PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 Feet Ring Road, BSK III Stage, Bengaluru-560 085
Department of Computer Science and Engineering
Session : Aug-Dec 2021

UE20CS203: Statistics for Data Science

Assignment Objective: To analyze a given data set. Perform Exploratory Data Analysis.

What is Exploratory Data Analysis (EDA)?

- How to ensure you are ready to use machine learning algorithms in a project?
- How to choose the most suitable algorithms for your data set?
- How to define the feature variables that can potentially be used for machine learning?

Exploratory Data Analysis (EDA) helps to answer all these questions, ensuring the best outcomes for the project. It is an approach for summarizing, visualizing, and becoming intimately familiar with the important characteristics of a data set.

Exploratory Data Analysis (EDA) is the first step in your data analysis process. Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how best to manipulate your available data sources to get the answers you need.

You do this by taking a broad look at patterns, trends, outliers, unexpected results and so on in your existing data, using visual and quantitative methods to get a sense of the story this tells. You’re looking for clues that suggest your logical next steps, questions or areas of research.

Developed by John Tukey in the 1970s, exploratory analysis is often described as a philosophy, and there are no hard-and-fast rules for how you approach it. EDA is used to tackle specific tasks such as:

- Spotting mistakes and missing data;
- Mapping out the underlying structure of the data;
- Identifying the most important variables;
- Listing anomalies and outliers;
- Testing a hypotheses / checking assumptions related to a specific model;
- Establishing a parsimonious model (one that can be used to explain the data with minimal predictor variables);
- Estimating parameters and figuring out the associated confidence intervals or margins of error.

Value of Exploratory Data Analysis

Exploratory Data Analysis is valuable to data science projects since it allows to get closer to the certainty that the future results will be valid, correctly interpreted, and applicable to the desired business contexts. Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and data scientists but can be very informative about a particular business.

EDA is performed in order to define and refine the selection of feature variables that will be used for machine learning. Once data scientists become familiar with the data set, they often have to return to feature engineering step, since the initial features may turn out not to be serving their intended purpose. Once the EDA stage is complete, data scientists get a firm feature set they need for supervised and unsupervised machine learning.

Tools and Techniques

Among the most important statistical programming packages used to conduct exploratory data analysis are S-Plus , R and Python libraries for data analysis.
(**NumPy**,**SciPy**,**Matplotlib**,**Pandas**,**ScikitLearn**,**Statsmodels**,**Seaborn**,**Bokeh**,**Blaze**,**Scrapy**,**Requests**,**BeautifulSoup**)

Sample Exploratory Data Analysis Case Studies

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

http://ucanalytics.com/blogs/exploratory-data-analysis-retail-case-study-example-part-3/

Data Sets :

Look for data sets online like

https://www.kaggle.com/datasets

https://www.tableau.com/learn/articles/free-public-data-sets

Possible Evaluation Scheme

1. Data Set

A. How well it meets the criteria mentioned in the guidelines?

B. No of attributes (columns) and rows (tuples)

C. How much value do the attributes contribute?

D. Missing Values, NANs

E. Categorical and Numerical column presence

2. Data Cleaning

A. Identification of missing data.

B. Missing data is categorical/numeric and how well are missing values replaced.
(Min of methods mentioned in the guidelines (More weightage))

C. Identify some different methods for missing values.

D. If you choose to drop any attribute instead of replacing it, valid reason for it. (Since attribute dropping should have strong reasons and one cannot simply attribute as it may carry important info pertaining to individual tuples)

3. Visualization

A. Is the graph chosen to represent the particular attribute/aspect appropriate?

Eg: Scatter plot for correlation visualization, then the next best method is line graph (histogram cannot be used)

B. How easy is to visualize the graph and draw conclusions without explanation? (How well is the data represented on graphs?)

C. Name of x-axis, y-axis and title of the graph. (Very very important aspect since a graph without these has no value)

D. legend for the graph if required

E. Extension of axis if required and not sticking to default graph options of python.
Eg: if the values are up to 7500 and x axis has last marked number on number line as 7000, it is wrong. It should up to 8000 or else high chances of wrong interpretations.

4. Insights/Results:

A. Hypothesis testing and results

B. Normalization

C. Checking for Normality

D. Graph conclusions

E. Accuracies

F. Reason for such accuracies/results

5. Identify the information nugget (key information that is not obvious) specific to the data set

Submission guidelines:

1.Submit a pdf report of your analysis along with the .ipynb file of your code.

2.Record a video(max-4 mins) explaining your solution using report or ppt.

3.Novelty, relevance, and clarity in the analysis will be taken into consideration while evaluating.