Predicting NBA Spread Using Various Models

Mason Boyles, Rikhil Fellner, Tucker Irwin

April 2024

Introduction

The sports betting industry is extremely large and has experienced a remarkable surge, with the global market reaching USD 83.65 billion in 2022 and projected to grow at a CAGR of 10.3 percent from 2023 to 2030 (Sports Betting Market Size and Share Analysis Report, 2030). Many sports bettors use their own intuition to decide on what to bet on, whereas others leverage historical data to make informed decisions. In order to create more accurate predictions, statistical methods are being incorporated into predictions for sports betting across a wide variety of sports, including basketball. Basketball is a sport that inherently generates many statistics, since scoring is frequent and relatively fast-paced, resulting in high volumes of possessions, shots, points, rebounds, and other statistical categories. The number-heavy nature of basketball analysis — combined with the tremendous worldwide popularity of leagues such as the NBA — makes it a popular candidate for gambling, especially statistically-backed gambling. The concept of "the spread," also known as the point spread, is central to NBA betting and predictions. It's a tool used by bookmakers to level the playing field, making games interesting from a betting perspective no matter how lopsided the teams may be in terms of how likely they are to win the game. By assigning a line for spread, bettors can choose to bet either that the favorite team will win by more than X points *or* the underdog will lose by less than X points / win the game. The spread has become a staple in NBA gambling, which is why we explore various machine learning models to attempt to evaluate performance on what is the best model for predicting point spread and to draw conclusions on the current state of machine learning models and their implications for sports betting.

Survey of Related Work

Prediction of spread for NBA games is something that has been attempted many times, and it will continue to be improved upon in the future as studies attempt to minimize of the margin of error for the spread prediction. This is important because sports books themselves want to have the best prediction of spread to set their lines, and individual bettors want to maximize their chance of making correct bets. One example of a study on spread is one from Bertasius et al.

from Dartmouth that evaluated the performance of models in predicting if a team will "cover" the spread (whether or not the team will make the scoring gap narrower / wider than what was predicted). This study split their features into home and away games, which we drew inspiration from for our spread predictions, but used mostly different models due to the fact that they were just classifying teams as "covering" the spread instead of predicting the spread itself. Even though the top-performing model had a success rate of around 57%, the model was still projected to result in an ROI that is 5 times that of investing in top hedge funds over the span of 1000 NBA games (Bertasius).

Another collaborative study out of the University of Auckland and the University of Texas attempted a similar prediction of NBA point spread, but this time predicting the spread itself. By looking at their results of mean absolute error and correlations for the models attempted, it was shown that the ones with the least error appeared to be linear and ridge regression, random forest, support vector machine, and gradient boosting with LightGBM, overall showing most success with ensemble methods as opposed to the linear ones attempted. Additionally through simulation of placing bets on the the safest bets (where margin between predicted spread and and posted spread is maximized) for games during 2018 it was found that the most profitable models were SVM, random forest, and LightGBM, which netted a total simulated revenue of 150,000 dollars on an initial 100 dollar investment. These results from this study were used as a benchmark in deciding which models to include and give most attention. Further, these values were used for reference in qualitatively analyzing the accuracy of the models produced in this project relative to previous work (Jayasuriuya). However, we ultimately refrained from using SVM and LightGBM due to how expensive it is to train them, finding that 5-fold cross-validation would take hours alone for each.

Data Collection and Manipulation

The primary data frame was accessed with nbastatR, a package in R that collects, harmonizes, and aggregates NBA data from a variety of sources including nbadraft.net, the NBA stats API, and HoopsHype. It was organized at the game level and includes box score and advanced statistics about both the home and away teams for every game in the 2018 season up until this year. We also joined in another data frame from nbastatR that gave us access to the dates of every game. We then downloaded these as CSV files and uploaded them to our repository. Next, we scraped data on the money line for every NBA game since the 2018 season (Historical NBA Scores and Odds Archives). This came from an archive of betUS odds that we cleaned up to include home_name, date, and home_moneyline variables. We then joined this using date and home_name to the main table to make a table called merged_table with all of the original stats as well as the home money line odds. Finally, we engineered some new variables that are generally related to the outcome of basketball games. First, we added a spread variable that we used as our target in modelling. This was calculated by the formula: $home\_points - away\_points$. We also

created a variable called home_ftr, which was calculated using the formula: $homeFreethrowAttempts/homeFieldgoalAttempts$. We also did the same for away_ftr. In addition, we decided to include pace variables in our dataset, which represents the number of possessions a team gets per minute in a game. The *home_pace* variable, the home team's pace, was calculated using the formula: $(homeFeildgoalAttempts + .44 * homeFreethrowAttempts - homeOffReb + homeTurnovers)/48$ where the numerator is the formula we used to find the number of possessions the home team had that game and the denominator represents the 48 minutes in a game. This was repeated using the away team's statistics to find the away_pace, away team's pace, variable. We believe this variable to be useful, as the pace a team plays at is an important offensive factor. We thought that the faster a team plays, the better the possibility they will take shots or score more, affecting the desired variables for prediction. Similarly, we believed that the assist-to-turnover ratio (ATR) was also an important offensive metric to consider while predicting the desired variables, such as spread. We defined this ratio as the number of assists a team gets in a game for every turnover they have. This metric was calculated using the following formula for the home_atr: $home\_assists/home\_turnovers$. This calculation was repeated to create the away_atr variable. On top of this, we were curious to see if this relationship between assists and turnovers would help us improve our models.

Models Run

We created multiple types of models using scikit-learn and then evaluated their effectiveness using 5-fold cross-validation. After surveying related work, we decided to create models using Linear Regression, Ridge Regression, Random Forests, Decision Trees, and Neural Networks. For each type of model, we used all of the numeric parameters to predict spread. In addition, we realized that in order for these models to be practical, they must be able to predict future games, and when predicting the spread of future games, statistics for those games will not be available. To account for this, we altered our table to include the average amount of each statistic for both the home and away teams. These were compiled in a table called full_averages_table which included each team's average for each statistic for each year at both home and away so that we could create predictions for each parameter when we model spreads for games in the future. Then, the moneyline parameter could be separately scraped from a sports book since it does not need to be estimated. After creating each of the models, we ran 5 fold cross validation on each of them using Mean Absolute Error (MAE) for scoring and then averaged the MAE from each fold to obtain results for each model. Figure 1 summarizes the results.

Results

Many of the models have similar MAE, with Decision Trees standing out as the most obvious model that would not be ideal for this scenario. However, we also plotted the histograms of the residuals for each model to see how the residuals

	Model	Mean Absolute Error
0	Random Forests	10.045945
1	Linear Regression	10.067568
2	Ridge Regression	10.045645
3	Decision Tree	15.325948
4	Neural Network	10.212540

Figure 1: Average MAE resulting from 5-fold cross validation for each model

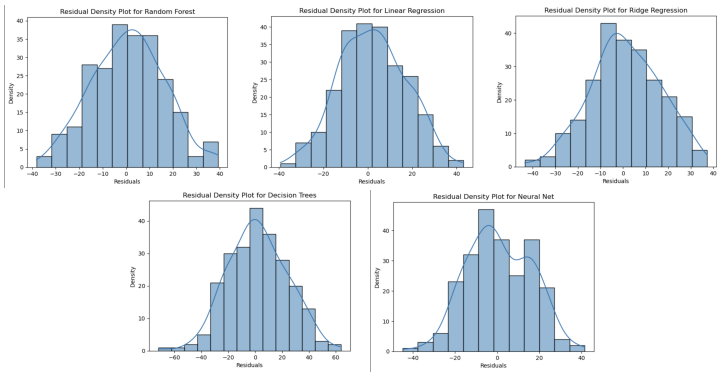

Figure 2: Histograms of the residual distribution of each model

were distributed. We saw that all models are roughly normally distributed and centered around 0, implying there was not a bias in the model. There was not significant variations between the residual plots of the models.

Conclusion

It is evident that there is still a margin of error that no model can fully minimize with current techniques. However, this can be attributed to inherent unpredictability of human athletic performance. This margin of error is why betting still exists, as if spread was reliably and accurately predictable to the point where the margin of error was very low, then sports books would no longer offer spread betting. That being said, there will continue to be work to minimize this margin of error further, and much of the room for progress with this comes with research into other parameters, such as player rest, team attitude, etc. Many of these parameters are not clearly quantified in current datasets that are available, which acts as a limiting factor for our predictions. In terms of the usefulness of the models we tested, they are still valid for those who want to set starting lines for spread and also for individuals who want to gamble based on the model's predictions. The performance of our model was very similar to the works surveyed, which indicated high profitability using those prediction models. Therefore, although there is a margin of error that is difficult to minimize, the models are still valuable in predicting NBA spread and could be used to gain an edge when setting lines or betting on spread.

Works Cited

Bertasius, Gediminas, et al. "Predictive Applications of Ensemble Methods: NBA Betting." *Predictive Applications of Ensemble Methods: NBA Betting*, www.cs.dartmouth.edu/~lorenzo/teaching/cs174/Archive/Winter2013/Projects/FinalReportWriteup/michelle.w.shu/. Accessed 30 Apr. 2024.

Historical NBA Scores and Odds Archives, www.sportsbookreviewsonline.com/scoresoddsarchives/nba/nbaoddsarchives.htm. Accessed 30 Apr. 2024.

Jayasuriya, Dulani and Liu, Jizhi Jacky and Dow, Kevin E., *Predicting United States National Basketball Game Spreads Using Machine Learning Techniques* (2024). The University of Auckland Business School Research Paper Series, Available at SSRN: https://ssrn.com/abstract=4766044 or http://dx.doi.org/10.2139/ssrn.4766044

"Sports Betting Market Size and Share Analysis Report, 2030." *Sports Betting Market Size and Share Analysis Report*, 2030, www.grandviewresearch.com/industry-analysis/sports-betting-market-report. Accessed 30 Apr. 2024.