Introduction:

Kobe Bryant marked his retirement from basketball by scoring 60 points in his final game as a member of the Los Angeles Laker team on Wednesday, April 12, 2016. Starting to play professional basketball at the age of 17, Kobe earned the sport’s highest accolades throughout his long career. Using 20 years of data on Kobe's shots made and shots missed.

EDA:

During Exploratory Data Analysis, we are assuming following features/variables are of no effect on the model.

* game\_event\_id
* recId
* game\_id
* season
* game\_date
* team\_id
* team\_name

Also, we assume the multi collinearity of following variables

* lat, loc\_x, loc\_y and lon --> So, we had gone with lat, lon
* matchup, opponent --> So, we had gone with opponent

Data Transformation:

To build the model and further data analysis, we have transformed the class variables into continuous variables and used those for multicollinearity check

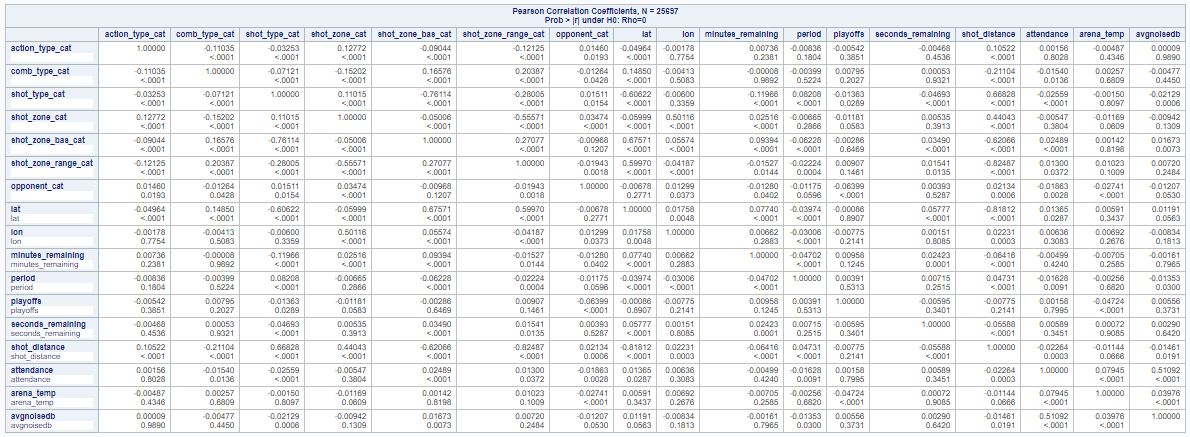
* action\_type\_cat
* comb\_type\_cat
* shot\_type\_cat
* shot\_zone\_cat
* shot\_zone\_bas\_cat
* shot\_zone\_range\_cat
* opponent\_cat

Outlier:

Based on Cooks’s D data and plain data analysis , we are seeing no outlier present in the selected variables. Here is the first 10 observations with highest cook’s D value. Since these values are less than 3, we assume that there are no outliers

|  |  |  |
| --- | --- | --- |
| **Obs** | **recId** | **cooks** |
| **1** | 2694 | 0.000777 |
| **2** | 27990 | 0.000524 |
| **3** | 29122 | 0.00045 |
| **4** | 1074 | 0.000356 |
| **5** | 17734 | 0.000352 |
| **6** | 28077 | 0.000343 |
| **7** | 13341 | 0.000336 |
| **8** | 19447 | 0.000336 |
| **9** | 17359 | 0.000333 |
| **10** | 28422 | 0.000331 |

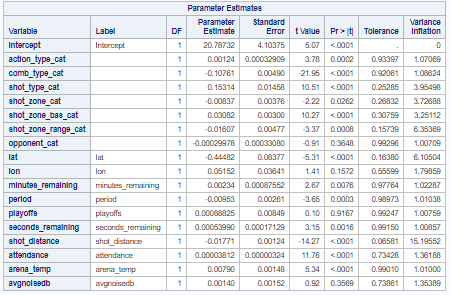
Multi Collinearity:



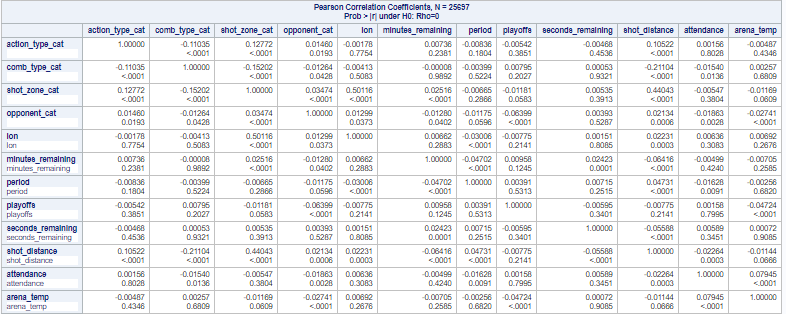
Upon inspection of these results, one is quickly drawn following conclusion:

* Negative correlation coefficient of lat and shot\_distance (-0.8810).
* Negative correlation Shot\_type and shot\_zone\_bas (-0.76114)
* Negative correlation Shot\_zone and shot\_zone\_range (-0.55571)
* Positive correlation lat and shot\_zone\_range (0.59970)
* Negative correlation shot\_distance and shot\_zone\_range (-0.82487)
* Negative correlation shot\_distance and shot\_zone\_base (-0.62066)
* Positive correlation attendance and avgnoisedb (0.5109)

We definitely have a case for further collinearity investigation here. This is further supported in our review of the parameter estimates results for VIF and Tol:



Based on above VIF and Tol table, we couldn’t get any conclusion, though higher VIF for shot\_distance, we had gone ahead without avgnoisedb, shot\_type\_cat, shot\_zone\_bas\_cat, shot\_zone\_range\_cat. Following is the result of the updated



VIF / Tol of the selected variables are below. Based on these values we can conclude that there is no multi collinearity between the selected variables and VIF is lesser than 10 and Tolerance is above 0.1

