Final Project:

What predictors most influence whether or not someone would have survived the sinking of the Titanic?

Helen Nguyen-Quach

Southern New Hampshire University

DAT 520: Decision Methods and Modeling

Professor Jason Eborn

October 23, 2022

Abstract

This research paper aims to identify variables that can lead to survivability of the Titanic. RStudio and machine learning algorithm was specifically used to create a decision tree to identify such variables. The dataset was extracted from Kaggle.com to use to create the top-down decision tree. It was concluded that variables such as gender and age was the determining factor for survivability on the Titanic.

Table of Contents

Introduction

We all know about the love story of Jack and Rose on the Titanic and how Jack could've survived if only Rose knew how to share. James Cameron, the renowned director, reminded the world of the tragedy of the Titanic as he invites the audience to sense the deafening silence as a the boat the size of three football fields with nine decks (Tousignant, 2012) splits into two as it sinks in the middle of the ocean. Cameron not only invited us to witness the chaos that happened that night in 1912, he also presented the inequality that occurred among the ticket class as 3rd class passengers were treated as inferior in everything including survivability.

Differential effects exist everywhere with the Titanic disaster being an example of such. At the time, the Titanic was considered the safest and "largest moving, human-made object in the world" (Main, 2013). Unfortunately, on April 14, 1912 the Titanic hit an iceberg four days after leaving Southampton, England and sank into the middle of the Atlantic Ocean. 68% (1,500 people) of the total number of people onboard died on the Titanic (Riley, n.d.). This leaves 706 survivors.

Research Question

In this research project, we would like to determine the predictors that could influence whether or not someone would have survived the sinking of the Titanic. Due to social norms as the time, safety was prioritized to woman and children. Therefore, we can assume the survivability of the Titanic was disproportionately dependent on gender and age. Nevertheless, the project is to validate this claim.

Data Appraisal

The data for this project was retrieved from Kaggle, a website that provides code and dataset for data science work and practices (Kaggle, n.d.). We will specifically look at the dataset that was collected from passengers of the Titanic.

A detailed description of the variables in the dataset is listed below:

- PassengerID: Passenger identification number
- Survived: 0 = did not survived, 1= survived
- Pclass (ticket class): 1 = 1st, 2 = 2nd, 3 = 3rd
- Name: Passenger's name
- Sex: male or female
- Age: in years
- Sibsp: number of siblings/spouses aboard the Titanic
- Parch: number of parents/children aboard the Titanic
- Ticket: Ticket number
- Fare: Passenger fare cost
- Cabin: Cabin number
- Embarked: Port of Embarkation – C= Cherbourg, Q= Queenstown, S=Southamptom

A snapshot of the variables collected on 6 passengers can be seen in Figure 1.

Figure 1.

```
> head(titanic_df)
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch
1           1        0      3                              Braund, Mr. Owen Harris   male  22     1     0
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0
4           4        1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0
6           6        0      3                                     Moran, Mr. James   male  NA     0     0
            Ticket     Fare Cabin Embarked
1        A/5 21171  7.2500  <NA>        S
2         PC 17599 71.2833   C85        C
3 STON/O2. 3101282  7.9250  <NA>        S
4           113803 53.1000  C123        S
5           373450  8.0500  <NA>        S
6           330877  8.4583  <NA>        Q
```

Technique

Data Preparation

The utility that we will be using for the project will be R Studio with libraries such as "dplyr," "rpart," and "rpart.plot." Certain steps were used to prepare the data. We selected categories that are relevant to project and dropped the other variables. Variables that were selected includes whether or not they survived, their ticket class ("Pclass"), gender, age, number of siblings or spouse on board ("SibSp"), and number or parents or children on board ("Parch"). Records that had "NA" were also dropped from our data. The codes that were used can be seen in Figure 2.

Figure 2.

```
> titanic_df=select(titanic_df, Survived, Pclass, Age, Sex, SibSp, Parch)
> titanic_df<- na.omit(titanic_df)
```

Next, we checked the structure of each variables of the dataset to ensure integrity. In Figure 3, we can see that the structure of variables "Survived" and "Pclass" are integers when they should be categorical. We then changed the structure of "Survive" and "Pclass" as categorical. We also classified "Pclass" as an ordinal variable with values of 3 being the least and 1 being the greatest.

Figure 3.

```
> str(titanic_df)
'data.frame':   714 obs. of  6 variables:
 $ Survived: int  0 1 1 1 0 0 0 1 1 1 ...
 $ Pclass  : int  3 1 3 1 3 1 3 3 2 3 ...
 $ Age     : num  22 38 26 35 35 54 2 27 14 4 ...
 $ Sex     : chr  "male" "female" "female" "female" ...
 $ SibSp   : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch   : int  0 0 0 0 0 0 1 2 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
  ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
> titanic_df$Survived = factor(titanic_df$Survived)
> titanic_df$Pclass = factor(titanic_df$Pclass, order =TRUE, levels = c(3,2,1))
```

Data Manipulation

A top-down decision tree was selected for this research project. Before we can run the data into our program to create a decision tree, we must split the data into training and testing sets. Figure 4 shows the code on how we split the data. Figure 5 shows how we created the train and test set with 80% of the data using to train.

Figure 4.

```
> train_test_split = function(data, fraction = 0.8, train = TRUE) {
+     total_rows = nrow(data)
+     train_rows = fraction * total_rows
+     sample = 1:train_rows
+     if (train == TRUE) {
+         return (data[sample, ])
+     } else {
+         return (data[-sample, ])
+     }
+ }
```

Figure 5.

```
> train <- train_test_split(titanic_df, 0.8, train = TRUE)
> test <- train_test_split(titanic_df, 0.8, train = FALSE)
```

Models and Checkpoints

To test for validity, a confusion matrix was created. Confusion tables categorize predictions and actual values (Tyagi, 2020). The categories include:

- True positive (TP): correctly classified as the class of interest
- True negative (TN): correctly classified as not a class of interest
- False positive (FP): incorrectly classified as the class of interest
- False negative (FN): incorrectly classified as not the class of interest

Accuracy will also be used by calculating (TP + TN)/(TP + TN + FP + FN).

Ethicality and Legality

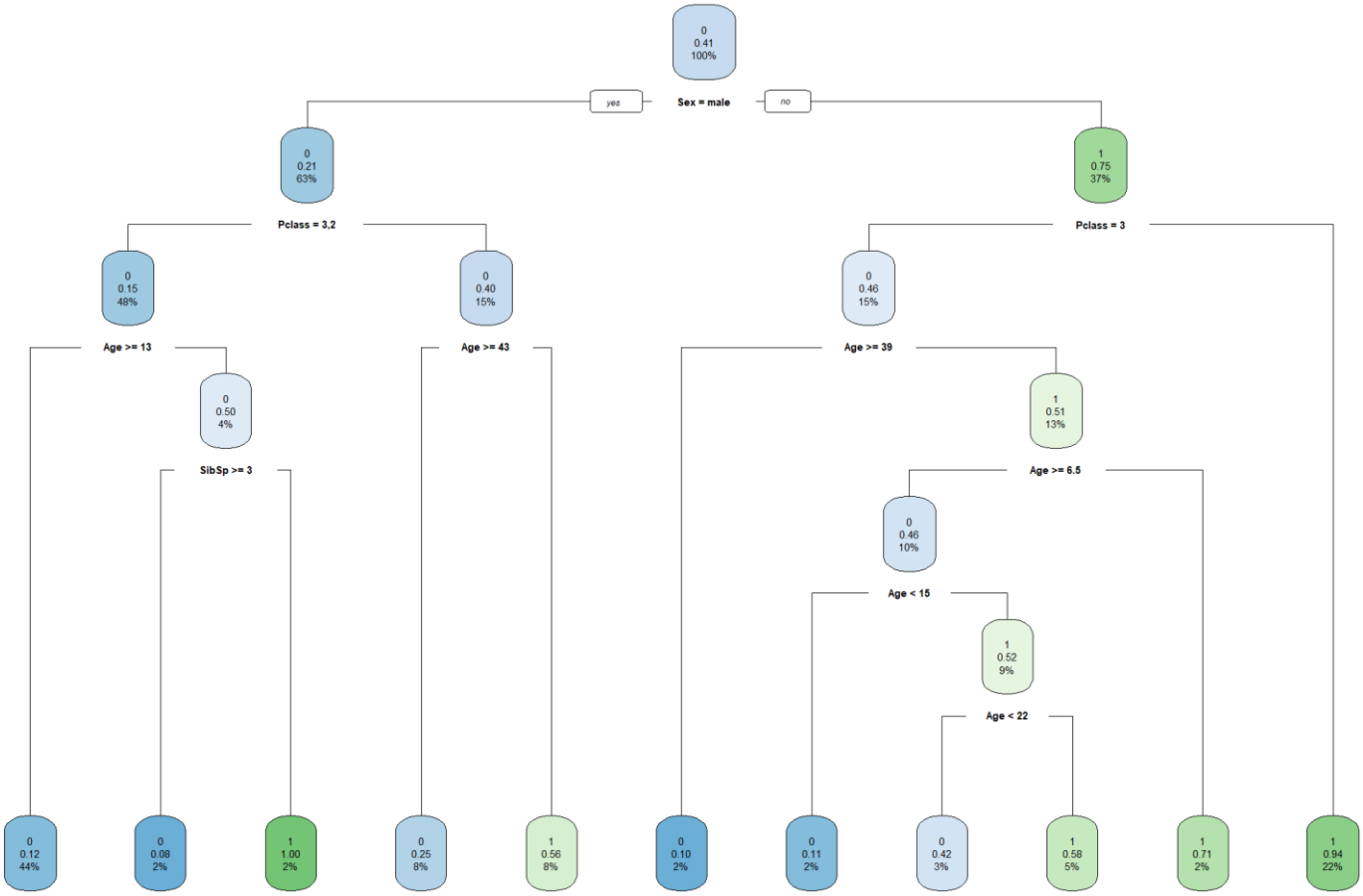
Measures were taken to ensure privacy. The names of the passengers on the Titanic were redacted. Other variables that can be used to identify a passenger were also dropped. The only remaining variables include passenger class, age, sex, number of siblings and number of parents.

Evaluation

RStudio with library “dplyr,” “rpart,” and “rpart.plot” was used in this project due to ease of use and accessibility. A decision tree can easily be made with this program.

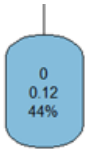
Decision Tree Model

The top-down decision tree can be seen below.



The top/root node of the tree showed the proportion of passengers 100% and the probability of survival is 41%. We can ignore the very top number (0,1) until the end. The first split, asks if the passenger is male or female. If they are then we go to the left branch. The first node on the left showed that there are 63% male and there was 21% probability of survival. We then split the tree to pclass, age, and siblings. In the first left branch in the final node as seen in Figure 3, we can see that 44% of the passengers are male, older than 13 and is classified as pclass 2 or 3. They had a 12% of survival.

Figure 3.



Overall, females had a higher survival rate with 75%. The highest survival rate (94%) was female of 1st or 2nd pclass. 22% of the passengers are in this category. 15% of passengers were female of 3rd class and their survival rate was 46%. Second or third class male passenger less than thirteen years of age, with three or more sibling with them, had the least chance of survival at 8%.

After training the model, we created a confusion matrix. A confusion matrix shows a table of actual and predicted values (Tyagi, 2020). Figure 4 shows the confusion matrix values and Figure 5 showed the predicted labels.

Figure 4.

```
> table
   predicted
     0  1
  0 71 16
  1 11 45
```

Figure 5.

True Negative (TN)	False Positive (FP)
False Negative (FN)	True Positive (TP)

A true negative predicted a negative outcome and was correct (71 cases). A false negative incorrectly predicted a negative (11 cases). A false positive incorrectly predicted a positive (16 cases). A true positive correctly predicted a positive (45 cases). Accuracy can be calculated by (TP + TN)/(TP +FP + FN + TN). Sensitivity can be calculated by (TP)/(TP + FN). Specificity can be calculated by (TN)/(TN + FP). The accuracy test showed 81%. The sensitivity test showed 80%. The specificity test showed 82%.

Results

As we can see from our top-down decision tree, the most important variable was whether the passenger was male or female, then which type of class they were in, their age, and finally the number of siblings they had onboard. 75% of woman survived the Titanic while they only made up 37% of the passenger. Males made up 63% of the passengers but only 21% survived. In the table below we can see the percentage of top three of those that survived and their passenger make up.

	Percentage that Survived	Make-up of passengers
Women of first and second class	94%	22%
Girls of third class less than 6.5 age	71%	2%
Women of 3rd class, greater than 22 years of age	58%	5%

Limitations

The weakness of this decision tree includes the variation in classifying age. The first branch had the separation at 13 years of age whereas the second brand had it at 43 years of age. The data would've been stronger if the age of separation was set at a certain number.

Another limitation with our data was how the passengers survived. The dataset didn't include if the passengers survived by having a seat on the lifeboats or by chance in the water. As described by a news piece by BBC, Fang Lang, a Chinese survivor of the Titanic clung to a wooden door because he wasn't able to escape on a lifeboat (BBC, 2021).

A variable that could be added in the dataset is distance from the lifeboats. It could be helpful to know if the survivability of the first class passengers were due to their accessibility or proximity to the lifeboats or was it because space was only made for the first class passengers. Were the lifeboats in close proximity to first class passengers or were their lifeboats close the 3rd class passengers?

Conclusion

The project was used to address predictive survivability based on data from the Titanic. With variables including sex, ticket class ("Pclass"), age, number of siblings or spouse on board ("SibSp"), and number or parents or children on board ("Parch"), the biggest indicator of survivability is gender and class. 94% of female passengers of the first and second class survived the Titanic but they only make up 22% of the passengers onboard. First class ticket male that was older than 43 years of age had a survivability of 56% but only made up 8% of the

passengers. This further proves that Titanic couldn't escape the social class injustices that existed during that time in history.

In retrospect, the second largest group of survivors was girls less than 6.5 years of age of the 3rd class. This group had 71% chance of survival. Children less than 13 years of age that were from the 3rd and 2nd class and had less than three siblings with them also had a highest rate of survival with 100%.

References

Feng, Z., Wang, Y., (2021, April 16). Titanic: Searching for the 'missing' Chinese survivors. BBC News. Retrieved from https://www.bbc.com/news/world-us-canada-56755614

Main, D. (2013, July 10). The Titanic: Facts About the 'Unsinkable' Ship. *Live Science.* Retrieved from https://www.livescience.com/38102-titanic-facts.html

Riley, P. (n.d.). Titanic Timeline and Facts. *Britannica.* Retrieved from https://www.britannica.com/story/titanic-timeline-and-facts.

Titanic – Machine Learning from Disaster (n.d.). Kaggle. Retrieved October 16, 2022, from https://www.kaggle.com/c/titanic

Tousignant, M., (2012, April 13). Kids were onboard the Titanic, too. *The Washington Post.* Retrieved from https://www.washingtonpost.com/lifestyle/kidspost/kids-were-onboard-the-titanic-too/2012/04/12/gIQANwAhFT_story.html

Tyagi, N. (2020, December 21). What is Confusion Matrix? *Analytic Steps.* Retrieved from https://www.analyticssteps.com/blogs/what-confusion-matrix