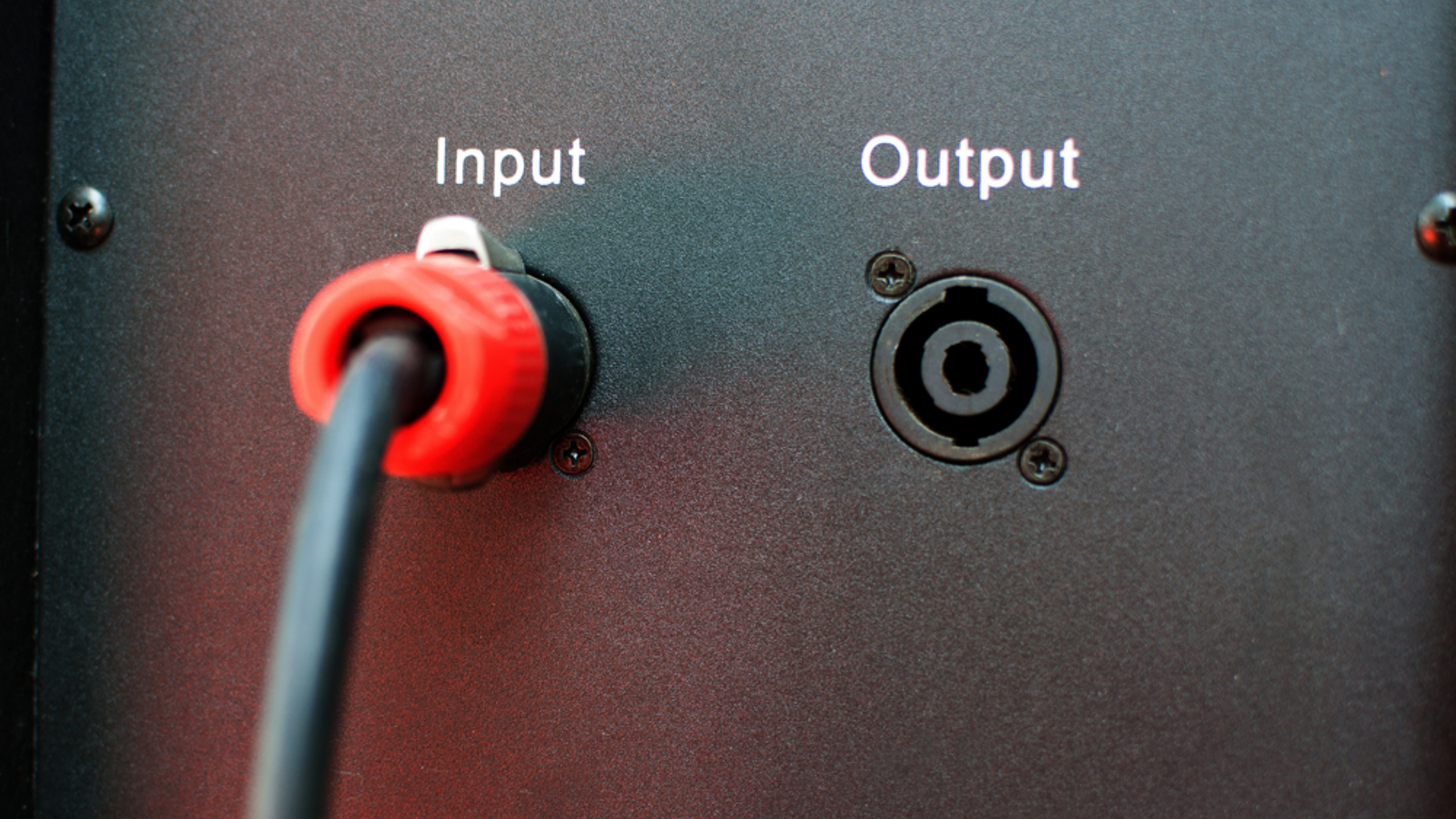
Performing Feature Normalization

Xavier Morera
HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera www.xaviermorera.com


Input
Output

Understanding Feature Normalization

Feature Normalization

Data preparation technique

Change values of numeric columns

- To a common scale

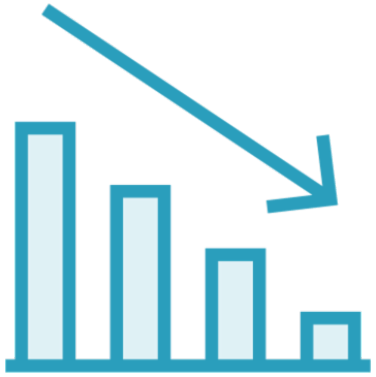
Without distorting differences in the ranges

Encoding to discrete values

Combine multiple features

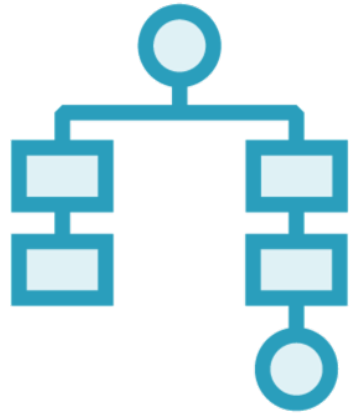
Benefits of Feature Normalization

Accuracy improvements



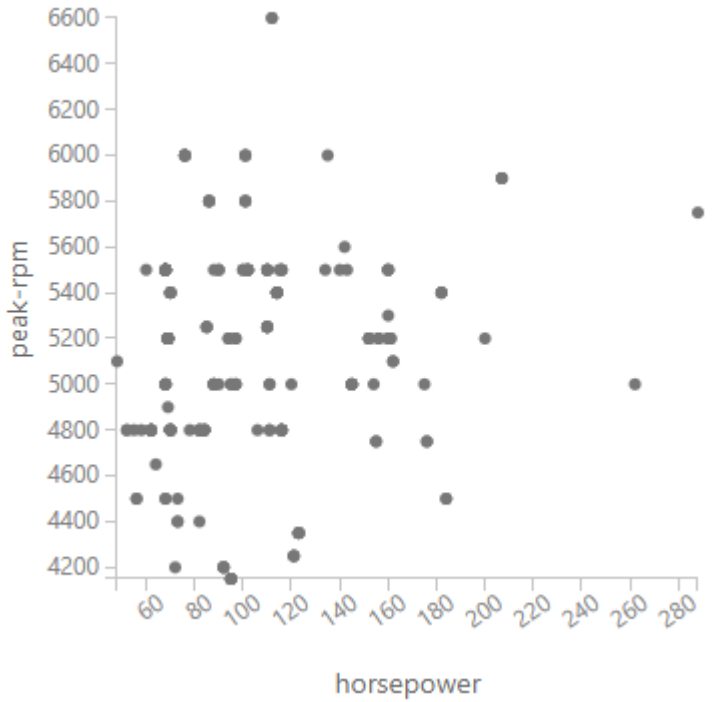
Overfitting risk reduction

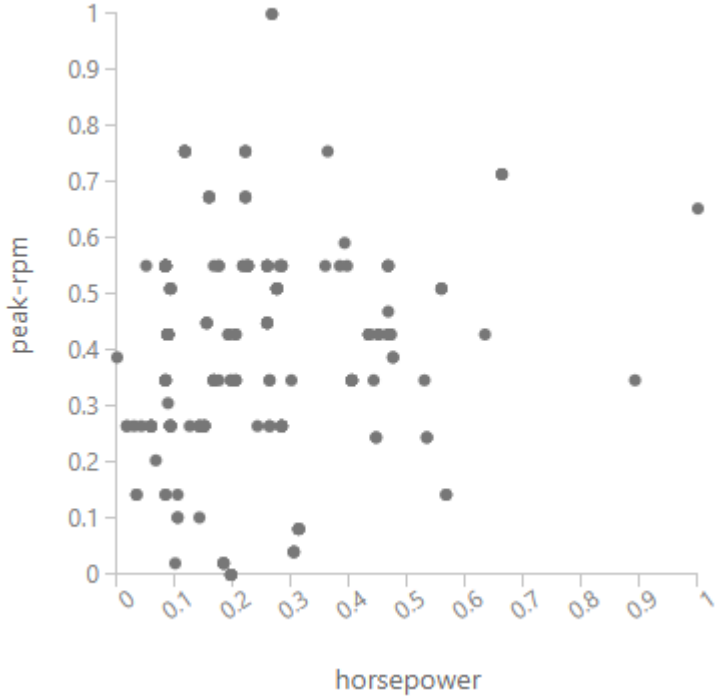
Speeds up in training



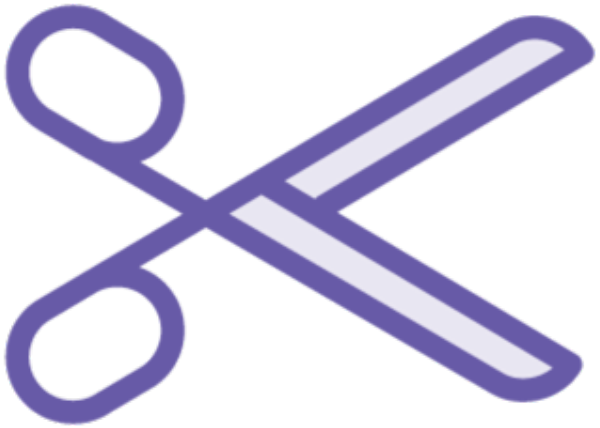
Improved data visualization

Feature Normalization





Clip Values



Detects outliers

- Clips or replaces values

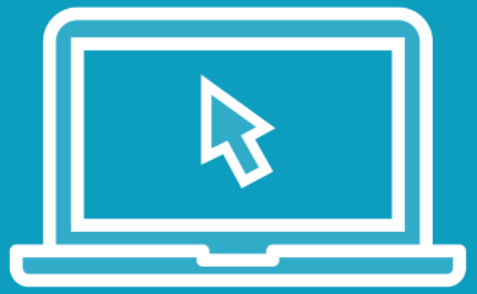
Set boundaries

- Upper and lower
- Constant or percentile

Substitute values

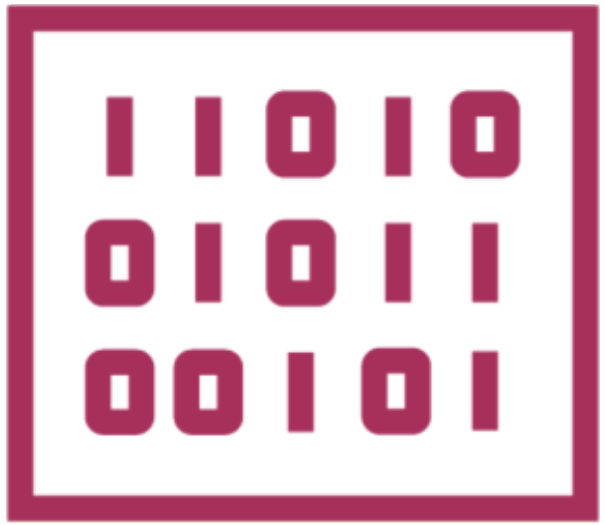
Generate new column

Demo



Detecting Outliers and Replacing Them Using the Clip Values Module

Group Data into Bins



Puts numerical data into bins

- Group numbers
- Change distribution of continuous data

Specify binning mode

- Manual or other methods, i.e. quantiles

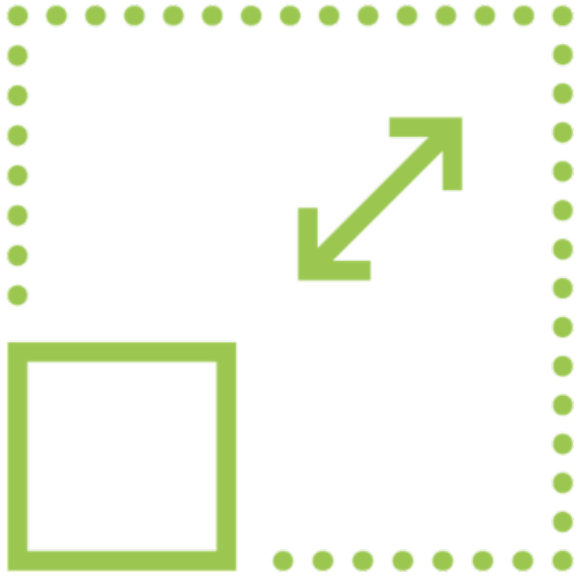
Binning on training data

- Same binning on testing and prediction

Demo

Binning Numeric Data Using the Group Data into Bins Module

Normalize Data



Rescales numeric data

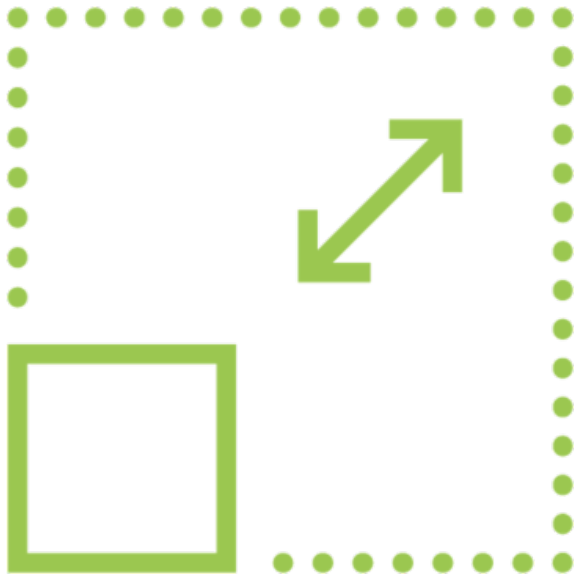
- To constrain dataset values
- To a standard range

Common scale

- Without distorting differences

May be required for some algorithms

Normalize Data



- Zscore
- MinMax
- Logistic
- LogNormal
- TanhZ

Demo

Rescaling Numeric Data Using the Normalize Data Module

Principal Component Analysis

Computes a set of features

- Reduced dimensionality
- For more efficient learning

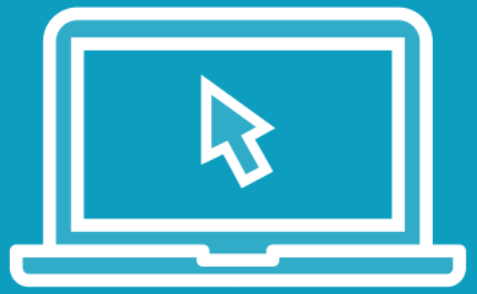
Reduce large set of variables

- While retaining most of the information

Combine features

- Provide better information
- Than if used separately

Demo



Reducing Dimensionality Using the Principal Component Analysis Module

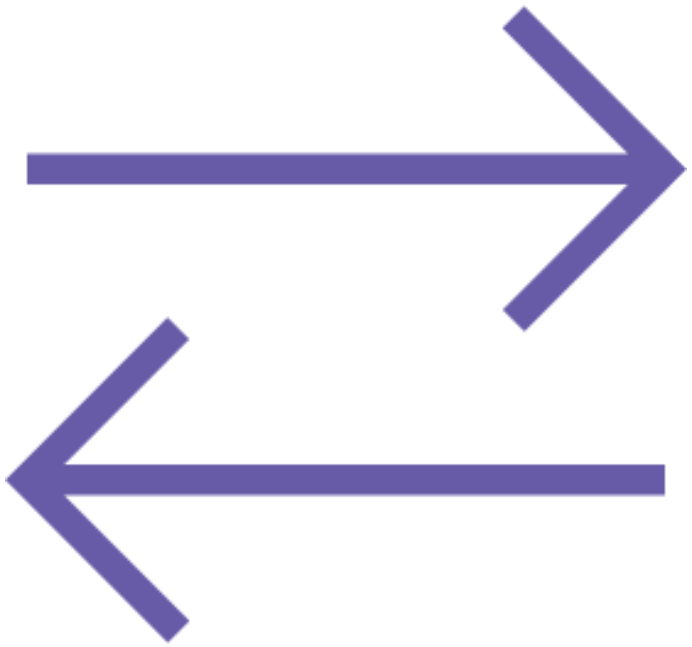
Features

$$\begin{bmatrix} 5.0 & 3.6 & 1.4 \\ 4.9 & 3.0 & 1.4 \\ 4.7 & 3.2 & 1.3 \\ 4.6 & 3.1 & 1.5 \\ 5.0 & 3.6 & 1.4 \end{bmatrix}$$

Features

Audi	3.6	1.4
BMW	3.0	1.4
MB	3.2	1.3
Tesla	3.1	1.5
LR	3.6	1.4

Encoding Features



Convert a categorical value

- Into a numerical value

Able to perform operations

One-hot encoding

- Vectors with 0 and 1
- Number of vectors depends on categories

Encoding Features

make	hp	pk rpm	price
audi	0,10	5,50	13,90
bmw	0,10	4,20	16,40
dodge	0,06	5,00	5,50

audi	bmw	dodge	hp	pk rpm	price
1	0	0	0,10	5,50	13,90
0	1	0	0,10	4,20	16,40
0	0	1	0,06	5,00	5,50

Demo

Encoding Features in the Automobile Price Data

Takeaway

What is feature normalization?

Modules

- Clip values
- Group values into bins
- Normalization
- Principal component analysis
- Encoding features