SUMMER PROGRAMMING RESEARCH PAPER – NBA FREE THROW PREDICTION

BY – KUNAL SACHDEV, R N PODAR SCHOOL, MUMBAI

MENTOR – Mr. IAN SHEN (MA in Statistics)

Table of Contents

1. INTRODUCTION

THE NBA

The National Basketball Association (NBA) is a professional basketball league in North America. The league is composed of 30 teams (29 in the United States and 1 in Canada) and is one of the major professional sports leagues in the United States and Canada. It is the premier men's professional basketball league in the world.

The league was founded in New York City on June 6, 1946, as the Basketball Association of America (BAA). It changed its name to the National Basketball Association on August 3, 1949, after merging with the competing National Basketball League (NBL). In 1976, the NBA and the American Basketball Association (ABA) merged, adding four franchises to the NBA. The NBA's regular season runs from October to April, with each team playing 82 games. The league's playoff tournament extends into June. As of 2020, NBA players are the world's best paid athletes by average annual salary per player.

NBA TEAMS:

2. OBJECTIVE

The overall goal is to **predict the free throw shooting accuracy** of NBA players. The scope is to **analyse the data** collected using various tools of the **Python** programming language, identify the **features that most affect the free throw shooting accuracy** of an NBA player, come up with a **regression model** to

predict the free throw shooting accuracy of NBA players, and **interpret** the coefficients of the generated model's predictor variables.

3. DATA STRUCTURE AND SOURCE

3.1 Data Set Information

- The data set contains information about various attributes of a player, such as field gaols attempted, assists, turnovers, steals, etc., per game for the regular NBA season
- Two data sets obtained are:
  - Players General Traditional 2018-19
  - Players General Traditional 2021-22
- Both datasets include 29 features describing the name of the player, and the player's other attributes throughout the season
- The features are presented in per game mode, i.e., feature per game – assists per game, turnovers per game, etc.
- The 2018-19 dataset contains 530 records whereas the 2021-22 dataset contains 605 records.
- The data structure for both the datasets is the same and is included in the **6.** .

4. EXPLORATORY DATA ANALYSIS

- The data set for the 2018-19 NBA season contains a total of **530 records** and the dataset for the 2021-22 NBA season contains a total of **605 records**.
- Both the datasets contain a total of 29 features that describe the performance of the player throughout the season.
- The numbers for these features are in per game mode, i.e., feature per game – field goals per game, rebounds per game, etc. The features also include the name of the player, the player's team, the number of games played, games won and games lost.
- The 2021-22 NBA season dataset is used for data analysis and model development as it is the latest non-covid NBA season.
- The 2018-19 NBA season dataset is for evaluating the model's ability to accurately predict the free throw accuracy of NBA players.

4.1 Check for Null Values

The data set contains zero null values.

4.2 Removing Outliers/Low Information Points

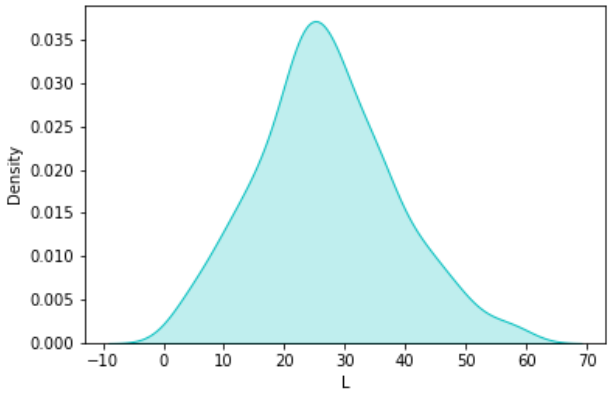
- Goal is to predict **the free throw accuracy** of NBA players. If a player has only attempted one free throw in the entire season and made that free throw, then his free throw percentage is shown as 100%. This can bias our results and lead to poor results from the model. Therefore, such values need to be eliminated.
- These values are not outliers per se; they are **low information points**. Some of these points might be outliers and some of them aren't, but we are choosing to exclude them because they can lead to faulty results.
- **Criteria set for qualifying as an outlier**: If the player has attempted fewer than 10 free throws in the regular NBA season, then the record for that player will be eliminated.
- **Dataset before removing outliers**: 605 rows, 29 columns

- **Dataset after removing outliers**: 460 rows, 29 columns
- Therefore, **145** players attempted less than 10 free throws in the NBA regular season.

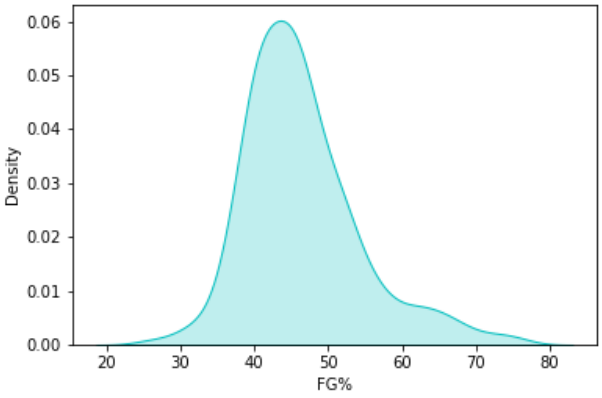
4.3 Check Normal Distribution of the Features Using KDE Plots

- A **kernel density estimate (KDE)** plot is a method for visualizing the distribution of observations in a dataset, similar to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions.
- One of the assumptions associated with a linear regression model is **normality**, i.e., for any fixed value of x, the value of y is normally distributed. Let's check if the distribution of the features in our data set are normally distributed.

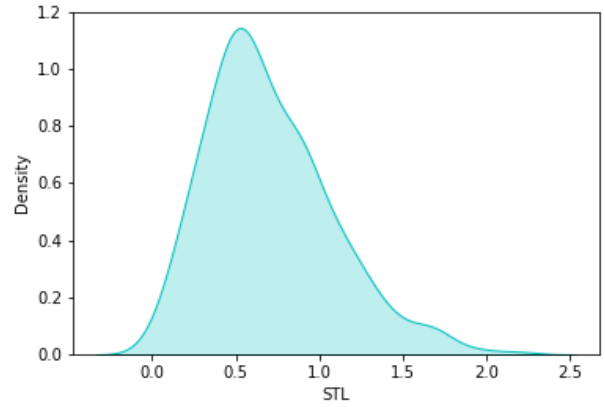
1. **KDE Plot for the feature "Losses" – Plot has a normal distribution curve**



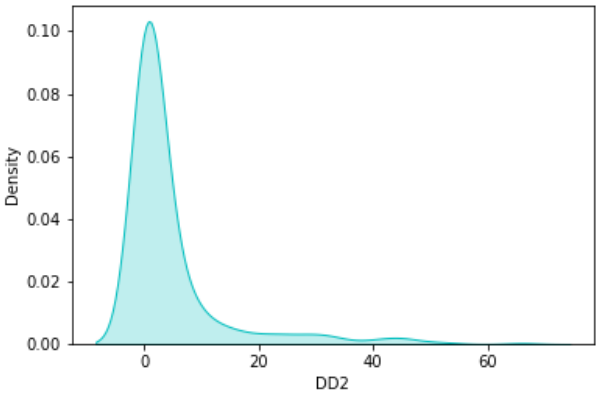
2. **KDE Plot for feature "FG%" (Field Goal Percentage) – Plot has a normal distribution curve**



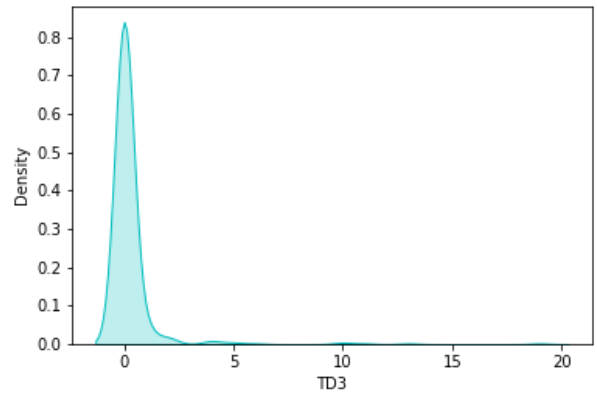
3. **KDE Plot for feature "Steals" – Plot has a normal distribution curve**



1. **KDE Plot for feature "DD2" (Double doubles) – Highly Right Skewed**



2. **KDE Plot for feature "TD3" (Triple doubles) – Highly Right Skewed**



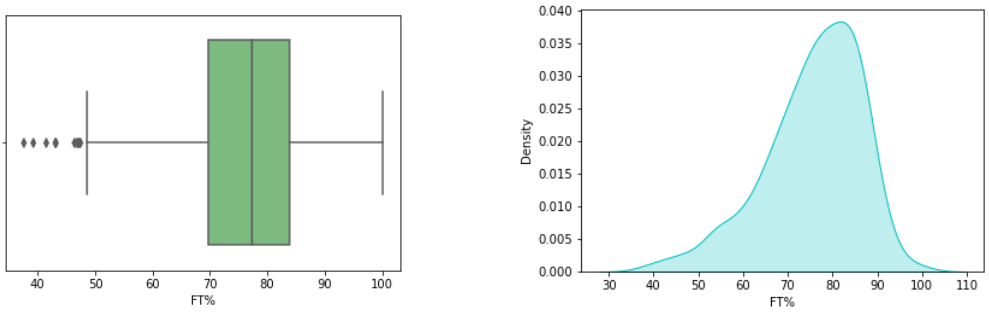
- **DD2 and TD3 were the most skewed by far of all the features.**

4.4 Dropping Unwanted Features

1. The features DD2 and TD3 do not show normal KDE plots.
2. The feature "FP", short for fantasy points, is directly derived from the features three-point field goals, two-point field goals, free throws made, rebounds, assists, blocked shots, steals, and turnovers.

3. The features DD2, TD3 and FP will therefore be excluded from our data set and from further analysis.
    - **Dataset before removing unwanted features**: 460 rows, 29 columns
    - **Dataset after removing the unwanted features:** 460 rows, 26 columns

4.5 Visualizing the Distribution of Free Throw Percentage

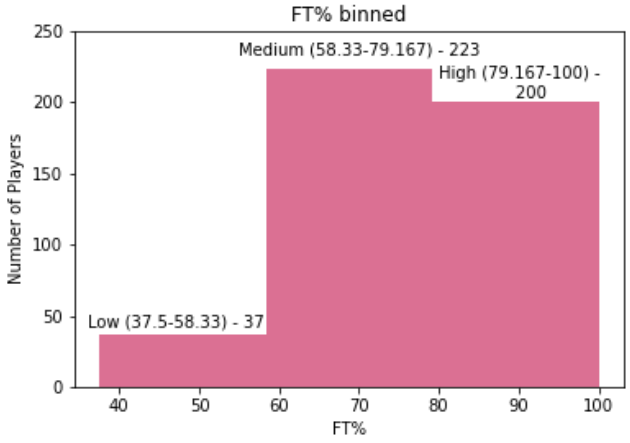


- The **median** free throw percentage seems to be around 76-77%.
- The **maximum** free throw percentage is a 100% and the **minimum** is somewhere around 47-48%.
- Except for some outliers are the lower side, the overall distribution of FT% is approximately normal.

4.6 Categorizing FT% into Equally Spaced Ranges and Visualizing the Same

- 4 equally spaced values were generated between the lowest (37.5) and highest (100) free throw percentages – 37.5, 58.33, 79.167, 100.
- It was found that the highest number of players had a free throw percentage in the middle range (58.33 – 79.167). We see that almost equal number of players have medium and high free throw percentages (223 and 200 respectively) but a very players have low free throw percentage.

Category	Range	Number of Players
Low	37.5 – 58.33	37
Medium	58.33 – 79.167	223
High	79.167 - 100	200



4.7 Correlation Coefficient, P-values, & Hypothesis Test

- A **correlation coefficient** is numerical value ranging from **-1 to 1**, with -1 and 1 being the strongest correlation and 0 being the weakest correlation. The correlation coefficient value signifies the strength of the relation between two variables.
- The **p-value** is a statistical measure to conclude if there exists a linear relationship between two variables. We test if the value of the coefficient is equal to zero (no relationship). The statistical test used is called **hypothesis testing**:
- **Null Hypothesis**: There exists no relationship between <feature> and FT%
- **Alternative Hypothesis**: There exists some relationship between <feature> and FT%
- This test results in a p-value. The **common threshold** for p-value is 0.05. If the P-value is lower than 0.05, we can reject the null hypothesis and conclude that there exists a relationship between the <feature> and FT%. A p-value of 0.04 would mean that 4% of the times we will falsely reject the null hypothesis, meaning 4% of the times we would have falsely concluded that there exists a relationship between the two variables being tested.

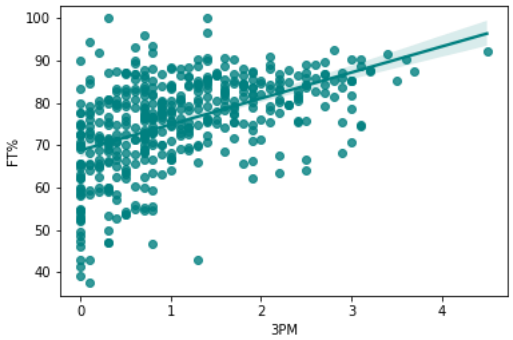
Feature	Correlation with FT%	Associated p-value
FG%	-0.3492	1.218e-14
OREB	-0.3488	1.328e-14
BLK	-0.1620	0.00048
REB	-0.1121	0.01618
DREB	0.0021	0.96333
PF	0.0442	0.34446
AGE	0.1029	0.02730
L	0.1356	0.00357
FTA	0.1656	0.00036
STL	0.1848	6.664e-05
W	0.1904	3.953e-05
+/-	0.1976	1.974e-05
TOV	0.2179	2.391e-06
GP	0.2182	2.321e-06
FGM	0.2825	6.908e-10
FTM	0.2872	3.469e-10
AST	0.2889	2.693e-10
MIN	0.3297	3.967e-13
PTS	0.3363	1.272e-13
3P%	0.3584	2.179e-15
FGA	0.3585	2.124e-15
3PM	0.4913	2.529e-29
3PA	0.4948	8.781e-30

- All the above features, except DREB & PF, give a p-value less than 0.05. Therefore, for all the features, except DREB & PF, we can conclude that there exists a relationship between that feature and FT%.

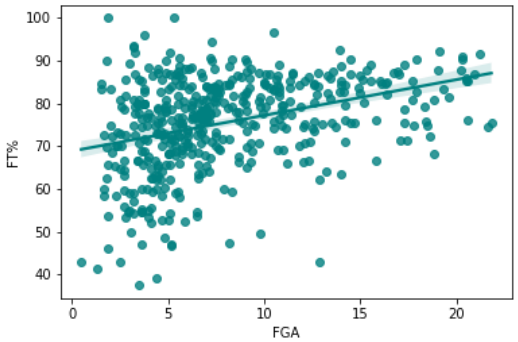
- We see that although some features' correlation coefficients with FT% come close to 0.5, such as 3PM & 3PA, most of the variables do not give a high correlation with FT%.

The scatterplot and line of best fit for some of the features are shown below:

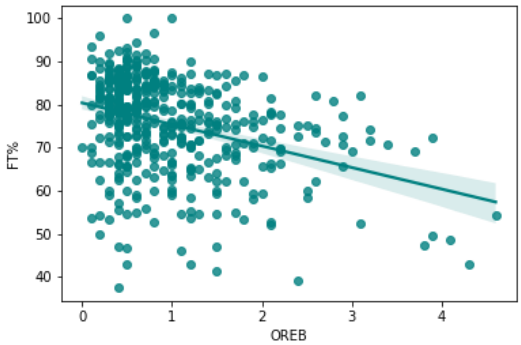
1. **Free Throw Percentage vs 3 Points Made**



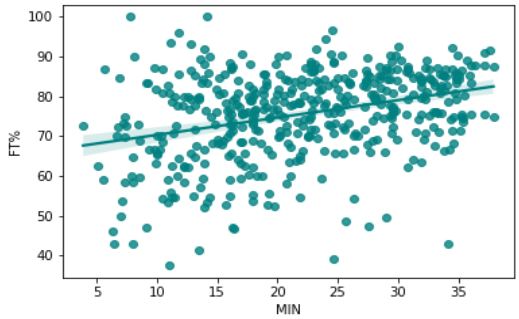
2. **Free Throw Percentage vs Field Goals Attempted**



3. **Free Throw Percentage vs Offensive Rebounds**



5. **Free Throw Percentage vs Minutes Played**



4.8 The ANOVA & Hypothesis Test

- An Analysis of Variance (ANOVA) test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.
- The ANOVA test returns a F-value, which tells us whether there exists a significant difference between the means of any groups you are testing.
- In our case, we will use the ANOVA test to check whether being in a certain team is associated with players' free throw accuracy OR check whether there is a significant difference in the free throw accuracy means of all the 30 NBA teams.
- The higher the F-value in an ANOVA, the higher the variation between sample means relative to the variation within the samples. The higher the F-value, the lower the corresponding p-value. If the p-value is below a certain threshold (e.g. $\alpha = .05$), we can reject the null hypothesis of the ANOVA and conclude that there is a statistically significant difference between group means.

- **Null Hypothesis**: There is no difference in means of the free throw accuracy of the 30 NBA teams
- **Alternative Hypothesis**: There is a difference in means of the free throw accuracy of the 30 NBA teams.

- **ANOVA Results: f-value = 0.9494, p-value = 0.5438**

- Since the p-value is greater than 0.05 and the f-value is very small, we fail to reject the null h ypothesis and conclude that we do not have sufficient evidence to say that there is a statistic ally significant difference between the means of free throw accuracy of the thirty groups.

4.9 EDA Summary

1. The data set has no null values
2. A total of 145 players out of 605 players attempted less than 10 free throws in the 2021-22 NBA Season.
3. Except for some outliers on the lower side, the overall distribution of FT% is approximately normal with a median at around 76-77%.

4. The highest number of players had a free throw percentage in the middle range (58.33 – 79.167) & almost an equal number of players have medium and high free throw percentages (223 and 200 respectively) but very few players have low free throw percentage (37).
5. For all the features, except DREB & PF, we can conclude that there exists a relationship between that feature and FT%.
6. Although some features' correlation coefficients with FT% come close to 0.5, such as 3PM & 3PA, most of the variables do not give a high correlation with FT%.
7. There does not exist a statistically significant difference between the mean free throw accuracy of the 30 NBA teams.

5. MODEL DEVELOPMENT

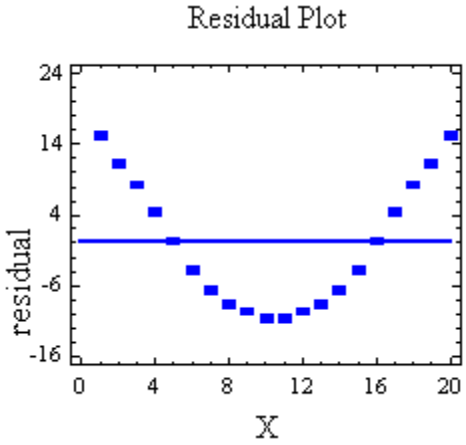
5.1 Checking Randomness in Residual Plot

- A residual plot shows the difference between the observed response and the fitted response values.

$$Residual\ (\epsilon) = y - \hat{y}$$

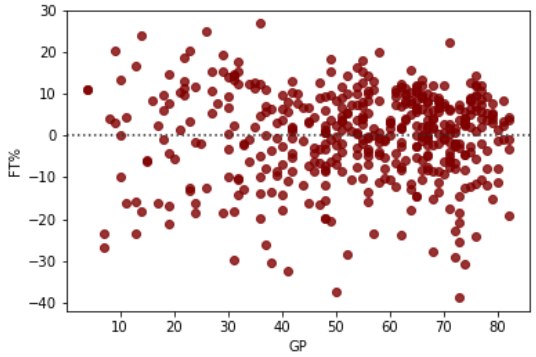
(Where y-hat is the predicted y-values)

- Since we will be developing a linear model, it is firstly necessary to ensure that a linear model is appropriate for our variables.
- This is can done by observing the residual plots of our features with our target variable (the variable that we are trying to predict) FT%. If there is no apparent pattern in the residual plot, then we can say that a linear model is appropriate for that variable.
- Residual pattern that does not have random arrangement of values:

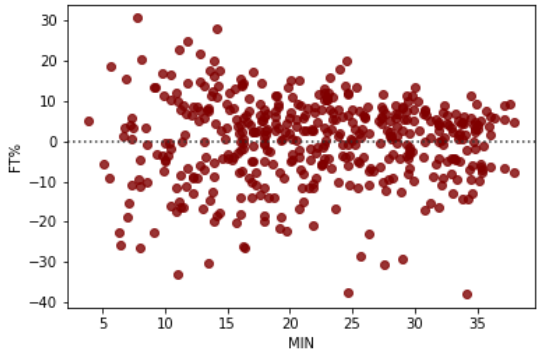


- All of our features show a random pattern in their residual plots with free throw percentage. **Therefore, we can say that a linear model is appropriate for our features with free throw percentage.**
- The residual plots of some of the features with free throw percentage are shown below:

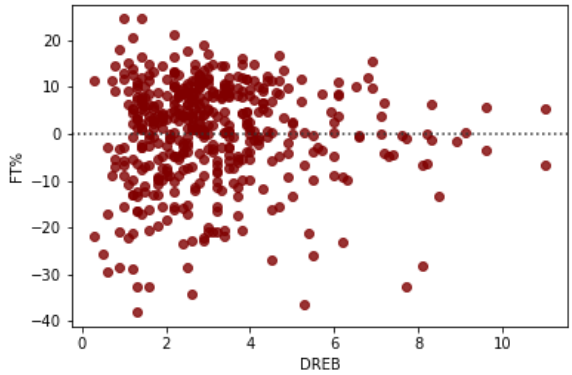
2. GP vs FT%



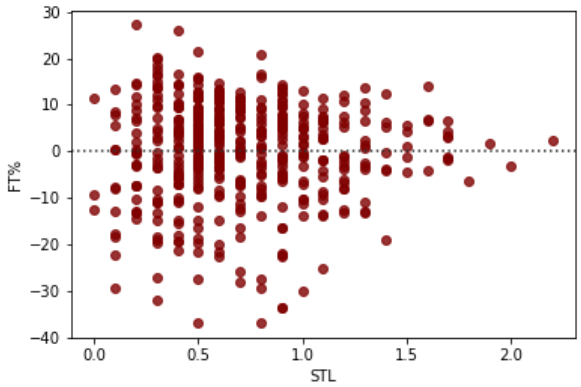
3. MIN vs FT%



4. DREB vs FT%



6. **STL vs FT%**



5.2 Lowest AIC Model: Multiple Linear Regression

5.2.1 AIC Analysis

- The **Akaike Information Criteria (AIC)** is a mathematical method for evaluating how well a model fits the data it was generated from. AIC is used to compare different possible models and determine which one is the best fit for the data.

$$\mathrm{AIC} = 2\mathrm{K} - 2\ln(\mathrm{L})$$

- **K** is the number of independent variables used and **L** is the log-likelihood estimate (a.k.a. the likelihood that the model could have produced your observed y-values). The default K is always 2, so if your model uses one independent variable your K will be 3, if it uses two independent variables your K will be 4, and so on.

- **R-Squared** (R^2 or the **coefficient of determination**) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model

- Since most of the variables did yield a high correlation with free throw percentage, I made a model using all of the variables. Then, following a descending order of correlations with FT%, I removed one variable at a time and calculated the AIC for each model. I then selected the model with the lowest AIC.

Feature	Correlation with FT%	AIC	R^2	Remarks
FG%	-0.349234	3325.4099	38.30%	AIC and R^2 considering all the variables. Before removing any of the variables, we get the lowest AIC and the greatest r^2 as well
OREB	-0.348768	3335.4076	36.67%	AIC and R^2 excluding FG%
BLK	-0.161951	3333.4528	36.66%	AIC and R^2 excluding OREB
REB	-0.112077	3331.4607	36.66%	And so on…
DREB	0.002149	3336.7668	35.65%	
PF	0.044177	3335.1354	35.60%	
AGE	0.102916	3333.5916	35.53%	
L	0.135587	3332.1374	35.46%	
STL	0.184837	3332.1374	35.46%	
W	0.1904	3330.1681	35.45%	
TOV	0.217883	3329.5504	35.26%	
GP	0.218158	3337.9314	33.78%	
FP	0.226785	3337.6590	33.53%	
FGM	0.282483	3354.5051	30.75%	
AST	0.288942	3370.8159	27.94%	
MIN	0.329733	3371.6989	27.49%	
PTS	0.33625	3369.6993	27.49%	
3P%	0.358391	3372.9780	26.65%	
FGA	0.358525	3383.8408	24.57%	
3PM	0.491301	3382.0427	24.54%	
3PA	0.494823	3380.3561	24.48%	

The model that contains all the features has the lowest AIC (3325.4099) and the greatest R^2 (38.30%) value as well. Therefore, let's construct a model with all the features and explore the model.

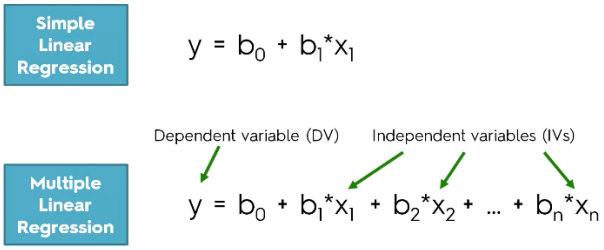
5.2.2 Parameters

The parameters that we will be using for our multiple linear regression model are:

FG%, OREB, BLK, REB, DREB, PF, AGE, L, STL, W, +/-, TOV, GP, FGM, AST, MIN, PTS, 3P%, FGA, 3PM, 3PA

5.2.3 Multiple Linear Regression Model

- A linear regression model is a model where the relationship between inputs and outputs is a straight line, i.e., a linear relationship.
- A multiple linear regression model is similar to a linear regression model but in this one there are more than one independent variables.



Dependent Variable – Free Throw Percentage (FT%)

Equation of our multiple linear regression model:

FT% = 84.1809 – (0.3824*FG%) + (1.0484*OREB) - (0.1188*BLK) - (4.6283*REB) + (3.964*DREB) + (1.1164*PF) + (0.0509*AGE) + (0.0696*L) - (3.6948*STL) - (0.0510*W) + (0.6529*+/-) - (2.2604*TOV) + (0.0186*GP) - (2.6410*FGM) + (0.3321*AST) + (0.4476*MIN) + (2.7584*PTS) + (0.0522*3P%) - (1.8311*FGA) - (1.1810*3PM) + (0.3548*3PA)

	coef	std err	t	P>|t|
const	84.1809	6.119	13.757	0.000
FG%	-0.3824	0.109	-3.513	0.000
OREB	1.0484	8.225	0.127	0.899
BLK	-0.1188	1.612	-0.074	0.941
REB	-4.6283	8.249	-0.561	0.575
DREB	3.9640	8.276	0.479	0.632
PF	1.1164	0.967	1.154	0.249
AGE	0.0509	0.101	0.506	0.613
L	0.0696	0.039	1.792	0.074
STL	-3.6948	1.706	-2.166	0.031
W	-0.0510	0.038	-1.326	0.185
+/-	0.6529	0.213	3.071	0.002
TOV	-2.2604	1.509	-1.498	0.135
GP	0.0186	0.018	1.025	0.306
FGM	-2.6410	2.445	-1.080	0.281
AST	0.3312	0.551	0.601	0.548
MIN	0.4476	0.156	2.872	0.004
PTS	2.7584	0.625	4.414	0.000
3P%	0.0522	0.048	1.085	0.279
FGA	-1.8311	1.019	-1.797	0.073
3PM	-1.1810	3.554	-0.332	0.740
3PA	0.3548	1.512	0.235	0.815

5.2.4 Interpretation of Coefficients

- **Constant**:

The expected free throw percentage for the average NBA player in this model is 84.181%. The p-value for the intercept term is less that 0.05, which tells us that the intercept term is statistically significantly different than zero.

- **Coefficients of Predictor Variables:**
  1. On average, each additional unit increase in FG% is associated with a decrease of 0.38 in the free throw shooting percentage. The corresponding p-value is less than 0.05, meaning that average change in free throw percentage for each additional unit increase in FG% is statistically significantly different than zero. Another way to put this might be that the predictor variable FG% has a significant negative relationship with the response variable free throw percentage.
  2. Similarly, for each additional steal per game the FT% decreases by 3.7 units. The corresponding p-value of 0.031 (< 0.05) tells us that steals per game has a significant negative relationship with free throw percentage.
  3. For each additional unit increase in +/-, minutes played, and point scored we see a rise in FT% by 0.652, 0.45 & 2.76 units respectively. The associated p-values of these variables tells us that the average change in FT% for each additional unit of these variables is statistically significantly different than zero or that these variables have a significant relationship with FT%.
  4. Similarly, because the rest of the variables have a p-value > 0.05, they do not have a significant relationship with FT%.

- **Personal Interpretation:**
  1. +/- is basically the team score when the player is on the court minus the team score when the player is off the court. We see that for a unit increase in +/-, the FT% increases by 0.65. This means that as +/- increases, i.e., the team score is more when the player is on the court, the FT% also increases. A possible explanation could be that players with a higher +/- are better players overall since their team scores more when they are on the field. Therefore, they play more and get fouled often, getting more free throws.
  2. The negative correlation observed between FT% and Steals/Rebounds could be because a player who steals the ball might tend to go directly for a layup. Alternately, after collecting the ball, he might pass the ball ahead for a fast break and therefore would not get a free throw.
  3. If you have a greater number of free throws then naturally you will score more points, which is evident in the positive correlation between with PTS and FT%.
  4. The positive correlation between minutes and FT% could be because if a player plays for a longer time in the game he would have a greater chance of being fouled and would therefore have more free throw opportunities.

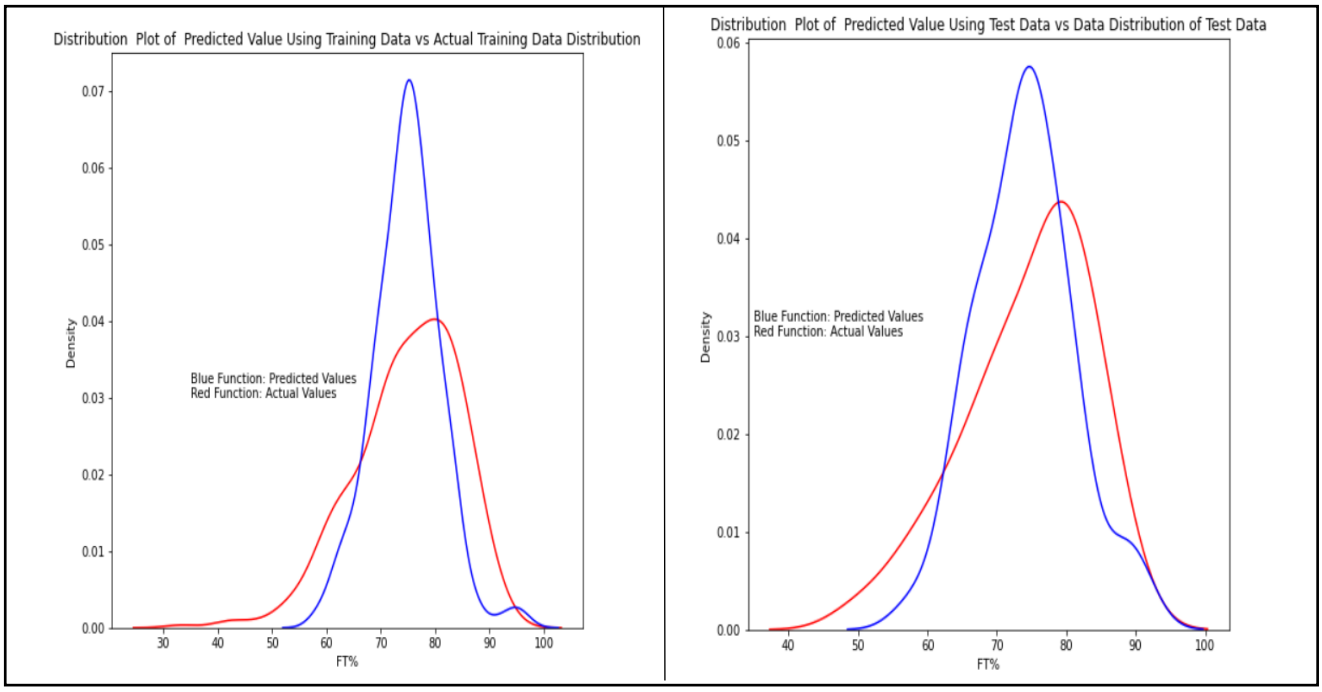
5.2.5 Train and Test

The NBA data was split randomly in a **90:10** ratio. 90% of the data was used to train the model and 10% of the data was used to test the model.

5.2.6 Predicting FT% for the 2018-19 NBA Season Using the Lowest AIC Model

- Now that we have built a model using the 2021-22 dataset, let's use our model to predict the FT% of NBA players in the 2018-19 NBA Season.

- Results: **R^2 for the training data** = 40.67 %
  R^2 for the testing data = 38.65 %



- **Left Graph** - The graph shows the Predicted **Training** FT% values in blue and the actual **Training** FT% values in red.

- **Right Graph** – The graph shows the Predicted **Testing** FT% values in blue and the Actual **Testing** FT% values in red.

- **We see that the model does a better job at tracking the Testing data than the training data.**

5.3 Model with variables having VIF<10: Multiple Linear Regression Model

5.3.1 VIF Analysis

- In the Lowest AIC Model, we observe that there is the problem of **multicollinearity**.
- Multicollinearity is the occurrence of **high intercorrelations among two or more independent variables** in a multiple regression model.
- Multicollinearity can lead to skewed or misleading results when you attempt to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.
- In our model, the parameters games played, wins, & loses are directly correlated. Also, 3P% is directly obtained from 3PA and 3PM.
- One way to **quantize this multicollinearity** is using **Variation Inflation Factor (VIF).** Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.
- In our VIF Analysis variables like GP, W and L have an infinite VIF because GP is directly calculated by adding W and L. Same is the case with REB, OREB, & DREB.

VIF Analysis	
Iteration 1	**R^2=39.59%**
Feature	**VIF**
+/-	2.599612607
BLK	5.344890031
STL	11.10601206
3P%	12.36575936
AST	15.67048068
PF	21.52064558
TOV	29.17571175
AGE	37.08921497
FG%	45.24293961
MIN	81.37024963
3PM	137.585495
3PA	199.1988257
PTS	326.554637
FGA	405.0430031
FGM	526.1529587
OREB	626.8481399
DREB	5326.714724
REB	9127.406919
W	inf
GP	inf
L	inf

5.3.2 Parameters

- One way to deal with multicollinearity is to remove one or more variables with very high correlations.
- The complete VIF Analysis Sheet is attached in the **6.** .
- After removing all variables with a VIF over 10, we are left with the parameters +/-, BLK, AST, 3P%, STL, REB, & GP:

			Iteration 9	**R^2=23.76%**			
Feature	+/-	BLK	AST	3P%	STL	REB	GP
VIF	1.04958	4.66425	4.87253	6.92502	8.51160	9.06216	10.17833

5.3.3 Multiple Linear Regression Model

	coef	std err	t	P>|t|
const	65.5799	1.903	34.458	0.000
BLK	-0.3909	1.668	-0.234	0.815
REB	-0.9955	0.293	-3.400	0.001
STL	-0.4035	1.673	-0.241	0.810
+/-	0.4687	0.155	3.030	0.003
GP	0.0719	0.027	2.619	0.009
AST	1.4970	0.342	4.381	0.000
3P%	0.2356	0.043	5.452	0.000

Equation of Multiple Linear Regression Model:

FT% = 65.5799 – (0.3909*BLK) – (0.9955*REB) – (0.4035*STL) + (0.4687*+/-) + (0.0719*GP) + (1.4970*AST) + (0.2356*3P%)

5.3.4 Interpretation of Coefficients

- **Constant**:
  The expected free throw percentage for the average NBA player in this model is 65.5799 %. The p-value for the intercept term is less that 0.05, which tells us that the intercept term is statistically different than zero.

- **Coefficients of Predictor Variables:**
  - On average, each additional block made is associated with a decrease of 0.39 in the free throw shooting percentage. The corresponding p-value is 0.815, which is not statistically significant at the alpha level of 0.05, meaning that average change in free throw percentage for each additional block is not statistically significantly different than zero. Another way to put this might be that the predictor variable block does not have a significant relationship with the response variable free throw percentage. We might say that although our model predicts that a player who has one more block per game than another player will have a lesser free throw percentage by 0.35, this might be due to random chance.
  - Similarly, for each additional steal per game the FT% decreases by 0.403 units. However, an observed difference of 0.403 units in FT% between two players whose number of steals per game differed by one might be completely due to chance because the p-value is greater than 0.05.
  - Each additional rebound on average is associated a decrease of 0.99 ~ 1 in the free throw percentage and the corresponding p-value of 0.001 shows that the average change in FT% for each additional block is statistically significantly different than zero.
  - Similarly, for each additional game played, assist per game, a unit increase in 3P%, or a unit increase in +/-, we see a rise in FT% by 0.072, 1.5, 0.235 & 0.47 units respectively. The associated p-values of these variables tells us that the average change in FT% for each additional unit of these variables is statistically significantly different than zero.

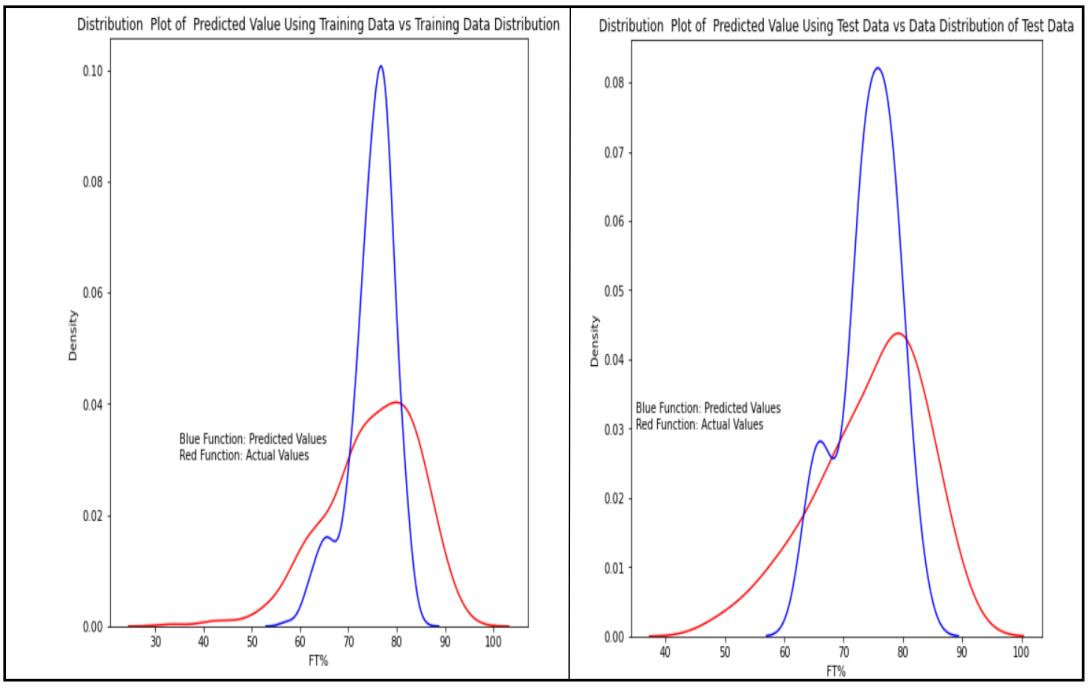
- **Personal Interpretation:**
  1. For a unit increase in 3P%, the FT% increases by 0.235 units. This could be so because a player who has a better 3-point shooting accuracy might just be a better shooter overall and would thus have a higher free throw shooting accuracy.
  2. As with the previous model, +/- and FT% are positively correlated. Thus, the previous explanation still applies.
  3. As with the previous model, the same explanation applies for Steals and Rebounds.
  4. We see that AST and FT% are positively correlated. A possible explanation could be that players who assist more often are more active in making plays with their teammates and get fouled more often. Therefore, they get more chances to shoot free throws.

5.3.5 Train and Test

The data was randomly split in a **90:10** ratio. 90% of the data was used to train the model and 10% of the data was used to test the model.

5.3.6 Predicting FT% for the 2018-19 NBA Season Using Model with variables having VIF<10

- Results: **R^2 for the training data** = 23.7 %
  R^2 for the testing data = 31.63 %



- **Left Graph** - The graph shows the Predicted **Training** FT% values in blue and the actual **Training** FT% values in red.

- **Right Graph** – The graph shows the Predicted **Testing** FT% values in blue and the Actual **Testing** FT% values in red.

- **The model yields a higher R^2 value for the training values than for the testing values.**

- **The model does a better job at tracking the Testing data than the training data, but does not do a good job at the tracking the data overall.**

5.4 Comparison & Conclusion

- It is evident that the lowest AIC model predicts the data better than the model with variables having the lowest VIF.
- Although this is the case, the **lowest VIF model should be preferred** because many variables in the lowest AIC model are directly derived from one another, causing multicollinearity.
- Multicollinearity generates high variance of the estimated coefficients and hence, the coefficient estimates corresponding to those interrelated explanatory variables **will not be accurate in giving us the actual picture.**

6. APPENDIX

6.1 Data Source

- The datasets for this project were obtained from nba.com - the official site of the NBA for the latest NBA Scores, Stats & News.
- Link to the Datasets:
  1. Players General Traditional 2018-19 –

     https://www.nba.com/stats/players/traditional/?sort=PLAYER_NAME&dir=-1&Season=2018-19&SeasonType=Regular%20Season
  2. Players General Traditional 2021-22 –

     https://www.nba.com/stats/players/traditional/?sort=PLAYER_NAME&dir=-1&Season=2021-22&SeasonType=Regular%20Season

6.2 Data Structure

Data_Structure.xlsx

6.3 AIC_VIF_Analysis

AIC_VIF_Analysis_Final.xlsx

6.4 Jupyter Notebook

Polygence_Research_
Paper_Python_Final.pd