Data Analysis

Q1. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

- Python would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
- R is more suitable for machine learning than just text analysis.
- Python performs faster for all types of text analytics.

Q2. How does data cleaning play a vital role in the analysis?

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps increase the accuracy of the model in machine learning.
- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
- It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

Q3. Differentiate between univariate, bivariate and multivariate analysis.

Univariate analyses are descriptive statistical analysis techniques which can be differentiated based on one variable involved at a given point of time. *For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.*

The bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot. *For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.*

Multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

Q4. Explain Star Schema.

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

Q5. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For example, a researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Q6. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

Q7. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

Q8. Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are

- False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
- False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

Q9. Can you cite some examples where a false negative important than a false positive? And vice versa?

Example 1 FN: What if Jury or judge decides to make a criminal go free?

Example 2 FN: Fraud detection.

Example 3 FP: customer voucher use promo evaluation: if many used it and actually if was not true, promo sucks.

Q10. Can you cite some examples where both false positive and false negatives are equally important?

In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.
Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

Q11. Can you explain the difference between a Validation Set and a Test Set?

A Training Set:

- to fit the parameters i.e. weights

A Validation set:

- part of the training set
- for parameter selection
- to avoid overfitting

A Test set:

- for testing or evaluating the performance of a trained machine learning model, i.e. evaluating the predictive power and generalization.

Q12. Explain cross-validation.

https://machinelearningmastery.com/k-fold-cross-validation/

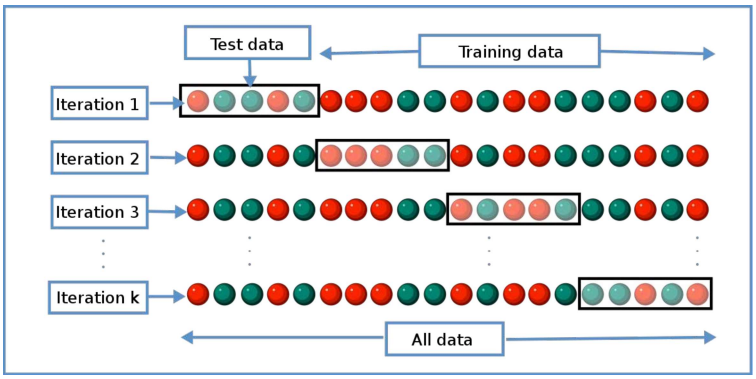
Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Mainly used in backgrounds where the objective is forecast, and one wants to estimate how accurately a model will accomplish in practice.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    a. Take the group as a hold out or test data set
    b. Take the remaining groups as a training data set
    c. Fit a model on the training set and evaluate it on the test set
    d. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores



There is an alternative in Scikit-Learn called Stratified k fold, in which the split is shuffled to make it sure you have a representative sample of each class and a k fold in which you may not have the assurance of it (not good with a very unbalanced dataset).