Feature Engineering

Types of Encoding

1 One hat Encoding

→ Assigning the values 1 or 0.

If for example

germany	France	~~Spain~~
1	0	0
0	1	0
0	0	1

↓

3 − 1 = 2

Columns coz we Removed Spain

↓ Dummy variable Trap.

↓

Performed with pandas and Sklearn

↓

pd.get_dummies

→ Skipping the Spain Column

because we can derieve it by last Raw which is 0,0 so it is understood that last raw will be 1 so it is Spain.

Disadvantages of one hat Encoding :-

Pincode Example :-

560011
560012 →
560013
|
|
and So on.

It by Previous Example i convert them into Something call dummy then 99 columns will get created.

→ If we perform one hat Encoding here. Then more columns will get Increased and it may cause curse of dimension. ←

↓

Leading us tie this.

→ Many numbers of categories So Dont need tie Apply one hat Encoding.

2

Label Encoding →

ORdinal categary

Education	
BE	2
masters	3
PhD	4
Statistician	1

3

One hat Encoding with multiple categaries ↓

naminal categary

1 Target Guided ordinal categaries.

classification prablem

f1	O/P	
A	1	→ mean → 0.73
B	1	0.6
C	0	0.4
D	1	
A	0	
B	0	

→ where the Value of A is 0 and 1 considering those features only.

→ when you consider the mean. you are finding the number of values which is for A=1.

→ Rearranging the Values by there Rank due to ordinal category.

due to ordinal.

LABel

4
3
2
1
4
3
2
1

Mean Encoding → Nominal

f1		O/P	
A	⟷	1	
B		0	0.73
C		1	0.6
D		1	
A		1	0.5
B		0	0.4
C		1	
D		1	

→ we will convert this into mean values.

f1	O/P
56011	1
56022	0
	1
	1
	1

↓

Finding out the mean and this values 56011 will be Replaced by the Values of mean.

Why Feature Scaling

Features
- magnitudes
- units

cm Height	Kg. weight	BMI
180	78	
170	84	

→ Not perform feature Scaling.

1. Decision Tree
2. Random Forest
3. X-G BOOST.

Handle missing Values in Categarical Variables

1 delete the Rows

2 Replace with the must Frequent Values.

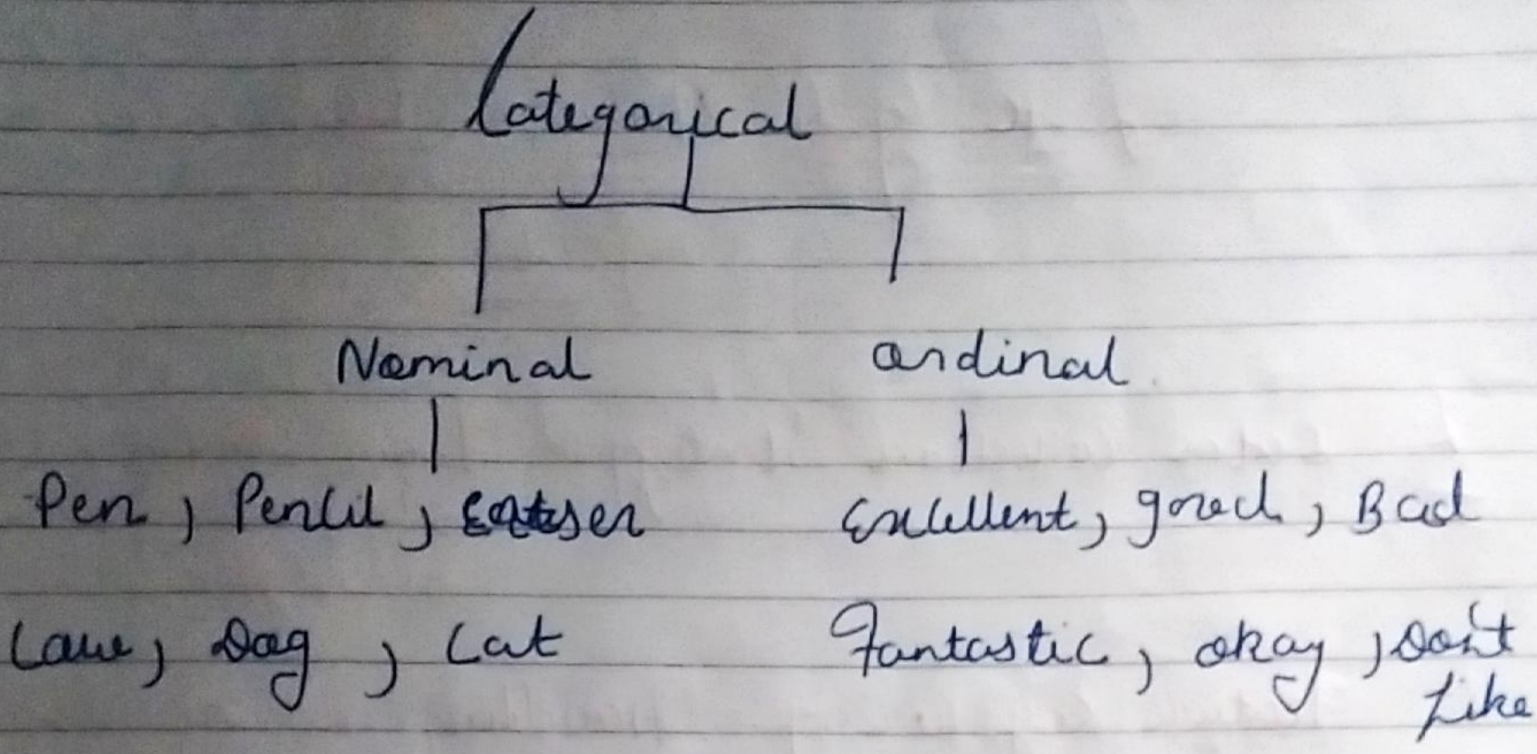
3 Apply Classifier Algarithm to predict.

4 Apply unsupervied ML.

Ordinal numbering Encoding or Label Encoding

Ordinal categarical Variables :-

→ ordinal data is categarical, Statistical data type where the variables have natural, ordered categaries and the distances between the categaries is not known.


Categorical
Nominal
ordinal
Pen, Pencil, Eraser
Cow, Dog, Cat
Excellent, good, Bad
Fantastic, okay, don't like.

Life cycle of a Data Science Projects

1. Data Collection Strategy :- From company side, 3rd party API's, Surveys

2. Feature Engineering :- Handling missing Values.

why are there missing Values? Survey - depression Survey.

1. They hesitate to put down the info.
2. Survey information are not that Valid.
3. men -- salary
4. Women --- age
5. people may have died -- NAN.

3. Data that will be missing

1) Continuous Data
2) Categorical Data.

What are the different types of missing Data

1 missing Compltely at random (MCAR)

2 missing at randam (MAR)

3 Nat missing at Randam (NMAR)

For getting The null Values in a Particular Rowe or Column.

df [df ['Embarked'].isnull()]

1 MCAR :- It means that there is no Relationship between the two particular thing in the Dataset ↓

No Relationship

2 NMAR :- There is absalutely a Relationship between the data missing and any ather Values. ↓

Having Relationship

3 MAR :- men :- hide Their Salary
women :- hide their Age.

Random Sample Imputation :-

→ It consists of taking Random observation from the dataset and we use this observation to Replace the nan Values.

When Should it be used?

→ It assumes that the data are missing complitely at Random.

Advantages

1. Easy to Implement

2. There is less distortion in Variance.

Dis-Advantages

1. Every Situation Randomness wont Work.

Capturing NAN Values with New Feature

You can use this by :- MNAR :-

Advantages :- (i) Easy to Implement
(ii) captures the Importance of missing Values.

Dis Advantages:- Creating Additional Features. [Curse of Dimensionality].

End of distribution Imputation

End of Distribution Imputation

Advantages :- Easy to Implement

→ Captures the Importance of missingness if there is one.

DisAdvantages :- Distorts the original distribution of the Variable.

→ If the number of NA is big, It will mask true outliers in the distribution

Arbitrary Value Imputation

The Technique was derived from kaggle competition. It consist of the method to Replace the NAN Values by arbitrary Values.

Advantages :- (i) Easy to Implement
(ii) Captures the importance of missingness if there is one

Disadvantages :- (i) Distorts the original distribution
(ii) Hard to decide which value to use.

Handling Categarical Missing Values

1 Frequent Category Imputation

Advantages :- (i) Easy to implement

Disadvantage :- (i) It may lead the distortion in the Relation of the most frequent label

(ii) Since we are using the more frequent labels, it may use them in an over represented way, if there are many nan's.

2 Adding a Variable to capture NAN