ISyE 6416 – Computational Statistics – Spring 2020

Final Report

Team Member Names: Yinjie Xu, Zhiwei Liao, Linlin Liu

Project Title: Predicting 2020 NBA Awards Winners

Contents

1 Problem Statement

The National Basketball Association (NBA) is the most popular professional basketball league in the world, where all the outstanding basketball players compete. Every June, after the playoffs, the NBA presents annual awards to honor and recognize the league's top performers of the year. The awards are determined by the voting results of a panel of sportswriters and broadcasters throughout the United States and Canada: each member of the voting panel casts a vote for first to fifth place selections, which worth 10, 7, 5, 3, 1 points, respectively. Starting from 2010, one ballot was also cast by fans through online voting. The player with the highest point total will win the award. Generally speaking, there are no official criteria for the awards. However, there are some unofficial (i.e., widely accepted) criteria, which includes but is not limited to player's performance statistics, teams' records, and sometimes even the players' social influence.

Our team is interested in predicting the NBA annual awards winners based on the quantifiable data. More specifically, our target is to predict the Most Valuable Player (MVP), Rookie of the Year (ROY), Defensive Player of the Year (DPOY) recipients based on the players' performance statistics and the teams' records. We use voting share as the explained variable, which is the ratio of points won by player over the maximum possible points won. By predicting the voting sharing instead of whether the player is the winner, we transform the binary classification problem into a regression problem. Six regression models we learned in ISyE 6416 are implemented and compared, which are linear regression, ridge regression, lasso regression, support vector machine (SVM), decision tree regressor and random forest regressor. By comparing our predicted voting shares to the actual voting shares, we identify the most appropriate regression model to predict each award. Also, by implementing the feature-selection models (lasso regression and random forest regressor), we can identify which features play more significant roles in the selection of awards winners.

2 Methodology

The goal of our project is to identify the best regression model to predict each award. We select the voting shares as the explained variable. Then, we will tackle the problem in three steps. Firstly, we implement feature selection to simplify models. Several features are removed based on their empirical explaining power and correlation matrix. Secondly, we build the 6 regression models using sklearn packages with the selected features for 3 awards, respectively. The parameters in each learning model are selected by K-fold cross-validation, where one year among all the years of training data is used as the validation set alternately. Lastly, multiple metrics are used to measure the accuracy of prediction, which include mean squared error, explained variance score, median absolute error, and mean absolute error. The best regression model to predict each award will be selected based on the error metrics over the test set.

3 Data Collecting and Preprocessing

3.1 Data Source

The data source of this project is Basketball-Reference[1], which is a website containing basketball statistics, scores, and history for worldwide basketball competitions. Our dataset consists of 3 parts: (1) the performance statistics of all NBA players (over 500) for all selected seasons; (2) the miscellaneous performance statistics of all teams for all selected seasons, among which we use the win ratio (games won / total games) as the indicator of the team's overall performance; (3) the voting statistics for all selected seasons, including the candidates' names and the shares won. Players who received zero vote will be assigned zero voting share.

We collect all the aforementioned statistics from season 1999-2000 to season 2018-2019, as our dataset. Besides, we also collect the current statistics of season 2019-2020 for prediction of 2020 MVP, ROY, DPOY. (Since the

NBA has suspended its season due to COVID-19, we can only obtain the performance statistics until March 11.) For most of the data we need, we directly download the data files from github/gmalim/NBA_analysis[2], which were obtained from basketball-reference.com. For the missing data, we collect from the website by ourselves.

3.2 Data Processing

3.2.1 Feature selection

Firstly, we remove features with no predictive power, such as player's name, position and age. The team name is replaced by its win ratio. Then, we remove features that are correlated with other features based on the correlation heat plot as shown in Figure 1. Next, we empirically drop irrelevant features for each award. For example, when predicting DPOY (defensive player of the year), we drop the features which evaluate offense ability only. In addition, to avoid influence caused by edge cases, we remove players who play less than a threshold number of games. The final selected features are listed in Table 1. The abbreviations are explained in the appendix.

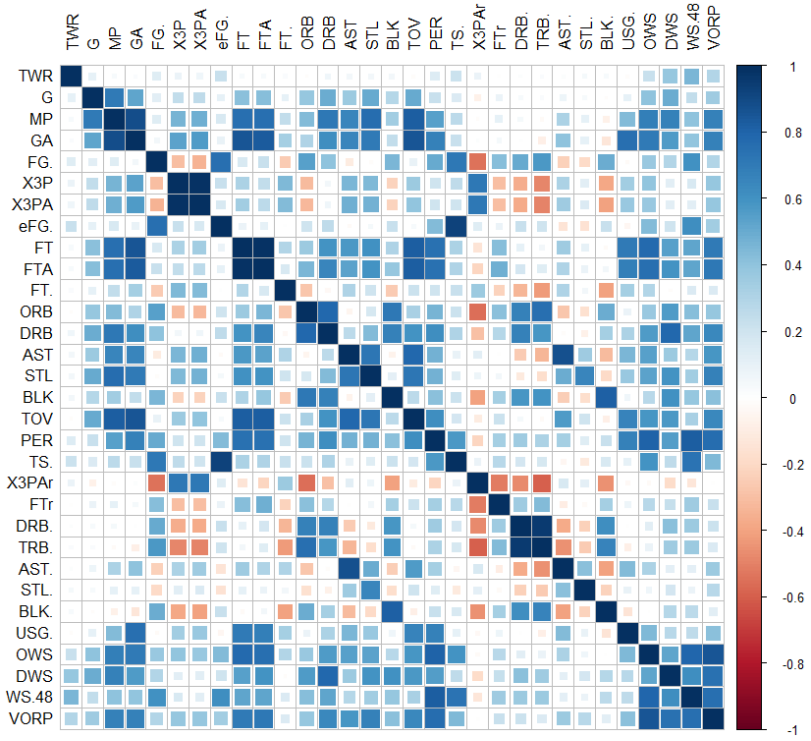

Figure 1: Features correlation heatplot

MVP	TWR, G, GS, MP, FG, FG, X3P, X3P%, X2P, X2P%, eFG%, FT%, ORB, DRB, AST, STL, BLK, TOV, PF, PTS, PER, ORB%, DRB%, AST%, STL%, BLK%, TOV%, USG%, OWS, DWS, WS, OBPM, DBPM, VORP
DPOY	TWR, G, GS, MP, ORB, DRB, TRB, STL, BLK, PF, PER, ORB%, DRB%, TRB%, STL%, BLK%, USG%, DWS, WS, WS/48, DBPM,'BPM, VORP
ROY	G, MP, FG, FGA, 3P, 3PA, FT, FTA, ORB, TRB, AST, STL, BLK, TOV, PF, PTS, FG%, 3P%, FT%, MP/G, PTS/G, TRB/G, AST/G

Table 1: Selected features

3.2.2 Train-test set split

We randomly choose 2 years as the test set and data of the rest 18 years is concatenated as the training set. The MVP voting shares distribution of the training set is plotted in Figure 2, which shows the sparsity of explained variable.

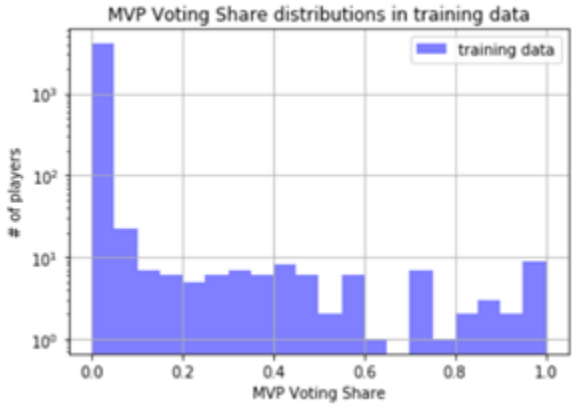

Figure 2: MVP voting share distribution

3.2.3 Data standardization

Features scaling is important because we are using features with different units, which do not contribute equally to the analysis and might end up creating a bias. In linear regression models, the weights will have more contribution to the features with much larger scale. Thus, before we fit the data into models, we first standardize the features.

4 Results and Evaluation

4.1 Error Metrics and Selection of Models

After tuning and determining the parameters for each model, we compare the predicted voting shares and the actual shares of all test years. The error metrics for MVP, DPOY, ROY prediction is shown in Table 2, 3, 4, respectively.

MVP Error Metrics	Mean Square Error		Mean Absolute Error		Median Absolute Error		Explained Variance Score		R^2 Score
	train	test	train	test	train	test	train	test	
Linear regression	0.00743	0.00749	0.0458	0.0466	0.0269	0.0270	0.394	0.454	0.394
Ridge regression	0.00778	0.00817	0.0452	0.0470	0.0272	0.0288	0.365	0.458	0.365
Lasso regression	0.00769	0.00818	0.0449	0.0479	0.0260	0.0278	0.373	0.403	0.373
SVM	0.00169	0.00484	0.0131	0.0364	0.0099	0.0155	0.863	0.646	0.862
Decision Tree	0.00242	0.00347	0.0128	0.0192	0.0020	0.0020	0.803	0.747	0.803
Random forest	0.00769	0.00818	0.0449	0.0479	0.0260	0.0278	0.373	0.403	0.944

Table 2: MVP Error Metrics

DPOY Error Metrics	Mean Square Error		Mean Absolute Error		Median Absolute Error		Explained Variance Score		R^2 Score
	train	test	train	test	train	test	train	test	
Linear regression	0.005204	0.007017	0.0305	0.0313	0.0176	0.0153	0.205	0.216	0.205
Ridge regression	0.005281	0.007311	0.0298	0.0330	0.0175	0.0195	0.193	0.182	0.193
Lasso regression	0.005287	0.007329	0.0299	0.0324	0.0175	0.0182	0.192	0.180	0.192
SVM	0.002748	0.006447	0.0131	0.0254	0.0072	0.0076	0.584	0.281	0.580
Decision Tree	0.003247	0.008083	0.0160	0.0205	0.0039	0.0039	0.504	0.097	0.504
Random forest	0.000529	0.004670	0.0058	0.0186	0.0001	0.0001	0.919	0.477	0.919

Table 3: DPOY Error Metrics

ROY Error Metrics	Mean Square Error		Mean Absolute Error		Median Absolute Error		Explained Variance Score		R^2 Score
	train	test	train	test	train	test	train	test	
Linear regression	0.00962	0.00732	0.0441	0.0392	0.0184	0.0175	0.338	0.347	0.338
Ridge regression	0.01018	0.00794	0.0410	0.0366	0.0156	0.0172	0.299	0.285	0.299
Lasso regression	0.00985	0.00752	0.0431	0.0382	0.0181	0.0188	0.322	0.327	0.322
SVM	0.00486	0.00445	0.0111	0.0137	0.0008	0.0007	0.671	0.602	0.666
Decision Tree	0.00717	0.00782	0.0247	0.0196	0.0100	0.0100	0.506	0.294	0.506
Random forest	0.00985	0.00752	0.0431	0.0382	0.0181	0.0188	0.322	0.327	0.909

Table 4: ROY Error Metrics

From the above tables, we can conclude that the SVM model performs the best in predicting all 3 awards, followed by the decision tree and random forest model. The value of R-square score and explained variance score of the last three models (SVM, decision tree, random forest) are closer to 1, which means the dependent variable y is better explained by the selected features. By contrast, these 2 metrics of the first three models (linear regression, ridge regression, lasso regression) are much less, indicating a poor fitting degree. Furthermore, the SVM model generally performs the best in mean square error, mean absolute error and median absolute error.

Based on the training and testing results, we select SVM as the regression model for DPOY and ROY prediction. SVM and decision tree are both used for MVP prediction, and the final rank is the average predicted rank over the 2 models.

4.2 Features Evaluation

The training results of the decision tree model reveal to us the features which play important roles in determining the winners of awards. The decision tree structures to predict voting shares for 3 awards are shown in Table 5, from which we can see PER, TWR, FG, OBPM and TOV are most important for determining MVP voting shares while DWS, DBPM, ORB% are essential factors for determining DPOY. As for the ROY, PTS/G, BLK and MP/G are the key factors. We also show the 10 most significant features based on the corresponding feature importance in the random forest model (Table 5). It's not surprising that these features are highly similar to the features listed by the decision tree model.

	MVP	DPOY	ROY
Decision Tree	|--- PER <= 2.26 | |--- PER <= 1.52 | | |--- value: [0.00] | |--- PER > 1.52 | | |--- FG <= 2.05 | | | |--- value: [0.08] | | |--- FG > 2.05 | | | |--- TWR <= 1.12 | | | | |--- WS <= 2.76 | | | | | |--- value: [0.12] | | | | |--- WS > 2.76 | | | | | |--- value: [0.57] | | | |--- TWR > 1.12 | | | | |--- value: [0.66] |--- PER > 2.26 | |--- TWR <= 1.12 | | |--- OBPM <= 2.64 | | | |--- value: [0.09] | | |--- OBPM > 2.64 | | | |--- X3P <= 2.11 | | | | |--- value: [0.40] | | | |--- X3P > 2.11 | | | | |--- value: [0.80] | |--- TWR > 1.12 | | |--- TOV <= 0.46 | | | |--- value: [0.28] | | |--- TOV > 0.46 | | | |--- value: [0.79]	|--- DWS <= 3.04 | |--- DBPM <= 1.31 | | |--- value: [0.00] | |--- DBPM > 1.31 | | |--- WS.48 <= 2.59 | | | |--- value: [0.06] | | |--- WS.48 > 2.59 | | | |--- value: [0.42] |--- DWS > 3.04 | |--- ORB. <= 1.63 | | |--- value: [0.12] | |--- ORB. > 1.63 | | |--- value: [0.74]	|--- PTS.G <= 2.71 | |--- BLK <= 7.34 | | |--- value: [0.01] | |--- BLK > 7.34 | | |--- value: [0.55] |--- PTS.G > 2.71 | |--- MP.G <= 2.26 | | |--- FGA <= 2.26 | | | |--- value: [0.58] | | |--- FGA > 2.26 | | | |--- value: [0.15] | |--- MP.G > 2.26 | | |--- value: [0.78]
Random Forest	PER 0.4896 TWR 0.0836 WS 0.0617 VORP 0.0581 TOV 0.0274 USG% 0.0250 PTS 0.0235 DWS 0.0215 eFG% 0.0169 2P 0.0166	DWS 0.2183 DBPM 0.1471 BLK 0.1022 WS/48 0.0683 USG% 0.0622 TWR 0.0557 WS 0.0423 ORB% 0.0322 STL% 0.0281 PER 0.0261	PTS/G 0.3974 MP/G 0.1242 FT% 0.0574 FG% 0.0558 AST/G 0.0456 3P% 0.0403 TRB/G 0.0343 BLK 0.0315 STL 0.0270 AST 0.0231

Table 5: Features Evaluation

From the above table, we are able to obtain an intuition on the selection criteria of the 3 awards. MVP is determined mostly based on the team's record and the player's offense ability, while more weights are put on the previous. In short, MVP is actually the "most valuable" player in the "most valuable" team. Similarly, DPOY is also determined based on these 2 aspects, which are evaluated more on the defense ability. The player who has the best defense statistics among all key defense players in each team will have the highest possibility of being selected as DPOY. For ROY, good statistics in both offense (points, assists) and defense (rebounds, blocks, steals) is the most essential, while high usage and minutes play per game are more like conditional terms. (Rookies who play more on the court are more likely to obtain high statistics. From the analysis of Table 5, the best-performed rookie is awarded based on the personal performance rather than the team's overall performance, while the selections of MVP and DPOY rely heavily on the term's record. This analysis well explained why Lebron James and Luka Dončić won the 2003 and 2019 ROY respectively even their teams didn't clinch the playoffs.

5 Prediction of 2020

Lastly, we apply the best-performed regressors on the current statistics of season 2019-2020 and make our prediction of 2020 NBA Awards winners. In the previous analysis, we have concluded that SVM and decision tree regressor perform the best in predicting MVP, while only SVM performs well in predicting DPOY and ROY. Hence, we predict the top-5 MVP candidates for 2020 in order of the average predicted rank over SVM and decision tree regressor. Similarly, we predict the top-5 DPOY candidates and ROY candidates using the shares predicted by SVM. Our final prediction of 2020 awards candidates are listed below. We will be excited to compare our prediction to actual winners and validate our learning once the 2020 NBA awards are announced.

NBA Most Valuable Player 2020: Giannis Antetokounmpo

Prediction : 2020 NBA MVP		
Avg. Rank	Name	Team
1.5	Giannis Antetokounmpo	MIL
2.5	Luka Doncic	DAL
3	James Harden	HOU
4	Kawhi Leonard	LAC
5.5	Anthony Davis	LAL

NBA Defensive Player of the Year 2020: Giannis Antetokounmpo

Prediction : 2020 NBA Defensive Player of the Year		
Rank	Name	Team
1	Giannis Antetokounmpo	MIL
2	Anthony Davis	LAL
3	Luka Doncic	DAL
4	Kawhi Leonard	LAC
5	Brook Lopez	MIL

NBA Rookie of the Year 2020: Ja Morant

Prediction : 2020 NBA Rookie of the Year		
Rank	Name	Team
1	Ja Morant	MEM
2	Zion Willamson	NOP
3	Luka Samanic	SAS
4	Eric Mika	SAC
5	RJ Barrett	NYK

6 Analysis of 6 Regression Models

After tuning and evaluating the 6 regression models, we compare their pros and cons, which are listed in Table 5.

Models	Pros	Cons
linear regression	• Closed form of solution • Easy to understand	• Assume linear relationship • Rely on assumption that features follow normal distribution and constant error variance
ridge regression	• Closed form of solution	• Cannot perform feature selection
Lasso regression	• Perform feature selection and give sparse weights	• Difficult to converge
SVM	• More accurate in high-dimensional spaces • Memory-efficient	• Computationally inefficient
Decision tree	• Ideal model to capture interactions between features • Easy-to-understand structure	• Unstable • Uninterpretable when the number of nodes is large
Random forest	• Generate internal unbiased estimate of the generalization error • Capture nonlinear interactions between features and target • Good at handling tabular data with numerical features	• Computationally inefficient

Table 6: Pros & Cons of models

7 Collaboration

Collecting data – Yinjie Xu
Preprocess data – Linlin Liu
Implementing regression models – Zhiwei Liao, Yinjie Xu
Running experiments and evaluating results – All
Final report writing – All

8 Conclusion

In this project, we investigate an interesting topic of predicting the NBA awards winners by implementing 6 regression models (linear regression, ridge regression, lasso regression, SVM, decision tree regressor and random forest regressor). Based on the testing results, we identify the best-performed regressors for predicting each award and apply them to predict the 2020 awards winners. Among the 6 models, SVM generally performs the best on predicting all awards. In addition, by observing the training results of decision tree and random forest models, we analyze the most essential features in determining each award winner. Lastly, we compare and analyze the pros and cons of the 6 regression methods based on our experience, which extend what we have learned in class to real application.

9 Reference

[1] URL: https://www.basketball-reference.com
[2] URL: https://github.com/gmalim/NBA_analysis/tree/master/data

10 Appendix

Player Total Statistics
G – Games played
GS – Games Started
MP – Minutes played
FG – Field Goals Per Game
FGA – Field Goals Attempts Per Game
FG% – Field Goal Percentage
3P – 3-Point Field Goals Per Game
3PA – 3-Point Field Goal Attempts Per Game
3P% – 3-Point Field Goal Percentage
2P – 2-Point Field Goals Per Game
2PA – 2-Point Field Goal Attempts Per Game
2P% – 2-Point Field Goal Percentage
eFG% – Effective Field Goal Percentage
FT – Free Throws Per Game
FTA – Free Throws Attempts Per Game
FT% – Free Throw Percentage
ORB – Offensive Rebounds Per Game
DRB – Defensive Rebounds Per Game
TRB – Total Rebounds
AST – Assists Per Game
ST – Steal Per Game
BLK – Blocks Per Game

TOV – Turnovers Per Game
PTS – Points Per Game

Player Advanced Statistics

PER – Player Efficiency Rating (A measure of per-minute production standardized).
TS% – True Shooting Percentage
3PAr – 3-Point Field Goal Attempts Rate
FTr – Free Throw Attempt Rate
USG% – Usage Percentage
WS – Win Shares (An estimate of the number of wins contributed by a player)
OWS – Offensive Win Shares
DWS – Defensive Win Shares
WS/48 – Win Shares Per 48 Minutes (An estimate of the number of wins contributed by a player per 48 Minutes)
BPM – Box Plus/Minus
OBPM – Offensive Box Plus/Minus
DBPM – Defensive Box Plus/Minus
VORP – Value Over Replacement Player