Predicting Premier League Match Results Based on First-Half Statistics

Omar Moustafa

900222400

Nour Kahky

900221042

DSCI 3415

Spring 2025

Introduction

The sport of football and its matches are heavily influenced by a long list of internal and external factors. However, because the game is divided into two halves, first-half statistics often provide strong indicators of both teams' performances and potential results. This study aims to investigate whether the events that occur in the first-half of a football game can be used to predict the final outcome. Specifically, can the collected statistics from the first-half be used to forecast if the second half will end with one of the teams winning or in a draw? Analyzing extensive data on the English Premier League, multiple machine learning methods will be employed to assess and evaluate the predictive capabilities of in-game features, including the number of attempted shots, completed passes, committed fouls, and duels, among others.

The objective of this study is to build accurate models and gain valuable insights on which in-game statistics and measures have the strongest influence in predicting the outcome of a football game. All in all, this investigation will portray the various aspects of a team's performance to understand how certain models deal with such an investigation along with the sport's most influential in-game factors.

Methods

To explain the relationship between first-half match statistics and final match outcomes thoroughly, the first step was merging multiple JSON datasets that contained match-level and event-level data. The information included, but was not limited to, team names, match dates and times, score lines, coaches, referees, and several in-game performance metrics such as shots, passes, duels, and fouls, to name a few. These metrics and variables were available for several national leagues and international tournaments. Therefore, it was decided to focus on only one competition, which led to the investigation of the fan-favorite English Premier League. Following this particular decision, a settlement on the research question was reached where the study would focus on first-half events within the Premier League to forecast how the second-half would end. Thus, second-half events were disregarded, and only events from the first halves were retained to align with the chosen research question: "**Can we predict whether the home team will win based on first-half events?**"

Then, relevant quantitative features and statistics were extracted for both teams, which were prefixed with team1_ or team2_ to distinguish them. Categorical variables, such as team names, were excluded from the model input was derived using the final score for each match and encoded as 0 for a team1 win, 1 for a team2 win, and 2 for a draw. To prepare the data for the upcoming machine learning methods, all numeric features were standardized using StandardScaler to ensure equal weighting and improve model accuracy and performance as much as possible.

Finally, the data was split into two sets based on an 80/20 proportion for training and testing to allow for the evaluation of the model's performance on new data. Following this, three classifier models were tested and compared. Those three classifier models were Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). The reasoning choosing these three particular modes to implement were for Logistic Regression to serve as a linear baseline, Random Forest to be the core model due to its flexibility and ability to model the more complex relationships between variables, and KNN was tuned using leave-one-out cross-validation where K = 5 to have a reasonable trade-off between bias and variance. Each one of these three models was run 10 times to account for the 80/20 splits, outputting the average accuracy of each model to ease and validate the comparisons between them.

Upon calculating each accuracy after 10 iterations of each model, the precision, recall, and F1-score were also calculated to thoroughly assess each model's productive capacity and capability. Lastly, the practical use of the model was illustrated by selecting random English Premier League matches from the test set and providing both the predicted results from the model and the actual match results, along with significant first-half statistics. This demonstrated how well the model could project real-world scenarios based solely on first-half event data.

Results

To summarize the results and performances of three classification methods—Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN)—to forecast Premier League game results based solely on first-half data was used to assess their effectiveness. The model with the highest average accuracy after 10 runs among the ones tested was Logistic Regression, which was closely followed by Random Forest and KNN, which saw a big drop-off in their success and accuracy compared to the two others.

Nevertheless, Random Forest was chosen as the final model because of its robustness in managing possible feature interactions and non-linear patterns in the data, as well as its balanced performance across all classes.

Logistic Regression Results:

Starting with the results of the Logistic Regression implementation, this model was evaluated on a test set, which included a total of 76 games of football. It achieved an overall accuracy of 99%, which is highly indicative of an extremely reliable prediction and performance. Even across 10 iterations, this model demonstrated great consistency with an average accuracy of 97.76%. All three classes—*team_1 win*, *team_2 win*, and *draw*—had extraordinarily high precision and recall values as well as F1-scores. Additionally, it can be noted that the model correctly identified all occurrences except for one, as indicated by the macro average F1-score of 0.99. Furthermore, this success is illustrated by the following confusion matrix (Figure 1), which also shows almost perfect agreement between expected and actual results.

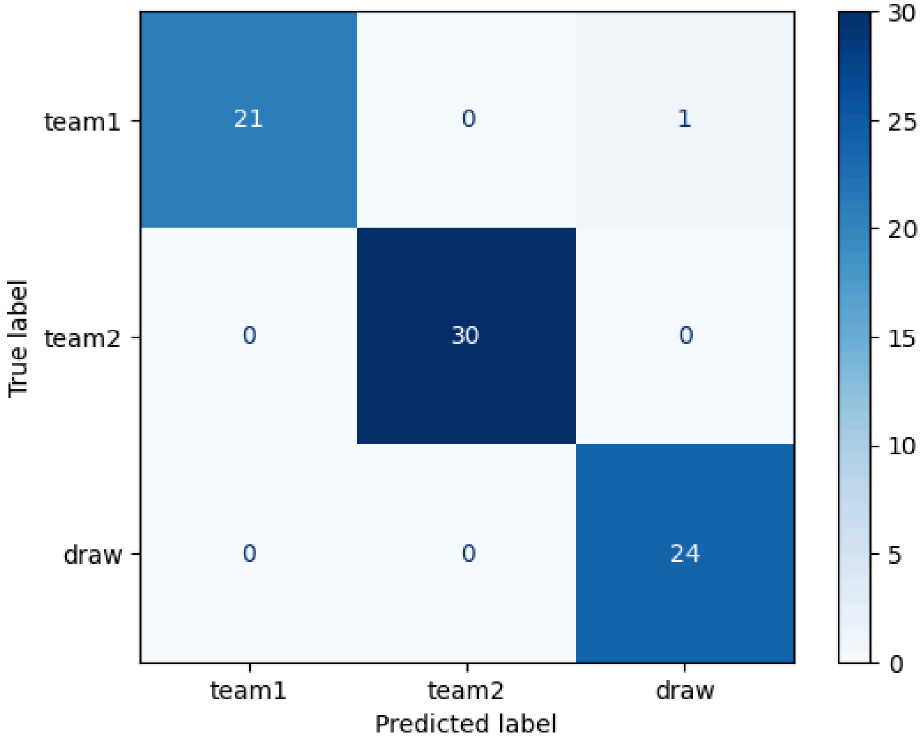

Figure 1

Random Forest Results:

Moving on to the Random Forest model, one simple run or instance had an accuracy of 87%, and over 10 runs, the average accuracy was equal to 83.68%. This model did well when it came to predicting wins for both teams, with an accuracy of 0.85 for team1 and team2 outcomes, and a recall of 1.00 and 0.97, respectively. Yet, it had a lower recall (0.62) for draws, indicating difficulty in accurately identifying these less frequent or more ambiguous outcomes. However, the model still had a precision of 0.94 for draws, yielding a balanced F1-score of 0.75. Feature importance analysis showed that the most salient single predictors were team2_score, team1_score, team2_Pass, and team1_Save attempt, suggesting the model's reliance on early scoring and possession-based measurements.

The confusion matrix (Figure 2), which is displayed directly below, shows that the most frequent misclassifications of draws were as wins for each team, suggesting overlapping patterns between draws and close wins. Overall, the Random Forest model as a whole produced a sufficient balance between predictiveness and interpretability in Premier League match data.

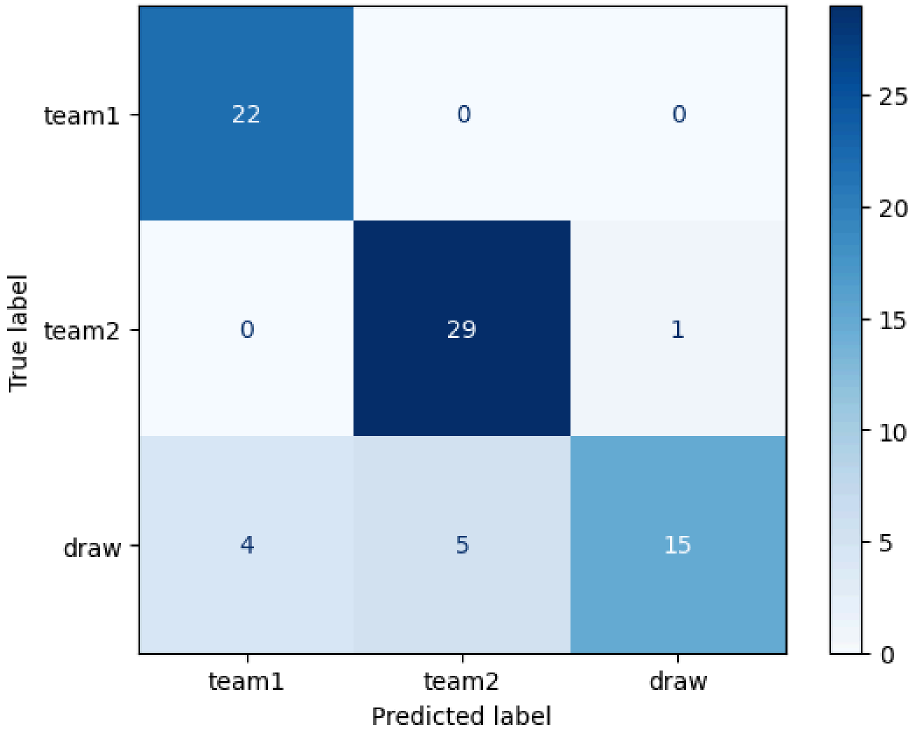

Figure 2

KNN Results:

The third and final classification model trained on this dataset was the K-Nearest Neighbors (KNN) model. KNN performed by far the worst among the three classifiers that were tested, as in a single test run, it achieved an overall accuracy rate of only 0.5,8, and over ten total runs, it achieved an average accuracy of 0.6211. To be fair, it did not do poorly when it came to predicting victories for both team1 (precision = 0.57, recall = 0.77) and team2 (precision = 0.61, recall = 0.83). However, it was inferior at predicting draws. The draw results recall dropped to a record-low 0.08, corresponding to an F1-score of 0.14, indicating that the model incorrectly classified nearly all of these cases. If anything, this model failed to correctly classify draws, as those that were correctly classified could be labelled as outliers when evaluating model accuracy. This is further supported by the following confusion matrix (Figure 3), where nearly all the true draws were misclassified as team1 or team2 wins.

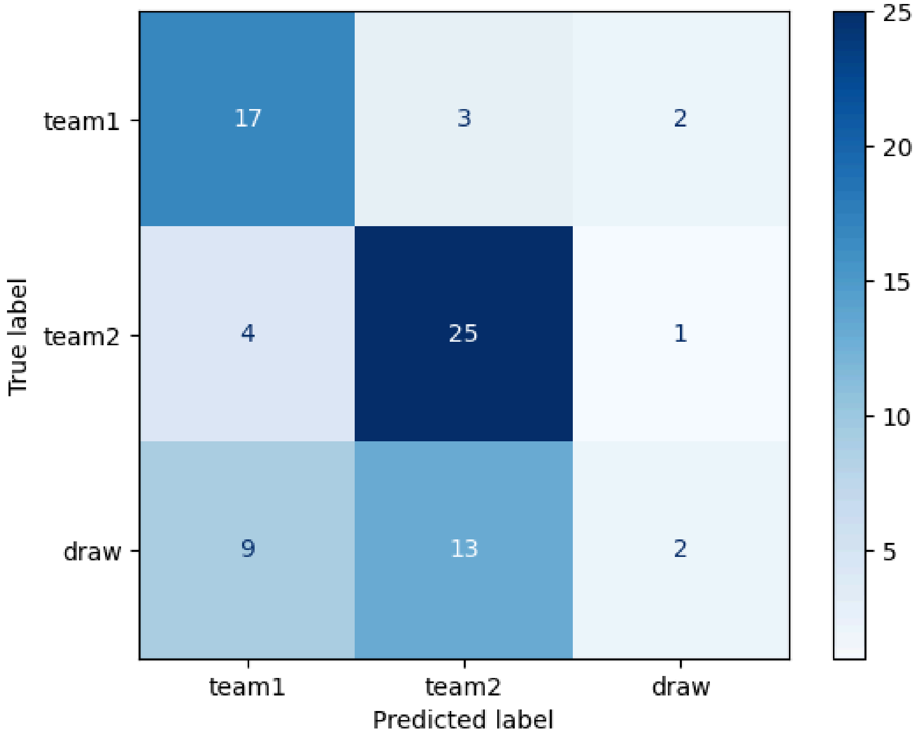

Figure 3

Overall, these particular results reveal that KNN could not differentiate between subtle patterns in the match data, especially for match results that are more uncommon or have similar but not identical statistics as team wins, namely, draws. This poor performance is likely due to the vulnerability to class imbalance often demonstrated by KNN, preventing it from accurately classifying this data and having a strong performance, relative to both the Logistic Regression and Random Forest models.

<u>Per-Team Accuracy:</u>

Following the training and repeated evaluation of the three different models previously described, assessing the model's performance on a per-team basis is a useful addition to the confusion matrix study. This indicates whether certain teams are relatively easy or hard to forecast, possibly due to their having consistent styles of play, more dominance in games, or even inconsistent patterns of performance and play. Therefore, the figure below displays the accuracy of the predictions made for all 20 teams in the Premier League, where those predictions, to reiterate, were made based on first-half data, offering a team-level view of model performance:

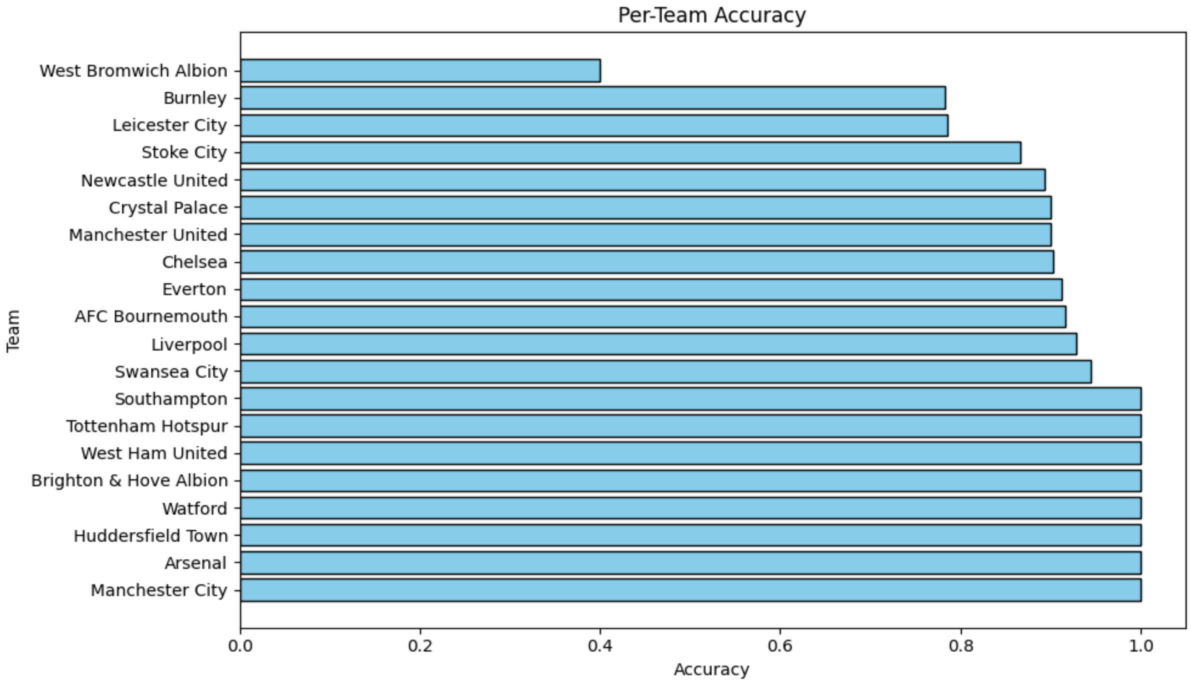

Figure 4

The plot above (Figure 4) shows that Manchester City, Arsenal, and Huddersfield Town are the most accurately predicted, along with a few others, all of which achieved perfect performances and predictions. This could be due to their statistical superiority across most of their games in a Premier League campaign, particularly in those games where they ultimately win. Looking at the other side of the plot, West Bromwich Albion has by far the worst accuracy, followed by Burnley and Leicester City, suggesting possibly a more unpredictable trend or more evenly matched first-half statistics that could be misleading the model, leading to this accuracy rate that is much lower than every other team in the league. Building on this, it can be inferred that those with the highest accuracy rates tend to play consistently well across both halves of the game, whereas those on the lower end of the spectrum could be playing well or alright in the first half but then turn off in the second half. This supports the interpretation that for teams such as West Brom, the model could have been misled, as some decent first-half statistics could have led the model to predict in their favor, which proved to be an inaccurate prediction in the end, and hence labelling those statistics as misdirection from the perspective of the model.

Sample Predictions vs Actual Results:

Finally, it came the time to put one of them to the test and make predictions on some real English Premier League games from the past. For this procedure, Random Forest was selected as the final predictive model.

Although the Logistic Regression model was the most accurate, Random Forest was chosen for several important reasons beyond just accuracy. For one thing, Random Forest offers greater interpretability of feature importance in the sense that a better understanding of the most essential first-half statistics will be gained. Secondly, it provides a suitable compromise between predictive performance and can handle linear and non-linear relationships, which is crucial in dynamic setups such as sports games. In addition, this particular model demonstrated solid class balance, a desirable trait when predicting less frequent outcomes, such as draws. Combining these reasons, Random Forest was deemed the most suitable model for practical use in forecasting football match outcomes based on first-half events.

Building on this decision, a small sample of past English Premier League games was utilized to test and evaluate the predictive capabilities of this Random Forest model. All three outcomes were correctly anticipated by the model, demonstrating strong reliability for such tasks. Below are the sample matches showing how the Model Prediction precisely matched the Actual Result, followed by brief analyses of how the first-half statistics aligned with outcomes:

Match 1: Brighton & Hove Albion vs Liverpool (May 13, 2018)

Model Prediction: Liverpool
Actual Result: Liverpool

First-Half Statistics:

- Brighton: Shots: 1, Fouls: 3, Passes: 122
- Liverpool: Shots: 15, Fouls: 1, Passes: 328

Analysis of Brighton vs Liverpool:

This first match demonstrates a typical footballing scenario where the pressure-forward attacking dominance pays a dividend for the winning team to come out victorious. Liverpool showed utter dominance in first-half shots (15 to 1) and the number of completed passes (328 to 122), which very likely led to the confident prediction made by the model. The very low foul tally also speaks to the control and discipline of Liverpool's back line, further adding to their profile. This implies that the model has figured out that future success is correlated with high attacking pressure and maintaining possession of the ball.

Match 2: Crystal Palace vs Manchester United (March 5, 2018)

Model Prediction: Manchester United
Actual Result: Manchester United

First-Half Statistics:

- Crystal Palace: Shots: 4, Fouls: 6, Passes: 126
- Manchester United: Shots: 2, Fouls: 2, Passes: 341

Analysis of Crystal Palace vs Man. United:

This second match appears to be more strategic, where having a successful all-around game is the difference-maker between the victor and the loser, rather than a straight all-out attack like the previous one. Although Manchester United had fewer shots (2 to 4), they completed over double the number of passes (341 to 126), indicating greater control in the midfield area, highlighting a potential scenario where they were more focused on creating build-up play with no rush, and potentially waiting for counter-attack opportunities. The fewer fouls committed by Manchester United also shows that they were defensively disciplined compared to the more aggressive Crystal Palace. This accurate forecast heavily demonstrates that the model can identify more general contextual patterns, including midfield dominance and strategic control, as indicators of eventual success. This is instead of just depending on the raw number of shots attempted and apparent attacking dominance from only one of the two teams.

Match 3: Watford vs Huddersfield Town (December 16, 2017)

Model Prediction: Huddersfield Town
Actual Result: Huddersfield Town

First-Half Statistics:

- Watford: Shots: 2, Fouls: 7, Passes: 180
- Huddersfield Town: Shots: 6, Fouls: 10, Passes: 154

Analysis of Watford vs Huddersfield Town:

This third and final match demonstrated that Huddersfield had a visible first-half advantage in shots numbers, a trait which the model should theoretically be able to use to stress as a proven indicator of coming out victorious. Although Huddersfield were less defensively disciplined and committed more fouls than Watford, they showed more offensive aspiration, tripling the number of attempted shots by Watford. Looking deeper into the first-half statistics of this particular game, Huddersfield Town totaled fewer completed passes than Watford. Since Huddersfield Town came out on top in the end, this would suggest that the number of completed passes is not the make-or-break factor and does not always translate into success in front of the goal. This particular case portrays the model's ability to trade off features according to context and interaction. It emphasizes how certain first-half statistical measures are more important than others when it comes to predicting how the second-half will end, such as shot volume outweighing pass completion, as many goals scored are known to be more correlated with attacking threats than midfield technicalities.

Overall Analysis of the Sample Predictions:

Collectively analyzing these sample predictions, they all ended up matching the real-game outcomes, strongly demonstrating that the Random Forest model can recognize intricate connections between first-half measures and game results. Its admirable trait of robustness can be recognized from the three fully correct predictions compared to the actual results across a variety of match styles, including sheer dominance by Liverpool, more strategic control by Manchester United, as well as more direct attack by Huddersfield Town, affirming its selection as the appropriate final model. These results further support the greater conclusion that first-half statistics can provide valuable insights and information about match course and trajectory when mined by a more sophisticated and recognized model, such as Random Forest.

Conclusion

In conclusion, this research examined the predictability of English Premier League match results from first-half statistics, primarily using a Random Forest classifier and modelling method. The model performed well across multiple evaluation metrics, as shown by the confusion matrices for all three classes—home win, away win, and draw—reflecting reliable identification of the more dominant or common outcomes in a game of English football, especially victories by the bigger teams in the division. Furthermore, the model's sample predictions for individual games turned out to be identical to or matched the real results, adding credibility to the perspective that the flow of the game in the first-half, specifically the number of shots attempted, fouls committed, and number of completed passes, provides a real idea of how the game will continue to go and eventually end.

The Per-Team Accuracy plot and analysis provided more details by identifying which teams the model predicted most and least accurately. Teams like Manchester City, Arsenal, Huddersfield Town, and a handful more had seemingly perfect prediction accuracies. This could reveal that their first-half trends and patterns are more consistent throughout their list of fixtures, which makes the predictions behind their results more plausible and reliable. Since there is little-to-no disparity in their first-half patterns, there is not much room for the prediction to go wrong. On the other hand, teams like West Brom, Burnley, and Leicester City had much lower team-level accuracies, indicating that their first-half performances are less consistent and could even be uncorrelated with their second-half performances due to certain external factors. This large discrepancy heavily emphasizes the genuine importance of including prior analysis of team-specific statistics before making predictions regarding overall outcomes.

Finally, the Random Forest model shows true promise for the use and implementation of first-half statistics and performance measures in the early predictions of final match results. If similar research and studies were to be conducted in the future, the accuracy could potentially be taken to higher levels by the inclusion of additional vital factors such as time-related, contextual, or player-level variables that help predict the trajectory and outcomes of games in the English Premier League.

References

Pappalardo et al., (2019) **A public data set of spatio-temporal match events in soccer competitions**, Nature Scientific Data 6:236, https://www.nature.com/articles/s41597-019-0247-7

Pappalardo et al. (2019) **PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach**. ACM Transactions on Intelligent Systems and Technologies (TIST) 10, 5, Article 59 (September 2019), 27 pages. DOI: https://doi.org/10.1145/3343172