Summary of Classification Predictors on the Titanic Survivor Dataset

Background and Motivation

This analysis was performed on the Titanic survival dataset, made available via the "Titanic - Machine Learning from Disaster" Kaggle competition (Kaggle 2022). The dataset contains records of over 1300 passengers of the "unsinkable" Titanic, categorizing the "training" dataset into those that survived and those that perished. The dataset contains details of each passenger's socioeconomic status, age, gender, and familial relationships, allowing analysts to identify the characteristics most strongly correlated with surviving disaster.

Approaches to Pre-Processing the Data

The original dataset provided by Kaggle contained 11 independent variables: two numeric (fare paid and passenger age) and nine categorical variables. Data was prepared in six tranches (with each step building on the last):

- Step 0: Basic preprocessing, including filling missing values and converting data that was captured as numeric into categorical
- Step 1: Feature engineering of numerics, namely by binning age and fare into discrete categories
- Step 2: Additional feature engineering, including the creation of a "Title" variable based on passenger name, a "Deck" feature based on ticket details, and indicators about family makeup
- Step 3: Create appropriate interaction effects.

Modeling Approach

Once the dataset was processed, models were developed using different categorization modeling techniques: Logistic regression (with and without regularization), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors. All IVs were MinMax scale. The best models were then aggregated using two approaches: one, a simple average of all models other than

QDA, which had a lower validation score; and two, a stacked classifier regression which created bespoke weightings for each of the model inputs. Results:

Model	Cross-Val Accuracy	Train ROC AUC	Test Accuracy
Logistic (no regularization)	82.27%	0.899	
Logistic (ElasticNet regularization)	83.28%	0.895	76.56%
LDA	83.05%	0.897	77.03%
QDA	77.11%	0.91	
KNN	82.38%	0.922	75.60%
Simple Average			77.03%
Stacked Classifier			76.56%

Implications

- Given the same inputs, the different categorization modeling approaches produced relatively consistently accurate results in both the training (cross-validation) and test populations
- The trope of "women and children first" does appear to be borne out in the data. That said, socioeconomic status ended up being an even more important feature: class of service (First Class vs Second vs Third) is one of the stronger predictors of death, as was size of the family traveling (traveling in a 5+ person family is another strong predictor of death)
- Age binning provided a useful way to divide the population, although there are relatively surprising results, e.g., older passengers less likely to survive. This may be driven by the lack of interaction effects in the models between age and gender or age and class of service. Two considerations here:
  - For the modeling types attempted this week, we must watch out for an explosion of independent variables. Thus, the value of additional interactions needs to be offset by the risk of overfitting
  - This may be less of a concern when tackling random forest classifiers, where these relationships may be inherently captured
- Composite results (in this case, simple averaging of models) yielded improved predictive power.

References

"Titanic - Machine Learning from Disaster." Kaggle. Accessed January 24, 2022. https://www.kaggle.com/c/titanic/overview.