Survival Rate in Titanic Disaster

1. Goal of the project

I would use the data on Titanic passengers information to conduct descriptive and predictive analysis. In the descriptive analysis, I would check survival rate in different categories of passengers. For example, is men's survival rate higher then women's survival rate? Is it more likely for children to be survived rather then grown-up? In the predictive analysis, I will use training data to build regression model and apply the model to testing data. Besides, I am going to calculate predictive accuracy using the model.

2. Description of data set

The data were collected from Kaggel (www.kaggel.com), a platform for predictive modeling and analytics competition on Nov 7, 2013. The data were passengers' and crews' information such as name, gender, age, class of cabins and so on.

The raw data downloaded from Kaggel is divided into two separate dataset-training and testing dataset. There are 891 observations in training data, while there are 418 observations in testing data. In terms of variables, there are 12 variables in training data. They are Passenger ID, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked. Except for Survived variable, other variables are also in the testing data because testing data would be used to predict what sorts of people were likely to survive according to the analysis of training data.

3. Description of variables

I would not go through whole variables in the data because meanings of some variables are clear such as Name, Sex and Passenger ID. I will pick out some unclear variables to explain.

- Sibsp: Number of Siblings/ Spouses
- Parch: Number of Parents/Children
- Age: Age is in years; fractional if age less then one. If the age is estimated, it is in the form xx.5
- Pclass: Socio-economic status, 1st - Upper; 2nd - Middle; 3rd - Lower
- Embarked: Port of Embarkation (C = Cherbourg; Q = Queentown; S = Southampton)
- Survived: 0 = No; 1 = Yes
- Ticket: Ticket number

4. Processing tasks

1. Split variable Name into first name and last name, and group the same first name or last name
2. Change variable Sex from male and female to 0 and 1, which is helpful for further analysis
3. In descriptive analysis, variable Age could be grouped to check survived rates in different age groups
4. Make dummy code for variable Sibsp and Parch.
5. Extract first two letters in the variable Ticket and make it as categorical variable to prepare for further analysis
6. Doing data processing for missing values, for example, count how many number of missing value in variable Cabin and then decide whether to delete the variable
7. Round variable Fare to 3 decimal places and group Fare into different groups such as 0-30, 30-60 and so on

5. Analysis tasks

1. Plot bar chart on gender to see how many female passengers are survived according to the gender
2. Plot the distribution of passenger's age to see
3. Plot bar chart on class of socio-economic status to display how many passengers are survived
4. Use training data to build logistic regression and check which variable is statistic significant and contributes the most to the dependent variable
5. Calculate predictive accuracy by using testing data and calculate false positive and false negative