Titanic Machine Learning (Kaggle)

using R

By:
Anish Anand

Titanic Machine Learning

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. In particular, 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

Titanic Machine Learning

Data Availability

The analysis has to be performed using the following file format given:

1. **train.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)

Titanic Machine Learning

2. test.csv- This is the testing set on which the machine learning model trained on training set has to 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

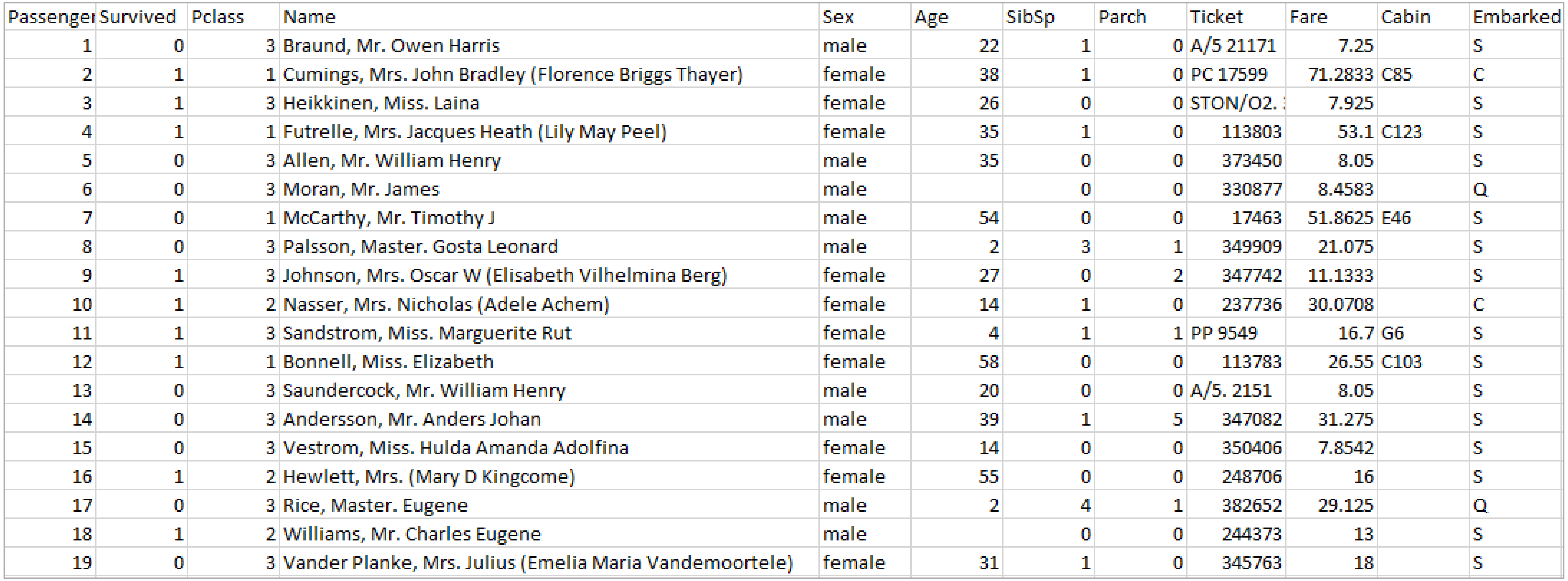
- Age is in Years; Fractional if Age less than One (1)
  If the Age is Estimated, it is in the form xx.5

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

Titanic Machine Learning

- 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.

Titanic Machine Learning

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 so as 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."

Titanic Machine Learning

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.

Titanic Machine Learning

Packages Information

1. Amelia :

It was named after American Aviation pioneer and author Amelia Earhart. This package uses a technique called ‘**multiple imputation**’ to impute the missing values. Amelia implements bootstrapping-based algorithm that gives essentially the same answers as the standard IP or EMis approaches, and is considerably faster than existing approaches and can handle many more variables

Assumptions Amelia makes:

Assumption 1 Data-set is multivariate normal

Assumption 1 Missing data values belong to MAR(Missing At Random).

Step 1: Installing and calling the package

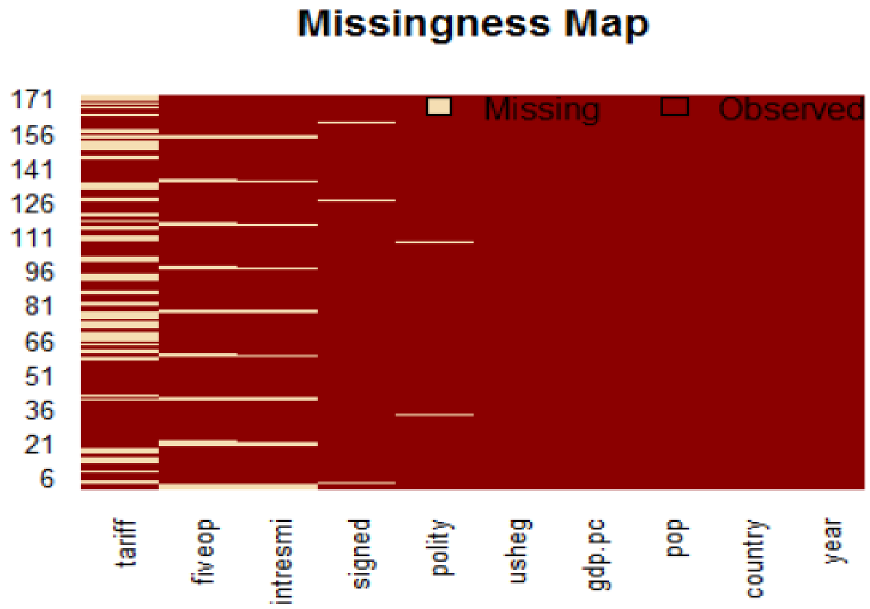
> *Install.packages(Amelia)*

> *library(Amelia)*

Step 2: Check whether the data-set has missing values

> *missmap(data)* #data is the data-set you use(here free trade)

##missmap for free trade data will look as follows



Titanic Machine Learning

For better output, convert the categorical values into numerical factors.
Eg. If the attribute is Gender with classes ‘male’ and ‘female’. Then this has to be replaced with factor levels male=0 and female=1, so that the attribute now will have 2 levels namely 0 and 1.
Similarly, change all the categorical values.

Step 4: Calling Amelia function
#For explanation purposes I have used “free trade” data-set available in Amelia package.
> Completed_data<-amelia(data,m=3,ts=”year”, cs=”country”, p2s=0, ords=”polity”, noms=”signed”, idvars=c(“year”,”country”))
#Variable explanation
data – the data-set with missing values.
m=3 number of imputed data sets we need…by default m=5
ts- attribute that is time series class
p2s print to screen. Decides what we to see in console. p2s=0 will not display the details of chain length during imputation. p2s=1, displays only the chain length. p2s=2 displays everything.
Ords, noms – if you want to compute factor levels, the attribute that has ordinal/nominal data type has to be mentioned to avoid continuous value imputations.
idvars – the attributes we need to retain but not used in missing data imputation models are mentioned here.

3. rpart (Recursive Partitioning)
The rpart programs build classication or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. The types of endpoints that rpart handles includes

Titanic Machine Learning

- classications (such as yes/no),
- continuous values (such as bone mineral density),
- poisson counts (such as the number of fractures in Medicare patients), and
- survival information (time to death/last known contact).

The rpart library includes tools to model, plot, and summarize the end results.

Attach the library so that the functions can be found.

library(rpart)

Decide what type of endpoint you have

- Categorical ==> method == "class“
- Continuous ==> method == "anova“
- Poisson Process/Count ==> method == "poisson“
- Survival ==> method == "exp"

Fit the model

fit <- rpart(y ~ x1 + x2 +x3 …, data=data1, method='poisson')

Print a text version of the tree

print(fit)

Print a summary which examines each node in depth

summary(fit)

Titanic Machine Learning

4. party: A Laboratory for Recursive Partytioning

The core of the package is ctree(), an implementation of conditional inference trees which embed tree-structured regression models into a well defined theory of conditional inference procedures. This non-parametric class of regression trees is applicable to all kinds of regression problems, including nominal, ordinal, numeric, censored as well as multivariate response variables and arbitrary measurement scales of the covariates. Based on conditional inference trees, cforest() provides an implementation of Breiman's random forests. The function mob() implements an algorithm for recursive partitioning based on parametric models (e.g. linear models, GLMs or survival regression) employing parameter instability tests for split selection. Extensible functionality for visualizing tree-structured regression models is available.

5.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, a large number of 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)

Titanic Machine Learning

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:

- **train.csv** (data containing attributes and known outcomes [survived or perished] for a subset of the passengers)
- **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. Data Munging

It is the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data, which may include further munging, data visualization, data aggregation,etc.

- First lets see the **missing values** in the data and treat them respectively. I'll use the **missmap** function from the **Amelia package** to display those(pictorial representation). The summary function in R also gives the count of NA's(missing values).
- we see that proportion of Age "missings" is likely small enough for reasonable replacement with some form of imputation, the cabin missings seem too extensive to make reliable imputation possible .
- Then we can also see relation of other fields to survival by plotting using histogram, boxplot, etc , to get a clear picture of how the fields affect the survival of a person on the ship.

Titanic Machine Learning

4.Feature Engineering

It is the process of using domain knowledge of the data to create features that make machine learning algorithms work. We will use feature engineering in our set to create new features, i.e title of person onboard, number of family members, child, mother, etc. so as to see, if the feature will be important enough in our model.

5. If we see clearly, we can also see that there are missing values in age of passengers. There are basically 3 approaches to deal with these types of problem :

- If number of missing values entries are very less as compared to that of size of data, we can remove them using *na.omit().*
- In other cases, we can replace the missing values using either mean, median or mode (seeing which of them could possibly relate to the distribution)
- However, the third case which we can use is to deploy an individual machine learning algorithm for finding missing values of the age of passenger using the relevant features. This will help in reducing error to great extent.

6. Fitting a Model

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.

Titanic Machine Learning

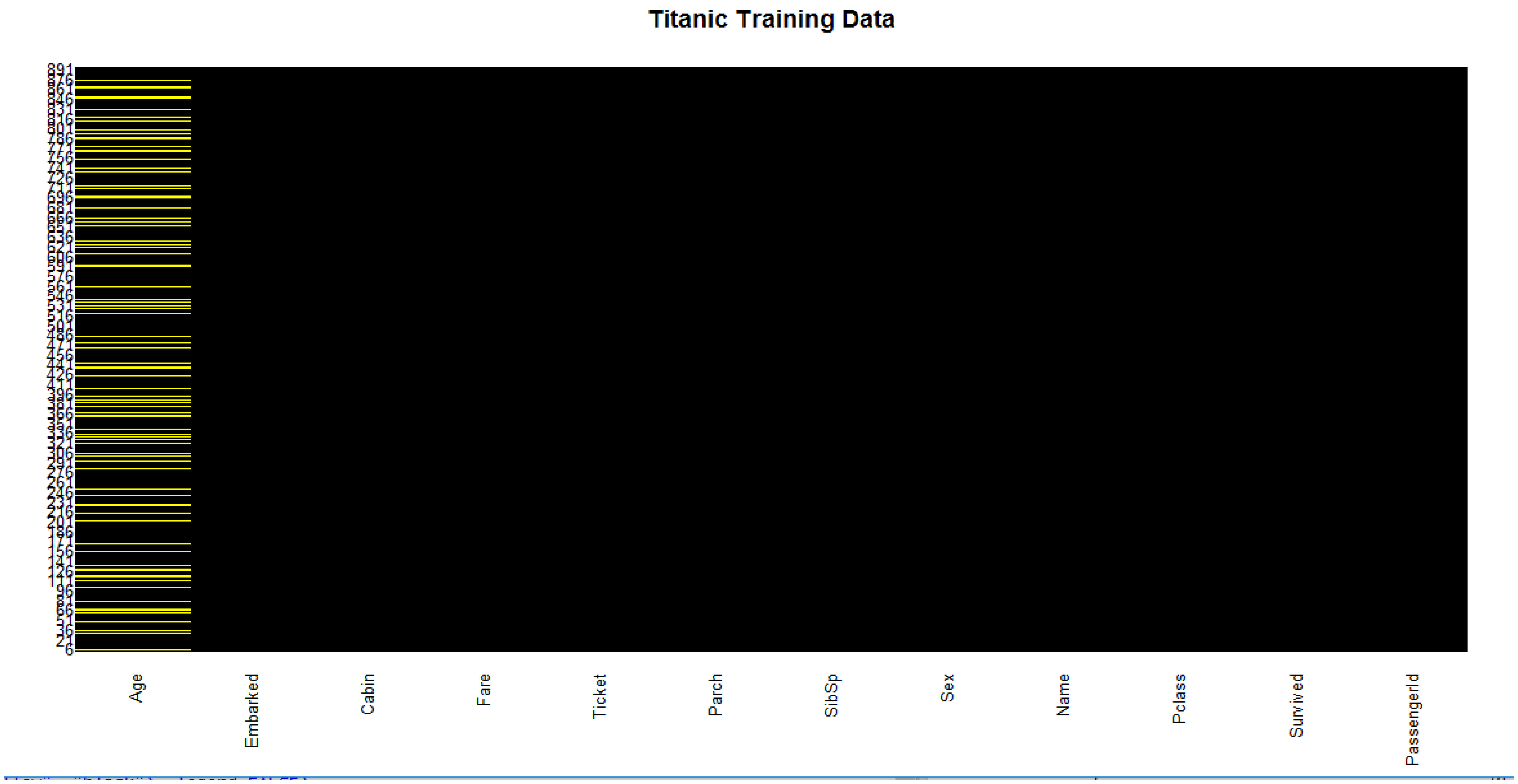
In my code, however, I have started from using an ensemble method randomForest. The basic approach is same. Initially, we need to include the raw parameters of training data in the model and then select the appropriate parameters using the variable importance fit *varImpPlot()* . Later, we can use *predict()* 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.

Titanic Machine Learning

Visualizations

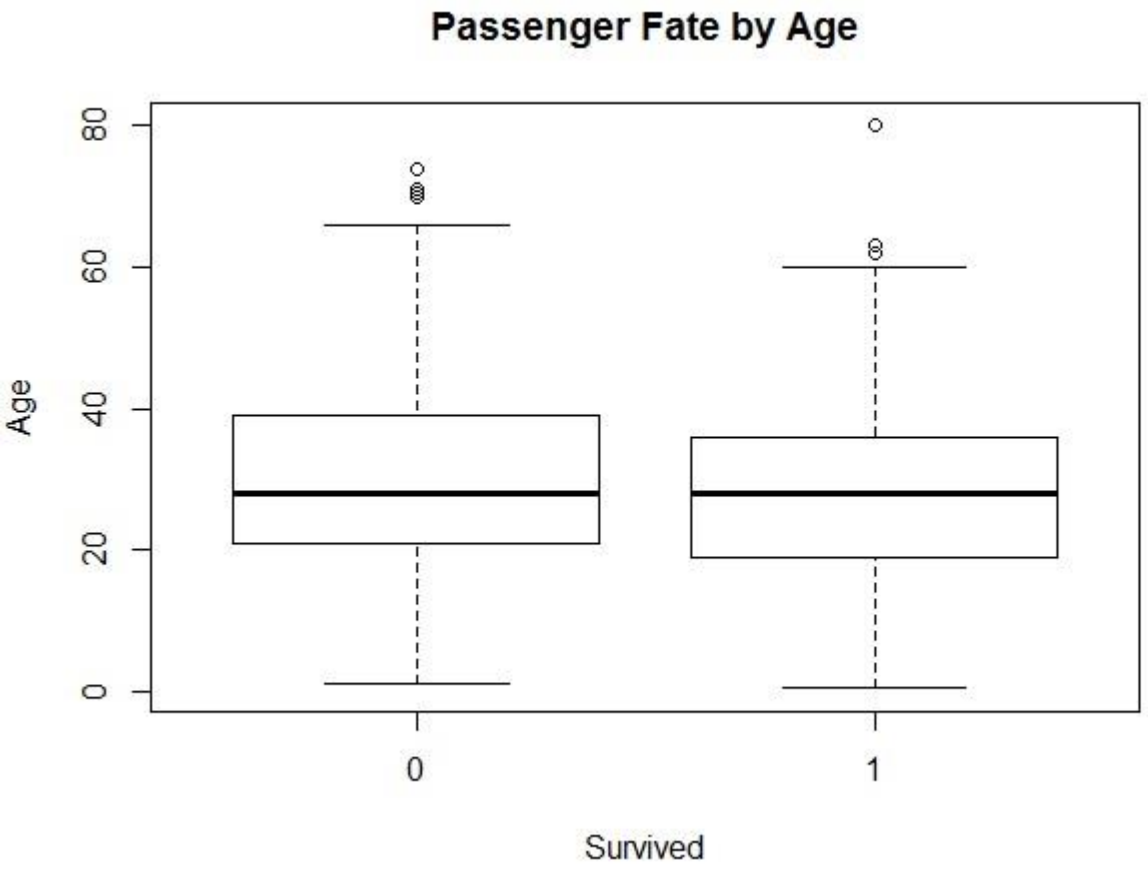
- **MISSING VALUE VISUALISATION USING AMELIA PACKAGE**

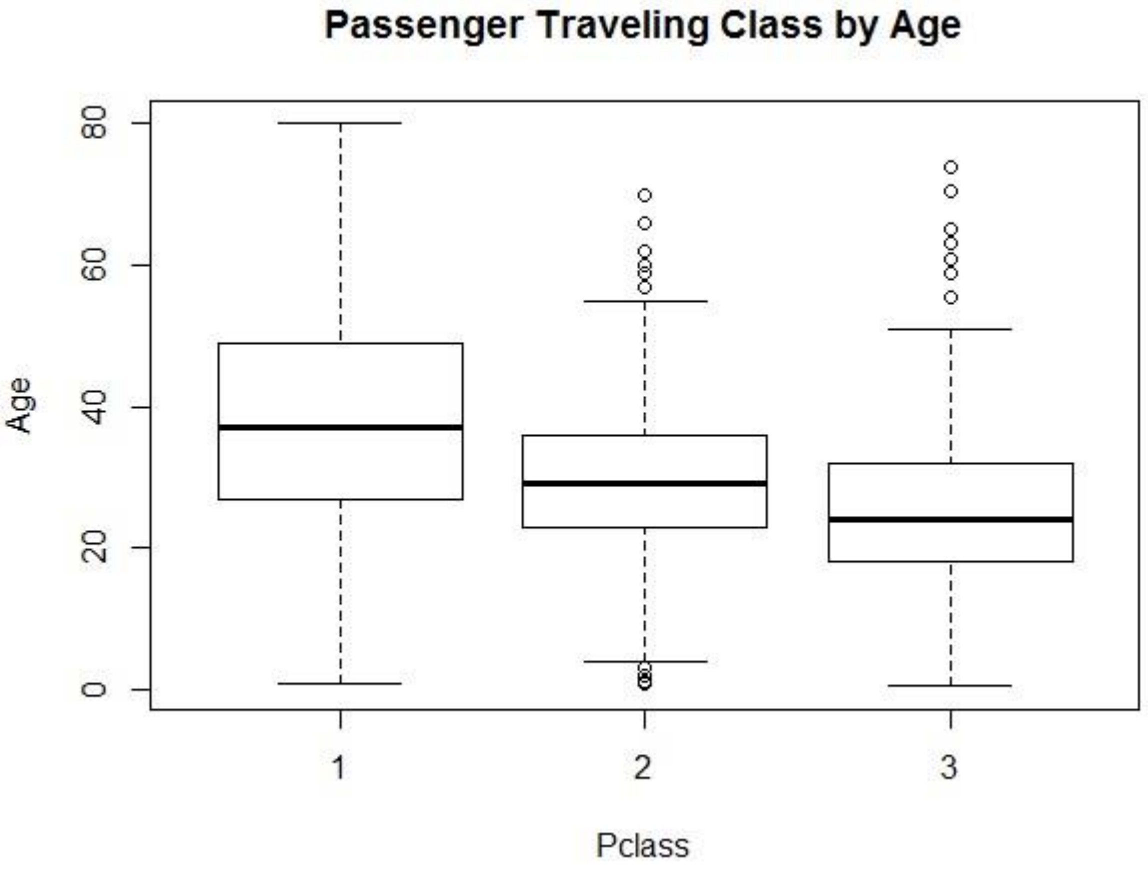


Prepared by : Anish Anand

Titanic Machine Learning

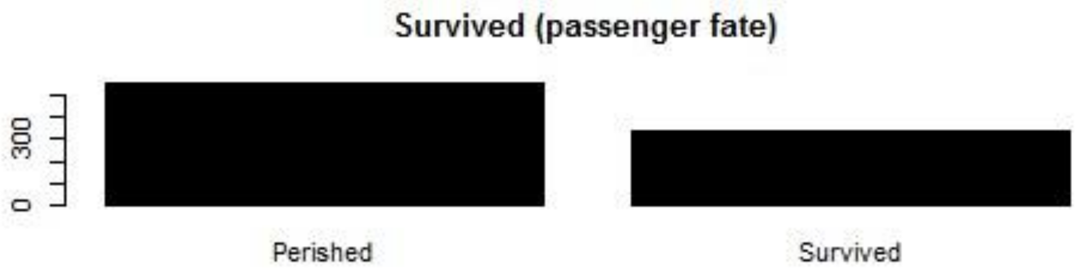
- **BOXPLOTS**



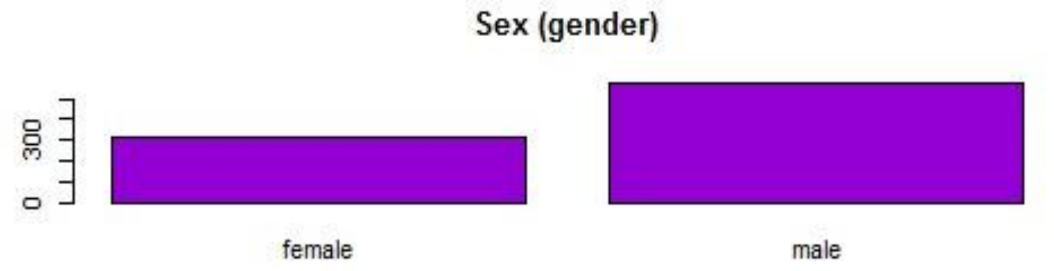


Titanic Machine Learning

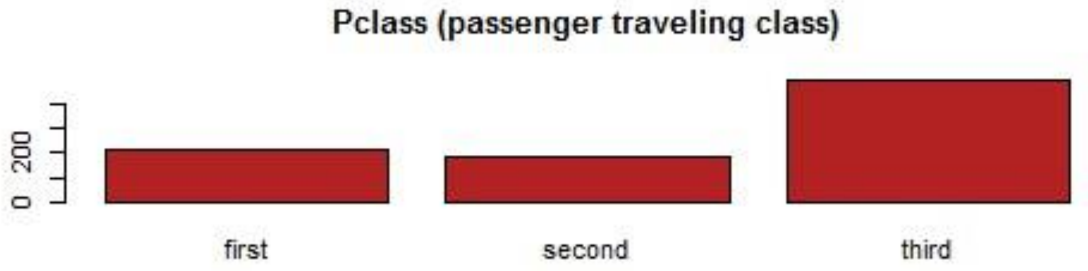
- **HISTOGRAMS**



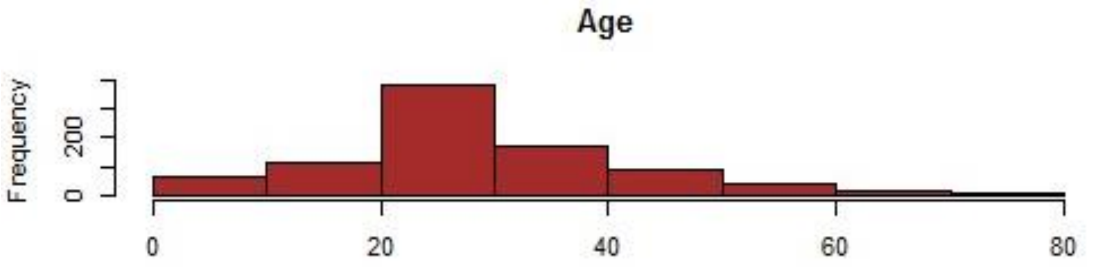
Survived (passenger fate)



Sex (gender)



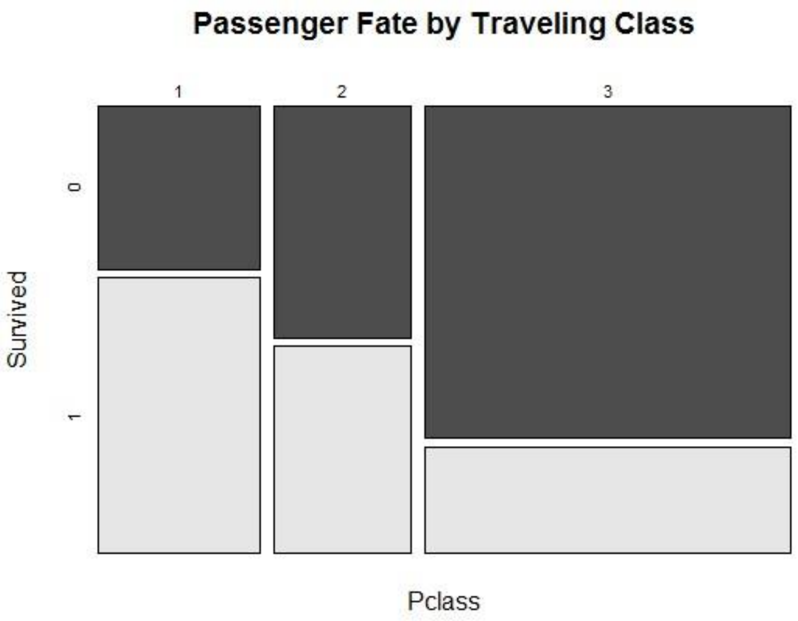
Pclass (passenger traveling class)

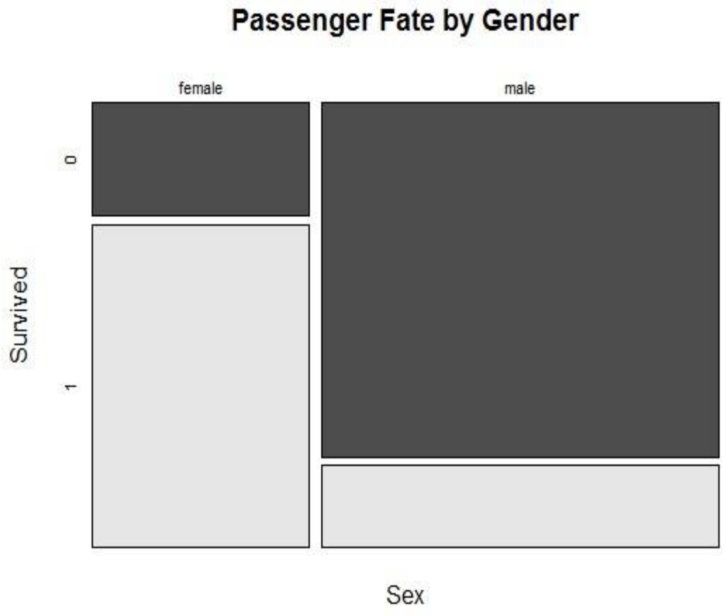


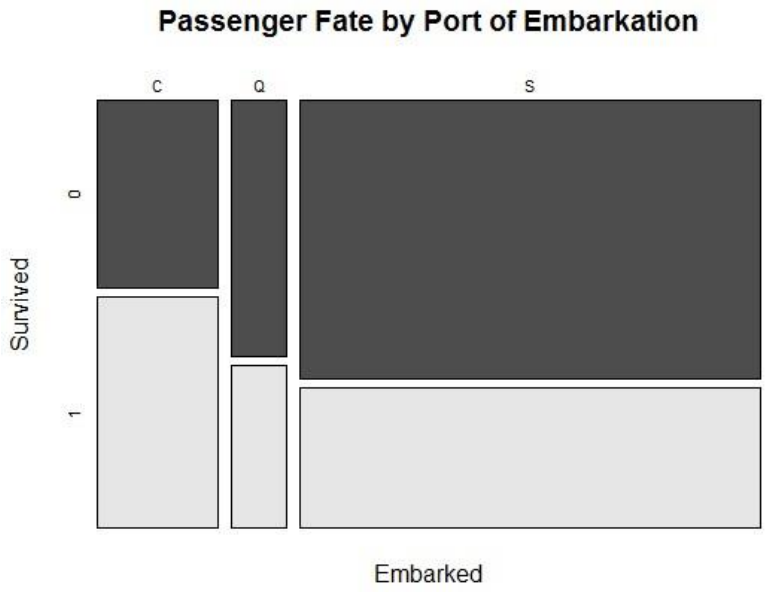
Age

Titanic Machine Learning

- **MOSAIC PLOTS**


Passenger Fate by Traveling Class
Survived
Pclass


Passenger Fate by Gender
female
male
Survived
Sex


Passenger Fate by Port of Embarkation
C
Q
S
Survived
Embarked