|  |
| --- |
| DS 6372 |
| Black Mamba |
| Applying Binomial Linear Regression models to predict success of Kobe’s field goals |

|  |
| --- |
| Hannah Kosinovsky, Yat Leung, and Javier Saldaña  4-10-2019 |

Kobe Shot Selection

# Introduction

At the age of 18, Kobe Bryant was a basketball sensation straight out of high school. Having practiced with the Los Angeles Lakers at the age of 17, his parents signed a contract on his behalf. The Lakers and the Hornets organized a trade in which Kobe Bryant would be drafted by the Hornets and then traded to the Lakers before the beginning of the season. Kobe then became the youngest guard ever drafted into the NBA. From that point on, Kobe played his entire career with the Los Angeles Lakers. He went on to earn astonishing records and accolades in professional sports. His talent and production has earned Kobe a legendary status among other great NBA legends. His skills has been studied by many in an attempt to try to replicate his performance to other players.

In this project, we have the ability to study the production of Kobe Bryant’s career. With the given dataset, we look to determine whether distance affected the odds ratio and probability of a successful field goal. In addition, we look to see if that same relationship varies when in a playoff game versus a regular season game. Finally, we take the model and, using stepwise forward selection, attempt to predict whether a field goal would be successful based on the given variables.

# Data Description

The dataset used for this project is sourced from the [Kaggle Kobe Bryant Shot Selection Competition](https://www.kaggle.com/c/kobe-bryant-shot-selection/data). The data set includes every single field goal attempt by Kobe Bryant during his 20-year career with the Los Angeles Lakers. Out of the 30,697 field goal attempts, 5000 (19.5% of entire data set) were extracted into a test data set. The remaining 25,697 attempts were used to train the models.

The data set contains 29 variables, which range between continuous and discrete variables. The variables pertain to information regarding the environment, location, field goal description, and opponents. The variables of interest will be *shot\_made\_flag* (describes whether a shot was successful), *shot\_distance* (distance from shot to basket), and *playoffs* (describes if regular season game or playoff game). Considering each attempt is independent of each other, we will assume independence for this study.

### Detailed Data Content Summary

As mentioned above, this dataset has 25697 observations and 29 variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Value Summary** | **Description** |
| recID | Num | vals 1-30692 | unique id for record |
| action\_type | Chr | ex:("Jump Shot" "Tip Shot") | type of shot taken |
| combined\_shot\_type | Chr | ex:("Jump Shot" “Layup) | type of combined shot taken |
| game\_event\_id | Num | values 2-653 | id of event assoc. with shot |
| game\_id | Num | vals 20000012-49900088 | id of a specific game |
| lat | Num | vals 33.2533-34.0883 | latitude of shot attempt |
| loc\_x | Num | vals -250-248 | (x)location of shot attempt by grid |
| loc\_y | Num | vals -44-791 | (y)location of shot attempt by grid |
| lon | Num | vals -118.5198- -118.0218 | longitude of shot attempt |
| minutes\_remaining | Num | vals 0-11 | num minutes remaining in period during shot |
| period | Num | vals 1-7 | period of game during shot attempt |
| playoffs | Num | vals 0 or 1 | binary: whether or not game is playoff |
| season | Chr | ex("2000-01") | game season ie 2000-2001 when shot occurred |
| seconds\_remaining | Num | vals 0-59 | num seconds remaining in period during shot |
| shot\_distance | Num | vals 0-79 | distance in feet of shot attempt |
| shot\_made\_flag | Num | vals 0 or 1 | binary: whether or not shot was made |
| shot\_type | Chr | ex("2PT Field Goal") | technical term for type of shot |
| shot\_zone\_area | Chr | ex("Left Side(L)") | zone location on court of shot attempt |
| shot\_zone\_basic | Chr | ex("Mid-Range") | approximate zone on court of shot attempt |
| shot\_zone\_range | Chr | ex("8-16 ft." ) | approximate zone on court of shot attempt |
| team\_id | Num | single val: 1610612747 | distance in feet of shot attempt zone location |
| team\_name | Chr | unique val:"Los Angeles Lakers" | NBA team league name |
| game\_date | Num | MMDDYY format | date game occurred |
| matchup | Chr | ex("LAL @ POR") | the two teams playing each other |
| opponent | Chr | ex("POR") | the opposing team of the game |
| shot\_id | Num | vals 2-30697 | id of shot attempt within game |
| attendance | Num | vals 11065-20845 | audience number during game |
| arenatemp | Num | vals 64-79 | temperature of arena during game |
| avgnoisedb | Num | vals 88.56-102.43 | Average noise in db during game |

Using this dataset, no causal inference can be made, since the data set was obtained through observing Kobe Bryant's 20-year professional career. Therefore, the findings can only aid in predicting if he will make the next shot or not and aren’t applicable to other players.

### Exploratory Data Analysis

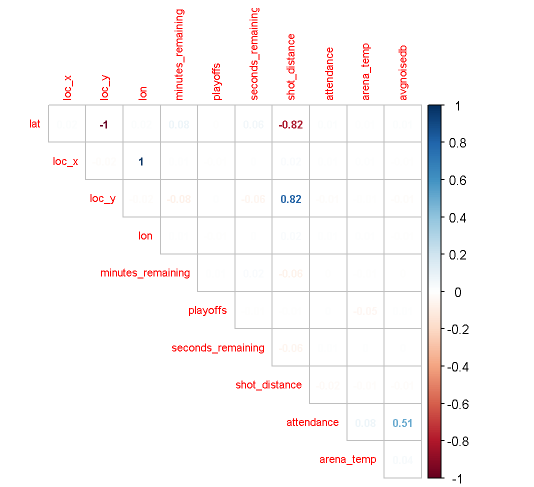
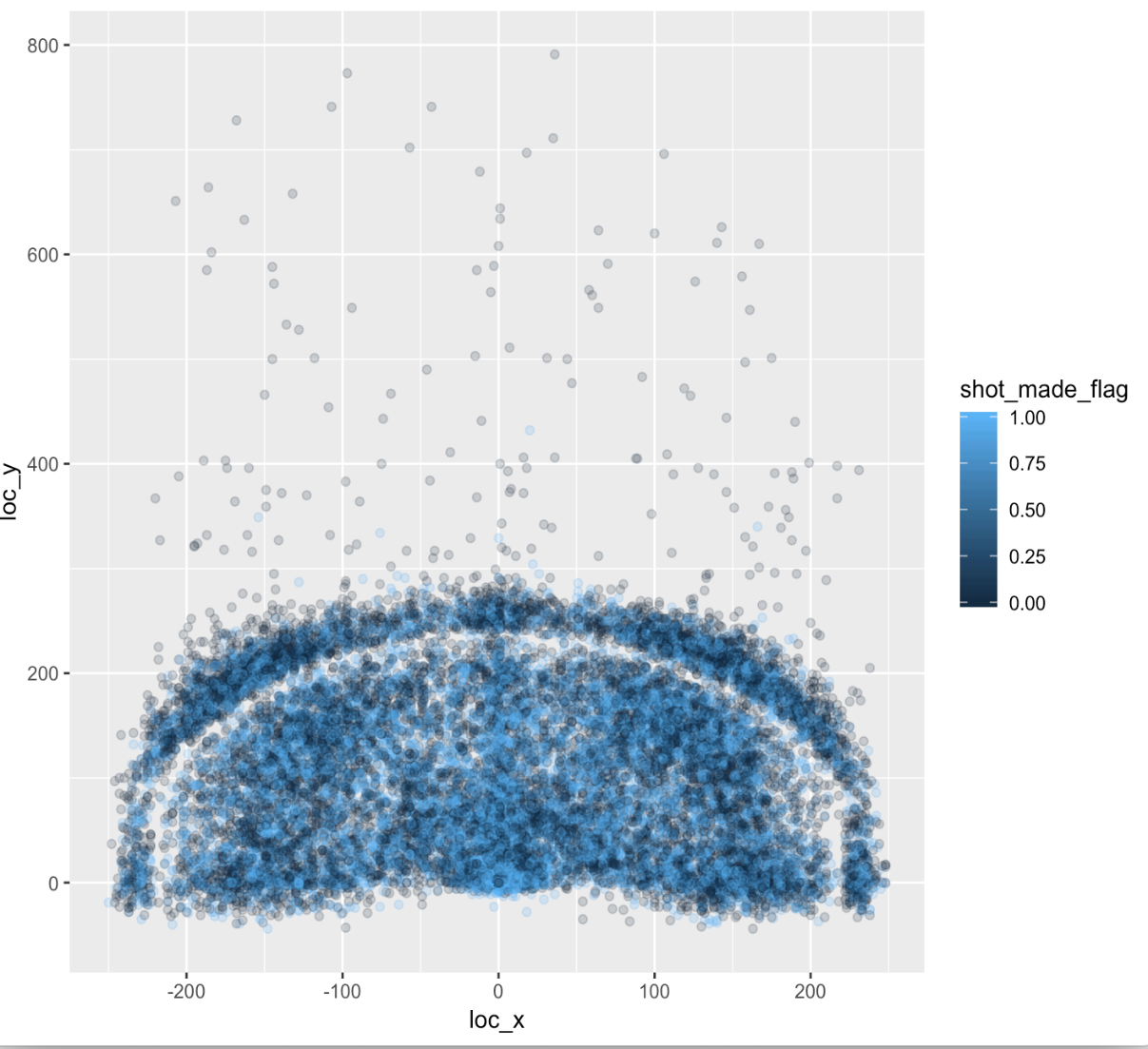
Since our analysis is based on logistic regression, our variables don’t have to be normality distributed to fit in the model, therefore no transformations are necessary.

Figure - Correlation Matrix

In this analysis, multicollinearity between variables is of concern so correlation between variables is examined. From Figure 1, we can see that loc\_y and shot\_distance as well and lat and shot\_distance are highly correlated and therefore we don’t need to include all those variables in the models we create. For our analysis we decided to remove lat from any models we fit to avoid multicollinearity between explanatory variables.

Since our dataset has over 1000 data points and we anticipate the central limit theorem to apply. The main analysis of this project consists of exploring the relationship between the distance and whether a shot was made. For an initial glance at this data, we decided to visualize the court distance locations versus the shot made flag variable. This can be viewed in Figure 2 below:



From looking at figure 2, we can preliminarily see a visual cluster closer around where the hoop would be at location (0,0) but the dark blue(shot not made) and light blue(shot made) indicators exist at all points of the court and don’t appear clustered in a particular way.

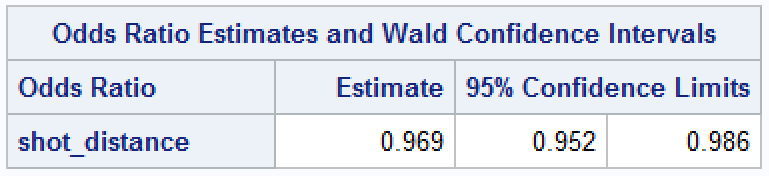
Figure - Plot of shots made and location

# Analysis:

Build models to provide arguments and evidence for or against the propositions below:

## Odds of Field Goal Made v. Distance from Basket

For our first hypothesis, we look to compile a model in which we will look to distance and determine if its effect is significance and what the odds are of a successful field goal. The first model looks at all the variables (regardless of collinearity) and presents an odds ratio of 1.010 with a 95% confidence interval of 0.991 and 1.029. This would suggest that with regards to distance, the odds of a successful field goal is 1.010 with a 95% confidence interval between 0.991 and 1.029. This would suggest the mean odd of a successful field goal increase by 1.010 times with every unit increase of distance. Most importantly, we point toward the fact that the p-value is above 0.05, which would suggest the distance variable is not significant in this model. The results are shown in Figure 3 below.

Full Model Simplified Model

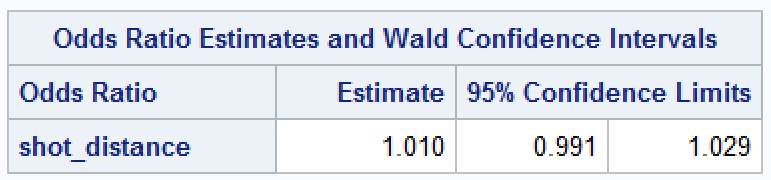
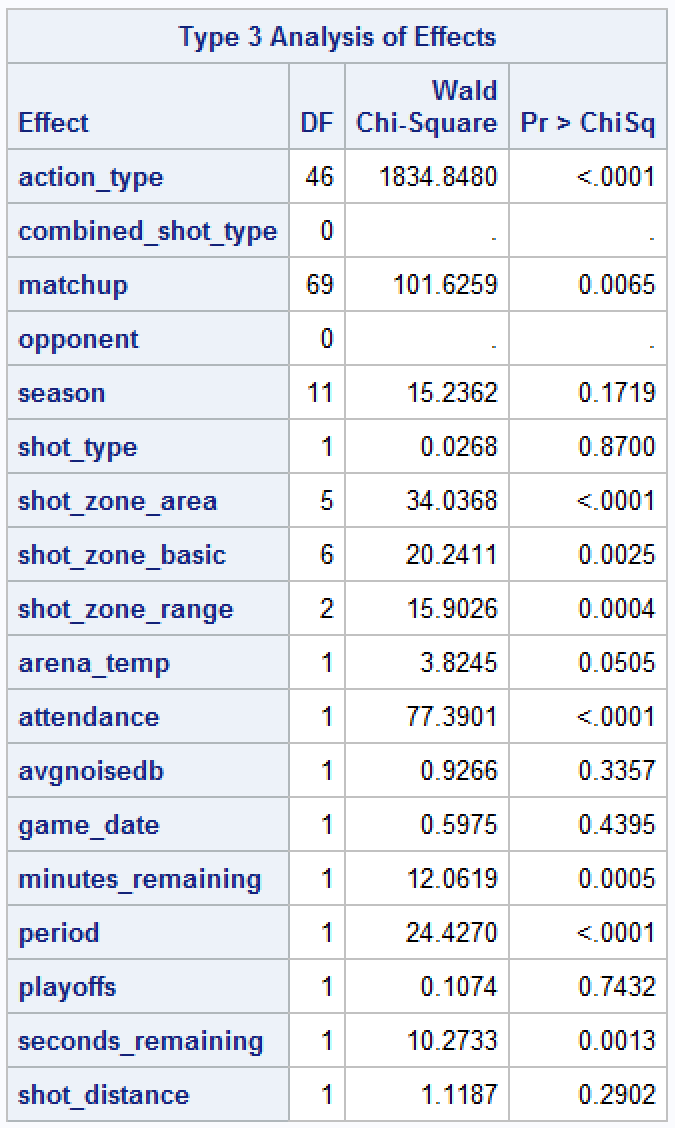
When looking at the simplified model, we see that the time variables have been collapsed into time\_remaining. This was done by converting the minutes\_remaining variable to seconds and then combining the columns. The action\_type was also collapsed and merged with combined\_shot\_type since both were indicators of whether the distance was short or other. Matchup was also reduced to 2 levels instead of the 70 it was previously under. This was performed since the opponent element already captures who the opposing team is, which would indicate the only value the matchup column would provide would be whether the game was home or away. When we performed the binomial regression under this simplified model, we see that the odds ratio, while still relatively close, now illustrates that the mean odds of a successful basket are no longer increasing. Furthermore, the distance is now significant (p-value < 0.05) in the simplified model.

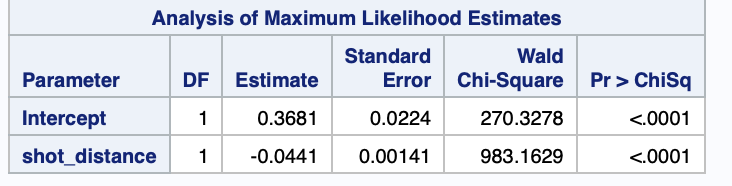
Figure - Odds Ratio Estimate output from SAS

Comparing both models, the simplified model would actually make the more sense. It is much more probable for the odds of a successful basket to actually decrease as the distance increases. While this is especially true for basketball players, athletes like Kobe Bryant are excellent players because they know how to play effectively within their means, which would make them the most efficient players as well. Kobe was well known for creating space between him and the defense in order to be able to improve his odds of making the field goal.

## Probability of Kobe making a shot respect to the distance

For this problem we want to use logistic regression since the response variable is whether or not Kobe makes the shot (binary). We can use a generalized linear model to predict the membership in the target group (whether or not a shot is made). Our goal is to model the probability that shot\_made\_flag = 1. Our x variable is distance so our p(x) modeled in the logistic regression model is the probability of a shot made at distance x.

When looking at the simple logistic regression model estimates in Figure 4, the p-values for the estimates of the intercept and slope are extremely small and therefore significant.



Since these are significant, we know that the probability of making a shot differs with various distances. In this case, the slope is negative, so the larger the distance is, the less chance there is that the shot is made.

Figure - Coefficient Estimate Output from SAS

## Relationship between the distance and odds of making the shot in the playoffs.

To determine the relationship between distance from basket and the odds of making the shot, we approached the problem by analyzing the quantitative continuous variables with component analysis. Without eliminating any potential multicollinear variables, the Scree plot below illustrated that approximately 52% of the variances can be explained with the first two dimensions. On the right, we plotted the variables and their contributions to each dimension to the left. (Figure 5 & 6)

|  |  |
| --- | --- |
| Figure - Scree Plot | Figure - Variables PCA |

Our assumption that lat and loc\_x were multicollinear. With lat removed from the set, we refitted the Scree plot resulted in 48% of the variances can be explained with the first two dimensions. (Figure 7 & 8)

|  |  |
| --- | --- |
| Figure - Scree plot without lat | Figure - Variables PCA without lat |

We performed a Pearson's Chi-squared test for variables independence for the qualitative variables *action\_type* and *combined\_shot\_type*. Our null assumption was that the two categorical variables were independent whereas the alternative would be that the pairs were not independent therefore exhibiting co-linearity. Since the resulting p-value was less than α 0.05, we rejected the null hypothesis that these two variables were independent. (Table 1)

Table

|  |  |  |
| --- | --- | --- |
| **Categorical variables** | | **p-value** |
| action\_type | combined\_shot\_type | < 2.2e-16 |

*X-squared = 128480, df = 270*

In our logistic model, we fitted the coordinations (loc\_x, loc\_y, lon), distance (shot\_distance), playoffs factor, and the distance factors (shot\_zone\_area, shot\_zone\_basic, shot\_zone\_range) then we performed a stepwise procedure to further tapering the variables for a more parsimonious model. Note that loc\_x, loc\_y, lon and playoffs were eliminated from the stepwise procedure

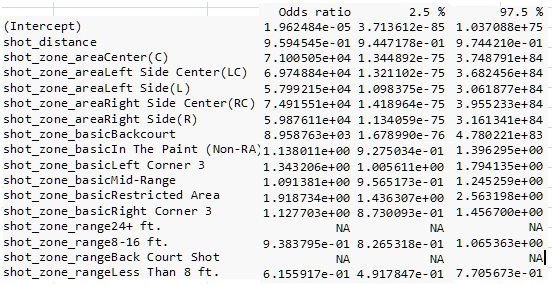
**Base logistic model**

glm(shot\_made\_flag~loc\_x+loc\_y+lon+playoffs+shot\_distance+shot\_zone\_area+shot\_zone\_basic+shot\_zone\_range,family=binomial(link="logit"),data=kobe)

**Stepwise logistic model**

glm(formula = shot\_made\_flag ~ shot\_distance + shot\_zone\_area + shot\_zone\_basic + shot\_zone\_range, family = binomial(link = "logit"), data = kobe)

### Odds Ratio

 Both *shot\_zone\_range24+ ft.* and *shot\_zone\_rangeBack Court Shot* did not have a odds ratio and the 95% CIs because these two variables were perfectly correlated. The odds of Bryant making the shot increased when the shot zone was between 8-16 ft (OR=9.38-01, 95% CI=[8.27-01,1.07] ) and when the shot zone was less than 8 ft (OR=6.16-01, 95% CI=[4.92-01,7.71-01] ). (Table 2)

Table

## Prediction

Using the stepwise model, we predicted Bryant’s odds of making the shot (shot\_made\_flag, binary 1/0). Because the predicted values were the probability values of making the shot (between 0 and 1), therefore we had to explicitly recode the probability values using a threshold of 0.5 (values > 0.5 would be classified as 1 else into 0).

### Sensitivity and Specificity

The sensitivity (True Positive Rate) based on our stepwise model was 0.826588 and the specificity (True Negative Rate) was 0.3344963 (See Figure 10 for the confusion matrix). The calculated area under the curve was 0.6144 (Figure 9) and the calculated log loss function at 13.57258.

|  |  |
| --- | --- |
| Figure - Sensitivity v. Specificity | Figure - Confusion Matrix |

# Code

## SAS Code

ods graphics on;

/\*Perform binomial linear regression using all variables\*/

title 'Binomial Linear Regression on Kobe Field Goals Data with All Relevant Variables';

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

class shot\_made\_flag action\_type combined\_shot\_type matchup opponent season shot\_type shot\_zone\_area shot\_zone\_basic shot\_zone\_range / param = ref;

model shot\_made\_flag (event='1') = action\_type combined\_shot\_type matchup opponent season shot\_type shot\_zone\_area shot\_zone\_basic shot\_zone\_range arena\_temp attendance avgnoisedb game\_date minutes\_remaining period playoffs seconds\_remaining shot\_distance / lackfit ctable;

oddsratio shot\_distance;

run;

/\*Perform binomial linear regression using combined variables\*/

title 'Binomial Linear Regression on Kobe Field Goals Data with Combined Variables';

proc logistic data = KOBEDATA outmodel=model2 plots=all;

class shot\_made\_flag action playoffs home\_away opponent season shot\_type shot\_zone\_area shot\_zone\_basic shot\_zone\_range / param = ref;

model shot\_made\_flag (event='1') = action period playoffs season time\_remaining shot\_distance shot\_type shot\_zone\_area shot\_zone\_basic shot\_zone\_range game\_date opponent attendance arena\_temp avgnoisedb home\_away / lackfit ctable;

oddsratio shot\_distance;

run;

/\*Use Stepwise Regression to find ideal model to predict shots\*/

title 'Stepwise Regression on Kobe Field Goals Data';

proc logistic data = KOBEDATA outest=betas outmodel=model3 plots = all;

class shot\_made\_flag action\_type combined\_shot\_type matchup opponent season shot\_type shot\_zone\_area shot\_zone\_basic shot\_zone\_range / param = ref;

model shot\_made\_flag(event='1') = action\_type combined\_shot\_type matchup opponent season shot\_type shot\_zone\_area shot\_zone\_basic shot\_zone\_range arena\_temp attendance avgnoisedb game\_date minutes\_remaining period playoffs seconds\_remaining shot\_distance

/ selection = stepwise details lackfit ctable;

output out=pred p=phat lower=lcl upper=ucl predprob=(individual crossvalidate);

run;

ods graphics off;

\*print betas and predictions;

proc print data=betas;

title2 'Parameter Estimates and Covariance Matrix';

run;

proc print data=pred;

title2 'Predicted Probabilities and 95% Confidence Limits';

run;

/\*Predict shot using Model 1\*/

proc logistic inmodel = model1;

title 'Kobe Field Goal Predictions Based on Binomial Linear Model\_1';

score data = KOBE\_PREDS out=KobePreds1;

run;

proc print data = KobePreds1;

run;

/\*Predict shot using Model 2\*/

proc logistic inmodel = model2;

title 'Kobe Shot Predictions Based on Binomial Linear Model\_2';

score data = KOBE\_PREDS out=KobePreds2;

run;

proc print data = KobePreds2;

run;

/\*Predict shot using Model 3\*/

proc logistic inmodel = model3;

title 'Kobe Shot Predictions Based on Binomial Linear Model\_3 (Stepwise Regression)';

score data = KOBE\_PREDS out=KobePreds3;

run;

proc print data = KobePreds3;

run;

/\*Perform simple logistic regression for prob of shot made\*/

PROC LOGISTIC DATA = Kobe DESCENDING;

MODEL shot\_made\_flag = shot\_distance / LACKFIT CTABLE;

TITLE 'Kobe Shot Data';

RUN;

## R code

#load data

require(openxlsx)

Kobe = read.xlsx(xlsxFile="~/Downloads/Project2\_2/project2Data.xlsx", sheet = 1, startRow = 1, colNames = TRUE)

str(Kobe)

require(ggplot2)

#visualize relationship bw shot made and distance

ggplot(data = Kobe)+

geom\_point(aes(x = loc\_x, y = loc\_y, color = shot\_made\_flag),alpha = 1 / 5)

library(readxl)

library(tidyverse)

library(aod)

library(caret)

library(glmnet)

library(corrplot)

library(MASS)

library(regclass)

library(FactoMineR)

library(factoextra)

library(pROC)

library(psych)

library(MLmetrics)

xlsx\_kobe <- read\_excel("C:\\Users\\Yat\\Documents\\MSDS\\MSDS 6372\\Project 2\\project2Data.xlsx")

#Take out an ID column from when data was imported

kobe<-xlsx\_kobe[,c(2:29)]

#Re-code the character columns into factors

kobe<-as.data.frame(unclass(kobe))

kobe$shot\_made\_flag<-as.factor(kobe$shot\_made\_flag)

kobe$shot\_type<-as.factor(kobe$shot\_type)

kobe$shot\_zone\_area<-as.factor(kobe$shot\_zone\_area)

kobe$shot\_zone\_basic<-as.factor(kobe$shot\_zone\_basic)

kobe$shot\_zone\_range<-as.factor(kobe$shot\_zone\_range)

kobe$game\_date<-as.factor(kobe$game\_date)

kobe$season<-as.factor(kobe$season)

kobe$period<-as.factor(kobe$period) #change period into factor?

kobe$playoffs<-as.factor(kobe$playoffs)

#subsetting variables (sans IDs)

kobe<-kobe[,c(1:2,5:20,22:25,27:28)]

kobe<-na.omit(kobe)

#data check

summary(kobe)

#numerical variables

kobe.NV<-kobe[,c(3:7,11:12,23:24)]

kobe.NV<-na.omit(kobe.NV)

fit<-prcomp(~., data=kobe.NV, cor=TRUE)

summary(fit) # print variance accounted for

loadings(fit) # pc loadings

plot(fit,type="lines") # scree plot

fit$scores # the principal components

biplot(fit,expand=10, xlim=c(-0.15, 0.05), ylim=c(-0.1, 0.05))

# Varimax Rotated Principal Components

# Extract, rotate and retain 5 PCs

component.retained <- principal(kobe.NV, nfactors=5, rotate="varimax")

component.retained

# Principal Axis Factor Analysis

axis.fit <- factor.pa(kobe.NV, 5)

axis.fit

#PCA

res.pca1 <- prcomp(kobe.NV, scale = TRUE)

fviz\_eig(res.pca1)

fviz\_pca\_var(res.pca1,

col.var = "contrib", # Color by contributions to the PC

gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

repel = TRUE # Avoid text overlapping

)

#numerical variables with lat removed

kobe.NV2<-kobe[,c(4:7,11:12,23:24)]

kobe.NV2<-na.omit(kobe.NV2)

fit2<-prcomp(~., data=kobe.NV2, cor=TRUE)

summary(fit2) # print variance accounted for

loadings(fit2) # pc loadings

plot(fit2,type="lines") # scree plot

fit2$scores # the principal components

biplot(fit2,expand=10, xlim=c(-0.15, 0.05), ylim=c(-0.1, 0.05))

# Varimax Rotated Principal Components

# Extract, rotate and retain 5 PCs

component.retained2 <- principal(kobe.NV2, nfactors=5, rotate="varimax")

component.retained2

# Principal Axis Factor Analysis

axis.fit2 <- factor.pa(kobe.NV2, 5)

axis.fit2

#PCA

res.pca2 <- prcomp(kobe.NV2, scale = TRUE)

fviz\_eig(res.pca2)

fviz\_pca\_var(res.pca2,

col.var = "contrib", # Color by contributions to the PC

gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

repel = TRUE # Avoid text overlapping

)

#action\_type and combined\_shot\_type

tbl.1<-table(kobe$action\_type, kobe$combined\_shot\_type)

chisq.test(tbl.1)

#shot\_made\_flag to distance, playoffs and coordinations

glm.fit<-glm(shot\_made\_flag~loc\_x+loc\_y+lon+playoffs+shot\_distance+shot\_zone\_area+shot\_zone\_basic+shot\_zone\_range,family=binomial(link="logit"),data=kobe)

summary(glm.fit)

#Stepwise based on the first logit regression

stepwise<- stepAIC(glm.fit,direction="both",trace = FALSE)

summary(stepwise)

#odds ratio calculation using library(epiDisplay)

exp(cbind("Odds ratio" = coef(stepwise), confint.default(stepwise, level = 0.95)))

#load the project2Pred.xlsx

pred\_kobe <- read\_excel("C:\\Users\\Yat\\Documents\\MSDS\\MSDS 6372\\Project 2\\project2Pred.xlsx")

#Take out an ID column from when data was imported

pred\_kobe<-pred\_kobe[,c(2:29)]

#Re-code the character columns into factors

pred\_kobe<-as.data.frame(unclass(pred\_kobe))

pred\_kobe$shot\_made\_flag<-as.factor(pred\_kobe$shot\_made\_flag)

pred\_kobe$shot\_type<-as.factor(pred\_kobe$shot\_type)

pred\_kobe$shot\_zone\_area<-as.factor(pred\_kobe$shot\_zone\_area)

pred\_kobe$shot\_zone\_basic<-as.factor(pred\_kobe$shot\_zone\_basic)

pred\_kobe$shot\_zone\_range<-as.factor(pred\_kobe$shot\_zone\_range)

pred\_kobe$game\_date<-as.factor(pred\_kobe$game\_date)

pred\_kobe$season<-as.factor(pred\_kobe$season)

pred\_kobe$period<-as.factor(pred\_kobe$period) #change period into factor?

pred\_kobe$playoffs<-as.factor(pred\_kobe$playoffs)

#subsetting variables (sans IDs)

pred\_kobe<-pred\_kobe[,c(1:2,5:20,22:25,27:28)]

pred\_kobe<-na.omit(pred\_kobe)

#dataframe for the shot\_made\_flag column in the pred\_kobe

pred\_shot\_made\_flag<-pred\_kobe

predictions <- predict(stepwise, pred\_shot\_made\_flag, type="response")

predictions<-as.data.frame(ifelse(predictions>0.5,"1","0"))

colnames(predictions) <- "Predicted shot\_made\_flag"

#actual and predicted values from model

threshold=0.5

predicted\_values<-ifelse(predict(stepwise,type="response")>threshold,1,0)

actual\_values<-stepwise$y

#confusion matrix using the training set

conf\_matrix<-table(predicted\_values,actual\_values)

conf\_matrix

#Sensitivity of the model

sensitivity(conf\_matrix)

#specificity of the model

specificity(conf\_matrix)

predicted\_prob<-predict(stepwise,type="response")

roccurve <- roc(actual\_values, predicted\_prob)

plot(roccurve)

#AUC

auc(roccurve)

#Log Loss Function

LogLoss(predicted\_values,actual\_values)