Titanic Dataset with survival prediction

Huong Tran

11/26/2021

Objective:

- The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City.
- There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history.
- This project is a competition on Kaggle with target of predicting which passengers survived the Titanic shipwreck by machine learning.

Overview about our data:

	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

Overview about our data:

	NA.train	Empty.train	Percent.train	NA.test	Empty.test	Percent.test
PassengerId	0	0	0	0	0	0
Pclass	0	0	0	0	0	0
Name	0	0	0	0	0	0
Sex	0	0	0	0	0	0
Age	0	177	19.87	0	86	20.57
SibSp	0	0	0	0	0	0
Parch	0	0	0	0	0	0
Ticket	0	0	0	0	0	0
Fare	0	0	0	0	1	0.24
Cabin	0	687	77.1	0	327	78.23
Embarked	0	2	0.22	0	0	0

Figure 2: Missing value in train and test data set

EDA:

Some insights about variable *Name*

- *Name* does not contain any missing and have 891 unique values.
- Contain information about title of person, which indicates their social class and profession.
- Extract *Title* from *Name* and combine it into train set.

```
head(train.org$Name, 3)
```

```
## [1] "Braund, Mr. Owen Harris"
## [2] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
## [3] "Heikkinen, Miss. Laina"
```

EDA

Some insights about variable *Name*

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

Figure 3: Unique Title

EDA

Some insights about variable *Name*

- *Title* relating to army and “Rev” are less likely to survive and they are male -> “Officer”.
- *Title* “Mme”, “Mlle”, “th”, “Lady”, “Ms” represent unmarried women -> “Miss”.
- “Jonkheer” -> “Mr”.
- The only one value of “Lady” and “Sir” are spouse -> “Mrs” and “Mr”.

EDA

Some insights about variable *Name*

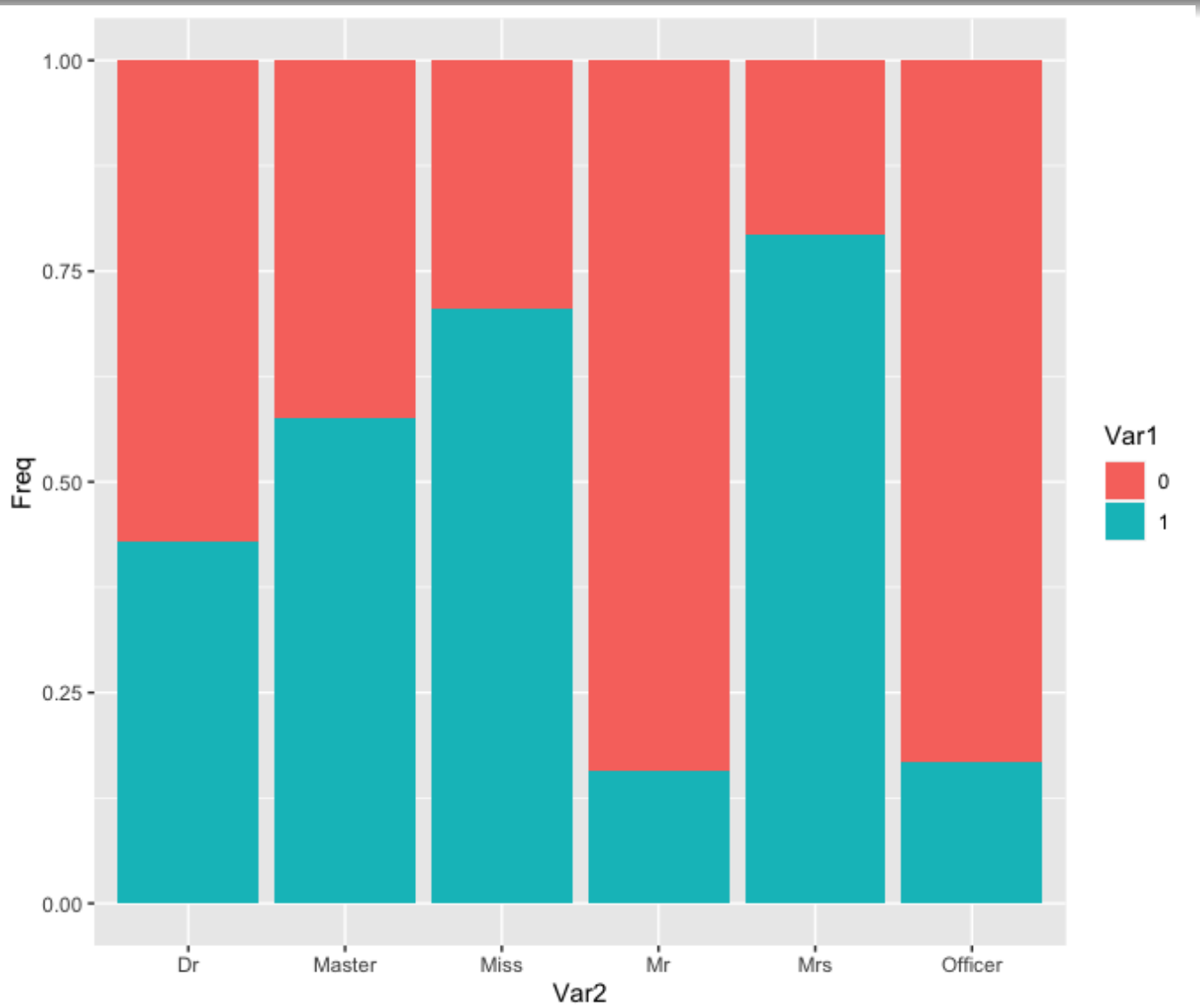


Figure 4: Ratio of survival based on *Title*

EDA

Some insights about variable *Name*

- Number of different levels in *Title* variable to 6.
- "Mr" and "Officer" has the least chance of survival.
- "Miss" has the highest chance of survival.
- this variable also contains information about *Sex*.

EDA

Age

- 177 missing value in train dataset, which accounts for 19.87% of total observation.
- 86 missing value in test dataset, which account for 20.57% of total observation.

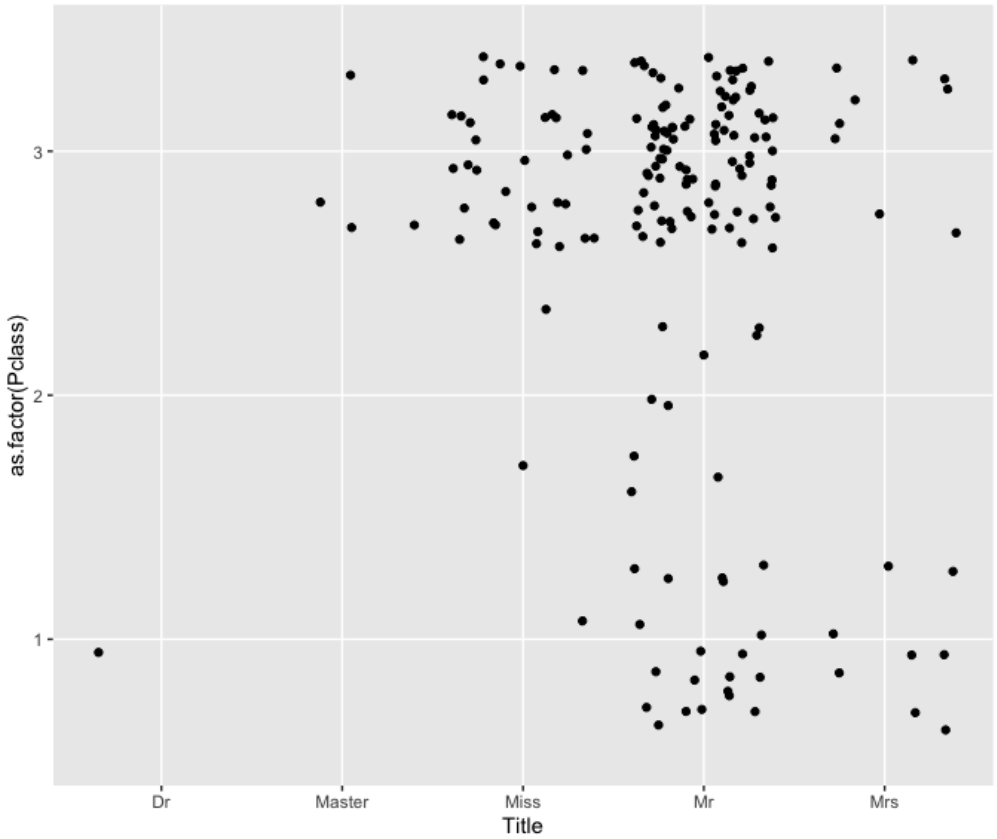


Figure 5: Passengers with missing value in Age

- Impute missing value by mean of age in social class *Pclass* = *3*, with *Title* "Mrs", "Mr" and "Miss"

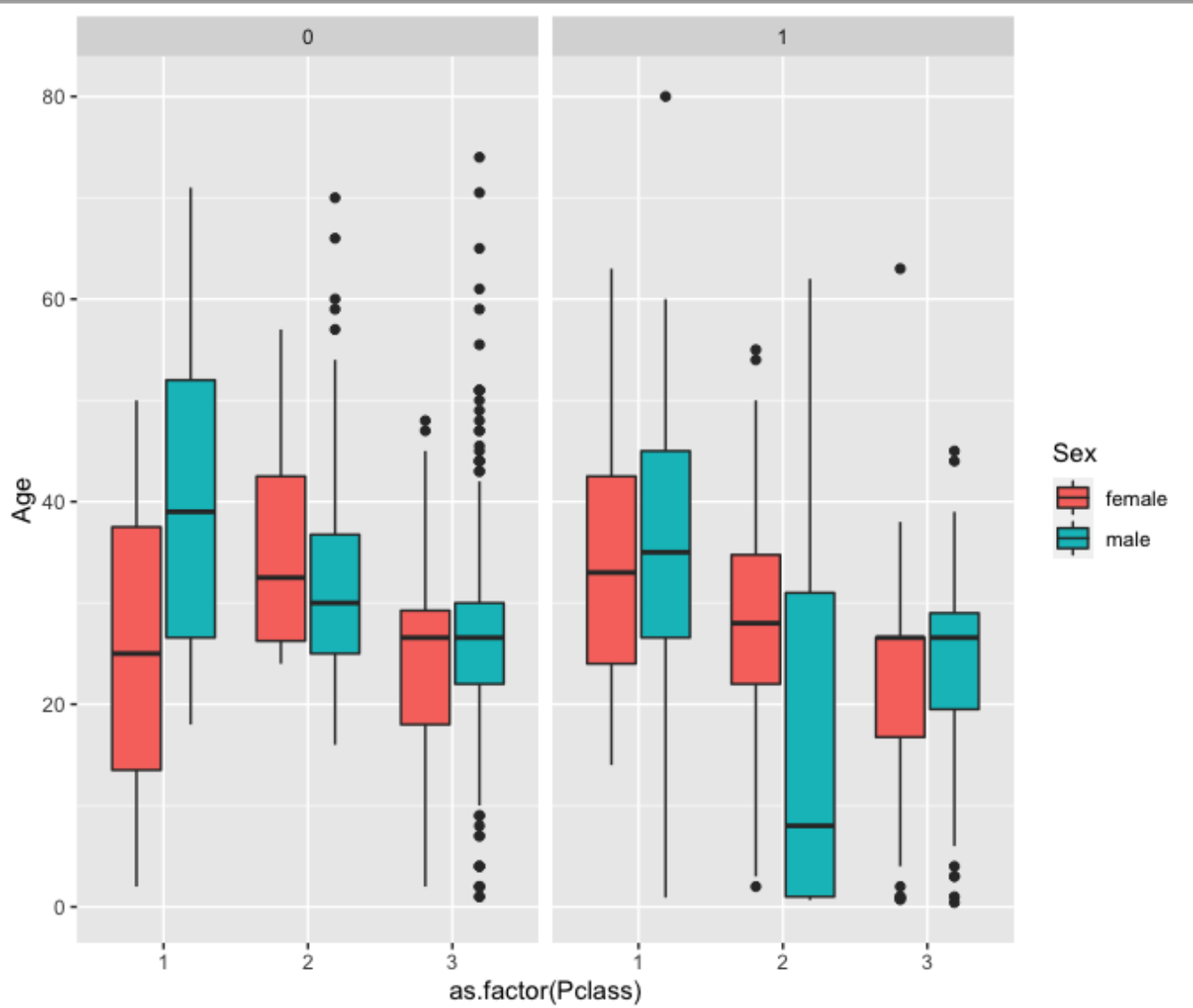


Figure 6: Boxplot of Age, different by *Pclass* and *Survived* and *Sex*

EDA

Age

- 50% of survival male in middle class was less than 10 years old.
- 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*.

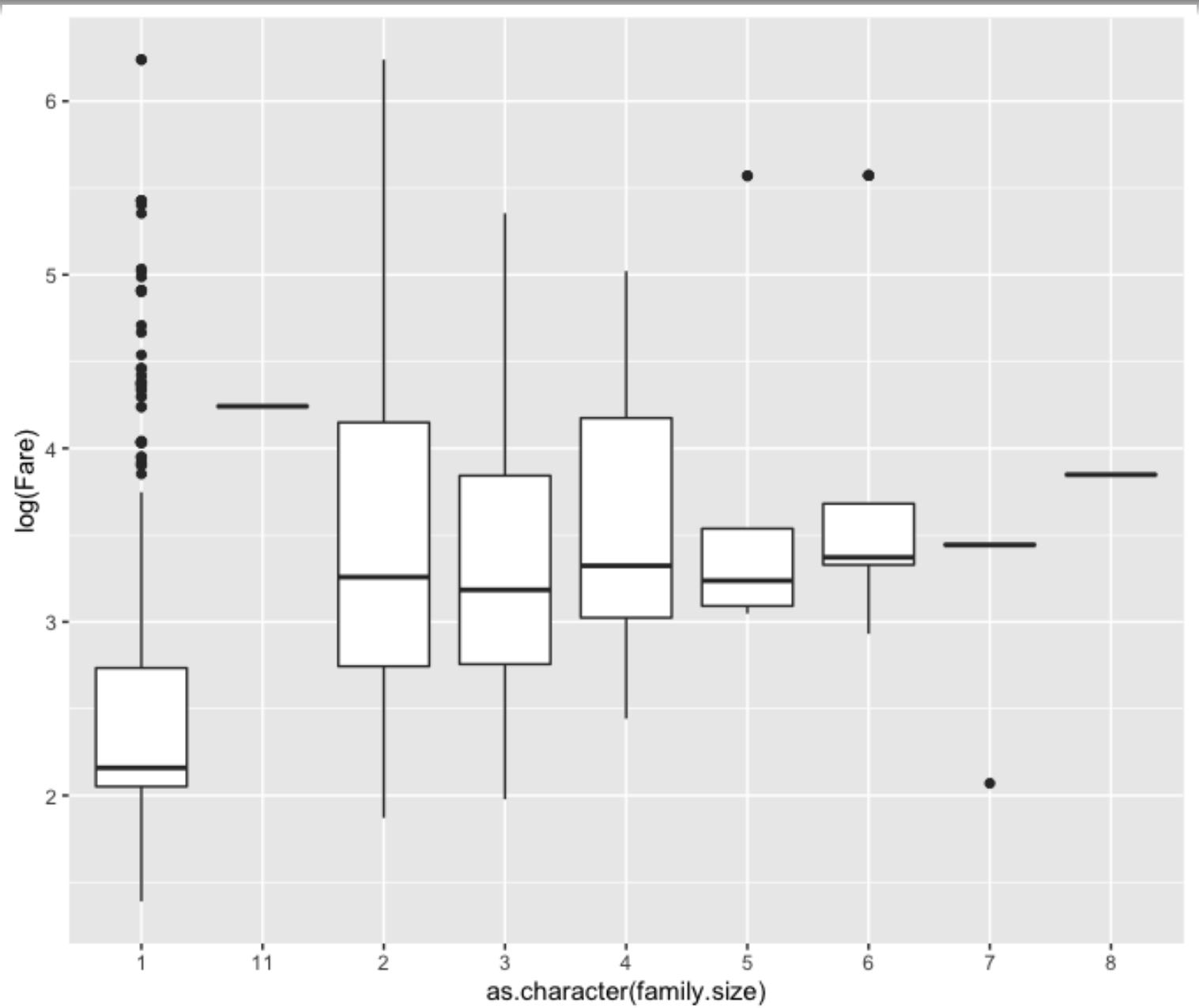
EDA

Information about Family size and total Ticket price of each family

- *SibSp*: Number of siblings / spouses aboard the Titanic.
- *Parch*: of parents / children aboard the Titanic
- They both contain information about family size, we can create a new variable as *family.size* to obtain information about passenger's family.

EDA

Information about Family size and total Ticket price of each family



EDA

Information about Family size and total Ticket price of each family

- Boxplot shows a positive trend of ln(*Fare*) and size of family.
- To get rid of the correlation, we will find the ticket price:

$$\text{price} = \frac{\text{Fare}}{\text{Family.size}}$$

- Missing value in *Fare* will imply tickets cost 0, which is impossible, we will impute missing value of *Price* by the mean based on *Pclass*.

EDA

What can Ticket class and Embarkation tell us?

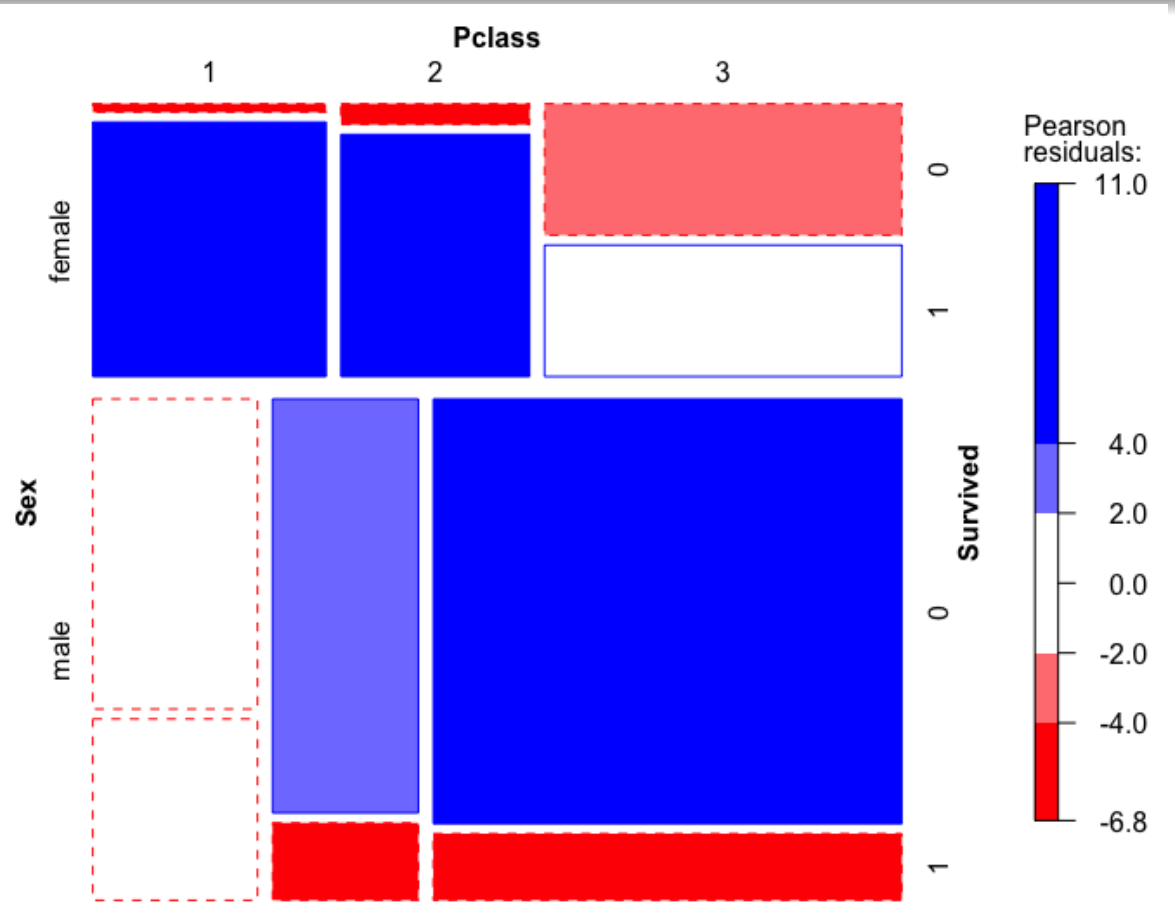


Figure 8: Mosaic plot of *Pclass*, *Sex* and *Survived*

- Women of upper class has the highest chance of survival.

EDA

What can Ticket class and Embarkation tell us?

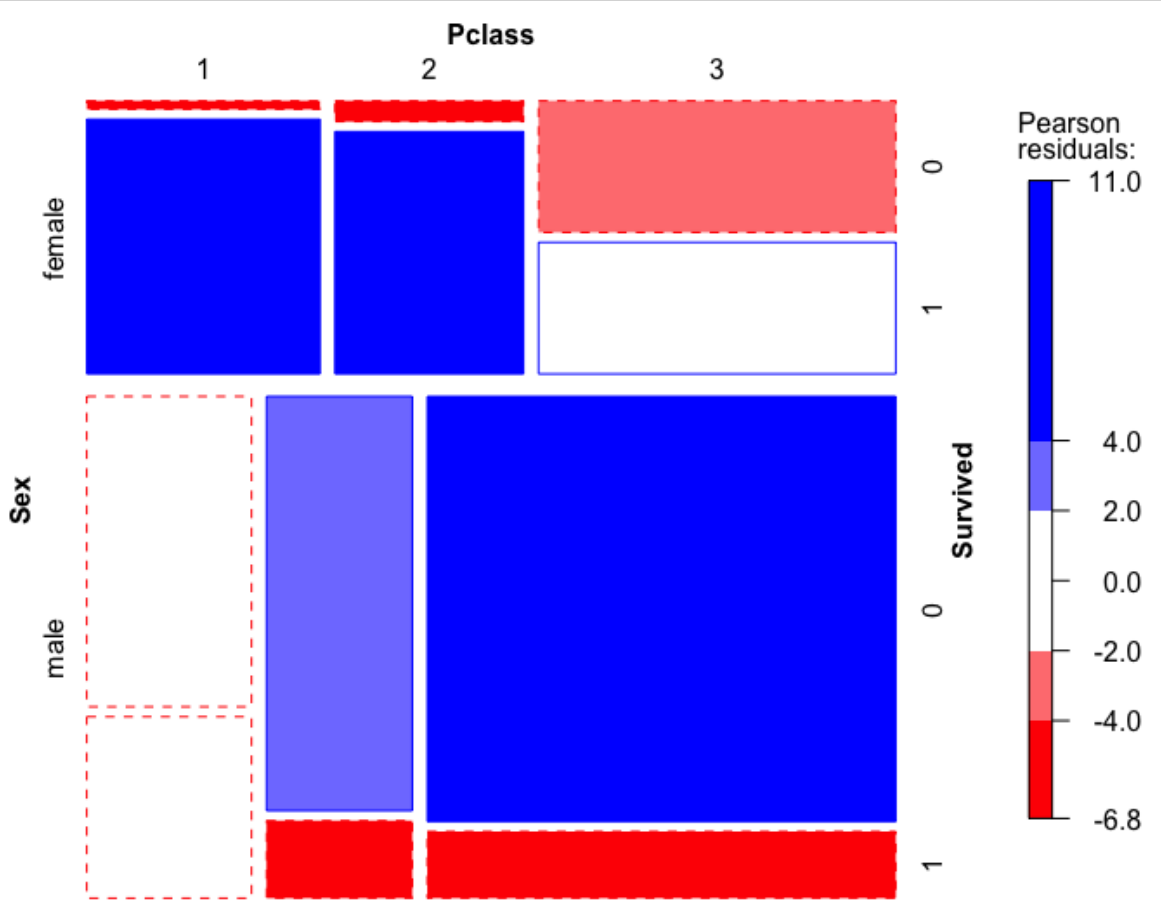


Figure 9: Mosaic plot of *Pclass*, *Sex* and *Survived*

- Men the the lower class the lowest chance of survival.

EDA

What can Ticket class tell us?

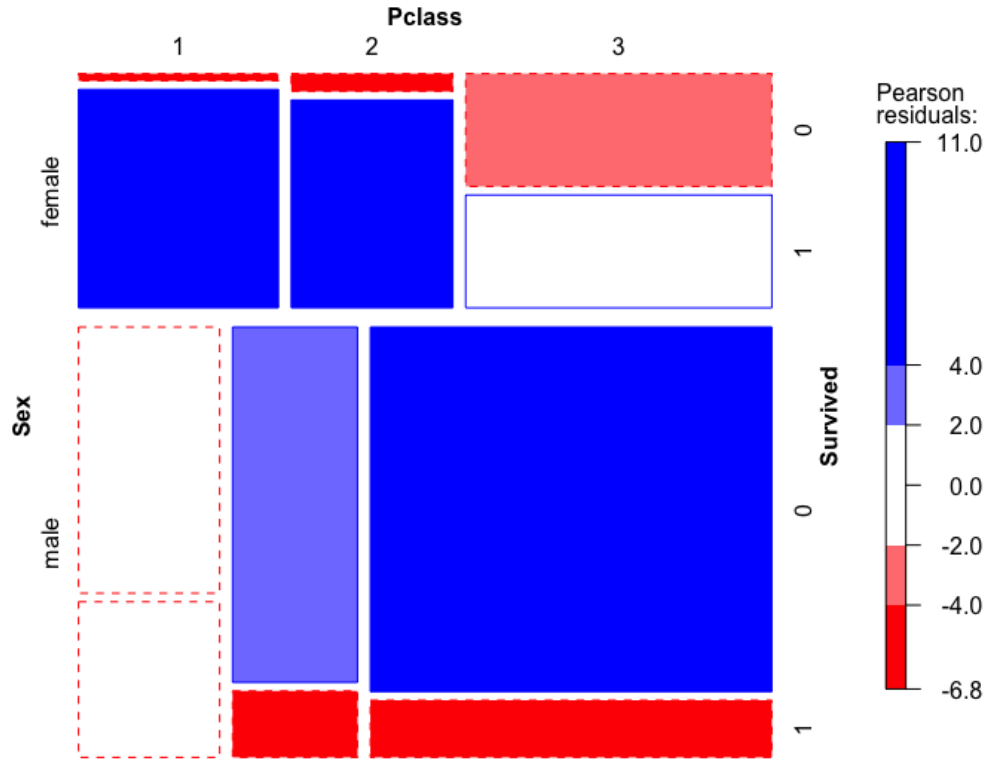


Figure 10: Mosaic plot of *Pclass*, *Sex* and *Survived*

- Total number of male is double the total number of female, but male has the lower probability of survival.

EDA

Where did they embarked?

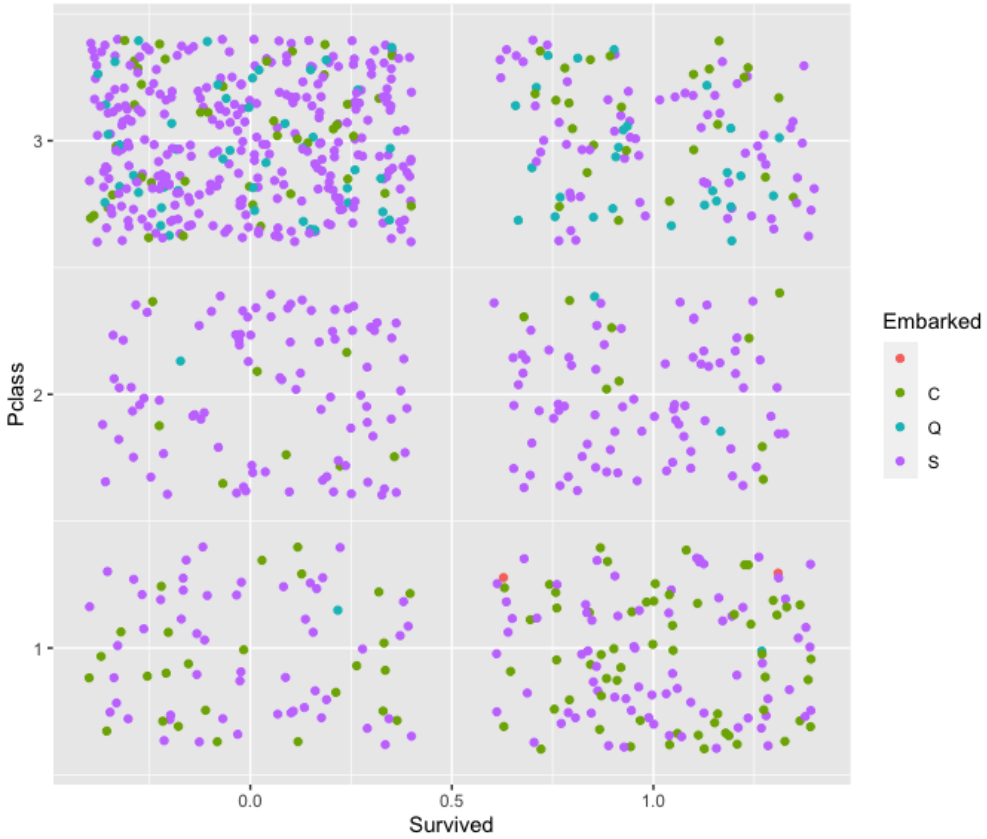


Figure 11: *Embarked* and Pclass

- The majority of passenger used Southampton (S) to embark.
- Only lower class (Pclass == 3) used Queenstown (Q).

EDA

Where did they embarked?

- *Embarked* has 2 missing values: 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.
- Passenger with Cabin in deck B used Cherbourg (C) and Southampton (S) as their embarkation -> used either S or C to impute missing data.

Model fitting

Variable Selection

- Split *train* data into: *train.mod* and *valid.mod*.
- Number of observation in train.mod is: 713
- Number of observation in test.mod is: 178

Model fitting

Variable Selection

- Forward and Backward selection suggests: *Title*, *Pclass*, *family.size*, *Age*, *Fare*, *price*, *Sex* as predictors.
- This model produced the AIC of 614.4.
- No need to use *Sex* variable as predictors.
- Drop *Fare* to avoid the collinearity with *family.size*.
- In conclusion, our model will use *Title*, *Pclass*, *family.size*, *Age*, *price*.

Model fitting

Logistic Regression

- Using cut point:

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

- Logistic Regression model provide AIC of 614.87 and the fitted accuracy is 82.04%.

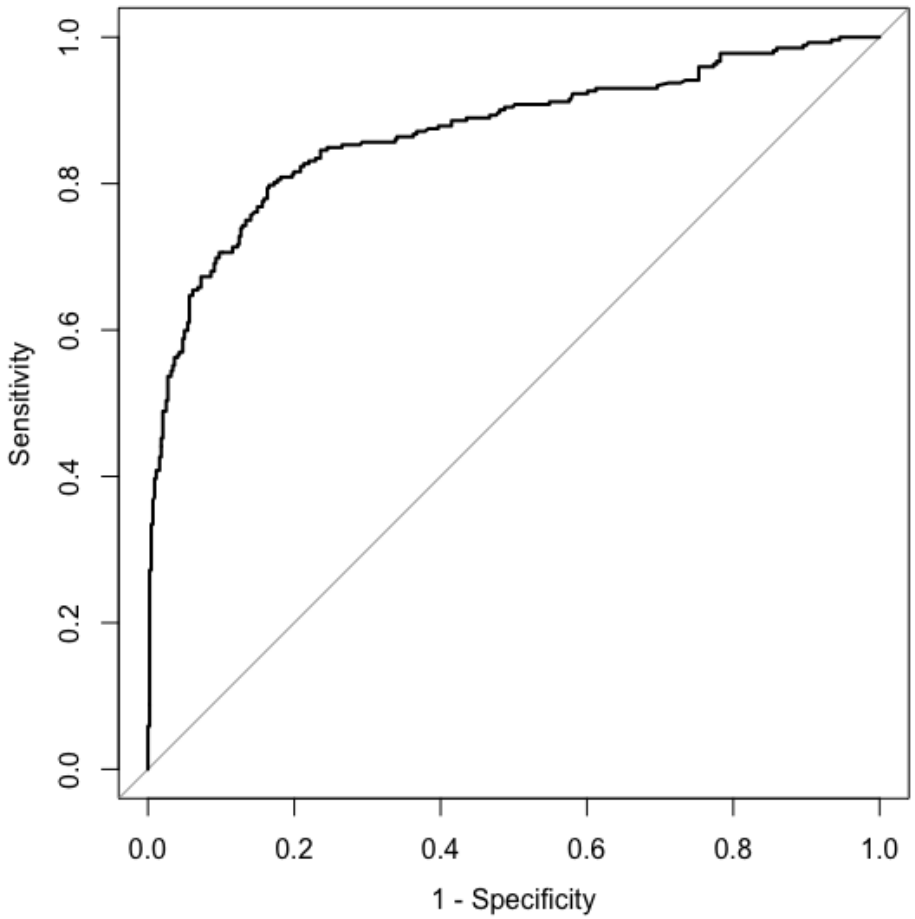


Figure 12: ROC curve of Logistic Regression

- AUC is 0.8696.

Model fitting

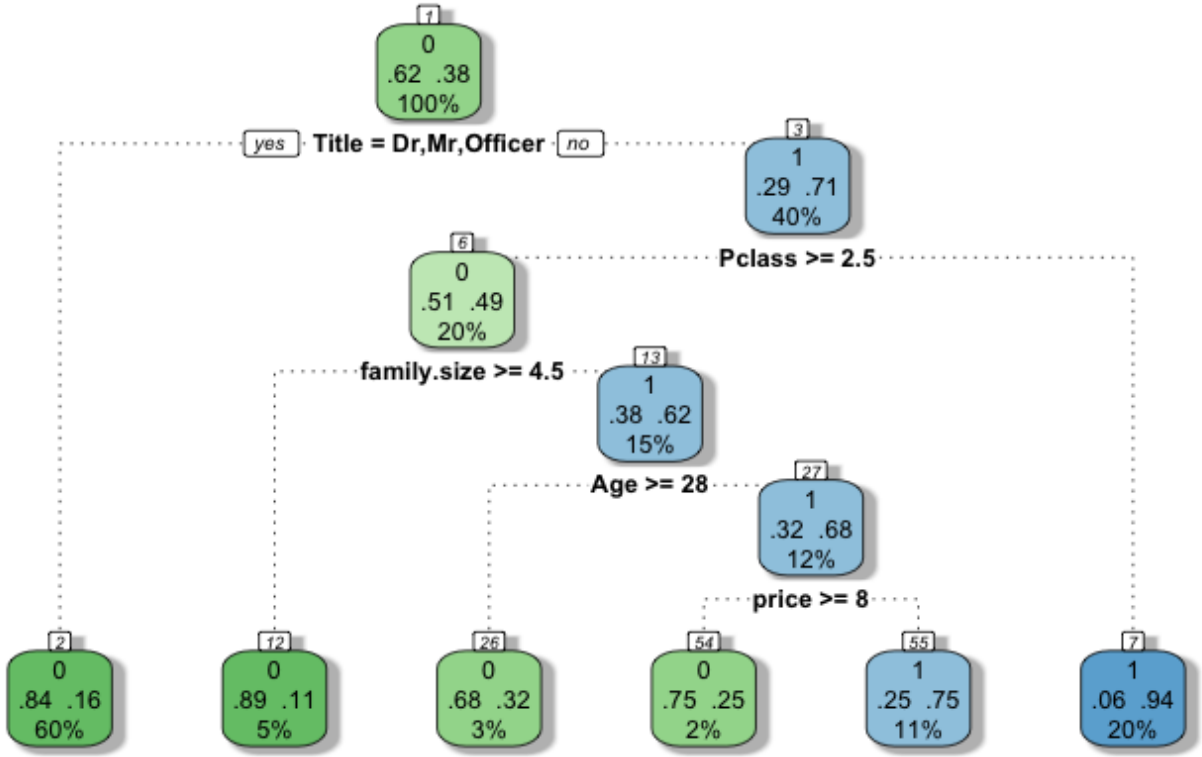
Logistic Regression

- 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.

Model fitting

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.



Rattle 2022-Apr-09 00:05:28 huongtran

Model fitting

Decision Tree

- *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.
- Smaller value of Pclass has higher probability of survive.
- The fitted accuracy of decision tree is about 84.57%, which is higher than logistic regression.

Model fitting

Random Forest

- Random forest is a supervised learning which creates several random Decision Tree and output is the aggregate of those trees.
- To get a good comparsion among all forests, we will create several forest with the same number of trees, final result is taken using the mean of result from these forests.

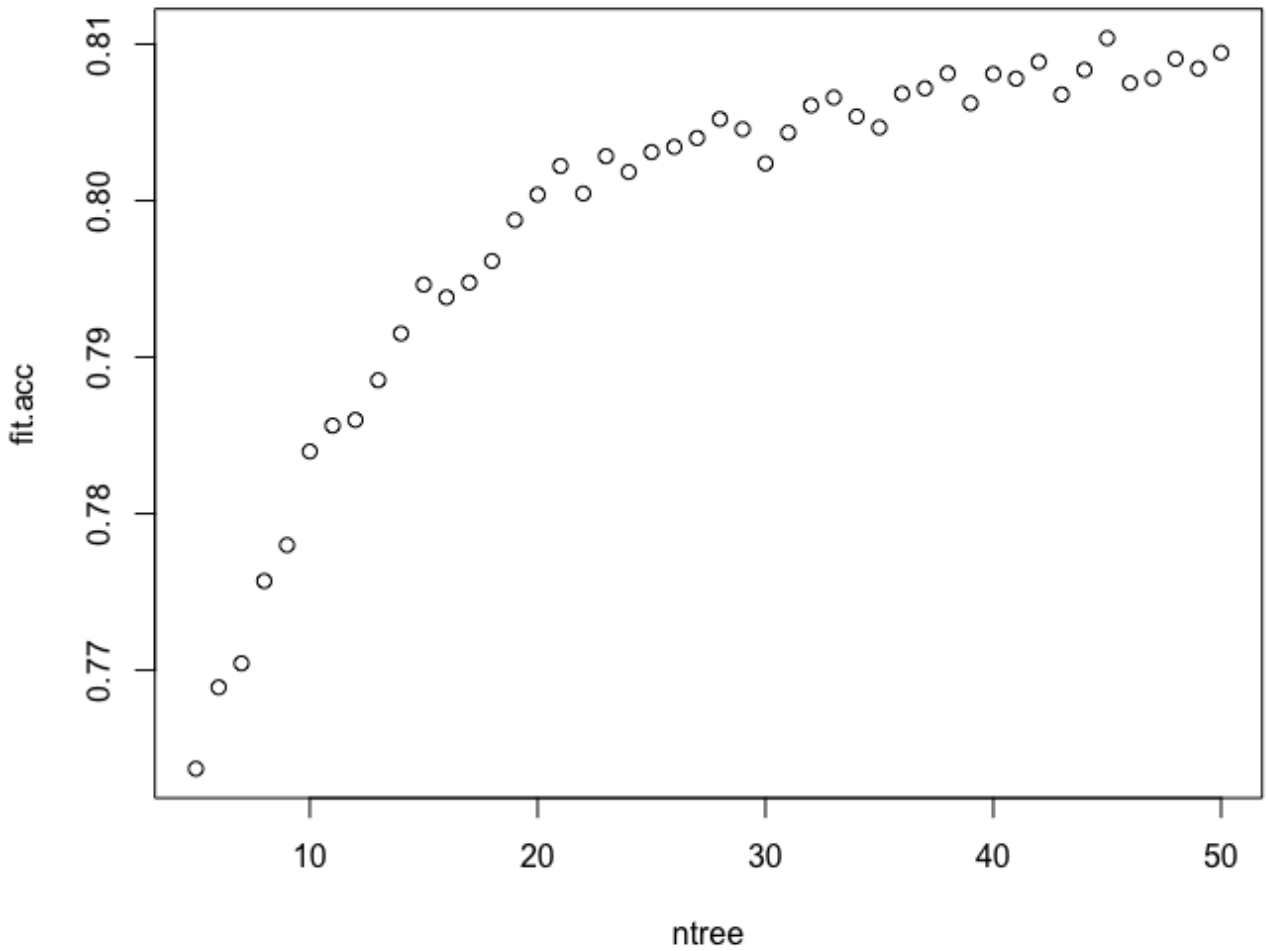


Figure 14: Fitted accuracy of random forest with different number of tree

Model fitting

SVM

- SVM (Support Vector Machine) is a supervised learning methods, which used to classified data.
- It creates a hyperplane to separate train data into 2 classes.
- The goal is to decide which class a new data point will be in.

Model fitting

SVM

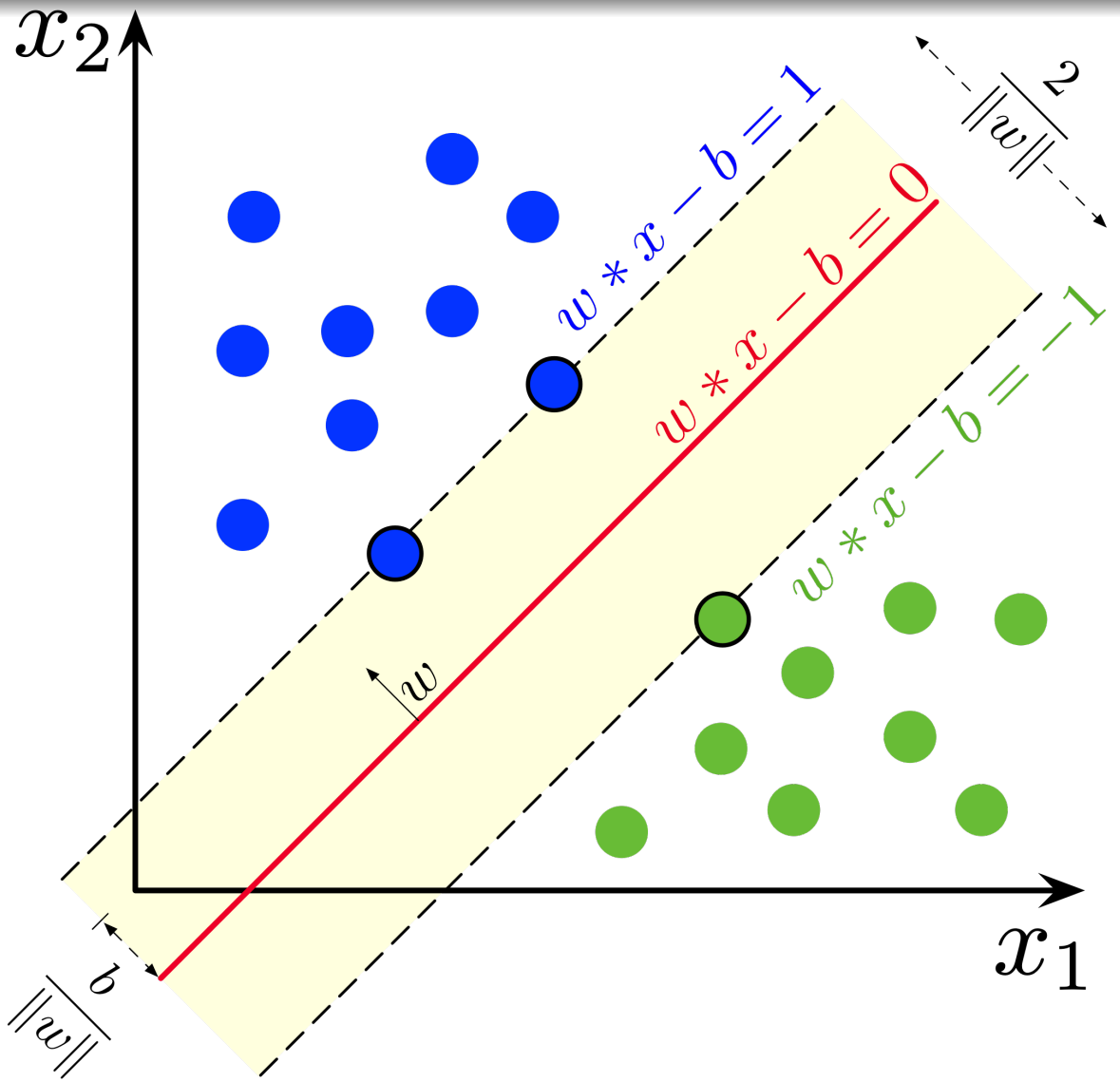


Figure 15: SVM

Model fitting

SVM

- Fitting train data into SVM model, we obtain the fitted accuracy is: 90%, which much higher that previous methods.

Model fitting

Model Comparision

	Fit accuarcy	Prediction accuracy
logistic	0.82	0.83
tree	0.85	0.86
random.forest	0.81	0.84
svm.mod	0.9	0.7

Figure 16: Comparision of fitted and predicted accuracy between models

Making prediction on *test* and submit to Kaggle:

- Cleaning test dataset with the procedure that we cleaned train dataset.
- 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 not very different between the two models:
  1. Logistic Regression 77.9%
  2. Decision Tree: 77.4%

Further question for improvement:

1. Ridge regression is expected to give better performance when handling collinearity.
2. Regression on *Age* using other variables can be applied and hopefully will bring better results.
3. *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.

Reference:

- Categorical Data Analysis, Third Edition, Alan Agresti.
- An introduction to Categorical Data Analysis, Third Edition, Alan Agresti.
- Machine Learning Algorithms -A Review, Batta Mahesh