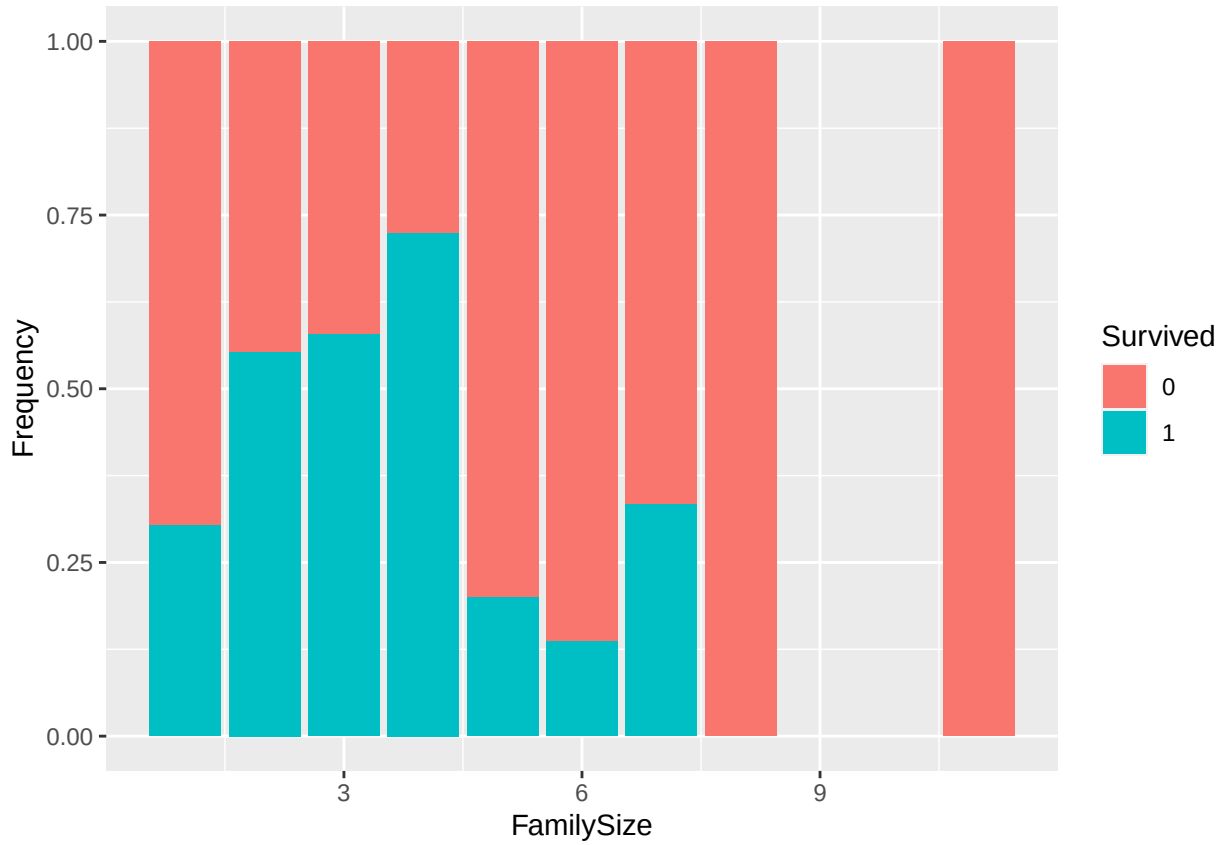

```
train_set %>% group_by(Embarked) %>% summarize(Survived = mean(Survived == 1)*100)
```

```
## # A tibble: 3 x 2
##   Embarked Survived
##   <fct>       <dbl>
## 1 C            55.4
## 2 Q            39.0
## 3 S            33.9
```

So, Passenger embarked from 'C' has greater chance of Survival about 55%.

5. **FamilySize as a function of Survival**

```
ggplot(train_set,aes(x = FamilySize,fill = Survived)) +
  geom_bar(position = "fill") + ylab("Frequency")
```




```
train_set %>% group_by(FamilySize) %>% summarize(Survived = mean(Survived == 1)*100)
```

```
## # A tibble: 9 x 2
##   FamilySize Survived
##        <dbl>    <dbl>
## 1          1     30.4
## 2          2     55.3
## 3          3     57.8
## 4          4     72.4
## 5          5     20
## 6          6     13.6
## 7          7     33.3
## 8          8      0
## 9         11      0
```

So, FamilySize between 2 & 4 has more than 50% chance of Survival.

Analysis for Objective 2

Before applying any method to find an algorithm for prediction purpose, the first step is to **partition the data into training and testing sets**. The **training set** is considered as the data whose outcome is known and is used to create the algorithm for predictions. The **testing set** is considered as the data whose outcome is unknown and the algorithm obtained using training set is applied in this data set for final predictions.

```
set.seed(42,sample.kind = 'Rounding')
index <- createDataPartition(train_set$Survived,times = 1,p = 0.5,list = FALSE)
train <- train_set[-index,]
test <- train_set[index,]
```

1. **Algorithm 1 : Logistic Regression**
   Let us use Logistic Regression to form the algorithm, but using only two features `Sex` and `Pclass` since these features has greater than 60% chance of survival.

```
model1 <- train(Survived~Sex+Pclass,method = 'glm',data = train)
pred1 <- predict(model1,test)
mean(pred1 == test$Survived)
```

```
## [1] 0.7780269
```

So, the **mean of correct predictions is quite high about 78%**.
Lets see if we can do better by including the other features.

```
model2 <- train(Survived~.,method = 'glm',data = train)
pred2 <- predict(model2,test)
m2 <- mean(pred2 == test$Survived)
m2
```

```
## [1] 0.7982063
```

So, the **mean of correct predictions has increased a little to about 80%**. So, `model2` is better than `model1`.
Creating Confusion Matrix for `model2`.

```
cm2 <- confusionMatrix(factor(pred2),test$Survived)$byClass['F1']
cm2
```

```
##        F1
## 0.8398577
```

The **F1 score is about 0.84** which is quite good, since it is far away from 0.

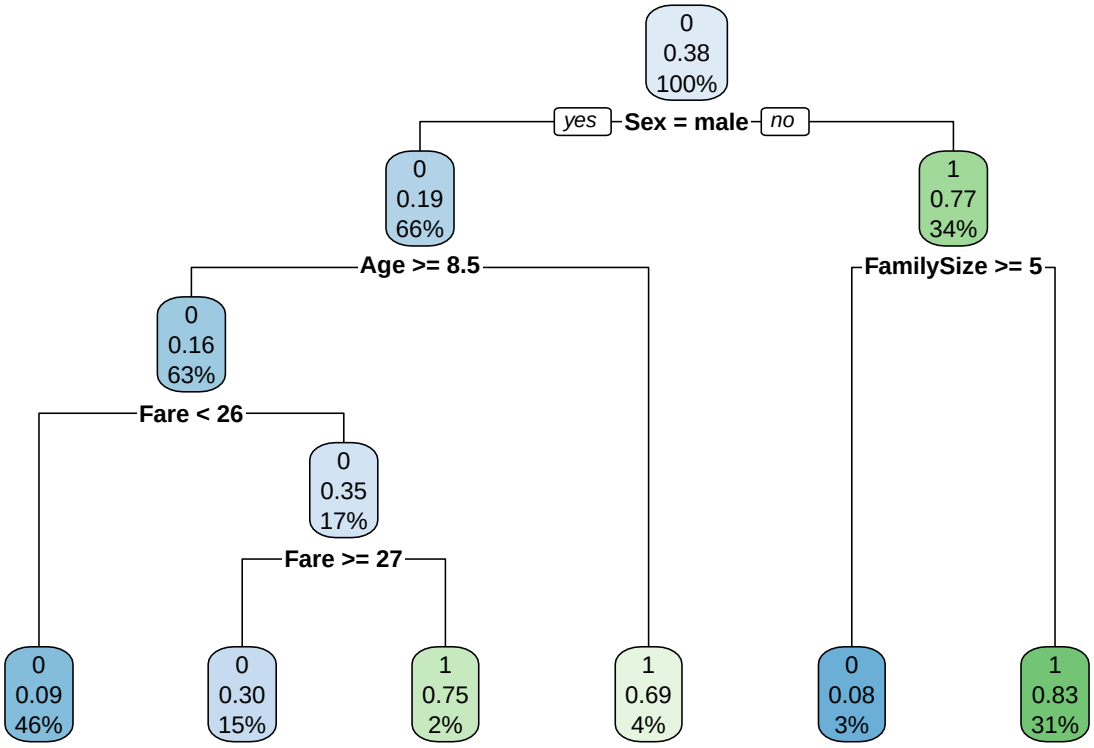
2. **Algorithm 2 : Decision Tree**
   Now, lets use Decision Tree method by using `rpart` function to obtain an algorithm for prediction and check whether it performs better than Logistic Regression or not.

```
model3 <- rpart(Survived~.,data = train,method = 'class')
pred3 <- predict(model3,test,type = 'class')
m3 <- mean(pred3 == test$Survived)
m3
```

```
## [1] 0.7869955
```

The **mean of correct predictions is about 79%** which is a little less than that obtained for Logistic Regression.
Thus, the decision tree is obtained as below.

```
rpart.plot(model3)
```



Creating Confusion Matrix for `model3`.

```
cm3 <- confusionMatrix(factor(pred3),test$Survived)$byClass['F1']
cm3
```

```
##        F1
## 0.8288288
```

The **F1 score is about 0.83** which is good, but less than that of obtained in Logistic Regression.
So, till now we can claim that Logistic Regression Model is preferable for prediction of passenger survival.

3. **Algorithm 3 : Random Forest**
   Random Forests is an approach which aims to improve prediction performance and reduce instability by averaging multiple decision trees.
   Since in Algorithm 2 we have used Decision Trees so lets check an algorithm made by Random Forest method using `randomForest` function.

```
model4 <- randomForest(Survived~.,data = train)
pred4 <- predict(model4,test)
m4 <- mean(pred4 == test$Survived)
m4
```

```
## [1] 0.8116592
```

The **mean of correct predictions is about 81%**, which is better than both Logistic Regression and Decision Tree models.
Lets see the important features according to this model.

```
varImp(model4)
```

	Overall
Sex	54.12
Pclass	12.14
Age	26.69
SibSp	7.07
Parch	5.70
FamilySize	11.01
Fare	31.76
Embarked	6.61

The most important feature is "Sex".
Creating Confusion Matrix for `model4`.

```
cm4 <- confusionMatrix(factor(pred4),test$Survived)$byClass['F1']
cm4
```

```
##        F1
## 0.8541667
```

The **F1 score is about 0.85** which is much better and greater than Decision Tree model and Logistic Regression Model.

Analysis for Objective 3

```
val <- c(m2,m3,m4,cm2,cm3,cm4)
matrix(val,ncol = 2,nrow = 3,byrow = FALSE,
       dimnames = list(c("Logistic Regression","Decision Tree","Random Forest"),
                       c("Mean Predictions","F1 Score")))
```

	Mean Predictions	F1 Score
Logistic Regression	0.79821	0.83986
Decision Tree	0.78700	0.82883
Random Forest	0.81166	0.85417

The **mean of correct predictions and F1 score is obtained highest for Random Forest Model about 81% and 0.85 respectively**. So, lets apply the model obtained from Random Forests on `titanic_test` data, which do not have any data about the Survival of Passengers.

```
test_set <- data_clean[length(train_set)+1:1309,]
Survived_pred <- predict(model4,test_set)[1:418]
result <- add_column(titanic_test,
                     Survived = as.numeric(as.character(Survived_pred)),
                     .before = "Pclass")
head(result)
```

Table 4: Predicted Titanic Data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	1	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000		S
894	1	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875		Q
895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625		S
896	0	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	310129	12.2875		S
897	1	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250		S

To find the proportion of passenger survival on the entire ship we will have to combine the resultant data, i.e., `result` and the data used for forming train and test sets, i.e., `data_clean`.

```
full_data <- bind_rows(titanic_train,result)
nrow(full_data)
```

```
## [1] 1309
```

```
mean(full_data$Survived == 1)*100
```

```
## [1] 37.12758
```

Conclusion :-

Therefore, according to the above predictions we can claim that **only about 37% out of a total 1309 passengers had survived** the Titanic Shipwreck according to the available data.