Win-Loss Prediction Based on Opening A Player Chooses for a Game of Chess

Akshay Jain, Nidhi Tholar, Sashank Pidur Kuppuswamy, San José State University

October 2021

Abstract

Introduction

The Data

We utilized a data-source[1] that provided thousands of games with respect to an Opening. We picked 15 opening/variation for our analysis. The data files were in text format (.txt). We transformed the data into a dataframe with the help of python string manipulation and regex. We also manually created a python dictionary that contained a mapping of an Opening to the move pattern.

Openings we chose: Four Knights, Caro Kann Classic, Qid4e3 (Queen's Indian,4e3, Sicilian Rossolimo, Sicilian dragon other6, KingsGambit, RuyLopezMarshall, ScotchGambit, GiuocoPiano, Reti2b3, Caro-Kann2c4, Caro-KannAdv, Nimzowitsch-Larsen, Caro-KannEx, Modern.

Missing values: The missing values were high for Elo rating of Players. Since it's an important factor in the analysis, replacing these missing values with mean or some other value, might have harmed the analysis. Hence, we decided to drop the rows containing missing values.

Elo Rating: We considered only those games where Elo Rating of both the players were above 2000

Openings: Though the data files were for a respective opening, there were some errors. We dropped the rows where the opening didn't belong to the 15 openings we chose.

Methods

Decision Tree: Since we wanted to analyze the amount of influence an opening had for a game's result, we decided to use Decision Tree.

Decision tree is a supervised learning method used for classification. It is used to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The library considered for the implementation is from sklearn.

Decision Tree Trained on Opening and Player's rating Here we considered features of opening, player's Elo rating and the result of the game to train the model. We utilized 80% of the data to train and the other 20% to test. Since Decision Tree doesn't work with categorical data, we encoded the opening name with numbers. We built the decision trees varying in depths to determine the outcome of the data considered. As the game of chess is a complex game and a game's result is affected by lot of factors other than opening. The classifier model's accuracy supported this inference. The accuracy we obtained with depth 12 to 40 was around 50%. The accuracy of the model showed that it is difficult to predict the game result based on opening and player's rating.

Decision Tree Trained on First X moves, Opening and Player's rating Since an opening would cover an average of 5 to 10 moves, we decided to extend the number of moves we considered to explore the effect of moves on the games result. In this approach, we selected first N moves from each record of games where first N moves represented a both the player's moves.

We used a python dictionary, where the combination of moves represented in a key, and the value marked against this was a list of games that had same first N moves. We built the decision tree for first 10 and 20 moves, but the accuracy of the model was not observed to improve.

Comparisons

Example Analysis

Conclusions

References

[1] https://www.pgnmentor.com/files.html#modking