A DECISION ANALYTICS APPROACH TO SPORTS BETTING

TUESDAY, MAY 10TH, 2022

Bennett Cohen
Senior Capstone Project
Department of Industrial Engineering & Operations Research
University of California, Berkeley

Contents

Introduction

The global sports betting industry is projected to grow to around 140 billion by 2028—excluding illegal, offshore betting—which is a huge industry itself in regulated countries[1]. In general, online (or in person) sportsbooks, often referred to as Vegas, set the payout odds given to bettors. The general rule is that if the bet is correct, bettors make some profit based on the payout odds. If the bet is incorrect, the sportsbook keeps the bettor's money. These odds are given in one of two forms and are generally used only for sports betting.

0.1 Understanding Betting Odds

American odds are the most common type of odds and are centered around betting or winning \$100. American odds greater than 0 indicate an "underdog," or that the event is less likely to occur (probability < 0.5). American odds less than 0 indicate a "favorite," or that the event is more likely to occur (probability > 0.5). For example, odds of +150 mean that a \$100 bet will profit \$150. Odds of -300 mean that a \$300 bet will profit \$100. In short, positive odds means potential profit is greater than the initial bet size, while negative odds mean potential profit is less. Under this logic, odds of -100 and +100 have the exact same result. This often called "even value" or simply "EV," which can be confused later on with Expected Monetary Value (EMV).

Decimal odds represent the total payout for every \$1 bet. For example, odds of 2.25 mean that a winning \$1 bet will be paid \$2.25. This is the total payout, including the returned money from the bet (\$1) so the profit is \$1.25. We can see that the profit is simply the decimal odds minus one. Decimal odds > 2.0 are underdogs, and Decimal Odds < 2 are favorites. This format is advantageous for calculations because all decimal odds are positive (actually, decimal odds

[1] https://www.bloomberg.com/press-releases/2021-10-19/sports-betting-market-size-worth-140-26-billion-by-2028-grand-view-research-inc

are always > 1).

The implied probability from these odds is 1/Decimal Odds. Decimal odds of 2.25 is equivalent to saying that Vegas believes there is a 44.44% chance of an event occurring. American odds can be converted to decimal odds using the following formulas[2]:

- If American odds are positive $\rightarrow$ Decimal Odds = (American Odds)/100 + 1
- If American odds are negative $\rightarrow$ Decimal Odds = -100/(American Odds) + 1

Table 1 shows the conversion between the main types of odds used for reference, but the best source for quick conversions is through an online calculator.

0.2 Types of Sports Bets

There are three main types of sports bets that every sportsbook uses. The "moneyline" is simply the outright winner of the event. In the event of a tie, the book gives the bettor their bet back. In basketball, there are no ties, so this is never an issue. The "spread" is the point difference between the two teams. Let's say the Team A is +9.5 points against Team B. At the end of the game, we take Team A's score and add 9.5 points to it. If this new score is greater than Team B's score, we win the bet. The "total" is the sum of the points scored by both teams. Bettors simply bet over or under the line (e.g. over 220.5). In spread and total bets, the "line" is the break point value for the odds, so 9.5 and 220.5, respectively. For both the spread and total, the lines are set at ½ point increments to avoid ties in the case of Team A losing by exactly nine points, or the sum being exactly 220 points. For this project, we only will focus on *moneyline bets*.

0.3 Vigorish or "Juice"

Like all gambling, Vegas creates the game such that they always have a certain advantage or edge. In Blackjack, the dealer only plays after the players have played. In sports betting, Vegas makes money through the vigorish, known more commonly as the "juice." The juice is effectively a commission on losing wagers by setting odds that violate the law of total probability. For example, total points odds are set at -110 by most sportsbooks. Using our formulas from before, we convert -110 to a probability of 52.4%. This mean that under Vegas' odds, there is both a

[2] https://www.aceodds.com/bet-calculator/odds-converter.html

52.4% probability of the sum being greater than the line and being less than the line, clearly violating the law of total probability. Practically speaking, let's say Person A bets $100 on the over and Person B bets $100 on the under. The profit for each bet is $91 using our formulas from earlier. If Person A is correct, Vegas pays them $91 and keeps the $100 from Person B. If Person B is correct, Vegas also pays them $91 and keeps the $100 from Person A. No matter the outcome, Vegas will profit. In general, Vegas changes their lines and odds throughout the day as more bettors place bets such that they can maximize the number of people on either side of the line—this will maximize profit.

0.4 Key Objective

Will all of our background information set, the final objective of this paper is to ask: Can we use classical decision analytics to build an optimal betting strategy for NBA moneyline bets?

Model Formulation

In order to model any bet, we first must introduce some notation to represent all the quantities and variables involved. We can define the following:

- p = Probability of Home Team Winning
- q = Probability of Away Team Winning
- Oj = Decimal odds of Team j to win, j ∈ h, a
- B = Bankroll (amount available to bet)
- s = Proportion of our bankroll to bet

We express our bet size as the product of our bankroll and a proportion of our bankroll (sB), such that the model can scale to any size bankroll, and we aren't limited to true dollars as our bet size. This is a very common approach in sports betting. From our earlier discussion on decimal odds, we see that $(Oj - 1)$ is our profit per dollar bet, so $(Oj - 1)sB$ is our profit from a bet of size sB.

We can then express each game as a decision tree with three possible alternatives. If we bet on the home team, and the home team wins with probability p, then we profit $(Oh - 1)sB$. But, if the away team wins, we lose our bet size of $-sB$. For the away team, the exact opposite results occur. If we don't bet at all, we neither make nor lose money.

i Method #1: Maximin

The first method we will attempt is Maximin, which tells us to select the alternative with the best-worst case scenario. We look at the worst case scenario for all decisions, and select the alternative with the best payoff of these worst cases. For visibility, we will say that our bet size is always \$1 (sb=1).

The starred options in Table 2 are the worst case scenarios for each decision. Taking the $\max(-1,-1,0)$, we find that our optimal choice is to *never bet.* Note that because we have no method to generate different bet sizes anyway, omitting sB doesn't change our result.

ii Method #2: Minimax Regret

The next method we will attempt is Minimax Regret, which tells us to find the maximum regret for each decision (the biggest opportunity we missed), and choose the alternative that minimizes this maximum regret. If we bet on the home team, the biggest regret is if the away team wins, and if we bet on the away team, the biggest regret is if the home team wins. If we don't bet, our biggest regret is not betting on the team that won, so our regret is the profit for either alternative.

From Table 2, we see that both scenarios for No bet are starred even though they might not be equal. This is because regardless of the outcome, $Oh-1 < Oh$ and $Oa-1 < Oa$, no matter what values they take on. From this table, we also see that the optimal choice is to never bet.

2.1 Maximizing Expected Monetary Value (EMV)

It's clear that neither Maximin nor Minimax Regret are able to provide strategies that allow us to make money, as both strategies tell us not to bet under any circumstances. The next approach we can consider is maximizing our expected monetary value (EMV). Using the same tree from earlier, we can perform our rollback procedure, under which we take the expected value at chance nodes, and the maximum value at decision nodes. Because p and q are unknown, and Oh and Oa are variable, we can only do this symbolically.

$$EMV = \max[pOh - 1 \quad , \quad 0 \quad , \quad qOa - 1]$$

Recall Oh is $\frac{1}{p'}$, where p' is the implied probability from Vegas. We then say that:

$$pOh - 1 = \frac{(p-p')}{p'}$$

In other words, the expected value of betting on the home team is a proportional difference between the true probability of the home team winning p and Vegas' predicted probability p'; it is a measure of how accurate Vegas is. Therefore, the betting strategy under an EMV-maximization paradigm is to bet on the team with the highest positive EMV. If both EMV values are non-positive, then we should not bet.

While this betting strategy seems simply at first, there is a fundamental issue: in order to do this, we must know the probability distributions of p and q. We don't actually know these probabilities, so must use techniques to assess them.

2.2 Method #3a: Vegas Knows Best

Vegas has the benefit of not only highly skilled data scientists, but also "wisdom of many." Vegas can adjust their odds throughout the day to reflect how people bet, leveraging the average of the bets to be their predicted probability of a team winning. This is analogous to asking 500 people to guess the number of jelly beans in a jar. While each individual guess might be off, the average guess is likely fairly close the true number. With this in mind, let's see if we can use Vegas' predicted probabilities as the true p or q and calculate our EMV from there. Recall that Vegas adds the juice to their odds, so we want to remove it by re-weighting the probabilities such that they don't violate the law of total probability.

$$p'' = \frac{p'}{p'+q'}$$
$$q'' = \frac{q'}{p'+q'}$$

Because the probabilities are re-weighted, the new p'' and q'' will always be less than the original value. We then plug our values for p'' and q'' for p and q in our EMV formula, using the fact that $pOh - 1 = \frac{(p-p')}{p'}$:

$$EMV = \max[pOh - 1 \quad , \quad 0 \quad , \quad qOa - 1]$$
$$EMV = \max[p''Oh - 1 \quad , \quad 0 \quad , \quad q''Oa - 1]$$
$$EMV = \max[\frac{(p''-p')}{p'} \quad , \quad 0 \quad , \quad \frac{(q''-q')}{q'}]$$
$$EMV = 0$$

Because $p'' < p'$ and $q'' < q'$, the first and third elements of this array will always be negative. Therefore, we say the best approach is to not bet. Intuitively, this should make sense. Vegas sets the odds such that the expected value is always negative for bettors. Similar to most other gambling games, in the long run, the expected value is such that we will always lose money. Our strategy recognizes this and tells us never to bet.

2.3 Method #3b: A Machine Learning Approach

In short, we need a better method to assess the probabilities of p and q other than what Vegas is giving us. To do this, we will implement machine learning models to predict $\hat{p}$ and $\hat{q}$ and use these predicted values as the true p and q to calculate our EMV. More specifically, we will use a classifier to predict the winner of each game and use the predicted probability values instead of labels. The majority of this is out of scope of the class, so I will not delve too much into detail on these models. I have already written another paper on the topic, which I will attach to this report.

The input statistics for all tested models were:

1. Season Stats: Box score statistics for every player and every team for all games played in a season

2. Recent Stats: The same as season stats, but only in the last 10 games. This allows us to capture momentum and hot/cold streaks well.

3. External Features: Precompiled Elo ratings, injury statistics, home court advantage, etc

I trained the models on data from 2015 to 2019, and tested it on the 2021 season. I skipped the 2020 season due to COVID, where statistics for players had much higher variance, and win percentages were highly abnormal. The baseline accuracy to which I compared performance to was Vegas' accuracy, or how often the favorite won, which is 63% in most seasons. My best performing model is $\approx$ 67% accurate. While this difference seems small, we must take note of the fact that not all games are thought of equally. Predicting the best team to beat the worst team in most games takes very little data or insight. Predicting very close games and predicting underdogs is what will drive profit. Recall that we want to maximize the difference between the true probability of a team winning (what we will predict) and Vegas' predicted probability. The higher this value, the higher the profit. Table 3 shows the performance of various models in accuracy, true positive rate, and false positive rate.

Three key insights were discovered from this part of the project. First, more data isn't always better due to changes in the style of the game in recent years. Second, the best models were ensemble methods (XGBoost and blending), and deep learning isn't applicable here. Third, the best predictors of outcomes are advanced and aggregate statistics. For a deeper understanding of these models, please refer to the attached paper.

2.4 Bet Sizing using Kelly Criterion

In the first notation, we denoted sB to be our bet size, but in subsequent models, have only ever used a constant bet size. This is not the ideal scenario, and is best understood through an example. If we predict $\hat{p} = 0.6$ for two games, but Oh for Game #1= 2.5 (+150) and Oh for Game #2= 2.0 (+100), should we bet the same on each game because $\hat{p}$ is the same?

$$EMV1 = (2.5 - 1)(.6) - 0.4 = 0.5$$
$$EMV2 = (2.0 - 1)(.6) - 0.4 = 0.2$$

This means that, in the long run, if we place the bet on each game over and over, we will make $0.50 on a dollar bet from the first game, and $0.20 on a dollar bet from the second game. Clearly, a smart strategy would be to bet more on Game #1 than Game #2. More generally, we can say that as EMV increases, our bet size should increase as well.

We will do this through the Kelly Criterion, which is a common investment sizing strategy based on an optimization of a logarithmic growth of our bankroll[1]. The Kelly Criterion says our bet size s should be set such that:

$$s = \frac{Oh * p - q}{Oh}$$

Recall that Oh is the total payout from a bet of $1 so the numerator of this is the expected total payout from making a certain bet. Therefore, the Kelly Criterion effectively tells us to bet a ratio of the expected total payout over the odds set by Vegas. Note that if the expected payout is negative, so is the expected profit, meaning this bet size would be negative, indicating we shouldn't bet. Even further, our bet size is a ratio of our confidence in an event versus Vegas' confidence. It matches our forecasting model against Vegas' head to head. We should note that this method often results in very large s values, sometimes up to 40% of our bankroll. A bet size that large is not realistic, so the canonical method is to scale the bet size by some constant factor, thereby implementing a Fractional Kelly Criterion.

[1] https://www.eecs.harvard.edu/cs286r/courses/fall12/papers/Thorpe$_{K}elly Criterion$2007.pdf

Data Analysis Methods

In order to test these methods in placing actual bets, we will simulate each strategy across the 2021 NBA season, which is approximately 1200 total games, or 1200 potential bets to place or not place. We will begin each simulation with $1000 in our bankroll, and scale all bets to 5% of the true Kelly Criterion size. In addition to the methods we discussed, we also will analyze betting only on favorites and betting only on underdogs.

Three main data sources were used for these methods:

1. NBA statistics and outcomes were pulled directly from NBA using their internal, but public API[1].

2. Historical odds for the simulation were pulled from a well-respected odds archive online, which stores opening and closing lines for various sportsbooks[2].

3. Live odds were pulled from The Odds API, which handles requests to nearly every sportbook in the world, as each book has different odds at different times[3].

[1] https://github.com/swar/nba-api
[2] https://www.sportsbookreviewsonline.com/scoresoddsarchives/nba/nbaoddsarchives.htm
[3] https://the-odds-api.com/

Results

For simplicity and to best understand the results, we will look at simulations sequentially, and analyze each. In Figure 2 (Simulation Results #1), the first the thing to note is the blue horizontal line, which represents the strategies for Maximin, Minimax Regret, and Maximizing EMV using Vegas' probabilities. Because all of the strategies says to never place a bet, we keep our constant bankroll of $1000 the entire season. The red line shows always betting the favorite, which is a common strategy that people generally believe will profit a small amount of cash. However, because Vegas is only correct 63% of the time, in the end, betting on the favorites means we eventually lose our money. The orange line is betting on the underdogs for every game, which is clearly a more volatile strategy as we will occasionally bet on a horrible team so if they win, we make a large profit, only to likely lose it soon after.

Let's now take a look at using our machine learning approach. Figure 3 (Simulation Results #2) shows our bankroll following a strategy of using the predicted probabilities as the true probabilities, and betting on the team with the highest positive EMV. We see that while it's better than both betting on the favorite or underdog only, it isn't actually a profitable strategy. We can see that it actually follows a similar pattern to betting on the favorite for the first half of the season, but as it learns, it improves its own predicted compared to Vegas'.

However, it still isn't a profitable strategy, so let's try to understand why. Assume for a given game we have predicted $\hat{p} = 0.25$ and $Oh = 5.0$, which is equivalent to +400 or 20% probability according to Vegas. Using our EMV formula, we find our $EMV = (5.0-1)(.25) - .75 = 0.25$, but this bet is still too risky. If a team only has a 25% probability according to our model, is it worth it to risk our money on this game, even if the payout is very high (we would profit 4x our bet size here). So how can we address this?

Figure 4 (Simulation Results #3) shows what implementing a utility function $U(x) = x^{\frac{1}{3}}$ does. Instead of maximizing EMV, we now choose the alternative with the highest positive utility.

$U(0) = 0$ so if both utility values for the bets are negative, then we won't bet. We see that this model does not solve our problem; we still end in ruin. Various utility functions were tried, but most resulted in a value similar to this. The reason is more simple than a utility function: if we don't think a team has any significant probability of winning a game, why would we bet on them?

So, let's now try to set a threshold for the predicted probability. Under this revised strategy, we will bet on a team if and only if both of these conditions are met:

1. The EMV of betting on the team > 0
2. The predicted probability $\hat{p} > 0.55$ if a home bet or $\hat{q} > 0.55$ if an away bet.

Finally, we can see that in Figure 5 (Simulation Results #5), the new strategy is a significant improvement over any of our last strategies. Not only does it make some money, but it nearly triples our bankroll by the end of the season, and peaks at 4.5x our starting bankroll. The model loses money towards the end of the season due to motivation. If a very bad team has no chance of making the playoffs, they will actually want to lose games to secure a better position in the NBA draft so that they can get better players next year. Our model does not account for that, but Vegas does.

Table 4 shows an overview of the number of bets placed and the accuracy the bets placed. We see the maximum number of possible bets to be 1,267 and our best performing model places the least amount of bets. More interestingly, we see that the bet accuracy is very close for the Favorite Only, Utility, and final models, yet only one of them ends up profiting. This is why these high level statistics do not tell the whole story. Different games have different values, so we must tune our model to be very good at high value games.

Discussion

0.1 Expected Value of (Im)Perfect Information

In order to understand the significance of our model, we can first look at the Expected Value of Perfect Information (EVPI). How valuable is it to know the outcome of the game before it happens? We first can reconstruct the decision tree to show that we know the winner and make the decision after (i.e. chance node before decision node). We use the rollback procedure to then calculate the EMV of the perfect information (PI) tree.

$$PI = \max[Oh-1 \quad , \quad Oa-1]$$
$$NI = \max[pOh-1 \quad , \quad 0 \quad , \quad qOa-1]$$
$$EVPI = PI - NI$$
$$EVPI = \max[Oh-1 \quad , \quad Oa-1] - \max[pOh-1 \quad , \quad 0 \quad , \quad qOa-1]$$

Recall that p and q are unknown, and Oh and Oa are known. The best way to understand this is through a heat map that has the values for Oh and Oa on the axis, and is filled with the EVPI using the formula above. We can look at an example where $p = q = 05$ in Figure 7. Here, teams are truly evenly matched. We begin in the top right corner (where odds for both teams are 1.0, which is the true 50% probability), but as soon as either team's odds increase, the EVPI increases dramatically. Intuitively, if two teams are evenly matched, but we know who is going to win, we will always bet on the team with the higher odds, because higher odds means higher payout. Thus, the value of knowing this will increase because our potential profit would increase. This is directly tied to the Minimax Regret from Method #2, because our biggest regret for not betting is not betting on the team that won.

Clearly,our best models all include a forecast, so it's also interesting to understand just how valuable our forecast is. First, we can create a third tree to represent a prediction Fh or Fa as predicting home or away.

We can consider the same evenly matched game from before, and calculate the resultant probabilities. For ease, we will say the our model is accurate $\frac{2}{3}$ for both the home and away team:

1. $P(H|Fh) = \frac{2p}{2p+q} = \frac{1}{1.5} = \frac{2}{3}$
2. $P(A|Fh) = \frac{q}{2p+q} = \frac{.5}{1.5} = \frac{1}{3}$
3. $P(H|Fa) = \frac{2q}{2q+p} = \frac{1}{2p+q} = \frac{2}{3}$
4. $P(A|Fa) = \frac{p}{2q+p} = \frac{1}{2p+q} = \frac{1}{3}$

Thus, our Expected Value of Imperfect Information can be represented below, and we once again use the heat map to represent the EVII in Figure 9.

$$EVII =$$
$$\frac{2}{3}[\max[\frac{2}{3}(Oh-1), 0, \frac{2}{3}(Oa-1)]] + \frac{1}{3}[\max[\frac{2}{3}(Oh-1), 0, \frac{2}{3}(Oa-1)]] - \max[pOh-1, 0, qOa-1]$$

We can see something that is quite different from the perfect information heat map. For an evenly matched game, if our predictions are roughly $\frac{2}{3}$ accurate, the expected value of having this forecast only increases the most as both ends increase towards infinity. More specifically, the bottom right quadrant is brighter than top right quadrant, indicating that if home odds are very high, and away odds are very low, the EVII is greater than if away odds are very high and home odds are very low. This is an interesting quirk that doesn't have a clear explanation at the surface, though I suspect has something to do with capturing home court advantage. This is definitely a concept to explore later on.

0.2 Issues, Improvement, and Limitations

While the results of the project thus far are extremely promising, there are a few things we didn't explore fully, mostly in in the interest of time. First, we only tried a few combinations of thresholds for $\hat{p}$ and EMV. In the future, we will simulate across a grid of hyperparameters and choose the values that maximize our bankroll or stability of bankroll. With this, we might end up overfitting to the intricacies of the 2021 season alone. As the 2022 season is ending soon, we can simulate across this season for a larger sample size. A more robust solution would be to implement a dropout, similar to in neural network regularization, such that for each game in the simulation, we skip the game with a designated probability. If there are certain games where we always make a lot of money due to a Vegas error, we can remove it. A large problem is that the bet size scales too quickly, leading to higher variability. Towards the end of the season, the bankroll increases quickly towards $4,500 and then quickly drops down. This is because our bankroll has increased so quickly that our bet sizes increase from a range of $50 to a range of $250. This is not realistic as we won't actually take on this risk in real life. Our approach to solving this can be to "delay" the Kelly Criterion by changing the scaling factor by a proportion of the bankroll today compared to the bankroll from a look back window of seven days. This will let us automatically moderate the bet size scaling. The last improvement we could make is to the model itself,tuning it to perform better on the high value game, but those methods are out of scope of this class.

Conclusion

Overall, it is clear to see that while the simple strategies from earlier in the semester limit us from profiting, introducing a robust forecasting method with an EMV-maximization paradigm allows us to profit greatly. Instead of introducing risk via a utility function, we can implicitly define it by thresholding our betting strategy on either the probability of an event occurring, or on the EMV (such that EMV > Z > 0).

The next steps that can be worked on is the introduction of a reinforcement learning algorithm to this problem. We can model the simulation as a Markov Decision Process (MDP) from one bet to the next, as the only information that matters to our agent is the incoming data for the current game, and how much money we can bet.

1. State will be a dictionary of all the cumulative stats that went into the initial classifier model, along with the bankroll and the odds set by Vegas.

2. Action will be a 1x3 vector, $[bh, ba, s]$, where bh and ba are binary variables for betting on the home or away team, and s is the proportion of our bankroll to bet. If $bh = 0$ and $ba = 0$, then we don't bet.

3. Reward will simply be the total profit from a bet.

In order to address some of the bet scaling logic implicitly, we will want to define some sort of "memory" in our agent that tracks the change in bankroll over last look back window. However, all of this is out of scope in the class, and will likely be the next iteration that I work on in my masters next year.

Appendix

American	Decimal	Type	Implied Probability	Profit on $100 Bet
+150	2.5	Underdog	0.4	$150
-300	1.33	Favorite	0.75	$33

Table 1: Conversion between different types of betting odds.

	Home Team Wins	Away Team Wins
Home Bet	$Oh - 1$	-1*
Away Bet	-1*	$Oa - 1$
No Bet	0*	0*

Table 2: Decision table for Method #1: Maximin

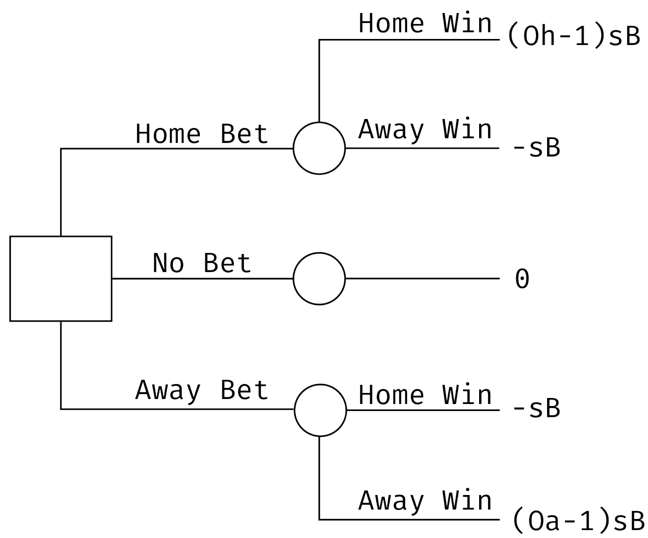

Figure 1: Decision tree for modeling a bet.

	Accuracy	TPR	FPR
Baseline Model	0.632	0.8634	0.6318
Logistic Regression	0.661	0.8162	0.5606
Random Forest	0.652	0.8311	0.6022
XGBoost	0.670	0.7961	0.5513
Neural Network	0.639	0.8634	0.6318
Blended	0.666	0.8239	0.5587

Table 3: Results of attempts at building ML classifiers for NBA games.

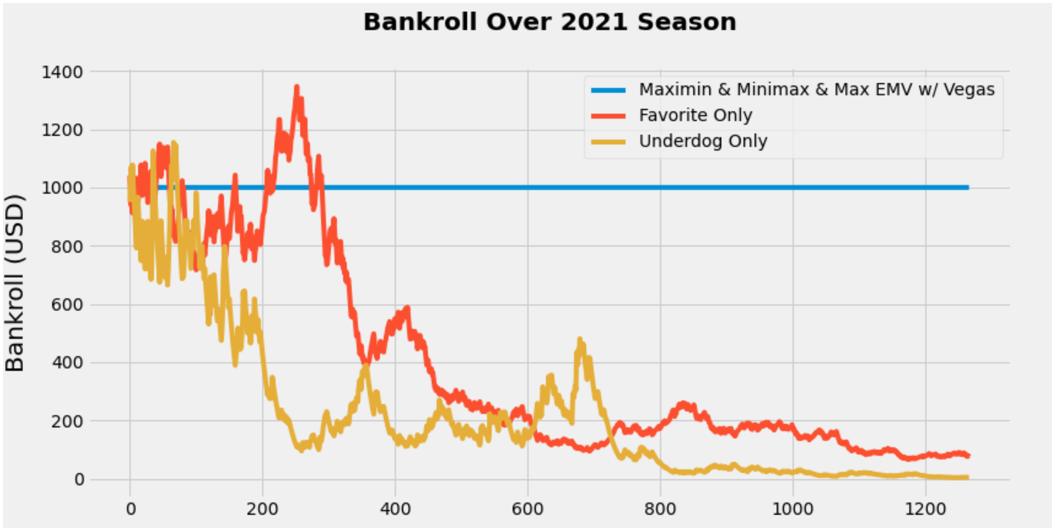

Figure 2: Simulation Results #1

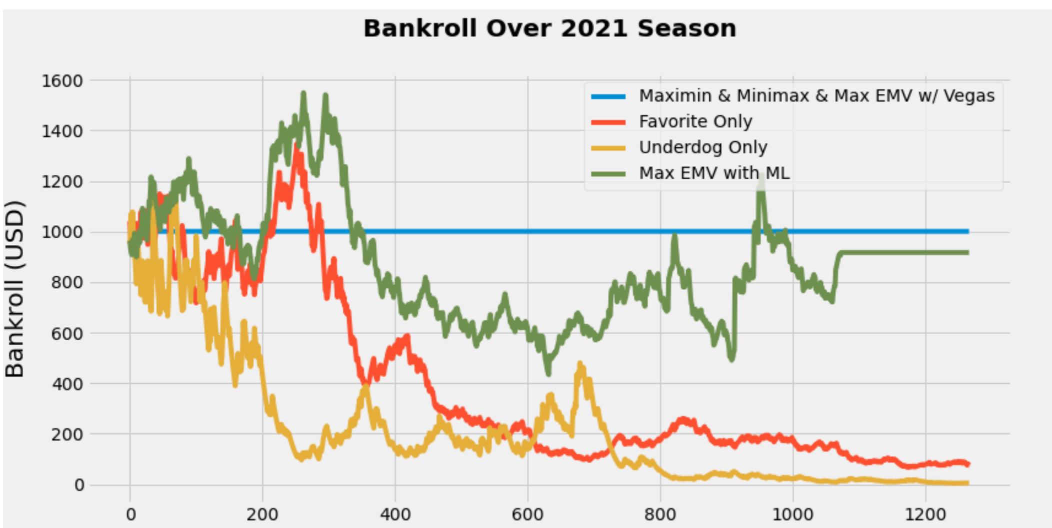

Figure 3: Simulation Results #2

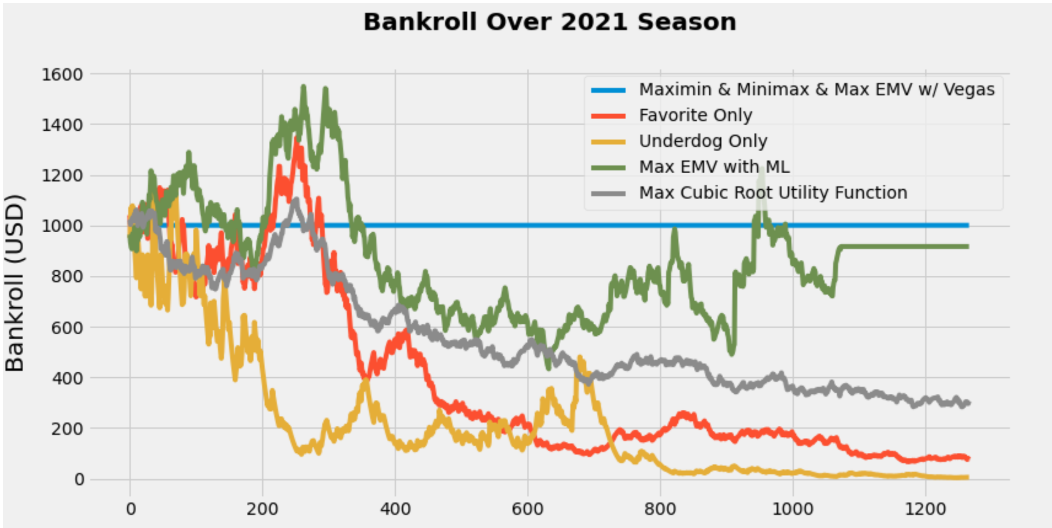

Figure 4: Simulation Results #3

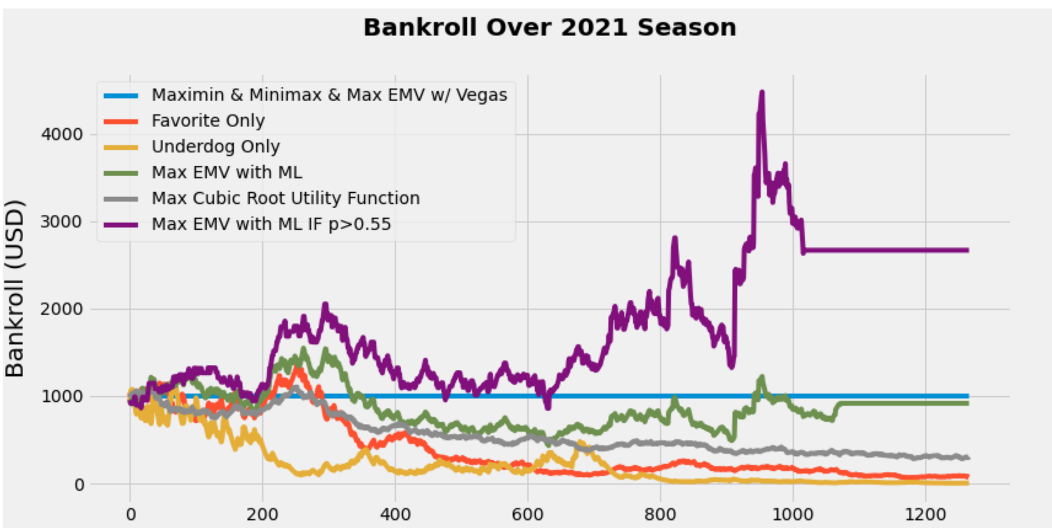

Figure 5: Simulation Results #4

	Bet Count	Bet Accuracy
Favorite Only	1267	64%
Underdog Only	1267	36%
EMV + ML	1074	50.6%
Max Utility Function	990	62.9%
EMV + ML + Threshold	616	63.1%

Table 4: Results of attempts at building ML classifiers for NBA games.

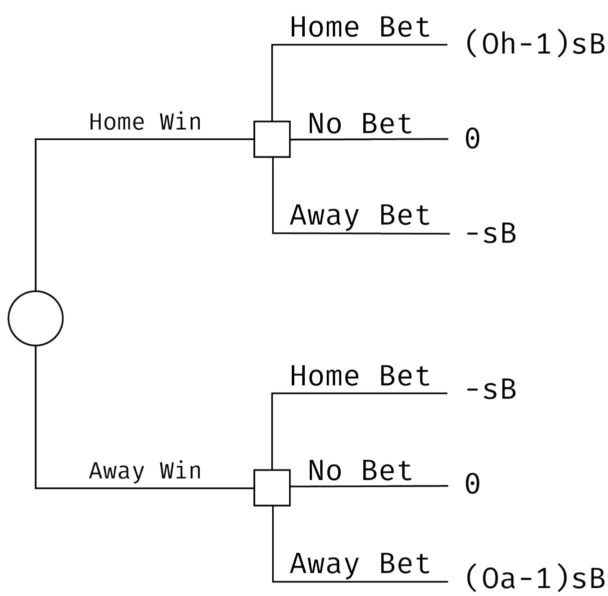

Figure 6: Perfect information decision tree.

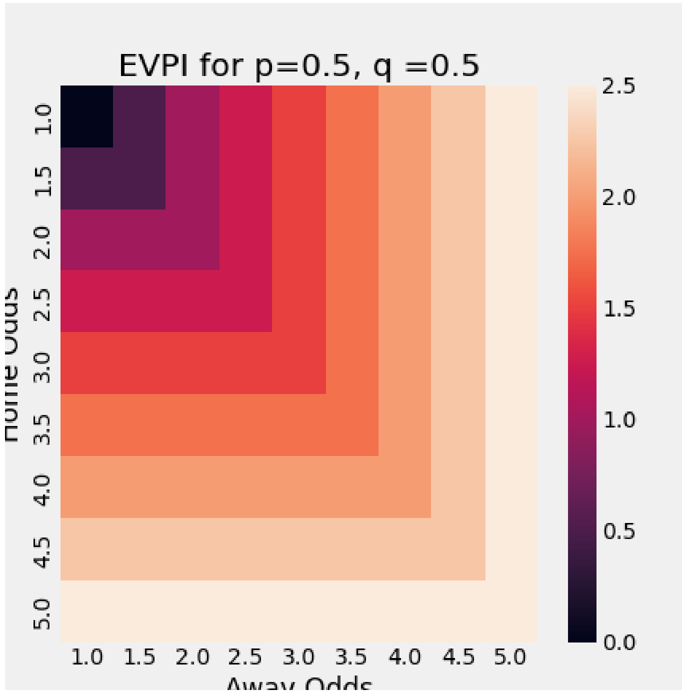

Figure 7: Heat map showing the EVPI for an even game.

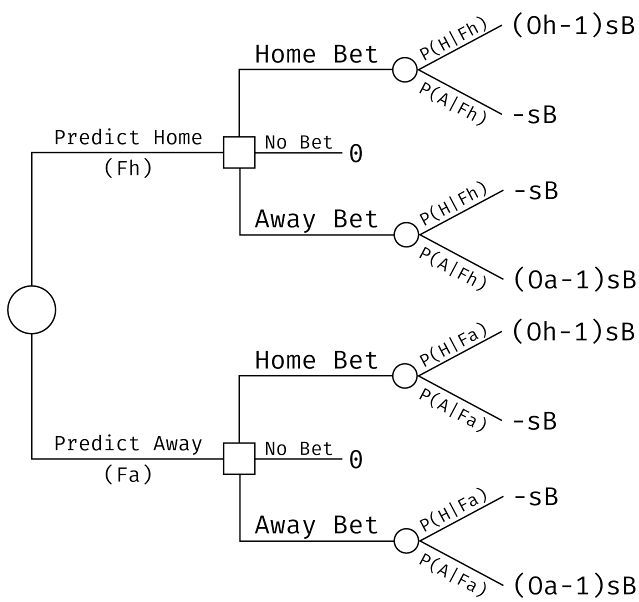

Figure 8: A tree reflecting the fact that our $\hat{p}$ value are forecasts, not true values.

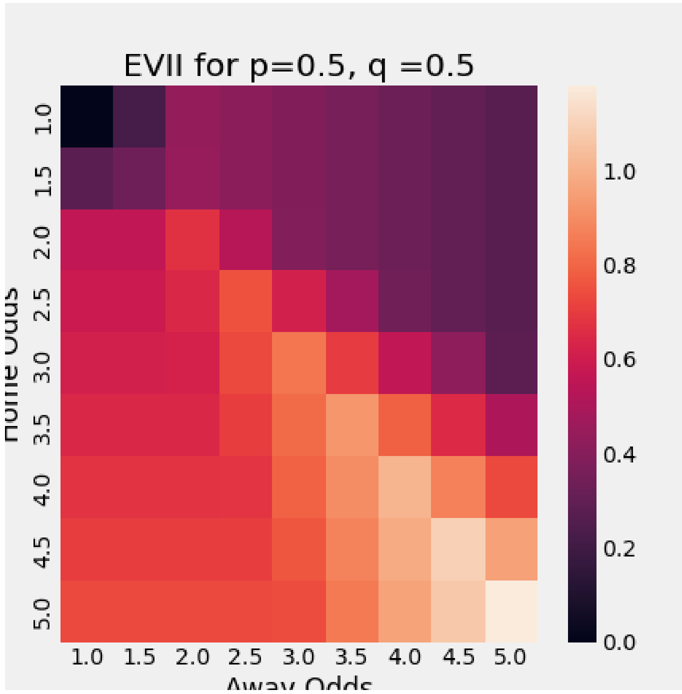

Figure 9: Heat map showing the EVII for an even game.