CAREER DURATION OF NBA ROOKIE PLAYERS

THE UNIVERSITY OF TEXAS AT ARLINGTON| FALL 2021

INSY-5339-002-PRIN OF BUSINESS DATA MINING

Abstract:

National basketball association (NBA) is one of the prominent worldwide basketball leagues. This dataset consists of each rookie players score point analysis upon which we will be predicting the career longevity for the rookie player. This project focuses on predicting if a rookie is successful or not and uses statistical analysis methods to decide if the player will last for 5 years or more in the NBA or not. Logistic regressionis one method for evaluating the model. From both an empirical and observational approach, our model predicts player's career duration. The default value for the threshold is 0.5. The player is considered as successful if the estimated probability is greater than 0.5.

Business Problem:

In the NBA, player acquisitions can have a significant impact on a team's success. Without the correct players, teams might easily fall short of their goals. Coaches and general managers (GMs) have begun to embrace a more statistics-based approach as analytical approaches have gained prominence in recent years.A traditional way to analyze player is by looking at previous stats with no future predicted analysis which is still an issue that GMs are struggling with today. These adopted traditional methods are limited to deciding whether the player is good/bad. We, as a digital consulting company are providing an added and ameliorated approach of predicting the longevity of these players. For instance, as per these 1-year statistics player A and player B were a good choice to be signed by any franchise. But through our analysis, we can predict if these stats are indeed correct or not. Our consultancy can help and guide the investors and sponsors to invest in right player by predicting that may be player A would have been the right choice but instead of player B, some other player has a promising longevity over 5 years.

Our project attempts to determine the following questions:

1. As a sports manager or coach, if given one year's data, they can tell if the player is good or not. But weas a digital consulting company focus to offer more by predicting if a player will survive in the league forat least for 5 years in NBA.

2. Which factors are significant in determining the performance of the rookie player (in their first year)?
3. Who will be benefitted by this analysis?
    i) A club franchise that invests a large amount of revenue into the players
    ii) Players can improve their performance in specific area

Data for Business Problem Analysis:

To work on the appropriate analysis of the above business problem, the data set containing information about NBA rookie players is used. This data set is a collection of information about 1340 rookie players gathered over a time span of 1 year. This data is primarily focused on the fields that can be used for analyzing the efficiency of any rookie player. The variables of the data set are further explained in detail. This data set has 1340 unique values. We divided the data into training, validation and test in the ratio of 70:20:10. We have used the testing data as the new data to compare and predict the career longevity. We as a consulting firm will use this one-year long performance statistics of these rookie players and predict their career longevity after time span of 5 years.

Independent variables:

- Index – indicates serial numbers.
- Name – indicates name of NBA rookie player
- Games played (gp) – indicates number of games played by a player
- Minutes played (min) – indicates the time in minutes a rookie player has played the game.
- Points per game (pts) – indicates number of points a player scored for number of games played.
- Field goals made (fgm) – indicates the number of points/baskets scored on any shot or tap other than a free throw.
- Field goal attempts (fga) – indicates the number of points scored depending on the distance of the attempt from the basket.
- Field goal percentage (fg) – indicates the percentage which is used to measure how well a player or team shoots the ball during a game.
- 3 points made (3p_made) – indicates the number of points scored when distance of a player's feet from 3point line (which must be completely behind the 3point line at the time of shot or jump).
- 3 points attempt (3pa) – indicates the number of points scored (up to 3 points) for a successful attempt upon scoring a 2-point field goal.
- Free throw made (ftm) - indicates number of points scored when a throw of unopposed attempt ismade by shooting the ball from behind the free throw line.
- Free throw attempts (fta) – indicates number of points scored when an undefended score attempt isawarded to a player after an opposing team commits a foul.
- Free throw percentage (ft) – indicates the percentage of number of free throws made by a player.
- Offensive rebounds (oreb) – indicates the number of points scored when a rebound is made by an offensive player on the same team.
- Defensive rebounds (dreb) – indicates the number of points scored when defensive player gains possession of the basketball after an offensive player misses a field goal.

- Rebounds (reb) – indicates the number of points scored when a team gains possession of the basketball.
- Assists (ast) – indicates the number of points scored when a player passes the ball to teammate in a way that leads to score by field goal.
- Steals (stl) – indicates the number of points scored when a defensive player takes/intercepts the ball from offensive player.
- Blocks (blk) – indicates the number of points scored upon a successful deflection of an attempted shot by a defender.
- Turnovers (tov) – indicates the number of points scored when a team loses possession of the ball to the opposing team before a player takes a shot.

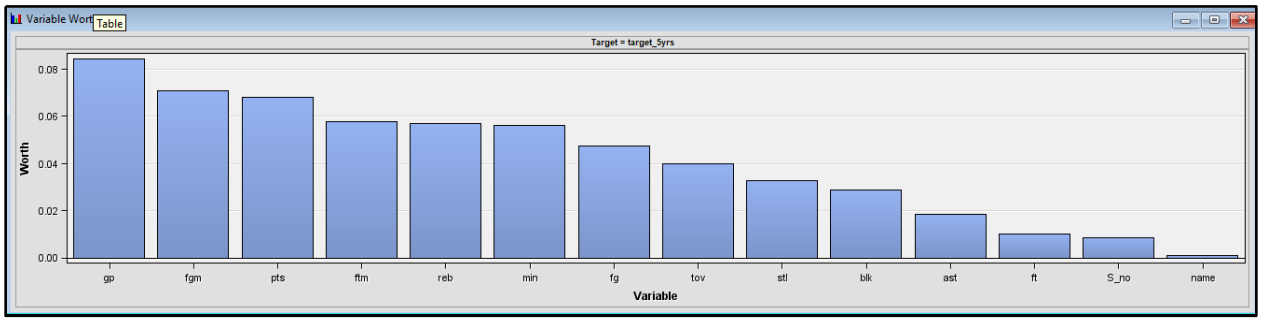
Dependent/Outcome variable:

Career duration more than 5 years (target_5yrs) – indicates 1 as yes for a rookie whose career in NBA is predicted to be more than 5 years in the team and indicates 0 for no for a rookie whose career in NBA is predicted to be not more than 5 years in the team.

Statistical Analysis:

As per our stat explorer analysis on the data we got the following results, and the mentioned top 7 variablesare more significant in predicting our target variable.

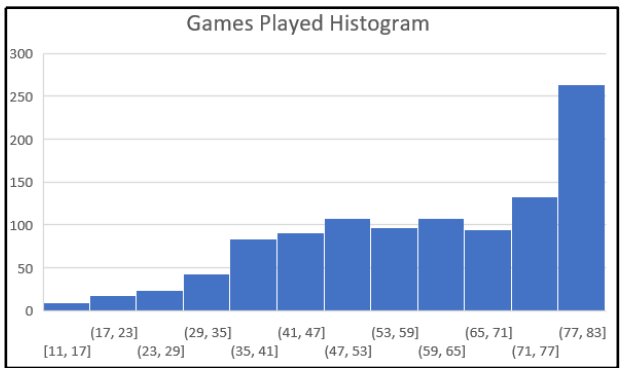
1. Games played (gp)
2. Field goals made (fgm)
3. Points per game (pts)
4. Free throw made (ftm)
5. Rebounds (reb)
6. Minutes played (min)
7. Field goal percentage (fg)

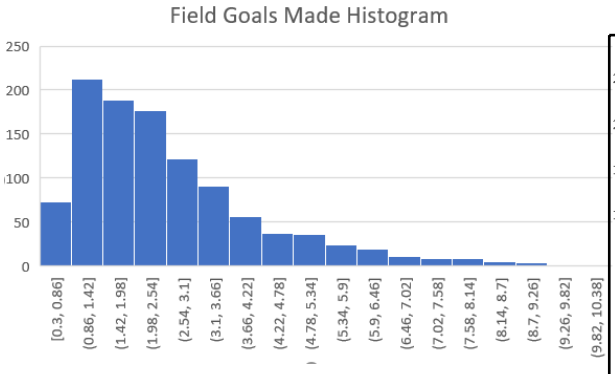


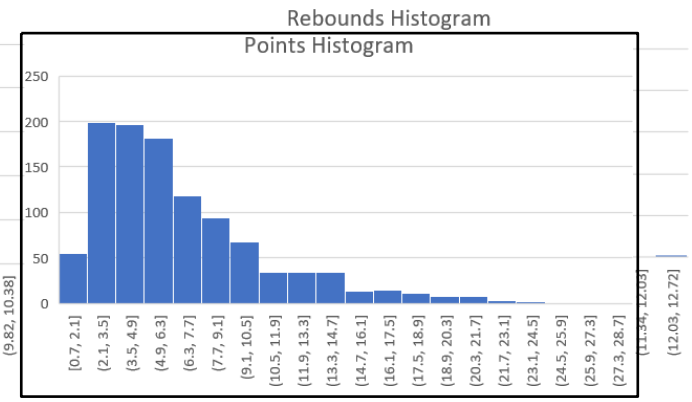
Data Pre-Processing:

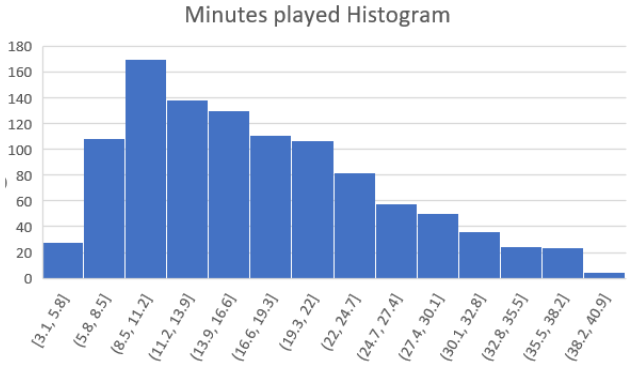
We have performed histogram analysis on our data to analyze the distribution and to find if our variables are skewed. From this we found that our data variables are skewed so we performed transformation of variables (to log10) in SAS EM.

The below are results screenshots of our initial histogram analysis.




Field Goals Made Histogram
250
200
150
100
50
0
[0.3, 0.86]
(0.86, 1.42]
(1.42, 1.98]
(1.98, 2.54]
(2.54, 3.1]
(3.1, 3.66]
(3.66, 4.22]
(4.22, 4.78]
(4.78, 5.34]
(5.34, 5.9]
(5.9, 6.46]
(6.46, 7.02]
(7.02, 7.58]
(7.58, 8.14]
(8.14, 8.7]
(8.7, 9.26]
(9.26, 9.82]
(9.82, 10.38]

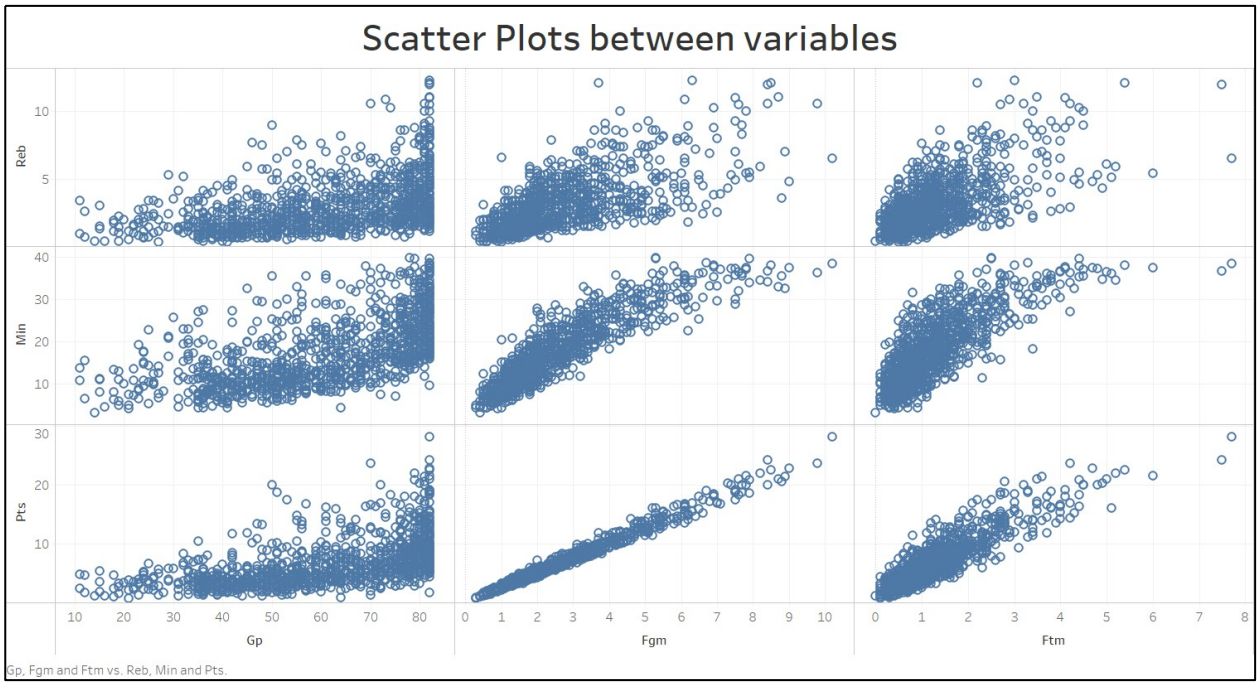

Rebounds Histogram
Points Histogram
250
200
150
100
50
0
[0.7, 2.1]
(2.1, 3.5]
(3.5, 4.9]
(4.9, 6.3]
(6.3, 7.7]
(7.7, 9.1]
(9.1, 10.5]
(10.5, 11.9]
(11.9, 13.3]
(13.3, 14.7]
(14.7, 16.1]
(16.1, 17.5]
(17.5, 18.9]
(18.9, 20.3]
(20.3, 21.7]
(21.7, 23.1]
(23.1, 24.5]
(24.5, 25.9]
(25.9, 27.3]
(27.3, 28.7]
(11.34, 12.03]
(12.03, 12.72]


Minutes played Histogram
180
160
140
120
100
80
60
40
20
0
[3.1, 5.8]
(5.8, 8.5]
(8.5, 11.2]
(11.2, 13.9]
(13.9, 16.6]
(16.6, 19.3]
(19.3, 22]
(22, 24.7]
(24.7, 27.4]
(27.4, 30.1]
(30.1, 32.8]
(32.8, 35.5]
(35.5, 38.2]
(38.2, 40.9]

Data Visualization:

Scatter plots:

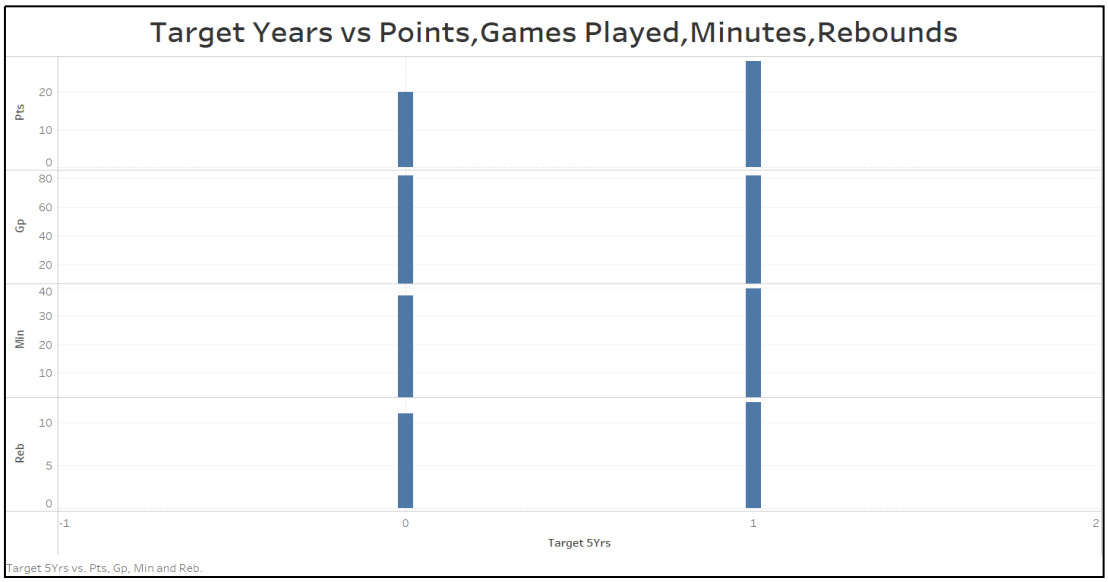
For our analysis, we used Scatter Plot to perform data Visualization. Scatter plot are used for identification of correlational relationships between variables.



From the plots, we can say that games played have correlation with the minutes played.

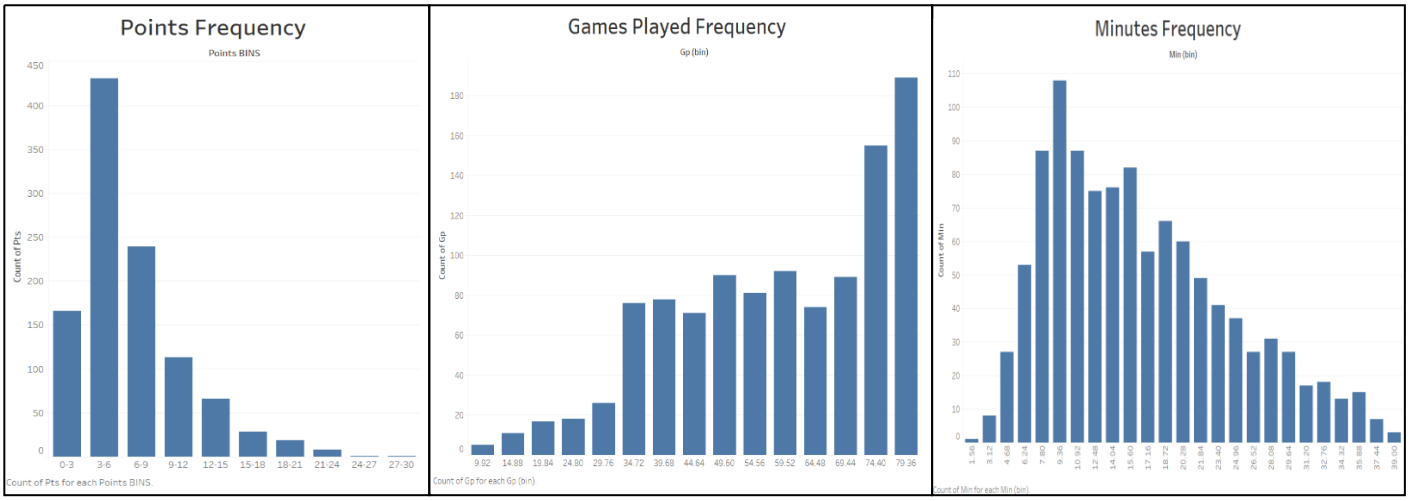
Bar chart:

Bar charts to visualize the cutoff values for players to analyze their career longevity with less than 5 years of career duration.



Histograms:

We used histograms to find the distributions of 3 most significant variables. The plots show that the data has been skewed.



Data Predictions techniques:

As mentioned in dataset definitions, the response variable is defined as binary output. Particularly in the case, Rookie of the Year Award would be defined as 1 (or successful) and 0 (or unsuccessful) for the rest. We have used two supervised learning techniques for our analysis:

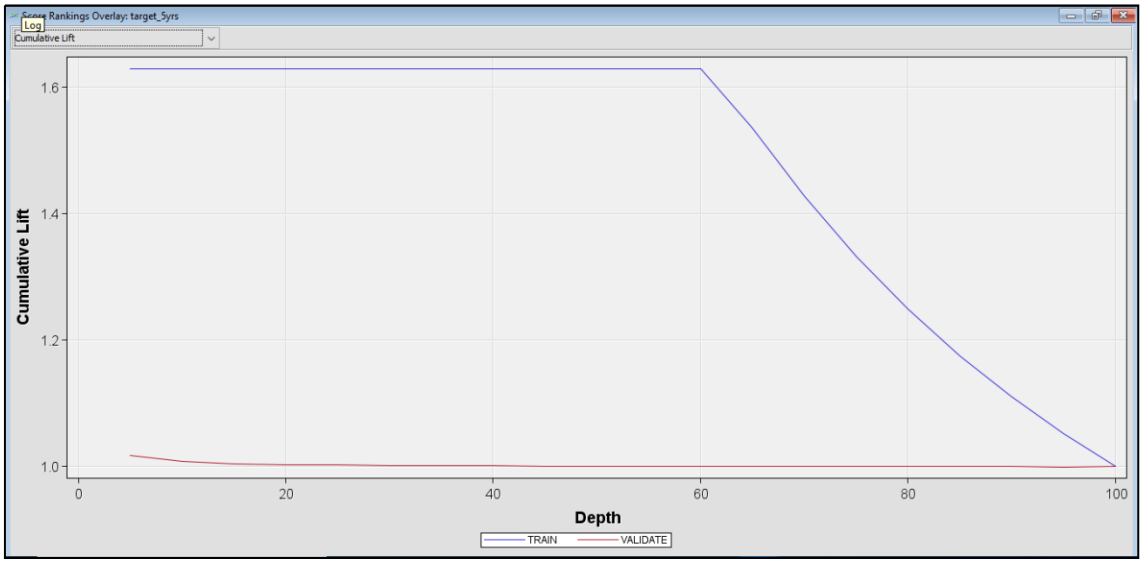
1. Logistic Regression
2. Linear Regression

1. Logistic Regression:

Logistic function is regarded as a sigmoid function. Consequently, the function approaches in determining the probability of either 0 or 1.

a. Logistic Regression (Model-None selected):

This technique is used to predict the target variable for rookie player being successful for 5 years.



From the above graph, we can see that 60% of our training data which is predicted as successful is 1.25 times likely to be positive and successful.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
target_5yrs	target_5yrs	_AIC_	Akaike's Information Criterion	1503.279	.	.
target_5yrs	target_5yrs	_ASE_	Average Squared Error	0.006007	0.256221	0.24175
target_5yrs	target_5yrs	_AVERR_	Average Error Function	0.016852	1.129606	0.840119
target_5yrs	target_5yrs	_DFE_	Degrees of Freedom for Error	11	.	.
target_5yrs	target_5yrs	_DFM_	Model Degrees of Freedom	739	.	.
target_5yrs	target_5yrs	_DFT_	Total Degrees of Freedom	750	.	.
target_5yrs	target_5yrs	_DIV_	Divisor for ASE	1500	428	216
target_5yrs	target_5yrs	_ERR_	Error Function	25.27856	483.4712	181.4657
target_5yrs	target_5yrs	_FPE_	Final Prediction Error	0.813068	.	.
target_5yrs	target_5yrs	_MAX_	Maximum Absolute Error	0.508364	1	0.999956
target_5yrs	target_5yrs	_MSE_	Mean Square Error	0.409537	0.256221	0.24175
target_5yrs	target_5yrs	_NOBS_	Sum of Frequencies	750	214	108
target_5yrs	target_5yrs	_NW_	Number of Estimate Weights	739	.	.
target_5yrs	target_5yrs	_RASE_	Root Average Sum of Squares	0.077502	0.506183	0.49168
target_5yrs	target_5yrs	_RFPE_	Root Final Prediction Error	0.901703	.	.
target_5yrs	target_5yrs	_RMSE_	Root Mean Squared Error	0.639951	0.506183	0.49168
target_5yrs	target_5yrs	_SBC_	Schwarz's Bayesian Criterion	4917.513	.	.
target_5yrs	target_5yrs	_SSE_	Sum of Squared Errors	9.009819	109.6628	52.21792
target_5yrs	target_5yrs	_SUMW_	Sum of Case Weights Times Freq	1500	428	216
target_5yrs	target_5yrs	_MISC_	Misclassification Rate	0.012	0.415888	0.361111

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
1000.851	25.279	975.5719	738	<.0001

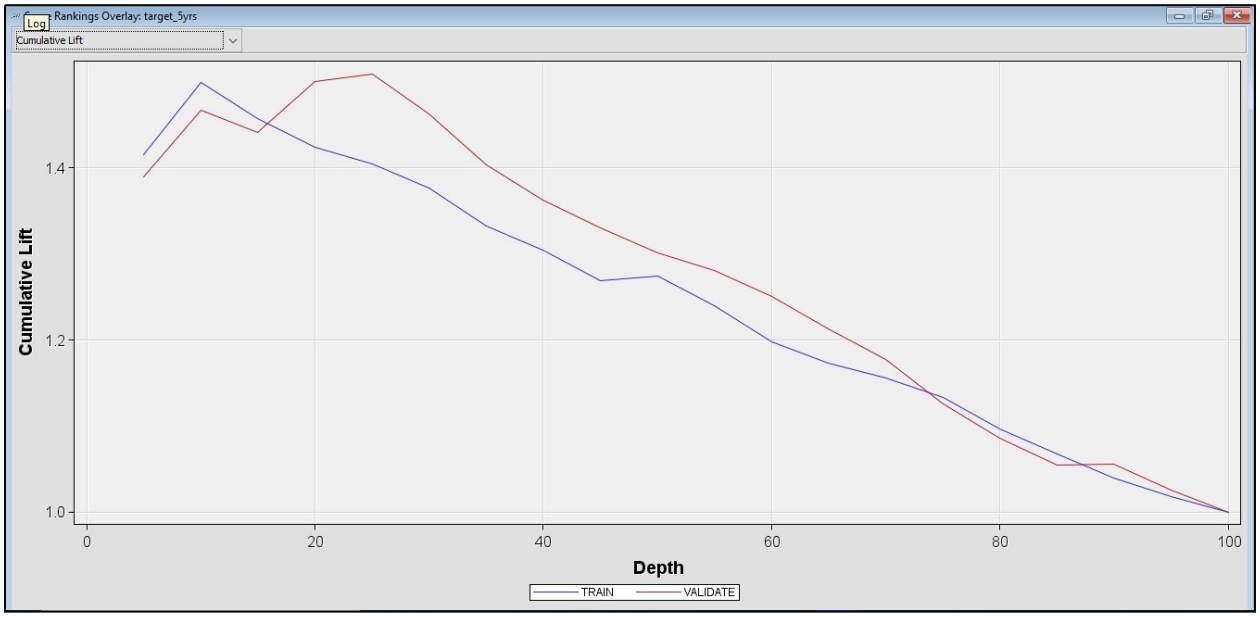
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LG10_ast	1	0.0220	0.8820
LG10_blk	1	0.0001	0.9929
LG10_fg	1	0.0000	1.0000
LG10_fgm	1	0.5088	0.4757
LG10_ft	1	0.0000	1.0000
LG10_ftm	1	0.0100	0.9205
LG10_gp	1	0.0004	0.9837
LG10_min	1	0.2797	0.5969
LG10_pts	1	0.0000	1.0000
LG10_reb	1	0.0000	1.0000
LG10_stl	1	0.0070	0.9332
LG10_tov	1	0.0315	0.8591
S_no	1	0.0118	0.9133
name	725	26.7980	1.0000

The P value, low misclassification rates and low average squared errors shows that the model issignificant.

b. Logistic Regression (Model – Backward regression):

This technique is also known for fitting regression models in which the choice of predictive variables is carried out byan automatic procedure. Initially this model takes all our variables into consideration and in each step, it gradually eliminates the variable from the final model based on its significance (prespecified



criteria)

From the above graph, we can see that 50% of our training data which is predicted as successful is 1.3 times likely to be positive and successful.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
target_5yrs	target_5yrs	_AIC_	Akaike's Information Criterion	1503.279	.
target_5yrs	target_5yrs	_ASE_	Average Squared Error	0.006007	0.256221
target_5yrs	target_5yrs	_AVERR_	Average Error Function	0.016852	1.129606
target_5yrs	target_5yrs	_DFE_	Degrees of Freedom for Error	11	.
target_5yrs	target_5yrs	_DFM_	Model Degrees of Freedom	739	.
target_5yrs	target_5yrs	_DFT_	Total Degrees of Freedom	750	.
target_5yrs	target_5yrs	_DIV_	Divisor for ASE	1500	428
target_5yrs	target_5yrs	_ERR_	Error Function	25.27856	483.4712
target_5yrs	target_5yrs	_FPE_	Final Prediction Error	0.813068	.
target_5yrs	target_5yrs	_MAX_	Maximum Absolute Error	0.508364	1
target_5yrs	target_5yrs	_MSE_	Mean Square Error	0.409537	0.256221
target_5yrs	target_5yrs	_NOBS_	Sum of Frequencies	750	214
target_5yrs	target_5yrs	_NW_	Number of Estimate Weights	739	.
target_5yrs	target_5yrs	_RASE_	Root Average Sum of Squares	0.077502	0.506183
target_5yrs	target_5yrs	_RFPE_	Root Final Prediction Error	0.901703	.
target_5yrs	target_5yrs	_RMSE_	Root Mean Squared Error	0.639951	0.506183
target_5yrs	target_5yrs	_SBC_	Schwarz's Bayesian Criterion	4917.513	.
target_5yrs	target_5yrs	_SSE_	Sum of Squared Errors	9.009819	109.6628
target_5yrs	target_5yrs	_SUMW_	Sum of Case Weights Times Freq	1500	428
target_5yrs	target_5yrs	_MISC_	Misclassification Rate	0.012	0.415888

The selected model is the model trained in the last step (Step 12). It consists of the following effects:

Intercept | LG10_blk LG10_fgm

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
1000.851	896.770	104.0809	2	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LG10_blk	1	8.6411	0.0033
LG10_fgm	1	57.6562	<.0001

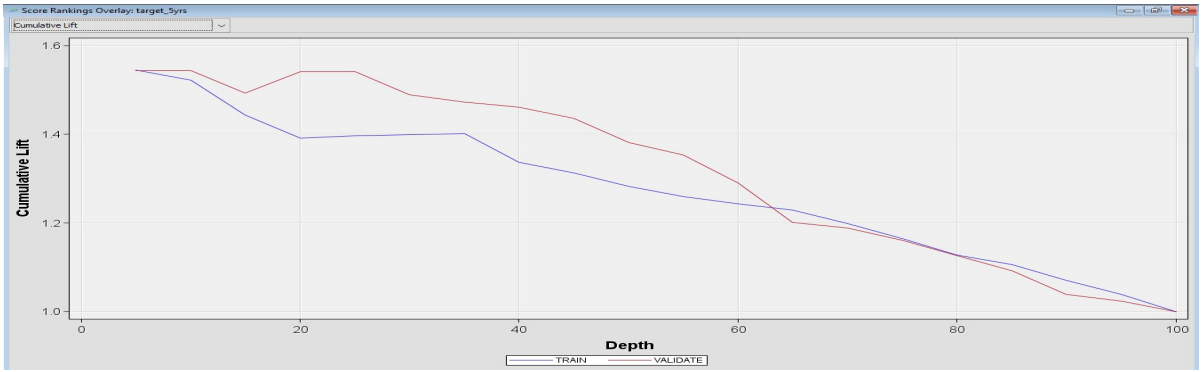
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-1.0865	0.1529	50.47	<.0001	
LG10_blk	1	1.6279	0.5538	8.64	0.0033	0.1721
LG10_fgm	1	2.3887	0.3146	57.66	<.0001	0.4301

The P value (<0.0001), low misclassification rates and low average squared errors shows that the model is significant.

c. Logistic regression (Model – Forward Regression):

This technique is a type of regression which begins with an empty model and adds in significant variables one by one which results in final best model



From the above graph, we can see that 35% of our training data which is predicted as successful is 1.4 times likely to be positive and successful

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
target_5yrs	target_5yrs	_AIC_	Akaike's Information Criterion	854.3114	.
target_5yrs	target_5yrs	_ASE_	Average Squared Error	0.190712	0.194183
target_5yrs	target_5yrs	_AVERR_	Average Error Function	0.562874	0.574334
target_5yrs	target_5yrs	_DFE_	Degrees of Freedom for Error	745	.
target_5yrs	target_5yrs	_DFM_	Model Degrees of Freedom	5	.
target_5yrs	target_5yrs	_DFT_	Total Degrees of Freedom	750	.
target_5yrs	target_5yrs	_DIV_	Divisor for ASE	1500	428
target_5yrs	target_5yrs	_ERR_	Error Function	844.3114	245.8149
target_5yrs	target_5yrs	_FPE_	Final Prediction Error	0.193272	.
target_5yrs	target_5yrs	_MAX_	Maximum Absolute Error	0.94746	0.947835
target_5yrs	target_5yrs	_MSE_	Mean Square Error	0.191992	0.194183
target_5yrs	target_5yrs	_NOBS_	Sum of Frequencies	750	214
target_5yrs	target_5yrs	_NW_	Number of Estimate Weights	5	.
target_5yrs	target_5yrs	_RASE_	Root Average Sum of Squares	0.436706	0.440662
target_5yrs	target_5yrs	_RFPE_	Root Final Prediction Error	0.439628	.
target_5yrs	target_5yrs	_RMSE_	Root Mean Squared Error	0.438169	0.440662
target_5yrs	target_5yrs	_SBC_	Schwarz's Bayesian Criterion	877.4118	.
target_5yrs	target_5yrs	_SSE_	Sum of Squared Errors	286.0687	83.11023
target_5yrs	target_5yrs	_SUMW_	Sum of Case Weights Times Freq	1500	428
target_5yrs	target_5yrs	_MISC_	Misclassification Rate	0.284	0.313084

Intercept LG10_fg LG10_ftm LG10_gp LG10_reb

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
1000.851	844.311	156.5391	4	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LG10_fg	1	6.6818	0.0097
LG10_ftm	1	4.1626	0.0413
LG10_gp	1	39.0288	<.0001
LG10_reb	1	2.9799	0.0843

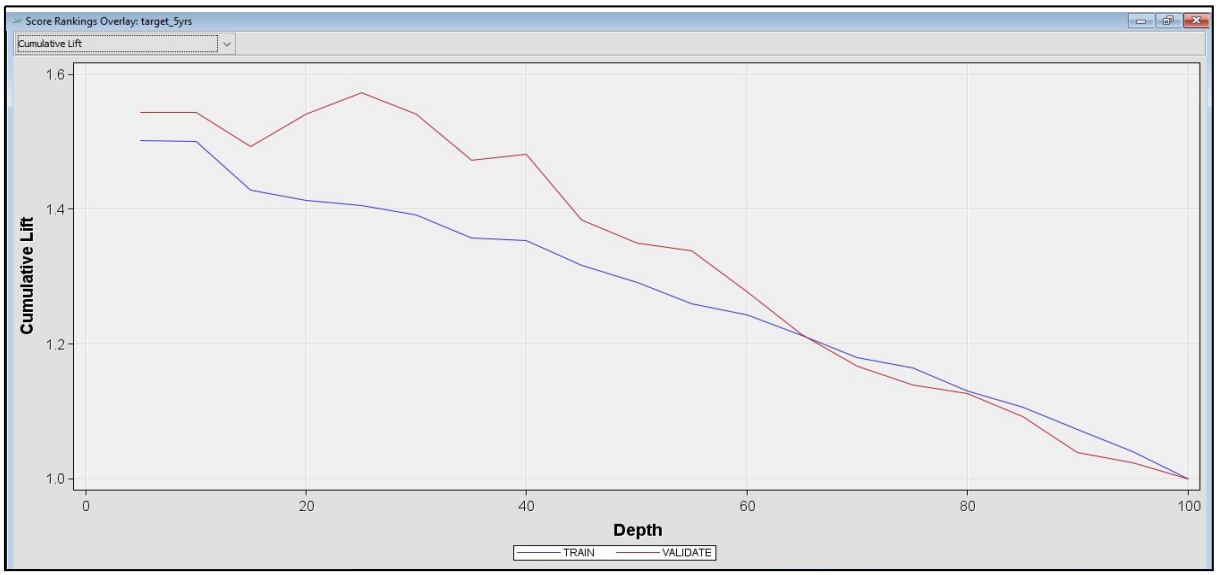
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-9.2991	1.6601	31.38	<.0001	
LG10_fg	1	2.4882	0.9626	6.68	0.0097	0.1498
LG10_ftm	1	0.9515	0.4664	4.16	0.0413	0.1482
LG10_gp	1	2.7371	0.4381	39.03	<.0001	0.3949
LG10_reb	1	0.6391	0.3702	2.98	0.0843	0.1290

The P value (<0.0001), low misclassification rates and low average squared errors shows that the model is significant.

d. Logistic Regression (Model- Stepwise regression):

This technique is a step-by-step iterative approach that involves selection of our significant variables fromthe data to be used in final model.



From the above graph, we can see that 35% of our training data which is predicted as successful is 1.4 times likely to be positive and successful

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
target_5yrs	target_5yrs	_AIC_	Akaike's Information Criterion	855.2975	.
target_5yrs	target_5yrs	_ASE_	Average Squared Error	0.191594	0.196768
target_5yrs	target_5yrs	_AVERR_	Average Error Function	0.564865	0.5811
target_5yrs	target_5yrs	_DFE_	Degrees of Freedom for Error	746	.
target_5yrs	target_5yrs	_DFM_	Model Degrees of Freedom	4	.
target_5yrs	target_5yrs	_DFT_	Total Degrees of Freedom	750	.
target_5yrs	target_5yrs	_DIV_	Divisor for ASE	1500	428
target_5yrs	target_5yrs	_ERR_	Error Function	847.2975	248.7107
target_5yrs	target_5yrs	_FPE_	Final Prediction Error	0.193649	.
target_5yrs	target_5yrs	_MAX_	Maximum Absolute Error	0.935575	0.968193
target_5yrs	target_5yrs	_MSE_	Mean Square Error	0.192621	0.196768
target_5yrs	target_5yrs	_NOBS_	Sum of Frequencies	750	214
target_5yrs	target_5yrs	_NW_	Number of Estimate Weights	4	.
target_5yrs	target_5yrs	_RASE_	Root Average Sum of Squares	0.437715	0.443585
target_5yrs	target_5yrs	_RFPE_	Root Final Prediction Error	0.440055	.
target_5yrs	target_5yrs	_RMSE_	Root Mean Squared Error	0.438887	0.443585
target_5yrs	target_5yrs	_SBC_	Schwarz's Bayesian Criterion	873.7778	.
target_5yrs	target_5yrs	_SSE_	Sum of Squared Errors	287.3912	84.21668
target_5yrs	target_5yrs	_SUMW_	Sum of Case Weights Times Freq	1500	428
target_5yrs	target_5yrs	_MISC_	Misclassification Rate	0.297333	0.327103

The selected model is the model trained in the last step (Step 5). It consists of the following effects:

Intercept LG10_fg LG10_ftm LG10_gp

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
1000.851	847.297	153.5530	3	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
LG10_fg	1	12.5699	0.0004
LG10_ftm	1	11.3511	0.0008
LG10_gp	1	43.7517	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-10.3892	1.5438	45.29	<.0001	
LG10_fg	1	3.1437	0.8867	12.57	0.0004	0.1893
LG10_ftm	1	1.3575	0.4029	11.35	0.0008	0.2114
LG10_gp	1	2.8646	0.4331	43.75	<.0001	0.4132

The P value (<0.0001), low misclassification rates and low average squared errors shows that the model is significant.

2. Multiple linear regression:

We have used Multiple linear regression to estimate how our dependent variable change as the independent variable(s) changes

This technique estimates how our dependent variable changes as the independent variable(s) change. Thismethod is used to estimate the relationship between two or more independent variables and one dependentvariable.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
target_5yrs	target_5yrs	_AIC_	Akaike's Information Criterion	-2365.97	.
target_5yrs	target_5yrs	_ASE_	Average Squared Error	0.005944	485.2574
target_5yrs	target_5yrs	_AVERR_	Average Error Function	0.005944	485.2574
target_5yrs	target_5yrs	_DFE_	Degrees of Freedom for Error	11	.
target_5yrs	target_5yrs	_DFM_	Model Degrees of Freedom	739	.
target_5yrs	target_5yrs	_DFT_	Total Degrees of Freedom	750	.
target_5yrs	target_5yrs	_DIV_	Divisor for ASE	750	214
target_5yrs	target_5yrs	_ERR_	Error Function	4.458333	103845.1
target_5yrs	target_5yrs	_FPE_	Final Prediction Error	0.804662	.
target_5yrs	target_5yrs	_MAX_	Maximum Absolute Error	0.541671	208.7005
target_5yrs	target_5yrs	_MSE_	Mean Square Error	0.405303	485.2574
target_5yrs	target_5yrs	_NOBS_	Sum of Frequencies	750	214
target_5yrs	target_5yrs	_NW_	Number of Estimate Weights	739	.
target_5yrs	target_5yrs	_RASE_	Root Average Sum of Squares	0.0771	22.02856
target_5yrs	target_5yrs	_RFPE_	Root Final Prediction Error	0.897029	.
target_5yrs	target_5yrs	_RMSE_	Root Mean Squared Error	0.636634	22.02856
target_5yrs	target_5yrs	_SBC_	Schwarz's Bayesian Criterion	1048.26	.
target_5yrs	target_5yrs	_SSE_	Sum of Squared Errors	4.458333	103845.1
target_5yrs	target_5yrs	_SUMW_	Sum of Case Weights Times Freq	750	214
target_5yrs	target_5yrs	_MISC_	Misclassification Rate	0.010667	0.425234

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	40.857031	3.142849	18.55	<.0001
Error	808	136.904526	0.169436		
Corrected Total	821	177.761557			

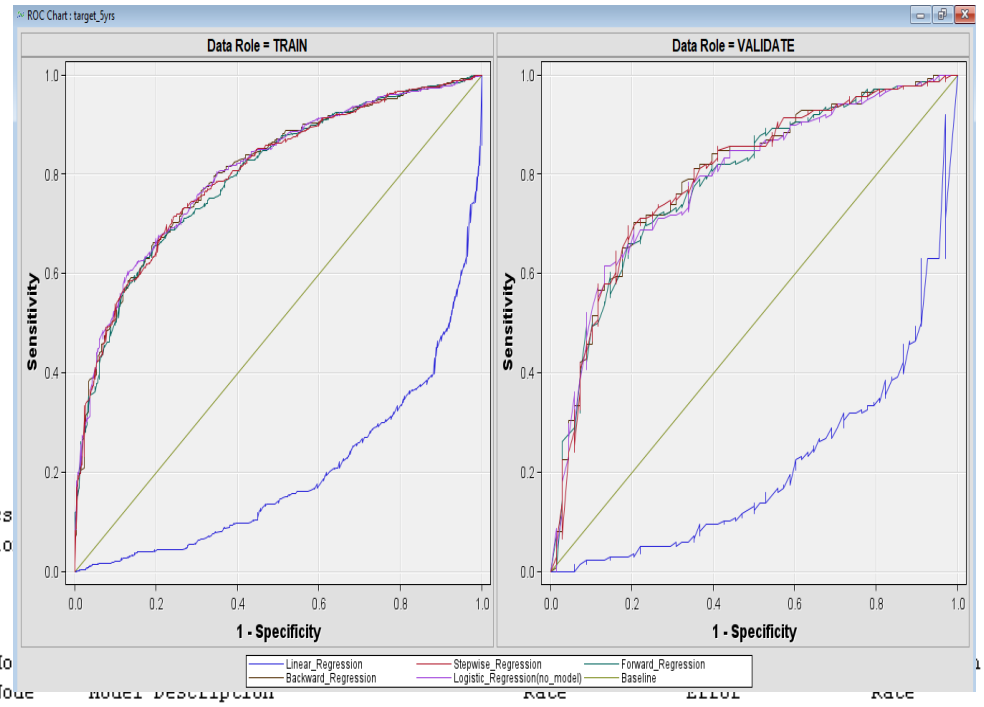
Model Fit Statistics

R-Square	0.2298	Adj R-Sq	0.2175
AIC	-1445.3993	BIC	-1442.9148
SBC	-1379.4350	C(p)	14.0000

From the above results we can see that our data does not fit the model as our R square value is not significant.

Model Comparison- Best model:

Upon performing the prediction techniques on our data set all our regression techniques have similar results and Stepwise Regression has least averaged squared error and misclassification rate, so we have selected "STEP WISE REGRESSION" model as our best model



Fit Statistics
Model Selectio

Selected Model	Mo Node	Model Description	Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate	Valid: Average Squared Error
Y	Reg4	Stepwise_Regression	0.24757	0.16597	0.25669	0.17173
	Reg5	Linear_Regression	0.25243	0.16655	0.24696	0.17966
	Reg3	Forward_Regression	0.26699	0.16692	0.25182	0.17378
	Reg2	Backward_Regression	0.27184	0.16494	0.24331	0.17189
	Reg	Logistic_Regression(no_model)	0.27184	0.16431	0.24574	0.17405

Business Problem Solution:

As we have been discussing throughout the analysis, there are lot of attributes that would define a player's performance. After running our models, we got the best model as:

Intercept LG10_fg LG10_ftm LG10_gp LG10_reb

The significant variables that would determine a rookie player's performance are FG, FTM, GP and REB.

For every 1 field goal made the player might have a better chance to have successful career or better performance, which is similar with other significant variables.

We as a digital consulting firm can now target the right group of rookie players for our client Houston rockets team. This helps management team to evaluate the team's budget for trading/extending the contract of that that rookie player. For instance, we have selected 3 players from our test data who are the current rookie players in Houston rockets team and compared them with our training data which comprises of rookie players who were successful for at least 5 years and not successful for 5 years.

From the results we can say that the current rookie player who has played a fewer number of games is successful compared with the rookie player who had failed despite playing a greater number of games. Hereby, we conclude that this prediction analysis is different from the traditional way of analyzing the player's performance and it also helps the coaches and players to evaluate the performance.