**Q: How AI Software Development life cycle differs from traditional software development**

Ans: Explanation of AI Software Development life cycle and traditional software development cycle

**Q: When people say that artificial intelligence (AI) and machine learning (ML) requires good data, what does that mean?**

**OR**

**“When the data is high-quality, scientists spend time working on the business problem and decision making, not tweaking data.” What are the attributes of a high-quality data?**

**Ans:**

Good quality data is fit for its purpose. Data quality has many dimensions:

·**Completeness:** The data has the expected comprehensiveness. For example, when someone gathers phone numbers, we expect it will include the area code.

·**Consistency:** All systems across the data ecosystem contain the same information

·**Accuracy:** How accurately does the data reflect the event in question or the real-world object?

·**Timeliness:** Is the data available when required. For instance, real-time clickstream data can demonstrate where in the buying process, customers are facing challenges and motivate them to continue through the purchase path.

·**Validity:** It conforms to the structure of its definition

·**Uniqueness:** Each data entry is one of its kind

**Q: How Big data is different from the data stored in traditional databases? Elaborate**

**Ans:** Big Data is different from traditional databases because of the following characteristics:

**1. Volume:**

* The name ‘Big Data’ itself is related to a size which is enormous.
* Volume is a huge amount of data.
* To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a ‘Big Data’. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
* Hence while dealing with Big Data it is necessary to consider a characteristic ‘Volume’.
* *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

**2. Velocity:**

* Velocity refers to the high speed of accumulation of data.
* In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
* There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
* Sampling data can help in dealing with the issue like ‘velocity’.
* *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

**3. Variety:**

* It refers to nature of data that is structured, semi-structured and unstructured data.
* It also refers to heterogeneous sources.
* Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

**4. Veracity:**

* It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
* Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
* *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

**5. Value:**

* After having the 4 V’s into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
* Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V’s.

**Q: What are the challenges associated with Machine Learning?**

**Source:** [**https://www.javatpoint.com/issues-in-machine-learning**](https://www.javatpoint.com/issues-in-machine-learning)

Although machine learning is being used in every industry and helps organizations make more informed and data-driven choices that are more effective than classical methodologies, it still has so many problems that cannot be ignored. Here are some common issues in Machine Learning that professionals face to inculcate ML skills and create an application from scratch.

### **1. Inadequate Training Data and Poor quality of data**

The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data. Although data plays a vital role in the processing of machine learning algorithms, many data scientists claim that inadequate data, inaccurate data, noisy data, and unclean data are extremely exhausting the machine learning algorithms. This leads to less accuracy in classification and low-quality results. Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.

### **2. Non-representative training data**

To make sure our training model is generalized well or not, we have to ensure that sample training data must be representative of new cases that we need to generalize. The training data must cover all cases that are already occurred as well as occurring.

### **3. Overfitting**

Overfitting is one of the most common issues faced by Machine Learning engineers and data scientists. Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. It negatively affects the performance of the model. The main reason behind overfitting is using non-linear methods used in machine learning algorithms as they build non-realistic data models.

**4. Underfitting:**

Underfitting is just the opposite of overfitting. Whenever a machine learning model is trained with fewer amounts of data, and as a result, it provides incomplete and inaccurate data and destroys the accuracy of the machine learning model.

### **5. Monitoring and maintenance**

As we know that generalized output data is mandatory for any machine learning model; hence, regular monitoring and maintenance become compulsory for the same. Different results for different actions require data change; hence editing of codes as well as resources for monitoring them also become necessary.

### **6. Getting bad recommendations**

A machine learning model operates under a specific context which results in bad recommendations and concept drift in the model. It generally occurs when new data is introduced or interpretation of data changes. However, we can overcome this by regularly updating and monitoring data according to the expectations.

### **7. Lack of skilled resources**

Although Machine Learning and Artificial Intelligence are continuously growing in the market, still these industries are fresher in comparison to others. The absence of skilled resources in the form of manpower is also an issue. Hence, we need manpower having in-depth knowledge of mathematics, science, and technologies for developing and managing scientific substances for machine learning.

### **8. Process Complexity of Machine Learning**

The machine learning process is very complex, which is also another major issue faced by machine learning engineers and data scientists. However, Machine Learning and Artificial Intelligence are very new technologies but are still in an experimental phase and continuously being changing over time. There is the majority of hits and trial experiments; hence the probability of error is higher than expected.

### **9. Data Bias**

Data Biasing is also found a big challenge in Machine Learning. These errors exist when certain elements of the dataset are heavily weighted or need more importance than others. Biased data leads to inaccurate results, skewed outcomes, and other analytical errors. However, we can resolve this error by determining where data is actually biased in the dataset.

### **10. Lack of Explainability**

This basically means the outputs cannot be easily comprehended as it is programmed in specific ways to deliver for certain conditions. Hence, a lack of explainability is also found in machine learning algorithms which reduce the credibility of the algorithms.

**Q: What is the difference between Supervised and unsupervised machine learning? Explain with examples**

**Q:** **You are working on an NLP model. So, you are dealing with words and sentences, not numbers. Your problem is to categorize these words and make sense of them. Your manager told you that you have to use embeddings.**

**Which of the following techniques are not related to embeddings? Explain the other terms related to embedding.**

**A. Count Vector**

**B. TF-IDF Vector**

**C. Co-Variance Matrix**

**Ans:** Covariance matrix is not an embedding technique. Covariance matrices are square matrices with the covariance between each pair of elements. It measures how much the change of one with respect to another is related.

**Other two are embedding techniques:**

**Count Vector**

* It is one of the simplest ways of doing text vectorization.
* It creates a document term matrix, which is a set of dummy variables that indicates if a particular word appears in the document.
* Count vectorizer will fit and learn the word vocabulary and try to create a document term matrix in which the individual cells denote the frequency of that word in a particular document, which is also known as term frequency, and the columns are dedicated to each word in the corpus.

**Example:**

**Document-1:** He is a smart boy. She is also smart.

**Document-2:** Chirag is a smart person.

The dictionary created contains the list of unique tokens(words) present in the corpus

**Unique Words:** [‘He’, ’She’, ’smart’, ’boy’, ’Chirag’, ’person’] Here, D=2, N=6

So, the count matrix M of size 2 X 6 will be represented as –

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **He** | **She** | **smart** | **boy** | **Chirag** | **person** |
| **D1** | **1** | **1** | **2** | **1** | **0** | **0** |
| **D2** | **0** | **0** | **1** | **0** | **1** | **1** |

**TF-IDF**

Term frequency denotes the frequency of a word in a document. For a specified word, it is defined as the ratio of the number of times a word appears in a document to the total number of words in the document.



Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. It measures the importance of the word in the corpus.

Document Frequency tells us about the proportion of documents that contain a certain word. IDF is the reciprocal of the Document Frequency.

The intuition behind using IDF is that the more common a word is across all documents, the lesser its importance is for the current document. A logarithm is taken to dampen the effect of (normalize) IDF in the final calculation.

The final TF-IDF score comes out to be:

**Q: You are a junior Data Scientist and are working on a deep neural network model to optimize the level of customer satisfaction for after-sales services with the goal of creating greater client loyalty. You are struggling with your model (learning rates, hidden layers and nodes selection) for optimizing processing and to let it converge in the fastest way. What is this problem called in ML language? Explain**

**Answer: Hyperparameter Tuning**

* Hyperparameters are configuration variables that influence the training process itself: Learning rate, hidden layers number, number of epochs, regularization, batch size are all examples of hyperparameters.
* Explanation

**Q: Imagine, you are working with a microblogging website and you want to develop a machine learning algorithm which predicts the number of views on the articles.**

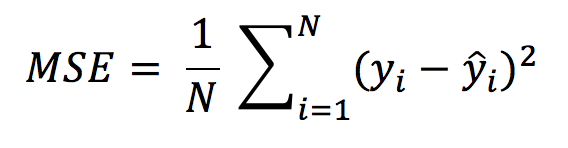
**Your analysis is based on features like author name, number of articles written by the same author in past and a few other features. Which of the following evaluation metric would you choose in that case? Explain the metric used.**

1. **Mean Square Error**
2. **Accuracy**
3. **F1 Score**

**Solution: Mean Square Error**

The number of views of articles is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metrics.

Mean Square Error is an absolute measure of the goodness for the fit.



* MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points.
* It gives you an absolute number on how much your predicted results deviate from the actual number.
* You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

Q: **You are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data. You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?**

* All categories of categorical variable are not present in the test dataset.
* Frequency distribution of categories is different in train as compared to the test dataset.

The OHE will fail to encode the categories which is present in test but not in train so it could be one of the main challenges while applying OHE. The challenge given in option B is also true you need to more careful while applying OHE if frequency distribution doesn’t same in train and test.

Q: **You’ve built a Decision tree model for a given dataset. You got delighted after getting training accuracy as 99%. But, the testing accuracy is 68%. What is the problem in this scenario? Explain how you can avoid it.**

**Answer:** The model has overfitted. Training accuracy of 99% means the classifier has mimicked the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on unseen sample, it couldn’t find those patterns and returned prediction with higher error. In random forest, it happens when we use larger number of trees than necessary. Hence, to avoid these situations, we should tune number of trees using cross validation.

Q:  **We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn’t. How?**

**Answer:** Using one hot encoding, the dimensionality (a.k.a features) in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let’s say we have a variable ‘color’. The variable has 3 levels namely Red, Blue and Green. One hot encoding ‘color’ variable will generate three new variables as Color.Red, Color.Blue and Color.Green containing 0 and 1 value.

In label encoding, the levels of a categorical variables gets encoded as 0 and 1, so no new variable is created. Label encoding is majorly used for binary variables.

Q: **What do you understand by Type I vs Type II error ?**

**Answer:** Type I error is committed when the null hypothesis is true and we reject it, also known as a ‘False Positive’. Type II error is committed when the null hypothesis is false and we accept it, also known as ‘False Negative’.

In the context of confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

Q: **A linear regression model is generally evaluated using MSE or Adjusted R². How would you evaluate a logistic regression model?**

**Answer:** Since logistic regression is used to predict probabilities, we can use AUC-ROC curve along with confusion matrix to determine its performance. **Explanation of AUC-ROC.**

**Q: Explain dimensionality reduction, where it’s used, and its benefits?**

Dimensionality reduction is the process of reducing the number of feature variables under consideration by obtaining a set of principal variables which are basically the important features. Importance of a feature depends on how much the feature variable contributes to the information representation of the data and depends on which technique you decide to use. Deciding which technique to use comes down to trial-and-error and preference. It’s common to start with a linear technique and move to non-linear techniques when results suggest inadequate fit. Benefits of dimensionality reduction for a data set may be:

* Reduce the storage space needed.
* Speed up computation (for example in machine learning algorithms), less dimensions mean less computing, also less dimensions can allow usage of algorithms unfit for a large number of dimensions.
* Remove redundant features, for example no point in storing a terrain’s size in both sq meters and sq miles (maybe data gathering was flawed).
* Reducing a data’s dimension to 2D or 3D may allow us to plot and visualize it, maybe observe patterns, give us insights.
* Too many features or too complex a model can lead to overfitting.

**Q: Explain over- and under-fitting and how to combat them?**

**Q: Why is ReLU better and more often used than Sigmoid in Neural Networks?**

**Q: What is the curse of dimensionality? List some ways to deal with it?**

The curse of dimensionality is when the training data has a high feature count, but the dataset does not have enough samples for a model to learn correctly from so many features. For example, a training dataset of 100 samples with 100 features will be very hard to learn from because the model will find random relations between the features and the target. However, if we had a dataset of 100k samples with 100 features, the model could probably learn the correct relationships between the features and the target.

There are different options to fight the curse of dimensionality:

* **Feature selection.** Instead of using all the features, we can train on a smaller subset of features.
* **Dimensionality reduction.** There are many techniques that allow to reduce the dimensionality of the features. Principal component analysis (PCA) and using autoencoders are examples of dimensionality reduction techniques.
* **L1 regularization.** Because it produces sparse parameters, L1 helps to deal with high-dimensionality input.
* **Feature engineering.** It’s possible to create new features that sum up multiple existing features. For example, we can get statistics such as the mean or median.

**Q:** **Why do we need a validation set and test set? What is the difference between them?**

When training a model, we divide the available data into three separate sets:

* The training dataset is used for fitting the model’s parameters. However, the accuracy that we achieve on the training set is not reliable for predicting if the model will be accurate on new samples.
* The validation dataset is used to measure how well the model does on examples that weren’t part of the training dataset. The metrics computed on the validation data can be used to tune the hyperparameters of the model. However, every time we evaluate the validation data and we make decisions based on those scores, we are leaking information from the validation data into our model. The more evaluations, the more information is leaked. So we can end up overfitting to the validation data, and once again the validation score won’t be reliable for predicting the behavior of the model in the real world.
* The test dataset is used to measure how well the model does on previously unseen examples. It should only be used once we have tuned the parameters using the validation set.

So if we omit the test set and only use a validation set, the validation score won’t be a good estimate of the generalization of the model.

**Q:** **Explain the differences between supervised, unsupervised, and reinforcement learning?**

**Q: What is an imbalanced dataset? list the ways to deal with it?**

**Q: What is regularization? Give some examples of regularization techniques?**

Regularization is any technique that aims to improve the validation score, sometimes at the cost of reducing the training score. Some regularization techniques:

* L1 tries to minimize the absolute value of the parameters of the model. It produces sparse parameters.
* L2 tries to minimize the square value of the parameters of the model. It produces parameters with small values.
* Dropout is a technique applied to neural networks that randomly sets some of the neurons’ outputs to zero during training. This forces the network to learn better representations of the data by preventing complex interactions between the neurons: Each neuron needs to learn useful features.
* Early stopping will stop training when the validation score stops improving, even when the training score may be improving. This prevents overfitting on the training dataset.

#### Q: You work as a machine learning specialist for a consulting firm where you analyze data about the consultants who work there in preparation for using the data in your machine learning models. The features you have in your data are things like employee id, specialty, practice, job description, billing hours, and principle. The principle attribute is represented as ‘yes’ or ‘no’, whether the consultant has made principle level or not. For your initial analysis, you need to identify the distribution of consultants and their billing hours for the given period. What visualization best describes this relationship?

**Ans: Histogram**

You are looking for a distribution of a single dimension: the consultants billing hours., “A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable.” The continuous variable in this question: the billing hours, binned into ranges (x-axis), at a frequency: the number of consultants at a billing hour range (y-axis).

#### Q: You work as a machine learning specialist for a state highway administration department. Your department is trying to use machine learning to help determine the make and model of cars as they pass a camera on the state highways. You need to build a machine learning model to accomplish this problem. Which modeling approach best fits your problem?

**A.** Multi-Class Classification  
**B.** Simulation-based Reinforcement Learning  
**C.** Binary Classification  
**D.** Heuristic Approach

**Correct Answer: A**

**Explanation**

**Option A is correct**. Multi-Class Classification is used when your model needs to choose from a finite set of outcomes, such as this car make and model classification image recognition problem.

#### Q:You work for the security department of your firm. As part of securing your firm’s email activity from phishing attacks, you need to build a machine learning model that analyzes incoming email text to find word phrases like “you’re a winner” or “click here now” to find potential phishing emails.  Which text feature engineering technique is the best solution for this task? ****Ans:**** N-Gram

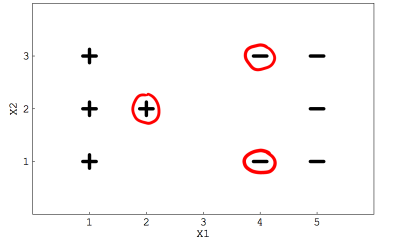
The N-Gram natural language processing algorithm is used to find multi-word phrases in the text, in this case, an email. This suits your phishing detection task since you are trying to classify an email as a phishing attack by having your model learn based on the presence of multi-word phrases. Further explanation….

#### Q: You work for a company that performs seismic research for client firms that drill for petroleum. As a machine learning specialist, you have built a series of models that classify seismic waves to determine the seismic profile of a proposed drilling site. You need to select the best model to use in production. Which metric should you use to compare and evaluate your machine learning classification models against each other?

**Ans.** Area Under the ROC Curve (AUC)

The area under the Receiver Operating Characteristic (ROC) curve is the most commonly used metric to compare classification models. Explanation….

**Q: Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.**

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/09/01154204/Image_18-19.png)

1. If you remove the following any one red points from the data. Does the decision boundary will change?
2. If you remove the non-red circled points from the data, the decision boundary will change?

Justify your answers

**Ans:**

1) These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.

2) rest of the points in the data won’t affect the decision boundary much.