International Journal of Engineering & Technology, 7 (3.33) (2018) 279-284

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Designing a Data Algorithm Prediction Model based on R

Qamber Ali Yagob Ali[1], Jun Lee[2], Hyoung-Jin Kim[3]*

[1]*Department of IT Applied System Engineering, Chonbuk National University*
[2]*Department of IT Applied System Engineering, Chonbuk National University*
[3]*Department of IT Applied System Engineering, Chonbuk National University*
**Corresponding author E-mail:kim@jbnu.ac.kr*

Abstract

This study is about data analysis and prediction model using R open source language, R is a language and environment for statistical computing and graphics so it's a full function programming language. In this study through using the Titanic datasets, we created a model that predict the survival rates of the test dataset, we used the Train dataset that has the survived variable with levels of "0"(perished) and "1"(survived) data and the test dataset with no survived levels to predict the survival rates on the test. In this study, we used the R functions, packages, and machine learning algorithms that provided in R, to combine, analysis and splitting the data we created utility functions to make features and predictive potential values for our new variables. We Analyzed the Training data set by cross-validation with 82% accuracy, visualization, and decision tree and then leveraged it to test data to predict the survival rate on test data.

Keywords: *Splitting data, utility functions, Cross validation, Analysis, Accuracy, Visualization, Prediction*.

1. Introduction

Globally, technology and development are constantly changing, and every year new technologies and method are being developed every year. Among them, data and data analysis techniques have become of interest to companies and schools and become an important issue. Much research has been done on prediction methods and prediction models for analyzing and analyzing big data and general data. This is because mass data itself requires a high probability of accuracy along with fast technology and analysis that finds the correlations and potential values of variables and proves the accuracy. In this paper, we use R statistical programming language to create utility functions to discover the correlation of valuable variable and splitting the data to knows what's going on between survival and perished using RStudio, which has recently been attracting attention as a big data analysis tools, We will analyze the Titanic data to create a survival prediction model. [1] In this study, we present a prediction model using the algorithms and packages of ggoplot2, random forest, rpart, caret, varImpPlot provided by R.

2. Analyze existing data

With the rapid development of IT technology, data production is so fast that momentarily goes beyond general data to big data. Thus, mass data is being an issue to corporations and public institutions, and various analytical techniques have been developed. In this study used the data of Titanic ship that crashed with an iceberg in April 15.1912 and almost 2.200 passengers have perished and 1.514 passengers survived, Dataset configurations are separated by rows and columns, columns are variable names of data, and rows are data and data types in variables[2]. We have two datasets of Train and test, the Train dataset has the survival variable that indeed we are analysing to find out how many peoples perished, what kind of peoples has been perished, what is going on with the survival rates. Train data set configuration is shown in (Figure 1).

```
> str(train)
'data.frame':   891 obs. of  11 variables:
 $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
 $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
 $ Name    : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket  : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin   : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
>
```

Fig. 1: The train data configuration

The Titanic and what differential it's our train data from our test data is the fact that we have these values that yes Mr. Own Bronde did not survive but Mi. Jon did survive and here all the data points that are associated (Figure 2.)

Filter

	Survived	Pclass	Name	Sex	Age	SibSp
1	0	3	Braund, Mr. Owen Harris	male	22.00	1
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1
3	1	3	Heikkinen, Miss. Laina	female	26.00	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1
5	0	3	Allen, Mr. William Henry	male	35.00	0
6	0	3	Moran, Mr. James	male	NA	0
7	0	1	McCarthy, Mr. Timothy J	male	54.00	0
8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1
12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0

Fig. 2: The sample of train data

The test data set configuration is the same as the train data set, but there is no survival data variable (Survived) in the test dataset. Test data set configuration is shown in (Figure 3).

```
'data.frame':   418 obs. of  10 variables:
 $ Pclass  : int  3 3 2 3 3 3 3 2 3 3 ...
 $ Name    : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 58 5 104 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
 $ Age     : num  34.5 47 62 27 22 14 30 26 18 21 ...
 $ SibSp   : int  0 1 0 0 1 0 0 1 0 2 ...
 $ Parch   : int  0 0 0 0 1 0 0 1 0 0 ...
 $ Ticket  : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
 $ Fare    : num  7.83 7 9.69 8.66 12.29 ...
 $ Cabin   : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Embarked: Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 3 2 3 1 3 ...
```

Fig. 3: The test data configuration

The data has been split into two groups:
Training set (train.csv)
Test set (test.csv)
The training set is used to build machine learning models. We used the feature engineering to create new features [3].
The test set would be used to see how well our model performs on unseen data. For each passenger in the test set, use the model the trained to predict whether or not they survived the sinking of the Titanic [4].
To better understand the data type and data structure in the variables, the following (Table 1) is used.

Table 1: The data variables structure

Variable	definition	key
Survived	Survival	Survived= (0=no, 1=yes)
Pclass	Ticket level	Pclass=1,2,3 A proxy for socio-economic status, 1 = 1st, 2 = 2nd, 3 = 3rd
Name	personal information	Baumann, Mr. John D
Sex	Sex	Male, female
Age	Age	Age
Sibsp	Family size	Sibsp: The dataset defines family relations , Spouse = husband
Parch	Passenger with family	Parch: The dataset defines family relations in this way. Parent = mother, father, Child=daughter, son, Some children, therefore parch=0 for them
Ticket	Ticket number	Ticket number
Fare	Ticket price	Passenger fare
Cabin	Cabin number	Cabin number
Embarked	Port of embarked	C = Cherbourg, Q = Queenstown, S = Southampton

3. Analysis method

Based on the train data set of Titanic, we used machine learning algorithms and packages applied to R statistical programming language, the train data set have 891 observations 11 variables, 594 peoples perished and 342 survived but, as mentioned above the test data set does not have the survival variable. to use the machine learning algorithms to use this data and learn from the data the patterns between who survived and who did not, in machine learning terms what we're trying to do here is build a classification model literally we're trying to tell the computer to build an asset from this data that enable it to either say with a certain amount of certainty to someone not survived do they perish or do they survived the sinking of the Titanic and what that allows the computer to do is to use a machine learning algorithms to use this data and learn from the data the patterns between who survived and who did not, this is a very key point we use machine learning or data science techniques when the pattern in the data signal that we're trying to understand is either too complicated or it would take too long for human being to divine it to do examine the data by hand understand the pattern create algorithms or logic to actually implement the patterns and also do the classes and do the classification, that's what machine learning really about. The first thing we're going to do, combined the Train and test datasets by using rbind( ) function to combine the data by rows and columns to handle the missing values. After loading, the datasets using the following code to combine this two into one set we need to add a variable to the test set "Survived" variable with None values to the test set to allow for combining data sets by using the data frame() & rep()functions, the data frame() function it allows us to create a data frame and rep()function it allows to replicate elements of vectors and lists .

test.survived <- data.frame(Survived = rep("None",nrow(test)), test[,])

And now we have test.survived data frame with 11 variables and 418 observations as will the train data frame have 11 variables now our datasets have the same variables that allow us to combine the data sates, the (Figure 4) is shown datasets with variables and observations.

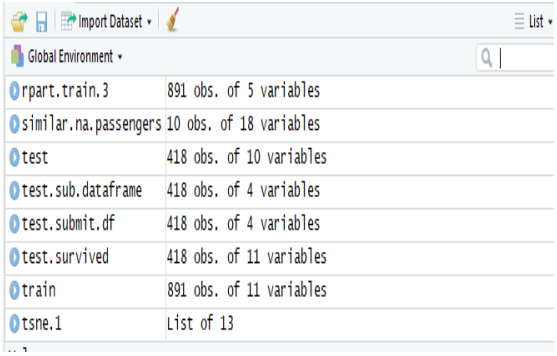


Fig. 4: The sample of datasets

Using the fallowing code to combine the datasets and what this is doing is, take the train data frame 891 observations of 11 variables as a table and append to it row by row the test.survived variables which are 418 observations. Now what that should give us, in the end, is a combined data frame (table) that has 1309 observations of 11 variables. (Figure 5) shown the combined data frame (data.combined).

data.combined <- rbind(train, test.survived)

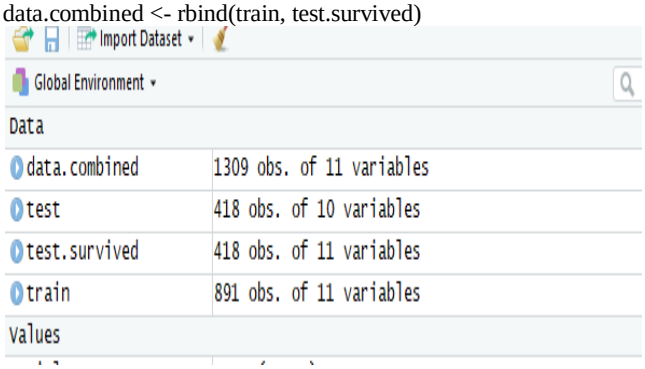


Fig. 5: The sample of combine data frame

In the existing data before to improve the data and adding the variables with potential values, the machine learning algorithms giving us 20.99% error rate that meant we 79% accuracy as shown in (Figure 6).

```
Call:
 randomForest(x = RF.train.1, y = RF.label, ntree = 1000, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 1

        OOB estimate of  error rate: 20.99%
Confusion matrix:
    0   1 class.error
0 536  13  0.02367942
1 174 168  0.50877193
```

Fig. 6: Accuracy of existing data

3.1. Proposal analysis

To plot the data as graph, use the fallowing code to convert the survived variable as factor.
data.combined$Survived <- as.factor(data.combined$Survived) and then use the ggplot2 package to plot survival rates by Pclass variable. [Fig 3] shows the result of survival by Pclass variable, the "0" red color mean perished and "1" blue color means survived. Using the following code to plot the survival rate by Pclass which have three levels (1, 2, 3) means the Pclass variable have three values. In section we used the ggplot2 package, ggplot2 is a data visualization package for statistical programming language R. Visualize the association of variables by x, y labeling. The basic elements that make up the grammar of ggplot2 are as follows.
1 data frame, the raw data that you want to plot.
2 Exterior elements such as color, size (aes).
3 The geometric shapes that will represent the data, points, lines, shapes, etc. However, geom_bar can be used according to the data.
4 stat_ Statistical summaries of the data that can be plotted, such as quantiles, fitted curves (loess, linear models, etc.), sums and so on [4].

```
ggplot(train, aes(x = Pclass, fill = factor(Survived)))+
geom_bar(width = 0.5)+
xlab("Pclass")+
ylab("Total Count")+
labs(fill = "Survived")
```

The survival rate is plotted using gglot2, and the survival rate is highest when the result graph Pclass = 1, Plass = 2 and the survival rate was predicted to be 50 to 50. If Pclass = 3, many people died. Survival rate graph is shown in (Figure 7) [5].

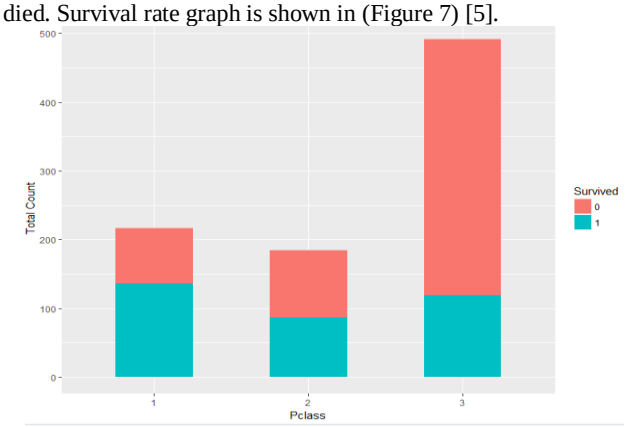


Fig. 7: The Survival rates by Pclass

If you look at the data in the variable, the Name variable is Mr., Miss., Mrs., and Master. There seemed to be a lot of information and possibility. In addition to the name, you can get information such as gender, age, marital status, occupation, etc. in the Name variable. The data in the Name variable is shown in (Figure 8).

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 310128
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450
6	0	3	Moran, Mr. James	male	0.00	0	0	330877
7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463
8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742
0	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736
1	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549
2	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783
3	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151
4	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082
5	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406
6	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706

Fig. 8: Shown the values of name variable

Used the next code to create a utility function to add a new variable named title with predictive potential values to the combined data frame

```
extractTitle <- function(Name) {
Name <- as.character(Name)
 if (length(grep("Miss.", Name)) > 0) {
 return ("Miss.")
 } else if (length(grep("Master.", Name)) > 0) {
return ("Master.")
} else if (length(grep("Mrs.", Name)) > 0) {
return ("Mrs.")
 } else if (length(grep("Mr.", Name)) > 0) {
 return ("Mr.") [6]
 } else {
return ("Other")
}   }
titles <- NULL
for (i in 1:nrow(data.combined)) {
  titles <- c(titles, extractTitle(data.combined[i,"Name"]))
}
data.combined$title <- as.factor(titles)
```

Now we add a new variable to data.combined data frame and use the ggplot package to gate more detail of correlation and easy understand of survival rates by Pclass and title by adding to data frame new variable we recognize lots of deference. the (Figure 9) shows the survival rates of Master, Mr., Mr., Miss., Mrs., by Pclass and title variables[5].
In the following code, we entered the entire test and train data, but only the first 891 rows were used because there are survival labels only on the train set. In order to show the newly created variable (title) functionality, the ggplot package was used to visualize the survival rate by analyzing the title, pclass, and survived directions [6].
The plotting of data shows that the Pclass is 1 and does not seem to have perished the master boys, Miss. The survival rate is good even if there is. And if Mrs. Pclass is 1 more than 50% died, and Mrs. has a high survival rate. If you have Pclass = 2, Master (young male children), Miss (single female) survival rate is very good. However, Mr. The survival rate is not good. Or Mrs. is more than 80% survival rate. Pclass = 3 Master (Young males) Survival rate is 50 to 50, and Miss has shown the same survival rate. Survival rate of 20% is very bad. Also, Mrs. Survival rate is more than 50%. We extended the relationship between Survived and Pclass by creating a Title variable and adding it to the data set [7].

```
ggplot(data.combined[1:891,], aes(x=title, fill = Survived)) geom_bar() +
facet_wrap(~Pclass) +
ggtitle("Pclass") +
xlab("title") +
ylab("Total Count") +
labs(fill = "Survived")
```

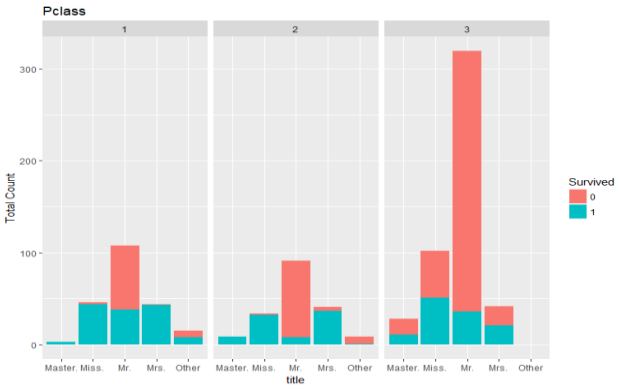


Fig. 9: The Survival rates by Pclass & title

Survival rates were visualized by comparing Sex, Pclass, and Survived orientation relationships with the title analysis in order to analyze the data and to find the survival rate correlation. We will use Ggplot2 to analyze the survival rate based on Pclass, Sex. The red color of the graph means perished, the blue color of it means survival [4].

Pclass 1 Women's survival rate is over 90%, and men 45% perished.

Pclass 2 Survival rate of female is 90%, but 80% of men perished

Plass 3 Women 55% is Survival and Men Survival Rate is not too good. Survival, Pclass, and Sex correlations were visualized as shown in (Figure 10).

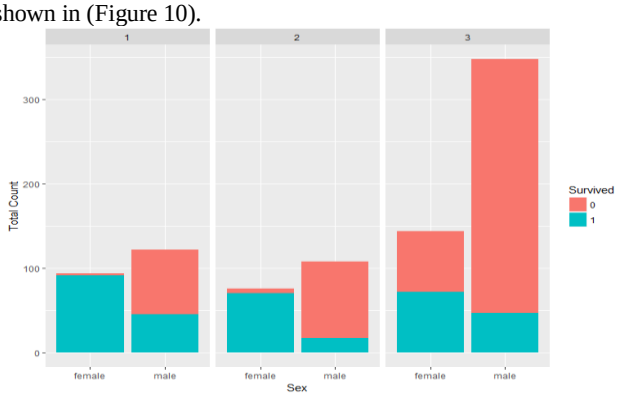


Fig.10: The survival rates by Pclass and sex

Now we are going to handling data to get more accuracy by adding a new variable called family size by using the next code. data.combined$family.size<- as.factor(Temporary.SibSp + Temporary.Parch + 1),

adding new variable may make more predictive, use the summary function to knows what going on with family size, summary(data.combined$family.size) In the following (Figure 11), you can see 1 ~ 11 people traveling together.

```
> summary(data.combined$family.size)
  1   2   3   4   5   6   7   8  11
790 235 159  43  22  25  16   8  11
>
```

Fig. 11: The sample of family size

By the adding family.size variable to the data frame used ggplot to Visualized the survival rates Figure 11 shows the survival rate that the big family is not good to survive ggplot(data.combined[1:891,], aes(x=family.size,fill=Survived)+
geom_bar() +
facet_wrap(~Pclass + title) +
ggtitle("Pclass, Title") +
xlab("family.size") +
ylab("Total Count") +
ylim(0,300) +
labs(fill = "Survived")

The results of the analysis are as shown in (Figure 12)

If Pclas = 1 & Title = Master, there is no one who perished regardless of family size. If Pclass = 1 & Title = Miss. The family size may be four looks some peoples perished. If Pclass = 1 & Title = Mr. More than 50% of passengers traveling alone have perished, and the survival rate of two or more family members is not good. Pclass = 1 & Title = Mrs. If there are less than three family members, it is safe, but those who travel with more than three family size die. Pclass = 2 & Title = Miss. The survival rate is very good regardless of the family member. Pclass = 2 & Title = Mr. 90% of passengers who traveled alone have perished and family size as big as many people perished. the passengers who's Pclass = 2 & Title = Mrs. looks no one perished survival rate is good. Pclass = 3 & Title = Master, and family size is more than four have perished. Pclass = 3 & Title = Miss. The passengers alone are seen as surviving 60%, but if the family size is more than two the survival rates is getting worse. Pclass = 3 & Title = Mr. Passengers who travel alone have survived more than 15%, and if family size more than two the survival rate getting worse. Pclass = 3 & Title = Mrs, if family size is four, survival rate is 50 to 50 bur if family size is more than 4 people as well as many people perished.

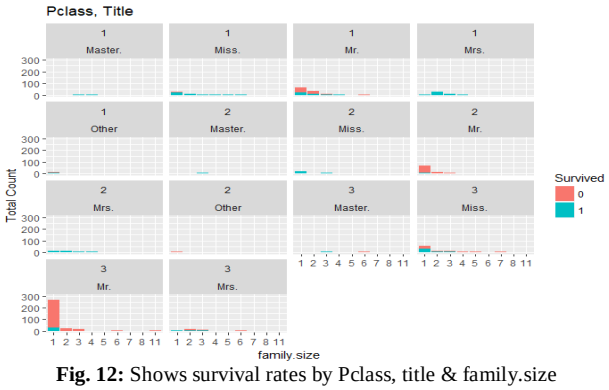


Fig. 12: Shows survival rates by Pclass, title & family.size

The following code uses the ggplot2 package to look at the correlation of Pclass survival between women and men by age. In Pclass 1, the female looks safe and the person who dies is small. The survival rate of Pclass 2 was lower than that of Pclass 1, and it was found that young passengers were safe and their mortality increased in their 20s or older.

3.2. Variable selection

It is common that there are a large number of input variables that are candidates for one target variable among the many variables of the big data. It is effective to select variables with high relevance to target variables. In this study, we analyzed all the data, found the variables that are related to the survival rate, and made the data frame different for easy analysis, so we are going to use the random forest algorithm to have a better variable selection and variables importance.

3.3. Radom Forest

Random forest algorithm evaluate the importance of variables and select the variable that can be used in the model [9].

Here, varImpPlot prints a result that allows you to precisely view the variable importance you select through the label variables that are dedicated to the random forest algorithm. The Random Forest can handle numeric, categorical, and associative variables [8]. Used the following code to create a temporary data frame with the variables most relevant to the survival rate and also create a private label variable [9]. Wrote the following code to Visualization the importance of the variables on the data frame that we made as a temporary data frame has the most correlation between survived and perished and shows the title as the most important variable (Figure 13) shows the importance of variables top-down10].

RF.train.2<-
data.combined[1:891,c("Pclass","title","SibSp","family.size")]

RF.2 <- randomForest(x = RF.train.2, y = RF.label, importance = TRUE, ntree = 1000)
varImpPlot(RF.2)[]. [11]

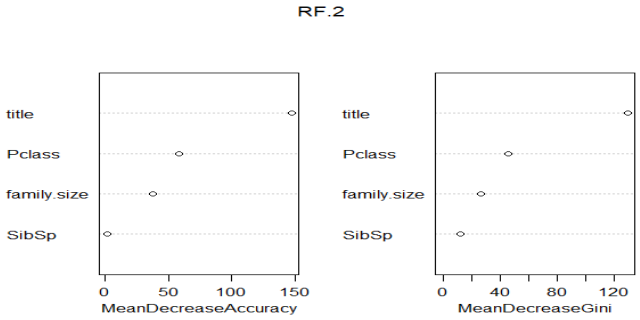


Fig. 13: The selection of variables by random Forest

So we got 81% accuracy as the fallowing error rate shows 19.64% error, OOB estimate of error rate: 19.64%. the selection of variable . Model RF.2 shows that the error rate is 19% and the model accuracy is 81%. (Figure 14) shown the sample of model accuracy [12].

```
OOB estimate of error rate: 19.64%
Confusion matrix:
    0   1 class.error
0 487|  62   0.1129326
1 113 229   0.3304094
```

Fig. 14: Shown the accuracy of model RF.2

3.4. Tree decision

During the processing and analyzing data, random forest algorithm given us the most predictive potential values of variables title, Pclass, and family size so used the following code by using 3 most important variable and figure out tree decision for our predict model (Figure 15) shows the survival prediction by a single tree[13]

features <- c("Pclass", "title", "family.size")
rpart.train.1 <- data.combined[1:891, features]
rpart.1.cv.1 <- rpart.cv(12422, rpart.train.1, RF.labe2, ctrl.3)
prp(rpart.1.cv.1$finalModel, type=1, extra = 1, under = TRUE)
Note the right branch of the tree is always no and left branch is yes so if the passenger title is Mr. or other it means perished 2. If Pclass=3 and family size is more than 4 peoples traveling together means perished.

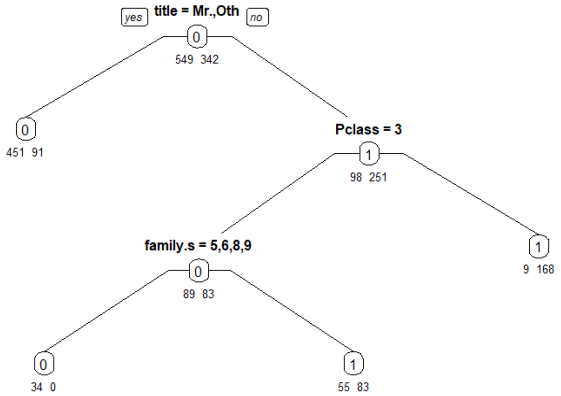


Fig. 15: Tree decision by a single tree

In order to easily identify the prediction model and analysis results, we visualized the data as shown in (Figure 16) using the prp() function provided in R and output it to decision tree.
prp(rpart.3.cv.1$finalModel, type = 0, extra = 101, under=TRUE)

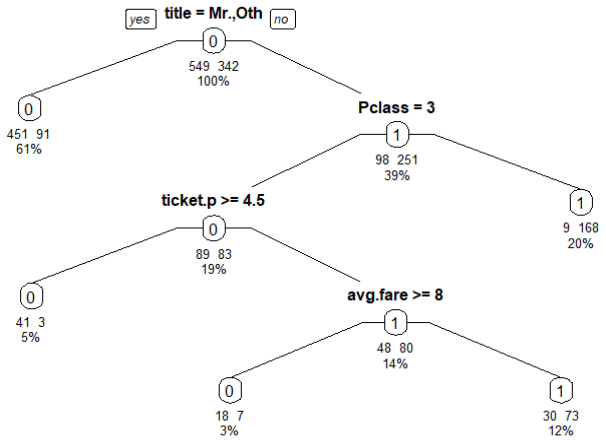


Fig. 16: The sample of by decision tree

In general, a train data set is a data set with variable values that can be predicted by analyzing the label or survival. In this study, the goal is to create a prediction model using the Train data set, which is the meaning of the RF.2 model, which will be used to create a prediction model for the test data set. This process was tested to identify the key variables associated with the survival rate of the Titanic data and to determine the accuracy of the model. We have demonstrated the significance and accuracy of the title, Pclass, and family.size variables that relate survival and death to the random forest machine learning algorithm. Therefore, we use the following code to retrieve the last 418 rows indexed from 892 to 1309 corresponding to the original test set, and create test.sub.dataframe by assigning three variables. As shown in the following (Figure 17), data.combined 1309 observations 18 variables and test.sub.dataframe are data sets with 418 observations 3 variables.

test.sub.dataFrame <- data.combined[892:1309, c("Pclass", "title", "family.size")]

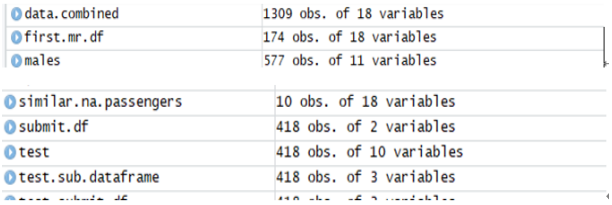
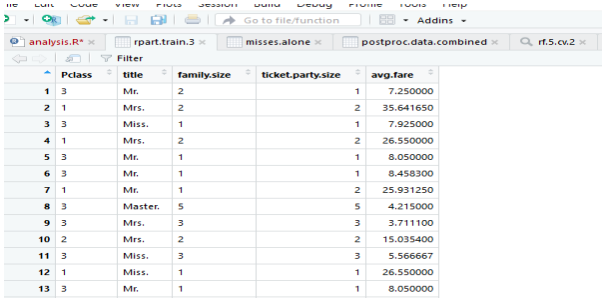


Fig. 17: Shown data sets bye observation and variable

To improve data analysis results and models, we used the following function rep() to create tichet.party.size variable, which means how many peoples travels with the same ticket . ticket.party.size <- rep(0, nrow(data.combined))
Or, the average ticket price is described as the follows.
avg.fare <- rep(0.0, nrow(data.combined))
It also created the possibility of a passenger traveling on a special ticket.
tickets <- unique(data.combined$Ticket)
In order to make it easier to predict survival, we would create ticket.party.size and age.fare variables as shown in the following code, and add them to the data to create a better decision tree. We need to make better variables with the possibility of this, so we created the desired variable as shown in the following code .

```
for (i in 1:length(tickets)) {
current.ticket<-tickets[i]party.indexes<-
which(data.combined$Ticket == current.ticket)
current.avg.fare <- data.combined[party.indexes[1], "Fare"] / length(party.indexes)
for(k in 1:length(party.indexes)) {
ticket.party.size[party.indexes[k]]<-length(party.indexes)
  avg.fare[party.indexes[k]] <- current.avg.fare
  }
 }
```

data.combined$ticket.party.size <- ticket.party.size
data.combined$avg.fare <- avg.fare
To create the most probable data frame, we used a c () function of R as follows to create a vector with five variables like, features.final.
features.final <- c("Pclass", "title", "family.size", "ticket.party.size", "avg.fare")
After taking the first 1 to 891 rows corresponding to the train data, take a specific row of features.final and create a data frame called rpart.train.3.
rpart.train.3 <- data.combined[1:891, features.final]
Using this, we created a data frame that is most relevant to the survival rate as shown in (Figure 18).

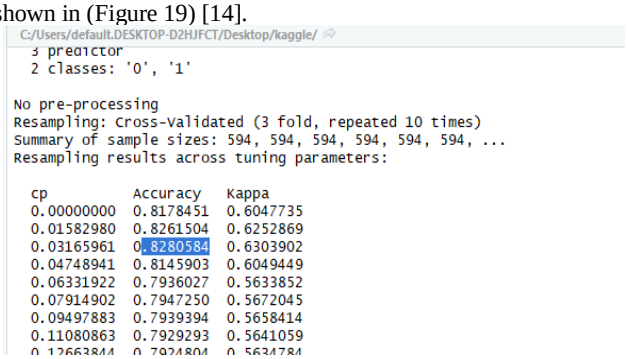


	Pclass	title	family.size	ticket.party.size	avg.fare
1	3	Mr.	2	1	7.250000
2	1	Mrs.	2	2	35.641650
3	3	Miss.	1	1	7.925000
4	1	Mrs.	2	2	26.550000
5	3	Mr.	1	1	8.050000
6	3	Mr.	1	1	8.458300
7	1	Mr.	1	2	25.931250
8	3	Master.	5	5	4.215000
9	3	Mrs.	3	3	3.711100
10	2	Mrs.	2	2	15.035400
11	3	Miss.	3	3	5.566667
12	1	Miss.	1	1	26.550000
13	3	Mr.	1	1	8.050000

Fig. 18: The sample of rpart.train.3 data frame

3.5. Cross validation

rf.label <- as.factor (train $ Survived) this comment convert the data type as a factor, and used the following code to create rpart.3.cv1 for cross validation.
rpart.3.cv.1 <- rpart.cv (12365, rpart.train.3, rf.label, ctrl.3) and the result of cross validation give us the 81.82% accuracy as shown in (Figure 19) [14].



```
C:/Users/default.DESKTOP-D2HJFCT/Desktop/kaggle/
  3 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (3 fold, repeated 10 times)
Summary of sample sizes: 594, 594, 594, 594, 594, 594, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.00000000  0.8178451  0.6047735
  0.01582980  0.8261504  0.6252869
  0.03165961  0.8280584  0.6303902
  0.04748941  0.8145903  0.6049449
  0.06331922  0.7936027  0.5633852
  0.07914902  0.7947250  0.5672045
  0.09497883  0.7939394  0.5658414
  0.11080863  0.7929293  0.5641059
  0.12663844  0.7924804  0.5634784
```

Fig. 19: The sample of predictive accuracy

4. Results

Finally, we split the training data using R functions, analysis and found the variables correlations. As well as mentioned at the beginning the original data have only 11 variable with 79% accuracy, we found the potential values of the variable and increase it to 18 variables to make more accurate and predictive. Created utility functions make more predictive data and convert the data types as needed for analysis.
Sued the machine learning algorithm and R packages to make highly accurate for the prediction model. used the packages to visualize our predictions at the end we got 82.8% accuracy and then used the following code to leverage it to the test data set and we predict 268 perished and 150people survived on test data as submission data .
test.submit.df <- data.combined[891:1309, features]
rpart.3.preds<-predict(rpart.3.cv.1$finalModel, test.submit.df, type="class")
table(rpart.3.preds)

References

[1] Practical data Processing and Analysis Using R
[2] http://www.cyclismo.org/tutorial/R/input.htmlhttps://ramnathv.github.io/pycon2014-r/visualize/ggplot2.html
[3] https://www.kaggle.com/c/titanic/data
[4] Practical data Processing and Analysis Using R
[5] Patent Big Data Analysis by R Data Language for Technology Management. International Journal of Software Engineering and Its Applications Vol. 10, No. 1(2016), pp, 69-78
[6] http://www.dodomira.com/tag/ggplot2/
[7] https://thebook.io/006723/ch10/03/04/03/
[8] http://hamelg.blogspot.com/2015/09/introduction-to-r-part-30-random-forests.html#!/2015/09/introduction-to-r-part-30-random-forests.html
[9] http://blog.heartcount.io/random-forest-ver-10
[10] https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf
[11] http://www.dodomira.com/2016/05/29/564/
[12] A Review Study on Big Data. Anshul Jatain et al, International Journal of Computer Science and Mobile Computing, Vol.6 Issue.6, June- 2017, pg. 8-13
[13] http://hamelg.blogspot.com/2015/09/introduction-to-r-part-29-decision-trees.html#!/2015/09/introduction-to-r-part-29-decision-trees.html
[14] http://dataaspirant.com/2017/02/03/decision-tree-classifier-implementation-in-r/