Slide-1 || বাদ

Slide-2

Data Analytics 6 steps:

1. Discovery:
   objective: To understand the business domain and objectives.
   Explanation: In this phase we gather information about the problem, set project goals and asses if analytics will help answer the business questions. This involves defining the project scope, identifying key stakeholders and formulating hypothesis.

2. Data Preparation:
   objective: To clean and organize the data.
   Exploration: This step involves preparing the data for analysis by cleaning, transformation, and integrating data from multiple sources, It ensures that the dataset is well-structured and ready for analysis.

3. Model Planning:
   objective: To select the right method and algorithms.
   Explanation: This phase selects algorithms, techniques, and tools to be used. You plan how the data will be processed, and create a blueprint for model building, including EDA to uncover patterns.

4. Model Building:
   objective: To develop and train model
   Explanation: This is where the actual development of the models take place. You use machine learning, statistical technique and other algorithm to build a predictive model, which continuosly iterating and validating them to improve performance.

5 Communicating Results:

objects: To present insights and finding to stakeholders.

Explanation: The results of the data analysis are communicated to stakeholders in a clear and actionable way, often through reports or visualizations. The step focuses on making sure that stakeholders understand the finding and how to apply them in decision-making.

6. Operationalizing:

objective: To deploy the model and monitor it's performance

Explanation: Once the model is validated and results are communicated, the next step is to deploy it in the operational environment. This stage also involves setting up processes for ongoing monitoring and maintainance to ensure the model performs well in production.

Sandbox: A data sandbox is an isolated environment created using real-world data, specifically designed for tasks like exploration and learning. The main feature of the sandbox is that it operates independently of live production environments. The sandbox allows analysts and data scientists to test and explore datasets without affecting live processes. It ensures an additional layer of protection to real world data.

A sandbox transformation refers to the process of manipulating or transforming data within a data sandbox environment.

⇒ ETL (Traditional approach), here data is first extracted, Transformed (cleaned, modified) and then loaded into a database or storage system.

Strengths: ① The data is pre-processed and optimized for storage and queing.

② Suitable for situations where structured and well organized data is required upfront.

Limitations: ① Transforming data before loading might result in data loss, especially if outliers or irregular data points are mistakenly removed.

⇒ ELT (Sandbox approach), where data is extracted and loaded first, and then transformation happens later inside the database.

Strengths: ① Raw data remains available in the sandbox, so transformation are flexible and can be done iteratively.

② Ideal for big data analytics and fraud detection where transformations should only occur after deeper analysis.

Limitations: ① Might require more storage since raw data is loaded first.

⇒ Fraud detection of credit card requires analyzing Raw data to capture anomalies and outliers. Early transformation in ETL might accidentally remove unusual but crucial transactions, mistakenly identifying them as errors. So, ELT is ideal for this kind of use cases where preserving raw data and exploring anomalies is crucial. ETL is well suitable for structured data analysis where data transformation is straightforward.

⇒ Butterworth filter is a type of filter that is designed to have a very smooth frequency response. This means it doesn't have sharp cuts but transitions smoothly, which helps avoid distortions in the signal. Imagine you are processing audio, and you want to remove high-pitched noise from a recording. You can apply a low-pass Butterworth filter with a cutoff frequency.

$$H(j\omega) = 1/(1+(\omega/\omega_c)^{\wedge}(2n))$$

; ω_c = cutoff frequency

n = order of the filter, determines how sharp or smooth the response will be.

$H(j\omega)$ = how much the filter will let a specific frequency pass through

Quiz-1

Q-1: # Data Analysis as linear Process (when objective is predetermined)

- This approach follows a strict sequence from defining the purpose, forming questions, data collection, data analysis, interpretation (interpreting the results) with findings, writing, Reporting, disseminating and finally evaluation.

Advantages:

1. Easier to manage and schedule because each step is clearly defined and follows logically from the one before.
2. Simpler to teach and understand because of it's straightforward, step by step nature.

Disadvantages:

1. Once a step is completed, it's often not revisited; this can limit the ability to adapt to new information or changes in the data environment.
2. The linear nature might prevent revisiting initial assumptions or incorporating new insights generated late in the process.

Necessary in situations:

1. This process works well in situations where requirements are clearly defined from the outset and stability exists in dataset and research objectives.

Example: Annual Financial Reporting

Financial reporting follows strict regulatory requirements and deadlines. The data and objectives are clear and the process needs to be systematic and predictable to meet the objective.

[II] Data Analysis as a Cycle (when iterative feedback are necessary)

— The approach views data analysis as a cycle with interconnected steps that include data collection, analysis, finding interpretations, reporting and disseminating results, followed by evaluation and formulation of new questions.

Advantages:

① Allows for returning to earlier stages based on findings and feedback, which can lead to more thorough understanding.

② Supports ongoing adjustments, making it highly effective in dynamic environments where data inputs and conditions change over time.

Disadvantages:

① less structured in terms of project management and timeliness, potentially leading to longer project duration.

Necessary in situations:

① This approach is more suited for exploratory research where ongoing insights and developments inform the analysis continuously.

Example: A tech company is developing a new app and uses ongoing user testing to refine features. User feedback and behavious can lead to a significant changes in the app's interface. Here, Cyclical process allows for iterative testing and reevaluation of data to continuously improve the product.

Data Conditioning: refers to the series of actions taken to prepare raw data for analysis. This preperation makes the data cleaner, more organized and more suitable for specific tasks.

Q-2: LMF: Local Maximum Fitting is an advanced algorithm developed to process time series data from remote sense (RS) images. The main goal is to mitigate the impact of atmospheric conditions such as clouds and haze that can obsecure true data.

Q-3: Low Pass Filter: It passes signals with a frequency lower than the cutoff frequency and attenuates signals with frequency higher than the cut-off frequency. (SMA) Simple moving Average → linear filters.

Q-4: High Pass Filter: Passes signals with a frequency higher than a certain cutoff and attenuates signals with frequency lower than the cut-off frequency.

Discretization: is the process that transforms a numerical feature into a discrete feature. Simply, it creates bins containing all the values of a feature.

1. Class Noise

This type of noise occurs when the **class label** assigned to an instance (data point) is incorrect or inconsistent.

Types of Class Noise:

- **Contradictory Instances**: The same data point is labeled with different class labels in the dataset, leading to inconsistencies.
  - **Example:**
    - Instance 1: (Att1 = 0.25, Att2 = red) → **Label: Positive**
    - Instance 2: (Att1 = 0.25, Att2 = red) → **Label: Negative**
      Both instances are the same (Att1 and Att2 have identical values), but one is labeled "positive" and the other "negative." This contradiction creates confusion in classification.
- **Mislabeled Instances**: The class label for an instance is simply wrong.
  - **Example:**
    - Instance: (Att1 = 1.02, Att2 = green) → **Label: Positive**
      However, based on business rules or actual data, this instance should have been labeled "negative." Mislabeling results in inaccurate training data, which can affect model performance.

2. Attribute Noise

This type of noise occurs when one or more **attributes** (input features) of an instance are incorrect, missing, or irrelevant.

Types of Attribute Noise:

- **Erroneous Values:** The attribute value is incorrect or falls outside the expected range.
  - **Example:**
    - Instance: (Att1 = 2.05, Att2 = green)
      Here, Att1 has a value of 2.05, which may be beyond the expected range (if Att1 is supposed to be between 0 and 1, for example), making it an erroneous value.
- **Missing Values:** One or more attribute values are missing from the instance.
  - **Example:**
    - Instance: (Att1 = 1.02, Att2 = ?)
      Here, Att2 is missing or unrecorded, making it harder for the classifier to make a proper decision.
- **"Don't Care" Values:** These values are irrelevant to the classification task but are still present in the dataset.
  - **Example:**
    - Suppose there is an attribute that has little or no impact on the classification task, like color (Att2 = green) in a dataset focused on numerical attributes.

☐ Techniques to identifies class noise:

i) Ensemble Techniques: Multiple classifiers are used to detect mislabled data.

Bagging: Bagging involves creating multiple models is parallel and then combining their outputs. After training prediction from each model are combined. Each model votes for a class for classification but for regression average is taken.

Boosting: involves building models sequentially by applying weights to instances in the dataset, with each subsequent model focusing more on instances that previous models misclassified.

ii) Distance based techniques: These technique rely on the assumption that similar instances will belong to the same class. It measures distance between instances to identify outliers.

• KNN

iii) ~~Sig~~ Single Learning based techniques: Single classifiers.

III Dimensionality Reduction Isomap

Step-1: Neighbourhood Graph:

ISOmap starts by constructing a neighbourhood graph where each point is connected to it's nearest neighbours.

Two ways to find neighbours

- ϵ-ball approach: For each x_i, another point x_j is close if and only if $\|x_i - x_j\| \le \epsilon$ or
- KNN approach: For each point x_i, x_j is close if it is among the ~~thi~~ k-nearest neighbours of x_i.

Construct a neighbourhood graph G from the given distance $d_X(i,j)$ using the specified method.

Step-2: Geodesic distance:

It then estimates the geodesic distances between all pairs of points in the graph. In practice, these distances are approximated using the shortest path through the Graph, which can be computed with algorithms like Dijkstra's.

Compute the shortest-path distance $d_G(i,j)$ between all vertices of G by using Dijkstra's algorithm.

Step-3: Multidimensional Scaling: MDS

Finally, ISOmap uses these geodesic distance estimates to embed the data into a lower-dimensional space through a process called MDS. MDS seeks to place each data point in a new, lower-dimesional space such that the distances between points are preserved as well as possible.

Apply MDS with $d_G(i,j)$ as input distances to find a k-dimensional representation Y of the original data.

☐ Local Linear Embedding (LLE): is an unsupervised learning algorithm designed to reduce dimensionality while preserving the geometric features of the original dataset.

Step-1: Finding the k-nearest neighbours for each data points. If k is chosen to be too small or too large, it will not be able to accomodate the geometry of the original data.

Step-2: A weight matrix W is computed where each element w_{ij} represents the contribution of the j-th data point to the i-th data point's neighbourhood. Weights are computed in such a way that they minimize the cost of reconstructing each point from it's neighbours, subject to the constraint that the sum of the weights for each point is 1.
weights are assigned as zero if a point is not a neighbour of the considered point.

$$\text{error}(W) = \sum_i \left(x_i - \sum_j w_{ij} x_j\right)^2$$

$x_i = \sum_j w_{ij} \cdot x_j$ Such that $\sum_j w_{i,j} = 1$ and $w_{i,j} = 0$ if x_j is not a linear neighbour of x_i

Step-3: Using the W, the algorithm seeks a set of points in a lower-dimensional space that best preserves the local neighbour structure. if y_i is the vector in the lower-dimensional space that correspond to x_i and Y is the new-data matrix whose ith row is Y_i, then this can be accomplished by finding a Y the minimizes the following equations.

$$\text{error}(Y) = \sum_i \left(Y_i - \sum_j w_{ij} Y_j\right)^2$$

CamScanner

SMOTE: Synthetic Minority Oversampling Technique

Q:5 # Why do we consider all attributes but not one.

- Preserving data distribution: SMOTE aims to generate synthetic samples that are realistic and representative of the underlying feature space of the minority class. By considering all attributes the synthetic samples maintain the multidimensional distribution of the data.
- Reflecting Complex Dependencies: Real world data often involves complex interactions and dependencies among features. By using all attributes, SMOTE ensures this dependencies.
- Balanceing the Data: It balances the class distribution.

Q:6 What Problem occurs in One-hot encoding?

- Multicolinearity occurs when two or more predictor variables in a statistical model are highly co-related meaning one can be linearly predicted. This can create problems in regression analysis because it makes it difficult to determine the effect of each individual predictor variable on the outcome.

⇒ SMOTE ALGO:

```
Populate (N,i,nnarray) (*Function to generate the synthetic samples *)
While (N ≠ 0)
    choose a random number between 1 and k, call it nn. The step chooses one of
    the k- ~~neigh~~ nearest neighbours of i
    for attr ← 1 to numattribute
        Compute: diff = Sample [nnarray [nn]] [attr] - Sample[i] [attr]
        Compute: gap = random number between 0 and 1
        Synthetic [newindex] [attr] = Sample[i] [attr] + gap * diff
    end for
    newindex++
    N = N - 1
endwhile
```

Pearson's co-relation:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Spearman's Rank Coefficient:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$;

ρ = Spearman rank correlation

d_i = the difference between the ranks of corresponding variables

n = number of observations.

Normalization
Standardization: (From PDF)

Normalization	Standardization
① Rescales values from 0 to 1.	① Centers around the mean and scales to the standard deviation of 1.
② Useful when the distribution of the data is unknown or not Gaussian.	② Useful when the distribution of the data is Gaussian or unknown
③ Sensitive to outliers	③ less sensitive
④ Retains the shape of original distribution.	④ changes the shape
⑤ May not preserve the relationships between the data.	⑤ Preserves
⑥ Equation: $\frac{(x - min)}{max - min}$	⑥ Equation: $\frac{(x - mean)}{\text{Standard deviation}}$

Regularization techniques are commonly used to mitigate multicollinearity in machine learning and regression models. These techniques add a penalty to the regression model's loss function, which discourages overly complex models by shrinking the regression coefficients. Regularization helps handle multicollinearity by reducing the influence of less important or highly correlated features.

Main Regularization Techniques:

1. Ridge Regression (L2 Regularization)

- **How it works**: Ridge regression adds a penalty equal to the square of the magnitude of coefficients (i.e., L2 norm). The objective function is modified as follows:

$$\text{Minimize } (RSS + \lambda \sum_{j=1}^{p} \beta_j^2)$$

  Where:

  - **RSS**: Residual Sum of Squares (original loss function in regression)
  - **λ** (lambda): Regularization parameter that controls the strength of the penalty
  - **β**: Coefficients of the regression model

- **Effect on Multicollinearity**: By shrinking the coefficients, ridge regression reduces the variance introduced by highly correlated features. It doesn't completely eliminate any coefficients but pulls them toward zero. This helps stabilize the estimates in the presence of multicollinearity.
- **When to use**: Ridge regression is useful when most variables are important but you need to handle multicollinearity. It's better suited when all features contribute to the output but with varying degrees of importance.

2. Lasso Regression (L1 Regularization)

- **How it works**: Lasso (Least Absolute Shrinkage and Selection Operator) regression adds a penalty equal to the absolute value of the coefficients (L1 norm). The objective function becomes:

$$\text{Minimize}\ (RSS + \lambda \sum_{j=1}^{p} |\beta_j|)$$

- **Effect on Multicollinearity**: Unlike Ridge, Lasso can shrink some coefficients to zero, effectively performing **feature selection**. It helps by eliminating less important or redundant features, especially if they are highly correlated with others.

- **When to use**: Lasso is useful when you expect some features to be irrelevant or redundant, and you want to shrink those coefficients to zero. It's beneficial for simplifying models and can be very effective in reducing multicollinearity by selecting only the most important features.