Supervised Learning: 5 Algorithms

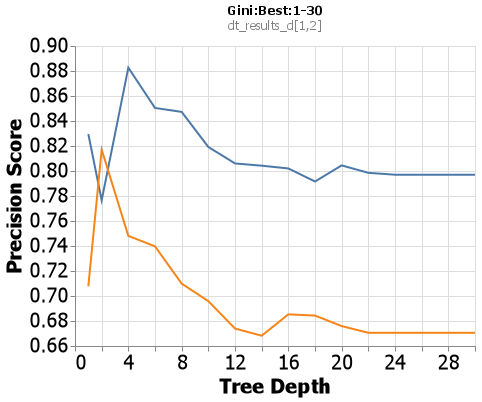
“Learn the rules like a pro, so you can break them like an artist” -Pablo Picasso

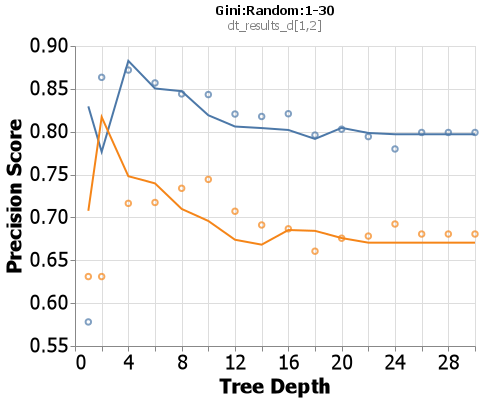
There are patterns and behaviors or “rules” to supervised learning algorithms. These must be understood before we can learn to “break them like an artist”. As we go through this paper, we will be exploring each of these models across 2 different datasets. These two datasets come from Kaggle and are called the “Product Subscription”, which we will call the 1st dataset, and “Bank Customer Churn” or the 2nd dataset. In order to evaluate how well a model is performing, we need to have a metric which is normally derived from a business case. For these two datasets we will propose a business case of “it is really bad if we predict more customers who churn/subscribe than those who didn’t” which would lead us to using the evaluation metric of precision.

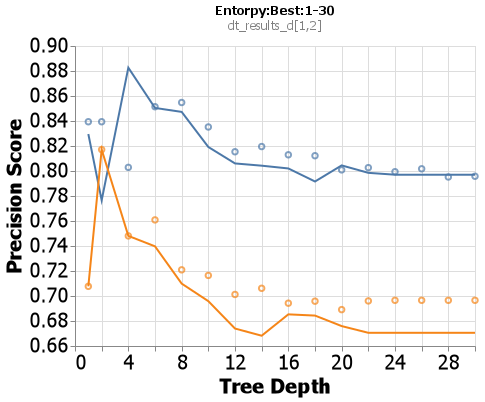
Now that we have defined our models, datasets and a performance metric, we can talk about the areas of analysis. We will start by examining the different parameters used in each of the models and how they tend to behave. Next, we will compare models within each dataset to determine which model was best at optimizing for our business case or the precision metric. This analysis was done in the python language and utilized the Sci-kit Learn (Sklearn), Altair, and Pandas packages and their dependencies found within.

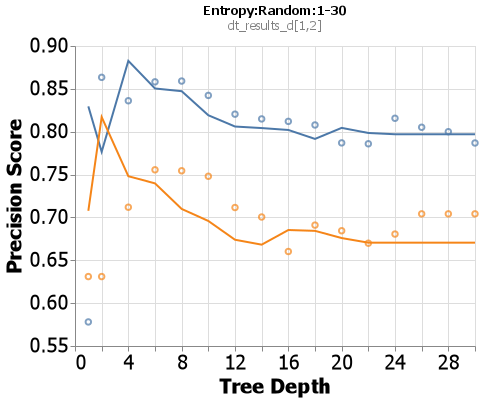
**Part 1: Model Behavior**

Decision Tree

When examining the Sklearn decision tree algorithm I chose to manipulate 3 different hyperparameters: criterion with values of gini and entropy, splitter with parameters of best and random, and max\_depth with parameters 1 through 30 by increments of 2. This allows us to look at how a tree with varying lengths can learn when it is constrained to a small network vs a large network, and how the other parameters affect this ability. The default parameter combination for sklearn’s decision tree is criterion = gini, splitter = best and max\_depth = None. Now max depth being set to None may not make sense, but it is essentially saying continue until all the final nodes are 100% one class, or until all the final nodes have 2 or less data points. With these default parameters being set we will treat the Gini:Best combination to be our base case with the blue line representing the first dataset and the orange line representing the second dataset and a line to denote the base case (see right).

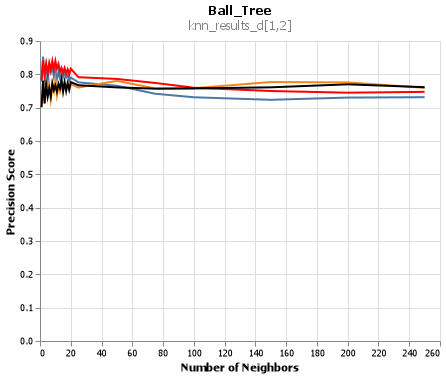
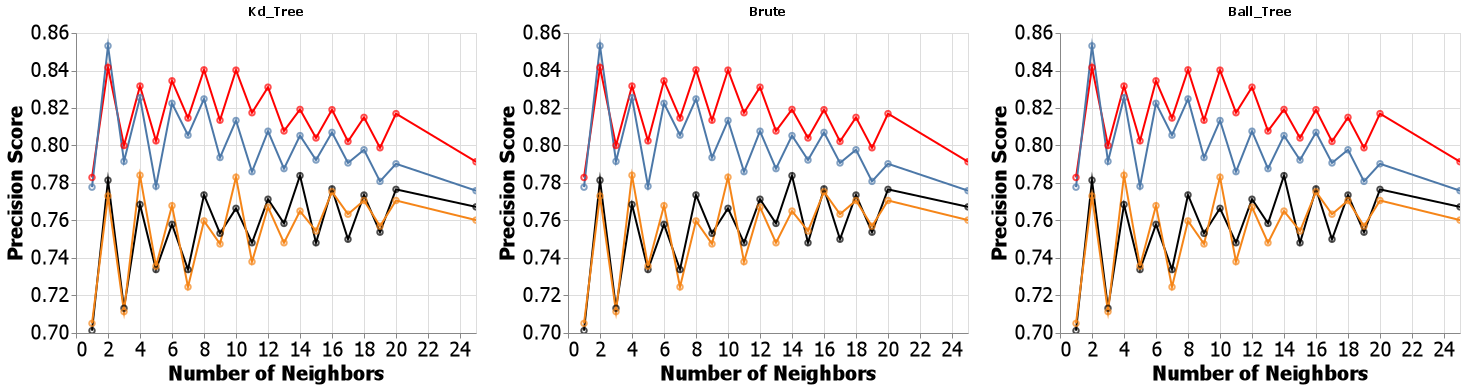
Now that we are familiar with our base case, we can put in another test case to compare, Gini:Random. When we look at depths 16 and on we see a little difference in the precision score, but no significant difference. When we look at 1-14 though, this is where the graph gets interesting. Here we changed the splitter parameter from Best to Random, this has caused the algorithm to not search for the best possible split on every single node. Once a node has been split well enough, it moves on. Within a depth of 1-6 we see a drastic difference, that is because the nodes on the trees haven’t had a chance to be optimized yet. But, when we look at 6-14 the Random splitter does a better job, most likely due to the better generalization of a random split then trying to go for the best gini impurity reduction.

On to a new criterion: entropy. When compared to the base case, we see the same plateau after a depth of about 12. But what differs is between 2&6, if you look closely, you will see entropy does a better job at being more generalized earlier on. This is because the entropy formula is and the equation for gini is . Due to the log in the entropy equation, it makes it more generalized because when you take the log of a big number it becomes much smaller, thus leveling out the drastic differences and providing a better generalization.

For this last combination we see a much more sporadic chart. There isn’t a clear plateau like in the previous 3, and the beginning starts off a little bit more extreme as well. This is most likely caused by the fact that we are using a random splitter so our entropy impurity values aren’t optimized, thus giving us inconsistent types of splits. This adds on to the idea that entropy is trying to generalize, but it is being fed bad splits by the random splitter.

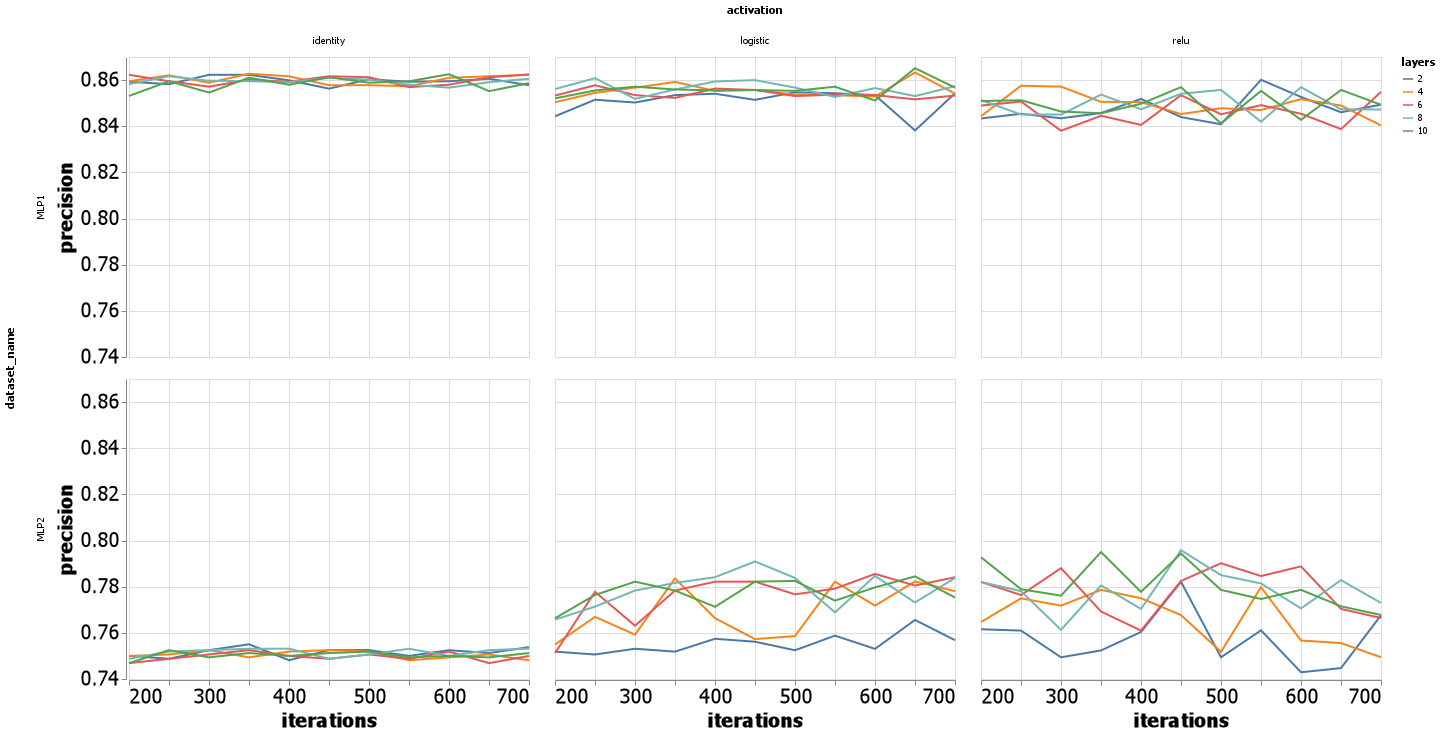
K-Nearest Neighbor (KNN)

Following the same pattern, we need to establish our base case. The hyperparameters I chose to examine are n\_neighbors, algorithm, and metric. The default for algorithm and metric are auto and Minkowski respectively. Even though I did not iterate over the Minkowski metric specifically, the Minkowski equation takes in a parameter p where when p=1 it is a Manhattan distance and when p=2 it is a Euclidean distance. Due to this fact, our base parameters will be auto and both Euclidean and Manhattan. Like decision tree, for the first and second datasets the lines will be blue and orange for Euclidean and black red and black for Manhattan respectively.

 Before we go in, I want to show a behavior that I found interesting. To the right is a chart ranging from 1 neighbor to 250 neighbors. As you can see, around 40 neighbors the graph begins to plateau. This was specifically for the ball\_tree metric, but all of the other metrics performed in this same manner. This is interesting because it tells us that anywhere past 30-40 neighbors and the information gained will only result in roughly a .01-.02 change. Even though this is specific to our model, it is interesting to know that there comes a point where no matter how many neighbors you want your model to look at, the model will perform the same. It isn’t work the extra computation time to search over the extra 220 neighboring points.

I know I said we would be examining the precision metric, but above is a graph that plots the precision metric against our other KNN models. I know it is extremely fishy, but I’ve checked my code many times over and I could find any reason this would be happening due to an error on my side. This leads me to believe the data was well clustered, which I would try to show but we cannot enter the 23rd dimension. It makes sense why this would occur, the ball\_tree and kd\_tree algorithms are very similar in the sense that they both work like little decision trees. The ball\_tree algorithm creates hyperspheres around certain clusters of points and tries to make those hyperspheres as pure as possible, then from within each sphere more spheres are formed while trying to maximize the distance between each sphere (hence the name “ball”). The kd\_tree splits the data on the mean of each feature for each class, thus creating something that might look like a bunch of boxes in a multi-dimensional space. The brute algorithm goes through the data and attempts to make predictions only by guessing and checking. It is a little rudimentary, but that is why they call that parameter “brute”. If the data was well partitioned and separated, it makes sense that these 3 algorithms would come to the same conclusion. If anything, this would tell you the model you have created has examined all possible outcomes and is tuned very well.

Artificial Neural Networks (ANN)

Below in the chart you will see how the variance of each of the lines increase with different activation functions. This was something that was seen across both of the datasets, starting with the identity function the variance is very small bouncing between .85 and .87. Then as we go through the logistic and relu functions we see the variance bounce between .75 and .8. In all reality the difference we see here isn’t big, but it is the spread of variance that makes it interesting. The identity function is a linear activation function, thus making it proportional to the input received and making the predicted outputs more normalized. The logistic function shows us that as the value of x increases it slowly approaches 1 and as x decreases it slowly reaches .5. The relu function is a value of 0 if the input is less than 0 thus essentially canceling out that value. So, it completely disregards all other inputs that are less than 0, thus would create more variation in the outputs if a portion of the inputs are less than 0. This wouldn’t be a problem though if all your inputs are greater than 0. With these 3 activation functions identified, we can begin to decern why the graph is behaving the way it is. When we append multiple layers to the identity activation function, any combination of linear equations is going to be linear, thus not allowing any of the nodes to behave differently and creating a constant small variance. As for the logistic and relu functions, they allow each layer to take on a nonlinearity behavior. This allows for some layers and nodes to go higher much quicker and some to stay stagnant. Once the output is reached for these two activation functions, the numbers would be across board as to what the prediction would be thus creating a wider range across the precision score metric.

Chart, line chart

Description automatically generated Since this is an iteration model, time is also something that should be examined. Below is the same chart from up above just with time as the y-axis. Because the identity function is linear, we don’t see any significant fluctuations in the training time across any number of layers. As for logistic and relu, it is not surprising to see their times fluctuate. Due to the logistic function working with exponentials, it is not surprising to see it grow quickly as the computation time needed for each node goes up at an ever-increasing rate. As for relu, we see it have a slight increase as the number of iterations increase, but not nearly as fast as the logistic function.

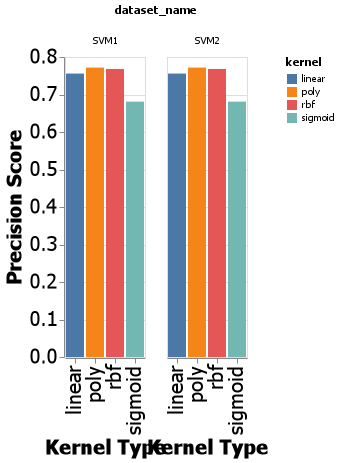
Boosting

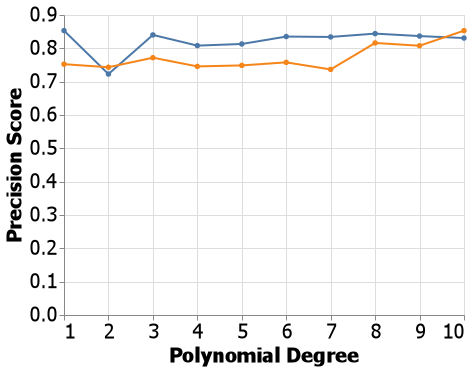
*Chart, scatter chart

Description automatically generated* For this algorithm I chose to examine the loss function, learning rates and the number of estimators within the boosting model. At first, when I started this analysis, I wanted to see what would happen when I do a wide range of estimators and I produced the chart below. It was interesting to find that even though they had different precision scores, there came a point in the number of estimators where the information gained was insignificant. So I went back and re-ran the algorithm to show more information from 0 to 50 as this seems to be where the charts leveled off. I also decided to decrease my max learning rate as a learning rate of 3 was leaping too far and looks like it got stuck in some local minima where the precision was very bad.

While exploring the results from my new tuning, I found that boosting does well very quickly. Only with a few estimators was I able to achieve a high precision score and stay there across any number of estimators and a learning rate of 1 or less. Once I reached a learning rate of 2 or 1.5 it was descending the gradient too quickly and would miss what most likely was the global minima.

Support Vector Machines (SVM)

 For the support vector machine, I chose to look at the kernel and degree parameters. Once I saw how the KNN model performed and the results were algorithm agnostic, I wondered how the SVM would do under different kernels. Even though they are all extremely similar, the sigmoid function stood out. Where the sigmoid function is essentially a two layer perceptron activation function, it creates values that are 0s or 1s which can give it the potential to over or underfit the data because it looses some of the complexity behind the data.

 Once I saw that the polynomial kernel was best suited for this type of problem, I chose to go back and look at what would happened when I select only the polynomial kernel but change its degree. With the blue line being the first dataset and the orange being the second, we can learn some information about the dataset by just looking at this graph. In the high dimensional space that the model for the first dataset lives in, it would be save to assume the data can be separated by only using a straight line. Notice how when the polynomial degree is “2”, our precision score drops significantly. When the highest degree is 2, it is impossible to make the line bend back down in the same relative direction before it started to bend back up again. Once we get into the 3rd order and higher it is possible to make the line quickly turn up and down to fit the data. We also see a huge jump in precision after the 8th order degree is reached for the second dataset. This tells us there is some complex relationship between many of these variables in which the data needs to be able to bend and curve at varying intervals and this can only be achieved when we work with a kernel of “poly and a degree of “8” or more with 10 producing some of the best results we have seen for the second dataset.

**Part 2: Model Cross Comparison**

Dataset 1: Subscriptions

Graphical user interface, calendar

Description automatically generatedAs mentioned earlier, the first dataset is based on the type of people who subscribe to a certain product. At this point, we have a long list of CSVs with the models and their hyperparameters with how they performed with various metrics to identify that performance. To the left is a table with all the best parameters for the precision metric. The Boost21 and SVM21 are the second versions of the boost and svm model I went back and ran more tests over because I realized the best parameters had more room for even finer tuning. The chart is ordered by greatest to least based on the Table

Description automatically generatedmetric column. From this we can see the decision tree was the best model for this dataset with a precision score of .88. To double check this I picked all of the best hyper parameters and ran them through a cross validation each with 10 folds focusing on general accuracy to get a broader view of how our model is behaving. To the right are the results. Seeing as how the standard deviation is only .01 or .02 I feel very confident that the values we are seeing are accurate and reliable enough to make these parameters our main parameters. As a special note, since the second versions of both the SVM and Boost models did better than the first version we will use those parameters in our final model.

Chart

Description automatically generatedChart, treemap chart

Description automatically generatedChart

Description automatically generated with medium confidenceAs we begin to cross compare the models and how well they performed with their best hyperparameters, we will be comparing them against the decision tree model as this model is the one that performed the best. But, before we can compare against the decision tree model, we need to examine it by itself. Below is an ROC curve for the decision tree model just to establish a base line. Even though this isn’t the best ROC curve out there, it is the one that did the best on our dataset. Even though we chose the precision metric for our analysis, we see that our decision tree model wasn’t well optimized for that metric, and it performed better with the recall metric. With a precision score of .88, this model will be hard to beat. Chart, treemap chart

Description automatically generated**Chart, treemap chart

Description automatically generated**

The first model we will explore will be the next runner up, our second revision of the boosting model. Its parameters were n\_estimators = 6, learning\_rate = .1 and loss = log\_loss, this gave us a precision score of .87 and a training time of .0069. Below you will see the confusion matrix, roc and precision recall curves. As to why this model did just a little bit worse, I think it can be found in the ensemble portion. Because we had so many models **Chart

Description automatically generated**going into the boosting model, they can pick up on smaller trends but that may not always be the best idea. Sometimes an ensemble model can pick up on a rule that may be specific to only a few, but then it is applied to the whole. Not the mention the boost model can create a tree that isn’t pruned by what the user knows to be the best limit. In our confusion matrix we see again that this model wasn’t well suited for the precision metric, as we performed better with a recall score. But, interestingly enough it did better for reducing the number of type 1 errors than the decision tree. Goes to show that more heads aren’t always better than one.

Chart, treemap chart

Description automatically generatedThe next runner up is our neural network. With parameters at layers = 10, iterations = 650, and activation = logistic this model comes in with a precision score of .86 and a training time of .14. Something I find interesting for this model comparison is in the confusion matrix. When we compare them the neural net was almost equally weighted in the precision recall balance. This makes sense because the neural network is designed to balance all outputs and make sure things are weighted properly to give us a very balanced model. This is good to know because if you are wanting a model that generally does well and has a high accuracy score, a neural net would be a good place to start.

Chart, treemap chart

Description automatically generatedOur svm model had parameters of kernel = poly, probability = True and degree = 1, the two metrics of time and precision were 1.84 and .85 respectively. Similarly, to the neural net, our confusion matrix for the svm seems to be very balanced as well. With a degree of 1, this helps me to further believe that this data is very well clustered and is best separated by linear where the clusters have low contamination.

For our last model on the first dataset we come to KNN with parameters metric = Euclidean, algorithm = ball\_tree, and neighbors = 2 with a precision score of .85 and a training Chart, line chart

Description automatically generatedtime of basically 0. This isn’t surprising though because we know KNNs to be lazy learners. Both the confusion matrix and roc curve are interesting compared to the decision tree. Below you will see how well the model did in predicting Chart, treemap chart

Description automatically generatedtype 1 errors. Out of all the models none of them did this poorly. The ROC curve as well is most bizarre, all the other curves smoothed out where this curve has a sharp turn. This tells me there was probably a cluster of points where most of them were 0, but there was enough 1s in there to bring down the score significantly. This is one of the drawbacks of a knn compared to a decision tree.

Dataset 2: Customer Churn

Attributions

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

Gladwell, Malcolm. Outliers: The Story of Success. New York: Little, Brown and Co., 2008