CSE 519 - Data Science Fundamentals

PROJECT PROPOSAL

How Good is a Chess Player?

Objective:

To build a model that will predict the ELO score of a chess player. Additionally, we are motivated to guess the type of game depending on the time limit of the game and to predict the quality of the play looking at the final board.

Background Research:

A chess rating system is a system used in chess to calculate an estimate of the strength of the player, based on his or her performance versus other players. They are used by organizations such as FIDE, the US Chess Federation (USCF or US Chess), International Correspondence Chess Federation, and the English Chess Federation.

The ELO rating system is a method for calculating the relative skill levels of players in zero-sum games such as chess. It was developed by the Hungarian physicist Arpad Elo in the 1950's and adopted by the world chess federation (FIDE) in 1970. Since its development, the system has been adopted with various modifications by many national chess federations. Today it is impossible to imagine tournament chess without a rating system.

The Elo rating system calculates for every player a numerical rating based on performances in competitive chess. A rating is a number normally between 0 and 3000 that changes over time depending on the outcomes of tournament games. When two players compete, the rating system predicts that one with the higher rating is expected to win more often. The more marked the difference in ratings, greater the probability that the higher rated player will win.

In this project we are concerned with predicting the chess rating score (ELO) of both players in a game, given a record of the match. With the benefit of powerful computers and large game databases, we can easily investigate approaches that might help in predicting chess results.

Similar work has been done in the past to predict the chess game result :

[1] has proposed a technique which uses the existing database of chess games and machine learning algorithms to predict the game results. Authors also developed various relationships among different combinations of attributes like half-moves, move sequence, chess engine evaluated score, opening sequence and the game result. The idea was to make the Multi-Variate Linear Regression algorithm learn from these evaluation scores for the same sequence of opening moves and game outcome, then using it to calculate the winning score of a side for each possible move and thus suggesting the move with the highest score.

[2] trains World Chess Federation (FIDE) rating systems using a training dataset of a recent eleven-year period with games from 2000 chess players. The system was then used to predict the outcome of chess games played by

the same players in the following half year. Accuracy between predicted results and actual game results is the primary indicator of whether the approach was a practical chess rating system.

Dataset:

Computer usage among chess players has become quite common in recent years and a variety of different programs, both commercial and public domain, are used to generate, access, and propagate chess game data.

Following datasets and sources have been identified for use:

A. **Lichess Game Database (Data from 2013)** [3]

Possible Features:

- Event
- Site
- Date
- Round
- White
- Black
- WhiteTitle
- BlackTitle
- Result
- BlackElo
- BlackRatingDiff
- ECO
- Opening
- Termination
- TimeControl
- UTCDate
- UTCTime
- WhiteElo
- WhiteRatingDiff
- Moves

B. **FICS Games Database (Data from 2008)** [4] (Similar Features as in Lichess)

C. **Kaggle Problem Statement - Chess Game Dataset (Lichess)** [5]

Possible Features:

- Game ID
- Rated (T/F)
- Start Time
- End Time
- Number of Turns
- Game Status (Out of time, Resign, Mate, Draw)
- Winner (White / Black)
- Time Increment
- All Moves in Standard Chess Notation
- Opening Eco (Standardised Code for any given opening)
- Opening Name
- Opening Ply (Number of moves in the opening phase)
- White Player ID
- White Player Rating
- Black Player ID
- Black Player Ratin

Format of Dataset in FICS and Lichess Game Database: Portable Game Notation (PGN)

PGN is a plain text in computer-processable format for recording chess games (both the moves and related data), supported by many chess programs. PGN is structured "for easy reading and writing by human users and for easy parsing and generation by computer programs." The chess moves themselves are given in algebraic chess notation. [6]

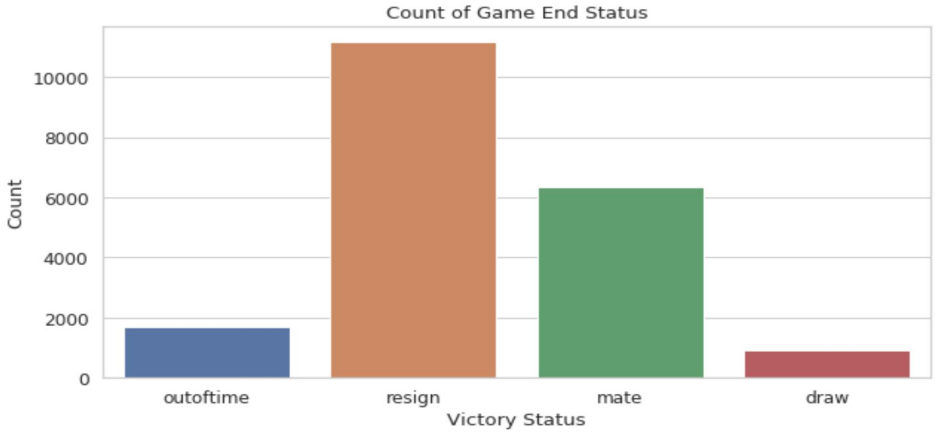
```
[Event "F/S Return Match"]
[Site "Belgrade, Serbia JUG"]
[Date "1992.11.04"]
[Round "29"]
[White "Fischer, Robert J."]
[Black "Spassky, Boris V."]
[Result "1/2-1/2"]

1. e4 e5 2. Nf3 Nc6 3. Bb5 a6 4. Ba4 Nf6 5. O-O Be7 6. Re1 b5 7. Bb3 d6 8. c3
O-O 9. h3 Nb8 10. d4 Nbd7 11. c4 c6 12. cxb5 axb5 13. Nc3 Bb7 14. Bg5 b4 15.
Nb1 h6 16. Bh4 c5 17. dxe5 Nxe4 18. Bxe7 Qxe7 19. exd6 Qf6 20. Nbd2 Nxd6 21.
Nc4 Nxc4 22. Bxc4 Nb6 23. Ne5 Rae8 24. Bxf7+ Rxf7 25. Nxf7 Rxe1+ 26. Qxe1 Kxf7
27. Qe3 Qg5 28. Qxg5 hxg5 29. b3 Ke6 30. a3 Kd6 31. axb4 cxb4 32. Ra5 Nd5 33.
f3 Bc8 34. Kf2 Bf5 35. Ra7 g6 36. Ra6+ Kc5 37. Ke1 Nf4 38. g3 Nxh3 39. Kd2 Kb5
40. Rd6 Kc5 41. Ra6 Nf2 42. g4 Bd3 43. Re6 1/2-1/2
```

Sample PGN Game Data with Minimum Headers

Exploratory Analysis:

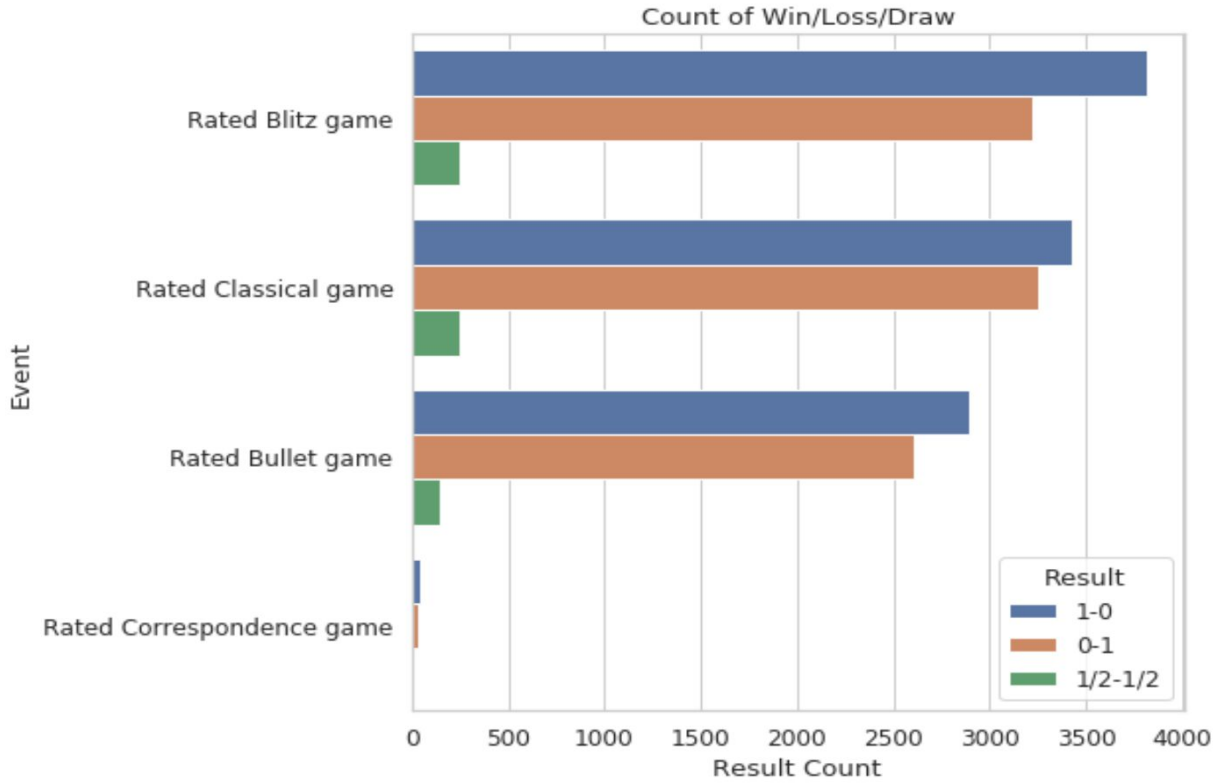
1. Game End Status

There are 4 ways, how a chess game can end - outoftime, resign, check-mate and draw between the two players. If the player feels that they are losing too much material or the upcoming checkmate is inevitable is the moment when player resigns. According to our dataset, most number of chess games end due to resigning.



2. Count of Wins / Losses / Draw

The below countplot shows the frequency distribution of win/loss/draw against the events of the game.
Blue Bar - White wins against Black
Orange Bar - White loses against Black
Green Bar - The match is a Draw

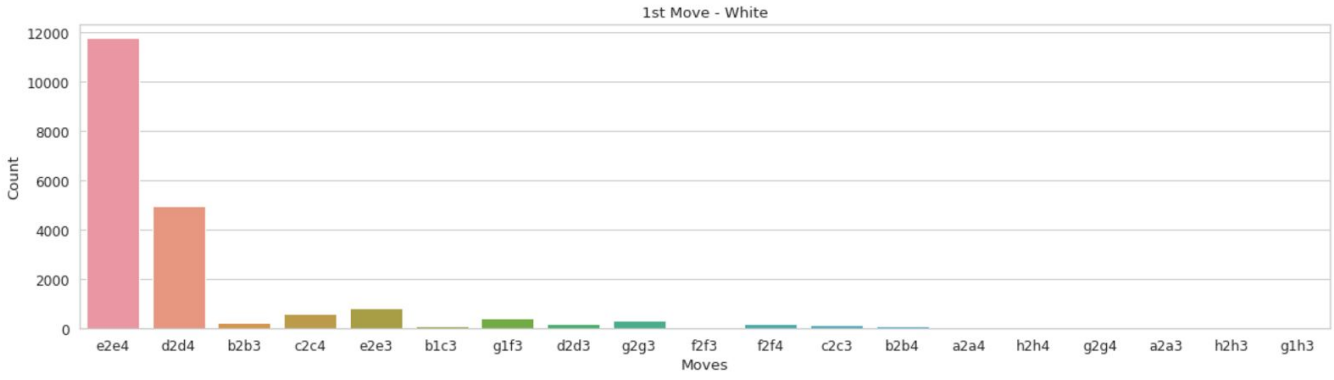


3. Moves

The first few moves of a chess game, known as the chess opening, are one of the most-studied aspects of the game, largely because of how important they can be. Given below is the analysis given in [7]:

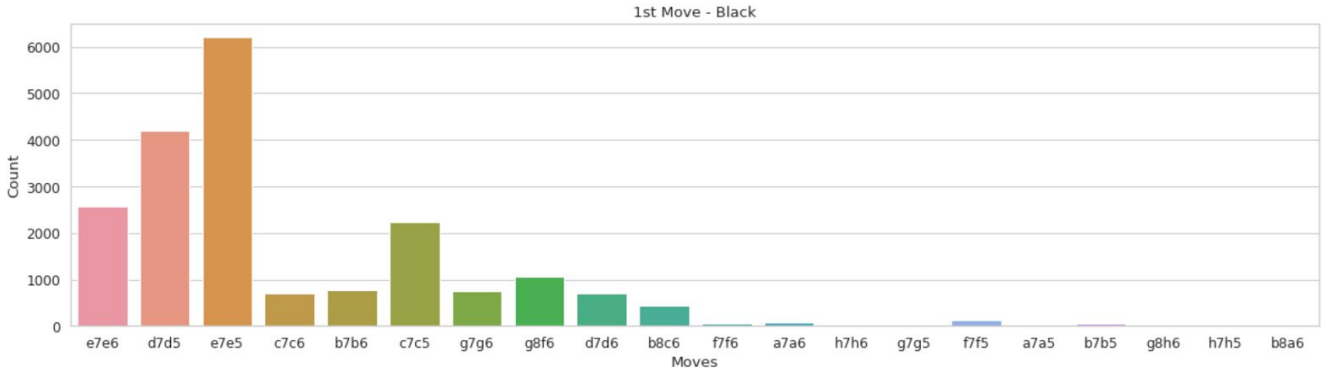
a. First Move - White

White has a small advantage at the beginning of the game. To maintain this advantage, White should press their advantage to take over the middle of the board as quickly as possible.



b. First Move - Black

Many of Black's opening moves are more defensive in nature and attempt to undermine White's initial advantage.



Challenges:

- ***Learning to compare positions***
  The evaluation function is the most important component of a chess program. It receives a chess position as an input, and provides a score as an output. This score represents how good the given position is (typically from White's perspective). [8]

- ***Predict Type of Game from transcript***
  Types of Games:
  - Classic: Chess is a game played by two people on a chessboard, with sixteen pieces (of six types) for each player. Each type of piece moves in a distinct way. The goal of the game is to checkmate, i.e. to threaten the opponent's king with inevitable capture. In addition, there are several ways that a game can end in a draw.
  - Rapid: Each player is given less time to consider their moves than normal tournament time controls allow. The rules specify a cumulative total time for moves for each side.
  - Blitz: Players have three to five minutes to make all of their moves.

  Based on the observed game, we need to predict the time limit of the game.

- ***Predict the ELO rating of the players from game transcript***
  Practical Uses:
  - For pairing purposes in tournaments, to avoid pairing candidates who are most likely to win a tournament during the earlier rounds of the tournament.
  - For tournament sectioning and prize eligibility. (Eg. Invitation to compete in the U.S. closed championships and to compete on the U.S. Olympiad team are based in part on players' USCF ratings). [9]

- ***Sub Transcript-level Analysis***
  Possible Analysis:
  - To tell the quality of play only from the final board position.
  - Predict the level of play from the time taken to play.

- ***Identify if player is a Human or a Computer?***

Humans identify potential long term positional ideas whereas Computers are better at calculating complex positions. Most of the games would end up in draws but in some situations (specially endgames where you need depth as well as complex strategic planning), human and computer partnership would triumph.
A famous example would be the Caruana vs Nakamura game where computers never identified the queen sacrifice.

- ***Identify the games where the weaker player wins by beating the stronger ones?***
  There exists some rare occasions where stronger players lose to weaker ones. Identifying such occasions after certain moves have been played is another task to focus on.

Approach:

- The dataset is being converted into a comma-separated values file from a portable game notation file for analysis. A single file contains millions of lines of text, this is too large and impractical to load into python at one time, so we must process the file in batches and then join these batches together.
- We plan to know the minimum number of moves required in a match to predict the rating of a player.
- We plan to generate different models for different types of training data and test them against a common test dataset, with this we can understand which dataset is better for prediction.
- We are going to make a model using Convolutional Neural Network for the prediction of rating of players based on the sequence of the board.

Future Work:

- Analyzing the data to understand the sequence of moves in chess and how it impacts the game.
- Cleaning and preprocessing of data. Here we are required to solve the anomalies, assign IDs etc.
- Build a prediction model using Convolutional Neural Networks and Long short-term memory architecture and compare the results with the ELO scoring system.
- Further analyze the data to guess the type of the game, the quality of the play from the final position of the board and also identify if the player is a human or a computer.

References:

[1] Paras Lehana, Sudhanshu Kulshrestha, Nitin Thakur and Pradeep Asthana - Statistical Analysis on Result Prediction in Chess [https://search.proquest.com/openview/5108034a9fa8212970332620bc1ace84/1?cbl=2026670&pq-origsite=gscholar]

[2] Zheyuan Fan, Yuming Kuang, Xiaolin Lin, Stanford University - Chess Game Result Prediction System [http://cs229.stanford.edu/proj2013/FanKuangLin-ChessGameResultPredictionSystem.pdf]

[3] Lichess Game Database [https://database.lichess.org/]

[4] FICS Games Database [https://www.ficsgames.org/download.html]

[5] Kaggle Problem Statement - Chess Game Dataset (Lichess) [https://www.kaggle.com/datasnaek/chess]

[6] Standard: Portable Game Notation Specification and Implementation Guide

[7] Randy Olsen - A data-driven exploration of the evolution of chess: Popularity of openings over time [https://spartanideas.msu.edu/2014/05/27/a-data-driven-exploration-of-the-evolution-of-chess-popularity-of-openings-over-time/]

[8] Omid E. David, Nathan S. Netanyahu, and Lior Wolf - DeepChess: End-to-End Deep Neural Network for Automatic Learning in Chess [https://www.cs.tau.ac.il/~wolf/papers/deepchess.pdf]

[9] Mark E. Glickman, Albyn C. Jones - Rating the Chess Rating System [http://glicko.net/research/chance.pdf]