Evaluating Machine Learning Projects

One of the salient features that helped in solving problems taking the article as a reference was "**do machine learning like the great engineer you are, not like the great machine learning expert you aren't.**"It's a fact most of the problems faced by even scientists are of an engineering nature as far as machine learning is concerned.Post that it becomes essential to understand whether Machine Learning could be at all be a part of theProblem or could be an overkill.If it's affirmative that machine learning algorithms is a mustthen one should think about solving the problems using specific algorithms but not wandering through complicated approaches or mathematical jargons.First and foremost repeated experiments should be conducted on various metrics by splitting the training data set into multiple chunks to ascertain the appropriate model.Datasets are composed of two main types of data: Numerical (eg. integers, floats), and Categorical (eg. names, laptops brands) Numerical data can additionally be divided into other two categories: Discrete and Continue. Discrete data can take only certain values (eg. number of students in a school) while continuous data can take any real or fractional value (eg. the concepts of height and weights).Probability Mass Functions gives the probability that a variable can be equal to a certain value, instead, the values of Probability Density Functions are not itself probabilities because they need first to be integrated over the given range.

We are predicting the natural log of the sum of all transactions per user in our project which is dealing with Google Analytics Customer Revenue Prediction.It can be solved by many algorithms like Classification,Bayesian Methods but we need to think about the dimension also since it have huge hit into the training phase.So thus a PCA over the training features enables to keep our Systems at place to tackle "Big Data".We should also watch for silent features.This is a problem that occurs more for machine learning systems than for other kinds of systems. Suppose that a particular table that is being joined is no longer being updated. The machine learning system will adjust, and behavior will continue to be reasonably good, decaying gradually. Sometimes one may find tables that are months out of date, and a simple refresh improves performance more than any other launch that quarter! The coverage of a feature may change due to implementation changes: for example a feature column could be populated in 90% of the examples, and suddenly drop to 60% of the examples.In our case we need to do a sum for all the transactions of the user and then do a log transformation on top and then we do a mean squared error on the transactionRevenue and Log of the PredictedRevenue to get a validation score.

Sometimes if data is still huge one may need to use distributed computing tools Apache Spark And HDFS to run the algorithms parallely using multiple machines which could be found As a bundle in AWS Elastic Map Reduce.
Cross-validation that another important pattern is frequently used to tune model parameters, for example, the optimal number of nearest neighbors in a *k*-nearest neighbor classifier. Here, cross-validation is applied multiple times for different values of the tuning parameter, and the parameter that minimizes the cross-validated error is then used to build the final model. Thereby, cross-validation addresses the problem of overfitting.