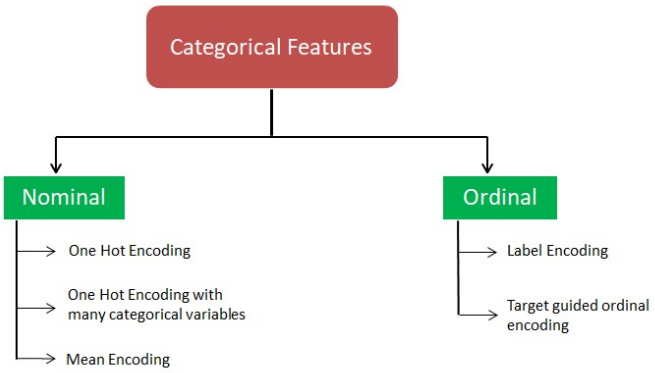
Encoding

Monday, January 30, 2023 7:30 PM



- Nomial encoding :
  - ONE HOT ENCODING
  - ONE HOT ENCODING WITH MANY CATEGORICAL
  - MEAN ENCODING

- Ordinal encoding :
  Ex: education column: 10, 12, ug, pg, phd
  rank : phd, pg, ug, 12, 10
  - LABEL ENCODING
  - TARGET GUIDED ORDINAL ENCODING

- **Nomial encoding**
- **ONE HOT ENCODING:**
  - Bascially applied to nominal categorical variables

	Germany	France	Spain
Germany	1	0	0
France	0	1	0
Spain	0	0	1

  - Dummy variable Trap
    Whenever we are doing one hot encoding we are supposed to delete one column, be it 1st or last or any column in between.
    If there are 5 columns or categories then the total number of columns that will get created will be 5 - 1 = 4
  - Disadvantages of one hot encoding:

- **ONE HOT ENCODING WITH MULTIPLE CATEGORIES:**
  - Multiple categories -> Many many categories
  - For Nominal categories only
  - If there are are 50 categories or columns, instead of taking 49 columns here we take only the top 10 repeating categories and that becomes(10-1=9) 9 columns wrt one hot encoding

- **MEAN ENCODING:**

F1	o/p
A	1
B	0
C	1
D	1
A	0
B	1

Finding the mean for each category(i.e A,B,C,D)
For example : mean of A : 0.73, B= 0.6, c =0.5 , 0.4

This is applicable for pincodes, we try to convert it into integer or float, and then take mean for each category (i.e each pincode) so the mean value will be replaced in place of pincode

C	1

Ordinal encoding

- **LABEL ENCODING:**
  - For ordinal category only
  - Higher the importance of the variable give a higher number
    Ex: UG PG MPHIL PHD
    1 2 3 4

- **TARGET GUIDED ORDINAL ENCODING:**
  - For ordinal category only
  -

F1	o/p
A	1
B	0
C	0
B	1
C	1

Now take the mean for mean category
Here, take mean for A, B, C
EX: mean of A = 0.4 --> rank 1
mean of B =0.6 ---> rank 2
mean of C = 0.75 --->rank 3
Higher the mean, higher the rank
NOW, we can assign labels based on the rank

labels	F1	o/p
1	A	1
2	B	0
3	C	0
2	B	1
3	C	1