MACHINE LEARNING
PROJECT 1

SYS 6016-02

Surviving the Wreck of the Titanic: The Hard Truth of Class and Love

Chao Dai
cd3rg@virginia.edu
Andrew Biedermann
amb4ht@virginia.edu
Siyuan Guo
sg5jn@virginia.edu
Sen Cong
sc9vv@virginia.edu
Instructor:
Abigail A. Flower
aaf4q@virginia.edu

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

Abstract

This project is mainly about analysis the data of Titanic to find suitable attributes for decision tree, then predict whether a human would survive in titanic. In the analysis part, boxplot of each attributes has been applied to find which attribute is significant. And in the data clean process, all attribute which lost too much data has been dropped. After data clean, new data used to feed decision tree. Decision tree has been tested through test sets and the accuracy reached 81%. Then, we analysis the decision tree and find several interesting results. Those result used to reach the conclusion of titanic.

1 Introduction

The sinking of the Titanic is one of the most notorious maritime disasters in modern history. Much of its renown stems from the enormous scale and purported safety of the vessel. Engineering journals of the day claimed the enormous ship was “practically unsinkable,” while passengers report to have heard boasts that “even God himself couldn’t sink this ship.” (1)

The Titanic was on pace to set a record time for the trans-Atlantic crossing on the evening of April 15th, 1912 when it struck an iceberg, which ruptured at least five hull compartments (2). Though some officers and passengers were initially skeptical of the incident’s severity, the dire nature of the situation soon became evident to everyone on board. Passengers rushed to the deck to secure a spot in the lifeboats for themselves and their loved ones. In an act of chivalry, women and children were given priority when loading the lifeboat, greatly increasing their chance of survival. However, sex was not the only factor which influenced an individual’s chance of survival.

We hypothesize that social class will play an important role in determining a passenger's probability of survival. We also expect that first class men will act as high-minded gentleman, yielding their safety to other, less principled

passengers. Finally, we hypothesize that members of a family will have a higher rate of survival than lone passengers, since they will have additional advocates to get them into a lifeboat. The above hypotheses are examined in the present study.

2 Methods

2.1 Data Source

Titanic data was originally obtained from Encyclopedia Titanica (3). There totally 3 different versions of data so far, the first 2 versions were kept on Encyclopedia Titanica. The 3rd first version was contributed by Thomas Cason of UVa, who had greatly updated and improved the titanic data frame. The 3rd version is the version we used in this research.

There are 1309 passengers and 14 variables in the data, as shown in table 1. The variables include passengers geographical information, such as name, sex, age, and survival. It also has inter personal relationship, such as number of siblings, spouses, parents, and children. (Indirect relationships, such as nephews, uncles, and in-laws, are not included in this data) There also exists ship related information, such as pclass, ticket, cabin, as well as fare. There is

one point worth noting, that ticket number is an important indicator for family relationship, as people from same family share the same ticket number.

Variable	Specification
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	survival Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embark	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Table 1: Meaning of each variable

2.2 Data Cleaning

Data was read from source data file called “titanic3.xls”, as a pandas dataframe in python. As shown in figure, that there are 4 variables have more than or close to half of values missing. To make results better representable, data of these 4 variables were deleted. The the other variables with missing data, imputations was performed, basically with mean imputation method. Briefly, all the available value of variable being imputed were pulled, the mean value was calculated, and all the missing spots were replaced with this mean value.

2.3 Feature Extraction

As almost all remaining variables are numerical variable, to make the analysis more consistent, we decided to extract 2 new features out of sex, so that it can become a numerical variable. The 2 new features are named ‘male’ and ‘female’, each coded as 0 and 1. For ‘male’ variable, males are coded as 1, and females are coded as 0. ‘Female’ variable is coded the other way around.

Figure 1: Missing values stats. Each bar represents one variable, blue color means counts of missing value.

2.4 Feature Selection

Influence of pclass over survival was plotted, as shown in figure 2, though the people who are survived are separated in all the three pclasses, the people who are dead are mostly in the third class. The average number of pclass in dead is around 2.8. So pclass really has a big influence on the survival rate.

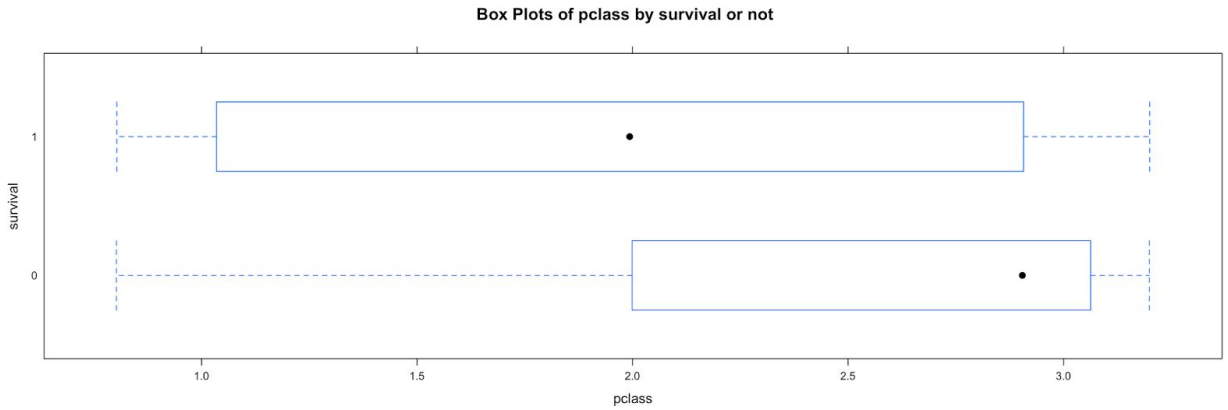

Figure 2: pclass vs survival

Gender is an important feature, as it directly reflects how women were protected and respected, which may have direct effect on survival. Figure 2 was plotted by using the original sex variable and survival variable. Death is coded 0 and survival is coded 1, the male survival average value is very close to 0, wheres in female, the average value is near 1. This box plot shows that the gender could be a significant feature that influences the survival pattern.

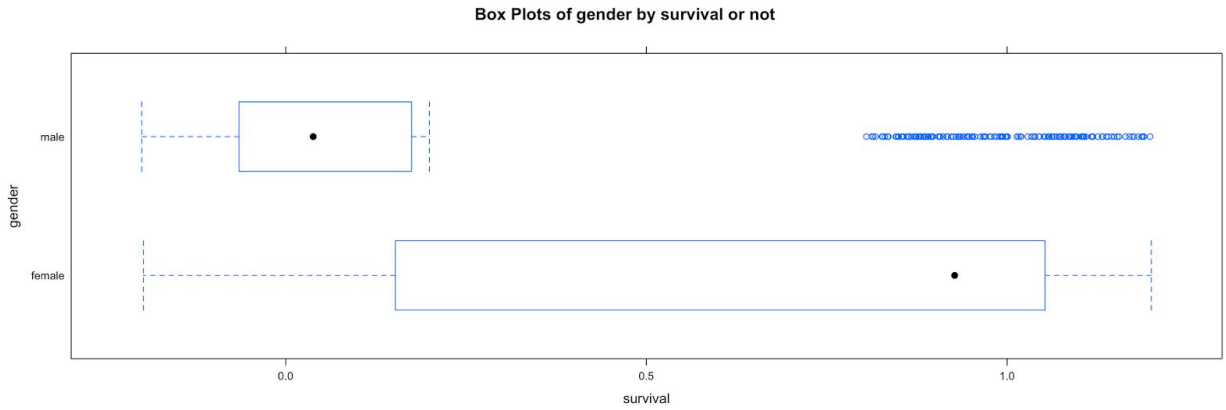

Figure 3: gender vs survival

Figure 4 is about the influence of age to survival rate. The average of survival people and dead people are nearly the same. The age feature is less significant compared with gender and pclass. However, there is an interesting finding, that the top 2 oldest passengers both survived. The other interesting finding is that the distribution of dead people is more condensed than that of survived. As a result, this feature may help distinguish extreme cases in our decision tree classifier.

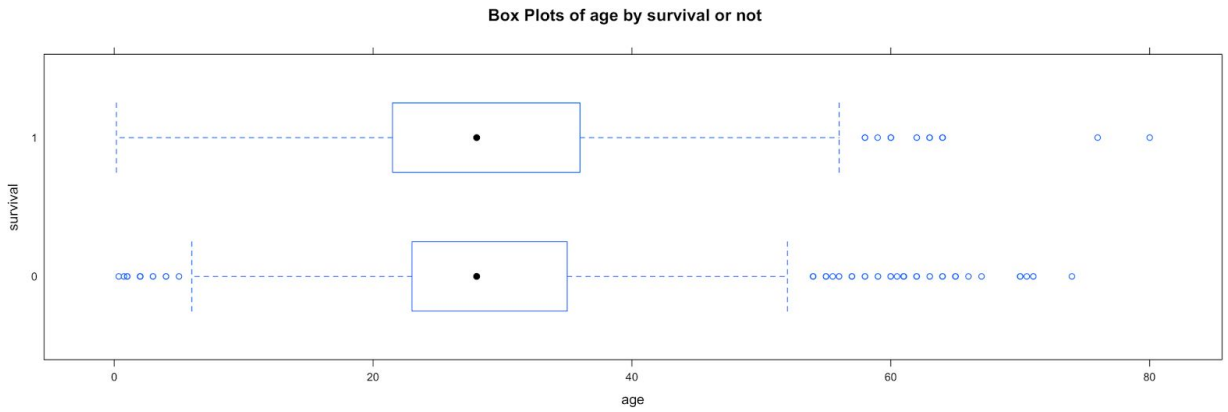

Figure 4: age vs survival

Sibling numbers and survival was plotted in figure 5, the average number of siblings doesn't differ in dead and survival passengers. However, it's interesting to note that all passengers with more than 4 siblings are all dead. So, siblings would also help in extreme cases, and due to this fact, that it might also interact with other features for more interesting findings.

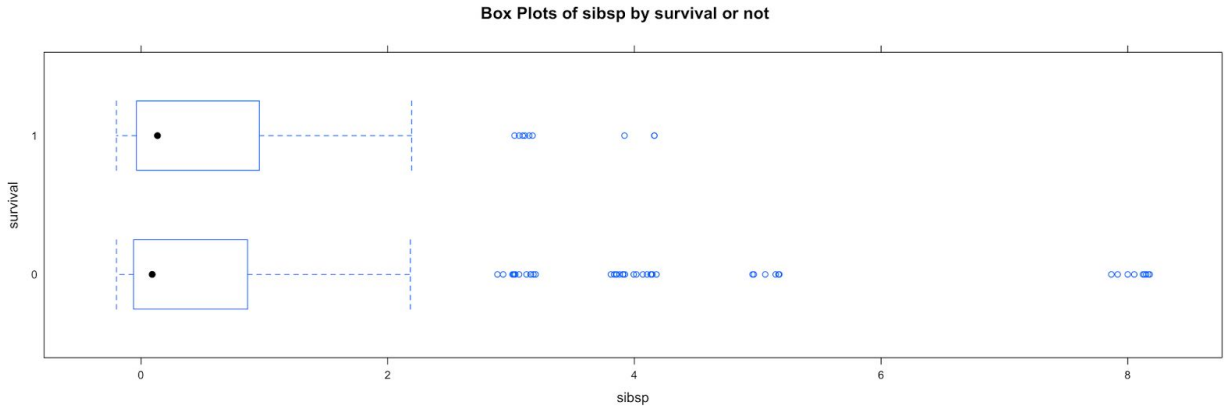

Figure 5: sibsp vs survival

Figure 6 shows that the number of family member (except siblings) has influence

on the survival rate. In the box plot we can see that the people who survived have a broader distribution than the people who dead. The average number is similar because many people have no family member. So we can make a hypothesis that people who have more family members will survive.

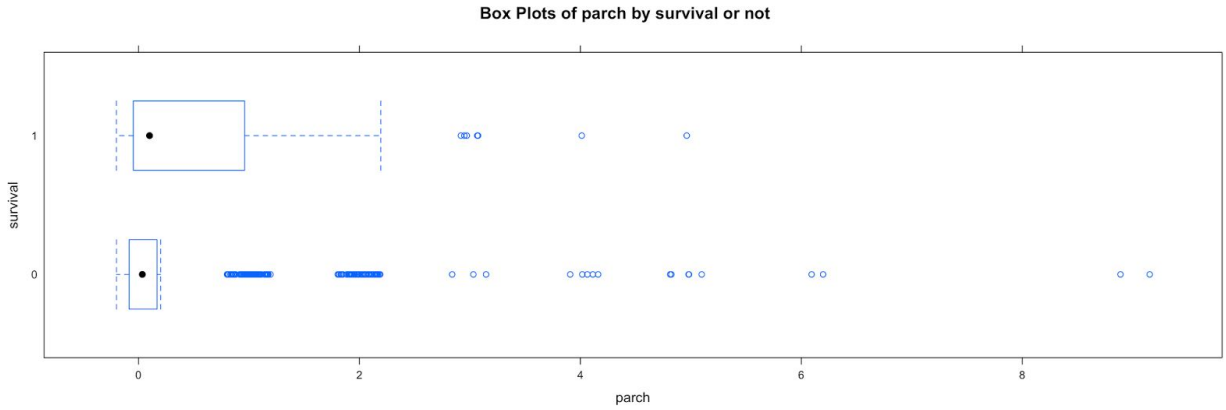

Figure 6: parch vs survival

The fare has a faire huge influence to the survive rate, according to figure 7. The average fare of survival people is much higher than of dead people. So we can make a hypothesis that the people who paid more have better chance of survival.

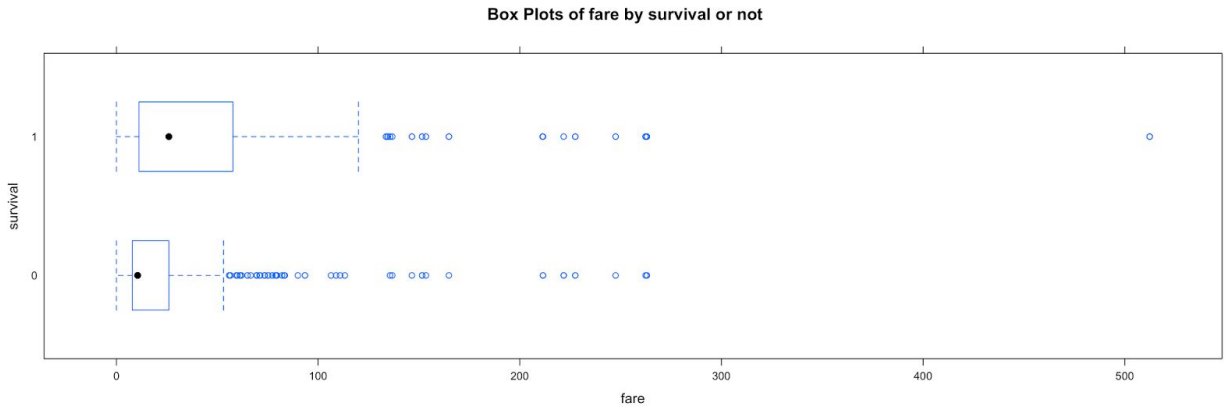

Figure 7: fare vs survival

2.5 Decision Tree Modeling

As mentioned above, that our hypothesis are family members, gender, and social classes related, so these features are definitely included in our analysis. However, to avoid any bias influenced by our predefined hypothesis, other features were also taken into consideration. Based on the feature selection analysis in section 2.4, we have decided to use pclass, age, sibsp, parch, fare, and gender as predictors for survival. For the purpose of validating the decision tree, the whole data set was divided into 2 parts, training set, and test set. ⅔ of samples are training set, and the rest are test set, the selection was random based. The decision tree was built using all these features, and prediction was carried out on test data, a confusion matrix and prediction accuracy were calculated.

To avoid having too complicated or less represented tree, the max depth tested by created trees with depth range from 1 to 10. The peak accuracy value appears when max depth is 3 or 4, and these 2 values are identical. As a result, to make the tree more easily readable, as well as still ensure accuracy, max depth of 3 was used for our decision tree.

The other important parameter in decision tree is the max features number. This was also experimented by using different combinations of features, and by

looking at accuracy, the best parameter can be identified. By doing this experiment, there is only one other tree with less feature has comparable accuracy with the full decision tree. The only difference between these 2 trees is the relatives number. However, as discussed above, the relatives number may not affect the overall accuracy, but it might provide some insight over extreme cases. Also, as no other tree has better accuracy than original decision tree, we would still like to use original decision tree.

3 Results and Discussion

3.1 Data cleaning

Data cleaning was performed as described in method section. The samples number remains unchanged, which is 1309. Due to missing data problem, data for cabin, boat, body, and home/destination are all dropped out of the data frame. As identified by figure 2, there are roughly 20% data missing in age. The missing data in age are imputed as described in method section, simply by mean value of all available ages. Although not visible, there is also 1 missing point in fare, which is also imputed as mean value.

3.2 Decision tree analysis

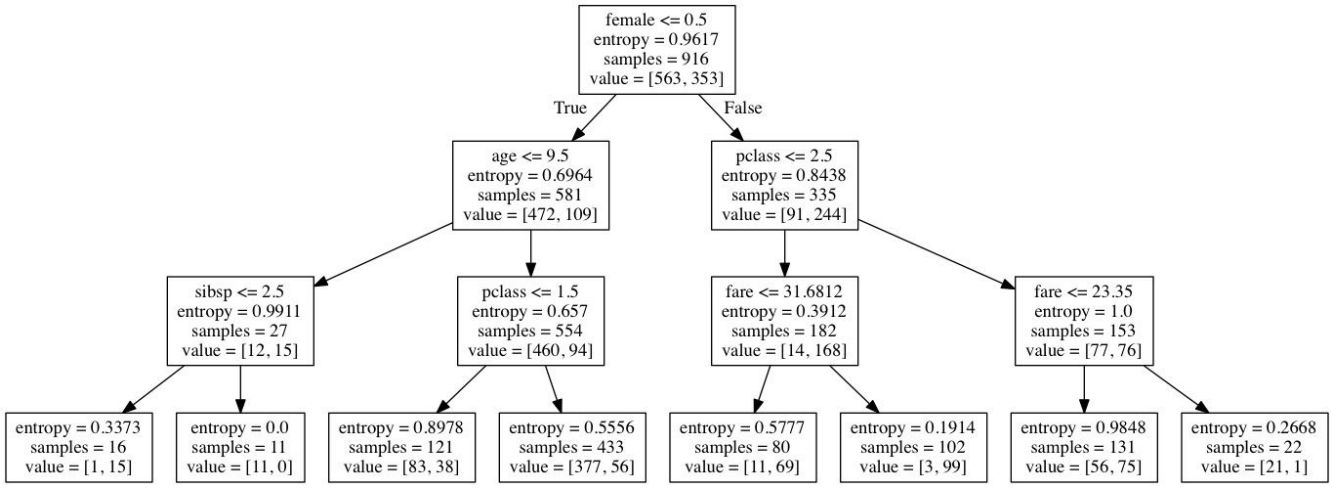

Figure 8: The original decision tree

The results of the decision tree analysis are shown in Figure 8. In the tree, a value of 1 in the root node indicates that a passenger is a female, and the value is interpreted as [number dead, number survived]. From the root node, we note that, as expected, females had a much higher chance of survival than males. Indeed, the survival rate for females is 73% while the survival rate of male passengers is a mere 19%, supporting the claim that women were given priority in boarding lifeboats.

We find that the conventional wisdom that children also received priority boarding privileges is also supported by the decision tree result. Male children, under the age of 10, had a survival rate of 56%, compared to just 17% for older males. Interestingly, the decision tree shows that no male children with more than 2 siblings survived the shipwreck, while almost all male children with 2 siblings or less survived. This result may not be significant when we consider the

size of the sample. Children were given priority boarding with women, and children would likely have remained with their families. Therefore, this results could be anecdotal, as it could merely represent the unlucky fate of 3-4 families.

In addition to the precedence given to children, the decision tree results suggest that class strongly influence survival rate. The survival rate for first class males was 31%, over twice the male survival rate of the lower classes at 13%. In primary accounts, several men ask to accompany their partners as protectors or rowers, or are allowed to board when no additional women are near the lifeboats. First class men, whose location in the upper cabins may have allowed them to reach lifeboats before men of lower class would be more likely to be granted a rare male spot in the boats.

First class privilege extended to female survival rates as well, as over 92% of first class women survived. Women in the lower classes were generally less fortunate, with a survival rate of about 50%. Interestingly, lower class women with a high fare had a very low survival rate, suggesting that the cabin location purchased by such individuals placed them at a significant disadvantage for survival.

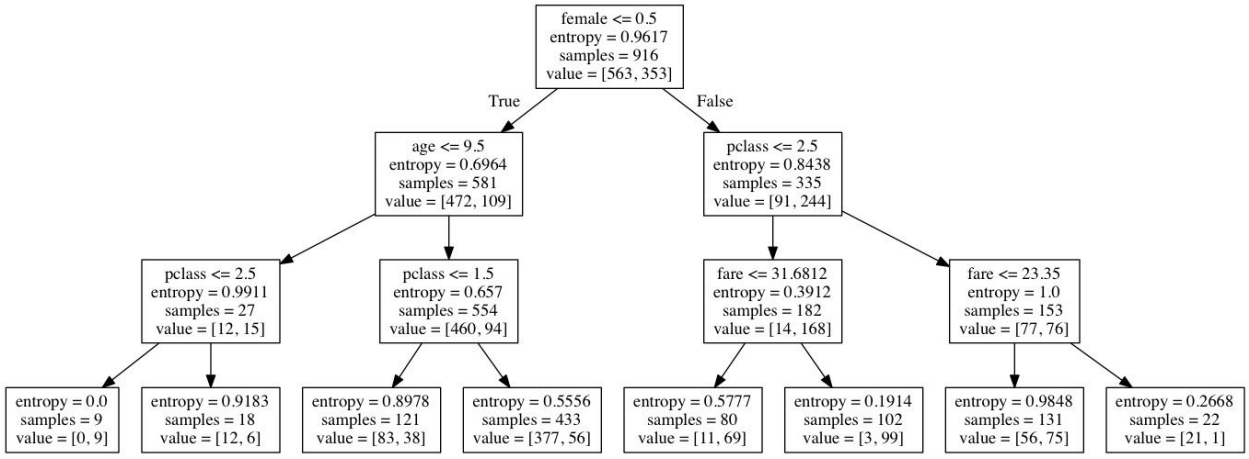

Figure 9: Non-sibling

In Figure 9 we reexamine the survival rate of men, focusing on the role of class, rather than sibling or spousal relations to determine survival rate. The data shows that no 1st or 2nd class male children survived, supporting our hypothesis that the earlier sibling result was likely a results of a few unlucky families. This result is counterintuitive as we would expect upper class children would be among the first admitted to lifeboats.

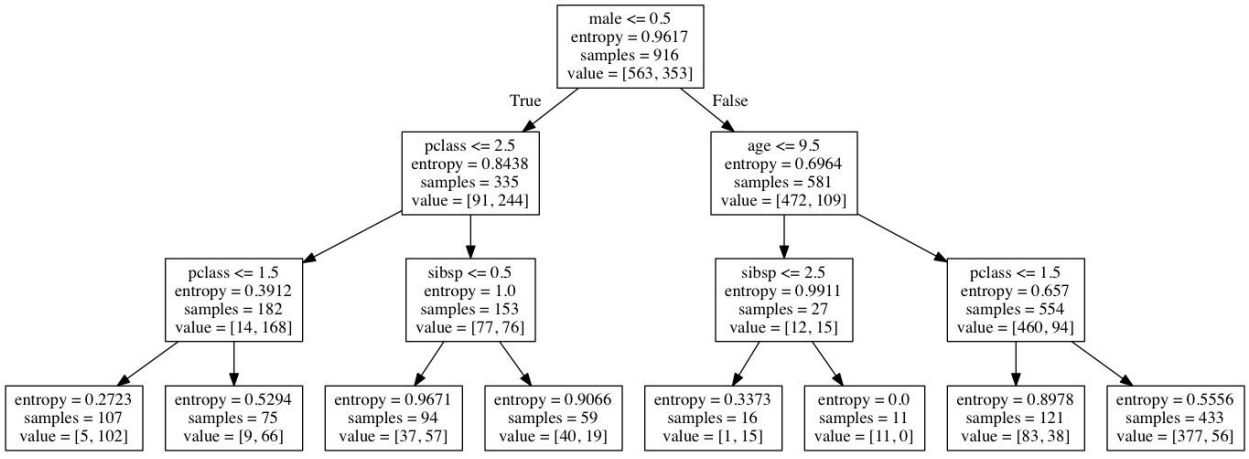

Figure 10: Non-fare

Finally, in Figure 10 we examine factors affecting female survival rates, by removing the fare attribute from the analysis depicted in Figure 8. Note that the root attribute in this case is male, so the orientation of the tree is reversed. The data again shows that women in the first 2 classes were likely to survive. However, an interesting result is obtained for 3rd class women. We find that women with no spouses or siblings had a much higher rate of survival than women with such relations, 61% to 32% respectively. This results is somewhat intuitive, as women with family might wait for a lifeboat on which they may take their husband or children, rather than fending for themselves. Since a 3rd class husband is much less likely to be allowed in a lifeboat, a devoted wife is therefore much more likely to remain on the ship with her husband, while women without such ties take her place.

4 Recommendations

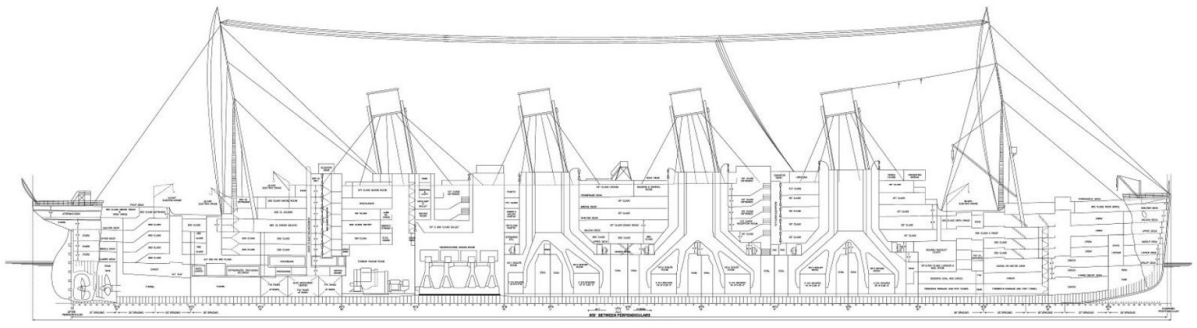

Figure 11: Side view of Titanic

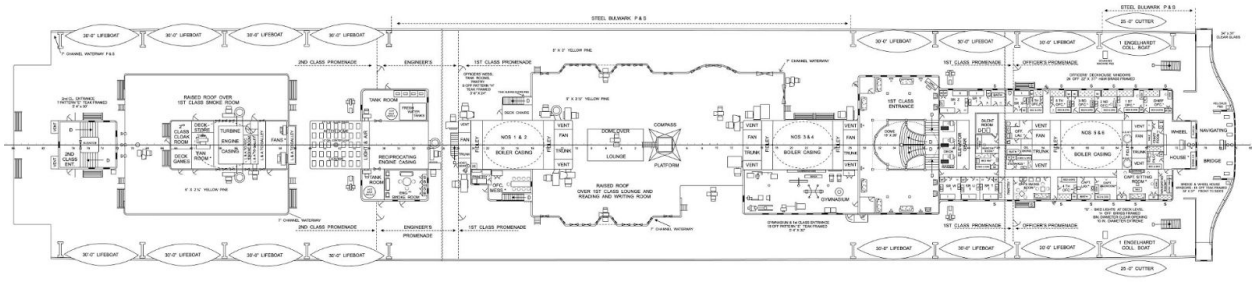

Figure 12: Top view of Titanic

From the decision tree presented in the previous section, it is obvious that first class passengers had a much higher rate of survival than lower class passengers. While this may be, in part, due to their higher socioeconomic status, careful examination of the Titanic's deck plans reveal that the location of first class cabins on the Titanic gives first class passengers a significant advantage in finding a lifeboat. First class cabins are distributed primarily through the middle and front of the ship, while lower class quarters are located toward the rear of the ship. As a result of their more central location, first class passengers could move quickly towards lifeboats in either the front or rear of the ship, and would have had almost exclusive access to lifeboats in the front of the ship.

Meanwhile, 2nd and 3rd class passengers, whose quarters are spaced more compactly would likely rush to the nearest lifeboats at the end of the ship. It is therefore plausible that the location of cabins, a characteristic which correlates

with social status, may play a significant role in survival rate. Such contributions of cabin location to passenger survival rate could be examined in future studies.

5 Conclusions

While the conventional wisdom that women and children were given priority to board lifeboats holds true for this analysis, social status was also a very important factor in survival. Though Benjamin Guggenheim was reported to say, “We’ve dressed up in our best and are prepared to go down like gentleman.” such sentiments were not felt by the majority of first class men, whose survival rate was nearly as high as married 3rd class women, and more than doubled the survival rate of their lower class compatriots (4). It was also generally found that women and children with siblings or spouses had lower rates of survival rates than those without family ties. Sadly, though we canonize the Guggenheims and Jack Dawsons who went down with the ship, in the majority of cases passengers who were out for themselves were more likely to survive.

References:

1. http://www.nbcnews.com/id/46916279/ns/technology_and_science-science/t/titanics-legacy-fascination-disasters/#.VtkIP3UrKV4
2. http://www.history.com/this-day-in-history/titanic-sinks

3. Hind, Philip. *Encyclopedia Titanica*. Online-only resource. from http://www.encyclopedia-titanica.org/
4. http://www.encyclopedia-titanica.org/titanic-victim/benjamin-guggenheim.html
5.