KOBE BRYANT SHOT SELECTION!!!

Allen Ansari, Yongjun Chu and [Solange Garcia de Alford](https://2ds.datascience.smu.edu/my/index.php?id=1394)

# Introduction

The objective of this project was to generate at least 2 models to best predict the shots made by Kobe Bryant on 5,000 occasions. To accomplish this our team:

1. Conducted extensive data analysis on the 30,697 records provided with data tracking the shots in Kobe Bryant’s 20 years’ career.
2. Built prediction models using two different statistical software packages, R and SAS:

## Logistic Regression Models using three different types of variable selection methods, Forward Selection, Backward Elimination and Stepwise Regression.

## Linear Discriminant Analysis (LDA) models with cross-validation using a training partition to devise the rule of the models’ algorithms and a test partition of the data to apply classification rules and predict the result of 5,000 shots made.

# Exploratory Data Analysis

## Data Description

## The original data set contains a total of 30,697 shot attempts by Kobe Bryant in his 20 years’ career with related data to his shots attempts. The data was partitioned as follows:

## 25,697 records which were used as the training set in the models.

## 5,000 records with the shot\_made\_flag values removed for use as test set in LDA predictions.

## The data contains 29 variables which listed as below including brief description.

|  |  |
| --- | --- |
| **action\_type:** The type of shot attempted, such as jump shot, dunk, etc.**combined\_shot\_type:** Classifies the shots under 6 larger categories: Bank Shot, Dunk, Hook Shot, Jump Shot, Layup, and Tip Shot.**Matchup:** The two teams in the specific match. Since Kobe was always on the Lakers and opponent contains all the information in matchup, we decided to reduced number of levels for this variable by dividing all games into Home and Away category.**Opponent:** Opponent in the specific match.**Season:** The basketball season (2000, 2001, etc.)**shot\_type:** includes categories 2pt or 3pt.**shot\_zone\_area:** Area from which shot was attempted (Right, Left, Center, Back Court, Right Center, Left Center)**shot\_zone\_basic:** Further area information (Mid-range, restricted area, in the paint, above the break 3, backcourt, left corner 3, right corner 3)**shot\_zone\_range:** Range (<8 ft, 8-16, 16-24, 24+, backcourt)**team\_name:** Name of Kobe’s team, the Lakers, so we decided to remove it from dataset.**arena\_temp:** average temperature of are*na***attendance**: Number of people who watched the game | **avgnoisedb:** Average noise level in the arena in decibels**game\_date:** Date of the specific match.**game\_event\_id:** **game\_id:** NBA game ID**lat:** The latitude of Kobe’s position during the shot attempt.**loc\_x:** The x-location on the court.**loc\_y:** The Y-location on the court.**Lon:** The longitude of Kobe’s position during the shot attempt.**minutes\_remaining:** The minutes remaining in the specific match**period**: The period in the specific match**playoffs:** binary, 1/0 values**recId****seconds\_remaining:** The seconds remaining in the specific match**shot\_distance: The** distance from which the shot was attempted, in ft.**shot\_id:** (from 1 to 30,697) of the attempted shot**shot\_made\_flag:** it is response variable and indicates if shot was successful (as 1) or not (as 0)**team\_id**: ID of Kobe’s team. Always the Lakers, so removed |

## Data Exploration

We found that transformations were not necessary for any of the variables. We found that we needed re-coding of several categorical variables into indicator variables for generating one of the LDA models when using SAS’ Proc DISCRIM’s Var statement. Proc LOGISTIC automatically re-codes the variables and presents these results in its resulting Design Variables matrix. Refer to Appendix D – Indicator Variables.

We explored the remaining predictors to find meaningful relationships with shot\_made\_flag. First, we looked for correlation between variables



Figure 1 shows there is high correlation between loc\_y and lat and shot\_distance which make sense because loc\_y, lat shows distance from basket. Additionally, there is high correlation between attendance and average noise db.

## Next, we examined location data as shown in Figure 2. By plotting loc\_y vs. loc\_x a visualization of the shots Kobe made and missed by location is depicted. It is rather difficult at first glance to discern any differences. One point that becomes obvious, however, is the impact of range. There are many more misses than makes at longer ranges, meaning the 3-point line and beyond. Within the 3-point area the data is too noisy to analyze.

A close up of a logo

Description automatically generated

Figure 2

Figure 3 provides a visualization of shot\_zone\_area, showing the on-court representation

of each zone. The accuracy of shots in each zone shown in figure 4. As expected, the accuracy for shots from the backcourt is extremely low.

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 3 Figure 4

Now, we are looking at shot\_zone\_basic. Figure 5 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 and surprisingly, Kobe’s left corner accuracy is higher than right corner accuracy.

A close up of a piece of paper

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 5 Figure 6

Plus other interesting facts on Kobe’s number of shots from different zone\_basics:

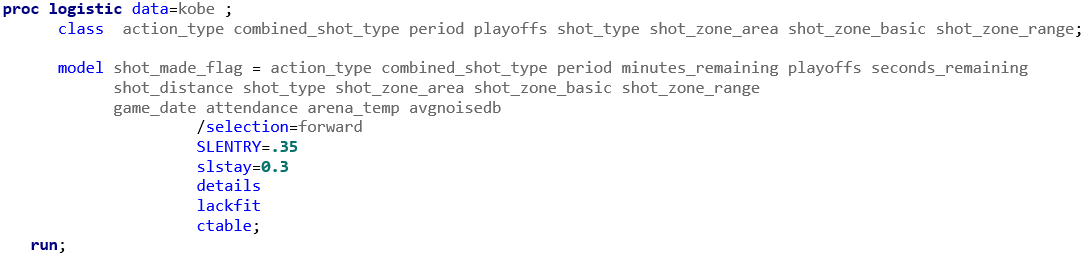
A close up of text on a black background

Description automatically generated

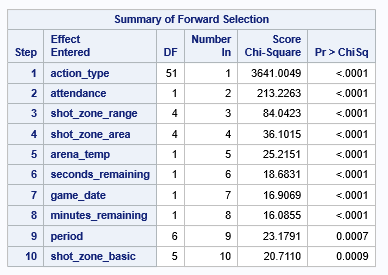
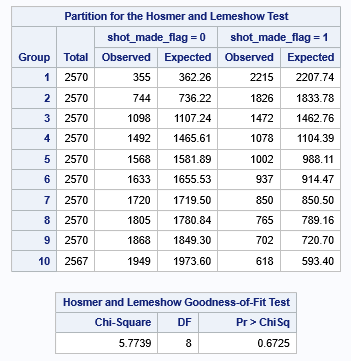
# Statistical Modeling

## Logistic Regression Model

We ran the model below with forward variable selection:



The model’s summary of forward selection is shown below.

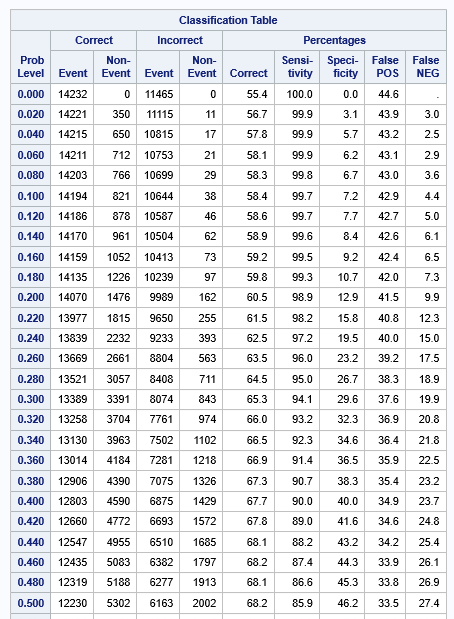
 

A question of interest in the project was what are the odds that Kobe would make a shot decrease with respect to the distance he is from the hoop. We found no evidence of this, the forward selection model found that shot\_distance was not a significant, as you can see in the results above. Similarly, the playoffs variable was not significant. We concluded that Kobe was performed well and similarly in all games.

1. CONCLUSION

# We failed to reject the null-hypothesis that the model has a good fit at alpha 0.05 p-value (0.6725) with a Hosmer and Lemeshow Goodness of Fit test.

# At the 50% probability level, our model had a 68.2% accuracy, a sensitivity of 85.9% and a specificity of 46.%. Please refer to the Classification Table below.



**Appendix D – Indicator Variables**

The screen prints below show the Design Variables table resulting from Proc LOGISTIC. The following variables were automatically re-coded: action\_type, combined\_shot\_type, period, playoffs, shot\_type, shot\_zone\_area, shot\_zone\_basic.

