Principles of Statistical Modeling: Miniproject I
A Short Tourist Flight over the lands of Basketball Statistics

MARK KOERNER

May 15, 2019

Introduction

The objective of this Miniproject is to analyze the structure of a chosen dataset. I chose the NCAA Men's March Madness dataset for this project, which is part of the annual NCAA ML Competition[Kaggle, 2019] to predict the winners for each game of the National Collegiate Athletics Association (NCAA) Basketball Championship playoff tournament. There are numerous datafiles documenting different aspects of the College basketball world which are part of the competition, such as rankings for each team, regular season and playoff statistics for past years dating back to 2003 as well as more information on the teams and game locations. I chose to focus mostly on the RegularSeasonDetailedResults.csv file, as it offers the most breadth and the most opportunity for exploration and feature extraction. In the first section, I will go into further detail regarding my motivation for choosing this particular dataset. In the second section, I will give more background on the game of basketball and introduce the general structure of the dataset, whereas I will take a deeper look at the raw random variables in the following section. Lastly, I will talk about feature extraction and give an outlook on the challenges for predicting actual games, followed by the conclusion.

1 Motivation

The first time I had come across this dataset was as part of the Kaggle Competition in 2018, when I was looking for a way to apply my then limited knowledge of Data Analysis with a field I have had a passion for almost my entire life. I started playing basketball when I was 7, and both my parents and my sister are avid basketball players, so I never had much of a choice in what sport to play. In all seriousness though, basketball is one of my biggest passions and was a large part of my life growing up. In addition, I had been following College Basketball in the U.S. for a couple of years, and thus

was familiar with the NCAA season structure and most of the rule peculiarities that are specific to the league. Even though I had a good amount of domain knowledge, I struggled with the structure of the dataset, in particular manipulating the dataset, extracting useful features and building a model. After not finding the time to participate in the competition this year, I figured it would be a useful exercise to take a more detailed look at the structure of the dataset before potentially participating again next year.

2 Background and Dataset Structure

In this section, I will describe and explore the factors that make teams win basketball games by looking at regular season game results. First, I will give a little background on the game of basketball and the NCAA, then describe the experimental design, and finally explore the provided game statistics.

2.1 Basketball Background

Basketball is a team game, where both teams, each comprised of 5 players, attempt to score points by throwing the ball in the respective hoop which the other team is trying to defend. The game is competitively played indoors on a rectangular court, with the two baskets on the opposing shorter ends of the court. Players are allowed to advance the ball by dribbling, i.e. bouncing the ball on the court without fully catching it between bounces, or passing to other players on the team. The NCAA rules dictate that the game is played in two 20 minute periods, at the end of which the team that has scored the most points wins. In case of a tie, additional 5 minute overtime periods are played until a winner is determined. Player substitutions are allowed when the ball goes out of bounds or when a foul is committed by a player. Generally speaking, a foul is committed when a player initiates unnecessary or excessive contact with a player on the other team. Each player is allotted 5 personal fouls per game, after which the player is expelled from the game.

2.2 Experimental Design

Before actually looking at the available data, I decided to take a step back to describe the scenario in which the data was collected in terms of the five components introduced in the course: the reality segment of interest (RSOI), the set of observation opportunities (OO), the observation procedure (OP), the observation acts (OA) and the data value space (DVS).

NCAA Basketball Game	
RSOI	All Division I NCAA Men's Basketball Games (in this case from the 2003 season until today)
OO	Any time an NCAA sanctioned game occurs between two Division I basketball teams.
OP	When it comes to just determining the winner of the game, all one would have to do is keep track of the points scored of either team and then declare the team with the higher point total the winner. During NCAA games, the scorer's table keeps track of the score and other statistics and publishes it to the scoreboard as well as to online websites providing real-time statistics for each game. While it would not be a proper scientific experimental setup, one would not even have to attend the games to be able to determine the winner.
OA	An observation act occurs when such a game is carried out.
DVS	The data value space for whether a team wins or loses is a boolean [win, lose], similar to die throwing. For most other statistics, the value is derived by counting the occurrences of certain actions for both teams. A realistic DVS for most statistics would be all natural numbers in the interval [0, 200] or just $\mathbb{N}$.

The actual competition expects a probabilistic output of one team winning over the other in the form of $[Team\,1\,loses, Team\,1\,wins]$ with the outcome in the interval [0,1], which is why the above setup was chosen. Note that the above setup is for games that have already happened, whereas the actual competition expects a future prediction where the described statistics are obviously not yet available. But before we go further into the actual competition, let us take a look at the impact of traditional basketball statistics on winning other than "whoever scores more points wins the game".

2.3 Dataset Structure

Before we go into the actual statistics, we have to first learn to navigate the dataset. The regular season dataset features 82041 observations broken down in 34 variables. Ignoring the Team IDs and the Game Location, the dataset is made up of 31 numerical random variables: the number of overtimes that were played *NumOT*, the season variable *Season*, the day of the season *DayNum* and 14 game statistics for both teams. All of the basketball statistics in this dataset are counted by volunteers on the scorer's table, meaning all are RVs are of the form:

$$X_i : \Omega \to \mathbb{N} \quad (i \in I) \tag{1}$$

In particular, the dataset is already ordered by the winning team and losing team for each game. The variable names for the game statistics are preceded by W for the winning team or a L for the losing team, where the *WTeamID* and *LTeamID* fields allow us to attribute the game outcome to a

specific team. The rest of the variables are described in more detail below according to the Kaggle descriptions [Kaggle, 2019]:

Variable Descriptions	
Score	Points scored by the respective team
FGM	Field Goals Made: the count of successful shots for each team
FGA	Field Goals Attempted: the count of attempted shots for each team
FGM3	Three Pointers Made: the count of successful shots taken behind the three point line
FGA3	Three Pointers Attempted: the count of attempted three point shots
FTM	Free Throws Made: the count of successful free throws made; a player is awarded a free throw after a foul when the player was in the process of shooting the ball or after the team has committed a certain number of fouls per period.
FTA	Free Throws Attempted: the count of attempted free throws
OR	Offensive Rebounds: the count of instances where the ball is retrieved by the attacking team after a missed shot attempt
DR	Defensive Rebounds: the count of instances where the ball is retrieved by the defending team after a missed shot attempt
Ast	Assists: the count of instances where a pass from a teammate is immediately followed by a successful shot attempt
TO	Turnovers: the count of instances where a member of one team loses the ball to a member of the opposing team without shooting the ball
Stl	Steals: the count of instances where a member of one team directly the ball to a member of the opposing team without shooting the ball
Blk	Blocks: the count of shots that the opposite team takes which are blocked or rejected before reaching the hoop
PF	Personal Fouls: the count of fouls committed by all members of a team

As for the variables that are not in-game event based statistics, they contain temporal and location information. The variable *Location* is a discrete variable with values in the set $S = \{H, A, N\}$, which denotes whether the game was played at home, away, or in a neutral location from the winning teams perspective. The variable *Season* contains the year in which that season's NCAA Basketball Tournament was held. This distinction is necessary as the season usually starts in the fall of the previous year. The *DayNum* is another natural number variable, which counts the number of days elapsed from an arbitrary starting date for that season. The variable is designed such that the regular season

always ends on Day 132, so the value for the regular season dataset is constrained to the interval [0, 132][Kaggle, 2019].

It should be noted that all of these statistics are counted by a team of professionals as well as volunteers for each game. While the dataset is clean from missing values and severe outliers as it is cleaned by the Kaggle Team prior to publication, the individual game statistics cannot be taken as the absolute truth. The possibility for human error, whether intentional or not, is fairly high especially given the extensive nature of the dataset. It has become basketball tradition to always distrust the game statistics collected at away games, especially at semi-professional or lower levels of play. With that being said, the NCAA holds teams to a very high standard, and as a whole the dataset should give a fairly accurate representation of the underlying trends and distributions.

3 Raw Data Exploration

In order to get a better sense of the game statistics, I decided to take a closer look at the sample distributions of the random variables across all games. In essence, I wanted to compare the distributions of game statistics for the winning team versus the losing team. In order to visually assess the sample distribution of a discrete numerical random variable, the easiest way is to look at a histogram displaying the counts for each value. As an example, figure 1 shows a histogram for the variable *WScore*. The distribution looks approximately Gaussian and is centered around 75 with most values falling in the interval [55, 95]. Histograms provide a good visual approximation for the distribution of discrete numerical variables, as the distribution of such random variables is defined by its probability mass function. The p.m.f. of a sample can be described as the fraction of observations with a particular value for all values in the sample space. Specifically, given the associated σ-field of a discrete numerical random variable Pot($\mathbb{N}$), i.e. the power set of its sample space, its distribution would be given by the function

$$p : \mathbb{N} \to [0, 1], \quad x \to P_{WScore}(\{x\}), \quad (x \in \mathbb{N}). \tag{2}$$

In practice, the sample distribution can thus be approximated by $count(WScore_i = x)/N$.
Yet, when comparing distributions, looking at two histograms makes it very awkward and laborious to detect differences. A much more elegant way is to estimate sample distributions via Kernel Density Estimation, and to examine the resulting graphs on the same grid.

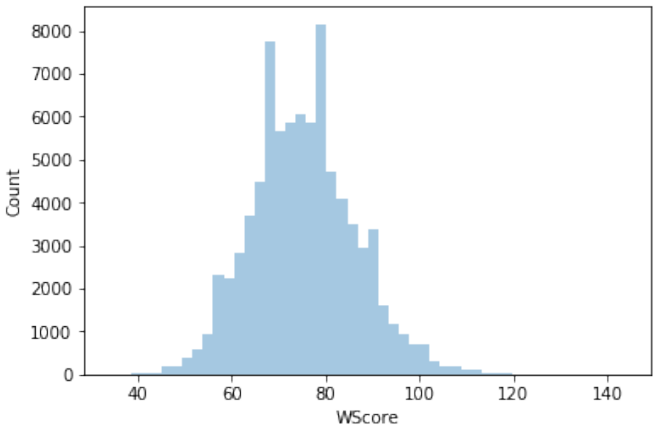

Figure 1: Histogram of Winning Team Game Score

3.1 Kernel Density Estimate

In essence, Kernel Density Estimation applies a kernel with a certain assumed distribution centered at each data point, and then locally averages and smoothes the resulting densities to approximate a density estimate for the entire sample. Equation (3) describes the Parzen density estimate for a Gaussian Kernel with centered mean and standard deviation λ [James et al., 2013].

$$\hat{f}_X(x) = \frac{1}{N} \sum_{i=1}^{N} \phi_\lambda(x - x_i) \tag{3}$$

where $\hat{f}_X$ is the kernel density estimate and ϕ denotes the Gaussian Kernel. In practice, λ is also called the bandwidth parameter and is either applied as a scalar to the standard deviation or replaces it entirely, and ultimately determines the relative smoothness of the density estimate [Waskom, 2018].

While certainly useful for estimating and comparing sample distributions, it must be noted that this method's estimates are based on prior assumptions about the structure of the underlying distribution. While the histogram in Figure 1 as well as histograms for other variables look to be approximately normally distributed, it is still a tricky assumption to make. In addition, kernel density estimation is usually reserved for continuous data, and its use in the context of the raw data is only for visualization purposes.

3.2 Graphical Comparison of Sample Distributions

Instead of going directly into the shooting statistics, which have a very direct impact on the outcome of a game in terms of the score, I wanted to first take a closer look at the supporting statistics such as

Assists, Turnovers, Steals and Rebounds. In particular, assists are actions that are directly associated with field goals and thus points scored, meaning that a large number of assists should potentially be associated with winning as more points are being scored. The scatterplot in Figure 2 confirms a positive correlation between *WScore* and *WAst*.

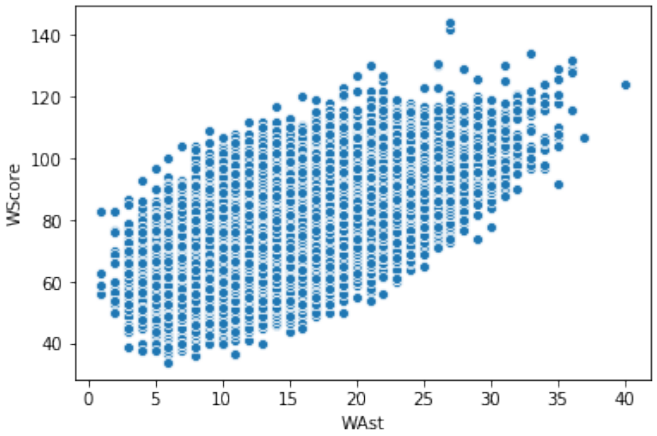

Figure 2: Scatterplot of WAst vs WScore

When it comes to the distributions, Figure 3 shows a clear difference in the kernel density estimates for WAst and LAst. While the distributions show a large overlap, the *WAst* distribution is centered around a higher value than the *LAst* distribution.[1] This does not mean, however, that the winning team's assist total is always higher than the losing team's assist total. Rather, the difference between the winning team's the losing team's assist total $AstDiff = WAst - LAst$ seems to be normally distributed itself. While most of the winning teams do seem to post higher assist totals than their counterparts, a relatively large amount of games have been won by teams that have posted lower assist totals than their counterparts.

[1]Notice the use of the word center as opposed to mean, as the center of the distribution is likely close to the sample mean but potentially different. Additionally, the difference in mean could be determined as "significant" through hypothesis testing, but this exploration section is mostly concerned with exploring patterns rather than labelling sample distributions

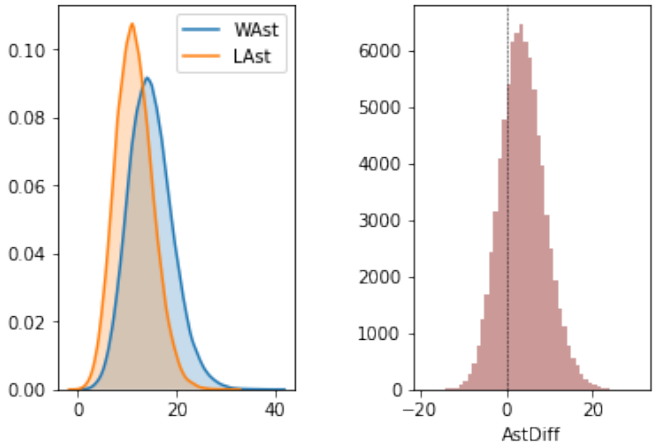

Figure 3: Kernel Density Estimates of WAst and LAst; Histogram of the assist difference between winning and losing teams

Similarly, turnovers result in the loss of possession of the basketball, meaning the team loses the chance to score points in that particular offensive possession. Therefore, a higher value for turnovers should have a negative effect on the number of points scored and ultimately could be a factor in losing. When looking at the scatterplot in Figure 5, there does not seem much of a correlation between the two random variables.

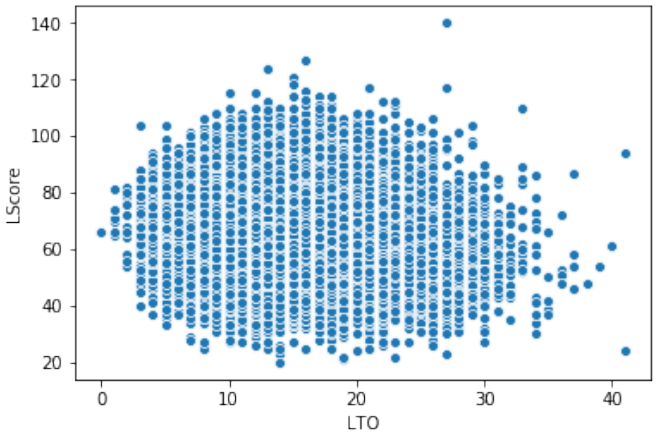

Figure 4: Scatterplot of LTO vs LScore

This is confirmed when looking at Figure 5, which shows the estimated distribution for both *WTO* and *LTO*. The distributions seem much more similar as compared to the assist distributions, and only show a slight difference in terms of the center value. It is important to note that a smaller number of turnovers is desirable, so lower values for winning teams are not a surprise.

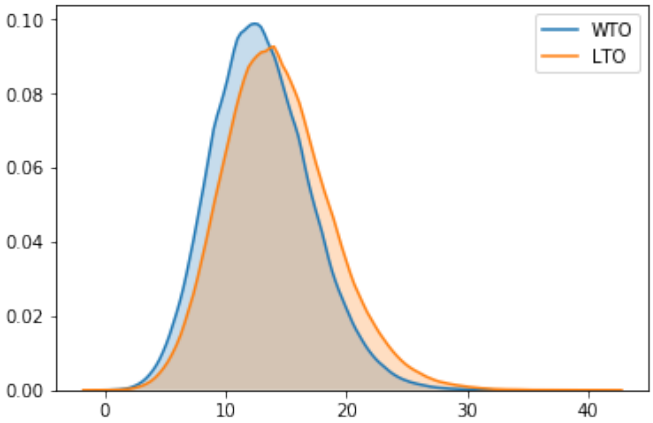

Figure 5: Kernel Density Estimates of WTO and LTO

In addition to those two random variables, I created similar plots for all game statistics. All of them confirmed my intuition, namely that winning teams on average record a higher number of positive actions and a lower number of negative actions within a game, which seems fairly standard.

One last interesting observation was that both winning and losing teams seemed to take a similar amount of shots each game, meaning the score difference originates from the winning team making more shots in the same number of attempts. In a sense, both *FGM* and *FGA* are different variants of the same random variable process. When looking at a particular game as a reality segment of interest, a player attempting a shot constitutes an observation act. *FGA* counts the number of such observation acts, whereas *FGM* counts the success events. In a sense, each shot is an event with binary outcomes [miss, make]. Thus, combining the two would create the field goal success rate for each particular game. As seen in Figure 6, the difference in the distributions is much larger as compared to assists and turnovers. Again, this is most likely due to the fact that made shots have a direct impact on the outcome of the game.

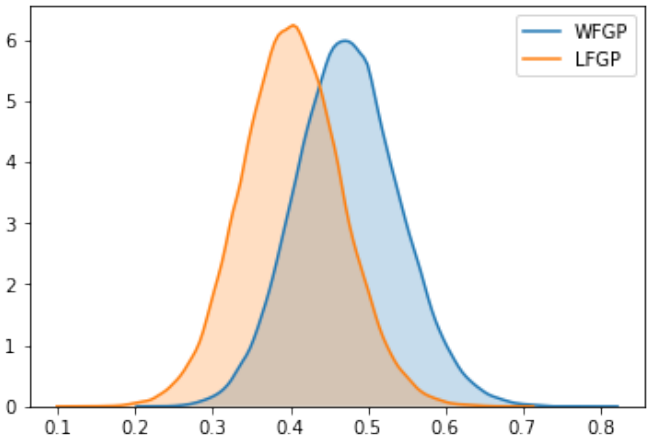

Figure 6: Kernel Density Estimates for Field Goal Percentage

4 Combinations of Random Variables

Besides the raw data, I also decided to extract some features. Most of these features would be described as advanced statistics in basketball terms, but can mostly be described as combinations of random variables. An easy example would be the Assist to Turnover ratio (ATR), which is calculated as Ast/TO. Both are independent random values, such that $Ast, TO : \Omega \to \mathbb{N}$. The sample space of ATR is thus comprised of any possible solution for Ast/TO. By substituting in the sample spaces and looking at the limits for both, we can approximate the resulting sample space:

$$\begin{aligned} \lim_{Ast\to\infty} \lim_{TO\to 0} \quad & \frac{Ast}{TO} \to \infty \\ \lim_{Ast\to 0} \lim_{TO\to\infty} \quad & \frac{Ast}{TO} \to 0 \end{aligned} \tag{4}$$

$$ATR : \Omega \to [0, \infty)$$

There are a couple of things to note here, namely that combining the variables in such a way moves the sample space from $\mathbb{N} \to \mathbb{Q}$. As both Ast and TO only feature positive integers, however, the sample space can be further defined as $\mathbb{Q} \in [0, \infty)$. In addition, the equation for ATR requires that $TO \neq 0$. In practice, when a team commits 0 turnovers, the value is usually replaced by 1 such that $ATR = Ast$.

In terms of basketball, the actions that cause Assists and Turnovers are somewhat related, as both can be committed by attempting to pass the ball to a teammate. Yet, there are numerous other actions that can lead to a turnover as well. Overall, the measure is seen as an indicator for how effective a

team is in creating scoring opportunities by sharing the ball. Figure 7 shows the distributions for ATR, again split by winning and losing team. To no surprise, the distributions look slightly right-skewed which is not out of the ordinary for such a ratio. The boxplot in Figure 8 confirms this sentiment. The distribution for the losing teams is centered below 1, whereas the winning teams' distribution is centered above 1, meaning that losing teams are likely to record more turnovers than assists.

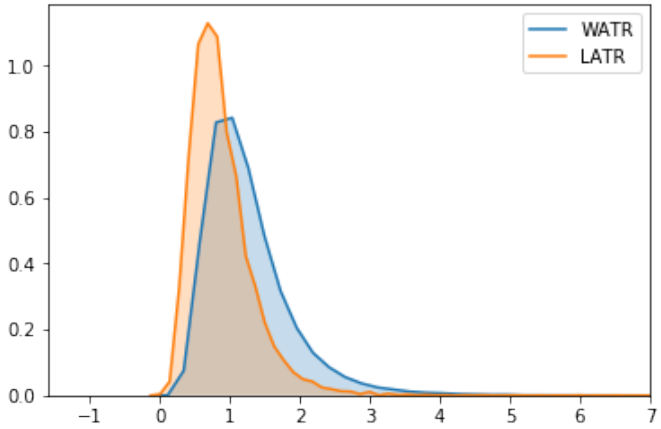

Figure 7: Kernel Density Estimate of Assist to Turnover Ratio split by game outcome

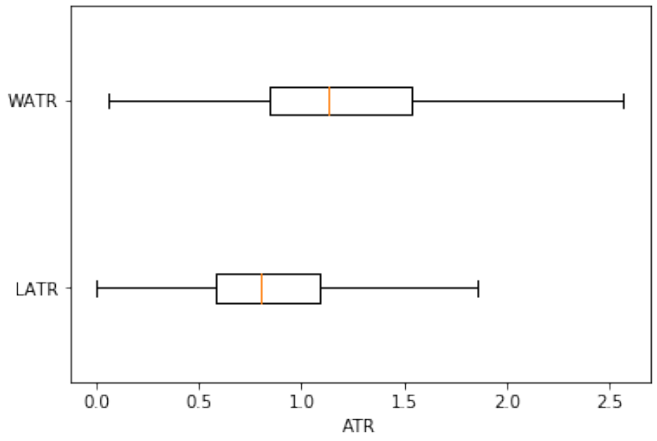

Figure 8: ATR Boxplot

Other than simple ratios between two independent random variables, there is also a basketball game characteristic that can be approximated from the data. A team can only score when it has the ball, so each duration of time where a team has the ball is labelled a possession, which can generally end in either a shot or a turnover. Possessions are not generally recorded as a game statistic, but can be inferred from the aforementioned statistics. Specifically, an often-used formula for Possession was presented by Ken Pomeroy [Pomeroy, 2004]:

$$Pos = (FGA - OR) + TO + 0.44(FTA) \tag{5}$$

Offensive Rebounds are subtracted from Field Goal Attempts, as they allow a team to retain possession. In addition, Free Throw Attempts are scaled because teams are awarded anywhere between 1 to 3 free throws depending on the preceding play, and 0.44 is a value that has historically yielded the most accurate approximations. The Pos statistic is then calculated for both teams, after which it is averaged as both teams are expected to have a fairly equal amount of possessions per game. Thus, possessions themselves are not a very helpful factor in assessing why teams win or lose games as they are the same value for both teams. Rather, Pos is usually used to calculate possession based efficiency statistics to differentiate between teams that post higher scores simply due to playing at a faster pace and teams that post high scores due to higher scoring efficiency.

Therefore, an often-used efficiency metric is the Points per Possession: $PPos = Score/Pos$. Similarly to ATR, the data value space is $[0, \infty)$. The kernel density estimates show a very clear split between the two distributions, with LPPos distribution centered around 1.0 and WPPos centered around 1.2. Considering most games have a possession value between [50, 80], this 0.2 difference is quite substantial.

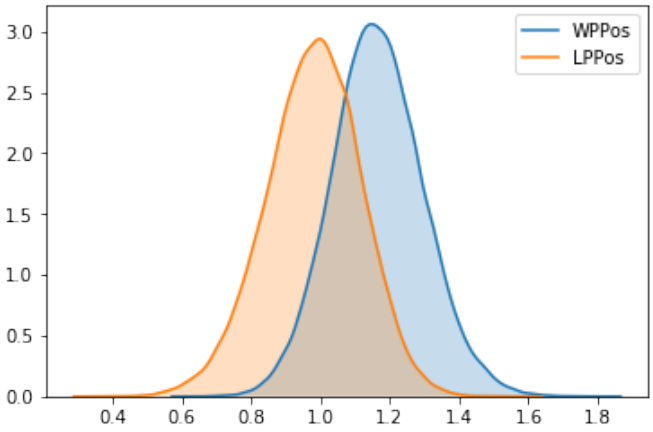

Figure 9: Kernel Density Estimate of Points Per Possession split by game outcome

Besides looking at the game outcome, this efficiency statistic also allows us to compare the progress college basketball has made over the years. Specifically, an important trend in professional basketball has been the increasing value put on the 3 point shoot for scoring efficiency reasons. Specifically, high-level teams average around a 45% 2 point shooting success rate and a 35% 3 point shooting success rate. Besides some other added benefits, the expected point value from shooting 3 point shots is higher given the success rates, which is why teams have taken increasingly more three point attempts over the years. Thus, it would be interesting to see whether college basketball teams have shown increased scoring efficiency over the years. Although the visual increase is very

slight in Figure 10, the overall Point per Possession have increased around 0.03 since 2003, which is still a fairly large increase considering the nature of the variable. It is interesting to note that both winning and losing teams exhibit the same pattern over the years, suggesting independence from game outcome.

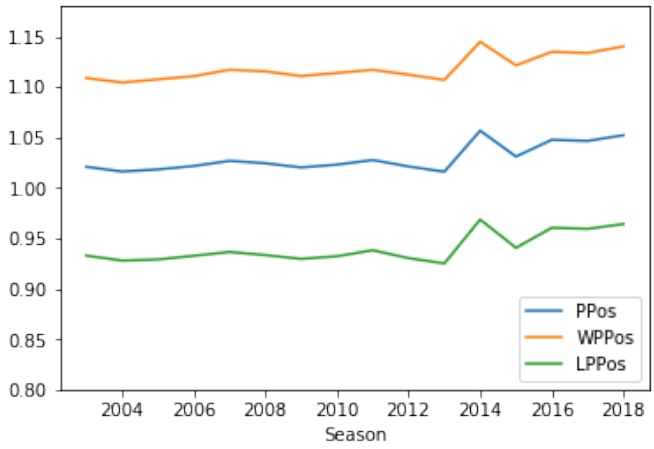

Figure 10: Average Points Per Possession split per Season

5 Modelling Considerations

After exploring some of the underlying trends in the data, the next step would be to use that information to model the outcome of the games. For now, all of the exploration initiatives have dealt with $P(X = x \mid Outcome = y)$, meaning we looked at the distributions of the random variables based on information about the game outcome. For a non-deterministic prediction of the game outcome, we are looking for the probability $P(Outcome = y \mid X = x)$. In addition, there are a couple of other challenges that present themselves with this particular dataset.

Since the dataset is ordered by winning and losing team, if we were to introduce a decision space $D = \{lose, win\}$ for the team whose ID is first mentioned, all decision values would be 1. An easy solution would be to duplicate each observation, rename W and L to T1 for Team 1 and T2 for Team 2 and the other way around for the duplicates. Not only would that create a perfectly class-balanced dataset, it would also artificially double the sample size.

Finally, the goal of the competition was to predict the outcome of future playoff games, for which none of the discussed game statistics are available yet. A basic approach would be to extract the average season statistics for each team and then to use those values to predict the outcome of the game. A lot of Kaggle competitors also use other 'advanced statistics' or the output of ranking systems

as variables, or adjust variables based on some measure of opponent strength. Those concepts are beyond the scope of this report, however.

6 Conclusion

This report analyzes the data generated over multiple seasons of NCAA Basketball in terms of random variables and their distributions. To no surprise, positive actions in a game are more prevalent for teams that ultimately win the game. In addition, I explored the distributions of certain random variable interactions that are commonly used in the context of advanced basketball statistics and confirmed their validity. Finally, I discussed the implications and challenges in estimating the likelihood of a team winning a specific game.

References

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

Kaggle. Google cloud & ncaa ml competition 2019 - men's, 2019. URL https://www.kaggle.com/c/mens-machine-learning-competition-2019/.

Ken Pomeroy. The possession, 2004. URL https://kenpom.com/blog/the-possession/.

Michael Waskom. Visualizing the distribution of a dataset, 2018. URL https://seaborn.pydata.org/tutorial/distributions.html.