CFG – Specialisation Exam

1.1 The role of a data scientist

The roles of a data scientist encompass various responsibilities:

a. Collaborate closely with the business to identify challenges and leverage data for proposing effective solutions that facilitate informed decision-making.

b. Develop algorithms and craft experiments to integrate, manage, query, and extract data, generating tailored reports for colleagues, customers, and the broader organization.

c. Utilize machine learning tools and statistical methods to create problem-solving solutions, employing techniques that harness the potential of data.

d. Evaluate and test data mining models to pinpoint the most suitable ones for implementation on a given project.

e. Maintain transparent and coherent communication, both in written and verbal forms, in order to comprehend data requirements and convey findings.

f. Construct concise and engaging reports that narrate compelling narratives about the interactions between customers or clients and the business.

g. Appraise the efficiency of data sources and data collection techniques, enhancing methods for gathering data.

h. Keep abreast of the latest technologies, techniques, and methodologies through constant horizon scanning, ensuring up-to-date expertise.

i. Conduct research to formulate prototypes and proofs of concept, deriving insights that drive innovation.

j. Identify possibilities to apply insights, datasets, code, and models across other functional areas within the organization, such as HR and marketing.

k. Sustain a sense of curiosity and enthusiasm for utilizing algorithms to tackle challenges, inspiring others to recognize the value of your contributions.

1.2 Outliers

a. Definition:

An outlier represents an observation situated at a noticeably unusual distance from the remaining values within a randomly selected sample taken from a population. To some extent, this explanation places the determination of what qualifies as abnormal in the hands of the analyst or a collective decision-making process. Prior to isolating

exceptional observations, it becomes imperative to establish the characteristics of standard or typical observations.

b. E**xamples:**

i. In a dataset indicating ages, an outlier could manifest as an age value far beyond the common range, like 180 years.

ii. When dealing with monthly incomes, an outlier might be an income value that starkly deviates from the typical income levels. In such a case, it might be beneficial to also investigate the Job title, Qualifications and years of experience.

c. Removal of Outliers

Generally, we eliminate outlier values when they arise from data entry errors, data processing anomalies, or when outlier observations are minimal in frequency.

In more detail, deciding whether to keep or toss outliers is important. However, throwing out outliers should be done cautiously and only when we're really sure they're mistakes. We usually can't be sure when we're analyzing data.

Sometimes, outliers show that something went wrong when collecting data. Other times, they affect the data, so keeping them helps us understand the whole picture.

Here's when you should or shouldn't get rid of outliers:

Get Rid of Outliers if:

- You're Certain They're Wrong

For example, if you know what range your data should be in, like people's ages, you can safely remove values that don't fit that range.

- You Have Lots of Data

When you have a lot of data, removing a doubtful outlier won't harm your results much.

- You Can Double-Check

If you can go back and check those strange data points, then you might remove them.

Don't Get Rid of Outliers if:

- Your Results Are Super Important

When results are really important, even small changes matter.

For instance, you might feel okay removing outliers from a dataset about people's favorite TV shows, but not when dealing with temperatures that make airplane seals fail.

-There Are Lots of Outliers

Outliers are rare. If, for instance, a big chunk—say, 25%—of your data is outliers, something interesting is likely going on. This is similar to what we talked about earlier.

Deciding whether to toss out outliers or not depends on the data and what those outliers mean in the big picture.

d. Possible issues that can be found in a dataset include:

1. Inconsistent Data: The presence of both structured and unstructured data within the dataset can lead to challenges. Unstructured data, which encompasses irregular, or inconsistent data, requires careful handling and preprocessing before effective analysis can take place.

2. Data Types: Inconsistent data types in a dataset create confusion and hinder accurate analysis. Misinterpreting categorical as numerical or mishandling dates can lead to flawed insights. Ensuring proper data typing is essential for maintaining data integrity and facilitating accurate processing and analysis.

3. Handling Bulk Data Quantity: Organizations often deal with enormous quantities of data, ranging from terabytes to petabytes, gathered from diverse sources. Formatting and preparing such large volumes of data for analysis can be challenging, even with tools designed to simplify the process.

4. Misspelling: Typing errors and misspelled words are common data issues. While automated tools can rectify spelling and grammatical errors for common words, addressing these issues within a vast database, especially in fields like addresses and names, can be demanding.

5. Duplication: Instances of duplicated data appearing multiple times within the same database can lead to redundancy and complicate analysis efforts.

6. Contradiction: Contradiction errors emerge when real-world entities are represented by different values within the dataset, often causing confusion and inconsistencies.

7. Incorrect Reference: Errors tied to incorrect data validation can result in data mismatches. For example, entering an incorrect department name might lead to validation errors that produce inaccurate outcomes.
8. Missing Data: Incomplete or absent data can introduce bias and inaccuracies if not appropriately managed.

9. Data Skewness: Asymmetrical data distributions can impact certain statistical analyses, necessitating suitable adjustments.

10. Data Scaling: Features with varying scales can disrupt algorithms reliant on proximity, like clustering or gradient descent.

1.3 a. Data Cleaning

Data cleaning, also known as data cleansing or data wrangling, is an important first step in analyzing datainvolves repairing or eliminating inaccurate, corrupted, improperly structured, copied, or partial data in a dataset.

b. Importance of Data Cleaning:

Data cleaning is important because if your data isn't good enough, then any analysis you do with that data will also be wrong. Even if you do everything else perfectly in the data analysis process, messy data will mess up your results.

Clean data is essential for accurate and reliable insights in data analysis. It also brings several other advantages:

1. Organization: Businesses gather a lot of information from clients, customers, etc. Cleaning this data keeps it organized, making storage more efficient and secure.

2. Error Prevention: Clean data isn't just for analytics; it avoids mistakes in daily operations. Accurate databases help teams avoid errors like sending personalized mail to the wrong person.

3. Productivity Boost: Regularly cleaning and updating data removes irrelevant information, saving time and effort in searching for what's needed.

4. Cost Savings: Bad data can lead to costly mistakes. Detecting and correcting issues early avoids bigger problems later on.

5. Enhanced Mapping: When building data infrastructures or applications, clean data streamlines the process, making data modeling and mapping easier.

In short, clean data isn't just about analysis—it improves efficiency, accuracy, and cost-effectiveness across the board.

c. Mistakes commonly seen in datasets.

Frequently encountered errors in datasets encompass:

1. Missing Data: Absences of information that can distort analysis and modeling outcomes.

2. Data Entry Mistakes: Typos, inaccuracies, and misspellings leading to inconsistencies.

3. Duplicates: Repetitive records causing redundancy and calculation errors.

4. Formatting Inconsistencies: Variations in data representation causing confusion.

5. Outliers: Unusual values distorting statistical analysis.

6. Incorrect Data Types: Misassigned data types leading to misinterpretation.

7. Contradictions: Conflicting data about the same entity resulting in inaccuracies.

8. Data Imbalance: Uneven class representation impacting classification models.

9. Biased Collection: Unequal demographic representation leading to skewed conclusions.

10. Incomplete Data: Limited population representation affecting generalization.

11. Data Leakage: Accidental inclusion of future information during model training.

12. Sampling Bias: Non-random sampling not reflecting the true population.

1.4 Unsupervised Learning:

a. Unsupervised learning is a branch of machine learning that discovers patterns within datasets without external guidance. It identifies patterns in unclassified and unlabeled datasets, making it valuable for uncovering patterns, groupings, and distinctions in unstructured data, such as customer segmentation, exploratory data analysis, or image recognition. Unsupervised learning algorithms can categorize, label, and group data points independently, relying solely on inherent patterns to do so.

b. Applications of unsupervised learning are diverse, including data exploration, customer segmentation, recommender systems, targeted marketing campaigns, as well as data preparation and visualization. For instance, it can unveil patterns in customer data for targeted marketing efforts and cluster similar products to enhance recommendation systems. In summary, unsupervised learning excels at identifying patterns, groupings, and disparities in unstructured data.

c. Real-world use cases encompass clustering, visualization, dimensionality reduction, finding association rules, and detecting anomalies. For example, this is useful in market basket analysis, where the goal is to discover frequently co-purchased products. Another instance involves...

d. Limitation: Unsupervised learning faces primary constraints due to its need for substantial data volume to be effective and the challenge of evaluating outcomes. With no predefined classes or attributes, determining result significance can be intricate.

1.5 Supervised Learning:

a. Supervised learning constitutes a category of machine learning wherein algorithms learn from labeled data. The algorithm is trained on datasets containing both input and corresponding output data. The core objective of supervised learning is to learn a function that maps inputs to outputs. Most supervised learning algorithms fall under classification or regression.

b. Supervised learning is employed when algorithms are trained on datasets comprising input data alongside corresponding output data.

c. In practical scenarios, supervised learning finds applications in image recognition, speech understanding, natural language processing, and recommendation systems. For instance, consider a spam detection system using supervised learning. This system can be trained on a dataset of emails labeled as spam or non-spam. Its aim is to develop a function capable of predicting whether new emails are spam, based on patterns extracted from the training data.

d. Data requirements for supervised learning vary by application. Typically, effective supervised learning algorithms demand labeled data, implying a specified target variable. Some preprocessing might be necessary to prepare data for analysis.