Predicting Survival on the RMS Titanic

Binary Classification

Joseph Mifsud
Hungry for Opportunity

January 14, 2021

Abstract

A study was performed to compete in the modeling passenger survival on the RMS Titanic. The model training dataset including 11 features for 891 passengers was downloaded from Kaggle.com. Missing values for specific features were modeled and estimated. Each feature was tested for significance and correlation with survival with Pearson's r metric. A logistic regressor was created from these significant features to estimate survival. The model was tested on a reserved sample of 491 passengers. The test yielded an accuracy of (0.78 ± 0.035) and an f1-score of (0.82 ± 0.040). Although variables were not analyzed for correlation with one another, this is a useful model for predicting the survival of passengers.

1 Introduction

1.1 Background

Binary classification is an important tool for every data scientist. Whether an object or action can be categorized into a feature of interest or not is inherently an task of binary classification. Binary classification could be use to help a farmer determine whether or not to plant a field or inform a physician if a growth is likely cancerous. This tool's use-case is very common. In demonstration of this machine learning technique, features of the passengers on the ill-fated RMS Titanic are used to estimate his or her survival.

1.2 Problem

"Which are the models and variables that produce the greatest f1-score for our test-set?"

1.3 Interest

Binary classification is an incredibly useful tool for many different types of activities. Binary classification is of interest to the marketing department when answering the question "Is this person a potential customer?" It can answer the banker's question of "Should I provide this loan?" In medicine, binary classification can be useful to answer the question "Is this a benign growth?" Even the farmer who wants to know if a field should be planted can be assisted by binary classification. This versatile tool is of interest of a wide range of individuals in every industry.

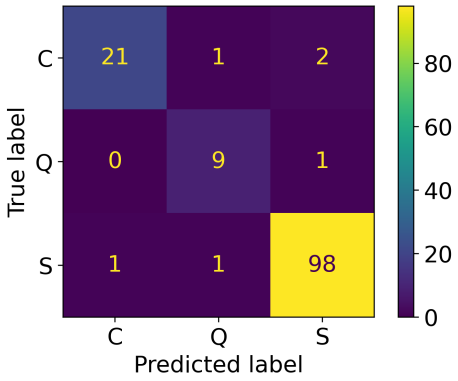
2 Data

2.1 Data Sources

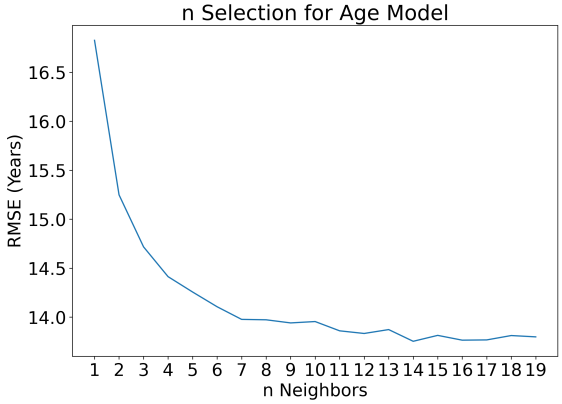
Kaggle.com provided the dataset for this analysis.

2.2 Data Cleaning

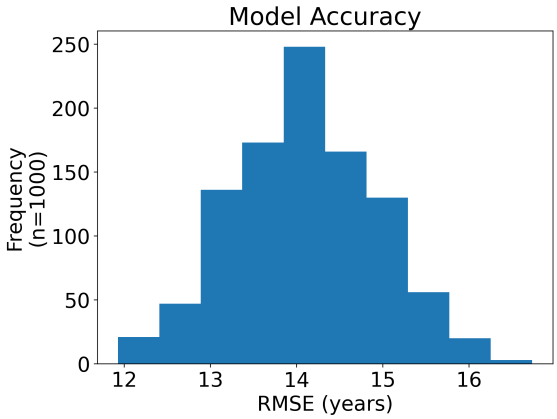
There were three features of the dataset which were missing values for a number of passengers: **Age**, **Cabin**, and **Embarked**. The most straightforward feature to address was which port the passenger embarked from. The passenger's fare and passenger class would be indicative of this or her port of embarkation. A decision-tree classifier was defined and fit to these features from 711 passengers. The decision-tree model was tested against a reserve sample of 178 passengers yielding an accuracy score of 0.90 and an f1-score of 0.90. The port of embarkation for the two target passengers was then estimated with this model. It was estimated that both of these passengers embarked at Southampton.



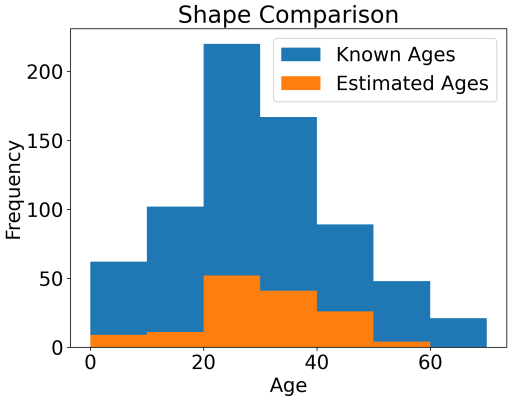
177 passengers in this dataset did not have an age. There are various techniques to handle missing values in data. One such technique is imputation. A kNN model of the age of the passengers was created. An iterative approach was used to determine the number of neighbors in the passenger model. The number of neighbor was varied in a for-loop between 1 and 20 neighbors. The RMSE for each number of neighbors was calculated and averaged over 500 iterations. The average RMSE was plotted against the number of neighbors.



The "elbow method' was use to select the number of neighbors in our kNN model. Four neighbors were selected to estimate the age of each passenger whose age information was missing from the data set. The model was tested against a randomly reserved set of 20% of the sample data. 1,000 tests were completed yielding a mean RMSE of (14.11 ± 0.85) years. The RMSE of this model represents roughly half of a human generation.



The shape of the modeled age distribution is similar in center, shape, and spread to the known age distribution. Although the RMSE of the model is a significant portion of the human lifespan, there is no obvious reason to reject the results of the model.



The last feature which included missing data was Cabin. 77% of all passengers have no record of which cabin they stayed in. It was thought that no appreciable amount of information could be mined from this feature as it is. Therefore, the cabin feature did not receive a modeling treatment. The feature was ultimately engineered into another feature **Deck**.

2.3 Feature Engineering

Within the data of the Kaggle dataset were some additional data. The data in the set was engineered to produce additional data for the model. Two features were engineered from the existing dataset: **Title** and **Deck**. An appreciable amount of passengers did not have a known value for cabin. An effort was made to salvage any information in this feature by engineering **Deck**. The **Deck** feature was created from the **Cabin** feature. Each known cabin number was prepended with a letter to indicate on which deck the cabin was located. This letter was stripped from

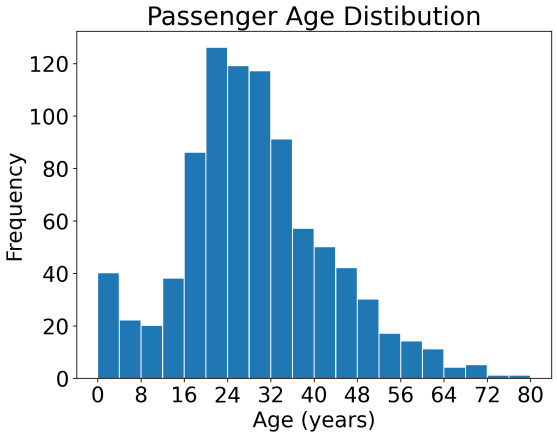
all passengers whose cabin information was known and included in the new feature. Accordingly for all passengers whose cabin, and therefore deck, was unknown a question-mark acted as placeholder for the passenger's deck.

Deck	Freq.
?	687
A	15
B	47
C	59
D	33
E	32
F	13
G	4
T	1

Each passenger's name includes a title. Example titles are "Mrs" and "Capt." This title was stripped from the passenger and included in the dataset as another feature.

Title	Freq.
Mr	517
Miss	185
Mrs	127
Master	40
Prestigious	22

The less common, and more prestigious, titles were replaced with the string 'Prestigious.' Both Mme and Ms were replaced with Mrs. Mlle was replaced with Miss.



A passenger's age in years is given by the Age feature. This continuous variable was binned into bins of equal width in years. The Rice rule was used to determine the number of bins. 20 bins were used to encode the Age feature. The width of each bin was 4 years. Each passenger was one-hot encoded to indicate in which age group bin he or she belongs.

A feature which was investigated but ultimately rejected for engineering was ticket number. 25% of passenger's ticket number have a short, cryptic, alphanumeric prefix. The sample size of any particular prefix was not large enough to make meaning for the population.

2.4 Feature Selection

The Kaggle data set includes features which were irrelevant or redundant to this study. The most striking example of irrelevance is PassengerId. The feature set of this study was established by calculating Pearson's r for correlation of each feature with survival. Features that were found to have a correlation of any strength at a significance level of 0.05 were included in the model feature set. 21 features were ultimately included in the model.

Feature	Correlation	P-value
Mr	-0.55	2.4×10^{-71}
male	-0.54	1.41×10^{-69}
Class3	-0.32	5.5×10^{-23}
Deck ?	-0.32	3.09×10^{-22}
Parch0	-0.15	1.0×10^{-5}
Southampton	-0.15	7.22×10^{-6}
SibSp0	-0.12	5.3×10^{-4}
SibSp8	-0.07	3.6×10^{-2}
Queenstown	0.01	9.134×10^{-1}
Parch2	0.08	2.5×10^{-2}
Master	0.09	1.1×10^{-2}
Class2	0.09	5.3×10^{-3}
Deck C	0.12	6.06×10^{-4}
0-4	0.13	1.02×10^{-4}
Parch1	0.13	5.9×10^{-5}
Deck D	0.15	6.23×10^{-6}
Deck E	0.15	1.33×10^{-5}
SibSp1	0.17	2.0×10^{-7}
Deck B	0.18	1.44×10^{-7}
Class1	0.29	3.2×10^{-8}
Miss	0.34	6.7×10^{-25}
Mrs	0.34	2.7×10^{-26}

It is likely there are correlations between these features. **Mr** and **male** are very similarly correlated with survival and probably with one another. Indeed, there may be correlations between deck and passenger class. Correlation between features was not tested for. Implementing this procedure would decrease model complexity but it isn't though to have an impact on model performance.

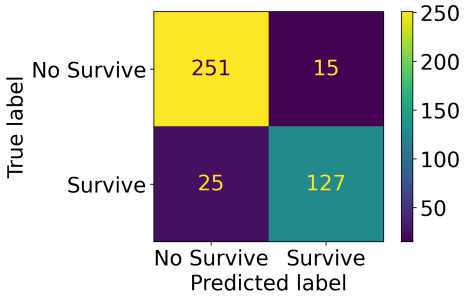
3 Methodology

Features from the feature set were one-hot encoded for each passenger. The data were then split

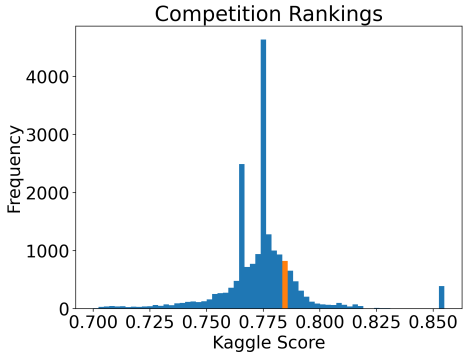
into a training and test fold. 20% of the data were held for the test fold. The training data were fit to a logistic regressor.

4 Results

The test yielded an accuracy of (0.78 ± 0.035) and an f1-score of (0.82 ± 0.040).



The overall performance of the model is above average. Among competitors in the Kaggle competition, this model scored as good as or better than 87% of the competing models.



There are 67% more type 2 errors than type one. A reduction in type 2 error would increase model performance.

5 Discussion

This model outperforms the lionshare of models on the Kaggle platform. Improvements can be made in multiple areas.

The difference between the test accuracy and the competition accuracy is greater than 10% of the competition accuracy. This could indicate the model is overfit in at least one region. **Age** stands out as a feature for further investigation. Depending on which passengers were selected for the training set, there were different outcomes for which age groups were correlated with **Survived**.

Further, 17 of the features included in the model have a very weak, but significant correlation with **Survived**. It is likely that eliminating some of these weaker features by testing for correlation between them would enhance the accuracy of the model.

Finally, testing other modeling techniques could improve performance. The Random Forest classifer as well as Grid Search classifier both lend themselves to this type of analysis. Only an analysis would determine which performs best.

6 Conclusion

Binary classification is an important tool for a plethora of use-cases. This particular model is both useful and competitive. There is still work to be done to improve the model. Gains in performance are likely to be made by testing other classifiers and pruning the feature set. Notwithstanding, the model is as good as or better thsn 87% of the models on the Kaggle leaderboard.