Exploratory Data Analysis Life cycle-An Overview

Nidhi Singh[1,] Manikant Roy[2]
IBSAR Institute of Management Studies Karjat[1] Lovely Professional University, Punjab

Abstract: Every Software development project follows the certain model such water fall model, spiral model or popularly known as Software Development Life Cycle, which breaks down the complex software project into many small chunks which is easy to handle and implement. Similarly Exploratory Data Analysis also requires to be done in phase wise manner. This paper presents the complete overview of exploratory data analysis life cycle.

Keywords: Data Analysis, Data Science, Analytics, Statistical analyses

I. INTRODUCTION

In exploratory data analysis process the very first thing to be understood is data itself. Understanding the data helps in analyzing the outcome which is to be achieved. In traditional, relation database management system data is in tabular format where as in case of data science domain data can be structured like typical RDBMS or Unstructured data or Semi-structured. In relation table column is known as **attribute** and row is a record where as in exploratory data analysis every row is an **observation** and column is **feature such as Age, Salary etc.** Data can be classified into various category like Categorical Data, Numerical Data, Nominal Data and Ordinal Data. Before applying any technique on data one needs to apply various transformation on data.

II LIFE CYCLE

The exploratory data analysis starts with preliminary steps
1. Understanding of main characteristics data set
2. Finding relationship between two variables
3. Extracting main variables
The next section explains the steps in involved in entire exploratory data analysis

III DATA WRANGLING

Once the data is imported inside the tool for processing, data wrangling is first task to be performed. Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics[1].Data Wrangling includes following steps

- Preprocessing
- Dealing with Missing Values
- Data Formatting
- Data Normalization
- Binning
- Turning categorical into numerical

III EXPLORATORY DATA ANALYSIS

Exploratory data analysis process uncovers the some important fact that should be taken into consideration while making any data driven research e.g. what are important factor for customer while doing online shopping? Etc.
EDA involves following steps

- Descriptive Statistics
- Group by
- ANOVA
- Correlation
- Correlation – Statistics

IV MODEL DEVELOPMENT

One the data is wrangled and Descriptive analysis is done the next phase is to build a model which will take the data and will predict the target values. In model development regression is used. There can be linear regression or multiple regression depending upon the problem statement. A Model will help us understand the exact relationship between different variables and how these variables are used to predict the result. There are following steps in model development

- Simple and Multiple Linear Regression
- Model evaluation using visualization
- Measure for in-sample evaluation
- Prediction and Decision Making

V MODEL EVALUATION

After the development of model, it is very important to test the model. So that one can see the performance of the model. Any model can be evaluated by applying following methods

- Overfitting and Under fitting
- Cross Validation
- Ridge Regression
- Grid Search

V CONCLUSION

Exploratory Data Analysis is very important for data driven decision which can be very useful from business point of view. Any data analysis project helps in either increasing the profit or mitigating the coming risk. Hence EDA plays a very important role in any of the businesses.

VI REFRENCES

[1] Wikipedia contributors. (2018, April 3). Data wrangling. In Wikipedia, The Free Encyclopedia. Retrieved 07:36, June 9, 2018, from https://en.wikipedia.org/w/index.php?title=Data_wrangling&oldid=834062041

[2]https://courses.cognitiveclass.ai/courses/coursev1:CognitiveClass+DA0101EN+2017/courseware/f20c6998fe634d849f618a84bcf55a72/