Survival Analysis On Titanic

– Manav Bhagat, Aayush Mainali, and Shreehar Joshi

Introduction

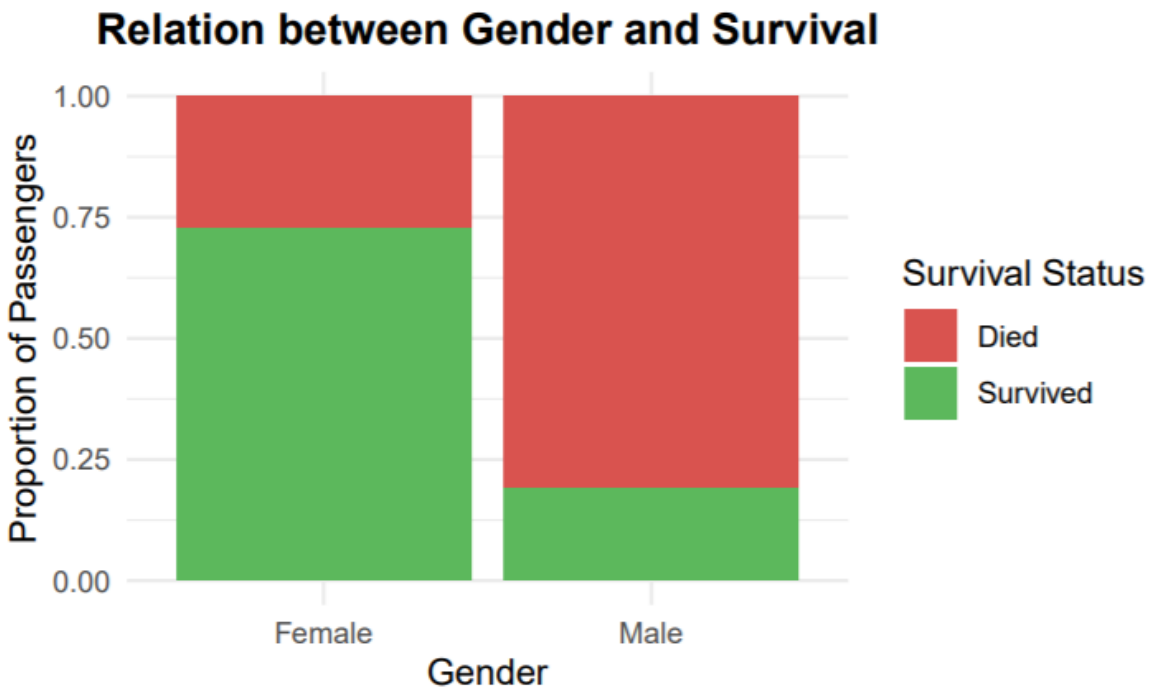
- Titanic, in full Royal Mail Ship (RMS) Titanic, was a British Passenger Luxury Liner
- Sank on April 14-15, 1912 during its maiden voyage en route to New York City from Southampton, England
- Only 705 passengers survived leaving over 1500 dead.
- We hypothesize that besides luck there were other factors that increased chances of survival of passengers.

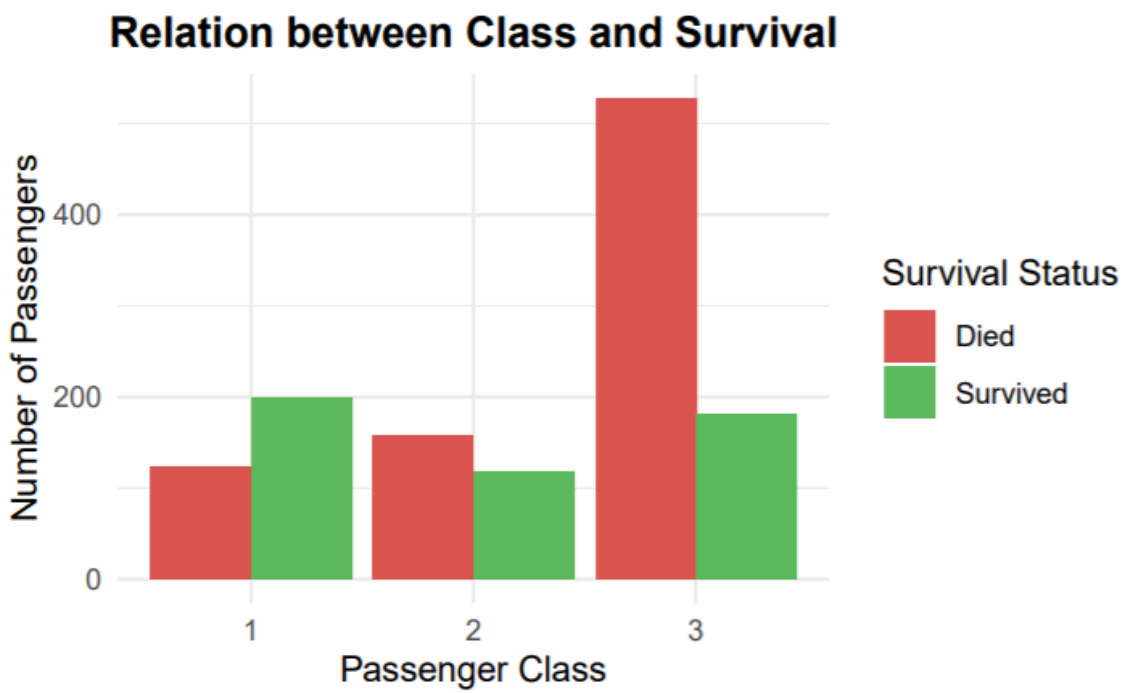
Overview of Data

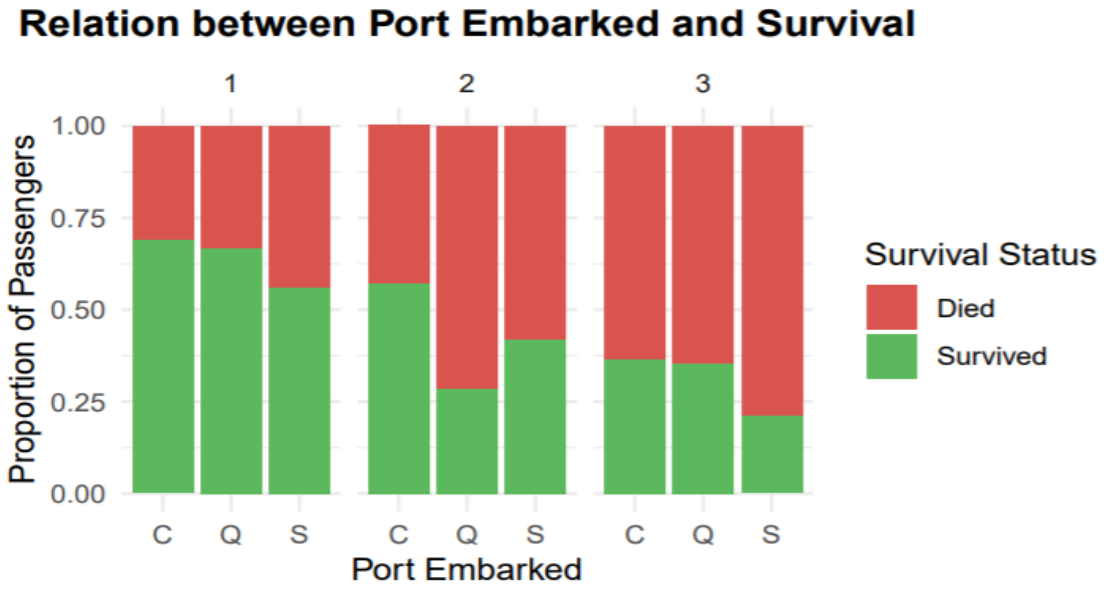
- Dataset created by Noah Rippner
- Retrieved from Data World
- 1310 observations of the passengers in Titanic
- 14 columns

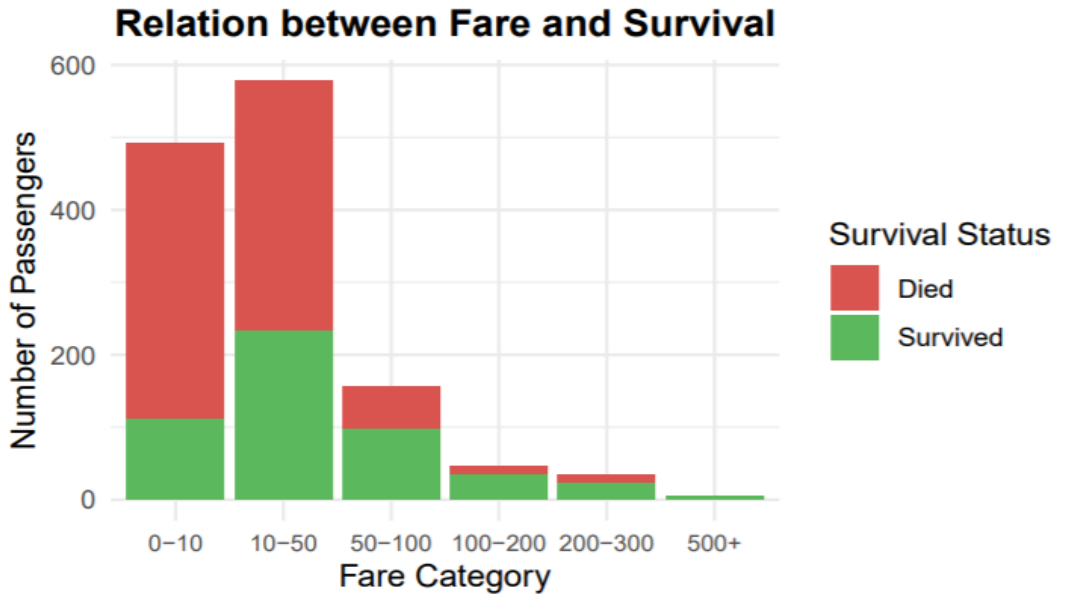
pclass: The class on which the passenger was traveling.
survived: The survival status of the passenger.
name: The name of the passenger.
sex: The sex of the passenger.
age: The age of the passenger.
sibsp: The number of siblings/spouses the passenger had aboard.
parch: The number of parents/children the passenger had aboard.
ticket: The ticket number of the passenger.
fare: The price of the ticket of the passenger.
cabin: The cabin address of the passenger.
embarked: The port from which the passenger had embarked.
boat: The lifeboat number of the passenger (if survived).
body: The body number of the passenger (if not survived and body recovered).
home.dest: The home/destination of the passenger.

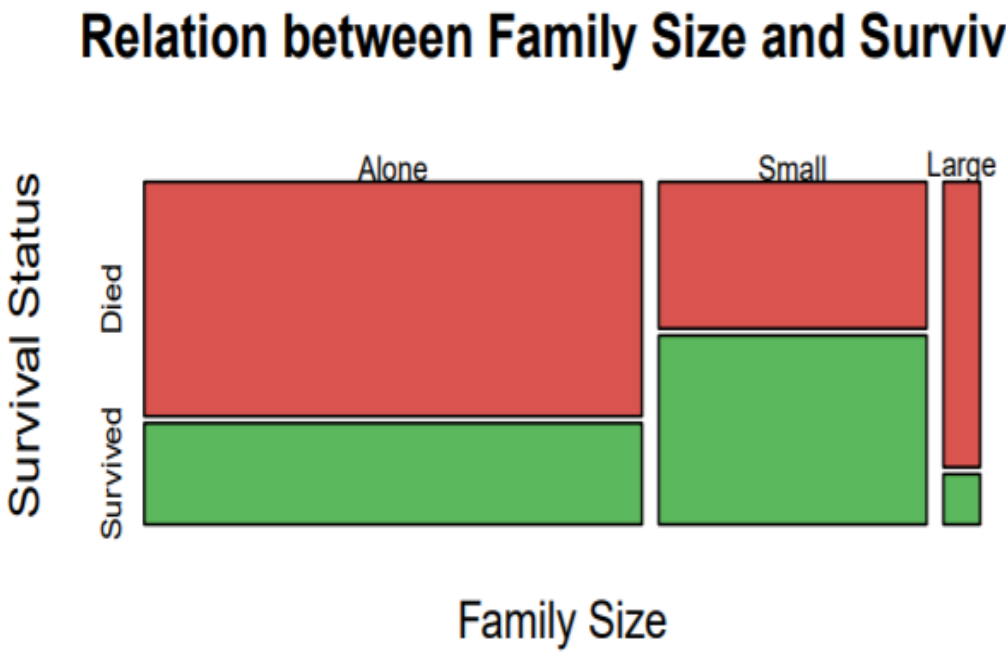
Highlights From EDA

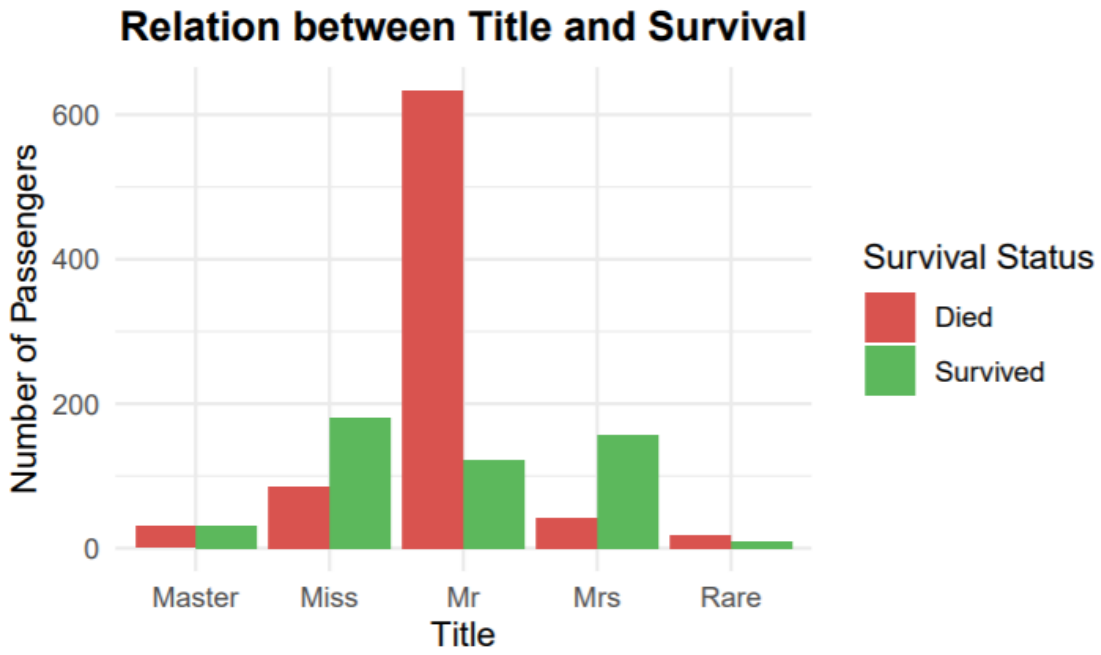





Relation between Port Embarked and Survival
1
2
3
1.00
0.75
0.50
0.25
0.00
Proportion of Passengers
C
Q
S
Port Embarked
Survival Status
Died
Survived

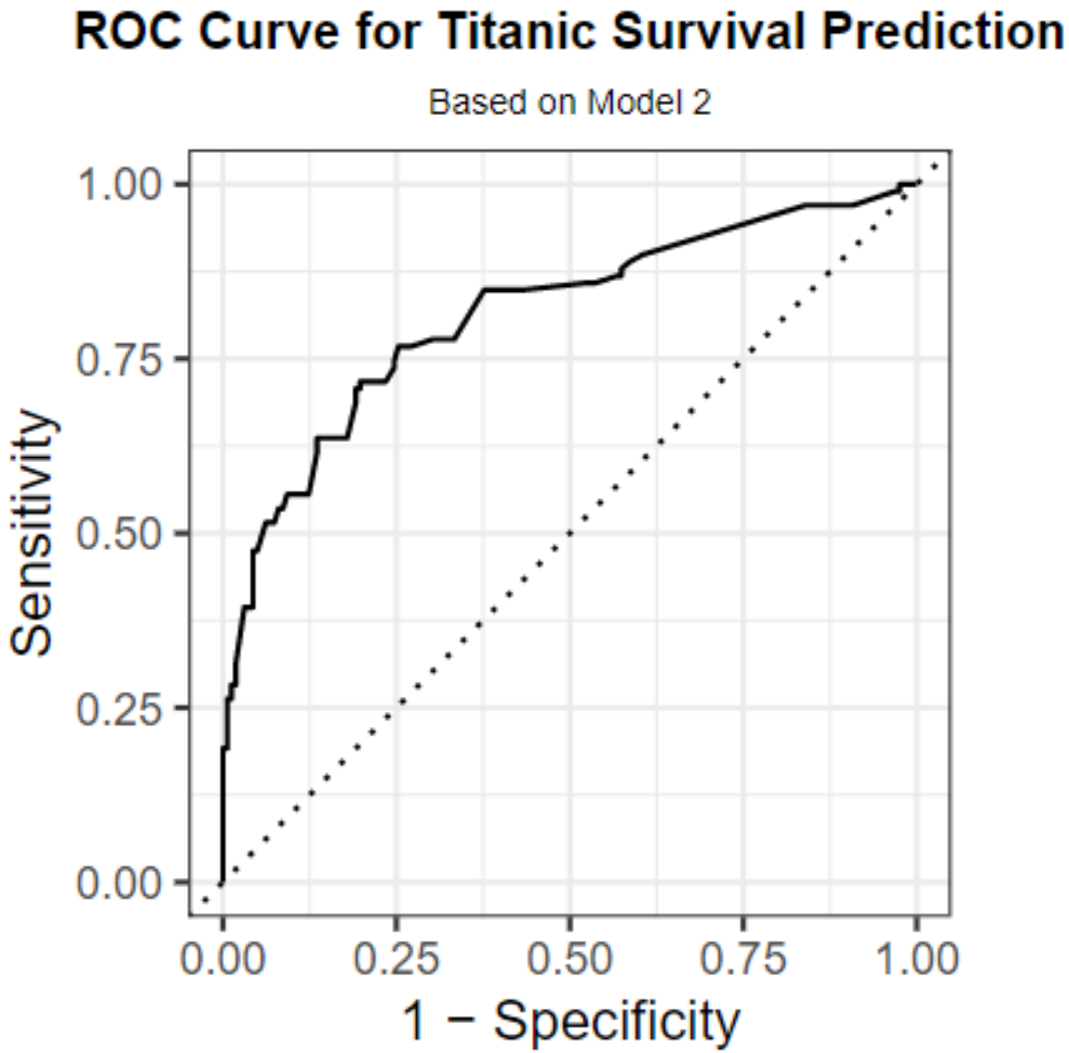

Relation between Fare and Survival
600
400
200
0
Number of Passengers
0–10
10–50
50–100
100–200
200–300
500+
Fare Category
Survival Status
Died
Survived


Relation between Family Size and Survival
Survival Status
Died
Survived
Alone
Small
Large
Family Size


Relation between Title and Survival
600
400
200
0
Number of Passengers
Master
Miss
Mr
Mrs
Rare
Title
Survival Status
Died
Survived

Modeling

- Three models devised.
- Model 1 consisted of pre-existing variables only as predictor variables.
- Model 2 and 3 consisted of mixture of pre-existing and feature engineered variables like family size, categorical fare and title as predictor variables.
- Model 3 was selected based on area under the curve for prediction on test set.
- Model 3 had area under the curve of 0.813 on test set.
- A cutoff of 0.5 was found to be the most suitable for further analysis.



Conclusion/Future Work

- Our hypothesis was correct. There were indeed certain features that increased the likelihood of passengers to survive.
- Females, children, class 1 and high fared passengers, passengers with small family members and passengers who embarked from Cherbourg were more likely at survive.
- Our model with feature engineered and pre-existing variable including port as the predictor variables showed the highest efficiency.
- For our future work we could add other features like age, home/destination of the passengers, cabin, etc. that we had completely ignored in our analysis.
- Other classifiers like Random Forest, Extra Trees, and Support Vector Machines could be used to compare the efficiency.