NATIONAL RESEARCH
UNIVERSITY

Big Data and Machine Learning Project

- NBA MVP Prediction -

By Said Magomedov

February 5, 2024

Contents

Abstract

The aim of this study is to examine the NBA Most Valuable Player (MVP) award, a highly debated topic in the sports community. After data preprocessing gradient boosting, random forest, elastic net linear regression and finally ensemble algorithms will be implemented. Performance evaluation will be based on RMSE. The choice of this topic is motivated by two factors. Firstly, the MVP award considers numerous variables and does not necessarily identify the best, but rather the most valuable player. Therefore, predicting the winner requires a comprehensive analysis of multiple factors (11 for this project), making machine learning techniques an appropriate tool. Secondly, the author has a personal interest in basketball, having played it since childhood. This is indeed helpful as understanding of some specific factors is crucial.

1 Introduction

Many of those who are somewhat far from basketball and generally not interested in this game may have the following questions: who earns the title of the league's Most Valuable Player, and how is this distinguished individual chosen?

Initially (actually it all happens at the end of the regular season), a comprehensive list of candidates, typically comprising 10-20 exceptional players from the league, is compiled. This selection is based on various criteria including statistical performance, team contribution, personal achievements, and the collective opinion and recognition of the players. Following this, the MVP is determined by evaluating who has made the most significant impact on their team throughout the season. This assessment considers consistent excellence across a spectrum of performance metrics, regardless of their team's overall standing.
The MVP award is decided through a voting process conducted by a panel of approximately 100 sportswriters and broadcasters from the United States and Canada. The recipient of the award is the player recognized for bringing the greatest value to their team, demonstrating exceptional performance and leadership throughout the NBA season.

Thus, this study aims to predict the MVP by modeling and forecasting the players' vote share: the highest share among all candidates $\Rightarrow$ the award is presented. The set of various game statistics throughout the season will serve as independent variables. They will be further elaborated on in the following chapter.

Given that the target variable is quantitative, this study addresses a regression problem. First, I build predictions on a test sample using five base machine learning algorithms: XGBoost, AdaBoost, CatBoost, Random Forest, Elastic Net linear regression. Second, I implement an ensemble stacking algorithm using simple linear regression in order to combine obtained base predictions together and probably get a higher forecast quality. Third, I analyze the feature importance and select only relevant and most significant variables. Then I repeat previous steps on a truncated sample to compare the results which I expect to become better and more precise.

2 Data Description

The target variable which is share of points won by a player in the contest, is defined as follows:

$$Share_{i,t} = \frac{PointsWon_{i,t}}{PointsMax_t}$$

The major difficulty through data preparation process was data web scrapping using the BeautifulSoup4 module since there was no open source datasets containing all the features I wanted to include in my model.
After parsing the tables from the official website of basketball stats and historical statistics `Basketball Reference`, I removed the independent variables that were not of interest to me and left only the following 11:

- G – number of games in which the player participated
- PTS – points per game
- TRB – rebounds per game
- AST – assists per game
- STL – steals per game
- BLK – blocks per game
- FG – field goal percentage
- 3P – 3-point FG
- FT – free throw percentage
- WS – win shares, an advanced statistic roughly speaking representing the player's contribution to team wins
- POS – categorical variable responsible for the player's gaming position: 0 for Guard, 1 for Forward and 2 for Center

As a result I obtained a dataframe with 499 observations and 18 columns. Why 18? I have got 1 dependent and 11 independent variables and 6 additional columns: Rank, Player, Age, Position, Tm and Year – player's prize place, name, age, gaming position, team and year of competition.
Note that although the number of observations is 499, number of unique players is only 144, because many players participate in the competition for several years in a row. Time period (variable Year) is from 1991 to 2023, i.e. 33 regular seasons.

As for treatment of outliers in the data, feauture of this study allows us to omit this step. Let's give an example: some player who plays as a center has significantly fewer points and assists than others, and he didn't shoot three-pointers at all or hit very few. Someone may want to remove this player from the sample, which would be a grave mistake, since this player will have much higher other important indicators (typical of a center): blocks and rebounds. That is the specificity of the game of basketball.

When checking the data for missing values, it was revealed that there are some NaNs in the 3P column. It is because some players have not even tried to shoot three-pointers -– all of them played as a power forward or center, who are not actually supposed to shoot from beyond the arc. I decided to replace them with zeros. Of course, one can argue the fact that the absence of attempts and misses in 100% of cases (when the value of 3P is zero) are generally speaking non-identical, since having no attempts is better than being a total "muff", but for the purposes of this study I will omit this minor detail.

In order to preliminarily assess the influence of independent variables on the dependent one and then make statistical hypotheses, I will construct scatter diagrams.

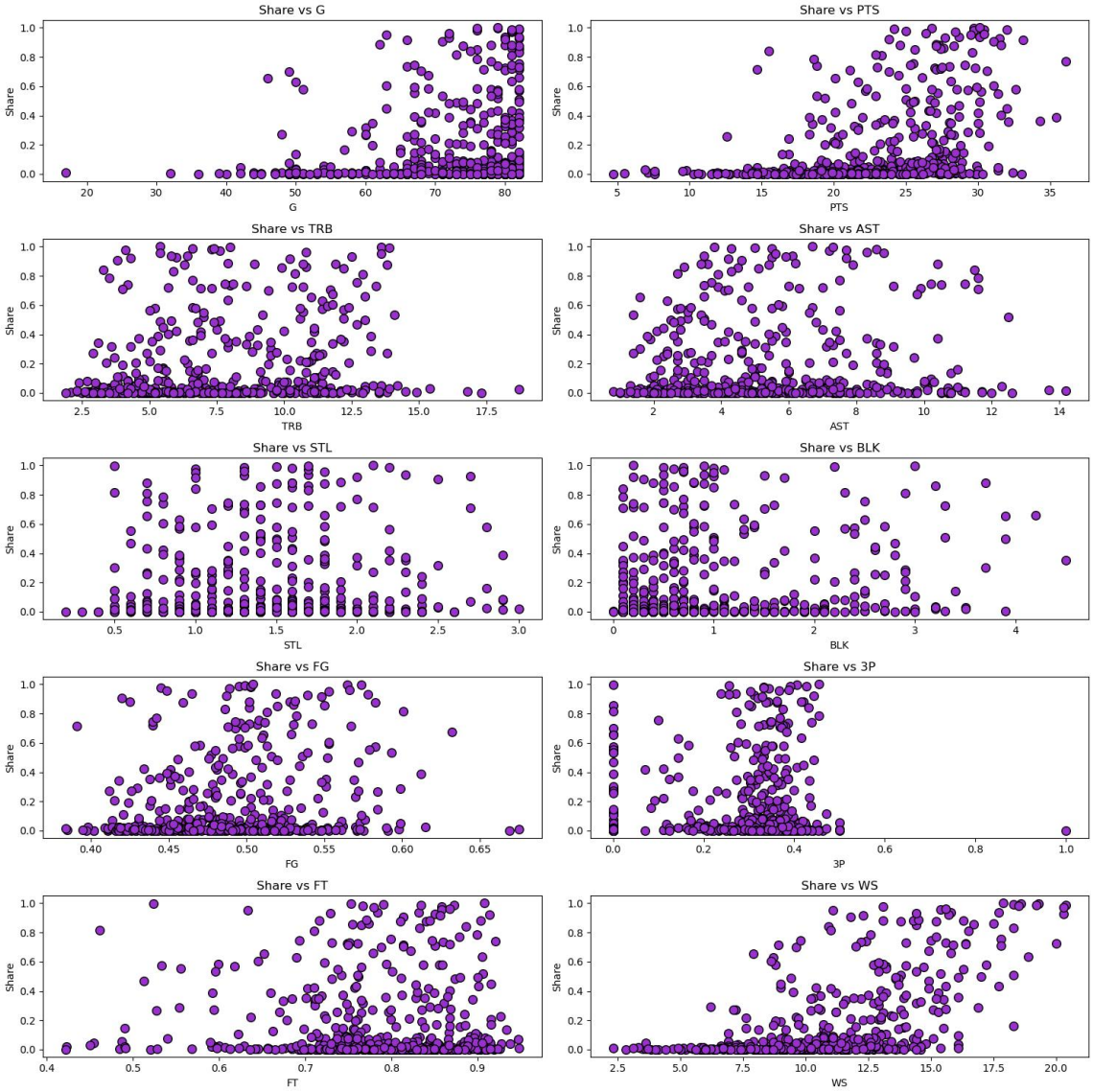

Figure 1: Scatterplots of Shares with respect to each independent variable

The only variables that show any significant relationship with our target variable are PTS, AST, FG and WS. Let's make the following hypothesis:

"All of these four stats have a strongly significant positive effect on Shares, while win shares have the greatest one."

It is also worth showing one interesting fact related to MVPs' age that I discovered during the data analysis: it has a slightly downward trend, which was calculated using the Hodrick-Prescott filter.

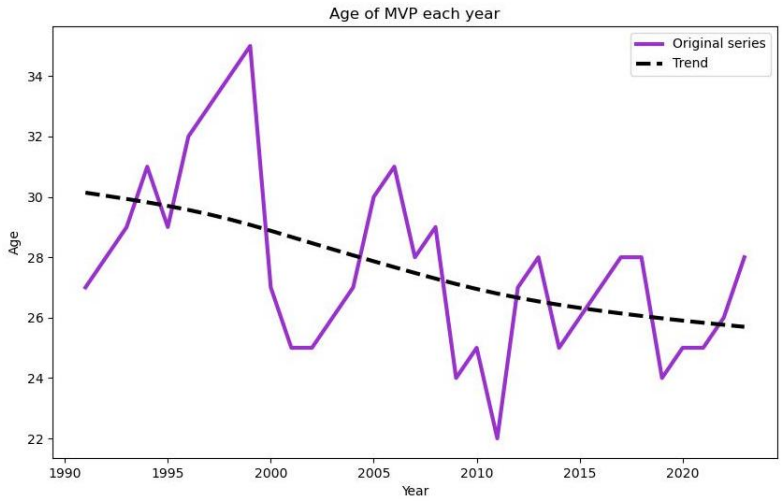

Figure 2: Plot of Age of MVP for each year with trend by Hodrick-Prescott filter

It may be the case that previously, gaming experience played a decisive role in player's success, but over time, priority began to be given to younger and physically stronger talents.

Now let's look at the correlation matrix of all our variables.

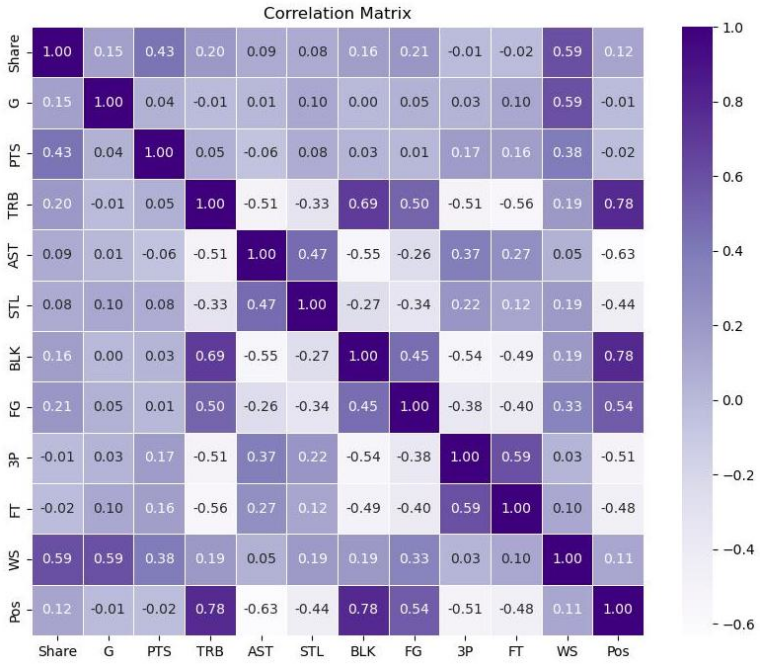

Figure 3: Sample Correlation Matrix

The correlation matrix allows to make a supposition about multicollinearity problem. For example, BLK and TRB factors are strongly correlated with POS. This fact makes sense as special types of player's actions are more frequent for particular gaming positions. To ensure the problem exists it would be reasonable to conduct VIF tests during the further

research stage. Nevertheless, Elastic Net is able to mitigate multicollinearity by constraining parameters.

Here the descriptive statistics of the whole data are revealed.

	Age	Share	G	PTS	TRB	AST	STL	BLK	FG	3P	FT	WS	POS
mean	27.72	0.17	72.65	22.55	7.51	5.27	1.39	0.96	0.49	0.30	0.78	10.82	0.76
std	3.89	0.27	10.49	5.20	3.27	2.69	0.53	0.89	0.05	0.13	0.09	3.33	0.76
min	19.00	0.00	17.00	4.70	1.90	0.80	0.20	0.00	0.38	0.00	0.42	2.30	0.00
25%	25.00	0.00	68.00	19.30	4.80	3.10	1.00	0.30	0.46	0.27	0.74	8.70	0.00
50%	27.00	0.02	77.00	22.90	7.00	4.90	1.40	0.60	0.48	0.33	0.80	10.70	1.00
75%	30.00	0.24	81.00	26.60	10.25	7.00	1.70	1.30	0.51	0.38	0.85	12.90	1.00
max	38.00	1.00	82.00	36.10	18.70	14.20	3.00	4.50	0.68	1.00	0.95	20.40	2.00

Table 1: Descriptive statistics of all variables

The share, BLK and POS variables have quite a high standard deviation, other factors seem to fluctuate not as much. The values of quartiles are represented in order to use 3IQR outliers diagnostics and omit irrelevant observations.

The following step of analysis is target variable density construction.

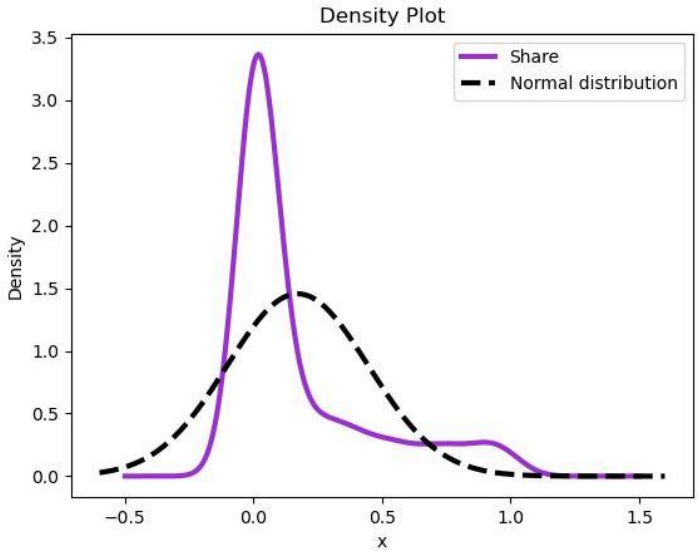

Figure 4: Density of the target variable Share along with Normal distribution

Distribution of Share is positively skewed and has thin tails in comparison with normal distribution. In case of traditional statistic approaches it may be reasonable to standardize variables.

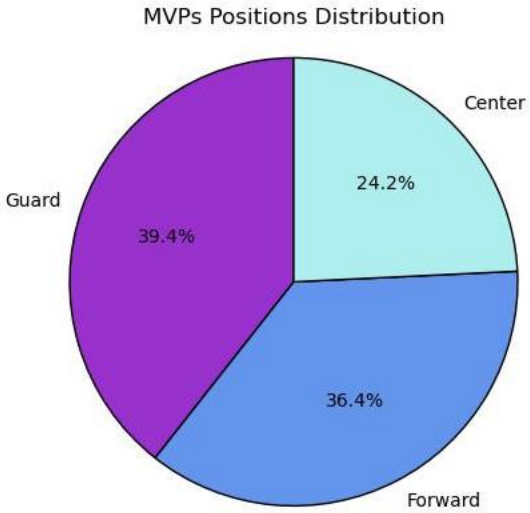

Figure 5: Shares of MVPs positions

The gaming position identifier is added to the list of variables as MVP award assumes the most prominent player, which may be dependent on position as well. For instance, guard players are generally the most active during the game. Consequently, the effect of such players is more obvious. This statement goes in line with the pie chart graph below.

According to the latter analysis, the following hypothesis can be made:

"The players with guard gaming position are more likely to get the MVP award in comparison with forward and center players."

3 Building Models

3.1 Full Sample

Goal of this study is to predict the MVP for the season 2022-2023, which will be the testing sample. At this stage a forecast for the last season was built using five base machine learning algorithms: XGBoost, AdaBoost, CatBoost, Random Forest and Elastic Net linear regression. Reported metrics are the cross-validation (CV) and out-of-sample (OOS) root mean squared errors. In the context of model selection, RMSE took precedence over R^2. This choice was driven by the research topic's underlying assumption that the variance of the explained variable is not as meaningful as the precision inherent in the forecasts.
The ensemble stacking is done using the linear regression as a combining model and in-sample predictions from the five base algorithms as regressors. Then the forecast for the test sample is computed based on the out-of-sample predictions. It is done this way since the whole training sample is not large enough to divide it into two independent ones: one for training base algorithms and another – for the meta-model.
It's worth noting that prior to constructing the models, the data was scaled using the `MinMaxScaler` technique to be able to correctly interpret the results (as we have linear regression as a base and stacking model).

Original	XGBoost	AdaBoost	CatBoost	Random Forest	Elastic Net	Stacking
0.915	0.457	0.589	0.389	0.422	0.405	0.661
0.674	0.408	0.650	0.450	0.391	0.503	0.784
0.606	0.277	0.210	0.233	0.207	0.250	0.175
0.280	0.143	0.177	0.182	0.153	0.233	0.155
0.046	0.318	0.279	0.250	0.291	0.299	0.236
0.030	0.147	0.151	0.159	0.111	0.126	0.148
0.027	0.144	0.136	0.192	0.188	0.254	0.073
0.010	0.208	0.198	0.253	0.186	0.324	0.171
0.005	0.128	0.085	0.136	0.122	0.188	0.025
0.003	0.139	0.131	0.149	0.123	0.283	0.090
0.002	0.082	0.054	0.061	0.076	0.009	0.007
0.001	0.101	0.090	0.062	0.074	0.084	0.053
0.001	0.088	0.081	0.082	0.076	0.057	0.044
CV RMSE	0.214	0.213	0.211	0.213	0.211	–
OOS RMSE	0.216	0.182	0.227	0.227	0.232	0.168

Table 2: Original values of Share in 2023 and its predictions by different algorithms

XGBoost and Random Forest both correctly predict the first two prize places. RMSE is almost the same for all base models, but AdaBoost demonstrates a relatively smaller one – 0.182.
3-fold cross-validation metrics are also very similar: Elastic Net and CatBoost perform better than others but unfortunately do worse on the test sample misclassifying the MVP by awarding the actual silver medalist .

Since the dataset contains some categorical variables, CatBoost seems to be a suitable approach. The method is designed to handle categorical features more efficiently than the classic gradient boosting.
However, according to the RMSE criterion, it turned out that the classic approach provides more precise forecast. The reason is probably the nature of almost all explanatory variables, they are numeric and, thus, the adjustment for categorical variable does not result in precision increase. It may be the fact that gaming position has no significant effect on the target variable.

The relatively modest performance of Elastic Net should not be overly emphasized, as its implementation was originally motivated by the aim to address multicollinearity in the data. Its primary purpose is to mitigate multicollinearity, thereby partially reducing dimensionality, while also facilitating the testing of stated hypotheses.

3.2 Truncated Sample

In this section, I trim the sample by narrowing down the set of regressors to the most important and significant ones. Subsequently, I iterate through the process of executing the

algorithms on the trunacted sample and contrast the resulting predicted values with those obtained from training models on the complete set of features.
In order to select those features let's look at the feature importances returned by each algorithm.

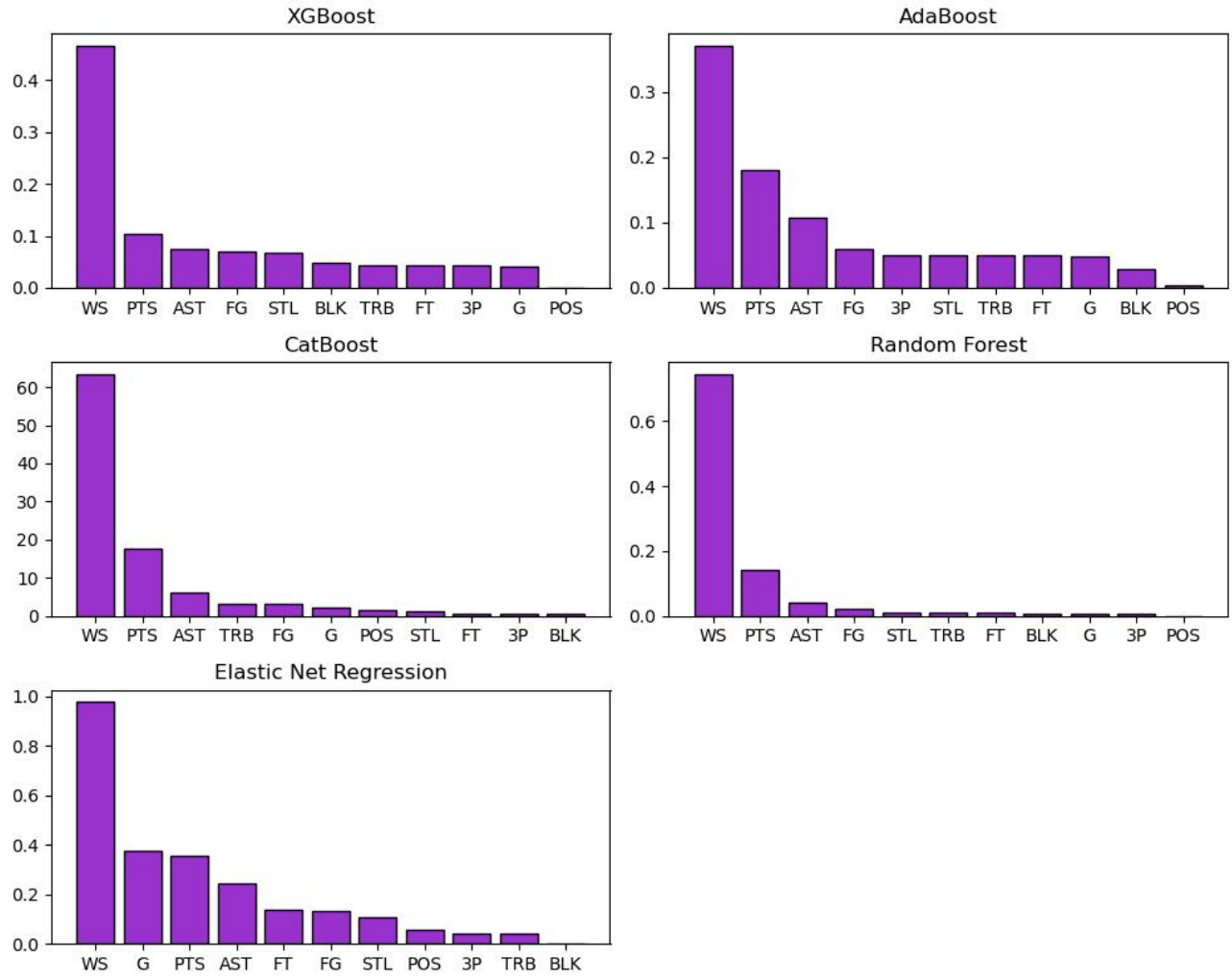

Figure 6: Feature Importances

BLK, TRB, 3P and POS demonstrate the least significance among all estimated models, thus they will be removed.
WS, PTS, AST and FG, on the contrary, turned out to be the most important – the first hypothesis is confirmed. Actually that is almost true since only the FG coefficient is negative in the Elastic Net linear regression:

G	PTS	TRB	AST	STL	BLK	FG	3P	FT	WS	POS
-0.378	0.357	0.044	0.245	-0.107	0.000	-0.131	-0.044	-0.137	0.977	0.056

Table 3: Elastic Net linear regression estimated coefficients

The second hypothesis is also confirmed since POS coefficient is very small (0.056) and is the least important in all other models except CatBoost which slightly increases it since this algorithm is supposed to catch the categorical features' influence better. The least important in this context means that it is better not to include this feature in the model, i.e. set it to zero – remember that POS equals zero for Guards, hence, the hypothesis is not rejected.

Original	XGBoost	AdaBoost	CatBoost	Random Forest	Elastic Net	Stacking
0.915	0.467	0.494	0.368	0.622	0.364	0.762
0.674	0.441	0.427	0.452	0.559	0.441	0.633
0.606	0.247	0.179	0.265	0.441	0.267	0.590
0.280	0.143	0.151	0.183	0.116	0.214	0.081
0.046	0.340	0.338	0.278	0.373	0.262	0.428
0.030	0.152	0.131	0.154	0.092	0.100	0.066
0.027	0.157	0.190	0.171	0.178	0.216	0.178
0.010	0.169	0.163	0.242	0.273	0.339	0.312
0.005	0.127	0.133	0.162	0.117	0.163	0.097
0.003	0.146	0.173	0.158	0.155	0.245	0.151
0.002	0.081	0.100	0.091	0.034	0.014	-0.000
0.001	0.103	0.103	0.086	0.070	0.114	0.065
0.001	0.081	0.097	0.091	0.070	0.088	0.061
CV RMSE	0.214	0.212	0.209	0.217	0.211	–
OOS RMSE	0.216	0.225	0.230	0.176	0.244	0.167

Table 4: Original values of Share in 2023 and its predictions by different algorithms on the truncated sample

Performance of XGBoost, CatBoost and Elastic Net linear regression has not changed at all (same predictions of the first two places and (almost) RMSEs on cross-validation and test sample). AdaBoost correctly predicted first two places but with a much higher out-of-sample RMSE (0.182 $\Rightarrow$ 0.225) – it is not a big problem since my aim was just to predict who will be the MVP and AdaBoost did it well on a truncated sample. Random Forest and Ensemble Stacking algorithms turned out to be the most outstanding ones on a truncated sample as they accurately predict the winners podium (actual MVP and silver and bronze medalists too) with a smaller OOS RMSE – 0.176 and 0.167 respectively.

4 Results and Further Development

To summarize let's note that:

1. Both hypotheses were strongly confirmed: (1) PTS, AST, FG and WS are the most important features and (2) Guards are more likely to get the MVP award.

2. Preferred base algorithm is Random Forest with the first two and three leaders predicted correctly on full and truncated sample respectively.

3. Ensemble Stacking did well only on truncated sample having the 5% smaller out-of-sample RMSE compared with Random Forest.

4. The feature selection led to a significant improvement in the predictive power of the models: OOS RMSE of Random Forest was reduced by 22.5% and winners podium was predicted precisely as well as when using Ensemble Stacking.

In order to further advance the research, it may be useful to explore the application of the implemented methodology for predicting awards in other sports. By applying the same methodology to different sports, I can gain a deeper understanding of its effectiveness and potential limitations. This could also provide valuable insights into the varying factors that contribute to award predictions in different sports.

Additionally, adjusting different train-test ratios can be a valuable area of exploration. By using the number of seasons in the test sample as a hyperparameter in Grid Search, I can optimize the model's performance. This could lead to more accurate and reliable predictions, ultimately enhancing the overall effectiveness of the methodology.