Analysis of NBA Top Players Based on Shooting Motions and Season Statistics

Javier Alberto Bernal Sigala
javieralberto.bernalsigala@studenti.unipd.it

Giovanni Dal Mas
giovanni.dalmas.1@studenti.unipd.it

Xianlong Wu
xianlong.wu@studenti.unipd.it

1. Introduction

The National Basketball Association (NBA) is a professional basketball league in North America comprising of 30 teams. Since its inception in 1946, the NBA has managed to attract billions of fans of all ages, making basketball the 5th most popular sport in the world and the second most popular in the US.

Statistics play a vital role in the analysis and evaluation of a NBA player's performance and their impact on the team. From points scored to rebounds grabbed and assists dished out, even the shooting motion, there is a vast array of statistics that are tracked and analyzed.

In this project, we will delve into the world of NBA statistics, exploring the various metrics that are used to measure a player's contribution to their team. We will use tools based on high dimensional data statistics in order to understand relevant aspects of the datasets, and answer the following questions.

- Predict a NBA player's salary:

  See if the salary is especially related to some subset of features and try to estimate it.

- Which components of the shooting motion are most relevant to the accuracy of the player?:

  We want to understand if it is possible to establish if a "shooting technique" is more accurate than another in general terms.

- Can we define a criterion to classify the position of a player based on his attributes?:

  Define which aspects a coach should pay attention to regarding a player to decide in which position he will be more efficient.

2. Datasets

The main dataset is a composition of 5 datasets from different sources. This section will explain in depth what each dataset consists of. Each of the tasks was approached using different combinations of these datasets.

Term	Explanation	Term	Explanation
G	Games	FT	Free Throw Per Game
GS	Games Started	FTA	Free Throw Attempted
MP	Minutes Played Per Game	PTS	Points per Game
FG	Field Goals Per Game	PF	Personal Fouls Per Game
FGA	Field Goal Attempts Per Game	TOV	Turnovers Per Game
FG.	Field Goal Percentage	BLK	Blocks Per Game
X3P	3-Point Field Goals Per Game	STL	Steals Per Game
X3PA	3-Point Field Goal Attempts Per Game	AST	Assists Per Game
X3P.	3-Point Field Goal Percentage	TRB	Total Rebounds Per Game
X2P	2-Point Field Goals Per Game	DRB	Defensive Rebounds Per Game
X2PA	2-Point Field Goal Attempts Per Game	ORB	Offensive Rebounds Per Game
X2P.	2-Point Field Goal Percentage	FT%	Free Throw Percentage
PER	Player Efficiency Rating	TS.	True Shooting Percentage

Figure 2. Features abbreviation meaning.

2.1. NBA 2K20 Ratings

NBA2k20 is a basketball simulation video game, developed by the company "2k sports" and released in 2020. Although the dataset [1] comes from a videogame, we must consider that this data are actually very faithful to the reality and can be indeed good metrics to describe ad measure different aspects of a player's game.

From this dataset, 37 numerical attributes were obtained (in a score range from 0 to 99) with which the ability to shoot from long distance, energy, strength and other characteristics of more than 100 players are evaluated.

Figure 1. 2KSports logo

2.2. NBA 2020-2021 Season Player Statistics

These are 3 datasets [2] that contains basic and advanced player statistics of nba players in the season 2020-2021.

The first is related to the performance of 480 players expressed in average number of games played, the second is similar, only the results are expressed per minute, while the third is focused on the direct effect of the player to win the game (these are the "advanced" statistics of NBA). Figure 2 states what the abbreviations of the features stand for.

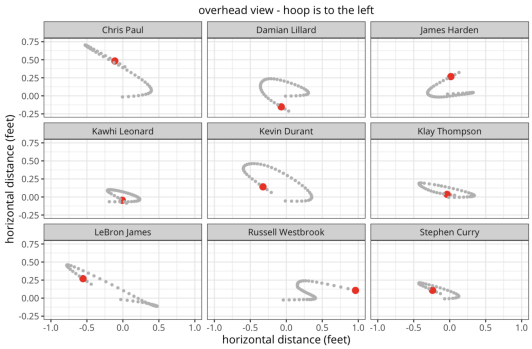

Figure 3. Overhead view from the recreated trajectory of the shooting motion. **Source:** taken from "NBA Player Shooting Motions: A Data Dump".

2.3. NBA Player Shooting Motions

These datasets[3] are described in the work "NBA Player Shooting Motions: A Data Dump" where they document and recreate the shooting motions of different top NBA players based on specific points of the shooting trajectory using a LOESS curve as shown in figure 3.

Here is contained a dataset with information on the physical variables of the ball's trajectory from the beginning of the shooting motion until the player releases the ball, there is information on more than 30 conditions such as "when the ball changes direction", or "how long does it take for the player to release the ball", "at what distance from the initial position is the ball released" and more. The second dataset is a detail of the path of the ball during the shooting motion, using the coordinates, velocities and acceleration in 3 axis, plus the distance changes with respect to the last step, and indicating at which time the features were extracted.

The datasets are about more than 190 top players, where it is described if the player hit, missed or was being defended during the shot, all the shots are relative to three points (25m distance).

2.4. NBA Team Statistics

Various NBA team statistics (Wins, Losses, Points, etc...) scraped from stats.nba.com for the 2000-2001 to 2020-2021 regular and playoff seasons. In this report, we restrict ourselves to the team statistics of the season 202-2021 as they are coherent with the time of other dataset. Furthermore, in this dataset, some of the features have the same name as that of the player dataset, these features are kept during the joining phase as they are team statistics that are different from the player ones.

2.5. NBA Player Salary

The original dataset [4] consists of 464 players and 15 features each of the year 2020. However, most of the features are already present in other dataset, thus only the columns of height, weight and annual salary of the players are taken into consideration. Furthermore, the height and weight columns includes double units of measure , namely feet/meters and pounds/kilograms, for the sake of convenience, we will keep the heights in meters and the weights in kilograms.

3. Experiment

In this section, the methodology used to solve each of the tasks will be discussed. For all the tasks discussed below, the train test split employs a split ratio of 0.7.

3.1. NBA Player salary

In this task, the NBA 2020-2021 Season Player Statistics, the NBA 2k20 Ratings and the NBA Player Salary are considered. After properly joining the dataset and data cleaning, we are left with a dataset with 77 players and 122 variables each player. Different regression models such as: Ridge, LASSO, Elasitic-net, Adaptive LASSO are used. All the models are fitted with 'glmnet' package and for the selection of the hyper-parameter, the 'cv.glmnet' is used. Furthermore, for the adaptive lasso, the parameter ω_j are simply determined by the ridge regression with a pre-defined search grid of 1000 lambdas. And finally, the models are fitted with the best hyper-parameters selected by the cross-validation and then prediction is made based on the newly fitted model. As for the assessment of the model, the mse is used.

3.2. Shooting Motion and Accuracy

For the analysis of the accuracy of the players vs. their shooting motions, a merge of the dataset that mentions the player's season statistics was made with the dataset of the shooting motion tracks.

It was necessary to do a re-sample since not all players had the same amount of time stamps, so 100 samples were taken from each (3 coordinates, 3 velocities, 3 accelerations and 3 distance changes in 100 samples) plus the total time and the players height, which gave a total of 1502 columns, since each coordinate of each time was expressed as a feature to model a player's shot. And there were 78 players after the cleaning of outliers and missing values.

The methods used were the same as in the previous subsection, group lasso was tested with two grouping methods, grouping them by time and grouping them by type of variable. A simple lasso regression on 4 subsets of the original dataset was also used, where groups were manually separated into a group of only coordinate components, only velocity components, only acceleration components, and only absolute values of velocity and acceleration. The least square error of the cross-validation was used as a metric to establish the lambda parameters.

```
 C PF PG SF SG
40 39 40 38 39
```

Figure 4. Classes distribution after oversampling

3.3. NBA Player position classification

For this task the datasets conveniently merged were the NBA 2020-2021 Season Player Statistics, the NBA 2k20 Ratings and the NBA Player Salary. The initial dataset contained 77 players and 85 features. Some of the variables needed to predict the player salary were dropped for this task they are not relevant (e.g. the team a player belongs to, that was one hot encoded to 30 different variables in the previous section). The task is a multiclass classification with 5 classes, the positions C = Center, PG = Point Guard, PF = Power Forward, SF = Small Forward and SG = Shooting Guard. The model used is the Multinomial Logistic Regression from the *glmnet* package, while the baseline to compare the performance of the model is a Random Forest from the *caret* package. After a first look at the distribution of the classes it was clear that the data needed to be balanced. Moreover, in general, all the classes were pretty poor represented (they all contained less than 20 examples each) and this could have been a problem for the cross-validation. For this reason we opted to oversample the whole dataset. The oversampling algorithm we chose is the RACOG (RApidly COnverging Gibbs)[5], that generates new examples with respect to an approximated distribution using a Gibbs Sampler scheme. Unlike other methods, RACOG was preferable for this type of task because it generates discrete examples, indeed very accurate and close to the real ones. Figure 4 shows the final distribution of the classes. For the fitting of the Multinomial Logistic Regression we used the 'glmnet' function with family = 'multinomial', nlambda = 100. The cross-validation was performed through the 'cv.glmnet' function with parameters family= 'multinomial', type.measure = 'class', lambda = grid (100 lambda between 10e-10 and 10), nfolds = 5. To assess the model we used the missclassification error and the accuracy.

4. Results

4.1. NBA Player Salary

The hyper-parameters are selected by cross-validation, an example of selection of lambda for different models is reported in the figure 5. And the selected parameters of different models are reported in the table 1.

Notice that the lambda of LASSO and Elastic-net are slightly different even if the alpha as the search grids are different.

The plot for parameter selections is shown at figure.6.

Most of the coefficients fitted for the Ridge regression are close to zero which is expected. We can see that the larger coefficients are related to the teams such as Golden

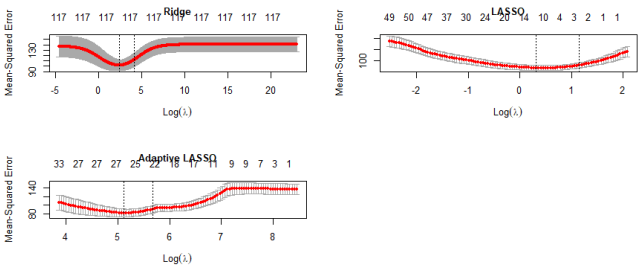

Figure 5. MSE of cross validation for different models (salary).

MODEL	α	λ
Ridge	0	10.94
LASSO	1	1.37
Elastic-net	1	1.3
Adp. LASSO	-	166.89

Table 1. Hyper-parameters of the model (rounded for salary).

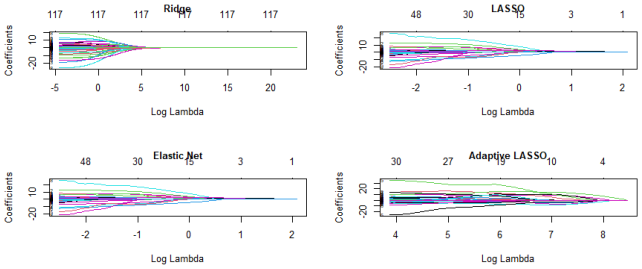

Figure 6. Growth of parameters (salary).

States Warriors which has more star players thus it is reasonable that they have a positive correlation with the salary and on the other hand, team like Memphis Grizzles has an negative correlation with the salary as they do not have a lot of star players but Ja Morant which is on a rookie contract that pays way lower than that of other types of contracts.

We can also see that some other predictors such as the Free Throw Attempt Rate also has a positive correlation with the response which also makes sense as usually those players get paid more are those ones that have longer playing time and harder to guard thus they will get fouled more and subsequently get a higher free throw rate.

As for LASSO, most of the important features are kept as shown in the figure7, some more features that worth to mentioned are: Age, small forward position, team Toronto Raptors, mid-range shots and turnovers. For the age, SF position and turnover, all three have positive correlation to the response, it is quite intuitive that the veterans get paid more which can explain age and SF as in our dataset, most of the players of the position SF 8 are around 28. As for the turnover, it can be explained as more playing time of the

	s1
(Intercept)	-5.600213e+01
weight	.
Age	1.739583e+00
Pos_C	.
Pos_F	.
Pos_PF	.
Pos_PG	.
Pos_SF	8.339730e-01
Pos_SG	.
Tm_ATL	.
Tm_BOS	.
Tm_BRK	.
Tm_CHI	.
Tm_CHO	.
Tm_CLE	.
Tm_DAL	.
Tm_DEN	.
Tm_DET	.
Tm_GSW	3.645935e+00
Tm_HOU	1.116113e+00
Tm_IND	.

Figure 7. Part of the features selected by LASSO (salary prediction).

salary	weight	Age	Player
32742000	99.8	31	Jimmy Butler
17150000	86.2	28	Evan Fournier
33005556	95.3	30	Paul George
32700690	93.9	30	Gordon Hayward
7265485	88.9	23	Brandon Ingram
32742000	102.1	29	Kawhi Leonard
30603448	98.4	29	Khris Middleton
7830000	92.5	22	Jayson Tatum

Figure 8. SF statistics.

higher paid players.

The team TOR 9 in our dataset does contains a star player Kyle Lowry, but there are two other players especially Pascal Siakam that get paid way lower than average player of the dataset. And finally, the last feature worth to be mentioned is the mid-range shooting ability of the player, it has some negative correlation too, which can reflects the trend of modern basketball, that two pointers are becoming less important compared to three pointers and the playing style has changed. To be more precise, in nowadays NBA, players tends to shoot more and shoot from a longer range, the good shooter not only are they more popular among fans but also among the teams.

As for the elastic-net, as reported in the table, since the alpha equals to 1 and thus it is essentially just a LASSO, and unsurprisingly, it selected the same features as the reported LASSO from above.

The last model is the Adaptive LASSO, features like Three Point Attempt Rates, Three Point Shooting Percentage and the Effective Field Goal Percentage start to emerge, which provides further evidence to the analysis of the trend of the playing style mentioned above.

salary	weight	Age	Player
33296296	93.0	34	Kyle Lowry
2351838	103.0	26	Pascal Siakam
9346153	89.4	26	Fred VanVleet

Figure 9. TOR statistics.

4.2. Shooting Motion and Accuracy

Regarding the results of shooting motion vs accuracy, only those related to the common lasso will be reported, this is because all the other methods used demonstrated similar results and establish the same conditions.

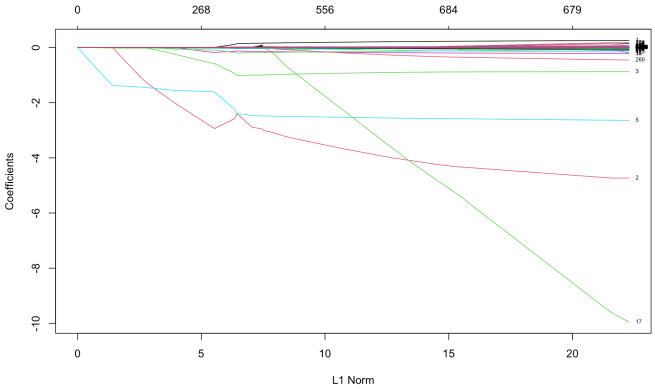

Figure 10. growth of the parameters with respect to the L1 norm. **Source:** Own creation using R.

In the figures 10 and 11 It is possible to see how the parameter that always grows at a different rate is the so-called ”rt” (release time), which indicates that it is a relevant feature for the regression. The second parameter to grow is the one related to the change of distance in the z axis, this is seen more clearly in the figure 11 in the plot of the coordinates.

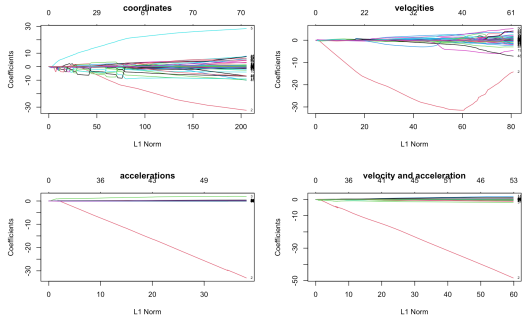

Figure 11. growth of the parameters of the 4 subsets separated. **Source:** Own creation using R.

Anyway, after calculating the mean square error using cross validation, all the algorithms proved that the number of optimal parameters to make the prediction was 0 or 1 (rt).

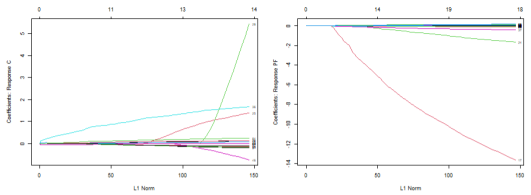

Figure 12. Growth of the parameters for C and PF with respect to the L1 norm.

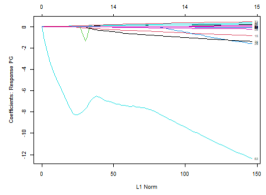

Figure 13. Growth of the parameters for PG with respect to the L1 norm.

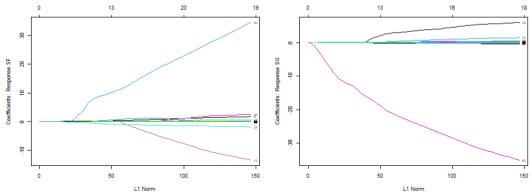

Figure 14. Growth of the parameters for SF and SG with respect to the L1 norm.

4.3. NBA Player Classification

From the plots at figures 12, 13 and 14 we can see how the coefficients for the different features changes in each position as lambda varies.

The coefficients for the relevant features for each position are shown at figure 15. From the table some interesting insights emerge. Center position is positively correlated with blocks (centers indeed tend to block more compared to other players). Point guards are positively correlated with steals, passes and assists while they have strong negative correlation with heights, which makes perfect sense since they are the playmakers of the team and they are usually shorter than the average player. The plot at figure 16 shows the selection of lambda by cross-validation and the miss classification error for models with different number of features. The minimum lambda obtained is 0.0002782559 while the 1se lambda is 0.007742637. The accuracy of the baseline model, the Random Forest, is 94,12%. The accuracy from the Multinomial Logistic regression reaches 91.53% and the details from the confusion matrix are described in figure 17.

48 x 5 sparse Matrix of class "dgCMatrix"

	C	PF	PG	SF	SG
(Intercept)	-1.52906431	18.43246427	3.552318e+00	-80.062294055	59.60657619
Age	.	.	.	.	-0.24528630
C	.	.	1.663455e-01	.	.
CS	-0.16551618	.	.	.	.
X3P	.	.	.	1.449018231	.
X3PA	.	.	.	.	0.15993022
X2P.	.	.	.	-10.037955412	.
FT	.	.	-7.137196e-01	.	.
FT.	.	-11.85146651	.	0.624454316	.
BLK	.	.	.	.	1.18115188
TOV	.	-1.29066794	.	.	0.07615503
PF	0.96981462	.	.	-1.509461065	0.40208265
TS.	1.26448553	.	.	.	.
X3PAr	.	.	-9.884922e-01	.	5.67096429
AST.	.	-0.02034517	2.975128e-01	.	.
STL.	.	.	6.422787e-02	2.085659474	.
BLK.	1.48634474	.	.	.	.
TOV.	.	-0.37187917	.	.	0.32155154
OWS	.	.	-1.180822e+00	.	0.38592141
VORP	-0.33046728	.	.	.	.
Close.Shot	0.08262982	-0.04628298	.	0.099271470	-0.10594274
Free.Throw	-0.13165947	.	.	0.247598952	.
Shot.IQ	.	-0.12351156	-2.028805e-02	.	0.13424546
Offensive.Consistency	.	.	-1.860916e-03	0.096746670	.
Speed	-0.14549494	.	.	.	.
Strength	0.04968687	.	.	-0.079729782	.
Vertical	.	0.01918656	.	-0.017392754	0.10034162
Hustle	.	0.09464307	1.918633e-01	-0.068883867	.
Overall.Durability	.	0.03583450	-1.843538e-01	0.049309485	.
Layup	-0.02844318	0.06798466	.	.	0.11771191
Driving.Dunk	.	.	-4.490835e-02	0.035558687	.
Post.Hook	.	0.01353011	.	.	.
Post.Fade	.	.	.	.	-0.09205786
Post.Control	.	0.11794562	.	.	.
Draw.Foul	.	0.12503415	.	-0.050428589	.
Hands	.	.	.	0.048719761	.
Ball.Handle	-0.01611227	.	.	.	.
Pass.Vision	.	.	4.090809e-01	.	.
Interior.Defense	.	.	-3.917251e-05	.	.
Perimeter.Defense	-0.11163994	.	.	.	.
Steal	.	.	.	0.002181408	.
Block	0.10947094	.	.	.	.
Help.Defense.IQ	.	-0.02149097	.	0.033312074	-0.11266851
Pass.Perception	.	-0.12060807	.	.	.
Defensive.Consistency	.	-0.07898824	-5.736745e-02	.	0.03435726
Defensive.Rebound	.	0.04587447	.	.	.
height	.	.	-1.093443e+01	27.527345496	-31.71808742
weight	0.20410436	.	-1.485468e-02	.	.

Figure 15. Non-zero coefficients for each position, for the minimum lambda.

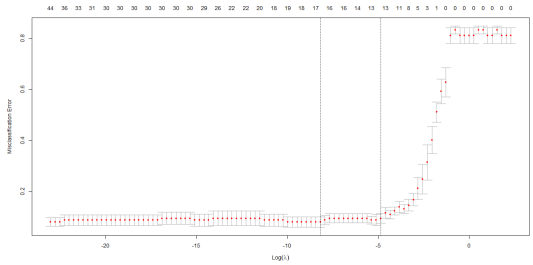

Figure 16. Non-zero coefficients for each position.

Confusion Matrix and Statistics

Prediction	Reference C	PF	PG	SF	SG
C	8	0	0	0	0
PF	0	15	0	1	1
PG	0	0	8	0	3
SF	0	0	0	11	0
SG	0	0	0	0	12

Overall Statistics

Accuracy	0.9153
95% CI	(0.8132, 0.9719)
No Information Rate	0.2712
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.8927
Mcnemar's Test P-Value	NA

Statistics by Class:

	Class: C	Class: PF	Class: PG	Class: SF	Class: SG
Sensitivity	1.0000	1.0000	1.0000	0.9167	0.7500
Specificity	1.0000	0.9545	0.9412	1.0000	1.0000
Pos Pred Value	1.0000	0.8824	0.7273	1.0000	1.0000
Neg Pred Value	1.0000	1.0000	1.0000	0.9792	0.9149
Prevalence	0.1356	0.2542	0.1356	0.2034	0.2712
Detection Rate	0.1356	0.2542	0.1356	0.1864	0.2034
Detection Prevalence	0.1356	0.2881	0.1864	0.1864	0.2034
Balanced Accuracy	1.0000	0.9773	0.9706	0.9583	0.8750

Figure 17. Confusion matrix for Multinomial Logistic Regression.

5. Conclusions

- From the prediction of player salary, we can conclude that a highly paid NBA player has the following characteristics. First, he has to be an experienced player and secondly he should has more playing time during the games and he must have the ability to shoot from long range. Furthermore, we can also conclude that the playing style has changed in NBA, as good shooter from longer range are preferred by the teams.

- With respect to the accuracy of the players compared to the shooting motion, it is possible to say that if any parameter is relevant in "the technique" this would be the time it takes them to make the shot, apart from this no other position parameter, acceleration or speed seems to be relevant to predict the accuracy of three-point shots in the season.

- The player position classification showed very interesting results. Despite the relatively small number of observations the model was able to extract the most relevant attributes to characterize each role on the basketball court. We believe that with the greater variety of metrics that a professional team possess, a model of this type could be a valid support for coaches managing their roster or for talent scouts looking for new adds to their franchise.

References

[1] Kaggle NBA 2k20 player dataset: https://www.kaggle.com/datasets/isaienkov/nba2k20-player-dataset

[2] Kaggle NBA 2020-2021 Season Player Statistics: https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18?select=nba_team_stats_00_to_21.csv

[3] NBA Player Shooting Motions: https://www.inpredictable.com/2021/01/nba-player-shooting-motions-data-dump.html

[4] Kaggle NBA Player Salary: https://www.kaggle.com/datasets/isaienkov/nba2k20-player-dataset

[5] Rpubs-imbalance: Oversampling Algorithms for Imbalanced Classification in R: https://rpubs.com/yoompubs/467234