Che Cobb, Joe Schueder, Anthony Egbuniwe

Professor Anthony Tanaydin

SMU-DS 6372

April 7, 2019

**KOBE BRYANT SHOT SELECTION**

**Introduction:**

Kobe Bryant marked his retirement from the NBA by scoring 60 points in his final game as a Los Angeles Laker in 2016. Using 2014- 2016 data as a retrospective study of Kobe's swishes and misses, predictive models were built for Kobe Bryant’s shots with Logistic Regression techniques. The goals was to find the most significant features in predicting his shots (shot\_made\_flag). The target variable for this project is binary, with 0-missed, 1-made. The objective is to predict Kobe’s shots with Logistic Regression models and compare the models’ performance.

**Data Description:**

The dataset, obtained from Kaggle, was sourced directly from the NBA. It contains every goal attempted by Kobe in his 20-year career, a total of 30,697 shots. Of these, 5,000 were randomly selected to serve as a test set in file project2pred.xlsx, with their shot success labels removed. The data contains many pieces of information, with 25 variables in total. The table (Table 1) shows all the variables and their disposition.

|  |  |  |
| --- | --- | --- |
| **Table 1** | | |
| **shot\_made\_flag**​: 1 if shot made, 0 if not made. | **lat**​: The latitude of Kobe’s position during the shot attempt. | ​**lon**​: The longitude. |
| **action\_type**​: The type of shot attempted, like a jump shot, dunk, etc. Total there are 57 distinct values. | ​**lon**​: The longitude. | **loc\_x**​: The x-location on the court. |
| **combined\_shot\_type**​: Classifies the shots under 6 categories: Bank Shot, Dunk,  Hook Shot, Jump Shot, Layup, and Tip Shot. | **loc\_x**​: The x-location on the court. | **loc\_y**​: The y-location on the court. |
| **game\_event\_id**​: The ID of the game event (attempted shot) in the specific match being played. Discarded for our purposes. | ​**loc\_y**​: The y-location on the court.  ​m 1 to 30,697) of the attempted shot. | **action\_type**​: The type of shot attempted, like a jump shot, dunk, etc. Total there are 57 distinct values. |
| **game\_id**​: The ID of the specific match. Also removed. | **minutes\_remaining**​: The minutes remaining in the specific match. | **combined\_shot\_type**​: Classifies the shots under 6 categories: Bank Shot, Dunk,  Hook Shot, Jump Shot, Layup, and Tip Shot. |
| **lat**​: The latitude of Kobe’s position during the shot attempt. | **period**​: The period in the specific match. | **game\_event\_id**​: The ID of the game event (attempted shot) in the specific match being played. Discarded for our purposes. |
| ​**shot\_zone\_range**​: Range (<8 ft, 8-16, 16-24, 24+, backcourt)  ecific match. Discarded. | **playoffs**​: Indicator variable whether the match was in the playoffs or not. | **shot\_type**​: 2pt or 3pt. |
| **opponent**​: Opponent in the specific match.  ​ | **season**​: The basketball season (2000, 2001, etc.) | ​**shot\_zone\_area**​: Area from which shot was attempted (Right, Left, Center, Back Court,  Right Center, Left Center) |
| **game\_id**​: The ID of the specific match. Also removed. | ​**seconds\_remaining**​: The seconds remaining in the specific match. | ​**team\_id**​: ID of Kobe’s team. Always the Lakers so discarded.21. team\_name: Name of Kobe’s team,the Lakers, so discarded. |
| **shot\_id**​: ID (from 1 to 30,697) of the attempted shot. | ​**shot\_distance**​: The distance from which the shot was attempted, in ft. | ​**matchup**​: The two teams in the specific match. Since Kobe was always on the Lakers, opponent contains all the information in the matchup. The matchup is thus discarded. |
| ​**shot\_zone\_basic**​: Further area information (Mid-range, restricted area, in the paint, above the break 3, backcourt, left corner 3, right corner 3) | **game\_date**​: Date of the sp |  |
| And several more like game\_event\_id, game\_id, team\_id, team\_name, game\_date, and matchup. | | |

After dispositioning all the variables, 18 predictors remain with one response, shot\_made\_flag.

The dataset is further cleaned by combining the 5 least attempted shots in action\_type into another category. Next, the remaining predictors were explored to find meaningful relationships with shot\_made\_flag. There are two approaches utilized to achieve the desired outcome.

1. Various Selection methods including LASSO (Least Absolute Shrinkage and Selection Operator) to select variables.

2. Using the findings of selection methods and domain knowledge to guide variable selection.

**Exploratory Data Analysis:**

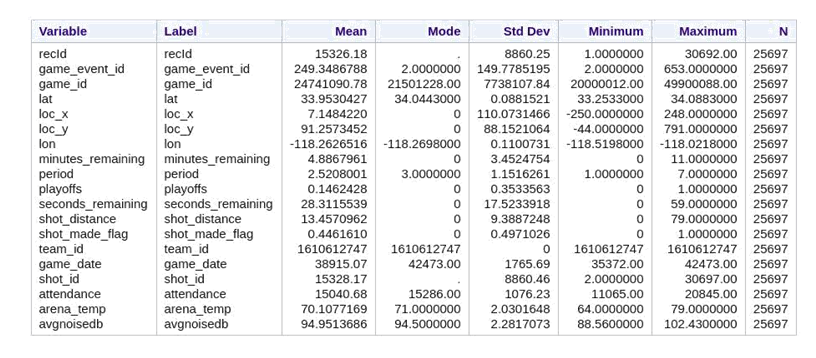
The first step was exploring the need for any potential transformations.No transformations are needed since it did not appear to make a difference in any models. Secondly, differentiate the variables. From the above description, the features to check are Numerical Features, Categorical Features and Statistical Variables behavior of those features, and finally checking for missing values. Numerical Features include lat, loc\_x, loc\_y, lon, arena\_temp, avgnoisedb,game\_date, game\_event\_id, game\_id, period, playoffs, season, shot\_id, attendance, minutes\_remaining, seconds\_remaining, shot\_distance, shot\_made\_flag, team\_id. Categorical Features include action\_type, combined\_shot\_type, period,playoffs, season, shot\_type, shot\_zone\_area, shot\_zone\_basic, shot\_zone\_range. Figure 2 shows the Variable’s Statistical Behavior. ​The statistical part of the data consists of the variablecalculation based on their lowest and highest value present, labeling of the data and other statistical presence. 

Figure 1 - Kobe Bryant Career Summary Statistics

An analysis was done to check for missing values in data.All the predictors in the trainingfile were examined to determine the presence of missing values in the data set. Missing value could harmful to a model due to overfitting and underfitting concerns. This analysis showed the variable season has 6,182 Number of missing values out of 25,697. All other variables were a 100% populated. See appendix for details. A key summary statistic shows Kobe’s overall accuracy. Actual Accuracy of the shot\_made\_flag by Kobe. The actual accuracy of the shot hits by theKobe Bryant is as follows based on training data. The percentage probability of the shot\_made\_flag by Kobe is 55.4% failure and 44.6% success.

|  |  |
| --- | --- |
| Shot made | 44.6% |
| Shot missed | 55.4% |

Figure 2 - Kobe Bryant Accuracy Summary

|  |  |
| --- | --- |
| **shot\_zone\_area Frequency and Accuracy ​** | |
| 1 | 2 |
| 3 | 4 |

Figure 3

The shot\_zone\_area does not vary as much as action\_type, but it is likely useful for prediction to some extent. The shot played by Kobe based on area shows as follows:

The fig.3 chart 1 shows the distribution of shot\_zone\_area by its shot frequency. The top frequency of the shot\_zone\_area is Central(C). The Central shot zone area is a high frequency area for most basketball games and also Kobe’s seemingly favorite shot zone area. Fig.3 chart 3 a visualization of shot\_zone\_area, shows the on-court representation of each zone. As expected, shots from the backcourt are at such great range that the accuracy is extremely low.

Though the variation in accuracy appears large, it must be noted that backcourt shots are rare. Of course, such shots are hardly ever attempted by any player, and Kobe is no exception. Among the remaining zones, Kobe appears to slightly prefer the right and highly prefers the center. He also shoots much better in the center, and slightly better in the right zones. A quick search confirms that Kobe shoots with either hand, but that his right is dominant.

A very important variable to consider when exploring Kobe’s data are shots played by Kobe the shot zone area. Let's have a preview below. Figure 3 chart 4 provides a visualization of the on-court locations, and Figure 3 chart 2 the accuracy and number of shots by location. Kobe’s accuracy by shot\_zone\_basic actually varies substantially even after accounting for the fact that shots from the corners and the backcourt are very rare. Surprisingly, Kobe’s left corner accuracy is higher than right corner accuracy.

It might be tempting to conclude shot\_zone\_basic should be included in our model. However, in actuality the variable is simply describing the influence of range on accuracy. It is required to analyze shot\_zone\_range in and shot\_distance in order to decide on shot\_zone\_basic’s conclusion. For instance, it could well be the case that shot\_zone\_range contains all the information of shot\_zone\_basic and more, or vice-versa.

The next figure shows the visual distribution of the shot zone range in percentage occupied by when playing shots by Kobe. This variable is a categorization of distances and is very helpful in answering questions about accuracy by distance.

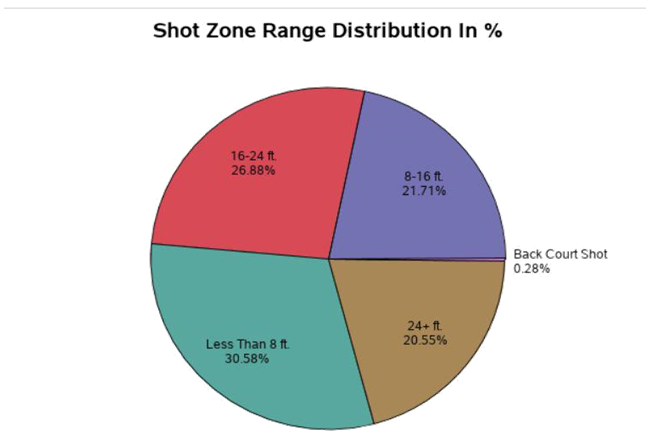
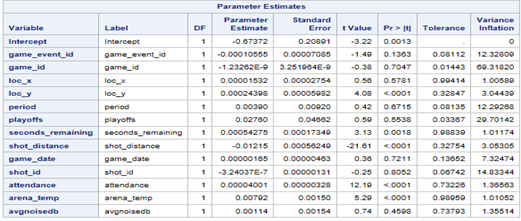


Figure 4- Shot Zone Attempts Distribution

An outlier is a data point that is distant from other similar points. They may be due to variability in the measurement or may indicate experimental errors. If possible, outliers should be excluded from the dataset. For outlier detection, only the continuous-valued variables can be considered. No outliers were identified.

****

**Figure 5**

Multicollinearityis a state of very high inter-correlations or inter-associations among theindependent variables. It is, therefore, a type of disturbance in the data, if present in the data the statistical inferences made about the data may not be reliable. In review of the Tolerance results, several variables – namely game\_event\_id, game\_id, period, playoffs and shot\_id – have values well below the desired 0.1 cutoff value. This finding is echoed in review of the Variance Inflation results, where these same variables reveal values far larger than the 10 cutoff for this column. For the sake of completeness, collinearity diagnostics are also reviewed.

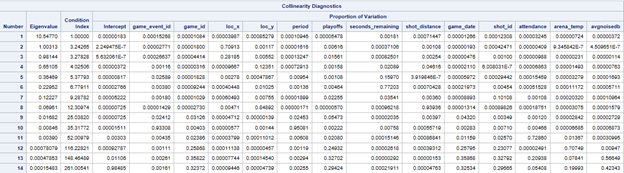


Figure 6

In review of the eigenvalue and condition index association, there is a large deviation in the final three factors, with the eigenvalue resulting very close to zero and the condition index resulting quite large in comparison. So, a prime case for multicollinearity was found.

**Model Building:**

​After the initial data exploration, aLogistic Regressionmodel is used to select variables and to predict the dataresult. Logistic Regression is the categorical regression analysis to conduct when the dependent variable isdichotomous (binary) that is yes/no analysis. Like all regression analyses, the logistic regression is predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The logistic regression equation is:

y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x)) --------------------(1)

Using this method, the model building happens on the training data name as Kobe.xlsx and prediction based on project2pred.xlsx test data set. LASSO Regression and correlation of variable methods were used to select the set of the final variables for the operations like model fitting and prediction. All variables were statistically significant at the 0.05 level.

The final selected set of variables is a combined\_shot\_type, action\_type, shot\_type, shot\_zone\_area, shot\_zone\_basic, period, playoffs, shot\_zone\_range, loc\_y loc\_x, minutes\_remaining, seconds\_remaining & shot\_distance were also significant.

**Model Evaluation:**

AUC (area under the curve) value is a measure of accuracy.The AUC value for this model is 0.7030 i.e. 70% accuracy. AMisclassification Rate is included in figure 7 c. This misclassification rate generally called a confusion matrix to evaluate the actual and predicted rate. The rate includes Specificity Vs Sensitivity. Specificity is the probability of being correctly classified shot made when the shot was actually made. Sensitivity is the probability of being correctly classified as missed if the shot was actually missed. Overall Misclassification Rate is 8,140/25,697 = 31.7%. Sensitivity is estimated to be 12,280/14,232 = 86.3% from this study. Specificity is 5,277/11,465 = is 46%. The ROC (receiving operating characteristic) Curve shows sensitivity vs specificity. A straight vertical line in chart 7a would mean unhelpful performance, while a line that went up the left side and across the top would show a very good model. This model is somewhere in between those two extremes and is considered an acceptable model. In the ROC chart since the AUC is roughly 70% the model will correctly predict made shots and missed shots.

|  |  |
| --- | --- |
| a | |
| b | c |

Figure 7

One of the most significant variables in determining whether Kobe will make a shot is the distance from the hoop. The analysis of this specific variable shows that as the distance from the hoop increases the probability of making the shot decreases. As shown in various images in figure 8 the probability of making a shot goes from around 55% at zero feet to close to 0% at 80 feet in a non-linear pattern. There are not significantly different odds and probabilities between the regular season and playoffs.

|  |  |  |
| --- | --- | --- |
| Pure Distance | Distance Grouping | Including Playoffs |
|  |  |  |
|  | | |

Figure 8

In simpler odds and probability terms, each foot further away from the goal decreases the probability by 1.1% based on converting logit to probability. A sampling of distances and probabilities with confidence intervals are shown in figure 9.

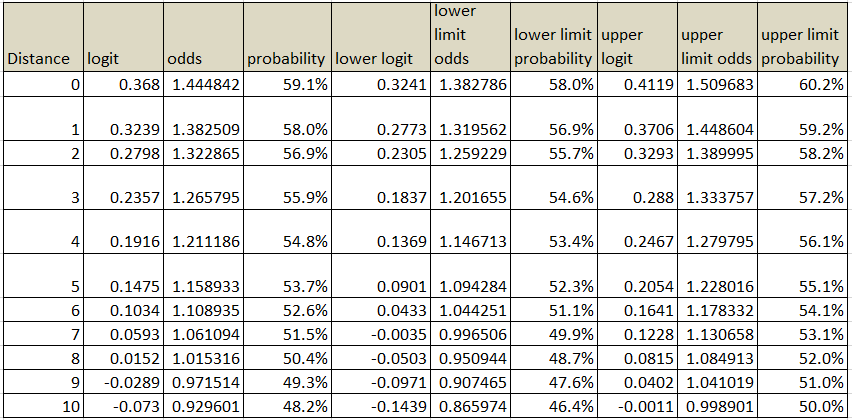


Figure 9

**Conclusion:**

To evaluate this model, the predictions were performed on a test set of data. Using the Log loss metric, the performance of this model was compared. The Log Lossvalue for this model is about 11.182. Unfortunately, performance is not as high as desired. Still, the work presented here goes a long way in showing how applicable modern statistics is to sports. One can easily imagine the implications. For example, team coaches can easily maintain a model for each of their players, and analyze which shots they need to improve on and which they excel, whether their performance dips with less time remaining in the match, who on the team should shoot longer ranges, where to position each player, etc. The possibilities are endless. Hopefully, this exercise will serve as a step towards more creative quantitative analysis in sports.

Appendix

1. Overall model SAS code

/\* Import a file kobe.xlsx \*/

%web\_drop\_table(WORK.IMPORT);

FILENAME MyFile '/home/chec0/New Folder/Kobe.xlsx';

PROC IMPORT DATAFILE=MyFile

DBMS=XLSX

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

%web\_open\_table(WORK.IMPORT);

data Kobe1;

set WORK.IMPORT;

run;

/\*

proc print data=Kobe1(obs=15);

run;

\*/

PROC SORT DATA=Kobe1;

by shot\_zone\_range;

run;

/\*Summary of the Variables \*/

proc means data=Kobe1 chartype mean mode std min max n vardef=df;

\* var lat loc\_x loc\_y lon minutes\_remaining period playoffs seconds\_remaining

shot\_distance shot\_made\_flag game\_date attendance arena\_temp avgnoisedb;

run;

/\* Frequency of getting output \*/

proc freq data=Kobe1;

table shot\_made\_flag;

run;

/\* Count missing values for numeric variables \*/

DATA Kobe2;

set Kobe1;

if missing(shot\_made\_flag) then delete;

run;

proc print data=Kobe2(obs=10);

run;

/\*Bar Chart : Target Variable Visualization \*/

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=Kobe1;

title height=14pt "Target Variable Class Distribution";

vbar shot\_made\_flag / fillattrs=(color=CX5f9ae7) datalabel stat=percent;

yaxis grid;

run;

ods graphics / reset;

title;

/\*Bar graph of shot\_zone\_area \*/

title'Accuracy of the shot by shot\_zone\_area';

proc sgplot data=Kobe1;

vbar shot\_zone\_area /;

yaxis grid;

run;

/\*Summary of the Variables \*/

proc means data=Kobe1 chartype mean std min max n vardef=df;

var shot\_distance shot\_made\_flag;

class shot\_type;

run;

/\*\*\* Analyze categorical variables \*\*\*/

title "Frequency for Categorical Variables";

proc freq data=Kobe1;

tables action\_type combined\_shot\_type shot\_type shot\_zone\_area shot\_zone\_basic

shot\_zone\_range / plots=(freqplot);

run;

/\*\*\* Analyze numeric variables \*\*\*/

title "Descriptive Statistics for Numeric Variables";

proc means data=Kobe1 n nmiss min mean median max std;

var shot\_made\_flag;

run;

/\* Define Pie template for shot\_zone\_range\*/

proc template;

define statgraph SASStudio.Pie;

begingraph;

entrytitle "Shot Zone Range Distribution In %" / textattrs=(size=14);

layout region;

piechart category=shot\_zone\_range / stat=pct;

endlayout;

endgraph;

end;

run;

proc sgrender template=SASStudio.Pie data=Kobe1;

run;

/\*Frequency of Shots by Shot\_Zone\_Area \*/

proc sgplot data=Kobe1;

title height=14pt "Frequency of Shots by Shot\_Zone\_Area";

vbar shot\_zone\_area / group=shot\_made\_flag groupdisplay=cluster datalabel;

yaxis grid label="Shots";

run;

/\*Frequency of shots by shot\_zone\_basic \*/

proc sgplot data=Kobe1;

title height=14pt "Frequency of shots by shot\_zone\_basic";

vbar shot\_zone\_basic / group=shot\_made\_flag groupdisplay=cluster datalabel;

yaxis grid label="Shots";

refline 6253 / axis=y lineattrs=(thickness=2 color=green)

label="Highest Accuracy" labelattrs=(color=green);

run;

/\* Combined Shot type vs Shot made by flag \*/

proc sgplot data=Kobe1;

hbar combined\_shot\_type / group=shot\_made\_flag groupdisplay=cluster;

xaxis grid;

run;

/\*Bar graph to show the total shots of particular action\_type\*/

proc sort data=Kobe1 out=BarChartTaskData\_action\_type;

by shot\_made\_flag;

run;

proc sgplot data=BarChartTaskData\_action\_type;

by shot\_made\_flag;

hbar action\_type / group=shot\_made\_flag groupdisplay=cluster datalabel;

xaxis min=100 grid;

run;

proc datasets library=WORK noprint;

delete BarChartTaskData\_action\_type;

run;

/\*Variable Selection using LASSO method \*/

proc glmselect data=kobe1;

class combined\_shot\_type action\_type shot\_type shot\_zone\_area shot\_zone\_basic

period playoffs shot\_zone\_range season opponent;

model shot\_made\_flag = combined\_shot\_type action\_type shot\_type arena\_temp attendance

shot\_zone\_area shot\_zone\_basic period shot\_zone\_range season avgnoisedb game\_date lat lon playoffs

loc\_y loc\_x minutes\_remaining seconds\_remaining shot\_distance shot\_id / selection=lasso(stop=none);

run;

/\*Outlier Identification \*/

title 'Outlier Identification for continuous-valued variable';

proc univariate data=kobe1 robustscale plot;

var arena\_temp attendance period avgnoisedb game\_date lat lon playoffs

loc\_y loc\_x minutes\_remaining seconds\_remaining shot\_distance shot\_id;

run;

/\*Correlation Analysis for Feature Selection: Select the variable which has significant value greater then 0.05 \*/

title 'Correlation between the variables';

proc corr data=Kobe1 pearson nomiss nosimple rank

plots=matrix(histogram);

var loc\_x loc\_y minutes\_remaining period seconds\_remaining

shot\_distance;

with shot\_made\_flag;

run;

/\*Multicollinearity\*/

proc reg data=Kobe1;

model shot\_made\_flag = period lon loc\_y loc\_x minutes\_remaining seconds\_remaining shot\_distance / tol vif COLLIN; /\*vif = variance inflation factor and tol= tolorance \*/

run;

/\* Lat and Lon scatter plot visualization \*/

title 'lat and lon scatter plot and outlier detection';

proc sgplot data=Kobe1;

scatter x=lat y=lon /;

xaxis grid;

yaxis grid;

run;

/\* loc\_x and loc\_y scatter plot visualization \*/

title 'loc\_x and loc\_y scatter plot and outlier detection';

proc sgplot data=Kobe1;

scatter x=loc\_x y=loc\_y /;

xaxis grid;

yaxis grid;

run;

/\* loc\_x and loc\_y scatter plot visualization by shot\_zone\_area \*/

proc sgplot data=Kobe1;

scatter x=loc\_x y=loc\_y / group=shot\_zone\_area;

xaxis grid;

yaxis grid;

run;

/\*shots attempted by shot distance \*/

title 'shots attempted by shot distance';

proc sort data=Kobe1 out=HistogramTaskData\_shot\_made\_flag;

by shot\_made\_flag;

run;

proc sgplot data=HistogramTaskData\_shot\_made\_flag;

by shot\_made\_flag;

title height=14pt "shots by shot\_distance";

histogram shot\_distance / scale=count nbins=13 fillattrs=(color=CX4b5ab8)

weight=shot\_made\_flag;

xaxis max=80;

yaxis max=4000 grid;

run;

proc datasets library=WORK noprint;

delete HistogramTaskData\_shot\_made\_flag;

run;

/\*Accuracy by seconds\_remaining \*/

title 'Accuracy by seconds\_remaining';

proc sgplot data=Kobe1;

title height=14pt "Accuracy by seconds\_remaining";

vbar seconds\_remaining / group=shot\_made\_flag groupdisplay=stack datalabel;

yaxis grid label="Accuracy";

run;

/\* Import 2nd File \*/

title 'Import test file';

FILENAME MyFile2 '/folders/myfolders/project2Pred.xlsx';

PROC IMPORT DATAFILE=MyFile2

DBMS=XLSX

OUT=Kobe3;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=Kobe3 varnum; RUN;

proc print data=Kobe3(obs=10);

run;

/\* Change the data type of the variable shot\_made\_flag from char to num \*/

title 'Change the data type of the target variable from char to num';

data Kobe4;

set Kobe3;

shot\_made\_flag\_new = input(shot\_made\_flag, 8.);

drop shot\_made\_flag;

rename shot\_made\_flag\_new = shot\_made\_flag;

run;

/\* Checked the mean of the target variable and in project2pred.xlsx replace NA values by 0.5 for further prediction \*/

title 'Checked the stats of the test file';

proc means data= Kobe2 mean mode;

var shot\_made\_flag;

RUN;

/\* Logistic Regression: shot\_made\_flag(event='1') because we are only intresting in knowing the probability of getting shots done or made\*/

/\* Load file project2pred in a file as Kobe3 to show the prediction of the variable\*/

/\*Model training and getting AUC value with the accuracy of 0.70 i,e 70% \*/

/\* I am using file name Kobe as to train the model and project2pred for prediction.\*/

title 'Training of the model';

proc logistic data=Kobe1 plots=all;

class combined\_shot\_type action\_type shot\_type shot\_zone\_area shot\_zone\_basic

period shot\_zone\_range season / param=glm;

model shot\_made\_flag(event='1')=combined\_shot\_type action\_type shot\_type

shot\_zone\_area shot\_zone\_basic period season shot\_zone\_range

loc\_y loc\_x minutes\_remaining seconds\_remaining shot\_distance /

link=logit selection=forward slentry=0.05 hierarchy=single technique=fisher;

output out=work.Logistic\_stats1 xbeta=xbeta predicted=pred / alpha=0.05;

score data =Kobe1 out=MyPred fitstat;

run;

proc print data=MyPred(obs=10);

run;

/\*Model fit to predict the value \*/

title 'Test the accuracy of the model';

proc logistic data=Kobe1;

class combined\_shot\_type action\_type shot\_type shot\_zone\_area shot\_zone\_basic

period shot\_zone\_range season / param=glm;

model shot\_made\_flag(event='1')=combined\_shot\_type action\_type shot\_type

shot\_zone\_area shot\_zone\_basic period season shot\_zone\_range

loc\_y loc\_x minutes\_remaining seconds\_remaining shot\_distance /

link=logit selection=forward slentry=0.05 hierarchy=single technique=fisher;

output out=work.Logistic\_stats1 xbeta=xbeta predicted=pred / alpha=0.05;

score data =Kobe4 out=MyPred\_test;

run;

/\* Show the result with shot\_id and shot\_made\_flag \*/

data pridicted\_value;

set MyPred\_test;

drop combined\_shot\_type action\_type shot\_type shot\_zone\_area shot\_zone\_basic period playoffs shot\_zone\_range lat lon

loc\_y loc\_x minutes\_remaining seconds\_remaining shot\_distance shot\_made\_flag rannum game\_event\_id game\_id

season team\_id team\_name game\_date matchup opponent attendance arena\_temp avgnoisedb F\_shot\_made\_flag P\_0;

label F\_shot\_made\_flag = 'shot\_made\_flag' P\_0= 'Actual\_Predicted\_value';

run;

proc print data=pridicted\_value(obs=50) label;

run;

/\* Mis-Classification Chart or Confusion Matrix \*/

proc freq data=MyPred;

tables F\_shot\_made\_flag\*I\_shot\_made\_flag;

run;

1. Distance analysis SAS code

\* Import a file kobe.xlsx \*/

%web\_drop\_table(WORK.IMPORT);

FILENAME MyFile '/home/jschueder0/project2Data.xlsx';

PROC IMPORT DATAFILE=MyFile

DBMS=XLSX

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

%web\_open\_table(WORK.IMPORT);

data Kobe1;

set WORK.IMPORT;

run;

/\*

proc print data=Kobe1(obs=15);

run;

\*/

proc discrim data=Kobe1 pool=test;

class playoffs;

var shot\_made\_flag shot\_distance;

run;

proc logistic data=Kobe1 outmodel=model1 PLOTS(MAXPOINTS=NONE)=all;

class playoffs;

model shot\_made\_flag(event='1')= playoffs shot\_distance

/ link=logit selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=pred xbeta=xbeta p=phat lower=lcl upper=ucl / alpha=0.05;

score data =Kobe1 out=MyPred fitstat;

run;

proc univariate data=Kobe1;

histogram;

run;

/\* Combined Shot type vs Shot made by flag \*/

proc sgplot data=Kobe1;

hbar shot\_zone\_range / group=shot\_made\_flag groupdisplay=cluster;

xaxis grid;

run;

proc logistic data=Kobe1 outmodel=model1 plots=all;

model shot\_made\_flag(event='1')= shot\_distance

/ link=logit

details

lackfit;

output out=pred xbeta=xbeta p=phat lower=lcl upper=ucl / alpha=0.05;

score data =Kobe1 out=MyPred fitstat;

run;

proc sort data=Kobe1;

by descending shot\_zone\_range;

run;

proc freq data=Kobe1 order=data;

format shot\_made\_flag ExpFmt. shot\_made\_flag RspFmt.;

tables shot\_zone\_range\*shot\_made\_flag / CMH chisq riskdiff(equal var=null) alpha = .05;

title 'Kobe Study ';

run;

proc logistic data = Kobe1 plots=all;;

class shot\_zone\_range / param = ref;

model shot\_made\_flag(event='1') = shot\_zone\_range/ lackfit ctable CLPARM=WAL CLODDS=WAL;

Output out=Probs Predicted=Phat lower=lcl upper=ucl

predprob=(individual crossvalidate) ;

run;

proc logistic data = Kobe1 plots=all;

class playoffs shot\_zone\_range / param = ref;

model shot\_made\_flag(event='1') = shot\_distance/ lackfit ctable CLPARM=WAL CLODDS=WAL;

Output out=Probs Predicted=Phat lower=lcl upper=ucl

predprob=(individual crossvalidate) ;

run;

1. Extra images



.