Analyzing Chess Blunders For Different Skill Levels

Elisabeth Fritze
Matriculation Number: 4183541
elisabeth.fritze@student.uni-tuebingen.de

Felix Pfeiffer
Matriculation Number: 4232052
f.pfeiffer@student.uni-tuebingen.de

Abstract

We analyzed 516 348 chess games played on lichess.org, to find patterns in the blunder behavior of players with different Elo ratings. We extracted distinct features from the gameplay to compare the different player groups. Further, we performed a correlation analysis on the features to find out how the gameplay lengths, blunder rates per piece, and blunder severity differ at different Elo ratings. We were able to confirm our assumption that weaker players blunder more often and with a higher degree of severity. The code can be found at https://github.com/LisbethIsolde/chess-blunders.

1 Introduction

Chess is known worldwide and has acquired a deep cultural significance. Millions of players engage in the game, trying to improve their skill level online. So, what differentiates a good chess player from a bad one? In this paper, we are investigating what role the blunder behavior of the player plays. To accomplish this, we scanned nearly three million games played on Lichess between January and February 2015, filtering for games analyzed by the Stockfish chess engine. This resulted in over half a million relevant games, that we are visualizing using different techniques like violin plots and heatmaps. We display both the blunders per piece and the blunder locations on the board for players in different Elo ranges. Additionally, we are trying to predict a player's Elo by using 16 handcrafted features, extracted from the gameplay, using linear regression. Relevant features were gameplay length, blunder severity, and piece usage.

2 Background

We gathered our data from lichess.org a free-to-play and open source online chess platform. All games played on the site are saved in PGN (Portable Game Notation) format and are free to download. The PGN files contain the gameplay in algebraic notation and metadata about the games, e.g., the timestamp of the game, players' Elo ratings, and the game mode. The Elo system [1] assigns an Elo rating number to every player that describes the relative skill level. The typical range is from around 600 to 2800. When users request a computer analysis of their game, the evaluation is added to the PGN. Lichess uses Stockfish, one of the most powerful publicly available chess engines, to represent the current position of the game as a numeric value. The value can be interpreted as material balance. A rating of 1.0 means that white is leading by an advantage of one pawn. Not only material but also the position of the pieces on the board is considered, since positioning is one of the most important characteristics of chess positions. Moves that worsen the evaluation significantly can be considered blunders. If the engine rating decreases by three points, it is equivalent to the loss of a minor piece (knight or bishop). Five points are equivalent to a rook, while the queen is worth nine points.

Project Report for *Data Literacy* 2021/22

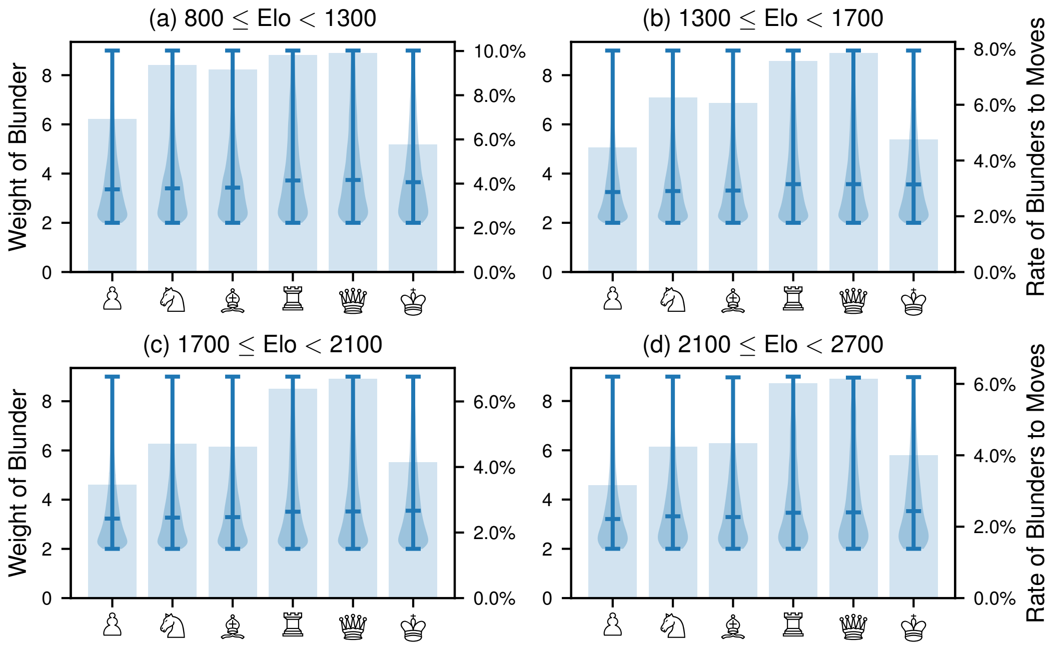

Figure 1: The violin plots show the distribution of blunder weights per piece. The bar chart in the background shows the relative blunder score per piece

3 Analysis

In order to get expressive data, we had to set limitations to the analyzed games. One requirement was that the gameplay had to last longer than 15 moves. We only looked at games where the deviation between the players was less than 100 Elo points, to exclude games where one player is significantly better than the other. We also ignored games where the player had an Elo below 800 or above 2600, due to the sparsity of data points in the extreme regions.

Our first point of interest was whether players at different skill levels blunder with different pieces. As seen in Figure 1 (a) players in the lowest Elo range have similar blunder-to-move ratios for all pieces, besides the king and the pawns. For higher rated ranges, the overall blunder ratio drops, while there is a slight change in the relationships between the different pieces. The knight and bishop are sometimes considered harder pieces to master, as they have less obvious moving ranges. This coincides with our finding that the blunder rates for the knight and bishop drop in Figure 1 (b), where the players have achieved a somewhat decent playing level.

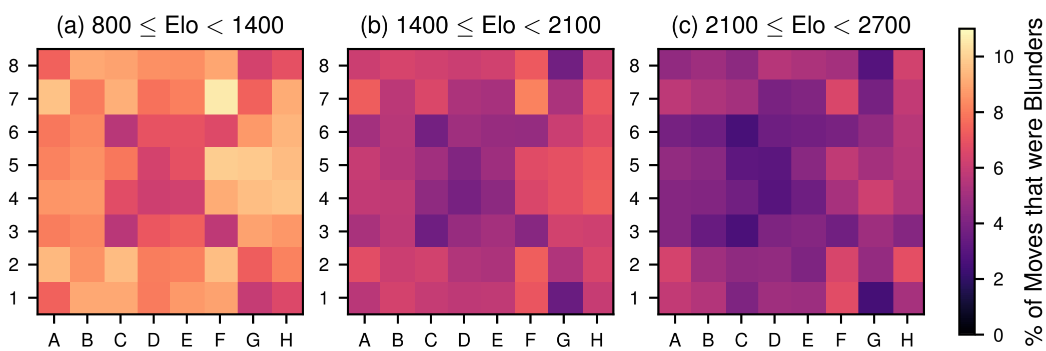

Figure 2: The heatmaps show the relative error frequency per square. In general, better ranked players blunder less, thought at similar spots on the board as lower ranked players. **g1** and **g8** are the squares the king moves to when kingside castling.

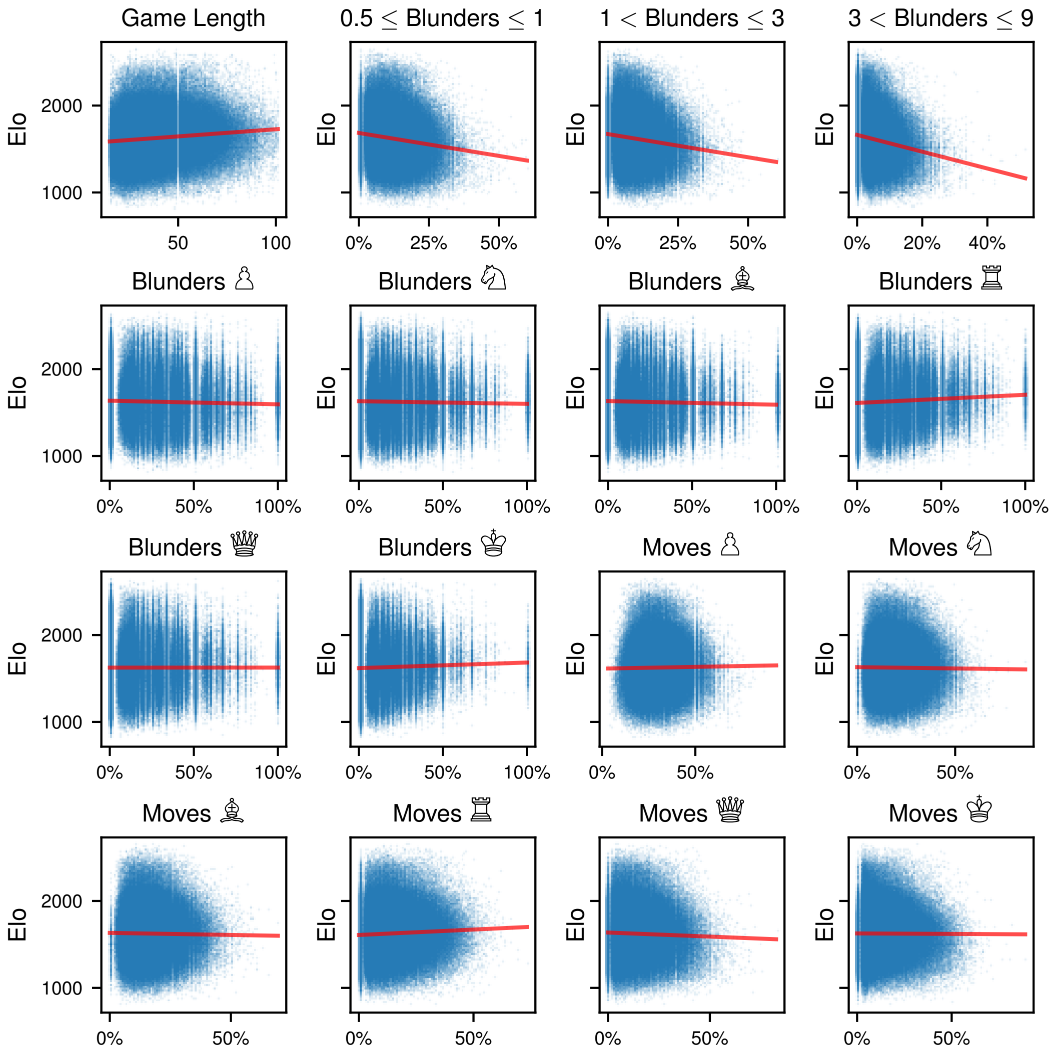

Figure 3: The figure shows the correlation between 16 different handcrafted features based on the gameplay and Stockfish evaluation with respect to the Elo. The blunders were divided into three severity domains. In addition, the number of moves of the individual figures was calculated as well as the relative blunder frequency per figure.

To create heatmaps mirroring the chessboard, we collected the destination squares for all blunders and moves. Then we calculated the ratio of blunders to moves for each square, resulting in a plot where one can recognize the regions of the board in which blunders are more likely to happen. An interesting finding is that pattern of the blunder ratio is similar among players of all Elos, though having different intensities. As most of the gameplay happens in the middle of the board, one could assume that players are more attentive when moving figures to or around the middle, whereas dangerous positions are easier to overlook when they occur outside the main area of the game. An interesting aspect of the heatmap is that the squares **g1** and **g8** are so apparent. This corresponds to the squares the king moves to when kingside castling. Since castling is one of the basic moves, performed in almost every game, one might conclude that kingside castling is a safe move to make. This complies with chess theory, generally considering kingside castling to enable safer gameplay than queenside castling [2].

In the correlation plots of Figure 3, blunders of severities from 0.5 and upwards (excluding moves that lead to mate in a fixed number of steps) were considered, except for the three plots at the top right. We can see that better players play longer games on average. Presumably, because their gameplay does not result in a mate position as quickly. The plot $0.5 \leq \mathit{Blunders} \leq 1$ visualizes all blunders that worsened the chess position by less than a pawn. Plot $1 < \mathit{Blunders} \leq 3$ contains all blunders that

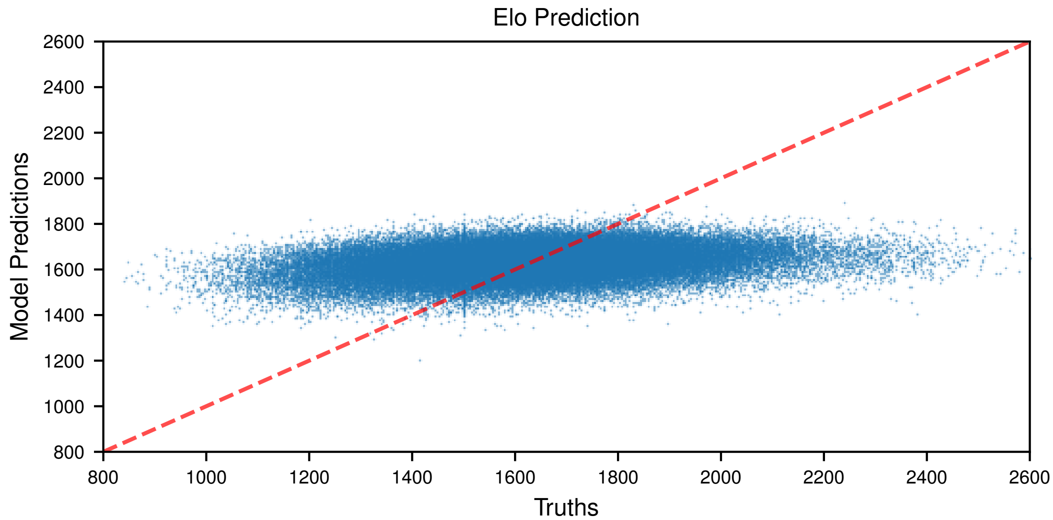

Figure 4: Linear regression to predict the Elo of a player based on the 16 features of Figure 3.

lost the player at least a pawn. The plot on the top right includes blunders where the chess position worsened in a way that is equivalent to losing at least a minor figure, which is practically always a game-deciding mistake. The rather obvious fact that worse players make more serious mistakes is easily recognizable. It is noticeable that the relative blunder frequency of the individual pieces, with the exception of the rook, is relatively similar across all skill levels. Furthermore, all players move their pieces almost equally often. Only with the rook and the queen are there slight deviations. Players with a higher Elo rating use their rook more, and their queen less frequently than players with a lower Elo rating.

In Figure 4 it becomes clear that linear regression on standardized data does not work very well with the features we have hand crafted. The absolute error deviation is 168.73 Elo points, which corresponds to an error of about 10.7%. One reason for this could be that the Elo is almost normally distributed, and thus a large part of the data points are centered around 1500 Elo. Moreover, it is probably very difficult to draw accurate conclusions about the playing strength of the players involved in a chess game based on a few features, since chess is a very complex game. We also ran a ridge regression model, where we noticed that the loss is smallest when the alpha value is approaching zero, which is equivalent to a linear regression.

4 Conclusion

Players with lower Elo ratings tend to blunder with all pieces, while blunders of more experienced players are somewhat limited to the higher-valued rook and queen. Surprisingly, there are no obvious differences in the blunder pattern on the board for different Elo ranges. To achieve better results for the regression, we would have needed a more complex regression model or further and more meaningful features. However, this is not a trivial task because the data is highly complex and there are no obvious choices for expressive features. Since PGN notation contains only the minimum of information needed to reconstruct a chess game, it is difficult to extract further features without having to recompute every single game from the algebraic notation. This is linked to considerable computational effort, which means a more in-depth analysis was not possible due to limited hardware resources.

References

[1] A. E. Elo. *The rating of chess players: Past and present.* Arco Publishing, New York, 1978.

[2] Mike Henebry. *Chess Words of Wisdom: The Principles, Methods and Essential Knowledge of Chess.* SCB Distributors, 2011