**Kobe Bryant Shot Selection**

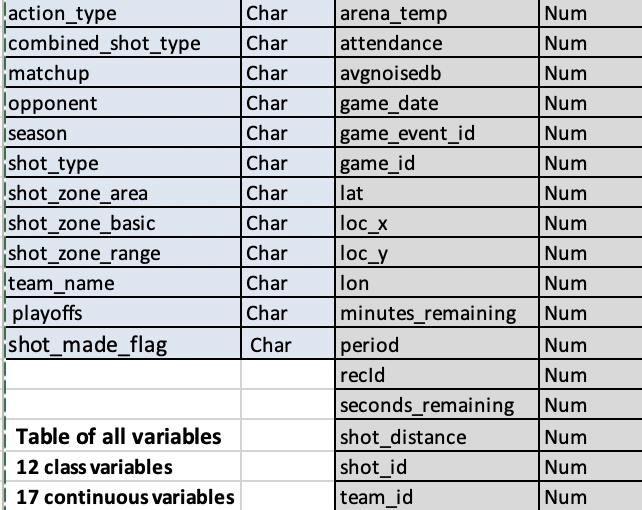
**MSDS 6372 - Project 2**

## Introduction

We have been given a subset of 20 years of data that details different statistics and metrics for each of the attempted shots that Kobe made in his career. We used this data to analyze and make multiple prediction models based on the likelihood of a shoot’s success or failure.

## Data Description

Using 20 years of data on Kobe’s shots made and shots missed, we received the data file project2KobeData.xlsx . In this dataset there are a total 29 variables which contain 12 Characters variables and 17 continuous variables.



## Exploratory Data Analysis

Before exploratory data analysis and checking correlation, we performed a deep dive into the data. Our objectives were to find missing values, impute data, identify outliers, and design strategies to clean and address these issues.

### Missing Values/Imputation

Using SAS we evaluated for missing values, none were missing, so there’s no need to impute data for missing values.

Table of checking missing values

A screenshot of a cell phone

Description automatically generated

### Normality

Plots of checking data normality

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

### 

Our analysis started by visualizing the data into box plots and bar charts to inspect the distribution of variables. Skewness was observed in the above 6 variables, however, samples are large enough, therefore, we may rely on the central limit theorem for normality assumptions in spite of visual indication.

### Data Omission

Additionally, we decided to not consider the following variables (game\_event\_id, game\_id, shot\_id, team\_id) as the values were either auto-generated or another variable provided sufficient information and made them irrelevant thus not improving our model.

### Transformations

We re-coded the below class variables into discrete numerical values for better analysis . We didn’t recode action\_type as it's a subset of combined\_shot\_type. We combined period, minutes\_remaining, seconds\_remaining to time.

### Outliers

Looking at the histogram we can see that shot\_distance is right skewed, additionally only 65 out of 20697 observations had a shot\_distance over 40 feet or roughly 0.3% of all observations. We’ll consider getting rid to outliers for shot\_distance over 40 feet.

### Multicollinearity

After removing the variables no longer in consideration ( action\_type, matchup,opponent,team\_name,game\_data,game\_event\_id,game\_id,lat,

lon, recid, shot\_id,team\_id) , We ran **proc corr** to check multicollinearity for the remaining variables, see the below correlation table:

#### Multicollinearity Analysis

### Strategy of selecting variables

#### Significant variables

Time, shot\_distance, attendance, arena\_temp, avgnoisedb, shottype, years, combinedtype, zonerange, zoneBasic. However, attendance is highly correlated to avgnoisedb and arena\_temp; shot\_distance is highly correlated to loc\_y and slightly correlated to time. Therefore, we will eliminate loc\_y, avgnoisedb, keep shot\_distance and attendance.

#### Selected variables

shot\_distance and attendance were selected, more research was required for time and arena\_temp variables. For those class variables, we’ll add them one at a time into the model and see how the model responds.

## Models

### LRM Shot Distance Model

In order to analyze the relationship between shot\_distance and the likelihood of kobe making a shot we performed a logistic regression model using shot\_distance as a explanatory variable, below you’ll see the results of the model:

Shot\_made\_odds=0.3637-0.0436\*(shot\_distance)

|  |  |
| --- | --- |
| **Analysis of shot\_distance with all distance data** | |
|  |  |
|  |  |
| **After removing outliers over than 40 feet** | |
|  |  |
|  |  |
| **Odds shot distance**   1. There is significant evidence to prove there is a strong relationship between the probability of successful shot\_made\_flag and shot\_distance as p-value is less than .0001. From the above plot, we may see that this relationship is negative with an odds ratio estimate is 0.958. 2. The probability of the shot decreases linearly from distance 0 to 40, after 40 feet the probability stops decreasing linearly. Looking at the histogram we can see that shot\_distance data is right skewed, additionally only 65 out of 20697 observations had a shot\_distance over 40 feet or roughly 0.3% of all observations. We re-ran the model without the outlier observations and as you can see the probability vs shots distance is almost linear when accounting for outliers.   **Model**: Shot\_made\_odds=0.3637-0.0436\*(shot\_distance)  **Interpretation**: For every foot increase of shot\_distance, the estimated odds of probability of Kobe making the shot decreases by a factor of e^-0.0436=0.958, which means the estimated probability of Kobe making a shot decreases (1-0.958=0.042) 4.2% for every unit/foot increase of shot\_distance. A 95% confidence interval for the multiplicative change is (0.955,0.961) or for the percentage decrease is (3.9%, 4.5%) The probability of success decreases by 4.2 for every foot kobe is further away from the net, and the probability of failure increase by 4.2.  Moreover, the odds of success (kobe making the shot) at a distance of 20ft is e^( 0.3545 - 20\*.0427) or about 60.7% whereas the odds of failure is 39.31% | |

### LRM Playoff Model

To analyze if there was a difference in the overall odds of kobe making a shot during playoffs as opposed to regular season we added **playoff** as an additional explanatory variable, below you’ll see the results of the model:

|  |  |
| --- | --- |
|  |  |
| **After removing outliers over than 40 feet** | |
|  |  |
| **After grouping the shots in 4 different categories by distance, see the below plots** | |
|  |  |
| **Odds shot distance playoffs**  When kobe is in a playoff game his overall likelihood increases by 3.6%, regardless of shot distance. However with a Wald Chi-Square statistic of 0.7866 and a p-value of 0.3751, it is not statistically significant, thus we fail to reject the null hypothesis. The confidence interval for the odds ratio agrees with our findings, with a 95% confidence interval of the odds ratio (-4.2%, 12.2%) being between 4.2% less likely to 12.2% more likely of making a shot regardless of distance. Looking at the plots they appear to be very similar, which shows further evidence that the difference is minor and not significant.  To draw a comparison a 20ft shot by Kobe during a playoff game has a 62.9% odds of success whereas during a non playoff game he had 60.7% odds of success | |

### Prediction Models

#### Logistic Regression Model (LRM) Prediction

From the multicollinearity analysis, we can see significant variables are: time, shotdistance, attendance, arena\_temp, avgnoisedb, shottype, years, combinedtype, zonerange, zoneBasic.

However, attendance is highly correlated to avgnoisedb and arena\_temp; shot\_distance is highly correlated to loc\_y and slightly correlated to time. Therefore, we will eliminate loc\_y, avgnoisedb, keep shot\_distance and attendance.

Final selected variables are: shot\_distance, attendance , we will do some more research for variables of time and arena\_temp

Among all significant variables: shottype, years,combinedtype, zonerange, zonebasic are factor variables re-coded to continuous numbers, we’ll enter them one at a time to see how the model responds.

Below are the 3 different types of models:

|  |  |
| --- | --- |
| **Stepwise selection** | **Backward selection** |
|  |  |
|  |  |
| Stepwise selection result: Sensitivity is 42.7 and specificity is 74.6, correct is 60.4  Loss Function: 0.66 with natural log | Backward selection result: sensitivity is 41.1, specificity is 75.9, correct 60.3 |
| **Forward selection** |  |
|  | **ROC Curve from LRM Stepwise Model**Forward selection result: sensitivity is 37.7, specificity is 77.8, correct is 59.8.  After deep consideration: we decided to choose stepwise selection for our logistical model with loss function 0.66 with natural log , see the attachment file ‘logistic loss function’ |

#### Linear Discriminant Analysis (LDA) Prediction model

Our best LDA model used the shot distance, attendance, shot type, and Kobe's year in the league. They was very little collinearity between the variables and it yielded an accuracy of 60.17% and a loss statistic of 0.664 with natural log

|  |  |
| --- | --- |
| **Generalized squared distance to Shot Made** | **Linear Discriminant function for Shots Made** |
|  |  |
| **Classification Summary** | **Error Count Estimates** |
|  |  |
| **Cross-Validation Summary** | **Error Count Estimate for CV** |
|  |  |

##### Sensitivity

Out of 9266 shots made the model predicted correctly 3949 correct giving a sensitivity rate of 42.62%

##### Specificity

Out of 11431 missed shots the model predicted correctly 8504 giving a specificity rate of 74.39%

##### Misclassification

The total misclassification rate or error count was 8244 out of 20697 observations, thus the misclassification rate was 39.87%

##### Loss

The loss function yielded a result of 0.664 with natural log

#### Quadratic Discriminant Analysis (QDA) Prediction model

Additionally we tried running a QDA model with the same explanatory variables, however the model performed marginally worse, with an overall accuracy of 59.73% and a worse fit with loss statistic of 0.672 with natural log

|  |  |
| --- | --- |
| **Generalized squared distance to Shot Made** | **Covariance Matrix Information** |
|  |  |
| **Classification Summary** | **Error Count Estimates** |
|  |  |
| **Cross-Validation Summary** | **Error Count Estimate for CV** |
|  |  |

##### Sensitivity

Out of 9266 shots made the model predicted correctly 4484 correct giving a sensitivity rate of 48.39%

##### Specificity

Out of 11431 missed shots the model predicted correctly 7878 giving a specificity rate of 68.92%

##### Misclassification

The total misclassification rate or error count was 8335 out of 20697 observations, thus the misclassification rate was 40.34%

##### Loss

The loss function yielded a result of 0.672 with natural log

### Comparisons (LRM vs LDA)

We decided to compare our best LRM and LDA models when it comes to sensitivity, specificity, misclassification and loss. As you’ll see below the models performed very similarly on all the metrics.

|  |  |
| --- | --- |
| **ROC Curve for Model from Stepwise Selection** | **ROC Curve for Model from Backward Selection** |
|  |  |
| Stepwise selection looks a little better than the Backward selection. | |

#### Sensitivity

Both the LRM and the LDA model had very similar sensitivities. The best LRM model had a sensitivity of around 42.7% while our best LDA model had a sensitivity of 42.71%

#### Specificity

Similar to sensitivity both models had comparable specificity, with the LRM model having a specificity of 74.6% and the QDA model having a specificity of 74.5%

#### Misclassification

The misclassification for each was 39.6 and 39.7 for the LRM and LDA model respectively. Again both models fared very similarly.

#### Loss

Both the LRM and LDA models had almost equivalent loss statistics of .66 with natural log, meaning the both had very similar fits.

### Conclusion

After looking at 20 years of data of Kobe's playing career we ended up analyzing the relationship that distance, attendance, shot type and years in the league had on the probability of Kobe making a shot. We developed two competing models, one Logistic regression and one Linear discriminant model to predict the odds of kobe making a specific shot. Both models ended up having very similar characteristics and fit with a loss statistic of 0.66 with natural log.

|  |
| --- |
| **Appendix: Code**  **proc** **import** datafile='\\smu.edu\Files\users$\jeolmos\Apps.SMU\Documents\6372\project\_2\project2KobeData.csv'  out=kobe\_data  dbms=csv  replace;  guessingrows=max;  **run**;  **proc** **import** datafile='\\smu.edu\Files\users$\jeolmos\Apps.SMU\Documents\6372\project\_2\predData.csv'  out=pred\_data  dbms=csv  replace;  guessingrows=max;  **run**;  Data prep  **data** kobe\_data;  set kobe\_data;  time = (period-**1**)\***12**\***60**+(**12**\***60**-(minutes\_remaining\***60**+seconds\_remaining));  **run**;  **data** pred\_data;  set pred\_data;  time = (period-**1**)\***12**\***60**+(**12**\***60**-(minutes\_remaining\***60**+seconds\_remaining));  **run**;  **data** kobe\_data;  set kobe\_data;  if shot\_distance <= **10** then shotdistance = **1**;  if shot\_distance <= **20** and shot\_distance > **10** then shotdistance = **2**;  if shot\_distance <= **30** and shot\_distance > **20** then shotdistance = **3**;  if shot\_distance > **30** then shotdistance = **4**;  /\*recode shot\_zone\_basic\*/  if shot\_zone\_basic = "Above the Break" then zoneBasic = **4**;  if shot\_zone\_basic = "Backcourt" then zoneBasic = **4**;  if shot\_zone\_basic = "In The Paint (Non-RA)" then zonebasic = **1**;  if shot\_zone\_basic = "Left Corner" then zonebasic = **3**;  if shot\_zone\_basic = "Mid-Range" then zonebasic = **2**;  if shot\_zone\_basic = "Restricted Area" then zonebasic = **1**;  if shot\_zone\_basic = "Right Corner" then zonebasic = **4**;    /\*recode shot\_zone\_range\*/  if shot\_zone\_range = "Less Than 8ft." then zonerange = **1**;  if shot\_zone\_range = "8-16 ft." then zonerange = **2**;  if shot\_zone\_range = "16-24 ft." then zonerange = **3**;  if shot\_zone\_range = "24+ ft." then zonerange = **4**;  if shot\_zone\_range = "Back Court Shot" then zonerange = **4**;  /\*recode action\_type\*/  /\*recode combined\_shot\_type\*/  if combined\_shot\_type="Bank Shot" then combinedtype=**1**;  if combined\_shot\_type="Dunk" then combinedtype=**2**;  if combined\_shot\_type="Hook Shot" then combinedtype=**3**;  if combined\_shot\_type="Jump Shot" then combinedtype=**4**;  if combined\_shot\_type="Layup" then combinedtype=**5**;  if combined\_shot\_type="Tip Shot" then combinedtype=**6**;  /\*recode matchup-- too many counts, hard to recode\*/  /\*recode opponent -- too many counts of opponents, hard to recode all\*/  /\*recode season\*/  if season="1996-97" then years=**0**;  if season="1997-98" then years=**1**;  if season="1998-99" then years=**2**;  if season="1999-00" then years=**3**;  if season="2000-01" then years=**4**;  if season="2001-02" then years=**5**;  if season="2002-03" then years=**6**;  if season="2003-04" then years=**7**;  if season="2004-05" then years=**8**;  if season="2005-06" then years=**9**;  if season="2006-07" then years=**10**;  if season="2007-08" then years=**11**;  if season="2008-09" then years=**12**;  if season="2009-10" then years=**13**;  if season="2010-11" then years=**14**;  if season="2011-12" then years=**15**;  if season="2012-13" then years=**16**;  if season="2013-14" then years=**17**;  if season="2014-15" then years=**18**;  if season="2015-16" then years=**19**;  /\*recode shot\_type\*/  if shot\_type="2PT Field Goal" then shottype=**2**;  if shot\_type="3PT Field Goal" then shottype=**3**;  /\*recode shot\_zone\_area\*/  if shot\_zone\_area="Left Side(L)" then area=**1**;  if shot\_zone\_area="Center(C)" then area=**2**;  if shot\_zone\_area="Right Side Center(RC)" then area=**3**;  if shot\_zone\_area="Left Side Center(LC)" then area=**4**;  if shot\_zone\_area="Right Side(R)" then area=**5**;  if shot\_zone\_area="Back Court(BC)" then area=**6**;  **run**;  **data** pred\_data;  set pred\_data;  if shot\_distance <= **10** then shotdistance = **1**;  if shot\_distance <= **20** and shot\_distance > **10** then shotdistance = **2**;  if shot\_distance <= **30** and shot\_distance > **20** then shotdistance = **3**;  if shot\_distance > **30** then shotdistance = **4**;  /\*recode shot\_zone\_basic\*/  if shot\_zone\_basic = "Above the Break" then zoneBasic = **4**;  if shot\_zone\_basic = "Backcourt" then zoneBasic = **4**;  if shot\_zone\_basic = "In The Paint (Non-RA)" then zonebasic = **1**;  if shot\_zone\_basic = "Left Corner" then zonebasic = **3**;  if shot\_zone\_basic = "Mid-Range" then zonebasic = **2**;  if shot\_zone\_basic = "Restricted Area" then zonebasic = **1**;  if shot\_zone\_basic = "Right Corner" then zonebasic = **4**;    /\*recode shot\_zone\_range\*/  if shot\_zone\_range = "Less Than 8ft." then zonerange = **1**;  if shot\_zone\_range = "8-16 ft." then zonerange = **2**;  if shot\_zone\_range = "16-24 ft." then zonerange = **3**;  if shot\_zone\_range = "24+ ft." then zonerange = **4**;  if shot\_zone\_range = "Back Court Shot" then zonerange = **4**;  /\*recode action\_type\*/  /\*recode combined\_shot\_type\*/  if combined\_shot\_type="Bank Shot" then combinedtype=**1**;  if combined\_shot\_type="Dunk" then combinedtype=**2**;  if combined\_shot\_type="Hook Shot" then combinedtype=**3**;  if combined\_shot\_type="Jump Shot" then combinedtype=**4**;  if combined\_shot\_type="Layup" then combinedtype=**5**;  if combined\_shot\_type="Tip Shot" then combinedtype=**6**;  /\*recode matchup-- too many counts, hard to recode\*/  /\*recode opponent -- too many counts of opponents, hard to recode all\*/  /\*recode season\*/  if season="1996-97" then years=**0**;  if season="1997-98" then years=**1**;  if season="1998-99" then years=**2**;  if season="1999-00" then years=**3**;  if season="2000-01" then years=**4**;  if season="2001-02" then years=**5**;  if season="2002-03" then years=**6**;  if season="2003-04" then years=**7**;  if season="2004-05" then years=**8**;  if season="2005-06" then years=**9**;  if season="2006-07" then years=**10**;  if season="2007-08" then years=**11**;  if season="2008-09" then years=**12**;  if season="2009-10" then years=**13**;  if season="2010-11" then years=**14**;  if season="2011-12" then years=**15**;  if season="2012-13" then years=**16**;  if season="2013-14" then years=**17**;  if season="2014-15" then years=**18**;  if season="2015-16" then years=**19**;  /\*recode shot\_type\*/  if shot\_type="2PT Field Goal" then shottype=**2**;  if shot\_type="3PT Field Goal" then shottype=**3**;  /\*recode shot\_zone\_area\*/  if shot\_zone\_area="Left Side(L)" then area=**1**;  if shot\_zone\_area="Center(C)" then area=**2**;  if shot\_zone\_area="Right Side Center(RC)" then area=**3**;  if shot\_zone\_area="Left Side Center(LC)" then area=**4**;  if shot\_zone\_area="Right Side(R)" then area=**5**;  if shot\_zone\_area="Back Court(BC)" then area=**6**;  **Run**;  /\*get rid of outlier of distance\*/  **data** clean;  set kobe\_data1;  if shot\_distance>**40** or shot\_distance=**40** then delete;  /\*Check the odds and probability between shot\_made\_flag and shot\_distance \*/  proc logistic data =clean descending outest=betas covout plots=all;  class shot\_made\_flag/ param = ref;  model shot\_made\_flag(Event='1') =shot\_distance/ctable pprob = .5;  output out = logisticOut predprobs=I p=predprob lower=lcl upper=ucl resdev=resdev reschi=pearres;  run;  /\* Get logistic predictions and Building a logistic regression model with significant variables\*/  proc logistic data=clean order=data plots=all;  class years/param=ref;  model shot\_made\_flag(event='1') = shot\_distance attendance years shottype/selection=stepwise ctable pprob= .5;;  output out=logisticout;  score data=pred\_data1 out=predlogisticout;  run;  /\*LDA \*/  **proc** **discrim** data=kobe\_data pool=yes crossvalidate;  class shot\_made\_flag;  var shot\_distance attendance shottype years;  priors "0" = **.5523** "1" =**.4477**;  **run**;  /\*QDA \*/  **proc** **discrim** data=kobe\_data pool=no crossvalidate;  class shot\_made\_flag;  var shot\_distance attendance shottype years;  priors "0" = **.5523** "1" =**.4477**;  **run**;  /\*train and test data set\*/  %let inputDataset = kobe\_data;  %let numObs = 20698; \*\*\* number of observations + 1;  **data** inDat; set &inputDataset; randNumber = ranuni(**11**); if \_n\_ < &numObs; **run**;  **data** train; set inDat; if randNumber <= **1**/**4** then delete; **run**;  **data** test; set inDat; if randNumber > **1**/**4** then delete; **run**;  /\*LDA with train and test dataset\*/  **proc** **discrim** data=train pool=yes crossvalidate testData=test testout=discrimOutLDA;  class shot\_made\_flag;  var shot\_distance attendance shottype years ;  priors "0" = **.5523** "1" =**.4477**;  **run**;  /\*QDA with train and test dataset\*/  **proc** **discrim** data=train pool=no crossvalidate testData=test testout=discrimOutQDA;  class shot\_made\_flag;  var shot\_distance attendance shottype years ;  priors "0" = **.5523** "1" =**.4477**;  **run**;  /\*Loss Calculation\*/  **data** discrimOutLDA;  set discrimOutLDA;  loss = shot\_made\_flag\*log(\_1)+(**1**-shot\_made\_flag)\*log(\_0);  **run**;  **proc** **summary** data=discrimOutLDA;  var loss;  output out=totals sum=;  **run**;  **proc** **print** data = totals;  **run**;  /\*saving loss data\*/  **data** discrimOutLDA;  set discrimOutLDA;  pred = \_INTO\_;  **Run**;  **data** loss\_data;  set discrimOutLDA;  KEEP recId shot\_made\_flag \_1 \_0 pred loss;  **run**;  **proc** **export** data=loss\_data  outfile='\\smu.edu\Files\users$\jeolmos\Apps.SMU\Documents\6372\project\_2\loss\_data.csv'  dbms=csv;  delimiter=',';  **Run**;  /\*Calculating prediction data\*/  **proc** **discrim** data=kobe\_data pool=yes crossvalidate testData=pred\_data testout=discrimOutPred;  class shot\_made\_flag;  var shot\_distance attendance shottype years;  priors "0" = **.5523** "1" =**.4477**;  **run**;  /\*saving prediction data\*/  **data** discrimOutPred;  set discrimOutPred;  shot\_made\_flag = \_INTO\_;  **run**;  **data** lda\_prediction;  set discrimOutPred;  KEEP recId shot\_made\_flag;  **run**;  **proc** **contents** data = lda\_prediction;  **run**;  **proc** **export** data=lda\_prediction  outfile='\\smu.edu\Files\users$\jeolmos\Apps.SMU\Documents\6372\project\_2\lda\_prediction.csv'  dbms=csv;  delimiter=',';  **run**; |