Logistic Regression Review

Jennifer Ahlport

December 9, 2020

Abstract

Within the field of data science and machine learning there are many solutions to the same problem. For the Kaggle Titanic competition, its creating a model that helps predict whether or not a passenger will survive based their features, such as age and sex. Using a simple logistic regression I'm able to get within the top 10% of scores, even going up against much more complicated models. This is due to spending time on data exploration and feature creation, to ensure that the features selected do a good job explaining the data, and not overfitting the data.

1 Introduction

When I started looking at the Kaggle introductory Titanic problem, I saw many direction to take this classification problem. In my first pass at the problem, I looked into how different classification models would compare, and got decent results. For this, I wanted to dive into how well a logistic regression would do when combined with more advanced techniques. The goal of this model is to predict whether or not a passenger survives based on their features.

2 Data Review

The first step with any machine learning problem is to understand the data. There are 891 instances in the training data set, with 10 different features for each instance.

1. Ticket Class (pclass): 1, 2, or 3
2. Name (name): Full name of the passenger, including titles
3. Sex (sex): Male or Female
4. Age (age): Age of the passenger, in years, between 0.42 and 80. Null for 177 instances in training set.
5. Siblings/Spouses (sibsp): An integer value of the number of siblings or spouses of the passenger that are on the Titanic, between 0 and 8.

6. Parents/Children (parch): An integer value of the number of parents/children of the passenger that are on the Titanic, between 0 and 6.

7. Ticket Number (ticket): A non-standardized string containing the ticket number.

8. Passenger Fare (fare): Passenger fare, float between 0 and 512.3292.

9. Cabin Number (cabin): Passenger cabin number, null for 687 instances in training set. Some instances have multiple cabin numbers listed.

10. Port of Embarkation (embarked): Classes include Cherbourg (C), Queenstown (Q), and Southampton (S), null for 2 instances in training set.

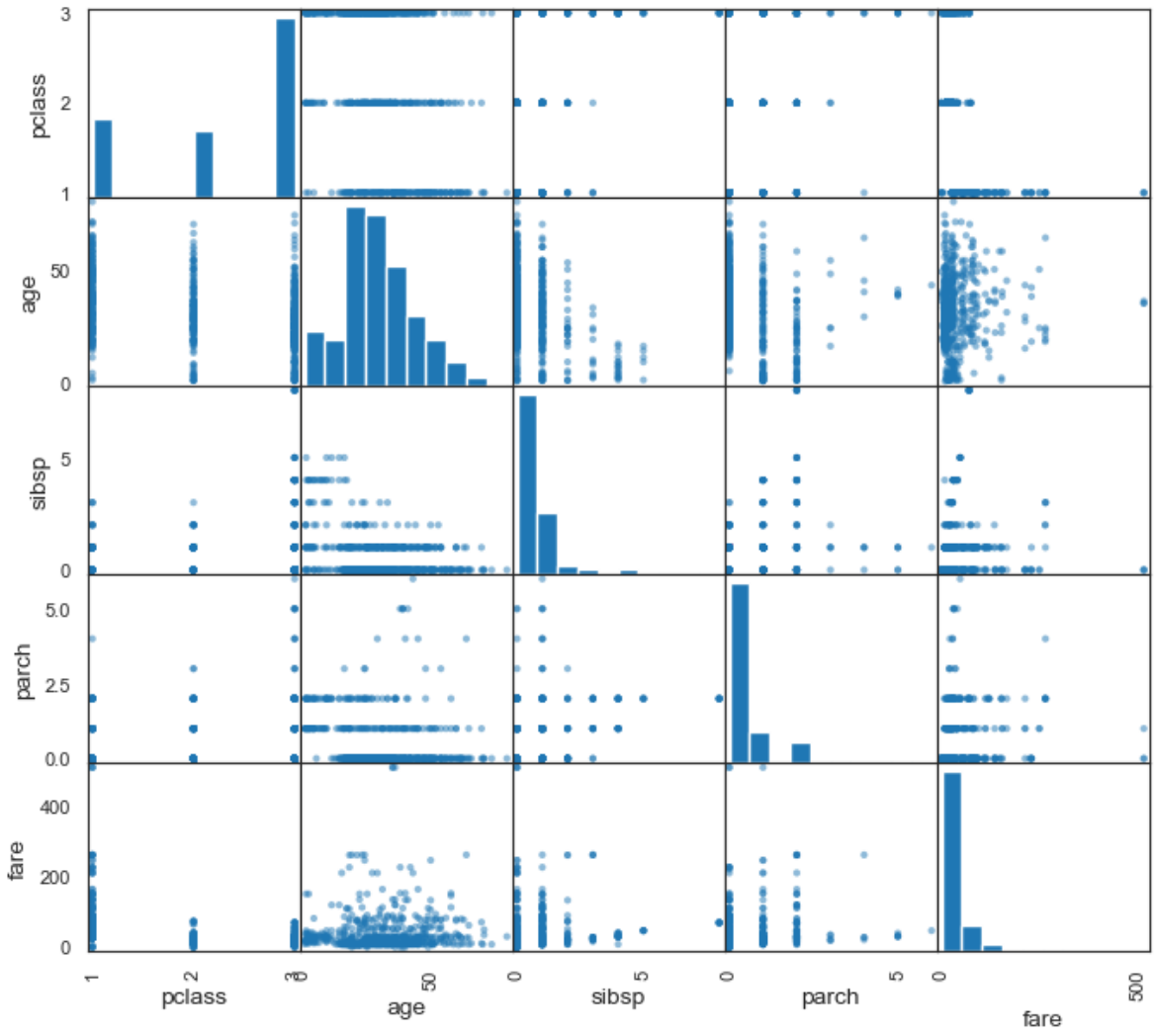

Figure 1: Figure shows a histogram for each numerical feature on the diagonal, and a scatterplot between each feature in the off-diagonals.

Figure 1 provides an overview of the numerical features, pclass, age, sibsp, parch and fare and their relationship with the other features. Given the clustering of data, it is hard to see any strong correlation between the features.

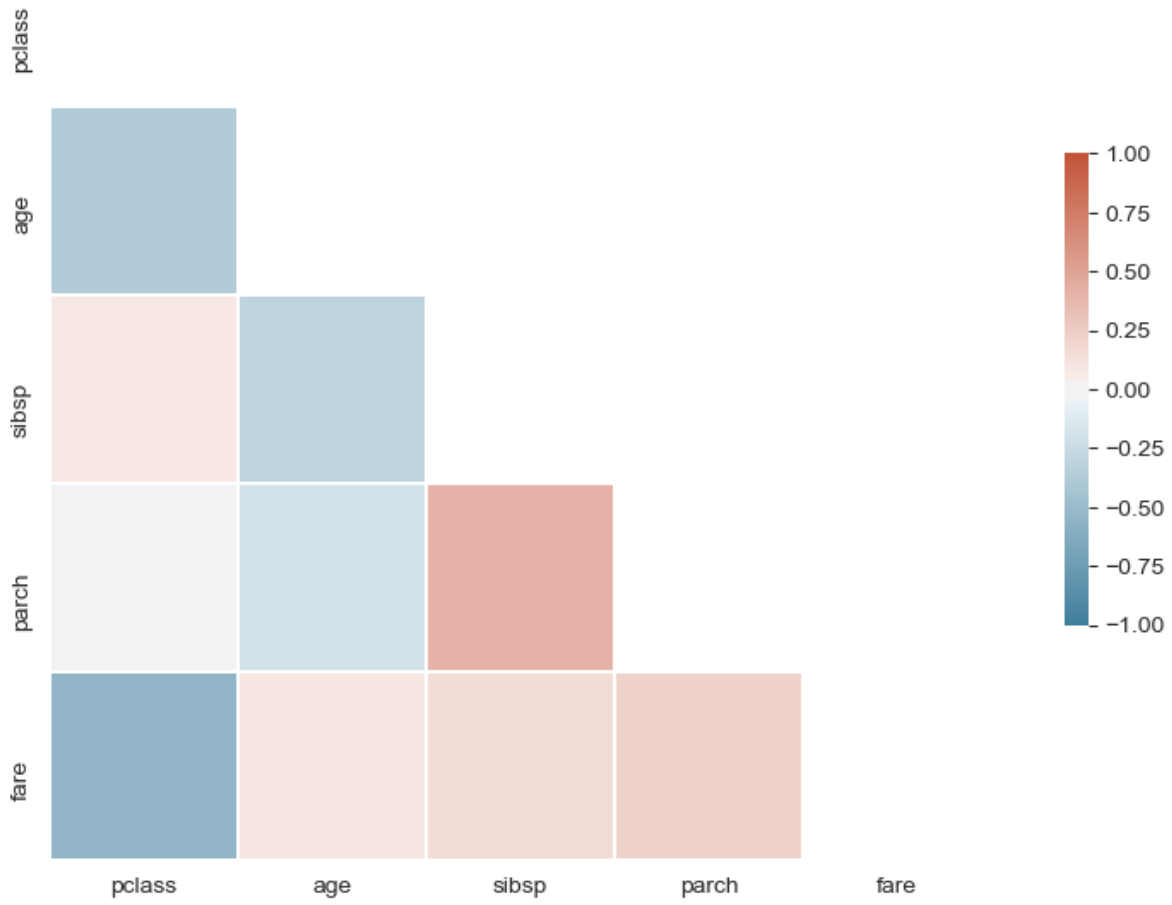

Figure 2: Correlation between the numerical features in the training data.

Figure 2 shows the correlation between the numerical features in the training data. In order to understand the data better, I've dug into some of the larger magnitude correlations. Passenger class and fare have a strong negative correlation (-0.55). This makes intuitive sense as tickets in the first class are expected to cost more than tickets in third class. The next largest correlation magnitude is the positive correlation between the number of Parents/Children and the Number of Siblings/Spouses, 0.41. As seen in Figure 11 and Figure 12, most passengers have no relatives on the boat with them, which is reflected by this high correlation. The next highest magnitude of correlation is between age and passenger class at -0.37. This states that those who are older typically travel in first class, while those who are younger were traveling in third class. This most likely represents the wealth difference between those who are older and those who are younger.

3 Feature Construction

The features included in the data training set need to be adjusted before they can be used in a classification model.

3.1 Ticket Class

The passenger class data is broken up into three categories, 1, 2 and 3, denoting which class of ticket the passenger bought.

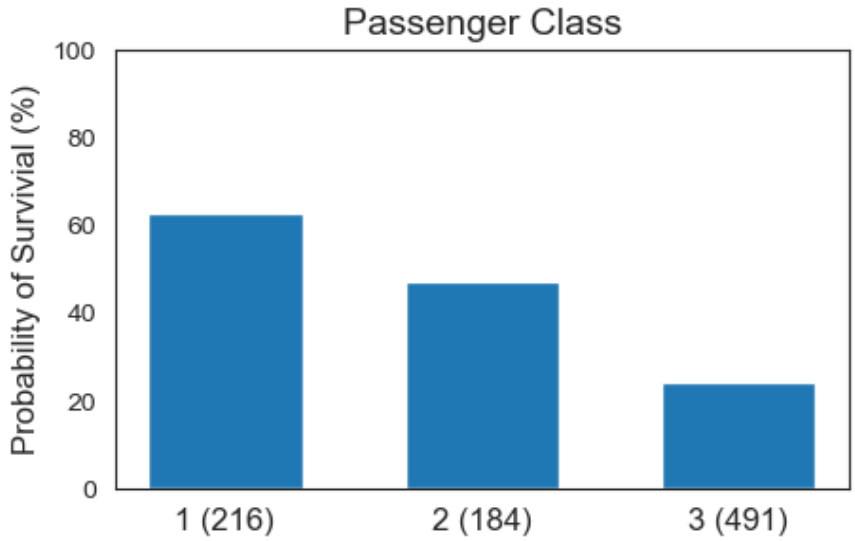

Figure 3: Survival by passenger class, with the counts for each category denoted in (#) in the x label.

Figure 3 shows the probability of survivial broken down by the different passenger classes, where survival is highest in first class and goes down for passengers in the second and third classes. Although there is a relationship between the passenger class and the probability of survival, it is not a linear relationship - the R^2 is only 0.11. Instead, I chose to transform the pclass feature into three boolean category features using the OneHotEncoder function to remove the numerical tie between the categories.

3.2 Name

The name feature contains the actual names of the passengers on the titanic. Given that these are unique, there is no pattern that can be discerned from the names themselves. However, one important pieces of information in the names is the titles of each passenger. Figure 4 shows the probability of survival based on the passenger title. The benefit of the title is that it includes information into the sex of the passenger, as well as the class of the passenger. In general, women are more likely to survive than men, and those in a higher socioeconomic class are more likely to survive. It was interesting to see that all passengers with the title "rev" for Reverend, did not survive. This makes intuitive sense - members of

the clergy sacrificing themselves so others can be saved. The less common titles are shown in Figure 5 and align with the general trend seen above. However, there are too few instances of each name to use them in a predictive model. For the model, I created boolean features for each of the common titles.

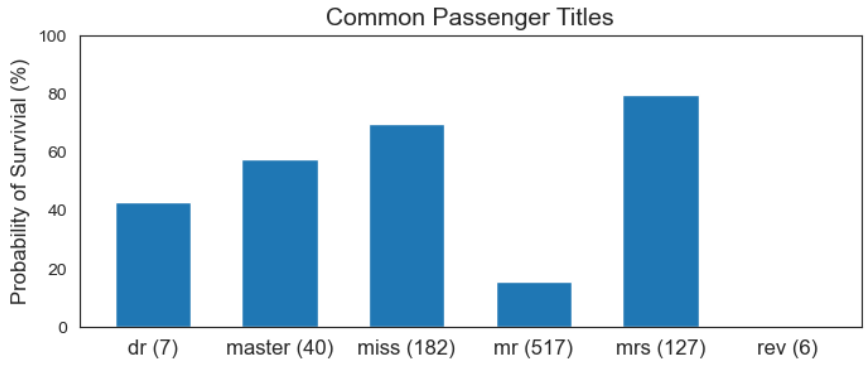

Figure 4: Survival by title for more common titles, with the counts for each category denoted in (#) in the x label.

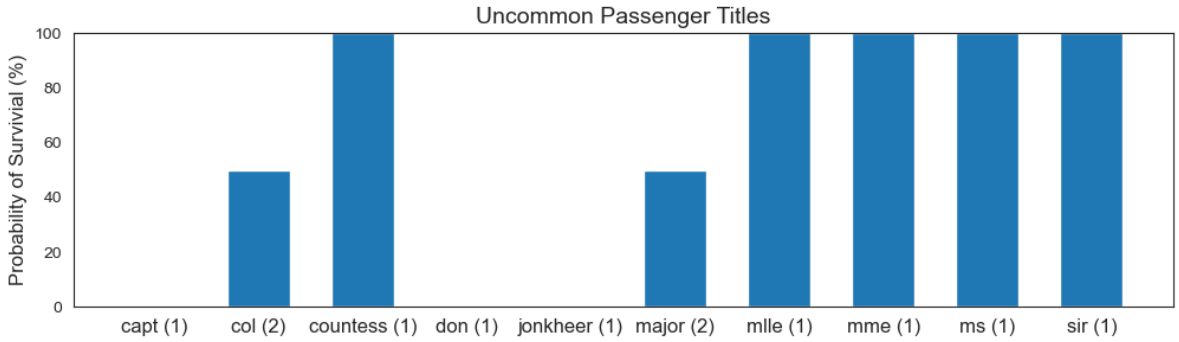

Figure 5: Survival by title for less common titles, with the counts for each category denoted in (#) in the x label.

3.3 Sex

Sex is one of the most important factors for determining if a passenger is likely to survive, with women survival probability being four times the rate of men, as seen in Figure 6. Figure 7 shows the sex breakdown when combined with the passenger class breakdown as well. These two plots show a similar story to the data we've seen before - females are more likely to survive than males, and those in class 1 are more likely to survive than class 2. However, the breakdowns show a noticeable difference between the women in class 2 than the men, where the women in class 2 are almost as likely to survive as the women in class 1, but the men in class 2 are less than half as likely to survive as the men in class 1.

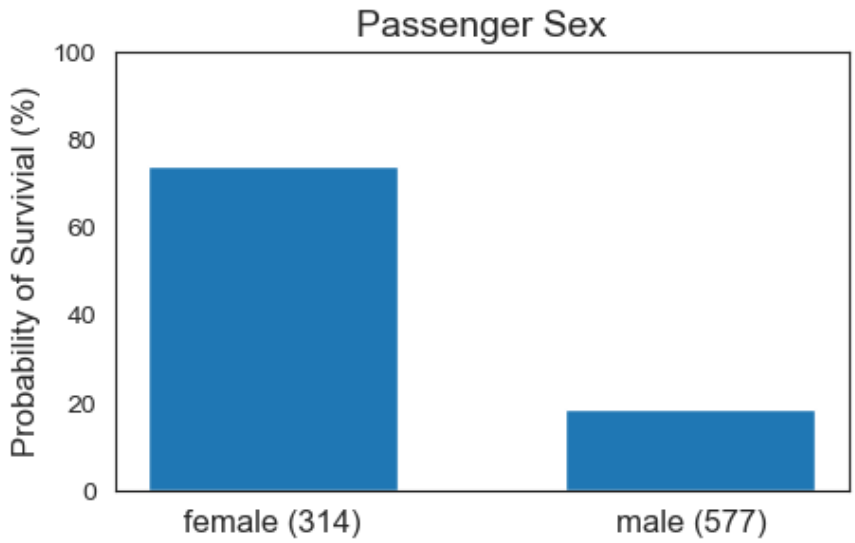

Figure 6: Survival by passenger sex, with the counts for each category denoted in (#) in the x label.

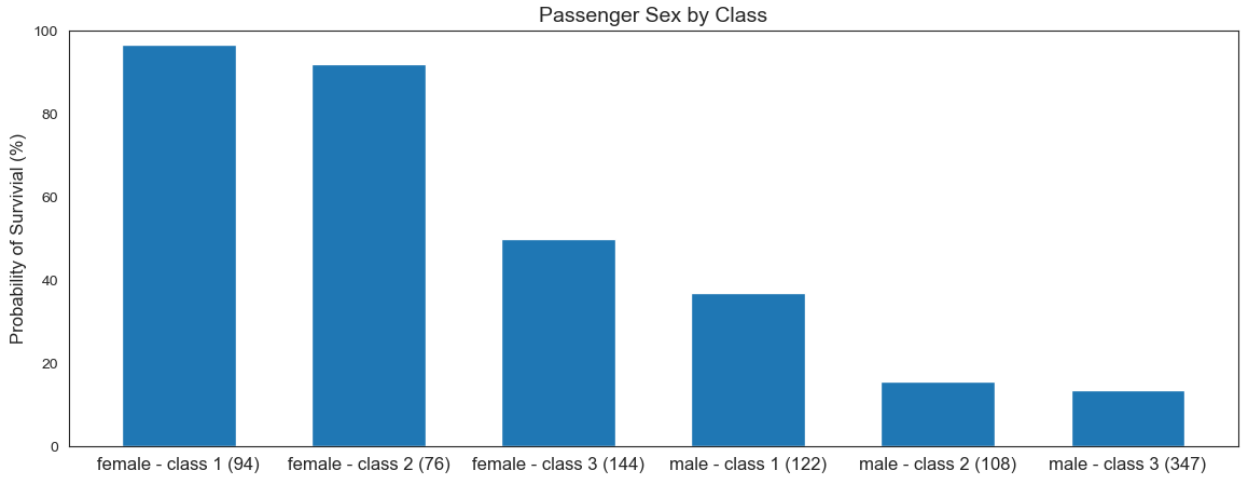

Figure 7: Survival by passenger sex and ticket class, with the counts for each category denoted in (#) in the x label.

3.4 Age

The age feature in this data gives the passengers ages in years, and ranges between between 0.42 and 80, with 177 null instances in the training set. Figure 8 shows the distribution of the ages of the passengers, with a mean of 29.7 years and a median of 28 years. There is no strong linear relationship between the ages and probability of survival, with an R^2 of 0.006.

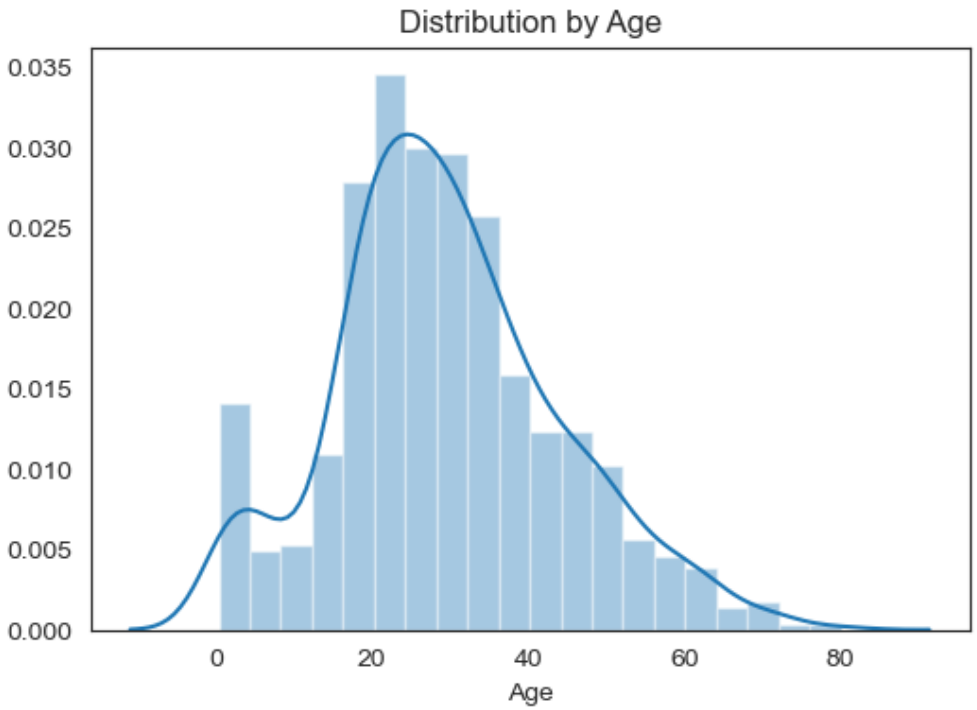

Figure 8: Distribution of passenger's ages.

Looking at passenger ages binned by decade, seen in Figure 9, there is an increase in survival probability for those under 10, but there was not a strong pattern beyond that. Breaking this down by gender, seen in Figure 10 shows an even stronger pattern. The survival probability is not significantly different by age for female passengers, but male passengers show a significantly higher survival rate in the 0-10 category than any others. This makes sense, given that the overall survival rate for women is higher and children were more likely to be saved than adults.

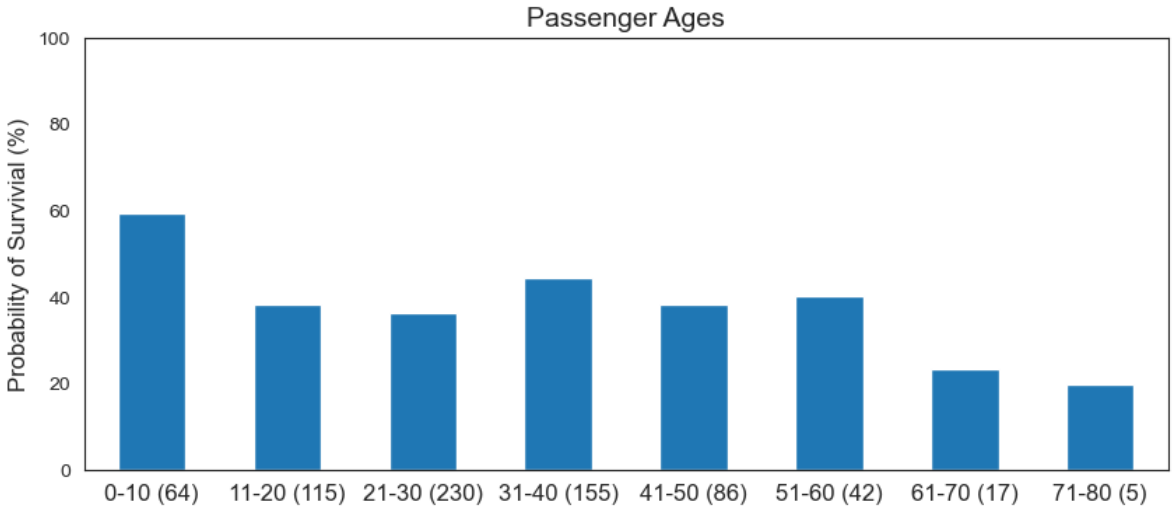

Figure 9: Survival by passenger age, with the counts for each category denoted in (#) in the x label.

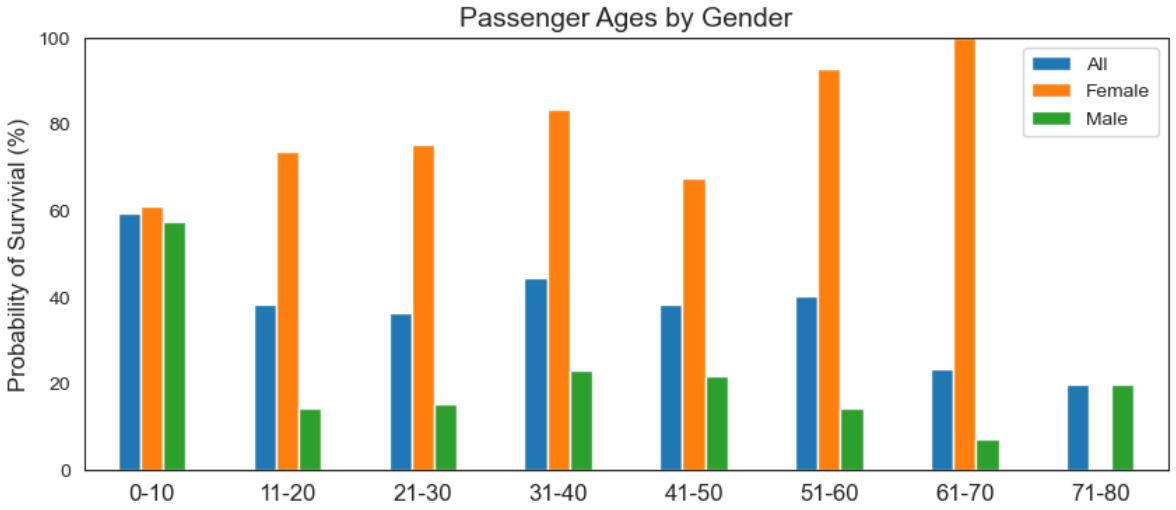

Figure 10: Survival by passenger age and sex.

3.5 Siblings/Spouses

The 91% of the passengers had one or fewer siblings or spouses on the titanic with them, as seen in Figure 11. The probability of survival is highest for those with 1 sibling or spouse on board, and it decreases significantly for those with 3+ siblings or spouses, although one cause of that may be the much smaller sample sizes. Due to the sample sizes for passengers with two or great siblings or spouses on board, I've chosen to create 3 sibling/spouse boolean features: 0, 1, and 2+.

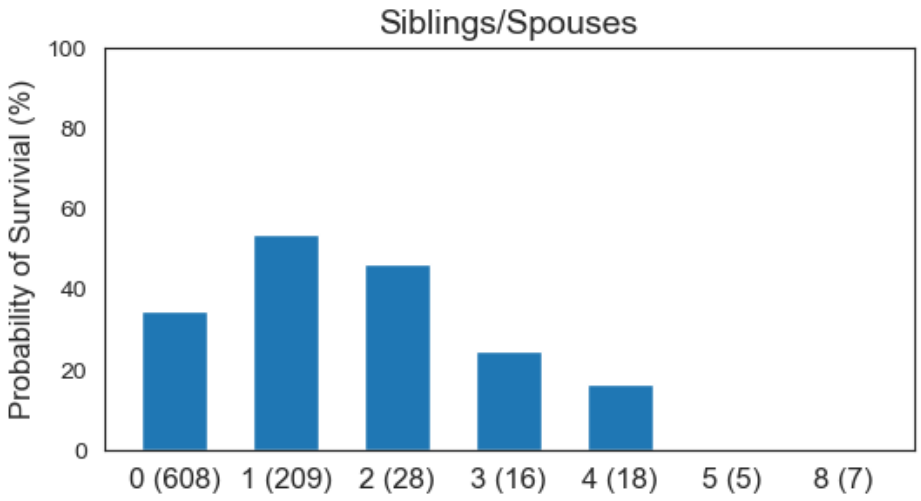

Figure 11: Survival by the number of parents/childred of passengers, with the counts for each category denoted in (#) in the x label.

3.6 Parents/Children

Similar to the number of siblings or spouses, 89% of passengers had one or fewer parents or children on the Titanic with them. The breakdown of parents or children does not show a strong trend, but does show that passengers without parents or children onboard with them had a lower chance of survival. Similar to the siblings and spouses features, I've created three boolean features for counts of 0, 1 and 2+.

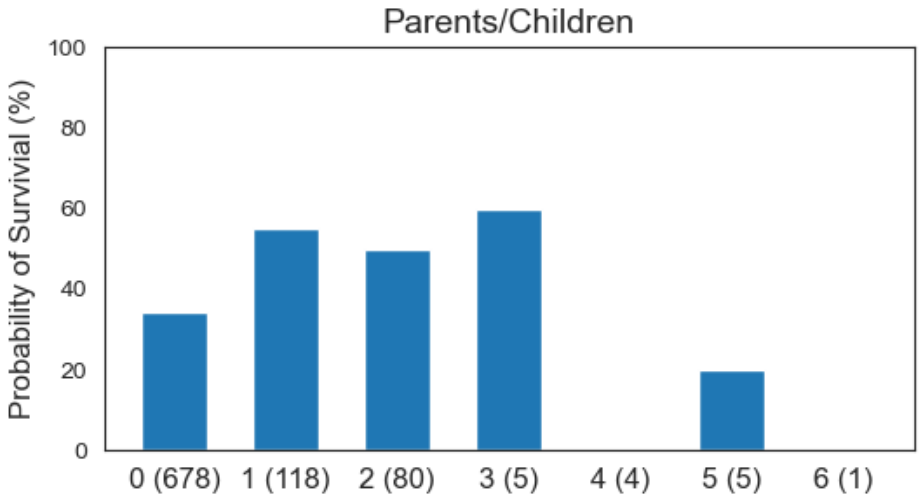

Figure 12: Survival by the number of parents/children of passengers, with the counts for each category denoted in (#) in the x label.

3.7 Ticket Number

The ticket numbers do not have a standardized pattern to them, even when conditioned on the port of embarkation. As a result, I am not using them at this time. Additionally, it is unclear how a ticket number would have an intuitive reason for predicting if a passenger survived or not, given that information such as fare, passenger class, and port of embarkation are already accounted for in other features.

3.8 Passenger Fare

Passenger fares are available for all instances in the training data and are heavily skewed to the right, as seen in Figure 13, ranging between 0 and 512, with a mean of 32.2 and a median of 14.45.

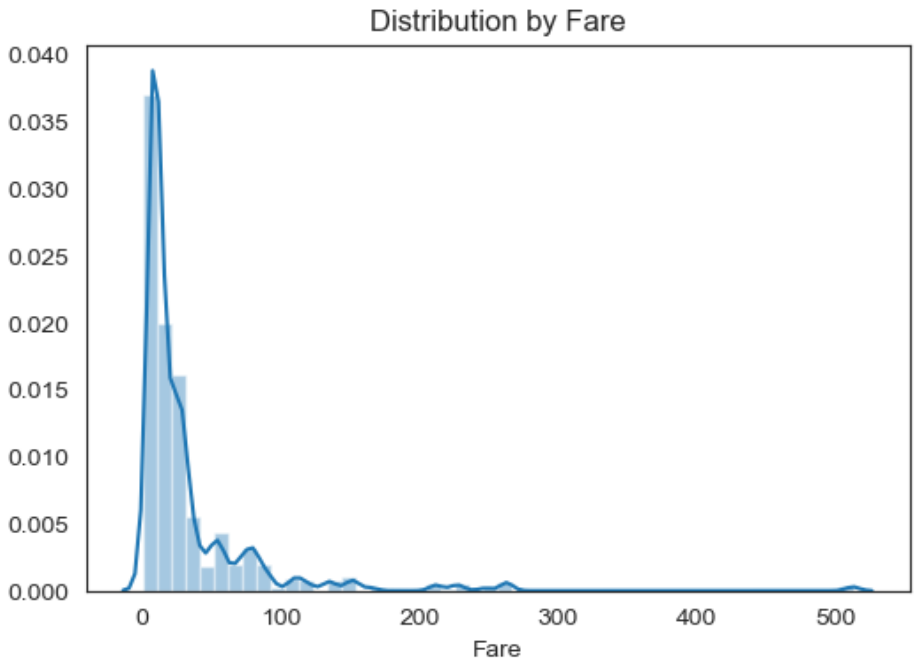

Figure 13: Distribution of passenger's fares is heavily skewed to the right.

Breaking the passengers fares into deciles, there is a clear trend showing male passengers who paid more for their ticket are more likely to survive. The trend isn't as pronounced for female passengers.

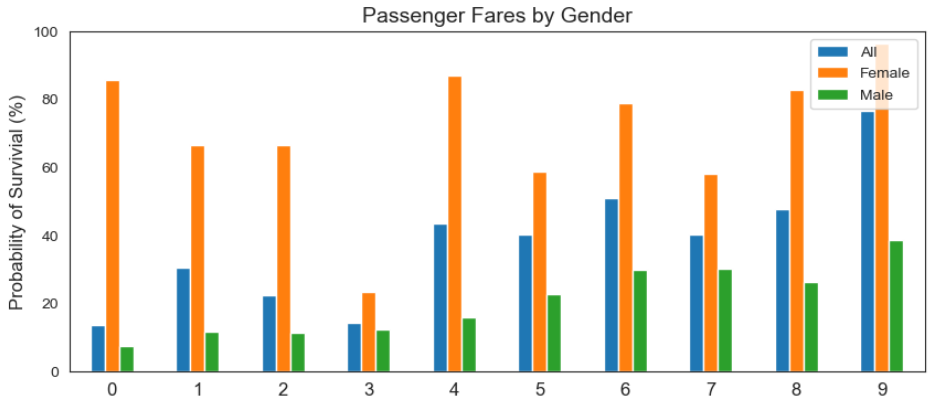

Figure 14: Probability of survival based on fare decile.

3.9 Cabin Number

The Titanic hit the iceberg on the starboard side at 11:40pm and sank in about two hours, with the bow sinking first. As a result, cabin location was very important in determining where the passengers were located when the sinking started.

There were 687 null instances in the training set for the cabin. Table 1 shows the breakdown of cabins defined vs not being defined by passenger class. Most of the passengers in first class have their cabin numbers defined and almost no passengers in third class have the cabin defined. Survival rate is higher for those with the cabin defined, even when conditioned upon the passenger class, shown in Figure 15.

Class	Cabin Defined	No Cabin Defined
1	176	40
2	16	168
3	12	479

Table 1: Provides the count of passengers in each passenger class that have a cabin number defined vs those that have a null value for the cabin number.

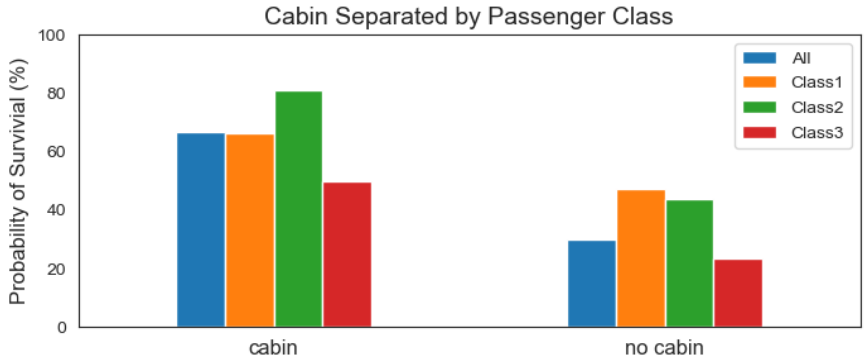

Figure 15: Survival by cabin defined or null, with the counts for each category denoted in (#) in the x label.

Focusing on just the passengers with a cabin defined, the pattern is less clear. Figure16 shows the probability of survival by a passenger with a cabin on the starboard side vs one on the port side and Figure17 show the probability of survival based on deck level, both conditioned on gender.

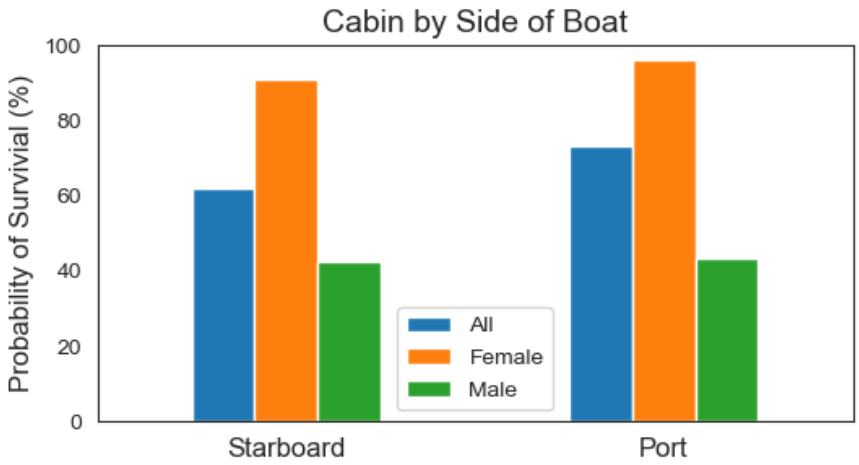

Figure 16: Probability of survival as determined by the side of the boat the passenger's cabin was on.

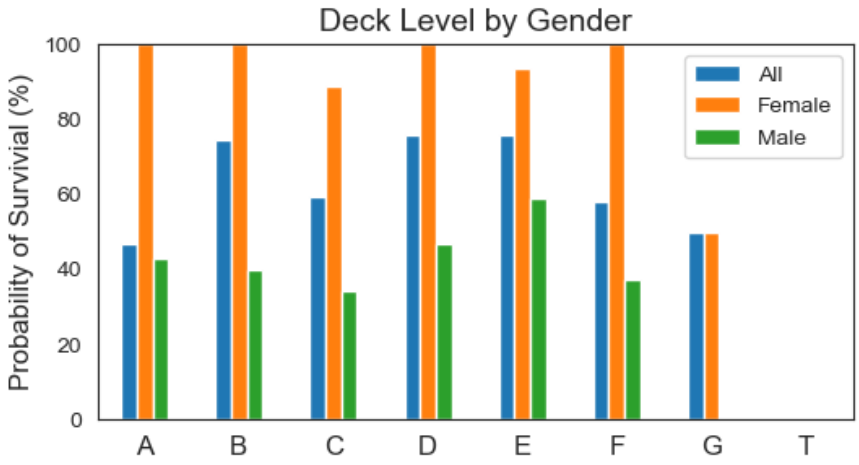

Figure 17: Probability of survival as based on the deck level of the passenger's cabin.

The patterns indicate that whether a passenger has a cabin defined is a stronger indication of survival than the deck level or side of the boat, but I will be testing features including both a binary feature for the passenger having a cabin or not, as well as the deck level and side of the boat to add a metric to this assumption.

3.10 Port of Embarkation

The embarkation location is broken up into three locations, denoted by the first letter of the city. These are Cherbourg, France (C), Queenstown, Ireland (Q), and Southampton, England (S).

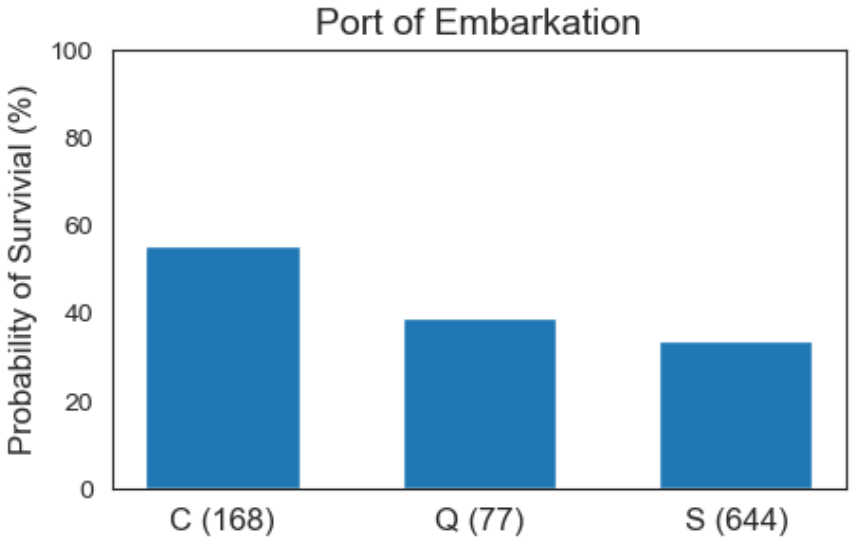

Figure 18: Survival by embarkation location, with the counts for each category denoted in (#) in the x label.

There is a difference in the survival based on the port of embarkation, although it is unclear why those boarding in Cherbourg have a higher probability of survival than those entering in Queenstown or Southampton. For the model features, I turned the initial embarked feature into three boolean category features using the OneHotEncoder function.

4 Candidate Models

For this project I'm testing two candidate models using all of the factors defined in Section 3:

1. A single logistic regression
2. Two separate logistic regressions, one for males and one for females.

The basis behind creating a model that is conditional on sex is due to the exploratory data analysis done in previous sections. Women were almost four times more likely to survive than me, and therefore many of the survival probabilities based on other features differed when looking at the two populations.

4.1 Hyperparameter Selection

For this model I've tuned the regularization hyperparameter, C. I've chosen to use L1 regularization because it better for feature selection instead of shrinking each parameter more evenly. This parameter is set up so that a smaller value results in a stronger regularization. Table 2 shows the results using k-folds cross validation with 5 folds and the F1 metric for scoring for the first candidate model. I'll be using C = 3 in the models, corresponding to the highest mean score. I will not be refining the hyper parameter further because the results do not differ much between parameters of 0.3 and 100.

K-Fold	**0.01**	**0.03**	**0.10**	**0.30**	**1**	**3**	**10**	**30**	**100**
1	0	0.293	0.739	0.774	0.738	0.743	0.730	0.719	0.719
2	0	0.521	0.735	0.752	0.761	0.761	0.742	0.742	0.731
3	0	0.409	0.726	0.731	0.731	0.731	0.722	0.722	0.722
4	0	0.381	0.682	0.689	0.689	0.688	0.683	0.683	0.683
5	0	0.500	0.701	0.756	0.824	0.820	0.812	0.812	0.812
Mean	0	0.421	0.717	0.740	0.748	0.749	0.738	0.736	0.733

Table 2: F1 score for regularization parameters between 0.01 and 100 with L1 regularization

4.2 Feature Importance

In Section 3 I created a series of features based on the initial data provided, however, not all features will be useful in the regression. Figure 19 shows the top 5 smallest and largest coefficients for the first candidate model, including all passengers. A positive coefficient is associated with a higher probability of survival, and a negative coefficient means a higher probability of not surviving. The feature with the largest negative coefficient in his model is if a passenger is male. This aligns with what was seen in the Section 3.3, where female passengers were 4 times more likely to survive than male passengers.

Having "Master" or "Rev" in the passenger's name are strong indicators of survival, with Rev being associated with not surviving, and Master associated with survival. Looking at Figure 4, passengers with Master in their name are 3.7x more likely to survive than passengers with Mr in their name. None of the 6 passengers with Rev in their name survived, which supports the strong indication of not surviving.

It's interesting that $\text{cabin}_{P}ortandcabin_{S}tarboardbothshowupinthetop5coefficients.Thisindicatesthath$

Passenger class is an important indicator for survival, with passenger class 1 or 2 showing up with almost identical coefficients (1.10 for class 1, 1.07 for class 2). Age is also important in survival, with those passengers who are over 10 being less likely to survive.

Ten of the factors have a coefficient of 0:

- Passenger Class 3
- Name includes Dr or Mrs
- Fare in quartile 1, 2 or 4
- Cabin on floor F or T, or no cabin noted

- Embarked in Queenstown

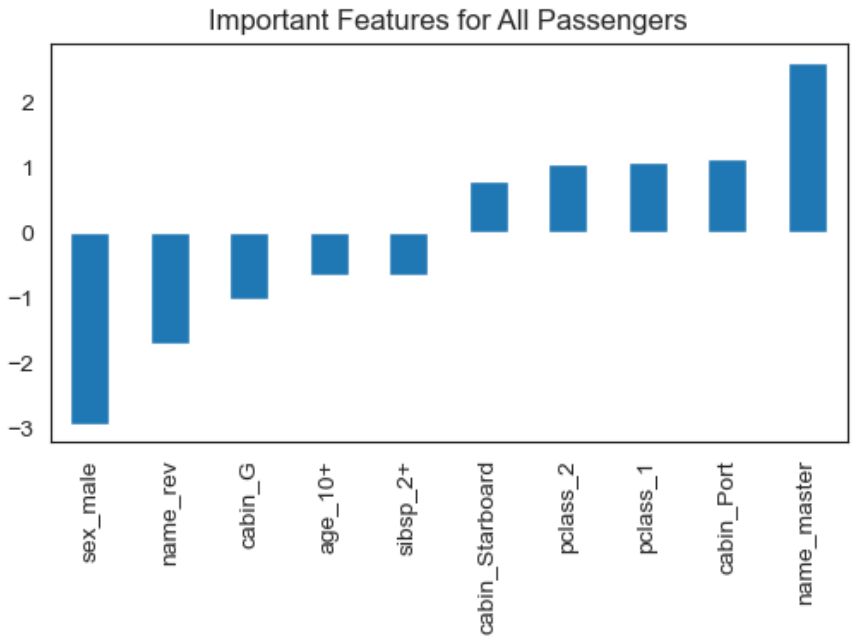

Figure 19: Coefficients from the logistic regression including all passengers.

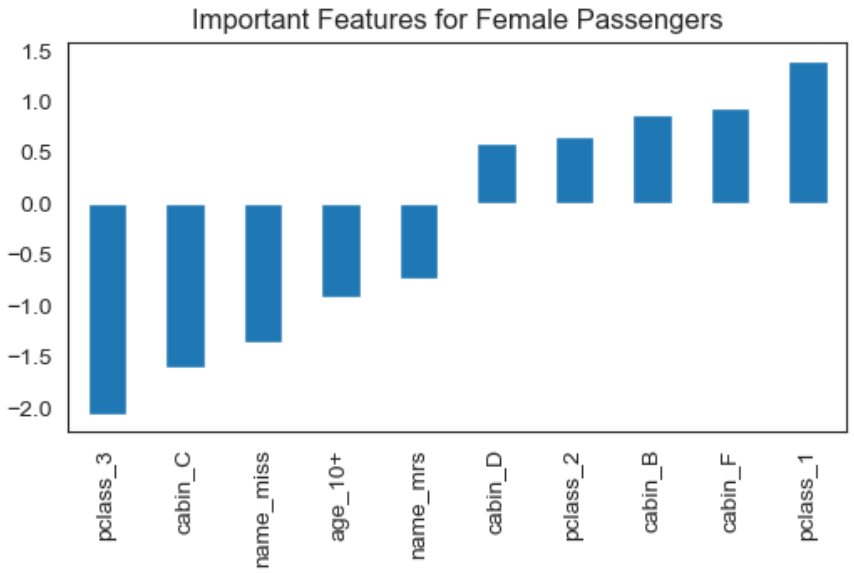

Figure 20: Coefficients from the logistic regression including female passengers only.

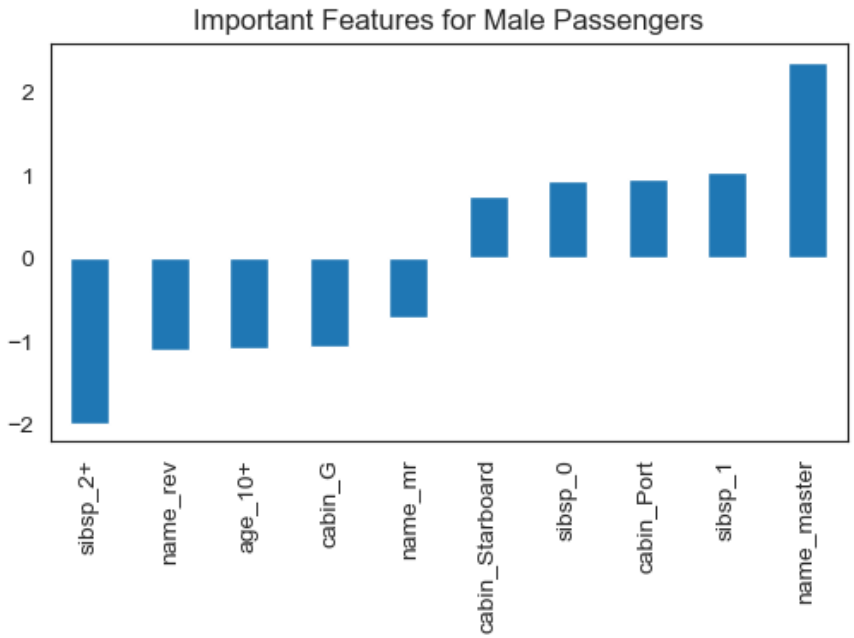

Figure 21: Coefficients from the logistic regression including male passengers only.