Modeling Kobe

Daniel Byrne

April 12, 2019

## Abstract

Sports data modeling has been a staple of video game development for many years. The major leagues leagues NBA, NFL, NHL have all partnered with various software development companies to produce ever more realistic video games depicting the leagues’s top stars acting and more importantly scoring as they do in real life. With the $20 billion video game industry fueled by the now growing e-sports segment, ever more realistic models are needed for developers to build the characters in their games. We explore a common dataset of Kobe Bryant’s depicting the shots he made and missed over his 20 year career and try to build a model that would predict with a high degree of probability the likelihood of his making or missing a shot. We show that our final model is moderately successful in predicting a shot. We speculate that additional data points commonly captured in sport statistics such as whether or not the shot was contested could add specificity to the model.

## Introduction

Kobe Bryan is a retired professional basketball player who spent 20 years with the Los Angles Lakers. Kobe entered the NBA directly out of Lower Menton High School in Pennsylvania. He won 5 NBA Championships, was selected to the All-Star team 18 times, and won 2 Olympic gold medals. He is also widely considered one of the greatest basketball players of all time.

Using 20 years of data on Kobe’s shots made and shots missed, we explore potential models that attempt to predict whether or not his shot went in. The data set project2Data.xlsx contains the location on the floor some surrounding circumstances of every shot he attempted in the NBA. Free Throw data is not included in the dataset. We attempt to build a model from this data that can predict with a reasonable degree of certainty whether the shot went in (shot\_made\_flag = 1) or missed (shot\_made\_flag = 0). We tested our final model against the held out project2Pred.xlsx dataset which was not used in training or testing the iterative models.

This type of model could be used in building a simulation or video game mimicking Kobe’s game. Sports data modeling has been a staple of video game development for many years. The major leagues leagues NBA, NFL, NHL have all partnered with various software development companies to produce ever more realistic video games depicting the leagues’s top stars acting and more importantly scoring as they do in real life. With the $20 billion video game industry fueled by the now growing e-sports segment, ever more realistic models are needed for developers to build the characters in their games.

## Data Description

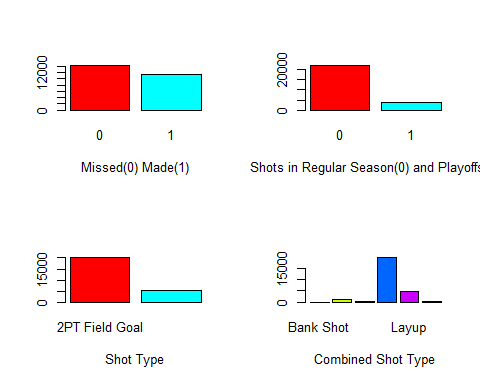
The field names are self explanatory. The predictors our analysis focused on are as follows:

|  |  |
| --- | --- |
| Data Label | Description |
| combined\_shot\_type | Type of shot combined with action |
| loc\_y | Vertical position on floor |
| minutes\_remaining | Minutes remaining in quarter |
| playoffs | Playoff game or not |
| seconds\_remaining | Seconds remaining in quarter |
| shot\_distance | Distance from goal |
| shot\_made\_flag | 1- shot went I, 0 - shot missed |
| shot\_type | 2pt or 3pt shot |
| attendance | The attendance in the stadium |
| arena\_temp | The average temperature during the game |
| avgnoisedb | The average noise level in dB during the game |

## Data Analysis

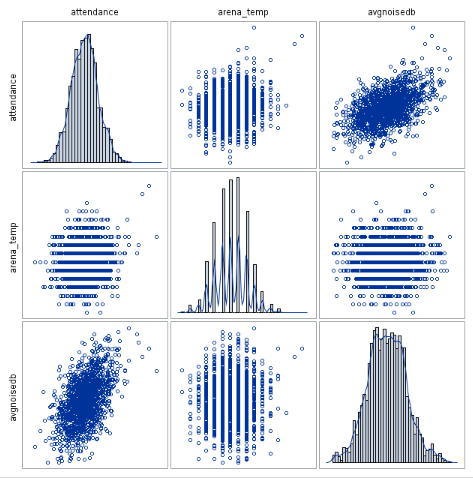
We evaluated a large but not exhaustive number of predictive variable combinations and potential models in our analysis. The following variables played a part in our final model

* Shot\_distance : We analyzed the hyphothesis that Kobe’s odds of making his shots decreased as the shot distance increased and whether or not this was a linear phenomenon.
* Shot\_type : we saw a statistically signifigant contribution from shot\_type in a number of our test models we led us to include this variable into our final model.
* Combined\_shot\_type : Likewise combined\_shot\_type showed a stitistically signifigant contribution in a number of models
* Playoffs : We included this predictor in a model used to evaluate Kobe’s performance in the regular season vs. the playoffs



We combined the following continuous variables into their principle components to include in out final model.

* Time\_remaining : We created this datapoint from minutes\_remaining\*60+seconds\_remaining
* Average Attendance
* Average Temperature
* Average Noise Level (dB)

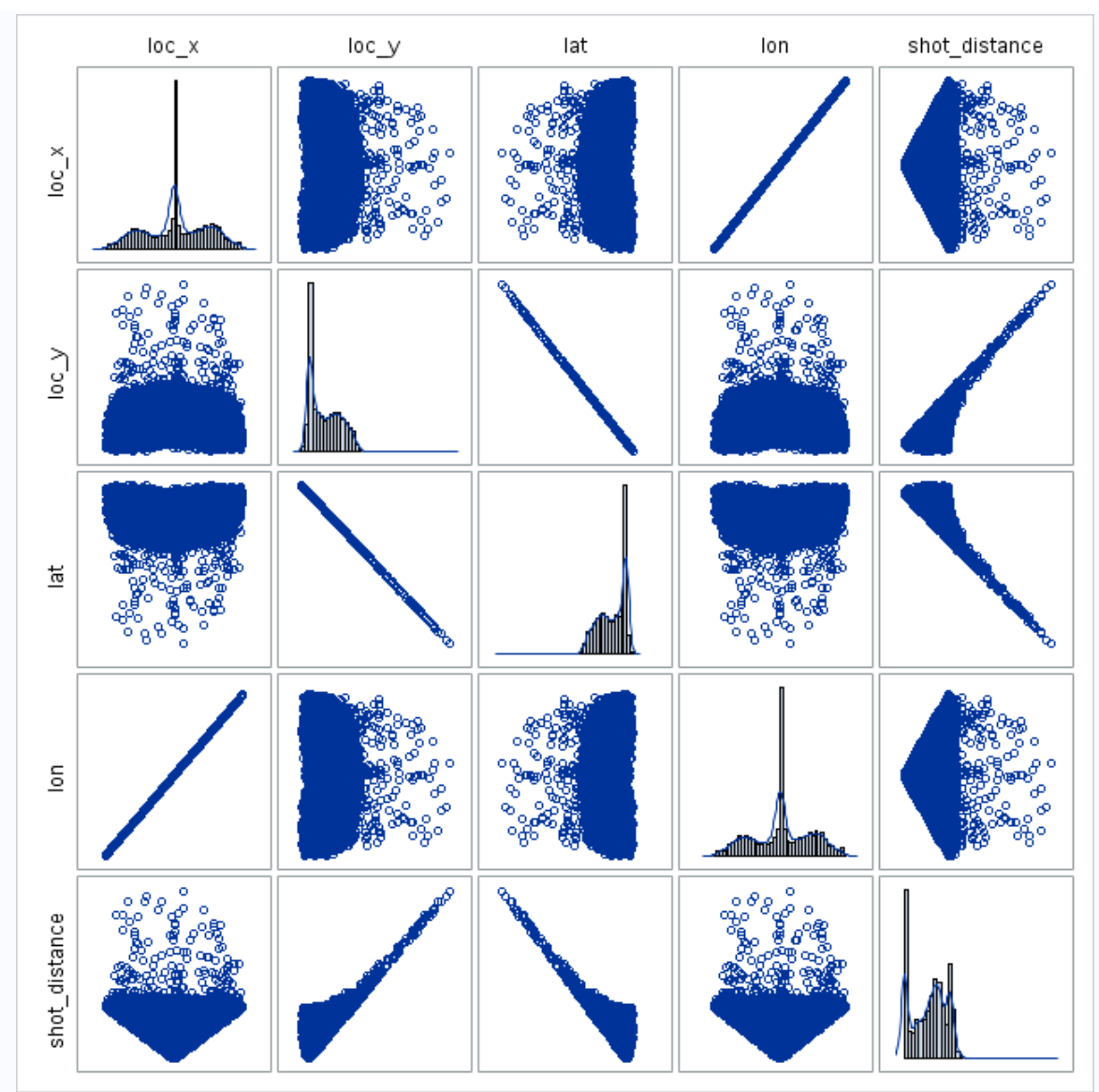


## Multicollinearity Analysis

High bivariate correlations were easy to spot bwhen we ran correlation calculations among our target predictors. We noticed signifigant correlations between loc\_y and shot\_distance, loc\_x and lon and loc\_y and lat. Coincidently we did not find models with both loc\_y and shot\_distance or with loc\_x and lon or loc\_y and lat to be good models due to their colinearity.

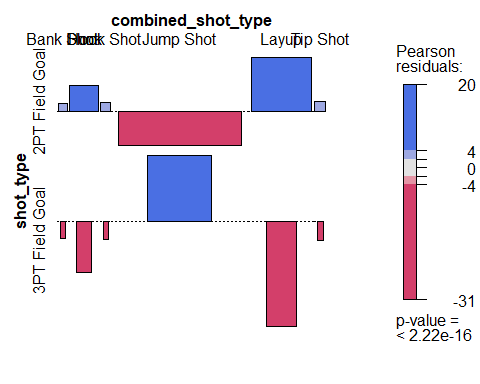
## loc\_x loc\_y lat lon  
## loc\_x 1.00000000 -0.01757819 0.01757819 1.00000000  
## loc\_y -0.01757819 1.00000000 -1.00000000 -0.01757819  
## lat 0.01757819 -1.00000000 1.00000000 0.01757819  
## lon 1.00000000 -0.01757819 0.01757819 1.00000000

## loc\_y shot\_distance  
## loc\_y 1.000000 0.818124  
## shot\_distance 0.818124 1.000000



There are also some similarities betweeb categorical variables although categorical variables cannot be colinear. They do not represent linear measures in Euclidean space. We use chi-square tests to determine independence of categorical variables.

However it is helpful to visualize some of the similaities. The graph below shows high colinearity between layups and dunks, and a high degree of similarity with bankshots and tipshots. All of which are understandable. However all of these shot types are also in the same categorical variable, and as such not directly colinear with any otehr variables.



## Outlier Analysis:

Based on Cooks’s D data and plain data analysis, we are seeing no outlier present in the selected variables. Here is the first 5 observations with highest cook’s D value. Since these values are less than 3, we assume that there are no outliers.

Following variables are selected for analysis

proc reg data=kobe;  
model shot\_made\_flag = lat lon time\_remaining period playoffs shot\_distance attendance arena\_temp avgnoisedb / r;

output out=kbCook cookd=cooks student=students rstudent=studresid;  
run;

|  |  |  |
| --- | --- | --- |
| **Obs** | **recId** | **cooks** |
| **1** | 2694 | 0.000777 |
| **2** | 27990 | 0.000524 |
| **3** | 29122 | 0.00045 |
| **4** | 1074 | 0.000356 |
| **5** | 17734 | 0.000352 |

## Analysis Questions

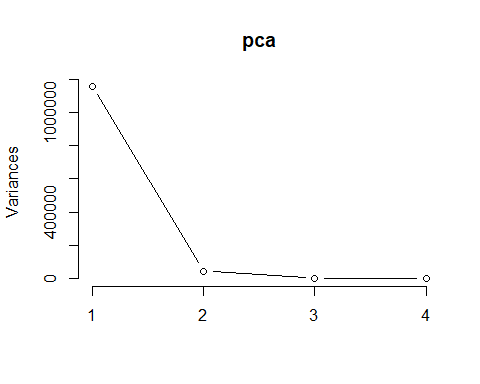
We set out to test the following hyphothesis:

1. The odds of Kobe making a shot decrease with respect to the distance he is from the hoop.
2. The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop.
3. 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.

To test these hyphothesis we evaluated several models. For the first we evelated a logisic regression model consisting of shot\_distance, shot\_type, combined\_shot\_type and a linear combination of the continuous variables time\_remaining, attendance, arena\_temp, avgnoisedb using principal componenet analysis. Principal component analysis is a statistical techinique that transforms a set of possibly correlated variables into a set of values of linearly uncorrelated variables.

Our PCA analysis revealed that the first orthoganol combination contributed nealy 95% of the variance of these variables where as the remaining transformations did not contribute signifigantly.

proc princomp cov prefix=k data=kobe out=kobe;  
var time\_remaining attendance arena\_temp avgnoisedb;  
run;



## Standard deviations (1, .., p=4):  
## [1] 1076.227723 208.308920 2.023717 1.961407  
##   
## Rotation (n x k) = (4 x 4):  
## PC1 PC2 PC3 PC4  
## time\_remaining 0.0008978239 9.999996e-01 -6.443052e-05 -9.139863e-06  
## attendance -0.9999989991 8.978220e-04 1.330830e-04 -1.085390e-03  
## arena\_temp -0.0001498820 -6.414719e-05 -9.998808e-01 1.543873e-02  
## avgnoisedb -0.0010831965 1.110603e-05 1.543888e-02 9.998802e-01

In final SAS model was then we only included k1, the first orthagonal combination of PCA analysis.

class shot\_type home combined\_shot\_type(ref='Jump Shot') /param=ref;  
 model shot\_made\_flag(event='1') = shot\_distance shot\_type k1 combined\_shot\_type

##   
## Call:  
## glm(formula = shot\_made\_flag ~ shot\_distance + combined\_shot\_type +   
## shot\_type, family = "binomial", data = kobe)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2953 -1.0223 -0.8908 1.2956 1.6256   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.488234 0.226523 6.570 5.04e-11 \*\*\*  
## shot\_distance -0.017137 0.003066 -5.589 2.29e-08 \*\*\*  
## combined\_shot\_typeDunk 1.071499 0.255857 4.188 2.82e-05 \*\*\*  
## combined\_shot\_typeHook Shot -1.250658 0.286975 -4.358 1.31e-05 \*\*\*  
## combined\_shot\_typeJump Shot -1.590542 0.226130 -7.034 2.01e-12 \*\*\*  
## combined\_shot\_typeLayup -1.219195 0.228348 -5.339 9.34e-08 \*\*\*  
## combined\_shot\_typeTip Shot -2.109243 0.283277 -7.446 9.63e-14 \*\*\*  
## shot\_type3PT Field Goal -0.171709 0.048310 -3.554 0.000379 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 35325 on 25696 degrees of freedom  
## Residual deviance: 33474 on 25689 degrees of freedom  
## AIC: 33490  
##   
## Number of Fisher Scoring iterations: 5

From the results in the table above, shot\_distance, combined\_shot\_type, and shot\_type all appear to be statistically signifigant.

## Wald test:  
## ----------  
##   
## Chi-squared test:  
## X2 = 914.9, df = 5, P(> X2) = 0.0

Furthermore the chi-squared test statistic of 914.9, with 5 degree of freedom and a p-value << 0 indicates that the overall effect of combined\_shot\_type is statistically significant as well.

**Model Comparison**:

**Loss Function:**

Based on the model run for above analysis, we compare the models against loss function. We found out LDA model as better than the Logistic model. Table below is showing only the first 5 records.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Better Model (LDA)** | | | **Worse Model (Logistic)** | | |
| **recId (first 5)** | **Actual Outcome for y** | **Py = Estimated Prob y = 1** | **Loss Function Term** | **Actual Outcome for y** | **Py = Estimated Prob y = 1** | **Loss Function Term** |
| 1 | 0 | 0.420650852 | -0.23706 | 0 | 0.395780639 | -0.21881 |
| 2 | 0 | 0.411259655 | -0.23008 | 0 | 0.391604564 | -0.21581 |
| 3 | 0 | 0.370070868 | -0.20071 | 0 | 0.366865655 | -0.1985 |
| 4 | 1 | 0.644965711 | -0.19046 | 1 | 0.92327104 | -0.03467 |
| 5 | 0 | 0.430796004 | -0.24473 | 0 | 0.399979941 | -0.22183 |
|  | **Loss Function Value for whole data** | | **-0.20022** | **Loss Function Value for whole data** | | **-0.21342** |