Basketball Analytics MVP Predictor

Math 189R Project by Sid Rastogi

Background

- The NBA MVP is a highly contentious award that is always heavily debated each year
- Voted for by a panel of sportswriters and broadcasters - criteria for decision can be very subjective
- No set criteria
- Want to develop a model that can be used to predict who the season MVP should be based on player season stat

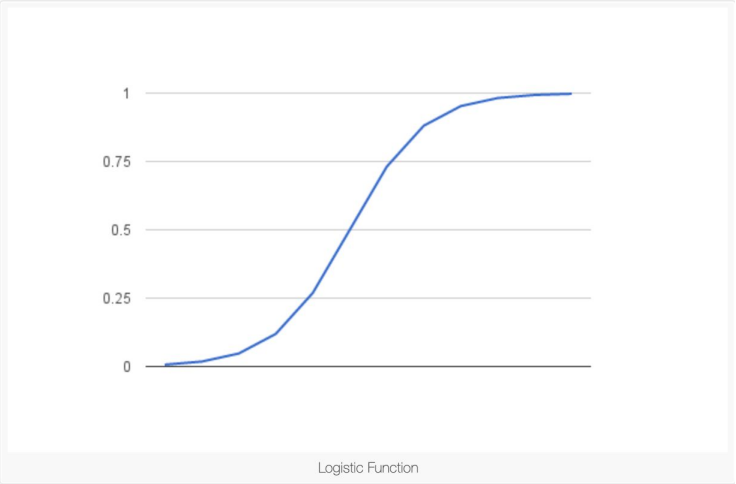
Objective

- Goal: Predict MVP of an NBA season
- Data: NBA player season total statistics per year from 1950
  - https://www.kaggle.com/drgilermo/nba-players-stats#Seasons_Stats.csv
  - Used data from 1982 onwards, since incomplete before then
  - Had to do a TON of data preprocessing :(

Solution

- Tried many models (neural network, SVM, etc.)
- Logistic regression worked!
  - Labeled players as MVP (1) or not (0) and used logistic binary classifier
- Used data from 1982-2016 seasons to train model
- Used data from 2017 season to test
  - Russell Westbrook was named MVP that season
- Trained on 38 features (removed 5 redundancies)

Model



- Logistic regression model for binary classification (MVP or not)
- Optimal weights θ found using gradient descent
- Sigmoid activation function

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_n X_n)}} = \frac{1}{1+e^{-(\beta_0+\sum \beta_i X_i)}}.$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

2016-17 NBA Awards Voting

« 2015-16 Awards Voting | 2017-18 Awards Voting »

Most Valuable Player

Share & more ▼ Glossary

				Voting					Per Game						Shooting			Advanced	
Rank	Player	Age	Tm	First	Pts Won	Pts Max	Share	G	MP	PTS	TRB	AST	STL	BLK	FG%	3P%	FT%	WS	WS/48
1	Russell Westbrook	28	OKC	69.0	888.0	1010	0.879	81	34.6	31.6	10.7	10.4	1.6	0.4	.425	.343	.845	13.1	.224
2	James Harden	27	HOU	22.0	753.0	1010	0.746	81	36.4	29.1	8.1	11.2	1.5	0.5	.440	.347	.847	15.0	.245
3	Kawhi Leonard	25	SAS	9.0	500.0	1010	0.495	74	33.4	25.5	5.8	3.5	1.8	0.7	.485	.380	.880	13.6	.264
4	LeBron James	32	CLE	1.0	333.0	1010	0.330	74	37.8	26.4	8.6	8.7	1.2	0.6	.548	.363	.674	12.9	.221
5	Isaiah Thomas	27	BOS	0.0	81.0	1010	0.080	76	33.8	28.9	2.7	5.9	0.9	0.2	.463	.379	.909	12.5	.234
6	Stephen Curry	28	GSW	0.0	52.0	1010	0.051	79	33.4	25.3	4.5	6.6	1.8	0.2	.468	.411	.898	12.6	.229
7T	Giannis Antetokounmpo	22	MIL	0.0	7.0	1010	0.007	80	35.6	22.9	8.8	5.4	1.6	1.9	.521	.272	.770	12.4	.210
7T	John Wall	26	WAS	0.0	7.0	1010	0.007	78	36.4	23.1	4.2	10.7	2.0	0.6	.451	.327	.801	8.8	.149
9T	Anthony Davis	23	NOP	0.0	2.0	1010	0.002	75	36.1	28.0	11.8	2.1	1.3	2.2	.505	.299	.802	11.0	.195
9T	Kevin Durant	28	GSW	0.0	2.0	1010	0.002	62	33.4	25.1	8.3	4.8	1.1	1.6	.537	.375	.875	12.0	.278
11	DeMar DeRozan	27	TOR	0.0	1.0	1010	0.001	74	35.4	27.3	5.2	3.9	1.1	0.2	.467	.266	.842	9.0	.166

Results

- Model had 99.68% accuracy on training data
- Surprisingly, the model actually predicted Russell Westbrook to win the MVP award in 2017!
- Furthermore, it named James Harden as 2nd most likely, which happened in real life as well
- Generally, good players were higher than bad players
- 10 of the top 20 players the model produced were named All-Stars that season

```
 ['Nikola Jokic' 0.9999999973806215]
 ['Giannis Antetokounmpo' 0.9999999976953009]
 ['Marcin Gortat' 0.999999998364862]
 ['Anthony Davis' 0.9999999985261354]
 ['Stephen Curry' 0.9999999992354277]
 ['Dwight Howard' 0.9999999993689178]
 ['Trevor Ariza' 0.9999999994034924]
 ['Kevin Love' 0.9999999996324713]
 ['LeBron James' 0.9999999997452249]
 ['Rudy Gobert' 0.9999999998079419]
 ['Nicolas Batum' 0.9999999998254964]
 ['Nikola Vucevic' 0.9999999998257108]
 ['Karl-Anthony Towns' 0.9999999999720947]
 ['DeAndre Jordan' 0.9999999999783511]
 ['Hassan Whiteside' 0.9999999999826057]
 ['DeMarcus Cousins' 0.999999999989584]
 ['Draymond Green' 0.9999999999938318]
 ['Andre Drummond' 0.9999999999985597]
 ['James Harden' 0.9999999999999762]
 ['Russell Westbrook' 0.9999999999999992]]
MVP Prediction: Russell Westbrook
Accuracy on training data: 99.68299972828548
```

Project

- Github link: https://github.com/srastogi1011/MATH-189R-The-Math-of-Big-Data

Drawbacks

- Dataset:
  - Doesn’t factor in team success
  - Only goes till 2017
  - Low number of positive results
- Model
  - Doesn’t account for time dependence of data
  - Overrates big men - 9/10 players in top 20 that weren’t named all-stars were centers

Looking forward...

- Model can definitely be improved to fix some of the drawbacks
  - Maybe using a neural network
- Dataset could be used to answer other interesting questions (NBA all-stars, types of players, etc.)
- Can do better feature selection/scaling
- Can test model on subsets of same data set to account for time (2000 onwards, 2010 onwards, etc.)

Thanks!

Any questions?