**Predicting NBA All Stars Using Season Stats**

SCMT 650: Applied Predictive Analytics

Dr. Michael Ketzenberg  
Group 3

Authors: Benjamin Kyle Clinton, Rahul Vincent Gerald, and Varun Sagar Gajendragadkar

**Table of Contents**

[Introduction](#_u6u1jqiy3glb) 2

[Previous Studies and What’s New](#_kg86z1jlslz2) 2

[Dataset Description and Initial Data Handling](#_ewcxgcaspweg) 3

[Dimension Reduction](#_xi70mmialt8o) 4

[Models Created](#_r6dtdt49h1xv) 4

[Logistic Regression](#_w37go6qfd89p) 5

[KNN Classification](#_579d6onercjj) 12

[Decision Trees](#_511mqmaplzdp) 15

[Random Forests](#_d07ms5nje5f6) 17

[Evaluation of Models & Conclusion  
  
 The models used are compared and evaluated based on the precision values found.](#_cf9j1guznmgh) 20

[Future Scopes](#_7hatejn8zj60) 20

[Works Cited](#_3fdg8lrgeptj) 20

[Appendix](#_yeavrgncvixf) 21

[Variable Definitions](#_jj0uafu3kdld) 22

[Formulas Used and Definitions](#_30egrunwy9eu) 23

# Introduction

Predictive analytics in sports is an interesting and rapidly growing application of machine learning. “Over the past few years, the world of sports has experienced an explosion in the use of analytics” (1). Machine learning algorithms are being used to create models that predict everything from the best defensive baseball positioning based on a batter’s characteristics to how successful a scouted rookie will be in the National Basketball Association (NBA). “The sports industry uses sports analysis to increase revenue, improve player performance and a team’s quality of play, prevent injury and for many more enhancements” (2). However, the use of predictive analytics in the sports industry does not stop on the sports teams side of things. Fans are huge consumers of predictive analytics. As the world shifts towards more data driven decisions, sports fans are increasingly taking advantage of machine learning, especially in the sports betting sphere. “Applying this tech to the sports betting industry opens a plethora of opportunities for the future market of tipping, betting and bookmaking alike” (3). It is obvious that predictive analytics has a place in the sports industry, and predictions will only continue to improve as technology advances and more data is collected.

Our project takes a specific look at the NBA. The goal of our project was to predict the NBA All Star team based on player stats from the season. We used readily available data from the NBA for the years 2000 to 2016. We built multiple models using different methods and algorithms, including logistic regression, decision trees, random forests, and K-nearest neighbors (KNN). We then evaluated and compared the different models. The best models and methods are discussed later in this report.

The information provided by the models is valuable. The models created could be used by the NBA All Star voting committee to determine which players should be All Stars, although the models might be poor at handling the politics often associated with this process. Fans could apply these models to make bets about the All Star team in advance of the voting, called futures. Our models have important practical applications, but that is not the only information they provided. We also gained more knowledge about the types of metrics the NBA uses to assess players and got a better understanding of predictive methods.

# Previous Studies and What’s New

“Using machine learning to predict the 2017 NBA All Star rosters” (4) used a decision tree model to attempt answering the All Star roster question. It is important to note that this study used Python to produce the model. This study also uses data dating back to 1980. This study identified FGM, FTA, Points, FGA, Mins, Rebounds, Assists, TS%, Blocks, and EFG% as the best predictors. This model predicted 20/24 predictions correctly.

The study, “Predicting the NBA All-Stars and NBA Awards with Machine Learning” (5) also attempts to answer our question using Python. The report is vague, and really more of an outline, but it is important to note that this study used deep learning methods to produce a model. That is a difference from our study as we did not explore deep learning methods. Deep learning will be discussed in our future scopes for this project.

“Using Machine Learning to Predict NBA All-Stars” (6) is yet another study that aims to answer our question. This study, like ours, used RStudio to build models. This study only used a logistic model with 700 records, split into 550 records for training and 150 records for testing. The model was able to predict 134/150 predictions correctly on the test set, which is almost 90% accurate. Interestingly, this study mentions how popularity plays a role in All Star selection but does not offer a solution to code popularity into the model.

An important difference from the first two studies discussed is the language we used to create our models. Using R as opposed to Python, our results will differ from the first two studies, although the methods are similar. The first two studies do not give any indication of how many records were used or any implication of training and testing set splits, which is disappointing. However, the last study discussed does explain the amount of records used and how the data is split. For our study, we are using more than 10 times the amount of records the third study used, discussed later in the report. We also explored more models than the previous studies discussed. Lastly, these previous studies report high measure of total accuracy. However, we know that in classification problems with rare events, like All Stars in the NBA, overall accuracy may not be the best measure for determining the effectiveness of the model. For that reason, our study focuses on precision, defined as true positive responses divided by total positive responses, and recall, defined as true positives divided by true positives and false negatives. Our goal with precision is to discover what proportions of positive responses were actually correct (7). The goal with recall is to determine what proportion of actual positives was identified correctly (7). Ideally, there is a perfect classifier with high precision and high recall, but that is not how it works in reality. Since one of hopes for this model is to be used with sports betting, we decided that we want to focus more on predicting true positives with the lowest false positive rate (so we do not make bad bets and lose money). Therefore, the precision measure will be most important when we compare models later in the report.

# Dataset Description and Initial Data Handling

The data used in this project was taken from 5 different data sets and merged into one single data set. The data sets are as follows:

* MVP.csv - contains the names of players who have won the MVP award for seasons 2000 to 2017
* NBA All Star Games.csv - contains the names of players who have been all stars from seasons 2000 to 2017
* Players.csv - this contains stats of all the players who have played from 1950, such as field goals, height, weight, games etc.
* Salary.csv - contains the salary of all the players from season 2000 to 2017

The data is cleaned first by removing any records with missing values, and then merged to create one data set called “seasonStats.csv,” which contains details of all stats of players by year and if they were an all star or not. If a player is an all star, the all star column is equal to 1, otherwise it is equal to 0. After cleaning the data, we were left with 7,612 records and 57 predictors. The percentage of All Stars with respect to the rest of the league is about 5% on a per year basis (about 6% of the entire dataset are All Stars).

## Dimension Reduction



This figure depicts the method used for dimension reduction. Using all 57 predictors was leading to some singularity errors. Also, some of the original predictors were not useful for the project, like name, year, and position. Therefore we arrived at this correlation matrix using 45 variables that were relevant. To reduce the number of dimensions we wanted to exclude variables that were highly correlated. We highlighted correlations that were greater than 0.8 and less than -0.8, as we determined that these levels of correlation were too high. We were able to explain the data set’s variance with only 27 predictors, thus resolving the singularity issues and reducing runtime on the algorithms.

# 

# 

# 

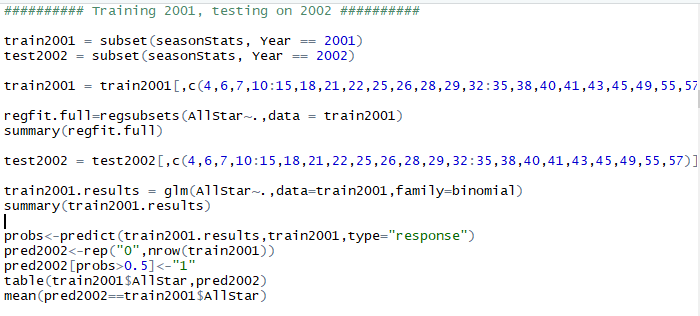
# 

# Models Created

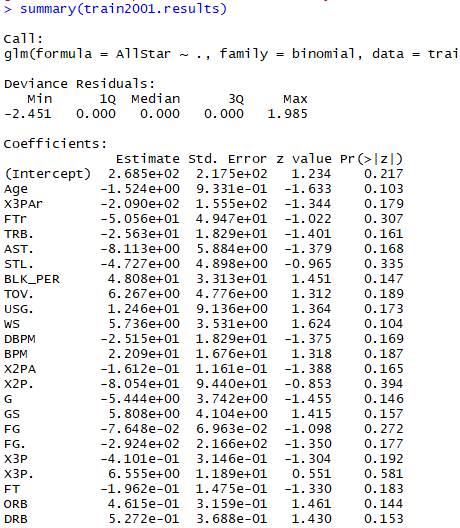
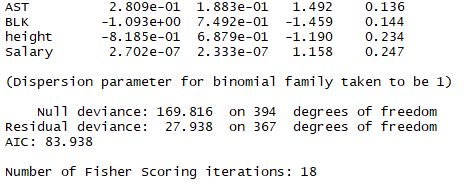
To predict the players who would be selected as an All Star, our group decided on using four classification techniques. The four used were logistic regression, K nearest neighbors, decision trees and random forests.To choose the best model, the three models were then compared based on its precision. Each model was executed using it’s specific library as well function in “R” Studio.

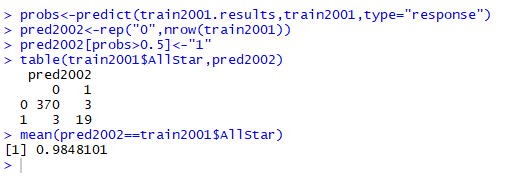
## Logistic Regression

Logistic regression is used to predict probabilities of binary outcomes, given some predictors. It is a popular classification model used in predicting the likelihood of an object belonging to a class based on certain predictors. This model has many applications and can be used in classifying customer churn, borrower default, customer return rate etc. It is more interpretable than other classification methods.

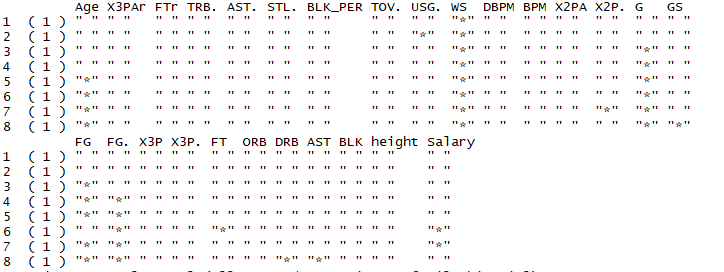
For classifying players as all stars, first we took every alternate year data as training and test sets and created 17 models. Example: year 2000 as training and 2001 as test, followed by year 2001 as training and 2002 as test. This is done using all 27 predictors.  
  
The code snippet was as follows (taking year 2000 as train and year 2001 as test):  


The summary is as follows:

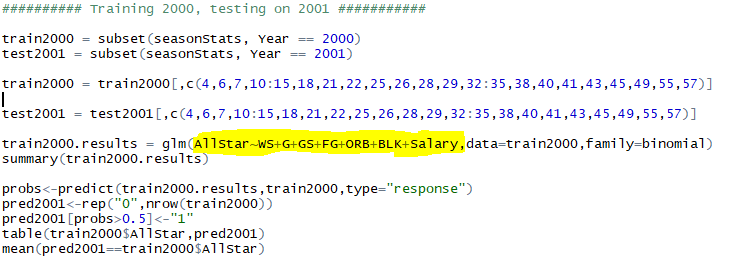
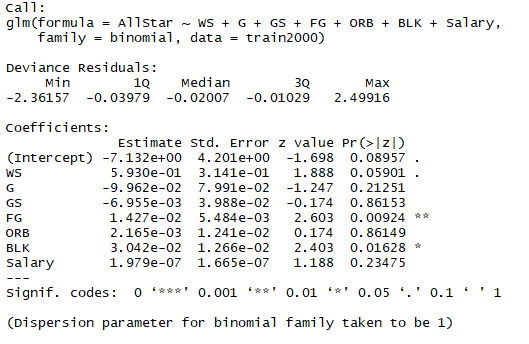
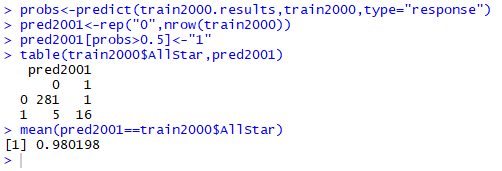
  
  
  
We see that, since all predictors were used we do not get to know which ones were good at predicting the class of an All Star.

The confusion matrix and accuracy found are as follows:  
  
The model has an accuracy obtained is 0.98.  
  
The same modelling function is used for 17 models on the remaining of the data sets (i.e. taking alternate years as training and testing data sets) and the following accuracy was obtained:

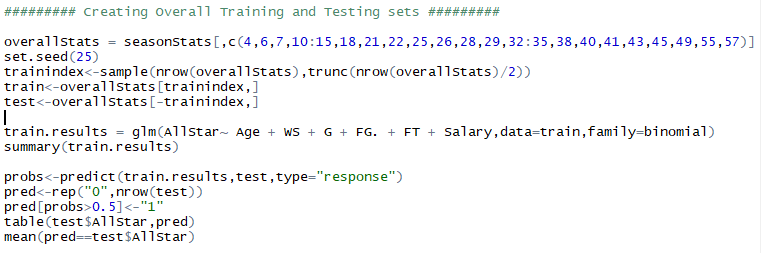
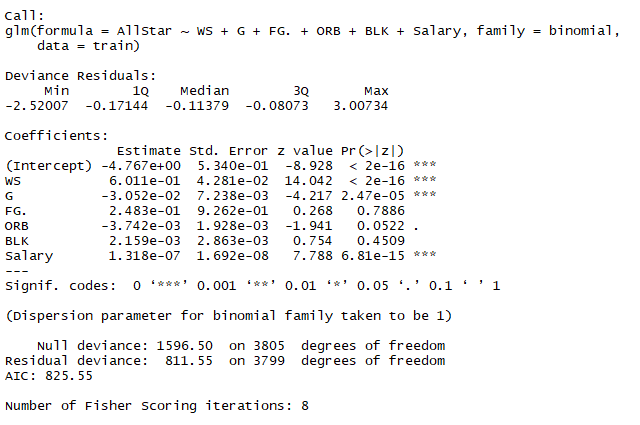
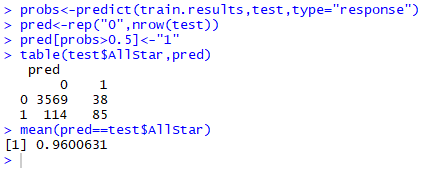
|  |  |  |
| --- | --- | --- |
| **Train Data Year** | **Test Data Year** | **Accuracy** |
| 2000 | 2001 | 0.89 |
| 2001 | 2002 | 0.95 |
| 2002 | 2003 | 0.94 |
| 2003 | 2004 | 0.95 |
| 2004 | 2005 | 0.96 |
| 2005 | 2006 | 0.93 |
| 2006 | 2007 | 0.94 |
| 2007 | 2008 | 0.92 |
| 2008 | 2009 | 0.93 |
| 2009 | 2010 | 0.96 |
| 2010 | 2011 | 0.97 |
| 2011 | 2012 | 0.92 |
| 2012 | 2013 | 0.95 |
| 2013 | 2014 | 0.96 |
| 2014 | 2015 | 0.95 |
| 2015 | 2016 | 0.96 |
| 2016 | 2017 | 0.95 |

This type of modelling did not give us a clear picture of the accuracy since all 27 predictors were used, where in a few of them may be irrelevant in predicting the All Star. Hence, the r “regsubset” function was used to identify only the predictors which are relevant in predicting which player will be an All Star.  
  
The code snippet is as follows:  
  
  
  
  
From the above screenshot it is clear that the following predictors are the best in predicting if a player is an All Star or not:

* WS - win shares
* G - games played
* GS - games started
* FG - field goals scored
* ORB - offensive rebounds
* BLK - blocks
* Salary - the salary earned that season

Now, the logistic regression model is use in predicting the All Star using only the above mentioned 7 best predictors. Also, note that as before 17 models are created using alternate years as training and test data sets. An example of the code snippet is as follows:  
  
   
  
Only the best predictors are used as shown in the screenshot above. The summary and confusion matrix is as follows:  
  
  
  
  
Win share, games played, field goals and blocks are good predictors in classifying if a player is an All Star or not. This model has an accuracy of 0.98.  
  
The same modelling function is used for 17 models on the remaining of the data sets (i.e. taking alternate years as training and testing data sets) using only the 7 best predictors and the following accuracy was obtained:

|  |  |  |
| --- | --- | --- |
| **Train Data Year** | **Test Data Year** | **Accuracy** |
| 2000 | 2001 | 0.94 |
| 2001 | 2002 | 0.96 |
| 2002 | 2003 | 0.95 |
| 2003 | 2004 | 0.96 |
| 2004 | 2005 | 0.97 |
| 2005 | 2006 | 0.96 |
| 2006 | 2007 | 0.95 |
| 2007 | 2008 | 0.97 |
| 2008 | 2009 | 0.94 |
| 2009 | 2010 | 0.97 |
| 2010 | 2011 | 0.96 |
| 2011 | 2012 | 0.95 |
| 2012 | 2013 | 0.95 |
| 2013 | 2014 | 0.95 |
| 2014 | 2015 | 0.97 |
| 2015 | 2016 | 0.96 |
| 2016 | 2017 | 0.97 |

Next, the logistic model was tested for the entire data set using only the best 7 predictors. The code snippet used is as follows:  
   
  
  
The summary as well as the confusion matrix is as follows:  
  
  
  
As seen from the summary, win shares, games played, field goals and blocks are good predictors of classifying players as all stars.  
  


The accuracy was found to be 0.96.  
  
From the confusion matrix we get the following:

True negative = 3569  
True Positive = 85

False negative = 114

False positive = 38

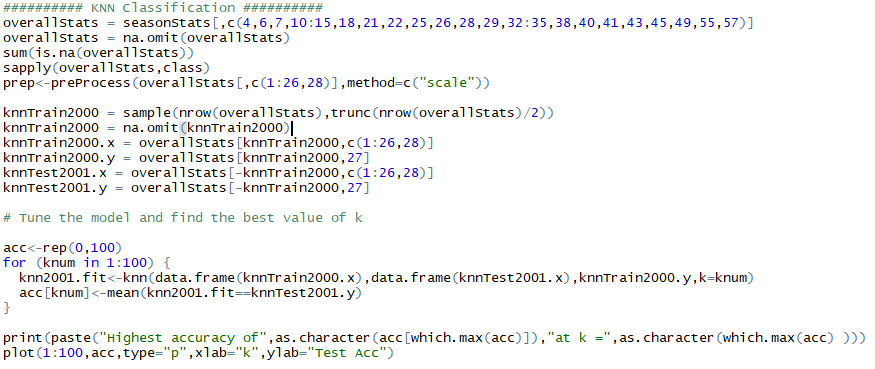
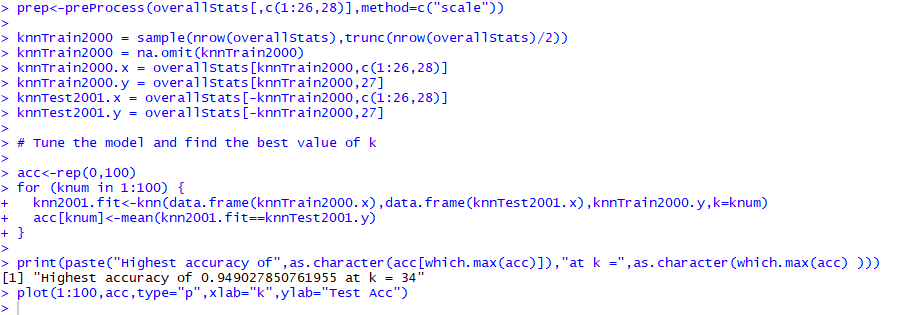
From the information above the precision of the model can be calculated using the formula:  
Precision = True Positives / Total Positives

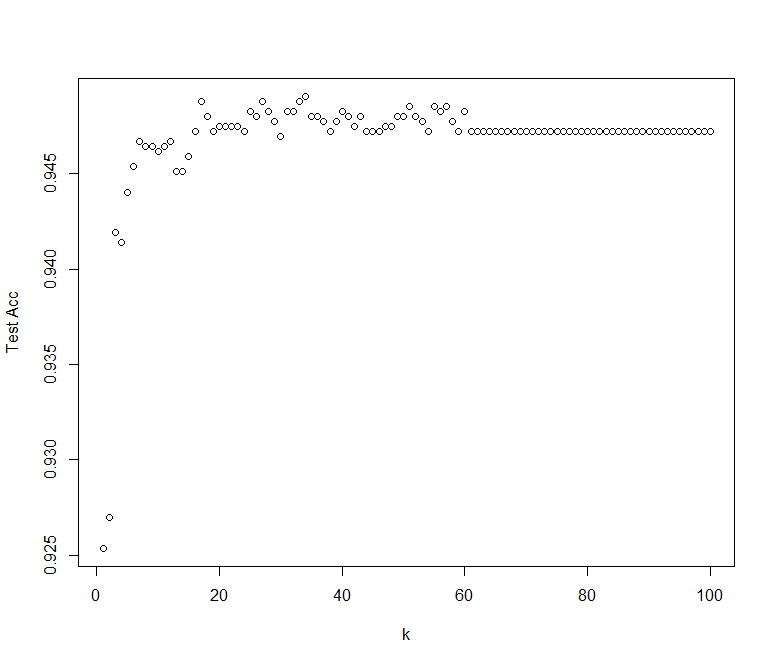
Precision = 85 / (85 + 38)  
Precision = 0.69

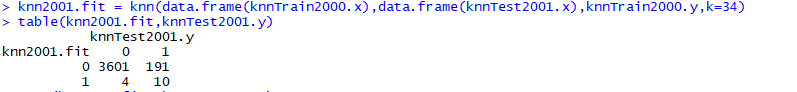
Hence, the precision of using logistic regression to predict and classify All Stars is 0.69.

## KNN Classification

KNN classification is used to predict the class of a given observation based on its nearest neighbors. It is compared with the k’th nearest neighbors. The class of the observation is decided based on the class of majority of the Kth nearest neighbors.  
  
Lower values of K provides a model which is flexible but has a high variance and low bias. High values of K used provides a model with low variance and high bias. Hence, the best K value is decided by taking the model with the best accuracy.

For this model, the entire data set was used. A snapshot of the code is as follows:  
  
  
  
The output was as follows:  
  


The best value of K was 34 and the model accuracy was 0.94. The graph of K value against accuracy is as follows:  
  


The confusion matrix for the best K = 34 for it is as follows:  
  


From the above confusion matrix we get the following:  
  
True negative = 3601

True positives = 10

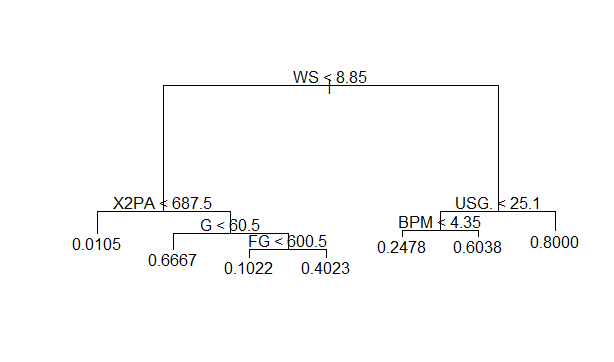
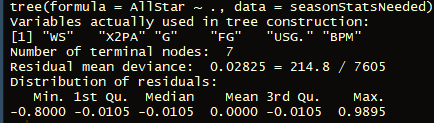
False negatives = 4

False positives = 191  
  
The precision of the model is calculated using the formula, precision = true positives / total positives  
  
Precision = 10 / (10 + 191)  
Precision = 0.0

## Decision Trees

Decision Trees are structures that use a tree like model of decisions and their possible consequences. While decision trees are relatively easy to implement and understand, they are often erroneous and are not the best method to predict.

After eliminating the variables that were pointed out to be highly correlated with others, the following output was obtained:



The important variables as found out by the model are:

* WS: Win Shares - Estimation of number of wins a player gets for his team
* X2PA: Number of 2 pointers attempted
* USG: The ball usage rate of a player
* G: Games played by a player
* BPM: Box Plus Minus

We then used stepwise classification, as the data for year 2000 was used as training data and data for year 2001 was used as testing data and so on. The following results were obtained:

|  |  |  |
| --- | --- | --- |
| **Train Data Year** | **Test Data Year** | **Accuracy** |
| 2000 | 2001 | 0.96 |
| 2001 | 2002 | 0.95 |
| 2002 | 2003 | 0.96 |
| 2003 | 2004 | 0.96 |
| 2004 | 2005 | 0.95 |
| 2005 | 2006 | 0.96 |
| 2006 | 2007 | 0.95 |
| 2007 | 2008 | 0.96 |
| 2008 | 2009 | 0.96 |
| 2009 | 2010 | 0.95 |
| 2010 | 2011 | 0.96 |
| 2011 | 2012 | 0.96 |
| 2012 | 2013 | 0.96 |
| 2013 | 2014 | 0.96 |
| 2014 | 2015 | 0.96 |
| 2015 | 2016 | 0.95 |
| 2016 | 2017 | 0.96 |

The following confusion matrix was found out after dividing the dataset equally into training and testing:



The accuracy was found out to be 0.95



From the confusion matrix we get the follow:

True negative = 3500  
True Positive = 126

False negative = 86

False positive = 94

From the information above the precision of the model can be calculated using the formula:  
Precision = True Positives / Total Positives

Precision = 126 / (126+94)  
Precision = 0.57

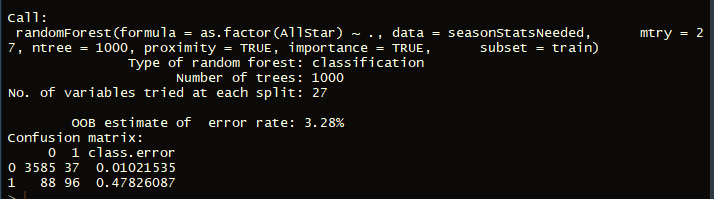
Thus the precision of Decision Trees in predicting All Stars is 0.57.

## Random Forests

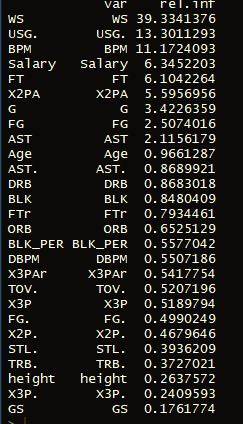
Random forests classification is a classification where multiple decision trees are constructed to give a mean of the prediction accuracy of all the decision trees.

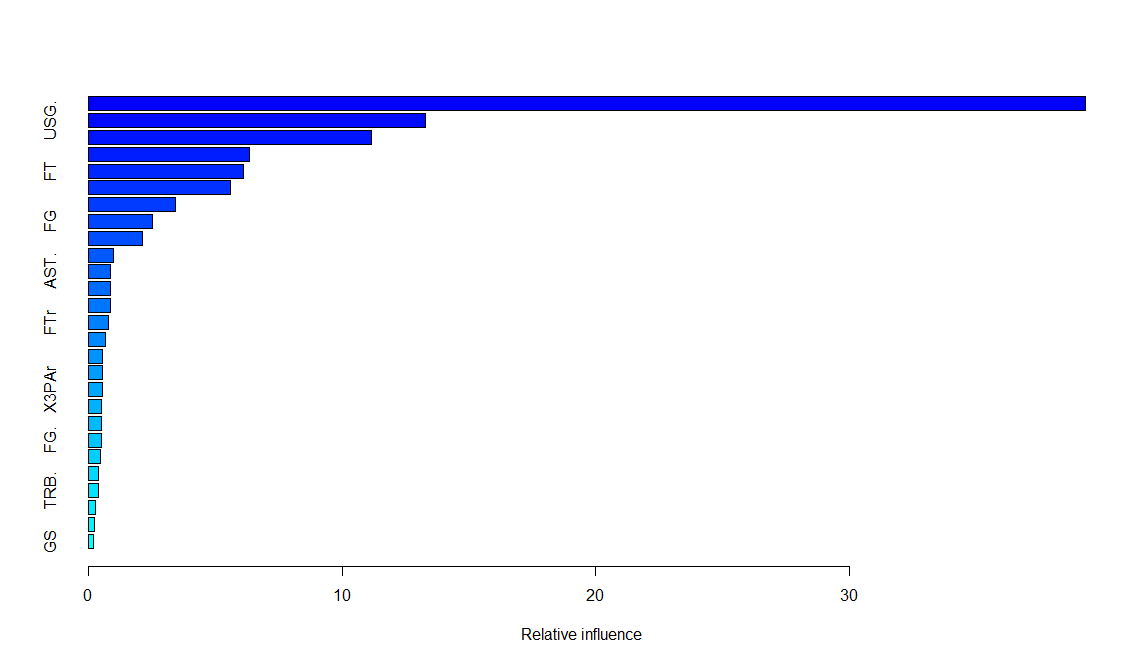
The dataset was divided into two equal parts. One was used as training and the other was used as testing data. The number of trees were chosen to be 1000.

The following output was obtained:



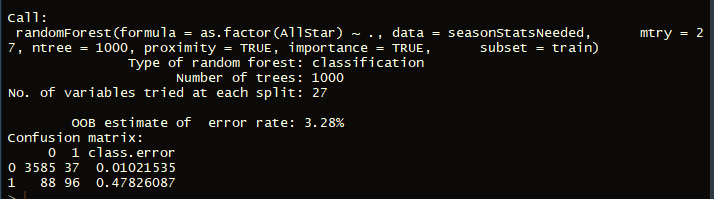
On Boosting, the following data was found:





According to the importance graph above, Win Shares, Box Plus Minus, Usage, 2 points attempted, free throws and salary were the most important predictors for this model.

For the entire dataset, the following confusion matrix was obtained:



The Accuracy was:



True negative = 3587  
True Positive = 96

False negative = 88

False positive = 37

From the information above the precision of the model can be calculated using the formula:  
Precision = True Positives / Total Positives

Precision = 96 / (96+37)  
Precision = 0.72

Hence the Precision of Random Forests to predict All Stars is 0.72

# Evaluation of Models & Conclusion The models used are compared and evaluated based on the precision values found.

|  |  |
| --- | --- |
| **Best Models from Methods** | **Precision** |
| Logistic Regression | 0.69 |
| KNN Classification | 0.05 |
| Decision Tree | 0.57 |
| Random Forest | 0.72 |

# 

According to the precision values obtained from our best models using the different methods, we can determine that Random Forest produces the best model for predicting the NBA All Star team. The 0.72 precision value means that when the Random Forest model predicts that a player will be an All Star, the model is correct 72% of the time. While 72% precision is good, there is still room for improvement to be discussed later. The main weakness of the Random Forest model is that the output is not vary interpretable compared to something like Logistic Regression. Currently, our model is limited to the data we have provided and catered to it, but we plan on overcoming this limitation and will discuss later. It is important to note that these precision values were calculated from the models using test data. With our validation accuracies being slightly higher than test accuracies, we believe that our models do not suffer from overfitting. Overall, the Random Forest model performed the best and has an acceptable precision value for the scope of our project goal.

# 

# Future Scopes

We would like to monetize these models in the future through the sports betting sphere. We also believe that our process and models could be adapted for use with other sports. It would be interesting to see how our models and process could also tackle problems outside of the sports betting world. In short, we would like to build a general program that can solve many problems with minimal user input. Expanding on our All Star selection problem, we would like to compare more machine learning methods and use more data to possibly improve the predictions. Lastly, we would consider using deep learning algorithms, such as ensemble, to make the best, most precise models possible.

# 

# Works Cited

(1)<http://analytics-magazine.org/beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i/>

(2)<https://www.samford.edu/sports-analytics/fans/2018/Why-is-Data-Analytics-So-Important-in-Sports>

(3)<https://www.bettinggods.com/machine-learning-in-sports-betting/>

(4)<https://medium.com/@griffinhoopes/using-machine-learning-to-predict-the-2017-nba-all-star-rosters-bbdd500f7ea5>

(5)<https://github.com/gmalim/NBA_analysis>

(6)<https://hoopsrenaissance.wordpress.com/2018/02/05/using-machine-learning-to-predict-nba-all-stars/>

(7)<https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>

# 

# Appendix

## Variable Definitions

Year - Season

Player - name

Pos - Position

Age - Age  
™ - Team

G - Games

GS - Games Started

MP - Minutes Played

PER - Player Efficiency Rating

TS% - True Shooting %

3PAr3 - Point Attempt Rate

FTr - Free Throw Rate

ORB% - Offensive Rebound Percentage

DRB% - Defensive Rebound Percentage

TRB% - Total Rebound Percentage

AST% - Assist Percentage

STL% - Steal Percentage

BLK% - Block Percentage

TOV% - Turnover Percentage

USG% - Usage Percentage

OWS - Offensive Win Shares

DWS - Defensive Win Shares

WS - Win Shares

WS/48 - Win Shares Per 48 Minutes

OBPM - Offensive Box Plus/Minus

DBPM - Defensive Box Plus/Minus

BPM - Box Plus/Minus

VORP - Value Over Replacement

FG - Field Goals

FGA - Field Goal Attempts

FG% - Field Goal Percentage

3P - 3-Point Field Goals

3PA - 3-Point Field Goal Attempts

3P% - 3-Point Field Goal Percentage

2P - 2-Point Field Goals

2PA - 2-Point Field Goal Attempts

2P% - 2-Point Field Goal Percentage

eFG% - Effective Field Goal Percentage

FT - Free Throws

FTA - Free Throw Attempts

FT% - Free Throw Percentage

ORB - Offensive Rebounds

DRB - Defensive Rebounds

TRB - Total Rebounds

AST - Assists

STL - Steals

BLK - Blocks

TOV - Turnovers

PF - Personal Fouls

PTS - Points

## Formulas Used and Definitions

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Accuracy = (TP + TN) / (TP + TN + FP + FN)  
Precision = (TP) / (TP + FP)