DeSales University

SAS Enterprise Miner Final Report

NBA Statistics Report

By: Shane Artis, Luke Faro, Elijah Eberly, and Larry Agyei

Dr. Sedat Cevikparmak

May 1, 2023

Link to Dataset: https://www.kaggle.com/datasets/nathanlauga/nba-games?select=games.csv

Executive Summary (What did you do? Why does it matter? What did you find? What does it mean?)

For our project, we looked at different NBA statistics from the years 2004 to 2022. With this data, our objective was to try and find a model that will help teams accurately predict which teams will win based on different variables. This is significant to many teams, as it will give them information on how to adjust their current players, whether or not to draft new players, or adjust current players' play style in order to give their team the best advantage in a head-to-head match-up and put their team in contention for a championship. We ran multiple different types of tests (for example, 3 different types of regressions). The three different types were regular logistic, stepwise logistic, and backward logistic regression. In order to further explore our data, we repeated the process with KNN models, as well as decision trees.

Project Motivation/Background (Why is this topic important?)

The topic that we chose to look deeper into was the NBA games dataset. Using the data sets at our disposal, we attempted to try and predict which team has the best chance of winning based on statistics like the number of rebounds, steals, blocks, and made baskets while also examining wins in losses in a season. We thought this data would be able to give us a better understanding of how different teams match up within their conference. This information is useful to those who are interested in the different statistics that determine success, and allows people to help predict which team may be in contention for a championship. The information is also useful to individual teams, as it provides coaches with a more in-depth view of the contributions that players make to their team. These findings can help improve the game and

identify the strengths and weaknesses of players to provide a more competitive atmosphere for the fans.

Data Description

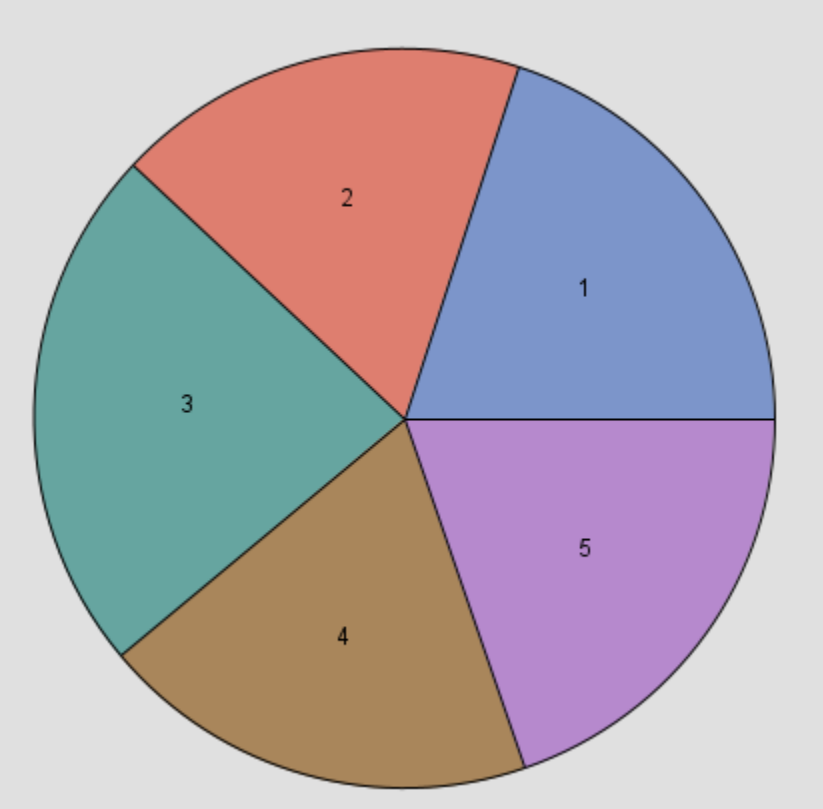
This dataset was collected from the NBA stats website. In this data set, there are five different Excel datasets that contain specific information about the NBA. The games.csv data set includes all of the games from the 2004 season all the way up to the 2022 season. It includes variables such as the date, the team, points for both home and away, assists both home and away, rebounds, etc. The next file that we have called games_details.csv, it includes a more in-depth view of the games.csv file. For example, the games_detail includes most of the details of the games dataset and all of the statistics of the players for a given game. In games_detail, we are able to see the players' name and the number of rebounds, steals, blocked shots, shots made, three-pointers made, etc, per game listed. Players.csv does not contain much information, only having four different variables: the players' first and last name in a column called player_name, the teamid, the playerid and the season they played in.

There are two more datasets that we have available for our use to conduct further analysis. They are the ranking.csv which includes information about the teams' standings, and the league that they play in (for example, the east or west), the season and the specific teams win-loss record, and the home and away records. Lastly, the teams.csv file contains the teams' teamid, the nickname of the team, the area they play in, the owner, as well as the head coach. Using these datasets at our disposal our data mining goal is to try and predict things like which team has the best chance of winning in a head-to-head matchup based on information obtained in

the games_detail.csv dataset like the number of rebounds, steals, blocked shots, percentage of made shots per player and the overall teams win and loss record.

Data preparation activities (i.e. noise, outliers, missing values etc.)

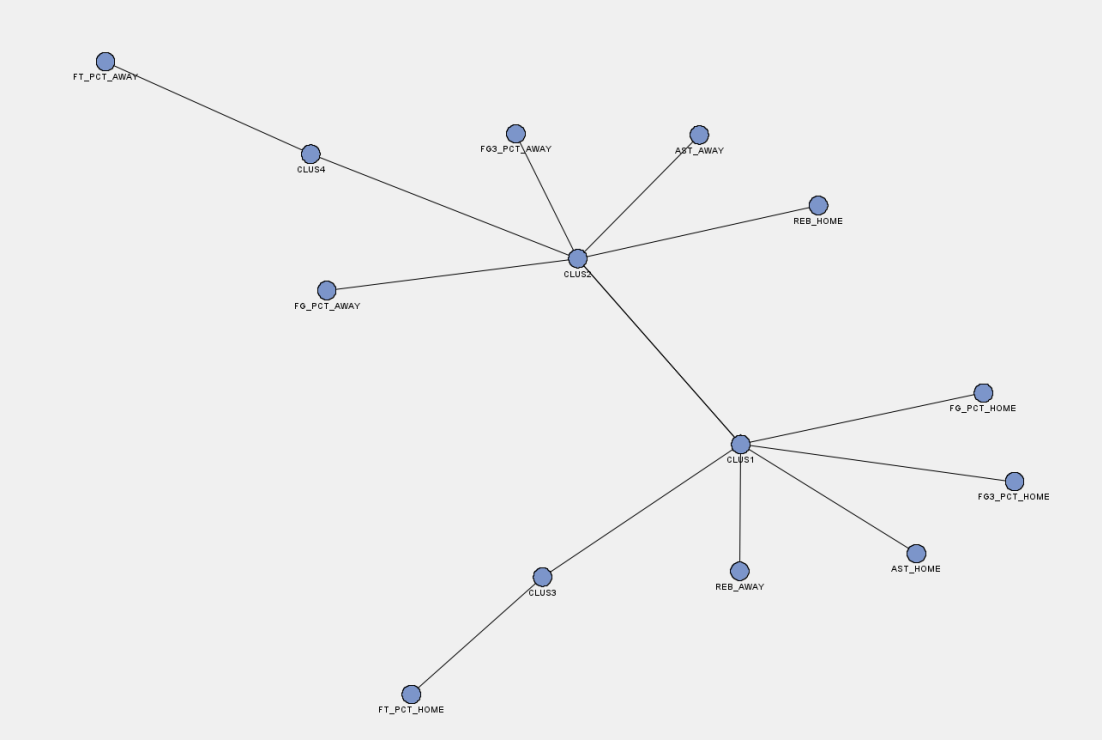
The first test that we chose to do was a cluster. After running the original cluster we ran another one except we set the final maximum number of clusters to create to 5. This was because the original cluster created twenty different ones which was too confusing and did not give us good information.



Mean Statistics

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	AST_away	AST_home	FG3_PCT_away	FG3_PCT_home	FG_PCT_away	FG_PCT_home	FT_PCT_away	FT_PCT_home	REB_away	REB_home
0.836201	0.011387	.	1	2042	0.84242	4.845343	5	2.509481	23.72233	27.18413	0.388788	0.44018	0.479804	0.519005	0.775688	0.787933	37.50294	40.53624
0.836201	0.011387	.	2	1821	0.812573	4.546519	5	2.118141	22.7117	25.53048	0.329512	0.368103	0.426713	0.459146	0.770767	0.705681	46.06864	47.85448
0.836201	0.011387	.	3	2370	0.840716	5.279148	1	2.530189	24.38819	20.32152	0.426779	0.314605	0.497575	0.437052	0.769797	0.765382	42.77131	39.27764
0.836201	0.011387	.	4	1940	0.845866	4.898198	2	2.27223	18.88144	18.21959	0.308087	0.263084	0.417702	0.402023	0.741118	0.759172	45.6366	46.5268
0.836201	0.011387	.	5	2007	0.836886	4.953419	2	2.118141	16.90583	23.14948	0.27345	0.401331	0.409605	0.487589	0.745478	0.782514	39.06378	44.23617

The clustering methods in the node perform disjoint cluster analysis on the basis of Euclidean distances. The results are computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. From the mean statistics above it does appear that the clustering method was effective as everything is grouped very similarly. There are not many outliers, as we expected given the data set and variables. Because we wanted to have a better visual representation of the clusters, we used the variable clustering node in order to identify how the different variables would be grouped together. This was an important step because this node helps with data reduction because it finds the best variables for analysis. Variable clustering removes collinearity, decreases variable redundancy, and helps reveal the underlying structure of the input variables in a data set.



For this project, we decided to focus on the games data set as one of the variables listed is if the home team won. This data is expressed by 1 resulting in a win and 0 resulting in a loss. Our goal is to have a model that has the best chance of correctly predicting the outcome of the games. For this to work we had to set our variables accordingly. Figure 1 below is our list of variables and indicates which variable we rejected and accepted.

Name	Role	Level
AST_away	Input	Interval
AST_home	Input	Interval
FG3_PCT_away	Input	Interval
FG3_PCT_home	Input	Interval
FG_PCT_away	Input	Interval
FG_PCT_home	Input	Interval
FT_PCT_away	Input	Interval
FT_PCT_home	Input	Interval
GAME_DATE_EST	Rejected	Interval
GAME_ID	Rejected	Nominal
GAME_STATUS_TEXT	Rejected	Nominal
HOME_TEAM_ID	Rejected	Nominal
HOME_TEAM_WINS	Target	Interval
PTS_away	Rejected	Interval
PTS_home	Rejected	Interval
REB_away	Input	Interval
REB_home	Input	Interval
SEASON	Rejected	Interval
TEAM_ID_away	Rejected	Interval
TEAM_ID_home	Rejected	Interval
VISITOR_TEAM_ID	Rejected	Nominal

Figure 1.

Home_team_wins is our target variable. We rejected all the variables that have no impact on a game's outcome like IDs, date, and text. We also had to reject both home and away points as they already answered the question of who won the game. The data we used consisted of over 25,000 different game outcomes, so we added a filter node to try and reduce the amount of data

and noise. We also ran a stat explore node to get a better sense of the data we were working with and the results can be seen in figure 2. below.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
AST	INPUT	2.276657	2.543779	81274	18726	0	2	24	1.713506	3.668508
BLK	INPUT	0.442196	0.797352	81274	18726	0	0	10	2.414958	8.218075
DREB	INPUT	3.172798	2.729215	81274	18726	0	3	21	1.255283	2.071077
FG3A	INPUT	3.236878	3.007763	81274	18726	0	3	22	1.042586	0.948942
FG3M	INPUT	1.159682	1.472011	81274	18726	0	1	12	1.584774	2.904668
FG3_PCT	INPUT	0.256892	0.290891	81274	18726	0	0.2	1	0.928964	0.044052
FGA	INPUT	8.164924	5.918276	81274	18726	0	7	37	0.867375	0.478916
FGM	INPUT	3.777604	3.175811	81274	18726	0	3	22	1.03024	0.981582
FG_PCT	INPUT	0.428783	0.253117	81274	18726	0	0.444	1	0.080959	-0.15455
FTA	INPUT	2.082671	2.753221	81274	18726	0	1	28	1.916667	4.944668
FTM	INPUT	1.615006	2.282961	81274	18726	0	1	23	2.084748	5.836113
FT_PCT	INPUT	0.423807	0.434191	81274	18726	0	0.5	1	0.230832	-1.7064
OREB	INPUT	0.932303	1.310615	81274	18726	0	0	13	2.014274	5.474772
PF	INPUT	1.860374	1.473122	81274	18726	0	2	6	0.548748	-0.40946
PLUS_MINUS	INPUT	-0.00324	11.3349	81274	18726	-56	0	54	0.118111	0.42006
PTS	INPUT	10.3299	8.602392	81274	18726	0	9	62	1.069451	1.152033
REB	INPUT	4.105101	3.420803	81274	18726	0	3	30	1.278377	2.179468
STL	INPUT	0.705219	0.943729	81274	18726	0	0	10	1.571711	3.017746

Figure 2.

We were able to identify some missing data from the games csv file. The file contained missing data points for pts home and away, field goal percentage home and away, three-pointers home and away as well as both assists and rebounds home and away. There were 18726 missing variables out of 81274 total data points. There are no major outliers that we were able to identify from our analysis of the raw data. We then added a data partition node where 40% of the data was used to train, 30% was allocated to the validate set, and the remaining 30% is the test set.

Model(s)/Enterprise Miner diagrams used

For the first test, we chose to do a normal logistic regression as our first model (shown in figure 3 below). Since our target variable is binary a logistic regression is used instead of a linear regression.

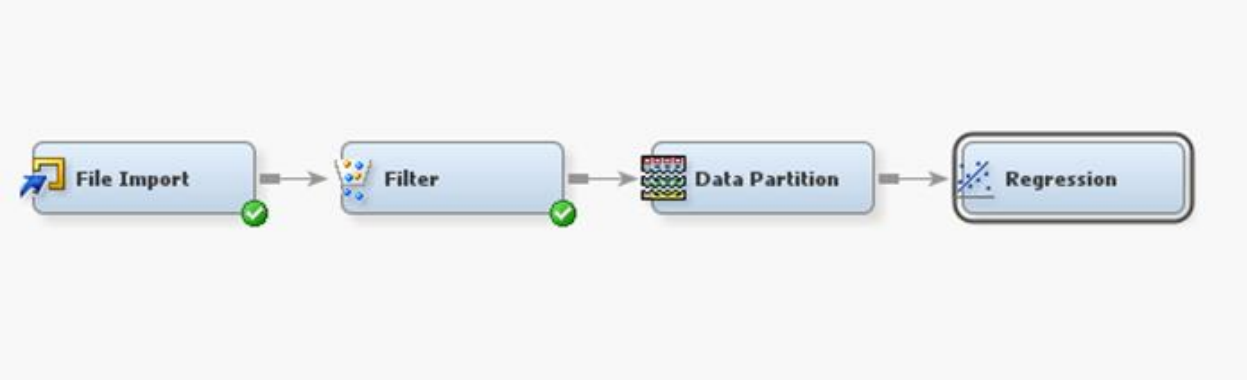


Figure 3.

The results of the linear regression test came back successful. Our R-Squared was 48.02% and Adjusted R-Squared value is 47.97%. As seen in figure 4 below. This indicates that at least 48% of the variance in the response variable can be explained by the explanatory variables.

Model Fit Statistics

R-Square	0.4802	Adj R-Sq	0.4797
AIC	-21114.9599	BIC	-21112.9361
SBC	-21035.4499	C(p)	11.0000

Figure 4.

In addition to calculating our R-Squared value we were also able to determine if the variables we selected are significant or not. To do this, we checked the statistics listed in the Analysis of Maximum Likelihood Estimates section letting us examine the p-values to determine the significance of each variable. Since we are satisfied with 95% confidence our alpha level is 0.05. So, if the p-value is less than our alpha level we can conclude that the variable is statistically significant. As seen in the figure below (Figure 5) we can come to the conclusion that all the variables are significant.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.6757	0.1020	6.62	<.0001
AST_away	1	-0.00827	0.000868	-9.53	<.0001
AST_home	1	0.00741	0.000886	8.36	<.0001
FG3_PCT_away	1	-0.5400	0.0378	-14.29	<.0001
FG3_PCT_home	1	0.5724	0.0374	15.32	<.0001
FG_PCT_away	1	-2.9505	0.0954	-30.91	<.0001
FG_PCT_home	1	2.7416	0.0954	28.75	<.0001
FT_PCT_away	1	-0.4789	0.0353	-13.58	<.0001
FT_PCT_home	1	0.4853	0.0364	13.32	<.0001
REB_away	1	-0.0105	0.000674	-15.54	<.0001
REB_home	1	0.00939	0.000675	13.90	<.0001

Figure 5.

Another key finding from our results is the mean square error (MSE). Since we partitioned our data into three groups (train, test, and validate) it is important that the MSE remains consistent among all the groups. Our training set had a MSE of 12.64%, the validate set's MSE was 12.77%, and the test set was 12.93%. From our MSE we can determine the

accuracy rate of the sets by subtracting it from 100%. The accuracy rate for all three sets is slightly more than 87%. This means that our data points are dispersed closely around the central mean and that the data is not very skewed, and has very few errors.

For our first test we are satisfied with the results but want to try and manipulate our models. The goal of our subsequent models wass to increase the R-Squared, Adjusted R-Squared, and the accuracy rate. Sticking with regression models, our next step was to try a stepwise regression. A stepwise regression works by starting with the most explanatory variable and adding more variables to the equation one at a time. With this we can determine the most significant variables to the model. We can also determine when to stop including additional variables as the effects of them are miniscule and not necessarily needed. This can help determine if all of our variables are truly significant.

After running the new regression, the first variable included was away team field goal percentage, followed by home team field goal percentage. After those two variables are included the R-Squared is 39.37% and the adjusted R-Squared is 39.36%. As we continue to add steps/variables to the model we begin to get closer to the same results we received from our normal regression model. We decided to stop adding variables at step 8 as the difference between step 8 and step 9 is less than half a percent, so we deemed the additional step unnecessary. As seen in Figure 6 below the last variable included in the model was the home team's free throw percentage. From the figure you can also see all the variables included up to this step. The R-Squared and adjusted R-Squared values are slightly lower than those from our original model, standing at 47.27% and 47.23%.

Step 8: Effect FT_PCT_home entered.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1160.825950	145.103244	1139.85	<.0001
Error	10171	1294.771692	0.127300		
Corrected Total	10179	2455.597642			

Model Fit Statistics

R-Square	0.4727	Adj R-Sq	0.4723
AIC	-20974.0826	BIC	-20972.3181
SBC	-20909.0290	C(p)	152.7554

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.6941	0.0954	7.28	<.0001
FG3_PCT_away	1	-0.5902	0.0377	-15.66	<.0001
FG3_PCT_home	1	0.6147	0.0372	16.52	<.0001
FG_PCT_away	1	-3.3027	0.0825	-40.02	<.0001
FG_PCT_home	1	3.0610	0.0809	37.86	<.0001
FT_PCT_away	1	-0.4837	0.0354	-13.66	<.0001
FT_PCT_home	1	0.4833	0.0366	13.20	<.0001
REB_away	1	-0.0107	0.000642	-16.70	<.0001
REB_home	1	0.00943	0.000638	14.79	<.0001

Figure 6.

We were willing to work with the slightly lower R-Squared values if we saw improvements in our MSE and accuracy rate. Unfortunately the MSE values for each set did not change resulting in the same accuracy rate as the original model. Because of this we concluded that the original model was better than the stepwise regression model.

Continuing our research using regression models, we decided to try a backwards regression in hopes to receive better results than the previous two models. The results from the backward regression were the exact same as the original regression model. Because of this we have decided that our best logistic regression is the original regression model.

The next test that we conducted was a decision tree. In an effort to increase readability, the maximum depth of our first decision tree was set to 4 for our first test. Everything else about the decision tree node, and the nodes used to prepare the data was left unchanged.

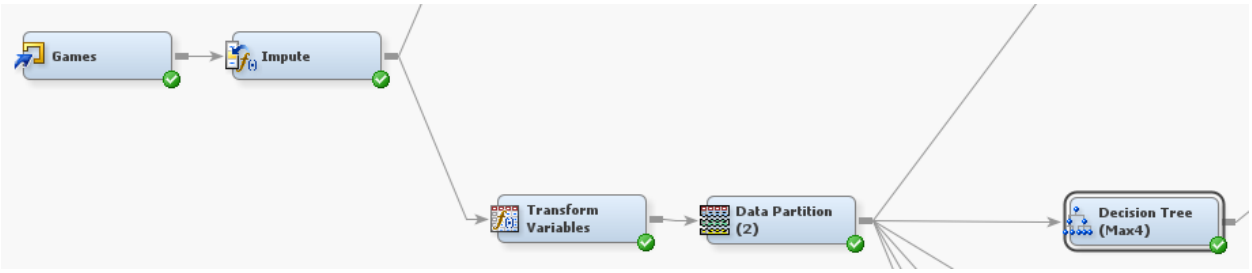


Figure 7.

After running the decision tree with the settings stated above, these were our results. With the maximum depth set to 4, our decision tree deemed that FG_PCT_away (The percentage of field goals the away team made) was the most important variable, being responsible for 8 of the splitting rules in the tree with the home percentage being a close second.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
FG_PCT_away		8	1.0000	1.0000	1.0000
FG_PCT_home		6	0.9889	0.9390	0.9495
REB_home		1	0.1394	0.1350	0.9686

Figure 8.

The decision tree returned an average square error (ASE) of 14% for the training set and 15% for both the validation and test set, as seen in Figure 9. This results in roughly a 85% accuracy rate. In its final testing phase, making it a fairly strong predictor of home team wins,

but not quite as strong as our linear regression model performed earlier which resulted in approximately 87% accuracy. In an effort to create a better model, we created additional decision trees with one tree being allowed to make 3 splits per decision and having a max depth of 6, and the other having a max depth of 4 while using the gini index to determine splits as opposed to the chi squared measurement for splitting. We then ran each model through a model comparison node to determine which method was the most successful. As shown below, the 3 split + depths 6 model yielded the best results.

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Tree3	Decision Tree (3Branch/Max6)	0.14677	0.12490
	Tree	Decision Tree (Max4)	0.15198	0.14199
	Tree2	Decision Tree (Gini/Max4)	0.15198	0.14199

Figure 9.

To round out our testing with the data, we also ran a MBR node (which is the same thing as a K-Nearest Neighbor (KNN) node). These are the results of our MBR/KNN model. It has been unaltered from SAS' default settings, and returned an ASE of 18%. in its final testing phase. It was a slightly worse predictor of home team wins compared to the decision tree and the linear regression model we've run so far.

Fit Statistics

Target=HOME_TEAM_WINS Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
_NW_	Number of Estimated Weights	10.00	.	.
NOBS	Sum of Frequencies	10632.00	7995.00	7996.00
SUMW	Sum of Case Weights Times Freq	10632.00	7995.00	7996.00
DFT	Total Degrees of Freedom	10632.00	.	.
DFM	Model Degrees of Freedom	10.00	.	.
DFE	Degrees of Freedom for Error	10622.00	.	.
ASE	Average Squared Error	0.16	0.18	0.18
RASE	Root Average Squared Error	0.40	0.43	0.42
DIV	Divisor for ASE	10632.00	7995.00	7996.00
SSE	Sum of Squared Errors	1699.85	1451.98	1411.54
MSE	Mean Squared Error	0.16	0.18	0.18
RMSE	Root Mean Squared Error	0.40	0.43	0.42
AVERR	Average Error Function	0.16	0.18	0.18
ERR	Error Function	1699.85	1451.98	1411.54
MAX	Maximum Absolute Error	0.94	1.00	1.00
FPE	Final Prediction Error	0.16	.	.
RFPE	Root Final Prediction Error	0.40	.	.
AIC	Akaike's Information Criterion	-19471.92	.	.
SBC	Schwarz's Bayesian Criterion	-19399.21	.	.

Figure 11.

As with our decision tree and linear regression models, we decided to create two more KNN models to try and yield a better model. For our additional models, one had its method set to scan, and the other had its maximum neighbors set to 20. We then set each model to run through the model assessment node, which yielded the following results:

Fit Statistics

Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	MBR3	MBR (Nearest 20)	0.17963	0.16210
	MBR2	MBR Scan	0.18152	0.15941
	MBR	MBR Default	0.18159	0.15990

Figure 12.

Increasing the number of neighbors yielded marginally better results when using the base model, and the scan method model had a very slightly higher accuracy rate.

Findings (What do your models and other analysis tell you?)

Throughout this entire process we have assumed the normal regression model has been the best, but we wanted to make sure. In doing so we included all our models in a model comparison node as shown in Figure 11. Like we assumed the original regression model was the best and also picked by SAS. Figure 12 below shows the average square error for all the models we have performed. Unfortunately none of our models were able to achieve less than 10 for their ASE, which we were hoping for. We believe this could be because of the amount of missing data in our dataset. We would love to be able to find a more complete set and see if that will give us a more confident result. The K-Nearest Neighbor model was the worst of all models for both the validate and train set. This was not much of a surprise to us as we expected it to be one of the weaker models we tried. Like stated earlier all three of the regression models maintained the same average mean square error. However, the original regression model had the highest R-Squared and adjusted R-Squared value resulting in it being the preferred model.

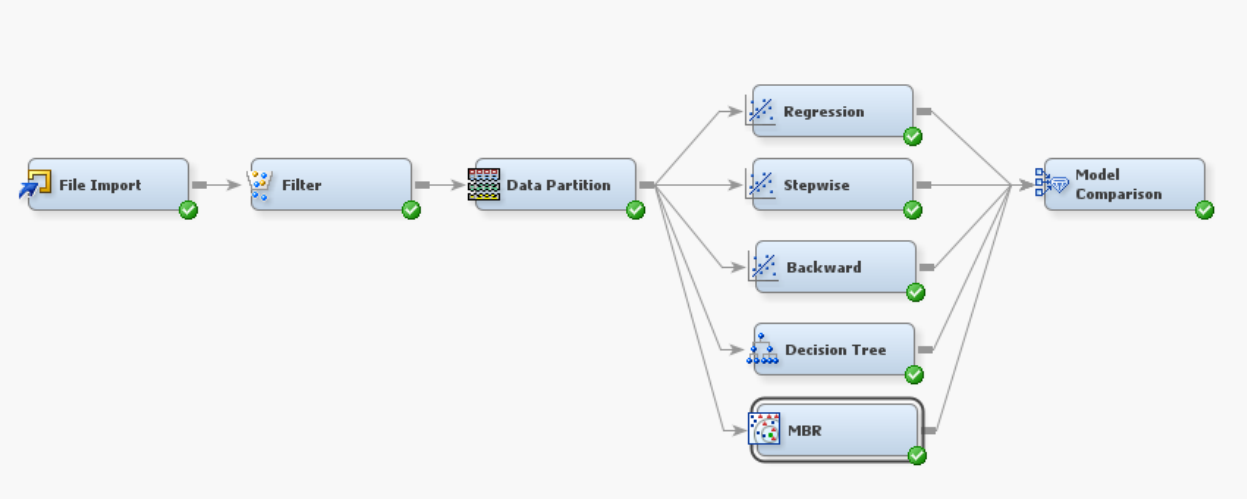


Figure 13.

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Reg	Regression	0.12766	0.12624
	Reg2	Stepwise	0.12766	0.12624
	Reg3	Backward	0.12766	0.12624
	Tree	Decision Tree	0.15133	0.14314
	MBR	MBR	0.17890	0.15995

Figure 14.

Managerial implications/conclusions (How should business operations be changed?)

As mentioned previously, the managerial implications of these statistics are useful for those interested in the different factors that determine success, and predict which team may be in contention for a championship. The information is also useful to individual teams, as it provides coaches with a more in-depth view of the contributions that players make to their teams. These findings can help improve the game and identify the strengths and weaknesses of players to provide a more competitive atmosphere for the fans and coaches so they know what to work on with their players and what they need to continue to do right. There isn't a whole lot of change that can be added right now in the NBA many of these teams already have their own data scientist who have advanced cameras and technology to get better stats. The main thing is that we need to continue running and finding different results so that we are able to compare models. Another suggestion is to continue working on more advanced technology so that we can be able to report the statistics of players and games quicker to coaches so they can make adjustments to

their lineups in-game instead of having to wait till the end of the game to see how their game style worked.