Data Mining for Business – ISGB 7967/BYGB 7967
Professor Michael Deamer

Zara Jillings, Lauren Murphy, Miley Li, Nick Sansone

NBA Statistics & Player Analysis Data Analysis

Final Project Report

Abstract:

Globally, sport is something that brings people together and shapes lives. The NBA Is a professional basketball league, and one of the largest and most popular leagues in the world. The average NBA team valuation is over $2B and there are hundreds of millions of fans globally. Our analysis is focusing on statistics from the NBA (National Basketball Association), to understand what factors contribute most towards position/role on the court and understand how the “NBA Bubble” which was implemented during COVID affected statistics. This is interesting because sport has such a wealth of statistics and what the statistics show can provide significant insight into player and team performance.

Introduction & Overview:

Our Data Mining analysis involves looking at NBA statistics. Basketball is the most popular sport in the United States and one of the most popular sports throughout the world, with the NBA, based out of America, being globally recognized with teams even playing in international events in Europe and Asia to showcase the game and league. Basketball is also a key Olympic sport where the US has historically dominated. Competition within the NBA is extremely high and qualifications to make it to this league are especially rigorous. Statistics are used as a measure to analyze players and teams, both in basketball and sport in general, the insights gleaned from effective data analysis of these statistics can influence key player and team decisions, such as draft picks, salaries, strategic game decisions and more which at the end of the day could mean the difference between winning and losing.

As with all sports, particularly professional sports, data is king. Coaches, players, fans, scouts, reporters, and others rely on data to both analyze historical performance and predict

future performance. The 2011 film “Moneyball” laid the foundation and pioneered the implementation of advanced data analysis on various factors of a sports organization. These techniques aim to improve performance and efficiency both financially and on the court or field. Deep analysis allows members of an organization to improve their decisions about players, coaching strategy, game play, partnerships, salaries, and fan engagement, just to name a few. Within sports leagues, there are extensive volumes of data on the individual players, teams, and games that provide a rich dataset for analysis. This is especially true for the National Basketball Association as they also collect a wealth of statistics, providing lots of potential information for us to analyze.

During a time where professional sporting events have largely been canceled, the end of the NBA season was able to take place through an ingenious implementation of a “bubble” to isolate NBA players from the rest of the US population. At a time when collective morale was plummeting, the opportunity to watch America’s favorite sport helped people sort of forget the unprecedented lifestyle that this pandemic has brought the world, if at least for a short while. Along with our collective passion for the sport, our ability to continue watching this league when times were at their worst intrigued our group to go at analyzing the performance of top NBA teams to understand what makes them successful inside and out. Further, we wanted to understand whether performance of teams this year was affected by the isolation and implementation of the NBA bubble.

Our team has experience in previously analyzing NBA and NCAA Basketball statistics to determine what makes teams successful, however, this did not use advanced data mining techniques. We are interested in identifying more robust insights and potentially unexpected

associations within the data using practices learned throughout this course. Three of our four team members are also college and national level athletes, having played for the Fordham Women's Basketball team, the New Zealand National Basketball team, and Boston College's Baseball team. This overall interest in athletics highlights why this analysis is both interesting and relevant. As data and analytics continues to grow as a field, it has become integral in sports organizations and how successful teams function as a whole. We aim to identify, interpret, and understand "interesting" correlations within our data and detail these insights further than what can just be inferred.

Data Description:

In order to tackle this problem and take it towards manipulation, we collected data from the metahub of public datasets, kaggle.com. The plethora of data frames available on this website supported us in compiling several different sets together to create a diverse and robust dataset that can be modeled and analyzed to illustrate these interesting and important correlations that we set out to find. Within our dataset, we had 2018-19, and 2019-20 regular season and playoff statistics that included factors such as: points, minutes, rebounds, assists, turnovers, steals per game (and more!), shooting percentages, efficiency metrics, position and other statistics. This provided a robust enough dataset to have many inputs to analyse and build models with that helped us answer our problem statement.

Problem Statement:

We want to analyze the performance of top NBA players, and understand what factors influence their position and overall success. We are interested if statistical performance (i.e.

average points, steals, rebounds, etc.) relates to a players position and naturally categorises them as a guard, forward or center.

We are also going to analyze how the NBA Bubble affected performance this year. Without fans in the stadium, home court advantage, and travelling it was a very different season which likely impacted statistical performance (we expect positively), like shooting a better percentage, or committing less turnovers.

We expect to identify obvious associations while also uncover relationships that we did not anticipate, which may provide new insight into the sport that proves valuable at a higher level. Ultimately, this insight will lay the foundation to explore other broader associations that impact player and team performance in the future such as the influence and impact of team payroll/salary cap, player experience, coaching etc.

Methodology:

Our data required initial analysis, merging and cleaning in both Excel and SPSS. As these data sets did not come from "official" NBA sites, there are some factors that are incomplete and through the merging of several data sources, many columns are redundant. Before we could analyze and model based on our compiled data set, some aspects had to be cleaned (limiting the number of decimal places given for percentage stats, removing unnecessary/random characters that followed a players name, removal of some columns and combination of others).

After our preprocessing and "cleaning" created a usable set, we discussed what factors we wanted to include as input variables in creating our clustering model- which variables hold impact and are significant, and which variables can be ruled out as "noise"?

We had a robust enough dataset to analyse, but there is always room for improvement and so we considered the possibility of merging with even more data, expanding the number of seasons and more, but ended up retaining the more narrow scope for this analysis.

In our analysis, we set all the fields as inputs except those typeless ones. We use Clustering Analysis to get the results. In particular, we use the K-Means node and the TwoStep node and generate two models based on the 2019-2020 NBA player/playoff datasets. We also include graphs such as histograms, scatter plot, web graph, etc. to give us clear visualizations, and compare the results based on the two models. We also ran other variants to test different aspects and see what else we could find out.

Results and Discussion:

Through analysis of our models, there are underlying insights that can be gained when digging deeper into the meaning of them. Upon initial inspection of the K-Means clustering, it appears that this is not indicative of position- however, the heterogeneity of clusters in predicting player position confirms a "theory" that our group began with. Analysis of the proportion of each cluster created through our model shows that every cluster is made up of every position, proving to be inconsequential for predicting position. In principle, there are attributes that guards, forwards, and centers all have in common that are reflected in their stats (i.e. centers/forwards have higher rebounds and lower assists than guards, and vice versa for guards/small forwards) which we wanted to uncover through exploration of our mining and modeling techniques. The results proved not to be as clear cut as expected- even though the "positions" are categorical and play different roles on the court, there is a lot of variation among the statistics within position (i.e. a rebounding guard, a distributing forward, a power-house

scoring center). We tried to further study this relationship by using different combinations of inputs, different numbers of clusters, and even different modeling techniques but all of these contributed toward the idea that positioning cannot be clearly defined by clusters. Our TwoStep clustering model also contributes toward this understanding- a histogram created based on the RPG parameter (rebounds per game) does show a separation of "big men" and "skill" players (centers and forwards are most prominent as RPG increases), however there is still high entropy for position as RPG decreases. Another output that proved to be valuable in showing this paradigm was a histogram based on the APG (assists per game) derived from the TwoStep model. This plots what is thought to be a leading statistic for these "skill" players, however, it illustrates the same relationship presented in the RPG histogram (guards and small forwards most prominent as APG increases, but contains high entropy as APG decreases).

To explore this further, we also ran a modified clustering analysis using only a select number of inputs (core statistics) we thought may have provided more meaningful results (FTA, 2PA, 3PA, PPG, RPG, APG, SPG, BPG, TOPG). Outputting RPG v PPG and APG v PPG with overlay of cluster and position (grouped from 7 positions to 3) showed that at the high end guards do have more assists and forwards more rebounds, etc. but with no clear clustering. The conclusions here are similar to those explored about as there is still high entropy and lots of crossover between player position.

We also took a deeper look at statistical categories from the 2019 and 2020 playoffs. Due to the impact that COVID-19 has had, the 2020 NBA playoffs took place in a "bubble" where only essential personnel (players, coaches, trainers, etc.) were allowed in. This prevented in person attendance and essentially eliminated the aspect of home court advantage and travel

between arenas. Two statistics that we predicted would be most affected were turnovers per game and shooting percentage, both by team, not individually. We also narrowed it down to the thirteen teams that participated in the playoffs both years because we felt that it would yield more accurate results. After looking at turnovers per game, we noticed that there was a significantly lower deviation from the mean in 2020 as opposed to 2019. We suspected that this could definitely be attributed to the factor that all of the teams played every game on the same court. The other statistic that we analyzed, shooting percentage, did not give us the results that we expected. We saw no significant correlations between 2019 and 2020. We did notice that most of the teams that had a higher seed in 2020 improved their shooting percentage and vice versa, but there was no evidence to support that any of our results were affected directly by the bubble.

Conclusions:

This analysis provided insight into the way the NBA functions in today's league. Based on our modeling and output nodes within SPSS, it initially appeared that there was no way to differentiate between positions using clusters. As this appears insignificant, it actually speaks a lot towards the "theory" our group had- as the game of Basketball changes, players have to transform and adapt their game to keep their competitive edge. There is no explicit categorical way to divide player position, implying that the game is becoming more fluid, meaning that players are no longer defined by position in their role on the court. Our histogram derived from the TwoStep model illustrates the slight differentiation among big men (forwards and centers) and "skill" players (guards and small forwards) on the factor RPG (rebounds per game), however it is still very mixed in the clustering of positions. This speaks to the necessity of players having

the ability to do everything on the court in today's game. Although an "interesting" insight, our team wanted to further question why. What has caused the game to change like this?

We found that it is a result of advanced analytics and their integration into the game of Basketball. Learning from these analyses helps a team to make informed decisions to enhance their overall performance. Today, one of the most important statistics is a players +/-; this is a measure of the point differential between one's team and the opposing team while that player is on the court. It is quite obvious why this holds so much importance as a high +/- implies that a team will score more than the opposing team while a specific player is on the court. However, this goes further as the +/- is simply a measure of general scoring, rather than individual statistics. Therefore, it causes teams to pursue players who rate high in this category with less regard for their other stats. This establishes a generalized and uniform player profile for those who are the most competitive and can lead a team towards success, blurring the lines between positional roles. With the implementation of advanced analytics and emphasis on this category, players have adapted towards being able to play all roles and positions on the floor, with some being better in measured categories than others. This directly reflects the insights gained through meticulous analysis of our data and models created. As future analytics continue to develop with the advancement and ubiquitous use of big data, the game will continue to change in ways that favor the competitive advantage given by analytical techniques in sports.

This analysis has opened the door for us to dive deeper into NBA statistics and the other interactions between players, teams, salary, staff, etc. This could include looking at more historical data to try and predict a players future performance, to predict a teams success based on individual players, to find the most powerful combination of features that impacts +/-. All of

this is interesting because not only does it take into account data, but it also then requires a strategic understanding to apply the results when making decisions. For example, having a team of the most high performing players statistically, may not be effective because they all want to score. You need a balance of players and skill sets, and then intangible things such as team chemistry and grit in order to be successful. Data can guide this decision, but someone who can interpret it needs to make the final call with a human-centric focus.

As a group of data scientists, we recognize that while the future of not only the NBA, but the world, will be reliant on data mining, that you still need to understand what the data can't capture and how that affects the results. We are excited to consider other developments and to see how future NBA performance may continue to change.

Appendix:

Task	*Completed by*
Abstract, Introduction, Problem Statement, etc.	All
Data Collection/Creation of Dataset	Nick Sansone
Data Preprocessing	Nick Sansone
Methodology	All
Data Modeling using K-Means	Miley Li
Data Modeling using Two Step	Miley Li
Data Modeling, Modified K-Means Clustering	Zara Jillings
Data Modeling comparing seasons	Lauren Murphy
Results & Discussion	All
Conclusions	Zara Jillings, Nick Sansone
Presentation	All