One Hot Encoding

One Hot Encoding is another popular technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. One hot encoding is the process of creating a dummy variables.

For example, We have a dataset that contains Countries, Age, and Salary here are our category The feature is country so we can convert this into the numeric form using One Hot Encoding.

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

As you can see here, 3 new features are added as the country contains 3 unique values – India, Japan, and the US. In this technique, we solved the problem of ranking as each category is represented by a binary vector.

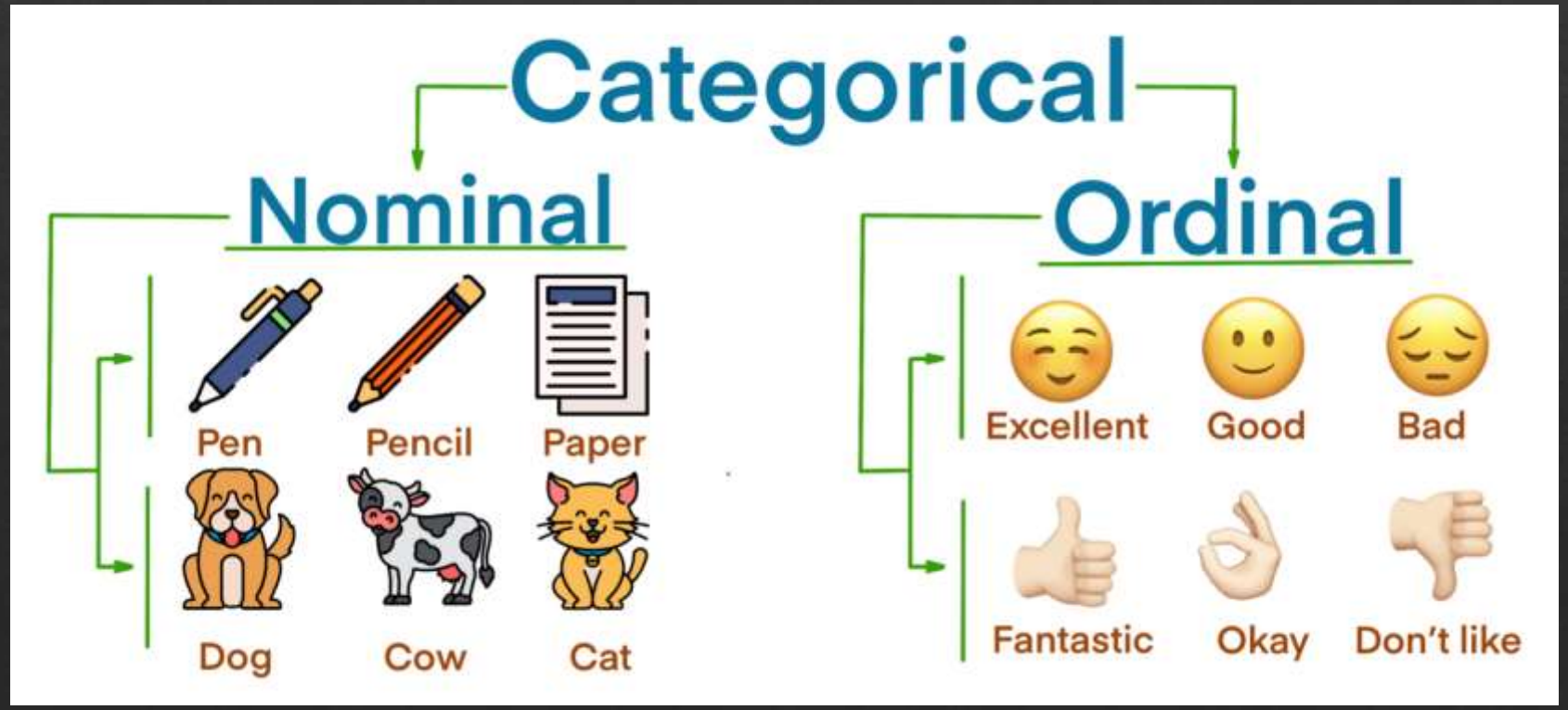
0	1	2	Age	Salary
1	0	0	44	72000
0	0	1	34	65000
0	1	0	46	98000
0	0	1	35	45000
0	1	0	23	34000

Problem with One Hot Encoding: One Hot Encoding results in a dummy variable trap as the outcome of one variable can easily be predicted with the help of the remaining variables. A Dummy variable Trap is a scenario in which variables are highly correlated to each other.

Solution: To drop one column or feature from the table or one dummy feature or variable should be dropped.

Apply One-Hot Encoding when:

- The categorical feature is not ordinal (like the countries)
- The number of categorical features is less so one-hot encoding can be effectively applied



Source: Medium

Label Encoding

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

For example, We have a dataset that contains Countries, Ages, and Salary here is our category The feature is country so we can convert this into the numeric form using Label Encoding.

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

As you can see here, the label encoding uses alphabetical ordering. Hence, India has been encoded with 0, the US With 2, and Japan with 1,

Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

Problem with Label Encoding:

When label encoding is performed, the country names are ranked based on the alphabet. Due to this, there is a high probability that the model captures the relationship between countries such as India < Japan < US.

Apply Label Encoding when:

- The categorical feature is ordinal (like Jr. kg, Sr. kg, Primary school, high school)
- The number of categories is quite large as one-hot encoding can lead to high memory consumption.

GitHub Link: https://github.com/AmeenUrRehman/Machine-Learning-Projects/tree/up-pages/Dummy_OneHot_Encoding