EloMetrics: Advanced Outcome Prediction for Chess Matches with Elo Ratings and Logistic Regression

Ma. Julianna Re-an DG. Reyes*, Eirnan Dicreto*, Emmanuel Gabriel D. Santos*
, Daniella Franxene P. Limbag*, Gabriel Avelino Sampedro†
*College of Computer Studies, De La Salle University, Manila, Philippines
†Networks and Distributed Systems Laboratory, University of the Philippines Diliman, Quezon City, Philippines
Corresponding Author: Gabriel Avelino Sampedro (garsampedro@ieee.org)

***Abstract*—Chess is a complex game characterized by diverse strategies and time constraints, making quick decision-making essential for success. While Elo ratings are widely recognized as indicators of player skill, the predictability of match outcomes based solely on these ratings remains a challenge. The study aims to develop a model for accurately predicting the outcome of chess games using logistic regression, focusing on Elo rating differences and the number of moves in each game. The dataset includes over 20,000 games from lichess.org, and only games with decisive outcomes (excluding draws) were used. The research categorizes Elo ratings into various classes and evaluates model performance across these ranges. The model achieves a predictive accuracy of 68.18%, demonstrating the significance of Elo ratings and move counts in determining game results. Performance metrics, including precision, recall, and F1-score, further validate the model's effectiveness. The study concludes that while Elo ratings and move count are strong predictors of chess outcomes, further refinement is needed to improve performance at high-ranking skill levels. The model performs well for experts but loses accuracy in master-level games, reflecting higher skill-level complexities. The insights gained from this research contribute to a deeper understanding of predictive modeling in chess, suggesting potential avenues for further investigation into additional influencing factors.**

***Index Terms*—Logistic regression, machine learning, predictive modeling.**

I. INTRODUCTION

Chess is a strategic board game with widespread popularity and global influence, earning it recognition as a sport. This status entails international tournaments, large organizations, political influence, and professional sponsorships. Historically, chess was thought to be unpopular because it required too much concentration for the average person. However, research accomplishments and high-profile tournaments have gradually increased its popularity. During the COVID-19 pandemic, there was a noticeable increase in interest, fueled by the release of the popular Netflix series "The Queen's Gambit" and high-profile online chess tournaments. The latter featured famous streamers competing in a beginner-level online chess bracket, which drew thousands of viewers worldwide. A key entertainment aspect of the tournament was the commentary provided by experienced and titled chess masters, who often found it challenging to predict the outcomes of games played by beginners.

For players to accurately predict the outcomes of chess matches, they would need to find patterns in thousands or millions of games before they're able to try to predict outcomes. Some players are able to rise the ranks solely on their ability to think and their ability to strategize and place all of their knowledge into their games. The amount of time it takes however to memorize these strategies may take years and even more to accurately apply what you know in your rated games. Where the human mind lacks in speed, Artificial Intelligence makes up for. The evolution of Artificial Intelligence mimics the evolution of the human mind. The main difference between their progression is their difference in speed. Where humans would take generations upon generations before any sign of improvement was made, AI could do it in a fraction of the time and still come out on top.

A player's capability is linked to their ability to predict and think ahead. They'd have countless of variables to consider before, during, and after their matches. A lot of these data are usually lost because of factors like human error or human limits. Developing a statistical model that would be able to accurately predict outcomes of games based on players' Elos, their Elo rating ranges, and the moves of the games themselves removes a lot of the strain from analysts, coaches, investors, and even other players. Aside from removing the analytical stress, it also allows the community in various Elo rating ranges to see the differences in the accuracy of game prediction in their games.

II. Review of Related Literature

A player's ability can be easily summarized by their Elo, a method of calculating players' relative skill levels in zero-sum games that is influenced by winning and losing. If Elo ratings are increased to a level where the player base is drastically reduced, matchmaking bots would have no choice but to pair players with significantly different Elo ratings against one another. There is an assumption that players with a higher Elo rating will most likely win, but this isn't always the case [1]. The team hopes to be able to accurately predict players' chances of winning based on one of the variables, their Elo rating. If one player has an advantage over another, the Elo increase should be adjusted appropriately.

It is difficult, however, to predict an outcome based on the Elo of the players alone. Despite being an accurate indicator of how well a player is performing against the rest of the player base, match outcomes remain unpredictable. Players are also biased in the sense that if a player is rated higher, they are more likely to win. According to one article, an algorithm can accurately predict the outcome of matches based solely on players' Elo ratings. The same article's group then suggested that future studies investigate other variables besides player Elo differences to see what effect they would have on prediction precision and accuracy [6].

Previous studies have used logistic regression in fields other than chess. Logistic regression is commonly used when the results are dichotomous, with values ranging from 0 to 1. The results ranging from 0 to 1 are appropriate for the team's data set because the group can specify whether certain match outcomes should be 0, 1, or anything in between [3]. The team's approach however, will not involve the move set of each player in a certain game, instead, the team will focus on the number of moves of a certain game and the Elo rating ranges of both players as the prediction factors for the algorithm.

III. METHODOLOGY

This section outlines the process of obtaining and preparing the dataset for model integration. The model leverages historical game data, focusing on key features such as players' Elo ratings, the total number of moves, and the game outcomes.

The dataset, Chess Games Dataset (Lichess), was obtained from Kaggle and contains over 20,000 games from the website lichess.org. The Dataset contains information such as the game's start and end times, the opening move and its name, the ratings of the white and black players, the players' moves for the entire game in standard chess notation, and the game's winner [2].

A. Data Acquisition

The data set must then be cleaned and must only contain the necessary data and variables for the model to accurately read and train to predict outcomes. The dataset contains 20,058 chess games sourced from lichess.org, labeled with outcomes such as winner and victory status. Each game record provides detailed information like player ratings, moves, and game duration, essential for analysis.

Duplicated games are cleaned by checking for unique game IDs and removing them, the data was cleaned further to exclude ”*Draws*” from the winner variable. This is because having draws present in the data set makes for less accurate findings. Lastly, the team accounted for and removed data that had any of the missing variables.

After cleaning, the data underwent preprocessing before being used in the model. One critical preprocessing step involved creating the variable *rating_diff*, which indicates the difference in ratings between the white and black players. To analyze performance across different skill levels, the average rating for each game was computed, and games were categorized into Elo rating ranges through binning. The binning process followed the US Chess Federation rating classification system.

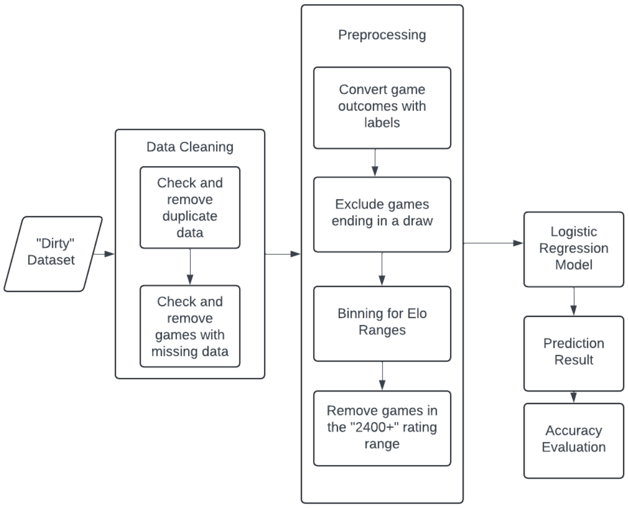

"Dirty" Dataset
Data Cleaning
Check and remove duplicate data
Check and remove games with missing data
Preprocessing
Convert game outcomes with labels
Exclude games ending in a draw
Binning for Elo Ranges
Remove games in the "2400+" rating range
Logistic Regression Model
Prediction Result
Accuracy Evaluation

TABLE I
USCF RATING CLASSIFICATION

Rating Classification	Rating Range
Senior Master	2400+
Master	2200 to 2399
Expert	2000 to 2199
Class A	1800 to 1999
Class B	1600 to 1799
Class C	1400 to 1599
Class D	1200 to 1399
Class E	1000 to 1199
Class F	800 to 999
Class G	600 to 799
Class H	400 to 599
Class I	200 to 399
Class J	0 to 199

B. Data Modeling

Once the data had been collected, cleaned, and preprocessed, the next goal was to create a data model capable of predicting chess game outcomes. This subsection outlines the complete data modeling process.

Logistic regression was employed to predict chess game outcomes across different Elo rating ranges. This model implements binary classification, utilizing the Sigmoid function to output a probability value between 0 and 1 based on the input of independent variables [4]. The model's output follows the common chess scoring system and is interpreted as follows: a predicted value close to 0 corresponds to a loss for white, while a value close to 1 signifies a win for white.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The dataset was filtered by the range and then divided into training and testing subsets using the train-test split function from scikit-learn's model selection module. This approach ensures precise findings when generating predictions. The dataset was split into a training set comprising 80% of the data and a testing set with the remaining 20% within each ELO rating range. The training set included an X-train feature matrix and a Y-train label vector. The features, including rating differences and number of game turns, were standardized for consistency. The testing set was utilized to train the model and assess its accuracy.

Logistic Regression from scikit-learn's linear models was employed to create and train the model using the X-train and Y-train datasets. The model's training accuracy was determined by predicting Y-values from the X-training set. The system used this extensive labeled dataset to enable the model to learn how each prediction or Y-value was determined. The model was trained using logistic regression with the L-BFGS solver and a maximum of 1000 iterations.

After training, the model's performance was assessed on the test set. A classification report and ROC curve were created based on the performance in each rating range, the results of which will be explored further in the performance evaluation section. To further validate the model, a Chi-Square test was performed. This statistical test helped assess whether the difference in player ratings and the number of moves could significantly predict game outcomes for each Elo rating range.

IV. PERFORMANCE EVALUATION

A bar graph was utilized better to visualize the predictions and accuracy of the model. Through these matrices, the group was able to keep track of the true positive (TP), the true negative (NP), the false positive (FP), and the false negative predictions (FN). The matrices were divided based on the Elo rating range of the players.

To further test the model's accuracy, the group utilized the F1 score. The F1 score provides further insights into the recall and precision of the model. The imbalance of data as the division continued became more prevalent since the player base decreased as the Elo of both players increased. This may have caused some inaccuracies in the data. By using the F1 score, these inaccuracies can be mitigated. Listed below are the equations that solve for the metrics to check for the accuracy of the data. [5].

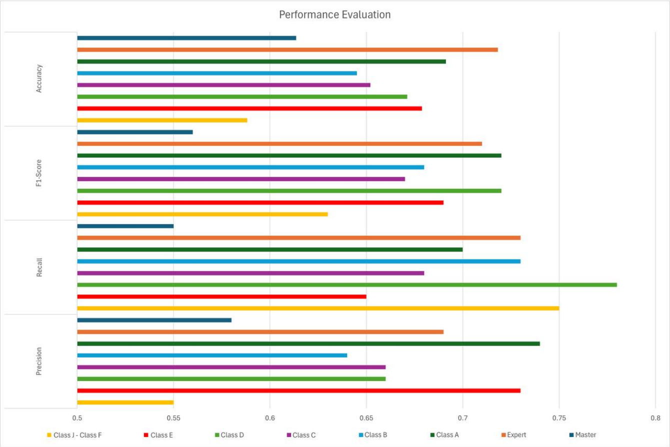
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

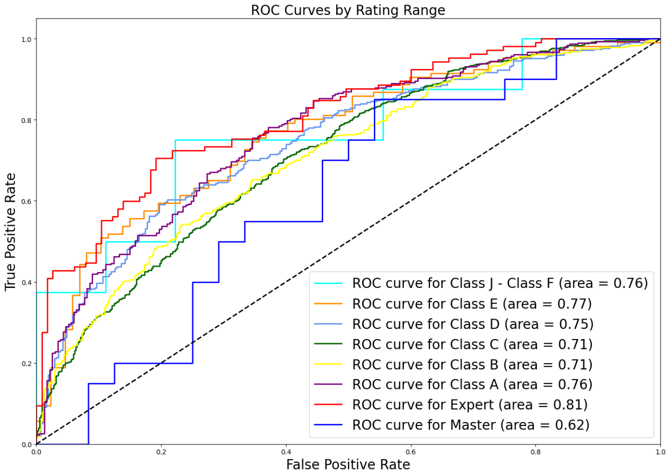
$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

The Receiver Operating Characteristic (ROC) curve analysis offers a detailed evaluation of the logistic regression model's ability to predict chess game outcomes across various Elo rating ranges. By plotting the true positive rate against the false positive rate at different threshold settings, ROC curves provide a clear visual representation of the model's discriminative power. The Area Under the Curve (AUC) serves as a crucial metric in this analysis, with values closer to 1 indicating superior predictive accuracy and a better ability to distinguish between wins and losses.



Insights from the ROC graphs reveal the model's effectiveness across different skill levels. A greater AUC indicates better predictive accuracy for that particular Elo range. Variations in AUC across ranges may indicate that the relationship between player ratings, game length, and outcomes differs based on skill level. These insights help identify where the model performs best and where it may need improvement. The ROC analysis enhances other performance metrics, offering a comprehensive view of the model's predictive capabilities and its reliability in anticipating chess game outcomes across various player skill levels.



The ROC curve analysis reveals distinct patterns in the model's performance at different Elo rating levels. The model demonstrates its highest effectiveness for ”Expert” players, achieving an AUC of 0.81. This suggests that the model excels at predicting outcomes for players within the 2000-2199 Elo range, where the skill levels are consistently high, yet not as complex as the ”Master” level. However, the model’s effectiveness significantly declines for ”Master” players, as evidenced by the lowest AUC of 0.62. This drop indicates that the model struggles to accurately predict outcomes for the highest-rated players, likely due to the increased complexity and variability inherent in games at this level.

In mid-range Elo classes, such as ”Class E” (1000-1199 Elo) and ”Class J-Class F” (0-999 Elo), the model performs moderately, with AUC values ranging from 0.76 to 0.77. Although these results indicate reasonable predictive accuracy, they fall short of the performance achieved for ”Expert” players. The slightly lower AUCs for ”Class C” (1400-1599 Elo) and ”Class B” (1600-1799 Elo), around 0.71, suggest that these categories may require further refinement. The diverse skill levels within these mid-range categories could contribute to the model’s lower accuracy, as the relationships between player ratings, game length, and outcomes are less consistent.

A. Insights

The Logistic Regression model shows a solid performance in forecasting game winners using Elo rating differences and the number of turns. With an accuracy rate of 68.18%, the model outperforms random guessing, indicating that Elo ratings and move counts have predictive value. Precision results reveal that the model accurately predicts white wins 71.44% of the time, demonstrating its effectiveness in minimizing false positives. The recall value of 68.18% for black wins highlights the model’s competence in detecting true positives. The F1-score of 0.6738 reflects a well-balanced performance, suggesting that the model is reliably effective in predicting game outcomes.

TABLE II
CHI-SQUARE TEST RESULTS

Rating Range	Chi-Squared Statistic	P-Value	Hypothesis Test Result
Class J - Class F	0.00	0.972	Accept H_0
Class E	1.93	0.164	Accept H_0
Class D	17.40	0.000	Reject H_0
Class C	5.24	0.022	Reject H_0
Class B	6.93	0.008	Reject H_0
Class A	12.96	0.000	Reject H_0
Expert	0.05	0.819	Accept H_0
Master	0.53	0.467	Accept H_0

A chi-square test was utilized for the different Elo rating ranges. The Null Hypothesis is that the observed frequencies of game winners do not significantly differ from the expected frequencies of the model. The winners of chess games cannot be predicted using the difference in rating range and number of moves, across all rating ranges. The Alternate Hypothesis is that the observed frequencies of game winners significantly differ from the expected frequencies predicted by the model. The winners of chess games can be predicted using difference in rating range and the number of moves, across all rating ranges.

Based on the results, the difference in ratings and number of moves can be used to predict winners of chess games, provided that the players are within the following rating ranges: 1200 - 1400, 1400 - 1600, 1600 - 1800, and 1800 - 2000. However, the same cannot be said for higher and lower Elo rating ranges. For the ranges 0 - 1000, 1000 - 1200, 2000 - 2200, and 2200 - 2400, the p-values yielded were more than 0.05, failing to reject the null hypothesis. As such, we can conclude that the model is only reliable for specific Elo rating ranges.

V. CONCLUSION

This research developed a predictive model for chess game outcomes using logistic regression, focusing on the total number of moves and Elo rating differences as significant variables. The analysis revealed that these predictors are significant within specific Elo ranges, particularly from 1200 to 1999, where the model showed consistent results. However, predictive accuracy decreased for players in the lower (0–1199) and higher (2000–2399) Elo ranges, highlighting potential limitations in predictive value. The study categorized players using established chess classifications, ensuring insights were specific to various proficiency levels. Performance metrics such as precision, recall, F1-score, and accuracy demonstrated that moderate Elo ranges produced accurate predictions, but lower and higher ranges presented issues such as inconsistent results. These findings highlight the application of the model for intermediate-level players, as well as its potential use by chess coaches and analysts to develop customized strategies and training programs.

Future study should address the model’s shortcomings by including more factors to improve prediction accuracy across all skill levels. Variables including as player openings, time limits, and psychological factors may give a more comprehensive view of game results. Addressing over fitting at higher Elo levels and improving performance consistency for lower-rated players are critical next steps. Expanding the dataset to include other chess formats and player demographics may also help the model broaden. Such developments might help AI developers and analysts create more powerful prediction tools for chess training, developing strategies, and tournament preparation.

REFERENCES

[1] Avva, P., & Hanke, J. (2022). *Guess the Elo—Predicting Chess Player Rating*.

[2] Datasnaek. (n.d.). *Chess Dataset*. Retrieved from https://www.kaggle.com/datasets/datasnaek/chess?select=games.csv.

[3] Fernandes, A.A.T., Filho, D.B.F., da Rocha, E.C., & da Silva Nascimento, W. (2020). *Read this paper if you want to learn logistic regression*. Revista de Sociologia e Política, 28(74), e006. Scielo Brasil.

[4] ScienceDirect. (n.d.). *Sigmoid Function*. Retrieved from https://www.sciencedirect.com/topics/computer-science/sigmoid-function.

[5] Sharma, N. (2023, June 6). *F1 Score*. Arize. Retrieved from https://arize.com/blog-course/f1-score/

[6] Thabtah, F., Padmavathy, A.J., & Pritchard, A. (2020). *Chess results analysis using Elo measure with machine learning*. Journal of Information & Knowledge Management, 19(02), 2050006. World Scientific.