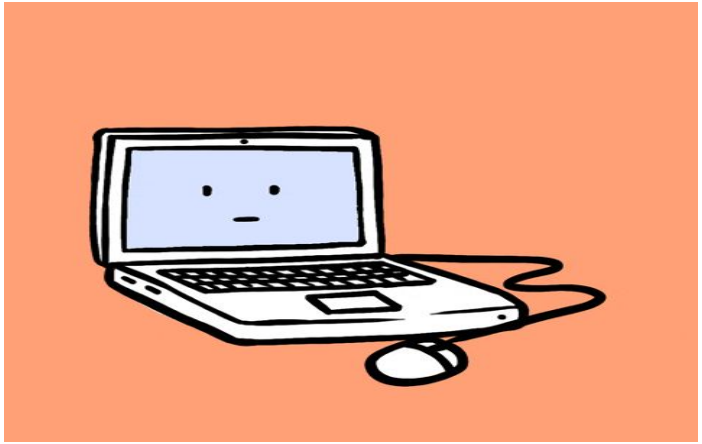
Feature Engineering

By: Omnia Hosny

What is feature Engineering?

Feature Engineering is the act of converting raw observations into desired features using statistical or machine learning approaches.



Definition Of Feature Engineering?

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.

In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

Why we need Feature Engineering?

The powerful of Feature Engineering is to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations

while also enhancing model accuracy.

Different types of data

1. Continuous: either integers (or whole numbers)
2. Categorical
3. Ordinal
4. Boolean
5. Datetime

Dealing with Categorical Variables

Encoding categorical features (bad)

- One-hot encoding
- Dummy encoding

One-hot vs. dummies

Index	Sex
0	Male
1	Female
2	Male

Index	Male	Female
0	1	0
1	0	1
2	1	0

Index	Male
0	1
1	0
2	1

One-hot vs. dummies

- One-hot encoding: Explainable features
- Dummy encoding: Necessary information without duplication

Another problem in get_dummies

If we have many values at categorical feature.

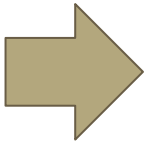
Country
USA
India
UK
Egypt

Types of numeric features

- Age
- Price
- Counts
- Geospatial data

Binarizing numeric variables

Work	No_of_children
Yes	2
No	1
No	0



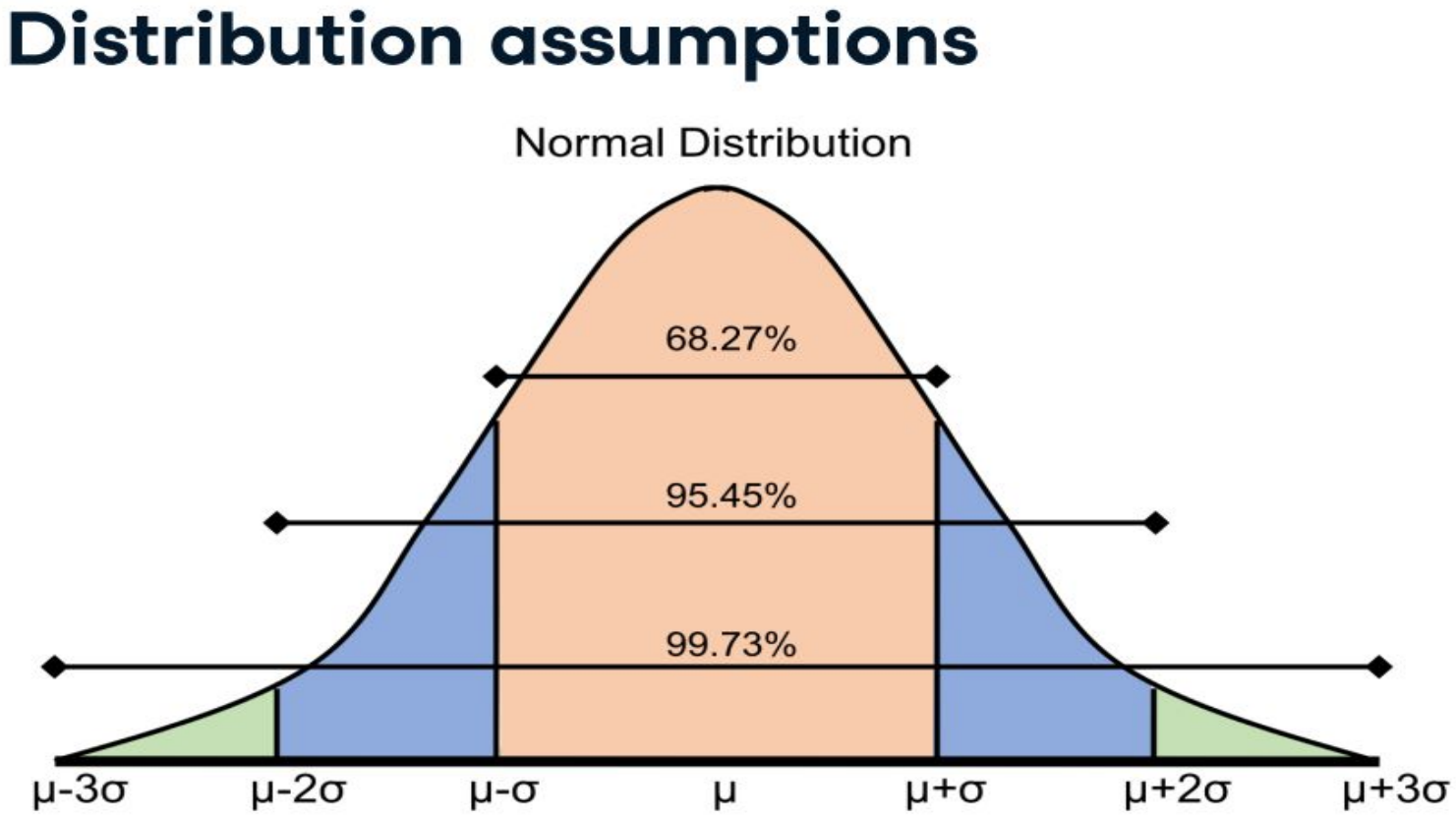
Work	No_of_children	Has_child
Yes	2	1
No	3	1
No	0	0

Binning numeric variables

AGE
21
29
35
33

AGE	transforms
21	20
29	20
35	30
33	30

Transformation



Why we need Scaling?

NO_Rooms	Length	Width	Price
2 =⅓ =0.133333	1000	300	1000000
3 =⅔ =0.	1780	400	2000000
4 =1	1600	500	5000000
1 =0	2000	300	3000000

Why we need scaling?

Real Life Datasets have many features with a wide range of values like for example let's consider the house price prediction dataset. It will have many features like no. of. bedrooms, square feet area of the house, etc.

As you can guess, the no. of bedrooms will vary between 1 and 5, but the square feet area will range from 500-2000. This is a huge difference in the range of both features.

Ways to do feature scaling

- Min-Max Scaling
- Normalization
- Standardization

Min Max Scaling

In min-max you will subtract the minimum value in the dataset with all the values and then divide this by the range of the dataset(maximum-minimum). In this case, your dataset will lie between 0 and 1 in all cases whereas in the previous case, it was between -1 and +1. Again, this technique is also prone to outliers.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalization

In scaling, you are changing the range of your data while in normalization you arere changing the shape of the distribution of your data.

$$X_{new} = \frac{X - X_{mean}}{X_{max} - X_{min}}$$

Standardization

In standardization, we calculate the z-value for each of the data points and replaces those with these values.

This will make sure that all the features are centred around the mean value with a standard deviation value of 1. This is the best to use if your feature is normally distributed like salary or age

$$X_{new} = \frac{X - X_{mean}}{\sigma}$$