In machine learning, we use various statistical methodologies to create models that can in simple terms, perform pattern recognition. The models that we trained to perform one of two tasks, they either do what’s called regression analysis or Classification More simply, we are either attempting to predict or forecast a value or attempting to label or categorize the data. Regardless of the results, the primary goal of both forms is to estimate the relationships between a different variables or items within our data.

Within machine learning there are many different methodologies or models we can use. With each of these paths, ranging from simple mathematical calculations, such as: Linear Regression, to slightly more complex machine learning models like Ensemble and SVM, all the way to replicating the way the human mind works with technology like Deep Learning, Natural Language Processing, Generative Adversarial Networks. Don’t worry, our focus will stay on the simpler side of the machine learning world.

As we stated before, the model that we have current found to perform the best out of our selections is the Histogram-based Gradient Boosting Classification model, or hist-gradient boosting model. This model is from the Ensemble library within our machine learning tool kit, the Sci-Kit Learn Python API. So what makes this model better at classifications than the others. Let’s take a step back and talk about what the Ensemble is.

In mathematics, the Statistical Ensemble, or simply just ‘Ensemble’, is an idealization consisting of a larger number of virtual copies of a system, considered all at once, each of which represents a possible state that the real system might be in. In other words, we create many virtual copies of our system that then take a small sample of our data, then each of these copies perform run some calculations on there subset and then generation predictions. We then take all of those sperate smaller predictions and combined them to create a prediction that should theoretically be better than each of the single predictions on their own.

In machine learning, we utilize the Ensemble method to create multiple base models or base learners, also known as weak learners. Each of these weak models have poor prediction capabilities on purpose. This is done on purpose to make them faster and more constricted in their functionality. This way we are able to make small tweakers to their processes to generate different results. Some models will generate a large amount of bias, some will have large amounts of variance, as well as the vice versa to the others. However, all of these weak models will perform quickly, allowing us to do many iterations and combinations of setups. Once done, all of their results will be combined into a single, better model that will have the best parts of all the smaller models.

Ensemble models are categorized in types: Bagging, Boosting, and Stacking/Blending. In bagging, we create many different weak learners, all who each have a random sample of the training data. These are all ran at the same time in the same way and then are ultimately combined into a single model at the end. With Boosting, the process is more iterative, with each weak learner running sequentially, learning from its predecessor and becoming its successors foundation. This method works to improve upon itself and builds up reducing the errors with each iteration. Stacking/Blending is similar to bagging but instead of utilizing the same model to train each random sample, we use several different models that are trained independently, before being combined into a single model at the end.

Our model, the histogram based gradient boosting model, is of the boosting variety of ensemble. As its name implies it’s a slightly different from the standard Gradient Descent Boosting method. Similar to the standard version after each run of the weak learner in the model, we check the performance and compare it to what was predicted to happen, this difference calculation is known as the loss function. The model then adjusts its parameters for and passes them along to the next weak leaner in the sequence. One by one, run by run we improve the model, training it. In our version, our model will start convert our data groups of similar items them sort them into groups called bins, a concept known as a histogram (thus the other part of the model’s name’s sake). These bins of data make it much easier and faster for the computer to perform its calculations on. Thus, why our model is a much faster and efficient version of its original.