Gregory Indick and Cameron Mitchell Final Project Report

IST 707 12/16/2021

1. Introduction

Every year, the National Basketball Association hosts its annual first year player draft where teams take turns selecting the players they deem the best out of all those who are draft eligible. This process takes place across most professional sports leagues in America, such as Major League Baseball and the National Football League, and the idea is to ensure competitive balance. If there were no draft, and teams were able to sign whichever rookies they wanted, chances are the same teams would sign the best players every year, and those teams would dominate the league. This would make the league quite boring and uninteresting to follow, and thus the fan base would dwindle and the sport would make a lot less money. The aspect that makes the NBA draft unique from other leagues is the draft lottery that determines the draft order each year. In other leagues, the team with the worst record from the previous season gets the number 1 pick, second worst gets the number 2 pick, and so on. In the NBA a lottery is used to determine the order. The team with the worst record has the best odds at obtaining the number 1 pick, but they are not guaranteed to get it. This is in place so that teams are less incentivised to not try to win towards the end of the season and try to get a better pick, as your record is not always directly correlated to your pick.

One might be able to infer a major issue that teams have when it comes to the draft; there are hundreds of draft eligible players each year, and the amount of data teams have on these players is quite vast. Teams have to decipher all of this information and determine who they deem to be the best players. Additionally, these decisions are not just about the talents of the players, rather teams must also take into account what their team specifically needs even if the

player their team needs is not necessarily the best overall player available. Clearly, these decisions are complicated, and there is a lot that goes into them. In this project, we analyzed the 2021 NBA draft, and we first aimed to group players based on statistical performance using clustering in order to get an idea of the different types of players in our dataset. After that, we used regression modeling to analyze which player metrics were most important to teams this year, as well as factoring in team needs and our generated clusters in order to get an idea of which players got drafted this past year, as well as try to see how well our model could predict the order of this past year's draft.

2. Dataset Description and Preprocessing

a. Dataset Description

While initially trying to find a dataset to use, we wanted to make sure that there were enough metrics so that each player was unique, and that when we performed our clustering techniques the clusters were clearly defined. This led us to grabbing two separate datasets, one from Real GM, and another from the College Basketball Reference. From these sources, it left us with one dataset containing player information, and the other focusing on team information. To obtain our data, we started by taking the top 1000 players from the NCAA 2020-21 season in terms of points per game. We then added to that the top 500 players in other statistical categories, such as rebounds/game, assists/game, steals/game, etc.. Naturally, there were lots of duplicates, as players could be in the top 500 in multiple categories, so we removed duplicate rows. We then added in the metrics for players that played on the G-League Ignite team, a team in the NBA's development league created last year for the top high school players to skip college and prepare for the NBA draft by playing professionally. Lastly, we decided that for the International players,

we would take the top 50 players in terms of points per game from each league that had a player get drafted. The final dataset containing player attributes was composed of 1727 rows, and 30 columns, all detailing player statistics. These included Player Name, Field Goal Percentage, Offensive Rebounds, and Free Throw Percentage among others. We also added a few dummy variables, such as whether the player played in the NCAA, G-League, or was an International prospect, and whether the player was drafted or not. The team dataset contained similar columns, with a total of 24, but only contained one row for each separate NBA team, which came to a total of 30.

b. Preprocessing

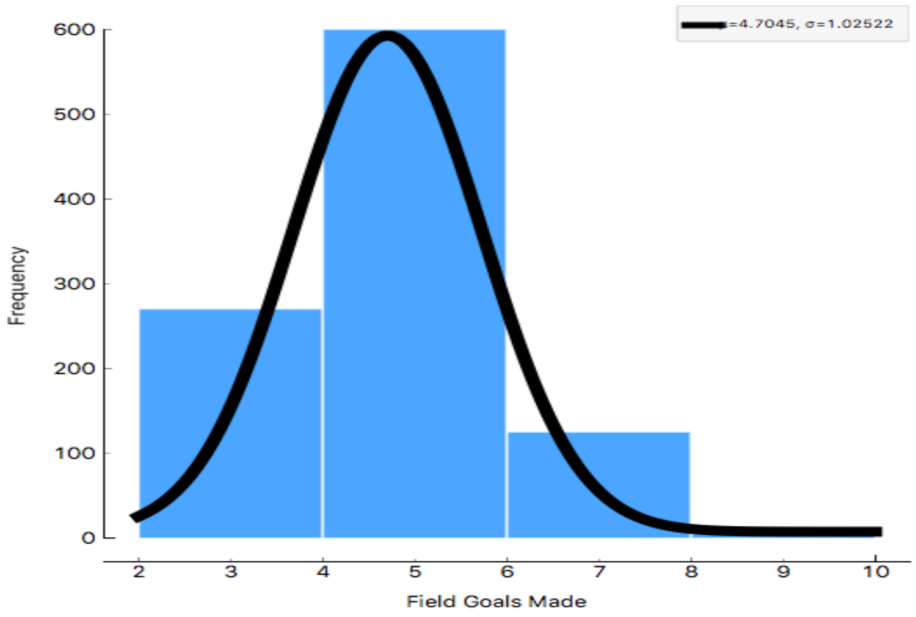
After choosing which datasets we wanted to include in our project, we decided which preprocessing steps needed to be done. Each dataset was examined at its face value, and then we searched for any null or missing values. When there proved to be only a few, we decided that keeping them would be the best course of action. This did not affect any processes we performed later on in the project, such as clustering or linear regression. We then looked at each column separately and decided whether or not discretization would need to be implemented. Since we wanted each player to have a unique statistical spread, we found that adding discretization to some columns would not be beneficial. Once the minimal cleaning was done to each data set, we then added our clustering results generated in Orange to a column within the player dataset. This would then be used to help our linear regression model we developed later in our project's life cycle.

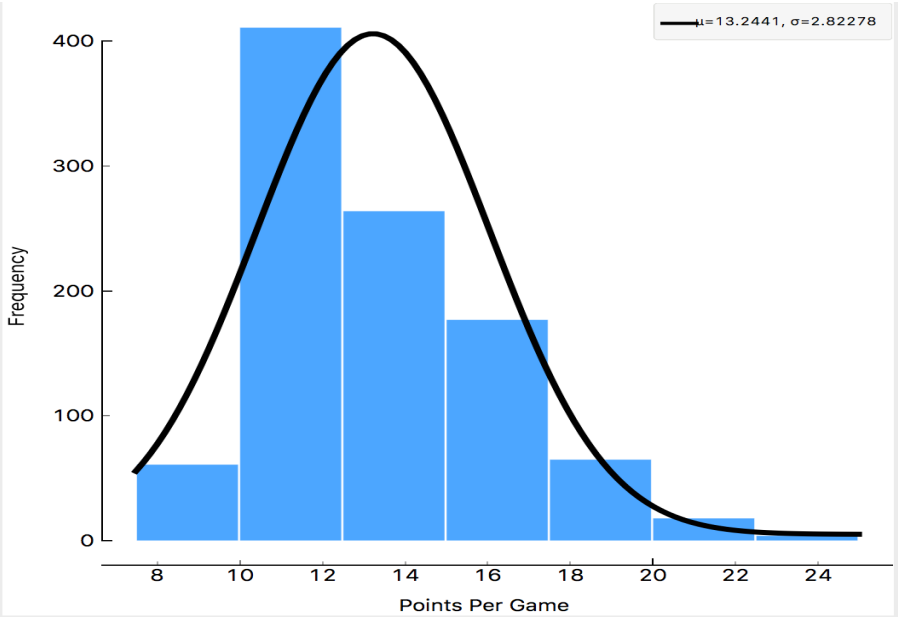
Next we decided to rank each team attribute based on how well they performed in that area, for each individual team. For example, the Charlotte Hornets were ranked 22 out of 30 for their Field Goal Percentage, and 9 out of 30 for their Three-Point Rank. This was done to every

team, so that we were easily able to tell which team was the best in each category, as well as the worst. Once this was done, we had our final datasets we would use moving forward with the rest of our project. Each one would play a specific role in both our clustering and linear regression model, and without the preprocessing steps we took to make sure the data was as clean as possible, we would not have gotten the results that we did.

c. Distributions

Our last step in the preprocessing stage was to examine some distributions that occurred within the player dataset. Since the team dataset contained such a small amount of rows, we decided to skip this set for the teams, since we believed we would not find any valuable information. Below are two distributions that occurred within the player dataset:





As seen above, both histograms contain a normal distribution throughout, and this was determined to be the case for every other distribution we tested as well. With no graphs sticking out to us as vital, we decided it was time to move on to the clustering phase, so that we could examine the players in a more thorough manner.

3. Clustering

a. Initial Approaches Attempted: Clustering and Decision Trees

With our datasets cleaned and ready to use, we imported them into Orange and began our introductory analysis. With our initial goal being that we wanted to select which teams needed players based on specific attributes, we decided that clustering and decision trees would be the best algorithms for this. For both the player and team dataset, we attached them to the decision tree, hierarchical cluster, and k-Means clustering module, respectively. We immediately noticed that there needed to be a target variable for the decision tree, and realized that this caused an issue. There were only 60 players out of our 1700 included in the data that were picked for the

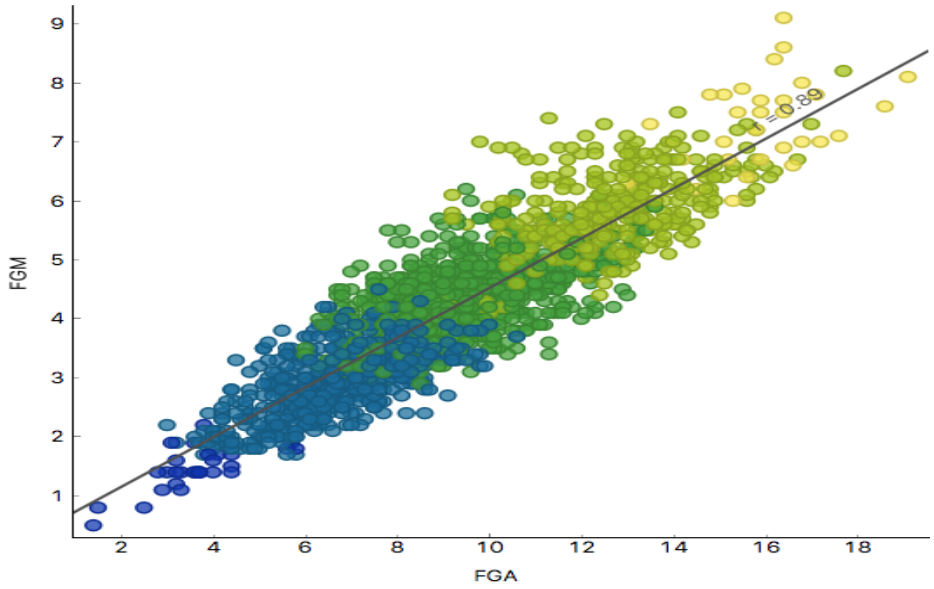
NBA draft, so we were unable to pick a target variable that would give us reliable outcomes for our decision tree.

With decision trees unable to give us any valuable output, we decided to then focus on hierarchical and k-Means clustering. For hierarchical clustering, we used Euclidean distance for the measuring, and did not prune any of the rows. We then tested each linkage to see which would give the most consistent results, and it turned out to be Average Linkage. Once this was done, we made sure that the "Pick" column was ignored since it did not provide any valuable information for the clustering results.

Before reviewing our results for the hierarchical clustering, we decided to set up our k-Means clustering first, so that we would be able to compare the two. For these clusters, we made sure all of the columns were normalized, and that the "Pick" column was ignored again. For our initial results, we ended up using a range of three to eight clusters, giving us three clusters to work with for each projection.

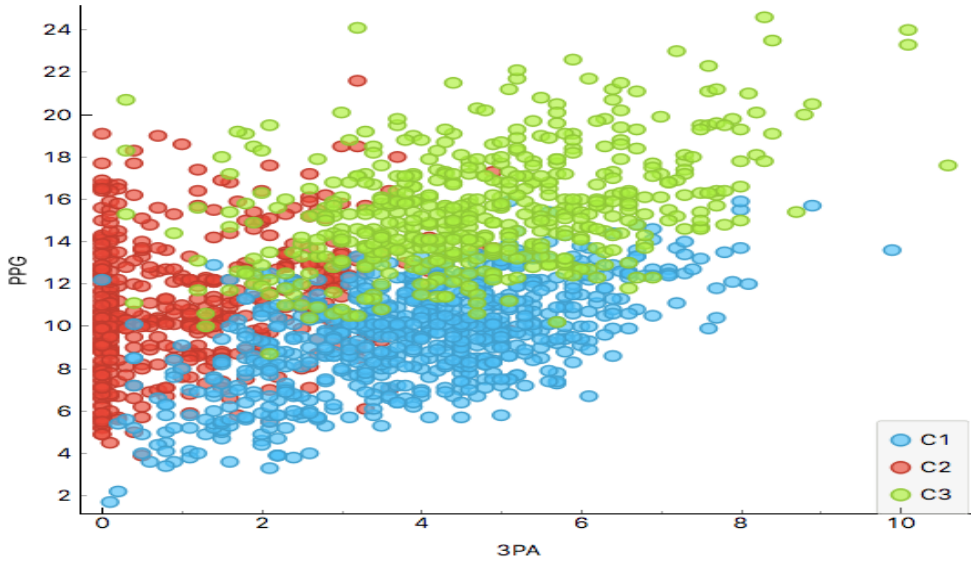
b. Initial Approaches Attempted: Preliminary Results

With decision trees being unviable, we solely focused on hierarchical clustering and k-Means clustering for our preliminary results. Hierarchical clustering was the first module we examined, and to our dismay, we did not find as much information as we were hoping for. While it gave some valuable insight into a regression cluster that can be seen below, other sets of clusters were not defined as we had expected.



As seen above, there is a steady regression increasing when comparing field goals made to field goals attempted. This, however, is not a projection that is too informative, since we were able to deduce that the more times you attempt to score a field goal, the more likely you are to score one. With hierarchical clustering not giving us the results we had expected, we then moved on to k-Means clustering.

With k-Means, we were generating better results compared to hierarchical clustering. The clusters were more defined, and there were more instances where comparisons of two different attributes provided valuable insight.



Above is an example of a k-Means cluster group, and how defined they were compared to our hierarchical groups. Each cluster is distinct in that it shows us what types of players excel at particular attributes. In this example above, it is a comparison of three point attempts and points per game. Cluster Three shows us players who are great at scoring, since they average more than 12 points per game, and also have many three point attempts. Cluster Two in contrast, shows players that do not attempt many three point shots, but can still score more than 12 points per game in some instances. This is just one example of how k-Means helped us examine our players more closely, and how we could use these clusters in our linear regression model for predicting the draft pick.

c. Initial Approaches Attempted: Conclusion

Once that preliminary clustering was complete, we came to the conclusion that both decision trees and hierarchical clustering were not the best methods to use going forward, and focusing primarily on k-Means clustering would give us the best results. This being the case, we also decided three clusters would not be enough, and for our final results we would need a higher cluster count, so that each player was clustered more according to their strengths and

weaknesses. Focusing only on k-Means clustering also gave us the opportunity to have a more focused goal, and would assist us more with our linear regression model.

d. Final Results: k-Means Clustering

After we examined our preliminary results more closely, the first aspect of our model that we decided to change was the amount of clusters that were to be generated. We made sure to keep our columns normalized, but decided that our range of clusters was too low. We changed our parameters from three to eight clusters, to five to eight clusters. This gave us six clusters to work with instead of three, and was the largest amount of clusters we could produce before the projections started to get less clear. We then examined our findings, and added them to our player dataset, so that we could see which clusters the drafted players ended up in.

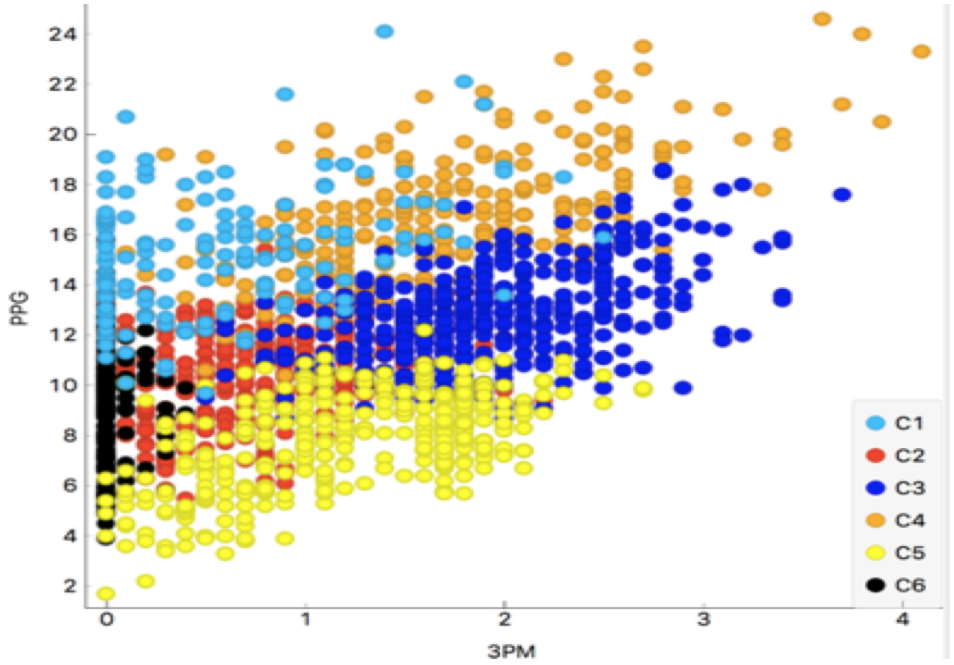
Cluster	Drafted Players
C1	10
C2	13
C3	11
C4	18
C5	5
C6	3

Above is a distribution of drafted players and which cluster they ended up in. Cluster Four had the most drafted players, while cluster Six had the least. Cluster One, Two and Three are also important, with there being a relatively even split of players between the three of them. Once we were able to see where the drafted players ended up, we decided to use this as an

opportunity to shift the focus of our project. While our initial goal was trying to predict the outcome of the 2021 NBA draft, we came to the conclusion that this would prove too difficult given the data and models we were using. We then singled out the 60 players who were drafted from our player dataset, and began to focus on what attributes were most important when being drafted. k-Means clustering would help in this area tremendously, which is why we then immediately started to examine our clusters more closely.

Below are some of the most informative clustering results we were able to find, and our deductions that were made about them:

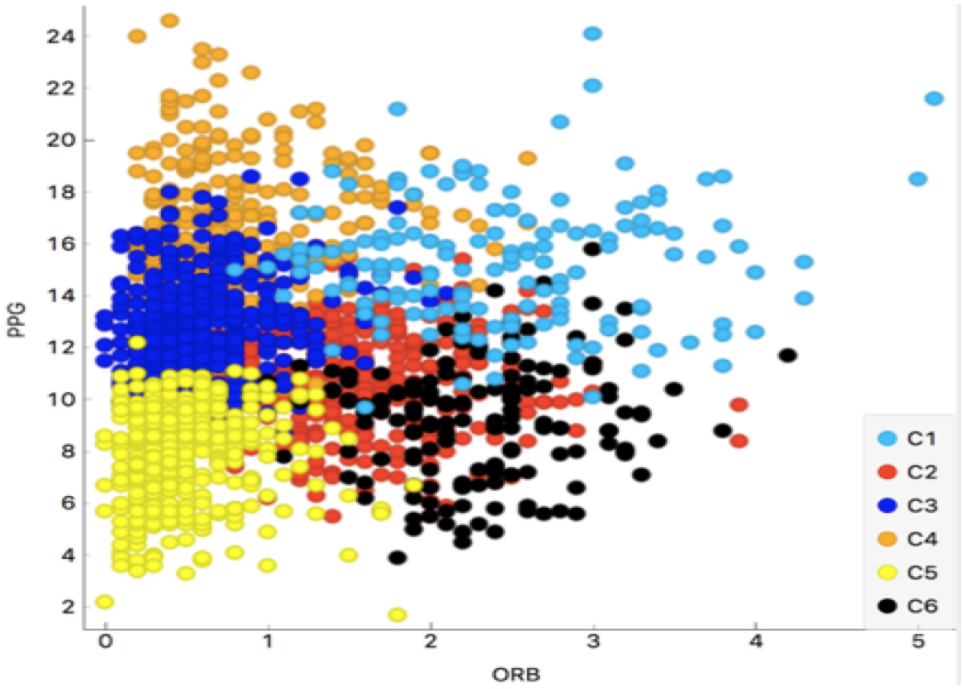
Cluster Four



Cluster Four contained 18 out of the 60 drafted players, which was the highest out of all the individual clusters. Looking at the projection above, which is comparing points per game to three pointers made, it is easy to deduce why. Cluster Four contained players who scored over 14 points per game on average, and also made multiple three pointers per game. This leads Cluster Four to be composed of high-scoring players, and most teams would want a player who can

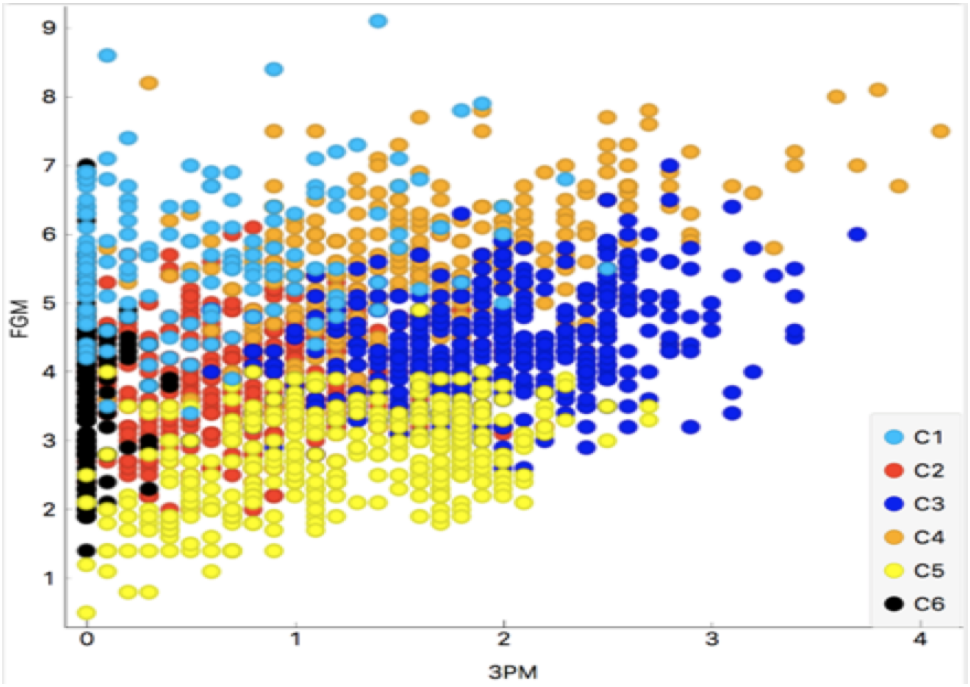
score large amounts of points per game. Comparing this to our player dataset, we found that the players within this cluster who were drafted were among the highest scoring players compared to the others.

Cluster Two



Cluster Two, on the other hand, is more of an all around player compared to that of Cluster Four. Cluster Two contained 13 players of the 60 that were drafted to the NBA, and in the projection above, which is points per game compared to offensive rebounds, it outlines why they might have been chosen. This graph shows Cluster Two at being relatively successful in both offense and defense, with rebounds and points per game being around the middle of the graph for both. This reinforces the idea of Cluster Two containing players that do not excel in only one field of play, but are good to above average in all areas of play.

Cluster Six



Cluster Six is the last important cluster, given that it contained the least amount of drafted players, with only three out of the 60 being chosen in this grouping. As shown above, when comparing three points made to field goals made, they do not score any three points per game, and can score anywhere from two to seven baskets per game. This does not lead to them being high scoring players, which counts for a lot in the draft. When comparing them to Cluster's Four and Two, they lack in both all around capabilities, as well as the amount of points they score per game. This leads to them being much less sought after compared to the other two groupings, which is backed up by the amount of players in Cluster Six being drafted.

e. Final Results: Assessment of k-Means Clustering

With all of our clusters examined, we took our findings and combined them with our player dataset. This was then added to our linear regression model, so as to hopefully improve the model as much as possible. Our clusters were much more fine-tuned compared to our preliminary results, and gave a better picture of how each player could be categorized. We hoped

this would bode well for our model, given that eachs player's skills and attributes were highlighted by our projections.

k-Means clustering turned out to be a very valuable tool, especially after changing the focus of our project to determining what metrics a player who gets drafted has. While not perfect, our clustering results grouped players by their attributes and gave us a clearer picture of the dataset we were working with. It showed us high-scoring players, low-scoring players, and players who were all around balanced players.With this information gathered, we were able to take it and all of the previous information we had collected, and put it into our linear regression model.

4. Linear Regression

Initially, our goal was to try and create a model/algorithm that would be able to predict the 2021 NBA draft out of all of the players in our dataset. Unfortunately, this task proved to be too noisy and difficult for reasons which will be discussed at the end of this paper. We attempted several methods to predict our drafted column, including creating a decision tree, a naive bayes model, and linear regression model, but none of the methods we tried gave us results that made sense, and the recall and F1 scores were around 50%. This led us to shift our approach slightly, and instead move towards an analysis of what metrics were most important amongst those players who were drafted. To do this, we filtered out the 60 drafted players and created a few models to determine what teams were looking for in players this year.

a. Player Metrics/Dummy Variables Model

First, we generated a model predicting draft pick that only contained player metrics and dummy variables to see which ones most impacted draft position. Since there were only 60 drafted players, we were able to obtain some advanced metrics to add to the ones we already had,

as well as player measurables including height, weight, and wingspan. We generated a variable called wingspan_diff, a player's height minus his wingspan, as this is really what teams are interested in when it comes to a player's wingspan. After obtaining these new variables, we generated the model below.

Table 1

Variable	Coefficient	P-Value
Intercept	173.485	0.196
FTM	-4.973	0.147
STL%	-7.569	0.064*
BLK%	2.516	0.097*
PPG	0.104	0.935
Height	-3.499	0.022**
Weight	0.292	0.045*
wingspan_diff	0.951	0.416
International	-9.141	0.315
G-League	-41.891	0.010***
PER	1.105	0.394
ORtg	-0.684	0.119
DRtg	1.476	0.019**
eFG%	0.019	0.979

In Table 1 above, the first thing to note is that a negative coefficient indicates that draft pick and that variable had a positive relationship, as 1 is the highest draft pick and 60 is the lowest. Some of the most important metrics for a player getting drafted this year according to

this model were his steal percentage, his height, his weight, his block percentage, and his offensive and defensive rating. Weight, defensive rating, and block rating were inversely related to draft pick, so players who were high in these categories actually got drafted lower in the draft. The variable that deserves the most attention here is the G-League variable, as it is the most significant and has the greatest coefficient in terms of absolute value. As was mentioned earlier, this was the first year of the G-League Ignite team, and three of their players were drafted: Jalen Green(#2 Pick), Jonathan Kuminga(#7 Pick), and Isaiah Todd(#31 Pick). Since only three players were drafted and two of them were in the top ten, it makes sense that the G-League variable is so strong in comparison to International and NCAA players. Whether the fact that these players were drafted so high due to playing in the G-League is difficult to determine, as who knows if their draft position would have been different had they gone to college. It remains to be seen just how valuable this G-League experience is for players, as a few years of data at least will be required to make an accurate determination. After this metric analysis, we decided to see if any of the clusters discussed earlier were significantly related to draft pick this year.

b. Cluster Model

This next model we generated contained only dummy variables for which cluster the player belonged to. For example, if the player was in cluster 1, he received a 1 for the variable Cluster1 and 0 in Cluster2, Cluster3, etc.. We did not include any player metrics in this model, as there likely would have been multicollinearity issues since these metrics were used to create the clusters. We decided to leave Cluster 4 out as the reference category since this cluster had the highest number of drafted players, as mentioned earlier. The results of the model are depicted in Table 2 below.

Table 2

Variable	Coefficient	P-Value
Intercept	23.389	5.83e-7***
Cluster_1	9.211	0.188
Cluster_2	10.842	0.095*
Cluster_3	11.247	0.099*
Cluster_5	6.211	0.486
Cluster_6	12.944	0.241

As to be expected, all of the clusters had positive coefficients in comparison to Cluster 4. This makes sense since Cluster 4 had the most players drafted and contained 4 of the top 6 picks. Cluster 2 and Cluster 3 were both significant, meaning there is a good chance their coefficient values are different from 0 when compared to Cluster 4. Clusters 5 and 1 appear to be the next best groups of players since they have the smallest coefficients, while Cluster 6 was the worst for a drafted player to be in. This means that the metrics that were key for players in Cluster 4, 1 and 5 were most important for determining a player's draft position this year. Lastly, we wanted to generate a model that would be able to predict the draft order of these 60 players, and we wanted to factor in the team needs aspect to accomplish this.

c. Team Needs Model

Our last model combined the aspects of the first two along with certain "ranks" for the team that made the pick at that particular spot. These rank values are the ones we discussed earlier where teams were given a value 1-30 based on their performance in different categories.

The results of this final can be seen below in Table 3, along with the predicted Top 10 of our model in Table 4.

Table 3

Variable	Coefficient	P-Value
Intercept	210.63111	0.1606
FTM	-7.7153	0.0757*
STL%	-7.0607	0.1215
BLK%	3.42907	0.0776*
PPG	1.24485	0.5712
Height	-4.34303	0.0109**
Weight	0.36864	0.0391**
wingspan_diff	1.14545	0.3851
International	-2.02131	0.8543
GLeague	-41.44261	0.0541*
PER	0.58395	0.7066
ORtg	-0.36524	0.4747
DRtg	1.59416	0.0269**
eFG%	-0.17951	0.8199
Cluster_1	-1.2099	0.9058
Cluster_2	5.24099	0.6101
Cluster_3	0.20699	0.9812
Cluster_6	-0.1425	0.9933
Cluster_5	-2.29009	0.8941
FG_RANK	-0.05557	0.8988
3P_Rank	0.0625	0.8571
TRB_Rank	-0.69383	0.0618*
AST_Rank	-0.50044	0.2949
BLK_Rank	-0.42428	0.2957
STL_Rank	-0.09555	0.7996

Table 4

R-Squared: 51.2%

Player	Actual Pick	Predicted Pick
Jalen Green	2	1
Tre Mann	18	2
Evan Mobley	3	3
Jalen Suggs	5	4
Jonathan Kuminga	7	5
Corey Kispert	15	6
Franz Wagner	8	7
Cade Cunningham	1	8
Kai Jones	19	9
Bones Hyland	26	10

As can be seen in Table 5, our model accurately predicted 6 of the top 10 players, and three of the other 4 predicted top 10 were picked in the top 20, so they were not too far off in their prediction. The one true outlier, Bones Hyland, was predicted 10th by our model yet he was picked 26th in reality. This is likely partially due to the fact that he ranked in the top 10 in both steal percentage and free throws made, as these both had highly negative coefficients in our model. Interestingly enough, Bones is off to a very strong start in his NBA career for the Denver Nuggets. In limited playing time through his first 17 games, he is averaging about 18 points, 2 steals, and 4 assists per 36 minutes played. These numbers are quite respectable for a first year player especially considering where he was drafted, so perhaps our model helped us discover a

future gem who when people look back on the 2021 NBA draft, people will say he should have been drafted higher.

5. Conclusion/Factors to Improve

After some thought and consideration, we determined there are a few factors that impacted our analysis and are partially responsible for our R-Squared being only around 50%, some that we could account for and some that we could not. While they may not represent all 50% of the remaining variability in draft pick, they certainly account for a good percentage of it. One aspect that we could potentially account for in the future is pre draft competition skill level. The skill level varies greatly in different leagues throughout the world and even in different conferences within college basketball. It is impossible to compare the stats of two players without noting who they obtained those stats against, as players may be able to put up good numbers against poor competition, but that does not mean they are NBA players. If we could have created some sort of control for this pre draft competition skill level, that would have helped our results greatly. Moreover, one aspect of draft evaluation that simply cannot be quantified is a player's character and his work ethic. Everyone in the NBA is talented, so a team is going to want a player who will work hard to improve his game. If a player does not have this drive to improve, he will eventually be passed by someone who comes after him, so he has no value to a team. There is no real way to quantify this character evaluation, so it is simply something that could not be controlled for in our project. Lastly, a player's injury history is also a major factor in his draft evaluation. This not only includes whether a player is currently injured at the time of the draft, but also if he has had a series of injuries that teams are worried could be chronic issues. A team is not going to want to invest a high draft pick in a player who may be not

be available throughout his career due to injury. Similar to player character, there is no straightforward way to quantify injury history, so we could not control for it. Overall, despite these factors, we feel that our analysis does an adequate job of analyzing the 2021 NBA draft, identifying the skillsets of different players, determining which factors were most important to teams, and predicting the order of the draft. We also discover just how important the factors mentioned in the conclusion are to our teams. As time goes on, we will be able to tell if our draft or the real draft was a better representation of the best players in this draft and just how valuable our identified skills were to teams, but as these players are still currently rookies, the truth remains a mystery.

Table Key: Models

Orange	Significant at 10%
Green	Significant at 5%
Yellow	Significant at 1%