Final Report

Yi Gao and Zhekun Wu
{yg185, zw244}@duke.edu

April 24, 2022

Press Release

NBA Draft Prediction Simulator is Now Available!

Figure 1: Stephen Curry got drafted with pick 7 in 2009

Basketball has been the second most popular sport in the world for a long time. More and more people become interested in playing and watching basketball games. In terms of basketball games, NBA is definitely one of the playgrounds for talented players that has the most

intense basketball matches and the most popularity in the world. Many great college players and also high school students are dreaming of being able to play in the NBA league. Every year in June or July around 60 players are selected by NBA teams to have the qualification to be trained as NBA players and play on the court in front of thousands of audience. It is known as the NBA Draft.

Before the NBA Draft, for those players who have great performance and unlimited potential, the NBA executives and scouts would spend enough time to inspect how these players play the games, collect all important stats of them, including not only the common statistics such as points, blocks, steals, assists, rebounds and turnovers per one game, but also some advanced stats such as effective field goal percentage and true shooting percentage, and analyze the player pool of that year based on the stats. These stats play an important role in deciding whether the player is good enough to attend NBA stage and also by comparing with all other players, deciding what rank one player could be among current year's available players.

Predicting the draft for our own favorite players is always an interesting and excited entertainment. If you are interested in simulating drafts for your favorite players or for all players of a specific year, our product is definitely an appropriate try for you! You do not need anything, just a year that your favorite player attended the draft. And you could see the predicted draft rank of all picked players in that year along with your favorite player!

Our dataset contains all needed stats including common stats and advanced stats of all players from 2009 to 2021. And the simulated draft prediction actually looks pretty good! Everyone who is interested in players draft situations for any of the recent years can try to have fun with it! Our product can also predict upcoming draft situations if the needed stats are provided. You might know the first pick of upcoming draft beforehand!

Frequently Asked Questions

1. **How does this system work?**

   Answer: Users can firstly download the project to the local machine with the following command:

   ```
   git clone https://github.com/YiiiGao/NBA_Draft_Prediction.git
   ```

   Then users need to enter the project directory and download the required packages with the following command:

   ```
   cd NBA_Draft_Prediction
   pip3 install -r requirements.txt
   ```

   Finally users can simply run the project with default setting with the following command:

   ```
   python3 main.py
   ```

   We also provide different options for users to explore our project. The argument list is shown below in Table 1.

Usage	Option	Default	Description
−−method	regression/ranking	regression	We provide two options for training and prediction. Users can choose the method they like.
−−year	2009 - 2021	2021	Users can select a year to make the prediction. The rest of the dataset will be used as the training set.
−−use_nn	True/False	False	Users can decide whether to use Neural Network for ranking. Note: Only available for ranking method.

Table 1: Argument list of the project

   For example, if the user wants to use ranking method, with Neural Network, to generate a prediction of college players in 2014, the following command is needed:

   ```
   python3 main.py --method ranking --year 2014 --use_nn True
   ```

2. **Which dataset does this system use?**

   Answer: The dataset used for both training and prediction is shown in the project repository, named *CollegeBasketballPlayers2009-2021.csv*. It is an unprocessed dataset, which includes the detailed information of all college players eligible for the NBA draft from 2009 to 2021. We carefully preprocessed this dataset, including feature selection, data imputation and encoding categorical features.

   - As for feature selection, since we focus on the performance of each player in college games, we carefully select 23 features shown in Table 2.

Abbreviation	Full Name
OREB	Offensive Rebounds Per Game
DREB	Defensive Rebounds Per Game
TREB	Total Rebounds Per Game
AST	Assists Per Game
BLK	Blocks Per Game
STL	Steals Per Game
PTS	Points Per Game
eFG	Effective Field Goal Percentage
MIN_per	Minutes Percentage
TS_per	True Shooting Percentage
FT_per	Free Throw Percentage
ORB_per	Offensive Rebounds Percentage
DRB_per	Defensive Rebounds Percentage
AST_per	Assists Percentage
BLK_per	Blocks Percentage
STL_per	Steals Percentage
TO_per	Turnovers Percentage
twoP_per	Two Points Percentage
TP_per	Three Points Percentage
yr	Year of College (Fr, So, Jr, Sr)
ast/tov	Assists/Turnovers
obpm	Offensive Box Plus/Minus
dbpm	Defensive Box Plus/Minus

Table 2: Feature list

- As for data imputation, since there is only one missing value of 'ast/tov' feature of a picked player within the dataset, we set the 'ast/tov' of that player to be 0.
- As for encoding, we only need to encode the feature 'yr'. We convert all four possible values into integer values, we set 'Fr' = 1, 'Jr' = 2, 'So' = 3, 'Sr' = 4, according to the lexicographical order.

3. **How is the system trained and how is the predictions made?**

   Answer: We provide two training methods, **regression** and **ranking**. The detailed descriptions are as follows.

   Regression

   Since we have approximately 60 picks in both training set and testset, we think regression models are better than classification models. Hence, nearly all regression models covered in class are trained and the best one with the highest NDCG score is chosen. Moreover, since the predicted scores are continuous, we need to sort all scores and transform them into rankings.

   Ranking

   We use the **pairwise approach** in **Learning to Rank** to train the system. Concretely speaking, the whole training process can be divided into the following steps.

- Transform the original dataset into 'player pairs'. This means given the statistics of player x and player y, we firstly concatenate features of x and y, then use 1 as the new label if the rank of x is larger than y, or use -1 otherwise.
- After we generated the new dataset, the ranking problem has been transformed into a binary classification problem. We train a list of classification models on the new training set and select the best one with the highest NDCG score to make predictions on the new testset.
- After we got the predictions, we calculate the scores of each player, and use the rank of scores as our final predictions.

4. **How will the system be evaluated?**

   Answer: We use Normalized Discounted Cumulative Gain, or **NDCG** as our primary evaluation metric. NDCG is an important measure of ranking quality, which can be computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{1}$$

   where DCG_p is the discounted cumulative gain at a particular rank position p, and $IDCG_p$ is the ideal discounted cumulative gain. DCG_p and $IDCG_p$ are computed as:

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \tag{2}$$

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)} \tag{3}$$

   where rel_i is the graded relevance of the result at position i, and REL_p is the list of relevant documents in the corpus up to position p.

   For example, if we make a prediction for players in 2021, the final results reach an NDCG score of 0.941 using regression, and an NDCG score of 0.916 using ranking, which indicate the great ranking qualities.

   We also use **mean absolute error** for regression, and **accuracy** for ranking, as the supplementary metrics.

5. **As for regression, which model does this system use? How is the parameters selected?**

   Answer: For the regression model, we selected seven models in total. For each model, we have tried enough times carefully tuning the parameters of the models in order to reach a relatively small enough MAE value and a relatively high enough NDCG score using **Randomized-Search-CV**. Models are related parameters are listed in Table 3. For parameters that are not listed in the table, we use default values in *sklearn*.

Model	Parameters
Linear Regression	-
Logistic Regression	penalty='l1', solver='liblinear'
Ridge	alpha=0.8, solver='lsqr'
Lasso	alpha=0.4, max_iter=200, selection='random'
XGBoost	learning_rate=0.05, max_depth=4, n_estimator=200
Random Forest	n_estimators=140, min_samples_split=6, min_samples_leaf=5, max_depth=50
K-Neighbors	n_neighbors=7

Table 3: Models and related parameters for regression

6. **As for ranking, which model does this system use? How is the parameters selected?**

   Answer: As is mentioned above, we have transformed the original dataset into data pairs, and transformed the original problem into a binary classification problem. Therefore, classification models are used to solve this problem. In addition, same as the regression method, we carefully searched best parameter settings using **grid-search**. Models and related parameters are listed below. For parameters that are not listed in Table 4, we use default values in *sklearn*.

Model	Parameters
XGBoost	max_depth=5, learning_rate=0.05
Random Forest	max_depth=6, criterion='entropy'
Decision Tree	-
K-Neighbors	n_neighbors=3
Gaussian Naive Bayes	-

Table 4: Models and related parameters for ranking

Meanwhile, we implemented a Neural Network using *pytorch* for ranking. The model architecture is shown in Figure 2. We simply add three linear layers with the ReLU activation function, and one Dropout layer to prevent overfitting.

7. **Why are a few of the predicted draft picks different significantly with the actual picks?**

   Answer: Our assumption is that since the age of one player is the same important as the statistics, NBA scouts and executives would accept those younger players who do not have outstanding statistics in college matches. Compared to older players, younger players are usually more talented and have more time to be trained. Therefore, they may be picked at a higher draft, and vice versa for the older players with a lower draft.

8. **What are the limitations of the dataset?**

   Answer: Our dataset only includes player statistics from 2009 to 2021, so it is not possible for the system to predict the rankings of players before 2009 (eg. Kobe Bryant) and in 2022 (eg. Paolo Banchero). Therefore, if more data is added, the system can be much

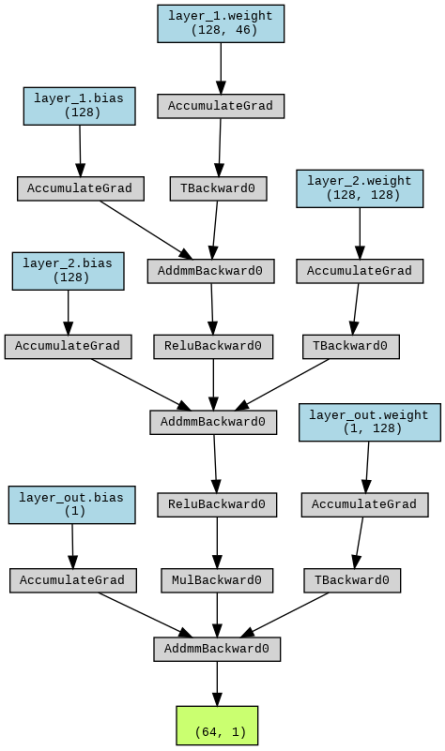

Figure 2: Model architecture

more powerful. In addition, this system does not support custom input testset, because our testset have to meet strict requirements (23 ordered features, few NaN values, etc.).

9. **Since you only focus on the college players, what about the international players and G-league players?**

   Answer: Unfortunately, we do not consider international players and G-league players because we do not have enough data. However, international players and G-league players do play an important role in the NBA draft. For example, Jalen Green, a player from G-league, was selected by the Houston Rockets with the second overall pick in the 2021 NBA draft. Therefore, we fixed this problem by making predictions on the relative rankings, instead of the actual picks of college players.

10. **Is it reasonable to consider only the statistics of every player in matches?**

    Answer: There are more factors to be considered such as injuries, physical and mental conditions. Physical conditions including heights, weights, agility and sprint ability. Mental conditions such as offensive and defensive mind.

11. **Is there any functionality that could be added in the future?**

    Answer: One possible functionality would be predicting the rank for one specific player in different year. For instance, if we want to know what rank Stephen Curry would get in 2021, we could take the information of Stephen Curry as input, and output the predicted rank for Curry in 2021 based on all the players statistics in 2021 draft.

Technical Details

The dataset, codes, sample predictions and instructions are all uploaded to `https://github.com/YiiiGao/NBA_Draft_Prediction`. Feel free to contact us if you have any questions!