Statistical Modeling and Analysis Results for Kobe Bryant Shot Selection
MSDS 6372

Submitted to:
Dr. Martin Selzer

Report prepared by:
Tanvi Arora
Rebecca Holsapple
Anjli Solsi

August 14, 2018

Introduction

This report summarizes the statistical modeling and analysis results for the data set on Kobe Bryant's attempted shots. Analysis of the data is limited to the knowledge and techniques learned in MSDS 6371 and 6372. The purpose of this report is to document the detailed analysis of the proposed questions and models regarding the data set.

For the first model, an explanation is made to quantify the relationship of the odds and probability of Kobe Bryant making a shot and the distance he is from the hoop. This model is built upon to describe the relationship between the distance Kobe is from the basket and the odds of him making a shot while accounting for whether it is a playoff or regular season game.

For the predictive analysis portion, this focuses on classifying shots as made or missed for the test dataset by creating models with known techniques from the training dataset. The methods used were: logistic regression and linear discriminant analysis (LDA). These models are compared using AUC (area under the curve) values, misclassification rate, sensitivity, specificity, and the log loss function.

The final section includes an Appendix with SAS code for each analysis question and screenshots with additional details.

Note: *Italicized font* represents supplemental tables and figures found in the Appendix.

Data Description

This dataset created was from an original one that contains the location and circumstances of every shot attempted by Kobe Bryant during his 20-year career. It has been adapted and provided to us in this format by the SMU MSDS program.

With 28 explanatory variables, describing numerous aspects of location and surroundings during the shots; there is a training dataset with 25,697 observations and the flag that identifies whether the shot was made or missed. The prediction dataset of 5,000 records had the same 28 explanatory variables, but the flag detailing a made or missed shot is blank.

Name	Type / Levels	Range	Description
recId	Num	continuous 1 - 30692	Unique record id
action_type	Char	55 levels	Different ways the shot was taken
combined_shot_type	Char	bank shot, dunk, hook shot, jump shot, layup, tip shot	Type of shot
game_event_id	Num	continuous 2 - 653	Game event id
game_id	Num	continuous 20000012 - 49900088	Game id
lat	Num	continuous 33.2533 – 34.0883	Location of shot (latitude)
loc_x	Num	continuous -250 - 248	Location of shot (x-axis)
loc_y	Num	continuous -44 - 791	Location of shot (y-axis)
lon	Num	continuous -118.5198 - -118.0218	Location of shot (longitude)

minutes_remaining	Num	continuous 0 - 11	Minutes remaining in the quarter
period	Num	continuous 0 - 7	Quarter of the game
playoffs	Num	0 (regular season), 1 (playoff)	Playoff or regular season game
season	Char	20 levels: 1996-97 – 2015-16	Season year
seconds_remaining	Num	continuous 0 - 59	Seconds remaining in the quarter
shot_distance	Num	continuous 0 - 79	Distance from hoop
shot_made_flag	Num	0 (missed), 1 (made)	Shot made or missed
shot_type	Char	2PT Field Goal, 3PT Field Goal	Points awarded for the shot
shot_zone_area	Char	Back Court, Center, Left Side Center, Left Side, Right Side Center, Right Side	Area on court where shot was taken
shot_zone_basic	Char	Above the Break 3, Backcourt, In the Paint, Left Corner 3, Mid-Range, Restricted Area, Right Corner 3	Zone on court where shot was taken
shot_zone_range	Char	16-24 ft, 24+ ft, 8-16 ft, Back Court shot, less than 8 ft	Categories of distance of shot taken
team_id	Num	1610612747	Id of team
team_name	Char	Los Angeles Lakers	Name of team
game_date	Num	continuous date from 1996-2015	Date when game was played
matchup	Char	74 levels of LAL @/vs Opponent	Team and opponent, specifies home or away
opponent	Char	33 levels of team names	Opponent name
shot_id	Num	continuous 1 - 30697	Unique shot id
attendance	Num	continuous 11065 - 20485	Number of people in arena
arena_temp	Num	continuous 64 - 79	Temperature in arena
avgnoisedb	Num	continuous 88.56 – 102.43	Average noise in arena during the game

Table 1 Description of the Explanatory Variables in the Dataset

Files used:
project2Data.xlsx – the training dataset
project2Pred.xlsx – the prediction dataset

Data Transformation

The original data in project2Data.xlsx did not contain any syntax errors or naming issues with SAS, so no such cleaning was performed. However, some variables transformed for usage in the models, and others were left out for specific reasons.

The transformed variables detailed in the following table. Most of the data transformations were performed to convert categorical variables into numerical values.

Original Variable	New Variable	New Value	Reason Transformed
minutes_remaining and seconds_remaining	tot_sec_remain	continuous 0 - 750	Combined both variables for total seconds remaining in a quarter
shot_type	points	2 or 3	Convert from char to num to show 2 or 3 points
matchup	loc_adv	0 (Away) 1 (Home)	Convert from char to num, since opponent variable exists only needed Home/Away info
shot_zone_area	Nshot_zone_area	continuous 1 - 6	Convert from char to num and assign values
shot_zone_basic	Nshot_zone_basic	continuous 1 - 7	Convert from char to num and assign values
action_type	Naction_type	continuous 1 - 55	Convert from char to num and assign values
opponent	Nopponent	continuous 1 - 33	Convert from char to num and assign values

Table 2 Details on Transformed Data

Exploratory Analysis

Due to the presence of multicollinearity and other factors, various variables have been excluded from the models created. Based on the plots in *Figure 1*, variables lon and loc_x and lat and loc_y are highly correlated. The linear relationship provided enough evidence to remove lat and lon from the model. The series of plots also demonstrates a correlation between shot_distance and shot_zone_range, which is the rationale behind eliminating shot_zone_range from the model. The variable recId and shot_id are both identifiers for the shots attempted by Kobe, and leaving shot_id as the unique shot identifier. The variables game_event_id and game_id provide a unique identification for games played, and as such were eliminated from the model. The variable combined_shot_type is a less detailed version of action_type, which resulted in removal from the model. The team_name and team_id remain the same for each data point, so both were omitted from the model.

After removal of the variables noted above, the model was run to validate the changes made. Based on the VIF values in Table 3, the variables are not highly correlated.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-0.76577	0.16773	-4.57	<.0001	0
recId	recId	1	-9.10429E-7	4.548867E-7	-2.00	0.0454	1.77846
loc_x	loc_x	1	0.00001672	0.00002754	0.61	0.5438	1.00619
loc_y	loc_y	1	0.00021985	0.00006052	3.63	0.0003	3.11599
period	period	1	-0.00921	0.00264	-3.49	0.0005	1.01066
playoffs	playoffs	1	0.01267	0.01117	1.13	0.2589	1.70662
tot_sec_remain		1	0.00004729	0.00001464	3.23	0.0012	1.01852
shot_distance	shot_distance	1	-0.01277	0.00060896	-20.96	<.0001	3.57875
attendance	attendance	1	0.00004007	0.00000328	12.21	<.0001	1.36658
arena_temp	arena_temp	1	0.00790	0.00150	5.27	<.0001	1.01351
avgnoisedb	avgnoisedb	1	0.00120	0.00154	0.78	0.4348	1.35414
game_date	game_date	1	0.00000296	0.00000182	1.63	0.1029	1.12604
points		1	0.02746	0.01015	2.70	0.0068	1.87607
loc_adv		1	0.00392	0.00608	0.64	0.5196	1.01267

Table 3 Parameter Estimates and VIF values

Figure 3 represents the model of the original data with transformed variables. We examined the graphs to verify the assumptions for the models. Looking at the variables in Table 3, most have statistically significant p-values, while some do not. While the residual plot does not show a random spread of data, the values are between 1 and -1, meeting the assumption for variance. The RStudent plot does not show observations that could be outliers, with the data having values between 2 and -2. The leverage plot displays some observations, but none are large enough to be considered with a maximum value of 0.004. The Cook's D plot shows a handful of observations with high leverage, but the actual values are not significant at 0.0003. Based on the histogram, there are two distinct plots that demonstrate a relatively normal distribution. Although the QQ-plot is not highly linear, the large dataset allows for the assumption of a normal distribution. No transformations were deemed necessary for the data.

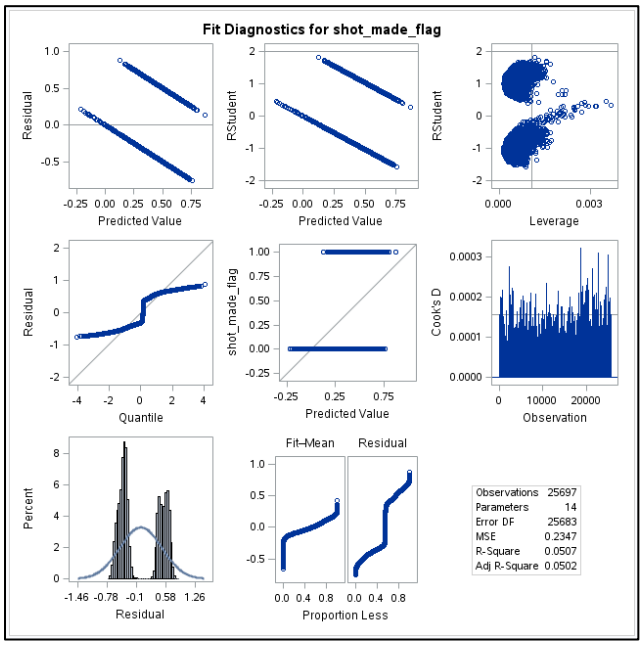


Figure 3 Fit Diagnostic Plots for Original Data

Preliminary Models

Analysis 1

The first question focuses on quantifying how the odds of Kobe making a shot decrease with respect to the distance he is from the hoop. There is evidence of this and based on the estimate of -0.0441 in Table 4, the negative relationship is apparent. For a 1-unit change in shot distance, the estimated likelihood of making the shot is -0.0441, and the odds of Kobe making the shot decreases by 0.957, with a 95% Wald confidence interval of (0.954, 0.960).

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3680	0.0224	270.2588	<.0001
shot_distance	1	-0.0441	0.00141	983.2257	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
shot_distance	0.957	0.954	0.960

Table 4 Analysis and Odds Ratio Estimates of shot_distance

Analysis 2

The second question focuses on quantifying whether the probability of Kobe making a shot decreases linearly in relation to the distance he is from the hoop. Figure 4 provides evidence in support of this. The left plot shows a decrease in the predicted probability of making a shot as the distance increases. Although not entirely linear, the trend is clear. The top left right on the right side also supports this, with red highlighting shots made.

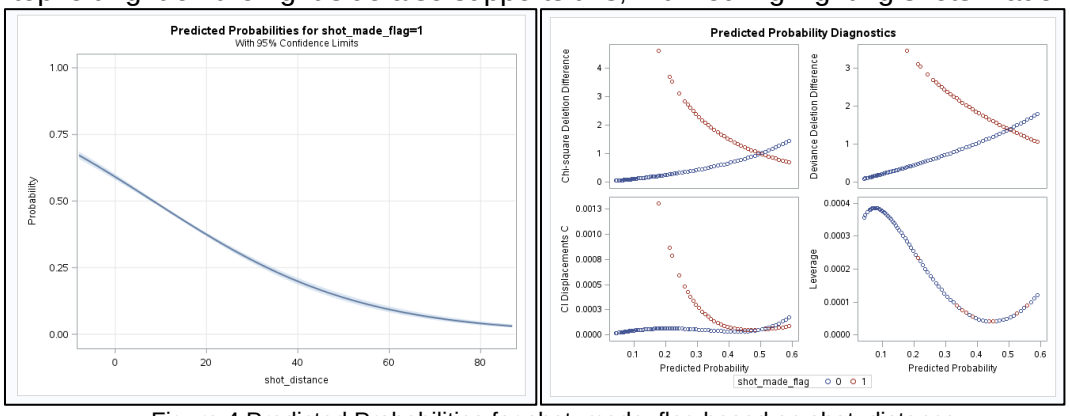


Figure 4 Predicted Probabilities for shot_made_flag based on shot_distance

Analysis 3

The third question's focus is on quantifying the relationship between the distance Kobe is from the basket and the odds of him making the shot and showing that changes in a playoff game versus regular season game. Based on Table 5, the estimate for shot distance has not changed from the first question, and playoffs have a p-value of 0.6228, meaning it is not statistically significant. There is evidence against the statement above. The odds ratio of Kobe making the shot in the playoff over the regular season is 1.018 with a 95% Wald confidence interval of (0.948, 1.093).

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3529	0.0380	86.1203	<.0001
shot_distance		1	-0.0441	0.00141	983.3929	<.0001
playoffs	0	1	0.0178	0.0362	0.2420	0.6228

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
shot_distance	0.957	0.954	0.959
playoffs 0 vs 1	1.018	0.948	1.093

Table 5 Analysis and Odds Ratio of shot_distance and playoffs

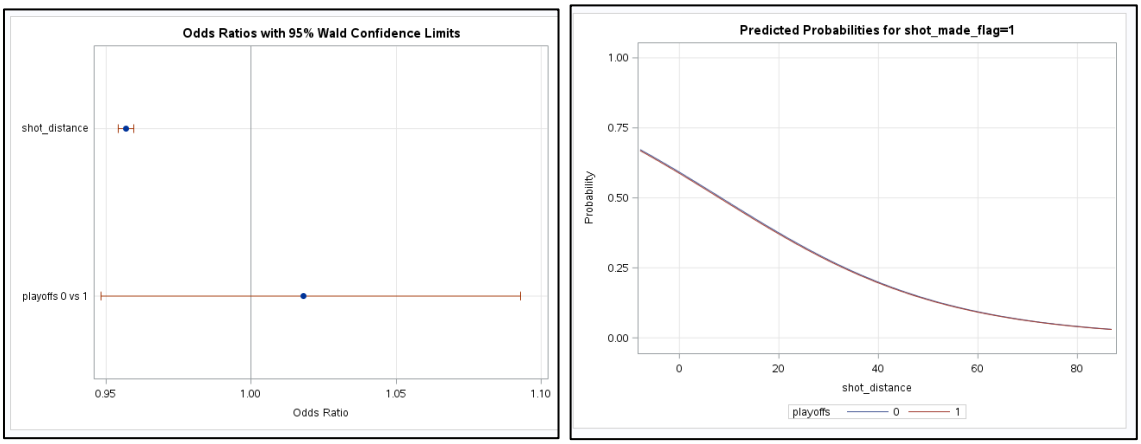


Figure 5 Odds Ratio and Predicted Probabilities for shot_made_flag based on shot_distance and playoffs

Predictive Models

Model 1 – Linear Discriminant Analysis with stepwise Quadratic test

The first is the Linear Discriminant Analysis (LDA) Model. Consider discriminant function,

$$\text{Shot\_made\_flag} = \beta_1 . {}_jX_1 + \ldots\ldots + \beta_k . {}_jX_k$$

where shot_made_flag is a binary response variable with outcome 0 or 1 and ${}_jX_1$ to ${}_jX_k$ are the predictor variables.

Based on the assumptions from the Exploratory Analysis section above, the RStudent plot does not show observations outliers, with the data having values between 2 and -2. The leverage plot displays some observations, but none are large enough to be considered with a maximum value of 0.004. The Cook's D plot shows a handful of observations with high leverage, but the actual values are not significant at 0.0003. Therefore, the assumptions for multivariate normality and standard deviation are met. Since these are all independent shots for Kobe, independence for the data is assumed. According to the Bartlett test below the homogeneity of within covariance is less than .05, indicating a quadratic test will be performed.

The DISCRIM Procedure Test of Homogeneity of Within Covariance Matrices		
Chi-Square	DF	Pr > ChiSq
3806.375578	153	<.0001

Table 6 Test of Homogeneity of Within Covariance Matrices

Both internal and external cross validation was used with LDA. For external cross validation the training dataset was split into a training dataset of 19417 observations and a test data set of 6280 observations using a random selection approach.

LDA used cross-validation on the following model:
shot_made_flag(event="1") = loc_x loc_y period playoffs tot_sec_remain shot_distance attendance arena_temp avgnoisedb game_date points loc_adv Nshot_zone_area Nshot_zone_basic Naction_type Nopponent Ncombined_shot_type

Model 2 – Stepwise LDA
Below are the predictors returned by applying the stepwise variable selection method for LDA.

Stepwise Discriminant Analysis - selection method for LDA

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE			
Total Sample Size	19417	Variable(s) in the Analysis	17
Class Levels	2	Variable(s) Will Be Included	0
		Significance Level to Enter	0.15
		Significance Level to Stay	0.15

Number of Observations Read	19417
Number of Observations Used	19417

Class Level Information

shot_made_flag	Variable Name	Frequency	Weight	Proportion
0	0	10699	10699	0.551012
1	1	8718	8718	0.448988

Stepwise Discriminant Analysis - selection method for LDA

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	Naction_type			0.0499	1018.85	<.0001	0.95013926	<.0001	0.04988074	<.0001
2	2	shot_distance		shot_distance	0.0222	440.75	<.0001	0.92904725	<.0001	0.07095275	<.0001
3	3	attendance		attendance	0.0087	169.53	<.0001	0.92100430	<.0001	0.07899570	<.0001
4	4	game_date		game_date	0.0021	40.46	<.0001	0.91908848	<.0001	0.08091152	<.0001
5	5	arena_temp		arena_temp	0.0012	23.64	<.0001	0.91797029	<.0001	0.08202971	<.0001
6	6	points			0.0009	18.14	<.0001	0.91711312	<.0001	0.08288688	<.0001
7	7	period		period	0.0007	14.18	0.0002	0.91644374	<.0001	0.08355626	<.0001
8	8	Nshot_zone_basic			0.0006	12.25	0.0005	0.91586568	<.0001	0.08413432	<.0001
9	9	tot_sec_remain			0.0006	10.72	0.0011	0.91535993	<.0001	0.08464007	<.0001
10	10	loc_y		loc_y	0.0004	7.81	0.0052	0.91499152	<.0001	0.08500848	<.0001

Table 8 Predictors returned by Stepwise LDA

Model obtained –

Shot_made_flag = loc_y period tot_sec_remain shot_distance attendance arena_temp game_date points Nshot_zone_basic Naction_type

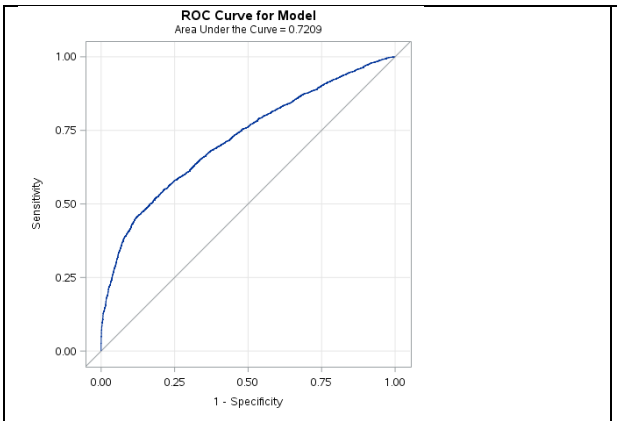
The output for QDA (quadratic discriminant analysis) on the new model can be seen in *Table 9*.

Model 3 – Logistic Regression with all predictors
The next model is the logistic regression model. Consider the Multiple Linear Regression Model,

$$\text{Shot\_made\_flag} = \beta_0 + \beta_1 . X_1 + \ldots\ldots + \beta_k . X_k$$

where shot_made_flag is a binary response variable with outcome 0 or 1 and X_1 to X_k are the predictor variables.

LR requires the same assumptions as LDA except for multivariate normality and equal variance/covariance matrices across groups. Based on the assumptions from the Exploratory Analysis section before, the RStudent plot does not show observations that could be outliers, with the data having values between 2 and -2. The leverage plot displays some observations, but none are large enough to be considered with a maximum value of 0.004. The Cook's D plot shows a handful of observations with high leverage, but the actual values are not large at 0.0003. It can be concluded that assumptions for normality and standard deviation have been met. Since these are all independent shots for Kobe, independence for the data can be assumed.



Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	1239	3077	456	1508	68.7	45.1	87.1	26.9	32.9

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	8609.295	7636.187
SC	8616.040	8297.209
-2 Log L	8607.295	7440.187

Actual by Predicted Using Test Data for logistic regression model from stepwiseclass

The FREQ Procedure

Frequency
Row Pct

Table of shot_made_flag by I_shot_made_flag

shot_made_flag(shot_made_flag)	I_shot_made_flag(Into: shot_made_flag) 0	1	Total
0	3097 87.66	436 12.34	3533
1	1492 54.31	1255 45.69	2747
Total	4589	1691	6280

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	FALSE POS	FALSE NEG	Misclassification Rate	AUC	log loss function
0.5	1239	3077	456	1508	68.7	45.1	87.1	26.9	32.9	0.312738854	0.7209	0.1658

Figure 6/Table 10 Output of Logistic Regression on all Predictors

Model 4 – Logistic Regression with stepwise selection

The next model is the logistic regression model. Consider the Multiple Linear Regression Model,

$$\text{Shot\_made\_flag} = \beta_0 + \beta_1 . X_1 + \ldots\ldots + \beta_k . X_k$$

where shot_made_flag is a binary response variable with outcome 0 or 1 and X_1 to X_k are the predictor variables.

LR requires the same assumptions as LDA except for multivariate normality and equal variance/covariance matrices across groups. Based on the assumptions from the Exploratory Analysis section before, the RStudent plot does not show observations that could be outliers, with the data having values between 2 and -2. The leverage plot displays some observations, but none are large enough to be considered with a maximum value of 0.004. The Cook's D plot shows a handful of observations with high leverage, but the actual values are not large at 0.0003. It can be concluded that assumptions for normality and standard deviation have been met. Since these are all independent shots for Kobe, independence for the data can be assumed.

Below are the predictors returned by applying the stepwise variable selection method for Logistic Regression.

Stepwise logistic with significance level and combined_shot/lat/lon

The LOGISTIC Procedure

Model Information		
Data Set	WORK.KOBE_SHOTS_TRAIN	
Response Variable	shot_made_flag	shot_made_flag
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	19417
Number of Observations Used	19417

Response Profile

Ordered Value	shot_made_flag	Total Frequency
1	0	10699
2	1	8718

Probability modeled is shot_made_flag='1'.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Naction_type		51	1	2735.6876		<.0001	
2	attendance		1	2	147.5359		<.0001	attendance
3	Nshot_zone_area		5	3	43.6327		<.0001	
4	arena_temp		1	4	24.6971		<.0001	arena_temp
5	tot_sec_remain		1	5	20.9379		<.0001	
6	Nshot_zone_basic		6	6	27.8698		<.0001	
7	period		6	7	24.8227		0.0004	period
8	game_date		1	8	11.2776		0.0008	game_date
9	shot_distance		1	9	11.2847		0.0008	shot_distance
10	loc_x		1	10	4.9322		0.0264	loc_x
11	Nopponent		32	11	46.9639		0.0427	

Table 11 Predictors returned by Stepwise Variable Selection using Logistic Regression

Logistic regression does not perform cross validations. So, external cross validation has been implemented by splitting the training dataset into a training dataset of 19417 observations and a test data set of 6280 observations. This was done using a random selection approach. The model obtained using this approach is below.

shot_made_flag(event="1") = loc_x period tot_sec_remain shot_distance attendance arena_temp game_date Nshot_zone_area Nshot_zone_basic Naction_type Nopponent

The logistic regression was applied on the test dataset using the above model.

Both the LDA and logistic regression methods are adequate for linear classification models. Both the LDA models discussed have almost similar statistics with the base LDA being slightly better with a lower log loss function value of 0.0508. Comparing the logistic models, the AUC statistic provides a measure of accuracy of the model in correctly predicting an observation is in a particular group. The AUC values are almost the same with the base model being slightly better with a value of 0.7213. The log loss function is used to compare the LDA and LR models, and the LDA provide a better model with a lower log loss function. Sensitivity is higher in LDA while specificity is lower, meaning that the true positives are better predicted with this model. The LDA model was deemed to be better at forecasting for this dataset.

Predictive Models	AUC	Mis-Classification Rate	Sensitivity	Specificity	Log Loss Function
Model 1 – LDA	N/A	0.3578	45.5	78.8	0.0508
Model 2 – stepwise LDA	N/A	0.3578	45.4	78.79	0.0593
Model 3 – Logistic Regression (Base)	0.7213	0.3146	44.8	87.0	0.5924
Model 4 – Logistic Regression (stepwise selection)	0.7209	0.3127	45.1	87.1	0.5919

Table 12 Predictive Modeling Techniques and Relevant Statistics

Conclusion

This report summarizes the statistical modeling and analysis results regarding attempted shots throughout Kobe Bryant's career. This involved modeling groups of variables as well as conducting predictive analysis to determine the best method to forecast whether Kobe will make or miss a shot.

For Analysis 1 and 2, the relationship of Kobe making a shot decreasing with respect to the distance from the hoop was quantified. Evidence was provided in favor of the relationship demonstrated by: for a 1-unit change in shot distance, the estimated likelihood of making the shot is -0.0441, and the odds of Kobe making the shot decreases by 0.957. The plot also supports this with the predicted probability of making a shot decreasing as the distance increases. Analysis 3 involved factoring in regular season games versus playoff games to the relationship. It was found that shot performance based on distance does not change in playoff games.

For the Predictive Analysis, out of the four models that were created, LDA, stepwise LDA, Logistic Regression, and Logistic Regression with stepwise selection, the LDA technique was chosen to complete the predictive analysis due to the smallest log loss function value of 0.0508. Each model's predictors were chosen using different selection methods. The strength of the LDA model come from considering almost all the variables, with categorical variables being converted to numerical values for analysis. These models demonstrate the types of factors that may be more heavily weighted than popular belief in Kobe Bryant making a shot. It would be interesting to continue the study with a more in-depth approach to the data, as well as comparing Kobe's data to other players, such as Steph Curry's and Lebron James's shots.

Appendix

Figures/Tables

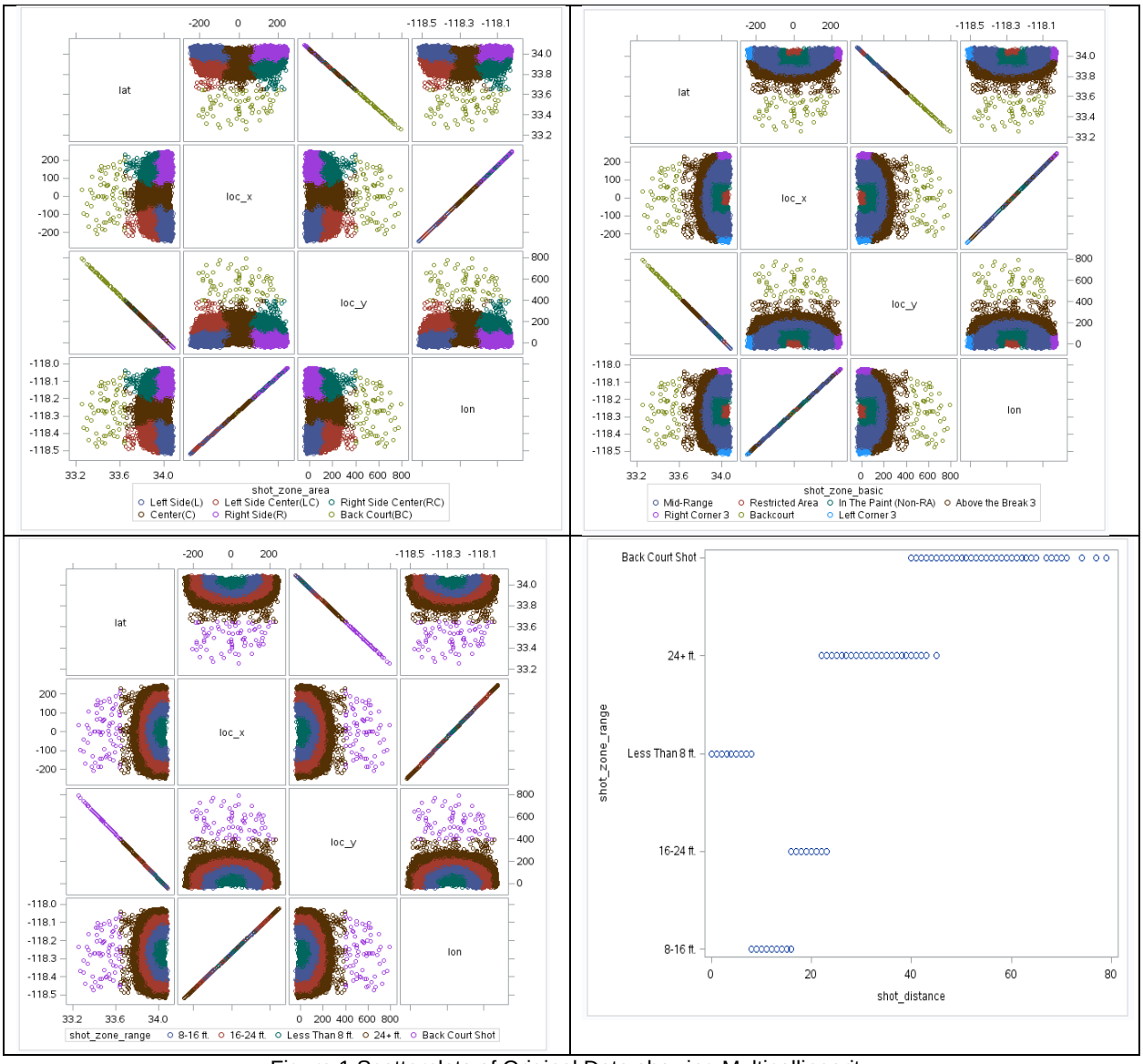


Figure 1 Scatterplots of Original Data showing Multicollinearity

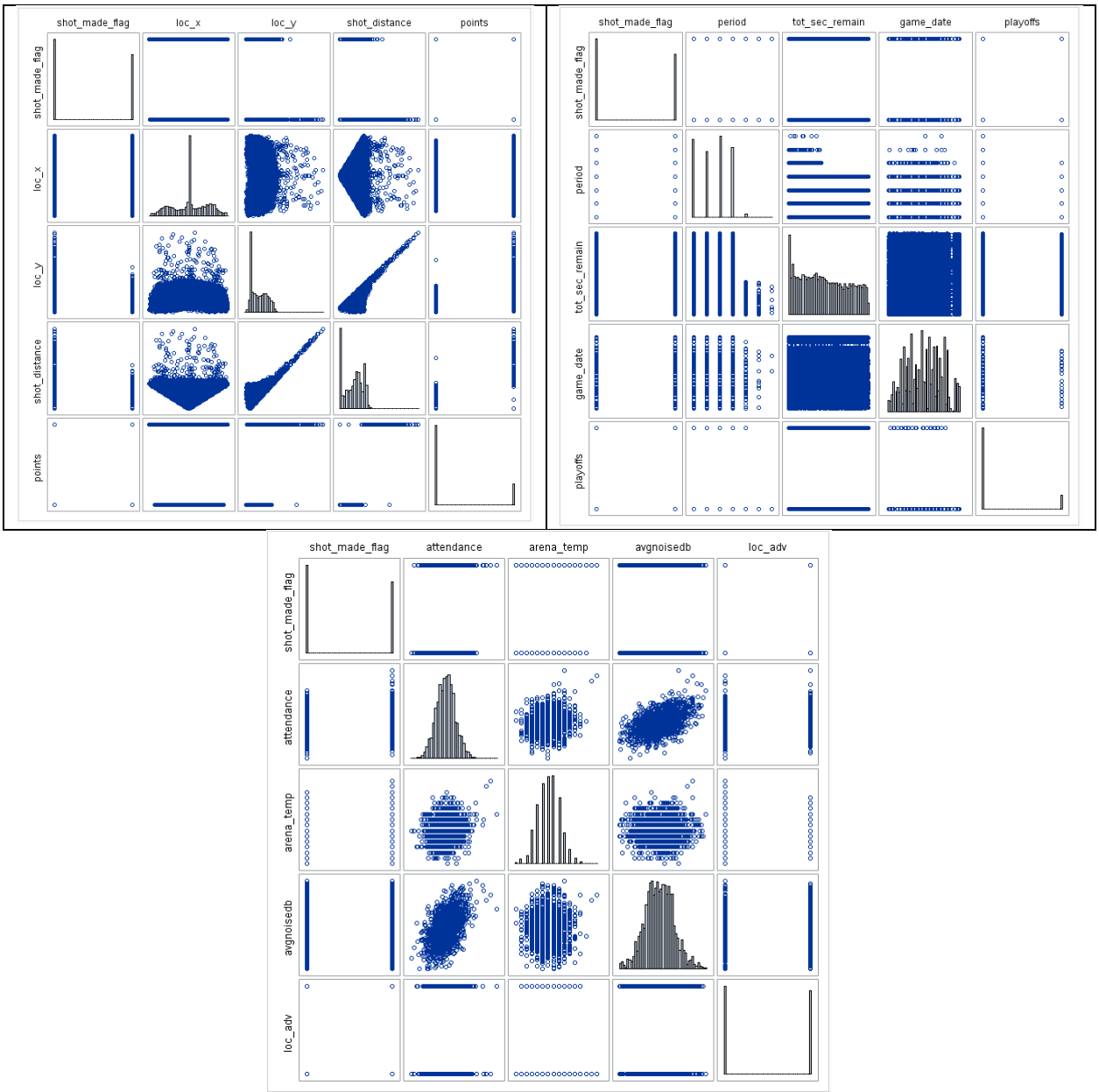
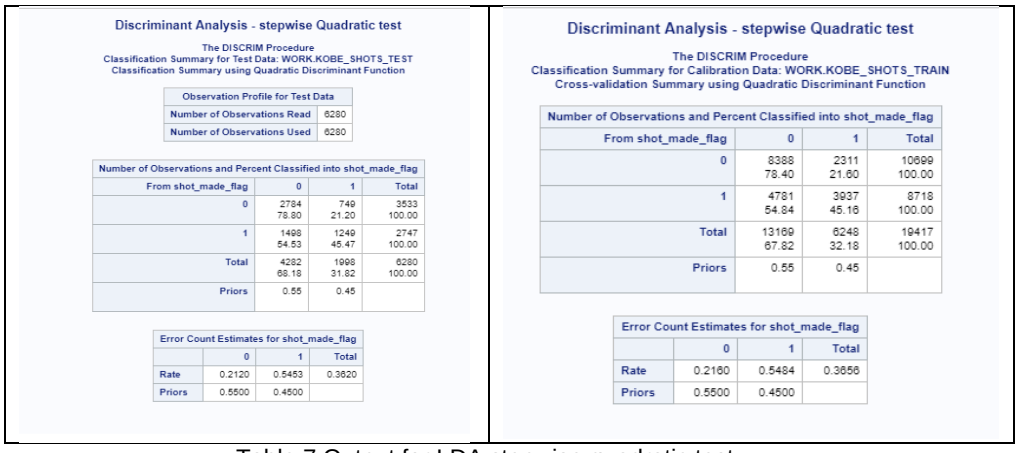


Figure 2 Scatterplot Matrix of the Original Data



Discriminant Analysis - stepwise Quadratic test

The DISCRIM Procedure
Classification Summary for Test Data: WORK.KOBE_SHOTS_TEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	6280
Number of Observations Used	6280

Number of Observations and Percent Classified into shot_made_flag			
From shot_made_flag	0	1	Total
0	2784 78.80	749 21.20	3533 100.00
1	1498 54.53	1249 45.47	2747 100.00
Total	4282 68.18	1998 31.82	6280 100.00
Priors	0.55	0.45	

Error Count Estimates for shot_made_flag			
	0	1	Total
Rate	0.2120	0.5453	0.3620
Priors	0.5500	0.4500	

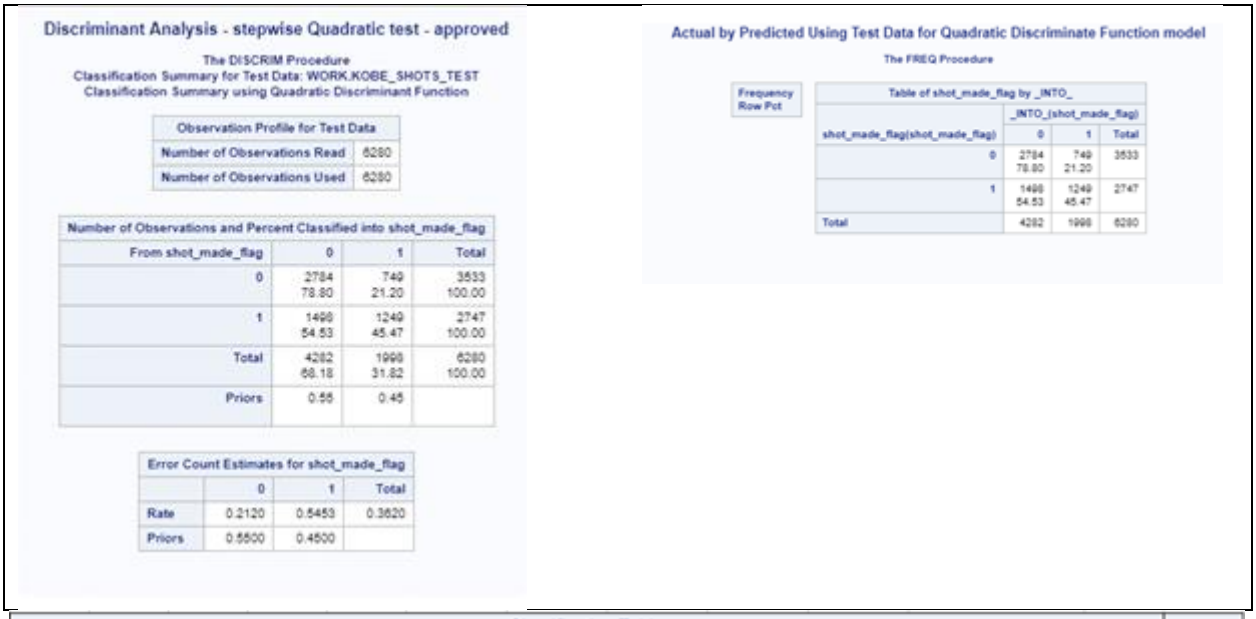
Discriminant Analysis - stepwise Quadratic test

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.KOBE_SHOTS_TRAIN
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into shot_made_flag			
From shot_made_flag	0	1	Total
0	8388 78.40	2311 21.60	10699 100.00
1	4781 54.84	3937 45.16	8718 100.00
Total	13169 67.82	6248 32.18	19417 100.00
Priors	0.55	0.45	

Error Count Estimates for shot_made_flag			
	0	1	Total
Rate	0.2160	0.5484	0.3656
Priors	0.5500	0.4500	

Table 7 Output for LDA stepwise quadratic test



Discriminant Analysis - stepwise Quadratic test - approved

The DISCRIM Procedure
Classification Summary for Test Data: WORK.KOBE_SHOTS_TEST
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	6280
Number of Observations Used	6280

Number of Observations and Percent Classified into shot_made_flag			
From shot_made_flag	0	1	Total
0	2784 78.80	749 21.20	3533 100.00
1	1498 54.53	1249 45.47	2747 100.00
Total	4282 68.18	1998 31.82	6280 100.00
Priors	0.55	0.45	

Error Count Estimates for shot_made_flag			
	0	1	Total
Rate	0.2120	0.5453	0.3620
Priors	0.5500	0.4500	

Actual by Predicted Using Test Data for Quadratic Discriminate Function model

The FREQ Procedure

Frequency Row Pct	Table of shot_made_flag by _INTO_		
	INTO(shot_made_flag)		
shot_made_flag(shot_made_flag)	0	1	Total
0	2784 78.80	749 21.20	3533
1	1498 54.53	1249 45.47	2747
Total	4282	1998	6280

Classification Table												
Prob	Correct		Incorrect		Percentages							
Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	FALSE POS	FALSE NEG	Misclassification Rate	AUC	log loss function
0.5	1249	2784	749	1498	64.22	45.467783	78.7998868	0.37487487	0.34983652	0.357802548		0.1465

Table 9 Output for QDA on new model

SAS Code

EXPLORATORY

```
*check correlation of shot_distance and shot_zone_range;
proc sgplot data=project2data;
scatter x=shot_distance y=shot_zone_range;
run;

*check correlations of the location variables to each group;
proc sgscatter data=project2data;
matrix lat loc_x loc_y lon / group=shot_zone_area;
run;


proc sgscatter data=project2data;
matrix lat loc_x loc_y lon / group=shot_zone_basic;
run;


proc sgscatter data=project2data;
matrix lat loc_x loc_y lon / group=shot_zone_range;
run;
```

PRELIMINARY MODELS

```
*exploratory analysis, data transformations assigned to new data set;
*create total seconds, change shot_type to num, change matchup to num;
data kobe_shots;
set project2data;
tot_sec_remain = (minutes_remaining*60) + seconds_remaining;
if (shot_type eq "2PT Field Goal") then points = 2; else points = 3;
*lattendance = log(attendance);
if (substr(matchup, 5, 1) eq "@") then loc_adv = 0; else loc_adv = 1;
run;
proc contents data=kobe_shots;
run;

*redo code , remove lat and lon and shotID to examine changes ----- add game_date;
*implement changes from new data;
proc reg data=kobe_shots plots(maxpoints=none);
model shot_made_flag = recId loc_x loc_y period
playoffs tot_sec_remain shot_distance attendance arena_temp avgnoisedb game_date points loc_adv / VIF;
run;
```

Data Preparation

```
*assign numerical values to categorical *;

proc format;
        value $num_shot_zone_area
        'Left Side(L)'=1
        'Left Side Center(LC)'=2
        'Right Side Center(RC)'=3
        'Center(C)'=4
        'Right Side(R)'=5
        'Back Court(BC)'=6
        ;
        value $num_shot_zone_basic
        'Mid-Range'=1
        'Restricted Area'=2
        'In The Paint (Non-RA)'=3
```

```
        'Above the Break 3'=4
        'Right Corner 3'=5
        'Backcourt'=6
        'Left Corner 3'=7
;
        value $num_action_type
        'Jump Shot'=1
        'Driving Dunk Shot'=2
        'Layup Shot'=3
        'Running Jump Shot'=4
        'Reverse Dunk Shot'=5
        'Slam Dunk Shot'=6
        'Driving Layup Shot'=7
        'Turnaround Jump Shot'=8
        'Reverse Layup Shot'=9
        'Tip Shot'=10
        'Running Hook Shot'=11
        'Alley Oop Dunk Shot'=12
        'Dunk Shot'=13
        'Alley Oop Layup shot'=14
        'Running Dunk Shot'=15
        'Driving Finger Roll Shot'=16
        'Running Layup Shot'=17
        'Finger Roll Shot'=18
        'Fadeaway Jump Shot'=19
        'Follow Up Dunk Shot'=20
        'Hook Shot'=21
        'Turnaround Hook Shot'=22
        'Jump Hook Shot'=23
        'Running Finger Roll Shot'=24
        'Jump Bank Shot'=25
        'Turnaround Finger Roll Shot'=26
        'Hook Bank Shot'=27
        'Driving Hook Shot'=28
        'Running Tip Shot'=29
        'Running Reverse Layup Shot'=30
        'Driving Finger Roll Layup Shot'=31
        'Fadeaway Bank shot'=32
        'Pullup Jump shot'=33
        'Finger Roll Layup Shot'=34
        'Turnaround Fadeaway shot'=35
        'Driving Reverse Layup Shot'=36
        'Driving Slam Dunk Shot'=37
        'Step Back Jump shot'=38
        'Turnaround Bank shot'=39
        'Reverse Slam Dunk Shot'=40
        'Floating Jump shot'=41
        'Putback Slam Dunk Shot'=42
        'Running Bank shot'=43
        'Driving Bank shot'=44
        'Driving Jump shot'=45
        'Putback Layup Shot'=46
        'Putback Dunk Shot'=47
        'Running Finger Roll Layup Shot'=48
        'Pullup Bank shot'=49
        'Running Slam Dunk Shot'=50
        'Cutting Layup Shot'=51
        'Driving Floating Jump Shot'=52
        'Running Pull-Up Jump Shot'=53
        'Tip Layup Shot'=54
        'Driving Floating Bank Jump Shot'=55
;
```

```
        value $num_opponent
        'POR'=1
        'UTA'=2
        'VAN'=3
        'LAC'=4
        'HOU'=5
        'SAS'=6
        'DEN'=7
        'SAC'=8
        'CHI'=9
        'GSW'=10
        'MIN'=11
        'IND'=12
        'SEA'=13
        'DAL'=14
        'PHI'=15
        'DET'=16
        'MIL'=17
        'TOR'=18
        'MIA'=19
        'PHX'=20
        'CLE'=21
        'NJN'=22
        'NYK'=23
        'CHA'=24
        'WAS'=25
        'ORL'=26
        'ATL'=27
        'MEM'=28
        'BOS'=29
        'NOH'=30
        'NOP'=31
        'OKC'=32
        'BKN'=33
;
value $num_combined_shot_type
        'Jump Shot'=1
        'Dunk'=2
        'Layup'=3
        'Tip Shot'=4
        'Hook Shot'=5
        'Bank Shot'=6
;
run;

*prepare training data for the model*;
data kobe_shots;
set project2data;
tot_sec_remain = (minutes_remaining*60) + seconds_remaining;
if (shot_type eq "2PT Field Goal") then points = 2; else points = 3;
if (substr(matchup, 5, 1) eq "@") then loc_adv = 0; else loc_adv = 1;
RandNumber = ranuni(11);
*Below categorical values converted to numerical values*;
Tshot_zone_area = put(shot_zone_area,$num_shot_zone_area.);
Tshot_zone_basic = put(shot_zone_basic,$num_shot_zone_basic.);
Taction_type = put(action_type,$num_action_type.);
Topponent = put(opponent,$num_opponent.);
Tcombined_shot_type = put(combined_shot_type,$num_combined_shot_type.);
Nshot_zone_area = input(Tshot_zone_area,8.);
Nshot_zone_basic = input(Tshot_zone_basic,8.);
Naction_type = input(Taction_type,8.);
Nopponent = input(Topponent,8.);
```

```
Ncombined_shot_type = input(Tcombined_shot_type,8.);
drop Tshot_zone_area Tshot_zone_basic Taction_type Topponent;
run;

* generate train data set where RandNumber > 0.25;

proc sort data=kobe_shots; by RandNumber; run;
data kobe_shots_train kobe_shots_test;
 set kobe_shots;
 if RandNumber <= 1/4 then output kobe_shots_test; else output kobe_shots_train;
 run;

*print contents of training and test datasets to verify the count and format of data*;
proc contents data=kobe_shots_train; run;
proc contents data=kobe_shots_test; run;
```

MODEL 1

```
*LDA - linear discriminant analysis **;
title 'Discriminant Analysis - Quadratic test';
proc discrim data=kobe_shots_train pool=no crossvalidate testData=kobe_shots_test testout=kobe_shots_qdiscout;
class shot_made_flag ;
var loc_x loc_y period
playoffs tot_sec_remain shot_distance attendance arena_temp avgnoisedb game_date points loc_adv Nshot_zone_a
rea Nshot_zone_basic Naction_type Nopponent Ncombined_shot_type;
priors "1" = 0.45 "0" = 0.55;
run;

title 'Actual by Predicted Using Test Data for Quadratic Discriminate Function model';
proc freq data=kobe_shots_qdiscout;
table shot_made_flag*_into_ / nocol nopercent;
run;
```

MODEL 2

```
title 'Stepwise Discriminant Analysis - selection method for LDA';
proc stepdisc data=kobe_shots_train ;
class shot_made_flag ;
var loc_x loc_y period playoffs tot_sec_remain shot_distance attendance arena_temp avgnoisedb game_date points
loc_adv Nshot_zone_area Nshot_zone_basic Naction_type Nopponent Ncombined_shot_type;
run;

title 'Discriminant Analysis - stepwise Quadratic test - approved';
proc discrim data=kobe_shots_train pool=no crossvalidate testData=kobe_shots_test testout=kobe_shots_sqdiscout;
class shot_made_flag ;
var  loc_y period  tot_sec_remain shot_distance attendance arena_temp  game_date points   Nshot_zone_basic
Naction_type  ;
priors "1" = 0.45 "0" = 0.55;
run;

proc print data=kobe_shots_sqdiscout(keep=shot_id _INTO_ _0 _1); run;
```

MODEL 3

```
title 'Logistic regression - from stepwise selection class model with ctable - approved';
proc logistic data=kobe_shots_test order=data plots=all;
        class shot_made_flag(param=ref) Nshot_zone_area Nshot_zone_basic Naction_type Nopponent;
```

```
        model shot_made_flag(event="1") = loc_x period  tot_sec_remain shot_distance attendance arena_temp
game_date  Nshot_zone_area Nshot_zone_basic Naction_type Nopponent / ctable lackfit clparm=wald link=glogit
pprob=0.5;
        output out = logistic6Out predprobs=I p=predprob resdev=resdev reeschi=pearres;
        Score data=kobe_shots_test out=kobe_shots_log6out;
run;

title 'Actual by Predicted Using Test Data for logistic regression model from stepwiseclass';
proc freq data=kobe_shots_log6out;
table shot_made_flag*i_shot_made_flag / nocol nopercent;
run;
```

MODEL 4

```
title 'Stepwise logistic with significance level and combined_shot/lat/lon';
proc logistic data=kobe_shots_train outest=betas covout;
        class shot_made_flag(param=ref) period  points loc_adv Nshot_zone_area Nshot_zone_basic Naction_type
Nopponent Ncombined_shot_type;
        model shot_made_flag(event='1') = loc_x loc_y period playoffs tot_sec_remain shot_distance attendance
arena_temp avgnoisedb game_date points loc_adv Nshot_zone_area Nshot_zone_basic Naction_type Nopponent
Ncombined_shot_type lat lon/ selection=stepwise ;
  run;

title 'Logistic regression - from stepwise selection class model with ctable ';
proc logistic data=kobe_shots_test order=data plots=all;
        class shot_made_flag(param=ref) Nshot_zone_area Nshot_zone_basic Naction_type Nopponent;
        model shot_made_flag(event="1") = loc_x period  tot_sec_remain shot_distance attendance arena_temp
game_date  Nshot_zone_area Nshot_zone_basic Naction_type Nopponent / ctable lackfit clparm=wald link=glogit
pprob=0.5;
        output out = logistic6Out predprobs=I p=predprob resdev=resdev reeschi=pearres;
        Score data=kobe_shots_test out=kobe_shots_log6out;
run;

title 'Actual by Predicted Using Test Data for logistic regression model from stepwiseclass';
proc freq data=kobe_shots_log6out;
table shot_made_flag*i_shot_made_flag / nocol nopercent;
run;
```

PREDICTIVE MODEL – LDA

```
data kobe_shots_pred;
set project2Pred;
tot_sec_remain = (minutes_remaining*60) + seconds_remaining;
if (shot_type eq "2PT Field Goal") then points = 2; else points = 3;
if (substr(matchup, 5, 1) eq "@") then loc_adv = 0; else loc_adv = 1;
RandNumber = ranuni(11);
Tshot_zone_area = put(shot_zone_area,$num_shot_zone_area.);
Tshot_zone_basic = put(shot_zone_basic,$num_shot_zone_basic.);
Taction_type = put(action_type,$num_action_type.);
Topponent = put(opponent,$num_opponent.);
Tcombined_shot_type = put(combined_shot_type,$num_combined_shot_type.);
Nshot_zone_area = input(Tshot_zone_area,8.);
Nshot_zone_basic = input(Tshot_zone_basic,8.);
Naction_type = input(Taction_type,8.);
Nopponent = input(Topponent,8.);
Ncombined_shot_type = input(Tcombined_shot_type,8.);
drop Tshot_zone_area Tshot_zone_basic Taction_type Topponent;
*shot_made_flag is converted to numeric value to match the model from the training data set;
if shot_made_flag eq 'NA' then Tshot_made_flag=input('0',8.);
drop shot_made_flag;
```

```
rename Tshot_made_flag=shot_made_flag;
*2 action types missing from training dataset. modifying test dataset to nearest available value*;
if shot_id = 22624 then Naction_type=34;  *mode value for "Finger Roll Layup Shot";
if shot_id =22578 then Naction_type=35;   *mode value is "Turnaround Fadeaway shot";
run;

proc contents data=kobe_shots_pred; run;


title 'Discriminant Analysis - Quadratic test';
proc discrim data=kobe_shots pool=no crossvalidate testData=kobe_shots_pred testout=kobe_shots_pdiscout;
class shot_made_flag ;
var loc_x loc_y period playoffs tot_sec_remain shot_distance attendance arena_temp avgnoisedb game_date points
loc_adv Nshot_zone_area Nshot_zone_basic Naction_type Nopponent Ncombined_shot_type;
priors "1" = 0.45 "0" = 0.55;
run;

title 'Actual by Predicted Using Test Data for Quadratic Discriminate Function model';
proc freq data=kobe_shots_pdiscout;
table shot_made_flag*_into_ / nocol nopercent;
run;

proc print data=kobe_shots_pdiscout(keep=shot_id _INTO_ _0 _1); run;
```