NBA Capstone Project Milestone Report
Jeffrey Ebert

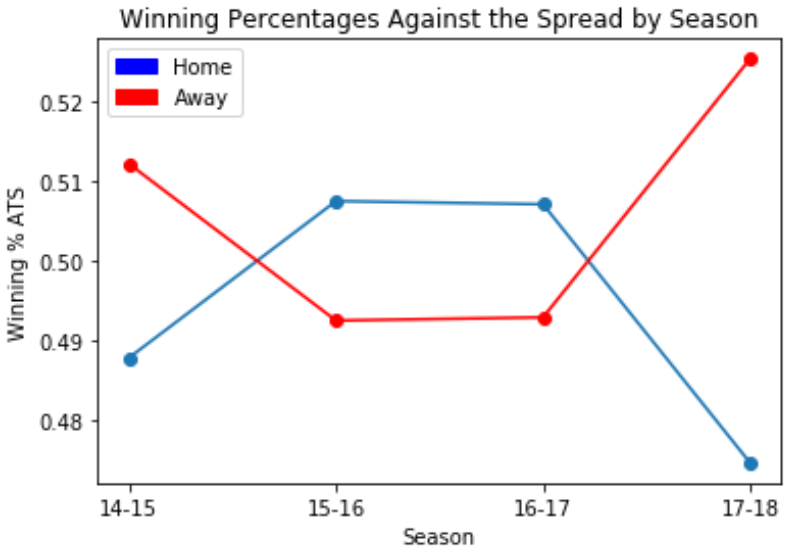
Evaluating Games and Trends in the NBA

The objective of this project is to determine what the most important factors are in an NBA team winning a game. I will evaluate important trends and patterns that occur for teams throughout the course of a season, and try to find value in predictive variables that could be beneficial for purposes of betting on NBA games. The rise of legalized sports gambling across the United States has extended the market of individuals and companies looking to provide sports insight to customers who are betting frequently and heavily on games. Soon, the sportsbooks themselves will be selling information about trends for a team and that take place during a season to try and make even more money than the massive amount of revenue that comes in from gambling on sports alone. As more and more states continue to review laws to legalize sports betting, there will be a greater audience and market for these types of forums. Every year that passes there is more data known and collected about sports than ever before, and basketball is becoming a worldwide phenomenon to watch, play and bet on, with the United States the forefront of it all.

The main dataset collected for this project was obtained with a web scraper, getting the scores from the regular season in the NBA during the 2014-2015, 2015-2016, 2016-2017, and 2017-2018 seasons. The dataset was thoroughly inspected to make sure all of the scores, teams and dates were recorded accurately. There were errors in the data collection caused by postponed games, where the web scraper returned the team names but no associated scores because they did not play. It was important to find these values in the dataset and delete the teams and dates of games that did not end up happening to ensure everything matched up correctly. The next dataset used in the project was uploaded from a public dataset on Kaggle.com. It included in-game statistics from every regular season game of these four seasons. These datasets were merged by their matching team names and dates of a game to connect each game score to its associated statistics. The last dataset obtained for this project came from sportsdatabase.com. Here, tables of NBA games point spreads were uploaded in order to perform analysis based on the recorded before game odds a team was projected to have of winning the game. This dataset was merged with the team scores and stats to form the main dataframe used throughout the code, df. Using all of this information, from this point many other variables were created. Some of these include the rest a team had, which was the amount of days between games played, whether a team was playing at home or on the road, whether or not a team won the game, whether that team won against the associated spread, lost, or pushed (tied the spread), a team's winning and losing streaks, their previous 10 game

performance and more. These were all created with various python functions that incorporated all of the data, often broken down by which team is playing and during which of the four seasons the game was in.

From here, I wanted to look into which variables were important measures in determining if a team wins a game or not. I knew it would be valuable to know that some variables could be important on their own, some more so when looked at it in the scope of something else happening, and it would also be important to understand what factors are not as important as we may think they are in determining who wins a game. This idea is especially significant in terms of gambling, where in order to accurately pick a winner the majority of the time one would need to eliminate previous biases related to what they may think is important in terms if winning a game. For example it is easy to believe that home teams are better at covering the spread in the NBA. This logistically makes sense, in sports we often hear about a team's home court advantage, how some teams are more comfortable shooting in their own arena and the opposing team does not do as well in an unfamiliar setting. In fact, from 2014-2018 home teams during the regular season actually won about 59% of the time, so most of this could be true. But, does this mean home teams usually cover the spread? Oddsmakers factor into the line where the game is being played, and how the teams playing have historically performed in that arena. So, let's look at the data:

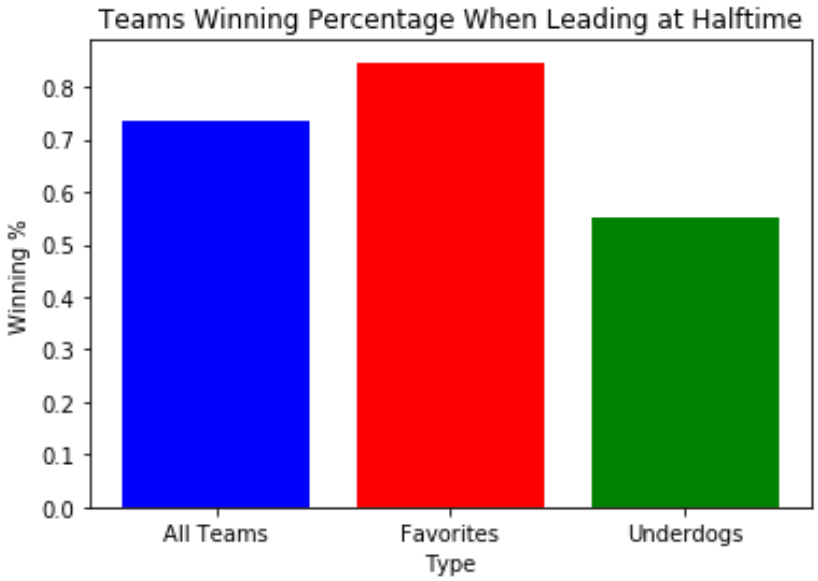


Based on our inferences, this is largely what we would expect. All of these against the spread (ATS) winning percentages at home and on the road hovered around the 50% mark. The greatest deviation came in 2017-2018 when home teams did not even cover 48% of the time. When making a bet, to conclude that any team will likely cover the spread because they are playing at home is naive. However, that does not mean that some teams don't have discernable

patterns that occur from year to year and during the season that reflect definitive trends in often winning ATS at home or on the road.

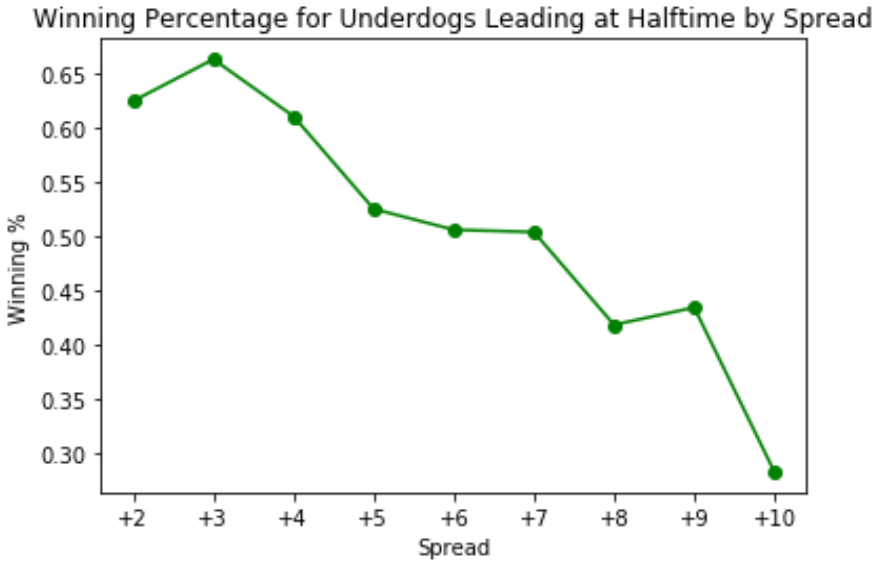
Portland was the only team in the league that had a winning percentage against the spread of greater than 50% for all four seasons. Three teams, Cleveland, Oklahoma City, and Memphis, has an ATS percentage of less than 50% all four seasons on the road. This demonstrates the overall variability, but that there can be consistency when looking at the correct data. There were also many stretches for teams throughout individual seasons where you could locate extended stretches of games where teams performed better on average at covering the spread depending on where they play, and the type of opponent.

We can also look at factors that take place throughout a game that can provide us information on how often certain teams win games based on a set of circumstances. This graph represents the winning percentages for all teams, favorites and underdogs when leading at halftime by any amount:



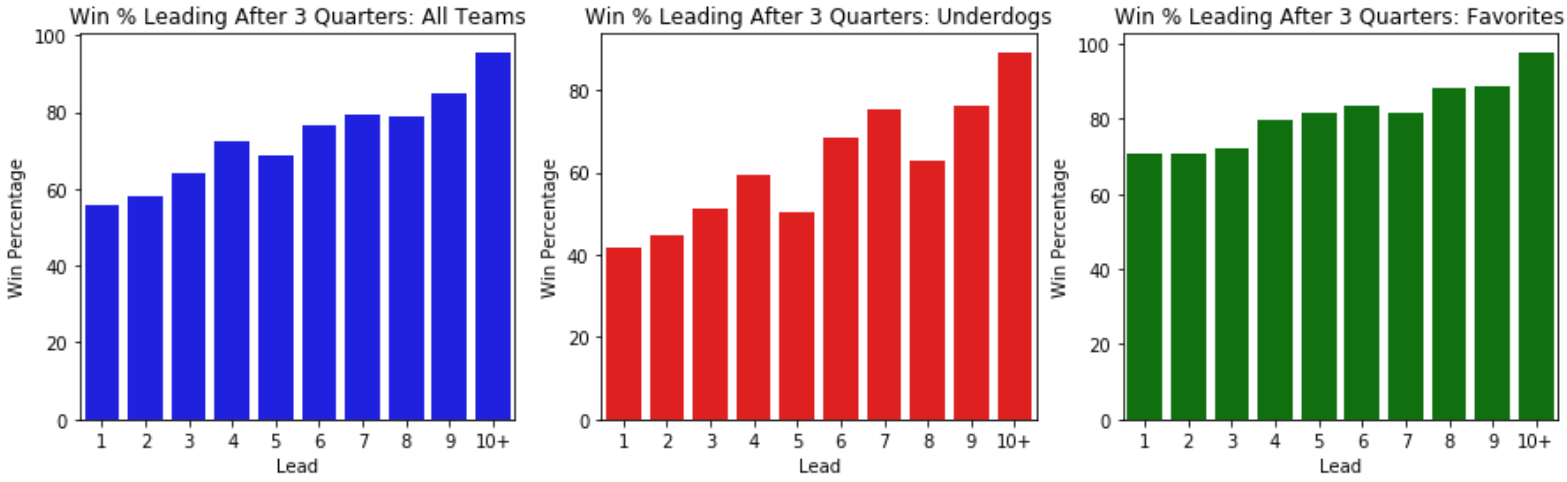
Teams leading at half win 74% of the time and favorites won over 84% of the time from 2014-2018. When an underdog is leading a game at the half compared to a favorite, they win at almost a 30% less rate, at under 55% of the time. Let's take a look at how this percentage fairs based on how big of an underdog a team is. We would expect the greater an underdog they are, no matter their halftime lead, they are less likely to win the game. This graph evaluates

how often an underdog leading at the half by any total wins depending on how many points they are favored to lose by before the game starts:

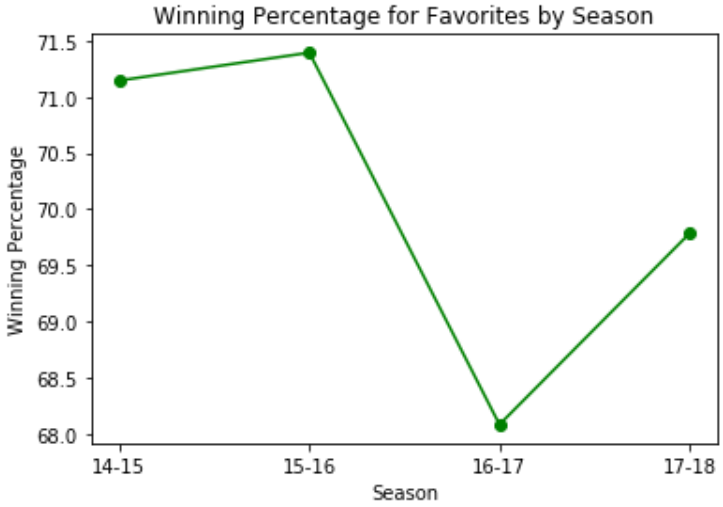


This graph is largely what we'd expect. When leading at half, five point underdogs and less win at more than a 50% rate, but larger underdogs of six or greater points have ended up losing more than half the time. This is not telling us everything going on here, because it does not factor in the size of the leads, but can still be valuable in the context of second half lines. At halftime, oddsmakers set a line for the second half of the game that can be wagered on. If a favorite is losing at halftime, many times the second half line would be close to an even spread for the remainder of the game. Knowing which types of teams end up closing out these games better could be advantageous, as it would be riskier to bet a larger underdog to win, especially since the percentage of wins drops dramatically at the +8 point mark.

The next analysis will look at when teams are leading after three quarters and how often they end up winning the game. How do these rates change depend on the lead they have, and whether they are an underdog or favorite?



As a team's third quarter lead increases, they end up winning the game more often. This is seen much more prominently in favorites than underdogs however. Underdogs are not consistently winning without at least a 6 point lead, whereas the percentage of games won with a lead is much more steady for a favorite. This could be simply reflective of how favorites win more often, which would mean the majority of the time they are just in being called a favorite. Let's see if that is true.



Like we predicted based on third quarter leads for favorites from the previous graph, for each year favorites are more likely to win. This shows they are right in being called favorites since they mostly win from a 68 to 72 percent rate year by year.

Another important question in understanding the makeup of what determines a winning team in the NBA is what statistics are most correlated with winning. Below is a correlation matrix showing the values of the relationships throughout all seasons of a team's average statistics with winning, as well as with each other.

	win	A_FGM	A_FGA	A_FG%	A_3PM	A_3PA	A_3P%	A_FTM	A_FTA	A_FT%	A_OREB	A_REB	A_AST	A_STL	A_BLK	A_TO	A_PF
win	1	0.11	-0.013	0.15	0.1	0.079	0.095	0.033	0.016	0.051	-0.024	0.042	0.12	0.065	0.066	-0.048	-0.055
A_FGM	0.11	1	0.59	0.77	0.38	0.27	0.42	-0.085	-0.16	0.19	-0.032	0.25	0.67	0.21	0.17	-0.0069	-0.092
A_FGA	-0.013	0.59	1	-0.068	0.24	0.29	-0.043	-0.14	-0.14	-0.017	0.37	0.54	0.23	0.11	0.019	-0.028	0.027
A_FG%	0.15	0.77	-0.068	1	0.28	0.098	0.56	0.0025	-0.092	0.25	-0.33	-0.11	0.63	0.16	0.19	0.0072	-0.14
A_3PM	0.1	0.38	0.24	0.28	1	0.94	0.52	-0.086	-0.12	0.092	-0.26	0.063	0.38	0.12	0.027	0.088	-0.073
A_3PA	0.079	0.27	0.29	0.098	0.94	1	0.22	-0.064	-0.069	0.024	-0.21	0.086	0.27	0.12	0.0043	0.11	-0.03
A_3P%	0.095	0.42	-0.043	0.56	0.52	0.22	1	-0.11	-0.19	0.2	-0.24	-0.045	0.41	0.044	0.064	-0.017	-0.13
A_FTM	0.033	-0.085	-0.14	0.0025	-0.086	-0.064	-0.11	1	0.92	0.28	0.17	0.14	-0.14	0.063	0.025	0.0094	0.23
A_FTA	0.016	-0.16	-0.14	-0.092	-0.12	-0.069	-0.19	0.92	1	-0.098	0.26	0.19	-0.2	0.064	0.021	0.068	0.25
A_FT%	0.051	0.19	-0.017	0.25	0.092	0.024	0.2	0.28	-0.098	1	-0.23	-0.13	0.16	0.021	0.022	-0.15	-0.048
A_OREB	-0.024	-0.032	0.37	-0.33	-0.26	-0.21	-0.24	0.17	0.26	-0.23	1	0.57	-0.22	0.0024	-0.0049	0.13	0.15
A_REB	0.042	0.25	0.54	-0.11	0.063	0.086	-0.045	0.14	0.19	-0.13	0.57	1	0.043	-0.19	0.23	0.18	-0.046
A_AST	0.12	0.67	0.23	0.63	0.38	0.27	0.41	-0.14	-0.2	0.16	-0.22	0.043	1	0.26	0.25	0.11	-0.14
A_STL	0.065	0.21	0.11	0.16	0.12	0.12	0.044	0.063	0.064	0.021	0.0024	-0.19	0.26	1	0.033	0.26	0.16
A_BLK	0.066	0.17	0.019	0.19	0.027	0.0043	0.064	0.025	0.021	0.022	-0.0049	0.23	0.25	0.033	1	0.11	-0.0068
A_TO	-0.048	-0.0069	-0.028	0.0072	0.088	0.11	-0.017	0.0094	0.068	-0.15	0.13	0.18	0.11	0.26	0.11	1	0.35
A_PF	-0.055	-0.092	0.027	-0.14	-0.073	-0.03	-0.13	0.23	0.25	-0.048	0.15	-0.046	-0.14	0.16	-0.0068	0.35	1

The top row or leftmost column is of the most importance to us because it shows the relationship between a certain average statistic and a win for that team. It appears that the statistics that are most correlated with a win are average field goal percentage (A_FG%) and assists per game (A_AST). These are again directly related to scoring and higher marks would determine a more efficient basketball team. The average statistics with the smallest correlations are free throw attempts per game (FTA) and field goal attempts per game (A_FGA). It is possible that the values representing team attempts for a certain type of shot have the least significant because it does not matter how many shots they take, and what is important how many they make and how many the other team make and attempt compared to their totals.

The team's statistics are not the only important numerical features occurring in the game, mainly because they are facing an opponent that will have its own strengths and weaknesses. Which team statistics compared to their opponents are significant in terms of winning and which aren't? This could be significant to know in terms of building models to predict wins based on using game statistics.

One important statistic often referenced during games is field goal percentage (FG%), which is equal to the percentage of made shots by a team during gameplay (not including free throws)

Null hypothesis (H0): There is no significant relationship between having a higher average field goal percentage than your opponent and winning

Alternative Hypothesis (H1): There is a significant relationship between having a higher average field goal percentage than your opponent and winning

The result of a one sample t-test yielded:
Ttest_1sampResult(statistic=13.903051918643442, pvalue=4.0362161127095995e-43)

Based on these values and our very small p-value of close to zero, we reject the null hypothesis that having a higher average field goal percentage than your opponent is not significant towards winning. Logistically this makes sense as teams that shoot better compared to their opponent would be thought to having an advantage. Teams with a higher average FG% win at a 59.79% rate, which seems very high.