DSC520 Final Project: Chess Dataset

Isabella Sturm

2021-03-06

When I was looking at datasets on Kaggle, I was hoping to find something I was interested in that would also be good for data analysis using the tools and techniques we learned in this course. I narrowed it down to three datasets: Video Games, Wine, and Chess. I ultimately choose the Chess Dataset because I thought it was the most interesting and I had just finished the Queen’s Gambit on Netflix and was feeling excited about the topic of Chess. A second reason I choose this dataset over the other others is because, at least compared to the video game dataset, I felt I had a clearer idea as to what questions I wanted to answer from the data provided. While I had good questions for the wine dataset, I had more questions than I felt I could answer from the data I could find.

Dataset: <https://www.kaggle.com/datasnaek/chess>

**Summarize the problem statement you addressed.**

Using the Chess Dataset I found on Kaggle, which was originally pulled from games on Lichess.com, I wanted to look into the factors that affect the outcome of a chess match. Looking at variables such as player ratings, number of moves, and outcome (mate, resign, out of time, resign), I wanted to find out how these factors affected who won the chess game. While white inherently has an advantage over black since they start, I wanted to see if a player with a lower rating starting as white was still more likely to win against a player playing black who as a higher rating. I was also curious if player’s rating is correlated with how quickly (in how many turns) they are able to finish the match.

**Summarize the methodology employed. Why do you think your method is appropriate? You should refer to the literature.**

I think in the analysis I did, the scatterplots show the greatest insights. This dataset would have been better represented with a binomial distribution model, if we ignore draws because there are two general outcomes – white either wins, or they lose. I performed a linear regression and while the summary of the linear model seemed to show there is significance in the variables I choose: ratings and turns, the model was really a poor fit and the predictions were not very accurate.

**Summarize the interesting insights that your analysis provided.**

There are some interesting insights in my analysis, particularly from looking at the plots. In general, a player with the higher rating will win regardless of who starts. The larger the gap in player ratings, the quicker the game will end. There doesn’t seem to be a clear pattern between who will win and the number of moves or the length of the game.

**Summarize the implications to the consumer (target audience) of your analysis.**

For a chess player looking to figure out how they can secure a win, the best way to do this based on my analysis is to ensure they are the better player in the match. Looking at the scatterplots, the player with a higher rating is more likely to win the game. While white inherently has an advantage as the starting player, if black is a better player (has a higher rating) they are still more likely to win.

**Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.**

I think the scatterplots I used provided useful insight into the data. However, as mentioned before, I used linear regression to create a model and make predictions but that was not the right choice for this data. To do better analysis on this data, something like a binomial distribution would be better to represent the data when looking at the factors that determine if white will win or lose a game. To build upon it, machine learning would be a great option. This dataset includes the moves and the openings used and text analysis on the moves would be great for prediction which moves or openings would be the best options for a greater chance in winning a chess match.