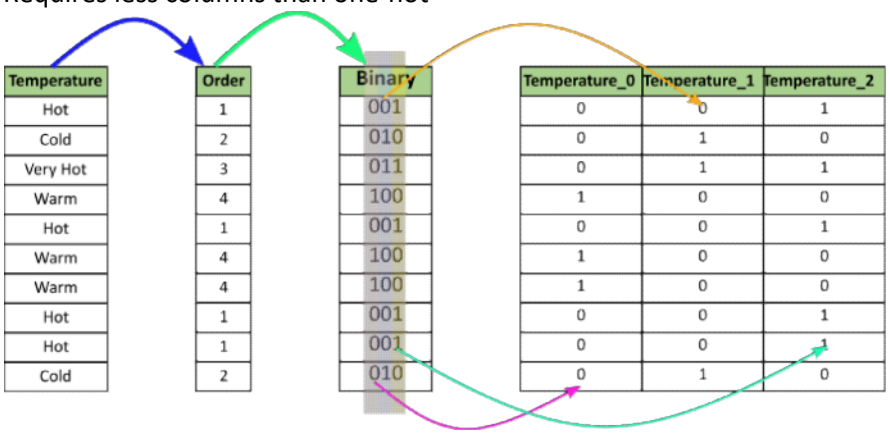
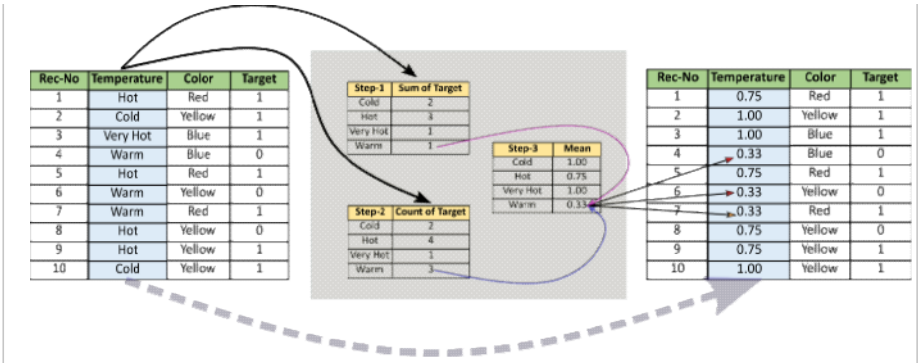
Encoding methods

09 October 2022 10:56 PM

https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding/#:~:text=Encoding%20categorical%20data%20is%20a,provided%20to%20the%20different%20models.&text=In%20the%20field%20of%20data,preparation%20is%20a%20mandatory%20task.
https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02
https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

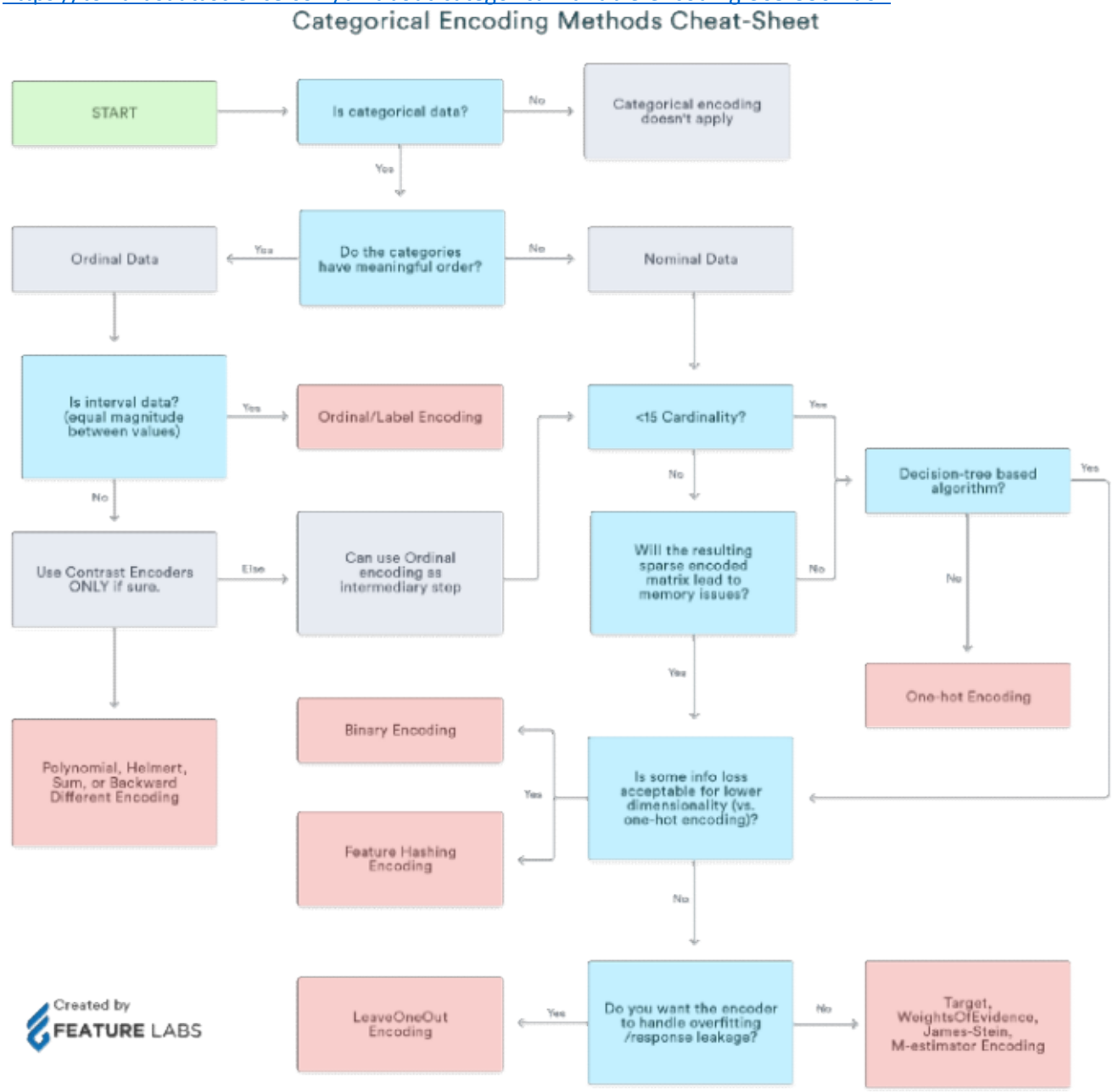
Encoding method	**Description**
Label/Ordinal encoding	• Converts each label into integer, with value representing sequence • Used for ordinal data
One-Hot Encoding	• Each class gets becomes a new variable, called dummy variables - maps each class into binary • Used for nominal data Note: Dummy encoding is the same but you remove one level
Effect/Deviation/Sum Encoding	• Classes are put into -1, 0, 1 format (has an additional -1 compared to one-hot) https://jamesmccaffrey.wordpress.com/2019/10/17/effect-coding-vs-one-hot-encoding-for-neural-networks/ • Usually used in multiple linear regression, when two categorical variables statistically interact • People tend to use one-hot over effect because it looks nicer and it is symmetrical • However, both works the same - no free lunch theorem
Binary Encoding	• Converts a category into binary digits • Requires less columns than one-hot  Note: Base-N Encoding is an extension of binary encoding, where the base is a hyperparameter
Mean/Target Encoding	• Converting a class into the mean of the target variable for its class (type of Bayesian encoding) • Brings out relation between similar categories • Does not affect volume of data • Helps in faster learning • BUT, tends to overfit

Temperature	Order	Binary	Temperature_0	Temperature_1	Temperature_2
Hot	1	001	0	0	1
Cold	2	010	0	1	0
Very Hot	3	011	0	1	1
Warm	4	100	1	0	0
Hot	1	001	0	0	1
Warm	4	100	1	0	0
Warm	4	100	1	0	0
Hot	1	001	0	0	1
Hot	1	001	0	0	1
Cold	2	010	0	1	0



Note: There are two types of categorical data - **Nominal** (no order) and **Ordinal** (order matters)

https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02



Dataset

- 33 variables; 32 features & 1 target variables