STA 5224: Final Project - Titanic Dataset

I. EDA:

1.1 Overview about the data set:

	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

Titanic dataset includes 11 features as predictors, which helps predict wheather a person would survived through the Titanic disaster.

```
## Number of observations:  891
##   Number of duplicated row:  0
##
```

At first, it seems that there is no missing value in this dataset, but in fact, there are some cells having value of empty string, and containing no information, those are considered as missing value.

	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

Cabin variables has 77% missing value in train set and 78.23% in test set, therefore I will drop this variable after exploration.

1.2 Some insights about the name?

In the variable "name", there is also title of the person, which indicates their social class and profession.

```
## Number of unique title is:  17
```

	Survived	Title	Sex
31	0	Don	male
150	0	Rev	male
151	0	Rev	male
250	0	Rev	male
450	1	Major	male
537	0	Major	male
627	0	Rev	male
648	1	Col	male
695	0	Col	male
746	0	Capt	male
849	0	Rev	male
887	0	Rev	male

The titles 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.

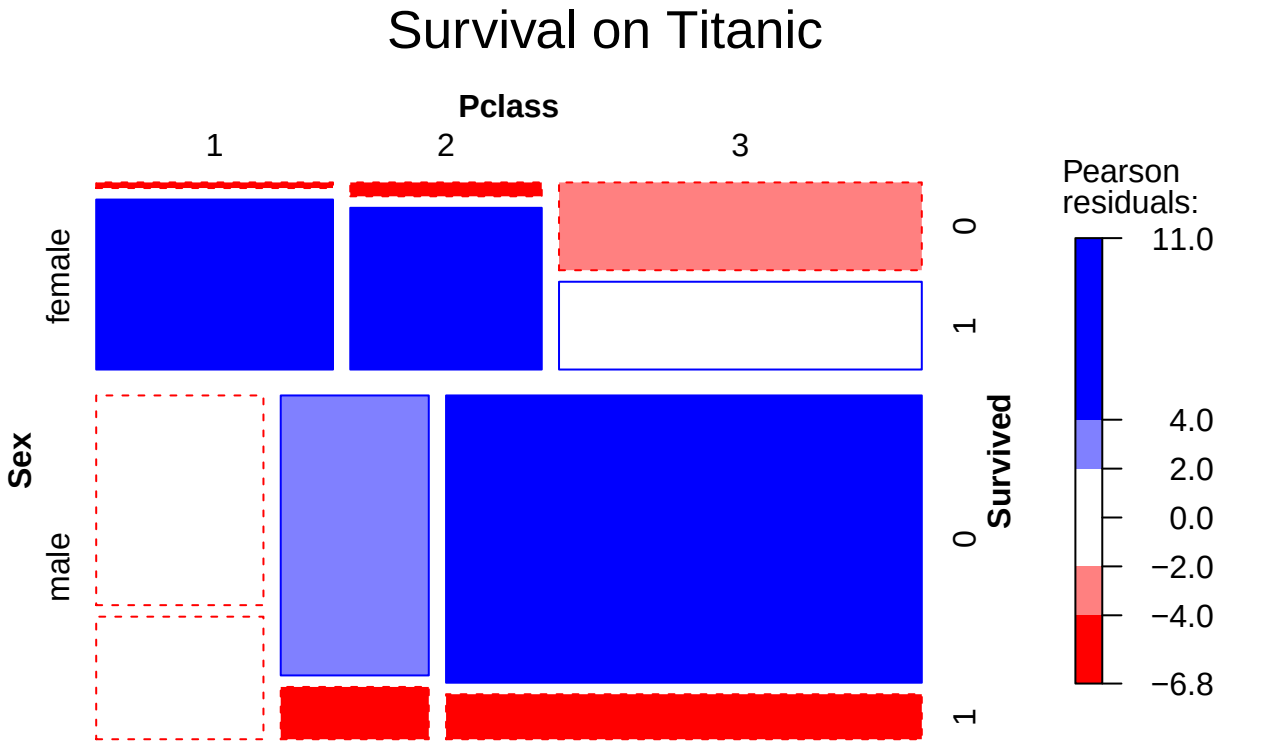
For the title listed above, they all represent for unmarried women, therefore, we should change them into "Miss. And"Jonkheer" will be changed into "Mr". Also the only one value of "Lady" and "Sir" are spouse, we will change their title into "Mrs" and "Mr".

```
## Unique Title are:  Mr Mrs Miss Master Officer Dr
```

After cleaning and merging "Title" variable, there are only 6 levels of this variable. In fact, those title also present passenger's sex. For example, a person with "Mr" value in "Title" should be a male.

1.3 What can Ticket class tell us?

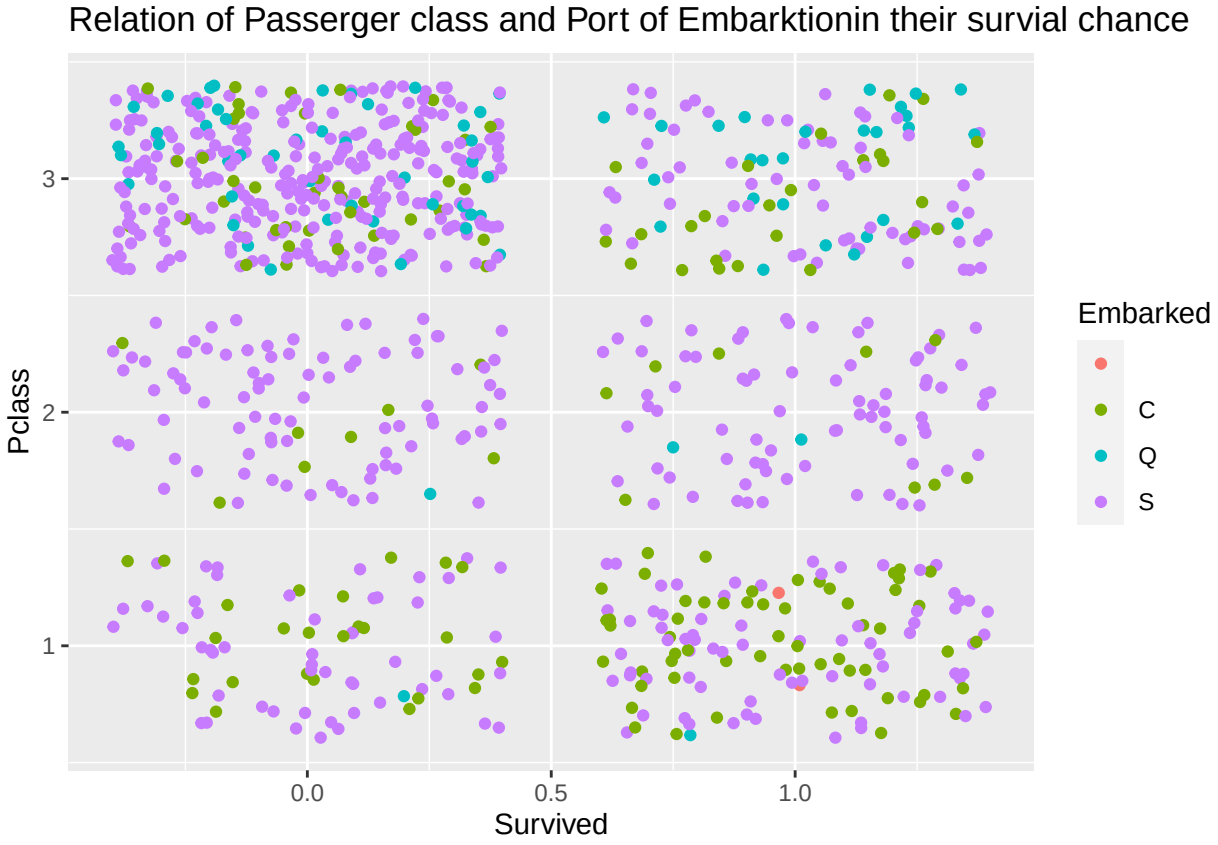
```
mosaic(~ Sex + Pclass + Survived, data = train,  main = "Survival on Titanic",
       gp = shading_Friendly, legend= T )
```



The mosaic plots 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.

1.4 Where did they embark?



From the table, we know that there are 2 missing values in variable "Embarked" of 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*, let see other variables in deck B: **? How can I impute the data here?**

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  47
##
##
##                | B$Embarked
##   B$Survived |           |         C |         S | Row Total |
## -------------|-----------|-----------|-----------|-----------|
```

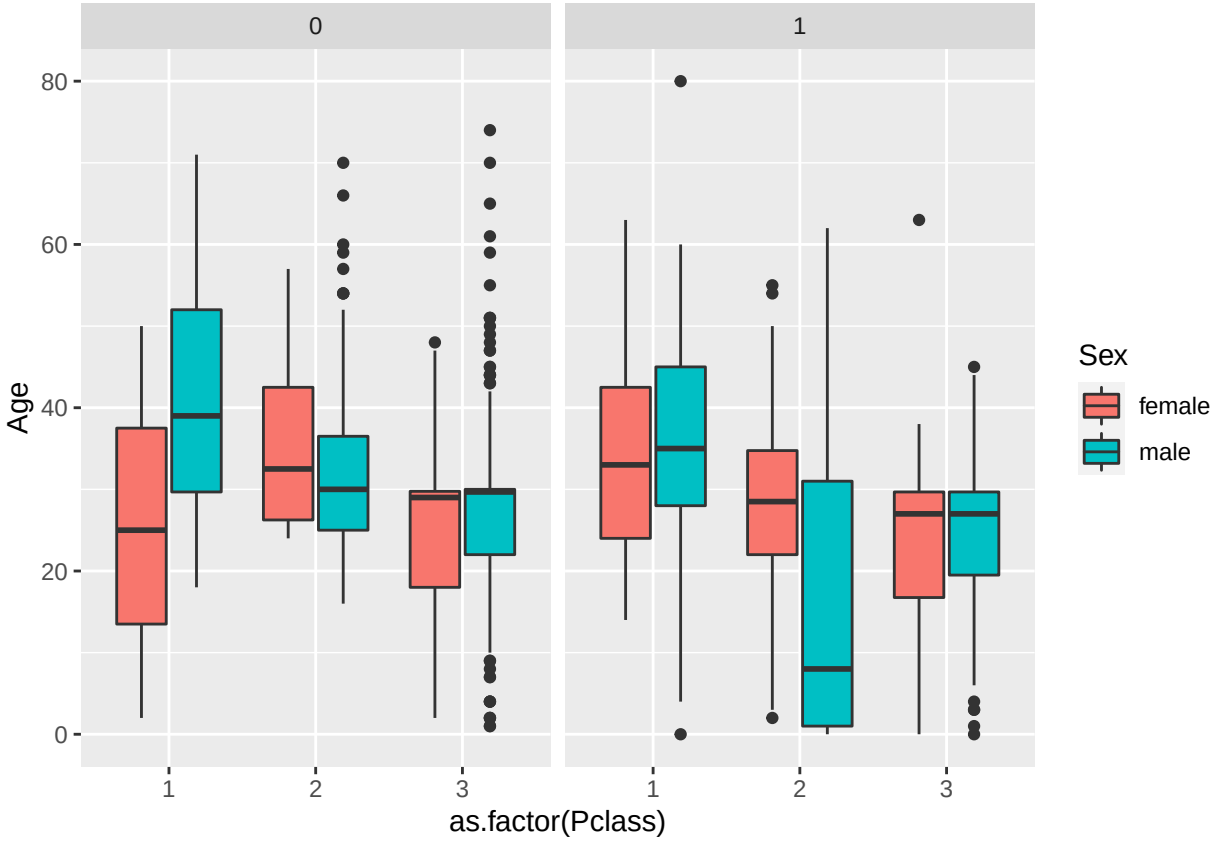
```
##              0 |         0 |         5 |         7 |        12 |
##                |     0.000 |     0.227 |     0.304 |           |
## -------------|-----------|-----------|-----------|-----------|
##              1 |         2 |        17 |        16 |        35 |
##                |     1.000 |     0.773 |     0.696 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |         2 |        22 |        23 |        47 |
##                |     0.043 |     0.468 |     0.489 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

Age:

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

```
train$Age <- as.integer(train$Age)
```

Those missing value are mainly are Mr. and Miss and Mrs, therefore, we can impute the missing data by mean of these group.



Surprisingly, 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.

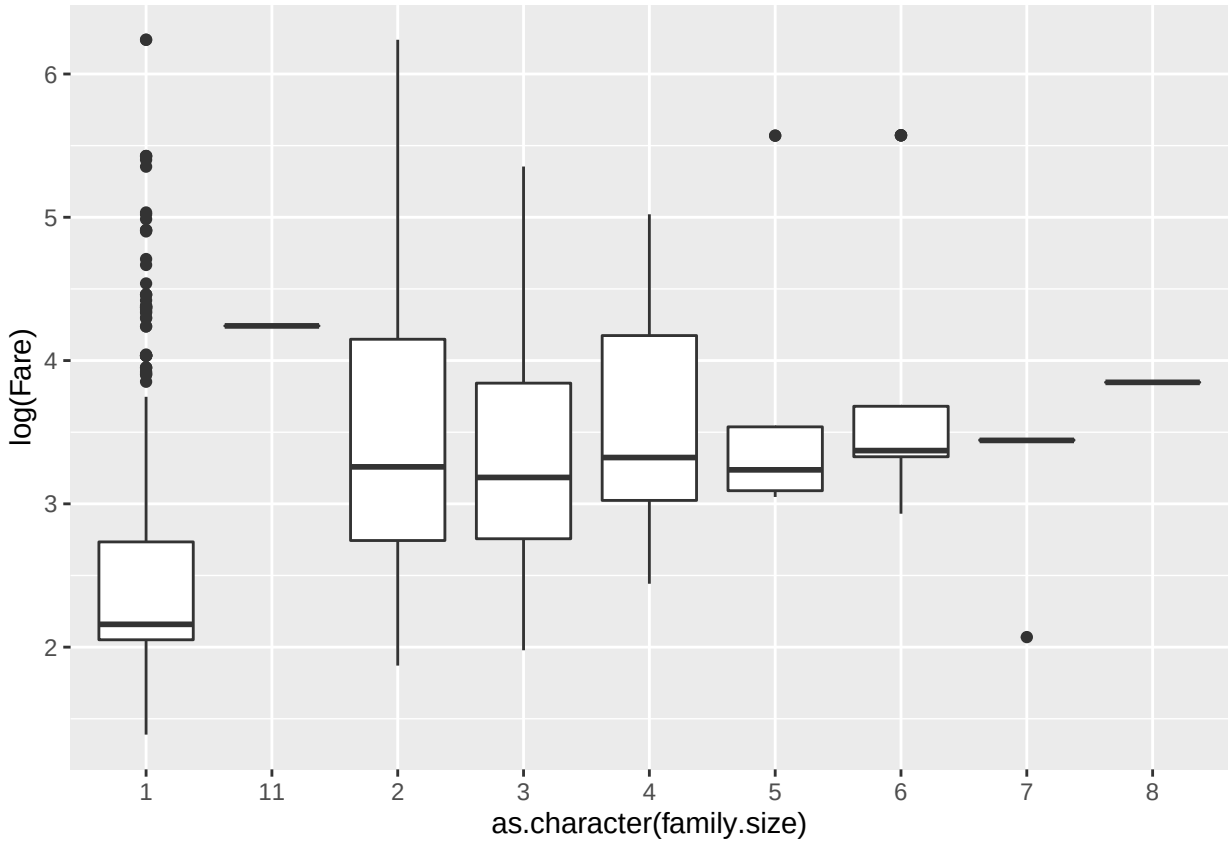
Family size:

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.

Surprisingly, the number of survival in family size from 2 to 4 is higher.

```
train$Fare <- as.numeric(train$Fare)
ggplot(train, aes(as.character(family.size), log(Fare))) + geom_boxplot()
```

```
## Warning: Removed 15 rows containing non-finite values (stat_boxplot).
```



There is positive trend of ln(Fare) and size of family, i.e, the large family size the more fare that ticket they paid. Possibly, it was because the Fare was given the same for all member in a family, not individually different. Remember, when we discuss about missing value of Cabin B28, the two people there had the same fare value. Let's check this logic:

```
fare <- train[which(train$family.size == 2),
              c("Ticket", "Fare", "family.size",
                "Pclass","Name", "Cabin")] %>% arrange(Ticket)
head(fare)
```

```
##   Ticket    Fare family.size Pclass
## 1 110813 75.2500           2      1
## 2 111361 57.9792           2      1
## 3 111361 57.9792           2      1
## 4 113505 55.0000           2      1
## 5 113505 55.0000           2      1
## 6 113509 61.9792           2      1
##                                          Name Cabin
```

```
## 1 Warren, Mrs. Frank Manley (Anna Sophia Atkinson)   D37
## 2                       Hippach, Miss. Jean Gertrude   B18
## 3  Hippach, Mrs. Louis Albert (Ida Sophia Fischer)   B18
## 4           Chibnall, Mrs. (Edith Martha Bowerman)   E33
## 5                        Bowerman, Miss. Elsie Edith   E33
## 6                    Ostby, Mr. Engelhart Cornelius   B30
```

It seems that family members had the same ticket number would have the the same Fare. Now, we will write a function to check how correct this assumption is:

As we expected, there are only 2 cases that does not agree with our assumption. Therefore, it is actually a correlation between the family size and fare. To get rid of it, I will find the price that each person has to pay for their ticket and also fill 0 in price by the mean based on their class.

Until this point, there is no dependence of ticket fare or price. But the variable "price" is actually dependent on *Pclass*, and this happens in practice when you have to pay more to get the best service.

Variable Selection:

After cleaning data, we will split train into 2 files: train.mod and valid.mod, to train and test the model.

```
library(caret)
set.seed(3456)
train <- train %>% select(-c("PassengerId", "Name", "mean.fare", "Cabin", "n", "Ticket"))
```

```
## Adding missing grouping variables: `Ticket`
```

```
train <- train[,-1]

trainIndex <- createDataPartition(train$Survived, p = .7,
                                  list = FALSE,
                                  times = 1)
train.mod <- train[trainIndex,]
valid.mod <- train[-trainIndex,]
```

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

Forward selection suggest the model including *Title, Pclass, family.size, Age, Fare, price, Embarked* as predictors. This model produced the AIC of 752.1, which is the same AIC as the smaller model with *Title, Pclass, family.size, Age, Fare, price*. Therefore, I will drop *Embarked* out of my model.

```
keep <- c("Survived","Title", "Pclass", "family.size", "Age", "Fare", "price")
```

```
## [1] "Mr"       "Mrs"      "Miss"     "Master"   "Officer"  "Dr"       "Officera"
```

```
## [1] "Mr"      "Mrs"     "Miss"    "Master"  "Officer" "Dr"
```

Models:

```
train.mod <- train.mod[, keep]
valid.mod <- valid.mod[, keep]
keep.test <- keep[-1]
test.mod <- test[, keep.test]
```

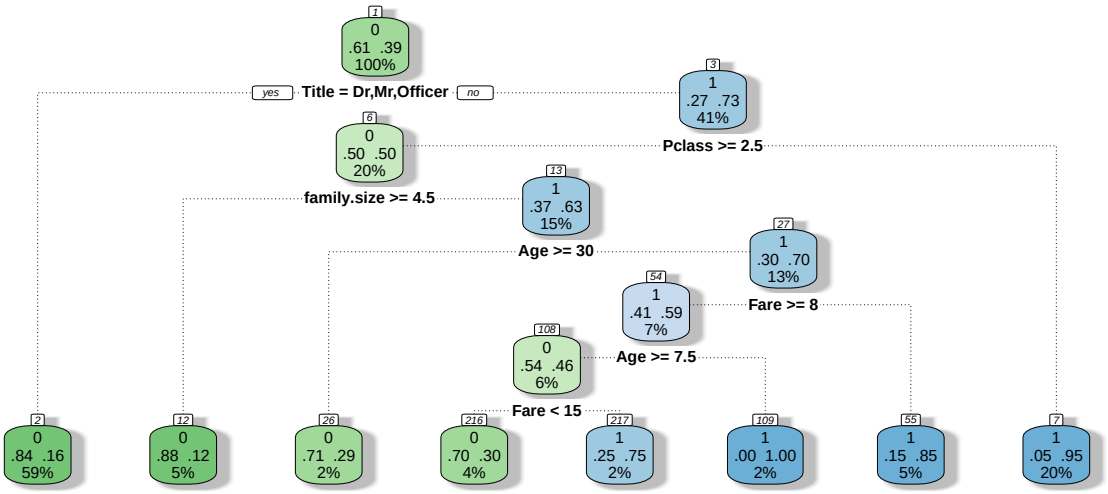
1. Logistic Model:

```
##
```

```
## Call:
## glm(formula = Survived ~ ., family = binomial, data = train.mod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5568  -0.5872  -0.3609   0.4919   2.6318
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.355818   1.151668   2.046 0.040799 *
## TitleMaster   3.333205   1.097964   3.036 0.002399 **
## TitleMiss     2.960671   0.954149   3.103 0.001916 **
## TitleMr      -0.331128   0.914050  -0.362 0.717154
## TitleMrs      3.466760   0.962435   3.602 0.000316 ***
## TitleOfficer -0.237215   1.222851  -0.194 0.846187
## Pclass       -1.158573   0.204345  -5.670 1.43e-08 ***
## family.size  -0.454559   0.095146  -4.777 1.78e-06 ***
## Age          -0.024760   0.010847  -2.283 0.022451 *
## Fare          0.007311   0.004870   1.501 0.133297
## price        -0.004446   0.010768  -0.413 0.679685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 834.27  on 623  degrees of freedom
## Residual deviance: 499.24  on 613  degrees of freedom
## AIC: 521.24
##
## Number of Fisher Scoring iterations: 5
```

2. Decision Tree:

- Decision Tree is a supervised learning method.
- Decision Tree is a graph to represent choices and their results in a form of a tree.

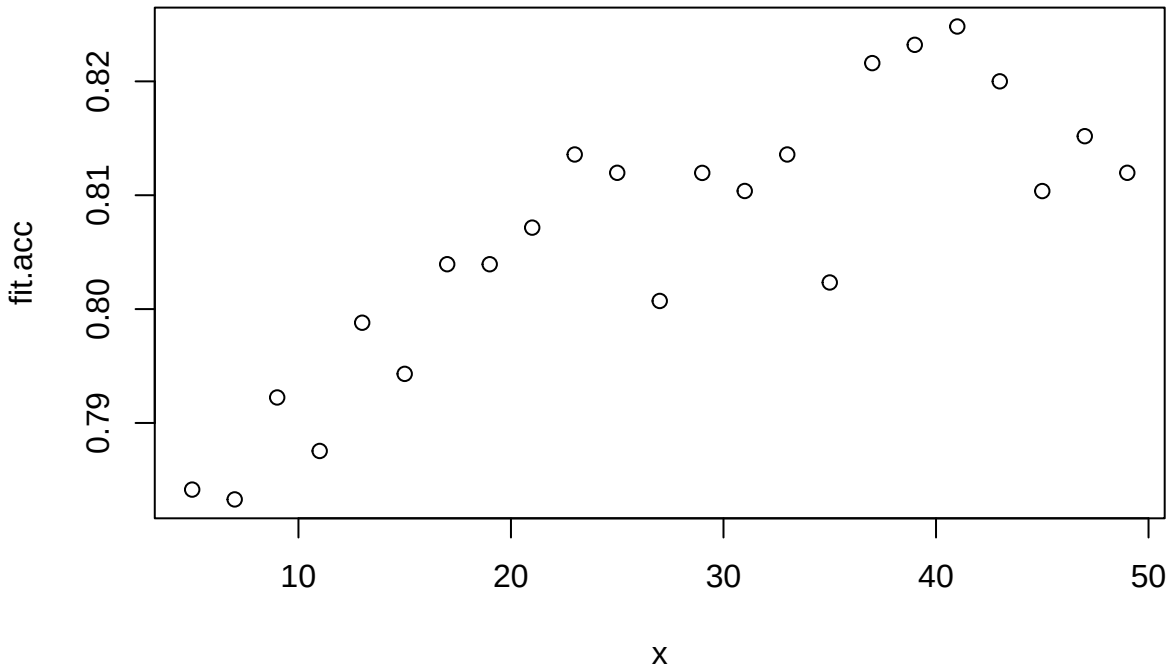


Rattle 2022−Apr−06 00:46:16 huongtran

```
tree.fit <- predict(tree, train.mod, type = "class")
tree.tab.fit <- table(train.mod$Survived, tree.fit)
tree.acc.fit <- sum(diag(tree.tab.fit)) / sum(tree.tab.fit)
```

3. Random Forest:

- Random forest creates several random Decision Tree output is the aggregate of those trees.




```
rf <- randomForest(as.factor(Survived) ~., data = train.mod, ntree = num.tree)
rf.acc.fit <- sum(diag(rf$confusion)) / sum(rf$confusion)
```

SVM:

- SVM (Support Vector Machine) is a supervised learning methods, which used to classified data.
- Figure in wiki

```
library(e1071)

svm.mod <- svm(as.factor(Survived) ~., data = train.mod, scale =  F)
svm.tab <- table(as.character(train.mod$Survived), svm.mod$fitted)
svm.fit.acc <- sum(diag(svm.tab)) / sum(svm.tab)

#make prediction:
```

Comparsions:

Now, we will make comparision using the validation sets that we created above

	Summary results		
	logistic	tree	random.forest
Fit accuarcy	0.8269231	0.8589744	0.8151881
Prediction accuracy	0.8202247	0.8239700	0.8014981

Interpretation of Losgistic model:

Summarizing Predictive Power: