

Milestone II Final Project Report

Project Title: Rugby Match Prediction and Player Cluster Analysis

Adam Bakopolus, Carolann Decasiano

Introduction

In the world of athletics, sports betting only continues to grow, becoming more and more prevalent in pregame shows and even during the game broadcasts themselves. As a result, there is an increasing population interested in finding features and models that can help predict results, leading to better gambling and betting outcomes. The problem and question that the supervised learning portion of this project hoped to solve was whether there were features in the Men's International Rugby Union space that could allow for effective prediction of match results. Again, with these features and model(s) identified, it will allow for better understanding of how two teams may match up against one another and how the final score would be a reflection of this team comparison. Individuals would be able to more reliably place wagers for a particular outcome when they understand the impact home field advantage, average points for and against, etc. play in this sport. With the 2023 Men's Rugby Union World Cup starting in September of this year, there was a particular interest to focus on this area to establish a model that could reliably predict results ahead of these matches. Rugby Union is also a sport with a small, but growing, fanbase and, as a result, had little prior work from which to draw. There was, therefore, also a motivation to develop and tune a novel model to leverage within this space.

For the supervised learning portion of this project, three models were trained and tuned, a Random Forest Regression model, a Support Vector Regression model, and a neural network. The objective of the supervised work was to leverage the features detailed later in this report to predict the point differential between the two teams participating in the match (the score of Team A minus the score of Team B). From this point differential, a winner could also be determined. The tuning of these three models leveraged Grid Search to identify the best combination of model parameters to use. When the "best" combination was determined, 5-fold cross-validation was performed, and the mean average error was the metric leveraged to determine the "best" overall model. The mean and standard deviation across all five folds was taken into account, and it was determined that the neural network was the best model to leverage for this supervised task centered around Rugby Union score prediction. All models had negative mean average errors around ~ -15 and training data accuracy ranging from 74-77%, so there was fairly strong predictive power across all three models, but the neural network was leveraged for the failure and ablation analyses detailed further below. This neural network also was used to predict the 2023 World Cup, selecting the host country of France, the number two team in the world, a slight upset selection on the game's biggest stage.

In the unsupervised portion, the goal of this task was to cluster players by analyzing playing style for potential roles on a roster. For coaches, or a fantasy rugby enthusiast, this application on rugby players' could be used to supplement their team's cohesion by identifying patterns and finding successful player combinations on match day. Two models were fit and evaluated for this task: KMeans and Agglomerative Clustering. They were fed PCA and PCA/UMAP combination outputs and tuned using GridSearch as well. Then they were evaluated based on three metrics: Silhouette, Calinski-Harabasz Index and Davies-Bouldin Score. Both models in various iterations had positive Silhouette scores, but the KMeans model using the PCA/UMAP outputs performed better by finding a balance between the local and global

structure of the data providing more discernible cluster densities. This model gave an optimal cluster of 9 player styles.

Related Work

This is a link to an existing study similar to our outlined approach in the project proposal:

<https://www.sciencedirect.com/science/article/pii/S2210832717301485#b0040>

This article provides a framework for machine learning-based team and individual sport performance and score prediction. While this article did acknowledge some prior work around rugby score prediction was already available, this study did not leverage any rugby-specific features and instead tried to globally predict scores across multiple leagues (e.g. Rugby Union, English Premier League, Australian Football). With the goal of our project centered specifically around Rugby Union and score differential and match winner prediction, this specific area of focus appears novel.

Additionally, this link also points to an already established study that focused around sports betting and match prediction for tennis: <https://content.iospress.com/articles/journal-of-sports-analytics/jsa200463>

This paper used the same models (Random Forest, SVM, Neural Network) for their prediction of match results as this project, which was an encouraging sign for the direction of the work. Our work again builds off of prior match prediction work, as not only will we attempt to predict a match winner, but we will also predict the point differential. This type of prediction and information would have major benefits as the reach of sports betting only begins to grow within the world of professional athletics.

A related work around unsupervised methods in sports analytics was done by Verna et. al. They applied k-means clustering to data of ball possession in areas of a field to determine playing styles of various English Premier League teams. They plotted the Within-Cluster-Sum-of-Squares against 1 through 20 clusters and selected a k using the elbow method. This work provided a basis for a use case in this report's unsupervised task. The difference is the attempt to cluster the players based on attacking and defensive features. So that playing styles could be used to build a roster.

Data Sources

For the supervised learning task, there was an initial hope to leverage ESPN Scrum and a web scraper to pull down historical match data for Men's International Rugby Union. However, unfortunately ESPN discontinued this domain a year or so ago and available web scrapers were unable to parse through ESPN's newly-implemented format. Fortunately, prior to this migration, data was pulled down into github and was available for download here:




https://github.com/octonion/rugby/tree/master/world_rugby/csv

To generate the required dataset for the supervised learning tasks, the CSVs corresponding to 2004-2022 match results were pulled into one dataframe. Initially, this resulted in 5,038 records once the data was filtered to Men's Rugby Union only. Prior to any processing or data cleaning, the dataframe contained detail around venue name, city, and country, the date of the match, which teams were participating, and the score of the match among other less important details related to the game played. In regards to the initial preprocessing needed for this dataframe, there was inconsistency in how team names were listed. For example, Canada A and Canada and Italy and Italy XV should be considered the same. Additionally, consistency around "and" and "&" and team nicknames (Maori All Blacks, New Zealand Maori, New Zealand) needed to be introduced. Therefore, to account for these types of situations, functions that cleaned up inconsistent naming issues for both the team and venue fields were developed.

Prior to detailing the final dataframe size and feature engineering/selection needed for the supervised task, one additional data source must be discussed. An important expected feature needed for effective prediction would be the world ranking points tied to a Men's International Rugby Union team. This was available through this site:

<https://www.world.rugby/tournaments/rankings/mru>

As opposed to a simple 1, 2, 3 ranking, teams are ranked based on points which allows for a better understanding of how two teams may match up against one another:

MEN'S FULL RANKINGS					
1	→ (1)	 Ireland	91.33	★	▼
2	→ (2)	 France	89.38	★	▼
3	→ (3)	 New Zealand	88.98	★	▼

From 2004 to 2023, Men's International Rugby Union points for each team (if available for the year) were added into an Excel document. This Excel, once all years were added in, was converted to a CSV and was also then made available as a dataframe. Now, with this dataset available, world rank scores could be included as features in the supervised learning task. These two datasets would form the basis for the feature engineering process detailed below.

For the unsupervised task, RugbyPass.com was used to gather the dataset. RugbyPass is a digital rugby platform for broadcasting and print content for Rugby fans. The site contains videos & analysis, live scores, stats, fixtures and results on rugby union matches and players. Their coverage is primarily split by northern and southern hemisphere competitions. Using python's BeautifulSoup package, player game stats were scraped, parsed and stored into csv files by year. Ultimately, 3911 rows were obtained from rugbypass of players for 2018-2022 seasons¹. These include player stats from their international and league teams, the latter were scraped separately after little amount of stats were available for their international sides.

Due to the format of the site, a separate web scrape of RugbyPass was done to gather players' weight, and height. That was then merged with the player game stats. The dataset contains 21 features describing the players position, attack (ex. clean breaks, metres run, kicks) and defense (ex. Tackles, turnovers won) features. See appendix for complete details of features. There were a few preprocessing steps that had to be taken to get the data into a good shape for input into our model.

Initial review of the data contained 487 players with missing their weight and height. Missing values were collected from other international and league team websites. The remaining missing values were filled with the averages for their positions. Seven rows were dropped because they had 0 or null values for attacking and defensive features. Lastly, rows were grouped by player and values were averaged since players contained multiple rows for their performance on their international and league teams.

¹ Very little of the dataset contained stats for 2020 due to the suspension of play caused by the Covid-19 pandemic.

Feature Engineering

With the two datasets listed above, and following the initial data preprocessing, there were a few additional steps to move from the raw input data to the final features used in the supervised machine learning methods. First, the second dataset was merged with the first to incorporate the world ranking information for both teams in the match. The match year, in addition to team names, was used to appropriately complete this merge. Additionally, if either the first or second team was eligible for home field advantage (based on the venue country being equal to the team name) then a binary column would be denoted as a 1. If there was not a home field advantage, the team would be listed as a 0, and if the teams played on a neutral site, both would be considered a 0 for home field advantage. Following the merge of the world ranking dataset, our initial record count was reduced to 3,841 records, which was a result of semi-professional and minor league/backup teams being included in the original dataset.

Lastly, another dataframe was created by melting the dataframe created above to help aid in a moving average calculation. A 3-game moving average for points for and points against for a team playing in the match was created by evaluating the three prior games played by the team. This was to get a sense for how prolific a team's recent offensive performances have been or how stout their defense has performed. When creating the moving averages and merging back into the original dataframe, the final cleaned and processed dataset was 3,596 records. The slight decrease was due to removing records for a team's 1st through 3rd games as their moving averages would not be calculable. Following this engineering, the eight features to be used in the supervised learning method are:

- Team A's world ranking points
- Team B's world ranking points
- Home Field Advantage for Team A (binary)
- Home Field Advantage for Team B (binary)
- Three-Game Moving Average of Points For for Team A
- Three-Game Moving Average of Points Against for Team A
- Three-Game Moving Average of Points For for Team B
- Three-Game Moving Average of Points Against for Team B

A sample of how some of these features would look for predicting the point differential between the two teams in the match (score_a minus score_b):

venue_country	match_year	team_a	team_b	score_a	score_b	world_rank_a	world_rank_b	home_a	home_b	points_for_moving_a	points_against_moving_a
Uruguay	2017	Tonga	Samoa	31	28	71.96	71.29	0	0	22.333333	21.666667
Wales	2017	Wales	New Zealand	18	33	82.56	94.78	1	0	20.333333	14.333333
Scotland	2017	Scotland	Australia	53	24	80.71	86.36	1	0	30.000000	20.000000
Wales	2017	Wales	Australia	21	29	82.56	86.36	1	0	20.333333	18.666667
Argentina	2017	Argentina	Australia	20	37	79.93	86.36	1	0	55.333333	22.333333

To start feature engineering on our unsupervised learning dataset, we began with using a heatmap to quickly examine correlations and determine if any features had little to no relevance in the dataset. Three were dropped that had more to do with player discipline than it had with their performance; penalties conceded, yellow and red card counts. We also did not include player position since the goal is to cluster players into playing style as opposed to their traditional positions. Since the plan was to input the data to a KMeans model it was important to also scale the features so players would be better evaluated and

assigned to clusters by the distance metric. For example, a 40 metre run is not comparable to 5 conversion goals.

Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) were the various techniques used for dimension reduction. The first technique attempted was PCA that transformed the correlated variables into a smaller number of principal components. In this case, four principal components were tested in our clustering models to represent our features, explaining 79% of the variance. Five components (83% of variance) were also tested in our k-means model and had a slightly lesser mean silhouette score, 0.45. With four components, we opted for a slightly better performing model versus increased retention of variance of the original feature space. Much of our research in PCA also references a common practice of using components that represent 80% of variance.

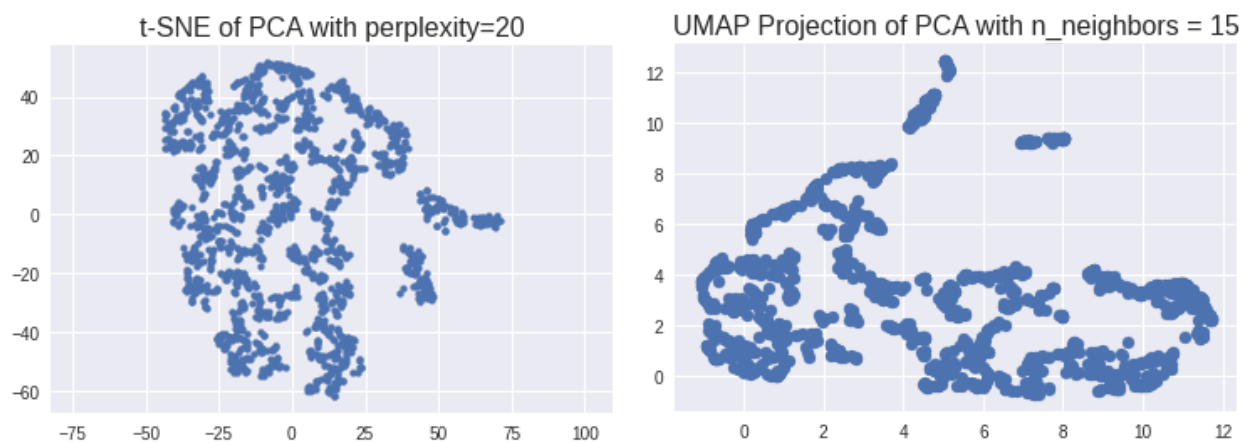


Figure 1. t-SNE and UMAP Plot Comparison

To visually explore the components, we used an additional dimensionality reduction technique, t-SNE, to project their relationships and to possibly identify clusters for our KMeans model. However, the resulting plot did not show a clear structure or discernable patterns. Even after conducting multiple iterations of t-SNE plots at varying perplexities, little insights were gleaned from them.

The next technique in our feature engineering was to use UMAP on the principal components, in the rest of this report will just be referred to as UMAP, to have more control over visualizing the local structure of the data by using the `n_neighbors` parameter. There was an attempt to use UMAP on the normalized data to get results with distinguishable clusters, but it did not perform any better than using KMeans on just the PCA.² From the plotted UMAP embeddings (Figure 1) there is better cluster separation when put next to the t-SNE plot.

Part A. Supervised Learning

Methods Description

The three supervised learning methods that were to be leveraged for this project were a Random Forest Regression model, a Support Vector Regression model, and a neural network. In terms of a supervised

² See links to notebook `rugbyplayer-unsupervised-cluster.ipynb` in appendix

learning workflow, the dataframe that was created and detailed in the Feature Engineering section above (with a sample snapshot included with some of the key fields displayed) was split into X and y tables. The X table included the following 8 features:

- Team A's world ranking points
- Team B's world ranking points
- Home Field Advantage for Team A (binary)
- Home Field Advantage for Team B (binary)
- Three-Game Moving Average of Points For for Team A
- Three-Game Moving Average of Points Against for Team A
- Three-Game Moving Average of Points For for Team B
- Three-Game Moving Average of Points Against for Team B

And the y table or Series was the point differential in the match, or the score of Team A minus the score of Team B.

The first model to tune was the Random Forest. Grid Search, with 5-fold cross-validation and negative mean absolute error as the scoring, was leveraged to identify the optimal max_depth, min_samples_split, and min_samples_leaf parameter values. These values were found to be 10, 2, and 20, respectively. A Random Forest Regression model, with 100 n_estimators and the values noted above, was used in cross-validation testing with the X and y data. Specifically, 5-fold cross-validation was run with the scoring again being the negative mean absolute error. To justify this approach, a random forest was a very suitable method to leverage for this type of project. The method is very intuitive and suitable for the type of regression task outlined for this project. Further, random forests are more powerful than a simple decision tree, leveraging many trees to arrive at a model less prone to overfitting. Further, 5-fold cross-validation was leveraged for this work as a more robust way of assessing model performance relative to a simple train/test split. We wanted to ensure that the model was not overfit and would respond well to new data, which is difficult to capture through just a simple train/test split. Lastly, negative mean absolute error is a standard method for evaluating the fit of a regression model. It simply assesses the predicted differential to the actual differential for each of the matches, with the larger score (-15 being larger than -16, for example) denoting a stronger model prediction.

Next, a Support Vector Regression model was tuned. Similarly, Grid Search was leveraged to assess the optimal C, kernel, and gamma parameters to use for the training data that was available. These values were found to be 1, 'linear', and 1, respectively. Therefore, in effect, our SVR model was behaving similarly to a simple linear regression model. Regardless, this "best" SVR model was used in cross-validation testing with the X and y data noted above. Again, 5-fold cross-validation was leveraged with the negative mean absolute error assessing the model performance. The decision to use cross-validation and the negative mean absolute error metric in SVR was for the same reason as noted above in the Random Forest Discussion. For why SVR was the second chosen model for the supervised task, it allowed for greater flexibility depending on the type of data that was available. The ability to tune a model to assess non-linear relationships within the data (rbf and sigmoid kernels) is powerful and more flexible than simply selecting a linear regression model. Ultimately, a linear kernel was selected, but SVR was still a suitable choice to cover situations where data may be linear or nonlinear.

Lastly, a neural network was tuned. Similarly to the other two models, Grid Search was leveraged to assess the optimal optimizer, learning rate, activation, batch size, and number of epochs. Additional Grid Search results found a single hidden layer of 14 neurons to be the most effective in predicting point differentials. This Grid Search returned values of RMSprop, 0.01, relu, 16, and 100, respectively. This "best" model was used in cross-validation testing with the X and y data noted above, with negative mean

absolute error serving as the scoring metric. The decision to use cross-validation and the negative mean absolute error metric was for the same reasons as detailed above. For why a neural network was chosen as the third and final model, it was well-documented in the literature that neural networks performed well in these types of point differential and score prediction tasks. As a result, tensorflow/keras was leveraged to build up a neural network model.

As will be seen in the below section, these methods provided 3 supervised learning models that performed relatively well across the board. However, the neural network would be deemed the best and was carried through to the failure and ablation analysis portions of the project.

Supervised Evaluation

Negative Mean Absolute Error was the chosen evaluation metric for the assessment of the three models detailed above. Mean Absolute Error is a popular metric for regression models and was a fitting choice, therefore, for this particular project. MAE assesses the average difference in the predicted score to the actual score. Unlike mean squared error, MAE does not square the difference between predicted and actual, which tends to more greatly punish predictions that are far off from actual. By using MAE, the goal was for the model not to be unfairly punished for not predicting blowouts or upsets, which are often difficult to accurately capture through a model. The negative portion of negative mean average error just notes that the MAE was set to a negative value, and therefore larger values would be more desirable when assessing model performance. When assessing a model, the negative mean average error was determined for each fold of the 5-fold cross-validation, and then the mean and standard deviation were used to assess the “best” of the three models to proceed with.

Additionally, while it’s not considered an adequate metric, for this project we also note the accuracy of the three models on the full training data to predict the correct winner of the match. While the true evaluation metric, negative MAE, assesses the ability of the models to predict the point differential, this is a similar view into how the models may perform in picking a simple winner, regardless of a score spread. The Random Forest, SVR, and Neural Network models had prediction accuracy of 77.1%, 74.1%, and 74.3%, respectively, suggesting that the models do perform fairly similarly with a relatively high degree of accuracy.

Table 1. Mean and Standard Deviation of Negative Mean Average Error for Optimized Random Forest, SVR, and Neural Network Supervised Learning Models For Predicting Match Point Differentials in Men’s International Rugby Union

Random Forest Regression	Support Vector Regression	Neural Network
-15.324 (0.609)	-15.077 (0.461)	-15.051 (0.493)

Again, while these models, as seen above by the mean negative MAE (and standard deviation in parentheses), performed fairly similarly, the neural network performed the best based on the chosen evaluation metric, and it was, therefore, leveraged in the below ablation and failure analysis work.

Ablation Analysis

To perform an ablation analysis, the “best” neural network model was again leveraged with one slight tweak - as opposed to having an expected input dimensionality of eight, it instead expected seven input features. A for loop was created that, for each iteration, removed one of the eight features noted above from the X table, built the model, and ran a 5-fold cross-validation on this new X and y data, using again the negative MAE as the evaluation metric. From the five-folds, the mean and standard deviation of

negative MAE was captured for when each one of the features was removed. The below table details the results.

Table 2. Mean and Standard Deviation of Negative Mean Average Error for Optimized Neural Network Model Following Ablation Analysis

Removed Feature	Negative MAE (<i>standard deviation</i>)
world_rank_a	-18.343 (0.740)
world_rank_b	-18.383 (0.550)
home_a	-15.089 (0.553)
home_b	-15.122 (0.497)
points_for_moving_a	-15.126 (0.497)
points_against_moving_a	-15.061 (0.364)
points_for_moving_b	-15.105 (0.495)
points_against_moving_b	-15.131 (0.482)

As can be seen above, the world rankings of teams A and B play a major role in the success of the model's score differential prediction. Intuitively, this is reasonable as there are very few notable upsets among national teams when they play against each other, so these two features have great predictive power. While the model's negative MAE decreases to ~ -18.4 when either of these features are removed, the removal of other features also has an effect, albeit significantly less impactful. This ablation analysis provided great insight into what is driving the model and its predictive success.

Sensitivity Analysis

To assess how the model may be sensitive to the choice of hyper-parameters, three sensitivity analyses were run: one assessing the impact of different optimizers (RMSprop vs. Adam and SGD), different learning rates (0.01 vs. 0.001, 0.005 and 0.1), and different activations (relu vs. tanh and sigmoid). A for loop was created that, for each iteration, tweaked the hyper-parameter of interest, built the model, and ran a 5-fold cross-validation on the X and y data, using again the negative MAE as the evaluation metric. From the five-folds, the mean and standard deviation of negative MAE was captured. With all else remaining constant outside of the one hyper-parameter included in the sensitivity analysis, the results are detailed below.

Table 3. Mean and Standard Deviation of Negative Mean Average Error for Optimized Neural Network Model Following Sensitivity Analysis

Optimizer Sensitivity	RMSprop -15.051 (0.493)	Adam -15.320 (0.594)	SGD -15.548 (0.720)	
Learning Rate Sensitivity	0.01 -15.051 (0.493)	0.001 -15.185 (0.529)	0.005 -15.127 (0.472)	0.1 -17.246 (1.193)
Activation Sensitivity	relu -15.051 (0.493)	tanh -16.259 (0.934)	sigmoid -15.838 (0.621)	

This affirms that the neural network model is sensitive to hyper-parameter selection and performs worse when the optimal optimizer, learning rate, or activation is changed. Notably, the model performs markedly worse when a large learning rate is leveraged. This is consistent with prior understanding of the learning rate, where high values may cause the model to converge too quickly upon a suboptimal result, which is clearly the case in the situation noted above.

Trade Off Considerations

For the supervised learning portion of this project, there is a noteworthy tradeoff between the training data size and accuracy of the model. The training data only includes approximately 3.6k Men's international matches. This does lead to the ability to perform very computationally intensive Grid Searches to arrive at optimally tuned Random Forest, SVR, and neural network models. However, this small sample size does raise accuracy concerns, especially with a fairly small negative MAE. Ideally, this negative MAE would be closer to 0, and this largely can be attributed to the insufficient amount of Rugby Union data available for processing. Therefore, while there were model preparation benefits, there is a considerable tradeoff where the accuracy and reliability of the models is a concern.

Failure Analysis

- Rugby World Cup 2019 match between Japan and Ireland

The neural network model predicts an ~24 point win for Ireland. However, Japan pulls off a stunning upset and wins by 7 points. This type of scenario highlights the huge reliability the model places on world rank. The world rankings suggested that a top-ranked Ireland would secure a lopsided win over a middle-of-the-pack Japan team. However, while this was a failed prediction, we likely wouldn't propose any future improvements. To predict such an event would require a likely very overfit model. These types of upsets are so rare that an incorrect prediction in these types of scenarios is acceptable to ensure greater generalizability of the model.

- 2019 Men's Internationals match between England and Ireland

The neural network predicts a ~2 point win for Ireland. However, England, the underdog based on world rankings, wins by 42 points. While these teams are fairly close to each other in terms of world rankings, such a blowout would be unexpected. As a result, this type of blowout win suggests that Ireland may have been resting their usual players or had key injuries that resulted in a team below full strength. As a future enhancement to the neural network model, it may be valuable to pull forward additional features that check the lineups fielded by the two teams to ensure that key players are available to better account for this type of scenario.

- Rugby World Cup 2019 Qualifying - Americas match between USA and Canada

The neural network predicts a very slight win for Canada in this match. However, the match ends in a tie. While a tie is a possibility in Men's Rugby Union, the model always predicts a winner and never predicts a match ending with a 0 point differential. This is likely driven by the limited dataset and the model being presented with very few ties to allow for this scenario to adequately be accounted for. As a future enhancement, more training data is required to allow for better model training, with the goal being that enough ties are present that this can be a trainable outcome.

2023 World Cup Prediction

As a fun next step in the process, the neural network was leveraged to predict the upcoming 2023 Men's Rugby Union World Cup. The specific results of the group stage, quarterfinals, semifinals, and final can

be found in the appendix, but, to summarize, the neural network predicted the host nation, number two ranked France to win the tournament. This was perhaps an expected outcome, as high world rankings play a significant role in the predictive power of the model. However, it does appear that homefield advantage may have played a role in France advancing over a team higher ranked than them.

Overall, this neural network does solve the question outlined in the introduction that yes, there are opportunities available in the supervised learning space to leverage modeling to better understand and predict sport outcomes, to perhaps gain an advantage in the betting landscape.

Part B. Unsupervised Learning

Methods description

Two commonly used models in sports analytics were used for the task of clustering players based on performance statistics: KMeans (centroid-based) and Agglomerative Clustering (hierarchical clustering).

KMeans was chosen based on literature mentioned earlier in this report and its success in identifying patterns in team performance. Similar to the Verna et. al, the first step was to use the UMAP embeddings in scikit-learn's KMeans function with the default parameter and use an elbow method to find our clusters. The elbow method gave an optimal clustering of 4. A KMeans model was run on $n_clusters$ of 4 with the remaining parameters left at the models default and evaluated using a silhouette score, resulting with a score of 0.453. To further analyze cluster performance, silhouette plots were used on cluster sizes ranging from 2 to 9. The results show many clusters had higher scores on average than the optimal provided by the elbow method with 9 being the highest at 0.493.

With the list of optimal clusters brought down to two possibilities, a GridSearch was utilized to tune the model and find the best combination of selections for $n_clusters$ (4 or 9), $init$, and max_iter . Based on the results of the GridSearch, it was safe to drop a clustering of 4 from our selection since it still performed much lower after tuning. Next, we applied an Agglomerative Clustering model to visually check the clusterings and see if they were similar to our KMeans. Figure 2, shows similar clustering across both, but the Agglomerative Clustering silhouette score was less, 0.47. Despite this lower score, and considering our task to find player performance patterns and reviewing the clusters we thought 9 was our best choice. Silhouette scores were more consistent over several iterations with 9 and the model had a high Calinski Harabasz and a low Davies Bouldin Score compared to other cluster sizes.

Table 4. Evaluation Scores for Various Cluster Sizes in KMeans Model

$n_clusters$	$silhouette_score$	$calinski_harabasz_score$	$davies_bouldin_score$
2	0.446	1405.764	0.914
3	0.505	2050.385	0.675
4	0.453	2086.065	0.806
5	0.463	2287.485	0.763
6	0.473	2337.568	0.717

7	0.490	2388.430	0.703
8	0.481	2610.783	0.740
9	0.493	2726.120	0.689

Unsupervised Evaluation

Various evaluation metrics were used on our model but most frequently used throughout was Silhouette Scores. It was used to measure how well each data point fit into its assigned cluster using euclidean distance. It was also used because it had less computational complexity when evaluating all the different parameter tunings done on our KMeans model. We are aware that scores can be misleading on overlapping clusters by assigning points to the wrong cluster, thus inflating the score. We can see this in figure 2 across each of our tested models. For example, PCA cluster 0 and 1 have heavy overlap and upon further analysis of its silhouette plot many data points in cluster 0 were incorrectly assigned showing negative values.

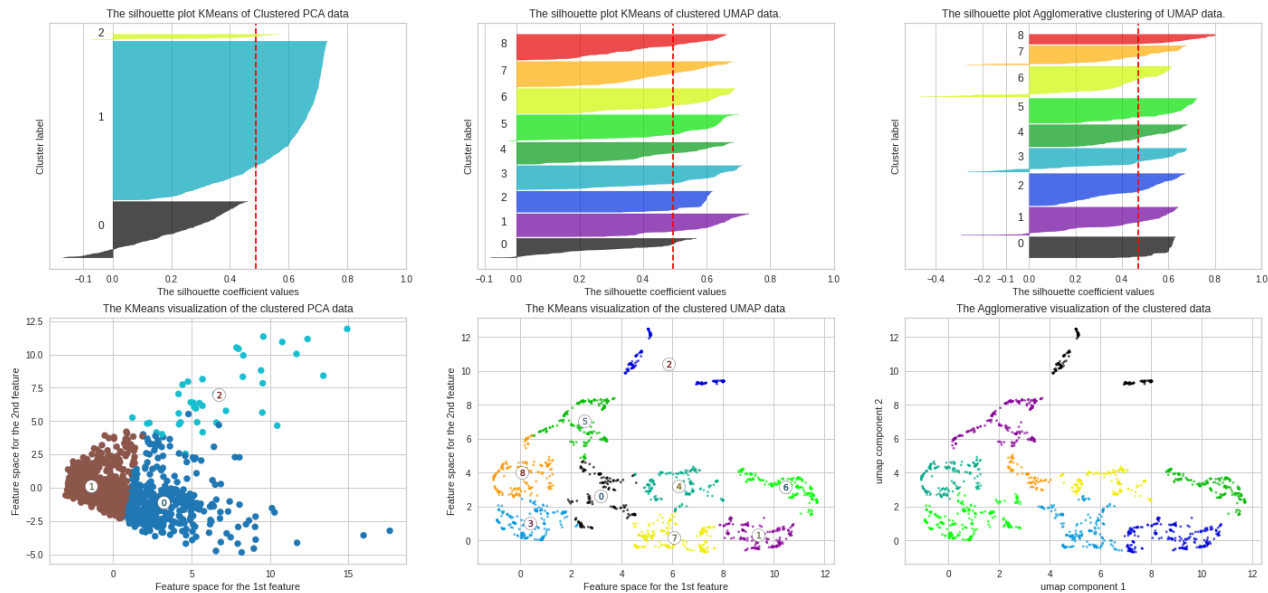


Figure 2. Cluster comparisons of best performing models for PCA and UMAP feature inputs in to KMeans. Third column, is the agglomerative clustering on UMAP features. First row, shows variation, spread, and additional details on cluster performance using silhouette score (dashed red line showing average score)

The other evaluation metrics, Calinski-Harabasz Index and Davies-Bouldin Score, were used as supplemental metrics to narrow down the list of cluster sizes after using the elbow method. With these metrics and our domain knowledge, it supported our choice to go with 9 clusters and our choice of using the KMeans model with the UMAP embeddings.

Table 4. Summary of Evaluation Metric over KMeans and Agglomerative Cluster Models

model	silhouette_score	calinski_harabasz_score	davies_bouldin_score
PCA KMeans	0.388	816.81	1.014
UMAP KMeans	0.493	2726.12	0.689

UMAP Agglomerative	0.470	2542.428	0.701
--------------------	-------	----------	-------

In the UMAP KMeans model, useful patterns were found in our player cluster assignments. The following are a few of them based on an initial review of the averaged player stats of the clusters³:

- Kicks and Passes are important features to clustering and on average players in this cluster are 86.2 kg and 180 cm tall (cluster 2)
- Clean Breaks and the number of defenders beaten are important to tries made as is the number of Try Assist and on average players in this cluster are 96 kg and 185 cm tall (cluster 6)
- Tackles are important to Turnovers Won and on average players in this cluster are 109 kg and 190 cm tall (cluster 1)

We mention average weight and height but it should be noted that when looking at the traditional positions in each cluster there were distinguishable differences between the majority and minority positions that made them up. So it's obvious that, for example, cluster 1 contained forward positions who are more involved around breakdowns forcing them to tackle more and who are typically the larger players. For any future use cases, it would be good to reference positions in these clusters to identify the most successful player combinations.

Sensitivity analysis

We assessed how sensitive our UMAP KMeans model with 9 clusters was to changes in two things: a model parameter and the input features. Utilizing GridSearch and 5-fold cross validation again, we tested performance on the initialization method. KMeans can be sensitive to the initial positions of its centroids, so both a k-means++ and random method were attempted. It showed very little change in average silhouette score over various splits between methods (k-means++: 0.489, random:0.490).

KMeans can also be sensitive to the scale or normalization of the input data so we tried min-max scaling on the features used in the PCA/UMAP combination. The result of scaling the input features was a lower score of 0.467. This makes sense when we conducted our initial exploratory data analysis and plotted each feature. Most of the features were highly skewed and the drawback of MinMax scaling is that it can compress values and lose information about those extreme values that the distance metric was sensitive to. The last thing tested was the stability of the model, as mentioned in the method description, the model provided the most consistent performance over repeated clustering processes.

Discussion

For Part A, I continued to learn and really become familiar with the various hyperparameters related to Random Forest and SVR models and how results may be impacted depending on the direction the values are tuned. Following this Milestone, I feel very comfortable developing, tuning, and training data on these models and how to best evaluate the effectiveness of the fit through cross-validation and metric calculation/scoring. I spent a lot of time tuning the models and testing various combinations of hyperparameters and again feel much more comfortable with this process and leveraging Grid Search relative to when I was first taking SIADS 542 Supervised Learning. Additionally, I also obtained valuable exposure to tensorflow and keras while developing my neural network model. This was my first time leveraging these libraries, and I hope to continue to leverage the lessons I learned in the future and within my current line of work. In thinking about the results of Part A, I was not necessarily surprised by the

³ Full table of average player stats can be found in the appendix

importance of the world rankings of each team and how important each was to the predictive power of the model, but I was disappointed that the other features looked to be fairly inconsequential to the success of the model. I suspected that home field advantage would be a more notable driver of team performance as well as recent match performance in the form of the moving averages for points for and against. Therefore, while this finding may not have been particularly shocking, it was surprising to me just how important these features became in predictive power over the others. In terms of challenges, the biggest issue encountered for this project was just a lack of data surrounding rugby matches. As a result, creativity was required to not only obtain match results but then feature engineer relevant information to use in the supervised learning method. For example, obtaining match data through github required persistence and strong research skills and developing the moving averages for points for and against was a relatively intensive process. Additionally, even though match data was obtained and features developed, we were still only left with ~3.6k matches. As a result, the models did suffer in their predictive power, but the lack of data led to the response of focusing heavily on parameter tuning. GridSearch was leveraged throughout the project and helped ensure that the models were optimal and could generalize well despite the lack of available training data. Lastly, if there was more time and resources available, I would have liked to have tried to include Women's International Rugby Union data as well as a way to increase the training data size and allow for more robust model generation. Additionally, I would have liked to spend more time with tensorflow/keras and understanding the underlying mechanisms that go into neural network model creation. However, given the time constraints, I feel as though the supervised learning portion of the project was a success and led to a model that can reasonably predict score differential and winners of Men's International Rugby Union matches.

For Part B, I became much more familiar with feature engineering and embeddings. Specifically, their various uses are based on the underlying structure of the dataset and the specific task. I initially understood PCA as just a way to reduce dimensionality and to avoid the curse of dimensionality, but seeing the results of it as input to clustering models like KMean and comparing it to the UMAP (or combination of PCA/UMAP) was very rewarding to see discernable clusters and improved scores. As seen in the notebook, I tried several workflows with various feature engineerings, methods for model selections, and hyperparameter tunings. I found visualizations and scoring methods together were really beneficial when evaluating my models at various stages of development. If scores did not make sense to me I would view the plotted clusters and silhouette plot to see if any changes to my model were actually effective. Or if I had no idea where to start with the number of clusters, I looked at the t-SNE and UMAP plots to make an estimate range of possible values. It was certainly a very iterative process attempting to build a model, with an iteration or two excluded from the method description because I did not have enough time to tune and evaluate. I was also able to web scrape which I had never done before by learning a new package.

I was surprised to see such a large number of clusters perform well on the data. A team is traditionally made up of a forward pack and a back line, with this in my mind I thought a cluster of 4-5 would have been the optimal solution. I say this because forwards can also be split into tight five and loose forwards, the latter usually averaging in weight to centres in the back line. Not only that but the back line can also be split by their kicking skills where often the 10 and full back provide relief in their teams 22 by kicking while the rest of the back line chase the ball and apply defensive pressure. It's certainly possible to have such a high clustering because with 15 positions on the field a myriad of player combinations arise as no two games are alike.

Similar to our supervised task, a challenge was obtaining a large enough dataset. With luck and extensive web searches I was able to get a reasonable amount of rows and features pertaining to player performance. I say with luck because shortly after scraping and saving the data to file, RugbyPass had

changed their website structure and rendered my scraper useless. Another challenge I faced in the unsupervised task was keeping track of various models and their iterations while trying to interpret evaluations. It required better structure in my code and helpful markdown. Without this it was difficult to compare each and determine which had better performance and fulfilled the task based on my domain knowledge. Like I mentioned, I was surprised by such a high number of clusters but an initial review did explain the assignments and made sense based on my rugby knowledge as a player. Which leads to how I would extend my solution given the time, I would do a further analysis of the clusters with other players, coaches, and rugby fans to gather qualitative data. I would also spend additional time on tuning the Agglomerative Clustering and the DBSCAN (this being the one of the iterations I excluded from the method description). Also given the time, I would use the clusters as features in our supervised task for match predictions. Then given the resources, I would obtain more player stats and features on players movement on the pitch in game.

Ethical Considerations

There were not many ethical issues to consider when providing a solution to Part A, but one important item to note is that developing score prediction models is a common practice within the gambling and sports betting space. It is important to gamble responsibly and, to address this issue for this particular project, I will not be leveraging this work for any type of gambling or betting and will simply predict results out of interest and to become more and more comfortable developing my own neural network models for future work.

Similar to Part A, Part B did not have many ethical considerations except for in the realm of fantasy sport betting. Although many rugby enthusiasts participate in the annual Six Nations Tournament fantasy league, no betting is involved. Aside from this annual tournament there are no other fantasy leagues for rugby year round.

Statement of Work

Adam led the development of the supervised learning methods, and Carolann led the development of the unsupervised learning methods. Each individual then shared equal responsibility in the final report write-up.

References

Christopher D. Long. "Rugby". Date Accessed: Jan 16, 2023.

(https://github.com/octonion/rugby/tree/master/world_rugby/csv)

P. Verma, B. Sudharsan, B. R. Chakravarthi, C. O'Riordan and S. Hill, "Unsupervised Method to Analyze Playing Styles of EPL Teams using Ball Possession-position Data," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 58-64, doi:

10.1109/ICACCS48705.2020.9074426. (<https://ieeexplore.ieee.org/document/9074426>)

Rory P. Bunker, Fadi Thabtah. "A machine learning framework for sport result prediction".

Applied Computing and Informatics, Volume 15, Issue 1, 2019, Pages 27-33.

(<https://www.sciencedirect.com/science/article/pii/S2210832717301485>)

Sascha Wilkens. "Sports Prediction and Betting Models in the Machine Learning Age: The Case of Tennis". Journal of Sports Analytics, Volume 7, Issue 2, 2021, Pages 99-117.

(<https://content.iospress.com/articles/journal-of-sports-analytics/jsa200463>)

World Rugby. "World Rugby Rankings - Men's Rankings". Date Accessed: Jan 20, 2023.

(<https://www.world.rugby/tournaments/rankings/mru>)

Rugby Pass. "Players- Team Rosters". Date Accessed: Feb 6, 2023.

(<https://www.rugbypass.com/players/>)

Appendix

A1. Features in our Unsupervised Learning Dataset

Scraped data files can be found here under players:

<https://drive.google.com/drive/folders/1tlq4LT7DaC75fZ0Dv3cej5z0Y5qgTjv0?usp=sharing>

feature	description	datatype
position	The position on a roster of 15	categorical
weight	Weight in kilograms	numerical
height	Height in centimeters	numerical
Points	Points scores by player	numerical
Tries	Number of tries made (worth 5 points)	numerical
Metres	Number of metres run with the ball	numerical
Runs	Number of runs with the ball	numerical
Defenders Beaten	Number of defenders a player got past	numerical
Clean Breaks	Number of times a player broke the defensive line to score	numerical
Passes	Number of passes made	numerical
Try Assists	Number of passes made to a scoring player	numerical
Kicks	Number of kicks in open play	numerical
Conversion Goals	Number of successful kicks converting tries with an additional 3 points	numerical
Penalty Goals	Number of successful kicks at post after a penalty was conceded	numerical
Tackles	Number of tackles made	numerical
Tackles Missed	Number of tackles missed	numerical
Turnovers Won	Number of turnover made to gain ball possession	numerical
Turnovers Conceded	Number of turnovers made to lose ball possession	numerical
Penalties Conceded	Number of infringement made my player	numerical

Yellow Cards	Number of Yellow Cards given to player who commits an offense under Law 9 – Foul Play of World Rugby	numerical
Red Cards	Number of Red Card given to a player who had already received two yellow card in a game or commits an intentional, dangerous, or reckless offense under Law 9 – Foul Play of World Rugby	numerical

A2. Notebooks

notebook_description	notebook_name	notebook_url
Code for the preprocessing, feature engineering and supervised task of project	Rugby Union - Supervised Learning.ipynb	https://drive.google.com/file/d/11hIGMQrmPPL7PBQwahTo0ykETDU-TLe8/view?usp=share_link
Webscraper for rugbypass.com gather player stats in to csv files	Rugbypass-scraper.ipynb	https://drive.google.com/file/d/1r1InoTWFvoD9if-FcQq1a-00a0geYP/view?usp=share_link
Code for the preprocessing, feature engineering and unsupervised task of project	rugbyplayer-unsupervised-cluster.ipynb	https://colab.research.google.com/drive/19Dd4l4V3LBCfj_5WUB1DGMdJru4UxzBp?usp=share_link

A3. 2023 Men's Rugby Union World Cup Predictions Leveraging Optimized Neural Network Supervised Model

2023 Men's Rugby Union World Cup - Group Stage Results

Group A	Record
France	4-0
New Zealand	3-1
Italy	2-2
Namibia	0-4
Uruguay	1-3

Group B	Record
Ireland	4-0
Romania	0-4
South Africa	3-1
Scotland	2-2
Tonga	1-3

Group C	Record
Australia	4-0
Georgia	1-3
Wales	3-1
Fiji	2-2
Portugal	0-4

Group D	Record
England	4-0
Argentina	3-1
Japan	2-2
Chile	0-4
Samoa	1-3

2023 Men's Rugby Union World Cup - Quarterfinals Results

France	South Africa
New Zealand	Ireland
Australia	Argentina
Wales	England

2023 Men's Rugby Union World Cup - Semifinals Results

Australia	Ireland
England	France

2023 Men's Rugby Union World Cup - Final Results

France	Ireland
--------	---------

Additional detail related to the predictions can be found within the available code and Jupyter Notebook.

Figure A4. Cluster results from UMAP KMeans showing average player statistics

feature	0	1	2	3	4	5	6	7	8
weight	106.30	109.52	86.31	117.69	98.43	86.46	95.37	115.11	107.02
height	186.64	190.49	179.72	193.79	184.18	178.49	185.04	192.12	184.67
Points	1.44	5.88	24.68	0.32	4.56	1.44	13.13	2.09	0.32
Tries	0.28	1.17	0.87	0.06	0.90	0.24	2.48	0.42	0.06
Metres	22.12	92.63	48.51	9.32	51.74	11.12	155.41	44.59	5.64
Runs	12.03	47.22	18.44	5.82	19.28	4.21	46.71	25.57	3.27
Defenders Beaten	0.66	2.39	3.40	0.17	2.37	0.57	7.99	0.86	0.08
Clean Breaks	0.20	0.94	1.18	0.06	0.92	0.21	3.34	0.31	0.01
Passes	4.69	19.70	167.08	1.64	10.93	17.87	28.06	8.68	1.33
Try Assists	0.80	3.19	3.50	0.21	2.13	0.62	6.00	1.25	0.13
Kicks	0.45	0.71	30.54	0.07	2.01	2.81	7.79	0.18	0.11
Conversion Goals	0.01	0.00	5.01	0.00	0.02	0.05	0.27	0.00	0.00
Penalty Goals	0.00	0.00	3.42	0.00	0.01	0.04	0.06	0.00	0.00
Tackles	20.09	71.38	26.97	11.30	22.46	5.13	37.24	44.77	6.04
Tackles Missed	3.00	8.47	6.54	1.31	3.88	1.20	7.72	4.94	0.82
Turnovers Won	1.04	4.83	1.39	0.53	1.29	0.24	2.72	2.31	0.29
Turnovers Conceded	2.08	6.70	5.80	0.87	3.29	0.96	7.81	3.11	0.54