머글들을 위한
Machine Learning 3

김영욱
부장/ PS / Microsoft

youngwook@outlook.com
Blog: Youngwook.com

문제 정의

어떤 모델을 만들 것인가?

미국 45개 월 마트 부서별 주간 판매액 예측 모델


kaggle™

kaggle Search Competitions Datasets Notebooks Discussion Courses ... Sign in Registe

Competitions

Documentation InClass

General InClass Sort by Grouped

All Categories Search competitions

19 Active Competitions

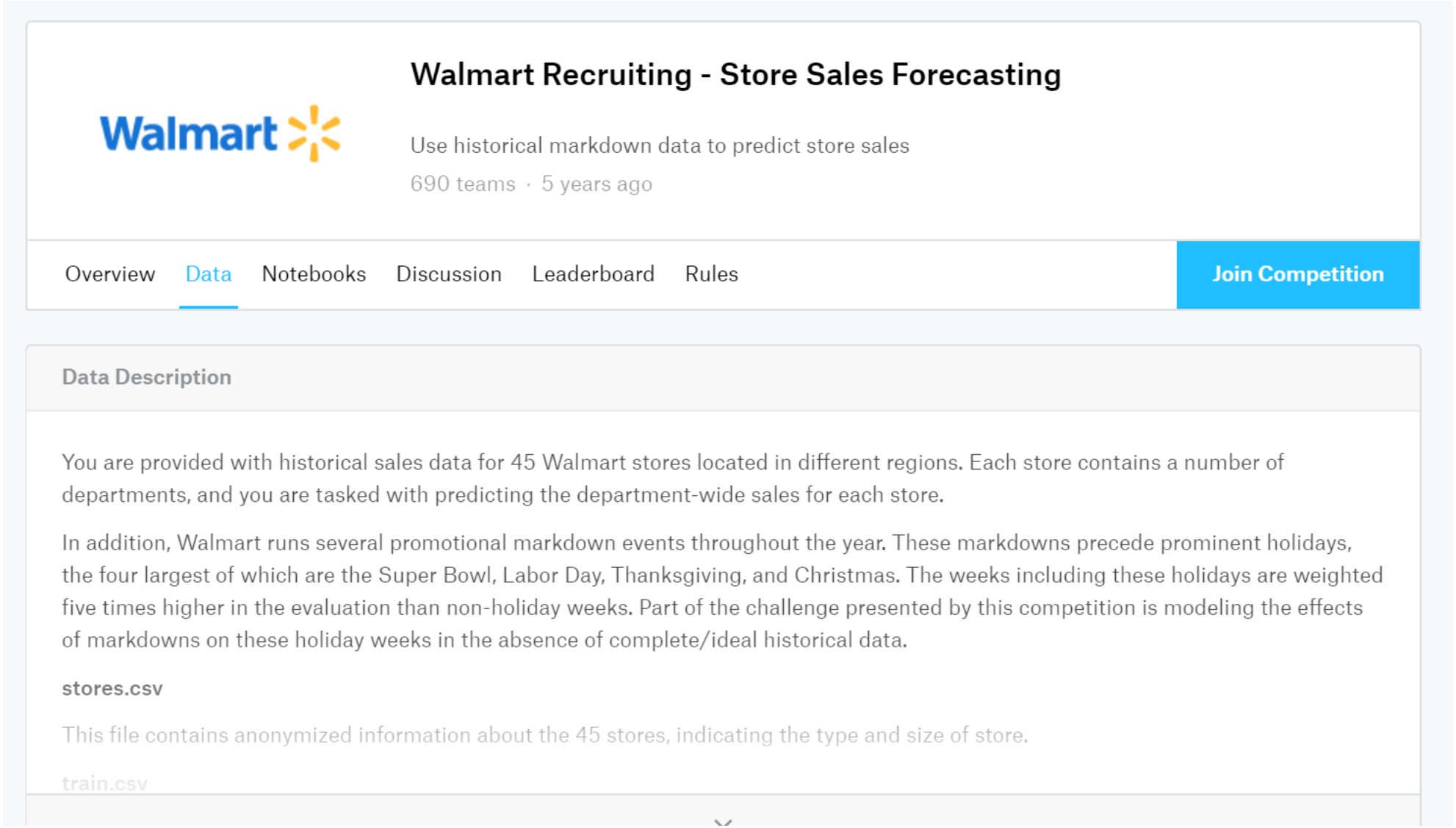
Severstal: Steel Defect Detection — $120,000
Can you detect and classify defects in steel? — 1,665 teams
Featured · Code Competition · a month to go · manufacturing, image data

Lyft 3D Object Detection for Autonomous Vehicles — $25,000
Can you advance the state of the art in 3D object detection? — 123 teams
Featured · 2 months to go · image data, object detection

RSNA Intracranial Hemorrhage Detection — $25,000
Identify acute intracranial hemorrhage and its subtypes — 399 teams
Featured · 2 months to go · health foundations and medical research, image data

The 3rd YouTube-8M Video Understanding Challenge — $25,000
Temporal localization of topics within video — 263 teams
Research · 16 days to go · video data, object detection

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data



데이터 전처리 과정

- 데이터 지원 형식 확인 및 변환
- 데이터 업로드
- 데이터 Merge
- 데이터 전처리

데이터 지원 형식

- 헤더가 있거나 (.csv) 없는 (.nh.csv) 쉼표로 구분된 값(csv)
- 헤더가 있거나 (.tsv) 없는 (.nh.tsv) 탭으로 구분된 값 (tsv)
- 일반 텍스트(txt)
- Excel file, Azure Table, Hive Table
- SQL Server
- SVMLight 데이터(.svmlight)
- 특성 관계 파일 형식(ARFF) 데이터(. Arff)
- Zip 파일(.zip)
- R 개체 또는 작업 영역 파일(.RData)

데이터 지원 유형

- 문자열
- 정수
- Double
- Boolean
- Datetime
- timespan

데이터 지원 용량

- 1.98 GB/file
- Total 10GB

45개 Walmart 부서별 주간 판매량

2010-02-05 ~ 2012-11-01

12열 X 8,190행

- feature.csv
  - 지점
  - 날짜
  - 온도
  - 연료비
  - 프로모션 * 5
  - 소비자 물가 지수
  - 실업률
  - 휴일 여부

5열 X 421,570행

- train.csv
  - 지점
  - 부서
  - 날짜
  - 판매량
  - 휴일 여부

3열 X 45행

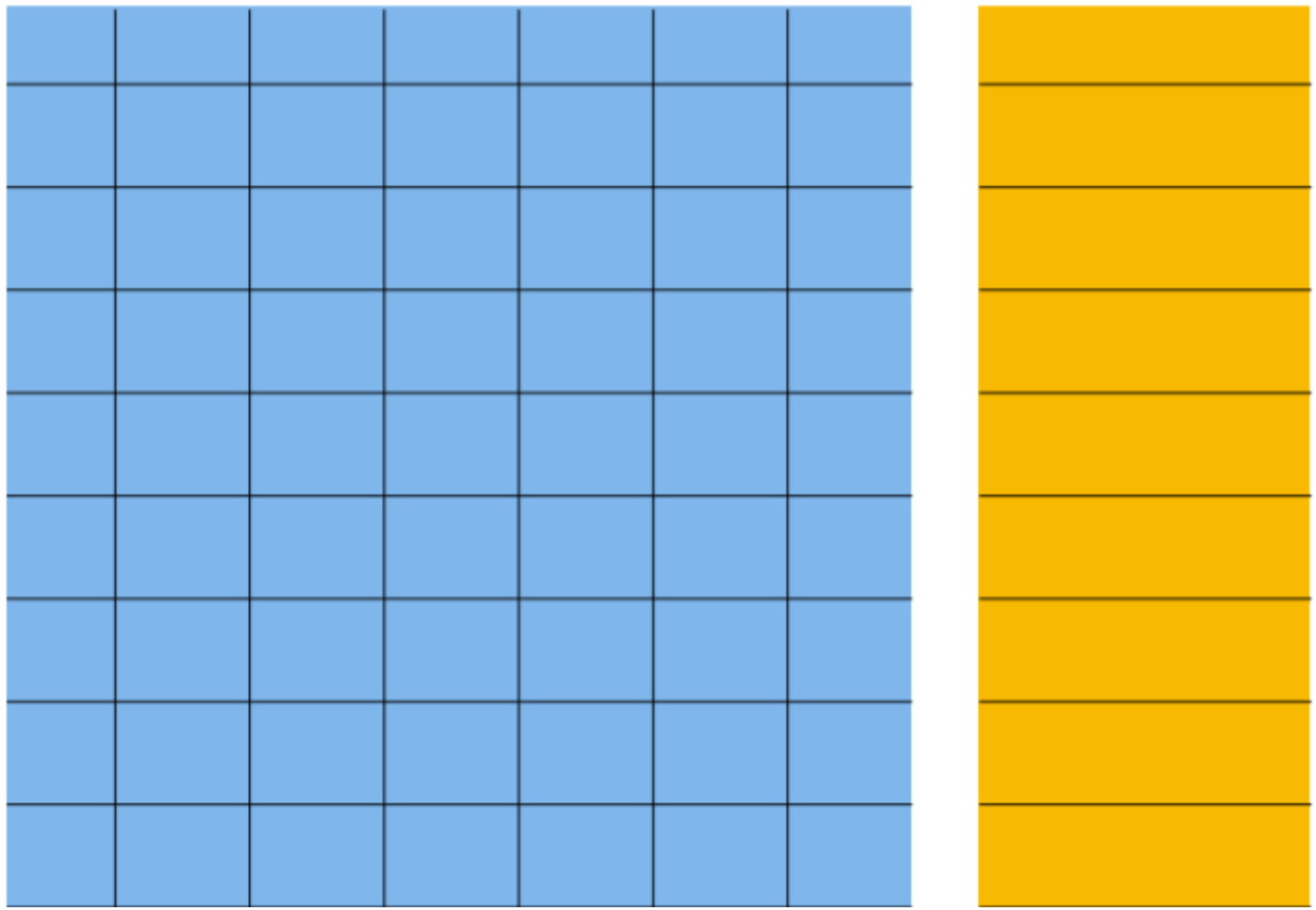
- stores.csv
  - 지점
  - 유형
  - 규모

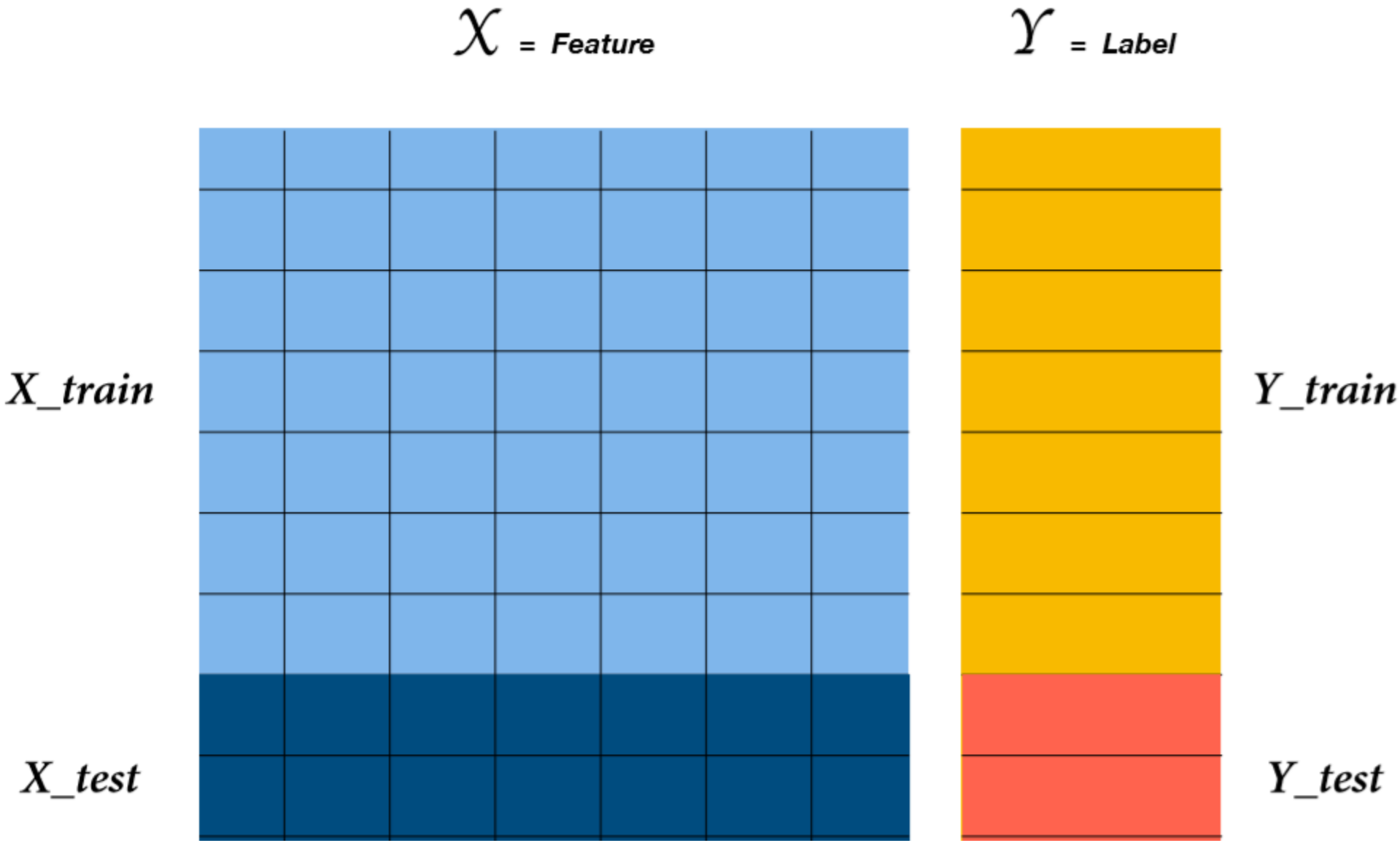
실습1:

Walmart data preprocessing

X = *Feature*

Y = *Label*

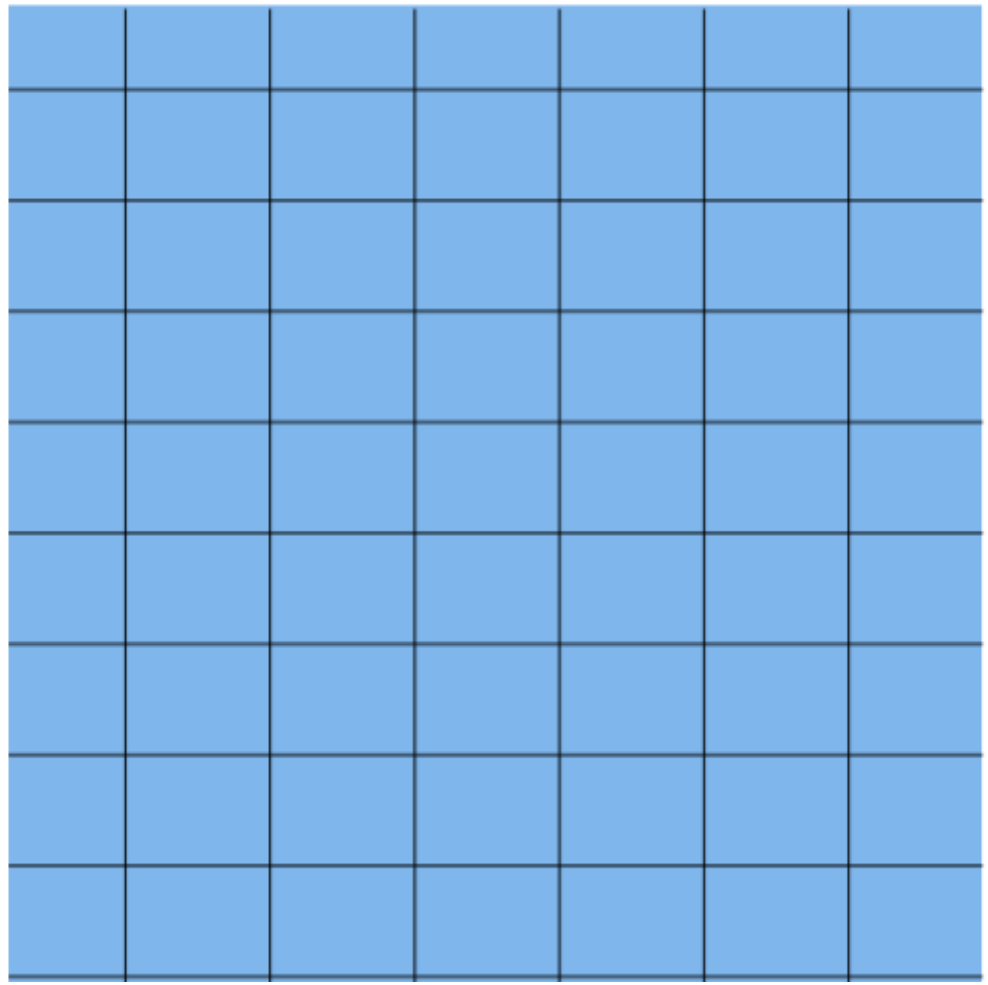



X = Feature
Y = Label
X_train
X_test
Y_train
Y_test

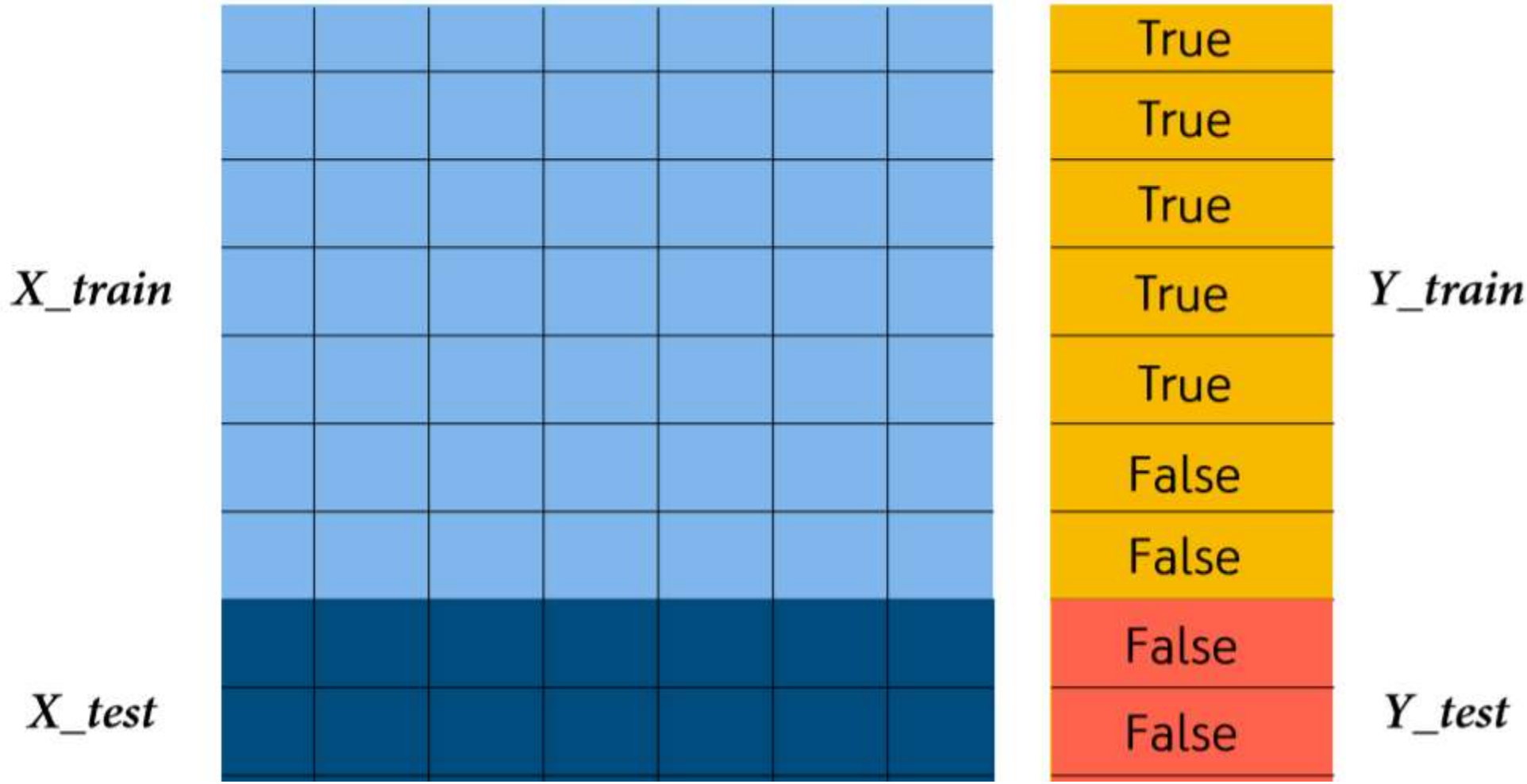
Random Split

X = *Feature*

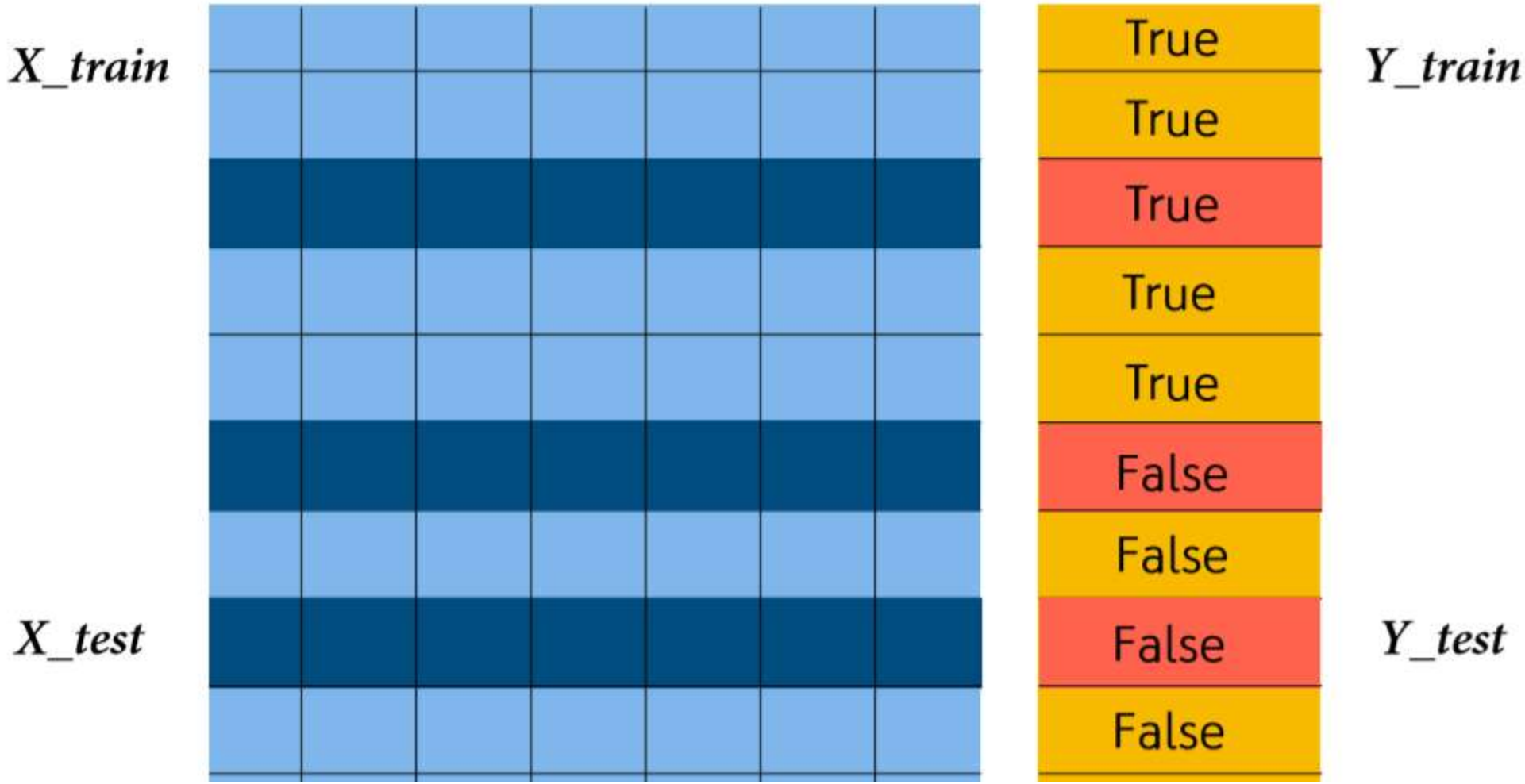
Y = *Label*



True
True
True
True
True
False
False
False
False


X_train
X_test
True
True
True
True
True
False
False
False
False
Y_train
Y_test

Random Split



Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points

Discovering structure

Three or more

Predicting categories

Predicting values

Two

ANOMALY DETECTION

- One-class SVM — >100 features, aggressive boundary
- PCA-based anomaly detection — Fast training

CLUSTERING

- K-means

MULTICLASS CLASSIFICATION

- Fast training, linear model — Multiclass logistic regression
- Accuracy, long training times — Multiclass neural network
- Accuracy, fast training — Multiclass decision forest
- Accuracy, small memory footprint — Multiclass decision jungle
- Depends on the two-class classifier, see notes below — One-v-all multiclass

REGRESSION

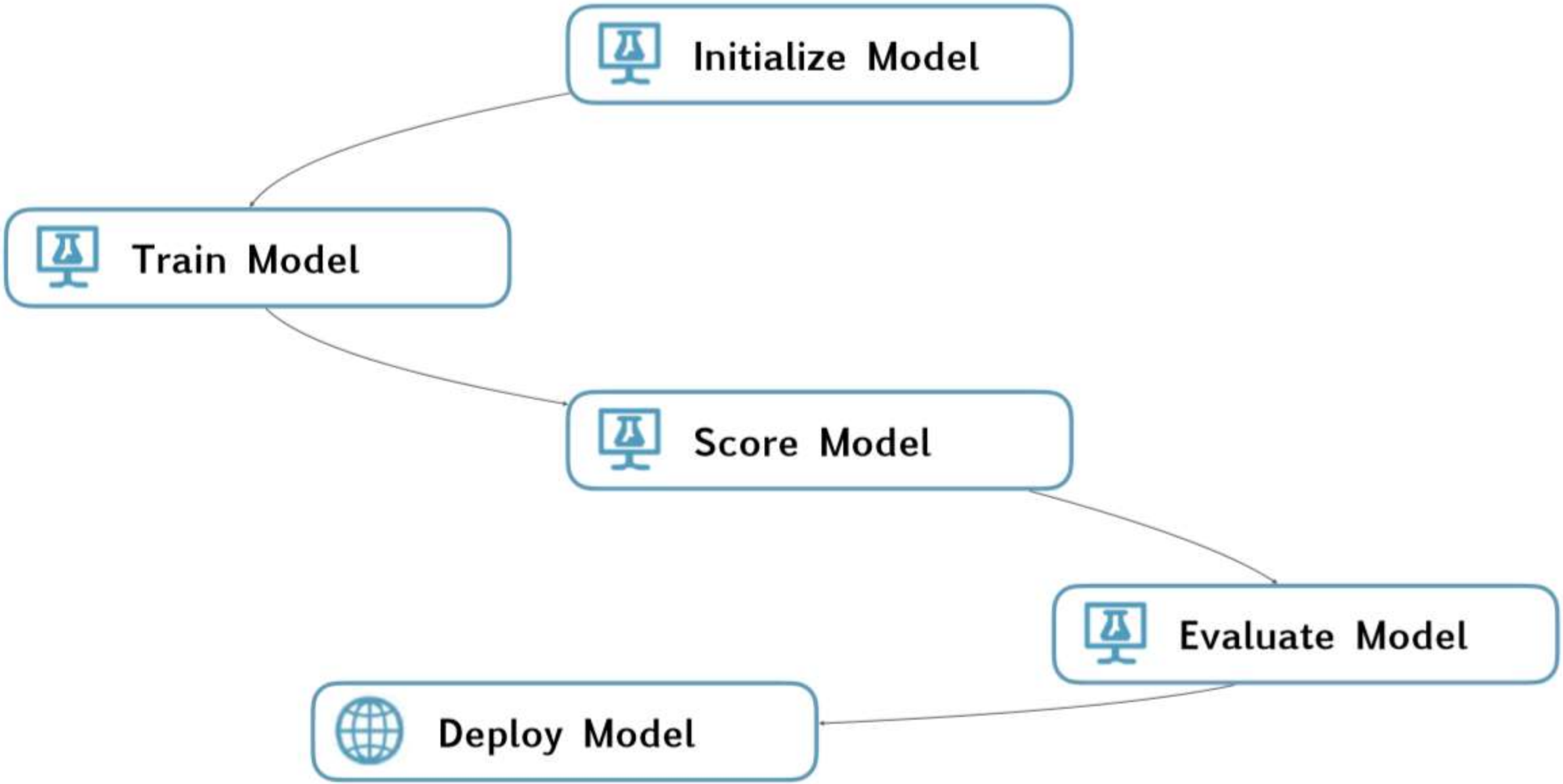
- Ordinal regression — Data in rank ordered categories
- Poisson regression — Predicting event counts
- Fast forest quantile regression — Predicting a distribution
- Linear regression — Fast training, linear model
- Bayesian linear regression — Linear model, small data sets
- Neural network regression — Accuracy, long training time
- Decision forest regression — Accuracy, fast training

TWO-CLASS CLASSIFICATION

- Two-class SVM — >100 features, linear model
- Two-class averaged perceptron — Fast training, linear model
- Two-class logistic regression — Fast training, linear model
- Accuracy, fast training — Two-class decision forest
- Accuracy, fast training — Two-class boosted decision tree
- Accuracy, small memory footprint — Two-class decision jungle
- >100 features — Two-class locally deep SVM

실습1:

Walmart Machine Learning


Initialize Model
Train Model
Score Model
Evaluate Model
Deploy Model

Microsoft Professional Program

감사합니다

Thank you~!