Elijah A Seeley

Chess Openings and Predictions – Final Report

Introduction

The future of chess analysis is upon us. With the movement of chess into the digital realm it is possible now to iterate over billions of chess games to reveal interesting and perhaps useful data. This project was intended to explore the opening moves in chess and detect their impact on the outcome of a game. While this project was completed with comparatively miniscule data, it can reveal what may be possible with larger datasets.

Audience

Being well versed in chess openings is not required for understanding this project. However, understanding how chess is played the importance of choosing an opening is required to understand the implications of what was explored. Without this introductory knowledge the significance of small changes may be lost to the reader. This future of this experiment is intended for intermediary players of chess as the future of this experiment will be gathering the accumulated games of these regular players.

What is the Point?

Some readers may be wondering why do this at all. Already we have chess engines that can determine the best move and websites like chessgames.com that can provide all the information we need about games. However, I would like to counterpoint that these engines and these collections can not possibly tell all the information. While chess is a game about accuracy and calculation, it is also a game about ideas. I like to think of chess as a conversation where one side promotes a thought, and the other side has a rebuttal. Billions of these conversations are had online, yet our understanding of openings is heavily left to the record of tournament games that are documented and slotted online for others to review and add to the database. This gets away from what it is like to be an average player and assumes that all players are going to play up to code. For anyone that has played low level chess – we know this is far from the case. By analyzing this mass amount of intermediate and beginner gameplay we can better serve the player by giving real metric on where players make blunders, what elo is playing what opening, and overall understanding the meta game of chess. Because of the nature of chess, these calculations are hardly far from reach. Future experimentation will hopefully bring us to a new point of, dare I say, enlightenment of the game for players that are not master level.

Goals

The goal of this project was to explore the openings of the dataset and subsequently develop models that could make predictions off related items, but not the most direct items.

The earliest model was an attempt to train an algorithm to predict which opening the players found themselves in without knowing what moves were played. This turned out to be unsuccessful. But perhaps if more overarching data was provided, such as demographical data, there would be more hope.

The final model would be trained to determine the winner of a game without knowing the moves played and how accurate the model could be at predicting this off of the opening alone.

The Data

The data was retrieved off from a user on Kaggle who was generous enough to provide a workable csv file by using Lichess's api to scrape some games from the Lichess server. Scraping from the server is not the most efficient way to grab this data as all games cannot be scraped at once and only a little bit at a time by individual users or clubs. Because of the collection method it is important to note that this dataset does not represent an average set of chess games.

Lichess.org is a free online chess platform that provides players a dataset of there own games, complete with analysis and trends. Analyses like those available on Lichess is already a trove of useful information for the player and it provides a good base on which further study can be founded.

The data was initially provided with over 20,000 games. However, we wanted to limit to games where the information would be useful to a model. We limited the data based off the following requirements:

- At least 5 moves were played in the game.
- The games' openings were popular enough to be played at least 200 times in this dataset

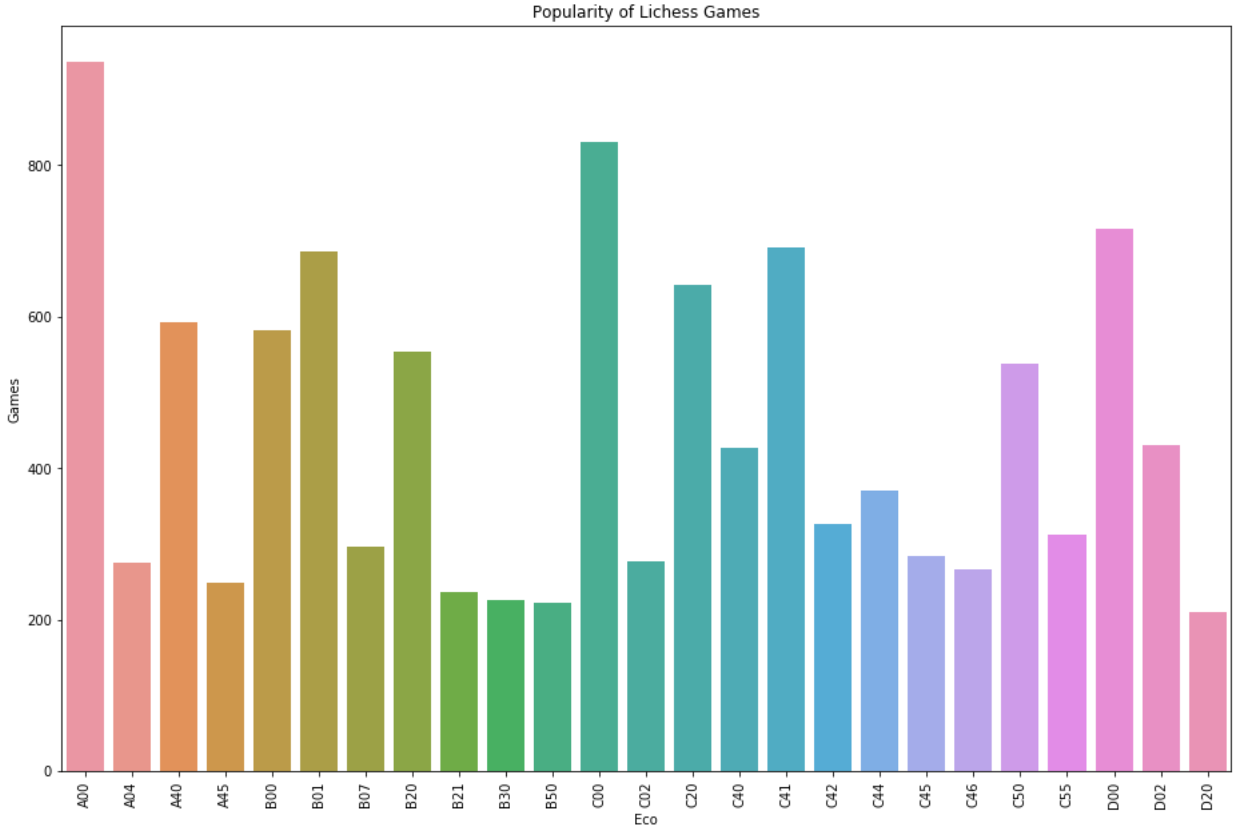
This left us with a final dataset of about 11,000 games to explore and train the models with.

The final features for each game are listed below [note: some of these were dropped and altered for the modelling step]

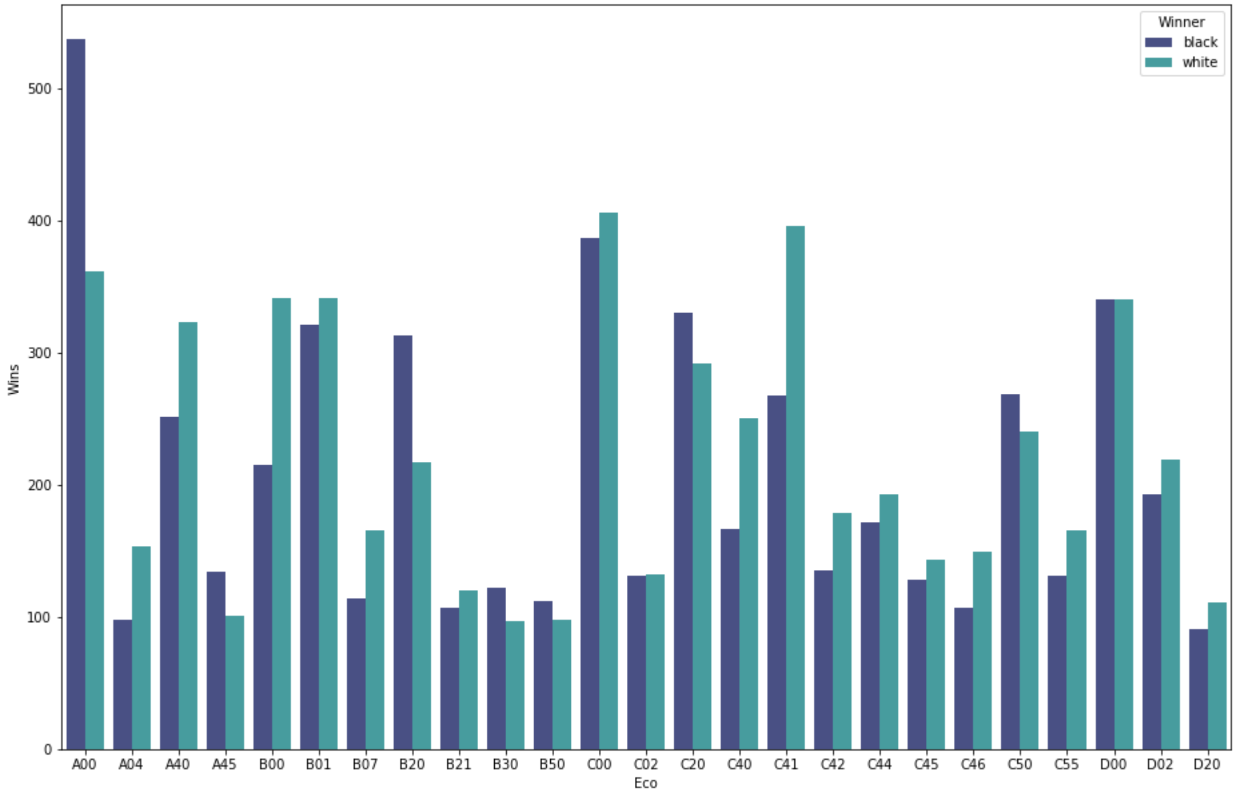
- Turns
- White Rating
- Black Rating
- Opening Move Count
- Victory Status (i.e how the game completed)
- Increment Code (generalized into categories: Bullet, Blitz, Rapid, Classical)
- Opening Eco
- Opening Name

EDA

The first order was to determine which openings the player used the most. Below is the initial graph of all the openings played with at least 200 games.



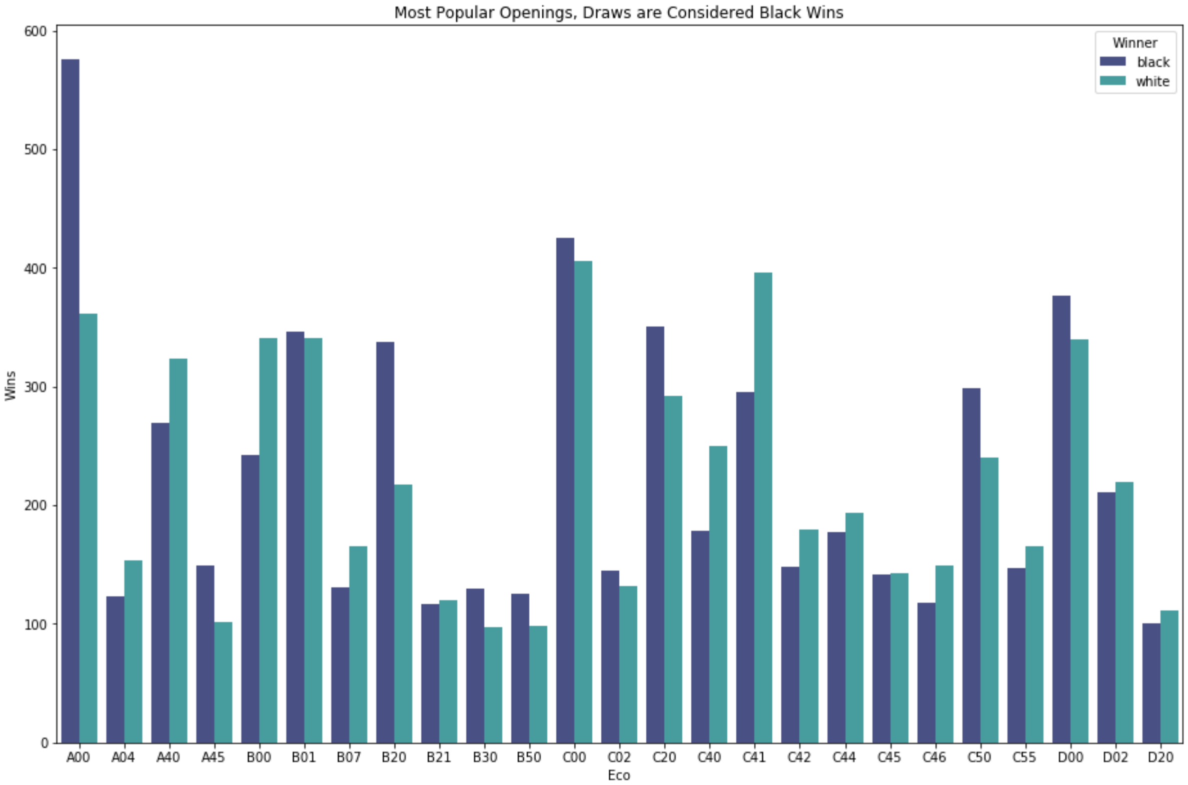
Here we see A00 is the most popular pick of these players. But why is that? A00 is considered the ‘uncommon opening’ category. This is where the flank openings that are not normally played are coded. Now because these games were scraped from average players it is safe to assume that this opening is so popular because of two factors. First, the players choose to play uncommon openings to because at their level it is feasible to still eek out a victory. Second, the players are not versed in theory and therefore end up playing openings that end up being irregular. The second and their most popular openings shown in this graph are the unsurprising C00 and D00 which are the king’s pawn and queen’s pawn respectively. From here it is important to know who is winning these games. The graph below shows white wins and black wins with the draws removed.



Of the 25 different openings here, it looks as though white is playing very well, winning in higher percentages in 17 of them. It’s also interesting to note here that while A00 is the most common opening for white to themselves into, it is also a clear victory for black most of the time (perhaps white should learn a little theory or stick to more practical openings). While looking into each of these openings and their theory is tempting, analyzing these openings individually is outside of the scope of this project.

It would be remiss to leave draws out of the equation, but the number of draws is so small in some areas that it makes little sense to include them everywhere. What I decided to do is to add draws to black’s victory count for the sake of the bar graph only. This gives a clearer representation of the openings strength for black as draws are considered okay for black as black begins the game down a tempo.

Below is the most popular openings, with draws considered wins for black.



Here we see that some openings that were even or even winning for white have lost their advantage. Some positions are highly favorable for one side. This is the type of data that I believe should be available for the average player. What openings to train for, what openings to play based on what is actually being played by players. In the future we would even be able to look at which move 10 is common for each opening and this can give the players a feel of control over their preparation and contribute to the evolving meta.

To make this point further, at a particular level at any given point of time a player could know what openings are most popular for their ranking. As an example, based on the dataset I went ahead and found the most popular openings in each 100 increment of elo, excluding A00. Here is what was found.

```
{'800s': 'D00', '900s': 'C20', '1000s': 'C20', '1100s': 'C20', '1200s': 'C
20', '1300s': 'C20', '1400s': 'C20', '1500s': 'C41', '1600s': 'C00', '1700
s': 'C00', '1800s': 'C00', '1900s': 'B01', '2000s': 'B01', '2100s': 'B01',
'2200s': 'B20', '2300s': '(B23, D35)'}
```

While this data is not representative of the chess community, this is an example of a metric that players can use to prepare for play based on their skill level. Different skill levels will play different complexities of openings. When doing a Pearson correlation between games' average rating and the opening move count (which is the amount of moves played before it was finalized)

I found a relationship of 0.292 between the average rating and the opening move count. While this is small correlation, it is enough push further that elo is directly correlated with which openings are being played the most, at least in this dataset.

Modelling

Going into this I was hopeful that the model would be able to predict the winner of each game with an accuracy above 50%. The models initially were logistic regression, KNN, decision tree, and SVM. While they all provided about the same accuracy on the initial set I decided that the decision tree would be the model we finalize on as the different depths sometimes presented slightly higher results which I found interesting.

Our model comes in two forms. The first version utilizes the dataset in its whole: this includes turns, player elo, victory condition, and increment code. The results for the decision tree at depth 9 are below.

```
                 precision    recall  f1-score   support

       Black       0.61      0.66      0.63      1579
       White       0.64      0.60      0.62      1632

accuracy                               0.63      3211
macro avg          0.63      0.63      0.63      3211
weighted avg       0.63      0.63      0.63      3211
```

Above 50%! Without knowing anything about the game itself the model was able to confirm an accuracy of 63% for electing the winning player. Now comes the part I find the most interesting regarding the models – the second version will be the same data, but with all features removed aside from the opening the game was played upon. Below is the result of the decision tree at depth 9.

```
                precision    recall  f1-score   support

       Black       0.73      0.19      0.30      1579
       White       0.54      0.93      0.69      1632

accuracy                               0.57      3211
macro avg          0.64      0.56      0.49      3211
weighted avg       0.63      0.57      0.50      3211
```

While the accuracy suffered and the recall is skewed, there is an obvious correlation between opening picked and the side of victory. What’s most interesting about this is that *the features that were dropped were assisting the model!* This means that arbitrary metrics like amount of moves played and the increment code were useful to the machine in establishing accuracy.

This gives hope for the future of chess analytics as we reimagine seemingly unimportant datapoints as metrics worth considering.

Conclusion

While this implementation is imperfect this project sets creates some points for moving forward with this form of chess analysis. We can see that metrics that seem unintuitive may be worth studying and I have a hopeful future for bringing mass game analyses to the average player. With information like this in massive form we may even be able to attract more players to the game, solidifying chess as it branches into esports by making it more approachable for players who see it now as some sort of black box of theory. In addition, we may be able to move away from using chess engine analyses as the main evaluation point of any game. More is to come once my hardware and data capacity is able to handle massive datasets.