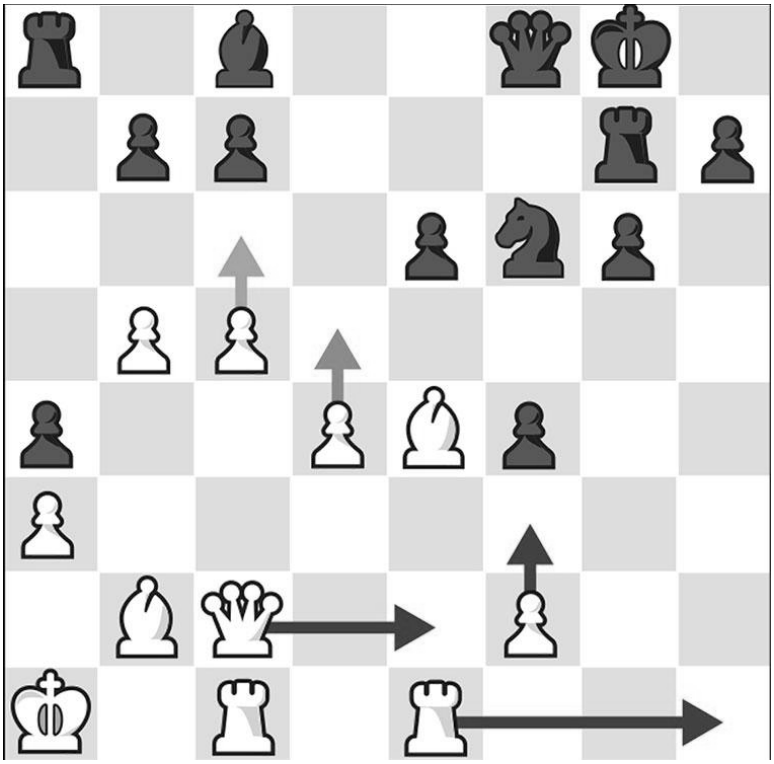
Breaking Down the Grandmaster's Strategy: An Analysis of the Game of Chess

Arunabh Sarkar, Nuray Ozden, Danielle Xu

ORIE 3120: Practical Tools for Operations Research, Machine Learning, and Data Science



I. Introduction

The game of chess traces its roots back to the 6th century in ancient India and later evolved towards the start of the 16th century with the development of chess theory and advanced chess strategy. For our analysis, it is essential to define a few terms and understand a couple of nuances that are key to the game of chess.

Chess players are rated based on their ability through a ranking system known as the "Elo Scale". This system allows players to identify their level of expertise and match against other players that are of similar caliber. These ratings range from scales around 0-1200 (which is considered novice division) all the way to 2700+ which is considered a super grandmaster. To be ranked amongst the greatest is extremely rare.

Chess matches also have alternative timing systems, which are defined as "increment codes." When matches are played, they can be played in a variety of timestamps. A general match will have the increment code of "10+0," which is a match where each player has 10 minutes on their clock, resulting in a 20 minutes fixed match. Other matches, such as a "0+5," are matches where players begin with 0 minutes on the clock. Every time a move is played, 5 seconds are added to their clock.

Our goal for this project was to hypothesize three questions and utilize machine learning models, visualizations, and data-driven algorithms to find answers to our questions and uncover the complexity that belies the game of chess.

II. Questions

1. Can the length of a game, in terms of moves played, be predicted based on the difference between players' ratings, the number of moves in an opening, and the time increment?
2. Based on previous chess games, can we make predictions about the number of moves and types of moves for future games?
3. How important are factors such as rating, number of turns, opening plays, and length of a game in regards to the outcome of a game?

III. Dataset Analysis and Preface

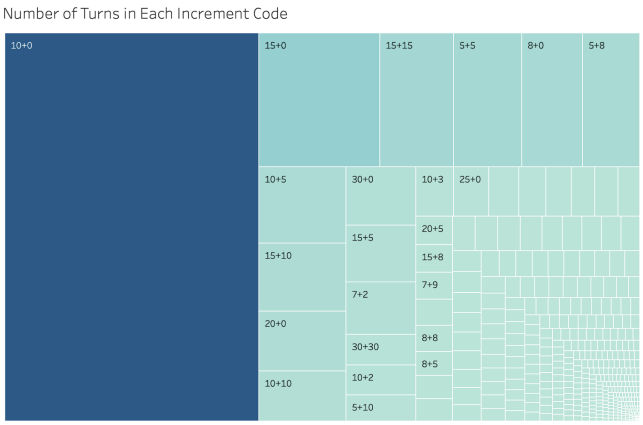
For the analysis and visualizations, we decided to use a dataset of 20,000 chess games from Kaggle collected from a chess-playing site Lichess.org. Variables in the dataset include the ratings of the two players, the winner, the opening move and play, whether the games are "rated", the date and time of the match's start and end, and more. The dataset had some known issues with the time variables. Multiple matches showed errors of having the same start and end time matches and a few also had incorrect start and end times that conflicted with the time limit set by the increment. These matches were omitted from our models and we cleaned our dataset before using them to fit our machine learning models.

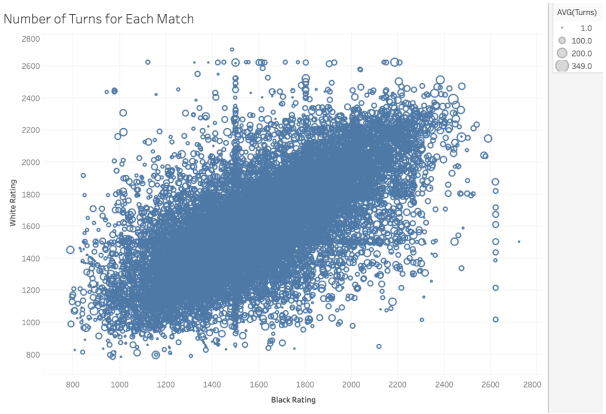
IV. Research Methods and Results

Linear Regression

1. *Preliminary Data Analysis*

The first visualization that we generated was to view how the time increment of a match affected the number of turns in a match. What we identified was that matches with a larger fixed time had more moves than matches with a low fixed time. The second visualization that we generated was to view how the different matchups in a game of chess would affect the number of moves in a given match. It was clear to see that games with a large amount of large mismatches had a lower number of average turns than games where opponents were evenly matched.





2. *Multiple Linear Regression*.

The goal of linear regression is to fit a linear model to a set of data such that the residual values are minimized. Residual values are the distance between the prediction for a certain Y variable (also known and referenced as Yhat) and the known Y value. By fitting a linear model to minimize the total distance from the residuals, we can calculate an R-squared value that tells us how well our linear model fits the data. The question that we wanted to answer using linear regression was whether or not we could accurately predict the length of a match (measured by moves) in matches where we knew the difference in rating between both players, the number of moves in an opening play, and the time increment of the match. This question was interesting because, in a match, the player is aware of the covariates since they are displayed on the screen and it is in their interest to understand how many moves the match might have so they can organize strategies based on knowing how many moves they may have left in the game. This is true for a game against any given player in any given setting.

In our dataset, we identified multiple variables that could affect the number of moves. To fit our model, we added variables that counted: the difference in

OLS Regression Results

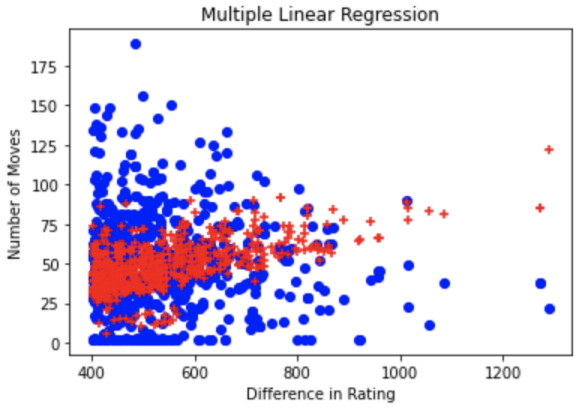
Dep. Variable:	turns	**R-squared (uncentered):**	0.717
Model:	OLS	**Adj. R-squared (uncentered):**	0.715
Method:	Least Squares	**F-statistic:**	427.2
Date:	Mon, 16 May 2022	**Prob (F-statistic):**	2.74e-183
Time:	00:56:39	**Log-Likelihood:**	-3314.4
No. Observations:	679	**AIC:**	6637.
Df Residuals:	675	**BIC:**	6655.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rating_difference	0.0621	0.004	14.795	0.000	0.054	0.070
opening_ply	3.8619	0.426	9.060	0.000	3.025	4.699
increment_code_min	-0.1361	0.047	-2.892	0.004	-0.229	-0.044
increment_code_sec	0.0403	0.061	0.664	0.507	-0.079	0.160

rating between players. The number of moves in the match's opening play, the number of minutes in the match, and the number of seconds added to the player's clock as a delay from the increment code. The resulting model, as seen above, met our standard of being well-fitted (having an R-squared > 0.7). The linear model fits the data at an R-squared value of 0.717, which is a relatively well fit. This means that approximately 71.7% of the change in turns is explained by the covariates.

3. ***Modeling Analysis***.

The resulting model for the multiple linear regression is below. The red values were our Yhat predicted scores and the blue values were our known Y values from the dataset. The visualization is a 2-D representation of the multiple linear regression.

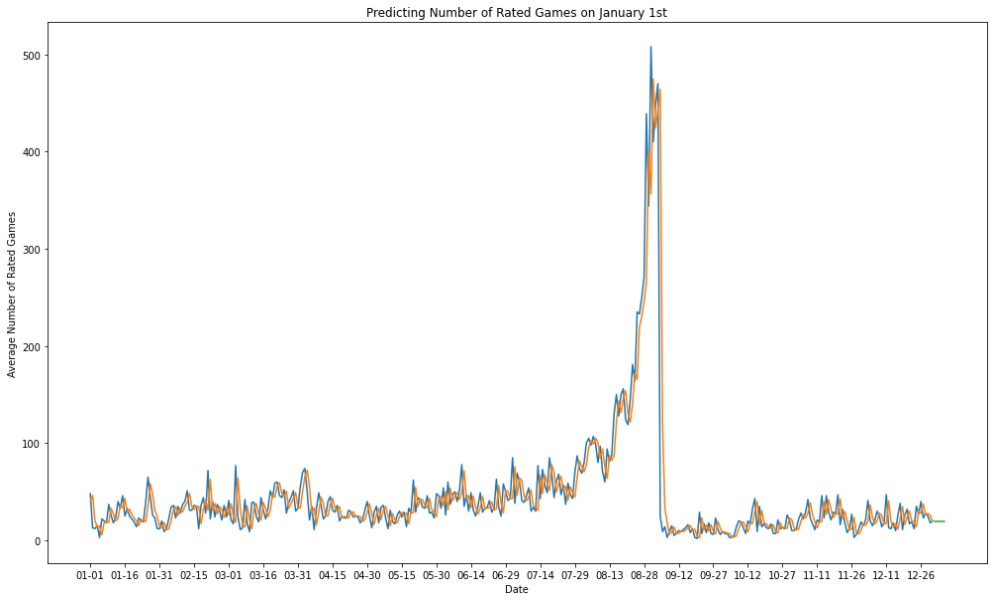


The linear model fits the data well due to the selected covariates. The difference in rating of players would mean that we could predict that heavily mismatched players would have relatively short games since one is more experienced. In matches where players are using long opening moves, there would be a fixed lower bound for the number of moves. Finally, matches with large fixed time increments would have more moves since there is a larger amount of time to strategize. There was no real effect to the time delay altering the moves played, but the minutes of fixed time had an effect on the regression nevertheless.

Time Series Forecasting

1.1 Forecasting Number of Rated Games

One descriptor given in the data set is the True/False "Rated", which describes whether the chess game is an official chess game. If the game is "rated", then the game's score impacts that user's rating on Lichess.org, and if it is "unrated," then the outcome has no impact on either player's rating. The rated games are typically more competitive whereas the unrated games are more casual and fun. Thus, we



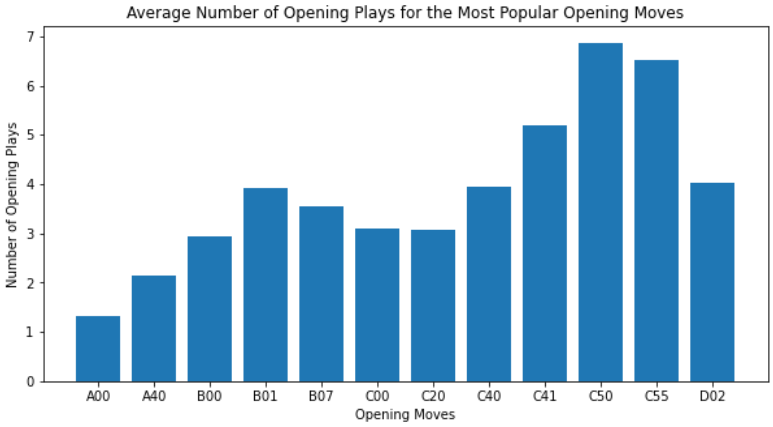
have created the forecasting plot below to predict the number of rated games on January 1st of 2018 for the more seasoned players. The data spans from 2013 to 2017 and includes multiple games for every day those 5 years, making the raw data below an accurate depiction of the average number of rated games played every day of the year (there aren't some days that have drastically more games than others). Simple Exponential Smoothing is the type of forecasting because there is no consistent trend or seasonality in the data. Lastly, there is a large spike around the late August mark due to a yearly tournament played in the early August/late September period.

1.2 Analysis

According to the plot, the predicted number of rated games on January 1st is 20. Overall, the raw data was not the best to work with due to the lack of trend and apparent variability day-to-day. There were also limited quantitative variables that changed over time to be examined. However, this method of forecasting is useful in determining the approximate number of rated games at the beginning of the new year for the more high-level Lichess players, so they have an understanding of the number of opportunities to play on that date. So if you're a player in Lichess.org playing in a rated game, you now have an idea of the number of games played on January 1st. This technique can be applied to predict future dates as well.

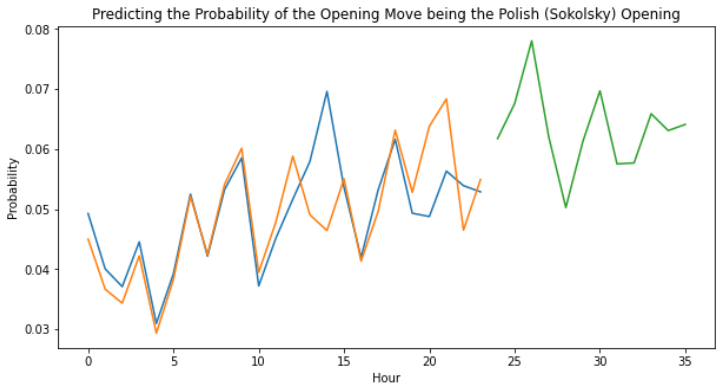
2.1 Looking at the Number of Opening Plays

Another predictive visual presentation that could aid players is the average number of opening plays for the ten most common opening moves. The dataset includes 66 different popular opening moves that the white player chooses to perform, and we have presented the 12 most popular moves (given in standardized code). The average number of moves in the opening play is presented with these moves, which is the number of moves in the opening phase of the game. While this is not a forecasting visualization since time is not a factor, these values are important because they can help black players self-predict how long the opening phase will be given the white player's opening move.



3.1 Forecasting Probability of a Specific First Move

While there was not much apparent seasonality for any of the data's variables, we noticed the probability of the white player opening with the most popular opening move, the Polish (Sokolsky) Opening, fluctuates throughout the day with a trend. Thus, the Holt-Winters method can be used to predict the probability the white player uses the Polish Opening for the first 12 hours of the day. As shown by the raw



data, the repeating peaks and troughs demonstrate the best period length to be 12, as to repeat the approximate cycle every 2 hours. While this period length is not perfect (does not follow the data around hour 13), it is quite accurate everywhere else. (Note: the hours after 24 represent the forecasted hours, between 12 am and 6 am).

3.2 Analysis

The time with the maximum probability of the white payer opening with the Polish Opening of 8% is at 2 am, and the lowest probability of 5% is reached at 5 am. It continues to fluctuate to a lesser degree for the next few hours. 5% may not seem substantial, but given there are 66 different opening moves listed in the data set this is a relatively large probability for a single move. This predictive information is useful for black players who want the probability that the white player will open with the most popular player so they can plan their moves accordingly.

Overall, this dataset is not the best for time-series forecasting due to the lack of seasonality and trend across a day or year. If it had included information about the number of users who joined the site each day, games in specific tournaments, or tracking individual players in their trajectory in the site, that would likely give fruitful forecasting results.

Logistic Regression

1. *Multiple Logistic Regression*

We also wanted to look at some of the different factors that affected the outcome of a chess game, which is important in helping players formulate their plan of action in a chess match. For a hypothetical example, if the number of opening plays was not significant in determining the outcome of a game, then that is one less factor the player would have to consider. To answer this question, multiple logistic regression was used in order to analyze the relationship between a white or black win and various other factors, explained more deeply below.

Logistic regression is a statistical program used to predict dichotomous binary outcomes, such as a win or loss. It is often used to portray the relationship between a categorical dependent variable and multiple independent variables. The independent variables used in the regression were the number of turns, number of opening plays, length of a game (TimeFrac), white ratings, and black ratings. TimeFrac was calculated by dividing the length of a game out of a fraction of a day; i.e. 1 day and 6 hours would equal 1.25.

These variables were chosen because they are continuous and discrete, and have little multicollinearity between each other. The dependent variable was the probability of a white win. Dummy variables were created: cases, where white won, were set to 1, and cases, when white lost, were set to 0. These values were stored in a column called "win." Cases, where white and black players drew, were omitted.

Logit Regression Results

Dep. Variable:	win	**No. Observations:**	10946
Model:	Logit	**Df Residuals:**	10940
Method:	MLE	**Df Model:**	5
Date:	Mon, 16 May 2022	**Pseudo R-squ.:**	0.1115
Time:	16:27:56	**Log-Likelihood:**	-6728.3
converged:	True	**LL-Null:**	-7572.9
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3200	0.132	2.425	0.015	0.061	0.579
turns	-0.0021	0.001	-3.334	0.001	-0.003	-0.001
opening_ply	0.0171	0.008	2.274	0.023	0.002	0.032
TimeFrac	-0.3285	0.730	-0.450	0.653	-1.759	1.102
white_rating	0.0038	0.000	31.873	0.000	0.004	0.004
black_rating	-0.0039	0.000	-32.646	0.000	-0.004	-0.004

2. ***Analysis***.

Looking at the model, the p-values that were < 0.05 suggested that all variables except TimeFrac were statistically significant, illustrating that the null hypothesis should be rejected. However, the low R-squared value of 0.1115 and small coefficients for each variable imply that the data is noisy and has high variability, though a positive coefficient indicates that the probability of a white win increases as that variable increases (for example, a white rating compared to its black counterpart will usually result in a white win). These outcomes are to be expected when dealing with a large dataset with broad deviations between each chess match, and since players can often be unpredictable, the data reflects the irregularity as such. Another regression model was also run for black wins and losses, with notably similar results, so it was omitted for sake of brevity.

A classification report (shown to the right) was also made in order to understand the reliability of the regression. Generally speaking, it appears that the predictions were accurate roughly 50-60% of the time. Nonetheless, the significant variables maintain their interpretation that player rating, number of turns, and number of opening plays have an effect on the winner of a match.

	precision	recall	f1-score	support
0.0	0.66	0.57	0.61	1587
1.0	0.64	0.73	0.68	1697
accuracy			0.65	3284
macro avg	0.65	0.65	0.65	3284
weighted avg	0.65	0.65	0.65	3284

Something to note is that the regression makes clear that the rating of a player is statistically significant, but makes no specification between different rating classes—like whether or not rating mattered more for lower-rated players playing against each other compared to higher-rated players. Another thing to consider is that this regression makes no distinction between the different types of chess games, such as Blitz, Bullet, or normal chess—chess variations that depend greatly on the amount of time a player has left to make a move. This would obviously affect the TimeFrac variable, which was labeled as not significant earlier. As the dataset comprises over 20,000 matches, the number of normal chess games exceeds that of the other specialized chess variants but looking into how the rating and length of matches based on categorizing each type of chess game would provide an interesting future analysis.

V. Conclusion

Our research goal was to hypothesize and answer three key research questions using machine learning models in order to help chess players whether that be in-game or out of game. Using linear regression, logistic regression, and time-series forecasting, we were able to manipulate our data, run a series of machine learning algorithms, and make statistical predictions on the game of chess for any given player in any given match. These questions provide information about what strategy is most likely to appear at a certain time of day, which could help players better manage their time and help to improve their overall approach and strategy. Even if knowing the information will not have a huge impact on the game, understanding the correlations and phenomena is nonetheless very interesting and could be useful for accurately scheduling tournaments or advising users on which types of factors are better to focus on depending on their time commitment. With these models, chess players on Lichess.org can make more informed, calculated, and data-driven decisions that will benefit them in their chess careers.

References

J, M. (2017, September 4). *Chess game dataset (lichess)*. Kaggle. Retrieved May 16, 2022, from https://www.kaggle.com/datasets/datasnaek/chess

The U.S. Chess Trust. Chess History – The U.S. Chess Trust. (n.d.). Retrieved May 16, 2022, from http://www.uschesstrust.org/chess-history/

University of Maryland. (n.d.). *History of chess*. UMD, Omeka RSS. Retrieved May 16, 2022, from https://libapp.shadygrove.umd.edu/omeka/exhibits/show/international-games-day/history-of-chess