How Important is Narrative? Using Machine Learning to Predict NBA MVP

Jake Reichel

PRINCETON UNIVERSITY

Motivation

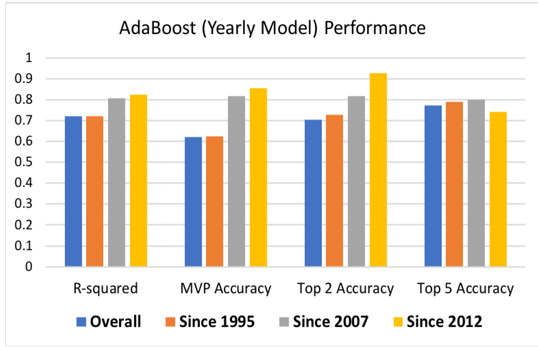
- Every season, the National Basketball Association (NBA) has a panel of voters determine who will be awarded the Most Valuable Player (MVP) award.
  - ❖ Award given to player with highest number of "vote shares"
- Voters change yearly, and process is not very clear
  - ❖ How important is the player's "narrative"?
  - ❖ Does a team's record play a role in determining MVP?
  - ❖ Can statistics, alone, predict who will win MVP?
- Predict 2019 NBA MVP

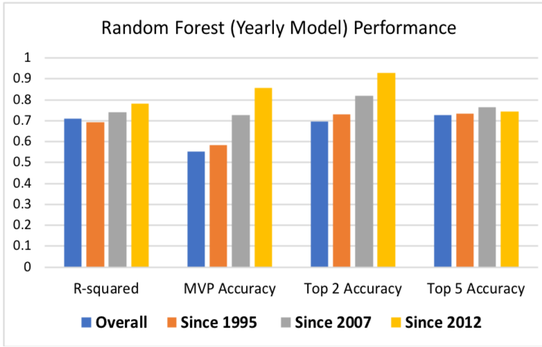
Process / Datasets Used for Training

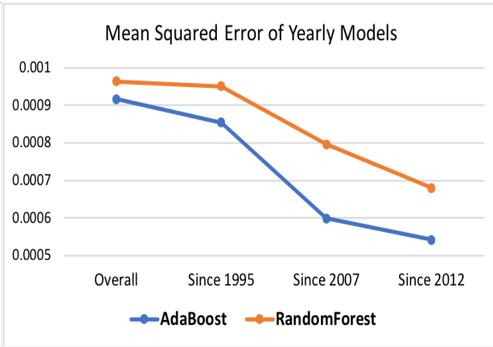
Iterative process used to continually retrain the models. Listed in order

1. **Full Dataset** – All players, all statistics. This was a baseline test. Found winning percentage was very important to model prediction.
2. **Reduced Dataset** – Limited to players who received votes for MVP. Attempt to limit the "noise." Did not improve the non-linear models.
3. **Redundancy Reduced Dataset** – Eliminated columns with high collinearity. Helped focus the models on different features.
4. **Scaled Dataset** – Each row was scaled based on all player's performances *for that season*. This allowed for much better interpretation of performance.
5. **Yearly Dataset** – This is the scaled model, without the "test" holdout. Instead, models were trained on all years, and tested on a single holdout.

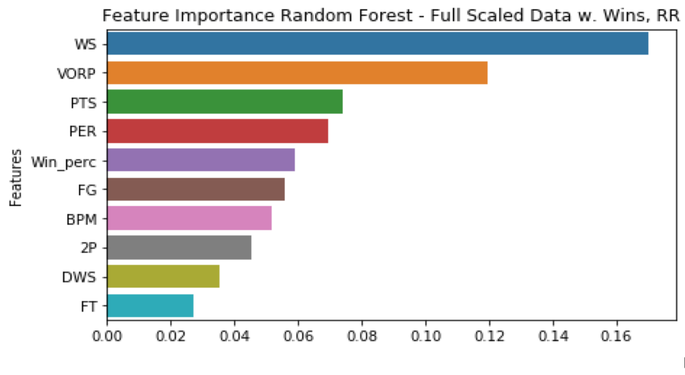
Performance and Evaluation Metrics

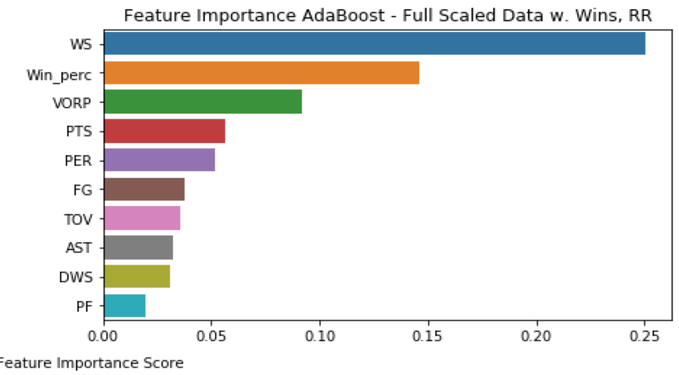






- Overall, both performed pretty well in predicting MVP: Random Forest – 58%, AdaBoost – 63%
- Both of the models have continually improved in performance in recent years.
  - ❖ Random Forest: *Since 2007* – 73% MVP, 82% Top 2; *Since 2012* – 86% MVP, 93% Top 2
  - ❖ AdaBoost: *Since 2007* – 82% MVP, 82% Top 2; *Since 2012* – 86% MVP, 93% Top 2





- Important feature differences account for difference in model performance, including some outliers.
- Nearly all important features (with exception of DWS) are *offensive* statistics.

2019 Predictions

	Ada-Player	Ada-Share	RF-Player	RF-Share
1	Giannis Antetokounmpo	0.808	James Harden	0.660281
2	James Harden	0.720	Giannis Antetokounmpo	0.623163
3	Nikola Jokic	0.148	Nikola Jokic	0.256289
4	Kevin Durant	0.093	Paul George	0.248621
5	Damian Lillard	0.091	Kevin Durant	0.248159

Outliers & Sources of Error

Error Sources:

- Lack of variable to account for popularity and narrative
- No available statistics to capture defensive performance

Outliers:

- 1999 – entirely wrong predictions. Lockout shortened season
- 2005 – Steve Nash MVP. "7 seconds or less" narrative
- 2017 – Russell Westbrook MVP. "Triple-Double" narrative

Conclusion(s)

- Statistics, alone, are mostly (86%) sufficient to predicting NBA MVP
- Voters choose MVP and runner-up based on statistics
  - ❖ Their other votes heavily factor in popularity of the players
- 2019 NBA MVP race is very close. Giannis Antetokoumnpo has the edge