Predicting the NBA MVP

Kobe Sarausad

5/29/2022

How are we doing this?

Outline

- The Problem

Outline

- The Problem
- The Data

Outline

- The Problem
- The Data
- The Model

Outline

- The Problem
- The Data
- The Model
- The Results

The Problem

- Picture this, you have $100 in Vegas, mid NBA season, and you want some money.

The Problem

- Picture this, you have $100 in Vegas, mid NBA season, and you want some money.
- You want a data-driven way of doing this

The Problem

- Picture this, you have $100 in Vegas, mid NBA season, and you want some money.
- You want a data-driven way of doing this
- Why not put some money on some juicy odds for the NBA MVP?

The Approach

Use the public's perception of the player is one of the predictors of the model.

The Data

We first went on by using basic season statistics obtained from basketball reference

Data Dictionary

The following are the variables that are in consideration for the model.

- `rank` - ranking based on the number of votes the player received for that particular year
- `season` - NBA season
- `player` - NBA player
- `player_id` - NBA player id (basketball reference)
- `age` - NBA player's age for that particular season
- `team` - NBA player's team for that particular season
- `fpv` - first place votes
- `mvppts` - points based on votes
- `ptsmax` - total amount of points given to the pool of MVPs
- `share` - the share of the max points the player received
- `games` - number of games player played during the season
- `mpg` - minutes played per game
- `ppg` - points per game
- `rbpg` - rebounds per game
- `astpg` - assists per game

Modeling

I started off with a simple model, using linear regression

Trying other models

I moved onto trying other regression models, such as:

- Logistic Regression

Trying other models

I moved onto trying other regression models, such as:

- Logistic Regression
- LASSO Regression

Trying other models

I moved onto trying other regression models, such as:

- Logistic Regression
- LASSO Regression
- Ridge Regression

Trying other models

I moved onto trying other regression models, such as:

- Logistic Regression
- LASSO Regression
- Ridge Regression
- Random Forest

Trying other models

I moved onto trying other regression models, such as:

- Logistic Regression
- LASSO Regression
- Ridge Regression
- Random Forest
- XGBoost

Settling on XGBoost

I ended up choosing XGBoost and sticking with it for the remainder of the project.

Results

SEASON	IMG	PLAYER	RANK	PREDICTED RANK	COMP
2010		LeBron James	1	1	-0.1
2010		Dwyane Wade	5	2	-0.1
2010		Kevin Durant	2	3	0.14
2011		LeBron James	3	1	-0.1
2011		Kevin Durant	5	2	1.24
2011		Manu Ginobili	8	3	-0.1
2012		LeBron James	1	1	-1.8
2012		Kevin Durant	2	2	-0.1
2012		Chris Paul	3	3	-1.14
2013		LeBron James	1	1	-0.1

1–10 OF 39 ROWS PREVIOUS **1** 2 3 4 NEXT

Reflection

This was a great learning experience, getting hands-on experience applying machine learning to real-world data. It was super helpful to have mentors by my side to lead me through the journey of the whole process. From scraping data, to compiling the final presentation, big thanks to them,

The whole process was a huge learning experience, especially realizing that the majority of the data you want in the world is not readily available on the web, and it's necessary that we spend time scraping and cleaning all this data to feed it to our models.

Future

I would love to apply other kinds of models to this data like neural nets and deep learning.