STA 5224: Final Project - Titanic Dataset

Huong Tran

Contents

1 I. Project Proposal:

This project is a competition on Kaggle with target of predicting which passengers survived the Titanic shipwreck by machine learning. Part 1 of the project will focus on cleaning data and obtaining some insights about data, together with variable selection process to create a meaningful dataset that can be used. Model application will be in the second part. Besides the statistical approach in logistic models, other machines learning model will be used. A typical model for supervised learning is decision tree, which is a series of sequential decisions represented as a tree to get a specific result. However, decision tree are prone to overfitting, especially when the tree is deep and random forest is a solution for this kind of problem. In general, a random forests model will creates several random decision trees and aggregate their result. Also, SVM works well with classification and regression, therefore it is good to represent SVM. Finally, the comparision between these models will be derived. In the last section, statistical inference and interpretation from logistic models will be discussed.

2 EDA (Exploratory Data Analysis):

2.1 Overview about Titanic Dataset:

The data is obtained from the Titanic competition from Kaggle. Train dataset is provided with labels, from which we build our model to predict if a person would survive through the disaster. Kaggle will evaluate model performance using Test dataset.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Bri...	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhel...	female	27	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

Figure 1: Train dataset

There are 11 independent variables helping predict "Survived" variable, which take value of 1 as the passenger survived throught the disaster. Dictionary of these variables can be found through Kaggle.

Table 1: Summary of missing values.

	NA.train	Empty.train	Percent.train	NA.test	Empty.test	Percent.test
PassengerId	0	0	0.00	0	0	0.00
Pclass	0	0	0.00	0	0	0.00
Name	0	0	0.00	0	0	0.00
Sex	0	0	0.00	0	0	0.00
Age	0	177	19.87	0	86	20.57
SibSp	0	0	0.00	0	0	0.00
Parch	0	0	0.00	0	0	0.00
Ticket	0	0	0.00	0	0	0.00
Fare	0	0	0.00	0	1	0.24
Cabin	0	687	77.10	0	327	78.23
Embarked	0	2	0.22	0	0	0.00

From table 1, *Cabin* variables has 77% missing value in train set and 78.23% in test set, therefore we will drop this variable after exploration.

2.2 Some insights about variable *Name*:

From 1, *Name* does not contain any missing and 891 unique values, which is not very informative. However, this variable having information about title of person, which indicates their social class and profession. At first, there are 17 unique titles as in table 3 and many of them represents for the same levels:

- The *Title* relating to army and "Rev" are less likely to survive and they are male, to increase the degreee of freedom for error term, we can merge this title as "Official" level.
- *Title* "Mme", "Mlle", "th", "Lady", "Ms" represent unmarried women, we change them into "Miss".
- "Jonkheer" will be changed into "Mr".

- The only one value of “Lady” and “Sir” are spouse, we will change their title into “Mrs” and “Mr”.

From figure 2, the cleaning process help reduce the number of different levels in *Title* variable to 6. “Mr” and “Officer” has the least chance of survival while “Miss” has the highest chance of survival.

Finally, this variable also contains information about *Sex*. For example, a person with “Mr” value in “Title” should be a male.

2.3 Age:

Recall from table 1, variable *Age* has 177 missing value in train dataset, which accounts for 19.87% of total observation. Also, in test dataset, there are 86 missing value.

Figure 3 shows that the majority of missing value distributed mostly in social class *Pclass = 3*, with *Title* “Mrs”, “Mr” and “Miss”. From that, missing value can be imputed by the mean of *Age* of this group.Imp

In figure 4, 50% of survival male in middle class was less than 10 years old. Also, more than 50-year-old man is the least likely to survive through the disaster. This suggests a way to divide age into 3 smaller groups: young, middle.age and old stored in variable *Age.char*.

People in the first class has the highest chance to survive, especially when they are in middle age group. In contrast, middle-age men is the most likely died in the disaster. And in fact, the young in second group will survive.

2.4 Information about Family size and total Ticket price of each family:

At first, both variable *SibSp* and *Parch* contain information about family size, we can create a new variable as *family.size* to obtain information about passenger’s family.

Boxplot 5 shows a positive trend of ln(*Fare*) and size of family, i.e, the large family size, the more fare that a ticket they paid. It seems that family members had the same ticket number would have the the same Fare. To get rid of the correlation, we will find the ticket price that each person had to pay by dividing *Fare* for number of family member. Also, since there is some missing value in *Fare*, this will imply tickets cost 0, which is impossible. In fact, the higher of socical class, the more expensive ticket, it is reasonable to impute missing value of *Price* by the mean based on *Pclass*.

2.5 What can Ticket class and Embarkation tell us?

Mosaic plot 6 shows that a women of upper class has the highest chance of survival while a men the the Lower class the lowest chance of survival. Also, although the total number of male is three times the total number of female, but male has the lower probability of survival. In fact, when the disaster hit, women and children were the first priority to go to the rescue boat.

Embarked has 2 missing values, from tabel 1 in the train set. The people of these two missing value have same information, but different name. In fact, this cabin belongs to Mrs. George Nelson, and Miss Amelie is her maid. Since these missing values are from cabin B28, other variables in deck B can be used for imputation.

In general, result of cross table shows that passenger with Cabin in deck B used Cherbourg (C) and Southampton (S) as their embarkation. The percentage of survival of port Cherbourg (C) is higher, therefore, we can impute the missing data above by C.

```
train$Embarked[train$Embarked == ""] <- "C"
```

Figure 7 shows that the majority of passenger used Southampton (S) to embark. However, only lower class (Pclass == 3) used Queenstown (Q).

3 Model Fitting:

3.1 Variable Selection:

After cleaning data, we will split train into 2 files: *train.mod* and *valid.mod* using function

`createDataPartition()`

from

`caret`

package, to for evaluate our model before applying it to give prediction on Test dataset.

```
## Number of observation in train.mod is:  713
##   Number of observation in test.mod is:  178
```

Forward selection below suggests the model including *Title, Pclass, family.size, Age, Fare, price, Sex* as predictors. This model produced the AIC of 614.4. However, as discussed above, *Title* variable already includes information about *Sex* and there is no need to use this variable as predictors. Also, to avoid the collinearity of *Fare*, this variable should be replace by *price*. In conclusion, our model will use *Title, Pclass, family.size, Age, price*.

```
## Survived ~ Title + Pclass + family.size + Age + Fare + price +
##     Sex
```

3.2 Logistic Regression:

Logistic Regression will return result of numeric value, which represent for the probability that a person survived through the disaster. Therefore, choosing a cut point will help obtain classification as 1 for survived and 0 for not survived. A resonable cut point is the estimate probability of survival, calculated as:

$$\hat{p} = \frac{\text{Total number of survival}}{\text{Total number of observation}}$$

```
## The fitted accuracy is:  0.8204769
```

Logistic Reg- sression model provide AIC of 614.87 and the fitted accuracy is 82.04$%%.

3.3 Decision Tree:

Decision Tree is a supervised learning method, which uses a graph to represent choices and their results in a form of a tree. Figure 8 illustrates how the tree looks like when model is fitted. *Title* is the most important factor, if a passenger has title of "Dr", "Mr", "Officer", there is 60% of chance that they could not survive. In constrast, smaller value of Pclass has higher probability of survive, since they were in first class, and obviously were helped to escape from the ship when disaster hit. The fitted accuracy of decision tree is about 84.57%, which is higher than logistic regression.

```
## Fitted accuracy for Decision Tree is:  0.8457223
```

3.4 Random Forest:

Random forest is a supervised learning which creates several random Decision Tree and output is the aggregate of those trees. To make the algorithm stable, we will create several forest with the same number of trees, final result is taken using the mean of these forests.

In figure 9, fitted accuracy increase as number of tree increase, accuracy peaks when ntree = 45, but it never exceed 82% of accuracy.

```
## Fitted accuracy of random forest is:  0.811594
```

3.5 SVM (Support Vector Machine):

SVM (Support Vector Machine) is a supervised learning methods, which used to classified data. In general, it creates a hyperplane to separate train data into 2 classes, which are labled, in our dataset, those lables are 0 and 1. The goal is to decide which class a new data point will be in. Fitting train data into SVM model, we obtain the fitted accuracy is: 90%, which much higher that previous methods.

```
## Fitted accuracy of SVM is:  0.9046283
```

3.6 Model comparision:

Model comparsion should be based on new data. Fitting new data from *valid.mod* dataframe can result a more objective comparison between our models. In table 2, although fitted accuracy of SVM is highest, but this algorithm does not work well with new data. It seems that Decision Tree is the best algorithm, which provides the highest accuracy, 86% of correct prediction.

Table 2: Accuracy for valid.mod dataframe

	Summary results			
	logistic	tree	random.forest	svm.mod
Fit accuarcy	0.8204769	0.8457223	0.8115940	0.9046283
Prediction accuracy	0.8314607	0.8595506	0.8426966	0.6966292

4 Interpretation of Losgistic Regresion:

Recall logistic regression model:

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = train.mod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5812  -0.6102  -0.3984   0.5254   2.5295
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.036758   1.070463   1.903 0.057081 .
## TitleMaster   3.405710   1.086281   3.135 0.001717 **
## TitleMiss     2.813440   0.942680   2.985 0.002840 **
## TitleMr      -0.184966   0.911529  -0.203 0.839199
## TitleMrs      3.435723   0.951745   3.610 0.000306 ***
## TitleOfficer -0.272468   1.207744  -0.226 0.821512
## Pclass       -1.036790   0.166819  -6.215 5.13e-10 ***
## family.size  -0.503103   0.102745  -4.897 9.75e-07 ***
## Age          -0.022752   0.009974  -2.281 0.022533 *
## Fare          0.012884   0.006137   2.099 0.035783 *
## price        -0.010904   0.007129  -1.529 0.126168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 947.99  on 712  degrees of freedom
## Residual deviance: 592.87  on 702  degrees of freedom
## AIC: 614.87
##
## Number of Fisher Scoring iterations: 5
```

Using cutpoint at

$$\hat{p} = 0.38$$

logistic regression provide overall proportion of correct classification is 82.04%. Figure 10 shows AUC is 0.8696.

```
## AUC of logistic regression is:  0.8696312
```

Restriction of Logistic Regression: - Since the choice of cut point is arbitrary, the result are sensitive to relative numbers of times that $y = 1$ and $y = 0$.

- Also, it collapses continuous predictive value $\hat{p}$ into binary ones. For example, predictive value of 0.37999 will result 0, which is not very convincible.

5 Make prediction on test dataset and submit to Kaggle:

Besides cleaning test dataset with the procedure that we cleaned train dataset, table 1 also say that there is 1 missing value in variable *Fare*, therefore, after calculating the ticket price, we can fill in this missing value by the following fomula:

$$\text{Fare} = \text{price} \times \text{family.size}$$

As discuss above, SVM does not seem to work well in newdata, therefore we will just make prediction for test dataset, using Logistic Regression and Decision Tree.

Kaggle returns the score for our submission at 77.9% for Logistic Regression and 77.4% for Decision Tree, which is not very different between the two models.

6 Further question for improvement:

1. In variable selection, we used least square regression model. In fact, variables in our data are correlated, as mentioned above and Ridge regression is expected to give better performance when handling collinearity.
2. We used mean to impute missing value for *Age* based on their age group. In fact, regression on *Age* using other variables can be applied and hopefully will bring better results.
3. In our logistic model, we did not consider the interaction of predictors. However, as we have seen above, *Title* "Master" in the first class has more chance to survive than the lower class. Therefore, interaction of some variables should be taken into account.

7 Appendix: Figures and Code

Table 3: Count of unique Title.

Title	count
Capt	1
Col	2
Don	1
Dr	7
Jonkheer	1
Lady	1
Major	2
Master	40
Miss	182
Mlle	2
Mme	1
Mr	517
Mrs	125
Ms	1
Rev	6
Sir	1
th	1

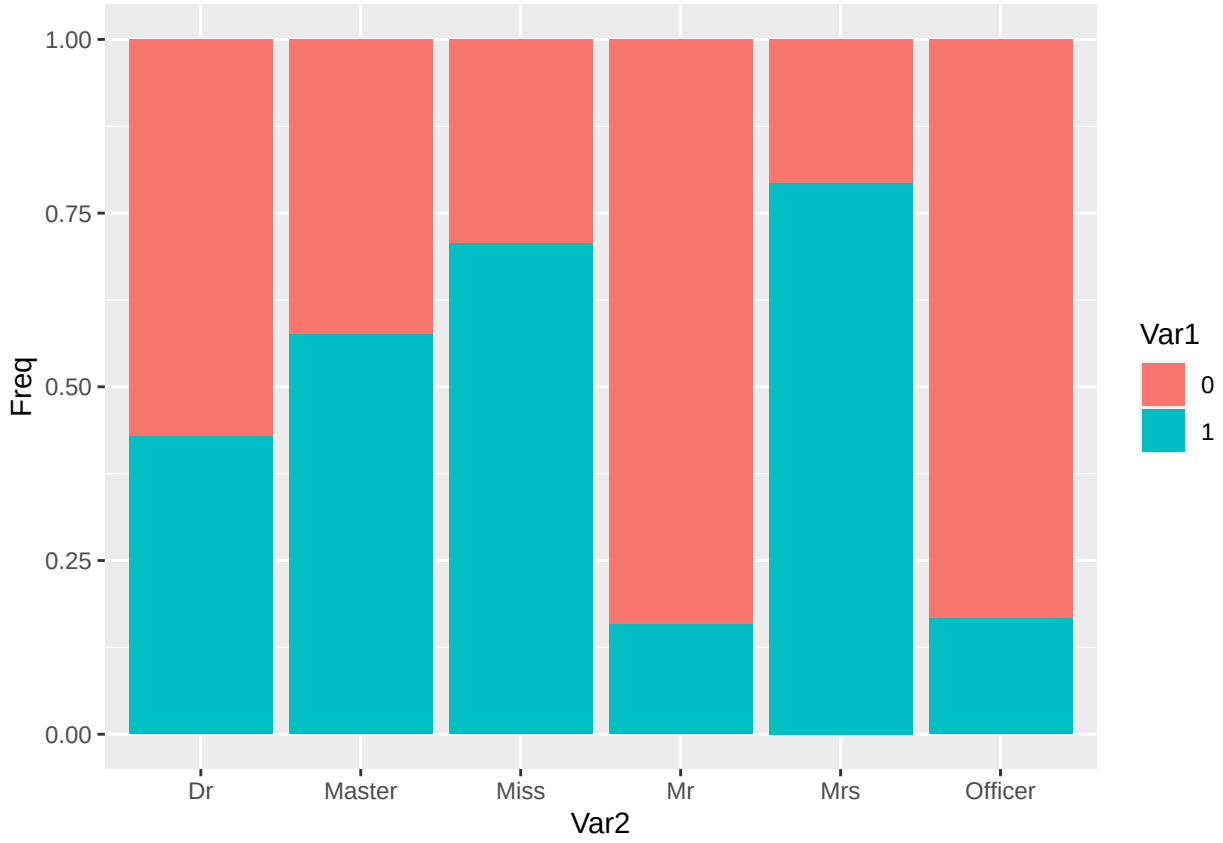

Figure 2: Title and Survival.

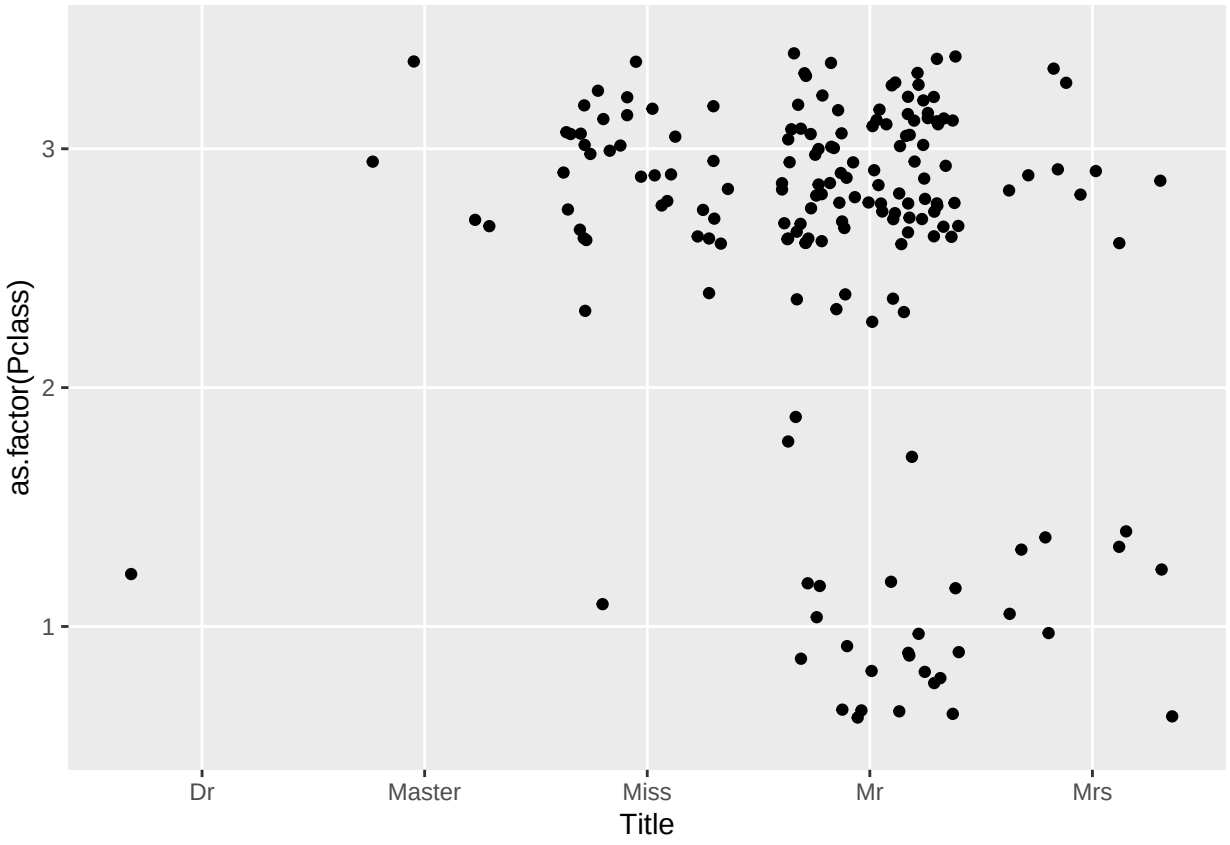

Figure 3: Relationship between Age and Title and Pclass

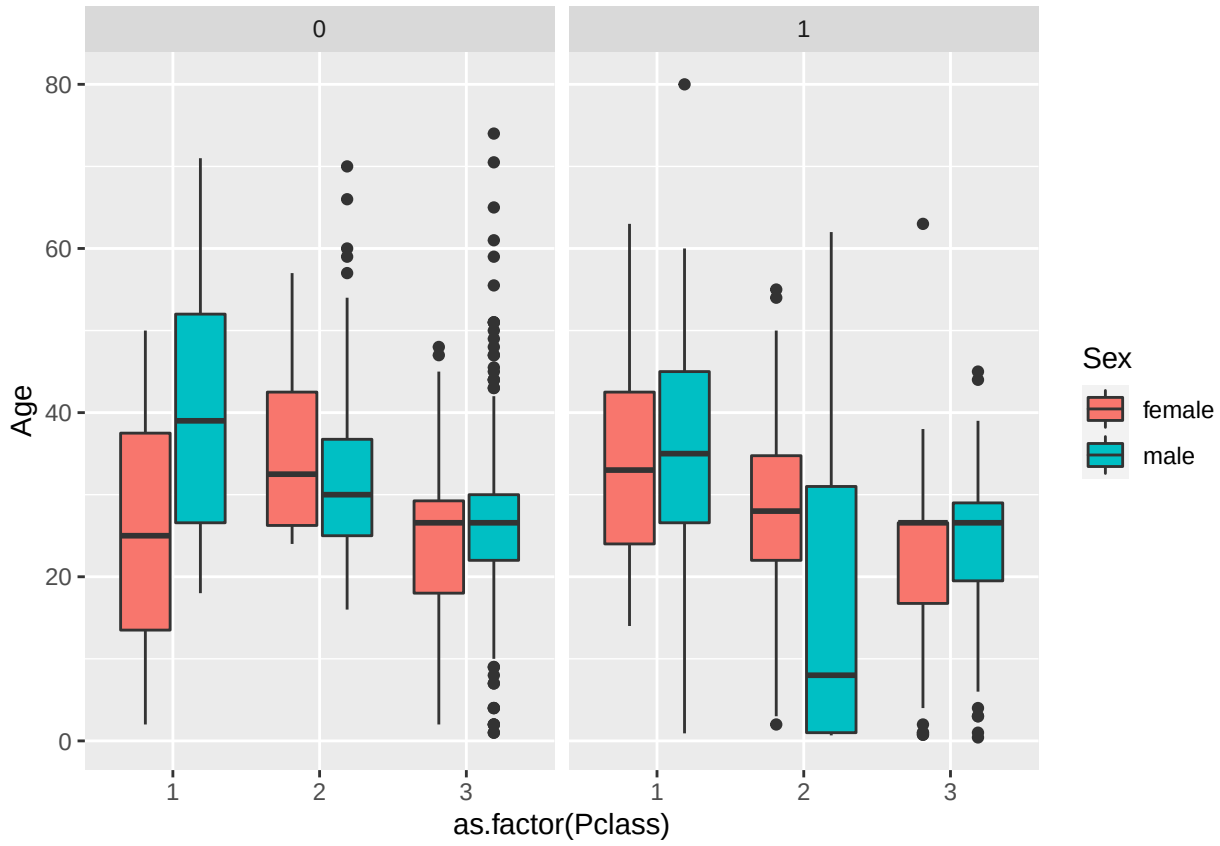

Figure 4: Boxplot of Age, by Pclass, Sex, Survival.

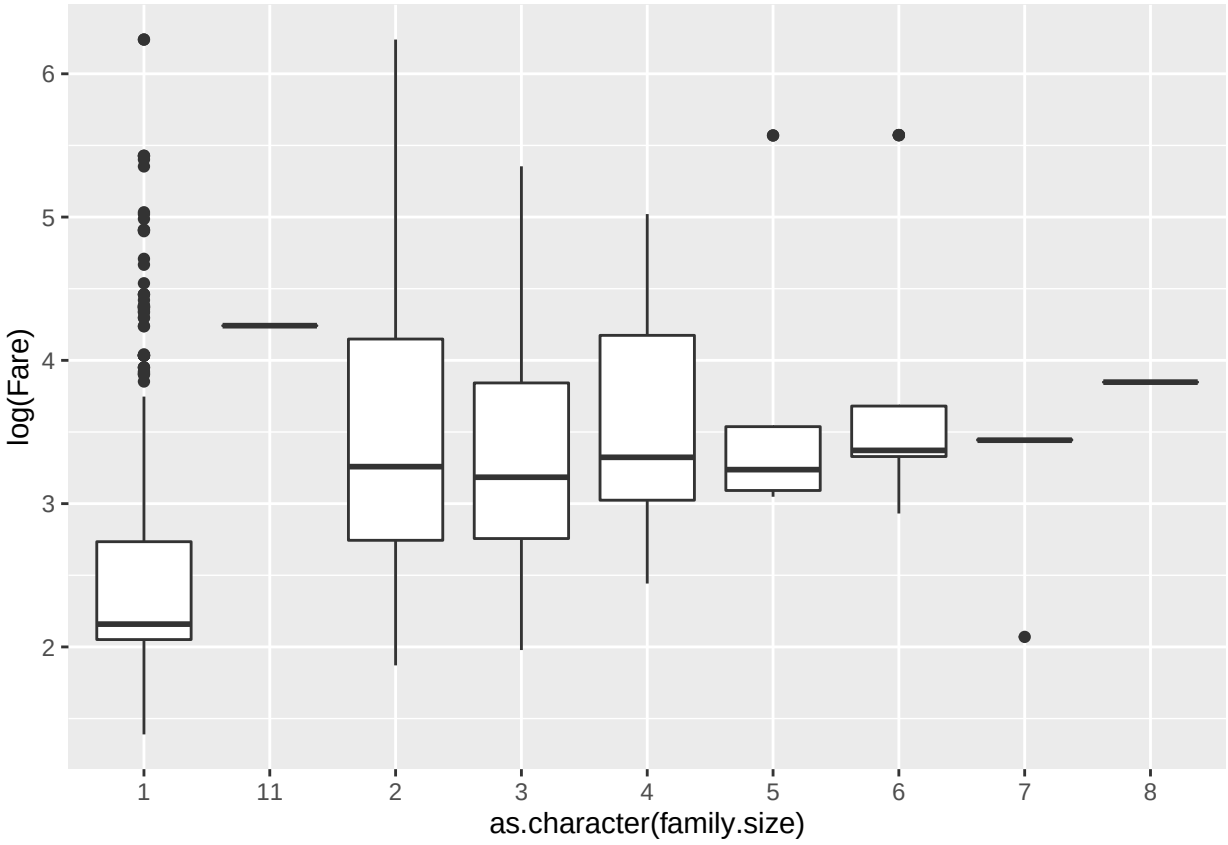

Figure 5: Family Size and Ticket Fare

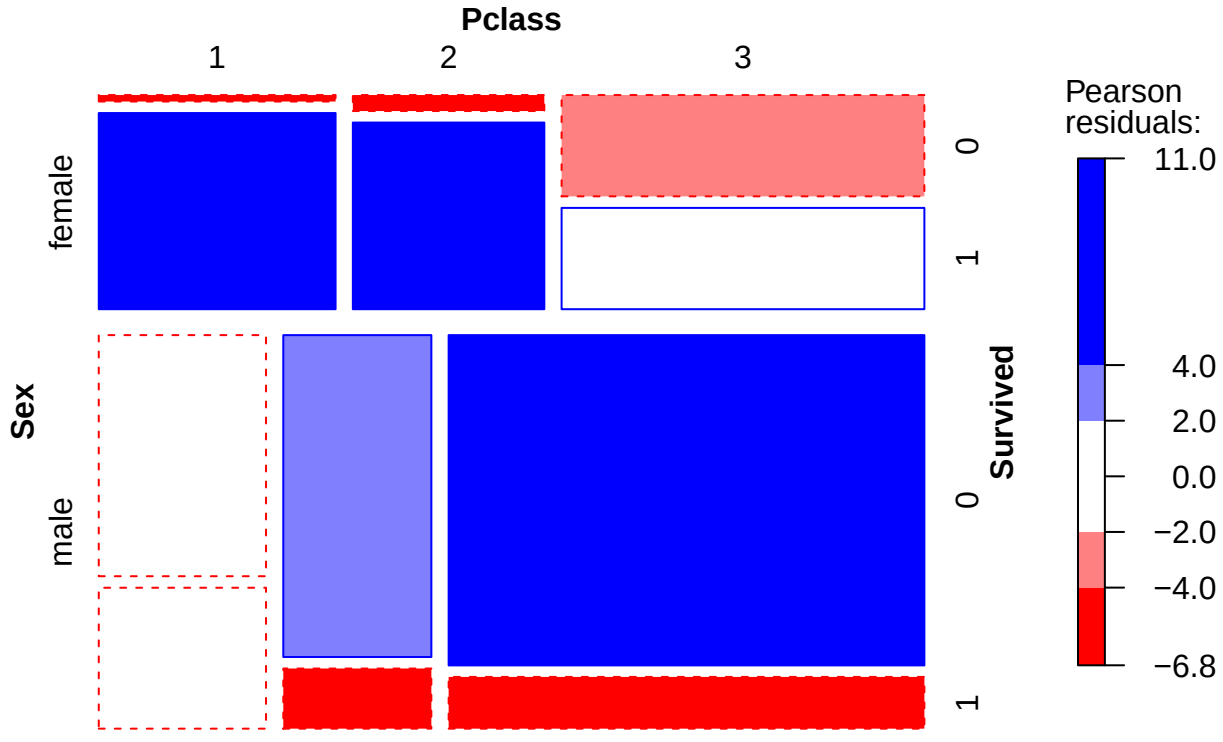

Figure 6: Pclass and Sex affect Survival.

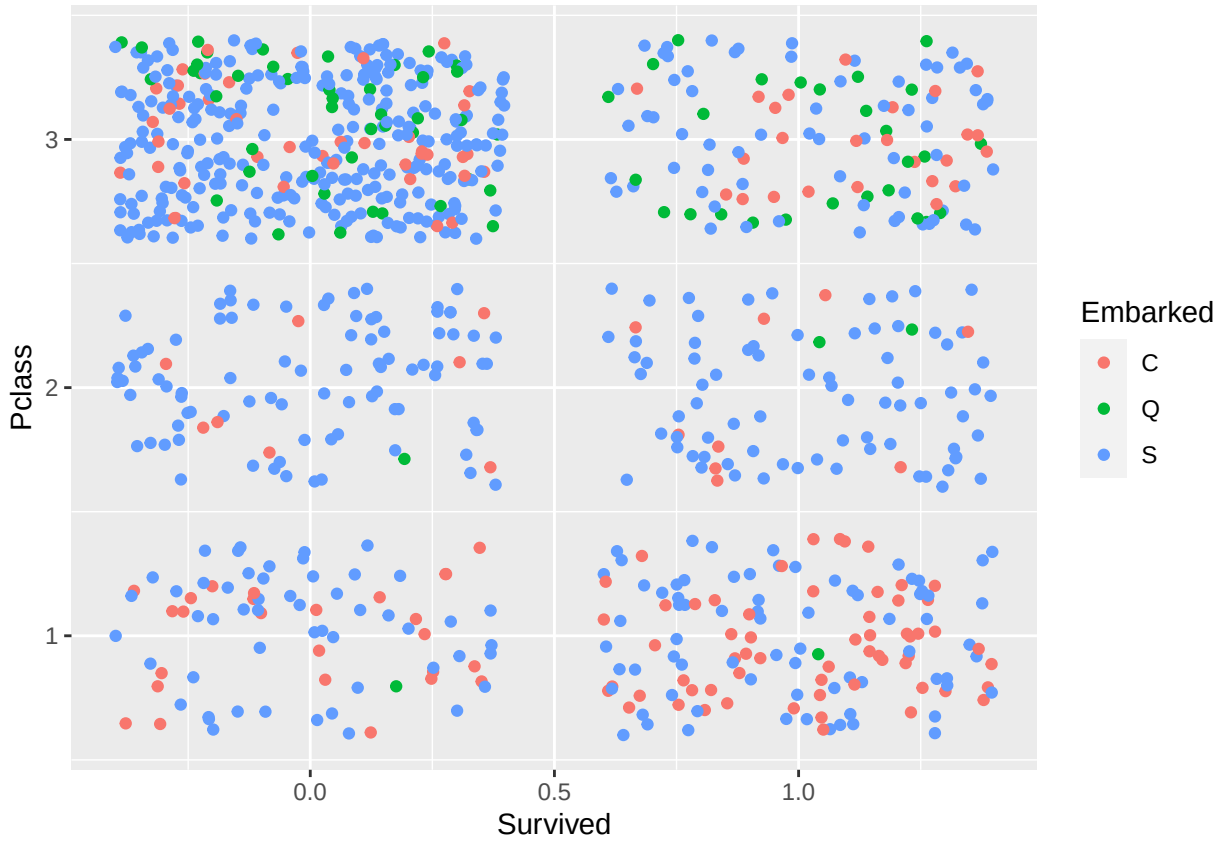

Figure 7: Relation of Passerger class and Port of Embarktionin their survial chance.

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  47
##
##
##                | B$Embarked
##   B$Survived |           |         C |         S | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            0 |         0 |         5 |         7 |        12 |
##              |     0.000 |     0.227 |     0.304 |           |
## -------------|-----------|-----------|-----------|-----------|
##            1 |         2 |        17 |        16 |        35 |
##              |     1.000 |     0.773 |     0.696 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |         2 |        22 |        23 |        47 |
##              |     0.043 |     0.468 |     0.489 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

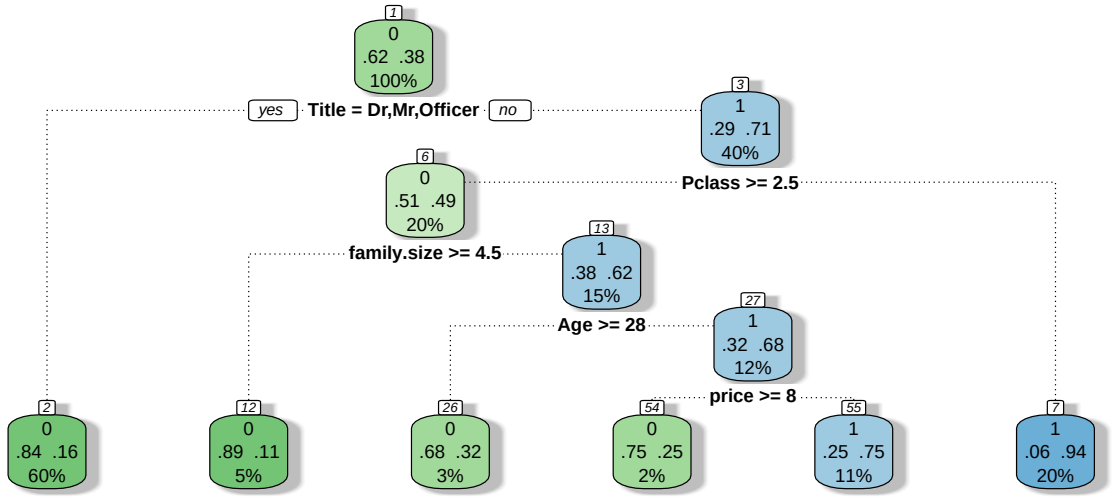

Figure 8: Decision Tree.

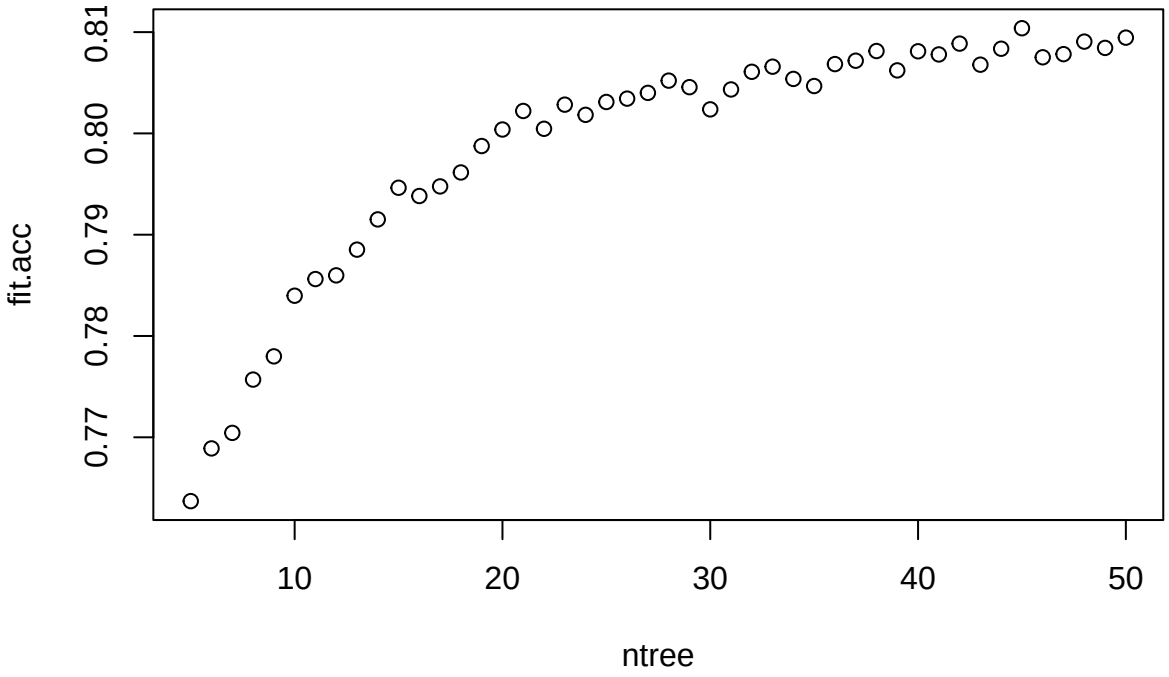

Figure 9: Accuracy based on number of tree in each forest

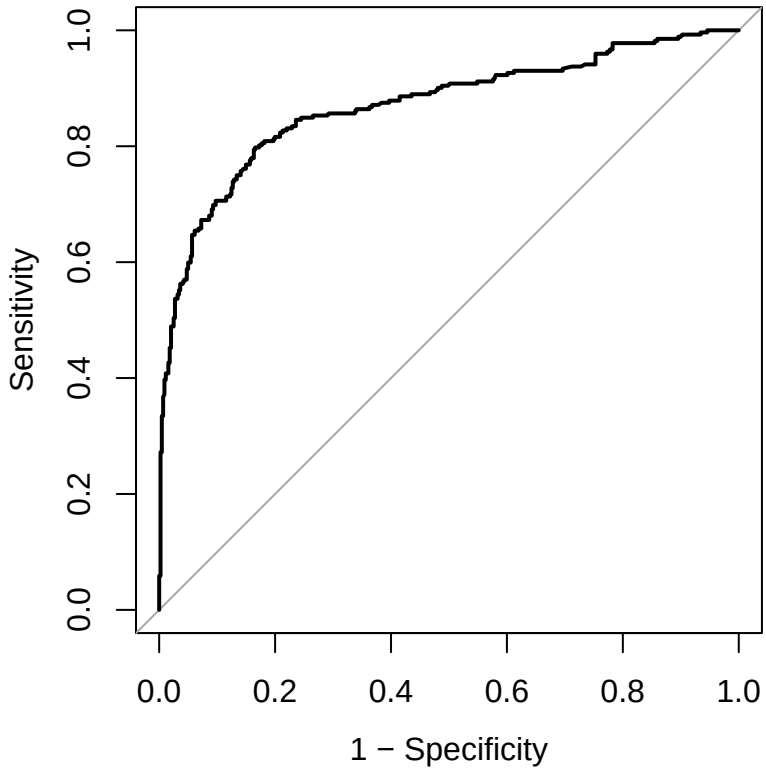

Figure 10: ROC curve of Logistic Regression.

References

[1] Categorical Data Analysis, Third Edition, Alan Agresti.

[2] An introduction to Categorical Data Analysis, Third Edition, Alan Agresti.

[3] Machine Learning Algorithms -A Review, Batta Mahesh