National College of Ireland

NBA Draft Analysis

BSH in computing

Data analysis 2020/2021

Alex Hussey

17764291

X17764291@student.ncirl.ie

Contents

Executive Summary

The National Basketball Association (NBA) draft is the primary gateway for young athletes to move on to the professional and to the elite levels of basketball. All college and international players worthy of transitioning to the next level of play dream of being the number one draft pick or even a lottery pick for one of the 30 NBA franchises.

This project looks at how each pick in the first round of the draft is used by each team, and how a prospect can look to be for a team given previous players statistics, number drafted, and team drafted to.

Who would be interested in this data and could use it in the future? The NBA General Managers (GMs) would be! The NBA is a multi-billion-dollar industry made up of 30 teams, with two divisions and 15 teams per division. The difference between being a good GM and a bad can come down to who you draft, what trades you make and of course what staff you need to bring in.

1.0 Introduction

1.1. Background

The National Basketball Association is the most watched sport of different cultures in the United States, with roughly 47% of African Americans watching the play offs in 2017 with an average of 17 million fans watching. Along with this, the players are considered to be the best paid athletes by average annual salary per player. With this much money invested into the league as a whole, why is there not a bigger influence on scouting for the future of the sport?

The NBA draft consists of 2 rounds of 30 picks, one first and one second round pick for all 30 teams in the NBA. All teams are able to trade their picks in order to move up or down in the draft and or to trade for a player. This can allow for all 30 teams to improve and reinforce their side.

For NBA teams, the draft can be a part to turn a team's future around, or to acquire more assets for the future or contend for a championship. For a player to register for the draft, must meet these conditions. For North American prospects, complete 1 year of college and be at least 19 years old, and for international players, either need to be 22 years old in the season or have played for a professional team outside of the NBA. Or a new rule introduced in the last two years, a player can sign for a G-League team, or minor league for NBA teams to develop young players.

With 2020 and covid being so prevalent, there was no college basketball in March, which is the best time for scouts to see the top prospects, which caused some players to go under rated and fall in the draft. With an NBA team needing to select the right player to turn their future around, the 2021 Draft is looking to be extremely important for many teams and players.

1.2. Aims

The goal of this project is to try and predict a new player's career entering the NBA draft, from being a superstar, a starter, bench player or a bust. This figure below will show a player's possible career path in the year 2021 and onwards.

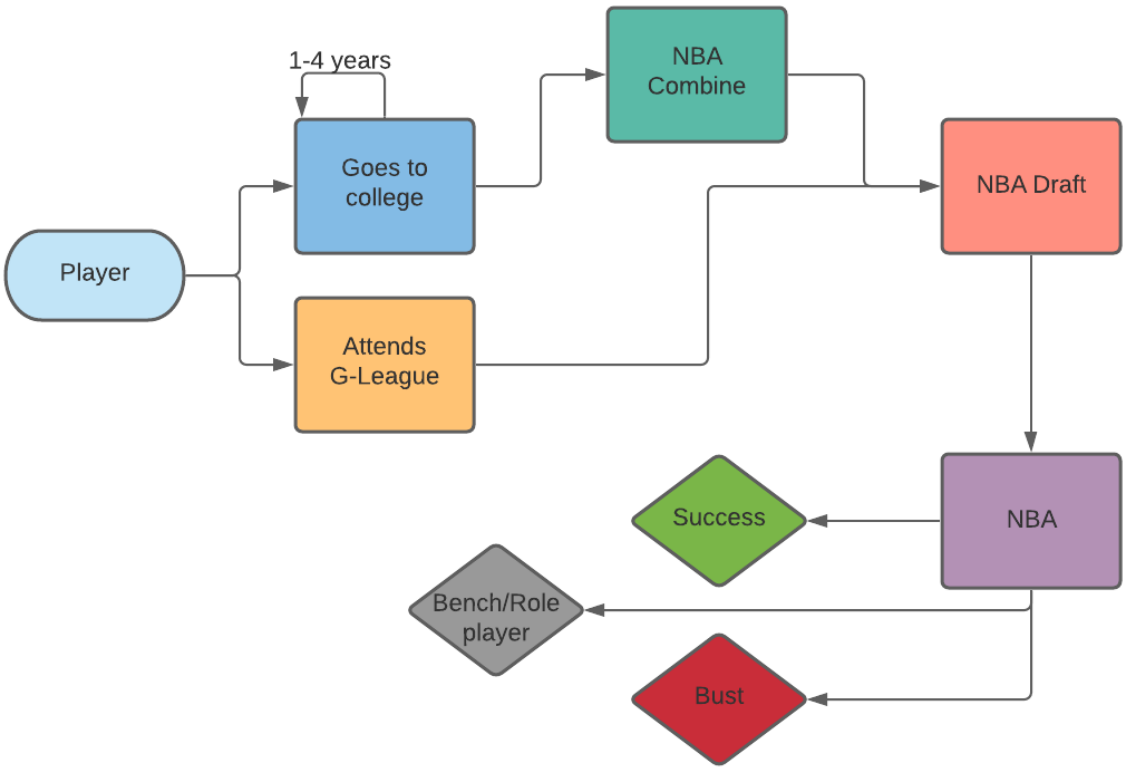


Figure 1: Path on an NBA players career path

For the 2021 NBA draft I hope to see if any generational player be drafted late or earlier than expected? Will a generational talent be picked that was overlooked by other teams? Will teams make the right choice in the draft based on where a team is picking? All these questions are to be answered but with the help of this project, I hope to be able to predict how good a player will be in the future, depending on what team the player was selected to, and what position/number in the draft the player was picked, compared to previous people at their team and position, for example, if the Cleveland Cavaliers get the no.1 pick based over the last 20 years, they have a 50% chance of drafting a bust or a generational talent.

I will create a machine learning program to look at other players drafted at the same number, position within 4 picks and to try predict a players rookie number and will look to develop it further to predict their career paths. This will not be 100% accurate however as there are certain players who have exceeded their expected output. For instance, players such as Michael Jordan and LeBron James were considered the best players in the NBA even at 36 years of age, while most players at 36 years old would be

on the verge of ending their careers. I will also look at which player is considered the best player from all 60 picks and undrafted players.

1.3. Technology

For this project RStudio will be used for both R and Python. RStudio is an open source development environment for programming data languages. Python will be used as the main coding languages for the project.

Visual Studio Code will be used for Python on the averaging algorithm.

1.4. Structure

At the start of the document, I will give a brief introduction, followed by an explanation to what the NBA draft is and how it is used. After this I will explain the reasoning behind choosing this project and why it needs to be addressed.

When this is explained, I will talk about the data I got, using and where I sourced the data from, followed by how I cleaned up the data to be processed and see the results along with code snippets in to help explain to those who are not familiar with how the data works or to get a better understanding of the data.

Following this I will show my analysis of the work I produced, why I chose certain approaches and decisions and why these methods worked for my project. Afterwards, the results of my findings will be shown along with tables, figures and more to help clarify any unclear information. Along with my results, I will also give my conclusions and further development and research that can be made on this project, such as if more can be done in the future to give a more accurate representation of a player's stats and career.

2.0 Data

While researching for the data online I came across the website "Basketball-reference.com" which had details about every single player to play basketball, from what year they were drafted, what pick and round they were selected in, what team they played for and even the stats of every season and game ever played. Looking through the website I found lots of details on these players I could look at and use but found they didn't have their own dataset readily available to the public. That's when I went onto "Kaggle.com" and searched up "NBA draft" to find datasets on the players and to try find data as good as what Basketball reference had.

I discovered two different datasets on the data I needed but some of the data was a missing on one of the datasets and was removed from consideration. The first including all seasons which included the players' names, team played for, age, college attended, year drafted along with which pick and round, games played, points scored, assists, and rebounds, along

with net rating which is a number used to determine how impactful a player is for his team in any given season. [1]

The second dataset looked at but was eventually dropped, was only until 2015 draft but included the win share per player, per year. Win share is defined as "a player statistic which attempts to divvy up credit for team success to the individuals on the team…the important things to note are that it is calculated using player, team and league-wide statistics and the sum of player win shares on a given team will be roughly equal to that team's win total for the season". [2] The higher a players win share is, the more the player contributed to the teams wins.

The reason for these selections was for the machine learning to look at the team a player played for and to try find an estimate of a player's stats for a season after being drafted. This may also be further developed to predict a player's career averages as long as the player is not injured.

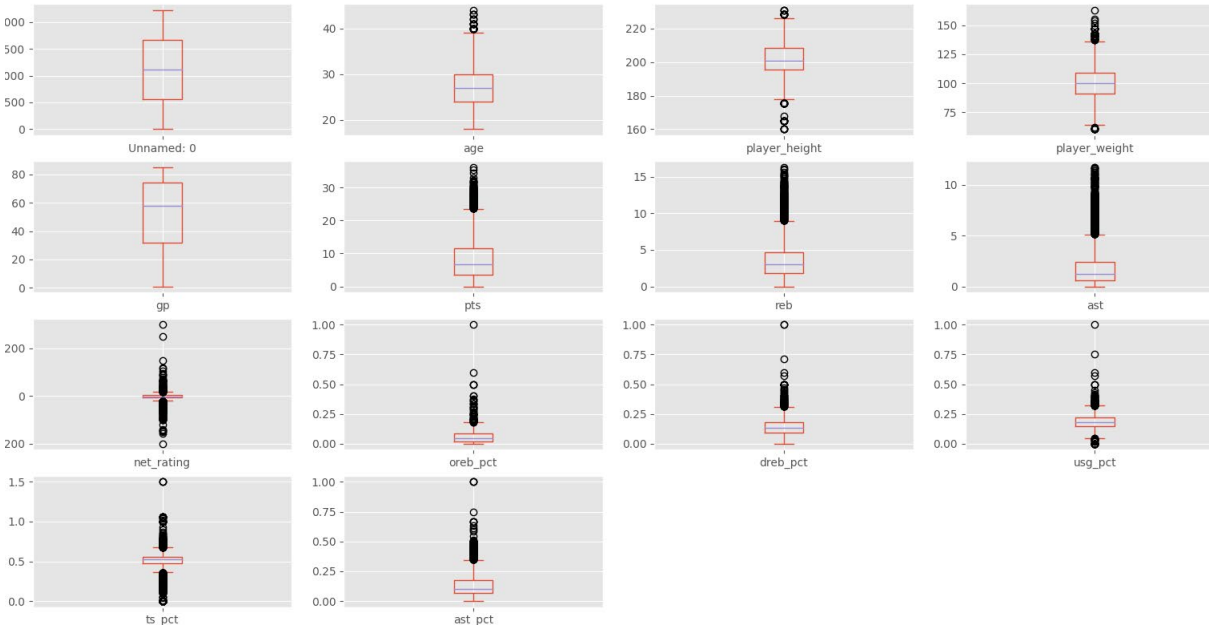


Figure 2: A plot of all the statistics and their expected averages and outliers

3.0 Methodology

The methodology used for this study was the KDD (Knowledge Discovery and Data Mining) approach. The KDD approach is used by researchers when using machine learning or artificial intelligence or even looking at databases and many more domains. The KDD follows a series of process from data selection to interpretation and evaluation. The KDD method allows for a step-by-step approach to turning the data you have into the knowledge desired; this can be seen in figure 2 below.

It is important to have a deep understanding of the applications domain which is being studied along with the goals of the study before selecting and proceeding to use the KDD process.

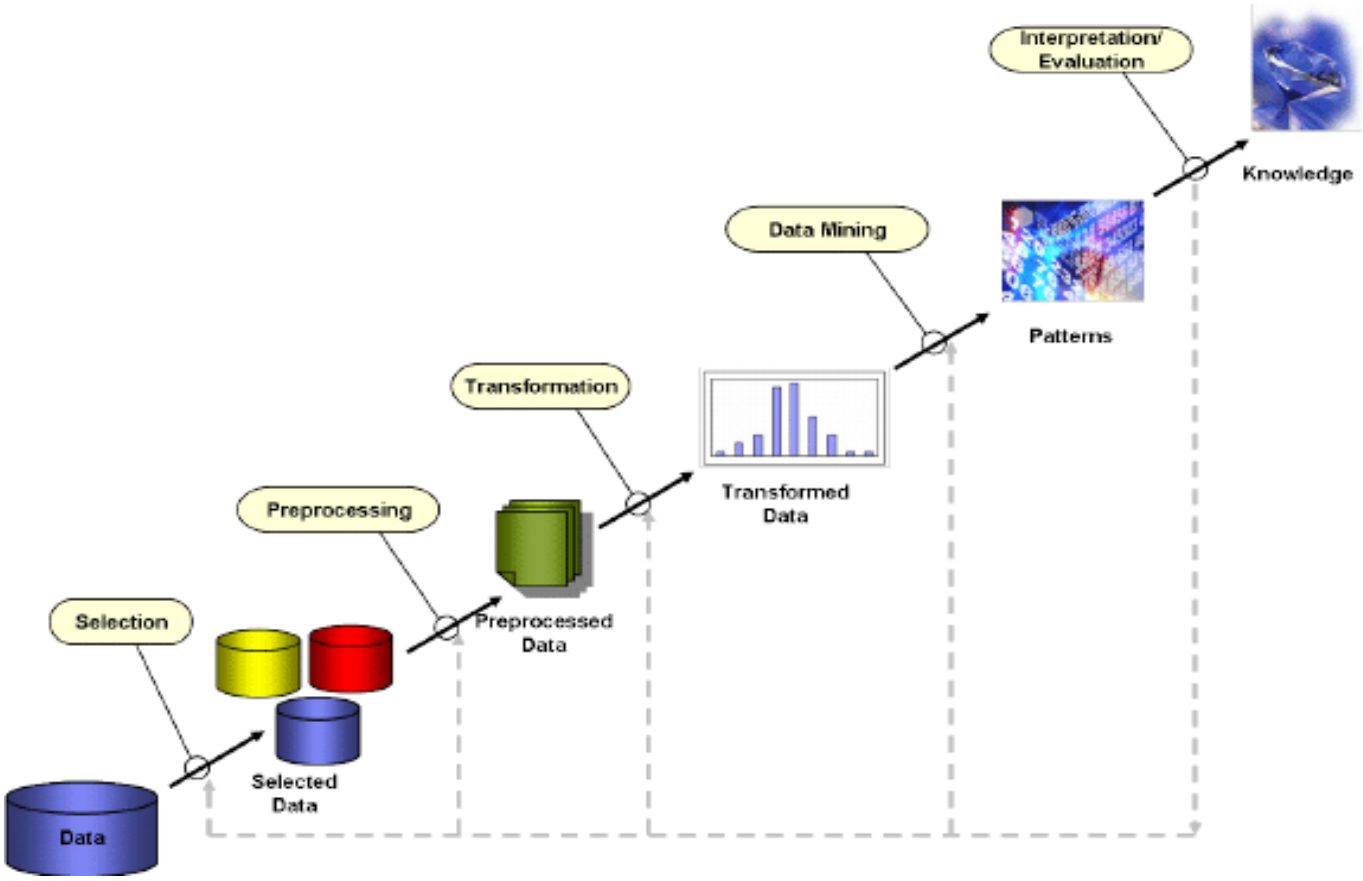


Figure 3: KDD Methodology approach. [3]

3.1 Data Selection

The first part of the KDD process is the data selection, where this step has the KDD target the datasets of the study. For this instance, data was required on the NBA draft and the NBA in order to accomplish the desired objective of this dataset.

The datasets selected for this project was taken from Kaggle, as previously mentioned, and cross referenced with the "Basketball-reference" website, which both contained details of the NBA and the NBA draft. This also allowed for combination of the datasets at certain points, as they contained similar data, but all datasets were not containing the same data. By combining two parts of a dataset for instance, the user would be able to have the player drafted numbers along with the position they played.

3.2 Pre-processing

The pre-processing or data cleaning section of the methodology involves the techniques of sorting the data into a more readable section. These techniques include but are not limited to dealing with outliers, missing data and accounting for noise in the data.

For this project, there was 3 different approaches I could have used to deal with outliers, and then there was the method of cleaning the data which was chosen as well. For the outliers, the options were to drop the record of the player who only played some games every season, drop the record of players under a career number of games or simply remove the outliers entirely. All of these options have their own pros and cons, for option one, there's a chance we drop some the career of a plyer who had a decent career, but because of injuries after 1 or 2 games, didn't play for the rest of the year. This could affect stats later in the data. The second option, the deflated stats will be kept and would affect the final calculations. And the final option is more even on its approach but can be considered the best option. For the approach I decided to discard the players who averaged less than 3 games for their career and for any season.

```
>>> (df.loc[(df['gp']==1 )& (df['net_rating']>20)])
       Unnamed: 0             player_name team_abbreviation   age  \
163           163            Bruce Bowen               MIA  26.0
1008         1008          Tyson Wheeler               DEN  23.0
1687         1687          Mario Bennett               LAC  26.0
1725         1725        Randell Jackson               DAL  24.0
1788         1788          David Wingate               SEA  37.0
2651         2651            Chris Owens               MEM  24.0
4226         4226          Bryon Russell               DEN  35.0
4710         4710          Renaldo Major               GSW  25.0
4733         4733              Von Wafer               LAC  21.0
5738         5738            Steven Hill               OKC  23.0
6015         6015             Ryan Bowen               OKC  34.0
7825         7825        DeAndre Liggins               MIA  26.0
8543         8543            Malcolm Lee               PHI  25.0
9591         9591         Gordon Hayward               BOS  28.0
9675         9675          Anthony Brown               MIN  25.0
9707         9707         Devin Robinson               WAS  23.0
9824         9824        Naz Mitrou-Long               UTA  24.0
10069       10069          Justin Patton               MIN  21.0
10188       10188           John Holland               CLE  30.0
10389       10389         Tahjere McCall               BKN  24.0
10405       10405              RJ Hunter               BOS  25.0
10503       10503           Kobi Simmons               CLE  21.0
10729       10729            J.P. Macura               CLE  24.0
11078       11078  Justin Wright-Foreman               UTA  22.0
11081       11081           Luka Samanic               SAS  20.0
11130       11130         Marques Bolden               CLE  21.0
```

Figure 4: An example of player who had a high net rating but only played 1 game

Following all of this, done on both datasets, allowed for these datasets to be merged into one dataset, which in turn allowed for machine learning algorithms to be used much more effectively.

3.3 Transformation

The transformation process entails the reduction of data along with the projection of the data; depending on the purpose of the research, this may include statistical or visual representation of the data.

The combined data was subjected to Principal Component Analysis (PCA) to determine whether there was a possible link between variables. Exploratory analysis of both the datasets was carried out using data visualization.

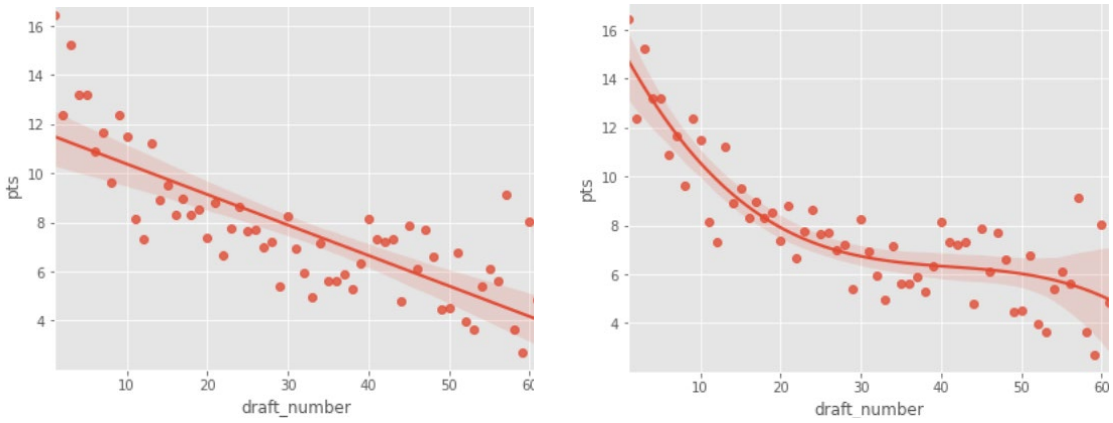
An example of PCA being used was for the players not drafted or over pick 60, they were turned into pick 61 for easier analysis purposes. The dataset also contains all of the player records from the 1996 season, and we can't include all players in the dataset as some players drafted before 1996 would have most of their stats off weight and could largely affect the accuracy of the draft analysis and players accomplishments. This is also unfortunate as we miss out on some great players stats, such as the great Michael Jordan and Shaquille O'Neil. Therefore we would need to discard all players drafted before 1995.

```
>>> df.draft_year.unique()
array(['1986', '1990', '1979', '1995', '1985', '1981', '1994', '1992',
       '1991', 'Undrafted', '1983', '1988', '1996', '1993', '1989',
       '1987', '1982', '1984', '1980', '1976', '1978', '1997', '1998',
       '1999', '2000', '2001', '2002', '2003', '2004', '2005', '2006',
       '2007', '2008', '2009', '2010', '2011', '1963', '2012', '2013',
       '2014', '2015', '2016', '2017', '2018', '2019'], dtype=object)
```

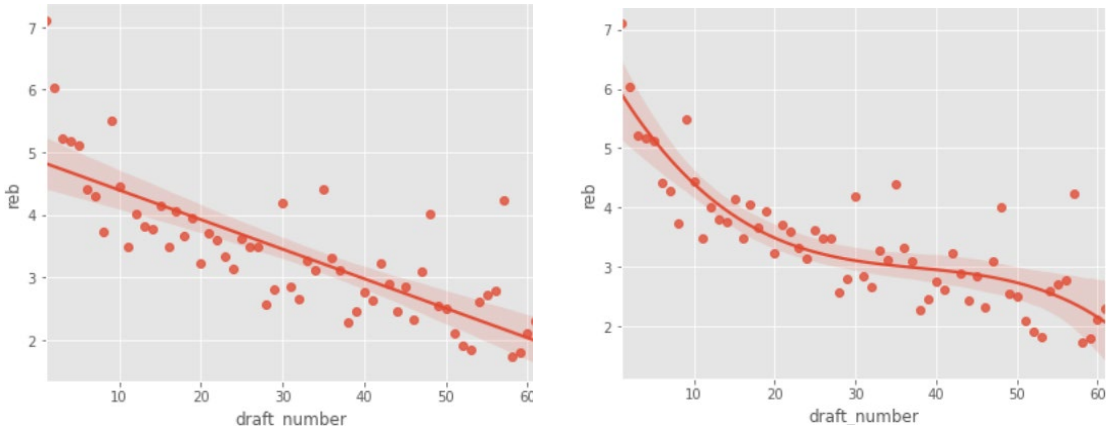
Figure 5: The total draft years before PCA

3.4 Data Mining

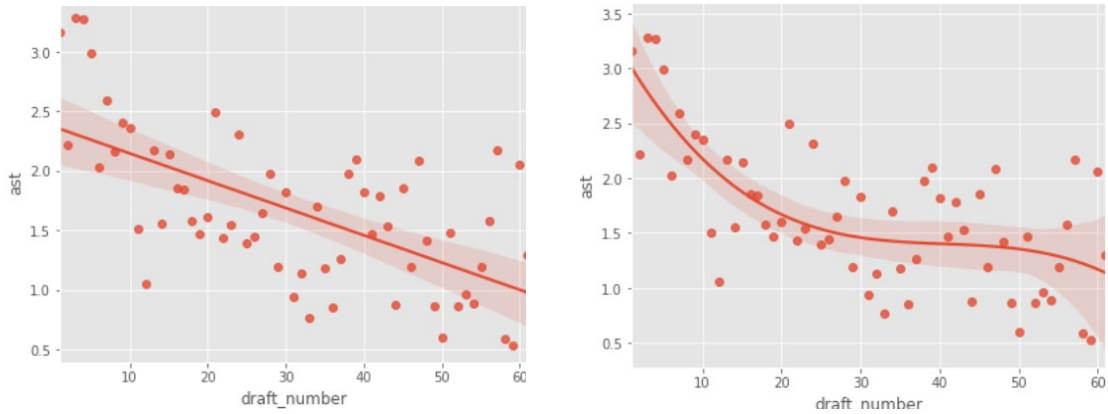
The selection of relevant machine learning algorithms for use on the modified dataset is part of the data mining process. This process also includes selecting parameters from the dataset that would be used in the machine learning model, as well as analysing for and identifying patterns and relationships in the data. To meet the report's objectives regression were applied. The way of using regression in Python is by the NumPy function which allows for many different high performance usages of single and multi-regression. There were 3 different graphs created using regression, these were points, rebounds and assists at number drafted. These results can be seen below.



Figures 6: Points scored per draft pick (count 1 on left, count 3 on right)



Figures 7: Rebounds secured per draft pick (count 1 on left, count 3 on right)



Figures 8: Assists created per draft pick (count 1 on left, count 3 on right)

3.5 Interpretation and Evaluation

This step of the KDD process may involve revisiting previous steps to reinforce, improve and reiterate these steps. This is the Interpretation step, while you then evaluate the overall result and knowledge gained from the information and from the previous steps. The area under the curve (AUC) was computed after visualizing the necessary models and examining the receiver operating characteristics curve (ROC). Using these models, it was used to attempt to predict the 2021 lottery picks for the NBA draft along with ESPN's player list.

Although this is always quite difficult and by the time of the draft, this paper will be published. In the results category I will put in what my prediction for the 1-14 picks and what the data produced recommends in order to see person reference and computer AI results.

4.0 Analysis

There are three different analytical approaches attempted in this study, those being Regression and cohort analysis. Regression as seen in the above examples is used to estimate a relationship between variables. This is used to check for a correlation between variables in order to identify patterns and trends, which is useful for making predictions and forecasts. This was used to look for the correlation between players stats and their position in the draft. As for cohort analysis, this approach is used to break data into small groups rather than looking at the whole dataset as one unit. Examples of this would be, people drafted at the first overall pick or players to play in the 2020 season.

These approaches were chosen as they are easy to use and very simple for what is needed from this study. Along with this, I have experience with dealing with these two analytical approaches. Other options looked at included the Monte Carlo and Factor analysis. Monte Carlo approach considers possible outcomes and calculates how likely they are to happen. The reason this was consider, was for the machine learning system to look at how likely a player drafted in the first 3 picks would be a bad player. This would instead be better used in the further development parts of this study. Factor analysis was the other method looked at but not used in this study but could be used if the study took a different route. Factor analysis condenses large datasets into smaller chunks with underlying construction of observed variables which can correlate to each other and discover hidden patterns, which can then be used to explore concepts which are harder to measure.

For NBA teams, there is 3 important stats that are visible to all, but the hidden numbers are just as important to know just how good a pick or player will be. The draft number is the average of all players drafted at that pick. For instance in figure 8 below, we can see that picks 1 and 3 have both the highest points and net rating, while picks 11 and 12 have some of the lowest statistics overall in the whole figure.

```
avg_per_pick[['draft_number','pts','reb','ast', 'net_rating']][:15]
```

draft_number	pts	reb	ast	net_rating
1	16.445417	7.110833	3.158333	1.318750
2	12.390541	6.038739	2.219820	-1.175225
3	15.210549	5.224051	3.286920	0.002532
4	13.204167	5.173750	3.267917	-0.549583
5	13.210040	5.118474	2.990763	-0.230522
6	10.873054	4.410180	2.026946	-1.995210
7	11.679399	4.293562	2.591416	-1.902575
8	9.618932	3.732039	2.166990	-2.564563
9	12.362150	5.496262	2.401869	-0.274766
10	11.487391	4.445652	2.355217	-0.627391
11	8.160510	3.484076	1.508280	-0.823567
12	7.292045	4.017045	1.053977	-2.628409
13	11.202094	3.805236	2.174869	-1.100524
14	8.889941	3.770414	1.553846	-3.080473
15	9.540230	4.151149	2.140805	-1.290230

Figure 9: The average stats of the first half of the NBA draft.

This can be a correlation to players lower in the draft would have a much lower impact on teams than those at the top of the draft. Of course this isn't always true, one instance of this is a player named "Giannis Antetokounmpo". In Figure 9, the stats show the historic rise of the player, starting in 2013/14 averaging 7 points, 5 rebounds and 2 assists with a massive -5 net rating all the way up to 30 points, 14 rebounds and 5 assists with a huge 16 net rating in the span of 7 years. This has allowed for the player two win back to back MVP's (Most Valuable Player) awards. His rise is impressive but going back and looking at the scouting and analysis of this player when he was drafted, no one expected him to ever get to where he is at now! This shows that if a technology or an A.I. is created that can accurately predict a player's worth, just how important this would be to organisations.

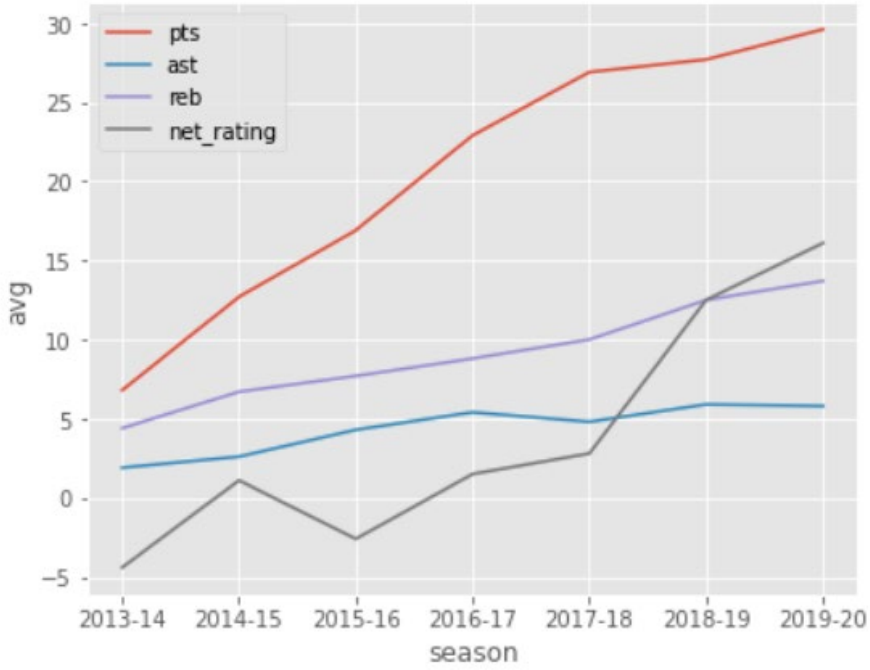


Figure 10: The stats of Giannis Antetokounmpo, drafted at pick 15.

5.0 Results

From the results shown in previous figures, the majority of the time, the 1st overall pick of the NBA has the highest chance of being the best player of the draft. This also applies to the majority of first round picks, however like in most situations there are a few outliers which shine through. An example of this being that in the 2021 season, a second round pick became the MVP of the league for the first time ever. If we are to look at the top of the draft (figure 9) and compare this to figure 11 below, we can see that the top 15 picks are generally much better than the last 10 picks in the draft, bar an instance in 2 selections.

```
avg_per_pick[['draft_number','pts','reb','ast']][-11:]
```

draft_number	pts	reb	ast
51	6.762500	2.096875	1.475000
52	3.981159	1.914493	0.866667
53	3.614286	1.835714	0.964286
54	5.402857	2.608571	0.888571
55	6.134426	2.711475	1.190164
56	5.632653	2.787755	1.579592
57	9.105405	4.227027	2.175676
58	3.651724	1.737931	0.593103
59	2.700000	1.800000	0.533333
60	8.044000	2.112000	2.056000
61	4.872002	2.302008	1.294791

Figure 11: The average stats of the last 10 picks and undrafted players of the NBA draft.

At the 57th and 60th pick there is two outliers who have overachieved their expected outputs for their teams. In both figures 12 & 13 below, their stats will be examined as to why this has occurred. Looking first at fig.12, we see that 2 players have high net rating but the main reason for the high statistical output are both “Manu Ginobili” and “Marcin Gortat”, both players who had long running and successful careers in the NBA. They may have been selected rather late in the draft but are both large outliers to the data available. As for fig.13, we look at the very last pick in the NBA draft and can easily spot the big reason for this outlier. “Isaiah Thomas”, who unlike everyone else at pick 60, averaged double digits in points and a high 4.5 assists. These two examples show that even second round picks are considered valuable for their own reasons. There may be a year or two where no player drafted in the second round goes on to do anything with their career but there is always a chance, they develop into excellent role players or even a league MVP.

player_name	draft_number	gp	pts	reb	ast	net_rating
Corsley Edwards	57	10.000000	2.700000	2.500000	0.300000	-14.600000
Drew Barry	57	20.000000	2.300000	1.100000	1.866667	2.266667
Jordan Bone	57	10.000000	1.200000	0.400000	0.800000	-2.900000
Kevin Hervey	57	7.000000	0.300000	0.300000	0.100000	-13.300000
Manu Ginobili	57	66.062500	13.150000	3.481250	3.781250	10.225000
Marcin Gortat	57	67.166667	9.183333	7.458333	1.058333	0.816667
Ryan Reid	57	5.000000	1.600000	0.400000	0.000000	15.000000
Torraye Braggs	57	11.000000	1.800000	2.150000	0.250000	-5.750000

Figure 12: The average stats of the 57th picks career.

player_name	draft_number	gp	pts	reb	ast
Alex Acker	60	15.000000	2.350000	1.000000	0.650000
Cory Jefferson	60	29.000000	3.250000	2.450000	0.150000
Isaiah Thomas	60	58.333333	16.522222	2.266667	4.566667
Kostas Antetokounmpo	60	2.500000	0.500000	0.400000	0.150000
Robert Sacre	60	47.250000	3.700000	2.850000	0.600000
Semih Erden	60	34.500000	3.750000	2.750000	0.400000
Tyrone Wallace	60	35.333333	5.366667	2.233333	1.333333
Will Blalock	60	14.000000	1.800000	1.100000	1.200000

Figure 13: The average stats of the 60th picks career.

While looking into the lower picks, and comparing them to the best player per pick, I decided to check just who is the best player per draft pick and to then compare them to the

other players who are the best at their draft number in the first round and compare their score which can be seen in figure 14 below.

```
avg_per_player.loc[avg_per_player.groupby('draft_number')['score'].idxmax()][:30]
```

player_name	draft_number	gp	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	score
LeBron James	1	73.882353	27.064706	7.447059	7.441176	7.205882	0.036176	0.182000	0.312882	0.587471	0.352941	34.393126
Kevin Durant	2	70.750000	26.883333	7.075000	4.125000	5.975000	0.022417	0.179750	0.296167	0.614417	0.192417	33.399405
James Harden	3	74.909091	25.154545	5.300000	6.263636	6.136364	0.025091	0.138818	0.297000	0.608545	0.306909	29.916734
Russell Westbrook	4	72.666667	23.408333	7.083333	8.225000	5.400000	0.052583	0.160417	0.320750	0.528167	0.398500	33.240575
Kevin Garnett	5	69.100000	17.350000	9.830000	3.630000	5.565000	0.071100	0.263900	0.240600	0.540250	0.182000	20.345175
Damian Lillard	6	75.750000	24.312500	4.187500	6.537500	3.125000	0.017500	0.101750	0.280875	0.577500	0.294500	38.488002
Stephen Curry	7	63.545455	22.854545	4.536364	6.481818	7.036364	0.022818	0.118455	0.272727	0.615545	0.298727	24.718230
Andre Miller	8	76.705882	12.047059	3.576471	6.347059	1.441176	0.039882	0.099412	0.201118	0.535824	0.337412	25.586498
Dirk Nowitzki	9	72.476190	20.109524	7.347619	2.338095	4.947619	0.031524	0.213905	0.260238	0.568000	0.118048	25.139365
Paul Pierce	10	70.684211	18.931579	5.436842	3.357895	3.689474	0.026895	0.160158	0.256421	0.565632	0.170842	22.782404
Klay Thompson	11	76.875000	19.387500	3.437500	2.300000	8.387500	0.015250	0.091875	0.247625	0.572875	0.108750	43.438639
Steven Adams	12	74.857143	9.871429	7.700000	1.228571	5.042857	0.131571	0.164714	0.147714	0.589429	0.064143	31.313844
Donovan Mitchell	13	72.666667	22.900000	4.066667	4.033333	5.633333	0.021000	0.096333	0.299333	0.546667	0.203000	33.501333
Bam Adebayo	14	71.333333	10.600000	7.766667	2.900000	1.466667	0.082000	0.213000	0.173000	0.598333	0.162667	25.633939
Giannis Antetokounmpo	15	74.571429	20.500000	9.114286	4.385714	3.857143	0.056143	0.226429	0.262143	0.583571	0.225429	40.315085
Nikola Vucevic	16	65.333333	15.755556	10.000000	2.533333	-2.255556	0.094111	0.263667	0.239778	0.528222	0.138667	24.189636
Jrue Holiday	17	64.181818	15.809091	3.900000	6.500000	1.363636	0.027636	0.097545	0.233364	0.525909	0.303091	22.440079
David West	18	68.933333	13.266667	6.320000	2.133333	1.720000	0.073800	0.191333	0.221067	0.542533	0.141000	23.609050
Jeff Teague	19	69.909091	12.381818	2.436364	5.790909	1.072727	0.013182	0.085636	0.215273	0.539727	0.314273	21.939581
Zydrunas Ilgauskas	20	64.846154	12.938462	7.338462	1.092308	2.107692	0.124538	0.191000	0.229923	0.531308	0.069385	26.693391
Michael Finley	21	72.928571	15.592857	4.285714	2.857143	2.850000	0.029071	0.111857	0.212929	0.525643	0.134929	26.397399
Mason Plumlee	22	74.571429	8.271429	6.142857	2.200000	-0.285714	0.097571	0.205857	0.174000	0.595571	0.149143	29.893189
Tayshaun Prince	23	72.642857	10.800000	4.150000	2.321429	1.228571	0.038786	0.120000	0.174357	0.509500	0.123643	22.243752
Serge Ibaka	24	72.818182	12.609091	7.354545	0.763636	5.272727	0.087727	0.187727	0.189545	0.570727	0.042091	30.989588
Nicolas Batum	25	65.750000	11.083333	5.125000	3.608333	2.166667	0.039833	0.145500	0.167750	0.554333	0.168333	26.684994
Taj Gibson	26	72.363636	9.418182	6.190909	1.009091	1.872727	0.101727	0.166818	0.173545	0.550364	0.066182	24.280945
Rudy Gobert	27	66.714286	11.200000	10.600000	1.300000	4.742857	0.122286	0.267714	0.155286	0.634000	0.067429	29.078019
Tony Parker	28	69.666667	15.372222	2.661111	5.588889	7.111111	0.012889	0.085167	0.248500	0.541056	0.313444	36.191162
Cory Joseph	29	65.666667	6.322222	2.377778	2.677778	1.044444	0.026778	0.099556	0.157444	0.504889	0.193444	24.006230
Jimmy Butler	30	64.000000	16.722222	4.988889	3.666667	5.344444	0.051778	0.112667	0.211444	0.566222	0.170333	31.793972

Figure 14: The best player of each pick in the first round of the NBA draft.

If you are to just look at the score, the best player in the league is “Klay Thompson” at the 11th overall pick while Russel Westbrook is the worst of the best at the 4th overall pick. This shows that the score system isn’t the greatest or the best as while Klay has won NBA championships, he was never considered a top 10 player in the NBA throughout his career, while Russel Westbrook has won an MVP award and has broken several records. Another interesting note I discovered in this score and in my results, a lot of recognisable players are on this list while being also the best at their draft number of the last 25 years. Players like “LeBron James”, “Stephen Curry” and “Kevin Durant”. I even tested this on people who know nothing about basketball and these 3 players were recognised on name basis only. One name many reading fig.14 may be asking “Where is Kobe Bryant?” and to my surprise as well, his name is not on the list. The reason for this would be his score would have been lower than that of the player ahead of him. I decided to have a deep dive look into the number where Kobe was drafted at and found that he was second behind “Donovan Mitchell”. The reason behind this may come down to two issues with the score system, the first being net score being a high factor in scoring and the data is looking at a player’s entire career up until the data was recorded, meaning if a player suffered injuries or played less games towards the latter half of their career, would affect the player’s final score. The deep dive of the 13th overall pick with Kobe Bryant’s statistics is able to be seen in figure 15.

```
avg_per_player.loc[avg_per_player.draft_number==13].sort_values('score')
```

player_name	draft_number	gp	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	score
Georgios Papagiannis	13	19.500000	3.850000	3.050000	0.700000	-4.950000	0.095000	0.199500	0.169500	0.501000	0.091000	-53.759000
Marcus Haislip	13	22.250000	3.300000	1.450000	0.150000	-3.075000	0.076250	0.116750	0.216750	0.481250	0.029000	-51.970000
Sean May	13	29.750000	6.825000	4.050000	0.950000	-4.550000	0.092000	0.199500	0.224000	0.488000	0.098750	-38.837750
Kendall Marshall	13	40.000000	4.725000	1.425000	4.325000	-6.400000	0.007000	0.075750	0.160000	0.490000	0.324750	-31.832500
Marcus Banks	13	43.500000	5.187500	1.212500	1.800000	-3.950000	0.015750	0.084500	0.194000	0.528500	0.219250	-28.173000
Jerome Robinson	13	43.500000	3.550000	1.550000	0.900000	-1.250000	0.009500	0.113500	0.149500	0.467000	0.099500	-27.876000
Brandon Rush	13	53.444444	6.088889	2.433333	0.933333	-3.855556	0.016778	0.118556	0.146333	0.513444	0.077444	-17.048000
Sebastian Telfair	13	56.400000	7.360000	1.600000	3.380000	-7.490000	0.014300	0.076400	0.199500	0.480100	0.269800	-14.674900
Tyler Herro	13	46.000000	13.100000	4.000000	2.000000	-1.200000	0.011000	0.126000	0.210000	0.535000	0.116000	-12.067000
Julian Wright	13	57.750000	3.925000	2.325000	0.800000	0.875000	0.072000	0.134500	0.157750	0.525750	0.092500	-10.307500
Courtney Alexander	13	62.333333	9.066667	2.200000	1.233333	-5.600000	0.035333	0.082333	0.216667	0.488000	0.100667	-6.808667
Tyler Hansbrough	13	61.142857	6.671429	4.157143	0.471429	-0.271429	0.117143	0.174714	0.184571	0.528000	0.046857	-3.742286
Derek Anderson	13	55.909091	11.227273	3.100000	3.254545	-0.554545	0.028182	0.104455	0.201455	0.528909	0.202818	-2.962818
Keon Clark	13	58.833333	6.633333	5.250000	0.800000	4.600000	0.084833	0.200333	0.159667	0.491833	0.063167	0.151500
Ed Davis	13	66.500000	5.890000	6.360000	0.720000	-3.110000	0.120500	0.237600	0.134700	0.566700	0.057600	0.512100
Zach Lavine	13	58.833333	18.150000	3.733333	3.566667	-7.766667	0.015333	0.115000	0.262500	0.546667	0.199000	0.690167
Thabo Sefolosha	13	62.000000	5.657143	3.671429	1.300000	4.164286	0.037857	0.153929	0.130214	0.537071	0.091500	0.778429
Corey Maggette	13	59.071429	15.500000	4.657143	1.978571	-3.364286	0.044500	0.147286	0.251214	0.566929	0.121071	2.008857
Markieff Morris	13	69.888889	11.333333	5.266667	1.644444	-1.922222	0.053000	0.175000	0.212000	0.525778	0.102000	10.313889
Richard Jefferson	13	69.470588	12.223529	3.923529	2.023529	0.135294	0.027765	0.127176	0.186000	0.561941	0.115706	11.830059
Corliss Williamson	13	69.909091	11.118182	3.881818	1.190909	1.709091	0.071636	0.119818	0.231273	0.535455	0.087455	11.889727
Kelly Olynyk	13	70.000000	9.628571	4.814286	1.857143	2.457143	0.059143	0.182000	0.193571	0.590714	0.134143	12.951714
Devin Booker	13	66.400000	22.660000	3.680000	4.800000	-5.800000	0.014800	0.095400	0.284800	0.565200	0.234000	15.969200
Kobe Bryant	13	67.300000	24.200000	5.200000	4.760000	2.600000	0.034200	0.126200	0.311600	0.543200	0.238700	28.348900
Donovan Mitchell	13	72.666667	22.900000	4.066667	4.033333	5.633333	0.021000	0.096333	0.299333	0.546667	0.203000	33.501333

Figure 15: The number 13th overall picks deep dive.

The most important part about the NBA draft and the lottery system, is getting the 1st overall pick in order to get the best player possible for your team. While looking at the players who were the 1st overall pick, we can see from the score system that roughly half of the picks end up with a positive score as shown in figure 16.

```
avg_per_player.loc[avg_per_player.draft_number==1].sort_values('score')
```

player_name	draft_number	gp	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	score
Anthony Bennett	1	37.750000	3.975000	2.850000	0.400000	-11.425000	0.076750	0.228750	0.186500	0.465500	0.051250	-55.560771
Markelle Fultz	1	31.666667	9.066667	3.400000	4.033333	1.000000	0.037333	0.098333	0.203667	0.463333	0.258667	-39.891521
Greg Oden	1	35.000000	7.633333	5.933333	0.466667	3.000000	0.147333	0.232000	0.184667	0.604333	0.037000	-36.880854
Kwame Brown	1	50.583333	6.166667	5.316667	0.875000	-3.941667	0.094250	0.208000	0.155083	0.511833	0.066000	-30.084354
Zion Williamson	1	18.000000	23.600000	6.800000	2.200000	9.600000	0.096000	0.113000	0.285000	0.623000	0.116000	-28.686521
Michael Olowokandi	1	55.555556	7.744444	6.444444	0.655556	-4.911111	0.082667	0.210000	0.182556	0.453556	0.042444	-23.659410
Andrea Bargnani	1	55.000000	14.270000	4.530000	1.230000	-5.720000	0.039200	0.149200	0.243100	0.524200	0.073800	-19.780021
Kenyon Martin	1	50.466667	10.726667	6.513333	1.606667	2.906667	0.069667	0.187800	0.183867	0.505533	0.090067	-16.862588
Andrew Bogut	1	50.428571	8.600000	8.192857	2.035714	3.071429	0.099500	0.254786	0.158071	0.548571	0.120786	-16.609235
Deandre Ayton	1	50.500000	17.650000	11.150000	1.850000	-2.950000	0.112500	0.234500	0.223500	0.589500	0.093500	-10.666021
Joe Smith	1	63.200000	9.773333	5.873333	0.933333	-0.433333	0.099000	0.185800	0.191867	0.494200	0.064867	-9.737121
Derrick Rose	1	54.181818	17.900000	3.209091	5.236364	-0.027273	0.027727	0.079727	0.277818	0.517909	0.281455	-8.434884
Elton Brand	1	62.235294	14.841176	8.111765	1.935294	-2.558824	0.100529	0.191059	0.215235	0.531706	0.108353	-4.407933
Yao Ming	1	60.750000	18.562500	8.912500	1.550000	2.012500	0.092625	0.227875	0.268000	0.593375	0.092250	2.942104
Kyrie Irving	1	58.666667	22.655556	3.822222	5.722222	1.433333	0.025444	0.096889	0.290667	0.572444	0.295556	3.461479
John Wall	1	63.666667	19.022222	4.244444	9.100000	-0.433333	0.017444	0.112889	0.272333	0.517444	0.408778	6.809368
Andrew Wiggins	1	75.500000	19.800000	4.416667	2.433333	-2.333333	0.037333	0.093500	0.253333	0.521167	0.111167	10.713646
Blake Griffin	1	62.200000	21.160000	8.450000	4.420000	4.950000	0.065600	0.204300	0.279200	0.552600	0.219400	12.381579
Allen Iverson	1	65.285714	26.064286	3.692857	6.000000	0.600000	0.022214	0.080571	0.311000	0.517000	0.274643	12.728765
Dwight Howard	1	69.062500	16.500000	12.087500	1.356250	3.650000	0.119125	0.287750	0.220750	0.606375	0.069250	13.839979
Anthony Davis	1	64.875000	23.950000	10.387500	2.275000	1.262500	0.080375	0.237625	0.278875	0.586625	0.112125	13.926104
Ben Simmons	1	71.333333	16.466667	8.233333	8.033333	3.733333	0.056333	0.169333	0.213000	0.582333	0.347667	19.049146
Karl-Anthony Towns	1	71.600000	23.120000	11.660000	2.980000	0.500000	0.094600	0.263400	0.260800	0.623600	0.143800	21.126679
Tim Duncan	1	73.263158	18.889474	10.768421	3.010526	9.747368	0.097105	0.262632	0.267316	0.548421	0.158895	26.893795
LeBron James	1	73.882353	27.064706	7.447059	7.441176	7.205882	0.036176	0.182000	0.312882	0.587471	0.352941	34.393126

Figure 16: The number 1 overall pick deep dive

Looking at this we can see that LeBron is the best player followed by Tim Duncan and Karl-Anthony Towns, with Karl and Tim roughly having the same statistics, however their net score is vastly different which causes a huge divide. I decided to look at the last 40 drafts and everyone who was drafted 1st overall and to see if they were the best player in their respective draft year. I took an image from online and decided to see who is the best player and the odds of them being the best player in the draft. In figure 17, we can see some recognisable players such as LeBron James (2003) and Shaquille O'Neil (1992). Looking at the figure, there is the biggest gap between 2005 and 2011 with none of the players becoming the best player in their respective draft class, however there are 4 players who were drafted first that have had very good careers.

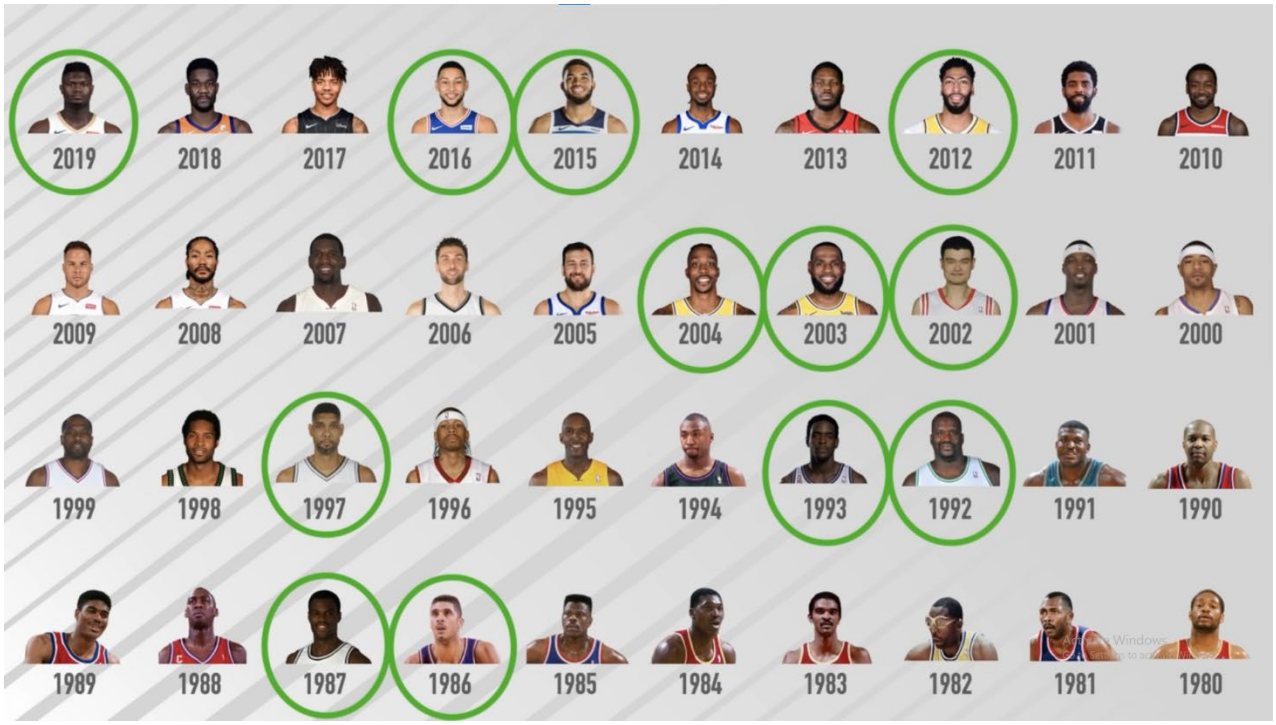


Figure 17: Last 40 first round picks 2019-1989, green circle means best. [4]

Next area looking at for the 1st overall pick is their odds of becoming what. For example, looking at figure 18, there is a 20% chance the 1st overall selection is a bust (very bad player), 20% of becoming an MVP, 30% chance of being the best player in the draft and a 75% chance of being an all-star (a mid-season break that the best player gets invited to). This would then mean there a team drafting the first overall pick has a 20% chance of drafting an MVP or a 30% of drafting the best player in the draft. An important question to ask is "If you have the 1st overall pick, would you be able to draft the best player?"

Figure 18: The number 1 overall pick percentages based off previous players [4]

NBA Draft lottery statistic predictions

Pick number	Points	Assists	Rebounds	Net Rating	Score max	Score min
1	12.93	6.2	8.7	1.26	34.393126	-55.560771
2	13.12	5.2	4.4	1.09	33.399405	-39.868295
3	12.98	0.91	4.66	1.28	29.916734	-42.669266
4	9.47	0.85	1.97	-1.11	33.240575	-36.649720
5	6.45	2.49	1.11	-2.76	20.345175	-51.728675
6	9.28	1.36	3.9	-1.67	38.488002	-45.363623
7	9.68	1.49	3.76	-3.3	24.718230	-22.636088
8	10.58	2.06	3.58	2	25.586498	-50.876296
9	9.42	2.1	3.12	-0.3	25.139365	-63.395647
10	5.97	1.42	2.3	-30	22.782404	-68.443543
11	4	1.04	1.5	-29.96	43.438639	-37.543836
12	3.12	1.44	0.91	-41.38	31.313844	-72.162727
13	2.08	.9	.98	-41.98	33.501333	-53.759000
14	2.12	.076	1.3	-43.52	25.633939	-43.774061

6.0 Conclusions

The initial goal of the study was to predict the statistics of a player being selected in the first round of the draft

We looked at the best NBA players based on their overall draft position and analysed why the outcomes were as they were. We make quite a lot of assumptions in our analysis; therefore the results aren't the most objective. We also ranked the players simply using a single metric which can be changed and updated a whole lot more for a much better understanding in the future. This is promising to know that from small research and limited data we can determine the best players, it can also be used to look at the worst players and if more data becomes available of players drafted fully before 1996 is available, it may give a clearer and more deep understanding of the best players drafted. Another thing to note is that the score system is a little off with certain picks having a much higher max score because of games played, net score and more. This can be improved in the future. In reality, where a player is drafted, is not solely going to decide their future or performance in the NBA. Other factors such as injuries, teams need and more can affect a player’s draft needs and expected outcomes.

A dataset can always be greatly improved. This project's dataset can be enhanced by acquiring NBA statistics prior to 1996 and employing text mining techniques to increase the model's overall prediction performance.

7.0 Further Development or Research

In the future, I'd want to expand and to utilize a more advanced machine learning approach to identify players “tiers”, which has become a polarizing topic among NBA fans. Along with this if more time was available and more resources, I would have loved to be able to look at players coming out of college and their first year in the NBA and compare the difference between the two, to see if there is an increase/decrease in efficiency, scoring, and more intangibles which creates a player to go as a higher selection. If I was to pursue a masters, I believe this would be the kind of work that would be done to excel this type of project.

The averaging system created as well was okay but there were problems towards the start and the end, especially with certain picks. If there was a better way of doing this, I would like to develop it and get an accurate measure.

8.0 References

[1] Cirtautas, J., 2020. *NBA Players*. [online] Kaggle.com. Available at: <https://www.kaggle.com/justinas/nba-players-data> [Accessed 30 July 2021].

[2] *NBA Win Shares | Basketball-Reference.com*. [online] Basketball-Reference.com. Available at: <https://www.basketball-reference.com/about/ws.html> [Accessed 30 July 2021].

[3] Guerra-Hernández, A., 2008. *Explorations of the BDI Multi-agent support for the Knowledge Discovery in Databases Process*. [online] Research gate. Available at: <https://www.researchgate.net/publication/236373188_Explorations_of_the_BDI_Multi-agent_support_for_the_Knowledge_Discovery_in_Databases_Process> [Accessed 30 June 2021].

[4] Youtube.com. 2021. *Can You Solve the Hardest Problem in NBA History?* [online] Available at: <https://www.youtube.com/watch?v=mRmMGeyruB0> [Accessed 20 July 2021].

Appendices

Project Proposal

9.1 Objectives

The main objective of this project is to try and understand which players are the best the NBA draft, how a player’s stats are or their expected outcome and if the player can develop or turn into a superstar, a starter, bench player or a bust.

The secondary objective of this project is to see how important a draft pick is. Who is the bets player per draft pick, and which is the best pick. Since most teams want the number 1 overall pick the most, seeing who are the best and worst 1st overall pick and to see if it is better to have the 1st overall pick or not.

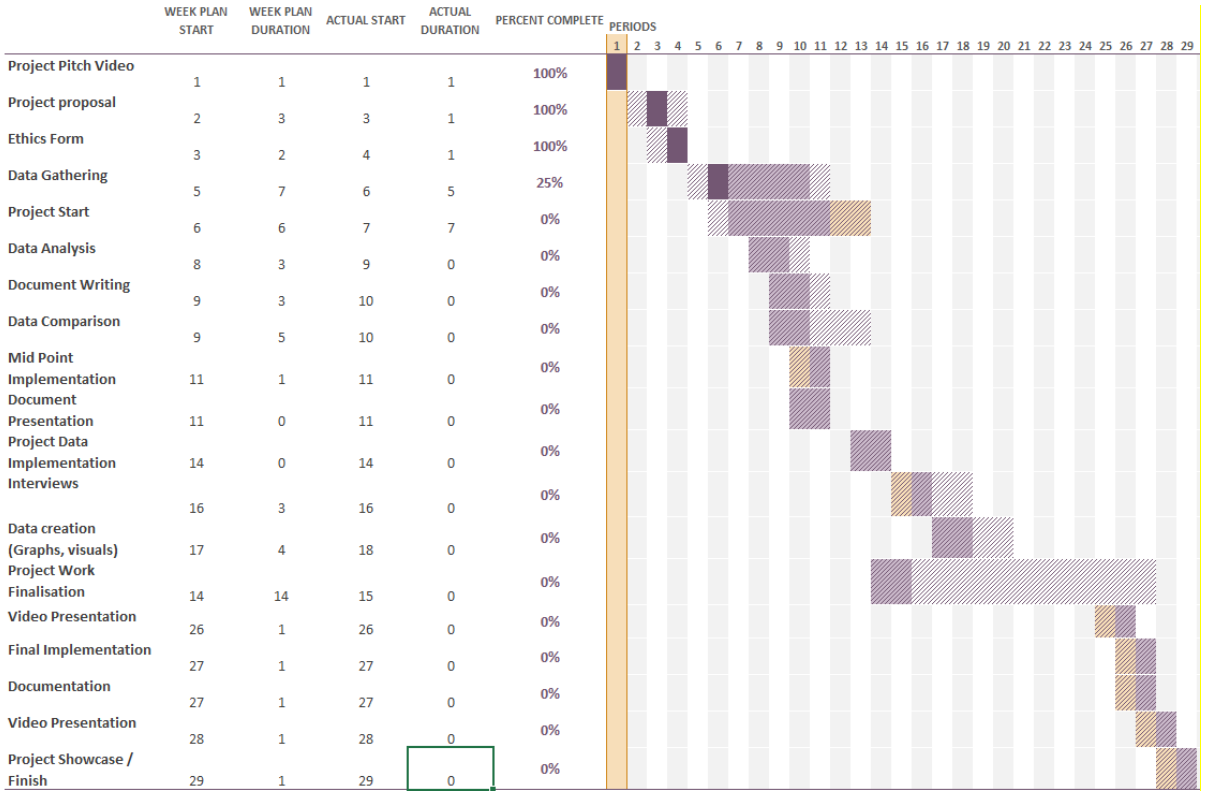
9.2 Background

Since 2020 when the last dance documentary was released, I feel in love with Basketball and have been interested in the small aspects of the NBA such as the draft. It was only in September last year that I stayed up all night to watch the NBA draft 2020 and watch some of the best young players get drafted. It was at this time I became infatuated with the idea of the NBA draft, the lottery and the NBA itself.

It wasn't until this year that I finally decided on what team to support and what it was like staying up and watching a game with the fans. Its safe to say I'm addicted to watching basketball and when my original idea was revoked, this was the next best thing I could do, as it's something I love.

9.3 Data Approach

Data will be targeted, selected, transformed and modelled to reveal knowledgeable patterns which will aid in the prediction and previous draft results will be analysed following the Knowledge Discovery and Data Mining methodology. Techniques such as data visualization and principal component analysis were used to explore the datasets.



9.4 Technical Details

RStudio, Python 3.9.2, excel, csv files and Visual studio code .

The computer specifications I am performing my project on are:

- AMD Ryzen 7 2700X Eight-Core Processor
- 16GB DDR4-2400MHz
- 1TB Solid State Drive

9.5 Evaluation

I would evaluate all the sources to see if the data collected is valid or not and can be useable. I will also be evaluating the data I collected in order to create my own data and use my own techniques that will be unique to the way the data is collected and presented. The data will be turned into a graph or some version of visual data to help give evidence of how the data is shown.

Reflective Journal October

At the start of the project, I was working on having a backup or two ideas', in case my original project was rejected I would have a fall-back option of my own choice and also be able to have something I was interested in. My original idea was on "power creep in gaming", where I would go through its origins, games affected, games that have dealt with it well and how to combat power creep for the future. But seeing as how this would be a nice for my data analytics market, I decided on having my back up ready as "NBA draft how to get it right". Considering the NBA is the biggest organised sport in America, and many different types of young prospects available, the data available and the possibility for a big organisation to use this data matric would help massively and could save or create millions of dollars for any given team.

I started to work on my original submitted idea first, looking at all the material available, the types of power creep and what games have been affected by speaking to general communities of these games, pro players in both the amateur and pro leagues, and attempted to get contact to the developers themselves. I have prepared some work and organised a meeting with my supervisor in order to future discuss and move on with the project. Until then, I will not be able to proceed.

Reflective Journal November

My project idea was approved with some questions asked as to how I would acquire the data proceed with the data along with what the data would be focusing on. Since there is no real data about this readily available, most if not all the data would have to be made from scratch. This will increase the difficulty but will help to improve my grade long term, I hope. Until I can source all the data, I will focus on more simple stuff to do, such as looking for data sets and the documentation and finally my other college assignments.

Reflective Journal December

For this month's reflective journal, I started to work on attaining the data for my project in order to start the actual project. First, I started to attain datasets from Kaggle.com of already previously made datasets of what I needed (characters stats) I decided I would start off the comparison for all the characters for league of legends at the last patch of 2020 and first of 2021. This would give the most recent and best available stats for the character for that game. Next, I went to work on attaining the same style of stats for Overwatch a game plagued by power creep and finally started to get the data for hearthstone, magic and Yu-Gi-Oh as they are card games with the same issues.

As it came closer to the middle and the end of the month, I focused on getting the midpoint presentation complete and then onto other assignments that also needed to be completed. As of writing this I still have two assignments taking up my interest before going back to complete my project.

Reflective Journal January

This month's journal focused on trying to work on and cleaning up the data, as well as starting to make my data set that needed to be created by myself, this dataset has a combination of roughly 10 games, with each character in the game, what kind of game they were, power creep probability from buffs and nerfs to each character, how many nerfs and buffs the character has gotten, how many patches have the character been in their own game for. This actually took up a surprisingly lot of time and made me feel a bit more uncomfortable when it comes to the project and time wise.

I will now look for February to start focusing on the data and decide if this project can be completed in the remaining time available considering I will have to also learn R and Python for the project.

Reflective Journal February

As of late February and discussions with my supervisor, I have been advised to procced with my secondary option as we have agreed the work necessary, and the workload would be too much to handle in the end. While this is unfortunate considering there was been roughly 3 months work put into the topic, I have started from scratch with a new topic “The NBA Draft”. I hope this will be the final chance I change my subject.

Reflective Journal March

With the subject change and focus now being put solely on acquiring the data and getting as close to as possible to where I was before changing subject. I feel my level of stress rising each time I sit down at my computer working on the project in such a short time frame left while having more assignments left to do for other subjects. From the 22nd I have started to work on these assignments as to not fall behind compared to others in my course.

Reflective Journal April

I have decided with the help of my supervisor to apply for an extension/ deferral for my final year project in order to get the project done and dusted with time left. This also helps as I have 4 assignments left to finish in both a group setting and as an individual, this now gives me the time and chance to work solely on these projects and then pick up from where I left off on this assignment. I have been working on the data and have currently started to learn Python as well in order to better understand and get better results. Next, I will be cleaning the data when I pick back up.

Reflective Journal May

Not much work was done this month in terms of project work, mostly using this time to recover from covid, stress and other mental health issues I have had during this troubling time. I have acquired all the data, cleaned it and have begun to get graphs and visualise the data, with an emphasis on doing so in Python and RStudio.

Reflective Journal June

This month I started to work on the document, writing it out and adding in the images as to allow for a better path to go in terms of what was needed and expected from the data. I had to change some of the original work to better get the data’s final work into the project’s idea. At the time of writing this, I am currently on section 5 (results) and will be working on finishing this section by the first week in July, having a road map of completion created.

Reflective Journal July

The final month, this month was used to look over the work, finalise the data and with time remaining add more work to better my grade. I created a system that gives a score to every player in the dataset to see who are the best players per draft pick and what their statistics tell. Another system created with the help of a friend was a prediction system that worked great in most instances but not in all, I worked on this system alone for 5 days but couldn't figure out a way to get the best results. Next, I finished the document completely and will be uploading everything on Sunday the 1st of August, a day before deferral submission date. I'm glad to finally have this done and dusted.