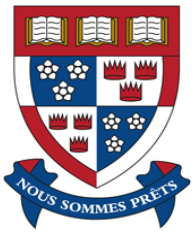

Simon Fraser University
Faculty of Statistics & Actuarial Science
Collaborated with: Canadian Sports Institute Pacific
Burnaby, British Columbia

A Method for Identifying Trending Amateur Players in the Context of the Canadian Women's National Soccer Team

Written By:

Barinder Thind
Matthew Reyers
Brad Smallwood

Supervised By:

Dr. Davis
Dr. Tsai

Contents

— Acknowledgements —

Thank you to Dr. Davis and Dr. Tsai for
allowing us the opportunity to be a part
of this class and to pursue this project.

Abstract

This paper explores methods to identify amateur female soccer players that are trending towards the Canadian Women's National Team. Our work utilizes athletic testing data from soccer Combine events. For athletes that did not complete all events, missing events are rectified through missing forest imputation. A scoring system is then developed that uses the best performances of each athlete across all of their tests. An athlete's score is then determined to be a weighted combination of their best performances, in which weights are derived through the use of gradient descent (via *optim()*). The results are then compared with predictions made by a Logistic Ridge Regression and a Logistic Principal Component Regression. Suggestions with respect to current top prospects and how to improve identification of future top prospects are then discussed.

Keywords: Soccer, Vector Optimization, Scoring System, Missing Forest Imputations

1 Introduction

The use of advanced statistics in sports for the purpose of player evaluation has seen a meteoric rise since the turn of the century. It has become evident that analytics are a vital tool that can play a prominent role in sculpting championship teams. Soccer is no exception and, while fitness tests have always been a part of the sport, the usage of them in the context of analytics can and has resulted in useful insights for coaches and managers [1]; the analysis done here falls under this domain.

2 Methodology

2.1 Data Description

The provided data had information about the fitness test scores for various amateurs and pro players from the Canadian National Women's Soccer Team. The particular tests were:

- 10 Meter Dash
- 30 Meter Dash
- 40 Meter Dash
- Broad Jump
- CMJ
- ASR
- Max Speed
- 30 by 15 Dash

Along with these test scores, information pertaining to a player's age, mass, height, position, and team were also provided. A correlation heatmap of the variables can be found in the appendix in Figure 6.

2.2 Imputation

The dataset prior to our imputation was sparsely populated. Imputation was essential to gain any insight from the data, regardless of the analysis methodology. Multiple imputation methods[1] were tested and a Missing Forest[2] consistently outperformed various cluster imputation and multiple imputation methods. Multiple Imputation Chained Equations (MICE) and K-Nearest Neighbor (KNN) imputation were the first two successful imputations we implemented. The methods were evaluated by introducing additional NA values to elements of a dataset at random that consisted of the player

[1] That was a pun

[2] https://arxiv.org/pdf/1701.05305.pdf

observation which had values in each of the fitness tests, and then imputing over this dataset. The results of the three imputations implemented can be seen in the table below.

Imputations			
Fitness Tests	MICE	KNN	Missing Forest
30 by 15	-2, 3	-2, 2	-10.8, 13.9
10m	-0.07, 0.23	-0.11, 0.08	- 0.082, 0.097
30m	-0.06, 0.22	-0.18, 0.12	-0.78, 1.84
40m	- 0.08, 0.24	-0.45, 0.34	-0.07, 0.06
Max Speed	-1.9, 1.3	-1.4, 1.9	-0.11, 0.06
ASR	-2.34, 4.17	-1.43, 2.36	- 0.216, 0.08
CMJ	-14, 34	-7.6, 5.1	-0.55, 1.80
Broad Jump	-11, 7.3	-21, 14	-1.98, 1.90

Table 1: Imputation Results of Our Three Methods

An additional benefit to using the Missing Random Forest is the ability to tune the parameters of the model in a similar way to that of a standard Random Forest. The out-of-bag imputation error is minimized by cross validation. Similar to a regular Random Forest, the model does not need to be cross validated on a train and test set due to the model being built on an "in-bag" sample and tested on an "out-of-bag" sample. In each iteration of the imputation process, the difference between the previous imputed data matrix and the new imputed data matrix is assessed. This is done for both continuous (age, fitness tests) and categorical (team) explanatory variables. The imputation process stops when both the differences for the continuous and categorical parts have increased.

The comma separated values in each cell of the table above represent the minimum and maximum imputation error for that fitness test. The closer these two values are to zero, the more successful that imputation scheme was. As seen in **Table 1**, the MICE and KNN imputations performed at near identical levels across most of the tests. The Missing Random Forest, however, outperformed both of the other imputation methods on all events except for the 30 by 15 test. Our final imputed dataset is therefore composed of the MICE imputed 30 by 15 scores, with all other scores imputed by Missing Forest.

2.3 Fitness Test Weighting

There are numerous methods for weighting [3][4] variables, however, these methods do not necessarily address the issue we were faced with here: the weight that was associated with the minimization was a vector of parameters rather than a single value. Let $\Psi = [\psi_1, \psi_2, ..., \psi_n]$ represent the aforementioned vector. Then, the minimization is defined as:

$$\textit{min}\ \{\sum_{j=1}^{k} \frac{\left(\frac{\sum_{i=1}^{n} u_i}{n} = x_j\right)}{k}\}\ \textbf{subject to}\ \mathbf{M1}\{\Psi^*\beta\}$$

Where u_i is defined as: $x_i - \bar{x}$, x_j is the j^{th} mean squared error[3], β is the matrix of values which corresponds to the eight fitness test variables, and the Ψ^* is a specific set of test values[4] chosen from the domain of Ψ. The algorithm is then carried out as defined in **Algorithm 1**. Of note here is that the algorithm was carried out through a model (defined **M1** in the algorithm) but the given model is not explicitly defined. There are significant ramifications dependent on the choice of **M1** however, they were not explored here[5]. Namely, the goal is to find the set of weights, Ψ, such that MSE is minimized. The question now shifts to which MSE should be minimized. MSE is a general measure for

[3] $\sum_{i=1}^{n} \frac{x_i - \bar{x}}{n}$
[4] Found and tested via gradient descent
[5] Although, there is some conjecturing on where research on this topic could go below

the performance of any model and so a better model might have an MSE lower than another model. It is also possible that this other model's minimum MSE *could* be associated with a different Ψ. In this case, the MSE is minimized subject to the *random forest* model and while the minimized MSE is lower than that of an unweighted MSE[6], there is no empirical evidence presented here to show that this is indeed the global minimum MSE across the set of all possible models $M1, M2, ..., Mn$.

Algorithm 1 Vector Minimization

procedure DONE THROUGH GRADIENT DESCENT VIA *optim()* IN R
  $\Psi^* \leftarrow \frac{1}{n}^7$
  $LB \leftarrow 0$
  $UB \leftarrow 1$
  $MSE_c = MSE_\beta$
While:
  if $MSE_c > MSE_f$, then $MSE_c = MSE_f$ *and* $\Psi^* = \Psi^f$ **given** $\exists\ MSE_f$
  if $\exists\ i \in \Psi^*[i]$ s.t $(i >$ UB$) \cup (i < LB)$ **then return** Ψ^*
  $\Psi^f \leftarrow optim()$
  goto *M1*
M1:
  Start Fold i
  Set : *Test*/*Train*
  Build Model on Train.
  Acquire MSE[i] via *Test*.
  goto *Next Fold (i + 1).*
  close;
  Return $MSE_f = avg(MSE[i])\ \forall\ i$.
  goto *5:*

A future consideration here is to explore this set of models and perhaps even consider the set of combinations of models (ensembles) in search of the global minimum. For these considerations, a metric could be defined to see whether or not the difference in minimal MSE is significant or not. Let the set of all minimal MSE's, ζ, be of size n. The hypothesis is then to consider the $\min\{\zeta_i\} = \alpha$ distance from the empirically discovered value of ζ_m[8] and to see whether this distance, δ, is different due only to random forces (for example, the randomization done during the cross-validation) or because there is a genuine difference in the ability of the models to make accurate predictions. Let $T_{(1)}$ be the first order statistic from ζ[9], the hypothetical global minimum. Then ζ_m can be compared as follows:

(I) $\alpha = T_{(1)} < T_{(2)} < ... < T_{(m)} = \zeta_m < ... < T_{(n)}$

(II) Let $\delta_i = T_{(i)}$ - $T_{(1)}$

(III) $\therefore \delta = \delta_1 + (\delta_2 - \delta_1) + ... + (\delta_m - \delta_1 - \delta_2 - ... - \delta_{m-1})$

(IV) Let the distribution of the sum of the first m order statistic distances (δ_i's) be defined as: $F_\Delta(\delta)$

(V) Then, applying the Central Limit Theorem [2]:

$$Z = \frac{\zeta_m - \alpha}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} \delta_i - n\alpha}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

We conjecture that you can normally approximate the test statistic to be [4]:

[6] Which is to say, the minimal MSE given $\Psi\beta$ is lower than the minimal MSE given just β
[8] Which is to say that ζ_m is the value corresponding to, for example, the lowest MSE in the *random forest* model here
[9] This is also α

$$T_\delta = \frac{\delta - T_{(1)}}{\frac{\sqrt{\frac{\sum_{i=2}^{n}(T_{(i)}-\alpha)^2}{n-1}}}{\sqrt{n}}}$$

(VI) Letting the confidence level be as desired, (preferably as some function, F, of some set of constraints specific to your context[10]), you can get an MSE value which you are satisfied with.

These considerations are to be used for future work. One important point to note here is that of computational constraints. The amount of flops required to compute even a single run of the algorithm is equal to the product of the number of folds, k, in the model and the number of combinations that *optim()* pulls for Ψ from its domain. The domain can be restricted[11] to deal with such an issue but then another problem arises: the probability that you are further away from α is the same or greater than it was prior to the domain restriction. Also, a note that the above equation requires you to know the minimum; this is obviously a theoretical task but for any pragmatic situation, you can set the first order statistics to be any desired minimum value and see whether what you have now is of significant distance away. Lastly, the final results of this weighting method can be found in **Section 3.2**.

2.4 Scores 1: Extreme Value Test Summations

One method to see how well players have performed is to compare their results for each exercise to the global maximum/minimum for that given exercise. For example, if you were to pick some random observation and noticed that each of their fitness test values were *far* away from the best scores, then you would say that this player is *far* away from the best player in the given set.

Using this logic, we can begin to develop a scoring system that can identify the amateur players that have better overall scores than the women's national team players.

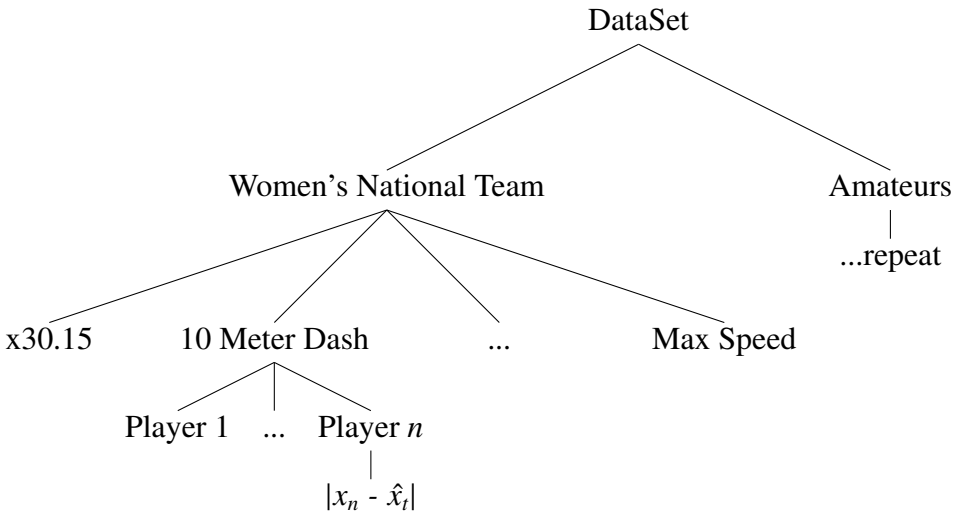

Table 2: Tree Diagram of Scoring Process

The diagram above depicts the overall algorithm of the scoring system. The data is first split on whether the players made the National Women's team or not, and then split again into individual events. For athletes who have been tested more than one time, their best performance in each event is considered. Next, the global maximums or minimums are found for each exercise. The choice of maximum or minimum depends on the kind of exercise it is. For example, if we are considering the *broad jump*, then we would take the global maximum for the set because a larger broad jump is better. For the 40 yard dash, however, we would want the global minimum. After we have these best scores, denoted by $\hat{x}_t$ where t is one of the eight exercises, we calculate the individual best to global best

[10] For example, a common one would be a computational constraint

[11] This is demonstrated in **Algorithm 1.** via the LB and UB values set in the preamble

distance for each player i belonging to the set of n unique players.

This process is done for each of those players in each of the subsets for *each* of the exercises. The reason we chose to use the L1 Norm is due to scaling issues. Since the differences are so small[12], using the L2 norm may create differences that do not accurately represent the magnitude of difference between the players. At this point, we have a baseline for the scoring process; summing up the values obtained for each player, we have a way to assess the performance of each player. However, it is worth considering the importance of each variable. Not all exercises are equivalent measures of potential soccer performance and are therefore not treated as such. Referring back to **Section 2.3**, we can use Ψ^* as the weights applied to the sum. This allows the final summation for a player i to be expressed as follows:

$$Score_i = \sum_{t=1}^{8} \psi_t |x_{it} - \hat{x}_t|$$

A sum of zero in the above formula would recognize an athlete that was the top performer in all events. The closer a score is to zero, the more closely aligned a player is with top performers in the weighted events. Implicitly this creates a cut-off: the highest value present in the Women's National Team subset can be seen as a cut-off for amateurs hoping to make the team. The results of using this threshold are less than ideal as evident in **Section 3.2**.

2.5 Scores 2: Logistic Ridge Regression and Leave-One-Out Prediction

While the fitness scores we've developed are useful in comparing athletes, they still do not allow us to gauge whether or not a player is trending to the Women's National Team (WNT). Furthermore, the fitness scores do not utilize some of the additional information we know about each player such as age, weight, or height. Intuitively, these predictors make sense to include. In order to determine which fitness tests and player characteristics are important when determining if a player will be selected for the WNT, a Logistic Ridge Regression was implemented. This model was chosen for its ability to address the collinearity between predictors and because it allowed us to estimate the probability that a player will make the WNT. The Logistic Ridge Regression is defined as follows:

$$logP(y|\beta, x) = \sum_{i=1}^{n} y_i log\left(\frac{1}{1+exp(-x_i\beta)}\right) + (1-y_i) log\left(\frac{exp(-x_i\beta)}{1+exp(-x_i\beta)}\right)$$

where x is a matrix with columns for each of the fitness test variables as well as age, height, and mass [3]. The maximum likelihood estimator for β is then

$$\hat{\beta}_{MLE} = \underset{\beta}{argmin}\left(- \sum_{i=1}^{n} log f(x_i|\beta)\right) + \frac{\lambda}{2} \sum_{i=1}^{k} \beta_i^k$$

where $\hat{\beta}_{MLE}$ is usually obtained via minimizing the negative log-likelihood. We chose to modify this approach to apply the additional constraint that for some $c > 0$, $\frac{\lambda}{2} \sum_{i=1}^{k} \beta_i^k < c$. The shrinkage model was chosen so that we could isolate only the fitness test predictors, as well as athlete characteristics, which were statistically significant predictors of our response variable. λ is the penalty parameter which is chosen by two fold cross validation. We chose such a low number of folds here because we perform this cross validation in the leave-one-out prediction (LOOP) process[13].

LOOP is a method of cross validation that can be used to predict for an entire dataset. The LOOP process is as follows: (1) Fit the model with $n-1$ observations in the dataset and then use this newly fit model to predict for the observation left out of the sample. (2) Continue this process until each

[12] The data was normalized
[13] https://arxiv.org/abs/1602.05801

observation has been predicted once. An important note to be made about this prediction method is that it has the potential to suffer from high variance. That is to say, the predictions may vary highly if a new sample was generated or collected. For our purpose, this was an acceptable short-coming as the goal was to develop a model that can be used in conjunction with the other scores.[14]

2.6 Scores 3: Logistic Principal Components Regression

The third set of scores are based on a principal components regression. The goal was to identify patterns or contrasts seen within each principal component that we decided to include in the prediction. The hope with dimension reduction here is more to see in what ways the different fitness tests interact and how we can use this information to make accurate inferences.

We will begin to define the model [5] by defining the logit transformation, and then the inverse logit transformation, as follows:

$$\theta_{ij} = log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$$

$$\pi(\theta) = (1 - exp(-\theta))^{-1}$$

where π_{ij} is the probability of success that a player will successfully make the WNT. From this we define the usual Logistic Regression model,

$$P(Y_{ij} = y_{ij}) = \pi(\theta_{ij})^{y_{ij}}(1 - \pi(\theta_{ij}))^{1-y_{ij}}$$

where $\theta_i = (\theta_{i1}, ..., \theta_{im})^T$ for $m < p$ and p is the number of predictors in the full model. The thetas themselves are linear combinations of the original p predictors. That is,

$$\theta_i = \mu + a_{i1}\boldsymbol{b}_1 + ... + a_{ik}\boldsymbol{b}_k$$

where $\boldsymbol{b}_1, ..., \boldsymbol{b}_k$ are the principle component loadings and the $a_{i1}, ..., a_{ik}$ are the principal component scores for the i^{th} player.

As with the Logistic Ridge Regression, we used the LOOP procedure where we fit a model on $n - 1$ observations to predict the n^{th} observation and repeat this process until each observation in the dataset has been predicted.

2.7 Stacked Clusters

Lastly, in order to combine the three scoring metrics and to identify the specific amateur players who are trending to make the WNT we implemented k-means clustering on the three scoring metrics. The three scoring metrics together compensate for the individual weaknesses of any one metric. For example, the fitness scores developed are well suited to measuring the athletic ability of a player but it does not utilize all of the information we have available to us on that player such as mass, height, and age. Similarly, if we had only used a single Logistic Regression model then it is unlikely that we'd have accurate probabilistic predictions on whether a player will make it to the WNT. To overcome this we implement stacking, a popular machine learning method to improve prediction accuracy. Specifically, we implemented this through clustering the three types of scores together using the K-Means algorithm. K-means is defined as,

$$\underset{C_1,...,C_k}{minimize}\left(\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij} - x_{i',j})^2\right)$$

[14] For future use, the model that is used could be changed and in practice, it would be a simple change (i.e. the scores gathered from this model would just be replaced by the improved model)

where each observation x_{ij} belongs to at least one of the K clusters $C_1 \cup C_2 \cup ... \cup C_k$ for player i over $j = 1, 2, 3$ which are the three player scores we defined above. This results in 3-dimensional clusters where the best players are the ones who score low on score 1 and high on both scores 2 and 3. For this stacked model to indicate that a player is elite and trending to the WNT, they must be a top performer in each of the 3 scoring systems outlined.

3 Results

3.1 Fitness Test Weighting

As mentioned earlier, the weighting was done with respect to the random forest model. The algorithm was run and the weights associated with the minimal MSE were extracted for use. The plot in **Figure 1** shows the results of the algorithm as it went through each of its iterations. It is interesting to note the cluster in the top left corner. We assumed that this cluster was a result of the way *optim()* looks to optimize over the grid of values. The model was also run on the original data set without weighting and the resulting MSE was **0.228**. The final MSE after the weighting, as shown in the plot, was around **0.18**.

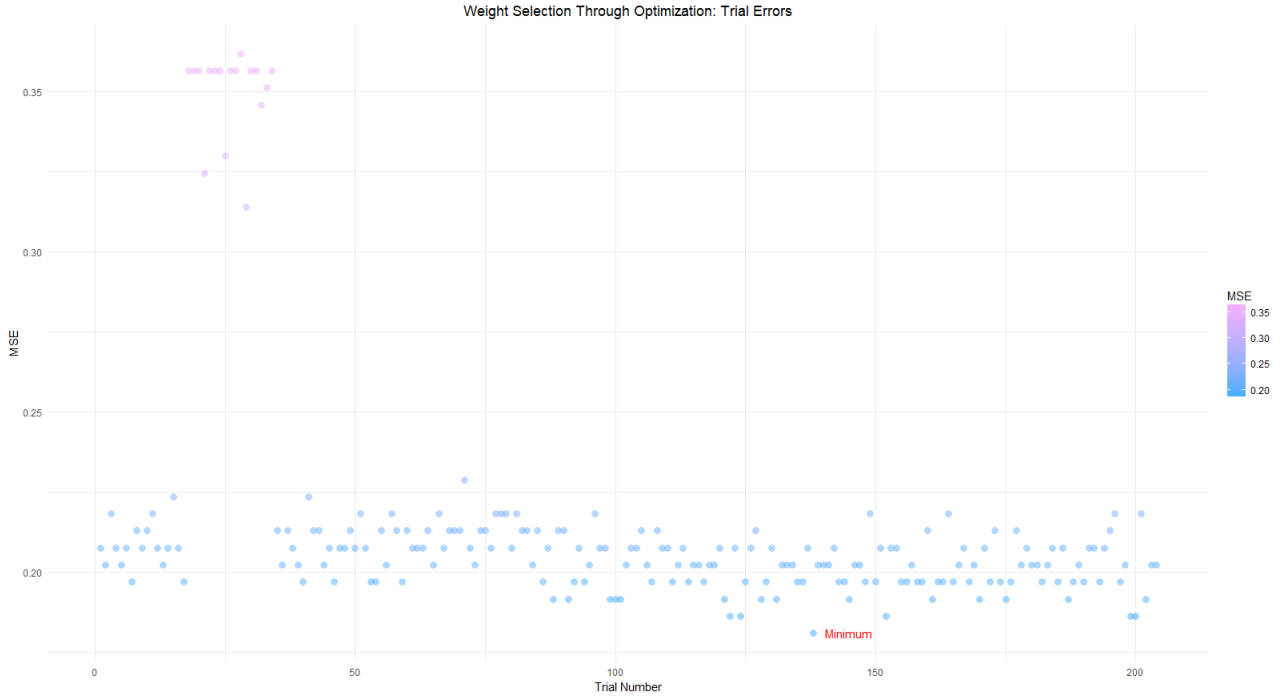

Figure 1: Optimization Error Plot

Lastly, the set of weights associated with this MSE[15] are: [0.1397415, 0.1228941, 0.1397415, 0.1250000, 0.1397415, 0.1397415, 0.1397415, 0.1250000].

3.2 Scores

The scoring metric was applied and, along with the weights above, the values were calculated for each player. As alluded to earlier, the implicit cut-off (the worst player in the National Women's team subset) was an issue with this data set as the player in question is only marginally better than the worst amateur. An explanation for this could be that this occurred due to the fact that we only looked

[15] I.e. the values for Ψ^*

at fitness scores. Some players who are not as athletic as others could still be very talented and thus make the team despite amateur-like fitness scores. In **Figure 2**, the distributions of the scores can be seen separated by team.

Next, we decided to consider the age plots of the various players and how they match up against the scores. This provided us a way to visualize how amateurs are performing relative to the pros.

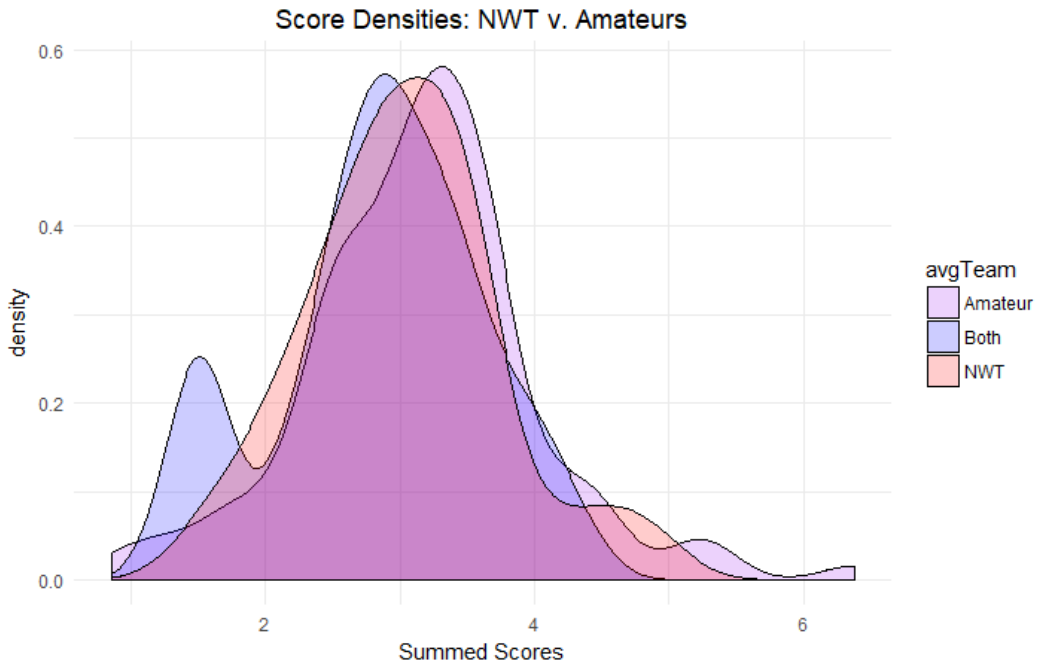

Figure 2: Distributions of Scores

The idea was that players who were of a young age but had values close to pro players were worth close consideration. The circled cluster in **Figure 3** illustrates the amateurs that performed particularly well for their age and are outside of the confidence intervals for the professional *local regression* line.

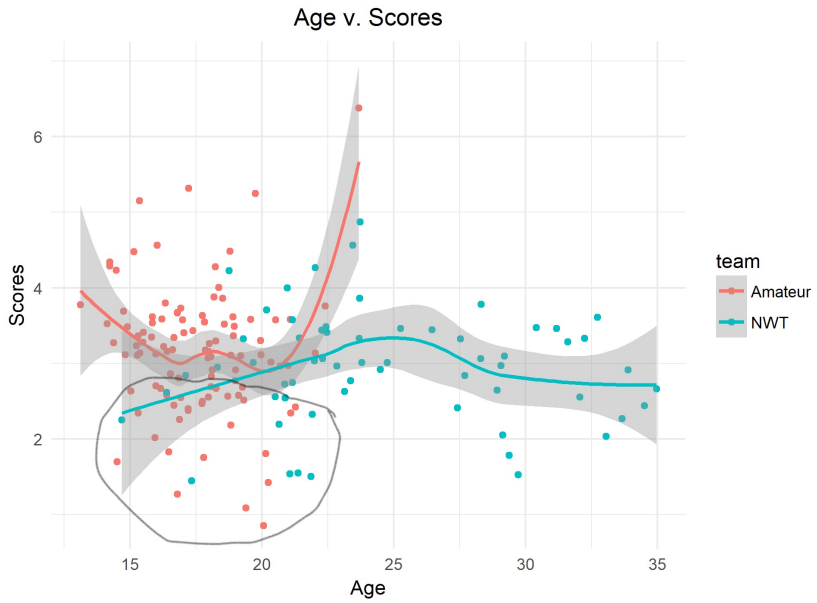

Figure 3: Age plot displaying score results

Next, we looked at the Logistic Ridge Regression and the scores achieved from its implementation. Since our prediction came from the LOOP process we do not have one single model we can evaluate and examine. Different penalty terms λ may have been chosen and different values for the β coefficients may have been estimated in each of the folds of the LOOP process. In order to evaluate the accuracy of these predictions we present two confusion matrices. The first will be a singular model where LOOP was not implemented, but instead cross validated on a regular train and test set. This is done to gauge the predictive power of a regular Logistic Ridge Regression on the dataset. We will then present the confusion matrix for the actual predictions we used where LOOP was implemented.

Table 3: Confusion Matrix of Leave-One-Out Predictions

Observed/Predicted	Women's National Team	Amateur	Class Error
Women's National Team	274	16	0.0551
Amateur	50	410	0.10869

As we see in the table above, the accuracy when using LOOP was quite high, at 91.2%. When not using LOOP, the results from a regular cross validation are:

Table 4: Confusion Matrix of Logistic Ridge Regression Predictions

Observed/Predicted	Women's National Team	Amateur	Class Error
Women's National Team	104	13	0.0563
Amateur	4	104	0.1111

The regular Logistic Ridge Regression produced a prediction accuracy of 89.6% which is very close to what was achieved with LOOP. Both of these results were obtained by taking 1 - P(Being an Amateur) as the accuracy of the model. Predicting whether a player was an amateur or not was approximately 8% in both cases.

Similarly, for the Logistic Principal Component Regression we took 1 - P(Being an Amateur) as the model accuracy of the model was only 28.1%.

Table 5: Confusion Matrix of Logistic Principal Component Regression

Observed/Predicted	Women's National Team	Amateur	Class Error
Women's National Team	218	105	0.325
Amateur	106	321	0.248

The final prediction accuracy of this method was 71.87%. While the principal component model is not as accurate as the logistic ridge model, we are able to use them to better understand the relationships in the data. The principal components are shown below in **Figure 4**.

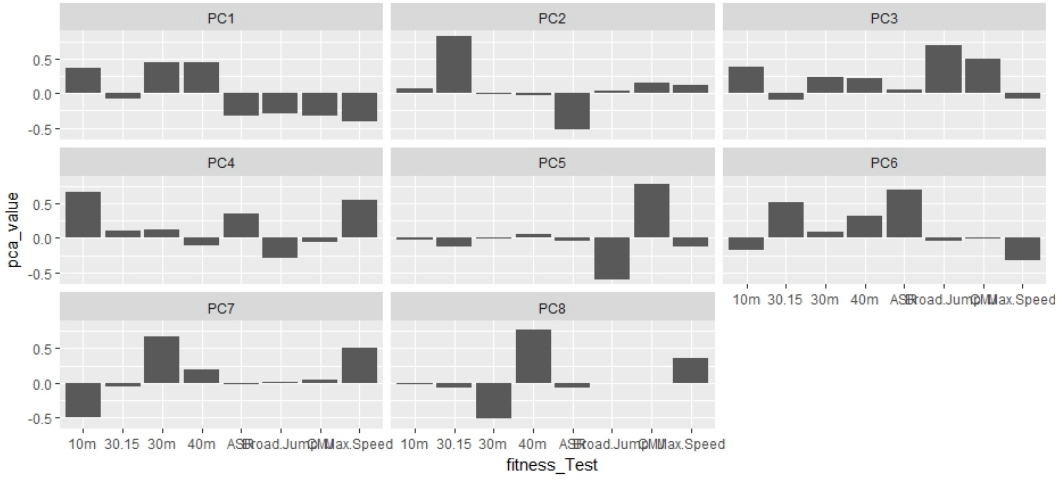

Figure 4: Principal Components of Fitness Tests

From the first principal component we can see that it is primarily a contrast between the fitness tests where we aim to minimize a time (speed based tests) and the tests where we aim to maximize (power based tests and max speed). Something the fitness scores we developed do very well is treat speed and power based tests equally. The same general range of values is assigned irregardless of which regime we are under, be it speed based or power based. Since this contrast explains the largest single proportion of the observed variation of the data, this may also validate the fitness scores we've designed.

3.3 Stacked Clusters

The stacked clusters allowed us to have more confidence in the amateurs that we selected as trending towards the national team. For example, if a player had a very low score but we predicted that they would not be on the national women's team, then this would cast doubt towards where they are trending. A visualization of the stacked clusters can be seen in **Figure 5**.

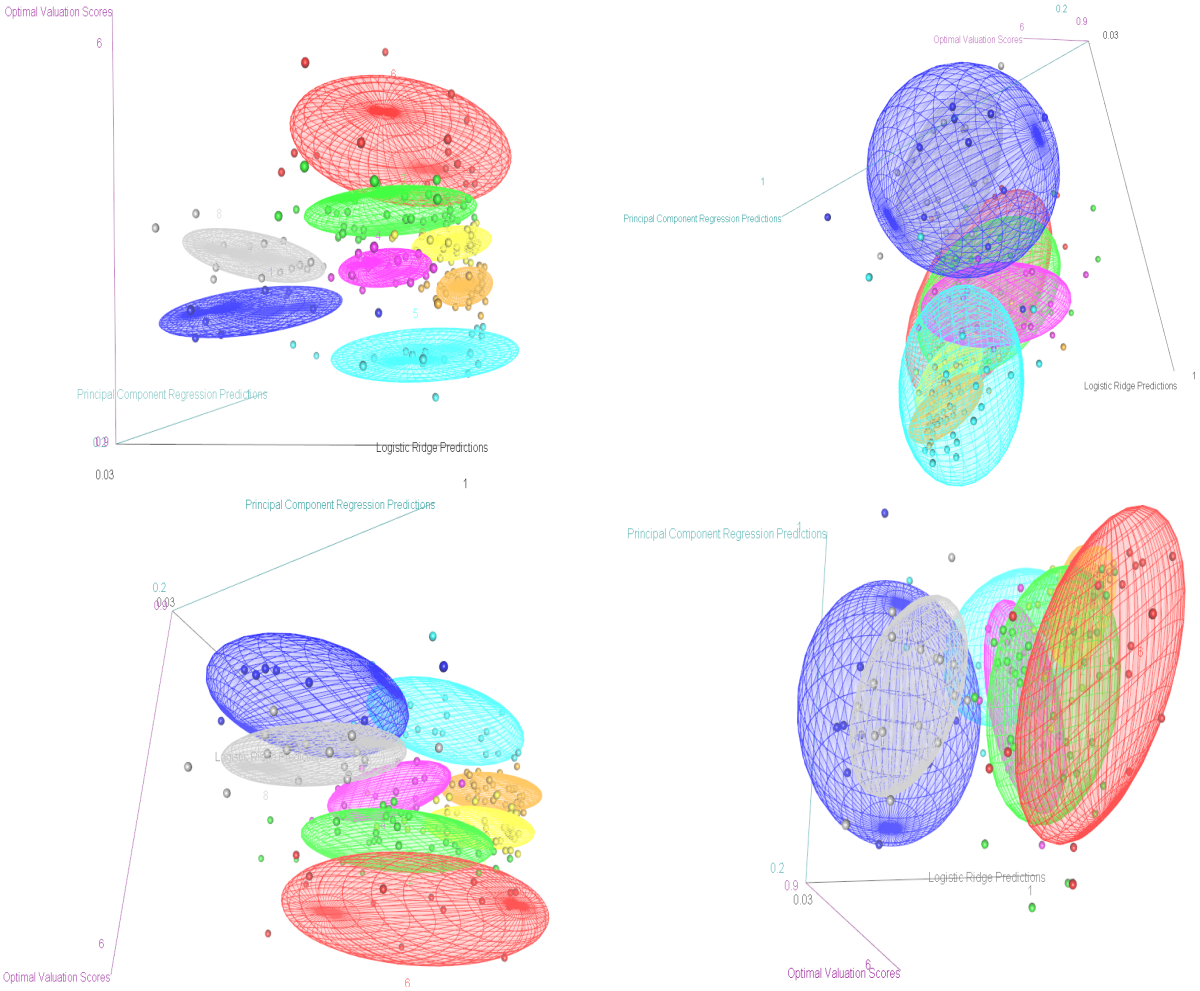

Figure 5: Three Dimensional Plots of Three Scoring Systems

In particular, consider the top left image; the indigo colored ellipse represents the group of players that performed the overall best across all clusters. These are the players with the lowest scores and with the highest probabilities when it comes to making the national team with respect to scoring metric 2 and 3. In contrast, the players in the red cluster are the ones performing the worst across all

metrics - they have the highest scores[16] and the worst probabilities of making the team according to the principal components regression and the ridge regression. We can now take a closer look at that indigo cluster to see which players they are exactly:

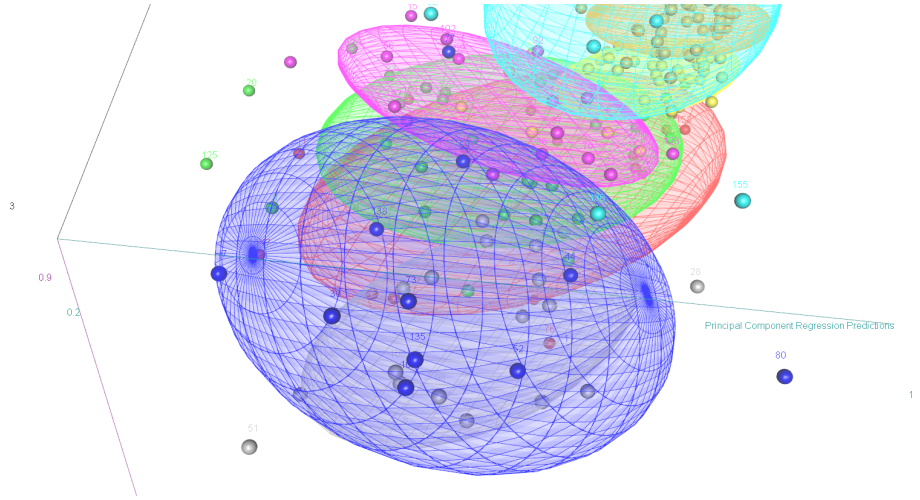

Figure 6: Players Corresponding to indigo Cluster

The player id's are: 7, 39, 44, 52, 59, 73, 77, 80, 108, 135, 138 of which only id = 138 is an amateur. Looking at the gray clusters, we find two more amateur players: 21 and 38. At this point, we would need to allow for more lee-way to identify a larger amount of players. We can consider the outliers in the light blue cluster which would have identified four more amateurs who were worth considering: 10, 16, 62, and 164. In fact, player 164 performed excellently according to our scoring system and should be considered a high priority. The last interesting note about the light blue cluster is that it largely corresponds to the circled group of players in **Figure 3**. However, note that many of those players were cut-off when using all three metrics. To be specific, the player id's are: 10, 14, 16, 19, 43, 61, 62, 90, 100, 120, 136, 155, 160, 164. Of these, 10, 16, 62, and 164 made the final cut. That's not to say that the others should not be considered; as mentioned earlier, we only have fitness data and that can only tell us so much.

4 Conclusion & Future Considerations

The project was fairly open-ended and what we created here is an all-purpose methodology that extensively examines fitness scores to provide information about which players are trending towards the National Women's Soccer Team. The methodology is fairly unforgiving to the average athlete but this is a conscious decision to compensate for the lack of information regarding talent[17]. The final results provide *Canadian Sports Institute Pacific* with information about which players are most worth considering and also details a way to repeat this scoring methodology in the future when more complete data is available.

There are also many avenues left unexplored. For the future, the optimization methods could be explored further (for which some groundwork has already been laid). Empirical tests could see which weights are the best for a given model. Also, there could be more work done on which classification model is picked; perhaps a better model would have provided more fruitful results for when stacking

[16] Which is bad

[17] That is to say, given this information, we can begin to include players that would be excluded here as their talent could compensate for an unflattering clustering or they might be clustered differently altogether

the scores together. Regardless, we hope that there is enough here to allow for future research that can improve on our conclusions!

5 References

[1] David Beaudoin and Tim Swartz. «Strategies for Pulling the Goalie in Hockey». In: 64 (Aug. 2010), pp. 197–204.

[2] George Casella and Roger L. Berger. *Statistical infence*. Duxbury, 2002.

[3] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2017.

[4] Stefan S. Ralescu Madan L Puri. «LIMIT THEOREMS FOR RANDOM CENTRAL ORDER STATISTICS». In: (1986), pp. 447–475.

[5] Jianhua Z. Huang Sy Seokho Lee and Jianhua Hu. «SPARSE LOGISTIC PRINCIPAL COMPONENTS ANALYSIS FOR BINARY DATA.» In: *Annals of Applied Statistics* 4 (2010), pp. 1579–1601.

6 Appendix

The following is the correlation heatmap:

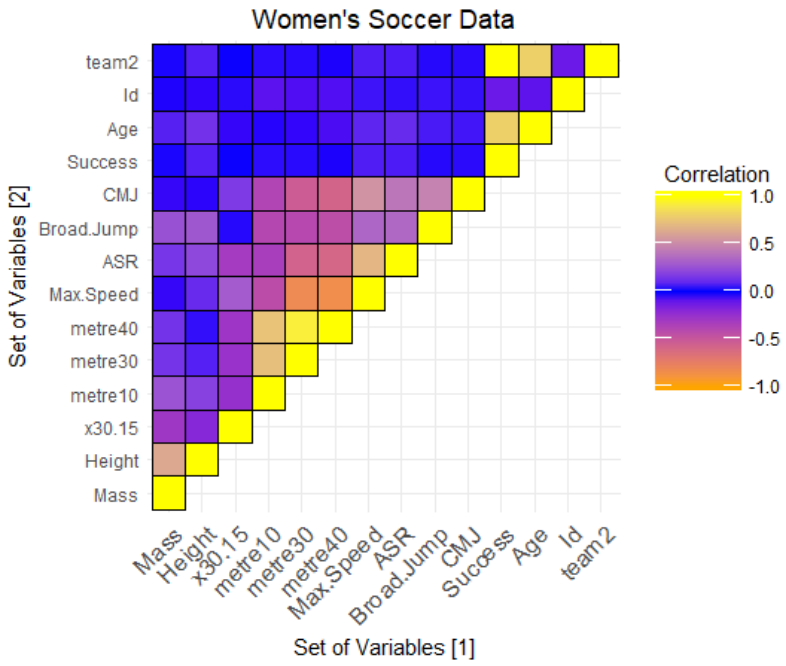

Figure 7: Correlation Heatmap