Executive Summary

■ Questions to investigate:

As a high jumper in high school, I was always interested in what my chances of success would have been had I stuck with it, given my body type. From Kaggle.com, a dataset was found containing information on historical data for each Olympian from the last hundred years.

The goal of this project was to establish initial models that would predict the probability of an Olympian to earn a medal while competing in the Olympics (i.e., gold, silver, or bronze), given the available predictors. Once satisfied with the initial models' accuracy, I was curious how they would perform when applied to different sports, for example, sports that were possibly less correlated with height, like swimming.

NOTE: This project's original strategy was to separate Olympian populations by the two chosen sports/events high jump and swimming, then further separate by sex and establish two predictive models for each of the four groupings. Unexpectedly, during data exploration, it was discovered that certain variables were weighted differently depending on sex. For example, Age was more significant for females and Height more significant for males in high jump. This interesting finding would interject a more complicated analysis than originally planned for, at this point it was decided to focus on only male Olympians from both sports for this project.

■ Purpose of Analysis:

Many organizations would benefit from accurate predictive models like these. One organization in particular, would be The U.S. Olympic & Paralympic Committee, or Team USA. This model could be used for scouting and training current Olympians. Further applications could be applied with other, more refined data sets, for high school athletes at an early stage to help them decide which events/sports to try. Yet another application of a successful model could be applied to online sports betting. With the advent of sport gambling legalization (in 2018 the Supreme Court ruled the Professional and Amateur Sports Protection Act (PASPA) unconstitutional), online sportsbook companies such as DraftKings or BetMGM could use such a model to set their odds.

■ Analysis:

Beginning with exploration of the high jump event to refine models, two approaches were chosen: 1) logistic regression and 2) artificial neural networking (ANN).

These two methods are appropriate and preferred because they were able to help guide which variables were more important (or significant) and if they should be included or not (i.e., p-value, edge weight, Garson's algorithm, etc.). Binary logistic regression was chosen because of its flexibility. Binary logistic regression is fairly non-parametric (meaning data requires few assumptions), assuming only that the variables are independent, little or no multicollinearity exists between predictors, the data is a relatively large sample size (>10 obs.), and for the response variable to be binary. This last parameter was satisfied by combining all the medals together (did or did not medal). Like logistic regression, ANNs also work

best with a large enough data set of non-linear data to train it on. On initial exploration of the variables, Height and Weight appeared slightly collinear (related to one another) but proved not to affect the final logistic models (the less significant Weight variable was regressed with the other predictor variables and compared using in various model versions but did not influence final misclassification errors).

High Jump

The model found to be most accurate for the high jump event was logistic regression (see Figure 1) with the predictor variables Age, Height, Weight (residuals), Team, and Year. This produces an area under the curve of approximately 0.98 (and a misclassification rate of 0.05, or 5.2%). This means the highest performing logistic regression model is approximately 95% accurate at predicting whether an Olympian competing in high jump will medal once they make the Olympics.

For comparison, an artificial neural network was created using the same data. This approach's best model also performed well, with a similar misclassification rate of about .052 (5.2%).

After the high jump model was established, it was refit to Olympic swimmers to understand the difference in variable influence between the two sports. As can been seen from Figure 1, using the area under the curve (AUC) to determine model fit, the variable *Team* is most influential (noted by the jump between the blue lines and yellow lines) for both sports/events. Year also has a strong influence on the models, but not as heavily in swimming as high jump (look at the model with the lowest AUC in each sport). In fact, for swimming, the model using only Year as a predictor did not predict any of the athletes from the dataset as medaling. This means the model is barely better than random guessing. On the other hand, a model using only *Team* as a predictor is better than any other single variable model (High Jump AUC: 0.93, Swimming AUC: 0.91). One could easily argue using a model with only Team as a predictor since it is simplest (most parsimonious).

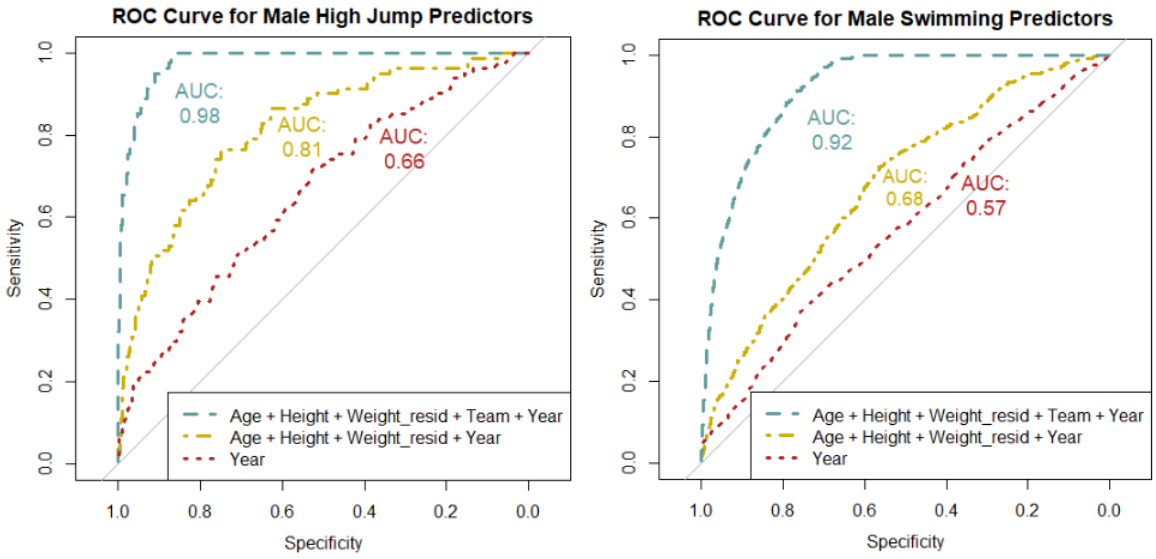

Figure 1 - Receiver operator characteristic (ROC) curve summarizes the trade-off between true positive rate and false positive rate. More effective models for predicting chances of medaling have curves closer to the upper-left corner.

Swimming

Unlike with high jump, the neural network model (ANN) was found to be most predictive for swimming. The best ANN model had a slightly improved (lower) misclassification rate of 0.063 compared to the logistic regression model's 0.070. With this sport, again *Team* was found overwhelmingly to be the most important variable. Removing Team from the ANN gives insight to the relationship among the remaining, less influential variables using what are called a Garson's Algorithm and Lek Profile (see Figure 2). Of the remaining variables, the Garson plot easily displays order of variable importance for the predictive ability of the neural network. Moving to the Lek profile, we get a better idea of the relationships between the variables. For example, we can see there is a sharp peak for in the 80th percentile for Age strongly associated with medaling. Note also that overall a male swimmer would want to be in the among the lowest percentiles for Age (youngest), Height (shortest), and Weight (weigh least) to maximize chances of medaling.

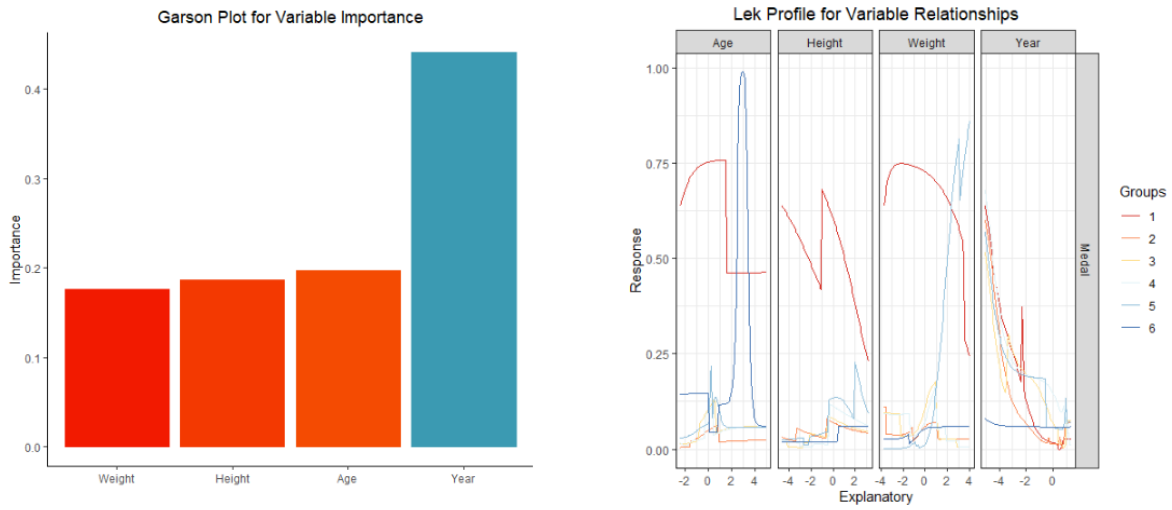

Figure 2 - Garson's algorithm plot (left) with Team variable removed to discern remaining variable importance more easily. On the right is the ANN model's Lek Profile to show the relationship between each predictor variable aside from Team and response.

■ Conclusions:

These models provide compelling conclusions to apply to future athletic datasets. Here, only two sports were examined for predictive ability; further research could be done to fit models to the remaining 64 sports as well as comparisons with female athletes. Examining these types of variable relationships would, no doubt, uncover new insights for success in competition.

The most effective predictive approach varied slightly for the two sports selected. For high jump, logistic regression and ANN were nearly identical at predicting which Olympians would medal (94.8% accuracy). As seen in Figure 1, AUC was used as a complement for assessment of fit to ensure this model's outperformance of a "no-information" model, or simply saying nobody/everybody medaled. For swimming, the best ANN model proved to be slightly more accurate than logistic regression in predicting which Olympians would medal (93.7% accuracy). Both models however, used the same predictors to get the most accurate model: Age, Height, Weight, Team, and Year. This information is useful when evaluating new data, especially in other sports, because if the model is found to be overfitting to the training data, the number of variables can be reduced, starting with the least important.