Executive Summary

Florin Andrei

5/5/2021

In preparation for an upcoming chess tournament, we need to help the home team maximize their performance. We've analyzed a large public data set of chess games, using statistical models that attempt to gauge the odds of winning based on criteria such as ELO rating, the opening strategy, the color of each player, etc. Based on the models, we are going to advise Player A of the home team on the best opening strategies in two key games.

Data set

The data set used to train the models is the Chess Game Dataset (Lichess) set on Kaggle:

https://www.kaggle.com/datasnaek/chess

The data set contains over 20,000 games collected on the public server lichess.org. The parameters describing each game include:

- ELO ratings of players
- the opening strategy
- game length and duration (moves and time)
- outcome (who won)

The key parameter here is the opening strategy. Some of the most popular openings in this data set are listed below, along with the number of times they occur in the data:

```
##                                    opening_name count
## 1                              Sicilian Defense   349
## 2                          Van't Kruijs Opening   342
## 3              Sicilian Defense: Bowdler Attack   290
## 4              French Defense: Knight Variation   260
## 5                                   Scotch Game   254
## 6 Scandinavian Defense: Mieses-Kotroc Variation   247
```

Analysis

We've employed two different models to analyze the data: logistic regression, and a neural network.

The data set was cleaned, removing from it games that employed rarely-used openings - only the top 39 most popular openings were kept (each occurring in at least 100 games). We've also removed the draws from the analysis - we're aiming to win.

The clean data set used for the analysis contained 6220 games, which was split 75:25 between a training set and a testing set. The training set was used to train the models, while the testing set was used for performance evaluation.

We define model performance as the ratio between correct predictions (using the testing set) and the total number of predictions. The goal of this process was to produce the best performing model, and then use that to recommend the best strategies in key games.

The performance of the models trained on the training set was:

- logistic regression: 67.7%
- neural network: 67.9%

The neural network performs slightly better, and was therefore chosen to make recommendations.

Results

Scenario

Our player has an ELO rating of 1500. They will play two key games:

- as white against a stronger opponent (ELO 1600)
- as black against a weaker opponent (ELO 1400)

The model assumes games of average length (moves) and duration (time). For each game, the model was used to generate predictions (who wins) for each opening strategy in our training data set. The openings were then ranked, keeping at the top the ones that maximize the odds of winning for our player.

Playing as white

When playing as white against the opponent of ELO = 1600, the model predicted these 10 top openings as the best for our player (the full list is included as a spreadsheet in the archive, see the file `best_white.csv`):

```
##                                           opening  win_prob
## 1                                    Queen's Pawn 0.4501371
## 2  Scandinavian Defense: Mieses-Kotroc Variation 0.4377557
## 3                                 Pirc Defense #4 0.4295530
## 4        Sicilian Defense: Smith-Morra Gambit #2 0.4198934
## 5                                    Italian Game 0.4196096
## 6       Queen's Gambit Refused: Marshall Defense 0.4164996
## 7                               Caro-Kann Defense 0.4140233
## 8               French Defense: Knight Variation 0.4138846
## 9                         Queen's Gambit Declined 0.4131132
## 10                               Philidor Defense 0.4128466
```

The estimated likelihood of winning for white is shown in the column `win_prob`. Since our player plays white, we obviously want to maximize this parameter.

All likelihoods are predicted below 0.5; this is likely due to the ELO handicap for our player.

Playing as black

When playing as black against the opponent of ELO = 1400, these are the best opening strategies recommended by the model (see full list in `best_black.csv`):

```
##                                   opening  win_prob
## 1                    Van't Kruijs Opening 0.3130230
## 2                            Giuoco Piano 0.3193277
## 3          Sicilian Defense: Old Sicilian 0.3266225
## 4                        Sicilian Defense 0.3342153
## 5                    Scandinavian Defense 0.3366948
## 6                            Owen Defense 0.3393778
## 7  King's Pawn Game: Wayward Queen Attack 0.3432906
## 8        Sicilian Defense: Bowdler Attack 0.3433769
## 9                             Scotch Game 0.3503893
## 10                            Indian Game 0.3510118
```

Since `win_prob` indicates the likelihood of white winning, and our player plays black this time around, the best strategies minimize the likelihood.

All `win_prob` values are below 0.5, which makes sense, since our player, playing black, is stronger.

Notes

The recommendations could be improved by taking into account different game lengths and durations.

The models do not look at all at the actual moves played in the training data set, though that information exists in the original data. From this perspective, this analysis only looks at metadata (anything but the actual moves).

We've made recommendations solely for our local chess team, but the models could predict outcomes for games between any players - we just have to feed the appropriate parameters (ELO, etc) into the models.