**Decision Trees**

The first feature returns the player whose piece is in the bottom left corner of the board. This feature is not very useful because having a piece in the bottom left corner of the board is not something that should determine if a player wins. This feature was implemented because it was required by the project guidelines.

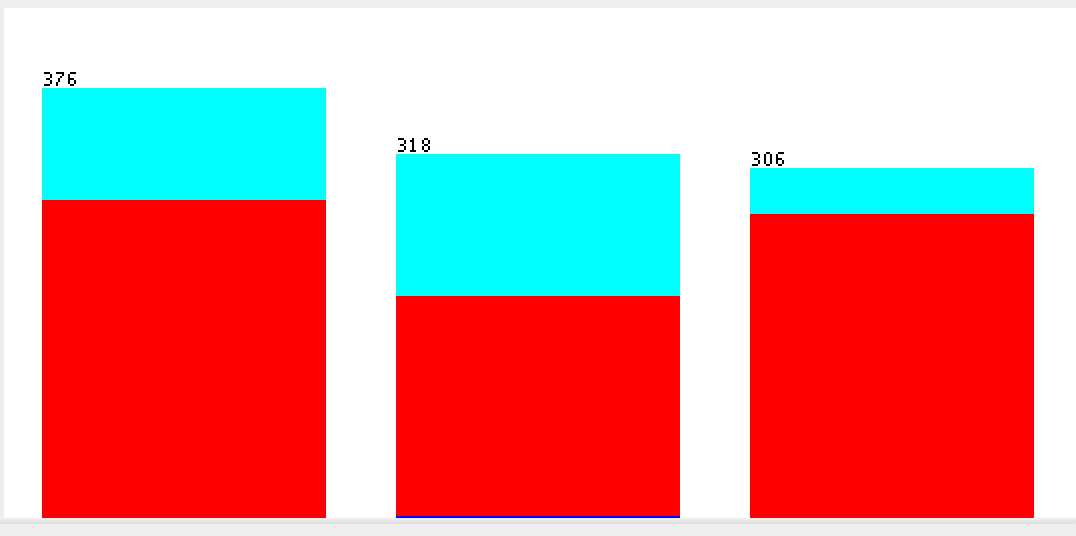
The second feature returns which player has more pieces that aren’t at the edge of the board. This feature should be useful because when a piece is at the edge of the board it has a lot less potential to be connected to. In general a piece toward the center would have a higher chance of being used to win than a piece on the side. This feature predicts that the player with the least pieces at the edge will win.

The third feature calculates a heuristic value based on how close all of a players pieces are to the center. This feature is similar to feature two because it also values pieces that are away from the edge of the board. This feature goes beyond feature two because each piece closer to the center has more weight in the heuristic calculation. The heuristic is calculated by adding one for every piece in an edge column. For every column closer to the center, a piece in that column is worth twice as much as a piece in the adjacent column closer to the edge of the board. Therefore, the number of pieces in a column is multiplied by 1, 2, 4, 8, 4, 2, or 1 respectively. The total value of all of the columns is evaluated and the player with the highest total is returned by the feature.

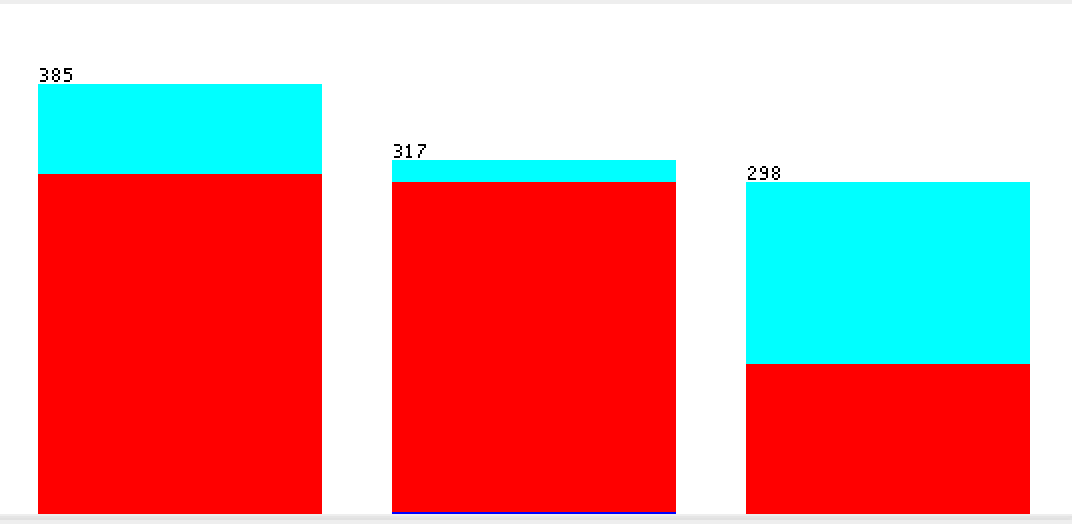
The fourth feature uses the same heuristic as feature three, but instead of returning whichever player has a higher heuristic, the feature returns the value of player one’s heuristic minus player two’s heuristic. This feature therefore returns a continuous function that represents how much of an advantage one player has over another because of a more centered positioning. This feature’s usefulness will be compared to the previous feature’s usefulness to determine whether it is better for this data set to use a continuous value to represent positioning advantage, or a categorized set of values.

The fifth feature tests who has “top control” of the columns of the board. “Top control” is defined as whoever has the highest piece on any given column. For every column where player 1 has top control, 1 is added to the top control variable and for every column that player 2 has top control, 1 is subtracted from this variable. Columns with no pieces in them do not count for top control at all and do not affect the top control score. When doing the first project, we found that often that your ability to make advantageous moves was dependent on how much control you had over the top edge of the board because you’re only able to make plays that will become new edges on the top, interacting with the old top of the board.

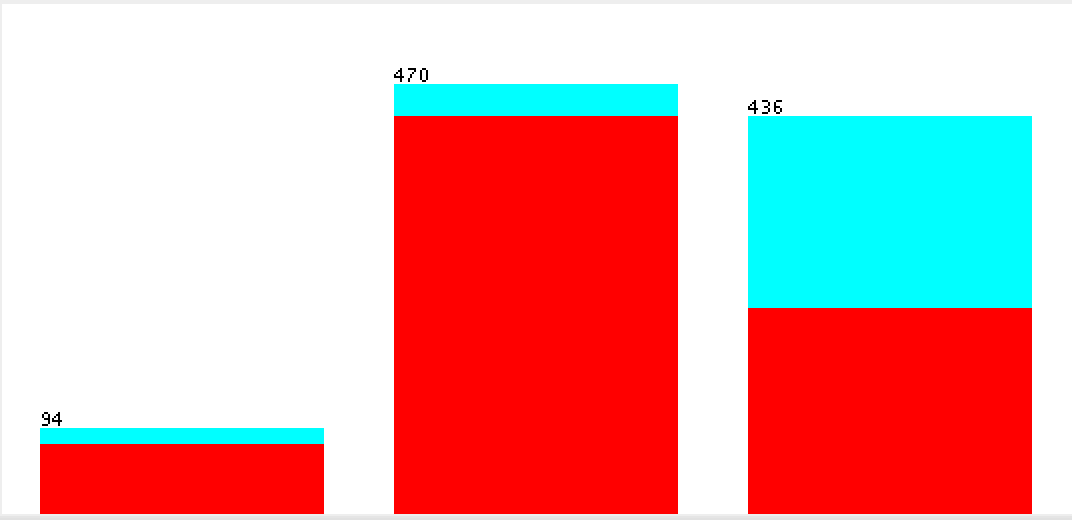
To determine which features are better than others we can look at how well each feature predicts the winner of a board. To do this we show a histogram of every value the feature returns with how many times player 1 or player 2 won a game at each value of the feature. Player 1 won about 75% of its games, so if the histogram shows that at each value of the feature, about 75% of the time player 1 won, then we know that feature is useless. Speaking of useless features, below is a graph of feature 1.



As you can see from the data, every value of feature one leads to only a slight change in the winrate of plaer one, ranging from around 65% to 85% instead of 75%. There is only around a 10% change in the winrate of player 1 based on the value of feature 1 so this heuristic is not very good at predicting the winner. Feature 2 on the other hand has a larger change in the winrate of player one based on its value. The graph of Feature 2 is shown below.

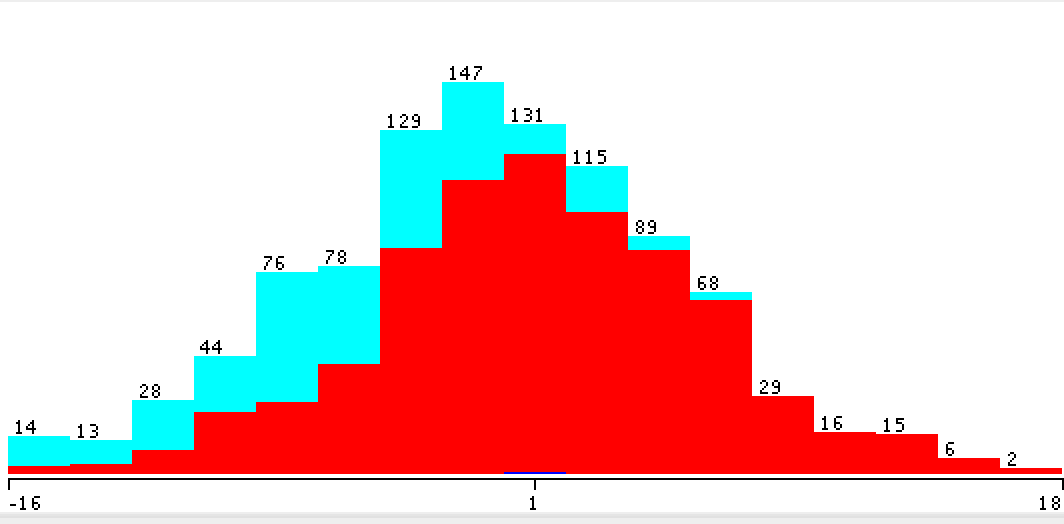


As shown by the graph, when feature 2 puts a data point into one category, player 1 wins 95% of the time, and when feature 2 puts a data point in a different category, player 1 wins 45% of the time. Both of these changes in the winrate of player 1 show that feature 2 can predict the winner much more accuratly than feature 1. The issue with feature 2 is that it has one category that it cannot predict the outcome from accuratly and it places over a third of the data into this category. Feature 2 is good, but only for two thirds of the data. Feature 3 on the other hand can almost as accuratly predict the outcome of many more data points than feature 2 which is shown by the graph of feature 3 below.

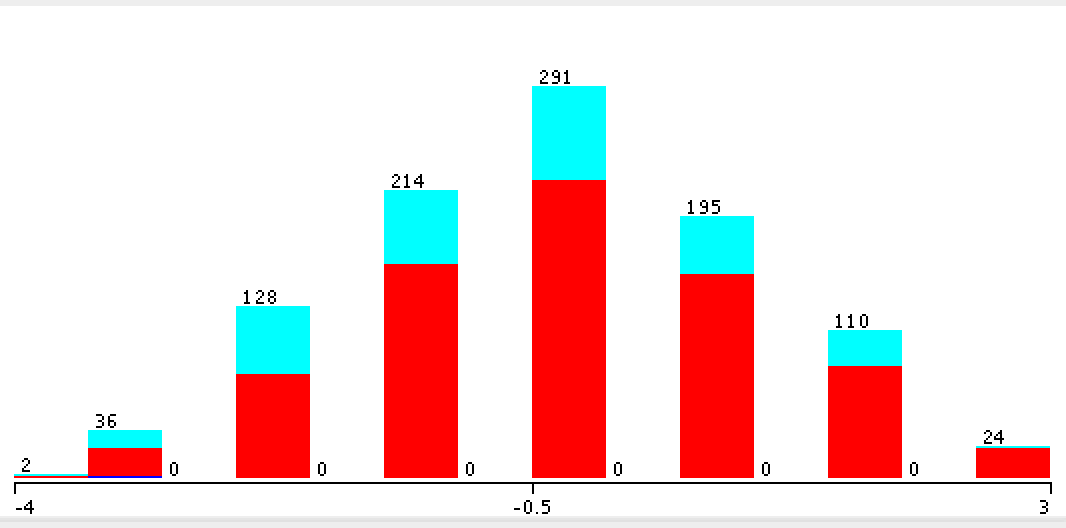


For each category in feature 3, there is almost a 25% change in the winrate of player 1 than if the predictions were random. Feature 3 can predict all of the data with this accuracy, wheras feature 2 could only predict two thirds of the data with this much accuracy.

Feature 4 creates a continuous output of values. The decision tree treats each value as a different category. Below is the graph of feature 4.



As shown by the graph, feature 4 has the potential to predict values more accuratly than any of the previous graphs. For some values feature 4 can 100% accuratly predict that player 1 will win, and that player 1 will win 15% of the time. There is only 30% of the data that feature 1 cannot predict accuratly but it can more accuratly predict the other 70% of the data than any of the other features. Feature 5 was not very good at predicting the winner and for every value of feature 5, there was not much change in the winrate of player 1 than if the winner was predicted randomly.



Based on the analysis of every feature, feature 4 is by far the most useful feature because it can predict wheather or not player 1 will win with more accuracy than any of the other features. The only place feature 4 falls short is it’s inability to accuraty predict 30% of the data. This is where feature 3 is usefull because it can predict the winner of all of the data almost as accuratly as feature 4 can predict the winner of 70% of the data. So the 30% of data that feature 4 cannot predict well can be predicted by feature 3, this will ensure all of the data is predicted with good accuracy.