Chess Openings and Elo: Patterns and Predictions of Game Outcomes

Kyle Tranfaglia

Department of Mathematics, Salisbury University

Data Science 470: Research Methods in Data Science

Dr. Kyle Teller

December 04, 2024

Abstract

This study explores the relationship between chess players Elo ratings, opening choices, and game outcomes, leveraging a dataset of over six million games from Lichess.org. Chess openings play a pivotal role in setting a game's strategic direction, and Elo ratings are widely used to measure player skill and predict outcomes. However, questions remain about how well Elo ratings alone predict results and whether other features, such as opening choices and event types, can enhance predictive accuracy. The analysis identified key trends in opening preferences across Elo groups and event types. While the Sicilian Defense, French Defense, Queen's Pawn, and Scandinavian Defense were consistently popular, other openings were popular for individual Elo groups. Event types also influenced preferences, with openings like the Italian Game, Philidor Defense, Zukertort, and Modern Defense appearing in specific formats. Two models were built to predict game outcomes. A logistic regression model using Elo-related features achieved 64.9% accuracy, outperforming a dummy model but highlighting limitations in using Elo alone. A random forest model incorporating categorical features had a slightly lower accuracy of 60.2%, indicating the limited utility of these additional features. Feature importance analysis reaffirmed Elo difference as the most critical predictor. These findings emphasize the central role of Elo ratings in outcome prediction and the need for additional contextual and comprehensive data to improve predictive accuracy.

Keywords: Chess, Elo Ratings, Chess Openings, Chess Events, Predictive Modeling

Chess Openings and Elo: Patterns and Predictions of Game Outcomes

Chess is one of history's most iconic board games, renowned for its complexity and strategic depth. It is a globally popular board game simulating a battle between two kingdoms. Played in turns with no hidden information, it relies purely on strategy, making luck virtually irrelevant (Chess.com, n.d., "Chess"). With millions of players worldwide and an ever-growing community on online platforms, chess has become an ideal subject for data-driven analysis. Among the fundamental aspects of chess are the opening moves, which set the game's initial strategy and significantly influence the middle and endgame dynamics. Some openings are more static and intuitive, while others are more dynamic and variable. The opening choice can reflect a player's aggressive or defensive style and varies across skill levels. A particular opening may not forecast a clear winner, but it certainly shapes the type of match to follow.

For the best game experience, players want a challenging match, and to ensure this, both players must have a similar skill level. The chess community relies on the Elo rating system to measure player skill; created by Arpad Elo, this system has become the standard for ranking players based on performance, assigning higher ratings to more skilled players (Chess.com, n.d., "Elo rating System"). While Elo is instrumental in creating balanced matchups, it offers only a partial view of a player's capabilities, particularly when predicting individual game outcomes. The Elo rating system estimates the outcome probability for players based on their prior game results. Although it's often considered a measure of absolute skill, Elo is more about relative performance, adjusting ratings after each game to reflect win/loss probabilities between players (Chess.com, n.d., "Elo rating System"). The Elo system serves to separate players by their expected performances in an attempt to match players of similar skill levels. This is crucial to any competitive game, and chess is no exception.

Elo alone is likely not a strong identifier when predicting game outcomes, although it does present some promising predictability. Since Elo is a more relative performance matrix of how well a player has performed against similar players, and all humans are prone to bad games and mistakes, predicting chess game outcomes is immensely complex. For instance, a player rated 100 points higher than their opponent is expected to win roughly five out of eight games, and a player with a 200-point advantage will likely win three out of four games (Chess.com, n.d., "Elo rating System"). Thus, Elo advantages tend to indicate a likelihood of victory. However, the extent to which Elo-related data can predict a game result is unclear. Many other features, such as time control, event type, opening choice, and even previous performance, are available for predicting game outcomes, but the issue is determining the most significant.

This study investigates patterns in opening choice among different Elo rating groups and event types and evaluates Elo's predictive power for game outcomes. The goal is to determine if frequent use of specific openings indicates a player's Elo and the extent to which event types and Elo groups impact the openings a player uses. With this information, a game outcome prediction model can be built using only Elo-related data and compared to another model that uses other player statistics, such as the event played, the time control, and the opening choice, along with Elo-related data. The updated model is expected to present the importance of specific game features in predicting game outcomes or inform a search for significant features. This study aims to research improvements to a dynamic and fair matchmaking system.

Background

Rating systems are fundamental in competitive gaming environments for assessing player skill and predicting match outcomes. The Elo rating system, one of the most well-established methods, provides a framework for assigning skill levels to players based on previous match

outcomes. The system's essential function is to estimate the probability of one player winning against another by comparing their respective Elo ratings using a logistic function: P(W) = 1/1+10((blackElo-whiteElo)/400) (Prats Rodríguez, 2023). This approach allows the Elo system to make probabilistic predictions about match results, with higher-rated players expected to have a higher likelihood of winning.

To create balanced and fair matchups, rating systems must prioritize several core principles: accuracy in predicting outcomes, computation efficiency, incentive compatibility to discourage gaming the system, and human interpretability to make ratings understandable and relevant to users (Ebtekar & Liu, 2021). Classical rating systems such as Elo, Glicko, and Glicko-2 adhere to these principles by modeling player skill as an evolving dynamic variable. However, these systems have flaws; for example, players can exploit "volatility farming" to artificially boost their ratings, undermining the system's fairness and credibility. Addressing these vulnerabilities is essential to preserving the integrity of skill-based matchmaking (Ebtekar & Liu, 2021). Despite these flaws, rating systems are crucial to establishing a balanced matchmaking system that keeps players engaged and challenged to a reasonable extent.

The Elo system's effectiveness improves as players engage in more matches, allowing the model to draw on a larger dataset for skill estimation. Ratings adjust after each game to reflect win probabilities and outcomes, meaning that a player with a higher rating will have a higher chance of winning. However, they will gain fewer rating points for a victory compared to a lower-rated player (Cornell University, 2022). Despite the Elo system’s predictive potential, it does not account for many factors that can influence outcomes, such as player fatigue, recent performance trends, or play style. Thus, while Elo provides a foundational model for skill

assessment, emerging challenges and limitations indicate the potential benefit of integrating more complex, contextual features into these models.

Research Questions

By examining a dataset of over six million games, the aim is to answer the following questions: How do opening choices and event types differ across Elo ratings, and are certain openings more likely in specific events or Elo ranges? How do rating differences and player Elo impact game outcomes and the predictability of the game results? Which game features, along with Elo difference, help to predict the game result, and how can these insights inform matchmaking criteria? The results of this study could enhance the understanding of strategic preferences in chess and the limitations of Elo as a predictive tool. Additionally, it may present insights into improving matchmaking by identifying potential beneficial features in predicting game outcomes and rejecting features that may be misleading or lack significance.

Methodology

This study utilized a dataset of 6.25 million chess games sourced from an open chess database by Lichess.org, covering games played online on the Lichess website during July 2016 (Revel, 2016). The dataset includes fifteen features: Event (type), White, Black (ID), Result (outcome), UTCDate (date of match), UTCTime (time of match), WhiteElo, BlackElo (Elo rating), WhiteRatingDiff, BlackRatingDiff (post-match Elo difference), ECO (opening code), Opening (name), TimeControl (plus increment), Termination (type), AN (moves as text). It is ideal for analysis given the immense dataset size, completeness, and abundant features. It is especially strong for player categorization and correlating features, such as Elo and openings or event types. The dataset was chosen to reveal player dynamics and patterns across Elo ranges and construct a predictive model for the game result.

In the preprocessing stage, Pandas was used to load and preprocess the dataset, including inspecting data for inconsistencies and calculating frequencies and distributions to identify patterns. Data visualization was performed using Matplotlib and Seaborn, focusing on box plots, bar graphs, histograms, and line graphs to illustrate feature distributions and relationships, such as the frequency of openings across Elo groups. Finally, Scikit-Learn was used to split the dataset into training and testing subsets. Models such as logistic regression and random forest classifiers were built to classify game outcomes. Performance metrics, including accuracy, cross-validation, and confusion matrices, were used to evaluate the models.

In the preprocessing stage, frequency analysis was conducted to identify the most common values for several key features, including generalized opening names, time controls, event types, terminations, and Elo groups. The analysis was limited to the top 50 most frequent values for each feature to manage the large number of unique entries in these columns. For Elo groups, the distribution showed a descending frequency pattern: intermediate players were the most common, followed by advanced, expert, master, and beginner players, with a notable drop in the number of beginners compared to other groups (Figure A1 in Appendix A).

A similar descending trend was observed for game events, with blitz games being the most common, followed by classical and bullet games. Tournament events (e.g., blitz tournament, classical tournament) occurred less frequently, while correspondence games were significantly less common than all other event types (Figure A2 in Appendix A). Termination reasons followed a distinct order, with normal outcomes being the most frequent, followed by time forfeits, abandonment, rules infractions, and uncompleted games (Figure A3 in Appendix A). Among time controls, the most popular formats included 300+0, 180+0, 60+0, 600+0, and 30+0 (in the format seconds+added time per move), indicating a preference for faster time

formats (Figure A4 in Appendix A). Generalized openings underlined popular opening strategies, with Sicilian Defense, Queen's Pawn, French Defense, Scandinavian Defense, and Queen's Gambit being the most frequently used (Figure A5 in Appendix A). This frequency analysis highlighted key patterns and provided a foundation for understanding player behavior and preferences. Identifying the descending frequency patterns amongst all features was very insightful, showing that none of the features are equally distributed and that trends are present.

A histogram of average Elo ratings was plotted to better understand the distribution of skill levels in the dataset, revealing a normal distribution with a mean Elo of approximately 1750 (Figure 1). This visualization provided a clear overview of the general skill levels represented in the dataset, with most games involving players near this average rating. Additionally, boxplots of average Elo ratings by event type were created to explore how skill levels varied across game formats. The boxplots revealed that the average Elo was highest in tournaments and lowest in correspondence games, with similar distributions in blitz, bullet, and classical (Figure 2). Classical tournaments and correspondence games also showed the least variance in Elo ratings, suggesting more consistent player skill levels in these formats. These visualizations were instrumental in identifying trends and patterns related to player skill and game formats, laying a foundation for deeper analysis.

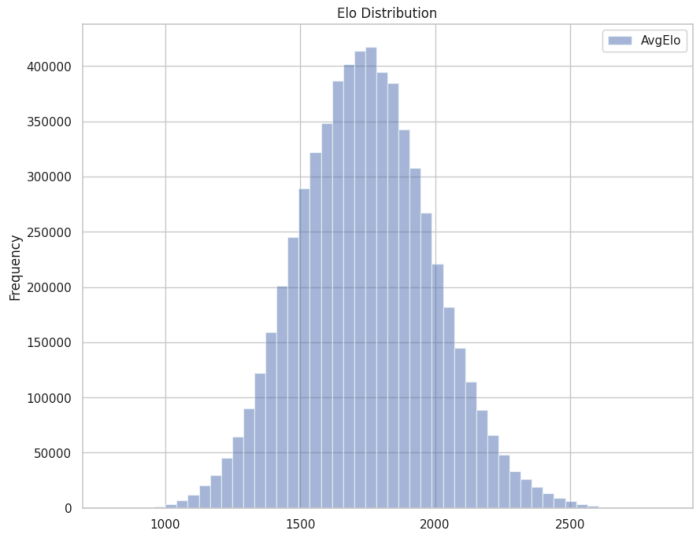

Figure 1: A histogram showing the frequency distribution of player Elo (match average)

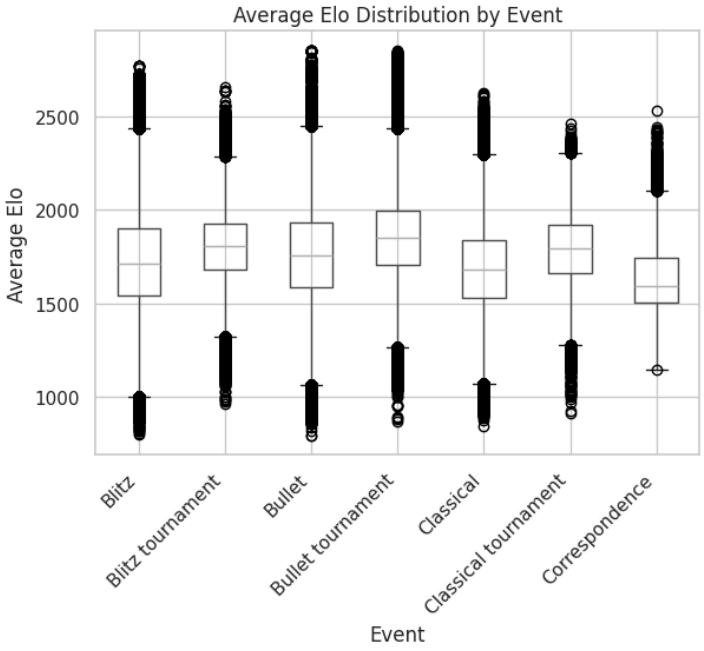

Figure 2: A boxplot showing the player Elo distribution (match average) for each event type

In the preprocessing stage, feature engineering and data transformation were applied to enhance the dataset and standardize key features for analysis. Two new columns, WhiteEloGroup and BlackEloGroup, were created to categorize player strengths into skill levels, "Beginner,"

"Intermediate," "Advanced," "Expert," and "Master," based on their Elo ratings (Beginner ≤ 1200, Intermediate ≤ 1800, Advanced ≤ 2000, Expert ≤ 2200, Master > 2200). A new feature, AvgElo, was introduced to represent the average Elo rating for each game, providing a measure of the match skill level. Another feature, EloGroup, categorizes games by the average player strengths based on the AvgElo values, allowing for analysis at a group level.

To standardize the representation of chess openings, the generalized_Opening column was created, capturing only the first two words of each opening name and removing any punctuation, such as colons and commas. This transformation grouped opening variations under a unified opening name, making analysis more consistent and manageable. Together, these engineered features and transformations enabled a deeper exploration of trends and patterns in the dataset, particularly concerning Elo ratings and opening choices.

Results

Event Types and Opening Choices By Elo

The analysis of opening choices by Elo rating revealed notable patterns and trends, highlighting the connection between player skill levels, strategic preferences, and event types. The Sicilian Defense emerged as the most popular opening across all Elo groups, except for Beginners, where it was less favored (Figure 3). Other widely used openings included the French Defense, Queen's Pawn, and Scandinavian Defense, which were common across most skill levels (Figure 4). However, specific openings showed stronger preferences in certain groups: the English Opening and Caro-Kann Defense were prevalent among Experts and Masters, while the Queen's Gambit stood out for Advanced players. Other unique choices were also observed, such as the Van't Kruijs Opening for Beginners and the King's Pawn Opening for both Beginners and Intermediates (Figure 3 and Figure B1 in Appendix B for more insights).

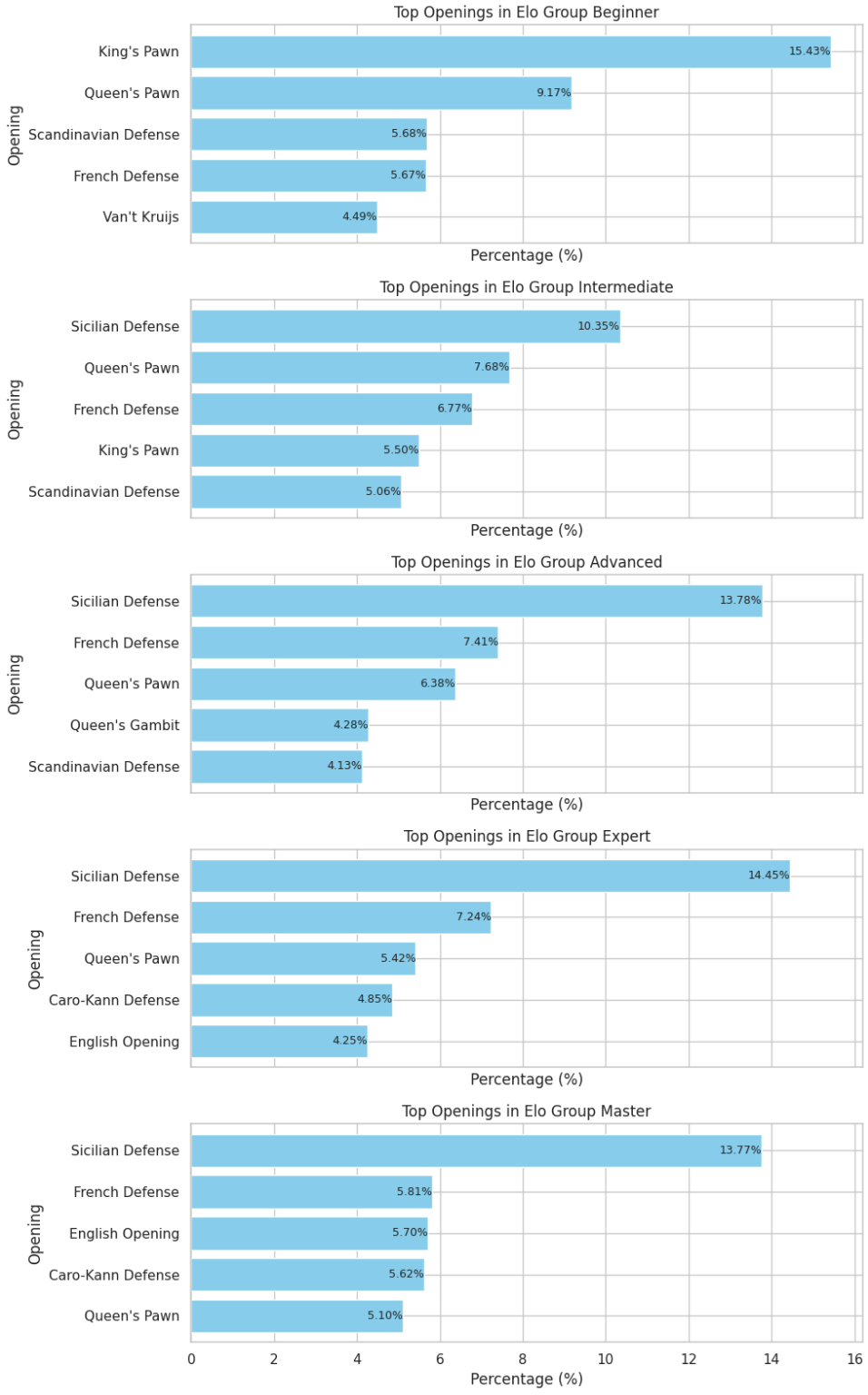

Figure 3: A bar graph showing each Elo group's five most frequent openings by percentage.

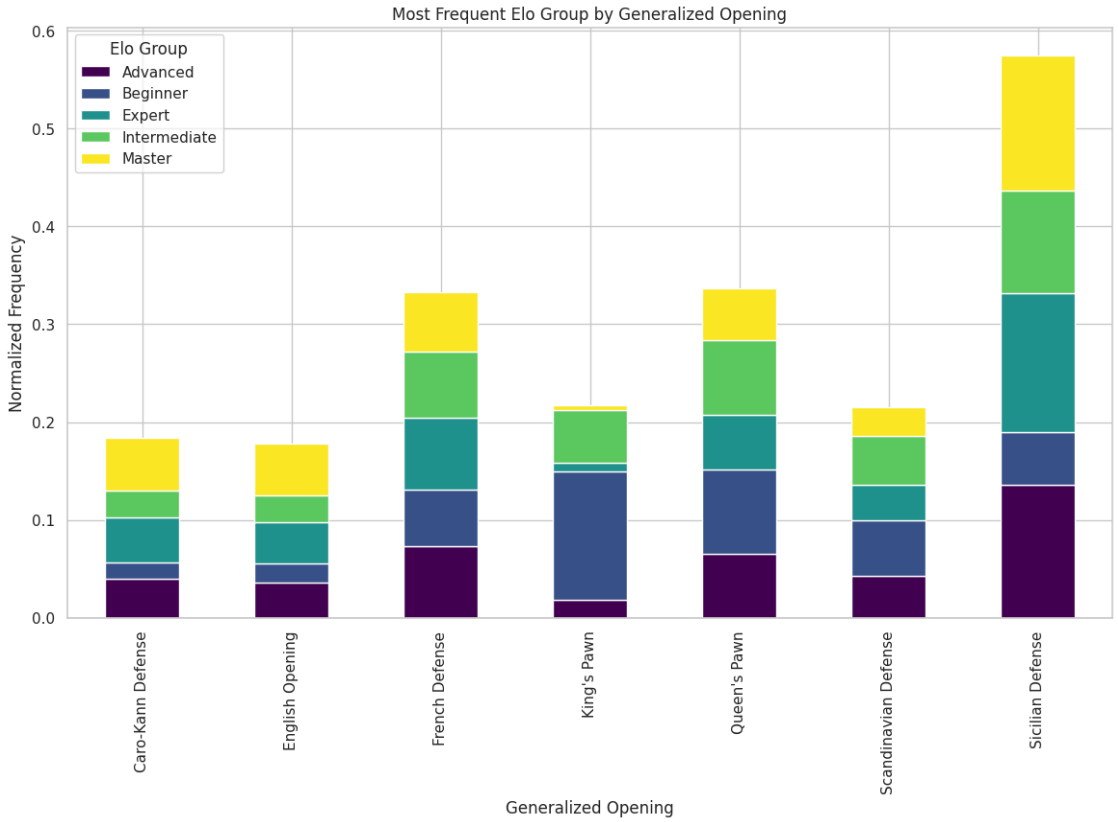

Figure 4: A bar graph showing the top seven openings and the distributions of Elo groups.

Further analysis revealed a significant connection between Elo groups and event types, with distinct variations in opening preferences across different formats. Lower-rated players were more prevalent in correspondence games, while higher-rated players dominated bullet tournaments. Additionally, as the average Elo for Black and White players increased, the event type changed linearly: correspondence had the lowest average Elo, followed by classical, blitz, bullet, classical tournaments, blitz tournaments, and finally, bullet tournaments with the highest average Elo (Figure 5). The most significant gaps in average Elo occurred after correspondence and before bullet tournaments, indicating a notable shift in player skill levels between these event categories.

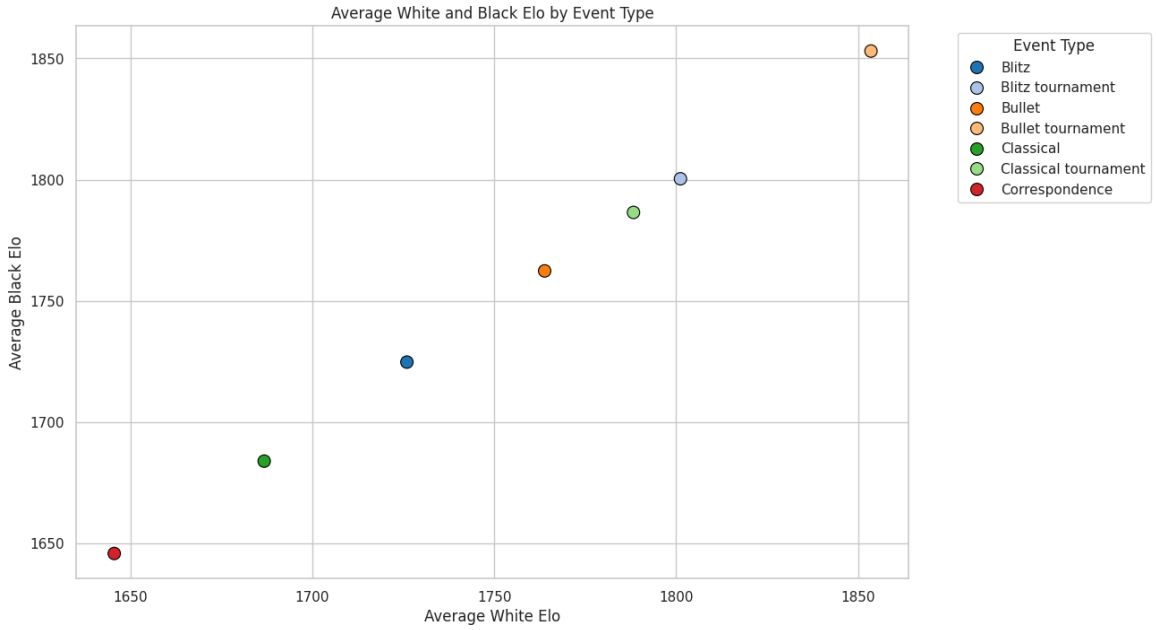

Figure 5: A line graph showing the average white and black player Elo for each event type.

Although the Sicilian Defense remained the most popular opening across all event types, other openings, such as the Italian Opening, Philidor Defense, Zukertort Opening, and Modern Defense, were event-specific (Figure 6). These findings underscore the refined relationship between Elo ratings, event types, and opening strategies, reflecting how players adapt their strategies based on skill level and play format (Figure B2 in Appendix B). The top three openings, Sicilian Defense, French Defense, and Queen's Pawn, were consistent across all event types, demonstrating their broad appeal and versatility at various skill levels and game speeds (Figure 6). Beyond these, certain openings stood out as event-specific preferences. For example, the Queen's Gambit appeared prominently in blitz and blitz tournaments, while the Caro-Kann Defense was favored in blitz tournaments and bullet tournaments. Finally, the Modern Defense emerged as a popular choice in bullet games, likely due to its flexibility in fast-paced games.

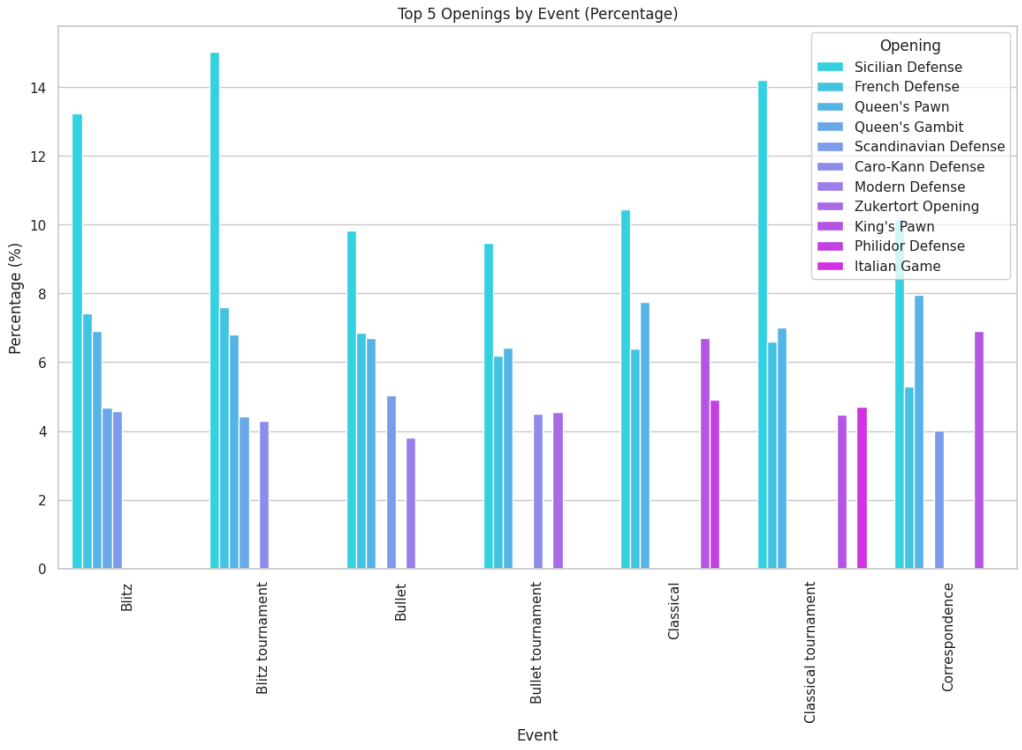

Figure 6: A bar graph showing each event type's top five openings by frequency percentage.

In classical formats, the top openings included the King's Pawn and Philidor Defense, with the Italian Game becoming a notable choice in classical tournaments, reflecting its popularity in traditional, methodical play. The King's Pawn was among the top five openings in correspondence games, suggesting a preference for classic, well-studied strategies in slower, more deliberate formats. These variations illustrate how players tailor their opening choices based on the event's time control and competitive nature while relying on universally strong openings like the Sicilian and French Defenses.

The consistency of the top three openings across event types underscores their general effectiveness and adaptability. Still, the distinct preferences for other openings in specific formats highlight the nuanced relationship between Elo ratings, event types, and opening

strategies. These findings suggest that players adapt their playstyle and choice of openings based on the demands and pacing of different game formats.

Game Outcomes by Elo Rating Difference and Predicting Outcomes with Elo

The analysis of game outcomes by Elo rating difference revealed a strong, approximately linear relationship between Elo difference and win probability for White. As White's Elo advantage increases, so does the win probability (Figure C1 in Appendix C). For example, at a 0 to 10 Elo difference, White's win probability is 50% (with a 5% chance of a draw). In contrast, with a 500+ Elo difference, the win probability rises to 93%, and the draw probability decreases to just 1% (Figure 7). A similar trend was observed across Elo groups, with slight variances: Intermediate players showed a win probability of 92% at the highest Elo difference range, while Masters reached 94%. Notably, no data was available for Beginners in the 451 to 500 or 500+ Elo difference ranges, preventing conclusions about this group (Figure C2 in Appendix C). The draw probability showed a linear decline as White's Elo advantage grew, starting at 4.5% for a 0 to 10 Elo difference and decreasing to 1% at a 500+ Elo difference (Figure 7). This indicates that as the Elo disparity between players increases, games are less likely to end in a draw and more likely to result in a decisive outcome for the higher-rated player.

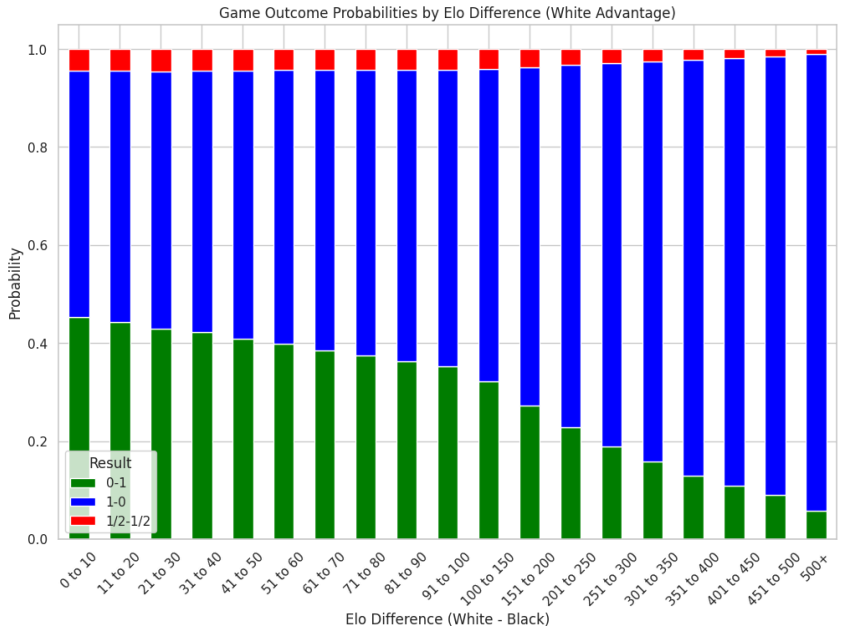

Figure 7: A bar chart showing game outcome probabilities for each Elo difference bin.

A comparison of observed win probabilities with those predicted by the logistic function P(W) = 1/1+10^((blackElo-whiteElo)/400) showed a close match, though the data slightly undervalued White's win probabilities compared to the function (Figure 8). This discrepancy may be attributed to how draws are accounted for in the dataset. Since draws are treated as distinct outcomes in the data, they increase the rate of "non-win" events for White, lowering the observed win probabilities. This effect persists even when draws are converted to binary outcomes (win vs. no win). Despite this, the calculated White win probability is a fairly accurate predictor of game outcomes, confirming its reliability and highlighting the complexities introduced by draws in real-world data.

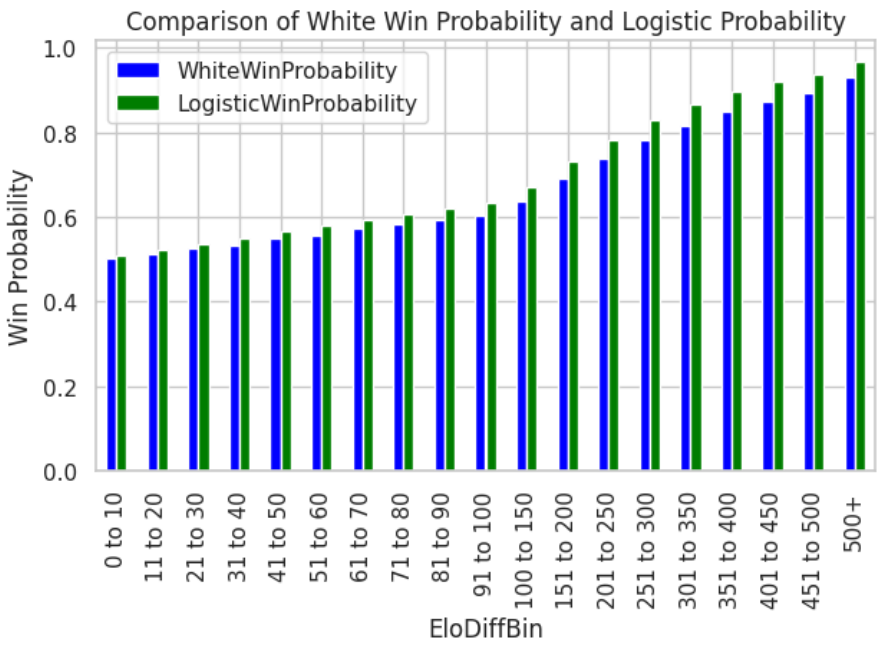

Figure 8: A bar graph comparing the White win probability calculated using the dataset and the White win probability calculated using the logistic equation for each Elo bin.

The predictive model for game outcomes was constructed using logistic regression, leveraging only Elo-related features: EloDifference, WhiteElo, BlackElo, and AvgElo. The model achieved an accuracy of 64.9%, outperforming a dummy model, which would have an accuracy of 50%. Notably, the model demonstrated near-equal rates of correct and incorrect classifications for Black and White wins, indicating a balanced performance across both outcomes (Figure 9). One contributing factor to this performance is the classification setup, where binary outcomes were defined as White wins versus non-White wins, categorizing ties as Black wins. This approach aligns with the Elo system's predictive emphasis on win probabilities but slightly favors White due to the inherent advantage of the first move in chess, especially at near-equal Elo ratings.

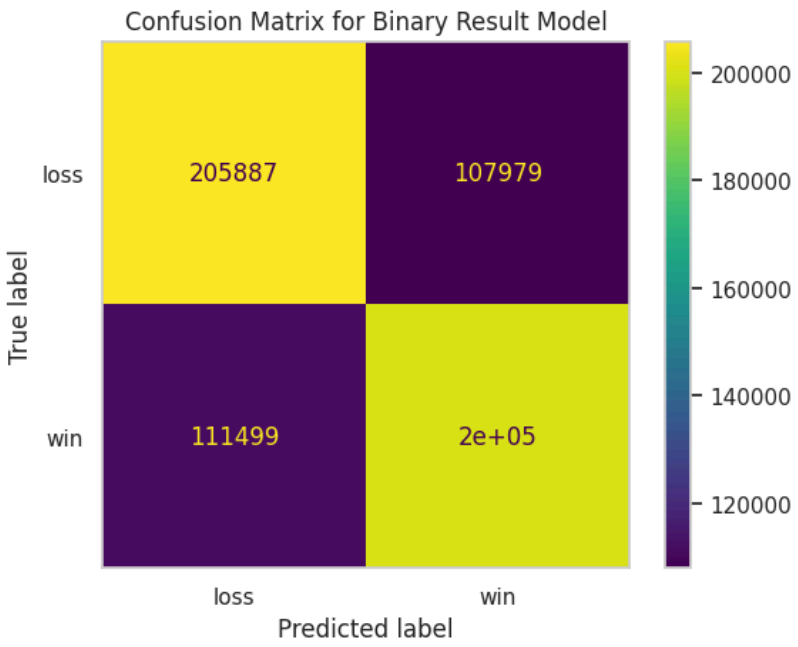

Figure 9: A confusion matrix showing the classification results for the binary (win vs no win) linear regression model using only Elo-related features.

While the model performed better than random, its accuracy suggests room for improvement. Incorporating additional features could enhance predictive power. Potentially useful features include game factors like opening choice, time control, and event type. These additions may capture contextual factors that Elo ratings alone cannot reflect, offering a more nuanced understanding of game dynamics and improving the model's ability to predict outcomes.

Predicting Game Outcomes with Available Game Features

Further analysis of Elo difference and game outcomes revealed distinct patterns: White wins are associated with a mean Elo difference greater than zero (approximately at the first quartile), Black wins with a mean less than zero (approximately at the third quartile), and ties with a mean of approximately zero. However, ties showed a much smaller data distribution, emphasizing their rarity compared to decisive outcomes (Figure 10). These insights informed the construction of a more comprehensive predictive model using a random forest classifier with

additional game features, including WhiteWinProbability (an engineered feature), Event, TimeControl, and generalized_Opening.

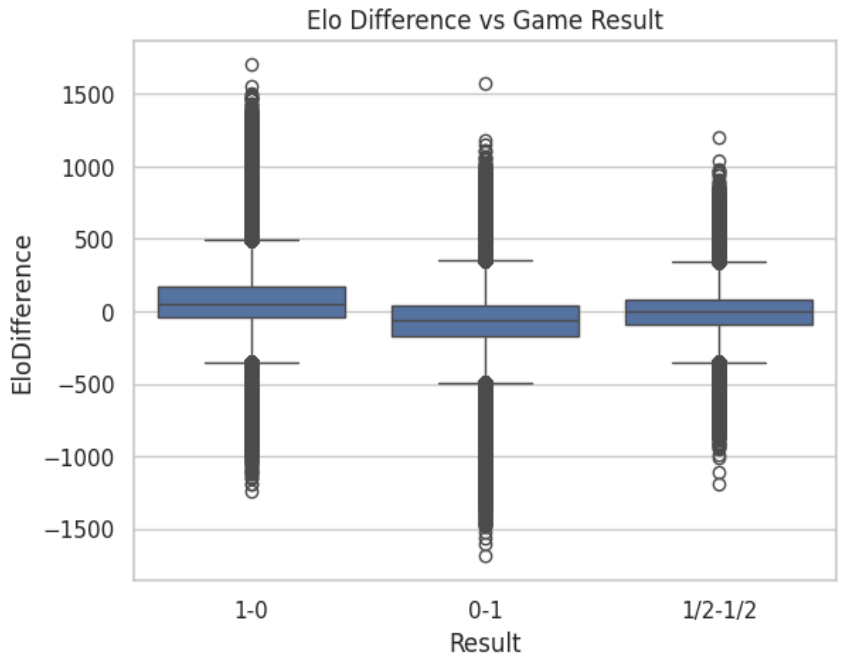

Figure 10: A box plot showing the data distributions of Elo difference for each game result type.

The model achieved an accuracy of 59.3% with a cross-validation accuracy of 60.2%, slightly underperforming compared to the Elo-based logistic regression model. Similar to the previous model, it correctly classified Black and White wins at near-equal rates, with ties again categorized as non-White wins, maintaining a performance advantage over a dummy model (Figure 11). Feature importance analysis highlighted EloDifference as the most influential predictor, with WhiteWinProbability and specific time controls and events (e.g., TimeControl_180+0, TimeControl_300+0, and Event_Blitz Tournament) contributing modestly.

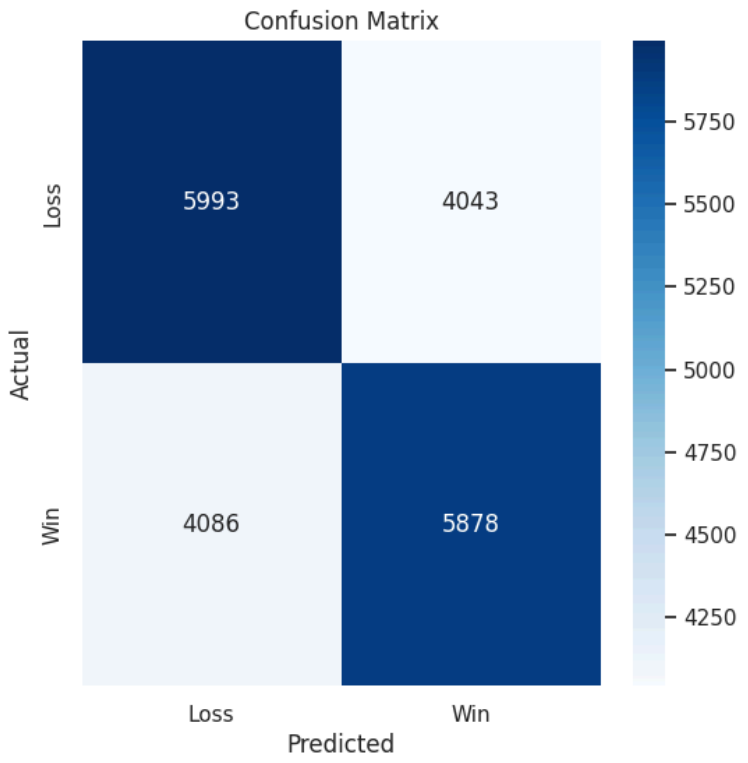

Figure 11: A confusion matrix showing the classification results for the random forest model using EloDifference, WhiteWinProbability, Event, TimeControl, and generalized_Opening.

While adding new features aimed to enhance performance, the model's overall accuracy remained modest, suggesting that further improvements require integrating features beyond Elo-related data, such as player history, move analysis, or psychological factors. The results underscore the limitations of the current dataset and the need for richer, contextual features to improve the predictive accuracy of game outcome models.

Discussion

The analysis of chess opening choices across Elo groups and event types revealed notable trends and relationships, reflecting how player skill and game formats influence strategic decisions. Across all skill levels and event types, the Sicilian Defense, French Defense, Queen's Pawn, and Scandinavian Defense stood out as the most popular openings, demonstrating their

broad applicability and effectiveness. However, certain openings were strongly associated with specific Elo groups, such as the English Opening and Caro-Kann Defense among Experts and Masters, the Queen's Gambit for Advanced players, and the Van't Kruijs Opening for Beginners. The event-specific analysis also highlighted unique preferences, with openings like the Italian Game, Philidor Defense, Zukertort Opening, and Modern Defense emerging in distinct formats, emphasizing the interplay between Elo ratings, event types, and opening strategies.

When modeling game outcomes, the logistic regression model built using Elo-related features achieved an accuracy of 64.9%, outperforming a baseline dummy model but still falling short of high predictive accuracy. This suggests that Elo ratings, while informative, are insufficient as standalone predictors. Adding categorical features such as generalized openings, time controls, and event types in a random forest model did not improve performance; instead, the cross-validated accuracy dropped slightly to 60.2%, indicating that these features may add noise without significant predictive power in the current dataset.

Feature importance analysis from the random forest model underscored EloDifference as the most influential predictor of game outcomes, far outweighing the contributions of other features like time controls and events. This finding reaffirms the centrality of Elo ratings in predicting game results but also highlights the limitations of existing data. While the engineered feature WhiteWinProbability, derived from Elo bins, offered a potential avenue for improvement, its reliance on Elo difference limits its practical utility and interpretability without added support from other potential predictive features.

Overall, these results emphasize that while Elo difference provides a solid foundation for predicting game outcomes, meaningful improvements likely require detailed data beyond Elo-related features. Incorporating more contextual information, specifically data beyond the

current dataset, could significantly enhance the predictive accuracy of models in future studies. These insights highlight the complexity of chess as a predictive problem. Elo-related features can not effectively predict the outcome of a chess match alone, as one sector of data can not encapsulate enough detail to foresee a result. There is a strong need for extensive and comprehensive data to capture the multifaceted nature of chess.

Future Work

Elo ratings provide a broad indication of a player's skill but may not capture critical nuances that could enhance game outcome prediction. Additional factors, such as players' recent game history, including win/loss streaks and frequency or intensity of recent matches, could be valuable predictive features. Psychological aspects, like player consistency under pressure (low time, opponent, losing position, etc) or fatigue, and in-game metrics, such as the frequency of poor moves or speed of moves, could also contribute to more accurate predictions. Integrating these data points into a predictive model would allow for a more comprehensive approach, recognizing the fluidity in player performance over time and potentially improving the robustness of predictions. Overall, such a complex game requires a complex model.

Beyond game outcomes, further work could investigate whether opening preferences and variations in opening choice over time provide predictive power for determining a player's Elo rating or Elo group. Players of specific skill levels often favor certain openings, and tracking a player's choice and adaptation of these openings could indicate their current Elo group and potential growth. Analyzing trends in opening evolution for individual players may offer insights into how changes in preferred openings correspond with Elo progression. Such models could be useful for player profiling, as they highlight how openings reflect players' style and strength.

A closing area of exploration involves players' time control and event-type preferences, which could serve as indirect indicators of skill and experience. Different time controls, such as bullet, blitz, and classical, emphasize varying aspects of player ability, from speed and reflex to calculation depth and patience. By examining patterns in a player’s time control and event-type participation over time, it may be possible to model and predict Elo groups or game outcomes using preferred formats. This line of research would likely help enhance matchmaking systems.

References

Chess.com. (n.d.). *Chess terms*. Chess.com. https://www.chess.com/terms/chess

Ebtekar, A., & Liu, P. (2021). Elo-MMR: A rating system for massive multiplayer competitions. *Communications of the ACM, 64*(8), 62–71. https://doi.org/10.1145/3442381.3450091

Prats Rodríguez, J. (2023). Chess cheats in check? *Significance, 20*(1), 4–5. https://doi.org/10.1093/jrssig/qmad008

Revel, A. (2019). *Chess games* (Version 1.0) [Data set]. Kaggle. https://www.kaggle.com/datasets/arevel/chess-games

Appendix A

Feature Frequency Distributions

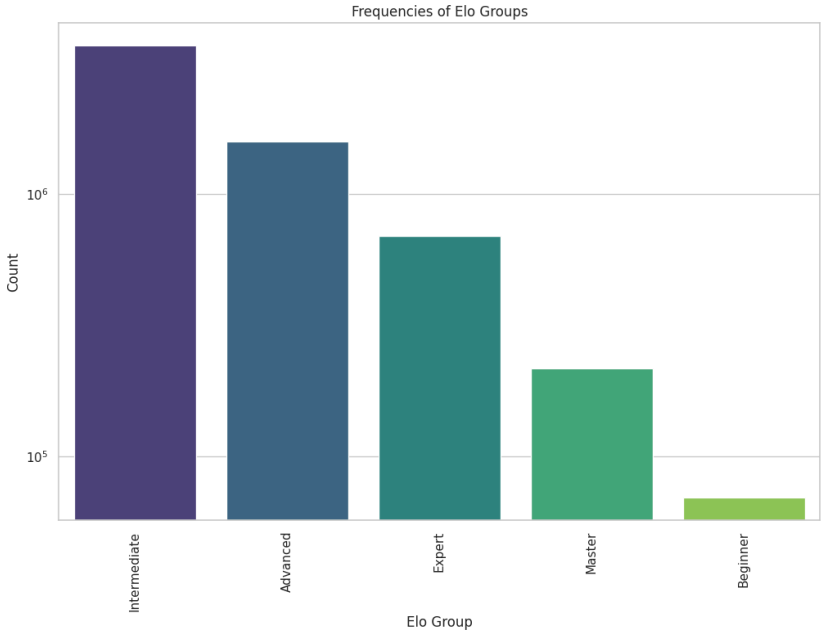

Figure A1: A bar chart showing the frequency of Elo groups in the dataset.

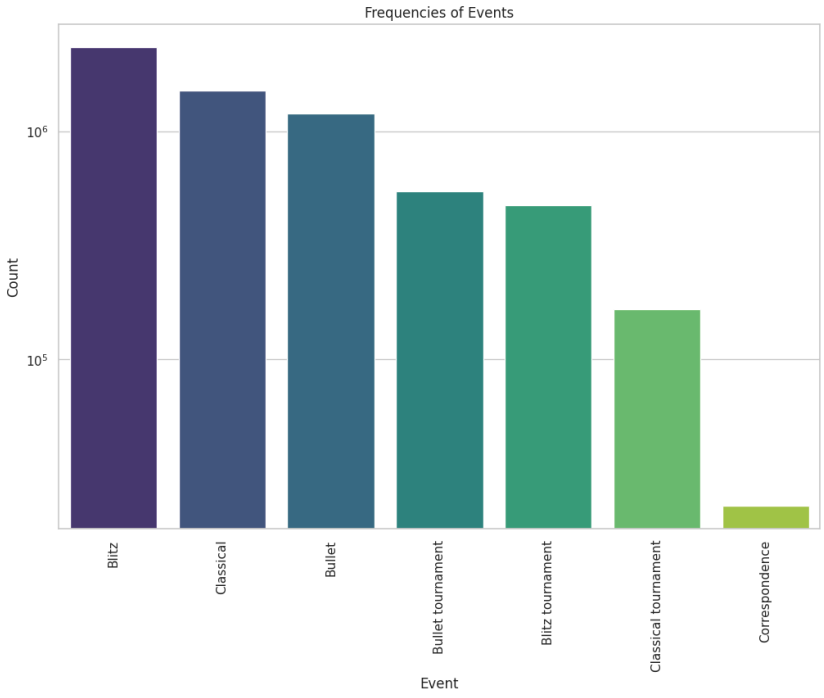

Figure A2: A bar chart showing the frequency of event types in the dataset.

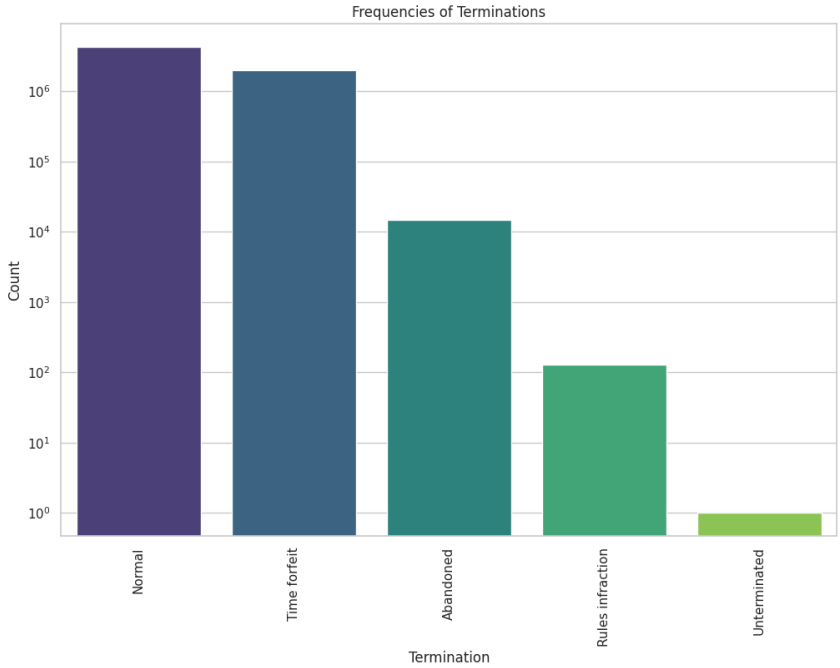

Figure A3: A bar chart showing the frequency of termination types in the dataset.

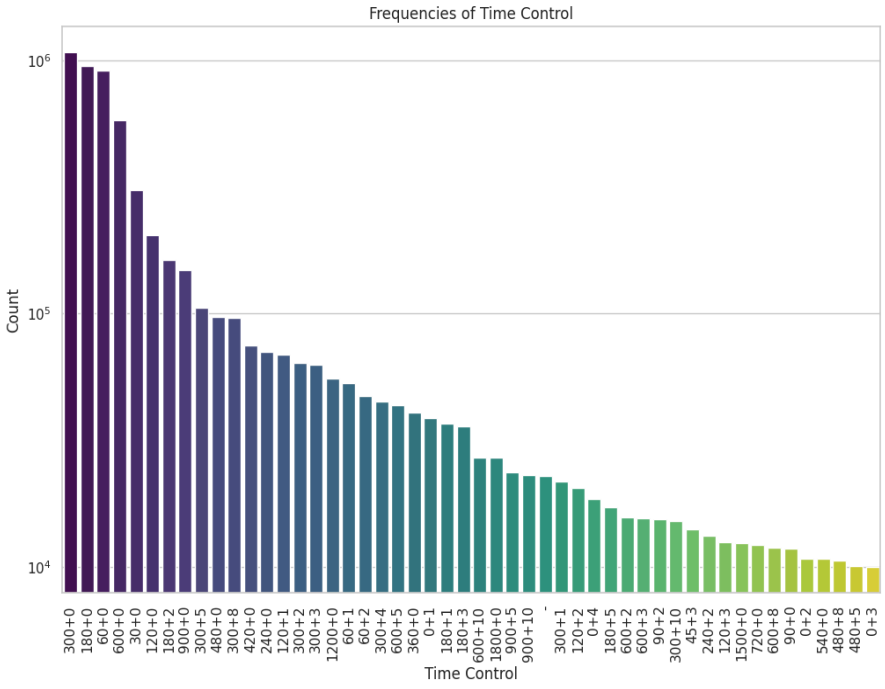

Figure A4: A bar chart showing the frequency of time controls in the dataset.

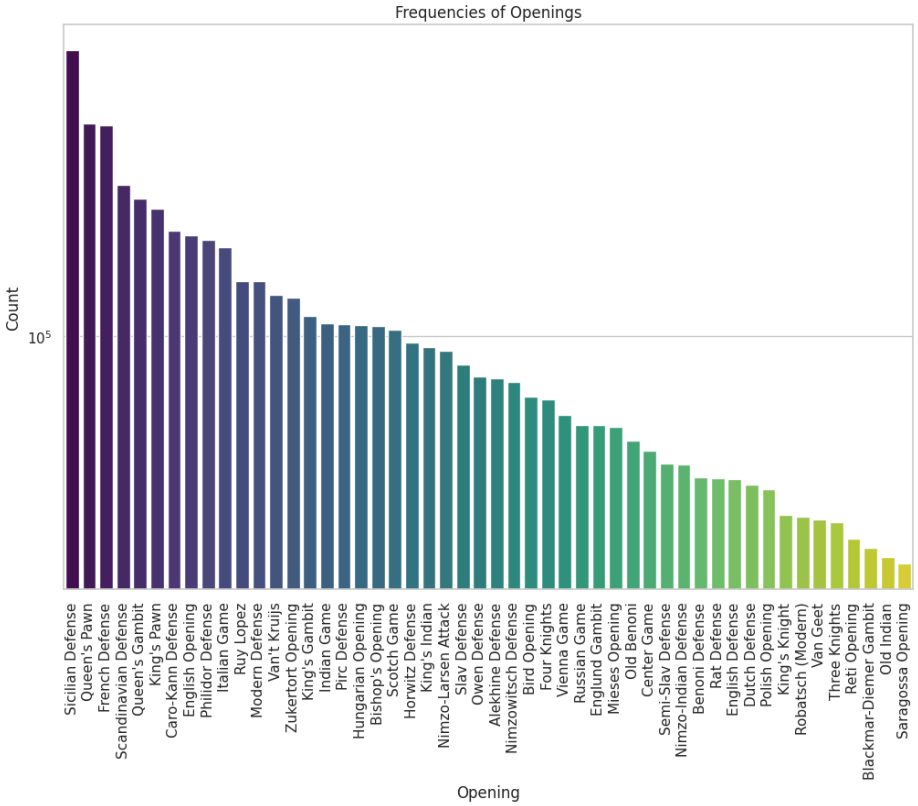

Figure A5: A bar chart showing the frequency of opening choices in the dataset.

Appendix B

Opening Choices and Event Types by Player Elo

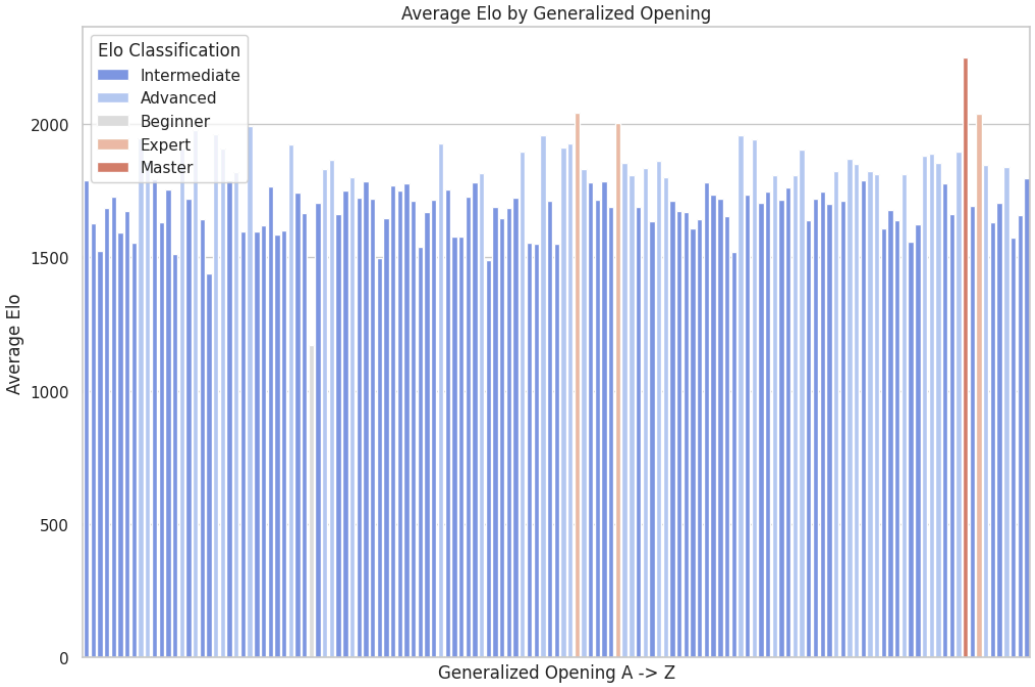

Figure B1: A histogram showing the average player Elo for each opening choice.

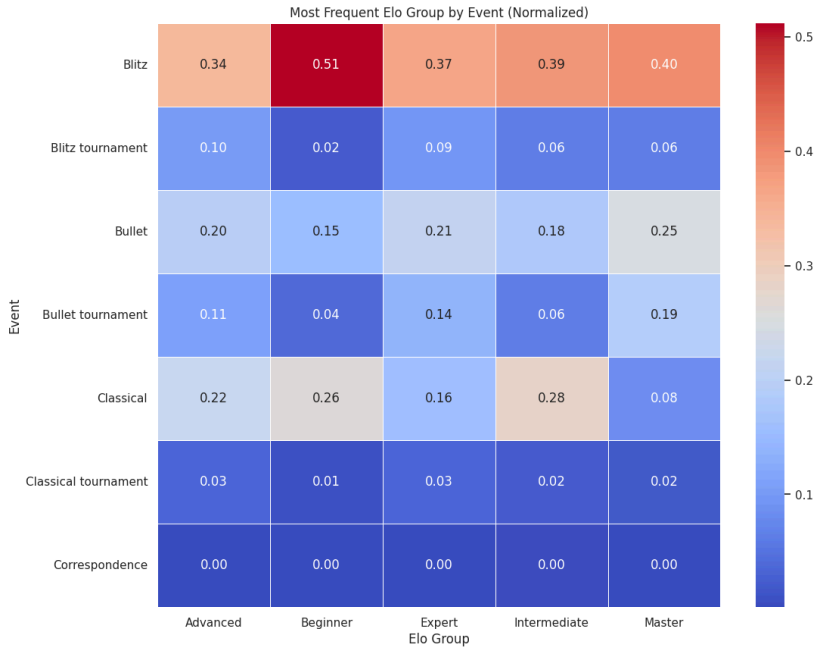

Figure B2: A heatmap showing player Elo group frequency distributions for each event type.

Appendix C

Game Outcomes by Player Elo Difference

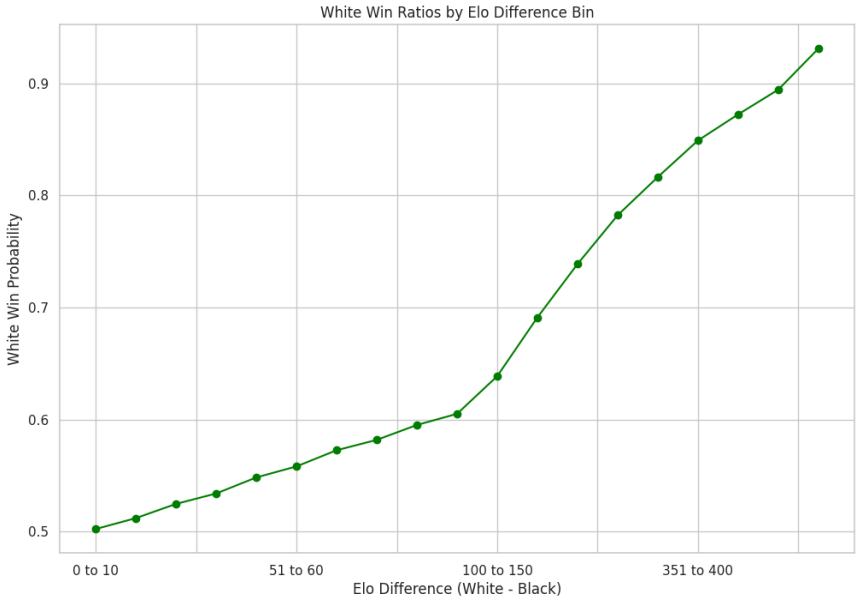

Figure C1: A line graph showing the white win probability using the dataset for each Elo bin.

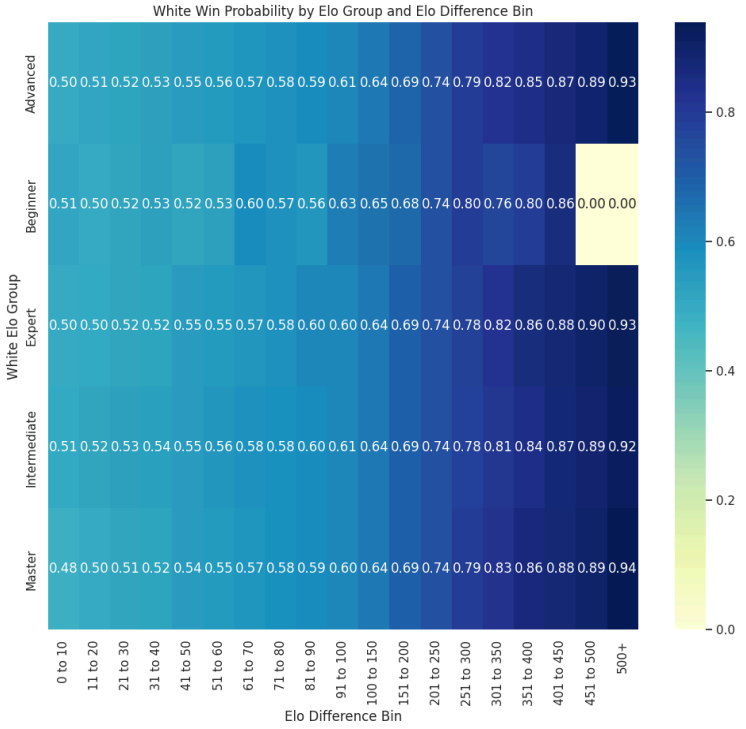

Figure C2: A heatmap showing the white win probability for each Elo group and difference bin.