Beating the Bookies

Zane Bookbinder, PJ Mullin, Homer LaBranche

Bowdoin College | Financial Machine Learning

Over the past few decades, sports betting has gained popularity because it offers entertainment value and the possiblity of making money quickly and easily. Our project aims to discern if statistical and gambling data can be combined to produce a model that reliability beats sportsbooks and turns a profit.

Methods

We combined two Odds datasets and one Stats dataset for this project. The main Odds dataset contains lines and odds for multiple sportsbooks including Pinnacle and Bovada and spans from 2012-2019. For games in this dataset we were able to calculate "Best", "Average", and "Worst" odds by comparing across sportsbooks. The Rotowire Odds dataset fills in the missing years from 2019-present, but doesn't break down the data by sportsbook. The Stats dataset comes from a Kaggle repository and includes a lot more information than we used.

We calculated the average statistics for each team's previous 8 games, including opponent stats (ex. average assists allowed by a team over the previous 8 games). We also added a Rest Days column and used the Geopy Python library to calculate Travel Distance from each team's previous game (0 miles if they played in the same arena two games in a row.)

We also used common NBA betting systems as manual indicators. "Bounce Back After Bad Shooting" suggests that a team will likely score a lot of points after a bad shooting game. "Fade Favorites After Blowout" suggests that after big wins, teams are over-valued and betting against them can be profitable. "The Tunnel System" relies on finding lines discrepencies between sportsbooks and betting twice on the same game. If both bets win (this happens about 12% of the time), we double our money. If only one bet wins, we lose about 5% of our original stake. We also tried the system "Overs for Three in Four," which says to bet the Over if a team has played three games in four nights. This one didn't work well at all, winning only 46.2% of bets.

We designed each model to output predicted Margins and Totals for each game. If the prediction differs from the betting line by more than the threshold (which we varied), our model places a bet using the Kelly Criterion.

Conclusions

- The profit margins are slim, and our maximum ROI was 9.8% on 328 bets (when the model was given 8000 out-of-sample games from which to find good value bets)
- With lots of statistical data from the previous eight games, we can consistently win over 50% of our bets—but barely
- Kelly Criterion increases winnings, but also increases losses
- The odds being used are crucial—for the same exact series of bets, using the best odds across sportsbooks can result in a profit while using the average or worst odds results in a loss
- Total lines vary more across sportsbooks than Spread lines, so there is more potential for profit with O/U bets

Future Considerations

The datasets we used were limited in both time-scale and informational value. For example, our models don't take into account player injuries, team vs. team history, coaching, etc. Additionally, it is possible that there is a better range of games to aggregate the data from than the past eight, as we didn't fully explore this theory. In the future, we'd like to combine more types of data to give our models more complete information, which would probably improve our results.

Introduction

Sports betting is now legal in 38 U.S. states and Americans wagered over $10 billion on sports bets in the month of February 2024 alone (American Gaming Association). Much of that money is being gambled for purely entertainment purposes, but plenty of people also attempt to make a living as professional sports bettors. Just like in the financial markets, ML techniques can be useful for predicting future outcomes. For example, the betting service sports-ai.dev claims to offer up to 17% Return On Investment (ROI) per bet.

Since sportsbooks take a profit on every bet, a win rate of at least 52.5% on Spread and Total wagers is required to profit in the long term. The lines for these bets are set originally by statisticians, but then fluctuate as people place money on both sides (much like the bid/ask model of financial markets). As a result, factors such as social media and recency bias can move the lines, leaving room for unbiased algorithms to profit.

In Dotan's paper (see References), he concludes that "much of the variance of an entire season can be derived from just an eight-game sample." We found that this method was just as effective as using statistics from a whole season, and captured recent performance well. Additionally, we calculated travel miles and rest days for each team, as we hypothesized that these factors affect team performance.

The Kelly Criterion is an equation that determines the optimal portion of money to place on a certain bet. It requires the expected win percentage (which we calculate using a tanh function based on the difference between the sportsbook line and our prediction) and the payout if the bet wins (usually between 90% and 100% of the original stake for Spread and Total bets). The equation is then:

$$f' = p - \frac{q}{b}$$

Where f' is the ideal fraction of the current bankroll to wager, p is the expected win probability, q is the expected loss probability (1-p), and b is the proportion of the bet gained if it wins.

Project Questions

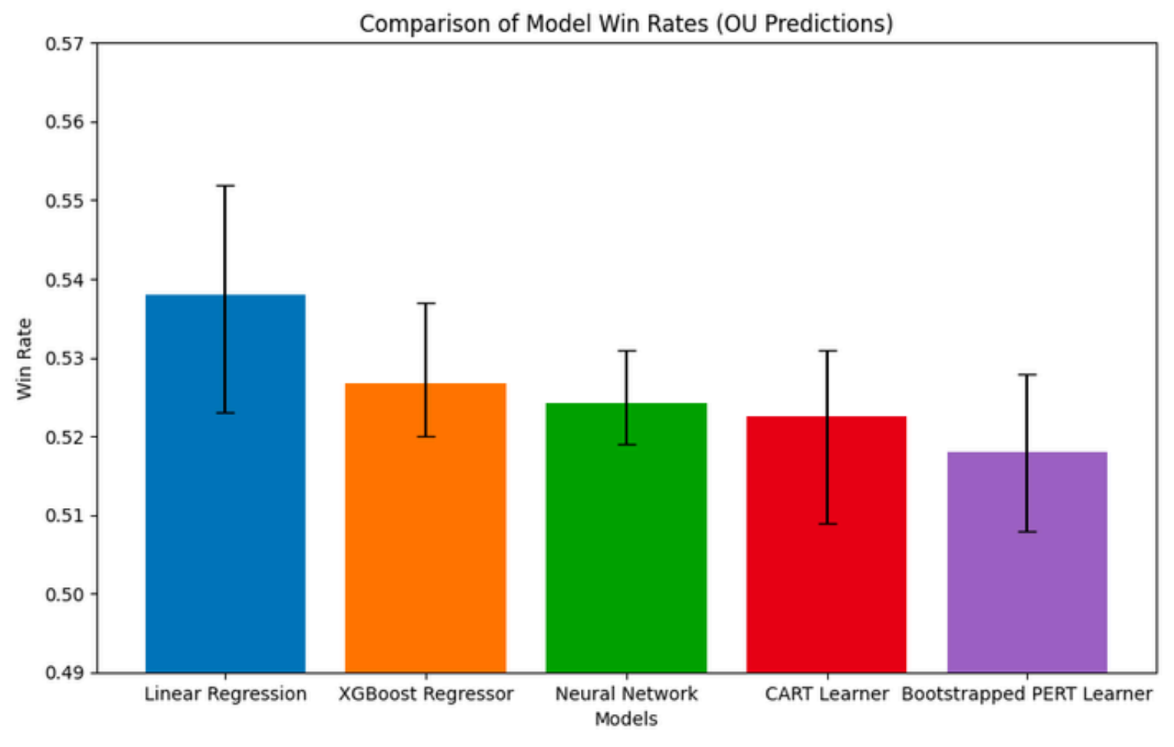
- Can we make money betting on NBA spreads and totals with 11 years of game-level data?
- Which type of ML model best predicts NBA game outcomes?
- What are the most significant features for predicting games?
- How beneficial is access to multiple sportsbooks?
- Are manual strategies/systems more or less profitable than ML models?
- Are ensemble models effective at improving predictions?

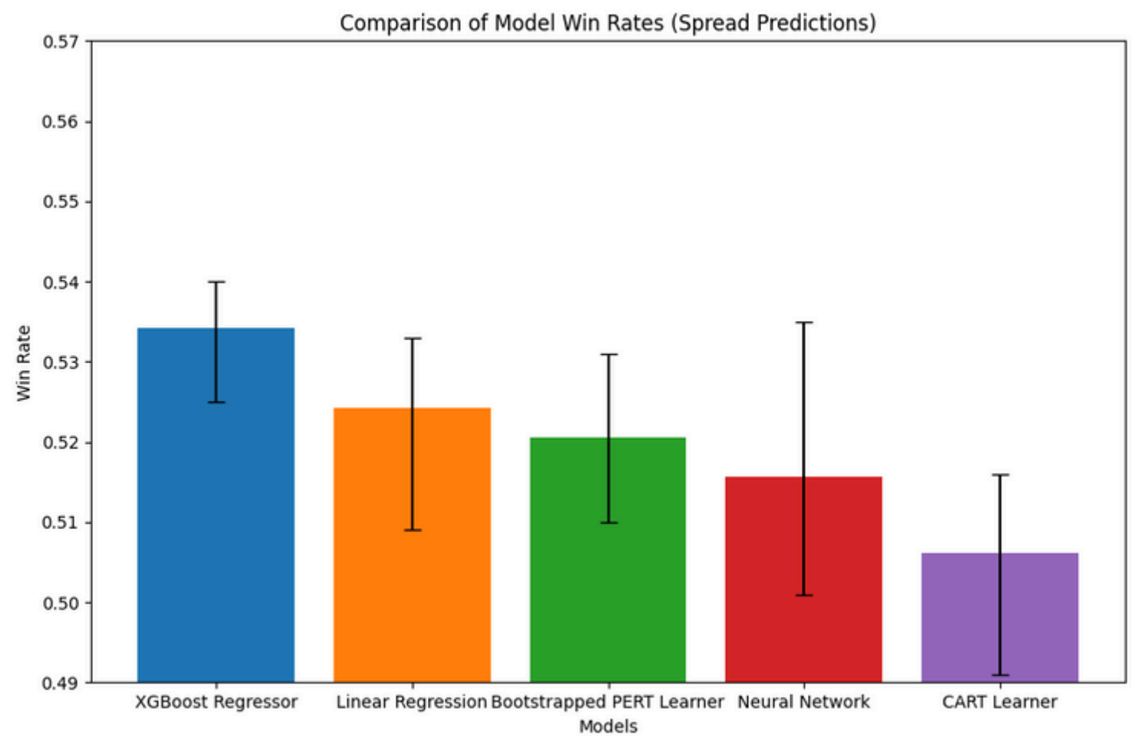
Results

We evaluated the performance of several models in predicting NBA game outcomes. Our models included:

- **Linear Regression (LinReg):** Used as the baseline model for performance
- **XGBoost:** Gradient boosting decision trees
- **Neural Network (NN):** Our network had 4 hidden layers (32, 32, 16, 8 nodes respectively). We used dropout (probability of 0.26), batch normalization, ReLu activation, MSE loss, and an Adam optimizer. The NN used batching to run through the entire training dataset each epoch, and the optimal number of epochs was found to be 200.
- **Classification and Regression Decision Tree (CART):** Standard decision tree model following the CART algorithm for features selection and splitting
- **Bootstrap Aggregation of Perfectly Random Decision Trees (PERT):** Ensemble learning using standard decision trees that split randomly

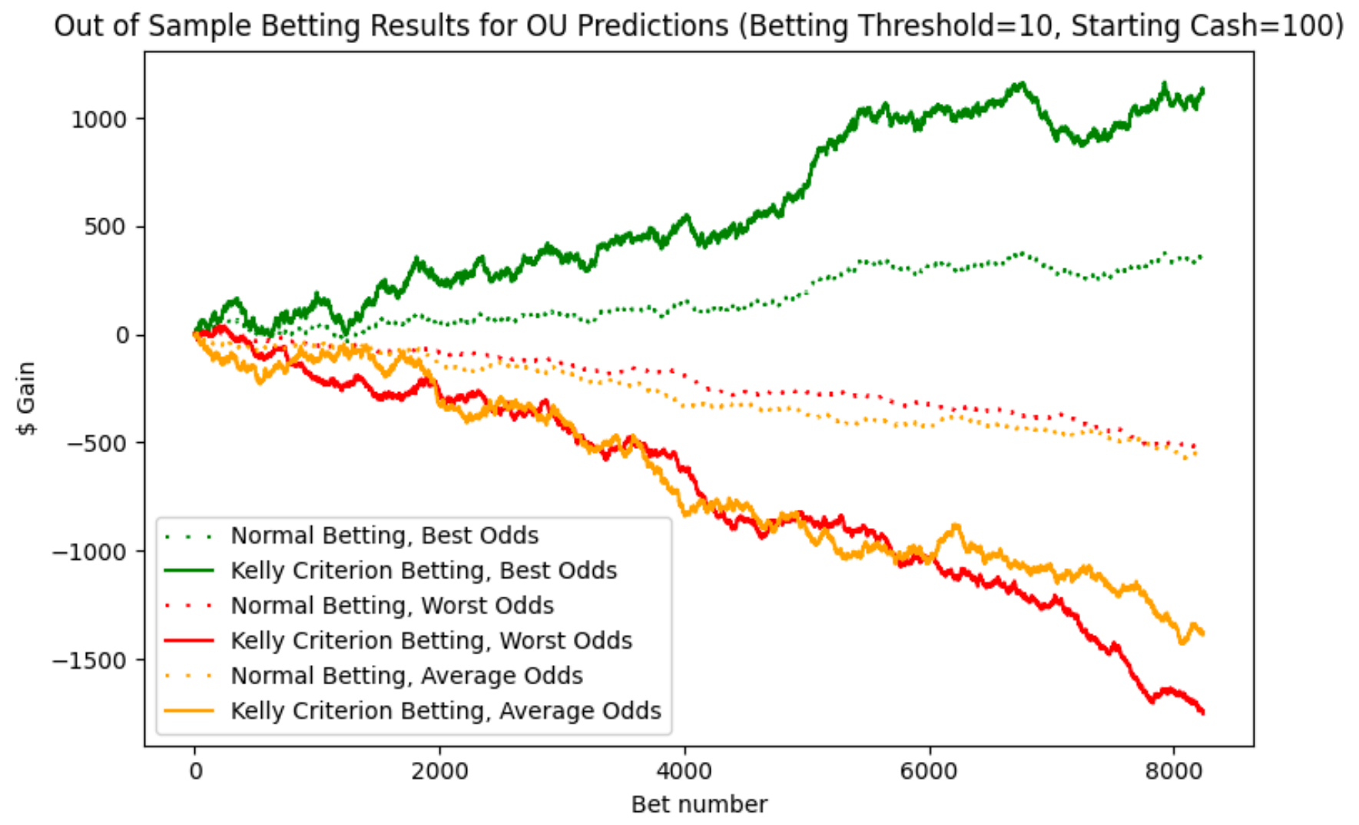
We trained all our models on the same data, used the same features for all models, and tested them on out of sample data. The primary metric for comparison was the Bet Win Rate for both Spread and Total wagers.





Linear Regression showed good consistency over all subsets of data. XGBoost proved to be competitive, but did not outperform LinReg. Despite the more sophisticated architecture and learning method, the NN performed worse than our LinReg model, further highlighting that complexity does not equal performance. Both CART and PERT performed reasonably but fell short of the three other models. None of our models outperformed basic LinReg. We suspect that this is due to the relatively-simple problem, and we think that the more complicated models might overfit the data. All models performed slightly better when predicting Totals compared to Spread, and all models were above 50% for both bet types.

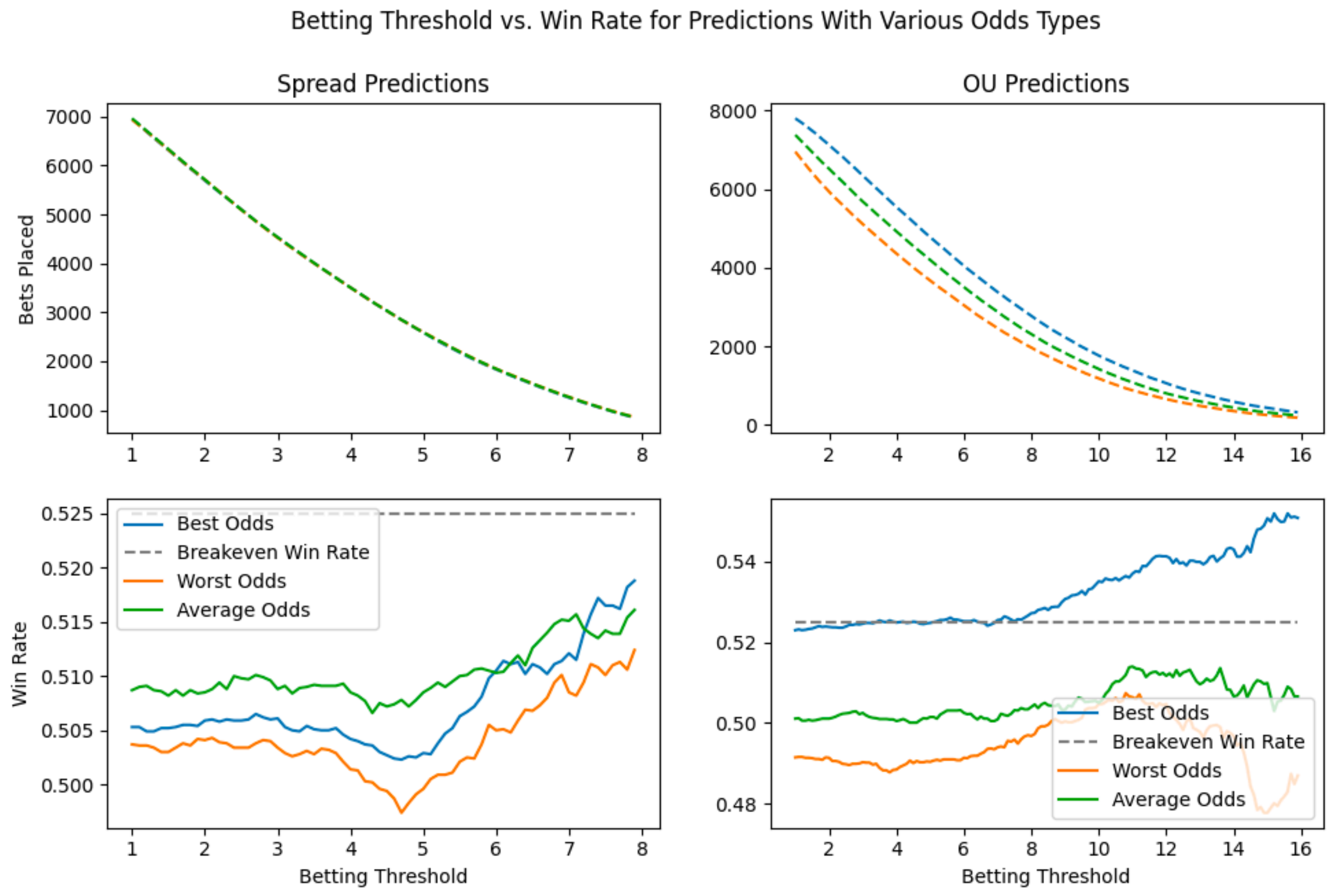
We analyzed the impact of using the Kelly Criterion compared to normal flat betting and found that while it increased gains, it also increases losses.



We looked at several manual trading strategies, shown in the table on the right. The Tunnel System (bet on both sides of an O/U) was the most effective, and only strategy that seemed able to make profit.

System/Strategy Name	Bet Win Percentage
Fade Favorites after Blowout	55.0%
Bounce Back after Bad Shooting	52.9%
Tunnel System	56.3%
Fade Teams after Scoring 135+ Points	54.7%
Over for Three in Four	46.2%

We found the most significant factor that determined our ability to make money was what odds the betting simulator had access to. With access to the best odds across all books, we found we could make money. Otherwise, all model types lost money.



References

- A Data Driven Approach to Finding an Edge in NBA Betting Markets (Reed Peterson — Carnegie Mellon)
- Predictive Models for NBA Sports Betting Report (Evan Hatton, Philip Rago, and Andrew Shanaj — WPI)
- Beating the Book: A Machine Learning Approach to Identifying an Edge in NBA Betting Markets (Guy Dotan — UCLA)
- Beating the Odds — NBA Analytics (Bodhi Nguyen and Matteo Santamaria — Stanford)
- Sports Betting: an application of neural networks and modern portfolio theory to the English Premier League (Vélez Jiménez, et al. - Cornell)

Datasets

- NBA Odds and Scores (ericqiu, Kaggle)
- NBA Game Lines (Rotowire)
- NBA Database (wyattowalsh, Kaggle)