Recap – Feature Engineering

Rina BUOY

AMERICAN UNIVERSITY OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

What is Feature Engineering ?

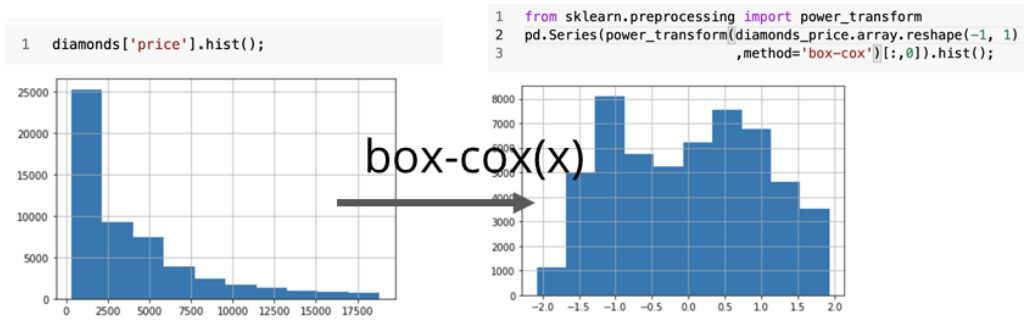
- Feature engineering is the process of transforming raw data into features that can be used to improve the performance of machine learning models.

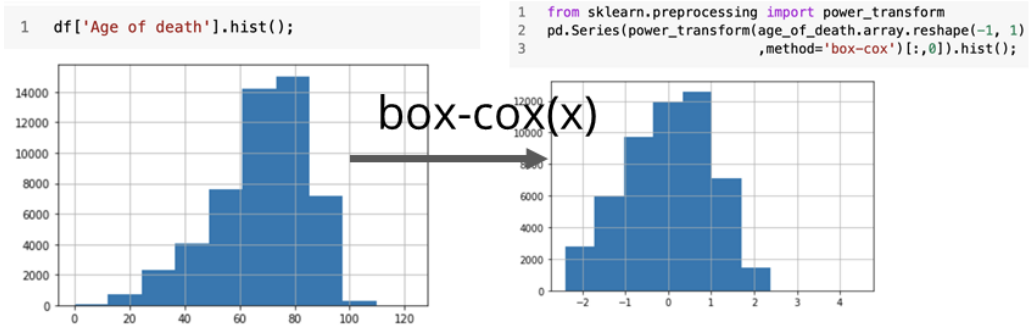
- It involves selecting, creating, or modifying features from the dataset to make them more suitable for the model, ultimately enhancing its predictive accuracy or interpretability.

- Feature engineering is a crucial step in the machine learning pipeline and can have a significant impact on the model's performance.

Feature Transformation

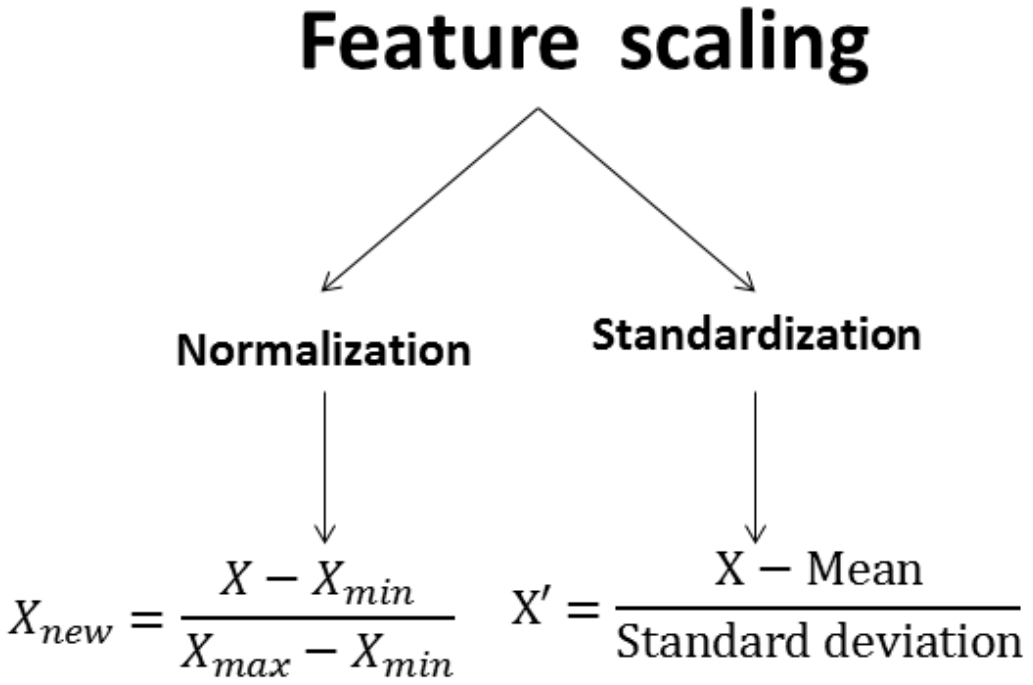
- Transforming features to make them more suitable for the model. This can include techniques like logarithmic transformation, polynomial transformation, or Box-Cox transformation to make the data more Gaussian-like.





Normalization and Scaling

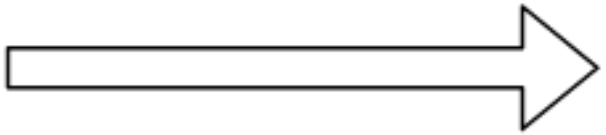
- Scaling features to a similar range or normalizing them to have a standard distribution. This is particularly important for algorithms sensitive to feature magnitudes, such as Ridge/Lasso algorithms.



Encoding Categorical Variables

Color
Red
Green
Blue

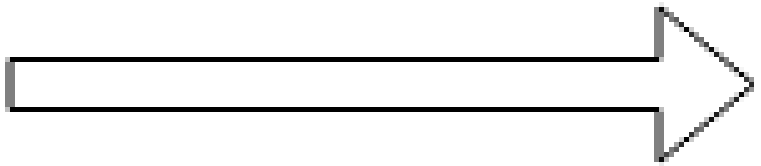
One-hot encoding



d1	d2	d3
1	0	0
0	1	0
0	0	1

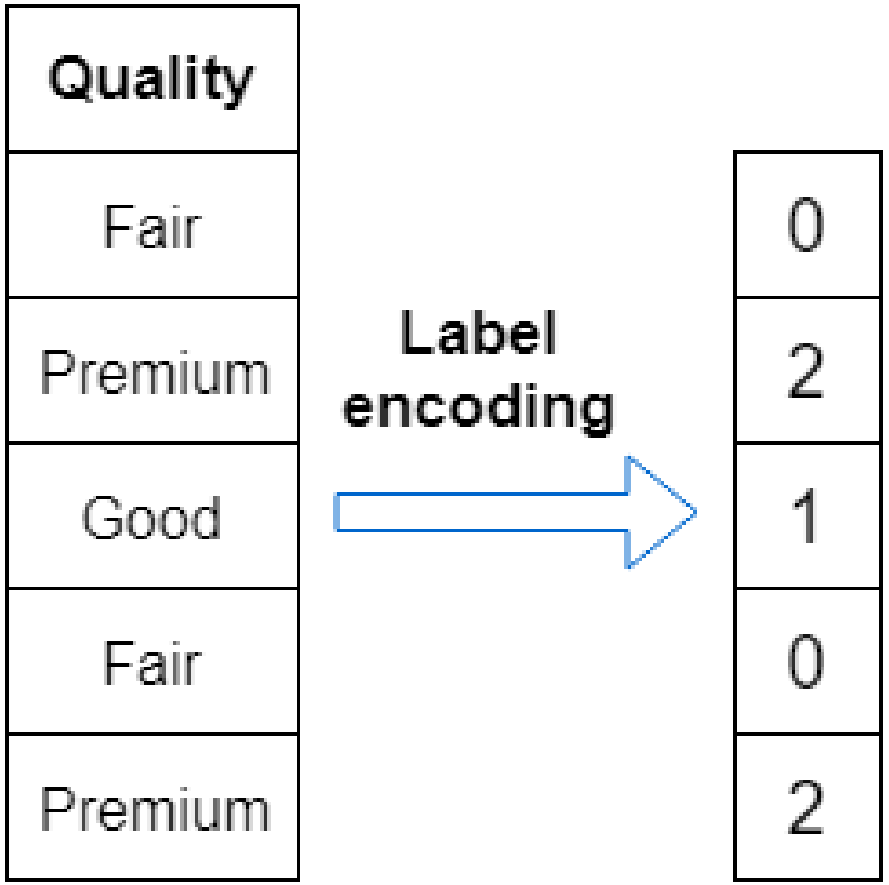
Encoding Categorical Variables

Color
Red
Green
Blue

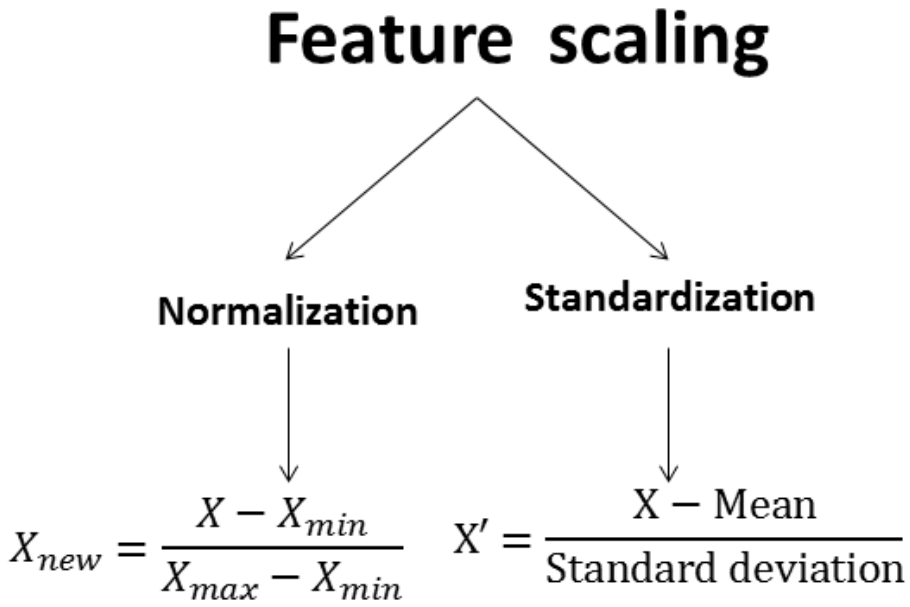


d1	d2
1	0
0	1
0	0

Encoding Categorical Variables



Training Statistics



Apply (DON'T RECALCULATE) on test data

Xmin, Xmax, Mean, Std on train data

Feature Interaction & Crossing

- Feature Interaction: Create new features by combining two or more existing features. For example, if you have features for "age" and "income," you could create a new feature representing the product of the two, which might capture some interaction between age and income level.

- Feature Crossing: Combine categorical features to create new features representing combinations of categories. For example, if you have features for "gender" and "age group," you could create a new feature representing combinations such as "male in the age group 20-30."

Feature Selection

- Choosing the most relevant features from the dataset to include in the model. This can involve methods like correlation analysis, feature importance ranking, or domain knowledge.