**Data Science for Sports Final Project Report**

**Ian Mabon, Fardin Hafiz, Loveness Manukopa**

**DNSC 6290**

**4/29/25**

**Introduction**

Sports betting has become a heavily discussed topic of conversation in recent years. This is a direct result of the widespread legalization of sports betting across the United States. Alongside the availability of betting apps and online betting, almost everyone has access to sports betting. 47 out of the 50 states have legalized sports betting, which means these sportsbooks have access to millions more people than before. The prevalence of sports betting made us interested in investigating the accuracy of these sportsbooks in predicting a correct scoreline. The main purpose of sportsbooks is to offer odds or a prediction of the result of a sports game. It is then upon the user to bet money on the results of the game. Sportsbooks set lines, which are a metric of which team will win and by how much. These lines are used to determine how much money a bettor would win if they won their bet. This makes the setting of lines and spreads to be very important in the entire process of sports betting.

Our project focused on how accurate these lines were for NBA basketball games. The purpose of this project is to see how often the sportsbooks were correct in guessing the winner of an individual NBA game. This was done by finding different NBA and basketball datasets that could shed some light on this investigation. By conducting different statistical tests and building models, we aimed to find out the general accuracy of different sports books. The results of this project could educate sports bettors into making more educated bets to win more bets and money. If a meaningful pattern could be found, it could help bettors make safer bets.

**Exploratory Data Analysis**

Before creating more intricate models, we conducted some EDA to gain a simple understanding of some metrics that could help determine what bet to make. These graphs are very simple in nature and just use the data provided to get a general understanding of some simple trends in the datasets. Figure 1 looks at the average spread deviation by each sportsbook. The purpose of this graph is to see how similar the spreads are for each separate sports book. The average spread for all the sportsbooks was taken, and this chart shows how much each book deviates from this average. As can be seen, some sports books have a positive variation from the spread, like Pinnacle Sports, 5Dimes, Heritage, Bookmaker, and InterTops. Some were very close to the average, like YouWager, JustBet, Sportsbetting, and BetOnline. Bovada had the biggest deviation from the average, and it was more negative than the average. It must be noted that even though the graph may make it seem like large differences, they are all generally similar in spread. This does not mean that the findings of this graph are not significant, as this average was taken across thousands of games, which means there are some definite trends.

Figure 2 investigates what percentage of games are won by either the home or away team. This graph was done by taking the total games played and dividing by the away winner and home winner. Doing this analysis provides some insight into which team usually wins in a matchup. As seen in Figure 2, 61.9% of the games are won by the home team and 38.1% are won by the away team. Figure 3 looks into how the total points scored have changed in NBA games throughout the years. The total points scored is a common metric that can be bet on, so it seemed intuitive to see how this trend has changed with time. This was done by making a line plot that compared the year with the average total points scored. Figure 4 demonstrates that total points have fluctuated by era, but in recent years, there has been a general uptick in total scoring. Lastly, another similar graph was produced in order to study the change in point differential over the years. Point differential is another metric that is commonly bet on, so seeing some trend in the data could help bettors have a better understanding of what they should bet on. Figure 4 investigates these trends by plotting the point differential by year in order to see the trend over time. It shows a similar trend to that shown in Figure 3. It should be noted that both Figures 3 and 4 have an outlier in 1961-1962, which was the year Wilt Chamberlain famously scored 50 points per game. This and other high scoring numbers that season made the 1961-62 season an outlier.

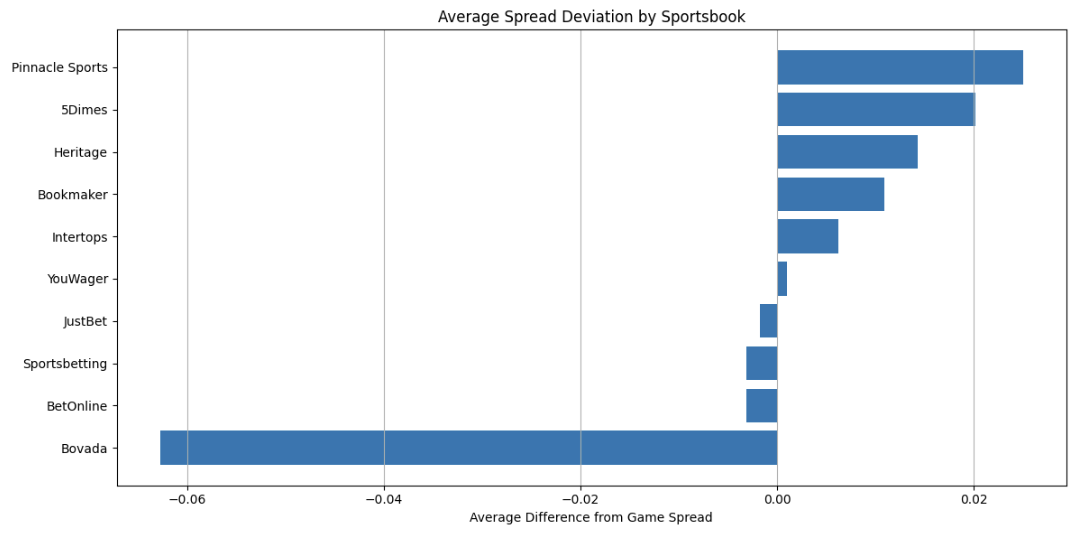


Figure 1: Average Spread Deviation by Sportsbook.

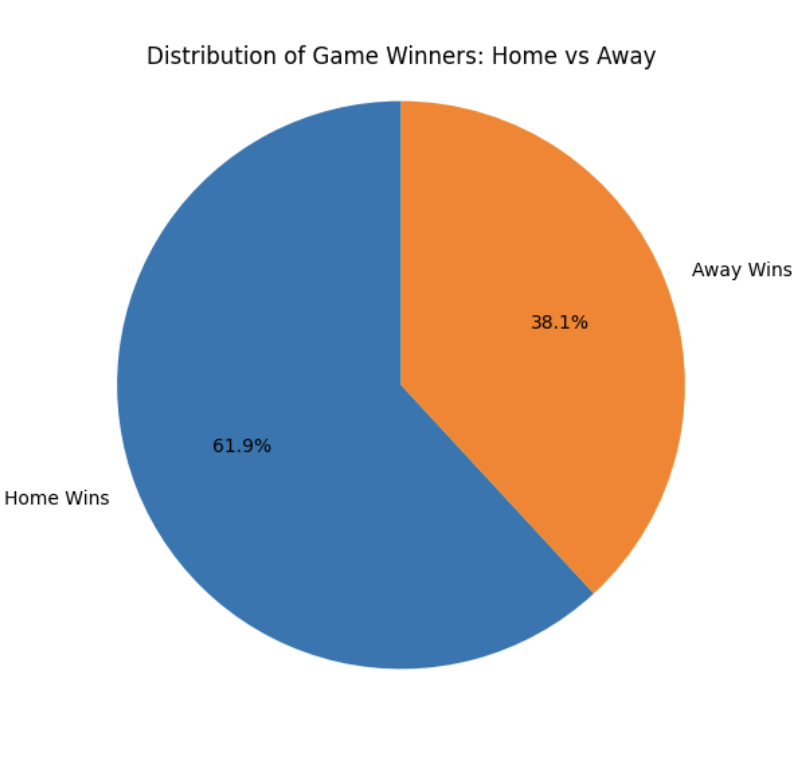


Figure 2: Pie chart of the Distribution of Game Winners

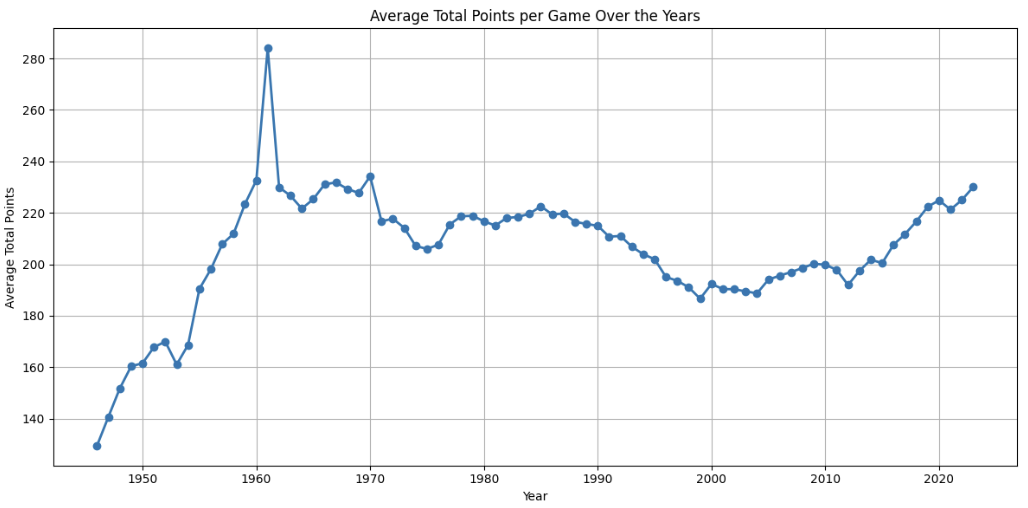


Figure 3: Average Total Points per game over the years

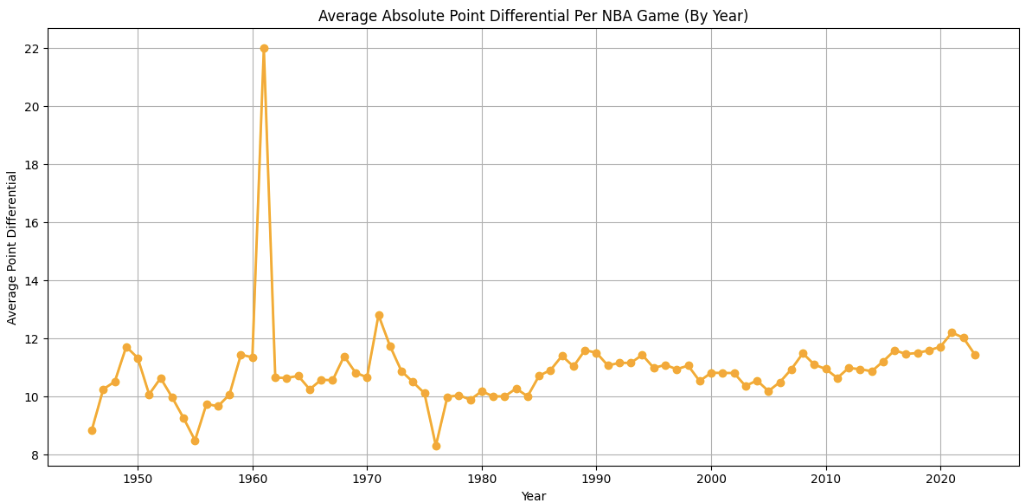


Figure 4: Average Points differential per game over the years

**Model Methodology and Preprocessing**

To explore the relationship between odds and win/loss outcome, it was decided to create decision tree classifiers and a logistic regression model using only odds as the predictor variables. First, the data was loaded into the environment and concatenated into one large dataset. Since there were 10 different sportsbooks with similar but not equal prices and spreads, each book was given its own dataset to distinguish each book from another. Each dataset created had values from the “nba\_betting\_money\_line,” “nba\_betting\_spread,” “nba\_betting\_totals,” and “nba\_games\_all” datasets downloaded from Kaggle joined on the “game\_id” column. Many of the betting datasets used the same column names like “price1” and “price2” so those were later changed to reflect from which dataset those prices (odds) came from. Unnecessary columns like game statistics and duplicate ids and games were also dropped since this project focused on odds being a predictor only. The hard part of this preprocessing was correctly joining the betting datasets to the game data since the game data was structured with a “Designated Team” which was home or away and had the win/loss indicator. Betting prices were structured in a home and away format, so numpy was used to correctly sort odds by looking at the team id and changing the value to reflect the correct team’s odds and point totals. A team abbreviation was also added for the purpose of being used in the Team Strength model built later. More columns were dropped and finally the datasets were ready for model creation.

**Model Creation and Analysis**

The first models created were decision tree classifiers chosen because the target variable was binary (Win = 1, Loss = 0) so a classification model was a simple choice. The X variables were “Designated Team ML Odds”, “Designated Team Spread”, "Other Team ML Odds", "Other Team Spread” while the y variable was the “wl” column. To assess accuracy, each dataset was split for 60% training data and 40% test data. One tree was created for each dataset which created 10 in total. Then confusion matrices and accuracy scores were given to each tree. This process was repeated for “heavy books” where teams were favored by -300 or more and for “massive books” where teams were favored by -1000 or more.

Looking at the trees created with all of the data, each dataset for each sportsbook contained around 4,000 games per tree, having the tree train on 2,400 games and testing their accuracies on 1,600 of them. Their accuracies were around 65%, meaning that the odds themselves were not reliable predictors for game outcomes. However, in comparison to a baseline model of randomly guessing outcome would produce an accuracy of 50%, they are marginally better meaning there is some significance odds have on game outcome. For the heavily favored matchups, each dataset contained around 1,000 games to train and test on. Using the same 60-40% split as the last trees, these trees had around a 70-80% accuracy, somewhat improving these metrics from the previous trees. Similarly, the massively favored books follow the same trend of improved accuracy and smaller datasets. The datasets only contained around 200 games (since being favored by -1000 is rare) and accuracy improves to around 80-90%. Looking at the confusion matrices for both heavily and massively favored games, a significant portion of the outcomes were predicted as wins and most were. This indicates that the more favored a team is, the more likely they are to win. Noise exists in all datasets since sports are inherently unpredictable and there is no way to know a game’s concluding score without collusion from the players.

Next, a logistic regression model was built with the aim of getting a reliable win probability for each matchup based on odds alone. Once the win/loss column was replaced with an actual binary value of 0 or 1, a constant was added, and the model was created. To get a win probability for each matchup, each game in the X\_const collection was predicted using the model built on that data and a new column was added to reflect each team’s probability to win. The same process for model creation for each sportsbook was adhered to and 10 models were produced, each with win probabilities from book specific models. Looking at each regression output, model reliability varied wildly with some models having significant coefficients based on their p values and others had p values of 0.999 or 1 making them completely insignificant. The variables that ended up being significant also weren't the same between models, with some favoring the “Designated Team ML Odds” and “Designated Team Spread” and others favoring “Other Team Spread.” The number of observations in each model varied as well which were between 9,000 and 14,000 games each model was built on. These models aren’t reliable, but they are a good baseline to compare to a traditional model to find win probabilities between two teams.

This traditional model that complemented the logistic model was the Team Strength Model as learned about in lecture since it produces a win probability for any given matchup based on a team’s relative “strength” as compared to the pseudo “average” team in the league. The X and W matrices were created for each sportsbook and team strengths were given to each team in every dataset. These team strengths are reflective of each team over a period of 11 years, so strengths are an aggregate of the franchises’ successes or failures over a long period of time. This could lead to inconsistencies in team strengths year by year even if a team does much better the next season through a trade, coaching changes, or injury statuses. Using those strengths specific to each sportsbook, a win probability was even to each matchup that the team strengths were trained on. Those probabilities were then compared to the probabilities created by the logistic regression odds model.

**Model Comparison**

To evaluate model effectiveness, the performance of the odds-based models were compared to the team strength models. The odds-based models consistently outperformed the strength-based models across every metric:

* Log Loss was lower for the Odds Model (0.5920) compared to the Strength Model (0.6587).
* Brier Score, measuring the accuracy of probability forecasts, was lower for the Odds Model (0.2034 vs. 0.2333).
* AUC (Area Under the ROC Curve) was higher for the Odds Model (0.7498) than the Strength Model (0.6481).
* Accuracy was also higher for the Odds Model (68.34% vs. 60.67%).

These results suggest that sportsbooks’ odds incorporate a wide array of dynamic, real-time information—such as injuries, travel schedules, player form—that traditional models based purely on historical team strength cannot fully capture. Thus, while team strength provides a general view of capability, odds offer a more current and sharper prediction for individual games.

**Conclusion**

Through this analysis, it became clear that sportsbook odds are quite effective at predicting NBA game outcomes, especially when there is a heavy or massive favorite. However, betting remains risky; favorites do lose, and often in surprising ways. Odds-based models, even when relatively simple, performed better than models built on historical team strengths.

For bettors, this means that while it is generally safer to bet on heavily favored teams, nothing is guaranteed. Upsets happen, and the margin of loss in those rare cases can be significant. Our findings also underscore the sophisticated nature of sportsbook algorithms, which aggregate vast amounts of information into the odds they offer.

Ultimately, this project demonstrates that betting odds themselves can serve as strong predictors and that understanding the nuances behind odds can help inform smarter sports analytics, even outside of betting. However, success in betting—or prediction in general—will always involve balancing analytical insights with the unpredictable nature of live sports. There is no perfect betting strategy found from this analysis. Betting remains risky, and while odds favor winners, variance (luck, injuries, unexpected performances) always plays a major role.