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

Paul Adams
Reannan McDaniel
Jeff Nguyen
Southern Methodist University

29 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, which included a graphical analysis of all shots taken, by loc__x and loc__y. This graphical analysis indicated a 2PT (2-point) Field Goal was recorded from the 3PT (3-point) range. Upon inspection of other attributes - such as action type and shot__zone__range - we verified this shot to be a member member of the 3-point level of shot__type. Under the assumption shots from beyond the 300 inch mark are more likely to have been incorrectly recorded as 2 points rather than an incorrectly recorded location y, we modified our programming to transform all shots where $loc_y > 300$ to be recoded as `3PT Field Gloal`.

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, a correlation heat map was created for visual inspection of correlation.

Correlation among Predictor Variables

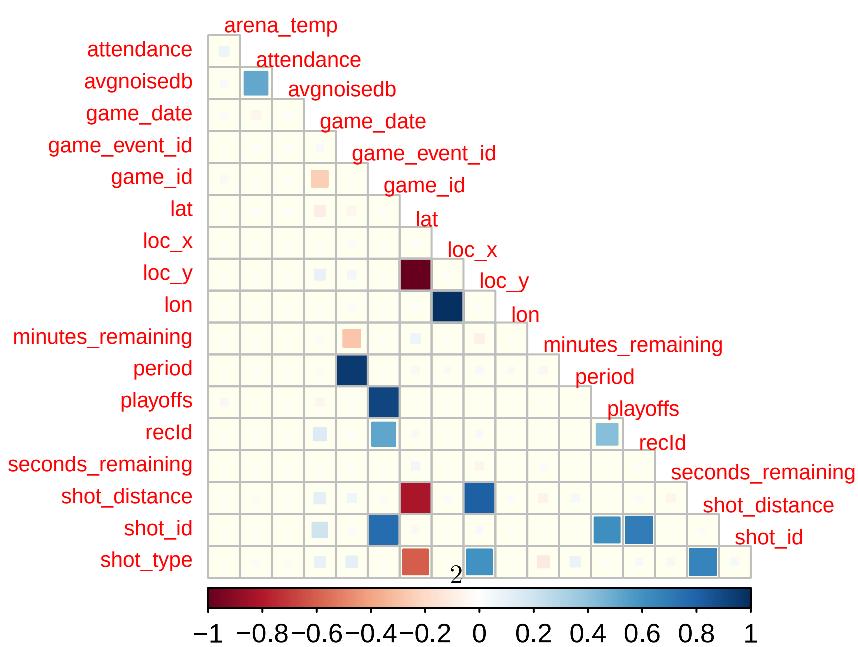

Table 1: Top 10 Collinear Terms

Correlation Predictor Variable	Correlation Response Variable	Correlation	p-Value
arena_temp	arena_temp	0.51092	$p < 0.0001$
attendance	arena_temp	0.51092	$p < 0.0001$
game_date	arena_temp	0.51092	$p < 0.0001$
game_event_id	arena_temp	0.51092	$p < 0.0001$
loc_x	arena_temp	0.51092	$p < 0.0001$
loc_y	arena_temp	0.51092	$p < 0.0001$
minutes_remaining	arena_temp	0.51092	$p < 0.0001$
playoffs	arena_temp	0.51092	$p < 0.0001$
recId	arena_temp	0.51092	$p < 0.0001$
seconds_remaining	arena_temp	0.51092	$p < 0.0001$

Post-Correlation Heat Map Variable Elimination

Following our correlation heat map, 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 derrive 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.

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` (QDA) must be used. `Wilks' Lambda` distribution is used to assess the nature of boundary linearity, which is a required understanding to develop a well-fit discriminant classification model. However, because of the large dimensions of the data set analyzed in this study, an approximation of Wilks' Lambda must

Table 2: Bartlett Test's Wilks' Lambda Approximation

	Metric Output
Chi Square Statistic	1037.24251
Degrees of Freedom	14
Wilks' Lambda	0.9511
p-Value	$p < 0.0001$

be used, rather than Wilks' Lambda itself. `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`. This method is applied herein to assess linearity. ## **Bartlett's Test:**

The result of this test returned statistically significant results, indicating the null hypothesis of linearity must be rejected in favor of the alternate, which is that the discriminant boundary is non-linear. Consequently, we proceed with a model based on `Quadratic Discriminant Analysis` to provide predictive responses from a discriminant model. However, we proceed with caution, as the quadratic version of the discriminant analysis is at greater risk for over-fitting to the data than Linear Discriminant Analysis as the boundary is required to conform more closely to the data rather than to the mean of the data. This was also taken into consideration when assessing the results of the Logistic Regression model development that occurs afterward.

Bartlett's Test of this data set yielded a significant p-value, where $p < 0.0001$, indicating that the proportion of distribution beyond the derrived 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.

Following the removal of predictor variables after visually inspecting the correlation heat map, we analyzed a correlation matrix. However, the matrix itself did not identify any remaining collinearity at a threshold of correlation necessitating removal of like-terms. Consequently, no further predictor variables are removed. Therefore, modeling data is broken into a 75% training / 25% testing data split for internal cross-cross validation. The objective of internal cross-validation is to develop a model using 75% of the data and test it on the remaining 25% in order to assess model fit statistics. Typically, following internal cross-validation, external cross-validation is performed.

Quadratic Discriminant Analysis: Internal Cross-Valdiation and Model Development

Following removal of significant levels of multicollinearity from the dataset and partitioning into a 75% training / 25% testing split, internal cross-validation is performed. The specifics of this test involves 25 folds of the data - meaning the 75% training data is divided into 25 partitions. The model is then trainied on 1/25th of the original 75%, then tested against the remaining 24/25ths, 1/25ths at-a-time. This test is repeated 5 times, with each repeat involving a different random partitioning of the 25 specified `folds` of the data. Finally, the model developed using the 75% training split is then applied to the 25% testing split and predictions are measured against the actuals of that split to develop model statistics such as `Accuracy`, `Misclassification`, `Precision`, `Sensitivity` and `Specificity`.

Quadratic Discriminant Analysis: External Cross-Valdiation and Model Development

After building a model using internal cross-validation, which applied 5 repeated internal cross-validations across the 25 folds of training data, a confusion matrix was constructed and analyzed. Next, we applied the model developed using the 75% training split to make predictions against the entire portion of data that includes values for `shot_made_flag` in order to assess how closely the model can predict against the entire data set compared to the actuals. Applying the model to the entire dataset as `external cross-validation` provides the model an opportunity to test against different data and more closely simulate a real-life scenario than internal cross-validation. Internal and external cross-validation is performed for later Logistic Regression models as well. Following external cross-validation of both models, the metrics are compared to determine the best model (Quadratic Discriminant Analysis versus Logistic Regression).

A confusion matrix is a table of results from cross-validation. Some key metrics provided by a confusion matrix include `Accuracy`, `Precision`, `Sensitivity` and `Specificity`. `Accuracy` is the number of all correct predictions divided by the number of all predictions. `Precision` is the ratio of the number of correctly classified positive predictions divided by the number of all positive

Table 3: Internal Cross-Validation Confusion Matrix

	Internal CV Statistics
Sensitivity	0.51431
Specificity	0.66761
Precision	0.56104
Accuracy	0.59826
Misclassification Rate	0.40174
Logarithmic Loss	0.70127
Area Under the Curve	0.40904

Table 4: External Cross-Validation Confusion Matrix

	External CV Statistics
Sensitivity	0.51187
Specificity	0.66993
Precision	0.55695
Accuracy	0.59917
Misclassification Rate	0.40083
Logarithmic Loss	0.70127
Area Under the Curve	0.59090

Quadratic Discriminant Analysis: Internal vs. External Cross-Validation

Using the two confusion matrix output tables immediately below, the performance across internal and external cross-validations of the QDA model can be compared. As indicated in those figures, the model performed highly similarly across both cross-validation techniques, indicating the model is consistent and reasonably fit, after controlling for the variables selected for modeling.

Logistic Model Development using Ordinary Least Squares

A preliminary, manual veriable 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:

$shot_{m}adeflag = shotdistance + attendance + combinedshottype + arenatemp + gameeventid + secondsremaining + shottype + gamedate + minutesremaining + locy + shotid$

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:

$shotmadeflag = combinedshottype + gameeventid + locy + minutesremaining + secondsremaining + shotdistance + shottype + gamedate + shotid + attendance + arenatemp$

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:

$shotmadeflag = combinedshottype + gameeventid + locy + minutesremaining + secondsremaining + shotdistance + shottype + gamedate + shotid + attendance + arenatemp$

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

Akaikes.Information.Criterion..Stepwise.Regression
26824.26