|  |  |
| --- | --- |
|  | **PROJECT 2 SPRING 2018:**  **KOBE BRYANT SHOT SELECTION !!!** |

**OVERVIEW:**

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. Drafted into the NBA at the age of 17, Kobe earned the sport’s highest accolades throughout his long career.

Using 20 years of data on Kobe's swishes and misses, can you predict which shots will find the bottom of the net?

**DATA:**

This data contains the location and circumstances of every field goal attempted by Kobe Bryant took during his 20-year career. Your task is to predict whether the basket went in (shot\_made\_flag = 1 for made shot and 0 for missed).

500 of the shot\_made\_flags have been removed from the data set, represented as missing values in the csv file. These are the test set shots for which you must submit a classification. You are provided a sample classification file with the shot\_ids needed for a valid classification. Simply replace the current classification in this file with your predictions and submit both your paper and the prediction file (csv format) to Canvas. I have the actual values of the shot\_made\_flag for these shot\_ids and will evaluate each team’s classifications based on the log loss objective function. Your goal is to ***minimize*** the log loss which is defined further below.

**“Leakage”** is the term that is used to describe the phenomenon when future events “leak” into the past to corrupt a predictive exercise. Since these shots have already occurred and are part of the public record, someone could simply find a record of each particular game and find out if Kobe made a particular shot, or at least if they won the game or what Kobe’s stats were in that match. *Each group is on the honor system to not use any information outside of the dataset to predict each of the missing shot flags.*

The field names are self explanatory and contain the following attributes:

|  |  |
| --- | --- |
| * 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) |

**DELIVERABLES:**

1 Paper (8 page limit with up to 5 page Appendix … 13 page max. (code should be in second appendix and can be as long as it needs to be. ) In addition you should submit your classifications via Canvas as well.

**Format of the Paper:**

Intro

Data Description

Exploratory Data Analysis

\*Address the need for any potential transformations

\*Address and identify outliers

\*Address and identify any multicollinearity

Interpretation Models / Questions

Build models to provide arguments and evidence for or against each proposition:

1. 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.)
2. 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.)
3. The relationship between the distance Kobe is from the hoop 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.)
4. **BONUS (up to 2pts)** Kobe’s shooting percentage is subject to a home field advantage. That is, Kobe’s shooting percentage is better or worse at home than when he is away.
5. **BONUS (up to 2 pts)** Is Kobe “clutch”? Kobe is often the coach’s choice to take the final shot of the period or game. What do we know about his shooting percentage in these moments? Better than average? Less than average? About the same? Of course, provide statistical evidence to support your findings.

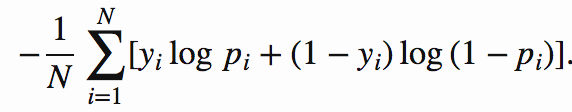
Predictive Model

Build a model to classify shots as missed or made.

You should produce at least 1 of each type of model:

1. A logistic regression model.
2. An LDA model.

Evaluation: Compare each competing model with the AUC, Mis-Classification Rate, Sensitivity, Specificity and objective / loss function. The loss function the model will be assessed on for the classification competition is the log loss:



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.)

**ASSESSMENT / EVALUATION:**

The grading scale will be based relative to the quality of the projects that come in. This means that the best paper will “set the curve” and all other papers will be judged as their “distance” from that paper.

I will say that good papers traditionally:

1. Are presented in an organized, neat and consistent fashion. (Labeled plots, figures and tables, consistently formatted, indented and labeled headers and sub headers, etc.) Note: given that your 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. Have no typos, misspelled words, grammatical mistakes, etc.
3. Use a variety of methods.
4. Usually have a creative use of a method or methods.
5. Have been contributed to by all group members and over time so that iterative changes can be made (as opposed to all at once … example: the night before.)

Don’t let this less structured grading scheme make you too nervous. If your group puts strong and consistent work into this project, you will do fine (B+ or better). Also, it is mathematically possible that everyone could get an A … it’s not that kind of “curve”. It all depends on the “distance” between the papers.

**SOFTWARE AND METHODS:**

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

**BONUSES:**

1. An extra 3 points will be given to the team that minimizes the objective / loss function.
2. (2 pts) Model Kobe’s shooting percentage over time. Does he appear to get better over time? Use your knowledge of the methods we have studied so far to answer this question the best way possible. (Again, you may have 1 additional page to answer this question.)
3. There are two more Bonus questions in the interpretation section above.