**Introduction:**

Kobe Bryant marked his retirement from the NBA by scoring 60 points in his final game as a Los Angeles Laker on Wednesday, April 12, 2016. Using 2014- 2016 data on Kobe's swishes and misses, we built predictive models for Kobe Bryant’s shots with Logistic Regression techniques. We try to find the most important features in predicting his shots (shot\_made\_flag). The target variable for this project is binary, with 0-missed, 1-made. The objective is to predict Kobe’s shots with Logistic Regression models and compare the models’ performance.

**Data Description:**

The dataset, obtained from Kaggle, was sourced directly from the NBA. It contains every

goal attempted by Kobe in his 20-year career, a total of 30,697 shots. Of these, 5,000 were randomly selected to serve as a test set in file project2pred.xlsx, with their shot success labels removed.

The data contains a piece of information, with 25 variables in total.

1. **shot\_made\_flag**: 1 if shot made, 0 if not made.

2. **action\_type**: The type of shot attempted, like a jump shot, dunk, etc. Total there are 57 distinct values.

3. **combined\_shot\_type**: Classifies the shots under 6 categories: Bank Shot, Dunk,

Hook Shot, Jump Shot, Layup, and Tip Shot.

4. **game\_event\_id**: The ID of the game event (attempted shot) in the specific match being played.

Discarded for our purposes.

5. **game\_id**: The ID of the specific match. Also removed.

6. **lat**: The latitude of Kobe’s position during the shot attempt.

7. **lon**: The longitude.

8. **loc\_x**: The x-location on the court.

9. **loc\_y**: The y-location on the court.

10. **minutes\_remaining**: The minutes remaining in the specific match.

11. **period**: The period in the specific match.

12. **playoffs**: Indicator variable whether the match was in the playoffs or not.

13. **season**: The basketball season (2000, 2001, etc.)

14. **seconds\_remaining**: The seconds remaining in the specific match.

15. **shot\_distance**: The distance from which the shot was attempted, in ft.

16. **shot\_type**: 2pt or 3pt.

17. **shot\_zone\_area**: Area from which shot was attempted (Right, Left, Center, Back Court,

Right Center, Left Center)

18. **shot\_zone\_basic**: Further area information (Mid-range, restricted area, in the paint, above

the break 3, backcourt, left corner 3, right corner 3)

19. **shot\_zone\_range**: Range (<8 ft, 8-16, 16-24, 24+, backcourt)

20. **team\_id**: ID of Kobe’s team. Always the Lakers so discarded.21. team\_name: Name of Kobe’s team, the Lakers, so discarded.

22. **game\_date**: Date of the specific match. Discarded.

23. **matchup**: The two teams in the specific match. Since Kobe was always on the Lakers,

opponent contains all the information in the matchup. The matchup is thus discarded.

24. **opponent**: Opponent in the specific match.

25. **shot\_id**: ID (from 1 to 30,697) of the attempted shot.

And several more like game\_event\_id, game\_id, team\_id, team\_name, game\_date, and matchup. So the remains us with 18 predictors and one response, shot\_made\_flag.

The dataset is further cleaned by combining the 5 least attempted shots in action\_type into another category.

**Exploratory Data Analysis:**

1. **Address the need for any potential transformations:**

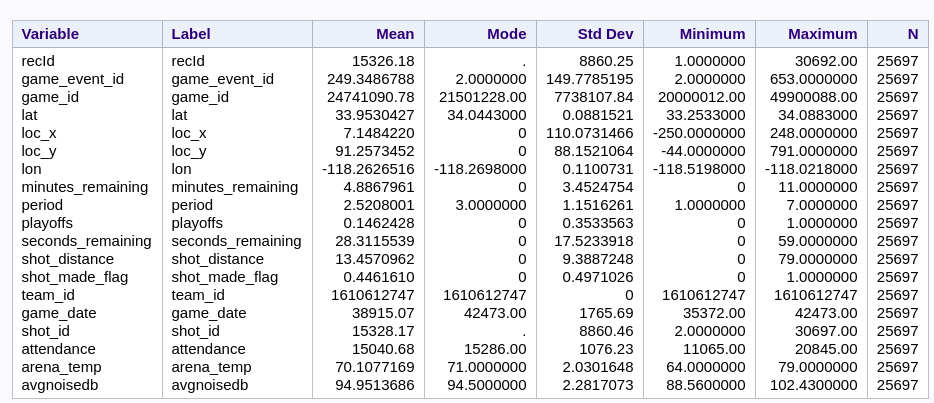
* **Differentiate the variables:**

From the above description, the features to check are -

**Numerical Features** - lat, loc\_x, loc\_y, lon, arena\_temp, avgnoisedb, game\_date, game\_event\_id, game\_id, period, playoffs, season, shot\_id, attendance,minutes\_remaining,seconds\_remaining, shot\_distance, shot\_made\_flag, team\_id.

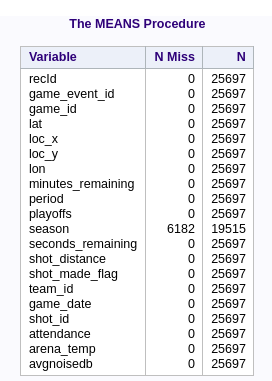
**Categorical Features** - action\_type, combined\_shot\_type, period, playoffs, season, shot\_type, shot\_zone\_area, shot\_zone\_basic, shot\_zone\_range.

* **Statistical Variable’s Behaviour:** The statistical part of the data consists of the variable calculation based on their lowest and highest value present, labeling of the data and other statistical presence. I have attached a fig consists of the all predictors information below:



**Fig.2**

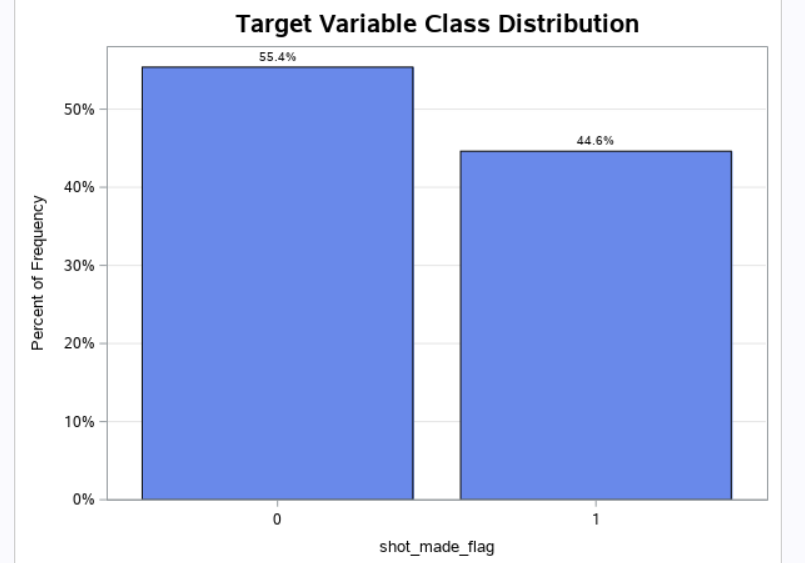
* **Checking Missing values in data:** With the consideration of all the predictors in the training file, we examine the presence of the missing values in the data set present. So that it could not be harmful to our model with overfitting and underfitting problem. Let us look at visual statistic:



**fig.3**

Here is the variable season has 6182 Number of missing values out off 25697.

* **Actual Accuracy of the shot\_made\_flag by Kobe:** The actual accuracy of the shot hits by the Kobe Bryant is as follows based on training data:



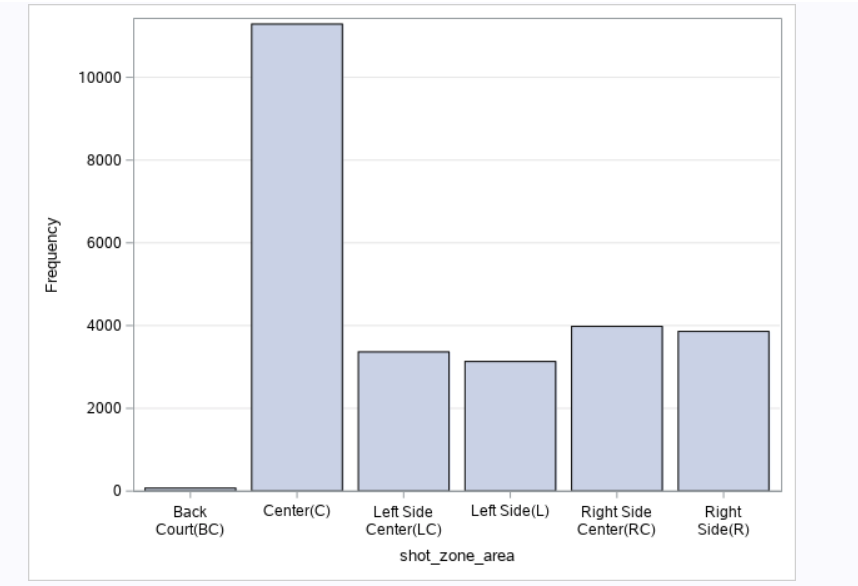
**Fig.4**

The percentage probability of the shot\_made\_flag by Kobe is 55.4% failure and 44.6% succeed.

* **shot\_zone\_area Frequency and Accuracy** :

The shot\_zone\_area does not vary as much as action\_type, but it is probably useful for

prediction to some extent. The shot played by Kobe based on area shows as follows:



**Fig.5**

The fig.4 shows the distribution of shot\_zone\_area by its shot frequency. The top frequency of the shot\_zone\_area is of only Central(C). The normal and most favorite shot zone area in the basketball game. And Kobe’s favorite shot zone area. Fig.6 a visualization of shot\_zone\_area, showing the on-court representation of each zone. As expected, shots from the backcourt are at such great range that the accuracy is extremely low.

Though the variation in accuracy appears large, we must note that backcourt shots are

rare. Of course, such shots are hardly ever attempted by any player, and Kobe is no exception.

Among the remaining zones, Kobe appears to slightly prefer the right and highly prefers the

center. He also shoots much better in the center, and slightly better in the right zones. A quick

search confirms that Kobe shoots with either hand, but that his right is dominant.

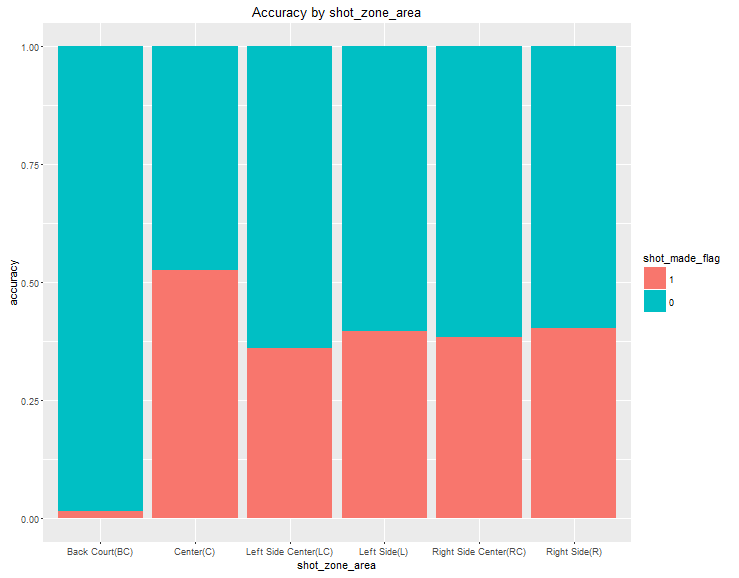
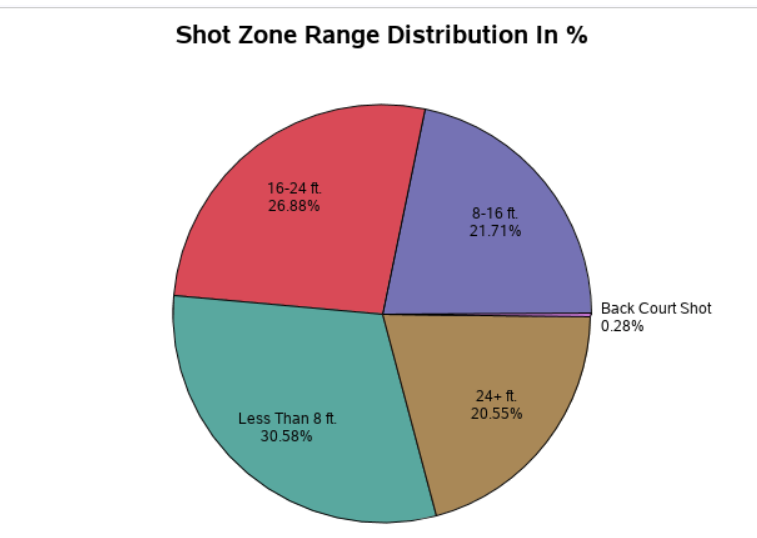


Fig.6

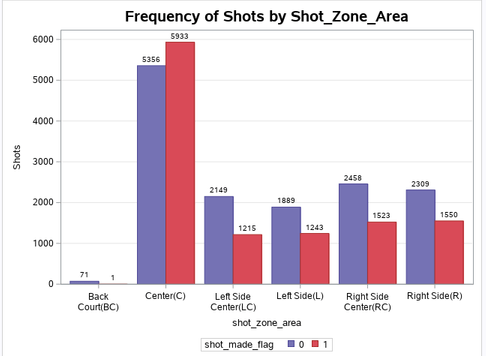
* **Shot\_zone\_range distribution in %:**

The next figure shows the visual distribution of the shot zone range in percentage occupied by when playing shots by Kobe.



**Fig.6**

The very important things need to be considered while having data exploration **on Kobe** is shots played by Kobe according to the zone area. Let's have a preview below:



* **Variations In shot\_zone\_basic:**

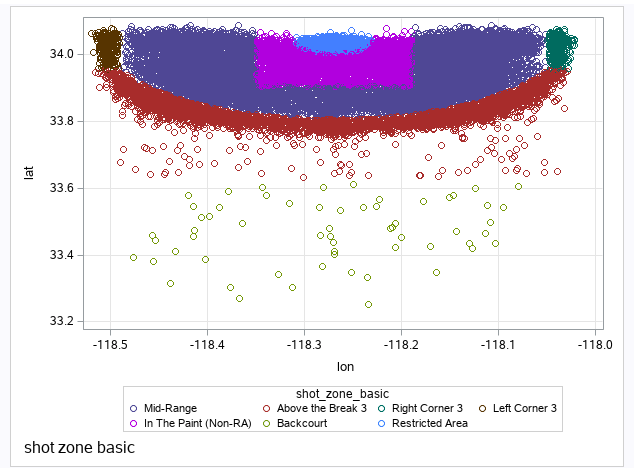
Fig.7 provides a visualization of the on-court locations, and Figure 6 the accuracy and number of shots by location.

Kobe’s accuracy by shot\_zone\_basic actually varies substantially even after we account

for the fact that shots from the corners and the backcourt are very rare. Surprisingly, Kobe’s left

corner accuracy is higher than right corner accuracy.

We are tempted to conclude shot\_zone\_basic should be included in our model. However, in actuality the variable is just describing the influence of range on accuracy. We will need to analyze shot\_zone\_range in and shot\_distance in order to decide on shot\_zone\_basic ’s conclusion. For instance, it could well be the case that shot\_zone\_range contains all the information of shot\_zone\_basic and more, or vice-versa.



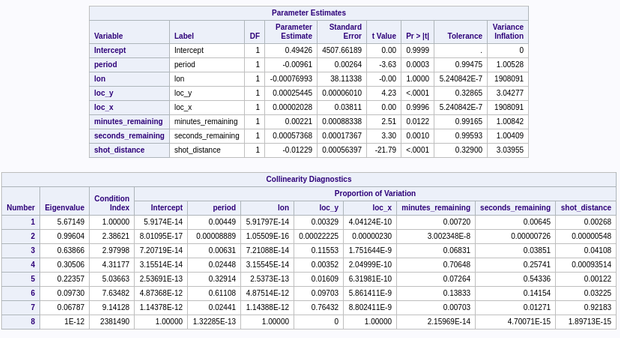
**Fig.7**

**2. Address & Identify the Outliers:**

An outlier is a data point that is distant from other common data points. For outlier detection, we can only consider the continuous-valued variables. We will not consider outlier for data moddelling because it may causes the overfitting of data or underfitting of data.

**3. Address and identify any multicollinearity:**

Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.



**Fig.8 Multicollinearity between the considered variable**

**Variable Selection:** Now we explore the remaining predictors, to find meaningful relationships with shot\_made\_flag. From here, we will proceed to take two approaches:

1. Using **LASSO[Least Absolute Shrinkage And Selection Operator]** to select variables.

2. Using our findings here and our intuition to guide our variable selection. And

3. With checking correlation factor

**Model Building:** So here we are usingLogistic Regressionmodel to fit them and to predict the data result.

Logistic Regressionis the categorical regression analysis to conduct when the dependent variable is dichotomous (binary) that is yes/no analysis. Like all regression analysis, the logistic regression is predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The figure below shows the insights presents in the logistic regression.

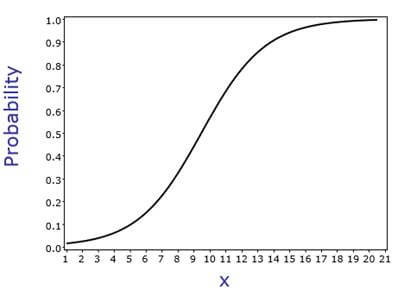


Fig.9

The logistic regression equation is :

y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x)) --------------------(1)

Using this method, the model building happens on the training data name as Kobe.xlsx and prediction based on project2pred.xlsx test data set. Using LASSO Regression and correlation of variable method we select the set of the final variables for the operations like model fitting and prediction. All variables were statistically significant at the 0.05 level.

The final selected set of variables is a combined\_shot\_type, action\_type, shot\_type ,shot\_zone\_area, shot\_zone\_basic, period, playoffs, shot\_zone\_range, loc\_y loc\_x, minutes\_remaining, seconds\_remaining & shot\_distance were also significant.

**Model Evaluation:**

1. **AUC Value:**

The AUC value for this model is 0.7030 i,e 70% accuracy.



Fig.10 AUC Value

**2. Specificity Vs Sensitivity:**

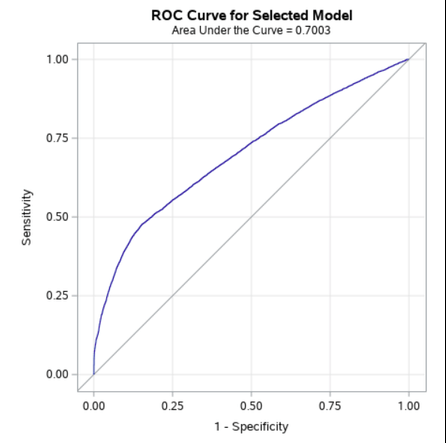


Fig.11 ROC Curve

**3. Misclassification Rate:**

The figure below shows the Misclassification rate and this misclassification rate generally called a confusion matrix to evaluate the actual and predicted rate.

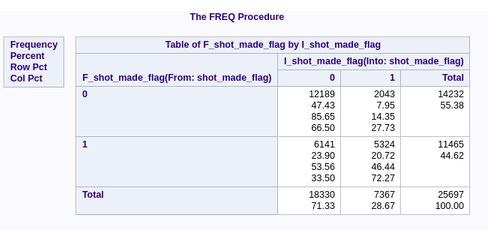


Fig.12 Misclassification Rate

**Conclusion:**

To evaluate this model, the predictions on the test set. Using this metric, the performance of this model could be compared. The **Log Loss** value for this model is about **11.182.** Unfortunately, performance is not spectacular. Still, the work presented here goes a long way in showing how applicable modern statistics is to sports. One can easily imagine the implications. Team coaches can easily maintain a model for each of their players, and analyze which shots they need to improve on and which they excel at, whether their performance dips with less time remaining in the match and more pressure, who on the team should shoot longer ranges, where to position each player, etc. The possibilities are endless. Hopefully, this exercise will as a step towards more creative quantitative analysis in sports.

.