Modeling Kobe

Daniel Byrne, Ganesh Kodi

April 12, 2019

## Abstract

Sports data modeling has been a staple of video game development for many years. The major leagues NBA, NFL, NHL have all partnered with software development companies to produce realistic video games depicting the league’s top stars acting and scoring as they do in real life.

With the $20 billion video game industry fueled by the now growing e-sports segment, ever more realistic models are needed for developers to build the characters in their games and to stay competitive in the market. We explore a common dataset of Kobe Bryant’s career shots, and try to build a model that would predict the likelihood of his making or missing a shot.

We show that our final models are poor, 35% to 47% accurate in predicting a shot will be made from anywhere on the floor. We further speculate that additional data points commonly captured in sport statistics such as whether or not the shot was contested could add specificity to the model.

## Data Description

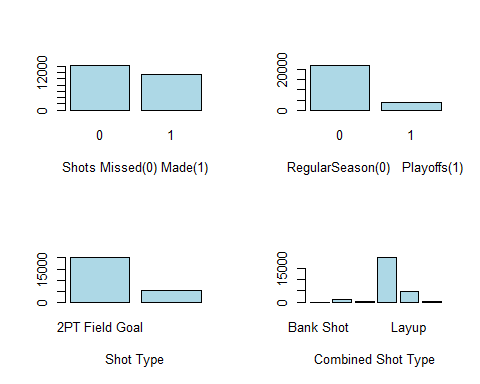
The field names are self-explanatory. The predictors our analysis focused on are as follows:

|  |  |
| --- | --- |
| Data Label | Description |
| combined\_shot\_type | Type of shot combined with action |
| loc\_y | Vertical position on floor |
| minutes\_remaining | Minutes remaining in quarter |
| playoffs | Playoff game or not |
| seconds\_remaining | Seconds remaining in quarter |
| shot\_distance | Distance from goal |
| shot\_made\_flag | 1- shot went I, 0 - shot missed |
| shot\_type | 2pt or 3pt shot |
| attendance | The attendance in the stadium |
| arena\_temp | The average temperature during the game |
| avgnoisedb | The average noise level in dB during the game |

## Data Analysis

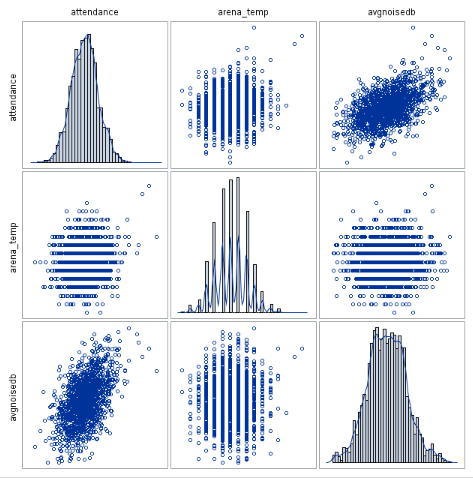
We evaluated a large but not exhaustive number of predictive variable combinations and potential models in our analysis. The following variables played a part in our final model

* Shot\_distance : We analyzed the hypothesis that Kobe’s odds of making his shots decreased as the shot distance increased and whether or not this was a linear phenomenon.
* Shot\_type : We saw a statistically significant contribution from shot\_type which led us to include this variable into our final model.
* Combined\_shot\_type : Likewise combined\_shot\_type showed a statistically significant contribution.
* Playoffs : Used to evaluate Kobe’s performance in the regular season vs. the playoffs



We combined the following continuous variables into their principle components to include in out final model.

* Time\_remaining : Created this datapoint from Minutes\_remaining\*60 + seconds\_remaining
* Average Attendance
* Average Temperature
* Average Noise Level (dB)



## Multicollinearity Analysis

High bivariate correlations were easy to spot when we ran correlation calculations among our target predictors. We noticed significant correlations between loc\_y and shot\_distance, loc\_x and lon and loc\_y and lat. Coincidently we did not find models with both loc\_y and shot\_distance or with loc\_x and lon or loc\_y and lat to be good models due to their collinearity.

## loc\_x loc\_y lat lon  
## loc\_x 1.00000000 -0.01757819 0.01757819 1.00000000  
## loc\_y -0.01757819 1.00000000 -1.00000000 -0.01757819  
## lat 0.01757819 -1.00000000 1.00000000 0.01757819  
## lon 1.00000000 -0.01757819 0.01757819 1.00000000

## loc\_y shot\_distance  
## loc\_y 1.000000 0.818124  
## shot\_distance 0.818124 1.000000

There are also some potential similarities between the categorical variables in the dataset although categorical variables cannot be colinear as they do not represent linear measures in Euclidean space. As such we use chi-square tests on our coefficients to determine independence and significance of the categorical variables.

## Outlier Analysis

Based on Cook’s D data and plain data analysis, we are seeing no outlier present in the selected variables. The table blow shows the first 5 observations with highest cook’s D value. Since these values are less than 3, we assume that there are no outliers.



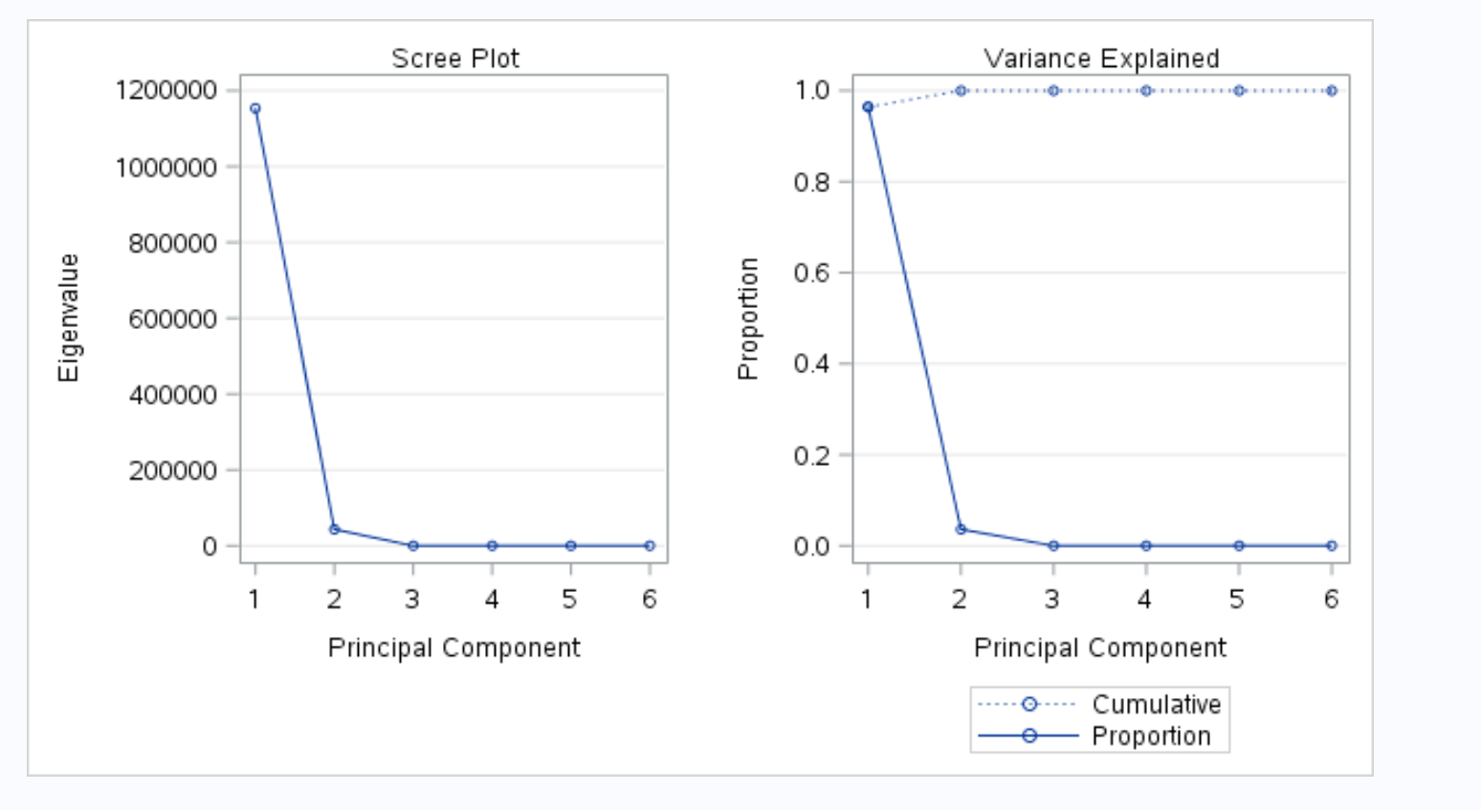
## Analysis Questions

We set out to test the following hypothesis:

1. The odds of Kobe making a shot decreases with the distance from the hoop.
2. The probability of Kobe making a shot decreases linearly with distance.
3. The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs.

### Odds of Making Decreases with Distance

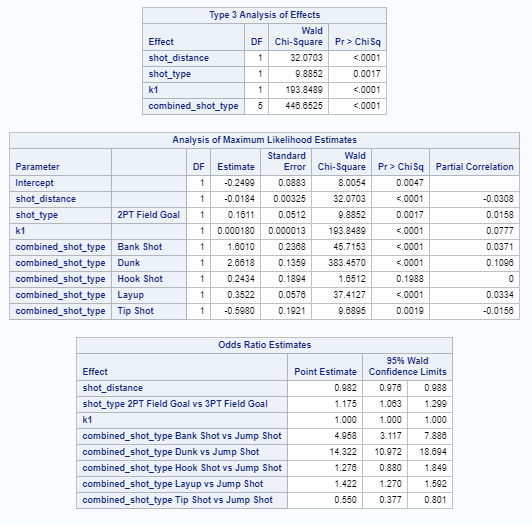
To test these hypothesis we evaluated several models. For the first we evaluated a logistic regression model consisting of shot\_distance, shot\_type, combined\_shot\_type and a linear combination of the continuous variables time\_remaining, attendance, arena\_temp, and avgnoisedb using principal component analysis. Our PCA analysis revealed that the first orthogonal combination contributed nearly 95% of the variance of these variables and the remaining transformations did not contribute significantly.



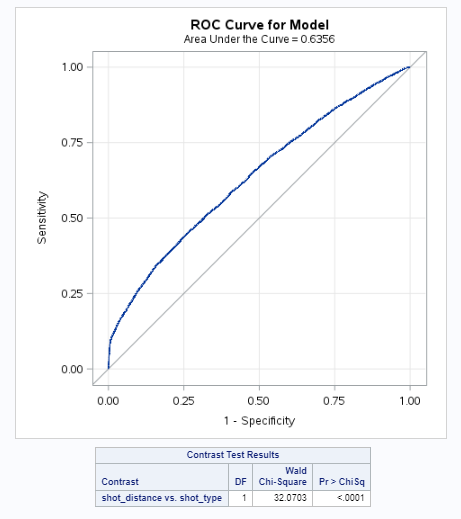
In final SAS model then we only included k1, the first orthogonal combination of PCA analysis.

model shot\_made\_flag(event='1') = shot\_distance shot\_type k1 combined\_shot\_type

Regression analysis revealed that shot\_distance, combined\_shot\_type, shot\_type, and k1 all appear to be statistically signifigant with p-values << 0.



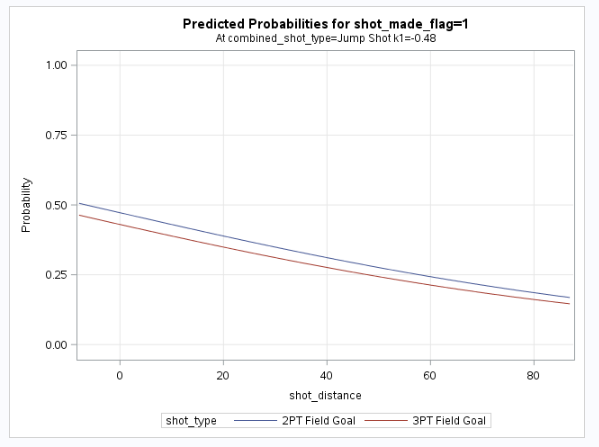
Furthermore, the Wald ChiSquare test statistics (5 df, p-value << 0.001) for combined\_shot\_type and (chisq =8.23 , p-value < .0041) for shot\_type indicates that the overall effect of the categorical variables are independent and also statistically significant. The Hosmer Lemeshow Fit test (ChiSq 40.318, 8df, Pr>ChiSq <.0001) also indicates a good fitting model.



The ROC curve revealed an AUC of .63, which is less than ideal. The usual target for this statistic is around 0.7. The logistic model shows that or a 1 unit increase in shot\_distance, the odds of Kobe making his shot decreases by about 2% (+/- .06% ).

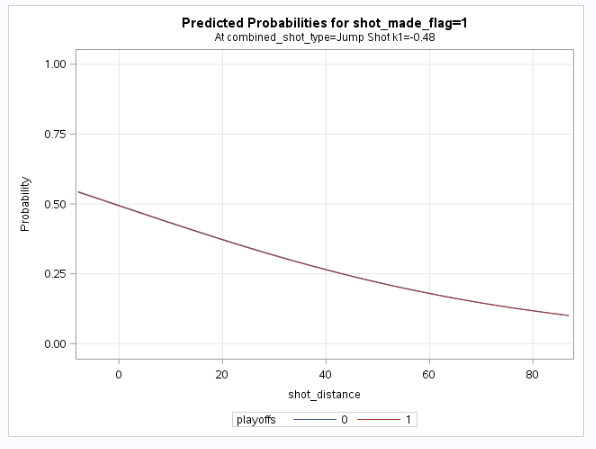
### The probability of making a shot decreases linearly

As part of our analysis we tested the hypothesis that the odds of Kobe making a shot decreases linearly as he is further away from the basket. To assess this we setup a contrast in our model to assess the effect of shot\_distance on a 2pt and 3pt jump\_shot while keeping k1 at its mean. The curve is mostly linear, with a slight asymptotic approach to zero after 40’ indicating that the probability drops off roughly linearly as shot\_distance increases up to 40’, or ~7’ less than half court.



### The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs

To assess this question, we added playoffs to our logistical model and evaluated the relationship of shot\_distance to shots made or missed in the playoffs and in the regular season.



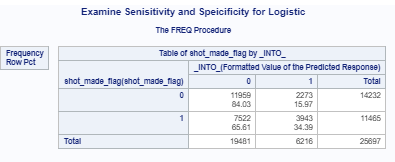
The graphic above shows the same linear relationship we observed before adding playoffs to the model. However, adding playoffs to the model shows adds no statistically significant impact to the probability of making shots (p-value = 0.8470).

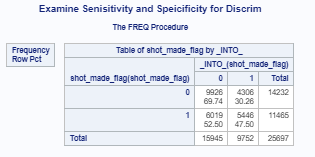
### Alternative Linear Discriminate Analysis Model

As an alternative to the logistic regression model, we also examined a model built upon the Linear Discriminant Analysis technique and compared it to our Linear Regression Model.

#### Sensitivity and Specificity

We compared the models’ Sensitivity and Specificity. The LDA model is more sensitive (47.36% vs 34.3%) but less specific (69.74% va 84.03% ) than Logistic model. Which means the Logistic model is more adept at recognizing missed shots, whereas the LDA model is more adept at identifying made shots.





#### Loss Function

We compared the model’s loss function’s against the . We found out LDA model has a lower log loss, .20022 than the Logistic model,.21342.

|  |  |
| --- | --- |
| Model | Loss |
| Logistic | 0.20022 |
| LDA | 0.21342 |

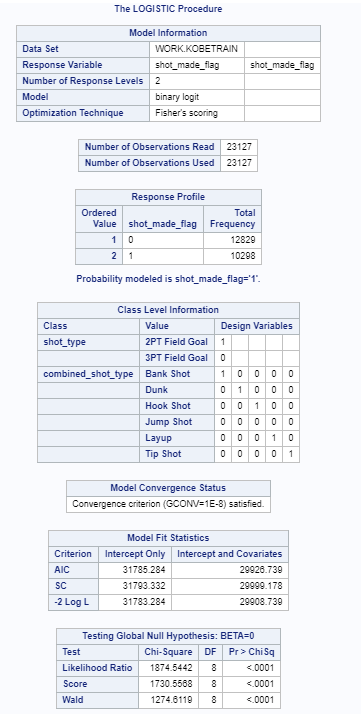
## Conclusions

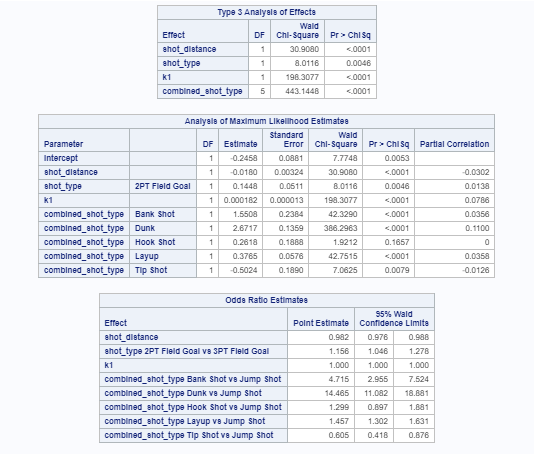
Our analysis concludes that it is difficult to build a model from Kobe’s shot statistics that can reasonably predict whether a shot went in. We showed a logistic regression model was about 34% accurate on our test data with a probability level set at .5 and a LDA model that was 47% accurate at .5 probability level.

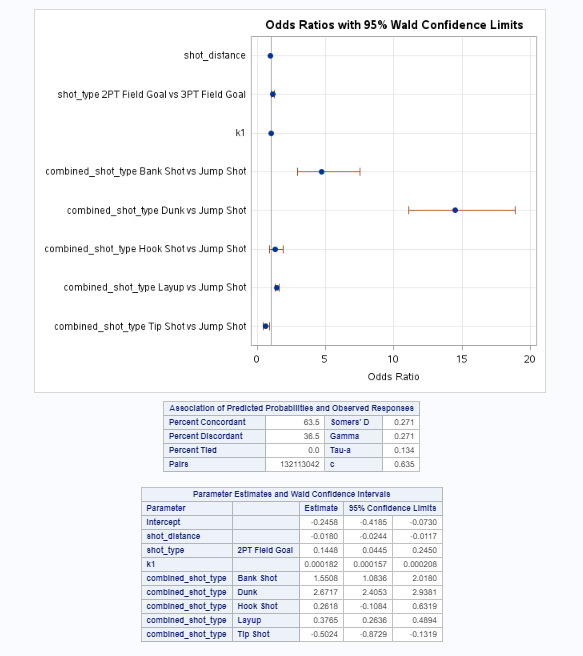
We speculate that additional data points commonly captured in sport statistics such as whether the shot was contested could add specificity to the model. We also do not rule out the need to explore more models based upon the data available in our dataset.

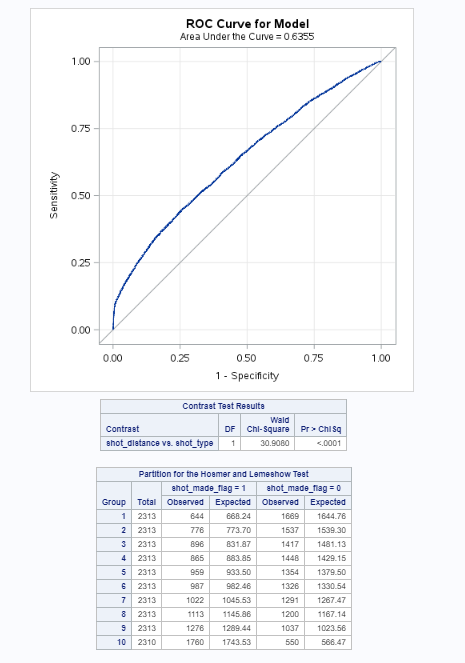
Appendix 1

Logistic Model artifacts:

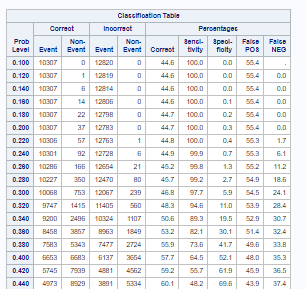


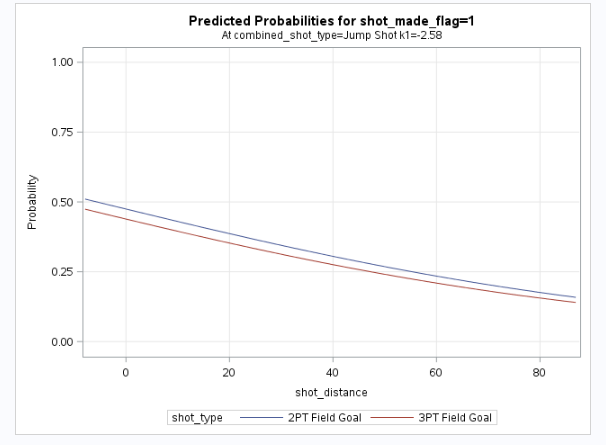




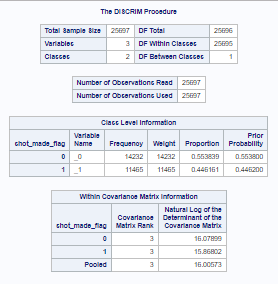


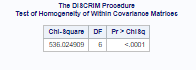
Classification table for initial few data:

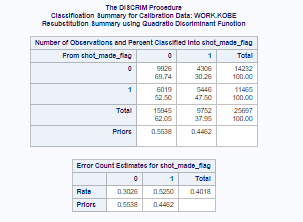




Linear Discrim Analysis artifacts:







Appendix 2

**Code:**

/\* Kobe Data Import \*/

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

FILENAME REFFILE '/folders/myfolders/project2Data.xlsx';

/\* Kobe Predict Data Import \*/

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

FILENAME REFFILE1 '/folders/myfolders/project2Pred.xlsx';

PROC IMPORT DATAFILE=REFFILE

  DBMS=XLSX

  OUT=WORK.kobeshots;

  GETNAMES=YES;

RUN;

PROC IMPORT DATAFILE=REFFILE1

  DBMS=XLSX

  OUT=WORK.kobeshotsPred;

  GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.kobeshotsPred ; RUN;

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

data kobeshots; set kobeshots;

if shot\_zone\_basic = "Above the Break" then szBasic = 4;

if shot\_zone\_basic = "Backcourt" then szBasic = 4;

if shot\_zone\_basic = "In The Paint (Non-RA)" then szBasic = 1;

if shot\_zone\_basic = "Left Corner" then szBasic = 3;

if shot\_zone\_basic = "Mid-Range" then szBasic = 2;

if shot\_zone\_basic = "Restricted Area" then szBasic = 1;

if shot\_zone\_basic = "Right Corner" then szBasic = 3;

if shot\_zone\_range = "24+ ft." then szRange = 3;

if shot\_zone\_range = "8-16 ft." then szRange = 2;

if shot\_zone\_range = "Back Court Shot" then szRange = 4;

if shot\_zone\_range = "Less Than 8ft." then szRange = 1;

if shot\_type = "2PT Field Goal" then shot\_type\_cat = 0;

if shot\_type = "3PT Field Goal" then shot\_type\_cat = 1;

if prxmatch ("/@/",matchup) > 0 then home = 1; else home = 0 ;

run;

/\*\* updated NA data in the prediction file \*/

data kobeshotsPred; set kobeshotsPred;

if shot\_made\_flag eq 'NA' then shot\_made\_flag='';

 shot\_made\_flag\_new = input(shot\_made\_flag,3.0);

  drop shot\_made\_flag;

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

if shot\_zone\_basic = "Above the Break" then szBasic = 4;

if shot\_zone\_basic = "Backcourt" then szBasic = 4;

if shot\_zone\_basic = "In The Paint (Non-RA)" then szBasic = 1;

if shot\_zone\_basic = "Left Corner" then szBasic = 3;

if shot\_zone\_basic = "Mid-Range" then szBasic = 2;

if shot\_zone\_basic = "Restricted Area" then szBasic = 1;

if shot\_zone\_basic = "Right Corner" then szBasic = 3;

if shot\_zone\_range = "24+ ft." then szRange = 3;

if shot\_zone\_range = "8-16 ft." then szRange = 2;

if shot\_zone\_range = "Back Court Shot" then szRange = 4;

if shot\_zone\_range = "Less Than 8ft." then szRange = 1;

if shot\_type = "2PT Field Goal" then shot\_type\_cat = 0;

if shot\_type = "3PT Field Goal" then shot\_type\_cat = 1;

if prxmatch ("/@/",matchup) > 0 then home = 1; else home = 0 ;

run;

/\*\* Transforming the time variables \*/

data kobeshots;

set kobeshots;

time\_remaining = 60\*minutes\_remaining+seconds\_remaining;

run;

data kobeshotsPred;

set kobeshotsPred;

time\_remaining = 60\*minutes\_remaining+seconds\_remaining;

run;

PROC SURVEYSELECT DATA=kobeshots outall OUT=kobe METHOD=srs SAMPRATE=0.1;

RUN;

PROC SURVEYSELECT DATA=kobeshotsPred outall OUT=kobePred METHOD=srs SAMPRATE=0.1;

RUN;

/\*\* Checking correlation between the selected variables \*/

proc sgscatter data=kobe;

matrix lat lon time\_remaining playoffs shot\_distance attendance arena\_temp avgnoisedb / diagonal=(kernel histogram);

run;

/\*\* Checking outliers \*/

proc reg data=kobe ;

model shot\_made\_flag = lat lon time\_remaining period playoffs shot\_distance attendance arena\_temp avgnoisedb / r;

output out=kbCook cookd=cooks student=students rstudent=studresid;

run;

proc sort data=kbCook out=outSortKB;

by descending cooks;

run;

proc print data=outSortKB (obs=10);

var recId cooks;

run;

/\*\* Running principal components for dataset \*/

proc princomp cov prefix=k data=kobe out=kobe;

var time\_remaining attendance arena\_temp avgnoisedb;

run;

proc princomp cov prefix=k data=kobePred out=kobePred;

var time\_remaining attendance arena\_temp avgnoisedb;

run;

PROC SQL;

CREATE TABLE WORK.kobetrain

AS

SELECT

DISTINCT \* FROM WORK.kobe kobe

where kobe.selected = 0;

QUIT;

PROC SQL;

CREATE TABLE WORK.kobetest

AS

SELECT

DISTINCT \* FROM WORK.kobe kobe

where kobe.selected = 1;

QUIT;

proc sort data = kobetrain;

by shot\_type;

run;

/\* shot\_type Analysis - • The odds of Kobe making a shot decrease with respect to the distance he is from the hoop \*/  
proc logistic data = kobetrain plots = all;

class shot\_type combined\_shot\_type(ref='Jump Shot') /param=ref;

model shot\_made\_flag(event='1') = shot\_distance shot\_type k1 combined\_shot\_type

/ ctable lackfit clparm=wald cl pcorr;

  contrast 'shot\_distance vs. shot\_type' shot\_distance 1 -1 0 0 0 0;

output out=logisticOut predprobs=I p=predprob resdev=resdev reschi=pearres;

run;

title "Examine Sensitivity and Speicificity for Logistic";

proc freq data=logisticOut; tables shot\_made\_flag\*\_into\_/nocol nopercent; run;

/\* playoffs Analysis - •  The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs \*/

proc logistic data = kobetrain plots = all;

class shot\_type playoffs /param=ref;

model shot\_made\_flag(event='1') = shot\_distance k1 playoffs combined\_shot\_type

/ ctable lackfit clparm=wald cl pcorr;

contrast 'shot\_distance vs playoffs' shot\_distance 2 0 -1 -1;

output out=logisticOut predprobs=I p=predprob resdev=resdev reschi=pearres;

run;

/\* LDA with Shot\_type principle component analysis and shot\_distance \*/

/\* Predicting for shot\_made\_flag for the prediction file data using LDA \*/

proc discrim data=kobe pool=test testdata=kobetkobePredest testout=discrimPredictKB out=discrimKB crossvalidate ;

class shot\_made\_flag ;

var shot\_distance k1 shot\_type\_cat;

priors '0' = 0.5538 '1' = 0.4462;

run;

title "Examine Sensitivity and Speicificity for Discrim";

proc freq data =discrimKB; tables shot\_made\_flag\*\_into\_/nocol nopercent; run;

/\*\* Predicting for shot\_made\_flag for the prediction file data \*/

proc logistic data = kobe plots = all;

class shot\_type home combined\_shot\_type(ref='Jump Shot') /param=ref;

model shot\_made\_flag(event='1') = shot\_distance shot\_type k1 combined\_shot\_type

/ ctable lackfit clparm=wald cl pcorr pprob=.4 .5 .6;

  contrast 'shot\_distance vs. shot\_type' shot\_distance 1 -1 0 0 0 0 ;

output out=logisticOut predprobs=I p=predprob resdev=resdev reschi=pearres;

Score data=kobePred out = logisticPred;

run;

contrast 'shot\_distance vs. shot\_type' shot\_distance 1 -1 0 0 0 0 ;

output out=logisticOut predprobs=I p=predprob resdev=resdev reschi=pearres;

Score data=kobeshotsPred out = logisticPred;

run;