DSC 550

Weeks 11 and 12

Project Milestone 5

Author: David Brehm

06/03/2021

**Case Study Report**

**Introduction**

I have always been interested in the NBA. Over the past several years, there has been an increased amount of focus on data and statistics in all sports leagues, but especially the NBA. This has likely had a large impact on my initial interest in the field of data science. For individual projects to practice my programming skills or learn new concepts, I tend to choose NBA data to work with.

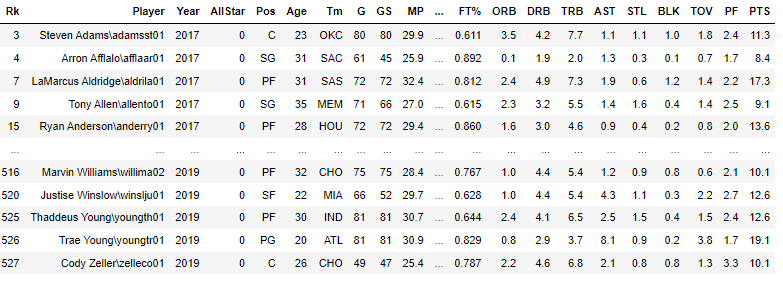
Every year, a number of NBA players are selected as all-stars for each conference. The fans, players, and media members all vote on this process to determine the all-stars for the year. There is usually controversy afterwards with certain players argued as being snubbed or being an all-star undeservedly.

For this project, I wanted to look at a few years of historical data to see how well basic NBA stats could predict which players are all-stars. I used 2017 through 2019 to train my models, and tested them on the current 2021 data.

**Data**

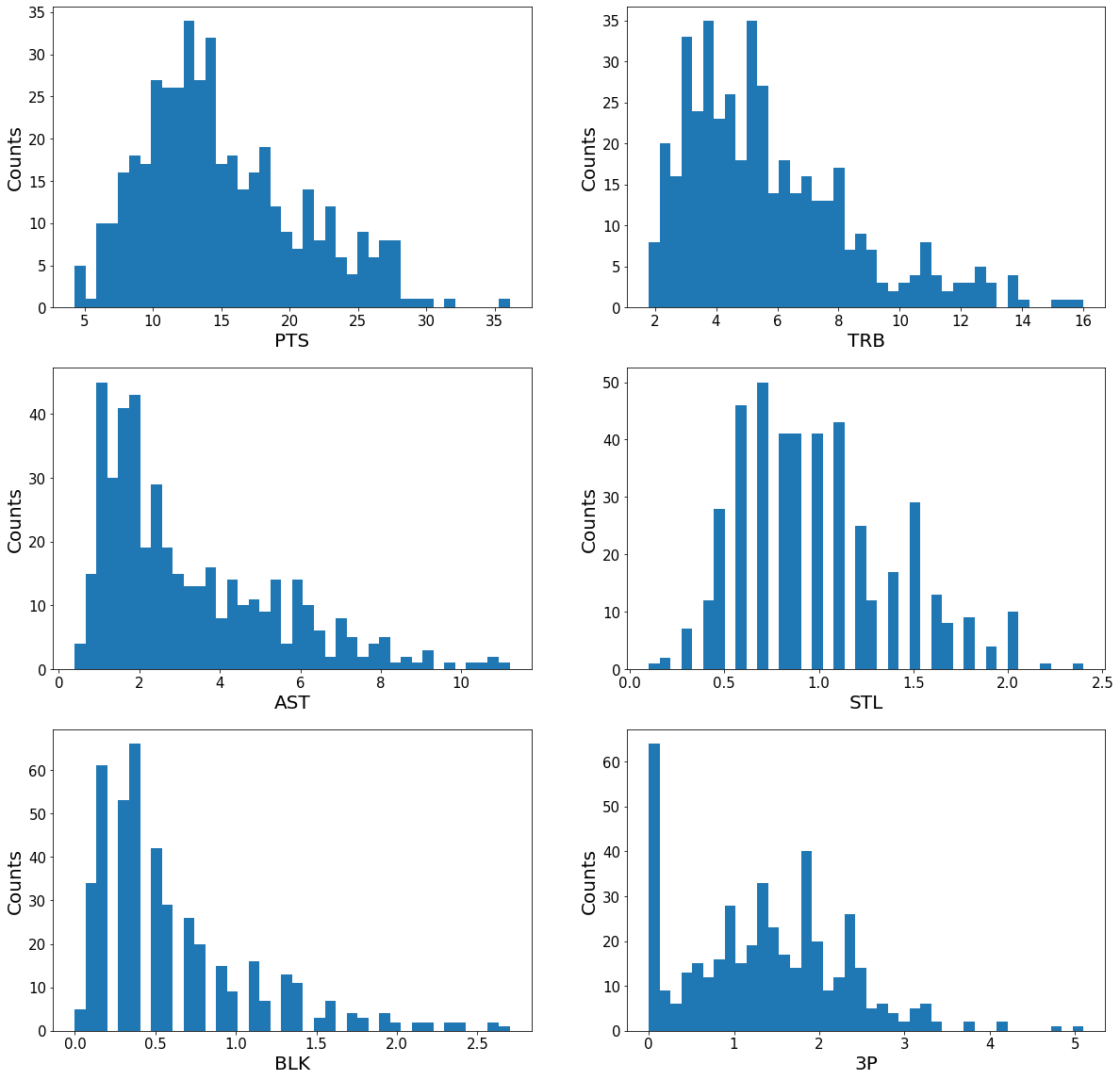
The data used for this project was all gathered from a site called Basketball Reference. This site contains a plethora of data and endless possibilities for analysis. For this project however, I decided to focus on basic player stats which consists of well-known stats such as points, rebound, and assists.

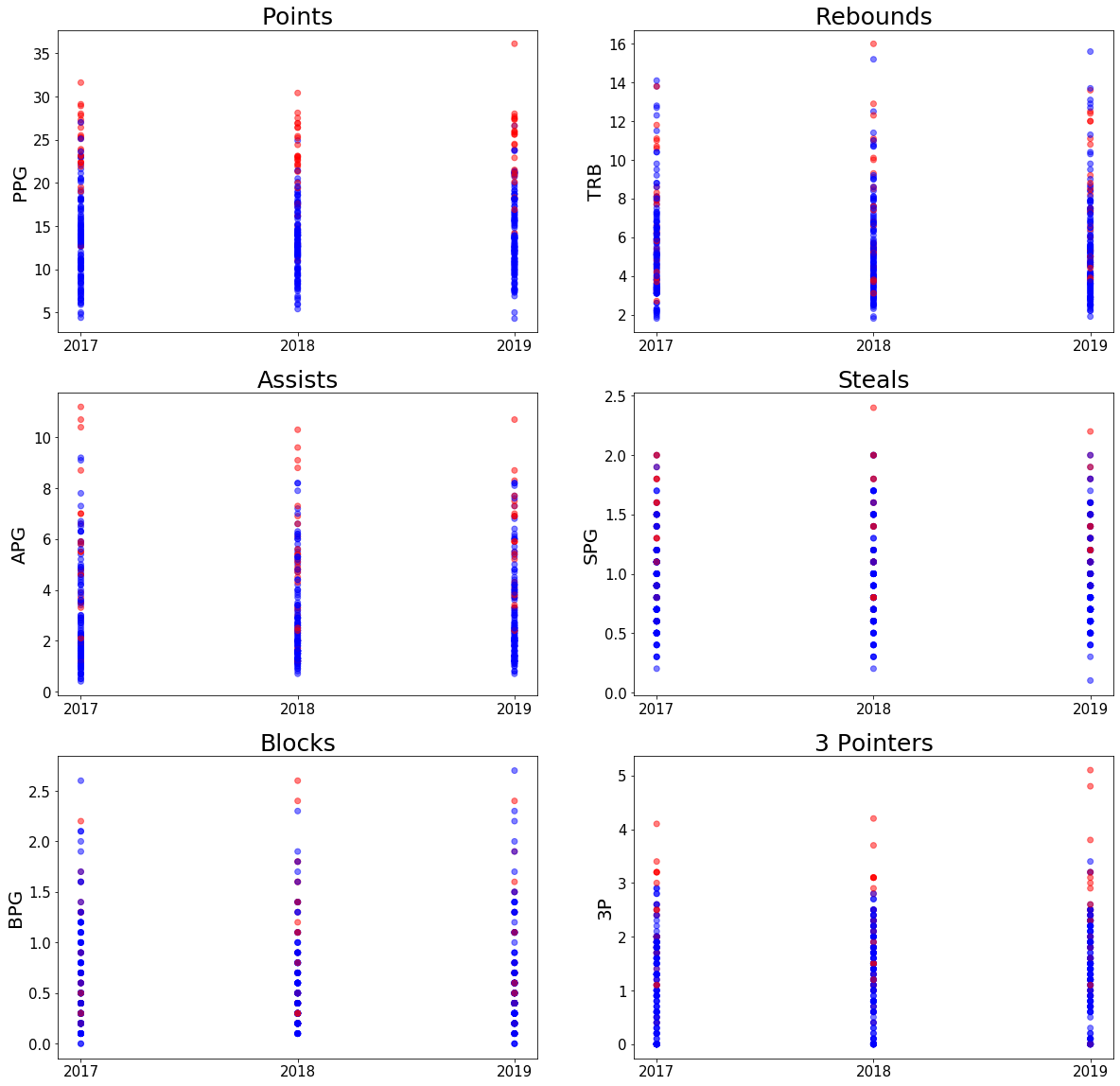
As an initial filtering process, I reduced the data down each year to players who had started at least half of the games. Many of the fields are either clearly not needed such as ‘Rk’, player name, or team name, and others are very similar which can be eliminated through feature reduction. The latter will be elaborated on in the following section.



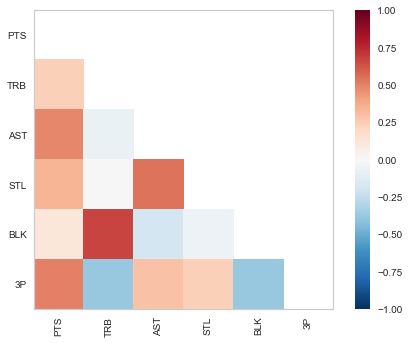
**Graphical Analysis**

Histograms and scatter plots were created to examine the distribution of data. For the scatter plots, all-stars are red. As expected, all-stars tend to do well in these basic box score stats. They do not tell the whole story, but performing well across the board here indicates clear production.



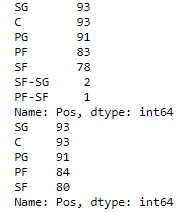
****

A correlation was also performed on the basic box score stats. This indicated that blocks and rebounds have a high correlation, which shouldn’t be alarming to anyone who follows the sport.

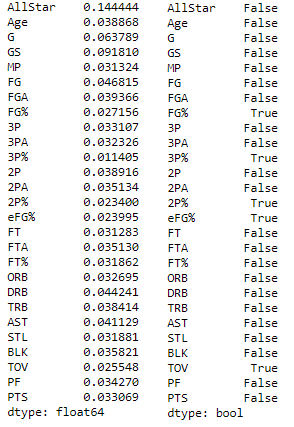


**Dimensionality & Feature Reduction and Feature Engineering**

Out of the features that passed the initial reduction, the only categorical variable remaining was player position. Examining this, there were a few unique points where a player had two positions. In these cases, I decided to keep the first position value out of the two. Dummy variable columns were also created for each of the five remaining positions.

****

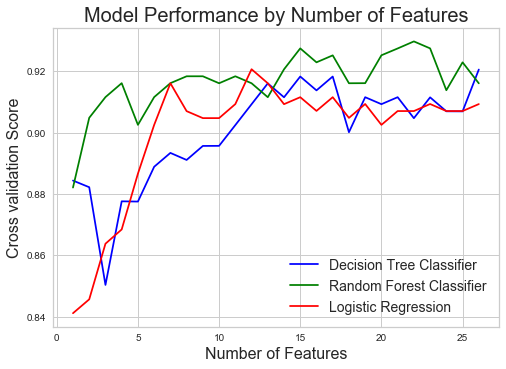
One method of feature reduction that was utilized was looking at a variance threshold. Using this, the FG%, 3P%, 2P%, eFG%, and TOV variables were eliminated due to low variance. After this, I used recursive feature elimination along with specific models. This will be described in the Model Selection and Evaluation section.



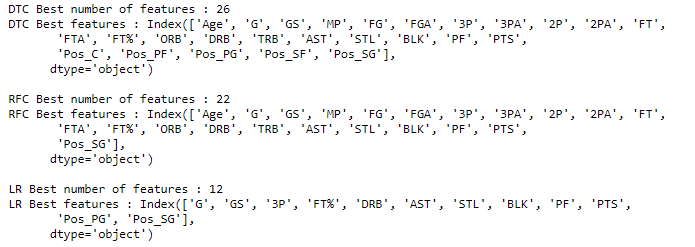
**Model Selection and Evaluation**

Three models were selected for this project. Logistic Regression, Decision Tree, and Random Forest. These were chosen because this is a binary classification problem, whether someone is an all-star or not. Those three algorithms are among the most popular binary classification algorithms along with Naive Bayes, KNN, and SVM.

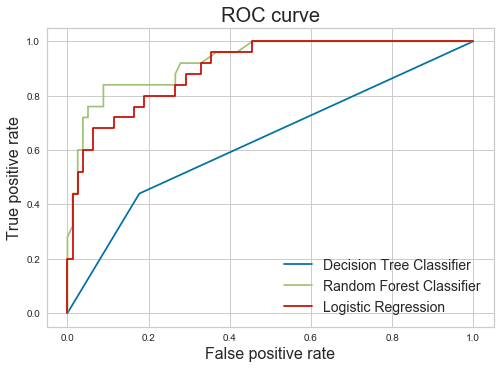
The algorithms were used to select the best features with recursive feature elimination, as mentioned previously.



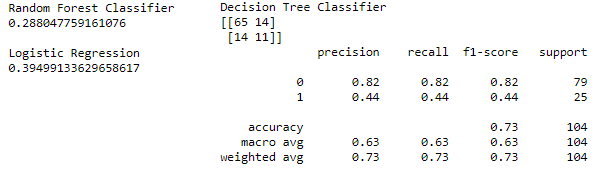
The best features can be tracked through this. From the above image, below were the best number of features and those exact features.



After the algorithms and their features were finalized, they were then tested with 2021 data. Receiver operating characteristic curves, or ROC curves, were created to examine the tradeoff between true and false positives. The ROC curve graph follows the random feature elimination cross validation scores, with Random Forest having he largest area under the curve.



The log loss for the two models that returned probability outputs (Random Forest and Logistic Regression) along with the confusion matrix and classification report for the Decision Tree Classifier were also examined.



**Conclusion**

This was an interesting project to work through. I did not have much experience with the Decision Tree and Random Forest algorithms before this program, so those were interesting to learn and see in action. As shown in the Model Selection and Evaluation section, the Random Forest Classifier was the best choice here to predict NBA all-stars from basic NBA statistics.

**Source**

*Basketball Statistics and History*. (n.d.). Basketball-Reference.Com. Retrieved June 3, 2021, from https://www.basketball-reference.com/