Capitalizing on Sport Betting Odds Biases with Machine Learning

Michael Liang

Dr. Joshua Wyatt Smith

Department of Mathematics and Statistics
Concordia University
Canada
2022

1 Introduction

The global sports betting markets have grown at a phenomenal rate in recent years due to broad deregulation, the removal of national monopolies, and, more importantly, the introduction of online gambling. Like every other industry, technology had a major influence on sports betting, in which companies have digitized and simplified their platforms to provide a pleasurable experience. Due to its economic importance and similarity to the financial market, the betting market has attracted much attention from academia, and some have studied the efficiency of the betting market. Some interesting results showed that betting odds in some betting markets failed to update their expectations.
This is considered a violation of semi-strong market efficiency. With this flaw in the betting market, this paper researches the possibility of exploiting the betting market inefficiencies by using some machine learning techniques. This research is relevant to bookmaker companies since it may give them a better understanding of the betting market and its loopholes, which might help them avoid revenue losses. Additionally, this research can help bookmakers better understand the potential risks and challenges associated with using machine learning in sports betting and how to mitigate these risks and challenges.

2 Literature Review

The efficient market hypothesis (EMH) is a theory in finance that suggests that financial markets are efficient, meaning that prices reflect all available information. Any new information is quickly and fully incorporated into prices. In recent years, the EMH has been applied to the field of sports betting, with some studies finding that sports betting markets are efficient and others finding evidence of inefficiency. For example, betting markets in Germany's first division significantly underestimated the loss of home advantage, while markets in the second division overestimated it. In addition, they found little adaptation of expectations over time in the second division and none in the first division. However, match outcomes differed significantly from bettors' opinions during the ghost games. The failure to update expectations violates semi-strong market efficiency and has persisted for some time due to unfamiliarity. This could create opportunities for profitable betting strategies.

Among ML approaches, Artificial Neural Networks (ANNs) are arguably the most widely utilized method for predicting sports outcomes. A recent study has shown that machine learning has great promise in sports betting. In a study by [Chen et al., 2020], machine learning algorithms were employed to forecast NBA game outcomes, and the findings revealed that the models' predictions exceeded traditional betting tactics. On a dataset of just 620 games, [Loeffelholz et al., 2009] employed neural network models to achieve a phenomenal accuracy of over 74%. As features, the writers used seasonal averages of 11 essential box score statistics for each squad. They also used average statistics from the previous five games and averages separately from home and away games but discovered no difference.

Extreme Gradient Boosting (XGBoost) is a popular and effective machine learning technique for supervised tasks such as classification and regression. XGBoost has been found to outperform alternative gradient-boosting algorithms and has become the go-to technique for many data science workloads. Recent studies have shown that XGBoost may be used to anticipate stock market movements, improve the accuracy of medical diagnoses, and analyse large-scale genetic data. Overall, XGBoost is a potent and adaptable tool in a data scientist's toolbox.

3 Data

This study examines profitable trading strategies for the online football gambling market using betting odds for the Premier League over a four-season period from 2019/20 to 2021/22. The data for the study can be found at `www.football-data.co.uk`. The paper focuses on three possible outcomes of any match: home victory, draw, and away win. In addi-

tion to the betting odds, the study uses data from home and visiting teams, aggregating quantitative metrics of their performance in the last ten matches. This time frame is considered important for predicting individual and club success. Aside from the betting odds, the predictive model uses match statistics data from both teams to predict the outcome of a match. Before training the model, we perform some feature engineering, such as aggregating different quantitative metrics of the team in its past ten last matches since most recent matches are often regarded as the most meaningful time windows for predicting individual and club success.

4 Methodology

4.1 Detecting arbitrage opportunities

Arbitrage chances in sports betting exist when several bookmakers provide different odds on the same event, allowing the punter to place bets with numerous bookies and ensure a profit regardless of the event's outcome. Arbitrage possibilities in sports betting can be difficult to locate and exploit since they need a comprehensive examination of the odds supplied by many bookmakers and the ability to place bets rapidly before the odds change. Furthermore, bookmakers may restrict or even outright forbid arbitrage betting, making it much more difficult to capitalize on these possibilities. Despite these obstacles, arbitrage betting may be a successful technique for those who can identify and execute these chances.

The predicted bookmaker benefit would be zero if odds were set exactly at their appropriate level by the real probabilities. Because of this, actual odds are slightly lower than fair odds to provide bookies with a profit margin. As a result, actual odds do not match genuine probabilities but rather slightly higher inferred probability (by P_i') [Vlastakis et al., 2009]. You may calculate the expected implied margin as

$$E(M') = (\sum_{i \in w,d,l} P_i') - 1 = (\sum_{i \in w,d,l} \frac{1}{d_i}) - 1 \quad (1)$$

To take advantage of arbitrage opportunities, more than one bookmaker is needed to realize an arbitrage bet. The set of bookmakers quoting odds is denoted by j. The goal of betting arbitrage is to exploit differences in the odds set by different bookmakers on the same event so that the margin is reversed to the punter's benefit, resulting in what is known as an "underground book". To accomplish this, a bettor must place a combined bet, which entails selecting the maximum odds per outcome from the set J of available bookmakers and betting on each outcome with the bookmaker who offers the highest odds for that

$$\tilde{M} = (\sum_{i \in w,d,l} \frac{1}{\max\limits_{j \in J} d_{ij}}) - 1 \quad (2)$$

where d_{ij} represents the odd on outcome i reported by bookmaker j and $\max\limits_{j \in J} d_{ij}$ represents the maximum odd on outcome i reported by all available bookmakers [Vlastakis et al., 2009].

4.2 Prediction model to detect any betting odds biases

In this subsequent step, our research will explore exploiting odds biases using the xGboost model for sports betting prediction. This model is an uncommon modeling technique in the context of sports betting, but it has been shown to be effective in other areas of machine learning.

1. Collect and analyze data on past match statistics from the 2019-2022 EPL seasons, which represent over 1520 matches. In addition, scrape from all the bookmakers the betting odds of these sessions. This data should include the odds offered for each outcome (e.g., home team win, away team win, draw) and any other relevant information, such as average market betting odds.

2. Preprocessing and feature engineering techniques are used to improve the machine learning model's performance, efficiency, and learning rate. In our case, some features may involve aggregating different quantitative metrics of a team's performance in its past N matches to create a more comprehensive picture of its current form and potential outcomes in the prediction match. The model can more accurately

predict the likelihood of different outcomes by engineering these features.

3. Train and test the XGboost model with a dozen features created from the match statistics data. While odds are a very relevant characteristic, incorporating them into a model naturally raises the undesired association with the bookmaker to forecast the chance of each result occurring based on the odds supplied by different bookmakers. Lastly, perform a grid search to find the optimal combination of hyperparameters.

4. Compare the predicted outcomes with the odds offered by different sports betting platforms and identify discrepancies that may represent arbitrage opportunities. Suppose our model's prediction of the event's probability (pM) is lower than the bookmaker's estimated probability of winning (pB). In that case, we can assume that the bookmaker has set the odds too high, and we may choose to bet on the event, creating an opportunity for profit[Hubáček et al., 2019].

5. Once possible arbitrage opportunities have been discovered, the Kelly Criterion may be utilized as a betting method to calculate the best amount to bet on a certain event. The Kelly Criterion is a mathematical formula that calculates the ideal amount to wager by considering the chance of success and the potential return. The Kelly Criterion formula is:

$$f^* = \frac{b_1 p_1 - q_1}{b_1} + \frac{b_2 p_2 - q_2}{b_2} + \frac{b_3 p_3 - q_3}{b_3}$$

where b_i is the betting decimal odds, p_i is the probability of winning and q_i are the probabilities of losing for each outcome. The overall optimal bet size is the sum of the fractions for each outcome.

5 Results & Discussion

The model's accuracy score of 60% indicates that it can correctly predict the outcome of a sports event in about 60% of cases. This is a relatively good accuracy score, but there is still room for improvement. One way to improve the model's accuracy is to optimize the hyperparameters and features used in the training process. This involves fine-tuning the model's settings and selecting the most relevant and informative features from the training data to improve the model's performance. By optimizing the hyperparameters and features, the model's accuracy score can be increased, resulting in more accurate predictions and potential profits from sports betting (consult codes for model tunning techniques).

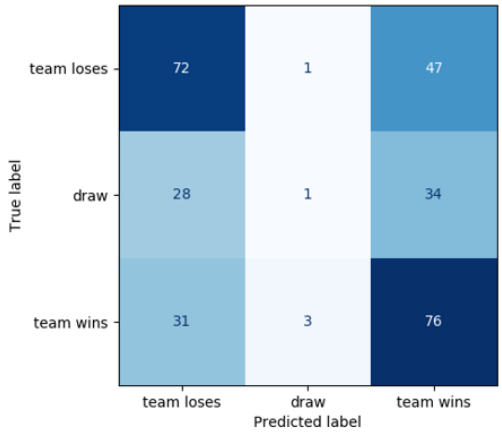


Figure 1: xGBoosting model's Confusion Matrix

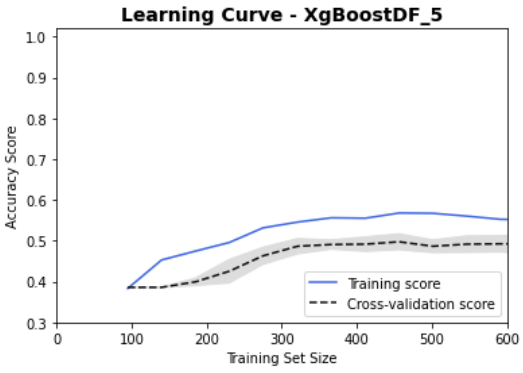


Figure 2: xGBoosting model's learning curve

After scraping the market with our method discussed in the previous section , there were about 256 matches with arbitrage opportunities with a profit margin of 0.55% (std =0.57).

The figure compares the distribution of profits from arbitrage opportunities, and all matches show that arbitrage opportunities are difficult to capture

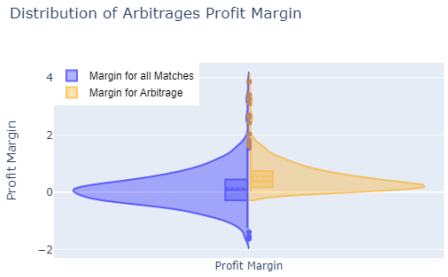
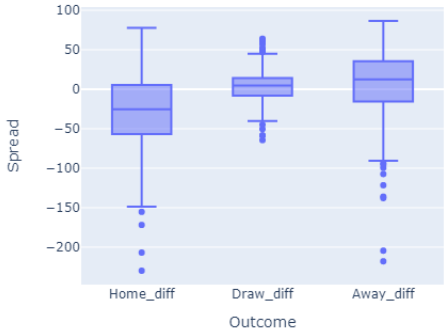


Figure 3: Profit Margin Distribution

and have minimal profit margins. The profit distribution of arbitrage opportunities follows a normal distribution with a negatively skewed shape and a fat tail. This indicates that some highly profitable arbitrage opportunities are limited in number and difficult to capture. Therefore, it is important to use advanced machine learning techniques and a robust betting strategy, such as the Kelly Criterion, to maximize profits from biases and odd opportunities in sports betting.



The figure contrasting the predicted odds and the bookmakers' odds shows the potential for arbitrage opportunities in sports betting. When the predicted odds (pM) are lower than the bookmakers' odds (pB), it indicates that the bookmakers have set the odds too high, and it may be a good opportunity to place a bet (all data points below y = x line). Due to potential inaccuracy in the model, implementing an error band threshold of 70% confidence interval can help to minimize false positive bets. By comparing the predicted and bookmakers' odds, it is possible to identify potential arbitrage opportunities and make informed betting decisions quickly.

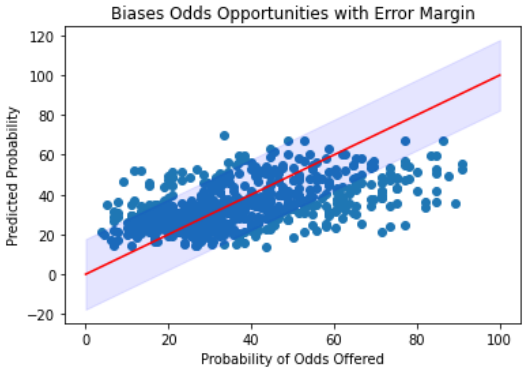


Figure 4: Probability Comparison between the model and the market

The table comparing the predictions of the xGB model trained with team statistical features and the market odds shows that there can be discrepancies between the two. In the case of Manchester City, the model predicts a lower probability of winning than the market odds, which may be due to biases in the betting odds. This discrepancy can potentially be exploited as an arbitrage opportunity, as the model's prediction may be more accurate than the market odds. The result of the match, with Brentford winning by 2-1, supports the hypothesis of potential biases (skewness) in odds, as Manchester City is the highly preferred team.

The simulation using the Monte-Carlo technique allows us to approximate the wealth of the portfolio after N simulations. We can observe that the profit fluctuates quite closely to the x-axis, which is an optimistic sign. However, much more simulations and betting is needed to have a betting accuracy of the strategy's profitability.

6 Shortfalls and limitations of the model

The main disadvantage of using the xGboost predictive model for sports betting is its weakness in predicting draw games. This is because it is harder to model, and the sample size of draw games is smaller than the other two outcomes. To improve the accuracy for draw games, it may be necessary to increase the dataset and incorporate more features,

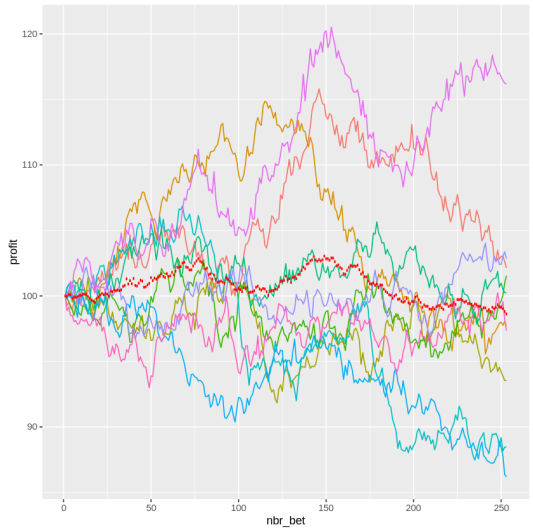


Figure 5: Estimating Kelly Criterion Profit

such as the likelihood distribution for each goal in a team. Additionally, the xGboost model is sensitive to the scale of the input data and requires careful preprocessing and feature engineering to achieve good performance. This can be time-consuming and requires expertise in data preprocessing techniques. Another disadvantage of using the Kelly criterion as a betting strategy is that it can lead to suboptimal betting decisions. This is because the Kelly criterion assumes that the bettor has perfect knowledge of the probabilities of the different outcomes, which is often not the case in real-world sports betting. Additionally, the Kelly criterion can be difficult to implement in practice, as it requires precise calculations of the probabilities and potential payouts for each event. Despite these drawbacks, the Kelly criterion can still be a useful tool for maximizing profits in sports betting. However, it is important to assess and refine the strategy to ensure its effectiveness carefully. Success in sports betting is not guaranteed, and a trading strategy that works well in one situation may not be effective in another.

7 Conclusion & Next Steps

In summary, this research uses a machine learning model and the Kelly Criterion betting strategy to exploit inefficiencies and biases in the betting market of the Premier League. The model is trained on past match statistics and betting patterns and is used to predict the likelihood of different outcomes in upcoming events. The predictions are compared to the odds offered by various sports betting platforms, and any discrepancies that may represent arbitrage opportunities are identified. The Kelly Criterion is then used to determine the optimal amount to bet on each event to maximize profits. The model's performance should be monitored and adjusted to adapt to changing market conditions and maximize profits. In the future, the same strategy could be applied to other leagues and sports to find more attractive margins.

Github link for code:
`https://github.com/mLiang99/exploiting-betting-odds-Sports-Analytics.git`

References

[Chen et al., 2020] Chen, M.-Y., Chen, T.-H., and Lin, S.-H. (2020). Using convolutional neural networks to forecast sporting event results. pages 269–285.

[Hubáček et al., 2019] Hubáček, O., Šourek, G., and Železnỳ, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2):783–796.

[Loeffelholz et al., 2009] Loeffelholz, B., Bednar, E., and Bauer, K. W. (2009). Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1).

[Vlastakis et al., 2009] Vlastakis, N., Dotsis, G., and Markellos, R. N. (2009). How efficient is the european football betting market? evidence from arbitrage and trading strategies. *Journal of Forecasting*, 28(5):426–444.