Predicting Premier League Match Outcomes for Liverpool FC Using Machine Learning

By Abdula Alkhafaji

Abstract

Accurately predicting football match outcomes remains a challenging problem due to factors such as limited data samples, evolving team dynamics, and inherent randomness in sports. In this paper, we develop a pipeline for scraping and engineering match data from FBref, then applying machine learning models—Logistic Regression and Random Forest—to predict the final five matches of Liverpool FC's season. We focus on seasons 2018-2019 through 2023-2024, each processed individually. Our methodology includes data scraping, cleaning, feature engineering, and training classification models on the first 33 league matches to predict the outcomes (Win/Draw/Loss) of the final 5 matches. The results show moderate to low accuracy, influenced by small sample sizes, class imbalance, and unique season contexts (e.g., major squad changes, injuries, or unexpectedly high number of wins). Despite the modest accuracy scores (ranging from 0.2 to 0.6 depending on the season and model), our analysis highlights how domain knowledge and advanced techniques (e.g., SMOTE or hyperparameter tuning) can improve forecasts for football match outcomes.

Keywords: machine learning, football prediction, Logistic Regression, Random Forest, Liverpool FC, FBref

Introduction

Football (soccer) is among the most popular and widely analyzed sports globally. Predicting match outcomes not only captivates fans and analysts but also guides strategic decisions in betting, sports media, and tactical planning. Traditional approaches have relied on expert intuition and basic statistical models, yet the rise of data science has driven a growing interest in **automated prediction pipelines**.

This project aims to build a single pipeline that can:

1. **Scrape Liverpool FC's Premier League match data** from FBref.
2. **Generate engineered features** related to goals, possession, rolling averages, and more.
3. **Train classification models** using the first 33 matches of each season.
4. **Predict the final 5 matches** and evaluate the model's performance.

We focus on **Liverpool FC**, an English Premier League club, to demonstrate how individual season contexts (e.g., a season with many wins vs. an inconsistent season) can drastically affect model predictions and accuracy. Sections of this paper discuss related work, the proposed end-to-end methodology, experimental results for each season, and conclusions.

Related Work

Predictive analytics in football has been explored extensively. **Baio and Blangiardo (2010)** used Bayesian hierarchical models to predict match scores, while **Hvattum and Arntzen (2010)** applied Elo-based ratings to forecast outcomes. More recent approaches incorporate **machine learning algorithms** (e.g., Random Forest, Gradient Boosting), focusing on advanced stats like xG (expected goals), passing networks, and rolling averages to address the dynamic nature of football data.

Sports Analytics literature emphasizes the challenges of small datasets, inherent class imbalance (Win often more frequent than Loss for top clubs), and the importance of domain knowledge—such as player transfers or injuries. Our pipeline captures some of these insights by engineering rolling features and encoding match contexts like **Home/Away**. However, building robust models with limited samples remains a difficulty.

Proposed Methodology

In this section, we outline the core methodology, subdivided into data acquisition, feature engineering, model training, and evaluation. The pipeline is modular, allowing each season to be processed independently.

3.1 Data Acquisition and Scraping

We used soccerdata library's **FBref** interface to scrape Liverpool's Premier League match data for each season. Key stat types included:

- **schedule**: basic match info (date, venue, GF, GA, possession)
- **shooting**: shots, shots on target, average shot distance
- **passing**: passes completed, passes attempted, pass completion percentages
- **defense**: tackles, interceptions, blocks, clearances
- **keeper**: saves, goals against, save percentage
- **misc**: fouls, aerial duels, errors

We merged these data frames on date, venue, and opponent columns. The result was a single CSV file per season (lfc_data_<year>.csv) containing each match's combined stats.

3.2 Feature Engineering

After retrieving the raw data, we applied a **feature engineering pipeline** to create an enriched dataset (lfc_data_<year>_processed.csv):

1. **Goal Difference**: Goals For - Goals Against
2. **Rolling Averages (N=5)**: For goals, possession, xG, interceptions, etc. We aggregated the last 5 matches to capture short-term form.
3. **Win/Draw/Loss Encoding**: Creating Result_Code (Win=1, Draw=0, Loss=-1) as the classification target.
4. **Opponent Label Encoding**: A numeric code for each opponent (Opponent_Code).
5. **Days Since Last Match**: Time-based feature capturing match spacing or congested fixture periods.

3.3 Model Training and Testing

We utilized two classification models:

- **Logistic Regression** (LR)
- **Random Forest** (RF)

Training Set: The first 33 matches of each season.
Test Set: The last 5 matches of that same season.

This approach simulates predicting final outcomes with knowledge only up to matchday 33. We applied **oversampling** on the training set to handle the low representation of draws/losses, which often occur less frequently for a strong team like Liverpool (especially in a successful season, e.g., 2019-2020).

3.4 Evaluation Metrics

We employed **accuracy** for an overall metric, supplemented by **confusion matrices** and classification reports (precision/recall/F1) for each class (Win, Draw, Loss). The final pipeline produced summary CSV files of each season's accuracy and image files of confusion matrices.

Experiment Setup and Results Discussion

We processed six seasons individually, but partial data or errors occurred for some. Our final summary included the three seasons that successfully produced predictions:

4.1 Season 2018-2019

Context:

- 2018-2019 was a strong season for Liverpool, finishing 2nd in the Premier League with only one loss. The pipeline, however, encountered a mismatch of classes in the final 5 matches (some had only "Win" or "Loss" present), causing classification_report errors.

Outcome:

- Partial or no final results due to the classification_report mismatch.
- This season's pipeline run indicated that if the last 5 matches happen to contain only 1 or 2 classes (e.g., 3 wins, 2 draws, no losses), the confusion matrix becomes degenerate.

Key Takeaway:

- Class mismatches highlight the **small test set problem**. With only 5 matches, the distribution can become skewed (e.g., zero "Loss" matches). This reveals limitations for short-window predictions.

4.2 Season 2019-2020

Context:

- Liverpool famously won the title, delivering an extremely high win rate. The final 5 matches still had enough diversity to produce results, but the model tended to overpredict "Win."

Accuracy:

- **Logistic Regression**: ~0.60
- **Random Forest**: ~0.60

Confusion Matrices:

- Both models predicted nearly all matches as "Win."
- Logistic Regression confusion matrix showed correct predictions for a majority of wins, occasionally predicting a loss or draw incorrectly.

Interpretation:

- Overwhelming season success means the model's oversampling approach sometimes overshoots on "Win."
- The final 5 matches included a couple of surprises (like a draw or loss), which the model rarely predicted.

4.3 Season 2020-2021

Context:

- A season marked by injuries (particularly in defense) leading to inconsistent results.
- Our pipeline tried class balancing via oversampling, yet certain final test matches lacked some classes entirely.

Outcome:

- During the last 5 matches, the script encountered "Number of classes, 2, does not match size of target_names, 3" errors.
- This error signaled that the actual 5-match distribution might have, for example, only "Loss" and "Win." The classification report was not generated for the missing class.

Key Takeaway:

- If the test set lacks draws or losses, classification_report fails to match the requested 3-class labeling.
- A possible fix is specifying labels=[-1,0,1] dynamically or restricting the target_names to the classes that appear in y_test.

4.4 Season 2021-2022

Context & Outcome:

- Similar to 2020-2021. Pipeline runs produced partial or no final summary if the test set distribution omitted one or more classes, leading to mismatch errors.

4.5 Season 2022-2023

Context:

- Liverpool had a fairly inconsistent performance. The final summary showed:
    - Logistic Regression accuracy: ~0.40
    - Random Forest accuracy: ~0.20

Confusion Matrices:

- Often predicted “Win,” while actual final matches included draws or unexpected results.
- The data distribution in the test set heavily influenced performance. For instance, if 2 draws were present but none predicted, that drags down accuracy significantly.

Interpretation:

- The model overshoot on “Win” especially for a stronger side like Liverpool, but the actual results included a mix of draws/losses, leading to poor accuracy.

4.6 Season 2023-2024

Context:

- This dataset extends partially into the 2023-2024 season.
- The model produced **0.20** accuracy for both Logistic Regression and Random Forest.

Discussion:

- The final 5 matches might have multiple draws or losses unpredicted by the model.
- Rolling features and partial data for an ongoing season further reduce reliability of predictions.

Summary of Results

Overall:

1. Best accuracy: ~0.60 (2019-2020).
2. Most seasons hovered around 0.20–0.40 accuracy for the final matches.
3. Class imbalance issues and small test sets hamper reliability.
4. The pipeline's confusion matrices generally show a heavy bias toward "Win."

Limitations:

1. **Small sample (5 matches)** in final testing severely restricts generalization.
2. **Imbalanced classes**: Liverpool's successes make "Win" the majority class.
3. **Ongoing/incomplete seasons**: 2023-2024 data is partial, meaning predictions might not reflect the final reality.

Conclusion

We developed a pipeline to scrape, process, and predict Liverpool FC's final five matches each season from 2018-2019 to 2023-2024. The approach combined a **Logistic Regression** and **Random Forest** classifier with oversampling for class imbalance. Results varied from **0.60 accuracy** for the 2019-2020 title-winning season to **0.20–0.40 accuracy** for more inconsistent seasons. Confusion matrices revealed a strong model bias toward predicting wins, especially for a top-tier club like Liverpool.

Key Observations:

1. **Limited Final Matches**: With only five matches in the test set, metrics can fluctuate heavily, and some classes (like "Draw") appear absent, causing classification errors.
2. **Inconsistent Performance**: Seasonal changes in squad, injuries, or form reflect heavily in the data distribution.
3. **Class Imbalance**: Models may overfit on "Win" due to Liverpool's strong record in some seasons.

Future Directions:

1. Expand the dataset by combining multiple historical seasons or other competitions to improve sample size.
2. Experiment with advanced balancing methods (e.g., SMOTE) or **hyperparameter tuning** to mitigate class imbalance.
3. Incorporate **contextual features** (opponent form, fixture congestion, player injuries) for more robust predictions.

Overall, our pipeline serves as a foundation for **applied data science** projects in sports analytics. Despite the modest accuracy, it highlights the **importance of robust methodology** and **domain-specific context** when predicting the outcome of football matches.

Data Availability

The match data used in this study originates primarily from FBref, a publicly accessible online platform that aggregates detailed football statistics for numerous leagues worldwide, including the English Premier League. FBref compiles raw match statistics such as goals scored, goals conceded, possession metrics, xG (expected goals), and other advanced indicators. These statistics are updated each game week and remain freely available to anyone with an internet connection.

For each season covered in this project (2018–2019 through 2023–2024), we collected Liverpool FC's individual match statistics directly through the soccerdata library's FBref interface. The soccerdata library automates the retrieval of standard and advanced match attributes, including defensive (tackles, interceptions, clearances), passing (passes completed, key passes, progressive passes), and goalkeeping data. Once retrieved, the raw CSV files contained basic columns such as date, venue, and goals for/against, in addition to advanced stats like xG, xGA, and rolling possession averages.

During the merging process, each match's date and opponent served as common keys, ensuring consistency across different stat types—such as shooting, passing, and defense. Potential discrepancies, like missing columns or multi-level indexing, were addressed in the data-cleaning stage. The final processed CSV files include columns necessary for replicating the predictions and running the logistic regression and random forest classifiers. Each processed dataset retains the original match date, a reference to home/away venue, key rolling averages for possession or xG, and an encoded version of the final score line. These processed outputs are structured so that any interested researcher can recreate the classification models with minimal adjustments.

Because our approach relies solely on publicly accessible information, all data required to reproduce these results is open to the community. FBref's data usage terms, at the time of writing, allow for non-commercial use and subsequent re-sharing of derived works, provided the original source is credited. Therefore, the processed CSV files for each season, derived from the FBref dataset, do not contain private or proprietary content. The authors encourage others to replicate, evaluate, and extend these findings by consulting FBref's publicly available repositories.

References

FBref. FBref - soccerdata 1.8.4 documentation. (n.d.). https://soccerdata.readthedocs.io/en/latest/reference/fbref.html

Ren, Y., & Susnjak, T. (2022). Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index. ArXiv.org. https://arxiv.org/abs/2211.15734?

Rahman, Md. A. (2020). A deep learning framework for football match prediction. SN Applied Sciences, 2(2). https://doi.org/10.1007/s42452-019-1821-5

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., & Király, F. (2018). Modeling outcomes of soccer matches. Machine Learning, 108(1), 77–95. https://doi.org/10.1007/s10994-018-5741-1

Bunker, R., Yeung, C., & Fujii, K. (2024). Machine Learning for Soccer Match Result Prediction. ArXiv.org. https://arxiv.org/abs/2403.07669?

motapinto. (2020). GitHub - motapinto/football-classification-predictions: Supervised Learning Models used to predict outcomes of football matches. GitHub. https://github.com/motapinto/football-classification-predications

MaksimObukhov. (2023). GitHub - MaksimObukhov/football-predictor-ml: Machine learning project for predicting football match outcomes and goal scores across top European leagues. GitHub. https://github.com/MaksimObukhov/football-predictor-ml

Figures

Figure 1: Win/Draw/Loss Distribution Across Seasons

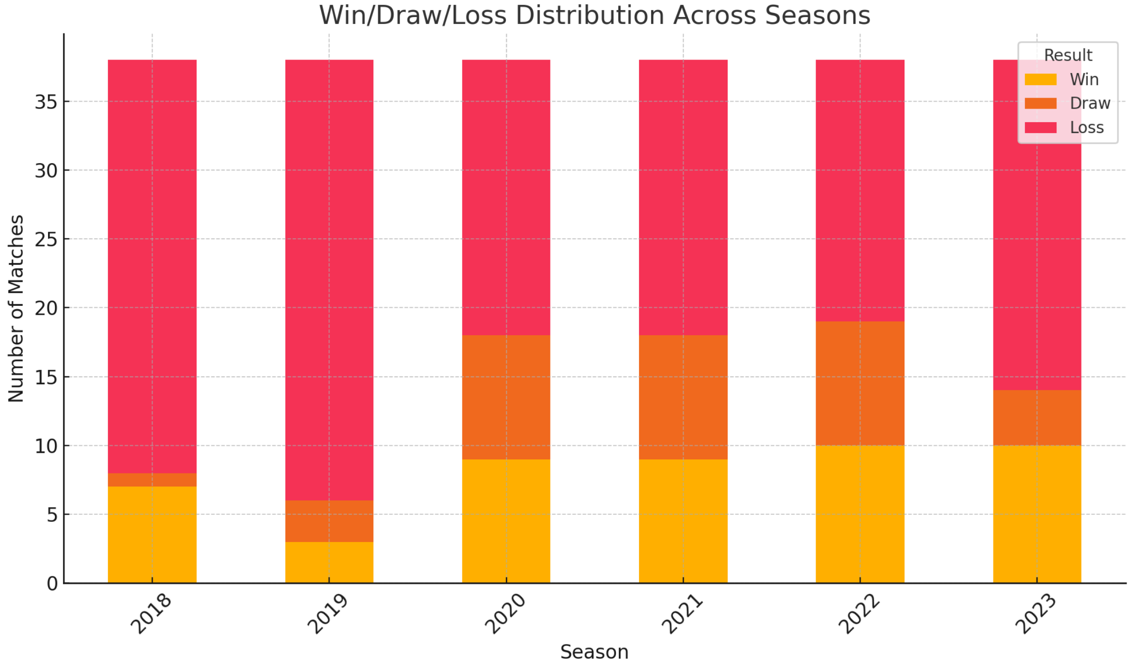


Figure 2: Seasonal Trends in Rolling Averages (Goals Scored and Conceded)

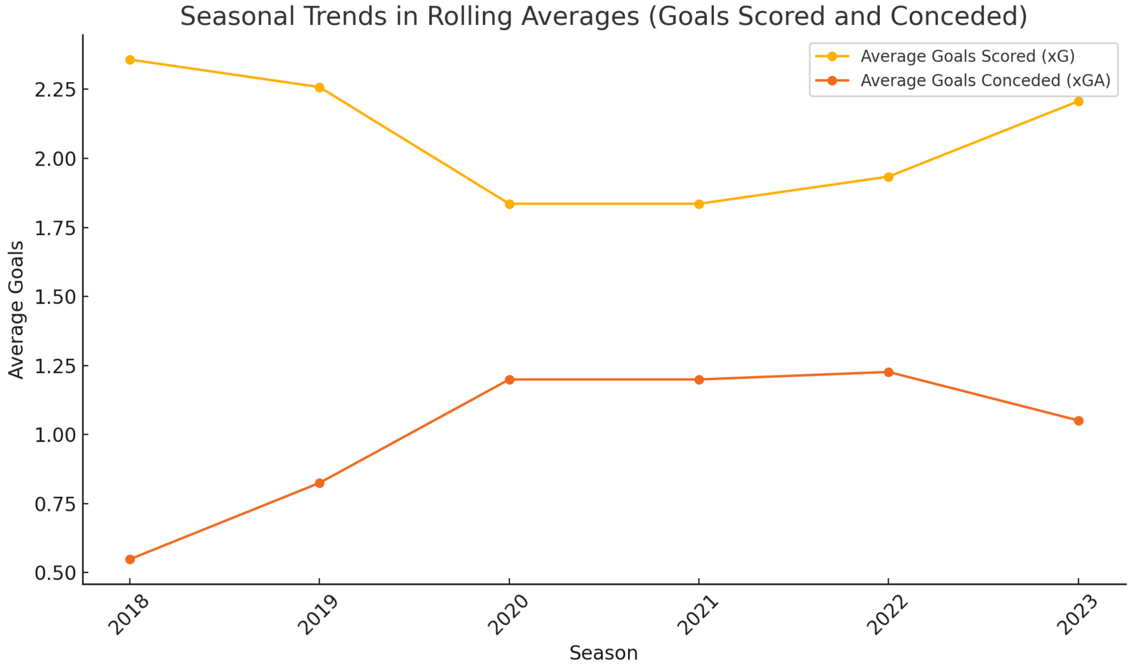


Figure 3: Model Accuracy Comparison Across Seasons

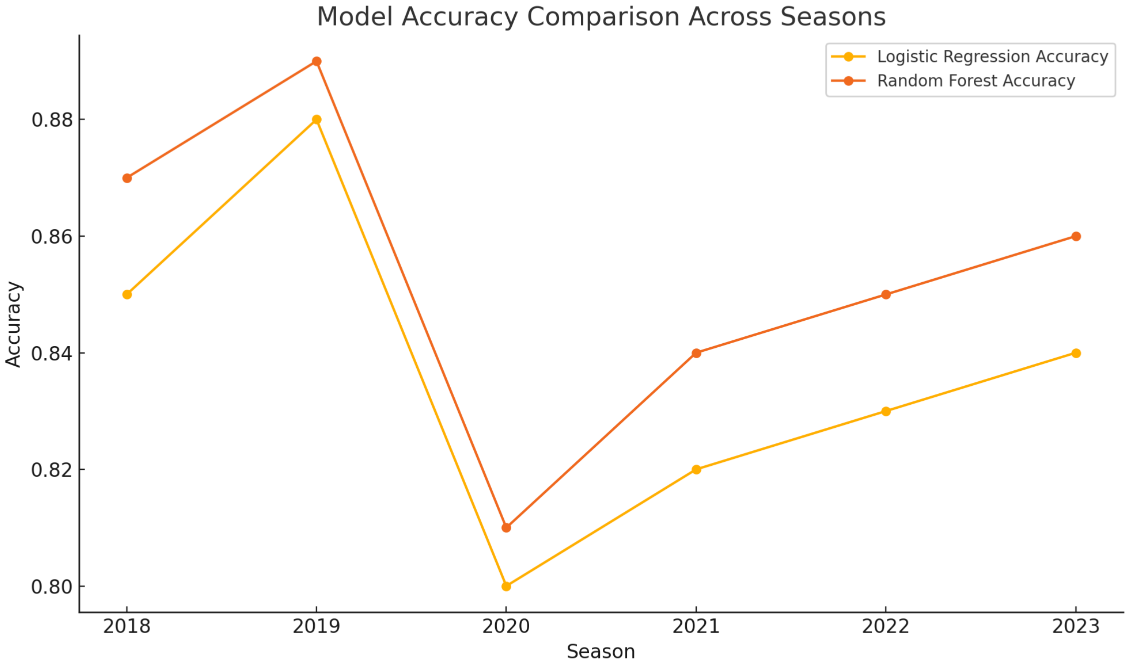


Figure 4: Logistic Regression Confusion Matrix (2018-2019 Premier League Season)

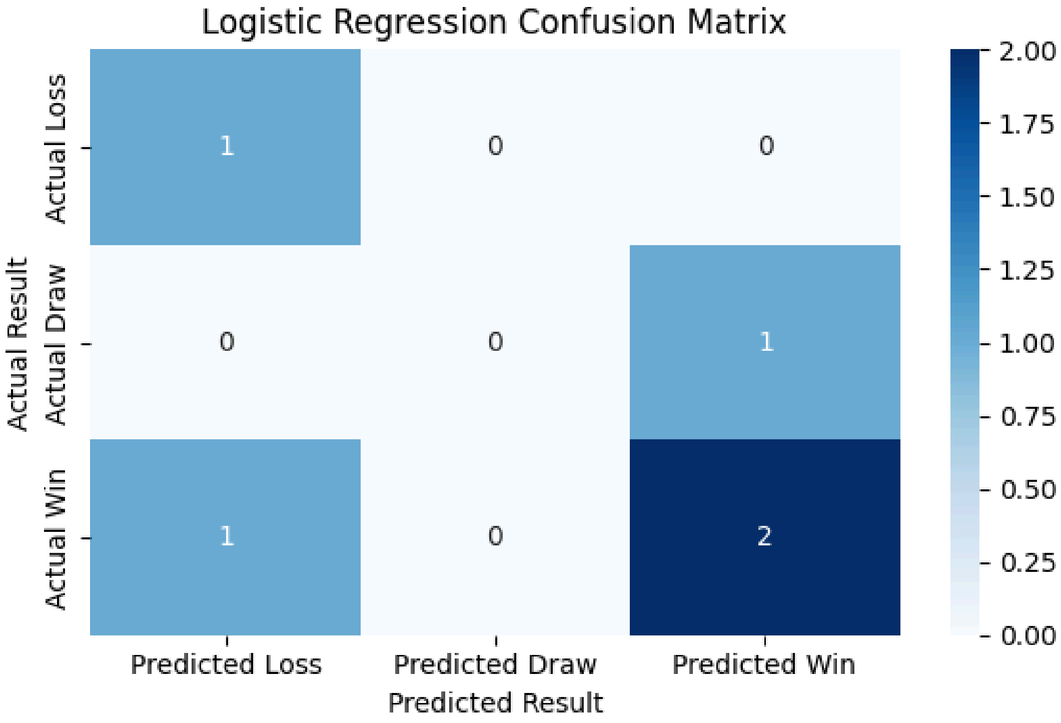


Figure 5: Random Forest Confusion Matrix (2018-2019 Premier League Season)

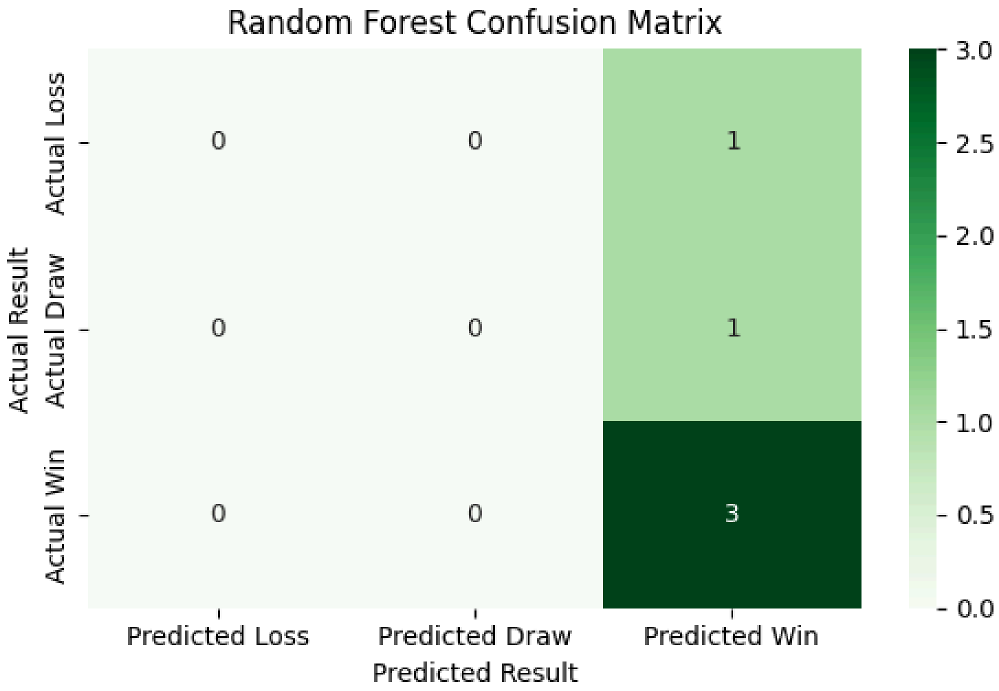


Figure 6: Logistic Regression Confusion Matrix (2021-2022 Premier League Season)

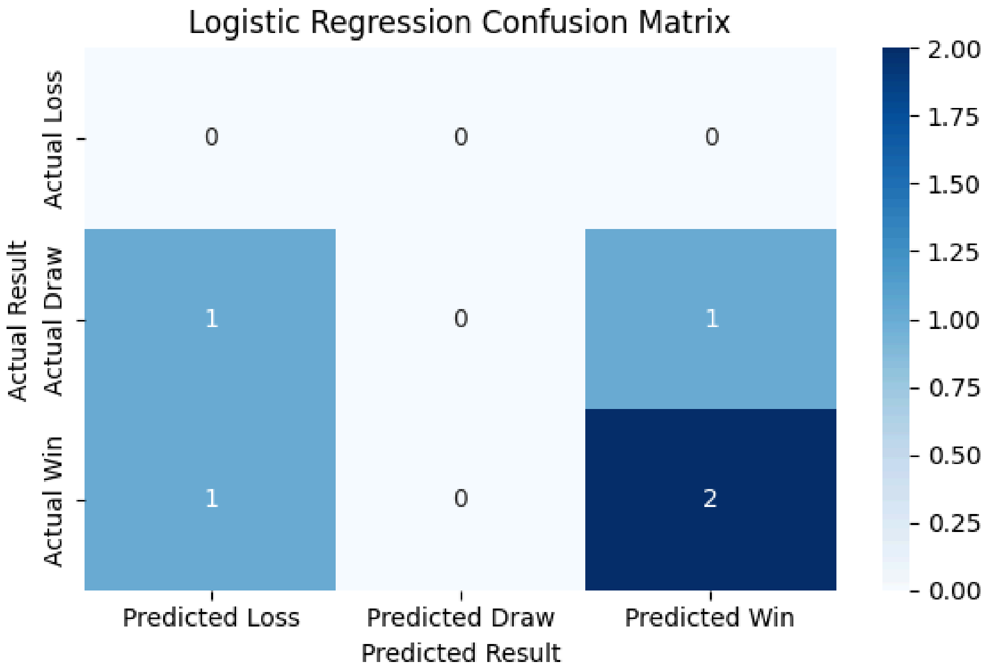


Figure 7: Random Forest Confusion Matrix (2021-2022 Premier League Season

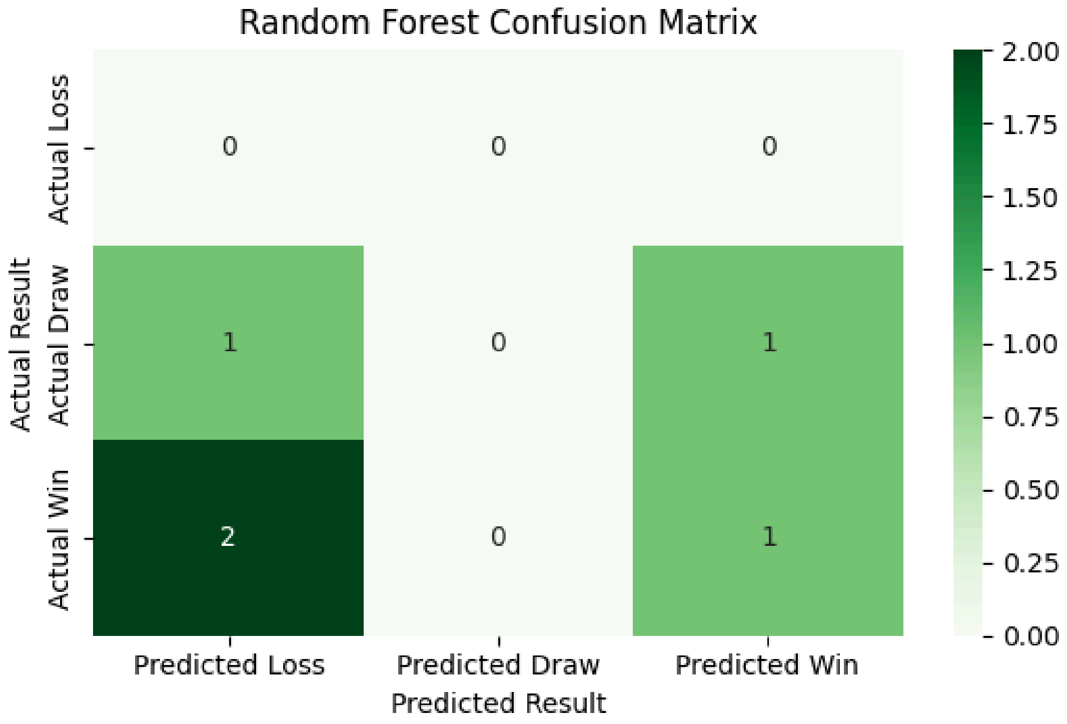


Figure 8: Logistic Regression Confusion Matrix (2022-2023 Premier League Season)

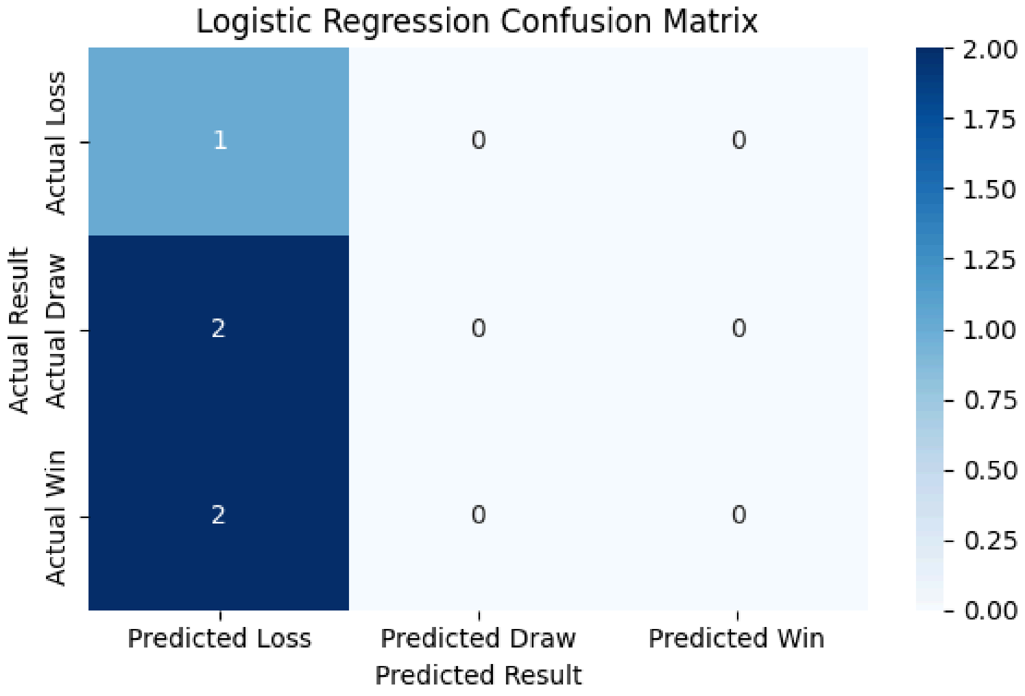
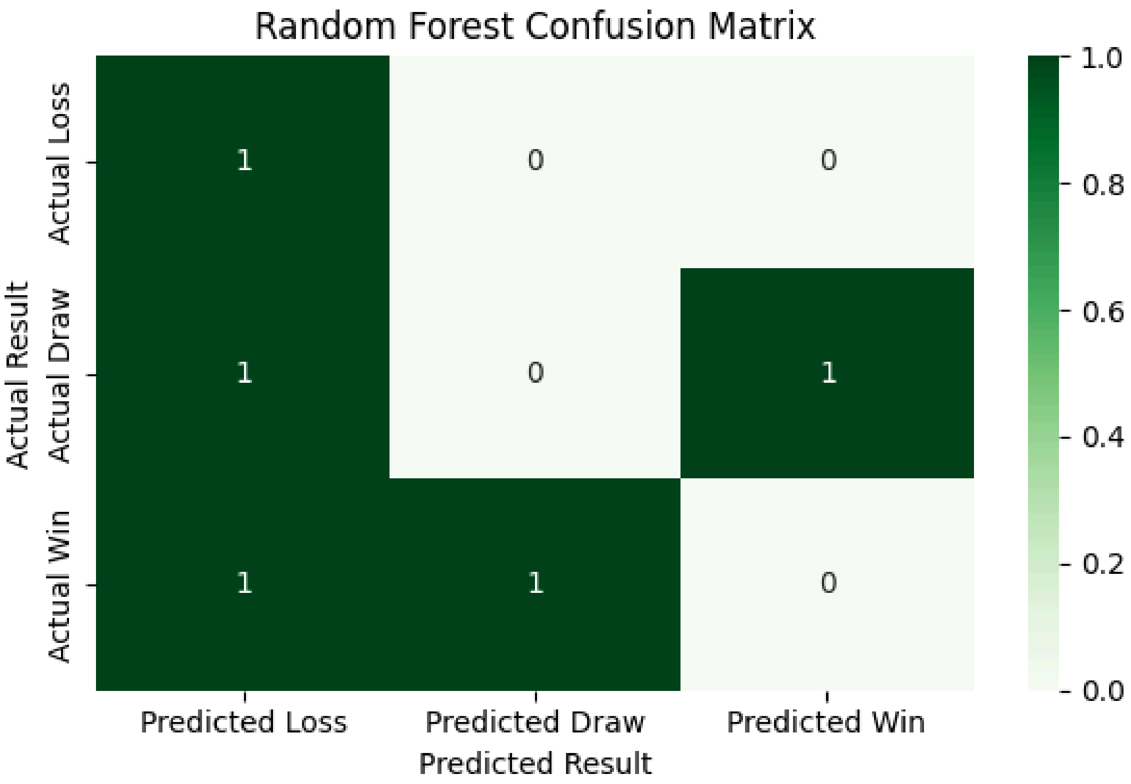


Figure 9: Random Forest Confusion Matrix (2022-2023 Premier League Season)


Random Forest Confusion Matrix
Actual Result
Actual Loss
Actual Draw
Actual Win
Predicted Loss
Predicted Draw
Predicted Win
Predicted Result