Anatomy of Data Analytics, Machine Learning and Deep Learning -- Demystified – Part III

Once you know the Deep and Shallow Learning methods and how to apply them, you are ready to dive in. However, to apply these methods appropriately, one needs to follow a systematic workflow. Following guidelines could be of help.

Problem Description

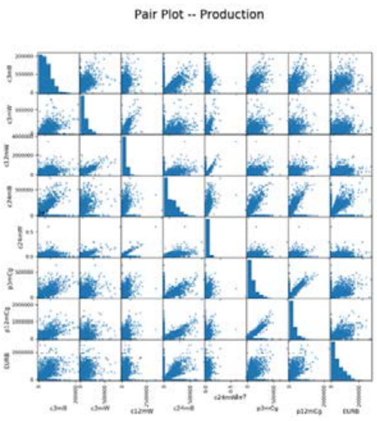
The most important part is to frame the problem. In this step a specific problem in the domain is identified and framed in a way such that the Machine Language techniques can seek a solution to the problem. The data related to these problems can be in tabular format, time series, textual format or may need speech and/or image recognition capabilities.

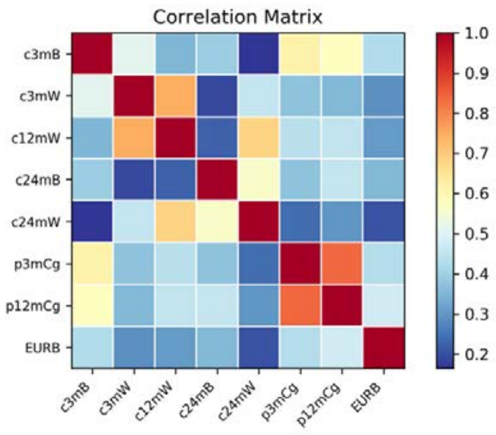
Feature Engineering

Next step in the workflow is to improve the data quality so that the Machine Learning techniques can truly derive the insights from the data and is not overly influenced by superfluous or less important parts of the data. Of course, this step needs the attention from Subject Matter Experts (SME). In this step, several aspects of data cleaning, filling missing data, identifying features and targets, combining explicit variables into compound/complex variables etc. are involved. Some of these are getting automated in the algorithms but others still need SME intervention.

Exploratory Data Analysis

Once a clean data set is available, next step is to make the features (input variables) as independent as possible i.e. they should not be correlated. If they are, they are not only redundant but also can negatively influence the ML algorithms to converge to non-insightful conclusions.





Model Selection

In this step, the practitioners need to select the appropriate technique based on the type of input (text, image, video etc.), kind of desired output (class/event, continuous), accuracy tolerance etc. Of course, there are variety of algorithms in the libraries but a knowledge of distinguishing them and appropriately applying algorithm for a specific problem need Data Science expertise.

Model Validation

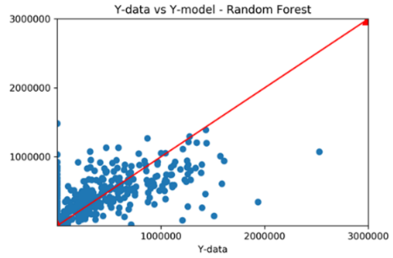
The step involves examining the quality of ML algorithm training. There are three common metrics for evaluating the goodness-of-fit are: 1. Average absolute error, or AAE 2) mean squared error, MSE and 3) pseudo-R2. AAE is defined as the average magnitude of difference between the true response and predicted response. MSE measures the average required difference between the observations and predicted values. Finally, pseudo-R2 is proportional to the variance of the response. One strategy is to use70 to 90% of the dataset to train the model and use the remainder 30-10% of the dataset to trial the model. The viability of the model as a predictor is gauged on the three metrics defined above. A different approach is to utilize k-fold cross-validation. Here, the training set is randomly split in k-different groups. Each of the k-groups is held out one at a time and the model is trained on the remaining k-1 groups.

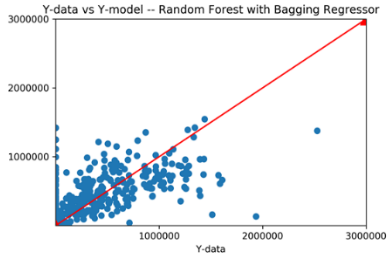
Parameter Tuning – Hyper Parameter Search

How do we know if the parameters used in the model are indeed the most optimal ones? It is difficult to select the optimal tuning parameters in the model that will lead to best goodness of fit. Therefore, an automated process is generally chosen where k-fold cross validation is performed on the sampled dataset and a final set of tuning parameters are determined using cross-validated RMSE. Finally, using these converged tuning parameters, the model is refit for the entire training set. Following is a sample list:

Parameter Name	**Parameter Values**
Activation	**Softmax, softplus, softsign, relu, tanh, sigmoid, hard sigmoid, linear**
Dropout	**Dropout rate, weight constraint**
Init mode	**'uniform', 'lecun_uniform', 'normal', 'zero', 'glorot_normal', 'glorot_uniform', 'he_normal', 'he_uniform'**
Learning Rate	**Learning rate, momentum, rho**
Neurons	**Number of neurons**
Optimizer	**'SGD', 'RMSprop', 'Adagrad', 'Adadelta', 'Adam', 'Adamax', 'Nadam'**

Improving Predictive Capability





Next question is how good is the trained model in forecasting? The training should be such that the model is optimized but not over-fitted. If over-fitted, the model starts to memorize and will excel in training but will be inaccurate in forecasting. This delicate balance can only be gauged with practice and experience with these algorithms.

Saving trained model and application on new datasets

Finally, the optimized and trained model can be saved to be used to forecast on other datasets with confidence.