24-SEP-2022

EDA and Feature Engineering

Data Scienve lifecycle

① Data Integration → Ensuring Projects dimension
② EDA (Analysis of Data) → Graphical or Data Analysis
③ Pre-Processing
④ Model/Algorithm Building → Different Algorithms
⑤ Evalnate & Validation of the model

→ Use statistics to analyse the data
→ core ML pipeline

Statistics

→ collect, organize, interpretation, and analysis of data.

we can find some insight by performing these actions.

e.g. Sales of Product → sales is going down.
what is the reason behind they down?

→ Is our product Unique?
Not paying attention to the customer
Leadership is not good.
Marketing strategy is not good.
Not looking to the Competitor.

Using all these we created one dataset
↓
need to analyse to find out the reason.
↓
conclusion.

① Project Manager
② Business Analysts
③ Data Scientist
↓
Domain Expert

} Look into the dataset and find conclusion

Note: Any domain requires EDA and feature engineering.

Data Integration

① from big data tools like HADOOP, MOSSL, Spark etc
② remote location like SQL, MongoDB etc
③ Some file format like CSV, XML, TSV etc

Types of data

Tendency of Data

- Batch data → Historical (Periodic) data
  - mini batch data (little more frequent)
- Streaming data → continuous data

① structured data ⇒ Table (RXC) ⇒ ML
② Unstructured data ⇒ videos, images, voice, text ⇒ DL
③ Semi structure data ⇒ XML, JSON

EDA + FE ⇒ Irrespective of type of data, we can perform.

Structured data

e.g.

feature1 weight	feature2 Height	feature3 BMI
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21
continuous	continuous	continuous

Two types of Structured Data

1. Numerical
   - continuous → continuous in nature. eg: Height
   - discrete → whole number eg: student in the class
2. Category
   - nominal → Order doesn't matters eg: Gender (M or F)
   - ordinal → Order matters eg: Degree (10th, 12th, Graduation, Master, PhD)

Implementation of the Data

Student Performance

Name	Age	Height	Sex	Weight	Education
Sunny	25	170	Male	70	GRDUH
Arjit	30	180	Male	80	PH
Priyam	35	160	Male	60	UH
Priya	20	150	Female	55	PhD
Aditi	27	145	Female	58	PH
↓	↓	↓	↓	↓	↓
Categorical	Numerical	Numerical	Categorical	Numerical	Categorical
Nominal	~~Discrete~~ Continuous	Continuous	Nominal	Continuous	Ordinal

UH → 0
PH → 1
PhD → 2

① First-level

① categorize the feature and check type of data

② 2nd-level

Univariate → One column → check Height column only.
Bivariate → Two column → check Height w.r.t Age column
Multivariate → More than two column → check sex w.r.t Height and Age column.

Independent and Dependent Variable

Age, Height & Sex → Based on it defined weight

Age, Height & Sex: Independent feature

weight: Dependent feature

→ Independent means independent analysis of the feature.
→ Dependent means one feature is dependent on one or more feature.

Pre-Processing Feature Engineering

① Missing values
② outliers detection
③ scaling

} changes in the column → Transformation next features

→ Engineering w.r.t feature.

Que first EDA is required or F.E or P.P?

Ans EDA → Pre-Processing → Import model

EDA and F.E

Real life example -
e.g Party

cook Biryani (Model)

① chicken
② Rice
③ Onion
④ Oil
⑤ Spices

} Raw Data to cook chicken

1 kg
1 kg
1 kg
1 kg
500gm

} → Directly we can't cook Biryani. We have to clean rice, cut onion etc as a preprocessing steps.

↓

cook Biryani (Model)

↓

taste Biryani (evaluation and validation)

EDA → Analysis of data

Preprocessing or F.E ⇒ cleaning the data
Preprocessing and feature Engineering are same.

⇒ 1st we will perform EDA, then F.E after that we can again perform EDA.

e.g

Name	Age	Education	Salary	Experience
Sunny	25	UG	25K	2
Deepak	30	PG	30K	3
Rushi	40	UG	40K	5
Aman	50	PhD	50K	10
Shalini	20	UG	35K	1

EDA - Analysis of Data

① Create profile of the data
② Statistical based analysis
③ Graph based analysis

Profile of the Data

1. Number of Rows
2. Number of columns
3. How many missing values
4. How many numerical values
5. How many categorical values
6. What is the type of the data
7. Duplicate value
8. How much RAM is consumed

Statistic based interpretation → univariate, bivariate, multivariate

1. Variance
2. Covariance
3. Standard deviation
4. Correlation
5. Chi square
6. t-test
7. z-test
8. Anova test
9. Mean/Median/Mode
10. Skewness
11. Kurtosis

Graph based analysis

1. Box Plot → can check outlier, distribution.
2. Scatter Plot → can check linearity, outlier
3. Pair Plot
4. Histogram → distribution of the data
5. KDE plot
6. P-P plot
7. Heatmap → correlation b/w variables
8. Countbar → count Row, column for each variable

Que Based on a EDA can we do a processing of the Data?

Ans Yes we can do. Preprocessing Steps for EDA and

1. Missing value can be handle
2. outlier can be handle
3. Scaling can be done
4. Transformation is possible

⑤ Encoding is possible
⑥ can handle imbalance data
⑦ feature selection is possible
⑧ Dimension Reduction (PCA, tSNE)

Missing value detection → Missing value handle
EDA Pre-Processing

outlier detection → outlier handling
EDA Pre-Processing

cat (man, woman) → encoding
EDA Pre-Processing

skewed range → Scale (within certain range)
EDA Pre-Processing.

. imp feature → handle imbalance data
EDA Pre-Processing

→ Feature selection

→ Dimension Reduction (PCA, tSNE)
Pre-Processing.

① no
② no → imbalance dataset
③ no
④ no
⑤ no

Encoding

$y = mx + c$

male, female

$y = 0 \times male + c$

most making sense
used to encode male and female to 0,1 so that machine can understand.

Note: In EDA, we do analysis, and in Pre-processing or Feature Engineering we do change the value.

Summary

EDA

1. Profile
2. Statistical Analysis
3. Graphical Analysis

Preprocessing (directly affect model)

1. Missing values
2. Outlier
3. Scaling
4. Transform
5. Encode
6. Imbalance
7. Drop / Duplicate
8. Feature Selection
9. Dimension Reduction (PCA, LDA, tSNE)
10. Split / merge / drop / Add

Automated tool in Python

1. Pandas Profiling
2. mito
3. Knime

25-SEP-2022

[Note: Before giving data to the model, we do EDA and Pre Processing.]

Pre-Processing and Feature Engineering ways

① Missing value Handle

(a) Random
(b) forward filling / backward filling
(c) statistical approach
  - mean
  - median
  - mode
(d) end of the distribution
(e) drop
(f) KNN-Imputer or diff imputation technique
(g) ML algorithm which supports missing values
(h) We can create our own ML mode and can predict missing values

② Outliers Handle

(i) Detect outliers

(a) Z-Score
(b) IQR
(c) Box Plot
(d) Scatter Plot
(e) Violin Plot

(ii) Handling of outliers

(a) Drop / Fill
(b) Replace with median
(c) Replace / ~~with~~ trimming

[Note: Does outliers effect mean?
→ Yes it effects the mean.]

③ Transformation

(a) Box-cox
(b) Power Transformation
(c) Log-Transformation
(d) Square
(e) Cube
(f) Yeo Johnson

④ Scaling

ⓐ Standardization
ⓑ Min-Max
ⓒ Unit Scaling

⑤ Encoding

ⓐ One hot
ⓑ Label encoding
ⓒ Binary encoding
ⓓ Target guided encoding
ⓔ Hash encoding

⑥ Imbalanced dataset

ⓐ Collect more data
ⓑ Under sampling
ⓒ Over sampling
ⓓ ~~over~~ & cluster based

Note: We can change preprocessing technique to improve accuracy.