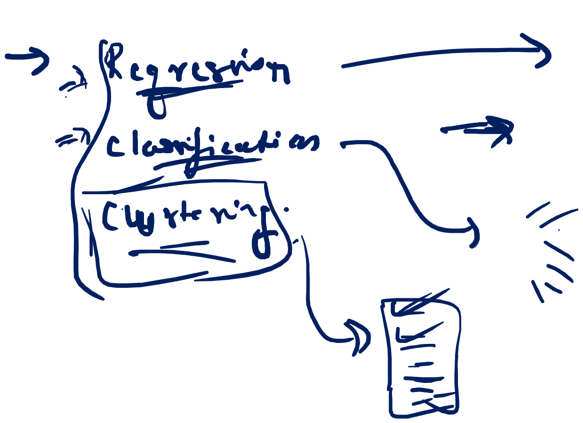
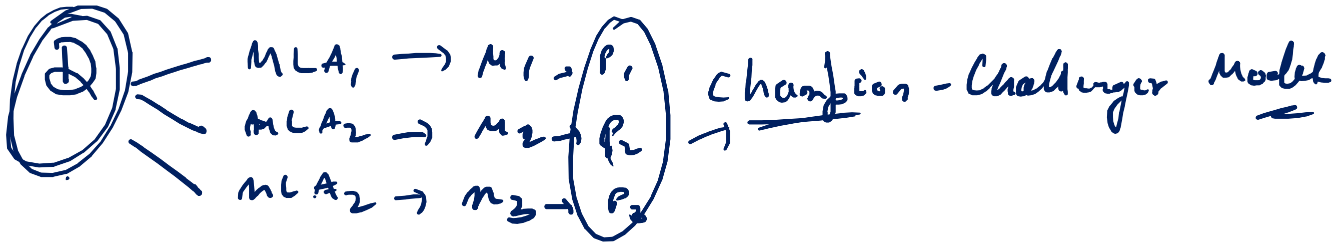
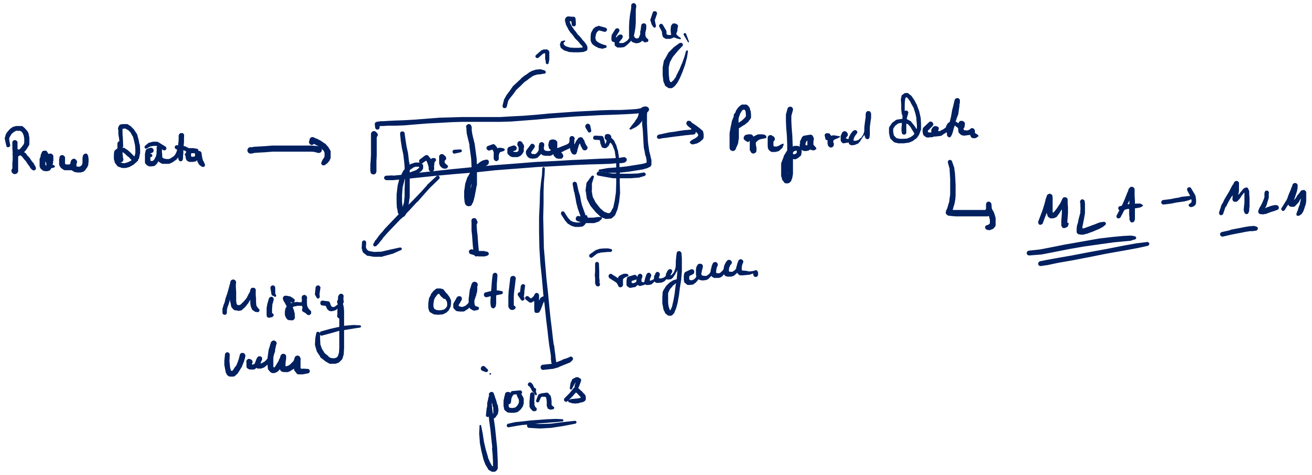

Regression
classification
Clustering

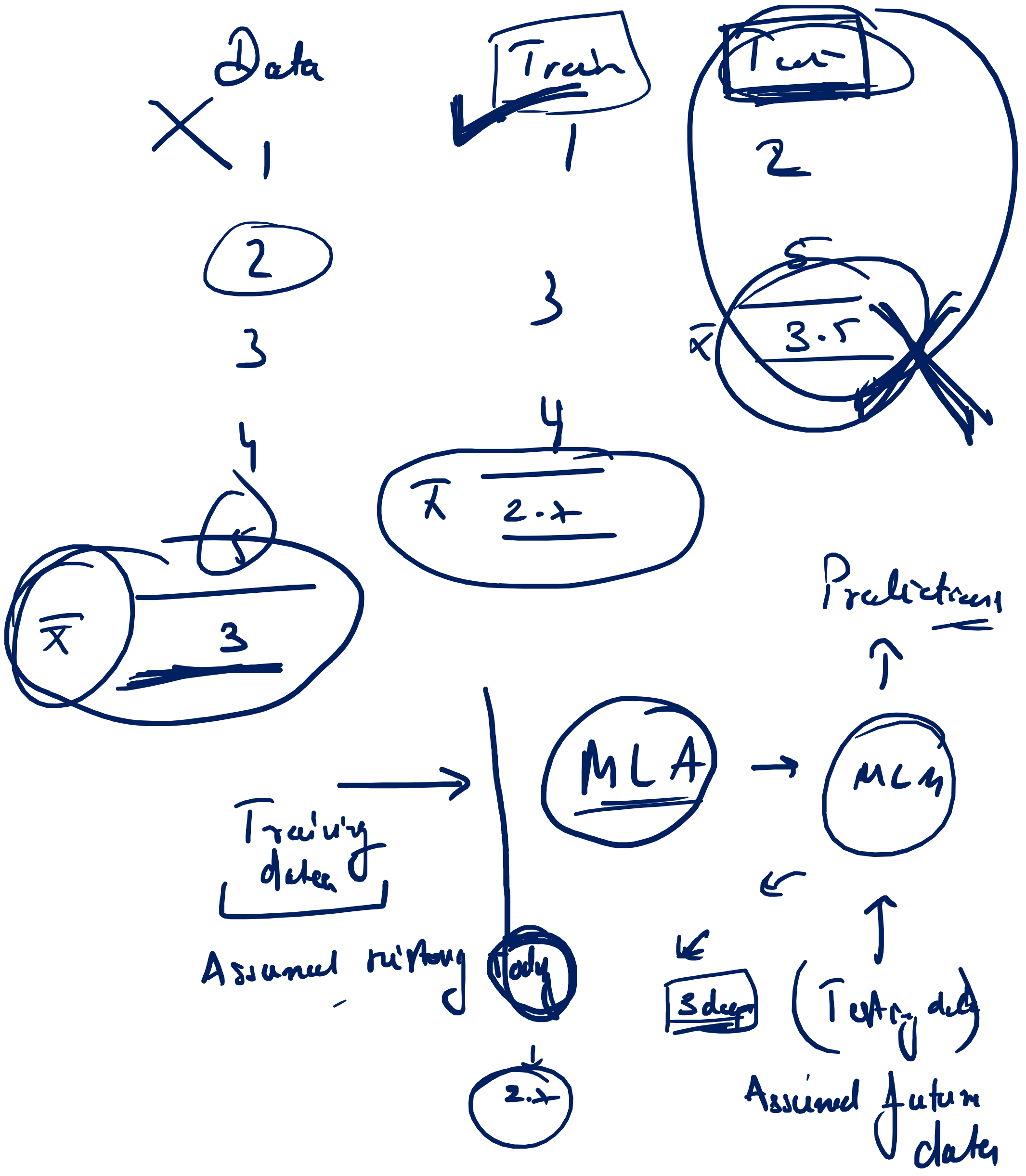

D
MLA1 → M1 → P1
MLA2 → M2 → P2
MLA2 → M3 → P3
Champion - Challenger Model


Raw Data
pre-processing
Prepared Data
MLA → MLM
Scaling
Missing value
outliers
Transform
joins

08:29PM

Data preprocessing

1. Import libraries
2. Load the data into python Env.
3. Making sure the dtypes are correct.
4. Extract target & features
5. Split the data into Train & Test
6. - Data Cleaning



→ Data Cleaning

1000 <

(1) Missing value Treatment

↳ excluded col-wise

→ If missing value ≤ 5% → Impute/replacing the value

Mean → continuous data | median ⇒ discrete data | mode = Categorical data

→ If missing values ≥ 10% → Drop the col | random Imputations (IQR)

↳ Business logic to replace

6%

(Q1.) If missing value b/w 5%– 10%

then it is upto the data scientist

to choose from any above case

If colA has val = (1)

↳ colB

② Outlier Treatment

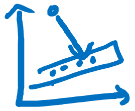
→ Outlier is good → Nothing to be done fraud detection

→ Outlier is bad → general pattern

→ If outliers ≥ 10% → split the dataset rowwise & create 2 models

with outl

w/o otlr



→ If outliers ≤ 5% → Impute/replace the with either LL or UL

Median method

$LL = Q_1 - 1.5(IQR)$

$UL = Q_3 + 1.5(IQR)$

Mean method

$LL = \mu - 3\sigma$

$UL = \mu + 3\sigma$

→ If outlier is b/w 5%-10%, then upto the data scientist to decide

7 → Analyze the statistical summary

8 → Feature Selection (Corr Analysis)

9 → OPTIONAL STEP → Data Scaling

↳ To bring data points in close Proximity

$(\mu, \sigma) \rightarrow$ fit()

a) Standard Scalar → Z-Score ⇒ $\boxed{\frac{x-\mu}{\sigma}}$ → Convert your data in a Standardized → N.D. $(\underset{\mu}{0}, \underset{\sigma}{1})$

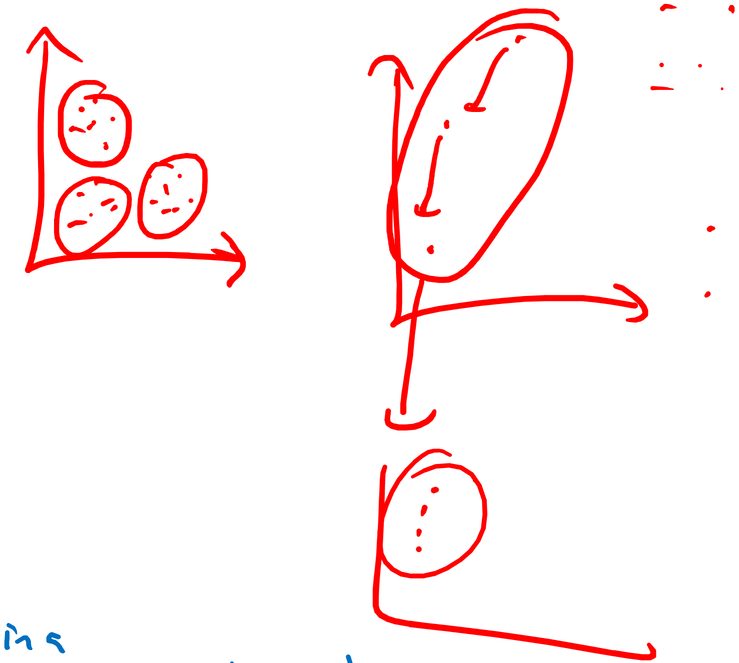
b) Min-Max Scalar → Normalization

↳ into a range of [0, 1]

$(x_{min}, x_{max}) \rightarrow$ fit()

$$\frac{x - x_{min}}{x_{max} - x_{min}} \Rightarrow [0, 1]$$

↳ Apply is done thru transform()



→ Data Encoding → Convert categorical cols to Numerical values such that the learning is not impacted

Label Encoding → only ~~for target~~

↳ Label Encoder() → Alphabetic order

→ Map() → Custom order

Categorical Variable

Ordinal

↓

has an in-built Rank

⇓

Numerical

↳ just use the Rank

Nominal

↓

No in-built Rank

Dummy Variables

↓

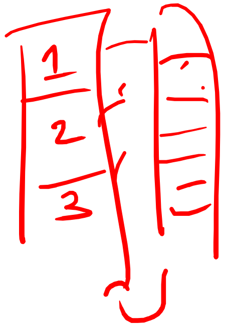
One-hot Encoding

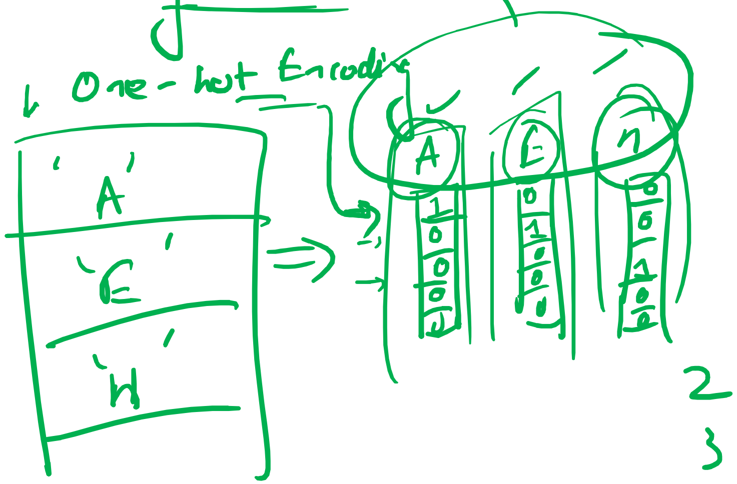
1 { High ~~Rating~~ 3

2 medium Rating 2

3 low ~~Rating~~ 1

Label Encoding





2

3

No

Revision

Gender
M
F
F
F
M
M

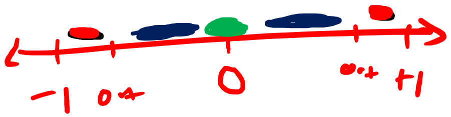
Male	Female
1	0
0	1
0	1
0	1
1	0
1	0

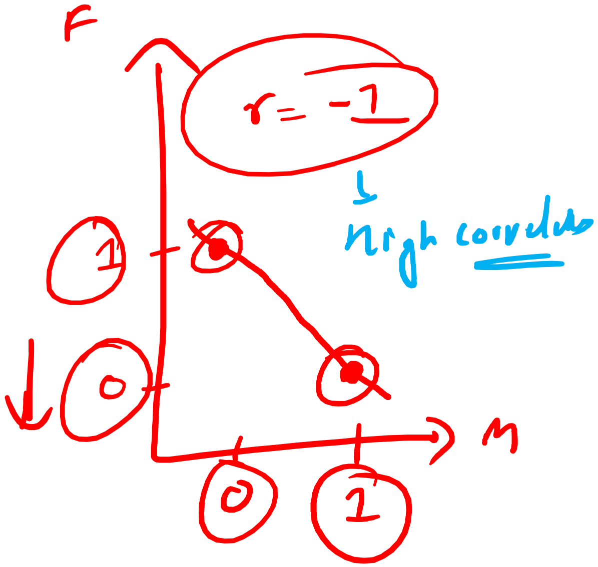
A

B

C

A	B	C
1	0	
0	1	
0	0	





$r \rightarrow [-1, +1]$

High Corr ⇒

≤ -0.7, -1

≥ 0.7, +1

No corr | $r \rightarrow \approx 0$ (-0.1, +0.1)

weak corr | $r \rightarrow$ (-0.7, -0.1) (0.1, 0.7)