**# Day19-Lab-FE**

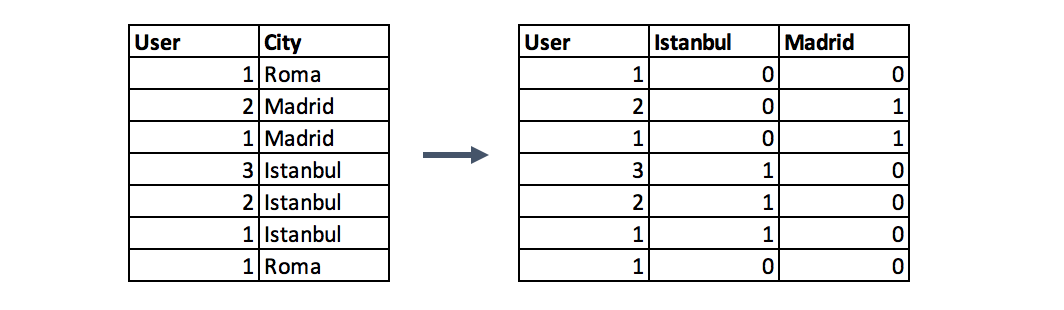
**# One-Hot / Dummy Encoding**

\* How it works ?

In one-hot encoding, categorical data are represented as vectors of zeros and ones. This is done by using a separate dummy variable for each category, and setting the value of the dummy variable to 1 if the observation belongs to that category and 0 otherwise.

\* Example :

Example, suppose we have a column containing 3 categorical variables, then in one hot encoding 3 columns will be created each for a categorical variable.



\* Assumptions :

one-hot encoding is the most commonly used method for nominal variables. It is simple to understand and implement, and it works well with most machine learning models.

\* Unapplicable cases:

Suppose we have a column which has 100 categorical variables. Now if we try to convert the categorical variables into dummy variable then we will get 99 columns. This will increase the dimension of the overall dataset which will lead to curse of dimensionality.

So basically, if there is a lot of categorical variables in a column then we should not apply this technique.

**# Label / Ordinal Encoding**

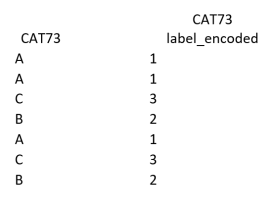
This is probably the simplest way to encode features for a machine learning algorithm.

\* How it works ?

In this method, the categorical data is converted into numerical data. Each category is assigned a numerical value.

\* Example :

Below table illustrates that PHD is considered as the highest degree, so the highest label is given to it and so on.



\* Assumptions :

Ordinal encoding is a good choice if the order of the categorical variables matters., The label is easily reversible and doesn’t increase the dimensionality of the data.

\* Unapplicable cases:

Label coding considers some hierarchy in the columns which can mislead to nominal features present in the data set.

**# Target Encoding**

\* How it works ?

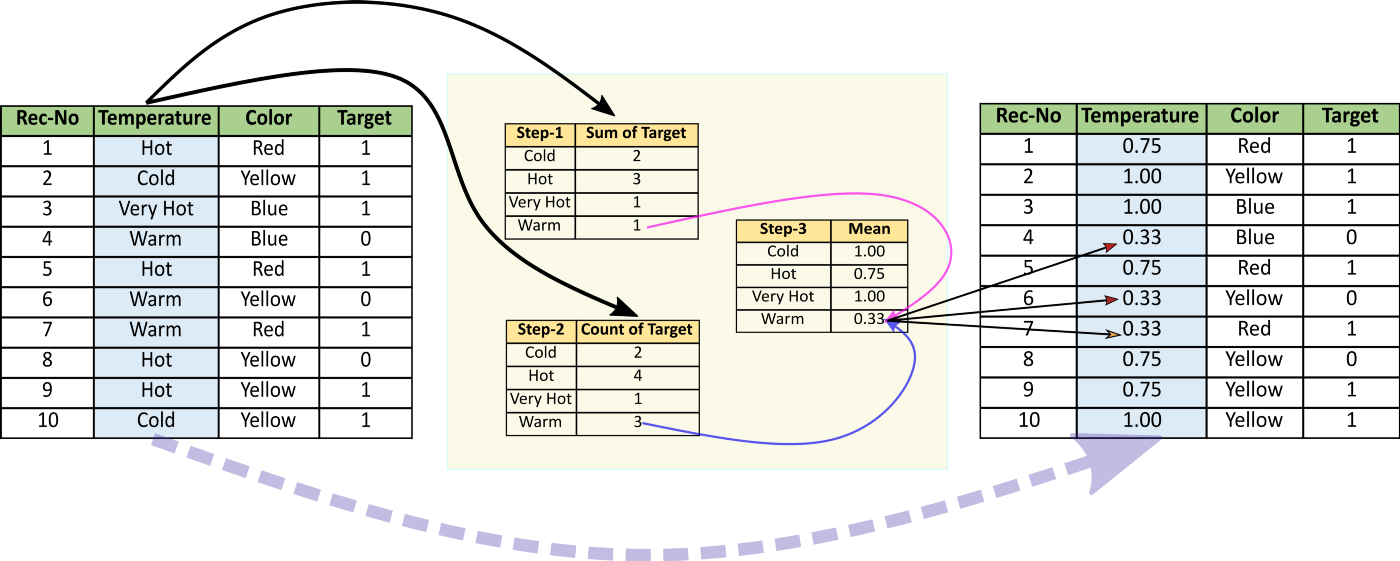
Target encoding is a technique for encoding the categorical values of the features by using the target value. The idea behind this technique is that if the feature is a good predictor of the target, then its values should be closer to the target.

1- Target-mean encoding: we replace the category with the mean of the target values. This method will usually be used with smoothing to avoid target leakage.

2- Leave-one-out encoding: this method is very similar to target mean encoding, but the difference is that in leave-one-out encoding, we take the mean of the target values of all the samples except the one we want to predict.

\* Example :

For example, in mean target encoding for each category in the feature label is decided with the mean value of the target variable on a training data.



\* Assumptions :

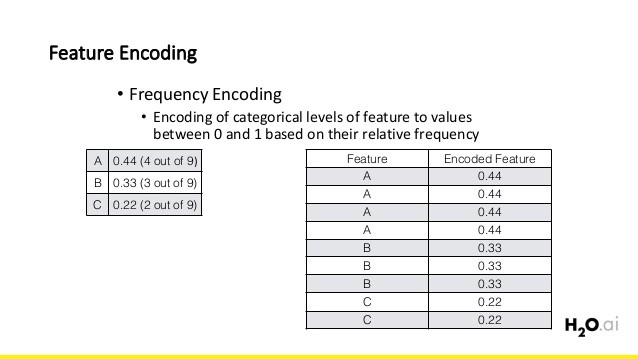
the target encoding is a supervised encoder that captures information about the label and potentially predictive features. This encoder does not increase the dimensionality of the feature space, but can lead to overfitting and is prone to target leakage.

**# Frequency / Count Encoding**

\* How it works ?

Count encoding is a way of representing categorical data using the count of the categories. Frequency encoding is simply a normalized version of count encoding.

\* Example :



\* Assumptions :

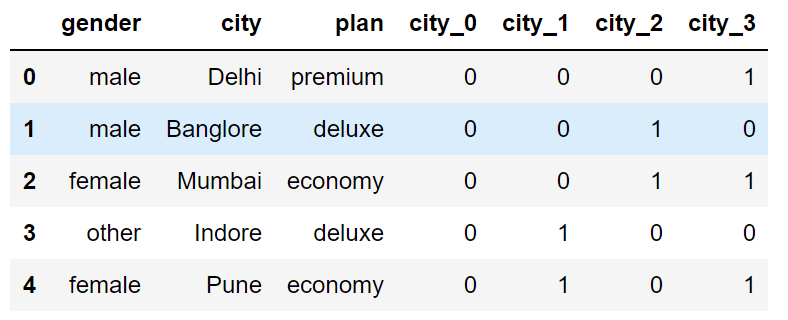
The frequency and count encoders are also supervised methods that do not increase the dimensionality of the feature space. However, these methods can only be used if the count refers to the target variable, otherwise, all categories that have similar cardinality will be counted the same.

**# Binary Encoding**

\* How it works ?

Binary encoding is a technique used to transform categorical data into numerical data by encoding categories as integers and then converting them into binary code.

\* Example :



\* Assumptions :

To fight the curse of dimensionality, binary encoding might be a good alternative to one-hot encoding because it creates fewer columns when encoding categorical variables.

**\*\*\* REFERENCES:**

https://towardsdatascience.com/6-ways-to-encode-features-for-machine-learning-algorithms-21593f6238b0

<https://ai-ml-analytics.com/encoding/>

<https://pub.towardsai.net/5-useful-encoding-techniques-in-machine-learning-f735567399f4>

<https://medium.com/analytics-vidhya/different-type-of-feature-engineering-encoding-techniques-for-categorical-variable-encoding-214363a016fb>

https://yogeshchauhan09.medium.com/categorical-data-encoding-in-machine-learning-8c5e30a19585