Machine Learning with Python

Marshall McQuillen

EDA Recap

After getting to know the data during EDA, some general statements about the data can be made with regard to survival status:

- A passenger's probability of surviving decreases as their socioeconomic status decreases.
- Given a passenger is male, he almost certainly died.
- Given a passenger is female, she most likely survived.
- Children had a higher probability of surviving than the elderly.
- A passenger's probability of surviving decreases as the number of siblings (or spouse) they have on board increases.

Machine Learning Outline

Using some of the insights (summarized above) that were extracted from the exploratory analysis, this part of the analysis will shift to machine learning, with the overarching goal being to correctly classify those who lived and those who died aboard the Titanic.

This section will be broken down into four parts:

1. **Create a Benchmark** - Logistic Regression with a few select coefficients
2. **Multi-Model Testing** - Cross validate a variety of classification algorithms, performing minor grid searches over each
3. **Algorithm Tuning and Explanation** - Choose the model with the highest accuracy, perform a more granular grid search and explain the mathematical foundations of the algorithm.
4. **Test Set Scoring** - Use the refined model on the test set.

Benchmark Model - Logistic Regression

The model that will be used as the benchmark will be logistic regression using the Age, Gender, Parch and Sibsp variables, as well as all the combinations of those variables, allowing for interaction effects to occur.

$$\hat{y} =$$