SAS Final Project

BY: Luke Faro, Larry Agyei, Elijah Eberly, Shane Artis

Project Background

- Group Interest in NBA statistics
- Obtained NBA data from Kaggle for the years 2004-2022
- Used the games.csv file to accurately predict team wins
- Data can be used to help coaches, players, and spectators

Data Description

- NBA games.csv file
- Data contained games from 2004 - 2022
- Some missing data, no outliers
- Used a filter and data partition node to reduce the noise
- Variables include:
  - date, team, points both home and away
  - assists and rebounds home and away

Data Prep

- Stat Explore Node
  - Gave insight on missing data (18726 out of 81274)
  - Mean, Skewness, and Standard Deviation
- Cluster / Variable Cluster
  - Set final maximum to 5
  - Helps remove collinearity, redundancy
  - Able to see how variables are grouped
-

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

Model Tests

- Logistic Regression
- Using Home_Team_Wins as Target Variable
- We used four different nodes
  - First the file import node
  - Filter node
  - Data Partition node
  - Regression node
- Resulted in a R-squared of 48.02%
- All variables were significant as they are less than .05

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

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

Model Tests

- Stepwise logistic regression
  - Starts with most significant adding more variables each time
- The most significant variables was field goal percentage home and away
- Field goal pct home and away resulted R-squared of 39.37%
- Stopped at step 8 as no more significant improvement
- Resulted in overall R-squared of 47.27%

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

Model Tests

- Backwards Logistic Regression
    - Removes least important variables and leaves the most important variables
- Results were the same as original model
- R-squared of 48.02%
- All variables were significant as they are less than .05

Model Fit Statistics

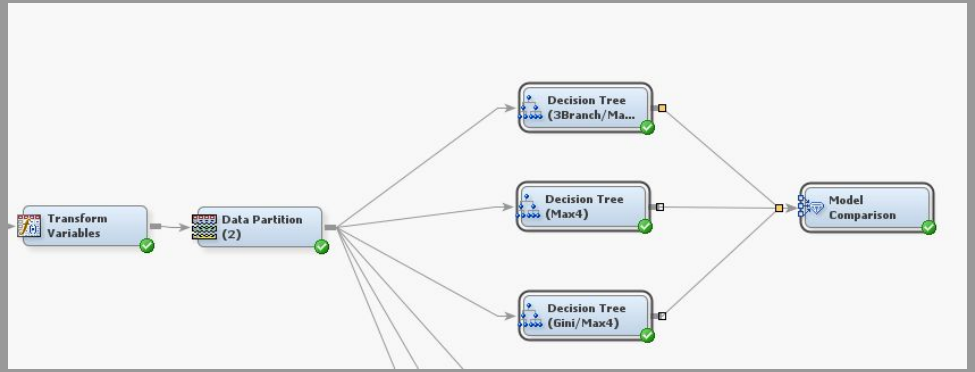
R-Square	0.4802	Adj R-Sq	0.4797
AIC	-21114.9599	BIC	-21112.9361
SBC	-21035.4499	C(p)	11.0000

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

Model Tests

- Decision Tree
- Set Maximum depth to 4 everything else left unchanged for first pass
- Used import file, transform variables, data partition and then a decision tree
- Field goal percentage away was most important followed by field goal percentage home
- Average Square Error (ASE) of 15% resulting in 85% accuracy rate



Fit Statistics

Target=HOME_TEAM_WINS Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	10660.00	7995.00	7996.00
MAX	Maximum Absolute Error	0.98	0.98	0.98
SSE	Sum of Squared Errors	1509.61	1206.92	1160.18
ASE	Average Squared Error	0.14	0.15	0.15
RASE	Root Average Squared Error	0.38	0.39	0.38
DIV	Divisor for ASE	10660.00	7995.00	7996.00
DFT	Total Degrees of Freedom	10660.00	.	.

Model Tests

- Decision Tree
- Set 3 splits per decision and having a max depth of 6 yielded improved accuracy
- Changing the split method from Chi Squared to Gini slightly decreased accuracy

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

Model Tests

- MBR / K-Nearest Neighbor
- Using default settings on first pass
- Returned an Average Squared Error of 18%
- Slightly worse predictor of home team wins when compared to decision tree and linear regression models
    - Even with a refined model

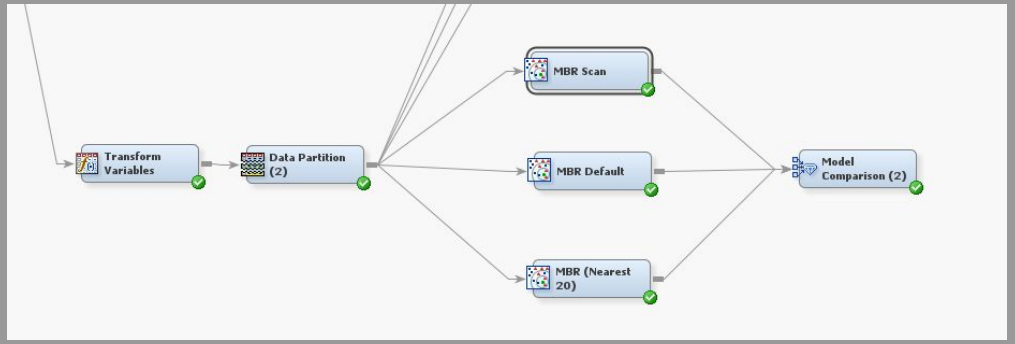
Fit Statistics

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	.	.

Model Tests

- MBR / K-Nearest Neighbor
- Increasing the number of neighbors collected improved accuracy
- Changing the input method to Scan yielded marginally better 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

Results

- Used a model comparison with all models
- Original logistic regression model worked best
- No models achieved less than 10 for ASE
- Possibly due to missing data
- MBR / KNN performed the worst
- Found that Field goal percentage was most important for a team winning the game
- Need to shoot more shoots

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

Results

- Our findings can be used for determine success and predicting possible championship teams
- Help coaches identify strengths and weaknesses
- Many teams now have data scientist to help because they have seen the benefits
    - Example: not directly related to basketball but the movie Moneyball

Q & A