NBAPhase3

March 19, 2020

Viraj Dhillon 862015754

For this phase of the project, I want to explore some analysis based on players stats. The first is that I want to be able to see what player statisitcs will increase a players minutes per game. I believe that a players playing time is mainly tied to the amount of points they average, and rebounds being a lower indicator of their playtime, and turnovers being the lowest indicator. We can use linear regression to see what the coefficients will be, and can see whether they are positive or negative.

Here we will be using the dataset that we've been using for the past two phases, "NBA1950-2019.csv". We will only be using players from 1980 and beyond because this is when the modern NBA started and statistics were more closely recorded.

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
statsDF = pd.read_csv("NBA1950-2019.csv")
statsDF = statsDF.drop(columns = ["Unnamed: 0", "Unnamed: 0.1"])
statsDF = statsDF[(statsDF["Season"] > 1981)]
statsDF = statsDF.dropna(subset=['Player'])
statsDF = statsDF.fillna(0)
statsDF.head()
```

[1]:

	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	\
287	Alaa Abdelnaby	PF	26.0	TOT	54.0	0.0	9.4	2.2	4.3	0.511	
288	Alaa Abdelnaby	PF	26.0	SAC	51.0	0.0	9.3	2.3	4.3	0.532	
289	Alaa Abdelnaby	PF	26.0	PHI	3.0	0.0	10.0	0.3	3.7	0.091	
290	Mahmoud Abdul-Rauf	PG	25.0	DEN	73.0	43.0	28.5	6.5	13.8	0.470	
291	Michael Adams	PG	32.0	CHH	29.0	0.0	15.3	2.3	5.1	0.453	

	…	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Season
287	…	0.7	1.4	2.1	0.2	0.3	0.2	0.8	1.9	4.7	1995
288	…	0.7	1.4	2.1	0.3	0.3	0.2	0.8	2.0	5.0	1995
289	…	1.0	1.7	2.7	0.0	0.0	0.0	1.7	0.7	0.7	1995
290	…	0.4	1.4	1.9	3.6	1.1	0.1	1.6	1.7	16.0	1995
291	…	0.2	0.8	1.0	3.3	0.8	0.0	0.9	1.4	6.5	1995

[5 rows x 30 columns]

Here we will be splitting up our data into testing and training. The predicted value we want to find is the average career amount of minutes a player plays in a game. The features we will be using is a players career average points, rebounds, assists, etc.

[2]:
```
y = statsDF.groupby('Player')['MP'].mean()
x = {'PTS': statsDF.groupby(["Player"])["PTS"].mean(),
     'STL': statsDF.groupby(["Player"])["STL"].mean(),
     'BLK':  statsDF.groupby(["Player"])["BLK"].mean(),
     'TRB': statsDF.groupby(["Player"])["TRB"].mean(),
     'TOV': statsDF.groupby(["Player"])["TOV"].mean(),
     'AST': statsDF.groupby(["Player"])["AST"].mean(),
     'GS': statsDF.groupby(['Player'])['GS'].mean().round()
    }
x = pd.DataFrame(data = x)
xTrain, xTest, yTrain, yTest = train_test_split(x, y, test_size = 0.3)
x.head()
```

[2]:

Player	PTS	STL	BLK	TRB	TOV	AST	GS
A.C. Green	9.233333	0.805556	0.394444	7.333333	1.077778	1.050000	50.0
A.J. Bramlett	1.000000	0.100000	0.000000	2.800000	0.400000	0.000000	0.0
A.J. English	9.850000	0.400000	0.150000	2.100000	1.350000	2.150000	9.0
A.J. Guyton	3.800000	0.333333	0.133333	0.700000	0.666667	1.566667	5.0
A.J. Hammons	2.200000	0.000000	0.600000	1.600000	0.500000	0.200000	0.0

Now let's start analyzing NBA stats and correlation to minutes played.

[3]:
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()

xTrain = xTrain[["PTS", "AST", "TRB", "STL", "BLK", "TOV", "GS"]]
xTest = xTest[["PTS", "AST", "TRB", "STL", "BLK", "TOV", "GS"]]
model.fit(X = xTrain, y = yTrain)
yPredict = model.predict(X = xTest)
model.coef_
```

[3]:
```
array([ 0.81665558,  0.92871676,  1.0039522 ,  3.86939867, -0.18441203,
       -0.50796908,  0.07366328])
```

To my surprise, my hypothesis wasn't fully correct. I stated that a players career points average

will be the highest coefficent, and rebounds will be a lower coefficient, and turnovers will be the lowest. But in reality, it turns out that steals is the highest coefficient, and then followed by total rebounds. This makes sense since steals give teams momentum and coaches won't take players out after they commit steals. The points category is actually the third highest coefficient. Not surprisingly, turnovers is the lowest coefficient as players who have higher turnovers will get less playing time.

First: Steals

Second: Total Rebounds

Third: Assists

[ ]: