**1. What are the three stages to build the hypotheses or model in machine learning?**

a) Model building b) Model testing c) Applying the model

* Model Building: Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.
* Model Testing: Most machine learning systems are based on neural networks. A neural network is a set of layered algorithms whose variables can be adjusted via a learning/testing process. The learning/testing process involves using known data inputs to create outputs that are then compared with known results.

Mentioned below are critical activities that I believe will be essential to test machine learning systems:

1. Developing training data sets

2. Developing test data sets

3. Developing validation test suites based on algorithms and test datasets

4. Communicating test results in statistical terms.

* Applying the model:

Pointers for Applying Machine Learning Model

1 – Begin with a priority problem, not a toy problem

2- You can give it data, but all of the context must come from you

Thinking through what information to “feed” your algorithm is not as easy as one might presume. While ML algorithms are adept in identifying correlations, they won’t understand the facts surrounding the data that might make it relevant or irrelevant. Here are some examples of how “context” could get in the way of developing an effective ML solution:

* Predicting eCommerce customer lifetime value
* Determining medical recovery time
* Recommending related products

3 – Expect to tinker, tweak, and adjust to find ROI

Building a ML solution requires careful thinking and testing in selecting algorithms, selecting data, cleaning data, and testing in a live environment. There are no “out-of-the-box” machine learning solutions for unique and complex business use cases. Even for extremely common use cases (recommendation engines, predicting customer churn), each application will vary widely and require iteration and adjustment. If a company goes into an ML project without resources committed to an extended period of tinkering, it may never achieve a useful result.

**2. What is the standard approach to supervised learning?**

The standard approach to supervised learning is to split the set of example into the training set and the test.

In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

**3. What is Training set and Test set?**

* In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an examples given to the learner.

In Machine Learning, a training set is a dataset used to train a model. In training the model, specific features are picked out from the training set. These features are then incorporated into the model. Thereby, if the training set is labeled correctly, the model should be able to learn something from these features.

* Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

The test set is a dataset used to measure how well the model performs at making predictions on that test set. If the prediction scores for the test set are unreasonable, we’ll have to make some adjustments to our model and try again.

**4. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?**

* The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model.

BAGGING

Several estimators are built independently on subsets of the data and their predictions are averaged. Typically the combined estimator is usually better than any of the single base estimator.

BOOSTING

Base estimators are built sequentially. Each subsequent estimator focuses on the weaknesses of the previous estimators. In essence several weak models "team up" to produce a powerful ensemble model.

**5. How can you avoid overfitting ?**

Overfitting refers to a model that models the training data too well. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.

This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.

Steps for avoiding overfitting:

* Add more data.
* Use data augmentation.
* Use architectures that generalize well.
* Add regularization
* Reduce architecture complexity.