The Best Chess Opening

Scott Melli

Using Open Source data from Lichess.org of over 90 million chess games to determine the best chess opening

Introduction

In this report, we delve into the world of **chess openings** with the aim of identifying the best strategy for success. Our analysis is based on a comprehensive dataset comprising over **90 million chess games** and over **200 gigabytes**. The dataset provides us with valuable insights into the playing habits, skill levels, and opening choices of players across different Elo ratings.

This report sets the stage for a comprehensive exploration of the best chess opening strategy. By leveraging a vast dataset and conducting thorough data analysis, we have uncovered valuable insights into the popularity and success rates of different openings across various skill levels. While there is room for improvement in the machine learning models, the data serves as a valuable tool for chess players seeking to enhance their game and maximize their chances of success.

The analysis also unveiled some of the most popular and successful openings for white players across various Elo bins. Notably, the **Queen's Gambit Accepted** emerged as a consistently popular and effective opening choice for white. However, we observed that the **success rates of openings varied across different Elo bins**, emphasizing the importance of tailoring the opening strategy to the player's skill level.

Queen’s Gambit Accepted

About the Data

The data used for this analysis consists of a vast collection of chess games from **lichess.org**, totaling over **90 million individual game records**. Originally presented as a single, massive **200 GB file**, the data had to be divided into smaller, more manageable pieces to overcome hardware and software limitations. This fragmentation allowed for efficient processing and analysis.

Each game record in the dataset contains valuable information that contributes to the understanding of chess openings and their impact on game outcomes. The data was organized into a structured format, primarily stored as a PGN Text file. We transformed this data into a CSV file which facilitates easy access and manipulation of the data for further analysis.

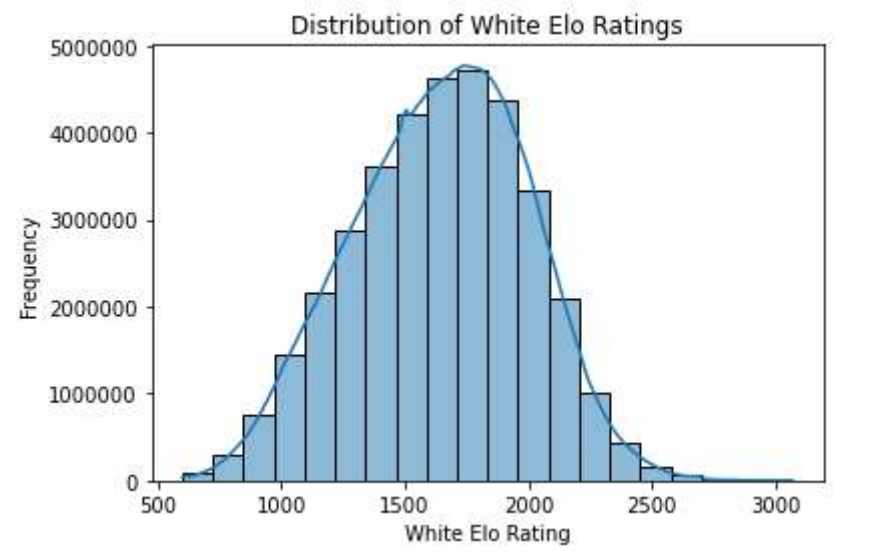
Key attributes extracted from the raw data include:

Game URLs: Unique identifiers or links to the individual chess games.
Events: Descriptions or names of the chess events in which the games took place.
White Elo Ratings: Elo ratings of the white players, representing their skill levels.
Black Elo Ratings: Elo ratings of the black players, indicating their relative skill levels.
Time Controls: Specifications for the time limitations or constraints imposed on the games.
Results: Outcomes of the games, such as win, loss, or draw.
Terminations: Indicators of how the games ended, such as checkmate or time forfeit.
Openings: The specific chess openings played at the beginning of each game.

The data represents a rich resource that allows for a comprehensive examination of chess openings and their impact on game outcomes. Its scale, variety, and organization provide a solid foundation for conducting meaningful analysis and drawing reliable conclusions.

Exploratory Data Analysis

The **exploratory data analysis (EDA)** conducted in this project provided valuable insights into the chess playing habits and opening strategies of players across different Elo ratings. Here is a summary of the key findings:



Elo Ratings: The majority of players in the dataset had an Elo rating around **1654**, suggesting that a significant portion of games were played within this skill level.

Popular Openings:

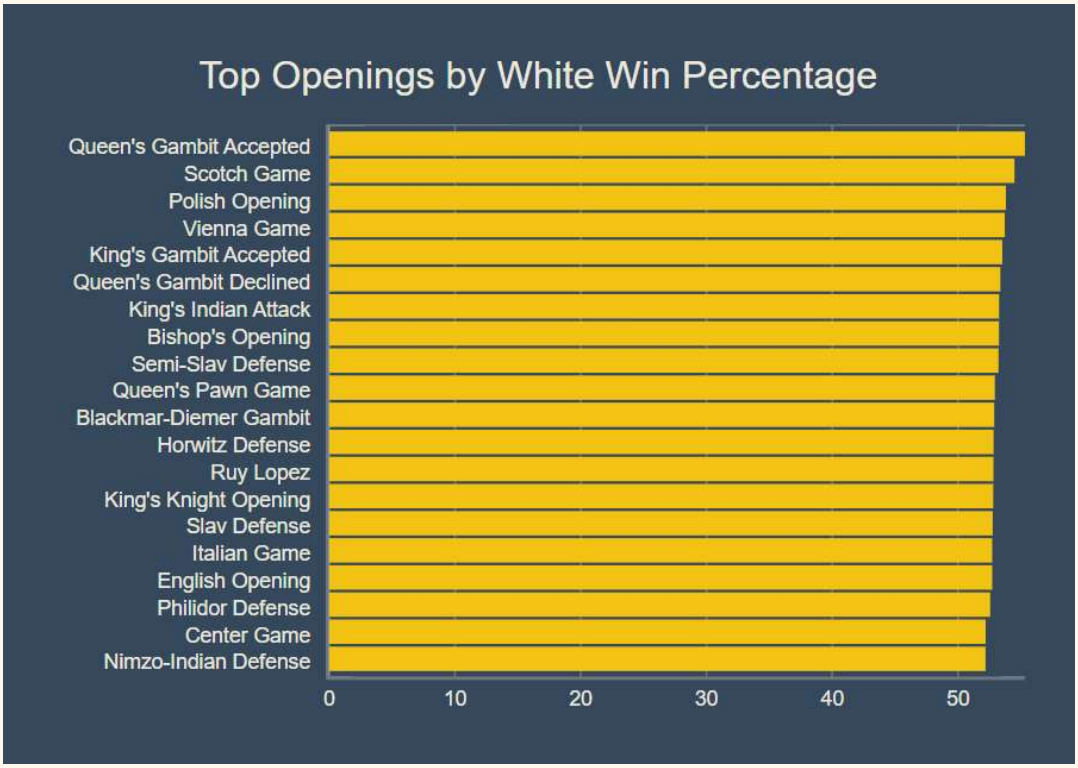
The Queen's Gambit Accepted emerged as one of the most popular and successful openings for white players across various Elo bins. However, the success rates of openings varied across different Elo bins, highlighting the importance of tailoring opening strategies to individual skill levels.

White Elo	Opening	Count	White Won
600-800	Benoni Defense	565	56.64%
800-1000	Queen's Gambit Accepted	11273	54.77%
1000-1200	Queen's Gambit Accepted	39198	55.77%
1200-1400	Queen's Gambit Accepted	67066	56.22%
1400-1600	Queen's Gambit Accepted	86178	55.82%
1600-1800	Queen's Gambit Accepted	84841	55.33%
1800-2000	Queen's Gambit Accepted	58930	54.66%
2000-2200	Center Game	12360	55.81%
2200-2400	Englund Gambit	2245	53.23%
2400-2600	Ponziani Opening	296	58.78%
2600-2800	Englund Gambit	42	82.14%
2800-3000	Hungarian Opening	6	100.00%
2800-3000	Ponziani Opening	1	100.00%

Success Rate Variation: The success rates of specific openings varied widely across different Elo bins. For example, the Benoni Defense opening had a success rate of 56.73% in the 600 - 800 Elo bin, while the Hungarian Opening achieved a perfect success rate of 100% in the 2800 - 3000 Elo bin.

Sample Size Limitations: It was observed that some of the popular openings in higher Elo bins had relatively small sample sizes. Consequently, the success rates for these openings might be skewed and may not reflect their true performance.

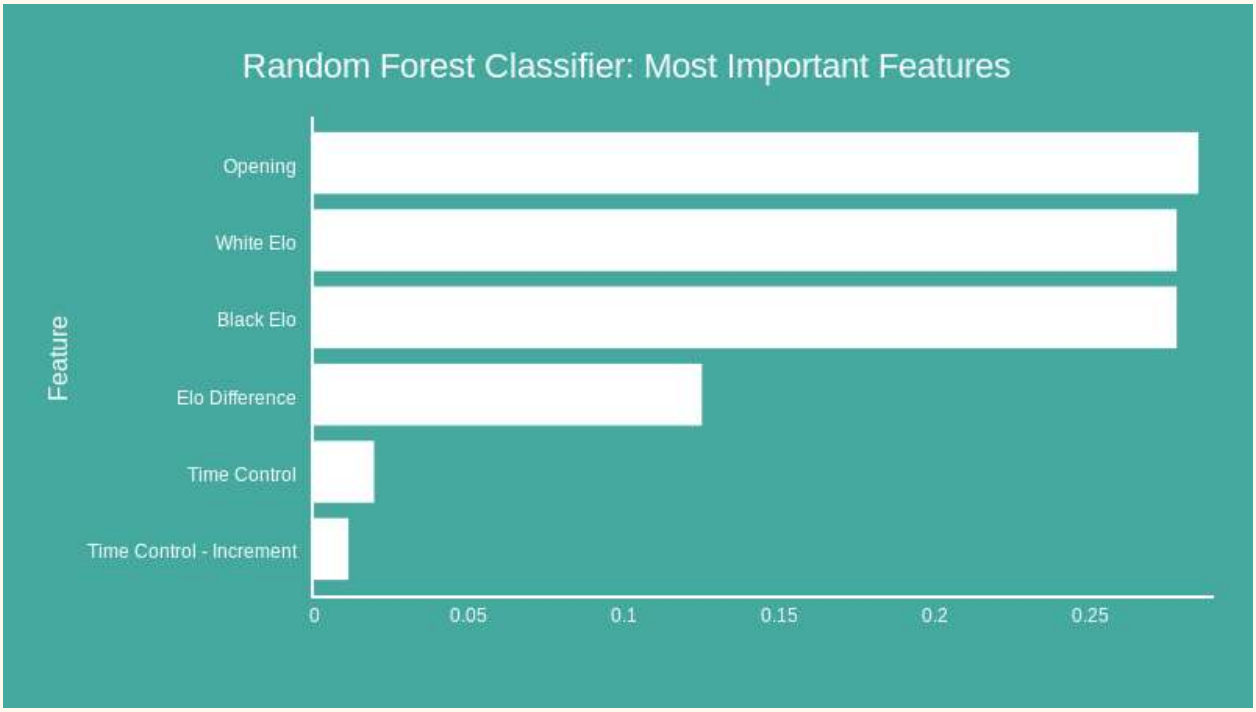
Overall, the EDA revealed important insights into the **popularity and success rates of different chess openings across various Elo ratings**. The findings emphasize the significance of considering one's own skill level and the performance characteristics of different openings when selecting an opening strategy in chess.

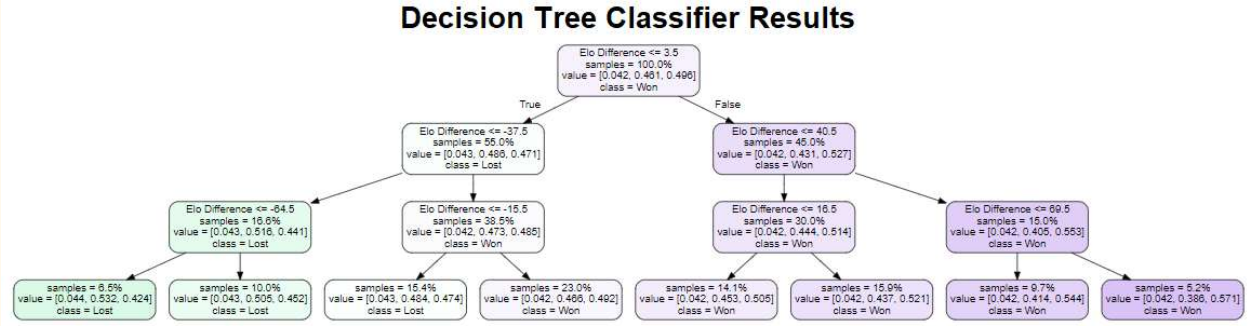


Machine Learning

Two models, a **decision tree** and a **random forest**, were trained and evaluated using a comprehensive dataset of over 90 million chess games. Along with **Ordinal Encoder**, these were the best choice because of the **categorical nature and abundance of the data**. The **decision tree model achieved a train accuracy of 0.5102 and a test accuracy of 0.5102**, indicating that it generalized reasonably well to unseen data. On the other hand, the **random forest model showed a higher train accuracy of 0.9531 but a lower test accuracy of 0.4812**, suggesting some level of overfitting and the need for further refinement.

The feature importance analysis revealed that the **opening choice** was the most influential predictor of game outcomes, followed by Elo ratings and Elo difference. This insight provides valuable guidance for chess players in strategizing their opening moves and improving their chances of success.





In conclusion, while the machine learning models showed potential, overfitting was identified as a challenge. Further refinement and adjustments are needed to enhance the models' performance on unseen data. By addressing overfitting and considering the influence of features such as opening choice, players can make more informed decisions and optimize their gameplay strategies in order to achieve favorable outcomes in chess games.