**New York Yankees Statistical Analysis Questionnaire**

Please answer the following questions. Please try to be concise in your answers:

1. Suppose you are asked to build a model that predicts salaries for arbitration-eligible players going through the arbitration process. You are given a list of possibly-relevant variables. Walk through your steps for this research. How would you test your model?
2. What questions in baseball do you want to investigate?
3. Describe a project you worked on (baseball or non-baseball, academic or personal, solo or in a group) that demonstrates the skills you would bring to the Yankees. Feel free to attach a pre-existing write-up or work sample if applicable.
4. Assume these fictional players are of equal stature: everyday players, playing the same position and possessing like handedness, having identical defensive skills, with similar age and experience in the big leagues, as well as comparable contract statuses. Which player (A, B, or C) would you argue had the best season and why? Do you think this player will provide the most offensive value moving forward or would you prefer one of the other players and why?

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Player** | **G** | **PA** | **AB** | **R** | **H** | **2B** | **3B** | **HR** | **RBI** | **BB** | **SO** | **SB** | **CS** | **AVG** |
| **A** | 151 | 640 | 581 | 99 | 159 | 29 | 0 | 42 | 127 | 59 | 139 | 2 | 0 | .274 |
| **B** | 143 | 602 | 485 | 108 | 140 | 37 | 3 | 21 | 87 | 117 | 151 | 17 | 5 | .289 |
| **C** | 159 | 677 | 597 | 87 | 160 | 31 | 0 | 30 | 102 | 80 | 76 | 28 | 2 | .268 |

1. The accompanying CSV file, matchupdata.csv, contains the results of 10,000 total at-bats.  There are three columns:

- "batterID": identity of the batter (1-100)

-"pitcherID" identity of the pitcher (101-200)

-"result": value of the result of the at-bat on a continuous scale

Next season, each batter in the dataset (1-100) will get exactly 300 at-bats. Each individual at-bat will be against a randomly selected pitcher from all the pitchers in the dataset (101-200), not just those that the given batter faced in the dataset.

Project each batter’s average result next season. Along with your projections, please provide a written explanation of your methodology and attach all code that you used to arrive at your answer.

1. The accompanying CSV files ‘pitchclassificationtrain.csv’ and ‘pitchclassificationtest.csv’ contain information from over 20,000 pitches from six different pitchers over three years. There are 12 columns:

- "pitchid": a unique identifier for each pitch.

- "pitcherid": identity of the pitcher (1-6). The identities are the same in both datasets. Pitcher 3 in the training set is the same pitcher as Pitcher 3 in the test set.

- "yearid": year in which the pitch occurred (1-3).

- "height": height in inches of the pitcher.

- "initspeed": initial speed of the pitch as it leaves the pitcher's hand, reported in MPH

- "breakx": horizontal distance in inches between where a pitch crossed the plate and where a hypothetical spinless pitch would have, where negative is inside to a right-handed hitter.

- "breakz": vertical distance in inches between where a pitch crossed the plate and where a hypothetical spinless pitch would have, where negative is closer to the ground.

- "initposx": horizontal position of the release point of the pitch. The position is measured in feet from the center of the rubber when the pitch is released, where negative is towards the third-base side of the rubber.

- "initposz": vertical position of the release point of the pitch. The position is measured in feet above the ground.

- "extension": distance in feet in front of the pitching rubber from which the pitcher releases the ball.

- "spinrate": how fast the ball is spinning as it leaves the pitcher's hand, reported in RPM

-"type": type of pitch that was thrown (will only appear in the training dataset).

Your goal is to give the most likely pitch type for all of the pitches in the test dataset (year 3) using information from the training dataset (years 1-2). Note that the pitchers in the datasets do not correspond with any specific real pitchers but are meant to be representative of real data. Please include the following with your final submission:

1. CSV with two columns: the pitchid and the corresponding predicted pitch type
2. write-up of your method and results, including any tables or figures that help communicate your findings
3. all code used to solve the problem