

Survival Rate in Titanic Disaster

1. Description of data set

The data were collected from Kaggle (www.kaggle.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 Kaggle 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.

In terms 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 than 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

2. Processing tasks

In the processing tasks, I mainly focus on data munging or data process to prepare for the exploratory analysis and statistical modeling. I cleaned some variables and split few variables into new variables.


2.1 Clean the Name column

Original Name column included first name, prefix and last name. I determined to split the column into separate three columns - FirstName, Prefix and LastName. The reason why I did this is that Prefix and LastName might be useful in the further analysis. For example, Steve probably was a last name whose survival rate was high.

	Name		
	Braund, Mr. Owen Harris		
	Cumings, Mrs. John Bradley (Florence Briggs Th...		
	Heikkinen, Miss. Laina		
	Futrelle, Mrs. Jacques Heath (Lily May Peel)		
	Allen, Mr. William Henry		
	LastName	FirstName	Prefix
	Owen Harris	Braund	Mr
John Bradley (Florence Briggs Thayer)	Cumings		Mrs
	Laina	Heikkinen	Miss
Jacques Heath (Lily May Peel)	Futrelle		Mrs
	William Henry	Allen	Mr

2.2 Clean the Ticket column

There are two steps in this processing part. Considering the fact that Ticket column contained ticket number and ticket mark, in the first step, I split the Ticket column into TicketNum and TicketMark column. In the second step, I used regular expression to clean the TicketMark because there were typo errors in the column. Specifically, the ticket mark A./5, A/5. and A/5 should be the ticket mark A/5. Following figures only show a part of all ticket marks and processed ticket marks. Ipython notebook has detail result.

TicketMark		Processed	TicketMark	
A./5.	2		A/4	6
A.5.	2		A/5	19
A/4	3		A/S	1
A/4.	3		A4	1
A/5	10		A5	2
A/5.	7		C	5

2.3 Clean and cut the Age column

The value of Age variable is in a range from 0.42 to 80. For the analysis, I cut Age variable into 4 different categories - Teen, Mid, Old and Unknown. The age of Teen, Mid, Old and Unknown are 0-17, 18-40, 41-100 and -1-0. There are some observations whose age was equal to 0. I think 0 represent missing values, so I used Unknown to stand for those missing values.

Following figure is the frequency table for AgeLevel, which shows how many observations in each category:

AgeLevel	
Mid	451
Old	150
Teen	113
Unknown	177

2.4 Clean the Cabin column

The Cabin column consisted of alphabets and numbers such as C85 and C123. I deleted numbers and kept alphabet. Then, I can make dummy code for cabin to be used in the logistic regression and classification tree. Also, there are some missing value, None in the Cabin column.

Following figure is the frequency table for Cabin, which shows how many observations in each category:

Cabin	
A	15
B	47
C	59
D	33
E	32
F	13
G	4
None	687
T	1

2.5 Create the Fare interval

Like Age column, I did the same to Fare column. I transformed Fare column to add a new column named FareLevel. 0-10 was classified into Low category. 10-50 was classified into LowMid category. 50-100 was


classified into Mid category and above 100 was classified into High category. FareLevel also can be made dummy code for further model and used in the exploratory analysis.

Following figure is the frequency table for FareLevel, which shows how many observations in each category:

FareLevel	
High	53
Low	321
LowMid	395
Mid	107

2.6 Clean the number of Sibling

There were 7 different values in the Sibling column (0, 1, 2, 3, 4, 5, 8). However, there were total 817 individuals who had 0 or 1 sibling, while only 74 people who had more than 1 sibling. Thus, I grouped the categories except 0 and 1 into category 2+, which referred to those individuals who had more than 1 sibling. The reason why I did this is that I think comparing category 0 and 1, number of each categories except 0 and 1 is too little, which cannot provides any useful effects on the analysis.

SibSp		Processed	SibSp	
0	608		0	608
1	209		1	209
2	28		2+	74
3	16		--	--
4	18			
5	5			
8	7			

3. Description of Analysis

3.1 Exploratory data analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods¹. Exploratory data analysis can be used to (1) check for missing data and other confounders; (2) summarized the data.

1) Bar chart for survival rate on different categories

I plotted histogram for survival rate on AgeLevel, Sex, Pclass and FareLevel variable. Another reason why I want to classify different continuous ages and fares into different categories is that I can plot bar chart on different categories of ages and fares.

From following figure 1, we can find that in AgeLevel, the number of middle age people is the largest among other three categories and most middle age people died in the titanic disaster. Also, we can find that the number of survived teenagers is more than the number of not survived teenagers. And this situation does not appear in middle age and old people. What we have seen in the bar chart in AgeLevel makes sense because in titanic disaster, survivors mostly are children and the elderly. Even though in the bar chart the largest number of survivors is in middle age people, the dataset does not include all passenger information in titanic disaster. In the histogram for death rate, I would further support my argument.

In terms of gender, we can easily found that most women were survived in the titanic disaster, comparing with men. Besides, the total number of men is about twice as many as total number of women.

Pclass refers to the social-economic status of passengers. The total number of class 3 is the largest among all class levels. It is possible that class 3 not only includes lower economic level passengers, but also has crews whose economic levels tend to be low. Only one forth people in class 3 were survived in titanic disaster. However, in class 1, the number of survival passengers is slightly more than the number of not survival passengers.

In the FareLevel, I think fare is related to the social economic status. For example, high social economic level people can afford high fare, while low social economic level people only are able to buy low or middle price. Thus, it is not surprised to find that the number of not survival people in Low and LowMid fare is the largest. In middle fare and high fare, the number of survival people is more than the number of not survival. By contrast, the number of dead people in low and low middle category is much more than the number of survived people.

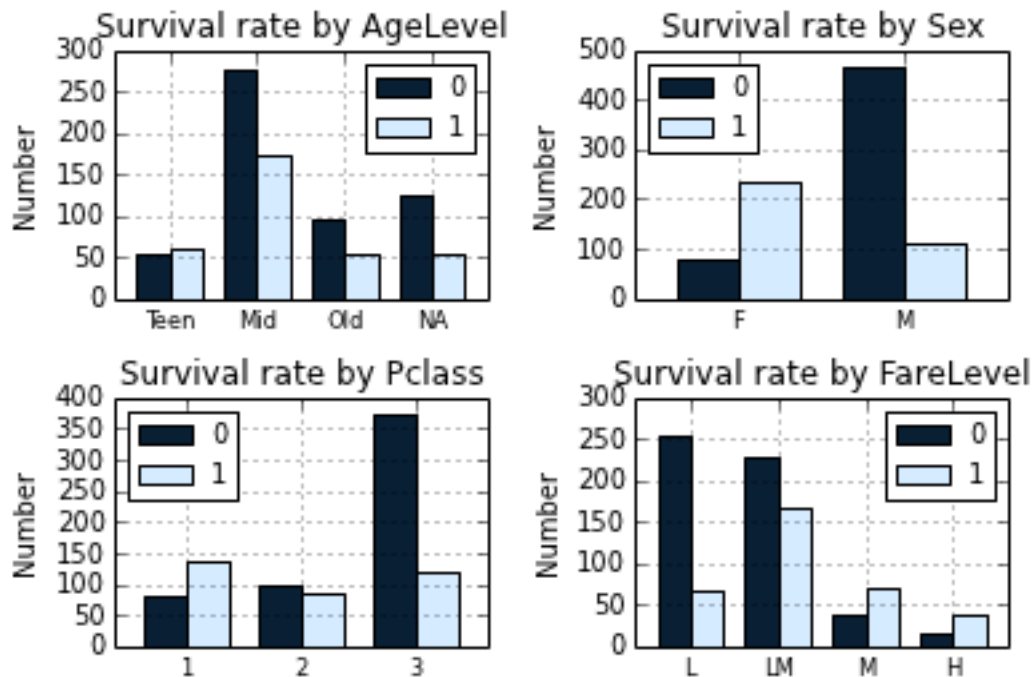


Figure 1. Bar chart for survival rate

2) Bar chart for death rate on different categories

After looking at the general situation about survival rate on different variables, I used pivot table function in python to calculate death rate on different categories. I think relative number is more useful than absolute number in comparing survival rate in different categories.

Following are bar charts for death rate on AgeLevel, Pclass, Sex and FareLevel variables:

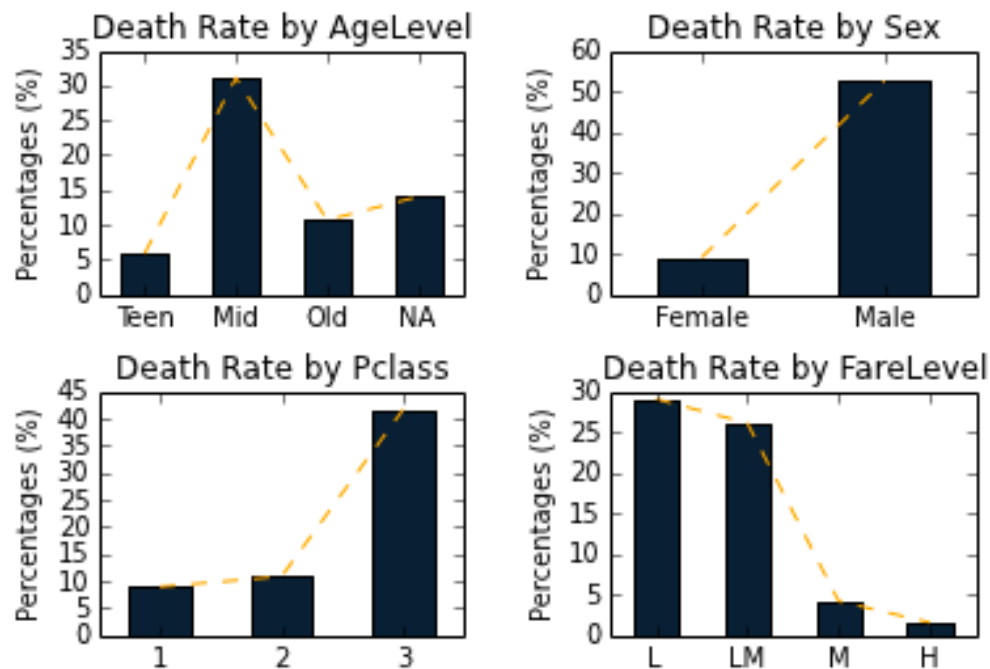


Figure 2. Bar charts for death rate

We can find that death rates in teenagers and old people are the lowest in AgeLevel. In Sex, men's death rate highly outnumbers its counterpart. In Pclass, death rate of class 3 is as high as 40%. Death rate in FareLevel decreases along with the increase of fare.

3) Scatterplot for age and fare

After understanding the survival rate on categorical AgeLevel and FareLevel, I used scatterplot to see the distribution of continuous variable age and fare. From figure 3, we can find that when the fare is between 100 and 500, most people were survived. Also, a large number of people whose ages were between 0 and 10 were survived, which proves that teenagers tend to survive in Titanic disaster.

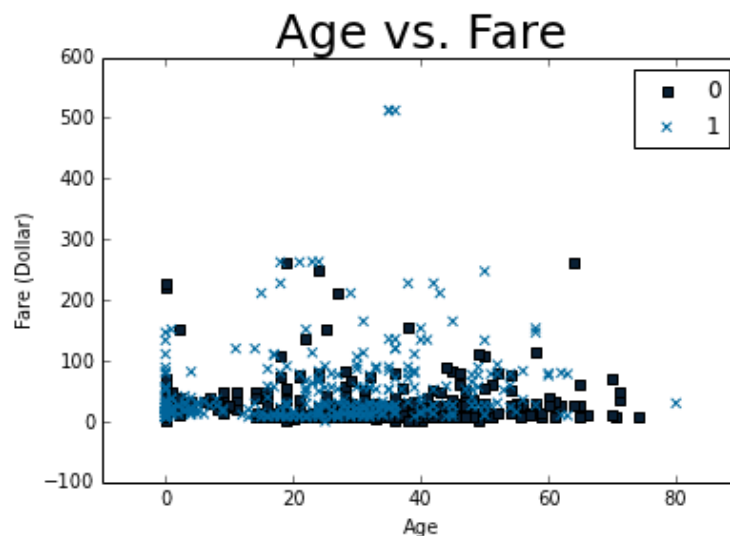


Figure 3. Scatterplot for age and fare

4) Distribution for age and departed ports

I used histograms to see the distribution of age and departed ports. In the scatterplot of age, we hardly see the exact number of people in different age. In figure 4, we could know most people's ages are range from 0 to 10. However, I think there are a large part is those people with unknown age.

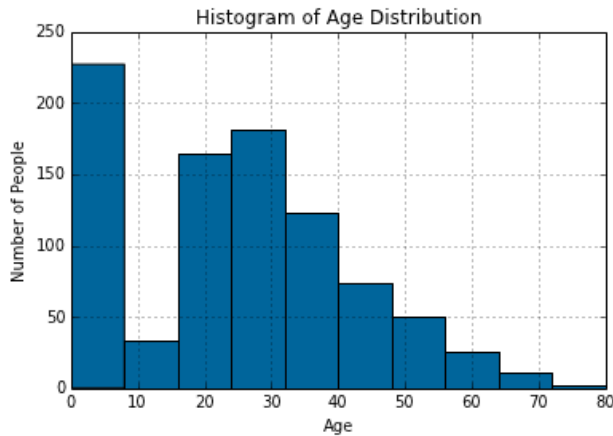


Figure 4. Histogram of age

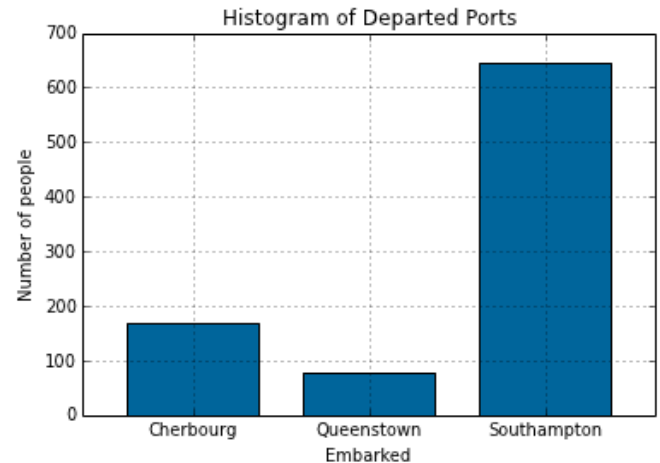


Figure 5. Histogram of departed ports

In figure 5, we are able to know that most people got into titanic in the Southampton port, which is the starting point.

3.2. Statistical Modeling

After conducting the exploratory data analysis, I begin to further the analysis in statistical modeling. Considering that the outcome variable, Survived, is binary and categorical, I ran the logistic regression, classification tree and random forest.

1) Dummy code for categorical variable

Before running the model, we have to conduct dummy code for categorical variables such as sex and pclass. In statistics, particularly in regression analysis, a dummy variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome². Therefore, for categorical variable such as Sex, Pclass, SibSp, Cabin, Prefix and so on, I conducted the dummy code and then put them in regression model.

2) Logistic Regression

Considering the fact that there are too many first names and last names in the passengers, I did not think that first name and last name can provide certain patterns related to the survival rate. Also, I would put different variables into the logistic regression to build different models. However, due to the time, I want the first logistic regression is not too complicated. That is also the reason why I did not add TicketMark into this logistic regression. My final logistic regression model was:

$$y = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Fare} + \beta_3 \text{Parch} + f(\text{Pclass}) + g(\text{Embarked}) + h(\text{Prefix}) + j(\text{SibSp}) + k(\text{Cabin}) + e$$

where β_0 is an intercept term and β_1 , β_2 and β_3 represent the change in age, fare and number of parents and children at the same Pclass, Embarked, Prefix, SibSp and Cabin. The terms $f(\text{Pclass})$, $g(\text{Embarked})$, $h(\text{Prefix})$, $j(\text{SibSp})$ and $k(\text{Cabin})$ refer to factor model with different levels for social-

economic status, departed ports, prefix, number of siblings and different cabins. The error term e means all sources of unmeasured and uncovered random variations in the titanic disaster.

Figure 6 is the output of logistic regression from python. Among variables, Parch, Sex_male, Pclass_3, Embarked_C, P_Master, P_Miss, P_Mrs, Sib_2+, Cabin_B, Cabin_D and Cabin_E are statistical significant because their p-values are all below 0.05. Take Parch for example. The coefficient of Parch is -0.3708. The interpretation for Parch is that when other variables stay the same, the odds of survive are multiplied by 0.69 for each increment of Parch. I used the odds to interpret because the link function of logistic regression is logit. Sometimes people prefer to look at a different measure of belonging to a certain class, known as odds³. Also, it makes sense for Pclass_3. Comparing with Pclass_1, the odds for survive for Pclass_3 are multiplied by 0.368, which suggests that the probability of survive for class 3 is less than class 1.

Logit Regression Results						
Dep. Variable:	Survived	No. Observations:	891			
Model:	Logit	Df Residuals:	866			
Method:	MLE	Df Model:	24			
Date:	Mon, 09 Dec 2013	Pseudo R-squ.:	0.3889			
Time:	21:24:52	Log-Likelihood:	-362.56			
converged:	False	LL-Null:	-593.33			
		LLR p-value:	1.564e-82			
	coef	std err	z	P> z	[95.0% Conf. Int.]	
Age	-0.0102	0.006	-1.598	0.110	-0.023	0.002
Fare	0.0041	0.003	1.456	0.145	-0.001	0.010
Parch	-0.3708	0.132	-2.808	0.005	-0.630	-0.112
Sex_male	-1.7809	0.880	-2.024	0.043	-3.506	-0.056
Pclass_2	0.0295	0.444	0.066	0.947	-0.841	0.899
Pclass_3	-1.0014	0.439	-2.280	0.023	-1.862	-0.141
Embarked_C	0.5377	0.251	2.138	0.033	0.045	1.031
Embarked_Q	0.1831	0.346	0.530	0.596	-0.494	0.860
P_Dr	1.1274	1.174	0.960	0.337	-1.174	3.428
P_Master	3.8136	0.946	4.030	0.000	1.959	5.668
P_Miss	1.6927	0.488	3.469	0.001	0.736	2.649
P_Mr	0.3926	0.816	0.481	0.630	-1.206	1.991
P_Mrs	2.1774	0.550	3.958	0.000	1.099	3.256
P_Ms	36.4045	6.71e+07	5.42e-07	1.000	-1.32e+08	1.32e+08
P_Sir	37.0345	6.71e+07	5.52e-07	1.000	-1.32e+08	1.32e+08
Sib_1	-0.1800	0.245	-0.735	0.463	-0.660	0.300
Sib_2+	-1.4178	0.389	-3.640	0.000	-2.181	-0.654
Cabin_A	0.8549	0.686	1.246	0.213	-0.490	2.200
Cabin_B	1.1419	0.564	2.024	0.043	0.036	2.248
Cabin_C	0.6485	0.516	1.256	0.209	-0.363	1.660
Cabin_D	1.5883	0.596	2.663	0.008	0.420	2.757
Cabin_E	1.9087	0.608	3.141	0.002	0.718	3.100
Cabin_F	0.9919	0.825	1.202	0.229	-0.625	2.609
Cabin_G	-0.2862	1.024	-0.279	0.780	-2.293	1.721
Cabin_T	-34.5035	7.27e+07	-4.75e-07	1.000	-1.42e+08	1.42e+08

Figure 6. Logistic Regression Report

Furthermore, I used the chi-square test to test how whether the logistic regression model is better than the null model⁴. The chi-square test was run on the R. Following is the result.

```
> with(model,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail=FALSE))
[1] 1.625689e-82
```

The p-value is much less than 0.05, which rejects the null hypothesis that the model is equal to the null model.

Moreover, I made confusion matrix for training data and ROC curve:

Confusion Matrix		Predicted	
		0	1
Actual	0	481	68
	1	82	260

- **Accuracy: 83.2%**
- **Sensitivity: 76.0%**
- **Specificity: 87.6%**
- **Positive Predictive Value: 79.3%**
- **Negative Predictive Value: 85.4%**

From the ROC, we can see the logistic regression preforms well. The circled line in the Figure 7 represents the logistic regression could predict accurately for the survival rate.

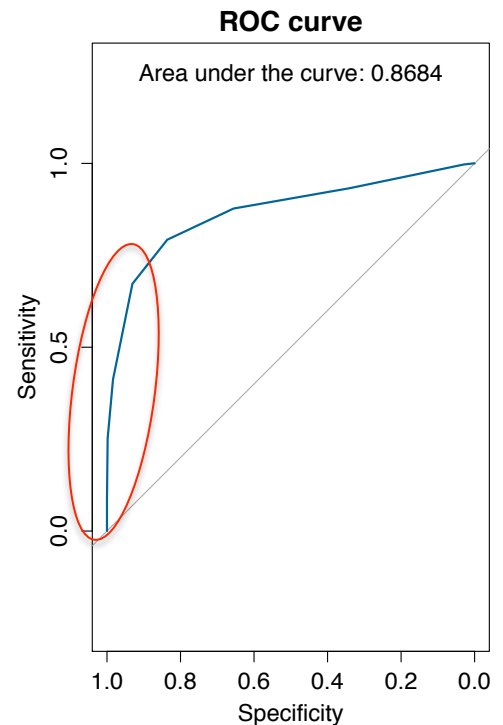


Figure 7. ROC curve

3) Classification Tree

If one had to choose a classification technique that performs well across a wide range of situations without requiring much effort from the analyst while being readily understandable by the consumer of the analysis, a strong contender would be the tree methodology developed by Breiman et al[5].

Classification tree is the best method to classified different categories. I used classification tree to find those important variables which probably influence the survival rate. First, I ran the classification tree by using python. Figure 8 is the output from python. I limited the max depth in DecisionTreeClassifier function to avoid overfitting. Also, I made the confusion matrix to calculate the sensitivity and specificity.

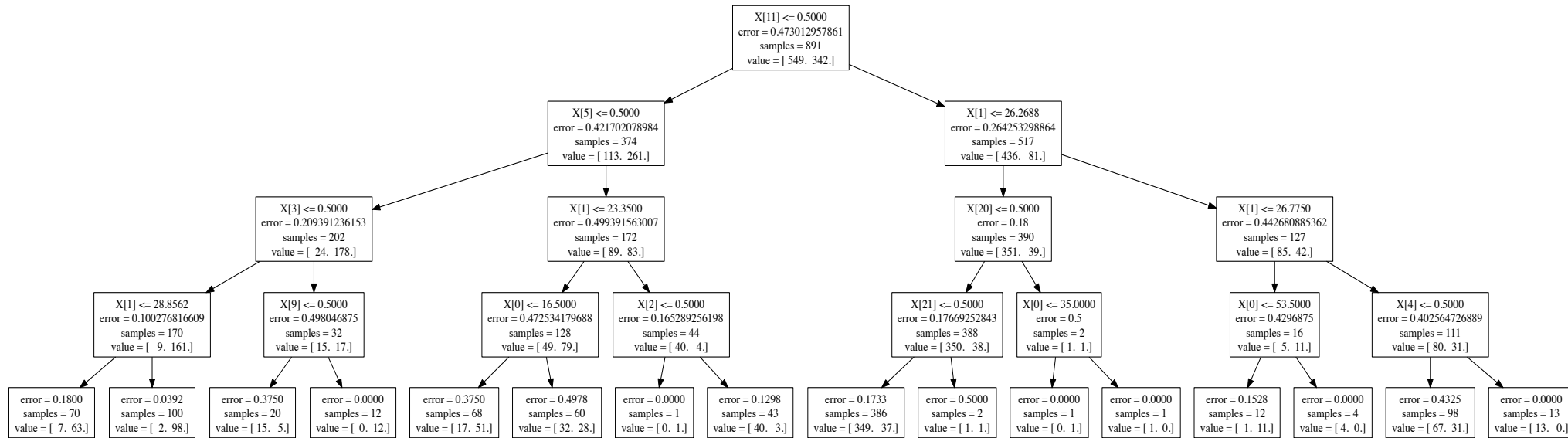


Figure 8. Classification tree by Python

Due to the space limit here, you can check the Ipython Notebook to find the clearer figure.

Confusion matrix is listed following. We can see that sensitivity is not very high, which suggests classification tree performed not very well in classifying not survival category. Later, I will compare the result produced by Python with the result produced by R.

Confusion Matrix		Predicted	
		0	1
Actual	0	522	27
	1	105	237

- **Accuracy: 85.2%**
- **Sensitivity: 69.3%**
- **Specificity: 95.1%**
- **Positive Predictive Value: 89.8%**
- **Negative Predictive Value: 83.3%**

Due to the time limit, I did not have enough time to find a better way to avoid overfitting such as pruning the tree by using minimal error rate. Therefore, I used R to rerun the classification tree. R has better function to build classification tree by using different parameters to prune the tree in order to avoid overfitting. Figure 9 is the classification tree on the training data. Also, like the logistic regression, I made the confusion matrix of the classification tree.

From the confusion matrix I made from the result produced by R, we can easily find the consequence is very similar. It strengthens my classification tree analysis. Like above confusion matrix, sensitivity here is even lower after pruning the tree.

Confusion Matrix		Predicted	
		0	1
Actual	0	523	26
	1	118	224

- **Accuracy: 83.84%**
- **Sensitivity: 65.5%**
- **Specificity: 95.3%**
- **Positive Predictive Value: 89.6%**
- **Negative Predictive Value: 81.6%**

The numbers under the node are the numbers of observations which were classified into 0 and 1. For example, the first node 0. Left number 436 represents that 436 observations were classified into 0, while right number 81 means that 81 observations were classified into 1. Thus, we can find that except for the bottom left node (32 28), the rates of other nodes for successfully classifying are above 70%.

Classification Tree for Titanic Survive

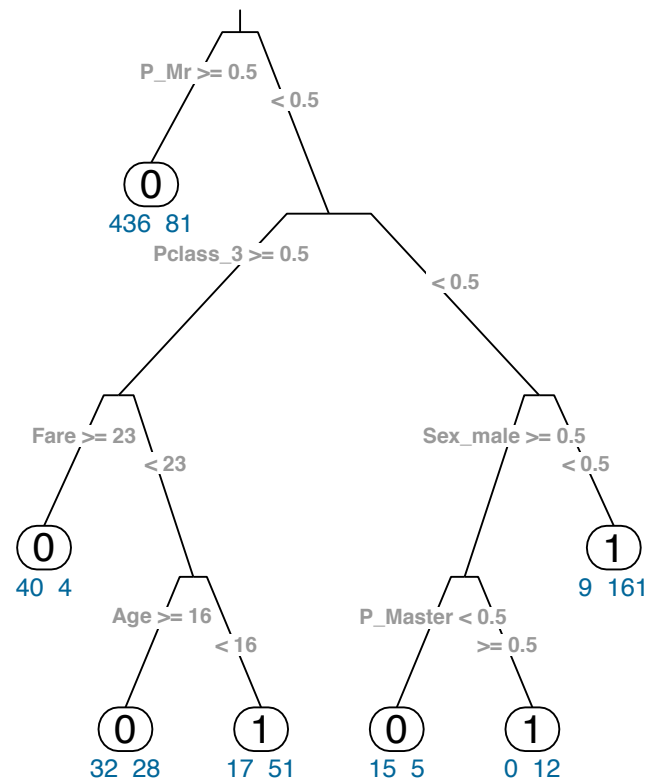


Figure 9. Classification tree by R

Some rules can be achieved by the classification tree:

- passengers whose prefix were Mr. tend not to be survived
- passengers whose prefix were not Mr., Pclass was 3 and fare for ticket is more than 23 tend not to be survived
- passengers whose prefix were not Mr., Pclass was 3 and fare for ticket was less than 23 but age was more than 16 tend not to be survived
- passengers whose prefix were not Mr., Pclass was 3 and fare for ticket was less than 23 but age was not more than 16 tend to be survived
- Female passengers whose prefix were not Mr. and Pclass was not 3 and tend to be survived
- Male passengers whose prefix were Master. tend not to be survived
- Male passengers whose prefix were Master. tend to be survived

4) Random Forest

Random forests is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them⁶. Random forest is helpful to reduce the variance and increase accuracy. Like classification tree, I used Python and R to run the random forest respectively. Then I would compare two results by using different tools. First, I used sklearn module in Python to build random forest and made the confusion matrix.

- **Accuracy: 83.8%**
- **Sensitivity: 73.7%**
- **Specificity: 90.2%**
- **Positive Predictive Value: 82.4%**
- **Negative Predictive Value: 84.6%**

Confusion Matrix		Predicted	
		0	1
Actual	0	495	54
	1	90	252

In the next step, I also used R to conduct random forest on the training data. The confusion matrix is following:

- **Accuracy: 82.4%**
- **Sensitivity: 70.5%**
- **Specificity: 89.8%**
- **Positive Predictive Value: 81.1%**
- **Negative Predictive Value: 83.0%**

Confusion Matrix		Predicted	
		0	1
Actual	0	493	56
	1	101	241

It is easy to see these two results are very similar. In the random forest, I set the number of total trees is equal to 500 in both Python and R to better average results produced by different classification tree. Comparing with the classification tree, we can find that the sensitivity increases to 70.5%, although specificity reduces to 89.8%. I think random forest are more reasonable and accurate than classification tree because random forest runs a bunch of classification tree and average them. The result of random forest tends to be more accurate and objective.

At last, I calculated mean decrease Gini from the random forest by using R. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest⁷.

	MeanDecreaseGini
Age	28.60
Fare	44.36
Parch	9.41
Sex_male	46.70
Pclass_2	5.67
Pclass_3	22.23
Embarked_C	5.66
Embarked_Q	3.01
P_Dr	0.85
P_Master	5.09
P_Miss	15.22
P_Mr	50.65
P_Mrs	14.07
P_Ms	0.09
P_Sir	0.20

	MeanDecreaseGini
Sib_1	6.33
Sib_2.	7.01
Cabin_A	1.06
Cabin_B	3.15
Cabin_C	2.79
Cabin_D	3.17
Cabin_E	5.05
Cabin_F	0.75
Cabin_G	0.53
Cabin_T	0.09

As we can see from above tables, a higher mean decrease Gini means that a particular predictor variable plays a greater role in partitioning the data into defined classes. Therefore, Fare, P_mr, Sex_male and Age are valuable variables in partitioning the data.

4. Conclusion

To conclude, logistic regression and random forest tree were better model to predict and classify the survival rate. Also, classification tree offers some rules for us to predict the survival rate. As I guessed before the analysis, Children, women and the elderly had more chance to be survived in titanic disaster. High social economic status were also likely to be survived. In random forest, mean decrease Gini helped prove that some variables which were statistically significant in logistic regression were special.

For exploratory analysis, I think it is perfect. I considered whole variables and analyzed their distributions to pick out the most useful and representative I supposed. Moreover, the quality of data is good and guaranteed because Kaggle has high reputation in organizing data analysis competition.

However, the analysis is not perfect and completed. I left few other variables such as first name, last name and ticket marks. I would add them in next logistic regression and classification tree to see interesting result. This is a preliminary analysis for the titanic disaster. In the future, I might consider to use ensemble or other methods to combining different classifiers to analyze the data in order to get more comprehensive and accurate result.

5. Reference

1. Wikipedia "Exploratory data analysis" Page. URL:
http://en.wikipedia.org/wiki/Exploratory_data_analysis. Accessed 12/4/2013
2. Wikipedia "Dummy variable" Page. URL: [http://en.wikipedia.org/wiki/Dummy_variable_\(statistics\)](http://en.wikipedia.org/wiki/Dummy_variable_(statistics))
3. Galit Shmueli et al. *Data Mining for Business Intelligence*. Vol. 139. Wiley, 2007
4. R Data Analysis Examples. URL: <http://www.ats.ucla.edu/stat/r/dae/logit.htm>. Accessed 12/10/2013
5. Galit Shmueli et al. *Data Mining for Business Intelligence*. Vol. 111. Wiley, 2007
6. Trevor Hastie et al. *The Elements of Statistical Learning*. Vol. 587. Springer, 2013
7. Metagenomics Statistics. URL: <http://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html>.
Accessed 12/10/2013