Chapter 2: Data Preparation

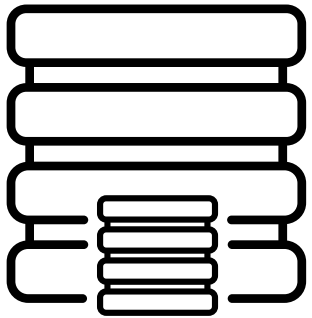
2.1 Data Exploration
2.2 Feature Extraction
2.3 Input Transformations
2.4 Feature Selection
2.5 Variable Clustering (Self-Study)
2.6 Best Practices

Chapter 2: Data Preparation

2.1 Data Exploration
2.2 Feature Extraction
2.3 Input Transformations
2.4 Feature Selection
2.5 Variable Clustering (Self-Study)
2.6 Best Practices

Overview of Data Preprocessing

The more data, the better!



High Cardinality

Missing Values

Non-linearities

Outliers

Redundant Inputs

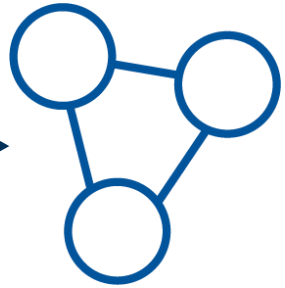
Misscaled Variables

Unstructured Data

Sparseness

Irrelevant Inputs

Messy data yields messy analytical models.



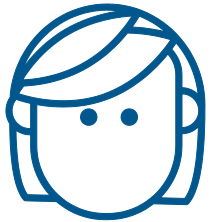
Essential Data Tasks

Essential Data Tasks

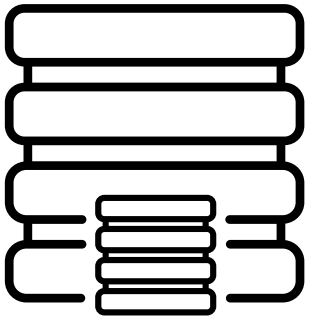
- Gather the data.
- Explore the data.
- Divide the data.
- Address rare events.
- Manage missing values.
- Replace incorrect values.
- Add unstructured data.
- Extract features.
- Manage extreme or unusual values.
- Select useful inputs.

Exploring the Data

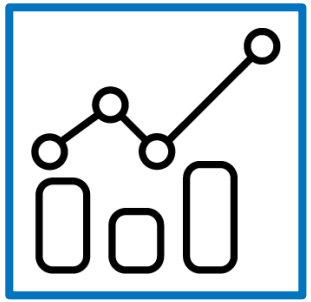
Get to know your data.



Ask lots of questions.



Use graphical and numerical methods.

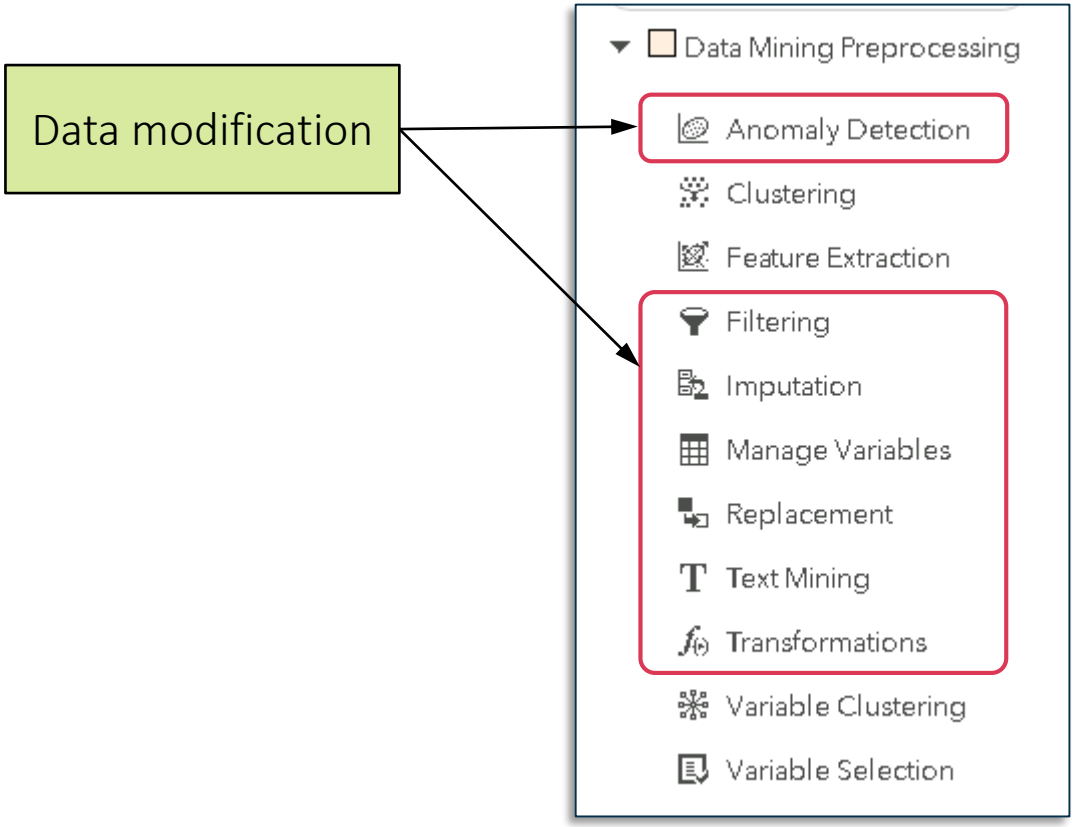


- outliers
- minimums
- maximums
- percent missing
- means
- ranges
- standard deviations
- distributions
- …

Exploring Source Data

In this demonstration, you use Data Exploration node in SAS Model Studio to assay and explore a data source.

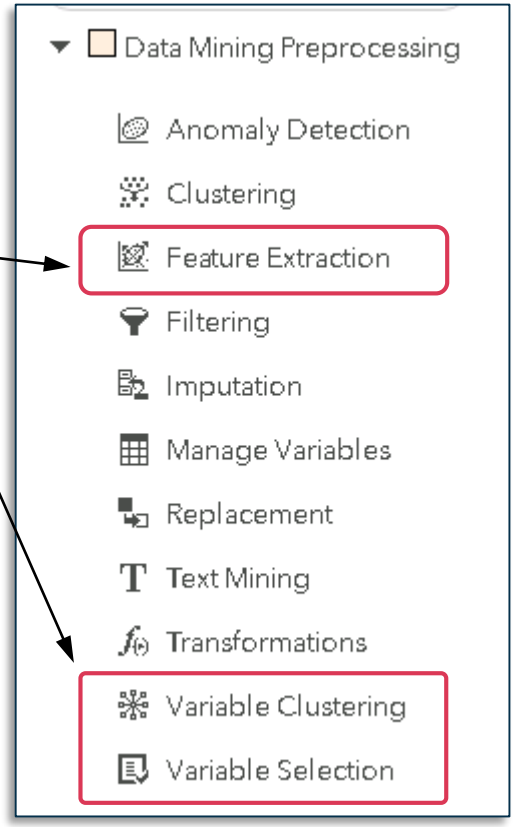
Data Preprocessing with Model Studio



Data Preprocessing with Model Studio

Data modification

Dimension reduction

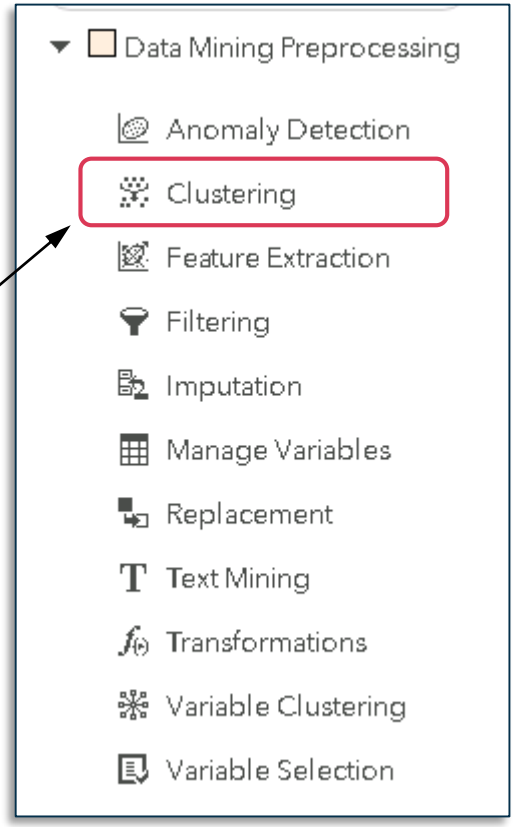


Data Preprocessing with Model Studio

Data modification

Dimension reduction

Unsupervised learning



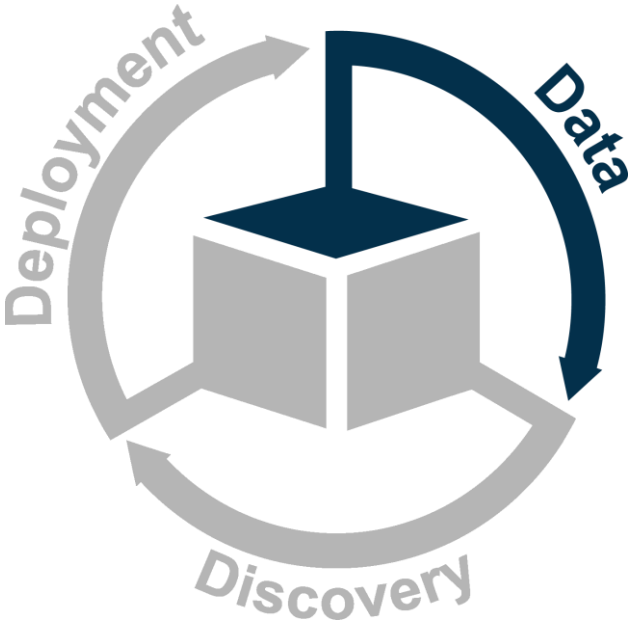
Modifying and Correcting Source Data

In this demonstration, you use the Data tab and Replacement node to modify a data source.

Chapter 2: Data Preparation

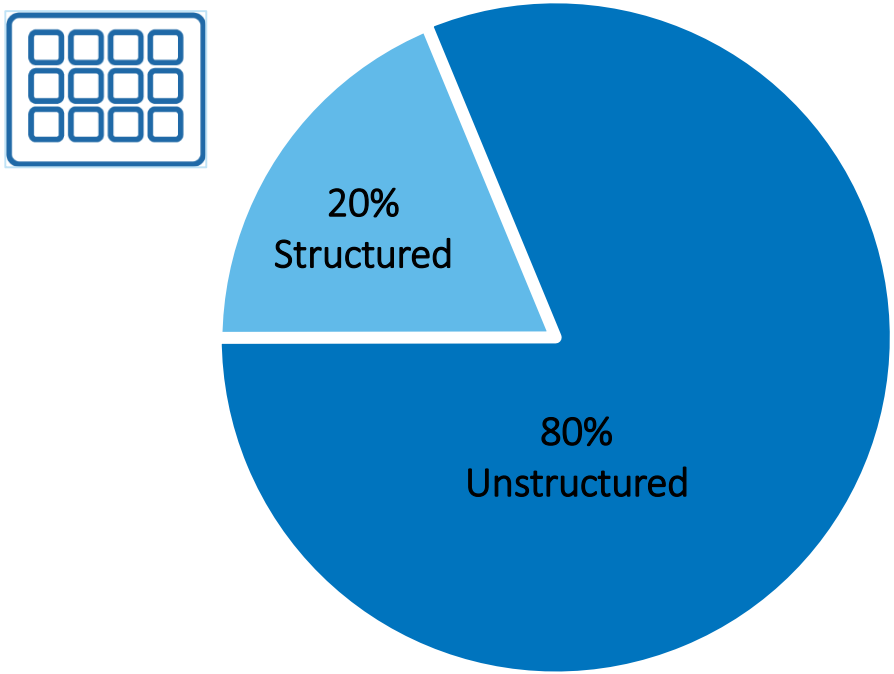
2.1 Data Exploration
2.2 Feature Extraction
2.3 Input Transformations
2.4 Feature Selection
2.5 Variable Clustering (Self-Study)
2.6 Best Practices

Essential Data Tasks



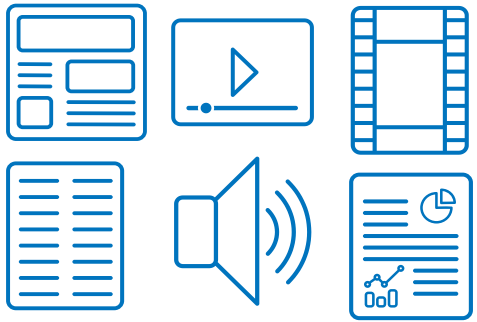
- Divide the data.
- Address rare events.
- Manage missing values.
- Add unstructured data.
- Extract features.
- Handle extreme or unusual values.
- Select useful inputs.

Text Mining



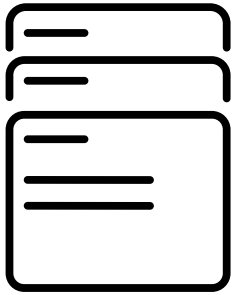
Unlocking the 80%!

Text mining helps extract meanings, patterns, and structure hidden in unstructured textual data.



Text Mining Feature Extraction

1. Text parsing



	Term 1	Term 2	Term 3	. . .
Doc 1				
Doc 2				
Doc 3				
. . .				

2. Transformation

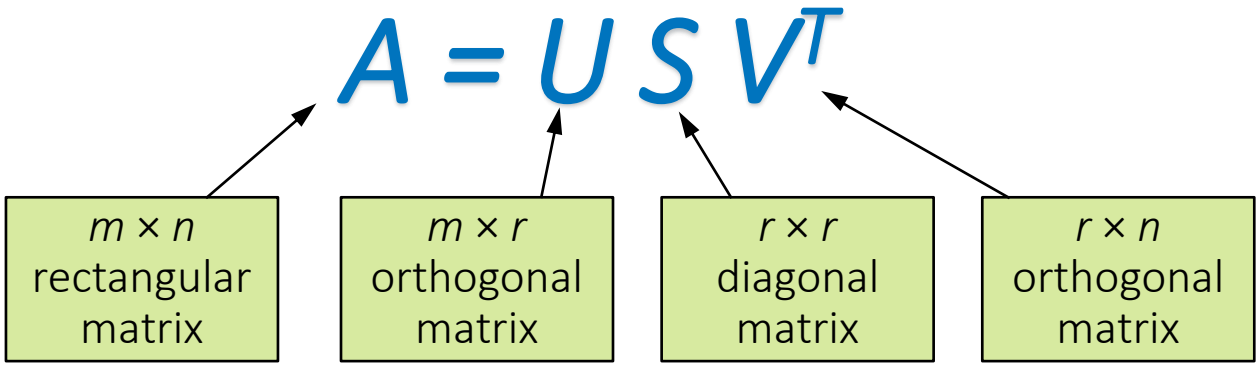
Structured Data

Inputs from Unstructured Data

ID	Var 1	Var 2	Var 3	. . .	SVD 1	SVD 2	SVD 3	. . .	Target
. . .									

Singular Value Decomposition (SVD)

- Singular value decomposition (SVD) projects the high-dimensional document and term spaces into a lower-dimension space.
- Singular value decomposition is a method of decomposing a matrix into three other matrices:



- The singular values can be thought of as providing a measure of importance used to decide how many dimensions to keep.

Adding Text Mining Features

In this demonstration, you create new features using the Text Mining node.

Chapter 2: Data Preparation

Essential Data Tasks

- Divide the data.
- Address rare events.
- Manage missing values.
- Add unstructured data.
- Extract features.
- Handle extreme or unusual values.
- Select useful inputs.

Input Transformations

Transformations stabilize variances, remove nonlinearity, and correct non-normality in inputs to improve the fit of the model.

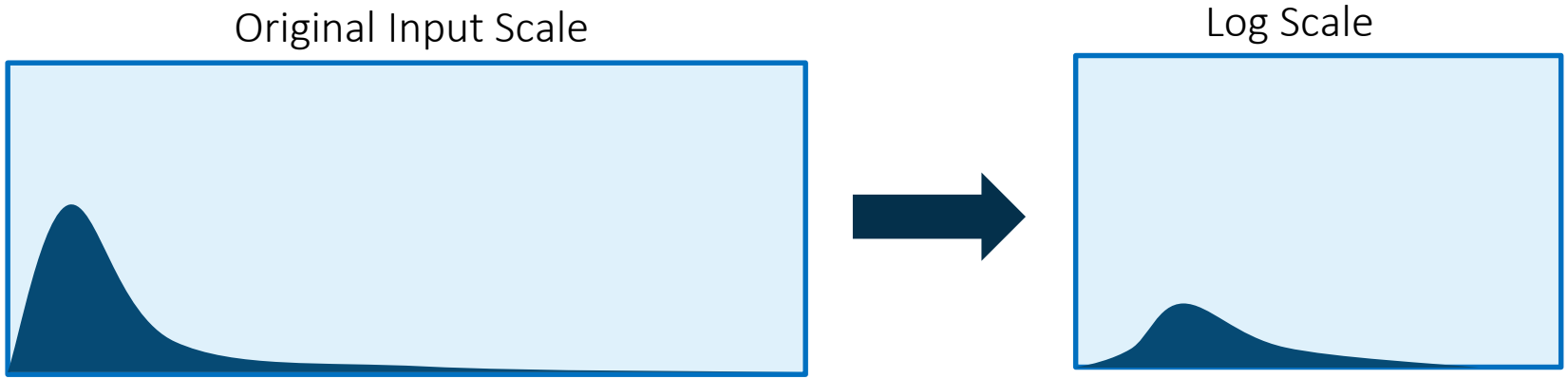
Mathematical Functions

- Centering
- Exponential
- Inverse
- Log
- Range
- Square
- Square root
- Standardize

Binning

- Bucket
- Quantile
- Tree-based binning

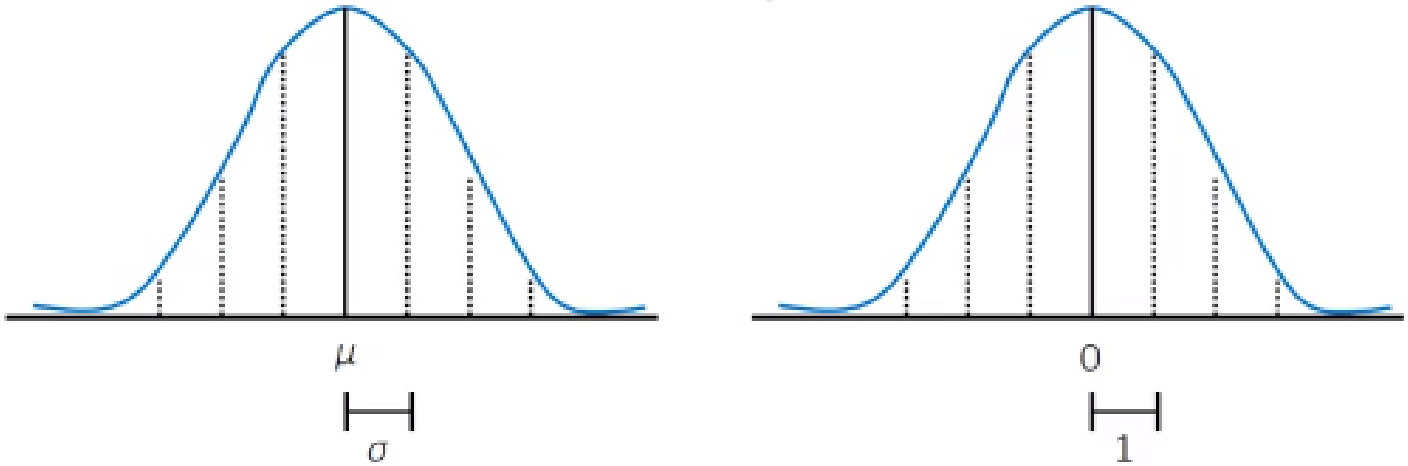
Transforming Inputs: Mathematical Functions



Transforming Inputs: Mathematical Functions

Standardize

$$z = \frac{x - \mu}{\sigma}$$



Transforming Inputs: Binning

$$0 \leq \text{Age} < \infty$$

Bin	Age Range
1	(0, 20]
2	(20, 50]
3	(50, 70]
4	(70 and greater)

2.01 Multiple Choice Poll

Why bin an input?

a. It can reduce the effects of an outlier.
b. It can classify missing values (into a category or bin).
c. It can generate multiple effects.
d. all of the above

2.01 Multiple Choice Poll – Correct Answer

Why bin an input?

a. It can reduce the effects of an outlier.
b. It can classify missing values (into a category or bin).
c. It can generate multiple effects.
d. all of the above

Transforming Inputs

In this demonstration, you use the Transformations node to apply a numerical transformation to input variables.

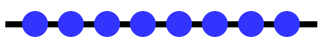
Chapter 2: Data Preparation

2.1 Data Exploration
2.2 Feature Extraction
2.3 Input Transformations
2.4 Feature Selection
2.5 Variable Clustering (Self-Study)
2.6 Best Practices

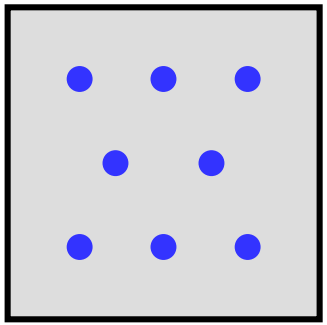
Essential Data Tasks

- Divide the data.
- Address rare events.
- Manage missing values.
- Add unstructured data.
- Extract features.
- Handle extreme or unusual values.
- Select useful inputs.

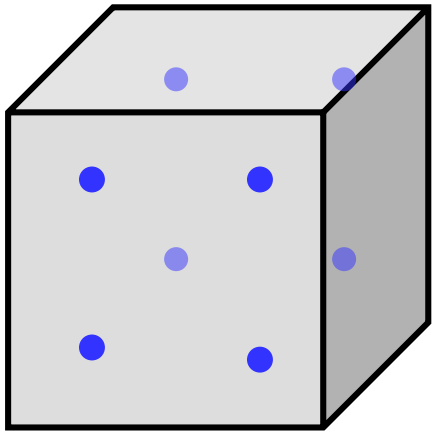
The Curse of Dimensionality



1–D

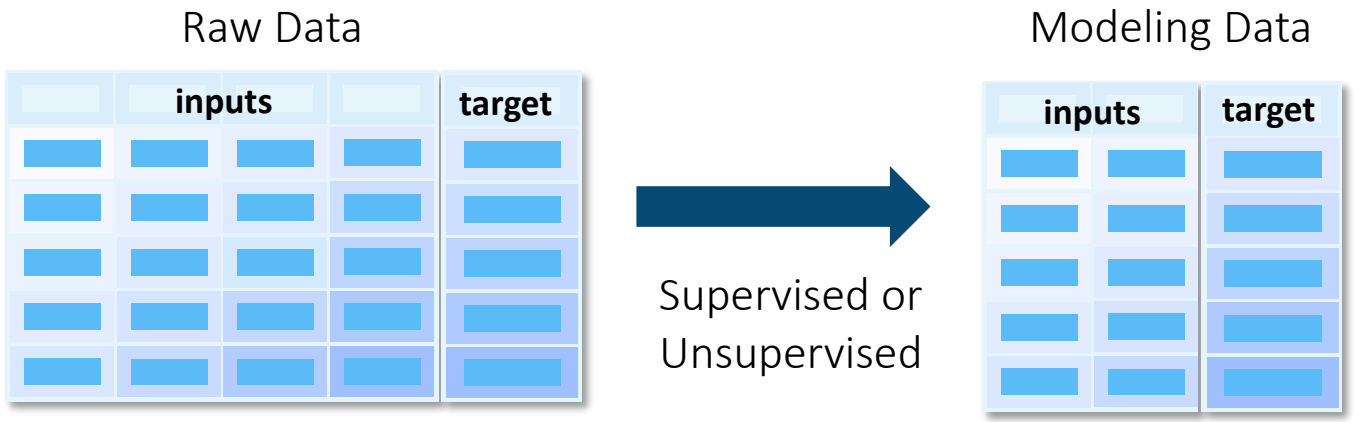


2–D



3–D

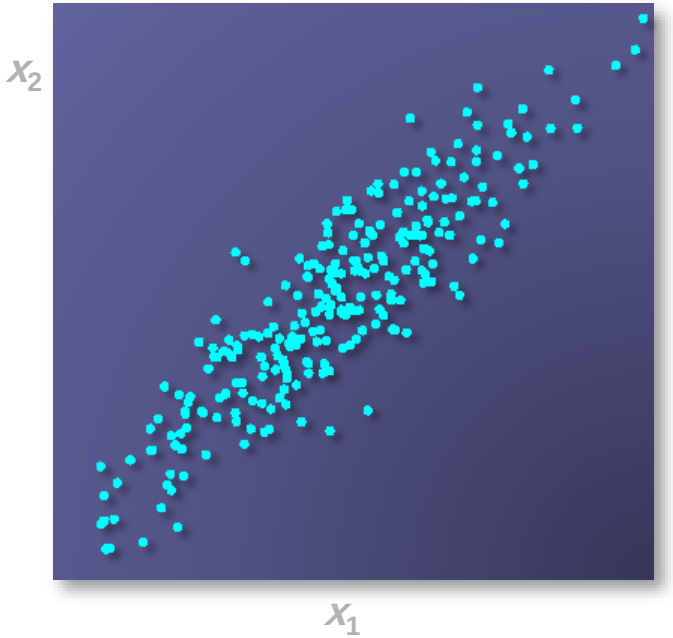
Feature Selection



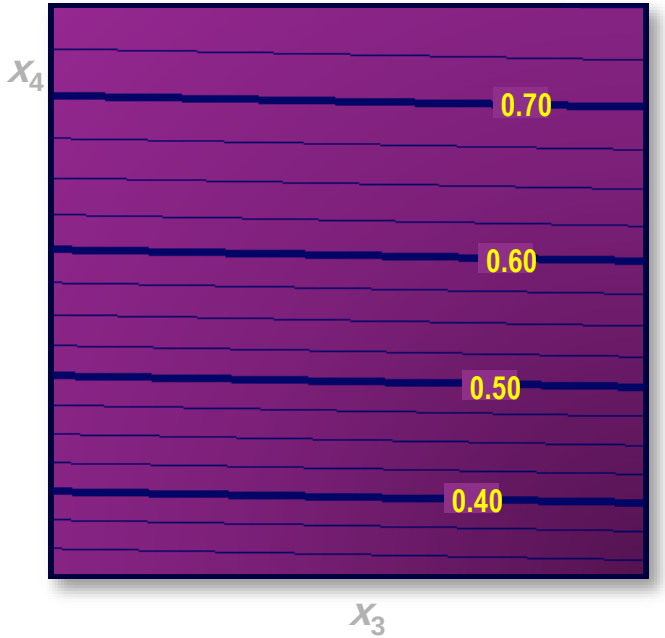
Using all available inputs usually leads to a model that does not generalize well to new data.

Feature Selection Strategies

Redundancy

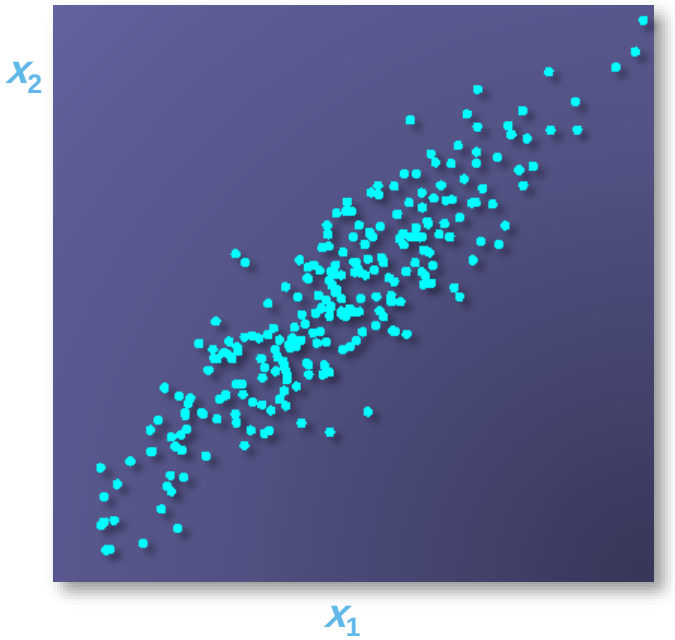


Irrelevancy

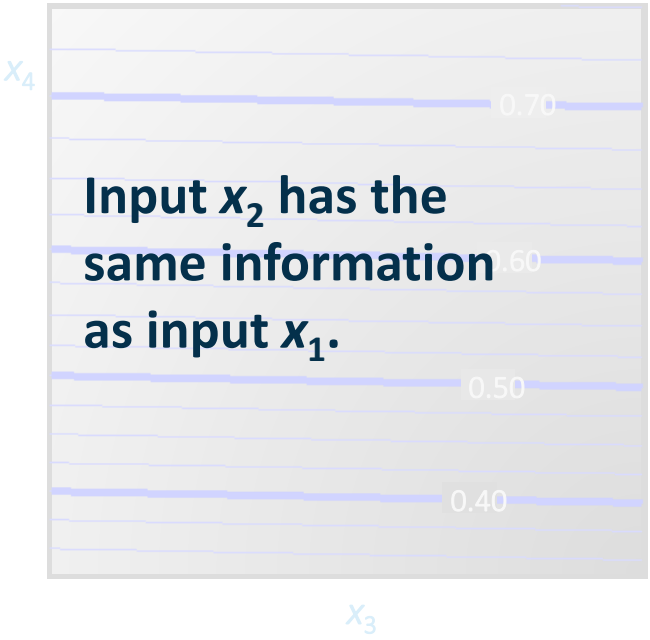


Unsupervised Selection

Redundancy



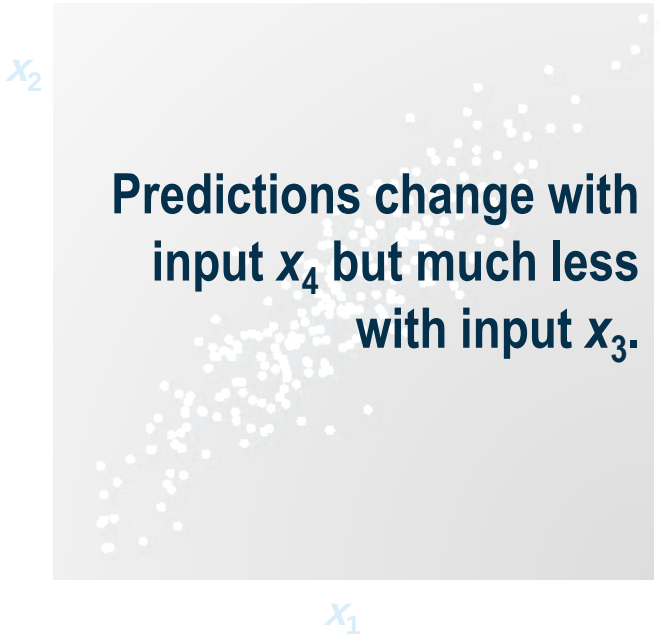
Irrelevancy



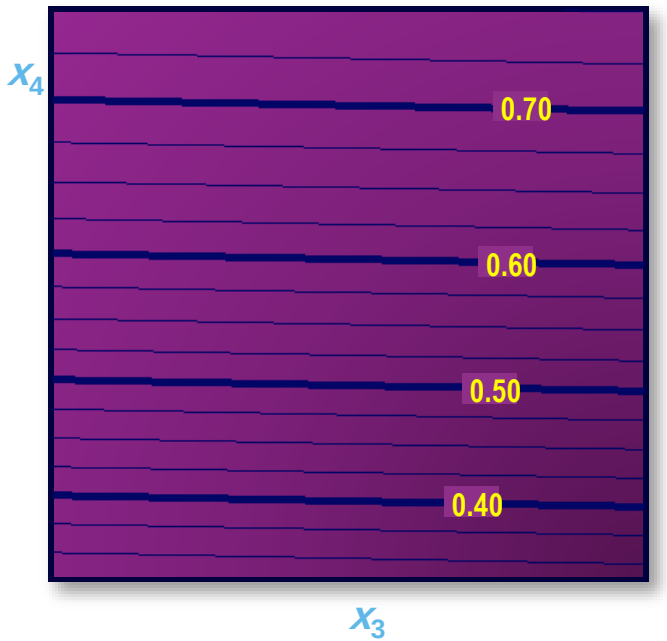
Example: x_1 is household income and x_2 is home value.

Supervised Selection

Redundancy



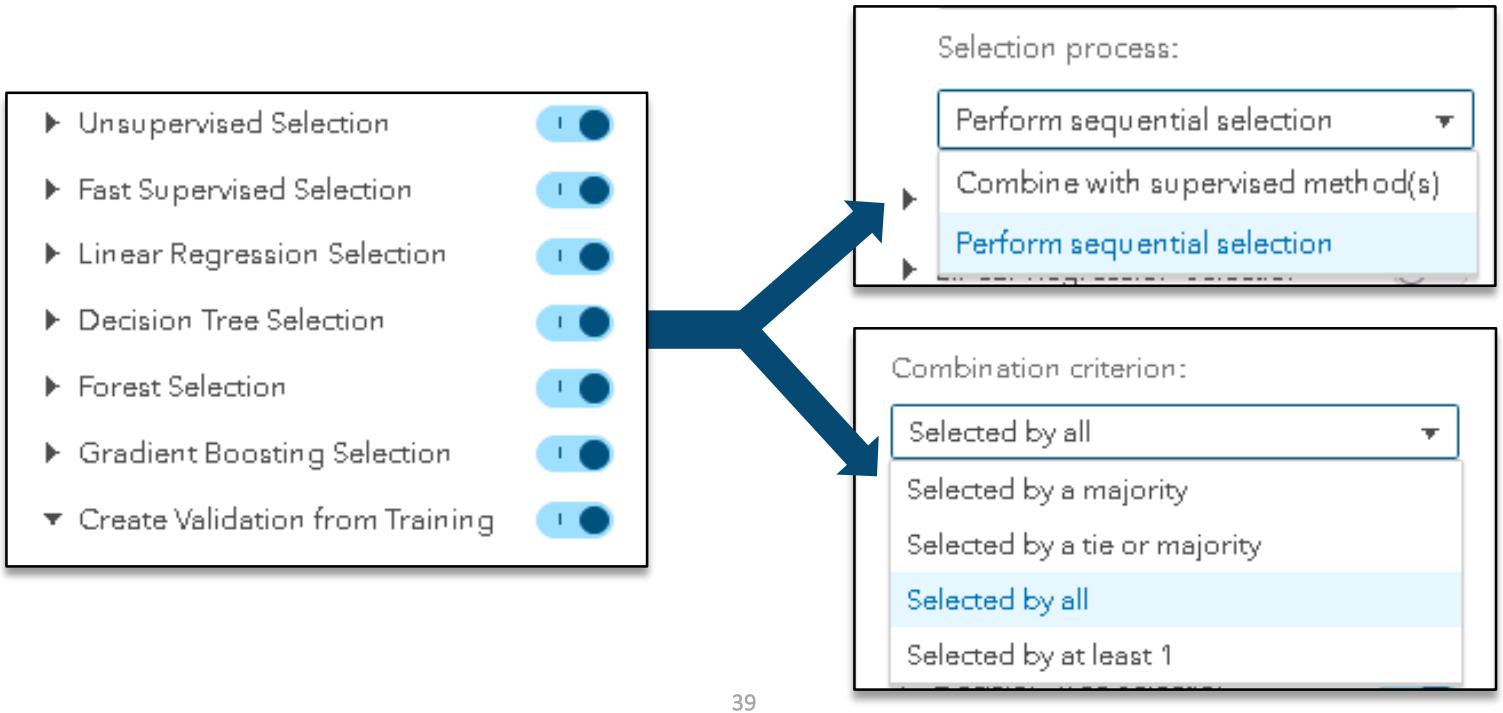
Irrelevancy



Example: Target is the response to direct mail solicitation, x_3 is religious affiliation, and x_4 is the response to previous solicitations.

Feature Selection in Model Studio

The Variable Selection node performs unsupervised and several supervised methods of variable selection to reduce the number of inputs.



Selecting Features

In this demonstration, you use the Variable Selection node to reduce the number of inputs for modeling.

Saving a Pipeline to the Exchange

In this demonstration, you save the Starter Template pipeline to the Exchange, where it will be available for other users.

Chapter 2: Data Preparation

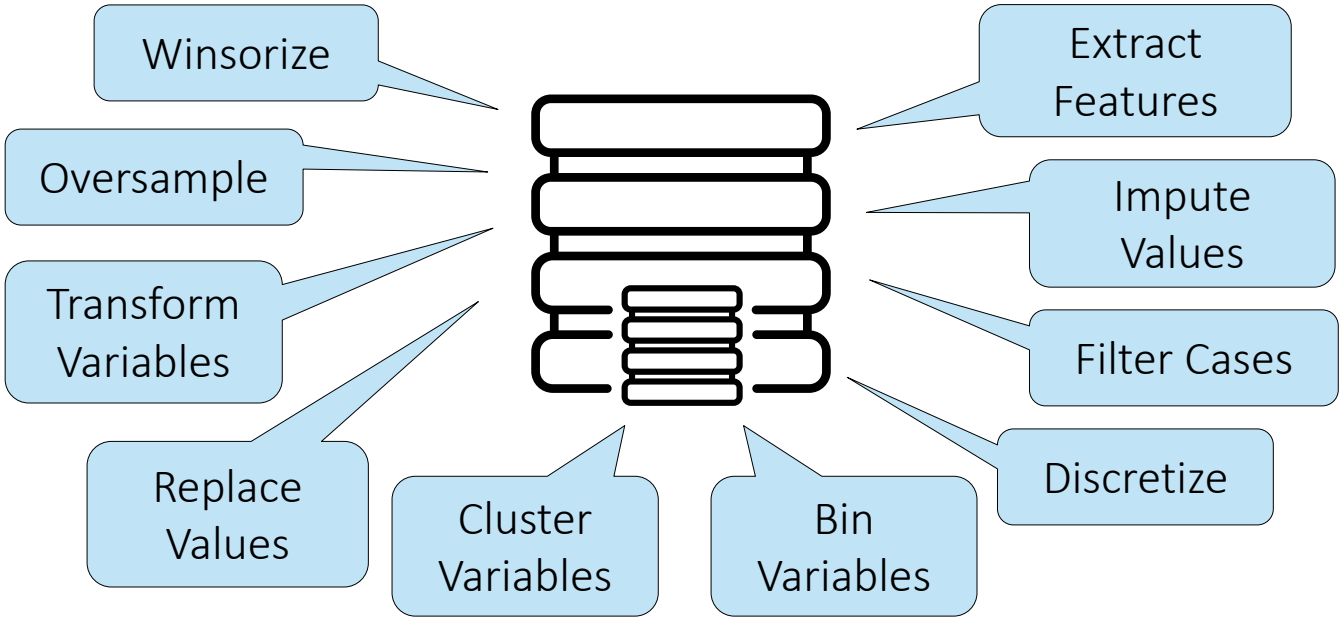
2.1 Data Exploration
2.2 Feature Extraction
2.3 Input Transformations
2.4 Feature Selection
2.5 Variable Clustering (Self-Study)
2.6 Best Practices

Chapter 2: Data Preparation

Data Preparation Best Practices

There is no single recipe!

There is no linear process flow!



Essential Discovery Tasks and Selecting an Algorithm

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize the complexity of the model.
- Regularize and tune the hyperparameters of the model.
- Build ensemble models.

Essential Discovery Tasks and Selecting an Algorithm

Essential Discovery Tasks and Selecting an Algorithm

1. What is the size and nature of your data?
2. What are you trying to achieve with your model?
3. How accurate does your model need to be?
4. How much time do you have to train your model?
5. Does your model have automatic hyperparameter tuning capability?

Open: Comparison of Modeling Algorithms pdf