Microsoft Azure Machine Learning

4th Bigdata Research Team Seminar

Google-MS-Amazon 머신러닝 서비스 비교분석

2017.03.28

김도연

CONTENTS

1. Azure ML 소개
2. Azure ML 시작
3. Dataset 준비
4. Regression Model
   4.1 Regression Model 성능 비교
5. Classification Model
   5.1 Classification Model 성능 비교
6. Azure ML 장.단점
7. Amazon-Google-MS 머신러닝 서비스 비교분석

Azure ML 소개

- Azure Machine Learning(Azure ML)은 MS Azure 클라우드 기반의 Predictive Analytics 서비스

- 특징

  1) 완전한 관리 : H/W, S/W를 별도 구매할 필요 없음

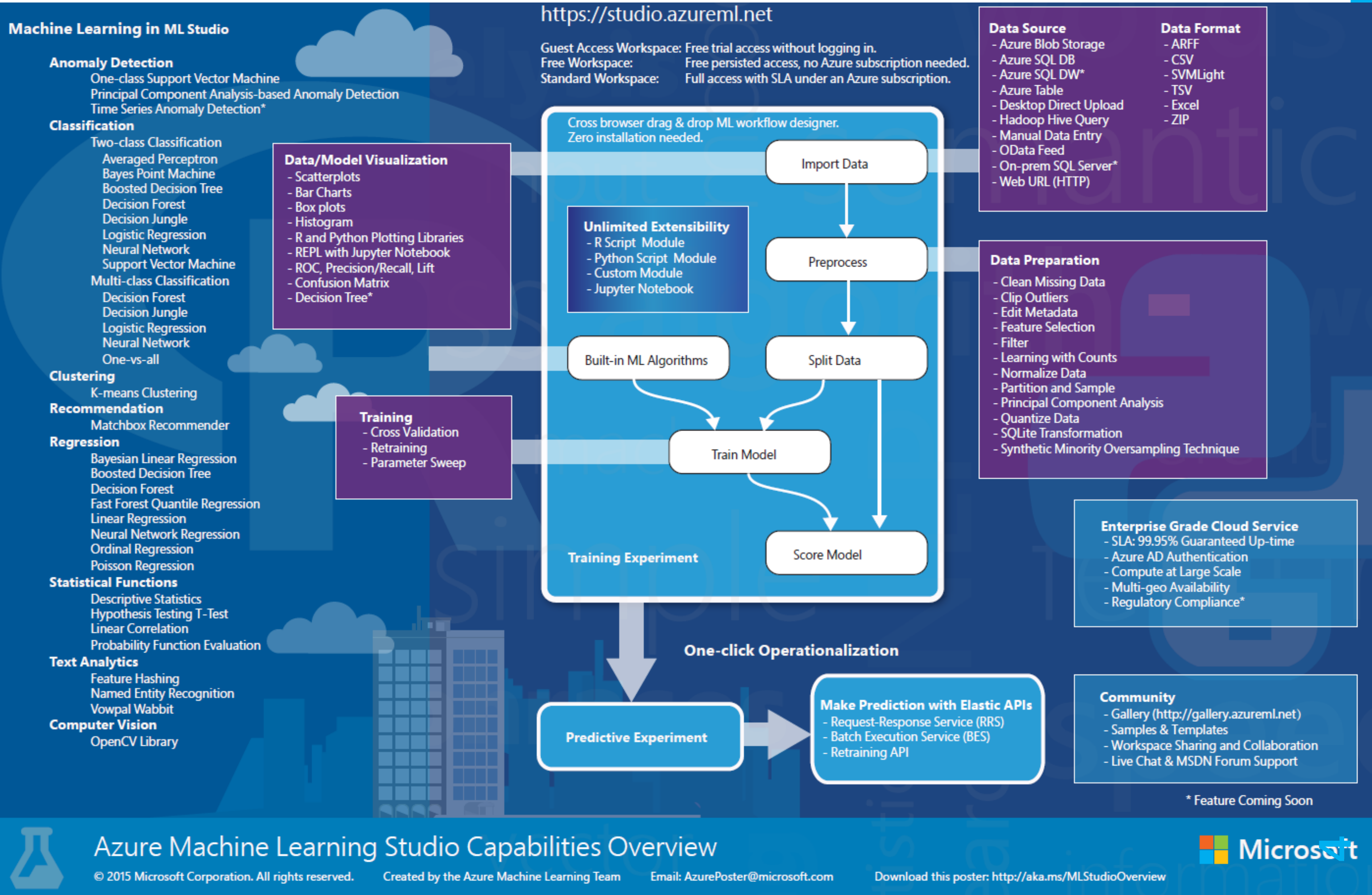
  2) 통합된 개발 : drag, drop, connect를 통해서 모델을 훈련 시킴

  -> 데이터 셋과 모듈을 시각적으로 연결하여 예측 분석 모델을 구성함

  3) 다양한 ML 라이브러리 제공

  4) R 및 Python 스크립트를 추가하여 확장 가능함

Azure ML Studio Overview



https://studio.azureml.net

Guest Access Workspace: Free trial access without logging in.
Free Workspace: Free persisted access, no Azure subscription needed.
Standard Workspace: Full access with SLA under an Azure subscription.

Machine Learning in ML Studio

Anomaly Detection
One-class Support Vector Machine
Principal Component Analysis-based Anomaly Detection
Time Series Anomaly Detection*

Classification
Two-class Classification
Averaged Perceptron
Bayes Point Machine
Boosted Decision Tree
Decision Forest
Decision Jungle
Logistic Regression
Neural Network
Support Vector Machine
Multi-class Classification
Decision Forest
Decision Jungle
Logistic Regression
Neural Network
One-vs-all

Clustering
K-means Clustering

Recommendation
Matchbox Recommender

Regression
Bayesian Linear Regression
Boosted Decision Tree
Decision Forest
Fast Forest Quantile Regression
Linear Regression
Neural Network Regression
Ordinal Regression
Poisson Regression

Statistical Functions
Descriptive Statistics
Hypothesis Testing T-Test
Linear Correlation
Probability Function Evaluation

Text Analytics
Feature Hashing
Named Entity Recognition
Vowpal Wabbit

Computer Vision
OpenCV Library

Data/Model Visualization
- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree*

Training
- Cross Validation
- Retraining
- Parameter Sweep

Cross browser drag & drop ML workflow designer.
Zero installation needed.

Unlimited Extensibility
- R Script Module
- Python Script Module
- Custom Module
- Jupyter Notebook

One-click Operationalization

Make Prediction with Elastic APIs
- Request-Response Service (RRS)
- Batch Execution Service (BES)
- Retraining API

Data Source
- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server*
- Web URL (HTTP)

Data Format
- ARFF
- CSV
- SVMLight
- TSV
- Excel
- ZIP

Data Preparation
- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

Enterprise Grade Cloud Service
- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance*

Community
- Gallery (http://gallery.azureml.net)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

* Feature Coming Soon

Azure Machine Learning Studio Capabilities Overview

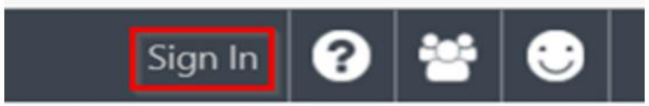
Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLStudioOverview

Microsoft

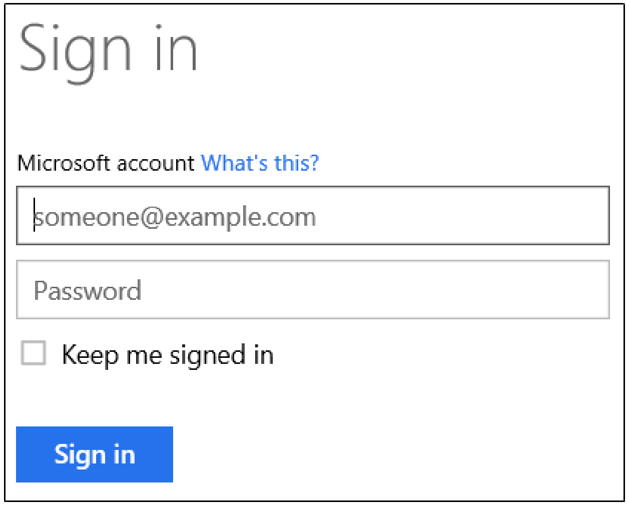
Azure ML 시작

- Azure ML 작업공간에 무료 계정 생성 후 로그인

1. 웹 브라우저 실행
2. http://studio.azureml.net 접속
3. 홈페이지의 위/오른쪽 코너의 Sign In 버튼 클릭



4. 마이크로소프트 계정 입력 후 Sign In 버튼 클릭



Sign in

Microsoft account What's this?

someone@example.com

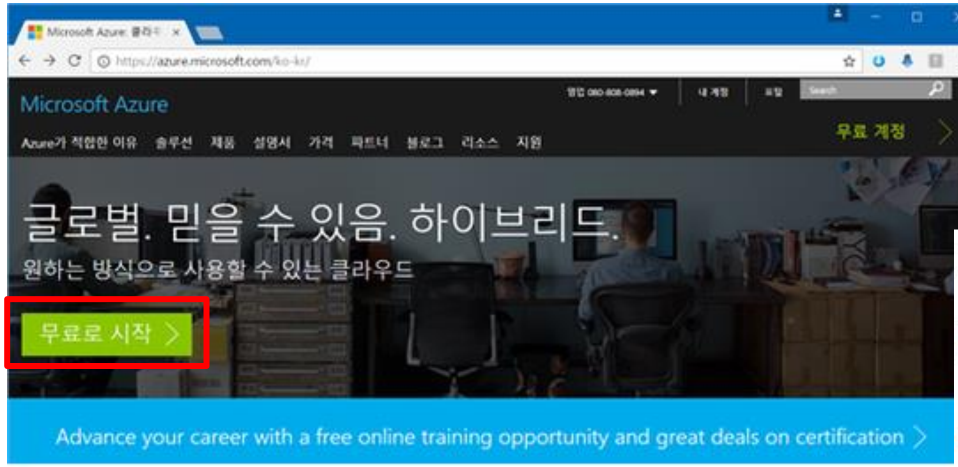
Password

□ Keep me signed in

Sign in

Azure ML 시작

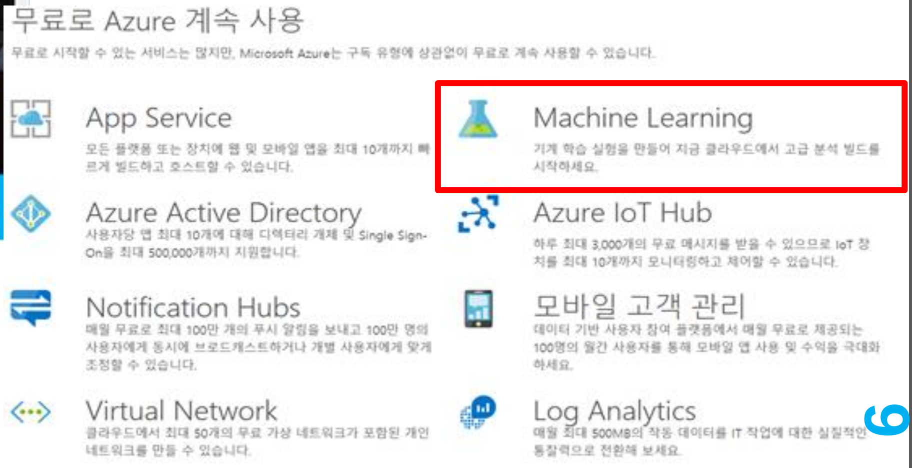
5. 만약, 마이크로소프트 계정이 없다면 …
   5.1 웹 브라우저(Internet Explorer) 실행
   5.2 Azure 관리 포탈 (https://azure.microsoft.com/ko-kr/) 접속
   5.3 '무료로 시작' 버튼 클릭 (10GB, 30일까지 무료)



글로벌. 믿을 수 있음. 하이브리드.

원하는 방식으로 사용할 수 있는 클라우드

Advance your career with a free online training opportunity and great deals on certification



무료로 Azure 계속 사용

무료로 시작할 수 있는 서비스는 많지만, Microsoft Azure는 구독 유형에 상관없이 무료로 계속 사용할 수 있습니다.

App Service
모든 플랫폼 또는 장치에 웹 및 모바일 앱을 최대 10개까지 빠르게 빌드하고 호스트할 수 있습니다.

Machine Learning
기계 학습 실험을 만들어 지금 클라우드에서 고급 분석 빌드를 시작하세요.

Azure Active Directory
사용자당 앱 최대 10개에 대해 디렉터리 개체 및 Single Sign-On을 최대 500,000개까지 지원합니다.

Azure IoT Hub
하루 최대 3,000개의 무료 메시지를 받을 수 있으므로 IoT 장치를 최대 10개까지 모니터링하고 제어할 수 있습니다.

Notification Hubs
매월 무료로 최대 100만 개의 푸시 알림을 보내고 100만 명의 사용자에게 동시에 브로드캐스트하거나 개별 사용자에게 맞게 조정할 수 있습니다.

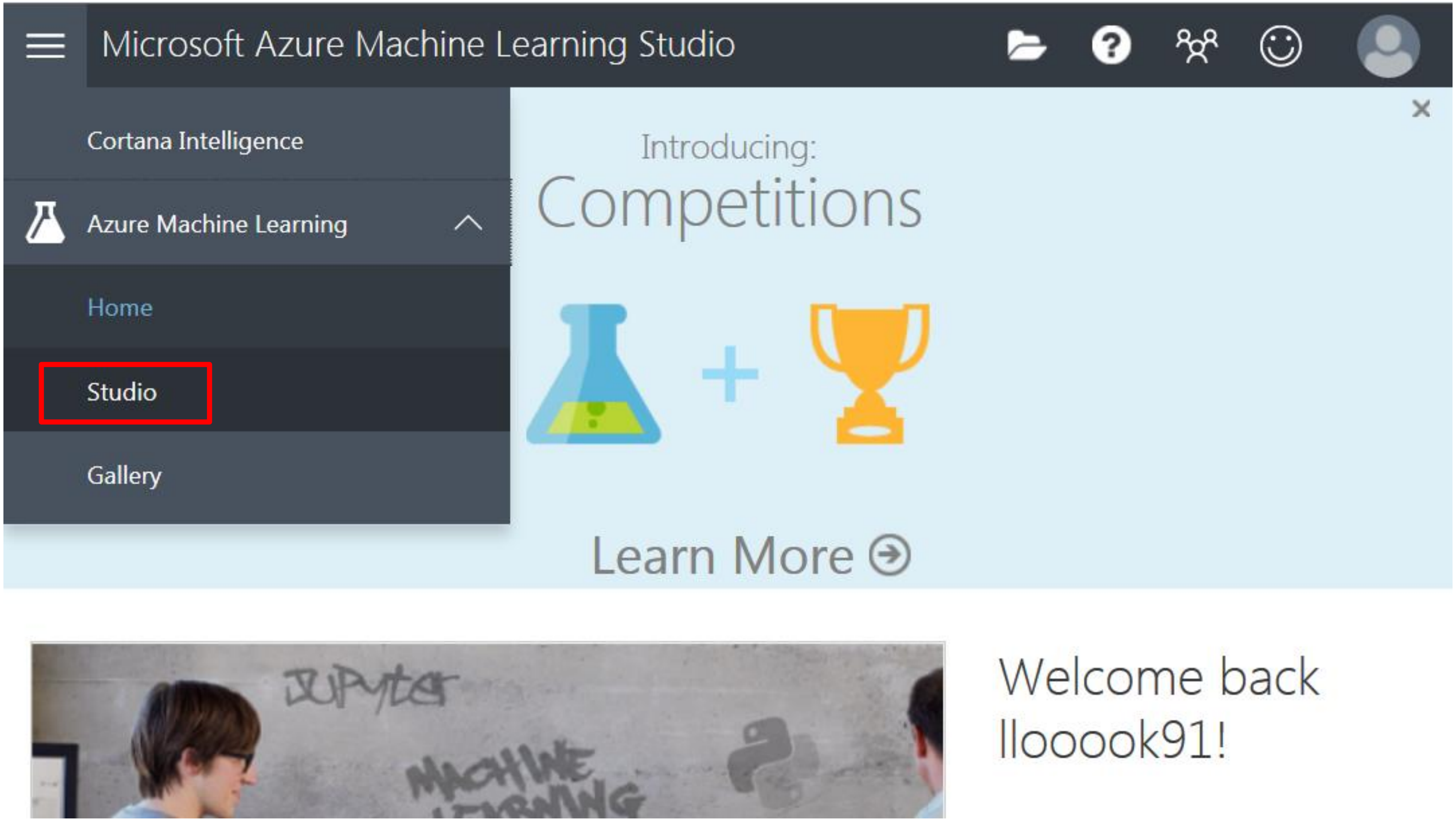
모바일 고객 관리
데이터 기반 사용자 참여 플랫폼에서 매월 무료로 제공되는 100명의 월간 사용자를 통해 모바일 앱 사용 및 수익을 극대화하세요.

Virtual Network
클라우드에서 최대 50개의 무료 가상 네트워크가 포함된 개인 네트워크를 만들 수 있습니다.

Log Analytics
매월 최대 500MB의 작동 데이터를 IT 작업에 대한 실질적인 통찰력으로 전환해 보세요.

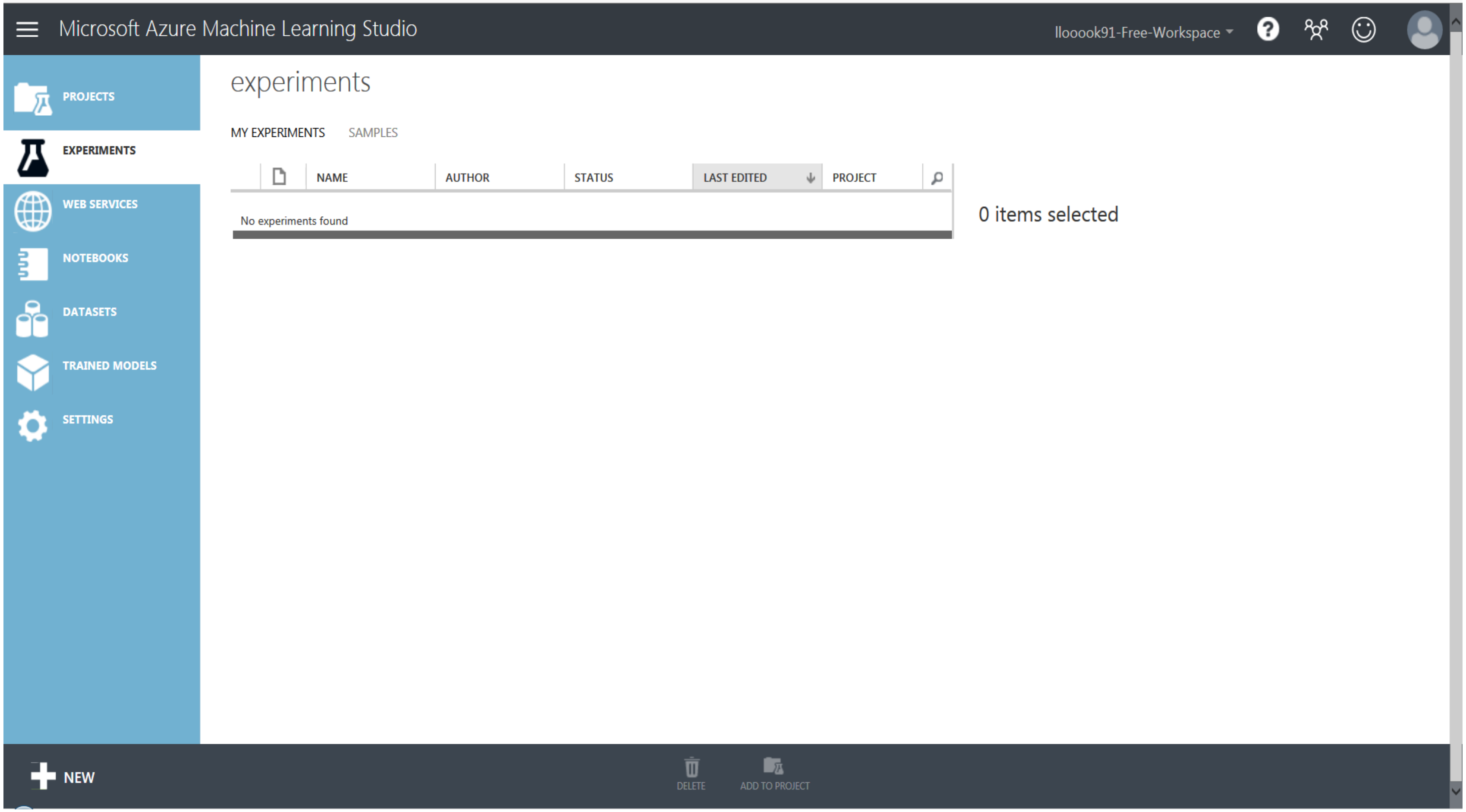
Azure ML 시작

6. Azure ML Studio 접속



Azure ML 시작

7. Azure ML Studio 화면



	NAME	AUTHOR	STATUS	LAST EDITED ↓	PROJECT
No experiments found					

0 items selected

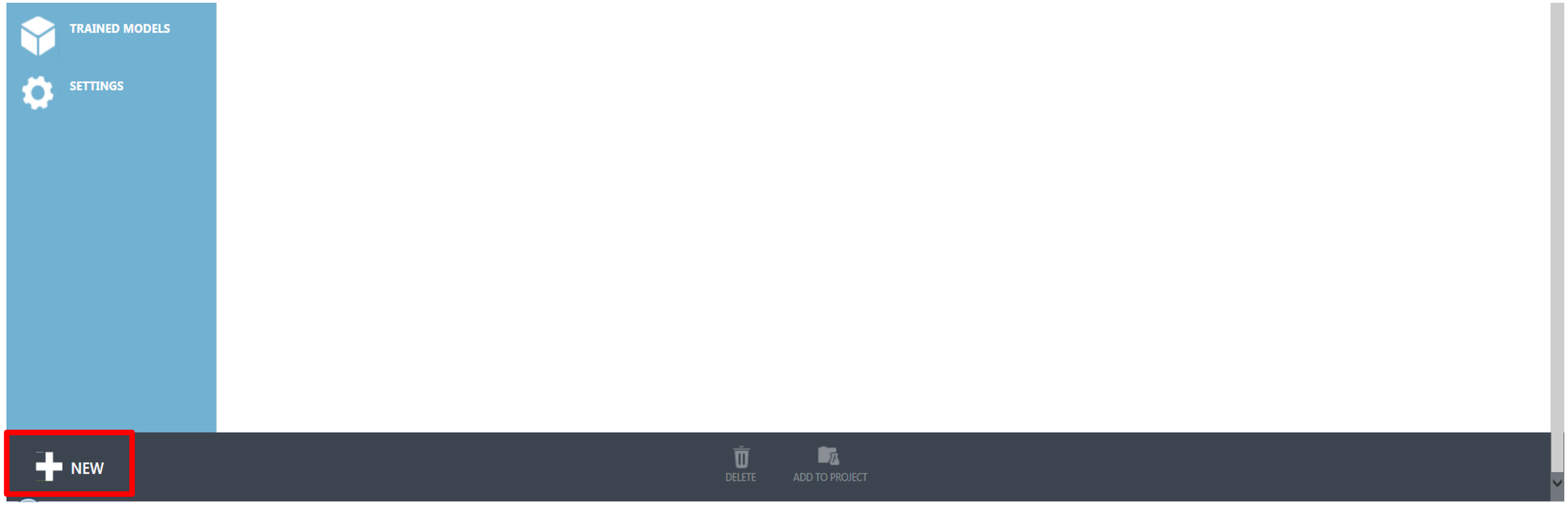
Azure ML 시작

8. Azure ML에서 실험 만들기

   : Azure ML 작업공간에서는 모델을 만들고 평가하고 분석하는 모든 작업이 실험(Experiment)이라는 단위로 이루어 짐

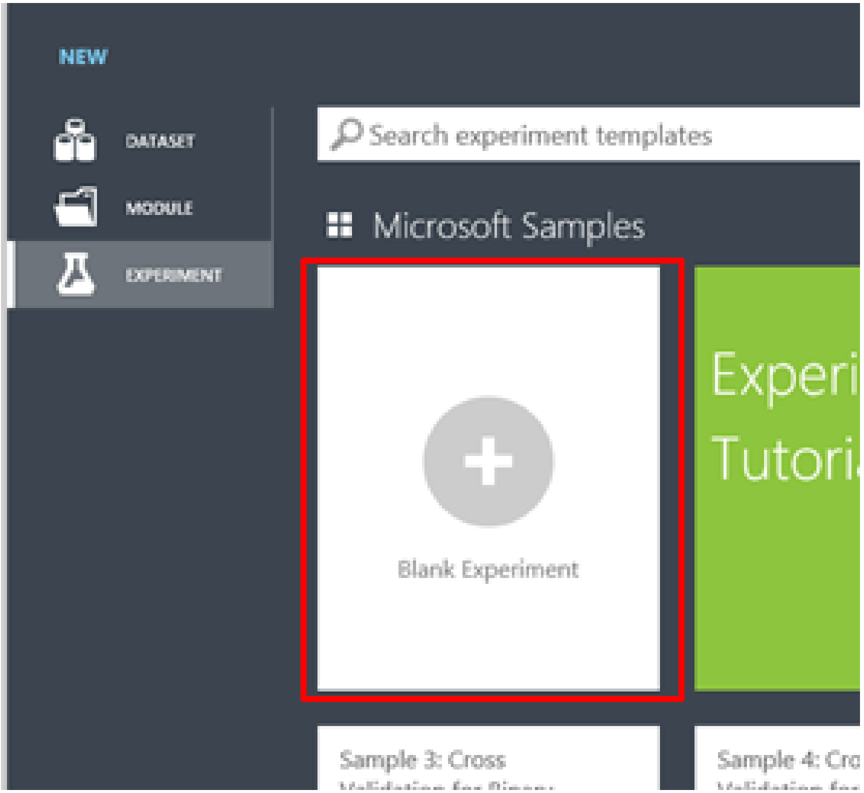
   실험은 모델과 관련된 데이터, 알고리즘 등을 포함

8.1 페이지 아래 왼쪽의 NEW 버튼 클릭



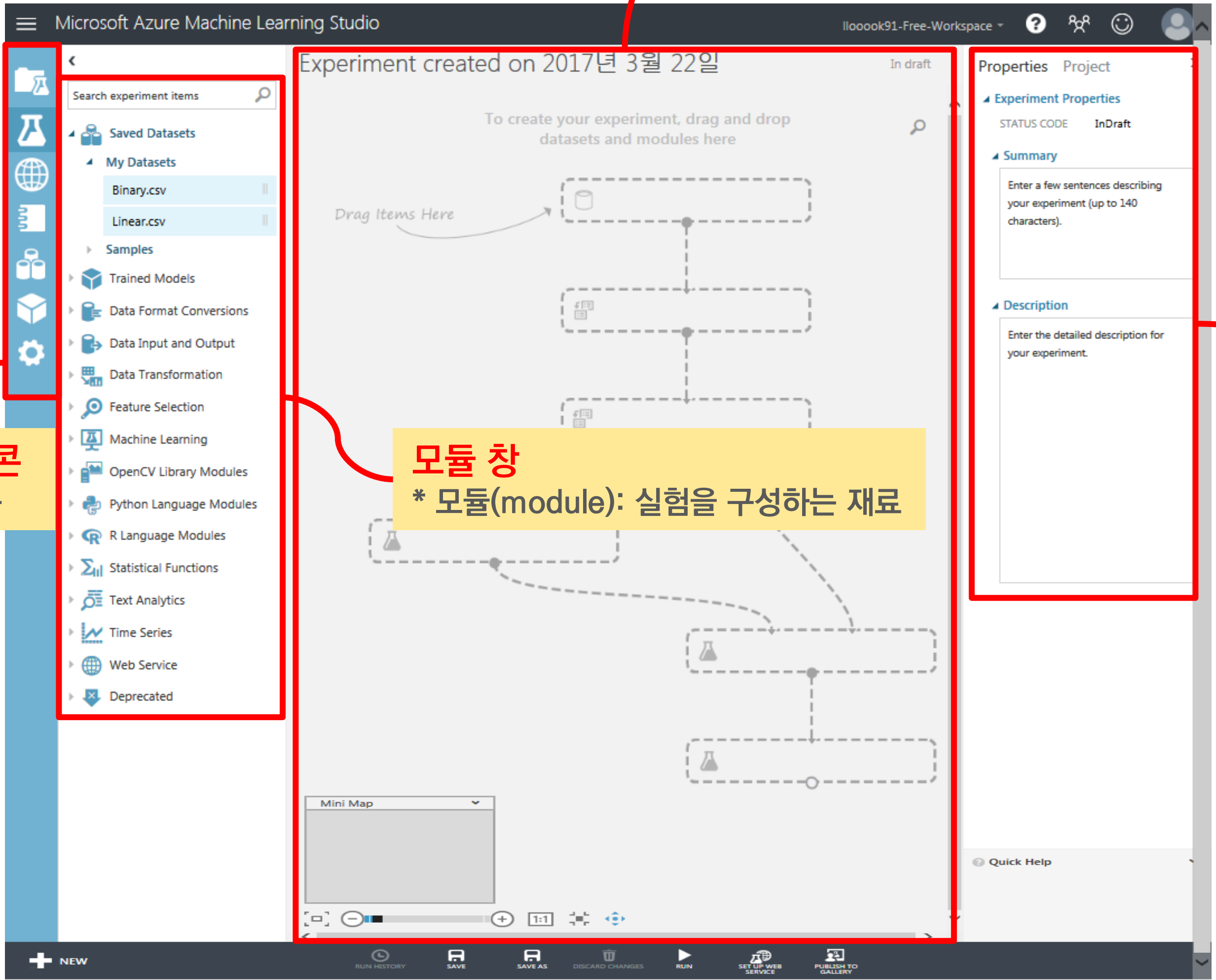
Azure ML 시작

9. EXPERIMENT가 선택된 상태에서 Blank Experiment 클릭



Azure ML 시작

10. 새로운 실험이 생성



Experiment created on 2017년 3월 22일

To create your experiment, drag and drop datasets and modules here

STATUS CODE InDraft

Enter a few sentences describing your experiment (up to 140 characters).

Enter the detailed description for your experiment.

캔버스(Canvas) : 실험을 구성하는 공간
* 실험의 구성: 모듈을 끌어다 놓고(drag and drop) 데이터의 흐름에 따라 서로 연결하는 과정

네비게이션 아이콘
: 작업공간으로 이동

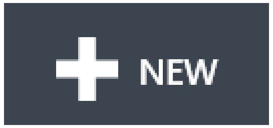
모듈 창
* 모듈(module): 실험을 구성하는 재료

속성(Properties)창
: 각 모듈의 속성을 지정

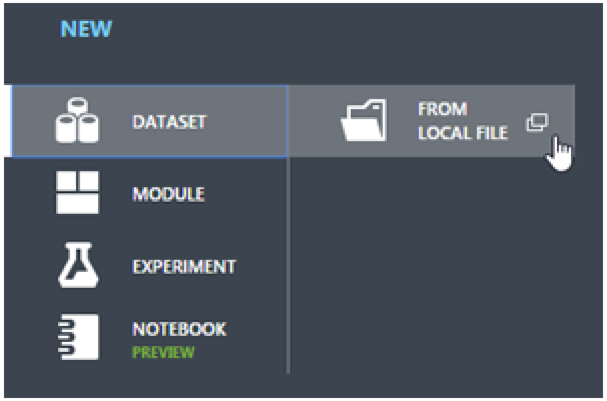
Dataset 준비

- **생성된 실험에 Dataset과 Module을 추가하여 분석 수행**

1. 앞서 실험을 만들 때와 마찬가지로 NEW 버튼 클릭



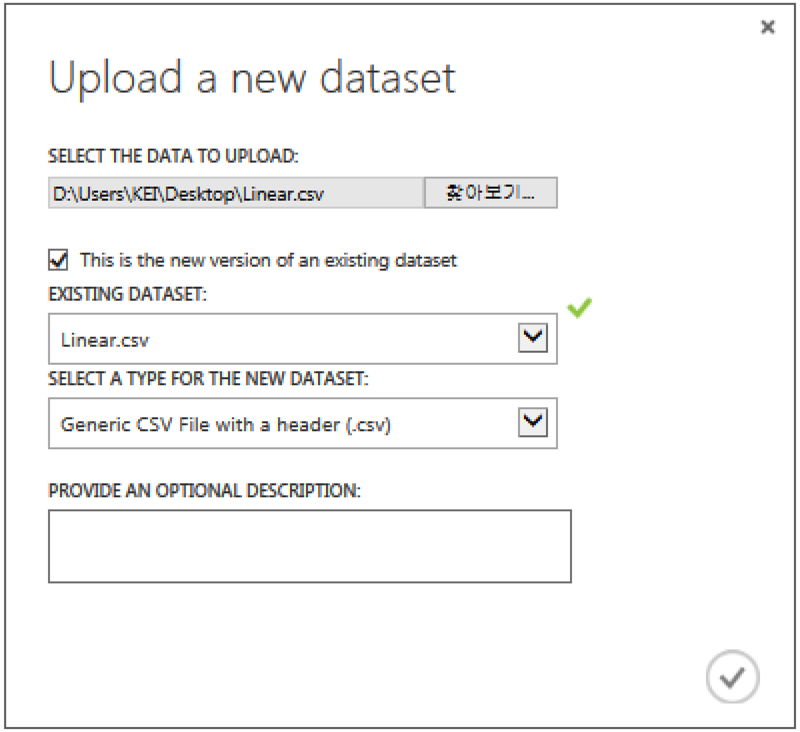
2. NEW 대화 창에 DATASET이 선택된 상태에서 FROM LOCAL FILE 버튼 클릭



Dataset 준비

- **생성된 실험에 Dataset과 Module을 추가하여 분석 수행**

3. Choose File 클릭
4. Dataset(Linear.csv) 불러옴



Upload a new dataset

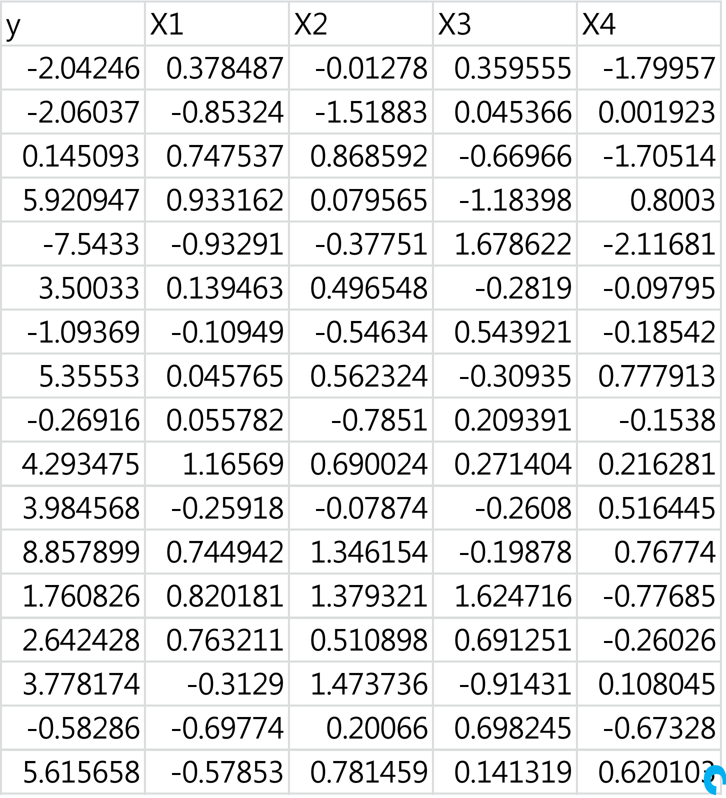
SELECT THE DATA TO UPLOAD: D:\Users\KEI\Desktop\Linear.csv

☑ This is the new version of an existing dataset

EXISTING DATASET: Linear.csv

SELECT A TYPE FOR THE NEW DATASET: Generic CSV File with a header (.csv)

PROVIDE AN OPTIONAL DESCRIPTION:

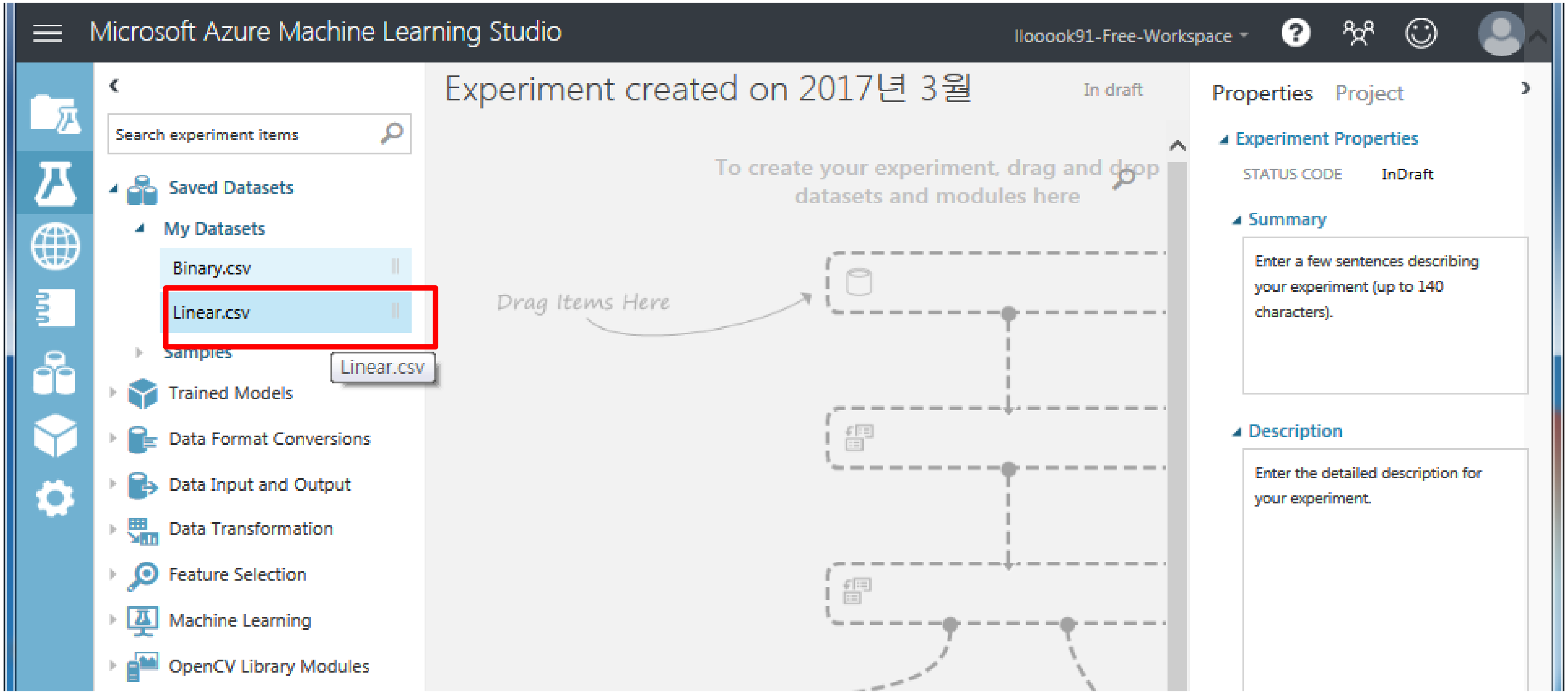


y	X1	X2	X3	X4
-2.04246	0.378487	-0.01278	0.359555	-1.79957
-2.06037	-0.85324	-1.51883	0.045366	0.001923
0.145093	0.747537	0.868592	-0.66966	-1.70514
5.920947	0.933162	0.079565	-1.18398	0.8003
-7.5433	-0.93291	-0.37751	1.678622	-2.11681
3.50033	0.139463	0.496548	-0.2819	-0.09795
-1.09369	-0.10949	-0.54634	0.543921	-0.18542
5.35553	0.045765	0.562324	-0.30935	0.777913
-0.26916	0.055782	-0.7851	0.209391	-0.1538
4.293475	1.16569	0.690024	0.271404	0.216281
3.984568	-0.25918	-0.07874	-0.2608	0.516445
8.857899	0.744942	1.346154	-0.19878	0.76774
1.760826	0.820181	1.379321	1.624716	-0.77685
2.642428	0.763211	0.510898	0.691251	-0.26026
3.778174	-0.3129	1.473736	-0.91431	0.108045
-0.58286	-0.69774	0.20066	0.698245	-0.67328
5.615658	-0.57853	0.781459	0.141319	0.620103

Dataset 준비

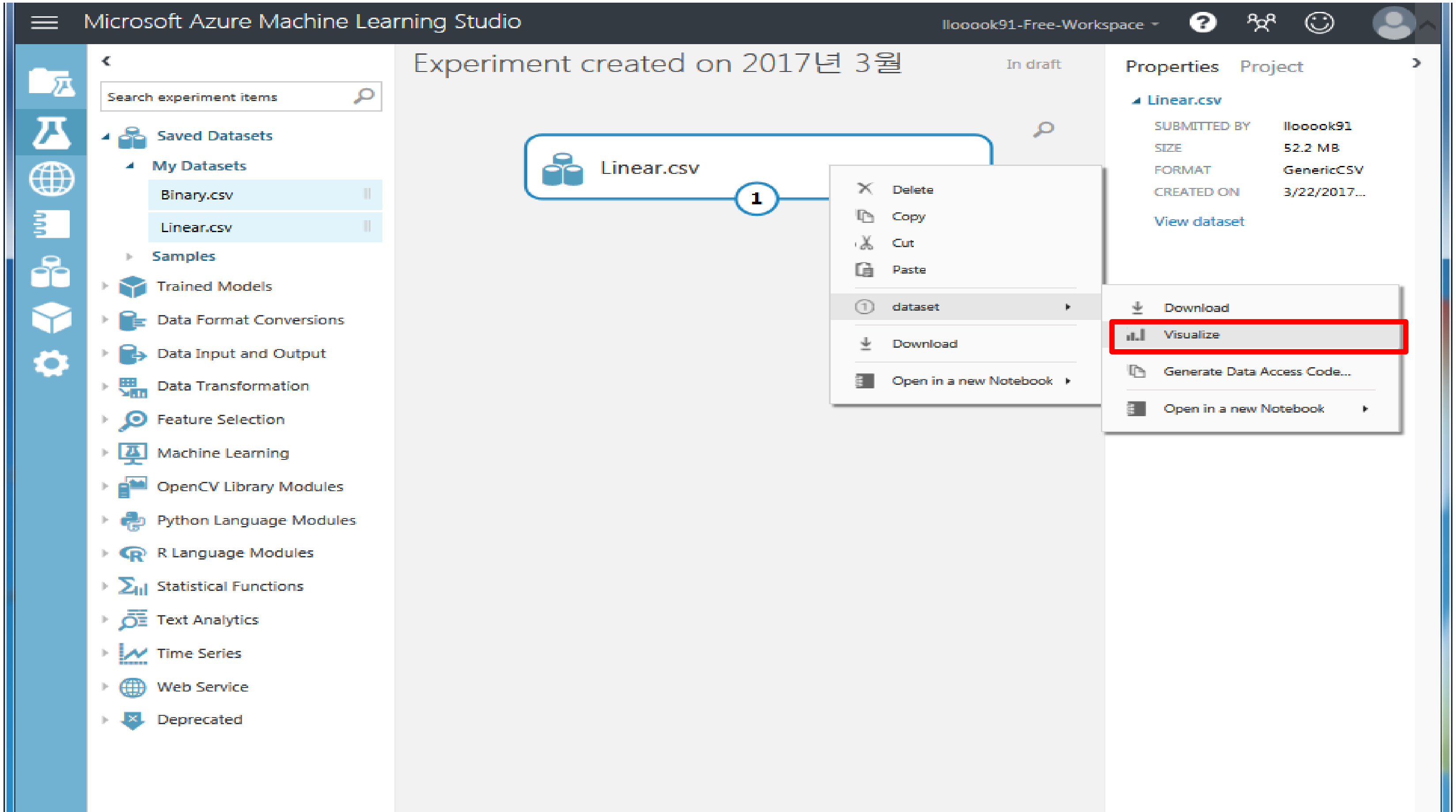
- **생성된 실험에 Dataset과 Module을 추가하여 분석 수행**

  5. Linear.csv라는 이름의 Dataset이 추가된 것을 확인함



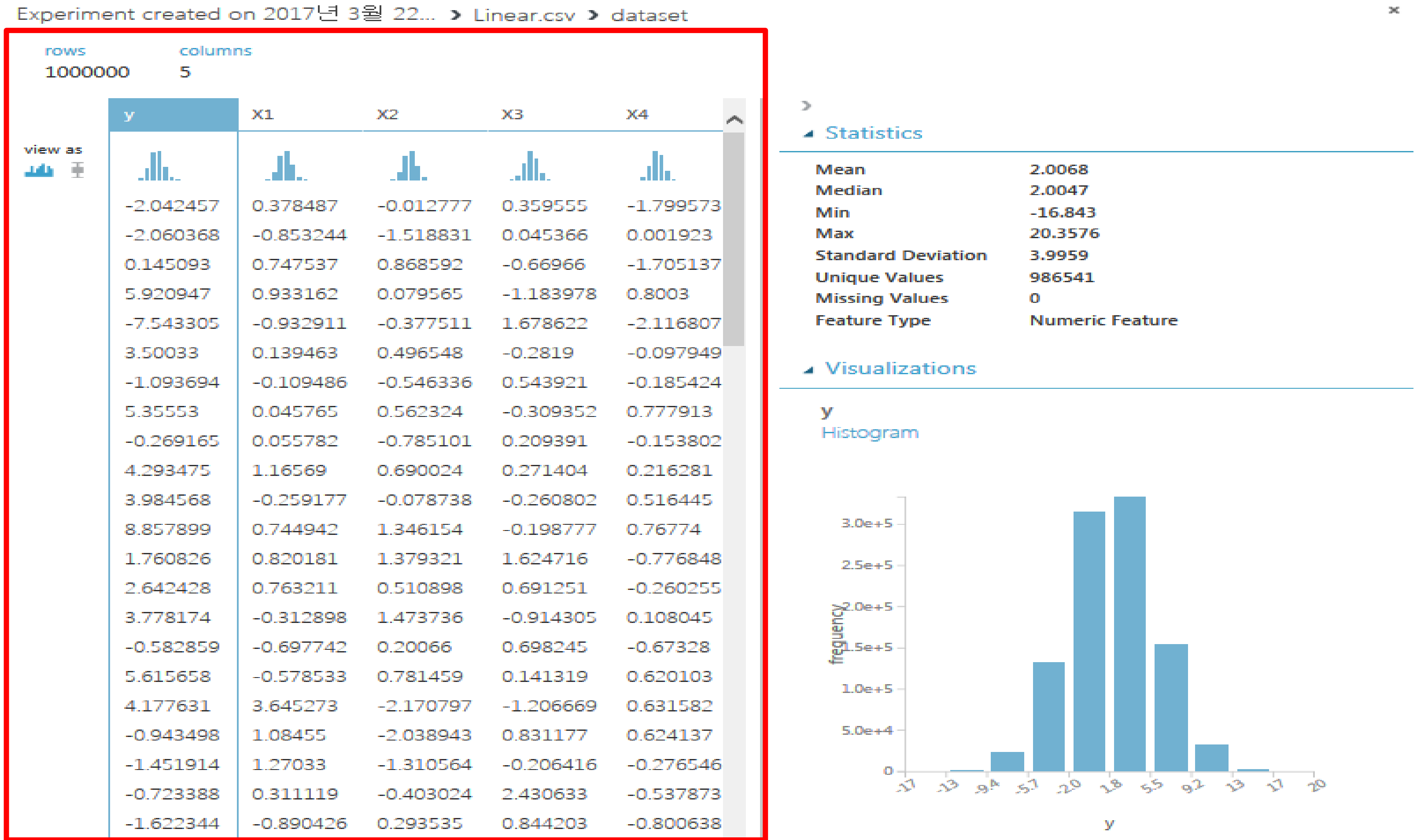
Dataset 준비

6. 추가한 Linear dataset을 캔버스로 끌어다 놓고 [dataset / Visualize] 선택



Dataset 준비

7. Linear dataset의 행과 열의 개수, 각 열에 대한 히스토그램 제공



Experiment created on 2017년 3월 22... › Linear.csv › dataset

rows	columns
1000000	5

y	X1	X2	X3	X4
-2.042457	0.378487	-0.012777	0.359555	-1.799573
-2.060368	-0.853244	-1.518831	0.045366	0.001923
0.145093	0.747537	0.868592	-0.66966	-1.705137
5.920947	0.933162	0.079565	-1.183978	0.8003
-7.543305	-0.932911	-0.377511	1.678622	-2.116807
3.50033	0.139463	0.496548	-0.2819	-0.097949
-1.093694	-0.109486	-0.546336	0.543921	-0.185424
5.35553	0.045765	0.562324	-0.309352	0.777913
-0.269165	0.055782	-0.785101	0.209391	-0.153802
4.293475	1.16569	0.690024	0.271404	0.216281
3.984568	-0.259177	-0.078738	-0.260802	0.516445
8.857899	0.744942	1.346154	-0.198777	0.76774
1.760826	0.820181	1.379321	1.624716	-0.776848
2.642428	0.763211	0.510898	0.691251	-0.260255
3.778174	-0.312898	1.473736	-0.914305	0.108045
-0.582859	-0.697742	0.20066	0.698245	-0.67328
5.615658	-0.578533	0.781459	0.141319	0.620103
4.177631	3.645273	-2.170797	-1.206669	0.631582
-0.943498	1.08455	-2.038943	0.831177	0.624137
-1.451914	1.27033	-1.310564	-0.206416	-0.276546
-0.723388	0.311119	-0.403024	2.430633	-0.537873
-1.622344	-0.890426	0.293535	0.844203	-0.800638

Statistics

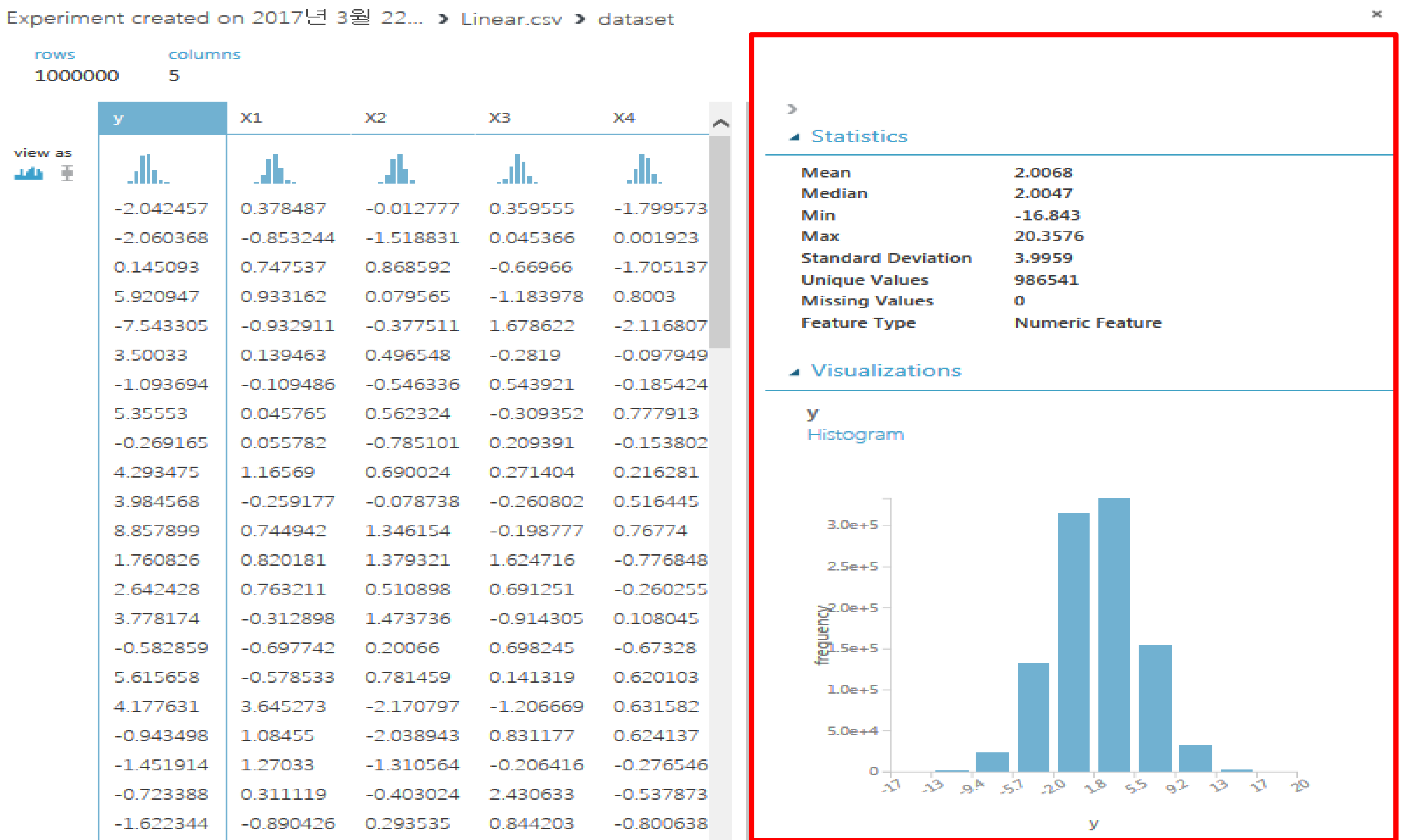
Mean	2.0068
Median	2.0047
Min	-16.843
Max	20.3576
Standard Deviation	3.9959
Unique Values	986541
Missing Values	0
Feature Type	Numeric Feature

Visualizations

y
Histogram

Dataset 준비

8. 해당 속성에 대한 다양한 통계값과 시각화 제공



Experiment created on 2017년 3월 22... > Linear.csv > dataset

rows	columns
1000000	5

y	X1	X2	X3	X4
-2.042457	0.378487	-0.012777	0.359555	-1.799573
-2.060368	-0.853244	-1.518831	0.045366	0.001923
0.145093	0.747537	0.868592	-0.66966	-1.705137
5.920947	0.933162	0.079565	-1.183978	0.8003
-7.543305	-0.932911	-0.377511	1.678622	-2.116807
3.50033	0.139463	0.496548	-0.2819	-0.097949
-1.093694	-0.109486	-0.546336	0.543921	-0.185424
5.35553	0.045765	0.562324	-0.309352	0.777913
-0.269165	0.055782	-0.785101	0.209391	-0.153802
4.293475	1.16569	0.690024	0.271404	0.216281
3.984568	-0.259177	-0.078738	-0.260802	0.516445
8.857899	0.744942	1.346154	-0.198777	0.76774
1.760826	0.820181	1.379321	1.624716	-0.776848
2.642428	0.763211	0.510898	0.691251	-0.260255
3.778174	-0.312898	1.473736	-0.914305	0.108045
-0.582859	-0.697742	0.20066	0.698245	-0.67328
5.615658	-0.578533	0.781459	0.141319	0.620103
4.177631	3.645273	-2.170797	-1.206669	0.631582
-0.943498	1.08455	-2.038943	0.831177	0.624137
-1.451914	1.27033	-1.310564	-0.206416	-0.276546
-0.723388	0.311119	-0.403024	2.430633	-0.537873
-1.622344	-0.890426	0.293535	0.844203	-0.800638

Statistics

Mean	2.0068
Median	2.0047
Min	-16.843
Max	20.3576
Standard Deviation	3.9959
Unique Values	986541
Missing Values	0
Feature Type	Numeric Feature

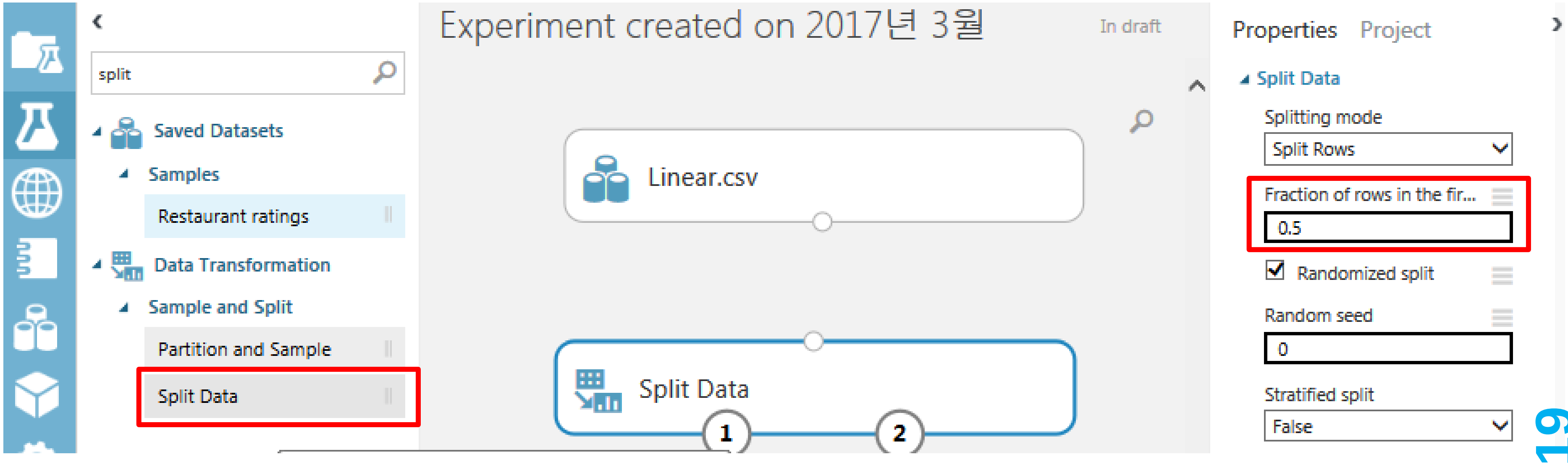
Visualizations

y
Histogram

Regression Model

Regression Analysis

- **모델을 만들기 위한 학습 데이터와 만들어진 모델을 평가하기 위한 평가 데이터로 분리**
  1. 작업공간 왼쪽 위의 검색 창에 데이터 분리를 위한 모듈 'split' 검색
  2. Split Data 모듈을 캔버스에 끌어다 놓음
  3. Properties 창에서 Fraction of rows in the first output dataset 에 0.75 입력 (왼쪽 출력포트: 75%, 오른쪽 출력포트: 25%)



split

Saved Datasets
- Samples
  - Restaurant ratings

Data Transformation
- Sample and Split
  - Partition and Sample
  - Split Data

Experiment created on 2017년 3월 — In draft

Properties | Project

Split Data

Splitting mode: Split Rows

Fraction of rows in the fir...: 0.5

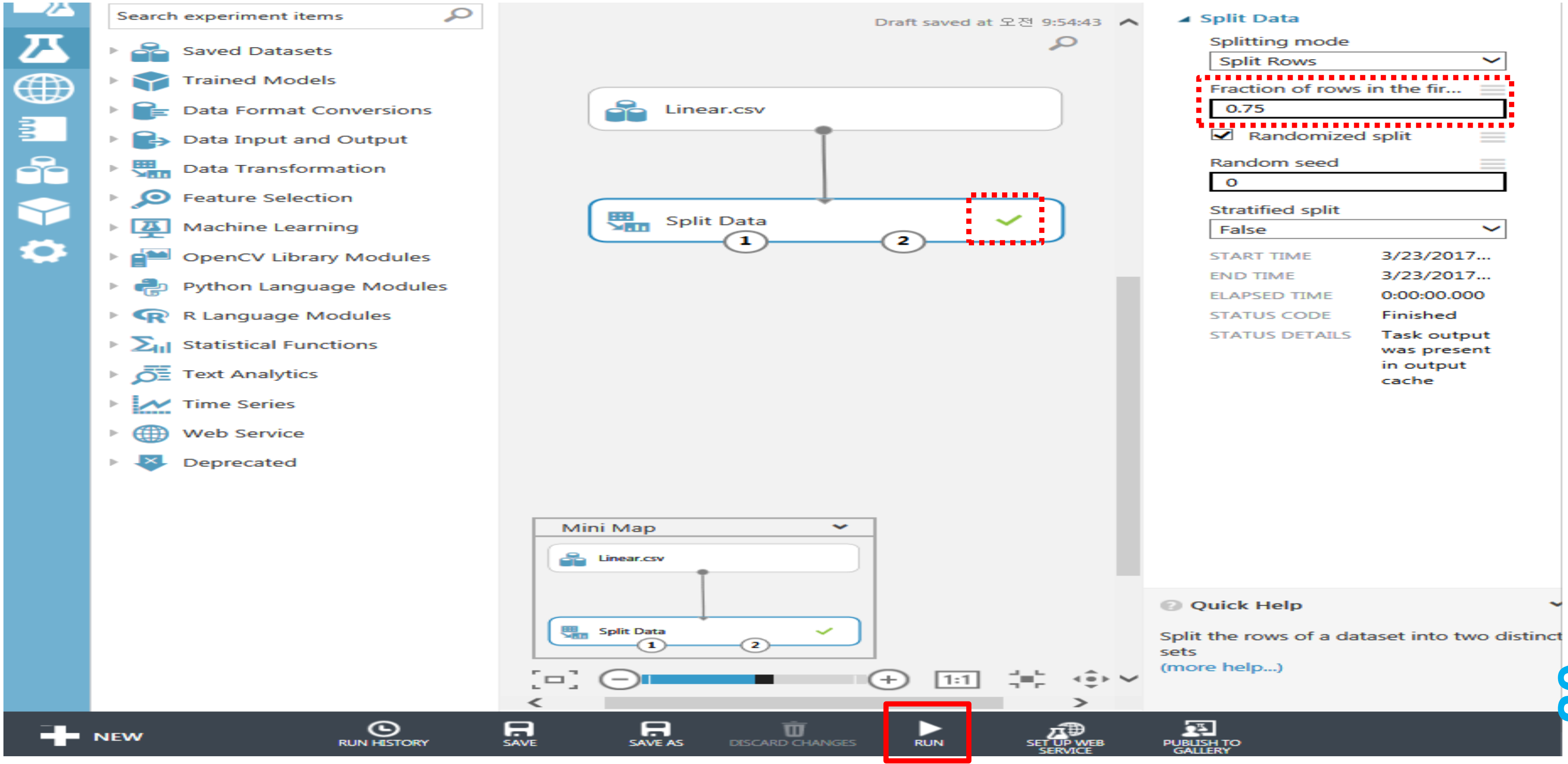
☑ Randomized split

Random seed: 0

Stratified split: False

Regression Analysis

4. 캔버스 아래쪽의 RUN을 클릭 실행이 완료되면 Split 모듈 오른쪽에 녹색 체크 표시



Split Data

Splitting mode: Split Rows

Fraction of rows in the fir...: 0.75

☑ Randomized split

Random seed: 0

Stratified split: False

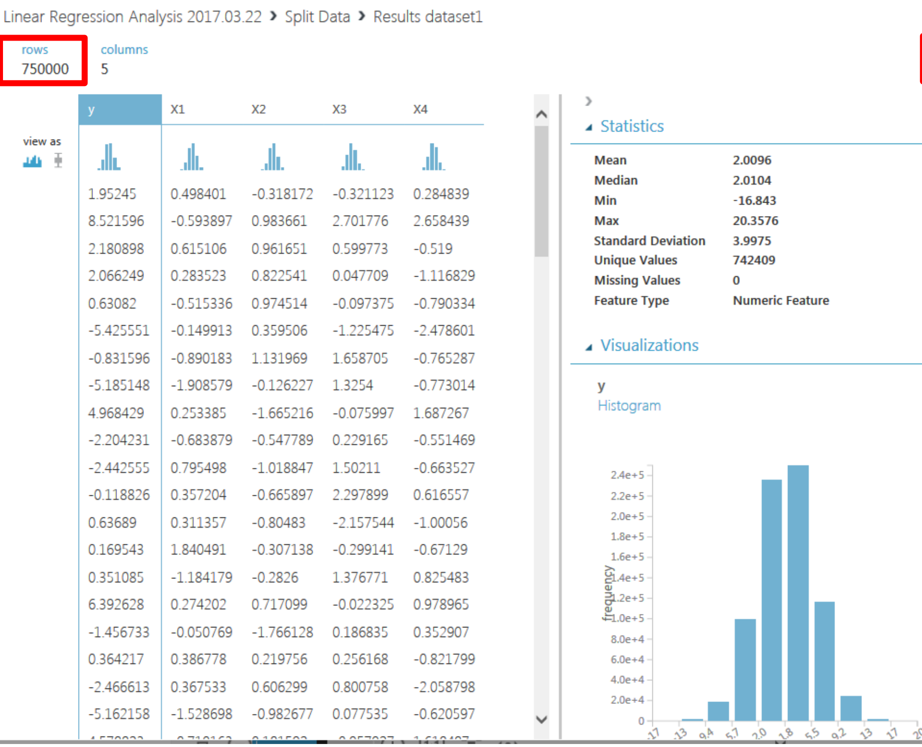
START TIME 3/23/2017...
END TIME 3/23/2017...
ELAPSED TIME 0:00:00.000
STATUS CODE Finished
STATUS DETAILS Task output was present in output cache

Quick Help

Split the rows of a dataset into two distinct sets
(more help...)

Regression Analysis

5. Split 모듈의 출력포트(왼쪽, 오른쪽) 클릭 -> Visualize 클릭
5.1 왼쪽 출력포트: 750,000개의 (75%)
오른쪽 출력포트: 250,000개의 (25%) 항목을 확인



Linear Regression Analysis 2017.03.22 › Split Data › Results dataset1

rows 750000 columns 5

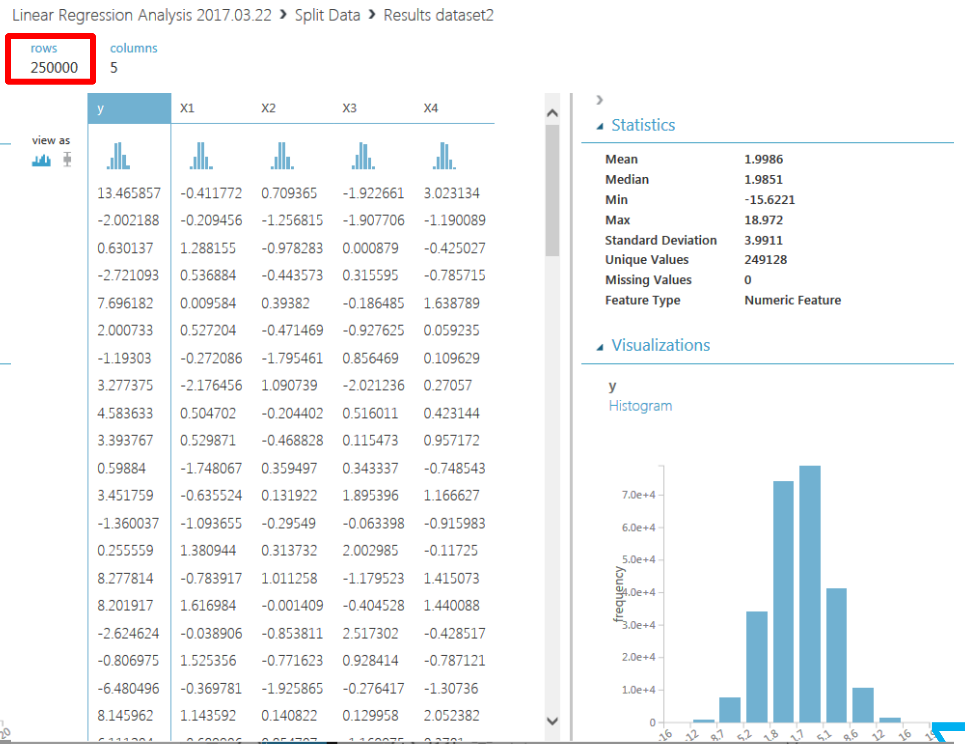
y	X1	X2	X3	X4
1.95245	0.498401	-0.318172	-0.321123	0.284839
8.521596	-0.593897	0.983661	2.701776	2.658439
2.180898	0.615106	0.961651	0.599773	-0.519
2.066249	0.283523	0.822541	0.047709	-1.116829
0.63082	-0.515336	0.974514	-0.097375	-0.790334
-5.425551	-0.149913	0.359506	-1.225475	-2.478601
-0.831596	-0.890183	1.131969	1.658705	-0.765287
-5.185148	-1.908579	-0.126227	1.3254	-0.773014
4.968429	0.253385	-1.665216	-0.075997	1.687267
-2.204231	-0.683879	-0.547789	0.229165	-0.551469
-2.442555	0.795498	-1.018847	1.50211	-0.663527
-0.118826	0.357204	-0.665897	2.297899	0.616557
0.63689	0.311357	-0.80483	-2.157544	-1.00056
0.169543	1.840491	-0.307138	-0.299141	-0.67129
0.351085	-1.184179	-0.2826	1.376771	0.825483
6.392628	0.274202	0.717099	-0.022325	0.978965
-1.456733	-0.050769	-1.766128	0.186835	0.352907
0.364217	0.386778	0.219756	0.256168	-0.821799
-2.466613	0.367533	0.606299	0.800758	-2.058798
-5.162158	-1.528698	-0.982677	0.077535	-0.620597

Statistics

Mean	2.0096
Median	2.0104
Min	-16.843
Max	20.3576
Standard Deviation	3.9975
Unique Values	742409
Missing Values	0
Feature Type	Numeric Feature

Visualizations

y
Histogram



Linear Regression Analysis 2017.03.22 › Split Data › Results dataset2

rows 250000 columns 5

y	X1	X2	X3	X4
13.465857	-0.411772	0.709365	-1.922661	3.023134
-2.002188	-0.209456	-1.256815	-1.907706	-1.190089
0.630137	1.288155	-0.978283	0.000879	-0.425027
-2.721093	0.536884	-0.443573	0.315595	-0.785715
7.696182	0.009584	0.39382	-0.186485	1.638789
2.000733	0.527204	-0.471469	-0.927625	0.059235
-1.19303	-0.272086	-1.795461	0.856469	0.109629
3.277375	-2.176456	1.090739	-2.021236	0.27057
4.583633	0.504702	-0.204402	0.516011	0.423144
3.393767	0.529871	-0.468828	0.115473	0.957172
0.59884	-1.748067	0.359497	0.343337	-0.748543
3.451759	-0.635524	0.131922	1.895396	1.166627
-1.360037	-1.093655	-0.29549	-0.063398	-0.915983
0.255559	1.380944	0.313732	2.002985	-0.11725
8.277814	-0.783917	1.011258	-1.179523	1.415073
8.201917	1.616984	-0.001409	-0.404528	1.440088
-2.624624	-0.038906	-0.853811	2.517302	-0.428517
-0.806975	1.525356	-0.771623	0.928414	-0.787121
-6.480496	-0.369781	-1.925865	-0.276417	-1.30736
8.145962	1.143592	0.140822	0.129958	2.052382

Statistics

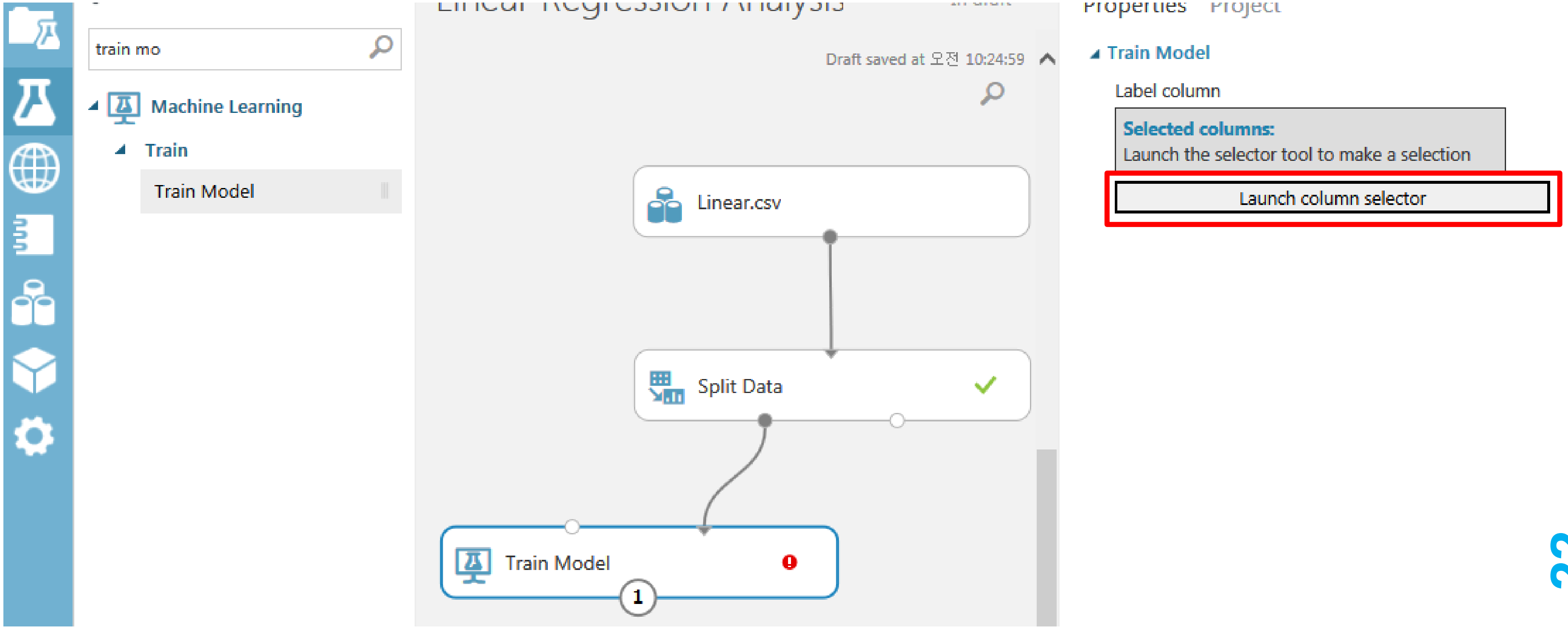
Mean	1.9986
Median	1.9851
Min	-15.6221
Max	18.972
Standard Deviation	3.9911
Unique Values	249128
Missing Values	0
Feature Type	Numeric Feature

Visualizations

y
Histogram

Regression Analysis

6. Train Model 모듈을 캔버스에 끌어다 놓고
   6.1 Split Data 모듈의 왼쪽 출력 포트와 Train Model 모듈의 오른쪽 입력 포트를 연결
7. Properties 창에서 Launch column selector 버튼 클릭

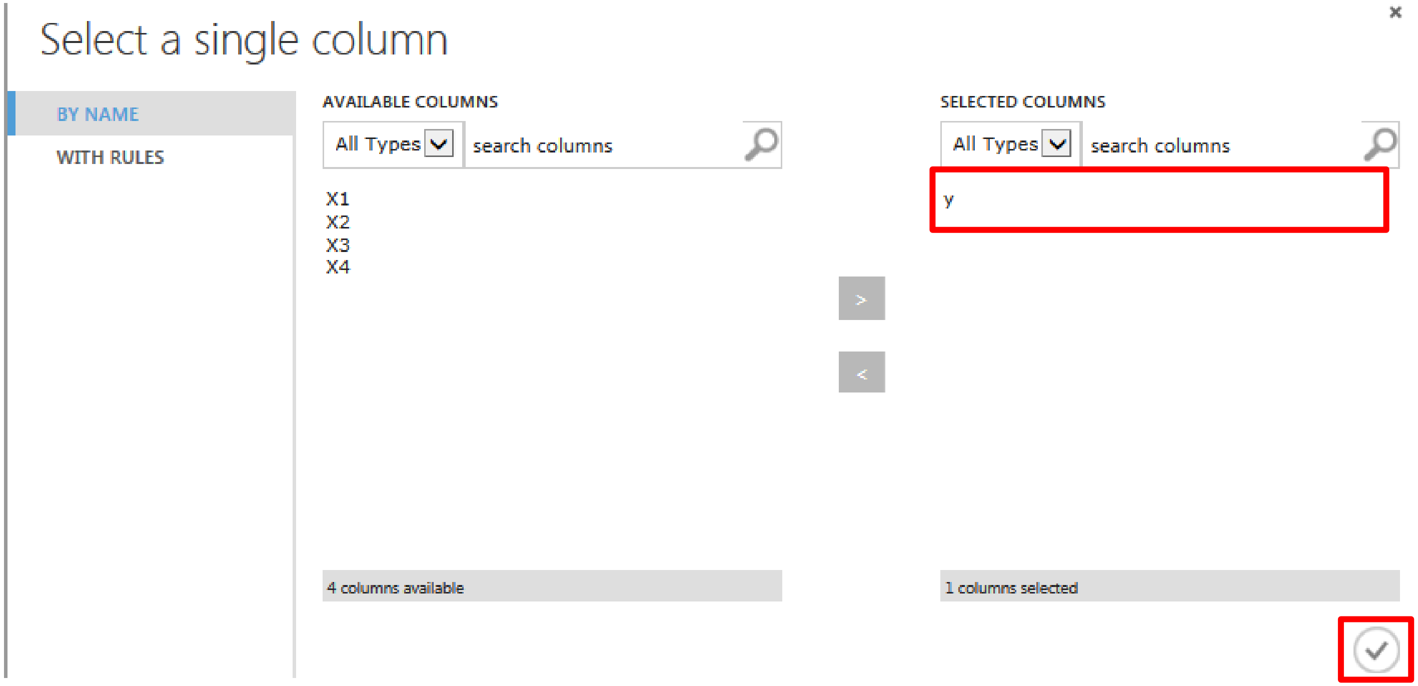


Label column

Selected columns:
Launch the selector tool to make a selection

Regression Analysis

8. 모델이 예측하고자 하는 속성 선택 (y 선택)



Select a single column

X1
X2
X3
X4

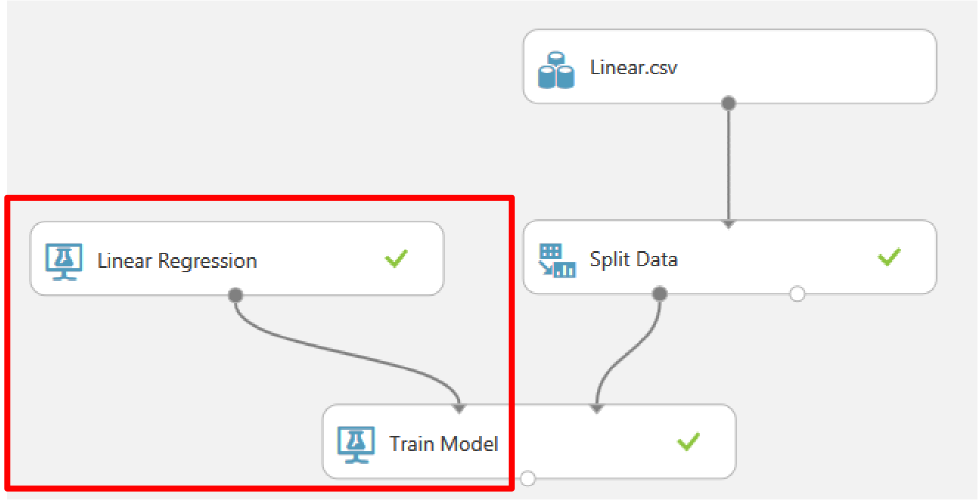
y

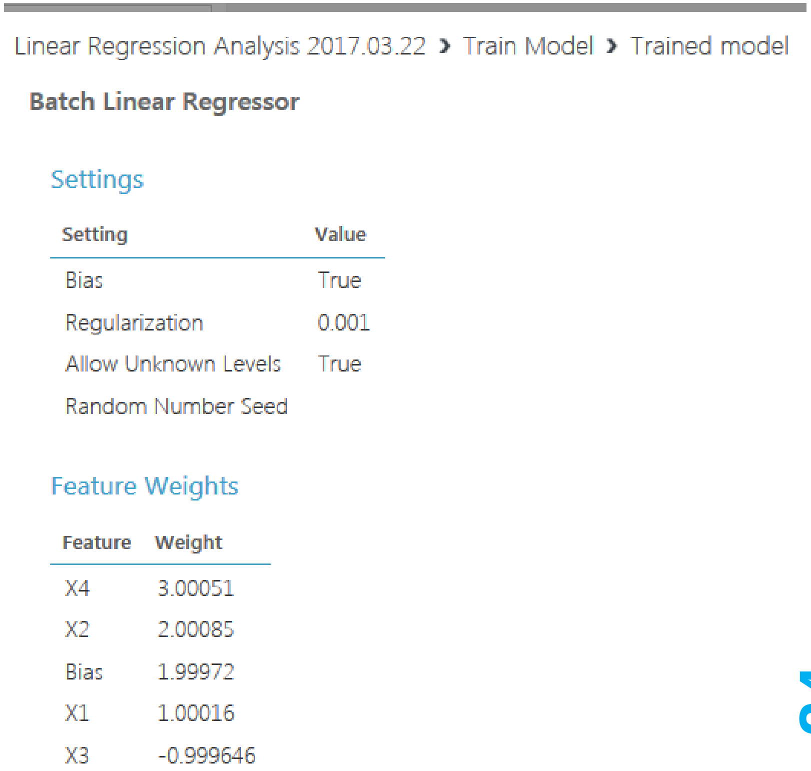
4 columns available

1 columns selected

Regression Analysis

9. 선형 회귀 알고리즘(linear regression)을 찾은 후, 이를 캔버스에 끌어 놓고 Train Model 모듈의 첫번째 입력 포트와 연결
9.1 Train Model 모듈 출력포트 Visualize해보면, 학습된 모델의 설정값과 속성의 가중치를 확인할 수 있음




Linear Regression Analysis 2017.03.22 › Train Model › Trained model

Batch Linear Regressor

Settings

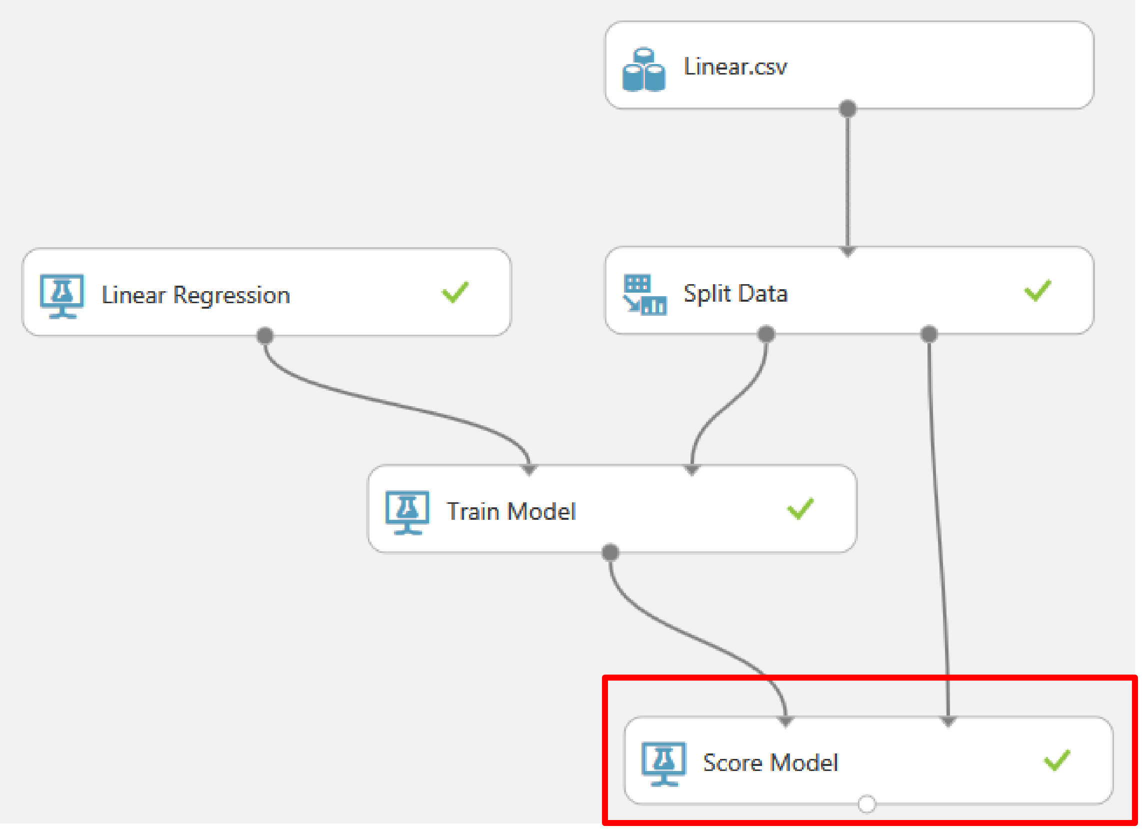
Setting	Value
Bias	True
Regularization	0.001
Allow Unknown Levels	True
Random Number Seed	

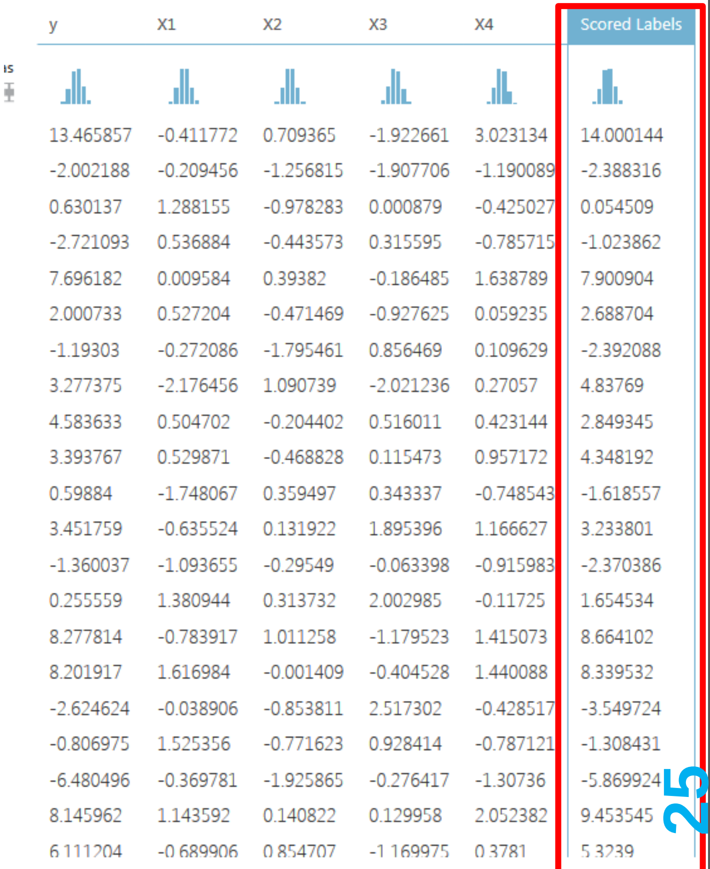
Feature Weights

Feature	Weight
X4	3.00051
X2	2.00085
Bias	1.99972
X1	1.00016
X3	-0.999646

Regression Analysis

10. Score Model을 캔버스에 끌어 놓고 앞서 만든 모델 및 평가 데이터와 연결
    10.1 Score Labels: y 속성을 예측한 결과

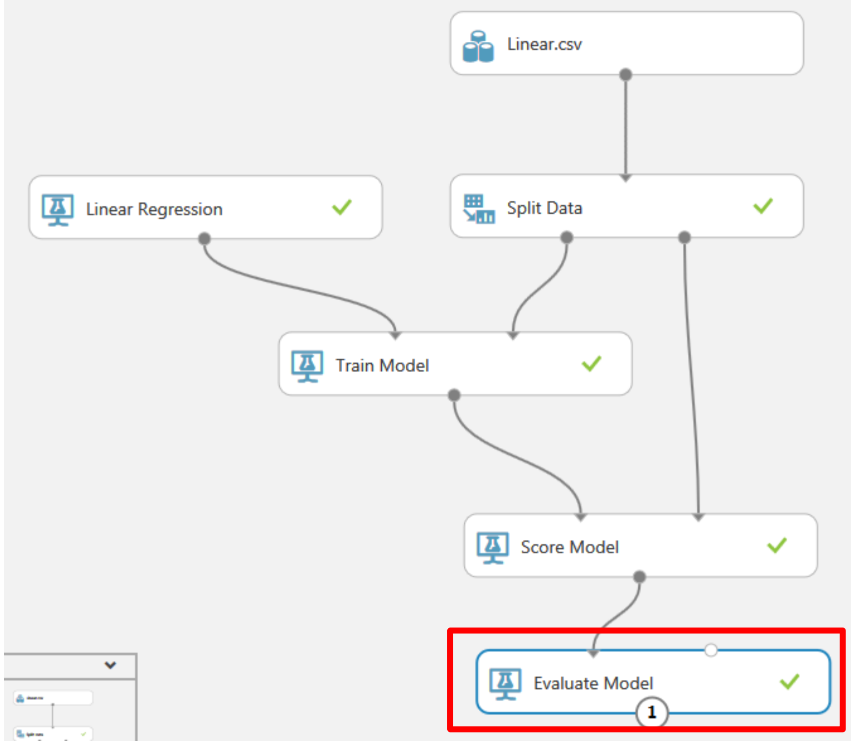


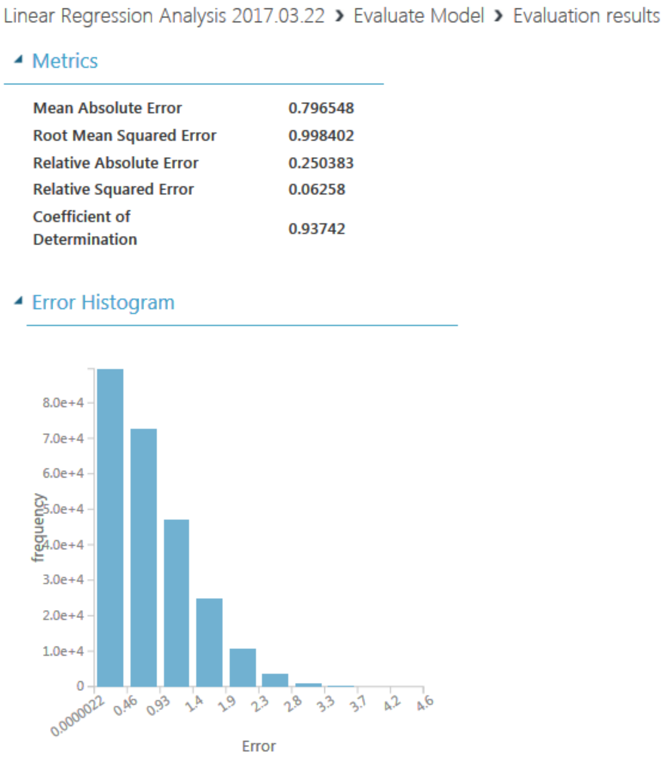


y	X1	X2	X3	X4	Scored Labels
13.465857	-0.411772	0.709365	-1.922661	3.023134	14.000144
-2.002188	-0.209456	-1.256815	-1.907706	-1.190089	-2.388316
0.630137	1.288155	-0.978283	0.000879	-0.425027	0.054509
-2.721093	0.536884	-0.443573	0.315595	-0.785715	-1.023862
7.696182	0.009584	0.39382	-0.186485	1.638789	7.900904
2.000733	0.527204	-0.471469	-0.927625	0.059235	2.688704
-1.19303	-0.272086	-1.795461	0.856469	0.109629	-2.392088
3.277375	-2.176456	1.090739	-2.021236	0.27057	4.83769
4.583633	0.504702	-0.204402	0.516011	0.423144	2.849345
3.393767	0.529871	-0.468828	0.115473	0.957172	4.348192
0.59884	-1.748067	0.359497	0.343337	-0.748543	-1.618557
3.451759	-0.635524	0.131922	1.895396	1.166627	3.233801
-1.360037	-1.093655	-0.29549	-0.063398	-0.915983	-2.370386
0.255559	1.380944	0.313732	2.002985	-0.11725	1.654534
8.277814	-0.783917	1.011258	-1.179523	1.415073	8.664102
8.201917	1.616984	-0.001409	-0.404528	1.440088	8.339532
-2.624624	-0.038906	-0.853811	2.517302	-0.428517	-3.549724
-0.806975	1.525356	-0.771623	0.928414	-0.787121	-1.308431
-6.480496	-0.369781	-1.925865	-0.276417	-1.30736	-5.869924
8.145962	1.143592	0.140822	0.129958	2.052382	9.453545
6.111204	-0.689906	0.854707	-1.169975	0.3781	5.3239

Regression Model 성능 비교

1. Evaluate Model을 캔버스에 끌어 놓고 Score Model과 연결
   1.1 Evaluate Model: 예측 값과 실제 값을 바탕으로 다양한 평가 지표 제공
2. 결과포트를 Visualize한 결과: RMSE를 비롯한 다양한 평가지표와 예측된 값과 실제 값 간의 차이를 시각화한 결과 확인




Linear Regression Analysis 2017.03.22 › Evaluate Model › Evaluation results

Metrics

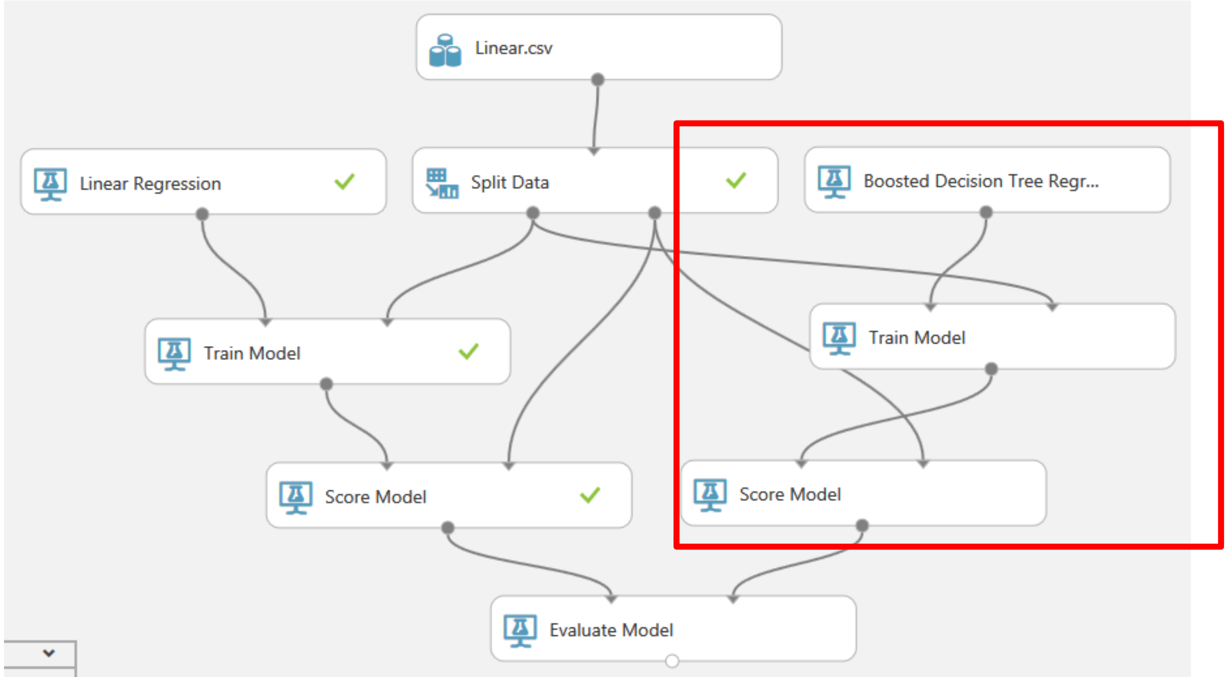
Mean Absolute Error	0.796548
Root Mean Squared Error	0.998402
Relative Absolute Error	0.250383
Relative Squared Error	0.06258
Coefficient of Determination	0.93742

Error Histogram

Regression Model 성능 비교

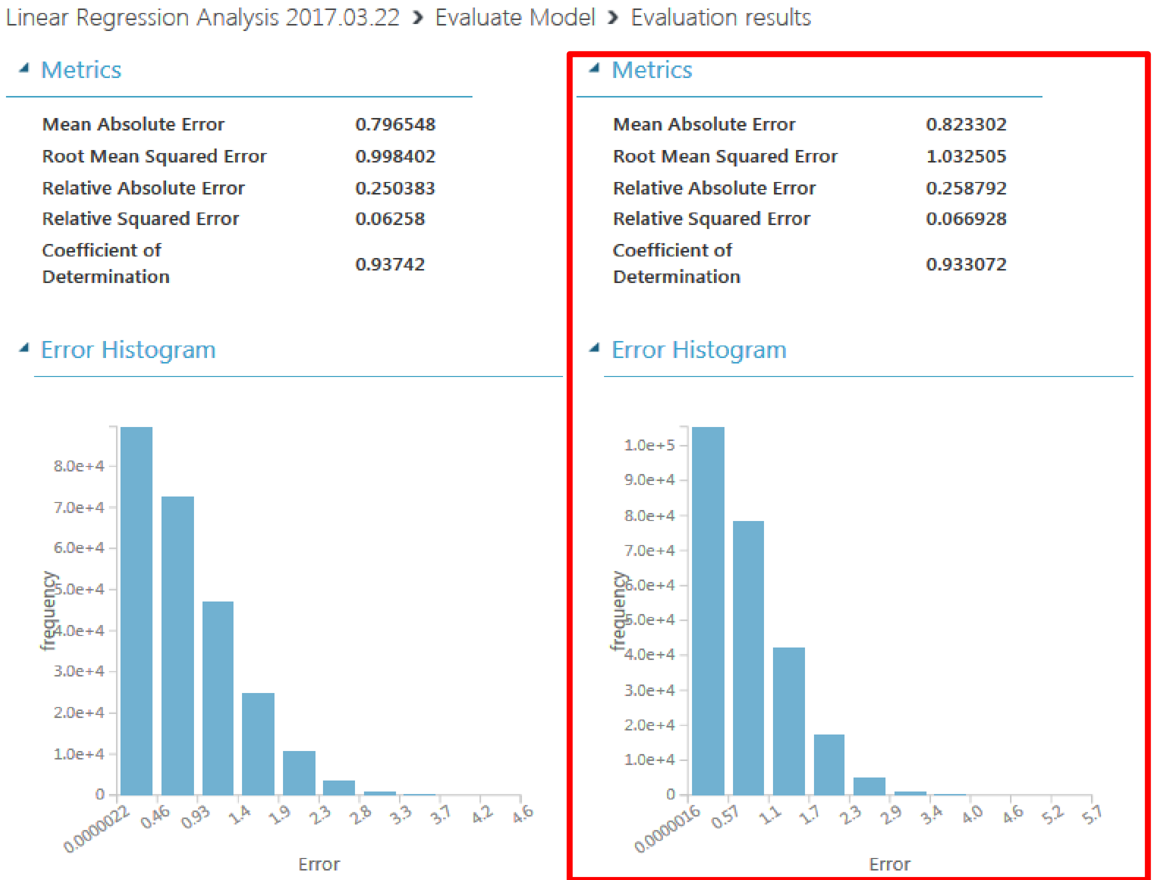
- **앙상블에 기반한 의사결정트리 모델을 추가해서 성능 비교**

3. Boosted Decision Tree Regression 모듈 및 해당하는 Train/Score Model 모듈을 추가
4. Score Model 모듈의 결과 포트를 Evaluate Model 모듈의 다른 쪽 입력 포트에 연결



Regression Model 성능 비교

5. 앙상블 트리 모델에 비해 선형회귀 모델의 RMSE가 낮으며, 에러의 분포 역시 전반적으로 훨씬 낮은 값을 보임



Linear Regression Analysis 2017.03.22 › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	0.796548
Root Mean Squared Error	0.998402
Relative Absolute Error	0.250383
Relative Squared Error	0.06258
Coefficient of Determination	0.93742

Error Histogram

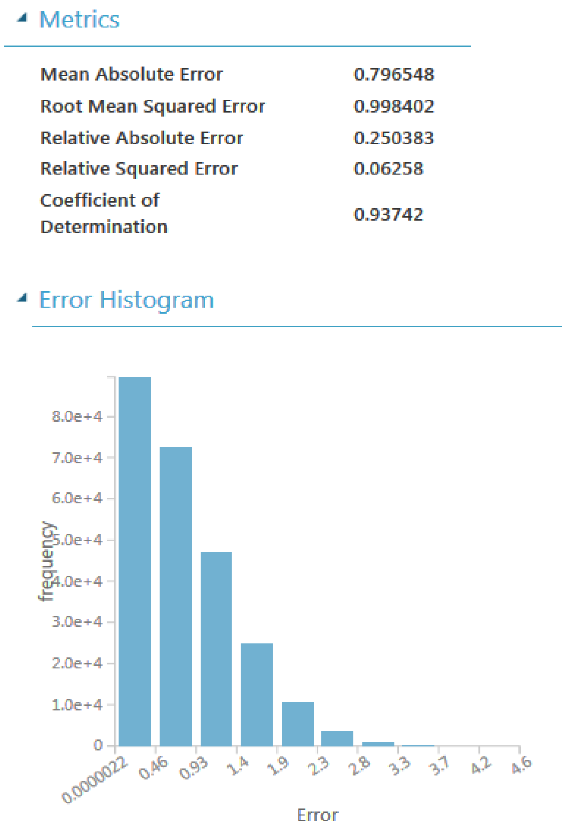
Metrics

Mean Absolute Error	0.823302
Root Mean Squared Error	1.032505
Relative Absolute Error	0.258792
Relative Squared Error	0.066928
Coefficient of Determination	0.933072

Error Histogram

Regression Model 성능 비교

Regression Analysis

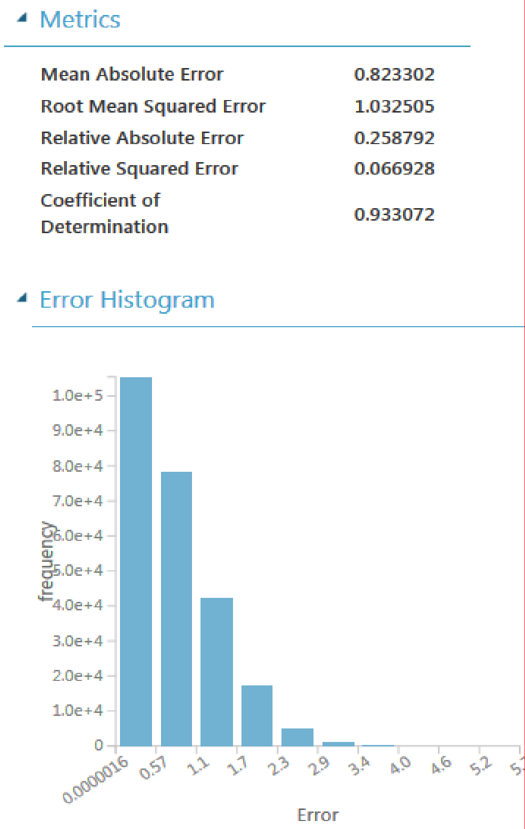


Metrics

Mean Absolute Error	0.796548
Root Mean Squared Error	0.998402
Relative Absolute Error	0.250383
Relative Squared Error	0.06258
Coefficient of Determination	0.93742

Error Histogram

Boosted Decision Tree

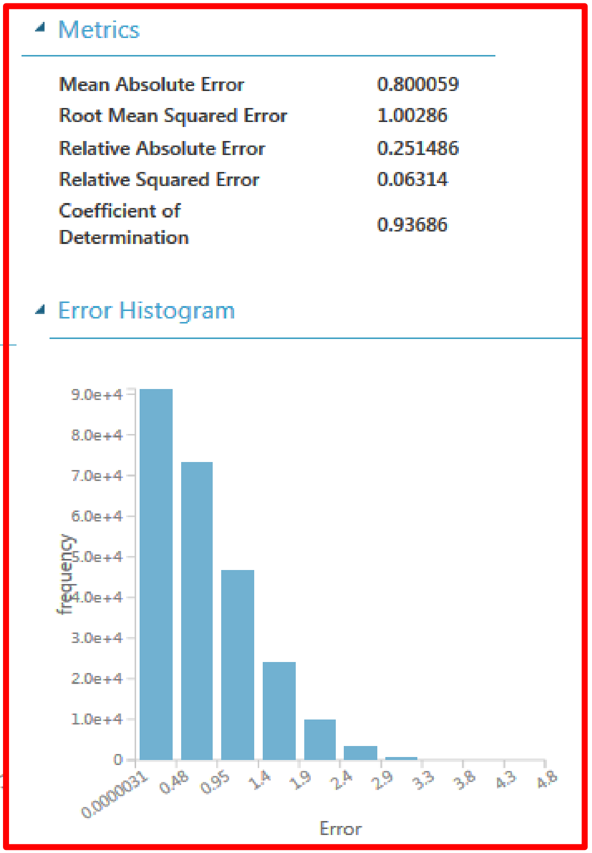


Metrics

Mean Absolute Error	0.823302
Root Mean Squared Error	1.032505
Relative Absolute Error	0.258792
Relative Squared Error	0.066928
Coefficient of Determination	0.933072

Error Histogram

Neural Network Regression



Metrics

Mean Absolute Error	0.800059
Root Mean Squared Error	1.00286
Relative Absolute Error	0.251486
Relative Squared Error	0.06314
Coefficient of Determination	0.93686

Error Histogram

6. 예측 정확도 : Regression 〉 Neural Network 〉 Boosted Decision Tree

Classification Model

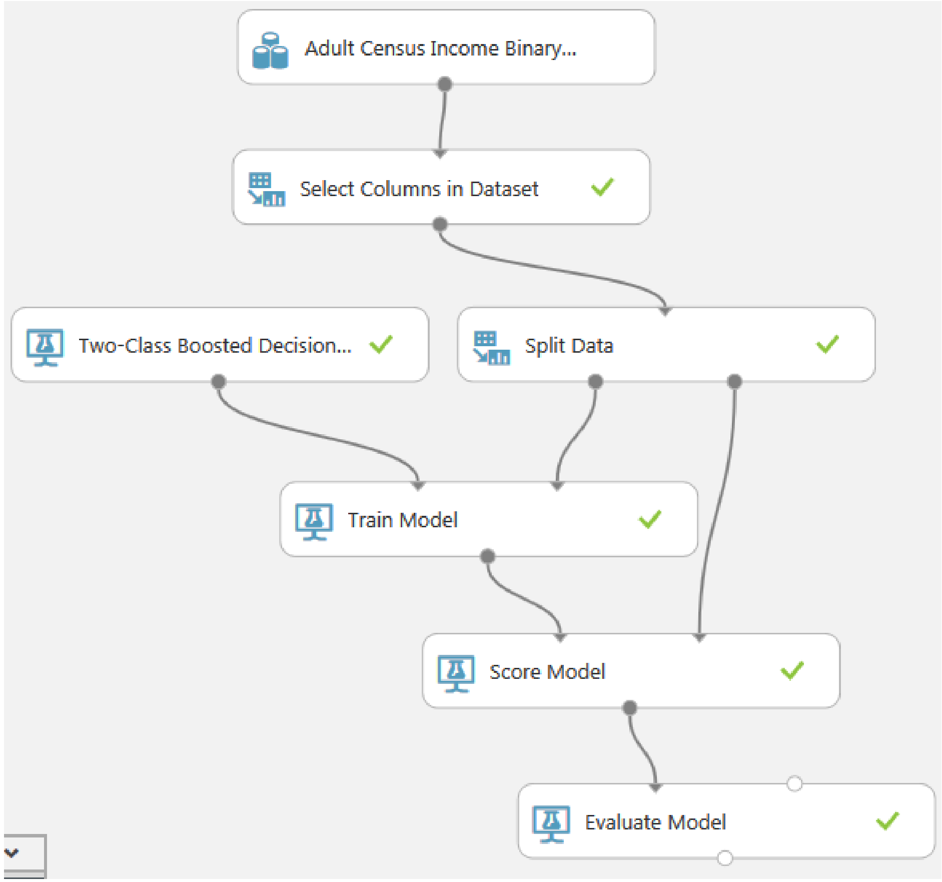
Classification Analysis

1. "Adult Census Income Binary Classification dataset" 불러오기

2. Select columns in Dataset
   : age, education, marital-status, relationship, race, sex, income(기준: 50K)

3. Split Data:
   Training set 80%, Test set 20%

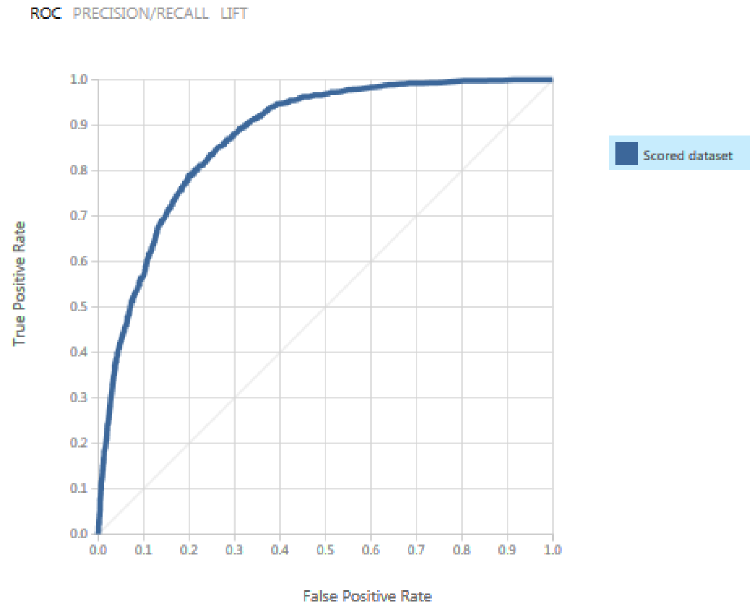
4. Classification method:
   Two-Class Boosted Decision Tree

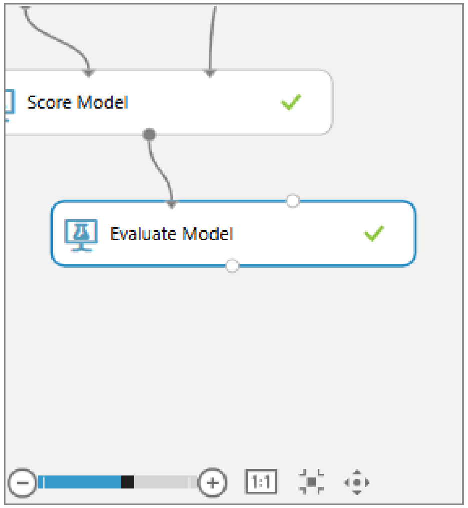

Adult Census Income Binary...
Select Columns in Dataset ✓
Two-Class Boosted Decision... ✓
Split Data ✓
Train Model ✓
Score Model ✓
Evaluate Model ✓

Classification Analysis

5. Evaluate Model
   : Visualize 클릭

6. ROC커브 를 통해
   모델 성능 평가
   : 수직축 (민감도)
   수평축 (특이도 – 1)



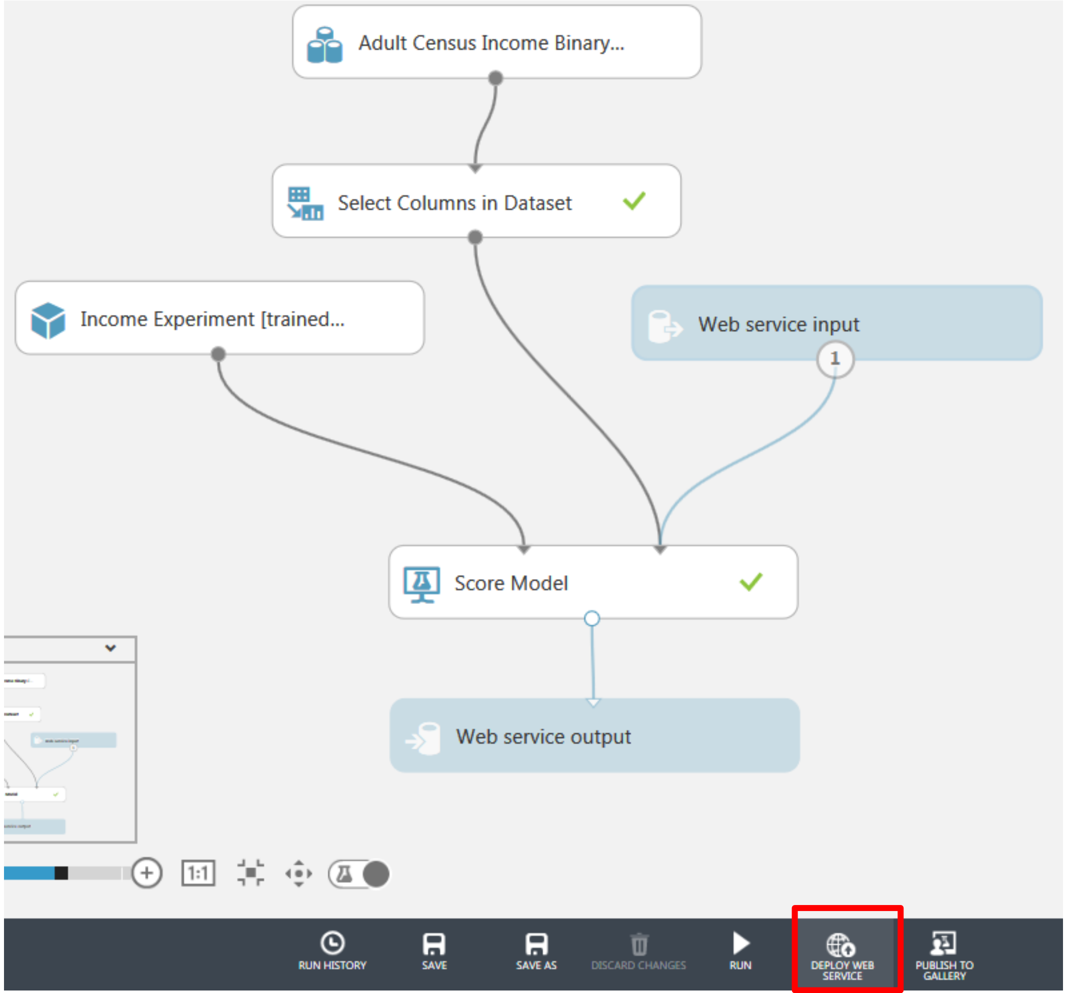


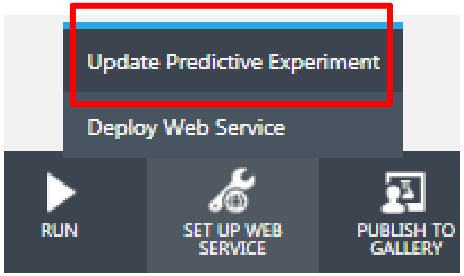
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
887	691	0.823	0.658	0.5	0.875
False Positive	**True Negative**	**Recall**	**F1 Score**		
462	4472	0.562	0.606		
Positive Label	**Negative Label**				
>50K	<=50K				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulativ
(0.900,1.000]	43	10	0.008	0.763	0.053	0.811	0.027	0.762	0.998	0.00
(0.800,0.900]	181	40	0.042	0.784	0.242	0.818	0.142	0.783	0.990	0.00
(0.700,0.800]	307	110	0.106	0.815	0.468	0.768	0.337	0.820	0.968	[illegible]
(0.600,0.700]	135	80	0.139	0.823	0.536	0.735	0.422	0.837	0.951	[illegible]
(0.500,0.600]	221	222	0.207	0.823	0.606	0.658	0.562	0.866	0.906	0.03

Classification Analysis

7. Update Predictive Experiment 클릭

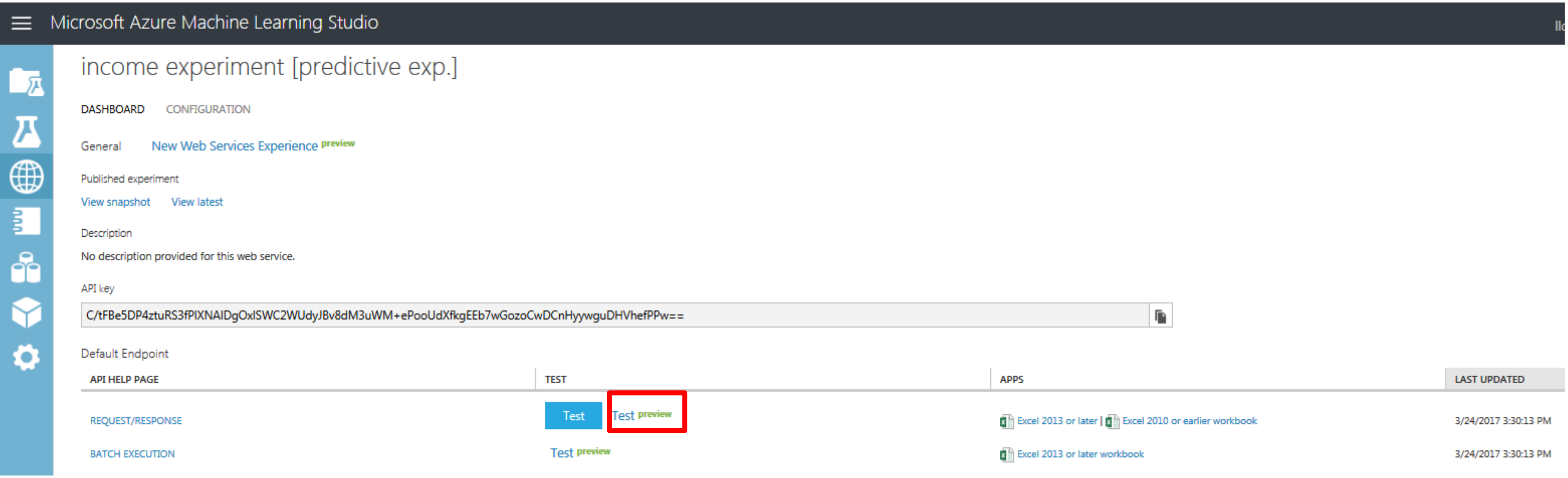




8. DEPLOY WEB SERVICE 클릭

Classification Analysis

9. Test preview 클릭



Microsoft Azure Machine Learning Studio

income experiment [predictive exp.]

DASHBOARD CONFIGURATION

General New Web Services Experience preview

Published experiment

View snapshot View latest

Description

No description provided for this web service.

API key

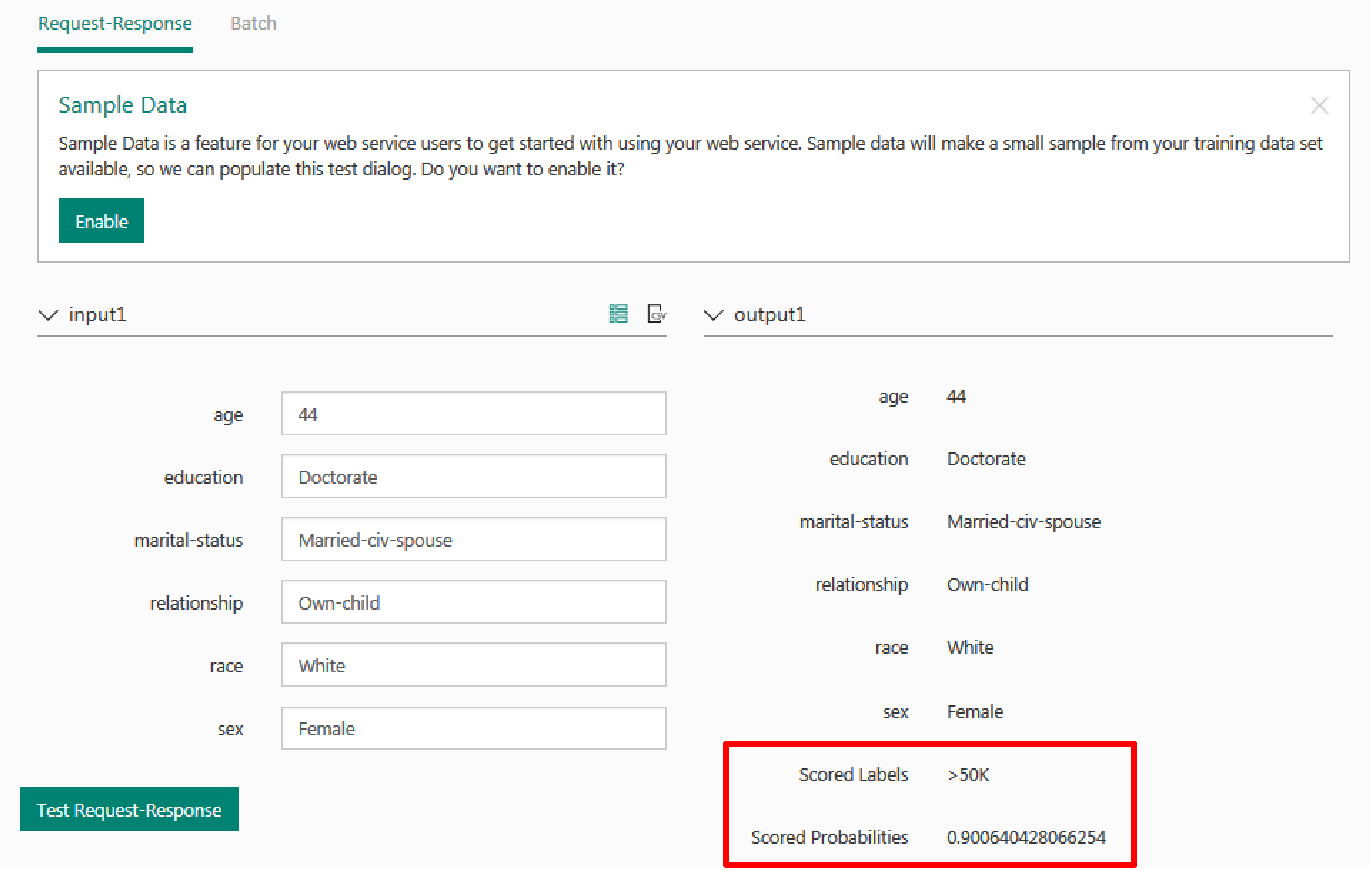
C/tFBe5DP4ztuRS3fPIXNAIDgOxlSWC2WUdyJBv8dM3uWM+ePooUdXfkgEEb7wGozoCwDCnHyywguDHVhefPPw==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test preview	Excel 2013 or later | Excel 2010 or earlier workbook	3/24/2017 3:30:13 PM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	3/24/2017 3:30:13 PM

Classification Analysis

10. Input 입력 후 output 확인
    Scored Labels : 〉50K
    Scored Probabilities: 90.06%



Sample Data

Sample Data is a feature for your web service users to get started with using your web service. Sample data will make a small sample from your training data set available, so we can populate this test dialog. Do you want to enable it?

input1

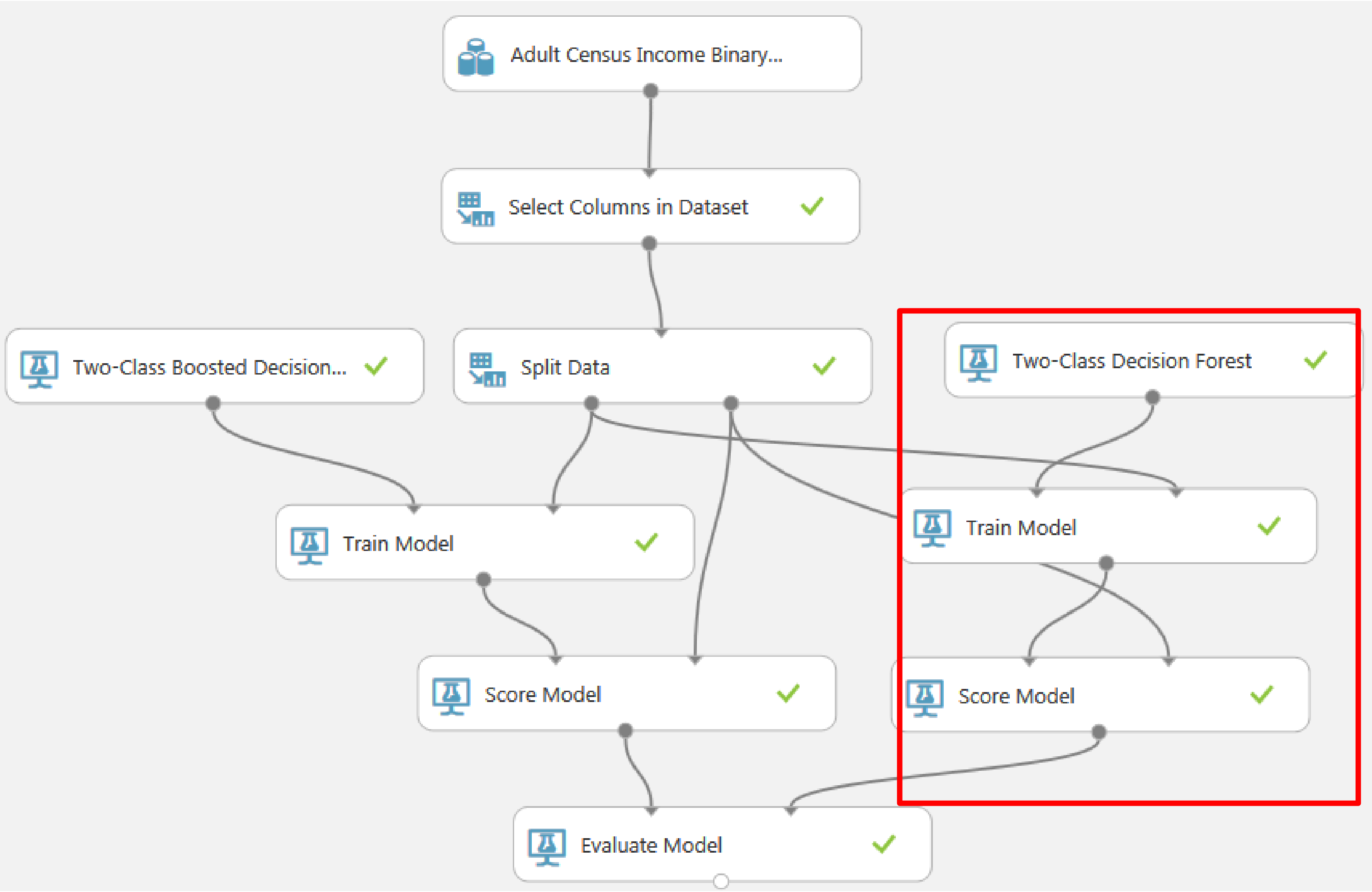
age	44
education	Doctorate
marital-status	Married-civ-spouse
relationship	Own-child
race	White
sex	Female

output1

age	44
education	Doctorate
marital-status	Married-civ-spouse
relationship	Own-child
race	White
sex	Female
Scored Labels	>50K
Scored Probabilities	0.900640428066254

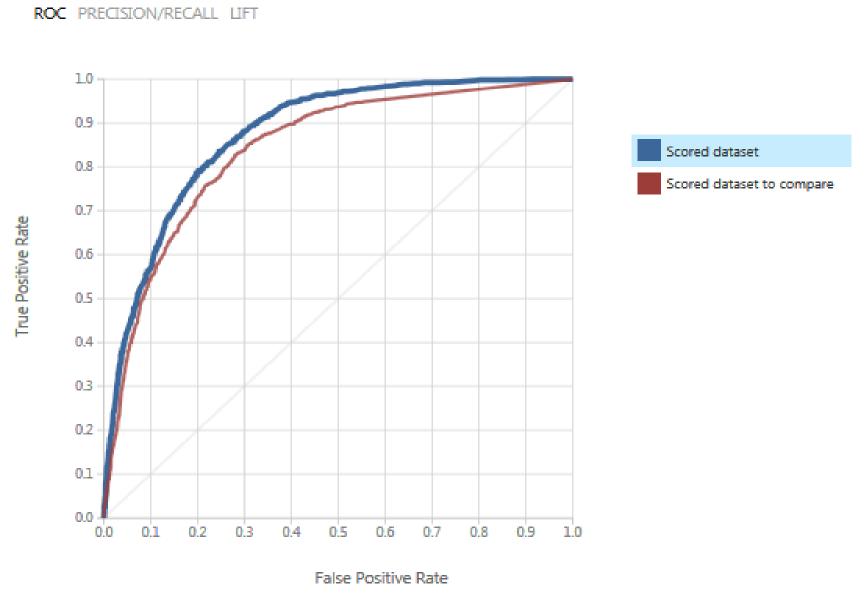
Classification Model 성능 비교

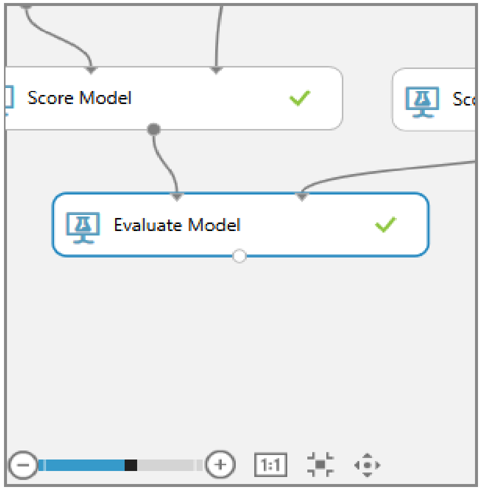
11. Two-Class Decision Forest Model 추가


Adult Census Income Binary...
Select Columns in Dataset ✓
Two-Class Boosted Decision... ✓
Split Data ✓
Two-Class Decision Forest ✓
Train Model ✓
Train Model ✓
Score Model ✓
Score Model ✓
Evaluate Model ✓

Classification Model 성능 비교

12. 예측 정확도 : Two-Class Boosted Decision Tree 〉 Two-Class Decision Forest





True Positive	False Negative	Accuracy	Precision	Threshold	AUC
887	691	0.823	0.658	0.5	0.875
False Positive	True Negative	Recall	F1 Score		
462	4472	0.562	0.606		
Positive Label	Negative Label				
>50K	<=50K				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	43	10	0.008	0.763	0.053	0.811	0.027	0.762	0.998	0.000
(0.800,0.900]	181	40	0.042	0.784	0.242	0.818	0.142	0.783	0.990	0.001
(0.700,0.800]	307	110	0.106	0.815	0.468	0.768	0.337	0.820	0.968	0.006

Azure ML 장.단점

장점:

1) H/W, S/W를 별도 구매할 필요 없음
2) drag, drop, connect를 통해서 모델을 훈련 시킴
3) 데이터 집합과 모듈을 시각적으로 연결하면 예측 분석 모델 구성
4) 35개의 샘플 데이터 셋과 68개의 샘플 실험 제공

편리성

5) 다양한 ML 라이브러리 제공 (36개)
6) R 및 Python 스크립트를 추가하여 확장 가능함
7) Predictive Experiment 기능
8) Open API 제공

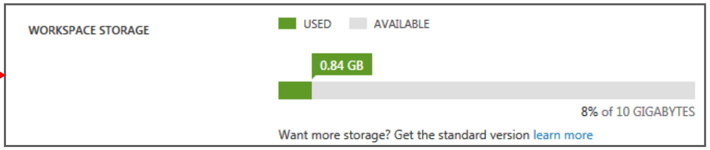
확장성 & 유연성

9) Running time 짧음 (ex. Linear.csv(100만건): 4분 소요)
Income Binary.csv(3만건): 15초 소요)
10) 다양한 data formats 지원 (csv, text, SQLtable, Rdata, zip 등)

기능성

단점:

1) 데이터 업로드 속도가 느림 (ex. Linear.csv(100만건): 10분 소요)
2) 동시에 3개 이상의 모델 성능 비교 어려움
3) 무료 사용기간이 짧고 사용 용량이 적음 (30일, 10GB)



WORKSPACE STORAGE

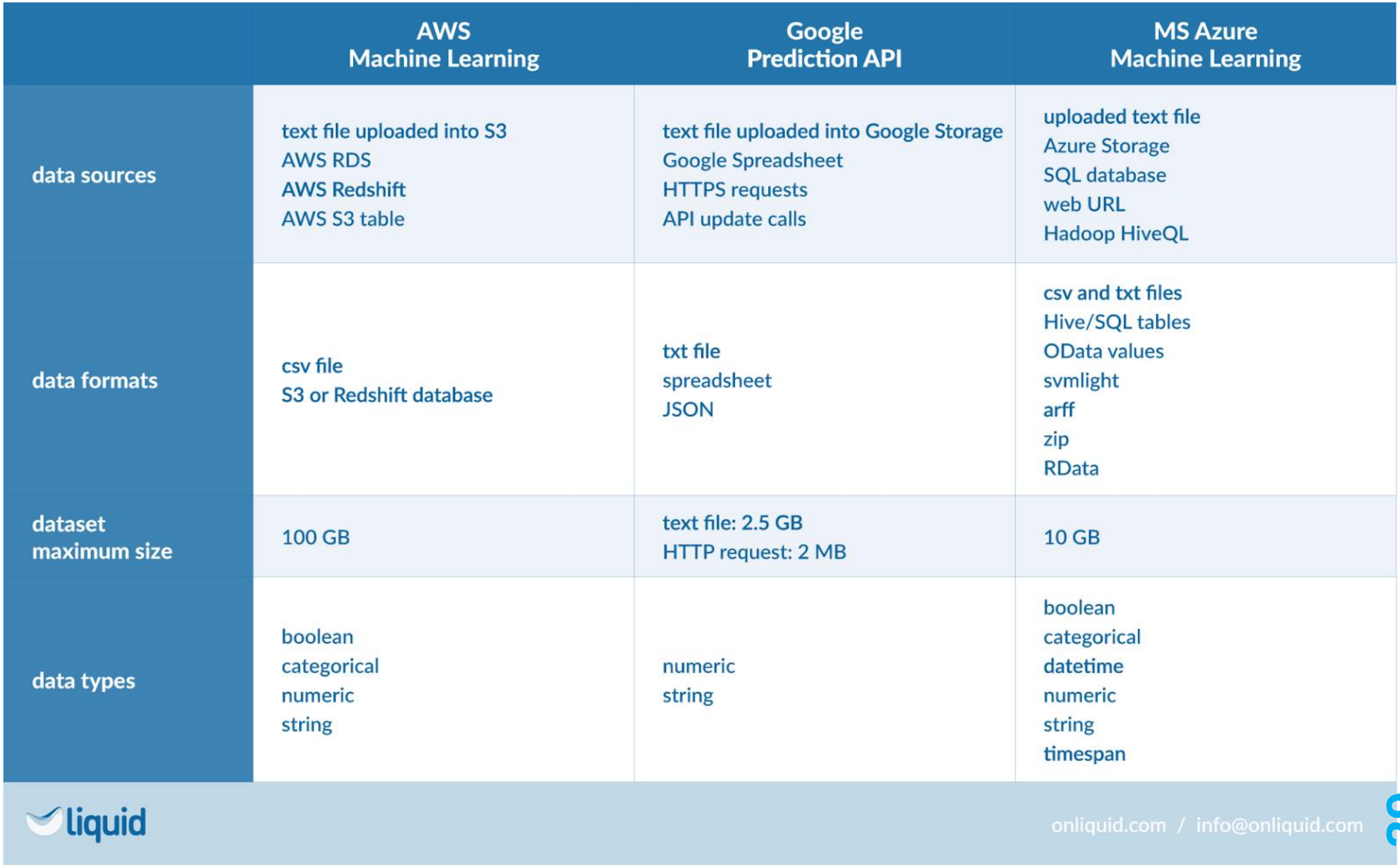
🟩 USED ▭ AVAILABLE

0.84 GB

8% of 10 GIGABYTES

Want more storage? Get the standard version learn more

Amazon-Google-MS 머신러닝 서비스 비교분석



	AWS Machine Learning	Google Prediction API	MS Azure Machine Learning
data sources	text file uploaded into S3 AWS RDS AWS Redshift AWS S3 table	text file uploaded into Google Storage Google Spreadsheet HTTPS requests API update calls	uploaded text file Azure Storage SQL database web URL Hadoop HiveQL
data formats	csv file S3 or Redshift database	txt file spreadsheet JSON	csv and txt files Hive/SQL tables OData values svmlight arff zip RData
dataset maximum size	100 GB	text file: 2.5 GB HTTP request: 2 MB	10 GB
data types	boolean categorical numeric string	numeric string	boolean categorical datetime numeric string timespan

liquid

onliquid.com / info@onliquid.com

참고자료

1. Microsoft Azure: https://azure.microsoft.com/ko-kr/
2. **Azure machine-learning studio: https://studio.azureml.net/**
3. Azure machine–learning 가격, 설명서: **https://azure.microsoft.com/ko-kr/services/machine-learning/**
4. Azure machine–learning 개요, 사용방법: **https://docs.microsoft.com/ko-kr/azure/machine-learning/machine-learning-algorithm-choice**

Thank you