Midterm 1

DS 740

Lyndsey Pohl

A model predicting the type of sport an athlete plays was developed through investigating multiple methodologies. In particular, this model predicts whether or not the sport played by an athlete uses some sort of ball in thinking that these sports might require a different set of skills such as balance or hand-eye coordination. The entire model selection process does improve the prediction process, as opposed to guessing which type of sport is played. The final model that was chosen employed logistic regression for 12 predictor variables. The methodology for building this model is detailed below.

A dataset with measurements on 202 athletes was used for building the model. The measurements on the athletes were: sex, height (Ht), weight (Wt), lean body mass (LBM), red cell count (RCC), white cell count (WCC), hematocrit (Hc), hemoglobin (Hg), plasma ferritin concentration (Ferr), body mass index (BMI), sum of skin folds (SSF), percent body fat (Bfat), and case labels (Label). The sports played by the athletes were: bball, netball, 400m run, sprint, row, swim, water polo, tennis, gym, and field. Due to the number of predictors being higher than some observations in each sport, the sport categories were grouped into sports using some sort of ball and those that did not. The sports with balls were: bball, netball, tennis, water polo, and field; and the sports with no balls were 400m run, sprint, row, gym, and swim. Along with this, some data cleaning was performed, any observations with missing measurements were omitted from the dataset, resulting in a slightly reduced dataset of 187 observations.

First examination of the distribution between the response variable (type of sport played) for each predictor showed that there were noticeable differences for Bfat, Wt, WCC, SSF, Hc, Hg, and BMI; suggesting that these may be important factors. Along with this, all distributions looked fairly normal, with only a few outliers in some predictors. Further investigation showed that approximately 9% of variables were highly correlated with other, suggesting that correlation was not a huge problem with this dataset.

The methods studied prior to assessing the model selection process included logistic regression, KNN, discriminant analysis, and decision trees. Some of the assumptions for discriminant analysis included same covariant matrices between the two categories of sports for LDA and normal distributions of each sport category for each predictor for QDA. Neither of these assumptions were met, and it would have been a large assumption to overlook these. In this case, discriminant analysis was not further used.

The decision tree models used in the model assessment process used the normal tree method with no pruning, bagging, or boosting. Another model used pruning to result in the optimal number of leaves at each iteration of the process. The third model used bagging and the fourth used random forests. As might be expected due to the low level of correlation between the predictors, there was no noticeable change in efficiency between the bagging and random forests methods, and both methods predicted SSF and Hg as important for accuracy and WCC and SSF important for node purity. Boosting was not used in the model assessment process.

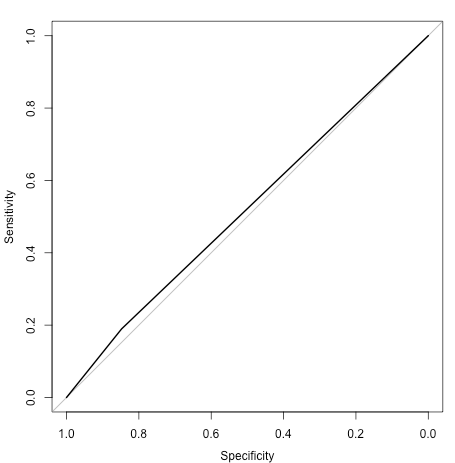
Logistic regression was also used in the model assessment process. Two models using this method were used: one with all 12 predictors, and one with only statistically significant predictors RCC, WCC, SSF, and Hc. A preliminary assessment of the models showed that there was only a slight increase in performance when using only the significant factors, but not by much.

KNN was the third type of method investigated in this study. Due to the amount of data prep needed for each iteration of the assessment process, only predictors WCC and SSF were used with this method. These predictors were used as there were highlighted as being important in the bagging decision tree and logistic regression methods.

Double cross validation was used to assess the model selection process for these seven models. Ten folds were used in this process, and the logistic regression model with all 12 predictors was picked as the ‘best model’ at each iteration. This assessment lead to an accuracy of approximately 72%, quite an increase from ‘guessing’ at which type of sport a player would play which gave an accuracy of approximately 50%.

The logistic model using all 12 predictors was then fit using the given data and a cross validation process with 10 folds. Since the logistic model gave a number between 0 and 1, a threshold value needed to be determined to find the best balance between true positives and negatives. An analysis of this showed that 0.4 was the ideal number that resulted in the highest area under the ROC curve (meaning that this yielded the best combination of high true positive rate and low false negative rate). Although this model was chosen as the ‘best model’ during the assessment process, the final fitted model only gave an accuracy of 51%, about the same as guessing the categories of the players. Figure 1 shows the true positive and false negative rates for the model, highlighting that this model is only slightly better than guessing.

Since this model gives such a low accuracy, it is not recommended for use. This does suggest that there is not enough variation between the two sports categories for the majority of the predictors. It is recommended that if a better model to predict whether or not an athlete plays a sport with some sort of ball is desired that different measurements (i.e. predictors) are used.



**Figure 1:** ROC curve for final fitted model using logistic regression and a threshold value 0.4.