Can We Predict the Next Olympian?

What is the RBC Training Grounds?
The RBC Training Grounds offers many young athletes the opportunity to discover their true potential. Participants between the ages of thirteen and thirty are eligible to attend the RBC Training Grounds event which consists of a series of tests measuring speed, power, endurance and strength. Each participant competes against the benchmarks set by eleven National Sport Organizations (NSO) which are used to determine an athlete's potential for an Olympic medal.

To participate in this program, participants can attend one of many qualifiers taking place throughout Canada. At each qualifier, participants will be graded based on athletic performance and will compete against each other to attend a regional finals. The top ten athletes from each test in the preliminary round (based on z-score) will be invited to participate in the final round of testing (final test). From the results of the final test, different sports will nominate participants for funding, resources, and coaching with the goal of being able to compete on the national team.

Data Collection Methodology
The RBC training grounds data was supplied by the Canadian Sports Institute, which contains information on last years participants. The data contains results on preliminary tests which include vertical jump, 10m sprint, 30m sprint, and isometric mid-thigh pull, as well as results from the final tests being 10m sprint, 30m sprint, 40m sprint, 30-40m sprint, upper body pull, upper body push, lower body pull, triple jump, single jump, arm/leg bike, and relative six second peak power measurements. The data also contained anatomic measurements of each participant which included height, weight, wingspan and gender. The participants were given ID numbers to track their results at each stage of the event. The data used within this analysis came from the RBC training ground events in Ontario, British Columbia, Quebec, Alberta, and Atlantic Canada as a whole.

Objective / Primary Questions of Interest
The overall objective of this data analysis is to predict the next Olympian using the data from prior RBC training ground participants.

The primary focuses are:

1. Predicting who will be nominated based on final test results.
2. Providing insight as to which test measurements are most important for increasing the chances of nomination.

Exploratory Data Analysis
This section will focus on exploring the data to see how the questions of interest can be answered.

The analysis had been conducted with the data being partitioned into two sets: male and female participants. This was done since sport teams are separated by gender and hence, nominations for different sports will be based on team selection for each gender. Given this, there were two hundred and seventy-four and one hundred and two male and female participants, respectively, which made it to the final testing round. Of these participants, one hundred and ten and sixty-six males and females, respectively, were nominated.

To begin, some important results from the exploration of preliminary data were concluded below.

The mean and standard deviations of test score measurements for both nominated and not nominated males and females are given below in **Tables 1** and **2**, followed by **3** and **4** respectively.

Table 1. Mean and standard deviations of final test measurements for males that were nominated.

	mean	standard deviation
triple broad jump (metres)	7.97	1.05
single broad jump (metres)	2.58	0.32
relative 6 sec. peak power	15.57	2.28
upper body pull	456.60	153.01
upper body push	354.22	116.62
lower body push	700.70	214.97
0-30m sprint (secs)	4.25	0.27
0-10m sprint (secs)	1.76	0.11
0-40m sprint (secs)	5.44	0.37
30-40m sprint (secs)	1.19	0.11
arm, leg, bike	70.56	6.03
height (cm)	176.41	19.76
weight(kg)	74.64	11.72
wingspan(cm)	183.50	12.00

Table 2. Mean and standard deviations of final test measurements for males that were not nominated.

	mean	standard deviation
triple broad jump (metres)	7.83	1.24
single broad jump (metres)	2.54	0.38
relative 6 sec. peak power	15.50	2.75
upper body pull	397.46	143.93
upper body push	321.54	104.30
lower body push	629.09	194.58
0-30m sprint (secs)	4.30	0.36
0-10m sprint (secs)	1.78	0.11
0-40m sprint (secs)	5.50	0.59

30-40m sprint (secs)	1.20	0.28
arm, leg, bike	67.66	8.32
height (cm)	174.54	9.10
weight(kg)	70.21	12.93
wingspan(cm)	179.84	11.37

Table 3. Mean and standard deviations of final test measurements for females that were nominated.

	mean	standard deviation
triple broad jump (metres)	7.15	1.33
single broad jump (metres)	2.30	0.39
relative 6 sec. peak power	13.12	2.39
upper body pull	404.22	173.73
upper body push	323.92	132.19
lower body push	653.83	246.41
0-30m sprint (secs)	4.59	0.44
0-10m sprint (secs)	1.89	0.24
0-40m sprint (secs)	5.90	0.60
30-40m sprint (secs)	1.31	0.16
arm, leg, bike	68.72	5.37
height (cm)	176.71	10.88
weight(kg)	74.81	15.46
wingspan(cm)	182.00	12.70

Table 4. Mean and standard deviations of final test measurements for females that were not nominated.

	mean	standard deviation
triple broad jump (metres)	6.94	1.20
single broad jump (metres)	2.27	0.37
relative 6 sec. peak power	13.72	2.62

upper body pull	325.39	135.90
upper body push	263.33	105.59
lower body push	538.95	191.04
0-30m sprint (secs)	4.61	0.40
0-10m sprint (secs)	1.89	0.16
0-40m sprint (secs)	5.92	0.54
30-40m sprint (secs)	1.31	0.15
arm, leg, bike	64.91	10.29
height (cm)	170.90	9.49
weight(kg)	66.48	14.95
wingspan(cm)	175.27	11.96

By conducting t-tests on each test measurement, the lower body pull and upper body pull test measurements showed significant differences in means between the males that were nominated and those that were not nominated, as seen in **Figure 1**. Lower body push, upper body push, and height measurements showed significant differences between the females that were nominated and those that were not nominated; **Figure 2** displaying these differences.

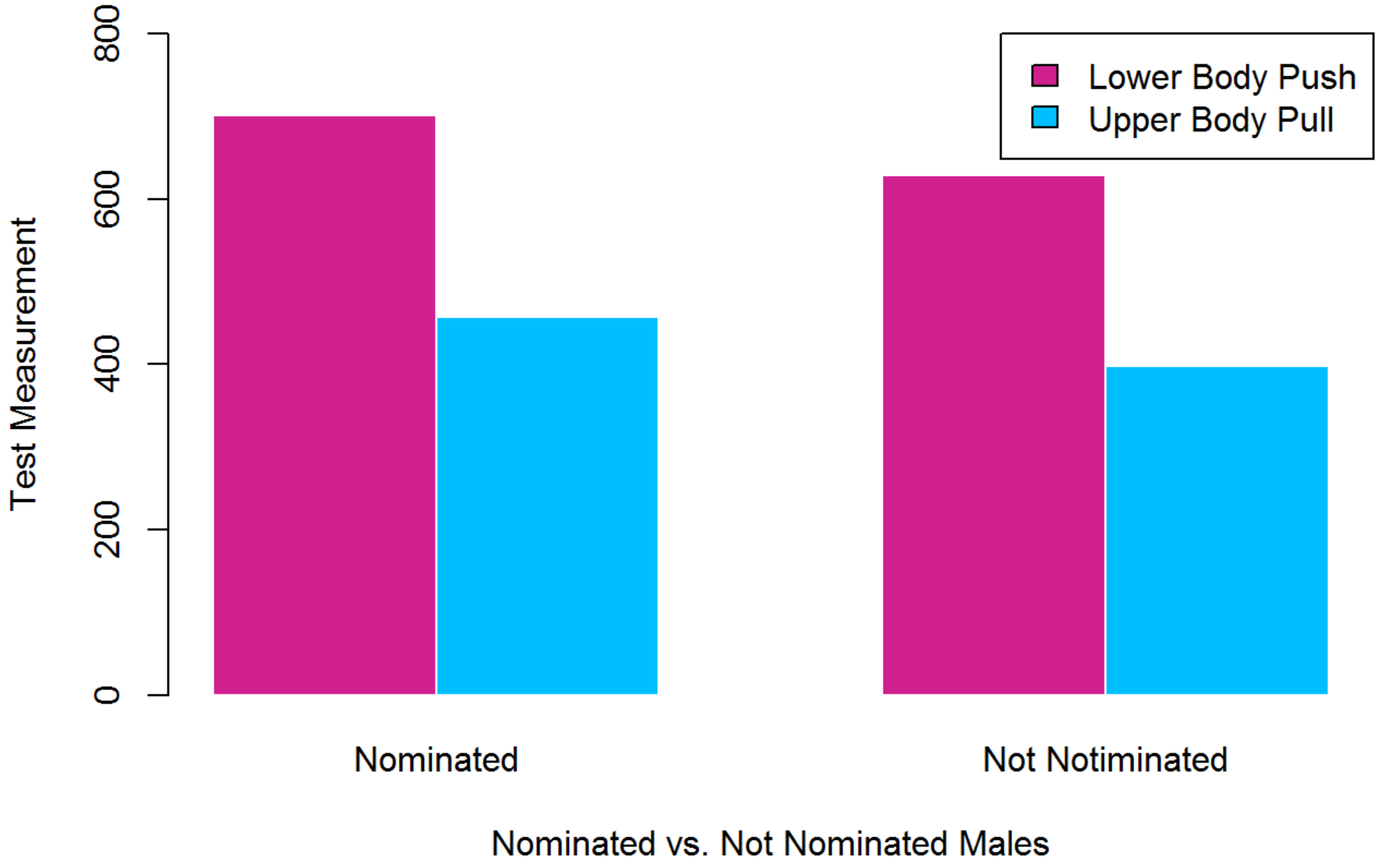

Figure 1. Test measurements that had significant differences in means between nominated and not nominated males

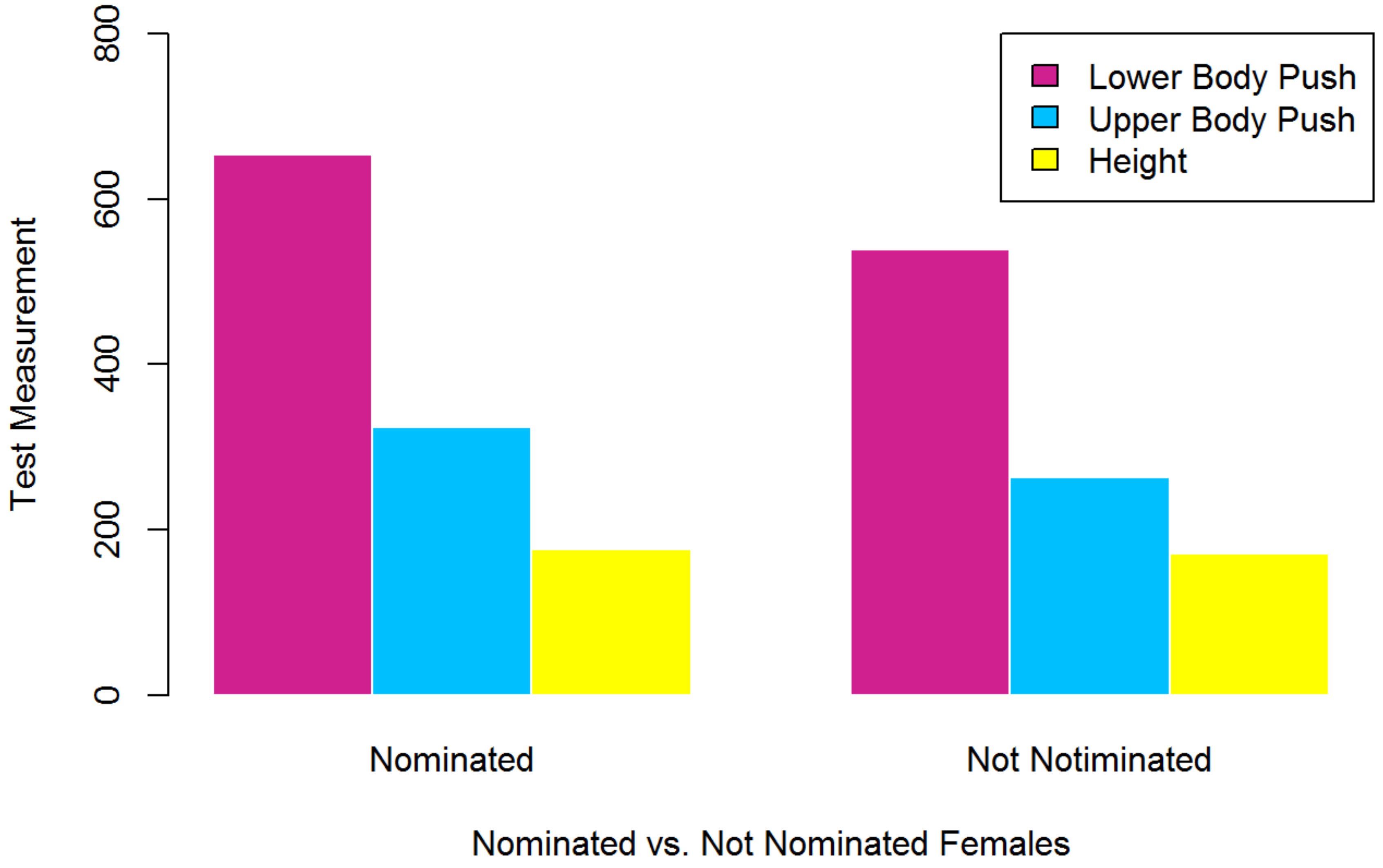

Figure 2. Test measurements that had significant differences in means between nominated and not nominated females

There was also significantly high correlation found between test measurements. In particular as seen in **Figure 3**, single and triple jump had high negative correlations with 0-10m sprint, 0-30m sprint, 0-40m sprint, and 30-40m sprint. As well, each sprint test had high positive correlations. **Figure 4** shows the high positive correlations between upper body pull, lower body push and upper body push and **Figure 5** displays the high negative correlation that was also found between relative six second peak power and the sprint test measurements. Based on each particular movement, these results could imply that an individual's genetic predisposition for limb length could influence the correlation since each movement requires different fulcrum lengths to be advantageous in each of the test movements. For example, individuals with longer limbs would tend to score higher in sprint versus jump tests since their stride length would be longer allowing for more distance to be covered in less time. Conversely, their jump test could be limited from the same physiological perspective due to the overall length of their legs increasing the amount of weight that would have to be carried through the motion. This is similar to the concept that heaiver people tend to have more difficulty doing pull-ups than lighter people.

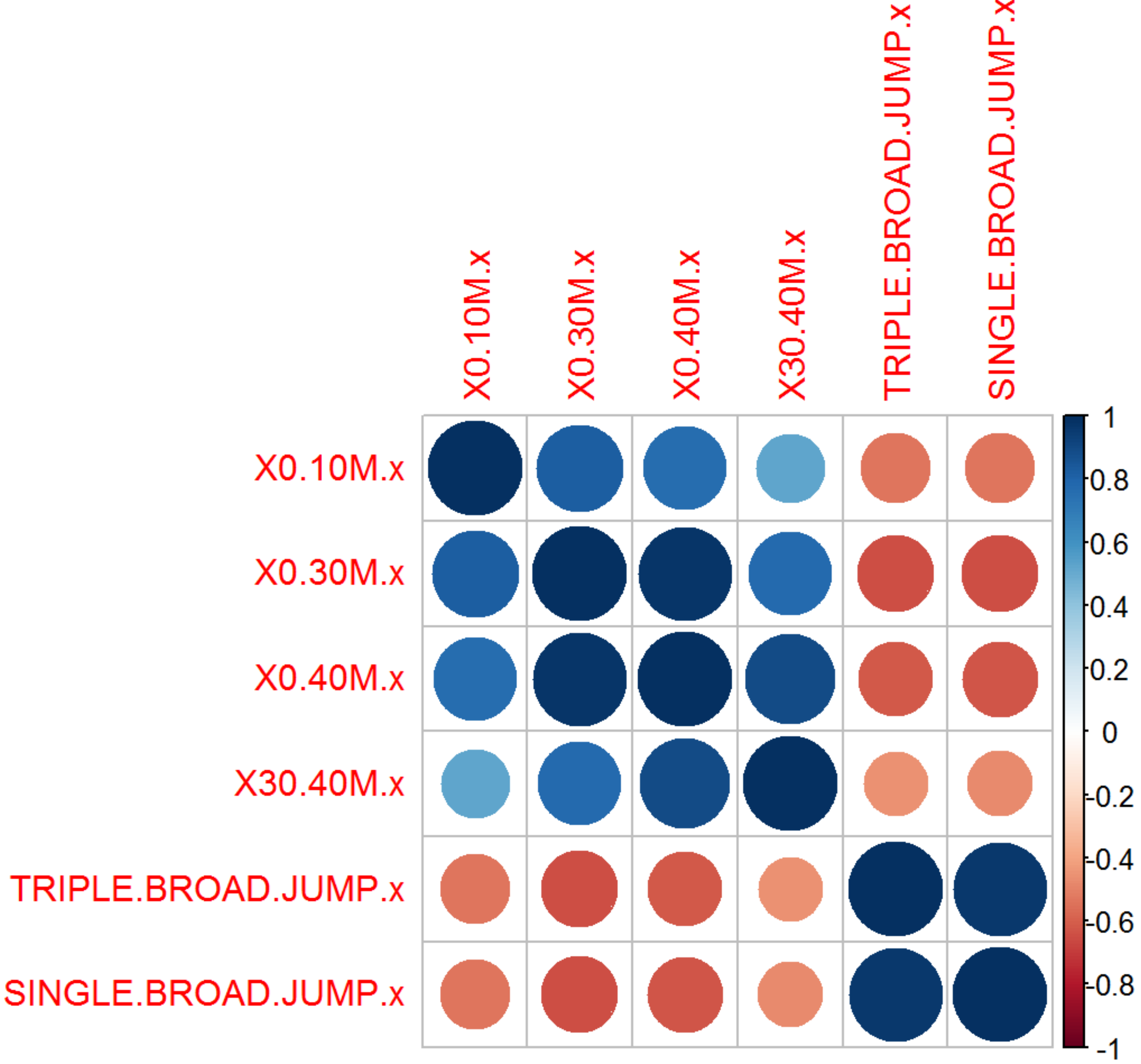

Figure 3. Correlation plot showing correlation between single jump, triple jump, 0-10m sprint, 0-30m spint, 0-40m sprint, and 30-40m sprint

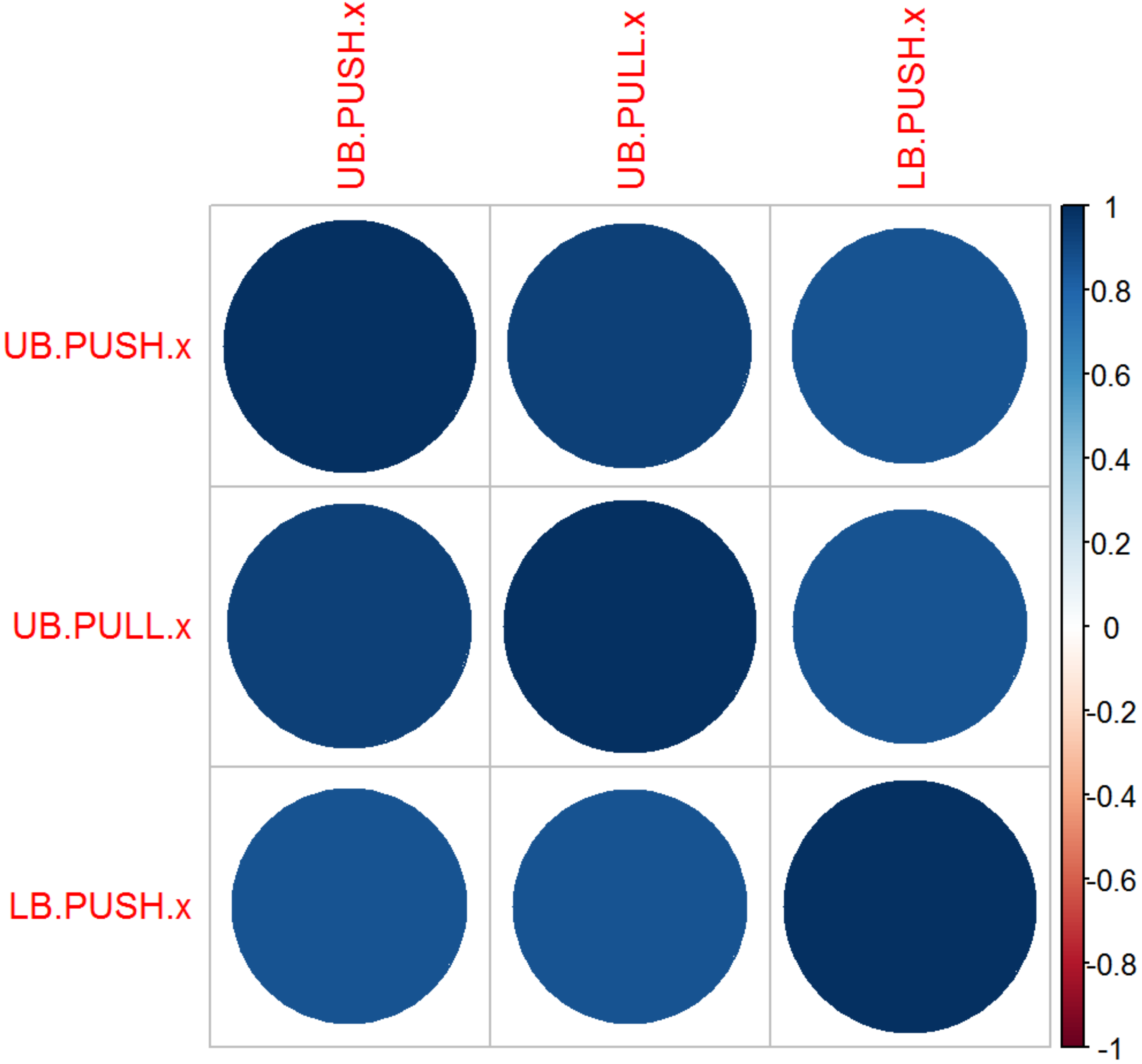

Figure 4. Correlation plot showing correlation between upper body pull, lower body push and upper body pull

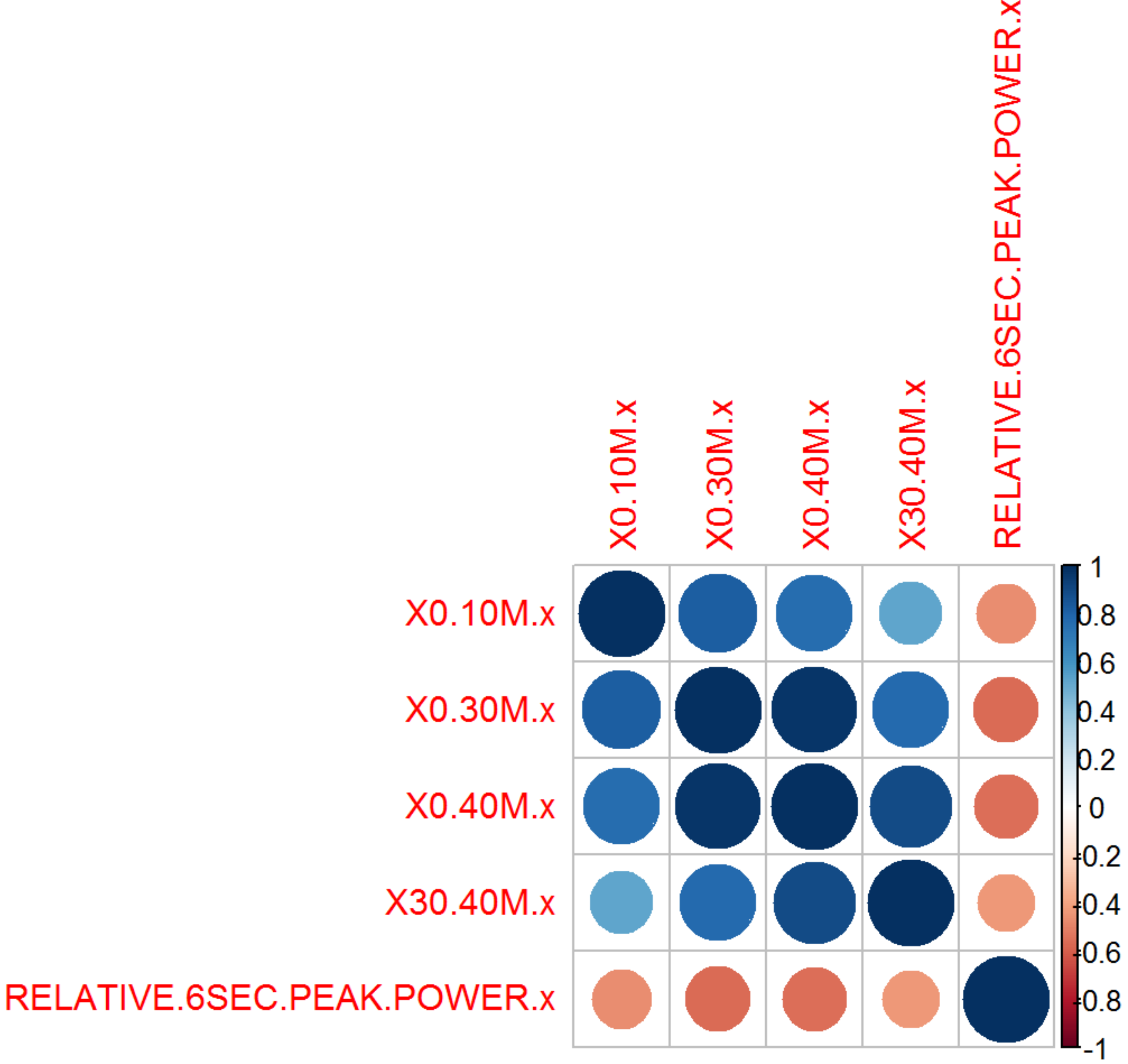

Figure 5. Correlation plot showing correlation between relative six second peak power and the sprint test measurements

To summarize, the results from the correlation tests seem to agree with the logical assumptions to be made about the relationships between these physical tests. In particular, as sprint time increases, natutally this means the athlete is slower implying their triple jump distance would decrease as they are not able to generate enough power to move themselves forward further in the short period of time. The same holds for the relative six second peak power test. Longer short distance sprinting times means less short bursts of initial power is available to the athlete. Clearly upper body push and pull positively improve each other as well as assist lower body pushing power.

What follows is the prediction model analysis focusing on answering the questions of interest. The above preliminary analysis was taken into account and proved to be beneficial as seen in later discussion.

To predict who will be nominated based on the final test results of participants, two different types of prediction models were compared for each gender to see what sort of test measurement criteria is needed to increase the chances of being nominated. Both models were constructed by randomly partitioning subsets of the male and female athletes RBC training grounds data as “training” and “testing” data, respectively. The training set was used to build the model and the testing set was later used to see how well the models performed for prediction.

The first model is a logistic regression model with all test measurements completed in the final testing round as predictors. In particular, the final test measurements used for prediction were triple broad jump, single broad jump, relative six second peak power, upper body pull, upper body push, lower body push, 0-30m sprint, 0-10m sprint, 0-40m sprint, arm/leg bike total, height, weight, wingspan, and age.

The logistic regression model has the following form for each gender:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0 + \beta_1 X_{triplejump} + \beta_2 X_{singlejump}} \\ + \beta_3 X_{relative6secpeakpower} + \beta_4 X_{upperbodypull} + \beta_5 X_{upperbodypush} \\ + \beta_6 X_{lowerbodypush} + \beta_7 X_{0-30msprint} + \beta_8 X_{0-10msprint} \\ + \beta_9 X_{0-40msprint} + \beta_{10} X_{arm/legbike} + \beta_{11} X_{height} \\ + \beta_{12} X_{weight} + \beta_{13} X_{wingspan} + \beta_{14} X_{age}$$

where π_i is the probability of participant i being nominated.

More specifically, using the training set of data, the following predictor estimates, in **Table 5** and **Table 6**, were obtained for each of the final tests for both the male and female participants:

Table 5. Estimates for male participants full logistic regression model. AIC approximately 328.

Predictors	Estimates	SE	z score	p-value
(Intercept)	7.427	5.161	1.439	0.150
TRIPLE.BROAD.JUMP.x	-0.434	0.482	-0.902	0.367
SINGLE.BROAD.JUMP.x	0.494	1.551	0.318	0.750
RELATIVE.6SEC.PEAK.POWER.x	-0.067	0.076	-0.884	0.377
UB.PULL.x	0.007	0.003	2.544	0.011
UB.PUSH.x	-0.006	0.004	-1.507	0.132
LB.PUSH.x	0.001	0.002	0.425	0.671
X0.30M.x	0.816	1.890	0.432	0.666
X0.10M.x	-3.978	2.016	-1.973	0.048
X0.40M.x	-0.331	1.002	-0.331	0.741
ARM.LEG.BIKE.x	0.070	0.033	2.112	0.035

HEIGHT.x	-0.025	0.019	-1.335	0.182
WEIGHT.x	0.001	0.026	0.057	0.955
WINGSPAN.x	-0.002	0.024	-0.081	0.936
AGE.x	-0.061	0.047	-1.294	0.196

Table 6. Estimates for female participants logistic regression model. AIC approximately 118.

Predictors	Estimates	SE	z score	p-value
(Intercept)	8.563	14.341	0.597	0.550
TRIPLE.BROAD.JUMP.x	1.217	0.933	1.304	0.192
SINGLE.BROAD.JUMP.x	-3.229	3.351	-0.963	0.335
RELATIVE.6SEC.PEAK.POWER.x	-0.718	0.252	-2.853	0.004
UB.PULL.x	-0.008	0.007	-1.178	0.239
UB.PUSH.x	0.015	0.009	1.579	0.114
LB.PUSH.x	0.004	0.003	1.318	0.187
X0.30M.x	-18.425	10.838	-1.700	0.089
X0.10M.x	2.045	2.799	0.730	0.465
X0.40M.x	12.679	7.593	1.670	0.095
ARM.LEG.BIKE.x	0.101	0.086	1.185	0.236
HEIGHT.x	-0.040	0.076	-0.527	0.598
WEIGHT.x	-0.015	0.049	-0.312	0.755
WINGSPAN.x	0.011	0.056	0.197	0.844
AGE.x	0.039	0.096	0.403	0.687

It is worth noting however, that as observed in the preliminary data analysis, the above group of predictors are highly correlated and hence, the predictors’ estimates do not provide stable estimates to how each specific test measurement will influence the probability of being nominated. Therefore, this model is best to be used to predict the chances of being nominated given all test measurements. Specific influences of test measurements will be discussed.

To test the accuracy of this model, the test data set was used to determine whether the model would predict the right conclusions about the participants in the test set; whether it was accurate in determining whether they were nominated or not. To do this, a Receiver Operating Characteristic Curve (ROC) was plotted, as seen in **Figure 6** and

Figure 7, to see how well the models predictions were in comparison to the actually results from the test data set participants.

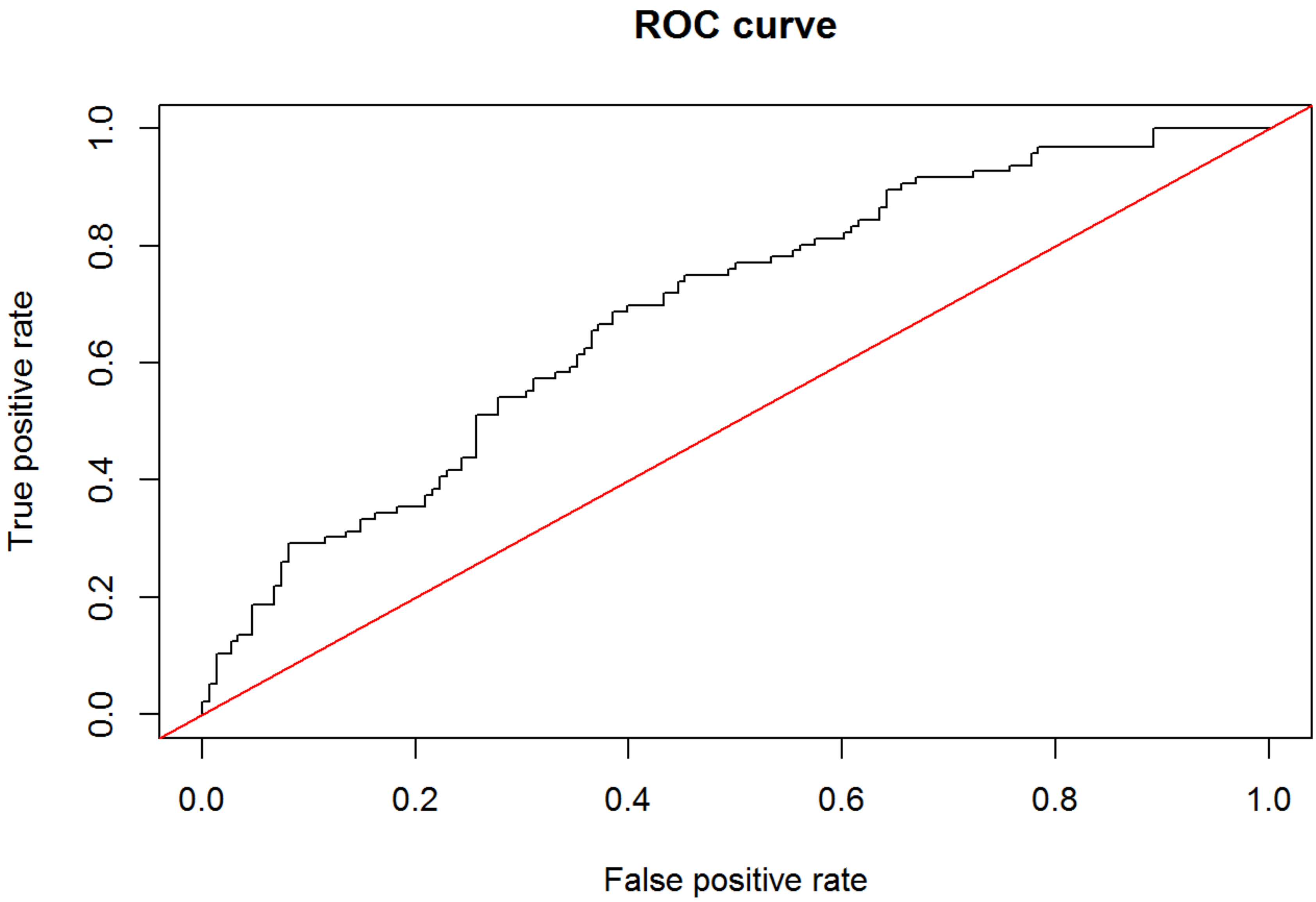

Figure 6. The black curve indicates the ROC of the full logistic regression for male participants.

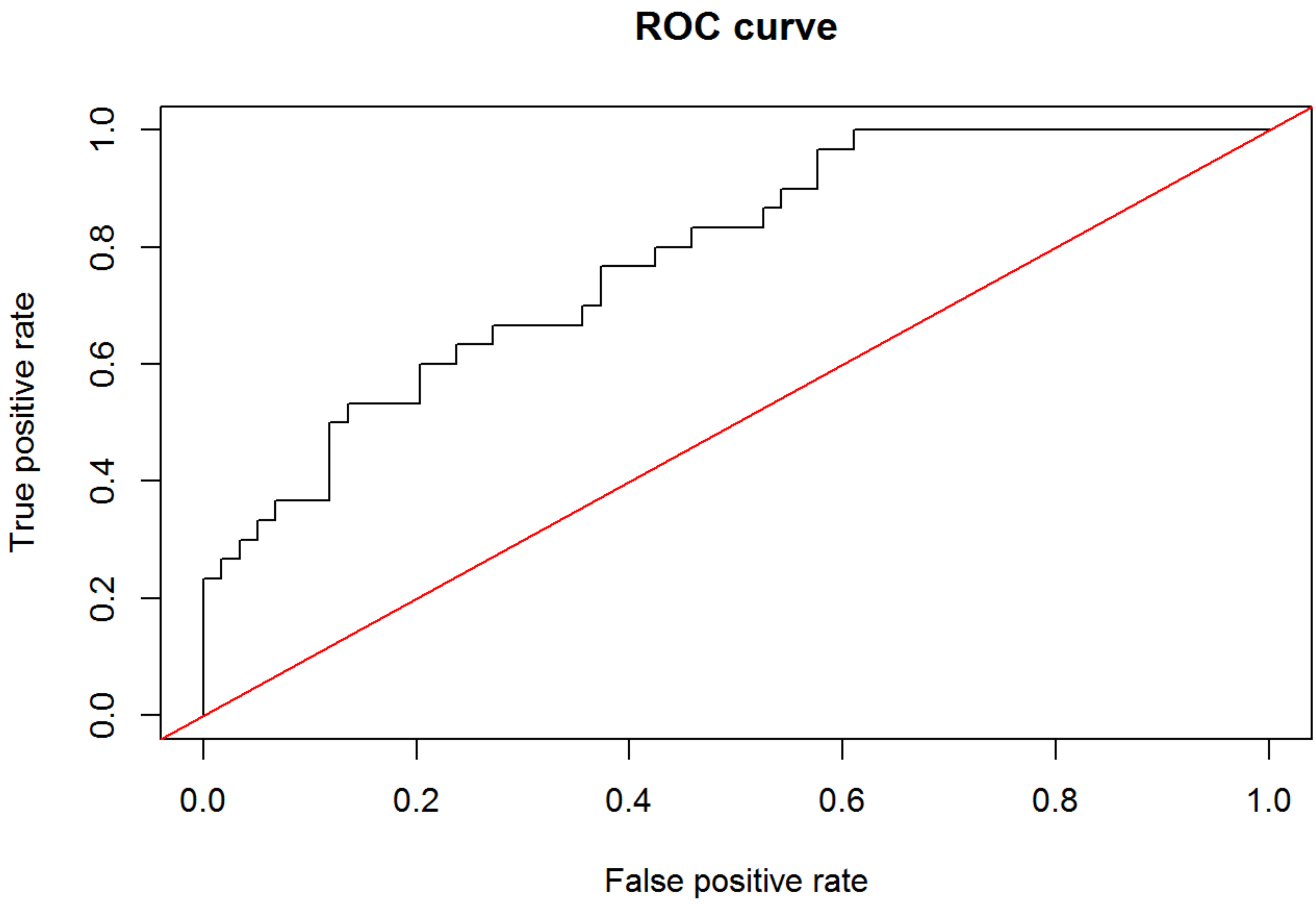

Figure 7. The black curve indicates the ROC of the full logistic regression for female participants.

With the ROC varying at different thresholds (i.e. different cut-offs needed to be classified as nominated), the plot implies how well the model is able to predict the outcome of a participant at different thresholds. From the results, the model is better than simply guessing (the red curve). By calculating the Area Under the Curve (AUC), the male model gives an accuracy of 69% meaning, the model got the classification of the participate (i.e. being nominated or not) right 69% of the time. The female model gives an accuracy of 78% based on the AUC which indicates that the prediction model works better for predicting female participant nominations compared to males nominations. However, this could be due to the fact that there is a significantly smaller population of female participants in the data set; in particular only 27% of the data is female. For both male and female prediction models, since high correlation exists between the predictors this may also be a factor as to what is lowering the prediction accuracy; the model may be slightly over fitted to the training data.

To see if this problem could be improved, many reduced logistic regression models were fitted and it was found that using only a specific combination of predictors for males and females proved to make the most logical sense. The reduced male model uses age, weight, arm/leg bike, 0-10m sprint, and upper body pull, whereas the reduced female model uses height, age, arm/leg bike and upper body pull. The reason for these predictors was due to their overall significance when applied individually to a logistic regression, taking into account the best predictor in each group of significantly correlated predictors. As well, the interaction terms for all of the chosen predictors for each male and female model were taken into consideration and only the significant interactions terms were included in the male and female reduced models, respectively. In particular, the interaction between arm/leg bike and upper body pull was significant for the male reduced model. The interaction between height and arm/leg bike was significant for the female reduced model. To summarize, below are the general forms for the male and female reduced models.

Male reduced model:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_{arm/legbike}+\beta_2 X_{0-10msprint} + \beta_3 X_{weight} + \beta_4 X_{upperbodypull} + \beta_5 X_{age} + \beta_6 X_{arm/legbike*upperbodypull}}$$

where π_i is the probability of the ith male participant being nominated.

Female reduced model:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_{upperbodypull}+\beta_2 X_{Arm/legbike} + \beta_3 X_{height} + \beta_4 X_{age} + \beta_5 X_{height*arm/legbike}}$$

where π_i is the probability of the ith female participant being nominated.

The estimates for each of these models respectively in displayed in **Table 7** and **Table 8**.

Table 7.Reduced model with uncorrelated, individually significant predictors for males. AIC approximately 324.

Predictors	Estimates	SE	z score	p-value
(Intercept)	3.593	5.601	0.641	0.521
ARM.LEG.BIKE.x	-0.009	0.068	-0.139	0.890
X0.10M.x	-1.509	1.367	-1.104	0.270
WEIGHT.x	-0.005	0.022	-0.209	0.834
UB.PULL.x	-0.008	0.011	-0.684	0.494
AGE.x	-0.070	0.045	-1.559	0.119
ARM.LEG.BIKE.x:UB.PULL.x	0.000	0.000	0.882	0.378

Table 8.Reduced model with uncorrelated, individually significant predictors for females. AIC approximately 119.

Predictors	**Estimates**	**SE**	**z score**	**p-value**
(Intercept)	28.541	50.382	0.566	0.571
HEIGHT.x	-0.193	0.297	-0.650	0.516
ARM.LEG.BIKE.x	-0.518	0.755	-0.685	0.493
UB.PULL.x	0.000	0.003	-0.142	0.887
AGE.x	0.026	0.062	0.420	0.674
HEIGHT.x:ARM.LEG.BIKE.x	0.003	0.004	0.748	0.454

To again test the accuracy of these models, the test data sets were used to determine whether the models would predict the right conclusions about the participants in the test sets. The ROC is displayed in **Figure 8** for the male model and **Figure 9** for the female model.

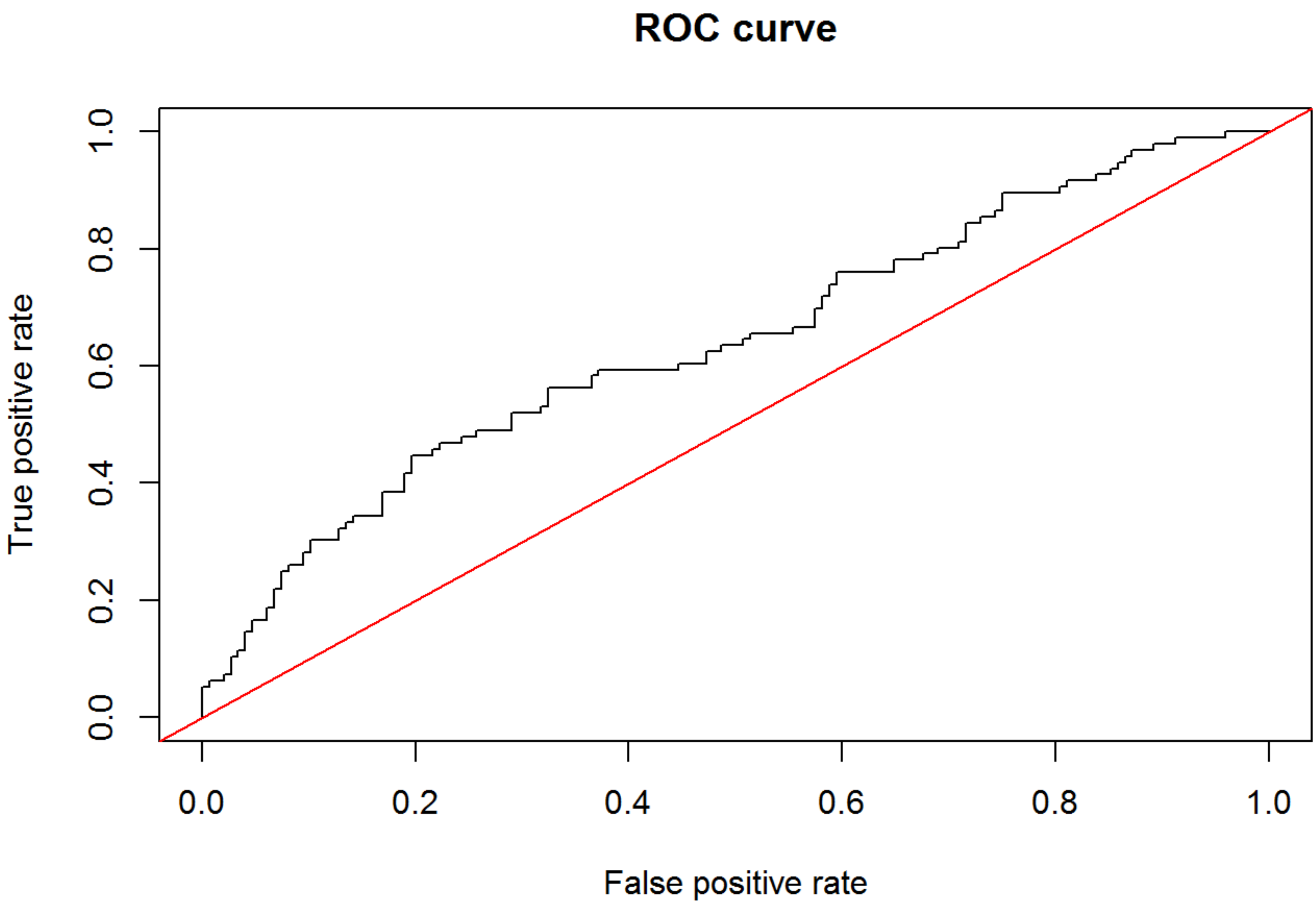

Figure 8. Reduced male model ROC curve.

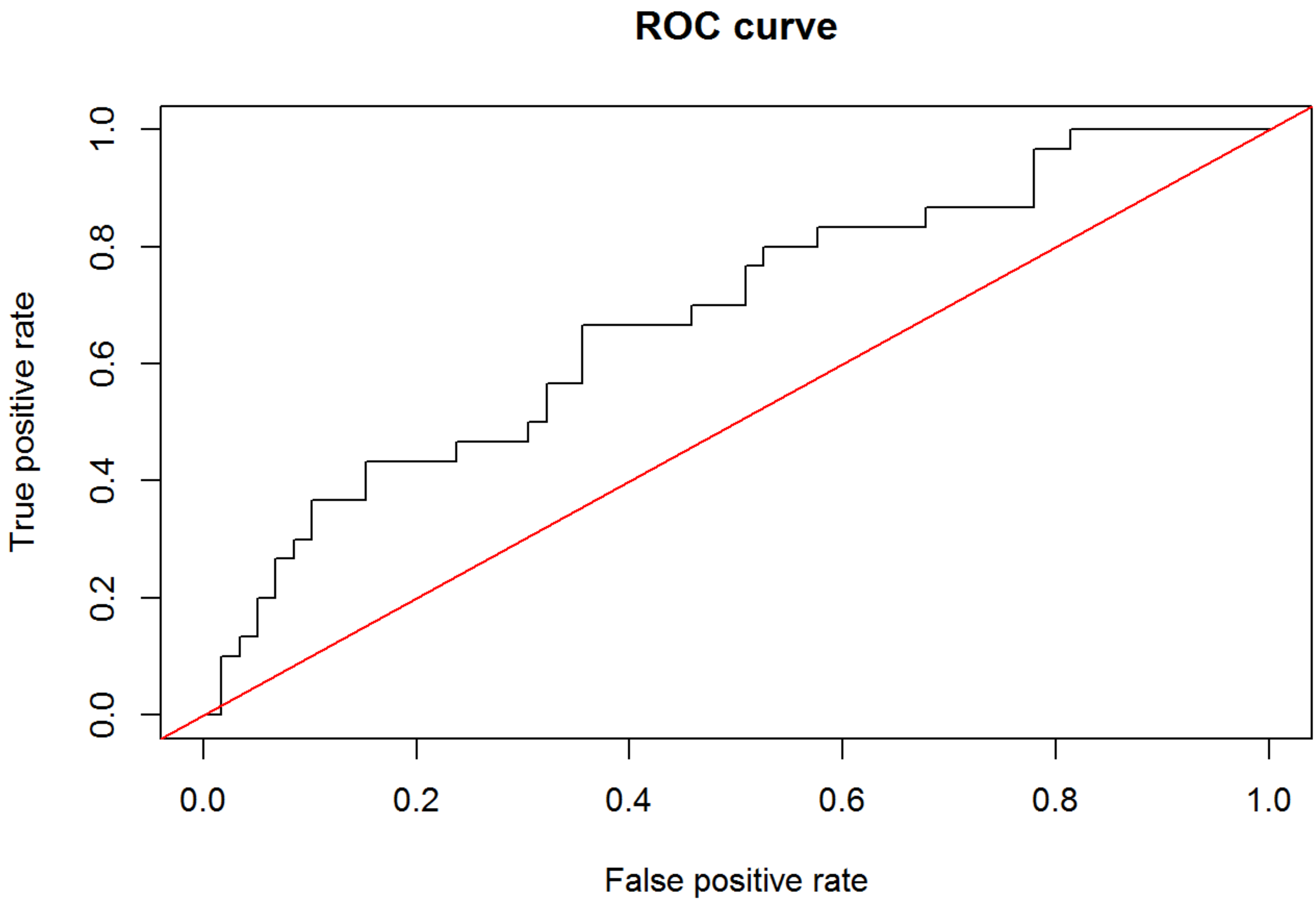

Figure 9. Reduced female model ROC curve.

Both reduced models performed worse in terms of accurate predictions compared to the full models. The male reduced model had an AUC of 64% and the female reduced model had an AUC of 68%. Even though the reduced models make more sense to use since there is less of a over fitting problem and no correlated predictors, in terms of being able to accurately predict whether an athlete will be nominated, the full models work better. Although the AIC value was slightly lower for the male reduced model implying that this model fitted the male data better, the same could not be said for the female reduced model. A Tree Classifier model method was then explored to see whether prediction accuracy could be further improved.

By using decision tree classifiers, it is possible to ignore the influences of highly correlated predictors hence, all test measurements were considered to grow the tree. For consistency purposes, the same training and set data sets were used to grow and validate the male and female trees respectively. Below **Figure 9** shows the male participant decision tree and **Figure 10** shows the female participant decision tree.

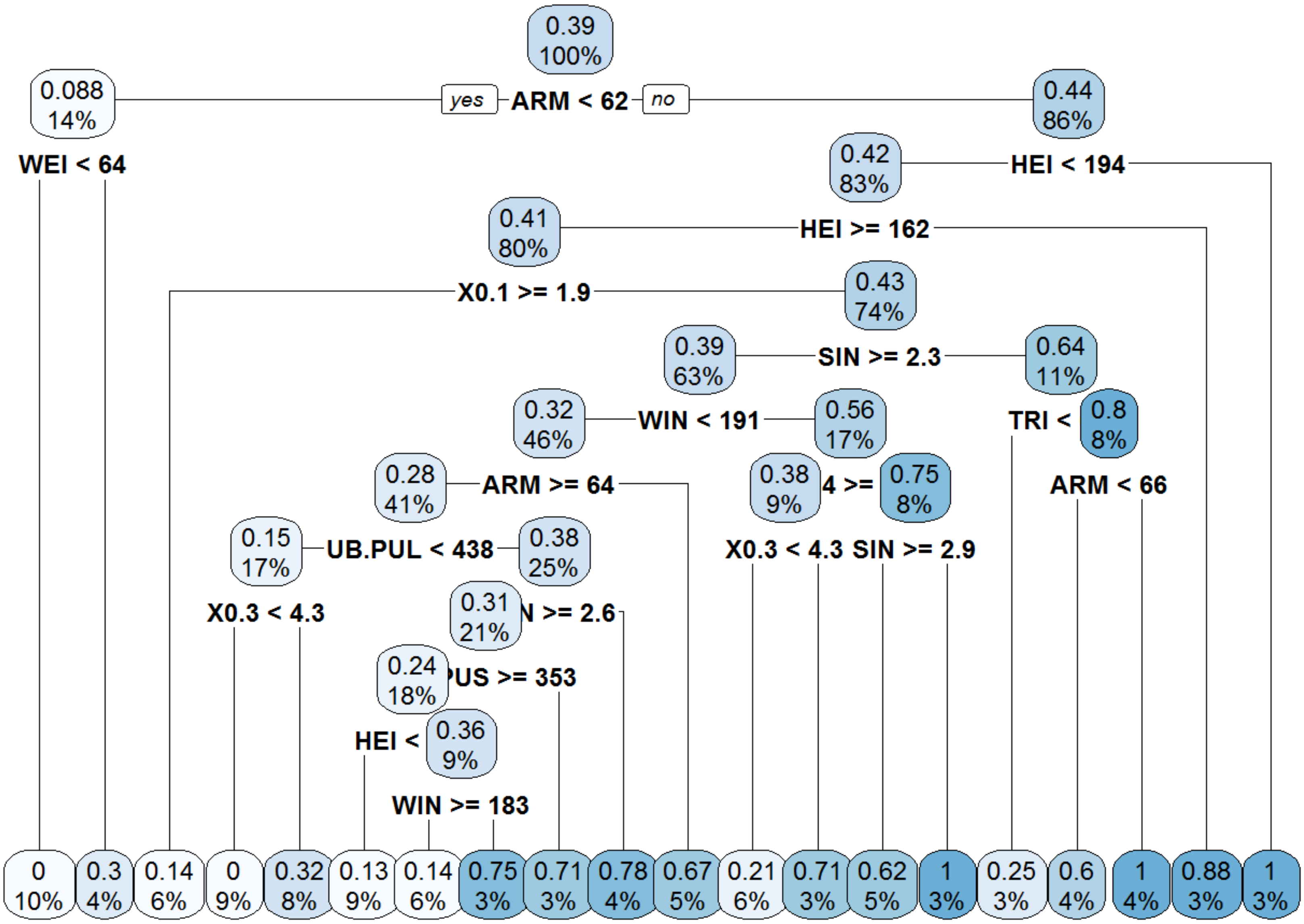

Figure 9. Decision tree classifer for male participants.

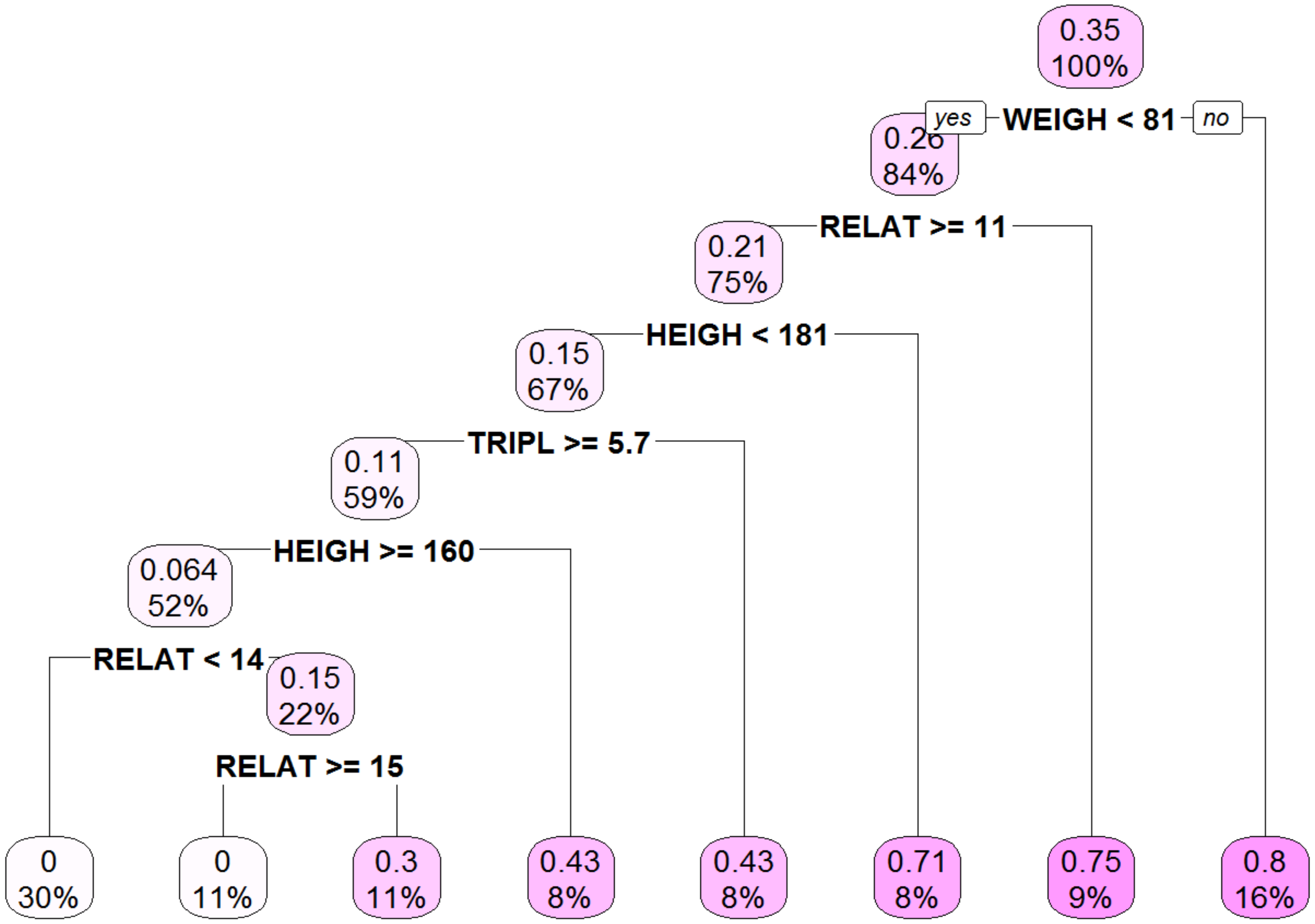

Figure 10. Decision tree classifer for female participants.

From **Figure 9** and **Figure 10**, the predictors the tree has chosen as significant predictors for males were arm/leg bike, height, upper body pull, 0-40m sprint, weight, upper body push, 0-10m sprint, relative six second peak power, 0-30m sprint, wingspan, triple jump, and single jump. Therefore, the tree did not find the lower body push or age measurements as important nodes to include in order to make decisions. For females, the classifier found only six measurements significant, in particular weight, relative six second peak power, height, triple jump, age, and upper body pull. Also noted from the trees are the predictors that are most significant to predicting the outcome of whether a participant will be nominated or not. The top three predictors for male participants are arm/leg bike, weight, and height whereas for females it is weight, relative six second peak power and triple jump. Most importantly, these trees provide insight as to what particular cut-offs are needed for certain tests as well as physiological measurements that a male or female athlete would need for an associated chance of being nominated.

To test how accurate these models are, the test data sets were used to see how well they were able to predict the participants outcomes; **Figure 11** and **Figure 12** show the ROC curves for the male and female participants trees, respectively.

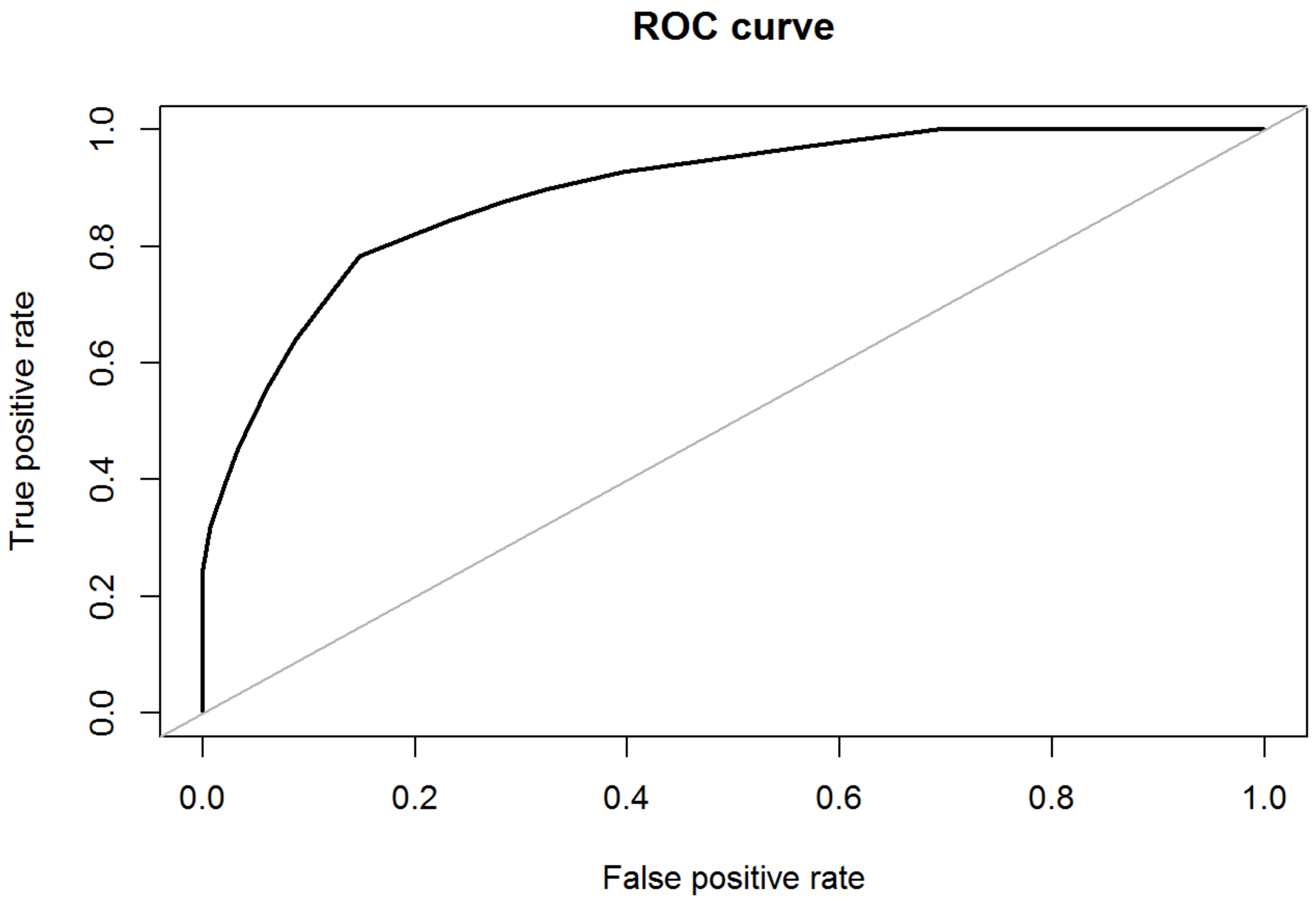

Figure 11. ROC curve for male participants.

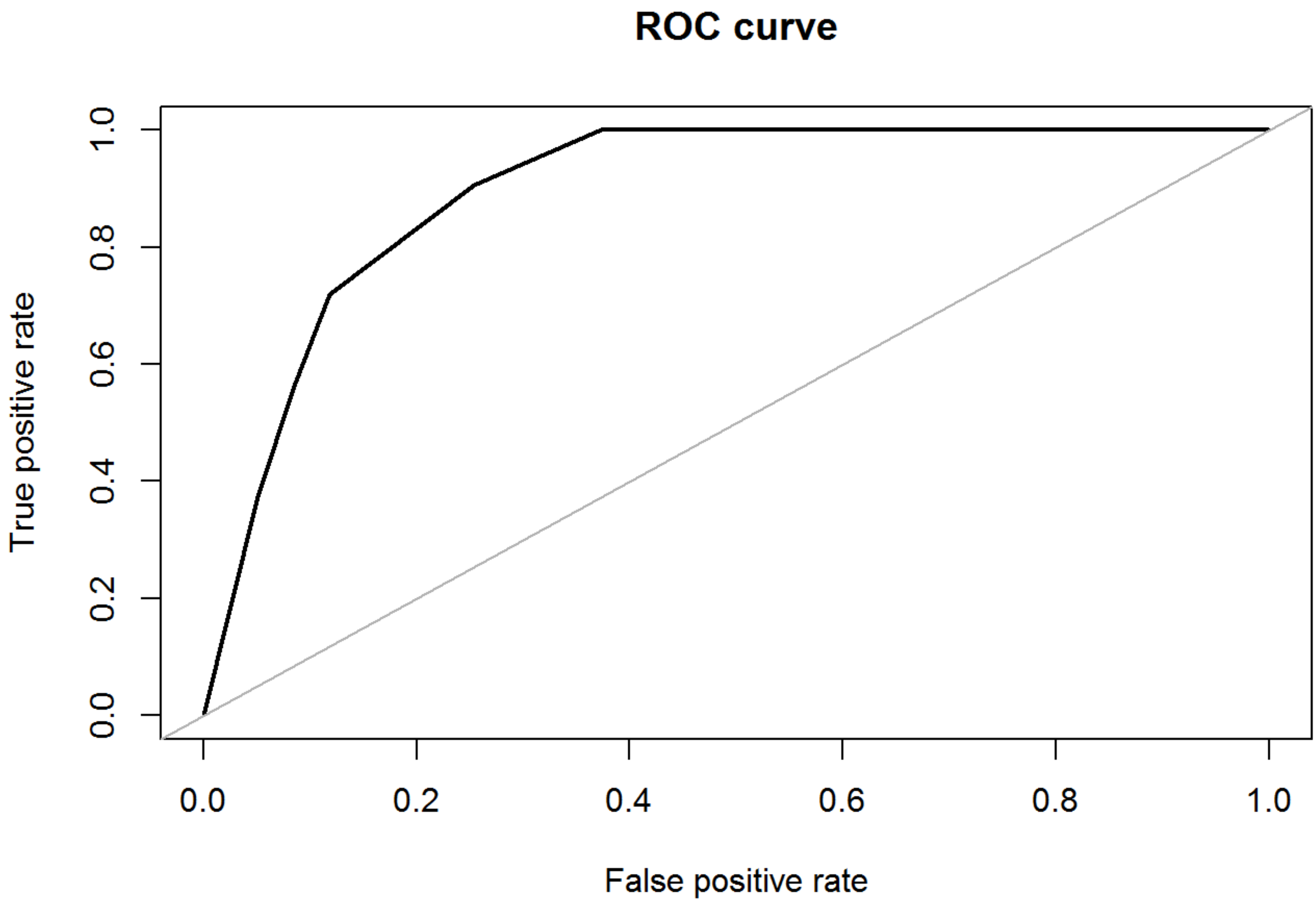

Figure 12. ROC curve for female participants.

After calculating the AUC of the ROC, the ROC curves suggests that the decision tree classifiers are much better than the regression models. In particular, the AUC is 92.4% for males and 90.7% for females. With this being said, this suggests that the decision trees seem to be very accurate at determining whether a participant will be nominated. To further analyze these decision trees, additional validation methods were applied such as bagging and random forests, however, these methods did not increase the level of accuracy since the data set was too small to effectively use these methods. Overall, the above decision trees seem to provide a good classification method for determining whether a male or female participant will be nominated.

Limitations

The main limitation apparent in this analysis is the data set size. The data set size is relatively small for applying certain prediction model methods such as bagging and random forests. The reason being is that there is not enough data to split into more test sets since each test set created would be too small. Since more data cannot be collected year to year, a possible solution to this would be to use more data from previous years as test sets. However, this could also pose a problem since benchmarks change each year because athletic performance generally increases. Therefore the model build based on the current years data may not be accurate for prediction purposes of the data from previous years. Further model selection considerations would be needed to accurately associate all the data.

Conclusion and Future Considerations

In conclusion, the decision trees seems to provide the highest accuracy in terms of prediction for both male and female participants. It would be interesting to see how these trees would perform with previous years RBC training grounds data. The full model logistic regressions perform better in terms of prediction accuracy compared to the reduced models, however, the reduced male model seems to fit the data better (lower AIC). In terms of ordering the test measurements from most to least influence on increasing the chances of being nominated, the following top three test measurements for both genders were gathered by the logistic regression analysis.

For males, the top three test measurements in order are: arm/leg bike, weight, and wingspan. For females, the top three test measurements in order are: arm/leg bike, height, and wingspan. These test measurements seem to be consistent with the most important predictors made by the decision trees. From a physiological stand point this is consistent to the anatomic make-up of men and women. These tests show that individuals who have measurements that are within the averages of **Table 1,2,3**, and **4** for each of the top three tests are more predisposed to have dimensions optimal for being successful as an Olympic athlete. The actual numerical increase in chances of being nominated with a certain test measurement is possibly not the most accurate with the logistic regressions and hence, further analysis needs to be done. Additionally, models with different response variables would be beneficial to consider particularly, whether accurate prediction models can be made to predict nominations for certain sports.