Date: / /

Data Science

Exploratory Data Analysis (EDA)

(1.2) EDA & objectives & why we need to perform EDA

EDA = involves using statistics and visualisations to analyze and identify trends in data sets.

Objectives -

EDA allows data scientists to get deep insight into a data set & at the same time provide some specific outcomes that a data scientist would want to extract from data set. It includes -

- Helps identify errors in data sets
- Gives better understanding of data set
- Helps detect outliers or anomalous event
- Helps understand data set variables & the relationship among them.

* delve into data set
* examine interrelationships among the attribute
* identify interesting subsets of observations.

MATRIKAS

Date: / /

Need to perform EDA -

The primary objective of EDA in order to perform EDA is to uncover underlying structure.

The structure of the various data sets determines trends, patterns & relationship among them. (Business)

Therefore, performing EDA allows data scientist to detect errors, debunk assumptions & much more; to ultimately select an appropriate predictive model.

(B) EDA v Hypothesis testing, why prefer EDA.

Hypothesis tests the relationships between the variables. While EDA is an approach for data analysis that employs a variety of techniques (mostly graphical & statistical) to maximize - insight into data set.

uncover underlying structure.

extract important variables

detect outliers & anomalies

develope accurate models.

MATRIKAS

Date: / /

4. Contigency Table - also called crosstabs - are used in statistics to summarize the relationship betn several categorical variables.

It is a special type of frequency distribution table, where two variables are shown simultaneously.

- It is essentially a display format used to analyze & record the relⁿ betⁿ two or more categorical variable. It is a categorical equivalent scatterplot used to analyse the relⁿ betⁿ two continuous variables

advantage - it allows one to more easily perform basic probability calculations

5. Row percentage

It is computed by dividing the count for a cell by the total sample size for that row

It is the percent of that each cell represents of the row total

Column Percentage

It is computed by dividing the counts for an individual cell by the total no. of counts for the column.

It is the percent that each cell represents of the column total.

MATRIKAS

Date: / /

6. ~~Normalised~~ overlay Histogram:

A normalised overlay histogram to enhance the pattern of the target variable and quantify the relationship between the numerical predictor & the target variable.

⑦ & Now ?

⑧ Now EDA would help to uncover the anomaly in training data.

→ EDA is beneficial to uncover strange or anomalous values and fields for eg: In the 'Account length' field. We can see that the values fall in the range of 50 - 150. So, we can see two outliers way above this range. We will delete these outliers and we can detect outliers using histogram & scatter-plots as well.

MATRIKAS

Date: / /

(9) Objectives and the methods of binning numerical variable.

Binning - It is the process of transforming numerical variables into categorical counterparts. (A)

Methods of Binning numerical variable:

1) Unsupervised Binning

(A) without considering the target class label into account.

(a) Equal Width Binning -

Algorithm divides continuous variable into several categories having bins or range of the same width.

$$w = \left| \frac{max - min}{n} \right|$$

w → width

max, min → in the list

n → no. of categories

(b) Equal Frequency Binning

Algorith divides the data into various categories having approximately the same number of values. The values of data are distributed equally into formed categories.

$$freq = n/n$$

freq → frequency of category

n → no. of values in data

n → no. of categories

MATRIKAS

Date: / /

2) Supervised Binning:

It is a binning that transforms a numerical or continuous variable into categorical variable considering the target class label into account.

1) (a) Entropy-based Binning:

It calculates entropy for target class labels, & it categorizes the split based on maximum information gain.

Q.10 - on predictive value *

Objective:

Make analysis more convenient by categorizing an attribute's numerical values into a reduced set of classes.

(11). Explain method of binning based on predictive value.

i) Binning based on predictive value partitions the numerical predictor based on the effect each partition has on the value of the target variable by e.g. ordinal categorical values with variable with more values (eg. low/high/medium).

MATRIKAS

Date: / /

ii) Boundries are inserted between bins to maximize difference in the target variables.

(12) Correlated variables -

i) Two variables X and y are linearly correlated if an increase in x is associated with either an increase in 'y' or a decrease in 'y'.

ii) The correlation coefficient 'r' quantifies the strength and direction of the linear relationship between x and y.

Using correlated variables will cause the model to become unstable and deliver unreliable results.

(Q. 14.) Using EDA to investigate correlated predictor variables.

Strategy for Handling Correlated Predictor variables at the EDA stage:

1. Identify any variables that are perfectly correlated (i.e. $r = 1.0$ or $r = -1.0$). Do not retain both variables in the model, but rather omit one.

MATRIKAS

Q. 7, 10, 13

Date: / /

2. Identify groups of variables that are correlated with each other. Then later during the modeling phase, apply dimension-reduction methods, such as Principal Components Analysis (PCA) to these variables.

This strategy applies to uncovering correlation among the predictors alone.

Q. 13.