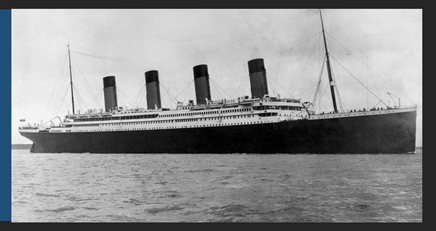
# Titanic: Surviving the Disaster

*Department of Mathematics, University of Wisconsin-La Crosse, WI 54601*

Nicole Nelson, Song Chen



## Introduction

The titanic was catastrophic as two-thirds of her passengers died in her maiden voyage. The list of those that passed has been published and was turned into a data science competition to see who can create a model that could predict the survival rate with the highest accuracy. The question then became who died and who lived and what correlations existed between them. Most people have watched the movie *Titanic* and saw that Rose lived and Jack died. Despite your opinions on the matter, however, using mathematical modeling it will show that even if there was room on the raft, Jack still most likely would've died and Rose would've lived. Now that a probable outcome is proven, how would you fare on the titanic?

## Data Sets & Variables

The predictions were created by analyzing two data sets: the training set (**Figure 1**) which contained 11 variables including the survival, and the testing set (**Figure 2**) which contained all variables except the survival. The model was created using the training set then evaluated with the test set. For missing data, the variable was replaced or disregarded. RStudio was used for analysis.

| PassengerId | Survival | Class | Name | Sex | Age | Sibling/Spouse | Parent/Child | Ticket | Fare Paid | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Passed | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | A/5 21171 | 7.25 | | S |
| 2 | Survived | 1 | Cumings, Mrs. Florence | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 3 | Survived | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.925 | | S |
| 4 | Survived | 1 | Futrelle, Mrs. Jacques Heath | female | 35 | 1 | 0 | 113803 | 53.1 | C123 | S |

**Figure 1**. Training set with known outcome

| PassengerId | Class | Name | Sex | Age | Sibling/Spouse | Parent/Child | Ticket | Fare | Cabin | Embarked | Survival |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 892 | 3 | Kelly, Mr. James | male | 34.5 | 0 | 0 | 330911 | 7.8292 | | Q | ? |
| 893 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47 | 1 | 0 | 363272 | 7 | | S | ? |
| 894 | 2 | Myles, Mr. Thomas Francis | male | 62 | 0 | 0 | 240276 | 9.6875 | | Q | ? |
| 895 | 3 | Wirz, Mr. Albert | male | 27 | 0 | 0 | 315154 | 8.6625 | | S | ? |

**Figure 2**. Test set with unknown outcome

This is what we want to predict

## Feature Engineering

More variables were created by using feature engineering — the modification of preexisting variables whether it is pulling apart, cutting apart and/or merging of data. The following variables were created from stories that were prevalent.

A famous maritime sayings is "women and children first". An child variable was created by making every passenger under the age of 18 a child and those above an adult. From **Figure 3** it is seen that this holds true for the titanic.

| Child | Sex | Survived |
|---|---|---|
| No | F | 75.28958% |
| Yes | F | 69.09091% |
| No | M | 16.57033% |
| Yes | M | 39.65517% |

**Figure 3**. Table of survival probabilities for children of both sexes.

| Fare2 | Survived |
|---|---|
| <10 | 19.94048% |
| 10-20 | 42.45810% |
| 20-30 | 42.64706% |
| 30+ | 58.75000% |

**Figure 4**. Table of survival probabilities for the 4 fare buckets

The next saying explored was "it pays to be rich". Fares were merged into four price ranges and it was found that the fare was directly correlated to survival odds.

Potentially the most important saying analyzed was that "prestige lives". By pulling apart and cutting up the name, the title of the person was isolated. It was again found that the saying held true that those with a higher title had better survival odds.

## Simple Models

To start, the first prediction was that everyone dies. This came back with a 63% accuracy which at the time was better than half the competition.

Next it was observed that about two-thirds of the passengers were male so the prediction only men died was made. This came back with an improved accuracy of 77%.

Since men had a terrible survival rate, women were looked at. It was guessed that third class women would have low survival chances, so it was predicted that men and lower class women would die, but this came back with the exact same accuracy of 77%.

## Decision Trees

Instead of examining each variable by hand, decision trees were used to automate the process. Decision trees are a classification model that scans through all the variables and finds the one that causes the largest split (most importance). This method resulted with an 79% accuracy using feature engineered variables.

Try reading the tree yourself using Mr. Kelly James. Does he survive?

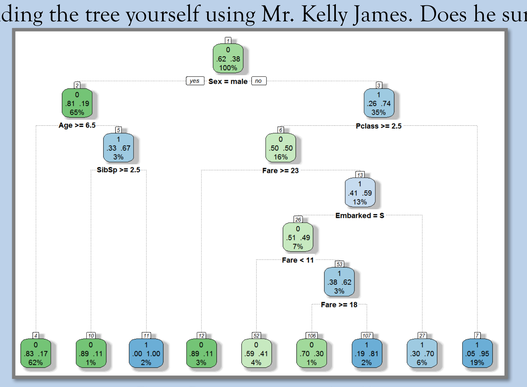



**Figure 5**. Decision tree using original variables

## Random Forest

Due to decision trees splitting on the variable of highest importance, the model is unable to distinguish if the variable should be the first split or later in the tree to reduce error. Random forest creates multiple (1000+ at times) decision trees and takes the average to determine which variable is most important at each split. Random forest came back with an accuracy of 80% (**Figure 6**).

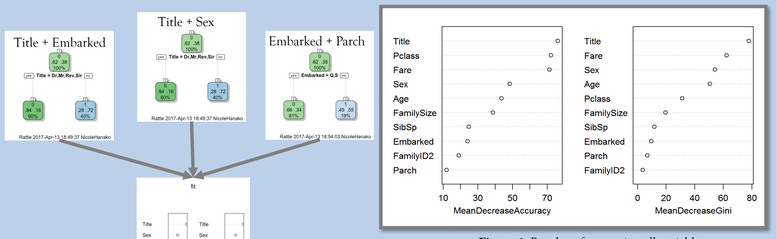



**Figure 6**. Random forest using all variables

**Figure 7**. An example of how decision trees are averaged into a random forest. The first two decision trees had a high accuracy while the third had a low accuracy. Once combined into random forest it was found that title was significant whereas embarked was not

## Classification Problem

From analysis it is clear that the question posed is a classification problem. Various methods were analyzed and it was found that conditional inference random forest and logistic regression (**Figure 8**) proved to be the most effective with an accuracy of 81% and 87% respectively.

| | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
|---|---|---|---|---|---|
| NULL | | | 799 | 1066.33 | |
| Pclass | 1 | 84.779 | 798 | 981.56 | < 2.2e-16 *** |
| Sex | 1 | 240.084 | 797 | 741.47 | < 2.2e-16 *** |
| Age | 1 | 20.638 | 796 | 720.83 | 5.549e-06 *** |
| SibSp | 1 | 13.549 | 795 | 707.28 | 0.0002325 *** |
| Parch | 1 | 0.854 | 794 | 706.43 | 0.3554687 |
| Fare | 1 | 0.736 | 793 | 705.70 | 0.3910806 |
| Embarked | 2 | 1.435 | 791 | 704.26 | 0.4879177 |
| Fare2 | 3 | 2.974 | 788 | 701.29 | 0.3956678 |
| Title | 10 | 45.596 | 778 | 655.69 | 1.698e-06 *** |
| FamilySize | 0 | 0.000 | 778 | 655.69 | |
| FamilyID2 | 21 | 54.478 | 757 | 01.21 | 8.419e-05 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Figure 8**. ANOVA table which was part of the logistic regression model. The p values close to zero denote that the variable is significant.

## Accuracy Measurement

To measure accuracy, models were submitted to Kaggle which was the server hosting the data competition. An immediate score was provided.

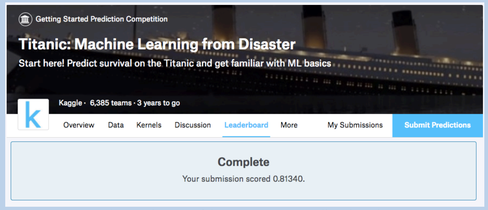



**Figure 9**. Screen clipping of the competition as well as the submission

## Conclusions

Of the classification methods logistic regression proved to be the most effective model in predicting the survival with an accuracy of 87%. Feature engineering was also important as more information was able to be extracted; it was seen that title was the key component in predictions. Following that, age, class, and sex were important factors in the decision.

After the last submission the student was ranked 102 of 6,338 competitors.


**Figure 10**. Screen clippings from the competition of the ranking

## References

Alice, Michy. "How to Perform a Logistic Regression in R." Rbloggers. N.p., 13 Sept. 2015. Web. 03 Nov. 2016.

Mick001. "Logistic Regression Tutorial Code." Gist. N.p., 13 Sept. 2015. Web. 03 Nov. 2016.

Stephens, Trevor. "Titanic: Getting Started With R." Trevor Stephens. N.p., 09 Jan. 2014. Web. 03 Nov. 2016.

"Titanic: Machine Learning from Disaster." Kaggle: Your Home for Data Science. N.p., 28 Sept. 2012. Web. 03 Nov. 2016.

Trevorstephens. "Trevorstephens/titanic." GitHub. N.p., 19 Jan. 2014. Web. 03 Nov. 2016.

## Acknowledgements

Dr. Song Chen & Dr. Chad Vidden
Colin Giles & Elizabeth McMahon

For more information contact
Dr. Chen at schen@uwlax.edu or Dr. Vidden at cvidden@uwlax.edu