Predictive Modeling with Sports Data

Homework 1

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting relevant code you want to include in your PDF. For Jupyter notebooks, you can save them as LaTeX or PDF before including them. Make sure your answers to each problem are clearly stated in the submitted PDF. The graders should not have to look through your code to find the solution.

Any code (Jupyter Notebooks and other source files) used to compute your answers should also be uploaded to Gradescope, along with a `readme` file telling the graders how to run your code if they need to, and if you are using any special packges. If the problem requires you to train and test a model, both the training and testing code should be submitted. Please indicate in your `readme` file if running your code requires special hardware (like a GPU, or 64GB+ of ram), a compilation step (if you decide to use Cython), or a signficant amount of time (longer than 5 minutes).

All of the questions in this homework will use the data in `soccer17.csv`. In the column names, `FT` and `HT` mean Full Time (result at the end of the game) and Half Time (result at the end of the first half), respectively. Also `HG` and `AG` mean Home Goals and Away Goals, respectively. The `Y` column denotes the year of the season, and may not match the year in the `Date` column.

1. (Goal Differentials) The goal differential for a team in a game is the number of goals they scored minus the number of goals the opponent scored. In this problem we will analyze the goal differentials of various teams in the given dataset. [Hint: If you are using LaTeX you can typeset the tables using `df.to_latex(filename)` and the `booktabs` package in LaTeX.]

   (a) Give a table containing each team in the English Premier League (EPL). The teams should be ranked in descending order by their average goal differentials per game in the 2017 season (`Y=17`). Each row of the table should include the team's

name, their average goal differential that year, and the number of games they played that year. It should also contain the number of wins, losses, and draws they had that year.

(b) Create another table with the same data as the previous, but with an added column `Points`, the total number of points (different from goals) scored by each team that year. A team scores 3 points for each win, 1 point for each draw, and 0 points for each loss. This new table should be ranked in descending order by points.

(c) Take the top 3 teams (in terms of average goal differential in 2017) from each league (as determined by the `Div` column) and put them together in a single table. This should contain the same data as in the previous part, along with the additional column `Div`. Sort the table in increasing order by `Div` and, within each league, in decreasing order by average goal differential.

2. (Goal Scoring) In this problem we will use all years in the given dataset.

(a) Consider the total number of goals scored in each game (sum of home goals and away goals).

i. Model the distribution of total goals scored using a parameteric distribution. State your choice of distribution, and the corresponding parameters.

ii. Plot a histogram of the total number of goals scored. Overlay your expected number of goals for each histogram bucket.

(b) Repeat question (a) for the total number of goals scored in the first half of each game.

(c) Repeat question (a) for the total number of goals scored in the second half of each game. [Note: All goals are scored in either the first half or the second.]

(d) Rank the leagues by average goals per game. Include the number of games in each league, and the average goals per game for each league.

(e) Considering games that have exactly 4 goals scored, determine whether there is evidence of "comeback tendency." You should use the approach presented in class that analyzes the number of drawn games. The columns `pH`, `pD`, `pA` contain market-based forecasts for the probabilities of a home win, a draw, and an away win, respectively.