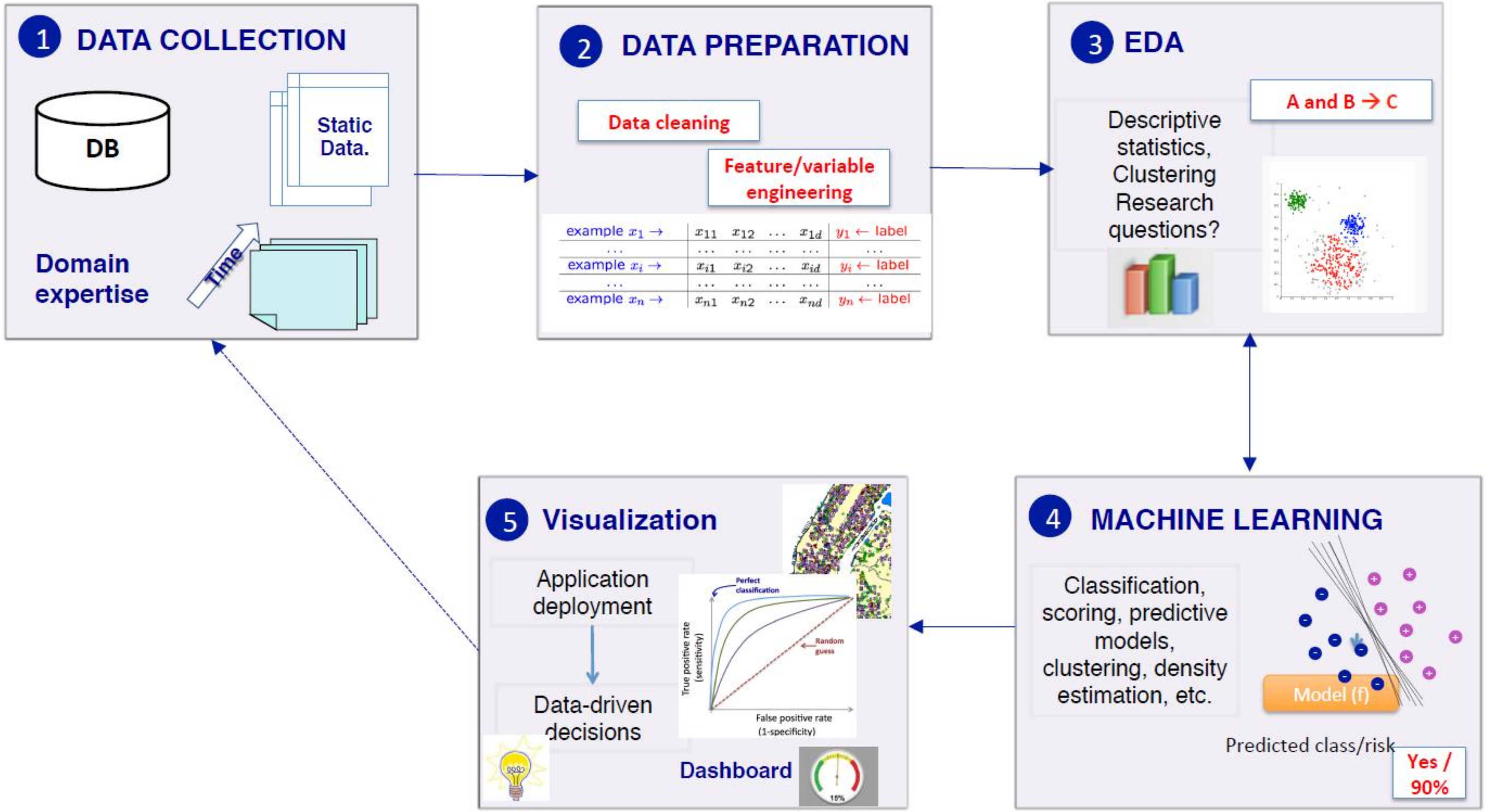
Machine Learning data processing

Credit card fraud detection Case Demo

YAO ZHAO

The Data Science process



1 DATA COLLECTION

DB

Static Data.

Domain expertise

Time

2 DATA PREPARATION

Data cleaning

Feature/variable engineering

example x_1 →	x_{11}	x_{12}	$\cdots$	x_{1d}	y_1 ← label
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
example x_i →	x_{i1}	x_{i2}	$\cdots$	x_{id}	y_i ← label
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
example x_n →	x_{n1}	x_{n2}	$\cdots$	x_{nd}	y_n ← label

3 EDA

A and B → C

Descriptive statistics, Clustering Research questions?

4 MACHINE LEARNING

Classification, scoring, predictive models, clustering, density estimation, etc.

Model (f)

Predicted class/risk

Yes / 90%

5 Visualization

Application deployment

Data-driven decisions

Dashboard

Perfect classification
Random guess
True positive rate (sensitivity)
False positive rate (1-specificity)
15%

Machine Learning

- Spam filtering
- Credit card fraud detection
- Digit recognition on checks, zip codes
- Detecting faces in images
- MRI image analysis
- Recommendation system
- Search engines
- Handwriting recognition
- Scene classification
- etc…

- Data collection
- Data Preparation
- Exploration Data Analysis
- Modeling
- Visualization

How to find datasets?

Big data competition platform

- Kaggle: https://www.kaggle.com/datasets

Colleges and Universities

- UCI: https://archive.ics.uci.edu/ml/datasets.html

Enterprises or public welfare organizations

- Google dataset: https://cloud.google.com/bigquery/public-data/

Government's open data

- The U.S. Government's open data: https://data.gov/

Searching datasets:

- https://datasetsearch.research.google.com/

Credit card fraud detection

- **Data collection**
- Data Preparation
- Exploration Data Analysis
- Modeling
- Visualization

www.kaggle.com/mlg-ulb/creditcardfraud

Download the dataset : creditcard.csv

- Data collection
- **Data Preparation**
- Exploration Data Analysis
- Modeling
- Visualization

Data Preparation

Eliminates duplicate and null values, corrupt data, inconsistent data types, invalid entries, missing data, and improper formatting.

Python libraries for Data Preparation

NumPy

- NumPy is the foundation for many other packages that hold the data science ecosystem like Pandas, Matplotlib and Scikit-learn.

Pandas

- Pandas offer developers fast, efficient and optimized objects for data manipulation in various academic and industrial fields.

Matplotlib

- The core package used for data visualization. Matplotlib offers various plots and figures developers can use to create different visualizations

Pandas

DataFrame : A Pandas DataFrame is a 2-dimensional data structure, like a 2-dimensional array, or a table with rows and columns.

pandas.read_csv(): Load a CSV into a DataFrame

pandas.DataFrame.head(): returns the headers and a specified number of rows, starting from the top.

pandas.DataFrame.tail(): viewing the *last* rows of the DataFrame

pandas.DataFrame.info():viewing more information about the data set.

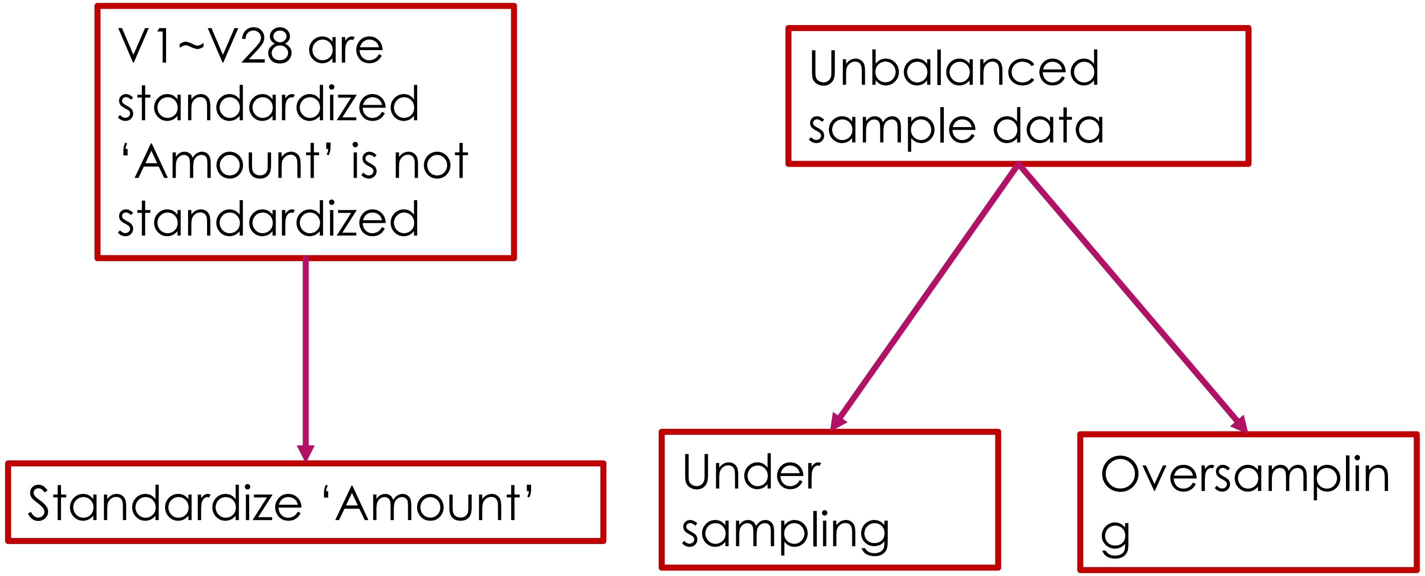
pandas.DataFrame.loc(): Access a group of rows and columns by label(s) or a boolean array.

DataFrame.describe(): Generate descriptive statistics.Descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.

- Data collection
- Data Preparation
- **Exploration Data Analysis**
- Modeling
- Visualization

EDA

Analyze and investigate the data set and summarize its key characteristics



V1~V28 are standardized ‘Amount’ is not standardized

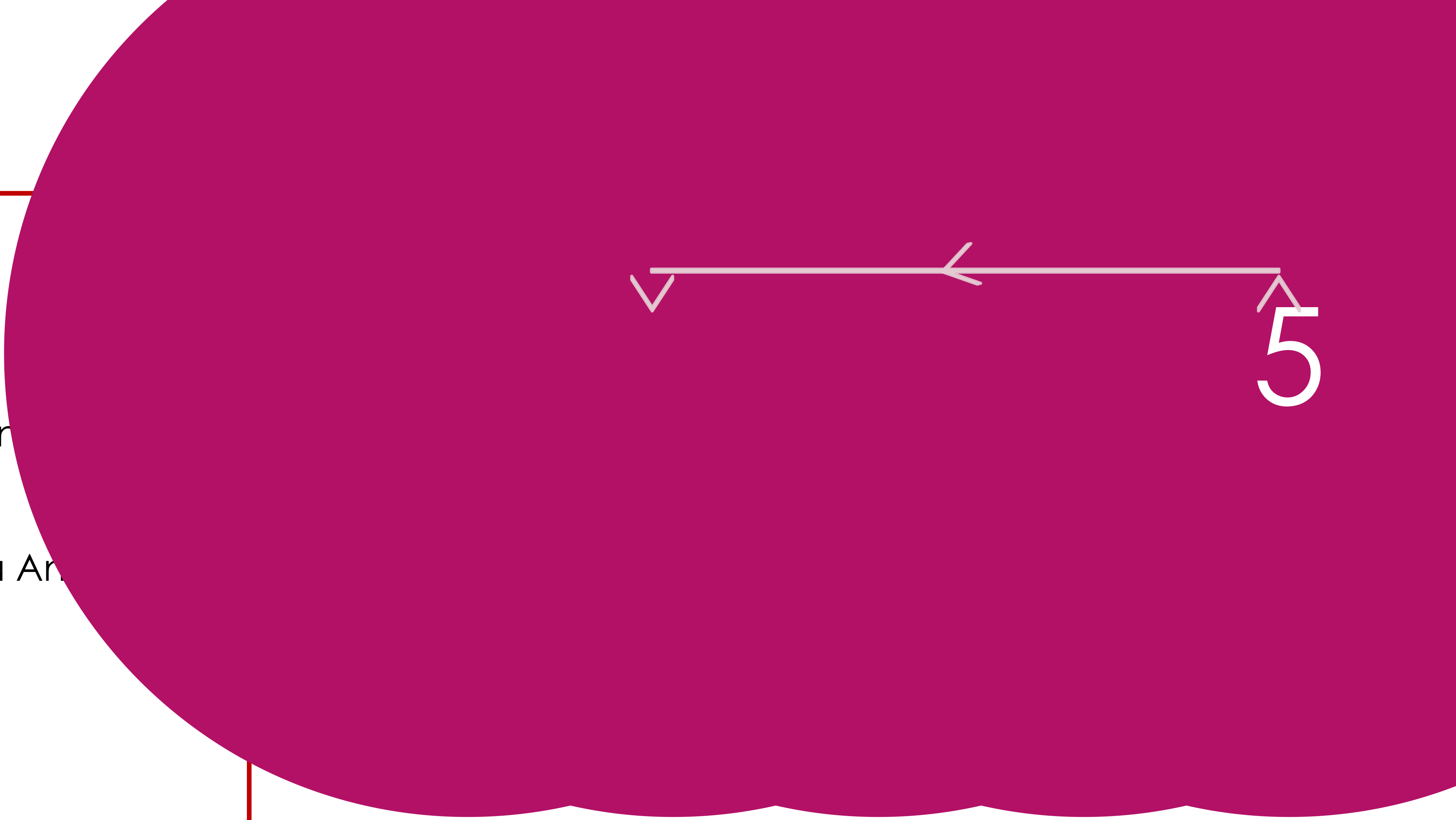
Standardize ‘Amount’

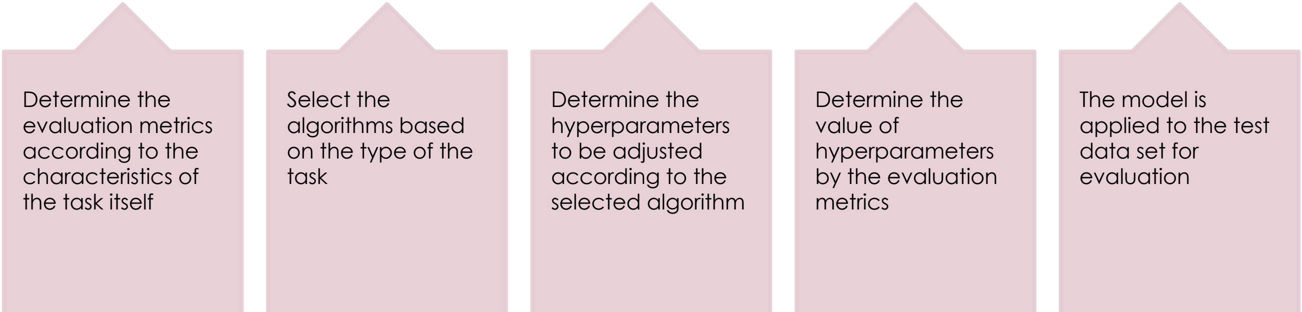
Unbalanced sample data

Under sampling

Oversampling

- Data collection
- Data Preparation
- Exploration Data An
- **Modeling**
- Visualization





Determine the evaluation metrics according to the characteristics of the task itself

Select the algorithms based on the type of the task

Determine the hyperparameters to be adjusted according to the selected algorithm

Determine the value of hyperparameters by the evaluation metrics

The model is applied to the test data set for evaluation

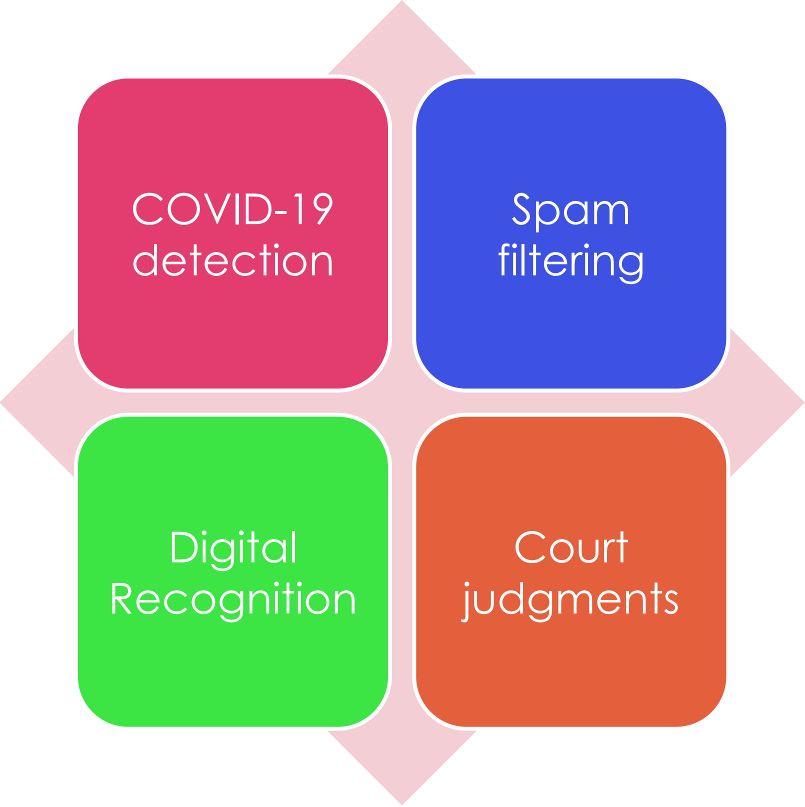
Modeling

1

		Actual Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

		Actual Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative



COVID-19 detection

Spam filtering

Digital Recognition

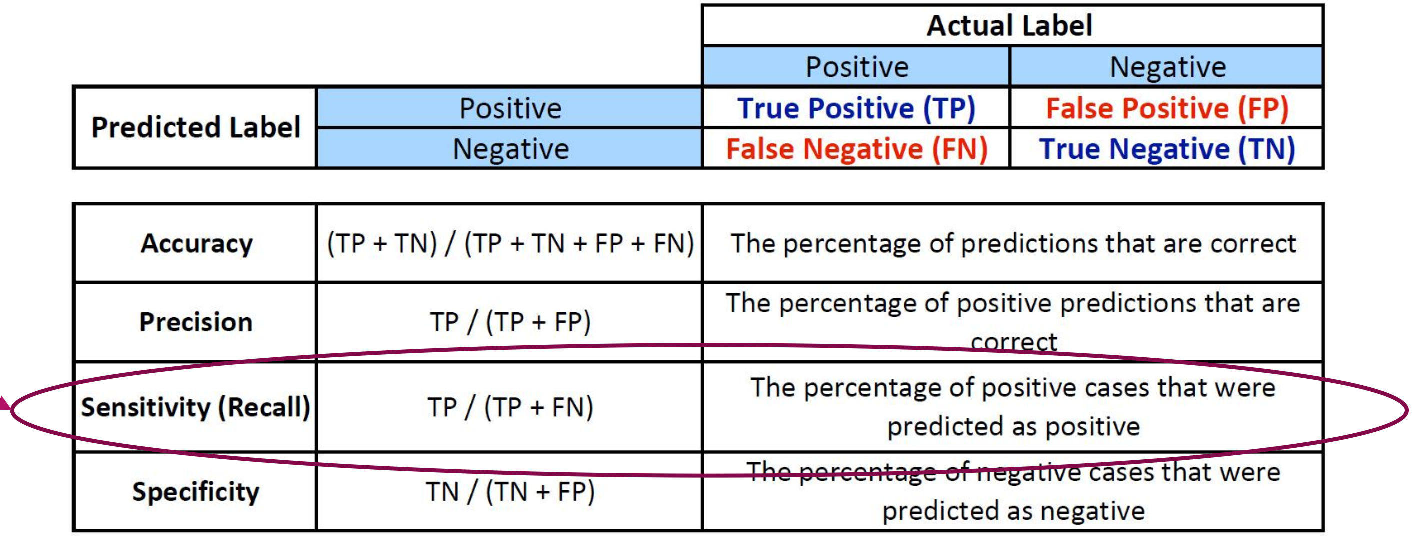
Court judgments

Modeling 1

1

Determine the evaluation metrics according to the characteristics of the task itself

Credit Card Fraud Detection



		Actual Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative

Modeling 2

2

Select the algorithms based on the type of the task

Supervised Learning

Training data: “examples” *x* with “labels” *y*.

$$(x_1, y_1), \ldots, (x_n, y_n),\ x_i \in \mathbb{R}^d$$

- **Classification**: *y* is discrete. To simplify, *y* ∈ {-1, +1}

$f\colon \mathbb{R}^d \to \{-1, +1\}$ (*f* is called a **binary classifier**)

Example: Approve credit yes/no, spam/ham, banana/orange.

Credit Card Fraud Detection

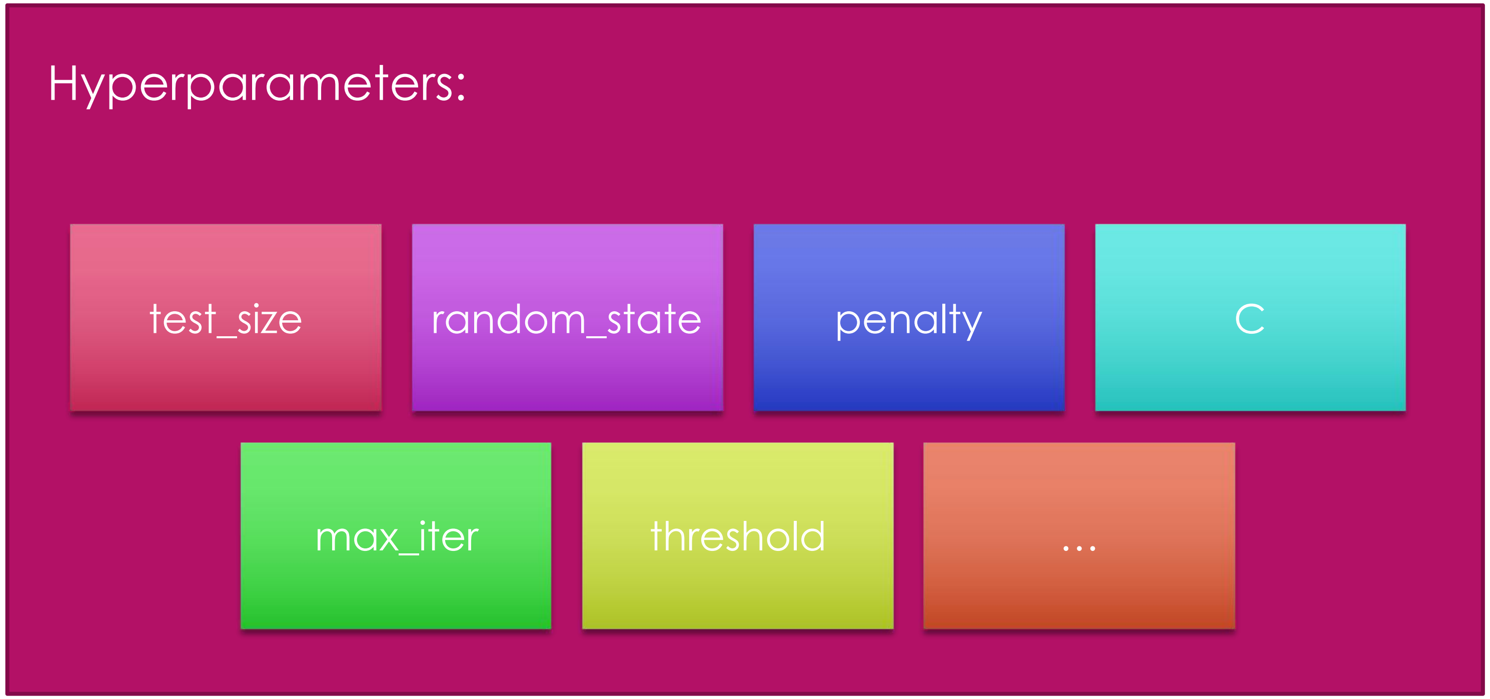
- Methods:
  - Logistic Regression
  - SVM
  - Neural network
  - decision tree
  - …

Modeling 3

3

Determine the hyperparameters to be adjusted according to the selected algorithm

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logisticregression#examples-using-sklearn-linear-model-logisticregression



Hyperparameters:

test_size

random_state

penalty

C

max_iter

threshold

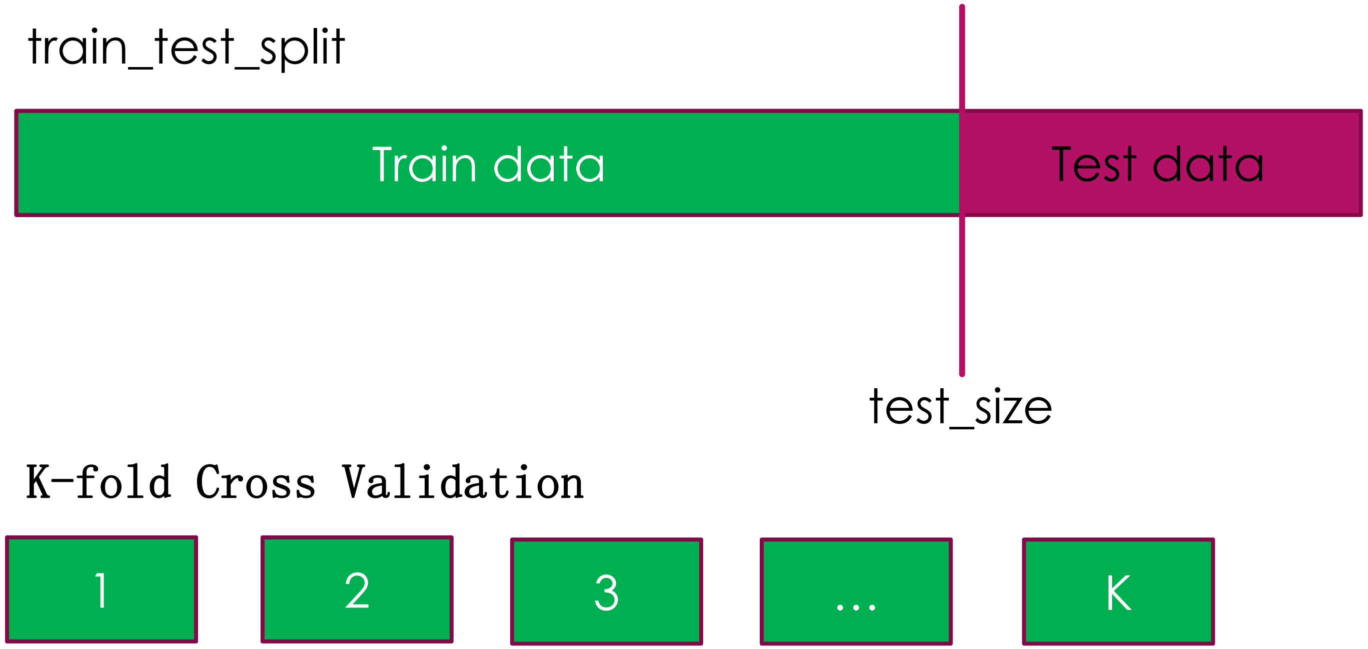
…

Modeling

4

4

Determine the value of hyperparameters by the evaluation metrics



train_test_split

Train data

Test data

test_size

K-fold Cross Validation

1

2

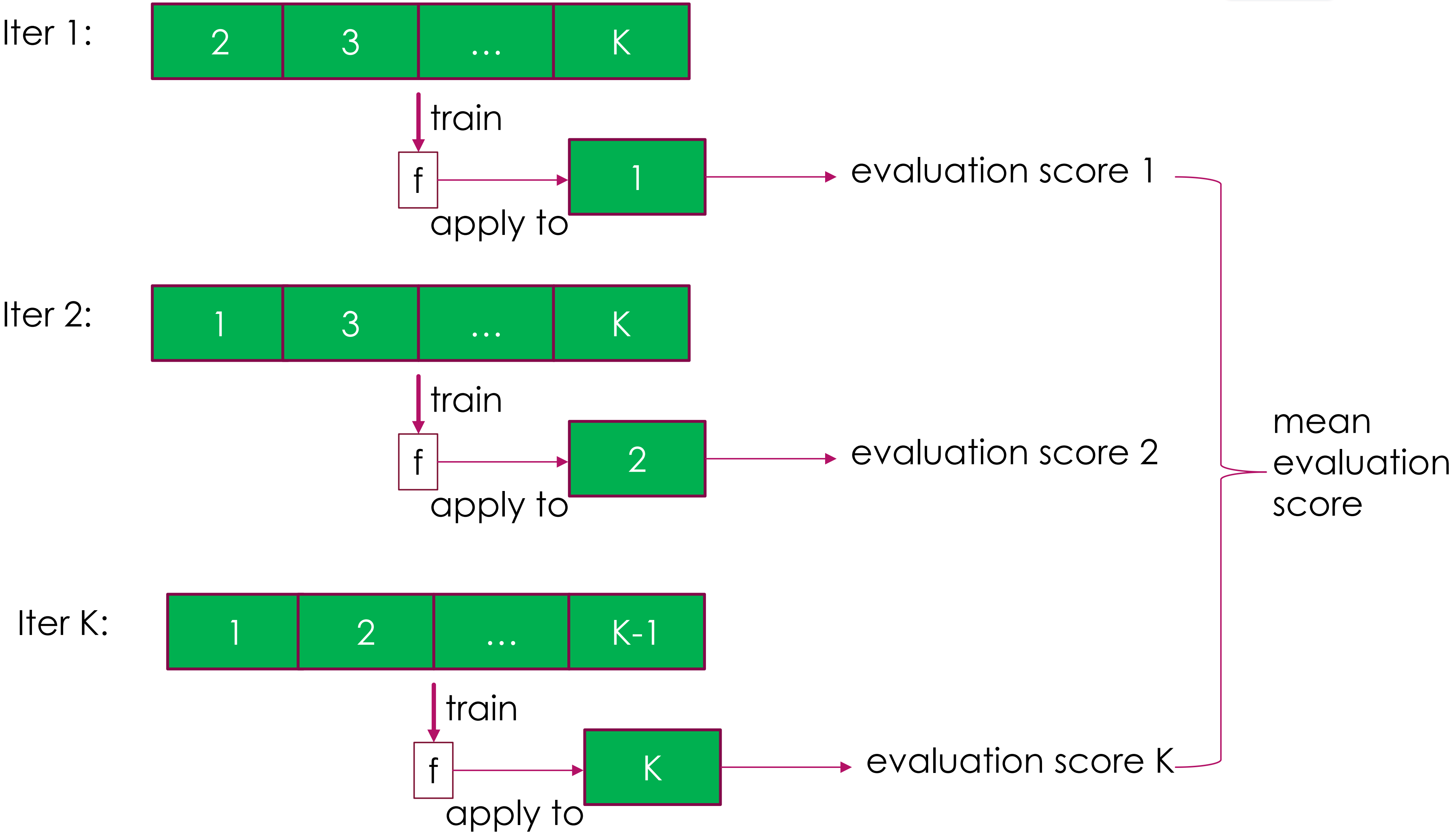
3

...

K

Modeling 4

K-fold
Cross
Validation



Iter 1:
2 | 3 | … | K
train
f
apply to
1
evaluation score 1

Iter 2:
1 | 3 | … | K
train
f
apply to
2
evaluation score 2

Iter K:
1 | 2 | … | K-1
train
f
apply to
K
evaluation score K

mean
evaluation
score

Modeling 5

5

The model is applied to the test data set for evaluation

Accept
or try other method of sampling
or other ML algorithms

Visualization

- Data collection
- Data Preparation
- Exploration Data Analysis
- Modeling
- **Visualization**

Visualization methods refer to the use of visual representation to display complex resource content after the original data is converted into visual elements and to deepen the user's understanding. Some important visualization techniques are histograms, Scatter Plots, timelines, Box and Whisker Plots, and treemaps.