A Study in Disaster

The Titanic Survivors

mauveSushi

Skidmore College

Abstract

This research aims to construct a logistic regression model to predict the survival outcome of passenger on the Titanic based on their various demographic and socioeconomic factors, such as age, gender, number of spouses and siblings, passenger class, etc. The findings of this research can be generalized to most shipwreck where the ship is considered big (having capacity of at least 1000 passengers) and the route is through Atlantic Ocean in the early 20^{th} century. We used logistic regression in tandem with backward feature selection to obtain several prediction models. We evaluate each obtained model using the Hosmer-Lemeshow goodness-of-fit test and then conduct a more vigorous feature selection process to come up with the best models. Despite the different approaches, all of the obtained models suggest a common trend that children, woman and first class passengers are among the one with highest chance of survival.

Introduction

On that fateful night April 15, 1912, the great Titanic carrying 2,200 people struck an iceberg and sank, brought down with it 1,500 poor souls. Such tragedy never fails to leave all of us unsettled, but as statisticians, we are naturally drawn to ask ourselves the question: *"How do demographic and socioeconomic factors affect chance of survival in disaster?"* Perhaps, based on just the titanic data set, such general question cannot be answered, since our findings are probably just generelizable to a specific type of disaster, which is shipwreck and to a time-frame limited to the early 20th century. Also, we have to consider the fact that Titanic is not just another ship... It is (even till today) one of the most luxurious ship ever been built (see figure below - the ship even had swimming pool and tennis court at the lowest level!). As such, we think the findings we obtain using this data set can only be generalized to big ship, i.e. ships whose capacity is at least 1000, crossing the Pacific ocean. Note that this number 1000 seems rather arbitrary, but we found that nowadays, on average, cruise ships with tonnage of around 40,000 usually have capacity of 1,000 passengers[1] and most of the bigger ships in early 20th century were around 40,000 or more in tonnage[2]. For this reason, we formulate our research question as follows:

"How do various demographic and socioeconomic factors affect chance of survival in shipwreck where the ship's capacity is at least 1000 passegners and the route is through the Atlantic Ocean in the early 20th century?"

In simpler words, we will attempt to create a model to be able to predict the survival outcome of a passenger, given his/her relevant background information. We constructed several different models with 2 major approaches, one using backward selection and one using a more rigorous feature selection process. Despite this, the models we obtain show a common trend that female, little kids and first-class passengers are more likely to survive the shipwreck. Following this, we will step by step present our data preparation and modelling process, as well as our findings.

[1] CruiserMapper, Cruise Ship Passenger Capacity
[2] Wikipedia, List of largest passenger ships

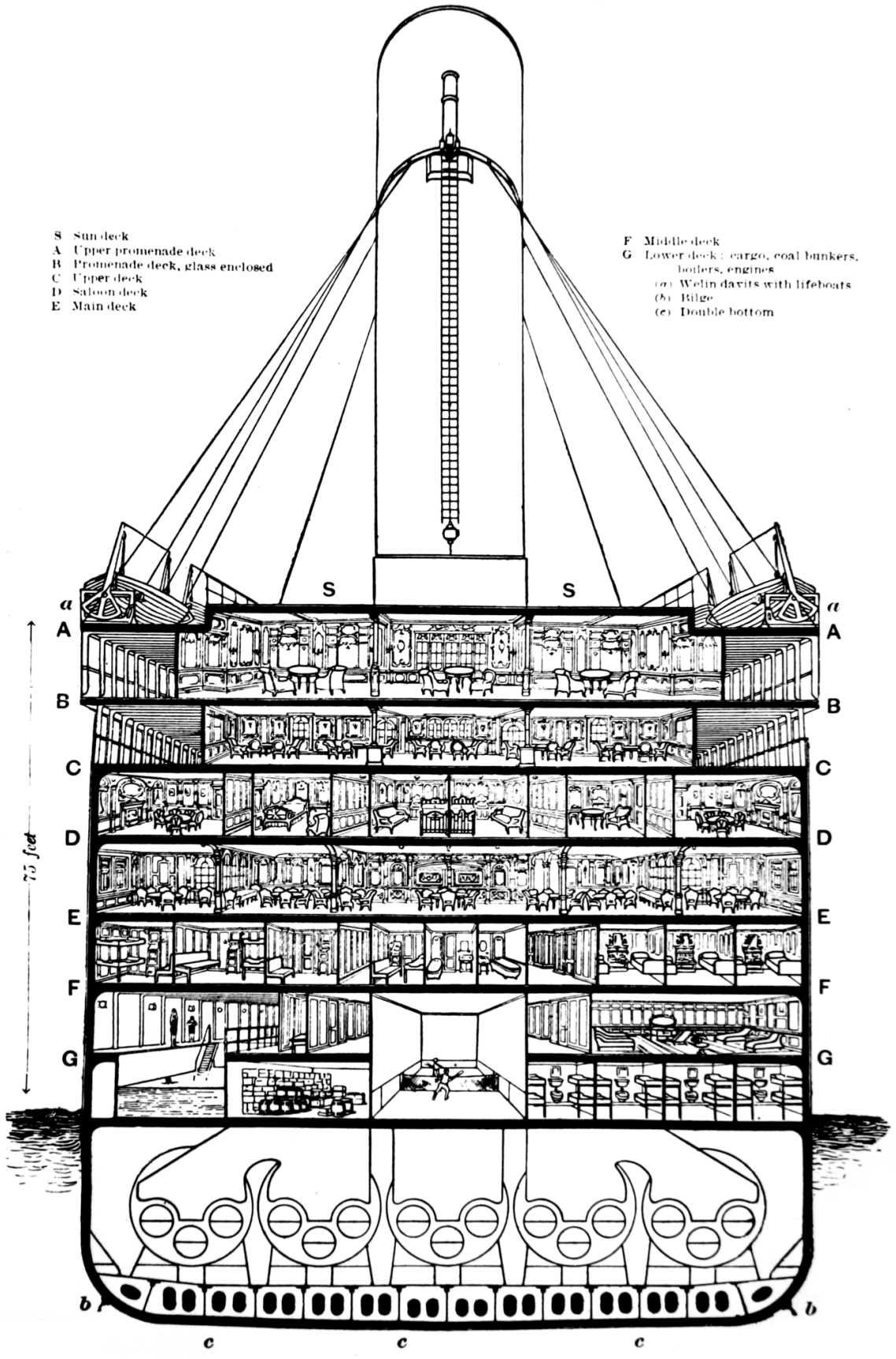

Figure 1: Cross Section of the Titanic

1. Data Exploration

Placeholder

1.1 Summary

1.2 Variable Selection

1.3 Data Preparation

1.4 Final Data Set Summary

1.5 Correlation Investigation

2. Methodology

Placeholder

2.1 Backward Selection

2.2 Logistic Regression

2.3 Homser-Lemeshow goodness-of-fit test

2.4 Akaike Information Criterion (AIC)

3. Model

Placeholder

3.1 Model obtained using backward selection

3.2 The Final Models

4. Conclusion

Placeholder

4.1 Findings

4.2 Further Study

Bibliography