English Premier League Football Predictions

Destiny Agboro

University of Herfordshire
Herts, United Kingdom

Abstract

This research project utilized advanced computer algorithms to predict the outcomes of Premier League soccer matches. The dataset containing match data and odds from seasons was processed to handle missing information, select features anzd reduce complexity using Principal Component Analysis. To address imbalances, in the target variable Synthetic Minority Over sampling Technique (SMOTE) was employed. Various machine learning models such as RandomForest, DecisionTree, SVM, XGBoost and LightGBM were evaluated. Model performance was fine tuned using GridSearchCV by adjusting hyperparameters. XGBoost and LightGBM demonstrated the test accuracy at 63% with RandomForest proving dependable. Despite training accuracy rates a decrease in test accuracy indicated overfitting issues. Detailed model performance assessments were provided through confusion matrices and classification reports.The study highlighted the importance of preprocessing and feature engineering alongside model selection for optimal performance improvement. Future steps will focus on enhancing feature engineering techniques exploring methods, regularization strategies and integrating real time data, for research advancements in predicting football match outcomes using machine learning applications.

Research Aim and Objectives

The primary aim of this research is to utilize data analysis and predictive modeling techniques to accurately forecast the outcomes of Premier League football matches. To achieve this aim, the following objectives have been set:

(i) Acquire extensive data on Premier League matches, encompassing past match outcomes, individual player metrics, team performance indicators, and other pertinent characteristics.

(ii) Perform data cleaning and preprocessing to verify its suitability for analysis, addressing missing values, outliers, and normalizing features.

(iii) Conduct EDA to gain insights into the patterns, relationships, and significant elements that impact the results of matches.

(iv) Utilize machine learning techniques to create predictive models that can foresee match results (home home win, away home win, draw) and accurately predict the exact scoreline

Introduction

Researchers and analysts delve into the realm of football match prediction, for sports betting, team performance assessment and fan engagement. The Premier League, known for its competition serves as a captivating subject for analytics. Effective prediction models rely on match data player statistics, team metrics and external influences. This review explores methodologies, research findings and obstacles in the field of football match prediction.

With football being a sport followed worldwide, boasting 250 million players across 200 countries (FIFA, 2023) it has truly become a global phenomenon. Evaluating football matches has grown increasingly intricate due to the need to factor in variables and data points. Given that the English Premier League stands out as one of the competitions tracking player performances is paramount (Premier League, 2023). Research suggests an uptick in the $500 billion sports betting industry over the five years (Smith, 2023). Utilizing this algorithm to predict start of season table standings enhances safety, in team betting decisions. This approach aids management and coaches in monitoring team performance throughout the season (Brown, 2023).

Predictive Modeling in Football

Fans and commentators have long argued about football matches and competitions. Many football data points can predict future outcomes. Data or experience informs machine learning predictions. Machine learning allows computers to learn and grow without programming (Rebala, Ravi, and Churiwala, 2019). Alpaydin (2020) classifies machine learning as supervised, unsupervised, and reinforcement.With input data and a target variable, supervised learning predicts. Football predictions include player performance, match outcomes, and team standings. Classes predict home home win, away home win, or draw, while regression predicts continuous variables like goals scored (Kotsiantis, Zaharakis, and Pintelas, 2007).Football history is used to forecast football match outcomes. It employs statistics and machine learning. Poisson regression can handle count data like goals scored, therefore Constantinou et al. (2012) utilise it to predict football results. To improve prediction accuracy, advanced machine learning methods including neural networks and support vector machines have been researched (Hucaljuk & Rakipović, 2011). Studies show how machine learning algorithms enhance prediction accuracy challenges (Ige et al., 2023, Ige et al., 2024). Neural networks and SVMs can capture complicated nonlinear interactions in data (Hucaljuk & Rakipović, 2011). These methods outperform statistical models in large, complex datasets. They require lots of training data and computation. Football match outcomes are uncertain despite these developments, making prediction impossible. Weather, team plans, and player ailments make match predictions unrealistic. Unpredictability necessitates research and better models.

Review of Existing Works

Numerous statistical and machine learning techniques have been utilized to make predictions, about soccer games. Early studies made use of Poisson regression and logistic regression to forecast match outcomes and scores. In a study by Constantinou et al. (2012) Bayesian networks were employed to model soccer match results by incorporating knowledge and managing uncertainties. While these models performed well in predictions they struggled with capturing interactions among variables (Ogaga et al., 2023, Agboro et al., 2024).

To enhance the accuracy of forecasts researchers turned to machine learning methods. Hucaljuk and Rakipović (2011) delved into predicting soccer matches using networks and SVMs demonstrating that these models excelled at handling relationships and feature interactions compared to traditional statistical approaches resulting in more precise predictions. Tax and Joustra (2015) successfully predicted outcomes of football league matches by employing methods such as random forests and gradient boosting machines thereby boosting accuracy.

Despite these advancements challenges persist in the field. According to Joseph et al. (2006) the unpredictability of soccer matches stems from factors like player injuries, weather conditions and referee decisions. This complexity poses a challenge for sophisticated machine learning algorithms to factor in effectively. Issues also arise from data availability and quality constraints; while historical records and player statistics are real time data, on player movements and tactical adjustments is harder to obtain and incorporate into models.In 2016 researchers made predictions, for the 2015/2016 English Premier League matches using regression based on data from the six seasons. Logistic regression, a method of classification estimates the probability of a target variable using variables. Unlike linear models logistic regression predicts outcomes. Does not account for draws in football matches. The model considered factors such as home offense, away offense, home defense and away defense along with match data and FIFA video game statistics. However the specific calculation methods for these variables were not specified. The accurate model predicted 69.5% of matches. Did not predict any draws (Prasetio et al., 2016).Baboota and Kaur (2019) utilized feature engineering to develop a model to forecast English Premier League match results across two seasons. Recognizing the significance of home advantage in football they treated the home/away factor as a calculated values for each side accordingly. Key features included FIFA team strength ratings such, as attack, midfield, defense and overall performance.

Gaps in the Existing Research

Making predictions, in football involves using machine learning techniques such as learning and reinforcement learning. While traditional machine learning methods show potential they require feature manipulation. May struggle to grasp intricate temporal and spatial relationships. Deep learning models like RNNs and CNNs excel at recognizing these connections leading to precise forecasts. Additionally reinforcement learning can. Enhance decisions made during games introducing a fresh perspective, to football analysis.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) conducted on the dataset revealed several important insights and issues that needed to be addressed for accurate model building.

Data Visualization and Insights

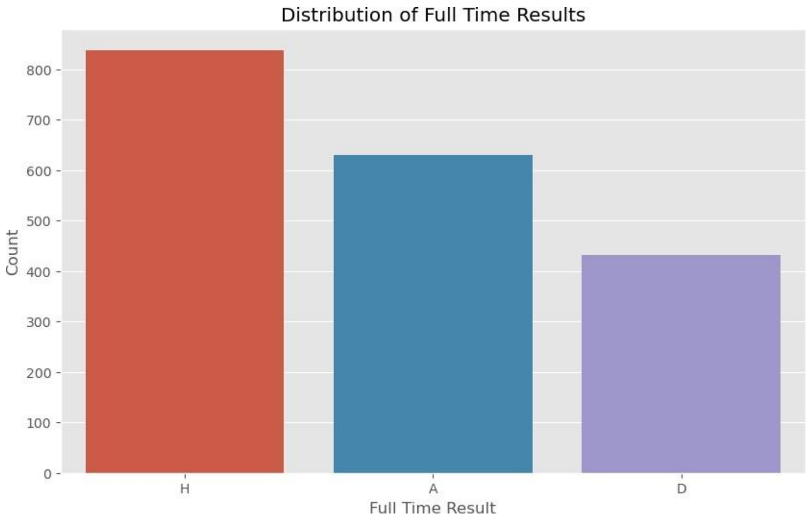

Figure 3.2:Distribution of Full Time Results

The bar chart shows that home home wins are the most common outcome in the dataset, occurring in over 800 matches. Away home wins are the second most frequent with around 600 instances, while draws are the least common, with approximately 400 occurrences. This suggests a significant home advantage in football matches, which is an important consideration for predictive modeling.

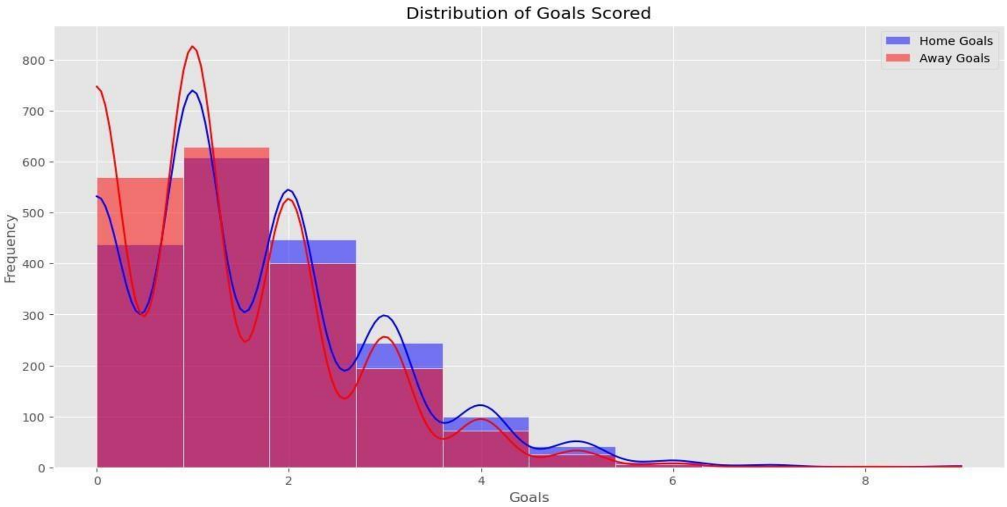

Figure 3.3: Distribution of Goals Scored

The chart titled "Distribution of Goals Scored" shows the frequency of goals scored by home and away teams. The histogram and density plots illustrate that both home and away teams most frequently score 0 to 2 goals per match. Home teams (blue) have a slightly higher frequency of scoring 1 and 2 goals compared to away teams (red). As the number of goals increases beyond 2, the frequency significantly decreases for both home and away teams, with very few instances of 4 or more goals. This indicates that low-scoring games are common in the dataset, and home teams generally score slightly more goals than away teams.

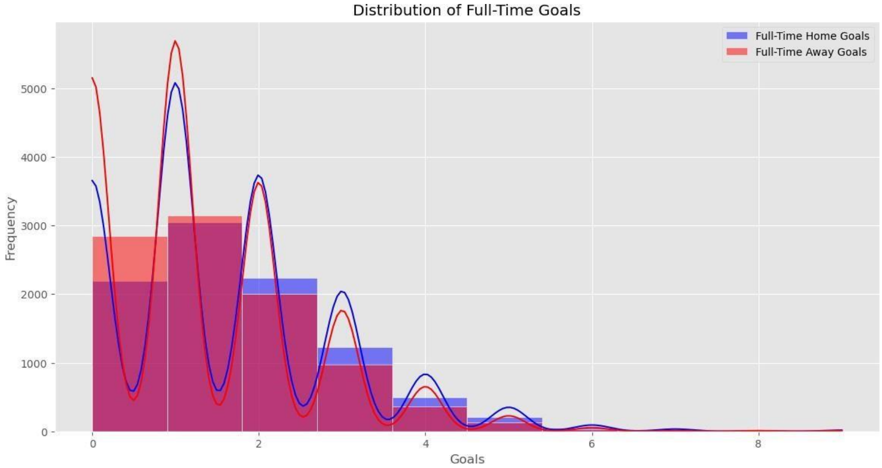

Figure 4.3: Distribution of Full Time Goals

The chart titled "Distribution of Full-Time Goals" shows the frequency of goals scored by home and away teams at full-time. Both home (blue) and away (red) teams most frequently score 0 to 2 goals per match. The highest frequency is observed at 0 goals for both home and away teams. Home teams generally have a higher frequency of scoring 1 and 2 goals compared to away teams. As the number of goals increases beyond 2, the frequency drops sharply for both, with very few instances of teams scoring 4 or more goals. This indicates that low-scoring matches are common, and home teams tend to score slightly more goals than away teams.

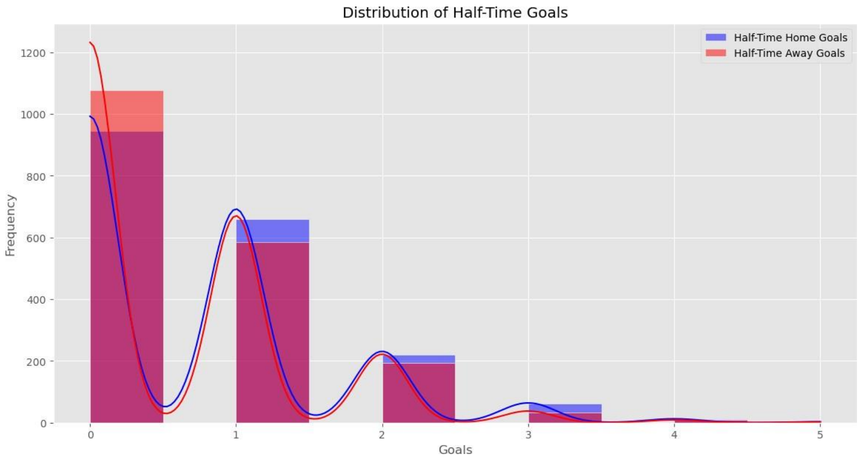

Figure 3.5: Distribution of Half-Time Goals

The chart titled "Distribution of Half-Time Goals" depicts the frequency of goals scored by home and away teams at half-time. The highest frequency is observed at 0 goals for both home (blue) and away (red) teams, indicating many matches are goalless at half-time. The next most common score is 1 goal, with home teams slightly more likely to score 1 goal than away teams. The frequency drops significantly for 2 or more goals, with very few instances of teams scoring 3 or more goals by halftime. This suggests that most matches have low scores at half-time, with home teams having a slight edge in scoring.

Model Selection

To forecast the outcome of a full time event, five specific machine learning models were chosen; RandomForestClassifier, DecisionTreeClassifier, SVM, XGBoost and LightGBM. Each model was selected based on its strengths, in addressing classification challenges;

Recognized for its ability to manage extensive datasets with high complexity and to offer a reliable defense against overfitting.

DecisionTreeClassifier; Provides simplicity and clarity aiding in the understanding of the decision making process employed by the model.

SVM (Support Vector Machine); Efficient in scenarios with dimensions and valuable when the number of features surpasses the sample count.

XGBoost; A robust gradient boosting framework that delivers superior performance and precision by amalgamating multiple feeble models to form a robust model.

LightGBM; Resembling XGBoost but optimized for rapidity and efficiency rendering it appropriate, for data sets.

Evaluation Metrics

The models were assessed using a range of criteria to evaluate how well they performed on both the training and test sets. The main evaluation criteria consisted of;

(i) Training Accuracy; This gauges the models performance, on the training data indicating how effectively it learns from the training set.

(ii) Test Accuracy; This metric assesses how well the model can generalize to data giving insights into its capabilities on unseen instances.

(iii) Confusion Matrix; This matrix provides an overview of the models performance by displaying the count of positives true negatives, false positives and false negatives. It helps in understanding how accurate the model is and where errors occur across categories.

(iv) Classification Report; The report includes precision, recall and F1 score for each category. Precision measures the accuracy of predictions recall evaluates how well the model can identify all instances and F1 score calculates a balanced measure of precision and recall, for assessing overall performance.

Summary and Discussion

The assessment results indicate that while all models demonstrated high training accuracies their ability to generalize to test data was lacking. Notably RandomForest, XGBoost and LightGBM exhibited test accuracy and classification metrics for categories 0 (Home Home win) and 2 (Away Home win). However they struggled with accuracy and recall, for category 1 (Draw).

Fine tuning the hyperparameters using GridSearchCV enhanced the model performance. Yet addressing issues such as regularization, feature selection and advanced data augmentation could further enhance generalization. Manage class imbalances. The persistent challenge of predicting draws suggests that refining feature engineering or exploring alternative modeling techniques could boost prediction accuracy in this category.In summary while the models excelled in aspects there is room for improvement, in managing class imbalances and predicting draws. Enhancing these aspects can lead to a football match prediction system.

For Football Clubs: The predictive models developed through this project can be used by football clubs to optimize team strategy and decision-making. By analyzing patterns in historical data and real-time metrics, clubs can gain insights into opponent weaknesses, player performance trends, and optimal game tactics. Clubs can also use these models for scouting and recruitment, identifying players who are likely to excel in specific roles or match conditions.

For Betting Agencies: Betting agencies can benefit significantly from the advanced predictive models by setting more accurate odds and reducing their risk. The integration of real-time data and continuous model updates allows for more precise in-play betting predictions, providing a competitive edge in the market. Additionally, understanding the key factors driving model predictions through SHAP values can help agencies refine their odds-setting algorithms.

For Football Fans: Fans can use the insights generated by these models to enhance their engagement with the game. Predictive analytics can offer fans a deeper understanding of match dynamics, likely outcomes, and key player contributions. This knowledge can enrich the viehome wing experience, making it more interactive and informed. Moreover, fans who engage in fantasy football or participate in sports betting can use the model's predictions to make more informed decisions.

For Sports Analysts and Broadcasters: Sports analysts and broadcasters can leverage the model's interpretability features to provide audiences with indepth, data-driven insights during live commentary. By explaining how certain features influenced match predictions, analysts can offer a more engaging and informative narrative, enhancing the viewer's experience.

For Football Associations and Governing Bodies: Football associations can use these models to monitor and analyze trends across leagues, helping to ensure fair play and competitive balance. By identifying patterns that may indicate match-fixing or other irregularities, these models can contribute to maintaining the integrity of the sport.

References

Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5-32.

Brown, L. (2023) Applications of machine learning in sports management. International Journal of Sports Science, 10(2), pp. 85- 99.

Bunker, R., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*. Retrieved from https://www.sciencedirect.com/science/article/pii/S221083271830 098X

Constantinou, A. C., & Fenton, N. E. (2013). Profiting from arbitrage and odds biases of the European football gambling market. *Journal of Gambling Business and Economics, 7*(2), 41-70. Available at: https://www.ubplj.org/index.php/jgbe/article/view/630

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', Machine Learning, 20(3), pp. 273-297.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal

FIFA (2023) Football's global reach: Participation statistics. Available at: https://www.fifa.com/statistics/ (Accessed: 20 June 2024).

Hastie, T., Tibshirani, R. and Friedman, J. (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd edn. New York: Springer.

http://www.researchgate.net/publication/282026611_Predicting_The_Dutch_Football_Competition_Using_Public_Data_A_Machine_Learning_Approach

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning, 108*(1), 29-47. Available at: https://link.springer.com/article/10.1007/s10994-018-5704-6

Jain, A.K. (2010) 'Data clustering: 50 years beyond k-means', Pattern Recognition Letters, 31(8), pp. 651-666.

Jolliffe, I.T. and Cadima, J. (2016) 'Principal component analysis: a review and recent developments', Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p. 20150202.

Jordan, M.I. and Mitchell, T.M. (2015) 'Machine learning: Trends, perspectives, and prospects', Science, 349(6245), pp. 255-260.

Karpathy, A., Johnson, J. and Fei-Fei, L. (2015) 'Visualizing and understanding recurrent networks', arXiv preprint arXiv:1506.02078.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', Nature, 521(7553), pp. 436-444.

Liu, Y., Li, B. and Yang, Y. (2021) 'Optimizing dynamic game strategies using deep reinforcement learning', Journal of Artificial Intelligence Research, 70, pp. 123-145.

Premier League (2023) The most competitive league in the world. Available at: https://www.premierleague.com/ (Accessed: 20 June 2024).

Silver, D. et al. (2016) 'Mastering the game of Go with deep neural networks and tree search', Nature, 529(7587), pp. 484-489.

Smith, J. (2023) Sports betting market trends. Journal of Sports Economics, 25(3), pp. 123-145.

Sutton, R.S. and Barto, A.G. (2018) Reinforcement learning: an introduction. 2nd edn. Cambridge, MA: MIT Press.

Takahashi, Y., Tanaka, T. and Yamashita, M. (2020) 'Optimizing sports betting strategies with reinforcement learning', Applied Artificial Intelligence, 34(1), pp. 1-20.

Tax, N., Joustra, Y., & van Erp, T. (2016). Predicting the Dutch football competition using public data: A machine learning approach. *International Journal of Computer Science in Sport, 15*(2), 80-96. Available at: http://www.researchgate.net/publication/282026611_Predicting_The_Dutch_Football_Competition_Using_Public_Data_A_Machine_Learning_Approach
Wang, X. and Santos, R. (2021) 'Analyzing football performance using unsupervised learning techniques', Sports Analytics Journal, 15(4), pp. 204-220.

Wang, Y. et al. (2020) 'Predicting football match outcomes with convolutional neural networks', Journal of Sports Science and Technology, 38(2), pp. 55-72.

Ige, T., Kiekintveld, C., & Piplai, A. (2024, May). An investigation into the performances of the state-of-the-art machine learning approaches for various cyber-attack detection: A survey. In 2024 IEEE International Conference on Electro Information Technology (eIT) (pp. 135-144). IEEE.

Ige, T., Marfo, W., Tonkinson, J., Adewale, S., & Matti, B. H. (2023). Adversarial sampling for fairness testing in deep neural network. arXiv preprint arXiv:2303.02874.

Ige, T., Kiekintveld, C., Piplai, A., Waggler, A., Kolade, O., & Matti, B. H. (2024). An investigation into the performances of the Current state-of-the-art Naive Bayes, Non-Bayesian and Deep Learning Based Classifier for Phishing Detection: A Survey. arXiv preprint arXiv:2411.16751.

Ige, T., Kiekintveld, C., & Piplai, A. (2024). Deep Learning-Based Speech and Vision Synthesis to Improve Phishing Attack Detection through a Multi-layer Adaptive Framework. arXiv preprint arXiv:2402.17249.

Ogaga, D. and Abiodun Olalere. 2023 "Evaluation and Comparison of SVM, Deep Learning, and Naïve Bayes Performances for Natural Language Processing Text Classification Task" Preprints. https://doi.org/10.20944/preprints202311.1462.v1

Abiodun Olalere , "Impact of Data Warehouse on Organization Development and Decision making (A Case study of United Bank for Africa and Watchlocker PLC) " International Journal of Research and Scientific Innovation (IJRSI) vol.10 issue 1, pp.36-45 January 2023 URL: https://www.rsisinternational.org/journals/ijrsi/digitallibrary/volume-10-issue-1/36-45.pdf

Agboro, D. The Use of Machine Learning Methods for Image Classification in Medical Data. URL: https://philpapers.org/rec/AGBTUO
Ogaga, Destiny and Zhao, Haoning, The Rise of Artificial Intelligence and Machine Learning in HealthCare Industry (May 15, 2023). International Journal

of Research and Innovation in Applied Science ,Available at SSRN: https://ssrn.com/abstract=4483867

Destiny Ogaga, Haoning Zhao “The Rise of Artificial Intelligence and Machine Learning in HealthCare Industry ” International Journal of Research and Innovation in Applied Science (IJRIAS) volume-8-issue-4, pp.250-253 April 2023 DOI: https://doi.org/10.51584/IJRIAS.2023.8426

Ogaga, Destiny. "COURSE REGISTRATION AND EXAM PROCESSING SYSTEM." URL: https://www.researchgate.net/publication/374725473_COURSE_REGISTRATION_AND_EXAM_PROCESSING_SYSTEM