Using Component Analysis to Developpe Performance Indicators in Internation Rugby Union: A Rugby World Cup 2023 Case Study

Arthur, Guillaume (N11371200)  
IFN703 Final Report

# Abstract

**A corpus of research on performance indicators in rugby union and rugby league suggests that dimension reduction techniques are efficient tools when analysing sporting datasets with many variables. Twenty-nine performance indicators from all 38 games of the Rugby World Cup 2023, collected by Rugbycology, were reduced to 8 orthogonal principal components. Logistic regression (match outcome) and linear regression (points difference) models were used to determine how well each component predicted success. The 8 principal components explained 83% of the variance in points difference and correctly predicted match outcomes. Results suggest that if a team increases their 'quick play on the counterattack' and 'gaining territory through kicks and ground play' component scores, they are more likely to win (OR=9.97 and OR=6.87, respectively). This highlights how PCA can be used to provide useful guidance on how teams can increase their chances of success by improving collections of variables, instead of focusing on variables in isolation.**

Table of Contents

[Abstract 1](#_Toc168750448)

[Introduction 2](#_Toc168750449)

[Literature Review 4](#_Toc168750450)

[The Project 6](#_Toc168750451)

[Definition 6](#_Toc168750452)

[Methods 7](#_Toc168750453)

[Samples 7](#_Toc168750454)

[Performance Indicators 7](#_Toc168750455)

[Statistical Analysis 7](#_Toc168750456)

[Results 8](#_Toc168750457)

[Principal Component Analysis 8](#_Toc168750458)

[Logistic Regression 11](#_Toc168750459)

[Linear Regression 12](#_Toc168750460)

[Models Comparison 14](#_Toc168750461)

[Components Analysis 14](#_Toc168750462)

[Discussion and Conclusion 15](#_Toc168750463)

[Appendix 18](#_Toc168750464)

[References 22](#_Toc168750465)

[Table 1: Principal component analysis results 10](#_Toc168750487)

[Table 2: Logistic Regression Results 12](#_Toc168750488)

[Table 3: Prediction Result for the logistic regression 12](#_Toc168750489)

[Table 4: Game outcome Prediction Linear Regression 13](#_Toc168750490)

[Table 5: Description of the performance indicators 18](#_Toc168750491)

[Table 6: Summary Statistics 19](#_Toc168750492)

[Table 7: Results of Backward Linear Regression 20](#_Toc168750493)

[Table 8: Loading, eigenvalue and % of explained variance for the 8 selected components 20](#_Toc168750494)

[Table 9: Wrong Predictions for each model 21](#_Toc168750495)

[Figure 1: Amount of total variance explained by the selected components (right) and cumulative sum of variance explained(left) 9](#_Toc168750528)

[Figure 2: Estimated correlations (over 0.5) between variables and components 10](#_Toc168750529)

[Figure 4: Odds ratios following the logistic regression 12](#_Toc168750530)

[Figure 5: Projection of the training and testing set over components 2 and 3. Annotations are the predicted probability of the event being a win for the wrong predictions. Losses are in blue and wins in yellow. 13](#_Toc168750531)

[Figure 6: Score gap predicted using all and 5 components 14](#_Toc168750532)

[Figure 7: Predicted score gap vs predicted probability of winning 14](#_Toc168750533)

[Figure 8: Results for components 2 (left) and component 3(right) 16](#_Toc168750534)

[Figure 9:Contribution of variables to the variance explained by each component 20](#_Toc168750535)

# Introduction

The continuous increase in computer capabilities has brought data analytics as a key companion for decision-makers in all areas of business and society. In the sports landscape, the study of on-the-field action reflects an increase in the adoption of data analysis. Sports have an important quantitative aspect. Be it in team or individual events, games are won or lost based on scores. Actions such as goals, tackles, and distance covered are all quantitative measures that help describe a game and the performance of a team or a player.

Data analytics takes such measurements and derives qualitative information that can then help inform game decisions. Instead of relying on general wisdom, it is now more common to apply the scientific method to gain a deeper understanding of a game. Data collection and analysis are useful tools that can be used to inform decisions by multiple stakeholders, including coaches, trainers, managers, or even players.

The enthusiasm for sports analytics has been building since 1950 and emerged from distinct communities: operations researchers, freelance sports journalists, and internet hobbyists [1]. "Operations Research" was a new branch of warfare in which statisticians analysed military operations to improve their efficiency. And what better thing could those OR officers do to relax than apply their knowledge to various sports? A 1959 article from a Royal Canadian Air Force officer, for example, described how right-handed batters had a higher batting average against left-handed pitchers, and vice versa [2]. He naturally suggested that managers should substitute a player depending on the handedness of the opposing pitcher.

Baseball focused a lot of interest from statisticians in the 1970s in the US through, not exclusively, the Society for Baseball Research founded in 1971, and the movement never lost momentum. In the nineties, it was the turn of American football with the publication of the pivotal work "The Hidden Game of Football" in 1988. This work illustrates the creation and use of performance metrics, such as expected points depending on the position on the field.

Booming in North America, the rest of the world was not ignorant. Charles Reep, a former wing-commander in Britain’s Royal Air Force, dismayed by the poor performance of his beloved Swindon team, started recording game actions on a notepad. This led to the study of more than 578 soccer games between 1953 and 1967 and can be seen as the spark of game analysis in football.

Rugby is a close-contact team "invasion" sport, in which teams try to infiltrate the opposition’s territory to score a goal. A team is composed of 15 players, using an oval-shaped ball on a rectangular field. Born in 1845 in England, it was split into rugby league and rugby union in 1895. Contrary to its counterpart, there is no limit to the amount of time a team can keep possession of the ball. Rugby Union remained an amateur sport until 1995, when the International Rugby Board declared it a professional sport.

Compared to the other sports mentioned previously, rugby is a relatively new federation. Since then, it has garnered interest worldwide, which has increased the need for performance analysis and game modeling.

Little work has been done in rugby union to model game outcomes co. This can be attributed to the lower amount of data available compared to other sports, such as soccer. This study aims to investigate how performance indicators at the team level can be used to study what combinations of metrics are important in relation to the game outcome.

## Literature Review

In Rugby, various approaches have been utilized to study the relationships between performance indicators and outcomes. Performance indicators are quantifiable metrics used to assess the performance of a player or team. In this review, our aim is to outline the research done exploring the relationship between one or multiple performance indicators and a positive outcome.

Our objective is to provide an overview of data analysis in rugby by showcasing results and highlighting potential unexplored paths.

Part one: Performance Indicators

The research in rugby union has primarily focused on single or combinations of performance indicators deemed relevant to team success, such as tackle success, while other studies have concentrated on physical and technical requirements, such as running. These studies often assume a linear, causal, and direct relationship between the performance indicator and the outcome. However, there appears to be little implementation of the conclusions drawn from these studies in practice by coaches, which may indicate a potential reluctance to adopt performance analyses, or more realistically, a lack of strength and consensus in the observations made [3].

Research is often conducted without sufficient consideration of the applicability and utility of the findings. Rugby performance is complex and requires coordination between players as well as individual excellence. In a recent literature review published by Colomer et al. [3], it was noted that there is a lack of data available from the period between 1997 and 2019; only 41 studies were included after screening. Most of the data were collected from games played between 2000 and 2008, and various levels of play were studied. These factors make it challenging to reach a definitive consensus and raise doubts about tactical analyses, especially considering that sports strategies are rapidly evolving.

Out of the 41 studies, only 22 contextualized the data, and a mere 5 utilized multiple variables, highlighting once again the lack of complex analyses. Overall, it is observed that little has changed over the covered time span regarding performance indicator analyses, with most studies relying on univariate measures of performance.

However, it was suggested that match outcomes were better predicted by relative datasets. Bennett et al.[4] described a model consisting of 10 performance indicators that have a significant relationship with match outcomes: kicks from hand, clean breaks, average carry distance, penalties conceded when the opposition has the ball, turnovers conceded, total meters carried, defenders beaten, and the ratio of tackles. Their models predicted the results with 80% accuracy at best, suggesting that a good proportion of the complexity is not captured using these performance indicators.[3]

Part two: overview of technique used for performance Analyses

Analysts can measure specific technical skills in each sport. These measurements help decision-makers understand different aspects of play. Further application of notational analysis can elucidate collective behavior, manifested in common styles of play, which can then model relative outcomes like match success. It was noted by Lago-Peñas et al.[5] and argued that this information can be used by coaches to counter opposition. However, the amount of data being collected is enormous, and no consensual framework of analysis exists. How do we interpret the information we receive? Multiple techniques can be employed [6].

The first step is to reduce the amount of data treated by selecting meaningful predictors and filters. In this process, large multidimensional datasets are condensed into smaller sets without losing information. Techniques commonly used in sports include Principal Component Analysis (PCA)and multidimensional scaling, both of which produce factors representing groups of similar variables and have been used to study performance indicators.

Although these techniques help us understand successful performance, coaches and analysts are more interested in which features can be modified to achieve better outcomes. Decision support analysis helps interveners identify interactions and factors, aiming to determine the probabilities of certain outcomes.

Modeling the probability of events in a team match is an area of vivid interest. Being able to account for multiple variables is a powerful tool. Logistic regression is used to predict dichotomous events (two possible outcomes). A major advantage of regression is that it provides the direction and strength of interaction between the variables and the outcome event. Parmar et al. [7] (redo)demonstrated how to combine PCA and logistic regression to model the probability of success in rugby league, noting that a team had a 91% chance of winning if outperforming its opponent on certain metrics [6].

In this study, we will have a look at performance indicators collected at the team level during the rugby world cup 2023.

Exploration of tactical performance has been conducted qualitatively, and three Rugby World Cup coaches reported unsuccessful performance connected with an irrational kicking game, weak defensive line, and losing possession of set-piece play [8]. Vaz et al. hypothesized that in a close game in the Super Twelve, a kicking-based game supported by an effective defensive structure is more likely to win than a possession-based one [9]. These studies hint at the necessity of a better understanding of tactical analysis in Rugby Union.

World cup format makes it harder for the favorite teams to prevail in the end; the knockout format makes upsets more likely. The winning team needs to finish among the best teams of its group in the knockout stage. The five teams of each group play only four games at that stage. Three direct elimination games follow before the best team can be crowned champion. Gaining a tactical advantage against a specific opponent therefore holds a lot of value.

# The Project

## Definition

Our goal will be to investigate how PCA can be used to study the relationship between performance indicators and match outcomes in international rugby union. We will more specifically study games during the Rugby World Cup 2023.

Reviewing existing literature, we can conclude that a common flaw in many studies is focusing solely on the relationships between single performance indicators and match outcomes. Such an approach often lacks depth and fails to capture the nuanced complexities of the game. We will take inspiration from the work of Parmar et al., who studied the relationship between game outcome and performance indicators as well as with the difference in scoring, in rugby league [10].

## Methods

### Samples

Data from 48 games played in the rugby world cup 2023 were collected by Rugbycology on a spreadsheet provided b Dimitri Perrin. These were extracted using Python (version 3.10). To enable clear comparisons between winning and losing teams, a draw was excluded. A summary of the statistics in performance from loosing sides and winning sides are presented in appendix (Table f).

### Performance Indicators

Relative (Team A minus Team B) frequencies for all action variables were used as predictor variables. Twenty-five indicators were used. PIs that related to scoring were excluded from the analysis, such as Score, Tries, Conversions, Goal Kicks Successful. The % of successful rucks, scrums, tackles, and goal kicks were also calculated. A description is presented in appendix (Table e). The dataset created contained more wins (36) than losses (11).

### Statistical Analysis

All data were analyzed using Python. The sklearn decomposition PCA method was used (sklearn version 1.4.1.post, ) to reduce the dimension of the dataset and overcome multicollinearity issues. Multicollinearity arises when dependent variables are correlated. Component scores were extracted using the **components\_ method**. The eigenvalue for each component were extracted using the **explained\_variance\_** method. After rotation, 8 components, which explained 84.85% of the variance, were retained for having eigenvalues >1. An eigenvalue of 1 means that the principal component would explain about one variable’s worth of the variability. The rationale is that each component should explain at least one variable’s worth of the variability.

Before the PCA analysis, the data was scaled with the standard scaler of the sklearn library, which removes the mean and scale to unit variance. This is necessary as PCA maximizes variance and would load on the largest variance otherwise.

Evaluation of the models Using the **sklearn.model\_selection.train\_test\_split** with random\_seed of 6 and the stratify parameter set to the outcome of a game for the logistic regression study.

The principal component scores saved from the PCA were run in both logistic (win/loss) and linear regression (statmodels version 0.14.1) using a random selection of 70% of the data. The library **sklearn.model\_selection.train\_test\_split** was used to split the data, with random\_seed of 6 and the stratify parameter set to the outcome of a game. The GLM method with the parameter family set as Binomial was used for the logistic regression, and the OLS method was used for the linear regression. The models were then used to predict match outcome and point difference on the remaining 30% of games. In both models, a constant was added to the data, before the split in training and testing, since the values of loading are not centered. That was done using the **add\_constant** method.

To assess more accurately how the predictive logistic model performed, we compared the number of correct predictions on matches in which the score was close. An even game is defined by a difference in score of less than 15 points (2 possessions). Eighteen games were classified as even.

## Results

### Principal Component Analysis

The 8 components are selected and the contribution from each factor to each component is presented in Figure 2. The 8 components explained 85% of the variance (Figure 1), with the first component explaining the most amount, with 36%. The percentage explained then drastically decreased to 11% and 10% for the two next components.

A graph with colorful squares

Description automatically generatedA graph with a line

Description automatically generated

Figure 1: Amount of total variance explained by the selected components (right) and cumulative sum of variance explained(left)

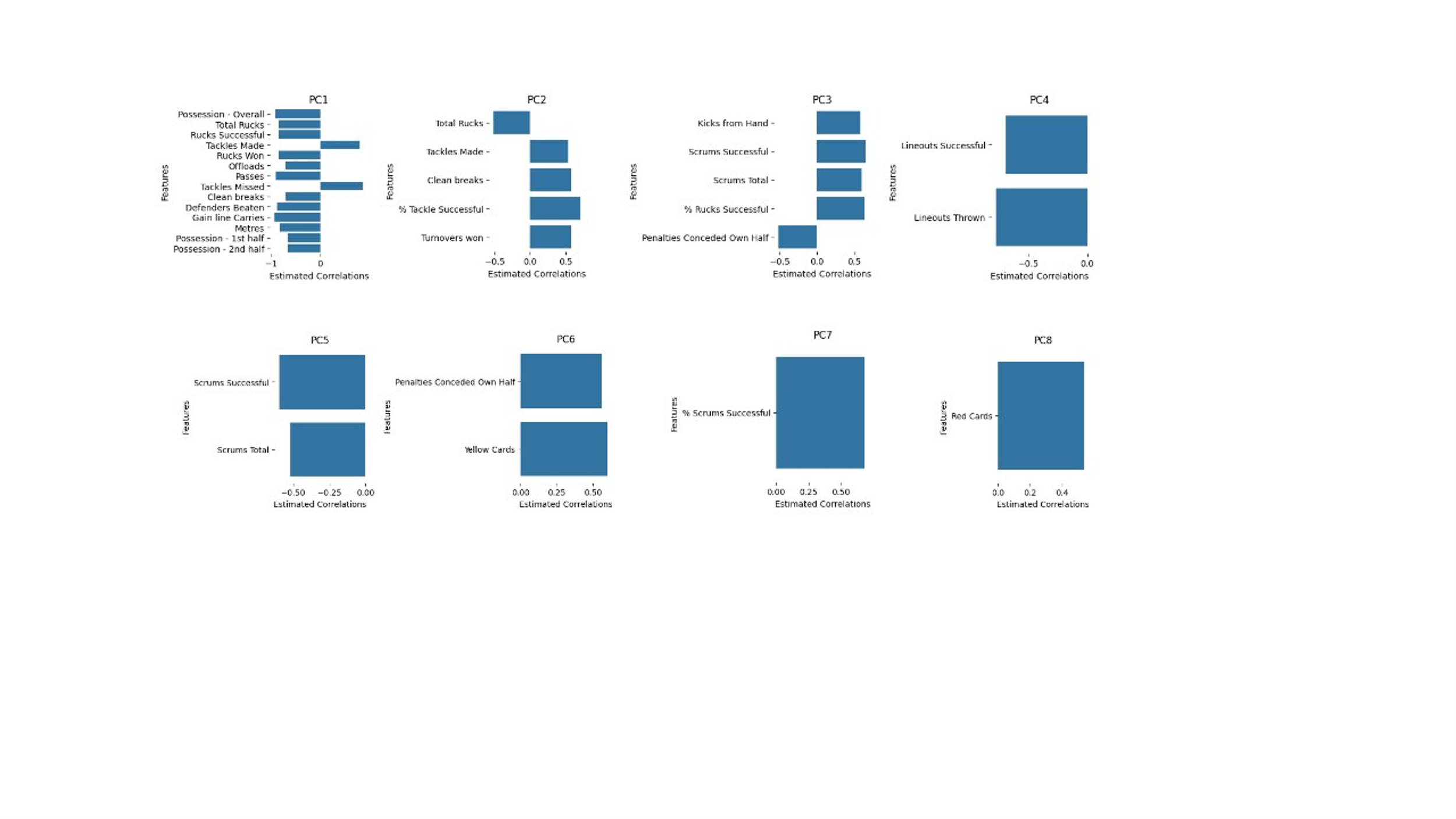
The contributions of PIs to the components scores through the estimated correlations. A positive PI improved the component score. A negative value reduced the component score (Figure 2). To characterize the said components, we also visualized the ratio of variance explained for each variable, in each component (appendix Figure 9). From that we can assign a meaning to each component. For the first component, the most important are Gain line Carries, Possession – Overall, Passes and Defender Beaten. Taking together, those metrics hint at the possession, and the ability of a team to keep the ball.

Figure 2: Estimated correlations (over 0.5) between variables and components

The second component is loaded with the number of turnovers won, the number of clean breaks, and more tackles. Making more total rucks than your opponent decreased the score of the component. Overall, metrics centered around the defense and possibly the ability of making quick breaks. It is logical to observe opposite loading for clean breaks and total rucks: a ruck starts when the play is broken with a tackle. A clean break is when a player goes through a gap in coverage and runs into space, something that is stopped by tackles. It seems to describe a pattern of play where a team does not seek contact, but rather progresses the ball by avoiding it.

Component 3 seems to describe a strategy of domination of territory, with the number of successful scrums, % of successful ruck, number of kicks from hand as well as penalties conceded own half.

Component 4 is centered around the lineouts, more precisely making less lineouts and being less successful at them increase the value of the component. A lineout is how the play restarts after the ball has been kicked out of play, for example.

Component 5 is centered around the scrum. Creating less scrums than the opponent or being less successful at it increased the component score.

Metrics associated with fowls have the most importance for component 6, with yellow cards, penalties in own half, penalties.

The component 7 is mainly explained by the % of successful scrums, and the main contributor to component 8 is receiving more red cards than the opponent. Results are summarize in Table 1.

Table 1: Principal component analysis results

|  |  |  |  |
| --- | --- | --- | --- |
| **Component** | **Meaning** | **Variance Explained** | **Main Contributors (r>0.5)** |
| 1 | Possession | 36% | Possession, Total Rucks, Rucks successful, Tackles made, Rucks win, Offloads, Passes, Tackles missed, Clean Breaks, Defender beaten, Gain line carries, meters |
| 2 | Quick breaks without contact | 12% | Total Rucks, Tackles made, Clean breaks, % Tackle successful, Turnover won |
| 3 | Gaining Territory through kicks and ball recovery | 11% | Kicks from hand, Scrums successful, Scrums total, %rucks successful, Penalties conceded own half |
| 4 | Lineouts | 7% | Lineouts successful, Lineouts thrown |
| 5 | Bad Scrum Performance | 6% | Scrum Successful, Scrums Total |
| 6 | Fowls | 4% | Penalties conceded own half |
| 7 | %Successful scrums | 4% | % scrums successful |
| 8 | Red Cards | 3% | Red cards |
| Unexplained |  | 15.15% |  |

### Logistic Regression

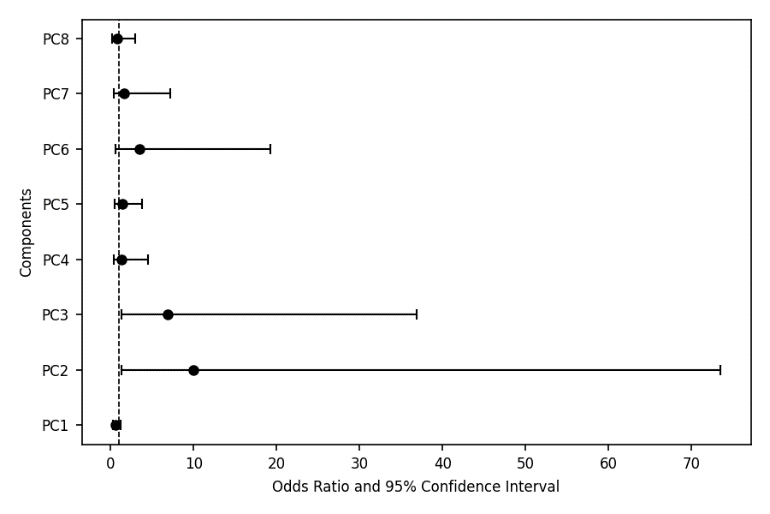
These same 8 components were then fitted through a logistic regression model. Table 1 presents the results of the analyses. Components 2 and 3 were shown to significantly predict the binary outcome of game, with an odd ratio of 9.97 and 6.87 respectively. This means that improving the value of component 2 by 1, while all other components staying equal, improved the probability of winning by 897%. Prediction using all 8 components correctly predicted a win 100 % of the time on the testing set and training set. However, the model correctly identified only 50% of the losses in the testing set. Those results show a lot at overfitting and highlight the challenges to dealing with class unbalance, as well as a low number of samples. Using 2 components yields similar results with 90% of win correctly predicted on the testing set and 92% on the training. Similarly, the model failed to correctly predict 50% of the losses (Table 2).

Figure 4: Odds ratios following the logistic regression

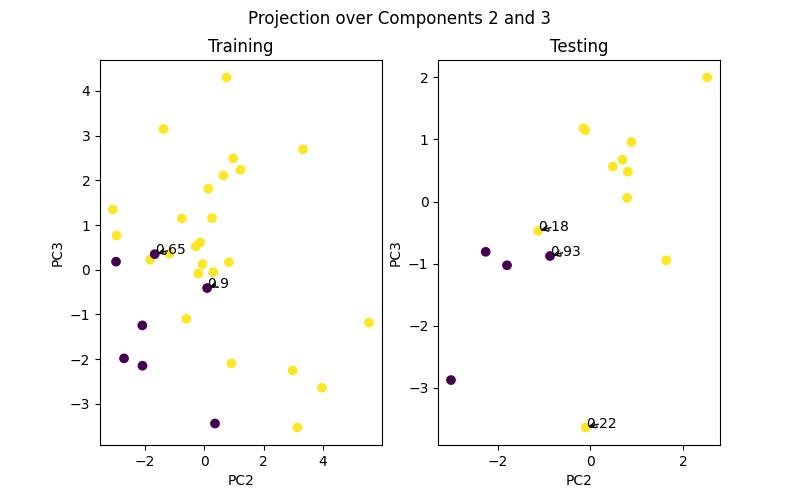
Table 2: Logistic Regression Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***β*** | **P-value** | **Odd Ratio** | **95% Confidence Interval of *β*** | | **% increase chance of Winning** | |
| **Lower** | **Upper** |  | |
| PC1 | -0.52 | 0.16 | 0.60 | 0.29 | 1.23 |  |
| **PC2** | **2.30** | **0.02\*** | **9.97** | **1.35** | **73.45** | 897% | |
| **PC3** | **1.93** | **0.02\*** | **6.87** | **1.28** | **36.82** | 587% | |
| PC4 | 0.24 | 0.71 | 1.27 | 0.35 | 4.56 |  | |
| PC5 | 0.33 | 0.52 | 1.39 | 0.51 | 3.77 |  | |
| PC6 | 1.24 | 0.16 | 3.46 | 0.62 | 19.24 |  | |
| PC7 | 0.46 | 0.55 | 1.59 | 0.35 | 7.22 |  | |
| PC8 | -0.27 | 0.70 | 0.77 | 0.20 | 2.97 |  | |

Table 3: Prediction Result for the logistic regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | % Win Correctly predicted  All components | % Loss Correctly predicted  All components | % Win Correctly predicted  2 components | % Loss Correctly predicted  2 components |
| **Training** | 100% | 50% | 92% | 50% |
| **Test** | 100% | 100% | 90% | 80% |

Figure 5: Projection of the training and testing set over components 2 and 3. Annotations are the predicted probability of the event being a win for the wrong predictions. Losses are in blue and wins in yellow.

Looking more precisely at the result along the projection over components 2 and 3(Figure 5)we observe that all the loss are aggregated in the bottom left corner. The model wrongly predicted result for performance at the border of wins and losses. This might indicate that more data might help the model PCA further differentiate these 2 categories. Annotations on the graphs display the probability of the performance being a win, according to the model. It does not look like most of those games were on the fence, around 50%. Looking at the misclassified games, they were not close encounter, with some battering, like Tonga winning by 21 points October 14th. However, it also included 2 games where the score difference was a single point. On average, the absolute score difference was 10.57. This is less than 2 possessions, which hint that our model often failed in contested matches. This a behaviour that we would expect, however, it’s in the most contested games that studying tactics and performance might be more useful.

### Linear Regression

Using a linear regression model and backward selection, we identified that the components 1,2,3,6 and 8 were sufficient to get best the prediction. We observe that the coefficients for the first and last components are negative, meaning that a decrease in possession increase the score difference, which seem counterintuitive at first glance (appendix, Table 3). For component 8, more red cards mean a negative score difference which makes sense.

Using the score difference yielded similar result as the logistic regression model, predicting 96% of the wins in the training and testing set (Table 4). Similarly, the model performed badly in predicting losses, with 33% for the training test and 80 % on the testing set. Looking more precisely at the predicted score vs actual score difference with all the components, we observe a decent R2 of 0.84 of and a RMSE of 11.24 for the training test and 12.83 for the testing set, hinting at some overfitting. Selecting a lower number of components seem to have reduced the overfitting, the RMSE for the training set increasing to 12.23, and the RMSE for the testing set decreasing to 11.02 (Figure 5).

Table 4: Game outcome Prediction Linear Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | R2 | % Win Correctly predicted  All components | % Loss Correctly predicted  All components | % Win Correctly predicted  3 components | % Loss Correctly predicted  3 components |
| **Training** | 0.84 | 96% | 66% | 96% | 40% |
| **Test** |  | 100% | 80% | 90% | 60% |

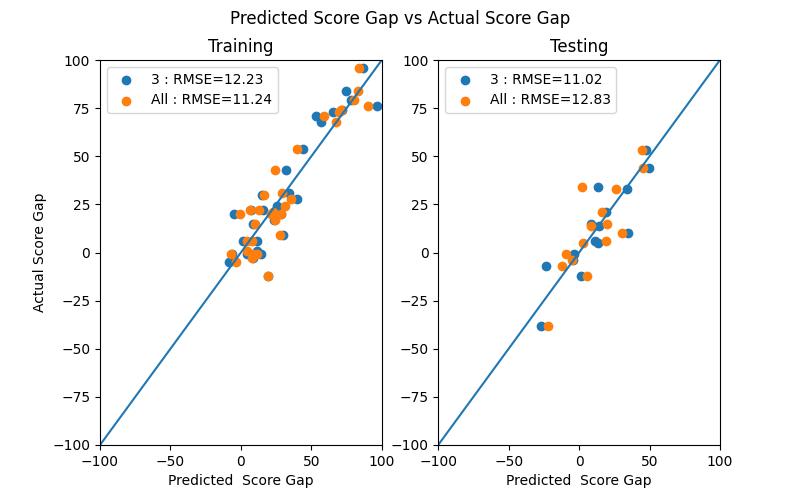


Figure 6: Score gap predicted using all and 5 components

### Models Comparison

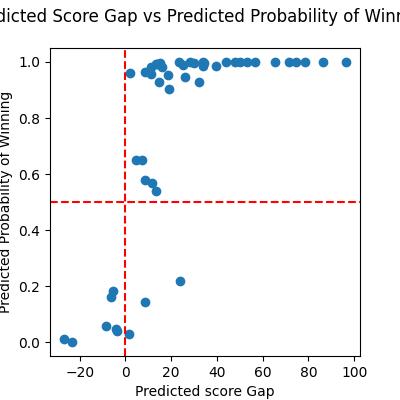
In aggregate, both models predicted the same wrong outcome 4 times (appendix Table g). There is a disagreement on the predicted winner at 3 occasions (Figure 6). The dataset being unbalanced, looking at how the model performed on even encounter might give us a good benchmark. There are 18 games which ended with a score difference of less than 14. Amongst them, 8 are loss and 10 are wins. The logistic model correctly identified all the wins, however 3 of the loss were predicted as wins.

Figure 7: Predicted score gap vs predicted probability of winning

### Components Analysis

PCA was conducted on 29 PIs with the contribution of each component to the principal component shown through estimated correlation. If a PI had a positive value, it improved the component score.

We obtained the best results using the logistic regression and 2 components. The next step is to describe both components, in Figure 8. We will set a correlation threshold of 0.5, to work with correlation of medium strength between each variable and components. We observe two different sets of PIs: for component 2, making more tackle than the opponent, creating more clean breaks, being more successful at making tackles and creating more turnover lead to an increase in the component value. On the contrary, making more rucks decreases the value. It’s a mixture of features relating to defence (Tackles made, % Tackles successful, Turnover won) and ball progression through carrying the ball (clean breaks and evading tackles that lead to a ruck). For the third component, more kick from hand, more total and successful scrums, as well as a greater % successful ruck increases its value. Conceded less penalties than your opponent in your own half also increases the value of the component. Those 2 combinations of attributes could define specific tactics: one made of successful carries and efficient defensive structures, and another made of more kicks and more infractions created. A scrum is allowed when a team commits a minor infringement, like passing the ball forward, or the ruck become unplayable. A team could also elect to go for scrum instead of a penalty. Taken with the fact that the away team taking more penalties increase the value of the component, this might suggest a tactic of teams putting pressure on the opposition, attracting minor fowl, and going for a scrum. This looks to be a more territorial approach, with a team kicking the ball to put it in the opposition half, and then regaining possession through forcing a fowl or keeping the play alive after a tackle. There is trend of strong play on the ground in that component.

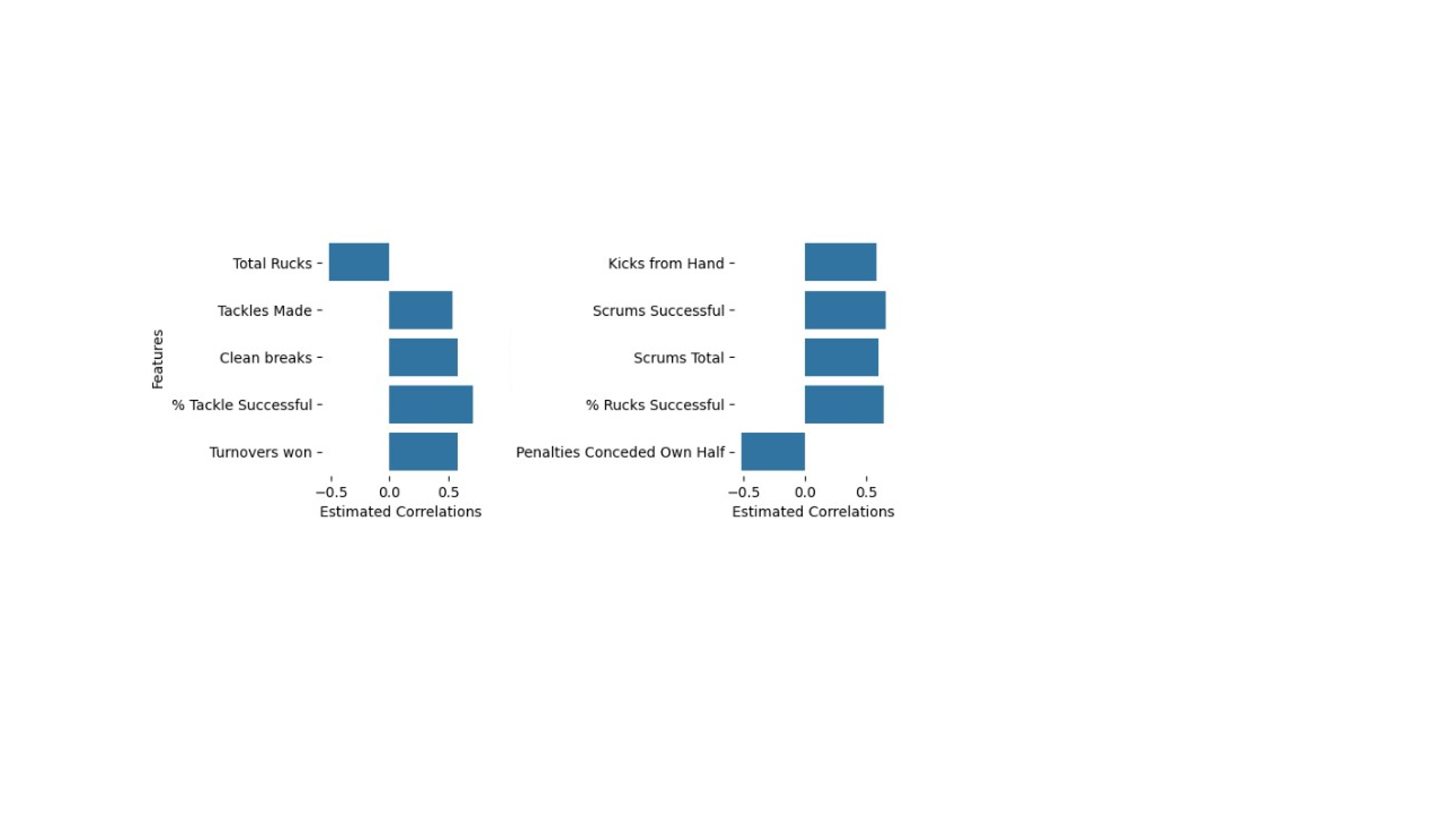


Figure 8: Results for components 2 (left) and component 3(right)

# Discussion and Conclusion

Identification of metrics that lead to success is a major part of performance analysis. Coaches are looking for ways to understand how to improve performance. A major roadblock in this process is attempting to provide reliable performance indicators. This study aimed to a) reduce the data set while retaining as much of the variance as possible using PCA, and b) assess the sustainability of the principal components in predicting match outcomes (logistic regression) and final points differences (linear regression).

The PCA created 8 components, which we aimed to group into C categories, explaining 84.85% of the variation in the dataset. These were possession (35%), defensive action and counterattack (11%), and infringement, with 15.15% of the variation not explained.

These variables have the potential to significantly impact the outcome of a game, as evidenced through regression results. We found that a team making fewer rucks, more tackles, more clean breaks, and creating more turnovers increases its chances of winning by a large margin. Similarly, for teams kicking more from hand, being more successful in the scrum and ruck, and forcing more penalties against them in the opposite half. However, the selected components did not account for a large proportion of the variance (around 10%). It is therefore hard to say they are key factors.

The logistic regression and backward linear regression retained components 2 and 3, with the linear regression including infringements, as well as the first component loaded with most of the variables. Those results are expected, as precisely predicting the difference in points is more difficult than predicting the outcome. The fact that not all the components were necessary to predict the outcome suggests that our model is too simplistic and will benefit from more data. Similarly, the World Cup is an event where the best teams will be overrepresented in the dataset because they will be involved in more matches. A victory might be due to the relative strength of the teams, and not necessarily tactics or strategies.

The use of PCA was suggested to be a sound approach when utilizing regression methods that suffer from multicollinearity issues. This case shows that even if sporting events can follow patterns to some extent, some wins can display unusual patterns of data, possibly due to tactics, a send-off, or injuries.

Future studies should include more data. Only 48 games were not enough, and during tests, trying to balance the number of wins and losses in the dataset had a major impact on the nature and loading of the components. Working with data from multiple World Cups will help test the relevance of our conclusions. A similar approach could be used to profile teams and study what metrics are related to a win for each of them. This is warranted, seeing the large confidence interval for the odds ratios, indicating a large variability of the predictor on the outcome depending on the game.

This study reproduced a method that could serve as guidelines on how teams can increase their chances of success during international events. By improving linear combinations of variables, as opposed to individual variables, the conclusions of this study are more realistic.

# Appendix

Table 5: Description of the performance indicators

|  |  |
| --- | --- |
| **Performance Indicator** | **Description** |
| Possession - Overall | The percentage of playing time the team had possession of the ball. |
| Possession – 1st Half | The percentage of playing time the team had possession of the ball during the first half. |
| Possession – 2nd Half | The percentage of playing time the team had possession of the ball during the second half. |
| Meters | Distance Covered with the ball |
| Defenders Beaten | The number of defenders evaded by the ball carrier |
| Clean breaks | A breach of the line of defenders by the player in possession of the ball |
| Gain line Carries | A crossing of the center of the pitch when there is a breakdown in open play, such as ruck, maul or scrum. Advancing across the gain line represents a gain in territory. |
| Passes | A pass is to transfer a ball to a teammate by throwing it. |
| Offloads | A short pass made by a player being tackled before he reaches the ground |
| Turnovers won | When a team regain possession of the ball by stealing it. |
| Kicks from Hand | Kicking the ball before it bounces |
| Rucks Won, Total Rucks, Rucks Successful | A ruck is formed when the ball is on the ground and two opposing players meet over the ball |
| Tackles Made, Tackles Missed | A tackling on the player in possession |
| Lineouts Successful | A lineout in which the team keeps possession of the ball |
| Lineouts Thrown | Number of Lineouts Taken |
| Scrums Successful, Scrums Total | A restart of the play during which the team in possession keeps possession. |
| Penalties Conceded Own Half | Number of Major fouls committed in the team’s own half |
| Penalties Conceded Opp Half | Number of Major fouls committed in the team’s opposition half |
| Penalties | Awarded for serious infringement. The team can then kick for goal, take a scrum, run or kick into touch |
| % Successful Tackle | Computed from the PI provided |
| % Successful Rucks | Computed from the PI provided |
| % Goal kicks Successful | Computed from the PI provided |
| % Scrums Successful | Computed from the PI provided |
| Form | Computed from the PI provided |
| Pre-game win probability | Computed from the PI provided |
| **Scoring Related – Not included** | |
| Score | The number of points |
| Tries | The number of 5 points try scored |
| Conversions | The number of 2 points conversions |
| Goal Kicks Successful | A kick of the ball through the uprights |

Table 6: Summary Statistics

|  |  |  |
| --- | --- | --- |
|  | **Mean ± std** | |
|  | **Losses** | **Wins** |
| **Score** | 25.6 ± 18.54 | 28.91 ± 24.55 |
| **Possession - Overall** | 50.94 ± 7.29 | 49.3 ± 7.49 |
| **Possession - 1st half** | 50.09 ± 9.13 | 50.0 ± 9.91 |
| **Possession - 2nd half** | 51.79 ± 10.14 | 48.62 ± 10.01 |
| **Tries** | 3.09 ± 2.92 | 3.7 ± 3.84 |
| **Conversions** | 2.43 ± 2.58 | 2.89 ± 3.4 |
| **Penalties** | 1.64 ± 1.67 | 1.45 ± 1.49 |
| **Metres** | 452.23 ± 178.1 | 491.62 ± 201.17 |
| **Defenders Beaten** | 23.04 ± 10.91 | 24.91 ± 11.54 |
| **Clean breaks** | 5.72 ± 4.22 | 7.38 ± 5.4 |
| **Gain line Carries** | 58.21 ± 20.2 | 60.02 ± 20.63 |
| **Passes** | 142.68 ± 55.14 | 140.6 ± 41.33 |
| **Offloads** | 6.51 ± 4.55 | 7.36 ± 4.87 |
| **Turnovers won** | 5.85 ± 2.41 | 5.81 ± 2.53 |
| **Kicks from Hand** | 24.85 ± 8.6 | 24.64 ± 8.22 |
| **Rucks Won** | 76.21 ± 21.6 | 75.11 ± 21.38 |
| **Tackles Made** | 122.36 ± 31.75 | 134.45 ± 41.96 |
| **Tackles Missed** | 23.28 ± 12.28 | 23.83 ± 9.81 |
| **Goal Kicks Successful** | 4.06 ± 2.67 | 4.34 ± 3.11 |
| **Goal Kicks Attempted** | 5.26 ± 2.89 | 5.66 ± 3.29 |
| **Rucks Successful** | 76.21 ± 21.6 | 75.11 ± 21.38 |
| **Total Rucks** | 79.85 ± 21.95 | 78.57 ± 21.45 |
| **Lineouts Successful** | 11.55 ± 4.27 | 11.13 ± 3.7 |
| **Lineouts Thrown** | 13.64 ± 4.46 | 13.21 ± 3.78 |
| **Scrums Successful** | 5.64 ± 2.52 | 5.81 ± 2.37 |
| **Scrums Total** | 7.47 ± 3.06 | 7.04 ± 2.62 |
| **Penalties Conceded Own Half** | 5.28 ± 2.37 | 5.85 ± 2.27 |
| **Penalties Conceded Opp Half** | 4.66 ± 2.09 | 4.21 ± 1.93 |
| **Yellow Cards** | 0.66 ± 0.76 | 0.49 ± 0.66 |
| **Red Cards** | 0.09 ± 0.28 | 0.09 ± 0.28 |
| **% Rucks Successful** | 0.95 ± 0.02 | 0.95 ± 0.03 |
| **% Goal kicks Successful** | 0.72 ± 0.28 | 0.72 ± 0.29 |
| **% Scrums Successful** | 0.76 ± 0.17 | 0.82 ± 0.15 |
| **% Tackle Successful** | 0.85 ± 0.05 | 0.85 ± 0.06 |

A group of blue and red graphs

Description automatically generated

Figure 9:Contribution of variables to the variance explained by each component

Table 7: Results of Backward Linear Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Coef** | **P-value** | **95% Confidence Interval of *coef*** | |
| **Lower** | **Upper** |
| **PC1** | **-4.76** | **0.00 \*** | **-6.51** | **-3.14** |
| **PC2** | **10.44** | **0.00 \*** | **7.92** | **13.82** |
| **PC3** | **5.32** | **0.00\*** | **1.76** | **8.52** |
| **PC6** | **3.59** | **0.03** | **0.31** | **6.86** |
| **PC8** | **-3.76** | **0.05** | **-7.49** | **-0.03** |

Table 8: Loading, eigenvalue and % of explained variance for the 8 selected components

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** | **PC8** |
| Possession - Overall | -0.92 | -0.11 | 0.27 | -0.01 | -0.02 | 0.07 | 0.08 | 0.05 |
| Possession - 1st half | -0.66 | 0.09 | 0.30 | 0.39 | 0.04 | -0.11 | 0.23 | 0.07 |
| Possession - 2nd half | -0.66 | -0.25 | 0.10 | -0.37 | -0.08 | 0.20 | -0.09 | 0.02 |
| Penalties | 0.34 | 0.20 | 0.44 | 0.05 | 0.14 | -0.43 | -0.06 | 0.31 |
| Metres | -0.83 | 0.31 | 0.00 | 0.23 | -0.15 | 0.12 | -0.10 | -0.10 |
| Defenders Beaten | -0.88 | 0.28 | -0.24 | 0.01 | -0.17 | 0.01 | 0.12 | 0.10 |
| Clean breaks | -0.70 | 0.58 | -0.05 | 0.16 | -0.02 | 0.09 | 0.04 | 0.02 |
| Gain line Carries | -0.93 | 0.06 | 0.16 | 0.18 | -0.14 | -0.01 | -0.02 | -0.04 |
| Passes | -0.90 | 0.01 | -0.09 | 0.08 | -0.02 | -0.01 | -0.10 | 0.08 |
| Offloads | -0.71 | 0.32 | -0.07 | 0.23 | -0.06 | 0.09 | -0.15 | -0.15 |
| Turnovers won | 0.15 | 0.59 | 0.44 | 0.13 | 0.44 | 0.01 | -0.16 | -0.23 |
| Kicks from Hand | 0.23 | 0.39 | 0.58 | 0.19 | 0.04 | 0.00 | 0.39 | -0.01 |
| Rucks Won | -0.84 | -0.49 | 0.19 | -0.01 | 0.07 | -0.07 | -0.05 | -0.03 |
| Tackles Made | 0.82 | 0.54 | -0.18 | 0.02 | -0.03 | 0.04 | -0.01 | 0.05 |
| Tackles Missed | 0.88 | -0.28 | 0.24 | -0.01 | 0.17 | -0.01 | -0.12 | -0.10 |
| Rucks Successful | -0.84 | -0.49 | 0.19 | -0.01 | 0.07 | -0.07 | -0.05 | -0.03 |
| Total Rucks | -0.84 | -0.51 | 0.14 | -0.02 | 0.02 | -0.08 | -0.03 | 0.00 |
| Lineouts Successful | -0.48 | 0.27 | -0.12 | -0.69 | 0.25 | -0.01 | 0.22 | 0.14 |
| Lineouts Thrown | -0.43 | 0.21 | -0.19 | -0.78 | 0.25 | -0.03 | -0.01 | 0.11 |
| Scrums Successful | 0.26 | 0.11 | 0.66 | -0.19 | -0.60 | 0.17 | -0.05 | 0.18 |
| Scrums Total | 0.25 | 0.05 | 0.60 | -0.16 | -0.52 | 0.16 | -0.32 | 0.36 |
| Penalties Conceded Own Half | 0.21 | -0.02 | -0.52 | 0.45 | 0.12 | 0.56 | -0.12 | 0.12 |
| Penalties Conceded Opp Half | 0.09 | -0.43 | -0.36 | 0.33 | -0.46 | -0.43 | 0.13 | -0.10 |
| Yellow Cards | 0.22 | -0.35 | 0.14 | -0.18 | -0.21 | 0.60 | 0.11 | -0.29 |
| Red Cards | -0.05 | -0.24 | 0.18 | 0.47 | 0.33 | 0.24 | 0.27 | 0.54 |
| % Rucks Successful | -0.33 | 0.14 | 0.64 | 0.08 | 0.48 | 0.13 | -0.20 | -0.22 |
| % Goal kicks Successful | -0.22 | 0.47 | 0.11 | 0.08 | -0.30 | -0.19 | -0.39 | -0.19 |
| % Scrums Successful | 0.08 | 0.13 | 0.38 | -0.12 | -0.24 | 0.02 | 0.68 | -0.31 |
| % Tackle Successful | -0.38 | 0.71 | -0.42 | -0.05 | -0.22 | 0.06 | 0.15 | 0.11 |
| Eigenvalue | 10.58 | 3.61 | 3.21 | 2.25 | 1.90 | 1.31 | 1.28 | 1.01 |
| % Variance explained | 0.36 | 0.12 | 0.11 | 0.07 | 0.06 | 0.04 | 0.04 | 0.03 |

Table 9: Wrong Predictions for each model

|  |  |  |
| --- | --- | --- |
| Game ID | Predict score | Predicted Probability of winning |
| 87849 | 8.63 | **0.32** |
| 8762 | **-0.92** | **0.37** |
| 87864 | **-10.19** | 0.67 |
| 87871 | **11.9** | **0.90** |
| 87872 | 19.77 | **0.48** |
| 87874 | -6.00 | **0.65** |
| 87875 | **3.43** | 0.17 |
| 87877 | **28** | **0.68** |
| 87879 | **3.24** | **0.64** |
| 87880 | **3.77** | 0.27 |
| 87881 | **4.75** | 0.10 |

# References

[1] E. Hintz, ‘Sports Analytics Before Moneyball’, Smithsonian, Museum of American History. [Online]. Available: https://invention.si.edu/sports-analytics-moneyball

[2] G. R. Lindsey, ‘Statistical Data Useful for the Operation of a Baseball Team’, *Oper. Res.*, vol. 7, no. 2, pp. 197–207, 1959.

[3] C. M. E. Colomer, D. B. Pyne, M. Mooney, A. McKune, and B. G. Serpell, ‘Performance Analysis in Rugby Union: a Critical Systematic Review’, *Sports Med. - Open*, vol. 6, no. 1, p. 4, Jan. 2020, doi: 10.1186/s40798-019-0232-x.

[4] M. Bennett, N. Bezodis, D. A. Shearer, D. Locke, and L. P. Kilduff, ‘Descriptive conversion of performance indicators in rugby union.’, *J. Sci. Med. Sport*, vol. 22, no. 3, pp. 330–334, Mar. 2019, doi: 10.1016/j.jsams.2018.08.008.

[5] M. G.-R. Carlos Lago-Peñas and G. Yang, ‘Styles of play in professional soccer: an approach of the Chinese Soccer Super League’, *Int. J. Perform. Anal. Sport*, vol. 17, no. 6, pp. 1073–1084, 2017, doi: 10.1080/24748668.2018.1431857.

[6] C. J. Wedding, C. T. Woods, W. H. Sinclair, and A. S. Leicht, ‘Operational Insights into Analysing Team and Player Performance in Elite Rugby League: A Narrative Review with Case Examples’, *Sports Med. - Open*, vol. 8, no. 1, p. 140, Dec. 2022, doi: 10.1186/s40798-022-00535-7.

[7] T. S. Neil Watson Sharief Hendricks and I. Durbach, ‘Integrating machine learning and decision support in tactical decision-making in rugby union’, *J. Oper. Res. Soc.*, vol. 72, no. 10, pp. 2274–2285, 2021, doi: 10.1080/01605682.2020.1779624.

[8] C. M. E. Colomer, D. B. Pyne, M. Mooney, A. McKune, and B. G. Serpell, ‘A qualitative study exploring tactical performance determinants from the perspective of three Rugby World Cup coaches’, *Int. J. Sports Sci. Coach.*, vol. 17, no. 4, pp. 734–741, Aug. 2022, doi: 10.1177/17479541221087384.

[9] L. Vaz, M. V. Rooyen, and J. Sampaio, ‘Rugby Game-Related Statistics that Discriminate Between Winning and Losing Teams in Irb and Super Twelve Close Games.’, *J. Sports Sci. Med.*, vol. 9, no. 1, pp. 51–55, 2010.

[10] N. Parmar, N. James, G. Hearne, and B. Jones, ‘Using principal component analysis to develop performance indicators in professional rugby league’, *Int. J. Perform. Anal. Sport*, vol. 18, no. 6, pp. 938–949, Nov. 2018, doi: 10.1080/24748668.2018.1528525.

[11] D. Rojas-Valverde, J. Pino-Ortega, C. D. Gómez-Carmona, and M. Rico-González, ‘A Systematic Review of Methods and Criteria Standard Proposal for the Use of Principal Component Analysis in Team’s Sports Science.’, *Int. J. Environ. Res. Public. Health*, vol. 17, no. 23, Nov. 2020, doi: 10.3390/ijerph17238712.