Predicting Survivability

Using data from the Titanic to build predictive models.

The Titanic data set contains information about 891 passengers that were aboard the Titanic when it sunk in 1911. It includes eleven characteristics such as Sex, Name, Cabin Class, Cabin number, Ticket Number, Fare, how many parents, spouses, and siblings with which the passenger was traveling, Age, and whether the passenger survived the tragedy.

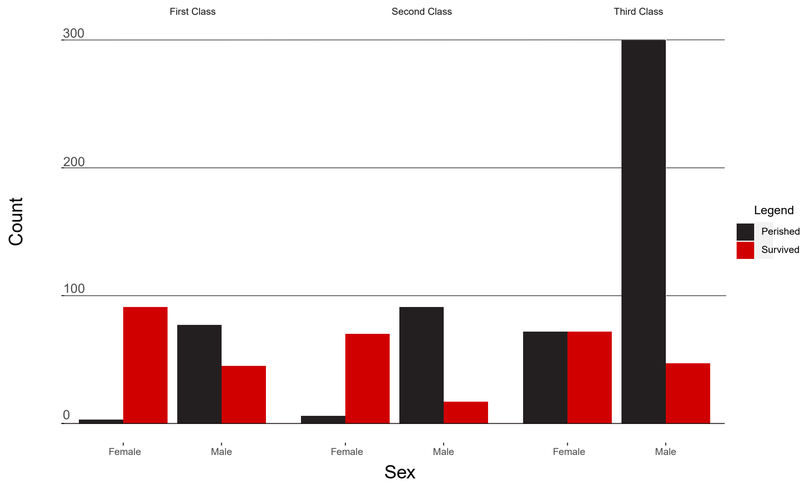
Shape of the Data

Distributions and Correlations

The main purpose of this data set is to teach concepts and techniques in machine learning. There are two data sets: a training set and a testing set. The training set contains the survival status of each passenger, but the testing set omits this variable. The goal is to train machine learning models using the training set and apply them to the testing set to predict survival. This is an excellent project for anyone who is interested in learning some machine learning techniques or anyone who is fascinated by history and predictive statistics.

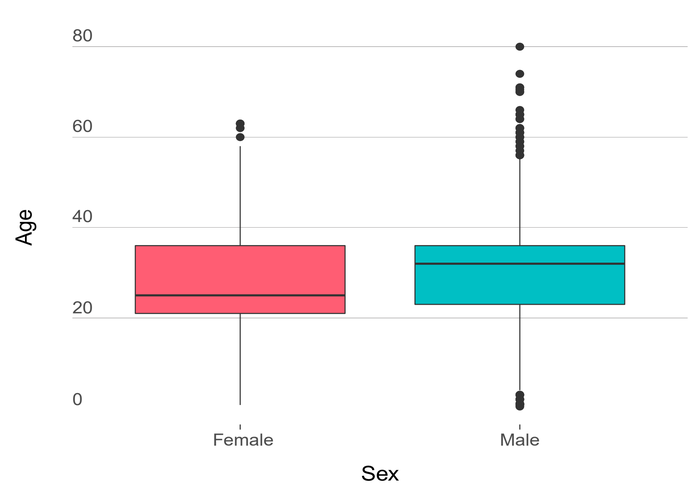
Survival Rates of Males and Females Across Class

The following chart shows the survival numbers between men and women divided by class.



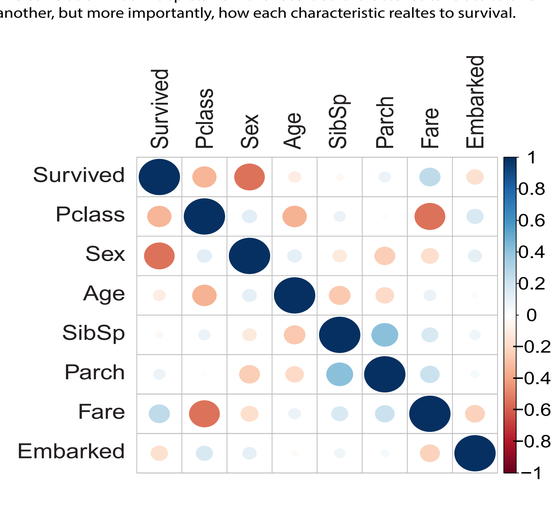
Distribution of Age by Sex

The average age of women on the Titanic was 27.7 years old and the average age of men was 30.8 years old.



Correlation Matrix

The correlation matrix depicts how the recorded characteristics relate to one another, but more importantly, how each characteristic realtes to survival.

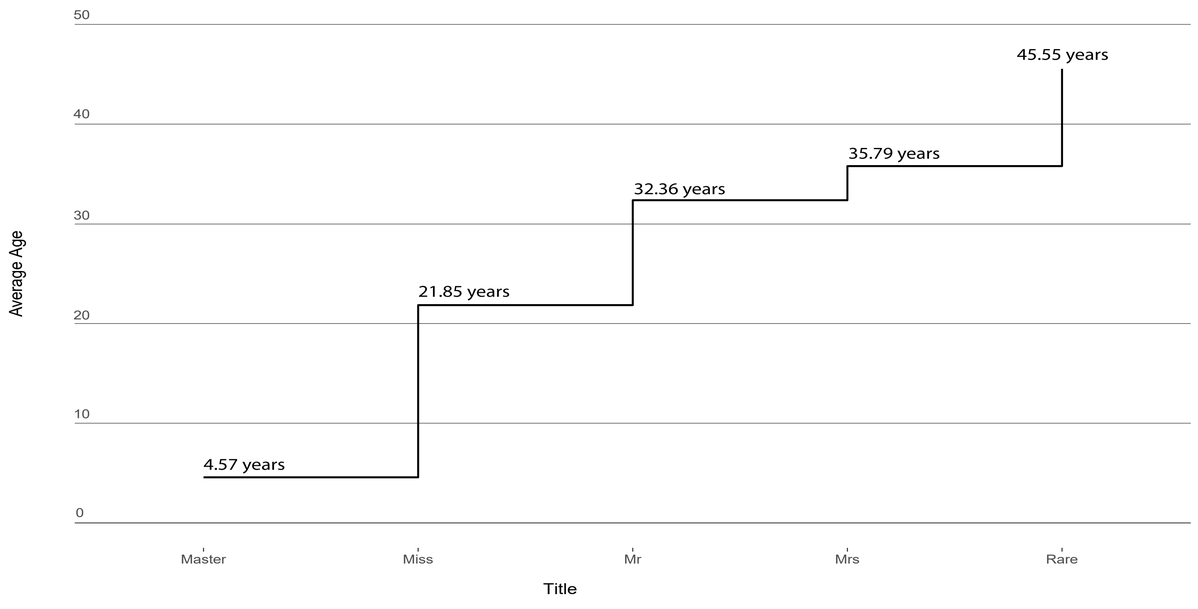


Feature Engineering and Predictive Models

The first question that must be addressed is how does one deal with several missing values in key characteristics such as age?

Average Age by Title

Each passenger's title was extracted from their name and the average age for each title was found and applied to missing values in Age.
'Rare' title indicates military, Clergy, or Nobility titles.

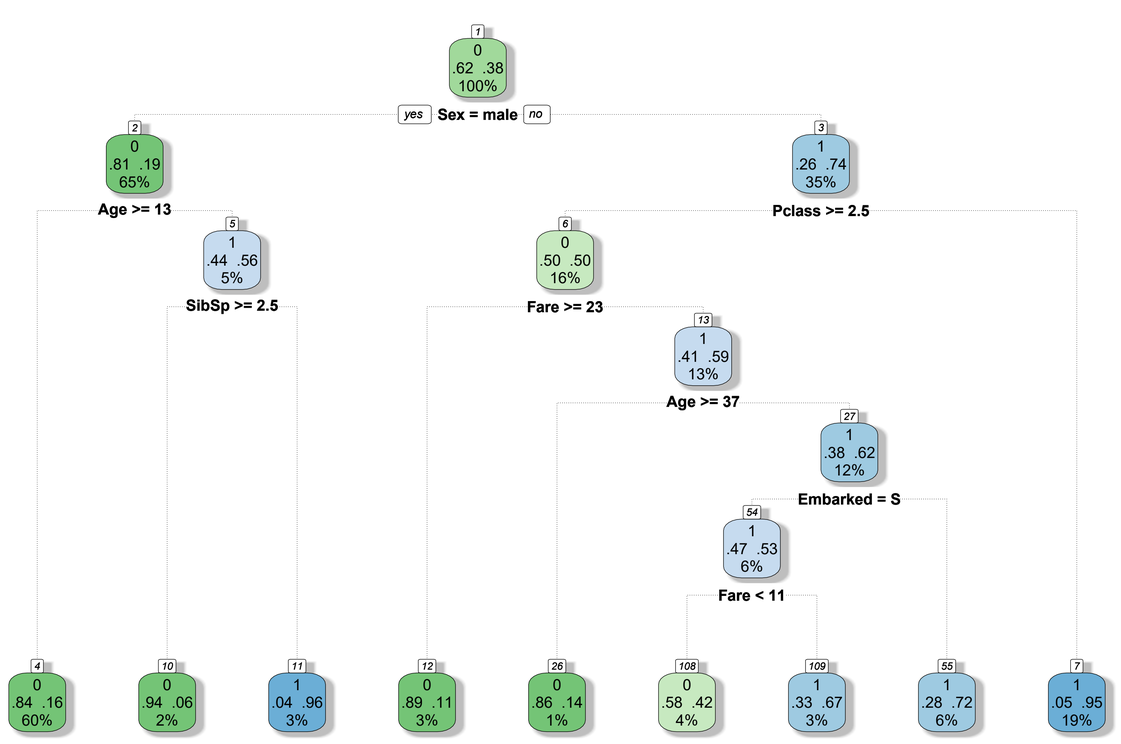


The model below shows the decision tree built to answer the question of what are effective ways of predicting survival? The boxes each represent a node while the number at the top of each box represents on what branch of the tree that node rests. The colors represent the majority predicted value of survival for that individual node: green predicts death while blue predicts survival. The darker the color, the more one-sided the prediction is. The decimal values are the percentages of values in that node that are predicted for each outcome: survival or death. The leftmost value will always be the percentage predicted death and the rightmost will always be the percentage predicted survived. Finally, the percentage at the bottom of each node is the percentage of total values that fall into that node, considering the previous node. For example, the first node at the top contains 100% of the values which then splits into 65% and 35% on the second level.

The model below shows the logistic regression model that was also built to answer the question of what are effective ways of predicting survival? This model was built around the age, sex, and class characteristics which will also help answer do women and children actually come first on a sinking ship? Further, sex and class are two of the most correlated characteristics to survival. The points represent observations and should be interpreted as either a zero or a one even though it looks like they are not perfectly on each line. Zero indicates death and one indicates survival. The curves show the probability of survival for each class as an individual's age increases. It is evident that young women in first class had the highest survival rate. The fact that a woman in third class had an equal or higher probability of survival than a man in first class shows that the women were prioritized when boarding lifeboats. It is also evident that the highest probabilities of survival, regardless of class or sex, occur on the left side of each chart meaning that children were also a priority in the effort to save as many

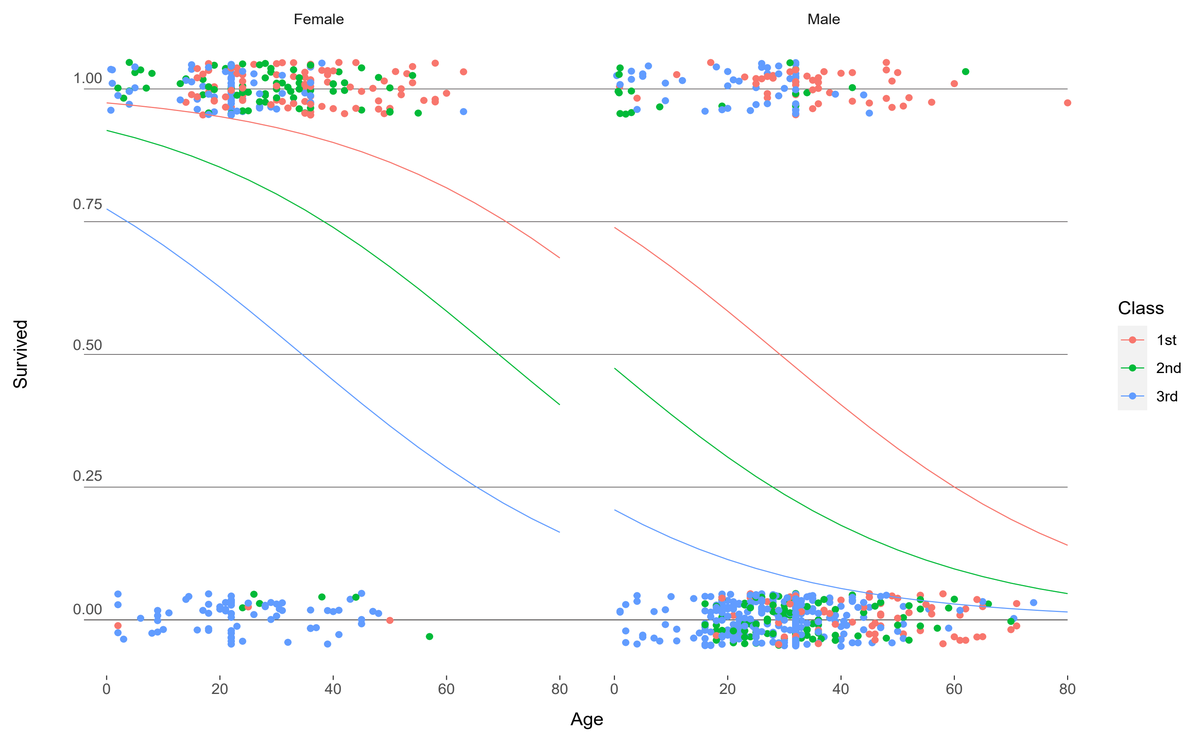
Decision Tree

This decision tree was built and trained using the Rpart package and pruned using the 1SE rule to a CP level of .00588.



Logistic Regression

The logistic regression model was built around age, sex, and class. The points show the survival status of each observation and the curves show the probabilities of survival for each age, sex, and class combination.



Author: Jack R. O'Connor IST 719

Data Source: https://www.kaggle.com/c/titanic
GitHub Repository: https://github.com/jro2014