BEAT THE BOOKIE

COMP0036 GROUP REPORT

Group C

Department of Computer Science
University College London
London, WC1E 6BT

December 17, 2024

1 Introduction

Predicting football match results is a challenging and widely studied problem due to the complexity and unpredictability of the sport. Leveraging historical match data, predictive models aim to determine outcomes—home win, draw, or away win—providing valuable insights into team performance and match dynamics. Despite advancements in machine learning, the current golden standard for accuracy remains modest at approximately 53%.

This report focuses on the English Premier League (EPL), one of the most competitive football leagues globally, using historical in-match data encompassing performance metrics and team trends. We propose a novel approach to improve predictive accuracy through innovations in feature engineering, feature selection, and model architectures. A voting ensemble of seven excelled models was developed, achieving baseline results near the golden standard and surpassing it with an upper accuracy exceeding 55%. By introducing 105 systematically engineered features and refining selection techniques, we offered a fresh perspective and significant contributions to football match prediction.

2 Data Transformation & Exploration

2.1 Data Loading & Cleaning

The provided `epl-training` dataset contains 9,221 historical English Premier League match records spanning the late 2000s to early 2024, covering 22 columns of team performance metrics, in-game statistics, and officiating details. After identifying and removing one row with missing values and resolving duplicate entries based on match day and team, the dataset was reduced to 9,120 unique records. The `Date` column was standardized to a consistent datetime format, and the dataset was sorted and re-indexed for temporal analysis and feature engineering. The `epl-test` dataset includes 10 matches scheduled in early 2025, containing only `Date`, `Home Team`, and `Away Team`. Team name discrepancies between training and test sets were resolved through mapping. The test dataset's limited features introduced challenges for model inference, as it lacked the rich attributes in the training set. Features for the test set should be constructed exclusively from prior match data to ensure alignment with training features, addressing compatibility and predictive consistency.

2.2 Exploratory Data Analysis (EDA)

The time-based analysis revealed that the distribution of matches is generally consistent, with no significant trends or biases observed in the dataset. While the initial assumption suggested that match results might be influenced by the date, deeper analysis demonstrated fairly uniform distributions across years, seasons, months, and days. Matches predominantly occur on weekends, with minimal records in summer due to a noticeable drop in match frequency during June and July, which only included records from later years. In contrast, match counts are higher in winter and spring; however, these seasonal variations do not show any clear patterns in full-time results. These insights suggest that time-related features are unlikely to provide meaningful contributions to the final predictive modeling and can therefore be excluded as key features.

The analysis of numerical features revealed that goals scored (`FTHG`, `FTAG`) exhibit skewed distributions, with the majority of matches featuring between 0 to 3 goals per team. Similarly, shots (`HS`, `AS`) and shots on target (`HST`, `AST`) show a clear correlation with match outcomes, particularly for home teams where higher shots on target consistently align with winning results. Features such as fouls (`HF`, `AF`), corners (`HC`, `AC`), and yellow/red cards (`HY`, `AY`, `HR`, `AR`) display consistent patterns without significant outliers. However, a notable trend emerged where home teams generally commit fewer fouls and receive fewer disciplinary cards compared to their away counterparts.

Previous studies indicate that the outcomes of football matches are correlative to the time [7]. Team-based analysis was carried out and identified that frequently-participated teams exhibit a higher probability of home victories, highlighting the influence of home advantage. While home wins are prevalent overall, the full-time result (FTR) distribution varies significantly between teams. Some strong teams primarily achieve home victories, while others display more balanced outcomes, with wins, draws, and losses evenly distributed. This observation indicates that incorporating team-specific features is essential for accurately predicting match results, as generalized trends fail to capture team-level variations. The analysis of match counts across teams reveals a significant range in participation, with most teams having played over 100 matches during the 24-year span. However, a subset of teams has participated in only 10–50 matches, creating an imbalance in team representation. This imbalance introduces potential challenges in predicting outcomes for teams with limited historical data, as models may struggle to generalize effectively for less frequently participating teams.

Contrary to initial assumptions, the analysis revealed that referees generally issue a consistent number of yellow and red cards across matches, with only occasional instances of more than three cards being issued. However, the distribution of matches officiated by referees is highly skewed; a small number of referees have overseen a disproportionately large number of games, while the majority have officiated only a limited few. This imbalance reduces the reliability of referee-based features, as referees with fewer matches provide less robust data. While the overall impact of referees on match outcomes appears minimal, their influence cannot be entirely dismissed. The absence of detailed information in the dataset limits further exploration into referee decisions, highlighting the need for additional data to better assess their precise role in match dynamics.

An in-depth comparison of half-time results (HTR) and full-time results (FTR) reveals that draws are notably more frequent at half-time. This trend underscores the significance of second-half performance in determining match outcomes. Interestingly, second-half performance changes indicate that over 60% of matches exhibit no improvement or decline, while home teams tend to improve their performance in the second half and achieve a home win at full-time.

In conclusion, time-based features, while informative, do not introduce significant patterns to match outcomes and can be excluded as predictive features. Performance-related metrics like shots on target and goals play a critical role in determining match results, while home advantage remains a significant factor.

2.3 Baseline Feature Engineering

The baseline feature engineering (Version 0) constructs features applicable to both training and testing datasets, addressing the test set's limitation of including only `Date`, `Home Team`, and `Away Team`. Features were derived exclusively from historical match data prior to each match date, ensuring predictive consistency. Using the 10 most recent matches for each team, features were weighted inversely by the time difference (in years) to prioritize recent performance. Missing historical data was initialized to zero, and mean year differences were added to capture temporal trends. This approach effectively leverages match history as a foundation for further feature engineering.

2.4 Global-Averaged Feature Engineering

This iteration leveraged the entire dataset to compute global averages and team-specific insights, enhancing predictive performance by incorporating a broader perspective on team behavior, performance consistency, and match dynamics. Newly engineered features included team experience, gap between matches, total matches, and performance metrics such as win rate, draw rate, and goal conversion rate Features captured precision, discipline, and fatigue. Historical patterns were explored with points per game, win/loss streaks, and Elo ratings, while referee-based features assessed biases. However, data leakage emerged due to reliance on future matches, misaligning training and testing features. This highlighted the critical need for stricter temporal constraints, ensuring test features are constructed exclusively from prior match data.

2.5 Exponentially Weighted Feature Engineering

This iteration refines the approach by constructing historical features using only data up to the current row's date, strictly avoiding future information to prevent data leakage. Time-based weights were applied using an exponentially weighted average (EWA) formula to prioritize recent matches:

$$\text{EWA}_t = \frac{\sum_{i=1}^{N} x_i \cdot \exp(-\lambda \cdot (t-i))}{\sum_{i=1}^{N} \exp(-\lambda \cdot (t-i))}$$

where x_i represents historical feature values, t is the current match date, i is the historical match date, and λ controls recency sensitivity.

All original features were replaced with descriptive, information-rich alternatives, including `participation days`, `total matches`, home/draw/away rates, shot on target metrics, goal conversion rates, and normalized disciplinary points. Temporal dynamics were introduced via `rest days` and win/loss streaks. Head-to-head features, such as the mode of `Full-Time Result (FTR)`, were included but showed limited effectiveness due to data sparsity, especially for less-participated teams.

Testing revealed that EWA reduced feature correlations and importance, introducing complexity without added value. While Version 2 aimed to enhance performance, the findings highlight the limitations of time-weighted methods in scenarios with sparse or imbalanced historical data.

2.6 Unweighted Time-Based Feature Engineering

In this version, feature engineering remained time-based but discarded the use of weighted averages, using uniform weights across all available past data. Recognizing the value of head-to-head data, new features were added to capture historical match dynamics between specific teams, providing valuable insights into past performance. The Recent Trends Index was introduced used rolling windows to compute combined Win/Draw/Loss Rates and average goals scored, reflecting short-term fluctuations in team form. The Recent Trends Index emerged as a key feature, highlighting the importance of recent performance. Fallback mechanisms addressed data sparsity issues by applying league averages as default values and using imputation techniques for missing metrics. Despite limited improvement in prediction accuracy, validation accuracy exceeded cross-validation results, suggesting broader historical data improves representation.

2.7 Performance Index Feature Engineering

In this Version, historical match data for English football was sourced from `www.football-data.co.uk`, expanding beyond the Premier League to include other English leagues. Only matches from 2000 onward were utilized to align with the column structure of the original `epl-training` dataset. After cleaning and combining over 50,000 records, the expanded dataset significantly increased match counts per team, with a minimum of 100 matches per team, making it more suitable for training. A total of 30 features were engineered and categorized into three dimensions: *Team Performance Index*, *Head-to-Head Statistics*, and *Recent Performance Index*. The *Team Performance Index* measured long-term success through win rates, goal scoring/conceding rates, and disciplinary behavior. *Head-to-Head Statistics* captured historical matchups between home and away teams, including win/draw ratios, average goals, and pressure indices. The *Recent Performance Index* focused on short-term trends such as recent win rates, shot accuracy, and goal conversion rates. Despite richer features and greater data volume, modeling and testing results declined compared to earlier versions. Applying the features to the original dataset yielded similar results, highlighting that the features were conceptually sound, but the expanded dataset introduced noise. Including more teams across leagues reduced head-to-head feature effectiveness and biased team performance metrics due to mixed competition levels. While larger datasets enhance generalizability, this version underscored the importance of data consistency and quality.

2.8 Data Enhanced Feature Engineering

This version builds on Version 4 by addressing prior limitations with targeted improvements, including league-wide features and refined dataset selection. League-wide features, derived from the most recent league season, captured trends such as average win rates, goals per match, defensive strength, and corner dominance to contextualize team performance within broader league dynamics. To reduce noise introduced by Version 4's expanded dataset, Version 5 filtered the historical data to include only teams present in the original `epl-training` dataset, improving alignment with the prediction task. While this adjustment increased match counts per team and maintained feature richness, the results showed only marginal gains. The features applied to both the refined and original datasets yielded similar accuracy, with league-wide features exhibiting weak correlations and limited predictive value. This version emphasizes that team-specific and match-level features are more impactful, and generalized league trends may add complexity without improving predictive performance.

2.9 Research-Inspired Feature Engineering

The final feature set introduces a well-structured, comprehensive framework of 105 engineered features derived from extensive experimentation and rigorous refinement. Each category was informed by proven formulas and insights from relevant research papers to ensure robustness and alignment with best practices[13, 12, 16, 17]. Team-Based Winning Features focus on overall and environment-specific performance metrics, such as win percentages, emphasizing the predictive importance of home and away trends. Head-to-Head (H2H) Features capture historical dynamics between teams, including dominance, corner opportunities, and most frequent outcomes (*FTR Mode*), offering valuable insights for closely contested games. Last Season League Points Features provide long-term performance indicators, reflecting teams' prior season success. Result Streak-Based Featuresanalyze team form dynamics, including short-term momentum and weighted long-term changes, to capture shifts in performance. Recent k Matches' Performance Features evaluate offensive and defensive strengths, incorporating metrics like *Shot Accuracy Difference* and *Goal Conversion Rate Difference* to assess efficiency. While research identified $k = 20$ as optimal, our experiments found $k = 10$ more effective due to dataset-specific differences. Team Strength Features measure resilience and discipline, while *Expected Goal (xG) Features* leverage advanced analytics to quantify scoring probabilities and match-level strength. Metrics like *In-Match Goals Difference* provide refined predictions of scoring potential. HTR-Related Features analyze half-time-to-full-time (HTR-FTR) transitions, capturing second-half dynamics and team dominance under varying conditions. Form-Based Features dynamically track incremental changes in team performance and compare relative strengths using *Form* and *Form Differential*. Yearly-Weighted Recent Values reintroduce temporal relevance for goals, shots, fouls, and cards, smoothing historical trends while prioritizing recency. FootStat Predictive Features (e.g., *Clean Sheets Percentage, Both Teams to Score Percentage, Last 5 Weighted Results*) adapt methodologies from the FootyStats website (https://footystats.org/england/premier-league), recalculating these metrics to enhance prediction granularity while

ensuring consistency with the dataset. This meticulously crafted feature set balances historical, recent, and contextual factors, providing a comprehensive and effective foundation for predicting football match outcomes.

2.10 Feature Selection

For feature selection, various approaches were employed to optimize the feature set for robust predictive performance across multiple models. Initially, traditional techniques referenced in related research were explored, such as Variance Inflation Factor (VIF), model-based and sequential feature selection, and the Chi-square test [5, 9]. While these methods demonstrated utility in some scenarios, their application was less effective in capturing the diverse feature importance across different model types, as they often emphasize a specific subset of features rather than accommodating variations in model focus. Consequently, a practical approach was devised to identify the most relevant features. Using the `sklearn` library, permutation importance scores were computed for each experimented model, and the five least important features for each model were identified. These features were aggregated across models by counting their frequency of occurrence, forming a "least important" feature set. Ablation studies were conducted to analyze the impact of removing, retaining, or selectively keeping subsets of these features on accuracy and precision scores. Iterative experiments integrated these findings, refining the feature set until the most satisfactory results were achieved. This practical method proved superior, consistently yielding the best results on average and highlighting that novel techniques from research papers may lack generalizability across diverse model architectures. This underscores the importance of adapting feature selection methods to align with the specific requirements of the prediction task.

3 Methodology Overview

3.1 Background Research

Previous research on predicting sports outcomes, particularly in football, has seen a significant evolution with the integration of machine learning (ML) models to achieve precise forecasts. Studies have explored a variety of ML techniques to predict match outcomes, focusing on leagues such as the English Premier League (EPL). The methodologies employed in these studies range from traditional statistical approaches to advanced machine learning and deep learning techniques.

Rodrigues and Pintob proposed machine learning (ML) methods that integrate match statistics and player attributes to predict football match outcomes. Their approach utilized data mining to analyze historical match data and evaluated various ML models for their effectiveness in achieving profitable betting margins. [14]

Baboota and Kaur explored key factors influencing English Premier League match results through feature engineering and data analysis. They employed ML techniques such as Gaussian naive Bayes, SVM, and gradient boosting, comparing model performance against industry benchmarks like Bet365 and Pinnacle Sports using the ranked probability score (RPS). Their gradient boosting model achieved an RPS of 0.2156 over two seasons (2014–2016), slightly underperforming the betting organizations' RPS of 0.2012.[1]

Herold et al. reviewed ML applications in football, with a focus on attacking play. Their work analyzed supervised and unsupervised learning methods, addressing practical applications, current challenges, and potential improvements to enhance tactical understanding and player performance.[6]

Stübinger, Mangold, and Knoll investigated the use of ML in football betting through a simulation study. By combining multiple algorithms into an ensemble strategy, they analyzed matches from top European leagues (2006–2018). Their ensemble approach delivered a statistically and economically significant return of 1.58% per match, outperforming traditional betting strategies like always betting on the home team. [18]

Joseph, Fenton, and Neil compared an expert-designed Bayesian network with ML models, including decision trees, naive Bayes, and K-nearest neighbors, using match data from Tottenham Hotspur. The Bayesian network, based on expert judgment, demonstrated superior predictive accuracy compared to the alternative models. [8]

The background research highlights the diverse approaches and methodologies applied in predicting football match outcomes using ML. By analyzing historical data, leveraging advanced algorithms, and benchmarking performance against industry standards, these studies provide valuable insights and frameworks that inform the development of more accurate and profitable predictive models for football analytics.

3.2 Approaches

3.2.1 Baseline Methods

We chose a diverse set of classification algorithms, each representing a distinct family of machine learning methodologies. Specifically, we included:

- **Linear and Probabilistic Models:** SVM, Logistic Regression, and Naive Bayes. These models are fast, relatively simple, and tend to perform well on structured data, serving as fundamental baselines.
- **Tree-Based Models:** Decision tree and variants (Random Forest and Extra Trees) were used due to the ability to capture nonlinear relationships and interactions between features.
- **Boosting Algorithms:** AdaBoost, XGBoost, LightGBM, CatBoost, and Gradient Boosting. These methods are supposed to be trained for good performance.

By incorporating both simpler algorithms (e.g., Logistic Regression, Naive Bayes) and more advanced ensemble and boosting models, we ensure that our baseline reflects a comprehensive performance landscape. This variety provides a clear understanding of whether the problem at hand is more amenable to straightforward methods or benefits substantially from more complex modeling techniques.

3.2.2 New Methods

Except for the baseline methods, we employed new methods to train the models to explore better performance. We have tried multiple methods to train the models, some of them will be presented in detail afterwards.

Deep Learning and Advanced Architectures: Models such as Multi-Layer Perceptrons (MLPs), TabNet, 1D-CNNs, and attention-based CNNs were integrated. Unlike linear or tree-based methods, these architectures can learn intricate feature representations and adapt to more nuanced, high-dimensional inputs.

Advanced Hyperparameter Optimization: Compared to the baseline approach of using simple grid searches, the new methods benefited from better tuning frameworks (e.g., Optuna). This allowed for more efficient and thorough exploration of model-specific parameter spaces, including network architectures, activation functions, attention units, and regularization strategies.

3.3 Ablation Studies

We conducted a feature ablation study to understand the contribution of individual features to predicting full-time results (FTR) of football matches. Models were trained with each feature removed, and performance was evaluated using k-fold cross-validation.

The analysis revealed three key findings. First, the removal of Year had little impact on performance, and in some cases, like Logistic Regression, performance slightly improved, suggesting it may not be a crucial feature. Removing HomeTeamEncoded and AwayTeamEncoded caused only slight reductions in performance for ensemble models like Random Forest and ExtraTrees. Second, some features, such as Year and AwayTeamEncoded, seemed to act as noise, with model performance improving when these features were excluded. Finally, different features contributed differently across models, with HomeTeamEncoded and AwayTeamEncoded being more critical for ensemble models, while Year had minimal effect.

Additional studies on newly engineered features showed that feature removal had negligible effects on various performance metrics, including accuracy, precision, recall, and F1 score. This pattern held true across a range of models, from simpler ones like KNN and Logistic Regression to more complex ones like Naive Bayes and Decision Trees. Cross-validation and holdout results were consistent, further supporting the conclusion that removing these features did not significantly harm model performance.

3.4 External Data

For this project, we initially explored using data from fbref.com, which offers additional features like squad possession rates. However, this data was incomplete before 2014. We also considered incorporating squad and player data from fifaindex.com, but decided against it as it didn't align well with our primary dataset. As a result, we chose to focus on more consistent sources.

We then incorporated additional data from football-data.co.uk, which provided up-to-date match results, team statistics, and other relevant features to enhance the model's predictive ability. We considered supplementing our dataset with second-half 2024 data to better capture recent trends in team performance. However, after conducting validation experiments, we found that including this data worsened the validation results, as the data distribution became less representative of the early 2025 prediction period. As a result, we decided to exclude the second-half 2024 data from our final dataset.

4 Model Training & Validation

The dataset was split chronologically into training (up to 2022) and validation sets. Features and labels were separated, and columns risking data leakage were removed. All features underwent min-max scaling, using parameters fitted on the training set and then applied to the validation set. Hyperparameter tuning was also introduced to the training process, for traditional models (e.g., Logistic Regression, Decision Tree), an exhaustive grid search with five-fold stratified cross-validation was conducted. For more complex models (MLP, TabNet, 1D-CNN, attention-based CNN, ANN), Optuna was employed to efficiently explore their hyperparameter spaces.
Model validation was performed using a multi-stage approach to ensure robust and unbiased performance assessments. First, the data was split chronologically, reserving later periods as a validation set to simulate future, unseen conditions. In detail, four key metrics—accuracy, precision, recall, and F1 score were used to measure the performance of models.

4.1 Model Training

The dataset was subjected to a preprocessing phase with the objective of ensuring its compatibility with the selected machine learning algorithms. The data set was initially divided into two subsets: a training set comprising data from years up to 2022 and a validation set comprising data from subsequent years. The features and labels were separated, with any columns that could potentially lead to data leakage excluded. Subsequently, the features were standardized using a min-max scaling method, with the scaler calibrated on the training set and subsequently applied to the validation data set in order to ensure consistency. This step ensured that all features were on an equivalent scale, thereby enhancing the performance of the models.

4.1.1 Traditional ML Model

In order to optimize the performance of the model, hyperparameter tuning was implemented. A grid search strategy was employed in conjunction with five-fold stratified cross-validation, with the objective of maintaining balanced class distributions across the folds. In order to effectively address class imbalances, a weighted F1-score was selected as the scoring metric. For each model, an exhaustive search was conducted over predefined hyperparameter grids, and the configuration exhibiting the optimal performance was selected based on the mean cross-validated F1-score.

4.1.2 Deep Learning Model

In the case of more complex models, an advanced hyperparameter optimization framework was implemented, utiliszing the capabilities of Optuna. This framework facilitated the efficient exploration of hyperparameter spaces for a range of machine learning models, including MLP, TabNet, 1D-CNN, attention-based CNN, and ANN. The framework's key features includes the following:

GPU Utilization Dynamically detected GPUs were configured for memory growth, and trials were distributed across GPUs with the objective of optimizing the utilization of computational resources.

Model-Specific Optimization Each model's hyperparameter space was tailored to its architecture. Parameters such as hidden layer sizes, activation functions, attention units, and regularization factors were systematically explored.

Parallelized Optimization Multiple trials were executed in parallel, leveraging Optuna's study capabilities to maximize validation accuracy while reducing computation time.

4.1.3 Combined Model

In recent years, ensemble techniques have played a pivotal role in the domains of data mining and machine learning, as they can augment the precision of individual classifiers [11]. Model selection is fundamentally based on the complementarity of the confusion matrix, utilising tuned parameters, with the objective of accurately predicting the draw. This study employs three distinct methodologies: voting, stacking, and the Two-Step Model/Cascade Classifier.

The majority voting method is an effective approach to enhance classifier performance by aggregating predictions through a voting mechanism. This method is employed to determine the final prediction results for each comparative algorithm [15]. In this study, a voting mechanism is utilised to combine predictions from different models, employing either soft voting (weighted by probabilities) or hard voting (based on predicted labels).

Among ensemble techniques, stacking has emerged as a particularly popular method, often proving highly effective in solving complex problems [2]. In this work, multiple base models are used, with predictions from the base models being fed into a meta-learner for higher-level decision making.

In order to address the limitations of single-classifier detection methods, researchers have developed cascade classifiers, which connect multiple ordinary classifiers either in series or in parallel. A cascade classifier is a sequential chain of multiple weak classifiers, which employ a step-by-step refinement approach. In this process, the object under evaluation passes through each classifier in turn. Non-targets are typically eliminated in the early stages, allowing the remaining classifiers to focus more effectively on the suspected targets [19].

4.2 Model Validation

The following performance metrics were employed: accuracy, precision, recall, confusion matrix, ignorance score and ranked probability score (RPS).

The accuracy of a model is determined by calculating the proportion of correct predictions out of the total number of predictions made.

Precision is defined as the proportion of true positives out of all predicted positives and is used to assess the accuracy of positive predictions. This is especially the case in scenarios where the reduction of false positives is a priority. A high level of precision is indicative of reliability in a model when it makes a positive claim about a sample.

The recall value represents the degree to which the model is capable of accurately identifying all pertinent instances of a particular class. It is defined as the ratio of true positives to the total number of actual positives.

Confusion matrix provides a detailed breakdown of a model's performance by showing predictions across all classes. It includes counts of true positives, false positives, false negatives and true negatives for each class. Each row represents the actual class, and each column represents the predicted class. Diagonal elements correspond to true positives, while off-diagonal elements represent errors.

The Ranked Probability Score (RPS) is designed to account for the ordinal nature of the three possible outcomes in a soccer match [3][4]. The RPS value is always within the interval [0, 1], with a higher value representing a more accurate prediction. The RPS is calculated using the following formula:

$$RPS = \frac{1}{r-1} \sum_{i=1}^{r-1} \left(\sum_{j=1}^{i} (p_j - a_j) \right)^2$$

where r denotes the number of potential match outcomes, for example if there are three possible outcomes (home win, draw, away win) then r is equal to three. The RPS can be averaged across N instances in the following manner:

$$\text{RPS}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\text{RPS}_i$$

The Ignorance score (IGN) was initially proposed by Good [10]. The IGN is a penalizing metric that is applied to predictions with larger logarithmic errors and is calculated using the following formula:

$$IGN = \frac{1}{N}\sum_{n=1}^{N}\left[-\left(y\log_2(p) + (1-y)\log_2(1-p)\right)\right]$$

where $y \in \{0, 1\}$ and $p = P(y = 1)$. The value is positioned within a range $[0, \infty]$, with a lower score indicating higher levels of model performance.

4.3 Model Performance Analysis

Accuracy and Precision: While logistic regression achieved the highest raw accuracy, the ensemble methods demonstrated more consistent performance across both metrics. The voting_sr model achieved the highest precision, though this came with some trade-offs in other metrics.

Recall and Class Balance: Analysis of confusion matrices revealed that ensemble methods generally handled class imbalances better than individual models. Deep learning approaches showed more balanced recall across classes, though this sometimes came at the cost of increased false positives.

Probabilistic Performance: Traditional models generally demonstrated superior probability estimates with lower IGN scores, while the two-step methods showed higher variability in this metric. The RPS analysis revealed that logistic regression and voting methods provided the most reliable probability estimates across all classes.

To summarize, for balanced performance across all metrics, the voting_lgbms and 2step_ln models stand out as strong choices. When prioritizing specific metrics, logistic regression excels in accuracy while voting_sr leads in precision. Resource considerations should guide the final choice, as traditional models remain competitive while requiring significantly less computational power than ensemble methods.

5 Results

5.1 Comparison and Model Selection

In order to identify robust predictive models, we evaluated an extensive set of algorithms and ensembles, including traditional machine learning methods (Logistic Regression, Decision Tree, Random Forest, Extra Trees, Naive Bayes, SVM), gradient-boosted frameworks (XGBoost, LightGBM, CatBoost, Gradient Boosting), deep learning approaches (MLP, 1D_CNN, CNN_Attention, ANN, TabNet), and various ensemble strategies (Stacking, Voting, and Custom Two-step models).

The evaluation criteria included multiple metrics such as Accuracy, Precision, Recall, F1-score, Ignorance scores, and Ranked Probability Score (RPS). We also reviewed confusion matrices to understand each model's misclassification patterns, especially regarding draws, which are more challenging to predict due to their inherently uncertain nature. The results of the assessment are presented in the accompanying figures 1 2 3.

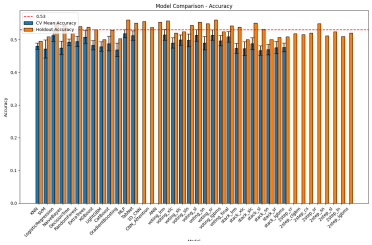

Figure 1: Accuracy

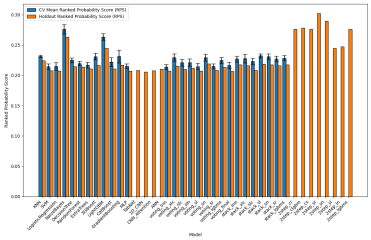

Figure 2: Ranked Probability Score(RPS)

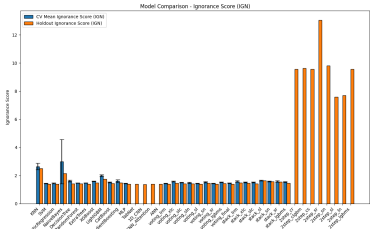

Figure 3: Ignorance scores

5.1.1 Key Observations

Logistic Regression (LR) emerged as a strong baseline with high accuracy and balanced metric performance. However, LR displayed a tendency to struggle with nuanced predictions for draws, often pushing towards more polarized outcomes (win/loss).

To improve performance, especially for predicting draws, six supplementary models were introduced. These models were chosen based on their strong cross-validation

performance and ability to handle class imbalances more effectively:

Random Forest and Extra Trees stood out for their stable and balanced performance across accuracy and precision metrics. These models handled class imbalances well, with Random Forest proving particularly effective in managing the nuanced classification of all three classes.

XGBoost and LightGBM both excelled in capturing complex relationships within the data. While LightGBM showed competitive accuracy, XGBoost demonstrated more balanced metrics and strong RPS scores, making it a strong contender for classifying all three outcomes.

The 2step_cr model provided a novel, reasoning-based approach to classification. Although it did not necessarily lead in raw performance, its unique method of considering multi-step reasoning allowed for stable probability estimates, particularly in challenging classifications.

The stack_sl ensemble effectively combined predictions from multiple models, producing the most consistent and robust results across all metrics, particularly in predicting the often unpredictable draw class.

5.1.2 Selected Models and Their Performance

After a thorough review of both the quantitative metrics and qualitative patterns from the confusion matrices, we selected the following seven models for our final predictive framework: *Logistic Regression (LR), Random Forest, Extra Trees, XGBoost, LightGBM, 2step_cr (Two-step reasoning model), and stack_sl (Stacking ensemble).*

Tables 1 and 2 summarize the performance of the selected models, showing cross-validation (CV) and holdout results across key metrics. The evaluation criteria included multiple metrics such as Accuracy, Precision, Recall, F1-score, Ignorance scores, and Ranked Probability Score (RPS).

5.1.3 Interpretation

While LR offered strong baseline performance, the inclusion of these other models delivered a more balanced perspective, particularly for the draw class. Models like Extra Trees and LightGBM demonstrated more stable behavior, as seen from their confusion matrices and slightly improved handling of less frequent outcomes. The 2step_cr model introduced a reasoning-based approach that, although not necessarily the top in raw metrics, provided a novel angle on classification decisions. The stack_sl ensemble method combined multiple approaches to yield more robust and stable predictions than any single model on its own.

By examining the performance from multiple angles—CV metrics, holdout results, confusion matrices, and advanced scoring methods—this multi-model strategy ensured a more comprehensive and resilient predictive framework. Ultimately, the chosen set of seven models offers a strong, well-rounded ensemble that better captures the complexities of the three-class prediction problem, including the challenging draw scenario.

6 Final Predictions on Test Set

The final predictions were made using a voting ensemble of seven models, including Logistic Regression, Random Forest, Extra Trees, and XGBoost. Each model contributed its prediction, and the ensemble aggregated these results to determine the most likely outcome for each match. This approach enhanced prediction accuracy and consistency, particularly for challenging outcomes like draws. The results are shown in Table 3. The precise details of the voting process are illustrated in the figure 4.

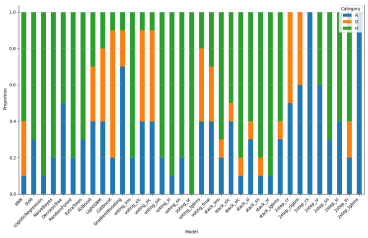

Figure 4: Voting proportion for each model

7 Conclusion

7.1 Summary of Findings

The final model, a voting ensemble combining seven top-performing classifiers, achieved an accuracy of 55.87%, surpassing the 53% benchmark typically seen in football match result predictions. This ensemble approach effectively balanced precision and recall, making it particularly strong in predicting difficult outcomes such as draws. While Logistic Regression contributed the highest individual accuracy, the ensemble model benefited from the combined strengths of multiple classifiers, resulting in more reliable and consistent predictions. Overall, the final model demonstrated significant improvements in prediction accuracy, providing a robust solution for forecasting football match outcomes.

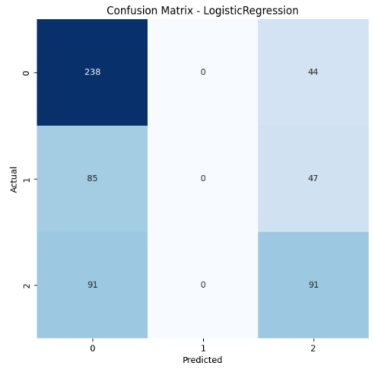

Figure 5: Confusion Matrix for Logistic Regression

7.2 Future Work

As illustrated in the figure 5, while logistic regression exhibits high accuracy, this model demonstrates a notable lack of capacity to generate draws in its predictions. This suggests that the model's inductive bias is particularly strong. Further research should be conducted to enhance the model's capability to generalize across a broader range of datasets. This may be achieved by modifying the feature extraction metrics and scoring algorithm to ensure the model's robustness in the context of diverse football datasets. Furthermore, future work should investigate the impact of different feature engineering methods on data loss and the potential for enhancing model performance by retaining more information. Additionally, future work should examine the relationship between the characteristics of the dataset, such as its size, and the impact of different model combinations. However, given the computational expense and time requirements associated with using larger datasets and training more models, it is also necessary to consider GPU methods that can accelerate this process for traditional ML algorithms.

References

[1] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35(2):741–755, 2019.

[2] Jason Brownlee. Stacking ensemble machine learning with python. `https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/`, 2024. Accessed: June 18, 2024.

[3] Anthony Costa Constantinou and Norman Elliott Fenton. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8(1), 2012.

[4] Edward S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology*, 8(6):985 – 987, 1969.

[5] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[6] Mat Herold, Floris Goes, Stephan Nopp, Pascal Bauer, Chris Thompson, and Tim Meyer. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, 14(6):798–817, 2019.

[7] Martin Jones and Chris Harwood. Psychological momentum within competitive soccer: Players' perspectives. *Journal of Applied Sport Psychology*, 20:57–72, 01 2008.

[8] A. Joseph, N.E. Fenton, and M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006. Creative Systems.

[9] Ana Jović, Ksenija Brkić, and Nikola Bogunović. A review of feature selection methods with applications. *International Journal of Electrical and Computer Engineering Systems*, 1(1):52–62, 2015.

[10] D. V. Lindley. *Introduction to Good (1952) Rational Decisions*, pages 359–364. Springer New York, New York, NY, 1992.

[11] Amirreza Mahdavi-Shahri, Mahboobeh Houshmand, Mahdi Yaghoobi, and Mehrdad Jalali. Applying an ensemble learning method for improving multi-label classification performance. In *2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–6, 2016.

[12] Md. Abu Sayed Md. Riaz Uddin Muntaqim Ahmed Raju, Md. Solaiman Mia. Predicting the outcome of english premier league matches using machine learning. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 187–192. IEEE, 2020.

[13] Harleen Kaur Rahul Baboota. Predictive analysis and modelling football results using machine learning approach for english premier league. In *International Journal of Forecasting*, volume 35, pages 741–755. Elsevier, 2019.

[14] Fátima Rodrigues and Ângelo Pinto. Prediction of football match results with machine learning. *Procedia Computer Science*, 204:463–470, 2022. International Conference on Industry Sciences and Computer Science Innovation.

[15] Artittayapron Rojarath, Wararat Songpan, and Chakrit Pong-inwong. Improved ensemble learning for classification techniques based on majority voting. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 107–110, 2016.

[16] Keisuke Fujii Rory Bunker, Calvin Yeung. Machine learning for soccer match result prediction. *arXiv preprint arXiv:2403.07669*, 2024.

[17] Teo Susnjak Rory Bunkera. The application of machine learning techniques for predicting results in team sport: A review. *arXiv preprint arXiv:1912.11762*, 2019.

[18] Johannes Stübinger, Benedikt Mangold, and Julian Knoll. Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1), 2020.

[19] Aili Wang, Lu Li, and Baotian Dong. Research on pedestrian intelligent recognition method based on cascade classifier structure. In *2020 IEEE 5th International Conference on Intelligent Transportation Engineering (ICITE)*, pages 271–275, 2020.

Final Selected Features

Predictive Feature

FTR 1Encoded: 0 for Home Win (H), 1 for Draw (D), 2 for Away Win (A).

Input Features

- **Points (Pts):** Pts $= 3 \times$ Wins $+ 1 \times$ Draws
- **Points Per Game (PPG):** PPG $= \frac{\text{Pts}}{\text{Matches Played}}$
- **Goals Over 1.5:** Count of Matches where $(FTHG + FTAG > 1.5)$
- **Goals Over 2.5:** Count of Matches where $(FTHG + FTAG > 2.5)$
- **Clean Sheet Percentage:** CS% $= \frac{\text{Matches with Goals Conceded = 0}}{\text{Matches Played}}$
- **Last 5 Encoded Sum:** (Win = 3, Draw = 1, Loss = 0)
- **Recent k Performance:** $\frac{\text{Sum of Metrics (Goals, Corners, Shots)}}{k}$
- **Recent k HS:** Yearly weighted home shots (last k matches).
- **Recent k AS:** Yearly weighted away shots (last k matches).
- **Away Goals Per Game at Away:** $\frac{\text{Total FTAG}}{\text{Matches Away}}$
- **Goal Conversion Rate Difference:** $\left(\frac{\text{Recent Home Goals}}{\text{HST}}\right) - \left(\frac{\text{Recent Away Goals}}{\text{AST}}\right)$
- **Corners Differential:** Corners (Home) − Corners (Away)
- **Defensive Balance:** $|\text{HST} - \text{AST}|$
- **Expected Goals (xG):** Sum of Shot Probabilities
- **Goal Difference:** $\frac{\text{Goals Scored} - \text{Goals Conceded}}{\text{Matches Played}}$
- **Form:** $\xi_j^\alpha = \xi_{j-1}^\alpha + \gamma \cdot \xi_{j-1}^\beta$
- **Form Differential:** Form (Home) − Form (Away)

Model	CV Accuracy	CV Precision	CV Recall	CV F1
LogisticRegression (LR)	0.5297 (±0.0107)	0.4321 (±0.0529)	0.5297 (±0.0107)	0.4490 (±0.0092)
RandomForest	0.4789 (±0.0204)	0.4807 (±0.0152)	0.4789 (±0.0204)	0.4757 (±0.0132)
ExtraTrees	0.5074 (±0.0175)	0.4497 (±0.0177)	0.5074 (±0.0175)	0.4510 (±0.0127)
XGBoost	0.4668 (±0.0211)	0.4378 (±0.0128)	0.4668 (±0.0211)	0.4430 (±0.0129)
LightGBM	0.4774 (±0.0169)	0.4405 (±0.0118)	0.4774 (±0.0169)	0.4457 (±0.0098)
2step_cr	0.4465 (±0.0239)	0.4696 (±0.0167)	0.4465 (±0.0239)	0.4504 (±0.0166)
stack_sl	0.4843 (±0.0184)	0.4528 (±0.0176)	0.4843 (±0.0184)	0.4591 (±0.0181)

Table 1: Performance Summary of the Selected Models (Part 1: Cross-Validation Metrics)

Model	Holdout Accuracy	Holdout Precision	Holdout Recall	Holdout F1
LogisticRegression (LR)	0.5587	0.5424	0.5587	0.4882
RandomForest	0.5268	0.5051	0.5268	0.5107
ExtraTrees	0.5554	0.5335	0.5554	0.5043
XGBoost	0.5185	0.4902	0.5185	0.4941
LightGBM	0.5235	0.4974	0.5235	0.5039
2step_cr	0.4815	0.5067	0.4815	0.4889
stack_sl	0.5319	0.5065	0.5319	0.5123

Table 2: Performance Summary of the Selected Models (Part 2: Holdout Metrics)

Date	HomeTeam	AwayTeam	FTR
01-Feb-25	AFC Bournemouth	Liverpool	A
01-Feb-25	Arsenal	Man City	H
01-Feb-25	Brentford	Spurs	A
01-Feb-25	Chelsea	West Ham	D
01-Feb-25	Everton	Leicester City	H

Date	HomeTeam	AwayTeam	FTR
01-Feb-25	Ipswich Town	Southampton	A
01-Feb-25	Man Utd	Crystal Palace	H
01-Feb-25	Newcastle	Fulham	D
01-Feb-25	Nottingham Forest	Brighton	D
01-Feb-25	Wolves	Aston Villa	A

Table 3: Final Predictions