Final Report for the Analysis of Chess Games and Chess Openings

Sahib Bajwa - 107553096 - CSCI 4502
University of Colorado Boulder
Boulder Colorado United States
sahib.bajwa@colorado.edu

Abstract

As one of the oldest, most complicated, and in my opinion, most fun games of all time, some would think that chess has been thoroughly covered on what moves are supposed to be played. Unfortunately this is not the case as the game has developed overtime and is still developing. For newer players chess can look daunting to take on due to said complexity. The goal of this analysis was to determine the best openings for players to take given their skill level. The hope is that given specific openings for specific players, the players will feel more comfortable playing the game as well as continuing to play due to the success they see playing said openings. This paper will cover the analysis of the Lichess data set bad stores around 20,000 games played on the Lichess website by chess players in 2016. This analysis will help me determine what opening a player should use given their rating, the color they are playing as, the opponents rating, and at the game type.

We will first cover the importance of this analysis and what the results will apply to eventually. Next, we will talk about related research in this space and key takeaways we can take from said research. Talking about the related work will allow us to gain better insight for the current problem as well as why this analysis will be important in the first place. Third, we will begin talking about the actual analysis conducted. This will include the data sets used, any tools used during the analysis, tasks we looked to complete during the analysis, as well as our reasons for picking these methodologies. We will then dive into the evaluation of the analysis Which will speak on our results as well as any interpretations we made about the data. Next, I will highlight key lessons I have learned throughout the project and the analysis which will entail what worked during the project as well as what did not. This section will also include analysis or development that could hopefully happen in the future. Finally, the paper will wrap up with a conclusion, a summary and a reiteration of our key tasks and findings of the analysis.

This topic is very complex, but because I enjoy what is being analyzed, and because I think the takeaways are very important, I think the results of this project are useful to many people who play chess. The growth of the game matters immensely to most people who play it, and if there is a way to get new players more comfortable with the game, then that work is very important unto itself.

Introduction

Chess is one of the most complicated games from a theoretical standpoint. The Shannon Number tells us that the conservative lower bound of the game-tree complexity of chess is 10^{120}.[1] To give this number some perspective, there are an estimated 10^{80} atoms in the universe[2]. It is very difficult to determine the best way to play the game, even given the use of AI (Stockfish, AlphaZero, Leela Chess Zero, etc.). Having a strong opening is one of the most important moments of a chess game as it sets the pace and structure for the rest of the game. Given that humans do not have the computational ability of a strong AI, finding out what the best opening is for a player would help them increase their chances of winning a large amount.

The Oxford Companion to Chess states that there are 1327 named openings in chess[3], and each opening can have multiple variants. One example of this is the popular "Sicilian Defense". Based on what situation the game is in, the "Sicilian Defense" can be modified into another opening. This includes another popular opening, the "Sicilian Defense: Bowdler Attack". Considering there are so many openings that exist and should be used in niche situations, helping players know and understand what opening to use would help them succeed in winning games.

I also believe that it would help lower rated players stay interested in the game as it would give them a strong starting point to build off. FM Steve Giddins said in 2008 that, "the average paler only needs to know a limited amount about the openings he plays. Providing he understands the main aims of the opening, a few typical plans and a handful of basic variations, that is enough." If a newer player learns the openings that are the best suited for them, they will have a much greater chance at becoming better at the game. Like I said before, I also think that once a newer or lower rated player sees the improvements in their play due to playing an opening suited for them, they will most likely choose to continue playing the game due to seeing improvement in their play.

But what is the reason to get new players into the game? The lifeblood of any game, board game, or sports game is its new players. Without new players the game will cease to continue forward. Any player within the chest space who deeply cares

[1] https://en.wikipedia.org/wiki/Shannon_number
[2] https://www.universetoday.com/36302/atoms-in-the-universe/#:~:text=At%20this%20level%2C%20it%20is,hundred%20thousand%20quadrillion%20vigintillion%20atoms.
[3] https://en.wikipedia.org/wiki/Chess_opening#:~:text=The%20Oxford%20Companion%20to%20Chess,the%20middlegame%20and%20the%20endgame.

about the life and longevity of the game wants newer players to join. This is so that the game still has relevancy as well as there being a chance that a new player joining will reach the top level of chess. If there is no easy barrier entry for a newer player who has just started playing the game, they will most likely decide to not continue. Since, as I have stated, most people in chess want more newer players, this study is very important for the entire game of chess even though it will most likely only effect newer or lower rated players.

Going along with this, it will not work to simply give a player an opening that is considered strong and tell them to use and learn it. This is because players at different ratings are better or worse at different openings. A complex opening that is highly successful may be a good option for a Grandmaster (a very highly rated player), but a newer/lower rated player may not understand the depth to the opening. Thus, at different ratings and skill levels, players should prefer and learn different openings.

There are a couple of potential challenges that come with doing this project. The first is that even though I could statistically determine a good chess opening for a specific player, there is no real way to tell if it will work as intended unless given to that player and tested. Considering everybody thinks differently, it is very hard to make general assumptions about a bracket of people even if there is a good amount of previous data. Even with this being said, I think that there can be a general structure set for openings given what rating bracket to players in. This will only really work for lower rated players like I have said before as higher rated players will most likely have figured out what openings they are good at or ones that they are comfortable with.

So, there are a few key ideas that I want to speak on that will make the idea of giving an opening to a player easier to understand. The opening sets the pace for a game as well as the games later structure. The pace of the game meaning how fast pieces are taken or how long a player has to think about a specific move. For structure of the game, I am speaking on how the pieces will end up in specific spots on the board. Obviously, if there are pieces in spots that are favorable for the player, it is most likely easier for them to play out the rest of the game. Getting to this favorable structure as well as having an easier time keeping up with the pace of the game is very important for newer players to learn. If they are given an opening that fits their skill level or playing ability, they should be able to match these two criteria more easily.

The final point I want to talk about is that this data will only be beneficial if given to players and results are seen from their end. If when this data is analyzed and specific openings are given to players and they respond well, then I will know that the analysis is correct and useful.

Thus, I decided to take a look at and analyze large chess games data sets in order to determine how and when to use certain openings. Due to the reasons stated in previous paragraphs, I believe this project to be important and useful. There are a lot of intricacies in chess, and helping players determine a good opening will allow them to be better prepared for the game as well as make the game overall much easier to win.

Related Work

• Chess AI

• There are Chess AI that are built around specific types of playstyles that prefer specific types of game states. Thus, they start a specific way in order to push the game towards that game state. Some of the most popular include: Stockfish, AlphaZero, Leela Chess Zero, and Komodo. A large list of chess engines and their ratings can be found at: https://ccrl.chessdom.com/ccrl/4040/

• Chess.com also has many videos speaking on different AI and their abilities to play the game. One of my favorites, and one of the most useful to me, is chess.com talking about AlphaZero (using Google's DeepMind), the variants that it can use, and its assessment abilities in game period while this does not specifically have to do with openings, it sheds light into the world of chess AI as well as how complicated/strong they have become. This article can be found here:

https://www.chess.com/news/view/new-alphazero-paper-explores-chess-variants

• Chess Theory

• There has been a lot of theory crafting about how to play in specific game states. One great example of this is the book FCO: Fundamental Chess Openings by Paul van der Sterren. Due to the volume of books on the subject, it is easy to find resources on specific openings or theory about chess game states.

• One of the largest websites regarding chess is chess.com. They have an entire page as well as training modes that help players learn easy openings. This differs from my analysis in that these are based on chess theory like I talked about above. They do not take in specific ratings or other factors (such as player starting color or player rating and starting color). despite this, I still think this page or simulations are very useful for new players. A link to one of the best pages for openings for beginning players is here: https://www.chess.com/article/view/the-best-chess-openings-for-beginners

• There is also lots of chess theory that is done on the individual level by high level chess players before specific games. This is usually not seen in/by lower-level players. One of the most recent examples of this that comes to mind is when Grandmaster, and considered best player in the world, Magnus Carlson was playing other Grandmaster Hikaru Nakamura at the Magnus Carlson Invitational. GM Hikaru Nakamura was able to theorize and practice an opening to the point where he played the game to an around 99% accuracy compared to what Stockfish AI believed would be the best moves for the game. Chess Theory is very researched and can be the deciding factor to who wins many high-level chess games. Considering how difficult, intensive, and nuanced this type of analysis by high level players is, it should not be used by lower-level players to dictate their play. This is another big reason why I think my analysis is very important and believe that it will help lower-level players more than studying how these high-level players do analysis.

Methodology

There was one main dataset set that I used in my analysis. This data set comes from Lichess and it is a record of 20,000 plus games played on the Lichess website in 2016. The data set held a good amount of information. The main information that it held was: turns per game, player ratings, game winner, white starting player rating, black starting player rating, all moves in standard chess notation, opening name, and number of moves in opening.
The data set can be found here:
https://www.kaggle.com/datasnaek/chess

There was one other dataset that I looked into heavily but note analysis was actually done on the data within it. This is because a lot of data had already been analyzed within it and displayed very well. This data set comes from FICS. The data set can be found here: https://www.ficsgames.org/download.html This data set includes information such as: winning percentage of computer versus computer in games, win percentage of human versus computer, most active players by ranking, And much more. The reason I did not use this data set as much was because most of the information I need is in the Lichess dataset and there is more information in this one that I do not need. I did think that it was worth noting this dataset as it would be very useful for further analysis of this topic. One section that I found very interesting in this dataset was the section talking about most common openings in games of computers computer, computer versus human, and human versus human. Although I could not interpret what the openings were because they were using a coding scheme for the openings (the coding scheme is known, I am just not well versed in it), it did reflect that some openings are used much more often than others when given human players. This was something that I noticed and learned from the Lichess dataset as well.

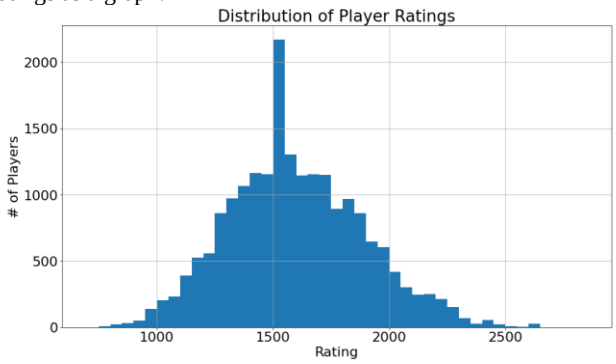
Some tasks that I wanted to complete on the Lichess dataset included: determining distribution of player ratings, Distribution of turns per game, most common openings for all players, winning percentages based on color, and preferred player openings by rating and color. All of this was mostly done through graphing using data frames (specifically, pandas, matplot, in math Python packages).

This was all done using Python as Python is a very easy to use computing language and worked well with the dataset I was using as a CSV. Using pandas, matplot, and math Python packages on data frames, I was able to graph and analyze this data that I would use for predictions later on for players.

Some analysis that I did that did not include a data set was surrounding AI. Using a simple Python chess library, I was able to simulate games of chess. Using these, I was able to test whether some of the information in the Lichess data set was correct and/or paralleled in execution. Although I would like to make the AI more sophisticated to take in openings to use, that was not really possible. Although I was not able to get it to work, there are many websites online, such as chess.com, that have this ability and can be used for such analysis.

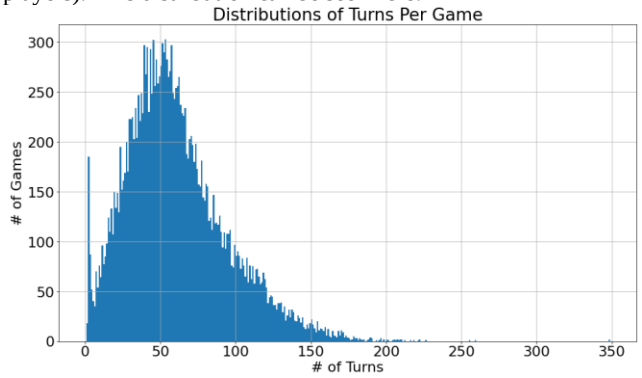
Evaluation

The first thing I did in my evaluation with the Lichess dataset was determined the distribution of player ratings within the dataset. This is important as I would like to know, and it would be very important to know, what types of players are most likely going to use the information given by this analysis. It also gives me a good idea what types of players I am looking at when looking at opening levels within games. Here is a distribution of player ratings as a graph:



As you can see, it is a mostly standard Bell curve. One point that is very easy to notice is that there is a large spike after 1500 rating. After seeing this distribution, I decided to separate players into three brackets. Players below 1000 rating, players within 1000 and 2000 rating, and players above 2000 rating. Although this may seem semi skewed as most of the players are within 1000 to 2000 rating, from personal experience and speaking to players, most people agree that 1000 to 2000 is usually similar skill level. I tend to disagree with this as I believe that above around 1600 rating there is a noticeable difference compared to players lower, but as this is usually accepted upon, I decided to focus my analysis in this way.

The next thing I decided to look into was how long games usually are. although this metric was not used in the final analysis of openings, I thought it would be important to look at how games usually play out in terms of length of games. Most games that take a very long time are either players not knowing how to end the game (i.e., less skilled players), or players who are playing against each other who know how to not lose easily (i.e., higher skilled players). This distribution can be seen here:



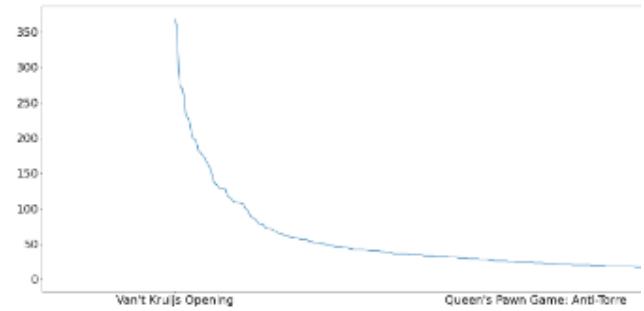
Most games and around the 50-turn mark. I think it is important to note though that some games and early at the four-turn mark. This is due to a scholar's mate. The scholars mate occurs when one player accidentally plays poorly at the beginning of the game, allowing them to be mated (mate/mated is another word for beaten used commonly in chess). This huge bump in losses on turn 4 is another reason why I think openings are very important for newer

players. If a player can understand how to use an opening even if they are not skilled at the rest of the game, it will help them get past these early points of failure.

The next thing I decide to look is what this project is really about, opening frequencies. Here are the five most popular openings among all players within the Lichess data set:

Van't Kruijs Opening	368
Sicilian Defense	358
Sicilian Defense: Bowdler Attack	296
French Defense: Knight Variation	271
Scotch Game	271

To add on top this comment here is a graph for the beginning few openings and their frequencies:



This graph actually goes on for a long time as there are many openings, but I think it easily shows that the most often used openings are used much more than the least often used. I thought that this was important to look into as of 1 opening is used very much by players, that must mean that it is either easy to use or considerably good when determining an opening to pick. Van't Kruijs Opening was the most popular and we will make another few appearances in this analysis.

The next thing that I decide to look into because I thought it was interesting would be the chance of winning depending on the color you started as. If are the white starting player, that means you get to move first. This is usually considered to be the better starting option. The numbers for winning percentages based on starting color are:

Chance White Wins: 49.8604%

Chance Black Wins: 45.4033%

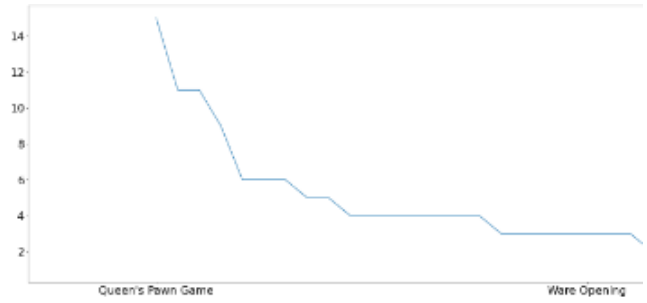
Chance a Tie Occurs: 4.7362%

This confirmed my earlier thoughts and assumptions that the player who was starting as the white color would win more often.

Finally, I decided to look at the playing percentages of openings based on rating bracket and starting color. There are six different tables for this as there are three rating brackets and two starting colors.

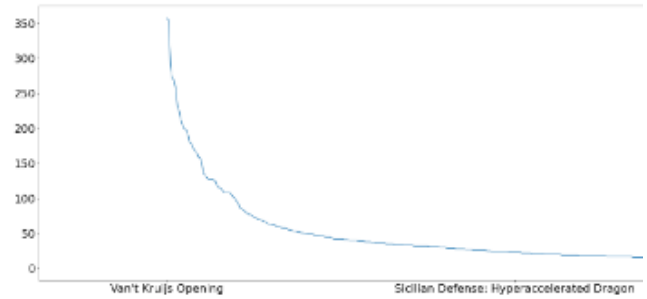
White's preferred openings under 1000 rating:

Queen's Pawn Game	15
Scandinavian Defense	11
Van't Kruijs Opening	11
King's Pawn Game	9
Englund Gambit	6



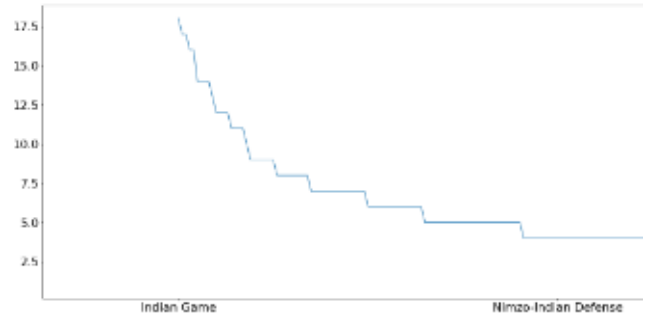
White's preferred openings from 1000 to 2000 rating:

Van't Kruijs Opening	357
Sicilian Defense	354
Sicilian Defense: Bowdler Attack	292
French Defense: Knight Variation	271
Scotch Game	269



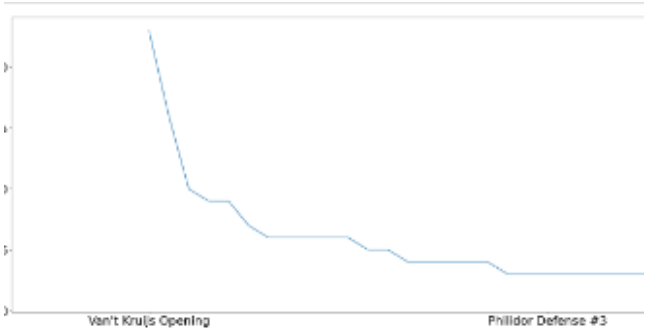
White's preferred openings above 2000 rating:

Indian Game	18
Queen's Pawn Game: London System	17
Scandinavian Defense: Mieses-Kotroc Variation	17
Horwitz Defense	16
Queen's Pawn Game: Mason Attack	16



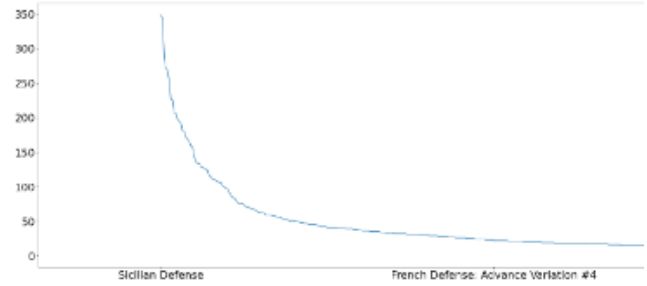
Black's preferred openings under 1000 rating:

Van't Kruijs Opening	23
Scandinavian Defense	16
King's Pawn Game: Leonardis Variation	10
Pirc Defense #4	9
Sicilian Defense	9



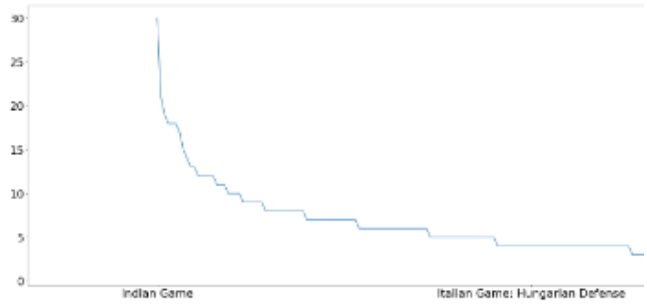
Black's preferred opening from 1000 to 2000 rating:

Sicilian Defense	349
Van't Kruijs Opening	344
Sicilian Defense: Bowdler Attack	289
French Defense: Knight Variation	271
Scotch Game	269



Black's preferred opening above 2000 rating:

Indian Game	30
French Defense: Knight Variation	21
French Defense: Exchange Variation	19
Sicilian Defense	18
Sicilian Defense: Old Sicilian	18



Through all of this analysis and graphing, I decided that starting with Van't Kruijs opening was a safe bet when giving players openings to try. Van't Kruijs is popular among both colors below 2000 rating. I do think that it is important to note that when you get above 2000 rating, Van't Kruijs is not seen. These players prefer to play either Indian Game openings, French Defense openings, or Queens Pawn Game openings. I believe this to be so because higher rated players are able to play more complex openings. This reflects what I thought before starting this project. Thus, using all this information, I decided to give seven players openings to try and play with. These are people that I knew personally and it was all conducted online due to the pandemic.

I started out this 'trial' by giving the players Van't Kruijs Opening to play against a low rating AI and chess.com. More information on Van't Kruijs Opening (technique specifics) can be found here: https://en.wikipedia.org/wiki/Van%27t_Kruijs_Opening this page also states that Van't Kruijs he is not popular among the higher rated players, which reflects what I determined in my initial analysis of the data set. I decided to do it this way because it was easiest considering the pandemic as well as chess.com having very well-structured AI per rating level. To be noticed that these players were from 1000 to 1500 rating. I took in their feedback into an Excel spreadsheet and I used that to make these determinations on Van't Kruijs Opening and giving players openings in general. Using Van't Kruijs made the game easier for most players, but the higher rated players (closer to 1500) stated that they would prefer to play their own way rather than using only one opening all the time. They stated that it impeded their ability to play the game the best they knew as when playing against specific openings, they knew to play a different opening.

I decided then to move on to giving the players the Queen's Pawn Game Opening. More information on the Queen's Pawn Game Opening (technique specifics and adaptations) can be found here: https://en.wikipedia.org/wiki/Queen%27s_Pawn_Game Given the same conditions as the last 'trial', the players began using this opening against the chess.com AI. Most players stated that this opening was easy to use, and simply feel like an adaptation of Van't Kruijs Opening. this made it easy to see why these two openings are very often used at lower ratings. Similar to the first 'trial', higher rated players still preferred to do their own openings and not stick strictly to one opening.

When asking them which opening they would prefer though, a majority of players said Van't Kruijs. The overlying reason for this was that it was easier to transition from the start of the game to later points in the game due to the opening. It allowed the Queen to come out from a diagonal we should apply more pressure to the center of the board compared to when using the Queen's Pawn Game, where the Queen could only come out if you moved a second piece after the first piece. The Queen could come out directly vertically but coming out at a diagonal was much better as it could go farther and could apply more pressure to the center of the board in different ways.

I finally ran a few simple AI simulations to see whether my data from the initial Lichess dataset was correct. These simple simulations determined that white does indeed have a greater chance of winning the match, and that if a player learns openings, they have a higher chance of winning the game period I did not implement openings into my AI, and they were much worse off with transitioning into later game states as there was no structure to the game. Although my AI was very flawed, I think it does give a good perspective into the Lichess data being correct. It also helped me reaffirm that the AI I was using from chess.com was most likely a good baseline to give versus other human players.

Given all of this data, I believe to have determined that lower rated players given an easy to use and understand opening, will succeed more and hopefully stick to the game of chess more often. I have also determined that players of higher rating prefer to vary their openings based on the game state and do not like to stick to 1

opening no matter what. This means that newer players should take easy to learn openings such as Van't Kruijs and apply to the game until the point where they understand chess at a deeper level and can expand to more openings.

Discussion

There are few lessons that I have learned during this project and I believe that these lessons I will carry on towards my future projects. The first is that even though you can statistically determine something, you need to be able to apply to real world situations for it to make sense. the second is that a conclusion from an analysis is a good way to measure real life expectations and should be weighted greatly when considering outcomes of real-world situations. One of the final things that I learned was that you think of more stuff to test once the project is over, and if you are able to realize these things during the project, your analysis will be much stronger. I realized this while writing this paper. When writing out the last section in my interpretation of the data, I realized that I did not give players a harder opening to compare to the use of an easy opening. This is because I assumed that the players would already not know how to do the harder opening. I should not have done this, and I wish I had realized this earlier.

Some things that went well included searching for information or deed on topic and researching information that might not specifically pertain to the topic, but potentially the game of chess as a whole. Researching background information helped me learn what has already been done in terms of chess opening theory, and it also reaffirmed my beliefs that there are not many ways for new players to learn what openings are best for them. It also helped me learn that new players are much more important to the game of chess than I realized.

For future work, I think the biggest thing to do is to plan more for my evaluation methods. I want to have more concrete ways to determine whether the data and analysis I have done is truly correct. This reflects in my project where I said previously, I did not give the players a hard opening to do to compare to the easy opening. I think spending more time on the evaluation methods would make my projects in evaluations much stronger in the long run.

Conclusion

Chess is very complicated game, and getting new players too enjoy and keep playing the game is somewhat difficult. To hopefully encourage players to keep playing and make their experience with the game more fun, determining a way for them to use specific openings tailored to them is important period that was the goal of this project.

My key tasks all revolve around determine the best opening for a player given there rating as well as their starting color. I determined what the rating brackets should have been for analysis, the distribution of played games in the dataset, and the preferred openings of players considering their starting color and skill bracket. My final key task was to create an AI that would hopefully reflect some of the key findings I found considering starting colored win percentages in the Lichess dataset.

My findings were relatively simple, newer and/or lower rated players prefer having an easy-to-understand opening given to them. This is in contrast to higher rated/more experienced players who prefer to use openings they know will be best suited for them in the game. I was also able to determine through the AI that the winning percentage numbers from the largest database or reflective of AI games. These findings culminate too hopefully having newer and/or lower rated players continue to play the game, as they are one of the main driving factors to chess succeeding as a thriving game in the future.

References

[1] Wikipedia. 2020. Shannon number. (November 2020). Retrieved December 13, 2020 from https://en.wikipedia.org/wiki/Shannon_number

[2] John Carl Villanueva. 2018. How Many Atoms Are There in the Universe? (April 2018). Retrieved December 13, 2020 from https://www.universetoday.com/36302/atoms-in-the-universe/

[3] Wikipedia. 2020. Chess opening. (December 2020). Retrieved December 13, 2020 from https://en.wikipedia.org/wiki/Chess_opening

[4] CCRL team. 2005. CCRL 40/15 Rating List — All engines (best versions only). (2005). Retrieved December 13, 2020 from https://ccrl.chessdom.com/ccrl/4040/

[5] Peter Doggers. 2020. New AlphaZero Paper Explores Chess Variants. (September 2020). Retrieved December 13, 2020 from https://www.chess.com/news/view/new-alphazero-paper-explores-chess-variants

[6] Paul van der. Sterren, Graham Burgess, and Paul van der. Sterren. 2011. *Fundamental chess openings*, London, England: Gambit Publications.

[7] Chess.com Team. 2018. The Best Chess Openings For Beginners. (July 2018). Retrieved December 13, 2020 from https://www.chess.com/article/view/the-best-chess-openings-for-beginners

[8] Mitchell J. 2017. Chess Game Dataset (Lichess). (September 2017). Retrieved December 13, 2020 from https://www.kaggle.com/datasnaek/chess

[9] Seberg Ludens. 2014. FICS Games Database - Download. (2014). Retrieved December 13, 2020 from https://www.ficsgames.org/download.html

[10] Andreas Stöckl. 2019. Writing a chess program in one day. (April 2019). Retrieved December 13, 2020 from

https://andreasstckl.medium.com/writing-a-chess-program-in-one-day-30daff4610ec

[11] Wikipedia. 2020. Van't Kruijs Opening. (December 2020). Retrieved December 13, 2020 from https://en.wikipedia.org/wiki/Van't_Kruijs_Opening

[12] Wikipedia. 2020. Queen's Pawn Game. (August 2020). Retrieved December 13, 2020 from https://en.wikipedia.org/wiki/Queen's_Pawn_Game

[13] University of Boulder. 2020. Honor Code. (April 2020). Retrieved December 13, 2020 from https://www.colorado.edu/sccr/honor-code

Appendix

"On my honor, as a University of Colorado Boulder student I have neither given nor received unauthorized assistance."

Sahib Bajwa

As I worked on this project alone, all the contribution for the project was my own. Any information received from outside sources was cited/credited.