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

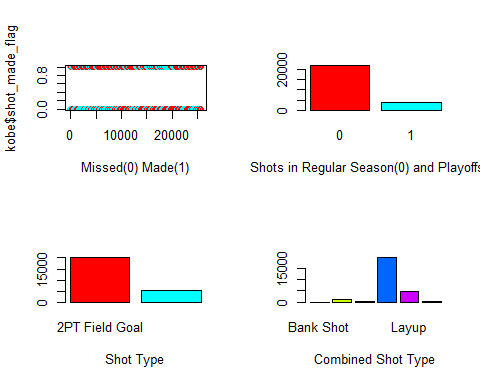
The field names are self explanatory and contain the following attributes:

|  |  |
| --- | --- |
| Data Label | Description |
| action\_type | Type of shot action |
| combined\_shot\_type | Type of shot combined with action |
| game\_event\_id | Game id used to identify shots within the same game |
| game\_id | Game id used to identify shots within the same game |
| lat | Horizontal position on floor |
| loc\_x | Horizontal position on floor |
| loc\_y | Vertical position on floor |
| lon | Vertical position on floor |
| minutes\_remaining | Minutes remaining in quarter |
| period | Quarter of play |
| playoffs | Playoff game or not |
| season | Season label |
| 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 |
| shot\_zone\_area | General location of shot |
| shot\_zone\_basic | General location of shot |
| shot\_zone\_range | General range from basket |
| team\_id | ID for LA Lakers |
| team\_name | Always LA Lakers |
| game\_date | The date the game was played on |
| matchup | The team and who was home vs away |
| opponent | The opposing team |
| shot\_id | Unique shot ID |
| 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

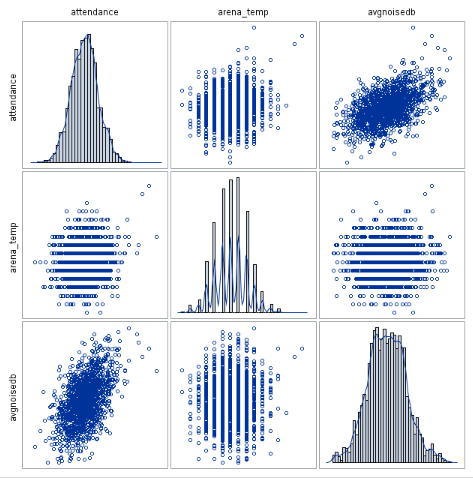
We evaluated a large but not exhaustive number of predictive variable combinations and potential models in our analysis. The following variables played a part in our final model

* 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)

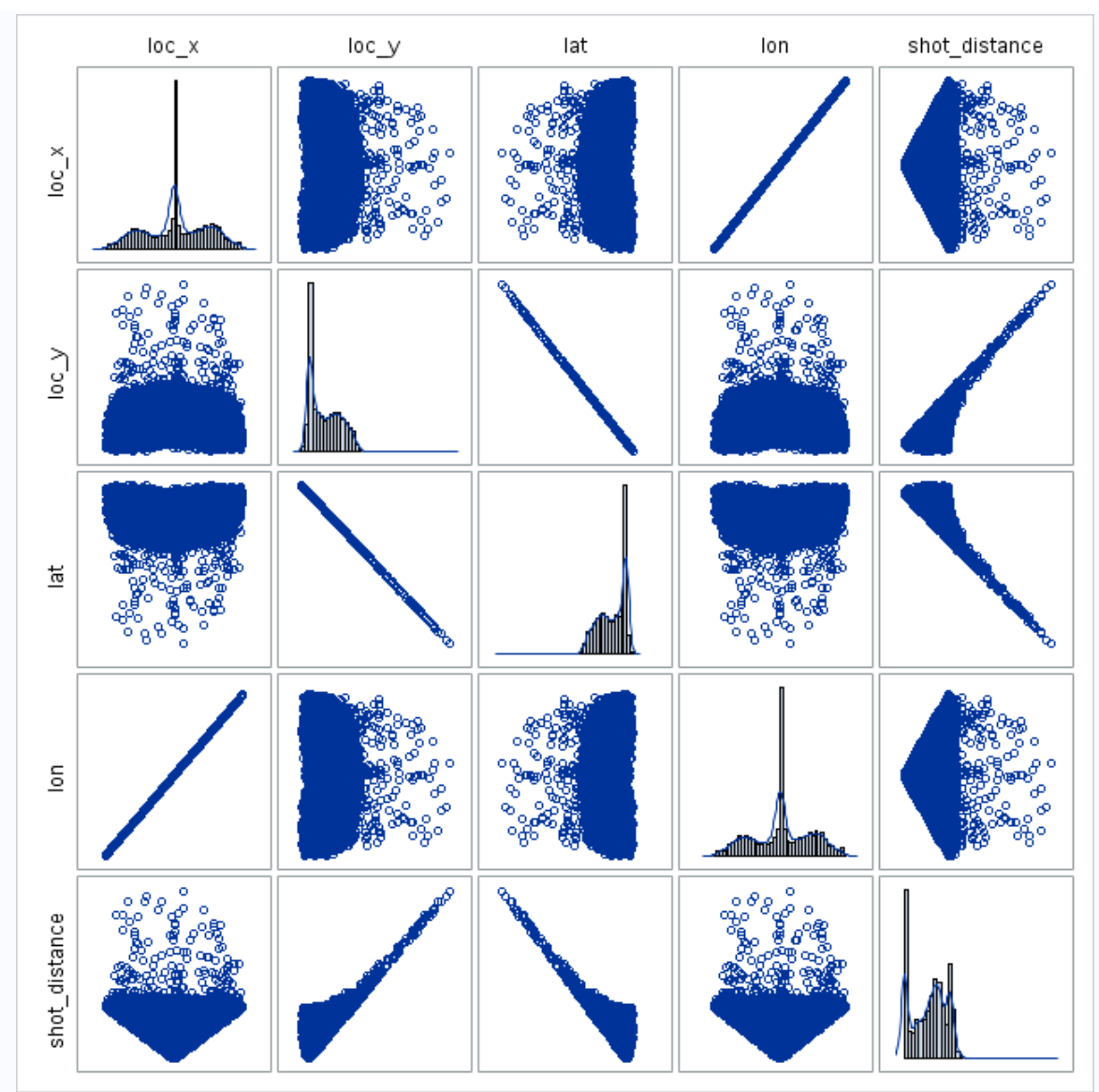


## Multicollinearity Analysis

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.

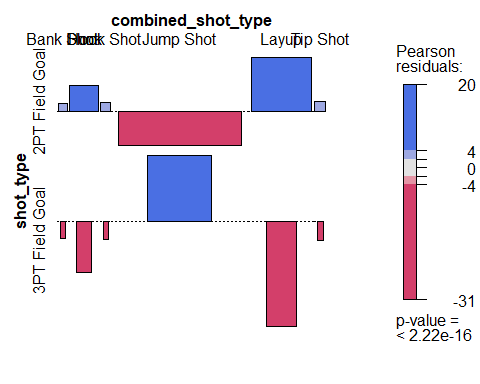
## loc\_x loc\_y lat lon  
## loc\_x 1.00000000 -0.01757819 0.01757819 1.00000000  
## loc\_y -0.01757819 1.00000000 -1.00000000 -0.01757819  
## lat 0.01757819 -1.00000000 1.00000000 0.01757819  
## lon 1.00000000 -0.01757819 0.01757819 1.00000000

## 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