678n, m

Performance Analysis and Data Mining of Premier League Teams (2016/17–2017/18)

Yasin Lester M00809695

A Data-Driven Approach to Performance Analysis and Insights Using J48 Decision Tree Algorithms"

Contents

[Introduction 1](#_Toc185012682)

[Data Analysis and Visualisation 2](#_Toc185012683)

[Data Mining Algorithm and Data Pre-processing 9](#_Toc185012684)

[Data Mining 9](#_Toc185012685)

[Data Ethics 11](#_Toc185012686)

[Conclusion 11](#_Toc185012687)

[References 12](#_Toc185012688)

# Introduction

This project uses the Premier League dataset to offer a whole study of club and player performance in the 2016/17 and 2017/18 football seasons. With a range of football match metrics including goals scored, clean sheets, assists, passes, and disciplinary records this Kaggle dataset (Nalla, Z., 2024) presents. These variables give thorough understanding of team and player performance, so enabling a better knowledge of trends and patterns that could guide strategic decisions in sports management and analytics.

The dataset features 42 variables and 240 cases, offering a range of measures including the number of wins, losses, goals scored, goals conceded, passes tried, tackles, and set pieces including penalties and free kicks. General classification for these factors is performance measures (goals, assists), disciplinary metrics (red and yellow cards), and defensive metrics (clean sheets, saves). With data visualisation and mining tools, the well-structured and orderly data makes perfect study environment. The data is structured and well-organised, making it ideal for examination with data visualisation and mining tools.

Although the dataset was presumed to be clean, it is critical to examine possible cleaning strategies to assure robustness and dependability. Hypothetically, cleaning procedures may have included correcting missing values using imputation methods, such as replacing them with the mean or median of related attributes. Box plots or Z-scores could have been used to find and rectify or remove outliers in performance indicators, such as exceptionally high or low numbers. To avoid conflicts during analysis, consistency checks would have been added, such as ensuring standard forms for variables like dates or categorical entries (e.g., team names).

To ensure compatibility with Tableau, the dataset was prepared in a structured format, such as a CSV file with labelled columns. Although no anomalies or inconsistencies were discovered, alternative actions to prepare the data for visualisation could have included normalising numerical values to increase scale comparability or encoding categorical variables for more sophisticated analysis.

Using exploratory data analysis, visualisation, and machine learning techniques, this dataset offers enormous possibilities for trend and pattern discovery. The results of the research might offer football managers, analysts, and other interested parties important business intelligence.

# Data Analysis and Visualisation

Does a strong relationship exist between the number of clean sheets achieved by Premier League teams and their total number of successful passes and tackles, and how can this data inform defensive strategies?

Figures 1 and 2 show the relationship between clean sheets and total successful passes and tackles under examination here. Figure 1 shows for Premier League teams in the 2016/17 and 2017/18 seasons a scatter plot showing the relationship between total passes and clean sheets. Leading in successful passes, Manchester City and Chelsea who had the most clean sheets 36 and 32, respectively showcased how defensively stable possessions-dominant teams are. Teams with less passes like Sunderland and Hull City have often maintained the fewest clean sheets, therefore underlining their lack of defensive power.

Figure 2 looks at how clean sheets and total tackles relate. Teams with higher tackle counts, like Manchester United and Chelsea, have had strong defensive records, therefore a good trend is obvious. Though tackling helps to create clean sheets, the correlation in Figure 2 is far less clear than in Figure 1, suggesting that possession through passing has more influence.

These results suggest that teams with a possession-oriented approach that restricts the capacity of the opposition to attack might enhance their defensive performance. To increase the performance of their team in both possession and tackling, managers could give top priority to selecting players with outstanding passing accuracy and defensive awareness, therefore producing better clean sheet statistics.

A graph with dots and lines

Description automatically generated

Figure 1

A graph with dots and numbers

Description automatically generated

Figure 2

2. Do teams with the highest number of goals also record the highest number of shots on target, and is there a strong relationship between the total passes and goals scored by Premier League teams?

The data presented in Figures 3 and 4 show a clear positive link between goals scored and both shots on target and total passes, implying that teams with more goals outperform in these categories. In Figure 3, Manchester City had the most goals (186), as well as the most shots on target (481), illustrating their attacking efficiency and ability to capitalise on opportunities. Liverpool follows closely with 160 goals and 473 shots on goal, showing the link between creating opportunities and scoring goals. Middlesbrough, on the other hand, has only 27 goals and 149 shots on goal, highlighting their problems in front of goal and inefficient conversion of chances.

Figure 4 shows a scatter plot with a substantial positive association between total passes and goals scored. Manchester City, Liverpool, and Chelsea lead the league in total passes (approximately 47,000, 43,000, and 39,000, respectively), as well as goals scored (186, 160, and 162 goals). The trend line emphasises the need of maintaining possession to generate scoring opportunities. Sunderland and Hull City, on the other hand, have fewer than 15,000 passes and fewer than 40 goals, highlighting the difficulties of generating offensive chances when possession is low.

Overall, Figures 3 and 4 show that teams with more shots on goal and more total passes are more likely to score. Adopting a possession-based strategy and improving shot accuracy may help underperforming teams increase their offensive production. Managers could consider focussing on players with great passing and finishing skills to capitalise on these relationships and improve their outcomes.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 3

A graph with numbers and letters

Description automatically generated

Figure 4

3. How does the distribution of goals conceded vary among Premier League teams, and is there a relationship between successful passes and a reduction in losses?

Figure 5 shows notable variation in the distribution of goals lost by Premier League clubs during the seasons 2016/17 and 2017/18. Most teams conceded between 57 and 80 goals; Sunderland and Hull City stood out as outlays, giving up 132 and 124 goals respectively, respectively, therefore emphasising their defensive problems. Manchester City, with their defensive strength and possession-based approach, only conceded 47 goals. With goals dropped consistently below the league average, Chelsea and Tottenham both performed rather well.

With teams conceding more goals facing more losses, Figure 6 illustrates a clear positive correlation between goals conceded and losses. With 132 goals given up, Sunderland lost the most 42; Manchester City suffered the fewest only 47 goals given up. Although neither statistic precisely links successful passes with reduced losses, past studies have indicated that defensive stability depends critically on possession. Teams who control possession, like Manchester City, can limit defensive exposure and lower goals surrendered, therefore lowering the number of defeats.

These results stress the need of ball control in reducing defensive exposure. Teams that retain possession can lessen the attacking possibilities of the opposition and hence address defensive issues. This knowledge should help coaches to enhance their strategies by stressing methodical build-up play and ball retention to lower defensive vulnerabilities.

A screen shot of a computer

Description automatically generatedFigure 5

A graph with a line and dots

Description automatically generated with medium confidence

Figure 6

4. Is there a correlation between the number of total goals and the number of wins for Premier League teams?

For Premier League teams in 2016/17 and 2017/18, the total number of goals scored, and the number of victories show a clear positive link in Figures 7 and 8. With Manchester City hitting a high of 32 wins in the 2017/18 season, Figure 7 displays a heat map of wins by team and season showing that Chelsea and Manchester City have routinely earned the most. These teams also scored the most goals, implying a strong relationship between match outcomes and offensive performance.

Figure 8 shows a scatter plot of goals scored and won, therefore offering more proof for this result. Teams with strong goal-scoring ability such as Liverpool (160 goals, 23 wins) and Manchester City (186 goals, 32 wins showcase the need of this talent in guaranteeing successes. Teams with low offensive output like Middlesbrough and Sunderland which scored less than 30 goals and won the fewest games five and six, respectively showcase the challenges these teams face.

These numbers clearly show that a team's capacity to win games depends much on its capacity to score goals. Clubs should give their attacking plan top priority to boost performance: they should raise scoring possibilities and acquire players with outstanding finishing ability. Managers can use this information to rank offensive improvements as they know that raising goal output directly helps to increase winning more games.

A close up of a screen

Description automatically generated

Figure 7

A graph with a line

Description automatically generated

Figure 8

5. Is there a correlation between the number of red cards a team receives and the number of goals and penalties they concede?

The data in Figures 9 and 10 examine whether the clubs with the highest number of red cards concede a significant number of goals and penalties during the 2016/17 and 2017/18 Premier League seasons. The number of red cards and the number of goals conceded are depicted in Figure 9. The highest number of red cards was exhibited by Sunderland, which also conceded the most goals (132), suggesting an absence of defensive discipline. Hull City and West Ham United both received numerous red cards and conceded numerous goals (123 and 102, respectively). In contrast, Manchester City had the fewest red cards (47), which suggests that a more disciplined defensive approach is associated with improved goal-prevention.

Red cards are juxtaposed with penalties in Figure 10. West Ham United and Watford, teams with a higher number of red cards, also conceded a significant number of penalties (16 and 15, respectively). Nevertheless, this correlation is not universal, as certain teams, such as Arsenal, received a greater number of red cards while conceding fewer penalties. The value of disciplined play in reducing risks in critical defensive zones is exemplified by Manchester City and Tottenham, which have low red card and penalty tallies.

Figures 9 and 10 show that teams with more red cards concede more goals and penalties, but the association is not absolute. Maintaining discipline, both in terms of avoiding red cards and sticking to defensive systems, is critical for reducing goals and penalties conceded. Teams should prioritise defensive organisation and player discipline to improve overall performance and mitigate the impact of costly fouls and dismissals. Managers can use this information to prioritise discipline and defensive control throughout training and match preparation.

A screen shot of a graph

Description automatically generated

Figure 9

A graph of a number of colored lines

Description automatically generated with medium confidenceFigure 10

# 

# Data Mining Algorithm and Data Pre-processing

For the analysis J48 Decision Tree technique was used for the Premier League dataset analysis due to its capability of handling both categorical and numerical data and is clear in visualising feature-outcome correlations. This decision tree technique generates a hierarchical model by splitting input into branches and predicting outcomes at the leaf nodes.

Key Reasons for Choosing J48:

* Interpretability: improves managerial understanding by means of open visualisation of relationships.
* Efficiency: Suited for the size of the dataset (240 cases), hence lowering the overfitting risk.
* Relevance: Visualisations showing great pattern modelling using J48 included scatter graphs of total passes against victories.

Main Variables for J48 Decision Tree:

* Class Variable: Season (categorical, intended for classification)
* Predictor Variables: Clean sheets, total passes, goals scored, shots on target, penalties conceded, red cards.

Data Pre-processing:

* Hypothetical Clean Dataset: Any possible issues including outliers or missing values were handled theoretically.
* Missing Values: May be imputed to avoid skewness using the median for numerical features.
* Outliers: To ensure exact model predictions, addressed by means of z-score normalisation or interquartile range (IQR).
* Feature Scaling: Although not required, it could be decided to normalise the impact of variables including goals scored.
* Using the J48 Decision Tree and fixing pre-processing issues will help the study offer Premier League club performance management strong, understandable information.

# Data Mining

The J48 Decision Tree method was utilised in this investigation because to its strong classification abilities and straightforward interpretability. It generates decision trees by optimising splits to reduce entropy, rendering it very proficient in elucidating relationships between variables and classifying intricate data sets. The main objective was to evaluate the performance of Premier League teams throughout many seasons to ascertain the factors contributing to success. The dataset underwent three rounds to improve the model's accuracy and interpretability.

Iteration 1: Baseline Analysis

* Initial Model: The complete dataset with 42 attributes was used to establish a baseline, generating a decision tree with 39 leaves and a size of 77 nodes. Initial accuracy was determined to be 44.2% through 10-fold cross-validation.
* Observations: Key variables like saves and interceptions were influential in the tree structure, underscoring their importance. Conversely, attributes such as own goals and big chance missed introduced noise, diminishing classification precision.
* Result: This phase underscored the propensity for overfitting when too many irrelevant features are included, setting the stage for more focused iterations.

Iteration 2: Feature Refinement

* Refinement: Ineffective attributes like own goals and dispossessed were removed, resulting in a streamlined tree with 38 leaves and a size of 75 nodes.
* Model Performance: Accuracy slightly decreased to 42.9%. The focus was shifted towards simplifying the tree to enhance interpretability rather than purely maximizing accuracy.
* Key Takeaway: This step highlighted the necessity of balancing model complexity with predictive accuracy, informing further refinement.

Iteration 3: Optimal Model

* Further refinements: The most compact and effective tree with 36 leaves and a capacity of 71 nodes is achieved by removing attributes such as "atto\_box\_goal," resulting in a reduction from the original 77 nodes. The highest accuracy of 47.5% was attained in this iteration.
* Model Efficiency: Demonstrated enhanced classification, particularly for high-performing seasons, while maintaining the model's simplicity.
* Final Evaluation: This model is the most appropriate for detailed analysis due to its effective balance of accuracy and simplicity of interpretation.

Model Evaluation and Insights

* Progressive Refinement: The model's size was effectively reduced from 77 to 71 nodes through iterative refinement, resulting in a simplified structure that preserves essential predictive attributes.
* Influential Characteristics: Consistent variables, including total clearance and saves, substantially influenced successful classifications across various seasons.
* Iterative Advantage: The model's overall efficacy was substantially enhanced by feature selection, which was instrumental in reducing overfitting.

Assessment of Model Fit

* Analysis Tools: The model's fit was assessed using a classification metric and a confusion matrix, resulting in a 47.5% accuracy. The model's efficacy and the dataset's complexity are both indicated by this accuracy.
* Practical Applications: The model is interpretable and concise, providing actionable insights for predicting and comprehending the performance of Premier League teams. This model is of significant value in the field of sports analytics.

In addition to improving the model's accuracy, this methodical approach provided a comprehensive understanding of the critical factors that determine success in the Premier League.

# Data Ethics

Ethical, legal, and professional considerations are essential when studying and utilising company data. These elements guarantee responsible, transparent, and compliant examination of the Premier League dataset. The following is a succinct evaluation of each principal domain:

1. Ethical Considerations

* Avoiding Bias: To ensure fairness, the analysis includes data from all teams, not just high-performing ones like Manchester City.
* Data Integrity: The dataset was not altered, ensuring accuracy and preventing misleading results.
* Avoiding Misleading Conclusions: Trends are clearly labelled as correlational, not causal, to prevent misinterpretation by stakeholders.

2. Legal Considerations

* GDPR Compliance: Although no personal data was used, hypothetical scenarios involving player-specific data would require anonymization and secure storage.
* Intellectual Property and Licensing: The dataset, sourced from Kaggle, follows licensing terms. Proper attribution ensures compliance with intellectual property rights.
* Proper Use of Data: The dataset was used solely for academic analysis, aligning with Kaggle’s usage terms.

3. Professional Considerations

* Transparency and Accountability: All results from Tableau and the J48 Decision Tree were presented honestly, with no exaggeration.
* Acknowledgement of Limitations: Limitations, like excluding external factors (e.g., player injuries), are highlighted to ensure stakeholder clarity.
* Responsible Communication: Insights are framed as correlations, not causations, ensuring managers make informed, objective decisions.

By addressing ethical, legal, and professional considerations, the analysis remains responsible, transparent, and credible, enabling stakeholders to rely on its insights for informed decision-making.

# Conclusion

Data visualisation and machine learning methods were used to examine the Premier League statistics for the 2016/17 and 2017/18 seasons, therefore exposing significant team performance trends. The results unequivocally showed that teams with a possession-based approach shown by high successful pass rates usually had better defensive records (clean sheets) and general team success (wins). While Sunderland underperformed, pointing deeper problems, Manchester City thrived in keeping possession, goal-scoring, and disciplined defence.

The employment of the J48 Decision Tree algorithm throughout the data mining process made a substantial contribution to these discoveries. The model was modified across three iterations to improve its accuracy and interpretability, eventually reaching an optimal performance of 47.5% accuracy in the final iteration. This iterative development underlined the necessity of proper feature selection, as measurements such as saves, total tackles, and total clearances were found to be indicative of team performance.  
  
These discoveries have significant implications for football managers, analysts, and strategists. To enhance team performance, managers may wish to investigate strategies that emphasise ball retention and robust defensive formations. Analysts can pinpoint areas for development, such as increasing player efficiency in important performance indicators like interceptions and clearances. Meanwhile, club officials should modify recruitment efforts to attract players whose skills match the discovered success characteristics, emphasising precision passing and tactical discipline.   
  
In conclusion, this analysis not only illustrates the advantages of integrating data visualisation and machine learning in sports analytics, but it also offers stakeholders a framework for making data-driven, informed decisions that could result in enhanced team performance and strategic success in competitive football.

# References

Nalla, Z., 2024. *Premier League Dataset*. Kaggle. [online] Available at: <https://www.kaggle.com/datasets/zaeemnalla/premier-league> [Accessed 15 November 2024].