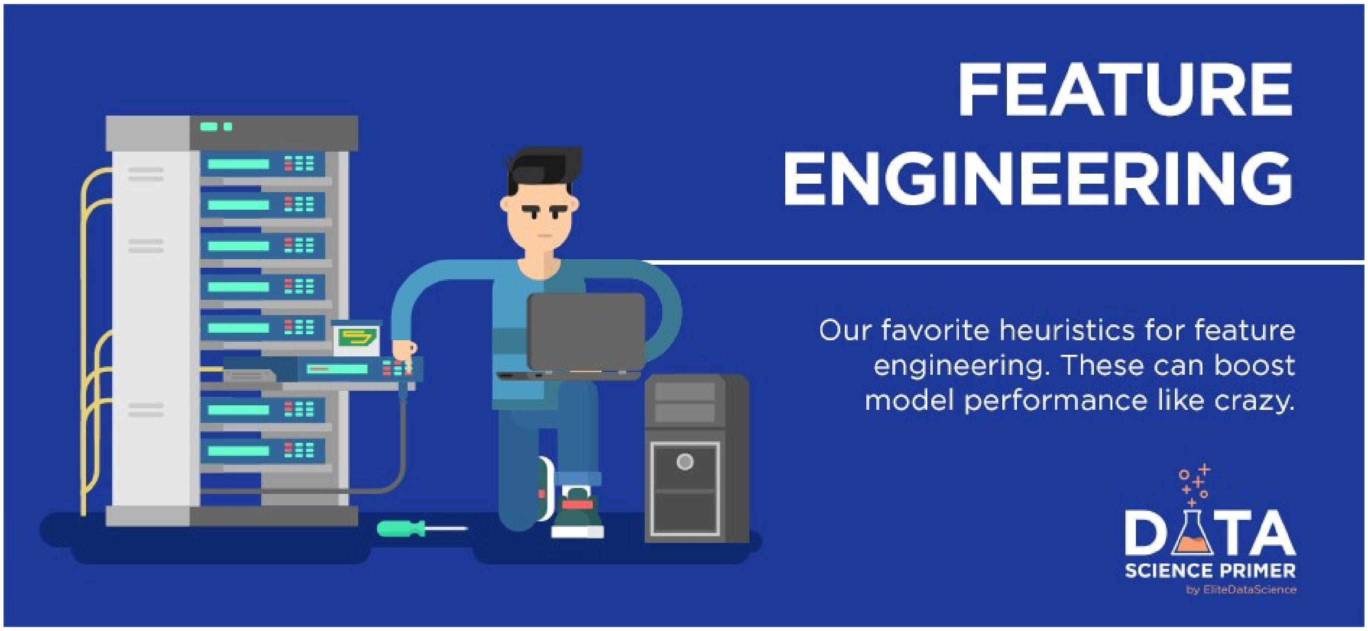
Different Type of Feature Engineering Encoding Techniques for Categorical Variable Encoding

Himanshu Tripathi Follow
Sep 20, 2019 · 4 min read


FEATURE
ENGINEERING
Our favorite heuristics for feature engineering. These can boost model performance like crazy.
DATA
SCIENCE PRIMER
by EliteDataScience

https://elitedatascience.com/wp-content/uploads/2018/05/Feature-Engineering-Banner-940px.jpg

“Let’s Create **New features** from existing ones.”

What we will be covering in this article

- What is Feature Engineering?
- Why Feature Engineering is important.
- What are encoding techniques? (Also types of Encoding Techniques)

What is Feature Engineering?

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive. The need for manual feature engineering can be obviated by automated feature learning.

Feature engineering is an informal topic, but it is considered essential in applied machine learning.

> *Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.*
>
> — Andrew Ng, Machine Learning and AI via Brain simulations

Why Feature Engineering is important?

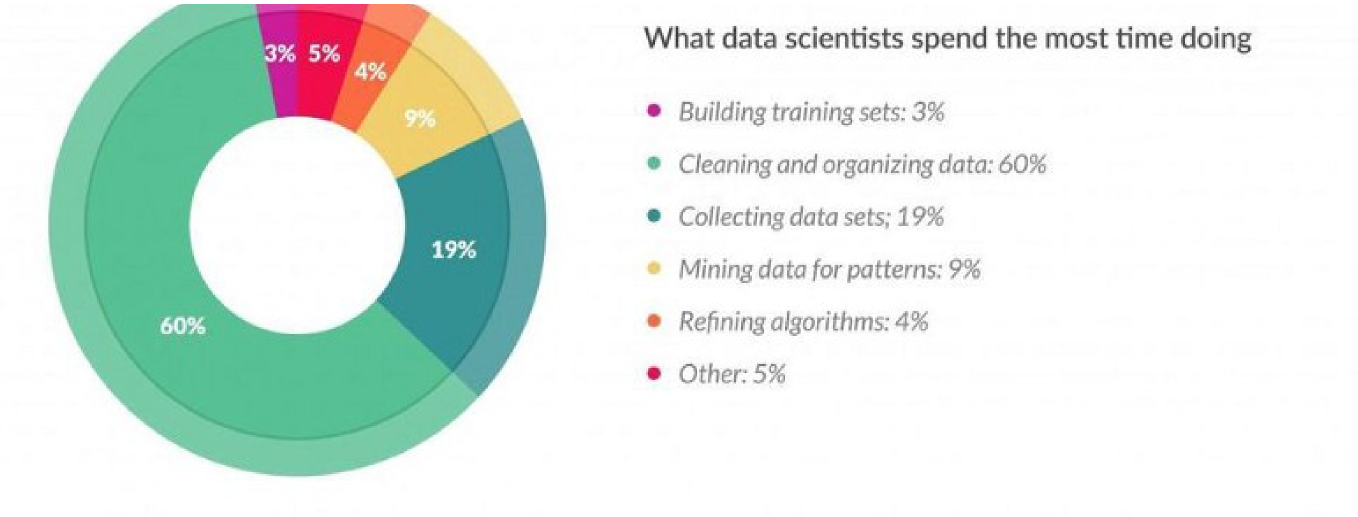
If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process. Feature Engineering is an art.

For solving any Machine Learning Problem there are some steps that we have to follow

- Gathering data.
- Cleaning data.
- **Feature Engineering.**
- Defining model.
- Training, testing model and predicting the output.

Basically, all machine learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristic to work properly.

Data scientists spend **80%** of their time on **Data Preparation:**



What are Encoding Techniques? And Types of Encoding Techniques

In many practical data science activities, the data set will contain categorical variables. These variables are typically stored as text values”. Since machine learning is based on mathematical equations, it would cause a problem when we keep categorical variables as is. Many algorithms support categorical values without further manipulation, but in those cases, it’s still a topic of discussion on whether to encode the variables or not. The algorithms that do not support categorical values, in that case, are left with encoding methodologies.

Encoding Methodologies

Nominal Encoding: — Where Order of data does not matter

In Nominal Encoding we have various techniques:

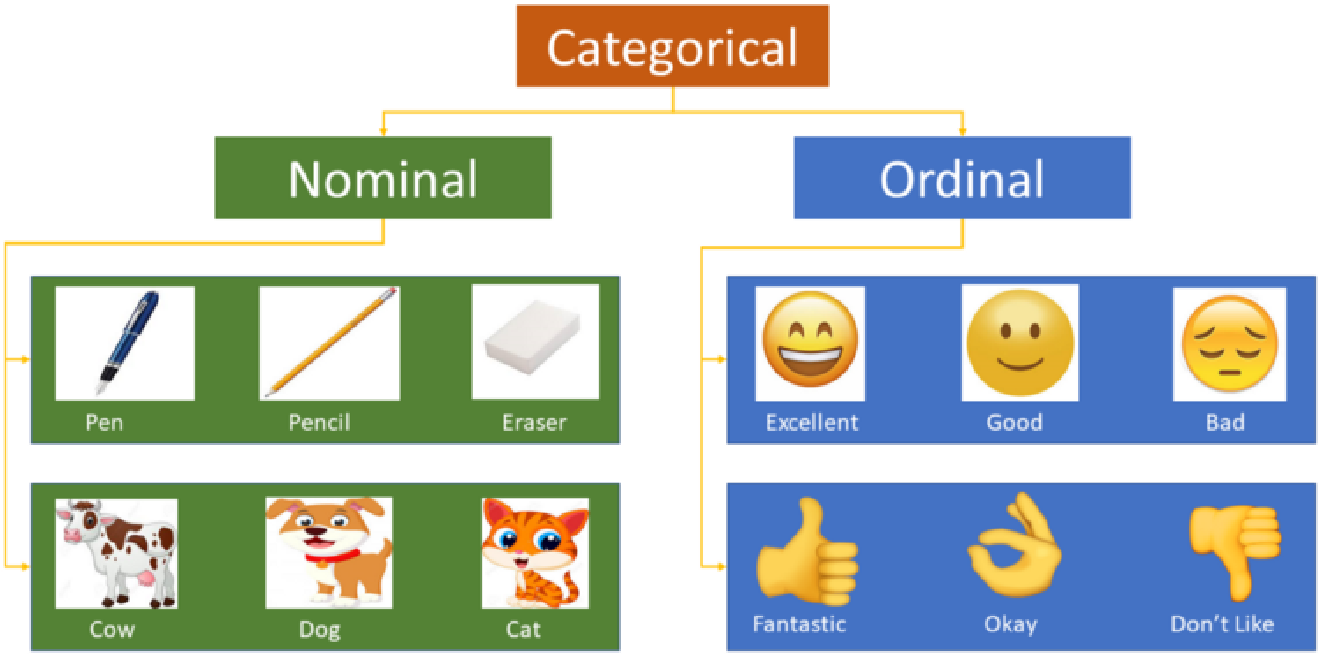
- One Hot Encoding
- One Hot Encoding With Many Categories
- Mean Encoding

Ordinal Encoding: — Where Order of data matters

In Ordinal Encoding we also have various techniques

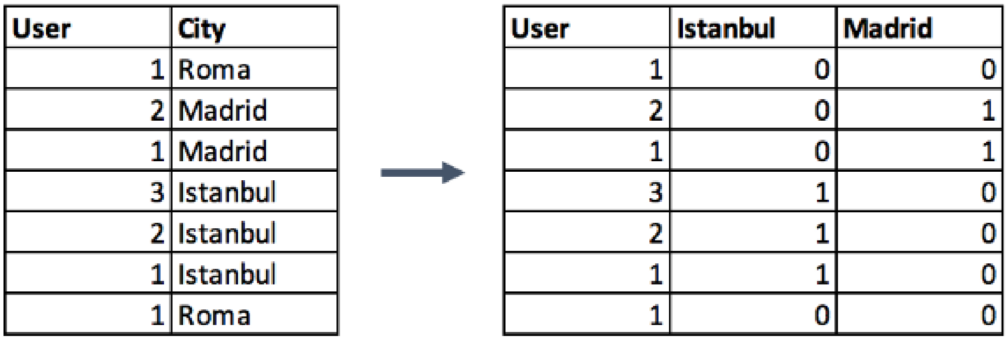
- Label Encoding

- Target Guided Ordinal Encoding



Let's Talk About Some Encoding Techniques: -

One Hot Encoding: — In this method, we map each category to a vector that contains 1 and 0 denoting the presence of the feature or not. The number of vectors depends on the categories which we want to keep. For high cardinality features, this method produces a lot of columns that slows down the learning significantly. There is a buzz between one hot encoding and dummy encoding and when to use one. They are much alike except one hot encoding produces the number of columns equal to the number of categories and dummy producing is one less. This should ultimately be handled by the modeler accordingly in the validation process.

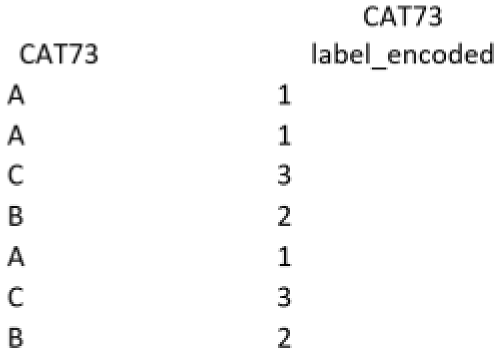


User	City
1	Roma
2	Madrid
1	Madrid
3	Istanbul
2	Istanbul
1	Istanbul
1	Roma

User	Istanbul	Madrid
1	0	0
2	0	1
1	0	1
3	1	0
2	1	0
1	1	0
1	0	0

One Hot Encoding

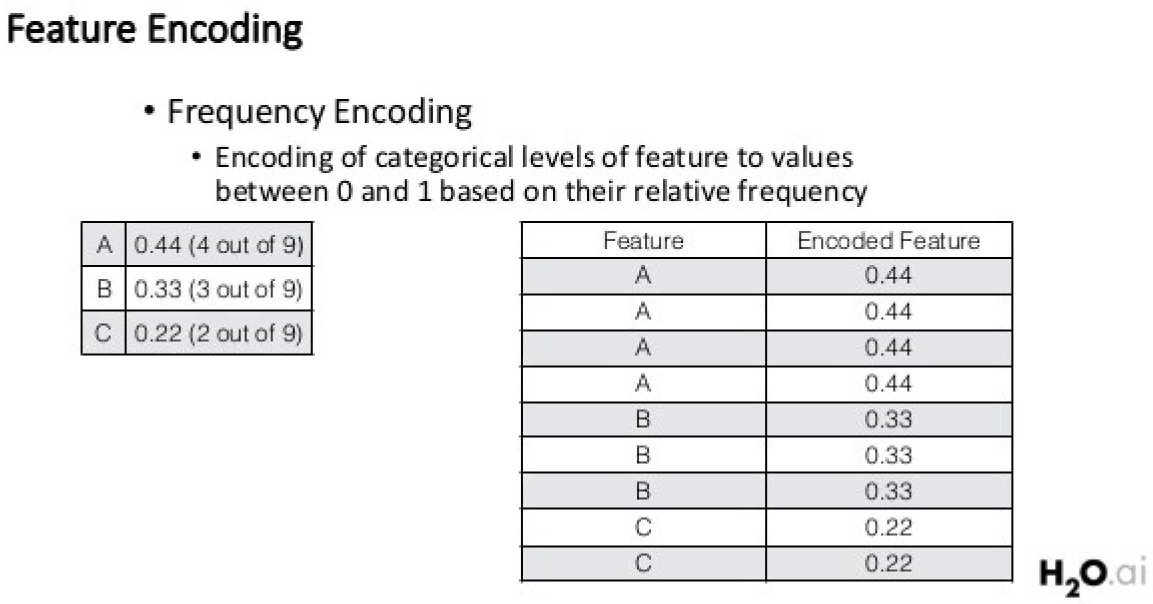
Label Encoding: — In this encoding each category is assigned a value from 1 through N (here N is the number of category for the feature). It may look like (Car<Bus<Truck ….0 < 1 < 2). Categories that have some ties or are close to each other lose some information after encoding.



CAT73	CAT73 label_encoded
A	1
A	1
C	3
B	2
A	1
C	3
B	2

label encoding

Frequency Encoding: — It is a way to utilize the frequency of the categories as labels. In the cases where the frequency is related somewhat with the target variable, it helps the model to understand and assign the weight in direct and inverse proportion, depending on the nature of the data.



Feature Encoding

- Frequency Encoding
  - Encoding of categorical levels of feature to values between 0 and 1 based on their relative frequency

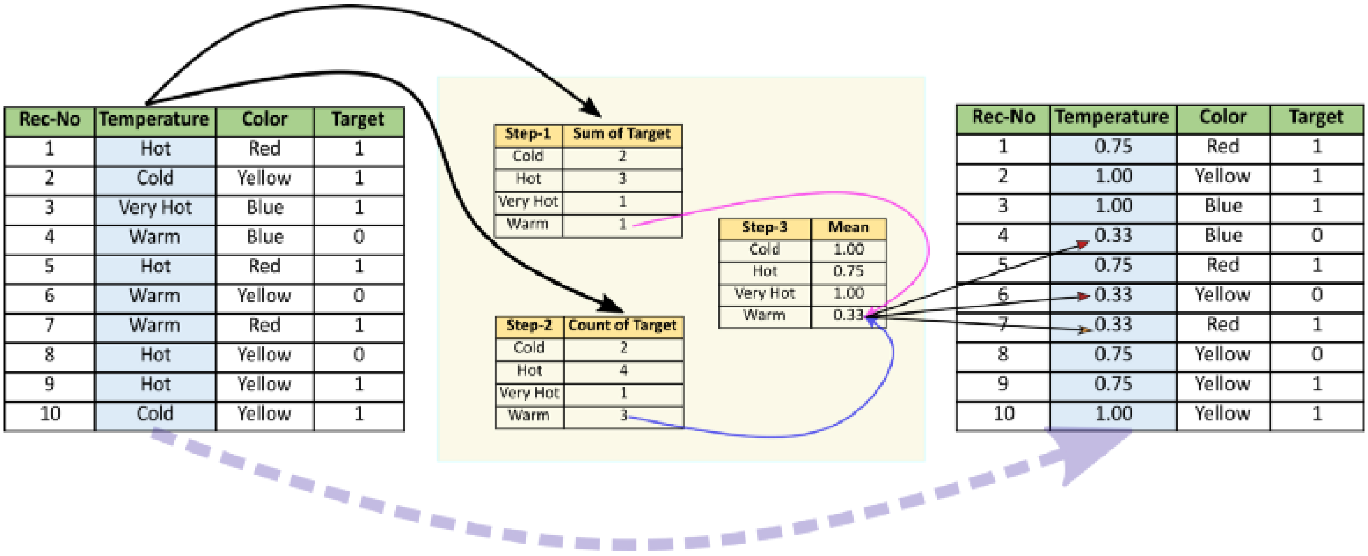
A	0.44 (4 out of 9)
B	0.33 (3 out of 9)
C	0.22 (2 out of 9)

Feature	Encoded Feature
A	0.44
A	0.44
A	0.44
A	0.44
B	0.33
B	0.33
B	0.33
C	0.22
C	0.22

frequency encoding

Mean Encoding: — Mean Encoding or Target Encoding is one very popular encoding approach followed by Kagglers. Mean encoding is similar to label encoding, except here labels are correlated directly with the target. 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.

The advantages of the mean target encoding are that it **does not affect the volume of the data** and helps in faster learning.



Rec-No	Temperature	Color	Target
1	Hot	Red	1
2	Cold	Yellow	1
3	Very Hot	Blue	1
4	Warm	Blue	0
5	Hot	Red	1
6	Warm	Yellow	0
7	Warm	Red	1
8	Hot	Yellow	0
9	Hot	Yellow	1
10	Cold	Yellow	1

Step-1	Sum of Target
Cold	2
Hot	3
Very Hot	1
Warm	1

Step-2	Count of Target
Cold	2
Hot	4
Very Hot	1
Warm	3

Step-3	Mean
Cold	1.00
Hot	0.75
Very Hot	1.00
Warm	0.33

Rec-No	Temperature	Color	Target
1	0.75	Red	1
2	1.00	Yellow	1
3	1.00	Blue	1
4	0.33	Blue	0
5	0.75	Red	1
6	0.33	Yellow	0
7	0.33	Red	1
8	0.75	Yellow	0
9	0.75	Yellow	1
10	1.00	Yellow	1

mean encoding

That's it for now.

If you found this article interesting, helpful and if you learn something from this article, please share your feedback.

Thanks for reading!

Check out my new Article "Practical Implementation of Recommendation System on Web"

References:-

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

· **https://www.datacamp.com/community/tutorials/encoding-methodologies**

And also, let's become friends on Twitter, Instagram, Github, and Facebook.

Machine Learning Data Science Feature Engineering Artificial Intelligence Deep Learning