Deep Learning and the NBA Draft: Predicting the Future Value of College Basketball Players

Cade Monroe May
Stanford University
Stanford, CA 94305
cademay@stanford.edu

Abstract

One of the many keys to running a successful NBA franchise is selecting talented players through the annual NBA Draft. A big question regarding prospective draft selections is one of whether a player's skills will translate well to the professional level. This project tackles this question using machine learning. The following is an exploration of building deep neural networks in the interest of predicting the future value of college basketball players. The supervised learning models were trained to predict players' potential future NBA contributions based on their college performances and profiles. The future NBA stats predicted include win shares, PER, and VORP (value over replacement). A wide range of models were trained for both regression and classification tasks.

1 Introduction

1.1 Motivation

Even though the conclusion of the NBA finals ushers in four months without official basketball games being played, the excitement around the league does not cease. Each summer the NBA offseason kicks off with a massive event: the NBA Draft. During this time, teams around the league hope to acquire a player with talents great enough to alter the fortunes of their franchise.

The NBA lottery is an element of the draft through which the order of selection is determined. The lottery is a probabilistic system in which the odds are purposely skewed in favor of teams with poor win-loss records. This design serves the purpose of allowing struggling franchises to rebuild and become competitive.

Preprint. Work in progress.

This year, the NBA lottery is a topic of particular interest because the system has been redesigned for the first time since its institution in 1985. Now, instead of the worst team having a 25% chance at winning the highest pick, the bottom three teams all have an equal, 14% chance.

1.2 Problem Statement

Historically, the top picks of the NBA draft have high likelihoods of blossoming into players with franchise-altering potential. However, fantastic players can be found throughout all 60 picks of the draft. For example, 4x NBA All-Star Jimmy Butler was taken with the 30th pick. Superstar Finals MVP and 3x All-Star Kawhi Leonard was taken with the 15th pick. Reigning NBA Defensive Player of the Year Rudy Gobert was taken 27th, and 3x NBA champion Draymond Green was taken 35th overall. All this to say, even without a top pick in the lottery, NBA teams can still find strong players through the draft. Accordingly, the aim of this project to employ machine learning models to predict the future value of NBA players based on their performances in college.

2 Related Work

There is a great deal of work being done in this domain. For example, in "Using machine learning to predict the long-term value of NBA draftees," Jesse Fischer reports on constructing a range of models in the interest of player evaluation. He specifies that many of his best models included per 40 college advanced stats and prospect age as features. Additionally, he also noted that including a player's actual draft pick number yielded substantial improvements in performance.(1) However, this project avoids using actual draft position as a feature because that would only be useful in less realistic scenarios.

3 Data

3.1 Data Collection

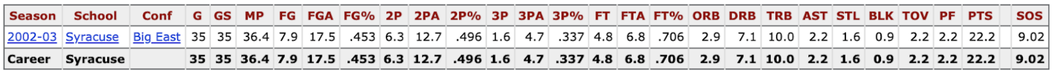
This data used for this project involves NBA and NCAA basketball statistics logged between 1991 and 2019. The statistics revolve around the nearly 1500 NBA players who, during that time period, entered the NBA after participating in college basketball.

The first step in the data-collection process was retrieving the names and advanced stats of all of the people who played in the NBA between the years 1991 and 2019. This data is readily available on basketball-reference.com.

After collecting and saving these annual NBA spreadsheets, all of the NBA players' names are compiled into a single list. What follows is the second step of the process, which involves iterating over this list, and for each name, attempting to identify the player's college stats, if available. This was done by sending a web scraper to each players college basketball profile on sports-reference.com. The web-scraper was creating using Python's BeautifulSoup library.

For each player who played in college, in addition to height, weight, basketball position, school, and conference, all of their per-game stats were extracted. An example of these stats can be found in Figure 2, shown below.

Figure 1: Sample Data from sports-reference.com/cbb



Season	School	Conf	G	GS	MP	FG	FGA	FG%	2P	2PA	2P%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS		SOS
2002-03	Syracuse	Big East	35	35	36.4	7.9	17.5	.453	6.3	12.7	.496	1.6	4.7	.337	4.8	6.8	.706	2.9	7.1	10.0	2.2	1.6	0.9	2.2	2.2	22.2		9.02
Career	**Syracuse**		**35**	**35**	**36.4**	**7.9**	**17.5**	**.453**	**6.3**	**12.7**	**.496**	**1.6**	**4.7**	**.337**	**4.8**	**6.8**	**.706**	**2.9**	**7.1**	**10.0**	**2.2**	**1.6**	**0.9**	**2.2**	**2.2**	**22.2**		**9.02**

3.2 Advanced Stats

Again, the goal of this project is to use the college-level stats described above to predict a player's future value to an NBA team. This is where the advanced NBA stats become important. In the NBA, there are many all-in-one stats that are used in order to attempt to capture and measure players' values. The machine learning models in this project were trained to predict some of these values.

The future NBA performance indicators that the models attempt to predict include win shares (WS), Player Efficiency Rating (PER), and value over replacement player (VORP). These statistics are commonly-used in basketball in order to estimate a player's true value to his or her team. The three mentioned above are from a class often referred to as "advanced statistics," as the calculation involved goes beyond that of typical box statistics such as points per game.

Wins shares, PER, and VORP are three different all-in-one stats that attempt to represent a player's value. Win shares attempt to represent the number of team-wins a player provided over the span of a season. PER represents a player's per-minute contributions, and is adjusted for league pace such that the average player's PER value is always 15.0. The idea behind the formula for PER is that it is increased by positive contributions (points, assists, etc.), and decreased by negative contributions (missed shots, turnovers, etc.). Finally, VORP "is an estimate of each player's overall contribution to the team, measured vs. what a theoretical "replacement player" would provide, where the "replacement player" is defined as a player on minimum salary or not a normal member of a team's rotation."

The ability to reliably predict these values can be incredibly valuable to NBA teams' drafting and recruiting processes. Machine learning can be used predict the future value and performance of college players based on a combination of attributes that may not be immediately obvious to the human eye. Accordingly, these machine learning predictions can be useful in ranking players.

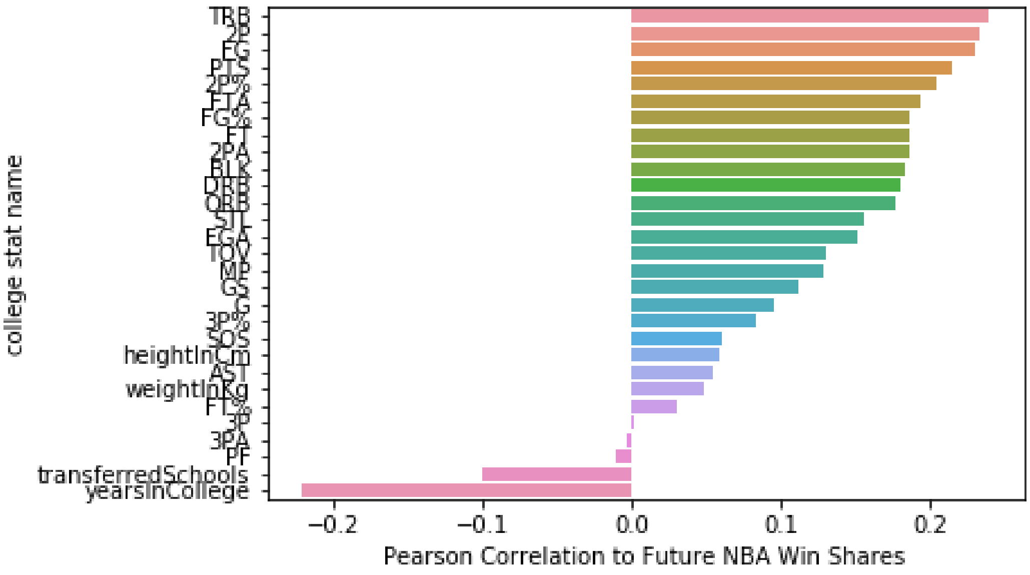
3.3 Correlation

A deeper look into the data at hand can provide a better understanding of the relationship between players' college performances and performance in the NBA. Consider Figure 1 below, which shows the Pearson correlation coefficients variety of college stats plotted against future NBA win shares. Pearson correlation is a formula that estimates the extent to which two variables may be linearly related. In this light, many of the top correlations shown in Figure 1 are understandable. Rebounds are an important part of basketball on both ends of the floor, and it is reasonable to predict that the most prolific rebounding college players are likely strong enough and savvy enough to contribute rebounds at the next level. High field goal counts and points are self-explanatory.

Additionally, on the other end of the chart, it is interesting to note that playing more years in college has a negative correlation with future NBA win shares. This resonates with the common stereotype that prospects who come out of college later have less room for improvement.

However, making selections based on these correlations alone is insufficient. Clearly, the optimal strategy is not to simply pick the player with the most rebounds and points. There are many complex relationships at play when it comes to a player's skills translating well into the NBA level. This summons an exploration of the transition to the NBA through deep learning applications.

Figure 2: Pearson Correlation Coefficients



4 Methods

4.1 Supervised Learning

The data described above was preprocessed for the purposes of being fed to supervised learning models. Specifically, this project involved the training of a wide variety of neural networks. Supervised learning involves training models on a set of input variables X and output variables y. The models are trained to approximate a function "f" such that f(X) = y. By building models in this fashion, they can be used to make predictions of a future outcome y for new input values X.

For this project, the neural networks were provided with input data, X, involving players' college stats, and were trained to output predictions, y, of win shares, PER, or VORP. During the development process, the primary focus was on win shares, so that statistic will also be the focus for much of the following discussion.

4.1.1 Data Preprocessing

The input set X was processed accordingly to a fairly standard protocol. The categorical variables (basketball position, school, and conference) were treated with a one-hot encoder so as to be interpretable my machine learning algorithms. The rest of the variables, all numerical, were processed using a scaler in order to allow the models to better learn the relationships.

The output set construction process involved the isolation of a given player's best season. That is to say, the output value, y, for each sample in the training dataset was given by the maximum win shares value out of all the seasons played by the corresponding player. This decision came after much consideration. The initial idea was to fit the models to predict a player's win shares 6-7 years in the future, as it is generally assumed that NBA players tend to enter their primes between the ages of 25-28. However, predicting a player's value a specific number of years in the future can be misleading in many cases. Some players underperform and fall out of the league before they reach year seven. Some players peak late. Some players peak early and then get injured. Accordingly, training the models to predict the value of a player's best future season captures most of these cases and makes the best use of the data available.

4.1.2 Hyperparameter Tuning

The models were designed as sequential deep neural networks using Python's Keras framework. The architecture selections were improved through the use of an exploration tactic known as grid search. Grid search is a form of hyperparameter tuning that involves iteratively evaluating a wide range of architectural designs, experimenting with parameters such as the number of hidden neural network layers, the number of hidden units in each layer, search optimizers, and weight regularizers.

4.2 Model Types

4.2.1 Regression

Some of the networks were trained as regression models, attempting to predict the specific number of win shares a college player may provide in their NBA future.

Many of the neural networks were trained as regression models. That is to say, they were trained to predict the value of a continuous variable. For these purposes, this involved predicting the specific number of win shares a college player may provide at the peak of their NBA future. The same goes for the models that were trained on PER and VORP.

4.2.2 Classification

Other neural networks were trained as classification models. Achieving this involved involve breaking the NBA performance values down into buckets. For example, one was trained as a binary classifier, aiming to predict whether a player would peak at a PER value greater than the 15.0 average or less than the 15.0 average. Again, the league-average PER is 15.0 by definition, so PER was a helpful statistic to use for classification purposes. This model could be understood as one that makes a simple prediction: will this player peak as an above average NBA player or a below average NBA player? This could be rather useful in the evaluation of late 2nd round picks, where oftentimes that hope is that the pick will become a serviceable player, and anything beyond that is a bonus.

Another strategy involved breaking PER down into 6 buckets. This spawned a multi-class classification problem wherein the models attempted to output a single number, 0 through 5, relative to an input player. The PER breakdown means that the model returns 0 for players with PER less than 5, 1 for players with PER values between 5 and 10, 2 for players with PER values between 10 and 15, 3 for players with PER values between 15 and 20, 4 for players with PER values between 20 and 25, and 5 for players with PER values over 25.0.

This strategy is interesting because it could be interpreted as a 5-star system, with an extra value reserved for 0 stars. It should be noted that in general, the most valuable player award tends to go to a player with a PER value greater than 30.0. Accordingly, as a player's PER value begins to exceed 25, their name may begin to be brought up in the MVP conversations. All this to say, again, this 6-class model is at the very least an interesting one, as players with PER values over 25 could most definitely be referred to as 5-star players, which aligns with the labeling system of this model.

A final strategy involved a 3-class breakdown of PER. Given a player's college stats, these models predicted 0, 1, or 2. In this setup, predicted PER values greater than 20.0 are assigned a 2, predicted

PER values greater than 10.0 are assigned a 1, and anything below 10.0 is assigned a 0. These labels could be interpreted as “great,” “good,” and “bad.”

4.3 Highest-Performing Neural Network Architecture

After running various iterations of hyperparameter tuning, a final model was selected based on its performance. What follows is a description of the best model.

The highest-performing model was a sequential neural network with a Dense input layer containing 128 units and "relu" activation. Additionally, it contains two hidden layers with 128 units each and "relu" activation. The model was compiled with using stochastic gradient descent, and is fit for 1200 epochs and a batch size of 64. Mean squared error loss was used for regression tasks and categorical cross entropy loss was used for classification tasks.

5 Results

5.1 Classification

For each of the 3 classification tasks, the models tested on test sets of size 249.

The overall accuracy of each model is shown in Table 1 below.

Additionally, Figures 4, 5, and 6 in the appendix show the confusion matrices for each of the models. These matrices show the percentage of correct predictions for the data points of each class, as well as the nature of the errors made for predictions that assigned an incorrect class. For example, in the binary confusion matrix, the bottom-right quadrant tells us that of all the players were actually above average, the model correctly identified 70% of them. This means that in reality, their future NBA performance was above average, and the model made a correct prediction of that based on their college statistics. However, the bottom-left quadrant shows that the model incorrectly classified 30% of the truly above-average players as A“below average.”

Binary Classification Accuracy	**3-Class Classification Accuracy**	**6-Class Classification Accuracy**
73.85%	76.60%	56.88%

Table 1: Classification Model Accuracies

The same goes for the confusion matrices for the 3-class classifier and the 6-class classifier.

5.2 Regression

A wide range of regression models were trained to predict future win shares, PER, and VORP. The results are shown below in Table 2, and in Figures 7, 8, and 9 of the appendix.

Additionally, a point of interest is that for the win shares model, despite the squared error loss, its predictions were within a +/- 2.5 buffer of the real win shares 74% of the time.

Win Shares Loss	**PER Loss**	**VORP Loss**
0.5769230769230769	0.6735787561180395	0.7289251173176051

Table 2: Regression Model Losses: Mean Squared Error

6 Conclusion

6.1 Discussion

The models trained for this project show a lot of promise when it comes to using deep learning to evaluate potential NBA talent. In the areas where the models do not perform well, they fail in predictable ways.

For example, consider NBA All-Star Damian Lillard. Lillard has performed at an All-Star level for the majority of his career, with over 24 PER the last two years, and over 20 for the last five. However, all of the models in this project vastly underrate him. This certainly makes sense when considering the correlation information discussed and shown in section 2.3 and in Figure 2. Not only did Lillard play all four years in college, but he also played at a small and largely unknown college with a low strength of schedule. Historically, star NBA players tend to enter the league after flashy one-year campaigns at powerhouse, athletically top-ranked universities. Again, this is corroborated by the deep inverse correlation between "years in college" and future NBA win shares shown in Figure 2.

Accordingly, while this is certainly a work in progress, there are clear points at which attention can be applied for improvement. This project could certainly benefit from more data, and in particular, more features. It is clear that the information provided to the models does not tell the whole story.

6.2 Future Work

There are many ways that this project could be developed further. In future iterations, I believe that incorporating advanced stats from college into the input data has the potential to improve the performance of the models. While the college advanced statistics were accessible from the same pages as the regular box stats, this approach was avoided for the following reason: temporal availability. That is to say, the tracking of advanced statistics is a relatively new development, so those numbers are only available for newer players. Accordingly, training the models on examples using those numbers would vastly decrease the number of data points in use. While an imputer could be used to fill these values in with the average of the corresponding column, this still would decrease the overall quality of the data since it would have to be done for so many players. This poses a challenge, as there are so few data points in the first place because the number of spots in the NBA is very limited. Only including players for which all of the modern advanced stats are available would have decreased the size of the training set to well under 1000 points.

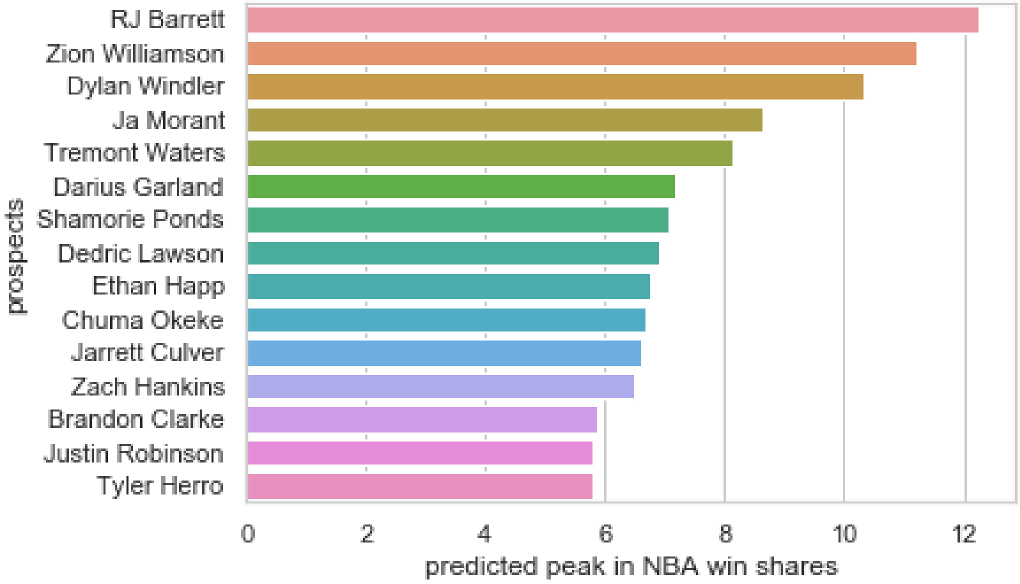
6.3 2019 Draft Predictions

In honor of draft day quickly approaching, this report concludes with a ranking of this year's NBA prospects. A trained model was run on the top 130 prospects for the 2019 draft, according to CBS Sports. Figure 3 below shows the 14 prospects it ranked highest in terms of expected future win shares.

The model's predictions seem to be largely in agreement with many analysts on certain fronts, as Zion Williamson, Ja Morant, and RJ Barrett are all being lauded as future stars.

One thing I'm very intrigued by is how much the model seems to like Dylan Windler and Sharmorie Ponds. The model projects both of them high into the lottery, despite their fairly low positions on many draft boards. This is also despite Windler playing three years in college and Ponds playing four, which, as discussed earlier, can be viewed as a red flag. Perhaps we have a couple of breakout stars on our hands.

Figure 3: 2019 Draft Prospect Ranking



References

[1] Fischer, Jesse. "Using machine learning to predict the long-term value of NBA draftees," http://www.tothemean.com/2014/06/17/machine-learning-predict-long-term-value-in-draft.html.

A Appendix

Figure 4:

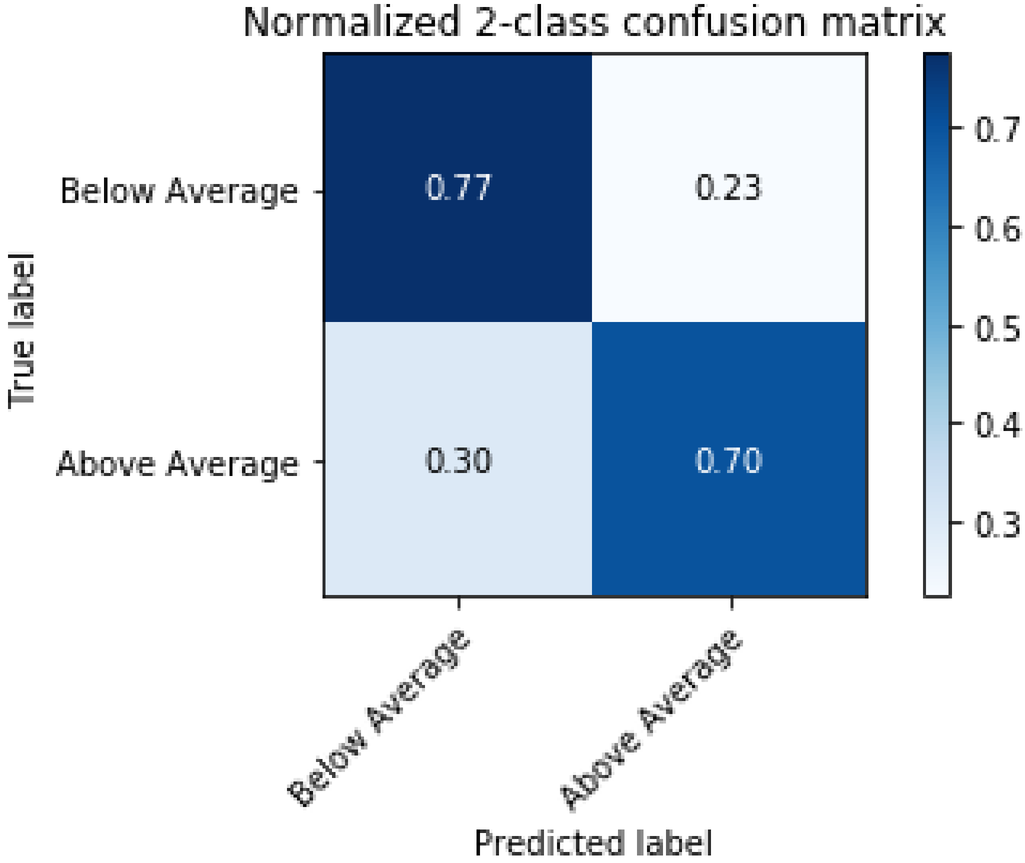


Figure 5:

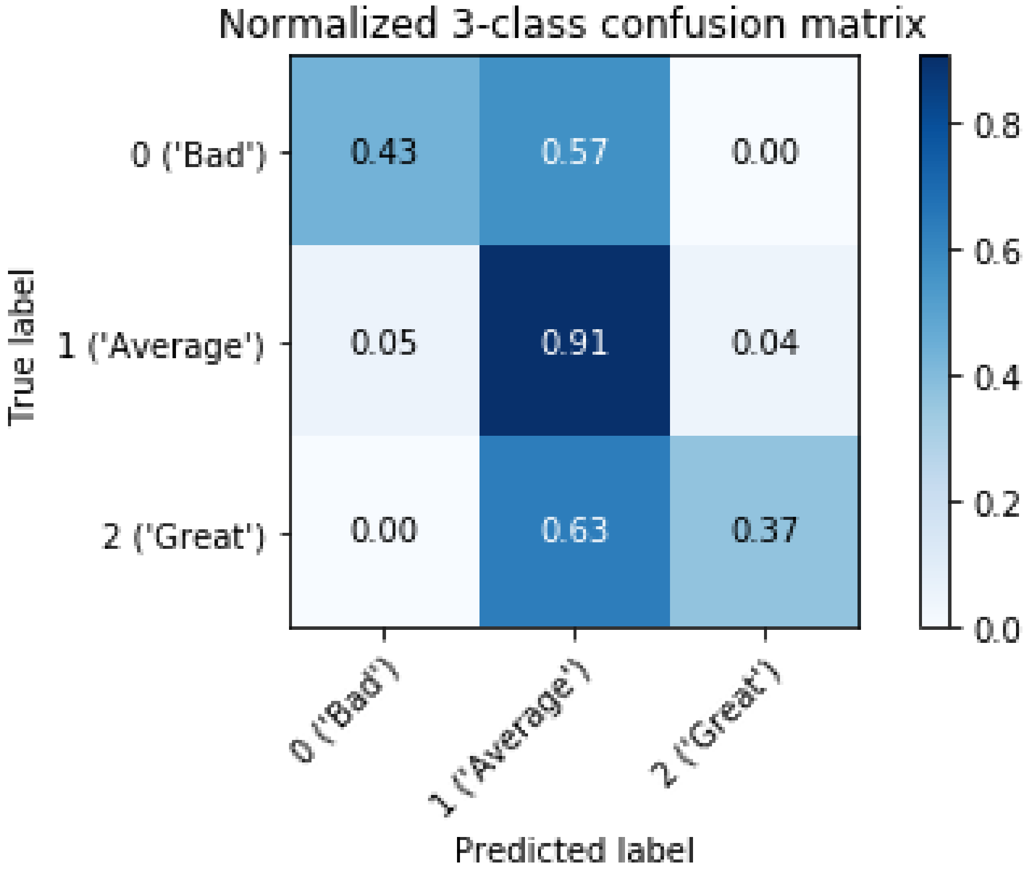
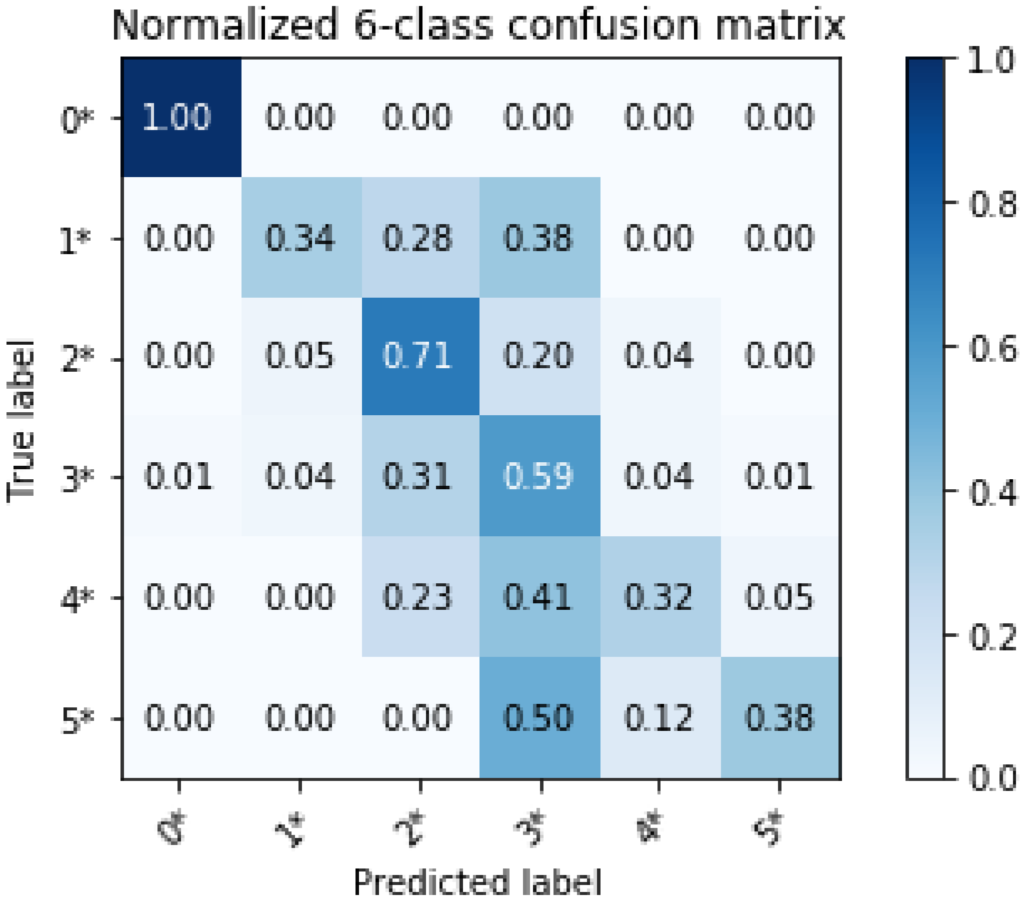

Normalized 3-class confusion matrix
True label
0 ('Bad')
1 ('Average')
2 ('Great')
0.43
0.57
0.00
0.05
0.91
0.04
0.00
0.63
0.37
0.8
0.6
0.4
0.2
0.0
0 ('Bad')
1 ('Average')
2 ('Great')
Predicted label

Figure 6:



	0*	1*	2*	3*	4*	5*
0*	1.00	0.00	0.00	0.00	0.00	0.00
1*	0.00	0.34	0.28	0.38	0.00	0.00
2*	0.00	0.05	0.71	0.20	0.04	0.00
3*	0.01	0.04	0.31	0.59	0.04	0.01
4*	0.00	0.00	0.23	0.41	0.32	0.05
5*	0.00	0.00	0.00	0.50	0.12	0.38

Figure 7:

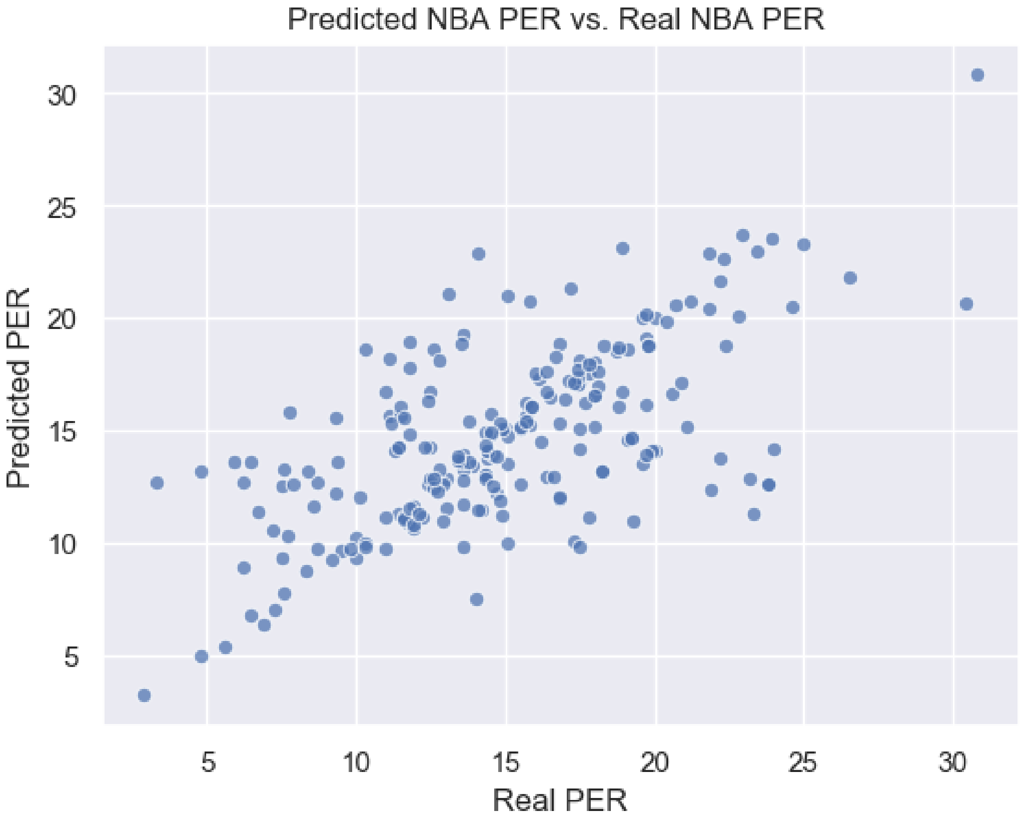

Predicted NBA PER vs. Real NBA PER
30
25
20
15
10
5
Predicted PER
5
10
15
20
25
30
Real PER

Figure 8:

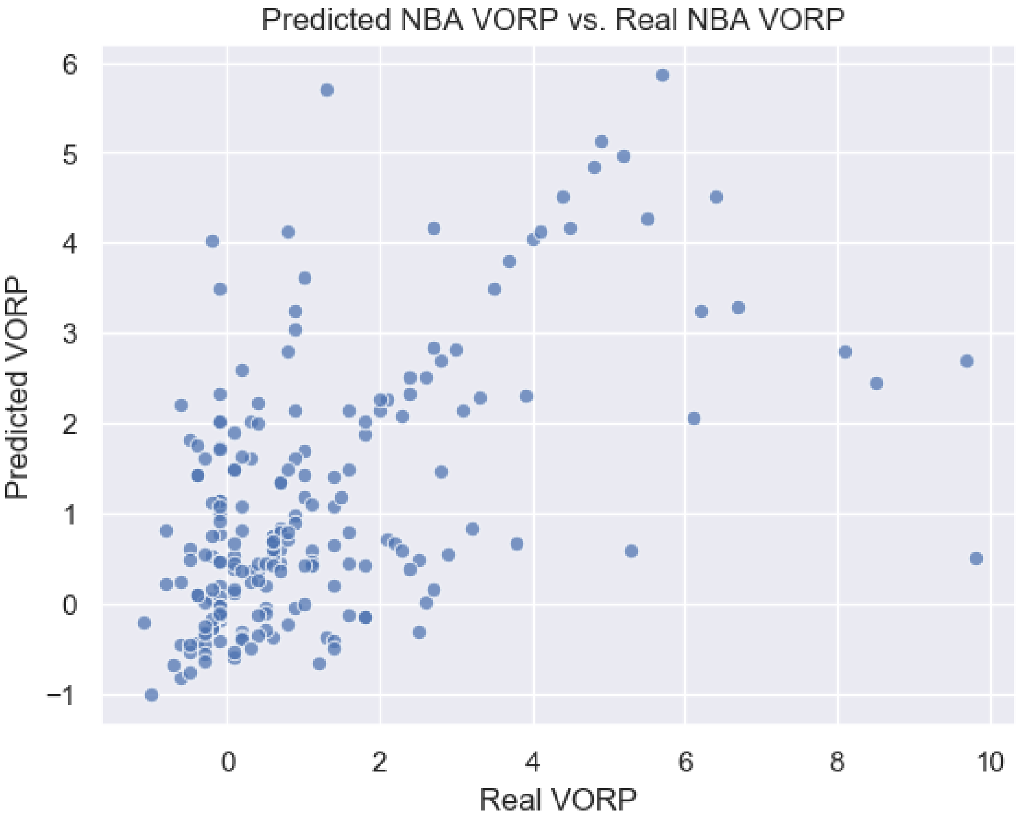
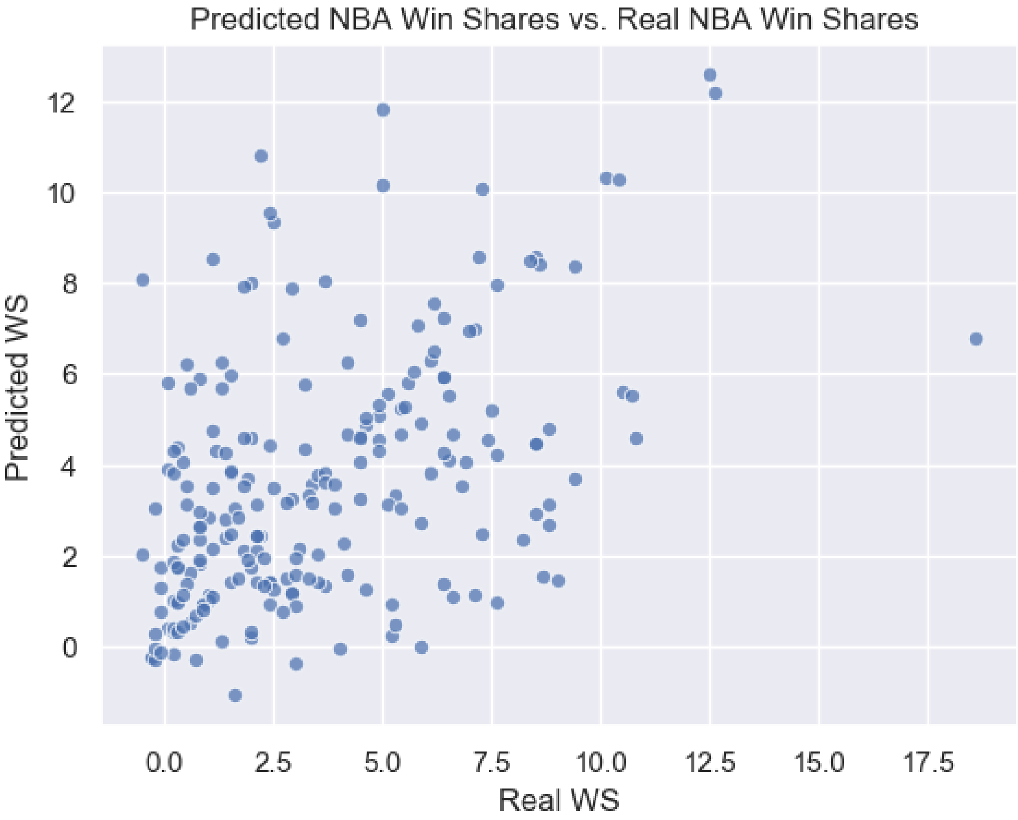

Predicted NBA VORP vs. Real NBA VORP
Predicted VORP
Real VORP
6
5
4
3
2
1
0
−1
0
2
4
6
8
10

Figure 9:


Predicted NBA Win Shares vs. Real NBA Win Shares
12
10
8
6
4
2
0
Predicted WS
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Real WS