Final Project - Kobe Shot Prediction

October 7th, 2017

By Leonardo Deartiaga & Brian Goff

Objective of Project

The goal of the project is to use data collected from Kobe Bryant's 20 year professional career in the NBA to build two different models that can be used to accurately predict if Kobe would have made a shot or not.

Data Preparation Performed

Prior to importing the data into SAS, we reviewed the data set for completeness. Upon review we found 5,000 rows missing a value for our target variable, shot_made_flag. Since this variable is critical to creating a model and our ability to asses the accuracy we removed all rows missing the shot_made_flag variable. This left us with 25,697 rows of data to use in training and validating our models.

In addition to determining missing data, we reviewed the dataset to determine which columns in the dataset were used as identifiers and could be rejected upon import.

Exploration of Variables

Data Exploration Technique:

The data exploration was completed outside of SAS. We used R, and R Studio with the tidyverse library for data manipulation and graphing. This provided us a quick way to clean, summarize, manipulate, and graph the data before moving over to SAS to run our models.

Variables analyzed:

Variable: Action_Type: The action type variable has 57 unique categorical variables describing the specific action taken during a shot.

The charts below show some of the distribution of the 57 action_type variables.

Top 10:

	Var1	n
1	Jump Shot	15836
2	Layup Shot	2154
3	Driving Layup Shot	1628
4	Turnaround Jump Shot	891
5	Fadeaway Jump Shot	872
6	Running Jump Shot	779
7	Pullup Jump shot	402
8	Turnaround Fadeaway shot	366
9	Slam Dunk Shot	334
10	Reverse Layup Shot	333

High:

Jump Shot	15836

Low:

Cutting Finger Roll Layup Shot	0
Turnaround Fadeaway Bank Jump Shot	0

I wonder if it can be re-categorized: Jumpshot or not (0) (1)

The distribution of the top 6 action_type variables shown in the graph below:

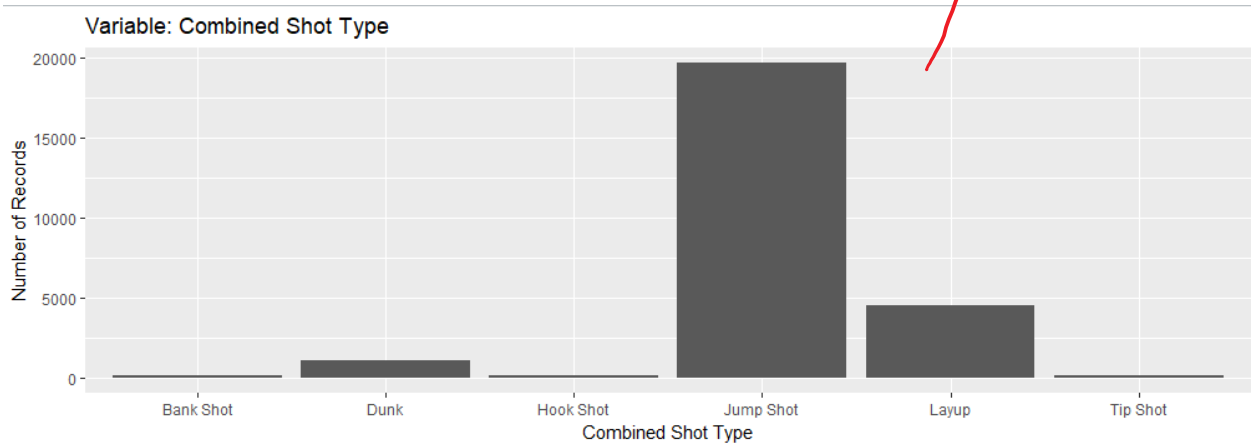
Variable: Combined_type_shot: The combined type shot variable has 6 unique categorical variables describing a shot type.

The table below shows the distribution of the combined_type_shot variable.

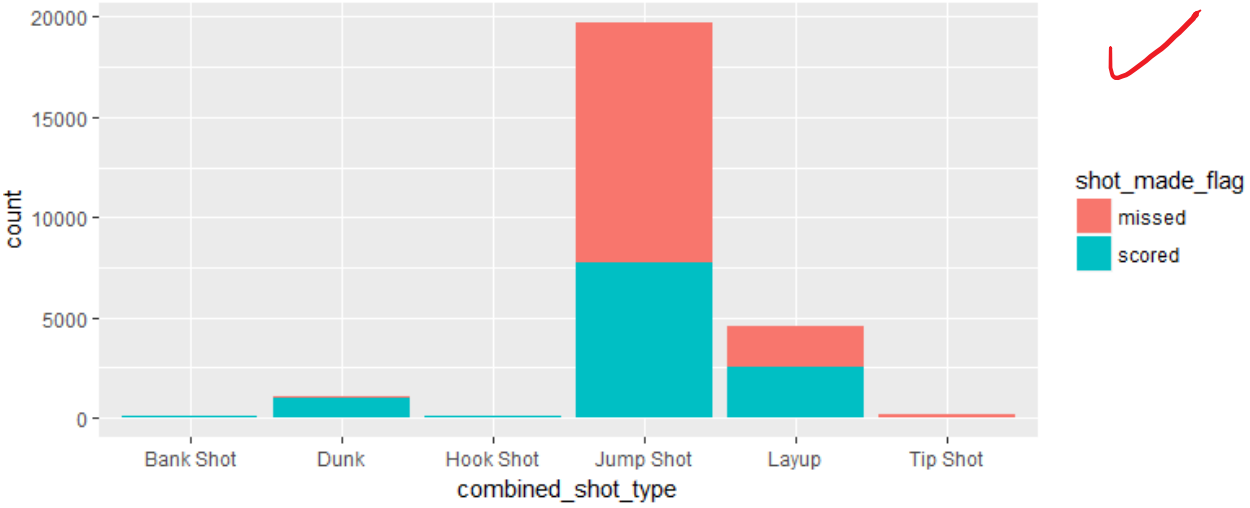
Jump Shot	19710
Layup	4532
Dunk	1056
Tip Shot	152
Hook Shot	127
Bank Shot	120

Maybe 3 categories?
1) Jumpshot
2) Dunk & layup
3) Other

The graph below shows the combined_shot_type variable distribution:



The graph below shows the combined_shot_type variable distribution split with shots made/missed.



Variables: lat,lon,loc_x,loc_y: These variables represent the location on the court where a shot was taken.

Lat statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
33.25	33.88	33.97	33.95	34.04	34.09	0.0881521

Lon statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
-118.5	-118.3	-118.3	-118.3	-118.2	-118.0	0.1100731

Loc_x statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
-250.000	-67.000	0.000	7.148	94.000	248.000	110.0731

Loc_y statistical summary:

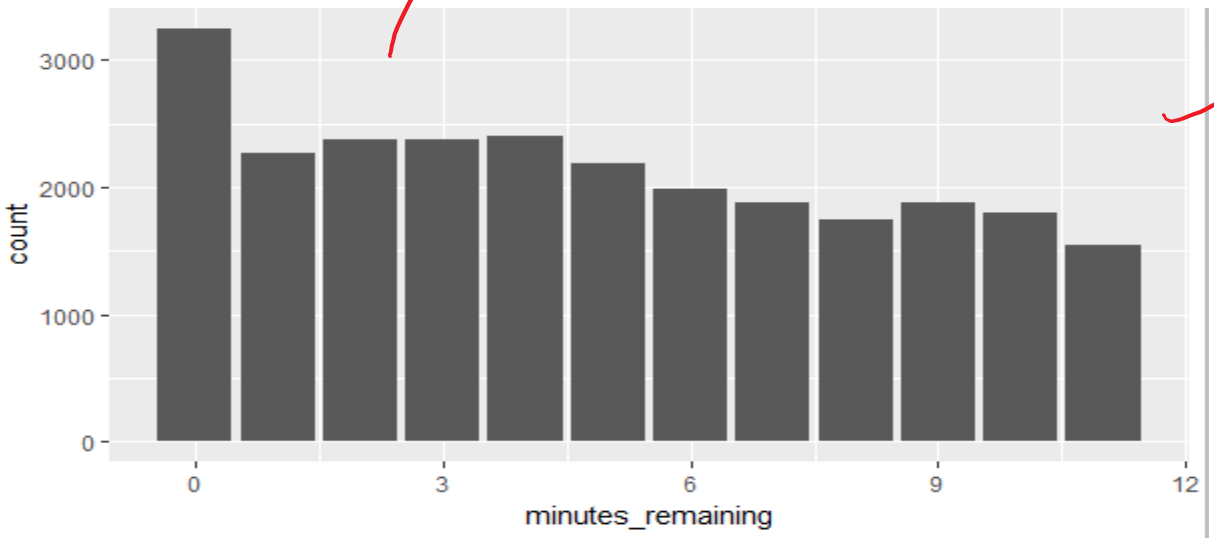
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
-44.00	4.00	74.00	91.26	160.00	791.00	88.15211

Variable: minutes_remaining: The minutes remaining variable is the amount of minutes left of a given period that a shot was taken.

Minutes Remaining statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
0.00	2.000	5.000	4.887	8.000	11.00	3.452475

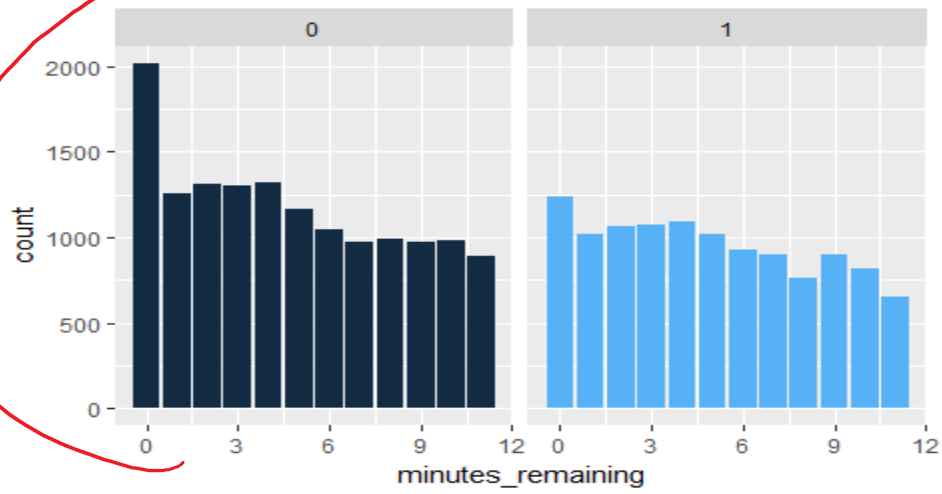
The graph below shows the distribution for attempted shots taken per minute remaining:



The graph below shows the distribution for shots missed, and shots made per minutes remaining:

Shot Missed: *0*

Shot Made: *1*

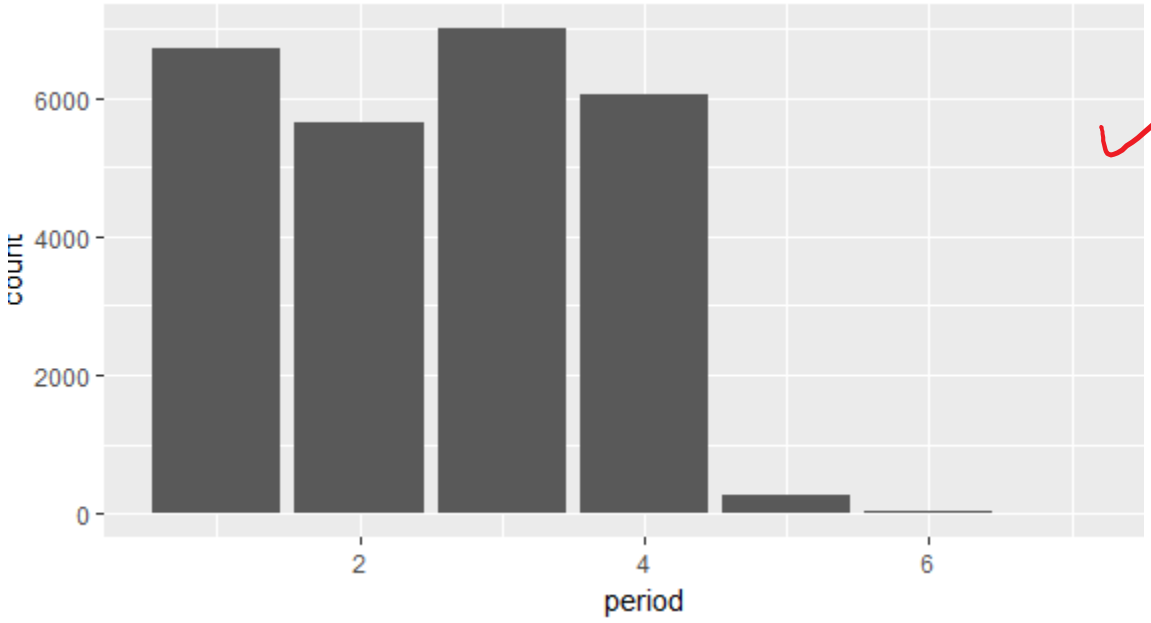


Variable: Period: There are 4 periods of a NBA basketball game. The max for the variable is 7. This most likely means that some games played went into overtime.

Period statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
1.000	1.000	3.000	2.521	3.000	7.000	1.151626

The graph below shows the distribution for the shots attempted during each period.



Variable: Playoffs : The playoff variable is a binary variable indicating if the shot was taken during a playoff game or not.

Regular vs Playoff game data distribution:

Regular Season games:	21939
Playoff games:	3758

0 = *Regular Season Game & Shot Missed*
1= *Playoff Game & Shot Made*

Below is a cross table between the playoff variable, and the shot_made_flag variable. The cell contents are listed below:

Cell Contents
N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

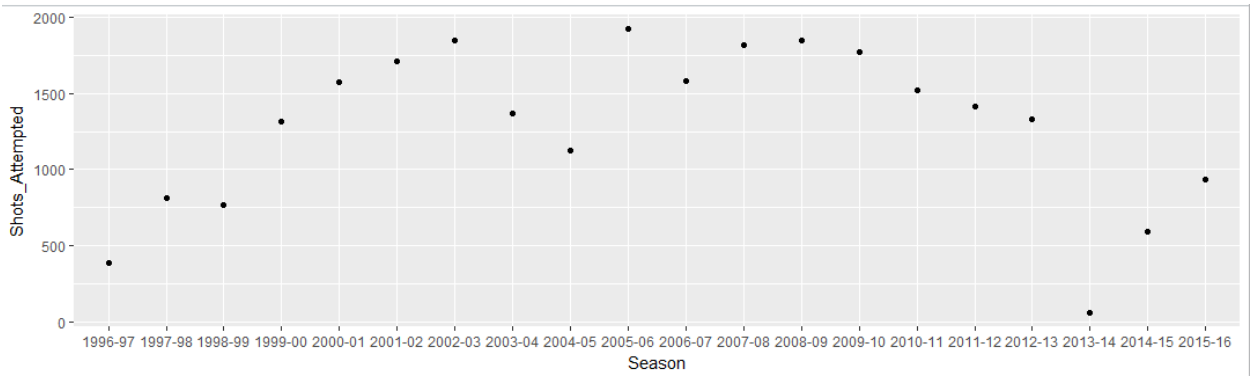
Total Observations in Table: 25697

playoffs	shot_made_flag 0	1	Row Total
0	12145	9794	21939
	0.003	0.003	
	0.554	0.446	0.854
	0.853	0.854	
	0.473	0.381	
1	2087	1671	3758
	0.015	0.019	
	0.555	0.445	0.146
	0.147	0.146	
	0.081	0.065	
Column Total	14232	11465	25697
	0.554	0.446	

Variable : Season:

Kobe played 20 season in the NBA, below is a scatterplot showing the amount of shots attempted in a given season. Notice, the 2013-2014 season where Kobe had a significant amount less than the other seasons. This is because he only played 6 games that year due to knee injury.

Shots attempting in a given season:

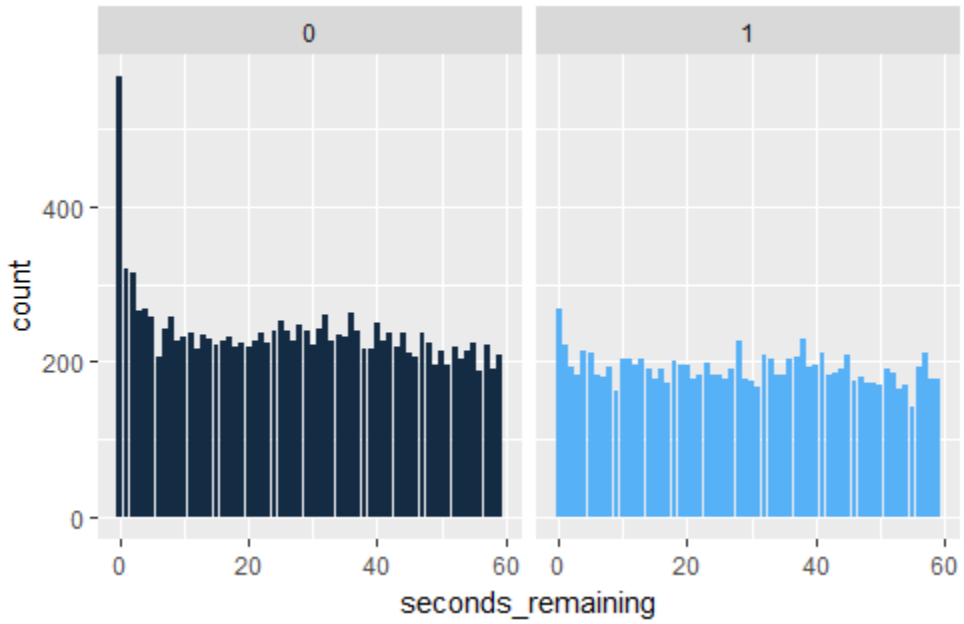


Variable : Seconds Remaining: The seconds remaining variable is the amount of seconds left of a given period that a shot was taken.

Seconds Remaining statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
0.00	13.00	28.00	28.31	43.31	59.00	17.52

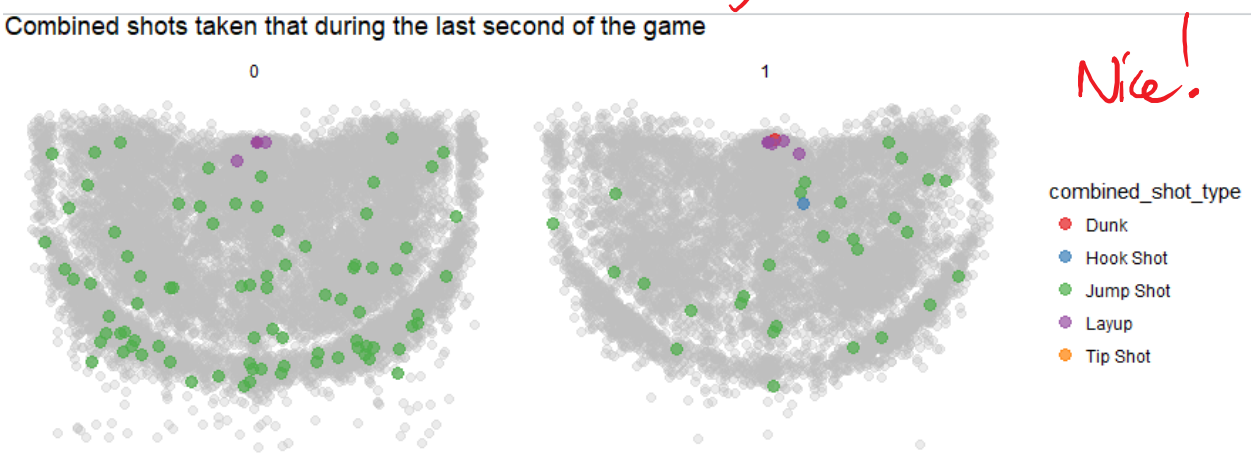
The graph below shows the distribution for shots missed, and shots made per minute remaining:
Shot Missed: 0
Shot Made: 1



I was interested in exploring the data for shots taken during the last second of every game. The graph and table below show the combined shot type, and distribution for the final second.

***0** = Missed*

***1** = Made*



Distribution of shot type for the last second of the game.

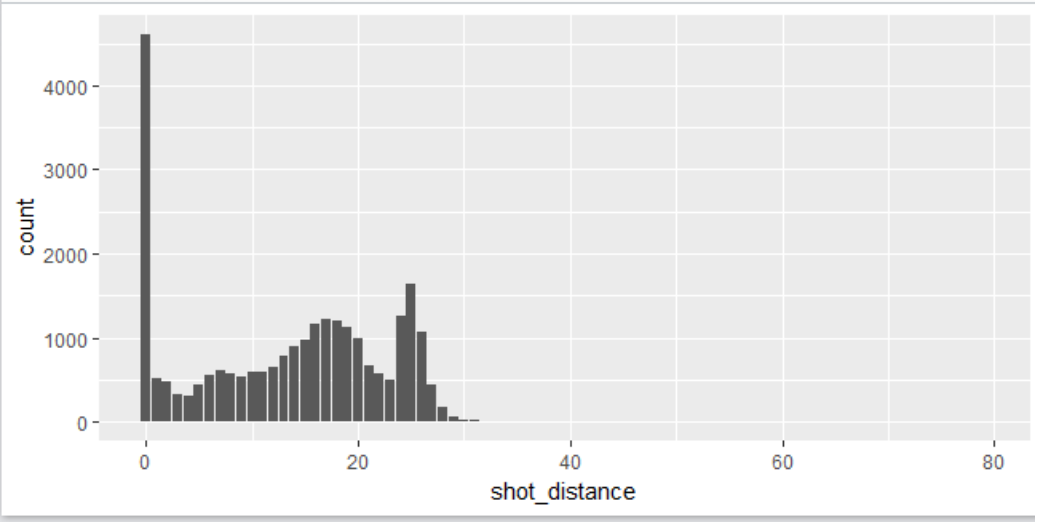
Shot Type	Shot Missed	Shot Made	Sum
Bank Shot	0	0	0
Dunk	1	5	6
Hook Shot	0	1	1
Jump Shot	90	33	123
Layup	90	33	123
Tip Shot	2	1	3
Sum	100	54	154

Variable: Shot Distance: The shot distance variable is the distance (measurement unknown) of a given shot from the basket.

Shot distance statistical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
0.00	5.00	15.00	13.46	21.00	79.00	9.38

The graph below shows the distribution of the shots taken by distance.



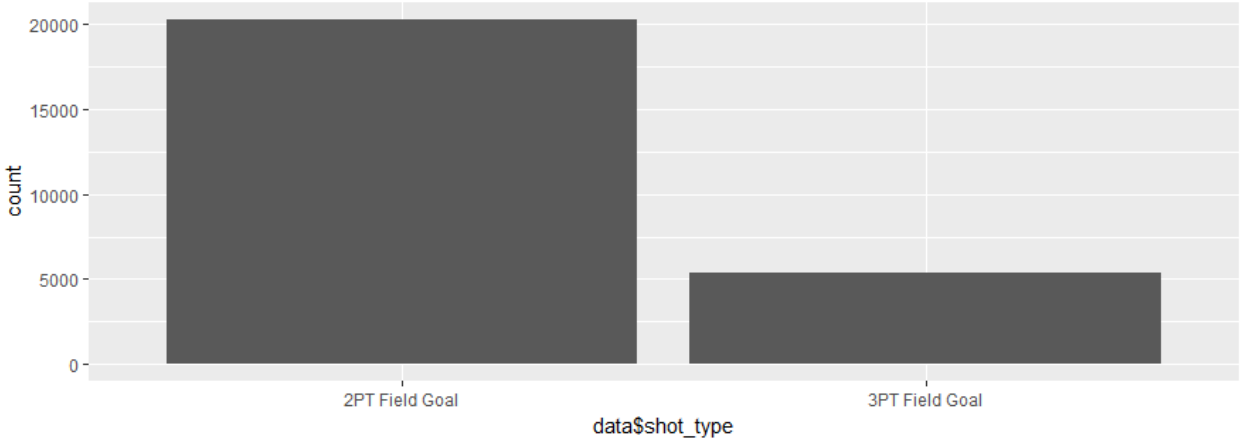
Variable: Shot Made Flag: The shot made flag is our binary target variable. Our objective is to predict if a shot will be made or missed using the other variables listed in this report as predictors.

There is a total of 25697 shots. Broken down by shots missed, and shots made as seen in the below chart:

0 = Missed	**1 = Made**	**Sum**
14232	11465	25697

Variable: Shot Type: The shot type variable is a categorical variable describing if a shot was worth 2 points, or 3 points.

The graph below shows the distribution of the shot_type variable.

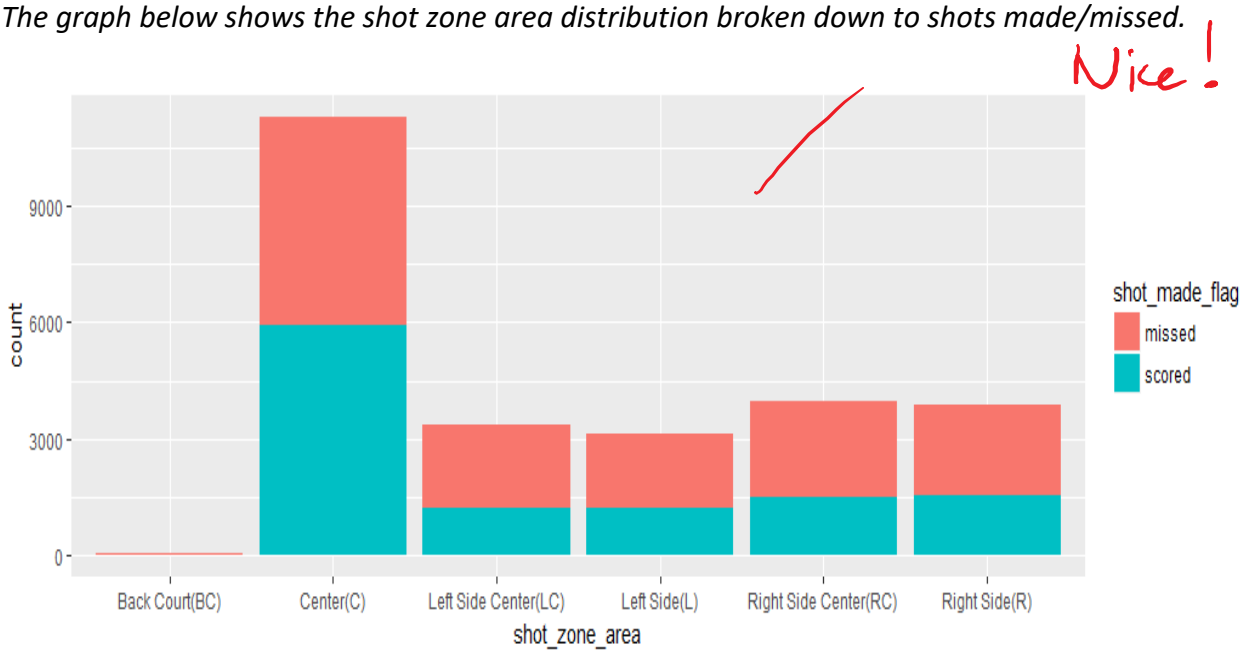


Variable:Shot Zone Area: The shot zone area variable is a categorical variable that describes where the shot was taken. There are 6 unique categories as seen below.

Shot zone area distribution:

Categories	**Distribution**
Back Court (BC)	72
Center (C)	11289
Left Side Center (LC)	3364
Left Side (L)	3132
Right Side Center (RC)	3981
Right Side (R)	3859

The graph below shows the shot zone area distribution broken down to shots made/missed.

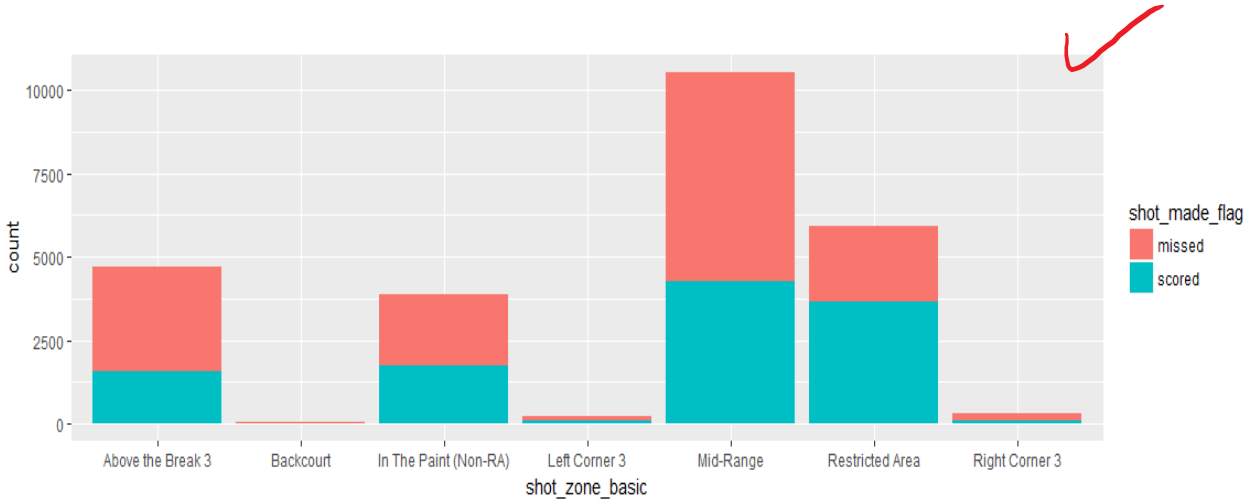


Variable:Shot Zone Basic: The shot zone basic variable is a categorical variable that describes the shot locations of the court. The distribution of the variable is show in the table below.

Shot zone basic distribution:

Categories	**Distributions**
Above the Break 3	4720
Back Court	60
In The Paint (Non-RA)	3880
Left Corner 3	240
Mid-Range	10532
Restricted Area	5932
Right Corner 3	333

The graph below shows the shot zone basic distribution broken down to shots made/missed.

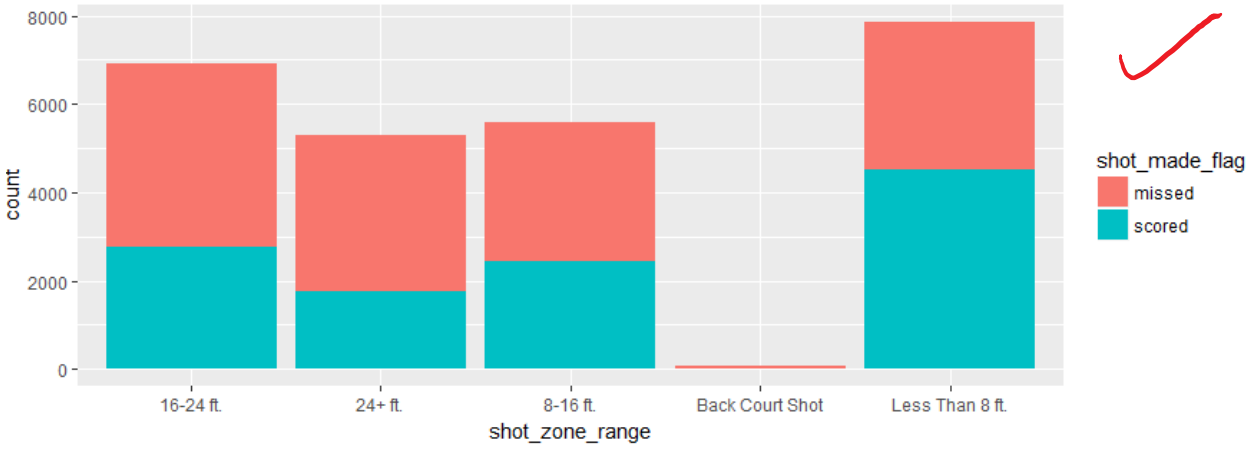


Variable: Shot Zone Range: The shot zone range is a categorical variable describing the range of a given shot from the basket. There are 5 unique categories and their distribution is shown below.

Shot zone range distribution:

Categories	**Distribution**
16 - 24 ft.	6907
24 + ft.	5281
8-16 ft.	5580
Back Court Shot	72
Less than 8 ft.	7857

The graph below shows the shot zone range distribution broken down to shots made/missed.



Variable: Opponent: The opponent variable is a categorical variable describing the name of the opponent played against.
The distribution of shots taken against each opponent is shown in the table below.

*Note, this is 10/66 rows.

The distribution of shots taken against each opponent:

Opponent	Shot_Result	Freq
ATL	missed	240
ATL	scored	198
BKN	missed	27
BKN	scored	18
BOS	missed	461
BOS	scored	322
CHA	missed	282
CHA	scored	218
CHI	missed	294
CHI	scored	222

The following variables were not analyzed for this section of the report :

Game_event_id: unique id for game event

Game_id: unique id for the game

Team_id: unique id for the Lakers

Shot_id: unique id for each shot

Team_name: variable for the Los Angeles Lakers. Kobe only played for one team during his career.

Game_date: variable that describes the date of game.

Matchup: variable for the teams who participated in the game.Kobe only played for the LA Lakers, and the opponent variable represents the opposing team. This variable is not needed.

Discussion on Data Problems

The dataset obtained was fairly clean. We ran into an issue with our target variable, but we were able to identify the missing values, and remove the corresponding rows from the data before importing the dataset into SAS.

Algorithm One - Neural Network

Description of Model

The Neural Network Model was able to predict if Kobe Bryant would make a shot with an accuracy rate of 68.66%. The baseline accuracy rate for the data set is 55.38%. This means our Neural Network Model outperforms the baseline model by 13.28%. The model was created performing the following steps:

After importing the data using the file import node, the drop node was then used to test eliminating various variables from the dataset to determine their impact on the misclassification rate. The dataset was then partitioned to provide us the ability to validate the models created. Next, the data was cleaned to remove any outliers in the filter node. The training dataset was then used to create an algorithm using the Neural Network node. The Neural network was built using 13 nodes in the hidden layer.

Nodes and Settings Used

1. File Import - advanced advisor set to "yes"
2. Drop Node - dropped variable "playoffs" & "Combined_shot_type"
3. Data Partition - Default Settings
4. Filter Node - Default Settings
5. Neural Network - Preliminary Training to "no", # of nodes set to 13, model selection set to "misclassification"

Variables Included

1. action_type
2. Lat
3. Lon
4. Loc_x
5. Loc_y
6. Minutes_remaining
7. Period
8. Seconds_remaining
9. Shot_distance
10. Season
11. Shot_type
12. Shot_zone_area
13. Shot_zone_range
14. shot _zone_basic

See appendix for variables roles and levels.

Misclassification Rate

The Neural Network Model had a misclassification rate of 0.3134 or 31.34%. The baseline misclassification rate for the data set is 0.4462 or 44.62%. The baseline misclassification rate was calculated using the naive model, which involved taking the value of the minority outcome and dividing it by the total number of observations.

This means our Neural Network Model outperforms the baseline model by 13.28%.

Algorithm Two - Stepwise Regression

Description of Model

The Stepwise Regression Model was able to predict if Kobe Bryant would make a shot with an accuracy rate of 68.63%. The baseline accuracy rate for the data set is 55.38%. This means our Stepwise Regression Model outperforms the baseline model by 13.25%. The model was created performing the following steps:

After importing the data using the file import node, The dataset was then partitioned to provide us the ability to validate the models created. The drop node was used to drop several variables that was not used in our model. The data was then cleaned to remove any outliers in the dataset by using the filter node. The data was run through the regression node with the selection model set to stepwise.

Nodes and Settings Used

1. File Import - advanced advisor set to "yes"
2. Data Partition - Default
3. Drop: The following variables were dropped:
    - Combined_shot, game_date, game_event_id, lat, loc_x, loc_y, lon, matchup, minutes remaining, period, playoffs, seconds_remaining, shot_id, shot_zone_basic, team_id, team_name, VAR1
4. Filter - default
5. Regression - Regression Type set to "linear Regression" ,
   Selection Model set to "Stepwise" ,
   Selection Criterion set to "Validation Misclassification"

Variables Included

The stepwise regression model selected the following variables for prediction:
Step 1: action_type
Step 2: shot_zone_range
Step 3: shot_zone_area

Misclassification Rate

The Stepwise Regression Model had a misclassification rate of 31.37%. The baseline misclassification rate for the data set is 0.4462 or 44.62%. The baseline misclassification rate was calculated using the naive model, which involved taking the value of the minority outcome and dividing it by the total number of observations.

This means our Stepwise Regression Model outperforms the baseline model by 13.25%.

What Was Learned

Stepwise Model:

The stepwise technique used for the regression model reduced the number of variables to 3. This gave us a clear understanding on what predictors would be useful, and not useful in predicting the target variable. We learned that the specific type of shot, the range from which Kobe was shooting from, and where he was located on the court were the best predictors for him making a shot or not. We also learned what were not strong predictors . For example, who he played against, or what year in his career he was in, didn't make much of a difference in predicting the target variable.

Neural Network Model:

Unlike the Stepwise Model, the Neural Network Model performed better when more information was fed into it. Much like the human brain it is trying to emulate, the more information lead to a better informed algorithm. The difference between the two models is insignificant and the stepwise model would likely be easier to use in a real world scenario since it requires less variables to make an accurate prediction than the neural network model.

Efforts to Minimize the Misclassification Rate

Stepwise Model Efforts:

The use of domain knowledge gained after performing the data exploration helped in deciding which variables would be dropped before running our data through a model. For example, there were a few variables that were almost duplicates of each other. Like the *action_type*, and *combined_type_shot* variables. The *action_type* variable is a more granular version of the *combined_type_shot* variable having 57 unique categories versus 6 unique variables that the *combined_type_shot* had. These two variables both described the type of shot being taken. Therefore, it was decided that *action_type* variable would be used instead of the *combined_shot_type* variable because of the level of detail the variable was able to provide.

Nice!

After the use of our domain knowledge the data was then reduced even further by using the stepwise technique available in SAS.

In the end, only 3 variables were selected by the stepwise regression model which were *action_type, shot_zone_range*, and *shot_zone_area*.

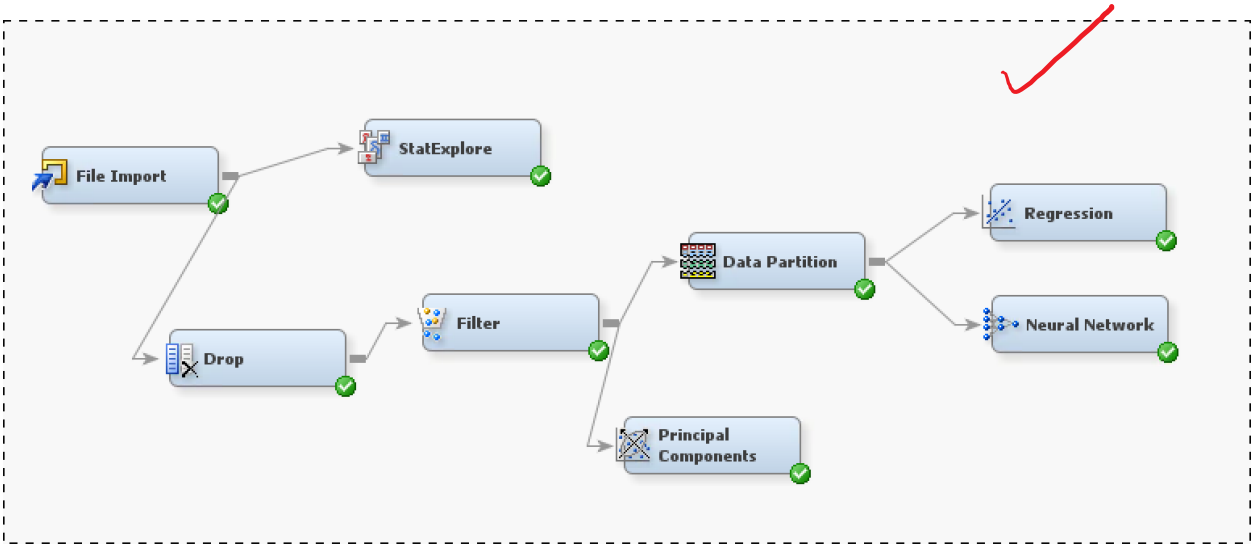
Neural Network Efforts:

In addition to the domain knowledge acquired as described above, we also used the transforming variables and principal components nodes to determine if eliminating skewed variables or creating new variables would reduce the misclassification rate. Ultimately, the models performed worse with the transformed and new variables so they were not used.

Additionally, the neural network model was run numerous times using different number of nodes to determine how many would produce the most accurate model.

Appendix

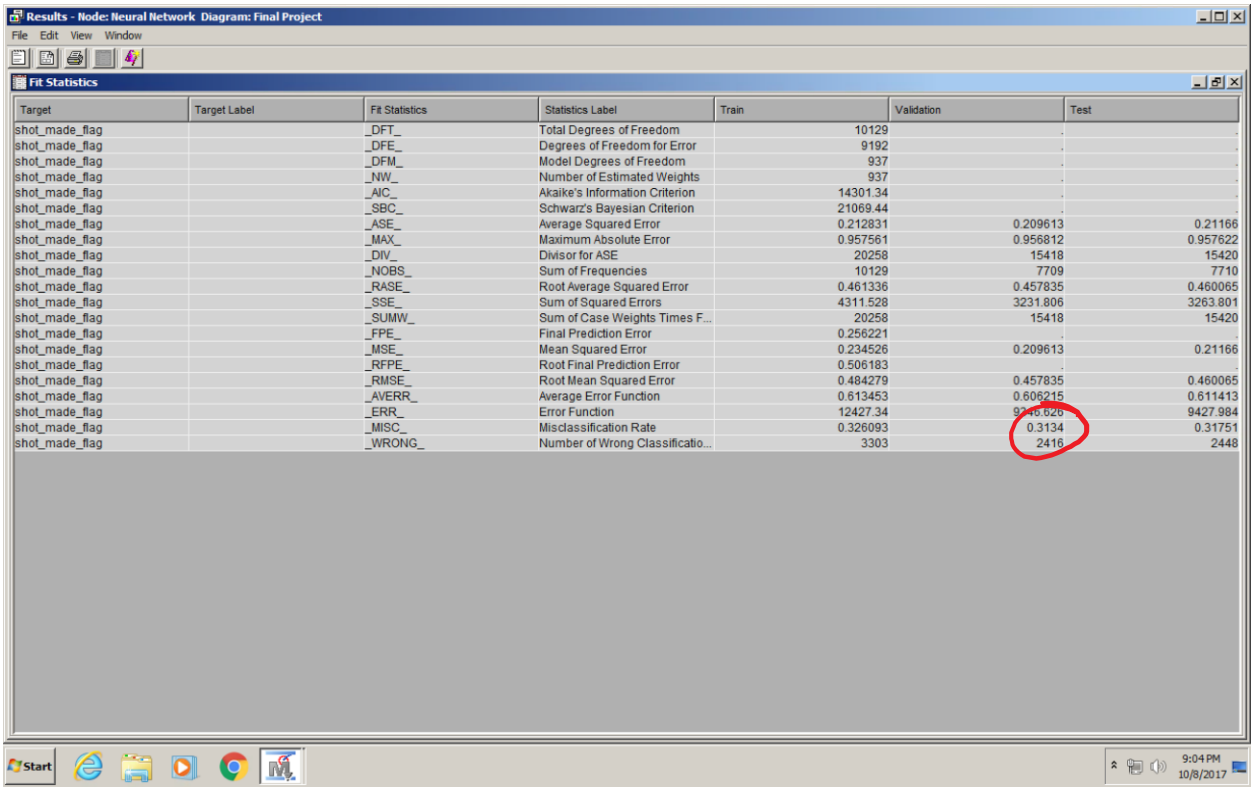
Screenshot of Neural Network Model Workspace



Screenshot of Neural Network Model Variables

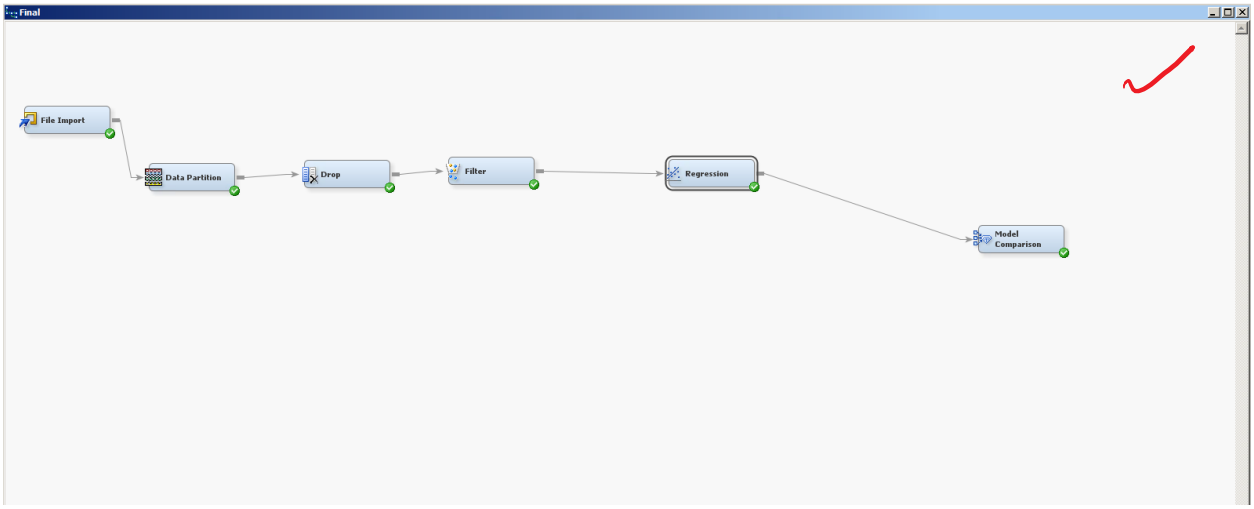
Name	Drop	Role	Level
VAR1	**Default**	Rejected	Nominal
action_type	**Default**	Input	Nominal
combined_shot_	**Yes**	Input	Nominal
game_date	**Default**	Rejected	Nominal
game_event_id	**Default**	Rejected	Interval
game_id	**Default**	Rejected	Interval
lat	**Default**	Input	Interval
loc_x	**Default**	Input	Interval
loc_y	**Default**	Input	Interval
lon	**Default**	Input	Interval
matchup	**Default**	Rejected	Nominal
minutes_remaini	**Default**	Time ID	Interval
opponent	**Default**	Rejected	Nominal
period	**Default**	Input	Nominal
playoffs	**Yes**	Input	Binary
season	**Default**	Time ID	Nominal
seconds_remaini	**Default**	Time ID	Interval
shot_distance	**Default**	Input	Interval
shot_id	**Default**	Rejected	Interval
shot_made_flag	**Default**	Target	Binary
shot_type	**Default**	Input	Binary
shot_zone_area	**Default**	Input	Nominal
shot_zone_basic	**Default**	Input	Nominal
shot_zone_rang	**Default**	Input	Nominal
team_id	**Default**	Rejected	Interval
team_name	**Default**	Rejected	Nominal

Screenshot of Neural Network Model Results

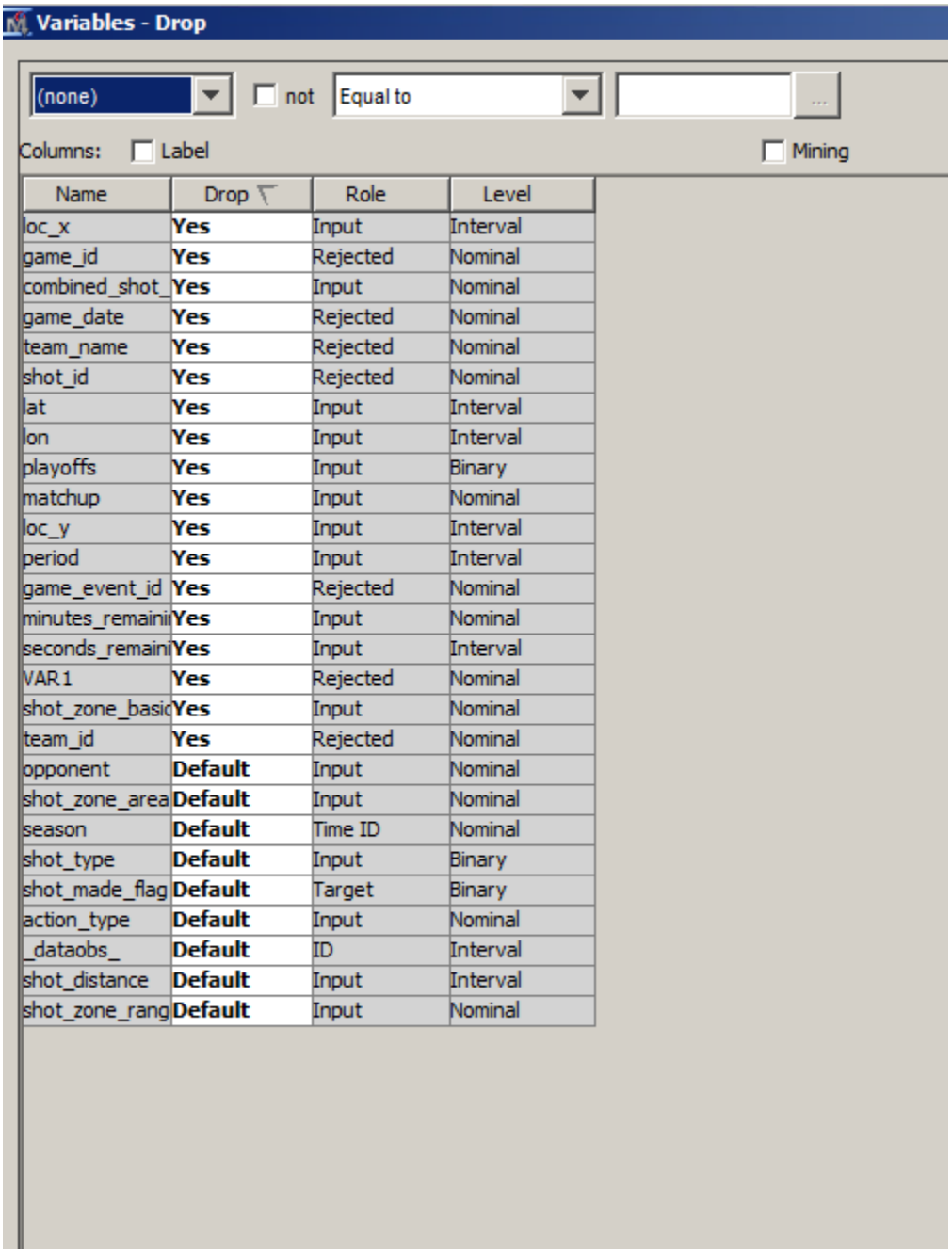


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
shot_made_flag		_DFT_	Total Degrees of Freedom	10129	.	.
shot_made_flag		_DFE_	Degrees of Freedom for Error	9192	.	.
shot_made_flag		_DFM_	Model Degrees of Freedom	937	.	.
shot_made_flag		_NW_	Number of Estimated Weights	937	.	.
shot_made_flag		_AIC_	Akaike's Information Criterion	14301.34	.	.
shot_made_flag		_SBC_	Schwarz's Bayesian Criterion	21069.44	.	.
shot_made_flag		_ASE_	Average Squared Error	0.212831	0.209613	0.21166
shot_made_flag		_MAX_	Maximum Absolute Error	0.957561	0.956812	0.957622
shot_made_flag		_DIV_	Divisor for ASE	20258	15418	15420
shot_made_flag		_NOBS_	Sum of Frequencies	10129	7709	7710
shot_made_flag		_RASE_	Root Average Squared Error	0.461336	0.457835	0.460065
shot_made_flag		_SSE_	Sum of Squared Errors	4311.528	3231.806	3263.801
shot_made_flag		_SUMW_	Sum of Case Weights Times F...	20258	15418	15420
shot_made_flag		_FPE_	Final Prediction Error	0.256221	.	.
shot_made_flag		_MSE_	Mean Squared Error	0.234526	0.209613	0.21166
shot_made_flag		_RFPE_	Root Final Prediction Error	0.506183	.	.
shot_made_flag		_RMSE_	Root Mean Squared Error	0.484279	0.457835	0.460065
shot_made_flag		_AVERR_	Average Error Function	0.613453	0.606215	0.611413
shot_made_flag		_ERR_	Error Function	12427.34	9346.626	9427.984
shot_made_flag		_MISC_	Misclassification Rate	0.326093	0.3134	0.31751
shot_made_flag		_WRONG_	Number of Wrong Classificatio...	3303	2416	2448

Screenshot of Stepwise Regression Model Workspace



Screenshot of Stepwise Regression Model Variables



Variables - Drop

(none) | not | Equal to | ...

Columns: Label | Mining

Name	Drop	Role	Level
loc_x	**Yes**	Input	Interval
game_id	**Yes**	Rejected	Nominal
combined_shot_	**Yes**	Input	Nominal
game_date	**Yes**	Rejected	Nominal
team_name	**Yes**	Rejected	Nominal
shot_id	**Yes**	Rejected	Nominal
lat	**Yes**	Input	Interval
lon	**Yes**	Input	Interval
playoffs	**Yes**	Input	Binary
matchup	**Yes**	Input	Nominal
loc_y	**Yes**	Input	Interval
period	**Yes**	Input	Interval
game_event_id	**Yes**	Rejected	Nominal
minutes_remaini	**Yes**	Input	Nominal
seconds_remaini	**Yes**	Input	Interval
VAR1	**Yes**	Rejected	Nominal
shot_zone_basic	**Yes**	Input	Nominal
team_id	**Yes**	Rejected	Nominal
opponent	**Default**	Input	Nominal
shot_zone_area	**Default**	Input	Nominal
season	**Default**	Time ID	Nominal
shot_type	**Default**	Input	Binary
shot_made_flag	**Default**	Target	Binary
action_type	**Default**	Input	Nominal
dataobs	**Default**	ID	Interval
shot_distance	**Default**	Input	Interval
shot_zone_rang	**Default**	Input	Nominal

Screenshot of Stepwise Regression Model Results

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
shot_made_flag		_AIC_	Akaike's Information C...	-15665	.	.
shot_made_flag		_ASE_	Average Squared Error	0.214477	0.211376	0.211962
shot_made_flag		_AVERR_	Average Error Function	0.214477	0.211376	0.211962
shot_made_flag		_DFE_	Degrees of Freedom f...	10190	.	.
shot_made_flag		_DFM_	Model Degrees of Fre...	50	.	.
shot_made_flag		_DFT_	Total Degrees of Free...	10240	.	.
shot_made_flag		_DIV_	Divisor for ASE	10240	7709	7710
shot_made_flag		_ERR_	Error Function	2196.241	1629.5	1634.228
shot_made_flag		_FPE_	Final Prediction Error	0.216581	.	.
shot_made_flag		_MAX_	Maximum Absolute Err...	0.983193	1	1
shot_made_flag		_MSE_	Mean Square Error	0.215529	0.211376	0.211962
shot_made_flag		_NOBS_	Sum of Frequencies	10240	7709	7710
shot_made_flag		_NW_	Number of Estimate ...	50	.	.
shot_made_flag		_RASE_	Root Average Sum of ...	0.463116	0.459757	0.460394
shot_made_flag		_RFPE_	Root Final Prediction ...	0.465383	.	.
shot_made_flag		_RMSE_	Root Mean Squared E...	0.464251	0.459757	0.460394
shot_made_flag		_SBC_	Schwarz's Bayesian C...	-15303.3	.	.
shot_made_flag		_SSE_	Sum of Squared Errors	2196.241	1629.5	1634.228
shot_made_flag		_SUMW_	Sum of Case Weights ...	10240	7709	7710
shot_made_flag		_MISC_	Misclassification Rate	0.325293	0.313789	0.316213