Titanic Machine Learning (Kaggle Dataset)

using R, for my Own Data Science project on EDX

By:
Adrian Zinovei

Business Problem/Objective

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. 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 sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led 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.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. We ask you to apply the tools of machine learning to predict which passengers survived the tragedy. This problem is a famous problem on Kaggle for beginner machine learning enthusiasts.

The YouTube link associated with problem description is given here: **https://youtu.be/9xoqXVjBEF8**

Data Availability

The analysis must be performed using the following file format given:

1. **titanic_training.csv:** This csv(comma separated value) file is the given training set on which we have to train our machine learning model. The dimension of file is **[891rows*12columns]**. The file contains following fields:
   VARIABLE DESCRIPTIONS:

- Survival - Gives the chances of survival of each passenger. It's a categorical variable with two values (0 = Died; 1 = Survived)
- pclass - Name of Passenger Class with notation (1 = 1st; 2 = 2nd; 3 = 3rd)
- name - Name of passenger
- sex - Sex of passenger
- age - Age of passenger
- sibsp - Number of Siblings/Spouses Aboard
- Parch- Number of Parents/Children Aboard
- Ticket - ticket Number
- Fare - Passenger Fare
- Cabin - Cabin

- Embarked- Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

2. titanic_test.csv- This is the testing set on which the machine learning model trained on training set must be tested. The dimension of file is **[418rows*11columns].** However, both sets contain the same fields, but the survival field in test.csv is missing and this is what, has to be found using the model.

SPECIAL NOTES:

- Pclass is a proxy for socio-economic status (SES)
  1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

- Age is in Years; Fractional if Age less than One (1) If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

- Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
- Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
- Parent: Mother or Father of Passenger Aboard Titanic
- Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

- Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

Passenger	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	7.25		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.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S

This is how the training data looks like.

Approach

Overview

One could hypothesize from stories of the Titanic's sinking that a passenger's survival was heavily dependent upon two factors:

- Recognition of the possibility that the ship could sink
- Access to a lifeboat

Firstly, I searched through Wikipedia to know more about the disaster to know the problem better and thinking if it might help me making better insights. The Wikipedia says that the Titanic reportedly struck an iceberg at 11:40 pm ship's time. The majority of its 2,224 passengers and crew had likely retired to their respective cabins for the evening by that time. Those on the upper decks had a shorter journey to the lifeboats, and possibly access to more timely and accurate information about the impending threat. Thus, any data relating to one's location on the ship could prove helpful to survival predictions.

The Titanic was designed to carry 32 lifeboats, but this number was reduced to 20 (enough for about 1,180 people) for its maiden voyage -- likely a cost-cutting measure influenced by perceptions that the additional boats would clutter the deck of a ship deemed "unsinkable."

Given that constraint, it is not surprising that a disproportionate number of men were apparently left aboard because of a **women and children first protocol** followed by some of the officers overseeing the loading of lifeboats with passengers

Basic approach will be to

- To plot the fields of the training set corresponding to their survival chances to get better insights on data.
- The data is raw and various feature engineering can be done on the dataset to get better insight on the data. Including new features like Title, family size, individual ticket fare, child or not, mother or not, deck name, deck no. etc.
- Moreover, we can do binning wherever necessary (e.g binning age into 0-15years,15-30years and so on).
- The missing values can be treated by using median or mean, or if possible, deploy a ML algorithm on missing values to get better accuracy of model.
- At the end, after getting relevant fields we can deploy ML model on the set to predict the problem objective.

Packages Information and Analysis

randomForest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

In the random forest approach, many decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model.

An error estimate is made for the cases which were not used while building the tree. That is called an **OOB (Out-of-bag)** error estimate which is mentioned as a percentage *Syntax*

randomForest(formula, data)

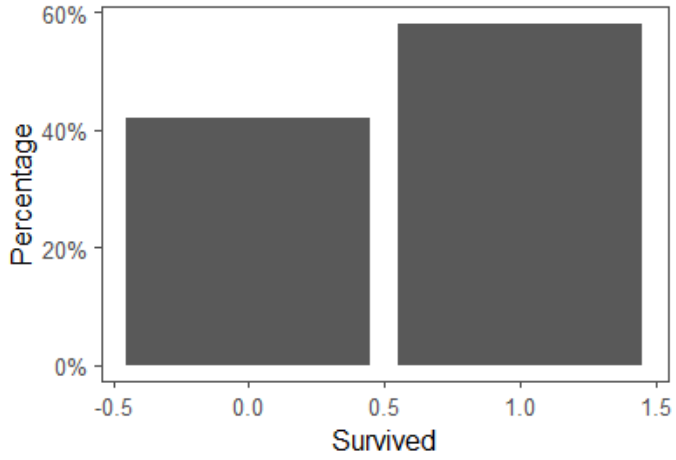
Steps

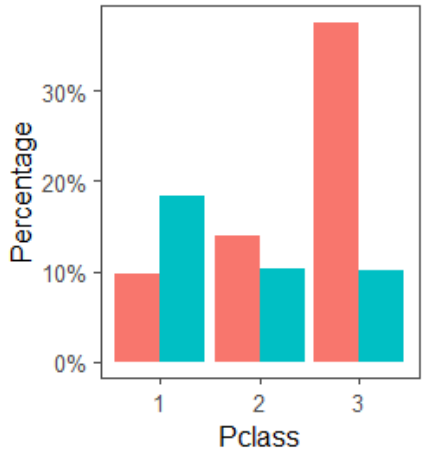
1. Install and load the relevant packages.
2. Getting the Data into R the data for the Titanic project is divided into two csv-format files:

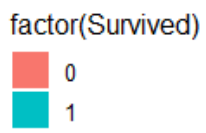
- **titanic_training.csv** (data containing attributes and known outcomes [survived or perished] for a subset of the passengers)
- **titanic_test.csv** (data containing attributes *without* outcomes for a subset of passengers)

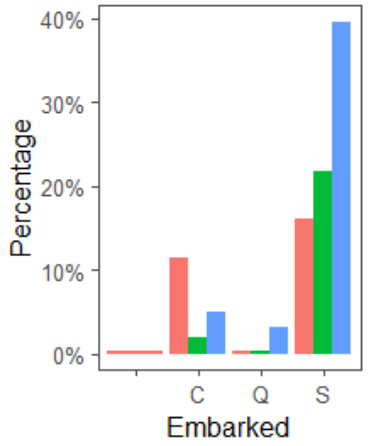
We import the above files in R using *read.csv()* and further bind the datasets by columns using *cbind()*, so that the changes or new variables creation in training set are created simultaneously in testing set.

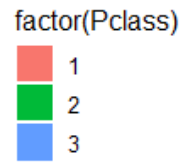
3. Plotting data by survivor passengers from different classes.

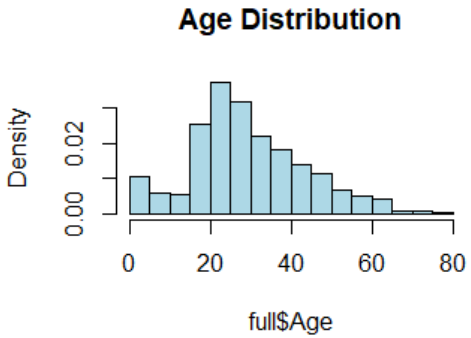




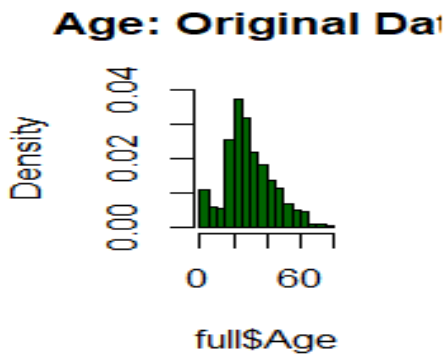








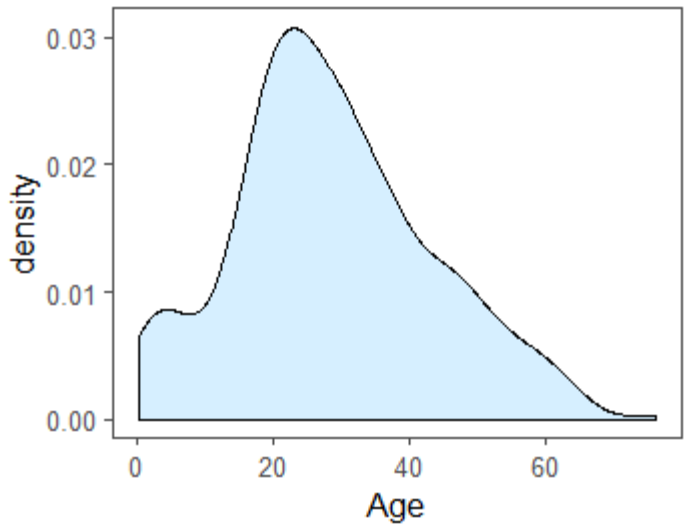
4. Make variables factors into factors: factor_vars <- c('PassengerId','Pclass','Sex','Embarked')



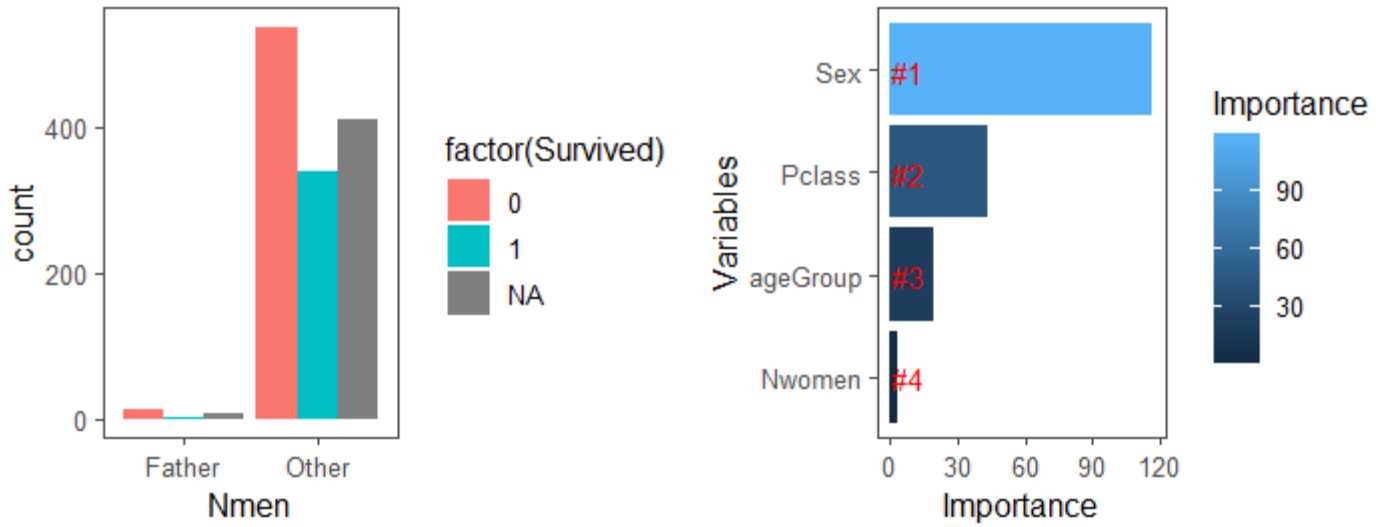

```
#Create new variable with different levels
completedata$ageGroup[completedata$Age<=2] <- "Baby"
completedata$ageGroup[(completedata$Age >2) & (completedata$Age<= 6)] <- "Toddler"
completedata$ageGroup[(completedata$Age >6) & (completedata$Age<= 12)] <- "Child"
completedata$ageGroup[(completedata$Age >12) & (completedata$Age<= 18)] <- "Adolescent"
completedata$ageGroup[(completedata$Age >18) & (completedata$Age<= 60)] <- "Adult"
completedata$ageGroup[(completedata$Age >60)] <- "Elderly"
```

5. ##Distribution of age of women

```
ggplot(completedata[completedata$Sex == "female", ], aes(x = Age)) + geom_density(fill = '#99d6ff', alpha=0.4) +
  geom_vline(aes(xintercept=median(Age)),
             colour='black', linetype='dashed', lwd=1) +
  theme_few()
```




```
#Creation of new attribute
#comparing mothers' chances across class
#Creation of new attribute
#Factorizing our predictive attributes
```



Fitting a Model and Conclusion

At the end, we will set a seed value and divide the training and testing set to their original self (number of rows). Then we will use a model for our data. There can be number of models and ensemble methods which can be deployed on a model.

Initially, we can use one of the simplest classification methods, Logistic regression for this problem too. We can start simple by passing essentially the features provided in the raw training data through the R function for fitting general linearized models (glm). To assess this

first model and the various binary logistic regressions that will appear in its wake, we will use the chi-square statistic, which is basically a measure of the *goodness of fit* of observed values to expected values.
In my code, however, I have started from using an ensemble method randomForest. The basic approach is same. Initially, we need to include the raw parameters of training data in the model and then select the appropriate parameters using the variable importance fit *varImpPlot()* . Later, we can use *predict()* on the test set to predict the results .
In order to increase the accuracy, we can use conditional inference tree random forest on the same.

Various other techniques are also used to increase accuracy, like cross validation, boosting techniques, hypertuning methods, etc. But there is not one specific model which is superior to the others, it is an iterative process and depends on the problem statement.

Result:

The result is saved in "**resultfinal_AdrianTitanic.csv**" with detailed predictions for passengers who survived and not.