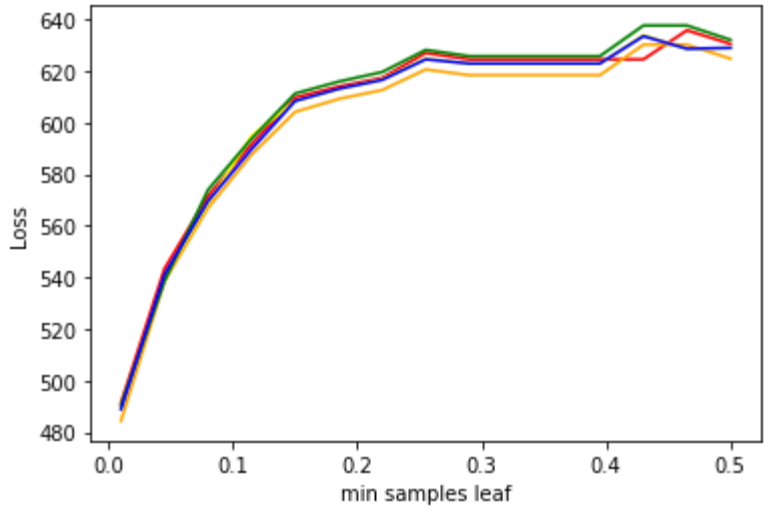
Madison Chamberlain

Below is my graph for prediction accuracy given different input values for the minimum number of samples per leaf. The different colors represent different train-test set ratios. The red line is when the train-test set ratio is 0.1, orange is 0.2, yellow is 0.3, green is 0.4 and blue is 0.5.



As you can see in the graph, all values of train-test set ratios have very similar accuracies, and they all follow the same pattern. For this specific run, the ratio 0.2 was the most accurate, but the difference between the accuracy for 0.2 and the other ratios appears to me so minor, that it is likely other runs would yield different most accurate ratios. With that being said, the accuracy is determined more by the value of the minimum number of samples per leaf rather than the train-test ratio. As you can see the loss is smallest as the minimum number of leafs per sample approaches zero. As the minimum number of samples per leaf increases, the prediction accuracy decreases. This makes sense because on the far right of the graph, the minimum number of samples needed per leaf is half of the samples. This means that the data can only be split one time, so the prediction is generalized to many data points, and thus not very specific or accurate. On the other hand, on the far left side of the graph where accuracy is high, the minimum number of samples per leaf is one, which means that the tree could potentially have a leaf node, and thus a different prediction for each data point. This will create a highly specific tree. Even the tree generated with a minimum sample of one needed to generate a leaf is not 100% accurate; this is likely due to overfitting. For all train-test ratios, the tree with the lowest value of minimum samples per leaf generates the most accurate predictions.

Min_samples_leaf is different from min_samples_split because min_samples_leaf guarantees a certain number of values per leaf node, while min samples split does not. An example of how this works is if we had two trees with 10 samples and for one tree set min_samples_split to 5, and for the other set min_samples_leaf to five. In the first tree, since there are greater than 5 samples, the data will split, but it may split where one leaf has one

sample and the other has 9, or ir could split evenly or anywhere in between. On the other hand, the min_samples_leaf scenario is guaranteed to have five samples per leaf. This may yield different accuracies given different datasets.