The Significance of the First Move Advantage in Chess

Labib Chowdhury

21/12/2020

Supporting Code: https://github.com/labib-c/first-move-advantage-analysis

Abstract

The purpose of this analysis is to determine whether the White player in Chess has a significant advantage due to having the first move, using the LiChess dataset from Kaggle (https://www.kaggle.com/datasnaek/chess). This was done using a logistic regression model through a binary outcome variable representing whether player White won the game. It was found that White does not have a significant advantage due to having the first move, and the first move advantage becomes less significant the longer a Chess game goes.

Keywords

Chess, First Move, White, Black, Logistic Regression, Experiment, Two Sided t-test

Introduction

In the world of Chess, one question which remains open is whether or not White has a significant advantage due to having the first move. The player who controls the White pieces begins the game on the attack while the player who controls the Black pieces must defend, which is where this apparent advantage originates. If both players play identically and symmetrically, then White will win a significant majority of the time. The issue with this is that players do not play identically or symmetrically most of the time, in fact White wins only 37% of the time compared to Black's 28%, according to the chessgames.com database of chess games (Chess Statistics). This paper aims to analyze whether White has a significant advantage in Chess due to having the first move, and this will be done through considering multiple factors including rating level, game type, and first move.

Chess remains as one of the most popular games of all time, claiming 600 million fans worldwide (Cowen, 2018) hence the significance of Chess cannot be understated. Analyzing whether White has a definite advantage over Black will help novice and expert players understand the game better and modify the strategies utilized by every player.

To identify whether White has a significant advantage over Black in Chess, a logistic regression model will be applied over the LiChess Chess Game Dataset (Jolly, 2017). Logistic regression is best used to describe the relationships between one dependent binary outcome variable against one or more independent predictor variables. The logistic regression model will also provide estimates identifying how a certain predictor affects the outcome in an intuitive manner.

An outcome variable of the form `white_won` will be used in the logistic model, where 1 represents games where the White player won and 0 represents games where White lost or tied. The Methodology section (Section 2) will dive deeper into which predictors were chosen and how the logistic regression model was developed. Results of the model can be seen in Section 3, Results, and further discussion on the outcome of the model will be highlighted in Section 4, Discussion.

Methodology

This section will discuss the dataset further as well as the development and selection of the logistic model.

Data

The data used in this analysis was retrieved from https://www.kaggle.com/datasnaek/chess. This dataset consists of 20,000 games of chess collected from LiChess.org. The data was assembled using the LiChess API (https://github.com/ornicar/lila), and consists of the most recent games taken from the top 100 teams on LiChess in 2017 (Jolly, 2017). The dataset contains 16 columns, including `white_rating`, `black_rating`, `moves`, `opening_name`, and `winner`, along with several other useful features. Additionally, this data is considered an experiment as LiChess randomly assigns players as Black or White. The population this dataset attempts to measure would be all players of online Chess, the frame being all LiChess members, and the sample would be the 20,000 most recent games taken from the top 100 teams on LiChess.

A table summarizing the characteristics of this dataset can be seen in Table 1 below (Revelle, 2020), along with a plot displaying the frequency of a White win, a Black win, and a draw (Figure 1).

Table 1: Characteristics of Chess Dataset

	n	mean	sd	min	max
id*	20058	9.549788e+03	5.525404e+03	1.000000e+00	1.911300e+04
rated	20058	NaN	NA	Inf	-Inf
created_at	20058	1.483617e+12	2.850151e+10	1.376772e+12	1.504493e+12
last_move_at	20058	1.483618e+12	2.850140e+10	1.376772e+12	1.504494e+12
turns	20058	6.046600e+01	3.357058e+01	1.000000e+00	3.490000e+02
victory_status*	20058	3.150065e+00	1.014535e+00	1.000000e+00	4.000000e+00
winner*	20058	2.044571e+00	9.750375e-01	1.000000e+00	3.000000e+00
increment_code*	20058	1.258813e+02	1.248025e+02	1.000000e+00	4.000000e+02
white_id*	20058	4.702474e+03	2.754084e+03	1.000000e+00	9.438000e+03
white_rating	20058	1.596632e+03	2.912534e+02	7.840000e+02	2.700000e+03
black_id*	20058	4.664879e+03	2.732066e+03	1.000000e+00	9.331000e+03
black_rating	20058	1.588832e+03	2.910361e+02	7.890000e+02	2.723000e+03
moves*	20058	9.493483e+03	5.459965e+03	1.000000e+00	1.892000e+04
opening_eco*	20058	1.412449e+02	8.535682e+01	1.000000e+00	3.650000e+02
opening_name*	20058	7.738153e+02	4.105686e+02	1.000000e+00	1.477000e+03
opening_ply	20058	4.816981e+00	2.797152e+00	1.000000e+00	2.800000e+01

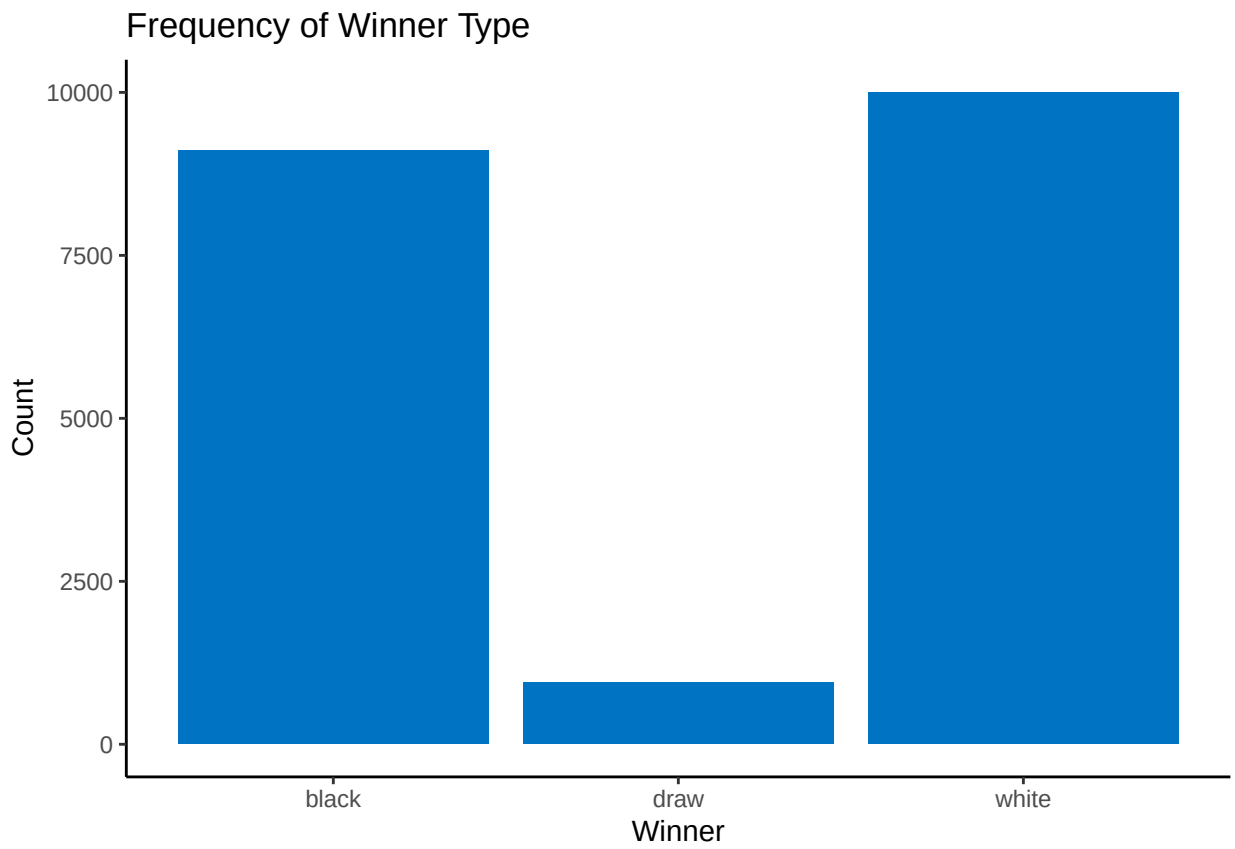

Figure 1: Frequency of Winners

To create a logistic model to analyze the relationship between a player controlling the White pieces and winning games, a binary outcome variable denoted as `white_won` was created where if player White won, then this variable would be represented as 1. Otherwise, if Black won or the game was a draw, the variable `white_won` would be represented as 0.

Two additional variables were also created to assist the analysis. A continuous variable `diff_rating` was created to represent the difference in player rating between White and Black. A variable named `game_type` was created to replace `increment_code` in the raw dataset. This variable represents the game type as an integer, where game type constitutes the time length of the game.

A table summarizing the characteristics of the cleaned data can be seen in Table 2.

Table 2: Characteristics of Cleaned Chess Dataset

	n	mean	sd	min	max
X1	20058	10029.500000	5790.3901855	1	20058
rated	20058	NaN	NA	Inf	-Inf
turns	20058	60.465999	33.5705848	1	349
victory_status*	20058	3.150065	1.0145353	1	4
winner*	20058	2.044571	0.9750375	1	3
white_rating	20058	1596.631868	291.2533757	784	2700
black_rating	20058	1588.831987	291.0361260	789	2723
opening_eco*	20058	141.244890	85.3568191	1	365

	n	mean	sd	min	max
white_won	20058	0.498604	0.5000105	0	1
diff_rating	20058	7.799880	249.0366667	-1605	1499
game_type	20058	13.824110	17.1601786	0	180

Model

The selection of predictors for the logistic model was chosen through backward step-wise variable selection. A full model comprised of all predictors was created and then through backward step-wise variable selection, predictors were dropped based on whether the new model reduced the Bayesian Information Criterion (BIC). The BIC is a criterion which penalizes models containing too many predictors (The Methodology Center). The full model and step-wise model were evaluated by comparing their AUC and ROC Curve, using the `pROC` package in R (Robin et al., 2011), seen in Figure 2.

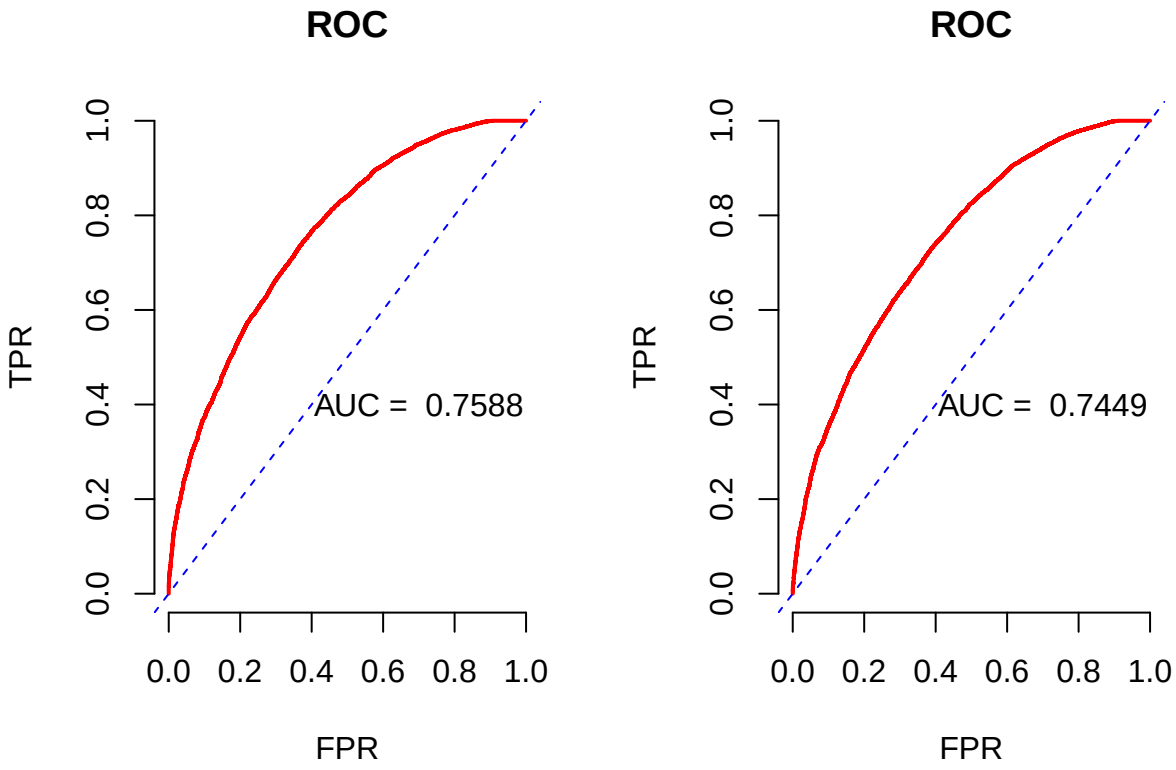

Figure 2: ROC Curve of Full Model and Reduced Model

It is notable that the reduced model has a lower AUC which implies that it is a worse model compared to the full model. Despite this, the Akaike Information Criterion (AIC) is lower, which suggests a better model fit (The Methodology Center). Thus, a model using the chosen significant predictors from the reduced model along with predictors from the full model that were deemed useful for the analysis will be used. The final model consists of predictors `turns`, `diff_rating`, and `game_type` to predict the outcome variable `white_won`. Therefore, we are modeling: (Eq. 1)

$$\log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i$$

where

- p = the probability of player White winning the chess game.
- K = the number of predictors, as well as the number of levels in each factored predictor.
- x_i = the value of the associated predictor.
- β_0 = the intercept estimate.
- β_1 = the change in log odds for every unit increased in the associated predictor.

A second model will be created to assess the relation between the first move made and the probability of White winning the game, where the outcome variable is `white_won` and the predictor is `opening_eco`. The variable `opening_eco` refers to a code representing the first move, where each code's corresponding move name can be found here: https://www.365chess.com/eco.php. This model produces a high AIC compared to the first model suggesting it is a poor model fit, but it will be used to assess the variable of interest `opening_eco`.

Results

The summary of the logistic model can be seen in Table 3.

Table 3: Summary of Logistic Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.2693184	0.0346774	7.766399	0.0000000
turns	-0.0046287	0.0004551	-10.171184	0.0000000
diff_rating	0.0036141	0.0000800	45.185224	0.0000000
game_type	-0.0013123	0.0009305	-1.410289	0.1584544

The table displaying the top and bottom coefficients of the logistic model using only the `opening_eco` variable as a predictor can be seen in Tables 4 and 5.

Table 4: Top Coefficients of Opening Moves

names	x
opening_ecoA67	14.99143
opening_ecoB74	14.99143
opening_ecoD47	14.99143
opening_ecoD19	14.99143
opening_ecoE95	14.99143
opening_ecoD74	14.99143

Table 5: Bottom Coefficients of Opening Moves

names	x
opening_ecoA71	-14.1407
opening_ecoA89	-14.1407

names	x
opening_ecoD58	-14.1407
opening_ecoA65	-14.1407
opening_ecoA24	-14.1407
opening_ecoE72	-14.1407

A two sided t-test was performed along with the logistic model, where the results can be seen in Table 6.

Table 6: Two Sided t-test Results

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.498604	-0.3953986	0.692553	20057	0.491684	0.5055241	One Sample t-test	two.sided

Discussion

Summary

In the previous sections, the raw data was cleaned to produce predictors used in the logistic model. The model, created through backward step-wise variable selection and identifying other useful predictors, attempts to model the probability that `white_won` using predictors `turns`, `diff_rating`, and `game_type`. A second model created solely to understand the relationship between `white_won` and `opening_eco` was also developed. Finally, a two-sided t-test was performed on the mean of `white_won` to identify if this was a significant result.

Conclusions

A two-sided t-test performed using the null hypothesis of "true mean is equal to 0.5" and alternate hypothesis of "true mean is not equal to 0.5" resulted in a p-value of 0.6926. As this p-value is greater than our α value of 0.05, we fail to reject the null hypothesis. This implies that we do not have evidence that player White would win Chess games greater or less than 50% of the time. Using this, the conclusion drawn is that White does not have a significant advantage due to having the first move. The t-test is visualized in Figure 3 using the R package `gginference` (Charalampos and Kleanthis, 2020).

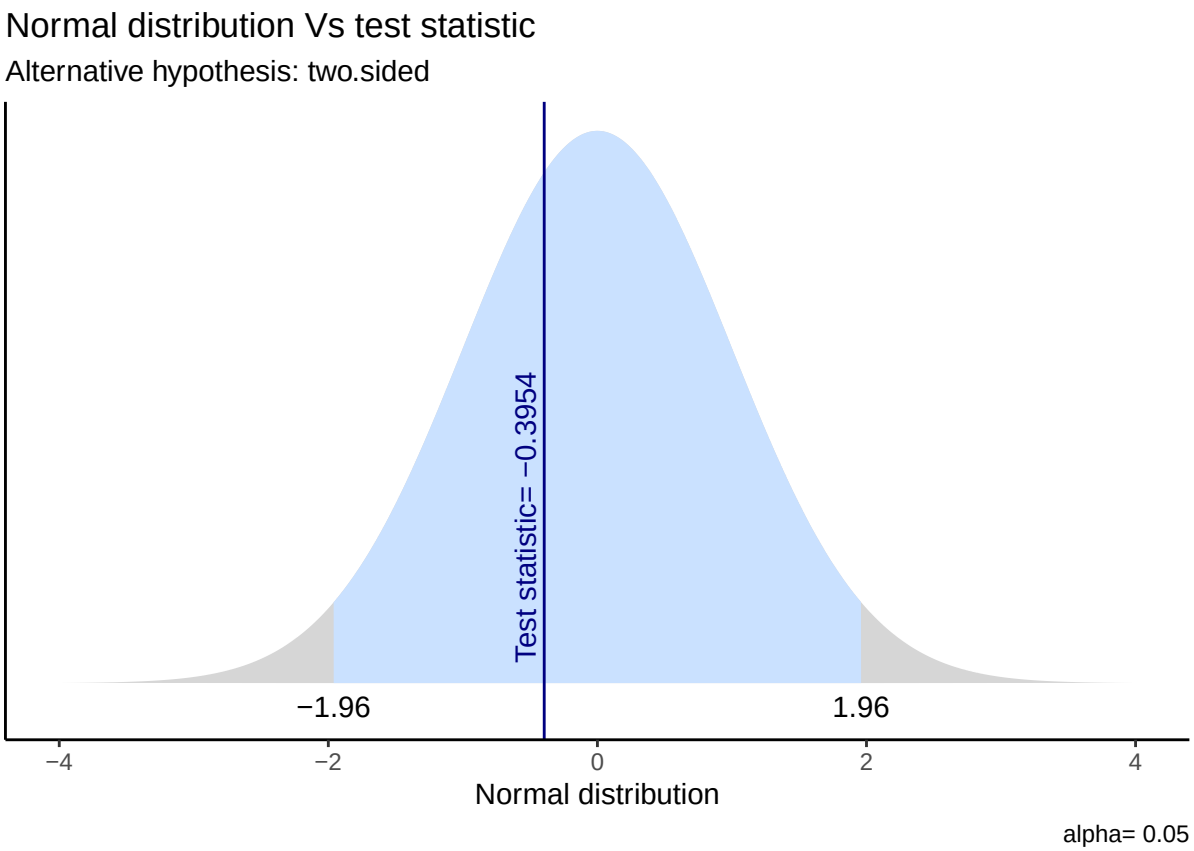

Figure 3: t-test Visualization

The coefficients of the first logistic model is seen in Table 3, and describes how the probability of White winning the chess game is affected by each predictor. It is notable that the coefficient for `turns` is -0.00046287 which explains the log odds of White winning the game decreases with every unit of increase in the number of turns. This implies that the longer the chess game goes, White loses their advantage from having the first move by a factor of -0.00046287. Similarly, the coefficient for `game_type` is -0.0013123. As `game_type` refers to the max time length of the game, the log odds of White winning the chess game decreases with every increased unit in `game_type` by a factor of -0.0013123. Combining these two results, it is evident that the significance of the first-move advantage decreases with longer games. One reason for this is that human players are prone to "blunders", where the odds of winning consistently shifts between players as more blunders are made by each player.

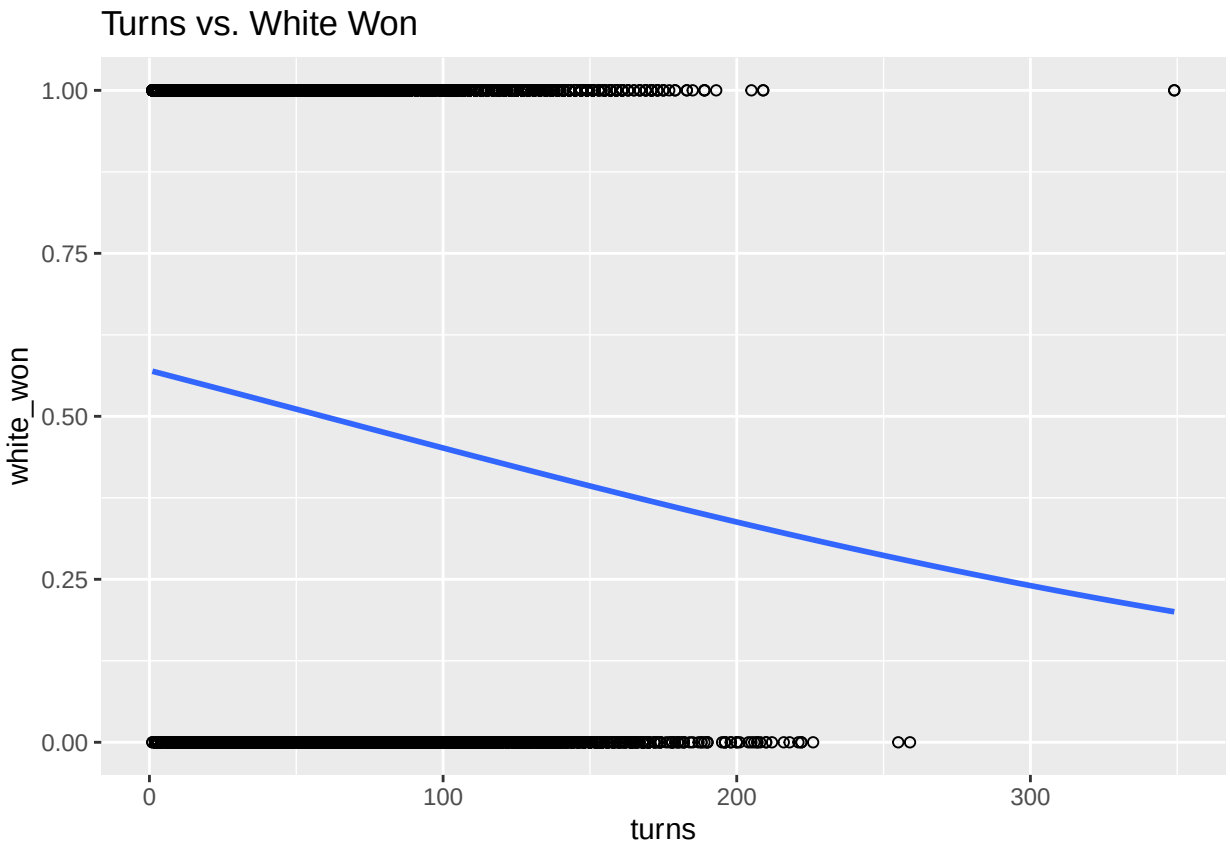

Figure 4: White winning in Relation to Turns

Figures 4 and 5 displays how the probability of White winning in relation to number of turns and game type. It is evident that the probability of White winning decreases steadily as the number of turns increases, while it is less evident the probability decreases with the increase in game type.

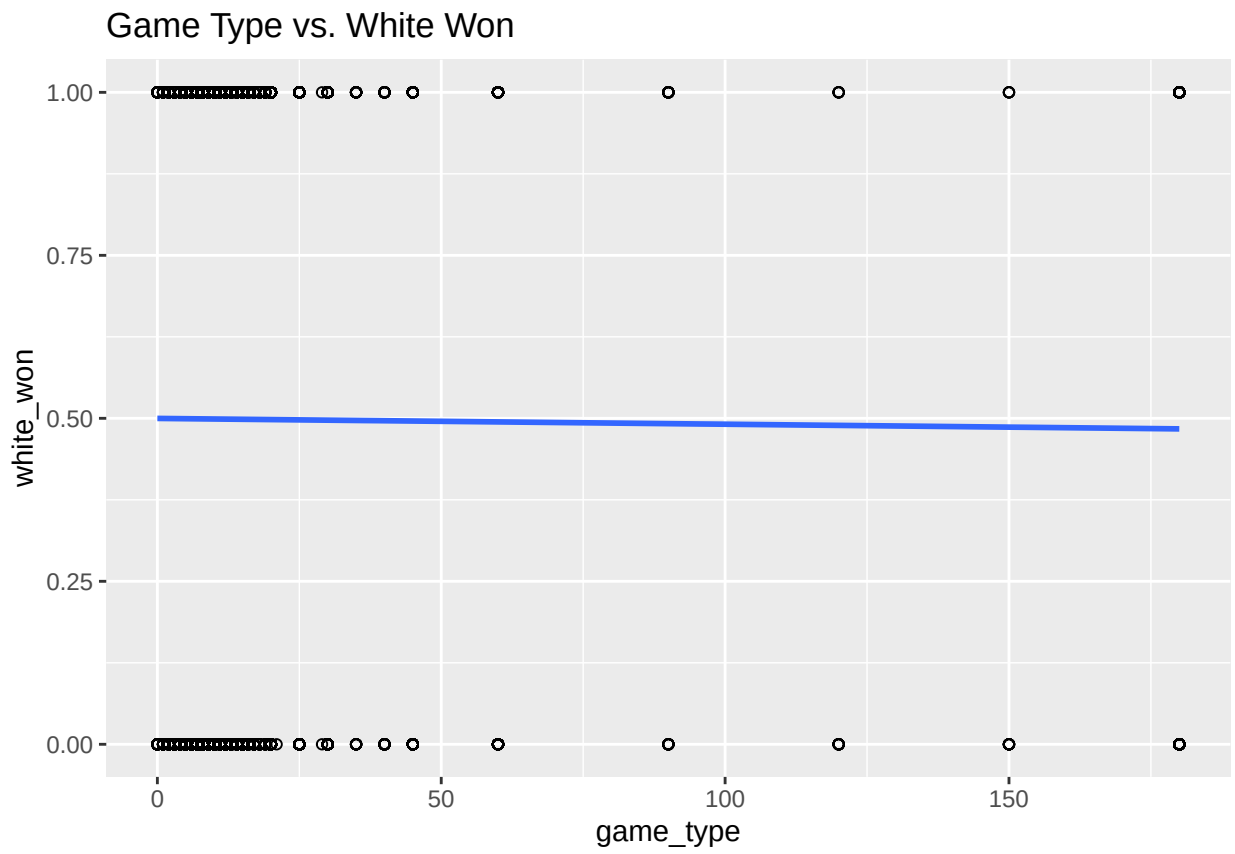

Figure 5: White winning in Relation to Game Type

The logistic model provides a positive coefficient for `diff_rating`, which is the difference between player White's rating and player Black's rating. This result is plausible as the higher the difference in rating, the greater advantage player White has. Figure 6 displays the sigmoid relationship between the difference in rating and player White winning; it is seen when the difference in rating is around 0 (i.e. players have relatively equivalent ratings), the odds of White winning is approximately 50%.

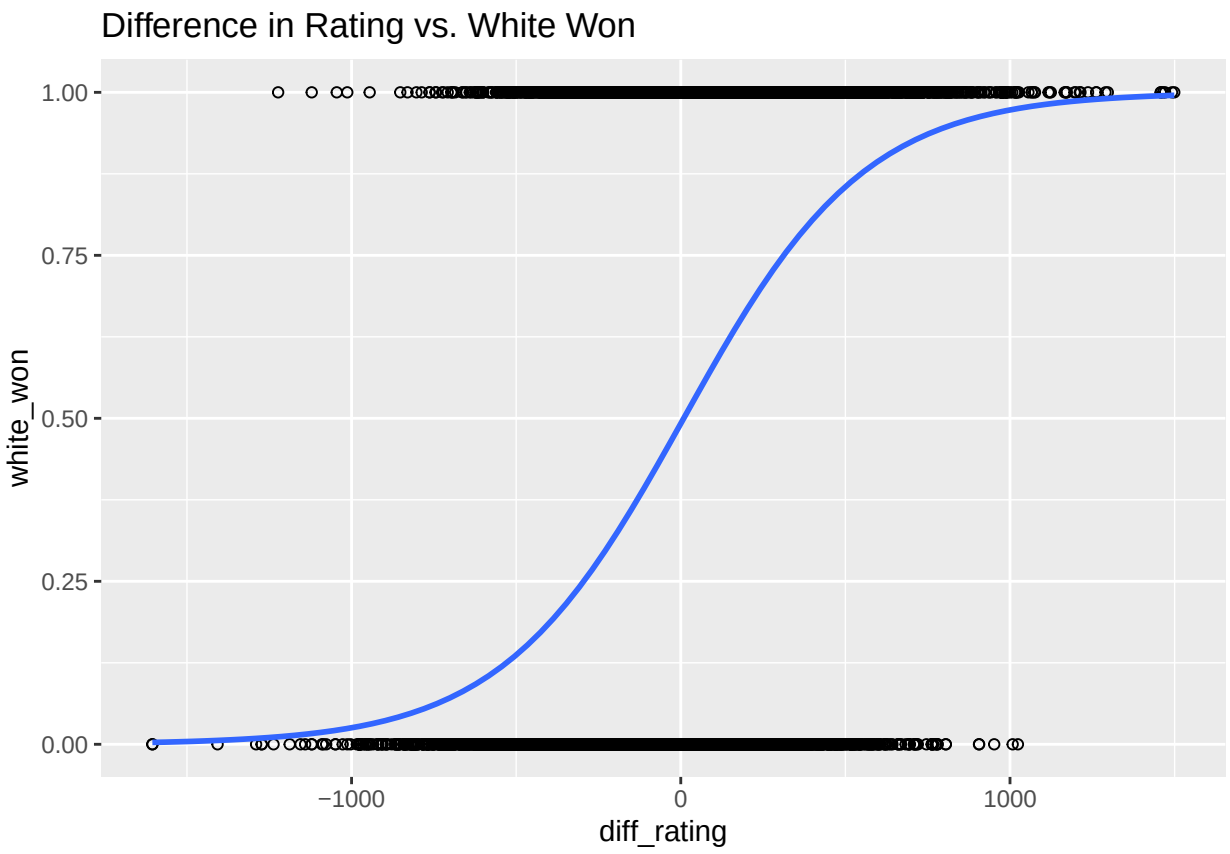

Figure 6: White winning in Relation to Difference in Rating

The second logistic model computes the coefficient for each of the opening moves in relation to their affect on the log odds of player White winning the game. Table 4 and 5 show the highest and lowest coefficients of the given factor levels of `opening_eco`. Using https://www.365chess.com/eco.php to decode the associated codes, it is seen that the best opening moves for White are Benoni, Taimanov variation (A67), Sicilian, Dragon, classical Variation (B74), and Queen's Gambit Declined semi-Slav (D47). Inversely, the worst opening moves for White are Benoni, classical (A71), Dutch, Leningrad (A89), and Queen's Gambit Declined, Tartakower (D58). It is interesting to see that slight variations of one move can result in it being considered a strong opening move or a weak opening move. For example, the Benoni Defense is considered a strong opening under the Taimanov variation while it is considered a weak opening move under the Classical variation.

In conclusion, player White does not win chess games at a significantly higher rate than player Black even with having the first move advantage. The longer the chess game goes for, the log odds of player White winning the game decreases. Using this information, player Black can aim to reduce the significance of player White's first move advantage through extending the game and avoiding blunders in the opening stages.

Weaknesses

One of the most significant weaknesses of this analysis is the sample size of the dataset. As stated previously, there are an estimated 600 million chess players worldwide and this study analyzes approximately 20,000 players on a single platform. Due to this, the results may not be generalizable to the population of Chess players.

Another weakness with this analysis is in relation to the logistic model. The model produced a rather high AIC score, which can imply a poor model fit. This is also supported through having an AUC between 0.5

and 1 as seen in Figure 2, which can imply the accuracy is not satisfactory.

The fact that Chess is typically a game played by two humans can also be considered a weakness in this topic. Chess is prone to human error, which may not be recorded in the dataset and so analyzing whether a specific player has an advantage may not be accurate.

Next Steps

In terms of the model, adding interaction terms or other predictors which can model the human error aspect of Chess can help produce a more accurate and meaningful model. Adding more complicated terms can also decrease the AIC and improve the accuracy. It may also be useful to group the data by average rating between player White and player Black as this can show whether the first move advantage is more significant in lower rated games.

The key next step would be to collect more data. This can be done by collecting games from numerous online Chess platforms rather than just LiChess.

References

Jolly, M. (2017). "Chess Game Dataset (Lichess)". Retrieved December 7, 2020 from https://www.kaggle.com/datasnaek/chess.

Charalampos Bratsas, Kleanthis (2020). gginference. Open Knowledge Greece. Retrieved December 19, 2020 from https://github.com/okgreece/gginference

Chess Statistics, Retrieved December 7, 2020 from www.chessgames.com/chessstats.html.

Cowen, Tyler. (13 Nov. 2018). "Chess Is the Killer App; How and Why a 1,500-Year-Old Game Has Conquered the Internet." Bloomberg, Retrieved December 7, 2020 from www.bloomberg.com/opinion/articles/2018-11-13/world-chess-championship-2018-is-made-for-the-internet.

Revelle, William. "Procedures for Psychological, Psychometric, and Personality Research [R Package Psych Version 2.0.12]." The Comprehensive R Archive Network, Comprehensive R Archive Network (CRAN), 16 Dec. 2020, cran.r-project.org/web/packages/psych/index.html.

The Methodology Center, "AIC vs. BIC", Retrieved December 18, 2020 from www.methodology.psu.edu/resources/AIC-vs-BIC/.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, Retrieved December 18, 2020 from http://www.biomedcentral.com/1471-2105/12/77/