My datascience project title

Irma Student and Soham Eye

April 4, 2023

1 Overview

2 Introduction

Context and motivation One of the fascinating aspects of chess is the contrast between the simple rules and foundations required to understand the game, and the extreme complexity and depth of the knowledge required to master it. Pawns have a similar juxtoposition as the most basic and numerous pieces on the board, they are often seen as enablers for more powerful pieces, but the influence of pawn moves on the outcome of a game cannot be understated.

In this data science analysis, we will explore how pawn moves in each file (a-h) influence the chances of winning, drawing or losing a chess game. We will use a large dataset of chess games from various levels, and apply statistical and machine learning techniques to extract meaningful insights and patterns. We will also discuss the limitations and challenges of this approach, and suggest some directions for future research.

Previous work Brief description of any previous work in this area (e.g., in the media, or scientific literature or blogs).

E.g. Recent surveys show that most students prefer final projects to final exams [3].

Objectives This paper seeks to discover if the number of pawn moves made in each file can inlfuence the outcome of a chess game, and if so, find the file(s) on which pawn moves are most influential. We want to know how the pawn activity changes the game result and will attempt to explain our results according to standard chess theory.

3 Data

Data provenance The data we analysed was uploaded to kaggle.com by user Adityhja1504 for use in the public domain, and was created using the chess.com API.

Data description The full dataset contains 66,879 records, and was initially formatted as a csv file. Each record has 14 fields describing identifying features of the game and each player (white and black). One such field is the game PGN, a standard format for encoding chess games, which includes in istelf much of the data in other fields, thus causing an overlap. For this reason, only the following fields were used in analysis.

- PGN
- white result (win, checkmated, resigned, timeout, insufficient, timevsinsufficient, stalemate, agreed, abandoned, repitition, 50 moves)

- black result (win, checkmated, resigned, timeout, insufficient, timevsinsufficient, stalemate, agreed, abandoned, repitition, 50 moves)
- white rating
- black rating

The fields were deemed irrelivent, their description and the author's orginal post of the data can be found on the kaggle website.

Data processing The initial step in processing was creating a field to record the number of pawn moves made by white and black in each file. *Black pawn moves* and *white pawn moves* hold this information as an 8 value array where the first value corresponds to the number of pawn moves made in the *a* file, the second to the *b* abnd so on.

In creating this data, it was neccessry to outline exactly what can be classified as one pawn move in a file; the definition was chosen to be as follows:

- One pawn move is recorded in file x for a colour when a pawn in file x of that colour is moved.
- If a pawn moves two squares on its initial push, this still counts as only one pawn move.
- If a pawn captures a piece in a file adjacent to x (including via *en passent*), this counts as a pawn move in file x.

Naturally, *sum white pawn moves* and *sum black pawn moves* were also computed, each holding the total number of pawn moves for each colour.

As shown above, the fields *white result* and *black result* have 11 possible values, 2 of which are a loss, and 7 of which are a draw. For this reason, *white simple result* and *black simple result* were created, each taking the values WIN, LOSS, DRAW, or ABANDONED, in order to simplify the results of the games.

The next stage in processing was cleaning the data, for which we had the sole objective of removing abnormal chess games. Abnormal games were classed as abandoned games, games with no pawn moves, games with equivalent start and end times, games with an event title (taken from PGN) which included the words, "ENDGAME" or "PUZZLE", games with no recorded moves, and games with rules other than basic chess (e.g. chess960).

4 Exploration and analysis

In chess theory, an understanding of pawns and pawn structure are often highlighted as what differenciates amatures from higher level player CITE, but can such a simple metric as pawn activity show the level of a player. By grouping the games by elo class CITE we can visualise the trend in the total number of pawn moves for white and black (figure 1).

It would be foolish to assume that there is a linear relationship between the number of pawns pushed and achiving a higher rating, instead a two logistic regression model was trained to classify games as a win, loss or draw, from the number of pawn moves made in each file. One model was trained for each color due to the slight advantage white obtains from playing first. Both models used a 80/20 train/test data split and have metrics outlined in figure 2.

Using the regression coeffients, we can plot the percentage change in the liklihood of the result of the game for a one unit increase in the number of pawn moves in that rank (figure 3). Percentage change is calculated as

$$p = (e^c - 1) \cdot 100$$

where c is a coefficient of the regression.

The low accuracy of the regression model removes some significance from the results, so a 400-Nearest-Neighbours classification was also implemented. We used a 60/20/20 split for training, testing

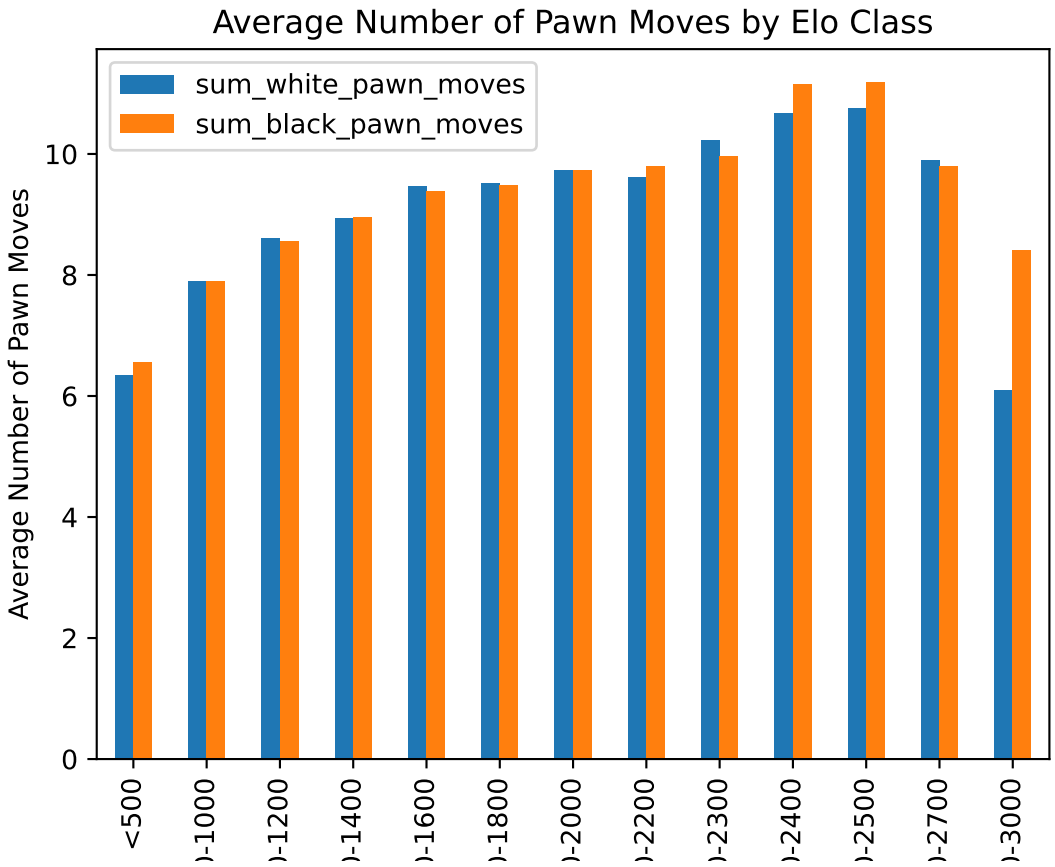
Average Number of Pawn Moves by Elo Class
sum_white_pawn_moves
sum_black_pawn_moves
Average Number of Pawn Moves
0
2
4
6
8
10
<500
)-1000
)-1200
)-1400
)-1600
)-1800
)-2000
)-2200
)-2300
)-2400
)-2500
)-2700
)-3000

Figure 1:

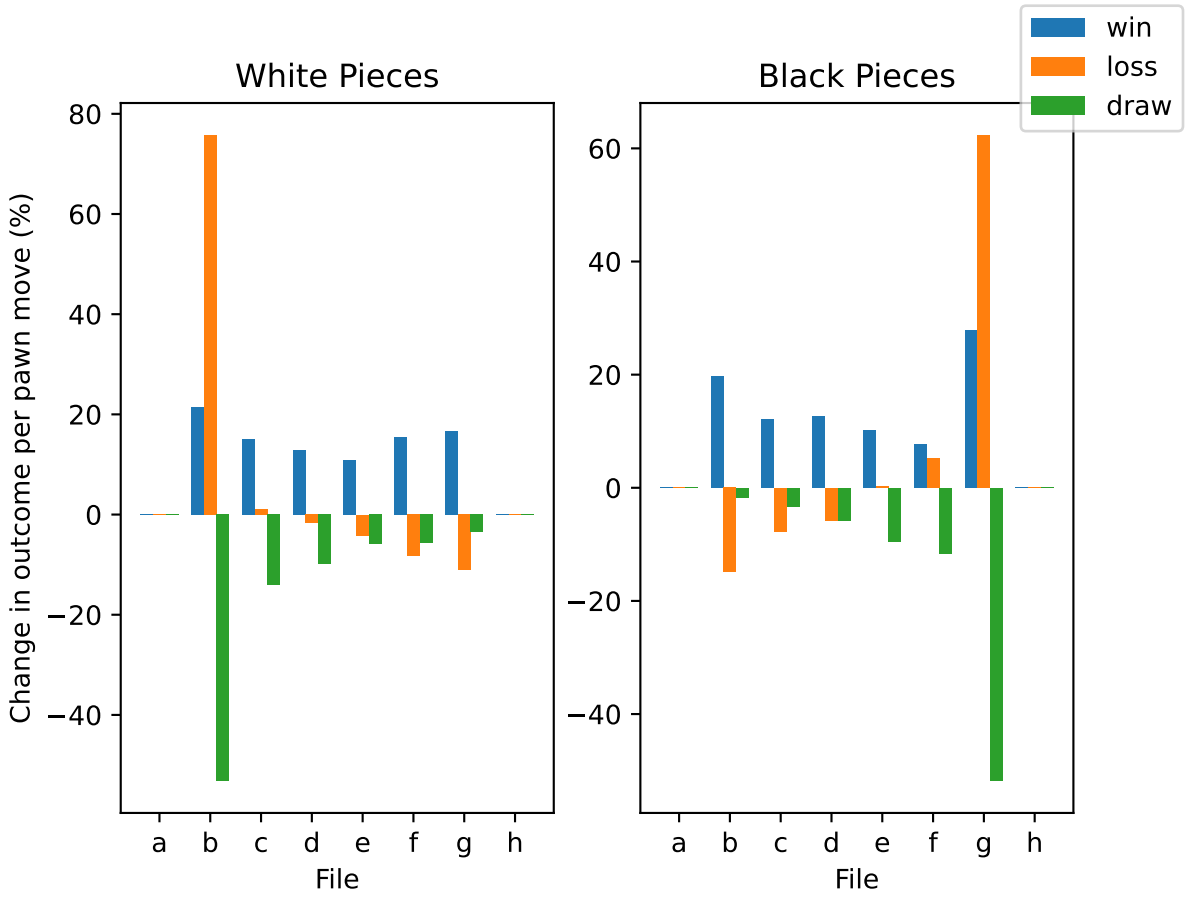

Figure 2: Logistic regression results for white and black pieces.

	White Score	Black Score
Accuracy	0.536876	0.554375
Precision	0.554671	0.534788
Recall	0.536876	0.554375
AUC-ROC	0.639057	0.645282

Figure 3:

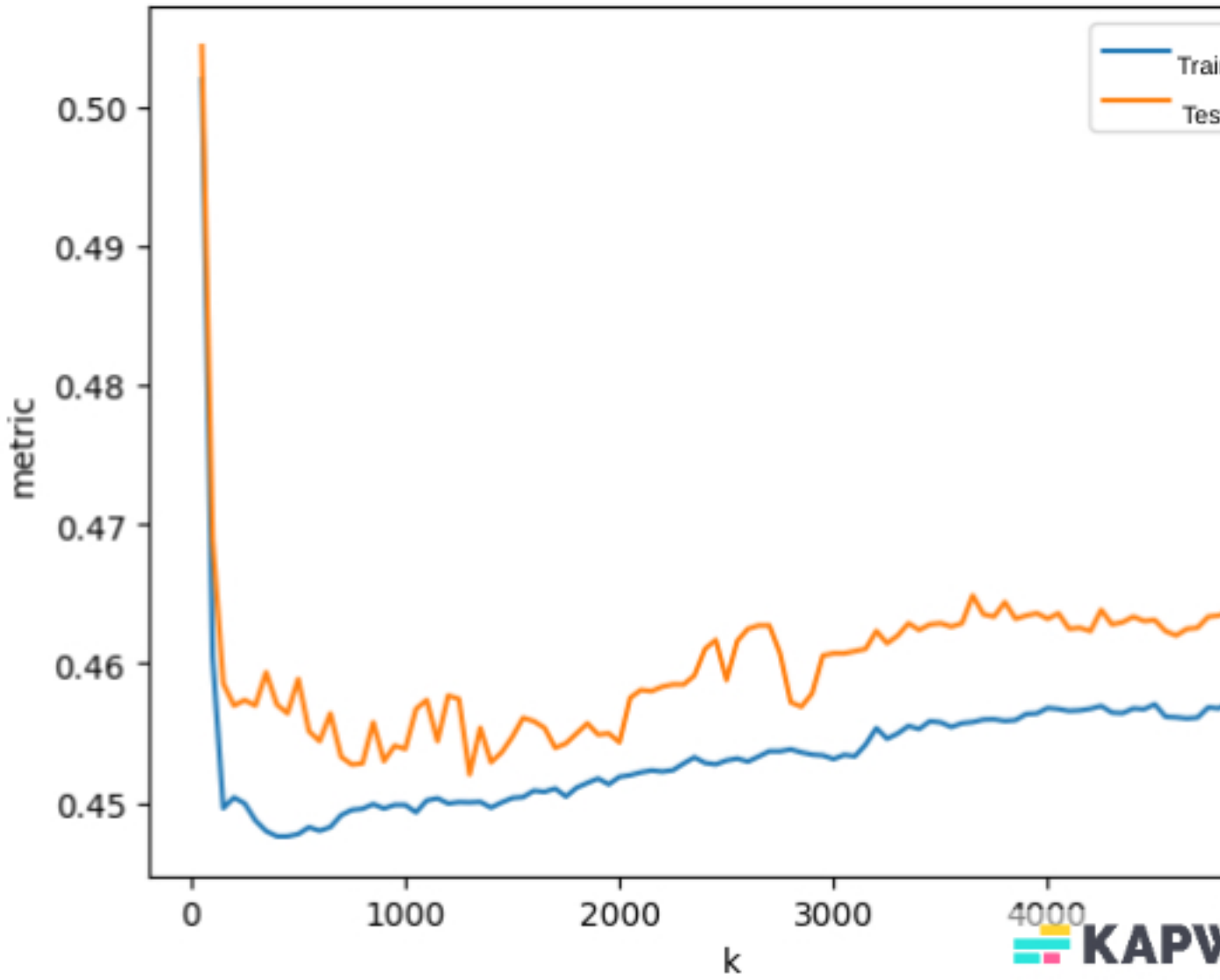

Figure 4:

	White Score	Black Score
Accuracy	0.558792	0.566343
Bias	0.441870	0.441390
Variance	0.212602	0.238316
Train Error Rate	0.433277	0.427884
Test Error Rate	0.441870	0.441390

Figure 5:

and validation sets. $k = 400$ was selected as the approprate hyperparameter for the model through testing values of k between 1 and 5000 and comparing train and test error rate (figure 4). Metrics were retrieved using a ten fold cross validation which gave a similar accuracy to the previous logistic regression model (figure 5).

5 Discussion and conclusions

Summary of findings The initial logistic regression model predicts that the most influential file for white is by far the b file, and for black the g file. In particular, for each pawn move white makes in the b file, this model describes a 22% increase in the liklihood of the game being classed as a win, a 77% increase in the liklihood of the game being classed as a loss, and a 52% decrease in the liklihood of the gam being classed as a draw. Likewise for each pawn move black makes in the g file, the model decribes a 27% increase in the liklihood of the game being classed as a win, a 61% increase in the liklihood of the game being classed as a loss, and a 49% decrease in the liklihood of the game being classed as a draw.

This indicates that controlling the b file is critical for white, and controlling the g file is critical for black, in order to increase the chances of winning. This also suggests that playing defensively (i.e. making less pawn moves) may reduce the liklihood of a win due to the correlation of pawn activity here and losing or drawing. Apart from these files, we can see that pawn activity in the center files (c, d, e, f) increases the likelihood of winning more than any other outcome for both colours (excluding black for the f file). This correlates to basic theory on the importance of conrolling the center in chess.lso suggests that playing offensively (i.e making more pawn moves) on these files has a greater benifit than playing passively.

Evaluation of own work: strengths and limitations Chess is a highly complex game and attempting to predicting the outcom eof a game is an unsolved problem with all the information on a game, nevermind juts the number of pawn moves per file. With this in mind, while the accuracy of model was 20% above random selection, this value is still far below the standard in most classification models. The low accuracy can also be attributed to the size of the dataset, which when reduced to only pawn move counts, lacks variance and holds a bias, as demonstated in figure 4. The lack of information on the influence of the a and h files can be attributed to these being the only files with only one neighbouring pawn, therefore less pawns capture into this file, reducing the overall number of pawn moves here.

Comparison with any other related work E.g. “Anscombe has also demonstrated that many patterns of data can have the same correlation coefficient” [1].

Wikipedia can also be cited but it is better if you find the original reference it for a particular claim in the list of references on the Wikipedia page, read it, and cite it.

The golden rule is always to cite information that has come from other sources, to avoid plagiarism [2].

Improvements and extensions

References

[1] Francis J Anscombe. “Graphs in statistical analysis”. In: *The American Statistician* 27.1 (1973), pp. 17–21.

[2] Wikipedia contributors. *Plagiarism – Wikipedia, The Free Encyclopedia*. Last accessed 22 July 2004. 2004. URL: `https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350`.

[3] Phil Space. “Why oh why must I do this project?” In: *The Daily Post* (2021). Retrieved on 28 February 2021. URL: `https://www.dailypost.com`.