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

Paul Adams
Reannan McDaniel
Jeff Nguyen

Master of Science in Data Science, Southern Methodist University, USA

Contents

1 Abstract:

1.0.1 *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.*

2 Introduction

2.0.1 This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space. This is a sample introduction. Nothing but a sample. But a necessary sample to preserve space.

3 Exploratory Data Analysis

3.1 Outlier Check

3.1.1 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`.

3.2 Variable Elimination

3.2.1 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.

3.3 Addressing Multicollinearity: Correlation Plot for Visual Data Exploration

3.3.1 To address multicollinearity among quantitative predictor variables, a correlation heat map was created for visual inspection of correlation. Please see Appendix A to view this correlation heat map.

3.4 Post-Correlation Heat Map Variable Elimination

3.4.1 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.

3.5 Addressing Multicollinearity: Correlation Matrix for Numerical Analysis

3.5.1 After deselecting the most obvious collinear terms through visually inspection of the correlation plot, 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.

4 Quadratic Discriminant Analysis

4.0.1 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.

4.0.2 `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 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.

4.1 Bartlett's Test

4.1.1 The result of the Bartlett's 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.

4.1.2 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.

5 Quadratic Discriminant Analysis: Internal Cross-Valdiation and Model Development

5.0.1 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`.

6 Quadratic Discriminant Analysis: External Cross-Valdiation and Model Development

6.0.1 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).

6.0.2 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`.

	Internal CV Statistics		External CV Statistics
Sensitivity	0.51431	Sensitivity	0.51187
Specificity	0.66761	Specificity	0.66993
Precision	0.56104	Precision	0.55695
Accuracy	0.59826	Accuracy	0.59917
Misclassification Rate	0.40174	Misclassification Rate	0.40083
Logarithmic Loss	0.70127	Logarithmic Loss	0.70127
Area Under the Curve	0.40904	Area Under the Curve	0.59090

8 Logistic Model Development using Ordinary Least Squares

8.0.1 Logistic Regression is a classification technique that is best suited for dichotomous response variables - in the case of the Kobe data the response is ‘0’ for shot missed, or ‘1’ for shot made. Compared to discriminant analysis techniques multiple explanatory variables, interactions, and categorical variables can be used allowing for a potentially more descriptive model. For this type of regression, coefficients are in log-odds where each coefficient needs to be exponentiated to yield odds ratios - this is done for ease of interpretation. Logistic Regression can also be used to generate predictions that yield the probability of an observation having the desired traits of the response variable as occurring or not.

8.1 Logistic Regression: Model Selection

8.1.1 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. These initial models are used by forward, backward, and stepwise model selection methods are used to help select a combination of variables that result in the lowest Residual Deviance and/or AIC. The selection method that generates the model with the lowest AIC/Residual Deviance is then used for internal cross validation to further tune the model which allows for better prediction. Below are models that each selection method generated:

Forward Selection Model: $shot\_made\_flag = action\_type + attendance + arena\_temp + game\_event\_id + season + seconds\_remaining + minutes\_remaining + loc\_y + game\_date + loc\_x$

Backward Elimination Model: $shot\_made\_flag = recId + action\_type + game\_event\_id + loc\_x + minutes\_remaining + season + seconds\_remaining + shot\_distance + game\_date + shot\_id + attendance + arena\_temp$

Stepwise Regression Model: $shot\_made\_flag = recId + action\_type + game\_event\_id + loc\_x + minutes\_remaining + season + seconds\_remaining + shot\_distance + game\_date + shot\_id + attendance + arena\_temp$

8.1.2 Based on the fit-statistics generated from each model selection method, the backwards and stepwise models are identical in fit-statistics with an AIC at 25167.77, and the residual deviance at 25001.77. The forward model selection out-performs both backwards and stepwise models with an AIC of 25166.48, but has a higher a residual deviance at 25001.48. Compared to the forward selected model, the backwards and stepwise models have lower residual deviances, although their residual deviances are very close in value to the forward model, their AIC values are larger compared to the forward selected model. Based on the evidence, the forward selected model will be used for the internal and external cross validation process.

Selection.Type	AIC	Residual.Deviance
Forwards	25166.48	25004.48
Backwards	25167.77	25001.77
Stepwise	25167.77	25001.77

	Internal CV Statistics
Sensitivity	0.51431
Specificity	0.66761
Precision	0.56104
Accuracy	0.59826
Misclassification Rate	0.31175
Logarithmic Loss	0.74464
Area Under the Curve	0.70166

	External CV Statistics
Sensitivity	0.86502
Specificity	0.46244
Precision	0.66501
Accuracy	0.68479
Misclassification Rate	0.31175
Logarithmic Loss	0.70127
Area Under the Curve	0.59090

8.2 Logistic Regression: Internal Cross Validation and Model Development

8.2.1 After identifying that the forward selected model is the best candidate based on it’s AIC and residual deviance, its features are tuned using an internal cross validation. A training set is generated by randomly selecting 75% of the observations, with the remaining 25% serving as the validation (test) set. The model is tuned using 25 folds and is repeated 5 times, this process is the same as described in section 5.0.1. Prior to the internal cross validation process, action types that rarely occur in the data are recoded to similar, but differ action types, i.e. “Running Tip Shot” (1 observation) to “Tip Shot” (many observations). If levels within the evaluation data exist but are not present in the training data, a trained model will have difficulty making predictions. Infrequently occurring action types are recoded to similar action types to avoid this situation.

8.3 Logistic Regression: External Cross Validation and Model Development

8.3.1 External cross validation is used to evaluate a model after it has been tuned in the internal cross validation process. External cross validation confusion matrix statistics, ROC/AUC, and misclassification rates can be compared to the internal cross validation statistics to help assess performance.

8.3.2 As with the “Quadratic Discriminant Analysis: External Cross-Validation and Model Development” section, confusion matrices are used to assess model performance where we will be focusing on `Accuracy`, `Precision`, `Sensitivity`, `Specificity`, `Misclassification Rate`, `AUC`, and `Log Loss`. When looking at model performance high values for: ‘Accuracy,Precision,Sensitivity,Specificity,AUC`are desirable; and low values for`Misclassification Rate,Log Loss‘ are desirable. For descriptions of the mentioned terms please refer to the “Quadratic Discriminant Analysis: External Cross-Validation and Model Development” section for more information.

8.4 Logistic Regression: Internal vs. External Cross-Validation

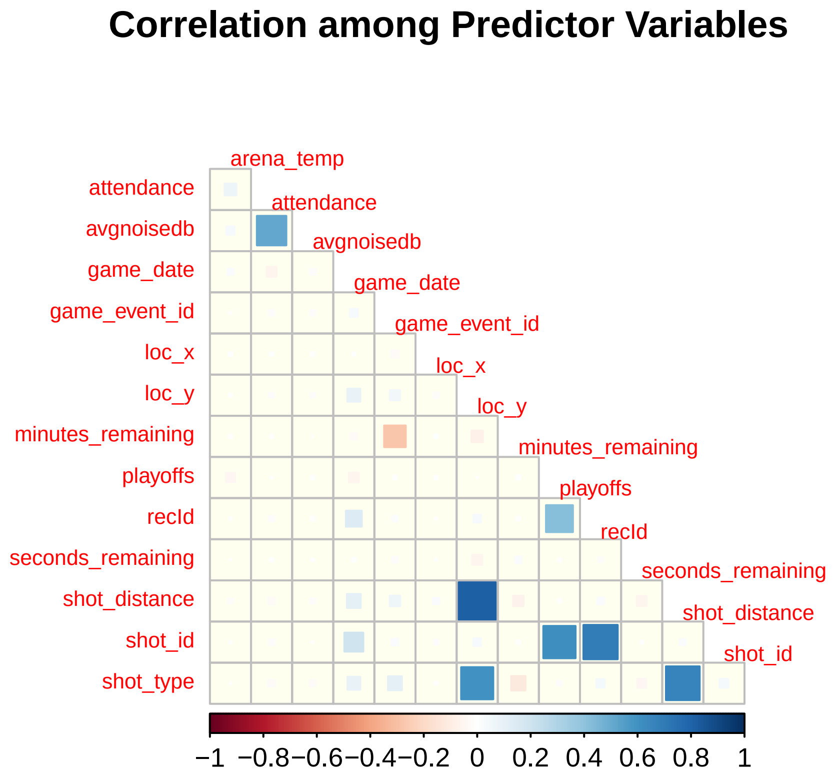
8.4.1 The misclassification rate, and Log loss for both internal and external cross validations are close in values However the external cross validated model has higher `Sensitivity`, `Precision`, and `Accuracy`; but lower `Specificity` and `Area Under the Curve` compared to the internal cross validated model. This suggests that model `Sensitivity`, `Precision`, and `Accuracy` improves, but its predictive performance decreases when evaluated using the external cross validation.

8.5 Logistic Regression: Fitted Model

	Coefficient (logit)	Odds Ratio
‘action_typeRunning Finger Roll Shot‘	-16.890156	0.0000000
‘action_typeRunning Pull-Up Jump Shot‘	-16.522812	0.0000001
‘action_typeRunning Reverse Layup Shot‘	-4.593293	0.0101195
‘action_typeTip Shot‘	-4.142288	0.0158865
‘action_typeJump Shot‘	-4.084338	0.0168343
‘action_typeDriving Jump shot‘	-4.071401	0.0170535

9 Appendix A: Images and Tables

9.0.1 Correlation Heat Map



9.0.2 Correlation Matrix - Top 10 Collinear Terms

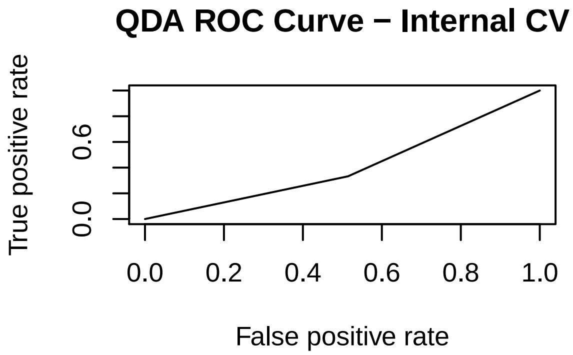
Correlation Predictor Variable	Correlation Response Variable	Correlation	p-Value
arena_temp	arena_temp	0.51092	$p < 0.0001$
attendance	arena_temp	0.51092	$p < 0.0001$
game_date	arena_temp	0.51092	$p < 0.0001$
game_event_id	arena_temp	0.51092	$p < 0.0001$
loc_x	arena_temp	0.51092	$p < 0.0001$
loc_y	arena_temp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

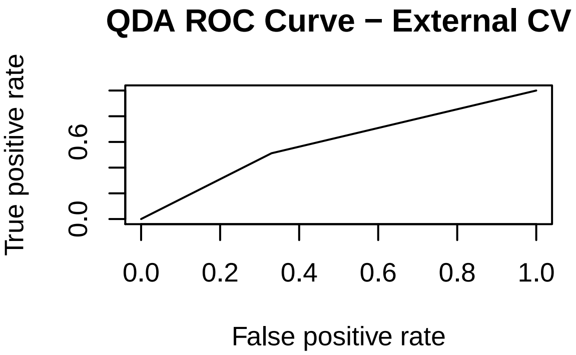
Table 1: (#tab:Bartletts Results for Appendix)Bartlett Test's Wilks' Lambda Approximation

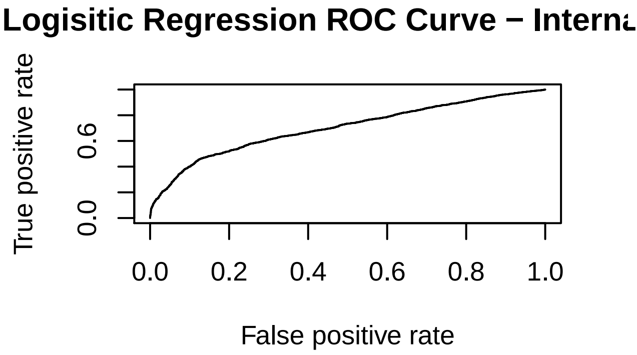
	Metric Output
Chi Square Statistic	1037.24251
Degrees of Freedom	14
Wilks' Lambda	0.9511
p-Value	p < 0.0001

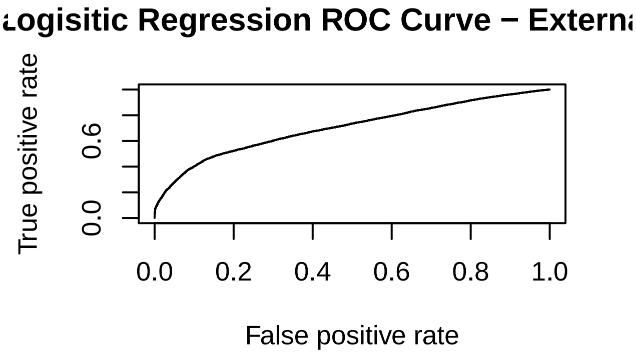
9.0.3 Bartlett Test's Wilks' Lambda Approximation

9.0.4 ROC Curves









9.0.5 Fitted Logistic Regression Model

	Coefficient (logit)	Odds Ratio
‘action_typeRunning Finger Roll Shot‘	-16.8901565	0.0000000
‘action_typeRunning Pull-Up Jump Shot‘	-16.5228116	0.0000001
‘action_typeRunning Reverse Layup Shot‘	-4.5932930	0.0101195
‘action_typeTip Shot‘	-4.1422879	0.0158865
‘action_typeJump Shot‘	-4.0843383	0.0168343
‘action_typeDriving Jump shot‘	-4.0714011	0.0170535
‘action_typeHook Shot‘	-3.8604300	0.0210589
‘action_typeLayup Shot‘	-3.8492906	0.0212948
‘action_typeTurnaround Hook Shot‘	-3.4742741	0.0309843
‘action_typeDriving Floating Jump Shot‘	-3.4017209	0.0333159
‘action_typeFinger Roll Shot‘	-3.1166502	0.0443053
‘action_typeFadeaway Jump Shot‘	-3.0098168	0.0493007
‘action_typeTurnaround Fadeaway shot‘	-3.0083465	0.0493732
‘action_typePullup Bank shot‘	-2.9362298	0.0530654
‘action_typeTurnaround Jump Shot‘	-2.9114822	0.0543950
‘action_typeRunning Layup Shot‘	-2.8570672	0.0574370
‘action_typeReverse Layup Shot‘	-2.8474095	0.0579944
‘action_typeRunning Finger Roll Layup Shot‘	-2.8130631	0.0600209
‘action_typeDriving Hook Shot‘	-2.7767584	0.0622399
‘action_typePutback Dunk Shot‘	-2.7693202	0.0627046
‘action_typePutback Layup Shot‘	-2.6855622	0.0681829
‘action_typeStep Back Jump shot‘	-2.5120193	0.0811043
‘action_typeFinger Roll Layup Shot‘	-2.3355224	0.0967599
‘action_typeAlley Oop Layup shot‘	-2.3236860	0.0979120
‘action_typeDunk Shot‘	-2.2667110	0.1036525
‘action_typeFloating Jump shot‘	-2.2488987	0.1055154
‘action_typeRunning Jump Shot‘	-2.2295025	0.1075819
‘action_typeDriving Layup Shot‘	-2.2065361	0.1100813
‘action_typePullup Jump shot‘	-2.2013807	0.1106503
‘action_typeJump Bank Shot‘	-2.0602324	0.1274244
‘action_typeJump Hook Shot‘	-2.0442775	0.1294737
‘action_typeTurnaround Bank shot‘	-1.9606411	0.1407681
‘action_typeDriving Reverse Layup Shot‘	-1.9127755	0.1476700
‘action_typeRunning Hook Shot‘	-1.8911388	0.1508999
‘action_typeDriving Finger Roll Shot‘	-1.8503630	0.1571801
‘action_typeCutting Layup Shot‘	-1.6524812	0.1915740
‘action_typeRunning Dunk Shot‘	-1.5748068	0.2070476
‘action_typeRunning Bank shot‘	-1.5252238	0.2175724
‘action_typeDriving Finger Roll Layup Shot‘	-1.1146155	0.3280414
‘action_typeFollow Up Dunk Shot‘	-1.0763881	0.3408243
‘action_typeFadeaway Bank shot‘	-0.6794888	0.5068761
‘season1997-98‘	-0.1355493	0.8732362
game_event_id	-0.0003291	0.9996709
game_date	-0.0002174	0.9997826
loc_x	0.0001627	1.0001627
attendance	0.0001668	1.0001668
loc_y	0.0005035	1.0005036
seconds_remaining	0.0025877	1.0025910
minutes_remaining	0.0125250	1.0126037
arena_temp	0.0375352	1.0382485
‘action_typeReverse Dunk Shot‘	0.0793841	1.0826201
‘action_typeDriving Slam Dunk Shot‘	0.1335033	1.1428250