Will Kobe Bryant Make His Next Shot: Linear 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, we used a correlation matrix.

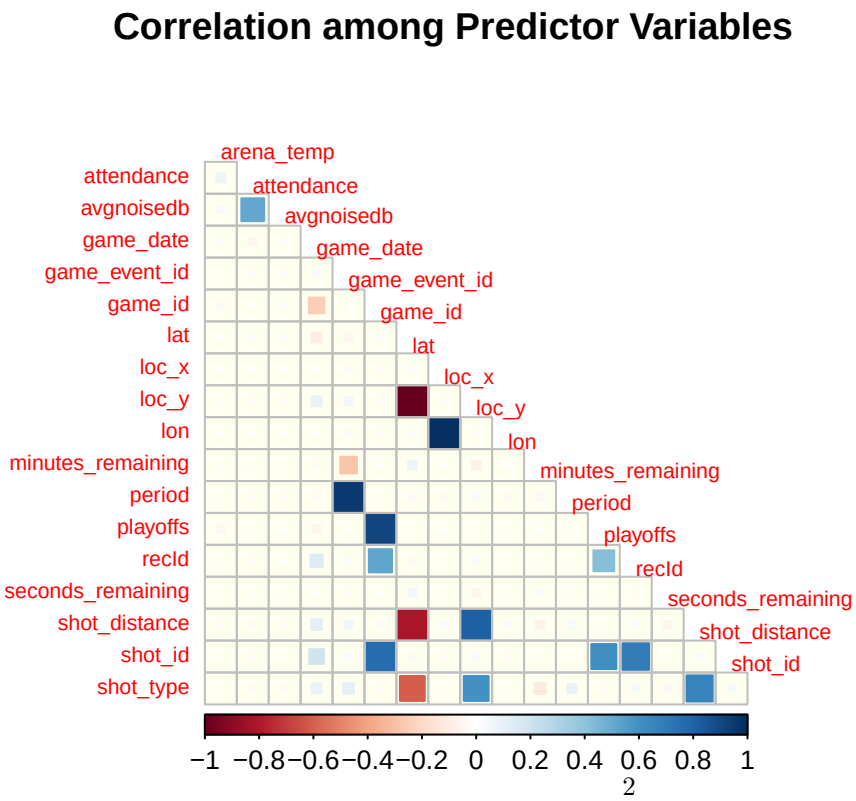


Table 1: Top 10 Collinear Terms

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 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 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.

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 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. ## **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 Test's Wilks' Lambda Approximation
Chi Square Statistic	1037.24251
Degrees of Freedom	14
Wilks' Lambda	0.9511
p-Value	$p < 0.0001$

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

Internal Cross-Valdiation and Model Development

External Cross-Validation

After building a model using internal cross-validation, which applied 5 repeated cross-validations across 25 folds of the data, a confusion matrix was constructed and analyzed. Next, we used the model to make predictions against the entire portion of the dataset that included values for `shot_made_flag` to assess how closely the model can predict against the entire data set compared to the actuals.

Although the confusion matrix was performed using internal cross validation.....this section needs to be removed:

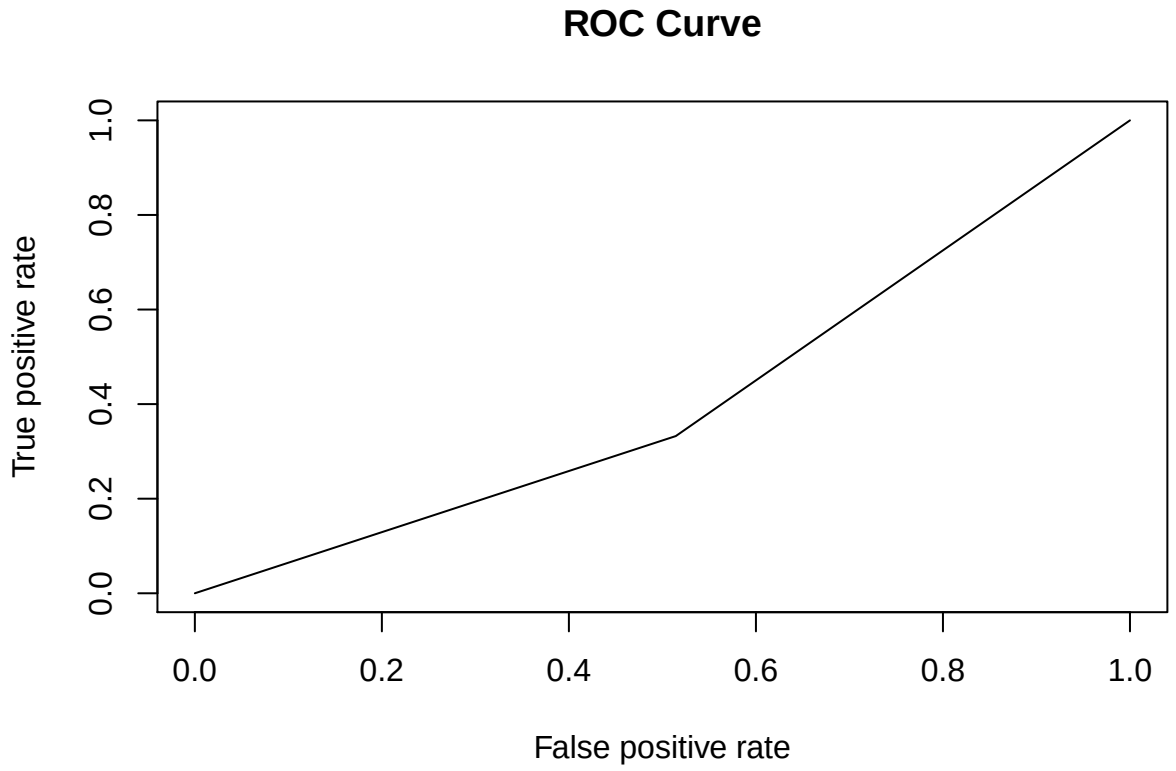
Quadratic Discriminant Analysis Misclassification Rate

Log Loss

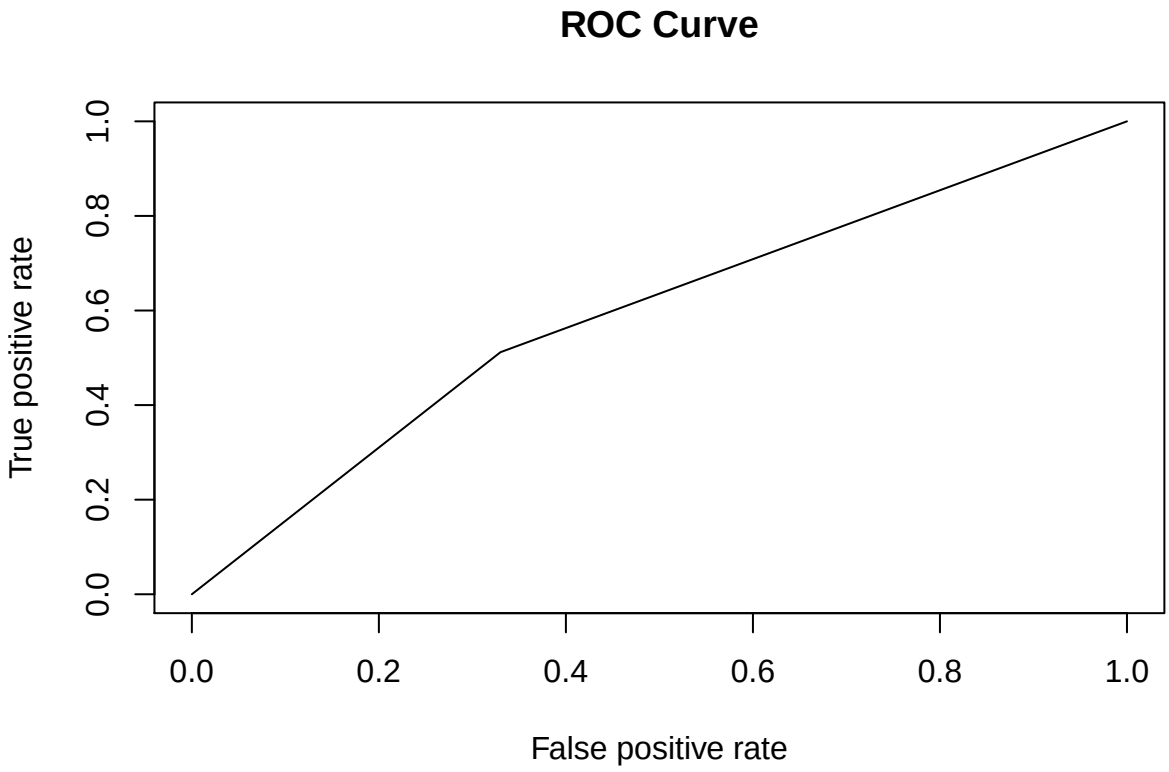
```
## function (actual, predicted)
```

```
## {
##     return(mean(ll(actual, predicted)))
## }
## <bytecode: 0x5592dd9ffd68>
## <environment: namespace:Metrics>
```

```
dfTrain.numeric$shot_made_flag <- ifelse(dfTrain.numeric$shot_made_flag=="made",1,0)
### Internal Cross-Validation
internal_cv.predicted.qda <- ifelse(internal_cv.predicted.qda=="made",1,0)

AUCpredStep.internal <- prediction(internal_cv.predicted.qda, as.numeric(subDF.Test.numeric$shot_made_fl
perf_step.internal <- performance(AUCpredStep.internal, measure = "tpr", x.measure = "fpr")
plot(perf_step.internal, main = "ROC Curve")
```




```
AUC.internal <- performance(AUCpredStep.internal, measure = "auc")
AUC.internal <- AUC.internal@y.values[[1]]

### External Cross-Validation
external_cv.predicted.qda <- ifelse(external_cv.predicted.qda=="made",1,0)

AUCpredStep.external <- prediction(external_cv.predicted.qda, as.numeric(dfTrain.numeric$shot_made_flag
perf_step.external <- performance(AUCpredStep.external, measure = "tpr", x.measure = "fpr")
plot(perf_step.external, main = "ROC Curve")
```




```
AUC.external <- performance(AUCpredStep.external, measure = "auc")
AUC.external <- AUC.external@y.values[[1]]
```

Confusion Matrix & Metrics Table

	Internal CV Statistics
Sensitivity	0.51431
Specificity	0.66761
Pos Pred Value	0.56104
Neg Pred Value	0.62463
Precision	0.56104
Accuracy	0.59826
Misclassification Rate	0.40174
Logarithmic Loss	0.70127
Area Under the Curve	0.40904

	External CV Statistics
Sensitivity	0.51187
Specificity	0.66993
Pos Pred Value	0.55695
Neg Pred Value	0.62868
Precision	0.55695
Accuracy	0.59917
Misclassification Rate	0.40083
Logarithmic Loss	0.70127

	External CV Statistics
Area Under the Curve	0.59090

Predictions from Quadratic Discriminant Analysis

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:

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

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