IWUCHUKWU COLLINS CHIJIOKE

SGA 0.8

WEEKLY CHALLENGE 1

08069817105

Question 1

Explain Data Science methodology

The first step in methodology is to understand the data he/she will be working with. When embarking on a project for example, a business project, it is important to identify the business problem and understand what the business is looking to achieve. This is actually the hardest part of the methodology as stated by John Rollins; a data scientist in IBM. He listed 10 stages of foundational methodology for data science namely;

1. **Business understanding;** this involves defining the problem, project objectives and solutions requirements from a business perspective.
2. **Analytic approach;** this is expressing the business problem in a statistical and machine learning context.
3. **Data requirements;** the choice of analytic approach determines the data requirements.
4. **Data collection;** this involves collecting the data that are relevant to the problem domain.

5. **Data understanding;** descriptive statistics and visualization techniques can help a data scientist understand data content, assess data quality and discover initial insight into the data.
6. **Data preparation;** this stage is most time consuming as it is unlikely to get a clean data ab initio. This stage involves feature engineering, data cleaning and combining data from different source. Data preparation done effectively increases the efficiency of the model prediction.
7. **Modeling;** data scientists use a training set historical data in which the outcome of interest is known to develop predictive or descriptive models using the analytic approach already described.
8. **Evaluation;** the data scientist evaluates the model quality and checks whether it addresses the business problem fully and appropriately. Doing so requires the computing of various diagnostic measures as well as other outputs such as tables and graphs – using testing set for a predictive model.
9. **Deployment;** after a satisfactory model has been developed and has been approved by the business sponsors, it is deployed into the production environment or a comparable test environment.
10. **Feedback;** by collecting results from the implemented model, the organization gets feedback on the model's performance and observes how it affected its deployment environment. Analyzing this feedback enables the data scientist to refine the model, increasing its accuracy and thus its usefulness.

Question 2

The concept of version control

Version control helps make work flow between numbers of data scientists working on a particular project. Version control enables changes to be made on a project without affecting the source code and the changes made can be tracked and traced. Version control makes backups of your work to be available and keeps history of your work and it enables you to compare your previous work and your new work, and see the changes made over time. A team member working on a project wants to experiment on the project but does not want to alter the source code, version controls allows that to be possible. Version control also makes collaboration of team members working on the same file at the same time.

They are two types of version control; centralized and distributed version control. In central version control, the entire project is stored on a central server, each person send their changes to the central copy of the project. In the distributed system, the entire project is mirrored on everyone's computer. Distributed system is faster and there is offline access to the server unlike the centralized system.

Question 3

What is the effect of improper data structure on analysis?

Data structures are the way we are able to store and retrieve data. A data structure is a specialized format for organizing, processing, retrieving and storing data. While there are several basic and advanced structure types, any data structure is designed to arrange data to suit a specific purpose so that it can be accessed and worked with in appropriate ways. That being said, the effect of improper data structure can cause a lot of problems like; bad predictions, ineffective machine learning model and algorithms, loss of data, unprocessed data and so on.

Question 4

As a data scientist, which type of database do you prefer and why?

It depends on your application and there are numerous types of database available and I don't think I will stick to a particular database in my life as a data scientist. Based on articles read and videos watched, data scientists work on both SQL and NoSQL database as per industry demand. If I am working on project that require large datasets and multi-rows, MySQL will be a preferred database to work with. If I want to work with Big Data and speed of processing is my primary emphasis then MongoDB or Hadoop which are NoSQL are better options.