Predicting NBA All-Stars

Xueru Xie
A15451232
x2xie@ucsd.edu

Kuo Lin
A15538652
k6lin@ucsd.edu

1. Introduction

In the sports field, players are always given various kinds of awards in each season based on their performances during the season. When it comes to the NBA, some of the players among the whole league will be voted by fan, player, media, and coaches to become star players based on their performance for each season [1]. While it seems that the player with greater performance will attract more attention and votes, we are curious about what aspect of the performance decides most why a certain player can be voted in all-stars.

Since NBA history has tens of years, and each year there are hundreds of players in the league, we have enough data to be collected. Each player's performance can be evaluated by its technical statistics during the game like score, rebound, assist, and other basketball terminologies. Whether a player is an all-star or not is just a binary value towards each player. Therefore, we perform a prediction model on this task to predict whether a player is an all-star or not based on its technical statistics.

2. Dataset

To collect all technical statistics and all-star information in NBA history, we use a Chinese website that collects all NBA technical statistics since 1946 [2]. We read all technical statistics on the website as our dataset. There are 52892 data points. Each data point represents the detailed technical statistics of a player for a certain season. The technical statistics include played games, played games as starters, minutes played per game, field goal percentage, field goals per game, field goal attempts per game, 3-point field goal percentage, 3-point field goals per game, 3-point field goal attempts per game, free throw percentage, free throws per game, free throw attempts per game, total rebounds per game, offensive rebounds per game, defensive rebounds per game, assists per game, steals per game, blocks per game, turnovers per game, personal fouls per game, points per game, wins, and losses, which are 23 features in total. Additionally, we read the all-star information for each season, except 1998-1999 since there is no all-star games due to the NBA lockout [3], on the website to add one more feature to these 52892 data points – the feature of whether the player is an all-star for that season.

During the process of reading data on the website, there are some reading errors like wrong names possibly caused by the website author. It is quite important to fix these errors to make analysis and prediction task succeed. Furthermore, the percentages are changed from percentage form to decimal form to make it easy to handle.

Before we move on to the exploratory analysis, we have to take a look at the missing values. Many players have missing data because they simply do not perform a certain action. For example, a player could have an empty statistic for his 3-point field goal percentage because he does not shoot any 3-pointers in a season. Some players have missing data because there are no records of his statistics. For example, the blocks per game statistic is not added until the 1973-74 season [4]. Thus, players before this season would have empty blocks per game value. In literal sense, missing data is different from 0 value data because one does not perform the action and the other does. However, in this study, we will treat both types of data the same way. We will fill in the missing values as 0 values because that would be the only option. Dropping missing data would not be a good option because there are too many players with missing data. In addition, some all-star players have missing statistics, so dropping those players would lead to inaccuracy of our result.

2.1 Univariate Analysis

To analyze the data itself for each single feature, it seems that most features have a distribution of skewed right (Figure 1). The possible reason for this is that most players in the NBA have few technical statistics since NBA games are mostly the battlegrounds for those star players.

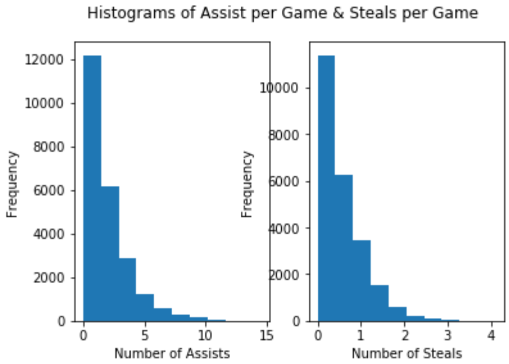


Figure 1 Histograms of two statistics.

Each game has limited amounts of time for players to play the game, so there are lots of players not able to play the game and play as backups on the team. Therefore, there are also lots of outliers for each feature (Figure 2).

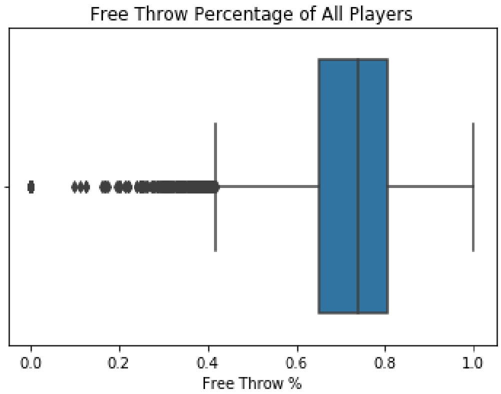
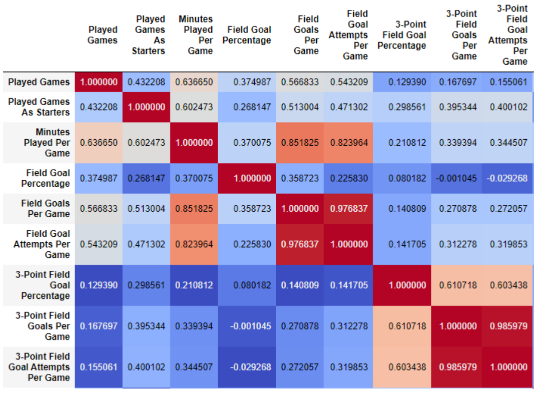


Figure 2 Boxplot of free throw percentage

In order to solve the rightly skewed and outliers' problems, we will be scaling the data by dividing by the maximum value. This method would stabilize our inputs because all the numbers are proportions instead of raw values.

2.2 Bivariate Analysis

Next, we want to investigate the relationship between the statistics to see if there are any possible correlations. Before we conduct a correlation matrix, we suspect that there are certain correlations between the statistics. For example, there should be a relationship between shot attempts per game and shot percentage (Figure 3). Here, shot attempts could be a 2-pointer or a 3-pointer. If a player keeps firing shots on the court, he will definitely make more shots than a player with fewer shot attempts because there are more samples for him.



	Played Games	Played Games As Starters	Minutes Played Per Game	Field Goal Percentage	Field Goals Per Game	Field Goal Attempts Per Game	3-Point Field Goal Percentage	3-Point Field Goals Per Game	3-Point Field Goal Attempts Per Game
Played Games	1.000000	0.432208	0.636650	0.374987	0.566833	0.543209	0.129390	0.167697	0.155061
Played Games As Starters	0.432208	1.000000	0.602473	0.268147	0.513004	0.471302	0.298561	0.395344	0.400102
Minutes Played Per Game	0.636650	0.602473	1.000000	0.370075	0.851825	0.823964	0.210812	0.339394	0.344507
Field Goal Percentage	0.374987	0.268147	0.370075	1.000000	0.358723	0.225830	0.080182	-0.001045	-0.029268
Field Goals Per Game	0.566833	0.513004	0.851825	0.358723	1.000000	0.976837	0.140809	0.270878	0.272057
Field Goal Attempts Per Game	0.543209	0.471302	0.823964	0.225830	0.976837	1.000000	0.141705	0.312278	0.319853
3-Point Field Goal Percentage	0.129390	0.298561	0.210812	0.080182	0.140809	0.141705	1.000000	0.610718	0.603438
3-Point Field Goals Per Game	0.167697	0.395344	0.339394	-0.001045	0.270878	0.312278	0.610718	1.000000	0.985979
3-Point Field Goal Attempts Per Game	0.155061	0.400102	0.344507	-0.029268	0.272057	0.319853	0.603438	0.985979	1.000000

Figure 3 Correlation heatmap of some columns

Additionally, we hypothesize that more shot attempts and more field goals made could relate to the chance of becoming an all-star because only star players could have the confidence and the infinite shooting privilege (Figure 4). Looking at figure 4, our assumption is not wrong because most of the all-star players have high field goal attempts per game and high field goals per game. However, there are a few exceptions based on the visualization. We believe that these outliers occur because fans and the organization do not evaluate the player as an all-star even though he scores or attempts to score a lot of points.

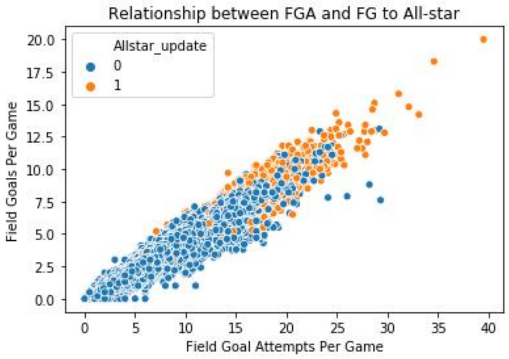


Figure 4 Scatterplot of field goal attempts and field goals with the option of all-star

Besides looking at positive data of a player. We want to investigate the opposite. Statistics like turnovers per game and personal fouls per game are the examples that we are considering. As we can see, more or fewer personal fouls per game do not make a player an all-star (Figure 5). When a player has too many personal fouls, that means he is not great at preventing the opponent from scoring. On the opposite, fewer personal fouls do not make an excellent player because the player may not have enough play time to foul the opponent, or the player simply does not confront the opponent in defense. This similar distribution occurs for turnovers per game as well (Figure 6).

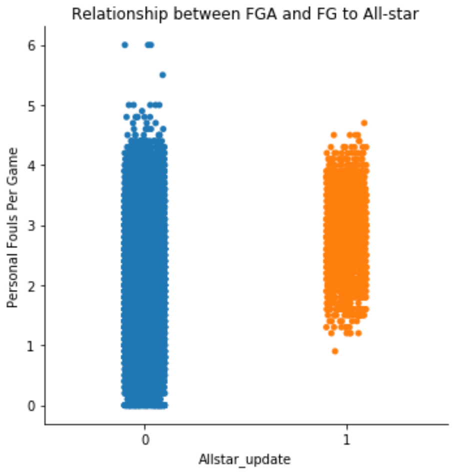


Figure 5 Catplot of fouls per game and all-star

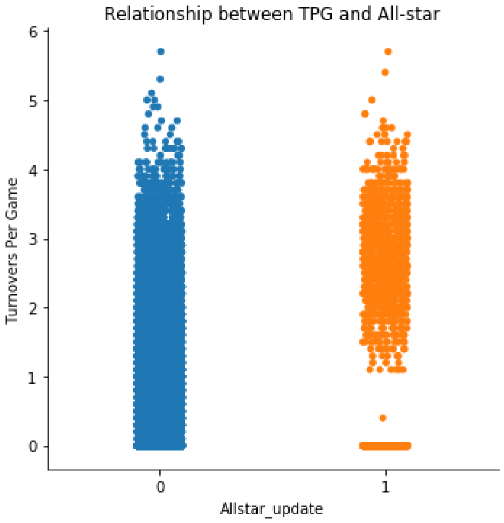


Figure 6 Catplot of turnovers per game and all-star

3. Predictive Task and Modeling

Our predictive task will be trying to predict whether an NBA player is selected as an all-star player, given his seasonal statistics. Since we are evaluating a player's seasonal statistics, we are using those data to predict whether the player is qualified as an all-star in that particular season.

Predicting whether a player is an all-star is a binary problem. So, the model will either output two values, 1 being an all-star and 0 is not. In order to evaluate our model, the best method would be the confusion matrix. The confusion matrix would show the true positive, true negative, false positive, and false negatives.

Our goal of this task is to predict the all-stars as accurate as possible. That means our maximum accuracy score should be 100%, theoretically. However, we need to set the minimum accuracy score to assess the validity of our model's predictions. We believe that the choice of the accuracy baseline is subjective because there is no minimum standard baseline score for this problem. Both of us are very familiar with the NBA, so we will set the baseline based on our NBA knowledge and viewing experience. We believe that players that score average 15 points per game in a season should be qualify as all-star players. Although this method is not accurate since there are many players that do not become the all-star with 15+ points per game, this is only a baseline assumption. It would be "good enough" since most of the all-stars average 15+ points per game in a season.

Before formally making our model and begin prediction tests, we need to compute the baseline accuracy, by means of setting the points threshold at 15 points per game in a season as illustrated above. First, we split the data set randomly such that 75% becomes training set and 25% becomes testing set. For each set, whether a player is an all-star or not becomes the label, and columns other than all-star information become the features. Then to compute the baseline accuracy, we check whether the feature of points per game for each player is above 15 or not. Finally, we compare our baseline result to the actual label of the test data, giving that the accuracy of the baseline is approximately 88.74%. Our model should perform better than the baseline because the baseline is only a vague generalization of the all-star players.

Now that our baseline is defined and the assessed, we will start choosing the predictive model. Since it is a classification task with the label either 1 or 0, we will not be using any regression models since they will not appropriately fit. In this study, we can use logistic regression and support vector machines (SVM) as our models. We choose these two models because they are excellent at classifying binary problems. Logistics regression can train the sigmoid function to maximize the product of features and theta when the label is positive and maximize it when the label is negative. SVM is able to separate compare different classifiers and seek the one furthest from the data points, in which way it can classify all points more correctly and penalize misclassified data points by moderating the distance to the data points. Our goal of using two models is to compare the results for both and select the best model that has the higher accuracy. For model optimization, we will be adjusting model parameters to somewhat avoid overfitting, scaling some/all the data to normalize and standardize the data, and modifying features to try to emphasize the effects of how data can have on the result. While we are optimizing the models and trying to find the max accuracy for both models, we really do a lot of wasted works since they are not improving the accuracy. But we are still beneficial at knowing how the model optimization and dataset modification can affect the model accuracy. To understand our model performance deeply, jump to the result section for further analyzing.

4. Literature

Although we look for data by ourselves on the statistics website [2] to do the research, there are also many similar researches on predicting NBA all-stars using technical statistics on the internet. On the internet, there are many kinds of websites that collect NBA statistics other than what we use. Examples are like the official NBA statistics website [5] and Basketball Reference website [6]. While we regard this prediction task as classification task, there are also research that consider it as a machine learning task [7]. The article uses “XGBoost, an implementation of gradient boosted decision trees” to build the model of prediction. It will build up a decision tree via learning the dataset and fit the dataset into the tree (Figure 7). While predicting, the tree will generate the new branches as predict results based on previous learnings. This kind of learning process is really popular in the current era.

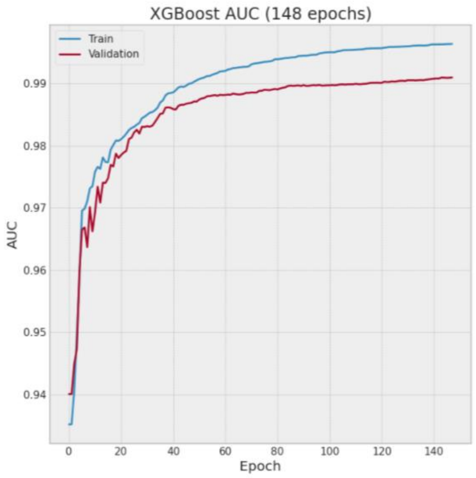


Figure 7 Graph of decision tree learning process picked from [7]

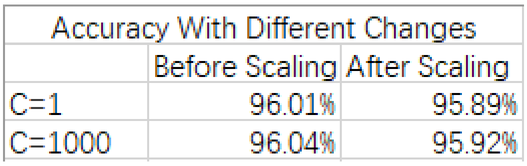
But one issue of this method is that when predicting, it needs to beat some threshold set by the author to evaluate a positive outcome for all-star prediction. The selection of the threshold decides the performance of the model. Instead, logistic regression model has the advantage of using only dataset to do the prediction task. Similar to our report, there are also researches comparing all-star and non-all-star seasonal technical statistics to see the different performance and the reasons of that between all-stars and non-all-stars [9]. In the report, authors also use players’ technical statistics as database and simply compare those using methods like ANOVA to try to find similarities and differences to do some predictions. Thus, actually our work is somewhat beyond this by including the prediction model to do it as a classification task. Despite of all these, the decision tree, which the machine learning author uses as the state-of-the-art method to do the prediction task, can perform better with accuracy of 97.5% than our method of logistic regression with accuracy of 96.0%. While the condition is to have selected a great threshold, the decision tree is still one advanced and efficient method of accomplishing prediction tasks nowadays.

Other than the modeling selection of the prediction task, we also learned to process our data furthermore from some researches. As we discuss early, there are lots of other features that can influence or reflected from the fact that a player is selected as an all-star. One feature that we do not include/have is the player’s salary. This feature could reveal a lot about a player because all-star players usually get to paid more than other players [10]. However, this does not mean that players with the most salaries must be predicted as all-stars. There are players with low salary and can be selected as an all-star. Other features mentioned in the article like age/experience, draft position, or height/weight can all be applied to our model in this report. Whether the features of dataset are inclusive and correlated enough or not is one important factor to the success of the prediction task.

5. Results and Conclusions

5.1 Logistic Regression Model

First, we use logistic regression with a normal regularization parameter to fit the train set and predict test set, giving that the accuracy reaches 96.01% (Table 1). This is already a high score for prediction test, while we can still make some changes to see if we can further optimize it. For adjusting model parameters, we tune the regularization parameter, the accuracy can improve by 0.01% or so, which is not a significant improvement, suggesting that tuning regularization may not be much useful (Table 1). Since Figure 2 suggests that there are some outliers in the feature data, we can further scale the data. For scaling of the data, we divide features of data points by their maximum values to stabilize our inputs. But the accuracy turns around 95.90% after scaling with different regularization parameters used (Table 1). The reason why the accuracy is lowered may be that data points are not that large and will become meaningless after being scaled to proportion, which affects the prediction.

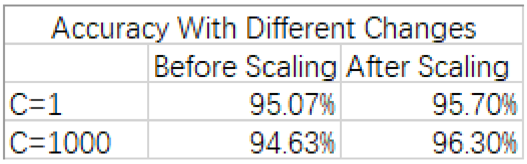


Accuracy With Different Changes		
	Before Scaling	After Scaling
C=1	96.01%	95.89%
C=1000	96.04%	95.92%

Table 1 Accuracies after various changes for model logistic regression

To further optimize the model, we also try modifying features of dataset. Modifying features would mean adding or subtracting statistics. An example of subtracting a feature would be to remove turnovers per game in the data. In Figure 6, we conclude that this statistic does not reveal too much information about a player being an all-star, so it could be beneficial to the model if we remove it. In terms of adding new features, there are plenty of options depending on the prediction result. If our result is many false negatives, that means many all-stars are not being predicted, then we should consider adding some features like player of the week/month, team of the year, MVP award, and more because these statistics tend to contain all-star players. Finally, we build our new dataset considering the correlations between whether a player is an all-star or not and all other features (Figure 3). Among all 23 correlations with the all-star information, we choose the median and pick those features who have the higher correlation than median as our new X features, since median can eliminate those features with small or negative correlations. While fitting the logistic regression model using the X with new features, the accuracy turns into around 95.35%, which is a further decrease. The possible reason may be the process of selecting features with high correlations eliminate features with low or negative correlations which may be actually useful for tuning the model and actually avoiding overfitting.

5.2 SVM Model



Accuracy With Different Changes		
	Before Scaling	After Scaling
C=1	95.07%	95.70%
C=1000	94.63%	96.30%

Table 2 Accuracies after various change for model SVM

Next, we can change to use SVM model to see if it can generate a better accuracy. The order of using different kinds of dataset is just the same as we use logistic regression model. It seems that SVM model does not perform really better than logistic regression with the same dataset are inputted (Table 2). Only when the regularization parameter is tuned much high to the scaling data can generate a higher accuracy than what the logistic regression model can generate (Table 2). But the increment of regularization parameter may also lead to the issue of overfitting, so maybe we cannot say SVM model performs better than logistic regression model at all.

5.3 Summary

The reason why SVM model performs worse than logistic regression for this dataset can be multiple. The number of kinds of NBA technical statistics are just too many leading to too many features for each data point, and the differences of statistics between each other can be very close, so data points here are not too far from each other and instead very concentrated, except some greatest players. Therefore, SVM model may not perform well since it is better at choosing a classifier that is furthest from data points. On the other hand, logistic regression does not have too many restrictions and can perform well on most classification tasks like this one.

In conclusion, we are satisfied with our model since we reach a high accuracy of above 95%. But we also discover that we still have a lot to learn after researching many other reports, we. In fact, for prediction task, it is always great to have plenty of datapoints like millions to be split by the train, test, and validation and predict much precisely. While NBA is an excellent basketball league overall the world, it only gathers no more than hundreds of people for each season, which cannot generate millions of datapoints for us to train our model. It is also not possible for model in this task to predict some randomly generated data since we will not have ways of validating our predictions. Despite all, we have at least accomplished a prediction model with high accuracy at encouraging more and more players to improve themselves and become all-stars.

6. Reference

[1] Wikipedia. *NBA All-Star Game.* https://en.wikipedia.org/wiki/NBA_All-Star_Game

[2] http://www.stat-nba.com/index.php

[3] The Shadow League. (2019). *The 20th Anniversary Of The NBA All-Star Game That Never Was.* https://theshadowleague.com/the-20th-anniversary-of-the-nba-all-star-game-that-never-was/

[4] Wikipedia. *Block (basketball).* https://en.wikipedia.org/wiki/Block_(basketball)

[5] https://www.nba.com/stats/

[6] https://www.basketball-reference.com/

[7] Porteous, C. (2020). *Using machine learning to predict NBA All-Stars, Part 2: Modelling.* Towards Data Science. https://towardsdatascience.com/using-machine-learning-to-predict-nba-all-stars-part-2-modelling-a66e6b534998

[8] Wikipedia. *Decision tree.* https://en.wikipedia.org/wiki/Decision_tree#:~:text=A%20decision%20tree%20is%20a%20flowchart%2Dlike%20structure%20in%20which,taken%20after%20computing%20all%20attributes).

[9] Sampaio, J. & McGarry, T. (2015). *Exploring Game Performance in the National Basketball Association Using Player Tracking Data.* PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132894#sec008

[10] Marks, B. & Off, G. (2019). *The Making of a Modern All-Star: What 28 years of data says about NBA's best players*. The Charlotte Observer. https://www.charlotteobserver.com/sports/charlotte-hornets/article226224515.html