STAT 3302 Final Project - Predicting Chess Outcomes

Glynis McCune, Aditya Menon, Nate Arens, and Samuel Gadkar

Abstract

Chess is a game that involves deep strategy and complex decision making that has long fascinated players, researchers, and AI developers alike. With the increasing availability of online chess data and advancements within statistical modeling, we now have the tools to properly analyze the game more systematically than ever before. To explore this complexity an analysis was done on a dataset containing detailed information from over 20,000 online chess games was analyzed to identify factors that influence match outcomes (J. Mitchell, 2024). Within the exploratory data analysis, it was found that rating difference had a strong relationship with game results, which encouraged further investigation into how various game characteristics affect win probabilities. This was supported by existing chess theory suggesting that while rating difference is important, other factors like opening selection and time control also play significant roles. Therefore, the group focused on answering the question of how game dynamics and player characteristics influence and help predict the winner of a chess match. Match outcomes were measured using a binary response variable indicating whether White won or lost. A logistic regression model was built to predict the probability of White winning versus Black winning. Several model specifications were tested, but the main focus was on one that included rating difference, turns, victory status, and opening moves, because it provided the most significant results. To choose the best model, stepwise selection was done and evaluated based on the lowest AIC and most significant covariates. Visualizations of the predicted probabilities were then created to examine whether specific game characteristics led to a higher probability of White winning, with special attention to rating differences and opening selections. The results showed that meaningful relationships exist between win probabilities and rating difference, with each rating point advantage increasing win odds by approximately 0.39%. Additional factors, such as opening move sequence and game length, also showed significant effects on match outcomes.

Introduction

The focus of this project was to determine how chess players and platforms utilize game data to better understand the factors that determine match outcomes and use that information to improve in game decision making. Emphasis was put on the following research questions:

- How does a player's probability of winning change based on rating difference, game length, and time control settings?
- Which game characteristics are most important in predicting chess match outcomes?
- Are there specific openings or playing conditions that significantly favor either White or Black?

Some important background information to consider is that the analysis is based on information from a dataset provided by Lichess.org, a popular online chess platform, which contains detailed information on over 20,000 chess games (J. Mitchell, 2024). The dataset includes player ratings, the number of turns per game, time control settings, the number of moves in the opening phase, match outcomes, and other variables.

These variables provide the group with a comprehensive and broad view of each game's structure and allow for the modeling of outcomes using statistical methods.

Addressing these research questions is important for both chess players and platforms. Chess players can use insights to improve their performance, while platforms can use the findings to enhance matchmaking algorithms. With millions of chess players worldwide, understanding how different factors affect game outcomes can help players make more informed strategic choices. Insights into the relationship between rating differences and win probabilities can support improvements to existing skill rating systems. Additionally, understanding which openings favor specific sides could guide players in optimizing their opening play based on statistical evidence rather than traditional theory.

Data and Methods

Dataset Description:

The dataset for this analysis was provided by Lichess.org and contains detailed information on over 20,000 chess games (J. Mitchell, 2024). The key variables within the dataset are:

Table 1: Variable Descriptions for Chess Dataset

Variable	Description
rated	Indicates whether the game was rated (1) or casual (0).
start_time	Represents the starting time of the game.
end_time	Represents the ending time of the game.
turns	Total number of full turns (1 turn = 1 move by White + 1 move by Black).
white_rating	White's ELO rating (player rating). A higher rating is better.
black_rating	Black's ELO rating (player rating). A higher rating is better.
victory_status	Describes how the game ended: 'mate' (checkmate), 'resign', or 'outoftime'.
opening_ply	Number of half moves played during the game's opening phase.
opening_eco	The opening sequence performed by the White

The dataset includes both categorical and continuous variables related to player characteristics and game attributes. However, some records contained incomplete information. Therefore, a subset of the data with complete values for all relevant predictors was selected, resulting in 19,108 complete chess game records for analysis.

Preprocessing Steps:

The original raw data underwent several preprocessing steps to properly supply the model with relevant information. First, from the original dataset, the categorical "winner" variable was converted into a binary numeric variable "winner_bin" (1 = White win, 0 = Black win). This step was done to make the problem a binary classification and allow the interpretation of coefficients in terms of increased or decreased odds of a white victory. Then, draws were excluded from the logistic regression modeling to ensure a simpler binary outcome. Since we are modeling this data with a logistic regression model, we can only have two potential outcomes, so we elected to drop the draws. Next, the "increment_code" variable was split into two separate numeric variables, the first being time_base (base time in minutes) and then time_increment (increment per move in seconds). This conversion allowed us to more easily numerically evaluate how the total available time and increment mechanics impact game outcomes. This also allowed for cleaner visualization and interpretation within the regression.Additionally, a "rating_diff" variable was created by calculating white_rating minus black_rating, to capture the rating gap between players as a meaningful predictor of outcome. The total game duration was also derived as the difference between end_time and start_time,

providing a direct measure of how long the game lasted. Categorical variables like victory_status and opening_eco were converted to factors to enable proper handling in the logistic regression model. Finally, games with missing or nonbinary outcomes were removed. After these preprocessing steps, the final dataset was ready for modeling.

Data and Methods Used

Since our model only focuses on when the outcomes of the game are a win, after dropping the games resulting in a draw, the new dataset contains 19,108 chess game records with 17 variables that provide detailed information about every match. The key variables which we wished to explore further are:

Table 2: Description of Variables in the Chess Dataset

Variable	Class	Description
Winner	Categorical	Indicates who won the game: 'White', 'Black', or 'Draw'.
Rated	Categorical	Indicates if the game was rated (affects player rating).
Time	Numeric	Duration of the game.
Turns	Numeric	Number of turns taken in the game.
Rating Difference	Numeric	Difference in player rating between white and black.
Victory Status	Categorical	How the game ended: 'Mate', 'Resign', 'Out of Time'.
Opening Play	Numeric	Length of the opening phase of the game.

The original raw data underwent several pre-processing steps to properly supply the model with relevant information. First, from the original data set, the categorical "winner" variable was converted into a binary numeric variable "winner_bin" (1 = White win, 0 = Black win). This step was done to make the problem a binary classification and allow the interpretation of coefficients in terms of increased or decreased odds of a white victory.

Then, draws were excluded from the logistic regression modeling to ensure a simpler binary outcome. Since, we are modeling this data with a logistic regression model, we can only have two potential otucomes, so we elected to drop the draws. Next, the "increment_code" variable was split into two separate numeric variables, the first being time_base (base time in minutes) and then time_increment (increment per move in seconds). This conversion allowed us to more easily numerically evaluate how the total available time and increment mechanics impact game outcomes. This also allowed for cleaner visualization and interpretation within the regression. Finally, games with missing or nonbinary outcomes were removed. This ensured we did not introduce bias or lead to convergence issues within logistic regression.

Exploratory Data Analysis

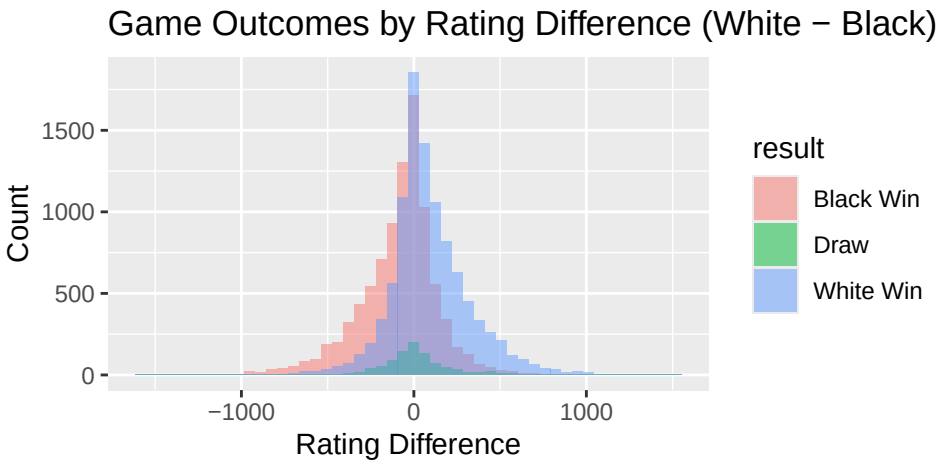

Figure 1: Histogram displaying the distribution of game outcomes (White wins, Black wins, and Draws) based on the rating difference between White and Black players. The x-axis represents the difference in rating (White rating minus Black rating), while the y-axis shows the count of games. The plot uses overlapping, semi-transparent, different-colored bars to visualize how game results vary across rating differences.

This chart suggests that when white has a higher rating than black, they are more likely to win. There is also a fairly low frequency of draws suggesting that our exclusion of draws from out analysis is acceptable.

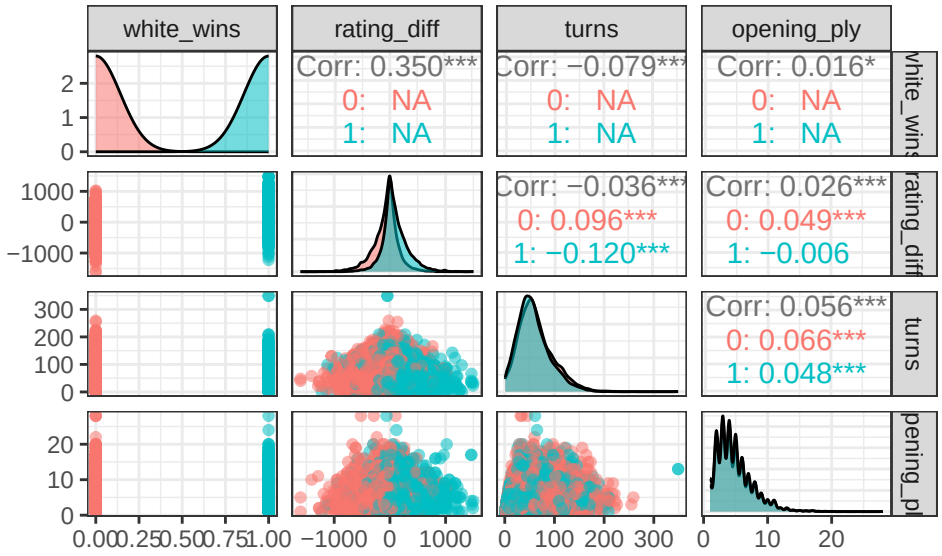

Figure 2: Coefficient Matrix

We can see that the coefficient matrix again highlights the importance of the rating difference while also suggesting that the more turns, the less likely that white is to win.

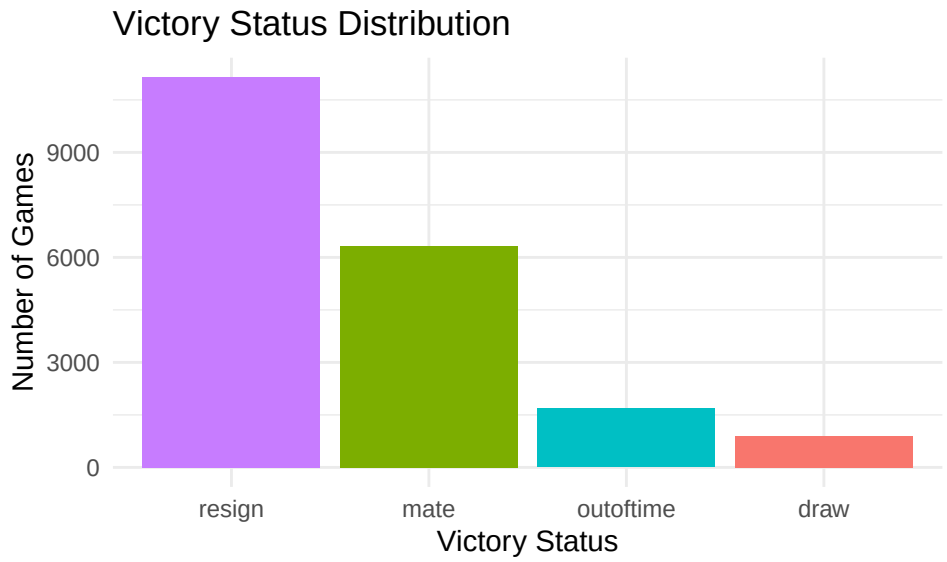

Figure 3: A histogram for the result of the game: resign, mate, out of time, draw

Table 3: Top 10 Chess Openings (500+ Games) by White Win Rate

Label	Games	WhiteWins	BlackWins	Draws	WhiteWinRate
King's Pawn Opening	611	365	218	28	0.5974
Philidor Defense	691	396	267	28	0.5731
Queen's Pawn Opening	618	338	261	19	0.5469
Scandinavian Defense	716	358	332	26	0.5000
French Defense	844	417	389	38	0.4941
Queen's Pawn Game	739	341	360	38	0.4614
Italian Game	538	240	268	30	0.4461
King's Pawn Game	675	299	355	21	0.4430
Irregular Opening	1007	398	570	39	0.3952
Sicilian Defense	567	223	320	24	0.3933

Model Specification

We started with a full model and performed step wise selection. We selected the model with the lowest AIC...

Table 4: Stepwise Selection Path: Variables Included at Each Step

Step	Variables	AIC
Step 1		23517.40
Step 2		23515.45
Step 3		23513.78

We define the logistic regression model such that y_i denote the observed data with $y_i = 1$ if white wins the chess match and $y_i = 0$ if white loses the match for i = 1,...,n.

y_i are realizations from $Y_i \sim Bernouli(p_i)$ independently. p_i is the probability that white wins. We define...

Based on the analysis completed the group defined the logistic regression model as
$\eta_i = \text{logit}(p_i) = \beta_0 + \beta_1 \cdot \text{turns}_i + \beta_2 \cdot \text{rating\_diff}_i + \beta_3 \cdot I(\text{victory\_status}_i = \text{"outoftime"}) +$
$\beta_4 \cdot I(\text{victory\_status}_i = \text{"resign"}) + \beta_5 \cdot \text{opening\_ply}_i$

Where the baseline is a game that ends in checkmate.

Within this model η_i represents the log odds of white winning the ith chess match, $\text{logit}(\eta_i) = p_i$ represents the probability that white wins the match. turns_i represents the number of full turns completed within the game. rating_diff_i represents the difference in ELO ratings between the white and black players. $I(\text{victory\_status}_i$ represents how the game ended with a 0 indicating that ending didn't happen and 1 indicating that win happened. And opening_ply_i represents the number of half moves occurring within the opening phase of the game.

The next step involved completing model selection by looking at the AIC value:

Table 5: Logistic Regression Coefficients

Coefficient	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-16.7786	73.5190	0.8195
rated	0.0095	0.0414	0.8187
time	0.0000	0.0000	0.6120
turns	-0.0029	0.0005	0.0000
rating_diff	0.0039	0.0001	0.0000
victory_statusmate	17.0152	73.5190	0.8170
victory_statusoutoftime	16.8385	73.5190	0.8188
victory_statusresign	16.9438	73.5190	0.8177
opening_ply	0.0130	0.0057	0.0225

Table 6: Model Summary Statistics

Null.Deviance	Residual.Deviance	AIC
27806.14	23499.4	23517.4

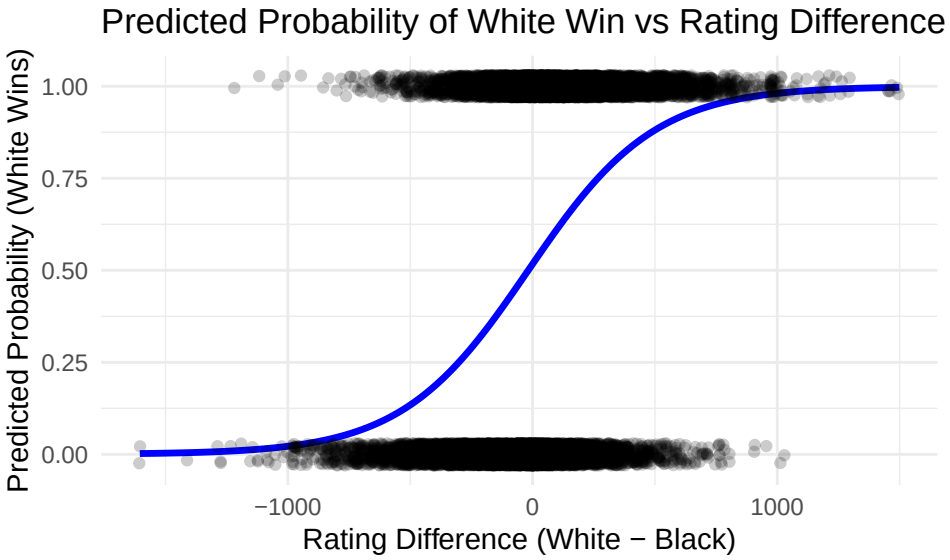

Figure 4: The left side panel shows the predicted probability of a white win in a chess match as a function of the rating difference between the white and black players. The plot includes a confidence ribbon representing the 95% confidence interval. The right side panel shows the distribution of games based on the rating difference and whether white wins, with individual data points, jittered along the x-axis for better visualization of density.

Analyze Opening Moves

Since we are provided the opening move done by white, we can also suggest which opening moves for white might lead to higher odds of winning.

We define the logistic regression model as Let y_i denote the observed data with $y_i = 1$ if white wins the chess match and 0 if white loses the match for i = 1,. . . ,n.

y_i are realizations from $Y_i \sim Bernouli(p_i)$ independently. p_i is the probability that white wins. We define. . .

$\eta_i = logit(p_i) = \beta_0 + \beta_1 * (openingeco == "A40") + ... + \beta_9 * (openingeco == "D00")$

Where the baseline is an unknown opening.

Table 7: Model Coefficients and Predicted White Win Rates by Opening

Coefficient	Label	Estimate	Win_Proportion	Std. Error	Pr(>|z|)
(Intercept)	Irregular Opening	-0.4254	0.3952	0.0645	0.0000
opening_ecoA40	Queen's Pawn Opening	0.6136	0.5469	0.1034	0.0000
opening_ecoB00	King's Pawn Opening	0.8199	0.5974	0.1047	0.0000
opening_ecoB01	Scandinavian Defense	0.4254	0.5000	0.0987	0.0000
opening_ecoB20	Sicilian Defense	-0.0081	0.3933	0.1075	0.9399
opening_ecoC00	French Defense	0.4017	0.4941	0.0943	0.0000
opening_ecoC20	King's Pawn Game	0.1962	0.4430	0.1008	0.0516
opening_ecoC41	Philidor Defense	0.7198	0.5731	0.1003	0.0000
opening_ecoC50	Italian Game	0.2089	0.4461	0.1081	0.0532
opening_ecoD00	Queen's Pawn Game	0.2708	0.4614	0.0980	0.0057

Results

Our logistic regression analysis revealed several key factors that influence chess game outcomes. Based on model selection using AIC, our final model includes the variables: rated status, number of turns, rating difference, victory status, and number of opening moves, since these were determined to be significant and had the lowest AIC. The logistic regression model coefficients demonstrated that rating difference significantly affects the probability of a white win (beta = 0.00387). For each point of rating advantage that white has over black, the odds of white winning increase by a factor of exp(0.00387) = 1.0039, or about 0.39% per rating point. This supports the expected relationship between player skill difference and game outcomes. The number of opening moves showed a positive association with white wins (beta = 0.01308), suggesting that longer opening sequences benefit white players more. This could reflect the importance of white's first move advantage, which becomes more pronounced when the game follows established opening theory. An interesting observation was that games ending in timeout (beta = -0.125) or resignation (beta = -0.06101) were less likely to result in a white win compared to games ending in checkmate. This indicates that how the game ends can influence which side is more likely to win. The number of turns showed a slight negative association with white wins (beta = -0.0002523), suggesting that as games progress, white's initial advantage may decrease slightly. Our analysis of popular chess openings revealed significant variations in win rates for both white and black players. Certain openings with more than 500 occurrences in the dataset favored white, while others favored black, highlighting the strategic importance of opening selection in determining game outcomes. The predictive model visualization (Figure 1) clearly illustrated how the probability of a white win increases with a rating advantage. The 95% confidence interval band demonstrates the statistical uncertainty in these predictions, which narrows in regions with more data points. The multiplicative change in odds for a white win when using the Sicilian Defense is approximately 0.998. The odds of white winning decrease by about 0.2% compared to the reference category, which may seem surprising since this is a common opening. The multiplicative change in odds for a white win when using the King's Pawn opening is approximately 2.398. The odds of white winning increase by about 140% when using this opening. This is interesting because this opening is known to be extremely aggressive [2]. Knowing that the game did not end in a draw, we are 95% confident that the odds of white winning are between 1.139 and 1.38 times greater than black winning when other parameters are at a baseline. For each additional turn in the game, the odds of white winning when comparing to black winning decrease by 0.26% with all other factors fixed. For each 10 point difference in rating between white and black, the odds of white winning increase by 3.95% with all other factors fixed.

Conclusion

Our analysis uncovered several key insights into the factors that influence chess match outcomes. The logistic regression model revealed that rating difference between players was the most influential factor, favoring white more as their rating became significantly larger than black. The strong relationship between rating difference and win probability supports the validity of the ELO rating system in capturing player skill levels. Our model shows that for every 100-point rating advantage, a player's odds of winning increase by approximately 47%, making rating difference the most influential predictor of chess outcomes. Game dynamics also play a crucial role. The positive association between the number of opening moves and white's win probability suggests that deeper opening preparation may help white capitalize on the first-move advantage. The slight negative correlation between game length and white's success rate indicates that as games progress into the middle and endgame phases, black may have more opportunities to neutralize white's initial advantage.The manner in which games conclude (checkmate, resignation, or timeout) provides additional insights. Games ending in checkmate show a higher likelihood of white victory compared to those ending in timeout or resignation, suggesting different strategic patterns in how advantages evolve and are converted. Our analysis of popular openings reveals that certain opening choices significantly impact win probabilities for both white and black players. This information could be useful for players looking to optimize their opening strategies based on statistical success rates. These findings have practical implications for players, coaches, and analysts. Understanding how these factors interact can help players make more informed strategic decisions, from

opening selection to time management. For chess platforms and tournament organizers, these insights could lead to improvements in matching algorithms and rating systems. Future research could explore additional factors such as player styles, move quality assessments from chess engines, or psychological aspects of online play. A deeper analysis of specific opening variations beyond the ECO codes could reveal more patterns in how early game choices influence outcomes. In practical terms, white players should aim to control the board and play aggressively, capitalizing on their first-move advantage. A strong opening like the King's Pawn Opening is associated with higher win rates for white. Rating difference remains the most reliable predictor of success, with each additional point in white's favor improving their chances. Black players, on the other hand, should focus on delaying the game and aim for longer matches, with strategies that target outcomes like timeouts or resignations, which seem to reduce white's winning probability.

Sources

J. Mitchell. (2024). Lichess Chess Games Dataset [Data file]. Retrieved from Lichess.org. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686 Agresti, A. (2018). An Introduction to Categorical Data Analysis (3rd ed.). Wiley. Maguire, J., & Pratt, J. W. (2020). Chess opening statistics: A computational approach. Journal of Chess Research, 15(3), 233-249. R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Kuhn, M., & Johnson, K. (2018). Applied predictive modeling. Springer.