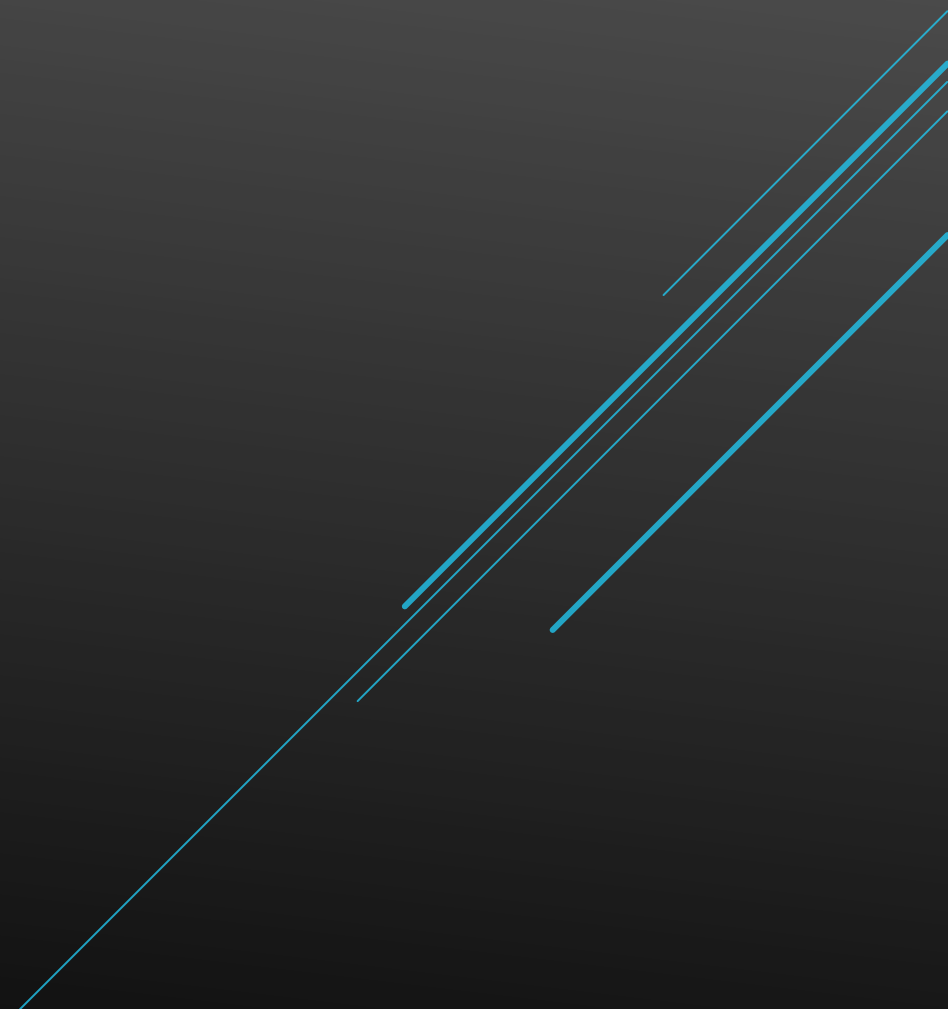
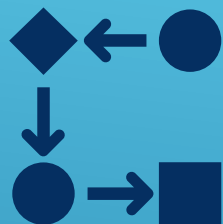


자동차 가격 예측

Azure Machine Learning Studio 활용

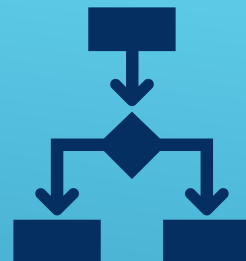




자동차와 관련된 26가지 지표와
가격 간의 상관관계를 분석하여
결과를 예측한다.



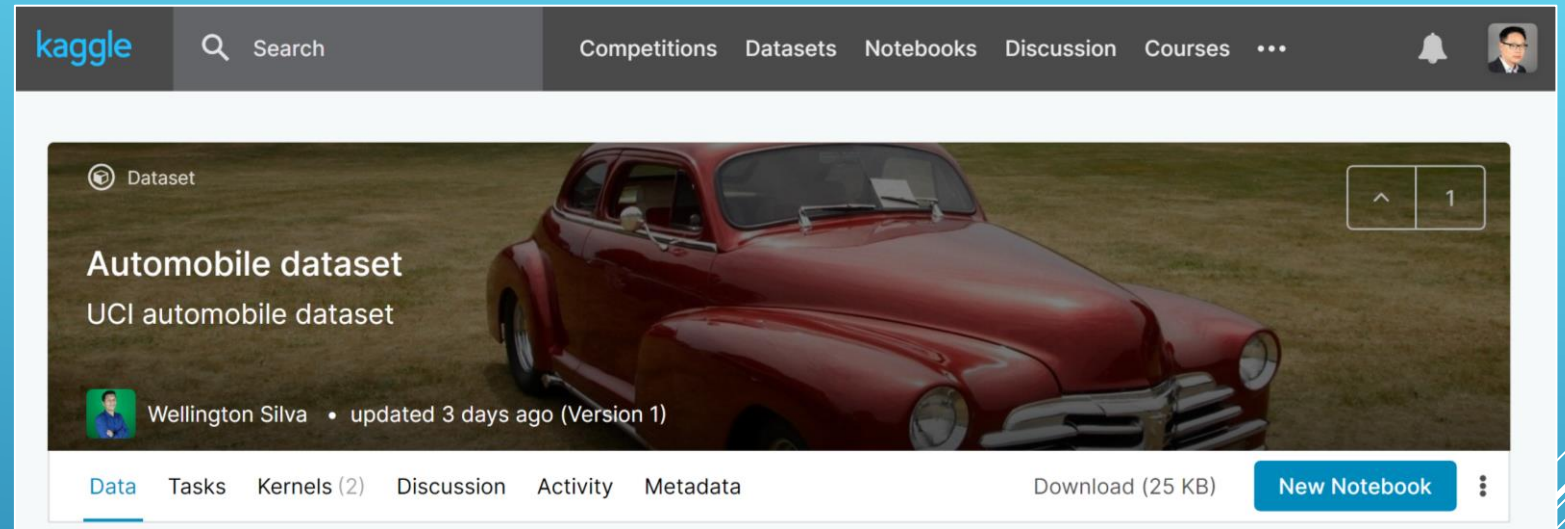
회귀 분석법에 대한 이론과 관련
모듈에 대한 지식을 습득한다.



예측 모델을 작성하여 웹에 배포한
후, 사용자가 직접 입력한 데이터에
대해 예측된 결과를 확인한다.

학습 목표

- ▶ UCI automobile dataset
- ▶ 26개 지표, 205개 데이터



분석할 데이터

Data Volume: 205 records, 26 variables

- | | |
|--|---|
| <p>Attribute Information</p> <ol style="list-style-type: none"> symboling: -3, -2, -1, 0, 1, 2, 3. normalized-losses: continuous from 65 to 256. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo fuel-type: diesel, gas. aspiration: std, turbo. num-of-doors: four, two. body-style: hardtop, wagon, sedan, hatchback, convertible. drive-wheels: 4wd, fwd, rwd. engine-location: front, rear. wheel-base: continuous from 86.6 to 120.9. length: continuous from 141.1 to 208.1. | <ol style="list-style-type: none"> width: continuous from 60.3 to 72.3. height: continuous from 47.8 to 59.8. curb-weight: continuous from 1488 to 4066. engine-type: dohc, dohc, l, ohc, ohcf, ohcv, rotor. num-of-cylinders: eight, five, four, six, three, twelve, two. engine-size: continuous from 61 to 326. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi. bore: continuous from 2.54 to 3.94. stroke: continuous from 2.07 to 4.17. compression-ratio: continuous from 7 to 23. horsepower: continuous from 48 to 288. peak-rpm: continuous from 4150 to 6600. city-mpg: continuous from 13 to 49. highway-mpg: continuous from 16 to 54. price: continuous from 5118 to 45400. |
|--|---|

Azure Machine Learning Service

Data -> Predictive model -> Operational web API in minutes

1. 데이터 취득: raw data를 모델에 추가하는 절차
2. 데이터 준비: 데이터를 학습에 용이하게 조작하는 절차
3. 모델 학습 및 평가: 실질적으로 학습을 시도하고 결과를 확인하는 절차
4. 웹 배포: 학습된 모델을 이용할 수 있도록 웹에 공개하는 절차

Blobs and Tables

Hadoop (HDInsight)

Relational DB (Azure SQL DB)

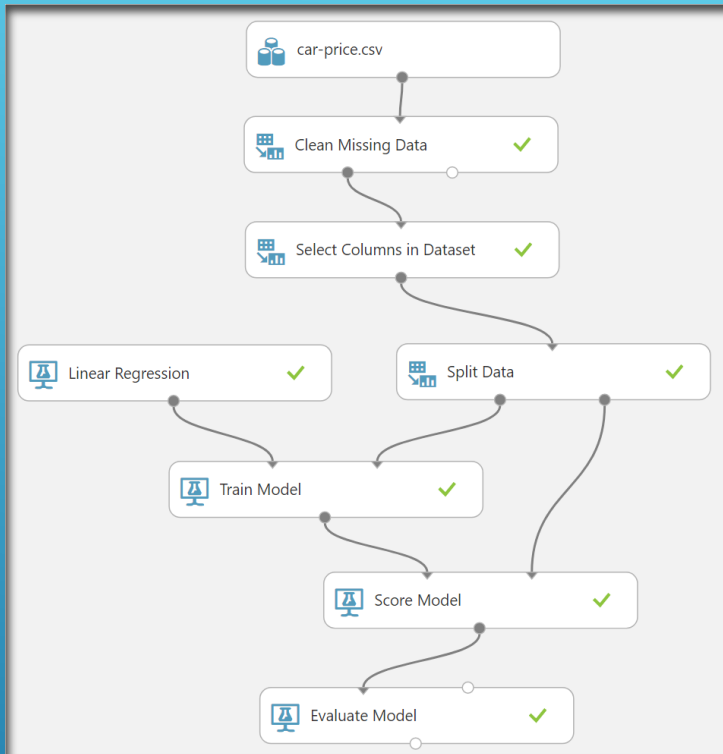
학습 절차

Integrated development environment
for Machine Learning

Model is now a web
service that is callable

Monetize the API through
our marketplace

Clients



- ▶ Clean Missing Data
- ▶ Select Columns in Dataset
- ▶ Linear Regression
- ▶ Train Model
- ▶ Score Model
- ▶ Evaluate Model

모델 개요도 및 신규 모듈 목록

- ① csv 파일을 업로드하고 새 실험을 생성한다.
- ② My Datasets 메뉴를 통해 csv 파일을 모듈로 추가한다.
- ③ 모듈을 우 클릭 후 visualize 기능을 통해 missing data를 확인한다.
- ④ Clean Missing Data 모듈을 추가하고 데이터셋에 연결한다.
- ⑤ RUN 기능 실행 후 visualize 기능을 통해 결과를 확인한다.

1. 데이터 취득 절차

CLEAN MISSING DATA 모듈



- ▶ 누락된 데이터를 정리하는 모듈.
- ▶ column selector를 이용해서 정리할 열을 선택.
- ▶ 최소, 최대 누락 값 비율 설정 가능(기본 0, 1).
- ▶ 사용자 지정 값으로 대체, 평균으로 대체, 중앙값으로 대체, 행 제거, 열 제거 등 다양한 정리 방법 설정 가능.

car-price > car-price.csv > dataset


rows
205

columns
26


view as




symboling




normalized-losses




make




fuel-type



aspiration



num-of-doors



3

alfa-romero

gas

std

two

3

alfa-romero

gas

std

two

1

alfa-romero

gas

std

two

2

164

audi

gas

std

four

2

164

audi

gas

std

four

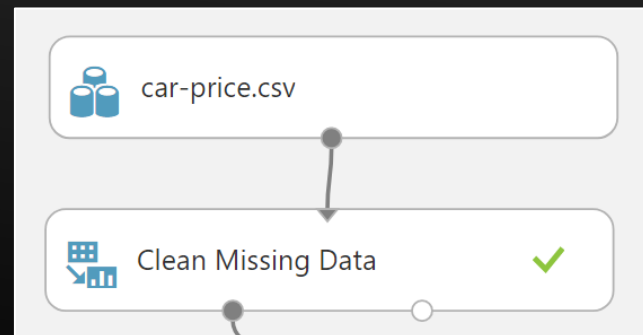
2

audi

gas

std

two



Cleaning mode

Custom substitution value ▼

Replacement value

0

☐ Generate missing val... ≡

- ① Select Columns in Dataset 모듈을 추가하고 연결한다.
- ② column selector로 bore, stroke를 제외한 모든 컬럼을 선택한다.
- ③ Split Data 모듈을 추가하고 분할 비율을 0.8로 설정한 후 연결한다.
- ④ RUN 기능을 수행한 후 visualize 기능을 통해 분할된 데이터셋을 확인.

2. 데이터 준비 절차

SELECT COLUMNS IN DATASET 모듈

- ▶ 데이터셋에서 특정 열을 선택/제외할 때 사용하는 모듈.
- ▶ 데이터셋을 입력받고 열 선택자를 통해 원하는 열을 선택/제외시킴.
- ▶ 선택된 열만 갖는 데이터셋을 출력.

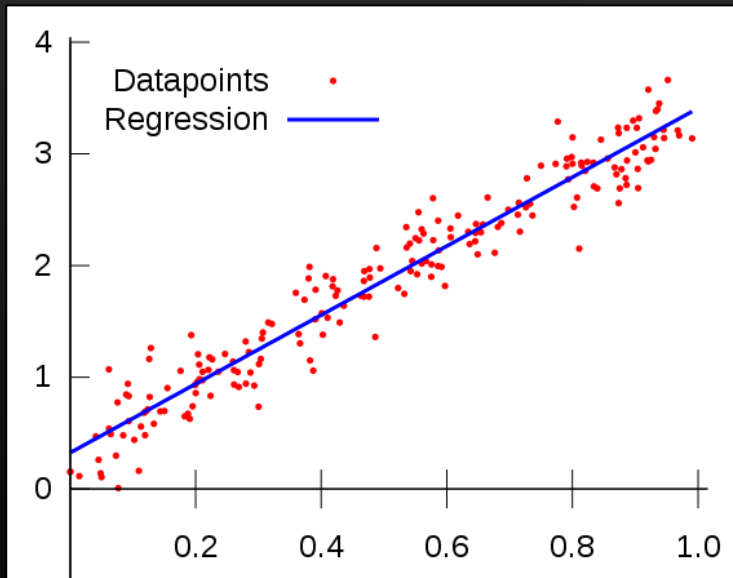
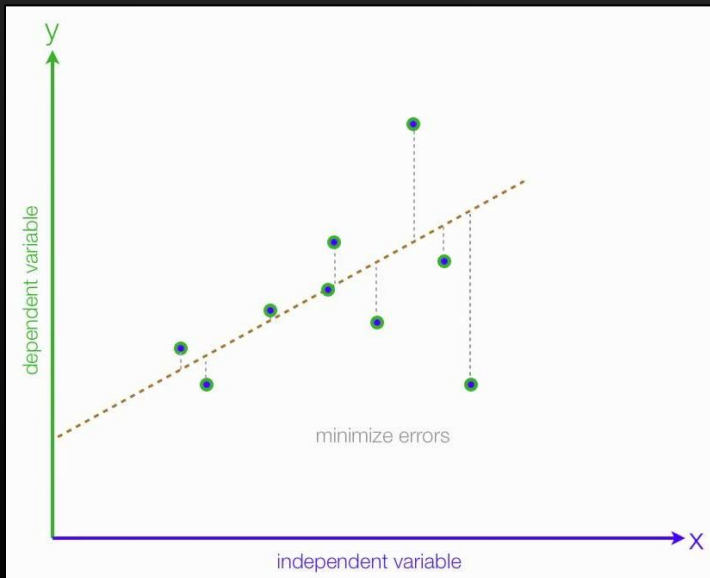
The screenshot shows the 'Select columns' module interface. It has a title bar with a close button (x). On the left, there are two tabs: 'BY NAME' and 'WITH RULES', with 'WITH RULES' being the active tab. The main area contains a checkbox labeled 'Allow duplicates and preserve column order in selection'. Below this, there is a 'Begin With' section with two buttons: 'ALL COLUMNS' (highlighted in blue) and 'NO COLUMNS'. Underneath, there is a row with a dropdown menu set to 'Exclude', another dropdown menu set to 'column names', and a text input field containing 'bore x' and 'stroke x'. To the right of the text input are '+' and '-' buttons. At the bottom right corner, there is a confirmation button with a checkmark icon.

- ① Linear Regression 모듈을 추가한다.
- ② Train Model 모듈을 추가하고 column selector로 price 컬럼을 선택.
- ③ 모듈을 연결하고 RUN 기능을 실행한 후 visualize 기능으로 확인한다.
- ④ Score Model 모듈을 추가하고 Trained model과 test dataset 연결.
- ⑤ RUN 기능을 수행한 후 결과를 시각화하여 확인한다.
- ⑥ Evaluate Model 모듈을 추가하고 Scored dataset을 연결한다.
- ⑦ RUN 기능을 수행한 후 결과를 시각화하여 확인한다.

3. 모델 학습 및 평가 절차

선형 회귀(LINEAR REGRESSION)

- ▶ 독립변수와 종속변수의 상관관계를 분석, 직선 그래프로 표현하는 방식.
- ▶ 독립변수(x)와 종속변수(y)를 2차원의 좌표평면 형태로 나타낸 뒤, 가장 오차가 적은 기울기와 y절편의 값을 구한다.
- ▶ 하나의 독립변수와 하나의 종속변수의 관계를 분석하는 단순 회귀, 여러 개의 독립변수와 하나의 종속변수의 관계를 분석하는 다중 회귀가 있음.

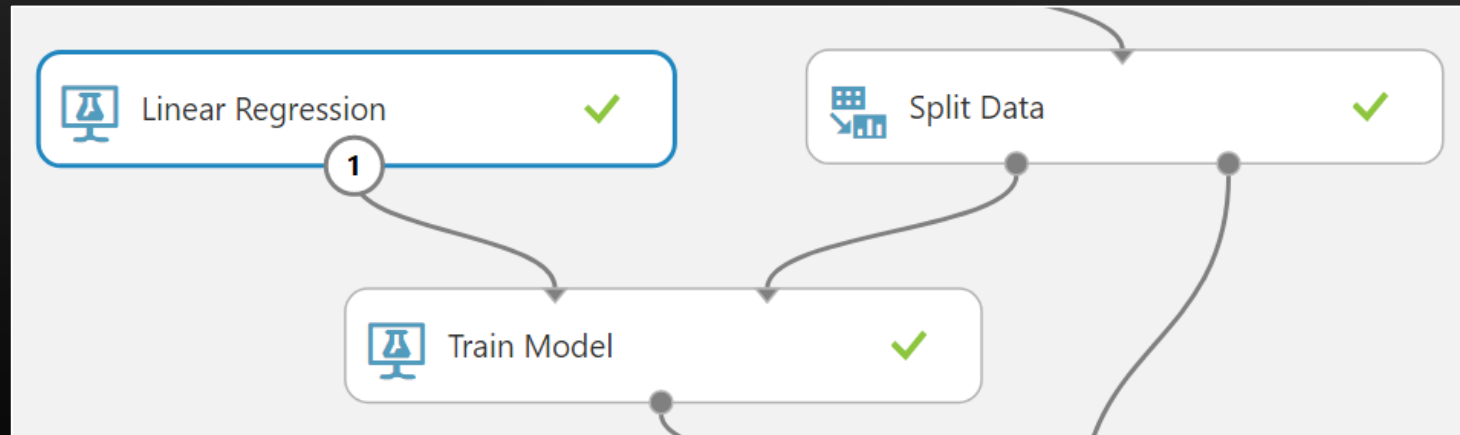


LINEAR REGRESSION 모듈

- ▶ Solution method: 회귀방정식을 구하는 방법. Online Gradient Descent(**경사하강법**), Ordinary Least Squares(**최소제곱법**) 두 종류.
- ▶ L2 regularization weight: L2 정규화 가중치. **이상값에 대한 가중치**를 설정하며 0의 경우 과적합이 발생.
- ▶ Include intercept term: 방정식의 절편 항을 확인하려면 체크.
- ▶ Random number seed: 동일한 실험에 대해 서로 다른 여러 실행을 할 때, **동일한 결과를 얻고자 할 때** 설정하는 seed 값.
- ▶ Allow unknown levels in categorical features: 선택 해제 시 누락 값이 오류를 발생시킴.



TRAIN MODEL 모듈

- ▶ 실질적으로 학습을 진행하는 모듈
- ▶ 학습 알고리즘 모델과 데이터셋을 입력받아 학습된 모델을 출력함.
- ▶ 사용자로부터 열을 선택받아 해당 열에 대해 학습을 진행.
- ▶ 군집화 학습은 **Train Clustering Model** 모듈, 회귀 또는 분류 학습은 **Train Model** 모듈을 사용

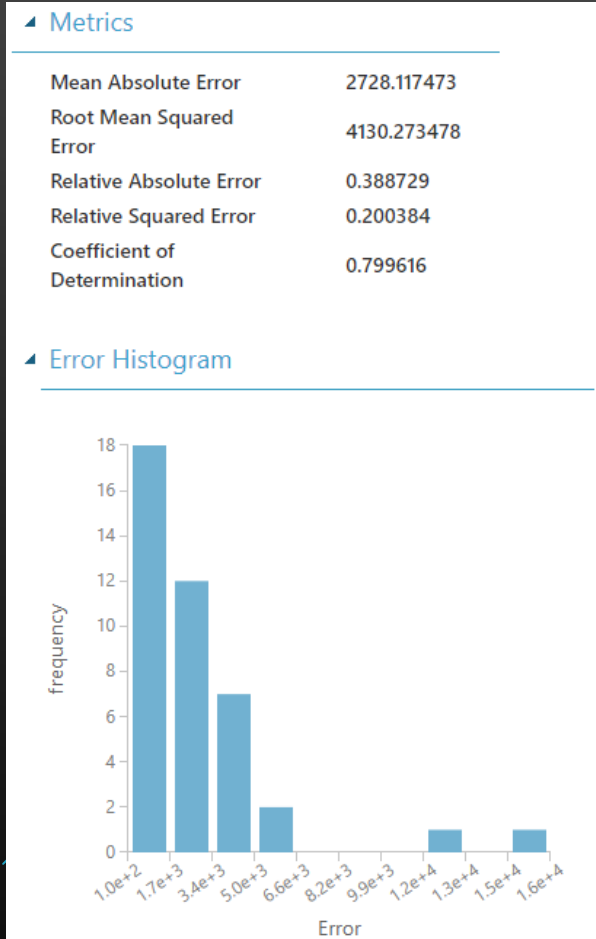


SCORE MODEL, EVALUATE MODEL 모듈

- ▶ 학습된 모델에 대해 각각 점수를 매기고 평가하는 모듈.
- ▶ Score Model에선 학습된 모델과 테스트셋을 입력받아 채점을 진행함.
- ▶ Evaluate Model에선 채점된 결과에 따라 오차를 분석함.

price	Scored Labels
	
13499	17502.986659
0	3169.04243
7898	3241.256779
13499	18196.221071
7738	6582.498686
34184	39631.296717

◀ Scored Model 시각화 결과 일부
Evaluate Model 시각화 결과 ▶



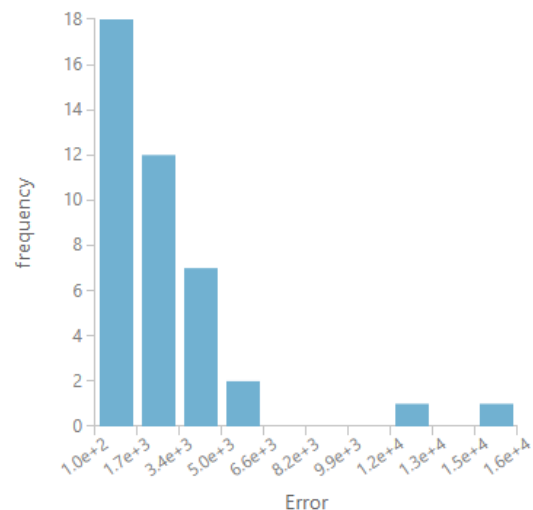
회귀 분석에서의 EVALUATE MODEL 지표

- ▶ Evaluate Model은 적용된 **학습 알고리즘의 종류에 따라 다른 결과 지표**를 나타냄.
- ▶ Mean Absolute Error(평균 절대 오차, MAE): 예측과 실제 결과 간의 차이(**오차**)들의 **평균**. 낮은 값일 수록 좋음.
- ▶ Root Mean Squared Error(평균 제곱근 오차, RMSE): **오차의 제곱의 평균에 제곱근을 적용**하여 표현한 지표. 정밀도를 표현하는데 적합.
- ▶ Relative Absolute Error(상대 절대 오차, RAE): 예측과 실제 결과 간의 상대적 절대 차이
- ▶ Relative Squared Error(상대 제곱근 오차, RSE): 오차의 제곱근의 총합을 실제값의 제곱근의 총합으로 나눈 결과
- ▶ Coefficient of Determination(결정계수): R^2 이라고도 하며 **예측 모델이 실제와 얼마나 잘 맞는지**를 0부터 1 사이의 수치로 표현한 값

Metrics

Mean Absolute Error	2728.117473
Root Mean Squared Error	4130.273478
Relative Absolute Error	0.388729
Relative Squared Error	0.200384
Coefficient of Determination	0.799616

Error Histogram



- ① SET UP WEB SERVICE 버튼에 마우스 커서를 올린 후 Predictive Web Service 메뉴를 클릭한다.
- ② RUN 후 DEPLOY WEB SERVICE 버튼을 클릭한다.
- ③ 새로운 창에서 Test 버튼 혹은 Test preview 텍스트를 클릭한다.
- ④ 샘플 데이터를 입력한 후 학습된 모델을 기반으로 예측된 결과를 확인.

4. 웹 배포 절차

WEB SERVICE TEST PREVIEW

- ▶ input 항목에 데이터를 입력한 후 **Test Request-Response** 버튼을 클릭하면 학습한 모델을 기반으로 예측 결과를 확인할 수 있음

input1		output1	
symboling	3	symboling	3
normalized-losses		normalized-losses	0
make	audi	make	audi
fuel-type	gas	fuel-type	gas
aspiration	std	aspiration	std
num-of-doors	four	num-of-doors	four
body-style	sedan	body-style	sedan
drive-wheels	fwd	drive-wheels	fwd
engine-location	front	engine-location	front
wheel-base	99.8	wheel-base	99.8
length	176.6	length	176.6
width	66.2	width	66.2
height	54.3	height	54.3
curb-weight	2337	curb-weight	2337
engine-type	ohc	engine-type	ohc
num-of-cylinders	four	num-of-cylinders	four
engine-size	109	engine-size	109
fuel-system	mpfi	fuel-system	mpfi
bore	3.19		
stroke	3.34		
compression-ratio	10	compression-ratio	10
horsepower	102	horsepower	102
peak-rpm	5500	peak-rpm	5500
city-mpg	24	city-mpg	24
highway-mpg	30	highway-mpg	30
price	13950	price	13950
Test Request-Response		Scored Labels	12991.0322265625