**Types of Encoding**

1. Nominal Encoding – No rank

* One Hot Encoding
* One Hot Encoding with many categorical variable
* Mean Encoding

1. Ordinal Encoding – Order present

* Label Encoding
* Target guided ordinal encoding

**ONE HOT ENCODING – Nominal Categories**

One Hot Encoding creates Dummy variable trap

|  |  |  |  |
| --- | --- | --- | --- |
| States | Germany | France | Spain |
| Germany | 1 | 0 | 0 |
| France | 0 | 1 | 0 |
| Spain | 0 | 0 | 1 |

When doing one-hot encoding, there is always a need to delete one column – it can be either first, second or last column, whichever column it may be. When Spain is 1, Germany and France are 0. Thus, with these 0 values, we can note that Spain is 1, the actual value here.

**Disadvantage of One Hot Encoding**

If we have many unique categories in the categorical feature such as PIN Code, one-hot encoding will create too many columns 🡪 curse of dimensionality

PIN Code – 100 categories 🡪 with encoding – 99 left 🡪 too many columns present in the data frame.

How to fix this?

**LABEL ENCODING – Ordinal Categories**

|  |  |
| --- | --- |
| **Education** | **Rank given** |
| BE | 1 |
| Master | 3 |
| PHD | 4 |
| Statistics | 2 |

Give higher value to the most important qualification so that model can understand the weight this particular value has.

**ONE HOT ENCODING WITH MULTIPLE CATEGORIES**

Feature having 50 categories 🡪 cannot create 49 columns

See the top 10 categories which are repeated in this feature

Apply One Hot encoding to the top 10 categories 🡪 get 9 columns

**TARGET GUIDED ORDINAL CATEGORIES – Ordinal Category**

|  |  |  |  |
| --- | --- | --- | --- |
| | **City** | | --- | | | **Purchased** | | --- | |
| New York | 1 |
| Los Angeles | 0 |
| Chicago | 1 |
| New York | 1 |
| |  |  | | --- | --- | | Chicago | 0 | | 0 |
| Los Angeles | 0 |

 **Compute Mean Target Value**:

* New York: (1 + 1) / 2 = 1.0
* Los Angeles: (0 + 0) / 2 = 0.0
* Chicago: (1 + 0) / 2 = 0.5

 **Rank Categories**:

* New York: 1.0 – Highest mean
* Chicago: 0.5
* Los Angeles: 0.0

 **Assign Ordinal Values**:

* New York: 2 – Ranked first
* Chicago: 1
* Los Angeles: 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| | **City** | | --- | |  | | | **Purchased** | | --- | |
| 2 | 1 |
| 0 | 0 |
| 1 | 1 |
| 2 | 1 |
| 1 | 0 |
| 0 | 0 |

MEAN ENCODING

F1 O/P Mean

A 1 A = 0.73

B 0 B = 0.6

C 1 C = 0.5

D 1 D = 0.4

A 0

B 1

C 1

D 1

These means are replaced in the feature🡪 instead of A, 0.73 will be inserted

This will be applicable in the PIN Code case so as to avoid getting numerous columns