Spline – special function defined by polynomials. Splines are popular curves because of the simplicity of their construction, accuracy of evaluation and capacity to approximate complex shapes.

Generalized additive model – generalized linear model in which the linear predictor depends linearly on unknown smooth functions of predictor variables.

8 – Tree- based methods

They can be used for regression and classification problems such as stratifying or segmenting the predictor space into a number of simple regions. These types are known as decision-tree methods are called like that because they can be summarized in a tree. The following methods grow multiple trees which are combined into a single consensus prediction:

Bagging – a way to decrease the variance of your prediction through generating additional data for training, using combinations with repetitions to form a multistep of the same size as the original data.

Boosting – an approach to calculate the output – an average of several different models. Wider range of input data using different models.

Random forest – an algorithm where you draw random samples of your training set and a random subset of features for training individual trees. More independent trees, therefore it is faster, due to each tree learning only from a subset of features.

9 – Support Vector Machines – SVM

It is a classification technique listed under supervised learning models in ML. It involves finding the hyperplane (higher dimensions – n-1 dimensional subspace of an n-dimensional space) that best separates two classes of points with the maximum margin.

The “support vectors” are the data points that “support” this hyperplane on either side.

10 – Unsupervised Learning

Is a technique where it is left on the learning algorithm to figure out patterns in the data provided. Clustering is an example of unsupervised learning in which different data sets are clustered into groups of closely related items. Below is the list of most widely used unsupervised learning algorithms:

Principal Component Analysis – helps in producing low dimensional representation of the data through the help of identifying a set of linear combination of features which have maximum variance and are mutually un-correlated.

k-Means clustering – partitions data into k-distinct clusters based on the distance to the centroid of a cluster.

Hierarchical clustering – builds multilevel hierarchy of clusters by creating a cluster tree.

This was a basic run-down of some basic statistical techniques that can help a data science program manager and or executive have a better understanding of what is running underneath the hood of their data science teams.