Page No.
Date

# One Hot coding

It is used for nominal datasets.

| ex: Color | Target |
|---|---|
| Yellow | 0 |
| Red | 1 |
| Blue | 1 |
| Yellow | 1 |
| Yellow | 0 |
| Blue | 0 |

How if we do ordinal coding then our model will think like colors have ranking ex: if Y=0, R=1, B=2 Model thinks Y < R < B

∴ we use One Hot encoding

| Color_Y | Color_R | Color_B | Target |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 |
| 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 |

∴ 100 → Yellow
010 → Red
001 → Blue

Page No.
Date

## ☆ Dummy Variable Trap

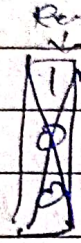
→ Remove 1 column

Col. count

Befor OHE → n

After OHE → n-1

Mostly we remove 1st column.




| | | |
|---|---|---|
| Y → | 1 0 0 | Y → 0 0 |
| B | 0 0 1 | B → 0 1 |
| R | 0 1 0 | R → 1 0 |

{ we do this to avoid multicolinarity

## ☆ OHE using most freq. variables.

In our example dataset

Brand [nominal column] → 40 brands

from this 40 brands select 4 frequent brands & for other create diff. column.