**Feature Encoding:**

* Nominal Encoding – Gender, State – No need to worry about order
  + One Hot Encoding
  + One hot Encoding with Many Variables
  + Mean Encoding
* Ordinal Encoding – Have some Rank/Order – Education/ Position in Company
  + Label Encoding
  + Target Guided Ordinal Encoding

**One Hot Encoding**

* Dummies Variables
* State – Germany, France, Spain
* Remove one column in this case, i.e we will drop Spain or Germany
* Cons – Features will increase in case of many categories

|  |  |  |  |
| --- | --- | --- | --- |
| **Value** | **Germany** | **France** | **Spain** |
| Germany | 1 | 0 | 0 |
| France | 0 | 1 | 0 |
| Spain | 0 | 0 | 1 |

**Label Encoding**

* Education – BE, ME, PHD
* BE – 1 , ME – 2, PHD – 3

**One Hot Encoding with Many Categories**

* Let’s say we have 50 categories
* Based on KDD Orange Competition
* We check top repeating categories
* Let’s say out of 50, 10 categories are mostly repeated
* Apply one hot encoding to 10 features – i.e., 9 columns will be created

**Target Guided Ordinal Encoding**

* We calculate the mean of the output columns based on category
* Category having high mean will be given higher rank
* Based on rank we define the labels, First rank with highest label

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **O/P** | **Mean** | **Rank** | **Encoding** |
| A | 1 | 0.33 | 2 | 1 |
| B | 0 | 0 | 1 | 2 |
| A | 1 | 0.33 |  |  |
| B | 1 | 0.5 |  |  |
| A | 1 | 0.33 |  |  |

**Mean Encoding**

* We calculate the mean of the output columns based on category
* Value will be replaced by mean
* Based on rank we define the labels, First rank with highest label

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **O/P** | **Mean** | **Encoding** |
| A | 1 | 0.33 | 0.33 |
| B | 0 | 0.5 | 0.5 |
| A | 1 | 0.33 |  |
| B | 1 | 0.5 |  |
| A | 1 | 0.33 |  |