The Significance of the First Move Advantage in Chess

Labib Chowdhury

21/12/2020

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

Abstract

Keywords

Chess, First Move, White, Black, Logistic Regression, Experiment

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.

A table summarizing the characteristics of this dataset can be seen in Table 1 below (Revelle, 2020).

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

Model

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.

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 1.

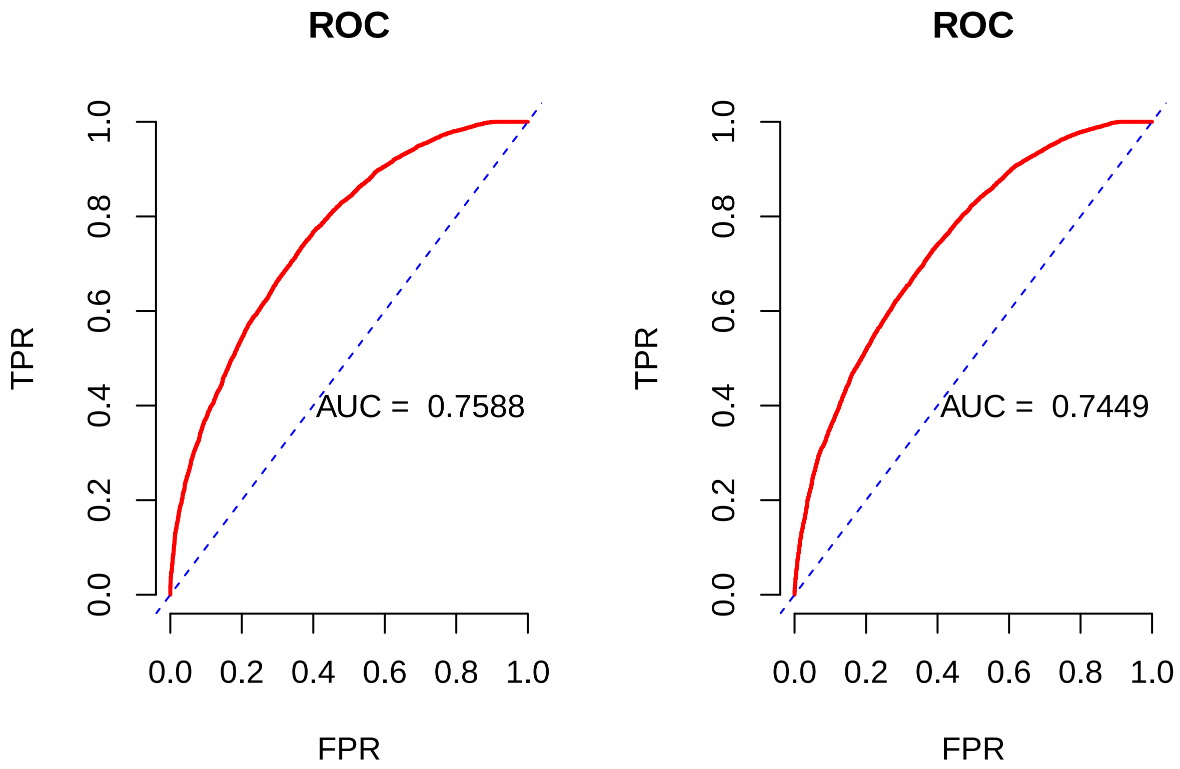

Figure 1: ROC Curve of Full Model and Reduced Model

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, every predictor in the reduced model is statistically significant. 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`, `victory_status` 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.
- $\beta_0 =$ the intercept estimate.
- $\beta_1 =$ the change in log odd for every unit increased in the associated predictor.

Results

do t test to see if white winning is significant

Discussion

References

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

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/