



When Timing Matters: Evaluating Temporal Leakage in Machine Learning Models of Football Pass Turnovers

Andrew Peters, Nimai Parmar, Michael Davies & Nic James

To cite this article: Andrew Peters, Nimai Parmar, Michael Davies & Nic James (03 Jun 2026): When Timing Matters: Evaluating Temporal Leakage in Machine Learning Models of Football Pass Turnovers, Research Quarterly for Exercise and Sport, DOI: [10.1080/02701367.2026.2677718](https://doi.org/10.1080/02701367.2026.2677718)

To link to this article: <https://doi.org/10.1080/02701367.2026.2677718>



© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC



[View supplementary material](#)



Published online: 03 Jun 2026.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

When Timing Matters: Evaluating Temporal Leakage in Machine Learning Models of Football Pass Turnovers

Andrew Peters^{a,b}, Nimai Parmar^a, Michael Davies^{a,b}, and Nic James^a

^aMiddlesex University; ^bLeicester City Football Club

ABSTRACT

The Expected Pass Turnovers (xPT) model advances turnover probability quantification in professional football, but the inclusion of post-pass descriptive features such as ball speed and distance moved introduces temporal leakage and limits real-time tactical utility. This study compares the original xPT framework with leakage-corrected alternatives across four modeling approaches: mixed-effects logistic regression, penalized logistic regression, random forest, and gradient boosting. Using 256,433 passes from the 2020–21 English Premier League, we evaluated the performance of leakage-inclusive and leakage-corrected feature sets using grouped cross-validation and cross-validated ROC-AUC, accuracy, sensitivity, specificity, F-measure, and Brier score. Removing post-execution features reduced ROC-AUC by 0.082–0.183 (mean = 0.136), with tree-based methods experiencing larger performance losses than logistic approaches. The best-performing alternative model, gradient boosting (ROC-AUC = 0.742), approached the default mixed-effects logistic model (ROC-AUC = 0.789), indicating that substantial predictive signal remains after leakage correction. Alternative models retained acceptable sensitivity and calibration, supporting their use in prospective tactical deployment. SHAP analysis showed that default models relied heavily on four post-pass descriptive variables, whereas leakage-corrected models shifted toward pressing intensity and tactical context variables that are available before pass execution. The findings suggest that default xPT-style models are well suited to retrospective analysis, while leakage-corrected models are more appropriate for real-time tactical decision-making.

ARTICLE HISTORY

Received 20 January 2026
Accepted 18 May 2026

KEYWORDS

Data leakage; Expected Pass Turnovers; football analytics; machine learning; pressing strategy; temporal validity

Expected Pass Turnovers (xPT), introduced by Peters et al. (2024), provide a quantitative framework for estimating the likelihood that a pass will result in loss of possession in professional football. By combining event-level information with player and match context, the xPT framework offers a way to identify passes and players associated with elevated turnover risk, with relevance for both retrospective analysis and tactical planning (Peters et al., 2024). In applied settings, such models can support analysts and coaches in understanding which passes are most vulnerable to defensive pressure and in identifying the contextual conditions under which risky pass choices (high probability of turnover) are more likely to occur.

A growing body of work in football analytics has used statistical and machine learning approaches (Table 1) to model passing behavior, event likelihood, and tactical performance (Carling et al., 2005; Cintia et al., 2015; Deshpande Sameer & Evans Katherine, 2020; Peters et al., 2024; Stival et al., 2023; Wong et al., 2025). Mixed-effects logistic regression is especially useful in this context because football data is naturally clustered: passes are nested within players, matches, and teams, and these dependencies need to be considered when estimating effects (Deshpande Sameer & Katherine, 2020; Peters et al., 2024). Penalised logistic regression extends this framework by applying regularization to reduce overfitting when predictors are correlated or numerous (Constantinou & Fenton, 2012; Beato et al., 2024), while tree-based ensemble methods such as random

forest and gradient boosting can capture non-linear relationships and higher-order interactions that may be difficult to specify a priori (Stival et al., 2023; Narayanan & Pifer, 2024). Table 1 summarizes the rationale for these modeling approaches and their prior use in football analytics.

At the same time, football analytics faces an important methodological challenge (Dong, 2022): the timing of model inputs must match the decision context in which predictions are intended to be used. Data leakage occurs when a model includes information that would not be available at the moment a decision is made, thereby inflating apparent performance and limiting practical deployment (Friedman, 2001; Kaufman et al., 2012). In the present context, pass-descriptive variables such as pass distance, pass angle, ball movement speed, and the proportion of distance the ball traveled toward the opposition goal are only fully known after the pass has been executed (Kaufman et al., 2012; Kuhnt & Schiffner, 2021). These variables remain appropriate for retrospective evaluation of completed actions, but they are not valid inputs for prospective models intended to infer live tactical decision-making.

This distinction matters because real-time pressing decisions depend on pre-execution information: the position of the ball carrier, the positioning of teammates and opponents, local defensive pressure, and the range of passing options available at the moment of the decision. A model that relies on post-execution descriptors may therefore perform well in

CONTACT Andrew Peters  andrewpeters1994@gmail.com  Faculty of Science & Technology, Middlesex University, NW4 4BT, London, UK.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/02701367.2026.2677718>.

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Table 1. Overview of modelling approaches applied in football analytics.

Algorithm	Description	Why Used	Football Analytics Examples
Mixed-effects logistic regression	Statistical model handling clustered or repeated-measures data with interpretable parameters.	Captures player- and match-level heterogeneity and allows inference on contextual predictors.	Applied to model pass completion and event likelihoods (Deshpande Sameer & Katherine, 2020; Peters et al., 2024).
Penalised logistic regression	Regression model incorporating regularisation (e.g., LASSO, elastic-net) to prevent overfitting.	Improves generalisation when datasets contain many correlated predictors or hierarchical effects.	Used for injury risk, shot success and broader predictive modelling in football (Constantinou & Fenton, 2012; Beato et al., 2024).
Random forest	Ensemble of decision trees trained on bootstrapped samples with random feature subsets.	Captures complex non-linear relationships and high-order interactions automatically.	Applied for passing difficulty, event classification and possession-sequence modelling (Stival et al., 2023).
XGBoost / Gradient boosting	Sequential tree-based algorithm minimising classification error via gradient descent.	Delivers state-of-the-art predictive accuracy and scalability for large spatiotemporal datasets.	Used for expected pass or possession value estimation and event prediction (Narayanan & Pifer, 2024).

prediction tests while still failing to represent the information actually available to players and coaches in the moment (Chassy, 2025; Herold et al., 2022). For that reason, evaluating leakage-free alternatives is not simply a technical exercise; it is necessary for determining whether xPT-type models are suitable for deployment in settings where tactical decisions must be made in advance of pass completion. Beyond evaluating completed passes, the leakage-corrected model can also be used to examine decision timing. For example, it can help assess whether a player might have reduced turnover risk by releasing the ball earlier, or whether carrying the ball momentarily helped move from a high-risk to a lower-risk situation. In this sense, the model provides a way to study pass release timing, pressure evasion, and decision-making under defensive constraint, rather than only the outcome of completed passes.

The present study addresses this issue by comparing leakage-inclusive and leakage-corrected versions of the xPT framework across four modeling approaches: mixed-effects logistic regression, penalized logistic regression, random forest, and gradient boosting (Lindberg & Soderbarg, 2020; Pappalardo et al., 2019; Peters et al., 2024; Rico-González et al., 2023). This design allows us to assess how much predictive performance is attributable to temporally misaligned features and how much remains when models are restricted to pre-execution contextual information only. It also provides a direct test of whether more flexible machine learning methods can retain useful predictive power once post-pass information is removed.

Leakage-corrected models are more appropriate for prospective tactical deployment because they estimate turnover risk using only pre-execution context. This makes them suitable not only for real-time pressing decisions, but also for evaluating pass timing, identifying players or situations that warrant increased defensive attention, and assessing how tactical context shapes the decision to release the ball.

Accordingly, this study has five aims. First, we assess the magnitude of performance change associated with removing temporally misaligned features from the original xPT framework. Second, we benchmark mixed-effects logistic regression against three contemporary machine learning algorithms under both leakage-inclusive and leakage-corrected feature sets. Third, we determine whether leakage-free models retain sufficient discrimination and calibration to be useful for real-time tactical application. Fourth, we provide practitioners with a clearer basis for choosing between retrospective and prospective modeling frameworks depending on the intended use

case. Fifth, we document the full analysis in a reproducible R workflow to support transparency and future extension.

Methods

Data source and overview

This study leveraged event and positional data from the 2020/21 English Premier League season, obtained via StatsBomb 360 (StatsBomb, 2024). The positional dataset consisted of freeze-frame observations with multiple rows per event, allowing player locations and local team structure to be captured at the moment of pass execution. Player *x* and *y* coordinates were standardized to pitch dimensions and expressed in meters to ensure consistent distance calculations across matches and venues. Following Peters et al. (2024), two sets of positional features were constructed to represent passing context and surrounding defensive structure.

Set 1 of positional features - the availability of unmarked teammates was quantified in four angular sectors relative to the ball carrier: left, right, front, and behind. Sectors were defined using standard angular partitions around the ball carrier, and teammate marking status was determined by the minimum distance to the nearest defender. A threshold of 2 m was used to classify a teammate as unmarked, consistent with the operational definition used in Peters et al. (2024). For each sector, binary indicators were created to reflect whether at least one unmarked teammate was present.

Set 2 of positional features - the local defensive pressure was quantified by counting opposition players within three concentric radii around the ball carrier. The base pressure radius was set to 4.5 m when the ball carrier was more than 24 m from goal. Inside 24 m, the radius increased linearly according to distance to goal, using the equation $0.15 \times \text{distance to goal} + 0.85$, to reflect the greater sensitivity of pressing pressure in more dangerous areas (Equation 1). Counts of opposition players were then computed for half, equal to, and twice this base radius, providing measures of pressure density in immediate, intermediate, and extended zones.

$$d = DFG = \text{distance from goal (in meters)}$$

$$PR(d) = \text{pressure radius (in meters)}$$

Table 2. Definitions of model features used in the turnover prediction analysis.

Feature	Definition	Type	Unit
Ball carrier x-coordinate	x-coordinate of the pass origin.	Event	Meters (m)
Ball carrier y-coordinate	y-coordinate of the pass origin.	Event	Meters (m)
Ball movement speed	Distance travelled by the ball divided by time to reach the receiver.	Event	Meters per second (m/s)
Closest defender	Distance from the ball carrier to the nearest opposition player.	Event	Meters (m)
Distance ball moved	Linear distance travelled by the ball during the pass.	Event	Meters (m)
Pass angle	Clockwise angle of the pass, measured from the forward direction.	Event	Radians (rad)
Pass type name	Provider-assigned label for passes from interception or loose-ball recovery; all others coded as normal.	Event	Categorical
Percent distance toward goal	Proportion of the pass distance gained toward the opposition goal.	Event	Percent (%)
Play pattern name	Provider-assigned phase of play, such as throw-in, kick-off, goal kick, free kick, counterattack, or corner; other events coded as regular play.	Event	Categorical
Three pressure radii	Number of opposition players within three radii around the ball carrier.	Positional count	Count (n)
Unmarked teammates	Binary indicators for whether at least one teammate was unmarked in each directional sector.	Positional binary	0/1

$$PR(d) = \left\{ \frac{0.15d + 0.85}{4.5}, \left| \frac{ifd \leq 24}{ifd > 24} \right| \right\}$$

Equation 1. Dynamic pressure radius calculation based on distance to goal.

The resulting positional data frame therefore contained binary indicators for unmarked teammates in each direction and counts of opponents within each pressure radius. These variables were then merged to the event data using the unique event identifier. Feature selection was informed by domain expertise from football practitioners, the feature set reported in Peters et al. (2024), and iterative model comparison during exploratory development. Variables were retained when they were tactically interpretable, consistent with the original xPT framework, and showed recurrent value across candidate models. This pragmatic approach was used to ensure that the retained predictors were both methodologically defensible and relevant to real-time turnover prediction. The final feature set is described in Table 2

The final analysis sample comprised 256,433 passes from the 2020/21 English Premier League season. The dataset contained 13 variables covering binary directional passing options, pass-descriptive features, pressing-intensity measures, and spatial coordinates. The binary option variables

showed that viable passing options were present in most events, particularly in the forward direction, while the numeric predictors captured both local pressure and ball trajectory characteristics. The distribution of the target outcome was imbalanced, making discrimination, calibration, sensitivity, specificity, and F-measure more informative than accuracy alone. A descriptive summary of all variables is presented in Table 3.

Event selection and filtering

To ensure the analysis focused on representative open-play passing events, we applied filtering criteria consistent with Peters et al. (2024). Only standard passes were retained, including successful passes, unsuccessful passes, and passes put out of play as operationally defined by the data provider. We excluded initial set-piece passes because they typically occur under different defensive conditions and do not reflect the same open-play decision context. We also excluded crosses because they serve a distinct tactical purpose and are not directly comparable to standard teammate-directed passes. Finally, goalkeeper passes were excluded to maintain focus on outfield decision-making.

Table 3. Model performance across default and alternative feature sets for expected pass turnovers.

Variable	Min Length	Max Length	Empty	Unique	% = 1 (Binary)	Mean	SD	Median
CATEGORICAL								
back_option	NA	NA	0	0	74%	NA	NA	NA
front_option	NA	NA	0	0	88.2%	NA	NA	NA
left_option	NA	NA	0	0	64.6%	NA	NA	NA
pass.type.name	0	12	228,318	3	NA	NA	NA	NA
play_pattern.name	11	14	0	8	NA	NA	NA	NA
right_option	NA	NA	0	0	64.8%	NA	NA	NA
NUMERIC								
ball_movement_speed	NA	NA	0	0	NA	13.94	52.98	12.74
distance_ball_moved	NA	NA	0	0	NA	19.2	12.03	16.32
pass.angle	NA	NA	0	0	NA	0.01	1.62	0
percent_distance	NA	NA	0	0	NA	-0.73	59.27	3.81
pressing_count_1	NA	NA	0	0	NA	0.33	0.55	0
pressing_count_2	NA	NA	0	0	NA	0.52	0.66	0
pressing_count_3	NA	NA	0	0	NA	1.37	1.18	1
x	NA	NA	0	0	NA	59.29	24.77	57.9
y	NA	NA	0	0	NA	39.51	23.48	39

Mixed-effects model overview

Expected Pass Turnovers (xPT) were modeled using a logistic mixed-effects framework with event and positional variables derived from 380 Premier League matches. Random intercepts were included for player ID, position group, and match ID to account for the hierarchical structure of the data and the non-independence of passes nested within players, roles, and matches. Playing positions were categorized into seven outfield groups: center back, full back, wing back, defensive midfield, attacking mid-field, forward and winger. As some playing positions may have overlapping roles and hence can be difficult to differentiate, practitioner analysts were consulted to identify what formations characterized these different positional groups from game to game (Table S1). This model served as the statistical baseline against which the machine learning approaches were compared.

Feature sets

We defined two feature sets. The default feature set was leakage-inclusive and included all contextual and pass-descriptive variables used in the original xPT specification. These comprised of the features outlined in Table 2, along with the random effect features: positional group, player ID and match ID. The alternative feature set was leakage-corrected and included the same variables except for the four pass-descriptive features: *distance_ball_moved*, *ball_movement_speed*, *percent_distance*, and *pass_angle*, as these are only fully observable after the pass has been executed. This alternative set therefore represents the information available at the moment of decision and is appropriate for prospective tactical use.

For the mixed-effects model, random intercepts were retained for player ID, match ID, and position group in both feature sets. For the penalized logistic regression, random forest, and gradient boosting models, player (player ID) and match identifiers (match ID) were excluded and all predictors were treated as fixed effects to avoid introducing identity-based leakage into the nonhierarchical models.

Machine learning algorithms and model specifications

To assess the impact of data leakage across modeling paradigms, we compared the mixed-effects logistic regression model with three complementary machine learning algorithms: penalized logistic regression, random forest, and gradient boosting (XGBoost). These methods were selected to represent a spectrum from interpretable statistical modeling to flexible ensemble learning, allowing us to evaluate whether the leakage effect was consistent across both parametric and non-parametric approaches.

Penalized logistic regression, implemented using elastic net regularization in *glmnet*, was used to reduce overfitting and manage correlated predictors. This approach is particularly useful when the feature set contains several related contextual variables and when model stability is important.

Random forest, implemented using *ranger* (Breiman, 2001), was used to capture non-linear relationships and interactions

without requiring explicit functional specification. Gradient boosting, implemented using *xgboost* (Chen & Guestrin, 2016), was included because of its strong performance in structured prediction tasks and its ability to model complex contextual effects with regularization and shrinkage.

Hyperparameter tuning and cross-validation

All models were tuned using grouped cross-validation stratified by match ID and grid search. Grouping by match ensured that passes from the same match were assigned to only one fold, reducing information leakage and preserving the dependence structure of the data. This approach preserved the dependence structure of the data and minimized leakage across resampling folds; no separate hold-out test set was reserved. Penalized logistic regression, random forest, and gradient boosting were tuned using 5-fold cross-validation with a grid size of 20, whereas the mixed-effects logistic regression model was tuned using 3-fold cross-validation because of its greater computational burden.

This grouped cross-validation strategy was used to estimate model performance under resampling while preventing the same match from contributing to both training and assessment folds. Optimal hyperparameters selected on the basis of cross-validated ROC-AUC for the default and alternative feature sets are reported in Supplementary Tables S2 and S3.

Model evaluation metrics

We evaluated performance using a set of complementary metrics because no single statistic captures all relevant aspects of classification quality. ROC-AUC was used as the primary discrimination metric because it summarizes the ability of the model to rank turnovers above non-turnovers across all thresholds. Accuracy was reported for completeness but interpreted cautiously because it can be misleading in imbalanced data.

Sensitivity and specificity were used to evaluate the balance between correctly identifying turnovers and avoiding false alarms. F-measure was reported as a threshold-dependent summary of precision and recall. Brier score was used as an overall measure of probabilistic prediction quality, reflecting both calibration and discrimination, with lower values indicating better agreement between predicted probabilities and observed outcomes. Calibration curves were also examined to assess whether predicted turnover probabilities aligned with empirical frequencies across the probability range. Feature importance was assessed using SHAP values and partial dependence plots for the gradient boosting models.

Results

Model performance: default feature set (leakage-inclusive)

Table 4 summarizes cross-validated performance for all four algorithms using the default leakage-inclusive feature set. The ensemble methods outperformed the logistic models on discrimination, with gradient boosting achieving the highest

Table 4. Model performance with default (leakage-inclusive) feature set. ROC-AUC, accuracy, sensitivity, specificity, F-measure, and Brier score are reported across all four algorithms.

Model	ROC-AUC	Accuracy	Sensitivity	Specificity	F-Measure	Brier Score
Mixed-effects logistic (glmer)	0.789	0.721	0.710	0.732	0.718	0.188
Penalised logistic (glmnet)	0.786	0.844	0.980	0.139	0.913	0.114
Random forest (ranger)	0.920	0.896	0.968	0.528	0.940	0.075
Gradient boosting (xgboost)	0.924	0.898	0.962	0.571	0.941	0.073

Table 5. Model performance with alternative (leakage-corrected) feature set.

Model	ROC-AUC	Accuracy	Sensitivity	Specificity	F-Measure	Brier Score
Mixed-effects logistic (glmer)	0.707	0.648	0.627	0.669	0.641	0.218
Penalised logistic (glmnet)	0.69	0.837	0.993	0.03	0.911	0.127
Random forest (ranger)	0.737	0.843	0.987	0.1	0.913	0.12
Gradient boosting (xgboost)	0.742	0.843	0.983	0.123	0.913	0.119

ROC-AUC (0.924) followed closely by random forest (0.920). Both substantially exceeded mixed-effects logistic regression (0.789) and penalized logistic regression (0.786), indicating that the inclusion of pass-descriptive features and non-linear modeling both contributed to improved predictive discrimination.

Gradient boosting also achieved the highest overall accuracy (0.898) and F-measure (0.941), while random forest produced similarly strong values (accuracy = 0.896; F-measure = 0.940). Penalized logistic regression showed very high sensitivity (0.980) but low specificity (0.139), suggesting a strong tendency to classify most passes as turnovers and therefore a high false-positive rate. By contrast, the tree-based models maintained a better balance between sensitivity and specificity, with gradient boosting reaching 0.962 sensitivity and 0.571 specificity. Brier scores were lowest for gradient boosting (0.073) and random forest (0.075), indicating stronger overall probabilistic quality.

Model performance: Alternative feature set (leakage-corrected)

Table 5 presents performance for the alternative leakage-corrected feature set, which excludes post-execution pass-descriptive variables. As expected, discrimination decreased across all models, with ROC-AUC values ranging from 0.690 to 0.742. Gradient boosting retained the strongest performance (ROC-AUC = 0.742), followed closely by random forest (0.737), while mixed-effects logistic regression and penalized logistic regression declined to 0.707 and 0.690, respectively.

Although discrimination fell, the tree-based models continued to identify passes completion with useful sensitivity, with values above 0.98 for penalized logistic regression, random forest, and gradient boosting. However, specificity declined markedly, particularly for penalized logistic regression (0.03), indicating a large number of false positives. The tree-based models also retained comparable Brier scores to the default

models, suggesting that their probability estimates remained consistent even after removal of leakage-prone features. Overall, the alternative models show that meaningful turnover prediction is still possible using only pre-execution contextual information, but performance is reduced once post-pass descriptors are excluded.

Quantifying data leakage impact

The performance decrement associated with removing leakage-prone variables is summarized in Table 6. Tree-based models were affected most strongly, with ROC-AUC drops of 0.183 for random forest and 0.182 for gradient boosting, compared with smaller reductions for mixed-effects logistic regression (0.082) and penalized logistic regression (0.096). On average, the decline in ROC-AUC was 0.136, or 15.9%, across models.

Figure 1 visualizes these differences through ROC curves under default and alternative feature specifications. In the default setting, the ensemble models show clear separation from the diagonal, reflecting strong discrimination. After removal of the post-execution pass-descriptive variables, all models shift closer to chance-level performance, but gradient boosting and random forest still outperform the logistic approaches. This pattern indicates that a substantial share of the apparent predictive gain in the default models is attributable to temporally misaligned information.

Feature importance and interpretation

To further interpret the gradient boosting model, SHAP values were used to quantify the relative contribution of each predictor to model predictions (Figure 2). SHAP provides a local and global explanation framework in which feature contributions can be compared across observations and mean absolute SHAP values can be used as a summary of overall importance. In the default

Table 6. Model performance decrement between default and alternative (leakage-corrected) models.

Model	Default ROC-AUC	Alternative ROC-AUC	AUC Decrement	Decrement (%)
Mixed-effects logistic (glmer)	0.789	0.707	0.082	10.4%
Penalised logistic (glmnet)	0.786	0.690	0.096	12.2%
Random forest (ranger)	0.920	0.737	0.183	19.9%
Gradient boosting (xgboost)	0.924	0.742	0.182	19.7%
Mean	0.855	0.719	0.136	15.9%

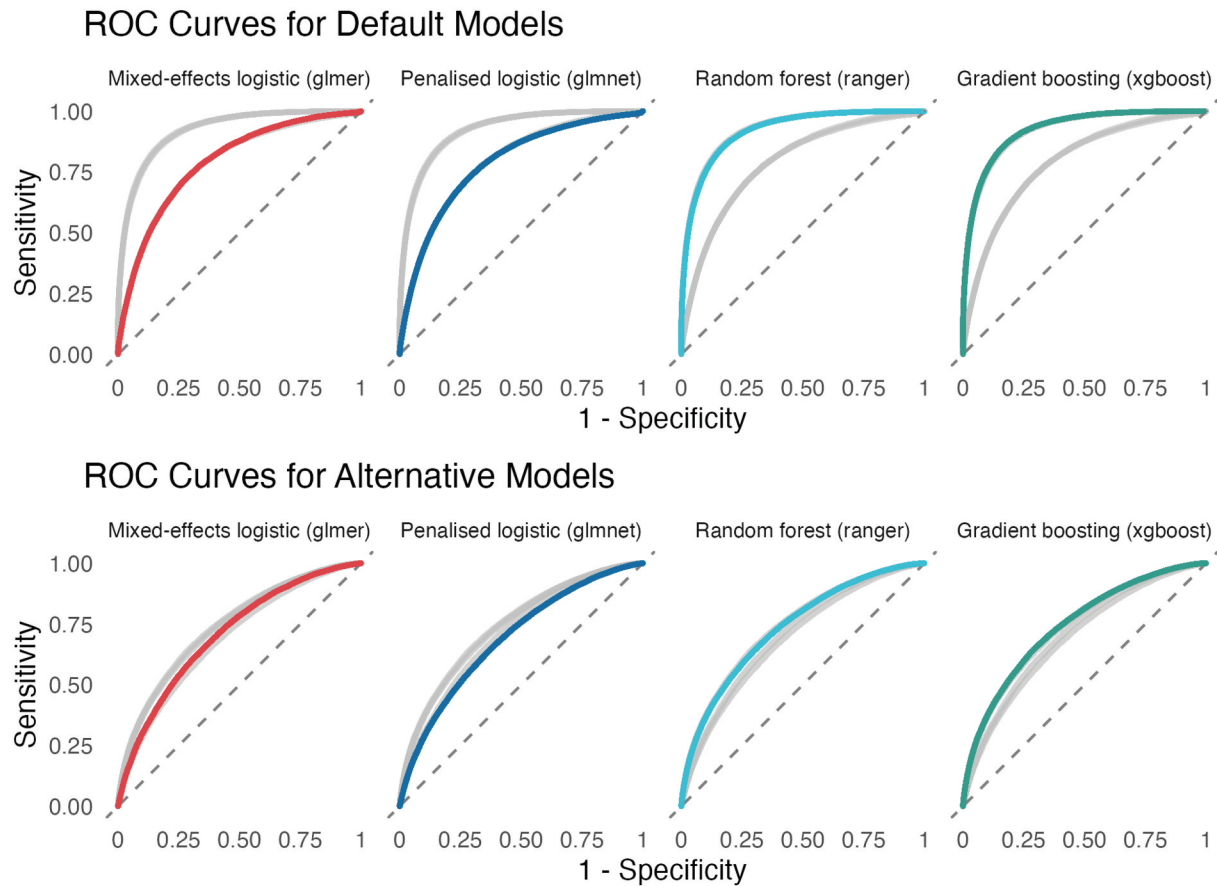


Figure 1. ROC curves comparing default leakage-inclusive and alternative leakage-corrected feature sets across the four models. Removing post-execution pass-descriptive variables reduced discrimination, with tree-based models showing the largest decline in ROC-AUC.

model, the four pass-descriptive variables: *distance_ball_moved*, *pass_angle*, *ball_movement_speed*, and *percent_distance*; dominated the feature importance ranking, indicating that these post-execution descriptors were the principal drivers of turnover prediction when available. In contrast, the contextual and spatial variables contributed less in the leakage-inclusive specification.

In the alternative model, the feature hierarchy changed substantially. Without access to post-execution pass descriptors, the model relied more heavily on pressing intensity variables, tactical context, and spatial position. This shift suggests that the leakage-corrected model is drawing on more plausible pre-execution information, but also that the prediction task becomes harder once outcome-adjacent features are removed. Because SHAP values can be influenced by correlated predictors, they should be interpreted as a measure of relative contribution to model predictions rather than as evidence of causal relationships between predictors and turnover risk.

Partial dependence plots

Partial dependence plots in Figure 3 provide a complementary view of these relationships. In the default model, the directional option variables show relatively flat associations with turnover probability, whereas pressing intensity variables display a non-linear pattern in which turnover risk rises sharply from 0 to 2 pressing counts and then begins to plateau. In the alternative model, the same pressure variables remain informative, but the partial dependence curves show a different level and shape after

removal of post-execution features. These patterns are consistent with the SHAP results and indicate that pressure context is a genuine pre-execution driver of turnover risk.

Calibration and practical deployment

Calibration curves in Figure 4 show that the tree-based models were generally better calibrated than the logistic models in both feature settings. In the default model, mixed-effects logistic regression deviated from the diagonal at lower predicted probabilities, whereas random forest and gradient boosting tracked observed frequencies more closely. Penalized logistic regression showed moderate calibration but tended to overpredict turnover risk at low probability values.

The alternative models displayed a similar ordering, with tree-based methods maintaining the most stable calibration profiles. This is important because tactical decisions require probability estimates that are not only discriminative but also reliable in absolute terms. For example, a predicted turnover probability of 0.70 should correspond to an observed turnover rate close to 70% if the model is to support actionable tactical decisions. Taken together, the calibration results suggest that gradient boosting and random forest offer the most practically useful balance of discrimination and probability reliability in both retrospective and prospective settings.

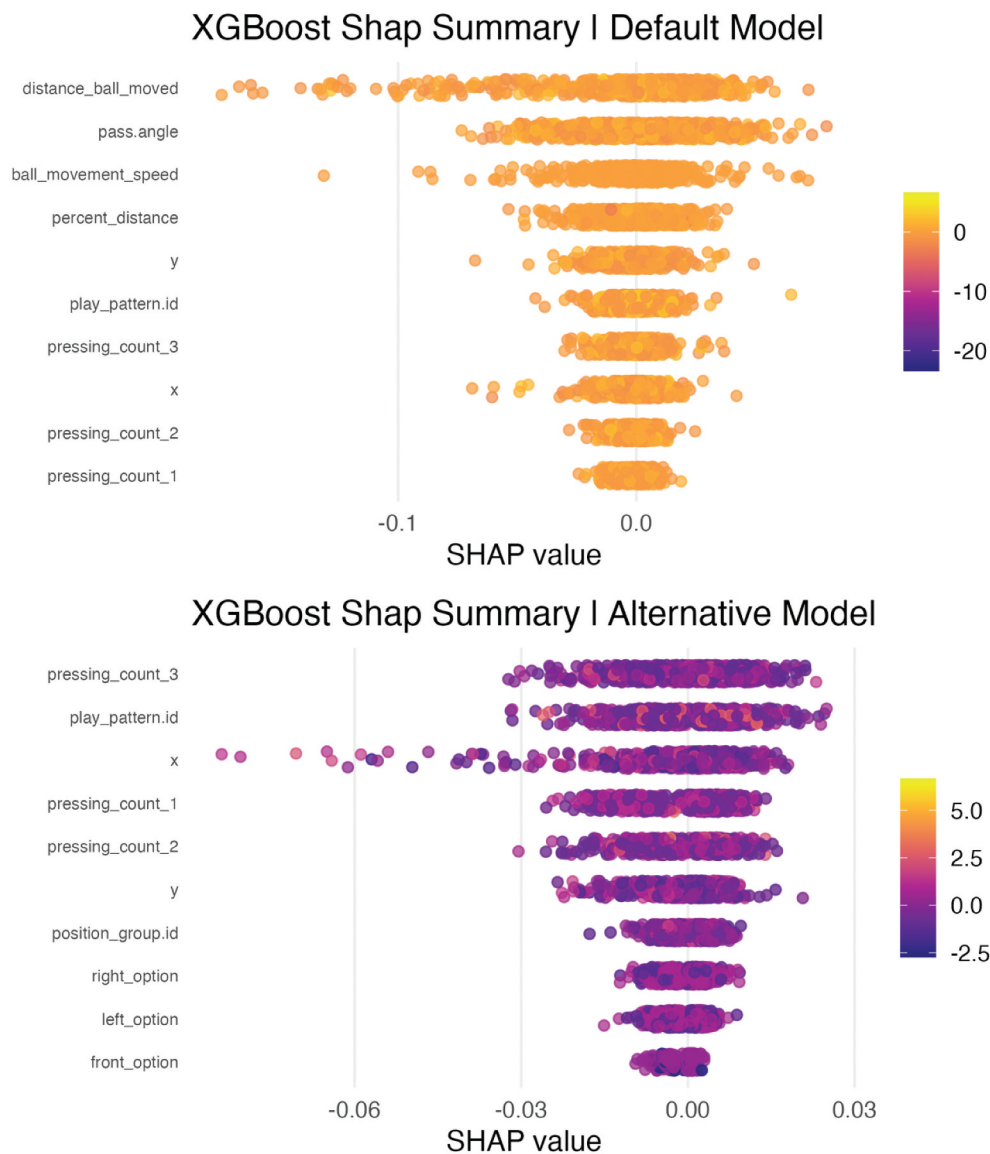


Figure 2. Shap summary plots for the gradient-boosting (XGBoost) model comparing default (top) and alternative (bottom) feature importance. Default models are dominated by post-pass descriptive features, whereas alternative models shift importance toward pressing intensity and tactical context variables, highlighting genuine pre-execution drivers of turnover risk.

Discussion

Data leakage as a methodological framework

This analysis quantifies the temporal misalignment inherent in models that incorporate features only available after the pass has been executed. The reduction in ROC-AUC after removing leakage-prone variables ranged from 0.082 to 0.183, with a mean decrement of 0.136, indicating that post-execution pass descriptors contribute substantially to apparent predictive performance. Tree-based methods were affected most strongly, with random forest and gradient boosting showing decrements of 0.183 and 0.182 AUC points, respectively, whereas the logistic approaches experienced smaller reductions of 0.082 to 0.096. This pattern suggests that more flexible models derive greater advantage from temporally misaligned information, but also that they retain meaningful predictive power even after leakage correction.

Importantly, the best-performing alternative model, gradient boosting, achieved a ROC-AUC of 0.742, which

approaches the default mixed-effects logistic model's ROC-AUC of 0.789. This indicates that the leakage-corrected framework remains practically useful and that substantial predictive signal is available from pre-execution contextual features alone. In other words, correcting leakage reduces apparent performance, but it does not eliminate the ability to identify pass completion risk in a meaningful way.

The SHAP analysis (Figure 2) further clarifies this distinction. In the default model, the dominant predictors were the four pass-descriptive variables: distance_ball_moved, pass_angle, ball_movement_speed, and percent_distance. These are precisely the variables unavailable at the moment of decision, confirming that the default model draws heavily on post-execution information. Once these variables are removed, the model reorganizes around pressing intensity and tactical context variables, including pressing_count_1, pressing_count_2, pressing_count_3, play_pattern.id and position_group.id, which better reflect the information available before the pass is played.

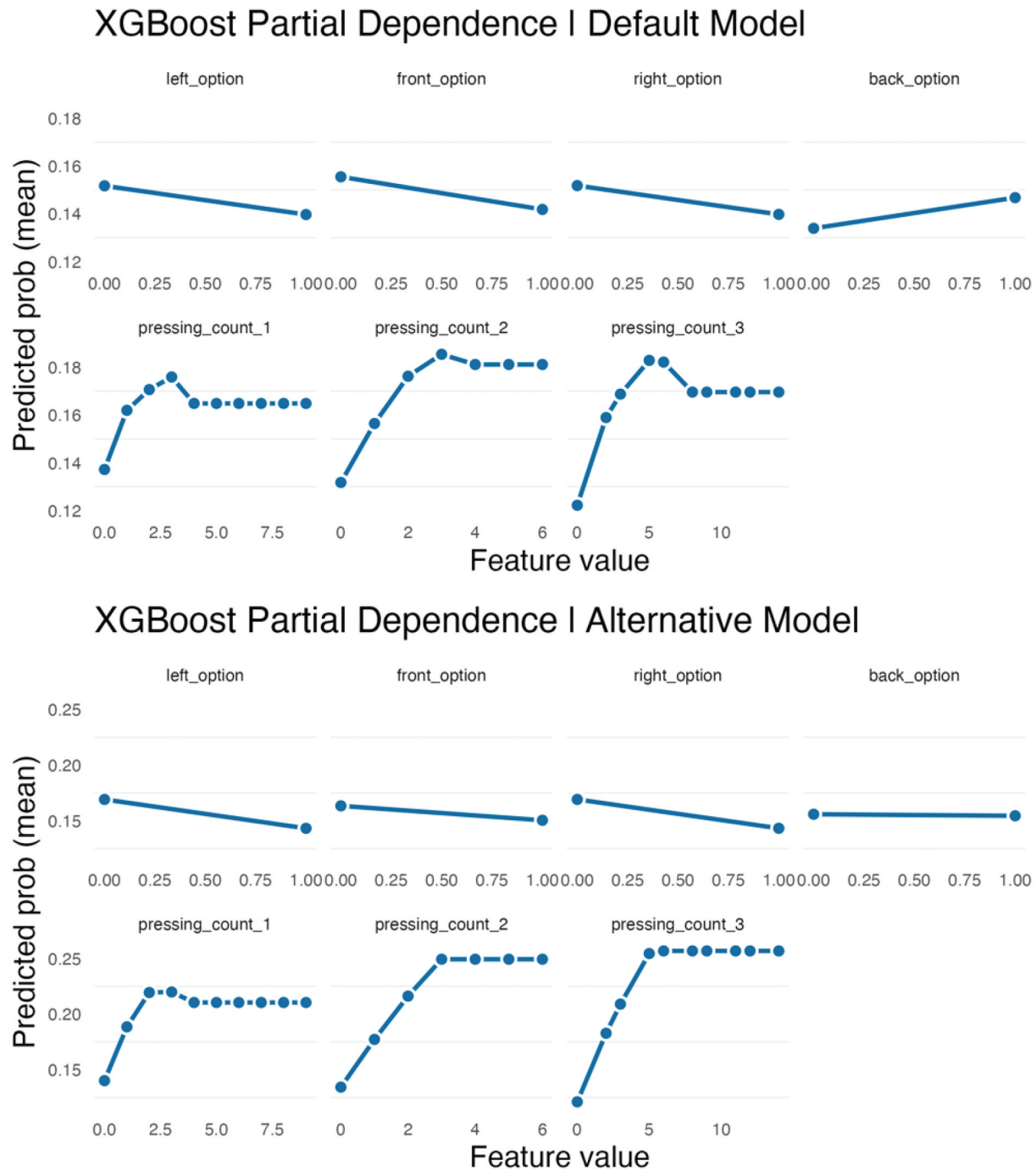


Figure 3. Partial dependence plots for selected features in XGBoost with default feature set. Plots illustrate marginal effects of passing options (back, front, left, right), pressing counts (1, 2, 3 seconds), and spatial coordinates on predicted turnover probability. Nonlinear relationships and threshold effects are evident, particularly for pressing intensity variables.

Partial dependence plots (Figure 3) reinforce this interpretation. In the default models, pass opportunity variables contribute little once the pass-descriptive terms are present, whereas pressing variables show non-linear relationships with turnover probability. In the alternative models, these contextual variables become more prominent, but the partial dependence curves shift in level and shape once post-execution features are removed. This is not a weakness of the model; rather, it reflects the reality that the prospective task is inherently more difficult than retrospective classification.

Calibration results (Figure 4) provide a further layer of support. Although the alternative models sacrifice some discrimination, they retain acceptable calibration and, in some cases, improved specificity relative to the default models. This is important because tactical deployment depends not only on whether a model can rank pass completion, but also on whether its predicted probabilities can be trusted in decision-making. A model with lower ROC-AUC but better temporal validity may be more valuable in practice than a model with inflated discrimination derived from unavailable post-event features.

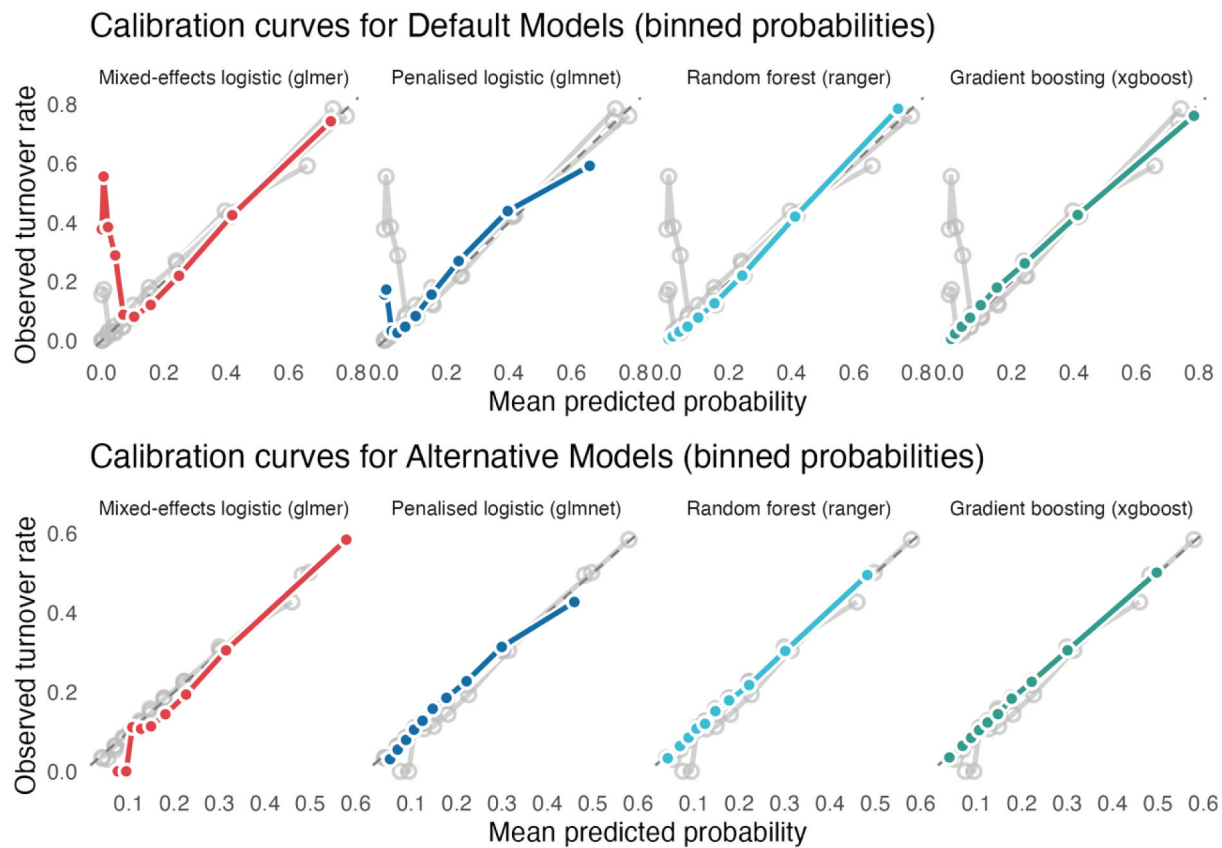


Figure 4. Calibration curves for default (top) and alternative (bottom) models across all algorithms, comparing predicted turnover probabilities with observed outcomes. Tree-based methods (random forest and gradient boosting) exhibit the strongest calibration, while logistic models show greater deviation at lower probabilities. Improved calibration in the alternative models supports more reliable threshold selection for tactical deployment.

These findings establish two complementary but distinct modeling applications:

- (1) Firstly, beyond evaluating completed passes, the leakage-corrected model can also be used to examine decision timing. For example, it can help assess whether a player might have reduced turnover risk by releasing the ball earlier, or whether carrying the ball momentarily helped move from a high-risk to a lower-risk situation. In this sense, the model provides a way to study pass release timing, pressure evasion, and decision-making under defensive constraint, rather than only the outcome of completed passes.
- (2) Secondly, default models with the full feature set are well suited to retrospective classification and analysis. They are useful for identifying which executed passes were risky, examining patterns in turnover-prone situations, and supporting post-match technical review. For these purposes, the original xPT framework and equivalent machine learning models remain valuable because the goal is to explain or classify completed actions.

The key implication is that leakage correction is not merely a technical adjustment. It changes the purpose of the model. A default model is retrospective by construction, whereas a leakage-corrected model is prospective by design. Practitioners should therefore choose between them based on the intended use

case rather than assuming that the highest-performing model is automatically the most useful one.

Relation to xPT

The original expected pass turnovers framework published by Peters et al. (2024) remains a strong contribution to football analytics. Our findings do not challenge its value; instead, they clarify the context in which it is most appropriately used. The original framework is highly suitable for post-hoc analysis, pattern recognition, and retrospective classification, where access to pass-descriptive variables is not problematic.

Our results extend that work by showing that a leakage-free version of the task still retains substantial predictive value. This is an important point because it suggests that the turnover risk signal is not created solely by post-execution pass descriptors. Rather, there is genuine pre-execution information in pressing intensity, spatial position, and tactical structure that can be used to support live decision-making.

Implications for machine learning in Sports

This work underscores a broader principle in sports analytics: model inputs must align with the timing of the decision they are intended to support. Machine learning methods are especially vulnerable to temporal leakage because their flexibility can make them highly effective at exploiting variables that are informative but

unavailable at decision time. The result may be impressive predictive performance that does not translate into practical utility.

Football has already learned this lesson when it comes to the metric—expected goals (xG). Pre-shot xG (or standard xG) is used more widespread because it aligns with the decision moment, whereas post-shot models are better suited to retrospective analysis of finishing and goalkeeping. The same logic applies here. Leakage-corrected turnover models are the appropriate analogue of pre-shot xG for pass decision timing: they may be less accurate in a strict predictive sense, but they are more valid for the tactical question at hand.

This distinction is important for both analysts and coaches. A model that is temporally aligned with the decision context is more likely to be trusted, implemented, and interpreted correctly. By contrast, a model that achieves higher AUC through unavailable information may be statistically stronger but operationally serve a different purpose.

Practical recommendations

For retrospective analysis and match review, the default models are preferable because they provide the highest discrimination and are well suited to identifying risky completed passes. They are useful for technical review, opposition analysis, and pattern recognition after the match (Tenga et al., 2010).

For tactical preparation and real-time decision-making

The alternative leakage-corrected models should be preferred. The loss in AUC is a reasonable trade-off for temporal validity, especially when the aim is to make pressing decisions before a pass is executed. Beyond evaluating completed passes, the leakage-corrected model can also be used to examine decision timing. For example, it can help assess whether a player might have reduced turnover risk by releasing the ball earlier, or whether carrying the ball momentarily helped move from a high-risk to a lower-risk situation. In this sense, the model provides a way to study pass release timing, pressure evasion, and decision-making under defensive constraint, rather than only the outcome of completed passes.

For interpretability and coaching communication

The mixed-effects logistic regression model remains attractive because it offers a transparent summary of contextual effects while still performing competitively relative to the more flexible machine learning methods. In environments where explanation matters as much as raw predictive performance, this may be the most practical option.

Limitations

This study is limited to a single season of English Premier League data. Although the sample is large, temporal generalization across seasons and cross-competition validation remain important next steps. In addition, the comparison was restricted to a binary distinction between leakage-inclusive and leakage-corrected feature sets. Intermediate feature combinations could reveal more nuanced trade-offs between predictive performance and temporal validity.

A further limitation is that this analysis relies on event and freeze-frame data rather than continuous tracking data. More

detailed movement information may improve the estimation of pressing context and spatial control in future work. Finally, turnovers in football are not limited to unsuccessful passes; they can also arise from poor ball control or unsuccessful dribble attempts. Extending the framework to a broader range of possession losses would provide a more complete account of turnover risk.

Future directions

Future work should explore tracking-based models that capture continuous movement, velocity, and interception potential (Spearman, 2018). Pitch control methods could also strengthen the representation of defensive influence and passing options in space. In addition, more recent tactical constructs such as rest defense (Peters et al., 2026) may offer a useful way to model the organization of the defensive structure at the moment of possession loss.

Another promising direction is to broaden the definition of turnover beyond passing errors alone. Including poor first touches, unsuccessful dribbles, and other possession-loss events would allow the framework to model possession vulnerability more comprehensively. Together, these extensions would improve both the tactical realism and the operational usefulness of the model.

Conclusion

This study systematically quantified and contextualized data leakage in turnover prediction models by comparing four machine learning algorithms across leakage-inclusive and leakage-corrected feature sets. The original Expected Pass Turnovers framework and equivalent default approaches remain feasible and valuable for retrospective classification and match analysis. However, for prospective tactical deployment, leakage-corrected models offer stronger methodological alignment, albeit with a defensible reduction in AUC performance.

The best-performing alternative model, gradient boosting, achieved a ROC-AUC of 0.742, which was close to the default mixed-effects logistic model's ROC-AUC of 0.789. This indicates that substantial predictive value is retained after correcting for temporal leakage. Tree-based methods experienced the largest performance drop when leakage-prone variables were removed, but their corrected versions remained competitive with simpler models and therefore retain practical value for real-time tactical guidance.

Overall, the findings show that the default and leakage-corrected frameworks serve complementary purposes rather than competing ones. Retrospective analysis benefits from larger feature sets and the highest discrimination, whereas real-time decision-support requires leakage correction to ensure that predictions are temporally valid and actionable at the moment of coaching decisions. In addition to informing real-time tactical guidance, the leakage-corrected framework may also help evaluate decision timing, including whether a player released the ball too early, too late, or after successfully reducing pressure through carrying or dribbling. The fully reproducible R pipeline used in this study also provides a transparent template for future validation, extension, and adaptation.

By foregrounding temporal alignment as a core methodological issue, this study contributes to more rigorous and deployable analytics in professional football. The distinction between retrospective and prospective modeling should inform turnover prediction and, more broadly, predictive sports analytics where the timing of information is critical.

Acknowledgments

CRedit: **AP:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing; **NP:** Writing – review & editing; **MD:** Writing – review & editing; **NJ:** Writing – review & editing.

Author contributions

CRedit: **Andrew Peters:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing; **Nimai Parmar:** Writing – review & editing; **Michael Davies:** Writing – review & editing; **Nic James:** Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Andrew Peters  <http://orcid.org/0000-0002-2586-1523>
 Nimai Parmar  <http://orcid.org/0000-0001-5540-123X>
 Nic James  <http://orcid.org/0000-0002-4181-9501>

Data availability statement

All analyses were conducted using R (version 4.5.1) in a fully reproducible, parallelized pipeline. The complete modeling framework—including data preprocessing, model training, evaluation, and visualization scripts—is openly available at: https://github.com/andypetes94/xTurnovers_Models

The repository includes:

- A shell orchestrator (`run_machine_learning_pipeline.sh`) for automated execution.
- `/run_machine_learning_pipeline.sh <input_csv> <output_dir> <model_type>`

where `<model_type> ∈ {default, alt}`.

- Modular R scripts for data loading (`turnover_pipeline_run.R`), model training (`turnover_pipeline.R`), and diagnostics (`turnover_evaluation_suite.R`).
- Generated figures, metrics, and model outputs.
- An anonymized synthetic dataset replicating the variable structure of the original data.

All scripts are version-controlled, documented with inline comments, and reproduce all results and figures presented in this paper. The exact dataset cannot be shared due to its commercial licensing by *StatsBomb*. To facilitate replication and further research, we provide an anonymized and randomly generated dataset based on the same variable structure.

Statement of ethics

This study used secondary, anonymized data obtained from a third-party provider, with no direct interaction with human participants. The research was conducted in accordance with established ethical guidelines for human research (e.g., Declaration of Helsinki) and did not require formal institutional ethics approval.

References

- Beato, M., Youngs, A., & Costin, A. J. (2024, May). The analysis of physical performance during official competitions in professional English football: Do positions, game locations, and results influence players' game demands?. *The Journal of Strength and Conditioning Research*, 38(5), e226–e234. <https://doi.org/10.1519/JSC.0000000000004717>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Carling, C., Williams, A. M., & Reilly, T. (2005). *Handbook of soccer match analysis: A systematic approach to improving performance*. Routledge.
- Chassy, P. (2025). Toward explainable data and sports analytics: Bridging prediction and understanding. *The American Statistician*, 79(2), 156–168. <https://doi.org/10.1080/00031305.2025.2541085>
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., & Maltagliati, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football analytics. *2015 IEEE International Conference on Data Science and Advanced Analytics*, Paris, France (pp. 1–10). IEEE.
- Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8(1), Article 1. <https://doi.org/10.1515/1559-0410.1418>
- Deshpande Sameer, K., & Katherine, E. (2020, June). Expected hypothetical completion probability. *Journal of Quantitative Analysis in Sports*, *De Gruyter*, 16(2), 85–94.
- Dong, Q. (2022). Leakage prediction in machine learning models when using data from sports wearable sensors. *Computational Intelligence and Neuroscience*, 2022, 1–9. <https://doi.org/10.1155/2022/5314671>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Herold, M., Müller, S., Weidlich, F., & Bestehorn, C. (2022). Shortcomings of applying data science to improve football. *Frontiers in Sports and Active Living*, 4, Article 1019990. <https://doi.org/10.3389/fspor.2022.1019990>
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15. <https://doi.org/10.1145/2382585.2382599>
- Kuhnt, S., & Schiffner, S. (2021). Cross-validation and prediction accuracy. *Computational Statistics*, 36, 1699–1721. <https://doi.org/10.1007/s00180-021-01072-7>
- Lindberg, E., & Soderbarg, A. (2020). Machine learning for passing success prediction in soccer [Master's thesis]. Royal Institute of Technology (KTH).
- Narayanan, S., & Pifer, N. D. (2024). An xG of their own: Using expected goals to explore the analytical shortcomings of misapplied gender schemas in football. *Journal of Sport Management*, 38(2), 92–109. <https://doi.org/10.1123/jsm.2023-0022>
- Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). Playmakers: A cohesive framework for tactical analyses in soccer. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA (pp. 2672–2681). <https://doi.org/10.1145/3292500.3330871>
- Peters, A., Parmar, N., Davies, M., & James, N. (2024). Expected pass turnovers (xPT): A model to analyse turnovers from passing events in

- football. *Journal of Sports Sciences*, 42(10), 1234–1245. <https://doi.org/10.1080/02640414.2024.2379697>
- Peters, A., Parmar, N., Davies, M., & James, N. (2026). A rule-based approach to classify counterpressing - analysis of its risks and relationship with rest defence. *International Journal of Performance Analysis in Sport*, 26(1), 206–222. <https://doi.org/10.1080/24748668.2025.2473799>
- Rico-González, M., Los Arcos, A., Buldú, J. M., Clemente, F. M., & Pino-Ortega, J. (2023). Machine learning application in soccer: A systematic review. *Sports Medicine*, 53(2), 255–275. <https://doi.org/10.1007/s40279-022-01689-w>
- Spearman, W. (2018). Physics-based modelling of pass trajectories in football. *13th MIT Sloan Sports Analytics Conference*, Boston, MA, USA. MIT.
- StatsBomb. (2024). *xPass: Expected pass completion*. <https://statsbomb.com>
- Stival, L., Pinto, A., Andrade, F. D. S. P. D., Santiago, P. R. P., Biermann, H., Torres, R. D. S., & Dias, U. (2023). Using machine learning pipeline to predict entry into the attack zone in football. *PLOS ONE*, 18(1), e0265372. <https://doi.org/10.1371/journal.pone.0265372>
- Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010). Effect of playing tactics on goal scoring in Norwegian professional soccer. *Journal of Sports Sciences*, 28(3), 237–244. <https://doi.org/10.1080/02640410903502774>
- Wong, A., Liu, K., & Thompson, W. (2025). Predicting soccer pass outcomes using machine learning and spatiotemporal features. *Frontiers in Sports and Active Living*, 7, Article 1638446. <https://doi.org/10.3389/fspor.2025.1638446>