Prediction on Titanic Survivors

André Jakob

17/06/2019

Introduction

This is the final Report in order to get the Data Science Professional Certificate provided by EdX and HarvardX. After studying the multiple data science and machine learning processes, this final work proposal was to choose a data set by my own and to write a report. Following the recommendation, I have researched a couple of data sets on kaggle, and finally downloaded the Titanic data set — whose original file can be found in https://www.kaggle.com/mysticvalley/titanic-original#titanic3.xls. The challenge here is to predict who survived and who did not based on other data available from the Titanic passengers.

##Method

Cleaning the data set

The first step of the analysis is to take a look on the data. It was downloaded as a xls file, which I put in a R vector. Some numeric observations was originally imputed as characters, but R has coerced those automatically to numeric.

```
dim(titanic)
```

```
## [1] 1309    9
```

A first look on the data set shows us that we have 1309 observations, in other words passangers, and 14 variables: “pclass”, “survived”, “name”, “sex”, “age”, “sibsp”, “parch”, “ticket”, “fare”, “cabin”, “embarked”, “boat”, “body” and “home.dest”. Not all of these variables seem useful to base the prediction, so I selected only 8 who seem meaningful. The variables elected as non important are: name, cabin, boat, body and destination. I discarded name and destination because sounds obvious to me that these variables could not interfere in the survival rate of the passenger. Cabin and boat could have been useful, but unfortunately there are too many NA’s in their columns - even more than actual data (see below), so it would be too risky infere their values. And the classification of body is already a sign that the person did not survive (even though there are many NA’s between the non survivals). There is some subjectiveness in my choices, but I do not see them as very controversial and believe my explanation above is enough.

```
kable(nas0)
```

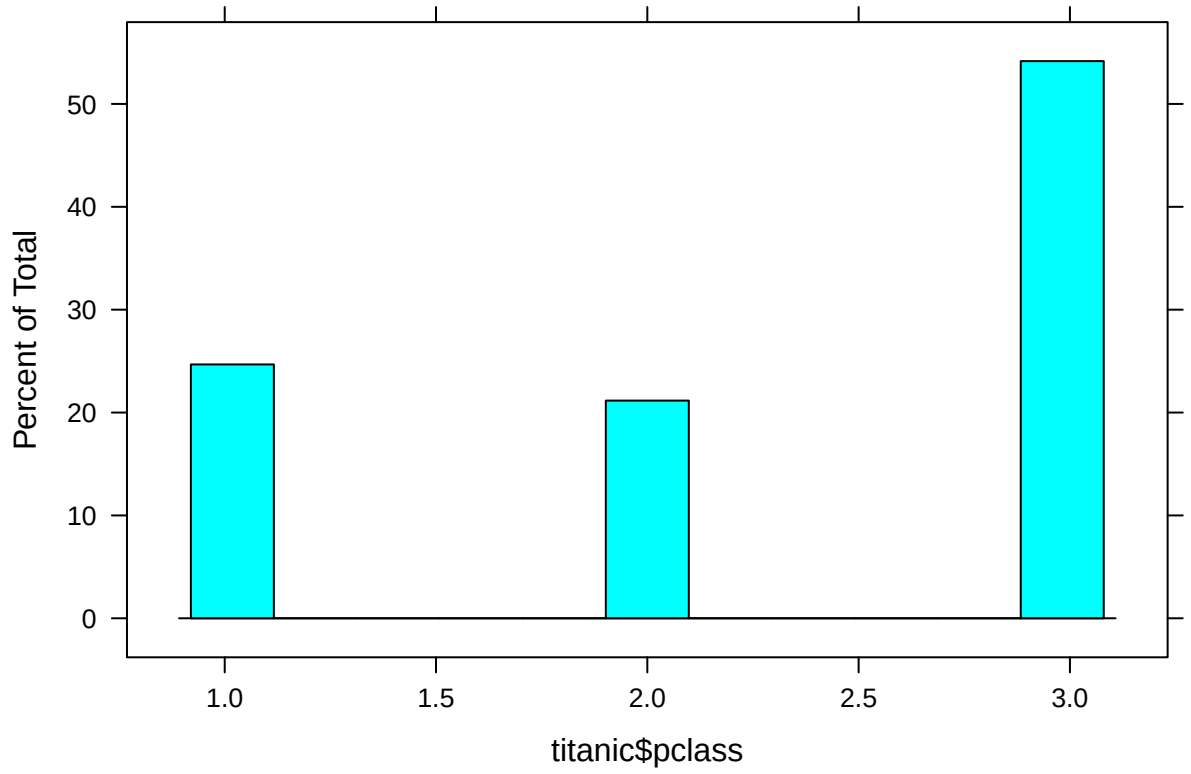
pclass	survived	sex	age	sibsp	parch	fare	embarked	cabin	boat
0	0	0	263	0	0	1	2	1014	823

Variables

So at the end we have a data set with 8 variables, which I briefly explain:

Pclass Referes to the class of the passengers, it has 3 possible values (1, 2 or 3).

```
histogram(titanic$pclass)
```



Survived Is the goal of prediction, and has 2 possible outcomes (1 for survived and 0 for not survived).

```
kable(summary)
```

survived	non_survived
500	809

Sex Male or female.

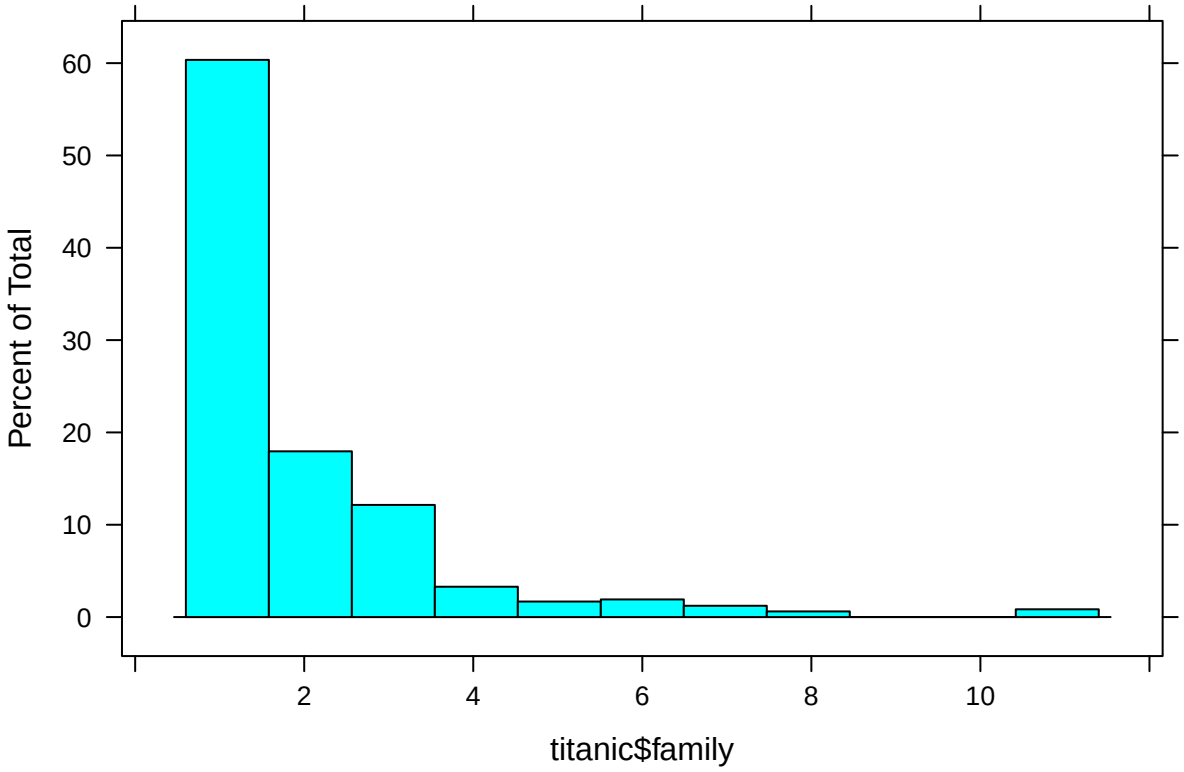
```
kable(sex)
```

Male	Female
843	466

Age Is a continuous distribution of the age in years from each passenger. As showed above, there are 263 NAs on the data set, so I will not plot this distribution nor analyze it further at this moment.

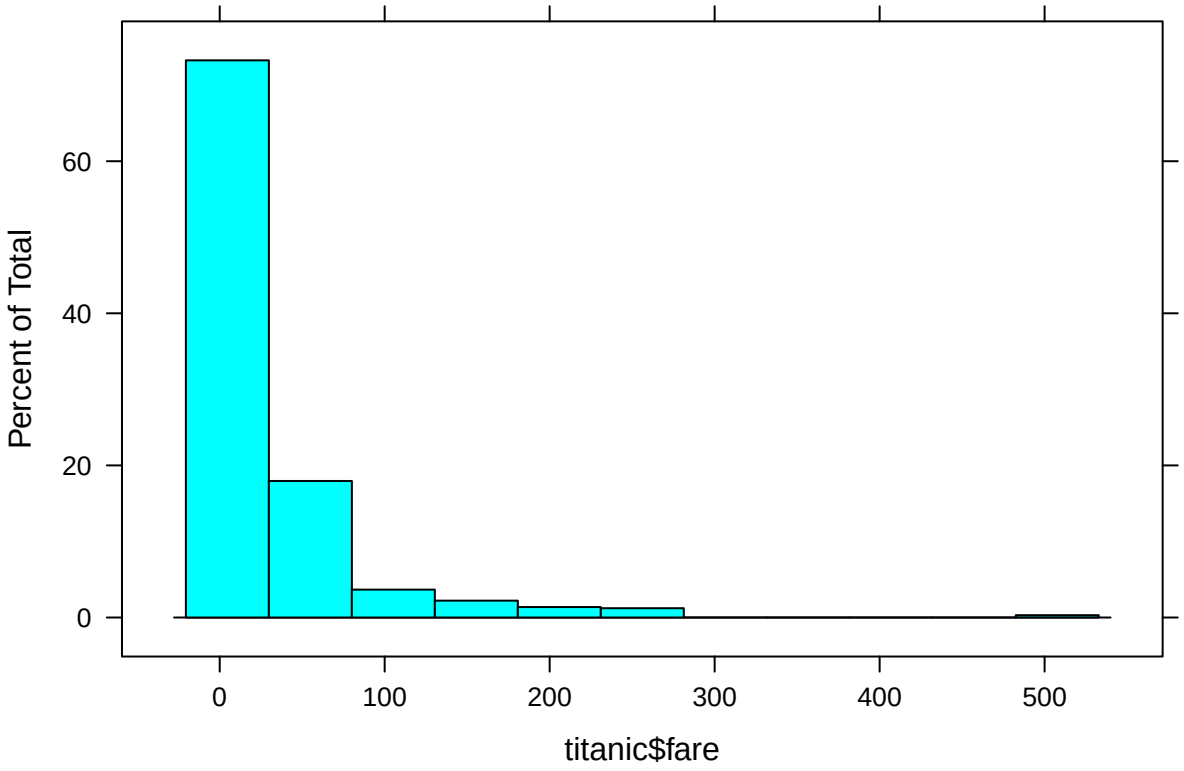
Sibsp and Parch I considered both as one since they are similar. The first one is the number of brothers, sisters or spouses while the second one is the number of parents and children each passenger has on board. At this point I created a new variable "family" to compute the size of the families on board, which is the sum of "parch" and "sibsp" plus 1 (the passenger himself).

```
histogram(titanic$family)
```



Fare Is the price each passenger paid for the ticket (probably in English pounds). It varies depending on many factors, most of them accordingly to the class of the ticket and the port that the passenger embarked because, among other issues, of the distance to be traveled. So I used the median of the fares of the other passengers with the same class and port to correct this NA. Then I reached the distribution of fares.

```
histogram(titanic$fare)
```



Embarked Is the port used by the passenger to embark on Titanic. It has 3 options (S for South Hampton, C for Cherbourg and Q for Queenstown). Filtering the data by similar observations on fare, class and family, I founded that Other passengers with the same characteristics have embarked in Queenstown port. So I have corrected the 2 NAs with "C" in the column embarked.

```
embarked<-titanic%>%filter(fare>=79&fare<=80&pclass=="1"&family=="1")
kable(embarked)
```

pclass	survived	sex	age	sibsp	parch	fare	embarked	family
1	0	male	24.00000	0	0	79.2	C	1
1	0	male	46.00000	0	0	79.2	C	1
1	1	female	38.00000	0	0	80.0	C	1
1	0	male	46.00000	0	0	79.2	C	1
1	1	female	62.00000	0	0	80.0	C	1
1	1	female	39.50444	0	0	79.2	C	1

Summary of Variables After filling this NAs, the distribution of the passengers by port is the table below.

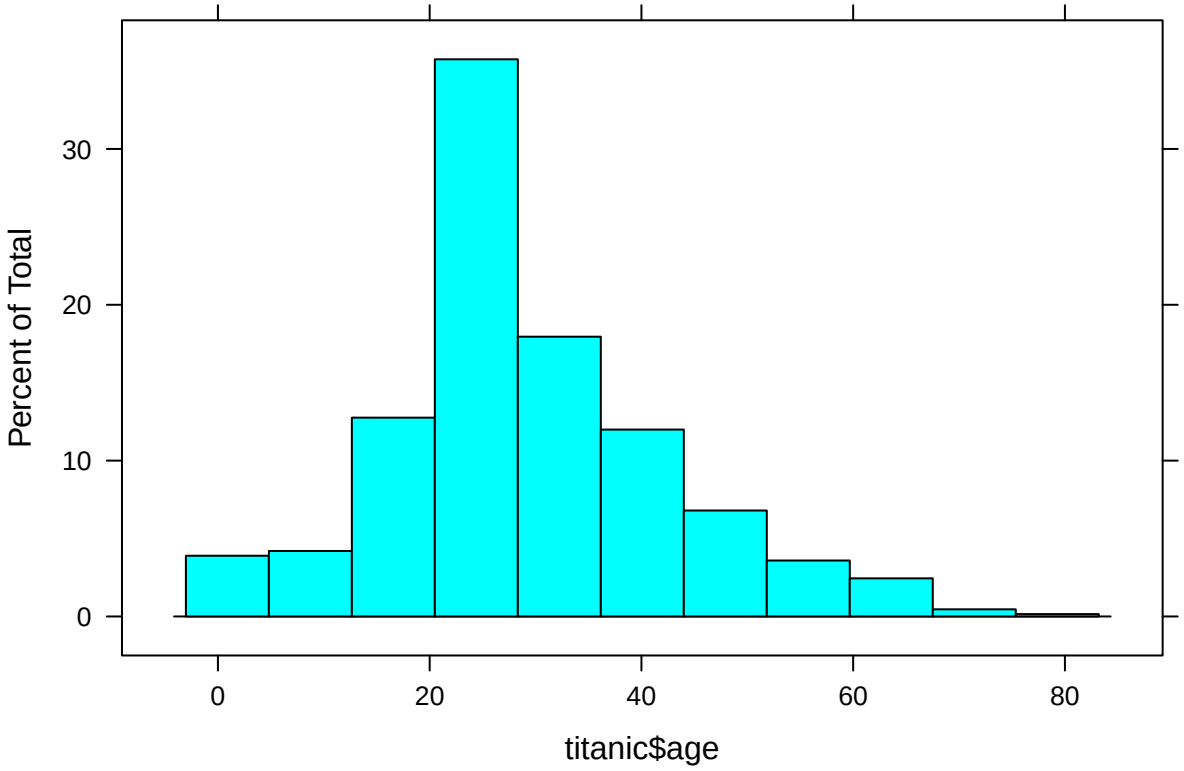
```
kable(ports)
```

South.Hampton	Cherbourg	Queenstown
914	272	123

First minor Machine Learning

As it is showed above, most of the variables have been cleaned and are ready for the prediction. There is though the column Age with several NAs in it (exactly 263). Unfortunately this seems to be an important criterion for survival and I can not give up on this. Since age usually follows a normal distribution, I could have been corrected the NA values with a median of the non NAs values. Nevertheless, I chose to run a first minor Machine Learning experiment to fill the NAs. Using the Regression Trees technique I created a fit model function to predict age from all other variables from the non-NAs (except survived, which is the final goal). Then I used predict function to fill the NAs, which gave me a age distribution that I show below.

```
histogram(titanic$age)
```



Splitting train and test set

Now that the data set is clean, I divided it into a training and a test set. It was common during the course to establish a proportion of 0.1 between training and test set. But here the entire data set is not so large (1309 rows), so I chose a higher proportion of 0.3. After separating both sets, I removed the "survived" color from the test set, in order to predict this information based on the training set.

```
kable(traintest[1,])
```

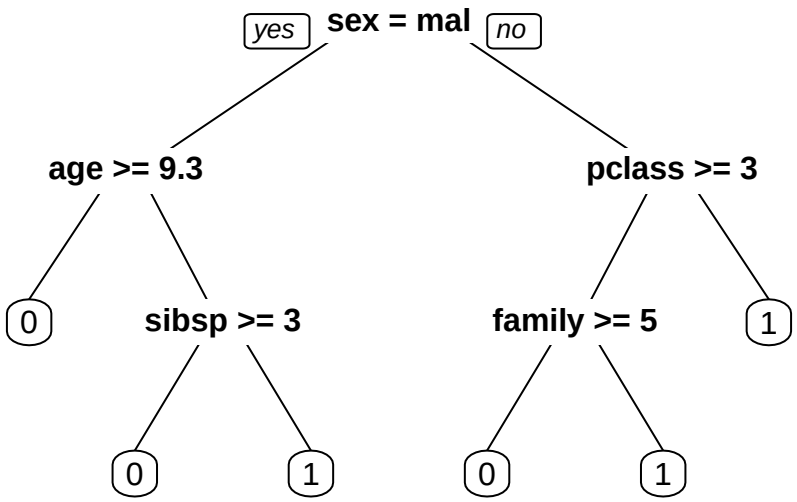
test	train
393	916

Regression Tree

Again I used the Regression Trees technique I created a fit model function to predict if the passenger has or has not survived from all other variables. Below I plotted the tree to show an example on who are the questions used by R for prediction.

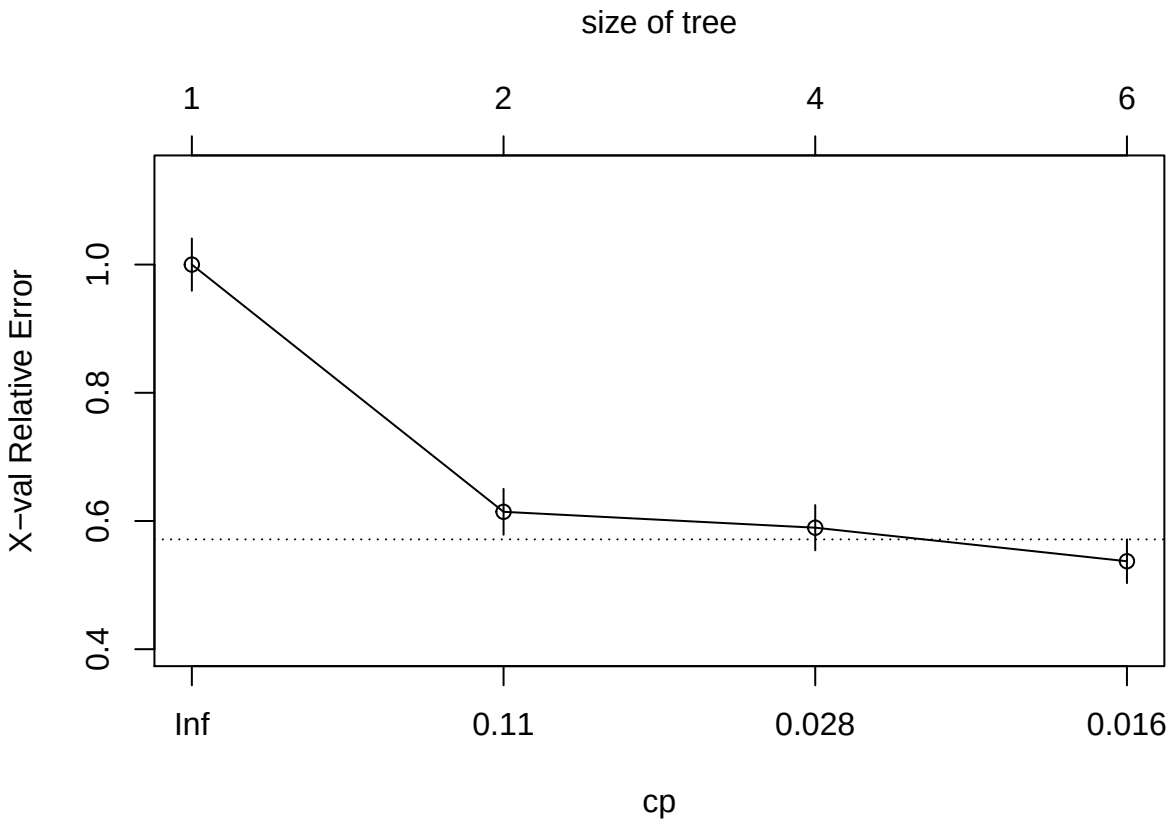
```
#I again made a fit modell using all variables available, this time to predict the data survived.
fit_survived<-rpart(survived~pclass+sex+sibsp+age+parch+fare+embarked+family,data=train,method="class")

prp(fit_survived)
```



If the model is plotted,it is possible to see how have R decided for the amount of braces of the regression tree. It has reached a total amount of 8 braces, but from 4 braces on the relative error has not decreased considerably.

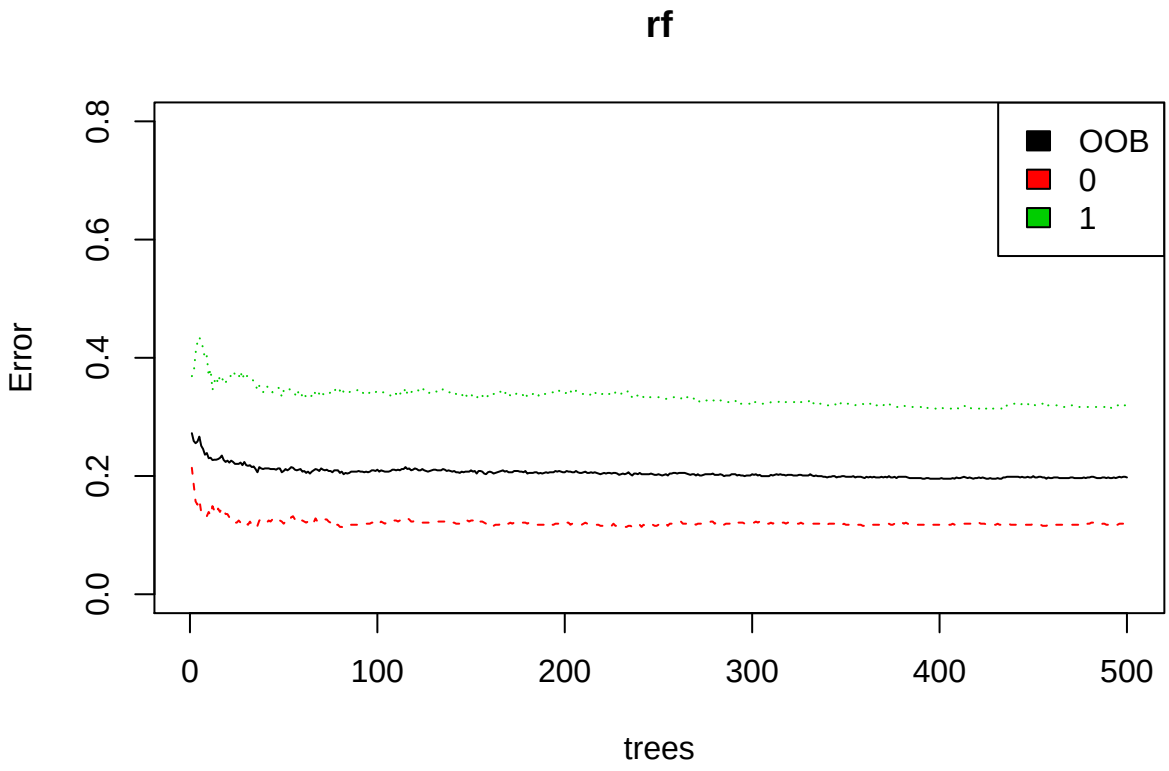
```
plotcp(fit_survived)
```



Random Forest

So I used the Random Forest technique, which is basically a combination of multiples regression trees, crossing multiple combinations. In the plot below we checked the errors from the random forest algorithm, in red for the non survivors, green for the survivors and black for the central line combining both errors. By this plot it is possible to say that our model predict better the destiny from those who have not survived than for those who did. It is explained because there are more people who have not survived the Titanic than otherwise, therefore the data available for the non survivors is larger, turning the prediction for them more accurate.

```
plot(rf,ylim=c(0,0.8))
legend('topright',colnames(rf$err.rate),col=1:3,fill=1:3)
```



Results

Below there is a table showing the minimum error reached for each outcome: 0.10 for non survivors and 0.33 for survivors, with a combined error of 0.19.

```
kable(error)
```

combined_error	non_survivor	survivor
0.1954148	0.1121157	0.3140496

Conclusion

At first view, the combined minimum error of 0.19 may not appear much impressive. I sustain otherwise, since one must keep in mind the characteristics of the data set. The provided information is general observations such Sex, Age, family on board, fare paid, class and port, which means a considerably little amount of data about the passengers of Titanic, which in its turn is also a relative little data set. Even though the Random Forest technique made possible to predict their destiny with satisfactory accuracy. A combined minimum error of 0.19 is relatively good prediction, and the prediction for non survivors can be classified as excellent: just 0.10.