Train In Data

Categorical Encoding

Categorical Encoding

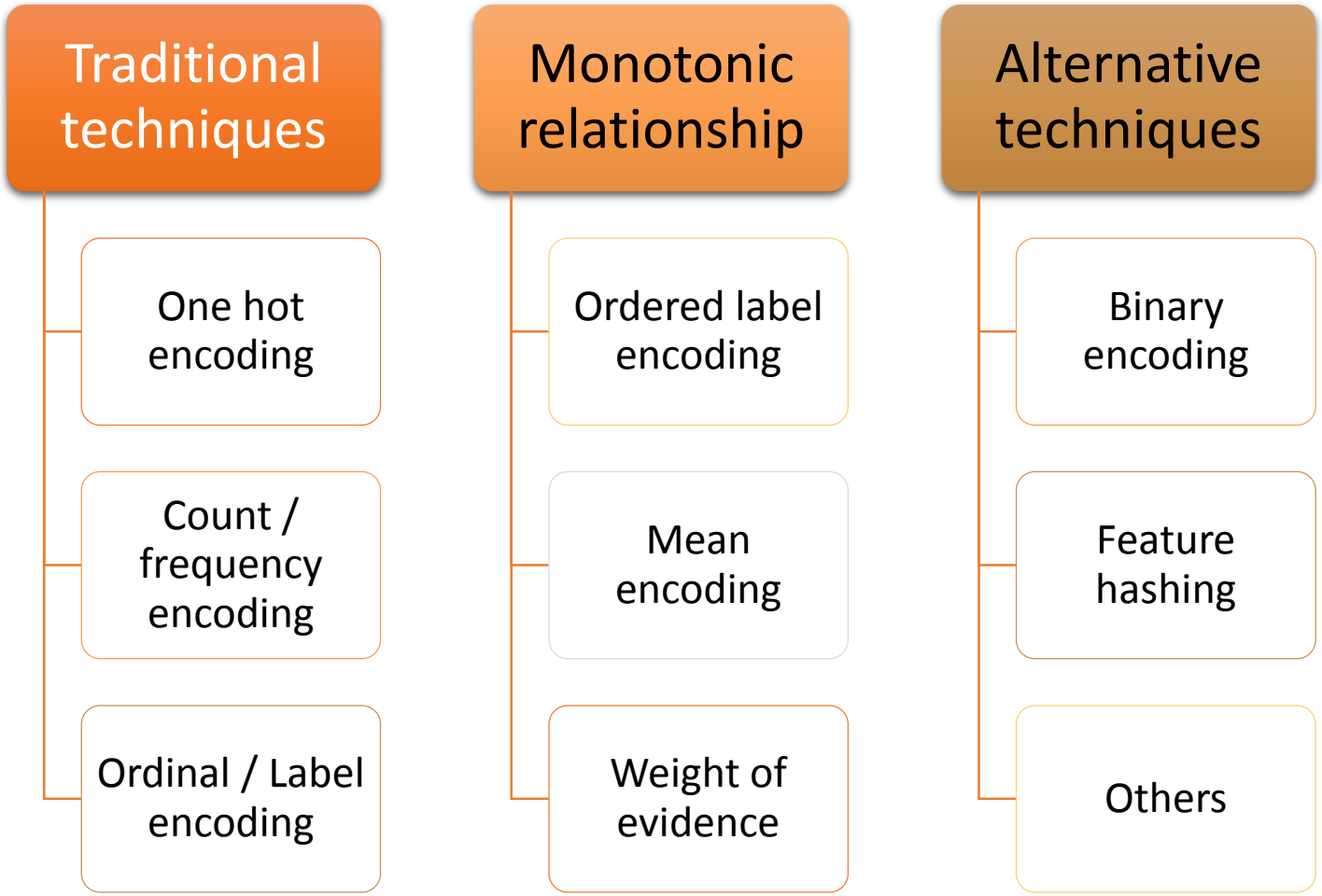
- Categorical encoding refers to replacing the category strings by a numerical representation

The goal of categorical encoding is:

- To produce variables that can be used to train machine learning models

- To build predictive features from categories

Categorical Encoding Techniques



Monotonic relationship

What is a monotonic relationship between the variable and the target?

- When variable increases and the target increases. Or,
- When variable increases and the target decreases.

Monotonic relationship

- Improve the performance of linear models
- May improve the performance of tree based models
  - Creates shallower trees
- Often, organisations want to include monotonic constrains
  - Insurance premiums decreasing with age

Encoding Techniques: Rare labels

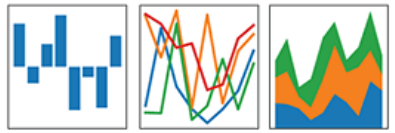
- ✓ One hot encoding of frequent categories
- ✓ Grouping of rare categories

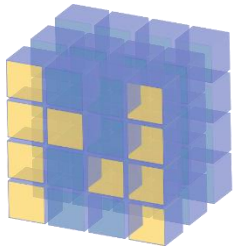
Particularly important for model deployment

Categorical Encoding Techniques

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$





NumPy

Feature-Engine

scikit learn

Category Encoders

Objectives

Understand the different techniques for categorical encoding.

- Learn multiple techniques
- Understand their impact on the variable and the machine learning model
- Learn how to implement it with pandas, Scikit-learn, and Feature-Engine, within a machine learning pipeline

Content

For each lecture:

- Presentation and video
- Accompanying Jupyter notebook
  - Explanation of the technique
  - Implementation in pandas and Numpy
  - Implementation in Scikit-learn (when possible)
  - Implementation in Feature-engine

Final Summary

- Final lecture comparing the performance of the different categorical encoding techniques with different machine learning models.

- Additional reading resources.

Train In Data

THANK YOU

www.trainindata.com