MDML Final Project Analysis Plan

Diana, Frank, Sue

2019/10/31

Dataset Description: NBA stats and injury records

In this project, we plan to scrape data from basketball reference websites and pro sports transactions.

The outcome of NBA stats is win or loss of the match and indicators include whether home or away, number of audience, distance from home city, field goal %, three point %, rebounds, assists and other team stats in the match. There will be at least 2 season stats scaped.

The injury records will contain team, date, player injured, injury type. The way of organizing data is stated in Part C.

Research Question

We are interested in the observed difference of winning probability between home matches and away matches. Therefore, our research questions are:

- Is there difference in winning probability between home matches and away matches?
- If there is, what are the underlying factors influencing team performance?

Also, we want to develop a model to predict injury of players. The corresponding research goals are:

- Predict risk of injury
- Identify impactful factors increasing risk of injury

Analysis Plan

Part A: Scrape data for 2 seasons, including match stats and injury records

A1: Scrape data to summarise match results and number of audience attended

Scrape data from *https://www.basketball-reference.com/leagues/NBA_2018_games-october.html*

A2: Scrape stats for all matches, including both basic and advanced, for both visitor team and home team

Scrape data from *https://www.basketball-reference.com/boxscores/201910220LAC.html* . For each match, there are four tables.

A3: Scrape injury records for all the 30 teams

Scrape data from *https://www.prosportstransactions.com/basketball/Search/Search.php* or use data scraped by others *https://www.kaggle.com/ghopkins/nba-injuries-2010-2018/version/1#*

Part B: Home and away match difference in winning probability and reasons

B1: Join data from A1 and A2

By transforming data from A2 to team stats and joining it with data from A1, create a dataset with relevent variables.

B2: Model fitting and performance evaluation

Here we plan to fit two models, for home matches and away matches respectively. Specifically, in the model for home matches, we will include all the home matches for all the NBA teams, using win or loss as the outcome and team stats for the home team as well as number of audience as the independent indicators.

$$Pr(Y_H = 1|\ ..) = logit^{-1}(\beta_{H1}X_1 + ... + \beta_{Hk}X_k + \beta_{HN}NumAudience + \epsilon_H),\ \epsilon_H \sim N(0, \sigma_H^2),\ indep.$$

While in the model for away matches, we will include all the away matches for all the NBA teams, using win or loss as the outcome and team stats for the away team as well as distance from home city, number of audience as the independent indicators.

$$Pr(Y_A = 1|\ ..) = logit^{-1}(\beta_{A1}X_1 + ... + \beta_{Ak}X_k + \beta_{AN}NumAudience + \beta_{AD}Distance + \epsilon_A),\ \epsilon_A \sim N(0, \sigma_A^2)\ indep.$$

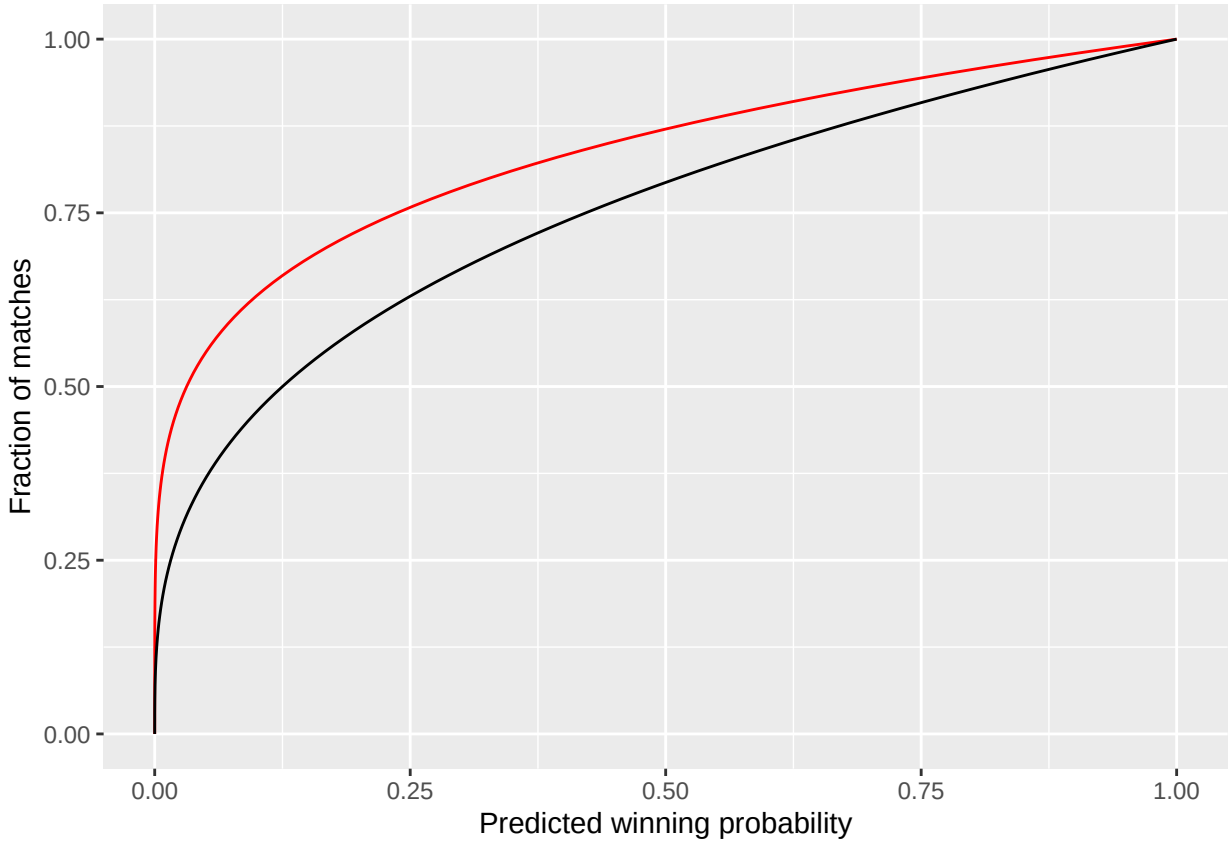
Benefits of developing two models for home and away matches respecitively instead of using home or away as an dummay variable:

- The result of a match for two teams are paired, one wins, then the other loses, which will violate the assumptions of independent epsilons in regression. So with two models, we can meet the assumption that both models have independent errors.
- Using home or away as dummy variable means model only allows difference in intercepts, but difference in slopes is also likely to exist. Using two models, coefficients are more flexible.

Then we will split data to training and testing datasets, fit models using logistic regression as stated above and calculate accuracy, precision, recall and AUC to measure the performance of models.

B3: Difference in winning probabilities between home and away matches respectively

We will plot fraction of matches against predicted winning probability for home and away respectively, with two lines in the same plot.



We expect a plot like this figure. Based on our observation and experience, there should be a difference

between two lines and the red line should refer to away matches and the black line refers to home matches. But of course, there is likely to be no difference or lines refer to reverse groups. This plot will answer the first research question: Is there difference in winning probability between home matches and away matches? For the same level of fraction of matches, black line on the right indicates higher winning probabiity.

B4: Impact of distance and number of audience

If winning probability is higher in home matches, look at the coefficients of models to see whether the distance and number of audiance can impact the winning probability.

B5: Further reasons: difference in team stats

Knowing that win or loss in away matches may be influenced by distance or number of audiance in the stadium, what is the furthur factors that are caused by distance or number of audience and meanwhile, directly cause the difference in winning probability?

We plan to use t tests to test are there any sinificant difference in team stats between home and away matches such as field goal %, three point %, rebounds, assists. Then we can find insights from those results to know is there any aspects that players cannot perform well in away matches compared to home matches.

B6: Further reasons: difference in slopes of team stats

We will further analysis the underlying reasons if winning probability is higher in home matches. Another possible explanation is that the impact of some team stats on the outcome gets changed when players play away. We will plot the predicted probability of winning against each team stat to see if there is any difference in slopes. The slopes suppose to be the same, but difference is likely to exist, showing that some team stats get more important or less important in away matches.

Part C: Injury prediction for individual players

C1: Join data from A2 and A3 and transform dataset

Join data from A2 with data from A3 by players, making each row represent a player in a specific match, with variable indicating minutes on court, points etc., and whether got injury or not.

For each player each match, develop variables as follows:

- Mean minutes on court in corresponding season before this match, excluding absence
- Mean of points and other stats before this match
- Total matches attended before this match
- Minutes on court in last five matches respectively
- Mean of points and other stats in last five matches respectively
- Home or away for last five matches respectively
- Home or away for this match
- Ranking of the opponent for this match
- Whether got injury or not

C2: Develop logistic regression model for prediction

Use the dataset created above to develop a logistic regression model. Split data into training and testing dataset and use training data to train the model. Then use testing data to test performance of the model.

C3: Identify impactful factors increasing risk of injury

Identify influencing factors by considering p-value and standardized coefficients.

Possible Questions

1. It is likely that performance of models are not good enough for us to do further analysis.
2. Distance and number of audience contribute to team stats, is it reasonable for us to include all of them in the same model? It seems that they are in two different layers, distance and number of audience cause other team stats. Should we include interaction?
3. Should we control number of covariates to increase interpretability? Or should we include as many covariates as we can to increase accuracy? Is this definitely going to increase accuracy?
4. There may be missing data because the process of web scraping will be pretty complicated since we need to scape from multiple websites and join them together. How can we deal with missing data? Will simply dropping missing data influence our predicted probability comparison?