Assessment of methods for predicting the NBA regular season MVP using Regression analysis and Classification

Aris-Konstantinos Terzidis

Semester Project in Data Science

Academic Advisor: Dr. Ioannis Katakis

Basketball analytics and MVP award

- Basketball analytics have changed the face of the game in the past 15 years. All teams have data analysts on their payroll. Data analytics play big part in the decision process for coaches, players and managers.
- Since the 1980-81 season, the NBA regular season MVP is decided by people of the media such as reporters and sportscasters from the USA and Canada.
- The votes are not solely based on performance merits. Other factors, such as expectations, team popularity, “the narrative” are equally if not more influential. Most of the times, the actual best player in terms of performance merits is voted as the MVP

Dataset

- Training dataset: Statistical attributes from seasons 1980-81 to 2017-18
- Predicting for season 2018-19
- Basic box score statistics (e.g. points, rebounds per game)
- Advanced statistics (e.g. PER, win-share)
- No feature engineering

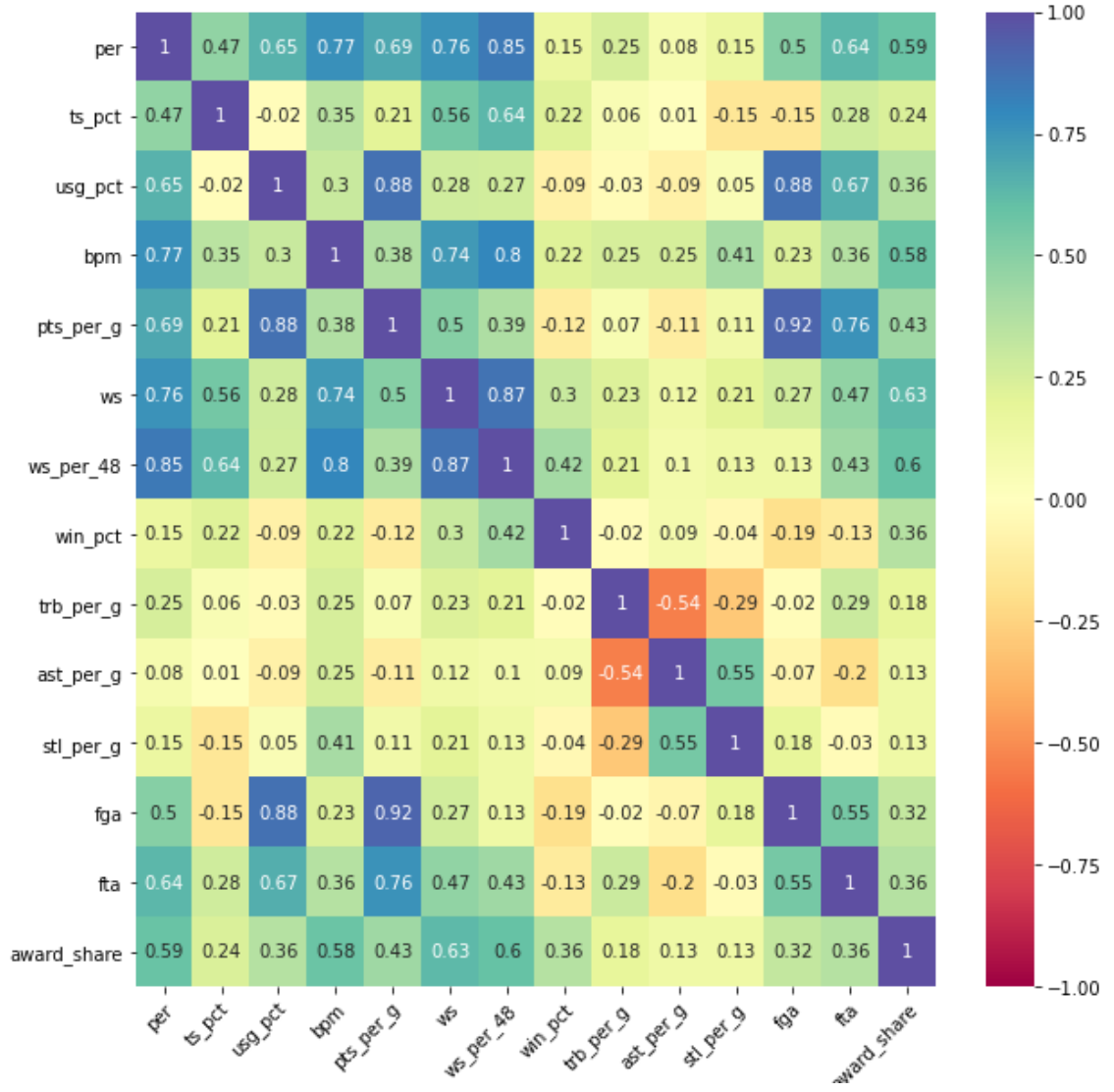
				Voting					Per Game						Shooting			Advanced	
Rank	Player	Age	Tm	First	Pts Won	Pts Max	Share	G	MP	PTS	TRB	AST	STL	BLK	FG%	3P%	FT%	WS	WS/48
1	Stephen Curry	27	GSW	131.0	1310.0	1310	1.000	79	34.2	30.1	5.4	6.7	2.1	0.2	.504	.454	.908	17.9	.318
2	Kawhi Leonard	24	SAS	0.0	634.0	1310	0.484	72	33.1	21.2	6.8	2.6	1.8	1.0	.506	.443	.874	13.7	.277
3	LeBron James	31	CLE	0.0	631.0	1310	0.482	76	35.6	25.3	7.4	6.8	1.4	0.6	.520	.309	.731	13.6	.242
4	Russell Westbrook	27	OKC	0.0	486.0	1310	0.371	80	34.4	23.5	7.8	10.4	2.0	0.3	.454	.296	.812	14.0	.245
5	Kevin Durant	27	OKC	0.0	147.0	1310	0.112	72	35.8	28.2	8.2	5.0	1.0	1.2	.505	.387	.898	14.5	.270
6	Chris Paul	30	LAC	0.0	107.0	1310	0.082	74	32.7	19.5	4.2	10.0	2.1	0.2	.462	.371	.896	12.7	.253
7	Draymond Green	25	GSW	0.0	50.0	1310	0.038	81	34.7	14.0	9.5	7.4	1.5	1.4	.490	.388	.696	11.1	.190
8	Damian Lillard	25	POR	0.0	26.0	1310	0.020	75	35.7	25.1	4.0	6.8	0.9	0.4	.419	.375	.892	9.2	.165
9	James Harden	26	HOU	0.0	9.0	1310	0.007	82	38.1	29.0	6.1	7.5	1.7	0.6	.439	.359	.860	13.3	.204
10	Kyle Lowry	29	TOR	0.0	6.0	1310	0.005	77	37.0	21.2	4.7	6.4	2.1	0.4	.427	.388	.811	11.6	.196

Predicting the MVP

- Predicting the MVP is a ranking problem
- Prediction of number of votes a player will be awarded
- Two techniques used:
  - Regression analysis
  - Classification

Regression analysis

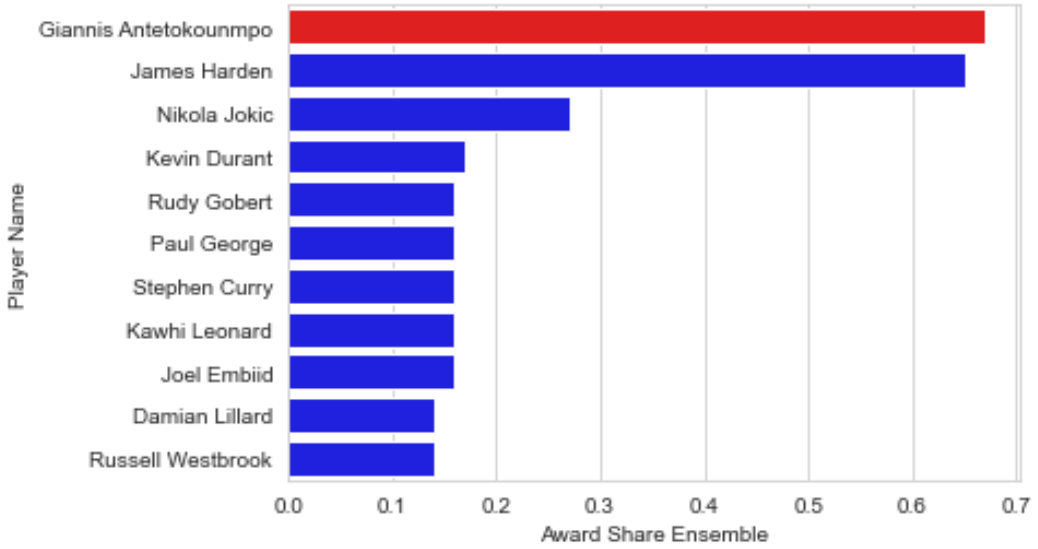
- Feature selection:
  - Correlation between features
  - Univariate selection based on statistical tests (F measure) with Select K Best method by Sklearn
  - Domain knowledge



Regression analysis

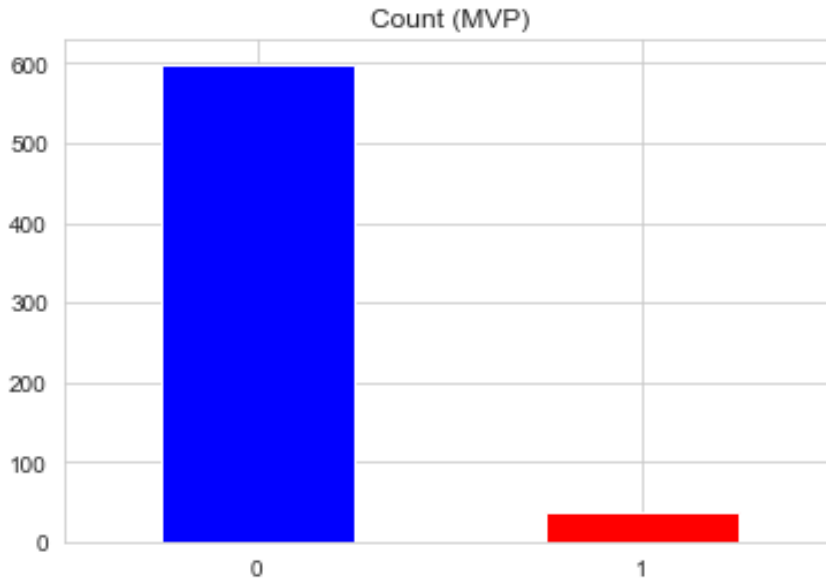
Regression Models	RMSE	R^2 (%)
Regression Ensemble	0.156	61.4
Random Forest	0.160	59.2
SVR	0.169	54.6
Linear Regression	0.181	48.7

- Models used: Linear Regression, SVM regressor, Random Forest regressor and a Stacking ensemble
- 10-fold cross-validation
- Metrics: R-square (R^2), Root Mean Square Error (RMSE).
- Best performer: Stacking ensemble



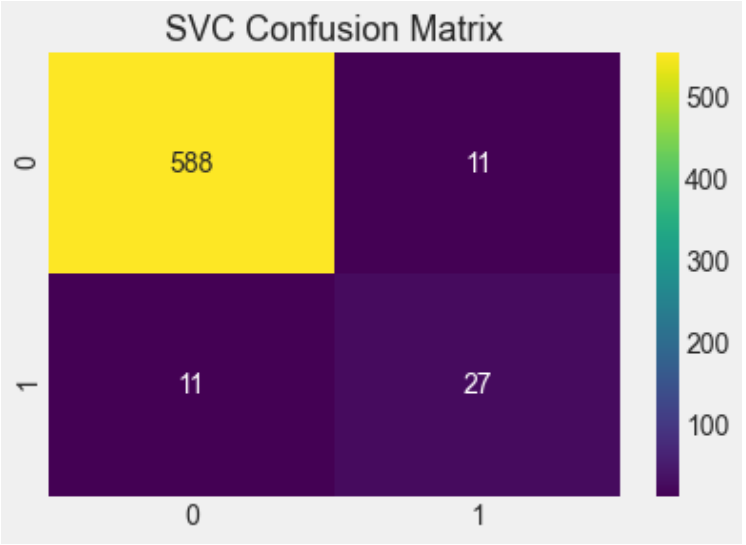
Classification

- Binary classification problem (MVP/ not MVP)
- Class imbalance: 94% negative class, because there can only be one MVP in each season
- Feature selection: same process as in regression (statistical test ANNOVA) resulted in choosing the same features as in regression analysis
- Predicting the probability a player is classified in the minority class



Classification

- Challenges: Only one MVP per season -> use of classifying models that can calculate probability
- Models used: Logistic Regression, Linear Discriminant Analysis, SVC, Random Forest, Naïve Bayes and Stacking classifier
- Precision as evaluation metric, due to the imbalance in classes and confusion matrix for better visualisation
- Leave-one-group-out cross-validation



Classification Models	Precision (%)
SVC	71.05
Random Forest	65.79
Classification Ensemble	63.16
Logistic Regression	60.53
LDA	60.53
Naïve Bayes	57.89

Final thoughts

- The MVP is not awarded exclusively on performance merits, which are reflected in the statistical data used in this experiment. Other factors, such as popularity, expectations and team success, are important too.
- Future project: scrape own data and create a data set which includes features that can reflect the abovementioned factors.

- Thank you for your attention!