OVERVIEW

You will be given a dataset ("soccermatches.csv") with data from club soccer matches. The fields are listed on the next page.

Your assignment is to write three different functions (as described below) to predict the results of soccer matches from an out-of-sample dataset of the same type (same variable names, also club soccer matches from similar period and similar years). The functions be written in either R or python. They will load the **out-of-sample** CSV and add the predictions to the dataframe. Do not use anything auxiliary like a model object, the functions must do the work to generate predictions.

You are expected to code your functions based on analysis of soccermatches.csv (example: you may fit a model on soccermatches.csv and hard code those coefficients into your functions). **The file soccermatches2.csv is the out-of-sample data you are predicting and is not to be used to fit a model.**

FUNCTION 1: Make pregame predictions of P(Home Team wins), P(Draw) and P(Away Team Wins). Note that these variables are Hwins, IsDraw and Awins in the dataset. Your goal is to minimize overall Brier Score on the out-of-sample CSV. The winner of the game can be determined by Hgoals and Agoals. Because these are pregame projections, you may not use Hgoals1H or Agoals1H as part of your prediction. Your predictions should be added to the dataframe as hProb, dProb and aProb.

FUNCTION 2: Make halftime predictions of P(Home Team wins), P(Draw) and P(Away Team Wins). Your goal is to minimize overall Brier Score on the out-of-sample CSV. In this function, you may of course use Hgoals1H or Agoals1H. Your predictions should be added to the dataframe as hProbHT, dProbHT and aProbHT.

FUNCTION 3: Make a pregame prediction of the total number of goals scored in the game. Note that this variable is "goals2H" in the dataset. Your goal is to minimize sum of squared errors in the out-of-sample CSV. Your prediction should be added to the dataframe as predGoals2H.

You should assume that the out-of-sample CSV will be titled "soccermatches2.csv". Your script will

1. Load soccermatches2.csv
2. Call the three functions and add predictions to dataframe
3. Write csv as [yourname]_output.csv

ADDITIONAL NOTES

1. Your submission will not be considered if you attempt to determine exact identity of games based on the rows in sample. (Example of something that is not in the spirit of the exercise: having a lookup table across vast database of soccer leagues to find results of games where Home Team had previously scored 33 goals, allowed 28, etc.)
2. If you feel that you can get the best possible results via some method that absolutely requires that you attach and load a model object, you can submit that as a supplement but you should still include a "standard submission" that accomplishes everything in self-contained functions.
3. The "in-sample" dataset "soccermatches.csv" exists for you to delve into and learn from. But your submission will not be able to ingest this file. That is, the function must do the work itself (as opposed to loading a CSV, building a model, predicting with that model).
4. Sample submissions are attached, one in R, one in python.

FINAL REMINDER: Before submitting please double check that we will be able to run your script (.r or .py file) without any additional auxiliary files (such as the original soccermatches.csv file or any model object files). Additionally, your file should not be expecting to load anything other than the soccermatches2.csv file.

VARIABLES

Hwins: Binary variable indicating whether home team won the match
Awins: Binary variable indicating whether away team won the match
IsDraw: Binary variable indicating whether match ended in a draw
goals2H: Number of goals scored in second half

[Above four variables are in principle redundant because they can be dervited from the rest of dataset but are added for avoidance of doubt since these are the variables you are assigned to predict]

Hgoals: Goals scored by home team
Agoals: Goals scored by away team
Hgoals1H: Goals scored by home team in the first half
Agoals1H: Goals scored by away team in the first half
HG_LY: Home team number of goals scored in previous season (“last year”)
HGA_LY: Home team goals allowed in previous season
HM_LY: Home team matches played previous season
AG_LY: Away team number of goals scored in previous season (“last year”)
AGA_LY: Away team goals allowed in previous season
AM_LY: Away team matches played previous season
HG_TY: Home team number of goals scored in current season (“this year”)
HGA_TY: Home team goals allowed in current season
HM_TY: Home team matches played current season
AG_TY: Away team number of goals scored in current season (“this year”)
AGA_TY: Away team goals allowed in current season
AM_TY: Away team matches played current season
Hrelegated: 0/1 indicator for whether home team was relegated in previous year
Arelegated: 0/1 indicator for whether away team was relegated in previous year
Hpromoted: 0/1 indicator for whether home team was promoted in previous year
Apromoted: 0/1 indicator for whether away team was promoted in previous year