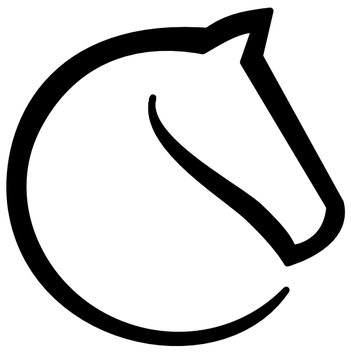
LiChess Project

Part *Data Processing* -- 2024/2025



Introduction

LiChess is a website dedicated to online chess. It is the second most visited site in the world with more than 3 million games played per day. It has the particularity to be based on free software, and to publish the games played on database.lichess.org. The games are in PGN format.

Data

You can start from the raw data, but there are multiple initiatives to extract LiChess data with specific focuses. One such extraction only keeps games that have been analyzed by Stockfish, a chess engine (see below), and reports several metrics of the engine's analysis in the dataset. The dataset is avaible here:

> https://www.kaggle.com/noobiedatascientist/lichess-september-2020-data

The CSV file (you can remove the .RDS) contains 3,739,909 games played in September 2020. The file contains 40 columns (most columns are doubled, once for white, once for black). We have in columns some interesting information, in particular:

- The ratings (number of ELO points) of both players. You can find on LiChess players of any level, from beginners to International Grandmasters.
- The type of game according to the time control: Blitz (between 2 and 10min), Fast (10-15min), Classic.
- The opening played. The openings, i.e. the beginning of the game, have been classified over time, and codes called ECO have been assigned to the main openings.
- LiChess also allows to request a (free) post-game analysis by Stockfish, a chess engine which annotates the game and detects major errors (*blunders*), errors (*mistakes*) or imprecise moves (*inacurracies*) of the players. The interest of this dataset is that it contains only games analyzed and annotated by Stockfish.

Other columns provide fine-grained information.

- The number of errors when the player is under time pressure (*ts: timescamble*)
- The number of moves that took a long time (>10% of the initial time) (*long_move*)
- The number of errors in time consuming moves
- The number of times the positional advantage (according to Stockfish's evaluation) has shifted from one player to the other (*game_flip*)

Work to be done

The work you are assigned in this project is to ask questions and answer them with quantitative analyses.

You are requested to answer the following questions, and you are encouraged to come up with another. Your analysis will be valued according to the relevance of the question.

Mandatory Questions:

- Q1: What is the rate of blunders, errors and inaccuracies per move, per level category (*) and on Blitz type games (Blitz type is by far the most played on these online sites). A game has two players, whose ELOs are most likely different. You will be able to classify a game into a category, either by considering the average ELO of both players, or by considering only the games where both players are in the same category.
- Q2: Win probability depending on opening:
  - Q2a: With which opening does White have the best chance to win, by level category (*) and by type of game (Blitz, Fast, Classic).
  - Q2b: same question with black. You don't need to write again the same but only the results with black.
- Q3: (difficult). Does a line of data in the file predict the outcome of the game (column *Result*), and with what probability? In other words, can any of the variables, such as the number of errors (mistakes, blunders, inacurracies, ts_blunders), the difference in ELO between the two players, etc., *explain* the outcome (win/loss)? You are free to define *explain* as you wish. It can be a correlation, linear or not, or any other relationship that allows this prediction.

  Note that the ELO is itself computed from a probability (normal distribution) of victory depending on the difference in ELO of the two players. For instance, for a difference of 100 ELO points, the higher ranked player is expected to win with probability 0.64. For a 200 points difference, it is 0.76.

  As we have more data than the ELO difference, your prediction should be more accurate than that.

Instructions

You will turn in a report with your analysis in the form of the questions you have about the dataset, and the answers you provide substantiated by quantitative results. Structure your answer in 3 points:

1. An introductory text that explains the main idea of the calculation, the assumptions, and defines the variables used if necessary.
2. Commented code that implements the idea.
3. The analysis of the obtained results, and in particular the statement of whether the results corroborate the initial hypotheses.

You will attach the programs that allowed you to calculate your quantitative results.

You can use Hadoop or Spark, and in the latter case, the notebook is ideal for presenting the results.

Appendix

(*) We can consider the following levels:

ELO	desc
[1200-1499]	occasional player
[1500-1799]	good club player
[1800-1999]	very good club player
[2000-2399]	national and international level (IM)
[2400-2800]	GMI, World Champions