In the field of data science, we have the opportunity of having so many different tools available to us to tackle all types of issues and locate several kinds of patterns within a data set. Author Rajat Harlalka explains the process, “When you look at machine learning algorithms, there is no one solution or one approach that fits all. There are several factors that can affect your decision to choose a machine learning algorithm.” With entering this field, I think that it is one positive outlook that I always imagine. Since this field is constantly evolving, there is no one single way to complete every goal; you have the space to determine what you or a team feels is the best approach at a problem. Looking into the process no matter what you choose for an algorithm, you always want to understand what you data consists of as well as any possible constraints within it. In understanding the data, it would involve possibly cleaning your data and or augmenting it to best fit a model. Once you are able to describe and gain perspective on your data, you can then categorize the problem. In this step, it is categorizing the input as well as the output. For example, with categorizing the input (Harlalka, 2018):

If you have labelled data, it’s a supervised learning problem

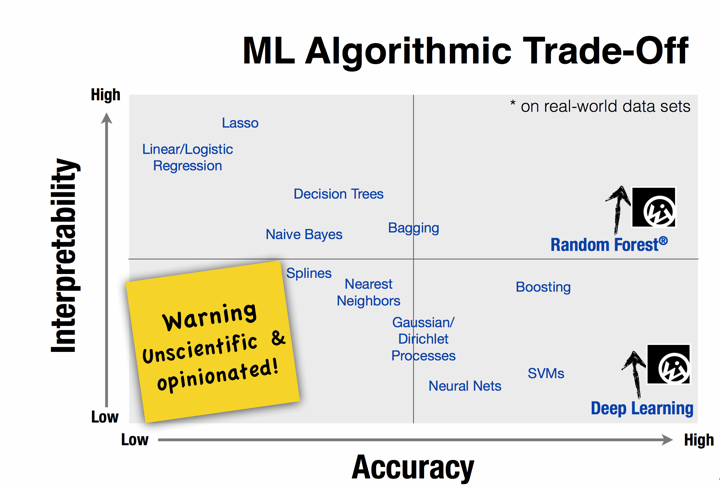
If you have unlabeled data and want to find a structure, it’s an unsupervised learning problem.

When it comes to categorizing the output, it follows the same idea that (Harlalka, 2018):

If the output of your model is a number, it’s a regression problem

If the output of your model is a set of input groups, it’s a clustering problem

Upon having all of this information, you can seek in discovering what the best machine learning algorithm is for you to use to reach your goal. For searching what the process of selecting a machine learning algorithm is, I came across a visual that looks at the give and take of selecting certain models to use.



Example of Machine Learning Trade-Off

By looking at this model, it shows how with some models you have the possibility of having more accuracy then interpretability.

In this quest to look for the most appropriate model, you would come across the hyper parameter problem that seeks candidate hyper parameters values that are expected to improve the prediction of the machine learning algorithm. In searching for these candidates to use for the parameters, you would use a grid search which is an exhaustive evaluation of discretization of the hyper parameter space on a regularly spaced grid. (Bermudez-Chacon, 2015) When you eventually decided on which value to use, it is important to understand that your choice can heavily impact its predictive power. When choosing a model and hyper parameter, it is a process that you cannot complete in a short time frame since you must understand what the goals are for you or a team and what the data involves as a whole.

References:

Bermudez-Chacon, R., Gonnet, G. H., & Smith, K. (2015). Automatic problem-specific hyperparameter optimization and model selection for supervised machine learning: Technical Report, 1–53. Retrieved from <https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/107673/eth-48308-01.pdf>

Harlalka, R. (2020, April 29). Choosing the Right Machine Learning Algorithm. Retrieved April 29, 2020, from <https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f>