Topic :

1. Reducing the Number of Categories in Categorical Variables
2. Converting a Categorical Variable to a Numerical Variable

1. Reducing the Number of Categories in Categorical Variables

- When a categorical variable has many categories, and this variable is destined to be a predictor, many data mining methods will require converting it into many dummy variables.
- In particular, a variable with m categories will be transformed into m – 1 dummy variables. This means that even if we have very few original categorical variables, they can greatly inflate the dimension of the dataset.
- One way to handle this is to reduce the number of categories by combining close or similar categories.

Dealing with High Cardinality Categorical Data

- High cardinality refers to a large number of unique categories in a categorical feature.
- Dealing with high cardinality is a common challenge in encoding categorical data for machine learning models.
- High cardinality can lead to sparse data representation and can have a negative impact on the performance of some machine learning models.
- Here are some techniques that can be used to deal with high cardinality in categorical features:

Combining Rare Categories

This involves combining infrequent categories into a single category. This reduces the number of unique categories and also reduces the sparsity in the data representation.

Target Encoding

Target encoding replaces the categorical values with the mean target value of that category. It provides a more continuous representation of the categorical data and can help capture the relationship between the categorical feature and the target variable.

workclass	target
State-gov	0
Self-emp-not-inc	1
Private	0
Private	0
Private	1

workclass	target mean
State-gov	0
Self-emp-not-inc	1
Private	1/3

workclass
0
1
1/3
1/3
1/3

Creating interaction variables

Interaction variables are new features created by combining two or more existing features. For example, if we have two categorical features, 'Gender' and 'Marital Status,' we can create a new feature, 'Gender-Marital Status,' to capture the interaction between the two features. This can help to capture non-linear relationships between the features and the target variable.

INTERACTION EFFECT

Gender	Marital	Gender_Marital
Male	Married	Male and Married
Male	Unmarried	Male and UnMarried
Female	Unmarried	Female and UnMarried
Male	Married	Male and Married

Binning numerical variables

Binning is the process of dividing continuous numerical variables into discrete bins. This can help to reduce the number of unique values in the feature, which can be beneficial for encoding categorical data. Binning can also help to capture non-linear relationships between the features and the target variable.

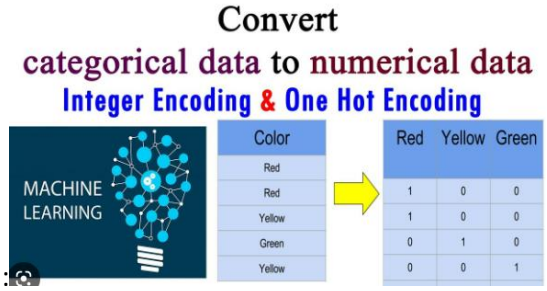
BINNING EFFECT

Age	Age_bin
5	0 - 10 year
1	0 - 10 year
21	20 - 30 year
25	20 - 30 year
36	30 - 40 year
39	30 - 40 year
55	>50 year
67	>50 year

2. Converting a Categorical Variable to a Numerical Variable

   Converting A Categorical Variable to A Numerical Variable Sometimes the categories in a categorical variable represent intervals. Common examples are age group or income group. If the interval values are known (e.g., category 2 is the age interval 20–30), we can replace the categorical value (“2” in the example) with the mid interval value (here “25”). The result will be a numerical variable that no longer requires multiple dummy variables.

   Examples:



Courtesy: https://www.youtube.com/watch?v=Hlmsz-HEqyY

	id	Gender	Age	Department	Rating
0	101	M	21	QA	A
1	102	M	25	QA	B
2	103	M	24	Dev	B
3	104	F	28	Dev	C
4	105	F	25	UI	B

	id	Gender	Age	Rating	Department_Dev	Department_QA	Department_UI
0	101	1	21	3	0	1	0
1	102	1	25	2	0	1	0
2	103	1	24	2	1	0	0
3	104	0	28	1	1	0	0
4	105	0	25	2	0	0	1

Courtesy: https://thinkingneuron.com/how-to-convert-categorical-string-data-into-numeric-in-python/