MSCI 623
Big Data and Analytics

Project Proposal

Using Supervised Machine Learning Model Predict 5 Year Career Longevity for NBA Rookies

Proposal Submitted By:

Prikshit Batta (20980469)

Problem Statement

The goal of this exercise is to use the player performance statistics to predict if the career of a NBA player is more than 5 years or not. In the given dataset (described in detail below), there are a number of variables capturing the performance statistics of the NBA players. We will be using this information to generate a binary decision on whether the player's career will be more than 5 years.

Dataset

The link of the dataset is - https://data.world/exercises/logistic-regression-exercise-1

A snapshot of a few sample rows of the dataset is added below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Name	GP	MIN	PTS	FGM	FGA	FG%	3P Made	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	TARGET_5Yrs
2	Brandon Ingram	36	27.4	7.4	2.6	7.6	34.7	0.5	2.1	25	1.6	2.3	69.9	0.7	3.4	4.1	1.9	0.4	0.4	1.3	0
3	Andrew Harrison	35	26.9	7.2	2	6.7	29.6	0.7	2.8	23.5	2.6	3.4	76.5	0.5	2	2.4	3.7	1.1	0.5	1.6	0
4	JaKarr Sampson	74	15.3	5.2	2	4.7	42.2	0.4	1.7	24.4	0.9	1.3	67	0.5	1.7	2.2	1	0.5	0.3	1	0
5	Malik Sealy	58	11.6	5.7	2.3	5.5	42.6	0.1	0.5	22.6	0.9	1.3	68.9	1	0.9	1.9	0.8	0.6	0.1	1	1

The following table provides a description of all the columns of the dataset.

	Description
Name	Name
GP	Games Played
MIN	MinutesPlayed
PTS	PointsPerGame
FGM	FieldGoalsMade
FGA	FieldGoalAttempts
FG%	FieldGoalPercent
3P Made	3PointMade
3PA	3PointAttempts
3P%	3PointAttempts
FTM	FreeThrowMade
FTA	FreeThrowAttempts
FT%	FreeThrowPercent
OREB	OffensiveRebounds

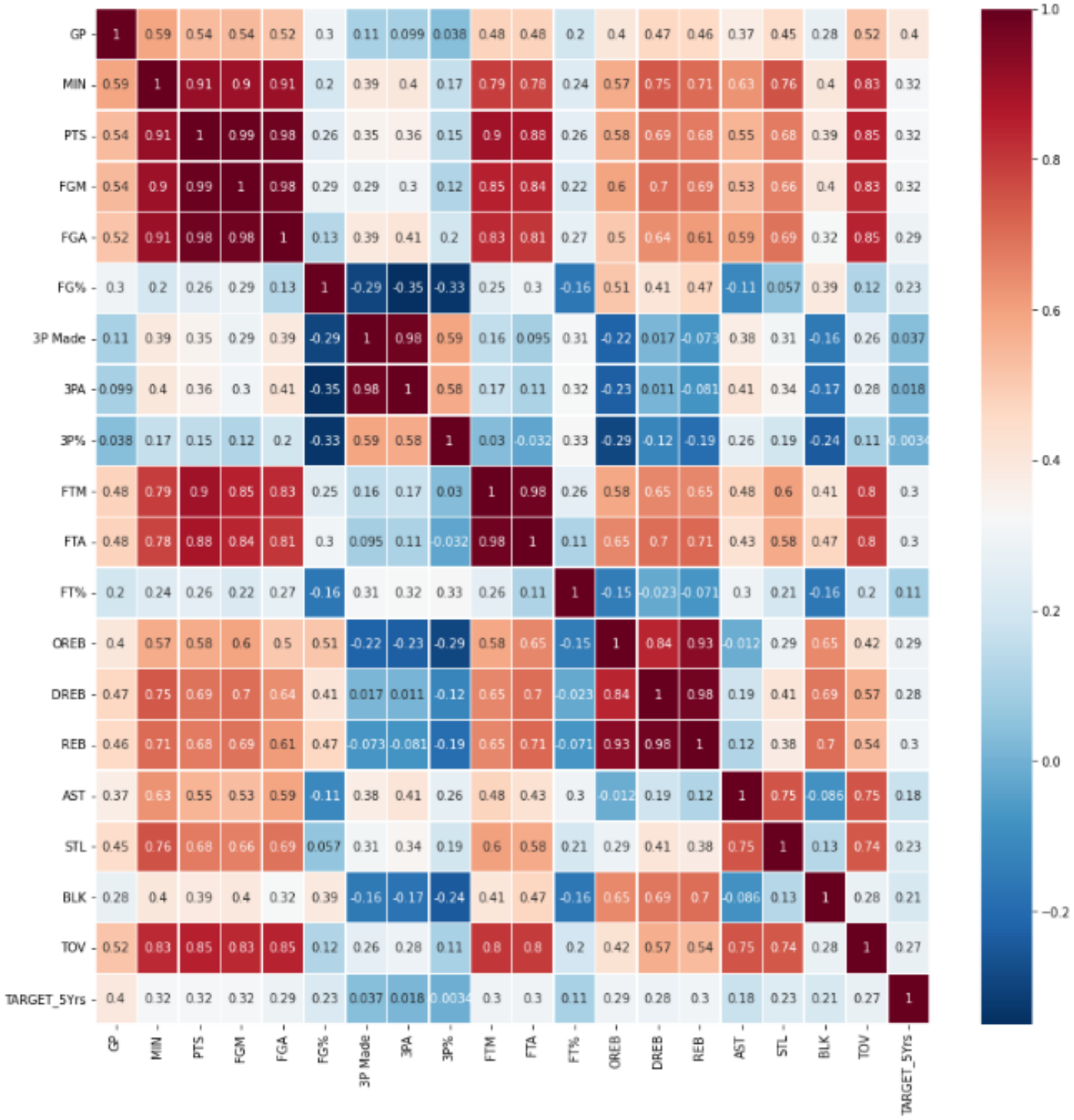
	Description
DREB	DefensiveRebounds
REB	Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
TARGET_5Yrs	Outcome: 1 if career length >= 5 yrs, 0 if < 5...

Observations from the data:

1. There are 1,341 rows and 21 columns in this dataset.
2. The datatypes of all the independent variables (features) is continuous.
3. The dependent variable, *TARGET_5yrs* is a categorical variable which takes 0 or 1 as values.
4. The player identifier variable, *Name* is a string. We will not be considering it as a feature for our model as the name of a player should not influence the duration of their career.
5. For the binary classification problem, the dataset is a balanced with around 62 % of the rows with the value of target variable (TARGET_5ys) as 1 and the remaining as 0.

Variables

The dependent variable (that will be predicted) in this project is Target_5Yrs. It is a categorical variable which takes a 1 or 0 value depending on whether a player has a career of 5 years or not. The following heatmap shows the Pearson's correlation of all the independent variables amongst themselves and with the dependent variable :



Key Observations: We used the above heatmap to identify good explanatory variables and have the following observations:

- **GP** (Games played) has 0.4 co-relation with **Target_5Yrs** variable which is very high and good corelation and intuitively it explains that if a player has a greater number of games played than that player can play more than 5 years and vice versa. Thus it makes GP(Games played) a very good factor for predicting the result.
- From the heatmap we can see that variables MIN, PTS,FGM,FGA,TOV has very similar corelation with variable **Target_5Yrs**. Moreover all of these variables have

very high corelation among all of them. So, there is no point of taking all these variables for consideration. We can only take one variable from MIN, PTS,FGM,FGA,TOV so that the outcome can be predicted easily.

- From the variables 3PA (3 Point Attempts) and 3P%(3 Point Attempts percentage) these both the variables are also highly correlated with each other and these variables has very less corelation with the variable **Target_5Yrs**. This intuitively makes sense as having less 3 pointer attempts shouldn't mean a player cannot be playing for more than 5 years.

Why do we need a Machine Learning Model to solve this problem?: The need of the Machine Learning algorithm here is because the relationship between the predictor variables and target variable is not very evident. None of the independent variables has a very good linear correlation with the target variable. So, we cannot use tools like SQL or excel to filter the players using simple rules. A few independent variables have a Pearson's correlation > 0.3 with the target variable which means that a machine learning model is best suited for solving this problem.