Date / /
Page No.
Shivalal

EDA
Exploratory Data Analysis

EDA is a critical step in data science. It involves analysing and visualising data to understand its key characteristics, uncovered patterns, and identifying relationship b/w variables.

Key aspects of EDA :-

① Distribution of data (Investigating the distribution of data points to understand their range, central tendencies [mean, median, mode] and its dispersion.

② Graphical representation (utilizing plots such as histograms, box plots, scatter plots and bar charts to visualize relationship within the data and distribution of variables.

③ Outlier Detection
[All noise is outlier but not all outlier are noise].
[Identifying abnormal values that deviates from other data points.

④ Correlation Analysis (checking the relationship b/w variables to understand how they might affect each other. This includes calculating correlation coefficients and creating correlation matrices.

⑤ Handling missing values (Detecting and deciding how to handle missing data points, whether by imputation or removal, depending on their impact & the amount of missing data.

Date / /
Page No.
Shivalal

⑥ Summary Statistics

Calculating key points statistics that provide insights into data trends and patterns.

⑦ Testing Assumptions

(Many statistics tests and models assume that data meets certain conditions such as normal data distribution. EDA helps to verify these assumptions.)

Importance of EDA -

①
② Identifying patterns and relationships.
③ Detecting anomalies and outliers.
④ Testing assumptions.
⑤ Feature selection
⑥ Optimizing model design.
⑦ Performing data cleaning
⑧ Enhancing Communication.

Types of EDA -

1. Univariate Analysis , Ex - Histogram
2. Bivariate Analysis , Ex - Scatter Plot
3. Multivariate Analysis , Ex - Heatmap

Depending on the no. of columns, we are analyzing or we can divide EDA into three types which are named above.

Date / /
Page No.
Shivalal

* Kernel Trick Contra

1. Univariate Analysis - Univariate Analysis focusses on a single variable to understand its internal structure. It is mainly concerned with describing the data and finding patterns existing in a single feature or attribute.
It involves summarising and visualising a single variable at a time to understand its distribution, central tendency, hidden patterns etc.
Common techniques of univariate analysis are as follows -
1. Histogram
2. Box plot
3. Bar chart
4. Statistical summary i.e. mean, median, mode, SD and Variance.

2. Bivariate Analysis - It involves ~~storing~~ exploring the relationship between two variables. It enables to find associations, correlation and dependencies b/w the pair of variables.
Some common techniques are -
1. Scatter Plot
2. Correlation Coefficients
3. Cross Tabulation
4. Line Graph
5. Covariance

3. Multivariate Analysis – It examines the relationship b/w more than two variables in the data set. It aims to understand how variables interact with one another, which is crucial for most statistical modelling techniques.
Example are given below –

1. Pair Plots
2. PCA i.e. Principal Component Analysis
3. ICA i.e. Independent Component Analysis
4. Fisher Discriminant Analysis
5. Relief Algorithm

Tools for performing EDA

1. Python libraries i.e, Matplotlib, Seaborne, numpy etc.
2. R packages i.e; GGPlot 2, ~~DELYR~~, tidyir, dplyr
3. Weka
4. Orange
5. Mini Tab
6. Rapid Miner

Steps for performing EDA –

Step 1 → Understand the problem and the data
Step 2 → Import and investigate the data.
Step 3 → Handle the missing values
Step 4 → Explore data characteristics.
Step 5 → Perform data transformation
Step 6 → Visualizing the data
Step 7 → Handling outliers
Step 8 → Communicate findings and insight.

Date / /
Page No.
Shivalal

- Multi-collinearity - It generally occurs when the independent variables in a regression model are correlated with each other.
This correlation is not expected as the independent variables are assumed to be independent.
If the degree of the correlation is high, it may cause problems while predicting results from the model.

- Few consequences of multi-collinearity

1. The estimators have high ~~co~~variances and covariances which makes the precise prediction difficult.

2. Due to the above consequence in pt 1, the confidence level/intervals tend to become wider which leads to the acceptance of the zero (null hypothesis) more often.

3. The standard errors can be sensitive to the small changes in data.

4. The coefficients become very sensitive to small changes in the model.

5. The impact of a single variable become difficult to distinguish from the other variable.

Variance Inflation factor (VIF)

Variance Inflation factor - It is used to test the presence of multi-colinearity in a regression model. For a regression model VIF is defined as -

$$VIF = \frac{1}{1 - R^2}$$

Where,

$$R^2 = \Sigma (Y_{calculated} - \bar{Y})^2$$

If the value of VIF is ~~greater than~~ less than or equal to 1

(i) $VIF \leq 1$;
~~It is non-correlated~~

(ii) $1 < VIF \leq 5$
Low Multicolinearity exists

(iii) $VIF > 5$;
~~High multi-colinearity exists~~

The inverse of VIF is ~~fact~~ called tolerance which is given as

$$\text{Tolerance} = \frac{1}{VIF} = (1 - R^2)$$

Note :- When $R^2 = 0$, no colinearity and high tolerance (no correlation b/w the variables)

Date / /
Page No.
Shivalal

Detection of multi-colinearity

There are several methods that are used for detection of multi-colinearity. For example.

1. Manual Method - VIF, T-Statistics, Pairwise Correlation
2. Automatic Method ↴
3. Recursive Feature elimination, Relief algorithm, (Attribute Selection methods, feature extraction method)

Missing data handling -

Missing values are data points that are absent for a specific variable in a dataset. They can be represented in various ways such as - null values, blank cells or special symbols like NA or unknown. These missing data points creates a challenging situation in data analysis and can lead to inaccurate or biased results

Effects of missing values

1. Reduce the data sample size, ~~data~~ model accuracy, and reliability.
2. Introduce bias.
3. It makes to perform statistical analysis difficult.

Types of missing values –

1. Missing completely at random (MCAR) – It is a special type of missing data in which the probability of a data point being missing is entirely random and independent of any other variables in the dataset.

2. Missing at random – It is a type of missing data where the probability of missing values depends on the value of other variables but not on the missing variable itself. This means that the missing mechanism is not entirely random but it can be estimated based on the available information.

3. Missing Not At Random – It is the most challenging type of missing data to deal with. It occurs when the probability of data point being missing is related to the missing value itself.

Methods for identifying missing data in pandas –

1. .isnull()
2. Not Null
3. info()
4. .isNA()
5. .dropNA()
6. .fillNA()
7. replace
8. drop-duplicates()
9. unique()

Use the company sales.csv data and do the following

Common representation of missing values -

1. Blank cells
2. Specific values like null, NaN, -9999.99 are used to represent the missing data.
3. ~~Codes or flags are~~
3. Non-numeric codes or flags can be used to indicate different types of missing values.

Methods for handling missing data values :-

1. Eliminating rows with missing values
   - Simple and efficient.
   - Not recommended for smaller datasets.

2. Imputation methods
   a) Replacing missing values with estimated values
   - Preserves sample size
   - Can introduce biasness

* There are some popular imputation methods which are in practice

- Mean, median and mode imputation
- Forward and backward fill.

3. Interpolation techniques
   - Estimates missing values based on the surrounding data. Ex - linear interpolation, Spline interpolation etc.

- More sophesticated than mean/median imputation because it captures relationship b/w variables.
- Requires computational resources.

Impact of missing data

1. Reduced data quality.
2. Reduced model performance.
3. Increased bias.
4. Loss of data integrity.
5. Unreliable ~~discuss~~ data summarization.

Outliers

An outlier is a data point that significantly deviates from the rest of the data. It can be either much higher or much lower than the other data points but its presence has significant impact on the performance of machine learning algorithms. They can be caused during measurements or execution.

Types of outliers -

1. Global outliers - They are isolated data points that are far away from the main body of the data. These are often easy to detect and eliminate.

Date / /
Page No.
Shivalal

2. Contextual outliers – These data points that are unusual in a specific context but have significance in different context. They are often more difficult to identify and eliminate as they require additional domain knowledge or expertise to determine their significance.

PAGE
DATE

30/01/24

D.V

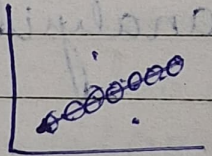
Methods of outlier detection:-

① Statistical method:-
a) Z-score
b) IQR, etc.

② Distance based methods: a) K-nearest neighbour (KNN).

→ KNN identifies outliers as data pts where k nearest neighbour are far away from them.

③ Local outlier factor (LOF) → calculates the local density of data pts & identifies outliers as those with significantly lower density compared to their neighbours.



④ Cluster based methods → K-means, ~~hierarchi~~ heirarchal clustering.

⑤ Other methods → (i) isolation forest → randomly isolates data pts by splitting features & identifies outliers as those isolated quickly & easily.

PAGE
DATE

2/9/24

S.S

Q. $A = P\left(1 + \frac{RT}{100}\right)$.

3/9/24

D.V

① One class Support vector machine :- learns a boundary around the normal data & identifies outliers as data pts. falling outside the boundary.

- Imp of outlier detection in ML :-

① To reduce bias from ML model.
② to improve performance of ML model.
③ to reduce data variance
④ to increase reliability of ML models, increase data analysis accuracy, and improve learning.

- Imbalanced data condition occurs when the distribution of data samples in diff. classes are uneven. In other words, some classes will have significantly higher no. of data samples. While the other classes have a lower count. The classes having larger no. of data samples are known as majority classes, while the other one are called minority classes.

$$\left[F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{P + R}\right]$$

PAGE
DATE

- Accuracy = $\frac{\text{Correct classification}}{\text{Total}}$

$$= \frac{T(P) + T(N)}{T(P) + T(N) + F(P) + F(N)}$$

- T(P) : true positive rate (TPR) of all the actual +ve's that were classified as +ve's.

Recall

$$\left[TPR = \frac{\text{Correctly classified +ve's}}{\text{All actual +ve's}}\right]$$

$$= \frac{T(P)}{T(P)} = \frac{T(P)}{T(P) + F(N)}$$

- F(N) : false -ve's : were actual +ve's misclassified as -ve's, which is why they appear in denominator.

- FPR → is the portion of all actual -ve's that were classified incorrectly as +ve's, also known as probability of false alarm.

$$FPR = \frac{\text{incorrectly classified actual-ves}}{\text{actual -ves}}$$

$$= \frac{F(P)}{F(P) + T(N)}$$

- Precision :

4/9/24

D.V

- **Precision** → is the proportion of all the models +ve classifications that are actually +ve.

Prec. = $\frac{\text{correctly classified actual +ve's}}{\text{Everything classified as +ve}}$

$= \frac{T(P)}{T(P) + F(P)}$

i.e. $f1 = 2 \times \frac{\frac{T(P)}{T(P)+F(P)} \times \frac{T(P)}{T(P)+F(N)}}{\frac{T(P)}{T(P)+F(P)} + \frac{T(P)}{T(P)+F(N)}}$

- **Matthew Correlation Coeff. (MCC):**

$$MCC = \frac{[T(P) \cdot T(N) - F(P) \cdot F(N)]}{\sqrt{[T(P)+F(P)] \cdot [T(P)+F(N)]}}$$

$$MCC = \frac{(T.P \times T.N - F.P \times F.N)}{\sqrt{(T.P + F.P)(T.P + F.N)(T.N + F.P)(T.N + F.N)}}$$

Range : (-1 to 1).

PAGE
DATE

Actual \ Estimated		
	15	5
	5	375

Actual \ Estimated	P	N
P	15 - T.P	5 - F.N
N	5 - F.P	375 - T.N

M.C.C = ?

$$\Rightarrow \frac{15 \times 375 - 5 \times 5}{\sqrt{20 \times 20 \times 380 \times 380}}$$

$$= \frac{5600}{\sqrt{400 \times 380 \times 380}} \Rightarrow \simeq 0.7$$

- G-mean = $\sqrt{T.P.R \times T.N.R}$

$$= \sqrt{\frac{T.P}{(T.P+F.N)} \times \frac{T.N}{(F.P+T.N)}}$$

Range: (0 to 1)

② Resampling (oversampling + undersampling): adjusting the balance b/w this method involves minority & majority classes through undersampling or oversampling, increasing (replicating) the minority class samples is known as oversampling whereas reducing the

This method ~~revolves~~ involves adjusting the balance b/w majority & minority classes. through undersampling or oversampling.

Increasing [replicating] the minority class samples is known as oversampling whereas reducing the majority class sample is known as undersampling.

Balance in Sample classifier

When dealing with the imbalance dataset, traditional classifier tend to favour the majority classes, neglecting the minority classes due to its lower representation

In sample methods are available which splits the training or imbalance training data into balanced subset.

These balance subset are utilized for training the base classifiers/base learners.

This reduces the adverse effect of imbalance ~~for~~ training data.

Scikit learn library provides a BalancedBagging-Classifiers to adjust address the imbalance data specifically

issue. It introduces parameters like → "Sampling_Strategy" determining the type of re-sampling and "re-placement" controlling whether the data sampling should occur with or without replacement

These parameters incorporates additional balancing during training phase.

This ensures a more suitable treatment of classes, particularly beneficial while handling imbalance dataset [similar parameter options are available with Random Forest classifier]

Synthetic Minority Oversampling Technique

SMOT addresses imbalance dataset by synthetic generation of new data points for the minority classes.

It enhances the data diversity by creating new artificial data points based on the distribution of each minority class.

In simpler terms, SMOT examines data points in the minority and select a random nearest neighbour using KNN Algorithm and generates synthetic data points randomly within the feature space

(ii) Threshold Moving → In classifiers, predictions are often expressed as probabilities of class memberships

Note: ROC or Area under ROC

↳ The conventional threshold for assigning predictions to classes is typically set at 0.5.

↳ However, Incase of imbalance class problems, this default threshold may not eat optimum result. To enhance classifier performance, it is essential to adjust the threshold value that efficiently discriminate b/w 2 classes.

Techniques such as ROC curves and to identify the optimal threshold

Using Tree Based Models

1. Decision - Tree
2. Random Forest
3. Gradient Boosted Tree
4. Xg Boost
5. EXtra Tree [Extremely Randomized Tree]
6. Using Anomaly Detection Algorithm

These techniques helps to identify minority classes data points as outliers [Rare datapoints]

In imbalance dataset, Model assumes majority class data points as normal data points and minority class data points as outlier data

Anomaly detection Algorithm prevents minority class data points as outliers

That is all about enlarging data for different Kinds of Purposes.

The data analysis process involves respecting, cleaning, transforming and modeling data to draw useful insights from it