A Needle in a Data Haystack Introduction to Data Science - 67978 - Project Milestone

The Project title: Player Profiling and Strategy Recommendation System

Team members info (name, email, TZ, CS id):

- Daniel Gibor (Bogatyrevich), daniel.gibor@mail.huji.ac.il, 334065281, danielgibor
- Amir Kelman, amir.kelman@mail.huji.ac.il, 208288811, Amir.kelman
- Itay Boiangou, itay.boiangou@mail.huji.ac.il, 322274697, Itay.boiangou
- Avihu Almog, avihu.almog@mail.huji.ac.il, 315709980, avihuxp

The Problem Description:

The system that we aim to create is a **recommended strategy system for chess players**. In the first stage, the system will profile the chess players based on the DB on their game record. Profiles could feature attributes such as favorite opening, history of games with other players, average game duration, and specific weaknesses like highest losing ratio to certain openings.The second stage will feature the recommendation system. By receiving profiles of 2 players - **we intend to provide a recommended strategy for both players**. Strategies could include recommendations such as making rapid moves in order to pressure the time stamp - make the opponent "run the clock", selecting the best openings to play against him, play aggressively, etc. We would like the recommendation analysis to be able to take into account the number and nature of previous matches between players and to be able to differentiate between the games played by friends and other players.

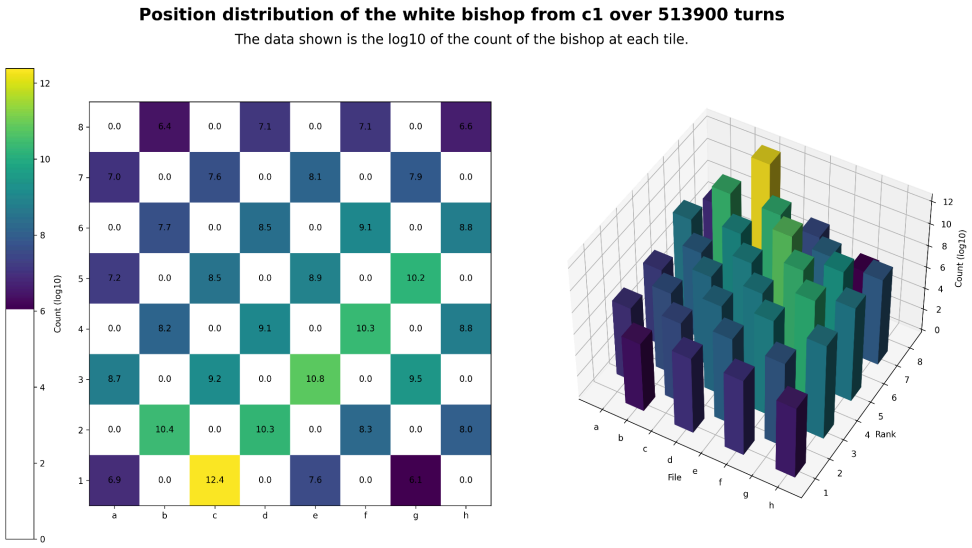
The Project Data:

The data published on the lichess.org open database website will be used for the project. We will analyze 10,680,708 played chess games on the website lichess.org, for January 2017 (around 9.5 GB). The file format is for each game - the name of the event it was played in, URL of the game on Lichess, the outcome of the game, handles of the players, date and time of the start of the game, ratings of the players before the game, changes in the player's rating as a result of the game, which classifies the opening used, details on the opening used in the game, the time control for the game, how the game ended and indicator of the evaluation of the position at specific moves (%eval). Each game also includes all the moves made in the game, encoded in PGN format.

Visualizations:

1. Bishop Positions:

In chess, bishops consistently remain on the same-colored tiles throughout games. We further hypothesized a bishop's position would be uniformly distributed across these squares over many games. Our analysis and visualization confirmed bishops stay on same-colored squares but revealed a non-uniform distribution of their positions. Bishops tend to remain near their initial positions, with movement patterns influenced by diagonal distance from the start and proximity to the board's center. This can be seen in the added figure (or even better - **a GIF** of the convergence!), in which we can see that 1) bishops stay on the same-colored tiles and 2) the distribution of visits is not uniform. Note that the mean and STD of rating difference between black and white are 0.62 and 199.63 points (respectively, as can be seen here), so it does not matter which bishop we chose, since 95%~ of the games had a rating difference of no more than 400 points, and 99.7%~ with no more than 600 points.



2. Winning rate vs player rating:

This line plot shows the relationship between player ratings and winning rates. As expected, the higher rated players tend to have higher winning rates. This is consistent with the assumption that player skill correlates with their success in matches. The plot shows that the average winning rate passes the 0.5 (red line) for all ratings above approximately 1300. And the green bars show the number of players in each rating bin. This visualization effectively confirms the hypothesis that higher rated players win more often, validating the reliability of the rating system.