|  |  |
| --- | --- |
|  | **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 Kobe.xlsx.

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 project2Pred.xlsx file. These are the test set shots for which you must submit a classification. You are provided a sample classification file, project2Pred.xlsx with the shot\_ids needed for your predicted classification. Provide you 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):

|  |  |
| --- | --- |
| 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

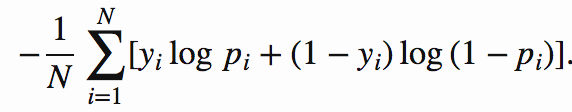
Build models to provide arguments and evidence for or against the propositions below:

* 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.)
* 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 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

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 (shot made or shot missed.)

##### Misclassification Rate for Classification Trees

Misclassification (Misc) comes from the number of incorrectly predicted observations. It is defined as



##### Residual Sum of Squares for Classification Trees

The residual sum of squares (RSS) for classification trees is defined as



where

*  is the actual response level
*  is the number of observations on the leaf  that have the response level 
*  is the posterior probability for the response level  on the leaf 
*  is the posterior probability for the actual response level  on the leaf 