Rikel Djoko

Pradeep Kumar

MSDS 6372

08/7/19

Introduction

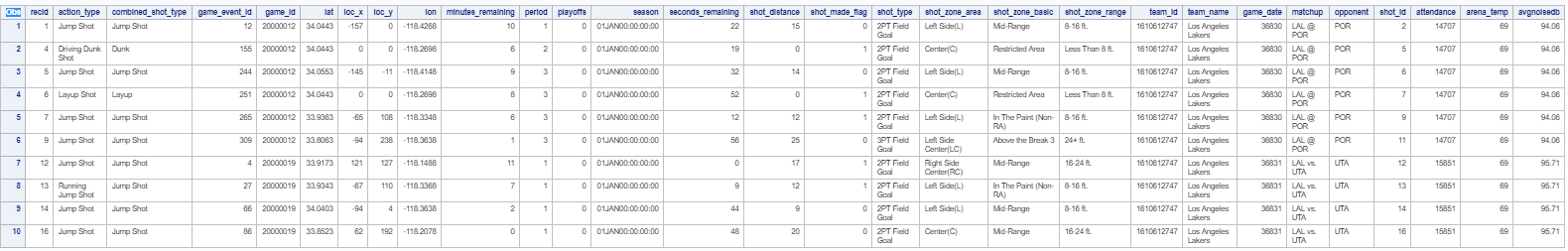
For most of us who follow Basketball, every shot is important and it’s a potential swish. Hence keep us in suspense for a brief moment until we find out if the shot will end up at the bottom of the net. As of today the shot outcome is simply a guessing game, in this paper we going to try a different approach to accurately predict the shot outcome using 20 years of data on Kobe’s swishes and misses and apply machine learning technique.

Data Description

The dataset for this project originates from the Kaggle platform (<https://www.kaggle.com/c/kobe-bryant-shot-selection/data>). The Data consists of 20697 shots with 22 features which describe location and circumstance of every field goal attempted by Kobe Bryant took during his 20-year career. Most of the feature are self-explanatory, see table 1 below and figure 1 for dataset overview/

***Table1: List of features***

|  |  |
| --- | --- |
| recId  action\_type  combined\_shot\_type  game\_event\_id  game\_id  lat – court location identifier (latitude)  loc\_x - court location identifier (x/y axis)  loc\_y- court location identifier (x / y axis)  lon - court location identifier (longitude)  minutes\_remaining – (in period)  period  playoffs  season  seconds\_remaining  attendance | avgnoisedb – avg noise in arena (decibels)  shot\_distance  shot\_made\_flag (this is what you are predicting)  shot\_type  shot\_zone\_area  shot\_zone\_basic  shot\_zone\_range  team\_id  team\_name  game\_date  matchup  opponent  shot\_id  arena\_temp (oF) |



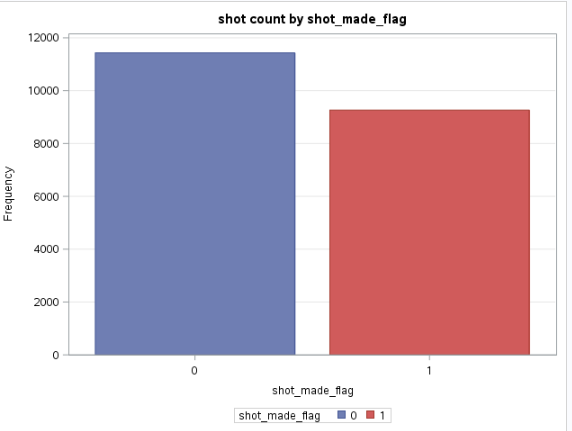
***Figure 1: dataset sample***

 Exploratory Data Analysis

In this section we will use multiple visualization technique to understand each feature and how it relate to the dependent variable, we’re also going to identify outliers address and identify multicollinearity and correlation between variable. And finally check assumption for logical regression.

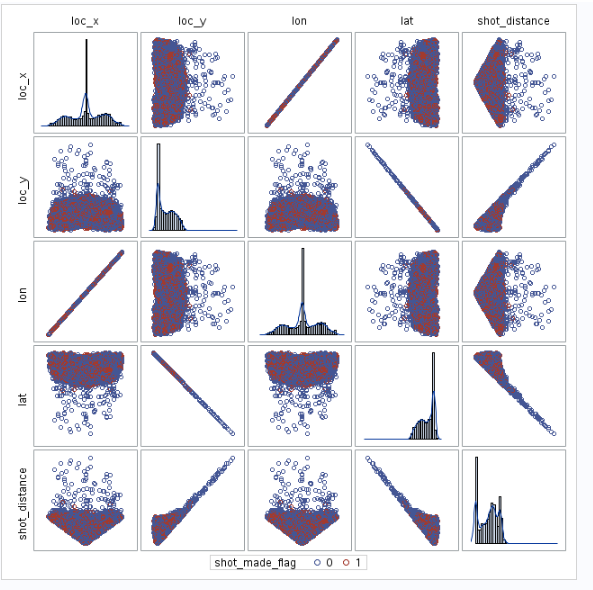
* Data visualization

Figure2 shows the frequency distribution of the shot across the 20697 shots, the delta between the two categories pretty significant.



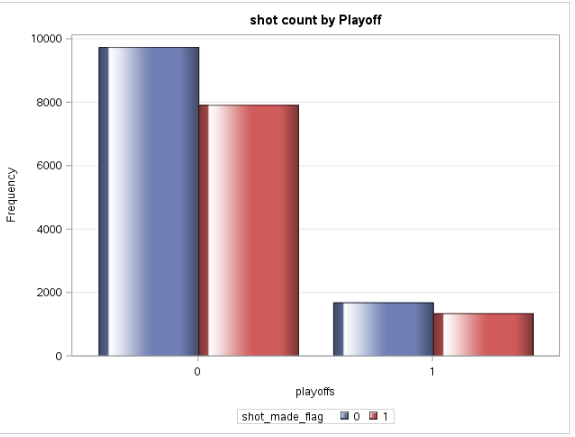
***Figure 2: short frequency distribution***

Figure3 below shows the matrix scatterplot of each of the continue variable, based on the trend of variation we can infer high correlation between lat and loc\_y, lon and loc\_x.

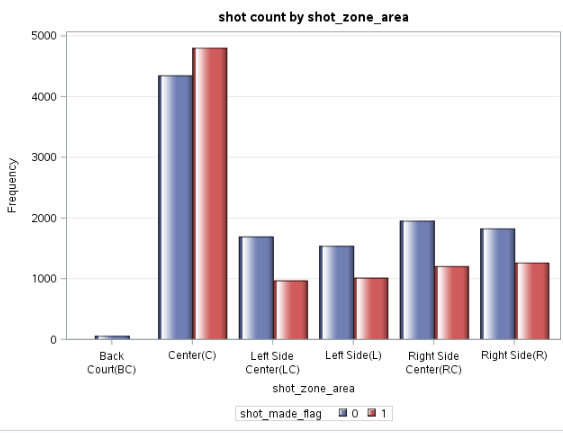


***Figure 3: matrix scatterplot***

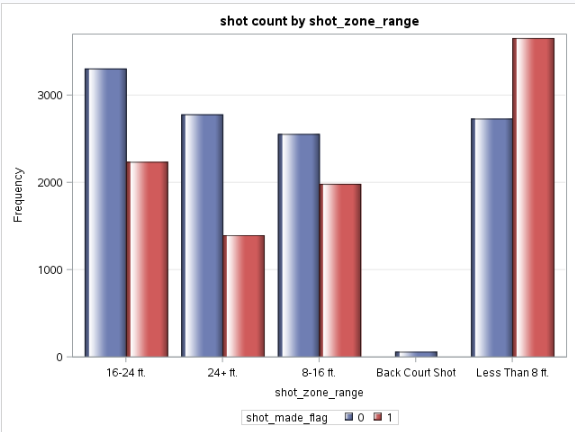
Figure 4-8 show the distribution of shot across different categorical, this will identify which of the categorical has less effect on the dependent variable.

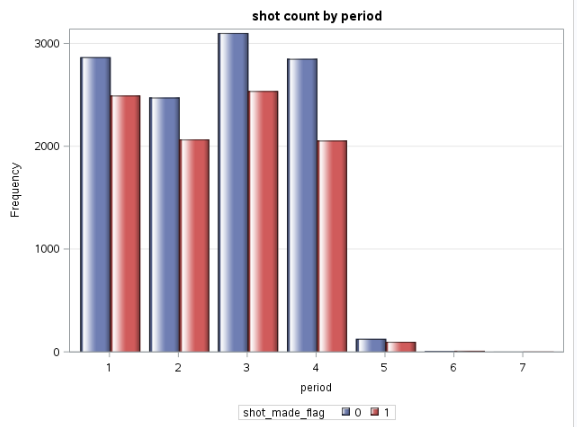


***Figure 4: short distribution based on playoff***



***Figure 5: short distribution based on shot\_zone\_area***

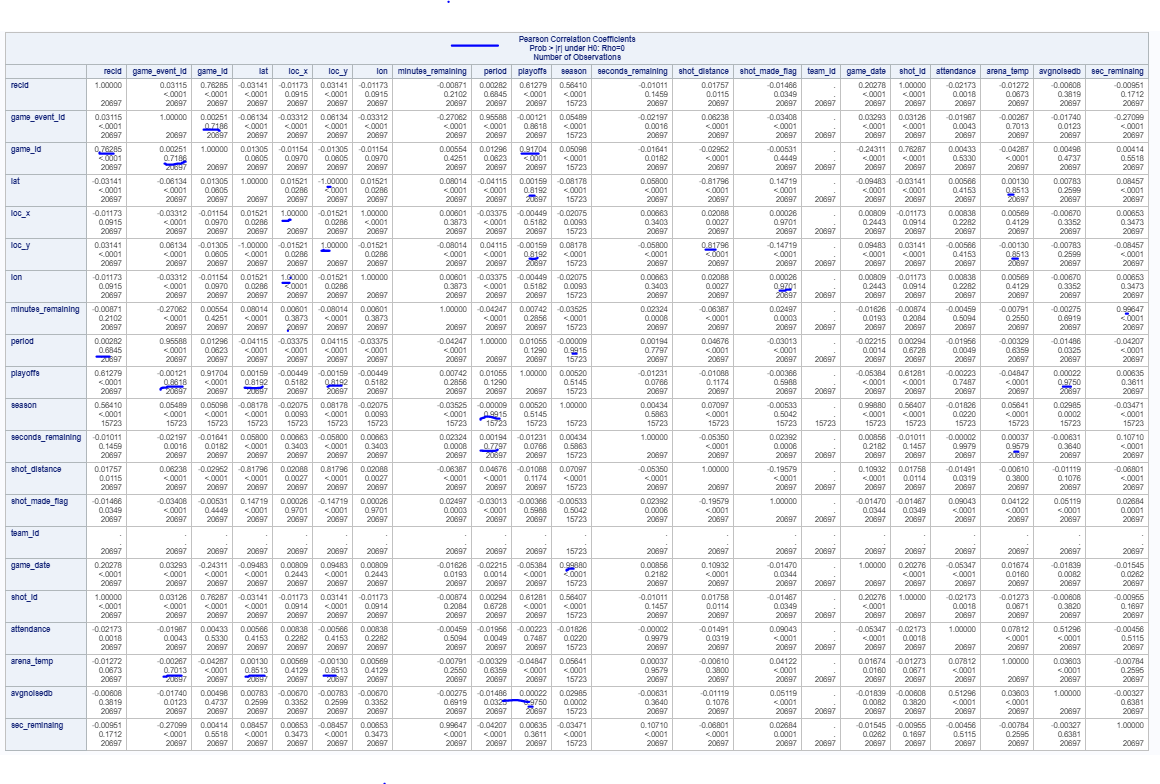




***Figure 6: short distribution based on short\_zone\_range Figure 7: short distribution based on period***

* Assumption:
* Binary response variable: shot\_made\_flag is a binary variable where 0 represent missed shot and 1 represent made shot.
* Model fitted correctly, no over fitting nor under fitting: not enough information to validate this assumption we will come back to it the next section using goodness of fit after building the model.
* Independent : each observation are independent from each other
* Linearity of independent variable and log odds: not enough information to validate this assumption we will come back to it the next section using the general test after building the model
* Feature Selection:

Figure 7 shows a Pearson correlation coefficient table across all variables.

Feature selection remove highly correlated feature ***Figure6: Pearson Correlation***

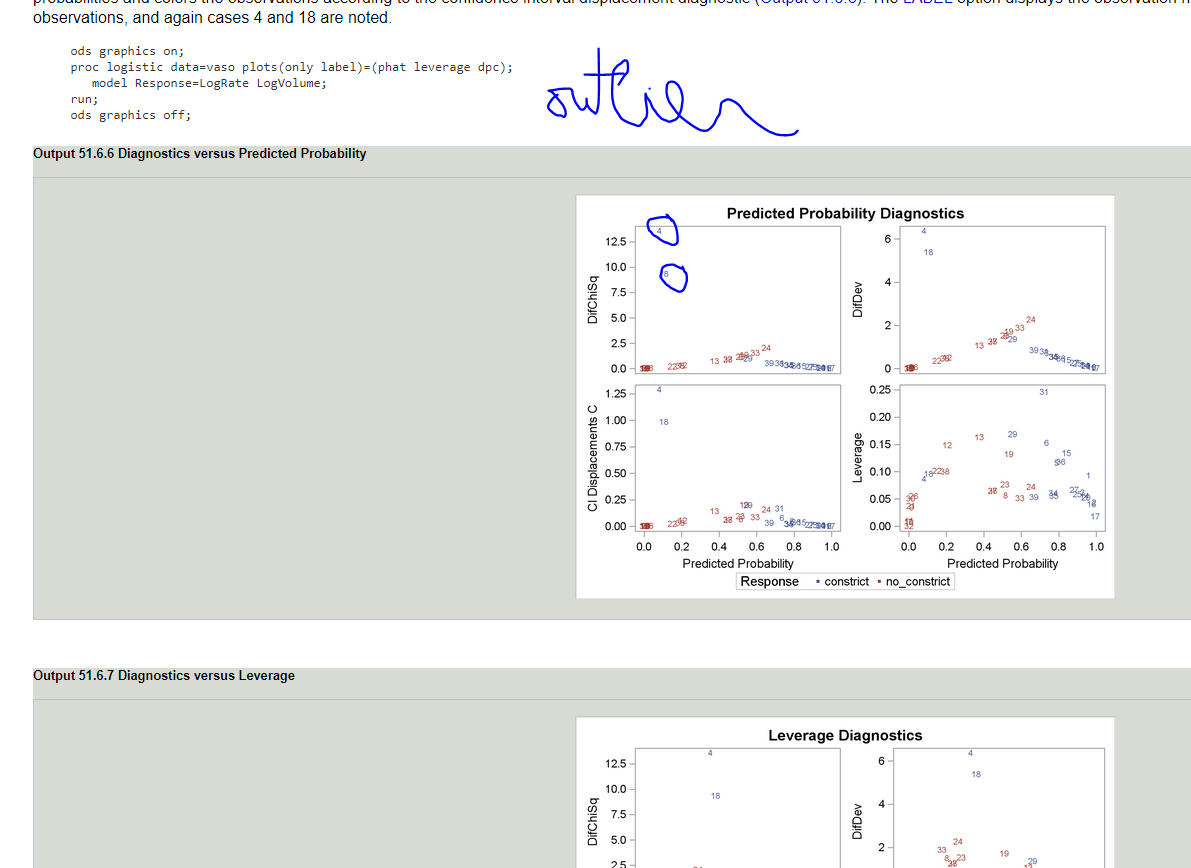
Based on the correlation coefficient we can remove some of the high correlated variable, since they are carry the same information.

* game\_Id and game\_event\_Id have a corr coefficient of 0.71 and based on the description it makes logical sense to use either of them in the final model.
* Arena\_temp is highly correlated with game\_event\_Id , lon and loc\_y, but logically it doesn’t make sense so we will keep each variable separately
* Game date and season are highly correlated, and it make logical sense since season is defined around the same date , thus we can keep season and remove date
* Avgnoisedb and playoff are highly correlated and it make sense since playoff are very intense. So we can keep playoff and remove avgnoisedb feature.
* Lon and loc\_x and lat and loc\_y are so highly correlated and make sense so we will keep loc\_x and loc\_y and remove Lon and Lat feature from the model.
* Since Minute remain and second remaining represent the time value at a different scale, so we add a new variable “sec\_reamaining” which is the sum of Minute and second remaining.

Build model

1. Remove outlier like LR: but now use a different technique

* Outlier detection during simple model build

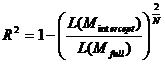


The second step of [logistic regression](http://www.statisticssolutions.com/resources/directory-of-statistical-analyses/what-is-logistic-regression) is to formulate the model, i.e. that variable X1, X2, and X3 have a causal influence on the probability of event Y to happen and that their relationship is linear.  We can now express the logistic regression function as logit(p)

Logistic Regression

The third step of regression analysis is to fit the regression line using maximum likelihood estimation.  Maximum likelihood is an iterative approach to maximize the likelihood function.  SPSS specifically -2\*log(likelihood function) ? min!

When it come to predicting the dichotomous dependent variable (bug is either dead or alive) then the cut-off is drawn at p = 0.5. This is at p > 0.5 we expect the bug to be dead.  The critical concentration for this to happen is  
Logistic Regression  
which is in this example 1.4/2.0 = 0.7.  
The last step is to check the validity of the logistic regression model.  Similar to regular regression analysis we calculate a R².  However for logistic regression this is called a Pseudo-R².  The measures of fit are based on the -2log likelihood, which is the minimization criteria for the maximum likelihood estimation.

The first R² value of the logistic regression is Cox & Snell’s R² (although other Pseudo R² exists, we focus on the 2 that are part of SPSS).  


This R² expresses the improvement of the full model with all variables included over the Block 0 model, that only includes the intercept

* Build models to provide arguments and evidence for or against the propositions below:
  + - The odds of Kobe making a shot decrease with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)
    - The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)
    - The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs
* Build a predictive model to classify shots as missed or made:
  + - A logistic regression model.
    - A Linear Discriminant Analysis (LDA) model
* The odds of Kobe making a shot decrease with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)

In the below table find the point estimate and CIs of odds ratio for every shot\_zone\_range.

The shot\_zone\_range = “less than 8ft” is the reference.



Shot distance is not used here to model since shot distance is not normally distributed and due to which the goodness of fit test (Hosmer and Lemeshow test) indicates that model is not a good fit.

Instead using shot\_zone\_range which is a similar measure at coarse level.

A sample calculation of odds ratio

W (successful shot) = eb0 + b1 (8-16)ft + b2 (16-24)ft + b3(24+ft)

Odds between “less than 8 ft” and “8-16 ft” is calculated as below

W(successful shot 8-16ft)  / W (successful shot less than 8 ft) = e(0.2899 – 0.5546) / e(0.2899) = 0.57

The confidence interval of the odds of successful shot in zone range 8-16ft to 8 ft is 0.52 to 0.63. This interval does not contain zero as a value so null hypothesis which states that odds of successful shots at range 8-16ft is same as odds of successful shot in range less than 8ft is rejected (p < 0.0001).

From the above model,

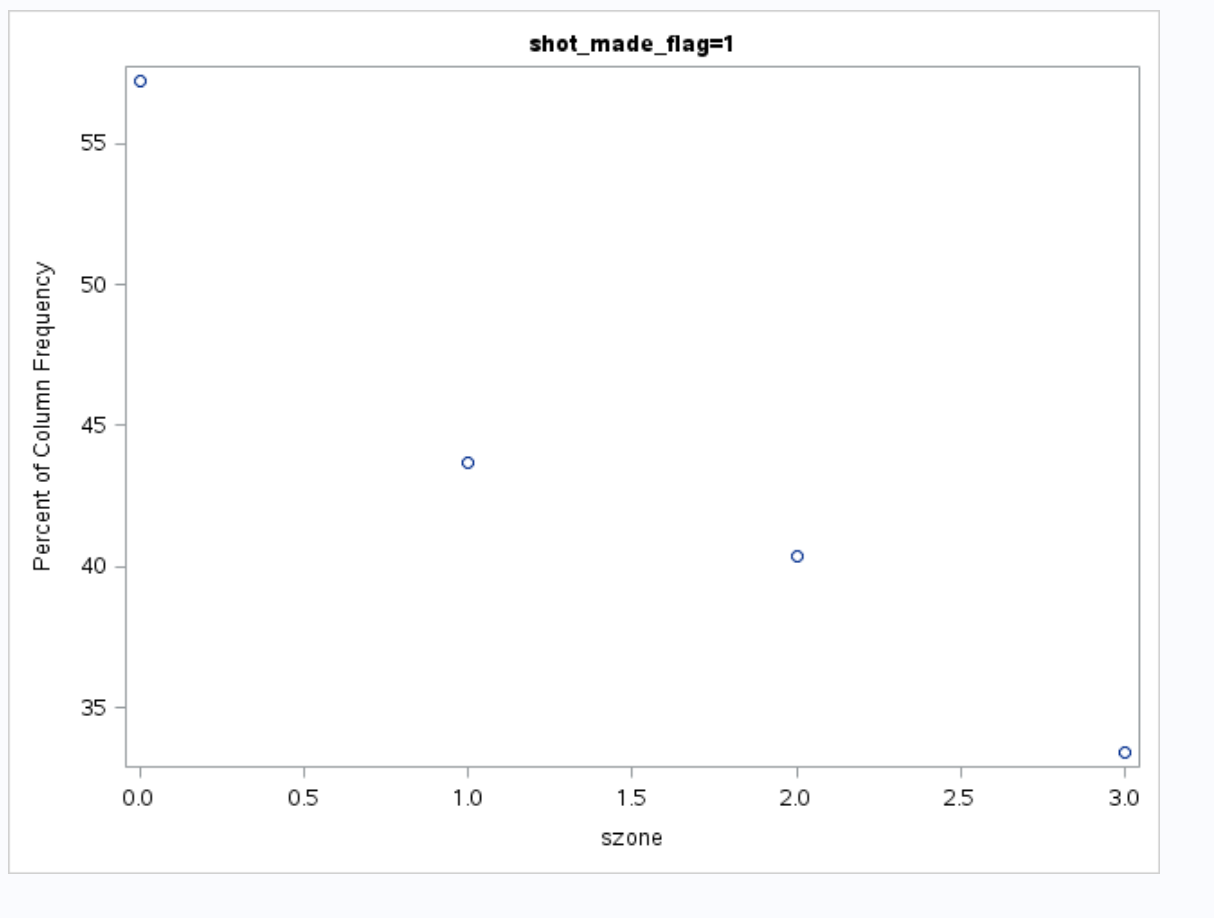
1. The odds of kobe making a successful shot if the distance from the hoop is in the range of 8-16ft is estimated to be 0.57 times that of hoop being with in 8ft range (p < 0.0001)
2. The odds of kobe making a successful shot if the distance from the hoop is in the range of 16-24ft is estimated to be 0.54 times that of hoop being with in 8ft range (p < 0.0001)
3. The odds of kobe making a successful shot if the distance from the hoop is in the range of 24+ft is estimated to be 0.37 times that of hoop being with in 8ft range (p < 0.0001)

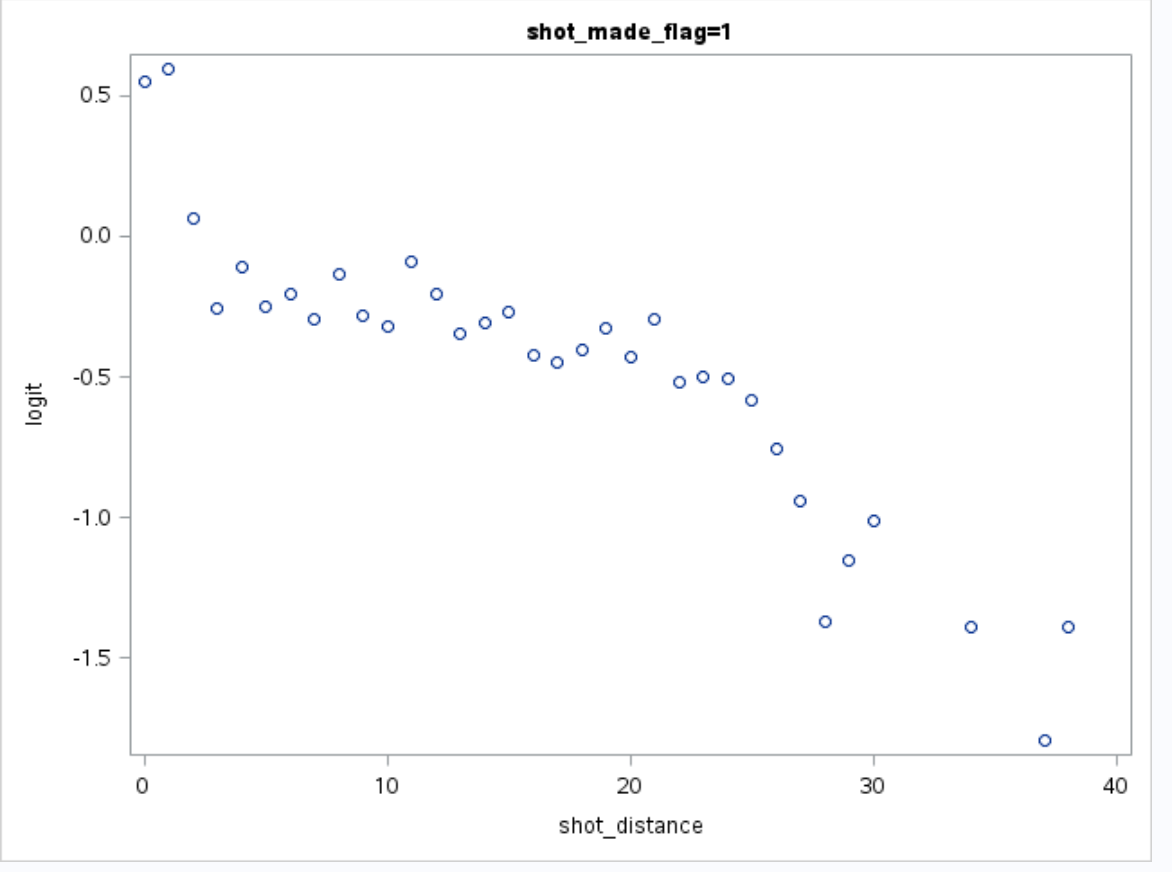
* The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)

The shot made variation w.r.t shot\_zone\_range coded values as below.

From the below plot we can see that proportions of successful shot made goes down linearly.

|  |
| --- |
|  |

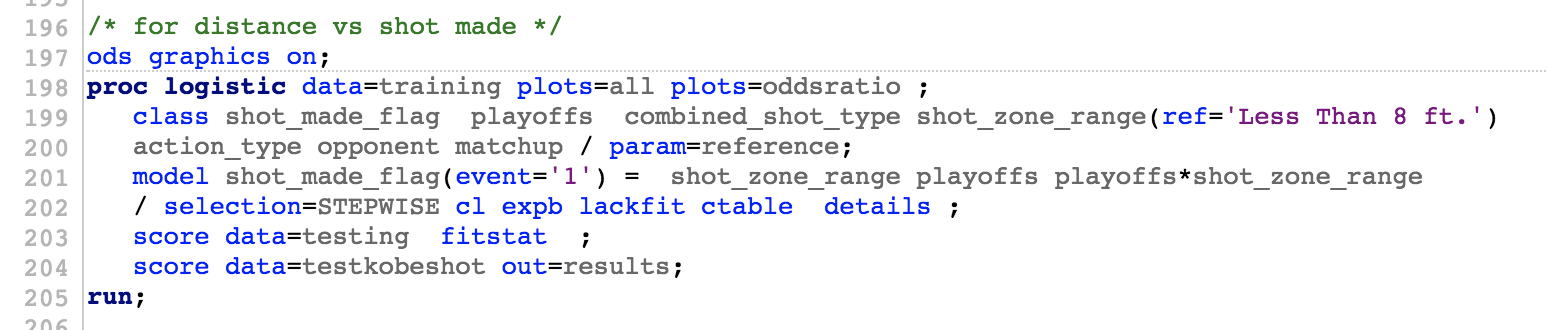


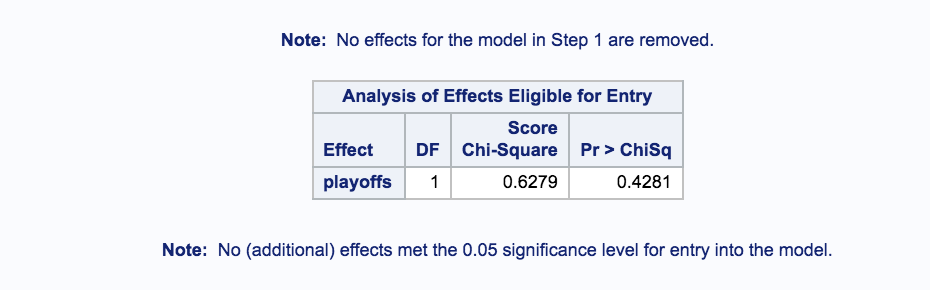


* The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs. Quantify your findings with statistical evidence one way or the other. (Tests, CIs, plots, etc.)

With stepwise selection model as shown below, we can see that playoffs gets eliminated at Step 1 since it does not qualify to be part of the model considering 0.05 to be the significance level for entry into the model.

So, playoffs do not have any influence on the odds of making a successful shot.





**Logistic Model and its parameter estimates**

| **Parameter Estimates and Wald Confidence Intervals** | | | | |
| --- | --- | --- | --- | --- |
| **Parameter** |  | **Estimate** | **95% Confidence Limits** | |
| **Intercept** |  | -3.2350 | -4.7341 | -1.7359 |
| **action\_type** | **Alley Oop Dunk Sh** | 2.9891 | 1.5609 | 4.4173 |
| **action\_type** | **Alley Oop Layup s** | 0.7866 | 0.0359 | 1.5372 |
| **action\_type** | **Cutting Layup Sho** | 0.7484 | -1.5253 | 3.0220 |
| **action\_type** | **Driving Bank shot** | 14.0612 | -2387.9 | 2416.1 |
| **action\_type** | **Driving Dunk Shot** | 3.2154 | 2.1978 | 4.2330 |
| **action\_type** | **Driving Finger Ro** | 1.4930 | 0.7889 | 2.1971 |
| **action\_type** | **Driving Floating** | -0.3788 | -3.1736 | 2.4160 |
| **action\_type** | **Driving Hook Shot** | 0.3979 | -1.0050 | 1.8008 |
| **action\_type** | **Driving Jump shot** | -0.5890 | -1.6684 | 0.4905 |
| **action\_type** | **Driving Layup Sho** | 0.8304 | 0.5685 | 1.0923 |
| **action\_type** | **Driving Reverse L** | 0.6624 | -0.0147 | 1.3395 |
| **action\_type** | **Driving Slam Dunk** | 13.9152 | -466.2 | 494.0 |
| **action\_type** | **Dunk Shot** | 1.0177 | 0.5174 | 1.5180 |
| **action\_type** | **Fadeaway Bank sho** | 1.9020 | -0.1624 | 3.9665 |
| **action\_type** | **Fadeaway Jump Sho** | -0.2028 | -0.4591 | 0.0536 |
| **action\_type** | **Finger Roll Layup** | 0.9023 | -0.2398 | 2.0444 |
| **action\_type** | **Finger Roll Shot** | 0.2897 | -0.8247 | 1.4040 |
| **action\_type** | **Floating Jump sho** | 0.7532 | 0.1171 | 1.3892 |
| **action\_type** | **Follow Up Dunk Sh** | 13.8989 | -893.0 | 920.8 |
| **action\_type** | **Hook Bank Shot** | 13.8974 | -1182.7 | 1210.5 |
| **action\_type** | **Hook Shot** | -0.8389 | -1.4715 | -0.2062 |
| **action\_type** | **Jump Bank Shot** | 0.8493 | 0.4288 | 1.2699 |
| **action\_type** | **Jump Hook Shot** | 1.2619 | -0.2866 | 2.8103 |
| **action\_type** | **Jump Shot** | -1.2137 | -1.4004 | -1.0270 |
| **action\_type** | **Layup Shot** | -0.7874 | -1.0315 | -0.5432 |
| **action\_type** | **Pullup Bank shot** | -0.4026 | -2.0160 | 1.2108 |
| **action\_type** | **Pullup Jump shot** | 0.5442 | 0.1929 | 0.8955 |
| **action\_type** | **Putback Dunk Shot** | -0.2896 | -3.0788 | 2.4996 |
| **action\_type** | **Putback Layup Sho** | 0.4579 | -1.2536 | 2.1694 |
| **action\_type** | **Reverse Dunk Shot** | 2.2532 | 1.0579 | 3.4485 |
| **action\_type** | **Reverse Layup Sho** | 0.2855 | -0.0893 | 0.6604 |
| **action\_type** | **Reverse Slam Dunk** | 13.9552 | -784.9 | 812.8 |
| **action\_type** | **Running Bank shot** | 1.5523 | 0.3239 | 2.7807 |
| **action\_type** | **Running Dunk Shot** | 1.6845 | -0.4221 | 3.7912 |
| **action\_type** | **Running Finger Ro** | -14.4740 | -1399.3 | 1370.4 |
| **action\_type** | **Running Hook Shot** | 1.4531 | 0.2096 | 2.6966 |
| **action\_type** | **Running Jump Shot** | 0.8455 | 0.5574 | 1.1336 |
| **action\_type** | **Running Layup Sho** | -0.0623 | -0.8522 | 0.7277 |
| **action\_type** | **Running Pull-Up J** | -14.5741 | -2413.5 | 2384.4 |
| **action\_type** | **Running Reverse L** | 0.1337 | -1.6690 | 1.9364 |
| **action\_type** | **Running Tip Shot** | -14.5425 | -2413.5 | 2384.4 |
| **action\_type** | **Slam Dunk Shot** | 3.5433 | 2.5295 | 4.5571 |
| **action\_type** | **Step Back Jump sh** | 0.0481 | -0.4839 | 0.5802 |
| **action\_type** | **Tip Layup Shot** | -0.2809 | -3.0669 | 2.5050 |
| **action\_type** | **Tip Shot** | -0.9267 | -1.4173 | -0.4361 |
| **action\_type** | **Turnaround Bank s** | 1.1437 | 0.3554 | 1.9319 |
| **action\_type** | **Turnaround Fadeaw** | -0.1343 | -0.4651 | 0.1964 |
| **action\_type** | **Turnaround Hook S** | 0.3078 | -1.4007 | 2.0163 |
| **shot\_zone\_range** | **16-24 ft.** | 0.3571 | 0.2069 | 0.5074 |
| **shot\_zone\_range** | **24+ ft.** | 0.2017 | 0.0446 | 0.3587 |
| **shot\_zone\_range** | **8-16 ft.** | 0.1449 | -0.00649 | 0.2963 |
| **shot\_zone\_range** | **Back Court Shot** | -13.3416 | -380.2 | 353.5 |
| **seconds\_remaining** |  | 0.00253 | 0.000490 | 0.00456 |
| **avgnoisedb** |  | 0.0362 | 0.0206 | 0.0518 |

The Parameters of the model are explained in terms of odds ratio.

The complete odds ratio table is as below.

Odds Ratio Estimates and its interpretation

Interpretation of a sample parameter is as below

Parameter: **action\_type Alley Oop Dunk Sh vs Turnaround Jump S= 19.86**

This means , when action type is “Alley Oop Dunk” with “Turnaround Jump” as reference action type, the odds of having successful shot gets multiplied by 19.86 times.

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **action\_type Alley Oop Dunk Sh vs Turnaround Jump S** | 19.868 | 4.763 | 82.875 |
| **action\_type Alley Oop Layup s vs Turnaround Jump S** | 2.196 | 1.037 | 4.651 |
| **action\_type Cutting Layup Sho vs Turnaround Jump S** | 2.114 | 0.218 | 20.532 |
| **action\_type Driving Bank shot vs Turnaround Jump S** | >999.999 | <0.001 | >999.999 |
| **action\_type Driving Dunk Shot vs Turnaround Jump S** | 24.914 | 9.005 | 68.924 |
| **action\_type Driving Finger Ro vs Turnaround Jump S** | 4.451 | 2.201 | 8.999 |
| **action\_type Driving Floating vs Turnaround Jump S** | 0.685 | 0.042 | 11.201 |
| **action\_type Driving Hook Shot vs Turnaround Jump S** | 1.489 | 0.366 | 6.054 |
| **action\_type Driving Jump shot vs Turnaround Jump S** | 0.555 | 0.189 | 1.633 |
| **action\_type Driving Layup Sho vs Turnaround Jump S** | 2.294 | 1.766 | 2.981 |
| **action\_type Driving Reverse L vs Turnaround Jump S** | 1.939 | 0.985 | 3.817 |
| **action\_type Driving Slam Dunk vs Turnaround Jump S** | >999.999 | <0.001 | >999.999 |
| **action\_type Dunk Shot vs Turnaround Jump S** | 2.767 | 1.678 | 4.563 |
| **action\_type Fadeaway Bank sho vs Turnaround Jump S** | 6.700 | 0.850 | 52.797 |
| **action\_type Fadeaway Jump Sho vs Turnaround Jump S** | 0.816 | 0.632 | 1.055 |
| **action\_type Finger Roll Layup vs Turnaround Jump S** | 2.465 | 0.787 | 7.724 |
| **action\_type Finger Roll Shot vs Turnaround Jump S** | 1.336 | 0.438 | 4.072 |
| **action\_type Floating Jump sho vs Turnaround Jump S** | 2.124 | 1.124 | 4.012 |
| **action\_type Follow Up Dunk Sh vs Turnaround Jump S** | >999.999 | <0.001 | >999.999 |
| **action\_type Hook Bank Shot vs Turnaround Jump S** | >999.999 | <0.001 | >999.999 |
| **action\_type Hook Shot vs Turnaround Jump S** | 0.432 | 0.230 | 0.814 |
| **action\_type Jump Bank Shot vs Turnaround Jump S** | 2.338 | 1.535 | 3.561 |
| **action\_type Jump Hook Shot vs Turnaround Jump S** | 3.532 | 0.751 | 16.616 |
| **action\_type Jump Shot vs Turnaround Jump S** | 0.297 | 0.246 | 0.358 |
| **action\_type Layup Shot vs Turnaround Jump S** | 0.455 | 0.356 | 0.581 |
| **action\_type Pullup Bank shot vs Turnaround Jump S** | 0.669 | 0.133 | 3.356 |
| **action\_type Pullup Jump shot vs Turnaround Jump S** | 1.723 | 1.213 | 2.449 |
| **action\_type Putback Dunk Shot vs Turnaround Jump S** | 0.749 | 0.046 | 12.178 |
| **action\_type Putback Layup Sho vs Turnaround Jump S** | 1.581 | 0.285 | 8.753 |
| **action\_type Reverse Dunk Shot vs Turnaround Jump S** | 9.518 | 2.880 | 31.454 |
| **action\_type Reverse Layup Sho vs Turnaround Jump S** | 1.330 | 0.915 | 1.936 |
| **action\_type Reverse Slam Dunk vs Turnaround Jump S** | >999.999 | <0.001 | >999.999 |
| **action\_type Running Bank shot vs Turnaround Jump S** | 4.723 | 1.383 | 16.131 |
| **action\_type Running Dunk Shot vs Turnaround Jump S** | 5.390 | 0.656 | 44.308 |
| **action\_type Running Finger Ro vs Turnaround Jump S** | <0.001 | <0.001 | >999.999 |
| **action\_type Running Hook Shot vs Turnaround Jump S** | 4.276 | 1.233 | 14.830 |
| **action\_type Running Jump Shot vs Turnaround Jump S** | 2.329 | 1.746 | 3.107 |
| **action\_type Running Layup Sho vs Turnaround Jump S** | 0.940 | 0.426 | 2.070 |
| **action\_type Running Pull-Up J vs Turnaround Jump S** | <0.001 | <0.001 | >999.999 |
| **action\_type Running Reverse L vs Turnaround Jump S** | 1.143 | 0.188 | 6.934 |
| **action\_type Running Tip Shot vs Turnaround Jump S** | <0.001 | <0.001 | >999.999 |
| **action\_type Slam Dunk Shot vs Turnaround Jump S** | 34.580 | 12.547 | 95.307 |
| **action\_type Step Back Jump sh vs Turnaround Jump S** | 1.049 | 0.616 | 1.786 |
| **action\_type Tip Layup Shot vs Turnaround Jump S** | 0.755 | 0.047 | 12.244 |
| **action\_type Tip Shot vs Turnaround Jump S** | 0.396 | 0.242 | 0.647 |
| **action\_type Turnaround Bank s vs Turnaround Jump S** | 3.138 | 1.427 | 6.903 |
| **action\_type Turnaround Fadeaw vs Turnaround Jump S** | 0.874 | 0.628 | 1.217 |
| **action\_type Turnaround Hook S vs Turnaround Jump S** | 1.360 | 0.246 | 7.510 |
| **shot\_zone\_range 16-24 ft. vs Less Than 8 ft.** | 1.429 | 1.230 | 1.661 |
| **shot\_zone\_range 24+ ft. vs Less Than 8 ft.** | 1.223 | 1.046 | 1.431 |
| **shot\_zone\_range 8-16 ft. vs Less Than 8 ft.** | 1.156 | 0.994 | 1.345 |
| **shot\_zone\_range Back Court Shot vs Less Than 8 ft.** | <0.001 | <0.001 | >999.999 |
| **seconds\_remaining** | 1.003 | 1.000 | 1.005 |
| **avgnoisedb** | 1.037 | 1.021 | 1.053 |

| **A** |
| --- |

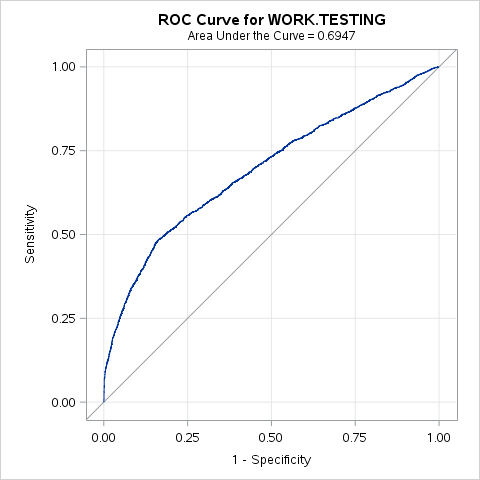
Evaluating Model Performance

This logistic model is 68% accurate. About 32% of the classifications can be error prone.

The **sensitivity** is the proportion of true positive = 46.7

The **specificity** is the proportion of true negative = 85.8

| **Classification Table** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Prob Level** | **Correct** | | **Incorrect** | | **Percentages** | | | | |
| **Event** | **Non- Event** | **Event** | **Non- Event** | **Correct** | **Sensi- tivity** | **Speci- ficity** | **False POS** | **False NEG** |
| **0.450** | 3024 | 6871 | 1134 | 3458 | 68.3 | 46.7 | 85.8 | 27.3 | 33.5 |



| **Fit Statistics for SCORE Data** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Set** | **Total Frequency** | **Log Likelihood** | **Error Rate** | **AIC** | **AICC** | **BIC** | **SC** | **R-Square** | **Max-Rescaled R-Square** | **AUC** | **Brier Score** |
| **WORK.TESTING** | 6208 | -3871.0 | 0.3215 | 7852.017 | 7853.018 | 8222.365 | 8222.365 | 0.120567 | 0.161335 | 0.694683 | 0.21284 |
| **WORK.TESTING** | 6208 | -3871.0 | 0.3215 | 7852.017 | 7853.018 | 8222.365 | 8222.365 | 0.120567 | 0.161335 | 0.694683 | 0.21284 |

Conclusion

Kobe Bryant’s successful shots do have some pattern. The success of shots depends on some of the features in the data set. Some action types like Reverse **Slam Dunk, Slam Dunk Shot have good success rate. These shots do not seem to have any interaction effect with shot\_zone\_range which means distance to hoops and action type do not have combined effect. The shot\_zone\_range itself is a significant effect which means all shots are affected by distance to hoops irrespective of type of action.**

**Kobe also gets affected by the average noise in the arena. The more cheering sound seem to make Kobe make successful shots.**

**The SAS code to this project can be found at**

[**https://github.com/pradeep17j/SMU\_MSDS/tree/master/AppliedStats\_Proj2**](https://github.com/pradeep17j/SMU_MSDS/tree/master/AppliedStats_Proj2)

APENDIX