Which Model Best Predicts Hall of Fame Status for NBA Players

Rocky Rowell

2024-12-16

Abstract

The study attempts to find the best regression / machine learning model to predict NBA all-stars' hall of fame status from their career totals in counting statistics such as points, rebounds, assists and more. The data includes the hall of fame status and career total counting statistics for all NBA all-stars up until 2017. The models tested in this study include: logistic regression, additive model, regression tree and random forest. The models are compared based on classification performance metrics like AUC, sensitivity and specificity. Along with this, the logistic regression and additive models will be compared by AIC values as well. After performing K-Fold Cross-Validation with 10 folds to get mean values for the performance metrics, the additive model appears to perform better than all other models with the logistic regression model not too far behind. Comparing those last two models by AIC, we find that the additive model is a better fit overall solidifying the conclusion that the additive model is the best model for predicting hall of fame status for NBA all-stars.

Introduction

The goal of this project is to test the performance of 4 separate regression / machine learning models for predicting the hall of fame status of NBA all-stars. This will be done with multiple career total counting statistics as the predictors. After running the models, they will be compared based on their classification performance metrics. These metrics include AUC, sensitivity and specificity.

The Predictors Used (Career Totals): Games Played, Offensive Win Shares, Defensive Win Shares, Win Shares, Field Goals Made, Field Goals Attempted, Free Throws Made, Free Throws Attempted, Total Rebounds, Assists, Personal Fouls, Points

The Models Used:
- Multiple Logistic Regression
- Additive Model
- Regression Tree
- Random Forests

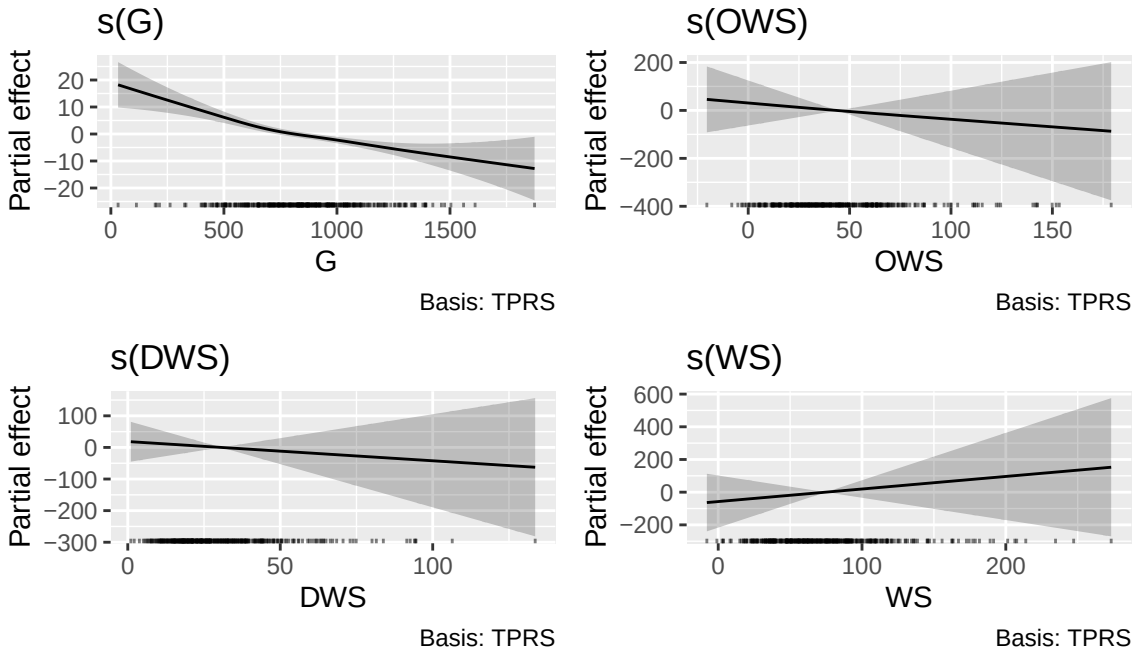
Data

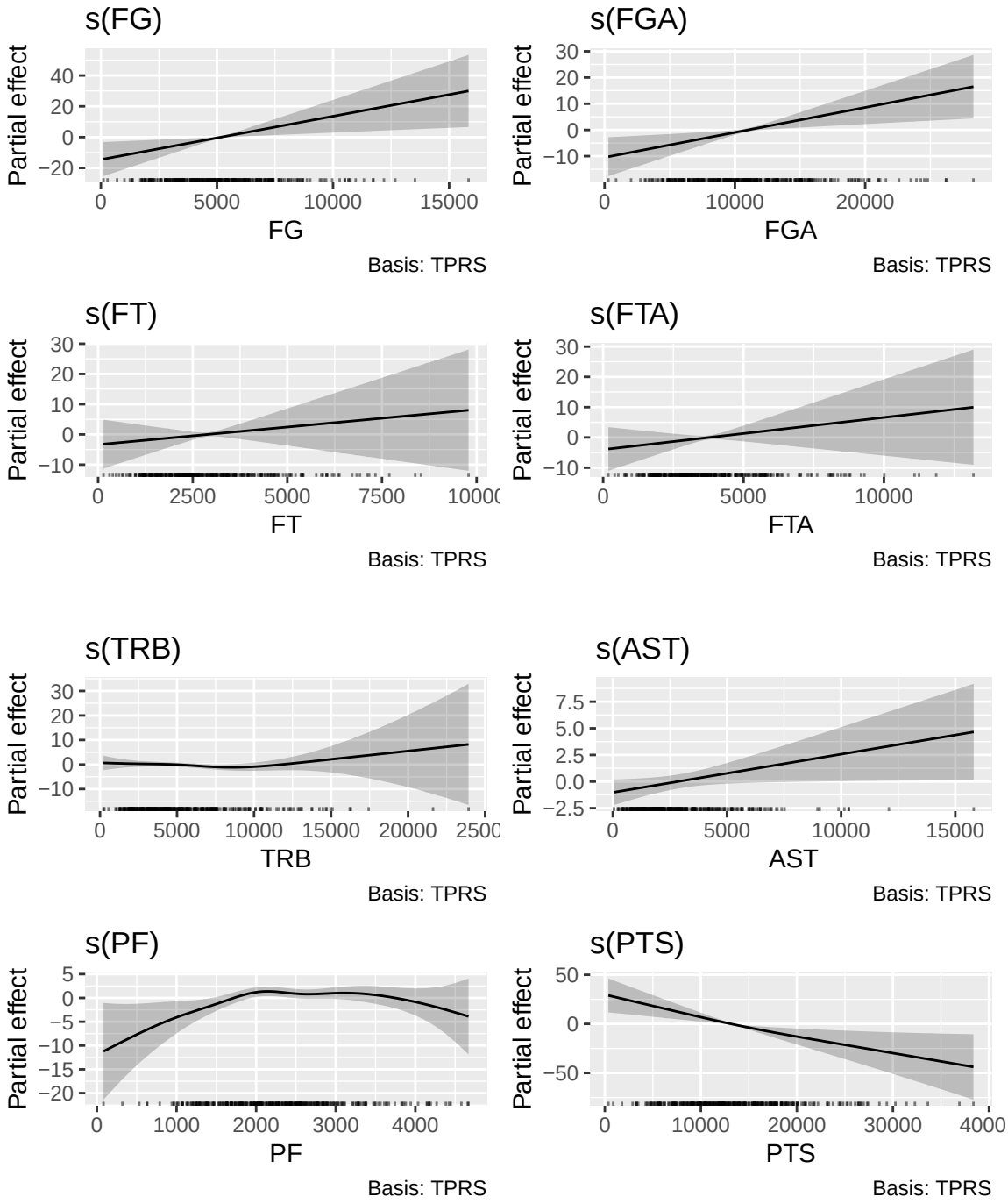
Cleaning the Data

The dataset used contains career totals in statistics listed above for all NBA all-stars up until 2017. This was combined with a separate dataset that was a list of all NBA Hall of Famers. A value of 1 was given to all players in the hall of fame dataset and 0 was given to all players not in the dataset. The last step of preparation for the study was to remove variables that were not needed or contained too many NA values due to the NBA not recording statistics such as blocks and steals till 1973 and the 3 point line not existing until 1979.

Testing Linearity of Predictors

Before going into the models and their output, it is important to analyze the linearity of these predictors. Since we have 1 parametric model, 1 semi-parametric model and 2 non-parametric model, the linearity of these variables can give us insight into the results we recieve later. The best visualization of this are the ones created for the smooth terms in the additive model:





For all the predictors used in this study, linearity seems to be good assumption except for personal fouls and possibly total rebounds. Due to this, we expect the logistic regression model and additive model to be better fits. To test this we will use K-Fold Cross Validation

Methods

K-Fold Cross-Validation

These are the mean values for the classification performance metrics: AUC, sensitive and specificity. This is done across 10 folds for all 4 models in the study. Below are the results:

Table 1: Model Performance Metrics

	AUC	Sensitivity	Specificity
Logistic Regression	0.903	0.948	0.639
Additive Model	0.906	0.924	0.731
Regression Tree	0.715	0.821	0.561
Random Forest	0.843	0.897	0.570

AIC for Logitstic Regression and Additive Model

```
## [1] "The AIC for Logistic Regression is: 246.28"
```

```
## [1] "The AIC for Additive Model is: 215.443"
```

Results

According to the table, the additive model is the best overall, especially in terms of accuracy, with logistic regression not too far off. All 4 models had good values for sensitivity and all struggled more with specificity, but the additive model and logistic regression model still have better values than the regression tree and random forest model. Comparing the two models based on AIC, the additive model is slightly better than the logistic regression model.

Based on these results, parametric and semi-parametric models are a better fit for the data than non-parametric models since the predictors are linear, except for personal fouls. Of those two models (logistic regression and additive model), the additive model is the best since it approaches the problem of personal fouls' non-linear form.

Conclusion

This study shows that model selection is very important. Despite random forests and regression trees being more advanced forms of machine learning, they failed to classify hall of fame status better than the less extensive models. These results rely on counting statistics' linear relationship with hall of fame status with slight variations for variables like personal fouls and total rebounds being better fit with the additive model. If someone wanted to predict hall of fame status for NBA all-stars, the additive model is the best model available.