

NBA Point Spread Predictive Modeling

Zach Diamandis



April 2019

Abstract

This paper attempts to predict the point differential (point spread) between the home team and away team in National Basketball Association (NBA) games. Two datasets are used for analysis, with each row representing a single game - one set consists of 538's CARMelo team ratings, and the second consists of average team box score statistics. The goal of the predictive model is to gain an edge over the point spreads set by Vegas for betting purposes. The strongest performing model was a stepwise two way interaction linear regression model fit on the team box score data. This model returned a profit of 2060 when tested over 30 random testing games with a threshold difference of 1. The neural net model on this same data was also strong, returning a profit of 1750 over these games with the same threshold.

1 Introduction and Background

Since the recent striking down of federal gambling prohibitions by the Supreme Court (2018), online gambling has continued to increase in popularity. In 2018, the commercial casino business was valued at 261 billion [1]. For each game in the National Basketball Association (NBA), sports gambling groups - often known as sportsbooks - assign probabilities to various outcomes for the game. While many possible outcomes to bet on exist, there are three primary ones: the point spread, the over/under, and the moneyline. The moneyline is simply a bet on the winner of the match, and the over/under is a bet on the total number of points in the game. The point spread, which is the response variable of choice in this project, is the difference in the two teams' scores for a given game. In this paper's analysis, it will always be the home team score minus the away team score. Reproduced below is an example of a recent set of these betting options.

+ NBA		BOSTON CELTICS @ MILWAUKEE BUCKS		
🕒 Apr 30 8:05 PM		SPREAD	TOTAL	MONEY LINE
571	 BOSTON CELTICS	+7½-110 <input type="checkbox"/>	o219-110 <input type="checkbox"/>	+280 <input type="checkbox"/>
572	 MILWAUKEE BUCKS	-7½-110 <input type="checkbox"/>	u219-110 <input type="checkbox"/>	-340 <input type="checkbox"/>

From this graphic, one can see that the Milwaukee Bucks are favored by 7.5 points over the Boston Celtics. The estimated total score is 219 points, and the Celtics are given about a 26.3% chance to win based off their moneyline of +280, which indicates a bet of 100 wins 280. This can be calculated by dividing 100 by 100 plus the moneyline of 280 = $100/380$.

2 Data

The first primary consideration of this project was selection of relevant datasets to predict point spread. Two datasets were selected, which are explained further below. One dataset uses a team rating system, and the other relies on raw team-level data from individual NBA games. For all data sets the response variable is an individual game's point spread. Both datasets are split into 80% training and 20% testing data.

2.1 Dataset 1

The first dataset used is 538's 2018-2019 CARMELLO rating system dataset. This dataset was favored over 538's ELO rating system and 538's 2015-2018 CARMELLO systems as basic initial models showed it performed much better in predictive tasks.

538's CARMELLO team rating system began in 2015 and was similar to ELO systems in many other competitive sports - a net sum game in which any increase in a winning teams' score leads to the same decrease in the losing teams'. However, it was adjusted so that the initial ratings for a given season were set taking into account offseason trades. This is done by use of 538's CARMELLO player projections system. Teams are also adjusted for playoffs experience. However, this system also came up short in several areas. For the 2018-2019 season, the system was adjusted to be based entirely off the CARMELLO player projections of the active team roster. This system had much better results [2]. Thus, moving forward, only the 2018-2019 CARMELLO rating system is considered. The mean 2018-2019 CARMELLO rating is approximately 1496, and the standard deviation is 135.

2.1.1 Predictors

The CARMELLO rating system only consists of two predictors - a score for each team - but implicitly considers several more because of the player projections.

The weight of each player projection is determined by depth charts and expected playing time. The CARMELO player rating system examines:

- Skills: Height, Weight, Draft Position
- Scoring: Usage Rate, TS %, FT %, 3PR, FTR
- Handling: Assist Rate, Turnover Rate
- Defense: Rebound Rate, Block Rate, Steal Rate, BPM, RPM

TS % refers to true shooting percentage, FT % to free throw percentage, 3PR to three point rate, FTR to free throw rate, BPM to box plus minus, and RPM to real plus minus. This calculation yields an offensive and defensive team rating, which are combined into a winning percentage via Pythagorean expectation to give the CARMELO rating. For an in-depth look at this specific calculation, see Appendix I.

2.2 Dataset 2

While Dataset 1 considers player level statistics, the Dataset 2 consists of team level statistics. Dataset 2 draws from the a dataset collected on kaggle.com, an online community for data science work [5]. Since the team statistics for an individual game are not known before the game ends, previous games' statistics must be used to predict subsequent games. Thus, a 10-day moving average was used for each predictor. The 10 day window was favored over a slower window, such as 30, as the predictive power was higher in initial testing. This is likely because the shorter window can better account for injuries and trades.

2.2.1 Predictors

The Dataset 2 predictors are:

- Scoring: Field Goals, Field Goals Attempted, 3P, 3PA, FT, FTA
- Handling: Assists, Steals, Turnovers
- Defense: ORB, DRB, Blocks, Total Fouls

Where 3P is the number of three pointers, ORB is offensive rebounds, and DRB is defensive rebounds. These predictors are recorded for both the home and away team.

Below, in **Table 1**, are a summary of some of the most important predictors in Dataset 2.

Table 1: Predictor Summary

	FieldGoals_x	X3PointShots_x	FreeThrows_x	Steals_x	Blocks_x
X	Min. :31.40	Min. : 2.600	Min. : 9.90	Min. : 3.700	Min. :2.000
X.1	1st Qu.:37.00	1st Qu.: 7.700	1st Qu.:15.50	1st Qu.: 6.900	1st Qu.:4.100
X.2	Median :38.60	Median : 9.100	Median :17.30	Median : 7.700	Median :4.700
X.3	Mean :38.61	Mean : 9.128	Mean :17.34	Mean : 7.757	Mean :4.829
X.4	3rd Qu.:40.10	3rd Qu.:10.500	3rd Qu.:19.00	3rd Qu.: 8.600	3rd Qu.:5.500
X.5	Max. :48.30	Max. :17.900	Max. :27.90	Max. :13.200	Max. :9.500

3 Methods

3.1 Overview

For each dataset, an initial baseline least squares regression is fit. Then, a random forest model is used. For Dataset 1, the rating system dataset, 538’s difference formula, explained below, is also used. Finally, for Dataset 2, a neural net model will be fit. Plots of the fitted vs. actual values for all six models are provided in Appendix II.

3.2 Dataset 1 Models

3.2.1 Least Squares Model

Least squares regression minimizes the sum of squared residuals. There is an assumption of an underlying linear relationship, and the games are assumed to be independent. The response is normally distributed and no major heteroskedasticity is visible. For this dataset, least squares regression yields the following:

$$\text{ScoreDiff} = \underset{(3.54)}{-10} + \underset{(.00167)}{.038} \cdot \text{HomeCarmelo} - \underset{(.00168)}{.030} \cdot \text{AwayCarmelo}$$

Both the home and away ratings are highly significant predictors ($p < 2e-16$), as expected. The impact of being home is seen by the larger, positive coefficient of HomeCarmelo compared to AwayCarmelo. The intercept is also significant and offsets this effect slightly with a negative sign.

3.2.2 Random Forest

The random forest model creates a multitude of split decision trees and outputs the mean prediction of these trees. The number of variables considered at each split was one. The only assumption is that sampling is representative, which it is assumed to be because individual NBA games do not vary greatly. Random forest regression on dataset 1 gives the following feature importances:

The random forest model shows that the removal of the HomeCarmelo variable would lead to more error than the removal of the AwayCarmelo variable, implying a greater importance and the significant effect of being home.

	IncMSE	IncNodePurity
HomeCarmelo	64.33	123571.2
AwayCarmelo	42.03	112081.3

3.2.3 Difference Formula

538 provides a simple formula with their analysis, shown below:

$$\text{ScoreDiff} = \frac{(\text{HomeCarmelo} - \text{AwayCarmelo} + 100)}{28}$$

This model assumes a fixed positive effect of 100/28 for the home team and uses the difference of the two CARMELO ratings. It is apparent from this formula that the CARMELO system would have difficulty predicting upsets.

3.3 Dataset 2 Models

3.3.1 Least Squares Stepwise Regression

Again, with the same first two assumptions taken for granted, the four assumptions of least squares regression are well met. Here, forward stepwise regression from an intercept model to a full two way interaction model is conducted by AIC criterion. This model was chosen over the ordinary least squares as several significant interaction variables were found. Some of the most significant are shown below in **Table 2**. The H denotes the home team, and the A denotes the away team.

Table 2: Significant Interactions

Statistics	Estimate	Std. Error	Pr($\geq t$)
FieldGoals.H:Turnovers.H	0.13235	0.03757	0.000430
FieldGoals.H:Blocks.H	0.14636	0.05190	0.004817
FieldGoals.H:X3PointShotsAttempted.H	0.04773	0.01607	0.002992
Steals.H:OffRebounds.A	-0.23029	0.07517	0.002195
DefensiveRebounds.H:Blocks.A	0.16968	0.06473	0.008783
OffRebounds.H:X3PointShotsAttempted.H	0.06035	0.02052	0.003286
X3PointShots.A:TotalFouls.A	-0.10923	0.03972	0.005972
OffRebounds.A:Turnovers.A	-0.15765	0.05463	0.003915

3.3.2 Random Forest

The most important features from the team data random forest model are summarized below. The number of variables considered to split a node was 5. This hyper-parameter was found through random search of values between 1 and 10.

	IncMSE	IncNodePurity
FieldGoals.H	6.377034	60590.41
X3PointShots.H	12.709596	58059.89
X3PointShotsAttempted.H	10.287064	55628.83
Assists.H	5.449298	61947.64
FieldGoals.A	6.572010	58009.70
X3PointShots.A	12.143188	57300.40
X3PointShotsAttempted.A	9.713036	55830.51
Assists.A	5.188024	61463.40

3.3.3 Neural Net

A Bayesian regularized neural net was used for this model (BRNN). Neural net models are comprised of many interconnected processing elements, referred to as neurons. The output of each neuron is computed by a non-linear function of the sum of its inputs. Neural nets are also adaptive models, such that the weights of neurons and inputs can be adjusted. The neurons are typically separated into several layers.

BRNN models are more robust than standard models, and use a method similar to ridge regression to convert a non-linear regression into a “well-posed statistical problem” [4]. Neural net models are difficult to visualize, as it serves as somewhat of a black-box algorithm. It does not rely on any underlying pattern assumption in the data. The number of neurons in the hidden layer was tuned to 4 via random search of values between 1 and 15. The sum of the squares of the biases and weights is 11.99. The sum of squared errors is 463.78. For a plot of this model, see Appendix III.

4 Results

4.1 Out of Sample Performance

The six models listed above were tested for performance on the training data, measured by RMSE. The results are summarized below, in **Table 3**

4.2 2019 Playoff Performance

From the above, the strongest Dataset 1 model is the ordinary least squares, although the 538 difference formula is extremely close. The strongest Dataset 2 models are the stepwise least squares and the neural net. For illustration, the predictions of these models on the first round of the playoffs, with comparisons to vegas, is shown below in **Table 4**.

Table 3: Holdout Performance

Model	RMSE Holdout
CARMELO LSR	13.59
CARMELO RF	14.5
CARMELO 538Diff	13.62
TeamData Stepwise LSR	12.91
TeamData RF	13.20
TeamDats NN	13.08

Table 4: First Round 2019 Playoffs

	CARMELO LSR	Step LSR	NN	Vegas	PointSpread(H-A)
DET at MIL(s)	+12.55	+4.36	+9.67	+13	+35
ORL at TOR(s)	+11.26	+2.28	+5.72	+9.5	-3
LAC at GSW(s)	+14.86	+11.08	+13.42	+13.5	+17
SAS at DEN(s)	+7.52	+5.03	+3.29	+5.5	-5
UTA at HOU(s)	+5.11	+2.12	+5.01	+6	+32
OKC at POR(s)	-.12	+.69	+5.12	+3.5	+5
BKN at PHI(s)	+8.98	+.44	+4.58	+8	-9
IND at BOS(s)	+6.23	+5.47	+8.40	+7	+10
RMSE	17.27	16.29	14.64	14.74	

4.3 Betting Performance

Since the purpose of creating this predictive model is to yield profit, evaluation versus vegas predictions is required. For point spread betting, vegas odds are set at -110, or 10:11. This means that 11/21, or 52.4% of bets must hit for profitability. The two best performing models, the stepwise and neural net, were evaluated on 30 randomly selected games from out of sample data on various betting thresholds. The vegas lines were retrieved by a manual search, primarily from <https://sportschatplace.com/> [3]. The threshold describes the difference required between the model’s prediction and vegas’s prediction for a bet to be placed. Results are summarized in **Table 5 Table 6**. A full list of both model’s predictions, the vegas line, and the real point spread for all 30 games is listed in appendix IV.

Table 5: 30 game Profit vs. Vegas

Threshold	1	2	3	4	5
Profit.NN	1750	1760	1460	1460	1360
Win/Loss.NN	23-5	22-4	19-4	19-4	18-4
Profit.Step	2060	1760	1460	1160	960
Win/Loss.Step	25-4	22-4	19-4	16-4	14-4

Table 6: 30 game RMSE vs. Vegas

Model	RMSE Holdout
Step	12.57
NN	12.70
Vegas	15.98

5 Conclusion

A purely quantitative analysis modeled through a neural net and stepwise linear regression out-competes Vegas predictions in terms of RMSE and profitability, achieving more accurate predictions on 30 test games. The most profitable betting model/threshold combination appears to be the stepwise model with a threshold of 1. However, in general the neural net and stepwise models made very similar predictions, as seen by the identical record under thresholds of 2 and 3.

Through such modeling, it is possible to gain an edge over the point spread predictions set by Vegas and return a profit. This finding may imply that the prediction lines set by Vegas reflect non-quantitative factors, such as public opinion or market popularity, to some degree.

References

- [1] Global Betting and Gaming Consultants. GBGC’s Global Gambling Report 2018 (2018).
- [2] <https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/>
- [3] <https://sportschatplace.com/>
- [4] Burden, F., Winkler, D. (2008). Bayesian regularization of neural networks. *Methods in Molecular Biology (Clifton, N.J.)*, 458, 25–44.
- [5] <https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018/version/5>

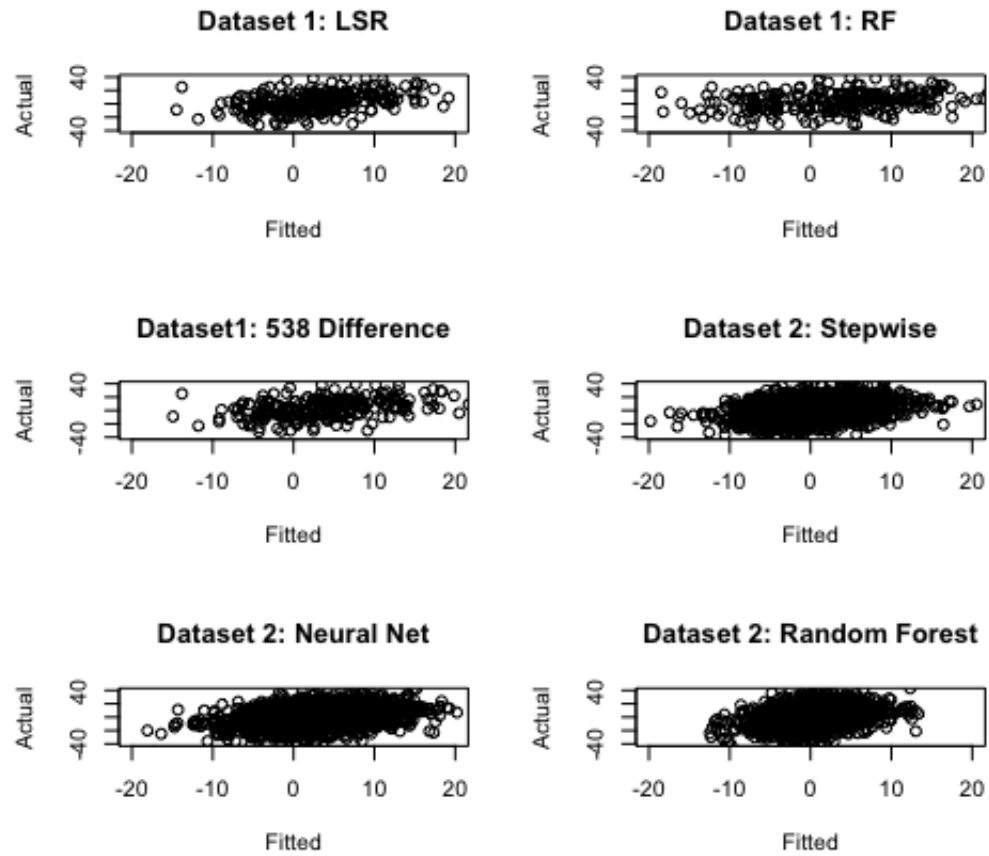
6 Appendix

6.1 Appendix I

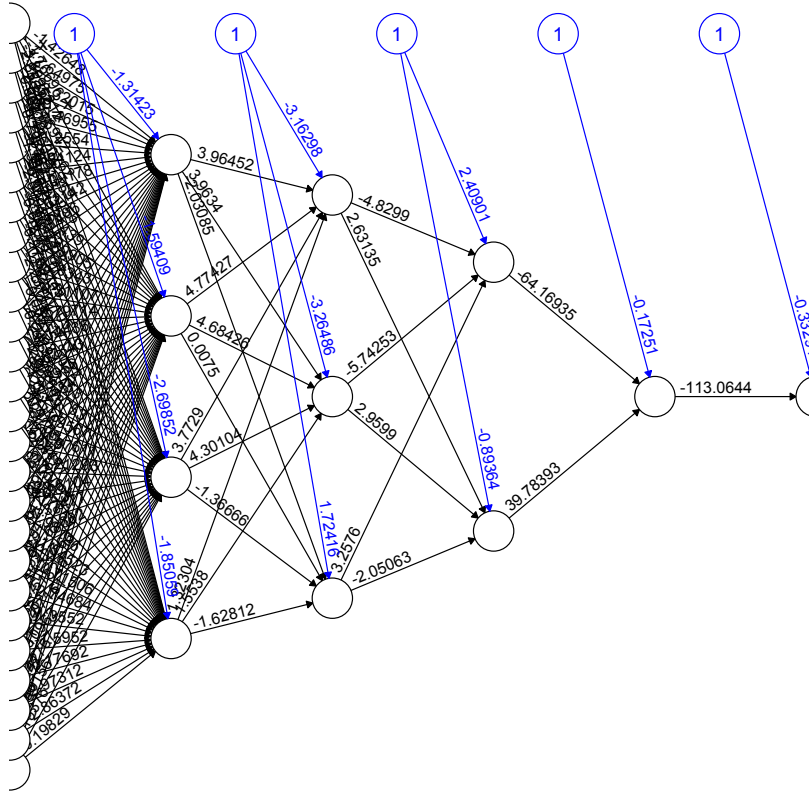
The mentioned statistics are used to create a player rating for each player. The player ratings are then combined by depth charts and manual estimation of expected usage rate. Then, Pythagorean expectation yields Winning Percentage, which leads to the CARMELLO combined rating by the below two formulas [1].

$$WP = \frac{(108+OR)^{14}}{(108+OR)^{14}+(108-DR)^{14}} \quad CARMELLO = 1504.6 - 450 \log \frac{1-WP}{WP}$$

6.2 Appendix II



6.3 Appendix III



6.4 Appendix IV

	StepwiseLSR	NeuralNet	Vegas	Actual
1	-2.76	2.65	-6.00	-4.00
2	-2.77	-0.70	14.50	5.00
3	-1.76	2.09	4.00	-13.00
4	16.43	16.82	-10.50	-21.00
5	0.80	5.06	4.00	-12.00
6	9.56	14.26	-10.50	10.00
7	-3.32	1.88	7.00	2.00
8	6.45	13.27	5.00	13.00
9	-5.02	-3.63	8.00	-28.00
10	3.73	4.24	-6.00	8.00
11	-1.13	-0.98	9.50	-24.00
12	7.69	11.00	3.00	5.00
13	1.89	5.31	-3.00	2.00
14	13.88	15.55	-4.50	16.00
15	3.72	5.73	-2.50	-5.00
16	1.08	0.29	9.50	-4.00
17	5.67	7.08	4.50	10.00
18	-7.14	-3.90	5.00	-17.00
19	0.09	-0.56	1.50	-19.00
20	4.33	7.92	-8.00	9.00
21	-0.18	0.84	-8.50	4.00
22	-4.75	-4.76	3.50	5.00
23	-0.22	0.89	5.00	-9.00
24	1.59	5.17	5.00	-3.00
25	-2.05	-0.42	0.00	-3.00
26	1.70	5.03	-1.00	22.00
27	4.75	3.57	10.00	11.00
28	0.45	-6.18	3.00	-20.00
29	1.88	3.69	1.00	13.00
30	8.17	10.52	-12.50	3.00