Predicting 5-Year Longevity in the NBA

Abhinav Singh

---

1. Introduction

The National Basketball Association (NBA) is a men's professional basketball league with 29 American teams and 1 Canadian team. It is widely considered to be the premier men's professional basketball league in the world. Players are the world's best paid athletes by average annual salary per player at about 5 million dollars annually. With thousands of colleges throughout the states, and thousands of foreign players aspiring to play in a league where athletes are guaranteed millions, the league is extremely competitive. With 30 teams receiving roughly two players per year, only 60 new spots each year are available out of the pool of thousands of athletes. And even if an athlete makes a ball club's roster, chances are that he will be gone within a few years.

With the advent of advanced machine learning, many ball clubs seek to utilize data analysis techniques to win the championship. This research project concerns a question that many players, recruits, coaches, and my statistics teachers would be interested in reading. The objective of this analysis is to answer how the number of games played, minutes, points, rebounds, assists, steals, blocks, and turnovers affect whether or not a rookie (first-year player) will last 5 or more years in the National Basketball Association.

To give some definitions to those unfamiliar with basketball, a rebound is when a player retrieves the ball after a missed shot. An assist is when the athlete passes to a teammate and he scores off the pass. A steal is when the player takes the ball away from the opponent's hands, and a turnover is when a player accidentally throws the ball to the other team.

2. Data

I used data from https://data.world/exercises/logistic-regression-exercise-1/discuss/logistic-regression-exercise-1/mezdgnjs. The user states that he aggregated data from the two following sites: https://data.world/gmoney/nba-rookies-by-min-1980-2016 and https://data.world/gmoney/nba-players-birthplaces. The data, with 1340 observations, includes a plethora of variables that will not be used. As mentioned earlier, the predictors are number of games played, minutes, points, rebounds, assists, steals, blocks, turnovers, and the outcome is whether the player lasts five years in the league. Each of these values are averages taken over the number of games the athlete played during the rookie season. The data from the previous link was fetched from NBA.com, the most reliable website for statistics. I did not have to do any sort of data cleaning or formatting as the data was already prepared to use, and all columns were filled with data.

From the dataset, the range of values of the number of games played is from 11 to 82, which is the maximum number of games possible, meaning that the player participated in all games of the season. The outcome variable is factored into its two dichotomous outcomes. A correlation

plot and histogram among the predictors is shown.

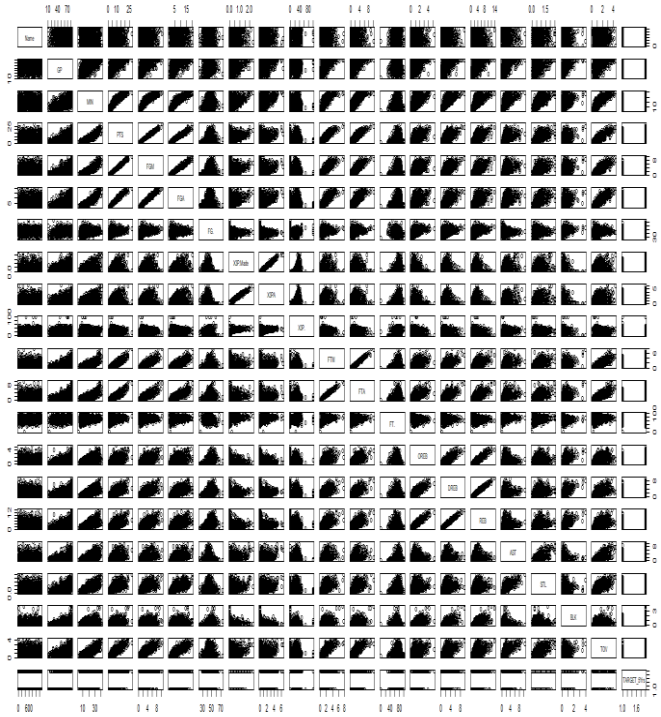
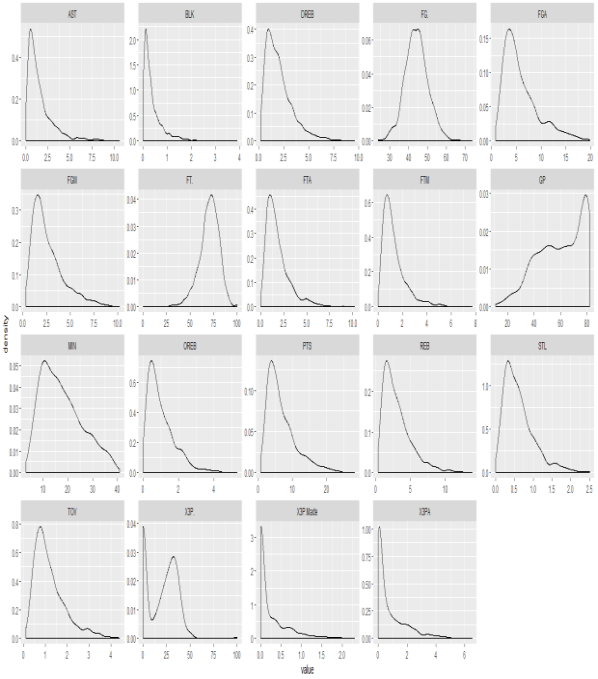


Figure 1 – Correlation Plot (above) and Histogram among Predictors (below)



3. Methodology

First, the necessary packages that were required were loaded and dataset was imported. Then, I checked the assumptions: the dependent variable is on a dichotomous scale, the observations are independent to a certain extent (this can't really be fixed anyways), and linearity of independent variables and log odds. After this, I started the logistic regression by creating the model, and found the confusion matrix. Then, I split the data into training and testing sets and determined the confusion matrix, and also utilized a cross-validation with 5 splits.

4. Discussion

Overall, the model does a relatively decent job at predicting whether or not a rookie will last 5 or more years in the NBA. Let's say someone goes out and presents a random person a photo of an NBA rookie, and asks a participant to predict whether the NBA player will last 5 years in the NBA. The chances of the person accurately guessing are 50%. The logistic regression revealed that without splitting into training/testing sets, the data was correctly classified about 70% of the time.

(unsplit data)	TARGET_5YRS	
GLM.PRED	0	1
0	269	159
1	240	672

Figure 2 – Classification rates

But this is an instance of overfitting, since the classifier has already seen the data. Thus, to reduce overfitting, after splitting the first 750 of 1340 observations into the training set, the data was correctly classified approximately 68% of the time in the testing set. Thus, there may be some overfitting. Using the first 750 observations for the training sets is fine because the data is random -- it's only sorted alphabetically. Moreover, after using a five-fold cross-validation, the accuracy was 70%, give or take one percent.

(testing)	TARGET_5YRS	
GLM.PRED	0	1
0	120	101
1	86	283

Figure 3 -Testing Split Classification Rates

```
Generalized Linear Model

1340 samples
   8 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1072, 1073, 1072, 1071, 1072
Resampling results:

  Accuracy   Kappa
  0.7007447  0.345133
```

Figure 4 – 5-Fold Cross-Validation Results

The pseudo R^2 showed that 25.1% of variance in the target could be accounted for by the predictors, which doesn't seem to be a very strong value. Additionally, with a $chi^2(8)=273$, $p<0.05$, the model overall is significant. The significance of predictors can be seen in the below table. Using an alpha level of 0.05, steals, blocks, and turnovers were the only variables that were not significant of the predictors chosen. The best predictor of one's longevity in the NBA seems to be the number of games played, with a z-score of 8.534. Points, having a z-score of 3.710, is the second best predictor of one's longevity in the NBA. These values make sense because games played and points are the most important statistics that sports analysts use to determine how good a player is. Minutes and turnovers were the only variables with negative coefficient values. For minutes, this suggests as the number of minutes increases by one, the log odds of lasting 5 years in the NBA decreases by 0.08, which doesn't make much sense – playing more minutes probably means that an athlete is a good player, and thus is more likely to last longer than 5 years in the NBA. Additionally, as the number of turnovers increases by 1 unit, the log odds of lasting 5 years in the NBA decreases by 0.43. This does make sense, because if a player commits more turnovers, then it will reflect poor performance.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.28153    0.23743  -9.609  < 2e-16 ***
GP           0.03840    0.00450   8.534  < 2e-16 ***
MIN         -0.08139    0.02841  -2.864 0.004178 **
PTS          0.18181    0.04900   3.710 0.000207 ***
REB          0.23398    0.07666   3.052 0.002271 **
AST          0.26427    0.10850   2.436 0.014870 *
STL          0.04536    0.30322   0.150 0.881085
BLK          0.46621    0.26502   1.759 0.078559 .
TOV         -0.43490    0.24706  -1.760 0.078360 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1779.5  on 1339  degrees of freedom
Residual deviance: 1506.7  on 1331  degrees of freedom
AIC: 1524.7
```

Figure 5 - Model Results

Finally, looking at the ROC, the model does seem to do a fair job at assessing true positive rates and false negative rates. The AUC of 0.76 suggests that this model does a somewhat mediocre job at predicting a player's longevity.

Overall, I am satisfied with my results. However, one of the ways to improve the model would be to find more observations, even though it is very time-consuming and expensive to go thorugh and individually search for each player's statistics. Additionally, this model may be improved by reducing the number of variables used, thus reducing overfitting. Perhaps we could use some variable selection methods such as subset selection and shrinkage. I also think it would be interesting to see how the data would change if you changed the output to a ten-year or a three-year longevity. That being said, this model would be extremely beneficial to basketball organizations, players, statistics teachers, and fans of the sport by finding a player's probability of longevity in the NBA given his statistics.

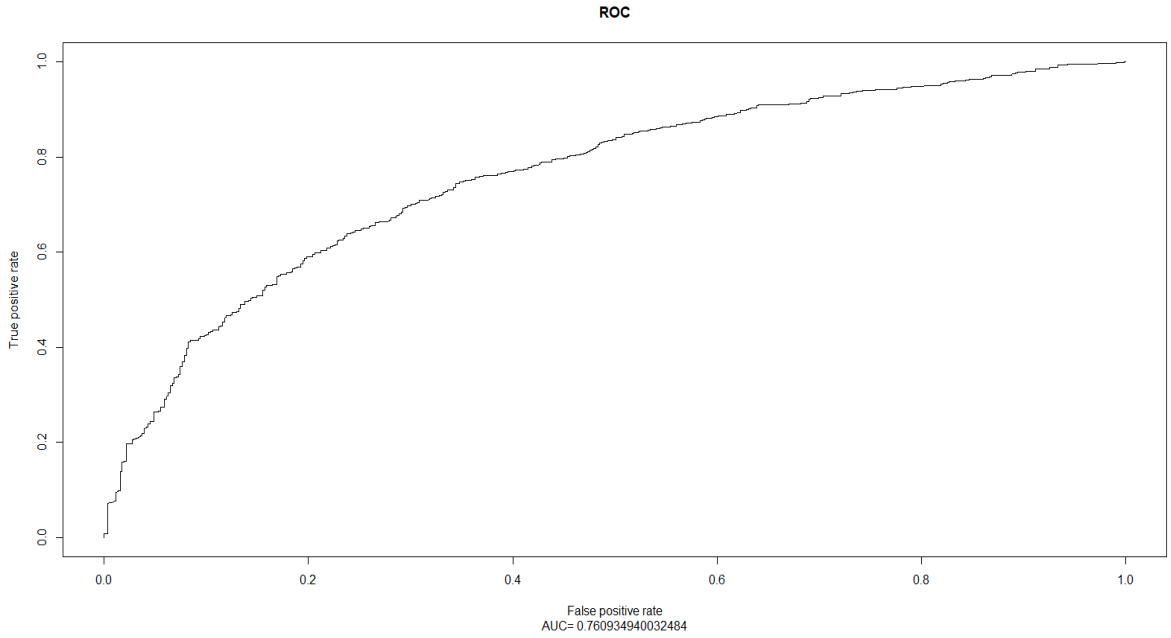


Figure 5 – ROC with AUC

5. References

[1] "National Basketball Association." *Wikipedia*, Wikimedia Foundation, 28 Nov. 2018, en.wikipedia.org/wiki/National_Basketball_Association.

[2] "Binary Classification Exercise Dataset - Dataset by Exercises." *Data.world*, 19 Sept. 2017, data.world/exercises/logistic-regression-exercise-1/discuss/logistic-regression-exercise-1/mezdgnj