Logistic Regression 5-Year Career Longevity for NBA Rookies

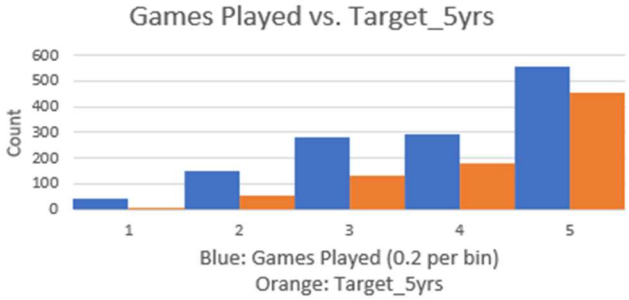
Introduction of the Problem

Every year about 38% of NBA players drop out before completing 5 years of their career. In order to find out the variables that affect the 5-year career longevity of NBA rookies, we are using logistic regression to identify their correlation among each other. In this report, we are going to employ logistic regression analysis to explore the variables that may predict if a player is going to stay in the league for more than 5 years and discover the relationship of the influential factors with the dependent variable. The study is based on a historical dataset derived from career stats of NBA players data, including several factors such as Games played, Minutes played, Field goals made, Blocks and various other details related to a player.

Exploratory Data Analysis

To see whether a player has 5-year career longevity or not we used binary values: 1 for yes, 0 for no. The dataset is quite balanced with a 3:2 ratio for the binary values. In this dataset, there are no categorical variables; all of them are numeric type data – except "Games Played" are all continuous type numeric data. To identify redundant variables, we prepared a correlation table as follows. After removing highly correlated variables we obtained the following variables with percentage.

Row Labels	Games Played	Minutes	Points/ Game	FieldGoalsMade	Field Goals %	3Point Made	3Point %	FreeThrow Made	Free Throw %	Rebounds	Assists	Steals	Blocks
0	31%	29%	27%	27%	34%	35%	38%	27%	37%	27%	30%	30%	26%
1	69%	71%	73%	73%	66%	65%	62%	73%	63%	73%	70%	70%	74%



On the left is the bar chart of Games played over the 5-year period and this shows how Target increases as the number of Games played increases.

Methodology

The data is cleaned and normalized by using Python. And then after removing the redundant variables, we took a small sample (10 samples) out for testing purposes and we trained the logistic regression with most of the data. We used R, Python, and Statstool for logistic regression. The regression was performed in a stepwise manner; the insignificant factors were removed in a backward manner. Finally, we get a logistic regression derived equation that describes the relationship very well based on the classification matrix as well as predicted data set.

Analysis, Outputs and Interpretation

Using the methodology described above; after a series of stepwise elimination of insignificant factors, we obtained the following coefficients with corresponding p values. At 5% significance, we chose to keep

some of the statistically insignificant ones too as they are not wildly insignificant. The confusion matrix looks good especially when it comes to identifying the players with higher odds of career longevity - which was the main objective of this analysis.

Coefficient	Value	p-Value
Constant	-3.2665467	0.000
Games Played	2.59193269	0.000
Minutes	-2.1178087	0.038
Field Goals Made	2.10674524	0.094
Field Goals %	1.32113538	0.038
Free Throws Made	2.21737734	0.079
Free Throws %	1.2813541	0.065
Rebounds	2.48409901	0.026
Assists	2.53874188	0.022
Blocks	1.76525445	0.092
Turnovers	-1.688776	0.138

The logistic regression is as follows:
Log [p/ (1 - p)] = - 3.27 + 2. 59*Games Played - 2.12*Minutes + 2.11*Field Goals Made + 1.32*Field Goal % + 2.22*Free Throws Made + 1.28*Free Throws % + 2.48*Rebounds + 2.54*Assists + 1.77* Blocks - 1.69*Turnovers.

The likelihood of 5-year career longevity is positively related to all the variables except the number of minutes and turnovers made. The odds of 5-year career longevity increase a lot with an increase in the number of games (2.59 times) the number of assists (2.53 times) and rebounds (2.48 times). Similarly, the odds significantly decrease with an increase in the total number of minutes played (-2.11 times) and turnovers (-1.69 times) made by the players.

Classification Matrix			Percent
	1	0	Correct
1	684	141	83%
0	237	266	53%

Conclusion

From our observation, we can conclude that a player's number of games played is statistically significant to his 5-year career longevity. However, we were surprised that a player's "field goal percentage" is a significant factor, whereas his "field goal made" is not. This might be due to the fact that "field goal percentage" is influenced by a player's attempt to score and it can be represented by the following equation: "field goal percentage = field goal made ÷ field goal attempt". This shows that a player's "field goal percentage" might be diluted by his increasing attempts to score while maintaining his actual field goal made. As for the techniques that a player acquired, such as rebounds, assists, steals, and blocks, we discovered an interesting fact that the attacking methods – rebounds and assists – are significant factors for a player's 5-year career longevity while defending methods – steals and blocks – show less significance. From our logistic regression equation, we could predict the probability that any NBA player would sustain his 5-year career longevity, given values for all of the independent variables.