Decision Analysis of Factors Impacting Survival of the Titanic Sinking

Tricia M. Lang

Southern New Hampshire University

DAT 511: Decision Methods and Modeling

Dr. Ozan Ozcan

07 August 2022

Abstract

This paper examines the factors that impact whether an individual would survive the sinking of the Titanic. Reasons why specific individuals survived, and others did not, has been assumed, speculated, and even glorified in movies. Through analysis of the data provided in the passenger manifest, cross-referenced with who did survive, the variables of greatest impact were determined.

A Classification and Regression Tree (CART) model was used to sort through the different factors available in the dataset, as well as combinations of factors, and determine which were most significant. The variables of title, passenger class, a factor of fare per person by passenger class, age, and total fare were identified as being relevant to predicting survival. The model created was determined to be valid and reliable through error matrix and ROC measures.

Table of Contents

Introduction ……………………………………………….…………………………………. 4

Data Appraisal ……………………………………………………………………………….. 5

Methods …………………………………………………………..………………………….. 6

Evaluation ………………………………………………………………………………….… 8

Model ………………………………………………………………………………………… 9

Results …………………………………………………………………...………………….. 10

Limitations …………………………………………………………….……………………. 12

Conclusion …………………………………………………………..……………………… 13

References ………………………………………………………...………………………… 15

# Introduction

In 1912, the boat that was touted as being unsinkable, the RMS Titanic, departed from Southampton, England, stopping at Cherbourg, France, and then Queenstown, Ireland, before heading across the Atlantic Ocean to New York. Five days later, before it could reach New York, the Titanic struck an iceberg and sank. Aboard were approximately 1315 passengers, of whom 500 survived (History.com Editors, 2022 and Salem Media, 2018). Information about the passengers is available from the passenger manifest, providing insight around which passengers survived and which did not. The data for passengers can identify which factors were most influential in predicting whether an individual would survive the sinking of the Titanic.

The benefit of answering the research question of which factors influenced a person’s ability to survive the sinking of the Titanic crash can be farther reaching than just that historical event. Knowing the outcomes given the data provides the opportunity to build and test models to identify whether they can predict outcomes accurately. This model specifically would be able to determine if the same factors have had the same impact on survival in other instances of ships sinking.

The target audience for this model could be individuals in disaster response or safety management. The model created using the Titanic data may point out weaknesses in rescue plans and identify which individuals based on specific factors may need more support to be able to exit a crisis. Identifying factors that negatively impact survival may assist in targeting those individuals during rescue operations. These factors may differ from other similar events and determining differences may help identify mitigating circumstances that change the outcome.

# Data Appraisal

The Titanic data set is from the passenger manifest of the ship. The manifest is the record of passenger names as well as other information, such as port of embarkation, cabin number, and fare. This dataset included 10 variables from the manifest and 3 variables were derived, fare per person and factor of fare per person/class. A list of the variables and their definitions can be seen in Table 1.

The fare per person was calculated by dividing the fare by 1 (for the individual) plus the number of siblings or spouses and the number of parents and children. As an example, if the fare were 21 and the individual had 2 children with them, the fare per person would be calculated as 20 (1 + 2). So, in this example, the fare per person would be 7, as each of the children would have the same divisor of 3, 1 for themselves, 1 for their parent, and 1 for their sibling.

Preliminary modelling indicated the factors of fare per person and passenger class were important factors in the model. The factor of the fare per person divided by the class was created to determine whether an interaction of these variables had an impact and if a greater expenditure per person for a ticket combined with the passenger class of the individual influenced whether the person survived. This factor would decrease both with a lower passenger class as well as a lower fare per person paid. As well, the factor would increase with a higher passenger class and a greater fare per person paid.

Early models indicated gender and age were significant factors, so to see if there were greater differences among males, the variable Title was derived from the data. The dataset included the passenger names in the format of Smith, Mr. John A. From this, the title of “Mr.” was able to be derived into its own data column. There were 17 unique titles, including military ranks, royal titles, and others.

In cleaning the data, some values were missing, which rendered some variables unusable. Cabin, for example, had many missing values that could not be reliably imputed. Some of the variables, such as age, had only a small number of missing values. These were imputed with the overall average age of all passengers in the train data set (30 years old). Port of embarkation also had very few missing values, which were imputed with “Unknown”, as they likely had little impact on the outcome given it was less than a quarter of one percent of all rows.

**Table 1.**

*Variables and Definitions in the Titanic Data Set*

|  |  |  |  |
| --- | --- | --- | --- |
| Categorical Variables | Definition | Numeric Variables | Definition |
| survival | Survival – 0- no, 1- yes | Age | Age in years |
| pclass | Ticket class- 1, 2, or 3 | sibsp | # of siblings / spouses aboard |
| sex | Sex- male or female | parch | # of parents / children aboard |
| Title | Titles of individuals | fare | Passenger fare |
| embarked | Port of Embarkation-  C, Q, S, U | Farepp | Fare per person |
| cabin | Cabin number | FareppXclass | Factor of Farepp divided by  class |
| ticket | Ticket number |  |  |

# Methods

A Classification and Regression Tree (CART) model was developed using a bottom-up approach. Using R and Rattle, the data set was partitioned with 70% of the data points included in the training data (n=623), 20% in the validation (n=179), and 10% in the test data (n=89). A secondary set of data with no outcomes included was used for validation of results compared to known outcomes.

An unpruned tree was created initially, using all variables that were not excluded due to missing values. From here, the tree was pruned in a manner described by Krueger (2022), looking at the most significant subtrees. As variables were isolated, the accuracy of the tree was noted as well as the relative significance of the variables. The variable that did not show any relevance was the port of embarkation, with no combinations identifying this factor as significant, and so it was excluded from further models.

As there are often ideas of women and children being those who filled the lifeboats first, the two variables of sex and age were looked at together. Sex of the individual was noted to be the most significant individual variable, with males largely not surviving. The different iterations of models differed on what age was a significant point, thereby at what decision point becoming a node, and depended on the other variables included in the model. Removing Age as a factor, the passenger class was shown to be of greater significance.

Fare became a curious variable through the pruning process, as results seemed odd that higher fare price would mean a lower chance of survival for female. This led to questions about identifying a fare per person, as those with higher fares may have been including others, such as spouses or children, which may have impacted leaving the boat. Additionally, passenger class was noted to be significant, particularly with fare. This led to identifying if there was an interaction of these variables that would impact the survival rate. With the idea that the higher the ticket price per person and the higher the passenger class would be a greater number, the ticket price per person was divided by the passenger class to create this interaction factor of FarePP and PClass. This factor was determined to be relevant to the decision tree.

Another variable was also examined, as the sex of the individual was identified as a strong determinant for survival. Titles were extracted from the passenger names to explore whether they had an impact on survival. Fewer than 23 of the 891 rows of data had titles outside that of a layperson. Examining this data, the divisions between the titles became a deciding factor for males, particularly. Once the model incorporated this variable, the variable of sex was no longer relevant. The titles of Don, Dr., Jonkheer, Mr., and Rev. were in the initial decision node with those with that title being partitioned to the less likely to survive side. The male titles that were grouped with the female titles included Capt., Col., Major, and Master. These titles are notable for being those of minors and military ranks.

# Evaluation

The tree model that was created used recursive partitioning, which allowed for the data to be divided based on identification of the decision nodes. Some methods for pruning the tree can be more reliable but much more involved, such as that described by Iorio, et al., (2019) where the length of the tree branch highlights the poorer fit variables, making tree pruning a visual task. Traditional pruning methods for the purposes of creating a model from data that has relatively few variables and can quickly assess the outcome is sufficient.

As noted in an article by Hoare (2020), aside from not pruning the tree there are two notable approaches to pruning- with the objective of having minimum error or the smallest tree. Having the smallest tree, the author notes, would be at the cost of an increase in error, albeit typically a small increase. The benefit to aiming for the smallest tree is the ease of interpretability.

Given the small data set, with few variables, an approach towards minimum error was taken. Within this approach is the concept of early stopping, which prevents the splintering of data into sample sizes that are too small to be valuable. For the Titanic data, the removal of variables that had little impact and were used only for the final divisions into very small samples were removed from the overall model, such as port of embarkation.

For other applications of this method, trees for any data set that contains a manageable number of variables can be manually pruned. Evaluating the error rate with different combinations of variables as well as identifying those whose samples contain a very small percentage of the data can be easily seen using the Rattle package in R. Statistics on the models including error rates and reliability can be seen nearly instantaneously with the various permutations of variable combinations.

# Model

The CART model was created using a bottom-up approach. To determine the variables that were relevant to the research question of “What factors influence survival of the Titanic”, multiple factors were evaluated and in different combinations. The result was the final model seen in Figure 1, which had an overall error rate of 16.8%.

**Figure 1.**

*Decision Tree for Survival of the Sinking of the Titanic*

Diagram

Description automatically generated

The decision tree shows the title as the first node, with 62% having the title of Don, Dr., Jonkheer, Mr. or Rev. Survival of an individual with one of these titles would then be more likely for those who had a Fare/Class Factor of greater than 16, were under the age of 52 years, and then had a Fare/Class Factor less than 31. This accounts for less than 4% of the passengers who had one of the titles in node 1, with 96% with one of the titles likely not surviving.

For individuals who did not have one of those titles from the first node, passenger class became the next most relevant factor. Those who were in first or second class, or upper or middle class socioeconomically, survival was highly likely. Third class passengers were then triaged by total fare, with those whose fare was greater than or equal to 23 likely not surviving. For those whose fare was less than 23, the fare per person was considered, with those whose fare per person was 8.1 or greater having a lower chance of survival than those whose fare was less than 8.1 per person.

# Results

The error matrices for this decision tree are noted in Table 2. The number of True Negatives were 95 with a rate of 87.2% and True Positives 53 with a rate of 76.8%. The false positive rate was 12.8%, the accuracy rate 83.15%, and the overall error rate of 16.8%. The precision of this model is 79.1%.

**Table 2.**

*Error matrix for the Decision Tree model on train.csv [validate] (counts):*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Predicted | |  |
| Actual |  | 0 | 1 | Error |
| 0 | 95 | 14 | 12.8 |
| 1 | 16 | 53 | 23.2 |

*Error matrix for the Decision Tree model on train.csv [validate] (proportions):*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Predicted | |  |
| Actual |  | 0 | 1 | Error |
| 0 | 53.4 | 7.9 | 12.8 |
| 1 | 9.0 | 29.8 | 23.2 |

Overall error: 16.8%, Averaged class error: 18%

A Receiver Operating Characteristic (ROC) curve was also created to identify the performance of this model, as seen in Figure 2. The curve shows the true positive and false positive rates. The closer the Area Under the Curve (AUC) is to 1 the better performing the model. The AUC is calculated to be 0.86. This indicates the model is good.

**Figure 2.**

*ROC Curve for Decision Tree*

Chart, line chart

Description automatically generated

A second set of data was provided from the same passenger manifest that included the same variables but did not indicate whether the individuals survived or not. This test data set was used to manually predict based on the model whether an individual survived or not. Filtering the different fields in an Excel spreadsheet of the data along the same parameters created in the decision tree, survival rates were identified and compared to known data for the Titanic from Salem Media (2018). The comparison of these results can be seen in Table 3 below.

**Table 3.**

*Comparison of Predicted Survival to Known Survival Rates by Passenger Class*

|  |  |  |
| --- | --- | --- |
|  | Predicted | Actual |
| First Class Passenger | 71% | 61% |
| Second Class Passenger | 34% | 42% |
| Third Class Passenger | 5% | 24% |
| Total Survival Rate | 28% | 37% |

The predictions for survival based on the model underestimated the third-class passenger survival rate significantly. The first-class passenger and second-class passenger survival rates were only slightly off, as well as the overall survival rate for passengers. This manual validation of the model highlights potential areas for refinement but also indicate satisfactory results.

# Limitations

There are several limitations noted for this model. It is tailored to the specific variables provided in the passenger manifest for the Titanic and cannot be easily applied to other shipwrecks or situations. This may be a benefit in identifying the differences in factors by following a similar process for other events and comparing which variables differ significantly.

This model also has an accuracy rate of 83.15% and an error rate of 16.8%. This indicates that there are likely other factors that have not yet been accounted for that may increase the accuracy of prediction. Further analysis of other information would need to be explored.

There is an underlying bias to be considered when evaluating these variables. A consideration when selecting variables for the preliminary testing is class. Identifying those of wealthier socioeconomic status as having more privilege to board a lifeboat can be a bias influencing the use of associated variables, possibly at the expense of other variables. Gender is also a factor that can be biased based on the stereotyped expectation of “women and children first” for getting to safety.

There also may be confounds with the variables identified as being significant. Physical abilities may impact a person’s socioeconomic status and reduce their ability to leave the ship promptly. The location of a cabin closer to a stairwell may increase the chances of survival but may also be more expensive rooms. This was not able to be evaluated as the data for cabin numbers was incomplete. Other confounds may exist and not thoroughly examined.

# Conclusion

The CART model created using data from the passenger manifest of the Titanic is able to predict survival based on the variables of title, age, passenger class, fare per passenger, and a factor of fare by passenger class with good reliability. Separating and evaluating different variables led to the creation of a decision tree that can be applied to predict survival of this event that can be validated with historical data. By pruning the decision tree, the interaction of variables became apparent and was able to be integrated as new variables.

The model that was created, along with the process to create it, can assist those who wish to take a bottom-up approach to determining significance of variables on a binary outcome. The use of R and Rattle to do so facilitated exchanging variables to quickly make new models and prune the tree. Methodical tree pruning and evaluation along the way led to a more effective model.

Comparing various shipwrecks with accuracy may lend to better interventions to reduce casualties in such events. An example of this is the article by Zielinski (2010), which compares the factors for survival from the Titanic to the sinking of the Lusitania. Being able to accurately identify which factors were most influential for survival of these events provides the opportunity to note the differences and what leads to the differences, such as dissemination of information around the emergency status, or the belief there is ample time to secure valuables or other possessions.

**References**

History.com Editors. (2022, April 12). *Titanic sinks*. History.Com. Retrieved August 5, 2022, from https://www.history.com/this-day-in-history/titanic-sinks

Hoare, J. (2020, November 19). *Machine Learning: Pruning Decision Trees*. Displayr.Com. Retrieved August 5, 2022, from https://www.displayr.com/machine-learning-pruning-decision-trees/

Iorio, C., Aria, M., D’Ambrosio, A., & Siciliano, R. (2019). Informative trees by visual pruning. *Expert Systems with Applications*, *127*, 228–240. https://www.sciencedirect.com/science/article/pii/S0957417419301800

Krueger, E. (2022, January 6). *Build Better Decision Trees with Pruning - Towards Data Science*. TowardDataScience.Com. Retrieved August 4, 2022, from https://towardsdatascience.com/build-better-decision-trees-with-pruning-8f467e73b107

Salem Media. (2018, May 15). *How Many People Were on the Titanic?* HistoryontheNet.Com. Retrieved August 5, 2022, from https://www.historyonthenet.com/how-many-people-were-on-the-titanic

Zielinski, S. (2010, March 2). *Titanic vs. Lusitania: Who Survived and Why?* Smithsonian Magazine. Retrieved August 5, 2022, from https://www.smithsonianmag.com/science-nature/titanic-vs-lusitania-who-survived-and-why-24622866/