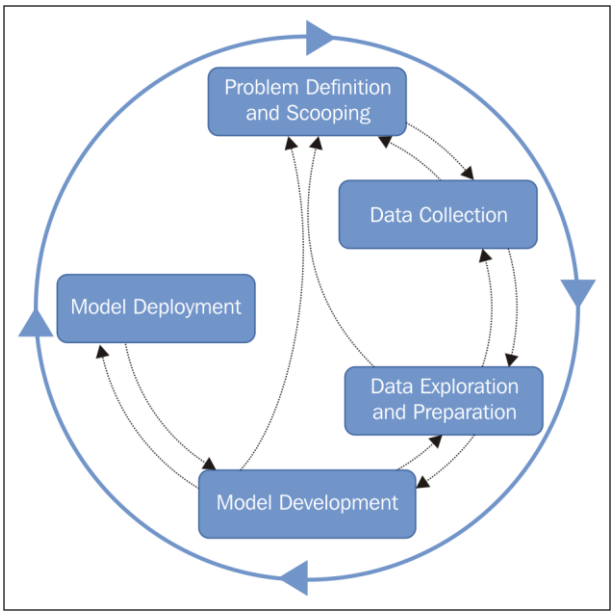
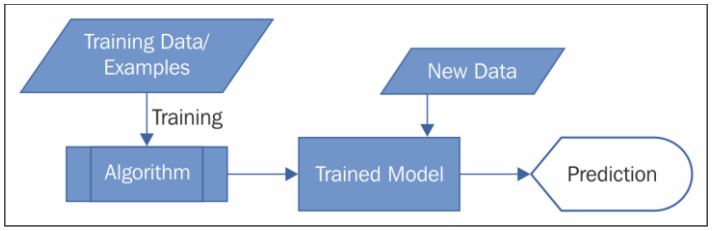
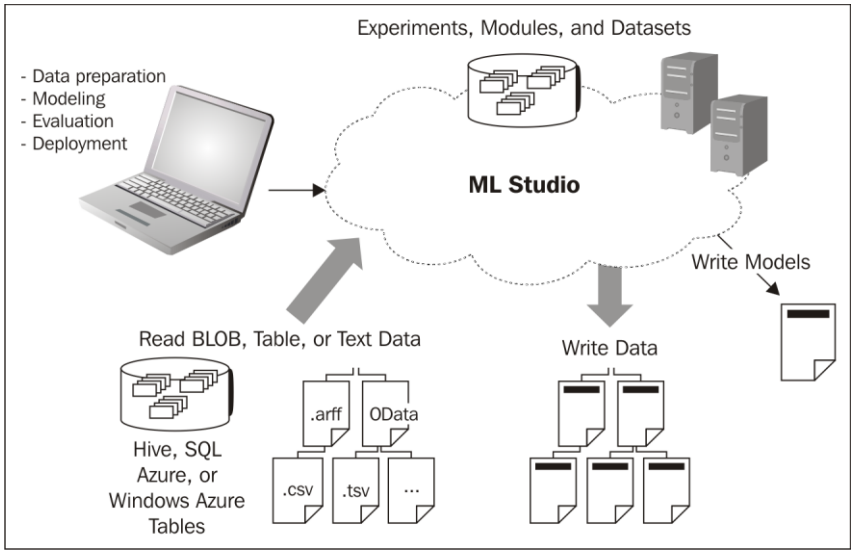
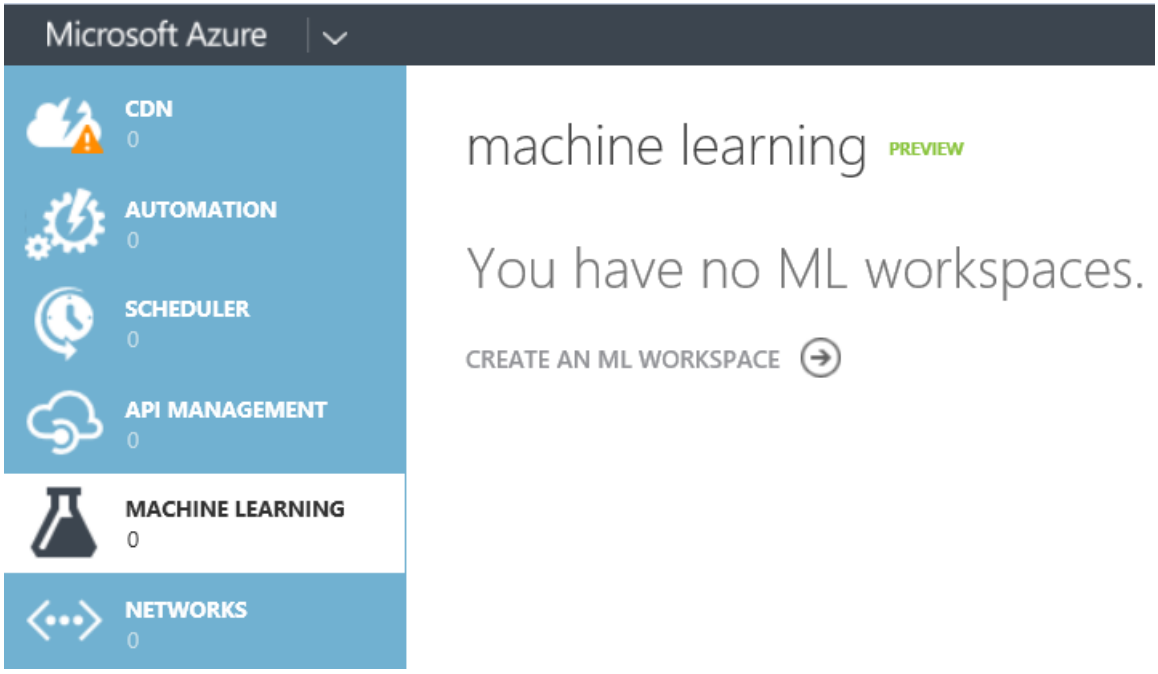
Chapter 1

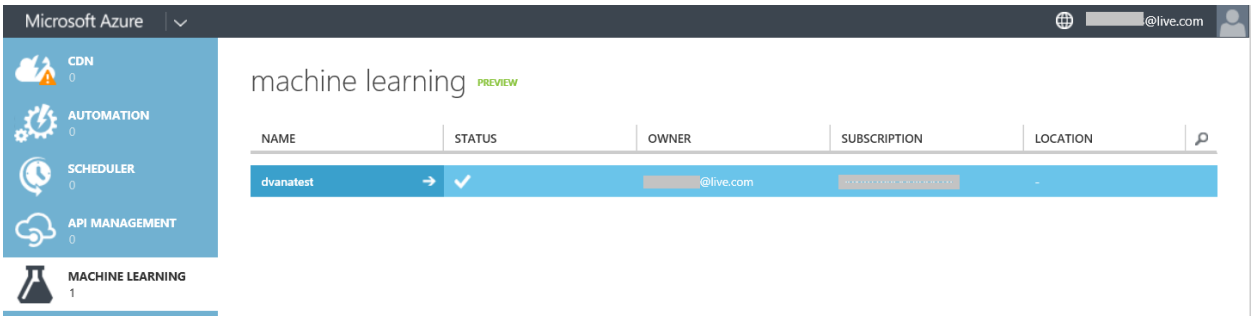


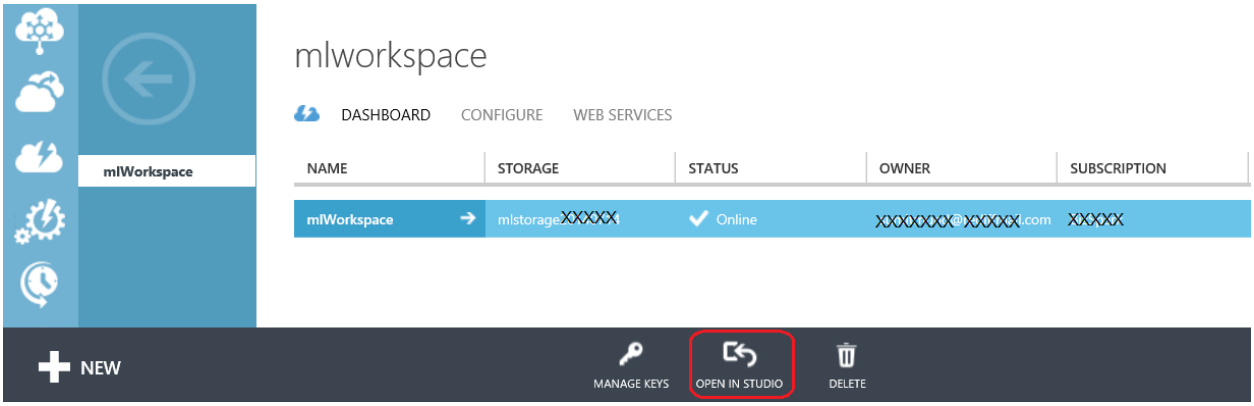


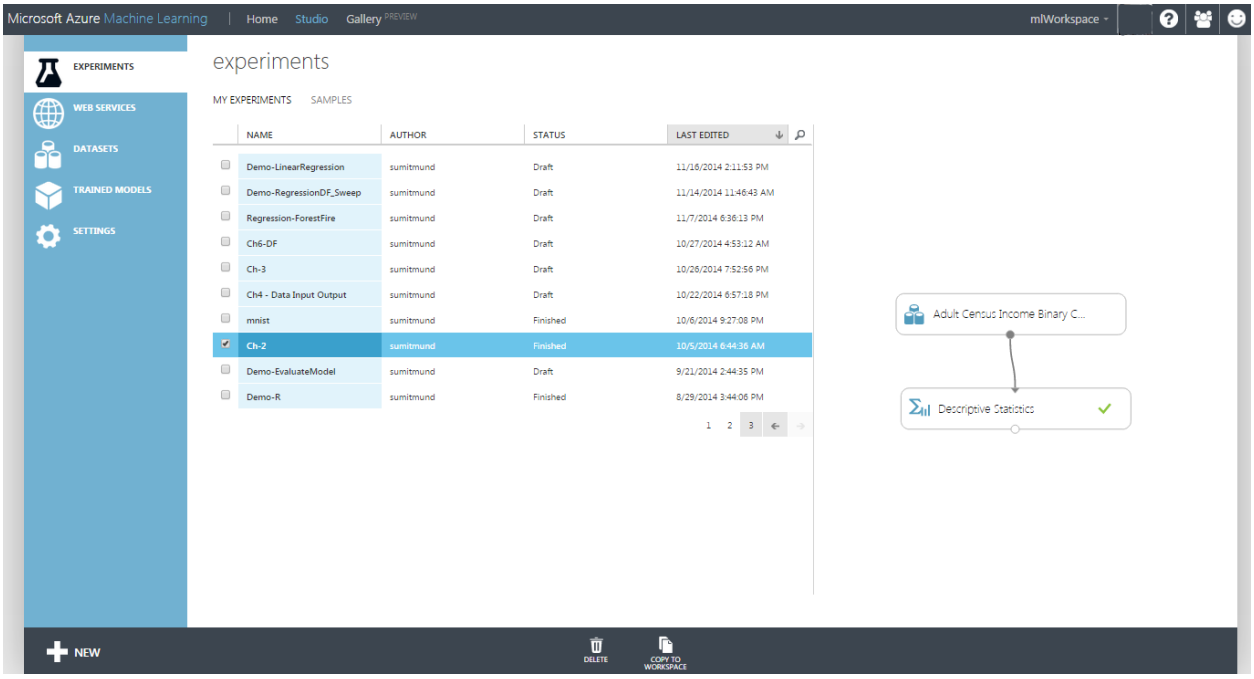
Chapter 2

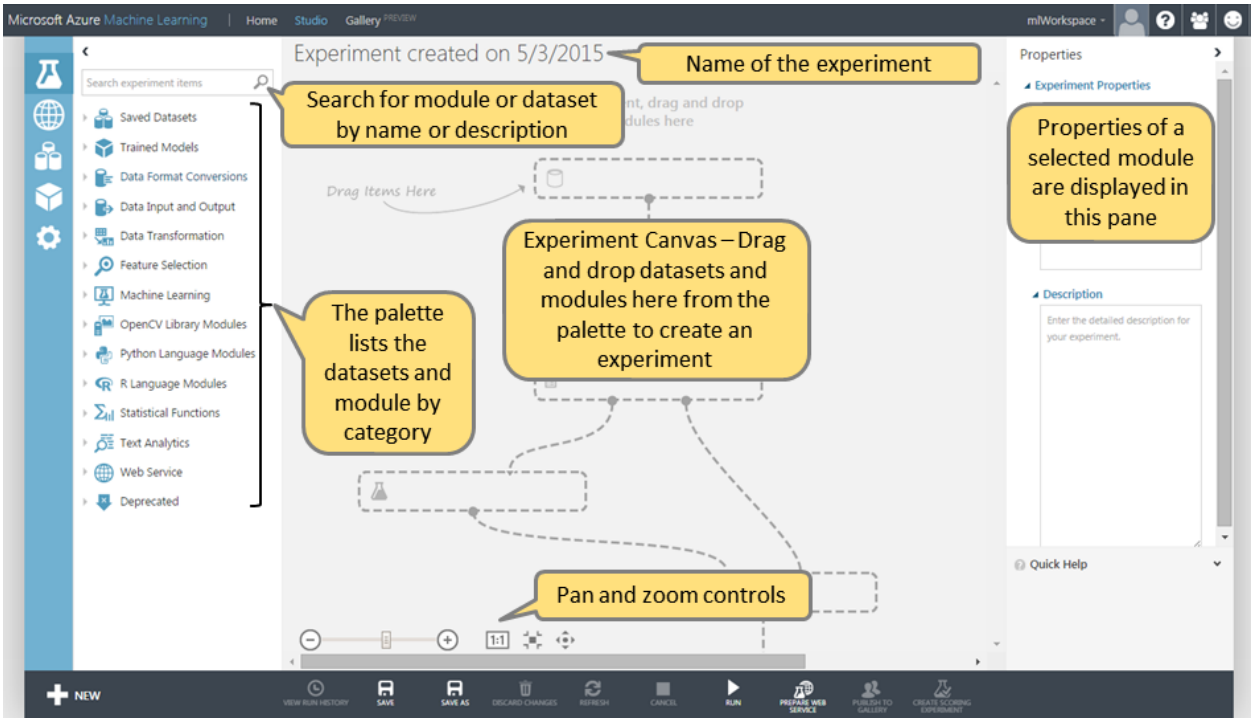

Experiments, Modules, and Datasets
- Data preparation
- Modeling
- Evaluation
- Deployment
ML Studio
Write Models
Read BLOB, Table, or Text Data
Write Data
Hive, SQL Azure, or Windows Azure Tables
.arff
OData
.csv
.tsv
...

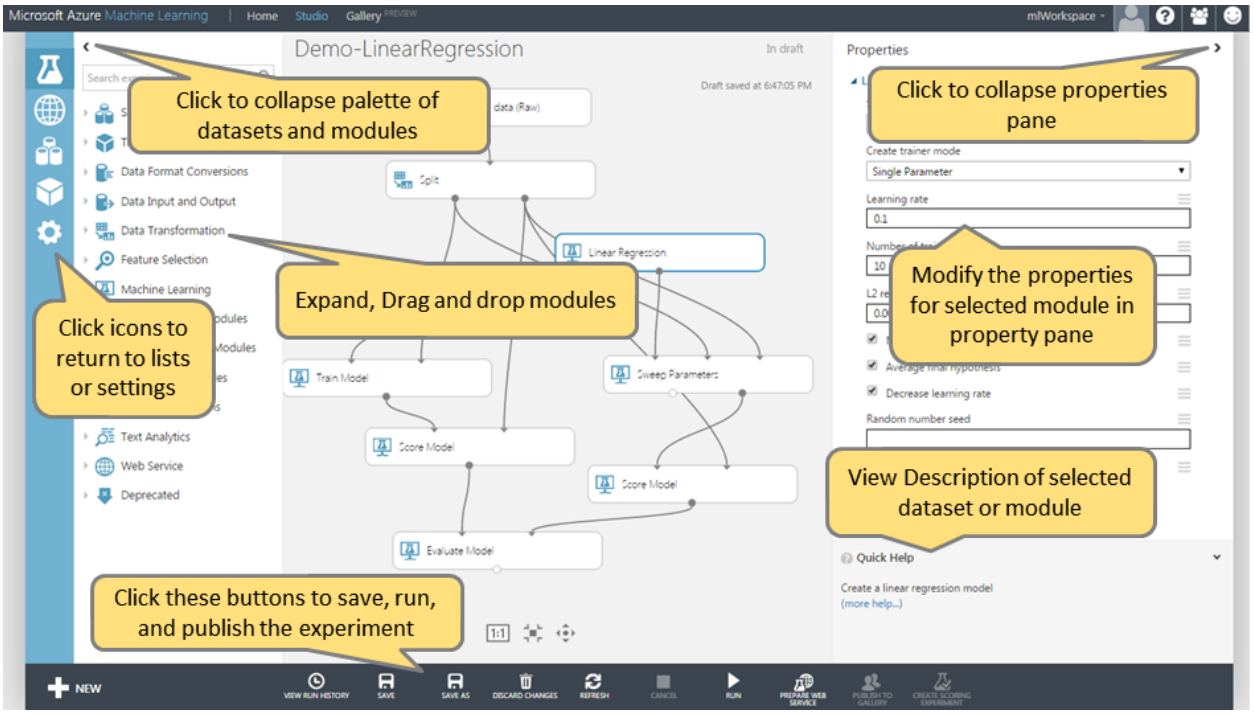

Microsoft Azure
CDN
0
AUTOMATION
0
SCHEDULER
0
API MANAGEMENT
0
MACHINE LEARNING
0
NETWORKS
0
machine learning PREVIEW
You have no ML workspaces.
CREATE AN ML WORKSPACE

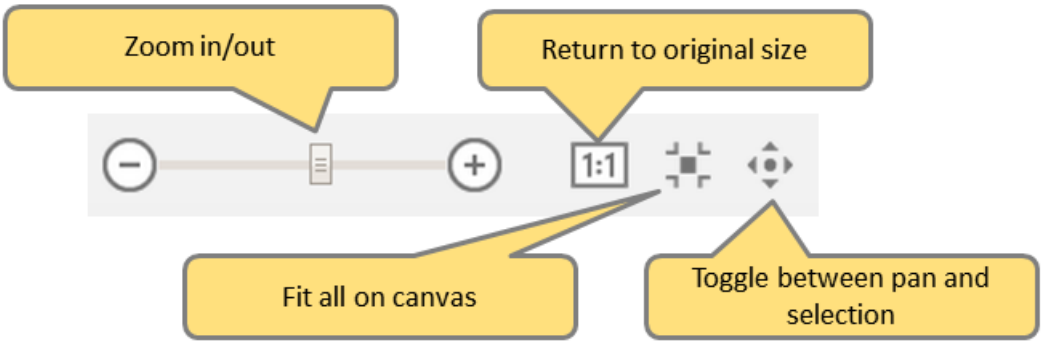

Microsoft Azure
@live.com
CDN
0
AUTOMATION
0
SCHEDULER
0
API MANAGEMENT
0
MACHINE LEARNING
1
machine learning PREVIEW
NAME
STATUS
OWNER
SUBSCRIPTION
LOCATION
dvanatest
@live.com
-

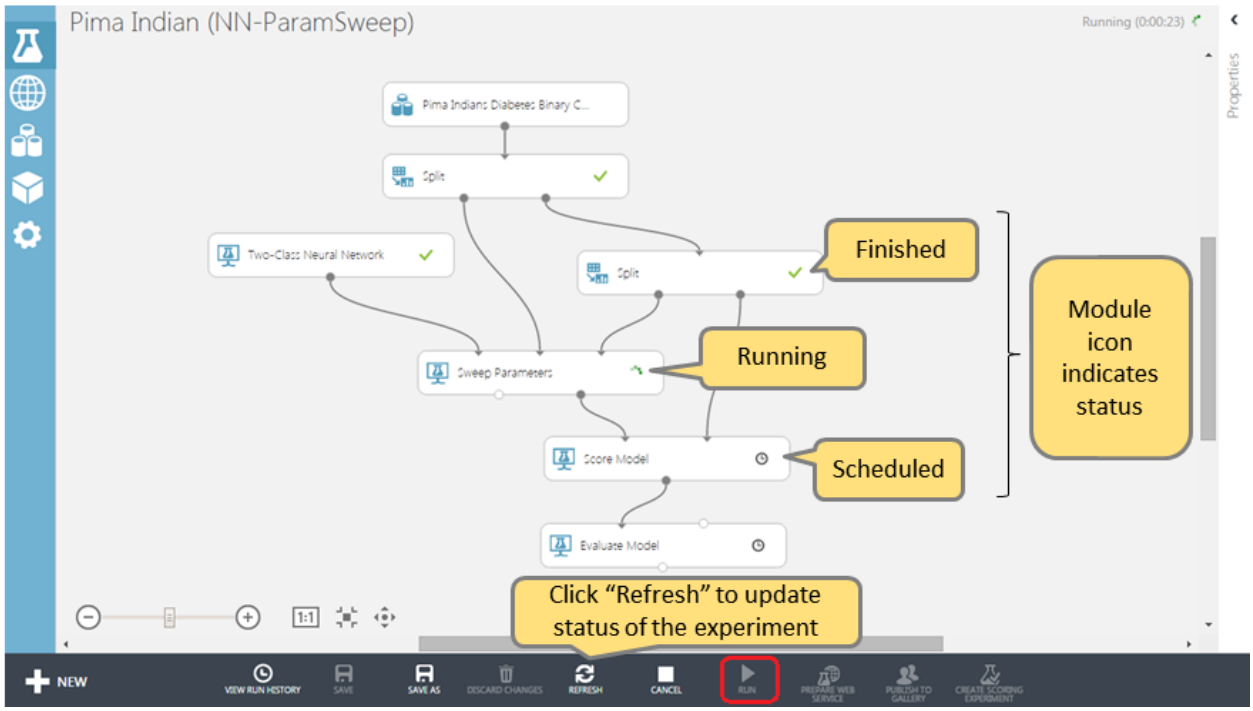

mlworkspace
DASHBOARD
CONFIGURE
WEB SERVICES
mlWorkspace
NAME
STORAGE
STATUS
OWNER
SUBSCRIPTION
mlWorkspace
mlstorageXXXXX
Online
XXXXXXXXXXXXXX.com
XXXXX
NEW
MANAGE KEYS
OPEN IN STUDIO
DELETE

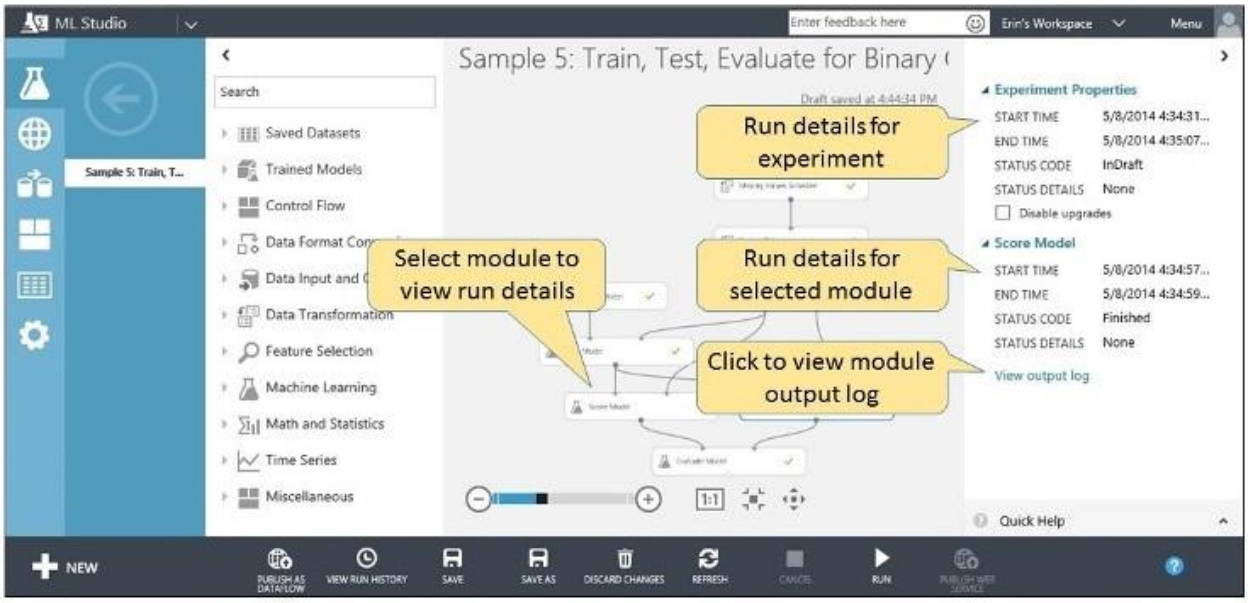

Microsoft Azure Machine Learning
Home
Studio
Gallery PREVIEW
mlWorkspace
EXPERIMENTS
WEB SERVICES
DATASETS
TRAINED MODELS
SETTINGS
experiments
MY EXPERIMENTS
SAMPLES
NAME
AUTHOR
STATUS
LAST EDITED
Demo-LinearRegression sumitmund Draft 11/16/2014 2:11:53 PM
Demo-RegressionDF_Sweep sumitmund Draft 11/14/2014 11:46:43 AM
Regression-ForestFire sumitmund Draft 11/7/2014 6:36:13 PM
Ch6-DF sumitmund Draft 10/27/2014 4:53:12 AM
Ch-3 sumitmund Draft 10/26/2014 7:52:56 PM
Ch4 - Data Input Output sumitmund Draft 10/22/2014 6:57:18 PM
mnist sumitmund Finished 10/6/2014 9:27:08 PM
Ch-2 sumitmund Finished 10/5/2014 6:44:36 AM
Demo-EvaluateModel sumitmund Draft 9/21/2014 2:44:35 PM
Demo-R sumitmund Finished 8/29/2014 3:44:06 PM
1 2 3
Adult Census Income Binary C...
Descriptive Statistics
NEW
DELETE
COPY TO WORKSPACE

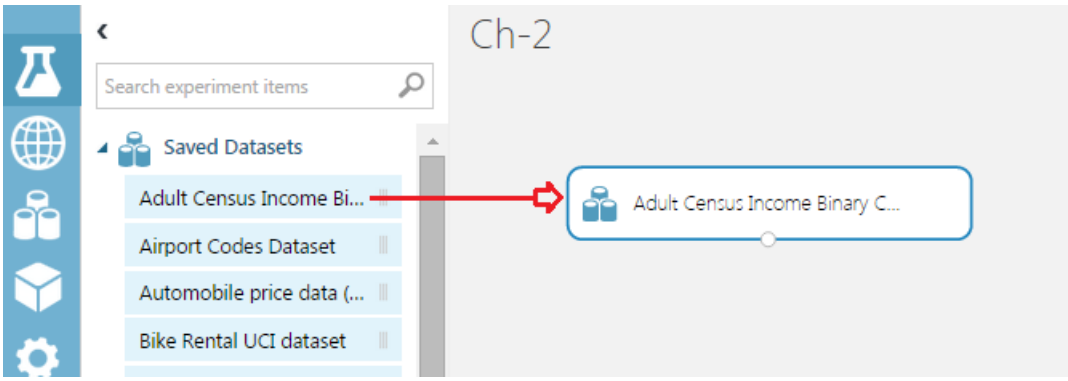

Microsoft Azure Machine Learning
Home
Studio
Gallery
mlWorkspace
Experiment created on 5/3/2015
Name of the experiment
Search for module or dataset by name or description
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Web Service
Deprecated
Drag Items Here
Experiment Canvas – Drag and drop datasets and modules here from the palette to create an experiment
The palette lists the datasets and module by category
Properties of a selected module are displayed in this pane
Properties
Experiment Properties
Description
Quick Help
Pan and zoom controls
NEW
SAVE
SAVE AS
RUN

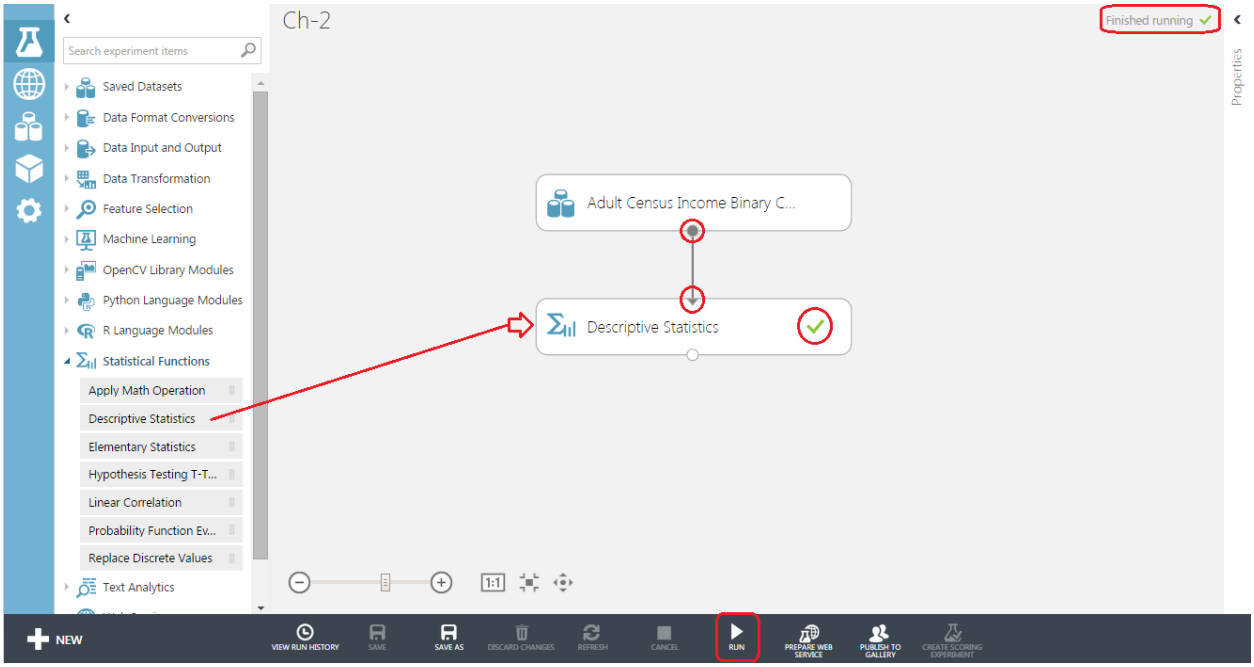

Microsoft Azure Machine Learning
Home
Studio
Gallery
mlWorkspace
Demo-LinearRegression
In draft
Draft saved at 6:47:05 PM
Click to collapse palette of datasets and modules
Click to collapse properties pane
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
Text Analytics
Web Service
Deprecated
Split
Linear Regression
Train Model
Sweep Parameters
Score Model
Score Model
Evaluate Model
Expand, Drag and drop modules
Click icons to return to lists or settings
Properties
Create trainer mode
Single Parameter
Learning rate
0.1
Decrease learning rate
Random number seed
Modify the properties for selected module in property pane
View Description of selected dataset or module
Quick Help
Create a linear regression model
(more help...)
Click these buttons to save, run, and publish the experiment
NEW
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
RUN

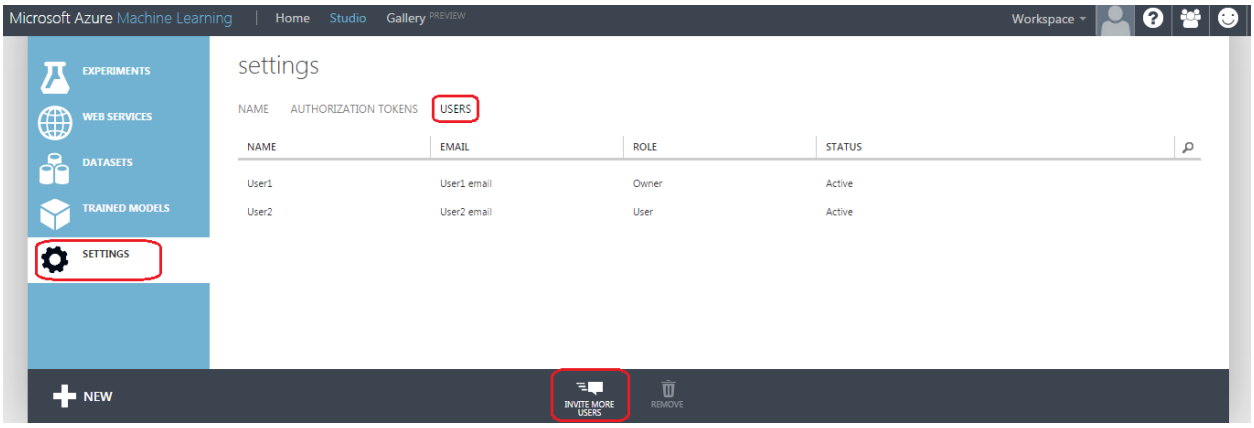

Zoom in/out
Return to original size
1:1
Fit all on canvas
Toggle between pan and selection


Pima Indian (NN-ParamSweep)
Running (0:00:23)
Properties
Pima Indians Diabetes Binary C...
Split
Two-Class Neural Network
Split
Finished
Sweep Parameters
Running
Score Model
Scheduled
Evaluate Model
Module icon indicates status
Click "Refresh" to update status of the experiment
NEW
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PREPARE WEB SERVICE
PUBLISH TO GALLERY
CREATE SCORING EXPERIMENT


ML Studio
Enter feedback here
Erin's Workspace
Menu
Sample 5: Train, Test, Evaluate for Binary
Search
Saved Datasets
Trained Models
Control Flow
Data Format Con
Data Input and
Data Transformation
Feature Selection
Machine Learning
Math and Statistics
Time Series
Miscellaneous
Run details for experiment
Select module to view run details
Run details for selected module
Click to view module output log
Experiment Properties
START TIME 5/8/2014 4:34:31...
END TIME 5/8/2014 4:35:07...
STATUS CODE InDraft
STATUS DETAILS None
Disable upgrades
Score Model
START TIME 5/8/2014 4:34:57...
END TIME 5/8/2014 4:34:59...
STATUS CODE Finished
STATUS DETAILS None
View output log
Quick Help
NEW
PUBLISH AS DATAFLOW
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PUBLISH WEB SERVICE

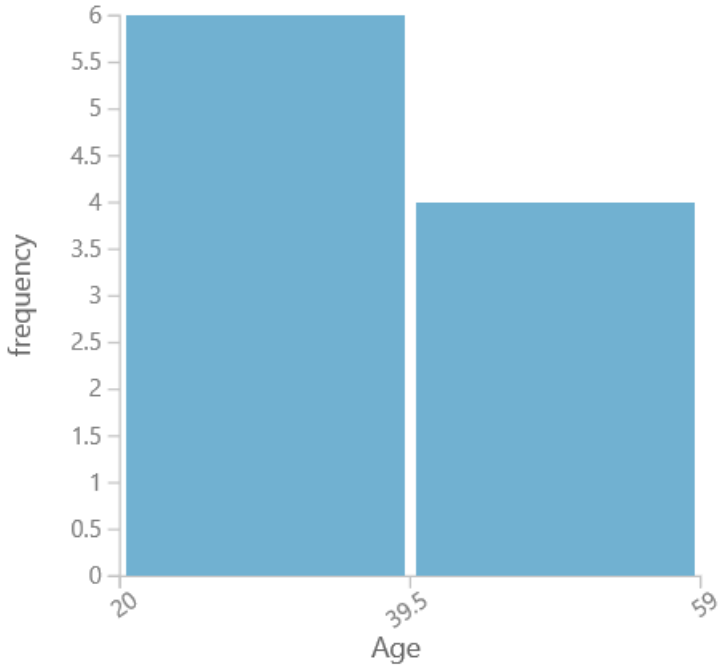

Ch-2
Search experiment items
Saved Datasets
Adult Census Income Bi...
Airport Codes Dataset
Automobile price data (...
Bike Rental UCI dataset
Adult Census Income Binary C...

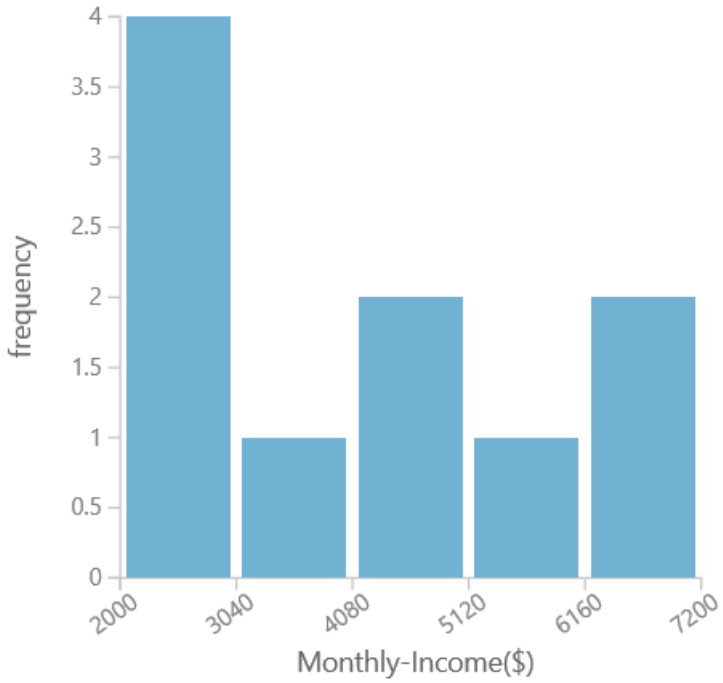

Ch-2
Finished running
Properties
Search experiment items
Saved Datasets
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Apply Math Operation
Descriptive Statistics
Elementary Statistics
Hypothesis Testing T-T...
Linear Correlation
Probability Function Ev...
Replace Discrete Values
Text Analytics
Adult Census Income Binary C...
Descriptive Statistics
NEW
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PREPARE WEB SERVICE
PUBLISH TO GALLERY
CREATE SCORING EXPERIMENT

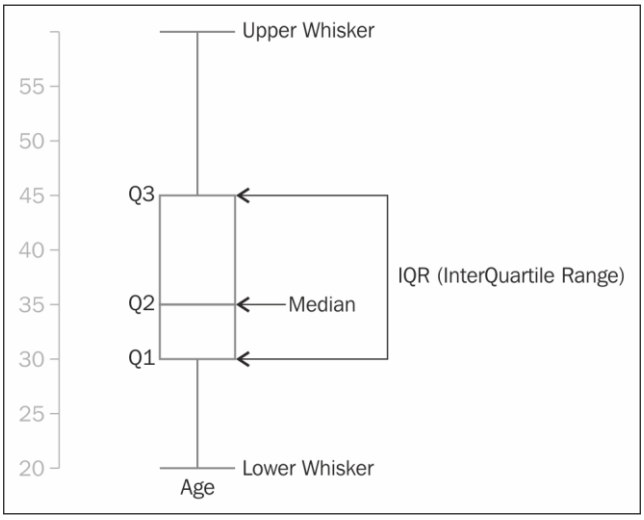

Microsoft Azure Machine Learning
Home
Studio
Gallery PREVIEW
Workspace
EXPERIMENTS
WEB SERVICES
DATASETS
TRAINED MODELS
SETTINGS
settings
NAME
AUTHORIZATION TOKENS
USERS
NAME
EMAIL
ROLE
STATUS
User1
User1 email
Owner
Active
User2
User2 email
User
Active
NEW
INVITE MORE USERS
REMOVE

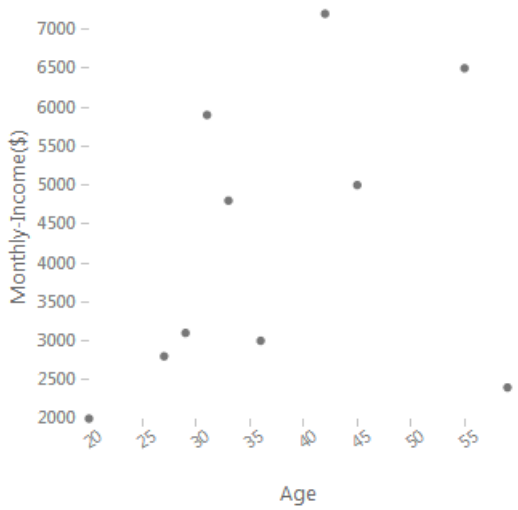
Chapter 3

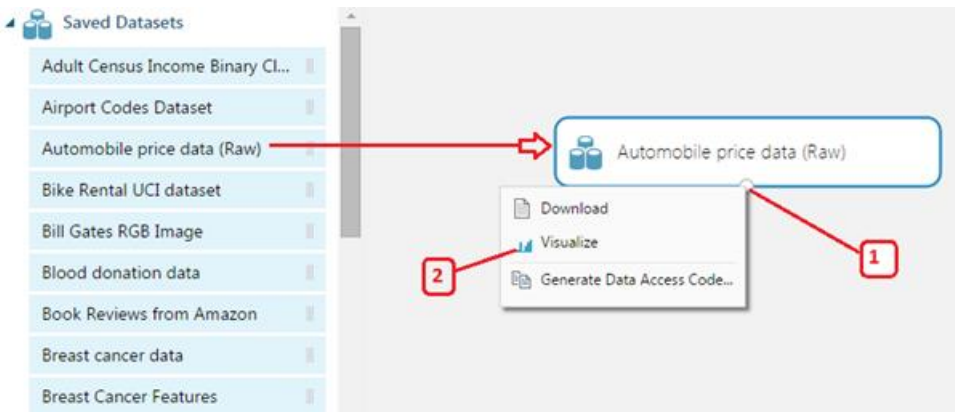
	Name	Age	Gender	Monthly-Income($)
Mean		37.7		4270
Median		34.5		3950
Min		20		2000
Max		59		7200
Standard Deviation		12.4637		1850.5555
Unique Values	10	10	2	10
Missing Values	0	0	0	0
Feature Type	String	Numeric	String	Numeric




frequency
4
3.5
3
2.5
2
1.5
1
0.5
0
2000
3040
4080
5120
6160
7200
Monthly-Income($)


Upper Whisker
55
50
45
40
35
30
25
20
Q3
Q2
Q1
Median
IQR (InterQuartile Range)
Lower Whisker
Age


Monthly-Income($)
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
20
25
30
35
40
45
50
55
Age


Saved Datasets
Adult Census Income Binary Cl...
Airport Codes Dataset
Automobile price data (Raw)
Bike Rental UCI dataset
Bill Gates RGB Image
Blood donation data
Book Reviews from Amazon
Breast cancer data
Breast Cancer Features
Automobile price data (Raw)
Download
Visualize
Generate Data Access Code...
1
2

rows 205 columns 26

view as

symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	heig
3		alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.
3		alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.
1		alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.
2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.
2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.
2		audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.
1	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.
1		audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.
1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.
0		audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52
2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.
0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.
0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.
0	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.
1		bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.
0		bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.
0		bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.
0		bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56.
2	121	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.
1	98	chevrolet	gas	std	two	hatchback	fwd	front	94.5	155.9	63.6	52

Statistics

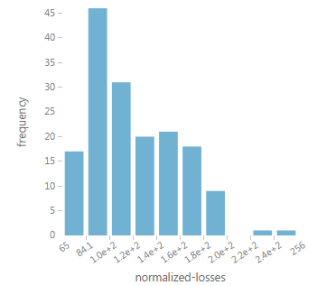
Mean	122
Median	115
Min	65
Max	256
Standard Deviation	35.4422
Unique Values	51
Missing Values	41
Feature Type	Numeric Feature

Visualizations

normalized-losses

Histogram

compare to None

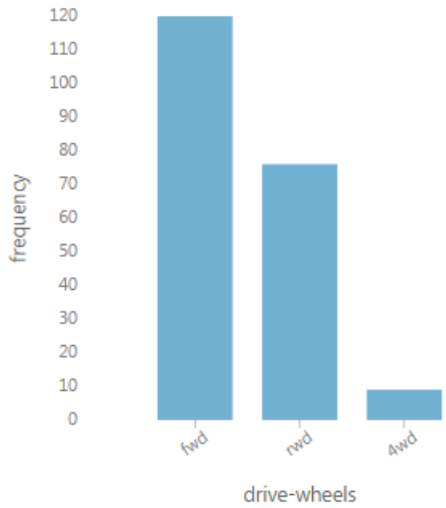


Visualizations

drive-wheels

Histogram

compare to None



frequency log scale

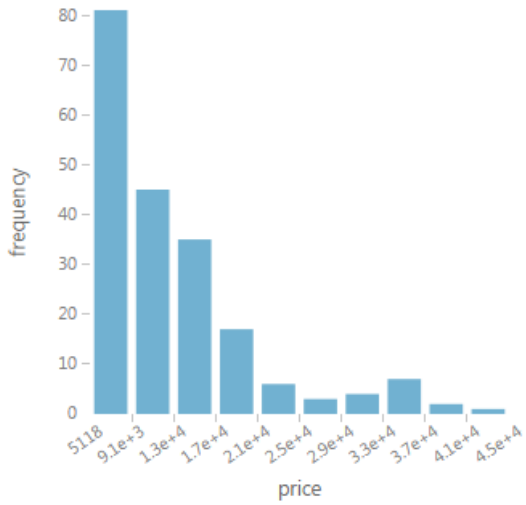
bins

10

price

Histogram

compare to None



- [ ] price log scale
- [ ] frequency log scale

bins

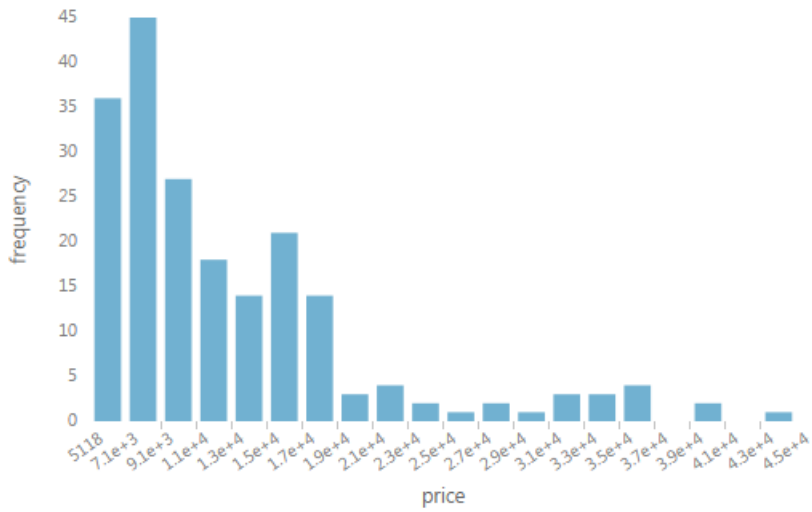
10

- [ ] cumulative distribution
- [ ] probability density

price

Histogram

compare to None



price log scale

frequency log scale

bins

20

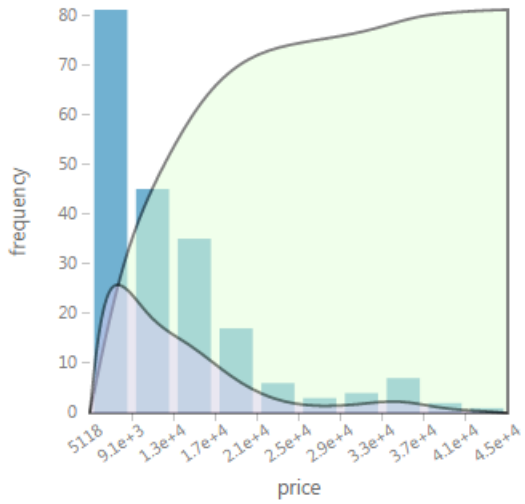
cumulative distribution

probability density

price

Histogram

compare to None



- [ ] price log scale
- [ ] frequency log scale

bins

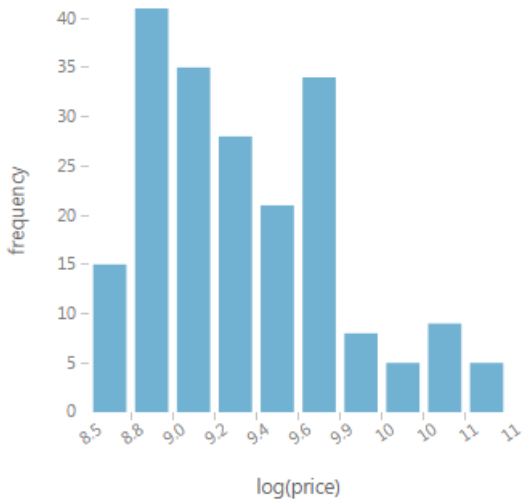
10

- [x] cumulative distribution
- [x] probability density

price

Histogram

compare to None

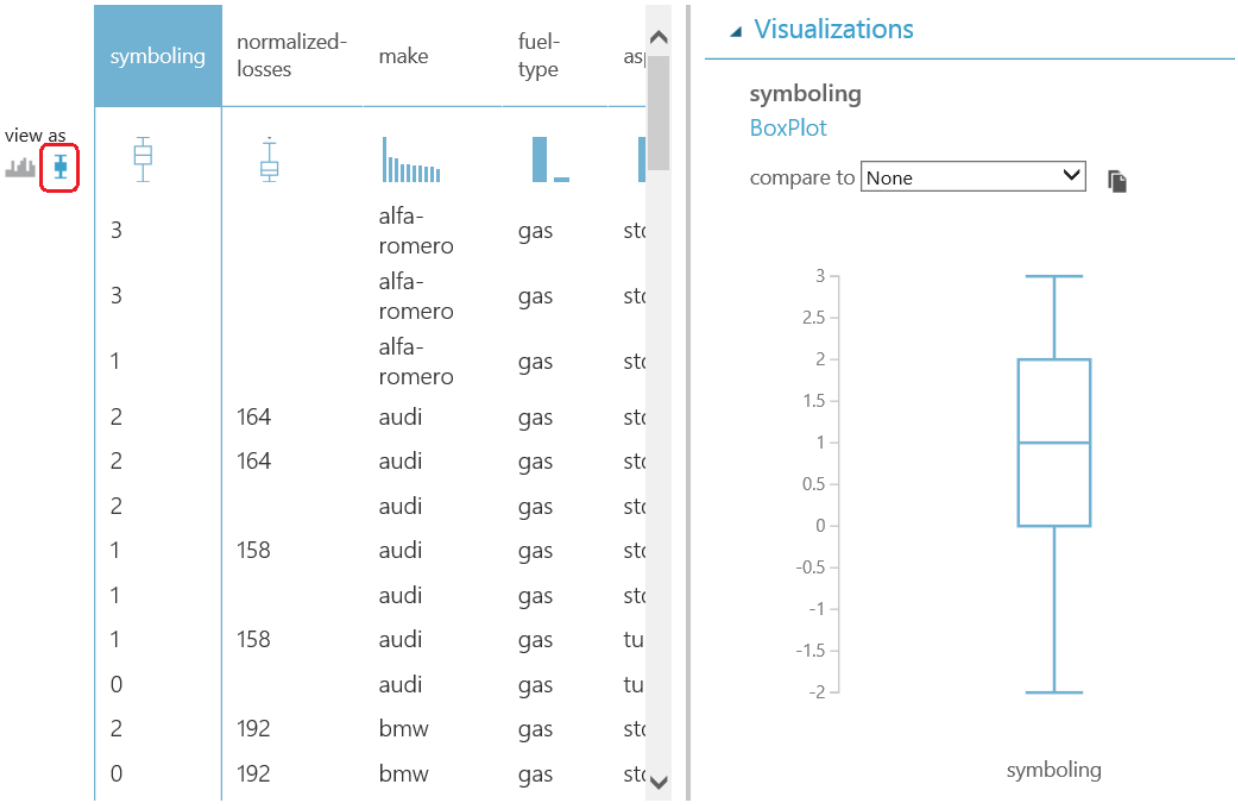


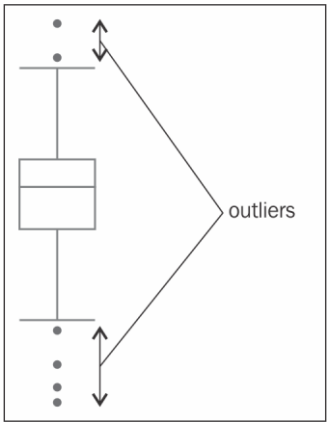
- [x] price log scale
- [ ] frequency log scale

bins

10

- [ ] cumulative distribution
- [ ] probability density


view as
symboling
normalized-losses
make
fuel-type
Visualizations
symboling
BoxPlot
compare to
None
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
symboling

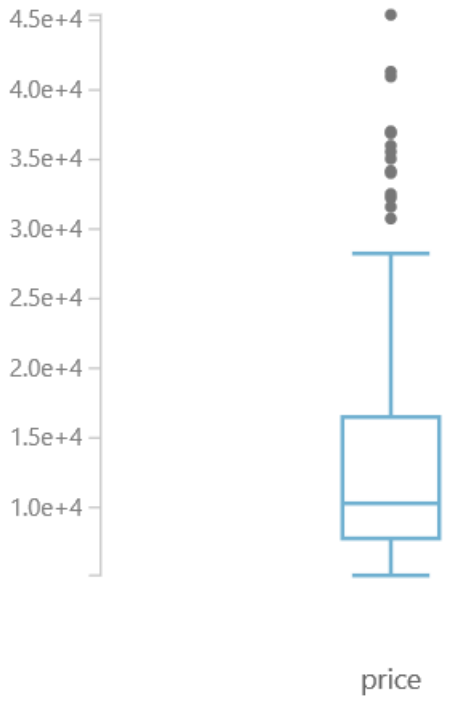

outliers

Visualizations

price

BoxPlot

compare to None



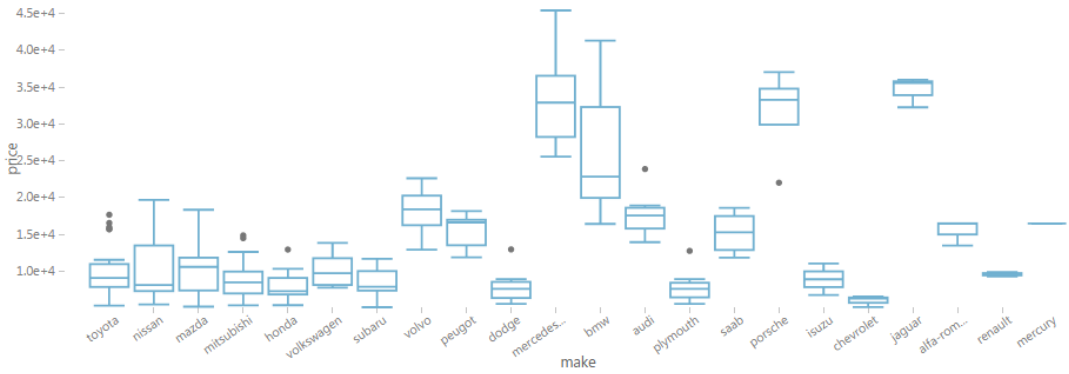
price log scale

Visualizations

price

MultiboxPlot

compare to make



price log scale

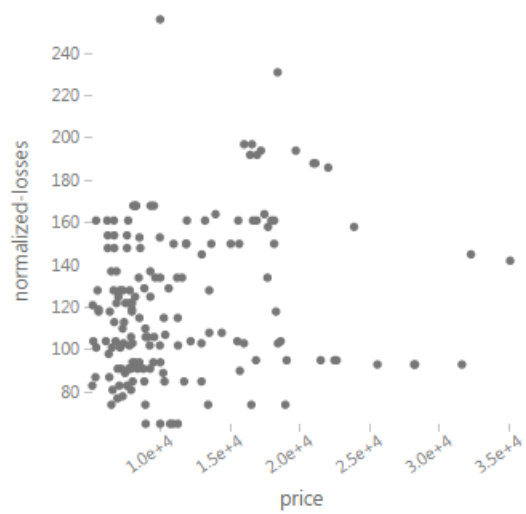
categories

55

price

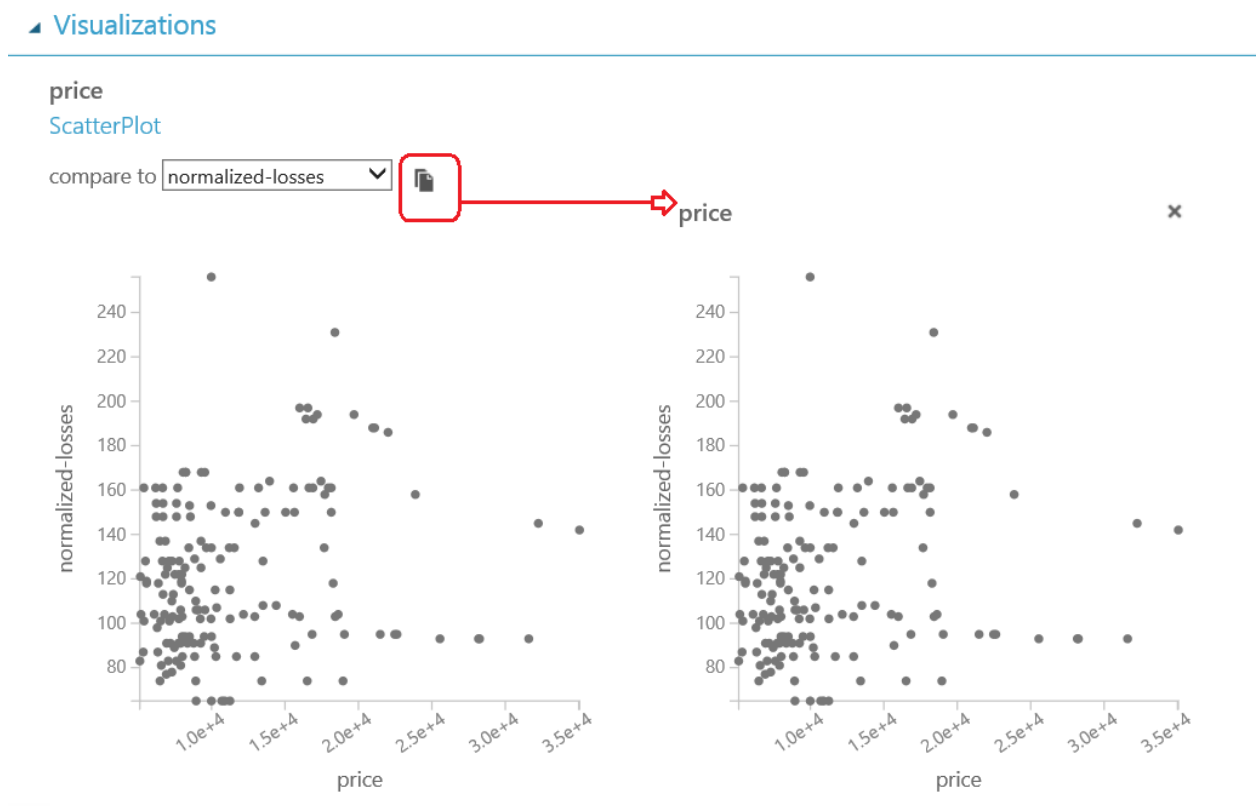
ScatterPlot

compare to normalized-losses

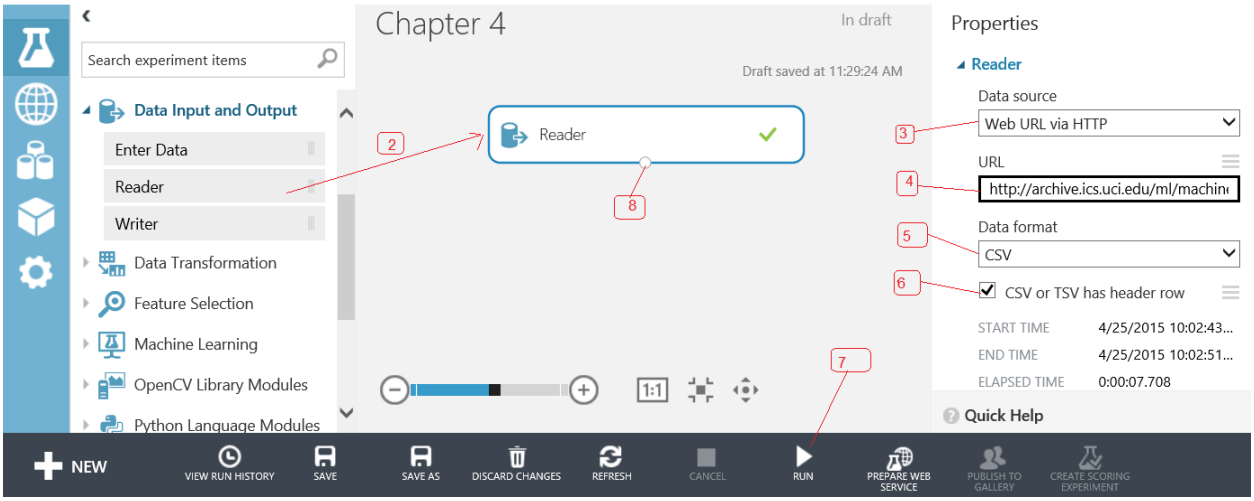


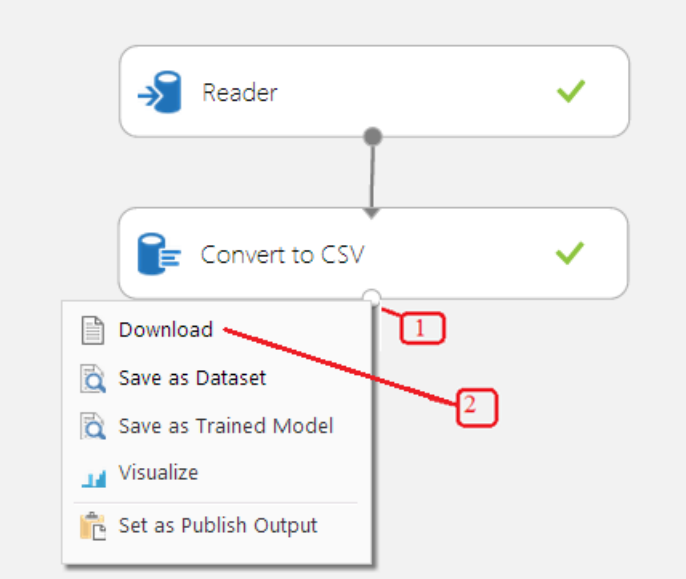
price log scale

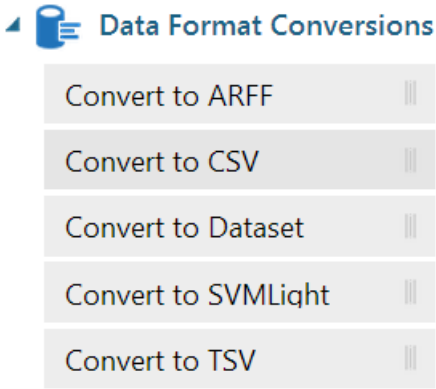
normalized-losses log scale

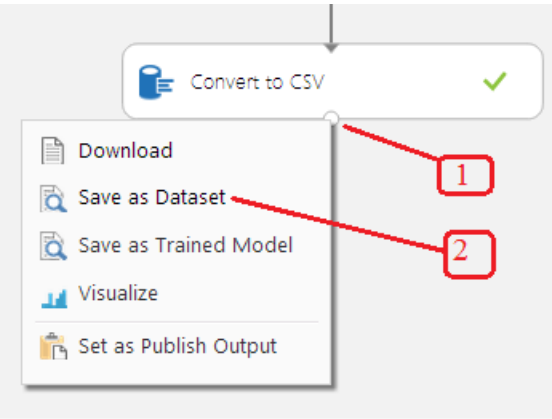

Visualizations
price
ScatterPlot
compare to
normalized-losses
price
normalized-losses
240
220
200
180
160
140
120
100
80
1.0e+4
1.5e+4
2.0e+4
2.5e+4
3.0e+4
3.5e+4
price
normalized-losses
240
220
200
180
160
140
120
100
80
1.0e+4
1.5e+4
2.0e+4
2.5e+4
3.0e+4
3.5e+4
price

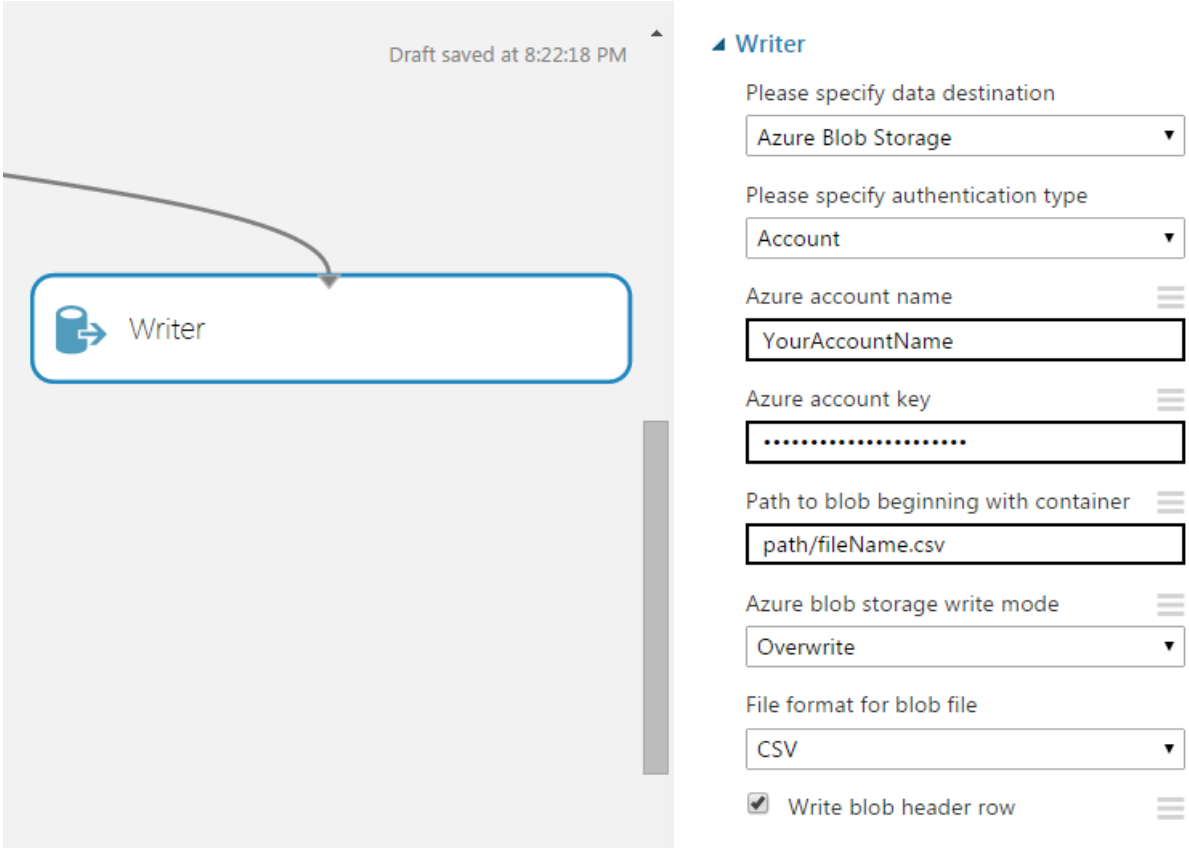
Chapter 4

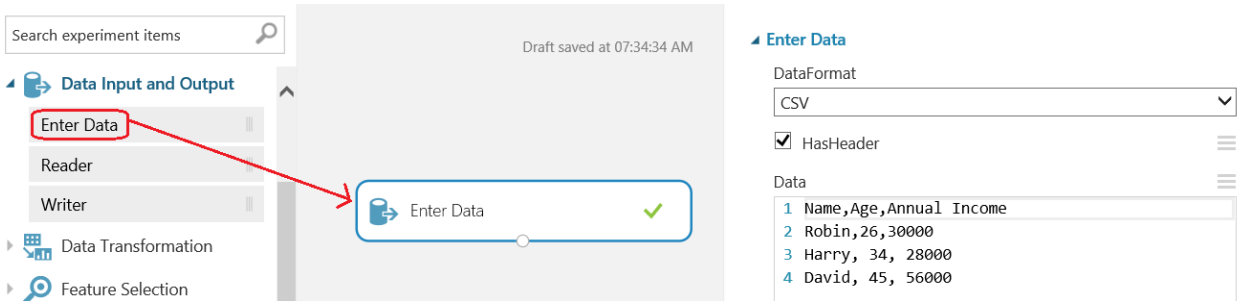

Search experiment items
Data Input and Output
Enter Data
Reader
Writer
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
Chapter 4
In draft
Draft saved at 11:29:24 AM
Reader
Properties
Reader
Data source
Web URL via HTTP
URL
http://archive.ics.uci.edu/ml/machine
Data format
CSV
CSV or TSV has header row
START TIME 4/25/2015 10:02:43...
END TIME 4/25/2015 10:02:51...
ELAPSED TIME 0:00:07.708
Quick Help
NEW
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PREPARE WEB SERVICE
PUBLISH TO GALLERY
CREATE SCORING EXPERIMENT
1
2
3
4
5
6
7
8


Reader
Convert to CSV
Download
Save as Dataset
Save as Trained Model
Visualize
Set as Publish Output
1
2

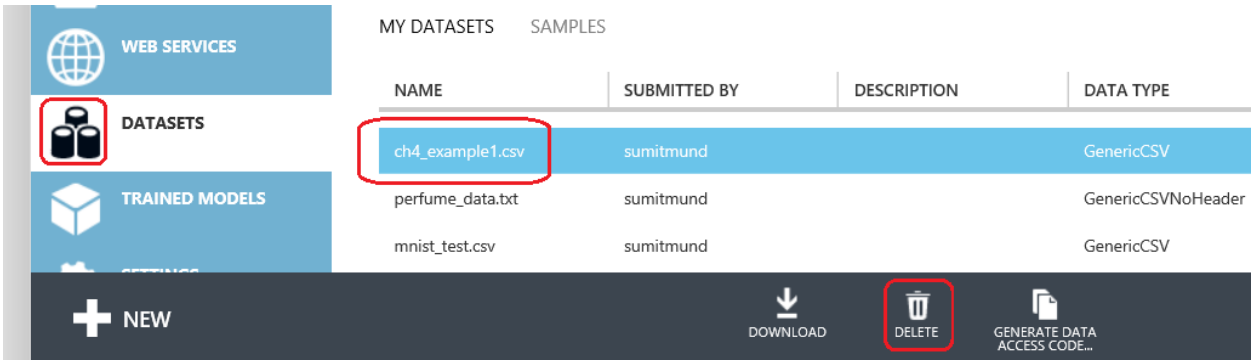

Data Format Conversions
Convert to ARFF
Convert to CSV
Convert to Dataset
Convert to SVMLight
Convert to TSV

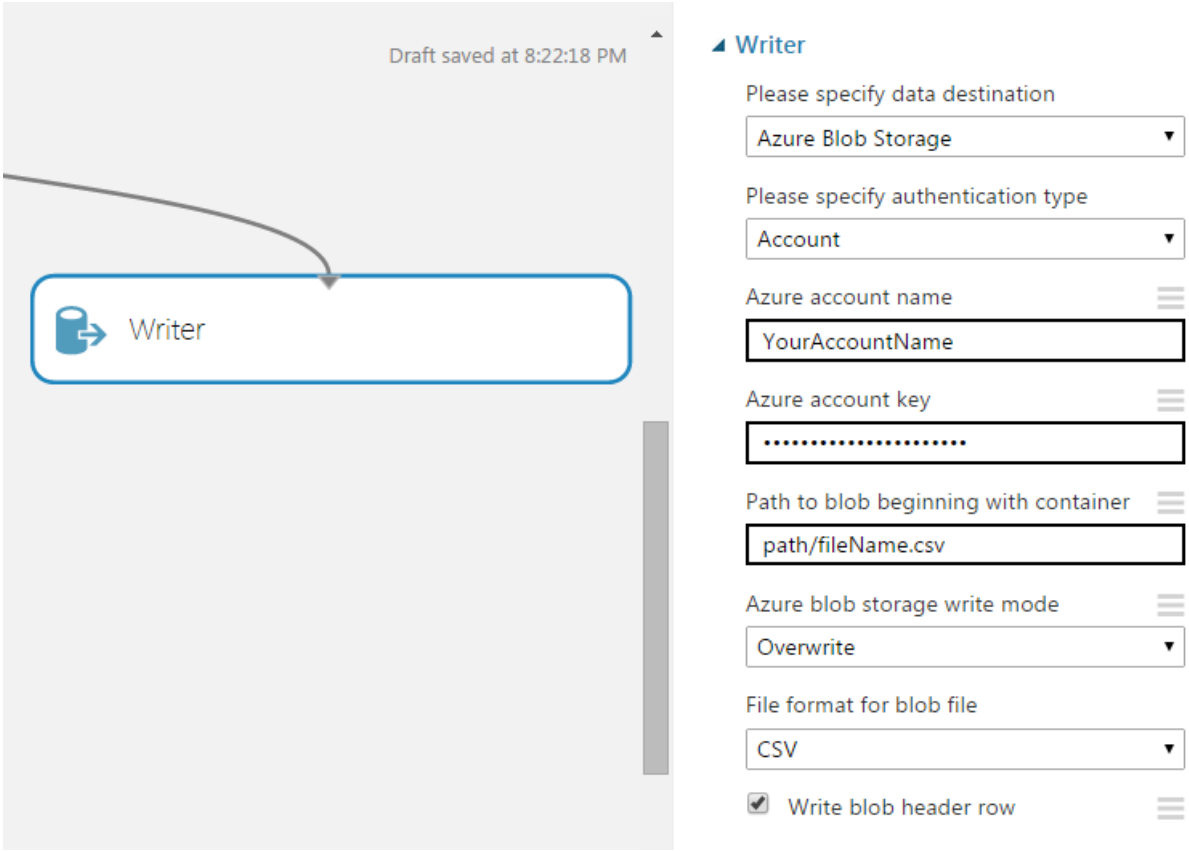

Convert to CSV
Download
Save as Dataset
Save as Trained Model
Visualize
Set as Publish Output
1
2


Draft saved at 8:22:18 PM
Writer
Writer
Please specify data destination
Azure Blob Storage
Please specify authentication type
Account
Azure account name
YourAccountName
Azure account key
Path to blob beginning with container
path/fileName.csv
Azure blob storage write mode
Overwrite
File format for blob file
CSV
Write blob header row

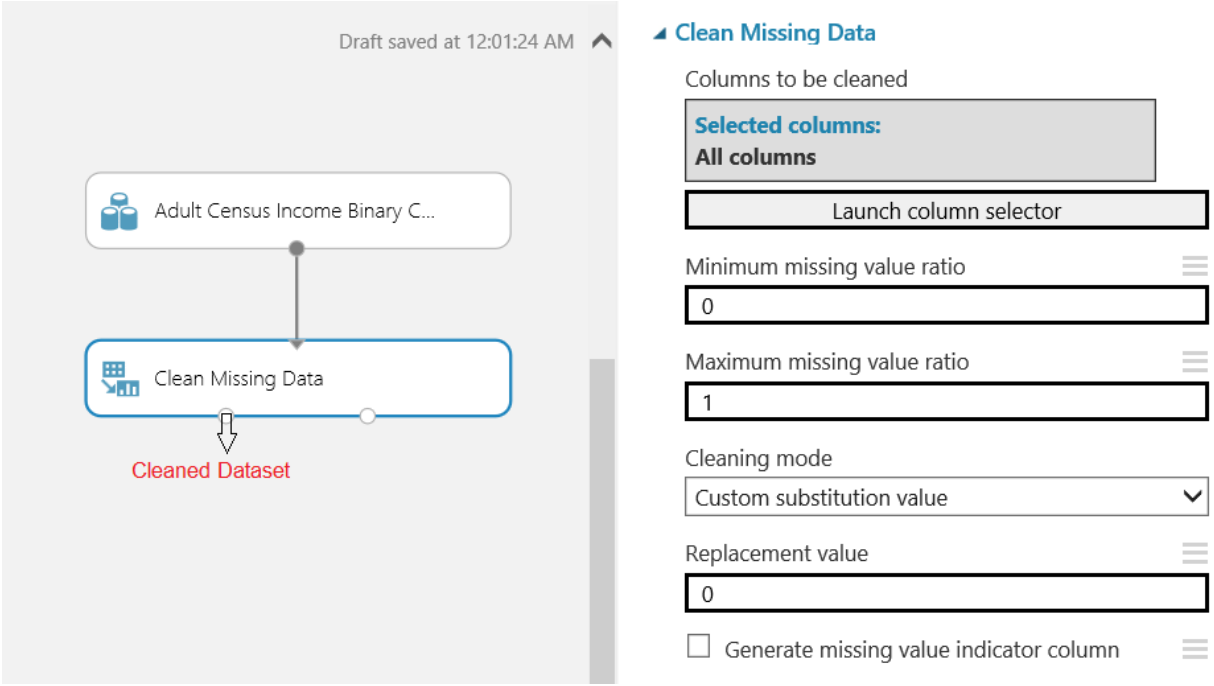

Search experiment items
Data Input and Output
Enter Data
Reader
Writer
Data Transformation
Feature Selection
Draft saved at 07:34:34 AM
Enter Data
Enter Data
DataFormat
CSV
HasHeader
Data
1 Name,Age,Annual Income
2 Robin,26,30000
3 Harry, 34, 28000
4 David, 45, 56000

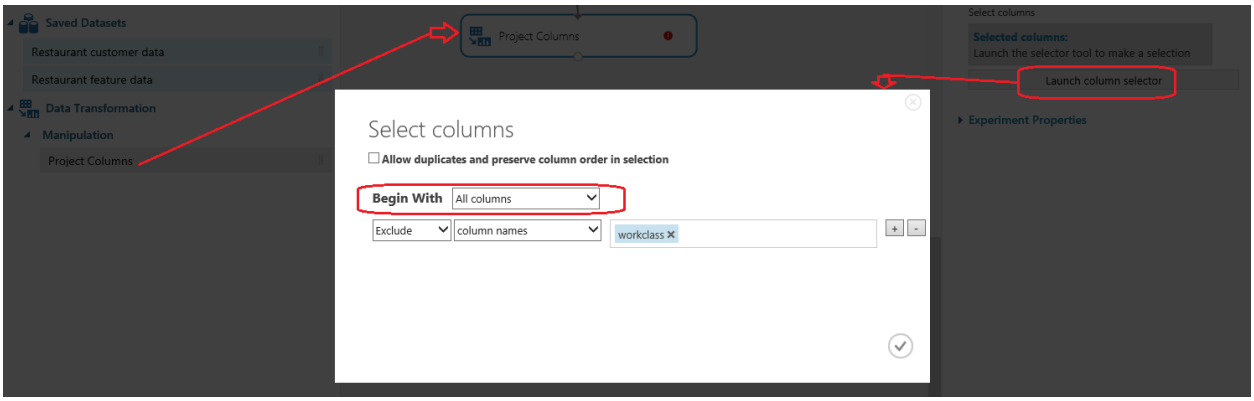

Saved Datasets
My Datasets
ch4_example1.csv
ch4_example1.csv

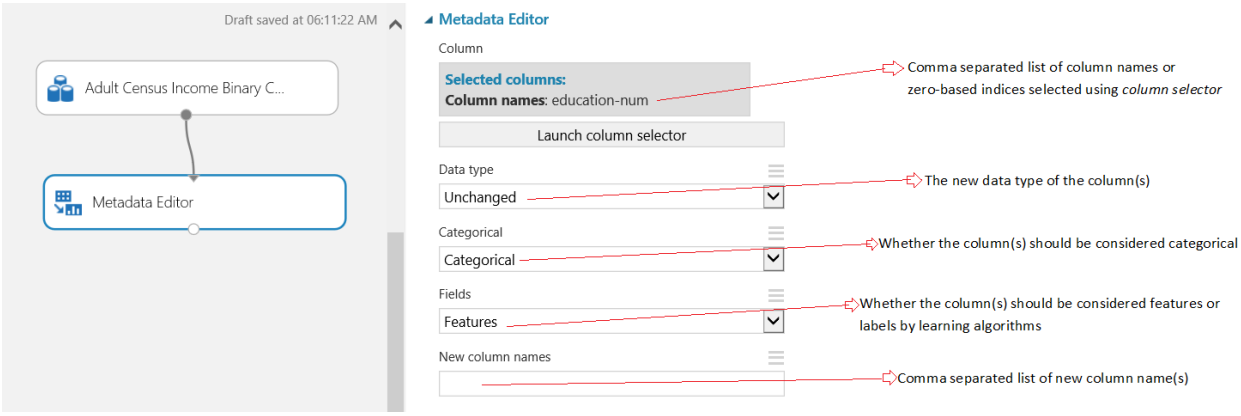

WEB SERVICES
DATASETS
TRAINED MODELS
MY DATASETS
SAMPLES
NAME
SUBMITTED BY
DESCRIPTION
DATA TYPE
ch4_example1.csv
sumitmund
GenericCSV
perfume_data.txt
sumitmund
GenericCSVNoHeader
mnist_test.csv
sumitmund
GenericCSV
NEW
DOWNLOAD
DELETE
GENERATE DATA ACCESS CODE...

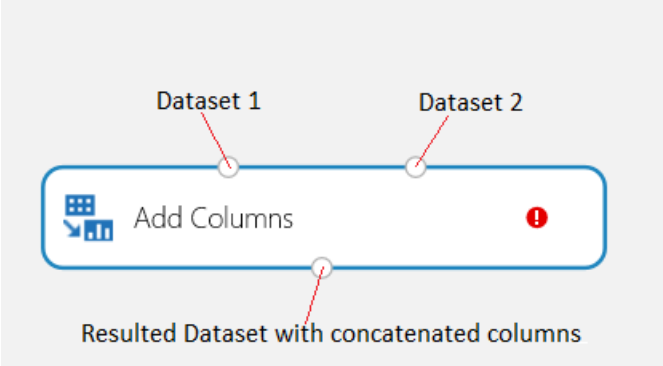

Draft saved at 8:22:18 PM
Writer
Writer
Please specify data destination
Azure Blob Storage
Please specify authentication type
Account
Azure account name
YourAccountName
Azure account key
Path to blob beginning with container
path/fileName.csv
Azure blob storage write mode
Overwrite
File format for blob file
CSV
Write blob header row

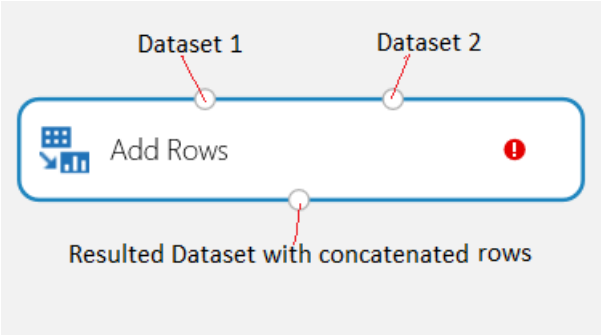
Chapter 5

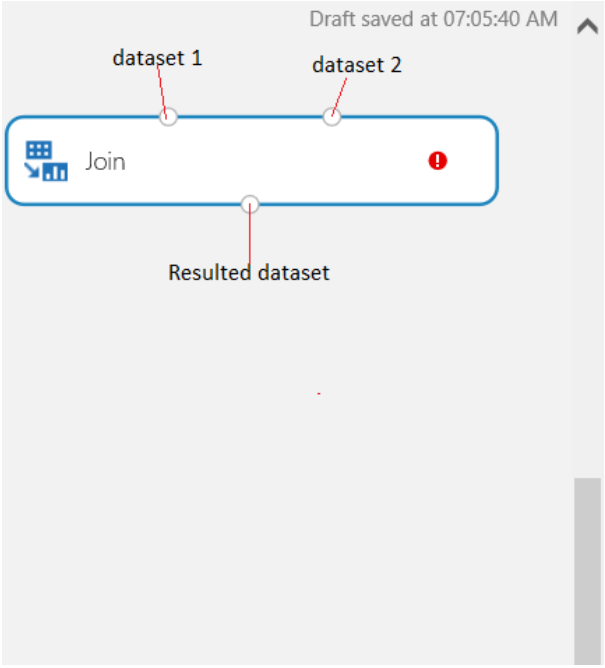

Draft saved at 12:01:24 AM
Adult Census Income Binary C...
Clean Missing Data
Cleaned Dataset
Clean Missing Data
Columns to be cleaned
Selected columns:
All columns
Launch column selector
Minimum missing value ratio
0
Maximum missing value ratio
1
Cleaning mode
Custom substitution value
Replacement value
0
Generate missing value indicator column

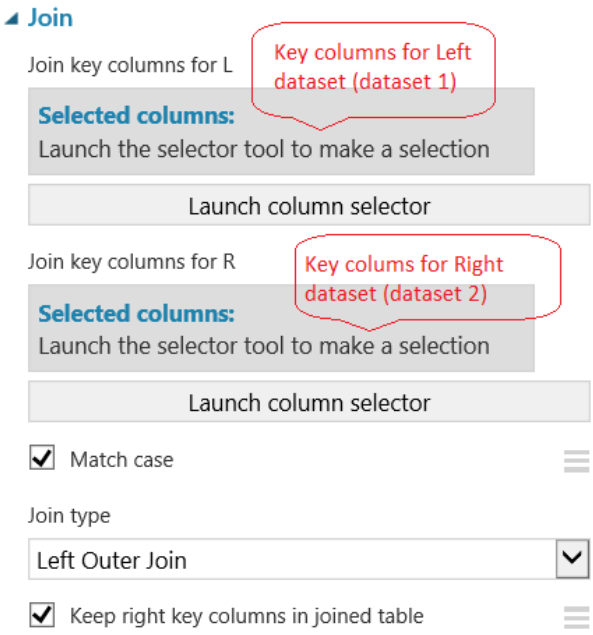

Saved Datasets
Restaurant customer data
Restaurant feature data
Data Transformation
Manipulation
Project Columns
Project Columns
Select columns
Allow duplicates and preserve column order in selection
Begin With
All columns
Exclude
column names
workclass
Selected columns:
Launch the selector tool to make a selection
Launch column selector
Experiment Properties

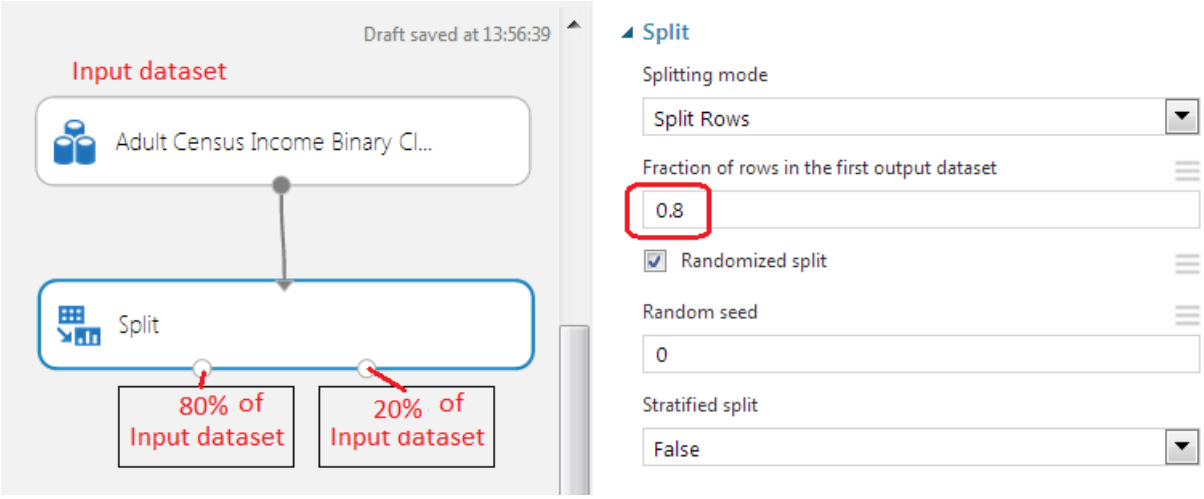

Draft saved at 06:11:22 AM
Adult Census Income Binary C...
Metadata Editor
Metadata Editor
Column
Selected columns:
Column names: education-num
Launch column selector
Data type
Unchanged
Categorical
Categorical
Fields
Features
New column names
Comma separated list of column names or zero-based indices selected using column selector
The new data type of the column(s)
Whether the column(s) should be considered categorical
Whether the column(s) should be considered features or labels by learning algorithms
Comma separated list of new column name(s)

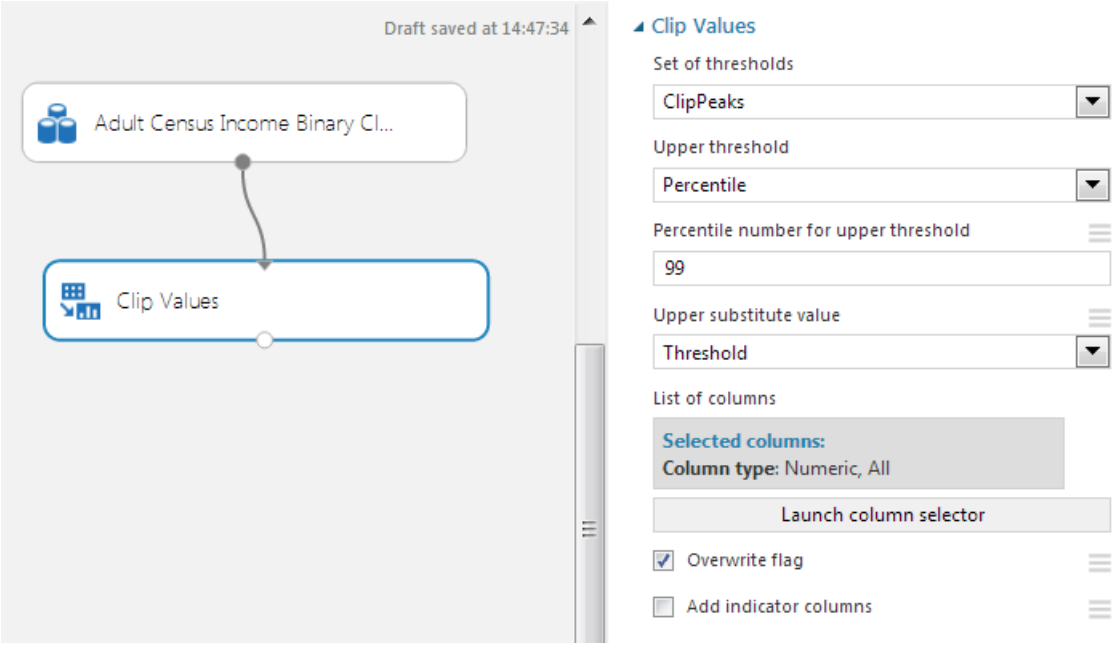

Dataset 1
Dataset 2
Add Columns
Resulted Dataset with concatenated columns

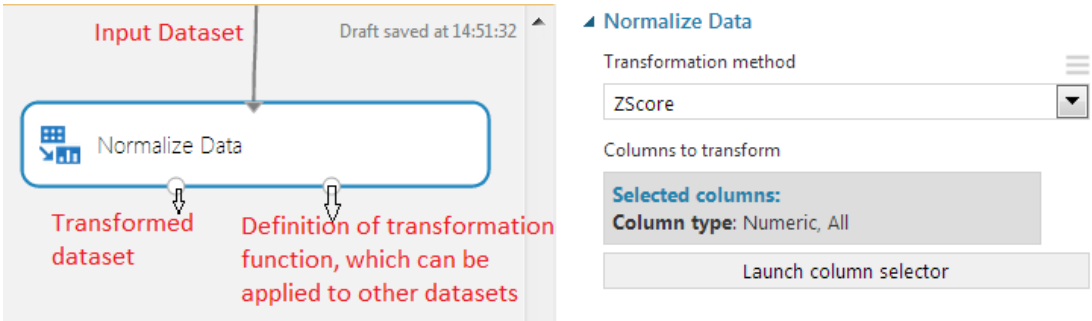

Dataset 1
Dataset 2
Add Rows
Resulted Dataset with concatenated rows

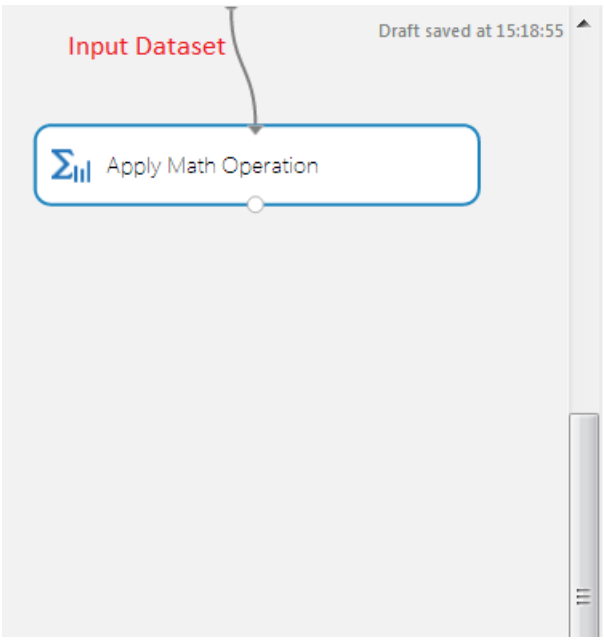

Draft saved at 07:05:40 AM
dataset 1
dataset 2
Join
Resulted dataset


Join
Join key columns for L
Key columns for Left dataset (dataset 1)
Selected columns:
Launch the selector tool to make a selection
Launch column selector
Join key columns for R
Key colums for Right dataset (dataset 2)
Selected columns:
Launch the selector tool to make a selection
Launch column selector
Match case
Join type
Left Outer Join
Keep right key columns in joined table


Draft saved at 13:56:39
Input dataset
Adult Census Income Binary Cl...
Split
80% of Input dataset
20% of Input dataset
Split
Splitting mode
Split Rows
Fraction of rows in the first output dataset
0.8
Randomized split
Random seed
0
Stratified split
False


Draft saved at 14:47:34
Adult Census Income Binary Cl...
Clip Values
Clip Values
Set of thresholds
ClipPeaks
Upper threshold
Percentile
Percentile number for upper threshold
99
Upper substitute value
Threshold
List of columns
Selected columns:
Column type: Numeric, All
Launch column selector
Overwrite flag
Add indicator columns


Input Dataset
Draft saved at 14:51:32
Normalize Data
Transformed dataset
Definition of transformation function, which can be applied to other datasets
Normalize Data
Transformation method
ZScore
Columns to transform
Selected columns:
Column type: Numeric, All
Launch column selector



Apply Math Operation

Category: Basic

Basic math function: Pow

Second argument type: Constant

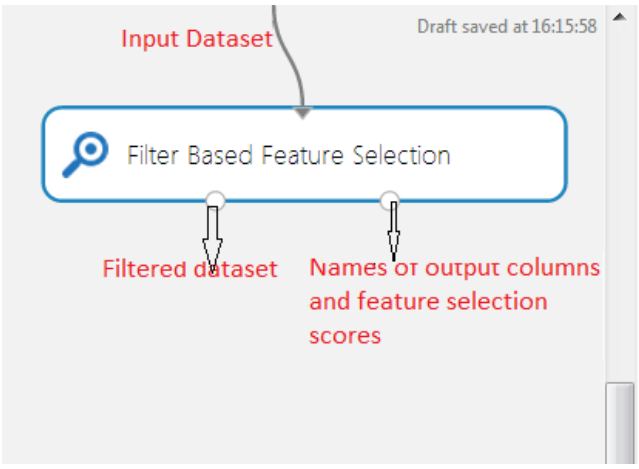
Constant second argument: 2

Column set

Selected columns:
Column type: Numeric, All

Launch column selector

Output mode: ResultOnly



Filter Based Feature Selection

Feature scoring method: Pearson Correlation

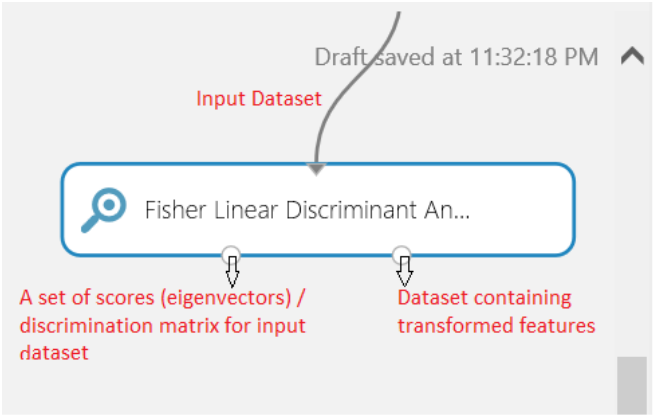
☑ Operate on feature columns only

Target column

Selected columns:
Column names: income

Launch column selector

Number of desired features: 3



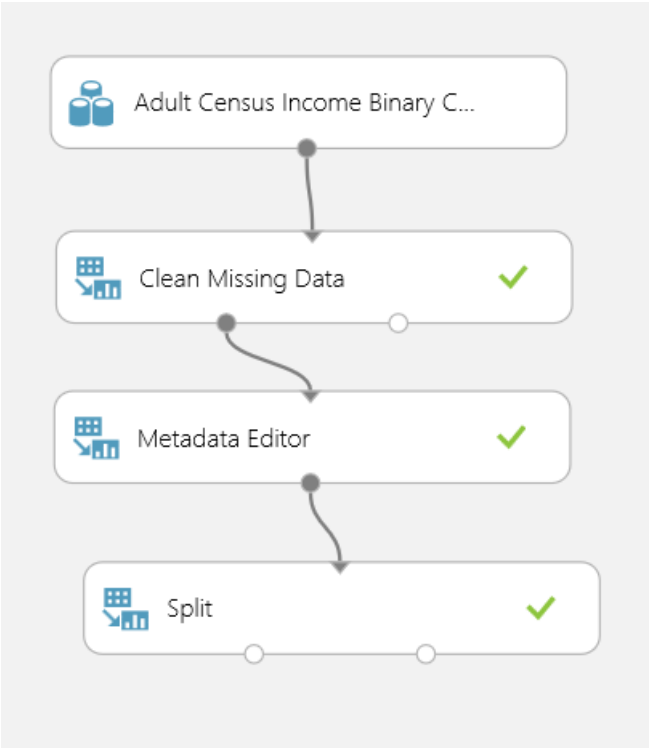
Fisher Linear Discriminant Analysis

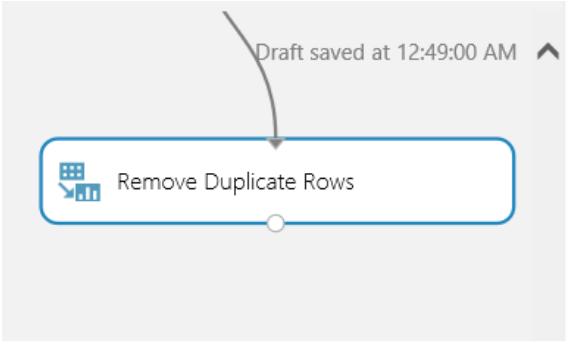
Class labels column

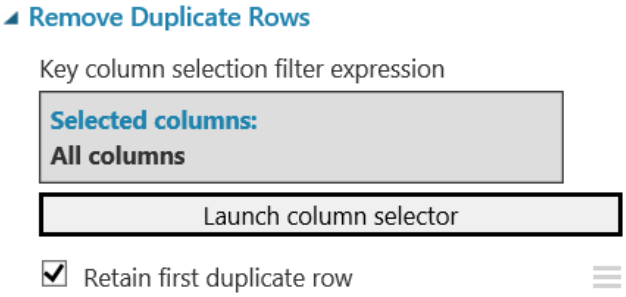
Selected columns:
Column names: income

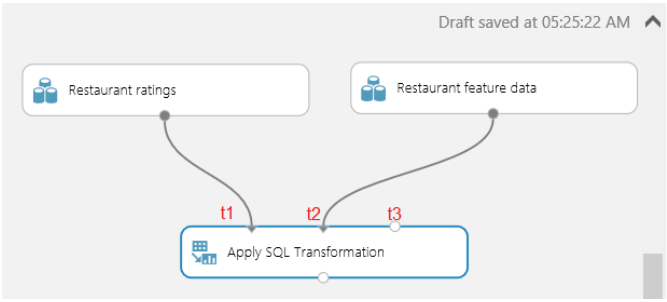
Launch column selector

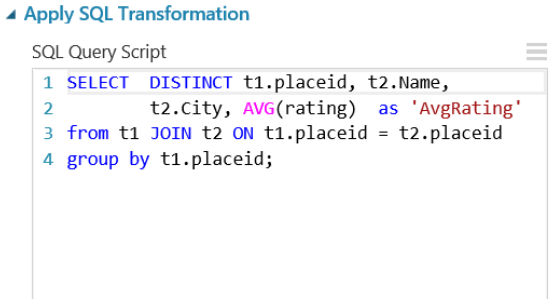
Number of feature extractors: 3


Adult Census Income Binary C...
Clean Missing Data
Metadata Editor
Split

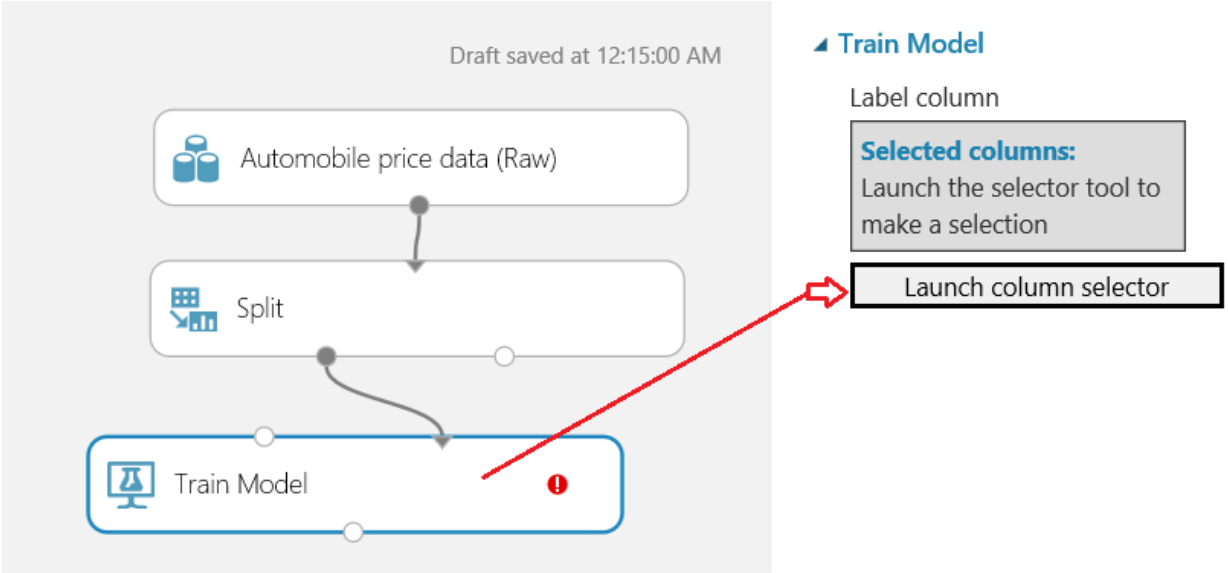

Draft saved at 12:49:00 AM
Remove Duplicate Rows

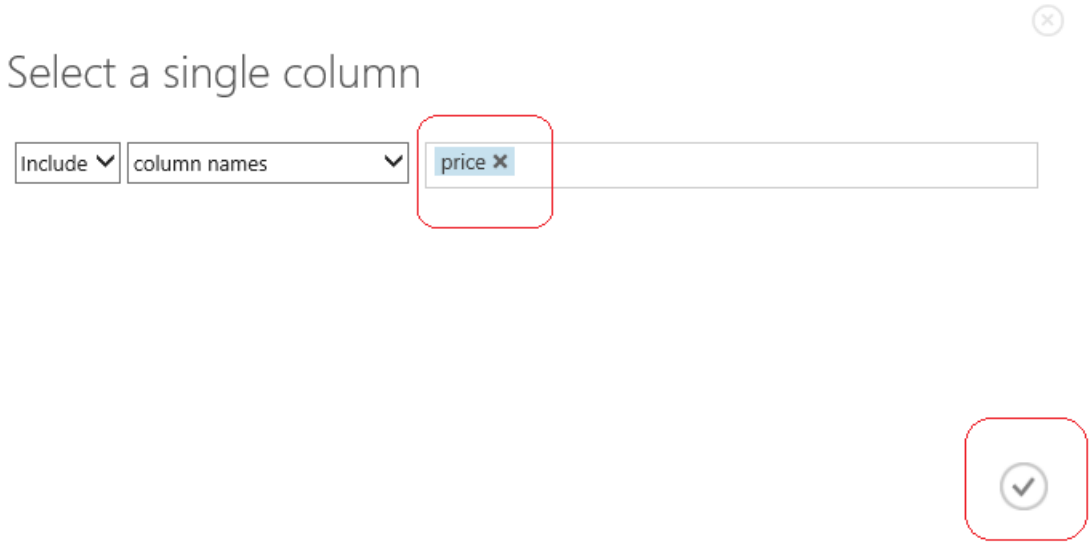

Remove Duplicate Rows
Key column selection filter expression
Selected columns:
All columns
Launch column selector
Retain first duplicate row

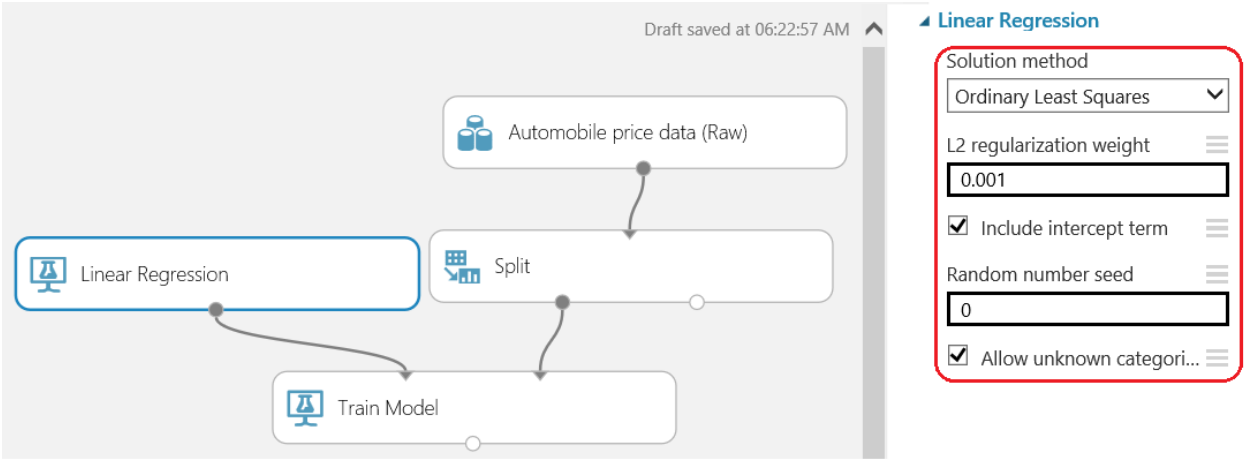

Draft saved at 05:25:22 AM
Restaurant ratings
Restaurant feature data
t1
t2
t3
Apply SQL Transformation

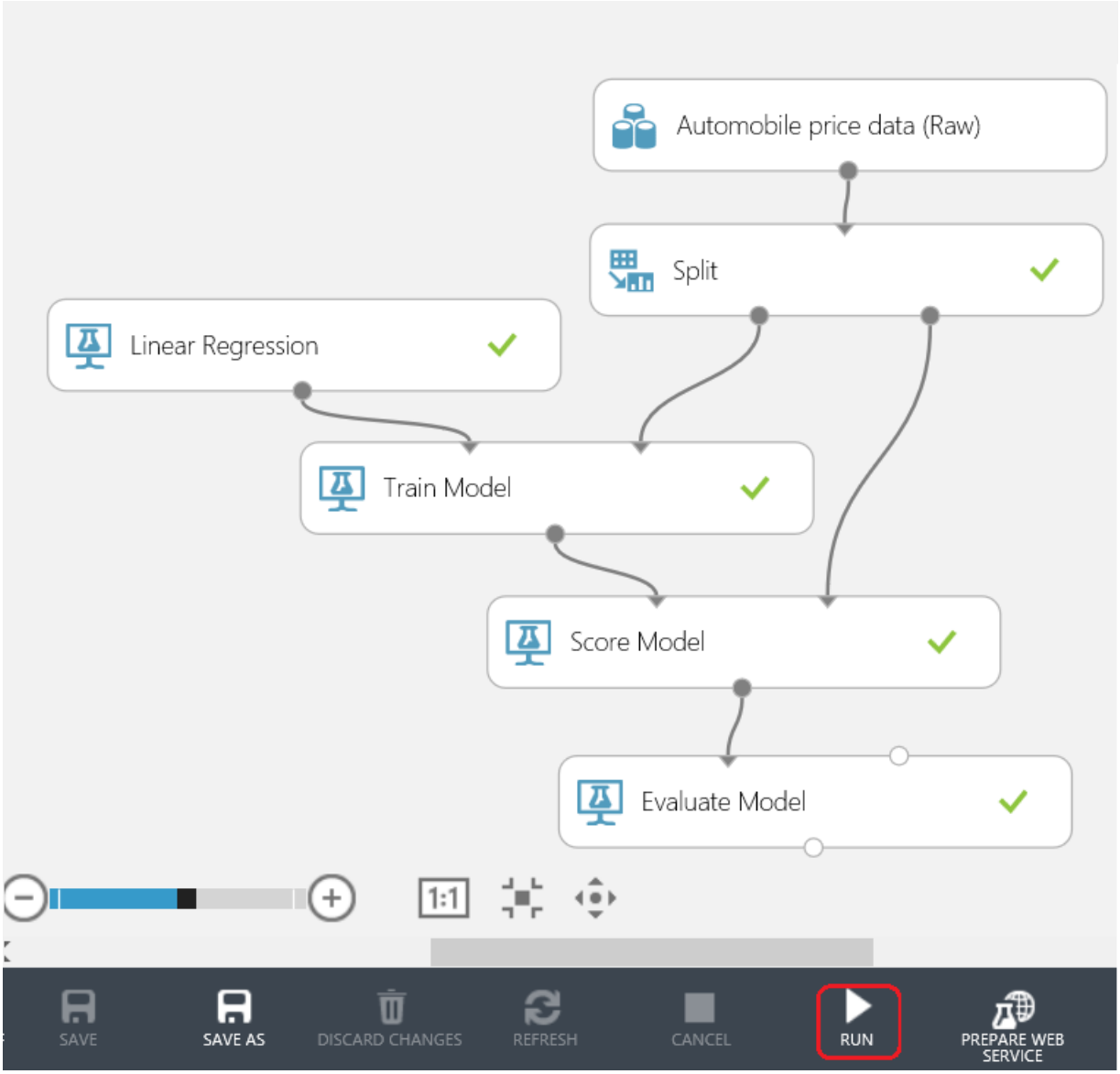

Apply SQL Transformation
SQL Query Script
1 SELECT DISTINCT t1.placeid, t2.Name,
2 t2.city, AVG(rating) as 'AvgRating'
3 from t1 JOIN t2 ON t1.placeid = t2.placeid
4 group by t1.placeid;

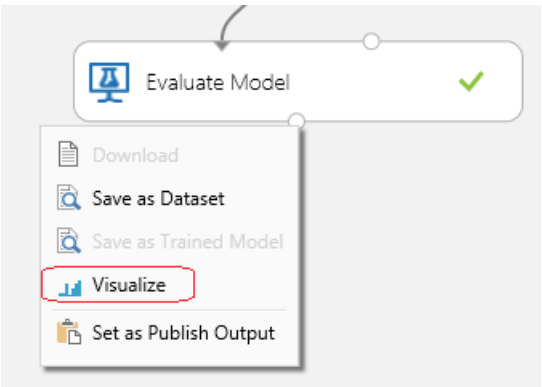
Chapter 6


Draft saved at 12:15:00 AM
Automobile price data (Raw)
Split
Train Model
Train Model
Label column
Selected columns:
Launch the selector tool to make a selection
Launch column selector


Select a single column
Include
column names
price


Draft saved at 06:22:57 AM
Automobile price data (Raw)
Linear Regression
Split
Train Model
Linear Regression
Solution method
Ordinary Least Squares
L2 regularization weight
0.001
Include intercept term
Random number seed
0
Allow unknown categori...


Automobile price data (Raw)
Split
Linear Regression
Train Model
Score Model
Evaluate Model
1:1
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PREPARE WEB SERVICE

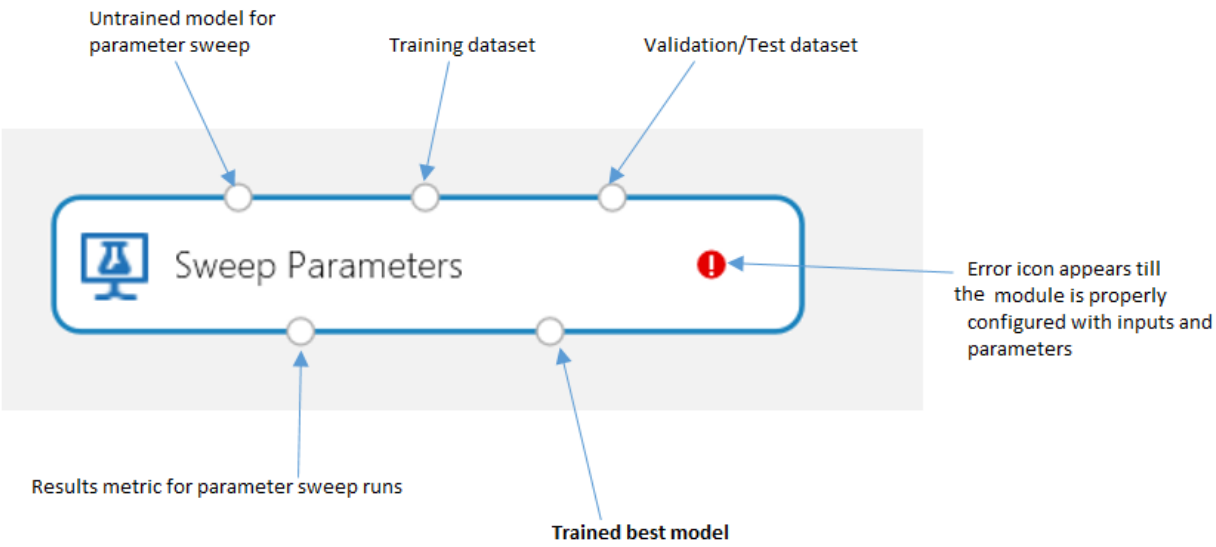

Evaluate Model
Download
Save as Dataset
Save as Trained Model
Visualize
Set as Publish Output

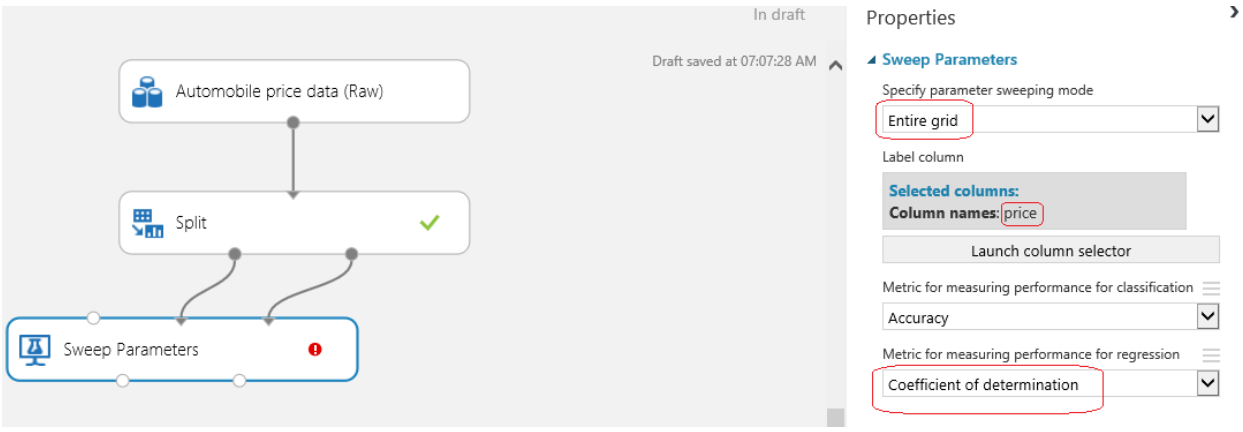
Demo-LinearRegression › Evaluate Model › Evaluation results

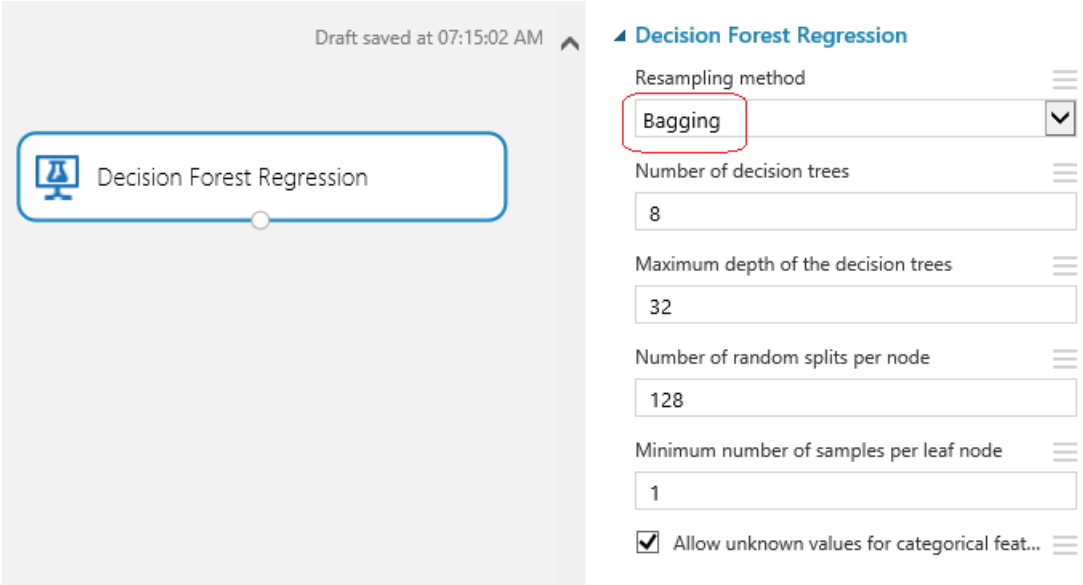
rows	columns
1	5

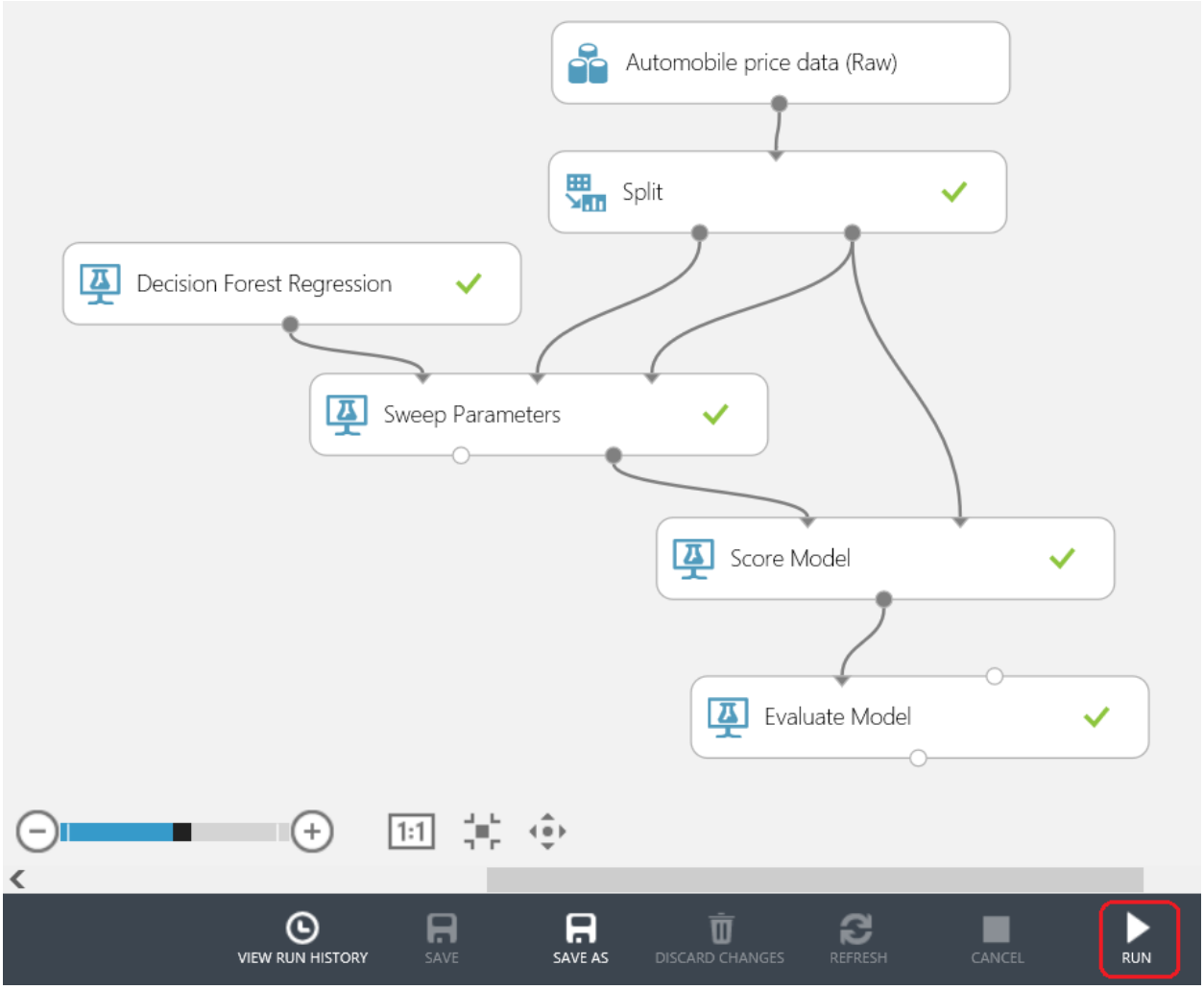
view as

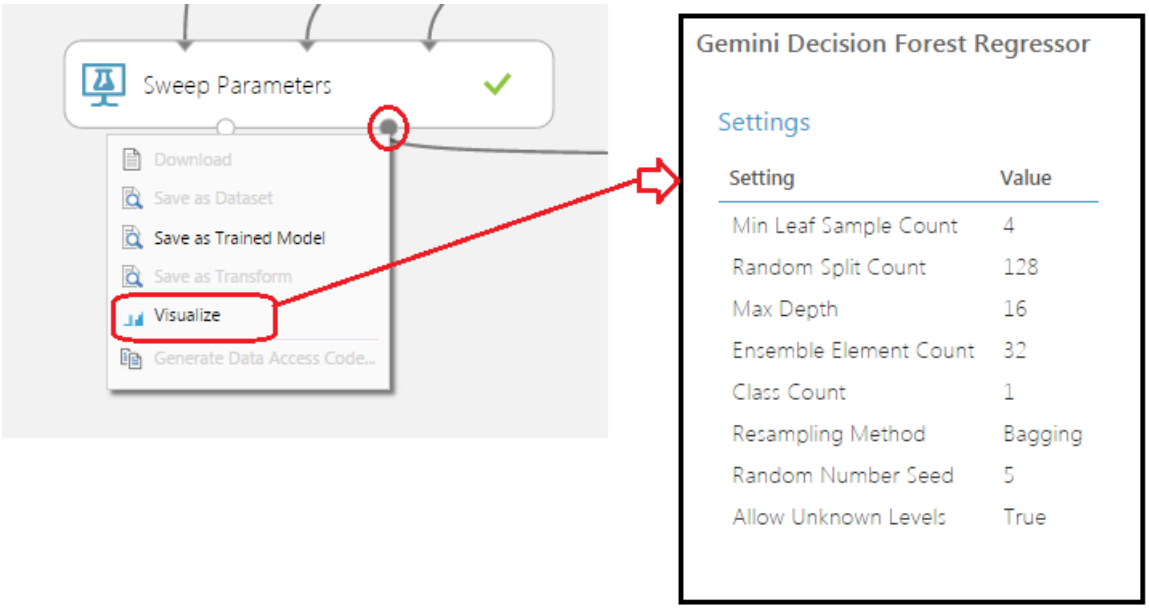
Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
1700.399749	2484.018779	0.189728	0.058297	0.941703

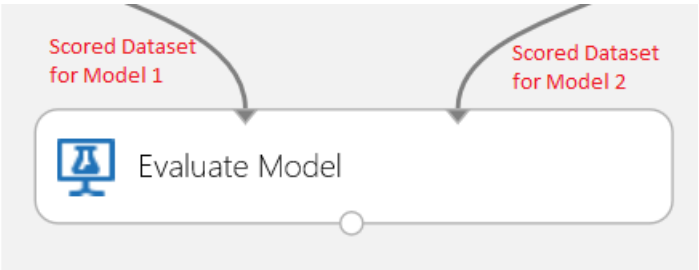


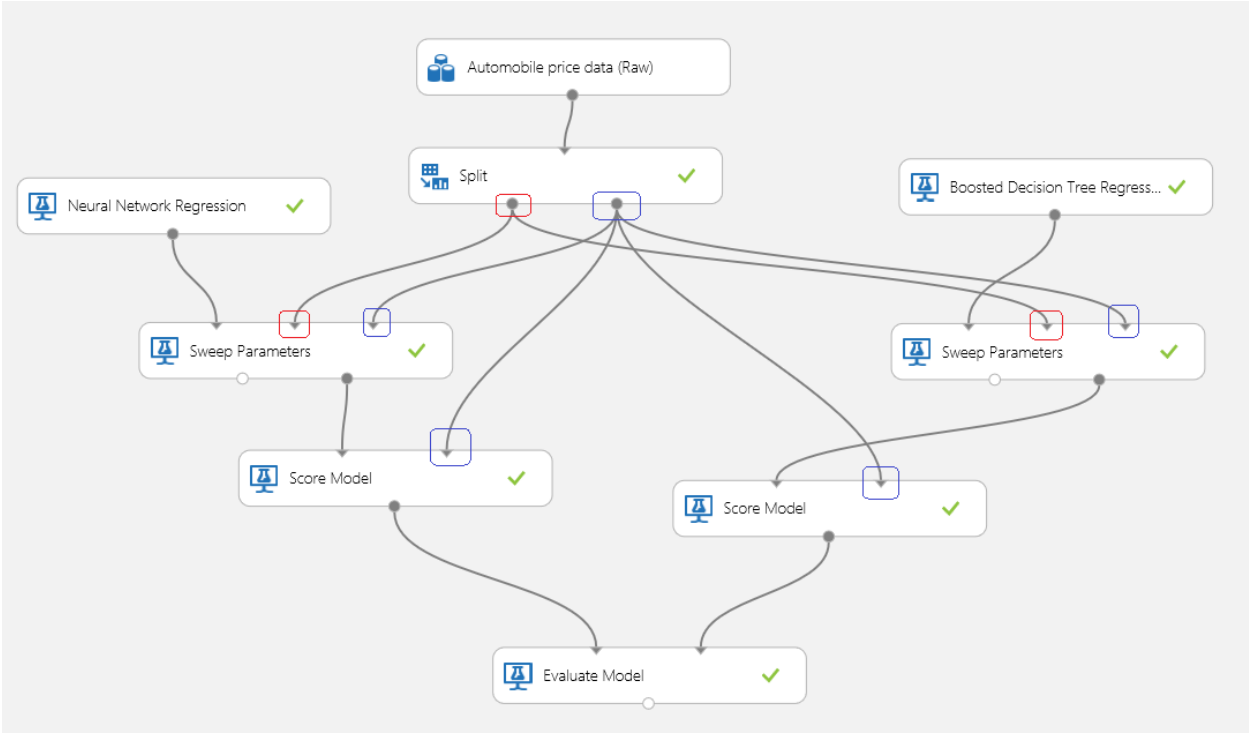



Draft saved at 07:15:02 AM
Decision Forest Regression
Decision Forest Regression
Resampling method
Bagging
Number of decision trees
8
Maximum depth of the decision trees
32
Number of random splits per node
128
Minimum number of samples per leaf node
1
Allow unknown values for categorical feat...


Automobile price data (Raw)
Split
Decision Forest Regression
Sweep Parameters
Score Model
Evaluate Model
1:1
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN

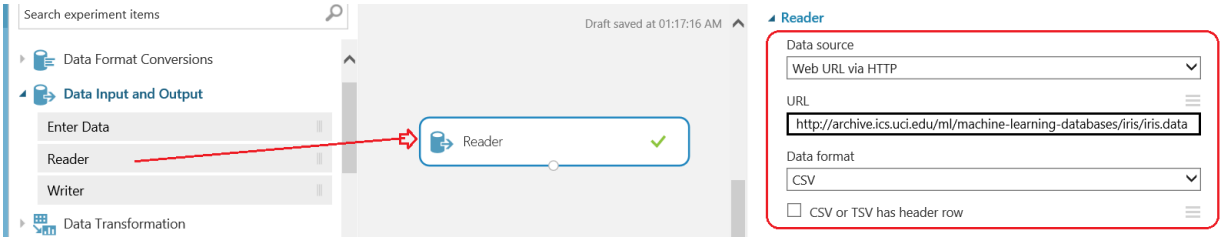

Sweep Parameters
Download
Save as Dataset
Save as Trained Model
Save as Transform
Visualize
Generate Data Access Code...
Gemini Decision Forest Regressor
Settings
Setting
Value
Min Leaf Sample Count
4
Random Split Count
128
Max Depth
16
Ensemble Element Count
32
Class Count
1
Resampling Method
Bagging
Random Number Seed
5
Allow Unknown Levels
True

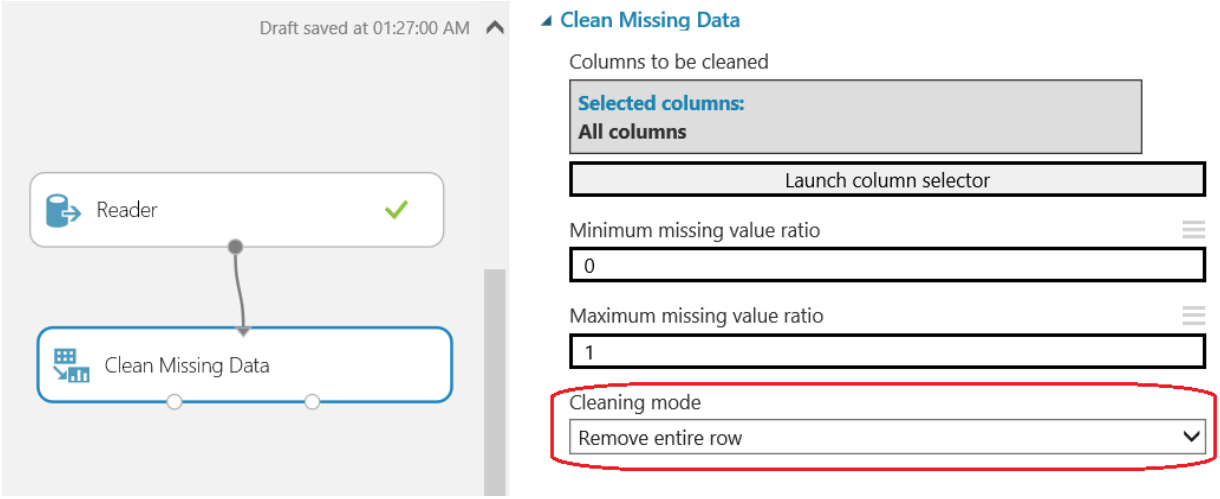

Scored Dataset
for Model 1
Scored Dataset
for Model 2
Evaluate Model

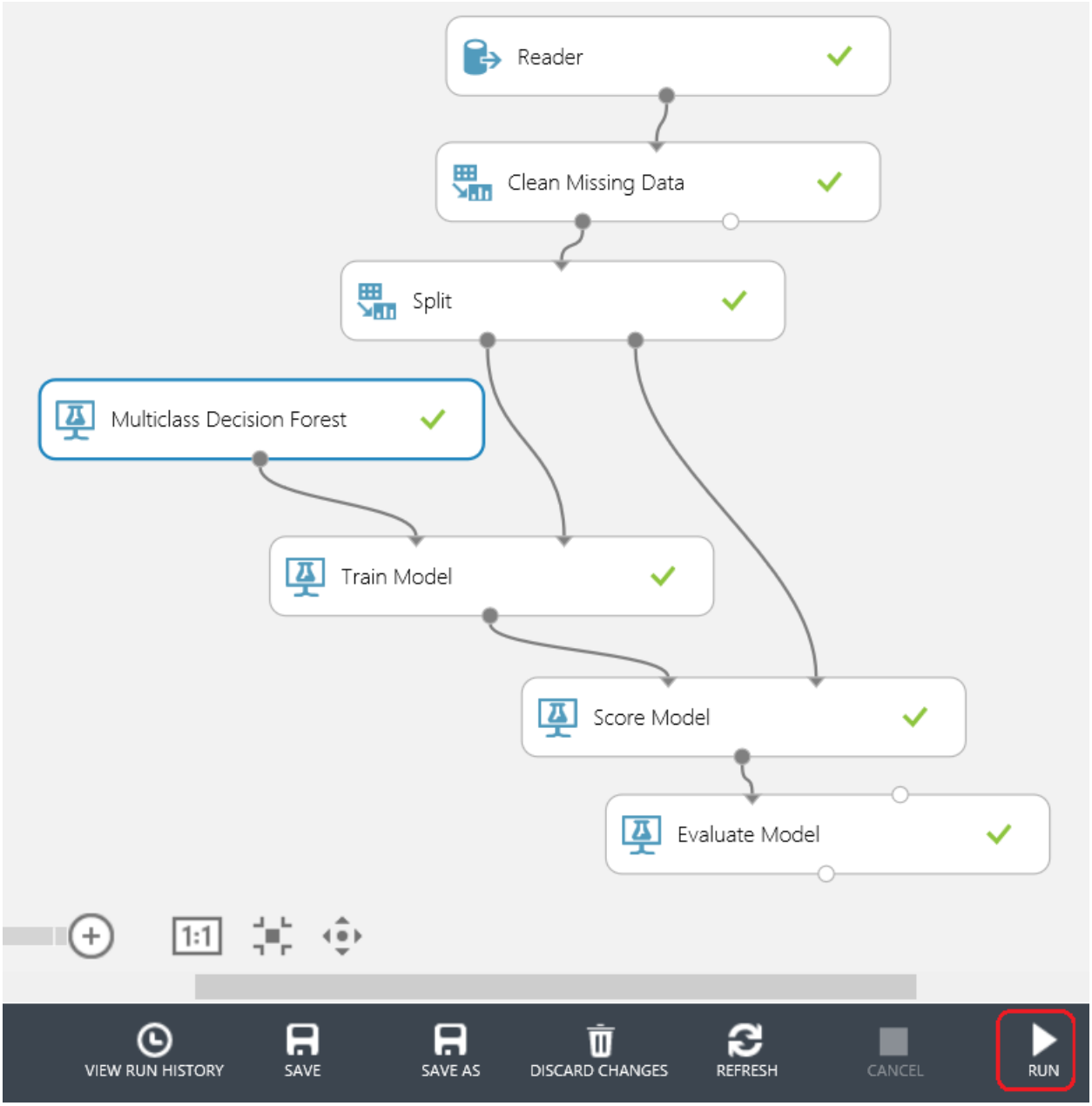


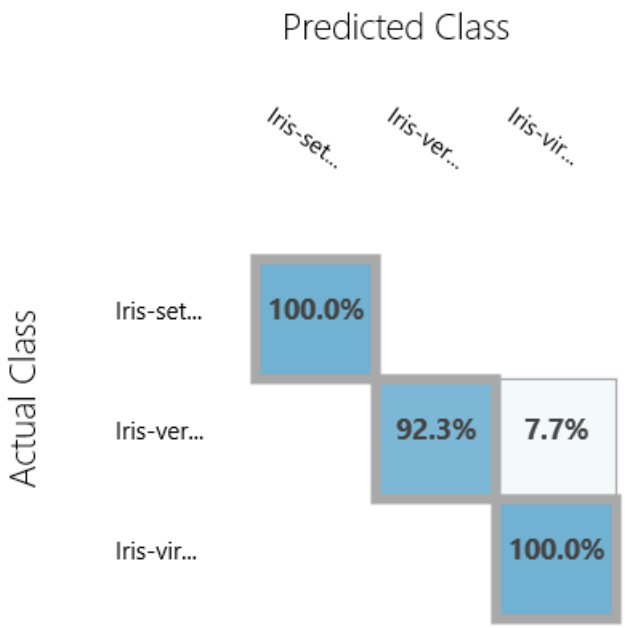
Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
5071.800876	7664.589306	0.565904	0.555025	0.444975
1340.164661	1958.367584	0.149534	0.036235	0.963765

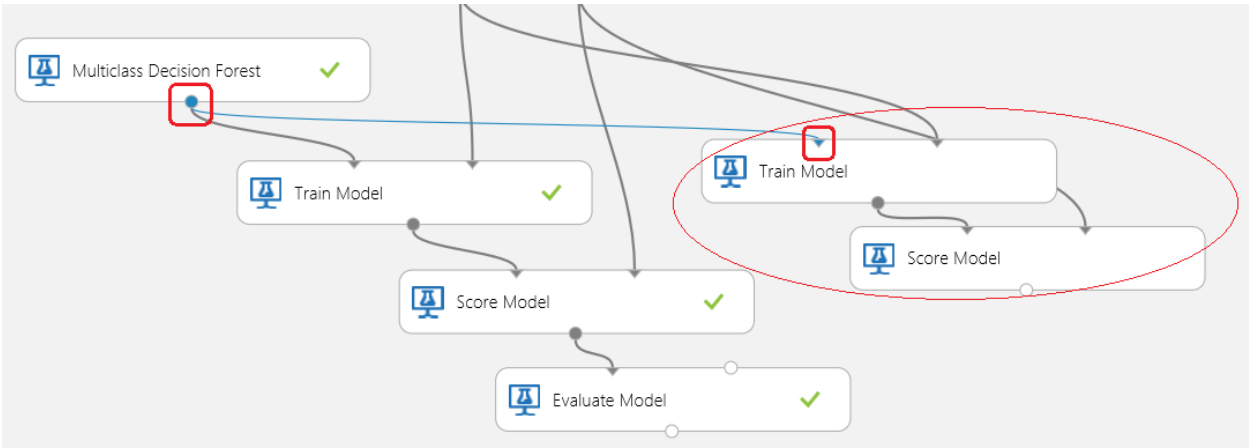
Chapter 7

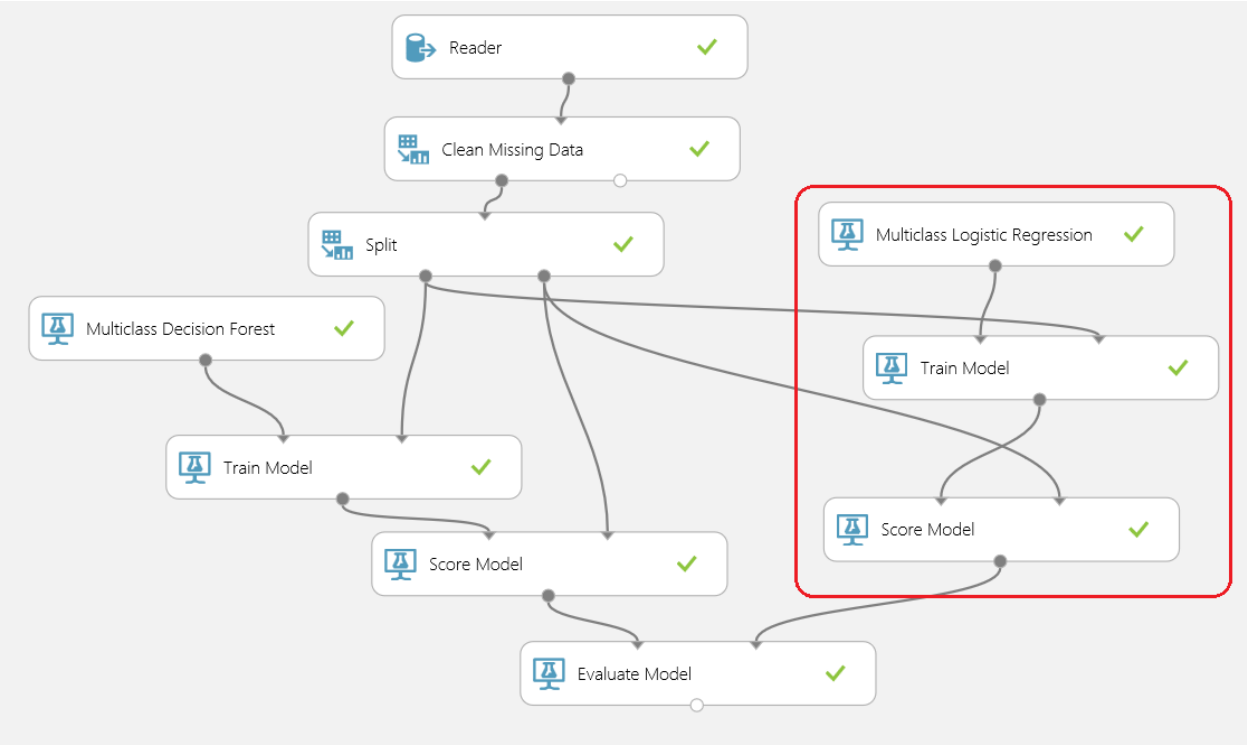

Search experiment items
Data Format Conversions
Data Input and Output
Enter Data
Reader
Writer
Data Transformation
Draft saved at 01:17:16 AM
Reader
Reader
Data source
Web URL via HTTP
URL
http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
Data format
CSV
CSV or TSV has header row

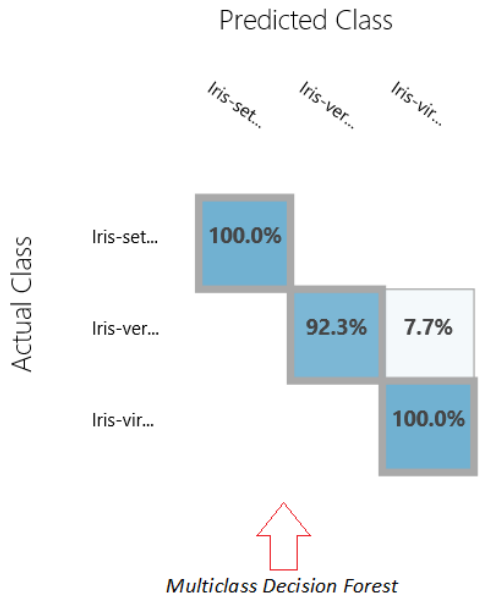

Draft saved at 01:27:00 AM
Reader
Clean Missing Data
Clean Missing Data
Columns to be cleaned
Selected columns:
All columns
Launch column selector
Minimum missing value ratio
0
Maximum missing value ratio
1
Cleaning mode
Remove entire row


Reader
Clean Missing Data
Split
Multiclass Decision Forest
Train Model
Score Model
Evaluate Model
1:1
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN


Predicted Class
Iris-set...
Iris-ver...
Iris-vir...
Actual Class
Iris-set...
Iris-ver...
Iris-vir...
100.0%
92.3%
7.7%
100.0%

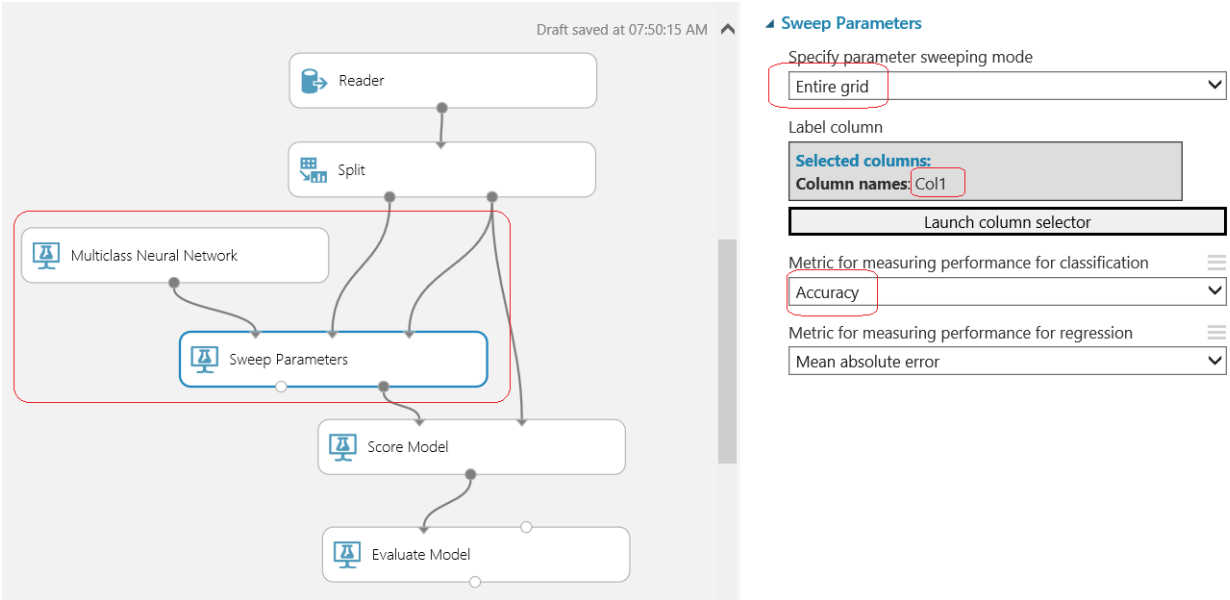

Multiclass Decision Forest
Train Model
Train Model
Score Model
Score Model
Evaluate Model

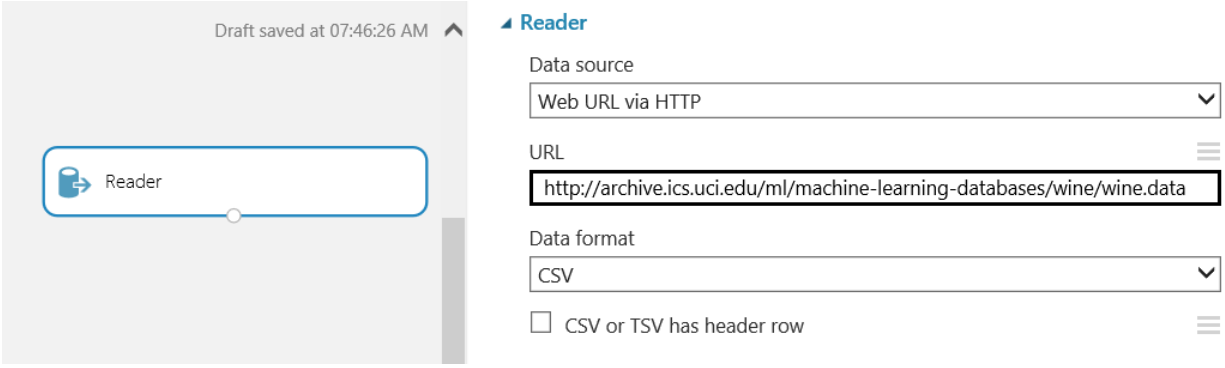


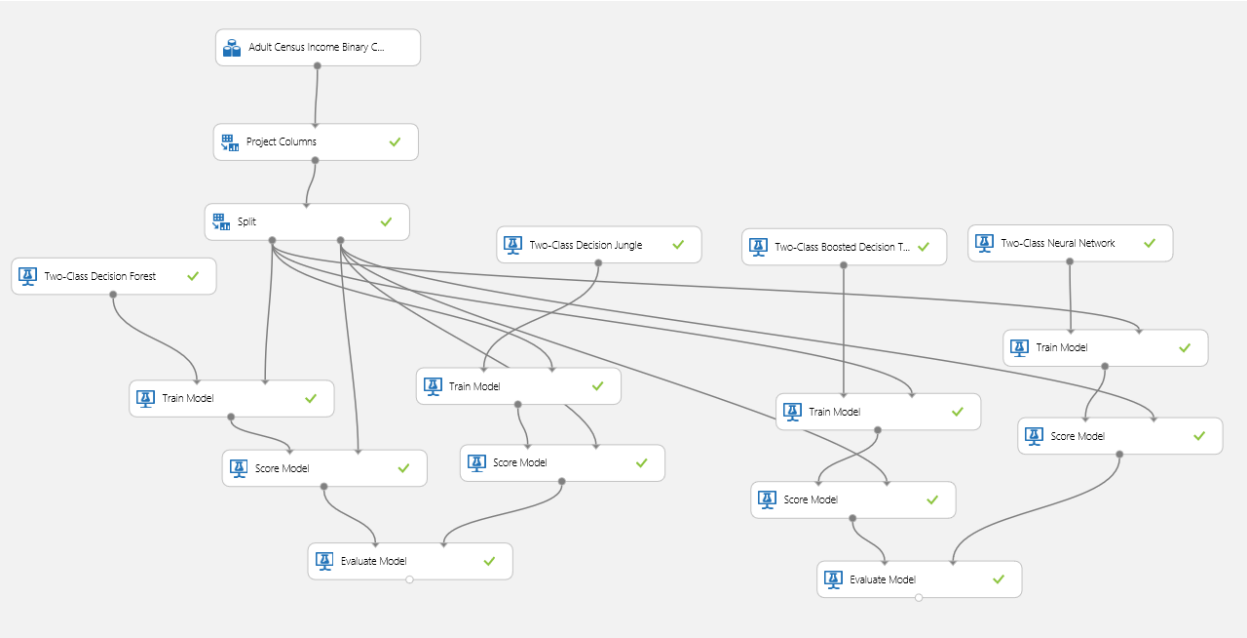


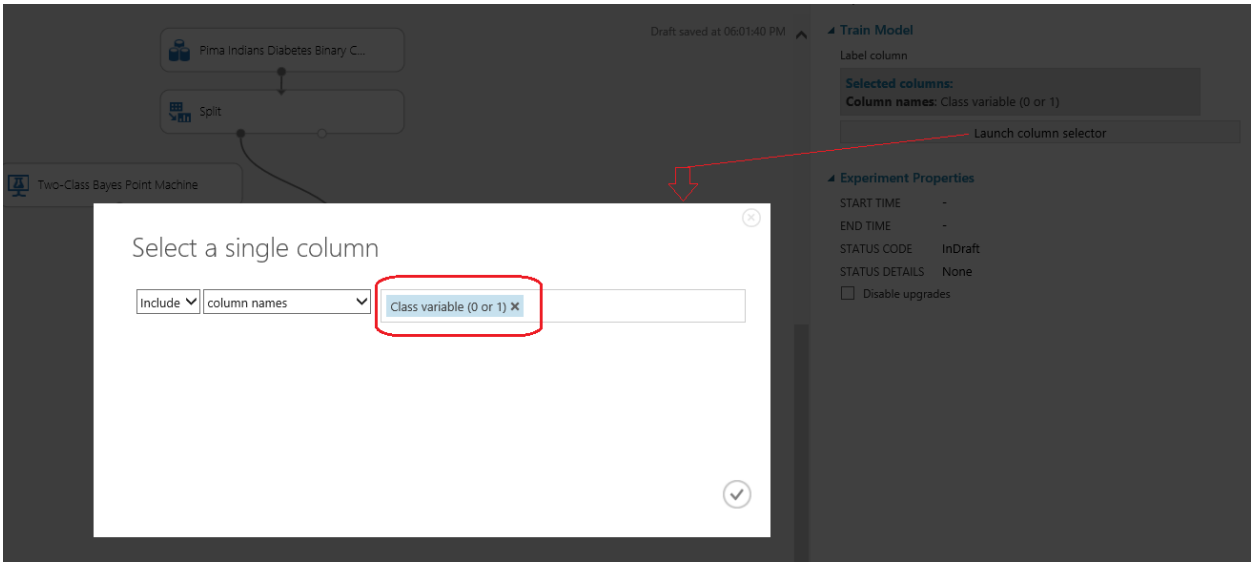
Multiclass Decision Forest

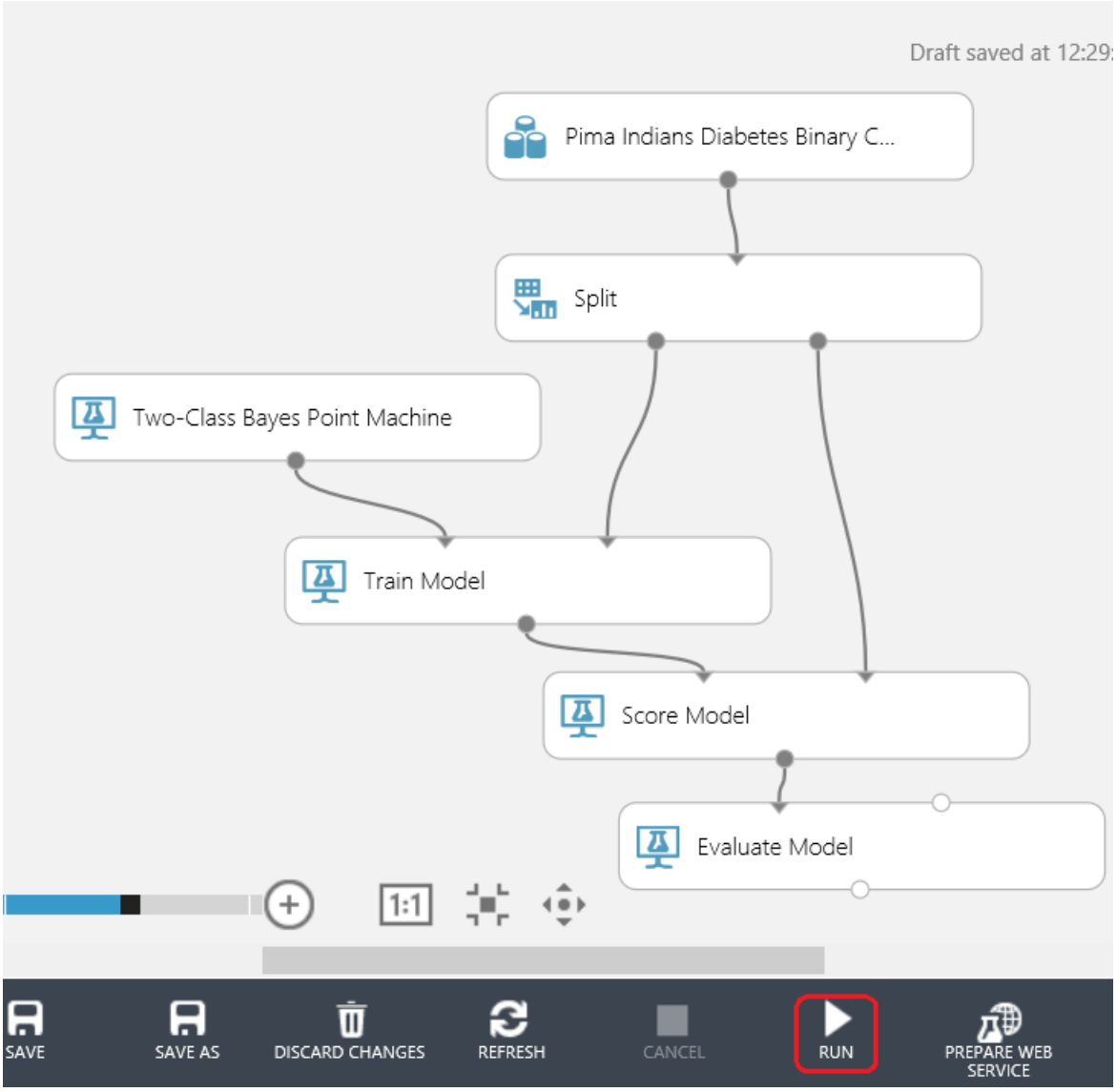
Multiclass Logistic Regression


Draft saved at 07:50:15 AM
Reader
Split
Multiclass Neural Network
Sweep Parameters
Score Model
Evaluate Model
Sweep Parameters
Specify parameter sweeping mode
Entire grid
Label column
Selected columns:
Column names: Col1
Launch column selector
Metric for measuring performance for classification
Accuracy
Metric for measuring performance for regression
Mean absolute error


Draft saved at 07:46:26 AM
Reader
Reader
Data source
Web URL via HTTP
URL
http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data
Data format
CSV
CSV or TSV has header row

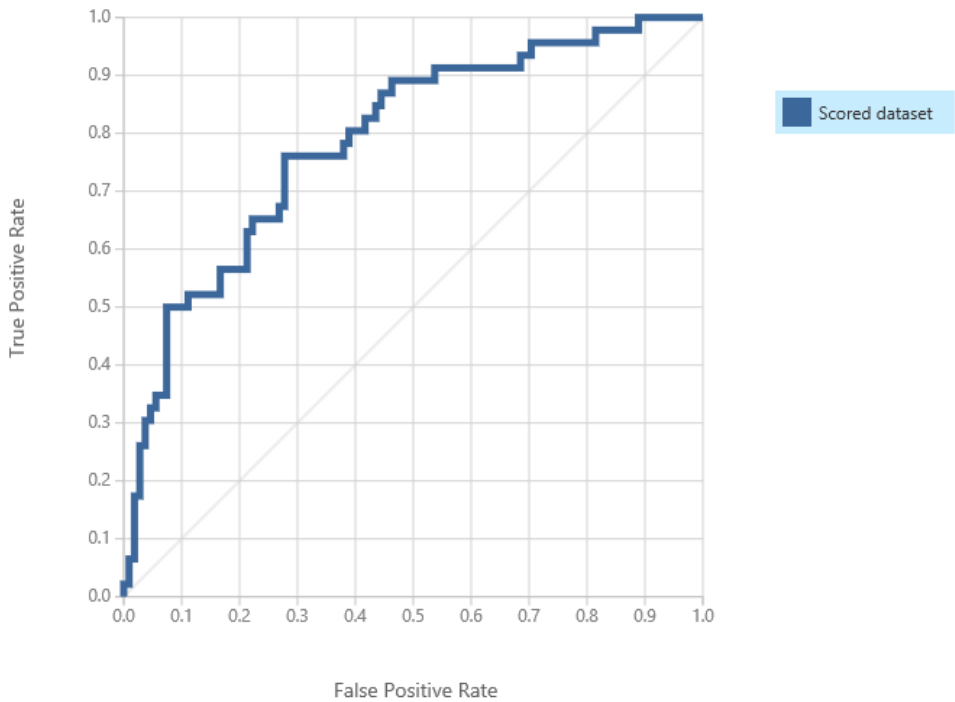

Adult Census Income Binary C...
Project Columns
Split
Two-Class Decision Forest
Two-Class Decision Jungle
Two-Class Boosted Decision T...
Two-Class Neural Network
Train Model
Train Model
Train Model
Train Model
Score Model
Score Model
Score Model
Score Model
Evaluate Model
Evaluate Model


Pima Indians Diabetes Binary C...
Split
Two-Class Bayes Point Machine
Draft saved at 06:01:40 PM
Train Model
Label column
Selected columns:
Column names: Class variable (0 or 1)
Launch column selector
Experiment Properties
START TIME -
END TIME -
STATUS CODE InDraft
STATUS DETAILS None
Disable upgrades
Select a single column
Include
column names
Class variable (0 or 1)

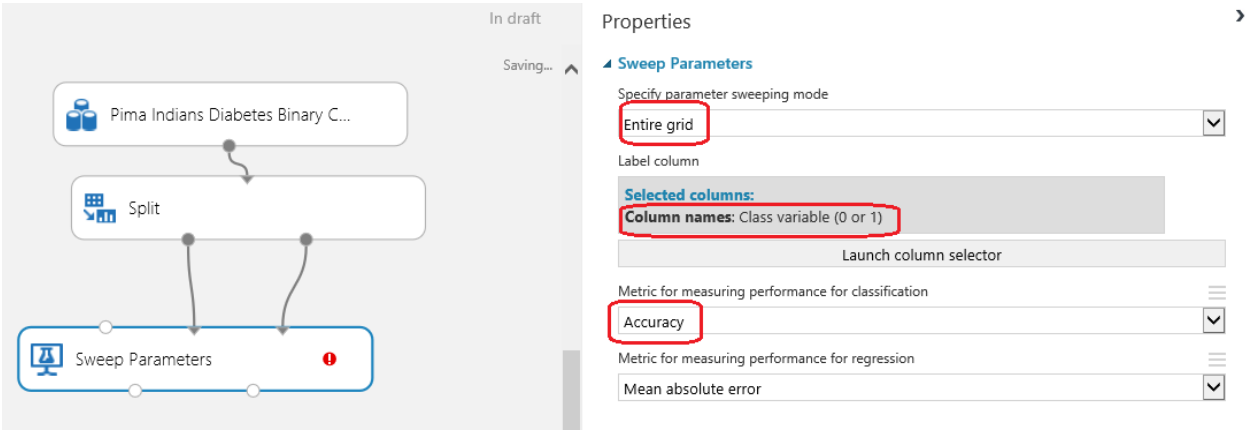

Draft saved at 12:29
Pima Indians Diabetes Binary C...
Split
Two-Class Bayes Point Machine
Train Model
Score Model
Evaluate Model
1:1
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PREPARE WEB SERVICE

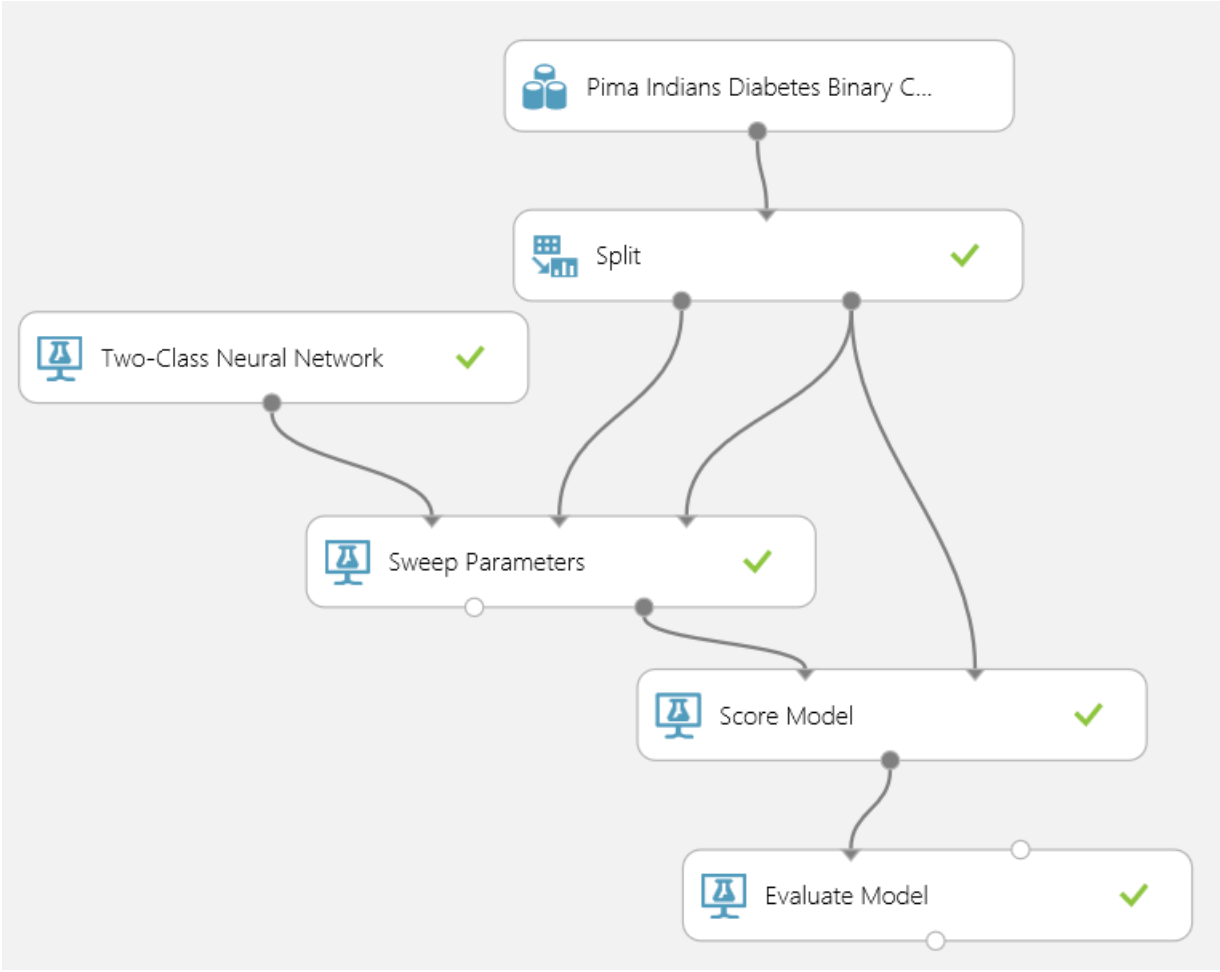
Classification - Pima Indians Diabetes Dataset › Evaluate Model › Evaluation results

ROC PRECISION/RECALL LIFT



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
23	23	0.779	0.676	0.51	0.788
False Positive	**True Negative**	**Recall**	**F1 Score**		
11	97	0.500	0.575		





Classification - Adult Census Income › Adult Census Income Binary Classification dataset › dataset

rows 32561 columns 15

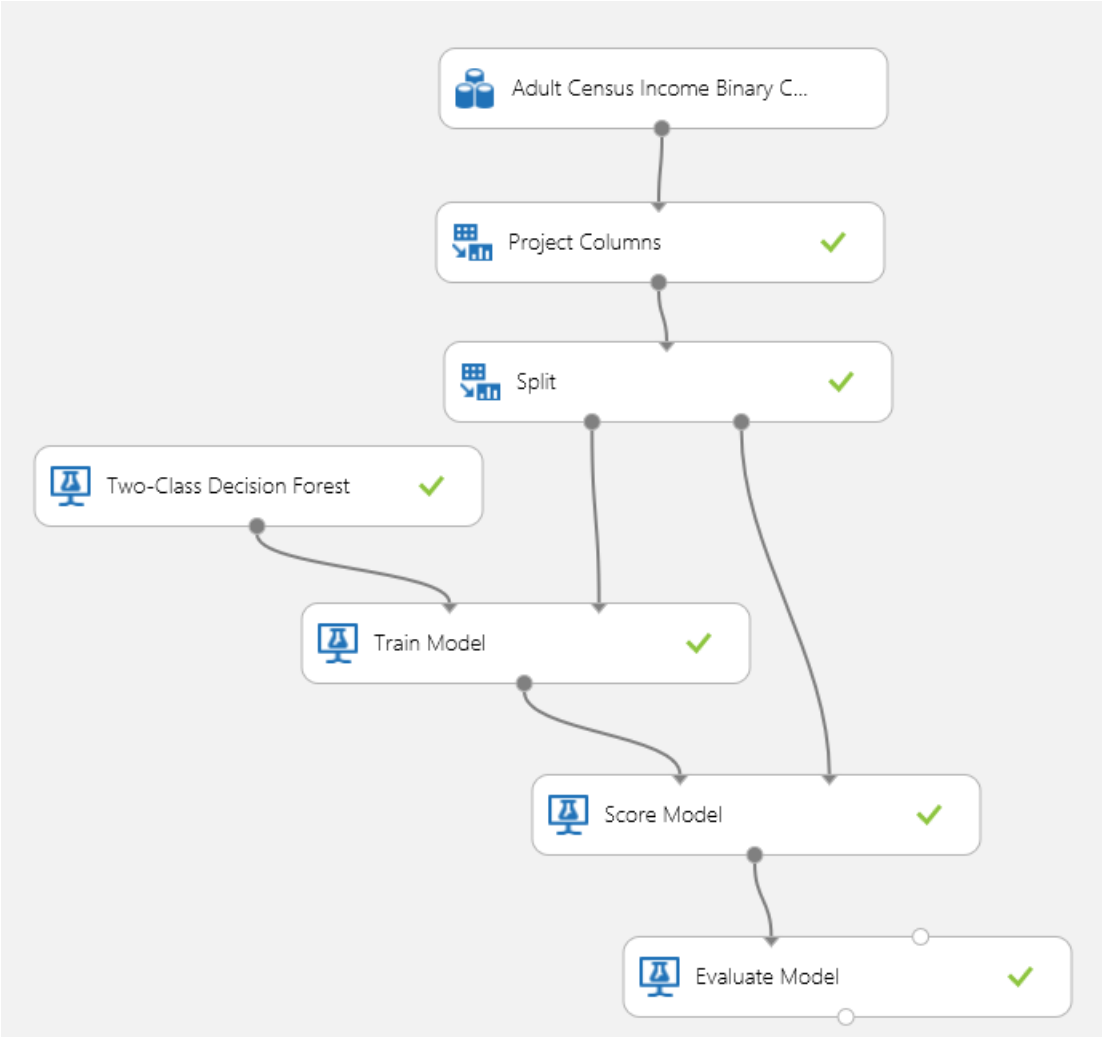
	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	ca ga
view as											
	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	21
	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0

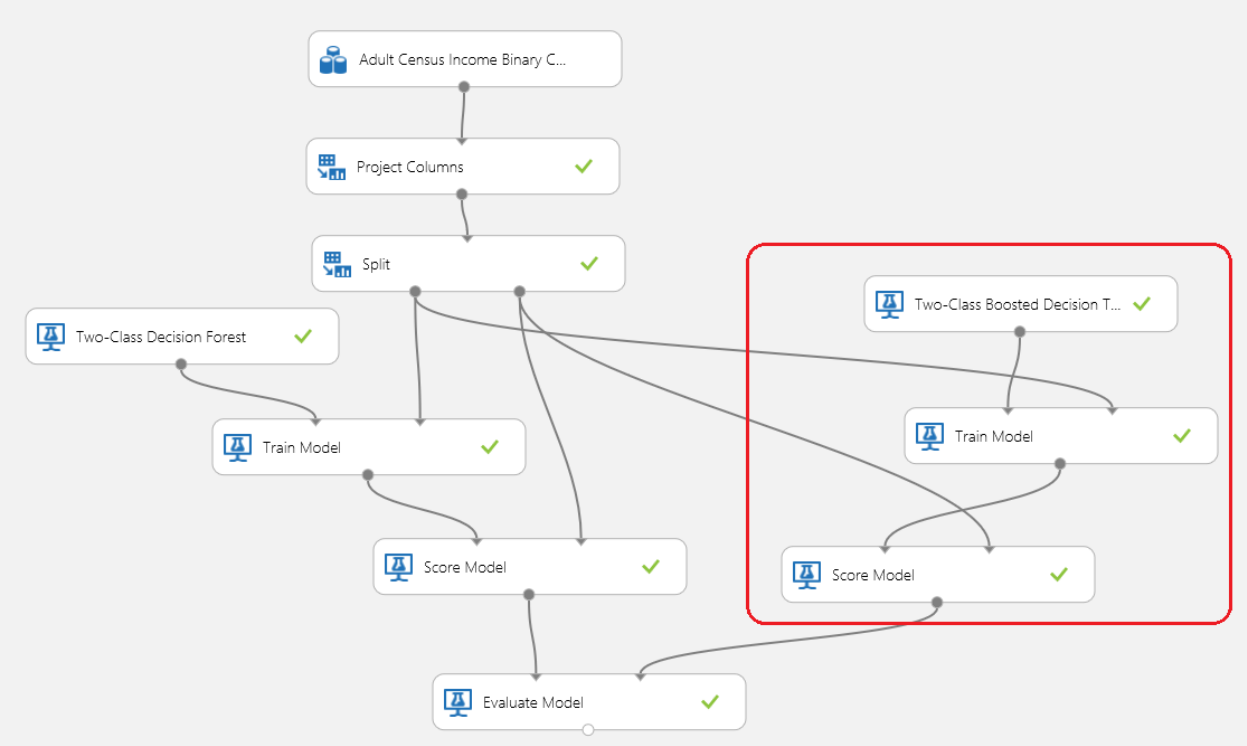
Statistics

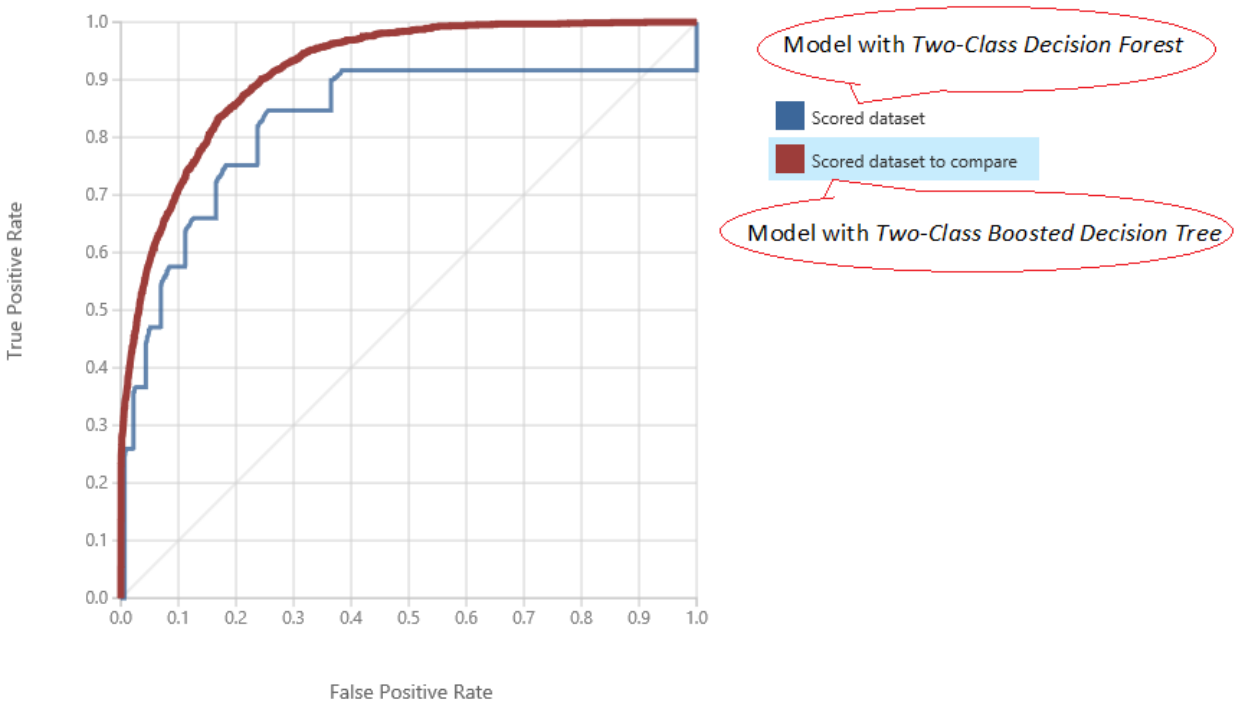
Unique Values	9
Missing Values	1836
Feature Type	String

Visualizations

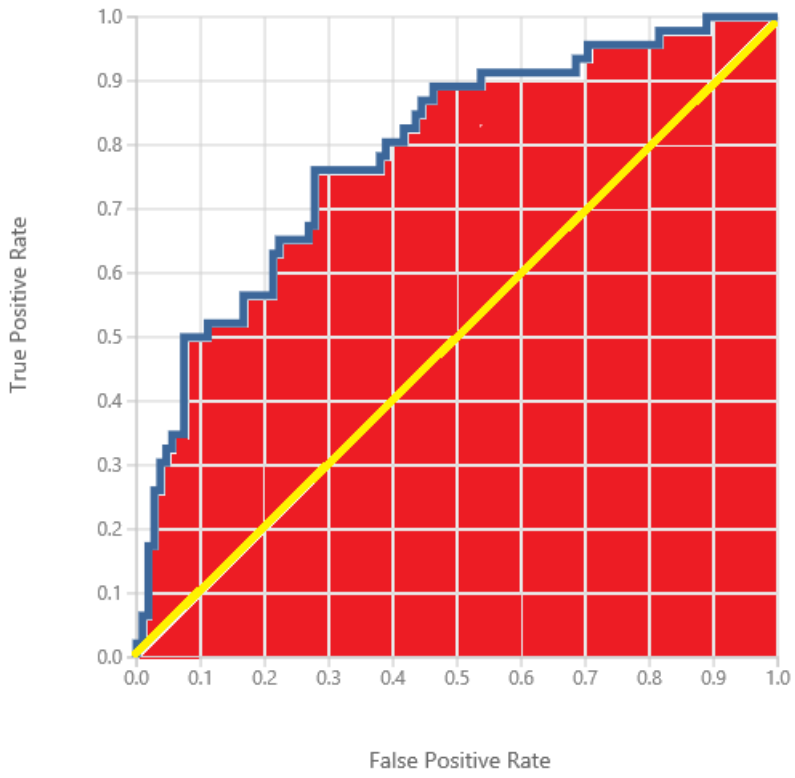

Classification - Adult Census Income (DF)
In draft
Draft saved at 11:56:17 AM
Adult Census Income Binary C...
Project Columns
Properties
Project Columns
Select columns
Selected columns:
All columns
Exclude column names:
workclass,occupation,native-country
Launch column selector
Experiment Properties
STATUS CODE
InDraft
Disable upgrades
Select columns
Allow duplicates and preserve column order in selection
Begin With
All columns
Exclude
column names
workclass
occupation
native-country

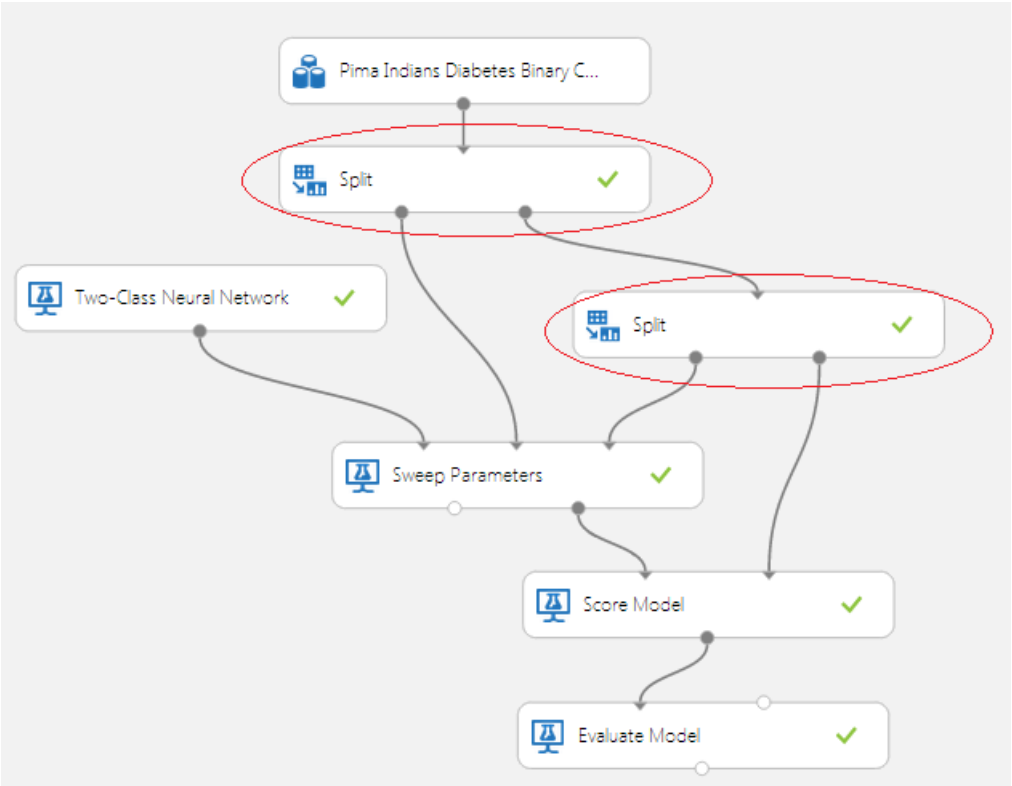

Adult Census Income Binary C...
Project Columns
Split
Two-Class Decision Forest
Train Model
Score Model
Evaluate Model

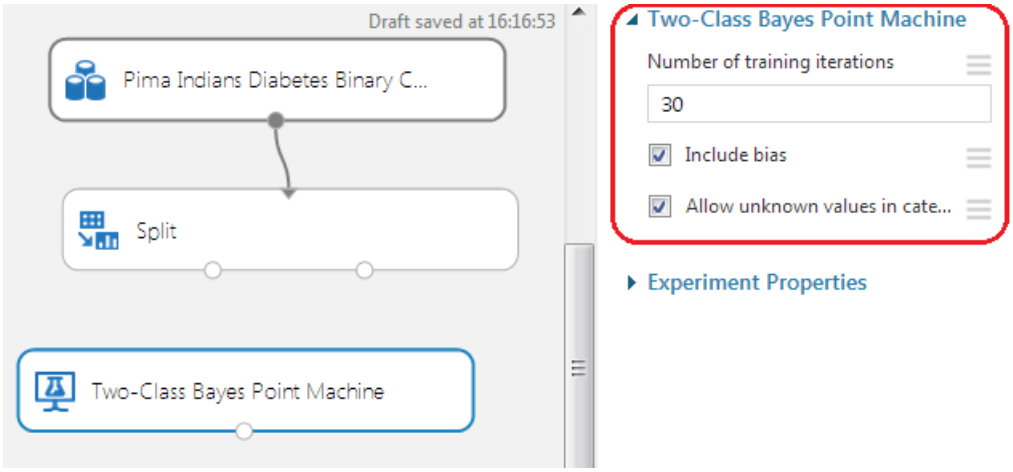

Adult Census Income Binary C...
Project Columns
Split
Two-Class Decision Forest
Two-Class Boosted Decision T...
Train Model
Train Model
Score Model
Score Model
Evaluate Model



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1017	561	0.859	0.741	0.5	0.916
False Positive	**True Negative**	**Recall**	**F1 Score**		
356	4578	0.644	0.689		


True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

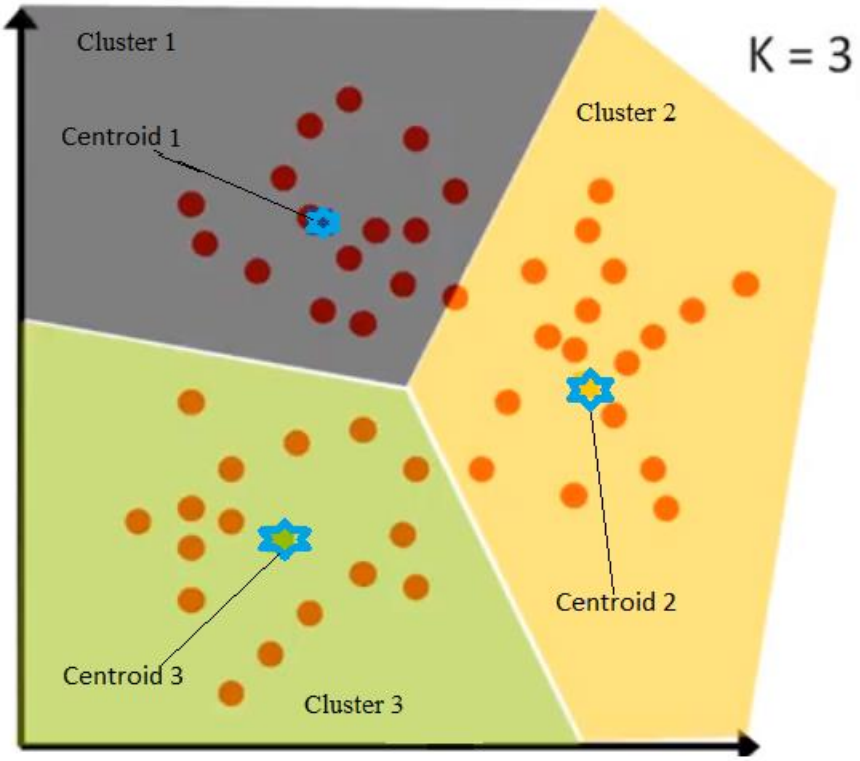

Pima Indians Diabetes Binary C...
Split
Two-Class Neural Network
Split
Sweep Parameters
Score Model
Evaluate Model



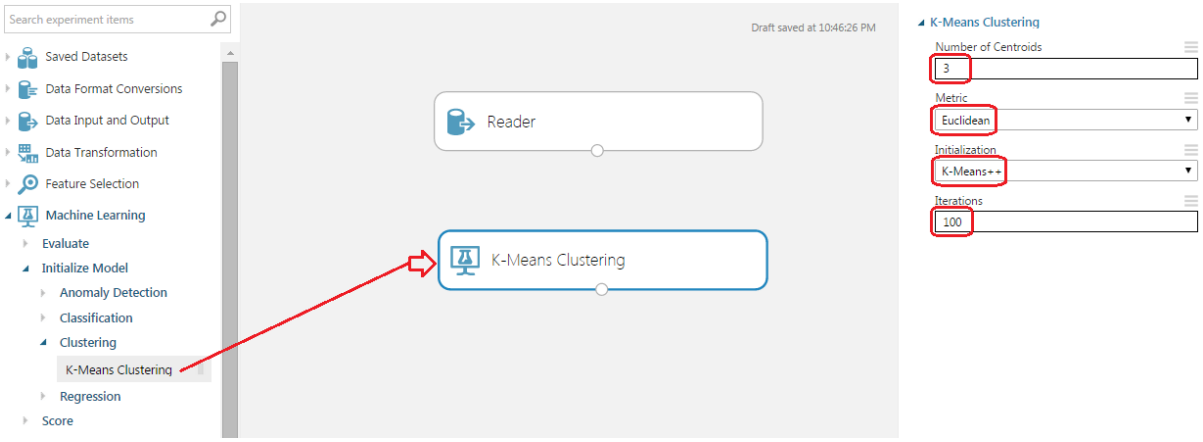
Actual \ Predicted	P	N
T	1	1
F	2	1

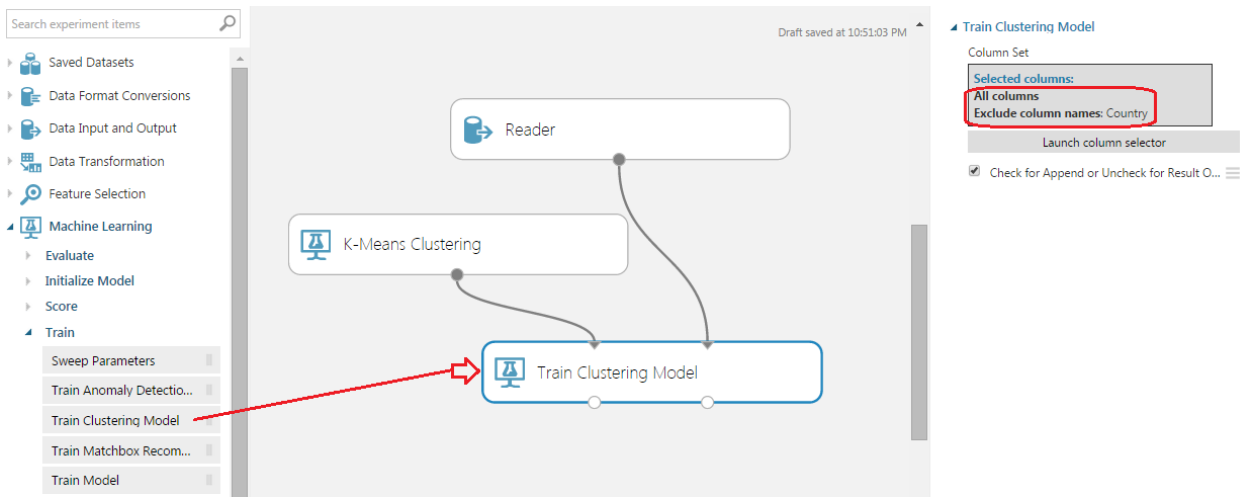
Actual \ Predicted	Low	Medium	High
Low	**0 (0%)**	1 (100%)	
Medium		**2 (66.6%)**	1(33.3%)
High			**1 (100%)**

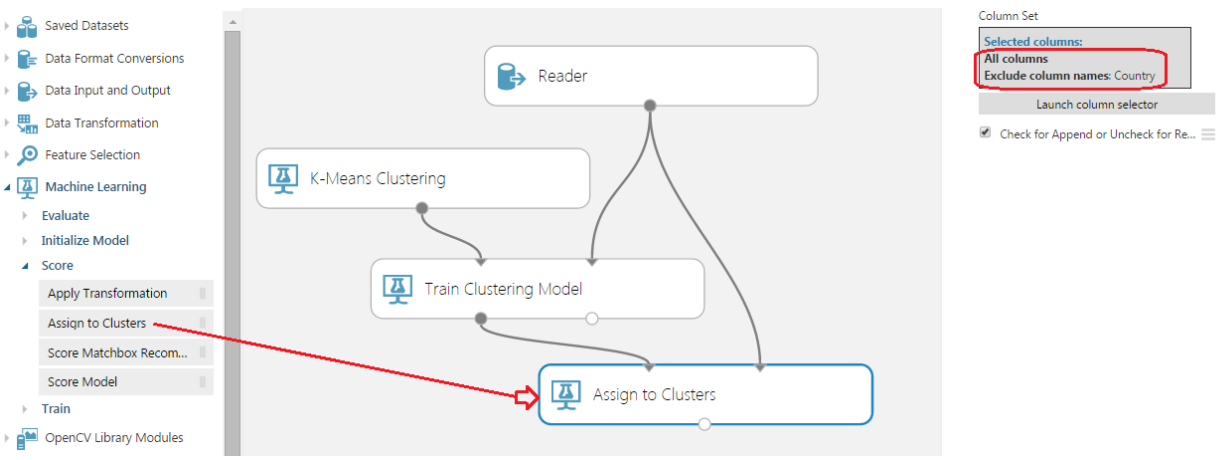
Chapter 8


Cluster 1
K = 3
Cluster 2
Centroid 1
Centroid 2
Centroid 3
Cluster 3

Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5

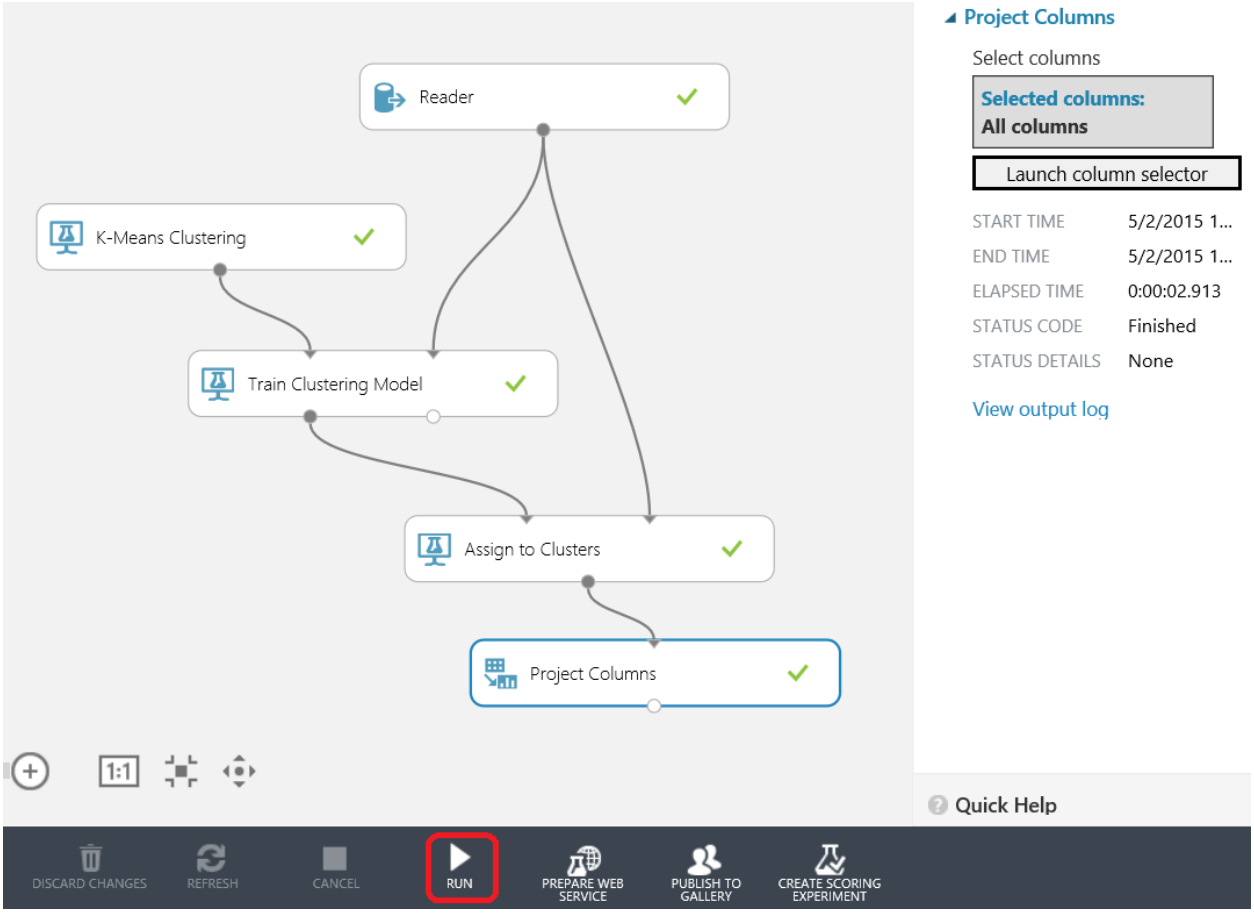




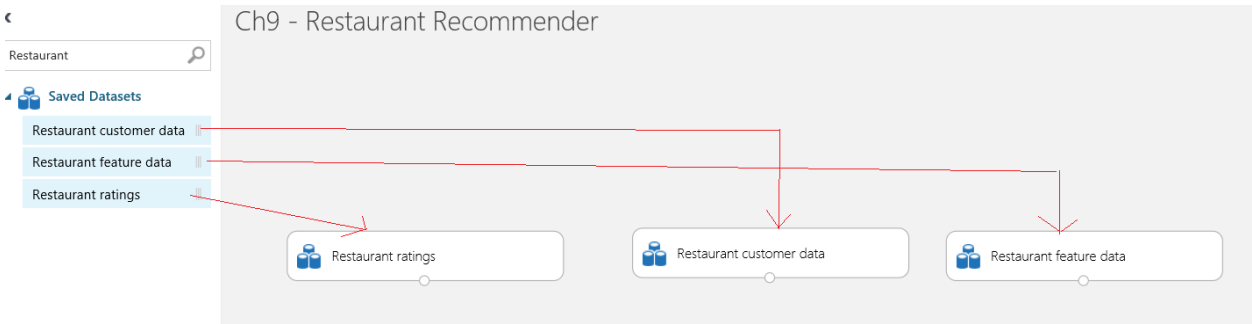


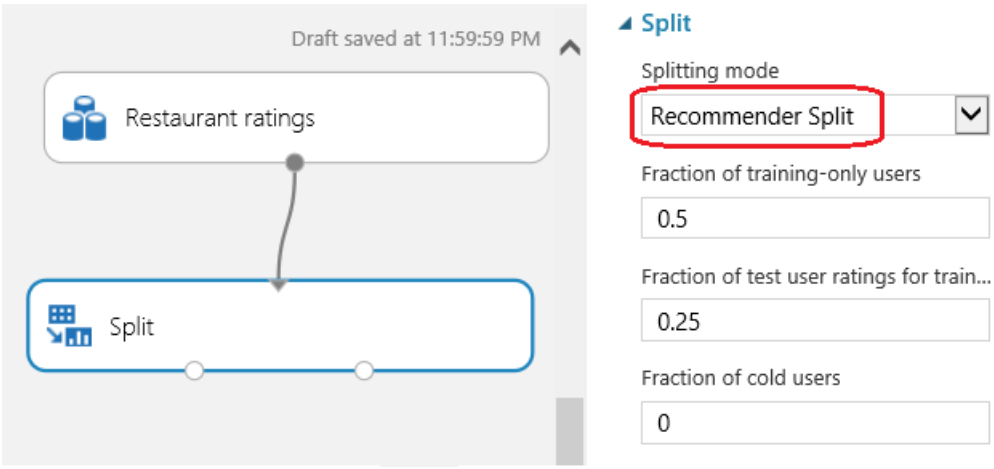
view as

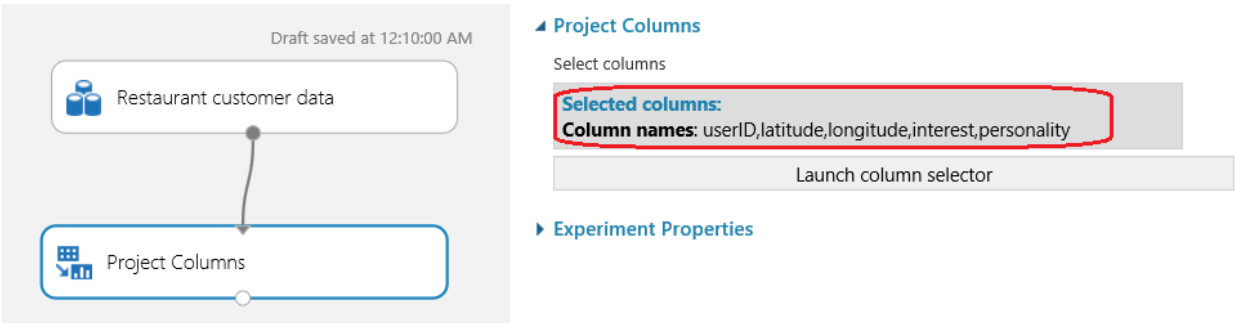
Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg	Assignments
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7	0
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3	1
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4	1
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2	0
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4	2
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4	1
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6	2
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4	1
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5	1
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5	0
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2	2

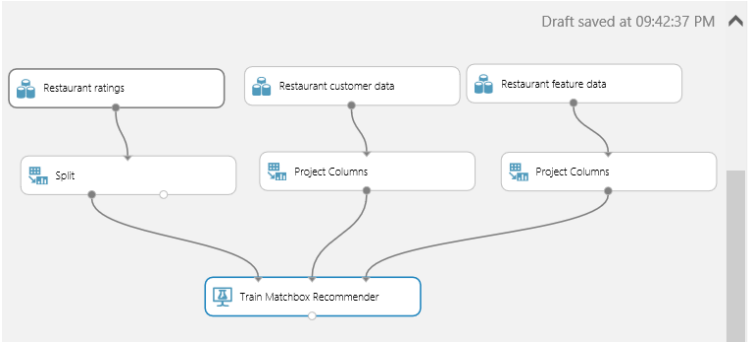

Reader
K-Means Clustering
Train Clustering Model
Assign to Clusters
Project Columns
Project Columns
Select columns
Selected columns:
All columns
Launch column selector
START TIME 5/2/2015 1...
END TIME 5/2/2015 1...
ELAPSED TIME 0:00:02.913
STATUS CODE Finished
STATUS DETAILS None
View output log
Quick Help
DISCARD CHANGES
REFRESH
CANCEL
RUN
PREPARE WEB SERVICE
PUBLISH TO GALLERY
CREATE SCORING EXPERIMENT

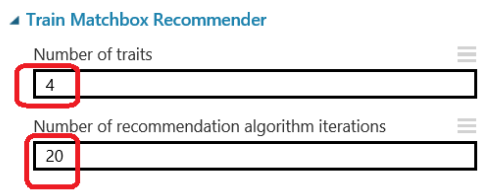
Chapter 9

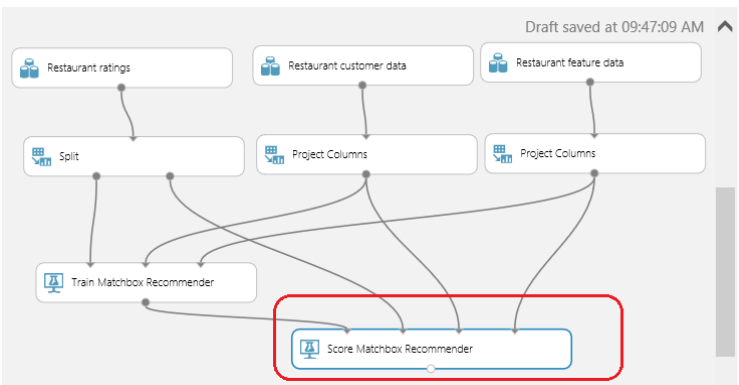

Restaurant
Saved Datasets
Restaurant customer data
Restaurant feature data
Restaurant ratings
Ch9 - Restaurant Recommender
Restaurant ratings
Restaurant customer data
Restaurant feature data

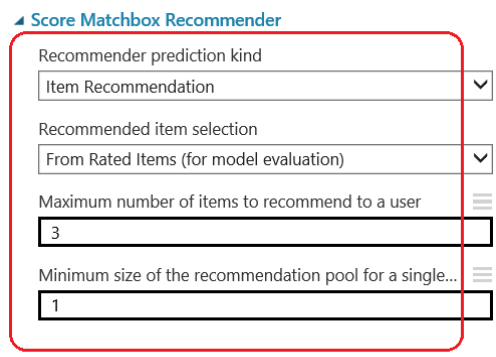

Draft saved at 11:59:59 PM
Restaurant ratings
Split
Split
Splitting mode
Recommender Split
Fraction of training-only users
0.5
Fraction of test user ratings for train...
0.25
Fraction of cold users
0
0


Draft saved at 12:10:00 AM
Restaurant customer data
Project Columns
Project Columns
Select columns
Selected columns:
Column names: userID,latitude,longitude,interest,personality
Launch column selector
Experiment Properties



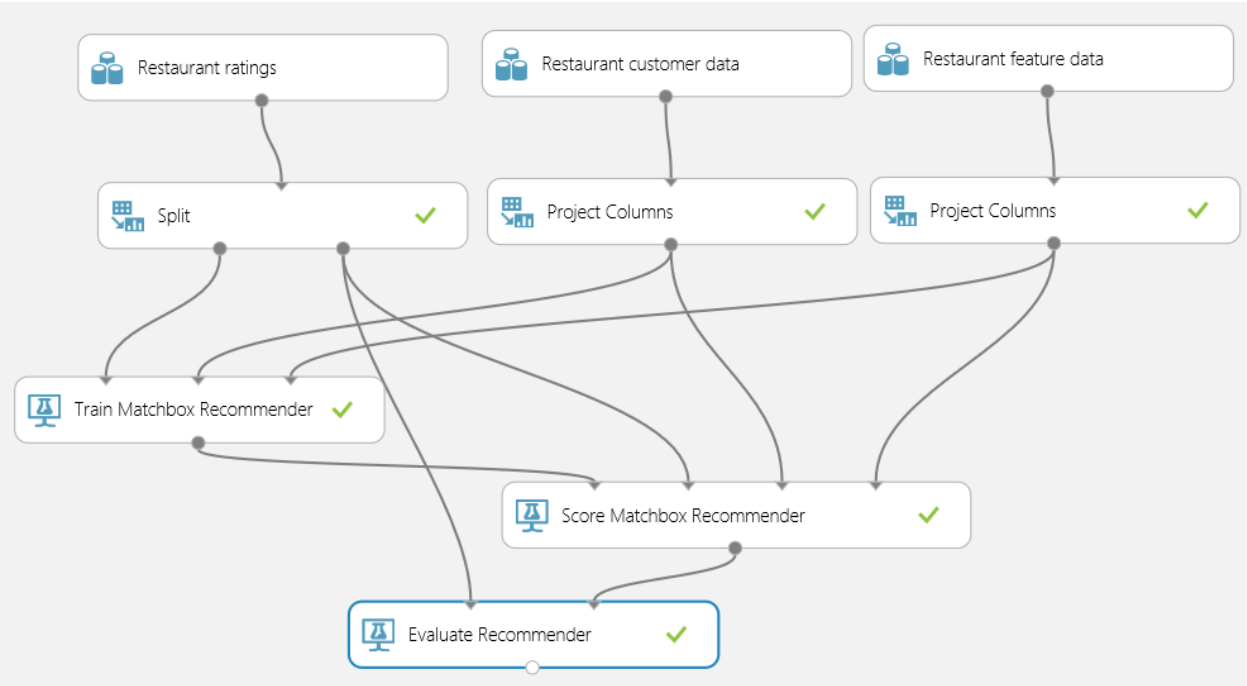


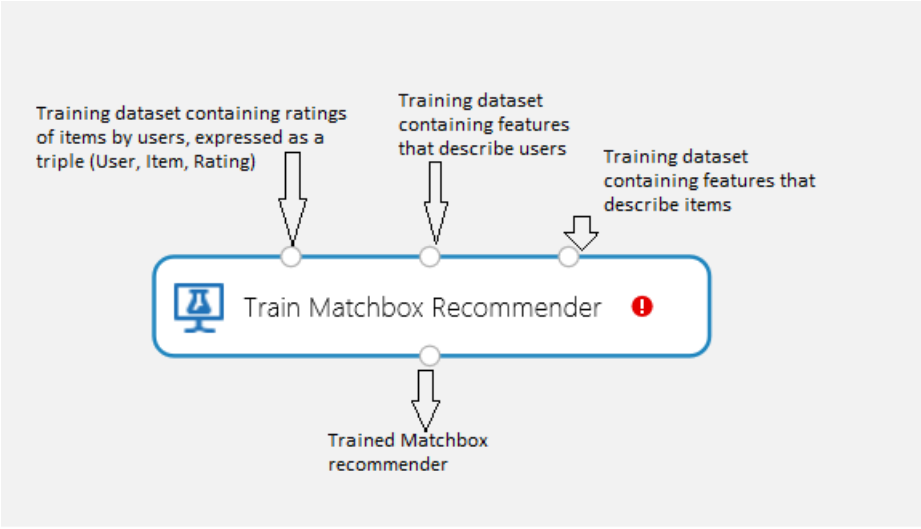


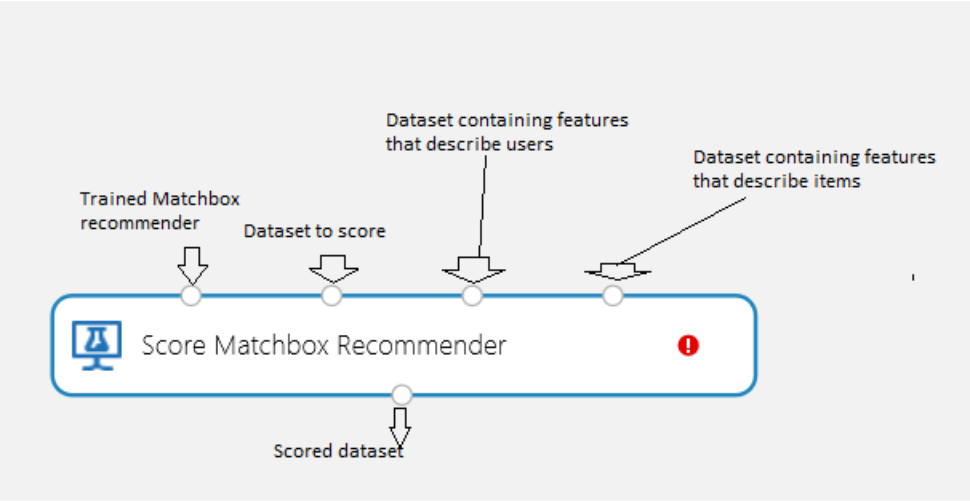


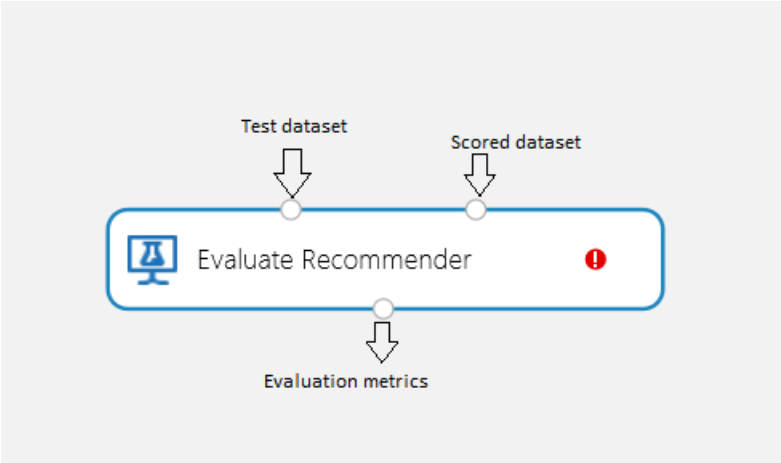
view as

User	Item 1	Item 2	Item 3
U1048	135034	135026	135065
U1117	135018	132766	135088
U1049	135052	132862	135051
U1088	135057	135071	135032
U1062	135052	135045	135062
U1035	134986	135018	132773
U1125	135062	135076	135038
U1013	135075	135079	132921

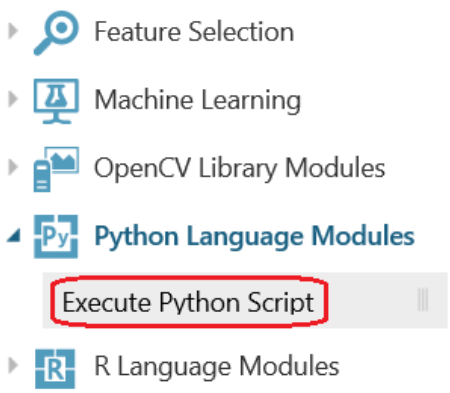

Restaurant ratings
Restaurant customer data
Restaurant feature data
Split
Project Columns
Project Columns
Train Matchbox Recommender
Score Matchbox Recommender
Evaluate Recommender

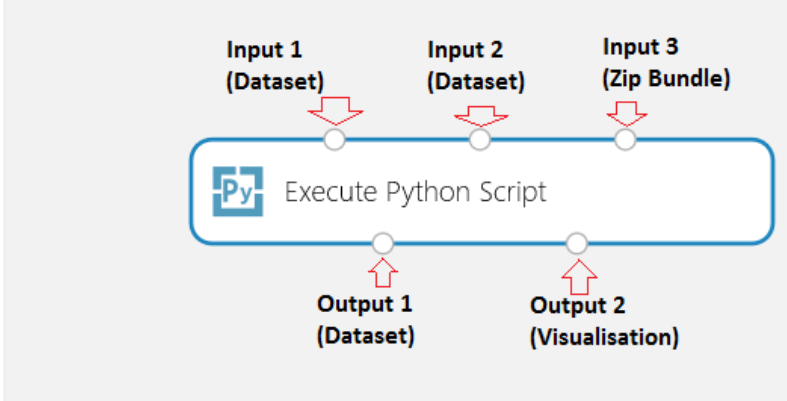

Training dataset containing ratings of items by users, expressed as a triple (User, Item, Rating)
Training dataset containing features that describe users
Training dataset containing features that describe items
Train Matchbox Recommender
Trained Matchbox recommender


Dataset containing features
that describe users
Dataset containing features
that describe items
Trained Matchbox
recommender
Dataset to score
Score Matchbox Recommender
Scored dataset


Test dataset
Scored dataset
Evaluate Recommender
Evaluation metrics

Chapter 10



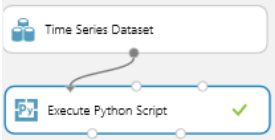


Python script

```
# The script MUST contain a function named azureml_main
# which is the entry point for this module.
#
# The entry point function can contain up to two input arguments:
#   Param<dataframe1>: a pandas.DataFrame
#   Param<dataframe2>: a pandas.DataFrame
                                  Input 1
def azureml_main(dataframe1 = None, dataframe2 = None):
                                                Input 2
###### Execution logic goes here ##########
# Code to get the result dataFrame
    resultDataFrame = ...

# Code to generate the visualisation (if any!)
    ...
# Return value must be of a sequence of pandas.DataFrame
    return resultDataFrame,
                        Output 1
```

In draft

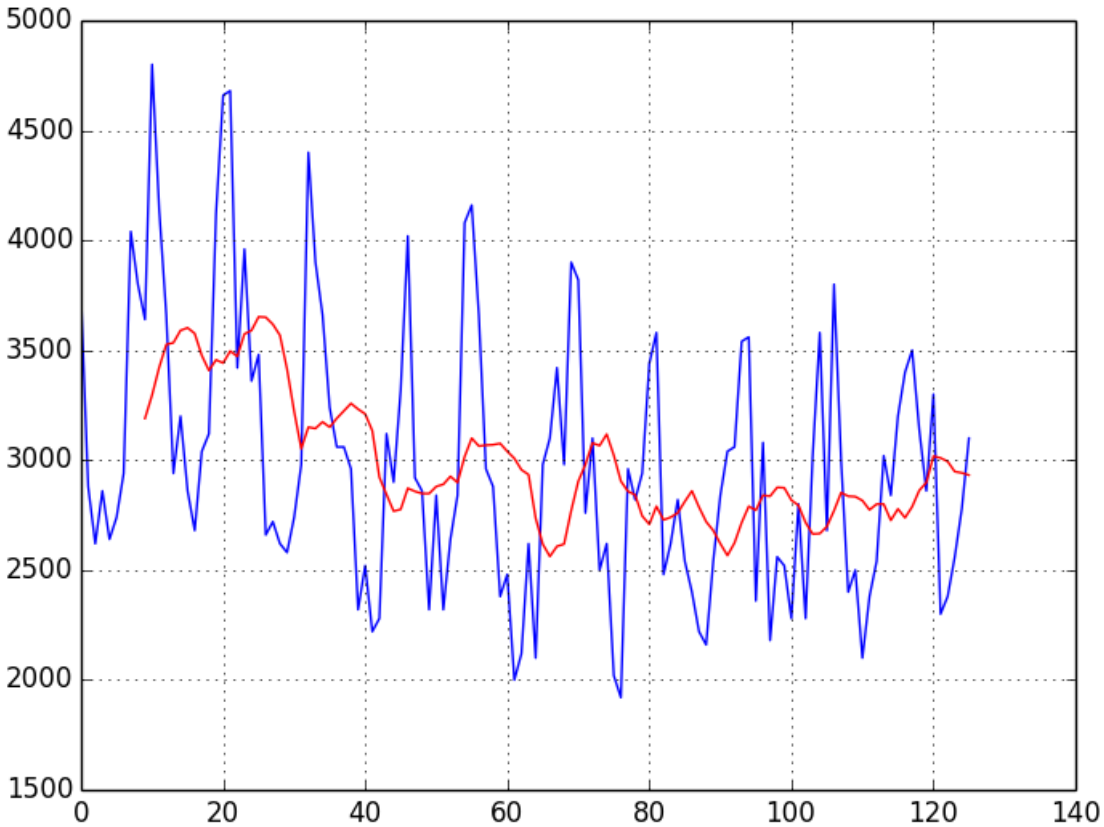
Draft saved at 07:31:53 PM

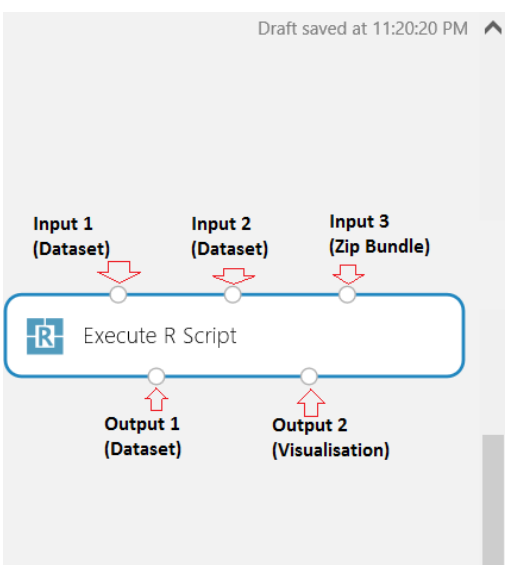


Properties

Python script

```python
def azureml_main(dataframe1 = None, dataframe2 = None):
##### Import package(s) #####
    import pandas as pd
    import numpy as np
    import matplotlib
# Change to use a non-interactive backend, Agg (for PNGs)
    matplotlib.use("agg")
    import matplotlib.pyplot as plt
# Assign the input data frame to a new one
    sampleTimeSeries = dataframe1
# Calculate moving average. Results in a series
    movingAverage = pd.rolling_mean(sampleTimeSeries['N1725'], 10)
# Add a new column for moving average data and assigne the value
    sampleTimeSeries['MovingAverage'] = movingAverage

#####  Visualisation ######
# Create New figure
    fig = plt.figure()
    ax = fig.gca()
####  Plot into specified axis
# Plot the time series
    sampleTimeSeries['N1725'].plot(ax=ax)
# Add line for moving average
    sampleTimeSeries['MovingAverage'].plot(ax=ax, color='red')
# Save the figure to image
    fig.savefig('TimeSeriesWithMovingAverage.png')
# Return value(pandas.DataFrame) -  time series data with moving average
    return sampleTimeSeries,
```

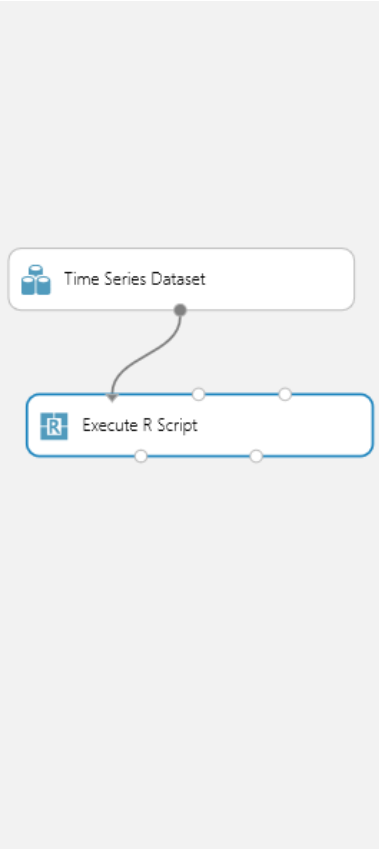




Execute R Script

R Script

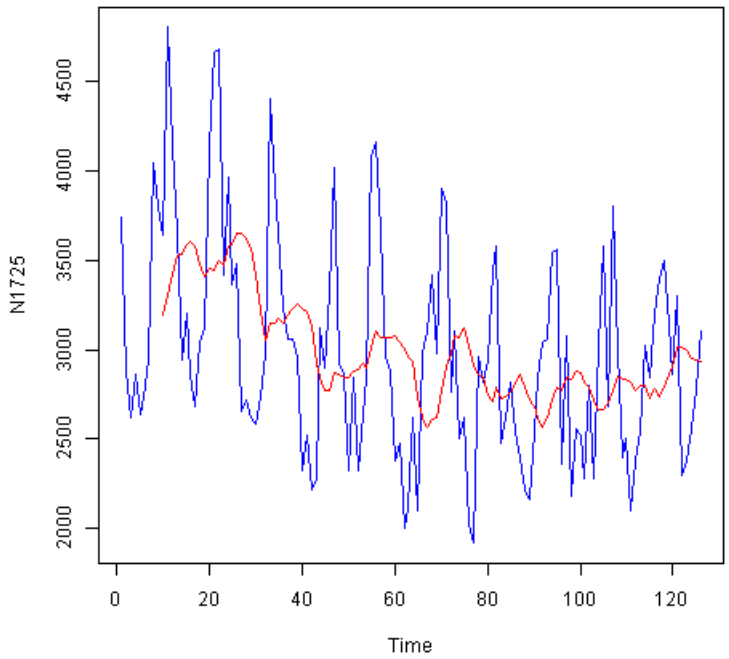
```
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1) # class: data.frame
dataset2 <- maml.mapInputPort(2) # class: data.frame

# Contents of optional Zip port are in ./src/
# source("src/yourfile.R");
# load("src/yourData.rdata");

# Sample operation
data.set = rbind(dataset1, dataset2);

# You'll see this output in the R Device port.
# It'll have your stdout, stderr and PNG graphics device(s).
plot(data.set);

# Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("data.set");
```



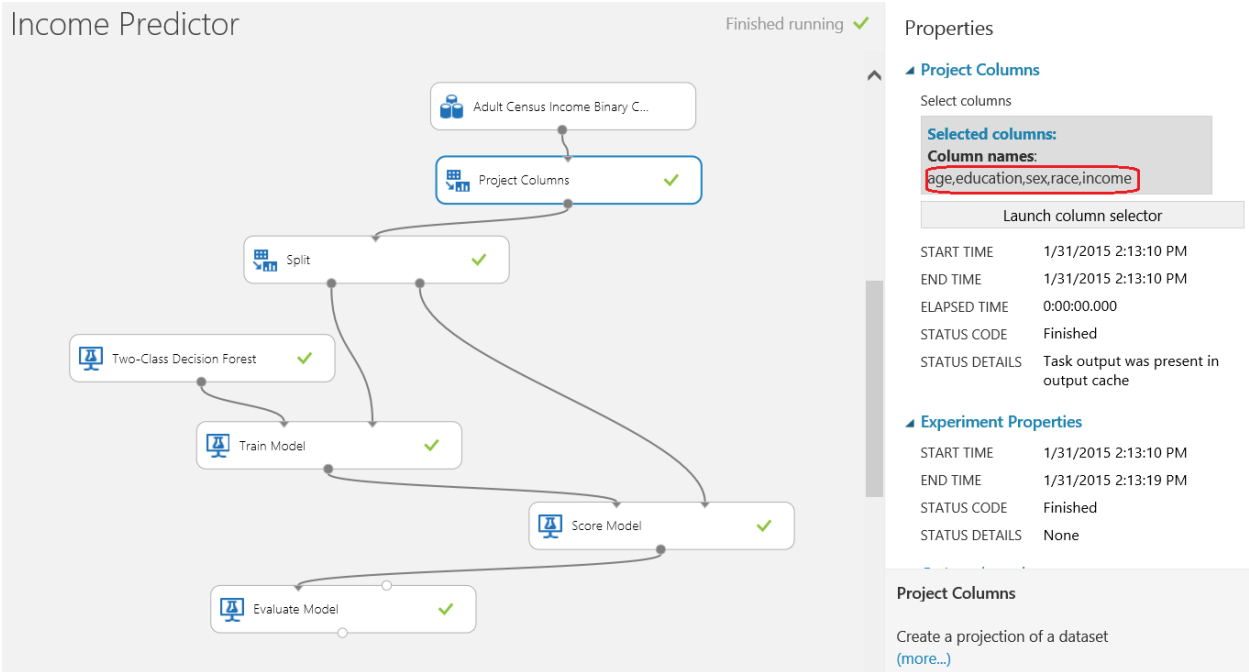
R Script

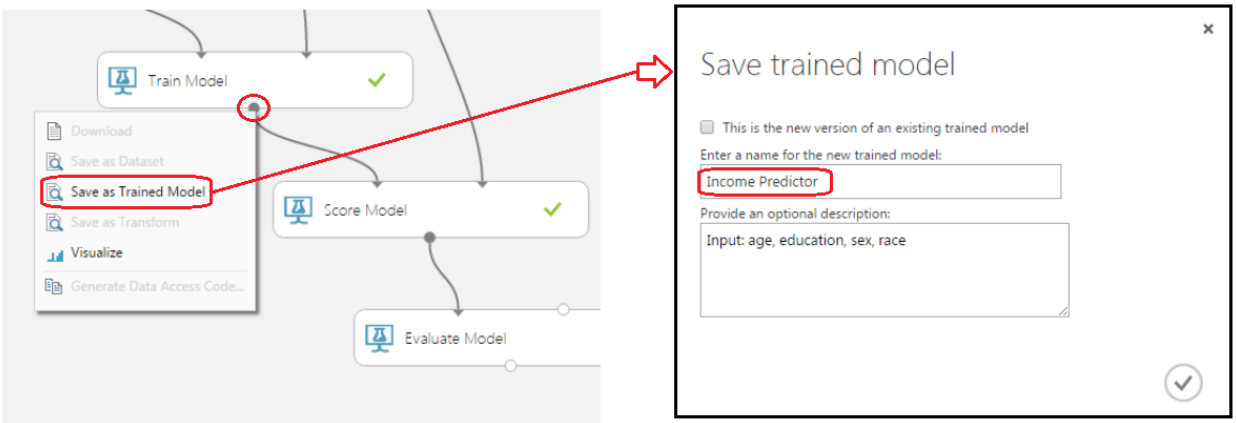
```r
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1) # class: data.frame

# Assign input dataset to new one
timeSeries <- dataset1

# Load the zoo package to the session
library(zoo)

# Calculate moving average with moving window 10
ma <- rollmean(timeSeries[2],10)
# Add a new column and assign moving average values to it
timeSeries$MovingAverage[10:126] <- ma

##### Visualisation ######
# Plot time series
plot.ts(timeSeries['N1725'], col='blue')
# Add a line for moving average
lines(timeSeries['MovingAverage'], col='red')

# Select data.frame to be sent to the output Dataset port -
# time series data with moving average
maml.mapOutputPort("timeSeries");
```

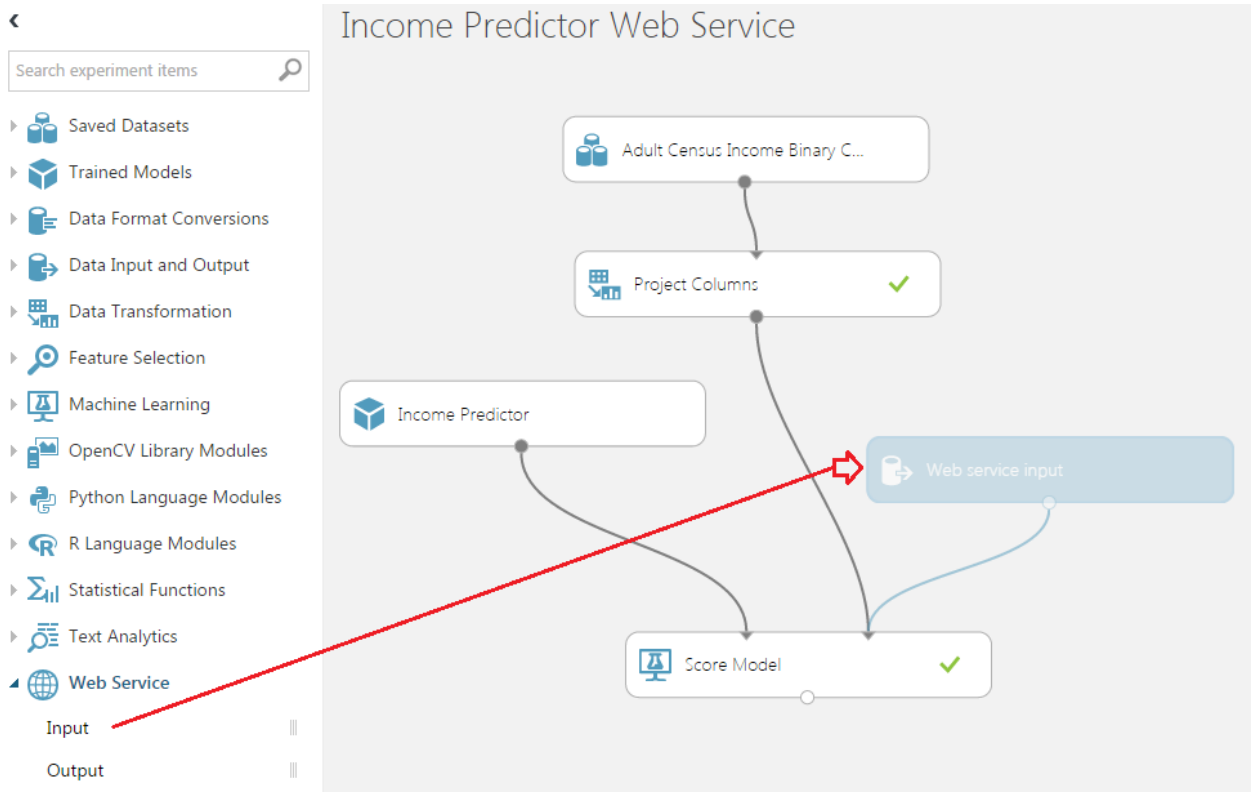
- Standard Output
- Standard Error
- Graphics Device

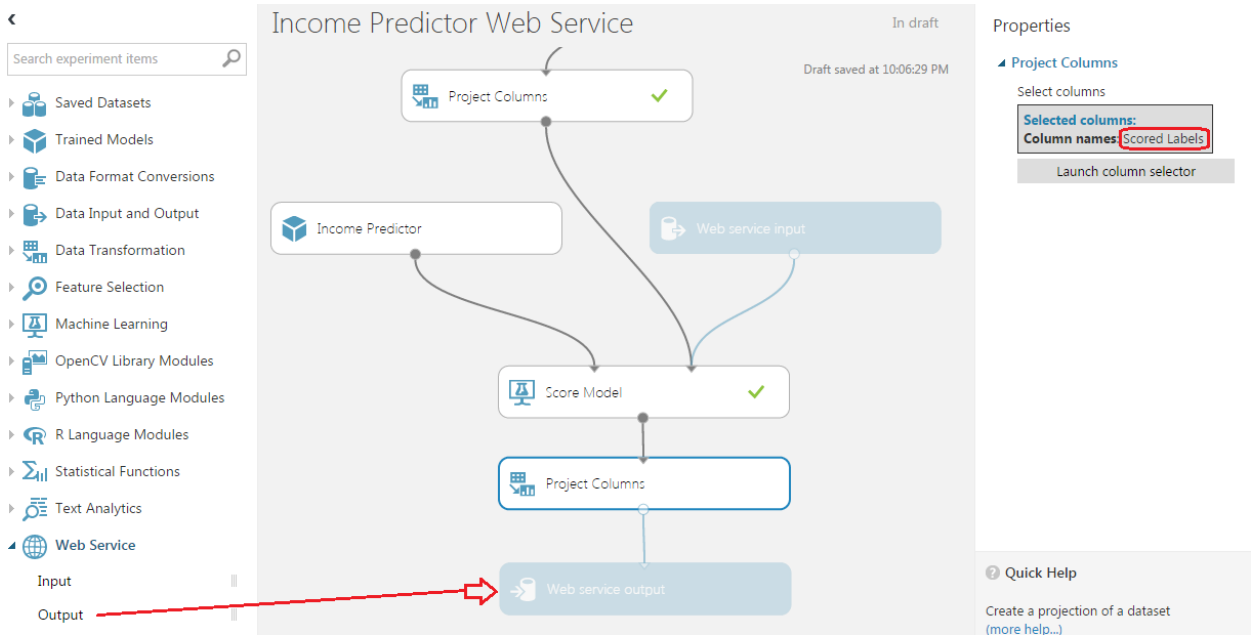


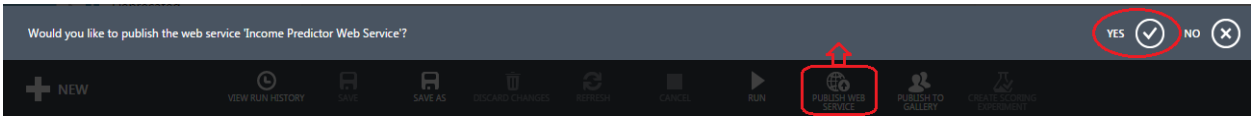
Chapter 11


Income Predictor
Finished running
Adult Census Income Binary C...
Project Columns
Split
Two-Class Decision Forest
Train Model
Score Model
Evaluate Model
Properties
Project Columns
Select columns
Selected columns:
Column names:
age,education,sex,race,income
Launch column selector
START TIME 1/31/2015 2:13:10 PM
END TIME 1/31/2015 2:13:10 PM
ELAPSED TIME 0:00:00.000
STATUS CODE Finished
STATUS DETAILS Task output was present in output cache
Experiment Properties
START TIME 1/31/2015 2:13:10 PM
END TIME 1/31/2015 2:13:19 PM
STATUS CODE Finished
STATUS DETAILS None
Project Columns
Create a projection of a dataset
(more...)


Train Model
Download
Save as Dataset
Save as Trained Model
Save as Transform
Visualize
Generate Data Access Code...
Score Model
Evaluate Model
Save trained model
This is the new version of an existing trained model
Enter a name for the new trained model:
Income Predictor
Provide an optional description:
Input: age, education, sex, race


Search experiment items
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Web Service
Input
Output
Income Predictor Web Service
Adult Census Income Binary C...
Project Columns
Income Predictor
Web service input
Score Model


Search experiment items
Saved Datasets
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Web Service
Input
Output
Income Predictor Web Service
In draft
Draft saved at 10:06:29 PM
Project Columns
Income Predictor
Web service input
Score Model
Project Columns
Web service output
Properties
Project Columns
Select columns
Selected columns:
Column names: Scored Labels
Launch column selector
Quick Help
Create a projection of a dataset
(more help...)


Would you like to publish the web service 'Income Predictor Web Service'?
YES
NO
NEW
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PUBLISH WEB SERVICE
PUBLISH TO GALLERY
CREATE SCORING EXPERIMENT

income predictor web service

DASHBOARD CONFIGURATION

1

General

Published experiment

View snapshot View latest 2

3

Description

No description provided for this web service.

API key 4

oyeUahaQIl5OBv/KXngd1SneS85OQ5EEVQMcoYDC6LoTzpu5eOIMIOY6N8QFR+Q8HPqJX/M+z4k+zPHYPk

Default Endpoint

API HELP PAGE	TEST	APPS
REQUEST/RESPONSE 5	Test 7	Download Excel Workbook
BATCH EXECUTION 6		8

Additional endpoints

Number of additional endpoints created for this web service: 0

Manage endpoints in Azure management portal

Test Income Predictor Web Service Service

Enter data to predict

AGE

49

EDUCATION

Masters

RACE

White

SEX

Male

'Income Predictor Web Service' test returned [">50K"]...

DETAILS CLOSE

NEW DELETE

'Income Predictor Web Service' test returned [">50K"]...

CLOSE

Result: {"Results":{"output1":{"type":"table","value":{"ColumnNames":["Scored Labels"],"ColumnTypes":["String"],"Values":[[">50K"]]}}}}

EXPERIMENTS

WEB SERVICES

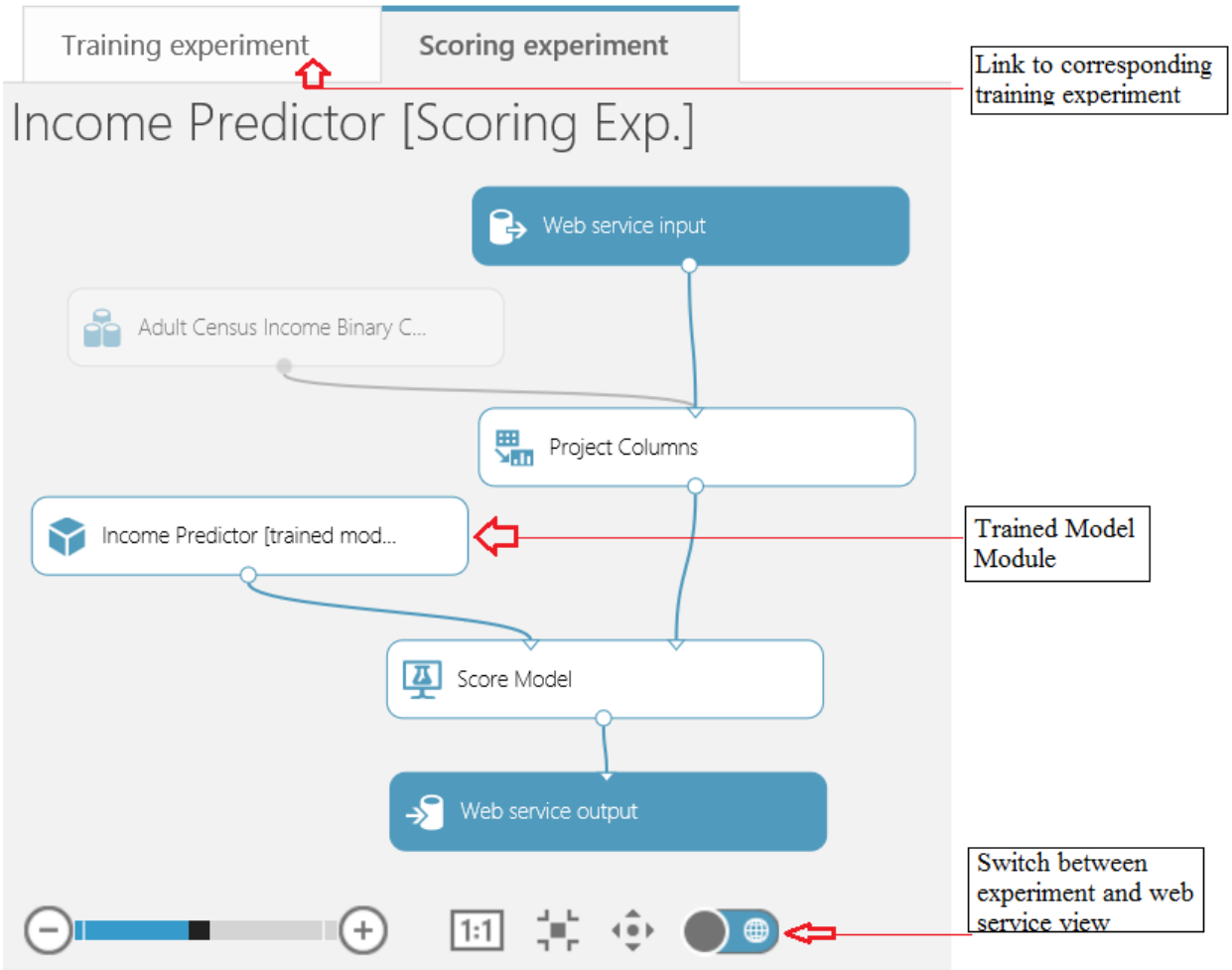
DATASETS

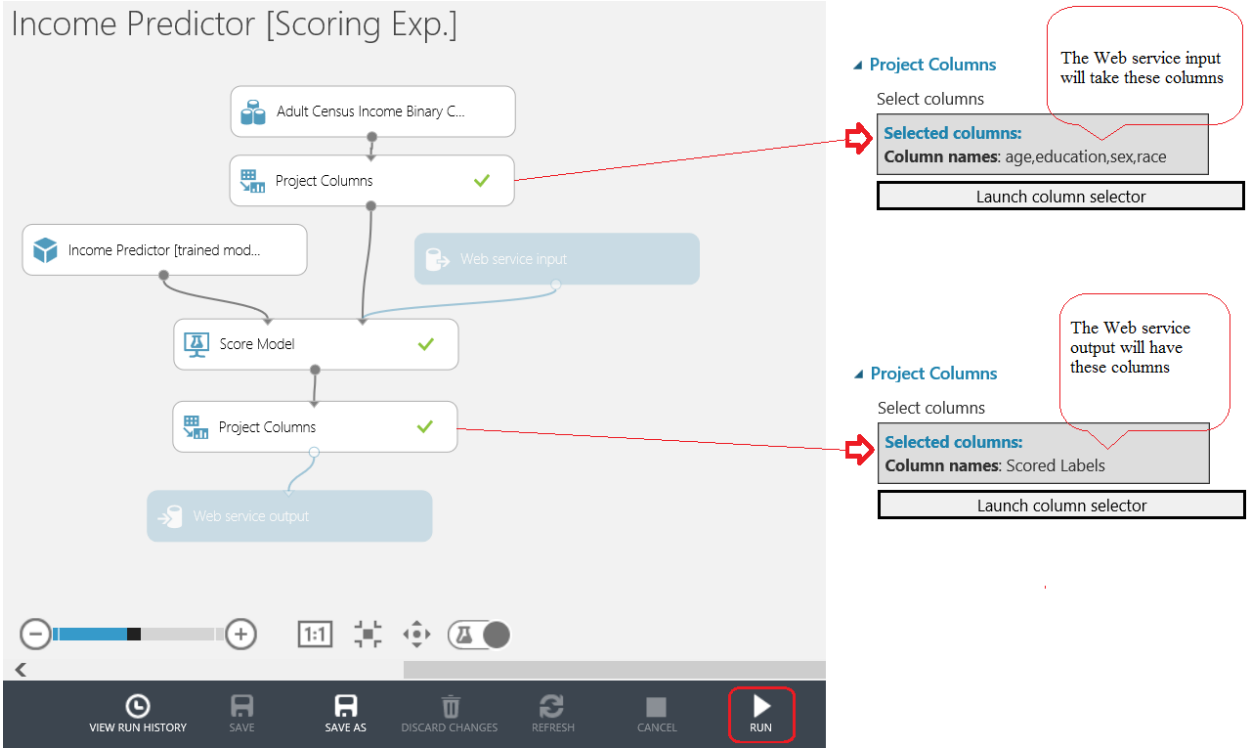
TRAINED MODELS

web services

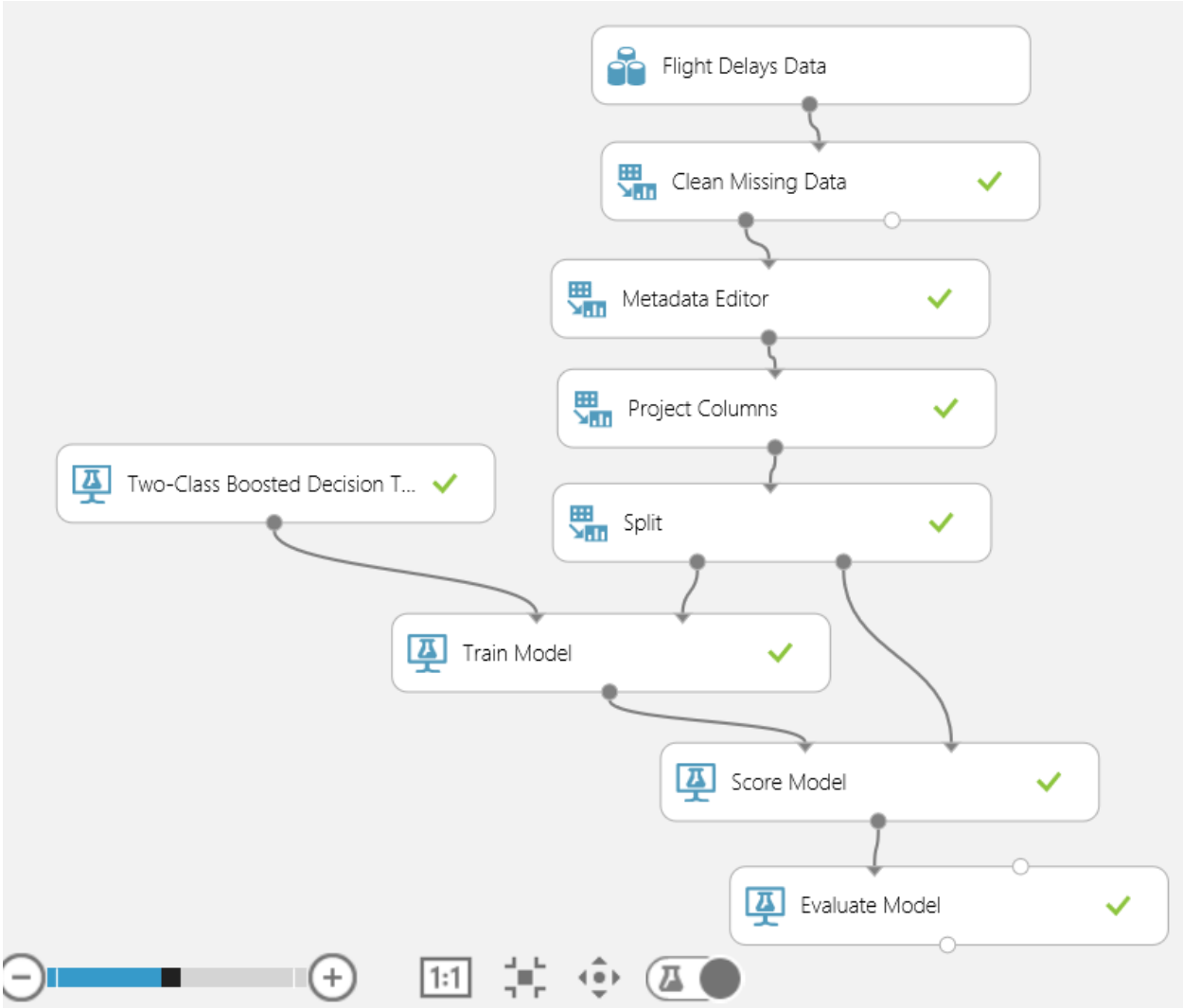
NAME	CREATED ON
Income Predictor Web Service	5/8/2015 10:17:24 PM

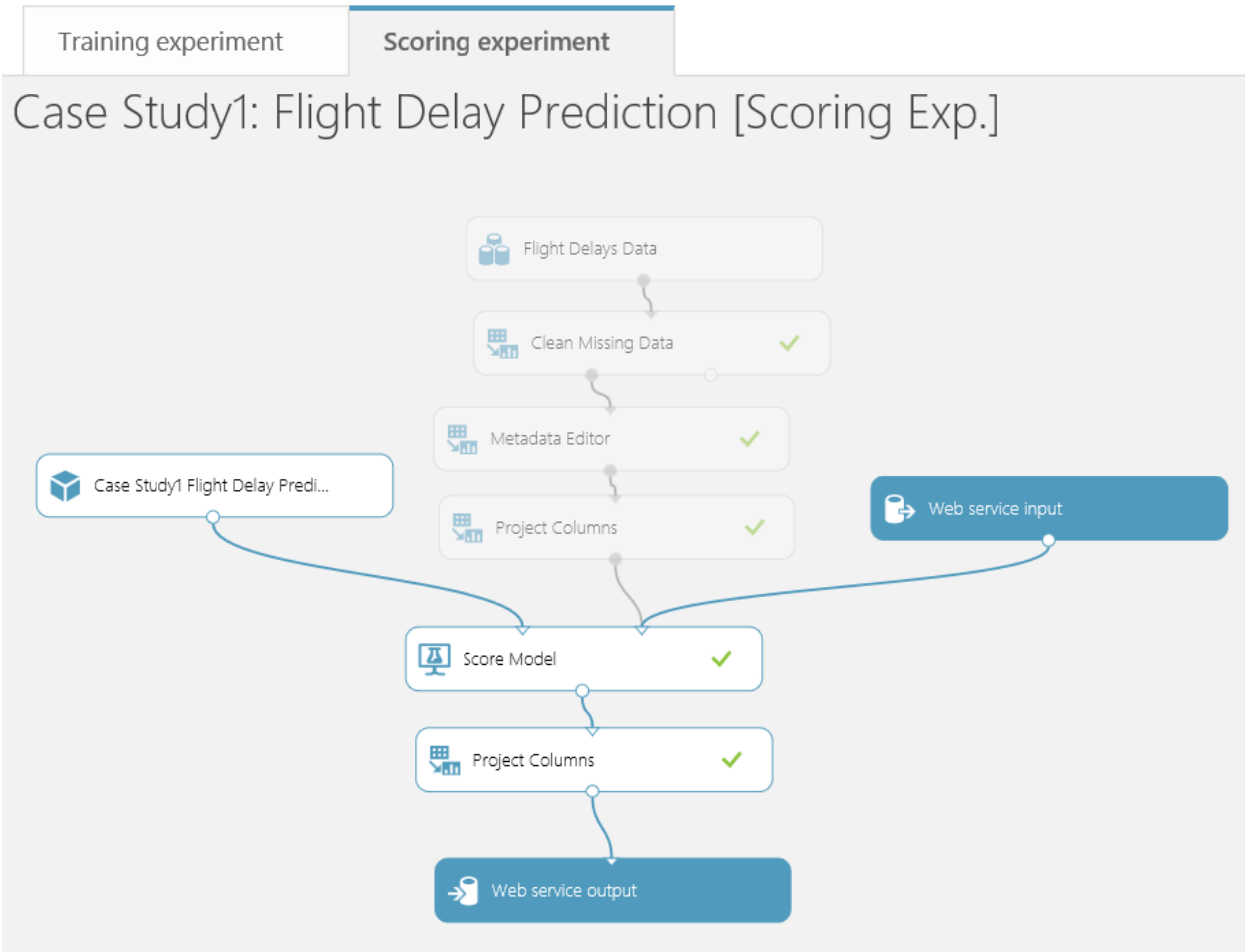

SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
PUBLISH WEB SERVICE
PUBLISH TO GALLERY
CREATE SCORING EXPERIMENT


Training experiment
Scoring experiment
Link to corresponding training experiment
Income Predictor [Scoring Exp.]
Web service input
Adult Census Income Binary C...
Project Columns
Income Predictor [trained mod...
Trained Model Module
Score Model
Web service output
Switch between experiment and web service view
1:1

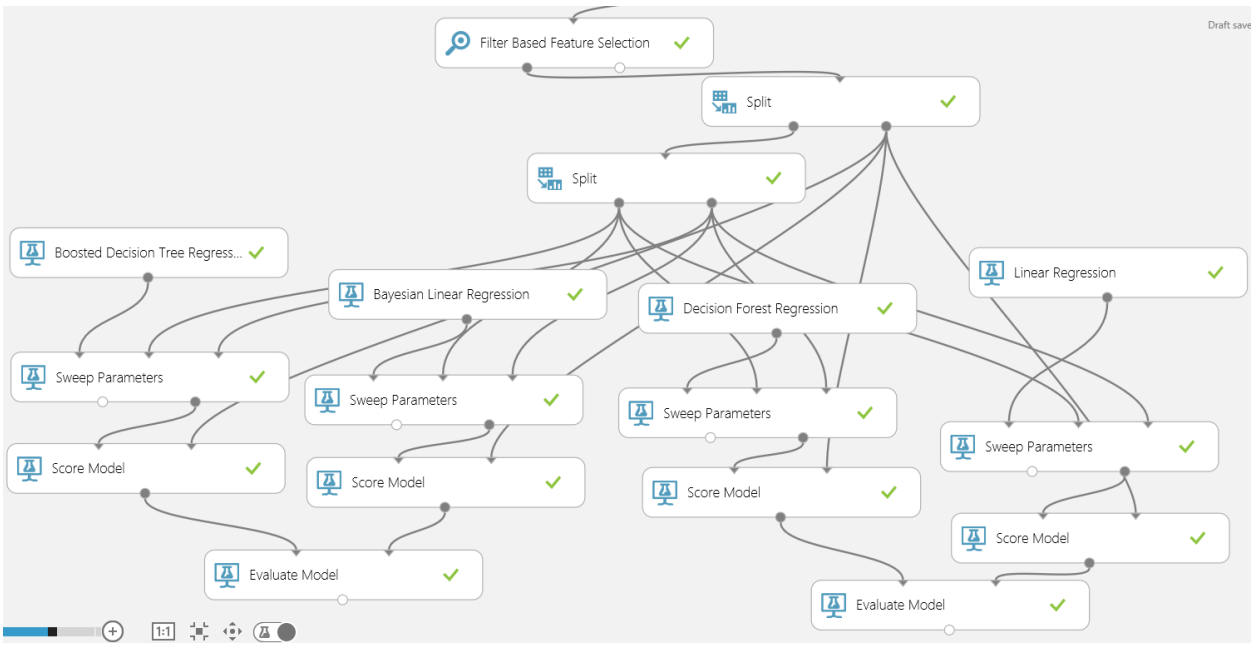

Income Predictor [Scoring Exp.]
Adult Census Income Binary C...
Project Columns
Income Predictor [trained mod...
Web service input
Score Model
Project Columns
Web service output
VIEW RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
REFRESH
CANCEL
RUN
Project Columns
The Web service input will take these columns
Select columns
Selected columns:
Column names: age,education,sex,race
Launch column selector
Project Columns
The Web service output will have these columns
Select columns
Selected columns:
Column names: Scored Labels
Launch column selector

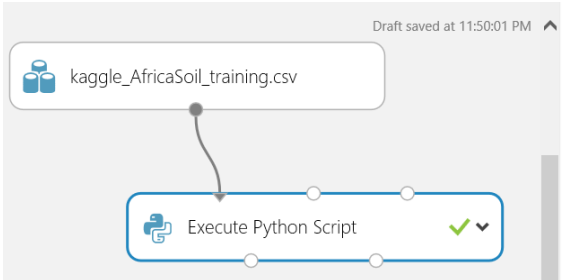
Chapter 12


Flight Delays Data
Clean Missing Data
Metadata Editor
Project Columns
Two-Class Boosted Decision T...
Split
Train Model
Score Model
Evaluate Model
1:1


Training experiment
Scoring experiment
Case Study1: Flight Delay Prediction [Scoring Exp.]
Flight Delays Data
Clean Missing Data
Metadata Editor
Case Study1 Flight Delay Predi...
Project Columns
Web service input
Score Model
Project Columns
Web service output

Chapter 13

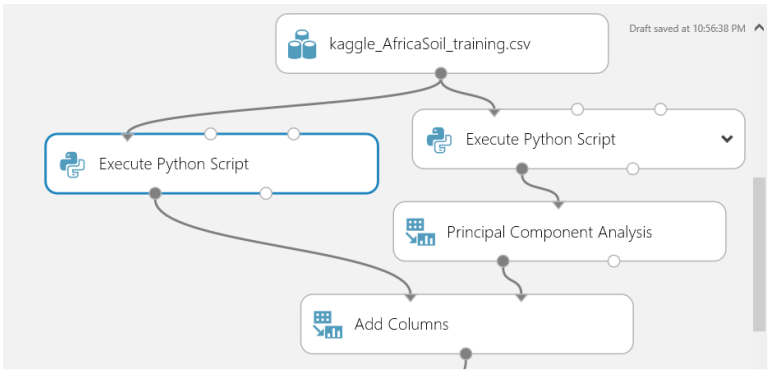




Execute Python Script

Python script

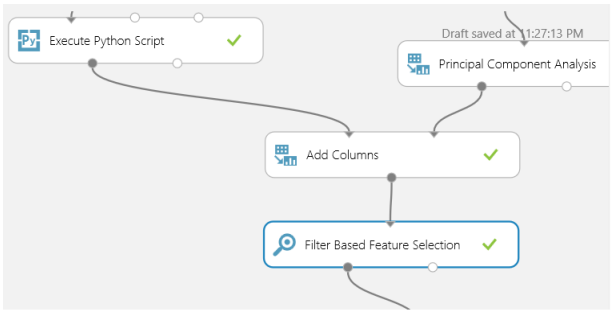
```
def azureml_main(dataframe1 = None, dataframe2 = None):
#Get all the columns
    cols = dataframe1.columns.tolist()
#Select columns with name starting with letter 'm'
    dataframe1=dataframe1[[col for col in cols if col.startswith('m')]]
#Return the modified dataset
    return dataframe1

```

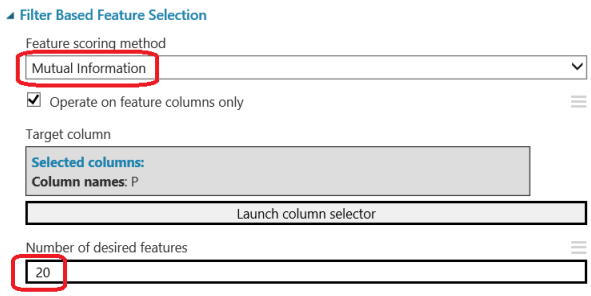


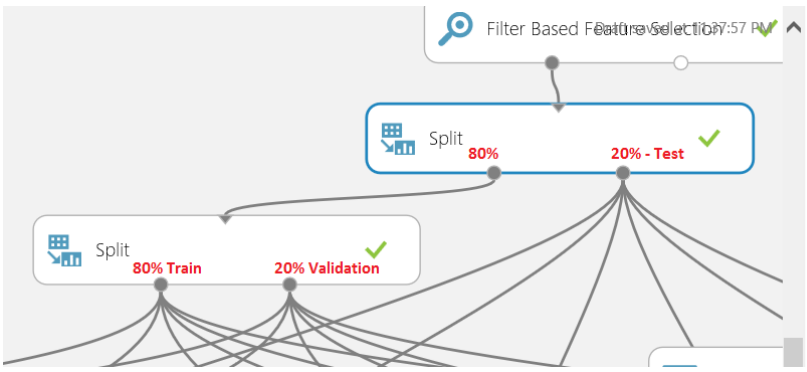
Execute Python Script

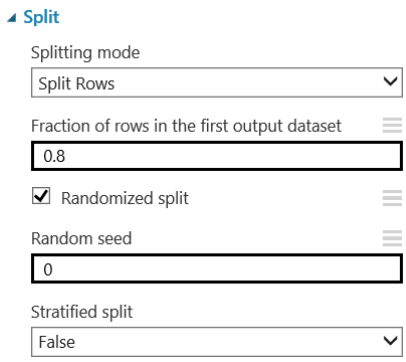
Python script

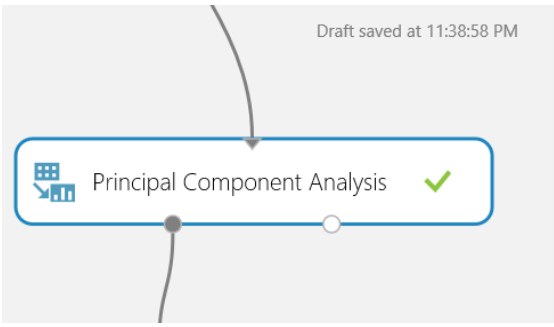
```
def azureml_main(dataframe1 = None, dataframe2 = None):
#Get all the columns
    cols = dataframe1.columns.tolist()
#Select columns with name starting with letter 'm'
    mCols=dataframe1[[col for col in cols if col.startswith('m')]]
#List of other columns to be excluded
    exCols=['PIDN', 'Ca',  'pH', 'SOC', 'Sand']
#Drop all the columns with name starting with letter 'm'
    dataframe1.drop(mCols, axis=1, inplace=True)
#Drop other columns - 'PIDN', 'Ca', 'pH', 'SOC', 'Sand'
    dataframe1.drop(exCols,axis=1, inplace=True)
#Return the modified dataset
    return dataframe1
```

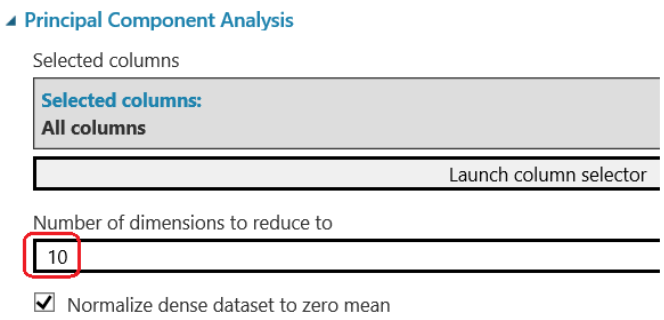

Execute Python Script
Draft saved at 11:27:13 PM
Principal Component Analysis
Add Columns
Filter Based Feature Selection


Filter Based Feature Selection
Feature scoring method
Mutual Information
Operate on feature columns only
Target column
Selected columns:
Column names: P
Launch column selector
Number of desired features
20


Filter Based Feature Selection
Split
80%
20% - Test
Split
80% Train
20% Validation


Split
Splitting mode
Split Rows
Fraction of rows in the first output dataset
0.8
Randomized split
Random seed
0
Stratified split
False


Draft saved at 11:38:58 PM
Principal Component Analysis


Principal Component Analysis
Selected columns
Selected columns:
All columns
Launch column selector
Number of dimensions to reduce to
10
Normalize dense dataset to zero mean