Steph Tung

Midterm Assignment – General Assembly

Imputing values for the missing values in the Age variable

Of the 291 observations, 177 were missing values for the age variable. This constitutes about 61% of our data – so dropping these rows would not be a reasonable approach. There were a few different options for imputing values: (1) use the original data's mean value, (2) use the most frequently occurring values, (3) impute values randomly that would keep the mean and standard deviation around the same as the original dataset. If I only used the mean, that would cause the original data's mean to increase as well and rather radically change the frequency distribution of our observations. Instead of a younger group of passengers with a few older passengers (explaining the positive skew of the data), we'd see a larger group of older passengers with a few younger passengers. If I had used the most frequently occurring values, this would've been difficult as well, since it would've required determining how many of each to use in assigning the 177 missing values, while still preserving the original sample's mean and standard deviation. To avoid manually assigning missing values (and potentially ending up with a very differently distributed dataset), I went with option 3, using NumPy to randomly assign values to all null age observations – while aiming for a mean close to 29.7 and a standard deviation of 14.5.

Building Logistic Regression Models

Model 1: For the first model, I used almost all the features – passenger class, age, number of siblings or spouses aboard, number of parents or children aboard, fare paid, point of embarkation, and sex (female or male) – just to see what the outcome would be.

One reason why the coefficients may have been so low for this model was that there may have been cases of multicollinearity between features. For instance, passenger class is a measure for socio-economic status, while fare paid and embarkation point may well have reflected that as well (wealthier versus not as well-off areas).

Table 1. Feature Importance (Model 1)

Feature	Coefficient	Absolute value
Passenger class	-0.067766	0.067766
Age	-0.096076	0.096076
Number of siblings & spouses	-0.217057	0.217057
Number of parents & children	0.048021	0.048021
Fare	0.007945	0.007945
Embarkation point	0.054056	0.054056
Female	0.034269	0.034269

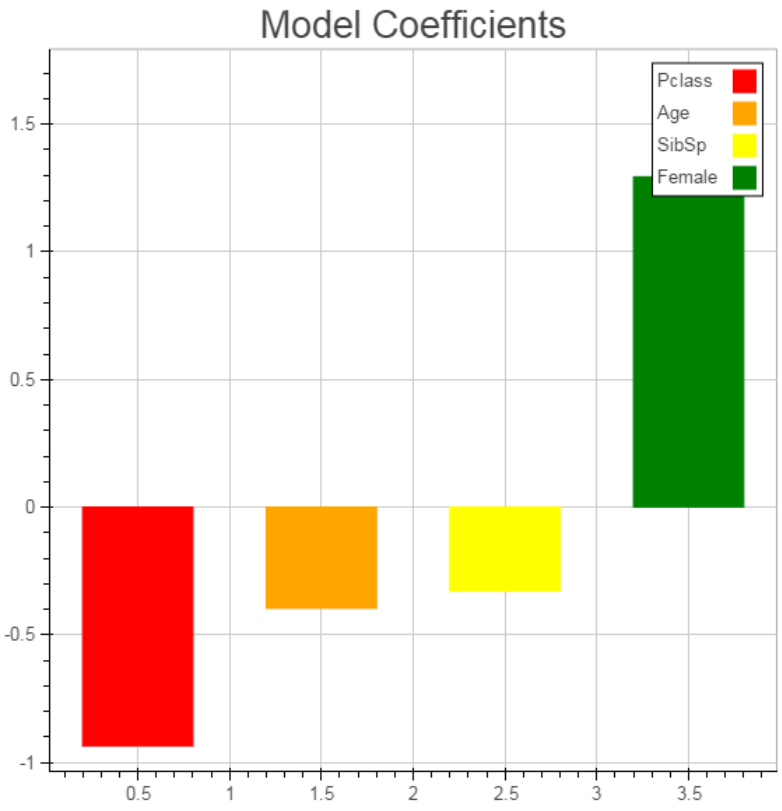
Model 2: Recognizing that the first model was overly complex, I removed the embarkation point, fare paid, and number of parents and children aboard features. Intuitively, it made more sense that women and

children would have a higher likelihood of surviving – and most likely, these would be women and children from wealthier backgrounds. Therefore, I kept female, passenger class, and age in. I tried number of parents and children, but it proved to be less important than number of siblings and spouses. This second model confirmed that indeed, the strongest predictive features of survival were being female and in an upper class (as indicated by passenger class). The negative coefficients on the other two features (see Table 2 and Figure 1) suggests that younger women with fewer families also had a better chance of survival – perhaps because families trying to stay together either could not fit or did not make it onto the lifeboats soon enough.

Table 2. Coefficients for Model 2

Feature	Coefficient	Absolute value
Female	1.294027	1.294027
Passenger class	-0.935045	0.935045
Age	-0.396121	0.396121
Number of siblings & spouses	-0.329025	0.329025

Figure 1. Model Coefficients



Cross-Validation

To score the performance of my first model in classifying survival, I used cross-validation scores (with 5 folds – to allow for multiple iterations of training / test groupings). The result was 0.612 – not significantly better than random guessing (in which case we'd have a 50/50 chance of correctly determining survival). When I calculated cross-validation scores again for my second model, which had far fewer features, the mean was 0.785 – a fairly sizable increase from the initial 0.612.

ROC, AUC

The ROC curve for model 2 is displayed in figure 1, below – based on our ROC curve, we have an AUC of 0.854. Given that model 2 performed better in its classification of survival rate (judging based on the CV score – though we also could have examined precision and recall performance), we would expect that the true positive rate is greater than the false positive rate.

If we wanted to improve precision, we would want more true positives than false positives – thereby being closer to the upper left corner. One threshold we might consider is 0.206 – since that would give us a TPR of 0.86 and an FPR of 0.37. This is rather low, but when we have a higher threshold, there's a very low TPR and FPR rate.

Figure 2. ROC Curve for Model 2