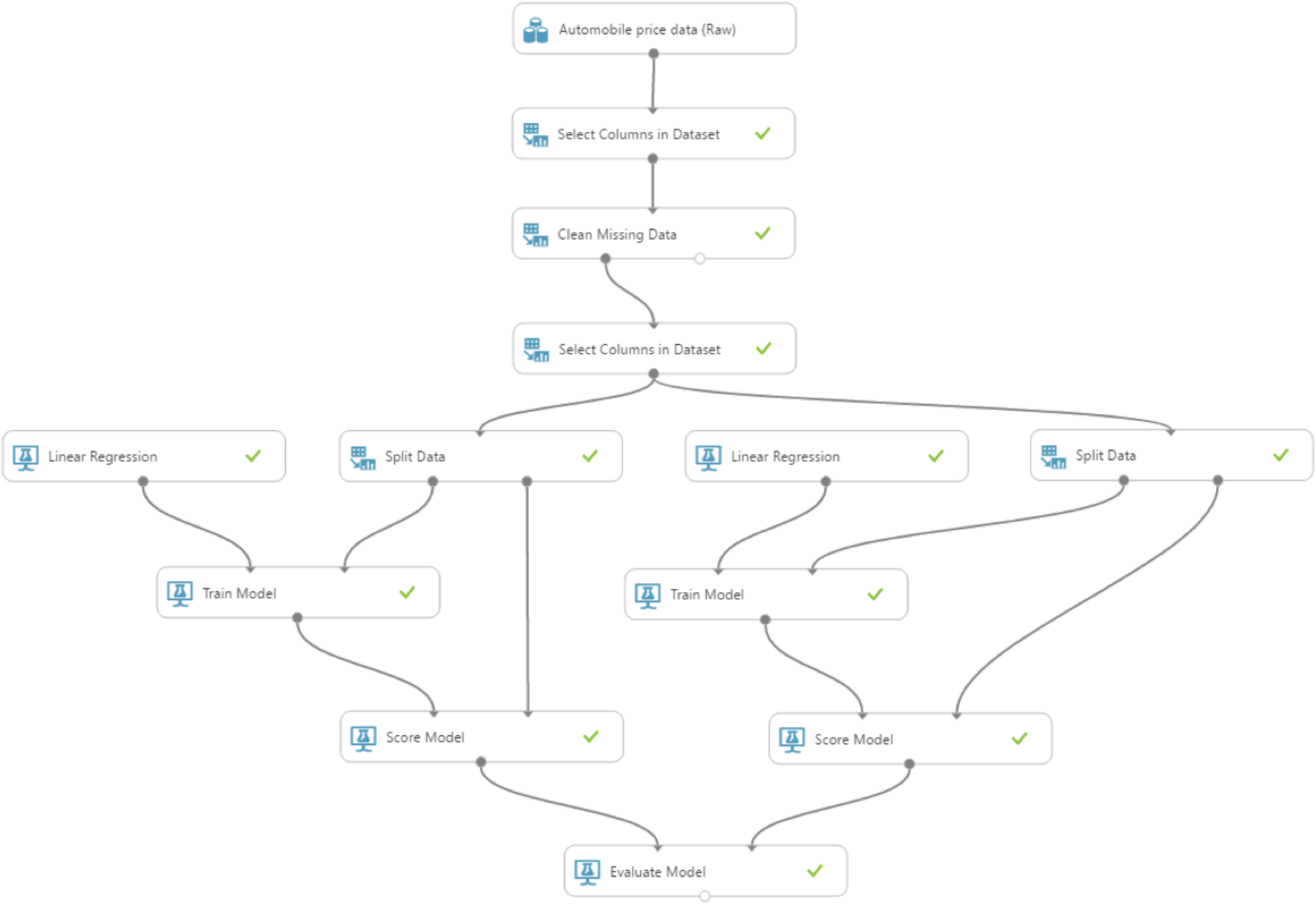
Machine Learning 기초 이론부터 Azure ML Studio 사용 실전기

3. Azure ML을 이용한 모델 제작 실습

백진헌 · 고려대학교
Microsoft Student Partner

Microsoft

Hello, Azure ML



5 steps of Machine Learning

데이터 수집

(1) 데이터 전처리
(2) 피처 정의

모델 만들기

(1) 러닝 알고리즘의 결정 및 적용

모델 테스트

(1) 새로운 데이터로 예측 실행

Session 3-1

들어주세요!

모델을 만드는 다양한 기법 소개

(1) 데이터 전처리
(2) 피처 정의

Session 3-2

따라해주세요!

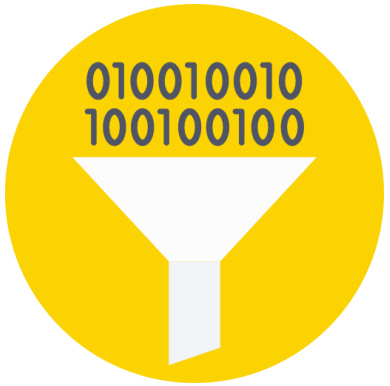
Machine Learning 학습 실험 제작

Session 3-1

모델을 만드는 다양한 기법 소개

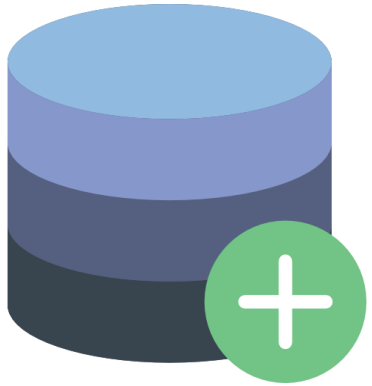
(1) 데이터 전처리
(2) 피처 정의

How to prepare Data?



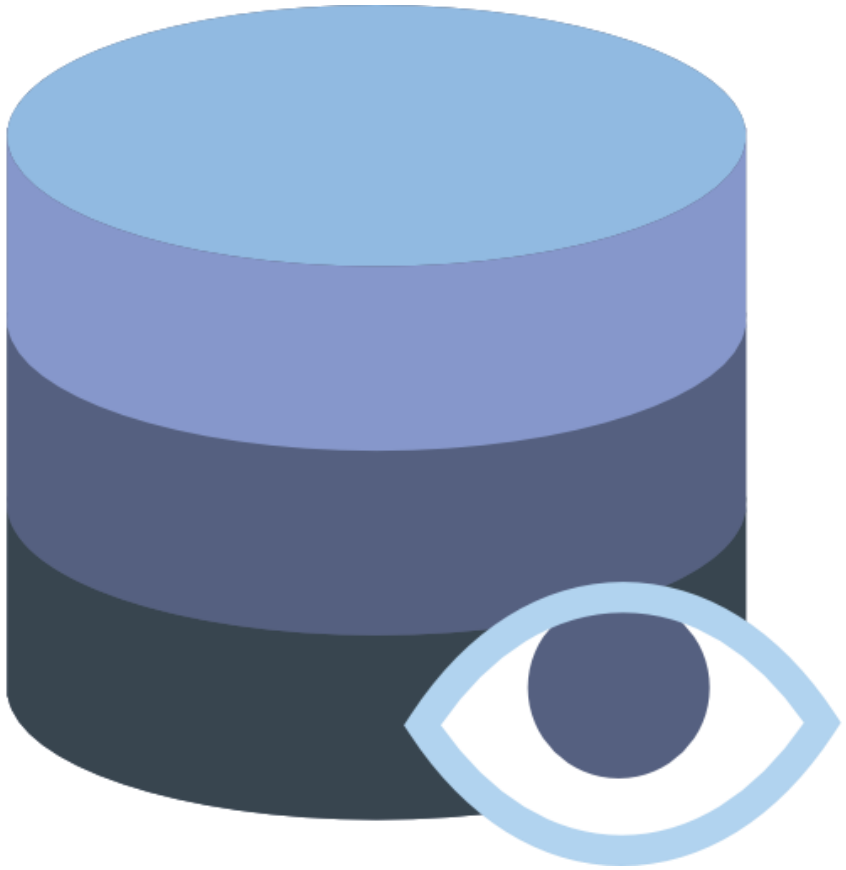
(1) 데이터 청소 및 처리

(2) 피처 선택



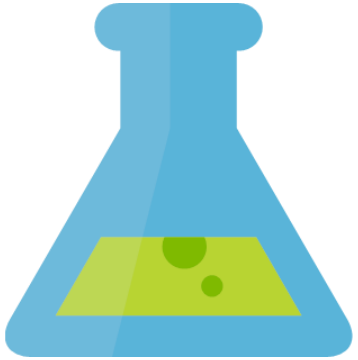
(3) 피처 공학

(1) 데이터 청소 및 처리



- 누락된 값
- 중복된 값
- 잘못된 값

(1) 데이터 청소 및 처리



Azure를 사용한 데이터 시각화

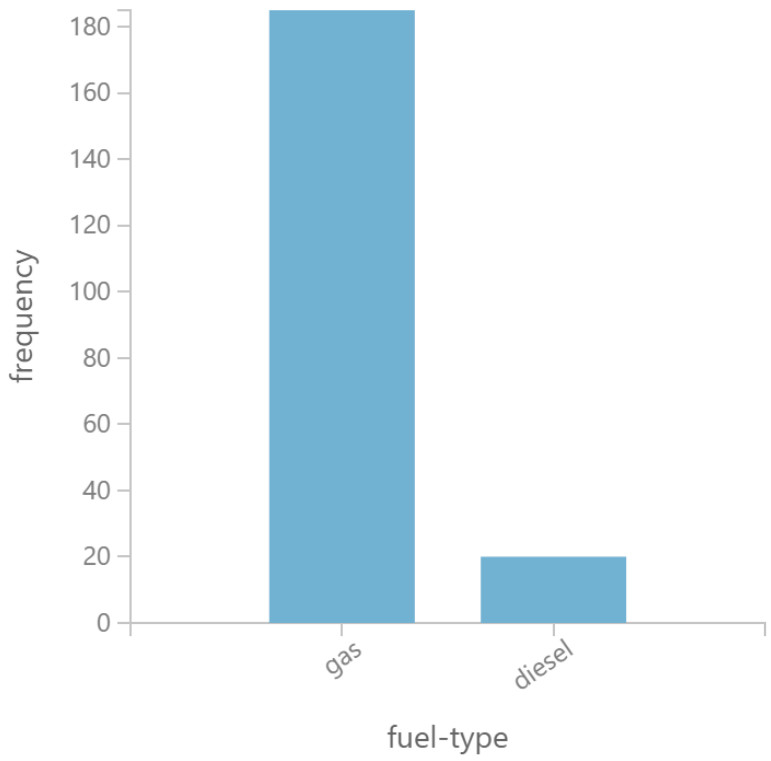
(1) 데이터 청소 및 처리

rows 205 columns 26

view as

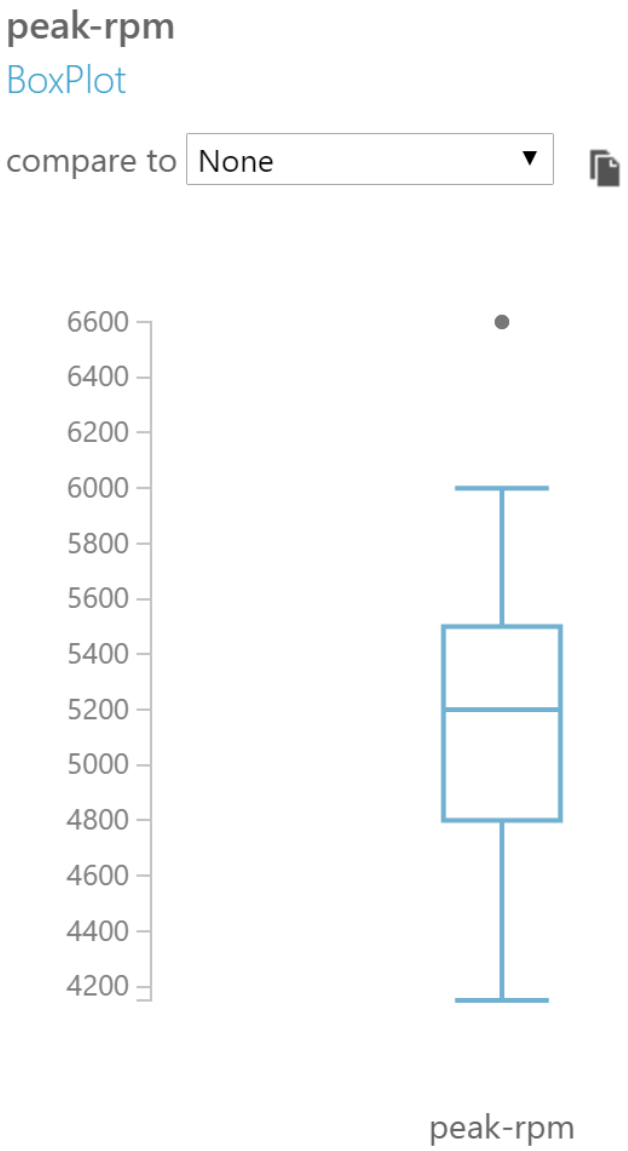
symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style
3		alfa-romero	gas	std	two	convertibl
3		alfa-romero	gas	std	two	convertibl
1		alfa-romero	gas	std	two	hatchback
2	164	audi	gas	std	four	sedan
2	164	audi	gas	std	four	sedan
2		audi	gas	std	two	sedan
1	158	audi	gas	std	four	sedan
1		audi	gas	std	four	wagon
1	158	audi	gas	turbo	four	sedan
0		audi	gas	turbo	two	hatchback

compare to None

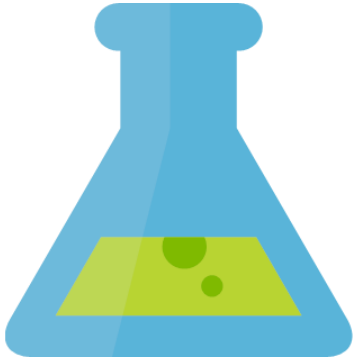


frequency log scale

(1) 데이터 청소 및 처리

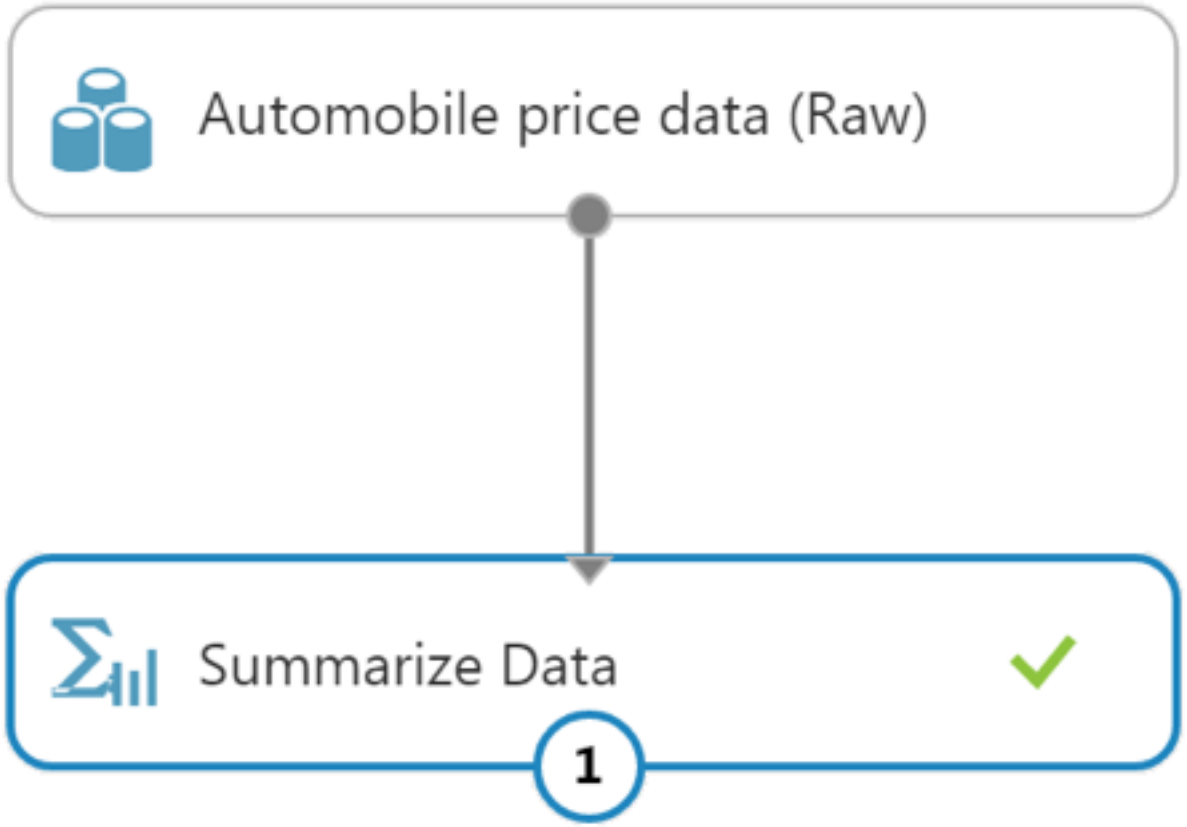

peak-rpm
BoxPlot
compare to
None
6600
6400
6200
6000
5800
5600
5400
5200
5000
4800
4600
4400
4200
peak-rpm

(1) 데이터 청소 및 처리



Azure를 사용한 데이터 분석

(1) 데이터 청소 및 처리



(1) 데이터 청소 및 처리

*누락된 값 처리

(1) 누락된 값들을 지정한 값으로 치환

(2) 누락된 값을 계산된 값으로 치환 (평균, 중앙값 등)

(3) 누락된 값이 있는 행 또는 열 제거

(1) 데이터 청소 및 처리

Clean Missing Data Module (Demo)

(1) 데이터 청소 및 처리

중복된 데이터 제거

(1) 데이터 청소 및 처리

Remove Duplicate Rows Module

(1) 데이터 청소 및 처리

*경계를 넘어가는 값 처리

(1) ClipPea: 상위 경계를 넘어서는 값을 찾아 깎아내거나 대치

(2) ClipSubpeaks

(3) ClipPeaksAndSubpeaks

(1) 데이터 청소 및 처리

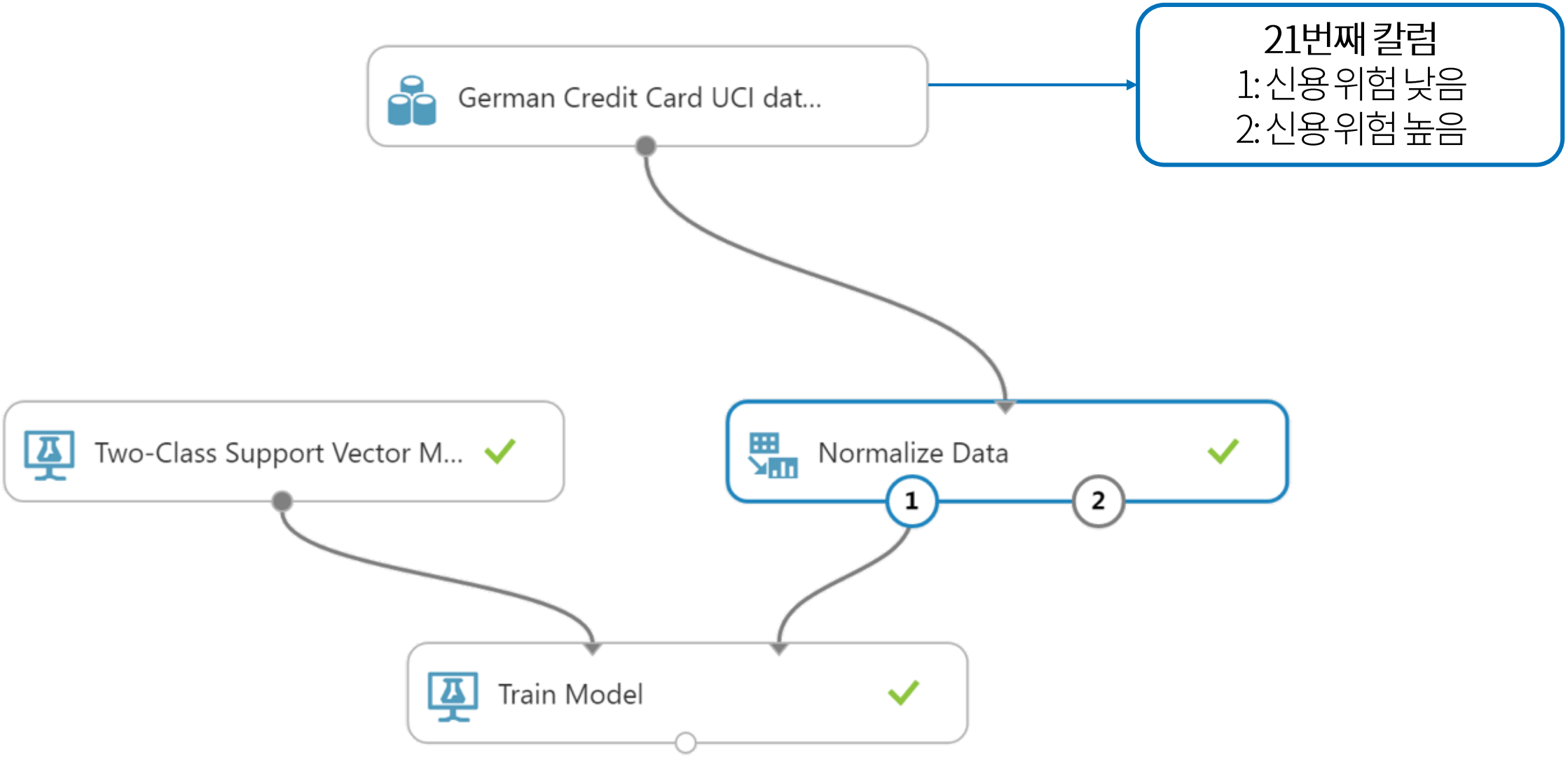
Clip Values Module
(Demo)

(1) 데이터 청소 및 처리

*피처 정규화

- ✓ 누락된 값 대치, 아웃라이어와 중복 레코드 제거 작업 완료?
- ✓ 데이터가 일관된 형태를 가질 수 있게끔 정규화

(1) 데이터 청소 및 처리

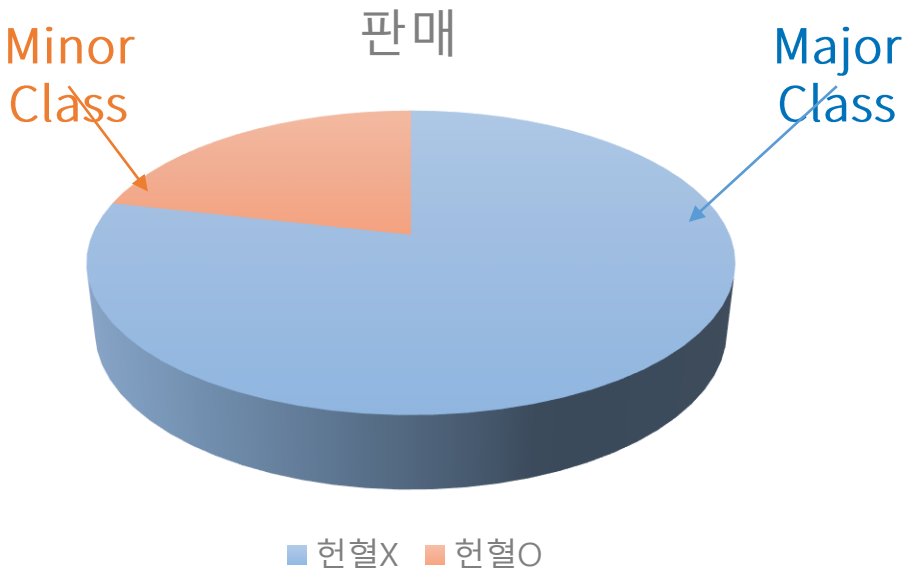

German Credit Card UCI dat...
21번째 칼럼
1: 신용 위험 낮음
2: 신용 위험 높음
Two-Class Support Vector M...
Normalize Data
1
2
Train Model

(1) 데이터 청소 및 처리

Normalize Data Module
(Demo)

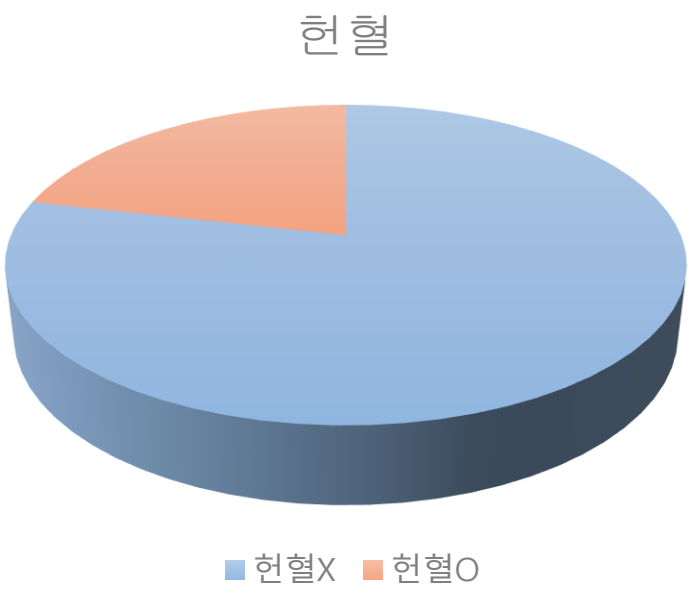
(1) 데이터 청소 및 처리

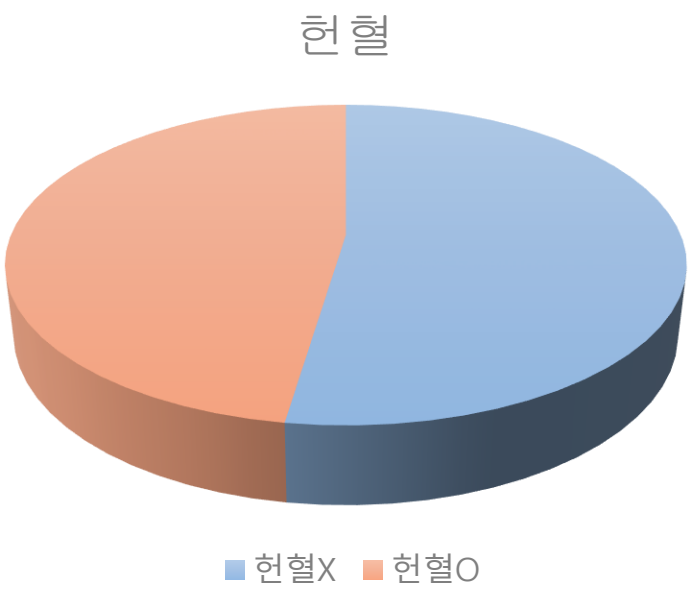
*클래스의 불균형 처리: SMOTE



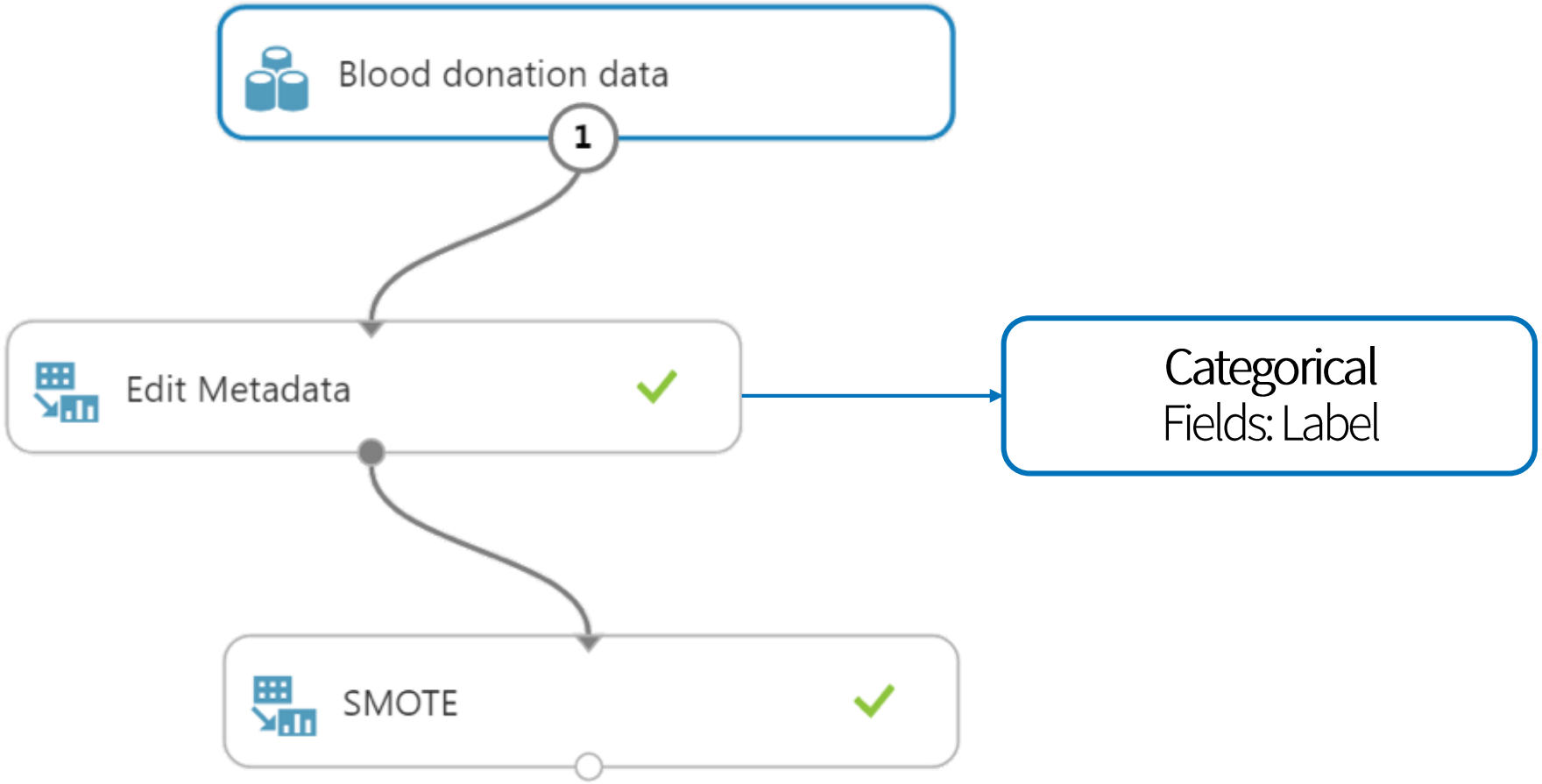
(1) 데이터 청소 및 처리

*클래스의 불균형 처리: SMOTE

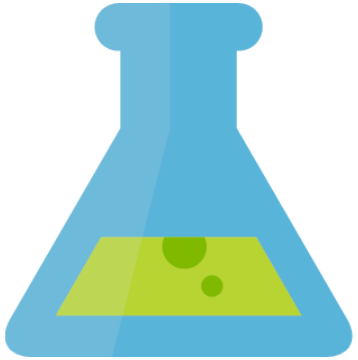




(1) 데이터 청소 및 처리


Blood donation data
1
Edit Metadata
Categorical
Fields: Label
SMOTE

(1) 데이터 청소 및 처리



SMOTE Module
(Demo)

(2) 피처 선택

가장 의미있는 정보를 피처로 선택

(2) 피처 선택

*필터링 (Filtering)

✓ 중복되었거나 정보를 제공하지 못하는 피처를 제거

*랩퍼 (Wrapper)

✓ 피처 선택을 위해 분류 모델을 이용 (의사결정 트리)

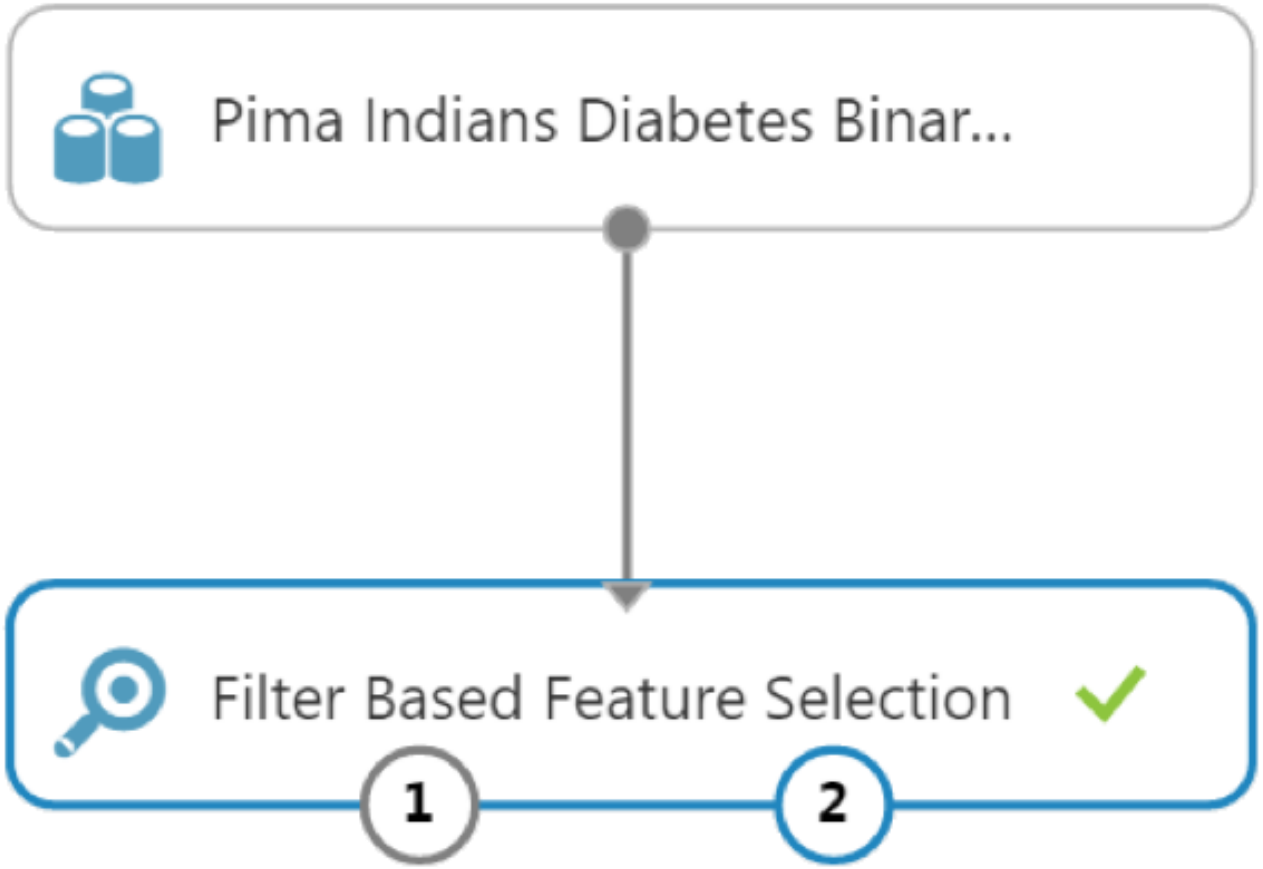
(2) 피처 선택

*필터링 (Filtering)

✓ 피어슨 상관 계수 (Pearson Correlation)

$$-1 \leq \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \leq 1$$

(2) 피처 선택


Pima Indians Diabetes Binar...
Filter Based Feature Selection
1
2

(2) 피처 선택

Filter Based Feature Selection Module
(Demo)

(3) 피처 엔지니어링

기존에 존재하는 **피처 셋**을 이용하여,
새로운 정보를 제공하는 피처 추가

(3) 피처 엔지니어링

출발지-도착지의 위도와경도, 육상 혹은 항공

➥ 출발지와 도착지 사이의 거리

스타벅스 기프티권

(3) 피처 엔지니어링

- ✓ 데이터를 통에 넣기 (데이터 양자화)
- ✓ 차원을 줄이기

휴식시간 (10분)

Session 3-2

Hello, Machine Learning
(Mission)

5 steps of Machine Learning

데이터 수집

(1) 데이터 전처리
(2) 피처 정의

모델 만들기

(1) 러닝 알고리즘의 결정 및 적용

모델 테스트

(1) 새로운 데이터로 예측 실행

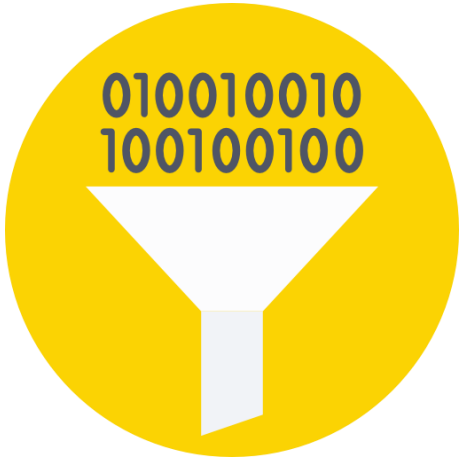
(1) 데이터 수집

Automobile Price Data

(1) 데이터 수집

같이 따라해봅시다!

(2) 데이터 전처리

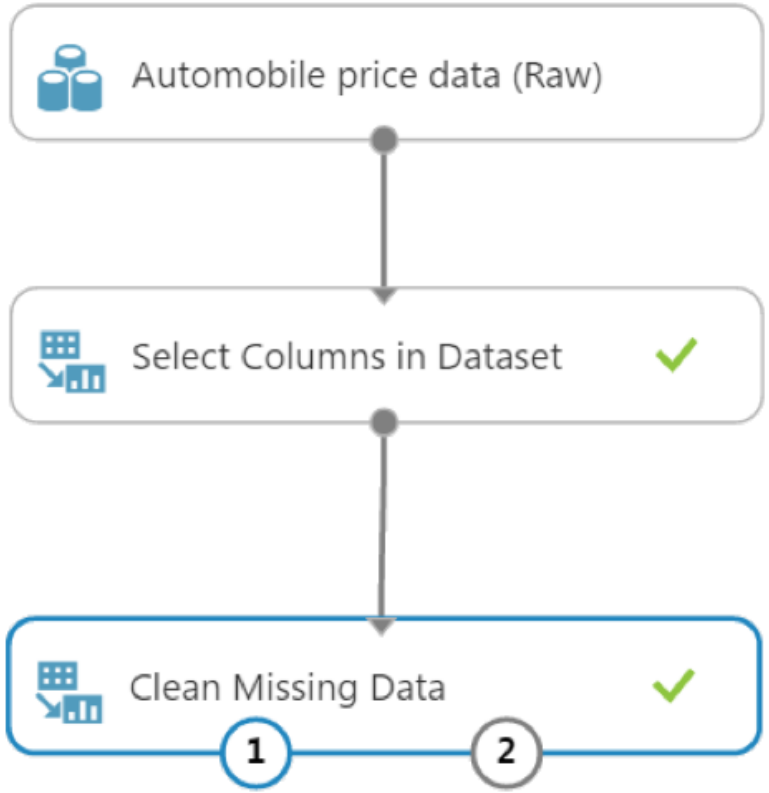


누락된 값이 너무 많은 행,
Normalized-losses 칼럼 제거

(2) 데이터 전처리

누락된 값 치환

(2) 데이터 전처리



Clean Missing Data

Columns to be cleaned

Selected columns:
Column type: Numeric, All

Launch column selector

Minimum missing value ratio

0

Maximum missing value ratio

1

Cleaning mode

Replace with mean

Cols with all missing values

Remove

Generate missing value indicator co...

Select Columns in Dataset

Select columns

Selected columns:
All columns
Exclude column names: normalized-losses

Launch column selector

(3) 피처 정의

가격을 예측할 때 주요하게 여겨지는 것들

(3) 피처 정의

make, body-style, wheel-base, engine-size,

horsepower, peak-rpm, highway-mpg, price

(4) 러닝 알고리즘의 결정과 적용

*분류

✓ 주어진 데이터 한 줄이 여러 분류 중 어디에 속하는지 분류

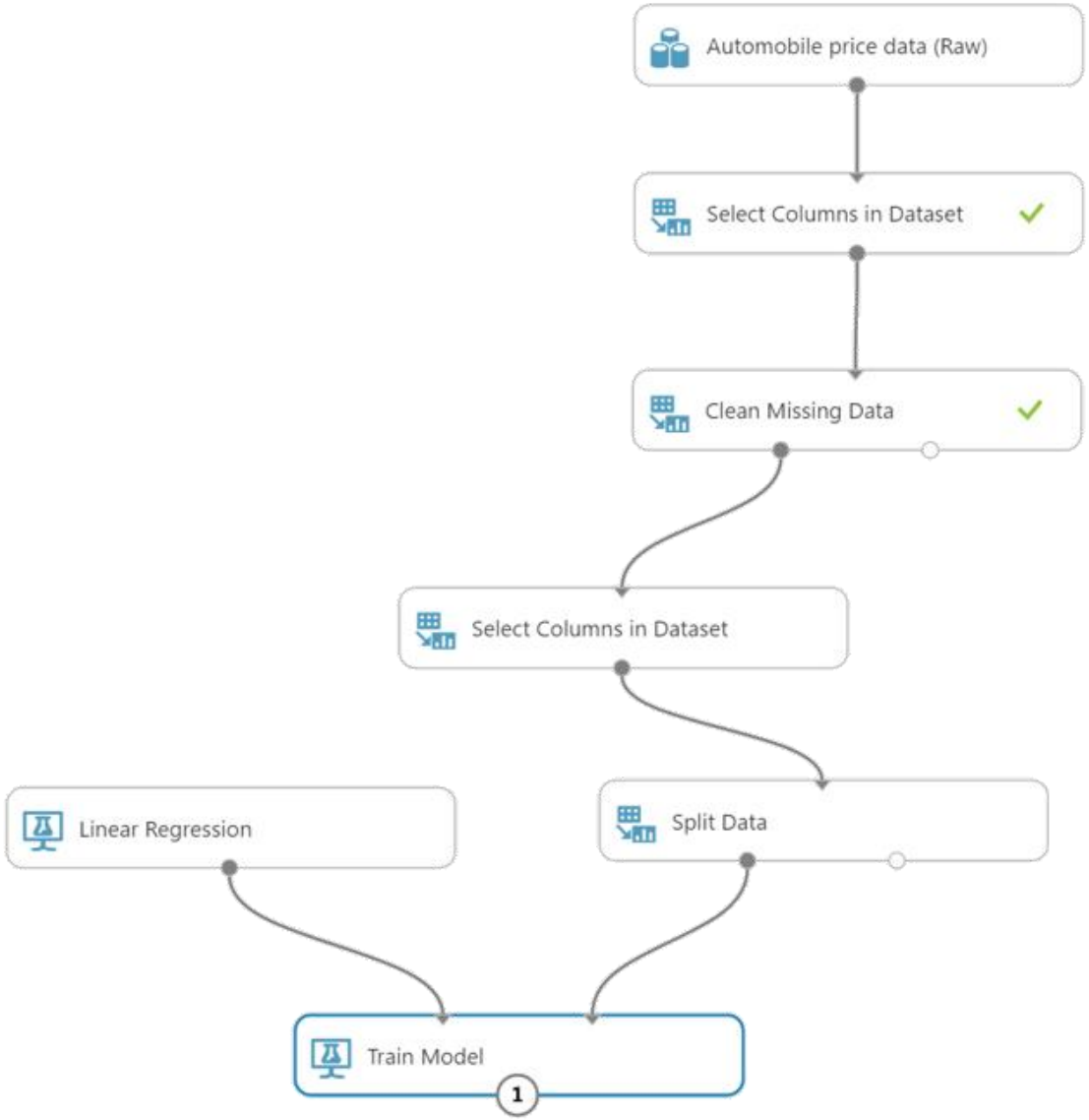
*회귀

✓ 자동차의 가격, 내일의 온도와 같은 연속된 출력 값을 예측

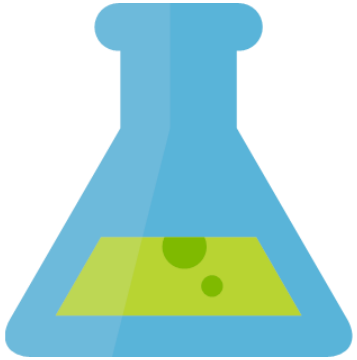
(4) 러닝 알고리즘의 결정과 적용

Linear Regression Module

(4) 러닝 알고리즘의 결정과 적용

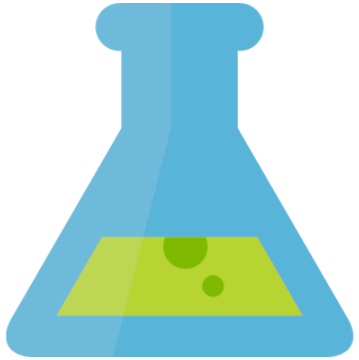


(5) 새로운 데이터로 예측 실행



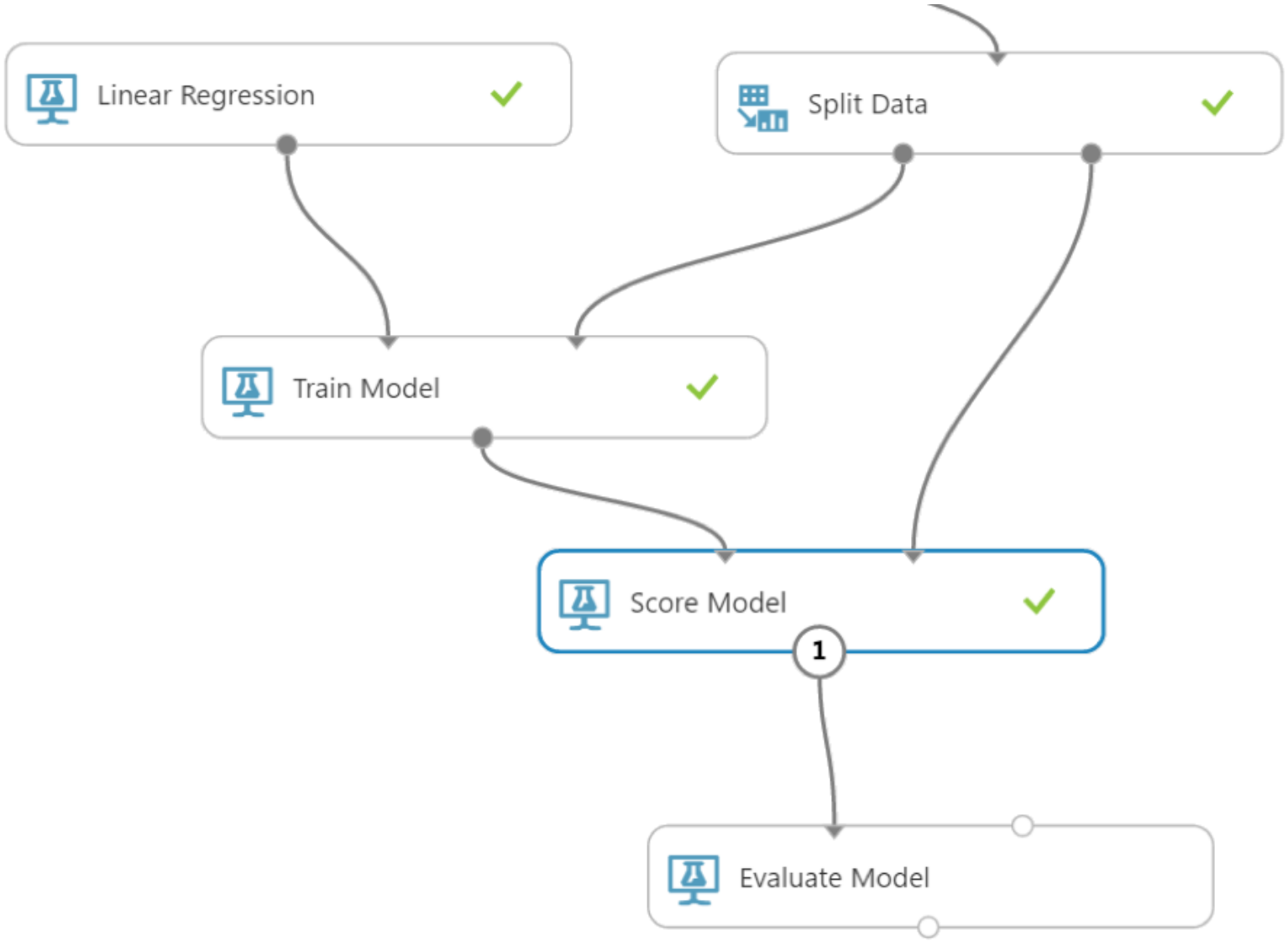
Score Model Module 사용
(Visualize를 선택하여 기존 값과 비교)

(5) 새로운 데이터로 예측 실행

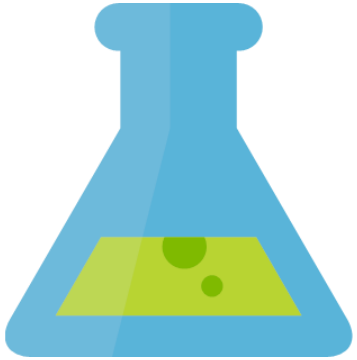


Evaluate Model Module 사용
(정확하게 예측했는지 통계 값 확인)

(5) 새로운 데이터로 예측 실행



(5) 새로운 데이터로 예측 실행



Evaluate Model Module 사용
(서로 다른 두 모델을 비교)

(6) 예측 실험 만들기를 통해 배포

*Mission

Evaluate Model Module
Coefficient of Determination가 가장 높은 사람

스타벅스 기프티권

Machine Learning 기초 이론부터 Azure ML Studio 사용 실전기

감사합니다

Microsoft