# Short Ball and Long Ball Playing Styles' Influence in Professional Soccer Matches

Matteo Deambrogio
matteod@mit.edu
31738849

Chris Tran
chris_tran@college.harvard.edu
61594876

Josh Zhang
JoshZhang@college.harvard.edu
21594970

Chloe Headland
cheadland@college.harvard.edu
71535320

May 17, 2025

# Introduction and Motivation

Modern professional soccer is shaped by a wide range of tactical philosophies, with teams around the world adopting various approaches to gain a competitive edge. Specifically, among these, two styles stand out for their contrasting strategies and widespread success: the short-passing, possession-based tiki-taka and the direct, vertical long-ball approach. While both styles have been used effectively by top clubs and national teams, the question of which style leads to more consistent performance and match success remains contested. This debate has gained renewed relevance as analytics continue to influence how managers design game plans and how clubs assess team performance.

Understanding this, the relationship between play style and match outcomes is critically important for coaches, analysts, and fans alike. In particular, identifying which tactical tendencies are more likely to lead to winning outcomes could inform in-game decision-making, recruitment strategies, and broader tactical evolution in the sport.

To date, there is no clear consensus in the academic literature on whether possession-based tiki-taka or direct long-ball football delivers superior results, owing largely to the scarcity of publicly available studies and the proprietary nature of most clubs' in-house performance analyses. At the same time, high-profile successes on the world stage—from Guardiola's Barça and Manchester City teams to Mourinho's and Klopp's more vertical, transitional sides—underscore that both approaches can yield top-level achievements when executed with the right personnel and tactical discipline.

## Research Question

In this project, we aim to explore how different playing styles influence team performance in professional soccer. Specifically, we compare the effectiveness of short-passing, possession-oriented tiki-taka strategies with that of direct, long-ball approaches. Using English Premier League and Italian Serie A match-level data from 2022 to 2024, we seek to answer the following question:

How do different playing styles — specifically the short-passing, possession-based tiki-taka versus the direct, long-ball approach — influence team performance outcomes in professional soccer matches?

# Data Description

The data we use in this study is built from detailed team–match passing statistics compiled by FBref. This dataset encompasses match-level data from two full seasons of the English Premier League and Italian Serie A: 2022–23 and 2023–24. Each record corresponds to a single team's performance in a specific match, meaning each match appears twice—once for each participating team. After merging both seasons, the final dataset consists of roughly 760 matches, resulting in approximately 1520 team–match observations.

This dataset provides itemized statistics on team passing behavior, including short (0-15 feet), medium (15-30 feet), or long (30+ feet) pass counts and accuracies, progressive passes, key passes, final-third entries, and more. These metrics allow us to

study how a team's stylistic tendencies—particularly around passing—correlate with performance outcomes such as goals, xG, and match results.

Compared to other publicly available soccer datasets, we believe this version is among the most analysis-ready and well-structured. After a thorough inspection of the raw data, including checks for consistency across matches and teams, we verified that the dataset contains no missing values. Furthermore, its uniform match structure and completeness make it especially well-suited for modeling match outcomes based on tactical styles.

To view all the variables in this dataset and their corresponding descriptions, please refer to Table 8.

# Exploratory Data Analysis

## Creation of New Variables

Below we create a list of new variables from our given dataset. Each will ultimately be used to help evaluate team performance later elaborated on.

### Percentage-Based Passing Metrics

While the dataset already includes completion rates for short, medium, and long passes, these metrics alone do not reveal how frequently each pass type is used. To capture stylistic tendencies more effectively, we compute the proportion of each pass type relative to a team's total pass attempts.

These percentage-based metrics allow us to distinguish between teams that favor quick, short-ball buildup and those that rely more heavily on direct long passes. By focusing on the distribution of pass types rather than their volume or success rate alone, we gain a clearer understanding of each team's tactical intent.

| Variable | Description |
|---|---|
| Short Pass % | Proportion of passes that are short |
| Medium Pass % | Proportion of passes that are medium |
| Long Pass % | Proportion of passes that are long |

Table 1: Percentage-Based Passing Metrics

### Team Playstyle Classification

To analyze how tactical decisions influence match outcomes, we first needed a systematic way to classify each team's playstyle. In particular, we sought to distinguish whether a team adopted a short-ball or long-ball approach on a match-by-match basis as teams often change their playstyles varying on a match-to-match basis.

Given that our dataset is already structured at the match level, no additional cleaning was required. In particular, we use unsupervised learning via $k$-means clustering for its simplicity, scalability, and effectiveness at separating observations based on Euclidean distance—making it well-suited for identifying contrasting styles of play.

Our clustering is based on the proportion of short and long passes attempted per match, two ratio-based features that capture a team's tactical intent. These variables require no further normalization and allow for intuitive interpretation.

We apply $k$-means clustering with two clusters (`centers = 2`; To represent each playstyle) and 1000 random initializations (`nstart = 1000`) to ensure robustness and minimize within-cluster variance.


Figure 1: K-means clustering of match-level team playstyles based on short and long pass proportions.

While there is no definitive quantitative metric to evaluate clustering accuracy in this context, visual inspection of Figure 1 confirms that $k$-means captures the dominant trends in team passing behavior, with some expected overlap due to match-specific variability.

Each team–match is then labeled with a binary indicator `short`:

- `short` $= 1$ indicates a short-ball, possession-based style.
- `short` $= 0$ indicates a long-ball, direct approach.

### Home Indicator

We include a binary indicator (`is_Home`) to capture whether a team is playing at home, as home-field advantage is a well-documented factor in soccer. Playing at home can influence team performance due to factors like crowd support, familiarity with pitch conditions, and reduced travel fatigue. This variable allows us to control for these contextual effects in our modeling of match outcomes.

### Elo-Based Power Index

To account for relative team strength, we construct a dynamic power index using the Elo rating system. Elo ratings provide a continuously updated measure of team quality based on both match outcomes and the strength of the opponent. We include

this variable to assess whether relative team strength influences performance, and to control for imbalances in matchups that could confound tactical effects.

The Elo system updates each team's rating using the formula:

$$R_{\text{new}} = R_{\text{old}} + K \cdot (\text{Actual Result} - \text{Expected Result})$$

, where $K$ is a sensitivity parameter (set to the default sensitivity of 20 in our case), and the expected result is computed as:

$$E = \frac{1}{1 + 10^{(R_{\text{opponent}} - R_{\text{team}})/400}}$$

This method allows us to capture the evolving strength of teams over time, providing an interpretable and empirical indicator of difficulty that complements our analysis of playstyle and performance. We ultimately add both team-match's elo and their opponent's for every row.

Notably, as all Elo models begin by assigning each team a standard baseline rating (1500 in our case), it takes time for the ratings to meaningfully reflect true team strength. Early match results may not fully capture relative quality until the system has had sufficient data to adjust. Therefore, we use the 2022-23 season exclusively to initialize and calibrate Elo ratings, and restrict all further analysis to the 2023-24 season, where team power indexes are more accurately established.

## Evaluating Performance and Predictor Selection

### Reponse Variable Selection Rationale

Ultimately, as we seek to to evaluate team performance, we use expected goals (`xG`) as our outcome variable, as it captures the quality of scoring opportunities created rather than the randomness of goal conversion. This makes it a more stable and informative indicator of attacking effectiveness over time.

### Predictor Variable Selection Rationale

To model how team tactics and context influence `xG`, we select predictors that reflect both the volume and efficiency of passing, the distribution of pass types (short, medium, long), and indicators of forward intent. These features allow us to quantify not only how much teams pass, but how and why. We also include contextual factors—such as home-field advantage, playstyle classification, and pre-match Elo ratings—to account for game-specific conditions and relative team strength.

## Predictor Variable Analysis

To prepare our predictor variables for modeling, we construct two distinct versions of the dataset: a standard version and an alternative version.

First, the standard dataset involves applying normalization (z-score standardization) to all predictors. Normalization ensures that all metrics are on a consistent scale, which is critical for assigning appropriate weight during modeling and for making model coefficients directly interpretable across predictors.

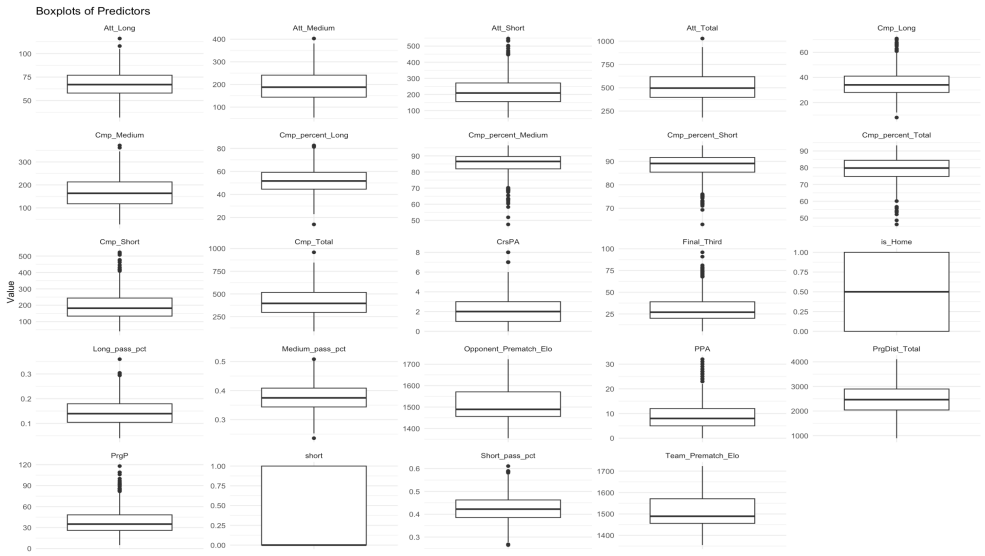


Figure 2: Boxplots of all raw predictor variables.

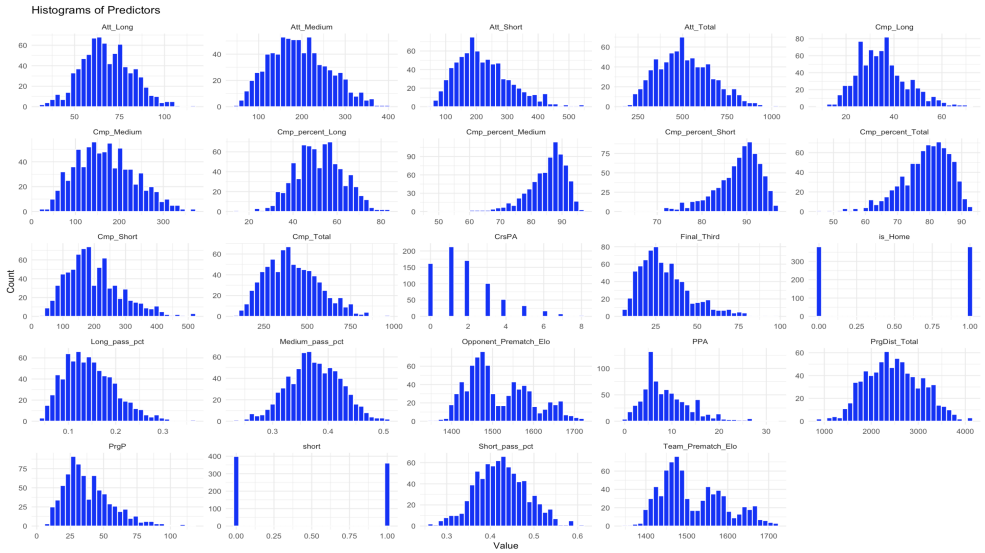


Figure 3: Histogram of all raw predictor variables.

Second, the alternative dataset follows a two-step process: we first apply transformations to variables with high skewness—using a natural logarithm for right-skewed variables and a square root for left-skewed ones—and then normalize. This approach addresses the non-normal distributions observed in Figure 2 and Figure 3. Skewed predictors can negatively affect model performance by introducing non-linearity, heteroskedasticity, and the influence of outliers. Reducing skewness prior to normalization improves the linearity and interpretability of these features, and makes their standardized values more statistically meaningful.

By comparing models trained on both datasets, we can assess the impact of distributional shape on performance. While we hypothesize that the alternative dataset may improve model stability and interpretability, we also admit that transformations may introduce noise or distort important relationships. As such, we proceed with both

datasets in our modeling analysis to ensure robustness. Note that we will use the language of standard and alternative, when referring to our datasets for the remainder of this paper.

# Methods

## Metrics of Evaluation

To evaluate and compare model performance across different specifications, we use two key metrics. First,we select the adjusted $R^2$ metric in order to judge its ability to understand the variance of Expected Goals, adjusted for the number of predictors. Unlike the standard $R^2$, the adjusted $R^2$ penalizes the inclusion of irrelevant features, making it especially useful for comparing models of differing complexity. From this, we can see how well our model understands $xG$'s variance and deviations.

Second, we report the Akaike Information Criterion (AIC score), which balances goodness of fit to complexity. A lower AIC indicates a better model in terms of explanatory power relative to its complexity. This is particularly important as we explore alternative variable sets, reskewed transformations, and more complex modeling techniques. Together, these two metrics provide a robust framework for evaluating both the explanatory value and efficiency of our models.

This is used for all linear based models , we will be conducting a multitude of linear-variant models to predict on $xG$ and conducting further analysis on the best model afterward.

## Baseline Linear Regression Methodology

Ordinary Least Squares (OLS) regression models were fit using the `lm` function in R. Our baseline model for predicting expected goals (`xG`) was an OLS regression using the full set of predictors identified in our exploratory data analysis.

In OLS regression, the model seeks to minimize the following loss function:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

where $y_i$ is the true expected goals value for observation $i$ and $\hat{y}_i$ is the predicted value from the model.

Formally, the regression equation is given by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \tag{2}$$

where:

- $y_i$ is the expected goals value for team–match observation $i$,
- $x_{ji}$ represents the $j$th predictor variable for observation $i$,
- $\beta_j$ are the coefficients to be estimated,
- and $\varepsilon_i$ is an error term assumed to be independently and identically distributed, with mean zero and constant variance.

This baseline model serves as a benchmark for evaluating the performance of more complex modeling techniques such as regularized regression and ensemble methods. This logic will be utilized for the baseline model and the stepwise regression.

**Stepwise Model Methodology**

Pivoting from our baseline model, we seek to devise a step-wise baseline model considering all interactions as the upper bound and just the intercept as the lower bound. All baseline model methods are appliable here as well. This will result in the optimal standard linear regression with our provided variables, when prioritizing AIC score. This is to find the optimal linear composition.

## Lasso & Ridge Model Methodology

To further investigate the relative importance of our selected predictors and assess model performance under regularization, we apply both Lasso and Ridge regression. These penalized regression techniques allow us to evaluate coefficient shrinkage, identify less informative or redundant variables, and compare overall model fit using metrics such as adjusted $R^2$, AIC, and Mean Squared Error (MSE). They also serve as a robustness check for the variable selection performed in our baseline linear regression.

Lasso (Least Absolute Shrinkage and Selection Operator) improves interpretability by shrinking some coefficients to exactly zero, effectively performing variable selection. In contrast, Ridge regression retains all features and distributes shrinkage across correlated predictors, which may offer better predictive performance in high-dimensional or multicollinear settings, albeit at the cost of interpretability.

Formally, both methods minimize a penalized loss function. For Lasso, the objective is:

$$L_{\text{Lasso}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

While for Ridge regression, the loss function is:

$$L_{\text{Ridge}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

To implement these models, we use the `glmnet` package in R. The optimal shrinkage parameter $\lambda$ is selected via 10-fold cross-validation using the `cv.glmnet` function, a standard approach that balances bias and variance and guards against overfitting. Once the optimal $\lambda$ is identified, we fit the regularized model using the `glmnet` function.

In addition to traditional model metrics listed earlier, we also report the best $\lambda$ value selected by cross-validation. This serves as a diagnostic for how much penalization is needed—lower values of $\lambda$ suggest stronger signal in the predictors and less need

for regularization, while higher values indicate overparameterization or noise in the feature space.

By applying both Lasso and Ridge, we gain complementary insights: Lasso highlights which variables are essential, while Ridge reveals how the full feature set contributes to prediction.

## Random Forest Methodology

To further strengthen our analysis, we introduce a Random Forest model using the full set of predictors. Random Forests aggregate the predictions of multiple decision trees to improve predictive accuracy and reduce overfitting. This method is particularly well-suited for our dataset, as it can automatically capture nonlinear relationships and complex interactions between variables beyond the capabilities that traditional linear models are able to achieve.

In addition, Random Forests are robust to multicollinearity and provide intrinsic measures of variable importance, offering valuable insights into which features most significantly influence expected goals. Unlike linear models, which can overweight dominant predictors, Random Forests help mitigate such bias by averaging across diverse tree structures.

We ultimately implement this model using the `randomForest` package in R.

Since Random Forests are not based on likelihood functions, we evaluate model performance using Adjusted $R^2$ and Mean Squared Error (MSE), capturing both explained variance and average prediction error. In the absence of a separate validation set, we use Out-of-Bag (OOB) error as a proxy for out-of-sample performance as well. We also analyze feature importance scores to interpret the model's internal decision structure.

The Random Forest algorithm minimizes the random forest mean squared error function over an ensemble of $T$ trees:

$$\mathcal{L}_{\text{RF}} = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t(x_i) \right)^2 \tag{5}$$

where $\hat{f}_t(x_i)$ is the prediction for observation $i$ from the $t$th decision tree in the ensemble, and $y_i$ is the true expected goals value. This reflects the average squared error across all trees in the forest.

As for the parameters within the Random Forest, we set `ntree = 1000`, well above the minimum of 100 trees typically required to ensure stability and reduce variance. This ensures the model converges to stable predictions across randomized splits.

For `mtry`, the number of predictors randomly considered at each split, we use the standard regression heuristic $\lfloor \sqrt{p} \rfloor$, where $p$ is the number of predictors. This encourages model diversity by preventing any single strong predictor from dominating splits, which in turn reduces correlation between trees, improves generalization, and mitigates overfitting.

Finally, Random Forests provide variable importance measures, allowing us to identify which predictors are most influential in determining expected goals.

## Mixed Modeling Methodology

Finally, we introduce mixed effects models to account for the possibility of hierarchical structure in the data. In previous models, the intercept was assumed to be fixed and shared across all teams, implying that each team operates under the same baseline level of expected goals. However, in reality, teams differ in inherent quality, tactics, and consistency—factors that may not be fully captured by observable predictors alone.

Mixed models allow us to account for this unobserved heterogeneity by incorporating random effects. Specifically, we include team-specific random intercepts to reflect that each team may start from a different baseline level of performance. This allows us to model both population-level effects (fixed effects) and group-level variation (random effects) simultaneously, improving both model realism and predictive power.

For this mode, we implement three model variations: using the full predictor set with and without interaction terms (up to second-degree), as well as versions built from the stepwise-selected variables. This helps us assess how interaction complexity and variable selection interact with team-level variation.

Although Adjusted $R^2$ is not defined for mixed models, we report the Conditional $R^2$ statistic as an parallel metric. Conditional $R^2$ reflects the proportion of variance explained by both fixed and random effects combined, providing a comprehensive measure of model fit.

## Baseline Logistic Model Methodology

We fit two simple logistic-regression baselines predicting the probability that the home side wins (`Home_better` = 1) from pass-ratio features. Model 1 has a right-hand variable the percentage of passes that were 'short' for the home team over the away team. Model 2 has the number of passes that were 'short' for the home team over the away team.

- **Model 1:**

$$\Pr(\text{Win}) = \sigma\big(\beta_0 + \beta_1 \,\texttt{HomePass\_to\_AwayPass\_perc}\big).$$

- **Model 2:**

$$\Pr(\text{Win}) = \sigma\big(\beta_0 + \beta_1 \,\texttt{HomePass\_to\_AwayPass\_short}\big).$$

## Logistic With Subset of Teams Methodology

Finally, we decided to re-run our base model, but using only a subset of the initial teams, specifically ones that are considered 'top teams'. The teams are shown in Table 2. This comes from the fact that we wanted to compare teams that are generally considered similarly strong, to avoid the effect that stronger teams do more passes and won more against weaker teams - regardless of their play style. The choices

of the teams were taken through a combination of personal experience and experts' opinions.

The two models fitted are repeated below for the sake of clarity, but are equivalent to the baseline models descripted above.

- **Model 1:**

$$\Pr(\text{Win}) = \sigma\big(\beta_0 + \beta_1 \, \texttt{HomePass\_to\_AwayPass}\big).$$

- **Model 2:**

$$\Pr(\text{Win}) = \sigma\big(\beta_0 + \beta_1 \, \texttt{HomePass\_to\_AwayPass\_short}\big).$$

Table 2: Team Categories by Playing Style

| Category | Teams |
|---|---|
| Tiki Taka | Manchester City, Arsenal, Brighton |
| Long Ball | Juventus, Roma |
| Other Top Teams | Bologna, Milan, Manchester United, Fulham |

# Results

## Baseline Linear Model Results

At this point, we have devised a baseline linear model with all our predictors to see how much explained variance we can initially capture. From Table 9, we see that we start with a baseline Adjusted $R^2$ of 0.3179 and an AIC score of 1717.797 when conducting a baseline linear regression with no interactions on the standard dataset.

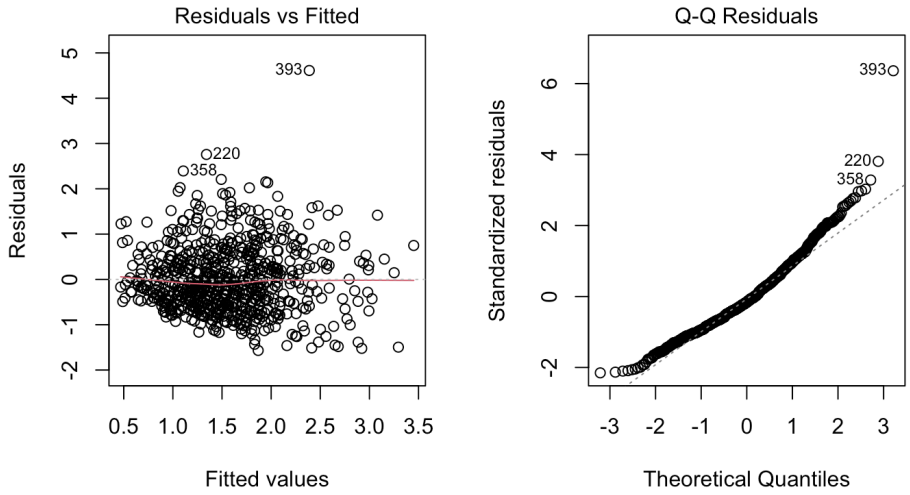


Figure 4: Baseline Model w/ Standard Dataset Model Diagnostics

Accordingly, we check if linear assumptions were violated by this baseline model by analyzing diagnostic plots from Figure 4. Analyzing the Residuals vs. Fitted graph, we confirmed the validity of our assumption of the linearity of the systematic component of our model, as there does not appear to be a non-linear trend in the residuals. Further, the plot supports the assumption of homogeneity of variance, as the residuals appear to be uniform in magnitude across the range of fitted values. From the QQ plot as well, we confirm normalization of residuals holds as well.

We quickly do a parallel process with the alternative dataset as well. As per Table 10, we find an Adjusted $R^2$ of 0.3155 and an AIC score of 1720.464 as well. Meaning we find a slightly less efficient model, likely due to the noise caused by reskewing prior to normalization. As such, we will not consider this model further due to inferiority and not analyze its linear assumptions validity.

We did a baseline model with full interactions on the standard set as well finding an Adjusted $R^2$ of 0.3512 and an AIC score of 1871.702 as a reference for the full interaction mixed model scenario. We will examine this further later onward if needed, but given that there is only a 4% explained variance increase and a jump of 150 in AIC score when interactions are applied, this model individually does not pose much use.

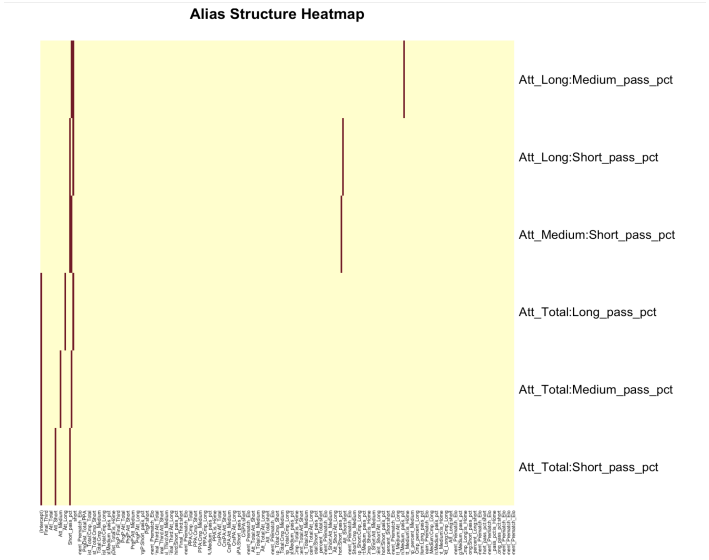


Figure 5: Alias Heat Map w/ Interaction Items

With the inclusion of interaction terms, we must check for multicollinearity among predictors to ensure model stability and interpretability. As per Figure 5, we devise heatmap using the alias structure, revealing 6 terms that are linearly dependent on others and may compromise model identifiably. Thus, while our baseline model is still sufficient as these are all interaction terms, allow us to remove these 6 interaction variables for future interactive models. Fortunately, the stepwise model and next 2 models handle multicollinearity on their own. We will do a likewise process for the alternative dataset, if the dataset starts to provide greater results.

### Stepwise Model Results

Pivoting from our baseline model, we sought to devise a step-wise baseline model considering all interactions as the upper bound and just the intercept as the lower bound as stated earlier. This will result in the optimal standard linear regression with our provided variables, when prioritizing AIC score. This is to find the optimal composition. With that, we find the following for our stepwise model.

- As per Table 11 and Figure 7, we find that our best case baseline regression to have an adjusted $R^2$ of 0.3411 and an AIC score 1679.8, while adhering to linear assumptions.
- Once more, as per table 12, we find a slightly lower adjusted $R^2$ of 0.3396 with a slightly higher AIC of 1681.533. This suggests that we created further noise when reskewing, which we will finalize later. Nonetheless, as its primary metrics are lacking as well, we will shy away from this model and stay with our stepwise model with the standard dataset as the best model at this point.

## Lasso & Ridge Models Results

Table 3: LASSO and Ridge Regression Results (Standard vs. Alternative Dataset)

| **Model** | **Dataset** | **Adjusted $R^2$** | **AIC** | **Best $\lambda$** |
|---|---|---|---|---|
| LASSO | Standard | 0.318 | -446.85 | 0.0305 |
| Ridge | Standard | -0.075 | 99.13 | 1.2909 |
| LASSO | Reskewed | 0.308 | -442.13 | 0.0369 |
| Ridge | Reskewed | -0.052 | 82.26 | 0.5601 |

From Table 3, we observe that the choice of $\lambda$ reflects each model's sensitivity to regularization. Ridge regression required higher $\lambda$ values but still resulted in negative adjusted $R^2$ scores, indicating poor model fit. This suggests that Ridge's uniform shrinkage was ineffective at filtering out irrelevant predictors, an expected outcome given Ridge's tendency to retain all variables. Since we included interaction terms across all predictors, this behavior does not necessarily point to flawed variable selection.

In contrast, Lasso achieved optimal sparsity at lower $\lambda$ levels, yielding positive and respectable adjusted $R^2$ values (0.318 for the standard dataset and 0.308 for the reskewed dataset). This demonstrates LASSO's strength in selecting informative variables and supports the quality of our overall predictor design. Thus, we maintain our variable selection.

Lasso also outperformed Ridge in both datasets across all evaluation metrics. This makes sense as we add a lot of unnecessary interaction items as predictors for these models. However, these models did ultimately result in the best AIC scores we've seen, at the expense of Adjusted $R^2$ to a certain degree. For the sake of simplicity and redundancy, we will only list the table and confirm linear assumptions of the best model of the batch, this ultimately being the Lasso regression, with an Adjusted $R^2$ of 0.318 and an AIC score of -446.85. We confirm this and linear validity with table 13, figure 8, and the same process we used on the baseline model.

### Elimination of the Alternative Dataset

Although the re-skewed models achieved a reasonably strong adjusted $R^2$s, it has been consistently performing slightly worse than its counterpart on the standard dataset. This suggests that the skew-reducing transformations, while theoretically intended to normalize distributions and enhance model interpretability, may have disrupted linear relationships or introduced unnecessary complexity. It is possible that the model's ability to detect signal from noise was better preserved in the original standardized feature space, where variable weights remained more aligned with their original structural relationships.

Given the consistent underperformance of models using the reskewed dataset and the absence of significant improvements in predictive power, we conclude that the reskewed feature set does not provide added value in this context. As such, we choose to pivot away from this dataset for the remainder of the analysis.

## Random Forest Model Results

Table 4: Random Forest Model Performance

| Metric | Value |
|---|---|
| Out-of-Bag Mean Squared Error (OOB MSE) | 0.5829 |
| $R^2$ | 0.2657 |
| Adjusted $R^2$ | 0.2418 |

As per table 4, we ultimately find that the random forest posted an OOB MSE of 0.5829 and an Adjusted $R^2$ of 0.2418, the worst of the all the models so far.

This suggests that the relationship between predictors and expected goals is largely linear and additive, limiting the advantage of a flexible, nonlinear model like Random Forest. The added complexity may have introduced unnecessary variance without capturing meaningful nonlinear structure. With such results, we find no need to analyze the prominence of each feature. This will be linked here if curious though (Figure 9)

## Mixed Model Results

For our mixed models, as per Table 5, we ultimately find these metrics for our results:

Table 5: Mixed Model Performance Across Specifications

| Model | AIC | Conditional $R^2$ |
|---|---|---|
| No Interactions (Baseline) | 1787.90 | 0.340 |
| Full Interactions | 1838.01 | 0.462 |
| Stepwise-Selected Variables | 3320.892 | 0.319 |

Ultimately, as shown in Table 14, the full interaction mixed model (Lack of 6 problematic interactive variables from earlier) emerges as our best-performing linear model to date, achieving a Conditional $R^2$ of 0.462 and an AIC of 1838.01. Additionally, Figure 10 confirms that this model satisfies key linear modeling assumptions.

This model explains approximately 11% more variance than our baseline full interaction model, while also posting a lower AIC—highlighting both improved explanatory power and a stronger model fit. Although the inclusion of all possible two-way interactions naturally increases model complexity, the AIC increase remains modest when compared to simpler models, suggesting that overfitting is being effectively managed and mitigated as a whole. Importantly, while models like LASSO and Ridge are explicitly designed for optimal variable selection and may outperform in AIC alone, the mixed model offers substantial gains in interpretability and variance explanation, especially when accounting for team-specific heterogeneity. Thus, of all the models we have devised, we found our best model.

## Linear-Variant Model Results

After evaluating four different linear-based models, we ultimately find that the full interaction mixed model performs best in explaining expected goals. Given its supe-

rior fit, we now focus our analysis on the statistically significant coefficients from this model.

By narrowing our interpretation to significant predictors, we reduce noise from irrelevant variables and better identify the most impactful relationships driving team performance. This allows us to draw more meaningful and reliable insights from the model's results.

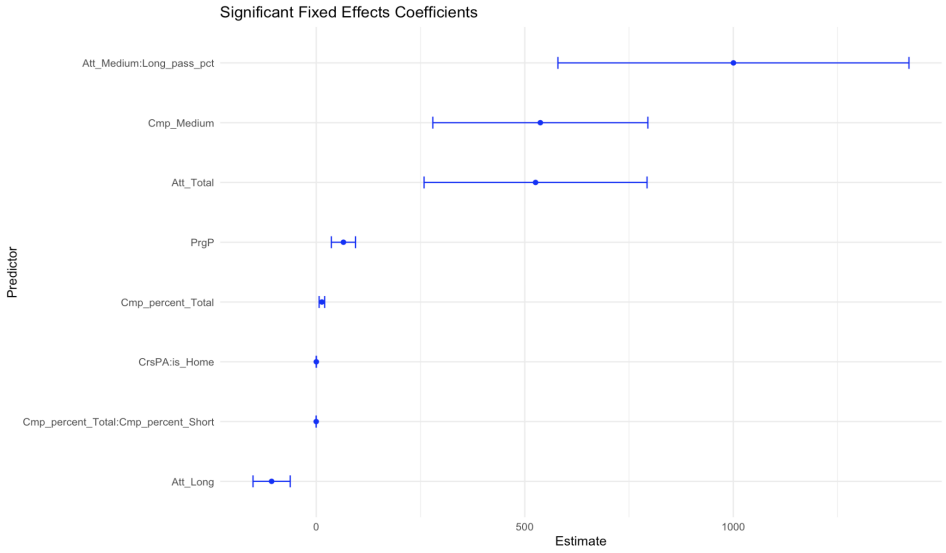


Figure 6: Significant Fixed Effects Coefficients w/ Standard Error

Based on the significant fixed effects coefficients visualized in Figure 6, we summarize key findings regarding team passing styles:

- **Long Pass Attempts (`Att_Long`)** show a statistically significant *negative* effect on expected goals. This suggests that excessive reliance on long balls reduces attacking efficiency, likely due to possession turnover and lower completion rates.
- **Completed Medium Passes (`Cmp_Medium`)** are positively and significantly associated with higher expected goals. This reinforces the value of medium-length passes in effective build-up play and controlled forward progression.
- While direct short passing metrics are not significant individually, the interaction between total pass completion and short pass completion (`Cmp_percent_Total : Cmp_percent_Short`) is significantly *negative*. Albeit, it is only ever so slightly negative. This may indicate small diminishing returns when teams over rely on conservative short-passing strategies without advancing play.
- The interaction between medium pass attempts and long pass share (`Att_Medium : Long_pass_pct`) is highly significant and extremely positive. This implies that long balls can be extremely effective when used sparingly and in tandem with structured medium-passing sequences, suggesting value in stylistic balance.

Overall, the results suggest that striking a balance between medium passes and a team's chosen style—whether short-ball or long-ball—is crucial. Completed medium

passes are strongly associated with increased Expected Goals, and their presence appears to mitigate the drawbacks of relying too heavily on either tactic in isolation.

Short-ball play enables teams to maintain possession and control, offering a safer, more passive approach. However, overuse can result in diminishing returns, as limited forward progression restricts attacking potential. When paired with medium passes, short-ball strategies benefit from added verticality while preserving structure and precision.

In contrast, long-ball strategies exhibit a high-risk, high-reward profile. Well-timed long passes, especially when supported by medium-passing sequences, can bypass defensive lines and generate quality scoring opportunities. Yet excessive reliance on long balls leads to a steep drop in efficiency, likely due to lower pass completion rates and reduced offensive cohesion.

## Baseline Logistic Model Results

Both pass-ratio coefficients are positive and highly significant ($p < 10^{-6}$), confirming that when the home team completes relatively more passes—whether overall or short—its win odds increase. The total-pass ratio (Model 1, slope = 0.772) has a slightly larger effect than the short-pass ratio (Model 2, slope = 0.651). Model 1 yields a lower AIC (1965.3 vs. 1976.0), indicating marginally better explanatory power using the total-pass ratio alone. Overall, this model suggests that winning probability and passes are positively correlated. However, further analysis is needed in order to understand whether there is causality from the effect.

Table 6 summarizes coefficient estimates, standard errors, statistical significance, and AIC for both models.

| **Term** | **Estimate** | **Std. Error** | **z value** | $\Pr(>\|z\|)$ | **AIC** |
|---|---|---|---|---|---|
| *Model 1: Total-Pass Ratio* | | | | | |
| Intercept | −0.6217 | 0.1073 | −5.796 | $6.8 \times 10^{-9}$ | 1965.3 |
| HomePass_to_AwayPass | 0.7724 | 0.0821 | 9.413 | $< 2 \times 10^{-16}$ | |
| *Model 2: Short-Pass Ratio* | | | | | |
| Intercept | −0.4915 | 0.0998 | −4.925 | $8.5 \times 10^{-7}$ | 1976.0 |
| HomePass_to_AwayPass_short | 0.6512 | 0.0735 | 8.855 | $< 2 \times 10^{-16}$ | |

Table 6: Baseline logistic-regression models predicting home-win probability from total- and short-pass ratios.

## Logistic With Subset of Teams Results

Table 7 shows the results from the baseline model when using only top teams. When restricting the teams analyzed, even the most basic assumption, that the percentage of short passes taken by the home team (with respect to the away team) and the number of short passes taken by the home team (with respect to the away team) are positively correlated with win probability - it is not statistically significant.

Table 7: Regression Results for Top Teams: Playing Style Effects

| | Model 3<br>% Short passes vs Total | Model 4<br># Short passes: Home vs Away |
|---|---|---|
| Intercept | −2.2077 (1.6033) | −0.5635 (0.4914) |
| HomePerc_to_AwayPerc_short | 2.4073 (1.5570) | — |
| HomePass_to_AwayPass_short | — | 0.6348 (0.3432) |
| AIC | 89.154 | 87.673 |

Note: Standard errors in parentheses.

# Conclusion and Discussion

## Conclusion

Our modeling framework explored both probabilistic match outcomes and performance-based metrics through logistic regression and Expected Goals (`xG`) linear modeling. Across these analyses, we sought to evaluate whether stylistic team choices—particularly regarding pass type and distribution—had tangible performance effects.

In our logistic regression results, we first established that both the total pass ratio and short pass ratio (home vs. away) are strongly predictive of win probability. Specifically, the total-pass ratio yielded a higher coefficient (0.772) and lower AIC (1965.3) than the short-pass ratio model, suggesting that general passing dominance is slightly more influential than just short passing. However, as expected, these results come with a clear caveat: passing behavior is also reactive. Teams in the lead often adopt lower-risk, short-passing strategies, and stronger teams naturally dominate possession. This makes it difficult to disentangle stylistic causality from game state or team quality effects.

To address this, we subsetted the data to only include matches between top-tier teams. Here, the predictive power of short-pass proportions on win probability disappeared. Neither the percentage nor the raw count of short passes (relative to the opponent) showed statistically significant effects. This indicates that once team strength is controlled for, tiki-taka–like passing in isolation does not reliably yield higher win odds. It also suggests that the earlier results may have been driven more by team quality than tactical style.

We then shifted to performance-based modeling using xG and a wide range of predictors, including newly engineered variables and team-level Elo ratings. Across several modeling approaches—linear regression, LASSO, Ridge, Random Forest, and mixed models—we found that no single passing style universally outperformed. Instead, the most effective attacking outcomes came from a balanced interplay between passing types, particularly when completed medium passes were involved.

Short-ball play appears to allow teams to control tempo and limit risk, but offers diminishing returns when overused. Long-ball strategies, when properly integrated with medium-passing build-up, show high-reward potential but also suffer from sharp declines in effectiveness when over-relied upon. These results highlight the nuanced tradeoffs between tactical aggression and control.

## Discussion

Although our best model (the full interaction mixed-effects model) explained roughly 46.2% of the variance in Expected Goals and offered the strongest linear fit to date, over half of the xG variation remains unexplained. Future work could benefit from incorporating defensive and positional data—such as pressure events, line height, and recovery metrics—to better capture the dynamics of goal-scoring opportunities. In this instance, we were limited to not including these metrics due to lack of public data provided.

Additionally, while our classification of team playstyles and strength (via $k$-means

and Elo) adds structure, it may still oversimplify in-game tactical fluidity. More dynamic modeling approaches such as Hidden Markov Models, RNNs, or sequence-based clustering may better capture how teams adjust their styles throughout a match. Finally, expanding the scope to include multiple seasons and leagues could improve generalizability and reduce the risk of dataset-specific conclusions.

## Final Thoughts

Taken together, at this point, our findings suggest that success in soccer is not dictated by one rigid passing ideology. Rather, it is shaped by context-aware decision-making that blends tactical preferences with adaptive progression, particularly through efficient medium-pass execution.

# Appendix

| Variable | Description |
|---|---|
| League | League match is from |
| Match Date | Match date |
| Match Week | Week of match |
| Home Team | Home team name |
| Home Formation | Formation used by home team |
| Home Goals | Name and time of goals scored by home team |
| Away Team | Away team name |
| Away Formation | Formation used by away team |
| Away Goals | Name and time of goals scored by away team |
| Team | Stats for the team of that game |
| Home Score | Score by home team |
| Expected Home Team Goals | Number of expected goals by home team |
| Home Team Yellow Cards | Yellow cards by home team |
| Home Team Red Cards | Red cards by home team |
| Total Completed Passes | Total passes completed by team |
| Total Passes Attempted | Total passes attempted by team |
| Total Pass Completion Rate | % of total passes completed by team |
| Short Passes Completed | # of short passes completed by team |
| Short Passes Attempted | # of short passes attempted by team |
| Short Pass Completion Rate | % of short passes completed by team |
| Medium Passes Completed | # of medium passes completed by team |
| Medium Passes Attempted | # of medium passes attempted by team |
| Medium Pass Completion Rate | % of medium passes completed by team |
| Long Passes Completed | # of long passes completed by team |
| Long Passes Attempted | # of long passes attempted by team |
| Long Pass Completion Rate | % of long passes completed by team |
| Total Distance Covered By All Pass | Total distance covered by all passes by team |
| Total Progressive Distance | Forward moving distance by team |
| Progressive Passes | # of passes moving ball min 10 yards forward |
| Assists | # of passes leading to a goal |
| Expected Assists | Expected assisted goals by team |
| Key Passes | # of passes that lead directly to a shot attempt |
| Passes Into The Final Third | # of passes into the final third of the field |
| Passes Into Penalty Area | # of passes into the penalty area |
| Crosses into penalty area | # of crosses into penalty area |

Table 8: Variables Included in the Soccer Match Dataset

Table 9: Baseline Linear Regression on Expected Goals w/ Standard Dataset

| **Variable** | **Coefficient** | **Std. Error** | **p-value** |
|---|---|---|---|
| Intercept | 6.50400 | 5.33183 | 0.2229 |
| Progressive Distance | 0.17010 | 0.07282 | 0.0198 |
| Progressive Passes | 0.06717 | 0.08145 | 0.4099 |
| Final Third Passes | -0.08832 | 0.06268 | 0.1593 |
| Progressive Passes Allowed | 0.36383 | 0.05493 | <0.001 |
| Crosses into Penalty Area | -0.02465 | 0.03510 | 0.4827 |
| Total Pass Attempts | -3.00101 | 2.13946 | 0.1611 |
| Total Passes Completed | 3.27480 | 1.98145 | 0.0988 |
| Total Completion % | -0.05202 | 0.05710 | 0.3626 |
| Short Pass Attempts | -0.97018 | 1.68764 | 0.5656 |
| Short Passes Completed | 0.76786 | 1.45090 | 0.5968 |
| Short Completion % | -0.02029 | 0.02946 | 0.4912 |
| Medium Pass Attempts | 0.92692 | 1.39746 | 0.5074 |
| Medium Passes Completed | -1.25191 | 1.18147 | 0.2897 |
| Medium Completion % | 0.01510 | 0.02791 | 0.5888 |
| Long Pass Attempts | 0.50552 | 0.27893 | 0.0703 |
| Long Passes Completed | -0.35100 | 0.23904 | 0.1424 |
| Long Completion % | 0.02439 | 0.01429 | 0.0883 |
| Short Pass Proportion | -0.65655 | 6.33179 | 0.9174 |
| Medium Pass Proportion | -0.21419 | 5.98767 | 0.9715 |
| Long Pass Proportion | -9.48695 | 6.64485 | 0.1538 |
| Short-Ball Indicator | -0.03704 | 0.09504 | 0.6968 |
| Home Indicator | 0.27617 | 0.05583 | <0.001 |
| Team Elo | 0.12528 | 0.03568 | 0.0005 |
| Opponent Elo | -0.11595 | 0.03268 | 0.0004 |

*Note:* $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.
Residual Std. Error = 0.7362; Adjusted $R^2 = 0.3179$; F-statistic = 15.74 on 24 and 735 DF; $p < 2.2 \times 10^{-16}$

Table 10: Baseline Linear Regression on Expected Goals w/ Alternative Dataset

| **Variable** | **Coefficient** | **Std. Error** | **p-value** |
|---|---|---|---|
| Intercept | 1.42229 | 0.06033 | <0.001 |
| Progressive Distance | 0.16313 | 0.07303 | 0.0258 |
| Progressive Passes | 0.11657 | 0.07820 | 0.1365 |
| Final Third Passes | -0.07927 | 0.06220 | 0.2029 |
| Progressive Passes Allowed | 0.32683 | 0.05080 | <0.001 |
| Crosses into Penalty Area | -0.00957 | 0.03449 | 0.7815 |
| Total Pass Attempts | -3.77372 | 1.54566 | 0.0149 |
| Total Passes Completed | 3.72332 | 1.46271 | 0.0111 |
| Total Completion % | -0.50960 | 0.36787 | 0.1664 |
| Short Pass Attempts | -0.80592 | 11.38001 | 0.9436 |
| Short Passes Completed | 1.15756 | 12.76655 | 0.9278 |
| Short Completion % | -0.17280 | 1.72420 | 0.9202 |
| Medium Pass Attempts | 1.30331 | 1.13869 | 0.2528 |
| Medium Passes Completed | -1.52139 | 1.03473 | 0.1419 |
| Medium Completion % | 0.16514 | 0.16483 | 0.3167 |
| Long Pass Attempts | 0.36829 | 0.30017 | 0.2202 |
| Long Passes Completed | -0.26225 | 0.26056 | 0.3145 |
| Long Completion % | 0.19248 | 0.16904 | 0.2552 |
| Short Pass Proportion | -0.16615 | 0.24077 | 0.4904 |
| Medium Pass Proportion | -0.03313 | 0.16432 | 0.8403 |
| Long Pass Proportion | -0.34362 | 0.30364 | 0.2581 |
| Short-Ball Indicator | -0.02160 | 0.09574 | 0.8216 |
| Home Indicator | 0.27620 | 0.05590 | <0.001 |
| Team Elo | 0.13752 | 0.03527 | <0.001 |
| Opponent Elo | -0.09806 | 0.03301 | 0.0031 |

*Note:* $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.
Residual Std. Error = 0.7375; Adjusted $R^2 = 0.3155$; F-statistic = 15.57 on 24 and 735 DF; $p < 2.2 \times 10^{-16}$

Table 11: Stepwise Linear Regression on Expected Goals w/ Standard Dataset

| **Variable** | **Coefficient** | **Std. Error** | **p-value** |
|---|---|---|---|
| Intercept | 1.48254 | 0.04258 | <0.001 |
| Progressive Passes Allowed (PPA) | 0.39108 | 0.04024 | <0.001 |
| Home Indicator | 0.27620 | 0.05429 | <0.001 |
| Team Elo | 0.12130 | 0.03391 | <0.001 |
| Opponent Elo | -0.07866 | 0.03286 | 0.0169 |
| Total Pass Attempts | -1.14118 | 0.25308 | <0.001 |
| Total Passes Completed | 0.92338 | 0.24717 | <0.001 |
| Progressive Distance | 0.13463 | 0.06581 | 0.0411 |
| PPA × Total Pass Attempts | -0.62772 | 0.23323 | 0.0073 |
| PPA × Total Passes Completed | 0.55684 | 0.23133 | 0.0163 |
| Home × Progressive Distance | 0.15438 | 0.05386 | 0.0043 |
| Team Elo × Opponent Elo | -0.07446 | 0.03133 | 0.0177 |
| Opponent Elo × Progressive Distance | 0.07498 | 0.03334 | 0.0248 |

*Note:* $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.
Residual Std. Error = 0.7235; Adjusted $R^2 = 0.3411$; F-statistic = 33.74 on 12 and 747 DF; $p < 2.2 \times 10^{-16}$

Table 12: Stepwise Linear Regression on Expected Goals w/ Alternative Dataset

| **Variable** | **Coefficient** | **Std. Error** | **p-value** |
|---|---|---|---|
| Intercept | 1.40665 | 0.04119 | <0.001 |
| Progressive Passes Allowed (PPA) | 0.35627 | 0.03562 | <0.001 |
| Team Elo | 0.13615 | 0.03365 | <0.001 |
| Home Indicator | 0.28392 | 0.05425 | <0.001 |
| Long Completion % | 0.03923 | 0.04415 | 0.3746 |
| Opponent Elo | -0.08048 | 0.03291 | 0.0147 |
| Total Pass Attempts | -1.05453 | 0.28700 | <0.001 |
| Total Passes Completed | 0.82901 | 0.28174 | 0.0034 |
| Progressive Distance | 0.12513 | 0.07275 | 0.0858 |
| Opponent Elo × Progressive Distance | 0.09608 | 0.03281 | 0.0035 |
| Home × Progressive Distance | 0.14499 | 0.05352 | 0.0069 |
| Team Elo × Opponent Elo | -0.07533 | 0.03112 | 0.0157 |
| PPA × Long Completion % | 0.07236 | 0.02697 | 0.0075 |

*Note:* $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.
Residual Std. Error = 0.7244; Adjusted $R^2 = 0.3396$; F-statistic = 33.52 on 12 and 747 DF; $p < 2.2 \times 10^{-16}$

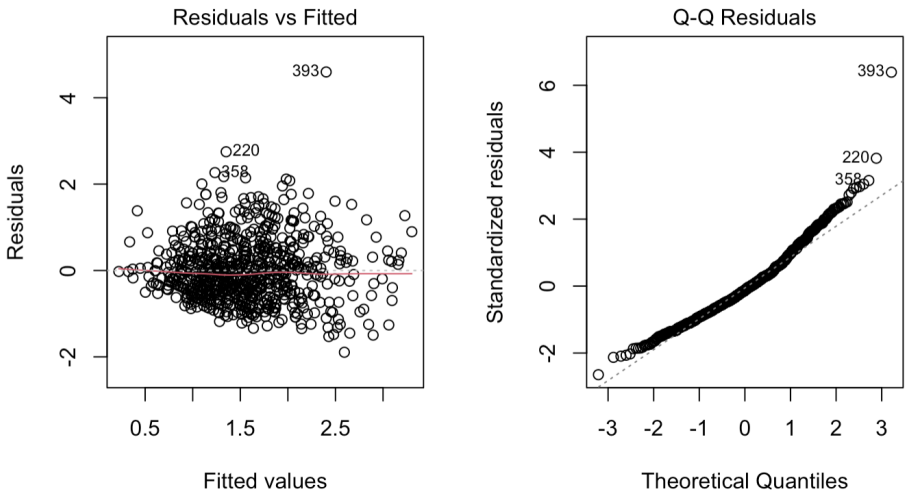


Figure 7: Stepwise Model on Standard Dataset Model Diagnostics

Table 13: Non-Zero LASSO Coefficients w/ Standard Dataset

| **Variable** | **Coefficient** |
|---|---|
| Intercept | 1.1049 |
| PPA | 0.1209 |
| Long Completion % (`Cmp_percent_Long`) | 0.0058 |
| Progressive Distance × Home Indicator | 0.0155 |
| Progressive Passes × Opponent Elo | 0.0577 |
| Final Third Passes × PPA | -0.0446 |
| PPA × Total Completion % | 0.0020 |
| PPA × Short Completion % | 0.0007 |
| PPA × Long Pass Proportion | 0.0001 |
| Total Completion % × Team Elo | 0.0002 |
| Short Completion % × Long Completion % | 0.00002 |
| Short Completion % × Team Elo | 0.0001 |
| Long Attempts × Opponent Elo | 0.0173 |
| Long Completion % × Team Elo | 0.0004 |
| Short Pass Proportion × Opponent Elo | -0.0229 |
| Medium Pass Proportion × Home Indicator | 0.5560 |
| Home Indicator × Opponent Elo | -0.0348 |
| Home Indicator × Team Elo | 0.0721 |
| Opponent Elo × Team Elo | -0.0099 |

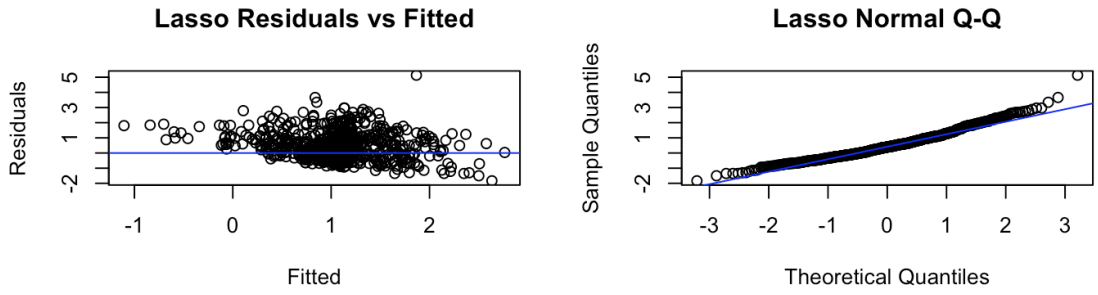


Figure 8: Lasso Model on Standard Dataset Model Diagnostics

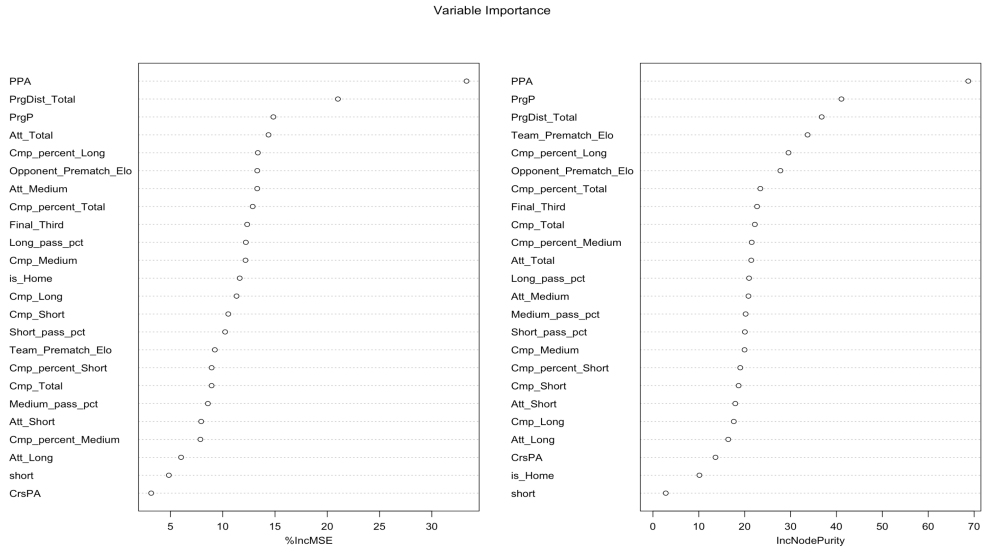


Figure 9: Random Forest Variable Importance

Table 14: Significant Fixed Effects from Mixed Model with Interactions

| **Variable** | **Estimate** | **Std. Error** | **p-value** |
|---|---|---|---|
| PrgP | 65.240 | 28.960 | 0.0247* |
| Att_Total | 525.900 | 267.300 | 0.0497* |
| Cmp_percent_Total | 13.590 | 6.648 | 0.0415* |
| Cmp_Medium | 537.400 | 257.800 | 0.0376* |
| Att_Long | -107.000 | 44.570 | 0.0168* |
| Cmp_percent_Medium | -8.236 | 4.421 | 0.0631. |
| PrgDist_Total:is_Home | 0.155 | 0.084 | 0.0565. |
| PrgP:Long_pass_pct | -68.170 | 35.400 | 0.0547. |
| CrsPA:is_Home | 0.207 | 0.0888 | 0.0200* |
| Att_Total:Cmp_Short | 128.500 | 73.400 | 0.0806. |
| Att_Total:Cmp_percent_Short | -4.633 | 2.603 | 0.0757. |
| Cmp_Total:Cmp_Short | -124.700 | 67.550 | 0.0654. |
| Cmp_percent_Total:Cmp_percent_Short | -0.0882 | 0.0423 | 0.0373* |
| Cmp_percent_Total:Cmp_Short | 5.316 | 3.179 | 0.0950. |
| Att_Long:Opponent_Prematch_Elo | 0.0173 | 0.0084 | 0.0397* |

*Note:* *p<0.05, .p<0.1

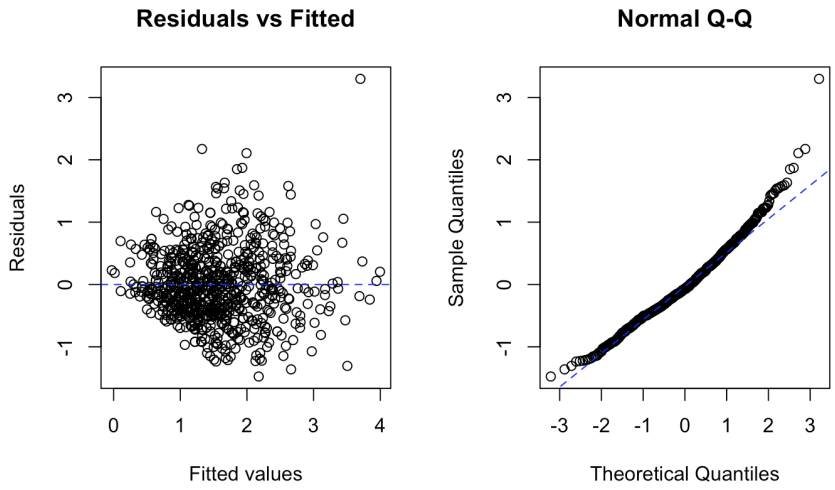


Figure 10: Mixel Model w/ Interactions on Standard Dataset Model Diagnostics