Predict Basketball Team Winning Record

Charles Chen[1] and Mason Chen[2]

1. Applied Materials, Sunnyvale, CA 94085, Charles.Chen.Training@gmail.com
2. Stanford OHS, Palo Alto, CA94306, Mason.Chen.Training@gmail.com

Abstract

This paper is to build an empirical model to predict the NBA team winning percentage based on their team offensive, defensive, and differential statistics by collecting historical data during 2003-2016. The raw data have been standardized through Z transformation to remove mean and large variance bias effect. A multiple linear and step regression model was derived to predict the team winning record. After trimmed the insignificant regression terms, the derived model can predict team winning percent with R-Square > 0.95. The multi-linearity concerns were addressed by looking at the Variance Inflation Factor > 10. The redundant terms were removed to avoid over-fit risk. The regression model has identified 3-point Percentage, Turn Over, and Point per Game most critical to the team offensive efficiency. This observation is consistent with modern basketball. In defense, how to defend the rebound and opponent's field goal percentage are most critical. Warriors' 2015-2016 team record has been identified as an extreme outlier since their winning formula and team statistics are significantly different from the remaining 29 teams. The 2nd-order and Interaction Terms were added to enhance the prediction accuracy. The nonlinearity terms have indicated the complexity of the basketball team behaviors. Defense Field Goal% * Defense Point per Game was identified as the most significant interaction term. Which may reflect the Best Defense is the start of a good Offense. The model built based on 2003-2016 data was further validated by the new season 2016-2017. The model accuracy was proved to be within +/-5% winning percent of the predicted target across all 30 teams. This model can provide NBA coaches and general managers how to draft, recruit, trade, or sign particular players to build a desired Championship team based on the winning % formula. This methodology can be applied to NBA play-off and other major professional sports like baseball, football, hockey, soccer.

Key words: Regression, Sports Analytics, Predictive Model, Statistics

1. Introduction

Sports are big part of our daily life. Every major city has their professional teams and local fans are very supportive to their local professional players as their heroes or role models. National Basketball Associate (NBA) is the largest basketball organization. Each year, each NBA team is fighting for the playoff spot to win the championship. This paper would try to formulate what could be the most deciding factors to formulate how to build a championship team by analyzing historical team statistics.

In major professional sports, the coach and team management are looking for ways to win more games to build their championship dynasty (such as 1980 Celtics, 1990 Lakers, 2000 Bulls, early 2010 Spurs, late 2010 Warriors) in order to attract more fans to support their business. The dynasty era was dominated by Centers, Forwards, now Shooters. Sports statistical modeling analytics [1-5] is becoming a critical approach to uncover the winning patterns hidden in sports data collected during each game played. The objective of this paper is to build a statistical model based on the past team offensive, defensive, and differential statistics in order to predict the NBA 2016-2017 Regular Season Team Record. There are several research talks presented in MIT Sloan Sports Analytics Conference [6-8]. These papers have used intensive Analytics to uncover players' playing patterns and help coach develop each player in order to create and maximize each player's values to their specific team.

In Figure 1, the authors have demonstrated the project scope of this paper: (1) use the 2003-2016 team offensive, defensive, differential statistics (input independent variables Xs) to build a transfer function to predict the 2003-2016 team record (Y); (2) use the same transfer function and 2016-2017 team offensive, defensive, and differential statistics to predict the new 2016-2017 regular season team record. 2003-2016 data collected is to provided sufficient sample size to build the predictive model.

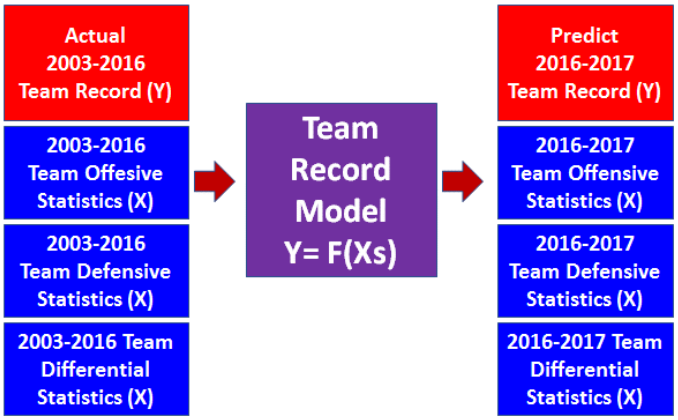

Figure 1: Build Predictive Model

2. Experimental Section

Author has laid out three subsections: (1) Raw Data Collection, and (2) Apply Z Transformation.

2.1 Raw Data Collection

Team statistics and record were collected [9,10] from the ESPN Sports NBA Website as shown in Figure 2.

Team Stats | Opposing Team Stats | Differential — Filter: NBA | 2015-16 Season

	FG			3PT			FT			Rebounds			Misc					
Team	M	A	Pct	M	A	Pct	M	A	Pct	Off	Def	Tot	Ast	TO	Stl	Blk	PF	Pts
Golden State Warriors	42.5	87.3	48.7	13.1	31.6	41.6	16.7	21.8	76.3	10.0	36.2	46.2	28.9	14.9	8.4	6.1	20.7	114.9
San Antonio Spurs	40.1	82.9	48.4	7.0	18.5	37.5	16.4	20.4	80.3	9.4	34.5	43.9	24.5	12.5	8.3	5.9	17.5	103.5
Oklahoma City Thunder	41.1	86.4	47.6	8.3	23.7	34.9	19.7	25.2	78.2	13.1	35.6	48.6	23.0	15.5	7.4	5.9	20.6	110.2
Miami Heat	38.4	81.7	47.0	6.1	18.0	33.6	17.1	23.0	74.4	9.8	34.3	44.1	20.8	13.2	6.7	6.5	18.3	100.0
Milwaukee Bucks	38.4	82.2	46.7	5.4	15.6	34.5	17.0	22.7	74.7	10.5	31.2	41.7	23.1	14.6	8.2	5.8	20.7	99.0
Los Angeles Clippers	38.3	82.4	46.5	9.7	26.7	36.4	18.2	26.2	69.2	8.8	33.3	42.0	22.8	12.4	8.6	5.6	21.3	104.5
Minnesota Timberwolves	37.7	81.3	46.4	5.5	16.4	33.8	21.4	27.0	79.2	10.0	31.5	41.6	23.4	14.4	8.0	4.6	20.7	102.4
Sacramento Kings	40.0	86.4	46.4	8.0	22.4	35.9	18.5	25.5	72.5	10.6	33.7	44.2	24.5	15.5	8.9	4.5	20.4	106.6
Cleveland Cavaliers	38.7	84.0	46.0	10.7	29.6	36.2	16.3	21.7	74.8	10.6	33.9	44.5	22.7	12.9	6.7	3.9	20.3	104.3
Washington Wizards	39.5	85.8	46.0	8.6	24.2	35.8	16.5	22.5	73.0	9.1	32.8	41.8	24.5	13.9	8.6	3.9	20.8	104.1
Atlanta Hawks	38.6	84.4	45.8	9.9	28.4	35.0	15.6	20.0	78.3	8.3	33.8	42.1	25.6	14.5	9.1	5.9	19.1	102.8
Orlando Magic	39.5	86.8	45.5	7.8	22.2	35.0	15.2	20.1	75.7	10.3	33.0	43.3	23.6	13.6	8.2	5.1	20.7	102.1
Brooklyn Nets	38.2	84.4	45.3	6.5	18.4	35.2	15.7	20.7	75.7	10.5	31.9	42.4	22.3	14.3	7.6	4.0	18.0	98.6
Houston Rockets	37.7	83.5	45.2	10.7	30.9	34.7	20.4	29.4	69.4	11.3	31.7	43.1	22.2	15.2	10.0	5.2	21.8	106.5
Toronto Raptors	36.7	81.3	45.1	8.6	23.4	37.0	20.8	26.7	77.7	10.2	33.2	43.4	18.7	12.1	7.8	5.5	19.6	102.7
Portland Trail Blazers	38.6	85.9	45.0	10.5	28.5	37.0	17.4	23.0	75.4	11.6	33.9	45.5	21.3	14.1	6.9	4.6	21.7	105.1
Indiana Pacers	38.3	85.2	45.0	8.1	23.0	35.1	17.4	22.8	76.4	10.3	33.9	44.2	21.2	14.3	9.0	4.8	20.0	102.2
Utah Jazz	36.1	80.4	44.9	8.5	23.9	35.5	17.1	23.0	74.4	10.7	32.5	43.2	19.0	14.2	7.7	5.2	20.2	97.7
New Orleans Pelicans	38.5	85.9	44.8	8.6	23.8	36.0	17.3	22.2	77.6	9.5	33.1	42.6	22.2	13.0	7.7	4.2	20.9	102.7
Dallas Mavericks	37.4	84.1	44.4	9.8	28.6	34.4	17.7	22.3	79.4	9.2	33.9	43.1	22.1	12.3	6.8	3.7	19.5	102.3
Denver Nuggets	37.7	85.4	44.2	8.0	23.7	33.7	18.5	24.1	76.6	11.5	33.1	44.6	22.7	14.2	7.4	4.8	21.0	101.9
Chicago Bulls	38.6	87.4	44.1	7.9	21.4	37.1	16.5	21.0	78.7	11.1	35.2	46.3	22.8	13.3	6.0	5.7	18.8	101.6
Memphis Grizzlies	36.8	83.6	44.0	6.1	18.5	33.1	19.3	24.7	78.3	11.2	30.5	41.6	20.7	12.7	8.8	4.3	21.7	99.1
Charlotte Hornets	37.0	84.4	43.9	10.6	29.4	36.2	18.7	23.7	79.0	9.0	35.0	43.9	21.7	11.9	7.3	5.3	18.1	103.4

Figure 2: Team Statistics Raw Data

2.2 Apply Z Standard Score Transformation

Prior to build a predictive model, Z transformation [11,12] is applied on team offensive, defensive, and differential statistics in Figure 3. Z transformation can eliminate any uneven influence (larger variance) among different team statistics categories in order to build an unbiased model. Otherwise, the predictive model may be dominated by any team statistics with larger variance.

	C33	C34	C35	C36	C37	C38	C39	C40	C41	C42	C43	C44	C45	C46	C47	C48
	O-FG%_1	O-3pt%_1	O-FT%_1	O-RB_1	O-Ast_1	O-TO_1	O-Stl_1	O-Blk_1	O-PF_1	O-Pts_1	D-FG%_1	D-3pt%_1	D-FT%_1	D-RB_1	D-Ast_1	D-TO_1
1	0.37	-0.15	0.71	-0.96	1.50	0.59	1.28	1.12	-0.90	0.04	-1.56	-1.11	-0.13	1.33	-0.20	1.2
2	-0.83	-0.99	0.85	0.67	0.86	-0.25	1.38	-0.89	1.28	0.80	-0.86	-1.26	-0.13	1.08	-0.99	1.4
3	0.05	-0.04	-0.03	-0.79	0.01	0.42	-0.24	-1.12	-1.75	-1.07	2.07	1.14	0.45	-0.18	1.57	-0.2
4	-0.83	0.52	0.91	0.09	-0.27	-1.60	-0.55	0.41	-1.67	0.19	-0.63	-0.31	0.88	0.45	0.65	-0.5
5	-0.71	1.03	0.83	1.49	0.23	-0.42	-1.87	0.89	-1.13	-0.28	-0.86	-0.60	-1.42	1.04	0.15	-1.8
6	0.49	0.52	-0.28	0.44	0.19	-0.76	-1.16	-1.24	0.03	0.43	-0.32	-0.46	-0.99	-1.34	-0.63	-0.6
7	-0.52	-0.49	1.03	-0.38	-0.08	-1.26	-1.06	-1.48	-0.59	-0.10	-0.09	-0.82	-1.35	0.99	-0.20	-0.2
8	-0.64	-0.88	0.23	0.50	0.19	0.34	-0.45	-0.18	0.58	-0.20	0.68	1.28	0.30	-0.52	0.86	-0.5
9	-0.83	-0.49	-2.56	1.49	-1.30	-0.67	-0.85	-1.48	-0.98	-0.18	0.68	0.12	1.88	-0.61	-0.56	-0.6
10	2.19	3.56	0.14	1.43	2.99	0.93	0.57	1.36	0.34	3.23	-1.33	-1.55	0.01	0.07	-0.06	-0.0
11	-0.01	-0.32	-1.82	-0.38	-0.04	1.18	2.19	0.30	1.20	1.01	0.53	0.56	-0.06	0.41	1.64	1.4
12	-0.14	-0.10	0.17	0.26	-0.49	0.42	1.18	-0.18	-0.20	-0.12	-0.94	-1.40	-0.42	0.36	-1.06	1.0
13	0.81	0.63	-1.88	-1.02	0.23	-1.18	0.77	0.77	0.81	0.48	-1.40	-1.11	-0.42	1.42	-0.77	0.6
14	-2.40	-2.00	0.66	-0.44	-1.94	-0.67	-0.65	-1.01	0.03	-1.42	1.61	-0.46	-1.86	1.33	1.71	-1.2
15	-0.77	-1.22	0.71	-1.26	-0.72	-0.93	0.98	-0.77	1.12	-0.94	0.30	0.85	0.81	-0.27	-0.27	1.2
16	1.12	-0.94	-0.40	0.20	-0.67	-0.50	-1.16	1.83	-1.52	-0.70	-0.79	-0.39	0.95	-1.19	-1.48	-1.1
17	0.93	-0.43	-0.31	-1.20	0.37	0.68	0.37	1.01	0.34	-0.97	0.14	-0.10	0.81	-0.37	1.57	0.7

Figure 3: Z Transformation on Team Statistics

3. Results and Discussion

3.1 Build Multiple Linear Regression Model

The multiple linear regression model was built by using the 2003-2016 Team Statistics and Team Record in Figure 4. ANOVA Table has listed the most significant variables (P-Value) < 0.5 and the responding Regression R-Sq (Adjusted) is 96.75% which indicated the built predictive model is reliable to predict the team record performance based on the few identified team statistics variables.

```
Analysis of Variance

Source       DF    Adj SS    Adj MS  F-Value  P-Value
Regression   10  0.813752  0.081375    87.33    0.000
  O-3pt%_1    1  0.003523  0.003523     3.78    0.067
  O-TO_1      1  0.011594  0.011594    12.44    0.002
  O-Pts_1     1  0.029247  0.029247    31.39    0.000
  D-FG%_1     1  0.004314  0.004314     4.63    0.045
  D-RB_1      1  0.000603  0.000603     0.65    0.431
  D-TO_1      1  0.005751  0.005751     6.17    0.022
  D-Pts_1     1  0.006720  0.006720     7.21    0.015
  C-RB_1      1  0.000539  0.000539     0.58    0.456
  C-Ast_1     1  0.001311  0.001311     1.41    0.250
  C-Stl_1     1  0.002288  0.002288     2.46    0.134
Error        19  0.017704  0.000932
Total        29  0.831457


Model Summary

         S    R-sq  R-sq(adj)  R-sq(pred)
0.0305254  97.87%     96.75%      92.91%
```

Figure 4: Multiple Linear Regression Model

Main effect regression coefficients and regression equation were listed in Figure 5. However, authors also checked any dependency among the identified input variables (Xs) to assess the multi-collinearity risk. In the VIF [13] (Variance Inflation Factor) column, five VIF index are above 10, which has shown significant concern on the Multi-Collinearity, which may inflate the regression R-Sq (Adjusted) and impact the regression equation. 3 out of five terms with VIF > 10 are also have P-values below 0.05. VIF concern may significantly impact the model reliability.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.49990	0.00557	89.70	0.000	
O-3pt%_1	0.01786	0.00918	1.94	0.067	2.63
O-TO_1	-0.0525	0.0149	-3.53	0.002	6.88
O-Pts_1	0.0983	0.0175	5.60	0.000	9.58
D-FG%_1	-0.0389	0.0181	-2.15	0.045	10.15
D-RB_1	-0.0147	0.0183	-0.80	0.431	10.44
D-TO_1	0.0461	0.0185	2.48	0.022	10.71
D-Pts_1	-0.0637	0.0237	-2.69	0.015	17.52
C-RB_1	0.0155	0.0204	0.76	0.456	13.00
C-Ast_1	-0.0137	0.0115	-1.19	0.250	4.14
C-Stl_1	-0.0248	0.0159	-1.57	0.134	7.83

Regression Equation

Win% = 0.49990 + 0.01786 O-3pt%_1 - 0.0525 O-TO_1 + 0.0983 O-Pts_1 - 0.0389 D-FG%_1
- 0.0147 D-RB_1 + 0.0461 D-TO_1 - 0.0637 D-Pts_1 + 0.0155 C-RB_1 - 0.0137 C-Ast_1
- 0.0248 C-Stl_1

Figure 5: Regression Equation and VIF

In addition to VIF, authors also checked the any residual or leverage outlier which may influence the regression model significantly as shown in Figure 6. Observed Data #10 was detected as residual outlier which has standard residual at 2.66. This Data #10 happens to be the Warriors Team which just broke Bulls' 72-win record. Warriors has created a new era on emphasizing 3-points, team assistance, and fast offensive flow. 2017-2018 Rocket team tried to duplicate the Warriors' winning mode (added Chris Paul for higher team assistance, increased 3-Point made). Rocket performed the best in the regular season and almost beat Champion Warriors in the Post-Season. Most NBA teams are looking for similar winning patterns by recruiting more 3-Point shooters, faster offense flow (heavy/slow centers are losing their playing time).

Fits and Diagnostics for Unusual Observations

Obs	Win%	Fit	Resid	Std Resid	
10	0.8900	0.8399	0.0501	2.66	R

R Large residual

	C1-T	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Team	Win%	O-FG%	O-3pt%	O-FT%	O-RB	O-Ast	O-TO	O-Stl	O-Blk	O-PF	O-Pts
1	Atlanta Hawks	0.585	45.8	35.0	78.3	42.1	25.6	14.5	9.1	5.9	19.1	102.8
2	Boston Celtics	0.585	43.9	33.5	78.8	44.9	24.2	13.5	9.2	4.2	21.9	105.7
3	Brooklyn Nets	0.256	45.3	35.2	75.7	42.4	22.3	14.3	7.6	4.0	18.0	98.6
4	Charlotte Hornets	0.585	43.9	36.2	79.0	43.9	21.7	11.9	7.3	5.3	18.1	103.4
5	Chicago Bulls	0.512	44.1	37.1	78.7	46.3	22.8	13.3	6.0	5.7	18.8	101.6
6	Cleveland Cavaliers	0.695	46.0	36.2	74.8	44.5	22.7	12.9	6.7	3.9	20.3	104.3
7	Dallas Mavericks	0.512	44.4	34.4	79.4	43.1	22.1	12.3	6.8	3.7	19.5	102.3
8	Denver Nuggets	0.402	44.2	33.7	76.6	44.6	22.7	14.2	7.4	4.8	21.0	101.9
9	Detroit Pistons	0.537	43.9	34.4	66.8	46.3	19.4	13.0	7.0	3.7	19.0	102.0
10	Golden State Warriors	0.890	48.7	41.6	76.3	46.2	28.9	14.9	8.4	6.1	20.7	114.9
11	Houston Rockets	0.500	45.2	34.7	69.4	43.1	22.2	15.2	10.0	5.2	21.8	106.5
12	Indiana Pacers	0.549	45.0	35.1	76.4	44.2	21.2	14.3	9.0	4.8	20.0	102.2
13	Los Angeles Clippers	0.646	46.5	36.4	69.2	42.0	22.8	12.4	8.6	5.6	21.3	104.5
14	Los Angeles Lakers	0.207	41.4	31.7	78.1	43.0	18.0	13.0	7.2	4.1	20.3	97.3
15	Memphis Grizzlies	0.512	44.0	33.1	78.3	41.6	20.7	12.7	8.8	4.3	21.7	99.1

Figure 6: Detect Regression Outliers

It's not surprised that the Warriors team statistics and team record is off the predictive regression chart as a residual outlier.

3.2 Improve the Predictive Model

In order to address the model adequacy concerns, authors have first trimmed the less significant terms (P-value above 0.2) as shown in Figure 7. Two terms were dropped from the previous regression model and R-Sq (Adjusted) has actually slightly been improved from 96.75% to 96.82% even R-Sq has been degraded. R-Sq (Adjusted) [14] is a better index to assess the multiple linear regression model. Authors would like to keep the remaining variables with P-values under 0.2 since little impact to trim the regression model further.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	0.813214	0.090357	99.06	0.000
O-3pt%_1	1	0.002988	0.002988	3.28	0.085
O-TO_1	1	0.011056	0.011056	12.12	0.002
O-Pts_1	1	0.055556	0.055556	60.91	0.000
D-FG%_1	1	0.004438	0.004438	4.87	0.039
D-RB_1	1	0.003519	0.003519	3.86	0.064
D-TO_1	1	0.005214	0.005214	5.72	0.027
D-Pts_1	1	0.007830	0.007830	8.58	0.008
C-Ast_1	1	0.001824	0.001824	2.00	0.173
C-Stl_1	1	0.002845	0.002845	3.12	0.093
Error	20	0.018243	0.000912		
Total	29	0.831457			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0302016	97.81%	96.82%	93.25%

Figure 7: Trim Regression Model

Authors further evaluated the model adequacy on the reduced model as shown in Figure 8. The removed two insignificant factors happen to be the higher dependent ones with VIF (Variance Inflation Factor) > 10. Only two factors still with VIF > 10. Though, these two factors are significant with P-values < 0.05. Authors decided to keep these two highly dependent factors in the regression mod

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.49990	0.00551	90.66	0.000	
O-3pt%_1	0.01532	0.00846	1.81	0.085	2.28
O-TO_1	-0.0501	0.0144	-3.48	0.002	6.57
O-Pts_1	0.1066	0.0137	7.80	0.000	5.93
D-FG%_1	-0.0394	0.0179	-2.21	0.039	10.13
D-RB_1	-0.0247	0.0126	-1.96	0.064	5.05
D-TO_1	0.0420	0.0176	2.39	0.027	9.81
D-Pts_1	-0.0674	0.0230	-2.93	0.008	16.80
C-Ast_1	-0.0157	0.0111	-1.41	0.173	3.92
C-Stl_1	-0.0272	0.0154	-1.77	0.093	7.53

Regression Equation

Win% = 0.49990 + 0.01532 O-3pt%_1 - 0.0501 O-TO_1 + 0.1066 O-Pts_1 - 0.0394 D-FG%_1
- 0.0247 D-RB_1 + 0.0420 D-TO_1 - 0.0674 D-Pts_1 - 0.0157 C-Ast_1 - 0.0272 C-Stl_1

Figure 8: Evaluate Model Adequacy

In Figure 9, two residual outliers were detected in the reduced model (#10, #15). The new #15 outlier is from Memphis Grizzles. This is an interesting finding. Authors could not well explain this new outlier pattern and which is out of this paper scope.

Obs	Win%	Fit	Resid	Std Resid	
10	0.8900	0.8418	0.0482	2.57	R
15	0.5120	0.4597	0.0523	2.12	R

R Large residual

	C1-T	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Team	Win%	O-FG%	O-3pt%	O-FT%	O-RB	O-Ast	O-TO	O-Stl	O-Blk	O-PF	O-Pts
4	Charlotte Hornets	0.585	43.9	36.2	79.0	43.9	21.7	11.9	7.3	5.3	18.1	103.4
5	Chicago Bulls	0.512	44.1	37.1	78.7	46.3	22.8	13.3	6.0	5.7	18.8	101.6
6	Cleveland Cavaliers	0.695	46.0	36.2	74.8	44.5	22.7	12.9	6.7	3.9	20.3	104.3
7	Dallas Mavericks	0.512	44.4	34.4	79.4	43.1	22.1	12.3	6.8	3.7	19.5	102.3
8	Denver Nuggets	0.402	44.2	33.7	76.6	44.6	22.7	14.2	7.4	4.8	21.0	101.9
9	Detroit Pistons	0.537	43.9	34.4	66.8	46.3	19.4	13.0	7.0	3.7	19.0	102.0
10	Golden State Warriors	0.890	48.7	41.6	76.3	46.2	28.9	14.9	8.4	6.1	20.7	114.9
11	Houston Rockets	0.500	45.2	34.7	69.4	43.1	22.2	15.2	10.0	5.2	21.8	106.5
12	Indiana Pacers	0.549	45.0	35.1	76.4	44.2	21.2	14.3	9.0	4.8	20.0	102.2
13	Los Angeles Clippers	0.646	46.5	36.4	69.2	42.0	22.8	12.4	8.6	5.6	21.3	104.5
14	Los Angeles Lakers	0.207	41.4	31.7	78.1	43.0	18.0	13.0	7.2	4.1	20.3	97.3
15	Memphis Grizzlies	0.512	44.0	33.1	78.3	41.6	20.7	12.7	8.8	4.3	21.7	99.1
16	Miami Heat	0.585	47.0	33.6	74.4	44.1	20.8	13.2	6.7	6.5	18.3	100.0
17	Milwaukee Bucks	0.402	46.7	34.5	74.7	41.7	23.1	14.6	8.2	5.8	20.7	99.0
18	Minnesota Timber>>	0.354	46.4	33.8	79.2	41.6	23.4	14.4	8.0	4.6	20.7	102.4

Figure 9 Evaluate Model Adequacy

In addition to the main effect, the interaction terms are also considered further in the multiple regression model. Two-way interaction terms are included and the step-regression model was conducted due to the 30 data cases available. Only one significant 2-way interaction term with P-Value < 0.3 as shown in Figure 10.

Regression Analysis: Win% versus O-3pt%_1, O-TO_1, O-Pts_1, D-FG%_1, D-RB_1, D-TO_1, ...

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	10	0.814449	0.081445	90.99	0.000
O-3pt%_1	1	0.003194	0.003194	3.57	0.074
O-TO_1	1	0.006727	0.006727	7.52	0.013
O-Pts_1	1	0.056375	0.056375	62.98	0.000
D-FG%_1	1	0.001739	0.001739	1.94	0.179
D-RB_1	1	0.001778	0.001778	1.99	0.175
D-TO_1	1	0.003828	0.003828	4.28	0.053
D-Pts_1	1	0.009000	0.009000	10.05	0.005
C-Ast_1	1	0.001519	0.001519	1.70	0.208
C-Stl_1	1	0.001962	0.001962	2.19	0.155
D-FG%_1*D-Pts_1	1	0.001235	0.001235	1.38	0.255
Error	19	0.017008	0.000895		
Total	29	0.831457			

Figure 10: Consider Interaction Term

Authors won't consider adding the quadratic terms to model simple (parsimony). A parsimonious [15] model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible. After built the reduced model, residual analysis was conducted to ensure model is adequate as shown in Figure 11. The predictive regression model has residuals which are normal

distribution (Normal Probability Plot, Histogram) in the left-hand side, with equal variance (upper right), and independent (lower right) [16].

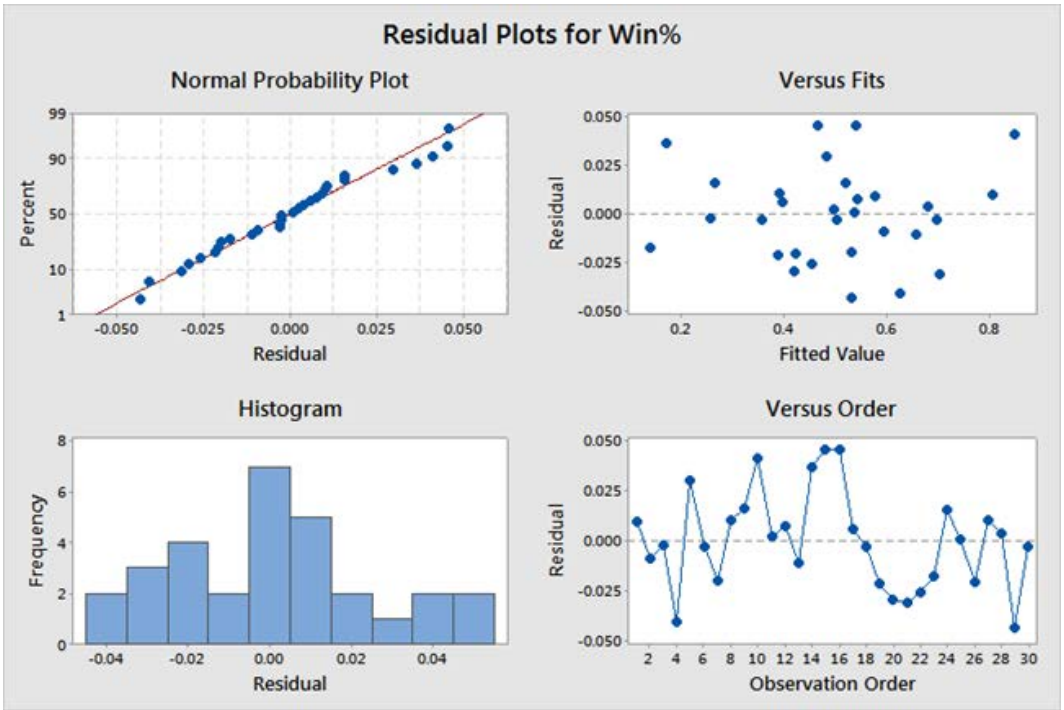

Figure 11: Residual Analysis

Figure 12 has listed the Predictive Regression Model Equation. Authors can use this equation and the top 9 factors in the team Offensive, Defensive and Differential Statistics to predict the Team Record Performance for any particular year.

```
Regression Equation

Win% = 0.50695 + 0.01586 O-3pt%_1 - 0.0427 O-TO_1 + 0.1075 O-Pts_1 - 0.0281 D-FG%_1
       - 0.0189 D-RB_1 + 0.0370 D-TO_1 - 0.0807 D-Pts_1 - 0.0144 C-Ast_1 - 0.0231 C-Stl_1
       - 0.01032 D-FG%_1*D-Pts_1
```

Figure 12: Predictive Regression Equation

In Figure 13, top three sensitive terms are listed for both offense and defense. Teams could enhance winning% by 10.75% if can increase team point average by just 1 point. It's not surprised that teams are finding more offense weapons. Reduce 1 "turn over" can get extra 4.27% winning chance. Since offense flow are much faster now, any careless turn over may cause opponents' fastscoring. The third offense term is 3-Point%. Each 3-pt % increase can earn 1.6% winning chance.

On the defense side, defensive points, defensive turn over, and defensive field goal are top three sensitive terms. As compared to offensive, the defense sensitivities are 15%-20% lower than the offense sensitivity on top two terms. There is significant VIF concern between defense points and defense field goals. This data may support the current trending: offense may be more critical than defense.

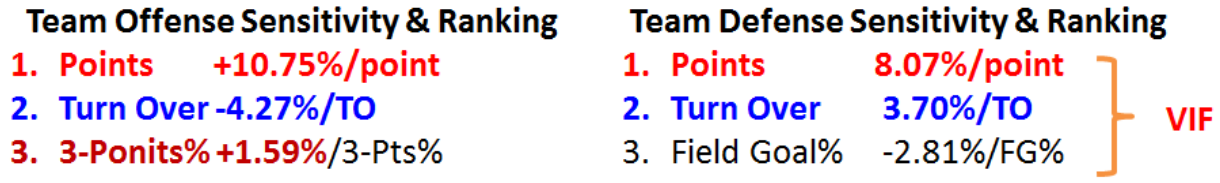

Figure 13: Top three sensitive terms in both Team Offense and Team Defense

In Figure 14, the top three offense and defense terms are listed across top six 2015-2016 NBA Teams in the regular season. In general, most top teams were ranked well in Offense categories. 73-Wins Warriors team performed best on Offense Points and 3-Point% with significant margin from the 2nd best team on these two categories. The 2nd best Spurs team performed well across all categories. Cavaliers were good on offense but weaker on defense. Clippers showed good ranking but winning % record. This data is consistent with the model prediction.

Ranking		Offense			Defense		
Team	Win%	Points	Turn Over	3-Points%	Points	Turn Over	Field Goal%
Warriors	0.890	1	25	1	19	15	3
Spurs	0.817	10	5	2	1	12	4
Cavaliers	0.695	8	8	8	4	22	14
Raptors	0.683	14	2	4	3	24	12
Thunders	0.671	2	27	17	15	26	5
Clippers	0.646	7	4	6	7	9	2

Figure 14: 2017-2018 Top Teams Offense and Defense

In Figure 15 modern NBA trending chart, (1) the average team scoring points were getting higher after 2004-2005 season. Before that season, the defense dominated the winning%. Champions like Spurs have emphasized defense in order to win more games. After 2004-2005, offense flow is becoming smoother and shorter. The increase was even more significant in the past two seasons 2016-2017, and 2017-2018 (3+ more points). The left chart also showed the more 3-pointers attempted after 2011-2012 season. The trend was even more significant after 2014-2015 season (the Warriors Dynasty Era). Both trendings are consistent with the model prediction on top offense sensitivity.

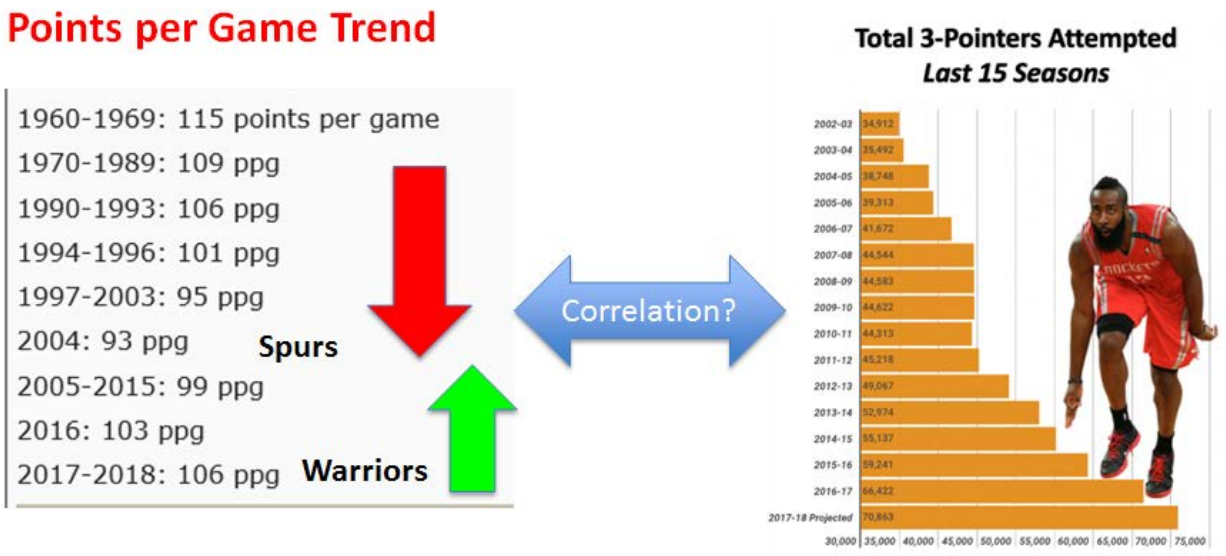

Figure 15: Modern NBA Trend

3.3 Validate the Predictive Model

After built the predictive model, authors have validated the model accuracy in Figure 13. The "Actual" column is team record performance (winning %) in the 2015-2016 season. The "Predicted" column is team record performance predicted by the regression model in Figure 12. It's amazing the delta (difference between actual winning % and the predicted winning %) is less than 5% across all 30 NBA teams.

	Team	Actual	Predicted	Delta
1	Golden State Warriors	0.890	0.849	-0.041
2	San Antonio Spurs	0.817	0.807	-0.010
3	Cleveland Cavaliers	0.695	0.698	0.003
4	Toronto Raptors	0.683	0.679	-0.004
5	Oklahoma City Thunder	0.671	0.702	0.031
6	Los Angeles Clippers	0.646	0.657	0.011
7	Atlanta Hawks	0.585	0.576	-0.009
8	Boston Celtics	0.585	0.594	0.009
9	Charlotte Hornets	0.585	0.626	0.041
10	Miami Heat	0.585	0.539	-0.046
11	Indiana Pacers	0.549	0.541	-0.008
12	Detroit Pistons	0.537	0.521	-0.016
13	Portland Trail Blazers	0.537	0.536	-0.001
14	Chicago Bulls	0.512	0.482	-0.030
15	Dallas Mavericks	0.512	0.532	0.020
16	Memphis Grizzlies	0.512	0.466	-0.046
17	Houston Rockets	0.500	0.498	-0.002

	Team	Actual	Predicted	Delta
16	Memphis Grizzlies	0.512	0.466	-0.046
17	Houston Rockets	0.500	0.498	-0.002
18	Washington Wizards	0.500	0.503	0.003
19	Utah Jazz	0.488	0.531	0.043
20	Orlando Magic	0.427	0.453	0.026
21	Denver Nuggets	0.402	0.391	-0.011
22	Milwaukee Bucks	0.402	0.396	-0.006
23	Sacramento Kings	0.402	0.423	0.021
24	New York Knicks	0.390	0.420	0.030
25	New Orleans Pelicans	0.366	0.388	0.022
26	Minnesota Timberwolves	0.354	0.357	0.003
27	Phoenix Suns	0.280	0.264	-0.016
28	Brooklyn Nets	0.256	0.258	0.002
29	Los Angeles Lakers	0.207	0.170	-0.037
30	Philadelphia 76ers	0.122	0.139	0.017

Figure 16: Validate Predictive Model

In Figure 14, the model accuracy is further displayed in the scatterplot (X: Actual Team Winning, Y: Predicted Team Winning). The offset is 0.01023 (around 1%), and the slope is 0.9795 (almost one). All the 30 team records (Even Warriors) are within 95% Prediction Interval (PI) [17] and the R-Sq (Adjusted) is 97.9% as calculated previously. The 95% prediction interval range is within 5% (s= standard deviation = 0.024 or 2.4%; 95% interval is around within +/- 2 Standard Deviations). No points were outside 95% Prediction Interval.

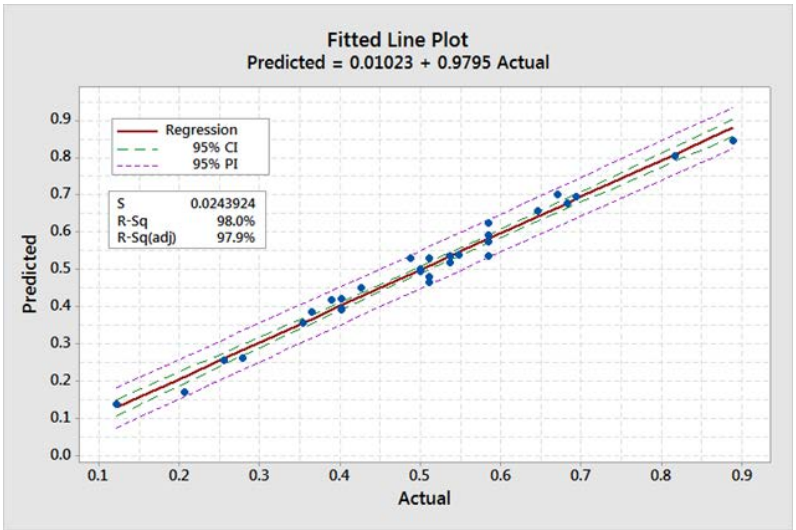

Figure 17: Model Accuracy

4. Conclusions

Authors have successfully built a predictive model which can predict the NBA Team Winning Record within 5% error based on the top 9 team statistics records. Authors have prepared a model flow chart to demonstrate the scope of the entire paper in Figure 15. Authors have addressed the model adequacy such as multi-collinearity (VIF), residual outliers, interaction terms, R-Sq vs. R-Sq (Adjusted), and residual analysis (normality, equal variance, independency). This paper has shown the power of applying the statistical regression model to predict the winning pattern. This paper can be further expanded to consider several situations such as player injury, match-up between two particular teams…

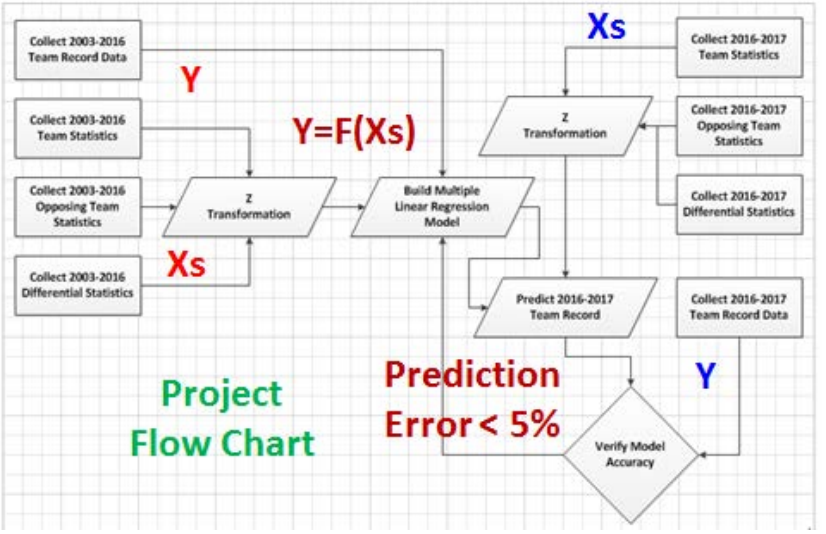

Figure 18: Model Flow Chart

Acknowledgment

Thanks to Dr. Ying Huang and Dr. Alex Wu for providing the regression statistics support. Thanks to Mr. Patrick Giuliano for revising our paper manuscript.

References (To be Added)

[1] Allan Maymin, Philip Maymin, and Eugene Shen (2012). "NBA chemistry: Positive and negative synergies in basketball". 2012 MIT Sloan Sports Analytics Conference.
[2] Min-hwan Oh, Suraj Keshri, and Garud Iyengar. (2015), "Graphical model for basketball match simulation". 2015 MIT Sloan Sports Analytics Conference.
[3] Mason Chen, Jason Li. (2017 April), "Predict Warriors' 73 Wins". 2017 IEOM Annual Conference Proceedings, Pages 614-625
[4] Mason Chen. (2017 November), "Predict NBA MVA", Frontiers in Artificial Intelligence and Applications, Volume 299 Fuzzy Systems and Data Mining III - Proceedings of FSDM 2017
[5] Anna Wu, Mason Chen, "Design Basketball Game and Rules", (2018 IEOM Annual Conference Proceedings P.1284-1289
[6] John, Forese, Jessica Gelman, Dan Reed, Michael Lorenc, and Ben Shields. (2016), "Modern NBA Coaching: Balancing Team and Talent". 2016 MIT Sloan Sports Analytics Conference.
[7] Avery McIntyre, Joel Brooks, John Guttag, Jenna and Wiens. (2016), "Recognizing and Analyzing Ball screen Defense in NBA". 2016 MIT Sloan Sports Analytics Conference.
[8] Joseph Kuehn. (2016), "Accounting for Complementary Skill Sets when Evaluating NBA players' Values to a Specific Team". 2016 MIT Sloan Sports Analytics Conference.
[9] NBA Team Statistics http://www.foxsports.com/nba/stats?category=SCORING
[10] NBA Team Record http://www.basketball-reference.com/leagues/NBA_2016_standings.html

[11] E. Kreyszig (1979). "Advanced Engineering Mathematics (Fourth ed.)". Wiley. p. 880, eq. 5. ISBN 0-471-02140-7.
[12] E. Kreyszig (1979). "Advanced Engineering Mathematics (Fourth ed.)". Wiley. p. 880, eq. 6. ISBN 0-471-02140-7.
[13] Allison, P. D. (1999). "Multiple Regression: A Primer". Pine Forge Press. p. 142
[14] Theil, Henri (1961). "Economic Forecasts and Policy". Holland, Amsterdam: North
[15] Herbert W. Marsh, John Balla, (1994 May), "Quality and Quantity", Volume 28, Issue 2, pp 185–217
[16] Steel, Robert G. D.; Torrie, James H. (1960). "Principles and Procedures of Statistics, with Special Reference to Biological Sciences". McGraw-Hill. p. 288.
[17] Geisser 1993, p. 6: Seymour Geisser (1993). "Predictive Inference: Chapter 2: Non-Bayesian predictive approaches". CRC Press.