STAT 425 Project Report

Titanic Disaster Data Analysis

Xiaodan Zhang

STAT 425-Applied Regression and Design

Feng Liang

University of Illinois at Urbana-Champaign

Dec.12, 2012

1. Description of the data

RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean on 15 April 1912 after colliding with an iceberg during her maiden voyage from Southampton, UK to New York City, US. The Titanic data we analyze in this project are demographic statistics for 1309 Titanic passengers and crew. Our goal is to determine how the variable survived depends on the other variables measured in the study and what sorts of people were more likely to be survived.

Number of Observations: 1309
Number of Variables: 13

Variable Names:
· Categorical Variables
survived: Survival (0 = No; 1 = Yes)
pclass: Passenger Class (1 = 1st, Upper ; 2 = 2nd, Middle; 3 = 3rd, Lower)
sex: Sex (male; female)
embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
· Numerical Variables
age: Passenger Age (In years)
fare: Passenger Fare (In pounds)
sibsp: Number of Siblings/Spouses Aboard
[Note: *Sibling*: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)]
parch: Number of Parents/Children Aboard
[Note: *Parent*: Mother or Father of Passenger Aboard Titanic
Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic]
· Variables Not Used In The Analysis:
Life_boat: Number or identifier of the boat the survivor was on
name: Passenger Name
ticket: Ticket Number
cabin: Cabin embarked
train: Training Set or Test Set (0 = Test Set; 1 = Training Set)

2. Data Analysis

I. Basic Summary of the data

· The 3-way table for "survived", "pclass", "sex" is shown in Figure 1 below.

```
                sex female male
survived pclass
0        1              5  118
         2             12  146
         3            110  417
1        1            139   62
         2             94   24
         3            106   76
```

Figure 1.

· The 5-number-summaries for "age" is "0.17 21.00 28.00 38.00 74.00".
The 5-number-summaries for "fare" is "0.0000 7.8958 14.4542 31.2750 512.3292".
And the histograms for "age" and "fare" are shown in Figure 2 below.

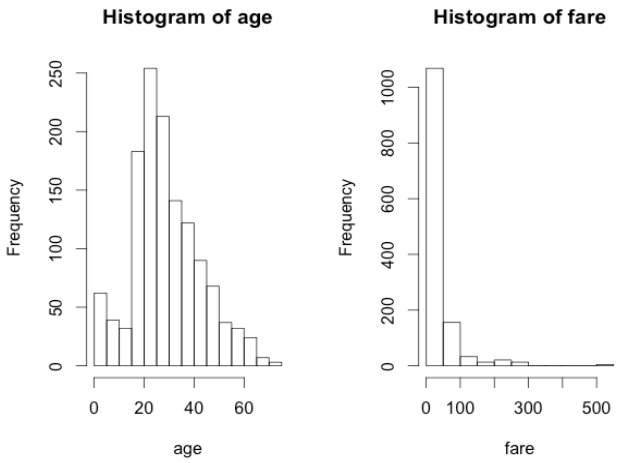

Figure 2.

· The frequency tables for "sibsp" and "embarked" are shown in Figure 3 below.

```
> ts
sibsp
  0   1   2   3   4   5   8
887 324  45  18  20   6   9
> te
embarked
  B   C   Q   S
 10 267 123 909
```

Figure 3.

· According to the R output, 17 people traveled with zero "fare".

· There were 4 people who paid the most expensive ticket that cost £512.3292:

```
> maxfare
     [,1]
[1,] "Ward, Miss. Anna"
[2,] "Cardeza, Mr. Thomas Drake Martinez"
[3,] "Lesurer, Mr. Gustave J"
[4,] "Cardeza, Mrs. James Warburton Martinez (Charlotte Wardle Drake)"
```

· There were 33 people (in 4 different families) whose family had more than 6 members on board and their names are listed as below:

```
> fams
      [,1]
 [1,] "Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)"
 [2,] "Asplund, Master. Clarence Gustaf Hugo"
 [3,] "Asplund, Miss. Lillian Gertrud"
 [4,] "Asplund, Master. Edvin Rojj Felix"
 [5,] "Asplund, Master. Filip Oscar"
 [6,] "Asplund, Mr. Carl Oscar Vilhelm Gustafsson"
 [7,] "Asplund, Master. Carl Edgar"
 [8,] "Andersson, Mr. Anders Johan"
 [9,] "Andersson, Miss. Ellis Anna Maria"
[10,] "Andersson, Miss. Ingeborg Constanzia"
[11,] "Andersson, Miss. Sigrid Elisabeth"
[12,] "Andersson, Mrs. Anders Johan (Alfrida Konstantia Brogren)"
[13,] "Andersson, Miss. Ebba Iris Alfrida"
[14,] "Andersson, Master. Sigvard Harald Elias"
[15,] "Goodwin, Master. William Frederick"
[16,] "Goodwin, Miss. Lillian Amy"
[17,] "Goodwin, Master. Sidney Leonard"
[18,] "Goodwin, Master. Harold Victor"
[19,] "Goodwin, Mrs. Frederick (Augusta Tyler)"
[20,] "Goodwin, Mr. Charles Edward"
[21,] "Goodwin, Mr. Charles Frederick"
[22,] "Goodwin, Miss. Jessie Allis"
[23,] "Sage, Master. Thomas Henry"
[24,] "Sage, Miss. Constance Gladys"
[25,] "Sage, Mr. Frederick"
[26,] "Sage, Mr. George John Jr"
[27,] "Sage, Miss. Stella Anne"
[28,] "Sage, Mr. Douglas Bullen"
[29,] "Sage, Miss. Dorothy Edith \"Dolly\""
[30,] "Sage, Miss. Ada"
[31,] "Sage, Mr. John George"
[32,] "Master Anthony William Sage "
[33,] "Sage, Mrs. John (Annie Bullen)"
```

· There were 86 people (in 12 different groups) who were in some big travel groups, i.e. more than 5 persons shared the same ticket, and their names are listed as below:

```
[1] "Allison, Miss. Helen Loraine"
[2] "Allison, Master. Hudson Trevor"
[3] "Allison, Mrs. Hudson J C (Bessie Waldo Daniels)"
[4] "Cleaver, Miss. Alice"
[5] "Daniels, Miss. Sarah"
[6] "Allison, Mr. Hudson Joshua Creighton"
[7] "Bing, Mr. Lee"
```

[8] "Ling, Mr. Lee"
[9] "Lang, Mr. Fang"
[10] "Foo, Mr. Choong"
[11] "Lam, Mr. Ali"
[12] "Lam, Mr. Len"
[13] "Chip, Mr. Chang"
[14] "Hee, Mr. Ling"
[15] "Fortune, Mr. Charles Alexander"
[16] "Fortune, Miss. Mabel Helen"
[17] "Fortune, Miss. Alice Elizabeth"
[18] "Fortune, Mr. Mark"
[19] "Fortune, Miss. Ethel Flora"
[20] "Fortune, Mrs. Mark (Mary McDougald)"
[21] "Panula, Master. Juha Niilo"
[22] "Panula, Master. Eino Viljami"
[23] "Panula, Mr. Ernesti Arvid"
[24] "Panula, Mrs. Juha (Maria Emilia Ojala)"
[25] "Panula, Mr. Jaako Arnold"
[26] "Panula, Master. Urho Abraham"
[27] "Riihivouri, Miss. Susanna Juhantytar Sanni"
[28] "Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)"
[29] "Asplund, Master. Clarence Gustaf Hugo"
[30] "Asplund, Miss. Lillian Gertrud"
[31] "Asplund, Master. Edvin Rojj Felix"
[32] "Asplund, Master. Filip Oscar"
[33] "Asplund, Mr. Carl Oscar Vilhelm Gustafsson"
[34] "Asplund, Master. Carl Edgar"
[35] "Andersson, Mr. Anders Johan"
[36] "Andersson, Miss. Ellis Anna Maria"
[37] "Andersson, Miss. Ingeborg Constanzia"
[38] "Andersson, Miss. Sigrid Elisabeth"
[39] "Andersson, Mrs. Anders Johan (Alfrida Konstantia Brogren)"
[40] "Andersson, Miss. Ebba Iris Alfrida"
[41] "Andersson, Master. Sigvard Harald Elias"
[42] "Skoog, Master. Harald"
[43] "Skoog, Mrs. William (Anna Bernhardina Karlsson)"
[44] "Skoog, Mr. Wilhelm"
[45] "Skoog, Miss. Mabel"
[46] "Skoog, Miss. Margit Elizabeth"
[47] "Skoog, Master. Karl Thorsten"
[48] "Rice, Master. Eugene"
[49] "Rice, Master. Arthur"
[50] "Rice, Master. Eric"
[51] "Rice, Master. George Hugh"

[52] "Rice, Mrs. William (Margaret Norton)"
[53] "Rice, Master. Albert"
[54] "Goodwin, Master. William Frederick"
[55] "Goodwin, Miss. Lillian Amy"
[56] "Goodwin, Master. Sidney Leonard"
[57] "Goodwin, Master. Harold Victor"
[58] "Goodwin, Mrs. Frederick (Augusta Tyler)"
[59] "Goodwin, Mr. Charles Edward"
[60] "Goodwin, Mr. Charles Frederick"
[61] "Goodwin, Miss. Jessie Allis"
[62] "Sage, Master. Thomas Henry"
[63] "Sage, Miss. Constance Gladys"
[64] "Sage, Mr. Frederick"
[65] "Sage, Mr. George John Jr"
[66] "Sage, Miss. Stella Anne"
[67] "Sage, Mr. Douglas Bullen"
[68] "Sage, Miss. Dorothy Edith \"Dolly\""
[69] "Sage, Miss. Ada"
[70] "Sage, Mr. John George"
[71] "Master Anthony William Sage "
[72] "Sage, Mrs. John (Annie Bullen)"
[73] "Ryerson, Miss. Emily Borie"
[74] "Ryerson, Miss. Susan Parker \"Suzette\""
[75] "Ryerson, Mrs. Arthur Larned (Emily Maria Borie)"
[76] "Chaudanson, Miss. Victorine"
[77] "Ryerson, Master. John Borie"
[78] "Ryerson, Mr. Arthur Larned"
[79] "Bowen, Miss. Grace Scott"
[80] "Hood, Mr. Ambrose Jr"
[81] "Hickman, Mr. Stanley George"
[82] "Davies, Mr. Charles Henry"
[83] "Hickman, Mr. Leonard Mark"
[84] "Hickman, Mr. Lewis"
[85] "Deacon, Mr. Percy William"
[86] "Dibden, Mr. William"

· The two missing values for variable “age” are associated with index 784 and 785. The median age for all male passengers in the 3rd class is 24.

· The one missing value for variable “fare” is associated with index 118. The median fare for all passengers in the 3rd class is 8.05.

II. Logistic Regression

Number of observations in training set: 891
Number of observations in test set: 418

(a) Fit a logistic model **survived ~ sex + pclass** on the training data.

The AIC value of this model is 833.8884, and the BIC value of this model is 854.0577 according to the R output.

```
> summary(g1)$coeff
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept)  2.2971232  0.2189839  10.489917 9.611309e-26
sexmale     -2.6418754  0.1840954 -14.350577 1.056551e-46
pclass2     -0.8379523  0.2447436  -3.423797 6.175280e-04
pclass3     -1.9054951  0.2141410  -8.898319 5.669887e-19
```

Figure 4. Coefficients of the 1st logistic regression model.

To interpret the coefficients in Figure 4, we recall that the response variable is log-odds. The intercept tells that the reference group is a female in the Upper Passenger Class. She had log-odds of surviving of 2.297. The odds of survival are $e^{2.297}$=9.944, which can be translated into a probability of survival of $\frac{9.944}{1+9.944} = 0.909$. The coefficient of "sexmale" is the difference in log-odds ratio of survival between males and females, -2.642. So, the log-odds of survival for males are 2.297-2.642=-0.345 and the odds of survival are $e^{-0.345}$=0.708 which can be translated into a probability of survival of $\frac{0.708}{1+0.708} = 0.415$. The coefficient of "pclass2" is the difference in log-odds ratio of survival between pclass1 and pclass2, -0.838. So, the log-odds of survival for passengers in the Middle Passenger Class is 2.297-0.838=1.459 and the odds of survival are $e^{1.459}$=4.302, which can be translated into a probability of survival of $\frac{4.302}{1+4.302} = 0.811$. And similarly, the coefficient of "pclass3" is the difference in log-odds ratio of survival between pclass1 and pclass3, -1.905. So, the log-odds of survival for passengers in the Lower Passenger Class is 2.297-1.905=0.392 and the odds of survival are $e^{0.392}$=1.480, which can be translated in to a probability of survival of $\frac{1.480}{1+1.480} = 0.597$. According to the above calculations, we find that female passengers in the Upper Class were more likely to survive with the largest probability of survival.

The 2-by-2 table containing the prediction on the test data is shown in Figure 5.

```
g1.pred   0   1
      0 213  53
      1  46 106
```

Figure 5.

The prediction accuracy is (213+106)/418≈76.32%.

(b) Fit a logistic model **survived ~ sex*pclass** on the training data.

```
> summary(g2)$coeff
                Estimate Std. Error   z value     Pr(>|z|)
(Intercept)    3.4122472  0.5867893  5.815115 6.059214e-09
sexmale       -3.9493901  0.6160608 -6.410715 1.448386e-10
pclass2       -0.9555114  0.7247579 -1.318387 1.873741e-01
pclass3       -3.4122472  0.6099995 -5.593852 2.220860e-08
sexmale:pclass2 -0.1849918  0.7939117 -0.233013 8.157513e-01
sexmale:pclass3  2.0957553  0.6572051  3.188891 1.428199e-03
```

Figure 6. Coefficients of the 2nd logistic regression model.
The AIC value of this model is 810.0969 and the BIC value of this model is 838.851 according to the R output.
Comparing model II (a) versus model II (b), we found that the model II (a) is nested in model II (b). And the smaller model (II a) has a higher AIC value, 833.8884, which is larger than that of the bigger model (II b) here, 810.0969.

To interpret the coefficients in Figure 6 above, we still recall that the response variable is log-odds. The intercept tells that the reference group is a female in the Upper Passenger Class. She had log-odds of surviving of 3.412. The odds of survival are $e^{3.412}$=30.326. Because there are interactions here, the effects of gender and passenger class alone only represent the effects for the reference group. So, the coefficients of "pclass" give the log-odds difference in survival between 1st (Upper) and 2nd (Middle) class (-0.956), and 1st and 3rd class (-3.412) male passengers. To get the effects for male passengers, we need to add on the male interaction terms. So, the log-odds difference in survival between male Upper and Middle class passengers is (-0.956-0.185=-1.141) and between male Upper and Lower class passengers is (-3.412+2.096=-1.316). In addition, the difference of log-odds between male and female within each class is -3.949 in the Upper Passenger Class; (-3.949-0.185=-4.134) in the Middle Passenger Class; (-3.949+2.096=-1.853) in the Lower Passenger Class. So, we find that the difference in survival between males and females drops off considerably in the Lower Passenger Class. Moreover, for males in the Middle class, they have odds of survival $e^{-1.141-3.949+3.412}$=0.187. For males in the Lower Class, they have odds of survival $e^{-1.316-3.949+3.412}$=0.157. Because odds are the ratios of survived/death and when it is greater than one, the person was more likely to survive through the disaster, we find that female passengers in the Upper class (pclass=1) were more likely to survive.

The 2-by-2 table containing the prediction on the test data is shown in Figure 7.

g2.pred	0	1
0	213	53
1	46	106

Figure 7.

The prediction accuracy is the same as that of part (a), 76.32%.
In addition, since both models give the same prediction accuracy, we prefer the one with a smaller AIC. That is, we prefer the interaction model in II (b).

(c) Add age to model II (b). Consider all interactions.
1. Original Model: **survived ~ sex*pclass*age** without using step() or drop1()
The AIC value of this model is 780.4776 and the BIC value of this model is 837.9857 according to the R output.
The 2-by-2 table containing the prediction on the test data is shown in Figure 8.

g3.pred	0	1
0	226	63
1	33	96

Figure 8.

The prediction accuracy is (226+96)/418≈77.03%.

2. Original Model: **survived ~ sex*pclass*age** using step() or drop1() functions

The sub-model I would select is **survived ~ sex + pclass + age + sex:pclass + sex:age + pcalss:age** after doing the stepwise selection since it gives the smallest AIC value.

The AIC value of this model is 779.3916 and the BIC value of this model is 827.315 according to the R output.

The 2-by-2 table containing the prediction on the test data is shown in Figure 9.

g4.pred	0	1
0	227	64
1	32	95

Figure 9.

The prediction accuracy is (227+95)/418≈77.03%.

3. Original Model: **survived ~ sex*pclass*log(age)**

The AIC value of this model is 762.714 and the BIC value of this model is 820.2221 according to the R output. And in fact, we did not drop any term after using step() or drop1() functions because if we drop the term sex:pclass:log(age), the AIC will increase from 762.71 to 767.08 and the p-value 0.06853 is significant at 0.10 significance level.

The 2-by-2 table containing the prediction on the test data is shown in Figure 10.

g3log.pred	0	1
0	238	72
1	21	87

Figure 10.

The prediction accuracy is (238+87)/418≈77.75%, which is the same as the result in part 1.

4. Original Model: **survived ~ sex*pclass*age** using step() or drop1() functions

The sub-model I would select is **survived ~ sex + pclass + age + sex:pclass + sex:age + pclass:age** after doing the stepwise selection since it gives the smallest AIC value.

And then, change age to log(age). So, the Model becomes **survived ~ sex + pclass + log(age) + sex:pclass + sex:log(age)+pclass:log(age)**

The AIC value of this new model is 767.0838 and the BIC value of this model is 815.0073 according to the R output.

The 2-by-2 table containing the prediction on the test data is shown in Figure 11.

g4log.pred	0	1
0	241	75
1	18	84

Figure 11.

The prediction accuracy is (241+84)/418≈77.75%.

Note: This way may not be reasonable or correct because generally, we should do transformations first before the selection steps. I mention it here just for reference and see how the resulting AIC, BIC and prediction accuracy values perform.

Finally, comparing the four model selection processes above, I would like to choose Model 3 with the highest accuracy among the four and an acceptable AIC value. Though the model in Part 3 gives a lower AIC value, the model in Part 4 gives a lower BIC value, they have the same prediction accuracy.
That is, I choose **survived ~ sex*pclass*log(age)** that has an AIC value of 762.714 and a 77.75% prediction accuracy as the answer to II (c).

(d) Add fare, embarked, sibsp, and parch into the model obtained at II(c).
Model: **survived ~ sex*pclass*log(age) + fare + embarked + sibsp + parch**
This model has an AIC value 736.6964 and the BIC value of this model is 822.9586 according to the R output.
After we call step() function with AIC criterion, we get the following new Model A:
survived ~ sex + pclass + log(age) + embarked + sibsp + sex:pclass + sex:log(age) + pclass:log(age) + sex:pclass:log(age)
The AIC value of the new model is 734.7276 and the BIC value of this model is 811.4051.
The 2-by-2 table containing the prediction on the test data is shown in Figure 12.

```
g6.pred   0   1
      0 220  50
      1  39 109
```
Figure 12.

The prediction accuracy is (220+109)/418≈78.71%.
However, if we call step() function with BIC criterion, we get the following new Model B:
survived ~ sex + pclass + log(age) + sibsp + sex:pclass + sex:log(age)
The AIC value of the new model is 741.4095 and the BIC value of this model is 784.5406.
The 2-by-2 table containing the prediction on the test data is shown in Figure 13.

```
g7.pred   0   1
      0 220  55
      1  39 104
```
Figure 13.

The prediction accuracy is (220+104)/418≈77.51%.
Therefore, Model A selected by AIC criterion gives a higher prediction accuracy value than Model B, which was selected by BIC criterion.

(e) The Table 1 below is a summary of the AIC value and Prediction Accuracy of each model appeared in previous parts.

Models	AIC	BIC	Prediction Accuracy
II (a)	833.8884	854.0577	76.32%
II (b)	810.0969	838.851	76.32%

II (c)	762.714	820.2221	77.75%
II (d)-AIC	734.7276	811.4051	78.71%
II (d)-BIC	741.4095	784.5406	77.51%

Table 1. Comparisons of the previous models.

In conclusion, the model in II (d) chosen by the stepwise selection with AIC criterion gives the smallest AIC value, 734.7276, and it also gives the highest prediction accuracy, 78.71% among the five models; the model in II (d) chosen by the stepwise selection with BIC criterion gives the smallest BIC value, 784.5406 among the five models.

III. Tree models

Divide the data into training and test. Explore various tree models with all the variables **except** Life_boat, name, ticket and cabin.

(a) Fit a tree model.

In this section, I want to fit a classification tree model.

First, I add "minsplit=5, cp=0.000001,maxdepth=30,method="class"" control parameters to the "rpart" function, and I run the commands several times, which give me the most frequent cp value 0.003898635 when the "xerror", the cross-validation estimate of misclassification error, has the minimum value. And in that case "nsplit" is 14.

So, then, I update the cp value to 0.003898635 and re-run the codes several times, which give me most frequent cp value 0.007797271 when the "xerror" has the minimum value. And in that case "nsplit" is 6.

Next, I update the cp value again to 0.007797271 and re-run the codes several times, which gave me the stable cp value 0.007797271 when the "xerror" had the minimum value. In addition, the cp value before it is 0.0204678 when "nsplit" is 5.

Therefore, as long as we choose a value between 0.0078 and 0.0204, we could get a tree model that give us the smallest CV errors with the optimal 6 splits.

So, I just pick 0.008 as our final cp value.

Then, I prune the tree and plot it as the following:

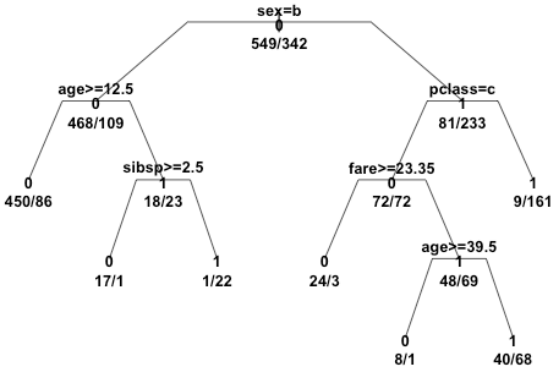

Figure 14. The Classification Tree Model Plot.

As we can see in the tree graph, there are totally 549 passengers who died in the disaster and 342 passengers who were alive. (Blue: dead; orange: alive)

First, let us investigate the left child of the big tree.

1) Among all male passengers, 109 of them were alive and 468 of them were died. Among all dead 468 males, 450 of them were 12.5 years old or older; 18 of them were younger than 12.5 years old.
2) Among all live 109 males, 86 of them were 12.5 years old or older and 23 of them were younger than 12.5 years old.
3) Among those 536 male passengers who were 12.5 years old or older, no one had 2.5 or more siblings/spouses on board.
4) But among those 41 male passengers who were younger than 12.5 years old, 17 of them had 2.5 or more siblings/spouses on board and were dead; 1 of them had 2.5 or more siblings/spouses on board and was alive; 1 of them had fewer than 2.5 siblings/spouses on board and was dead; 22 of them had fewer than 2.5 siblings/spouses on board and were alive.

Next, let us investigate the right child of the big tree.

1) Among all female passengers, 233 of them were alive and 81 were dead. Among all 233 live female passengers, 72 of them were in the Lower Passenger Class and 161 of them were in the Middle or the Upper Passenger Class.
2) Among all 81 dead female passengers, 72 of them were in the Lower Passenger Class and 9 of them were in the Middle or the Upper Passenger Class.
3) Among all dead 72 female passengers who were in the Lower Passenger Class, 24 of them had ticket fare that cost 23.35 pounds or more; 48 of them had ticket fare that cost fewer than 23.35 pounds.
4) Among all live 72 female passengers who were in the Lower Passenger Class, 3 of them had ticket fare that cost 23.35 pounds or more; 69 of them had ticket fare that cost fewer than 23.35 pounds.
5) Among the 48 dead female passengers who had ticket fare that cost fewer than 23.35 pounds, 8 of them were 39.5 years old or older and 40 of them were younger than 39.5 years old.
6) Among the 69 live female passengers who had ticket fare that cost fewer than 23.35 pounds, 1 of them was 39.5 years old or older and 68 of them were younger than 39.5 years old.

Consequently, we have the following comments about some interesting splits from the above investigations of the tree model we have got:

1) The “sibsp>=2.5” split in the left sub-tree kind of makes sense because younger male passengers with more than 2.5 siblings/spouses were more likely to care for the female members in the family and let them board the life boats first. So, these male passengers were more likely to survive if they boarded the lifeboats first without caring about others.
2) The “fare>=23.35” does not seem to make sense here because female passengers in the Lower Passenger Class with more expensive ticket were far more likely to

die and there were even more female passengers in the same class with cheap ticket survived than dead.

3) The "sex=male" split makes sense here because female passengers were all tended to have higher priorities of boarding the lifeboats than male passengers and as a result, were more likely to survive.
4) The "age>=39.5" split also makes sense here because female passengers at that age were inclined to already have kids. So, when the space was limit in a lifeboat, they gave up the chance of boarding it and let their kids board the boat instead.

The 2-by-2 table containing the prediction on the test data is shown in Figure 15.

```
g8.pred   0   1
      0 216  49
      1  43 110
```
Figure 15.

The prediction accuracy is (216+110)/418≈77.99%.

(b) Fit a random forest.

The following is the random forest model we fit.

Call:
randomForest(formula = survived ~ ., data = train.data, proximity = T, importance = TRUE, ntrees = 500, method = "class")

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 2

OOB estimate of error rate: 16.39%

Confusion matrix:

	0	1	class.error
0	509	40	0.07285974 = (40/549)
1	106	236	0.30994152 = (106/342)

Majority voting within the forest of 500 generated trees does classification in random forests.

The 2-by-2 table containing the prediction on the test data is shown in Figure 16.

```
g9.pred   0   1
      0 223  59
      1  36 100
```
Figure 16.

The prediction accuracy is (223+100)/418≈77.27%.

(c) Comparing the prediction accuracy values in Part (b) and (c) above, we find that the single tree model has a larger prediction accuracy than the random forest model.

References

Kaggle.com (2012). *Titanic: Machine Learning from Disaster*
Retrieved from:
http://www.kaggle.com/c/titanic-gettingStarted/data