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.

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.

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 an fig consists of the all predictors information below:

Variable	Label	Mean	Mode	Std Dev	Minimum	Maximum	N
recId	recId	15326.18	.	8860.25	1.0000000	30692.00	25697
game_event_id	game_event_id	249.3486788	2.0000000	149.7785195	2.0000000	653.0000000	25697
game_id	game_id	24741090.78	21501228.00	7738107.84	20000012.00	49900088.00	25697
lat	lat	33.9530427	34.0443000	0.0881521	33.2533000	34.0883000	25697
loc_x	loc_x	7.1484220	0	110.0731466	-250.0000000	248.0000000	25697
loc_y	loc_y	91.2573452	0	88.1521064	-44.0000000	791.0000000	25697
lon	lon	-118.2626516	-118.2698000	0.1100731	-118.5198000	-118.0218000	25697
minutes_remaining	minutes_remaining	4.8867961	0	3.4524754	0	11.0000000	25697
period	period	2.5208001	3.0000000	1.1516261	1.0000000	7.0000000	25697
playoffs	playoffs	0.1462428	0	0.3533563	0	1.0000000	25697
seconds_remaining	seconds_remaining	28.3115539	0	17.5233918	0	59.0000000	25697
shot_distance	shot_distance	13.4570962	0	9.3887248	0	79.0000000	25697
shot_made_flag	shot_made_flag	0.4461610	0	0.4971026	0	1.0000000	25697
team_id	team_id	1610612747	1610612747	0	1610612747	1610612747	25697
game_date	game_date	38915.07	42473.00	1765.69	35372.00	42473.00	25697
shot_id	shot_id	15328.17	.	8860.46	2.0000000	30697.00	25697
attendance	attendance	15040.68	15286.00	1076.23	11065.00	20845.00	25697
arena_temp	arena_temp	70.1077169	71.0000000	2.0301648	64.0000000	79.0000000	25697
avgnoisedb	avgnoisedb	94.9513686	94.5000000	2.2817073	88.5600000	102.4300000	25697

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:

The MEANS Procedure

Variable	N Miss	N
recId	0	25697
game_event_id	0	25697
game_id	0	25697
lat	0	25697
loc_x	0	25697
loc_y	0	25697
lon	0	25697
minutes_remaining	0	25697
period	0	25697
playoffs	0	25697
season	6182	19515
seconds_remaining	0	25697
shot_distance	0	25697
shot_made_flag	0	25697
team_id	0	25697
game_date	0	25697
shot_id	0	25697
attendance	0	25697
arena_temp	0	25697
avgnoisedb	0	25697

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:

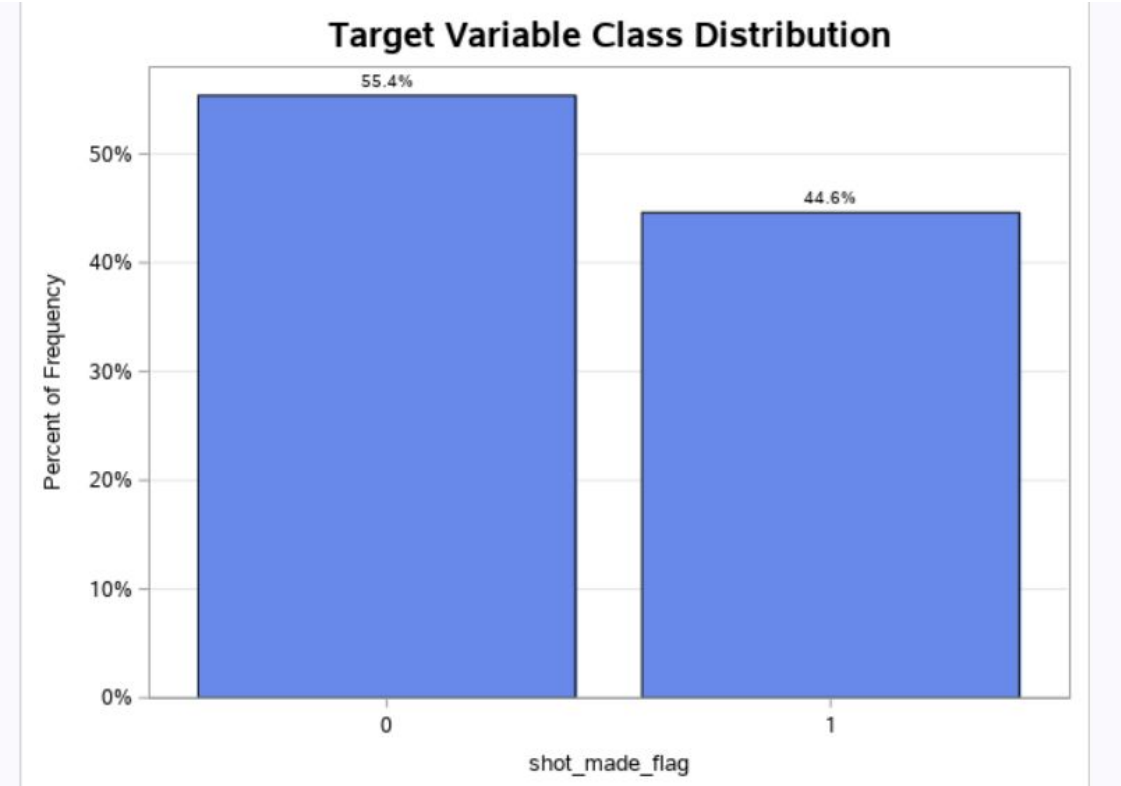


Fig.4

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:

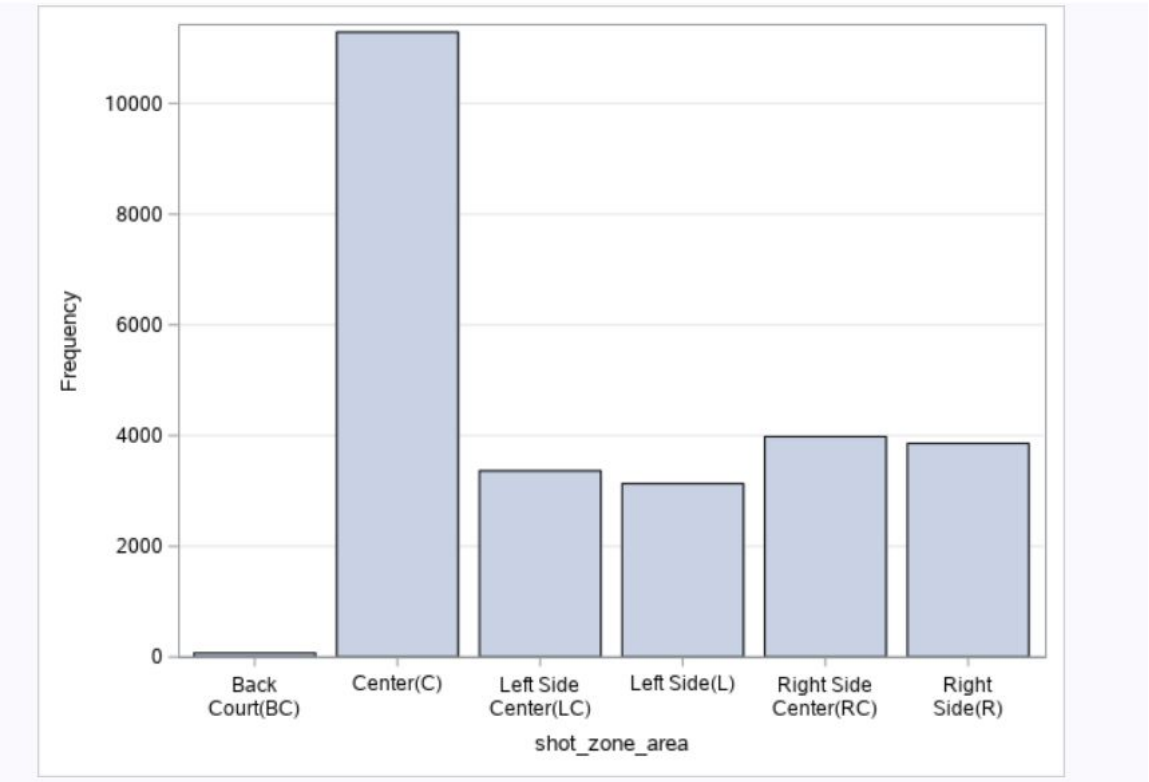


Fig.5

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.

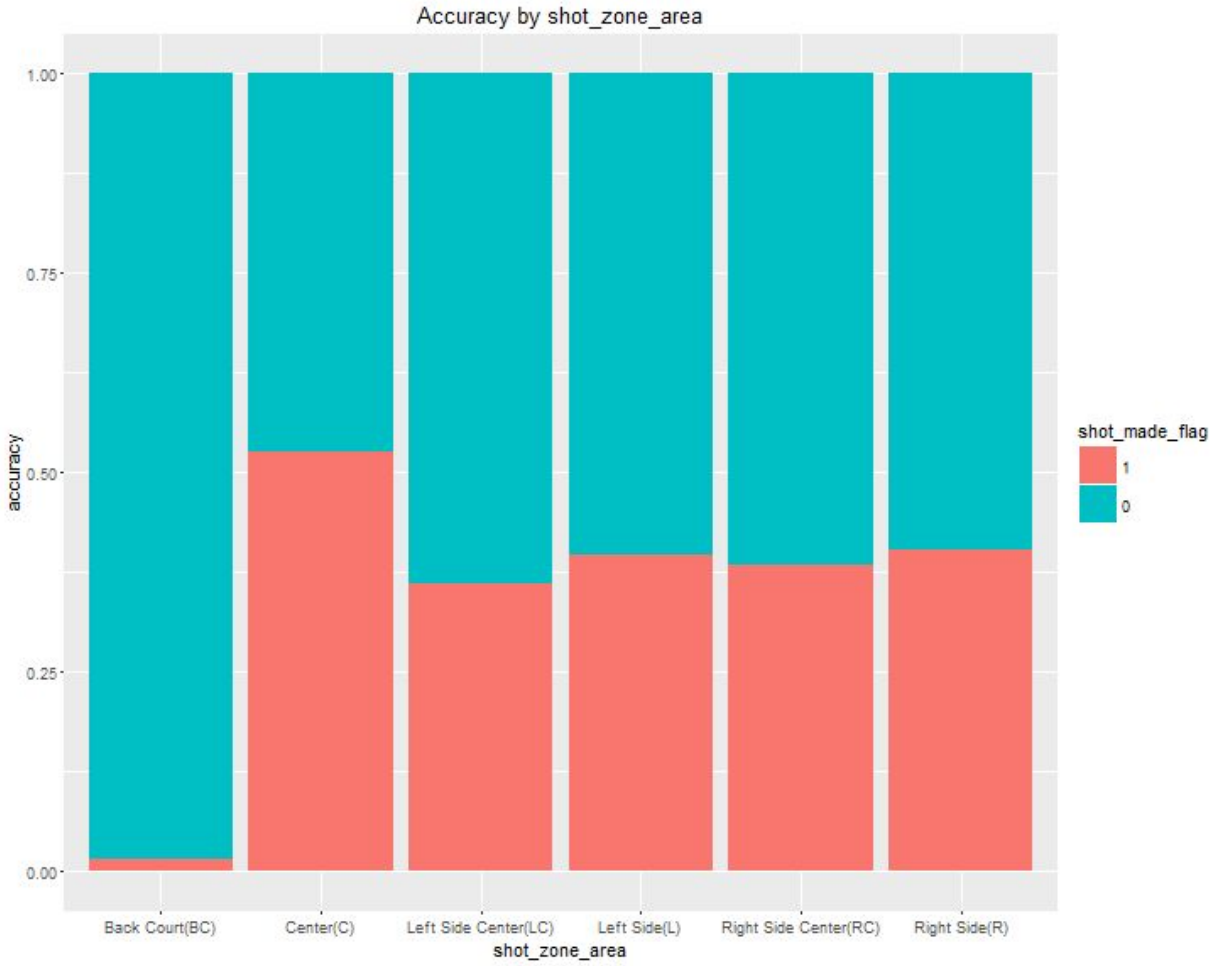


Fig.6

- **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.

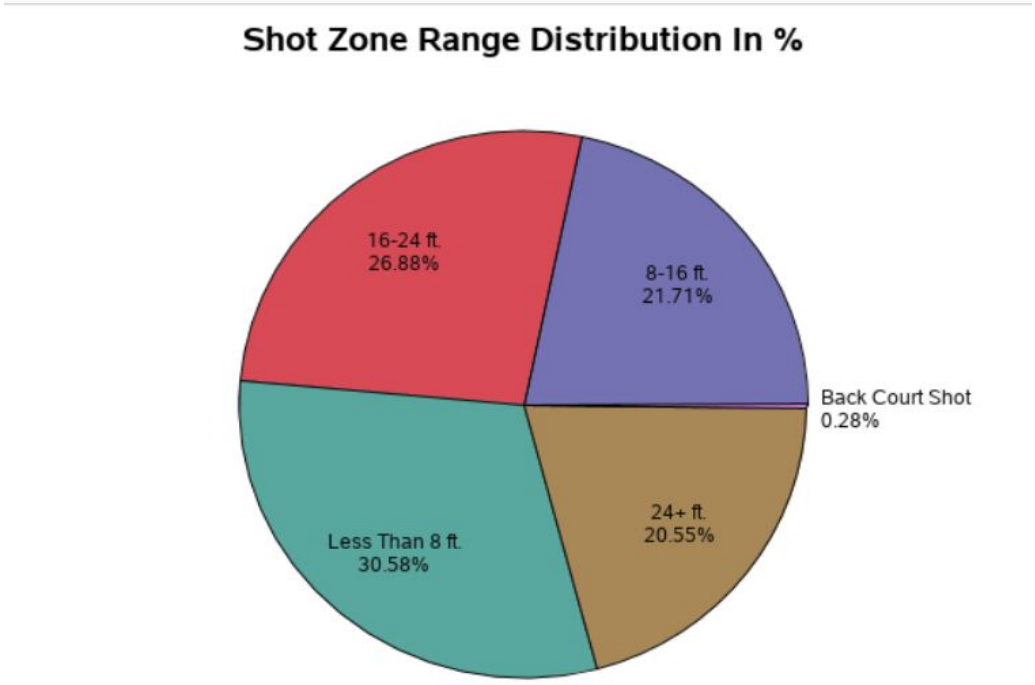
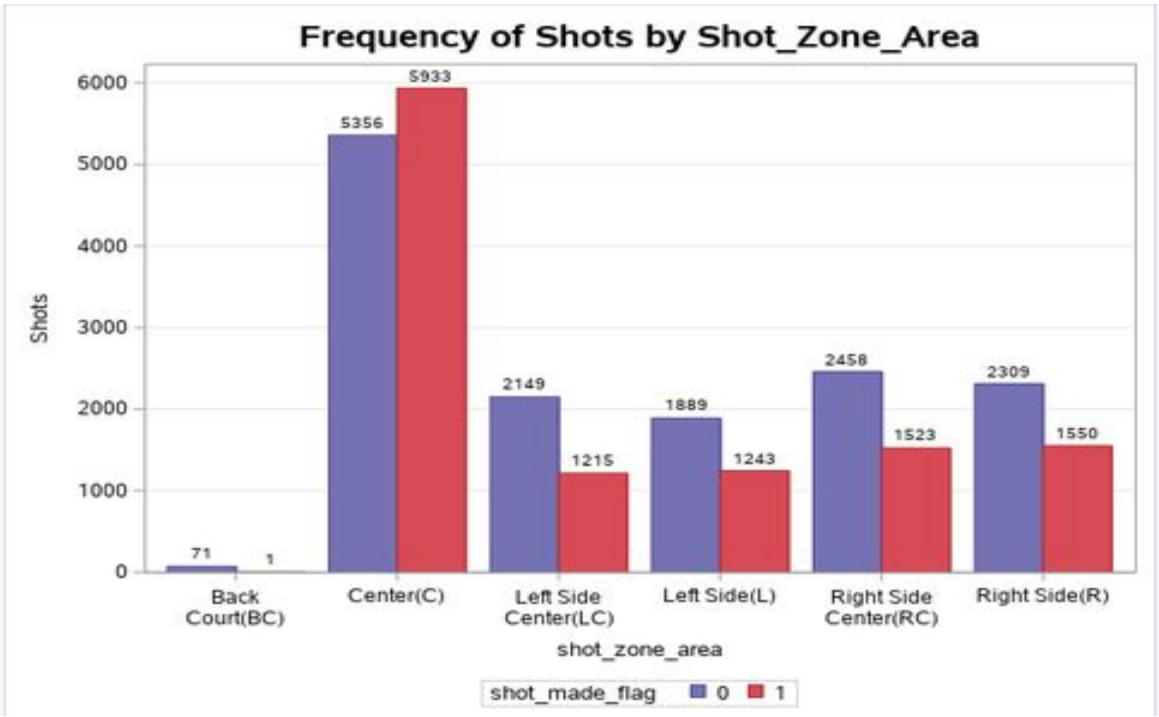


Fig.6

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.

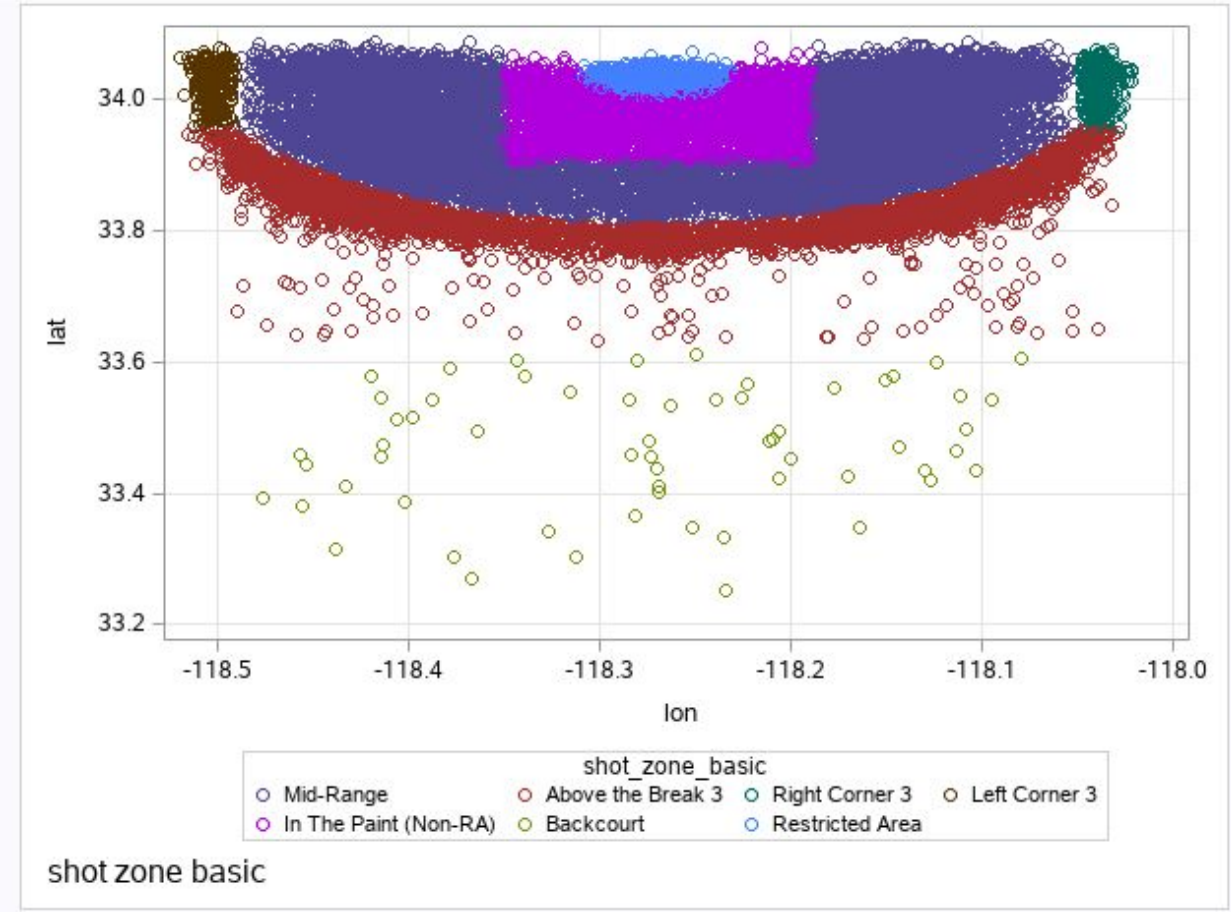


Fig.7

2. Address & Identify the Outliers:

An outlier is a data point that is distant from other similar points. They may be due to variability in the measurement or may indicate experimental errors. If possible, outliers should be excluded from the dataset. For outlier detection, we can only consider the continuous-valued variables.

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.

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.49426	4507.66189	0.00	0.9999	.	0
period	period	1	-0.00961	0.00264	-3.63	0.0003	0.99475	1.00528
lon	lon	1	-0.00076993	38.11338	-0.00	1.0000	5.240842E-7	1908091
loc_y	loc_y	1	0.00025445	0.00006010	4.23	<.0001	0.32865	3.04277
loc_x	loc_x	1	0.00002028	0.03811	0.00	0.9996	5.240842E-7	1908091
minutes_remaining	minutes_remaining	1	0.00221	0.00088338	2.51	0.0122	0.99165	1.00842
seconds_remaining	seconds_remaining	1	0.00057368	0.00017367	3.30	0.0010	0.99593	1.00409
shot_distance	shot_distance	1	-0.01229	0.00056397	-21.79	<.0001	0.32900	3.03955

Collinearity Diagnostics										
			Proportion of Variation							
Number	Eigenvalue	Condition Index	Intercept	period	lon	loc_y	loc_x	minutes_remaining	seconds_remaining	shot_distance
1	5.67149	1.00000	5.9174E-14	0.00449	5.91797E-14	0.00329	4.04124E-10	0.00720	0.00645	0.00268
2	0.99604	2.38621	8.01095E-17	0.00008889	1.05509E-16	0.00022225	0.00000230	3.002348E-8	0.00000726	0.00000548
3	0.63866	2.97998	7.20719E-14	0.00631	7.21088E-14	0.11553	1.751644E-9	0.06831	0.03851	0.04108
4	0.30506	4.31177	3.15514E-14	0.02448	3.15545E-14	0.00352	2.04999E-10	0.70648	0.25741	0.00093514
5	0.22357	5.03663	2.53691E-13	0.32914	2.5373E-13	0.01609	6.31981E-10	0.07264	0.54336	0.00122
6	0.09730	7.63482	4.87368E-12	0.61108	4.87514E-12	0.09703	5.861411E-9	0.13833	0.14154	0.03225
7	0.06787	9.14128	1.14378E-12	0.02441	1.14388E-12	0.76432	8.802411E-9	0.00703	0.01271	0.92183
8	1E-12	2381490	1.00000	1.32285E-13	1.00000	0	1.00000	2.15969E-14	4.70071E-15	1.89713E-15

Fig.8 Multicollinearity between the considered variable

Model Building: So here we are using **Logistic Regression** model to fit them and to predict the data result.

What is **Logistic Regression**?

Logistic Regression is 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.

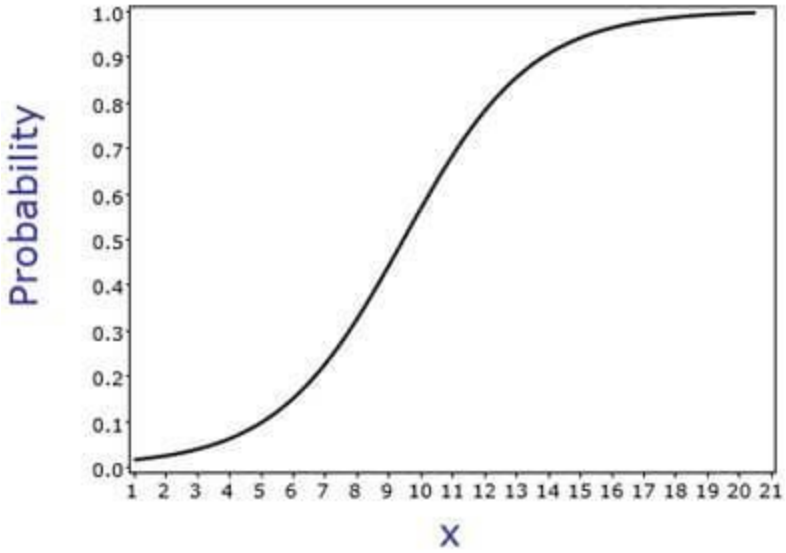


Fig.9

The logistic regression equation is :

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)}) \quad \text{----------(1)}$$

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.

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.KOBE1	25697	-15552.9	0.3171	31295.9	31296.61	32070.54	32070.54	0.151419	0.202682	0.70308	0.209513

Fig.10 AUC Value

2. Specificity Vs Sensitivity:

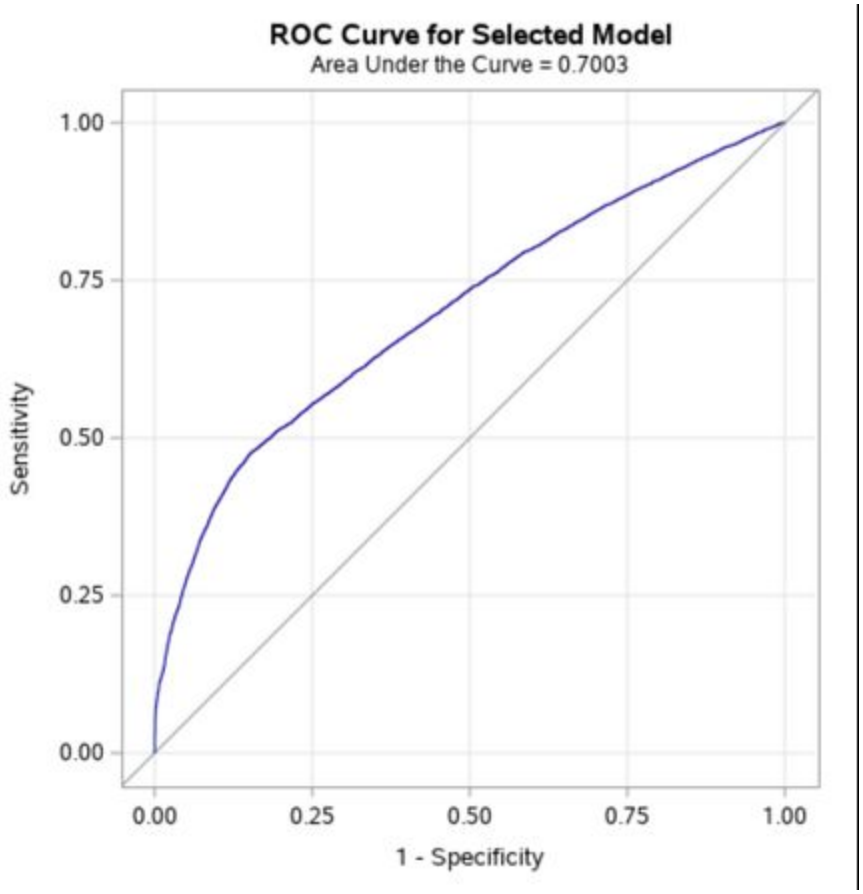


Fig.11 ROC Curve

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.

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of F_shot_made_flag by I_shot_made_flag		
	I_shot_made_flag(Into: shot_made_flag)		
F_shot_made_flag(From: shot_made_flag)	0	1	Total
0	12189 47.43 85.65 66.50	2043 7.95 14.35 27.73	14232 55.38
1	6141 23.90 53.56 33.50	5324 20.72 46.44 72.27	11465 44.62
Total	18330 71.33	7367 28.67	25697 100.00

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.

.