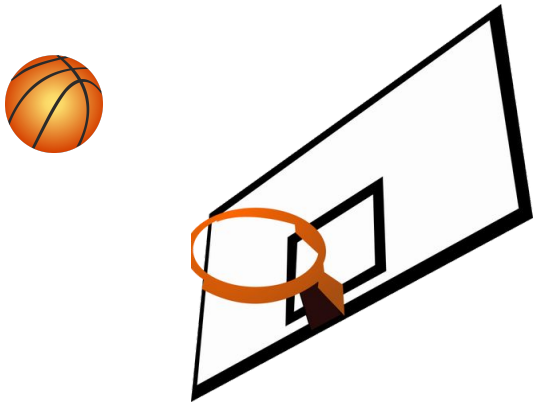
NBA GAME PREDICTIONS

By Philip Bowman



Overview

1. Goal
2. Trust the Process
3. Summary
4. Questions

Goal: To create a model that predicts the winner of an NBA game using historical game data.

Trust the Process: Bird's Eye

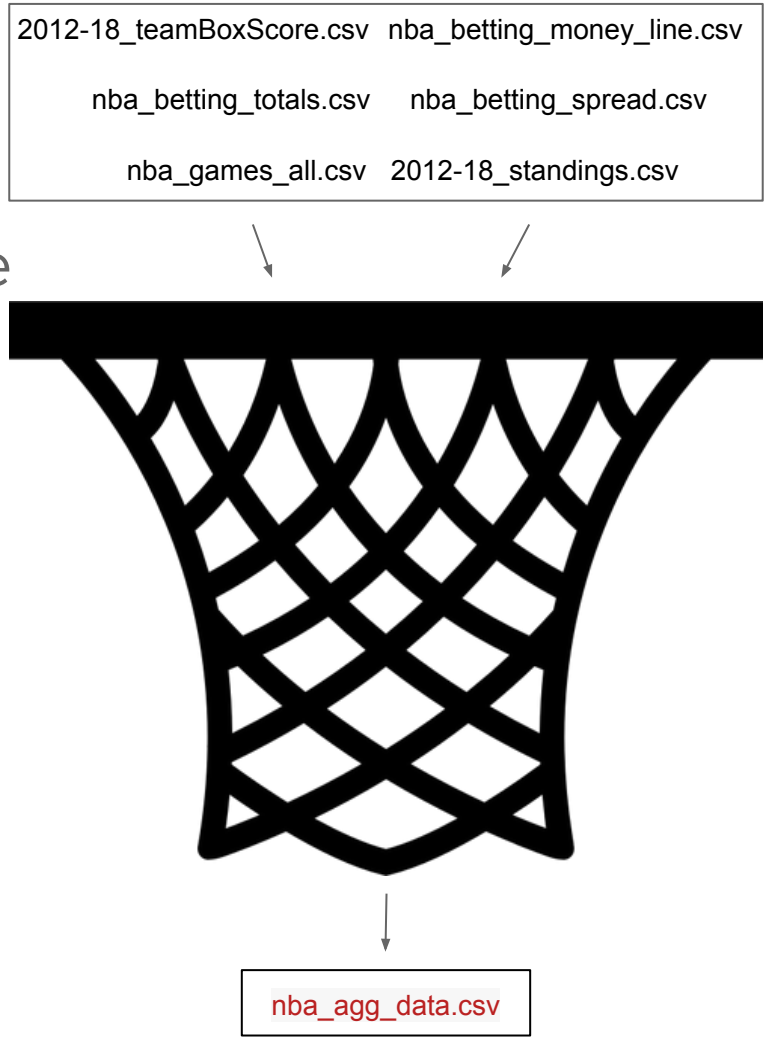
1. NBA Data Aggregation
2. NBA Data Cleaning and Exploration
    a. Data Cleaning
    b. Data Exploration
    c. Feature Engineering
3. NBA Modeling
4. NBA Model Testing

Data Aggregation: Sources

- Paul Rossotti's NBA Enhanced Box Score and Standings (2012 - 2018) (kaggle)
- Evan Hallmark's NBA Historical Stats and Betting Data (kaggle)

Data Aggregation

- Combine various datasets into one
- Regular season only
- 2012-2013 season to 2017-2018 season
- Merge/join, drop, repeat
- Average betting data
- Shift standings data
- Adversarial format



Data Aggregation: nba_agg_data.csv

- 14758 rows and 211 columns

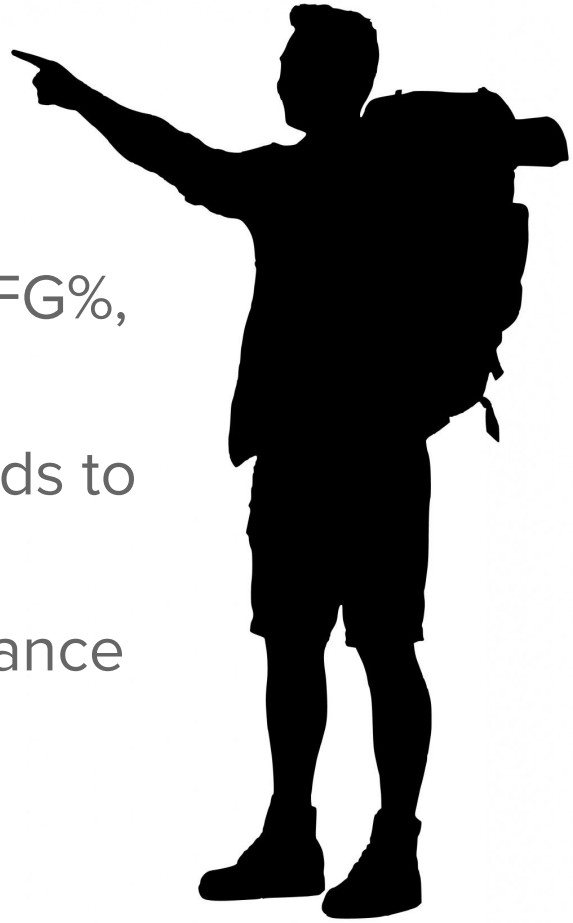
gameID	seasID	gmDate	gmTime	seasTyp	teamAbbr_A	teamConf_A	teamDiv_A	teamLoc_A	teamRslt_A	teamMin_A	teamDayOff_A	teamPTS_A	teamAST_A	teamTO_A	teamSTL_A	teamBLK_A	teamPF_A	teamFGA_A	teamFGM_A	teamFG%_A
10020	2012	2012-11-02	19:30	Regular	ATL	East	Southeast	Home	Loss	240	0	102	23	13	12	4	26	85	40	0.4706
10020	2012	2012-11-02	19:30	Regular	HOU	West	Southwest	Away	Win	240	2	109	22	21	8	2	18	90	38	0.4222
10039	2012	2012-11-04	19:00	Regular	ATL	East	Southeast	Away	Win	240	2	104	20	11	12	1	20	83	41	0.494
10039	2012	2012-11-04	19:00	Regular	OKC	West	Northwest	Home	Loss	240	2	95	27	21	4	9	21	71	33	0.4648
10057	2012	2012-11-07	19:30	Regular	ATL	East	Southeast	Home	Win	240	3	89	24	17	8	3	15	87	38	0.4368
10057	2012	2012-11-07	19:30	Regular	IND	East	Central	Away	Loss	240	2	86	18	15	10	7	14	85	35	0.4118
10067	2012	2012-11-09	19:30	Regular	ATL	East	Southeast	Home	Loss	240	2	89	20	13	4	5	16	81	34	0.4198
10067	2012	2012-11-09	19:30	Regular	MIA	East	Southeast	Away	Win	240	2	95	22	15	8	2	22	76	37	0.4868
10092	2012	2012-11-11	15:30	Regular	ATL	East	Southeast	Away	Loss	240	2	76	20	22	3	9	12	72	30	0.4167
10092	2012	2012-11-11	15:30	Regular	LAC	West	Pacific	Home	Win	240	3	89	21	15	12	6	22	82	38	0.4634
10095	2012	2012-11-12	22:00	Regular	ATL	East	Southeast	Away	Win	240	1	95	20	17	14	4	22	79	35	0.443
10095	2012	2012-11-12	22:00	Regular	POR	West	Northwest	Home	Loss	240	2	87	15	20	13	6	19	83	30	0.3614
10114	2012	2012-11-14	22:30	Regular	ATL	East	Southeast	Away	Loss	240	2	88	21	12	11	4	18	78	34	0.4359
10114	2012	2012-11-14	22:30	Regular	GS	West	Pacific	Home	Win	240	4	92	23	23	7	1	22	71	33	0.4648
10124	2012	2012-11-16	22:00	Regular	ATL	East	Southeast	Away	Win	240	2	112	24	15	8	5	14	76	42	0.5526
10124	2012	2012-11-16	22:00	Regular	SAC	West	Pacific	Home	Loss	240	3	96	25	14	11	0	22	85	39	0.4588
10151	2012	2012-11-19	19:30	Regular	ATL	East	Southeast	Home	Win	240	3	81	18	15	12	4	18	89	34	0.382
10151	2012	2012-11-19	19:30	Regular	ORL	East	Southeast	Away	Loss	240	1	72	17	19	6	5	19	82	31	0.378
10164	2012	2012-11-21	19:30	Regular	ATL	East	Southeast	Home	Win	265	2	101	30	14	12	5	25	92	40	0.4348
10164	2012	2012-11-21	19:30	Regular	WAS	East	Southeast	Away	Loss	265	2	100	20	23	8	6	27	89	38	0.427
10174	2012	2012-11-23	19:00	Regular	ATL	East	Southeast	Away	Win	240	2	101	29	22	6	15	18	78	42	0.5385
10174	2012	2012-11-23	19:00	Regular	CHA	East	Southeast	Home	Loss	240	2	91	17	15	16	9	15	83	31	0.3735
10183	2012	2012-11-24	19:00	Regular	ATL	East	Southeast	Home	Win	240	1	104	26	19	5	2	18	75	38	0.5067
10183	2012	2012-11-24	19:00	Regular	LAC	West	Pacific	Away	Loss	240	1	93	19	15	13	3	22	75	32	0.4267
10208	2012	2012-11-28	19:30	Regular	ATL	East	Southeast	Home	Win	240	4	94	28	9	9	9	18	90	36	0.4
10208	2012	2012-11-28	19:30	Regular	CHA	East	Southeast	Away	Loss	240	2	91	18	17	6	7	21	76	30	0.3947
10224	2012	2012-11-30	19:30	Regular	ATL	East	Southeast	Home	Loss	240	2	111	25	12	12	5	18	78	38	0.4872
10224	2012	2012-11-30	19:30	Regular	CLE	East	Central	Away	Win	240	3	113	21	19	8	3	25	85	41	0.4824
10263	2012	2012-12-05	19:30	Regular	ATL	East	Southeast	Home	Win	240	5	108	24	14	13	5	19	95	42	0.4421
10263	2012	2012-12-05	19:30	Regular	DEN	West	Northwest	Away	Loss	240	2	104	24	20	10	4	25	79	38	0.481
10279	2012	2012-12-07	19:30	Regular	ATL	East	Southeast	Home	Win	240	2	104	23	14	7	1	19	85	41	0.4824
10279	2012	2012-12-07	19:30	Regular	WAS	East	Southeast	Away	Loss	240	3	95	22	12	6	4	17	79	36	0.4557
10282	2012	2012-12-08	20:00	Regular	ATL	East	Southeast	Away	Win	240	1	93	20	16	10	3	16	77	34	0.4416
10282	2012	2012-12-08	20:00	Regular	MEM	West	Southwest	Home	Loss	240	1	83	15	16	8	5	19	77	33	0.4286
10299	2012	2012-12-10	19:30	Regular	ATL	East	Southeast	Away	Loss	240	2	92	16	14	8	1	16	75	32	0.4267
10299	2012	2012-12-10	19:30	Regular	MIA	East	Southeast	Home	Win	240	2	101	23	16	10	3	18	67	39	0.5821
10318	2012	2012-12-12	19:00	Regular	ATL	East	Southeast	Away	Win	240	2	86	16	14	11	6	11	80	35	0.4375

Data Cleaning

- Relatively clean data
- Interpolating missing numerical data
- Manipulation
- Not (immediately) dealing with outliers
- First game limitation

Data Exploration

- Home team wins more than away team
  - 58% vs. 41%
- More “good” stats (PTS, AST, STL, BLK, FG%, etc.) generally leads a higher win rate
- More “bad” stats (TOV, PF) generally leads to a lower win rate
- The lower the spread, the higher the chance of winning
- Recent performance
- Defensive rating (lower is better)



Feature Engineering

- New variables
- Team differences
- Moving Window or: How I Missed the Game Winner
- Eliminating features (401 to 25)
  - High correlation
  - Low correlation
  - Categorical Features

Feature Engineering: target_features.csv

teamWin_A	teamHome_A	gameBack_A	ptsAllow_A	teamPTS1_A	teamDrtg_A	team3PM_Diff_A	teamDRB_Diff_A	teamPTS1_Diff_A	teamPTS2_Diff_A	teamPTS3_Diff_A	teamBLKR_Diff_A	teamAST/TO_Diff_A
1	1	25	99.9	23.96	107.333232	-3.16	0.64	-2.44	-2	-1.92	0.047316	-0.21538
1	1	11	100.6	27.96	102.574628	2.72	-2.4	2.68	-2.4	-1.32	2.186764	0.330328
1	0	4.5	99.1	26.68	103.68694	1.44	1.52	0.68	0.88	-1	-3.871384	-0.253176
0	1	0	96.7	27.84	109.03038	2.6	3.36	1.4	5.44	2.48	-0.11268	0.246612
0	0	32.5	100.5	24.44	105.592256	1.24	-4.84	1.4	-4.88	-0.48	-2.072304	0.708708
0	0	8	98.8	20.68	110.333836	-2.24	-1.48	-5	-2.8	-2.16	0.119056	-0.639784
1	1	2	99.9	27.08	109.708712	-1	-1	0.32	-1.44	-0.4	-0.687388	-0.159724
1	1	10.5	100.5	25.8	102.124856	2.6	-1.12	1.12	0.96	2.4	2.654176	0.023436
0	0	31	103	24.76	113.7335	-0.92	-5.84	-1.8	-3.36	-5.52	-1.074064	-1.552772
1	1	10.5	96.9	30.24	96.096864	3.12	6.28	9.64	4.04	6.36	3.979936	0.40044
1	0	2	96.1	25.8	106.53012	-1.84	-0.2	-2.64	2	2.44	0.187748	0.142456
0	0	4.5	104.4	27.2	111.074704	0.64	-1.76	-0.8	-2.16	2.88	2.08082	0.41068
0	0	18	95.8	23.92	102.225376	1.48	-0.8	1.48	1.56	0.68	0.64336	0.069576
0	0	12.5	106.1	23.92	112.531308	1.16	1.52	-1.52	3.08	-1.72	-3.473584	-0.041764
1	1	0	98.8	24.92	104.309932	5.08	-0.36	0.96	-1	-6.16	0.40288	-0.025856
1	1	17	95	27.36	105.523112	0.88	3.04	5.72	3.64	-0.04	2.705352	-0.009412
1	0	1	96.4	29.28	101.4231	3.36	5.2	2.68	7.76	3.84	7.35228	0.059924
0	0	26.5	105.1	26.6	117.157764	2.44	-2.68	-0.76	2.24	0.04	-4.968372	-0.958864
1	1	3	105.4	25.4666666667	107.8447266667	-0.8666666667	-4.2666666667	-4.5333333333	6.2666666667	-1.2666666667	0.1635866667	0.3122066667

...

lastFiveWin%_B	teamDrtg_B	team3PM_Diff_B	teamPTS1_Diff_B	teamPTS3_Diff_B	teamBLKR_Diff_B	teamAST/TO_Diff_B	ptsAgnst_Diff	ptsScore_Diff	ptsAllow_Diff	spread_Diff	lastFiveWin%_Diff
0.8	104.215632	3.16	-1.04	4.04	4.166312	0.937584	485	-3.2	3.3	7.9	-0.2
0.2	102.803096	-0.08	-3.24	1.88	3.842984	-0.197548	442	11.3	3.7	-20.6	0.4
0.2	106.9431	1.72	-1.12	2.92	-1.199628	-0.525708	122	3.1	2.3	9.1	0.4
0.8	99.364076	2.32	5.04	2.56	-2.592392	1.11486	-475	-2.4	-3.6	-13.4	-0.2
0.4	111.331616	-3.44	-6.68	0.72	-3.7914	-0.631244	-386	-3.1	-1.5	12	-0.4
0.6	105.38168	3.36	-0.12	-2.08	0.784964	0.542904	-271	-3.6	-1.8	15.4	-0.4
0.8	106.125712	-2.28	-2.48	0.4	-1.41852	0.499224	-188	3.6	4.5	-9	-0.4
0.2	117.091212	-3.92	-3.2	-2.84	-1.387248	-0.337732	-516	0.8	-5	-19.9	0.6
0.6	112.038372	2.08	-2.44	-0.96	-1.00156	-0.802488	292	-4.1	3.2	26.2	-0.6
0.2	106.099296	-3.12	-3.6	-5.84	-0.153648	-0.188756	-498	0.5	-6.3	-34.1	0.8
0.4	111.41374	-2.56	-0.88	-0.64	-6.005752	-0.39942	-83	1.4	-7.5	-11.2	0.4
0.2	108.129656	-0.48	-2	-6.6	-1.67438	-0.024232	-254	5.7	-2.8	-12.4	0.6
0.6	102.091036	0.44	0.48	3.04	-0.420656	0.021344	443	2.8	6.6	14.1	-0.4
1	100.915988	-0.56	0.24	0.48	-1.617248	0.615536	402	5.3	7.1	11	-0.6
0.4	118.192824	-1.36	1.4	-0.36	4.11022	-0.510096	-366	-2	-7.1	-10.9	0.2
0.8	107.242948	0.8	-0.36	0.96	-6.21382	0.195628	-303	2	-3.8	-10.1	0
0.6	107.203528	1.24	-2.36	-1.92	1.82548	-0.136712	-192	9.5	0	-15.1	0.4
0.4	107.85512	2.84	-4.92	-3.84	-1.614304	-0.355576	310	-2.6	3.3	21.9	0
0.6	98.4062842105	2.7894736842	-2.2105263158	-0.3684210526	-3.3683315789	-0.3437684211	-47	4.2	9.7	-13	0

Modeling: Setup

- Separate target from features
- Split training, validation and holdout data (72%/8%/20%)
- Create Standardized and Min-Maxed training/validation sets
- Set up empty list for storing models
- Set up empty dictionary for storing type of data for each model

Modeling: Method and the Models

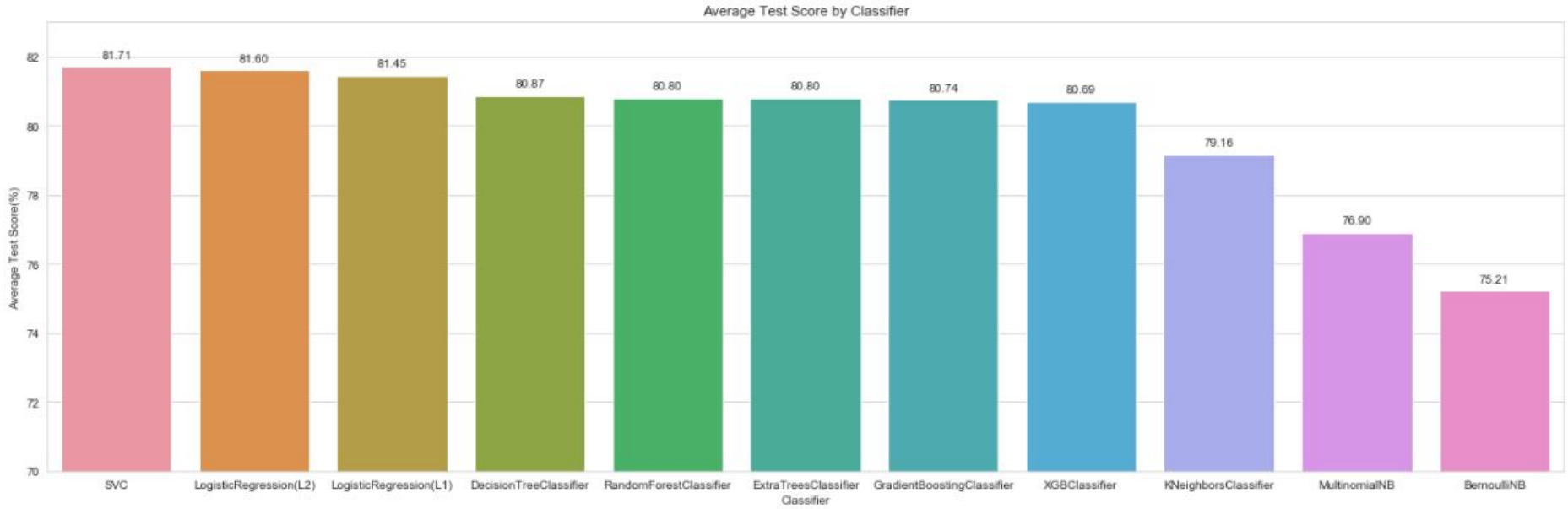
Method

- Find appropriate dataset using model defaults
- Assign parameters to be tested for model
- Find optimal parameters for model using GridSearchCV
- Store optimal model, fit time and dataset used for later inspection

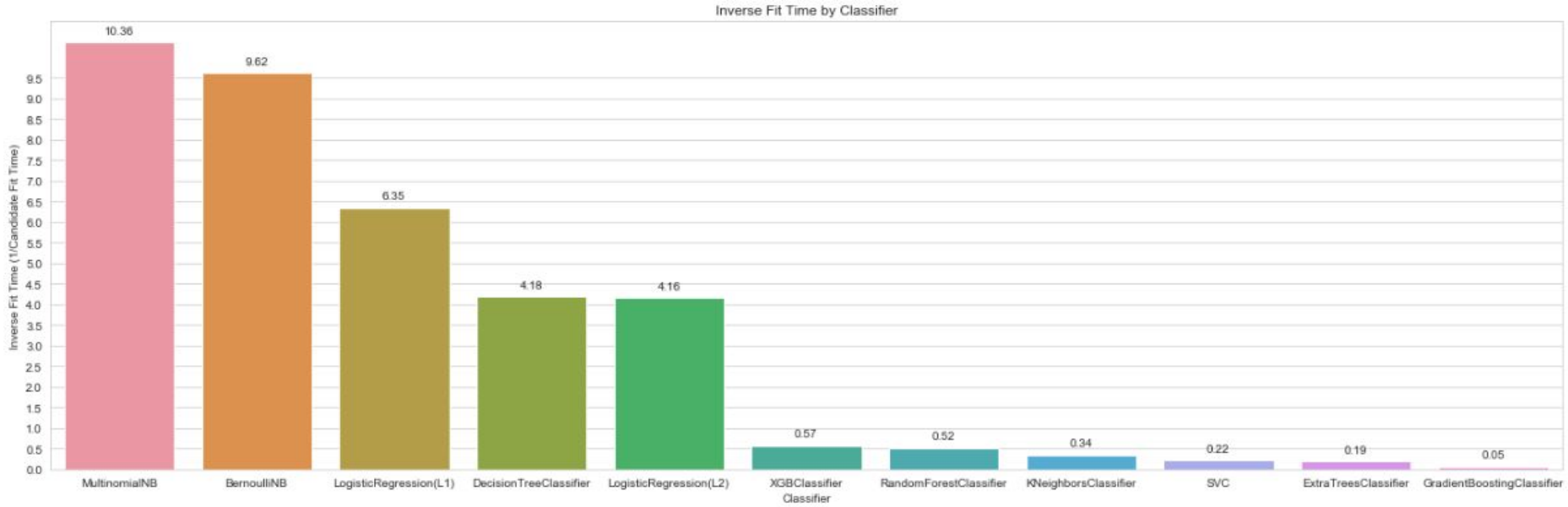
Models

1. Logistic Regression
2. K-Neighbors Classifier
3. Naive Bayes
4. Decision Tree
5. Random Forest
6. Support Vector Machines
7. Gradient Tree Boosting
8. eXtreme Gradient Boosting
9. Extremely Randomized Trees

Modeling: Average Score Results



Modeling: Inverse Fit Time Results



Modeling: Model Selection

- Logistic Regression (L2)
- Great speed (0.24s/candidate)
- High accuracy (81.6%)
- Dataset: Standardized
- Can’t be used for first games

Exact Parameters:

```
LogisticRegression(C=0.4, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=1000,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='newton-cg', tol=0.0001, verbose=0,
                   warm_start=False)
```

Model Testing or: Model in the Wild

- Utilize an NBA API to gather raw game stats from stats.nba.com endpoints
- Gather spread data from OddsShark
- 2019-2020 season only
- Manipulate data to fit exact format of training data
- Retrain model using all training data (2012-2018)
- Test performance on 2019-2020 games

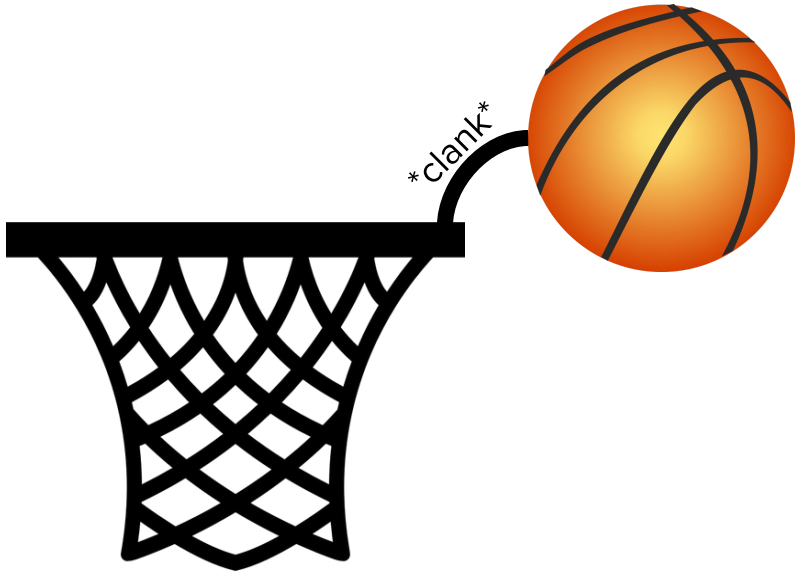
Model Testing: Result

63.8%...

This model doesn't perform well at all outside of the training data.

WHY?

Moving Window or: How I Missed the Game Winner



- Feature engineering step
- Find best K number of prior games
- Mimicking what should be unknown data
- Outcome leakage

Summary: The model chosen is not good. This project is a testament to how important the preparation steps are before modeling.

Probably going to miss using my model...

For the Future

- Focus on generalizing model
- Look for predictive aggregate statistics without comparing to actual game data

Questions

LINKS

- Project Files: https://github.com/philbowman212/Thinkful_repo/tree/master/projects/supervised_capstone
- Sources/Resources:
  - https://www.kaggle.com/pablote/nba-enhanced-stats (game data)
  - https://www.kaggle.com/ehallmar/nba-historical-stats-and-betting-data (betting data)
  - https://github.com/swar/nba_api (nba api)
  - https://www.oddsshark.com/nba/database (OddsShark betting data)