Exploratory Data Analysis

Shashankgoud Patil[1], Nagaraja G.S [2]

Email: shashankgvp.cs16@rvce.edu.in,Email: nagarajags@rvce.edu.in

---

Abstract: Data is that the most essential resource in today's world. This knowledge in its raw kind, cannot be taken simply . This paper explains the various steps that needs to be carried out for data pre-processing before the data is ready to undergo the process of machine learning .It also gives us an idea about various stages the data has to at the go during the process of machine learning. The objective of this paper is to present the idea of the various transformation the data has to undergo before it is ready to be learned by the machine. Machines do not perceive the information in its raw kind, so the date should endure a countless pre-processing, thus in machine learning the information should beneath go a countless pre-processing before it goes to train an explicit machine learning model. The paper gives summery of the various steps that have to be under gone for better prediction by learned model.

The result of performing the underlying exploratory data analysis is that the gives any organization maximum value by helping scientists understand whether the findings they h ave generated are correctly interpreted and if they relate to the appropriate business contexts. . The steps performed in exploratory data analysis are significant to the data scientists to verify that the outcomes they produce are legitimate, accurately deciphered, and pertinent to the ideal business settings.

Keywords: Exploratory Data Analysis, Data pre-processing, Regularisation ,Dimensionality reduction , Pandas, Data visualization.

1.0 Introduction

The quality of data and amount of useful information the training data contains are the key factors To determine how well the machine learning algorithms can learn. Thus is absolutely critical that we examine and pre-processed data before we feed it to the algorithm. The proposed framework is that data needs to be preprocessed and visualized in order to determine the kind of machine learning model that can be used to train the data efficiently. From point of view of building models, by envisioning the data we can locate the hidden patterns, discover if there are any clusters within data and we can analyse if the data is linearly separable/too much overlapped etc. From this underlying analysis we can easily rule out the models that won't be suitable for such a data and we will actualize just the models that are appropriate, without burning through our significant time and the computational resources. The machine learning is inalienably an iterative process. Modeling can be bulky when you are performing the process again and again to guarantee your model is optimized and can generalize well. Extra the time you spend on model selection and model tuning; the process can easily become a disappointing one. Good exploratory data analysis combined with relevant data visualization is fundamental for pinpointing the right direction to take. It both shortens the machine learning process and provides more exactness to its result. Data visualization tools enable data

scientists to quickly recognize and focus on the most significant data and most significant ways to proceed. Even during the modeling process, model graphs can assist with accelerating the model-creation process by displaying the model maps conceptually. While evaluating the models, visualizing the results of hyperparameter tuning can help data scientists narrow down the groupings of hyperparameters that are most significant. In this paper we have discussed essential data pre-processing techniques that will help us build good machine learning models.

2.0 Literature Review

Before mainstream data visualization tools for machine learning were developed, the machine learning process was much more abstract. Verifiably, ggplot2 in R gave truly necessary visualization tools for exploratory data analysis. But today, with the suite of data visualizations that are accessible in Python, such as seaborn, scikit-learn, and matplotlib, exploratory data analysis that forms the initial part of the machine learning process should be possible substantially more effectively. At the same time, with TensorFlow, model-building and model-tuning processes become a lot more intuitive. As opposed to investing on the existing system of scrutinizing values, with the assistance of both 2-dimensional and interactive data visualizations, data scientists can give more consideration to the comprehensive view of the data at each level of the machine learning process concentrate more on the significance of the data, the model design, and the model performance.

We need data visualization in light of the fact that a visual outline of information makes it simpler to identify patterns and trends than looking through large number of rows on a spreadsheet. It's the way the human mind works. Since the purpose of data analysis is to gain insights, data is significantly more valuable when it is visualized. Even if a data analyst can pull insights from data without visualization, it will be increasingly hard to impart the meaning without visualization. Charts and graphs make communicating data findings easier even if you can distinguish the patterns without them. The below figure indicates the various steps the data has to undergo during the process of machine learning

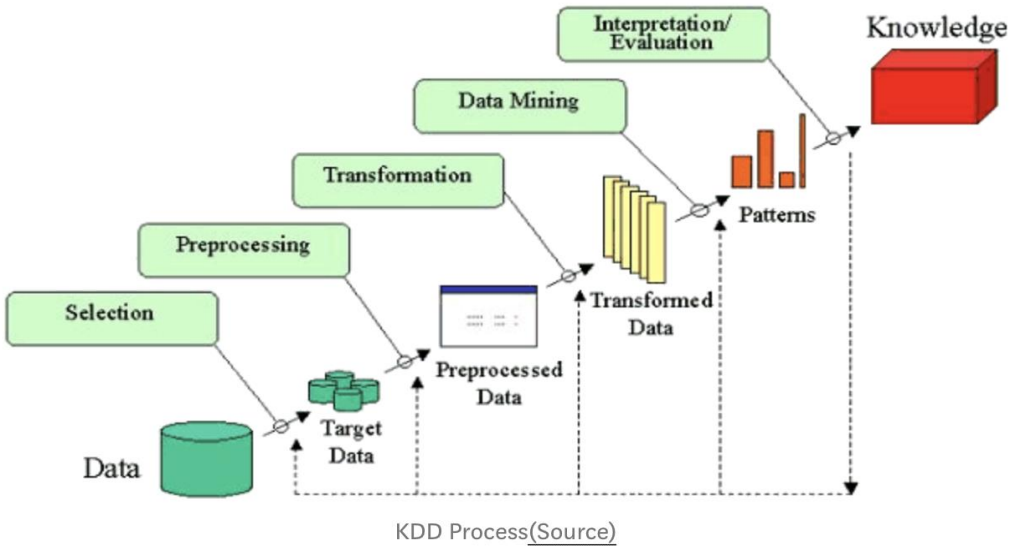

Fig. 1. Data Life Cycle

EDA IN PYTHON

Utilizing python to do exploratory data analysis is easy to learn. Its data handling capacity is much higher and it is an open source language. The developer can understand the code .it offers a assortment of libraries and some of them uses great visualization tool. Visualization process it simpler to make the reasonable report.

Pandas

It provides powerful package for data analysis. We can clean, investigate and change the information the data. Data can be put away in CSV format in computer. Cleaning, visualizing and storing the data can be easily done. It is built on the top of the NumPy package. Plotting functions from Matplotlib and machine learning algorithm in Scikit-learn.

Jupiter Notebook

It gives ability to execute the code in a specific cell. It gives the console based methodology for processing. It provides web based application process. It includes input and output of the calculation. It gives rich media portrayal of the object.

3.0 Proposed Approach

The steps that have to be carried how to perform data pre-processing are as follows :-

- **Importing libraries and Loading data into data frames:** The below snippet shows how the libraries must be imported and how the data from the CSV file should be loaded into the data frame. Loading of data into the data frame is the fundamental steps of exploratory data analysis.

```
In [3]: #importing Libraries
        import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd
        import seaborn as snsdataset #visualisation

        #Loading the Data
        dataset = pd.read_csv('input_cm1996.csv')
        dataset.head()
```

Out[3]:

	CARRIER_METHOD_ID	ZONE_ID	WEIGHT	SHIP_CHARGE_COMMERCIAL	SHIP_CHARGE_RESIDENTIAL
0	1996	2	0.25	4.25	5.25
1	1996	2	0.50	4.25	5.25
2	1996	2	0.75	4.25	5.25
3	1996	2	1.00	4.25	5.25
4	1996	2	1.25	4.25	5.25

Fig. 2. Importing Libraries and loading data

- **Analysing Required Data :** We check the data types obtained to see whether there is any kind of type conversion required to change the data type of the data for better visualization. Not all the columns in data frame are required perform the process of learning and thus few of the columns can be eliminated.

- **Dealing with Missing Data :-**Missing data may cause faultlessness in the learning process. Does in order to avoid this problem there are ways of dealing with the missing data are :-
    1. **Identify missing values :** The very first step will be to identify the missing values .For a large data frame we use isnull function to return a data frame with Boolean values specifying whether a hey cell contains a numeric value or if the data is missing .
    2. **Eliminate the examples with missing data :-**One of the easiest way to deal with missing data is dropping the corresponding feature or the training example in which the data is missing. dropna function can be used for eliminating the examples with missing data.
    3. **Imputing missing data :-**Eliminate the examples with missing data is not feasible we can try to approximate the value of the missing data. SimpleImputer function provide strategies to replace the examples of machine data.

```
from sklearn.impute import SimpleImputer
import numpy as np
imr = SimpleImputer(missing_values=np.nan, strategy='mean') >>> imr = imr.fit(df.values)
imputed_data = imr.transform(df.values)
imputed_data
```

Fig. 3. Handling Missing Data

- **Dealing with Categorical Variables :** Categorical data is unavoidable in real world data set. Categorical data can be distinguished into ordinal and nominal features .The ordinal features can be sorted or ordered .Nominal features do not imply any order**.**
    1. **Mapping Ordinal Features :**In this we convert the categorical string values to integers and perform manual mapping in order to derive the correct order of labels offer size feature.
    2. **Encoding class labels :**Here class labels are encoded as integers. We use a similar approach to that of mapping ordinal features As we know that class labels are not ordering doesn't matter what integer value assign to them until we are performing mapping in the later stage.
    3. **One hot encoding :**Here we use the label encoder to encode string labels to integer , the idea here is to create a dummy feature for each unique value in a nominal feature column.

- **Detecting Outliers :**Different from the normal data ,these point may be too high or too low .it is usually recommended to identify this outliers add eliminate them.

- **Graphical EDA:** To see how it looks graphically using various floating matrix such as histogram, stem plots ,box plots ,Scatter plots heat Maps and 3D surface plots etc**.** EDA, which analyses the data sets to help sum up their data Statistical features which focus on the same four key Aspects such as central tendency measurements, measures of spread, shape of the distribution and existence of outliers.

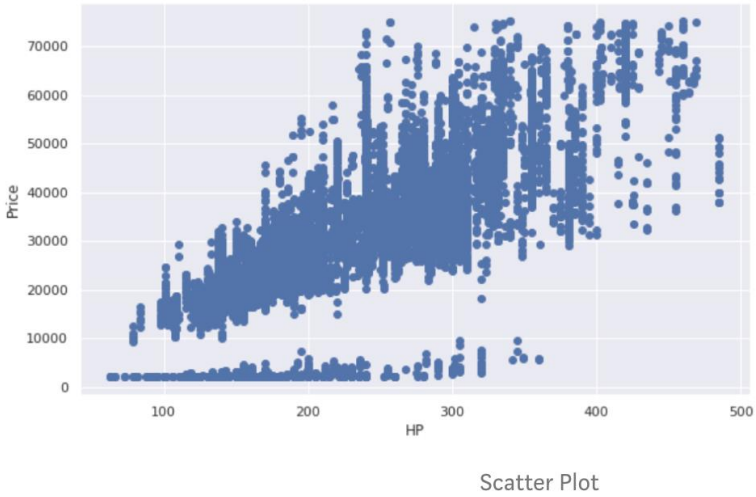

Fig. 4. Example of Graphical EDA

- **Selecting Meaningful Features:** Due to the complexity of the data sometimes over fitting takes place which means that the model fixed parameters to closely which regard to particular training data set ,but does not generalise for new data, such models have high variance. Common solutions to reduce generalization error are :
    1. Collect my training data .
    2. Introducing a penalty for complexity via regularisation
    3. Choosing a simpler model with your parameters
    4. Reducing the dimensionality off the data

L1 and L2 Regularisation:L2 Regularisation is one of the approaches to reduce the complexity of the model my penalising large individual weights.L2 norm of weight vector is defined as:

$$L2: \quad \|\boldsymbol{w}\|_2^2 = \sum_{j=1}^{m} w_j^2$$

Weather approach to reduce model complexity is L1 regularisation

$$L1: \quad \|\boldsymbol{w}\|_1 = \sum_{j=1}^{m} |w_j|$$

L2 Regularisation and L1 regularisation yield sparse feature vectors and most feature vectors will be zero. Sparsity can be useful if we have a high dimensional data set it many features that are irrelevant ,especially useful in the cases where we have more relevant features than the training examples.

- **Sequential Feature Selection Algorithm**: One of the ways to reduce the complexity of the model and avoid over fitting is dimensionality reduction via feature selection, which is useful

for unregularized models. There are 2 categories in dimensionality reduction via features selections that the feature selection and feature extraction. In feature selection select subset of original features whereas in feature extraction ,we derive information from the feature set to construct a new features subspace.

Steps of Machine Learning:

Data analytics and visualization are involved in every step of machine learning and a process has to be followed before making the data available for the purpose of training the data .

1. **Data Exploration**:-The first objective is to visualize the data, which gives us a better representation and a proper idea of what kind of data we will be using, The very next step is to make sure that the data is not incomplete that is there are no missing values in the data. Next we can visually try to find the correlation of the data in order to see which model will be appropriate to train the data.

2. **Data Cleaning:** -The second objective is to actually check the data for potential issue which may lead to problems in training the model and ensure that the issues are fixed.

3. **Model Building**: -The third objective is to determine the most suitable model and train the data. the we are supposed to visualize the results, perform model diagnostics, residual diagnostics, ROC curves etc. Model can be supervised or unsupervised model. We can use classification, regression model to get the output.

4. **Present Results**: -Representation of the data is very important to give the impact of the model. the representation can be done using charts, graphs, tables etc. It can identify the area which needs improvement. It can clarify the factor very well

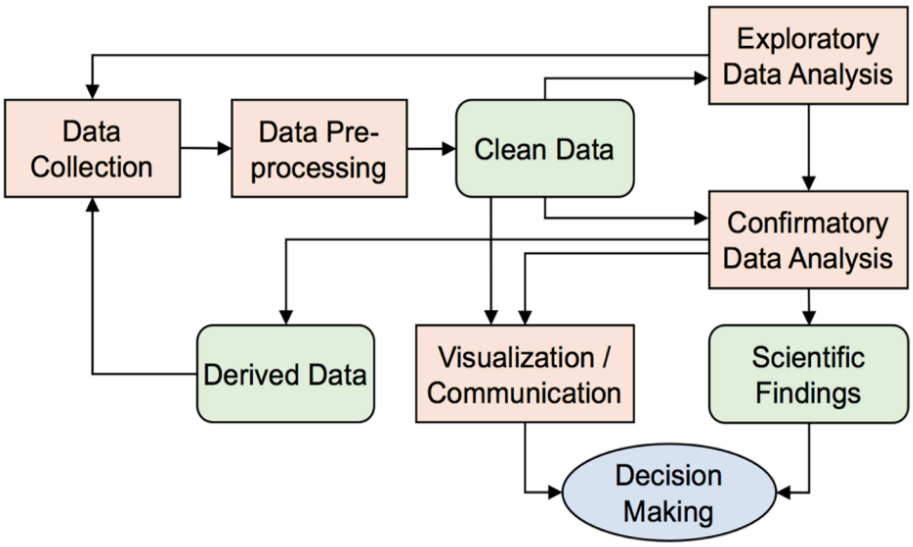

Fig. 5. Data Flow chart

4.0 Results and Discussion

The quality of data and amount of useful information it contains are the key factors To determine how well the machine learning algorithms can learn. Thus is absolutely critical that we examine and pre-processed data before we feed it to the algorithm using the described data pre-processing techniques. In an ideal world, data analysts have access to all their required data without concern for where it's stored or how it's processed analytics just work. The outcomes are ultimately throw-away, and all that you should be left with is a greater understanding and intuition for the data and a long list of hypotheses to explore when modelling. Exploratory data analysis gives any organization maximum value by helping scientists understand whether the findings they h ave generated are correctly interpreted and if they relate to the appropriate business contexts. In this paper we have looked at useful techniques to make sure that we handle missing data correctly. Before we feed data to a machine learning algorithm, we also have to make sure that we encode categorical variables correctly, and we also saw how we can map ordinal and nominal feature values to integer representations. Moreover, we briefly discussed L1 regularization, which can help us to avoid overfitting by reducing the complexity of a model. As an alternative approach to removing irrelevant features, we used a sequential feature selection algorithm to select meaningful features from a dataset.

Limitations of data Analysis, the traditional Exploratory data analysis process has three serious and related downsides:

1. **Complexity:** Data complexity can be determined only through visualization which is not a fair means. This means the data engineering team develops highly specialized, sometimes non-transferrable skills for managing its dataset.
2. **Brittleness**: we have to choose the right method of data visualization to understand what type of variable we are dealing with. For that we can use various software by importing them.
3. **Inaccessibility:** More importantly, EDA is all but inaccessible to smaller organizations without dedicated data engineers. On-premise EDA imposes further infrastructure costs. Smaller organizations may be forced to sample data or conduct manual, ad hoc reporting.

4.0 Conclusion and Future Work

The main take away from this is that always start exploratory data analysis with a open mind to discovery. EDA allows us to understand the dataset better exploration of the dataset and understanding the features and issues of the dataset. It also provides the basis of research hypothesis. The tools provided in this paper will allow the researcher to gain a better understanding of the dataset and also to generate novel hypothesis. EDA is a significant advance to take before plunging into machine learning or measurable demonstrating on the grounds that it gives the setting expected to build up a fitting model for the current issue and to effectively decipher its outcomes. EDA is significant to the information researcher to verify that the outcomes they produce are legitimate, accurately deciphered, and pertinent to the ideal business settings. We further intend to get a comprehensive view of the data at each level of the machine learning process concentrate more on the significance of the data, the model design, and the model performance

References

1. Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253
2. Matthew Ntow-Gyamfi and Sarah Serwaa Boateng, "Credit Risk and Loan Default among Ghanaian Banks: An Exploratory Study," Management Science Letters, Vol. 3, 2013, pp.753–762.
3. Exploratory data analysis – From Wikipedia, the free encyclopedia [Online], Available: https://en.wikipedia.org/wiki/Exploratory_data_ analysis
4. John T. Behrens, "Principles and Procedures of Exploratory Data Analysis," Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160.
5. X. Francis Jency, V. P. Sumathi, Janani Shiva Sri, "An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients," International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-4S, November 2018, pp.176-179.
6. Introduction to Machine Learning using Python [Online], Available: https://www.geeksforgeeks.org/introduction-machine-learning-using-python/
7. Pwint Phyu Khine, Zhao Shun Wang, 'Data Lake: A New Ideology in Big Data Era', 2017 4th International Conference on Wireless Communication and Sensor Network [WCSN2017], At Wuhan, China
8. Benjamin S. Baumer, 'A Grammar for Reproducible and Painless Extract-Transform-Load Operations on Medium Data', arXiv:1708.07073v3 [stat.CO] 23 May 2018
9. . Bogumil M. Konopka, Felicja Lwow, Magdalena Owczarz, Łukasz Łaczmański, "Exploratory Data Analysis of a Clinical Study Group: Development of a Procedure for Exploring Multidimensional Data," PLOS ONE, [Online] https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC6107146/pdf/pone.0201950.pdf, August23, 2018, pp. 1-21.
10. Brillinger DR. 2010. Exploratory data analysis. In International Encyclopedia of Political Science. Sage: New York; 530–537.
11. Yannan Zhang, "Visualization Research of environment monitoring spatial and temporal data based on R language", *Dissertation of wuhan university*, 2016.