Final Project

Problem Description:

Win Probability for Basketball Games.

As analysts we would like to predict the outcome off each game with the NBA. Analyzing the winning outcome for NBA's basketball games. Additionally, calculating the total points of each individual game through the given team statistics.

Analysis Questions:

How can we predict the number of points made by a basketball team in a single game?

Can the shots attempt, rebounds, steals, personal fouls, away/home game etc. Predict the outcome of the game (e.g., win/loss)?

Label quantitative/continuous variable.

PtsTeam: Total points made in the game by the NBA.

Label at least one qualitative/classification variable.

OutcomeGame: The outcome of the game is abbreviated with an L for the loosing or W for winning team.

Sources:

Official NBA Stats. NBA Stats. (n.d.). https://www.nba.com/stats/. ..

Variables:

SlugTeam - The team abbreviation of the base team we are analyzing the wins for NBA Teams.

SlugSeason- Competition year.

IsB2B: If the game played was or is going to be back-to-back for the NBA Teams.

LocationGame: If the game was at the home or away arena.

CountDaysRestTeam: The number of days the team had to rest before the game being played.

SlugOpponent: The abbreviation of the opposition team the NBA Teams are playing.

SlugTeamWinner: The team who won the game.

SlugTeamLoser: The Team abbreviation of the losing team.

OutcomeGame: The outcome of the game is abbreviated with an L for the loosing or W for winning team.

FgmTeam: The field goals made in the game for the NBA Teams.

FgaTeam: Field goals attempted in the game for the NBA Teams.

pctFGTeam(accuracy): The accuracy of field goals attempted by the NBA Teams, calculated from: FGM/ FGA.

Fg3mTeam: 3-pointer shots made in the game by the NBA Teams.

Fg3aTeam: 3-pointer shots attempted in the game by the NBA Teams.

PctFG3Team: Accuracy of 3-pointer shots made in the game by the NBA Teams.

TrebTeam: Total rebounds had by the NBA Teams.

StlTeam: Total steals had by the NBA Teams.

TovTeam: Total turnovers made by the NBA Teams.

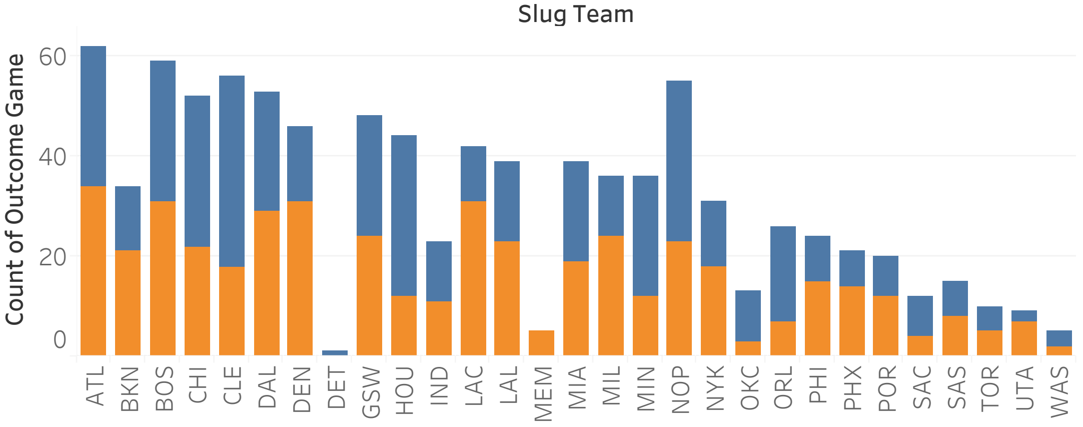
PfTeam: Personal fouls had by the NBA Teams.

PtsTeam: Total points made in the game by the NBA Teams.

Final Project

Intoduction | Breakdown of the dataset. | Model 1: Predictive Mo.. | Model 2: Predictive Mo.. | Model comparison o.. | Model 3: Logistical Reg.. | Model 4: Classifi..

W/L Breakdown by team



KPI for Points

🟩 Above Bench..

🟥 Below Bench..

KPI for Points

✓ Above Bench..

✗ Below Bench..

Outcome Game

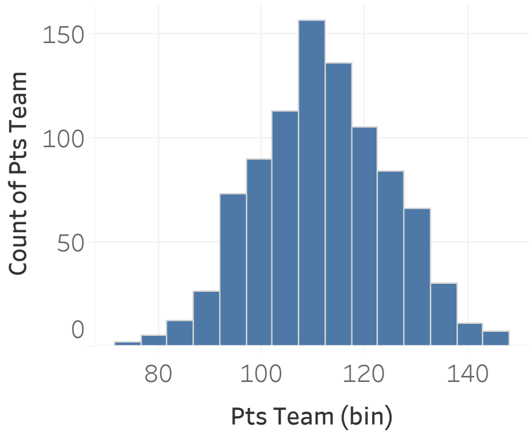
🟦 L

🟧 W

kpi

	Date Game				
Name Tea..	Janu..	Febr..	March	April	Dece
Atlanta H..	✗	✗	✗	✓	✓
Boston C..	✗	✗	✓	✗	✗
Brooklyn ..	✓	✓	✗	✗	✗
Chicago B..	✓	✗	✗	✗	✓
Cleveland..	✗	✗	✗	✗	✗
Dallas Ma..	✗	✗	✗	✗	✗
Denver N..	✓	✗	✗	✓	✓
Detroit Pi..			✗		
Golden St..	✗	✗	✗	✗	✗
Houston ..	✗	✗	✗	✗	✓
Indiana P..	✗	✗	✓	✓	
LA Clippers	✓	✗	✓	✓	✓
Los Angel..	✗	✗	✗	✗	✓

distribution of points

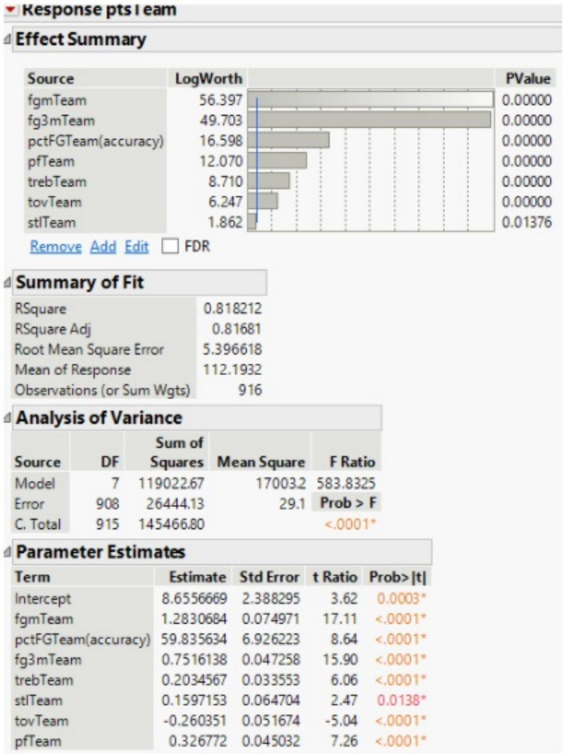


Final Project

Regression Results for continous variable - **total points scored**

The regression model will help us predict total points made in the game by the NBA Teams. The model was created using the normal regression model. All the variables were used in this model. The r-square for the model is 82%, and the p-value for the whole model is .001 smaller than the alpha .05 which makes this a good model. The Variables included in the model are: field goals made, accuracy of field goals, 3pointers made, total rebounds, steals, turn overs, personal fouls. All of them are significant with a p-value < .05.

The formula is:
8.6557 +1.2831 (fgmTeam) + 59.8356 (pctFGTeam (accuracy)) + 0.7516 (fg3mTeam) + 0.2036 (trebTeam) +0.1597 (stlTeam) - 0.2604 (tovTeam) + 0.3268 (pfTeam)



Response ptsTeam

Effect Summary

Source	LogWorth		PValue
fgmTeam	56.397		0.00000
fg3mTeam	49.703		0.00000
pctFGTeam(accuracy)	16.598		0.00000
pfTeam	12.070		0.00000
trebTeam	8.710		0.00000
tovTeam	6.247		0.00000
stlTeam	1.862		0.01376

Summary of Fit

RSquare	0.818212
RSquare Adj	0.81681
Root Mean Square Error	5.396618
Mean of Response	112.1932
Observations (or Sum Wgts)	916

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	119022.67	17003.2	583.8325
Error	908	26444.13	29.1	Prob > F
C. Total	915	145466.80		<.0001*

Parameter Estimates

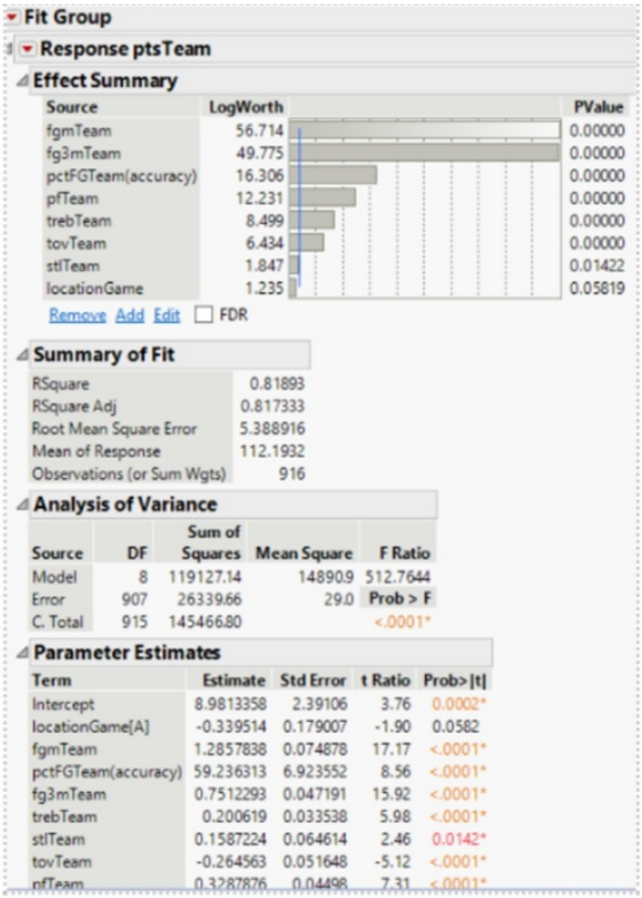
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.6556669	2.388295	3.62	0.0003*
fgmTeam	1.2830684	0.074971	17.11	<.0001*
pctFGTeam(accuracy)	59.835634	6.926223	8.64	<.0001*
fg3mTeam	0.7516138	0.047258	15.90	<.0001*
trebTeam	0.2034567	0.033553	6.06	<.0001*
stlTeam	0.1597153	0.064704	2.47	0.0138*
tovTeam	-0.260351	0.051674	-5.04	<.0001*
pfTeam	0.326772	0.045032	7.26	<.0001*

Final Project

Backward Regression Results for continous variable - **total points scored**

The regression model will help us predict total points made in the game by the NBA Teams. The model was created using the stepwise regression with a backwards direction with a p-value threshold. This model has an R-square of 82% data. The p-value of the whole model is .001 smaller than our alpha .05, meaning the model is a good model. The Variables included in the model are: field goals made, accuracy of field goals, 3pointers made, total rebounds, steals, turn overs, personal fouls, and location of game. All of them are significant with a p-value < .05 besides location of game.

The formula for total points made in the game by the NBA teams:
8.98 - 0.3395 (locationGame(A)) + 1.2857 (fgmTeam) + 59.2363 (pctFGTeam(accuracy) + 0.7513 (fg3mTeam) + 0.2006 (trebTeam) + 0.1587 (stlTeam) - 0.2646 (tovTeam) + 0.3288 (pfTeam)



Fit Group

Response ptsTeam

Effect Summary

Source	LogWorth	PValue
fgmTeam	56.714	0.00000
fg3mTeam	49.775	0.00000
pctFGTeam(accuracy)	16.306	0.00000
pfTeam	12.231	0.00000
trebTeam	8.499	0.00000
tovTeam	6.434	0.00000
stlTeam	1.847	0.01422
locationGame	1.235	0.05819

Summary of Fit

RSquare	0.81893
RSquare Adj	0.817333
Root Mean Square Error	5.388916
Mean of Response	112.1932
Observations (or Sum Wgts)	916

Analysis of Variance

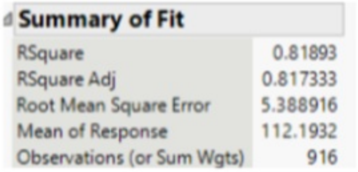
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	119127.14	14890.9	512.7644
Error	907	26339.66	29.0	Prob > F
C. Total	915	145466.80		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.9813358	2.39106	3.76	0.0002*
locationGame[A]	-0.339514	0.179007	-1.90	0.0582
fgmTeam	1.2857838	0.074878	17.17	<.0001*
pctFGTeam(accuracy)	59.236313	6.923552	8.56	<.0001*
fg3mTeam	0.7512293	0.047191	15.92	<.0001*
trebTeam	0.200619	0.033538	5.98	<.0001*
stlTeam	0.1587224	0.064614	2.46	0.0142*
tovTeam	-0.264563	0.051648	-5.12	<.0001*
pfTeam	0.3287876	0.04498	7.31	<.0001*

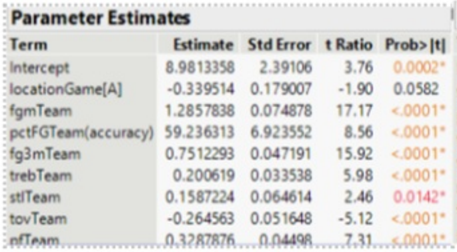
Final Project

Backward stepwise regression



Summary of Fit

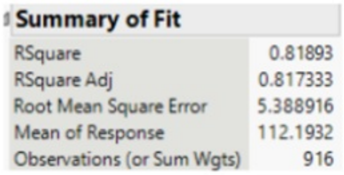
RSquare	0.81893
RSquare Adj	0.817333
Root Mean Square Error	5.388916
Mean of Response	112.1932
Observations (or Sum Wgts)	916



Parameter Estimates

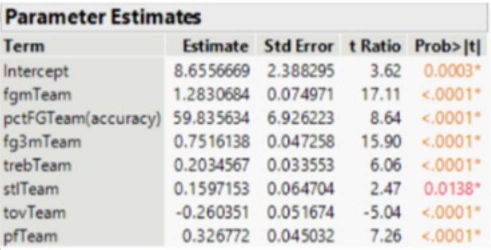
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.9813358	2.39106	3.76	0.0002*
locationGame[A]	-0.339514	0.179007	-1.90	0.0582
fgmTeam	1.2857838	0.074878	17.17	<.0001*
pctFGTeam(accuracy)	59.236313	6.923552	8.56	<.0001*
fg3mTeam	0.7512293	0.047191	15.92	<.0001*
trebTeam	0.200619	0.033538	5.98	<.0001*
stlTeam	0.1587224	0.064614	2.46	0.0142*
tovTeam	-0.264563	0.051648	-5.12	<.0001*
pfTeam	0.3287876	0.04498	7.31	<.0001*

Linear regression



Summary of Fit

RSquare	0.81893
RSquare Adj	0.817333
Root Mean Square Error	5.388916
Mean of Response	112.1932
Observations (or Sum Wgts)	916



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.6556669	2.388295	3.62	0.0003*
fgmTeam	1.2830684	0.074971	17.11	<.0001*
pctFGTeam(accuracy)	59.835634	6.926223	8.64	<.0001*
fg3mTeam	0.7516138	0.047258	15.90	<.0001*
trebTeam	0.2034567	0.033553	6.06	<.0001*
stlTeam	0.1597153	0.064704	2.47	0.0138*
tovTeam	-0.260351	0.051674	-5.04	<.0001*
pfTeam	0.326772	0.045032	7.26	<.0001*

Comparing the regression and backward stepwise regression

The best model to predict the total points made in the game by an NBA Team is the model with the normal regression because all its variables are significant. Both models have very similar adjusted r-squared values, backward regression = adj r squared= .81681 and the normal regression= adj r squared= .81733 so we are deciding based off the p-values of the variables in the model given the adjusted r squared values are almost equivalent.

Therefore, the predicted equation is:
8.6557 +1.2831 (fgmTeam) + 59.8356 (pctFGTeam (accuracy)) + 0.7516 (fg3mTeam) + 0.2036 (trebTeam) +0.1597 (stlTeam) - 0.2604 (tovTeam) + 0.3268 (pfTeam)

Final Project

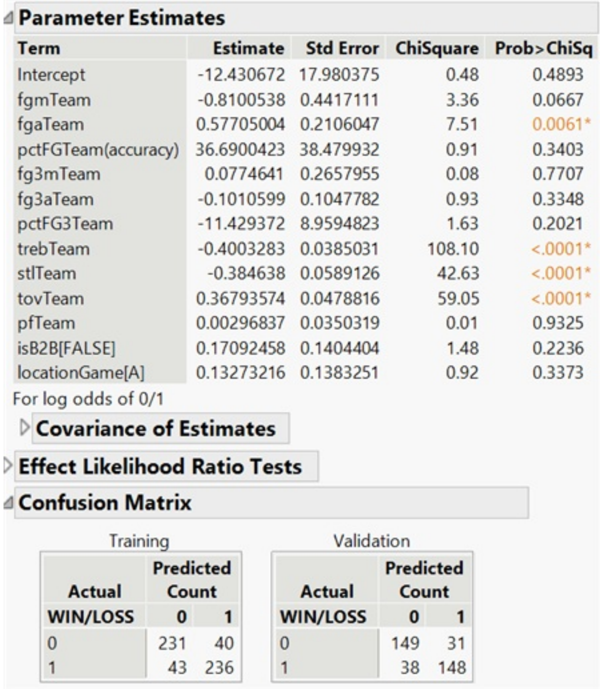
Model 1: Predictive .. | Model 2: Predictive Mo.. | Model comparison o.. | Model 3: Logistical Reg.. | Model 4: Classification .. | Model 5: K nearest neigh.. | Model 3, 4 &5 comparison

Logistical Regression for categorical variable - **win or loss**

A logistical regression model was conducted to predict if an NBA team will win or loose. The Variables included in the model are: field goals attempted, accuracy of field goals, 3pointers made, 3pointers attempted, accuracy of 3pointers, total rebounds, steals, turn overs, personal fouls, field goals made, location of game and if the game is back to back.

The probability chi squared is significant because it is lower that our alpha=.05.

From all our predictors only field goals attempted, total rebounds, steals, turnovers, are the predictors who have a p-value < our alpha of .05.



Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-12.430672	17.980375	0.48	0.4893
fgmTeam	-0.8100538	0.4417111	3.36	0.0667
fgaTeam	0.57705004	0.2106047	7.51	0.0061*
pctFGTeam(accuracy)	36.6900423	38.479932	0.91	0.3403
fg3mTeam	0.0774641	0.2657955	0.08	0.7707
fg3aTeam	-0.1010599	0.1047782	0.93	0.3348
pctFG3Team	-11.429372	8.9594823	1.63	0.2021
trebTeam	-0.4003283	0.0385031	108.10	<.0001*
stlTeam	-0.384638	0.0589126	42.63	<.0001*
tovTeam	0.36793574	0.0478816	59.05	<.0001*
pfTeam	0.00296837	0.0350319	0.01	0.9325
isB2B[FALSE]	0.17092458	0.1404404	1.48	0.2236
locationGame[A]	0.13273216	0.1383251	0.92	0.3373

For log odds of 0/1

Covariance of Estimates

Effect Likelihood Ratio Tests

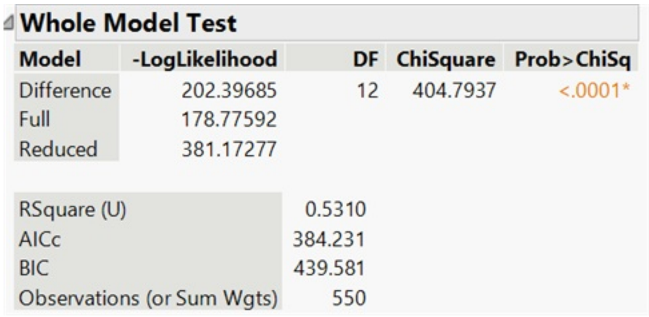
Confusion Matrix

Training

Actual	Predicted Count	
WIN/LOSS	0	1
0	231	40
1	43	236

Validation

Actual	Predicted Count	
WIN/LOSS	0	1
0	149	31
1	38	148



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	202.39685	12	404.7937	<.0001*
Full	178.77592			
Reduced	381.17277			

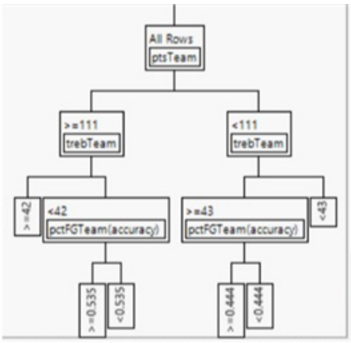
RSquare (U)	0.5310
AICc	384.231
BIC	439.581
Observations (or Sum Wgts)	550

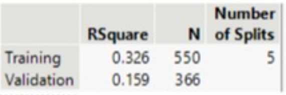
Final Project

Decision Tree categorical variable - **win or loss**

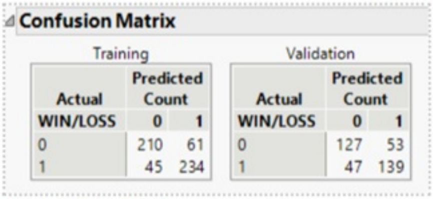
The decision tree model tried to help us understand the outcome of the game an L for the loosing or W for winning team. We used all the variables in the model resulting in an r-square of 32.6% for the training dataset and 15.9% for the validation dataset. The number of splits is 5.

The error rate for the validation dataset is 27.32% and for the training dataset is 23.87%. The sensitivity rate for validation dataset is 234/279= 83.87% and for the training dataset is 139/186=74.73%.





	RSquare	N	Number of Splits
Training	0.326	550	5
Validation	0.159	366	



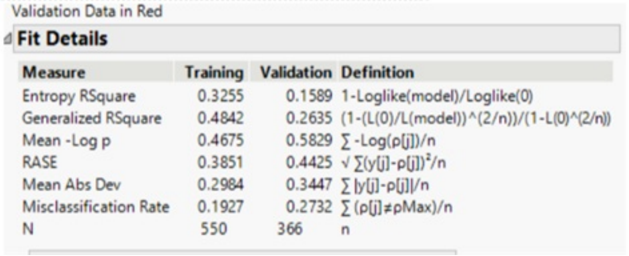
Confusion Matrix

Training

Actual	Predicted Count	
WIN/LOSS	0	1
0	210	61
1	45	234

Validation

Actual	Predicted Count	
WIN/LOSS	0	1
0	127	53
1	47	139

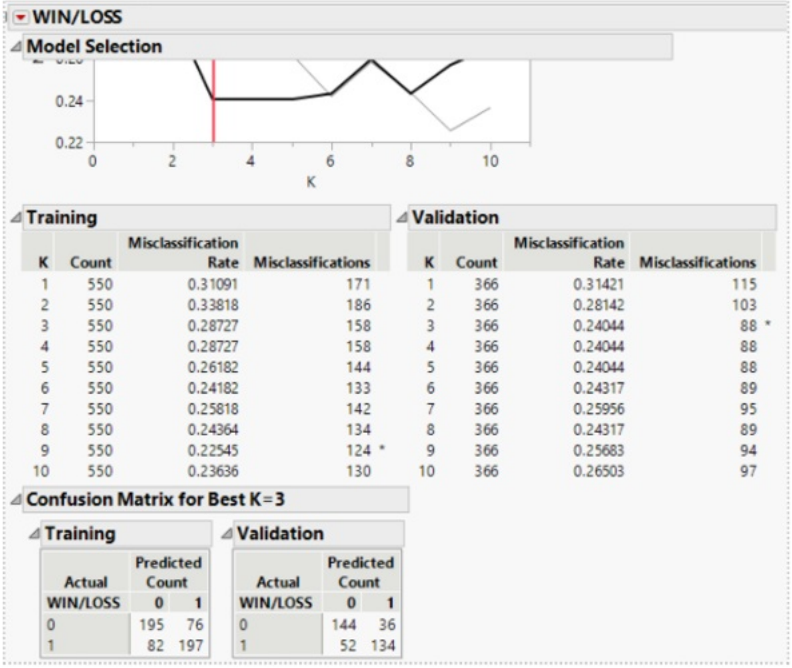


Validation Data in Red

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3255	0.1589	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4842	0.2635	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4675	0.5829	∑ -Log(ρ[j])/n
RASE	0.3851	0.4425	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2984	0.3447	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1927	0.2732	∑ (ρ[j]≠ρMax)/n
N	550	366	n

Final Project



WIN/LOSS

Model Selection

Training

K	Count	Misclassification Rate	Misclassifications
1	550	0.31091	171
2	550	0.33818	186
3	550	0.28727	158
4	550	0.28727	158
5	550	0.26182	144
6	550	0.24182	133
7	550	0.25818	142
8	550	0.24364	134
9	550	0.22545	124 *
10	550	0.23636	130

Validation

K	Count	Misclassification Rate	Misclassifications
1	366	0.31421	115
2	366	0.28142	103
3	366	0.24044	88 *
4	366	0.24044	88
5	366	0.24044	88
6	366	0.24317	89
7	366	0.25956	95
8	366	0.24317	89
9	366	0.25683	94
10	366	0.26503	97

Confusion Matrix for Best K=3

Training

Actual	Predicted Count	
WIN/LOSS	0	1
0	195	76
1	82	197

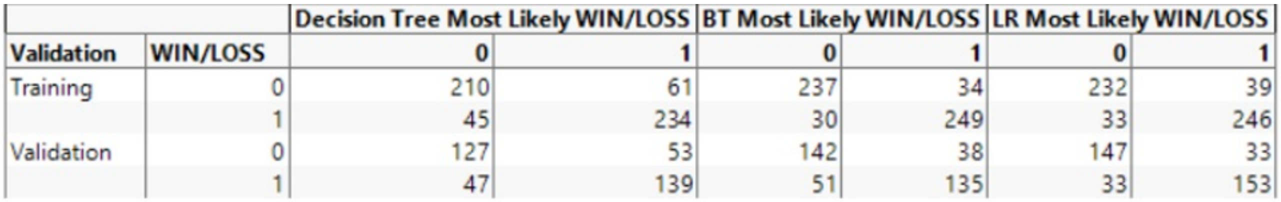
Validation

Actual	Predicted Count	
WIN/LOSS	0	1
0	144	36
1	52	134

K nearest neighbor categorical variable - **win or loss**

The K Nearest Neighbor model is helping us understand the outcome of the game by running ten different models. According to the analysis, k = 3 is the best model. The validation data set for model k = 3 had 88 misclassifications giving the model an error rate of 24%.

Final Project



		Decision Tree Most Likely WIN/LOSS		BT Most Likely WIN/LOSS		LR Most Likely WIN/LOSS	
Validation	WIN/LOSS	0	1	0	1	0	1
Training	0	210	61	237	34	232	39
	1	45	234	30	249	33	246
Validation	0	127	53	142	38	147	33
	1	47	139	51	135	33	153

The best model to predict the total points made in the game by an NBA Team is the decision tree because 1. it is simpler to interpret 2. the sensitivity rate in the decision tree model is 10% bigger than the k nearest model, where as the error rate difference from the decision tree and k nearest model is only around 3%.

Error rate for validation on the decision tree is 27.32%
Sensitivity for validation on the decision tree is 83.87%

Error rate for validation in the k nearest neighbor is 24%
Sensitivity for validation in the k nearest neighbor is 72%

Error rate for validation in the logistical regresion is 18.85%
Sensitivity for validation in the logistical regression is 82.25%

Final Project

Continous Analysis

How can we predict the number of points made by a basketball team in a single game?

Given the linear regression the best model to calculate the number of points made by a basketball team the significant predictors that will help us calculate the total points are field goals made, accuracy of field goals, 3pointers made, total rebounds, steals, turn overs, personal fouls. All of them are significant with a p-value < .05. The equation that will help us calculate the points scored in a game is 8.6557 +1.2831 (fgmTeam) + 59.8356 (pctFGTeam (accuracy)) + 0.7516 (fg3mTeam) + 0.2036 (trebTeam) +0.1597 (stlTeam) - 0.2604 (tovTeam) + 0.3268 (pfTeam). This can help coaches on how to improve the total points, the areas they need to focus on.

Categorical Analysis

Can the shots attempt, rebounds, steals, personal fouls, away/home game etc. Predict the outcome of the game (e.g., win/loss)?

Clearly the amount of points a team earns determines the outcome. However, we wanted to look at how other game statistics/ variables can predict game outcome. We ran three models, the partition tree model, K-Nearest Neighbors model, and a Nominal Linear Regression model. From our analysis we determined that the Nominal Regression Model was the best fit because of its low error rate and high sensitivity rate. The model determined that the following variables were good predictors for determining game outcome: Field Goals attempted ,Total rebounds , Total Steals , Total Turnovers

We can see that the above game statistics can help analysts determine game outcomes for teams giving a better perspective on team performance. In addition, coaches and scouts alike can focus on these statistics when scouting players and developing their teams.