SCENARIO BASED QUESTIONS IN ML

Q1. You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)

Answer: Processing a high dimensional data on a limited memory machine is a strenuous task, your interviewer would be fully aware of that. Following are the methods you can use to tackle such situation:  
Since we have lower RAM, we should close all other applications in our machine, including the web browser, so that most of the memory can be put to use.  
We can randomly sample the data set. This means, we can create a smaller data set, let’s say, having 1000 variables and 300000 rows and do the computations.  
To reduce dimensionality, we can separate the numerical and categorical variables and remove the correlated variables. For numerical variables, we’ll use correlation. For categorical variables, we’ll use chi-square test.  
Also, we can use PCA and pick the components which can explain the maximum variance in the data set.  
Using online learning algorithms like Vowpal Wabbit (available in Python) is a possible option.  
Building a linear model using Stochastic Gradient Descent is also helpful.  
We can also apply our business understanding to estimate which all predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.  
Note: For point 4 & 5, make sure you read about online learning algorithms & Stochastic Gradient Descent. These are advanced methods.

Q2. Is rotation necessary in PCA? If yes, Why? What will happen if you don’t rotate the components?

Answer: Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that’s the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn’t change, it only changes the actual coordinates of the points.  
If we don’t rotate the components, the effect of PCA will diminish and we’ll have to select more number of components to explain variance in the data set.

Q3. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

Answer: This question has enough hints for you to start thinking! Since, the data is spread across median, let’s assume it’s a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

Q4. You are given a data set on cancer detection. You’ve build a classification model and achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?

Answer: If you have worked on enough data sets, you should deduce that cancer detection results in imbalanced data. In an imbalanced data set, accuracy should not be used as a measure of performance because 96% (as given) might only be predicting majority class correctly, but our class of interest is minority class (4%) which is the people who actually got diagnosed with cancer. Hence, in order to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine class wise performance of the classifier. If the minority class performance is found to to be poor, we can undertake the following steps:  
We can use undersampling, oversampling or SMOTE to make the data balanced.  
We can alter the prediction threshold value by doing probability caliberation and finding a optimal threshold using AUC-ROC curve.  
We can assign weight to classes such that the minority classes gets larger weight.  
We can also use anomaly detection.  
Know more: Imbalanced Classification

Q5. Why is Naive Bayes so ‘naive’ ?

Answer: Naive Bayes is so ‘naive’ because it assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.

Q6. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?

Answer: Prior probability is nothing but, the proportion of dependent (binary) variable in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (not spam) is 30%. Hence, we can estimate that there are 70% chances that any new email would be classified as spam.  
Likelihood is the probability of classifying a given observation as 1 in presence of some other variable. For example: The probability that the word ‘FREE’ is used in previous spam message is likelihood. Marginal likelihood is, the probability that the word ‘FREE’ is used in any message.

Q7. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?

Answer: Time series data is known to posses linearity. On the other hand, a decision tree algorithm is known to work best to detect non – linear interactions. The reason why decision tree failed to provide robust predictions because it couldn’t map the linear relationship as good as a regression model did. Therefore, we learned that, a linear regression model can provide robust prediction given the data set satisfies its linearity assumptions.

Q8. You are assigned a new project which involves helping a food delivery company save more money. The problem is, company’s delivery team aren’t able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?

Answer: You might have started hopping through the list of ML algorithms in your mind. But, wait! Such questions are asked to test your machine learning fundamentals.  
This is not a machine learning problem. This is a route optimization problem. A machine learning problem consist of three things:  
There exist a pattern.  
You cannot solve it mathematically (even by writing exponential equations).  
You have data on it.  
Always look for these three factors to decide if machine learning is a tool to solve a particular problem.

Q9. You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?

Answer: Low bias occurs when the model’s predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results.  
In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. Then, these samples are used to generate a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).  
Also, to combat high variance, we can:  
Use regularization technique, where higher model coefficients get penalized, hence lowering model complexity.  
Use top n features from variable importance chart. May be, with all the variable in the data set, the algorithm is having difficulty in finding the meaningful signal.

Q10. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?

Answer: Chances are, you might be tempted to say No, but that would be incorrect. Discarding correlated variables have a substantial effect on PCA because, in presence of correlated variables, the variance explained by a particular component gets inflated.  
For example: You have 3 variables in a data set, of which 2 are correlated. If you run PCA on this data set, the first principal component would exhibit twice the variance than it would exhibit with uncorrelated variables. Also, adding correlated variables lets PCA put more importance on those variable, which is misleading.

### **Q11. You are given a data set consisting of variables having more than 30% missing values? Let’s say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?**

* Assign a unique category to the missing values, who knows the missing values might uncover some trend.
* We can remove them blatantly.
* Or, we can sensibly check their distribution with the target variable, and if found any pattern we’ll keep those missing values and assign them a new category while removing others.

### **Q12. Write an SQL query that makes recommendations using the pages that your friends liked. Assume you have two tables: a two-column table of users and their friends, and a two-column table of users and the pages they liked. It should not recommend pages you already like.**

|  |  |
| --- | --- |
| 1  2  3  4  5 | SELECT f.user\_id, l.page\_id  FROM friend f JOIN like l  ON f.friend\_id = l.user\_id  WHERE l.page\_id NOT IN (SELECT page\_id FROM like  WHERE user\_id = f.user\_id) |

### **Q13. There’s a game where you are asked to roll two fair six-sided dice. If the sum of the values on the dice equals seven, then you win $21. However, you must pay $5 to play each time you roll both dice. Do you play this game? And in the follow-up: If he plays 6 times what is the probability of making money from this game?**

* The first condition states that if the sum of the values on the 2 dices is equal to 7, then you win $21. But for all the other cases you must pay $5.
* First, let’s calculate the number of possible cases. Since we have two 6-sided dices, the total number of cases => 6\*6 = 36.
* Out of 36 cases, we must calculate the number of cases that produces a sum of 7 (in such a way that the sum of the values on the 2 dices is equal to 7)
* Possible combinations that produce a sum of 7 is, (1,6), (2,5), (3,4), (4,3), (5,2) and (6,1). All these  6 combinations generate a sum of 7.
* This means that out of 36 chances, only 6 will produce a sum of 7. On taking the ratio, we get: 6/36 = 1/6
* So this suggests that we have a chance of winning $21, once in 6 games.
* So to answer the question if a person plays 6 times, he will win one game of $21, whereas for the other 5 games he will have to pay $5 each, which is $25 for all five games. Therefore, he will face a loss because he wins $21 but ends up paying $25.

### **Q14. We have two options for serving ads within Newsfeed:** **1 – out of every 25 stories, one will be an ad** **2 – every story has a 4% chance of being an ad**

### **For each option, what is the expected number of ads shown in 100 news stories?** **If we go with option 2, what is the chance a user will be shown only a single ad in 100 stories? What about no ads at all?**

* The expected number of ads shown in 100 new stories for option 1 is equal to 4 (100/25 = 4).
* Similarly, for option 2, the expected number of ads shown in 100 new stories is also equal to 4 (4/100 = 1/25 which suggests that one out of every 25 stories will be an ad, therefore in 100 new stories there will be 4 ads)
* Therefore for each option, the total number of ads shown in 100 new stories is 4.
* The second part of the question can be solved by using Binomial distribution. Binomial distribution takes three parameters:
  + The probability of success and failure, which in our case is 4%.
  + The total number of cases, which is 100 in our case.
  + The probability of the outcome, which is a chance that a user will be shown only a single ad in 100 stories
* p(single ad) = (0.96)^99\*(0.04)^1

(note: here 0.96 denotes the chance of not seeing an ad in 100 stories, 99 denotes the possibility of seeing only 1 ad, 0.04 is the probability of seeing an ad once in 100 stories )

* In total, there are 100 positions for the ad. Therefore, 100 \* p(single ad) = 7.03%

### **Q15. How would you predict who will renew their subscription next month? What data would you need to solve this? What analysis would you do? Would you build predictive models? If so, which algorithms?**

* Let’s assume that we’re trying to predict renewal rate for Netflix subscription. So our problem statement is to predict which users will renew their subscription plan for the next month.
* Next, we must understand the data that is needed to solve this problem. In this case, we need to check the number of hours the channel is active for each household, the number of adults in the household, number of kids, which channels are streamed the most, how much time is spent on each channel, how much has the watch rate varied from last month, etc. Such data is needed to predict whether or not a person will continue the subscription for the upcoming month.
* After collecting this data, it is important that you find patterns and correlations. For example, we know that if a household has kids, then they are more likely to subscribe. Similarly, by studying the watch rate of the previous month, you can predict whether a person is still interested in a subscription. Such trends must be studied.
* The next step is analysis. For this kind of problem statement, you must use a classification algorithm that classifies customers into 2 groups:
  + Customers who are likely to subscribe next month
  + Customers who are not likely to subscribe next month
* Would you build predictive models? Yes, in order to achieve this you must build a predictive model that classifies the customers into 2 classes like mentioned above.
* Which algorithms to choose? You can choose classification algorithms such as Logistic Regression, Random Forest, Support Vector Machine, etc.
* Once you’ve opted the right algorithm, you must perform model evaluation to calculate the efficiency of the algorithm. This is followed by deployment.

### **Q16. How do you map nicknames (Pete, Andy, Nick, Rob, etc) to real names?**

* This problem can be solved in n number of ways. Let’s assume that you’re given a data set containing 1000s of twitter interactions. You will begin by studying the relationship between two people by carefully analyzing the words used in the tweets.
* This kind of problem statement can be solved by implementing Text Mining using Natural Language Processing techniques, wherein each word in a sentence is broken down and co-relations between various words are found.
* NLP is actively used in understanding customer feedback, performing sentimental analysis on Twitter and Facebook. Thus, one of the ways to solve this problem is through Text Mining and Natural Language Processing techniques.

### **Q17. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?**

* There are two ways of choosing a coin. One is to pick a fair coin and the other is to pick the one with two heads.
* Probability of selecting fair coin = 999/1000 = 0.999  
  Probability of selecting unfair coin = 1/1000 = 0.001
* Selecting 10 heads in a row = Selecting fair coin \* Getting 10 heads + Selecting an unfair coin
* P (A) = 0.999 \* (1/2)^10 = 0.999 \* (1/1024) = 0.000976  
  P (B) = 0.001 \* 1 = 0.001  
  P( A / A + B ) = 0.000976 / (0.000976 + 0.001) = 0.4939  
  P( B / A + B ) = 0.001 / 0.001976 = 0.5061
* Probability of selecting another head = P(A/A+B) \* 0.5 + P(B/A+B) \* 1 = 0.4939 \* 0.5 + 0.5061 = 0.7531

### **Q18. Suppose you are given a data set which has missing values spread along 1 standard deviation from the median. What percentage of data would remain unaffected and Why?**

Since the data is spread across the median, let’s assume it’s a normal distribution.  
As you know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

### **Q19. You are given a cancer detection data set. Let’s suppose when you build a classification model you achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?**

You can do the following:

* Add more data
* Treat missing outlier values
* Feature Engineering
* Feature Selection
* Multiple Algorithms
* Algorithm Tuning
* Ensemble Method
* Cross-Validation

### **Q20. You are working on a time series data set. Your manager has asked you to build a high accuracy model. You start with the decision tree algorithm since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than the decision tree model. Can this happen? Why?**

* Time series data is based on linearity while a decision tree algorithm is known to work best to detect non-linear interactions
* Decision tree fails to provide robust predictions. Why?
  + The reason is that it couldn’t map the linear relationship as good as a regression model did.
  + We also know that a linear regression model can provide a robust prediction only if the data set satisfies its linearity assumptions.

### **Q21. Suppose you found that your model is suffering from low bias and high variance. Which algorithm you think could tackle this situation and Why?**

Type 1: How to tackle high variance?

* Low bias occurs when the model’s predicted values are near to actual values.
* In this case, we can use the bagging algorithm (eg: [Random Forest](https://www.edureka.co/blog/random-forest-classifier/" \t "_blank)) to tackle high variance problem.
* Bagging algorithm will divide the data set into its subsets with repeated randomized sampling.
* Once divided, these samples can be used to generate a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).

Type 2: How to tackle high variance?

* Lower the model complexity by using regularization technique, where higher model coefficients get penalized.
* You can also use top n features from variable importance chart. It might be possible that with all the variable in the data set, the algorithm is facing difficulty in finding the meaningful signal.

### **Q22. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?**

Possibly, you might get tempted to say no, but that would be incorrect.  
Discarding correlated variables will have a substantial effect on PCA because, in the presence of correlated variables, the variance explained by a particular component gets inflated.

### **Q23. You are asked to build a multiple regression model but your model R² isn’t as good as you wanted. For improvement, you remove the intercept term now your model R² becomes 0.8 from 0.3. Is it possible? How?**

Yes, it is possible.

* The intercept term refers to model prediction without any independent variable or in other words, mean prediction  
  R² = 1 – ∑(Y – Y´)²/∑(Y – Ymean)² where Y´ is the predicted value.
* In the presence of the intercept term, R² value will evaluate your model with respect to the mean model.
* In the absence of the intercept term (Ymean), the model can make no such evaluation,
* With large denominator,  
  Value of ∑(Y – Y´)²/∑(Y)² equation becomes smaller than actual, thereby resulting in a higher value of R².

### **Q24. You’re asked to build a random forest model with 10000 trees. During its training, you got training error as 0.00. But, on testing the validation error was 34.23. What is going on? Haven’t you trained your model perfectly?**

* The model is overfitting the data.
* Training error of 0.00 means that the classifier has mimicked the training data patterns to an extent.
* But when this classifier runs on the unseen sample, it was not able to find those patterns and returned the predictions with more number of errors.
* In Random Forest, it usually happens when we use a larger number of trees than necessary. Hence, to avoid such situations, we should tune the number of trees using cross-validation.

**Q25. ‘People who bought this also bought…’ recommendations seen on Amazon is based on which algorithm?**

E-commerce websites like Amazon make use of Machine Learning to recommend products to their customers. The basic idea of this kind of recommendation comes from collaborative filtering. Collaborative filtering is the process of comparing users with similar shopping behaviors in order to recommend products to a new user with similar shopping behavior.

Q.26 Suppose that you have to train your neural networks over a dataset of 20 GB. You have a RAM of 3 GB. How will you resolve this problem of training large data?

Ans.

* We will train our neural network with limited memory as follows:
* We first load the entire data in our numpy array.
* Then we obtain the data through passing the index to the numpy array.
* We then pass this data to our neural network and train it in small batches.

*You can’t afford to miss Neural Network for data science interview preparation. Learn it through the DataFlair’s latest guide on [Neural Networks](https://data-flair.training/blogs/artificial-neural-network/) for Data Science Interview.*

Q.27 If through training all the features in the dataset, an accuracy of 100% is obtained but with the validation set, the accuracy score is 75%. What should be looked out for?

Ans. If the training accuracy of 100% is obtained, then a verification of overfitting is required in our model.

Q.28 Suppose that you are training your machine learning model on the text data. The document matrix that is created consists of more than 200K documents. What techniques can you use to reduce the dimensions of the data?

Ans. In order to reduce the dimensions of our data, we can use any one of the following three techniques:

* Latent Semantic Indexing
* Latent Dirichlet Allocation
* Keyword Normalization

Q.29 In a survey conducted, the average height was 164cm with a standard deviation of 15cm. If Alex had a z-score of 1.30, what will be his height?

Ans. Using the formula, X= μ+Zσ, we determine that X = 164 + 1.30\*15 = 183.5. Therefore, the height of Alex is 183.50 cm.

Q.30 While reading the file ‘file.csv’, you get the following error:

Traceback (most recent call last):

File “<input>”, line 1, in<module>

UnicodeEncodeError: ‘ascii’ codec can’t encode character.

How will you correct this error?

Ans. In order to correct this error, we will read the csv with the utf-8 encoding. pd.read\_csv(“‘file.csv”, encoding=’utf-8′)

Q.31 Assume that you have to perform clustering analysis. The first step towards any data science problem including clustering is data cleaning. However, in this case of clustering analysis you have a lesser number of data points. What strategy would you use while performing data cleaning prior to the clustering operation?

Ans. Capping and Flooring would be the most appropriate strategy for performing data cleaning prior to the clustering operation.

Q.32 Assume that you are given a data science problem that involves dimensionality reduction as a part of its pre-processing technique. You are required to reduce the original data to k dimensions using PCA and then use them as projections for the main features. What value of k would you select – high or low to decrease the regularization?

Ans. In order to preserve the characteristics of our data, the value of k will be high, therefore, leading to less regularization

.

Q.33 For a given dataset, you decide to use SVM as the main classifier. You select RBF as your kernel. What would be the optimum gamma value that would allow you to capture the features of the dataset really well?

Ans. In SVM, the gamma parameter denotes the influence of the points that are either near or far away from the dividing hyperplane. When the gamma is high, the model will be able to capture the shape of the data quite well.

*Master SVM concepts with DataFlairs best ever tutorial on [Support Vector Machines](https://data-flair.training/blogs/svm-support-vector-machine-tutorial/)*

Q.34 Suppose that you have to perform transformation operation on an image. The operation is a basic rotation. The point to be rotated has the coordinates (2,0) to a new coordinate of (0,2). How will you perform this operation?

Ans. In order to rotate the image from the point (2,0) to the point (0,2), we will perform matrix multiplication where [2,0] will be represented as a vector that will be multiplied with the matrix [ [0,-1] , [1,0] ]. As a result of their dot product, we will obtain the new coordinate point of (0,2).

Q.35 Assume that while working in the field of image processing. You have to deploy Finite Difference Filters. However, these filters are very vulnerable to additional noise. What will you do to reduce the noise to the point of minimal distortion?

Ans. In order to reduce the noise to the point of minimal distortion while using the Finite-Difference Filters, we will make use of Smoothing. Smoothing is used in image processing to reduce noise that might be present in an image which can also be used to produce an image that is less pixelated.

Q.36 Assume that for a binary classification challenge, we have a fully connected architecture comprising of a single hidden layer with three neurons and a single output neuron. The structure of the input and output layer is as follows –

Input dataset: [ [0,1,1,0] , [1,1,0,0] , [1,0,0,1], [1,1,0,0] ]  
Output: [ [0] , [1] , [0] ]

For performing model training, the weights have been initialized for both the input and output layer as 1. Based on this, will the model be able to learn from the patterns?

Ans. No. Since you have initialized the weights with 1, all the neurons will try to do the same thing as they will never converge.

Q.37 Suppose that you are training your Artificial Neural Network. After you have created your model, you evaluate it. However, during model evaluation, you find out that the training loss/validation loss remains constant. What could be the reason behind this constant figure loss between training and validation test?

Ans. There are two important reasons that would contribute towards the training/validation loss stagnation. Firstly, the architecture of the model is not properly defined. Secondly, the input data has noisy characteristics.

*Learn everything about [Machine Learning and its Algorithms](https://data-flair.training/blogs/machine-learning-algorithm/)*

Q.38 You have a data science project assignment where you have to deal with 1000 columns and around 1 million rows. The objective of the problem is to carry out classification. You are required to reduce the dimensions of this data in order to reduce the model computation time. Furthermore, your machine suffers from memory constraints. What will you do in this situation?

Ans. Considering memory constraints, developing a machine learning model would prove to be a laborious task. However, one can carry this out with the following steps:

* Since we are low on our RAM, we can preserve the memory by closing the other miscellaneous applications that we do not require.
* We will then perform sampling on our data randomly. This sample will be a much smaller version of the bigger dataset.
* We will then reduce the dimensionality by removing the correlated variables. Furthermore, using PCA, we will select those features that can explain maximum variance in our data.
* We will further create a linear model using stochastic gradient descent.
* Using domain knowledge, we will further drop the predictor variables that do not have much effect on the response variable. This will further lead to a reduction in the number of dimensions.

Q.39 Your company has assigned you a new project that involves assisting a food delivery company to prevent losses from occurring. The main reason behind the loss is that the food delivery team is not able to deliver food to their customers in the stipulated time-frame. As a part of their policy, they are then required to deliver food without any charge. This is resulting in losses on the company’s part. How can you fix this problem using machine learning algorithm?

Ans. Considering that this question does not have any pattern or required data, it does not qualify for a machine learning problem. It is clearly a route optimization problem that will require a different set of algorithms.

Q.40 Suppose that you have been assigned the task of analyzing text data that is obtained from the news sentences that are structured. Now, you have to detect noun phrases, verb phrases as well as perform subject and object detection. Which grammar-based text parsing technique would you use in this scenario?

Ans. In this scenario, we will make use of Dependency and Constituent Parsing Extraction techniques to retrieve relations from the textual data.

Q.41 You are working on a Data Science problem in which you have spent a considerable amount of time in data preprocessing and analysis. You are now required to implement a machine learning model that would provide you with a high accuracy. Knowing that boosting algorithms are coveted by data scientists for their high accuracy, you decide to develop five Gradient Boosting Models. However, the models do not surpass even the standard benchmark score. You then create an ensemble of these five models but you do not succeed. Where exactly did you go wrong?

Ans. Ensemble Learning involves the notion of combining weak learners to form strong learners. The underlying ensemble models only provide accurate results when they are uncorrelated. If the ensemble models in the scenario above do not yield an accurate output then we conclude that the models are correlated.

Q.42 Suppose that you are working on neural networks where you have to utilise an activation function in its hidden layers. The output that we obtain is -0.0002. What type of activation could have been used in order to obtain such type of an output?

Ans. Since, the output obtained is -0.0002 which is between -1 and 1, the activation function which has been used in the hidden layer is tanh.

*Wait! It is time to revise your [neural network concepts](https://data-flair.training/blogs/artificial-neural-networks-for-machine-learning/).*

Q.43 Assume that you are working at DataFlair and you have been assigned the task of developing a machine learning algorithm that predicts the number of views an article attracts. During the process of analysis, you include important features such as author name, number of articles written by the author in the past etc. What would be the ideal evaluation metric that you would use in this scenario?

Ans. Number of views that an article attracts on the website is a continuous target variable which is a part of the regression problem. Therefore, we will make use of mean squared error as our primary evaluation metric.

Q.44 Assume that you are working with categorical features wherein you do not know about the distribution of the categorical variable present in the validation set. Now, you wish to apply one hot encoding on the categorical features. What are the various challenges that you can encounter once you have applied one hot encoding on the categorical variable belonging to the train set?

Ans. Applying One Hot Encoding to encode the categories present in the test set but not in the train set, will not involve all the categories of the categorical variable present in the dataset. Secondly, there could be a possible mismatch between the frequency distribution of the categories present in the training set and the validation set.

Q.45 Suppose that you have to work with the data present on social media. After you have retrieved the data have to develop a model that suggests the hashtags to the user. How will you carry this out?

Ans. We can carry out Topic Modeling to extract significant words present in the corpus. To capture the top n-gram words and their combinations. And, for learning repeating contexts in the sentence we train a word2vec model.

### Q11. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?

Answer: As we know, ensemble learners are based on the idea of combining weak learners to create strong learners. But, these learners provide superior result when the combined models are uncorrelated. Since, we have used 5 GBM models and got no accuracy improvement, suggests that the models are correlated. The problem with correlated models is, all the models provide same information.

For example: If model 1 has classified User1122 as 1, there are high chances model 2 and model 3 would have done the same, even if its actual value is 0. Therefore, ensemble learners are built on the premise of combining weak uncorrelated models to obtain better predictions.

### Q12. How is kNN different from kmeans clustering?

Answer: Don’t get mislead by ‘k’ in their names. You should know that the fundamental difference between both these algorithms is, kmeans is unsupervised in nature and kNN is supervised in nature. kmeans is a clustering algorithm. kNN is a classification (or regression) algorithm.

kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

kNN algorithm tries to classify an unlabeled observation based on its k (can be any number ) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn’t use training data to make generalization on unseen data set.

### Q13. How is True Positive Rate and Recall related? Write the equation.

Answer: True Positive Rate = Recall. Yes, they are equal having the formula (TP/TP + FN).

Know more: Evaluation Metrics

### Q14. You have built a multiple regression model. Your model R² isn’t as good as you wanted. For improvement, your remove the intercept term, your model R² becomes 0.8 from 0.3. Is it possible? How?

Answer: Yes, it is possible. We need to understand the significance of intercept term in a regression model. The intercept term shows model prediction without any independent variable i.e. mean prediction. The formula of R² = 1 – ∑(y – y´)²/∑(y – ymean)² where y´ is predicted value.

When intercept term is present, R² value evaluates your model wrt. to the mean model. In absence of intercept term (ymean), the model can make no such evaluation, with large denominator, ∑(y - y´)²/∑(y)² equation’s value becomes smaller than actual, resulting in higher R².

### Q15. After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he’s true? Without losing any information, can you still build a better model?

Answer: To check multicollinearity, we can create a correlation matrix to identify & remove variables having correlation above 75% (deciding a threshold is subjective). In addition, we can use calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value = 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.

But, removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variable so that the variables become different from each other. But, adding noise might affect the prediction accuracy, hence this approach should be carefully used.

Know more: Regression

### Q16. When is Ridge regression favorable over Lasso regression?

Answer: You can quote ISLR’s authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.

Know more: Ridge and Lasso Regression

### Q17. Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?

Answer: After reading this question, you should have understood that this is a classic case of “causation and correlation”. No, we can’t conclude that decrease in number of pirates caused the climate change because there might be other factors (lurking or confounding variables) influencing this phenomenon.

Therefore, there might be a correlation between global average temperature and number of pirates, but based on this information we can’t say that pirated died because of rise in global average temperature.

Know more: Causation and Correlation

### Q18. While working on a data set, how do you select important variables? Explain your methods.

Answer: Following are the methods of variable selection you can use:

Remove the correlated variables prior to selecting important variables

Use linear regression and select variables based on p values

Use Forward Selection, Backward Selection, Stepwise Selection

Use Random Forest, Xgboost and plot variable importance chart

Use Lasso Regression

Measure information gain for the available set of features and select top n features accordingly.

### Q19. What is the difference between covariance and correlation?

Answer: Correlation is the standardized form of covariance.

Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we’ll get different covariances which can’t be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale.

### Q20. Is it possible capture the correlation between continuous and categorical variable? If yes, how?

Answer: Yes, we can use ANCOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

### Q21. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?

### Answer: The fundamental difference is, random forest uses bagging technique to make predictions. GBM uses boosting techniques to make predictions.

### In bagging technique, a data set is divided into n samples using randomized sampling. Then, using a single learning algorithm a model is build on all samples. Later, the resultant predictions are combined using voting or averaging. Bagging is done is parallel. In boosting, after the first round of predictions, the algorithm weighs misclassified predictions higher, such that they can be corrected in the succeeding round. This sequential process of giving higher weights to misclassified predictions continue until a stopping criterion is reached.

### Random forest improves model accuracy by reducing variance (mainly). The trees grown are uncorrelated to maximize the decrease in variance. On the other hand, GBM improves accuracy my reducing both bias and variance in a model.

### Know more: Tree based modeling

### Q22. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?

### Answer: A classification trees makes decision based on Gini Index and Node Entropy. In simple words, the tree algorithm find the best possible feature which can divide the data set into purest possible children nodes.

### Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. We can calculate Gini as following:

### Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).

### Calculate Gini for split using weighted Gini score of each node of that split

### Entropy is the measure of impurity as given by (for binary class):

### Entropy, Decision Tree

### Here p and q is probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% – 50%. Lower entropy is desirable.

### Q23. You’ve built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven’t you trained your model perfectly?

### Answer: The model has overfitted. Training error 0.00 means the classifier has minimised 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 situation, we should tune number of trees using cross validation.

### Q24. You’ve got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?

### Answer: In such high dimensional data sets, we can’t use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

### To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.

### Among other methods include subset regression, forward stepwise regression.

### Q25. What is convex hull ? (Hint: Think SVM)

### Answer: In case of linearly separable data, convex hull represents the outer boundaries of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create greatest separation between two groups.

### Q26. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn’t. How ?

### Answer: Don’t get baffled at this question. It’s a simple question asking the difference between the two.

### 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.

### Q27. What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?

### Answer: Neither.

### In time series problem, k fold can be troublesome because there might be some pattern in year 4 or 5 which is not in year 3. Resampling the data set will separate these trends, and we might end up validation on past years, which is incorrect. Instead, we can use forward chaining strategy with 5 fold as shown below:

### fold 1 : training [1], test [2]

### fold 2 : training [1 2], test [3]

### fold 3 : training [1 2 3], test [4]

### fold 4 : training [1 2 3 4], test [5]

### fold 5 : training [1 2 3 4 5], test [6]

### where 1,2,3,4,5,6 represents “year”.

### Q28. You are given a data set consisting of variables having more than 30% missing values? Let’s say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?

### Answer: We can deal with them in the following ways:

### Assign a unique category to missing values, who knows the missing values might decipher some trend

### We can remove them blatantly.

### Or, we can sensibly check their distribution with the target variable, and if found any pattern we’ll keep those missing values and assign them a new category while removing others.

### 29. ‘People who bought this, also bought…’ recommendations seen on amazon is a result of which algorithm?

### Answer: The basic idea for this kind of recommendation engine comes from collaborative filtering.

### Collaborative Filtering algorithm considers “User Behavior” for recommending items. They exploit behavior of other users and items in terms of transaction history, ratings, selection and purchase information. Other users behaviour and preferences over the items are used to recommend items to the new users. In this case, features of the items are not known.

### Know more: Recommender System

### Q30. 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).

### 

### Q31. You are working on a classification problem. For validation purposes, you’ve randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?

Answer: In case of classification problem, we should always use stratified sampling instead of random sampling. A random sampling doesn’t takes into consideration the proportion of target classes. On the contrary, stratified sampling helps to maintain the distribution of target variable in the resultant distributed samples also.

### Q32. You have been asked to evaluate a regression model based on R², adjusted R² and tolerance. What will be your criteria?

Answer: Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance is desirable.

We will consider adjusted R² as opposed to R² to evaluate model fit because R² increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R² would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R² because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R² and still provide fairly good predictions, as compared to a stock market data where lower adjusted R² implies that model is not good.

### Q33. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?

Answer: We don’t use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, euclidean metric can be used in any space to calculate distance. Since, the data points can be present in any dimension, euclidean distance is a more viable option.

Example: Think of a chess board, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal movements.

### Q34. Explain machine learning to me like a 5 year old.

Answer: It’s simple. It’s just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bend position. The next time they fall down, they feel pain. They cry. But, they learn ‘not to stand like that again’. In order to avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.

This is how a machine works & develops intuition from its environment.

Note: The interview is only trying to test if have the ability of explain complex concepts in simple terms.

### Q35. I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?

Answer: We can use the following methods:

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

Also, the analogous metric of adjusted R² in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Know more: Logistic Regression

### Q36. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?

Answer: You should say, the choice of machine learning algorithm solely depends of the type of data. If you are given a data set which is exhibits linearity, then linear regression would be the best algorithm to use. If you given to work on images, audios, then neural network would help you to build a robust model.

If the data comprises of non linear interactions, then a boosting or bagging algorithm should be the choice. If the business requirement is to build a model which can be deployed, then we’ll use regression or a decision tree model (easy to interpret and explain) instead of black box algorithms like SVM, GBM etc.

In short, there is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

### Q37. Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?

Answer: For better predictions, categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

### Q38. When does regularization becomes necessary in Machine Learning?

Answer: Regularization becomes necessary when the model begins to ovefit / underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

### Q39. What do you understand by Bias Variance trade off?

Answer: The error emerging from any model can be broken down into three components mathematically.

Bias error is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have a under-performing model which keeps on missing important trends. Variance on the other side quantifies how are the prediction made on same observation different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training.

### Q40. OLS is to linear regression. Maximum likelihood is to logistic regression. Explain the statement.

Answer: OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. In simple words,

Ordinary least square(OLS) is a method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.