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

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.

Introduction

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.

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. Please see Appendix A to view this correlation heat map.

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

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.

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 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 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.

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 predictions. `Sensitivity` (also called `Recall`) is the number of correctly classified positive

	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

	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

Logistic Model Development using Ordinary Least Squares

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.

Logistic Regression: Model Selection

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

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	25168.23	25004.23
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	0.31136
Logarithmic Loss	0.74323
Area Under the Curve	0.70361

	External CV Statistics
Sensitivity	0.86230
Specificity	0.46417
Precision	0.66502
Accuracy	0.68406
Misclassification Rate	0.31136
Logarithmic Loss	0.70127
Area Under the Curve	0.59090

Logistic Regression: Internal Cross Validation and Model Development

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.

Logistic Regression: External Cross Validation and Model Development

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.

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`, and `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.

Logistic Regression: Internal vs. External Cross-Validation

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.

Logistic Regression: Fitted Model

The model was selected for prediction based on having the lowes AIC and residual deviance compared to backwards and stepwise selected models. The forward selected model was tuned using k-fold internal cross validation, the forward selected model generated is listed below where Logit coefficents for each feature can also be found in the "Fitted Logistic Regression Model" portion of the appendix

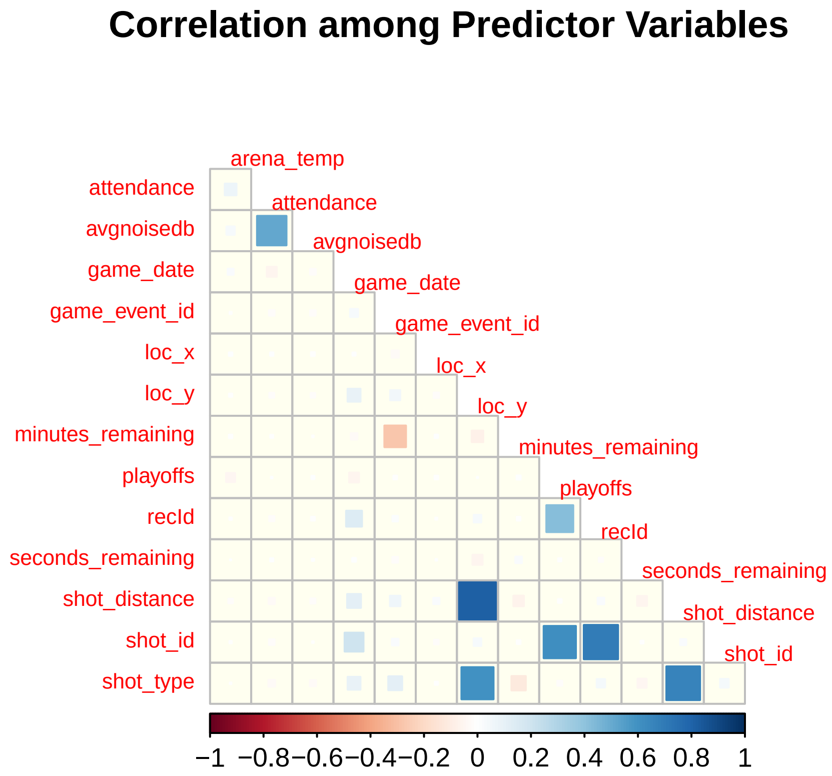
$$shot\_made\_flag = action\_type + attendance + arena\_temp + game\_event\_id + season + seconds\_remaining + minutes\_remaining + loc\_y + game\_date + loc\_x$$

Several notable features that this model generated are several of the `action_types`. The odds ratios of Kobe missing a shot are revealed once the coefficients are exponentiated. In the table below we can see that the odds ratios are all very small. This suggests that when Kobe performs one of the listed levels of `action_type`, his chances of missing the shot are very low. The frequency of the `action_type` can also be ascertained by looking at the standard errors from the model output. Small standard errors for a coefficient suggest that there are a large number of observations with that attribute with a similar outcomes, and a large standard error would suggest that there are fewer observations with a similar outcome. When moving from the base `action_type` of "Alley Oop Dunk Shot" to "Jump Shot" for every one unit increase in "Jump Shot" the logit value for `shot_made_flag` is approximate -4; in terms of odds ratios, compared to using the "Alley Oop Dunk Shot," Kobe's chances are missing the shot when using the "Jump Shot" is at a factor of approximately 0.016. The "Jump Shot" also has a standard error of approximately 0.723, this value is much lower compared to the other top shots suggesting that Kobe uses the "Jump Shot" frequently and rarely misses this shot type when he chooses to use it. When looking at the "Running Finger Roll Shot" the odds ratio of missing this shot when starting from a base level of "Alley Oop Dunk Shot" is approximately 4.5e-8 with a logit standard error of approximately 308; this suggests that when Kobe chooses to use this shot he will rarely use it, but due to the high standard error the number of times that Kobe uses this shot is very low. When a frequency table of the `action_type` for the whole dataset is created, the total number of times the "Running Finger Roll Shot" occurs is 4 - compared to 12712 observations for "Jump Shot."

	Coefficient (logit)	Odds Ratio
‘action_typeRunning Finger Roll Shot‘	-16.915327	0.0000000
‘action_typeRunning Pull-Up Jump Shot‘	-16.791811	0.0000001
‘action_typeRunning Reverse Layup Shot‘	-4.523732	0.0108485
‘action_typeTip Shot‘	-4.194048	0.0150851
‘action_typeJump Shot‘	-4.111135	0.0163892
‘action_typeDriving Jump shot‘	-4.109410	0.0164175

Appendix A: Images and Tables

Correlation Heat Map



Correlation Matrix - Top 10 Collinear Terms

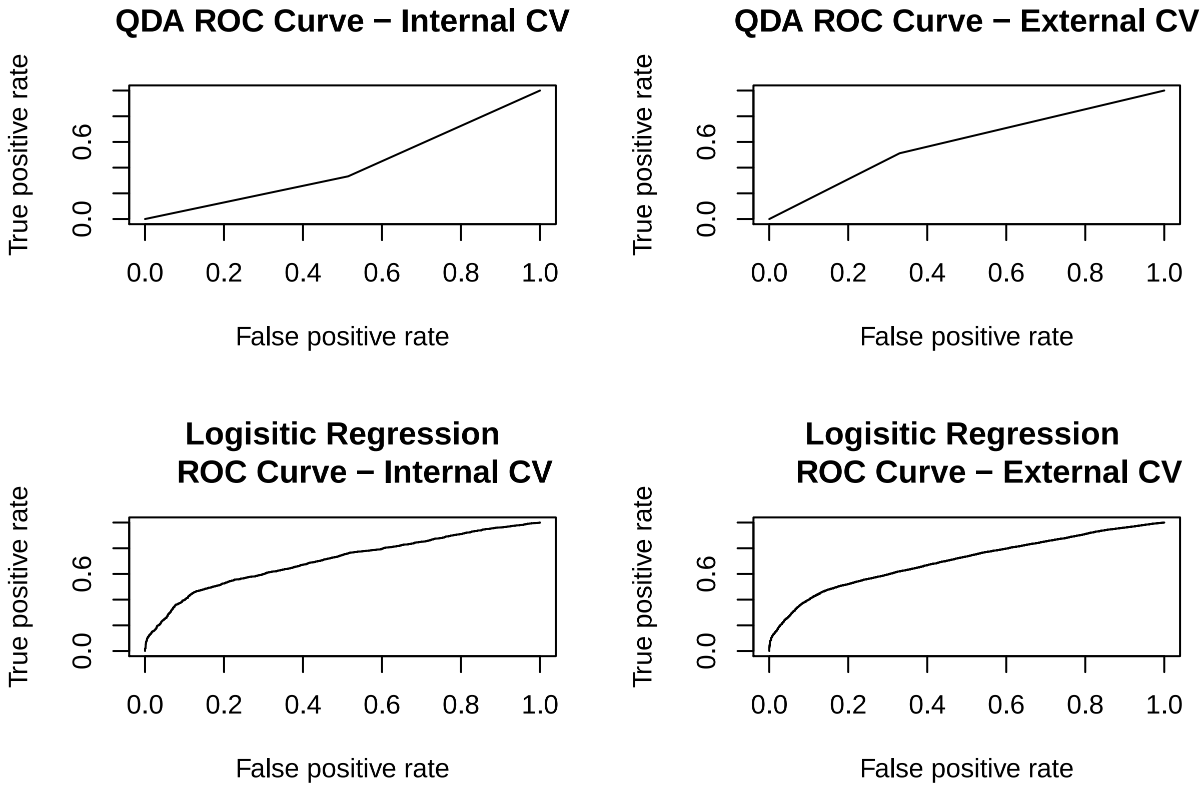
Correlation Predictor Variable	Correlation Response Variable	Correlation	p-Value
arena_temp	arena_temp	0.51092	p < 0.0001
attendance	arena_temp	0.51092	p < 0.0001
game_date	arena_temp	0.51092	p < 0.0001
game_event_id	arena_temp	0.51092	p < 0.0001
loc_x	arena_temp	0.51092	p < 0.0001
loc_y	arena_temp	0.51092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

Bartlett Test's Wilks' Lambda Approximation

ROC Curves



Fitted Logistic Regression Model

	Coefficient (logit)	Odds Ratio
(Intercept)	-0.5406091	0.5823934
‘action_typeAlley Oop Layup shot‘	-2.3125586	0.0990076
‘action_typeCutting Layup Shot‘	-1.9200208	0.1466039
‘action_typeDriving Bank shot‘	10.6747918	43251.7007167
‘action_typeDriving Dunk Shot‘	0.4927331	1.6367836
‘action_typeDriving Finger Roll Layup Shot‘	-1.1244474	0.3248319
‘action_typeDriving Finger Roll Shot‘	-1.8877604	0.1514105

(continued)

	Coefficient (logit)	Odds Ratio
‘action_typeDriving Floating Bank Jump Shot‘	9.9403844	20751.7192693
‘action_typeDriving Floating Jump Shot‘	-3.6656405	0.0255878
‘action_typeDriving Hook Shot‘	-2.8993874	0.0550569
‘action_typeDriving Jump shot‘	-4.1094104	0.0164175
‘action_typeDriving Layup Shot‘	-2.2506621	0.1053295
‘action_typeDriving Reverse Layup Shot‘	-1.9553431	0.1415159
‘action_typeDriving Slam Dunk Shot‘	0.1657104	1.1802312
‘action_typeDunk Shot‘	-2.2960202	0.1006587
‘action_typeFadeaway Bank shot‘	-0.6729202	0.5102165
‘action_typeFadeaway Jump Shot‘	-3.0523906	0.0472458
‘action_typeFinger Roll Layup Shot‘	-2.3467101	0.0956834
‘action_typeFinger Roll Shot‘	-3.0929933	0.0453660
‘action_typeFloating Jump shot‘	-2.2745444	0.1028438
‘action_typeFollow Up Dunk Shot‘	-1.1106209	0.3293544
‘action_typeHook Bank Shot‘	10.2687917	28819.0438353
‘action_typeHook Shot‘	-3.9216063	0.0198092
‘action_typeJump Bank Shot‘	-2.0558189	0.1279880
‘action_typeJump Hook Shot‘	-2.1372195	0.1179824
‘action_typeJump Shot‘	-4.1111346	0.0163892
‘action_typeLayup Shot‘	-3.8709890	0.0208378
‘action_typePullup Bank shot‘	-2.9677237	0.0514202
‘action_typePullup Jump shot‘	-2.3056159	0.0996974
‘action_typePutback Dunk Shot‘	-2.7340623	0.0649549
‘action_typePutback Layup Shot‘	-2.6452777	0.0709856
‘action_typeReverse Dunk Shot‘	0.0718912	1.0745384
‘action_typeReverse Layup Shot‘	-2.8641128	0.0570337
‘action_typeReverse Slam Dunk Shot‘	10.3104940	30046.2750438
‘action_typeRunning Bank shot‘	-1.5071851	0.2215327
‘action_typeRunning Dunk Shot‘	-1.6818534	0.1860289
‘action_typeRunning Finger Roll Layup Shot‘	-2.8650540	0.0569801
‘action_typeRunning Finger Roll Shot‘	-16.9153274	0.0000000
‘action_typeRunning Hook Shot‘	-1.9042649	0.1489321
‘action_typeRunning Jump Shot‘	-2.2720781	0.1030977
‘action_typeRunning Layup Shot‘	-2.8672407	0.0568556
‘action_typeRunning Pull-Up Jump Shot‘	-16.7918113	0.0000001
‘action_typeRunning Reverse Layup Shot‘	-4.5237315	0.0108485
‘action_typeSlam Dunk Shot‘	0.7787543	2.1787564
‘action_typeStep Back Jump shot‘	-2.5994483	0.0743146
‘action_typeTip Shot‘	-4.1940480	0.0150851
‘action_typeTurnaround Bank shot‘	-1.9673093	0.1398326
‘action_typeTurnaround Fadeaway shot‘	-3.0814350	0.0458934
‘action_typeTurnaround Finger Roll Shot‘	10.0616178	23426.3747389
‘action_typeTurnaround Hook Shot‘	-3.6153476	0.0269076
‘action_typeTurnaround Jump Shot‘	-2.9350977	0.0531255
attendance	0.0001723	1.0001723
arena_temp	0.0383777	1.0391236
game_event_id	-0.0003392	0.9996608

(continued)

	Coefficient (logit)	Odds Ratio
seconds_remaining	0.0027254	1.0027291
minutes_remaining	0.0120657	1.0121388
loc_y	0.0005358	1.0005359
game_date	-0.0000377	0.9999623
loc_x	0.0001772	1.0001772
playoffs	0.0045673	1.0045778

Action Type Frequency Table

Var1	Freq
Alley Oop Dunk Shot	76
Alley Oop Layup shot	59
Cutting Layup Shot	6
Driving Bank shot	1
Driving Dunk Shot	196
Driving Finger Roll Layup Shot	47
Driving Finger Roll Shot	52
Driving Floating Bank Jump Shot	1
Driving Floating Jump Shot	3
Driving Hook Shot	13
Driving Jump shot	19
Driving Layup Shot	1335
Driving Reverse Layup Shot	67
Driving Slam Dunk Shot	38
Dunk Shot	176
Fadeaway Bank shot	22
Fadeaway Jump Shot	693
Finger Roll Layup Shot	21
Finger Roll Shot	23
Floating Jump shot	75
Follow Up Dunk Shot	10
Hook Bank Shot	5
Hook Shot	61
Jump Bank Shot	223
Jump Hook Shot	16
Jump Shot	12712
Layup Shot	1734
Pullup Bank shot	10
Pullup Jump shot	318
Putback Dunk Shot	3
Putback Layup Shot	9
Reverse Dunk Shot	52
Reverse Layup Shot	276
Reverse Slam Dunk Shot	15
Running Bank shot	35
Running Dunk Shot	14

(continued)

Var1	Freq
Running Finger Roll Layup Shot	5
Running Finger Roll Shot	4
Running Hook Shot	28
Running Jump Shot	620
Running Layup Shot	42
Running Pull-Up Jump Shot	1
Running Reverse Layup Shot	6
Slam Dunk Shot	264
Step Back Jump shot	93
Tip Shot	121
Turnaround Bank shot	50
Turnaround Fadeaway shot	299
Turnaround Finger Roll Shot	1
Turnaround Hook Shot	8
Turnaround Jump Shot	739