Assignment 1

MIT 805

Connor McDonald
u16040725

Big Data

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science
University of Pretoria
21 September 2021

1 Overview

The dataset being used for this assignment is comprised of 6.25 million online chess games recorded over a month and published online by lichess.org. To understand the dataset, there are a few things the reader should familiarise themselves with, which are as follows: A chess players skill level is judged on a rating system developed by Arpad Emmerich Elo called the "ELO rating", it is used to measure the relative skill levels of players in zero-sum games such as chess. Secondly there are three ways in which a game of chess can terminate namely, a checkmate in which the player wins, a stalemate in which the game is declared a draw and lastly a time forfeit in which the player wins when the opponent runs out of time. Lastly, there are many different ways in which the time of a game is controlled, which also leads to different types of games. The most common time control method is when each player is given a certain timeframe, when it is their move their clock will begin running, when it is their opponent's move their clock will stop, if a players clock finishes, they lose the game. Another method is similar to the first method, however, whenever a player makes a move they gain 5 seconds to their clock, this encourages fast moves and more aggressive games. Lastly, a game can have no time limit and only will end in stalemate or with one player winning.

The inspiration behind the collection of this dataset is detailed below:

- To analyse different play styles of players based on their ELO rating
- Identify opening trends based on player ELO ratings.
- Identify the most efficient openings.
- Identify the most common and most frequent moves.
- Analyse how play styles change based on the type and level of time constraint
- Ultimately using these findings to train new players

1.1 Expectations

It is expected to see a win/lose ratio slightly favouring white as white always has the first move in chess and is often considered to be at an advantage because of this. Furthermore, it is expected to see a lot of noise in low-rated players as these games are often a lot less methodical and thought out as the games between higher rated players. This could make it difficult to identify trends and may result in low rated players being filtered out of the dataset, provided it does not "downsample" too drastically. Lastly, I expect to see an uneven distribution of openings as some openings tend to be more popular than others. Due to their popularity I expect to see some sort of correlation between the frequency of an opening and the win-rate associated with that opening.

2 Technical Aspects

This dataset contains 6.25 million online chess matches recorded in the month of July 2016, and was stored in a 4.38GB *.csv* file. It contains 15 columns which are described below:

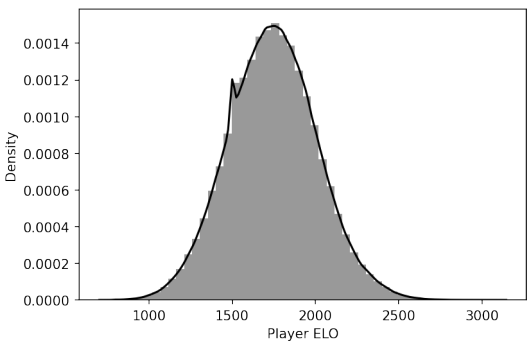
Table 1: Dataset Explanation

Column Name	**Description**	**Data Type**
Event	Game type	string
White	White's ID	string
Black	Black's ID	string
Result	Game Result (1-0 White wins) (0-1 Black wins)	string
UTCDate	UTC Date	date
UTCTime	UTC Time	time
WhiteElo	White's ELO	integer
BlackElo	Black's ELO	integer
WhiteRatingDiff	White's rating points difference after the game	integer
BlackRatingDiff	Blacks's rating points difference after the game	integer
ECO	Opening in ECO encoding	string
Opening	Opening name	string
TimeControl	Time of the game for each player in seconds. The number after the increment is the number of seconds before the player's clock starts ticking in each turn.	string
Termination	Reason of the game's end.	string
AN	Movements in Movetext format	string

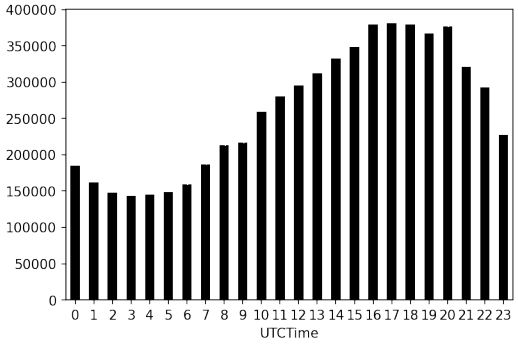
3 The V's of Big Data

3.1 Veracity

Veracity refers to the biases, noise and abnormality in data. Fortunately this dataset has no missing or invalid values and therfore from that point of view it can be considered a high quality dataset. However, this dataset deals with human behaviour in a game with an incomprehensible amount of outcomes. For these reasons it is expected that there will be a lot of noise especially amongst players with lower ELO ratings, as a large amount of their moves may be considered random and non-traditional. Fortunately the distributions of both white and black ELO ratings are normally distributed (Figure 1) with a mean around 1700 which is considered a highly skilled player, therefore if need be, a significantly large sample of the data can be taken which excludes low ranked players.



(a) ELO distribution



(b) Games Per Hour

Figure 1

3.2 Variety

The data contained in this dataset contains various datatypes sich as string, date, time and integers and was collected through platforms such as the lichess.org website as well as the lichess mobile app. However, due to the data being structured it has inherently less variety than an unstructured dataset, but it still provides multiple layers of value as it allows for different analyses to be conducted from different perspectives such as how game type affects openings or how ELO rating affects game type etc.

3.3 Volume

In terms of volume, this dataset contains 6.25 million rows and 15 columns for a total of 93.8 million entries. Before any adjustments or filtering, the dataset is 4.38GB in size. However, to reduce noise and provide high quality analyses, all games in which at least one player has an ELO rating less than 1500 will be filtered out. In doing so the dataset maintains a large amount of its original information with around 4.5 million rows and a size of 3.1GB. This is supported by Figure 1a which shows that majority of the data falls to the right of 1500.

3.4 Velocity

As stated in the overview, this data was collected over the month of July in 2016 with a continuous streaming process. By dividing the total number of rows by the number of seconds in 31 days we arrive at an average of about 2-3 games per second. However, when looking at the distributions, it was observed that the daily amount of games was fairly uniform with around 200 000 games per day, but the games per hour were not as consistent and showed a clear peak in the late afternoon to early evening seen in Figure 1b. The “off-period” observed in the early hours of the morning highlight the possibility for batch processing to be utilised, however the intervals between games are still frequent enough to make use of real-time analytics and stream processing should the need arise.