NBA Game Wins Prediction

SPA DRP Spring 2023

NBA Analytics and Machine Learning

Haoquan Fang

Introduction

- The National Basketball Association, or NBA, a professional basketball league
- We have 30 teams in the league
- We have 82 games for each team in each season

source: https://www.espn.com/nba/

Predict the number of game wins a team will have in a particular season using its performance statistics in the previous season with Machine Learning Algorithms

NATHAN LAUGA · UPDATED 5 MONTHS AGO

Data Collection

NBA games data

Dataset with all NBA games from 2004 season to dec 2020

source: https://www.kaggle.com/datasets/nathanlauga/nba-games

- Online open-source dataset from kaggle, generated by using webscrapping from NBA stats website

Table Name	Description
games.csv	all games from 2004 season to 2022, includes game dates, seasons, ids, home teams, visitor teams, and total number of points scored by both teams
games_details.csv	all statistics of players for a given game, including Field Goals Made/Attempt, Free Throws Made/Attempt, Offensive/Defensive Rebounds, etc.
players.csv	player details, including corresponding name and team
ranking.csv	ranking of NBA given a day, include the team wins and losses at a specific date and season
teams.csv	team details, including team name, nickname, abbreviation, location, start year, etc.

Data Preprocessing - Labels

ranking.csv

	TEAM_ID	LEAGUE_ID	SEASON_ID	STANDINGSDATE	CONFERENCE	TEAM	G	W	L	W_PCT
0	1610612747	0	22019	2020-02-03	West	L.A. Lakers	48	37	11	0.771
1	1610612746	0	22019	2020-02-03	West	LA Clippers	50	35	15	0.700
2	1610612743	0	22019	2020-02-03	West	Denver	50	34	16	0.680
3	1610612762	0	22019	2020-02-03	West	Utah	49	32	17	0.653
4	1610612745	0	22019	2020-02-03	West	Houston	49	31	18	0.633

- Select the latest date of every season
- Get the number of game wins for every team in every season
- Drop the rows if the total games played is not 82

labels

	TEAM_ID	SEASON_ID	W
0	1610612737	2002	35
1	1610612737	2003	28
2	1610612737	2004	13
3	1610612737	2005	26
4	1610612737	2006	30
...	...	...	...
501	1610612766	2015	48
502	1610612766	2016	36
503	1610612766	2017	36
504	1610612766	2018	39
505	1610612766	2021	43

Data Preprocessing - Features

games_details.csv

	GAME_ID	TEAM_ID	TEAM_ABBREVIATION	TEAM_CITY	PLAYER_ID	PLAYER_NAME
0	21900741	1610612753	ORL	Orlando	1628411	Wes Iwundu
1	21900741	1610612753	ORL	Orlando	203932	Aaron Gordon
2	21900741	1610612753	ORL	Orlando	202696	Nikola Vucevic
3	21900741	1610612753	ORL	Orlando	203095	Evan Fournier
4	21900741	1610612753	ORL	Orlando	1628365	Markelle Fultz

...	OREB	DREB	REB	AST	STL	BLK	TO	PF	PTS	PLUS_MINUS
...	0.0	2.0	2.0	2.0	1.0	0.0	0.0	1.0	9.0	2.0
...	2.0	10.0	12.0	5.0	2.0	0.0	1.0	2.0	16.0	13.0
...	3.0	4.0	7.0	5.0	0.0	0.0	1.0	1.0	22.0	11.0
...	0.0	3.0	3.0	1.0	0.0	0.0	3.0	1.0	17.0	23.0
...	0.0	1.0	1.0	14.0	2.0	0.0	2.0	3.0	12.0	9.0

- Sum up all player statistics in all game in one season. Regard this as the team statistics in one season.
- We don’t average the statistics become some players’ statisitcs will skew the data pretty much.
- Drop rows with NaN values

Data Preprocessing - Features

features (19 in total)

	TEAM_ID	NEXT_SEASON	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	FTM	FTA	...	DREB	REB	AST	STL
0	1610612737	2004	2857.0	6609.0	320.764	422.0	1256.0	104.083	1555.0	2004.0	...	2543.0	3548.0	1666.0	633.0
1	1610612737	2005	2997.0	6821.0	357.705	309.0	997.0	114.160	1456.0	2059.0	...	2387.0	3510.0	1649.0	642.0
2	1610612737	2006	3196.0	6997.0	358.522	450.0	1205.0	138.741	1804.0	2404.0	...	2427.0	3568.0	1759.0	633.0
3	1610612737	2007	3105.0	6986.0	369.110	385.0	1161.0	127.166	1865.0	2438.0	...	2528.0	3595.0	1728.0	663.0
4	1610612737	2008	3457.0	7668.0	382.692	436.0	1250.0	122.879	2129.0	2758.0	...	2885.0	4074.0	2084.0	696.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
442	1610612766	2015	3191.0	7612.0	370.551	551.0	1736.0	154.944	1547.0	2057.0	...	3070.0	3965.0	1833.0	552.0
443	1610612766	2016	3528.0	8122.0	425.323	987.0	2745.0	266.812	1859.0	2352.0	...	3403.0	4246.0	2027.0	691.0
444	1610612766	2017	3318.0	7589.0	387.285	866.0	2526.0	188.367	1741.0	2146.0	...	3111.0	3904.0	2024.0	621.0
445	1610612766	2018	3386.0	7537.0	390.534	872.0	2379.0	212.096	1738.0	2334.0	...	3078.0	3954.0	1869.0	594.0
446	1610612766	2021	3210.0	7117.0	352.732	1095.0	2976.0	211.146	1299.0	1729.0	...	2704.0	3559.0	2173.0	640.0

Data Preprocessing - Scaling

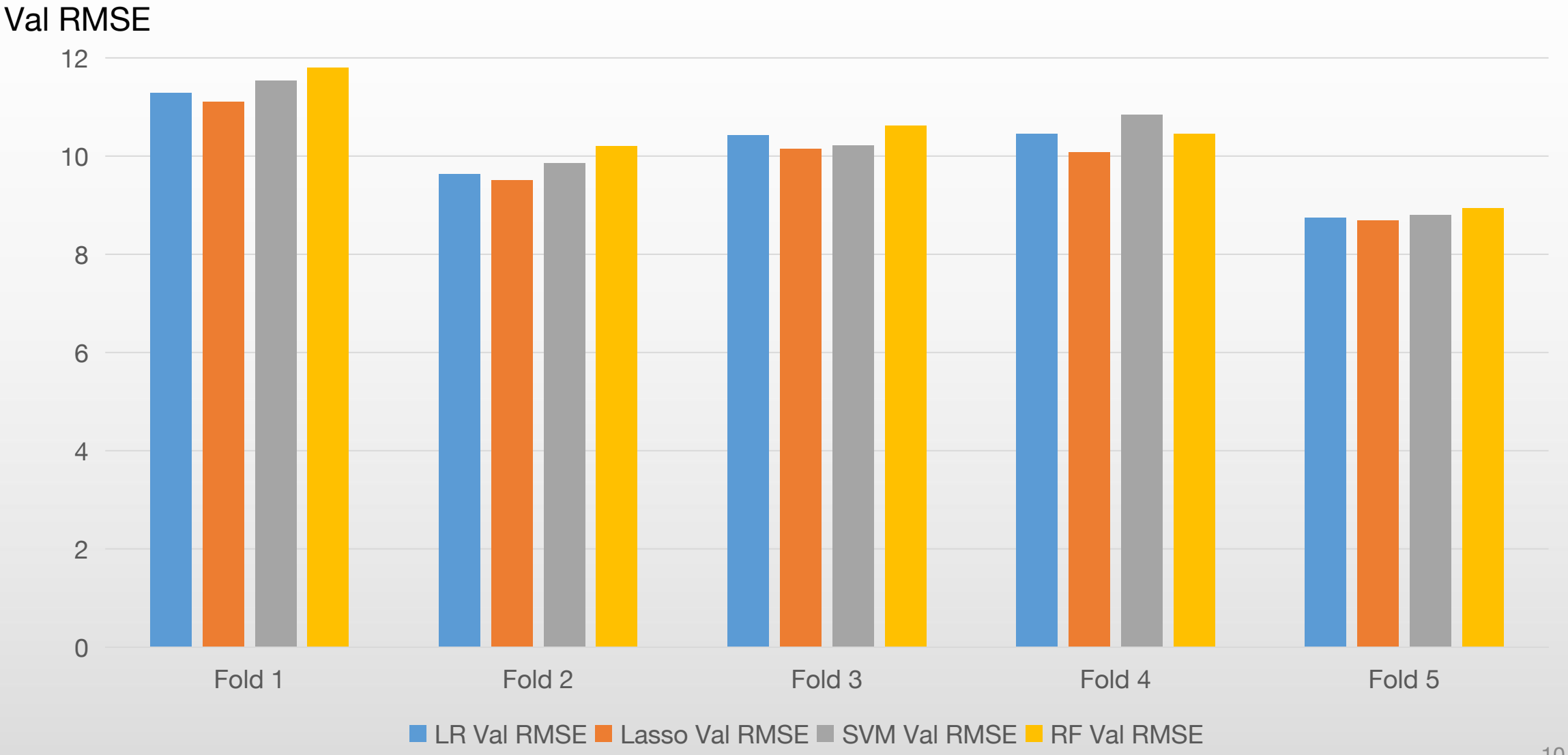
features (μ = 0, σ = 1)

	TEAM_ID	NEXT_SEASON	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	FTM	FTA	...	DREB
0	1610612737	2004	-1.411461	-1.264087	-1.498915	-1.106395	-1.073258	-1.705786	-0.471266	-0.614104	...	-0.976307
1	1610612737	2005	-1.059369	-0.977735	-0.786667	-1.577773	-1.486932	-1.490412	-0.839727	-0.461249	...	-1.413032
2	1610612737	2006	-0.558896	-0.740009	-0.770914	-0.989593	-1.154715	-0.965047	0.455471	0.497565	...	-1.301051
3	1610612737	2007	-0.787756	-0.754867	-0.566770	-1.260740	-1.224992	-1.212437	0.682503	0.592057	...	-1.018299
4	1610612737	2008	0.097503	0.166323	-0.304900	-1.047994	-1.082842	-1.304063	1.665067	1.481392	...	-0.018870
...	...	...	...	...	...	...	...	...	...	...	...	...
442	1610612766	2015	-0.571471	0.090682	-0.538987	-0.568274	-0.306604	-0.618743	-0.501040	-0.466808	...	0.499041
443	1610612766	2016	0.276064	0.779548	0.517056	1.250493	1.304966	1.772191	0.660172	0.353048	...	1.431282
444	1610612766	2017	-0.252074	0.059616	-0.216343	0.745744	0.955180	0.095600	0.220995	-0.219461	...	0.613822
445	1610612766	2018	-0.081058	-0.010621	-0.153700	0.770773	0.720392	0.602756	0.209830	0.303023	...	0.521438
446	1610612766	2021	-0.523687	-0.577923	-0.882550	1.701013	1.673918	0.582452	-1.424055	-1.378376	...	-0.525584

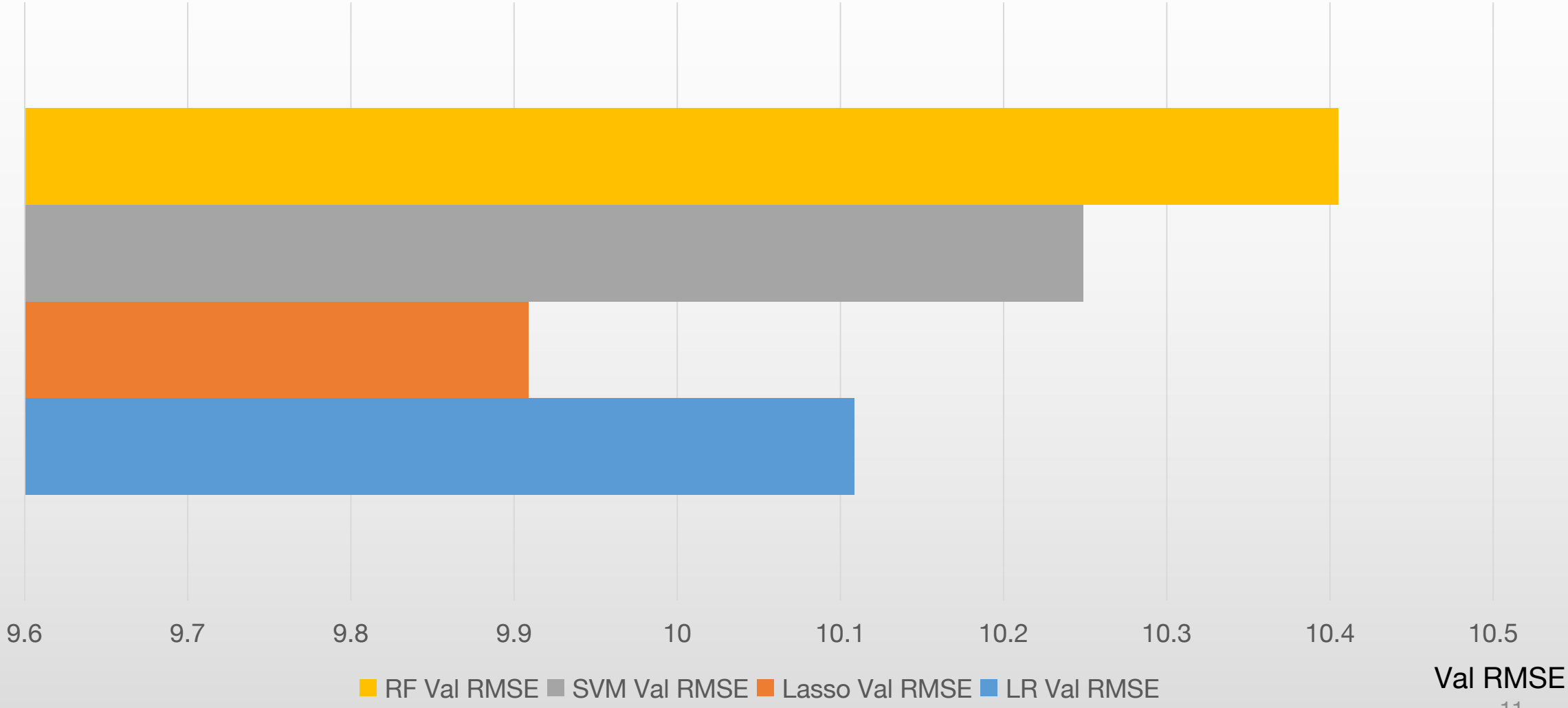
Model Selection

- Linear Regression
- Lasso Regression
- Support Vector Machine Regression
- Random Forest Regression

Model Training and Evaluation



Model Training and Evaluation



Feature Selection

Feature Name	Feature Importance
PLUS_MINUS	7.50363662
STL	0.94379254
DREB	0.72932052
FG3_PCT	-0.66973622
FG_PCT	-0.5065314
...	...
FGM	0
FGA	0
FTM	0
FTA	0

Drop 10 featues, Keep 9 features

Feature Selection

Model	Mean Val RMSE (Before)	Mean Val RMSE (After)
Linear Regression	10.11	9.89
Lasso Regression	9.91	9.83
SVM Regression	10.25	9.93
RF Regression	10.41	10.36

Real World Application (2018)

	team_name	wins_pred_2018	wins_2018
0	Rockets	58.0	53
1	Warriors	57.0	57
2	Raptors	52.0	58
3	76ers	51.0	51
4	Jazz	50.0	50
5	Celtics	48.0	49
6	Thunder	47.0	49
7	Spurs	45.0	48
8	Trail Blazers	45.0	53
9	Pacers	45.0	48
10	Pelicans	44.0	33
11	Nuggets	43.0	54
12	Timberwolves	43.0	36
13	Wizards	42.0	32
14	Cavaliers	41.0	19

- Use team statistics from 2017 to predict number of game wins in 2018 (assume we don’t know the result)
- Didn’t use statistics from later year because data are not so complete (perhaps due to COVID)
- Use Lasso Regression and round the result

Conclusion, Limitation, and Future Work

- Easy models like linear regression and lasso regression are preferred
- Features like Plus-Minus and Steals (positive), 3-Point Field Goal Percentage and Field Goal Percentage (negative) might be more considered when predicting game wins
- RMSE is still a bit high (slightly below 10), might due to poor features chosen
- Would try more feature engineering to find more valuable features