Catch Shooting Stars

Predicting NBA All-Stars from Rookie Season Performance

Chongqi Wu*
California State University, East Bay, Hayward, California, USA

Hongwei Du
California State University, East Bay, Hayward, California, USA

This work employs multiple machine learning algorithms to predict whether an NBA player will become an All-Star in his career based on his rookie season performance. The best performing model aggregates all the algorithms used, which generates an overall accuracy of 91.39%. Based on this aggregate model, we predict who will become All-Star among players drafted from 2013 to 2015.

* Corresponding Author. E-mail address: chongqi.wu@csueastbay.edu

I. INTRODUCTION

When an NBA team drafts a player, it is largely based on his talent and potential. Talent and potential, however, do not always predict the future success of a player. It occurs all the time that a high draft pick did not pan out as projected. Even with significant effort and manpower invested in talent evaluation and scouting, the list of busted high draft picks in the NBA goes on and on: Hasheem Thabeet, Marvin Williams, Shawn Bradley, Jay Williams, Adam Morrison, Kwame Brown, Greg Oden, Michael Olowokandi, Darko Milicic, and of course, Sam Bowie, best known as the man drafted before Michael Jordan. The success of a player depends on much more than talent. Factors such as effort and dedication, injury, off-the-court issues, coaches are essential. To be fair, it is extremely challenging and difficult for anyone to properly measure and evaluate factors like effort and dedication, potential to injury. It is thus not surprising that we see so many busts.

The difficulty in predicting the future of a player does not prevent teams from doing so. When a team drafts and keeps a player, the team is investing in this player with the belief that he will be valuable for the team. If we can predict hit-or-flop of a player relatively more accurately, it means tremendously for a team. NBA teams invest millions of dollars on the team roster. Accurate projection on players' future implies that teams can identify diamond in the rough and sign them relatively cheap. On the other hand, it helps teams stay away from over-hyped and over-priced players. Obviously, the earlier such accurate predictions can be made, the better off the teams will be. In this study, we use NBA players' rookie season performance and build machine learning models to project whether their future career is a success. To measure a player's performance in his rookie season, we have adopted typical basketball statistics of a season, including points per game, rebounds per game, assists per game, minutes per game, number of games played, among others. How to define success of a player in NBA is another challenge. To keep it simple and meaningful, we consider whether being an All-Star as the measurement of career success. Another often used success measurement is the longevity of a player in the league. One may argue that the dichotomous All-Star-or-bust measurement oversimplifies. On one hand, there are plenty of successful players who had never made it to All-Star, such as Robert Horry, Lamar Odom, Monta Ellis, and current head coach of Golden State Warriors Steve Kerr. On the other hand, there are All-Star players, who, according to some, do not deserve the accolade, such as B.J. Armstrong, James Donaldson, Dale Davis, and Christian Laettner. By no means are these players bad players. Indeed, these players had decent careers. Sometimes, they may just fall short of the expectation. Christian Laettner is a classic example, who were on Dream Team I with Michael Jordan, Charles Barkely, Magic Johnson, Larry Bird, etc. The expectation was for him to be like his teammates on the original Dream Team. Apparently, he fell short. Looking back his 13-year career with 12.8 points per game and 6.7 rebounds per game, there is nothing to be ashamed of. Almost all All-Stars are at least good, if not outstanding or great players.

Overall, there are just less than 15% of NBA players who have ever been All-Star. From the perspective of an NBA general manager, the false positive (a player did not become All-Star but was predicted so) is more acceptable as long as he turns out to be a decent player. False negative (a player did become an All-Star but the model failed to predict so) is a much more serious issue. Our model is likely to miss players like that. To tackle this issue along with the issue of hit-or-miss dichotomy, we consider five machine learning algorithms and a total of 30 models. We draw our conclusions aggregately. To be more specific, we count the number of models who predict a player will be an All-Star. In a dichotomous sense, we predict that a player would be an All-Star if 26 or more models predict so, even though the most accurate prediction is to conclude a player would be All-Star given all 30 models predict so. However, a more meaningful way of using our prediction is to rank all players based on the number of models that predict them to be All-Star. We call this number the number of votes a player received from our prediction model. Even if some rookie players missed the threshold of 26 votes, they could turn out to be good or even great players. By ranking all players based on the votes they received from our model, we can more effectively control the damage of false negative. Kobe Bryant had an unassuming rookie season. He was not predicted to be an All-Star based on the threshold of 26 votes. But still, 23 out of our 30 models predicted that he would be All-Star.

To the best of our knowledge, our work is the first of the kind to predict a professional basketball player’s career based on his rookie season performance. Overall, our models have achieved admirable accuracies at 91.39%. Based on our models, we also boldly project which players drafted between 2013 and 2015 will likely be All-Star.

II. RELATED LITERATURE

The application of quantitative methods, including machine learning and analytics algorithms to NBA and many other sports is prevalent nowadays. Berri (1999) developed an econometric model that links the player’s statistics in the NBA to team wins and measures each player’s marginal product or contribution to team wins. Sampario et al. (2007) applied discriminant analysis to examine the differences in game-related statistics between basketball guards, forwards and centers. Sill (2010) employed regularization and ridge regression to enhance the accuracy of popular NBA player evaluation technique “Adjusted +/- (APM)”. Cooper et al. (2009) evaluated effectiveness of basketball players with DEA. Similarly, Moreno and Lozano (2014) used a Network DEA approach to assess the efficiency of NBA teams. Page et al. (2007) studied the relationship of skill performance by position and game outcome with a hierarchical Bayesian model and examined position characteristics parameters’ relative importance to game outcome.

Some studies question the validity of quantitative research in NBA and in sports in general. Martinez and Martinez (2011) argued that the existing player valuation systems are deficient because they fail to rate intangibles and that qualitative thinking is prominent and should be considered in valuating such intangibles. It is true that quantitative methods have their limitations in their applications to sports, particularly with respect to “intangibles”. But with more and more data becoming available, we strongly believe that quantitative methods will be performing better and better, and “intangibles” will become more and more tangible.

As far as content is concerned, Coates and Oguntimein (2010) is the closest to our study. They examined NBA players drafted in 1987 through 1989 and studied how their collegiate performance affected their draft positions and the longevity of their professional career. They have found that some college productivity significantly influences draft position and draft position also affects the length of a player’s career, with earlier draftees having longer careers. In our work, we use players’ rookie year performance to predict their career success, defined as whether they become All-Star.

Another important and relevant work is Stiroh (2006), which showed that individual performance improves significantly in the year before signing a multi-year contract but declines after the contract is signed. This supports our choice of rookie season performance which is free from the contract-year effect.

III. ANALYSES AND RESULTS

3.1. Data Preparation

The NBA has its first All-Star game in 1951. Due to data availability and quality, we collected information on 2763 players, starting from 1952 NBA draft. This list does not include undrafted free agents. Nor does it include players drafted but failed to play or to make team roster for one reason or another. Among the undrafted free agents, three players have become All-Star thus far. They are John Starks, Brad Miller and Ben Wallace. However, vast majority of undrafted free agents have barely played a few games in the league. The data on them is sporadic and of poor quality. We do not think it would be appropriate to include only those three undrafted free agents while ignoring all the others. Thus, all undrafted free agents are excluded from this study.

Out of this list, there were 164 players who played 5 or less games in their rookie season. Their statistics are statistically insignificant, and thus excluded from this study. Moreover, an All-Star player played at least 6 games in his rookie season. There are two All-Star players who played six games in their rookie season: Charlie Scott and Michael Redd.

Among the remaining 2599 players, 126 players were drafted in 2013 – 2015. As of this writing, none of them have become All-Star. They still have plenty of time to prove themselves. Therefore, we exclude them from training and validation dataset. Instead, after we train our predictive models, we will predict, among them, who will be All-Star in the future.

Of the remaining 2473 players, 356 players have been All-Star, whereas the other 2117 are non-All-Star players. The ratio of All-Star players is 14.4%. This is a highly imbalanced dataset. To handle this imbalanced dataset, we first divide it into training data and validation data. In the validation dataset, there are 106 All-Star players and 317 non-All-Star players, all of whom are randomly selected from the original dataset of 2473 players. Then with the remaining 250 All-Star players and 1800 non-All-Star players, we create 6 training datasets. All those 250 All-Star players are always included in any of the 6 training datasets. On the other hand, in each training set, there are 300 non-All-Star players, exclusively and randomly selected from the pool of 1800. This is a standard approach to handle imbalanced datasets. There are 45.5% All-Star players in each training set, which significantly reduces the impact of imbalanced data.

With 6 training sets, we will have 6 models for each algorithm. The final prediction result will be based on all these six predictions, one from each model. Thus, ensemble method is used. Let n be the number of models that predict a player will be an All-Star and k be the threshold number of votes for All-Star. If n >= k, we then make the prediction that this player will be an All-Star. We can choose different k values from 4-6 and evaluate which choice generates the most accurate prediction.

3.2. Feature Selection and Algorithms

We decide to include 23 features in our study, most of which are typical in quantitative studies of basketball players, such as GP (games played in a season), MPG (minutes per game), and PPG (points per game). The complete list of all 23 features is in Appendix A.

The obvious exclusions from the list are 3-point related statistics, SPG (steals per game), and BPG (blocks per game). Our data starts with 1952 NBA draft. Three-point line was not adopted in the NBA until the 1979-80 season. SPG and BPG data had not been officially collected until the 1973-1974 season. Similarly, the old NBA statistics did not differentiate offensive and defensive rebounds until 1970s. Therefore, only RPG (rebounds per game) is included in our analysis.

In many studies, the position at which a player was drafted is frequently used as a feature. We restrict our attention to game-related data, and thus excluding it from our study.

In the end, we also include the square term of each feature as independent or explanatory variables in our models. As a result, there are a total of 46 independent variables in our model. Feature selection is more of an art than science. The square terms are one of the most often used non-linear terms in machine learning. In addition, the rule of thumb in machine learning is that the sample size should be at least 10 times of the number of independent variables. The purpose is to mitigate the overfitting problem. In each of our training set, we have 550 data points. A choice of 46 explanatory variables is in line with the aforementioned rule of thumb. Our results indicate that overfitting is largely under control in our models.

In our study, we have employed five algorithms: k-nearest neighbors (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LogReg), Random Forest (RF), and linear Support Vector Machine (SVM). In KNN, we choose k to be 10. Actually, the accuracies of KNN are very stable regardless of the choice of k. Part of the IPython script used to generate the results can be found in Appendix E.

3.3. Accuracies of Individual Algorithm

In our study, we consider both in-sample accuracy and out-of-sample accuracy. The former is the accuracy with respect to the training dataset and the latter is with respect to the validation dataset. Recall that we have 6 training sets, which result in 6 different models for each algorithm. The accuracy of each algorithm is thus the average accuracy of all 6 models under the same algorithm. Table 1 below summarizes both in-sample and out-of-sample accuracies of the five algorithms used.

Table 1 shows that the performance of all algorithms is rather consistent between in-sample and out-of-sample accuracy except RF. Apparently, overfitting occurs with RF. RF's in-sample accuracy is significantly higher than that of any other algorithm and its own out-of-sample accuracy. Indeed, RF has the worst out-of-sample accuracy among the five algorithms, whereas GNB has the best out-of-sample accuracy. Fortunately, there is no marked difference among all five out-of-sample accuracies. In two occurrences (LogReg and GNB), the out-of-sample accuracy is even higher than in-sample accuracy. This occurs because the percentage of All-Star players in the validation set (25.1%) is much lower than that of training set (45.5%).

Next, we investigate the aggregate out-of-sample accuracy of each algorithm. There are 6 models resulting from 6 training sets under each algorithm. Let k be the threshold value, ranging from 4 to 6. If there are k or more models predicting a player to be an All-Star, then the aggregate model of this algorithm or the ensemble method will predict so; otherwise, the aggregate model will predict that he will not be All-Star. Table 2 below summarizes the out-of-sample accuracy of each algorithm based on different threshold values.

TABLE 1. ALGORITHM ACCURACIES.

Algorithm	Accuracy	
	In-Sample Accuracy	*Out-Of-Sample Accuracy*
LogReg	0.7753	0.7872
KNN	0.8003	0.7979
GNB	0.7473	0.8258
RF	0.9827	0.7778
SVM	0.7967	0.7928

TABLE 2. AGGREGATE OUT-OF-SAMPLE ACCURACY.

Algorithm	k = 4	k = 5	k = 6
LogReg	0.7920	0.8038	0.8203
KNN	0.8109	0.8180	0.8251
GNB	0.8251	0.8298	0.8274
RF	0.8156	0.8369	0.8392
SVM	0.8014	0.8132	0.8203

TABLE 3. OVERALL ACCURACY OF AGGREGATE MODEL.

k	Overall Accuracy
15	0.8184
16	0.8269
17	0.8326
18	0.8383
19	0.8439
20	0.8476
21	0.8500
22	0.8544
23	0.8613
24	0.8706
25	0.8694
26	0.8751
27	0.8852
28	0.8965
29	0.9034
30	0.9139

With ensemble methods, the difference in accuracy among five algorithms is even narrowed. In fact, RF becomes the most accurate algorithm, albeit with very small margin. As the threshold value increases, almost all algorithms have better accuracy with the exception of GNB. The accuracy of GNB peaks at k = 5. This is mainly because there are much fewer All-Star than non All-Star players. When we increase the threshold value, some actual All-Stars will be predicted to be non All-Stars (false negative). In the meantime, we are able to rule out more non All-Star players who were originally predicted as All-Star (false positive). In other words, false negative increases a little but false positive decreases more. Arguably a more important observation is that after aggregation, the accuracy improves for all algorithms. This leads us to the aggregation of all 30 models, 6 for each of the five algorithms, which will be discussed in greater details later on.

3.4. Aggregate Model

Earlier, we notice that ensemble method improves prediction accuracy for each algorithm. We thus aggregate all 30 models across all five algorithms. The aggregate model here is similar to ensemble learning. We call our model the aggregate model rather than an ensemble learning method because it combines predictions from five different learning algorithms and it ranks all players instead of just making a binary prediction or classification.

We examine the overall accuracy of this aggregate model with all 2473 players. Table 3 below summarizes the overall accuracy of our aggregate model, depending on the choice of the threshold value k. Note that there are 30 models in the aggregate model. The threshold value, k, ranges from 15 to 30. If k or more models predicts that a player will be an All-Star, then the aggregate model will predict so; otherwise, the aggregate model will predict that this player will not be an All-Star.

Once again, due to the fact that there are much fewer All-Star players, the overall accuracy improves as the threshold value k increases. When k is equal to 30, the overall accuracy peaks at 91.39%. In other words, the best accuracy is achieved when we predict All-Star if and only if all 30 models predict so. This fact is also reflected in the confusion matrices for k = 25 through 30.

In the confusion matrix for k = 30, there are only 7 false positive and 206 false negative. When k is 25, there are 173 false positive and 150 false negative. As we lower the bar for All-Star prediction, false negative decreases. But it is not sufficient to compensate for the increase in false positive. False positive goes up from 7 to 173. Note that when k = 26, our model predicts there are 351 All-Star players, which is the closest to the actual number of All-Star players of 356.

3.5. False Positives, False Negatives and Predictions

In the Appendix B, we list all the false positives for k = 25. That is, we list all the non All-Star players who receive 25 or more votes from our 30 models. One vote means that one model predicts this player to be All-Star. Recall that n is the number of votes a player received. Even though these 172 players have not been All-Star, most of them were good, if not great players. Among players drafted between 1995 and 2009, Joe Smith, Marcus Camby, Marc Jackson, Bobby Jackson, Derek Anderson, Keith Van Horn, Jason Williams, Andre Miller, Lamar Odom, Shane Battier, Jason Richardson, Luis Scola, Kirk Hinrich, Ben Gordon, Emeka Okafor, Andrew Bogut, Mario Chalmers, Eric Gordon are all very good players. More recently, players like Greg Monroe, Kemba Walker and Bradley Beal still have very good chance to be All-Star in the next few years. It is not All-Star or bust. Most teams would love to have these players. Our prediction model provides the benefit of identifying potentially good and very good players even if they will not be All-Star in their career.

There are other reasons why some players did well in their rookie season but failed to materialize their potential down the road. In some cases, they had to end their career pre-maturely. Injury is one of the most important reasons. Jay Williams is probably the most well-known case who ended his career after his stellar rookie season when he was severely injured in a motorcycle accident. Some off-the-court issues may also dramatically hinder a player to be a star, such as legal troubles and drug problems. Richard Dumas and O.J. Mayo come to mind.

TABLE 4. CONFUSION MATRICES.

k = 30	Predicted to be All-Star	Predicted to be Non All-Star	
All-Star	150	206 (false negative)	356
Non All-Star	7 (false positive)	2110	2117
	157	2316	2473

k = 29	Predicted to be All-Star	Predicted to be Non All-Star	
All-Star	171	185	356
Non All-Star	54	2063	2117
	225	2248	2473

k = 28	Predicted to be All-Star	Predicted to be Non All-Star	
All-Star	186	170	356
Non All-Star	86	2031	2117
	272	2201	2473

k = 27	Predicted to be All-Star	Predicted to be Non All-Star	
All-Star	192	164	356
Non All-Star	120	1997	2117
	312	2161	2473

k = 26	Predicted to be All-Star	Predicted to be Non All-Star	
All-Star	199	157	356
Non All-Star	152	1965	2117
	351	2122	2473

k = 25	Predicted to be All-Star	Predicted to be Non All-Star	
All-Star	206	150	356
Non All-Star	173	1944	2117
	379	2094	2473

In Appendix C, we list most obvious false negatives for k = 10. That is, these players have become All-Star or even some of the greatest players in the NBA history while receiving no more than 10 votes from our models. The biggest miss is Steve Nash. The two-time MVP received zero vote. Some other obvious misses include Jermaine O'Neal, Michael Redd, Mo Williams, Joe Johnson, Devin Harris, Rashard Lewis, Gerald Wallace, Jimmy Butler, Andrew Bynum, Jeff Teague, Draymond Green, Dirk Nowitzki, David West, Paul George, Manu Ginobili, Zach Randolph, David Lee, Kyle Lowry, Peja Stojakovic, Tracy McGrady, DeMar DeRozan, and Kawhi Leonard. Shortened seasons due to strike (1998-99 and 2011-12) may skew our data which leads to misclassification. Players affected by shortened regular reason include Dirk Nowitzki, Rashard Lewis, Jimmy Butler and Kawhi Leonard. Injuries, coach's decision and many other factors may limit the number of games and minutes per game of a player. This also contributes to misclassification. For example, Charlie Scott and Michael Redd, both All-Star, played only 6 games in their respective rookie season. Kyle Lowry played only 10 and Rashard Lewis played 20. In most quantitative study of NBA players, it is typical to only consider players who played 50 or more games in a season.

In the end, we make predictions on players drafted between 2013 and 2015. None of them have been All-Star as of this writing. It provides very good opportunity to validate our model and prediction. The prediction can be found in Appendix D, which lists all the players who received 5 or more votes from our aggregate model. Victor Oladipo, Michael Carter-Williams, Andrew Wiggins, and Jahlil Okafor received 30 votes. Nilola Jokic and Karl-Anthony Towns received 29 votes. Other players who receive more than 25 votes are Jordan Clarkson, Kristaps Porzingis, Elfrid Payton, Emmanuel Mudiay, Devin Booker and D'Angelo Russell.

VI. SUMMARY

Our work applies multiple machine learning algorithms to predict career success of an NBA player defined as being All-Star based on his rookie season performance. Specifically, the machine learning algorithms used in this study are k-nearest neighbors, Gaussian Naïve Bayes, Logistic Regression, Random Forest and Support Vector Machine. We divide our dataset into 6 training sets and 1 validation set. All these algorithms yield similar and comparable out-of-sample accuracies. We have found that ensemble method improves our prediction accuracy across all five algorithms. We then aggregate all our 30 models as our final and predictive model. In this aggregate model, we use the n value, the number of All-Star votes a player received from our 30 models, to make final prediction whether a player will be All-Star. The threshold value is k. If n is greater than or equal to k, then our aggregate model will predict that he will be All-Star. With k = 30, the aggregate model achieves the highest overall accuracy of 91.39%.

We briefly analyze the reasons for false positive and false negative. These reasons include injuries, off-the-court issues, and shortened schedules, among others. These reasons may result in limited playing time, limited number of games played, or poor performance, thus misclassification.

We argue that false positive is more acceptable because the actual career of a basketball player is not dichotomous: All-Star or bust. Many non All-Star players have had very productive career. If our model predicts them to be All-Star (false positive), it is not the end of the world. Indeed, it provides an opportunity to identify diamond in the rough by listing all players who receive reasonable number of votes from our 30 models. On the other hand, false negative is far worse an issue. Our model does a reasonably good job in reducing false negative. Only a handful of All-Star players received 5 or less votes from our model.

We strongly believe that there is great future in applying advanced machine learning and analytics to the field of sports. 91.39% is by no means the best possible result. In our study, we have only considered one way of handling our imbalanced dataset. It may be worth investigating other methods such as resampling. Feature selection is another area for further improvement. Typically, the availability of data limits what we can do. Even so, there are many different ways of modeling non-linearity. Another possibility to improve our model and prediction accuracy is fine tuning machine learning algorithm parameters. It will also be interesting to consider other classification algorithms.

Acknowledgment

This research was partially supported by the CSUEB Faculty Support Grant for the engagement of students in coursework related research.

REFERENCES

Berri, D.J., “Who Is ‘Most Valuable’? Measuring the Player’s Production of Wins in the National Basketball Association”, *Managerial and Decision Economics*, 20(8), 1999, 411-427.

Coates, D and Babatunde, O., “The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes?”, *International Journal of Sport Finance*, 5(1), 2010, 4-26.

Cooper, W.W., Ruiz, J.L. and Sirvent, I., “Selecting Non-Zero Weights to Evaluate Effectiveness of Basketball Players with DEA”, *European Journal of Operational Research*, 195(2), 2009, 563-574.

Martinez J.A. and Martinez L., “A Stakeholder Assessment of Basketball Player Evaluation Metrics”, *Journal of Human Sport & Exercise*, 6(1), 2011, 153-183.

Moreno, P. and Lozano, S., “A Network DEA Assessment of Team Efficiency in the NBA”, Annals of Operations Research, 214(1), 2014, 99-124.

Page, G.L., Fellingham, G.W. and Reese, C.S., “Using Box-Scores to Determine a Position’s Contribution to Winning Basketball Games”, Journal of Quantitative Analysis in Sports, 3(4), 2007, 1-16.

Sampaio. J., Janeira, M., Ibanez, S. and Lorenzo, A., “Discriminant Analysis of Game-Related Statistics between Basketball Guards, Forwards and Centres in Three Professional Leagues”, *European Journal of Sport Science*, 6(3), 2006, 173-178.

Sill, J., ‘Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing’, Presented at the MIT Sloan Sports Analytics Conference, Boston, Massachusetts, March 6, 2010.

Stiroh, K.J., ‘Playing for Keeps: Pay and Performance in the NBA’, *Economic Inquiry*, 45(1), 2006, 145-161.

APPENDICES

Appendix A: Features

1. GP: the number of games he played in his rookie season.
2. MPG: the average number of minutes per game.
3. FGM: the average number of field goals made per game.
4. FGA: the average number of field goals attempts per game.
5. FG%: the field goal percentage, equal to FGM/FGA.
6. FTM: the average number of free throws made per game.
7. FTA: the average number of free throws attempts per game.
8. FG%: the free throw percentage = FTM/FTA.
9. PPG: points scored per game.
10. RPG: rebounds per game.
11. APG: assists per game.
12. PF: personal fouls per game.
13. TM: total minutes played = GP*MPG.
14. PP48: points per 48 minutes = PPG/MPG*48.
15. RP48: rebounds per 48 minutes = RPG/MPG*48.
16. AP48: assists per 48 minutes = APG/MPG*48.
17. PF48: personal fouls per 48 minutes = PF/MPG*48.
18. FTM48: free throws made per 48 minutes = FTM/MPG*48.
19. FTA48: free throw attempts per 48 minutes = FTA/MPG*48.
20. EFF: efficiency = PPG + RPG + APG – (FGA – FGM) – (FTA – FTM)/2 – PF/2
21. PRO: production = EFF*GP
22. EFT_G: effort per game = FGA + FTA/2 + RPG + APG + PF
23. EFT_S: effort in the season = EFT_G * GP

Appendix B: False Positives for k = 25

First Name	Last Name	Draft Year	Pick	n
John	Hummer	1970	15	30
Elmore	Smith	1971	3	30
Ernie	DiGregorio	1973	3	30
Clark	Kellogg	1982	8	30
Arvydas	Sabonis	1986	24	30
Nick	Anderson	1989	11	30
Lionel	Simmons	1990	7	30

Eddie	Miller	1952	49	29
Ed	Fleming	1955	16	29
John	Barnhill	1959	79	29
Ray	Scott	1961	4	29
Al	Butler	1961	17	29
Charlie	Hardnett	1962	21	29
Art	Heyman	1963	2	29
Rod	Thorn	1963	3	29
Jim	Barnes	1964	3	29
Ron	Boone	1968	147	29
Billy	Paultz	1970	103	29
Dave	Robisch	1971	44	29
Brian	Taylor	1972	23	29
John	Williamson	1973	96	29
Mike	Sojourner	1974	10	29
Al	Skinner	1974	160	29
James	Edwards	1977	46	29
Mychal	Thompson	1978	1	29
Phil	Ford	1978	2	29
Terry	Tyler	1978	23	29
Michael	Brooks	1980	9	29
Larry	Smith	1980	24	29
Jay	Vincent	1981	24	29
Benoit	Benjamin	1985	3	29
Hot	Rod Williams	1985	45	29
Chuck	Person	1986	4	29
Ron	Harper	1986	8	29
Willie	Anderson	1988	10	29
Kevin	Edwards	1988	20	29
Sherman	Douglas	1989	28	29
Travis	Mays	1990	14	29
Billy	Owens	1991	3	29

Stacey	Augmon	1991	9	29
Walt	Williams	1992	7	29
Clarence	Weatherspoon	1992	9	29
Sean	Rooks	1992	30	29
Brian	Grant	1994	8	29
Joe	Smith	1995	1	29
Damon	Stoudamire	1995	7	29
Keith	Van Horn	1997	2	29
Derek	Anderson	1997	13	29
Lamar	Odom	1999	4	29
Emeka	Okafor	2004	2	29
O.J.	Mayo	2008	3	29
Jason	Thompson	2008	12	29
Tyreke	Evans	2009	4	29
Brandon	Jennings	2009	10	29
Jack	Stephens	1955	7	28
George	Lee	1959	26	28
Dan	Anderson	1965	89	28
Tom	Boerwinkle	1968	4	28
Jo Jo	White	1969	9	28
Wil	Jones	1969	69	28
Lloyd	Neal	1972	31	28
Swen	Nater	1973	16	28
Jim	Chones	1973	31	28
Scott	May	1976	2	28
Richard	Washington	1976	3	28
Cliff	Robinson	1979	11	28
Darrell	Griffith	1980	2	28
Kelvin	Ransey	1980	4	28
Quintin	Dailey	1982	7	28
Vern	Fleming	1984	18	28
Wayman	Tisdale	1985	2	28

Gerald	Wilkins	1985	47	28
Walter	Berry	1986	14	28
Armen	Gilliam	1987	2	28
Kenny	Smith	1987	6	28
Grant	Long	1988	33	28
Dino	Radja	1989	40	28
Jim	Jackson	1992	4	28
LaPhonso	Ellis	1992	5	28
Bryant	Reeves	1995	6	28
Marcus	Camby	1996	2	28
Kerry	Kittles	1996	8	28
Bobby	Jackson	1997	23	28
Luis	Scola	2002	55	28
Eric	Gordon	2008	7	28
Greg	Monroe	2010	7	28
Monk	Meineke	1952	35	27
Slick	Leonard	1954	10	27
Si	Green	1956	1	27
Al	Tucker	1967	6	27
Lucius	Allen	1969	3	27
Mack	Calvin	1969	187	27
Darnell	Hillman	1971	8	27
Clifford	Ray	1971	40	27
Mike	Gale	1971	47	27
Tom	Henderson	1974	7	27
John	Lucas	1976	1	27
Ron	Lee	1976	10	27
Mitch	Kupchak	1976	13	27
Mike	O'Koren	1980	6	27
Mike	Gminski	1980	7	27
Frank	Johnson	1981	11	27
Mitchell	Wiggins	1983	23	27

Ledell	Eackles	1988	36	27
Vernon	Maxwell	1988	47	27
Richard	Dumas	1991	46	27
Todd	Day	1992	8	27
Calbert	Cheaney	1993	6	27
Lamond	Murray	1994	7	27
Marc	Jackson	1997	37	27
Andre	Miller	1999	8	27
Jason	Richardson	2001	5	27
Jay	Williams	2002	2	27
Ben	Gordon	2004	3	27
Andrew	Bogut	2005	1	27
D.J.	Augustin	2008	9	27
Jonny	Flynn	2009	6	27
Marcus	Thornton	2009	43	27
Brandon	Knight	2011	8	27
Kemba	Walker	2011	9	27
Bucky	Bockhorn	1958	17	26
Chico	Vaughn	1962	28	26
Howard	Komives	1964	15	26
Dave	Stallworth	1965	6	26
Erwin	Mueller	1966	20	26
George	Thompson	1969	66	26
Pete	Cross	1970	23	26
Leonard	Gray	1974	26	26
David	Vaughn	1975	63	26
Rick	Robey	1978	3	26
Ron	Brewer	1978	7	26
Allen	Leavell	1979	104	26
Wes	Matthews	1980	14	26
Don	Collins	1980	18	26
Darwin	Cook	1980	70	26

Albert	King	1981	10	26
Herb	Williams	1981	14	26
Steve	Stipanovich	1983	2	26
Sam	Bowie	1984	2	26
Sam	Perkins	1984	4	26
Rex	Chapman	1988	8	26
Rony	Seikaly	1988	9	26
Gary	Grant	1988	15	26
Kendall	Gill	1990	5	26
Lindsey	Hunter	1993	10	26
Jalen	Rose	1994	13	26
Shane	Battier	2001	6	26
Kirk	Hinrich	2003	7	26
Josh	Childress	2004	6	26
Al	Thornton	2007	14	26
Mario	Chalmers	2008	34	26
Dion	Waiters	2012	4	26
Guy	Sparrow	1955	19	25
Joe	Strawder	1964	36	25
Art	Harris	1968	16	25
Don	Adams	1970	120	25
Freddie	Boyd	1972	5	25
Dave	Twardzik	1972	26	25
Ray	Williams	1977	10	25
Greg	Kelser	1979	4	25
Spud	Webb	1985	87	25
Chris	Morris	1988	4	25
Rod	Strickland	1988	19	25
Brian	Shaw	1988	24	25
Pooh	Richardson	1989	10	25
Mahmoud	Abdul-Rauf	1990	3	25
Dennis	Scott	1990	4	25

Toni	Kukoc	1990	29	25
Tyus	Edney	1995	47	25
Ron	Mercer	1997	6	25
Jason	Williams	1998	7	25
Kenneth	Faried	2011	22	25
Bradley	Beal	2012	3	25

Appendix C: False Negatives for k = 10

First Name	Last Name	Draft Year	Pick	n
Gene	Shue	1954	3	0
Bill	Bridges	1961	32	0
Jeff	Mullins	1964	7	0
Paul	Westphal	1972	10	0
Michael	Adams	1985	66	0
Steve	Nash	1996	15	0
Jermaine	O'Neal	1996	17	0
Michael	Redd	2000	43	0
Mo	Williams	2003	47	0
Larry	Jones	1964	20	1
Joe	Johnson	2001	10	1
Devin	Harris	2004	5	1
Kermit	Washington	1973	5	2
Theo	Ratliff	1995	18	2
Ricky	Pierce	1982	18	3
B.J.	Armstrong	1989	18	3
Jim	King	1963	13	4
John	Block	1966	27	4
Steve	Mix	1969	61	4
Fred	Brown	1971	6	4

Reggie	Lewis	1987	22	4
Jayson	Williams	1990	21	4
Rashard	Lewis	1998	32	4
Gerald	Wallace	2001	25	4
Jimmy	Butler	2011	30	4
Alex	English	1976	23	5
Danny	Ainge	1981	31	5
Andrew	Bynum	2005	10	5
Jeff	Teague	2009	19	5
Draymond	Green	2012	35	5
Darrall	Imhoff	1960	3	6
Eddie	Miles	1963	5	6
Bob	Love	1965	36	6
Rickey	Green	1977	16	6
Jim	Paxson	1979	12	6
Dirk	Nowitzki	1998	9	6
David	West	2003	18	6
Kyle	Korver	2003	51	6
Larry	Costello	1954	12	7
Sam	Jones	1957	8	7
Mark	Eaton	1982	72	7
Jamaal	Magloire	2000	19	7
Paul	George	2010	10	7
Richie	Regan	1953	4	8
Cliff	Hagan	1953	11	8
Jerry	Sloan	1965	4	8
Jon	McGlocklin	1965	27	8
Doug	Collins	1973	1	8
Detlef	Schrempf	1985	8	8
Tyrone	Hill	1990	11	8
Manu	Ginobili	1999	57	8
Zach	Randolph	2001	19	8

David	Lee	2005	30	8
Kyle	Lowry	2006	24	8
James	Donaldson	1979	73	9
Peja	Stojakovic	1996	14	9
Tracy	McGrady	1997	9	9
Mehmet	Okur	2001	37	9
Chris	Kaman	2003	6	9
DeMar	DeRozan	2009	9	9
Kevin	Duckworth	1986	33	10
Chris	Gatling	1991	16	10
Kawhi	Leonard	2011	15	10

Appendix D: Predictions on Players Drafted from 2013 to 2015

First Name	Last Name	Draft Year	Pick	n
Victor	Oladipo	2013	2	30
Michael	Carter-Williams	2013	11	30
Andrew	Wiggins	2014	1	30
Jahlil	Okafor	2015	3	30
Nikola	Jokic	2014	41	29
Karl-Anthony	Towns	2015	1	29
Jordan	Clarkson	2014	46	28
Kristaps	Porzingis	2015	4	28
Elfrid	Payton	2014	10	27
Emmanuel	Mudiay	2015	7	27
Devin	Booker	2015	13	27
D'Angelo	Russell	2015	2	26
Trey	Burke	2013	9	24
Nerlens	Noel	2013	6	23
Zach	LaVine	2014	13	21
Myles	Turner	2015	11	17
Mason	Plumlee	2013	22	16
Kelly	Olynyk	2013	13	15

Jabari	Parker	2014	2	13
Cody	Zeller	2013	4	12
Ryan	Kelly	2013	48	12
Willie	Cauley-Stein	2015	6	11
Bobby	Portis	2015	22	9
Frank	Kaminsky	2015	9	7
Tim	Hardaway	2013	24	6
Jusuf	Nurkic	2014	16	6
Giannis	Antetokounmpo	2013	15	5
Marcus	Smart	2014	6	5

Appendix E: IPython Script

```
# import relevant python packages and libraries
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn import metrics


#load training, validation, and prediction datasets
f_train = open('training.csv')
f_valid = open('validation.csv')
f_pred = open('pred.csv')
f_train_valid = open('train_valid.csv')
data_train = np.loadtxt(f_train, delimiter=',')
data_valid = np.loadtxt(f_valid, delimiter=',')
data_pred = np.loadtxt(f_pred, delimiter=',')
data_train_valid = np.loadtxt(f_train_valid, delimiter=',')

X_train = []    # create a list to store six sets of values of independent variables of training set
y_train = []    # create a list to store six sets of values of dependent variable of training set
n = 550   # each training set is of size 550
for i in range(6):
    X_train.append(data_train[(i*n):(i+1)*n, 2:])
    y_train.append(data_train[(i*n):(i+1)*n, 1])
X_valid = data_valid[:, 2:]
```

```
y_valid = data_valid[:, 1]
X_pred = data_pred[:, 2:]
X_train_valid = data_train_valid[:, 2:]
y_train_valid = data_train_valid[:, 1]

# Logistic Regression
logreg = []
y_pred = []
y_valid_pred = []
for i in range(6):
    logreg.append(LogisticRegression())             # create six Logistic Regression classifiers
    logreg[i].fit(X_train[i], y_train[i])           # train the Logistic Regression classifiers
    y_pred.append(logreg[i].predict(X_train[i]))    # generate in-sample prediction
    print ('in-sample accuracy of model {} is     '.format(i+1), metrics.accuracy_score(y_train[i],
y_pred[i]))                                         # print in-sample accuracy

    y_valid_pred.append(logreg[i].predict(X_valid))# generate out-of-sample accouracy
    print ('out-of-sample accuracy of model {} is '.format(i+1), metrics.accuracy_score(y_valid,
y_valid_pred[i]))                                   # print out-of-sample accuracy
  print

# Aggregate Logistic Regression Validation
for k in range(3, 6, 1):
    y_pred = []
    for j in range(len(y_valid)):
        if y_valid_pred[0][j] + y_valid_pred[1][j] + y_valid_pred[2][j] + y_valid_pred[3][j] \
        + y_valid_pred[4][j] + y_valid_pred[5][j] > k:
          y_pred.append(1)
        else:
          y_pred.append(0)

    print metrics.accuracy_score(y_valid, y_pred)
```