Machine Learning in Business Decision Making

Dr. Waleed M.Ead

waleedead@bsu.edu.eg

Chapter 1: Introduction

1.1 Machine Learning in Business Decision Making

1.2 Essentials of Supervised Prediction

1.3 Introduction to SAS Viya

Chapter 1: Introduction

1.1 Machine Learning in Business Decision Making

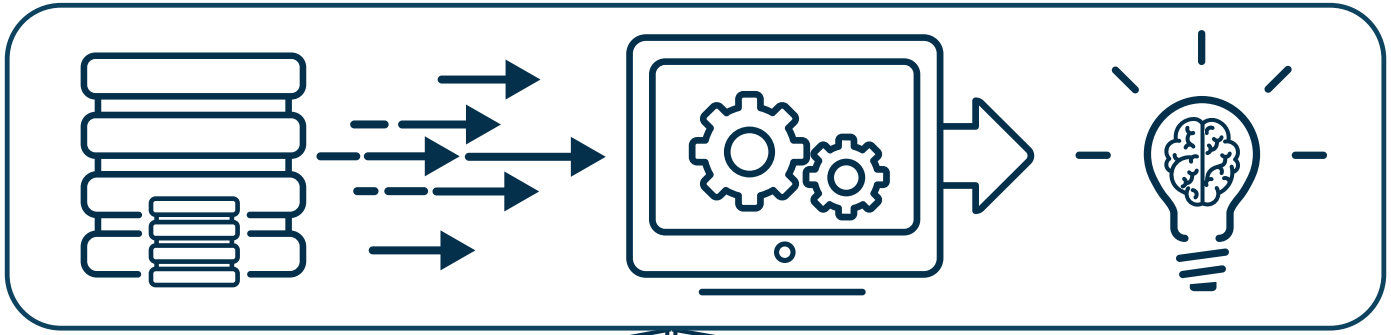
1.2 Essentials of Supervised Prediction

1.3 Introduction to SAS Viya

> “In the new world, it is not the big fish which eats the small fish, it’s the fast fish which eats the slow fish.”

Klaus Schwab
Founder and Executive Chairman
World Economic Forum

Machine Learning



Automate

Provide automation to the model building process by minimizing human intervention.

Customize

Build powerful models using state-of-the-art algorithms from SAS in conjunction with open source tools.

Accelerate

Fast response time for sophisticated analytics applied to data of any size or complexity.

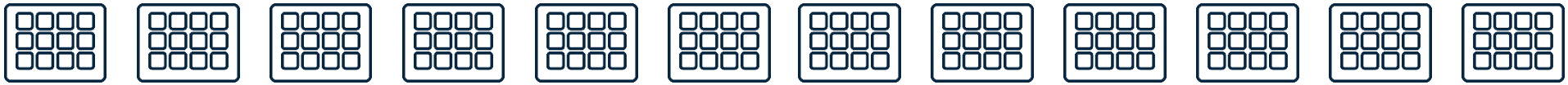
Today's Business Challenges

Fraud

Targeted Marketing

Financial Risk

Churn



Today's Tools

Fraud

Targeted Marketing

Financial Risk

Churn

Today's Tools

Fraud

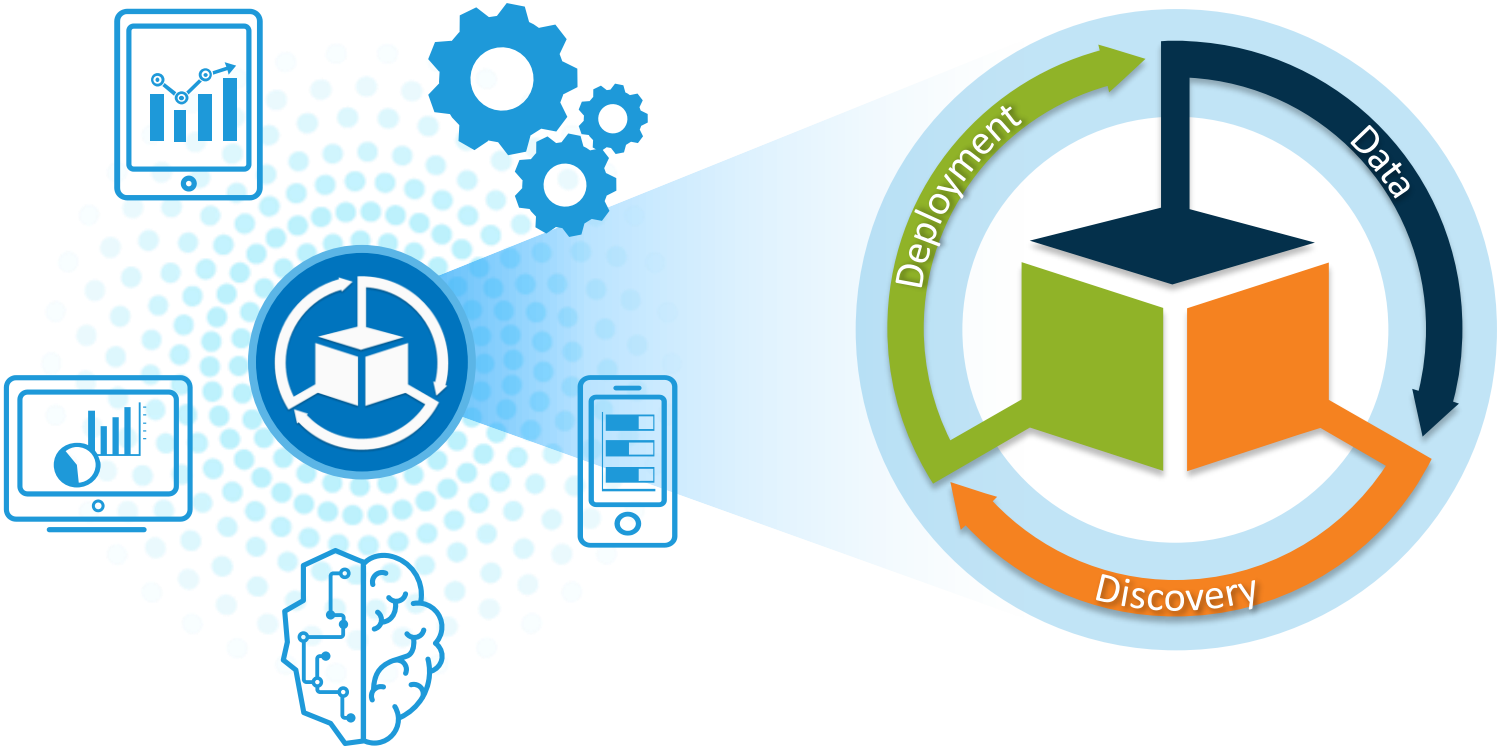
Targeted Marketing

Financial Risk

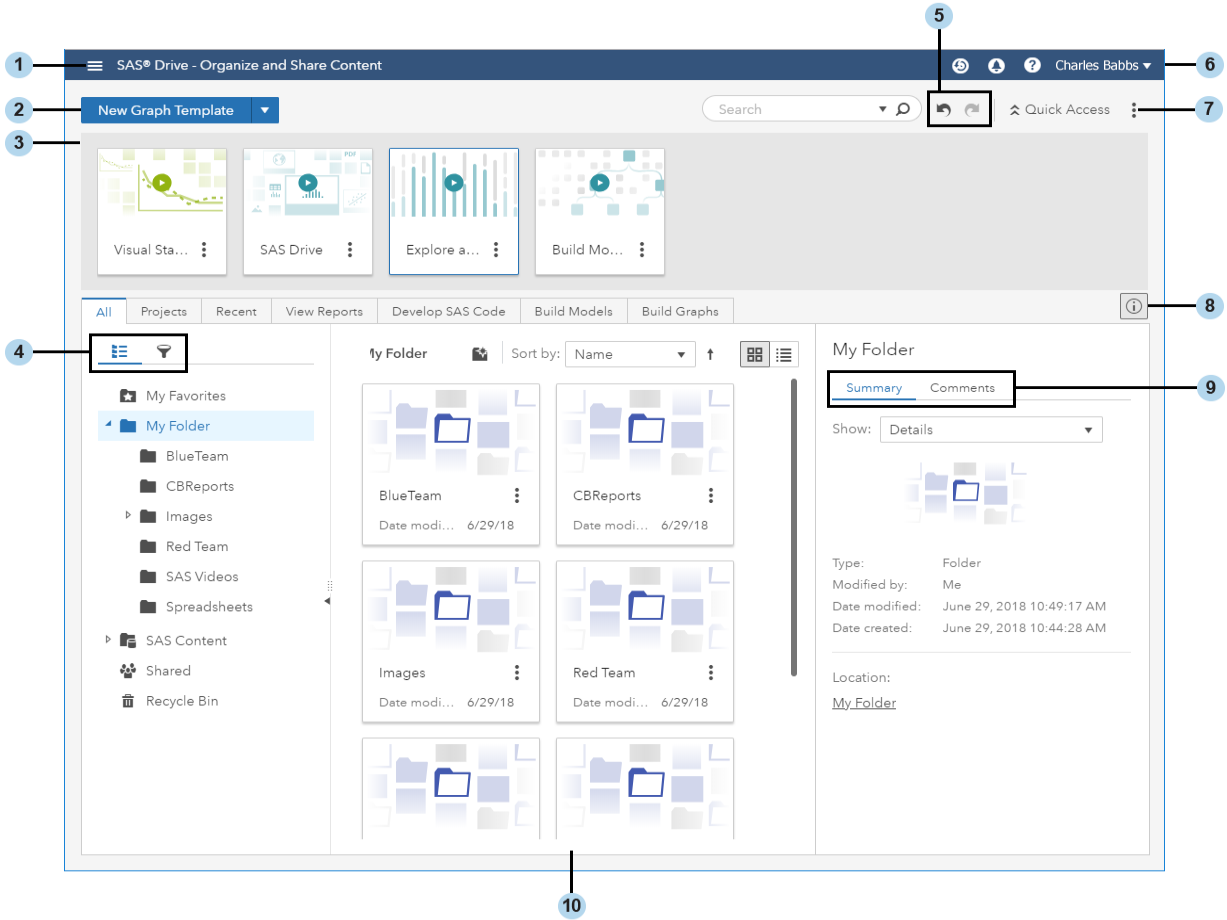
Churn

SAS® Viya®

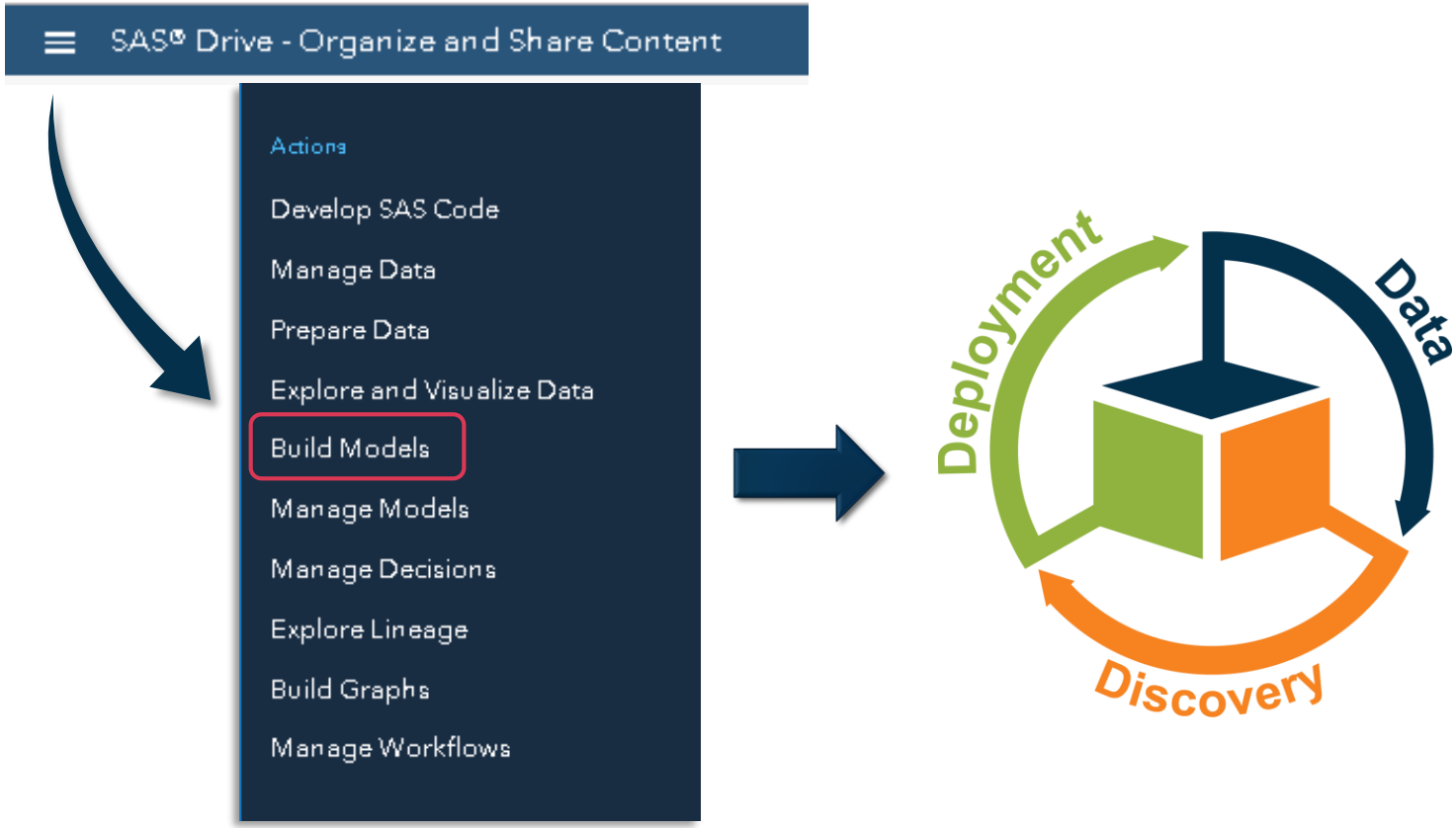
The Analytics Life Cycle



SAS Drive

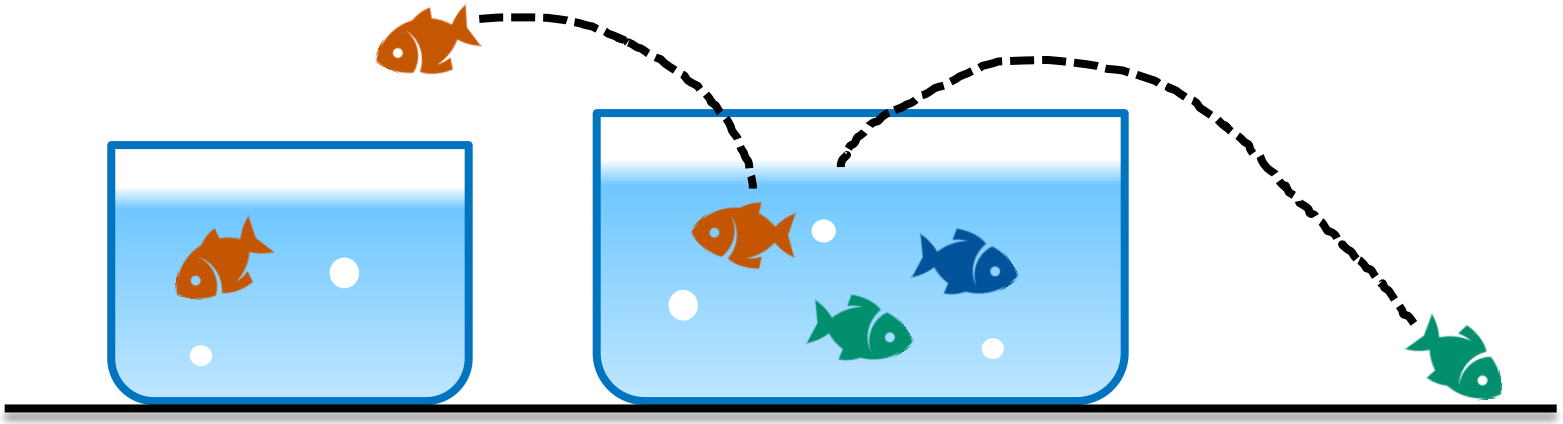


Common Interface for Entire Analytics Life Cycle



Business Challenge: Customer Churn

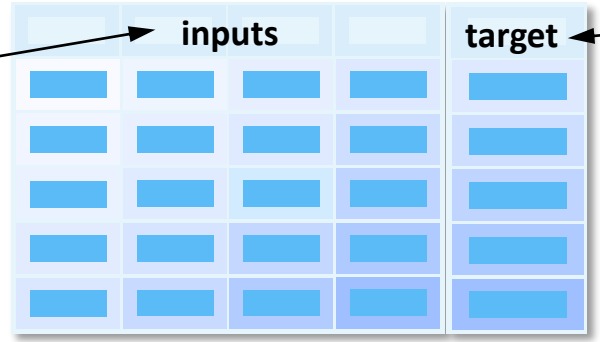
Customer churn, also known as *customer attrition*, is when an existing customer, subscriber, user, or any return client stops doing business or ends the relationship with a company.



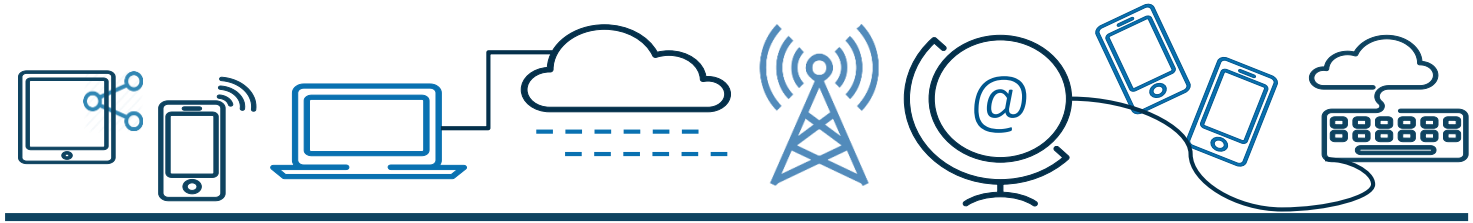
Customer Churn Scenario: Analysis Goal

A fictitious telecommunications company seeks to determine which customers might be likely to churn.

Demographics, customer care, and prior usage of services

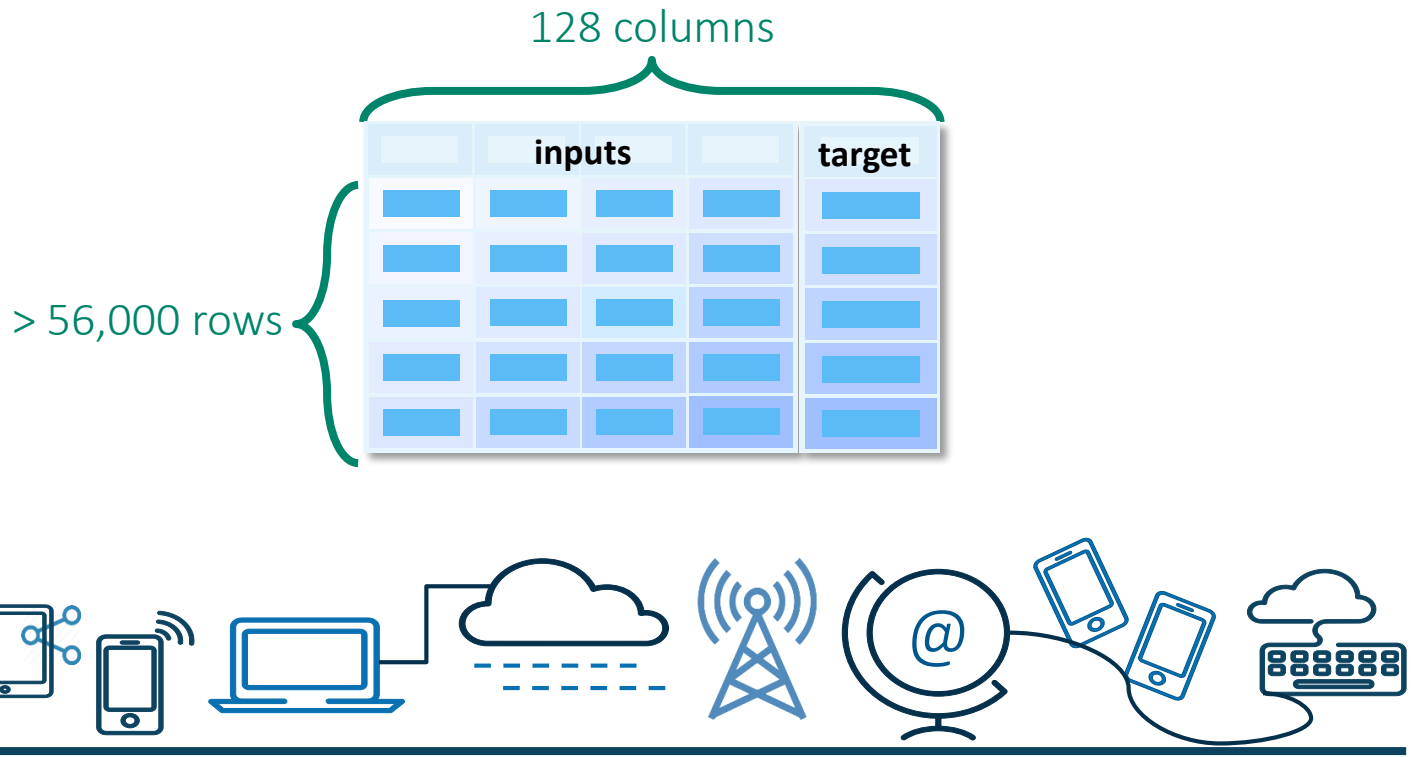


Did the customer churn in the past 10 months?

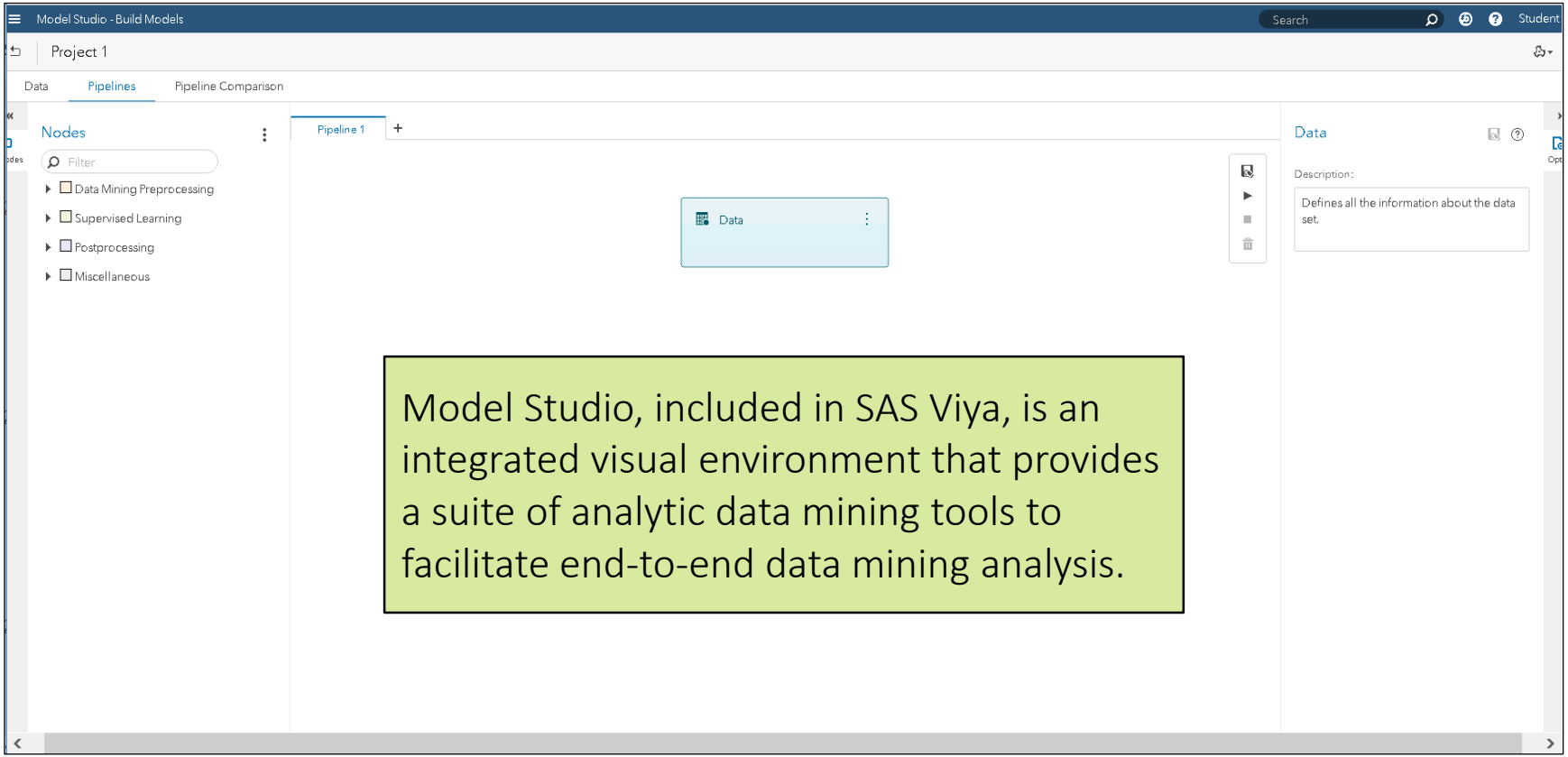


Customer Churn Scenario: Analysis Data

Raw Data: commsdata.sas7bdat



Model Studio



Model Studio, included in SAS Viya, is an integrated visual environment that provides a suite of analytic data mining tools to facilitate end-to-end data mining analysis.

Creating a Project and Loading Data

In this demonstration, you create a new project in Model Studio based on the **commsdata** data set.

Chapter 1: Introduction

1.1 Machine Learning in Business Decision Making

1.2 Essentials of Supervised Prediction

1.3 Introduction to SAS Viya

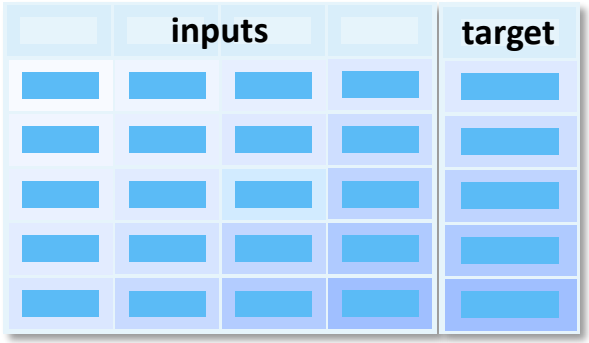
Predictive Model

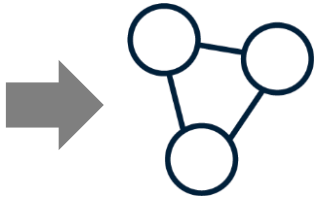
Predictive modeling

Supervised prediction

Supervised learning

Training Data





a concise representation of the input and target association

Predictive Model

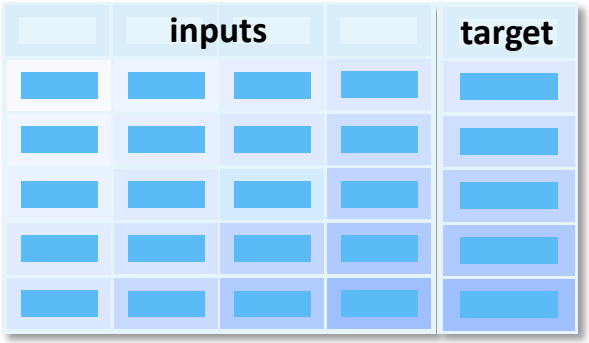
predictors

features

explanatory variables

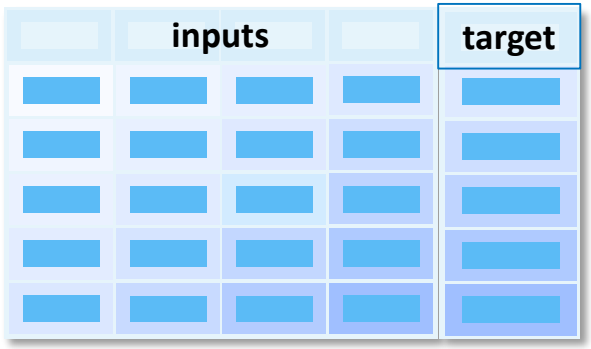
independent variables

Training Data



Predictive Model

Training Data



response

outcome

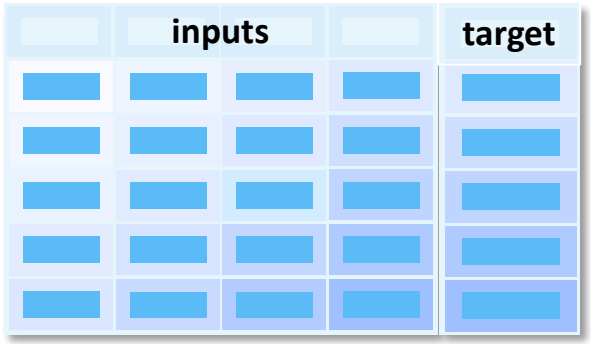
dependent variable

Predictive Model

Variables :

- Numeric
- Categorical

Training Data

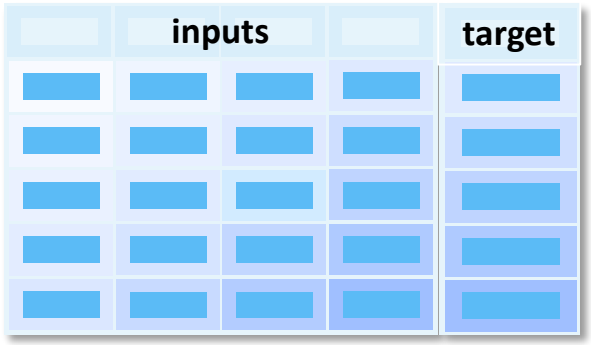


Predictive Model

Variables :

- Numeric
  - Continous (e.g Income)
  - Discrete (e.g No of items purchased)

Training Data



Predictive Model

Variables :

- categorical
  - Nominal

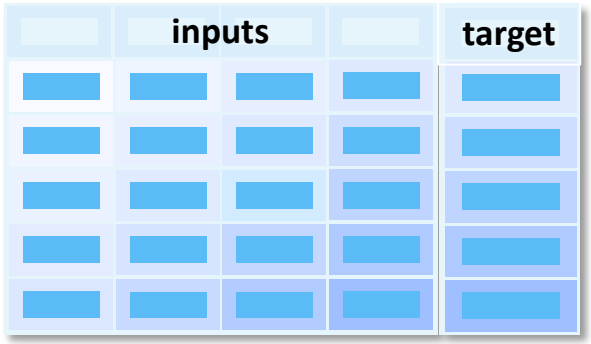
(e.g occupation)

  - Ordinal

(e.g shirt size)

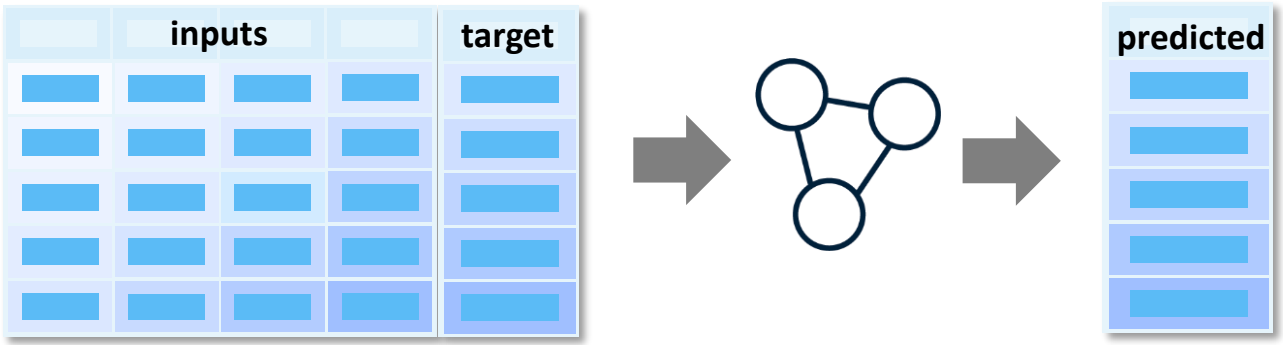
  - Binary (Y ,N)

Training Data



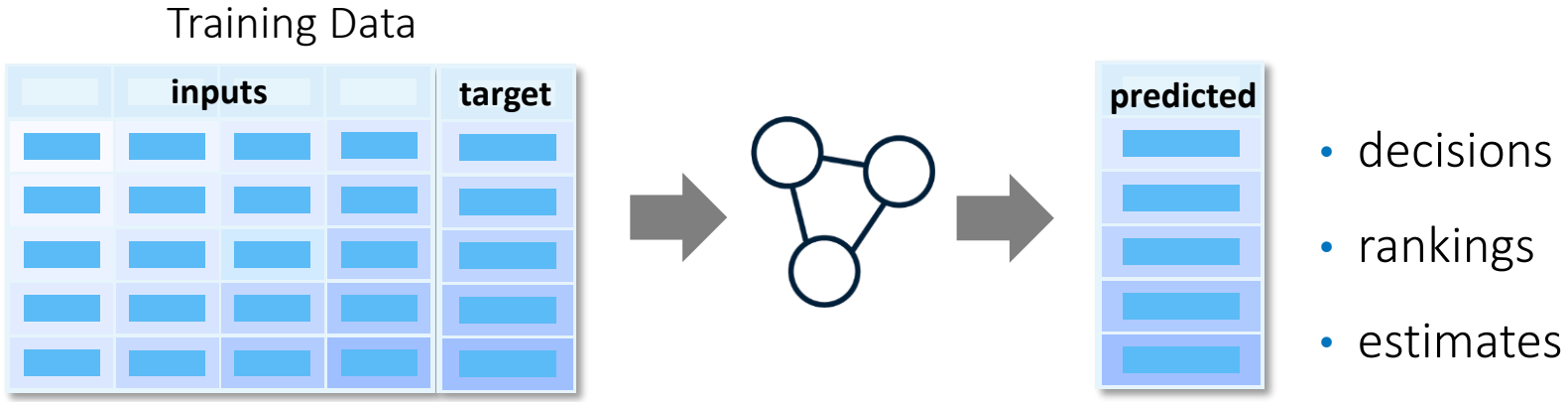
Predictions

Training Data



output of the predictive model given a set of input measurements

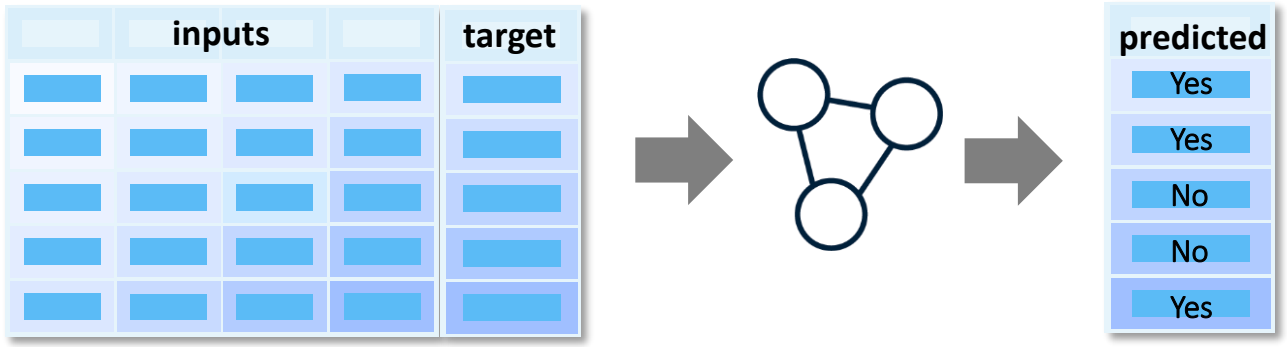
Prediction Types



- decisions
- rankings
- estimates

Decision Predictions

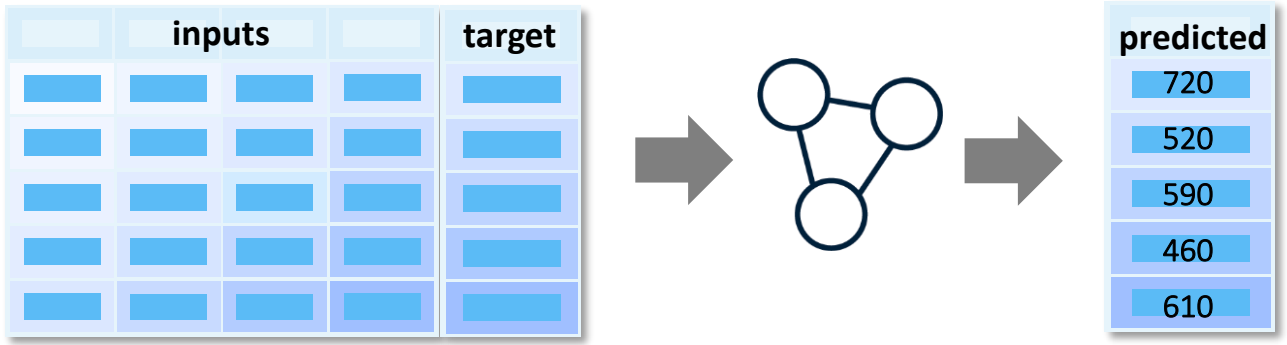
Training Data



A predictive model uses input measurements to make the best decision for each case.

Ranking Predictions

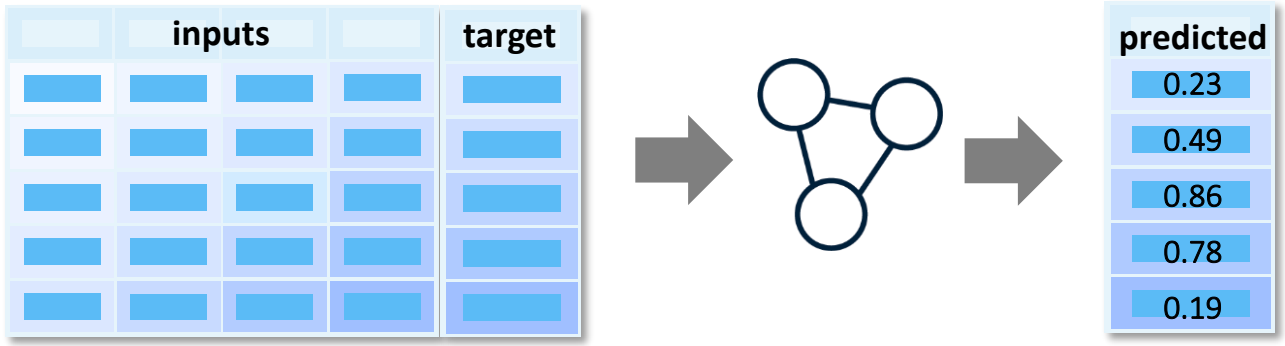
Training Data



A predictive model uses input measurements to optimally rank each case.

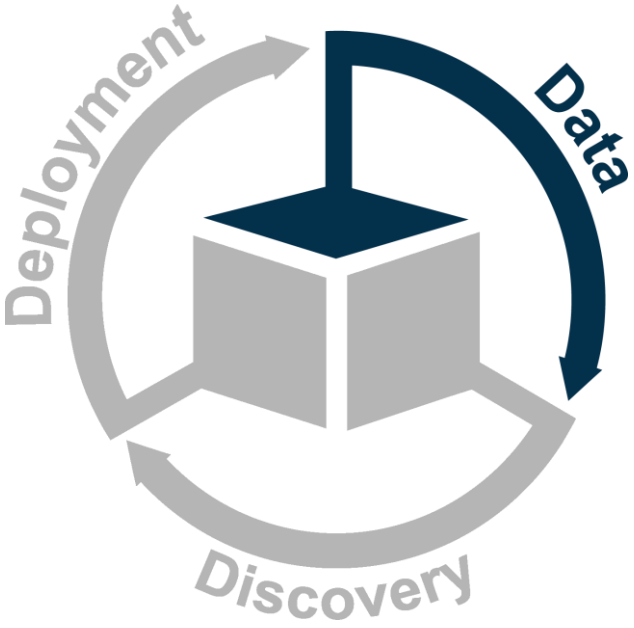
Estimate Predictions

Training Data



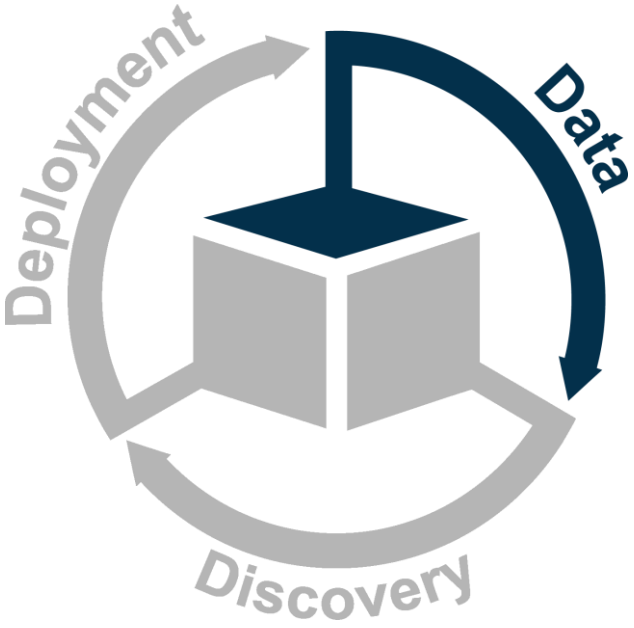
A predictive model uses input measurements to optimally estimate the target value.

Essential Data Tasks



Importance of Data Preparation

Essential Data Tasks

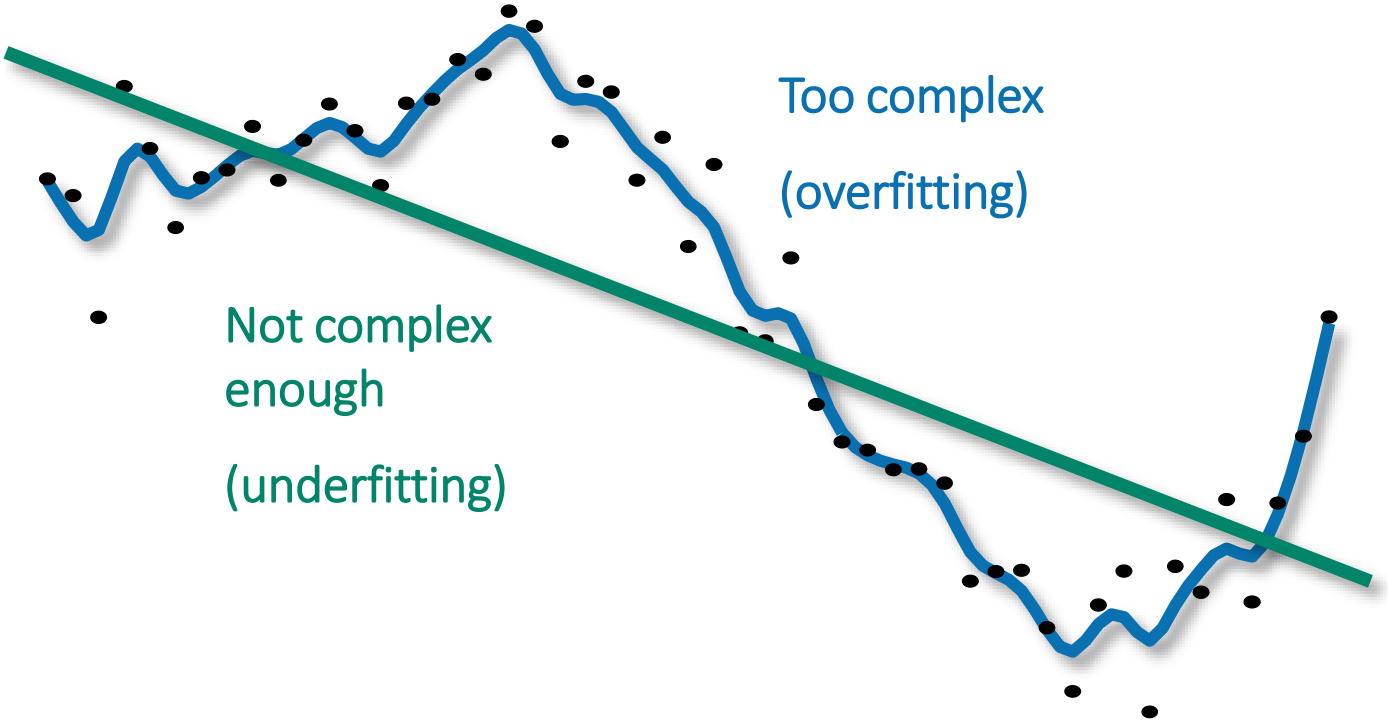


- Gather the data
- Explore the data
- Divide the data.
- Address rare events.
- Manage missing values.
- Add unstructured data.
- Extract features.
- Handle extreme or unusual values.
- Select useful inputs.

Essential Data Tasks

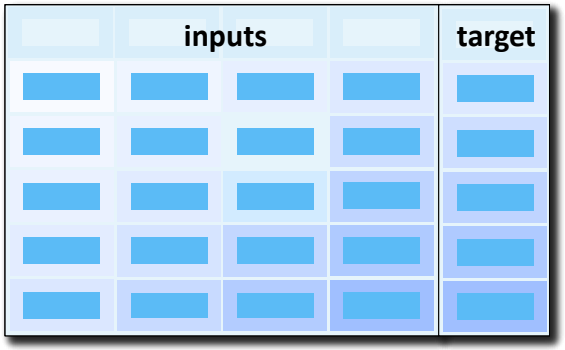
- Divide the data.
- Address rare events.
- Manage missing values.
- Add unstructured data.
- Extract features.
- Handle extreme or unusual values.
- Select useful inputs.

Accuracy versus Generalizability



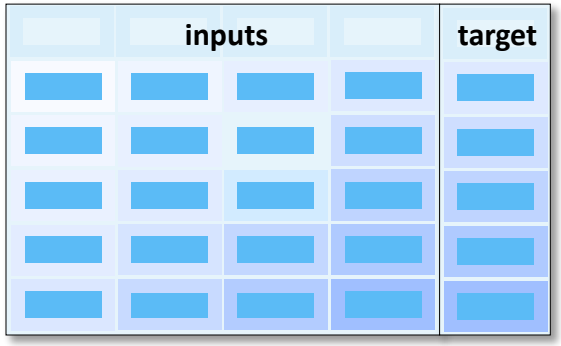
Partitioning the Input Data Set

Training Data

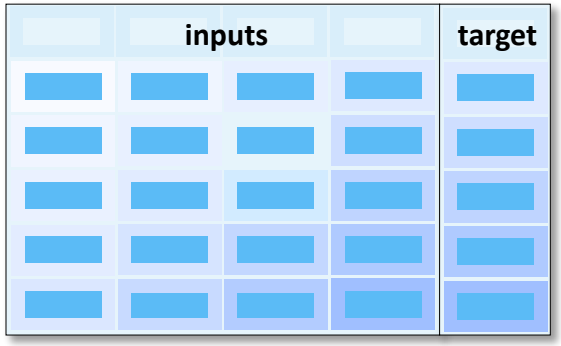


Partition available data into training, validation, and test sets.

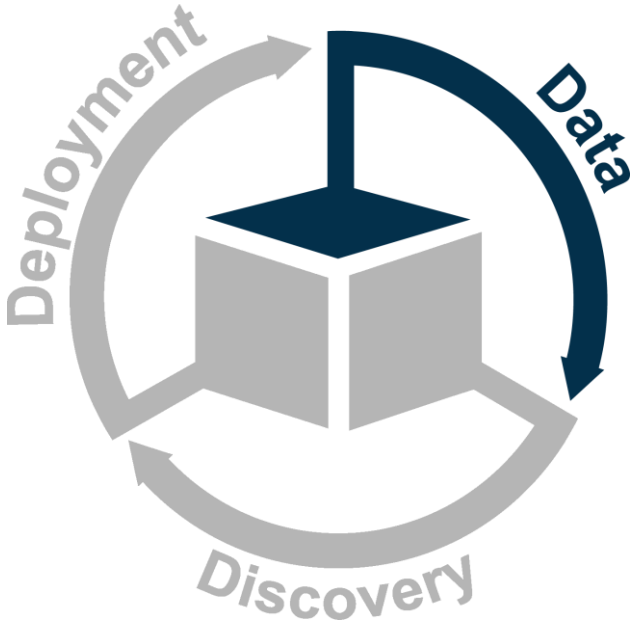
Validation Data



Test Data (Optional)



Essential Data Tasks

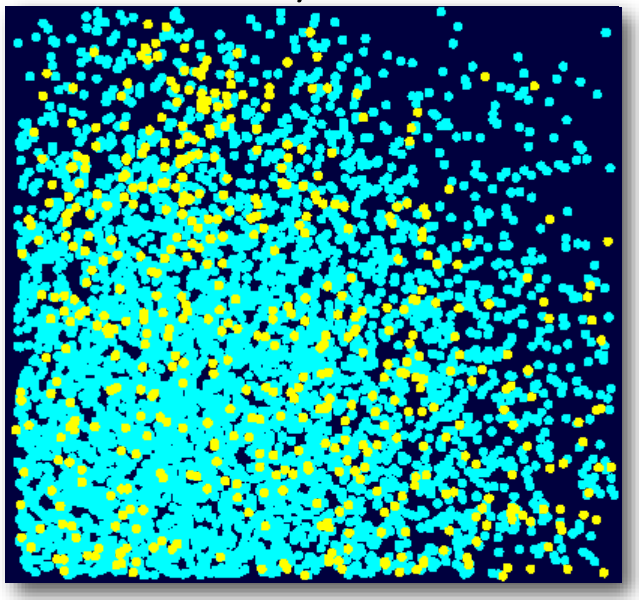


- Divide the data.
- Address rare events.
- Manage missing values.
- Add unstructured data.
- Extract features.
- Handle extreme or unusual values.
- Select useful inputs.

Event-Based Sampling

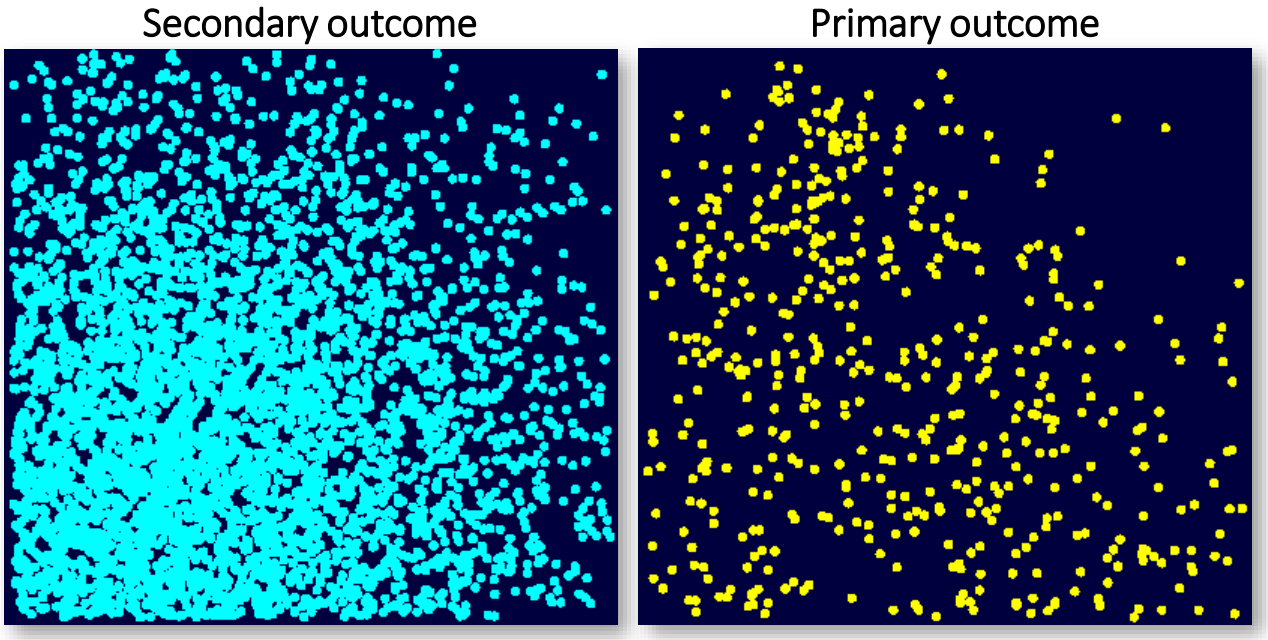
Secondary outcome

Primary outcome



Target-based samples are created by considering the primary outcome cases separately from the secondary outcome cases.

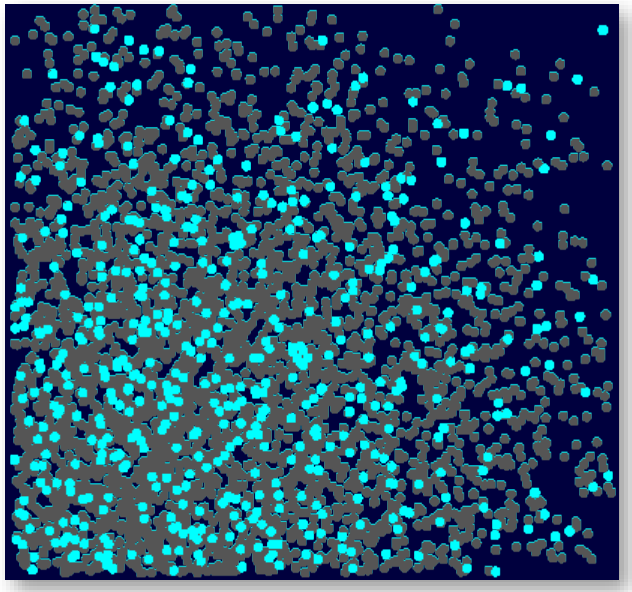
Event-Based Sampling



Target-based samples are created by considering the primary outcome cases separately from the secondary outcome cases.

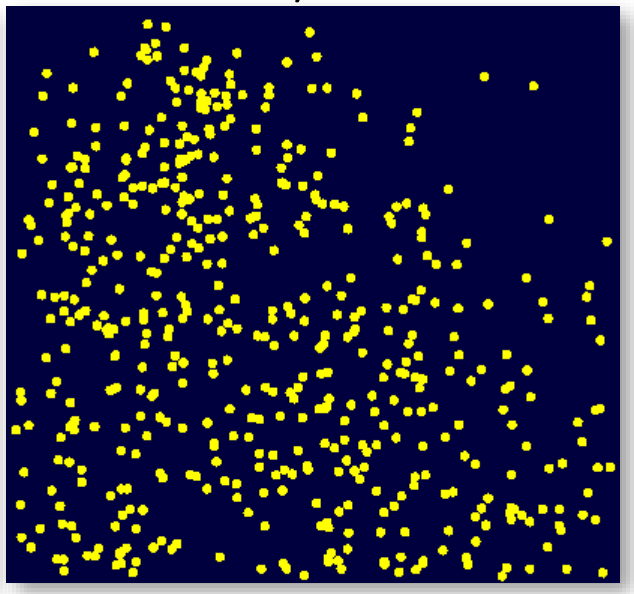
Event-Based Sampling

Secondary outcome



Select some cases.

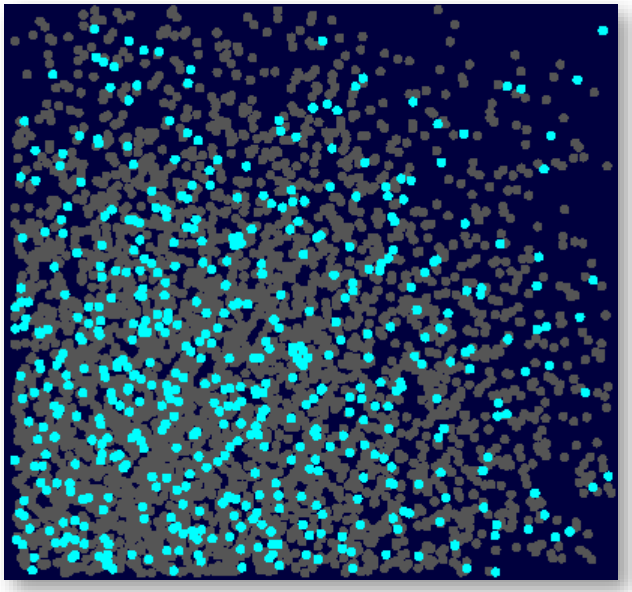
Primary outcome



Select all cases.

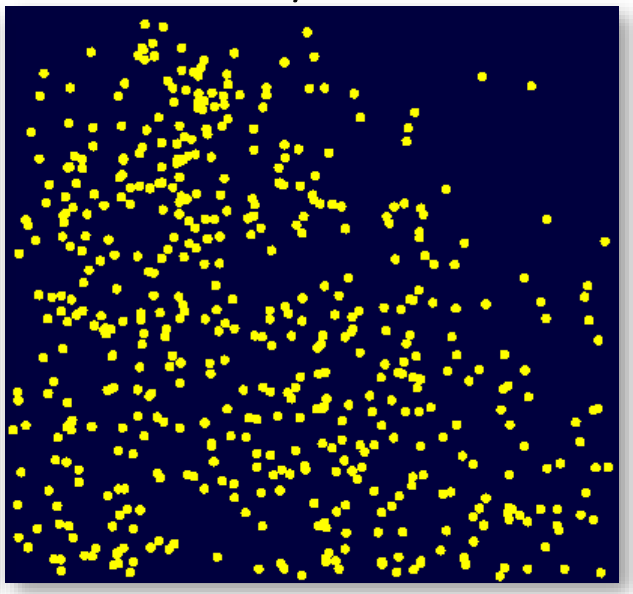
Event-Based Sampling

Secondary outcome



Select some cases.

Primary outcome



Select all cases.

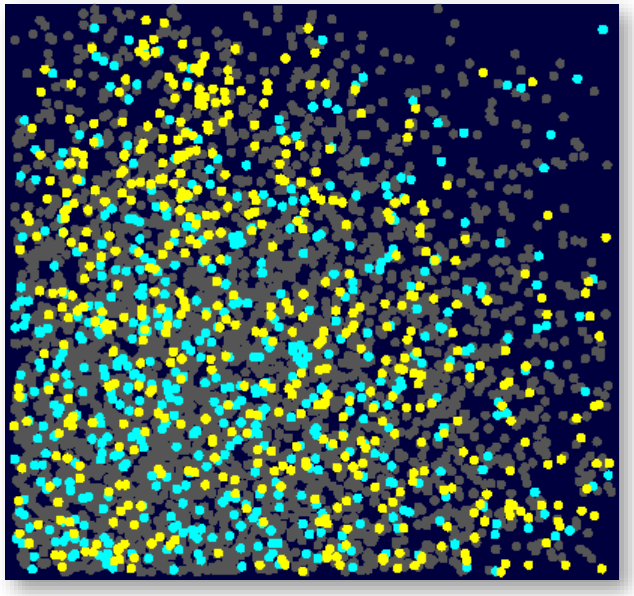
The Modeling Sample

+ Similar predictive power with smaller case count

– Must adjust assessment measures and graphics

– Must adjust prediction estimates for bias

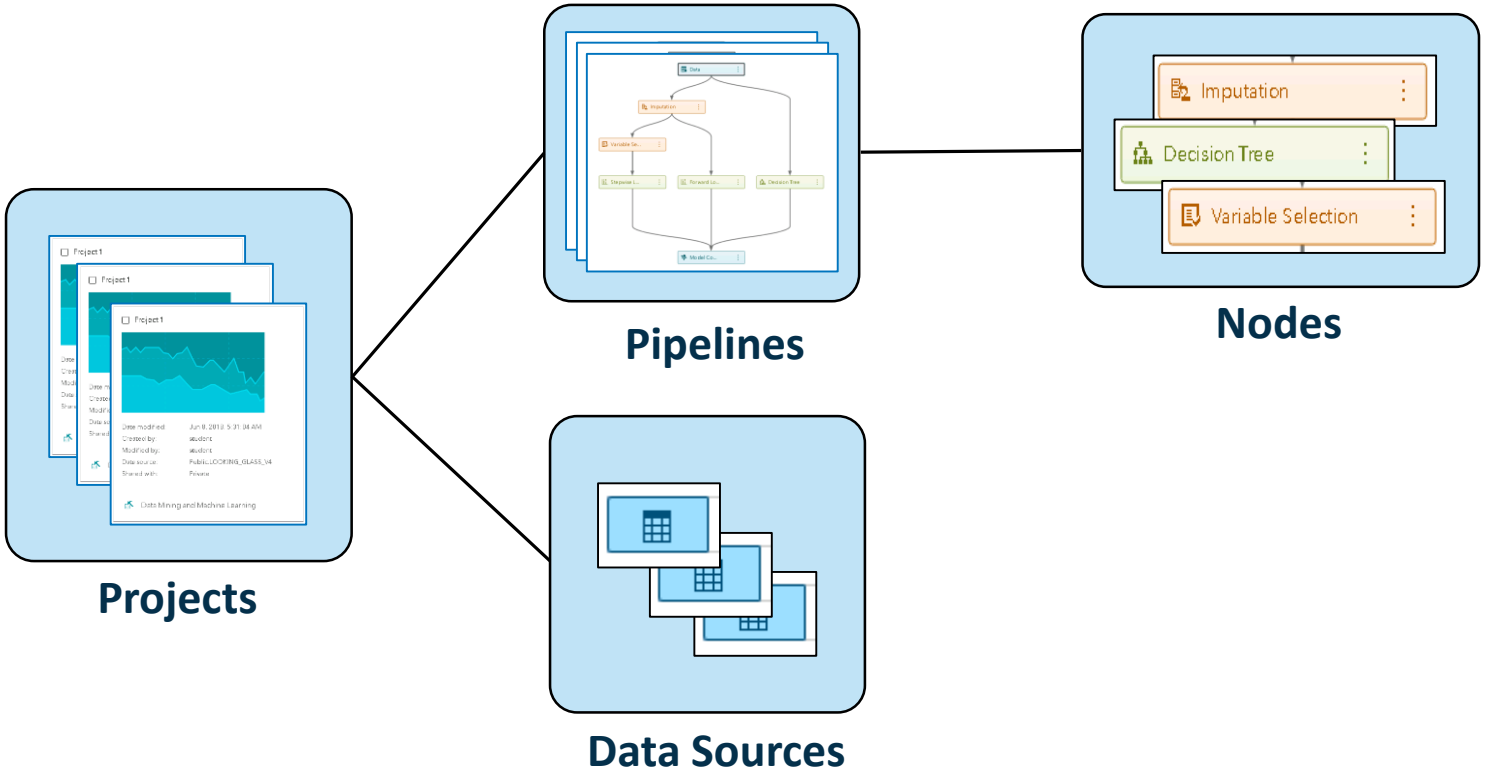
+ Model Studio automatically adjusts for event-based sampling



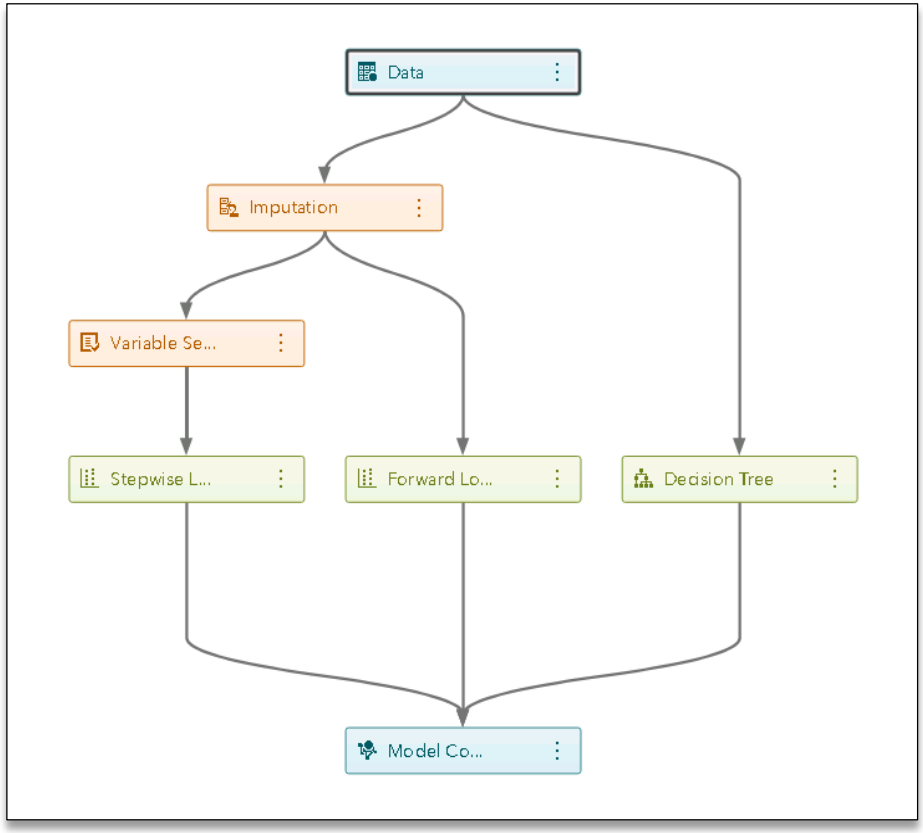
Modifying the Data Partition

In this demonstration, you modify metadata roles of some variables, explore the advanced project settings, and change the data partition properties.

Analysis Elements in Model Studio



Pipelines



- Pipelines are structured flows of analytic actions.
- Pipelines contain the nodes that process data and create models.
- Custom pipelines can be saved to ***the Exchange*** for others to use.

Pipelines Templates

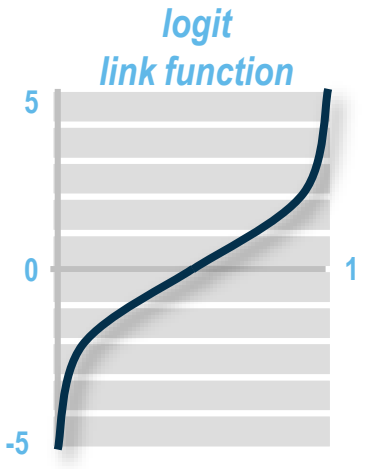
- Pre-populated pipeline templates are available for speedy model building.
- Three levels of pipeline templates (basic, intermediate, and advanced) are available for both class and interval targets.
- The advanced pipeline template is available with autotuning functionality.
- Each increasing level of pipeline template adds more data preprocessing and models.
- Regression (Linear/Logistic) is part of all the three pipeline template levels.

Note: You will build a basic pipeline, which consists of regression and imputation.

Logistic Regression

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 . x_1 + \hat{\beta}_2 . x_2$$

logit scores



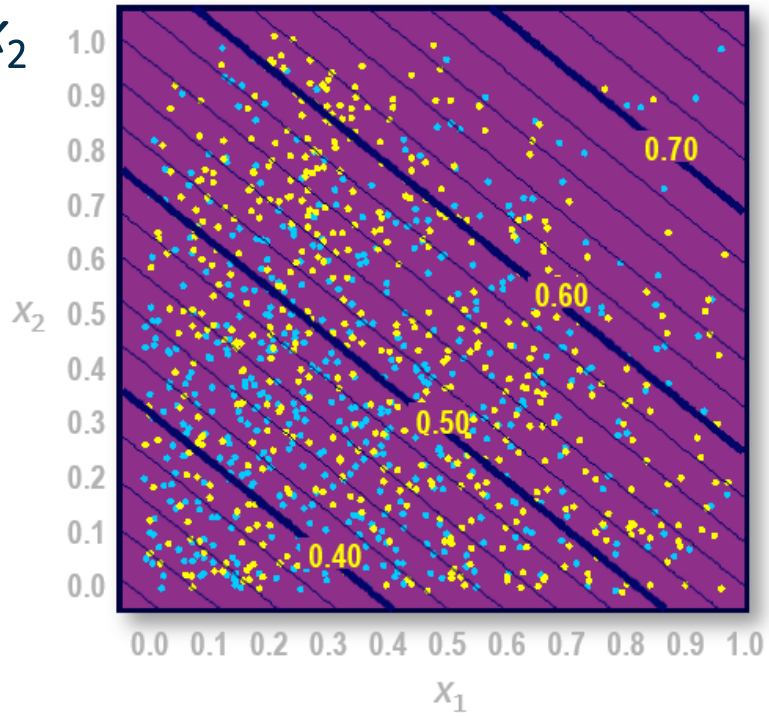
The logit link function transforms probabilities (between 0 and 1) to logit scores (between $-\infty$ and $+\infty$).

Logistic Regression Example

$$\text{logit}(\hat{p}) = -0.81 + 0.92 \cdot x_1 + 1.11 \cdot x_2$$

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

Using the maximum likelihood estimates, the prediction formula assigns a logit score to each x_1 and x_2.

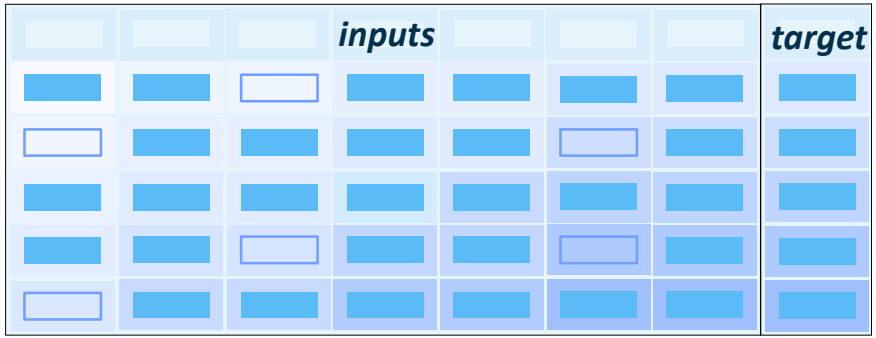


Essential Data Tasks

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

Missing Values: Problem 1

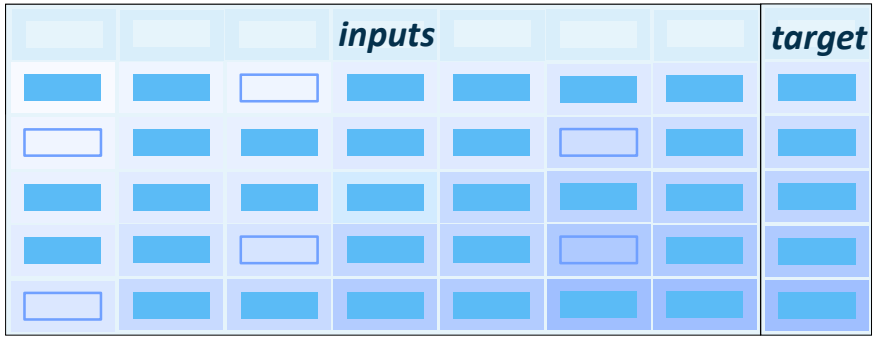
Training Data



Problem: Complete case analysis means that training data cases with missing values on inputs are ignored.

Missing Values: Problem 1

Training Data



Problem: Complete case analysis means that training data cases with missing values on inputs are ignored.

Missing Values: Problem 1

Training Data

Consequence: Missing values can significantly reduce your amount of training data for regression modeling.

Missing Values: Problem 2

$$\text{logit}(\hat{p}) = -0.81 + 0.92 \cdot x_1 + 1.11 \cdot x_2$$

Predict: $(x_1, x_2) = (0.3, ?)$

Problem: What if the scoring data also have missing values?

Missing Values: Problem 2

$$\text{logit}(\hat{p}) = -0.81 + 0.92 \cdot x_1 + 1.11 \cdot ?$$

Predict: $(x_1, x_2) = (0.3, ?)$

$$\text{logit}(p) = ?$$

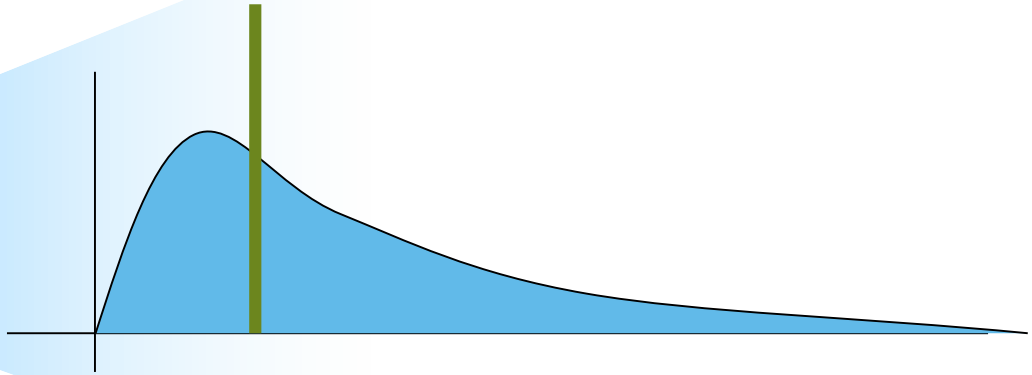
Consequence: Prediction formulas cannot score cases with missing values.

Managing Missing Values

- Naïve Bayes
- Decision trees
- Missing indicators
- Imputation
- Binning
- Scoring missing data

Managing Missing Values

- Naïve Bayes
- Decision trees
- Missing indicators
- Imputation
- Binning
- Scoring missing data



Building a Pipeline from a Basic Template

In this demonstration, you build a new pipeline from a basic template for class target..

Chapter 1: Introduction

1.1 Machine Learning in Business Decision Making

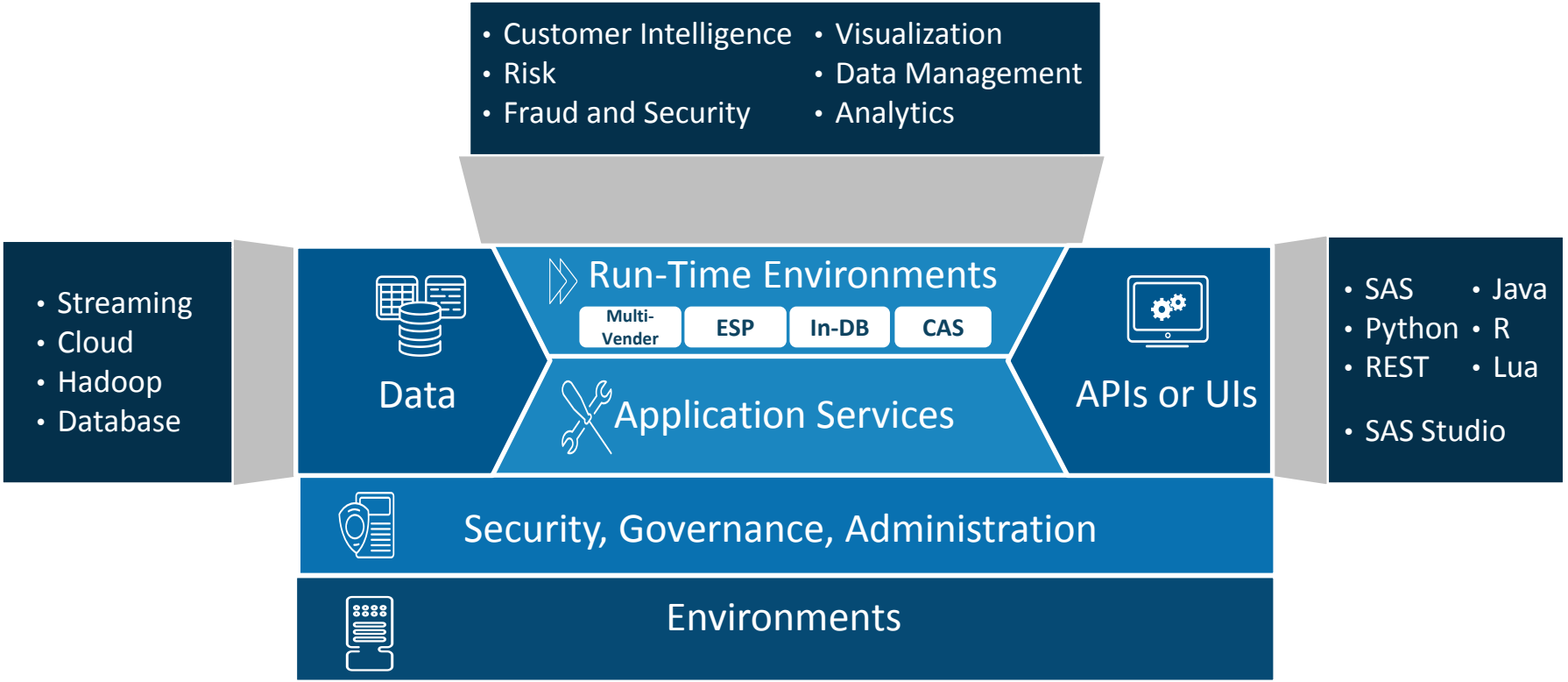
1.2 Essentials of Supervised Prediction

1.3 Introduction to SAS Viya

SAS Viya on the SAS Platform

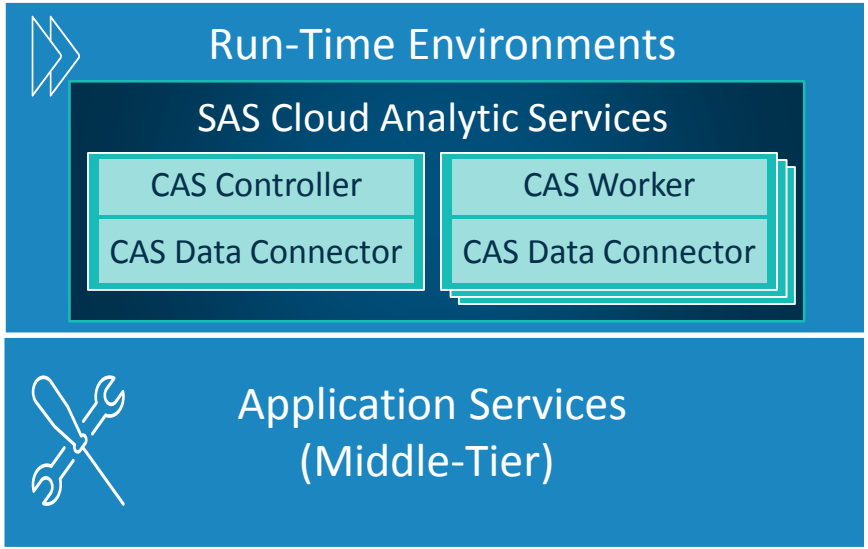
SAS Viya is an open, cloud-enabled, analytic run-time environment with a number of supporting services, including SAS Cloud Analytic Services (CAS). CAS is the in-memory engine on the SAS Platform.

SAS Platform Architecture

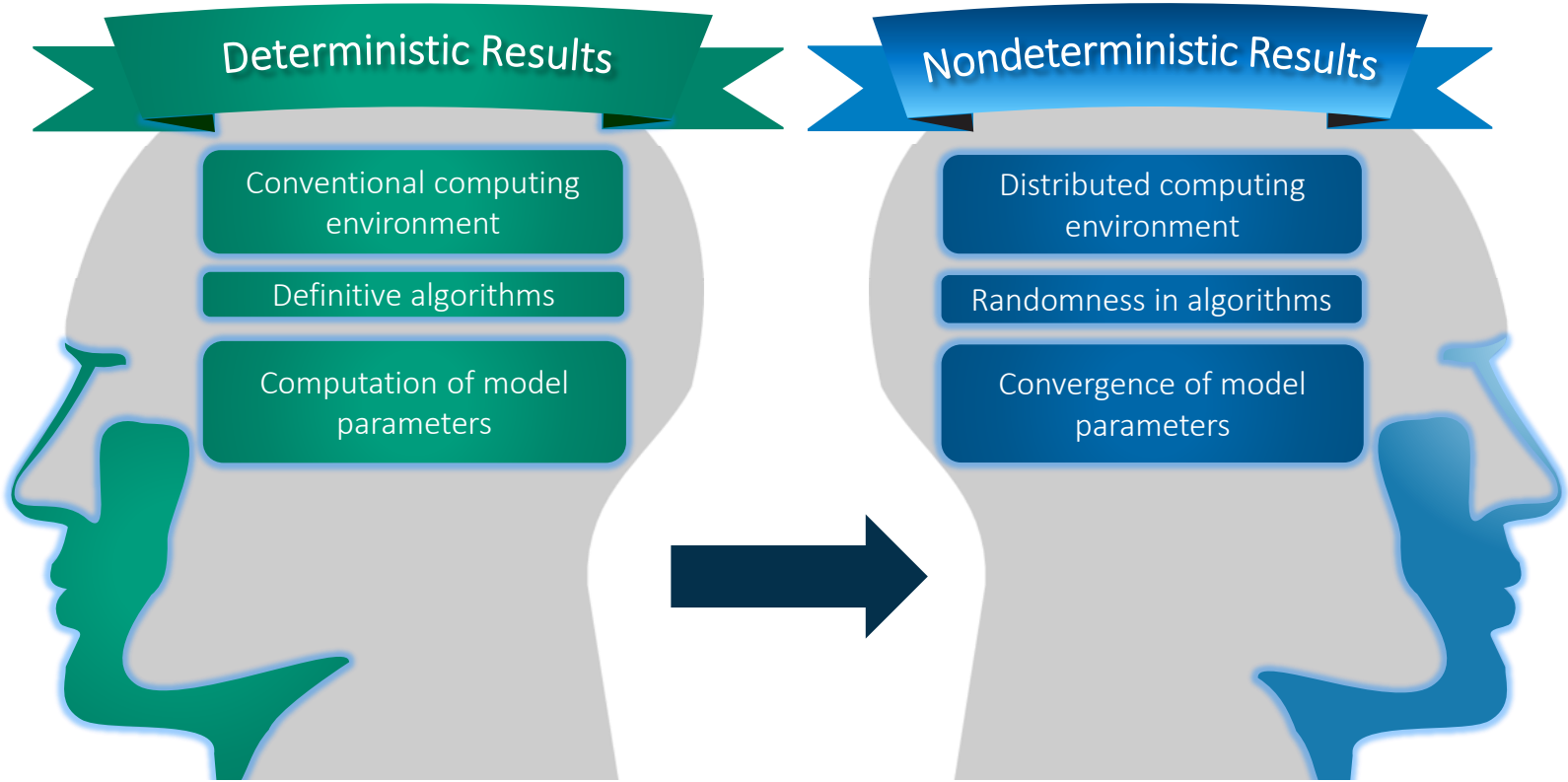


SAS Cloud Analytic Services

Cloud Analytic Services (CAS) is an in-memory, distributed, analytics engine. It uses scalable, high-performance, multi-threaded algorithms to rapidly perform analytical processing on in-memory data of any size.



A Mindset Shift



SAS Viya Infrastructure

SAS Viya is open to any public or private cloud platform.

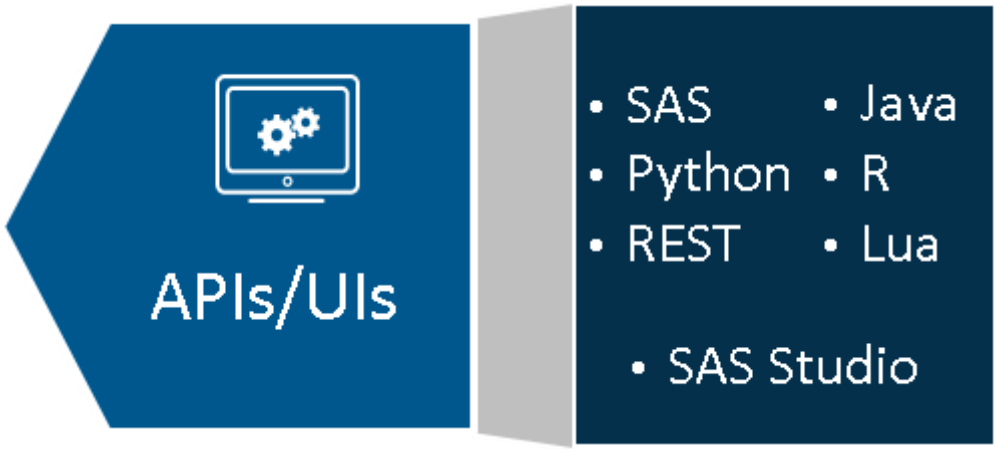
- SAS Viya and SAS®9 can coexist on the same hardware (physical or virtual).
- Multi-tenancy is supported.
- SAS Viya integrates with existing security practices.

Data Sources and SAS Viya

A variety of data sources can be accessed. These include native access to cloud application and data sources, enterprise on-premises data sources, relational and unstructured data, Hadoop, and various file formats (XML, JSON, CSV).

Interfaces to SAS Viya

Although SAS Viya can be used by various SAS applications, it also enables you to access analytic methods from SAS, Python, Lua, and Java, as well as through a REST interface that uses HTTP or HTTPS.



Products on SAS Viya

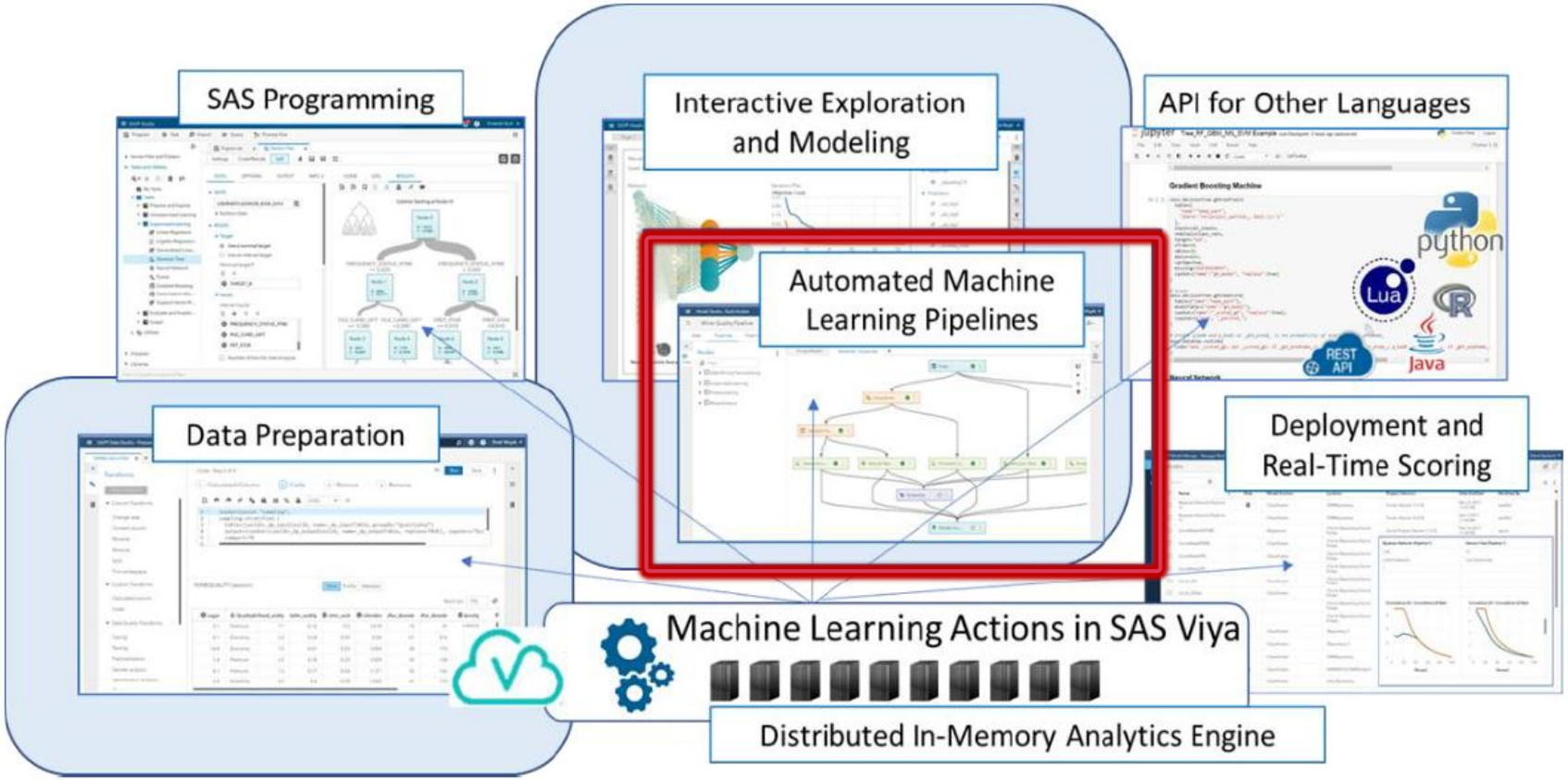
SAS products are licensed on the SAS Viya platform.

- SAS® Model Manager
- SAS® Decision Manager
- SAS® Event Stream Processing
- SAS® Scoring Accelerators

- SAS® In-Database Technologies
- SAS® Data Preparation
- SAS® Data Quality

- SAS® Visual Forecasting
- SAS® Econometrics
- SAS® Visual Investigator
- SAS® Visual Analytics
- SAS® Optimization
- SAS® Visual Text Analytics
- SAS® Visual Data Mining and Machine Learning
- SAS® Visual Statistics

SAS Visual Data Mining and Machine Learning



SAS Viya Consistency

Different Interfaces, Same Results

Submit
CAS Action

Return
Action Results

SAS Visual Data Mining and Machine Learning

SAS Visual Analytics

SAS Visual Statistics

… …

CAS-enabled Procedures: TREESPLIT, LOGSELECT, GENSELECT, …