CSCI 3360 Data Science I Course Project Proposal
Project Team I

Ayush Kumar, Chloe Phelps, Jason Furdui, Annie Lim

November 06, 2020

The dataset we are choosing to work with is a collection of nearly 20,000 chess games played on the chess website known as lichess (kaggle link: `https://www.kaggle.com/datasnaek/chess`). This dataset includes many features from different games including the ratings of the various players, the opening played, how long the game lasted, whether or not the game was rated, and the time control played. It also includes the standard algebraic notation for all the moves that were played in the game.

Our standard analysis will include plotting the winning percentage given the rating difference, based on the opening played, and time control. We can also create various columns of data from the game PGN strings analyzing if a player wins a queen, trades their queen, or loses material in the opening. The overall data analysis tends towards this being a classification task, with the response variable being predicting the winner of the game.

We propose the following models for classifying the outcome of the games:

1. A Decision Tree Model
2. Logistic Regression
3. KNN based on Dataset with Reduced Dimensions
4. Naive Bayes Classifier

The evaluation metrics for each of the methods will include accuracy, precision, recall, and f1-score. The initial analysis and feature engineering will be done by all members of team in collaboration over zoom. Then each individual member of the team will implement and evaluate one of the models. After models are implemented and evaluated the group will reconvene to aggregate the results and generate the final report and presentation.