NBA MVP Predictor

Introduction

Question:
Can machine learning algorithms predict who the next NBA MVP will be?
What features are most important in an MVP campaign?

Data obtained via basketball-reference.com.

Preprocessing, machine learning, and visualization done in Python, powered by Jupyter Notebook.

The Data - Preprocessing

20 Files

2/year (2010-2019)

Files converted into pandas dataframes from .csv files obtained from basketball-reference.com and stiched together into 1 dataframe.

3373 Entries

After removing players w/<40 games

Each entry has 38 features made up of per game and advanced statistics for each season, as well as name, age, position, team, year, and a binary classifier for an MVP season.

10 MVPs

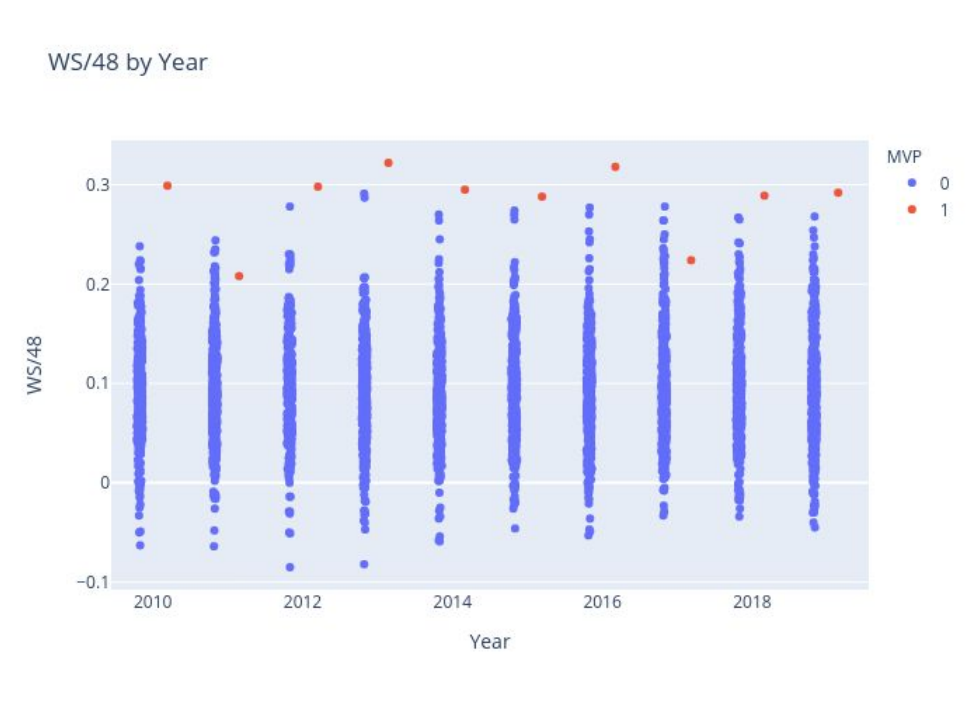
Large class imbalance can lead to overfitting

Because of this imbalance, data resampling was necessary.

Preprocessing

- Merge per game and advanced statistics for each year.
- Drop players with fewer than 40 games played.
- Fill undefined 3 point shooting (0 attempts) with a 0%.
    - Top caliber players who don’t take 3 point shots would be left out otherwise (e.g. Rudy Gobert)
- Clean rows with null values.
- Add a variable for the year and a dummy variable for who won MVP that season.
- Concatenate the yearly dataframes.

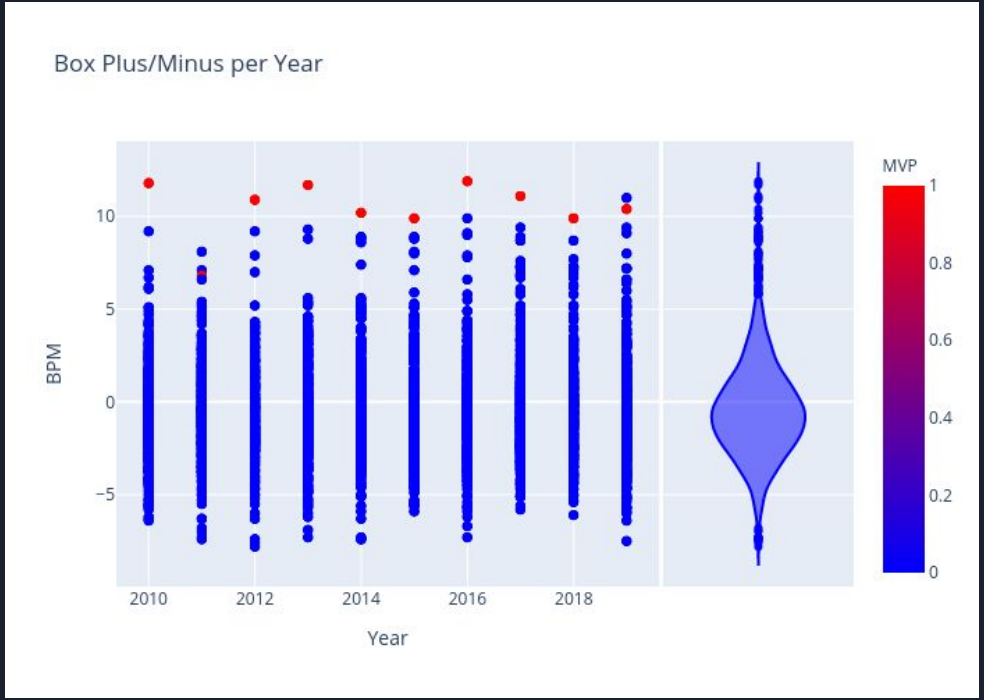
Exploratory Data Analysis



The MVP for each year lead in Winshare/48 in 8 out of 10 seasons in the dataset. Winshare/48 is a measure of the amount of wins a player contributed to if all players played the same number of minutes.

For an interactive version visit: https://plotly.com/~dylanstaub/3/

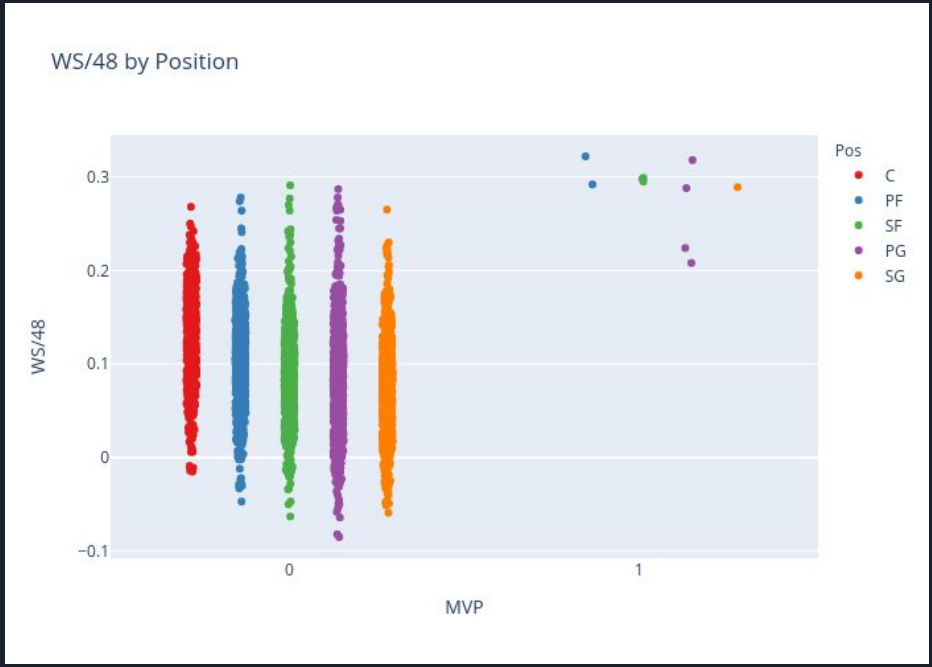
Exploratory Data Analysis



The leader in BPM/year won MVP 8 out of 10 seasons as well. Box Plus/Minus is the difference between the amount of points a players team and the opposing team scored while the player was on the court.

For an interactive version visit: https://plotly.com/~dylanstaub/12

Exploratory Data Analysis

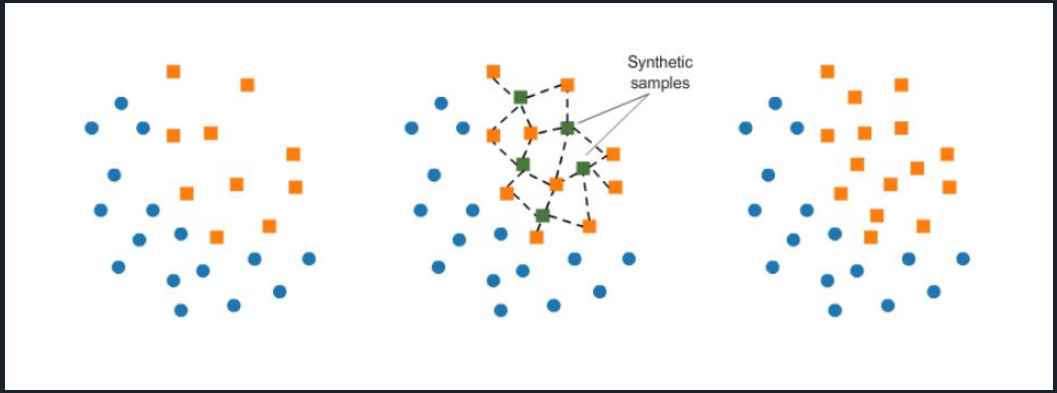


This graph separates the MVPs from the other players at their positions, sorted by color. No Center in the dataset has won and the most winners comes from the position with the highest range of WS/48 and lowest minimum WS/48.

For an interactive version visit: https://plotly.com/~dylanstaub/14/

Class Imbalance

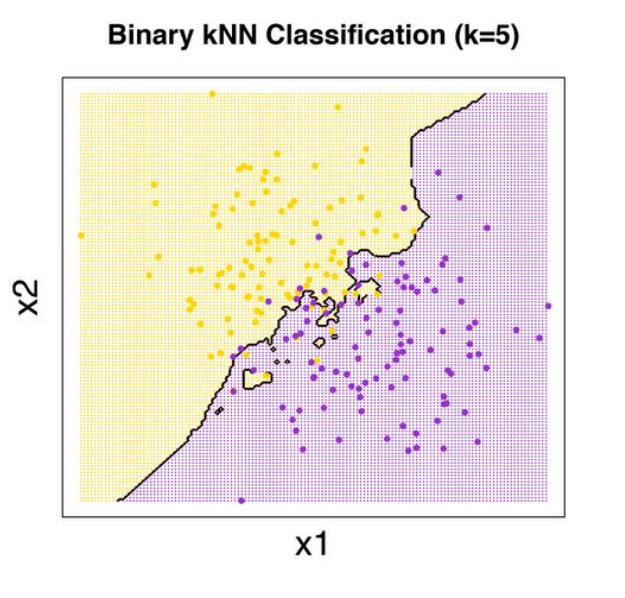
- Because there is only 1 MVP per season there is a massive class imbalance which can lead to overfitting in models.
- To combat this I used random under sampling to reduce the majority class and SMOTE to oversample the minority class to balance the two.
- SMOTE works by creating “look-a-like” data points based on the existing MVPs.



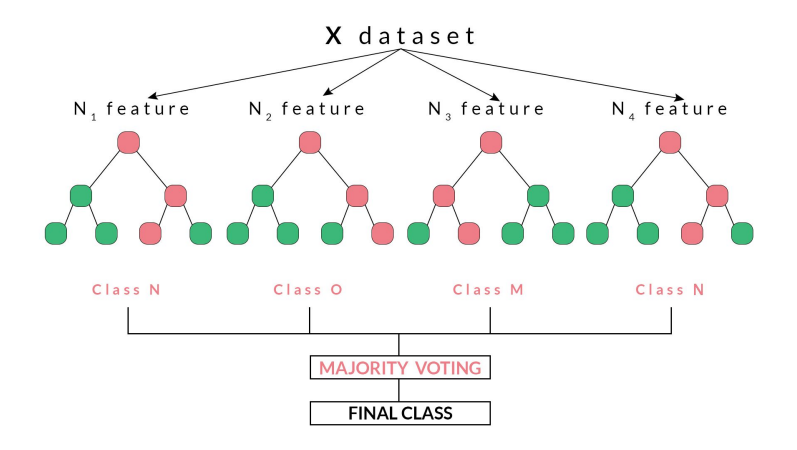
- Ended up with 28 data points in each class in the training data.

Creating Models

K Nearest Neighbors



Random Forest



K Nearest Neighbors Results

K-values chosen based on lowest error rates.

K = 2

True Label	Predicted Label 0	Predicted Label 1
0	987	22
1	0	3

Accuracy: 98%

Incorrect Predictions: 22

K = 6

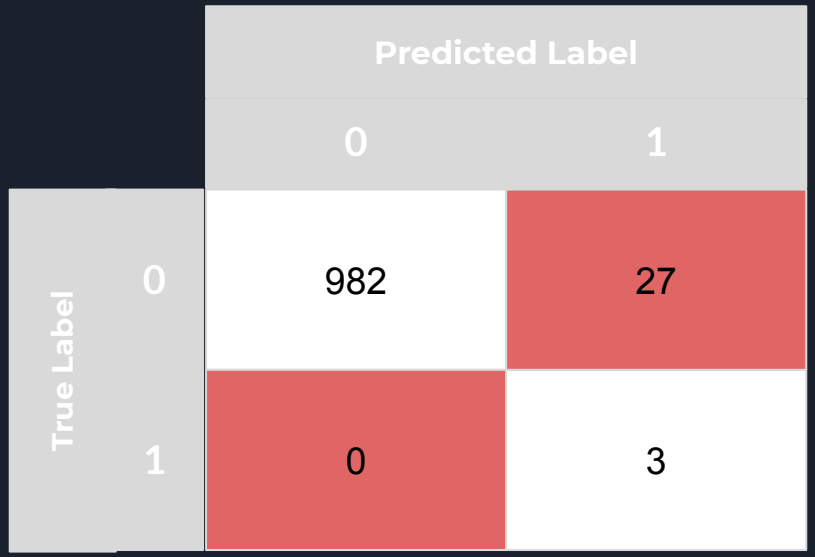
True Label	Predicted Label 0	Predicted Label 1
0	985	24
1	0	3

Accuracy: 98%

Incorrect Predictions: 24

High accuracy and no false negatives which is good. We don't want an MVP to not be classified as such.

Random Forest Results



Accuracy: 97%
Incorrect Predictions: 27

Slightly more incorrect than KNN but also very high accuracy and no false negatives.

Model Creation Conclusions

Both KNN and Random Forest created accurate predicting models, and neither of them seem to overfit thanks to the resampling methods.

Now we can see which features had the biggest impact on classification and test the models on data from the NBA season thus far to see how the predictor well the predictor works in the real world!

Feature Selection

Using Random Forests built in feature_importances_ we can see the impact each feature had on classification.

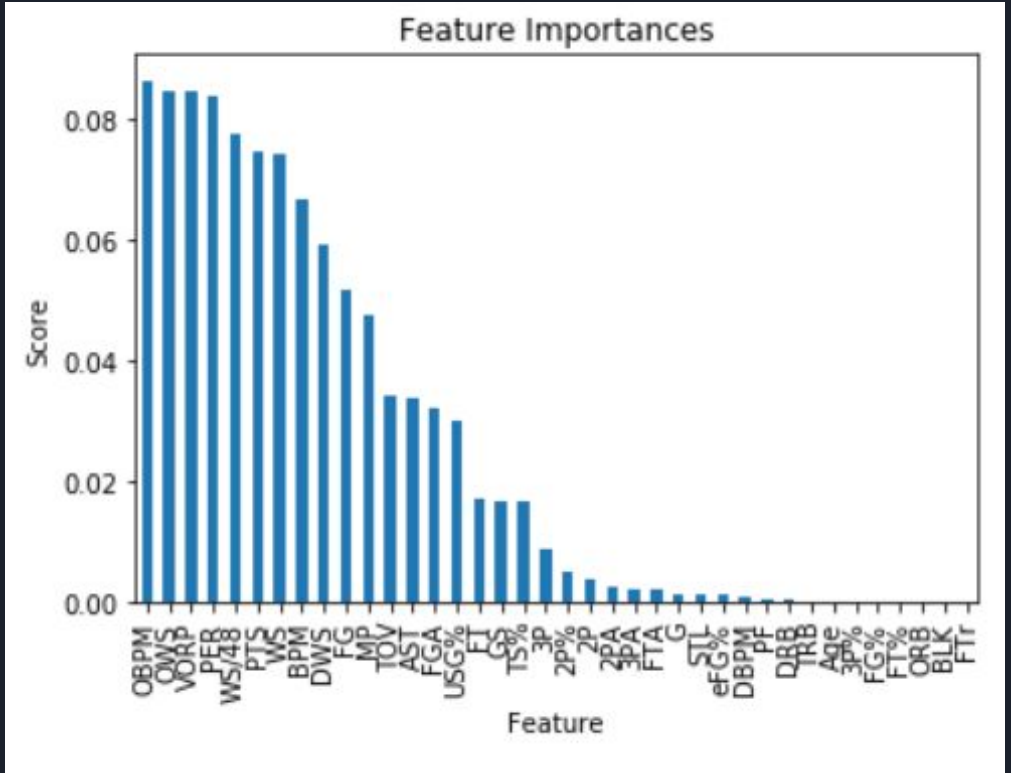
Higher score indicates more of an impact.

OBPM	0.086412
OWS	0.084506
VORP	0.084425
PER	0.083770
WS/48	0.077583
PTS	0.074699
WS	0.074344
BPM	0.066491
DWS	0.059262
FG	0.051750
MP	0.047450
TOV	0.034200
AST	0.033665
FGA	0.032183
USG%	0.030034
FT	0.017134
GS	0.016710
TS%	0.016681
3P	0.008748
2P%	0.005000
2P	0.003861
2PA	0.002468
3PA	0.002129

Feature Selection

The top 7 or so features are all similarly good at predicting MVP but predictability falls off pretty quickly after that.

There is not a solely defensive stat in the top 8 most important features indicating MVP voters value offensive prowess over defensive.



Predicting the 2020 MVP

I got another dataset from basketball-reference.com with all of the regular season data up until the COVID-19 break. A good test will be to see if my models can predict the MVP without a full season's worth of data.

Predicting the 2020 MVP

Using KNN=2

Rk	Player
12	Giannis Antetokounmpo\antetgi01
81	Jimmy Butler\butleji01
121	Anthony Davis\davisan02
132	Luka Dončić\doncilu01
196	James Harden\hardeja01
248	LeBron James\jamesle01
259	Nikola Jokić\jokicni01
292	Kawhi Leonard\leonaka01
295	Damian Lillard\lillada01
521	Trae Young\youngtr01

Using KNN=6

Rk	Player
12	Giannis Antetokounmpo\antetgi01
81	Jimmy Butler\butleji01
121	Anthony Davis\davisan02
132	Luka Dončić\doncilu01
196	James Harden\hardeja01
248	LeBron James\jamesle01
259	Nikola Jokić\jokicni01
292	Kawhi Leonard\leonaka01
295	Damian Lillard\lillada01
301	Kyle Lowry\lowryky01
521	Trae Young\youngtr01

Using RFC

Rk	Player
12	Giannis Antetokounmpo\antetgi01
121	Anthony Davis\davisan02
132	Luka Dončić\doncilu01
196	James Harden\hardeja01
248	LeBron James\jamesle01
259	Nikola Jokić\jokicni01
292	Kawhi Leonard\leonaka01
295	Damian Lillard\lillada01
472	Karl-Anthony Towns\townska01
521	Trae Young\youngtr01

Obviously there can only be 1 MVP but these are the players that are having MVP caliber seasons according to the models.

Predicting the 2020 MVP

However, using the predict_proba method in the Random Forest Classifier we can see which player had the highest percentage of MVP decision trees, indicating a higher likelihood of accuracy.

When looking at this metric Giannis Antetokounmpo looks poised to repeat his MVP win of 2019.

This information matches up with the Vegas odds on the MVP race, indicating my model will likely be proven correct.

	0	1	Player
0	0.033	0.967	Giannis Antetokounmpo
3	0.054	0.946	James Harden
4	0.168	0.832	LeBron James
2	0.181	0.819	Luka Doncic
6	0.181	0.819	Kawhi Leonard
1	0.208	0.792	Anthony Davis
7	0.229	0.771	Damian Lillard
5	0.407	0.593	Nikola Jokic
8	0.481	0.519	Karl Anthony Towns
9	0.483	0.517	Trae Young