**Research Scenario**

Chess is a well-known game fancied by plenty of people all over the world. People not only like playing chess but also like watching chess competitions to learn from each other. During the watching, someone may notice that they could simply predict the winner by only look at the information about the players and games. The winner maybe decided by turns, players and opening, etc. Someone on Kaggle.com kept recording chess games with basic information. For this project, I will explore whether the winner is related to features of games.

**Dataset description**

The dataset records over 20000 chess games from a selection of users on the site Lishess.org. The original dataset has 16 features. However, some of features are not related to the project, I only keep some of the features and remove the rest of them. Id is the id of the game; turns means both players have made their moves. The only exception to this rule is when White plays a move that ends the game. After black moves and White’s clock starts again, a turn is complete; victory\_status means the result of the game including out of time, resign, mate, draw; winner includes white, black or draw; white\_rating means score of white players; black\_rating means score of black players; opening\_ply means the group of initial moves of a chess game. For this dataset, I will explore what features (or a feature) decide the winner of a chess game.

Text

Description automatically generated

**Graphical Summary**

For preprocess the data, I deleted all rows with NA values, check the type of data, plot distributions of numeric and non-numeric features, check the outliers by boxplot. For the original dataset, opening\_ply and turns were a little bit right skewed, the white\_rating and black\_rating were like normal distribution. There are many outliers in all four numeric features. I remove the outliers form those features, now all features are like normal distribution with few outliers:

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

The non-numeric features were unbalanced, there are few samples for draw and out of time of victory\_status, and few samples for draw for winner. So, I delete the draw and out of time rows. The result distribution become:

Shape, square

Description automatically generated

The dataset has 16257 rows after the preprocess.

**Research Questions:**

How do features (or feature) increase the possibility of winning by white chess players?

Are features related to winners?

**Techniques to answer the question:**

For answering this question, I will use multiple logistic regression. Firstly, I dichotomize the winner viable. I created a new variable named white in which white = 1 if white chess player wins the game and white = 0 if black chess player wins the game. Secondly, I will do the overall test to make sure there is at least one feature relates to the winner. Finally, if there is at least one feature exists relates to winner, I will perform individual simple logistic regression to find out which one is related to winner.

**Conclusion:**

By the overall test, there is at least one feature influence the event that white play win the game. (P < 0.05)

By individual test:

There is an association between turns and the possibility of event that white win the game after adjusting other features at alpha = 0.05. The odds ratio for white win is 0.997 for every 1-unit increase in turns, which means for every 1-unit increase in turns, there will be 0.3% lower likelihood the white win the game. We are 95% confident that the true odds ratio is between 0.996 and 0.998.

There is an association between white\_rating and the possibility of event that white win the game after adjusting other features at alpha = 0.05. The odds ratio for white win is 1.004 for every 1-unit increase in white\_rating, which means for every 1-unit increase in white\_rating, there will be 0.1% higher likelihood the white win the game. We are 95% confident that the true odds ratio is between 1.003868 and 1.004263.

There is an association between black\_rating and the possibility of event that white win the game after adjusting other features at alpha = 0.05. The odds ratio for white win is 0.996 for every 1-unit increase in black\_rating, which means for every 1-unit increase in black\_rating, there will be 0.4% lower likelihood the white win the game. We are 95% confident that the true odds ratio is between 0.9956567and 0.9960470.

There is an association between opening\_ply and the possibility of event that white win the game after adjusting other features at alpha = 0.05. The odds ratio for white win is 1.025 for every 1-unit increase in opening\_ply, which means for every 1-unit increase in opening\_ply, there will be 2.5% higher likelihood the white win the game. We are 95% confident that the true odds ratio is between 1.008892 and 1.040774.

AUC is 0.7261 in this case which indicates a fine fit.

A picture containing text, group, line, bunch

Description automatically generated