EECS 510 Final Project Report

Data Mining for NBA Finals MVP Prediction

Tzu-Chien Fu, Sheng-Feng Hsu and Qianying Yu

June 12, 2018

1 Abstract

This project is to build a framework to predict who will win the 2018 NBA Most Valuable Player (MVP) Award. Our model contains two components: In the first part, we create a web crawler to get the record of NBA game in the past years. We then use these historical data to build statistical regression models to predict the vote share for the MVP prediction. In the second part, we collect data from Twitter using Natural Language Processing (NLP) methods including keyword, name entity frequency, and sentiment analysis to determine the athletes' overall popularity. Finally, the results are reasonable and sensible.

2 Introduction

The National Basketball Association (NBA) is a men's professional basketball league in North America. According to the report of average TV viewer-ship from the statistic website Statista [1], NBA became one of the most popular sport in recent years. NBA produces considerable statistical information about each player, team, games, and seasons. Hence, everyone could take advantage of the abundant labeled data through the use of data mining techniques. To deal with this complexity and achieve better predictions, a large number of statistical machine learning methods have been implemented on these data. Sports data mining assists coaches and managers in result prediction, player performance assessment, player injury prediction, sports talent identification, and game strategy evaluation [2]. Also, the MVP prediction is nowadays very popular among fans around the world, which is particularly contributed to the expansion of sports betting.

To predict the 2018 NBA MVP winners by analyzing past statistical data of NBA games, we could use regression models to predict the vote share of the player in the MVP race. An vote share (award share) is a statistic in baseball, basketball and other sports. It is calculated by the number of points a player received over the total points of all first-place votes for the MVP award. In statistical modeling, regression analysis is a set of statistical processes that estimate the relationship between variables. Regression analysis is widely used for prediction, and its use has a significant overlap with machine learning. Regression analysis is also used to understand which of the independent variables are related to the dependent variable and explore the form of these relationships. Therefore, regression is what we focus on because we want to predict the MVP vote share.

However, Except for the data, the finals MVP award isn't just based on performance of individual player in the finals, but also based on fan support. We choose Twitter as our social media mining platform and collect tweets including finals MPV 2018 information to conduct NLP processing. Our goal of this part is to determine people's attitude toward whether they want a player to win the finals MVP. We use Natural Language Processing (NLP) methods including name entity frequency and Sentiment Analysis and the result is helpful for our final decision.

In this project, we will design a MVP prediction framework based on machine learning and data mining techniques. There are two components in our framework: First, We use historical game data to build statistical regression models to predict the vote share for the MVP prediction. Second, we collect data from Twitter and conduct NLP methods to determine the athletes' overall popularity. Afterwards, we combined the vote share prediction with the Twitter prediction to get our final MVP prediction.

We organize the remainder of the report as follows. Section 3 presents the method of using historical data to predict the final MVP. Section 4 discusses techniques for mining the tweets and perform the sentiment analysis. Section 5 presents our results of the MVP prediction, followed by conclusions in Section 6.

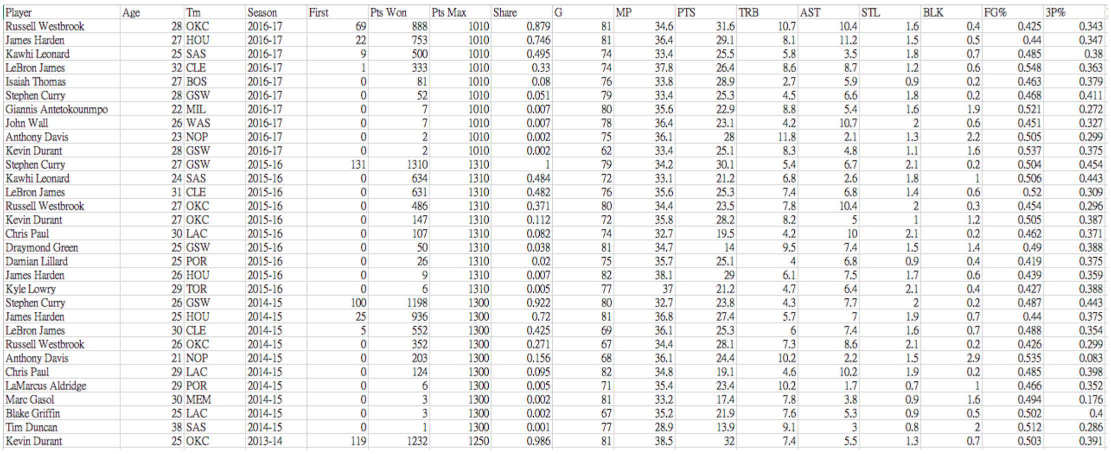
3 Vote Share Prediction by Historical Data

3.1 Data Collection

Data is the core of any data mining projects. Collecting sufficient amount of data with high quality is the fundamental and crucial step towards the success of the data mining problems. Our major data source comes from published data on a website (i.e., basketball-reference.com). The website provide basketball stats and history statistics, scores, and history for the NBA competition. In order to collect data efficiently, there are some automatic data collection tools.

However, data directly downloaded from web pages are insufficient or structureless, so we use Python with beautiful soup library to parse data and collect information from the website. We create a web crawler to get the record of NBA game in the past years. Meanwhile, we fix the "dirty" data to make our future work easier. To be more specific, the original data contains some redundancies and missing data. Then we adapt the data to meet certain specifications (e.g., no blanks allowed) and correct corrupt or inaccurate records.

The data we collected consists of NBA MVP voting data from the 1980-81 to the 2016-17 season. Our data contains some features such as players' age (Age), team (Tm), vote share (Share), games played (G), field goal percentage (FG%), 3 points field goal percentage (3P%),



Player	Age	Tm	Season	First	Pts Won	Pts Max	Share	G	MP	PTS	TRB	AST	STL	BLK	FG%	3P%
Russell Westbrook	28	OKC	2016-17	69	888	1010	0.879	81	34.6	31.6	10.7	10.4	1.6	0.4	0.425	0.343
James Harden	27	HOU	2016-17	22	753	1010	0.746	81	36.4	29.1	8.1	11.2	1.5	0.5	0.44	0.347
Kawhi Leonard	25	SAS	2016-17	9	500	1010	0.495	74	33.4	25.5	5.8	3.5	1.8	0.7	0.485	0.38
LeBron James	32	CLE	2016-17	1	333	1010	0.33	74	37.8	26.4	8.6	8.7	1.2	0.6	0.548	0.363
Isaiah Thomas	27	BOS	2016-17	0	81	1010	0.08	76	33.8	28.9	2.7	5.9	0.9	0.2	0.463	0.379
Stephen Curry	28	GSW	2016-17	0	52	1010	0.051	79	33.4	25.3	4.5	6.6	1.8	0.2	0.468	0.411
Giannis Antetokounmpo	22	MIL	2016-17	0	7	1010	0.007	80	35.6	22.9	8.8	5.4	1.6	1.9	0.521	0.272
John Wall	26	WAS	2016-17	0	7	1010	0.007	78	36.4	23.1	4.2	10.7	2	0.6	0.451	0.327
Anthony Davis	23	NOP	2016-17	0	2	1010	0.002	75	36.1	28	11.8	2.1	1.3	2.2	0.505	0.299
Kevin Durant	28	GSW	2016-17	0	2	1010	0.002	62	33.4	25.1	8.3	4.8	1.1	1.6	0.537	0.375
Stephen Curry	27	GSW	2015-16	131	1310	1310	1	79	34.2	30.1	5.4	6.7	2.1	0.2	0.504	0.454
Kawhi Leonard	24	SAS	2015-16	0	634	1310	0.484	72	33.1	21.2	6.8	2.6	1.8	1	0.506	0.443
LeBron James	31	CLE	2015-16	0	631	1310	0.482	76	35.6	25.3	7.4	6.8	1.4	0.6	0.52	0.309
Russell Westbrook	27	OKC	2015-16	0	486	1310	0.371	80	34.4	23.5	7.8	10.4	2	0.3	0.454	0.296
Kevin Durant	27	OKC	2015-16	0	147	1310	0.112	72	35.8	28.2	8.2	5	1	1.2	0.505	0.387
Chris Paul	30	LAC	2015-16	0	107	1310	0.082	74	32.7	19.5	4.2	10	2.1	0.2	0.462	0.371
Draymond Green	25	GSW	2015-16	0	50	1310	0.038	81	34.7	14	9.5	7.4	1.5	1.4	0.49	0.388
Damian Lillard	25	POR	2015-16	0	26	1310	0.02	75	35.7	25.1	4	6.8	0.9	0.4	0.419	0.375
James Harden	26	HOU	2015-16	0	9	1310	0.007	82	38.1	29	6.1	7.5	1.7	0.6	0.439	0.359
Kyle Lowry	29	TOR	2015-16	0	6	1310	0.005	77	37	21.2	4.7	6.4	2.1	0.4	0.427	0.388
Stephen Curry	26	GSW	2014-15	100	1198	1300	0.922	80	32.7	23.8	4.3	7.7	2	0.2	0.487	0.443
James Harden	25	HOU	2014-15	25	936	1300	0.72	81	36.8	27.4	5.7	7	1.9	0.7	0.44	0.375
LeBron James	30	CLE	2014-15	5	552	1300	0.425	69	36.1	25.3	6	7.4	1.6	0.7	0.488	0.354
Russell Westbrook	26	OKC	2014-15	0	352	1300	0.271	67	34.4	28.1	7.3	8.6	2.1	0.2	0.426	0.299
Anthony Davis	21	NOP	2014-15	0	203	1300	0.156	68	36.1	24.4	10.2	2.2	1.5	2.9	0.535	0.083
Chris Paul	29	LAC	2014-15	0	124	1300	0.095	82	34.8	19.1	4.6	10.2	1.9	0.2	0.485	0.398
LaMarcus Aldridge	29	POR	2014-15	0	6	1300	0.005	71	35.4	23.4	10.2	1.7	0.7	1	0.466	0.352
Marc Gasol	30	MEM	2014-15	0	3	1300	0.002	81	33.2	17.4	7.8	3.8	0.9	1.6	0.494	0.176
Blake Griffin	25	LAC	2014-15	0	3	1300	0.002	67	35.2	21.9	7.6	5.3	0.9	0.5	0.502	0.4
Tim Duncan	38	SAS	2014-15	0	1	1300	0.001	77	28.9	13.9	9.1	3	0.8	2	0.512	0.286
Kevin Durant	25	OKC	2013-14	119	1232	1250	0.986	81	38.5	32	7.4	5.5	1.3	0.7	0.503	0.391

Figure 1: Example of our historical data

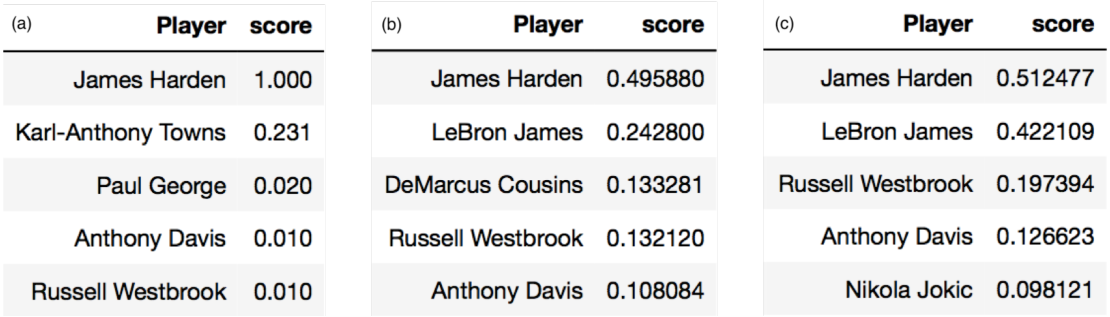
total rebounds (TRB), assists (AST) , steals (STL), blocks (BLK). (see Fig. 1 for examples). We use the data to build and train our statistical models for vote share prediction.

3.2 Regression

To address the task of the vote share regression, we applied three different tree-based models: Decision Tree, Random Forest Classifier and Gradient Boosting Regressor. Tree-based methods involve segmenting the feature space into a number of simple regions. To perform a prediction for a given testing data, we typically use the mean or the mode of the training data in the region to which it belongs. In addition, the set of splitting rules used to segment the feature space can be represented in a tree. Therefore, these type of methods are known as decision tree methods. To predict the response of a decision tree, we follow the tree from the root node down to a leaf node. At each node, we decide which branch to follow using the rule associated to that node. We continue until arriving at a leaf node. The prediction is the value associated to that leaf node.

Decision tree methods are simple and useful for interpretation. However, they typically too tied to the training set due to the overfitting. That is, the model doesn't generalize to new data well, which is the point of prediction. Hence, we also introduce random forests and gradient boosting for vote share regression. Each of them involves producing multiple trees which are then combined to yield a single consensus prediction. In other words, each tree can uses a subset of the data to give a different answer. The final regression is the most common amongst the trees.

To build high accuracy prediction models for predicting the 2018 NBA MVP of the final series, we could employ the historical data to train the models mentioned above. However, the final series only has four to seven games every year, and it's hard to build an effective model by using only a limited sample size. We designed our models by using all the data in the regular season from 1981-80 to 2016-17 and the vote share of the season MVP. As the problem mentioned above, the data of final series only contained four to seven games, and our training model used



(a)

Player	score
James Harden	1.000
Karl-Anthony Towns	0.231
Paul George	0.020
Anthony Davis	0.010
Russell Westbrook	0.010

(b)

Player	score
James Harden	0.495880
LeBron James	0.242800
DeMarcus Cousins	0.133281
Russell Westbrook	0.132120
Anthony Davis	0.108084

(c)

Player	score
James Harden	0.512477
LeBron James	0.422109
Russell Westbrook	0.197394
Anthony Davis	0.126623
Nikola Jokic	0.098121

Figure 2: Results of 2017-18 season MVP prediction using:(a) Decision tree method, (b) Random forest method, and (c) Gradient boosting method. Note that we sorted players by their vote share score.

the data of full season (mostly eighty-two games), so we normalized the data, which is counted to the number, by simply multiply to normalization factor.

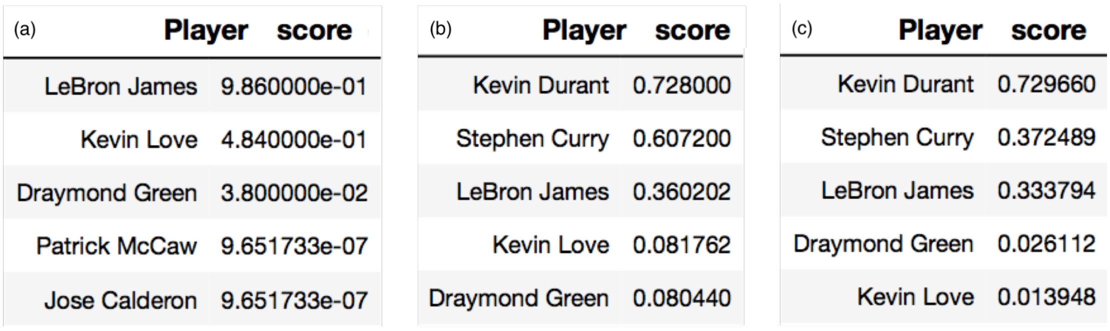
After preprocessing the data we obtained from the web crawler, we could use Python with the scikit-learn (sklearn) library to construct our models. The training data in the model is the historical data in regular season (from 1981-82 to 2016-17), and the testing data is 2017-18 regular season and 2017-18 final series stats.

3.3 Feature Selection

The accuracy of predictions will depend on proper manual or automatic selection of the most significant, highly correlated features. In prediction, features selected are required to be comprehensive and representative. The initial idea is putting all the features of our data into our models. We also use feature selection algorithms provided by the Padas library for telling us how important the features are for the final result. After we figure out which features are more important in our feature list, we can select a smaller but more effective list of features. Trimming the features down to a smaller list could improve the efficiency of our trees and possibly improve accuracy as well.

3.4 Results of Vote Share Prediction

To assess the performance of our vote share prediction models, we apply the historical data in regular season from 1981-82 to 2016-17 captured by our web crawler for training. The training dataset contains 15,798 data collected from the Internet. On the other hand, the 2017-18 regular season and 2017-18 final series players' stats was used for testing. The results of 2017-18 season MVP prediction are shown in Fig. 2. We found that all methods predict that James Harden will win the season MVP award this year. Recall that, since the final series only has four to seven games every year, it's hard to build an effective model by using a limited data. Therefore, we directly adapted our season MVP prediction models to predicting the final series MVP. From the results in Fig. 5, we see that decision tree method have a bad result due to overfitting our regular season data. However, both the random forest and the gradient boosting method improved the model and showed promising results.



(a)

Player	score
LeBron James	9.860000e-01
Kevin Love	4.840000e-01
Draymond Green	3.800000e-02
Patrick McCaw	9.651733e-07
Jose Calderon	9.651733e-07

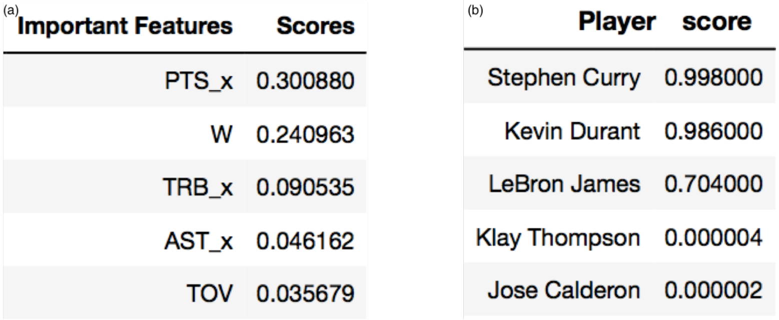
(b)

Player	score
Kevin Durant	0.728000
Stephen Curry	0.607200
LeBron James	0.360202
Kevin Love	0.081762
Draymond Green	0.080440

(c)

Player	score
Kevin Durant	0.729660
Stephen Curry	0.372489
LeBron James	0.333794
Draymond Green	0.026112
Kevin Love	0.013948

Figure 3: Results of 2017-18 final series MVP prediction using:(a) Decision tree method, (b) Random forest method, and (c) Gradient boosting method. Note that we sorted players by their vote share score.



(a)

Important Features	Scores
PTS_x	0.300880
W	0.240963
TRB_x	0.090535
AST_x	0.046162
TOV	0.035679

(b)

Player	score
Stephen Curry	0.998000
Kevin Durant	0.986000
LeBron James	0.704000
Klay Thompson	0.000004
Jose Calderon	0.000002

Figure 4: Results of feature selection:(a) Top five important features in the feature list (i.e., points per game, wins, total rebounds, assists, and turnover), and (b) Prediction result using a decision tree model trained based on the top five important features.

In addition, to get a more reasonable result using decision tree to predict the final series MVP, we try to select a more effective list of features to train our model. The feature selection result is shown in Fig. 4. Note that the feature of ”wins” was the second important in our feature list. That is, it would be considered in our model when we predict the vote share score. The result of the vote share prediction also verify that the accuracy of predictions will depend on the selection of significant and highly correlated features

4 Natural Language Processing with Twitter

4.1 Data Collection

Twitter is the best platform for analyzing the athletes' overall population. So we conduct our social media mining and natural language processing based on tweets during finals. First we collected tweets with the hash-tags of #NBAfinal2018, #NBAFinals and #nbafinal18 and tried to find the most mentioned player. But the disadvantages of doing this is very obvious. We could get some information we don't really want. For example, people could mention a player just to make jokes that have nothing to do with the NBA. And we also can't get enough the information about people's attitude on the finals MVP.

4.2 Key Word Selection

Due to the lack of useful information, we finally combined the name or nickname by fans of the players and MVP as the keyword to collect tweets from the beginning to the actual ending of games in final. Since our regression model has shrunk the candidates for finals MVP to threes plays, what we have to do is to search tweets with key words "Curry MVP", "Durant MVP" or "KD MVP" for Warriors and "LeBron MVP" or "LBJ MVP" for Cavaliers from May 31th, 2018 to June 8th, 2018. Here we use Tweepy to collect data and store tweets of each play in test files.

4.3 Sentiment Analysis

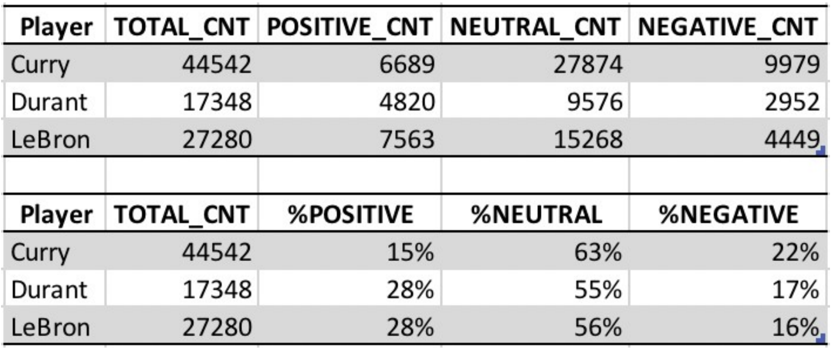
However, just counting the Name Entity frequency is not enough since people could express their anti-attitude on a player's wining chance by mention both the key word of their name and MVP. For example:

- At least **Curry** won't win finals **MVP**.
- **Curry** is trying his absolute best to never get finals **MVP**.

The number of such tweets has increasingly suddenly after Curry's poor performance in the third game, where he only got 11 points but Durant got over 40 points. Therefore, the sentiment analysis works here for preventing the misunderstanding of people's attitude through tweets. To implement sentiment analysis on the tweets we collected, first we need to clean the tweet text by removing links, special characters using simple Regex statements. And then we use TextBlob, the python library, for processing textual data. After deriving the attitude of the one who tweets, we could determine whether the tweet is positive, negative or neutral.

4.4 Results of NLP

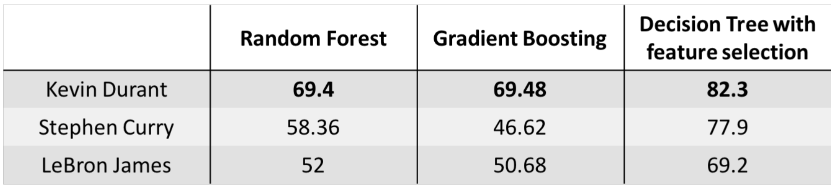
After calculation of the count of tweets on different attitude of each person, we got the statistic data in figure 5. From the numbers and percentage of it, we can see that although Curry owns the more popularity than Durant in Warriors, tweets that have negative attitude on him also get most number and percentage. Meanwhile, tweets about Durant and LeBron winning MVP have very close components. We use a variable p to describe the probability of a random person's attitude toward weather a player could win the finals MVP, where $p = \%positive + \%neutral - \%negative$. So $p_C = 56\%, p_D = 66\%, p_L = 68\%$, which means that the probability of a random fan hoping Stephen Curry, Kevin Durant or LeBron James win the finals MVP could separately be 56%, 66% or 68%. Since the finals MVP always belongs to a player in the winner team, our result of NLP analysis consider Kevin Durant or LeBron James as the winner of the finals MVP.



Player	TOTAL_CNT	POSITIVE_CNT	NEUTRAL_CNT	NEGATIVE_CNT
Curry	44542	6689	27874	9979
Durant	17348	4820	9576	2952
LeBron	27280	7563	15268	4449

Player	TOTAL_CNT	%POSITIVE	%NEUTRAL	%NEGATIVE
Curry	44542	15%	63%	22%
Durant	17348	28%	55%	17%
LeBron	27280	28%	56%	16%

Figure 5: Results of Sentiment Analysis of tweets during finals 2018



	Random Forest	**Gradient Boosting**	**Decision Tree with feature selection**
Kevin Durant	**69.4**	**69.48**	**82.3**
Stephen Curry	58.36	46.62	77.9
LeBron James	52	50.68	69.2

Figure 6: Final series MVP prediction using the average of the score of vote share regression and NLP analysis (the value showed in percentage).

5 Conclusion

The final version of our project is slightly different from the result of the meeting. There are many reasons that why we change our prediction models. At the beginning of our Project Preliminary Report, we proposed to build models which could predict the outcome of a game. However, the final goal of our project is to predict the finals MVP based on the performance of a player and the power of social media. Moreover, we want to build models not only predict the finals MVP but also the MVP of the regular season. However, the game outcome prediction models would not help us to achieve the goal. The final version of our prediction models could be viewed as a group of voters, which have comprehensive and representative information of all season or final series performance of the player. After our linearly combining (i.e., take an average of) the results of our regression models and NLP analysis with social media, our prediction of the finals MVP 2018 comes out to be Kevin Durant (see Fig. 6), which is successful according to the final awarding fact.

References

[1] https://www.statista.com/statistics/240377/nba-finals-tv-viewership-in-the-united-states/

[2] Haghighat, Maral, Hamid Rastegari, and Nasim Nourafza. ”A review of data mining techniques for result prediction in sports.” Advances in Computer Science: an International Journal 2.5 (2013): 7-12.

[3] Miljković, Dragan, et al. ”The use of data mining for basketball matches outcomes prediction.” Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on. IEEE, 2010.

[4] Lloyd Smith, Bret Lipscomb, Adam Simkins, “Data Mining in Sports: Predicting Cy Y oung A ward Winners”, Journal of Computing Sciences in Colleges, vol. 22, issue 4, April 2007, pp. 115-121, Consortium for Computing Sciences in Colleges

[5] Hosmer, D.W. and Lemeshow, S. ”Applied logistic regression,” Wiley-Interscience, 2000

[6] Michael Baulch, “Using Machine Learning to Predict the Results of Sporting Matches”, Department of Computer Science and Electrical Engineering, University of Queensland

[7] Alan McCabe, Jarrod Trevathan, “Artificial Intelligence in Sports Prediction”, Proceedings of the Fifth International Conference on Information Technology: New Generations, 2008, pp. 1194-1197, IEEE Computer Society

[8] Baghal, Tarek. ”Are the” Four Factors” Indicators of One Factor? An Application of Structural Equation Modeling Methodology to NBA Data in Prediction of Winning Percentage.” Journal of Quantitative Analysis in Sports 8.1 (2012).