|  |  |
| --- | --- |
|  | **PROJECT 2:**  **KOBE BRYANT SHOT SELECTION !!!** |

**OVERVIEW:**

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, can you predict which shots will be successful?

**DATA:**

The original data set contains the location and circumstances of every shot attempted by Bryant during his 20-year career. Your task is to predict whether the basket went in (shot\_made\_flag = 1) or missed (shot\_made\_flag = 0). The data for estimation is in project2KobeData.xlsx file, sheet modelData.

For this exercise, 5000 of the shot\_made\_flags have been removed from the original data set and are shown as missing values in the project2KobeData.xlsx file, sheet predData. These are the test set shots for which you must submit a classification. You are provided this sample classification file, project2Pred.xlsx, sheet predData, with the recIds needed for your predicted classification. Provide your estimated probabilities and predicted classifications in this file and submit both your paper and the prediction file. I have the actual values of the shot\_made\_flag for these missing shot\_ids and will evaluate the classifications. Your goal is to provide the best predictions possible.

Each group is on the honor system to not use any information outside of the dataset to predict each of the missing shot flags.

**DATA CONTINUED**

The field names are given below (Data descriptions are available in Kaggle):

|  |  |
| --- | --- |
| recId  action\_type  combined\_shot\_type  game\_event\_id  game\_id  lat – court location identifier (latitude)  loc\_x - court location identifier (x/y axis)  loc\_y- court location identifier (x / y axis)  lon - court location identifier (longitude)  minutes\_remaining – (in period)  period  playoffs  season  seconds\_remaining  attendance | avgnoisedb – avg noise in arena (decibels)  shot\_distance  shot\_made\_flag (this is what you are predicting)  shot\_type  shot\_zone\_area  shot\_zone\_basic  shot\_zone\_range  team\_id  team\_name  game\_date  matchup  opponent  shot\_id  arena\_temp (oF) |

**DELIVERABLE:**

Students will submit a paper with an 8 page limit with a separate Appendix up to 5 pages. Code should be in a second appendix and can be as long as necessary. A separate file with predicted classifications also should be submitted.

**PAPER REQUIREMENTS**

Introduction

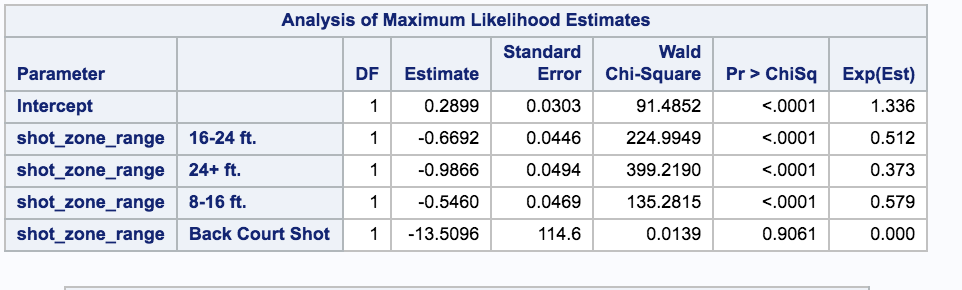
Data Description

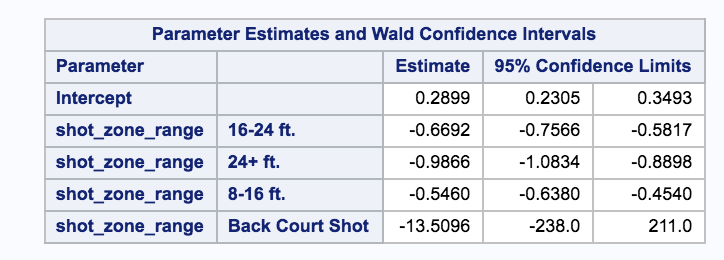
Exploratory Data Analysis

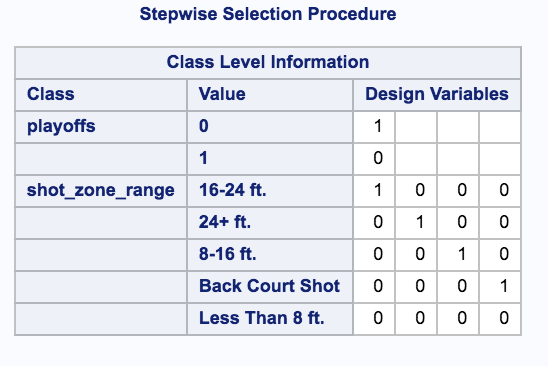
* Address the need for any potential transformations
* Address and identify outliers
* Address and identify any multicollinearity

Build models to provide arguments and evidence for or against the propositions below:

* The odds of Kobe making a shot decrease with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)







Shot distance is not used here to model since shot distance is normally distributed and due to which the goodness of fit test (Hosmer and Lemeshow test) indicates that model is not a good fit.

W (successful shot) = eb0 + b1 (8-16)ft + b2 (16-24)ft + b3(24+ft)

Odds between “less than 8 ft” and “8-16 ft” is calculated as below

W(successful shot 8-16ft)  / W (successful shot less than 8 ft) = e(0.2899 – 0.5546) / e(0.2899) = 0.57

The confidence interval of the odds of successful shot in zone range 8-16ft to 8 ft is 0.52 to 0.63. This interval does not contain zero as a value so null hypothesis which states that odds of successful shots at range 8-16ft is same as odds of successful shot in range less than 8ft “ is rejected (p < 0.0001).

From the above model,

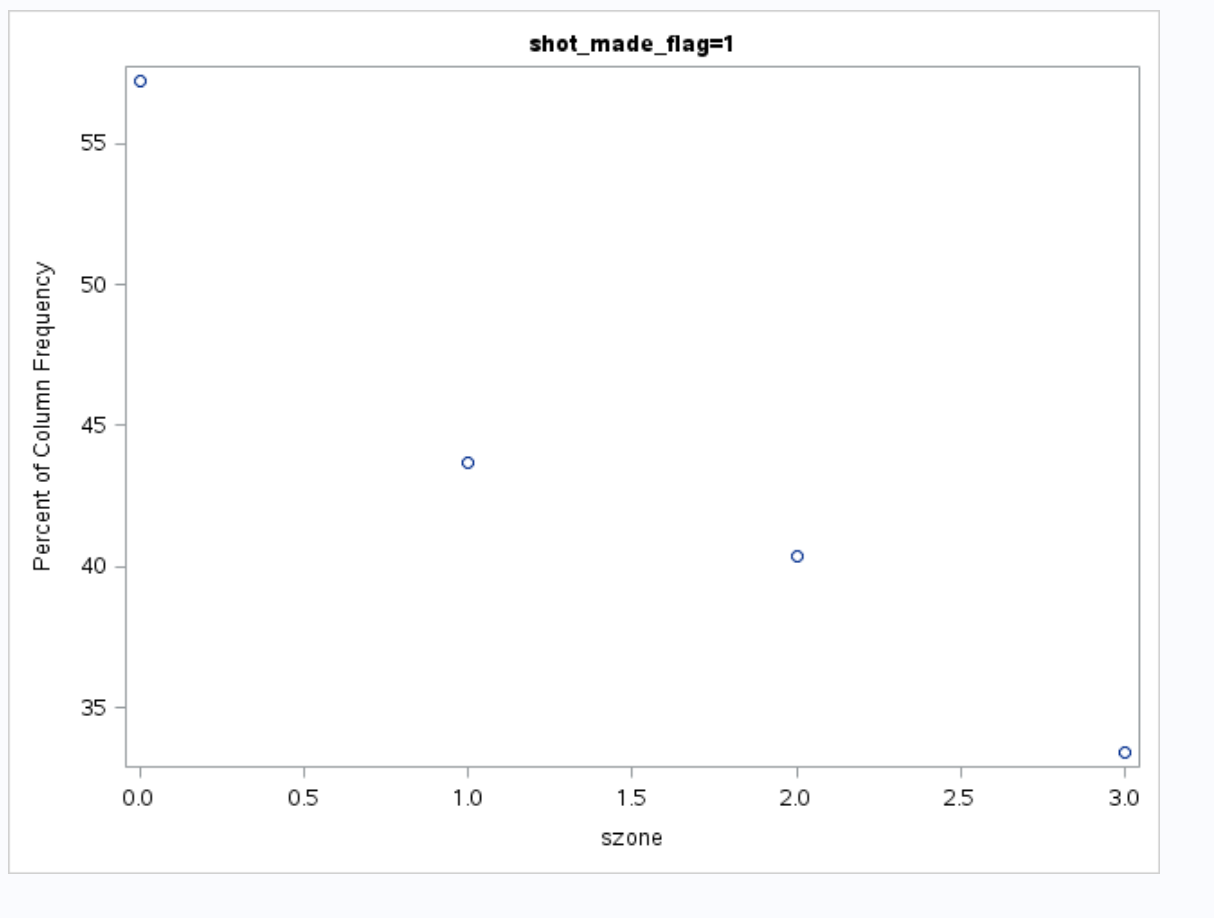
1. The odds of kobe making a successful shot if the distance from the hoop is in the range of 8-16ft is estimated to be 0.57 times that of hoop being with in 8ft range (p < 0.0001)
2. The odds of kobe making a successful shot if the distance from the hoop is in the range of 16-24ft is estimated to be 0.54 times that of hoop being with in 8ft range (p < 0.0001)
3. The odds of kobe making a successful shot if the distance from the hoop is in the range of 24+ft is estimated to be 0.37 times that of hoop being with in 8ft range (p < 0.0001)

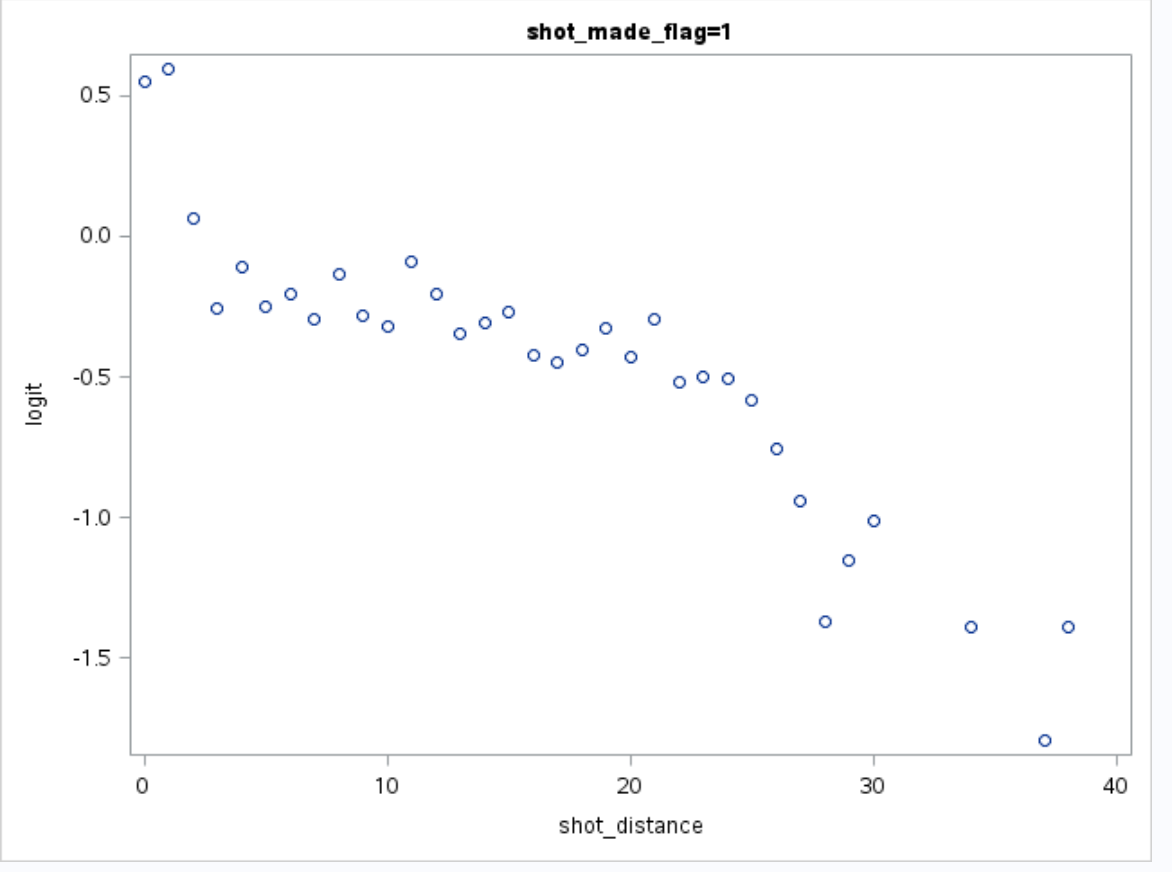
* The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)

The shot made variation w.r.t shot\_zone\_range coded values as below.

From the below plot we can see that proportions of successful shot made goes down linearly.

|  |
| --- |
|  |





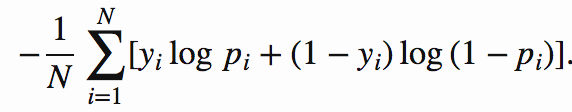
* The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs. Quantify your findings with statistical evidence one way or the other. (Tests, CIs, plots, etc.)

**FORMAT OF THE PAPER - CONTINUED**

Build a predictive model to classify shots as missed or made. You should produce at least 1 of each type of model:

* A logistic regression model.
* A Linear Discriminant Analysis (LDA) model.

Evaluation: Compare each competing models with the AUC, Mis-Classification Rate, Sensitivity, Specificity and objective / loss function. The log loss function of the model should be used to assess the model fit:



Where N is the total number classifications, yi is the shot\_made\_flag and pi is the probability from the model of each outcome for shot made.

**ASSESSMENT / EVALUATION:**

Good papers traditionally have the following characteristics:

1. They are presented in an organized, neat and consistent fashion. (Labeled plots, figures and tables, consistently formatted, indented and labeled headers and sub headers, etc.) Given that each group has 3 members, the paper should only have one look and feel. Titles, headers, sub headers, figures, tables, etc. should all look the same and have numbering that is consistent.
2. There are no typos, misspelled words, grammatical mistakes, etc.
3. They use a variety of methods.
4. Creative methods are used.
5. They have input from all group members and are developed iteratively over time as opposed to all at once such as the night before.

The group with the lowest log loss score will be awarded an additional 3 points for the project.

**SOFTWARE AND METHODS:**

You may use any software and must use only the methods we have studied thus far in the course. That being said, you can use innovative techniques inside of those methods like model averaging, cross validation or creating new variables from the ones in the data set. If you have any questions about this please let me know and we can discuss your ideas.