INFSCI 1091: Winning in Sports with Data

Homework 2

1. Home field advantage is a well-established phenomenon in sports, even though its roots are still widely unknown. However, different sports might exhibit different levels of home field advantage. Collecting the appropriate data from sports-reference.com rank the 4 major sports (NFL, NBA, NHL, MLB) in the US:
    a. Rank the leagues in terms of the extend of home field advantage they exhibit. (20 points)
    b. Examine whether there have been any significant changes in the home field advantage over the past 10 years. (20 points)

    Explain your answers and approach in both cases.
    Hint: One way to attack this problem is by using the rating methods we discussed in class.

2. Fivethirtyeight.com has opened its prediction data for everyone. Their NFL probabilistic predictions can be found here:
    https://projects.fivethirtyeight.com/nfl-api/nfl_elo.csv
    Calculate the Brier score, the accuracy (assuming a win probability threshold of 0.5) and the reliability curve for the predictions for the NFL seasons between 2010-2017. For the Brier score decompose it to its 3 parts (30 points).
    Note: In order to get the full points you need to write your own code for calculating the above metrics, i.e., do not rely on an already existing function in any programming language, which you should submit as part of your solution.

3. On courseweb I have provided you with the play-by-play data for all the games in two NBA seasons. Using these data build a simple in-game win probability model using logistic regression. Your model should take as input the current *state* of the game, and provide a win probability for the home team. The state is essentially the independent variables that you will choose for your model. Provide a description of the variables that you chose, the

trained model as well as its evaluation. Part of the question is to identify the right evaluation metric for this type of model. (30 points)