# Overview of Data

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.  During the ship’s maiden voyage, the Titanic sank after colliding with an iceberg resulting in the death of 1,502 passengers and crew out of the 2,224 people on board. One of the major findings from this event was a lack of proper safety considerations to account for such a large-scale event, mainly the number of lifeboats on the ship. Our goal is to review this information and determine through binary classification if we can determine survival of a passenger knowing the values for certain variables. While the data is fairly straight forward some cleansing and imputation was required in order to conduct an appropriate binary classification. More details associated with data cleansing and imputation will be discussed along with derived variables and their purpose in this analysis.

# Exploratory Data Analysis

While reviewing the provided Train and Test data sets it became clear that there were multiple observations lacking a value for Age. The only other variable with a missing value was Fare, that was isolated to the provided Test set. Figure 1 below provides a visualization of the null values among the data for both the Train and Test data sets that were reviewed.

|  |
| --- |
|  |
| ***Figure 1: Null Values Within Data Sets*** |

Given that Age, or variables closely correlated to age, are likely meaningful towards determining survival, imputation was conducted on the column using the mice package in R. Multivariate Imputation by Chained Equations (MICE) contains a function mice() which allows for the imputation of missing data *m* times. Within the imputation we chose to utilize the random forest (rf) method. We were unable to identify any sources of record indicating why Age values were missing within the data, our assumption is these values were missing at random (MAR) and did not depend on any missing variables. Given the MAR assumption for Age we will modify the full data using the imputed model of the observed data. Figure 2 below illustrates density within the original data for Age and density from the imputation of Age. The resulting density distributions for both raw and imputed values have similar trends and led to high levels of confidence for their use in the binary classification exercise. The provided Train and Test data were combined into a full data set so more of the known values could be leveraged in the imputation. Given that the data were split for us we felt it was prudent to incorporate as many samples as possible to generate values for Age where it was not explicitly present. Going forward we will assume the resulting values from imputation are plausible and valid for the binary classification.

|  |
| --- |
|  |
| ***Figure 2: Imputation for Age*** |

The only other raw data column lacking observations was Fare within the Test data set. While submitting our resulting model to Kaggel is a secondary exercise, we still took time to determine what value for Fare was most appropriate given the other known information for that observation. While many variables are available, the location from which a passenger Embarked and the class of the passenger’s ticket (Pclass) had the most practical significance.

Figure 3 below contains a density plot of the Fares for all passengers that Embarked from Southampton (Embarked = S) and were in third class grouping (Pclass = 3). The results skew heavily to the right, with most of the observed values falling closer towards approximately $8. Given the distribution of the fares for those conditions a median value of $8.05 for Fare was used for the missing observation.

|  |
| --- |
|  |
| ***Figure 3: Replacement for Fare*** |

# Derived Variables

In addition to imputation, we also considered age bins as a categorical variable. We noted that the proportion of people who survived varies mostly for very young children and seniors. For the vast majority of people, survival rate is 0.35-0.40. For children, age 6 and under, survival rate is about 0.70. However, survival rates begin to drop at age 60 and get close to zero at age 65. Figure 4 below shows the ratio of survival for equally spaced 10-year age bins next to the ratio of survival for the aforementioned manually created bins: children, age 7 to 63, and those above 63. You will notice a more prominent survival ratio disparity when grouped into fewer bins.

|  |
| --- |
|  |
| ***Figure 4: Survival by Age*** |

Since each passenger has a unique name, some kind of simplification/reduction was deemed necessary in order to extract meaningful information that may be useful for prediction. One such reduction is a passenger’s ‘title’. Title can refer to marital status, as in ‘Miss’ vs ‘Mrs.’, age, as in ‘Mr.’ vs. ‘Master’, or rank, such as ‘Captain’ or ‘Countess’. In addition, French titles were reconciled with their English counterparts, such as ‘Madame’ and ‘Mademoiselle’ to ‘Mrs.’ and ’Miss’. Royalty and military ranks were assigned to an ‘uncommon’ group given how few were present in the analysis. Table 1 below provides a numerical breakdown of how many passengers had these titles in their name.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sex** | **Master** | **Miss** | **Mr** | **Mrs** | **uncommon** |
| **Female** | 0 | 264 | 0 | 198 | 4 |
| **Male** | 61 | 0 | 757 | 0 | 25 |

***Table 1: Resulting Titles from Names***

In Figure 6 below survival rate by the newly derived Title is shown. Interestingly it appears that men and those with uncommon titles generally survived when compared to their female counterparts, this is interesting given the larger count of men on the ship.

|  |
| --- |
|  |
| ***Figure 5: Survival by Title*** |

# Review of Correlation

Before conducting any prediction based on the resulting raw and derived data, we reviewed the impact of any collinearity. From the correlogram in Figure 6 we can observe that the only variables with a strong correlation to each other are SibSp and Parch to the derived variable Family. This is expected given Family is a result of the addition of the other two values.

|  |
| --- |
|  |
| ***Figure 6: Correlogram of Data*** |

# Kaggel Results

Manny: Write up on results from Kaggle and meaning.

"Survived" "Pclass" "Sex" "Age" "SibSp" "Parch" "Fare" "Embarked" "Title" "AgeBin" "Family"

Care was taken not to remove too many data points given that the data set is broad, and the more information removed from the model the harder more extreme Sale Price might be to predict.

# Logistic Regression within Train Data

## Full Model

Stuff

## Reduced Model

Stuff

"Survived" "Pclass" "Sex" "Age" "SibSp" "Parch" "Fare" "Embarked" "Title" "AgeBin" "Family"

# Conclusion

X Model is being used because…

Appendix

# Data Dictionary

The data was provided by Kaggle and corresponds to the passengers who sailed on the titanic.

The data consists of two sets, a training set containing 891 observations and a testing set containing 418 observations. The observations consist of 12 base variables which include information from the size of families to amenities and the quality of many aspects of the cabin on the ship. Important variables that will recur in this analysis include:

1. PassengerID: Unique identifery of passengers on board
2. Survived: Survival; 0 = No, 1 = Yes
3. Pclass: Ticket class; 1 = 1st, 2 = 2nd, 3 = 3rd
4. Name: Passenger’s full name and title
5. Sex: Sex; ‘male’ or ‘female’
6. Age : Age in years
7. Sibsp: # of siblings / spouses aboard the Titanic
8. Parch: # of parents / children aboard the Titanic
9. Ticket: Ticket number
10. Fare: Passenger fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation; C = Cherbourg, Q = Queenstown, S = Southampton

The following variables were derived from the core table:

1. Title: Status or title of the passenger
2. AgeBin: Categories corresponding to passenger’s age
3. Family: The sum of #of siblings/spouses and # of parents/children

(Where did the data come from? How big is it? How many observations? Where can we find out more? What are the specific variables that we need to know to understand with respect to your analysis?)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Variables** | **Description** | **Key** | **Notes** |
| 1 | ID | Unique numeric identifier corresponds to passengers | Numeric:  [1 – 891]:train;  [892 –1309]:test | Unique identifier for passengers |
| 2 | Survived | Notes whether or not a person survived | Numeric; 0 = Did not survive;  1 = Survived |  |
| 3 | Pclass | Class of ticket belonging to passenger | Numeiric: 1 = 1st, 2 = 2nd, 3 = 3rd | Should be categorical |
| 4 | Sex | Sex of passenger | Character; ‘male’ or ‘female’ |  |
| 5 | Age | Age of passenger in years | Numeric | Age of infants less than 1 year old was guessed and reported as a decimal; Ex: 0.75.  [0.17 – 80].  177 NAs in train  86 NAs in test |
| 6 | Sibsp | Number of siblings or spouses aboard the titanic | Integer | [0 -8]; no 6 or 7 |
| 7 | Parch | Number of parents or children aboard the titanic | Integer | [0-6] |
| 8 | Ticket | Ticket number | Character | Many are numbers, some contain descriptions |
| 9 | Fare | Passenger Fare / Ticket Price | Numeric | 1 NA in test  [0 – 512.33]; presumed to be in American dollars |
| 10 | Cabin | Passenger’s Cabin number | Character; usually in the form of *LETTER##,* EX: A55, | Almost Unique for every passenger. One passenger can be in multiple cabins; 827 blanks for train; 327 blanks for test. |
| 11 | Embarked | Port of Embarkation; | Character;  C = Cherbourg,  Q = Queenstown,  S = Southampton | 2 blanks for train |
| 12 | Title | Extracted from name and corresponds to rank or marital status | Character; Ex: Mr., Mrs., Miss, | Some titles were uncommon, such as Countess, Captain; these are placed under ‘uncommon’ |
| 13 | AgeBin | Ages of passengers by increments of 10 years | Character;  (0-10], (10-20],…(80,90], ‘unknown’ | Missing values are given their own category, they are not designated NA |
| 14 | Family | Sum of: number of parents/children and number of siblings/spouses | Numeric |  |

# Correlation Matrix of Train Data





# Kaggel Score

