Football League Predictions

Rohit Singh, C S Ganapathy, Vishakha Kumari, Keshav Kumar

Abstract

Football is one of the world’s most popular sports, attracting the interest of fans, coaches, media, and sports analysts. Predicting match outcomes is a complex challenge due to the many variables influencing results. This unpredictability, coupled with the analytical opportunities it presents, has fueled the development of predictive models to support sports analytics. Predicting football match outcomes is a complex challenge, given the multitude of variables influencing game results. Despite the sport's inherent unpredictability, the availability of comprehensive data from matches and player attributes presents an opportunity to develop advanced predictive models. In this study, we apply machine learning techniques to analyze various statistics from previous matches along with player attributes from both teams to forecast match outcomes. Multiple predictive models were tested, with experimental results showing promising accuracy and insights for sports analysis. We applied multiple machine learning algorithms, including Random Forest, and XGBoost, to predict match outcomes. Our methodology involved data cleaning, feature selection, and model training. We identified key predictive variables, such as goals scored, recent performances, and home game which significantly impact match results.

Our findings demonstrate that machine learning models can achieve high accuracy in predicting football match outcomes, with the MLP (Neural Network) showing the most promising results. This research underscores the potential of machine learning in sports analytics, providing valuable insights for teams, coaches, and analysts. By leveraging historical data and advanced ML techniques, we can develop more accurate and reliable predictive models, enhancing decision-making in football. Furthermore, this study paves the way for future research to incorporate real-time data and additional features, such as player health and weather conditions, to improve predictive accuracy. The integration of machine learning in sports analytics not only enhances our understanding of game dynamics but also contributes to the development of strategies that can optimize team performance and match preparation.

2. Introduction- Related Work

As technology is rapidly increasing, there is a massive outburst in the availability of football league datasets , which leads to easy and secure access to resources for data scientists. We can classify the league data into different classes, While the unpredictability of sports is well known, the football world occasionally witnesses results that defy expectations—such as Leicester City’s stunning English Premier League title win in the 2015/16 season.

A detailed investigation [1] was conducted to explore the factors contributing to this remarkable success and to improve future prediction methods. Key findings highlighted Leicester's exceptional goalkeeper performance and their efficiency in counter-attacking. Additionally, several Leicester players consistently intercepted passes with a high probability of completion (over 80%). This case study also led to the development of a model to predict a team’s shots and goals during a game. The analysis revealed that models incorporating shot types (e.g., counter-attacks, shots from crosses into the penalty area) achieved more accurate predictions.

Another study [2] analyzed data from the 2010/11 Italian Serie A season, using 300 games for training and 80 for testing. One key conclusion was that teams frequently relying on aerial plays were more likely to draw or lose matches. Further research [3] explored machine learning to forecast football match outcomes based on match and player attributes. A simulation study, covering all matches from the top five European football leagues and their second divisions between 2006 and 2018, found that an ensemble approach significantly improved prediction accuracy and offered valuable insights for analyzing match outcomes.

The analysis revealed that studies with lower model performance often lacked variables that effectively capture key characteristics of players and the dynamics of the game. Additionally, it is crucial for models to be trained on data spanning multiple seasons, as teams tend to undergo significant changes each season, affecting performance and outcomes. To overcome these limitations, we have used the concept of rolling averages which is a metric which helps in predicting the team’s performance. The imputation technique is used to get rid of all the inappropriate values.

3. Material and Methods

3.1 Dataset Description

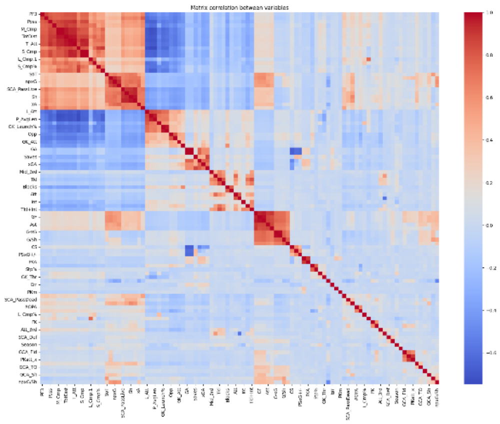
The data for this project was collected by scraping historical football match statistics from the website FBRef. A dataset of 1660 football matches across five seasons, from 2020/2021 to 2024/2025, was collected. However, since only seven games had been played in the ongoing 2024/2025 season at the time of data collection, data from this season was excluded from the analysis. The remaining 1520 matches pertain to the top tier of English football, officially known as the Premier League. Out of these games, the home team won 666 times (20.94%), 340 matches ended in a draw (10.69%), and the away team won 514 times (16.16%).

3.2 Evaluation Parameters

We have considered the variables with the highest positive correlation are those that evaluate the overall quality of a football team. These include passing stats like 'M_Cmp'(Medium Passes Completed), 'T_Att'(Total Passes Attempted), 'TotDist'(Total Passing Distance) and 'xA' (Expected Assists). Similarly, for goal scoring 'Gls'(Goals scored), 'GF'(Match Goals For), 'Ast'(Assists), 'G/Sh'(Goals per shot) and 'xG'(Expected Goals). It can also be observed that the variables like 'L_Cmp'(GK Passes Completed), 'SCA_Fld'(Fouls drawn that lead to shot attempt), 'Att_3rd'(tackles in attacking third) and 'Attendance' have low correlation with Goals. These are the metrics considered to predict the performance/outcome of a team's performance in a league.

3.3 EDA (Exploratory Data Analysis)

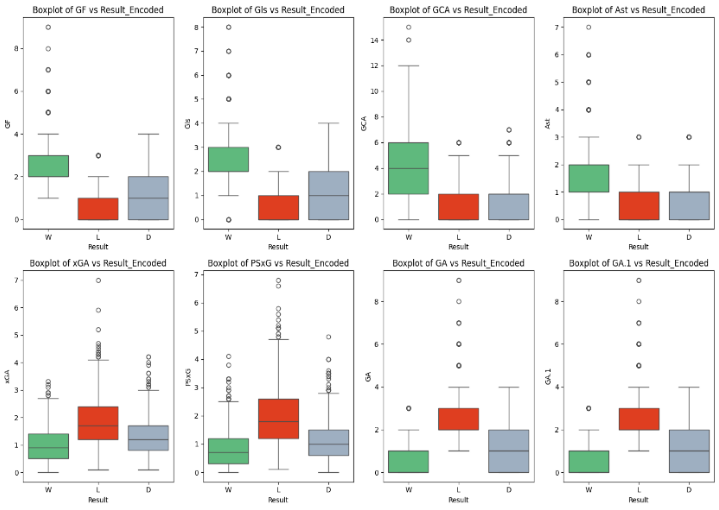
Among the available variables, an analysis was conducted to determine which were most strongly related to each other, which variables had the highest predictive power for the "goal" attribute, and which could be excluded. To assess the relationships between variables, a correlation matrix was created for all numerical variables.



The variables with the highest positive correlation are those that evaluate the overall quality of a football team. These include passing stats like 'M_Cmp'(Medium Passes Completed), 'T_Att'(Total Passes Attempted), 'TotDist'(Total Passing Distance) and 'xA' (Expected Assits). Similarly, for goal scoring 'Gls'(Goals scored), 'GF'(Match Goals For), 'Ast'(Assits), 'G/Sh'(Goals per shot) and 'xG'(Expected Goals). It can also be observed that the variables like 'L_Cmp'(GK Passes Completed), 'SCA_Fld'(Fouls drawn that lead to shot attempt), 'Att_3rd'(tackles in attacking third) and 'Attendance' have low correlation with Goals.

The correlation matrix also played a crucial role in identifying variables to be removed from the initial dataset. In classification tasks, variables with high correlation to others need to be excluded to avoid overestimating their importance, which could negatively impact result predictions. When two variables are identical, one becomes redundant and does not contribute valuable information to the model training. As an example, the variables "S_Cmp" and "S_Att", representing the Short passes completed and attempted, were removed due to their correlation exceeding 0.9.

A crucial step before entering the forecasting phase is identifying variables that can most effectively predict the match outcome, in our case result. To accomplish this, multiple visualizations were created to analyze the relationship between each variable and match outcomes: win, draw, and loss. These visualizations revealed four key variables GF, Gls, GCA, and Ast that are most predictive of a game's result, as shown below. In addition to identifying these predictive variables, an analysis was conducted to find variables with minimal relevance for predicting the target outcome. This analysis indicated that xGA, PSxG, GA, and GA.1 contribute little value to the model and are therefore less useful for training the model.



3.3.1 Goalkeeper Performance

An analysis was conducted to identify the best goalkeeper teams per season and the relationships among various performance metrics. To visualize these relationships, a bar graph was created for each season. This structured approach provided valuable insights into goalkeeper performance, enabling the identification of top goalkeepers for each team across multiple seasons. The findings enhance our understanding of the factors that contribute to a goalkeeper's success and highlight the teams that consistently excel in this area.

The performance of a goalkeeper can greatly influence match outcomes, particularly through their ability to stop goals. The key metrics used to evaluate goalkeeper performance include:

Saves: The total number of shots the goalkeeper has blocked.
Clean Sheets (CS): Matches in which the goalkeeper has prevented the opposing team from scoring.
Penalty Saves (PKsv): The number of penalty kicks saved by the goalkeeper.
Crosses Stopped (Stp): The number of crosses that the goalkeeper successfully intercepted or cleared.

By analyzing these statistics, we can identify patterns in goalkeeper performance and its relationship with team success across different seasons.

This analysis provides a deeper understanding of how goalkeeper performance impacts team success. By examining key metrics such as Saves, Clean Sheets, Penalty Saves, and Crosses Stopped, we can identify the best goalkeeper teams and predict match outcomes. Visualizations such as heatmaps, line plots, and bar charts further clarify these relationships, while machine learning models offer predictive insights for future matches. This exploration underscores the importance of goalkeepers in football, not just in individual performance, but in contributing to overall team success across seasons.

4. Model Development

During the model development stage, data were divided into training and test sets. To address the risk of overfitting, where a model fits too closely to the training data and performs poorly on new data, four seasons were used for training and one season for testing. The training set included three seasons (2020/2021 to 2022/2023), comprising a total of 1,140 games, while the test set consisted of the 2023/2024 season, with 380 games. To identify the optimal classification model, various algorithms with distinct characteristics were tested to determine which one best fits the data.

- Random Forest
- Xgboost
- CatBoost
- LightGBM
- Gradient Boosting
- Stacking-Combines predictions from multiple base linear and nonlinear models (e.g., Random Forest, XGBoost, and LightGBM) into one meta-model to capture complex patterns
- MLP (Neural Network)

4.1 Random Forest

Random Forest is the most popular bagging technique used in current scenarios. This technique considers decision trees as base-learners with applying bagging on top and also uses column sampling strategy with aggregation. The trees considered here are of reasonable depth. The number of base learners to use is a hyper parameter in RFs. The training data considered has been up sampled before it was trained. The best value for hyper-parameter is found out using GridSearchCV with 5-fold cross validation technique. Thereafter we considered the range of estimators for which there is performance raise on the cv data and tried out all the values in that range to obtain best performance on the test data.

4.2 Xgboost

Xgboost is almost like an extension to GBDT. It implements GBDT with column sampling strategy. Row sampling is inbuilt in GBDT in “sklearn” implementation. The number of base learners to use is a hyper parameter in Xgboost. The training data considered has been up sampled before it was trained. The best value for hyper-parameter is found out using GridSearchCV with 5-fold cross validation technique. Thereafter we considered the range of estimators for which there is performance raise on the cv data and tried out all the values in that range to obtain best performance on the test data.

4.3 CatBoost

It is a gradient boosting algorithm that builds decision trees sequentially to reduce errors in each iteration. One of the standout features of CatBoost is its ability to handle categorical data and missing values without the need for extensive preprocessing1. This makes it particularly user-friendly and efficient for feature engineering tasks.

CatBoost is known for its high performance and scalability, and it also supports GPU acceleration to speed up training processes. It can be used for a variety of tasks, including regression, classification, forecasting, and recommendation systems1. Compared to other popular boosting libraries like XGBoost and LightGBM, CatBoost offers unique advantages in handling categorical data, making it a powerful tool for machine learning projects.

4.4 Light GBM

LightGBM (Light Gradient Boosting Machine) is an open-source framework for gradient boosting developed by Microsoft. It is designed to be highly efficient, scalable, and able to handle large-scale data with high performance. LightGBM is particularly well-suited for tasks that involve large datasets and require quick training times.

4.5 Gradient boosting

Gradient boosting is an iterative technique for optimizing models by minimizing a loss function. Unlike other ensemble methods like bagging, which build multiple models independently and aggregate their predictions, gradient boosting builds models sequentially. Each new model aims to correct the errors of the previous ones, gradually improving the overall performance.

4.6 Stacking

Stacking, also known as stacked generalization, is an ensemble learning technique that combines the predictions from multiple base models to form a stronger, more accurate meta-model. This approach leverages the strengths of different algorithms to capture complex patterns in the data that individual models might miss.

4.7 MLP (Neural Network)

An MLP is a type of artificial neural network (ANN) that consists of multiple layers of neurons. It is designed to model complex relationships between inputs and outputs by learning from data through backpropagation. MLPs are particularly powerful for supervised learning tasks like classification and regression.

Input Layer:

The input layer receives the raw input data. Each neuron in this layer corresponds to one feature of the input data. For example, if the input data has 10 features, the input layer will have 10 neurons.

Hidden Layers:

One or more layers situated between the input and output layers. Each hidden layer consists of neurons that apply weights and activation functions to the inputs they receive, transforming the data in progressively more abstract ways. The number of hidden layers and the number of neurons in each layer can vary depending on the complexity of the task.

Output Layer:

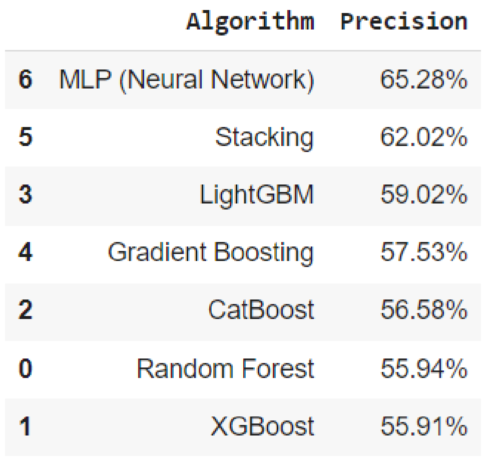
The output layer generates the final predictions. The number of neurons in this layer corresponds to the number of output variables. For binary classification, it typically has one neuron (with a sigmoid activation function), and for multi-class classification, it has as many neurons as there are classes (with a softmax activation function).

Our primary goal is to describe each team using features that highlight its strengths and weaknesses relevant to match outcomes. The dimensions of attacking and defensive performance are both clear and intuitive. A team that consistently scores many goals demonstrates a strong attack. Likewise, a team that regularly prevents its opponents from scoring likely has a strong defense. The stronger a team's attack and defense, the higher the likelihood of it prevailing over an opponent. We can determine a team's current attacking and defensive strengths by aggregating relevant recent performances over a specific period. Rolling averages are statistical tools that allow us to analyze a team's current attacking and defensive strengths by focusing on its performance metrics over the most recent matches (e.g., the last 3 matches). This approach smoothens short-term fluctuations and highlights recent trends, providing a dynamic assessment of a team's form. Rolling averages help us analyze team strength as follows:

- *Attacking Strength* describes a team’s ability to score goals. Metrics like Goals For (GF), Shots (Sh), Shots on Target (SoT), and Expected Goals (xG) are aggregated over the last 3 matches to measure a team’s ability to create and convert chances. If a team has an average of 2.5 goals and 7 shots on target in the last 3 matches, it indicates strong recent attacking performance.
- *Defensive Strength* describes a team’s ability to prevent goals by the opponent. Metrics like Goals Against (GA), Expected Goals Against (xGA), and Defensive Actions (e.g., tackles, interceptions) over the last 3 matches provide insights into a team's ability to prevent the opponent from scoring. If a team has conceded only 0.5 goals on average in the last 3 matches, it shows solid defensive performance.
- *Recent performance* characterizes a team’s current condition in terms of its aggregate performance over recently played matches. Rolling averages capture the current form, which is often a better predictor of near-future performance than season-long averages, especially for dynamic metrics like shots or xG. It accounts for changes like tactical adjustments, player injuries or recoveries and opponent strength in recent matches.
- *Home team advantage* refers to the advantage a team has when playing at its home venue. Home team advantage can be quantified using the dataset of 1,520 matches, the analysis reveals that 20.94% of matches result in a home team victory, 10.69% end in a draw, and 16.16% are won by the away team.

5. Prediction of results

The forecasting results were satisfactory, with MLP(Neural Network) emerging as the best-performing algorithm, achieving a precision of over 65.28%. Stacking and LightGBM followed closely, also showing strong predictive performance. While all algorithms achieved reasonable precision, simpler models like Random forest and XGBoost performed less effectively in handling the complexities of the dataset.



	Algorithm	Precision
6	MLP (Neural Network)	65.28%
5	Stacking	62.02%
3	LightGBM	59.02%
4	Gradient Boosting	57.53%
2	CatBoost	56.58%
0	Random Forest	55.94%
1	XGBoost	55.91%

6. Conclusion

Predicting sports event outcomes is an inherently complex challenge. The primary difficulty in this study was not the creation of a new machine learning algorithm but rather the effective integration of domain knowledge throughout the modeling process. This encompasses every stage, from data collection and integration to model development and refinement. The central premise is that innovative feature engineering, grounded in domain expertise, is critical for improving prediction accuracy. Success depends not merely on the choice of algorithm but on how effectively soccer-specific insights are incorporated into feature design. By leveraging domain knowledge to create meaningful features, models can better capture the intricacies of the game, ultimately enhancing predictive performance.

7. Contributions

Every author has put up lots of efforts in designing the models to work with great performance. The contributions made by every author are listed below,

Author	Contribution
Rohit Singh	Data collection, Model building
C S Ganapathy	Data cleaning and Model building
Vishakha kumari	Feature selection and model building
Keshav Kumar	Model Evaluation and Model building

8. References

[1] Hector Ruiz. 2017. "The Leicester City Fairytale?": Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons. https://dl.acm.org/doi/10.1145/3097983.3098121

[2] Paola Zuccolotto. 2014. Football Mining with R. https://www.researchgate.net/publication/257569396_Football_Mining_with_R

[3] Johannes Stübinger. 2020. Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. https://www.mdpi.com/2076-3417/10/1/46

[4] Daniel Berrar. 2018. Incorporating domain knowledge in machine learning for soccer outcome prediction. https://link.springer.com/article/10.1007/s10994-018-5747-8?getft_integrator=sciencedirect_contenthosting

[5] Fátima Rodrigues. 2022. Prediction of football match results with Machine Learning. https://www.sciencedirect.com/science/article/pii/S1877050922007955