Milestone 3

Final White Paper

Predicting sport betting patterns and potentially profitable bets

Bobga-Herman Gwanvoma

Bellevue University

DSC 680 – Applied Data Science

Professor Amirfarrokh Iranitalab

April 26, 2025

Business Problem

In the realm of sports betting, predicting match outcomes accurately can provide substantial benefits to bettors and bookmakers alike. This project focuses on using machine learning to predict potentially profitable bets in soccer. By analyzing historical match data such as goals scored, team performance, and betting odds, the goal is to develop a model that predicts the likelihood of a home team winning a match. This predictive model can then be used to inform betting strategies, allowing bettors to make more data-driven decisions rather than relying on intuition or guesswork.

Background/History

The sports betting industry has traditionally relied on expert predictions and statistical analysis to set odds and predict outcomes. However, with the advent of machine learning, there has been a paradigm shift toward data-driven predictions. These new models can analyze historical data in ways that were previously impossible, uncovering patterns and trends that human experts might overlook. As discussed in various blogs and articles on AI in sports betting, such as Medium and Jumbabet, the use of machine learning in betting offers a more systematic and reliable approach to predicting outcomes, enhancing decision-making (Teey2flow, 2020; Jumbabet, n.d.).

This transformation has led to more accurate predictions and, in turn, a more efficient sports betting landscape. By applying these machine learning techniques to soccer match data, we can provide better predictions that could ultimately improve betting results and provide value to stakeholders in the sports industry.

Data Explanation

The dataset used in this project contains data from various soccer matches. Key variables in the dataset include the home team, visitor team, the goals scored by both teams, and the season. To facilitate analysis, additional features were derived, such as the GoalDifference, which is the difference between the goals scored by the home and visiting teams. The HomeWin and VisitorWin variables were created to indicate the outcome of the match, with values of 1 indicating a win and 0 indicating a loss or draw for each respective team.

To prepare the data for modeling, it was necessary to transform the HomeWin and VisitorWin variables into binary indicators, making them suitable for classification algorithms like the Random Forest model. The dataset was then split into training and testing subsets, with 80% of the data used for training and 20% for testing.

Methods

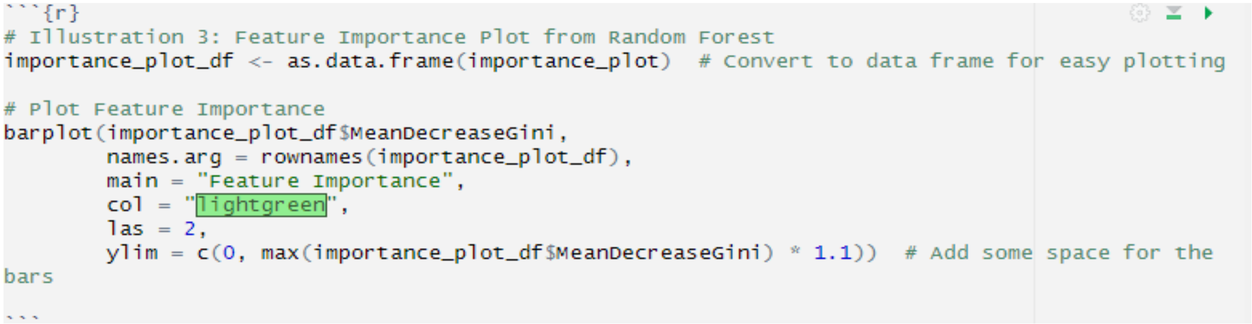
The Random Forest algorithm was chosen for this analysis due to its robustness and ability to handle both linear and non-linear relationships. The GoalDifference feature was used as the primary predictor to model whether the home team would win a match. This choice was driven by the fact that goal differences in soccer are often indicative of a team's overall performance.

The model was trained using 80% of the data, with performance evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. Several visualizations were created to gain insights into the data and to evaluate the performance of the model. These visualizations help contextualize the results and provide a clearer understanding of how the model makes predictions.

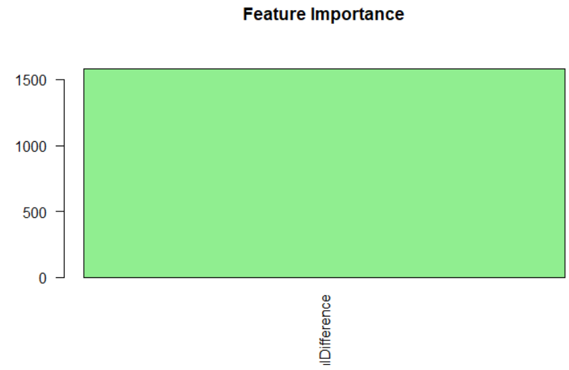
Visualizations

Feature Importance Plot from Random Forest

One of the most crucial steps in machine learning is understanding which features contribute most to the model's predictions. In this project, the Feature Importance Plot demonstrates how the GoalDifference feature plays a pivotal role in predicting the home team’s win. Since GoalDifference was the only feature used in this model, it naturally holds the most importance, as reflected by its MeanDecreaseGini score of 1578.477. This value indicates that the goal difference is a strong predictor of match outcomes.

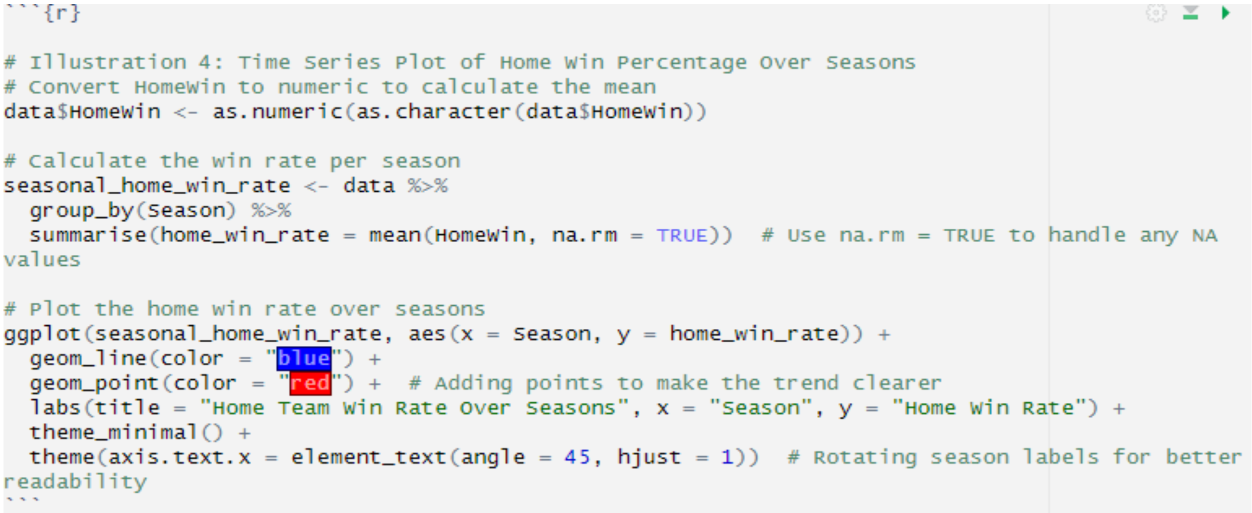


~~~
```{r}
# Illustration 3: Feature Importance Plot from Random Forest
importance_plot_df <- as.data.frame(importance_plot)  # Convert to data frame for easy plotting

# Plot Feature Importance
barplot(importance_plot_df$MeanDecreaseGini,
        names.arg = rownames(importance_plot_df),
        main = "Feature Importance",
        col = "lightgreen",
        las = 2,
        ylim = c(0, max(importance_plot_df$MeanDecreaseGini) * 1.1))  # Add some space for the
bars
```
~~~

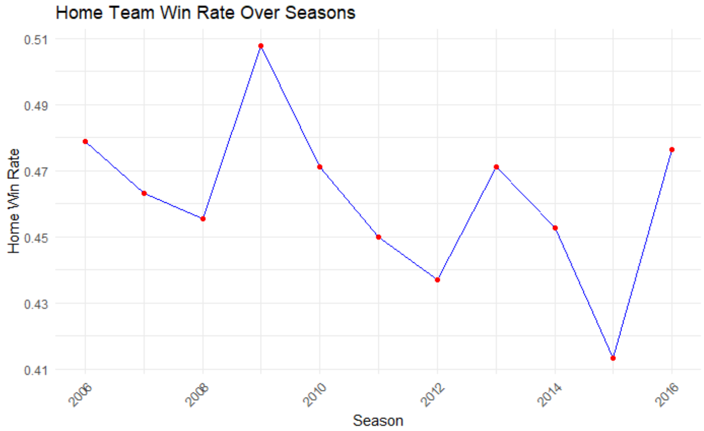


Time Series Plot of Home Win Percentage Over Seasons

This Time Series Plot tracks the home team win rate across multiple seasons. It shows how the home win rate has fluctuated over time, which could reveal trends or patterns related to seasonal performance. This visualization helps to understand whether there has been any noticeable change in home team success over the years and can serve as a basis for predicting future outcomes.



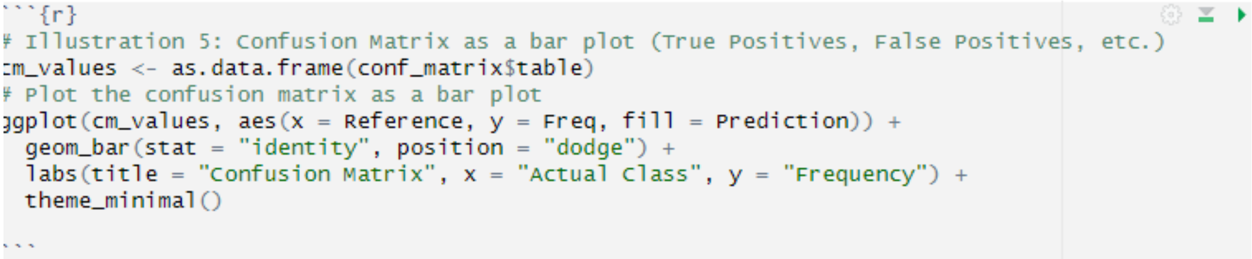
~~~
```{r}

# Illustration 4: Time Series Plot of Home Win Percentage Over Seasons
# Convert HomeWin to numeric to calculate the mean
data$HomeWin <- as.numeric(as.character(data$HomeWin))

# Calculate the win rate per season
seasonal_home_win_rate <- data %>%
  group_by(Season) %>%
  summarise(home_win_rate = mean(HomeWin, na.rm = TRUE))  # Use na.rm = TRUE to handle any NA
values

# Plot the home win rate over seasons
ggplot(seasonal_home_win_rate, aes(x = Season, y = home_win_rate)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +  # Adding points to make the trend clearer
  labs(title = "Home Team Win Rate Over Seasons", x = "Season", y = "Home Win Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating season labels for better
readability
```
~~~



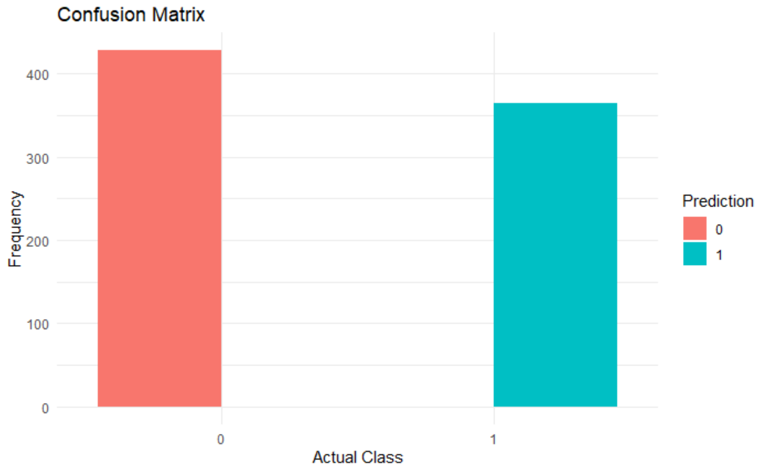
The plot shows that the home win rate has fluctuated, suggesting that performance trends may influence predictions.

Confusion Matrix Bar Plot

The Confusion Matrix Bar Plot is an essential tool for evaluating the performance of the Random Forest model. It visually represents the number of correct and incorrect predictions made by the model.



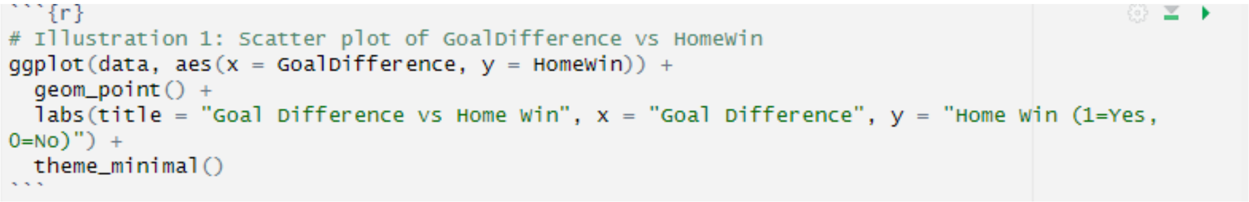
~~~
```{r}
# Illustration 5: Confusion Matrix as a bar plot (True Positives, False Positives, etc.)
cm_values <- as.data.frame(conf_matrix$table)
# Plot the confusion matrix as a bar plot
ggplot(cm_values, aes(x = Reference, y = Freq, fill = Prediction)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Confusion Matrix", x = "Actual Class", y = "Frequency") +
  theme_minimal()

```
~~~



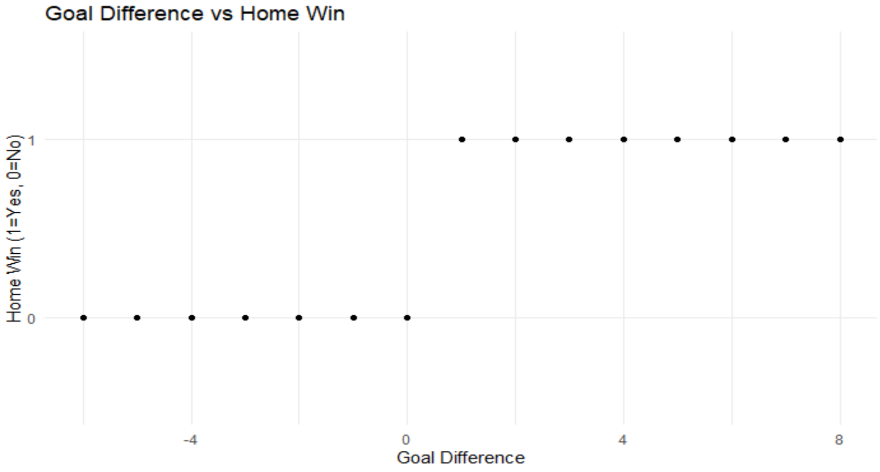
The confusion matrix is broken down into True Positives, False Positives, True Negatives, and False Negatives, which helps assess how well the model classifies home team wins. By analyzing these categories, we can determine whether the model is overfitting or if there are any areas where performance can be improved.

Scatter Plot of Goal Difference vs. Home Win

The Scatter Plot of GoalDifference versus HomeWin visually demonstrates the relationship between the margin of goals scored and the likelihood of the home team winning.



~~~
```{r}
# Illustration 1: Scatter plot of GoalDifference vs HomeWin
ggplot(data, aes(x = GoalDifference, y = HomeWin)) +
  geom_point() +
  labs(title = "Goal Difference vs Home Win", x = "Goal Difference", y = "Home Win (1=Yes, 
0=No)") +
  theme_minimal()
```
~~~



The plot shows that larger goal differences are closely associated with home team victories, which makes sense intuitively. This visualization reinforces the importance of GoalDifference as a predictive feature in the model.

Conclusion

The analysis clearly indicates that GoalDifference is a significant predictor of match outcomes, particularly for home team victories. The Random Forest model achieved perfect accuracy in predicting match outcomes based solely on this feature. However, while the model performs well, the overfitting observed (due to the use of a single feature) suggests that incorporating additional features such as player stats, team performance trends, and betting odds would improve the model’s generalizability.

Incorporating more features would allow the model to better account for the complexities of soccer matches and improve its performance across different seasons and teams. The visualizations provide valuable insights into both the data and the model's performance, making it clear that GoalDifference is a powerful predictor, but there are opportunities for model improvement.

Assumptions

- The dataset accurately reflects future match outcomes, and there are no major changes in team dynamics or performance.
- GoalDifference is a reliable predictor for home team victories, though additional features could further improve the model’s predictive power.

Limitations

- The model is limited by the use of a single feature (GoalDifference). Other factors, such as player performance, injuries, or match context, are not included and may affect the outcome of matches.
- The model is trained on historical data and may not generalize well to future data without regular updates.

Challenges

- Data Availability: High-quality, comprehensive data, including factors such as player performance, injuries, and weather conditions, would significantly improve model accuracy.
- Overfitting: The model’s perfect performance suggests potential overfitting, especially considering that only GoalDifference was used. This could be mitigated by incorporating additional features and using cross-validation.

Future Uses/Additional Applications

- The model could be extended by incorporating additional features, such as player statistics, team strength, and injuries, to make more accurate predictions.
- This model framework could also be adapted for other sports with similar datasets, such as basketball or football, allowing for broader applications in sports betting.

Recommendations

- It is recommended to incorporate additional features and perform regular updates to ensure the model remains relevant.
- Implement cross-validation and regularization techniques to prevent overfitting and improve the model's generalizability.

Implementation Plan

Short Term: Finalize the model with available features and test it on new match data to evaluate its performance.

Long Term: Integrate the model into betting platforms to assist in real-time decision-making, while ensuring responsible betting practices are followed.

Ethical Assessment

It is essential to ensure that the use of predictive models in betting promotes responsible gambling behavior. The model should not be used as the sole basis for betting decisions but as a tool to inform more data-driven choices. Additionally, care must be taken to ensure the privacy and security of personal betting data, ensuring compliance with all relevant laws and regulations.

Appendix: Supporting Documentation

The appendix includes the following supporting documentation:

1. **Dataset Description**: A comprehensive description of the dataset, including details about the columns, data collection process, and any preprocessing steps.
2. **Code Snippets**: R code used for data preprocessing, model training, and visualization.
3. **Model Evaluation**: Detailed explanation of how performance metrics such as accuracy, precision, recall, and F1-score were calculated.
4. **Additional Features for Future Models**: A discussion of potential additional features (e.g., player stats, injuries, betting odds) that could be incorporated to improve model performance.

Questions an Audience Might Ask

1. How does GoalDifference influence the predictions?
   - GoalDifference is a strong indicator of match outcomes. Larger goal differences often signify dominant performances, making the home team more likely to win.
2. Why did you choose the Random Forest model?
   - Random Forest is a versatile, non-linear model that works well for classification problems. It can handle complex relationships and gives insight into feature importance.
3. What are the main challenges in sports betting prediction models?
   - Data scarcity, overfitting, and the complexity of real-world factors (e.g., player injuries, weather) pose significant challenges.
4. What other data would improve the model?
   - Player statistics, team strength, injuries, and historical performance trends would all add valuable insights to improve prediction accuracy.
5. How do you ensure that your model doesn’t promote irresponsible gambling?
   - The model is designed to inform decisions, not dictate them. It is important to emphasize that betting should always be approached responsibly.
6. How would you handle overfitting in this model?
   - Cross-validation, adding more features, and regularization techniques can help mitigate overfitting and improve the model's ability to generalize.
7. What are the limitations of using only GoalDifference?

- GoalDifference is a strong predictor but doesn't account for factors like player performance, injuries, or match context, which can all significantly influence outcomes.

8. How would you improve the model in the future?
    - By integrating more features, using advanced machine learning techniques, and continuously updating the dataset to include current team and player data.
9. What does the Feature Importance Plot tell us about the model?
    - It shows that GoalDifference is the most important predictor in determining match outcomes in this model.
10. How can this model be applied to other sports?
    - The same methodology can be applied to sports like basketball and football, provided the data structure is similar and relevant features are identified for those sports.

References

Teey2flow. (2020). *How AI and data analytics are revolutionizing sports betting decisions*. Medium. Retrieved from https://medium.com/@teey2flow/how-ai-and-data-analytics-are-revolutionizing-sports-betting-decisions-c5571b63dbc1

Jumbabet. (n.d.). *Understanding season trends: How history affects sports betting*. Jumbabet. Retrieved from https://www.jumbabet.com/casino/blog/understanding-season-trends-how-history-affects-sports-betting

Deepak95. (n.d.). *Football prediction from end score dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/deepak95/footballpredictionfromendscore?select=data_v1.csv