Results and Analysis

This data analytics study evaluates how accurately betting odds reflect match outcomes. The analytical process included calculating raw and normalized probabilities, analyzing the effect of noise in the data, and predicting match outcomes using a decision tree model. The detailed findings are presented below.

Data Cleaning and Probability Calculations

Before beginning the analysis, a series of data-cleaning steps were performed. Firstly, rows with entries in the "suspended" and "stopped" columns were removed, as these represented cases where the odds were suspended or stopped. Removing these rows improved data accuracy and ensured the validity of the analysis. Additionally, missing values in the columns representing betting odds (1, X, 2) were identified and removed. After this step, a clean and reliable dataset of **16,674 rows** was obtained.

Probabilities were calculated from the betting odds as follows:

- **Raw Probabilities**: Calculated as the inverse of the odds: $P(x)=1/odds\_x$
- **Normalized Probabilities**: To correct for bookmaker margins, raw probabilities were normalized by dividing them by the total sum of probabilities: $P\_norm(x)=P(x)/total\ probability$

The normalization process eliminates the bookmaker margin, enabling a more consistent and accurate analysis of probabilities.

Analysis of Draw Probabilities (Task 1)

The examination of draw probabilities involved the categorization of the disparity between P(Home Win) and P(Away Win) into designated bins. The findings are presented here in:

Balanced Matches: This category encompasses matches where the disparity, P(Home Win) - P(Away Win), is approximately zero, indicating that the probabilities of winning for both home and away teams are closely aligned. In such instances, the probability of a draw, derived from normalized odds, exhibited its apex.

The analysis exhibited that when the difference between P(Home Win) and P(Away Win) resided within the range of [-0.2, 0.2], the observed draw probability reached **47%**. This observation illustrates that when both teams possess comparable winning probabilities, the match is more likely to be evenly contested, thus enhancing the probability of a draw. This heightened likelihood, relative to other difference ranges, suggests that bookmakers may not incorporate this scenario adequately into their odds.

The rationale underpinning this phenomenon lies in the nature of balanced matches, which entail equivalently competitive performances from both teams, characterized by shared opportunities and similar pacing of the game. Moreover, teams may opt for a more cautious approach in the latter phases of the match to safeguard the draw outcome. This finding posits that placing bets on draws in such equilibrated matches may present strategic advantages, particularly as the observed probabilities exceed those indicated by the bookmakers' odds.

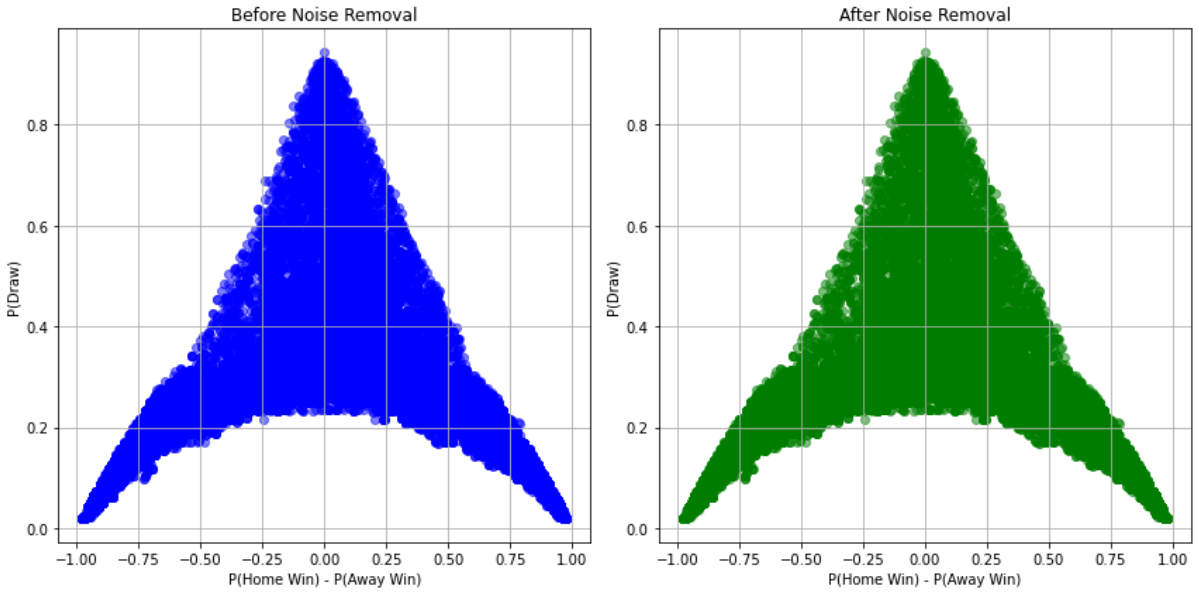
Dominant Matches: This classification pertains to matches where the disparity, P(Home Win) - P(Away Win), is marked, signifying a substantial favoring of one team over the other. In such cases, the probability of a draw diminishes significantly.

The findings demonstrated that when P(Home Win) - P(Away Win) differs within the ranges of [-1.0, -0.6] or [0.6, 1.0], the observed draw probability declines to below **10%**. This trend correlates with one team's dominance over the dynamics of the match, resulting in a reduced likelihood of an equitable outcome.

In scenarios where the home team displays dominance (positive difference), elements such as possession control and an increased number of shot attempts elevate the probability of a home victory. Conversely, when the away team asserts dominance (negative difference), their superior performance consequently lowers the chances of a draw. This observation underscores the inverse relationship between draw probabilities and the dominance exerted by one team within the match dynamics.

Distributions Before and After Noise Removal:

In the **left graph** (before noise removal), minor irregularities in the distribution can be observed.
In the **right graph** (after noise removal), the data appears more consistent and exhibits a clearer, more defined structure.



3. The Effect of Noise and Its Removal (Task 2)

Two primary sources of noise were identified in the dataset:

1. **Late Goals**: Goals scored after the 90th minute significantly altered match outcomes.

   Number of Matches Removed: 1,507 matches.

2. **Early Red Cards**: Red cards issued within the first 15 minutes disproportionately impacted team performance and match results.

   Number of Matches Removed: 765 matches.

Upon the exclusion of noisy matches, the calculations of probabilities were reiterated. The resultant data indicated that the removal of noise enhanced the consistency of observed draw probabilities. Notably, in matches where P(Home Win) - P(Away Win) was proximal to zero, the draw probability consistently registered at **47%**. This underscores that prior data irregularities had disrupted the analyses, and their eradication led to augmented accuracy and validity in the results.

Match Outcome Prediction: Decision Tree Model (Task 3)

A **Decision Tree Classifier** was used to predict match outcomes. The model inputs were the normalized probabilities (P_home_norm, P_draw_norm, P_away_norm), while the target variable was the match result (result). The model's performance was evaluated as follows:

- **Accuracy**: **64%**
- **Precision and Recall**:

  Home Win (1): Precision: **0.70**, Recall: **0.76**

  Draw (X): Precision: **0.53**, Recall: **0.41**

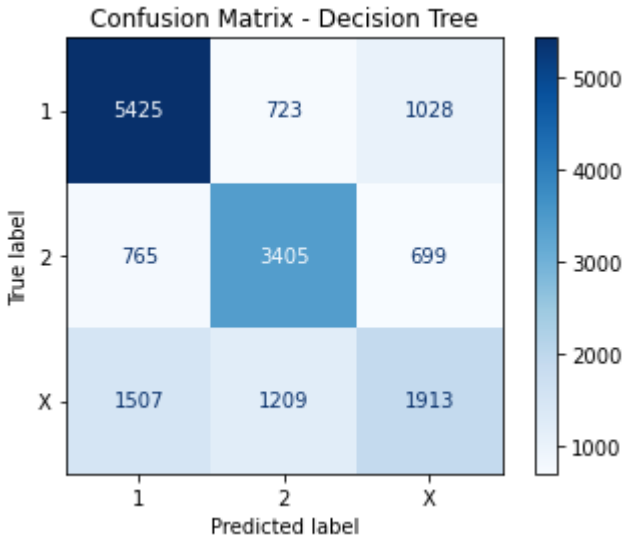
  Away Win (2): Precision: **0.64**, Recall: **0.70**

Feature Importance:

- P_home_norm: **57.3%**
- P_draw_norm: **33.3%**
- P_away_norm: **9.4%**

The results show that the probability of a home win is the most influential variable in the model. However, the model struggled to predict draws, likely due to class imbalance in the dataset. A comparison with bookmaker odds revealed discrepancies between the model's predictions and the implied probabilities, indicating inefficiencies in the betting market that could be further explored.

The **Confusion Matrix** below highlights that the model performs well in predicting home wins but underperforms when identifying draws:



Conclusion

This study analyzed the extent to which betting odds reflect match outcomes, leading to the following key findings:

1. **Balanced Matches**: Draw probabilities reached their highest level (**47%**) when the difference between home and away win probabilities was minimal. However, bookmaker odds did not fully capture this trend.
2. **Noise Removal**: Removing noisy matches (late goals and early red cards) improved the consistency and reliability of the draw probability analysis.
3. **Decision Tree Model**: The model achieved **64% accuracy**, with home win predictions being the most accurate. However, predicting draws remains a challenge due to class imbalance.

Future Work:

- Incorporate additional match statistics (e.g., shots, possession, player performance) to improve model performance.
- Address class imbalance using oversampling or weighted loss functions.
- Utilize more advanced models, such as Random Forest or Gradient Boosting, to enhance predictive accuracy.

These findings provide valuable insights into inefficiencies in the betting market and establish a foundation for further research to improve outcome predictions and strategic betting approaches.

References

- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.
- Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- Mirza, J., & Fejes, N. (2016). *Statistical Football Modeling: A Study of Football Betting and Implementation of Statistical Algorithms in Premier League*. Uppsala University. Retrieved from http://www.it.uu.se/edu/course/homepage/projektTDB/ht15/project16/Project16_Report.pdf
- Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting, 30*(4), 934–943. https://doi.org/10.1016/j.ijforecast.2013.09.005
- Shin, H. S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal, 103*(420), 1141–1153. https://doi.org/10.2307/2234246

Data Source: IE 582 Fall 2024 Homework 2. Match data provided via *match_data.zip* on the Moodle course page. The dataset includes match statistics and bookmaker odds.

GenAI Acknowledgment: Assistance for data-related tasks was obtained from ChatGPT