

# Will Kobe Bryant Make His Next Shot: Linear Discriminant Analysis and Logistic Regression using R

*Paul Adams*

*Reannan McDaniel*

*Jeff Nguyen*

*Southern Methodist University*

*28 November 2019*

## Abstract:

*This project investigates the correlation between multiple potential explanatory variables and Kobe Bryant's ability to make a shot while playing for the NBA team Los Angeles Lakers using data gathered from 1996-2015.*

## Exploratory Data Analysis

### Outlier Check

First, we performed a brief outlier check indicated a 2PT Field Goal was noted from the 3PT (3-point) range. Regardless of the actual score, the location matters more. Therefore, after confirming these values to be within 3PT-range, all shots will be encoded as 3-point once beyond 300 inches.

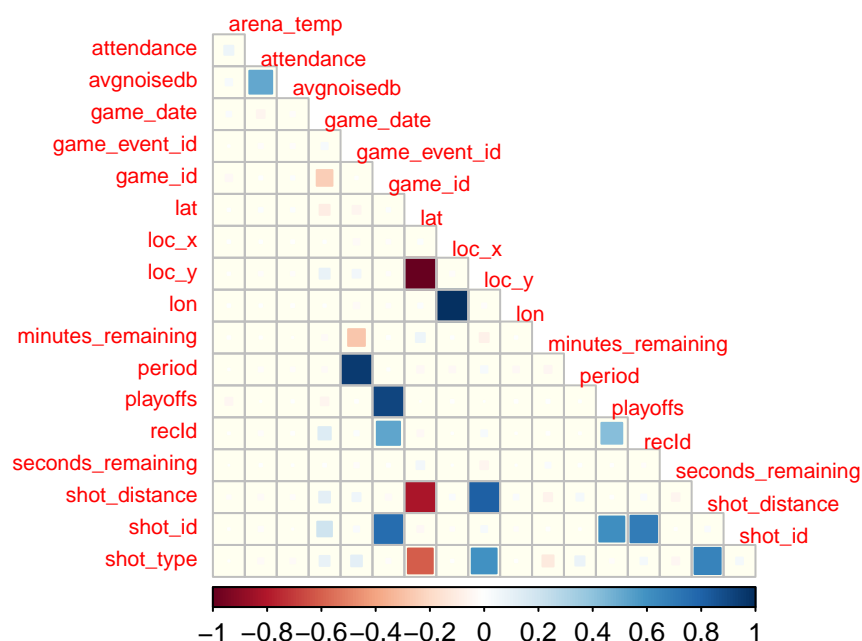
### Variable Elimination

Next, we removed one-level factors. These will never change so are not useful to the model; including can cause issues with model sensitivity since linear trajectories will be down-weighted. Therefore, their significance will be lessened by the constant state of the additional parameters. While this is may not be significant, it is not condusive to model quality.

### Addressing Multicollinearity: Correlation Plot for Visual Data Exploration

To address multicollinearity among quantitative predictor variables, we used a correlation matrix.

### Correlation among Predictor Variables



## Post-Correlation Plot Variable Elimination

Following our correlation plot, we decided to eliminate some collinear terms. However, some of the collinearity is useful to capture the instances where the terms are unique. For example, `combined_shot_type` (factor variable) is collinear with `shot_distance` (quantitative variable), but it also accounts for the method Kobe may use to make a shot. For example, distance may be relatively the same between 10 and 11 feet, but the factor levels used to derive their short or far indications may differ. This difference could be whether Kobe makes a potentially more accurate heel-planted shot or if he is forced to lean forward and take a riskier shot at basket; the difference in distance may only be one foot, but the difference in technique could measure significant relative to the odds of success.

## Addressing Multicollinearity: Correlation Matrix for Numerical Analysis

###Following the removal of the most obvious collinear terms visually performing a correlation plot analysis, a correlation matrix for analyzing the remaining results. Collinear quantitative data was preliminarily removed following correlation plot analysis to desaturate the model to an extent that allows more distinction among significance measures for terms in the correlation matrix.

Row	Column	Correlation	p-value	NA
arena_temp	arena_temp	0.510916827054281	0	0
attendance	arena_temp	0.510916827054281	0	0
game_date	arena_temp	0.510916827054281	0	0
game_event_id	arena_temp	0.510916827054281	0	0
loc_x	arena_temp	0.510916827054281	0	0
loc_y	arena_temp	0.510916827054281	0	0
minutes_remaining	arena_temp	0.510916827054281	0	0
playoffs	arena_temp	0.510916827054281	0	0
recId	arena_temp	0.510916827054281	0	0
seconds_remaining	arena_temp	0.510916827054281	0	0

## After the first round of

## Quadratic Discriminant Analysis

As requested within the requirements of this study, a Linear Discriminant Analysis must be assessed and provided. Discriminant analysis is an operation that compares a categorical response variable against measures of quantitative predictor variables. As a result, analysis for this section is performed on the numerical predictors, which include `recId`, `game_event_id`, `game_id`, `loc_x`, `loc_y`, `minutes_remaining`, `seconds_remaining`, `shot_distance`, `shot_made_flag`, `shot_type`, `game_date`, `shot_id`, `attendance`, `arena_temp`, `avgnoisedb`, controlling collinearity by eliminating a member of each collinear pair prior to model development.

###Linear Discriminant Analysis requires a linear boundary between the predictor variables, respective of the response. If the boundary between predictors and response is not linear, Quadratic Discriminant Analysis must be used. Wilks' Lambda distribution is used to assess the nature of boundary linearity, which can be used for discriminant analysis. However, because of the large dimensions of the data set analyzed in this study, an approximation of Wilks' Lambda must be used. Bartlett's Test is an approximation of Wilks'

Lambda that can be used for models with large dimensions by applying a measure against the Chi-Square distribution. Provided is a test statistic and p-value.

#### Bartlett's Test:

	Chi_Square_Statistic	Degrees_of_Freedom	Wilks_Lambda	p_value
Wilks' Lambda	1037.243	14	0.9510987	p < 0.0001

Bartlett's Test of this data set yielded a significant p-value, where  $p < 0.0001$ , indicating that the proportion of distribution beyond the derived test statistic is beyond that which could be explained by chance. Therefore, we must reject the null hypothesis that the boundary for analysis is linear; the boundary is non-linear. Thus, an analysis using Quadratic Discriminant Analysis is applied.

```
## mean.kobe.qda.posterior...1.. mean.kobe.qda.posterior...2..
## 1                0.4578203                0.5421797

## shot_made_flag_Posterior proportion_Posterior
## 1                0                0.4578203
## 2                1                0.5421797
```

#### Quadratic Discriminant Analysis Misclassification Rate

x
40.083%

#### Log Loss

```
## function (actual, predicted)
## {
##   return(mean(ll(actual, predicted)))
## }
## <bytecode: 0x0000000022e84290>
## <environment: namespace:Metrics>
```

#### Confusion Matrix & Metrics Table

Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	Precision	Accuracy	Misclassification	Logarithmic_Loss
0.51187	0.66993	0.55695	0.62868	0.55695	0.59917	0.40083	0.70127

## Predictions from Quadratic Discriminant Analysis

## Logistic Model Development using Ordinary Least Squares

A preliminary, manual variable elimination process was performed during the analysis of multicollinear terms in preparation for model development. Below we perform logistic regression using Ordinary Least Squares (OLS). In preparation for the model development, a starting model and a finishing model must be developed to provide the scope of variable selection.

### Forward Selection

Forward selection produced a model that produced an Akaike's Information Criterion score of 27,378.

#### Forward Selection Model:

$$\text{shot}_m\text{ade}flag = \text{shotdistance} + \text{attendance} + \text{combinedshottype} + \text{arenatemp} + \text{gameeventid} + \text{secondsremaining} + \text{shottype} + \text{gamedate} + \text{minutesremaining} + \text{locy} + \text{shotid}$$

#### Forward Selection - Akaike's Information Criterion for Logistic Regression:

Akaikes.Information.Criterion..Foreward.Selection
26824.26

### Backward Elimination

Backward elimination produced a model that produced an Akaike's Information Criterion score of 27,378.

#### Backward Elimination Model:

$$\text{shotmade}flag = \text{combinedshottype} + \text{gameeventid} + \text{locy} + \text{minutesremaining} + \text{secondsremaining} + \text{shotdistance} + \text{shottype} + \text{gamedate} + \text{shotid} + \text{attendance} + \text{arenatemp}$$

#### Backward Elmination - Akaike's Information Criterion for Logistic Regression:

Akaikes.Information.Criterion..Backward.Elimination
26824.26

## Stepwise Regression

Stepwise Regression produced a model that produced an Akaike's Information Criterion score of 27,378.

### Stepwise Regression Model:

$\text{shotmadeflag} = \text{combinedshotttype} + \text{gameeventid} + \text{locy} + \text{minutesremaining} + \text{secondsremaining} + \text{shotdistance} + \text{shotttype} + \text{gamedate} + \text{shotid} + \text{attendance} + \text{arenatemp}$

### Stepwise Regression - Akaike's Information Criterion for Logistic Regression:

Akaikes.Information.Criterion..Stepwise.Regression	
	26824.26