

# Foundations of expected points in rugby union: A methodological approach

Journal of Sports Analytics  
Vol. 11(0): 1–14  
© The Author(s) 2025  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/22150218251365220  
journals.sagepub.com/home/san



Guillermo Martinez-Arastey<sup>1</sup> , Naomi Datson<sup>2</sup> , Neal Smith<sup>1</sup>   
and Matthew Robins<sup>3</sup>

## Abstract

This study explores the feasibility of an Expected Points metric for rugby union, aiming to shift performance analysis from descriptive indicators to a predictive metric of possession quality. Notational analysis was conducted on 132 Premiership Rugby matches, producing a dataset of 35,199 unique phases of play containing variables such as team in possession, pitch location, play type, score differences, time remaining and scoring outcomes. Four machine learning algorithms were explored to predict scoring outcomes: multinomial logistic regression, random forest, support vector machine and k-nearest neighbors. After extensive feature engineering and hyperparameter optimisation, the best-performing model achieved 39.7% accuracy, below a literature-derived baseline for practical usability (44.3%), making it unsuitable for applied contexts. A key challenge was predicting minority scoring outcomes due to severe class imbalance. SMOTE was explored to address this imbalance, resulting in a lower accuracy (35.7%) but an improved 34.4% F1-score. This study highlights the limitations of modelling scoring outcomes in open-play team sports, challenging the predominant positivist paradigm in sports performance analysis. The methodology provides critical foundational groundwork and a benchmark for future research to build upon. It recommends exploring advanced samplers for minority classes, expanded feature sets and alternative modelling techniques, such as recurrent neural networks.

## Keywords

Sports performance analysis, key performance indicators, machine learning, predictive modelling, match analysis

Received: 14 February 2025; accepted: 14 July 2025

## Introduction

The analysis of performance in invasion team sports begins with assessing the successful execution of individual plays (Yurko et al., 2019). Conventional studies have approached this problem by relying on descriptive methods to establish key determinants of success and draw generalised conclusions on which on-pitch actions are most influential to winning (Bremner et al., 2013; James et al., 2005). A common methodology has been narrowing down multiple performance factors into a concise set of key performance indicators with strong associations to scoring or winning matches, such as territory gained (Hunter and O'Donoghue, 2001), attacking and defensive profiles (Hendricks et al., 2013), set piece outcomes (Jones et al., 2004) or tackling success (Hughes et al., 2012).

However, reducing rugby union analysis to isolated measures oversimplifies the complexity of human behaviour in sports, as it assumes linear and predictable outcomes from simple cause-and-effect observations (Colomer et al.,

2020). This descriptive approach fails to account for situational context. For example, a 10-metre carry may hold different value if performed at a team's own 22-metre line or the opponent's 22-metre line. Therefore, establishing success factors using descriptive methods on, for instance, carries (Bishop and Barnes, 2013) or metres gained (Watson et al., 2017), in isolation conflates explanatory power with true predictive capability (Shmueli, 2010).

<sup>1</sup>Department of Sports and Exercise Science, University of Chichester, Chichester, UK

<sup>2</sup>Department of Sport and Exercise Sciences, Manchester Metropolitan University Institute of Sport, Manchester, UK

<sup>3</sup>School of Natural Sciences, University of Kent, Kent, UK

### Corresponding author:

Dr Guillermo Martinez-Arastey, Department of Sports and Exercise Science, University of Chichester, College Lane, Chichester, PO19 6PE United Kingdom.  
Email: gmartaras@gmail.com



This paradigm has resulted in poor generalisability and inconsistent results, causing a profound lack of scientific consensus, with over 392 unique performance indicators identified in the literature (Colomer et al., 2020). This reflects the absence of a common framework capable of capturing how complex interactions between performance indicators in a given situation influence match outcomes. Identifying success determinants in the sport may first require the quantification of the expected situational value of match scenarios. Predictive modelling may bring research in rugby union a step closer to that goal. However, over the last two decades, rugby union has lagged behind other sports in the application of contemporary data analytics methods, with 80% of published articles omitting these critical contextual considerations (Colomer et al., 2020).

This methodological challenge has been exacerbated by issues with data accessibility. Unlike sports like American football, where organisations like the National Football League (NFL) publish vast, free play-by-play datasets (Romer, 2006), major rugby bodies have not offered similarly detailed records. Researchers are therefore forced to rely on time-consuming manual notational analysis from match footage (Bremner et al., 2013; Vahed et al., 2016). This has restricted studies to small sample sizes, averaging  $67 \pm 91$  matches, undermining their ability to produce statistically significant conclusions with broad applicability (Bishop and Barnes, 2013; Colomer et al., 2020).

Sports such as American football (Romer, 2006; Yurko, 2017) or rugby league (Kempton et al., 2016) have recognised this methodological limitation and progressed by developing standardised metrics to account for contextual factors influencing performance. These efforts, supported by the vast publicly and commercially available data in these sports, have resulted in the development of an Expected Points metric. Expected Points assigns a singular points value to each game scenario. This value is derived from the probabilities of all possible scoring outcomes (and their associated point values) occurring next, given the current play's context (Burke, 2008; Carter and Machol, 1971).

Carter and Machol (1971) introduced the concept of Expected Points in American football by analysing 8,373 plays from 56 games of the 1969 NFL season to quantify possession value at specific field locations. Expected Points were calculated by adding the products of each possible scoring outcome's true value and its probability of occurrence, expressed as  $EP = \sum_i V_i \times P_i$ , where  $V_i$  represents the point value of outcome  $i$ , and  $P_i$  is the probability of its occurrence. This formula implies that an essential prerequisite for calculating Expected Points is, first, being able to reliably predict the probability of each scoring event ( $P_i$ ). Carter and Machol (1971) used this approach to estimate the impact of different actions on match scores and assess the effectiveness of technical and tactical decisions. Expected

Points were then used to develop strategic recommendations for various game situations (Katz and Burke, 2016).

The theoretical framework established by Carter and Machol (1971) on Expected Points methodology in American football inspired numerous subsequent studies to expand on this foundational work and improve the scientific rigour of its calculation (Goldner, 2017; Romer, 2006). It also gained traction beyond academia, extending to NFL clubs, media and its fan base (Causey, 2015). Burke (2008) popularised the concept of average net point advantage through their website Advanced Football Analytics, challenging conventional performance metrics by arguing that the value of performance metrics is relative to field position.

The proliferation of Expected Points across American football literature produced several approaches for deriving expected values, such as dynamic programming (Romer, 2006), absorbing Markov chain models (Goldner, 2017), bootstrapping (Causey, 2015), linear regression (Burke, 2008) and logistic regression (Yurko, 2017). Adaptations of the Expected Points metric also emerged. Burke (2010) developed Expected Points Added (EPA) to quantify the change in Expected Points between plays and assess their effectiveness. Katz and Burke (2016) also developed positional, player-level EP metrics by distributing EPA among all players involved in a play, including a Total Quarterback Rating (QBR). Expected Points research also ventured into other sports, such as rugby league (Kempton et al., 2016), ice hockey (Thomas, 2006), basketball (Cervone et al., 2016), Australian rules football (O'Shaughnessy, 2006) and association football (Green, 2012).

Despite its similarities with other invasion team sports, rugby union has yet to fully embrace model-based analytical approaches prevalent in the NFL and rugby league. The primary aim of this study is to explore the extension of the Expected Points framework to rugby union, assessing its feasibility and practicality. It builds on the hypothesis originating from NFL studies (Burke, 2008; Carter and Machol, 1971; Yurko et al., 2019) that the development of a model that reliably estimates the points value of any given match situation has the potential to change the way rugby union is analysed and understood. As illustrated in Table 1, this approach assigns a quantifiable points value to each unique match situation by multiplying the points awarded from each scoring method (e.g., +3 points for a scored penalty kick) by their modelled probability. The aggregation of all these products represents the overall estimated points value for that specific match situation.

This study represents a foundational proof-of-concept for the feasibility of reliably deriving such probabilities. The objective is not to deliver a deployable Expected Points metric but to rigorously document the methodological process, establish a performance benchmark and

**Table 1.** Example illustrating the estimated value of a possession through expected points, assuming reliable outcome probabilities ( $P_i$ ) could be modelled.

Scoring outcome	Awarded points	Modelled probability	Expected Points
Scored try & conversion	+7	26%	+1.82
Scored try	+5	20%	+1.00
Scored penalty kick	+3	15%	+0.45
Scored drop goal	+3	3%	+0.09
End of half (no scoring)	0	15%	0
Conceded drop goal	-3	2%	-0.06
Conceded penalty kick	-3	10%	-0.30
Conceded try	-5	4%	-0.20
Conceded try & conversion	-7	5%	-0.35
Estimated possession value		100%	+2.48

transparently identify the primary obstacles to creating such a model for rugby union. The aim is to provide the critical groundwork that will guide future research in the shift of the research paradigm in sports performance analysis in rugby union from descriptive statistical methods to predictive modelling techniques (Shmueli, 2010).

The successful development of an Expected Points metric in rugby union has the potential to provide practitioners with a standardised, universally interpretable framework to benchmark performance, enabling consistent evaluations and coherent comparisons of possession value across teams and match scenarios. The quantification of scoring probabilities across a range of contextual factors could help coaches adapt tactics, exploit opponent weaknesses, prioritise specific strategies and make more informed decisions on play selection. The analysis of Expected Points fluctuations between plays could indicate how a team's actions affect their scoring chances. Coaching practices could be informed by a detailed evaluation of a team's over- or under-performance relative to the Expected Points value of particular contexts. Player performance analysis could also gain greater consistency with the evaluation of individual actions based on their relative contribution to the team's overall Expected Points.

The development process of Expected Points in rugby union presented in this study also aims to provide transparency and reproducibility of its methodology to inspire future sports performance analysis research. Previous Expected Points studies have lacked comprehensive explanations of their statistical methodologies (Carter and Machol, 1971) or have failed to share model performance evaluations necessary to demonstrate the generalisability of results (Romer, 2006; Yurko, 2017). This study aims to overcome such gaps by presenting a detailed account of data pre-processing (Kotsiantis et al., 2006), feature engineering (Zheng and Casari, 2018), hyperparameter optimisation (Feurer and Hutter, 2019), cross-validation

(Kohavi, 1995) and model performance evaluation processes (Powers, 2020).

## Methods

### Participants

Event-level data for 35,199 phases of play was collected through the notational analysis of all 132 matches played during the 2018/19 English Premiership Rugby season. Twelve rugby union clubs competed in the round-robin competition, playing each opposing team twice over 22 rounds. This match sample was 95% larger than the mean sample size (67 matches) reported in rugby union literature (Colomer et al., 2020), and also exceeded the 8,373 plays analysed by Carter and Machol (1971) and 11,112 by Romer (2006) in early NFL Expected Points studies.

### Data collection

Full match video recordings were analysed using Sportscode Elite (Version 10, Hudl, Nebraska, United States) by notating key data points at every breakdown and start of play. Full match video recordings were obtained from publicly available broadcasts. The use of this footage for notational analysis falls under fair dealing principles for non-commercial research purposes. The operational definitions in Table 2 ensured the validity and reliability of the descriptive variables notated (Williams, 2012). These definitions were cross-validated using the existing literature, such as the Rugby Union Video Analysis Consensus publication by Hendricks et al. (2020). Descriptive variables included phase sequence number, team in possession, pitch location, play type, match score, points difference, match clock and disciplinary cards.

Each notation corresponded to an individual phase of play, defined as the instance when the scrum-half retrieved the ball from the breakdown to begin a new phase. For scenarios that did not begin from a breakdown (e.g., match start, set pieces or turnovers), each data point reflected the moment the ball was first collected by the new team in possession. Upon scoring, the event (e.g., tries, penalty kicks, drop goals or end of the half) was assigned to all phases since the previous scoring event. Phases by the scoring team were labelled with the positive scoring outcome (e.g. scored try), while those by the conceding team were attributed the opposite outcome (e.g. conceded try).

Due to the possibility of observational errors, notational analysis was repeated on 13 randomly selected matches representing 10% of the dataset to test intra-observer and inter-observer reliability (O'Donoghue, 2007). Intra-observer reliability testing indicated high consistency,

**Table 2.** Operational definitions of data variables.

Variable	Definition
Phase	Period between subsequent rucks. A ruck is formed when at least one player from each team is in contact, on their feet and over the ball, which is on the ground (Hendricks et al., 2020).
Team in Possession	Team handling the ball at a given point in time (Ungureanu et al., 2019).
Location	The location on the pitch where the phase began, determined by the metre lines dividing the rugby union pitch.
Side	The location on the pitch where the phase began, determined by the lineout lines dividing the rugby union pitch.
Play Type: Scrum	A set piece for restarting play after penalties with scrum option, offsidess, unplayable mauls or incorrect lineout throws. A scrum is formed when eight players from each team engage with their opponents so that the heads of the front rows are interlocked (Hendricks et al., 2020).
Play Type: Lineout	A set piece restarting play after the ball has been taken out or kicked to touch. A lineout is formed on the mark of touch with teams forming a single line parallel to the mark of touch on their side of the lineout between the 5-metre and 15-metre lines (Hendricks et al., 2020).
Play Type: Quick Tap	A quickly taken penalty where a player taps the ball with their foot and surges forward (ESPN, 2022).
Play Type: Restart Kick	A match is started, or restarted, with a drop kick from behind the centre of the halfway line (Nakagawa, 2006).
Play Type: Kick	A kick in open play which does not go into touch, whether intentionally or not (Eaves et al., 2005).
Play Type: Turnover	When one side takes possession of the ball from their opponents (ESPN, 2022).
Score	The exact score on the scoreline at the time the phase began.
Time	The exact match clock time in minutes and seconds.
Cards: Yellow	A player who receives a yellow card from the referee has to leave the pitch for ten minutes and sit in the Sin Bin (ESPN, 2022).
Cards: Red	A player is sent off for the remainder of the game or for a period of 20 minutes, depending on the infraction (ESPN, 2022).
Outcome: Try	A score awarded when the ball is touched down on the ground by a player across the try line. It includes penalty tries awarded by the referee for defensive foul play during try-scoring plays (ESPN, 2022).
Outcome: Penalty Kick	An uncontested kick awarded to a team for a major infraction by the other team, taken directly at goal (ESPN, 2022).
Outcome: Drop Goal	A kick between the posts by an attacking side, where the ball must hit the ground before being kicked (ESPN, 2022).

with a percentage error of 1.4%. Inter-observer reliability also produced an acceptable error rate of 4.42% and a kappa coefficient of 0.907 (95% CI: 0.902–0.912,  $p$ -value  $\leq 0.05$ ).

### Data analysis

Analysis was conducted using the programming language Python (Python Software Foundation, Python Language Reference, version 3.10). Python scripts were written in Google Colaboratory (Google LLC, Mountain View, California), a cloud-based Jupyter notebook service (Kluyver et al., 2016). The dataset was exported as a comma-separated values file from Sportscodelite and imported into Google Colaboratory for exploratory data analysis, data pre-processing, model training and model performance evaluation.

### Exploratory analysis and preliminary modelling

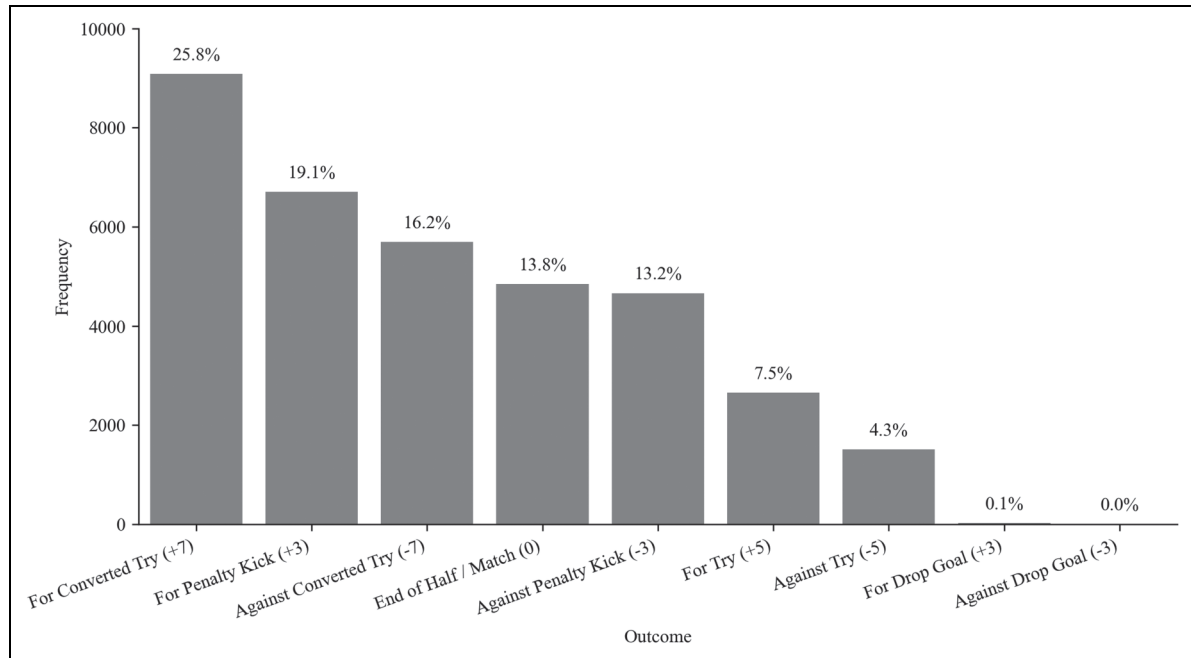
Initial exploratory data analysis primarily assessed the distribution of scoring outcomes to identify any bias or skewness, as class imbalance could increase the likelihood of classification errors (Japkowicz and Stephen, 2002). The analysis showed class imbalance with 25.8% of phases leading to converted tries, while drop goals (both scored

and conceded) only accounted for less than 0.1% of phases (see Figure 1). Moreover, penalty kicks were scored in 19.1% of phases and conceded in 13.2%.

Preliminary modelling conducted using default hyperparameter configurations from Scikit-Learn (Pedregosa et al., 2011) showed comparable results across models (Table 3). The support vector machine (SVM) achieved the highest accuracy (32.1%) and F1-score (29.9%), while the other models performed similarly, with accuracies ranging from 25.2% (k-nearest neighbors) to 30.6% (multinomial logistic regression). A key observation was the substantial model bias towards predicting majority classes, such as tries scored and the end of the half, impacting both precision and recall across all outcomes.

### Feature engineering

Following preliminary observations, the rugby union dataset was refactored to improve predictive power. Tries with and without conversions were consolidated, and phases leading to drop goals were also excluded due to their rarity (0.1% of phases). As shown in Figure 2, these changes reduced class imbalance, with the imbalance ratio increasing from 0.17 to 0.40 (He and Garcia, 2009).



**Figure 1.** Frequencies of different scoring outcomes showing an imbalanced dataset.

**Table 3.** Accuracy and F1-score across four classification models.

Model	Accuracy (%)	F1-score (%)
Multinomial logistic regression	30.6	26.9
Random forest	29.4	28.4
K-nearest neighbors	25.2	24.8
Support vector machine	32.1	29.9

### Predictive modelling

The development of an Expected Points model in rugby union was treated as a classification problem given the discrete nature of scoring outcomes (Yurko, 2017). This characteristic makes regression models inappropriate for predicting point scores, as their residuals fail to conform to the assumption of normality. A more effective strategy involves developing classification models that treat each scoring method as a distinct category, independent of its point allocation (Yurko et al., 2019).

Four classification algorithms were selected: multinomial logistic regression for its effectiveness in identifying linear relationships (Kleinbaum et al., 2008), random forest for its handling of high-dimensional data and categorical variables (Ho, 1995), support vector machine for its proficiency with imbalanced datasets (Wu and Chang, 2005) and k-nearest neighbors for its effectiveness with non-linearly separable data (Cover and Hart, 1967). While more advanced techniques such as gradient boosting machines or deep learning models exist, the primary goal of this foundational study was to first assess feasibility

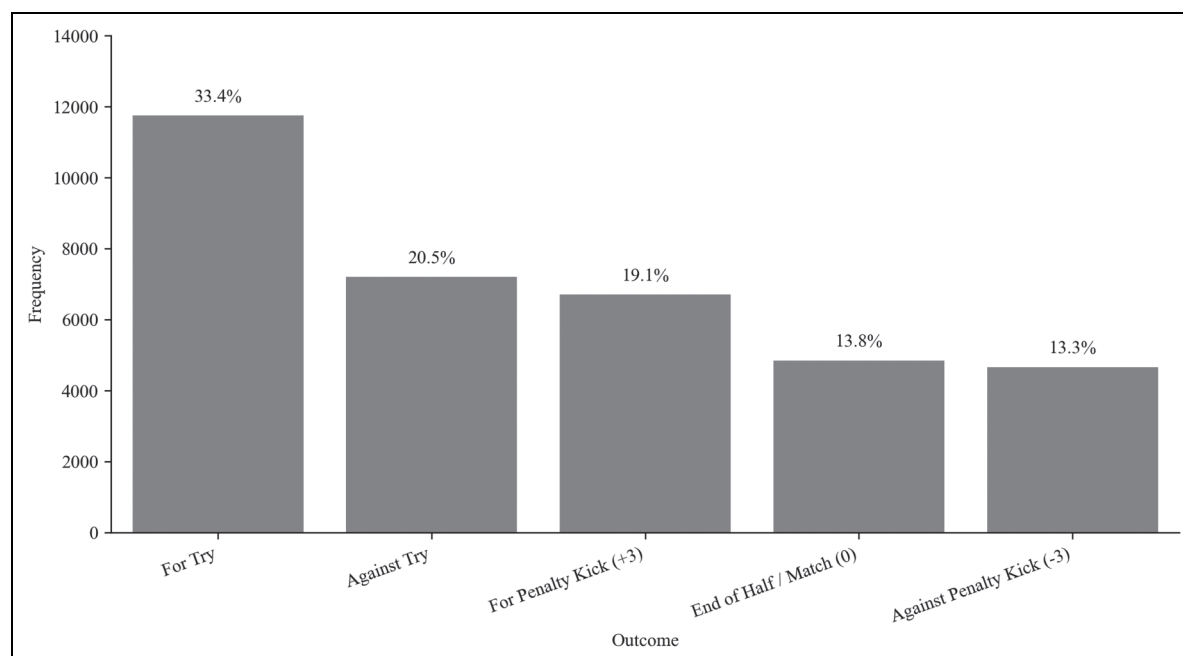
using these well-established models. The exploration of more complex architectures is identified as a key avenue for future work.

Models were trained on a sub-sample of the rugby union data to ensure their performance was only evaluated against a separate, unseen subset (Guyon, 1997). A stratified group 10-fold cross-validation method was used to prevent data leakage, ensuring phases from the same match were only present in a single subset. Models were iteratively trained and validated, and later evaluated for generalisability on a separate test subset (Davis and Goadrich, 2006).

Hyperparameter optimisation was performed using Optuna (Akiba et al., 2019) to identify optimal model configurations. Optuna is an optimisation framework that dynamically explores model configurations by pruning unpromising models and applying parallelisation to concurrently execute multiple trials (Akiba et al., 2019). The number of trials was adapted to each model's computational demands: multinomial logistic regression used 1,000 trials per iteration in the group 10-fold cross-validation method (10,000 total), random forest and KNN used 500 trials per fold (5,000 total), while SVM used 100 trials per fold (1,000 total).

The optimisation process explored a balance between model complexity, regularisation techniques, optimisation algorithms and decision boundary characteristics (Table 4). Multinomial logistic regression was tuned for regularisation strength, type and solver algorithm (Hosmer et al., 2013). Random forest hyperparameters included ensemble size, splitting criteria and class imbalance





**Figure 2.** Frequencies of different scoring outcomes after reducing the classes of the dependent variable.

strategies (Breiman, 2001). Support vector machine optimisation covered regularisation strength, kernel types, gamma values and decision function shapes (Andrew, 2000). K-nearest neighbors was optimised for number of neighbours, weighting schemes, search algorithms, and distance metrics (Cunningham and Delany, 2007).

Each model's best-performing configuration was evaluated using classification reports and confusion matrices to identify biases. The detailed model development, tuning and evaluation provide an exhaustive comparison of these four classification algorithms. This helps identify their predictive potential while highlighting the inherent challenges of modelling Expected Points in rugby union.

The modelling process was conducted in two stages. First, all four algorithms were trained and evaluated on the data's original, imbalanced class distribution to establish a performance baseline. This allowed for a transparent diagnosis of the challenges inherent to the dataset. Second, in direct response to the baseline models' poor performance on minority classes, a targeted experiment was conducted on the best-performing model (random forest). This experiment used the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance and assess its impact on predictive performance.

## Results

### Baseline model performance

The best-performing model was a random forest classifier composed of 770 decision trees. It achieved an accuracy of

39.7%  $\pm$  2.8 ppts and an F1-score of 29.3%  $\pm$  2.5 ppts (see Table 5). While this was a 17 percentage point improvement over the no-information rate (22.7%), it fell short of the 44.3% baseline established as a minimum threshold for practical application. The hyperparameter optimisation process for the random forest model resulted in the following configuration: a maximum depth of 45 levels, a maximum features ratio of 0.6, a minimum of 2 samples required for internal node splitting and a minimum of 13 samples mandated at each leaf node. Additionally, the model incorporated a minimum impurity decrease threshold of 0.0056 and a minimum weighted fraction of 0.001. The 10.4 percentage point difference between accuracy and F1-score indicates a bias towards certain classes in the random forest model.

The random forest model was closely followed by the SVM model, which demonstrated a 39.0%  $\pm$  2.7 ppts accuracy and a 33.0%  $\pm$  2.9 ppts F1-score. The optimal SVM model configuration used an RBF (Radial Basis Function) kernel with a C value (regularisation parameter) of 0.38, a gamma (kernel coefficient) of 0.1, a tolerance for stopping criterion of 0.0006, a one-versus-rest decision function shape and break ties enabled.

On the other hand, the multinomial logistic regression model (37.8%  $\pm$  2.7 ppts accuracy; 31.4%  $\pm$  2.6 ppts F1-score) and KNN (36.9%  $\pm$  2.1 ppts accuracy; 32.4%  $\pm$  2.1 ppts F1-score) showed marginally lower predictive power. The multinomial logistic regression configuration that achieved best results used a SAG solver with an L2 penalty, a C value of 0.006 and a tolerance of 0.000004; while the best KNN model configuration used a Ball Tree algorithm with 72 neighbours, a leaf size of 72, the

**Table 4.** Hyperparameters explored for different models.

Hyperparameter	Values tested
<i>Multinomial Logistic Regression</i>	
Regularisation (C)	0.00001 to 100,000
Regularisation type (penalty)	Lasso (l1), Ridge (l2) and Elastic Net
Solver	Newton-CG, LBFGS, SAG, SAGA
<i>Random Forest</i>	
Number of Decision Trees	Range between 50 and 1,000
Criterion	Gini, Entropy
Maximum Depth	2 to 100
Minimum Samples Split	Range between 2 and 15
Minimum Samples Leaf	Range between 1 and 15
Max Features	None, Sqrt, Log2, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Class Weight	None, Balanced, Balanced Subsample
Min Weight Fraction Leaf	Range between 0.0 and 0.5
Bootstrap	Yes, No
Minimum Impurity Decrease	Range between 0.0 and 0.1
<i>Support Vector Machine (SVM)</i>	
Regularisation (C)	0.1 to 1
Kernel	Linear, Polynomial, Radial Basis Function (RBF), Sigmoid
Gamma	Scale, Auto, 1, 0.1, 0.01, 0.001
Tolerance	0.0001 to 0.001
Decision Function Shape	One-Over-One (OVO), One-Over-Rest (OVR)
Break Ties	Yes, No
<i>K-Nearest Neighbors (KNN)</i>	
Number of neighbours (k)	1 to 400
Weights	Uniform, Distance
Algorithm	Ball Tree, KD Tree, Brute, Auto
Leaf size	1 to 300
Distance metric	Minkowski, Manhattan, Euclidean, Chebyshev
Minkowski metric power (p)	1, 2

Manhattan distance metric and uniform weights for all points in each neighbourhood. However, the small performance difference among the top configurations suggested that Optuna reached a performance plateau during optimisation of these two models.

### Classification report

A further evaluation using the classification report in Table 6 highlighted significant biases towards predicting the majority class. The random forest showed a considerable preference for predicting tries scored (83.1%  $\pm$  5.1 ppts recall; 36.9%  $\pm$  1.9 ppts precision) and end of half (65.7%  $\pm$  9.9 ppts recall; 54.0%  $\pm$  7.4 ppts precision), leading to a failure to predict any penalty kicks. This bias was less pronounced in the SVM and the multinomial logistic regression models, which showed a reduced recall for tries scored (73.7%  $\pm$  6.5 ppts SVM; 73.2%  $\pm$  4.5 ppts multinomial logistic regression) and the end of half (40.6%  $\pm$  9.3 ppts SVM; 42.2%  $\pm$  9.8 ppts

**Table 5.** Hyperparameter values of the top-performing configurations for each model.

Hyperparameter	Best Model
<i>Random Forest</i>	
Number of Decision Trees	770
Criterion	Gini
Maximum Depth	45
Minimum Samples Split	2
Minimum Samples Leaf	13
Max Features	0.6
Class Weight	None
Min Weight Fraction Leaf	0.001
Bootstrap	Yes
Minimum Impurity Decrease	0.0056
Accuracy	39.7% $\pm$ 2.8 ppts
F1-score	29.3% $\pm$ 2.5 ppts
<i>Support Vector Machine</i>	
Regularisation (C)	0.38
Kernel	RBF
Gamma	0.1
Tolerance	0.0006
Decision Function Shape	One-Versus-Rest (OVR)
Break Ties	Yes
Accuracy	39.0% $\pm$ 2.7 ppts
F1-score	33.0% $\pm$ 2.9 ppts
<i>Multinomial Logistic Regression</i>	
Regularisation (C)	0.006
Regularisation type (penalty)	L2
Solver	SAG
Tolerance	0.000004
Accuracy	37.8% $\pm$ 2.7 ppts
F1-score	31.4% $\pm$ 2.6 ppts
<i>K-Nearest Neighbors</i>	
Number of neighbours (k)	72
Weights	Uniform
Algorithm	Ball Tree
Leaf size	72
Distance metric	Manhattan
Minkowski metric power (p)	1
Accuracy	36.9% $\pm$ 2.1 ppts
F1-score	32.4% $\pm$ 2.1 ppts

multinomial logistic regression) compared to the random forest model. The reduced bias enhanced these models' precision across most classes compared to the random forest classifier. However, while this allowed the SVM and multinomial logistic regression models to outperform the random forest model's F1-score by 3.7 and 2.1 percentage points, respectively, the high recall but low precision for predicting tries scored continued to indicate an over-prediction of the majority class. The KNN model displayed similar patterns to those observed with the other models.

### Confusion matrix

The confusion matrix in Table 7 shows further details on the effectiveness across models in predicting different scoring

**Table 6.** Classification reports for different models.

Model	Outcome	Precision (%)	Recall (%)	F1 (%)
RF <sup>1</sup>	For Try	36.9	83.1	51.1
	For PK <sup>5</sup>	0.0	0.0	0.0
	End of Half	54.0	65.7	59.2
	Against PK	0.0	0.0	0.0
	Against Try	41.1	14.0	20.0
SVM <sup>2</sup>	For Try	39.6	73.7	51.5
	For PK	33.5	7.4	11.9
	End of Half	55.0	40.6	46.6
	Against PK	21.7	1.2	2.3
	Against Try	32.0	35.3	33.5
MLR <sup>3</sup>	For Try	38.7	73.2	50.7
	For PK	25.5	5.9	9.5
	End of Half	48.4	42.2	44.9
	Against PK	24.2	0.2	0.4
	Against Try	31.8	31.4	31.5
KNN <sup>4</sup>	For Try	39.1	73.0	50.9
	For PK	28.4	15.1	19.6
	End of Half	56.5	22.2	31.7
	Against PK	25.8	6.0	9.8
	Against Try	30.8	28.2	29.4

<sup>1</sup> RF: Random Forest<sup>2</sup> SVM: Support Vector Machine<sup>3</sup> MLR: Multinomial Logistic Regression<sup>4</sup> KNN: K-Nearest Neighbors<sup>5</sup> PK: Penalty Kick

outcomes. It indicates that models over-predicted tries scored by between +125.2% (random forest) and +85.9% (SVM) relative to their true occurrence ( $n=11,743$ ), accounting for between 75.2% (random forest) and 62.1% (SVM) of all predictions despite only representing 33.4% of all phases in the dataset. On the other hand, tries conceded were significantly under-predicted by the random forest model, with -60.8% fewer predictions ( $n=2,827$ ) than true instances ( $n=7,205$ ). Most of these missed true instances of tries conceded (90.4%) were predicted by this model as tries scored, further highlighting a bias towards the latter outcome. While the other three models showed a more balanced distribution of predictions for tries conceded, ranging from -7.9% (KNN) to +10.0% (SVM) compared to true instances, they produced a considerable number of false positives (69.4% KNN; 68.3% MLR; 68.0% SVM).

Penalty kicks were significantly impacted by the models' bias towards predicting tries scored. The random forest model made no penalty kick predictions, despite these two scoring outcomes combined representing 32.4% of the phases in the dataset (19.1% scored; 13.3% conceded). Other models also significantly struggled with penalty kick predictions, under-predicting penalty kicks scored by a margin ranging from -76.8% (SVM) to -46.4% (KNN) and penalty kicks conceded by between -98.8% (MLR) and -76.4% (KNN) compared to their true instances. As with tries conceded, the few penalty kick predictions

made by the models also produced a high number of false positives for both scored (74.9% MLR; 71.8% KNN; 68.2% SVM) and conceded (84.5% MLR; 79.3% SVM; 74.6% KNN).

Predictions for the end of the half showed mixed results across models. The random forest model over-predicted phases leading to the end of the half by 21.6% compared to its true instances ( $n=4,848$ ). The majority of these incorrect predictions corresponded to true instances of tries scored (47.2%) and conceded (21.9%). In contrast, all other models under-predicted the end of the half by amounts ranging from -60.8% (KNN) to -12.9% (MLR). The bias towards predicting tries scored was a major contributor to under-predicting the end of the half, with the models predicting tries scored for between 43.0% (SVM) and 55.3% (KNN) of true instances leading to the end of the half.

### Addressing class imbalance with SMOTE

In response to the baseline random forest model's inability to predict penalty kicks, a targeted experiment was conducted using SMOTE. While this led to a 4 percentage point drop in overall accuracy to 35.7%, it achieved a substantial 5.1 percentage point increase in the weighted F1-score to 34.4%, a more appropriate metric for this imbalanced classification problem. As shown in Table 8, the application of SMOTE increased the model's ability to predict penalty kicks from a 0.0% F1-score to a more effective 26.1% (scored) and 21.9% (conceded) F1-score. This demonstrates a successful rebalance, where a moderate decrease in performance on majority classes resulted in a considerable gain in predictive power on rare but meaningful minority classes, making the model more practically relevant.

### Feature importance

Feature importance metrics were extracted from the multinomial logistic regression and the random forest models, given that the support vector machine and k-nearest neighbors do not inherently produce feature importance measures, as these methods rely on distance or similarity calculations rather than explicit parameters or splitting criteria tied to individual features. The analysis of standardised coefficient magnitudes showed that pitch location was the most influential feature for the multinomial logistic regression model, accounting for 44.9% of the model's importance. Disciplinary actions, such as yellow cards (14.7%) and red cards (13.8%), and time remaining (9.1%) also played important roles in the model's predictions. On the other hand, phase sequence number (1.9%), points



**Table 7.** Confusion matrix for the best models of each classification algorithm.

Model	True Value	Predicted Value					Total
		For Try	For PK	EoH	Ag. PK	Ag. Try	
RF	For Try	9,765	0	1,280	0	698	11,743
	For PK	5,672	0	500	0	534	6,706
	EoH	1,592	0	3,181	0	75	4,848
	Ag. PK	3,805	0	338	0	519	4,662
	Ag. Try	5,609	0	595	0	1,001	7,205
	<b>Total</b>	26,443	0	5,894	0	2,827	35,164
SVM	For Try	8,658	487	772	52	1,774	11,743
	For PK	4,677	494	249	46	1,240	6,706
	EoH	2,084	65	1,967	12	720	4,848
	Ag. PK	2,531	193	224	55	1,659	4,662
	Ag. Try	3,883	315	370	101	2,536	7,205
	<b>Total</b>	21,833	1,554	3,582	266	7,929	35,164
MLR	For Try	8,593	600	843	18	1,689	11,743
	For PK	4,862	395	322	18	1,109	6,706
	EoH	2,242	17	2,042	0	547	4,848
	Ag. PK	2,558	220	343	9	1,532	4,662
	Ag. Try	3,918	343	671	13	2,260	7,205
	<b>Total</b>	22,173	1,575	4,221	58	7,137	35,164
KNN	For Try	8,574	1,075	322	230	1,542	11,743
	For PK	4,368	1,013	132	182	1,011	6,706
	EoH	2,680	271	1,080	72	745	4,848
	Ag. PK	2,419	525	133	280	1,305	4,662
	Ag. Try	3,894	710	232	337	2,032	7,205
	<b>Total</b>	21,935	3,594	1,899	1,101	6,635	35,164*

Abbreviations: MLR = Multinomial Logistic Regression, RF = Random Forest, SVM = Support Vector Machine, KNN = K-Nearest Neighbors, PK = Penalty Kick, EoH = End of Half, Ag. = Against.

\*N=35,164 after the exclusion of drop goals.

**Table 8.** Comparison of random forest performance with and without SMOTE.

Scoring Outcome	Metric	Without SMOTE	With SMOTE	Change
Overall	Accuracy	39.7%	35.7%	−4.0 ppts
Overall	F1-score	29.3%	34.4%	+5.1 ppts
<i>Per-Class F1-score</i>				
For Try	F1-score	51.1%	42.4%	−8.7 ppts
End of Half	F1-score	59.2%	53.8%	−5.4 ppts
Against Try	F1-score	20.0%	24.3%	+4.3 ppts
For PK	F1-score	0.0%	26.1%	+26.1 ppts
Against PK	F1-score	0.0%	21.9%	+21.9 ppts

difference (1.4%) and pitch side (0.7%) all showed minimal contribution to its predictions.

The random forest model presented a different feature importance distribution. The model heavily relied on the seconds remaining variable (72.7%), defined as 80 minutes minus the seconds elapsed from the beginning of the match. Given the model's low overall accuracy, this heavy reliance indicated a possible under-representation of more complex features in the dataset, forcing the

model to concentrate on the most easily discernible pattern. As a result, the model over-emphasised time-sensitive scoring outcomes, such as the end of a half. Points difference was the second most influential feature (12.6%), while features such as type of play (0.6%), phase sequence number (0.3%), cards (0.1%) and pitch side (0.1%) had minimal impact on the model's decisions.

## Discussion

This study presents a comprehensive methodological approach to developing an Expected Points framework for rugby union. While the primary objective was to build a predictive metric representative of possession quality, the results did not achieve the necessary performance for practical application. However, the study's main contribution lies in the invaluable foundational groundwork. It presents a transparent methodological benchmark, a rich dataset and a clear identification of the core challenges that must be overcome to transition the analysis of performance in rugby union from descriptive performance indicators to predictive modelling.

Following extensive notational analysis of 35,199 phases of play, four classification algorithms were evaluated. Despite hyperparameter optimisation, the top-

performing model, a random forest classifier, achieved an accuracy of only 39.7% and an F1-score of 29.3%. All baseline models displayed considerable bias towards predicting tries at the expense of penalty kicks. While the SVM model produced a more balanced F1-score (33.0%), the overall performance was limited by factors like feature representation and class imbalance. However, a subsequent experiment using SMOTE demonstrated that this imbalance could be partially mitigated, improving the F1-score by 5.1 percentage points.

Comparing these results against current Expected Points models was challenging due to the absence of model evaluation results (e.g., accuracy, F1-scores or confusion matrices) in existing literature (O'Shaughnessy, 2006; Romer, 2006; Yurko et al., 2019). Consequently, alternative sports performance analysis literature was used to set an accuracy baseline. Studies across various sports have developed three-class models to predict match outcomes (wins, losses or draws) with accuracy scores between 52.4% and 67.5% (Hubáček et al., 2019; McCabe and Trevathan, 2008). The average performance of these models (54.9%) represented a 21.6 percentage point improvement from random guessing (33.3%). In this study, a 21.6 percentage point improvement from the no-information rate (22.7%), i.e., the accuracy that a model must exceed to outperform random guessing after accounting for class imbalance (James et al., 2013), resulted in a minimum accuracy baseline of 44.3%. Unfortunately, neither the random forest (39.7%) nor the support vector machine (39.0%) models were able to reach that baseline. This result confirms that while the methodological framework is sound, the resulting model is not yet reliable enough for direct application in professional coaching or tactical decision-making.

Despite insufficient predictive power, the Expected Points framework presented in this study carries important theoretical implications. The study recommends transitioning from fragmented individual indicators to a holistic approach of measuring performance through contemporary modelling techniques (Colomer et al., 2020). It identifies opportunities in rugby union research by drawing parallels from other invasion team sports, such as American football (Carter and Machol, 1971; Yurko, 2017), Australian rules football (O'Shaughnessy, 2006) or rugby league (Kempton et al., 2016).

Research in these sports has embraced predictive modelling and Expected Points, concepts that have received little attention in rugby union performance analysis literature (Hughes et al., 2012; James et al., 2005; Ortega et al., 2009). An opportunity exists to revolutionise the understanding of performance in rugby union by applying established methodologies from other sports. However, Expected Points literature often lacks sufficient information on model performance evaluation (Burke, 2008; Romer, 2006; Yurko, 2017). The absence of essential model evaluation metrics fails to demonstrate the generalisability and

reliability of current Expected Points models in informing tactical decisions and limits their comparability to new models. This study addresses this gap by providing a detailed account of modelling steps and performance evaluation results, establishing a clear benchmark for future Expected Points models in rugby union.

The primary limitation identified was the severe class imbalance inherent to rugby union data, likely originating from the point-maximising nature of on-pitch actions, such as opting for tries over penalty kicks (Romer, 2006). This uneven class distribution caused the baseline models to over-predict majority outcomes at the expense of minority ones. However, the successful application of SMOTE indicated that this problem can be addressed with appropriate resampling techniques. Additionally, feature expansion may also help mitigate model bias by improving feature representation of key factors influencing penalty kicks, such as referee decisions to award penalties (Mascarenhas et al., 2005), player discipline (Mitchell and Tierney, 2021) or weather conditions (Crewther et al., 2020).

A second limitation observed was the complexity in modelling open-play sports like rugby union. The sport's characteristics, involving tactical execution, technical ability, physicality and continuous play, create a dynamic environment where complex relationships between variables influence each phase's outcome, such as player skills (Ziv and Lidor, 2016), tactics (Roberts et al., 2008), weather (Kearney and Riddiford-Harland, 2012), morale (Cotterill and Fransen, 2016) and fatigue (Duthie et al., 2003). These elements are typically absent from notational analysis datasets due to their limited observability (Hughes and Franks, 2004). Consequently, models trained on core contextual variables struggle to learn these underlying interactions. Future research could explore the addition of contextual data related to player fitness, fatigue, team morale or psychological resilience to capture player endurance and its influence on decision-making and scoring. Team-level data on formations, player roles and strategic plays could also enrich the modelling process.

Furthermore, the selection of four classification algorithms explored represents only an initial evaluation of the predictive potential in rugby union phase-level data aimed at laying the foundation for Expected Points modelling in rugby union. Future research could expand upon this work with a more comprehensive exploration of advanced modelling techniques that include methods such as gradient boosting machines (Ke et al., 2017) or recurrent neural networks, such as long short-term memory networks (Hochreiter and Schmidhuber, 1997). Inspired by the work of Yurko et al. (2019) in the NFL, this study framed the task as a classification problem. However, future approaches could instead investigate techniques that assume different data structures and emphasise the temporal and sequential characteristics of rugby play. These approaches may improve upon the predictive performance achieved by the classification models in this study.

The study lays the foundation for the development of Expected Points in rugby union. Building on the application of SMOTE demonstrated in this study, future research should explore more advanced resampling techniques, such as Borderline-SMOTE or ADASYN, which may offer similar benefits with less artificial noise (Han et al., 2005; He et al., 2008). Combining these advanced samplers with cost-sensitive learning could also be used to apply heavier penalties for misclassifying less frequent scoring outcomes (Ling and Sheng, 2008). Additionally, data expansion could also improve the feature representation of the complex dynamics in rugby union. New variables may include the relative strength between teams, such as recent form, head-to-head records or win-loss ratios (McCabe and Trevathan, 2008); player-specific metrics, such as fatigue levels and individual player skills; environmental conditions, such as weather or pitch conditions; and psychological factors, such as team morale. One avenue for future research is the collection of spatial and temporal features from wearable sensors or GPS units. Another is the derivation of momentum-related features from existing data (Blum and Langley, 1997), such as calculating the cumulative number of phases in a possession, the number of seconds since the last scoring event or the percentage possession using a 20-minute rolling window. This greater emphasis on spatiotemporal factors could effectively account for changes in playing dynamics and scoring patterns over the course of a match.

## Conclusion

The development of Expected Points in rugby union requires a paradigm shift towards more rigorous and reliable approaches for quantifying the expected situational value of different match scenarios. The study presents a theoretical performance analysis framework designed to better capture complex feature relationships than association-based statistics. However, extensive notational analysis, feature engineering and hyperparameter optimisation resulted in a baseline model with insufficient accuracy for practical application, primarily due to the limiting factors of class imbalance and feature representation. This study demonstrated that these limitations can be partially mitigated with methods such as the application of resampling techniques. The use of SMOTE sacrificed a small amount of overall accuracy for a meaningful improvement in the model's F1-score and its ability to predict a wider range of scoring events, suggesting that a more balanced model is achievable.

The complexity observed in measuring and interpreting all variables affecting rugby union scoring challenges the positivist paradigm that has predominated sports performance analysis literature (Mackenzie and Cushion, 2013). Research within this paradigm has often aimed to establish causal relationships between isolated performance


indicators (Jones et al., 2004; Ortega et al., 2009). This epistemological approach has reduced the complexity of sports performance by presenting it in an overly descriptive, systematic and unproblematic way that ignores the confounding variables and contexts directly and indirectly influencing success (Cushion, 2007).

The development of an Expected Points metric for rugby union represents an area of opportunity with important implications for tactical decision-making, player development and fan engagement. A standardised measure of performance that accounts for the situational context and potential impact of each action on the pitch could revolutionise the way the sport is understood, analysed and played.


This study presents a reproducible methodology aimed at inspiring future research to build upon its methods and advance towards a more data-driven and engaging future for rugby union. Exploring the sport's spatiotemporal characteristics by incorporating positional and momentum-related metrics could substantially enhance the predictive capabilities of datasets through improved feature representation. The integration of enriched datasets with time-series models or boosting techniques could result in predictive performance exceeding that observed in this study. Such enhanced predictive power could lead to the development of reliable Expected Points models that provide novel insights into the sport's key determinants of success.

## ORCID iD

Dr Guillermo Martinez-Arastey  <https://orcid.org/0000-0002-2034-3823>

Naomi Datson  <https://orcid.org/0000-0002-5507-9540>

Neal Smith  <https://orcid.org/0000-0003-2286-7572>

Matthew Robins  <https://orcid.org/0000-0003-2370-9429>

## Ethical approval

Ethical approval was not required for this study as it involved the analysis of publicly available data.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

## Data availability

Data supporting the findings of this study are openly available at <http://doi.org/10.5281/zenodo.13851563>. 1Department.

## References

Akiba T, Sano S, Yanase T, et al. (2019) Optuna: A next-generation hyperparameter optimization framework. In:

- Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp.2623–2631.
- Andrew A (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods by Nello Cristianini and John Shawe-Taylor. *Robotica* 18: 687–689.
- Bishop L and Barnes A (2013) Performance indicators that discriminate winning and losing in the knockout stages of the 2011 rugby world cup. *International Journal of Performance Analysis in Sport* 13(1): 149–159.
- Blum AL and Langley P (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97(1-2): 245–271.
- Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32.
- Bremner S, Robinson G and Williams MD (2013) A retrospective evaluation of team performance indicators in rugby union. *International Journal of Performance Analysis in Sport* 13(2): 461–473.
- Burke B (2008) Expected points. <http://archive.advancedfootballanalytics.com/2008/08/expected-points.html>.
- Burke B (2010) Expected points (ep) and expected points added (epa) explained. <http://archive.advancedfootballanalytics.com/2010/01/expected-points-ep-and-expected-points.html>.
- Carter V and Machol RE (1971) Operations research on football. *Operations Research* 19(2): 541–544.
- Causey T (2015) Expected points part 1: Building a model and estimating uncertainty. <http://thespread.us/expected-points.html>.
- Cervone D, D'Amour A, Bornn L, et al. (2016) A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association* 111(514): 585–599.
- Colomer CM, Pyne DB, Mooney M, et al. (2020) Performance analysis in rugby union: A critical systematic review. *Sports Medicine-Open* 6(1): 4.
- Cotterill ST and Fransen K (2016) Athlete leadership in sport teams: Current understanding and future directions. *International Review of Sport and Exercise Psychology* 9(1): 116–133.
- Cover T and Hart P (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1): 21–27.
- Crewther BT, Potts N, Kilduff LP, et al. (2020) Performance indicators during international rugby union matches are influenced by a combination of physiological and contextual variables. *Journal of Science and Medicine in Sport* 23(4): 396–402.
- Cunningham P and Delany SJ (2007) k-nearest neighbour classifiers. *Multiple Classifier Systems* 34(8): 1–17.
- Cushion C (2007) Modelling the complexity of the coaching process. *International Journal of Sports Science & Coaching* 2(4): 395–401.
- Davis J and Goadrich M (2006) The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd international conference on Machine learning*, pp.233–240.
- Duthie G, Pyne D and Hooper S (2003) Applied physiology and game analysis of rugby union. *Sports Medicine* 33(13): 973–991.
- Eaves JS, Hughes DM and Lamb LK (2005) The consequences of the introduction of professional playing status on game action variables in international northern hemisphere rugby union football. *International Journal of Performance Analysis in Sport* 5(2): 58–86.
- ESPN (2022) Glossary of rugby union terms. <http://en.espn.co.uk/statsguru/rugby/page/97263.html>.
- Feurer M and Hutter F (2019) Hyperparameter optimization. *Automated Machine Learning: Methods, Systems, Challenges* : –.
- Goldner K (2017) Situational success: Evaluating decision-making in football. In: *Handbook of statistical methods and analyses in sports*. Chapman and Hall/CRC, 199–214.
- Green S (2012) Assessing the performance of premier league goalscorers. <https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/>.
- Guyon I (1997) *A scaling law for the validation-set training-set size ratio (Technical Report)*. AT&T Bell Laboratories. 1–11.
- Han H, Wang WY and Mao BH (2005) Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: *International conference on intelligent computing*, pp.878–887. Springer.
- He H, Bai Y, Garcia EA, et al. (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, Ieee, pp.1322–1328.
- He H and Garcia EA (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263–1284.
- Hendricks S, Roode B, Matthews B, et al. (2013) Defensive strategies in rugby union. *Perceptual and Motor Skills* 117(1): 65–87.
- Hendricks S, Till K, Den Hollander S, et al. (2020) Consensus on a video analysis framework of descriptors and definitions by the rugby union video analysis consensus group. *British Journal of Sports Medicine* 54(10): 566–572.
- Ho TK (1995) Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1, pp.278–282. IEEE.
- Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8): 1735–1780.
- Hosmer DW, Lemeshow S and Sturdivant RX (2013) *Applied Logistic Regression*, 398. Hoboken, New Jersey: John Wiley & Sons.
- Hubáček O, Šourek G and Železný F (2019) Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning* 108: 29–47.
- Hughes M and Franks IM (2004) *Notational Analysis of Sport: Systems for Better Coaching and Performance in Sport*. London: Routledge Taylor & Francis Group.



- Hughes MT, Hughes MD, Williams J, et al. (2012) Performance indicators in rugby union. *Journal of Human Sport and Exercise* 7(2): 383–401.
- Hunter P and O'Donoghue P (2001) A match analysis of the 1999 rugby union world cup. In: *Books of abstracts Fifth World congress of performance analysis in sports*, pp.85–90.
- James G, Witten D, Hastie T, et al. (2013) *An Introduction to Statistical Learning with Applications in R*. Vol. 112, 1st Edition New York: Springer.
- James N, Mellalieu S and Jones N (2005) The development of position-specific performance indicators in professional rugby union. *Journal of Sports Sciences* 23(1): 63–72.
- Japkowicz N and Stephen S (2002) The class imbalance problem: A systematic study. In: *Intelligent data analysis*, Vol. 6, pp.429–449. Elsevier.
- Jones NM, Mellalieu SD and James N (2004) Team performance indicators as a function of winning and losing in rugby union. *International Journal of Performance Analysis in Sport* 4(1): 61–71.
- Katz S and Burke B (2016) How is total qbr calculated? we explain our quarterback rating. <https://www.advancedfootballanalytics.com/index.php/home/stats/stats-explained/expected-points-and-epa-explained>.
- Ke G, Meng Q, Finley T, et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in neural information processing systems (NIPS)*, pp.3146–3154.
- Kearney PE and Riddiford-Harland DL (2012) Changing weather conditions and the effect on rugby league match play. *Journal of Sports Science & Medicine* 11(2): 327.
- Kempton T, Kennedy N and Coutts AJ (2016) The expected value of possession in professional rugby league match-play. *Journal of Sports Sciences* 34(7): 645–650.
- Kleinbaum DG and Klein N, (2008) *Logistic Regression: A Self-Learning Text*. 2nd Edition. New York: Springer.
- Kluyver T, Ragan-Kelley B, Pérez F, et al. (2016) Jupyter notebooks—a publishing format for reproducible computational workflows. In: *Positioning and power in academic publishing: Players, agents and agendas*, pp.87–90. IOS press.
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*, Vol. 14, pp.1137–1145. Montreal, Canada.
- Kotsiantis S, Kanellopoulos D and Pintelas P (2006) Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30(1): 25–36.
- Ling CX and Sheng VS (2008) Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning* 2011: 231–235.
- Mackenzie R and Cushion C (2013) Performance analysis in football: A critical review and implications for future research. *Journal of Sports Sciences* 31(6): 639–676.
- Mascarenhas DR, Collins D, Mortimer PW, et al. (2005) Training accurate and coherent decision making in rugby union referees. *The Sport Psychologist* 19(2): 131–147.
- McCabe A and Trevathan J (2008) Artificial intelligence in sports prediction. In: *Fifth International conference on information technology: New generations (itng 2008)*, pp.1194–1197. IEEE.
- Mitchell S and Tierney GJ (2021) Sanctioning of breakdown infringements during the knockout stage of the 2019 rugby world cup. *International Journal of Sports Science & Coaching* 16(2): 407–414.
- Nakagawa A (2006) Re-examination of importance of kick-off and 50m restart kick play in rugby football games. *International Journal of Sport and Health Science* 4: 273–285.
- O'Donoghue P (2007) Reliability issues in performance analysis. *International Journal of Performance Analysis in Sport* 7(1): 35–48.
- Ortega E, Villarejo D and Palao JM (2009) Differences in game statistics between winning and losing rugby teams in the six nations tournament. *Journal of Sports Science & Medicine* 8(4): 523.
- O'Shaughnessy DM (2006) Possession versus position: Strategic evaluation in afl. *Journal of Sports Science & Medicine* 5(4): 533–540.
- Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12: 2825–2830.
- Powers DM (2020) Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- Roberts SP, Trewartha G, Higgitt RJ, et al. (2008) The physical demands of elite english rugby union. *Journal of Sports Sciences* 26(8): 825–833.
- Romer D (2006) Do firms maximize? evidence from professional football. *Journal of Political Economy* 114(2): 340–365.
- Shmueli G (2010) To explain or to predict? *Statistical Science* 25(3): 289–310.
- Thomas AC (2006) The impact of puck possession and location on ice hockey strategy. *Journal of Quantitative Analysis in Sports* 2(1): 1–19.
- Ungureanu AN, Brustio PR, Mattina L, et al. (2019) “how” is more important than “how much” for game possession in elite northern hemisphere rugby union. *Biology of Sport* 36(3): 265–272.
- Vahed Y, Kraak W and Venter R (2016) Changes on the match profile of the south african currie cup tournament during 2007 and 2013. *International Journal of Sports Science & Coaching* 11(1): 85–97.
- Watson N, Durbach I, Hendricks S, et al. (2017) On the validity of team performance indicators in rugby union. *International Journal of Performance Analysis in Sport* 17(4): 609–621.
- Williams J (2012) Operational definitions in performance analysis and the need for consensus. *International Journal of Performance Analysis in Sport* 12(1): 52–63.
- Wu G and Chang EY (2005) Kba: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering* 17(6): 786–795.
- Yurko R (2017) Nfl expected points with n—scrapr: Part 1 - an introduction to expected points. <https://www.cmusports>



- analytics.com/nfl-expected-points-nflscrapr-part-1-introduction-expected-points/.
- Yurko R, Ventura S and Horowitz M (2019) nflwar: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports* 15(3): 163–183.
- Zheng A and Casari A (2018) *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, CA: O'Reilly Media, Inc..
- Ziv G and Lidor R (2016) On-field Performances of Rugby Union Players: A Review. *Journal of Strength and Conditioning Research* 30(3): 881–892.