ENGG5103 Techniques for Data Mining

Project Report

Survival Prediction on the Titanic

LIN Huangjing 1155072677

WANG Xi 1155083228

1. Background and problem

1.1 Background

This is one of the challenges from Kaggle website, with the Challenge Title "*Titanic: Machine Learning from Disaster*"[1].

The sinking of the RMS Titanic is one of the deadliest peacetime shipwrecks in history. It happened on the night of 14th April 1912 through to the early morning of 15th April in the North Atlantic Ocean on her maiden voyage from Southampton to New York City[2]. The Titanic sank quickly after bumping an iceberg, which led to the deaths of more than 1,500 people out of 2, 224 passengers and crew.

This tremendous disaster can be ascribed to various factors, such as misjudgment and wrong operation, delaying warming, improper material used to build the hull. However, deficient lifeboats prepared on the Titanic for the passengers and crew is the direct reason that caused such loss of lives. In fact, over 1,000 passengers and crew were still on board when the Titanic sank and almost of those who had to jump or fall into the water drowned quickly due to the effects of hypothermia. In the process of rescue, women, children and the upper-class were given priorities to board the safe refuge, which means those people highly likely survived in the shipwreck.

In this challenge, we are asked to analyze what category of people have higher chance to survive the tragedy, which would do benefit to make better safety regulations for ships and rescue strategy afterwards.

1.2 Problem and Data

The target of this challenge is to predict whether a passenger would survive or not. In this project, we aim to find effect methods to address this classification problem with high accuracy.

The dataset provided includes training data and testing data where the information of each person on the Titanic was offered[3], consisting of PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Far, Cabin and Embarked. Obviously, the survival labels are given in

training data, but not in testing data.
The variable descriptions are as follows:

Table. 1.1 Variable decription

Survived	Suvival (0 = No; 1 = Yes)
Pclass	Passerger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Name	Name
Sex	Sex
Age	Age
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)

There are some samples of training data below:

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

Fig.1.1 Training data samples

2. Data preprocessing

2.1 Data cleaning

a. Fill in missing values

Missing value is one of the problems in the raw data. Four attributes have such problem. They are Age, Ticket, Cabin and Embarked respectively.

As for Age column, the data complement method depends on another attribute “Parch”. Because generally when the passenger is alone, she or he is highly likely a young person. Therefore, we calculate the average ages when Parch is larger than 0 and when it is exactly 0. Then once filling in NA values of Age, we also consider different situations of Parch.

When it comes to Embarked and Fare, we employ the mode values “S” and “8.05” which occur most frequently respectively.

b. Reduce redundant features

“Name” attribute consists of the full name of each passenger as well as different titles,

such as ‘Mr.’, ‘Mrs.’, ‘Miss’, ‘Master.’ and so forth which can be simply replaced by ‘Sex’ attribute to identify gender. In addition, names make little difference to survival. As a result, “Name” is eventually removed.

Name	Sex
Braund, Mr. Owen Harris	male
Cumings, Mrs. John Bradley (Florence Briggs T	female
Heikkinen, Miss. Laina	female
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female

Fig. 1.2 Preprocessed data samples

c. Remove irrelevant features
“PassengerId is only about the sequence number for records, hence we delete this column from raw data. Likewise, “”

2.2 Coding

a. Feature selection
After our deliberation, PassengerId, Name, Ticket and Cabin are discarded, and we keep the rest 7 features: Pclass, Sex, Sibsp, Parch, Fare as well as Embarked for following processing.

b. Data transformation
As “Sex” and “Embarked” are non-digit values, we convert “male”, “female” into 1 and 2, and “C”, “Q”, “S” to 1, 2 and 3 respectively.

c. Data Normalization.
In order to make balance among all features when calculating similarity, we utilize zero mean normalization method. After normalization, for each feature, its mean value is zero and unit variance.

	A	B	C	D	E	F	G	H
1	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
2	0.826913	-0.73728	-0.61058	0.43255	-0.47341	-0.50216	0.585625	0
3	-1.56523	1.354813	0.612754	0.43255	-0.47341	0.786404	-1.94121	1
4	0.826913	1.354813	-0.30474	-0.47428	-0.47341	-0.48858	0.585625	1
5	-1.56523	1.354813	0.38338	0.43255	-0.47341	0.420494	0.585625	1
6	0.826913	-0.73728	0.38338	-0.47428	-0.47341	-0.48606	0.585625	0
7	0.826913	-0.73728	0.167653	-0.47428	-0.47341	-0.47785	-0.67779	0
8	-1.56523	-0.73728	1.836086	-0.47428	-0.47341	0.395591	0.585625	0
9	0.826913	-0.73728	-2.13974	2.246209	0.767199	-0.22396	0.585625	0
10	0.826913	1.354813	-0.22829	-0.47428	2.007806	-0.42402	0.585625	1
11	-0.36916	1.354813	-1.22224	0.43255	-0.47341	-0.04293	-1.94121	1

Fig. 2.3 the training samples after preprocessing

3. Classification methods

Titanic challenge is a typical classification task. So as to address this problem. We tried several canonical classification approaches, like Random Forest, Naïve Bayes Algorithm,

Multiple Layer Perceptron and K-Nearest Neighbors which were introduced in class. Next, we will firstly talk about the main idea behind these techniques as well as their pros and cons.

3.1 Random Forest[4]

Random Forest is an ensemble learning method which can be used to solve classification, regression and other tasks. Random Forest operates by constructing range of decision trees at training step and outputting the class which is the mode of the classes for classification task or mean prediction for regression task of the individual trees [5].

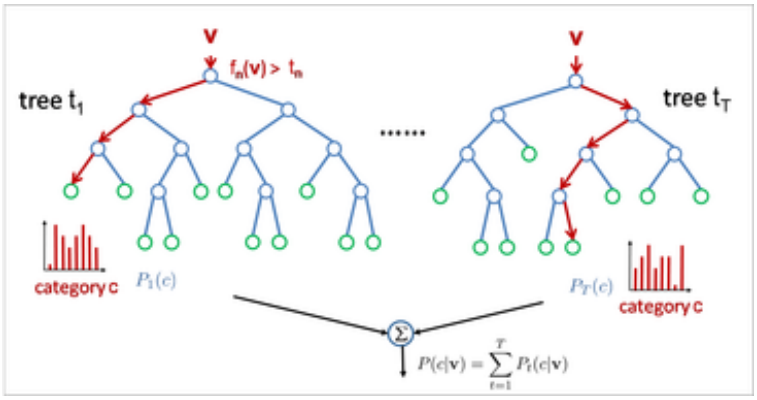


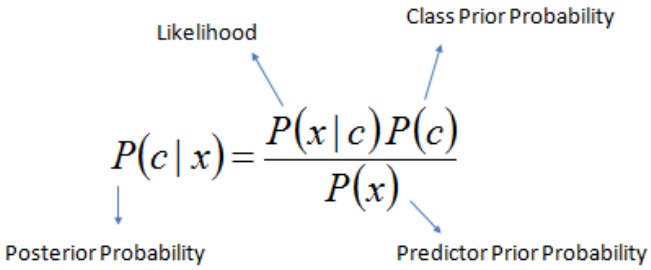
Fig. 3.1 Random Forest models

Random Forest has plenty of advantages. First of all, it can process high dimension data without feature selection. In other words, it is not necessary for researchers to pick up feature, which would help to save much effort and time. Besides, this method can be readily implemented in distributed system. It is high effective. In addition, it employs the technique of voting by major, therefore, it is strongly robust. On the contrary, there are also some drawbacks. For example, the robustness is proportional to the number of trees. That is to say, the more trees, the more robust. However, if the trees are deficient, it might be instable. Moreover, random forest is not effect to handle unbalanced dataset as the result may be partial to the class with more samples.

In this project, we constructed 500 decision trees by randomly selecting features of passengers, then classification results were used to be voted by major to determine the final class of the random forest classifier.

3.2 Na ïve Bayes Algorithm[6]

Na ïve Bayes classifiers are kind of simple probabilistic classifiers based on Bayes' theorem with independence assumptions between the features in machine learning [7].



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 3.2 Principle of Naïve Bayes Classifier

It is a very efficient approach with fast training and test. Furthermore, it is not sensitive to irrelevant data. By contrast, the limitation of Naïve Bayes method is that the features must be mutual independent.

In the project, each of the selected features does not dependent on any other features, so Naïve Bayes method can be used to such case. For parameters setting, we choose two data distributions, one is "kernel" which means kernel smoothing density and the other one is "normal" which stands for normal Gaussian distribution.

3.3 Multi-layer Perceptron (MLP)[8]

A multi-layer perceptron is a feedforward artificial neural network that consists of input layer, hidden layers and output layer[9]. MLP utilizes a supervised learning technique called backpropagation during training phase. Deep networks enjoy the great reputation of having significantly excellent representational power. It performs well with strong capability for classification tasks driven by large dataset. Besides, it can interpret high-level features which is always called semantic features.

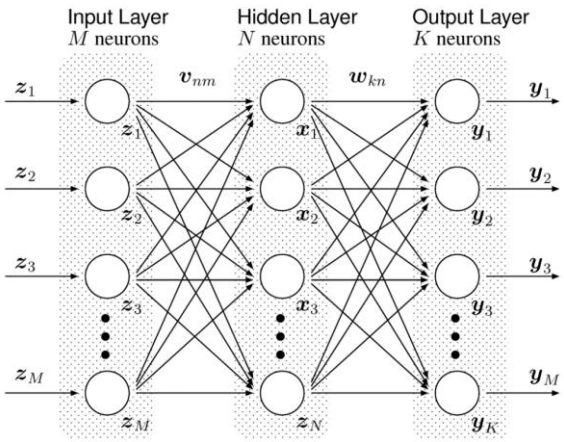


Fig. 3.3 MLP models

As for its demerits, it is hard to be trained due to deep layers. And overfitting problems occur when the training data is not huge enough.

We choose Deep belief network (DBN)[10] to train the classifier. DBN is a generative graphical model composed of multiple layers with connections between the layers but not between nodes within each layer. The number of layers is 100.

3.4 K-Nearest neighbors (KNN)[11]

KNN is among the simplest of machine learning algorithms, a lazy learning strategy[12]. This non-parametric method is easy to implement and of high effect. Undeniably, it also has some limitations. For starter, it requires large memory to store all the training data, thus it is extremely memory-consuming. And also, when the training data is huge, it shows low speed.

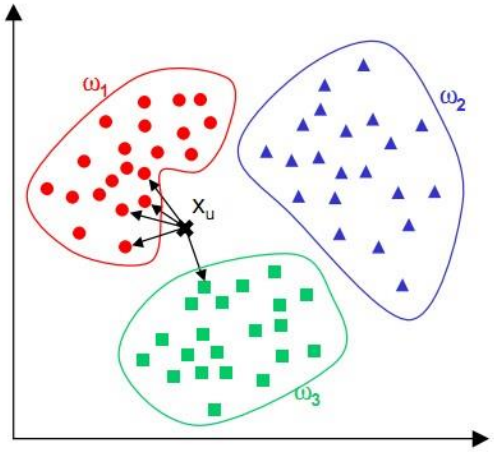


Fig 3.4 K-Nearest Neighbors

The most of computation of KNN is for similarity (distance) calculation. Therefore, it is merely suitable for small dataset instead of large one.

There are generally many distance measurement, such as Euclidean distance, City block, Minkowski distance, Cosine similarity, Canberra metric and so forth. When we employ KNN, we test two distance measurements, Euclidean distance and City block. As for the value of K, we set it 5.

4. Results and analysis

4.1 Results

During training phase, due to lack of groundtruth of testing data, we initially split training data into two groups, one third is for training and the rest is for validation.

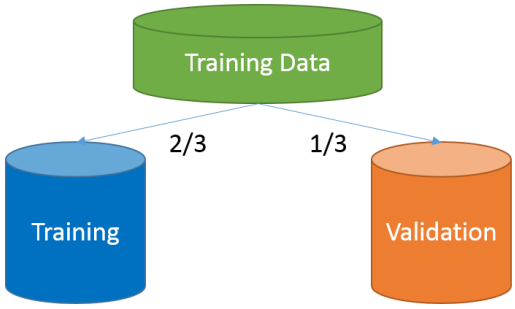


Fig 4.1 Training data

When different classifiers were trained, testing data are feed to output the predictions. We uploaded our results on Kaggle website to evaluate their accuracy. The table below shows the results of all classifiers.

Table. 4.1 Results of different classification methods

Method	Accuracy (Training Data)	Accuracy (Testing Data)
Random Forest	0.8152	0.76555
Na ïve Bayes Algorithm (Normal)	0.7849	0.73206
Na ïve Bayes Algorithm (Kernel)	0.7524	0.71292
Multi-Layer Perceptron	0.7785	0.77512
K-NN (Cityblock)	0.8066	0.76077
K-NN (Euclidean)	0.8045	0.78469

As the table demonstrates, during validation on training dataset, although Random Forest shows the best performance in accuracy (also called degree of fitting), however not achieve the best testing result. The reason might be ascribed to overfitting during training. KNN with Euclidean distance performs stably in validation and testing, which produces accuracy with 80.5% in training data and similar accuracy with 78.5% in testing data. Besides, Multi-Layer Perceptron also displays its stability with 0.3% drop of accuracy (77.5%) in testing step. Moreover, MLP is the second best approach we tried for this classification problem, followed by Random Forest and KNN (Cityblock), with 76.6% and 76.1% respectively. On the other hand, Na ïve Bayes algorithm has the worst performance, with merely 73.2% and 71.3% of accuracy using two different data distributions.

4.2 Feature analysis

Apart from the accuracy, we also analyze the relationships between features and survival.

a. Feature importance

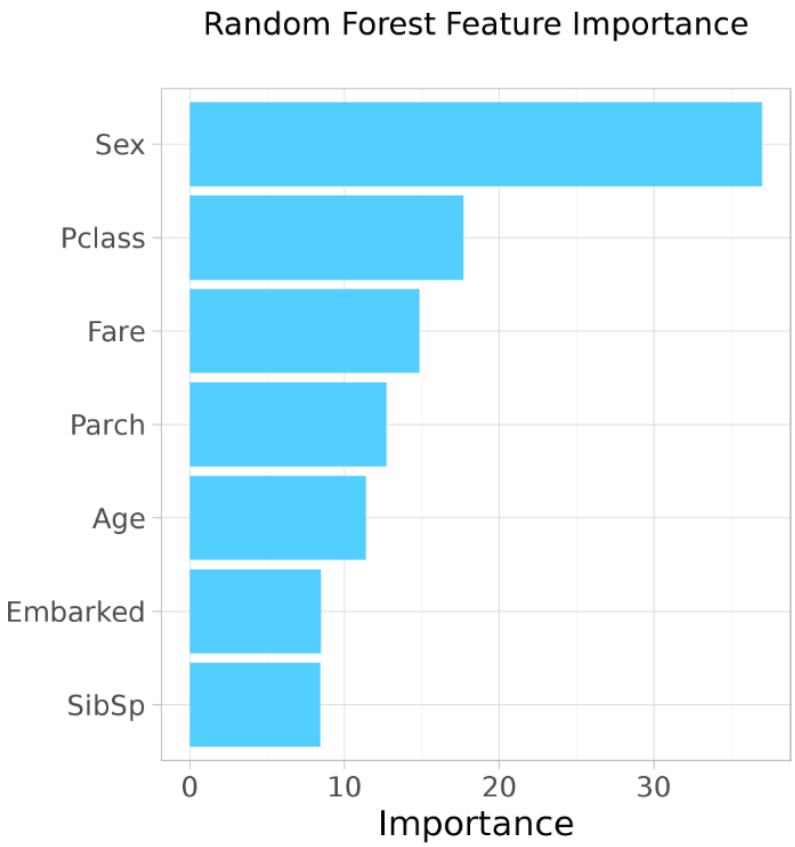


Fig4.2 Feature importance

The feature importance that the figure shows above is resulted from Random Forest method. Obviously, Sex and passenger class (Pclass) occupies lion's share, followed by Fare, Parch and Age. On the contrary, Embarked and Sibsp are the last less important attributes among all features.

b. Relationship between Sex/Age and Survival

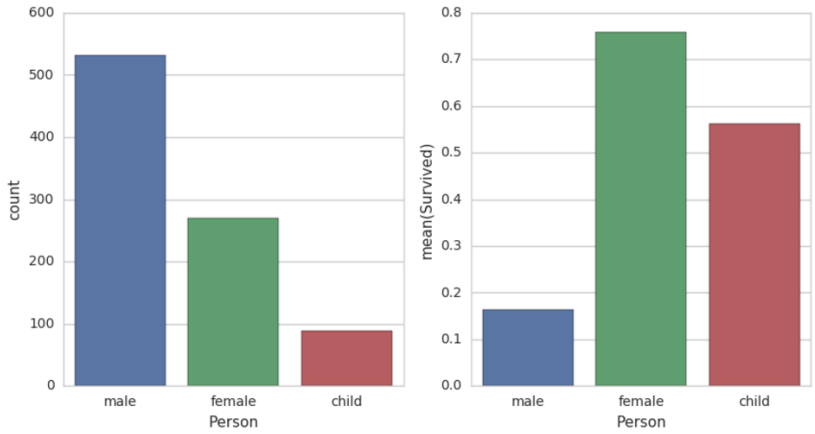


Fig4.3 Relationship between Sex and Survival

From the figure, it is quite easy to discover that the number of male is much larger than female. However, the percentage of survival among female considerably overwhelms that of male. Children and young people are more likely to survive.

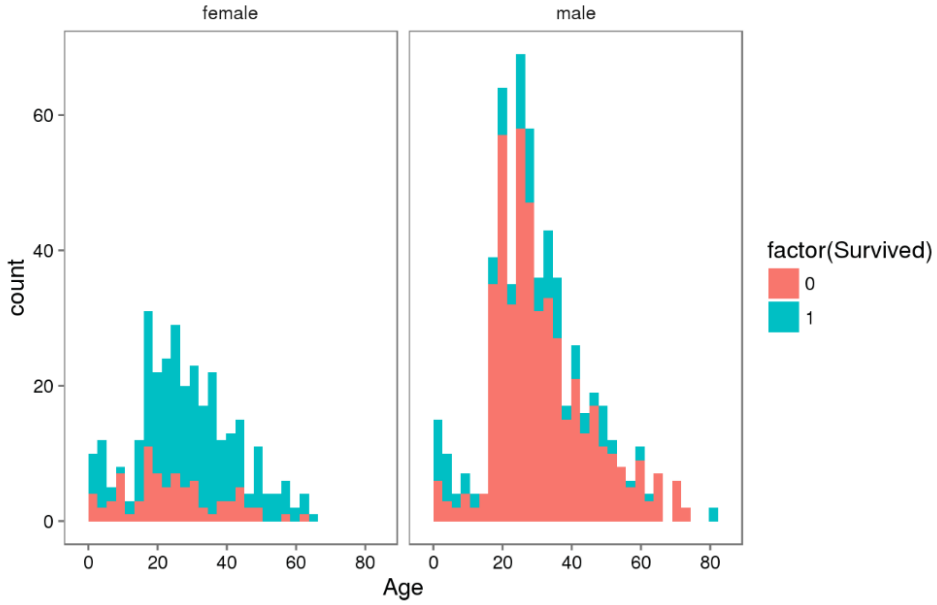


Fig4.4 Relationship between Age and Survival

Because we know that women have the higher chance than men do from Fig4.3. Here we only focus on the Age. In the right figure regarding male, children and young person range from 0 to 45 years old seems to highly likely survive the disaster.

c. Relationship between Pclass and survival

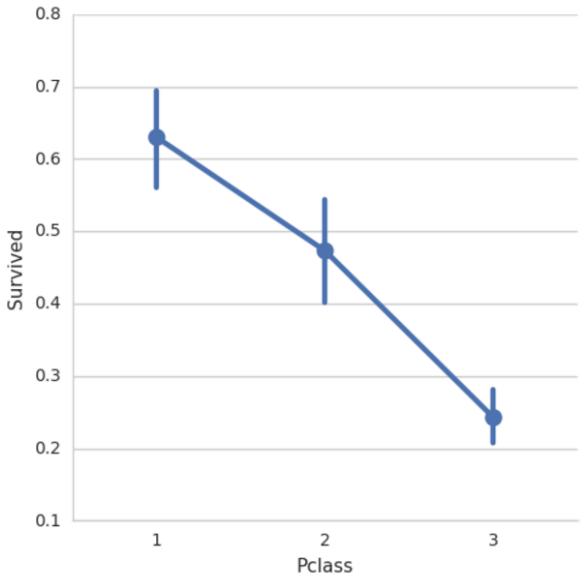


Fig 4.5 Relationship between Pclass and Survival

As the Fig 4.5 demonstrates, the upper-class of passengers have probability with around

63% to survive. Furthermore, the survival rate decreases dramatically with the drop of class level.

d. Relationship between Family size and survival

As Sibsp means the number of siblings/spouses aboard and Parch stands for the number of children/parents aboard, thus these two items could be combined together to form a new feature Family size. The relationship between family size and survival is shown below.

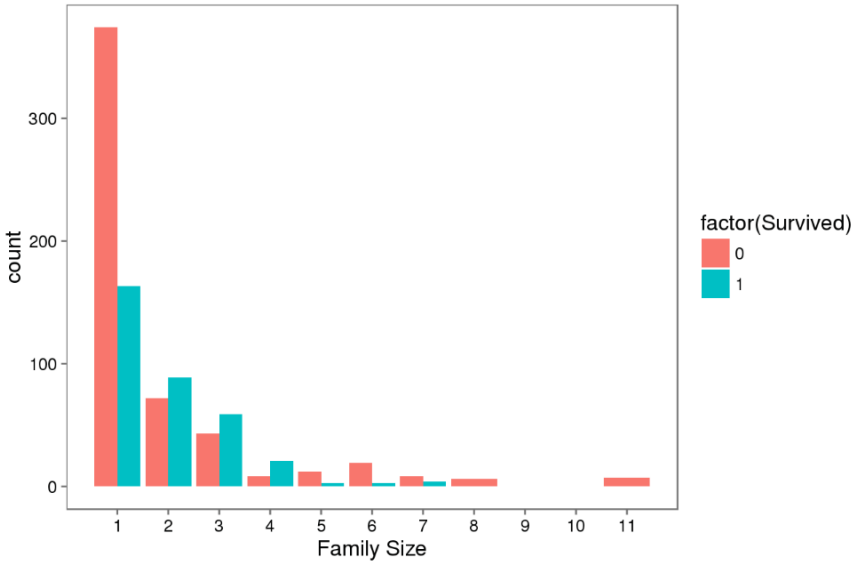


Fig 4.6 Relationship between family size and survival

Undoubtedly, the majority of passengers are alone, however with very low survival rate compared to other groups. For a better visualization, different family sizes are divided to several groups and the percentages of survival are also calculated as below.

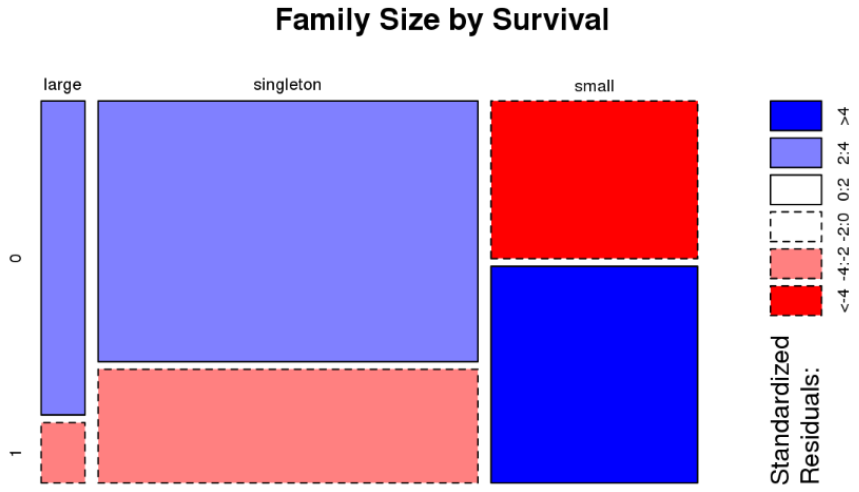


Fig. 4.7 Family size by Survival

It is clear that there's a survival penalty to singletons and those with family sizes above 4.

e. Relationship between Fare and Embarked

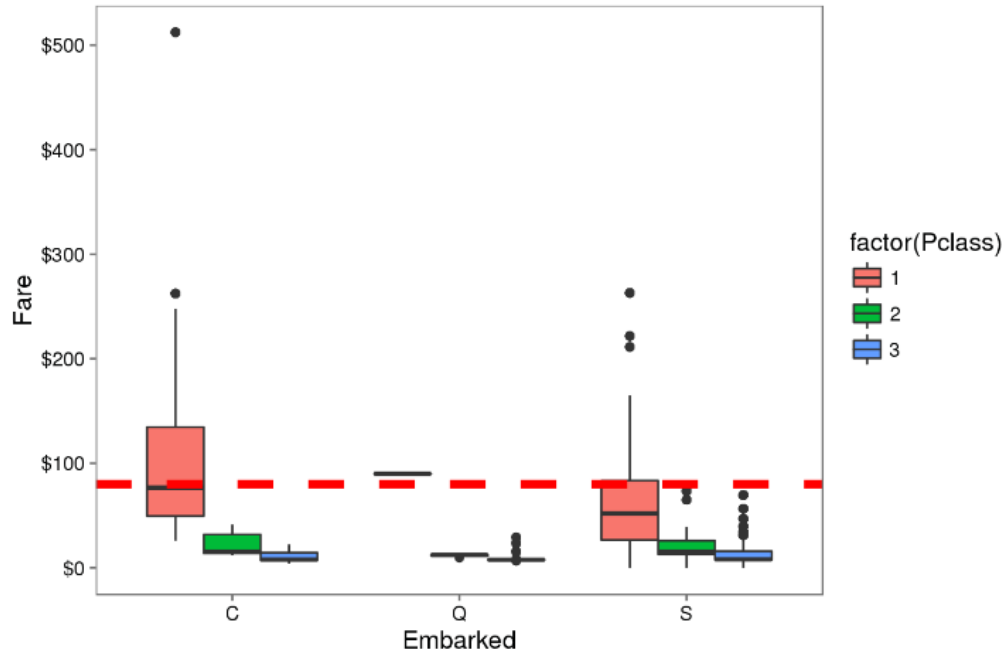


Fig. 4.8 Relationship between Fare and Embarked

As there are some missing values in Embarked column, it is wise to dig the possible relationship among the port that the passenger embarked and their fare. The median fare for a first class passenger departing from Charbourg ('C') coincides nicely with the $80 paid by our embarkment-deficient passengers. Hence, we think we can safely replace the NA values with 'C'.

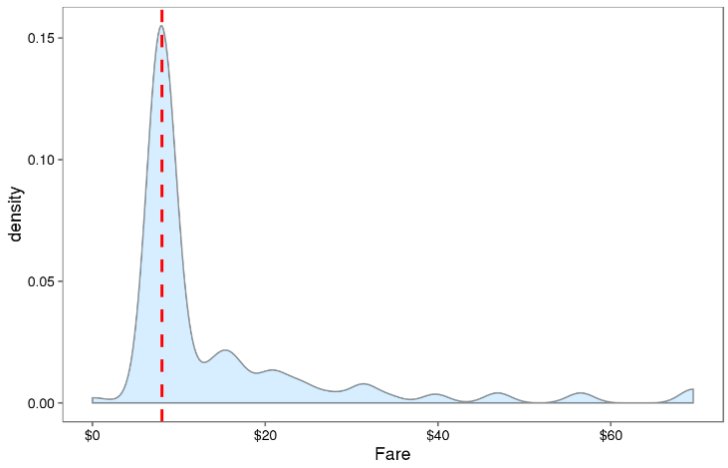


Fig. 4.9 The density of Fare

From this visualization, it seems quite reasonable to replace the NA Fare value with median for their class and embarkment which is $8.05.

5. Conclusion

We compared 4 different classification approaches, and found that KNN with Euclidean distance has the best performance (testing accuracy with 78.5%) on prediction of survival on the Titanic, followed by Multi-layer Perceptron. We also analyze the feature importance and the relationship between features and survival. Surprisingly, the gender and passenger class make great significant difference on survival. In other words, the female and the upper-class have higher probability to survive, which coincides to the fact that during the rescue in history, the woman and the upper-class were given priority to aboard on lifeboats. Besides, most of the passengers without families on aboard or the family size over 4 died in the shipwreck. The first class departing from Cherbourg paid $80 which can be used to fill in missing values in Embarked. Moreover, $8.05 occurs most frequently in Fare. Hence it can be used for data implementation.

There are still weakness of our work. First of all, the data preprocessing is not enough. In addition, the robustness of classification methods is not strong, which leads to big fluctuation between validation and testing. Lastly, the accuracy of classifiers need to be improved in the following work.

6. References

[1]https://www.kaggle.com/c/titanic
[2]https://en.wikipedia.org/wiki/Sinking_of_the_RMS_Titanic
[3]https://www.kaggle.com/c/titanic/data
[4]Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
[5] https://en.wikipedia.org/wiki/Random_forest
[6]Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural networks, 1989, 2(5): 359-366.
[7] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[8]Russell S J, Norvig P, Canny J F, et al. Artificial intelligence: a modern approach[M]. Upper Saddle River: Prentice hall, 2003.
[9]https://en.wikipedia.org/wiki/Multilayer_perceptron
[10]https://en.wikipedia.org/wiki/Deep_belief_network
[11]Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.
[12] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm