Data Science 201 Project Proposal 2021: Team 9

NBA Draft Analysis of ESPN Top 100 High School Recruits

Student A: Sydney Branstetter , Student B: Madison Gannon, and Student C: Tyler Mogensen

I. Introduction

There is a lot of money, decisions, and time that goes into the sport of basketball. Although it is said that there is equal opportunity to get into the draft; after looking at many factors we can see that there are many key things that can increase your chances of being successful. Unlike other sports, there are far fewer positions on the basketball court and each player can make a larger individual impact on the court. Also unlike other sports, there are more subtle differences in the positions, which means that more players can play more positions, again, making it more difficult to be drafted and make a difference for an NBA team. Seeing as there are also the best player-owner agreements throughout sports, many multisport athletes lean toward basketball as there is typically smaller injury risk as well as a large payout if they can make it in the league. Our decision to use data regarding the NBA and what factors determine how you can be successful is important because there is much more information to go more in-depth on.

The inspiration for our project and research comes from the following. We want to dive deeper into the NBA and what makes certain players more successful than others. The analysis will show the key components that we researched that will help younger athletes understand the steps to take in order to be successful in the NBA. Our research will also help managers and coaches understand what it means for a player to truly be helpful for a team. More wins leads to more opportunity to win the Finals and lead to becoming a more successful coach, team, and franchise.

II. Problem definition

The problem that we are trying to solve is the accurate analysis of the NBA draft. We are focusing on the many aspects that make a basketball player more eligible and have higher chances of being successful in the draft. Some problems that we are focusing on are:

- Can we accurately predict whether they will be helpful for a team to win games?
- Can we make a proper model to represent a correlation between a player's skills and their draft?
- Which variables have the highest impact on the draft results of a player?

We are using data collected from players that have already been drafted and had some time in the NBA to determine if other similar players in the current draft will be helpful. Some important factors that will be particularly useful is their college, their draft pick, and their high school rank.

III. Survey on what has been done and what is different in your project.

The current ESPN model for the NBA draft uses College performance, International performance, Scout ranking, AAU/FIBA Juniors performance, Combine measurables which certainly emcompasses many aspects that many would deem important when it comes to basketball performance. However, these metrics do not always translate well, so we are planning on using data from past players with similar rankings, college performances, and other career milestones that we can use to compare and predict players in the current draft to see if they will likely have long productive NBA careers.

IV. Proposed method

Data Collection

Our data collection comes from a csv file called players.csv that we found on data.world (https://data.world/the-pudding/hype/workspace/project-summary?agentid=the-pudding&datasetid=hype). There was a lot of data to start with and the data's original shape was

1873 x 30. We wanted to make it a little more manageable to be able to get better results from the data, so we narrowed it down to: Average plus minus, Average wins added, Average added value rank, Average plus minus rank, The average player impact plus minus value, Average wins added rank, AvgPlusMinusRank, The average league rank of value over replacement player, The average value over replacement player value. We also renamed all of the columns to make the data easier to understand as opposed to the raw data we were given.

Data Analysis

We will be analyzing several different variables from 1998 and 2013. In order to do this, we will be making a model to get a multiple linear regression that will help us visualize the data and make predictions about the stats. We will split the data into test and training data to find the best fit model for the data. We will analyze the models to find correlations between variables that show what makes a basketball player successful in the NBA draft. Our data could have been more helpful if it were to include college statistics that would help us further compare players and the competition that they played against and how far beyond their level that they were playing. This

Data Visualization

To present the data, we will create a correlation matrix heat map to demonstrate the correlation between the different variables. We will also create OLS regression to show the p-values for the different variables and show other values like the R-square to explain the model. Another visualization we will make to explain the data is a pairplot. All of these visualizations will help explain the data in the model.

V. Experiments/ Evaluation

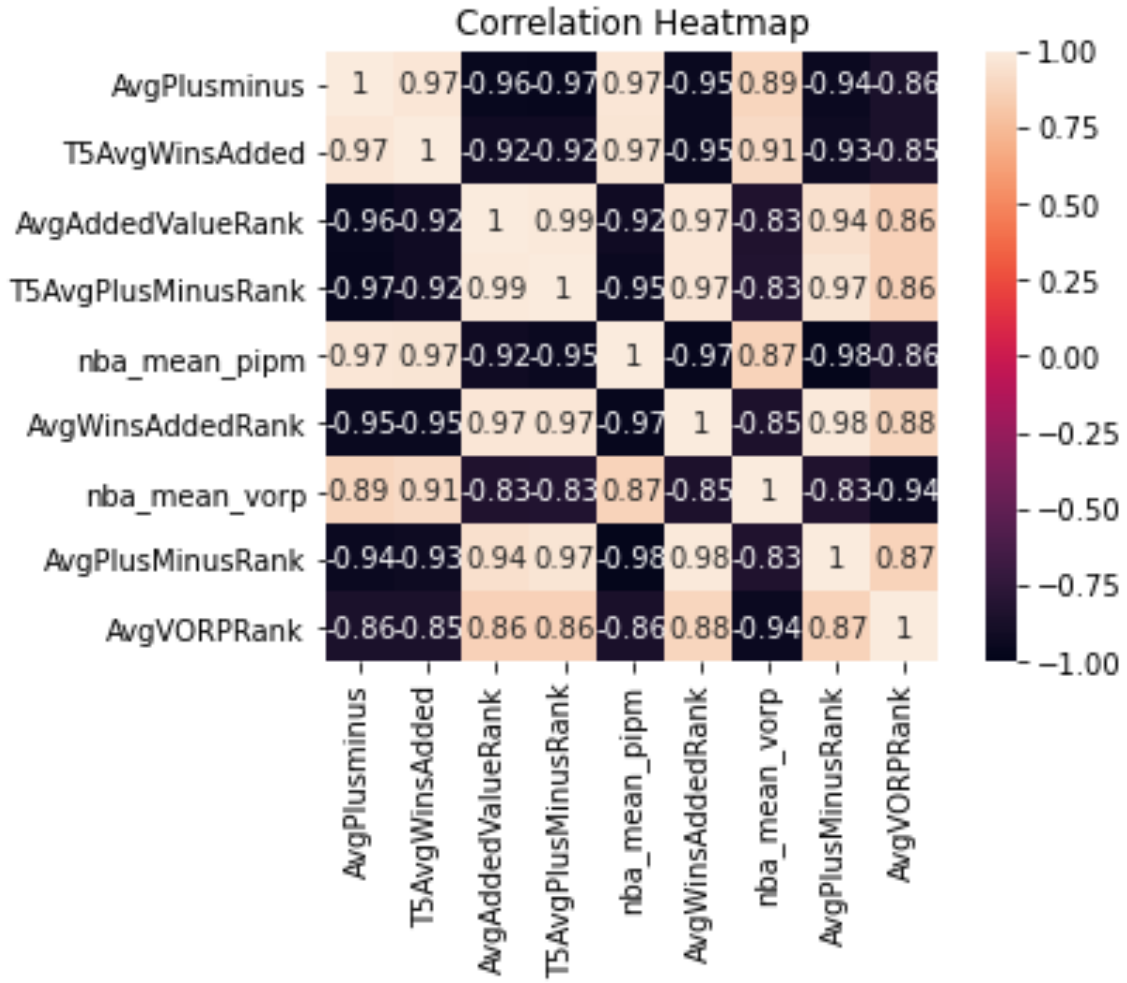
There are many ways to analyze and make predictions about the data. Some techniques include manipulating and cleaning the data to have a simplified model, creating test and

training models, and creating visualizations to understand the data in the model. We used some of these techniques to analyze the variables involved in the NBA draft.

We started out with cleaning the data to make a simple model that was easier to understand and analyze. We took out variables that had null values, little correlation, and ones that did not help predict the data.

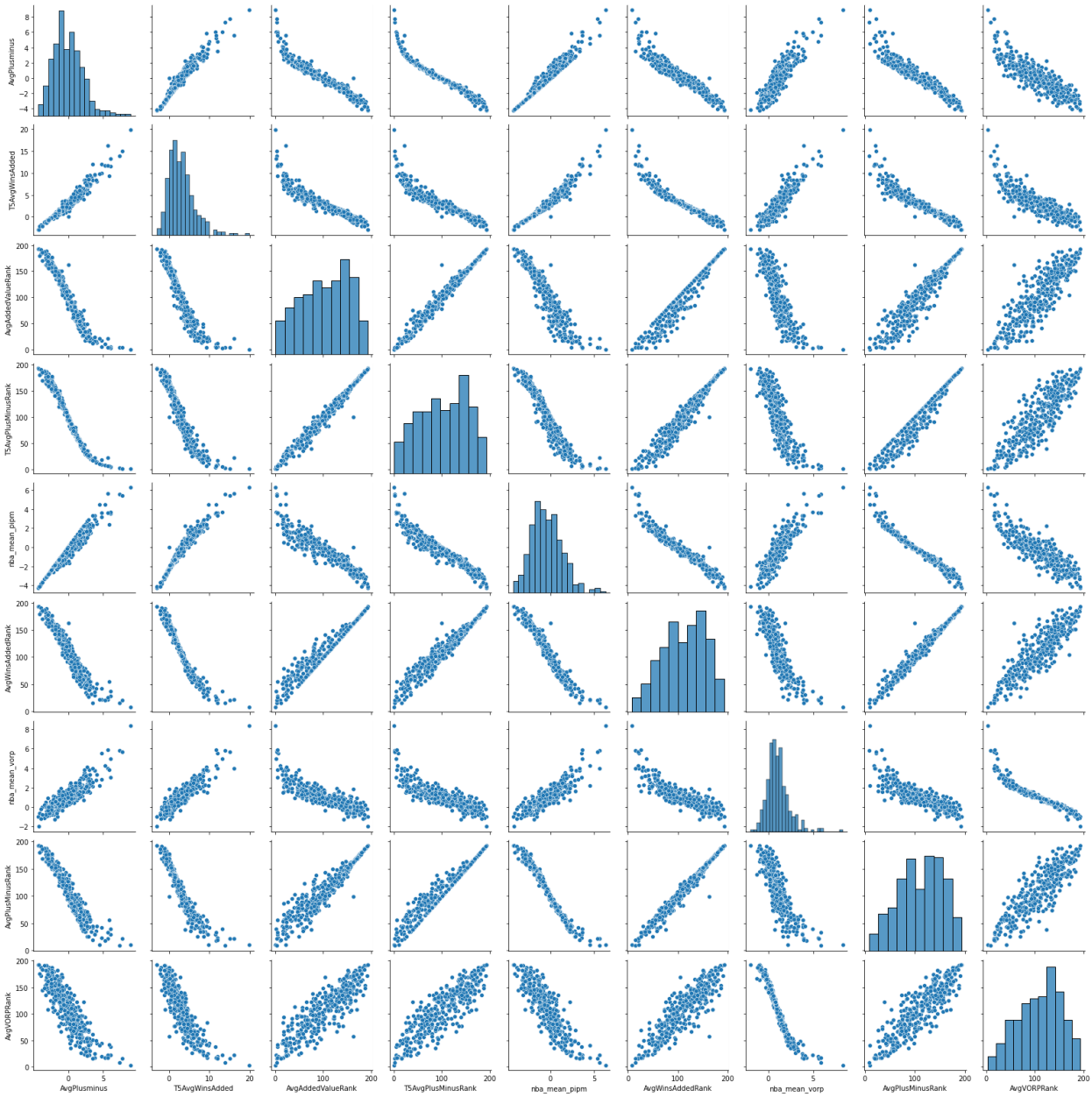
Correlation Heat Map

We created this correlation map to show which variables had the most correlation with each other. The map is a great visualization to explain how most of the variables are highly correlated.



Pair Plot

Below we made a pair plot to show the relationship between each variable in our model. This visualization was a great way to see how each of the variables affect each other. As we can see, many of the variables are linear, and strongly positively or negatively correlated with few outliers.



Regular Regression Model

We used OLS regression after we filtered our data to have p-values only under 0.05. The model shows the 9 variables that have the highest correlation.

OLS Regression Results

Dep. Variable:	AvgWinsAdded	**R-squared:**	0.998
Model:	OLS	**Adj. R-squared:**	0.998
Method:	Least Squares	**F-statistic:**	1.183e+05
Date:	Fri, 10 Dec 2021	**Prob (F-statistic):**	0.00
Time:	09:29:28	**Log-Likelihood:**	1712.7
No. Observations:	1873	**AIC:**	-3405.
Df Residuals:	1863	**BIC:**	-3350.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0003	0.003	0.133	0.894	-0.005	0.005
AvgPlusminus	2.2207	0.027	82.282	0.000	2.168	2.274
T5AvgWinsAdded	1.1875	0.013	93.797	0.000	1.163	1.212
AvgAddedValueRank	-0.0584	0.003	-19.203	0.000	-0.064	-0.052
T5AvgPlusMinusRank	0.0697	0.003	21.528	0.000	0.063	0.076
nba_mean_pipm	-2.5395	0.033	-77.262	0.000	-2.604	-2.475
AvgWinsAddedRank	0.0675	0.003	20.567	0.000	0.061	0.074
nba_mean_vorp	-0.1428	0.022	-6.626	0.000	-0.185	-0.101
AvgPlusMinusRank	-0.0806	0.003	-23.295	0.000	-0.087	-0.074
AvgVORPRank	-0.0032	0.001	-6.017	0.000	-0.004	-0.002

Omnibus:	925.001	**Durbin-Watson:**	2.119
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	224930.403
Skew:	1.132	**Prob(JB):**	0.00
Kurtosis:	56.638	**Cond. No.**	2.00e+03

Training Regression Model

We split the data into test and train data sets using a 80-20 split. After testing and training the data, we were able to get results using OLS regression. This shows us the p-values that were all under 0.05 and our R-squared value for the train data.

Dep. Variable:	AvgWinsAdded	**R-squared:**	0.998
Model:	OLS	**Adj. R-squared:**	0.998
Method:	Least Squares	**F-statistic:**	8.915e+04
Date:	Fri, 10 Dec 2021	**Prob (F-statistic):**	0.00
Time:	09:11:28	**Log-Likelihood:**	1305.1
No. Observations:	1498	**AIC:**	-2590.
Df Residuals:	1488	**BIC:**	-2537.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0005	0.003	0.159	0.874	-0.005	0.006
AvgPlusminus	2.1680	0.031	68.843	0.000	2.106	2.230
T5AvgWinsAdded	1.1957	0.015	80.999	0.000	1.167	1.225
AvgAddedValueRank	-0.0585	0.004	-16.467	0.000	-0.065	-0.051
T5AvgPlusMinusRank	0.0670	0.004	17.298	0.000	0.059	0.075
nba_mean_pipm	-2.5154	0.038	-65.735	0.000	-2.590	-2.440
AvgWinsAddedRank	0.0686	0.004	18.293	0.000	0.061	0.076
nba_mean_vorp	-0.1405	0.025	-5.697	0.000	-0.189	-0.092
AvgPlusMinusRank	-0.0794	0.004	-19.614	0.000	-0.087	-0.071
AvgVORPRank	-0.0033	0.001	-5.419	0.000	-0.005	-0.002

Omnibus:	716.095	**Durbin-Watson:**	1.984
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	142228.529
Skew:	1.085	**Prob(JB):**	0.00
Kurtosis:	50.686	**Cond. No.**	1.88e+03

Categorization

We also decided that it would be ideal to try and find players that would be the most useful when it comes to adding wins to a team. We did this by filtering the data to find those that averaged more wins than most of the other players in the league. If we were to dig deeper into these players to find similarities to those in the upcoming draft, we would be able to make more informed decisions when it comes to who to pick for the best value.

	name	HSrank	draft_year	draft_rd	draft_pk	college
0	Al Harrington	1.0	1998.0	1.0	25.0	NaN
1	Rashard Lewis	2.0	1998.0	2.0	32.0	NaN
3	Dan Gadzuric	4.0	2002.0	2.0	34.0	University of California, Los Angeles
4	Stromile Swift	5.0	2000.0	1.0	2.0	Louisiana State University
7	Joel Przybilla	8.0	2000.0	1.0	9.0	University of Minnesota

If we were to do further analysis with players in the upcoming draft, we could group these players with their college and high school statistics with the upcoming draft class and run some unsupervised learning to create clusters. We could then search for these players and see what other players come up with. We could use that as well as some of the other previous analysis to make informed decisions on where and when to draft certain players. We can do the same efficient players to see where they are usually picked.

VI. Conclusions

When creating our model, we were able to eliminate variables that had a high p-value and end with p-values of 0.00 for all the variables. This meant that we only used variables that would help us predict the success of the NBA draft. We were also able to see how

those final variables correlated with each other and affected each other by using visualizations such as a correlation heat map and a pair plot. There is more than likely some overfitting that occurred in both of the models, so there are likely fewer better predictors that would skew our data less.

In conclusion, we see our experiment and project as a work in progress. Our model has an R-squared value of 0.998 which shows that the model was a very close fit with the data that we were given in our data set. However, there could be concerns when it comes to new data introduced. This is typically fixed when splitting testing and training data, but the data overall is very similar, whereas it could differ given data for upcoming players.

If we were giving more data for the players prior to them playing in the NBA we could have made further discoveries about what makes them valuable and how they compare to their peers. Our experiment was successful in the fact that we were able to see what makes a player valuable for their team in the NBA. We can take with us the fact that if we can find players that are similar to those that were great picks and find a way to pick them consistently, then the team that we are helping will tend to have more wins, better players, and more consistency within their organization.

VII. Distribution of team member effort.

Each team member equally contributed to our project. We worked simultaneously to make sure we all shared our ideas and ended on the same page. We split up some of the work. Tyler worked a lot on the regression models, training and testing data, and the visualizations and the conclusion. Sydney and Maddie worked on the final report and final presentation to explain all of the analysis of the results of the models.