STAT 847: Analysis Assignment 1

DUE: Thursday, February 2 2023 by 11:59pm EST

NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark. Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible. Furthermore, if you submit your assignment to Crowdmark, but you do so incorrectly in any way (e.g., you upload your Question 2 solution in the Question 1 box), you will receive a 5% deduction (i.e., 5% of the assignment's point total will be deducted from your point total).

This assignment is all about cleaning (wrangling) text data of a custom structure, please show all your code and comment it accordingly. There are a total of 50 points possible. Look at the end of the Week 02 notes for a start on this assignment.

1. [7 points] Create a data frame or a tibble of tags/metadata portion of the games in chess_classic_games.pgn. One row should represent one game, and one column should represent one tag.

Tags that appear in some games, but not a given game, should be left as NA in any game that doesn't have them.

For example, the first five rows should look like classic_first_five.csv

Show the next five lines and the skim() of the dataset.

2. [3 points] Add two columns to the left end of the data set (hint: rbind() can do this, and so can select()). These added first two columns should have the first line in the file chess_classic_games.pgn which includes a tag and the moves for the given game, respectively. For example, the first five values of tag_line should be 1, 21, 41, 61, 81, and the first five values of moves_line should be 19, 39, 59, 79, 99.

3. [3 points] Find the quantile(. . . , probs=c(0.01,0.05,0.25,0.5,0.75,0.90,0.99,0.999,1)) of all player Elos. Make sure to include both black and white players.

4. [4 points] Create a linear model of proportion of time the first player (white) wins as a function of the amount of higher rating the first player has over the second player. Use this model to determine how many rating points playing first is worth (i.e., how many rating points worse does the player using white have to be in order to win exactly half of the time).

Show the summary() of the lm()

5. [3 points] Repeat the linear model from part 4, but filter() so that only games between players where both players have a 2000 rating or better. Does the pattern hold.

Show the summary() of the lm()

*Note: A logistic regression would be better for each of these, we will return to this in class or in a later assignment.

6. [6 points] Add the first move to the cleaned dataset. Also add an indicator variable that is 1 if the first move is either e4, and 0 otherwise. The moves_line variable from part 2 will be useful. Get two separate table()s, one of each of the new variables.

The first move of the first five games is d4, f4, d4, e4, and d4, respectively.

7. [4 points] Plot the proportion games where WhiteElo is in (-infty, 850), [850, 900), [900, 950)... starts the game by playing e4 as broken line graph (Hint: plot(..., type='b'). You may use ggplot instead, but it will be worth no extra points.

8. [6 points] A question mark on a move indicates that it is a blunder. A double question mark indicates it is an extreme blunder. Make a multivariate model of the number of question marks that appear as a function of the average of WhiteElo and BlackElo. (Hint: str_count())

Show the summary() of the lm()

9. [4 points] A time control of 300 + 1, for example, indicates that each player started the game with 5 minutes (300 seconds) to play all their moves in the match, and that each move played added (i.e., incremented) 1 second to their clock. Add log(starting time) (Hint: + I(log(x))) in seconds to your model from question 7.

Show the summary() of the lm() comment on the differences between this model and the question 7 model.

10. [5 points] Chess 960 is a variant of chess where the starting positions of some of the pieces are randomized before the match. Repeat steps 1 and 2 for the chess_960_database. Show the skim() of the resulting dataset.

11. [5 points] Repeat questions 6 and 7 for the chess_960_database. (Get opening move, and plot the proportion starting e4). Comment on the differences in the popularity of e4 as an opening move. (You do not need to demonstrate any chess knowledge to make this comparison. You can if you wish, but you are not being marked on chess knowledge)