NBA2023-2024 Dataset Guidelines

Shirong Xu

November 2024

Dataset Summary

The dataset contains 2,460 entries and 24 columns, with the following structure:

Column	**Description**
Team	Team name (e.g., GSW, PHX, LAL)
Match Up	Description of the match (e.g., GSW vs. PHX or GSW @ PHX). If GSW vs. PHX, then GSW is home team and PHX is guest team. If GSW @ PHX, then GSW is guest team and PHX is home team.
Game Date	Date of the game (e.g., 10/24/2023)
W/L	Win or Loss indicator (W or L)
MIN	Minutes played
PTS	Points scored
FGM	Field goals made
FGA	Field goals attempted
FG%	Field goal percentage
3PM	Three-point field goals made
3PA	Three-point field goals attempted
3P%	Three-point field goal percentage
FTM	Free throws made
FTA	Free throws attempted
FT%	Free throw percentage (some entries are non-numeric)
OREB	Offensive rebounds
DREB	Defensive rebounds
REB	Total rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal fouls
+/-	Plus/minus statistic

Table 1: Dataset Structure

Here's a quick look at the first few rows of the dataset:

Team	Match Up	Game Date	W/L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-
GSW	GSW vs. PHX	10/24/2023	L	240	104	36	101	35.6	10	43	23.3	22	28	79	18	31	49	19	11	6	11	23	-4
PHX	PHX @ GSW	10/24/2023	W	240	108	42	95	44.2	11	33	33.3	13	17	77	17	43	60	23	5	7	19	22	4
LAL	LAL @ DEN	10/24/2023	L	240	107	41	90	45.6	10	29	34.5	15	20	75	13	31	44	23	5	4	12	18	-12
DEN	DEN vs. LAL	10/24/2023	W	240	119	48	91	52.7	14	34	41.2	9	12	75	9	33	42	29	9	6	12	15	12

Figure 1: Examples in the dataset. The **first two rows** correspond to **the game between GSW and PHX**. The first row shows that GSW, the home team, scored 104 points but lost the game. This row contains all the game data for GSW. The second row presents the information for the visiting team, PHX, which played against GSW.

Objective

A natural task in this dataset is to predict the outcome of the game (Win or Loss) based on historical data. For example,

NYK	NYK @ BOS	11/13/2023	L
BOS	BOS vs. NYK	11/13/2023	W

Figure 2: The game between NYK and BOS. In this game, BOS is the home team and NYK is the visiting team. The outcome is BOS wins the game. The question is *Can we predict this outcome based on the data of two teams* ***before Nov 13, 2023****.*

Therefore, we can simply analyze the game data of NYK and BOS before Nov 13, 2023.

Team	Match Up	Game Date	W/L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-
NYK	NYK vs. BOS	10/25/2023	L	240	104	36	97	37.1	18	41	43.9	14	26	53.8	17	30	47	24	9	0	11	22	-4
NYK	NYK @ ATL	10/27/2023	W	240	126	43	90	47.8	20	44	45.5	20	28	71.4	12	33	45	30	9	6	16	21	6
NYK	NYK @ NOP	10/28/2023	L	240	87	33	90	36.7	7	37	18.9	14	18	77.8	16	43	59	19	7	3	19	17	-9
NYK	NYK @ CLE	10/31/2023	W	240	109	38	86	44.2	13	34	38.2	20	25	80	12	36	48	18	9	1	14	17	18
NYK	NYK vs. CLE	11/01/2023	L	240	89	32	92	34.8	5	30	16.7	20	30	66.7	16	34	50	18	10	3	14	16	-6
NYK	NYK @ MIL	11/03/2023	L	240	105	38	96	39.6	10	39	25.6	19	25	76	16	40	56	18	6	1	11	22	-5
NYK	NYK vs. LAC	11/06/2023	W	240	111	41	88	46.6	12	31	38.7	17	18	94.4	18	30	48	28	13	1	20	13	14
NYK	NYK vs. SAS	11/08/2023	W	240	126	46	100	46	19	42	45.2	15	22	68.2	13	34	47	28	7	0	4	14	21
NYK	NYK vs. CHA	11/12/2023	W	240	129	47	87	54	15	36	41.7	20	30	66.7	10	30	40	25	8	4	9	18	22
NYK			0.55	240	109.6	39.33	91.78	42.98	13.22	37.11	34.93	17.67	24.67	72.78	14.44	34.44	48.89	23.11	8.667	2.111	13.11	17.78	6.333
BOS	BOS @ NYK	10/25/2023	W	240	108	37	77	48.1	12	39	30.8	22	26	84.6	7	39	46	18	6	11	13	22	4
BOS	BOS vs. MIA	10/27/2023	W	240	119	45	95	47.4	16	39	41	13	19	68.4	16	39	55	20	7	6	15	19	8
BOS	BOS @ WAS	10/30/2023	W	240	126	51	102	50	19	53	35.8	5	7	71.4	15	36	51	31	11	6	18	21	19
BOS	BOS vs. IND	11/01/2023	W	240	155	54	95	56.8	20	35	57.1	27	28	96.4	11	46	57	27	5	2	13	19	51
BOS	BOS @ BKN	11/04/2023	W	240	124	43	90	47.8	15	45	33.3	23	27	85.2	10	40	50	22	4	6	11	17	10
BOS	BOS @ MIN	11/06/2023	L	265	109	36	92	39.1	11	39	28.2	26	34	76.5	11	34	45	20	13	2	18	26	-5
BOS	BOS @ PHI	11/08/2023	L	240	103	36	91	39.6	15	47	31.9	16	19	84.2	9	34	43	28	8	6	13	13	-3
BOS	BOS vs. BKN	11/10/2023	W	240	121	41	94	43.6	19	52	36.5	20	28	71.4	17	35	52	29	5	3	8	13	14
BOS	BOS vs. TOR	11/11/2023	W	240	117	47	86	54.7	15	44	34.1	8	13	61.5	5	41	46	30	1	4	13	14	23
			0.77		120.2	43.33	91.33	47.46	15.78	43.67	36.52	17.78	22.33	77.73	11.22	38.22	49.44	25	6.667	5.111	13.56	18.22	13.44

Figure 3: Before November 13, 2023, both the New York Knicks (NYK) and the Boston Celtics (BOS) played 9 games, with winning probabilities of 0.55 and 0.77, respectively. Furthermore, the statistics for BOS almost entirely dominate those of NYK. For instance, BOS averaged 120.2 points in their last 9 games, while NYK averaged only 109.6 points. Additionally, NYK played against BOS on October 25 and lost, which serves as another strong indicator for predicting the outcome of their upcoming game.

Based on the historical data before November 13, it seems reasonable to conclude that BOS would win the game on November 13 against NYK, which aligns with the actual outcome. Based on a similar idea, I created a dataset where, for each game, the features are constructed as **the difference in averaged game statistics** before the current game date, and the label is the current game result. Here the averaged game statistics basically reflects the capability of a team.

Additionally, I also take the home advantage into consideration by introducing a binary variable. The prediction testing accuracy (training-testing split 70-30) is roughly 67% by a random forest classifier. To improve this result, I then consider using **weighted averaged game statistics**. Specifically, for the game on November 13 between BOS and NYK, we should assign higher weights to the game statistics that occur closer to November 13. With an appropriate weighting scheme, the testing accuracy may improve to 70%. For this dataset, you may use 67% and 70% as baselines, and I look forward to seeing better results in your application.

Guidelines

I think the classification task on this dataset will be highly sensitive to the features you construct. If you can construct more informative features, the prediction performance will be improved significantly. In the following, we provide several possible improvements or possible feature engineering steps you can try.

1 **Home advantage**. It is always expected that home team will perform better compared to they do as a visiting team.

2 **Stability**. In the above example, I only consider the averaged game statistics before each game occurs. However, some teams could be highly unstable, which, to some extent, reflects their inability.

3 **Previous Competitions**. For example, if Team A and Team B played 3 games before, and Team A won all of them, but when calculating the averaged statistics before the 4th game, Team B appears to be better than Team A, it raises an interesting question about prediction. In this case, we face a choice between trusting empirical evidence (past wins) or relying on current statistics.

4 **Weighting Statistics**. I found that assigning higher weights to those games occur recently will benefit predicting the outcomes. The question is how to construct the weighting function.

5 You may consider more.

When analyzing this dataset, the key question to consider is how to collect data about the two teams before the event at time point A and integrate it into meaningful features in order to predict the outcome of the game. When creating features, they must be **logical and based on common sense**. For example, in a game between NYK and BOS, if we find that NYK's average score is lower than BOS's, we can use the point difference as a feature for prediction, as it reflects the scoring ability gap between the two teams. Finally, you need to try different models and strictly follow the process to select the best predictive model. For instance, you can use K-fold cross-validation to select the best Random Forest model.