DAY-6

Feature Transformation:

Data → Numerical
→ Categorical data

Nominal

→ No relation or No order between Variables

eg: parle-g, monaco, krack-jack.

Ordinal

~~No~~ There is a order or relation b/w the variables.

eg: good, better, best
eg:

Ordinal Encoding

→ ~~Nominal~~ ordinal Data

for output column
Target column
eg: Yes or No,
Churn prediction,
we use label encoder

eg: Education

High school
UG
PG
PG
UG
HS
HS

Ordinal data.
PG > UG > HS

HS => 0
UG => 1
PG => 2

Tranformed based on order.

Label Encoding → output column

Ordinal Encoding → input column

Onhot encoding :

→ why? ↳ Nominal chategorical

↳ it doesnot have any orde

color

Yellow
Green } No order
Red

if we encode with ordinal or Label encode. ML model will prioritize more on having greater numbers.

No matter how much unique categories you have in a column.

→ Dummy Variable Trap.

we remove one column (generally first column)

	Y	R	G
yellow	1	0	0
red	0	1	0
green	0	0	1

remove

⇒ Multicollinearity:

In ML we should not have any relⁿ b/w Independent columns.

by doing OHE :

	Y	R	G	
yell	1	0	0	Σ=1
reel	0	1	0	Σ=1
Green	0	0	1	Σ=1

Summation = 1

Σ=1 of each row

this is not correct for linear models
- → linear regression
- → logistic regression.

then How would I be able to represent Yellow column.

~~Y~~	R	G
~~1~~	0	0
~~0~~	1	0
~~0~~	0	1

→ if both are zero then its yellow, problem solved.

OHE using most frequent variables

Brand nominal → 40 diff brand cars

↓

OHE

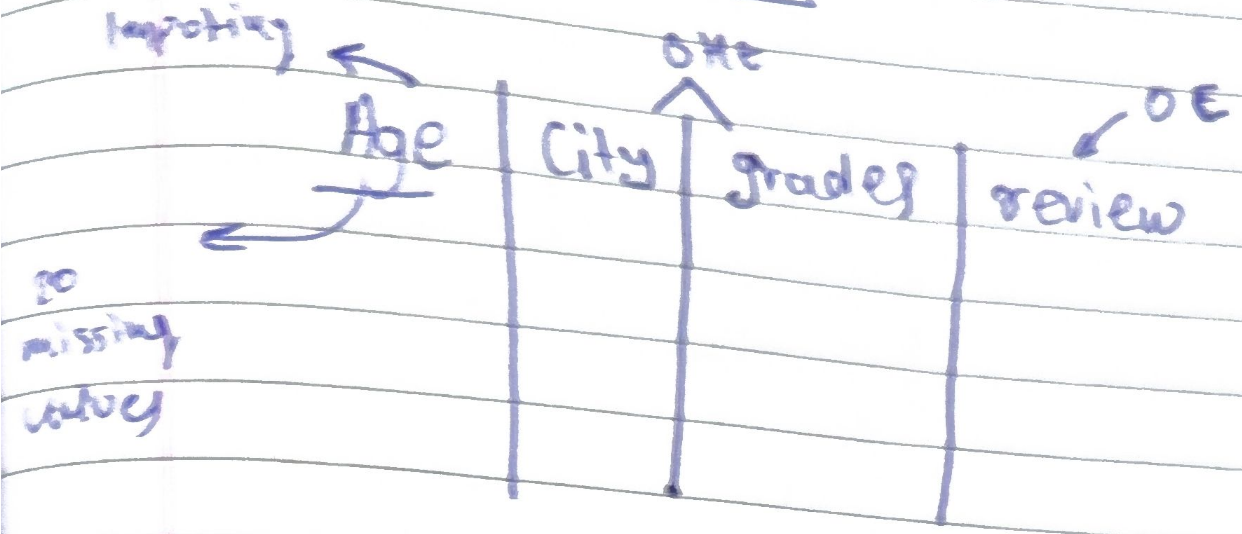
↓

Increase dimensions

↓

creat most freq categories to "others"

Column Transformer:



Now, here we have to transform all the features with some different techniques

We will be getting many numpy arrays and it may cause trouble while making Machine learning pipelines.

So, we use column transformer