Train In Data

Unseen categories
Basic encoding

Rare labels & high cardinality

- **Rare labels** are those that appear only in a tiny proportion of the observations in a dataset.

- **High cardinality**: lots of unique categories.
  - Usually causes categories to be unseen.

Rare label example

Obs	Vehicle Make
1	Mercedes
2	Ford
3	Renault
4	Seat
5	Renault
6	Ford
7	Kia
8	Kia
9	Nissan
10	BMW

Train Set

Obs	Vehicle Make
1	Mercedes
5	Renault
6	Ford
8	Kia
9	Nissan

Test Set

Obs	Vehicle Make
2	Ford
4	**Seat**
3	Renault
7	Kia
10	**BMW**

Rare label in train set

Train Set

Obs	Vehicle Make
1	Mercedes
5	Renault
6	Ford
8	Kia
9	Nissan

Rare label in train set - ohe

Train Set

Obs	Vehicle Make
1	Mercedes
5	Renault
6	Ford
8	Kia
9	Nissan

Merced	Renault	Ford	Kia	Nissan
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Rare label in test set - ohe

Test Set

Obs	Vehicle Make
2	Ford
4	**Seat**
3	Renault
7	Kia
10	**BMW**

Merced	Renault	Ford	Kia	Nissan
0	0	1	0	0
0	0	0	0	0
0	1	0	0	0
0	0	0	1	0
0	0	0	0	0

Rare label in train set – ordinal or count

Train Set

Obs	Vehicle Make
1	Mercedes
5	Renault
6	Ford
8	Kia
9	Nissan

Ordinal
1
2
3
4
5

Count
1
1
1
1
1

Rare label in test set – ordinal or count

Test Set

Obs	Vehicle Make
2	Ford
4	**Seat**
3	Renault
7	Kia
10	**BMW**

Ordinal
3
NAN
2
4
NAN

Count
1
NAN
1
1
NAN

How do the Python libraries handle unseen categories?

Scikit-learn OneHotEncoder

handle_unknown : *{'error', 'ignore', 'infrequent_if_exist'}, default='error'*

Specifies the way unknown categories are handled during `transform`.

- 'error' : Raise an error if an unknown category is present during transform.
- 'ignore' : When an unknown category is encountered during transform, the resulting one-hot encoded columns for this feature will be all zeros. In the inverse transform, an unknown category will be denoted as None.
- 'infrequent_if_exist' : When an unknown category is encountered during transform, the resulting one-hot encoded columns for this feature will map to the infrequent category if it exists. The infrequent category will be mapped to the last position in the encoding. During inverse transform, an unknown category will be mapped to the category denoted `'infrequent'` if it exists. If the `'infrequent'` category does not exist, then `transform` and `inverse_transform` will handle an unknown category as with `handle_unknown='ignore'`. Infrequent categories exist based on `min_frequency` and `max_categories`. Read more in the User Guide.

Feature-engine OneHotEncoder

- All zeroes
- Note that if encoding into k-1, then the unseen category will be blended with the dropped category.

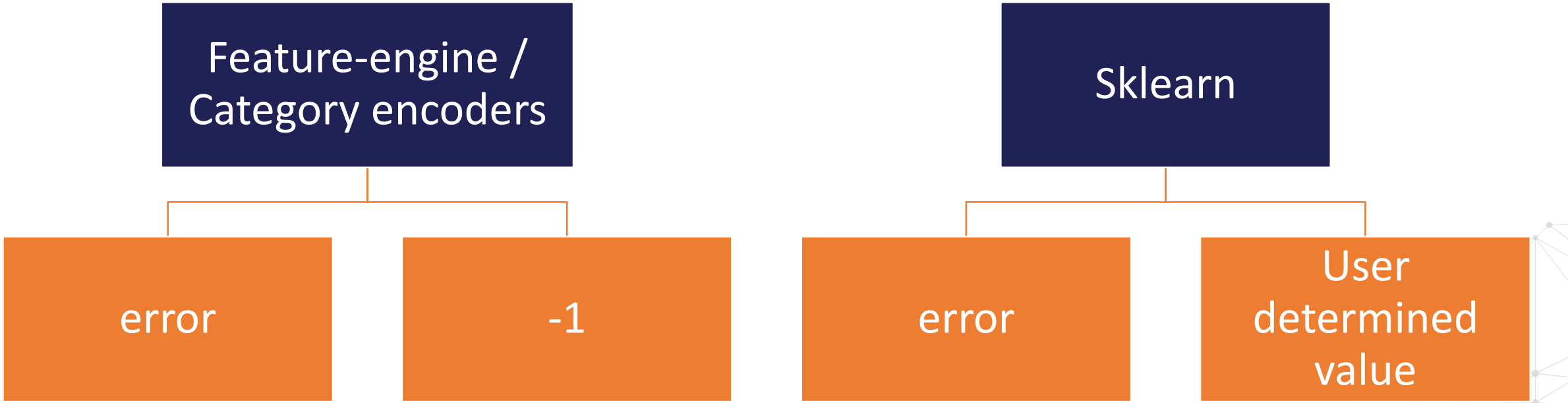
Category encoders OneHotEncoder

handle_unknown: str

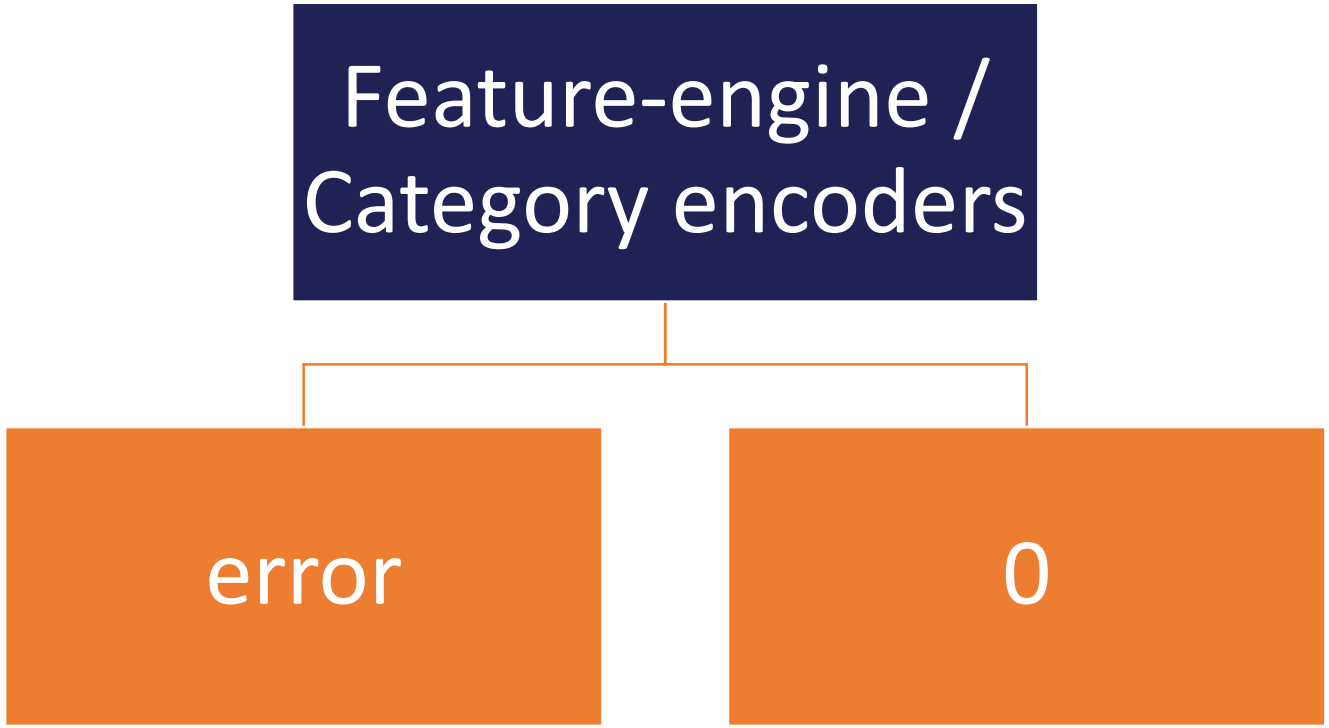
options are ‘error’, ‘return_nan’, ‘value’, and ‘indicator’. The default is ‘value’.

‘error’ will raise a *ValueError* at transform time if there are new categories. ‘return_nan’ will encode a new value as *np.nan* in every dummy column. ‘value’ will encode a new value as 0 in every dummy column. ‘indicator’ will add an additional dummy column (in both training and test data).

OrdinalEncoder



Count encoder



What to do?

My advice:

- Try to do something that lets you know if there are unseen categories.
  - ➢ You learn more about your data.
  - ➢ You don’t blend unseen with other levels of the variable by mistake.

Train In Data

THANK YOU

www.trainindata.com