Azure ML Tutorial

- Azure Bootcamp in Troy -

2016. 4. 16

Ingyu Lee
(inlee@troy.edu)

Contents

I Introduction

1. What is AzureML?
2. AzureML Workflow
3. AzureML Key Terminology and Concept

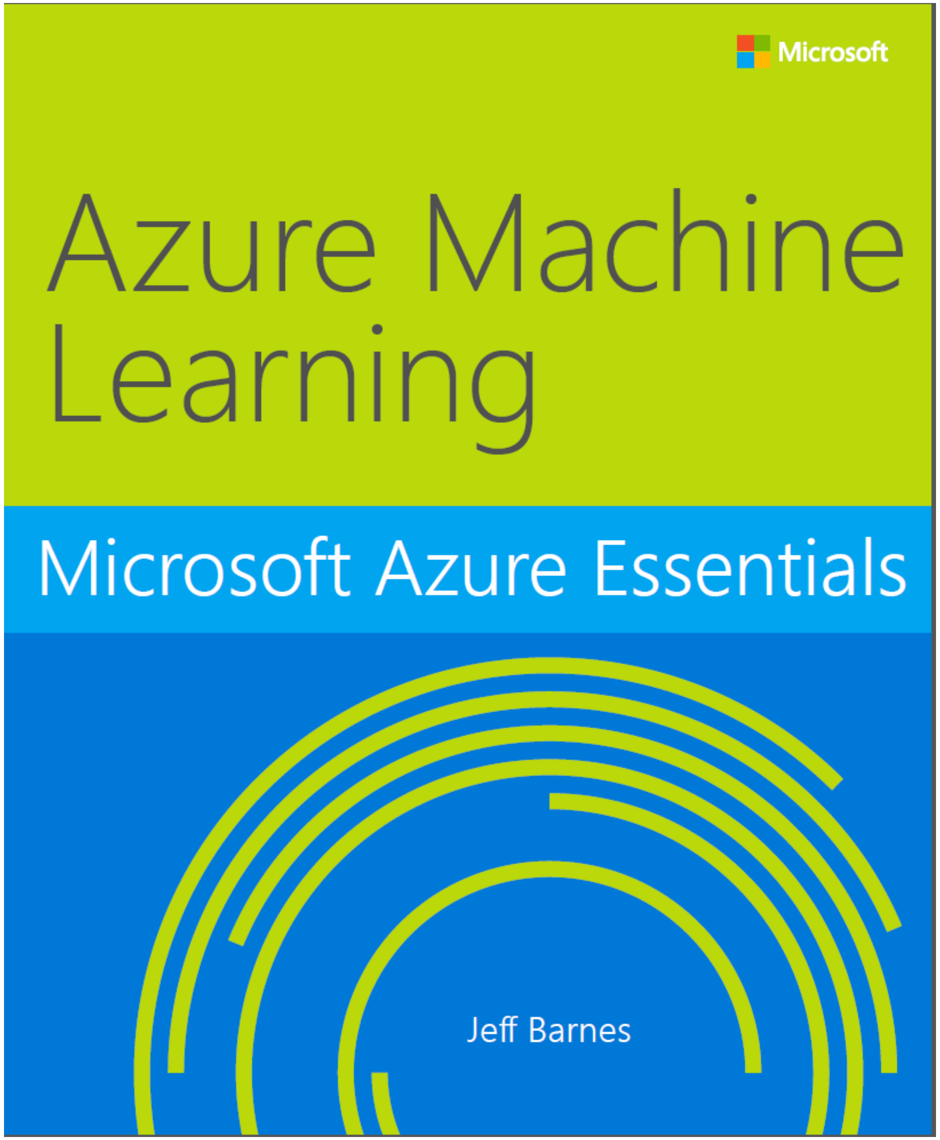
II The First Experiment of AzureML

1. Overview of AzureML
2. Create a Model
3. Build a Model
4. Test and Evaluate the model
5. Publish as a Web Service

III AzureML Hands-On

1. Classification
2. Cluster analytics
3. The Azure ML Matchbox recommender
4. Exercises

IV Lessons / Resources



Microsoft

Azure Machine Learning

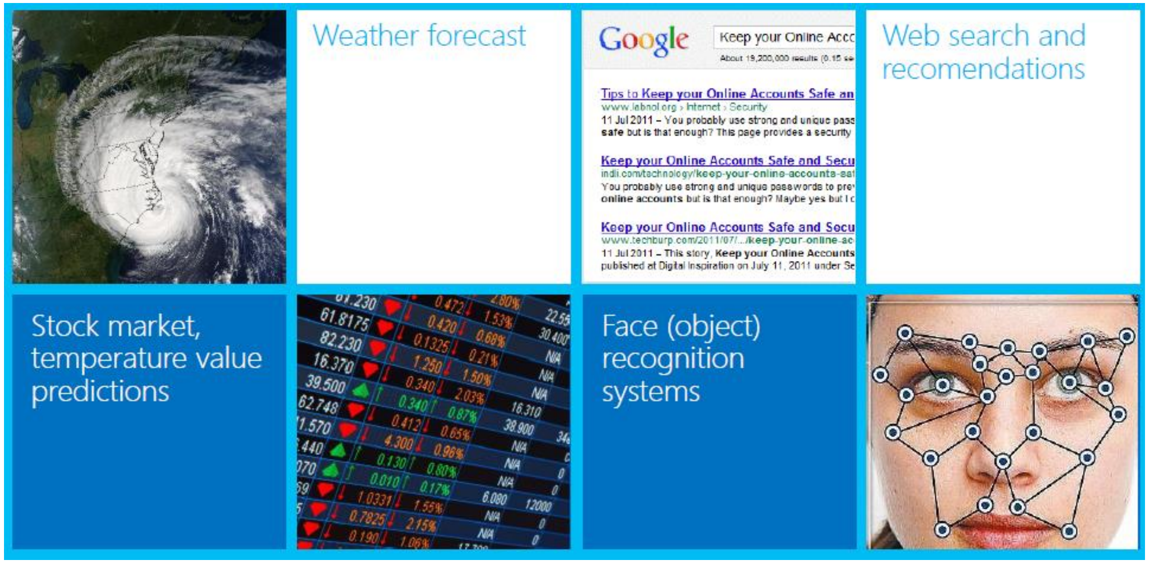
Microsoft Azure Essentials

Jeff Barnes

Book, Microsoft Azure Essentials: Azure Machine Learning,
http://www.microsoftvirtualacademy.com/ebooks#9780735698178

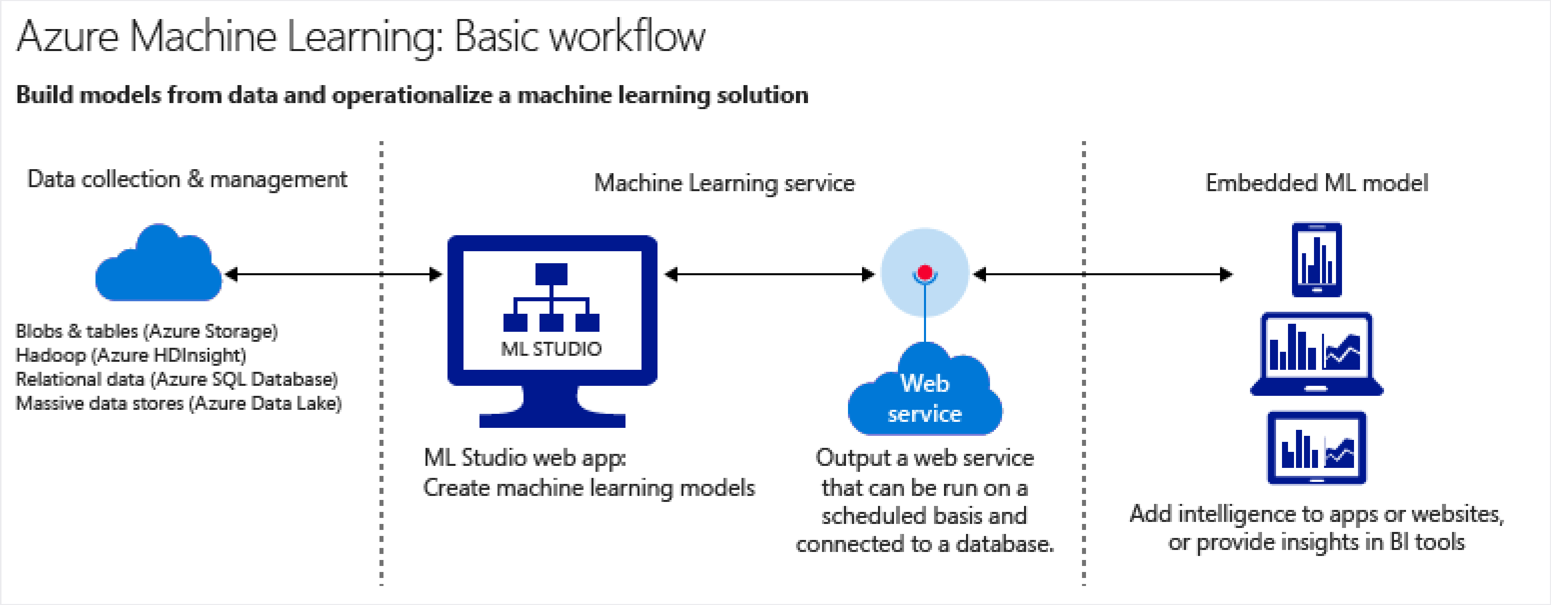
Part I What is Azure ML(Machine Learning)?

1. What is machine learning?
   a. Machine learning uses computers to run predictive models that learn from existing data in order to forecast future behaviors, outcomes, and trends.
   b. These forecasts or predictions from machine learning can make apps and devices smarter.
   c. When you shop online, machine learning helps recommend other products you might like based on what you’ve purchased.
   d. When your credit card is swiped, machine learning compares the transaction to a database of transactions and helps the bank do fraud detection.


Weather forecast
Google
Keep your Online Acc
Tips to Keep your Online Accounts Safe an
Keep your Online Accounts Safe and Secu
Keep your Online Accounts Safe and Secu
Web search and recomendations
Stock market, temperature value predictions
Face (object) recognition systems

Part I What is Azure ML(Machine Learning)?

2. What is Machine Learning on Microsoft Azure?
   a. Azure Machine Learning is a cloud-based predictive analytics service that makes it possible to quickly build and deploy predictive models as analytic solutions.
   b. Azure Machine Learning not only provides tools to model predictive analytics, but also provides a fully-managed service you can use to deploy your predictive models as ready-to-consume web services.
   c. Azure Machine Learning provides tools for creating complete predictive analytics solutions in the cloud: Quickly create, test, operationalize, and manage predictive models.


Azure Machine Learning: Basic workflow
Build models from data and operationalize a machine learning solution
Data collection & management
Machine Learning service
Embedded ML model
Blobs & tables (Azure Storage)
Hadoop (Azure HDInsight)
Relational data (Azure SQL Database)
Massive data stores (Azure Data Lake)
ML STUDIO
ML Studio web app:
Create machine learning models
Web service
Output a web service that can be run on a scheduled basis and connected to a database.
Add intelligence to apps or websites, or provide insights in BI tools

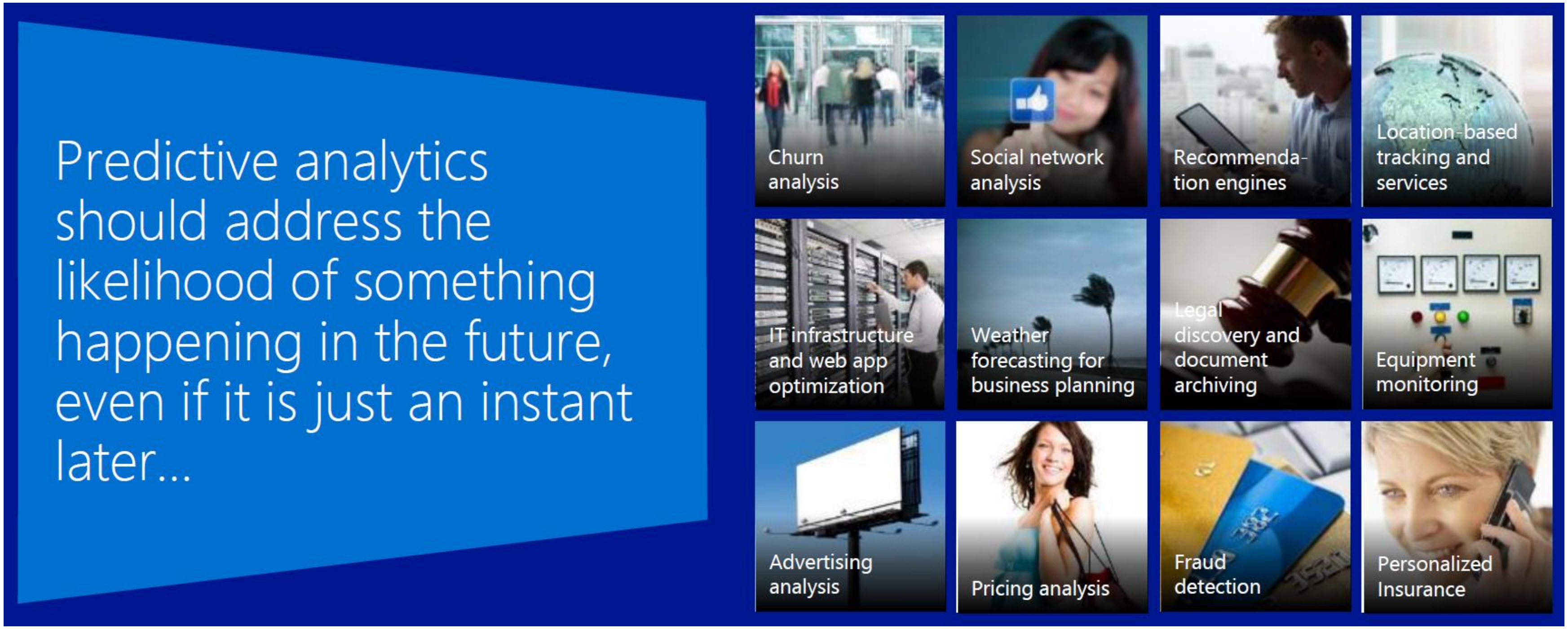
3. What is predictive analytics?

a. Predictive analytics uses various statistical techniques – in this case, machine learning – to analyze collected or current data for patterns or trends in order to forecast future events.

b. Azure Machine Learning is a particularly powerful way to do predictive analytics:

1) You can work from a ready-to-use library of algorithms, create models on an internet-connected PC without purchasing additional equipment or infrastructure, and deploy your predictive solution quickly.

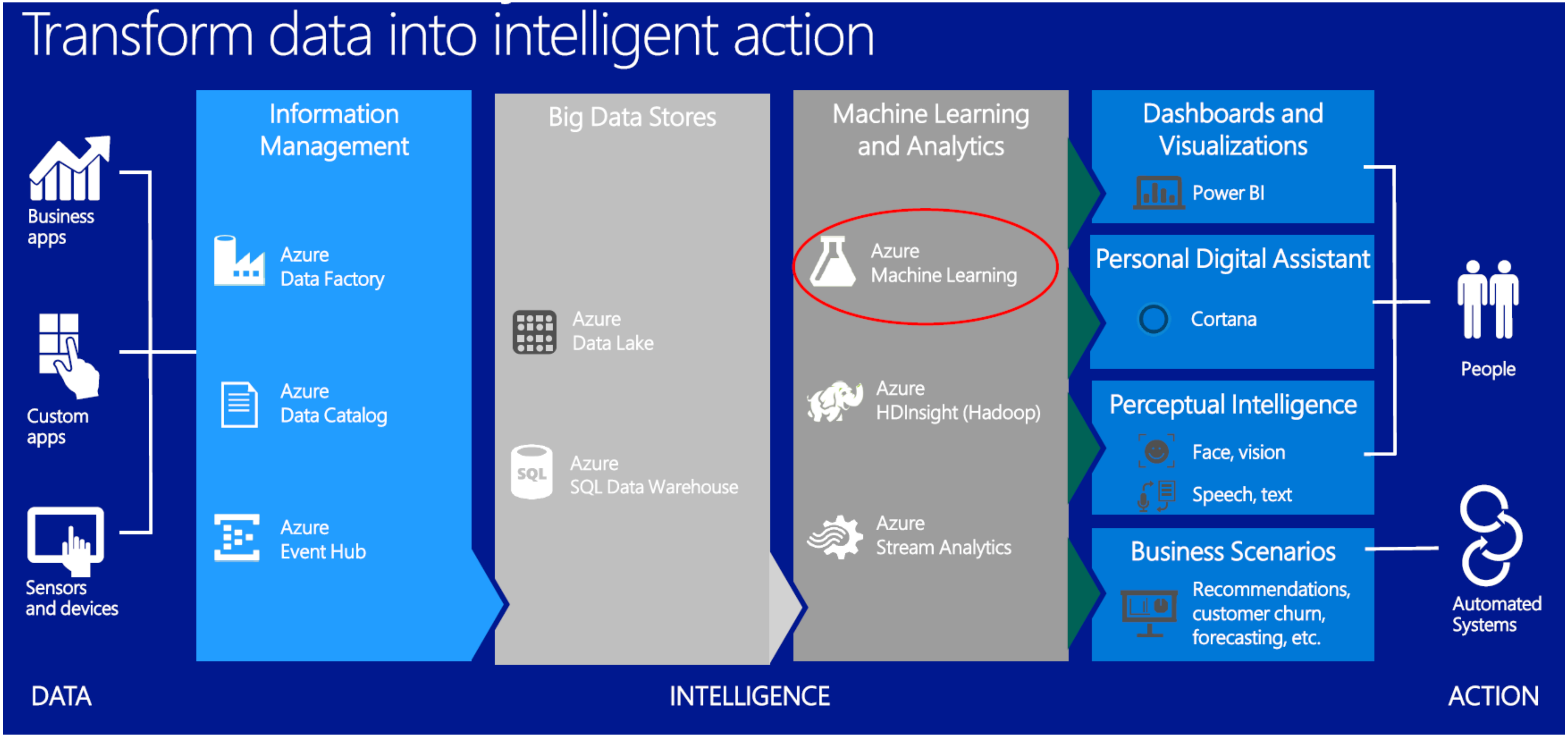
2) You can also find ready-to-use examples and solutions in the Microsoft Azure Marketplace or Cortana Intelligence Gallery.



Predictive analytics should address the likelihood of something happening in the future, even if it is just an instant later...

Churn analysis
Social network analysis
Recommenda-tion engines
Location-based tracking and services
IT infrastructure and web app optimization
Weather forecasting for business planning
Legal discovery and document archiving
Equipment monitoring
Advertising analysis
Pricing analysis
Fraud detection
Personalized Insurance

Part I Cortana Analytics Suite



Transform data into intelligent action

Information Management

Azure Data Factory

Azure Data Catalog

Azure Event Hub

Big Data Stores

Azure Data Lake

Azure SQL Data Warehouse

Machine Learning and Analytics

Azure Machine Learning

Azure HDInsight (Hadoop)

Azure Stream Analytics

Dashboards and Visualizations

Power BI

Personal Digital Assistant

Cortana

Perceptual Intelligence

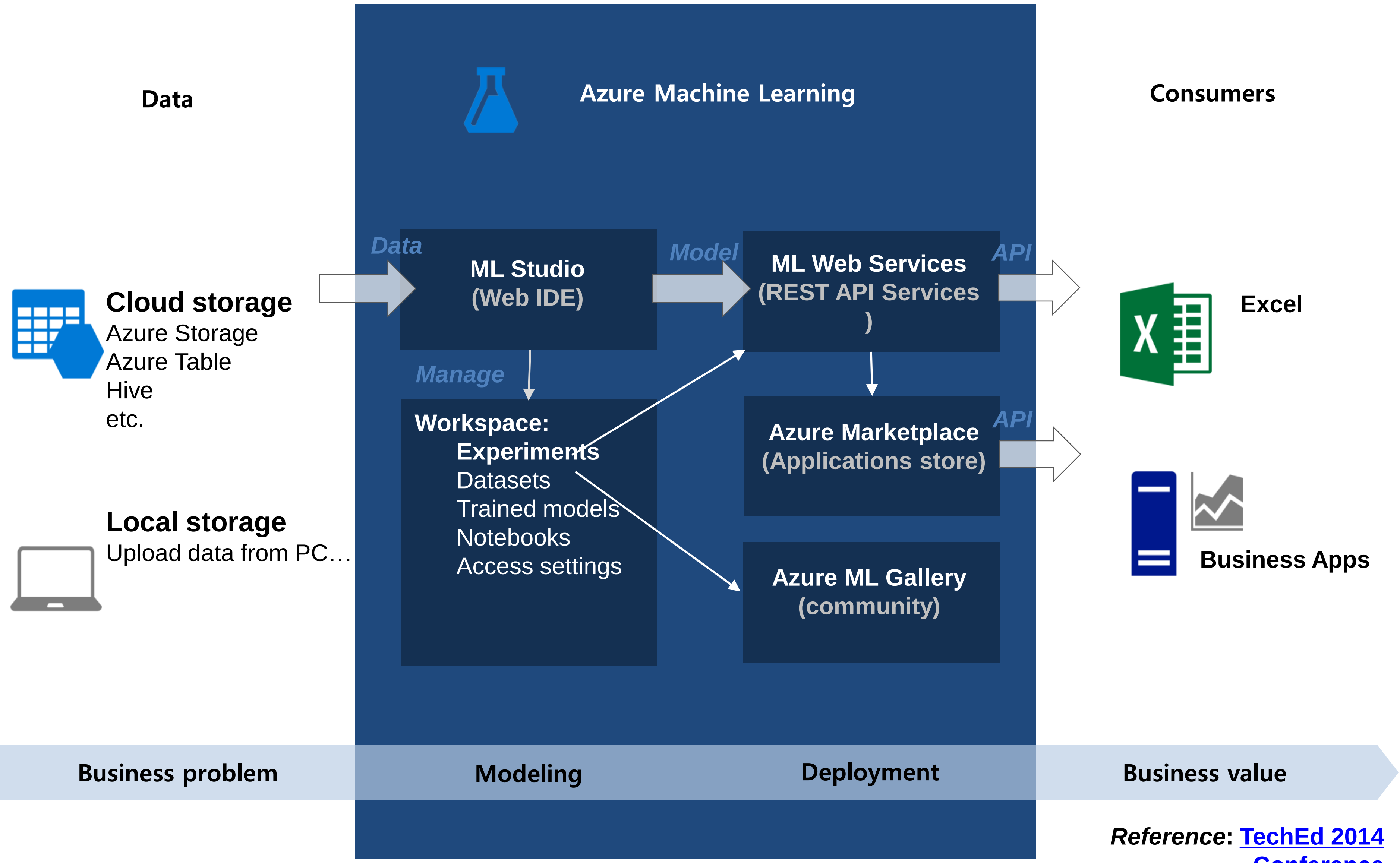
Face, vision

Speech, text

Business Scenarios

Recommendations, customer churn, forecasting, etc.

Part I Azure ML Workflow



Data

Cloud storage
Azure Storage
Azure Table
Hive
etc.

Local storage
Upload data from PC…

Azure Machine Learning

Data

ML Studio
(Web IDE)

Model

ML Web Services
(REST API Services
)

API

Manage

Workspace:
Experiments
Datasets
Trained models
Notebooks
Access settings

Azure Marketplace
(Applications store)

API

Azure ML Gallery
(community)

Consumers

Excel

Business Apps

Business problem

Modeling

Deployment

Business value

Reference: TechEd 2014 Conference

Part I Key Machine Learning Terminology and Concepts

1. Data exploration:
   - Process of gathering information about a large and often unstructured data set in order to find characteristics for focused analysis. Data mining refers to automated data exploration.

2. Descriptive analytics:
   - Process of analyzing a data set in order to summarize what happened. The vast majority of business analytics – such as sales reports, web metrics, and social network analysis – are descriptive.

3. Predictive analytics:
   - Process of building models from historical or current data in order to forecast future outcomes.

4. Supervised learning:
   - Algorithms are trained with labeled data – in other words, data comprised of examples of the answers wanted. For instance, a model that identifies fraudulent credit card use would be trained from a data set in which data points indicating known fraudulent and valid charges were labeled.

5. Unsupervised learning:
   - Is used on data with no labels, and the goal is to find relationships in the data. For instance, you might want to find groupings of customer demographics with similar buying habits.

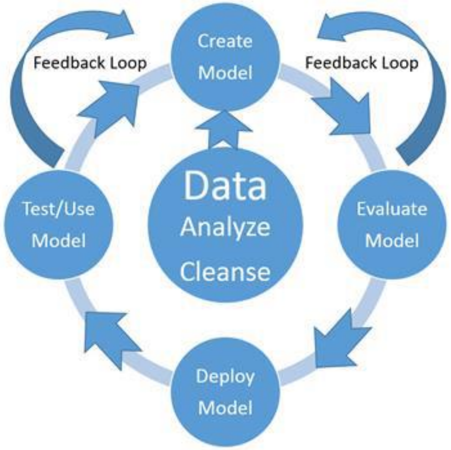
Part I Key Machine Learning Terminology and Concepts

1. Machine learning model
   - Abstraction of the question you are trying to answer or the outcome you want to predict. Models are trained and evaluated from existing data.

2. Training from data
   - In Azure Machine Learning, a model is built from an algorithm module that processes training data and functional modules, such as a scoring module.
   - In supervised learning, if you're training a fraud detection model, you'll use a set of transactions that are labeled as either fraudulent or valid. You'll split your data at randomly, and use part to train the model and part to test or evaluate the model.

3. Evaluation data
   - Once you have a trained model, evaluate the model using the remaining test data. You use data you already know the outcomes for, so that you can tell whether your model predicts accurately.

Azure Machine Learning Workflow



Part I Key Machine Learning Terminology and Concepts

1. Classification:
   - Organizing data points into categories based on a data set for which category groupings are already known.

2. Regression:
   - Predicting a continuous value based on independent variables, such as predicting the price of a car based on its year and make.

3. Clustering:
   - Partition items into homogeneous groups. Typically used to predict grouping classifications for a given variable.

4. Recommendation:
   - The Netflix contest: Build a better recommender system from Netflix data. Use the crowd's votes to complete the missing entries.

5. Anomaly Detection:
   - Forecast of a value or values from a machine learning model.

- Occam's Razor: The best models are simple models that fit the data well.

- William of Ockham, English friar, philosopher, and theologian (1287-1347) said that among hypotheses that predict equally well, we should choose the one with the fewest assumptions.

Part II Azure Machine Learning Studio

1. Projects: Sets of related Experiments, Trained Models, Datasets, Transforms.

2. Experiments: Experiments that have been created, run, and saved as drafts. These include a set of sample experiments that ship with the service to help jumpstart your projects.

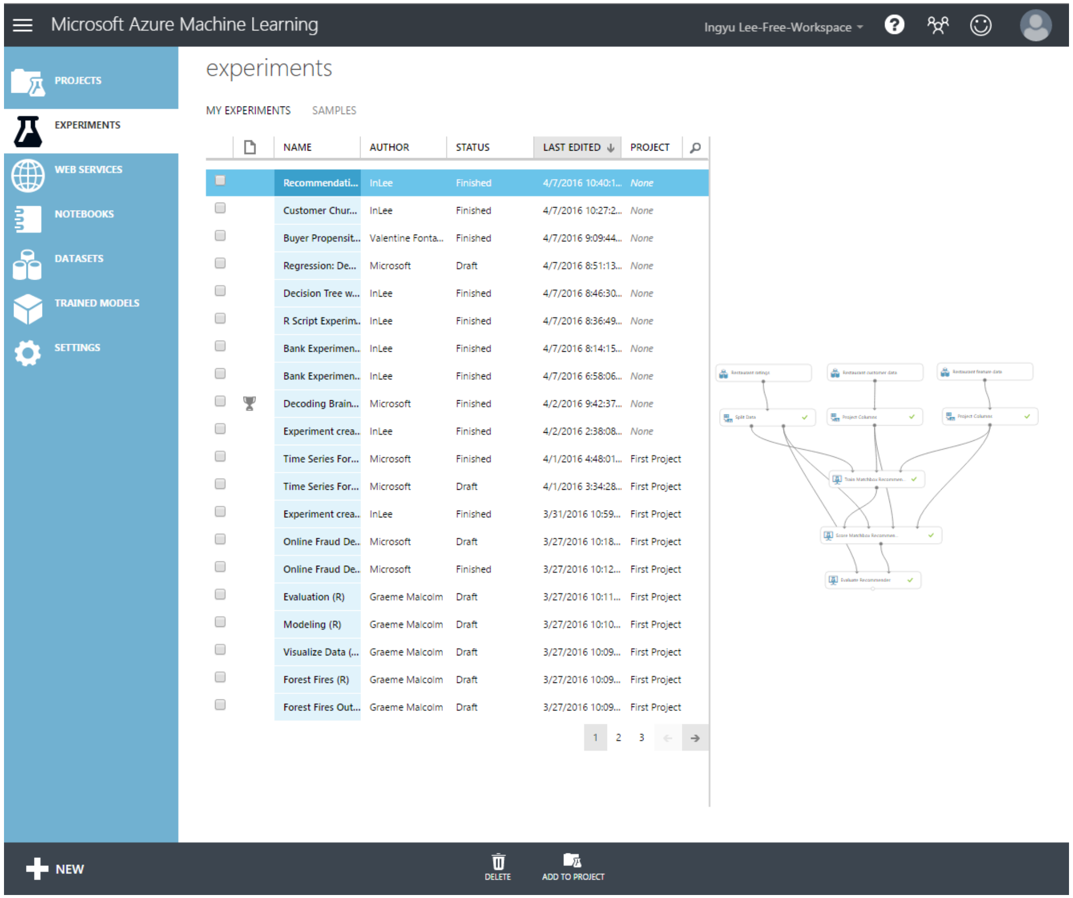
3. Web Services: A list of experiments that you have published as web services.

4. Notebooks: Jupyter notebooks that you have created.

5. Datasets: A list of sample datasets that ship with the product, and uploaded data. You can use these datasets to learn about Azure Machine Learning.

6. Trained Models: List of any trained models that you saved from your experiments.

7. Settings: Configure your account and resources. Invite other users to share your workspace in Azure Machine Learning.



experiments

		NAME	AUTHOR	STATUS	LAST EDITED ↓	PROJECT
□		Recommendati...	InLee	Finished	4/7/2016 10:40:1...	None
□		Customer Chur...	InLee	Finished	4/7/2016 10:27:2...	None
□		Buyer Propensit...	Valentine Fonta...	Finished	4/7/2016 9:09:44...	None
□		Regression: De...	Microsoft	Draft	4/7/2016 8:51:13...	None
□		Decision Tree w...	InLee	Finished	4/7/2016 8:46:30...	None
□		R Script Experim...	InLee	Finished	4/7/2016 8:36:49...	None
□		Bank Experimen...	InLee	Finished	4/7/2016 8:14:15...	None
□		Bank Experimen...	InLee	Finished	4/7/2016 6:58:06...	None
□	🏆	Decoding Brain...	Microsoft	Finished	4/2/2016 9:42:37...	None
□		Experiment crea...	InLee	Finished	4/2/2016 2:38:08...	None
□		Time Series For...	Microsoft	Finished	4/1/2016 4:48:01...	First Project
□		Time Series For...	Microsoft	Draft	4/1/2016 3:34:28...	First Project
□		Experiment crea...	InLee	Finished	3/31/2016 10:59...	First Project
□		Online Fraud De...	Microsoft	Draft	3/27/2016 10:18...	First Project
□		Online Fraud De...	Microsoft	Finished	3/27/2016 10:12...	First Project
□		Evaluation (R)	Graeme Malcolm	Draft	3/27/2016 10:11...	First Project
□		Modeling (R)	Graeme Malcolm	Draft	3/27/2016 10:10...	First Project
□		Visualize Data (...	Graeme Malcolm	Draft	3/27/2016 10:09...	First Project
□		Forest Fires (R)	Graeme Malcolm	Draft	3/27/2016 10:09...	First Project
□		Forest Fires Out...	Graeme Malcolm	Draft	3/27/2016 10:09...	First Project

1 2 3 ← →

Part II Azure Machine Learning Studio

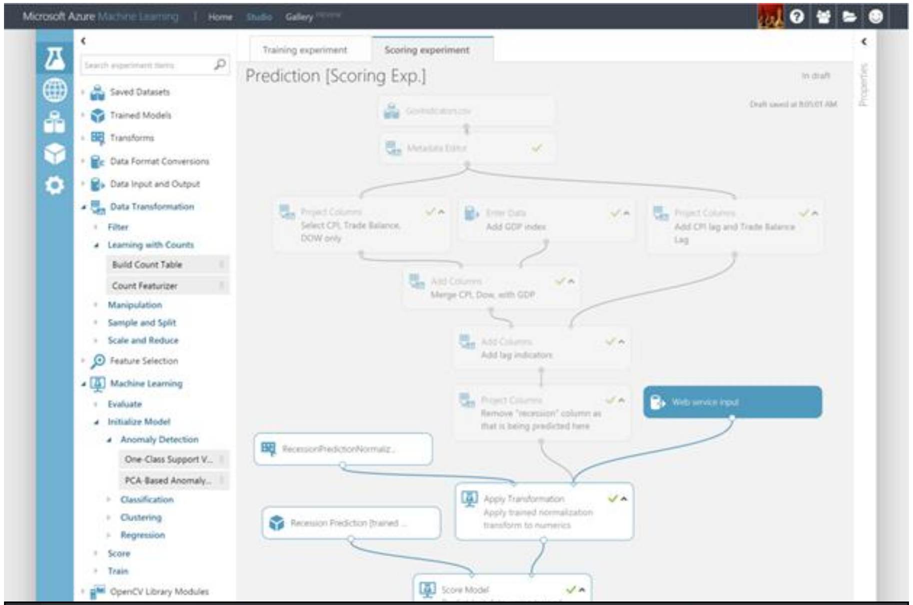
With Azure Machine Learning Studio, you can

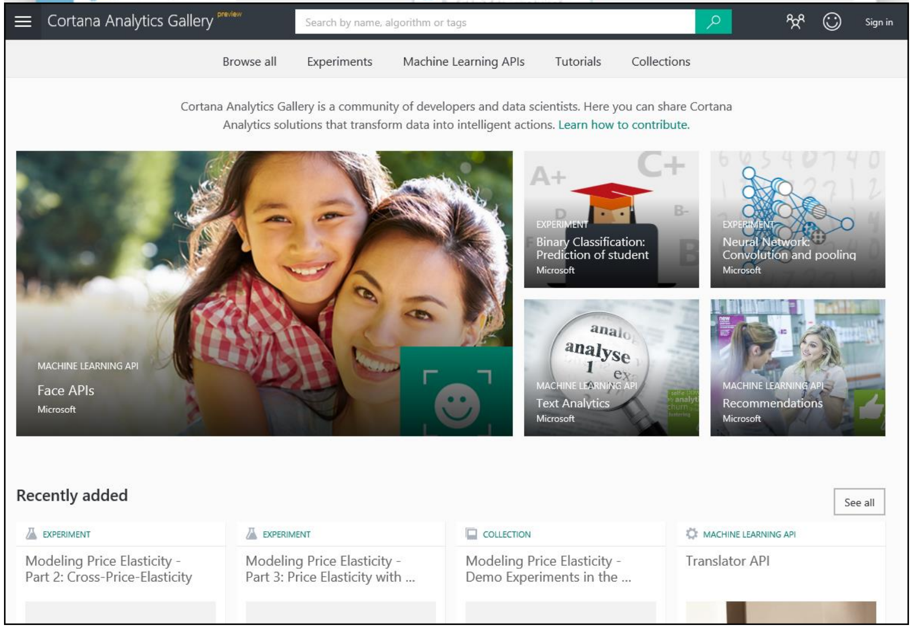
a. Create predictive models in Machine Learning Studio, a browser-based tool, by dragging, dropping, and connecting modules.

b. Use a large library of Machine Learning algorithms and modules in Machine Learning Studio to jump-start your predictive models.

c. Choose from a library of sample experiments, R and Python packages, and best-in-class algorithms from Microsoft businesses like Xbox and Bing. Extent Studio modules with your own custom R and Python scripts.

d. In Cortana Intelligence Gallery you can try analytics solutions authored by others or contribute your own using Azure services including Machine Learning, HDInsight (Hadoop), Stream Analytics, and Data Lake Analytics, as well as Azure big data stores and data management services.





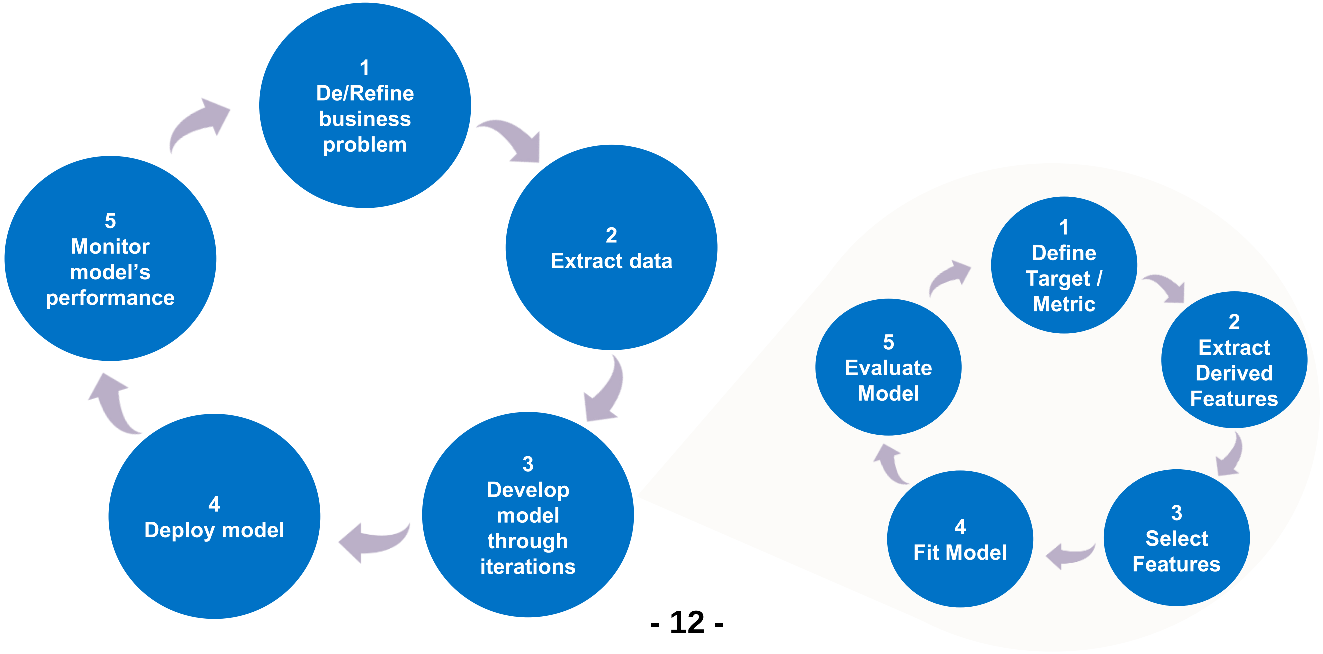
Cortana Analytics Gallery preview
Search by name, algorithm or tags
Sign in
Browse all
Experiments
Machine Learning APIs
Tutorials
Collections

Cortana Analytics Gallery is a community of developers and data scientists. Here you can share Cortana Analytics solutions that transform data into intelligent actions. Learn how to contribute.

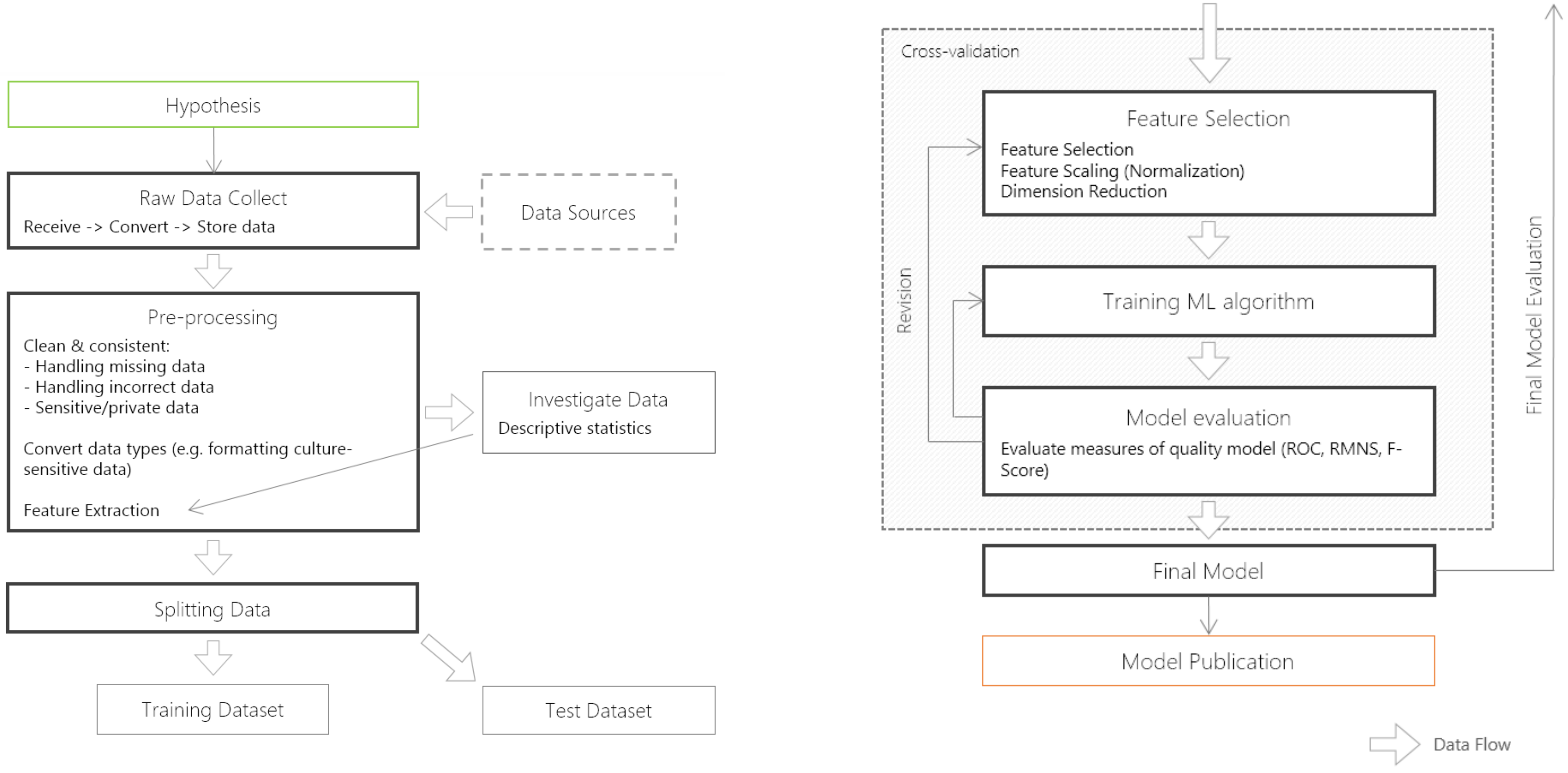
MACHINE LEARNING API
Face APIs
Microsoft
EXPERIMENT
Binary Classification: Prediction of student
Microsoft
EXPERIMENT
Neural Network: Convolution and pooling
Microsoft
MACHINE LEARNING API
Text Analytics
Microsoft
MACHINE LEARNING API
Recommendations
Microsoft
Recently added
See all
EXPERIMENT
Modeling Price Elasticity - Part 2: Cross-Price-Elasticity
EXPERIMENT
Modeling Price Elasticity - Part 3: Price Elasticity with ...
COLLECTION
Modeling Price Elasticity - Demo Experiments in the ...
MACHINE LEARNING API
Translator API

Part II First Experiment with Azure ML

1. Goal
   - Create a linear regression model that predicts the price of an automobile based on different variables such as make and technical specifications.
2. Steps
   a. Create a model
      Step 1: Get data
      Step 2: Preprocess data
      Step 3: Define features
   b. Train the model
      Step 4: Choose and apply a learning algorithm
   c. Score and test the model
      Step 5: Predict new automobile prices

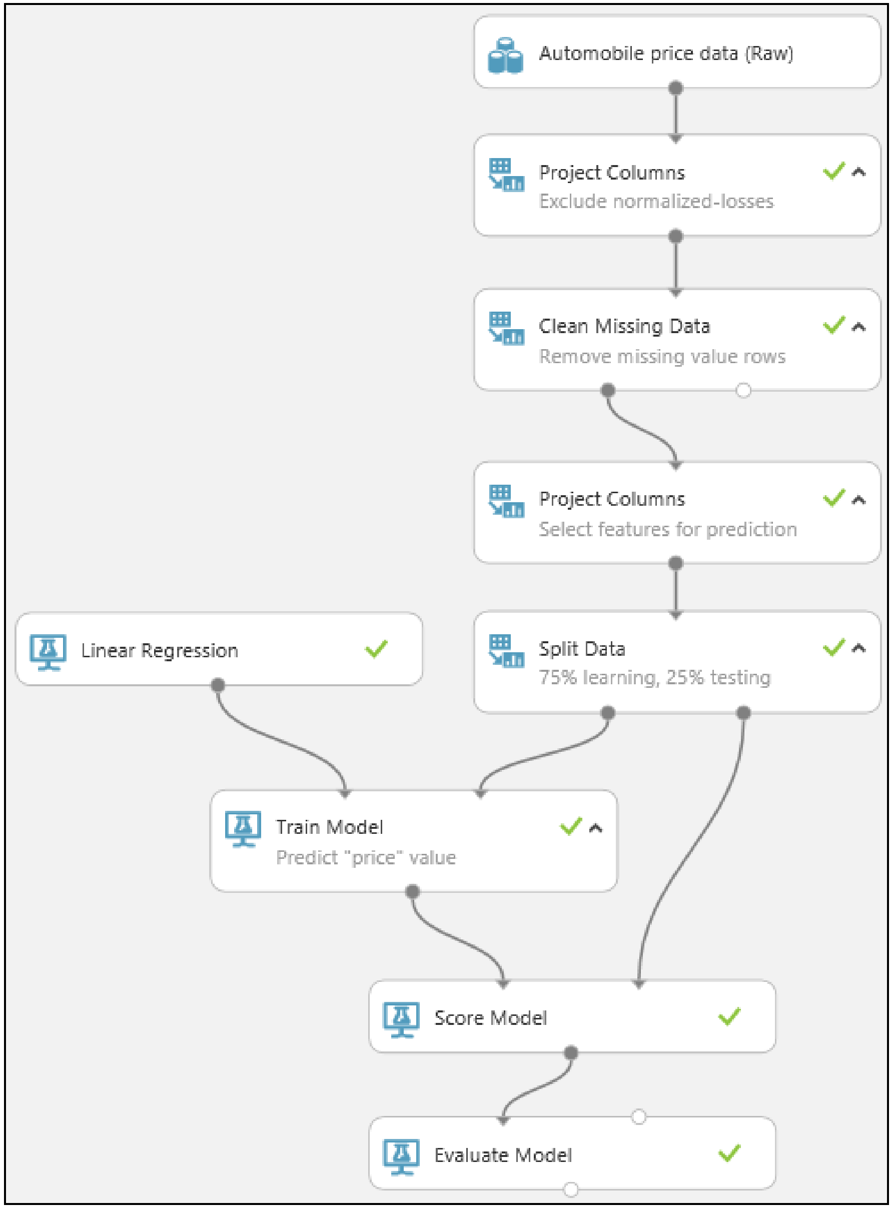


Part II First Experiment with Azure ML



Source: http://0xCode.in/azure-ml-for-data-scientist

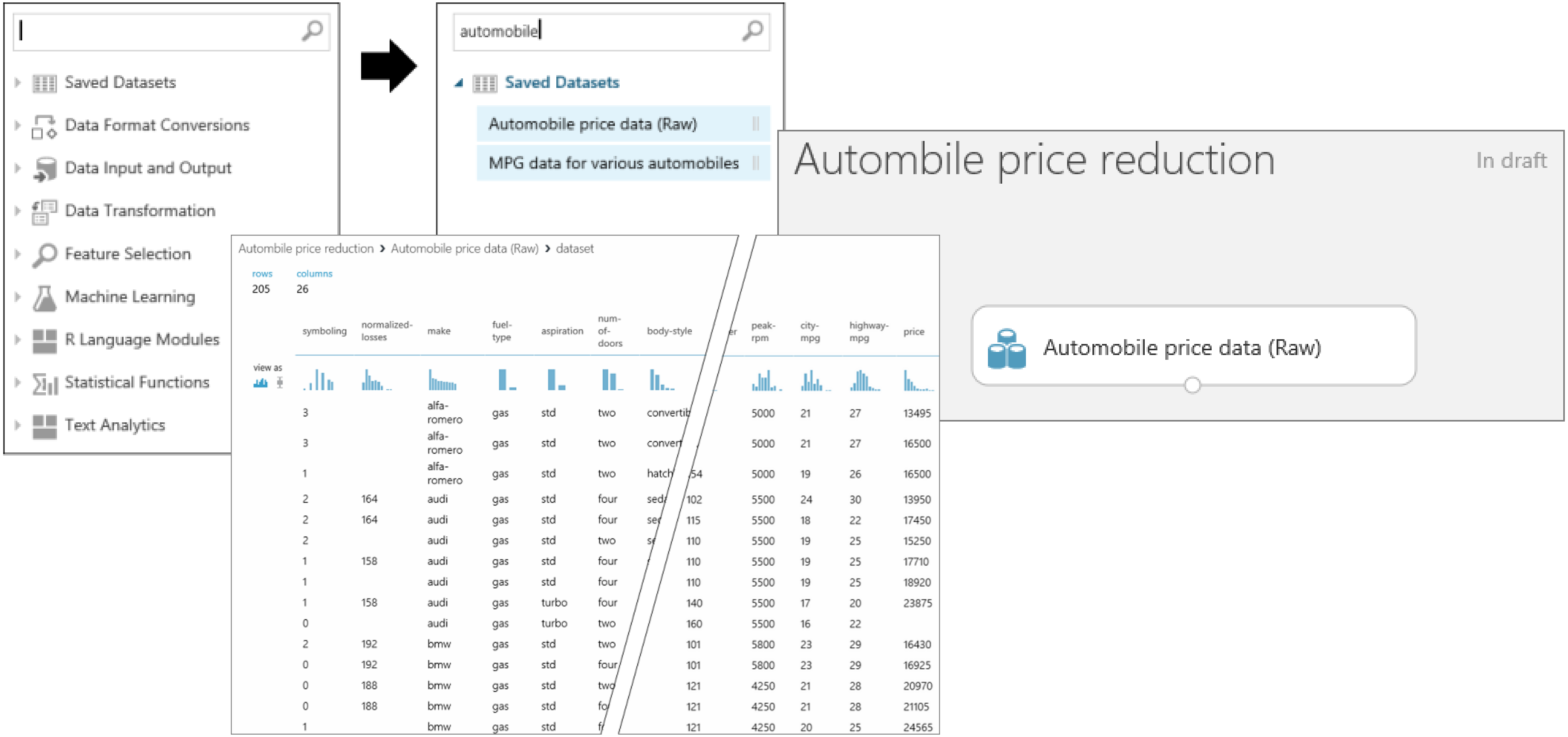
Part II First Experiment with Azure ML


Automobile price data (Raw)
Project Columns
Exclude normalized-losses
Clean Missing Data
Remove missing value rows
Project Columns
Select features for prediction
Linear Regression
Split Data
75% learning, 25% testing
Train Model
Predict "price" value
Score Model
Evaluate Model

Part II First Experiment with Azure ML

Step 1. Get Data

a. Start a new experiments.
b. Type automobile in the search box.
c. Drag the dataset to the experiment canvas.
d. Check the data with Visualize.



Autombile price reduction › Automobile price data (Raw) › dataset

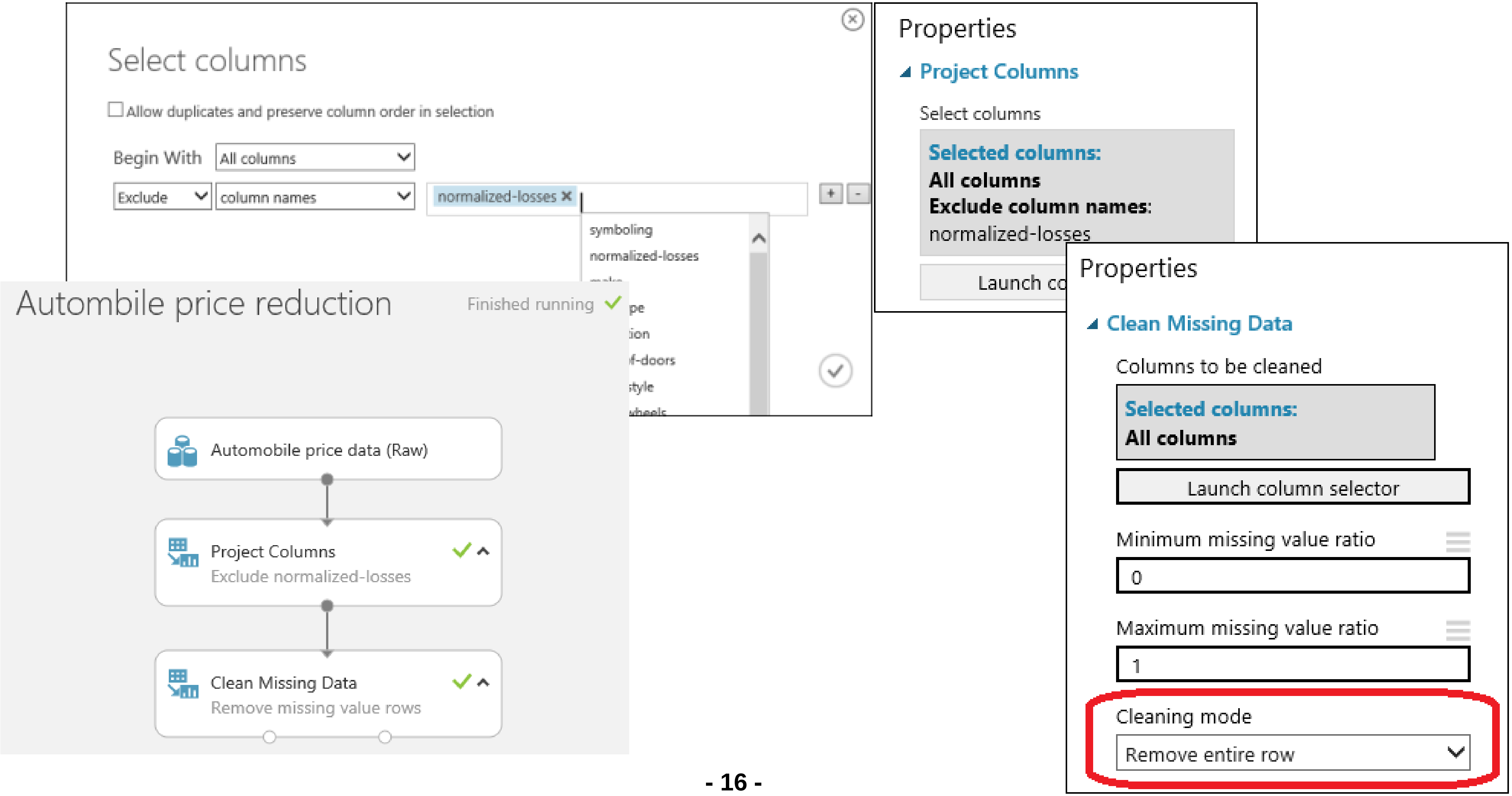
rows 205 columns 26

symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	er	peak-rpm	city-mpg	highway-mpg	price
3		alfa-romero	gas	std	two	convertib		5000	21	27	13495
3		alfa-romero	gas	std	two	convert		5000	21	27	16500
1		alfa-romero	gas	std	two	hatch	54	5000	19	26	16500
2	164	audi	gas	std	four	sed	102	5500	24	30	13950
2	164	audi	gas	std	four	se	115	5500	18	22	17450
2		audi	gas	std	two	s	110	5500	19	25	15250
1	158	audi	gas	std	four		110	5500	19	25	17710
1		audi	gas	std	four		110	5500	19	25	18920
1	158	audi	gas	turbo	four		140	5500	17	20	23875
0		audi	gas	turbo	two		160	5500	16	22	
2	192	bmw	gas	std	two		101	5800	23	29	16430
0	192	bmw	gas	std	four		101	5800	23	29	16925
0	188	bmw	gas	std	two		121	4250	21	28	20970
0	188	bmw	gas	std	fo		121	4250	21	28	21105
1		bmw	gas	std	f		121	4250	20	25	24565

Part II First Experiment with Azure ML

Step 2. Preprocess Data

a. Type project columns.
b. Click Launch column selector. Select normalized-losses to remove.
c. Drag the Clean Missing Data. Select remove for missing row.



Select columns

□ Allow duplicates and preserve column order in selection

Begin With: All columns

Exclude | column names | normalized-losses ✕

Properties

Project Columns

Select columns

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

Autombile price reduction

Finished running ✓

Project Columns ✓
Exclude normalized-losses

Clean Missing Data ✓
Remove missing value rows

Properties

Clean Missing Data

Columns to be cleaned

Selected columns:
All columns

Minimum missing value ratio: 0

Maximum missing value ratio: 1

Cleaning mode: Remove entire row

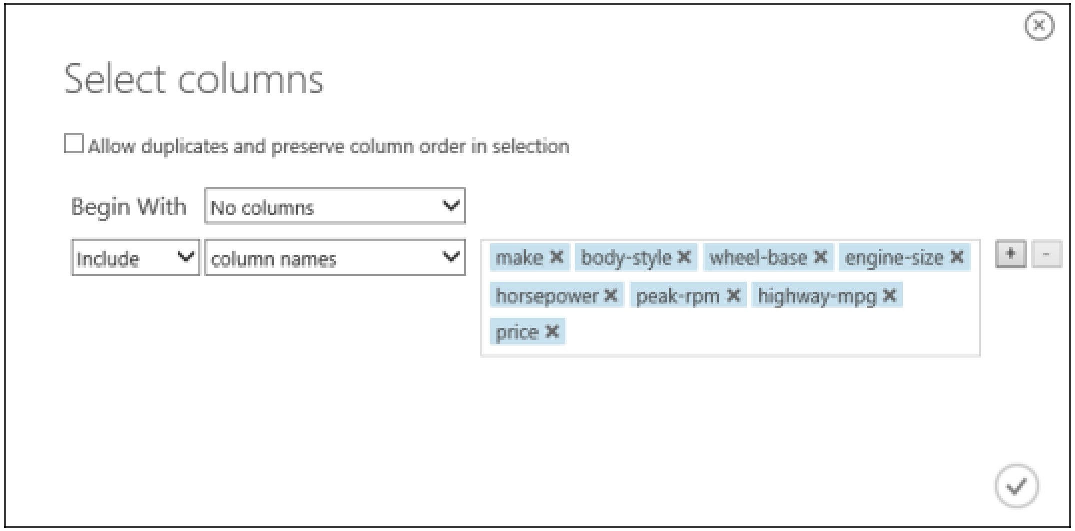
Part II First Experiment with Azure ML

Step 3. Define features

a. Drag Project Columns.

b. Launch column selector.

c. Select no columns for Begin with and then select include and column names.



Select columns

□ Allow duplicates and preserve column order in selection

Begin With: No columns

Include | column names | make ✖ body-style ✖ wheel-base ✖ engine-size ✖ horsepower ✖ peak-rpm ✖ highway-mpg ✖ price ✖

Part II First Experiment with Azure ML

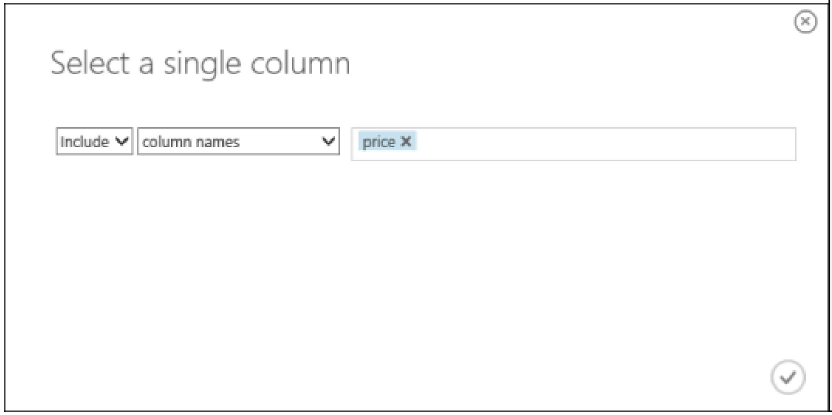
Step 4. Choose and apply a learning algorithm

a. Split data (80:20).

b. Select learning algorithm (Linear Regression).

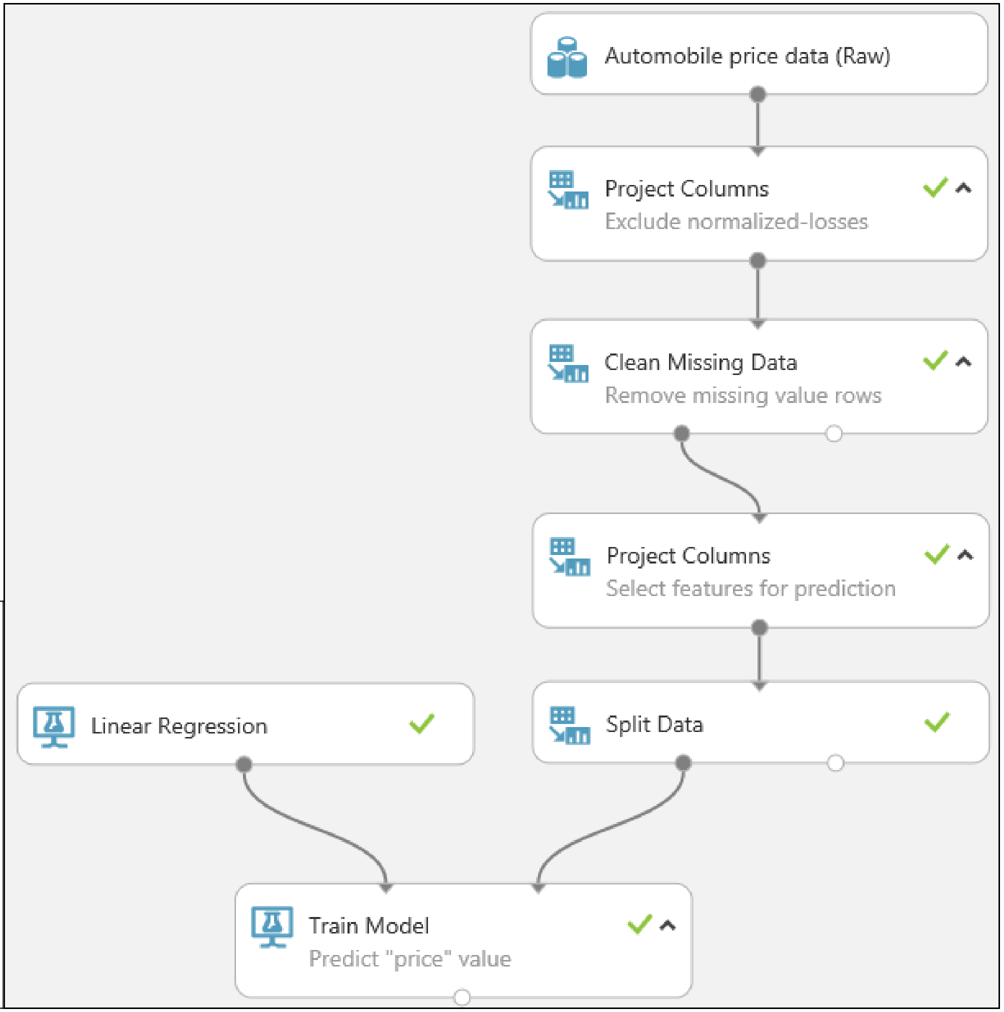
c. Select Train Model with price column.

d. Run the experiment.



Select a single column

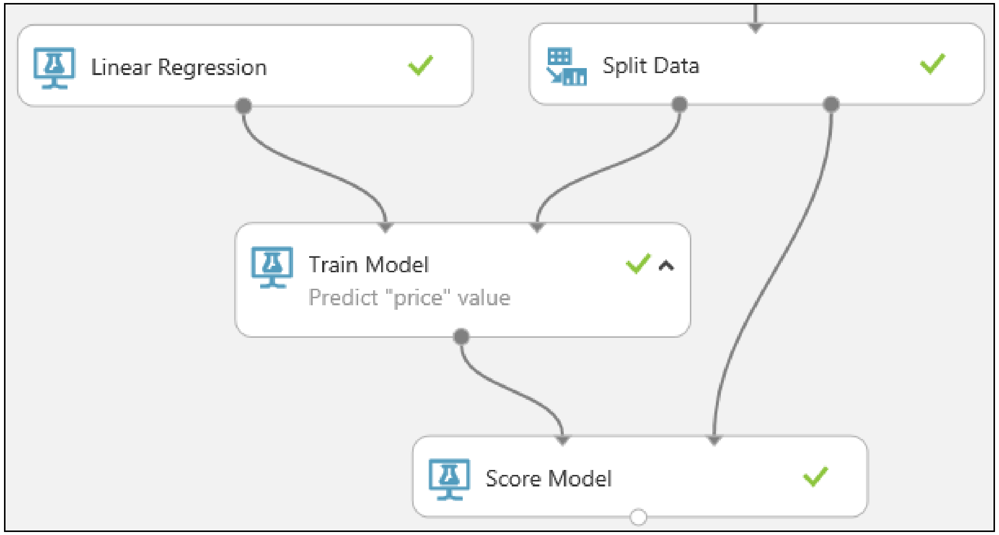
Include | column names | price ✕

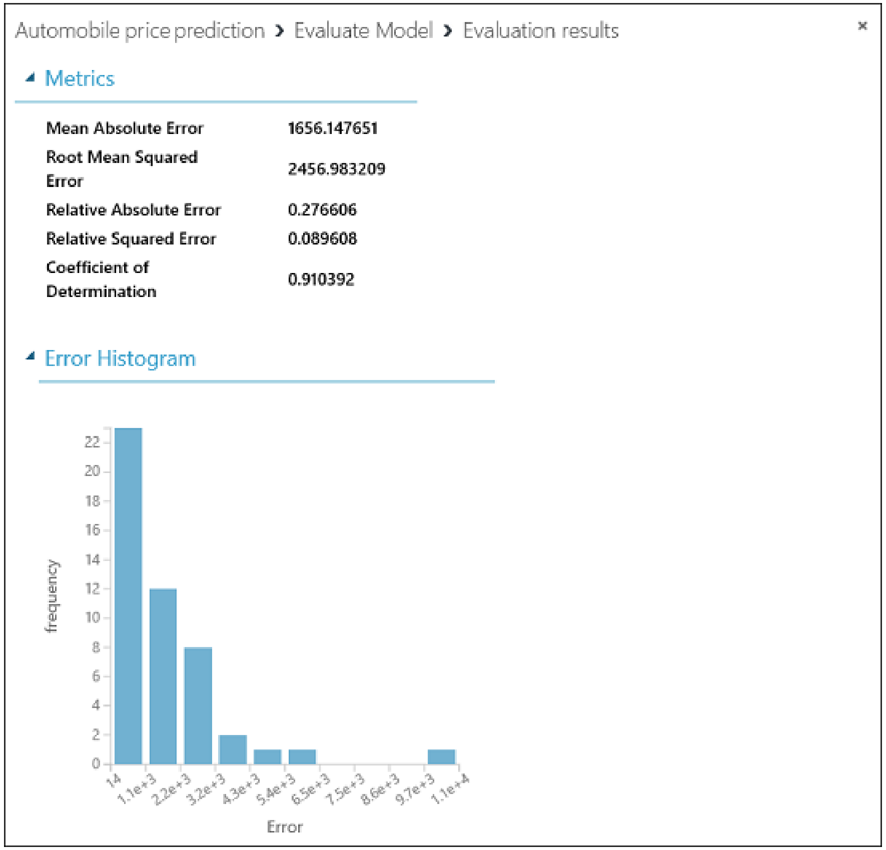

Automobile price data (Raw)
Project Columns
Exclude normalized-losses
Clean Missing Data
Remove missing value rows
Project Columns
Select features for prediction
Linear Regression
Split Data
Train Model
Predict "price" value

Part II First Experiment with Azure ML

Step 5. Predict new automobile prices

a. Find and drag Score Model.
b. Run the experiment.
c. Visualize Score Model.
d. Evaluate Model.



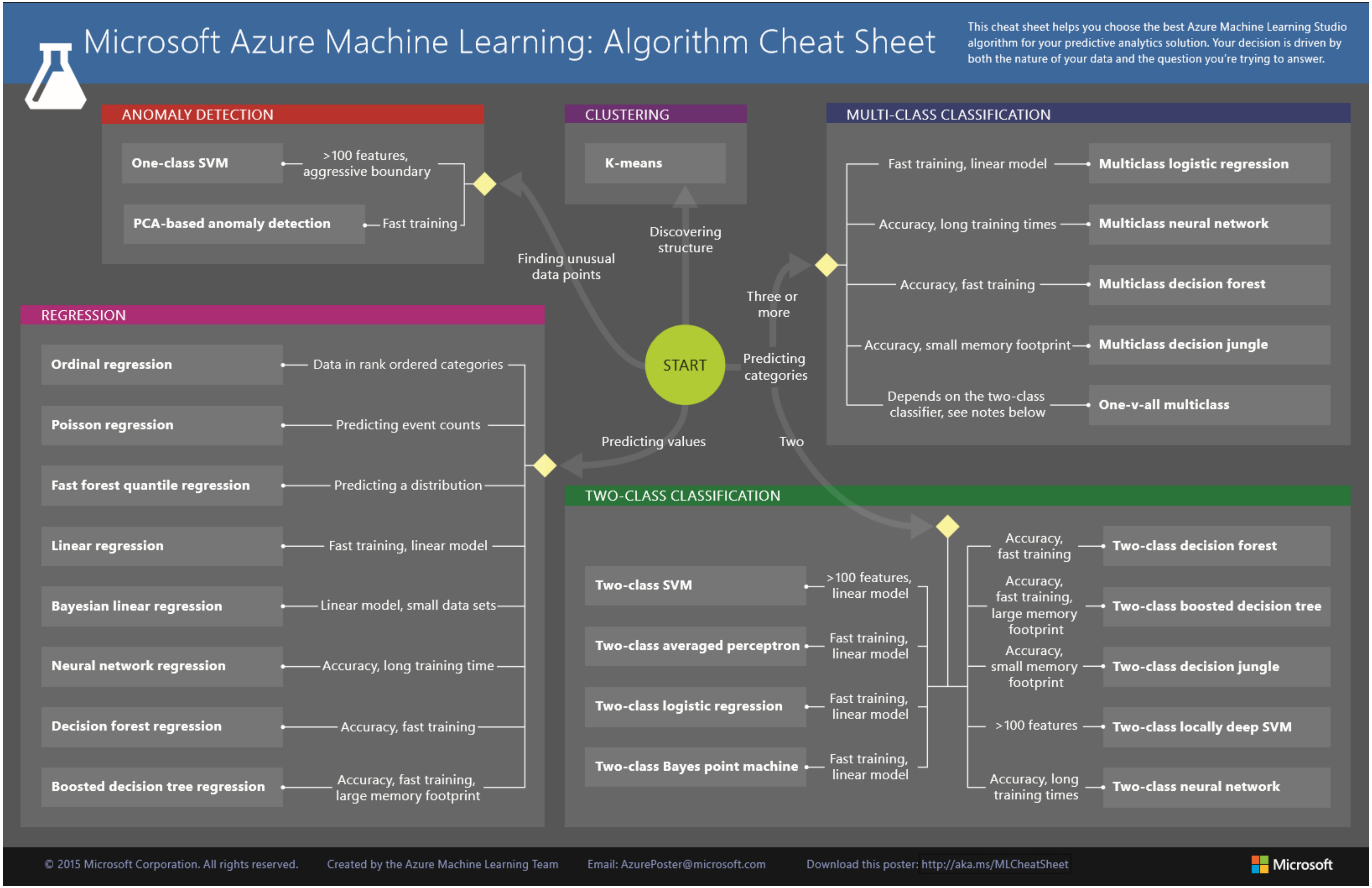

Automobile price prediction › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	1656.147651
Root Mean Squared Error	2456.983209
Relative Absolute Error	0.276606
Relative Squared Error	0.089608
Coefficient of Determination	0.910392

Error Histogram

Part II Azure ML Algorithm Cheat Sheet



Microsoft Azure Machine Learning: Algorithm Cheat Sheet

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

START

Finding unusual data points

Discovering structure

Predicting categories

Three or more

Two

Predicting values

ANOMALY DETECTION

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

CLUSTERING

- K-means

MULTI-CLASS CLASSIFICATION

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

REGRESSION

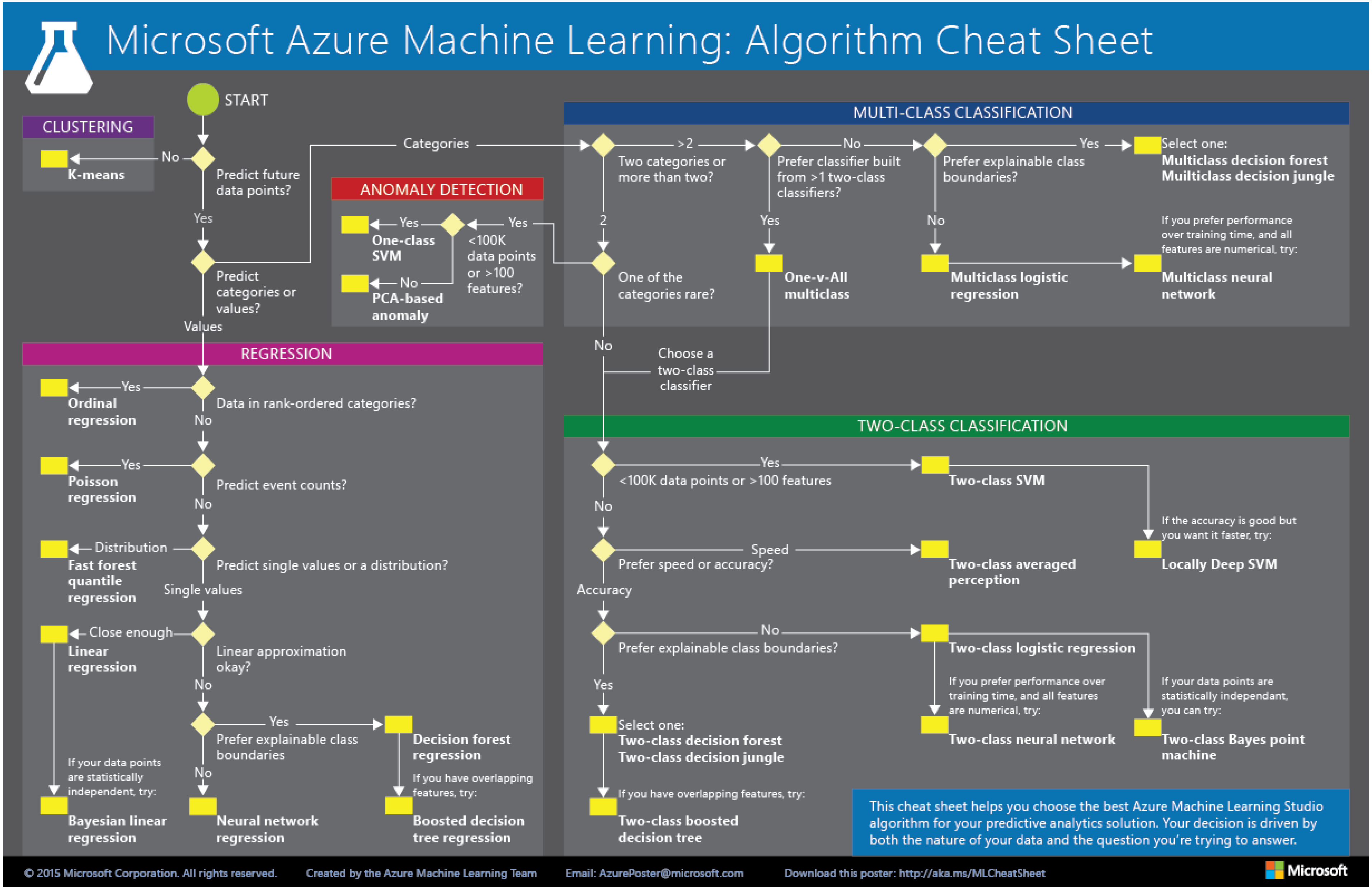
- Ordinal regression — Data in rank ordered categories
- Poisson regression — Predicting event counts
- Fast forest quantile regression — Predicting a distribution
- Linear regression — Fast training, linear model
- Bayesian linear regression — Linear model, small data sets
- Neural network regression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

TWO-CLASS CLASSIFICATION

- Two-class SVM — >100 features, linear model
- Two-class averaged perceptron — Fast training, linear model
- Two-class logistic regression — Fast training, linear model
- Two-class Bayes point machine — Fast training, linear model
- Accuracy, fast training — Two-class decision forest
- Accuracy, fast training, large memory footprint — Two-class boosted decision tree
- Accuracy, small memory footprint — Two-class decision jungle
- >100 features — Two-class locally deep SVM
- Accuracy, long training times — Two-class neural network

© 2015 Microsoft Corporation. All rights reserved. Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLCheatSheet Microsoft

Part II Azure ML Algorithm Cheat Sheet



Microsoft Azure Machine Learning: Algorithm Cheat Sheet

START

Predict future data points?
No → CLUSTERING: K-means
Yes ↓

Predict categories or values?
Categories → Two categories or more than two?
Values ↓

ANOMALY DETECTION

<100K data points or >100 features?
Yes → One-class SVM
No → PCA-based anomaly

REGRESSION

Data in rank-ordered categories?
Yes → Ordinal regression
No ↓

Predict event counts?
Yes → Poisson regression
No ↓

Predict single values or a distribution?
Distribution → Fast forest quantile regression
Single values ↓

Linear approximation okay?
Close enough → Linear regression
If your data points are statistically independent, try: Bayesian linear regression
No ↓

Prefer explainable class boundaries
Yes → Decision forest regression
If you have overlapping features, try: Boosted decision tree regression
No → Neural network regression

MULTI-CLASS CLASSIFICATION

Two categories or more than two?
>2 → Prefer classifier built from >1 two-class classifiers?
2 → One of the categories rare?

Prefer classifier built from >1 two-class classifiers?
Yes → One-v-All multiclass
No → Prefer explainable class boundaries?

Prefer explainable class boundaries?
Yes → Select one: Multiclass decision forest, Muilticlass decision jungle
No → Multiclass logistic regression
If you prefer performance over training time, and all features are numerical, try: Multiclass neural network

One of the categories rare?
Yes → ANOMALY DETECTION
No ↓

Choose a two-class classifier

TWO-CLASS CLASSIFICATION

<100K data points or >100 features
Yes → Two-class SVM
If the accuracy is good but you want it faster, try: Locally Deep SVM
No ↓

Prefer speed or accuracy?
Speed → Two-class averaged perception
Accuracy ↓

Prefer explainable class boundaries?
No → Two-class logistic regression
If you prefer performance over training time, and all features are numerical, try: Two-class neural network
If your data points are statistically independant, you can try: Two-class Bayes point machine
Yes → Select one: Two-class decision forest, Two-class decision jungle
If you have overlapping features, try: Two-class boosted decision tree

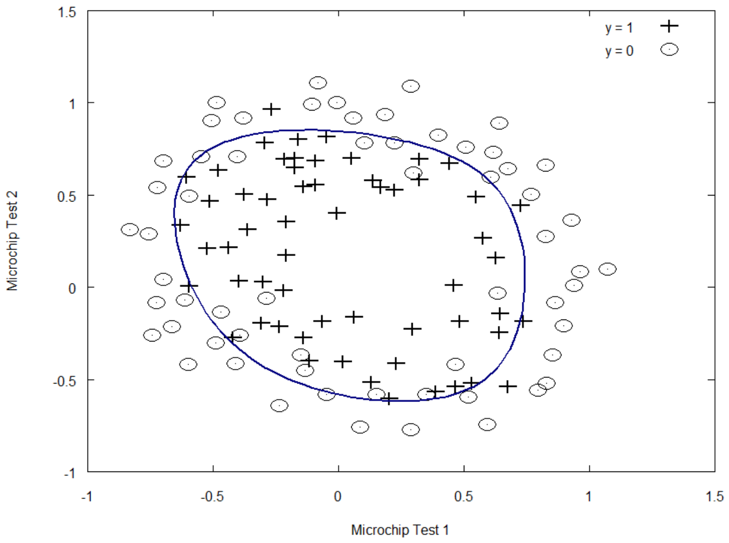
This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

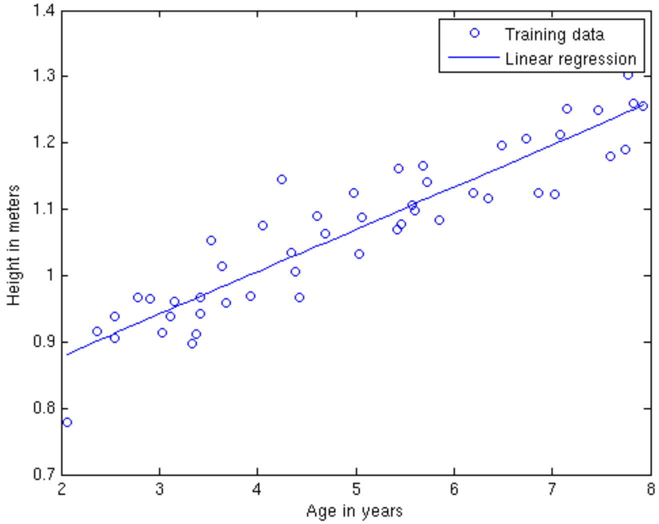
Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Part III Azure ML Hands-On: Classification

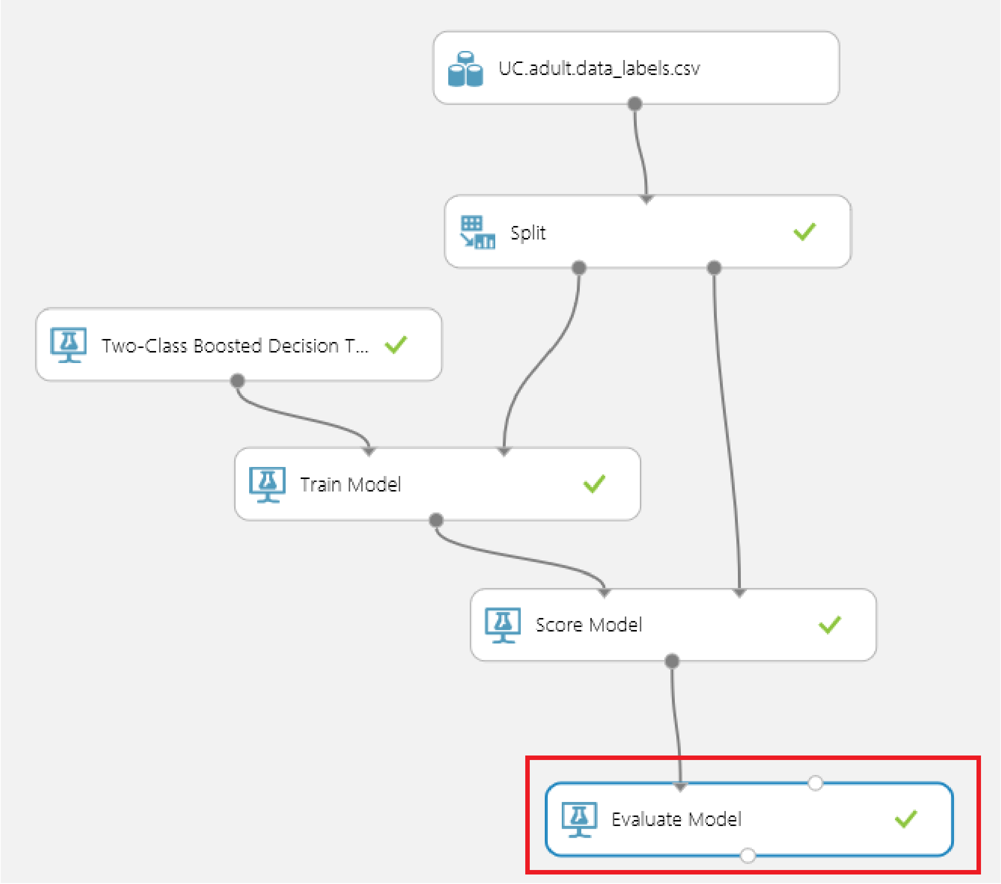
1. Supervised learning
   a. Classification: used for predicting responses that can have just a few known values, such as “married,” “single,” or “divorced,” based on the other columns in the dataset.
   b. Regression: predict one or more continuous variables, such as profit or loss, based on other columns in the dataset.
      a. Simple linear regression: A single key variable is predicted based on one or more other variables in the dataset.
      b. Multiple linear regression: More than one key variable is predicted based on one or more other variables in the dataset.
      c. Multivariate linear regression: multiple correlated dependent key variables are predicted, rather than a single key variable.





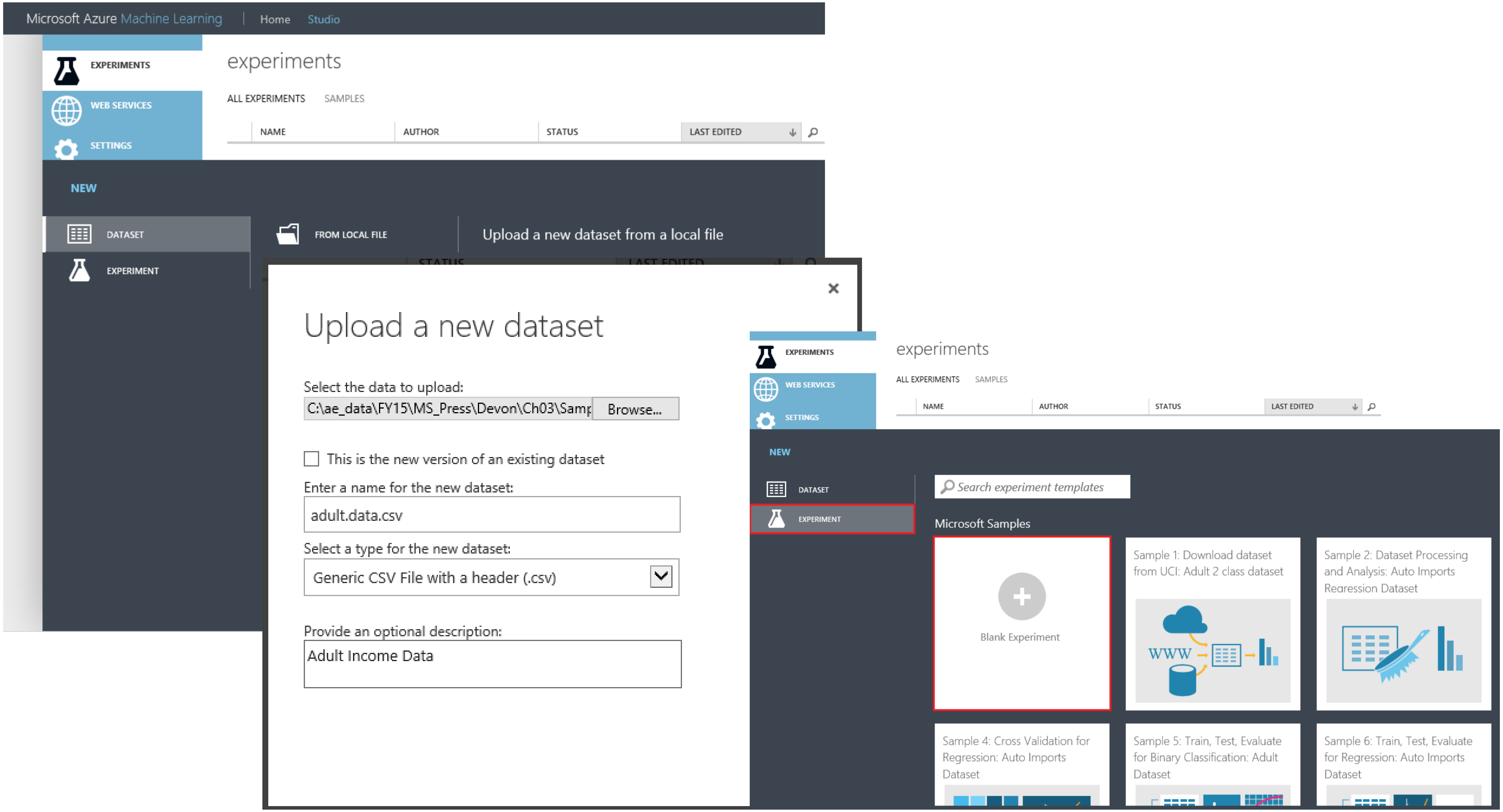
Part III Azure ML Hands-On: Classification

1. **Goal:** To predict whether a person's income exceeds $50,000 per year based on his demographics or census data.
2. **Features:**
   - Age, Workclass, Fnlwgt, Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-loss, Hours-per-week, Native-country, Income.



Part III Azure ML Hands-On: Classification

Step 1: Get the data


Microsoft Azure Machine Learning | Home Studio
EXPERIMENTS
WEB SERVICES
SETTINGS
experiments
ALL EXPERIMENTS
SAMPLES
NAME
AUTHOR
STATUS
LAST EDITED
NEW
DATASET
EXPERIMENT
FROM LOCAL FILE
Upload a new dataset from a local file
×
Search experiment templates
Microsoft Samples
Blank Experiment
Sample 1: Download dataset from UCI: Adult 2 class dataset
WWW
Sample 2: Dataset Processing and Analysis: Auto Imports Regression Dataset
Sample 4: Cross Validation for Regression: Auto Imports Dataset
Sample 5: Train, Test, Evaluate for Binary Classification: Adult Dataset
Sample 6: Train, Test, Evaluate for Regression: Auto Imports Dataset

Upload a new dataset

Select the data to upload:

C:\ae_data\FY15\MS_Press\Devon\Ch03\Samp Browse...

☐ This is the new version of an existing dataset

Enter a name for the new dataset:

adult.data.csv

Select a type for the new dataset:

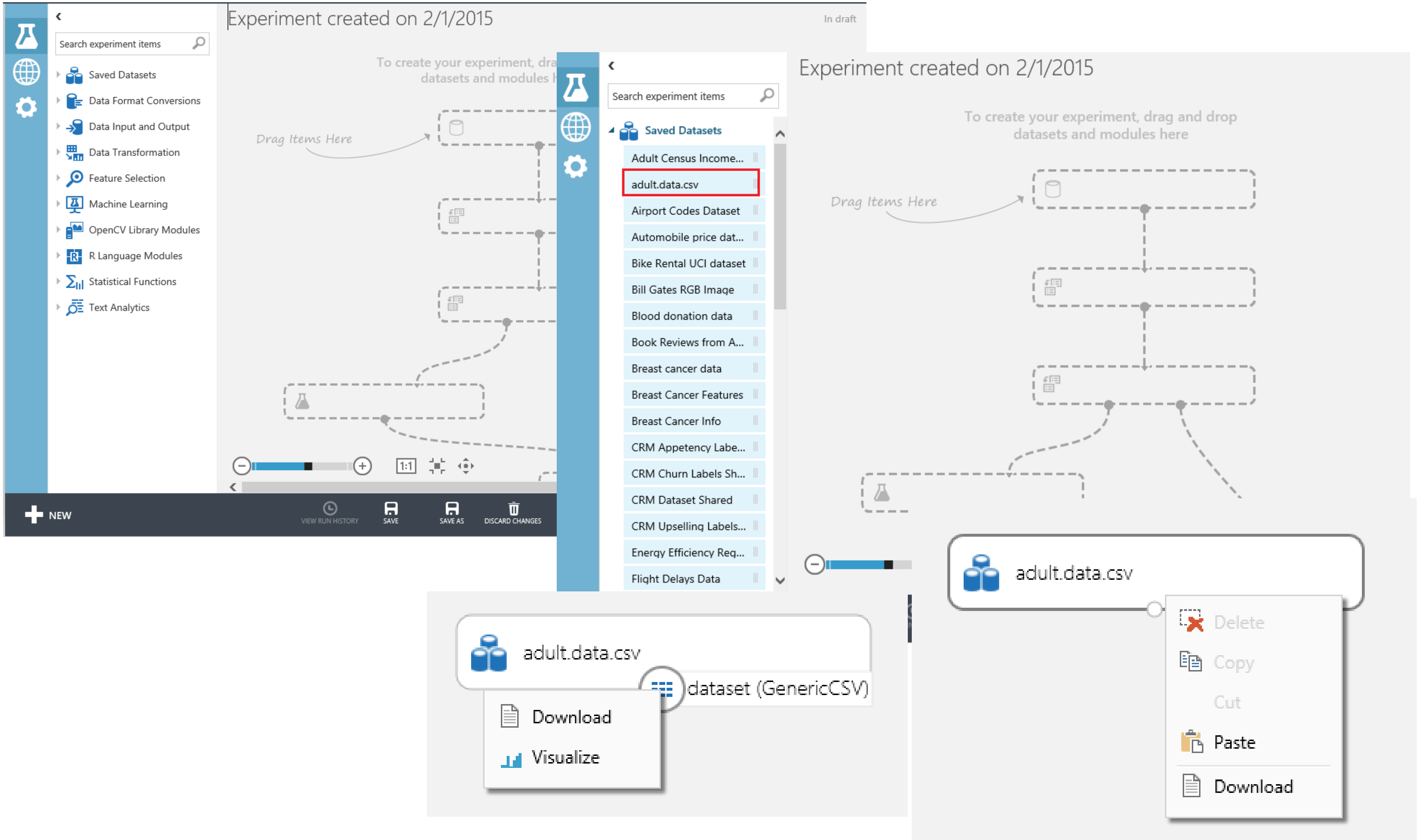
Generic CSV File with a header (.csv)

Provide an optional description:

Adult Income Data

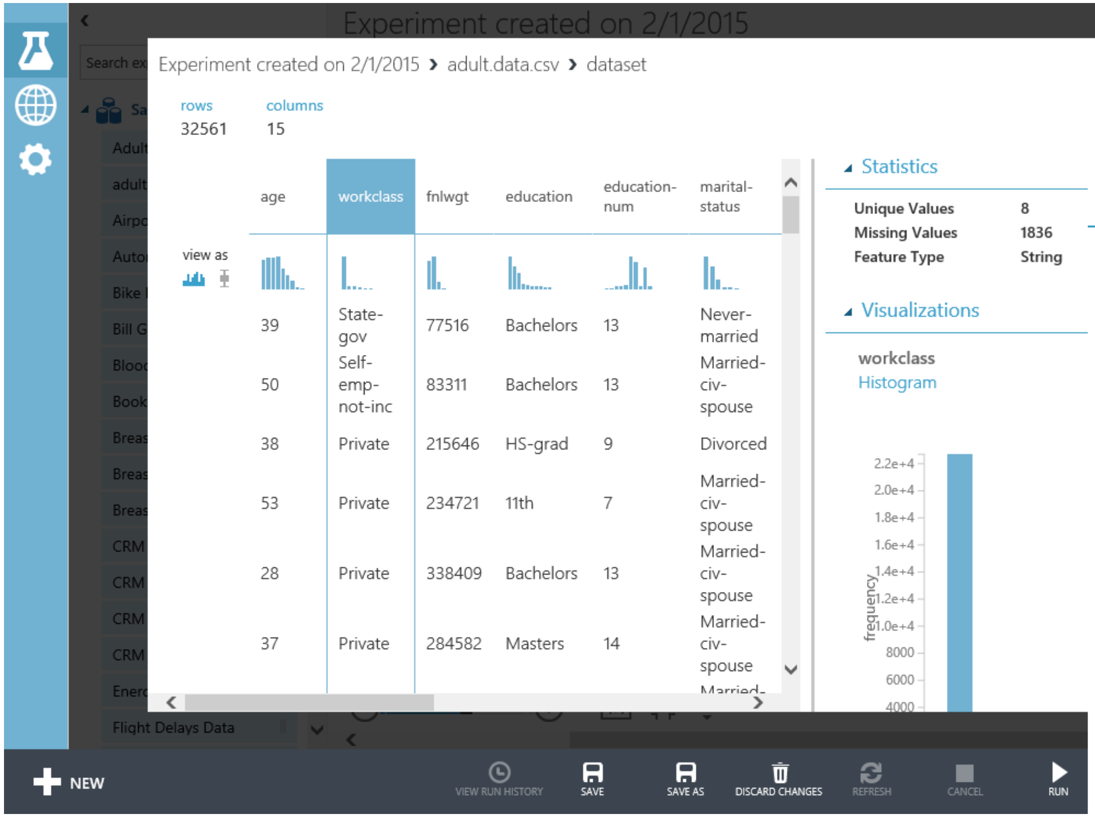
Part III Azure ML Hands-On: Classification

Step 2: Visualize the data



Part III Azure ML Hands-On: Classification

Step 2: Visualize the data



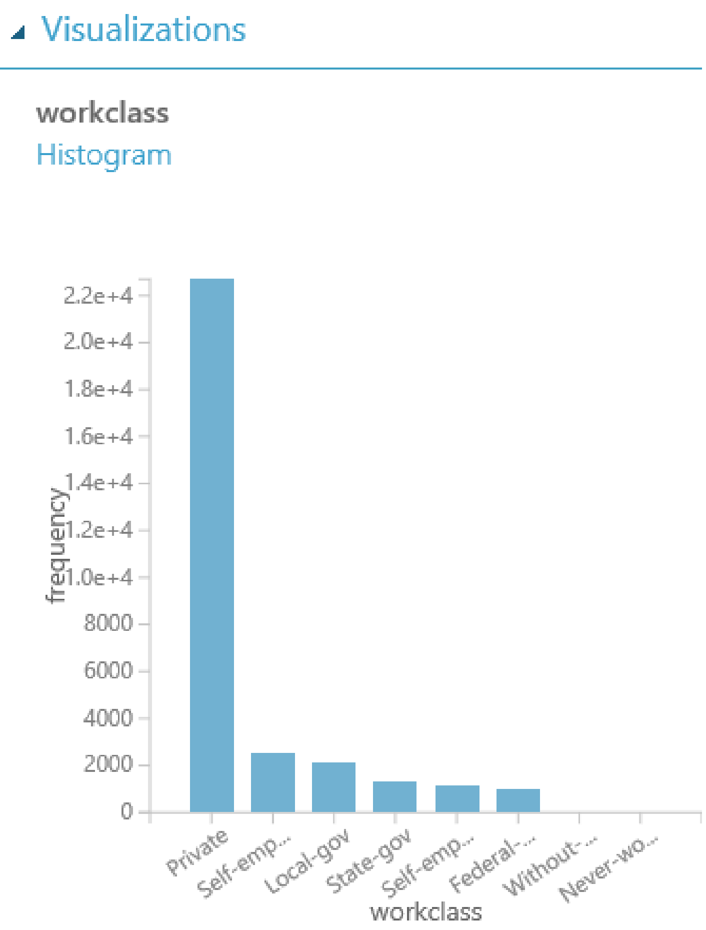
Experiment created on 2/1/2015 › adult.data.csv › dataset

rows	columns
32561	15

age	workclass	fnlwgt	education	education-num	marital-status
39	State-gov	77516	Bachelors	13	Never-married
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse
38	Private	215646	HS-grad	9	Divorced
53	Private	234721	11th	7	Married-civ-spouse
28	Private	338409	Bachelors	13	Married-civ-spouse
37	Private	284582	Masters	14	Married-civ-spouse

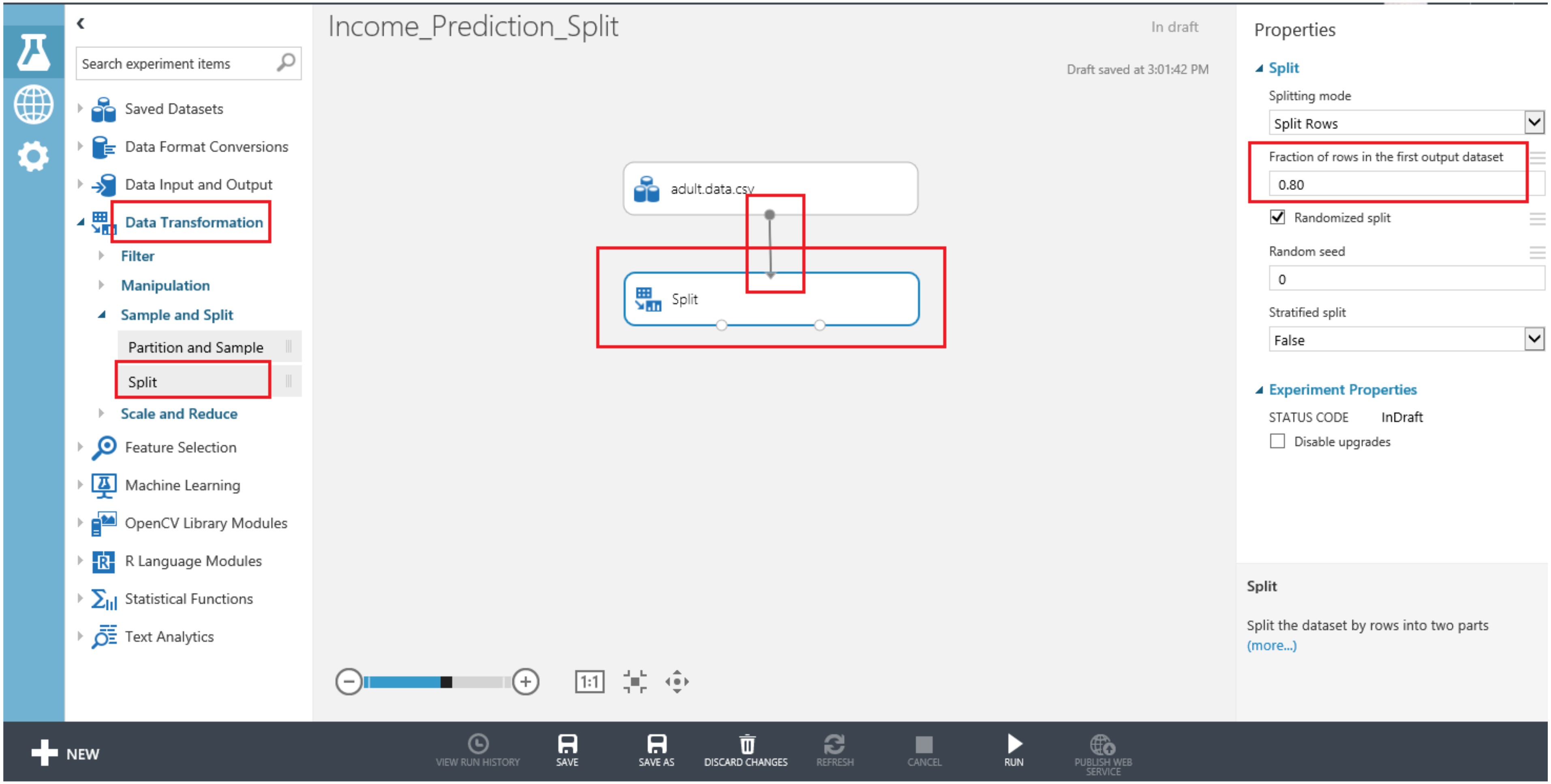
Statistics

Unique Values	8
Missing Values	1836
Feature Type	String



Part III Azure ML Hands-On: Classification

Step 3: Split the data



Income_Prediction_Split

In draft

Draft saved at 3:01:42 PM

Properties

Split

Splitting mode: Split Rows

Fraction of rows in the first output dataset: 0.80

☑ Randomized split

Random seed: 0

Stratified split: False

Experiment Properties

STATUS CODE: InDraft

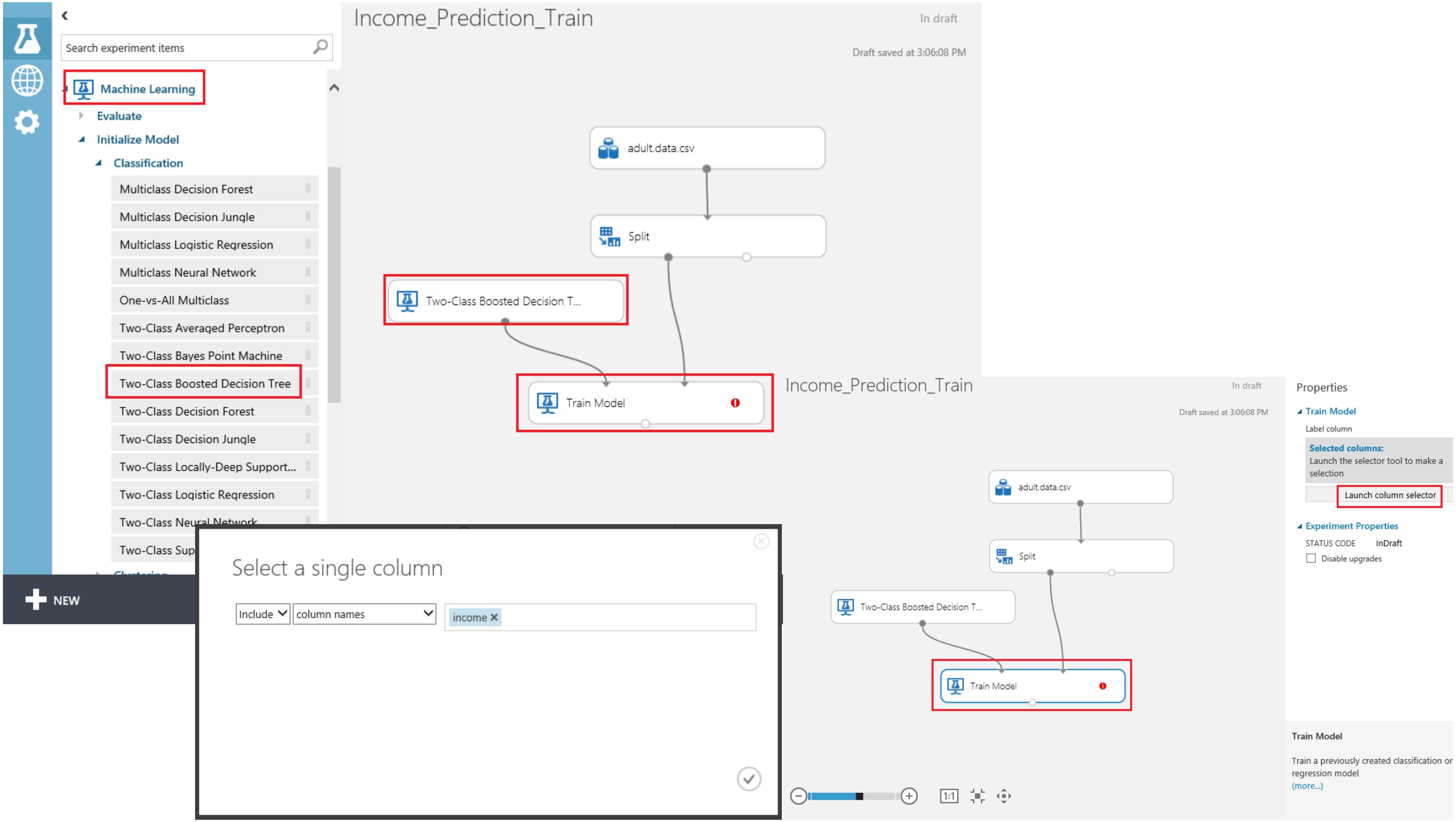
□ Disable upgrades

Split

Split the dataset by rows into two parts (more...)

Part III Azure ML Hands-On: Classification

Step 4: Train Model. Choose Two-Class Boosted Decision Tree. The income will be dependent.


Search experiment items
Machine Learning
Evaluate
Initialize Model
Classification
Multiclass Decision Forest
Multiclass Decision Jungle
Multiclass Logistic Regression
Multiclass Neural Network
One-vs-All Multiclass
Two-Class Averaged Perceptron
Two-Class Bayes Point Machine
Two-Class Boosted Decision Tree
Two-Class Decision Forest
Two-Class Decision Jungle
Two-Class Locally-Deep Support...
Two-Class Logistic Regression
Two-Class Neural Network
Two-Class Sup
NEW
Income_Prediction_Train
In draft
Draft saved at 3:06:08 PM
adult.data.csv
Split
Two-Class Boosted Decision T...
Train Model
Income_Prediction_Train
In draft
Draft saved at 3:06:08 PM
adult.data.csv
Split
Two-Class Boosted Decision T...
Train Model
Properties
Train Model
Label column
Launch column selector
Experiment Properties
STATUS CODE
Disable upgrades
1:1

Select a single column

Include | column names | income ✕

Selected columns:
Launch the selector tool to make a selection

STATUS CODE InDraft

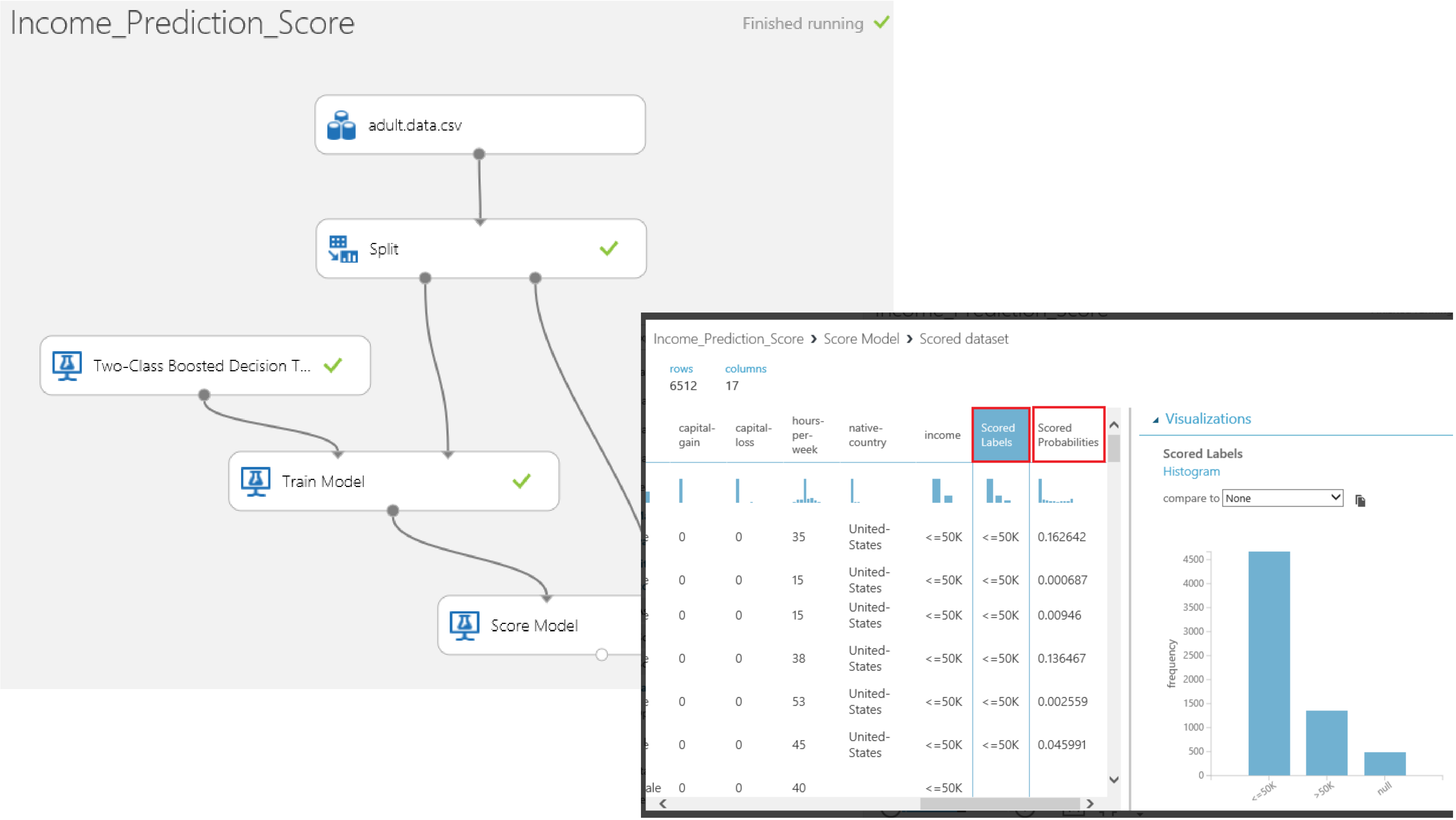
□ Disable upgrades

Train Model

Train a previously created classification or regression model
(more...)

Part III Azure ML Hands-On: Classification

Step 5: Score Model with Scored Labels and Probabilities.



Income_Prediction_Score › Score Model › Scored dataset

rows	columns
6512	17

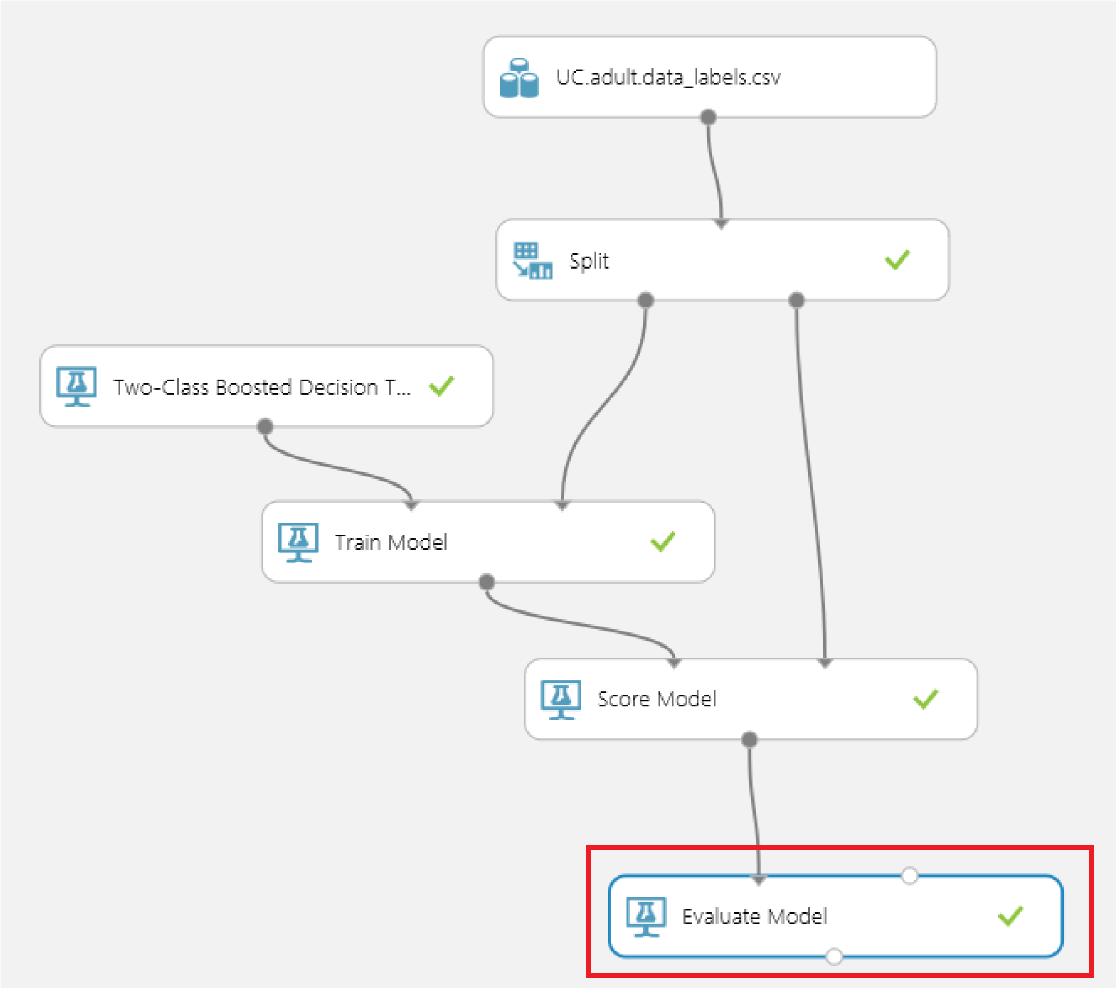
	capital-gain	capital-loss	hours-per-week	native-country	income	Scored Labels	Scored Probabilities
e	0	0	35	United-States	<=50K	<=50K	0.162642
e	0	0	15	United-States	<=50K	<=50K	0.000687
e	0	0	15	United-States	<=50K	<=50K	0.00946
e	0	0	38	United-States	<=50K	<=50K	0.136467
e	0	0	53	United-States	<=50K	<=50K	0.002559
e	0	0	45	United-States	<=50K	<=50K	0.045991
ale	0	0	40		<=50K		

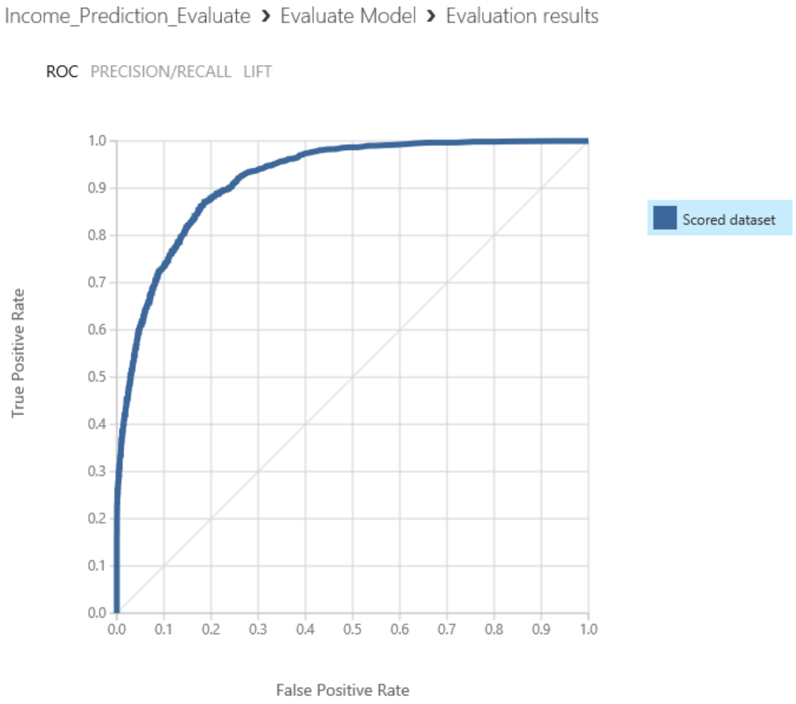
Scored Labels
Histogram

compare to None

Part III Azure ML Hands-On: Classification

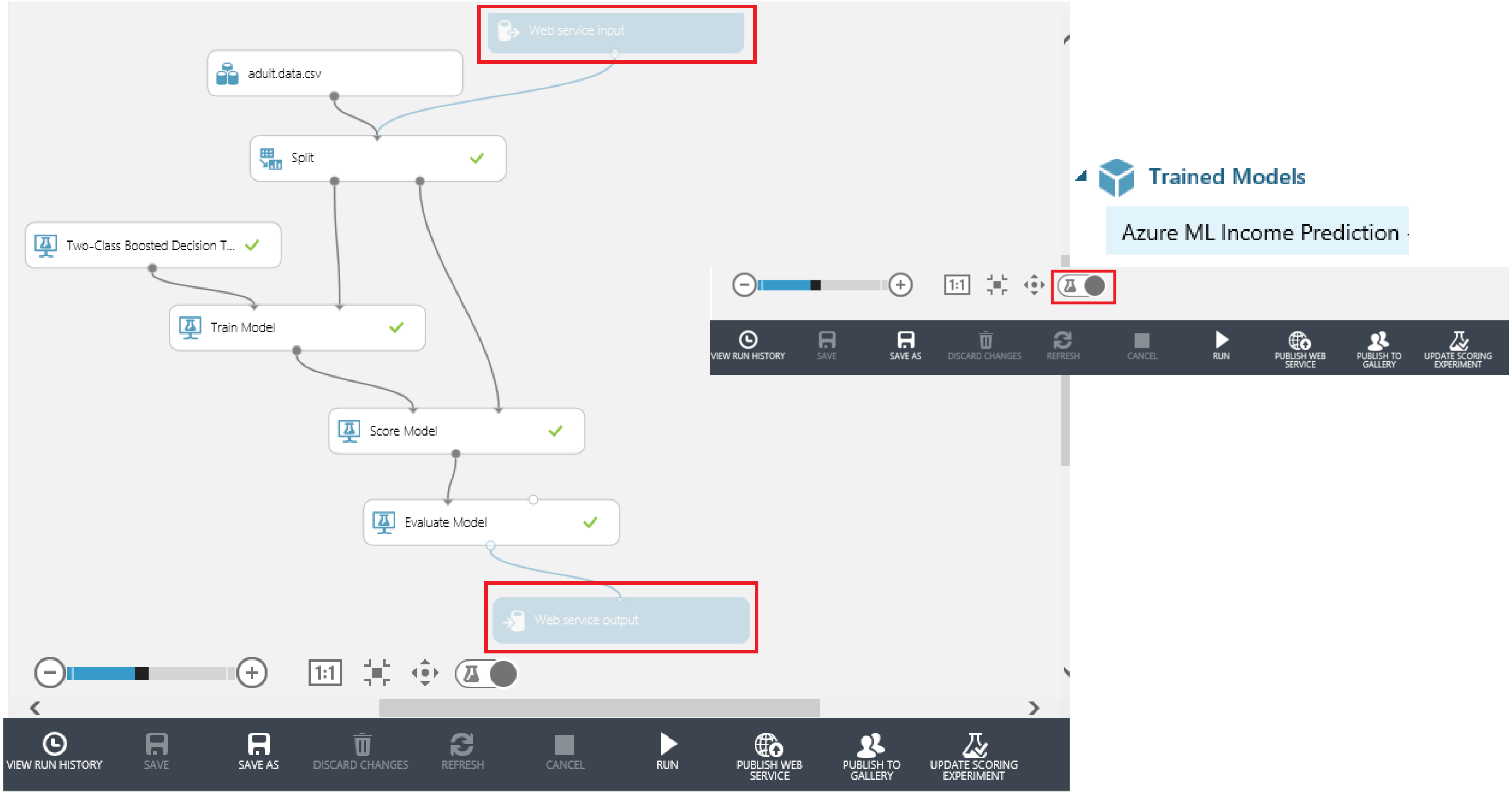
Step 6: Evaluate Model. Check the ROC curve.





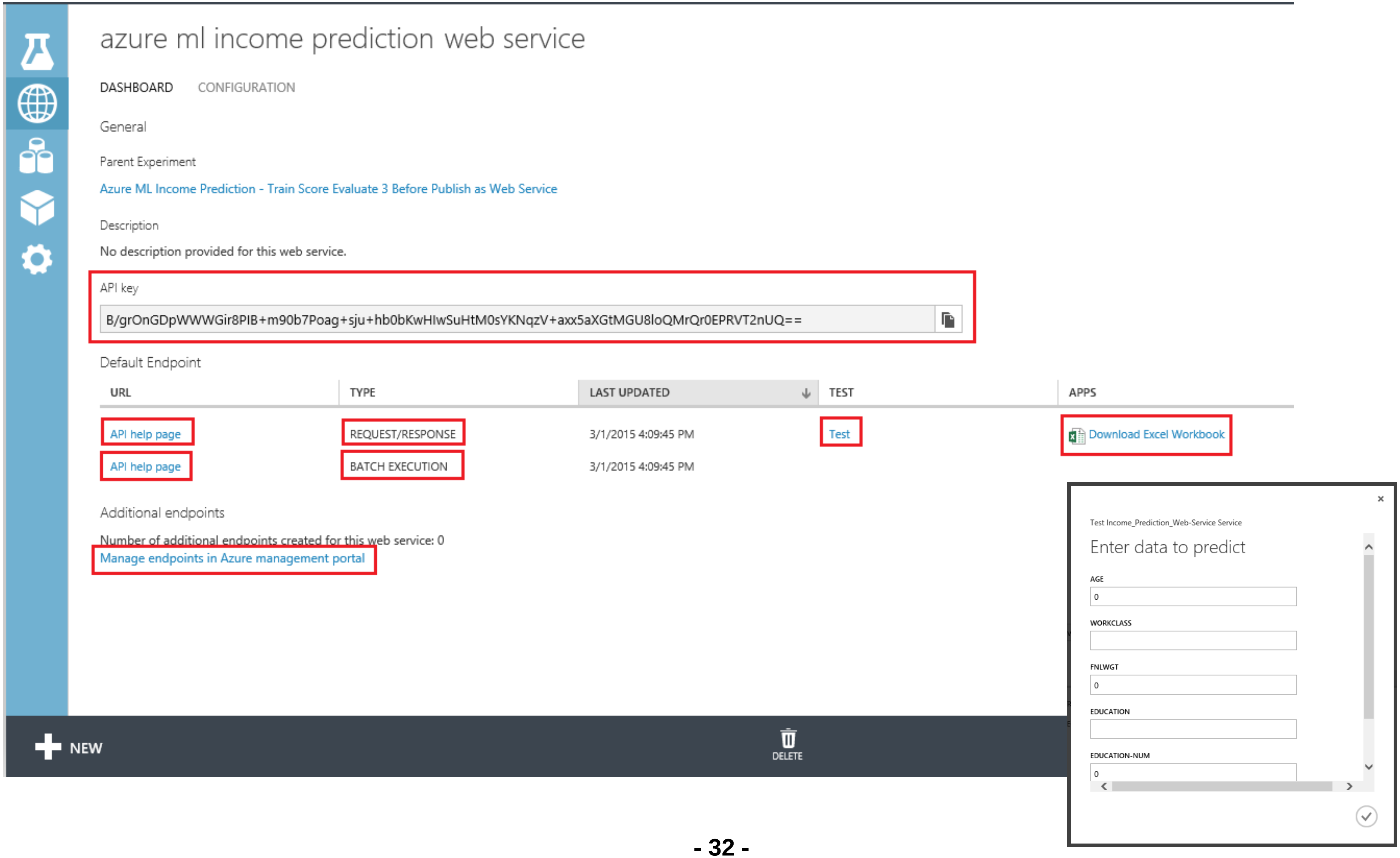
Part III Azure ML Hands-On: Classification

Step 7: Publish as a Web service.



Part III Azure ML Hands-On: Classification

Step 8: Test Web service.



azure ml income prediction web service

DASHBOARD CONFIGURATION

General

Parent Experiment

Azure ML Income Prediction - Train Score Evaluate 3 Before Publish as Web Service

Description

No description provided for this web service.

API key

B/grOnGDpWWWGir8PIB+m90b7Poag+sju+hb0bKwHIwSuHtM0sYKNqzV+axx5aXGtMGU8loQMrQr0EPRVT2nUQ==

Default Endpoint

URL	TYPE	LAST UPDATED	TEST	APPS
API help page	REQUEST/RESPONSE	3/1/2015 4:09:45 PM	Test	Download Excel Workbook
API help page	BATCH EXECUTION	3/1/2015 4:09:45 PM		

Additional endpoints

Number of additional endpoints created for this web service: 0

Manage endpoints in Azure management portal

Test Income_Prediction_Web-Service Service

Enter data to predict

AGE: 0

WORKCLASS:

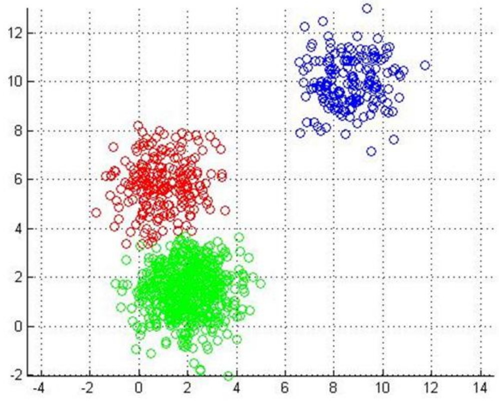
FNLWGT: 0

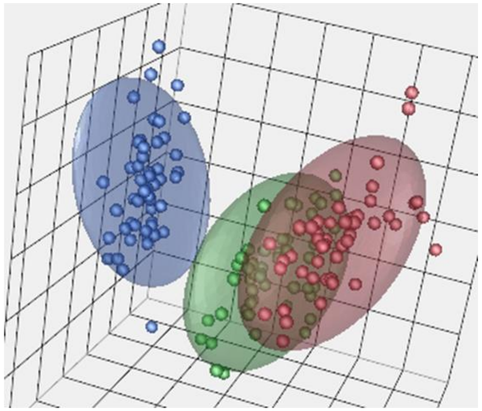
EDUCATION:

EDUCATION-NUM: 0

Part III Azure ML Hands-On: Clustering

1. K-Means Clustering module: The K-means method finds a specified number of clusters for a set of D-dimensional data points. It starts an initial set of K centroids, and then uses algorithms to iteratively refine the locations of the centroids. The algorithm terminates when the centroids stabilize or when a specified number of iterations are computed.
2. Train Clustering Model module: This module takes an untrained clustering model, such as that produced by the K-Means Clustering module, and an unlabeled data set. It returns a trained clustering model that can be passed to the Assign to Clusters module. It also returns labels for the training data.
3. Assign to Clusters module: This module takes a trained clustering model, produced by the Train Clustering Model module, and an unlabeled data set. The module then returns the cluster assignments (indexes) for the input data.

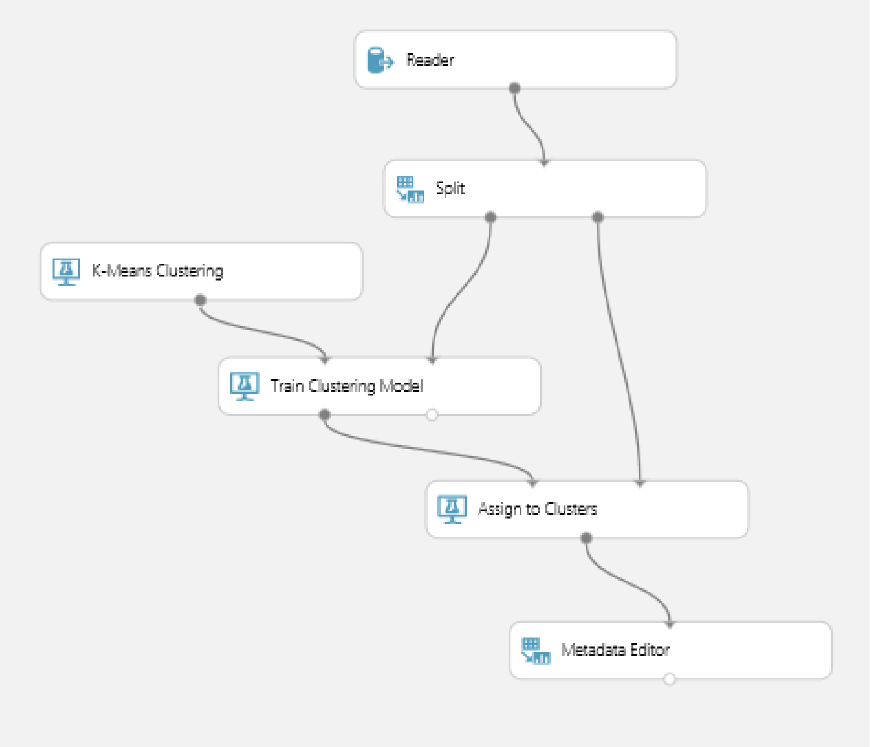




Part III Azure ML Hands-On: Grouping wholesale customers

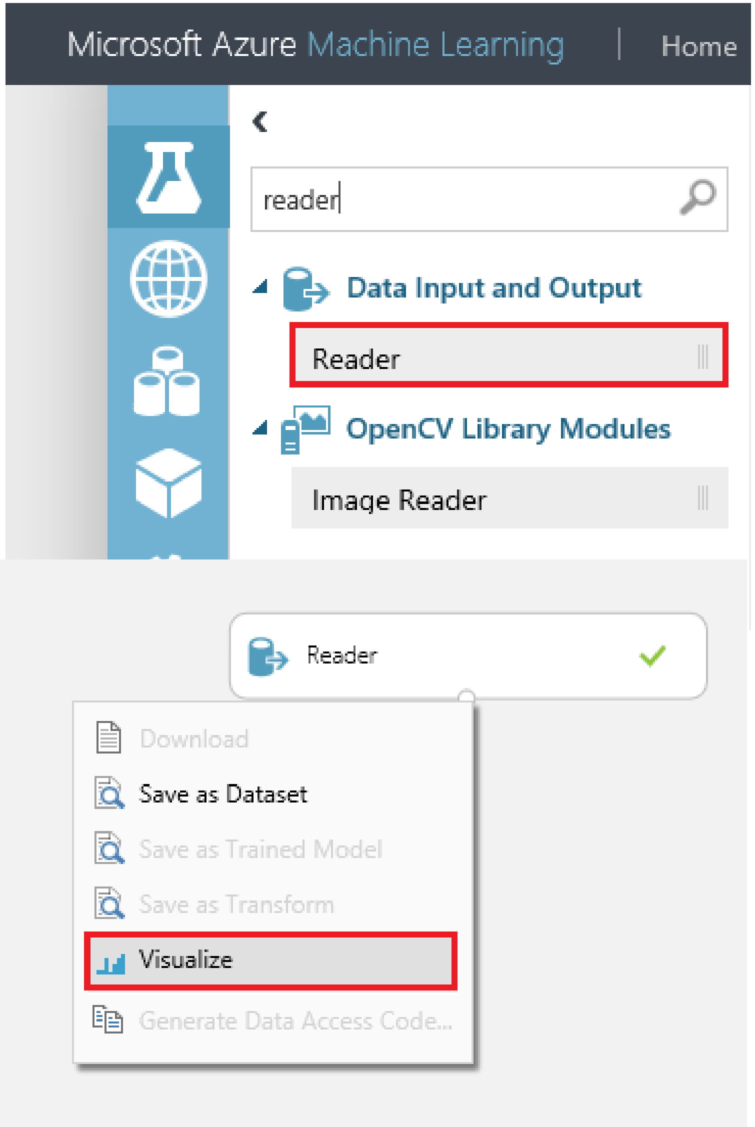
1. Datasets
   - Wholesale customers dataset from the UCI Machine Learning Repository located at http://mlr.cs.umass.edu/ml/datasets/Wholesale+customers

2. It includes the annual spending on the following product categories
   a. Channel: Retail channel types such as hotel/restaurant/café.
   b. Region: Customer region code.
   c. Fresh: Annual spending on fresh products.
   d. Milk: Annual spending on milk products.
   e. Grocery: Annual spending on grocery products.
   f. Frozen: Annual spending on frozen products.
   g. Detergents-Paper: Annual spending on detergents and paper products
   h. Delicatessen: Annual spending on delicatessen products.



Part III Azure ML Hands-On: Grouping wholesale customers

Step 1: Get the data from URL and explore the data.



Microsoft Azure Machine Learning | Home

reader

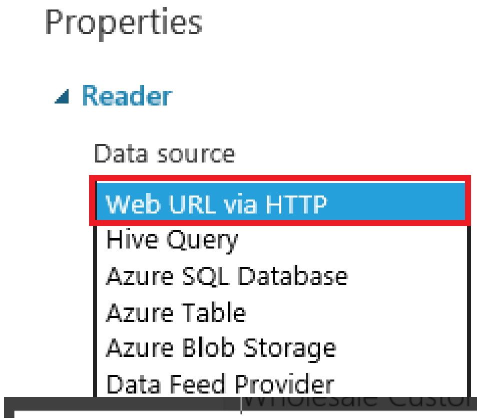
Data Input and Output

- Reader

OpenCV Library Modules

- Image Reader

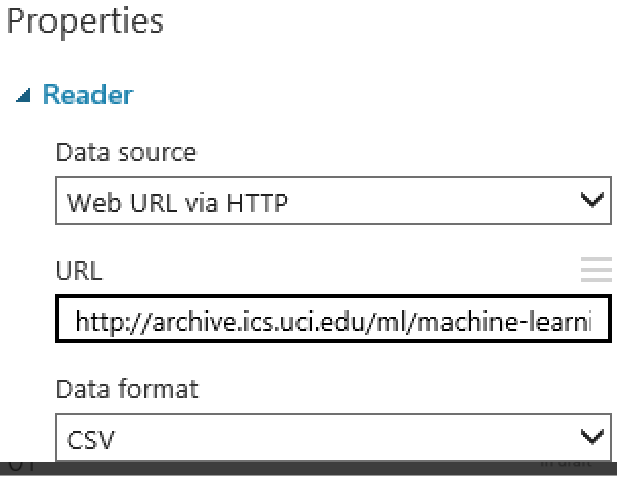
Reader ✓



Properties

Reader

Data source



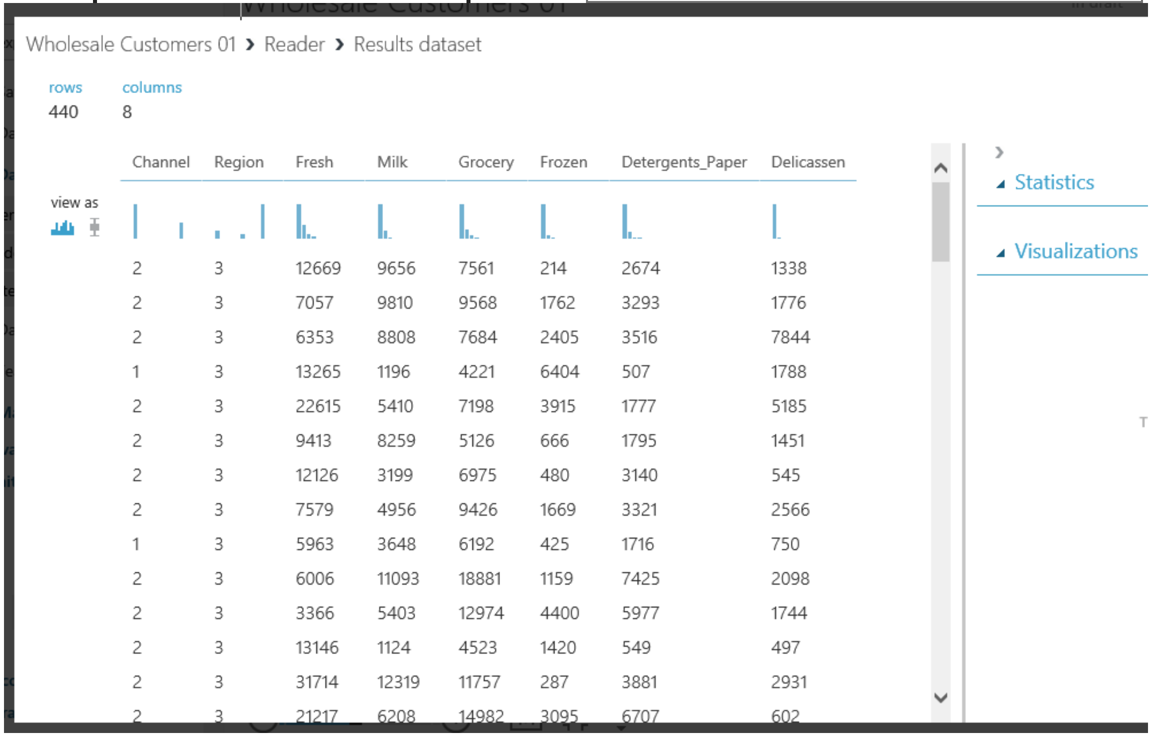
Properties

Reader

Data source: Web URL via HTTP

URL: http://archive.ics.uci.edu/ml/machine-learni

Data format: CSV



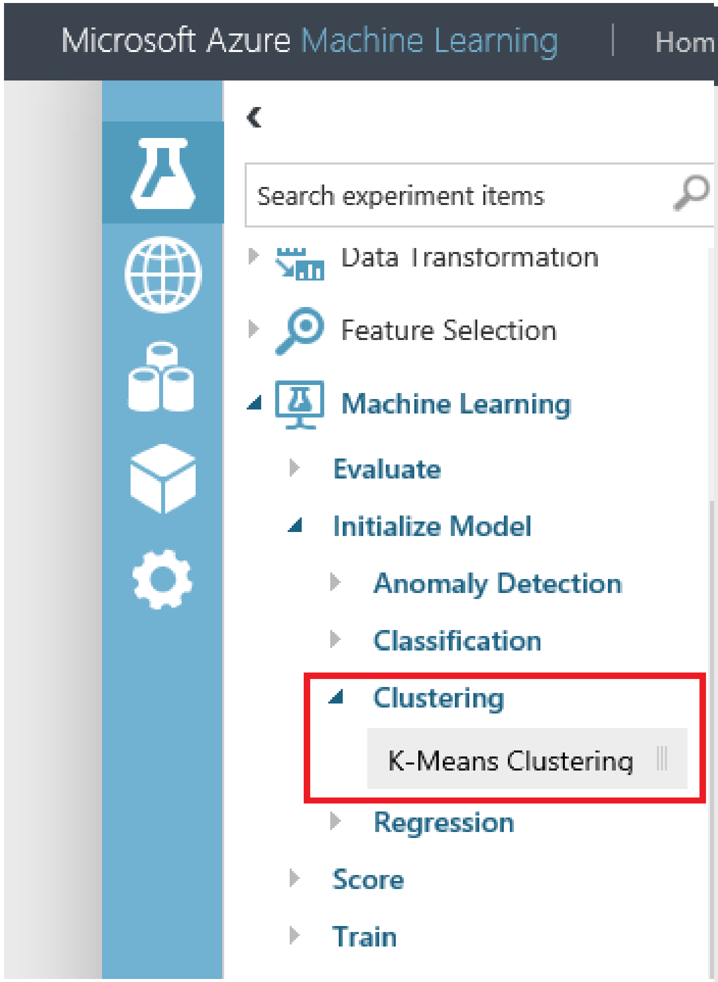
Wholesale Customers 01 › Reader › Results dataset

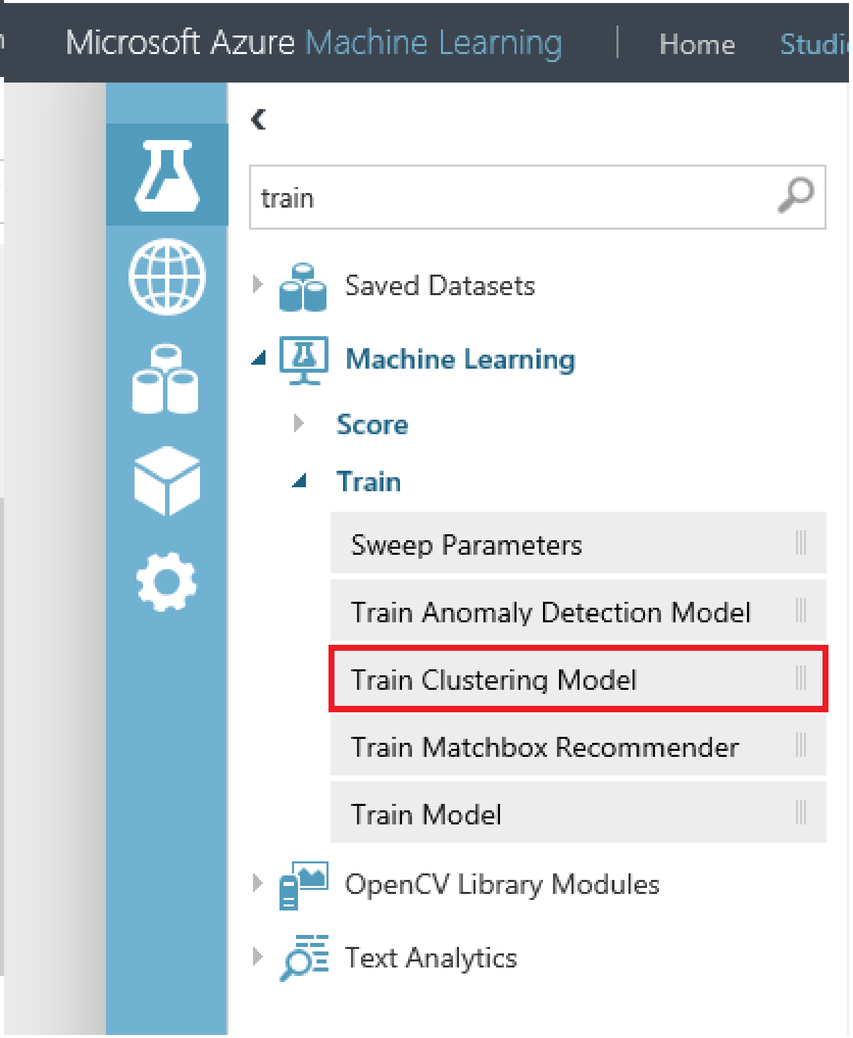
rows 440 columns 8

Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
2	3	12669	9656	7561	214	2674	1338
2	3	7057	9810	9568	1762	3293	1776
2	3	6353	8808	7684	2405	3516	7844
1	3	13265	1196	4221	6404	507	1788
2	3	22615	5410	7198	3915	1777	5185
2	3	9413	8259	5126	666	1795	1451
2	3	12126	3199	6975	480	3140	545
2	3	7579	4956	9426	1669	3321	2566
1	3	5963	3648	6192	425	1716	750
2	3	6006	11093	18881	1159	7425	2098
2	3	3366	5403	12974	4400	5977	1744
2	3	13146	1124	4523	1420	549	497
2	3	31714	12319	11757	287	3881	2931
2	3	21217	6208	14982	3095	6707	602

Part III Azure ML Hands-On: Grouping wholesale customers

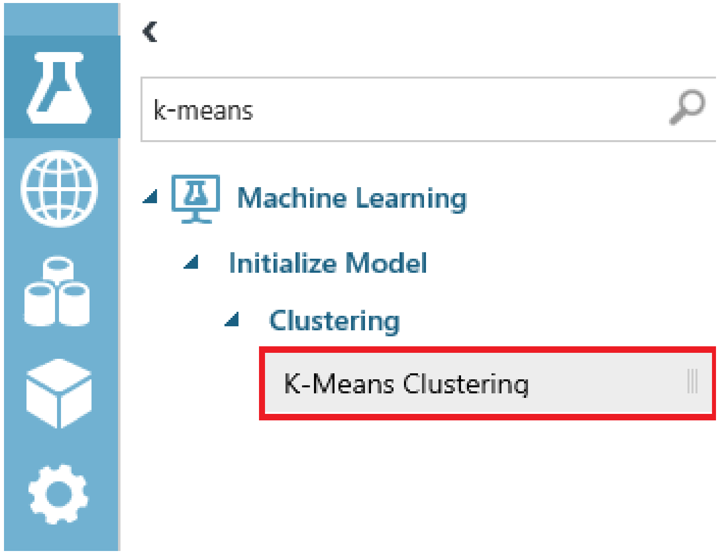
Step 2: Choose the K-Means Clustering Model and Train the Clustering Model.

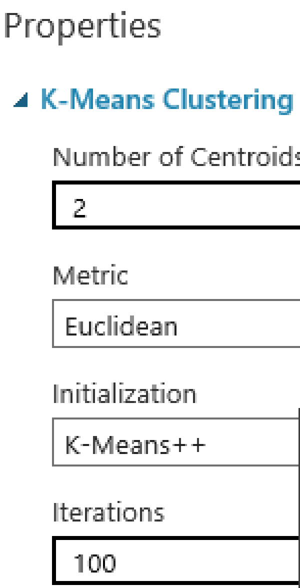




Part III Azure ML Hands-On: Grouping wholesale customers

Step 3: Setup the K-Means Clustering properties.


k-means


Properties

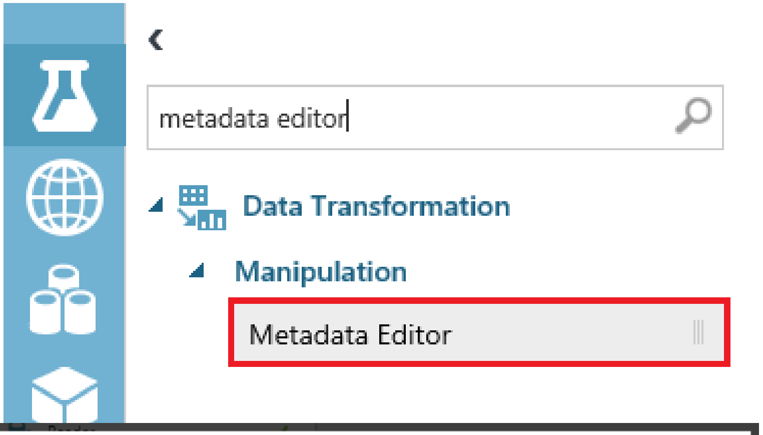
K-Means Clustering

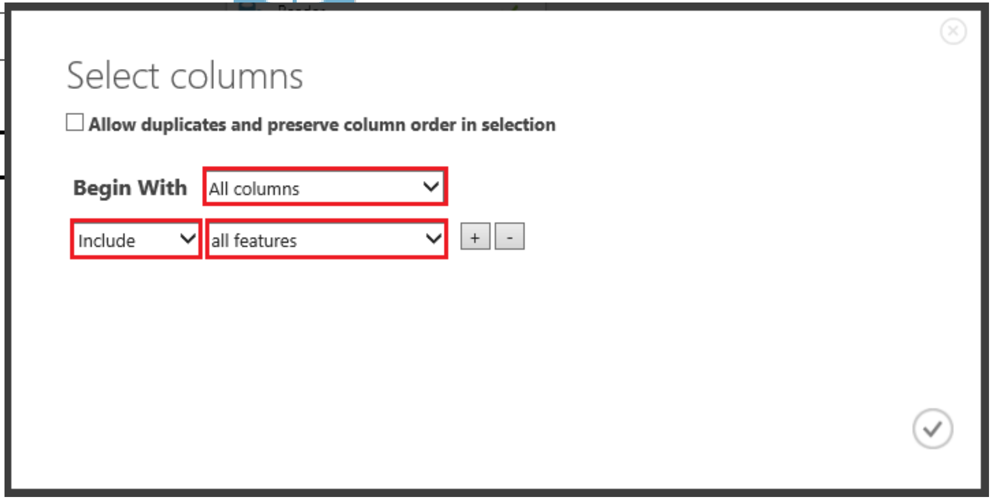
Number of Centroids: 2

Metric: Euclidean

Initialization: K-Means++

Iterations: 100


metadata editor


Select columns

□ Allow duplicates and preserve column order in selection

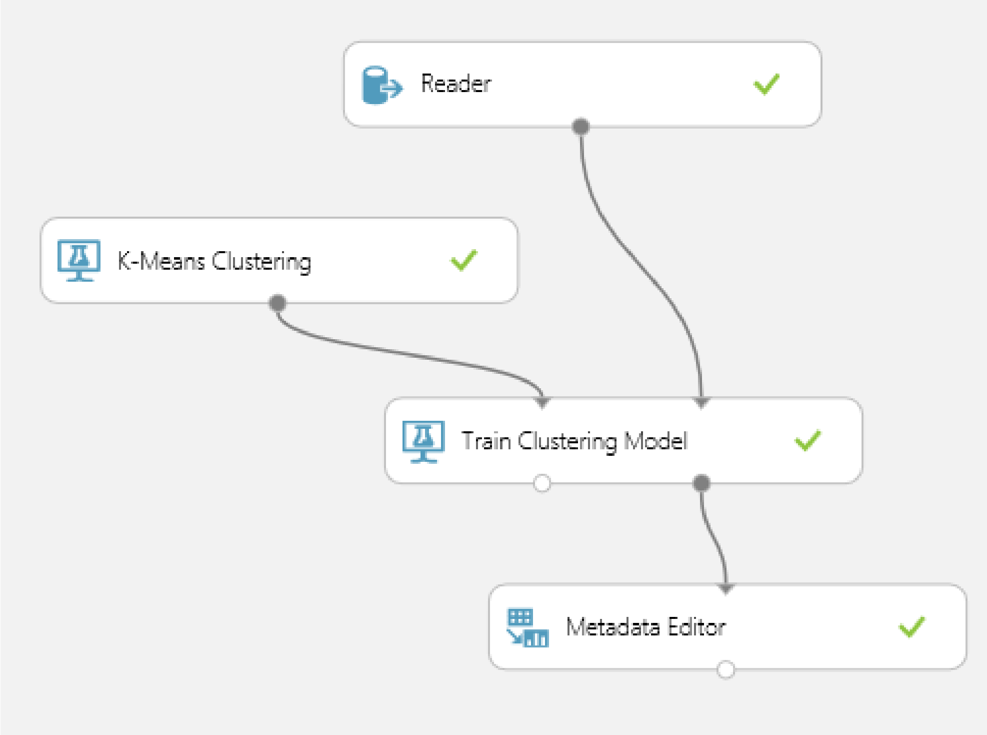
Begin With: All columns

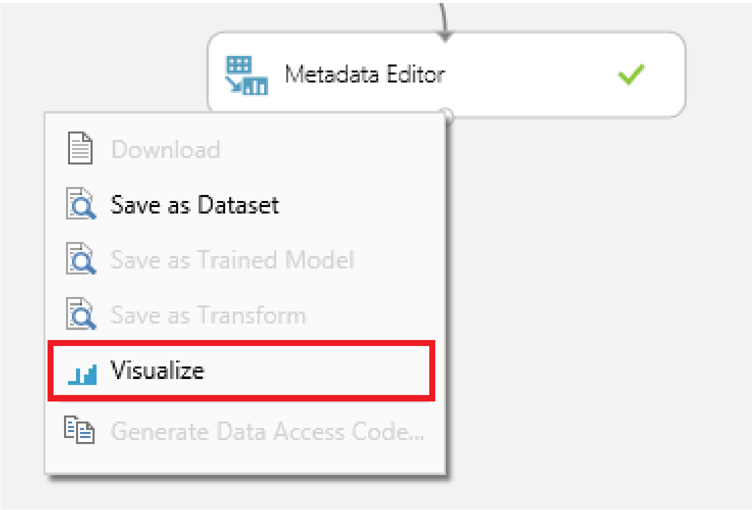
Include: all features

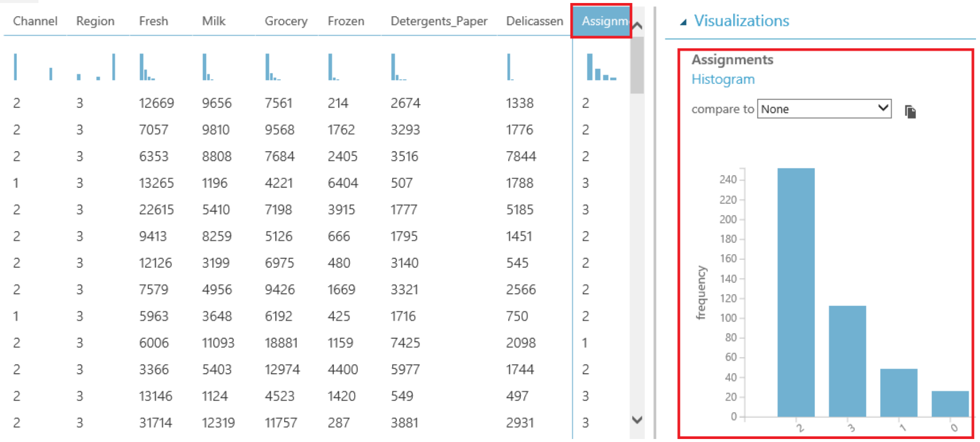
- Euclidean: distance between two points is the length of line segment connecting them.
- Cosine: Measure of similarity between two vectors of an inner space that measure the cosine of the angles between them.

Part III Azure ML Hands-On: Grouping wholesale customers

Step 4: Visualize the Training Results.



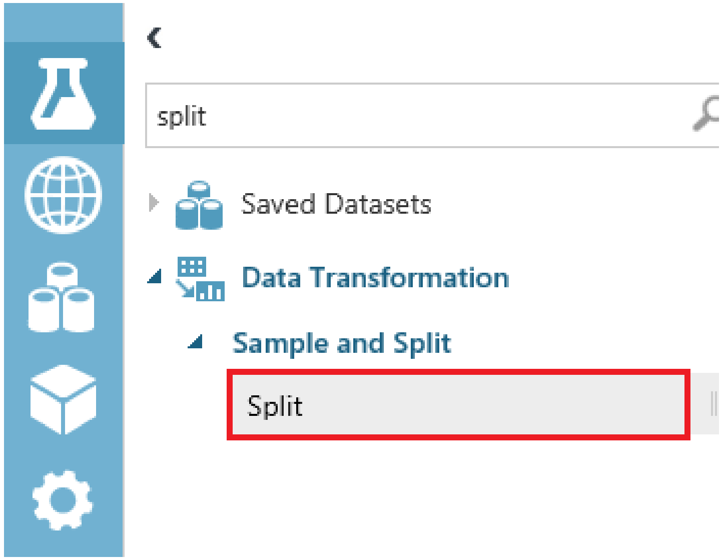


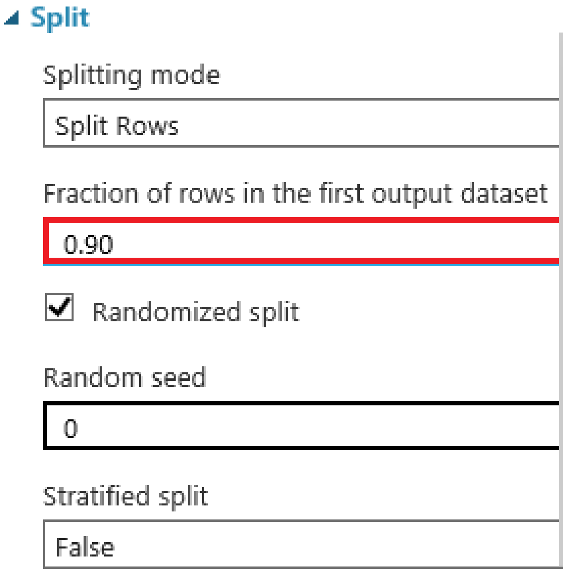


Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Assignm
2	3	12669	9656	7561	214	2674	1338	2
2	3	7057	9810	9568	1762	3293	1776	2
2	3	6353	8808	7684	2405	3516	7844	2
1	3	13265	1196	4221	6404	507	1788	3
2	3	22615	5410	7198	3915	1777	5185	3
2	3	9413	8259	5126	666	1795	1451	2
2	3	12126	3199	6975	480	3140	545	2
2	3	7579	4956	9426	1669	3321	2566	2
1	3	5963	3648	6192	425	1716	750	2
2	3	6006	11093	18881	1159	7425	2098	1
2	3	3366	5403	12974	4400	5977	1744	2
2	3	13146	1124	4523	1420	549	497	3
2	3	31714	12319	11757	287	3881	2931	3

Part III Azure ML Hands-On: Grouping wholesale customers

Step 5: Split the data and set up for testing.





Split

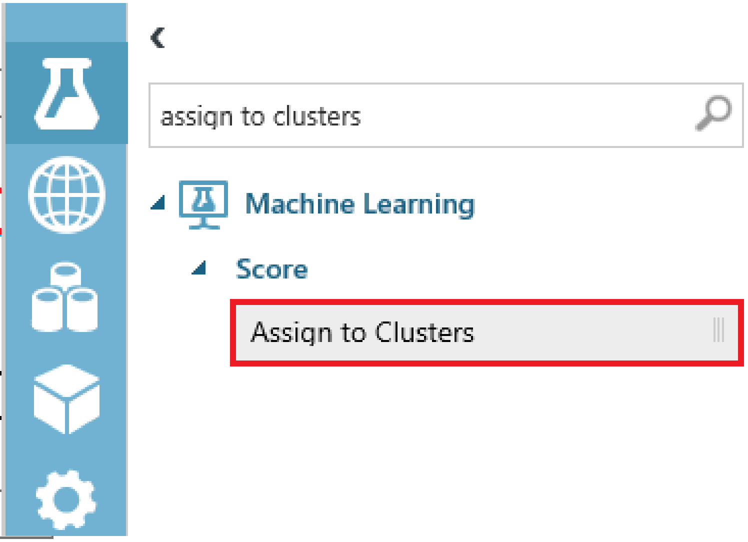
Splitting mode: Split Rows

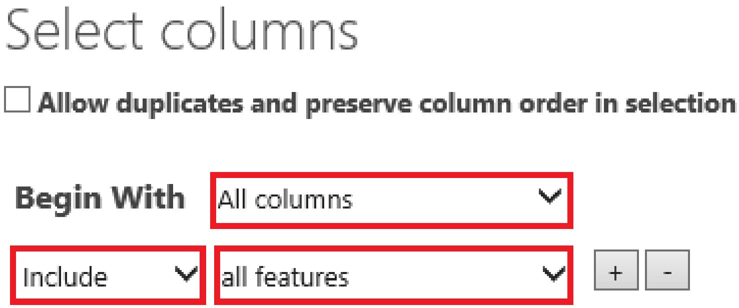
Fraction of rows in the first output dataset: 0.90

☑ Randomized split

Random seed: 0

Stratified split: False





Select columns

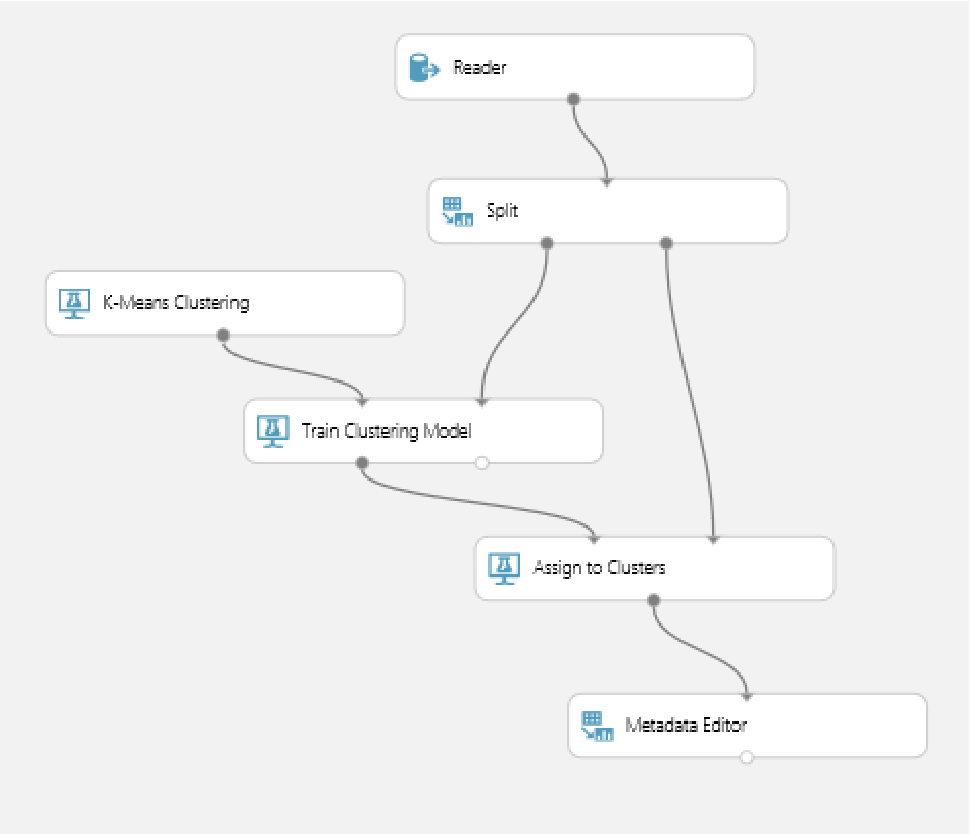
☐ Allow duplicates and preserve column order in selection

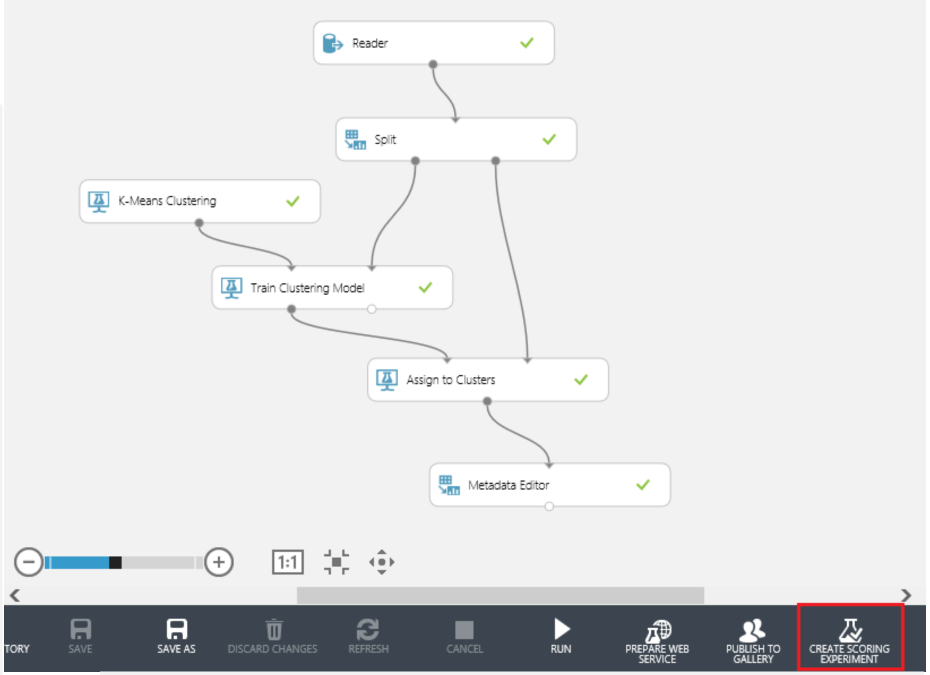
Begin With: All columns

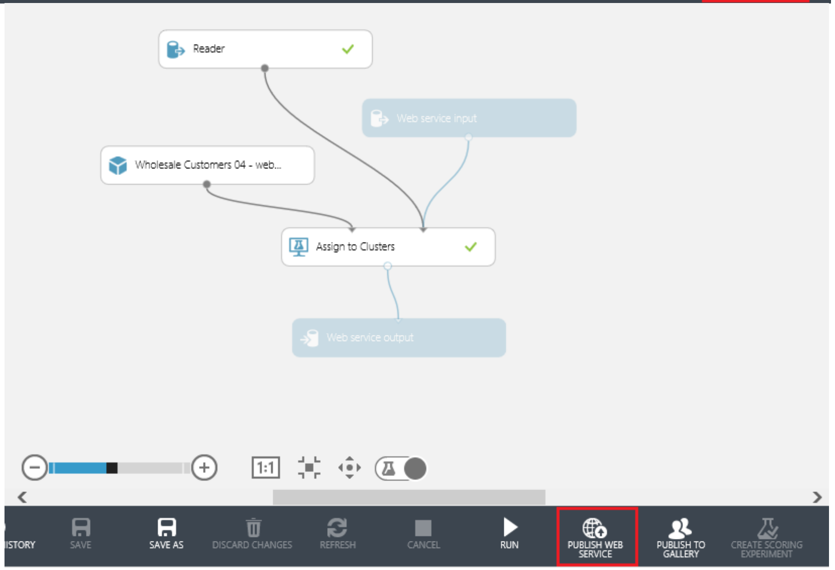
Include: all features

Part III Azure ML Hands-On: Grouping wholesale customers

Step 6: Publish as a Web Service.

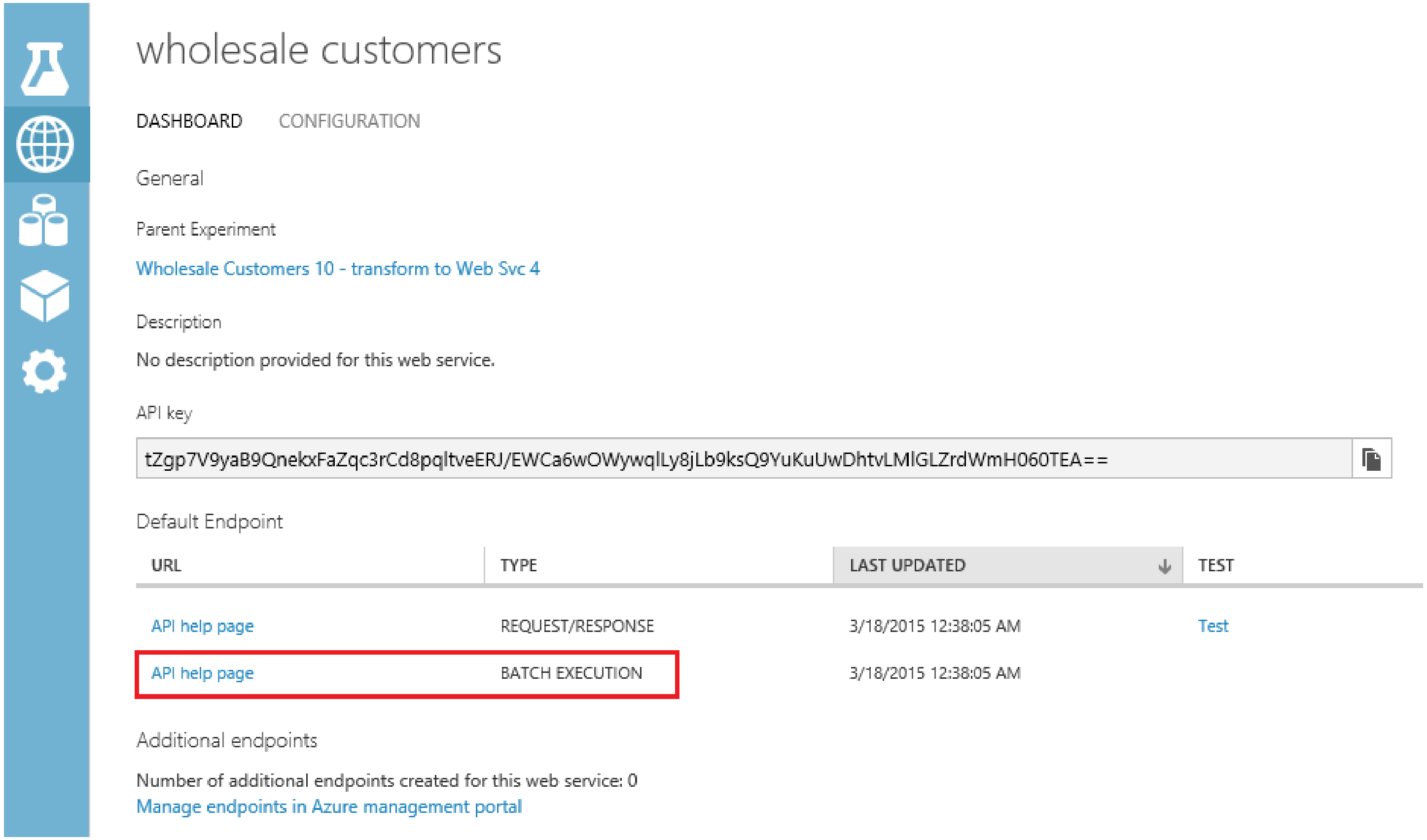






Part III Azure ML Hands-On: Grouping wholesale customers

Step 7: Test the Web Service.



wholesale customers

General

Parent Experiment

Wholesale Customers 10 - transform to Web Svc 4

Description

No description provided for this web service.

API key

tZgp7V9yaB9QnekxFaZqc3rCd8pqltveERJ/EWCa6wOWywqlLy8jLb9ksQ9YuKuUwDhtvLMlGLZrdWmH060TEA==

Default Endpoint

URL	TYPE	LAST UPDATED	TEST
API help page	REQUEST/RESPONSE	3/18/2015 12:38:05 AM	Test
API help page	BATCH EXECUTION	3/18/2015 12:38:05 AM	

Additional endpoints

Number of additional endpoints created for this web service: 0
Manage endpoints in Azure management portal

Part III Azure ML Hands-On: Recommendation

1. Create a model: A model is a container of your usage data, catalog data, and the recommendation model.

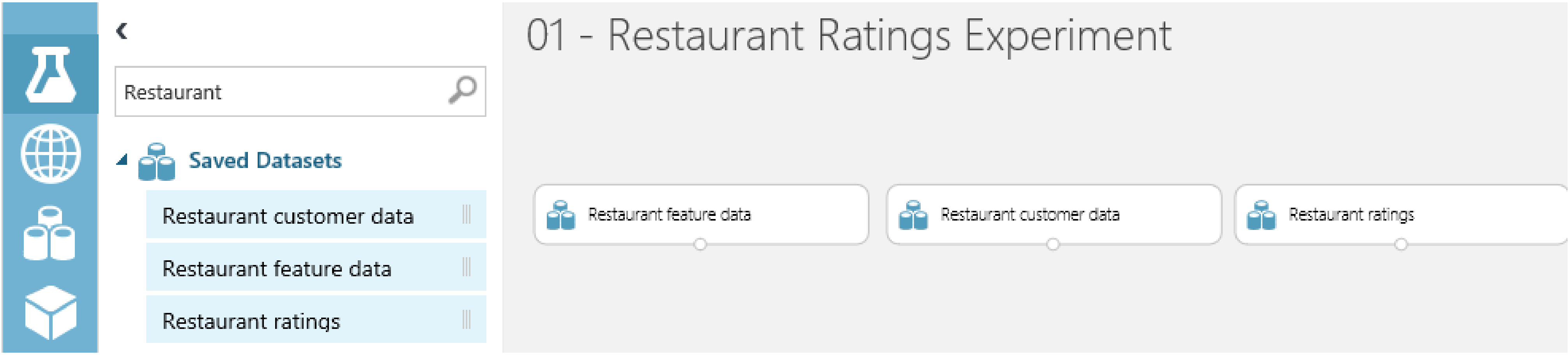
2. Import catalog data: This is an optional step. A catalog contains metadata information on the items, if you do not upload catalog data, the recommendation's services will learn about your catalog from the usage data.

3. Import usage data
   - By uploading a file that contains the usage data.
   - By sending data acquisition events. Usually you upload a usage file to be able to create an initial recommendation model (bootstrap). You would use this usage until the system gathers enough data by using the data acquisition format.

4. Building a recommendation model:
   - This is an asynchronous operation in which the recommendation system takes all the usage data and creates a recommendation model. This operation can take several minutes or several hours depending on the size of the data and the build configuration parameters. When triggering the build, you will get a build ID. Use it to check when the build process has ended before starting to consume recommendations.

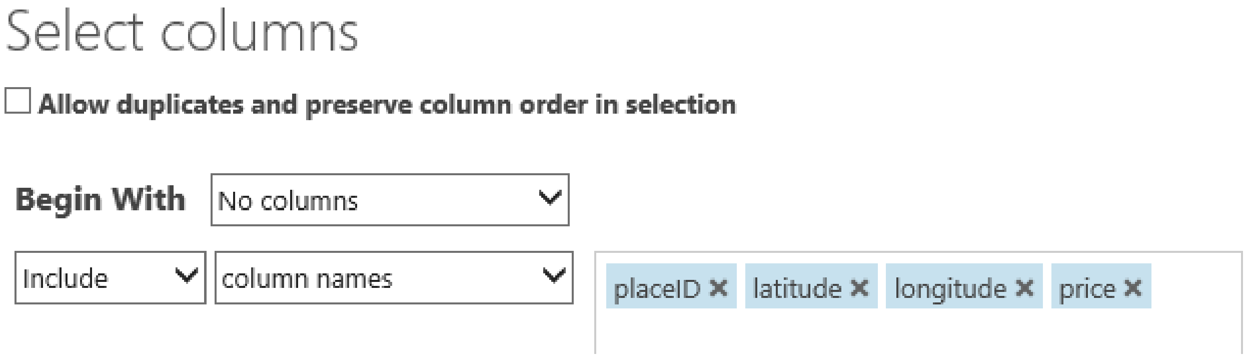
Part III Azure ML: Building the restaurant ratings recommender

1. Data sets
   a. Restaurant features: This includes information about each restaurant, such as placeID, location, name, address, state, whether alcohol is served, smoking policy, dress code, accessibility, and price.
   b. Restaurant customers: This includes a wide variety of personal attributes include userID, smoker, drink level, dress preference, marital status, birth year, interests, personality traits, religion, favorite color, weight, and budget.
   c. Restaurant ratings: This includes key fields like userID, placeID, and rating.



Part III Azure ML: Building the restaurant ratings recommender

2. Select Columns
   a. We can filter out unnecessary columns in the Restaurant Feature Data and Restaurant Customer Data datasets.
   b. Restaurant features: placeID, latitude, longitude, price.
   c. Customer Data: userID, latitude, longitude, interest, personality.
3. Split Module
   - Specify that half of this dataset is to be used for training and the other half of scoring the new recommendation model.

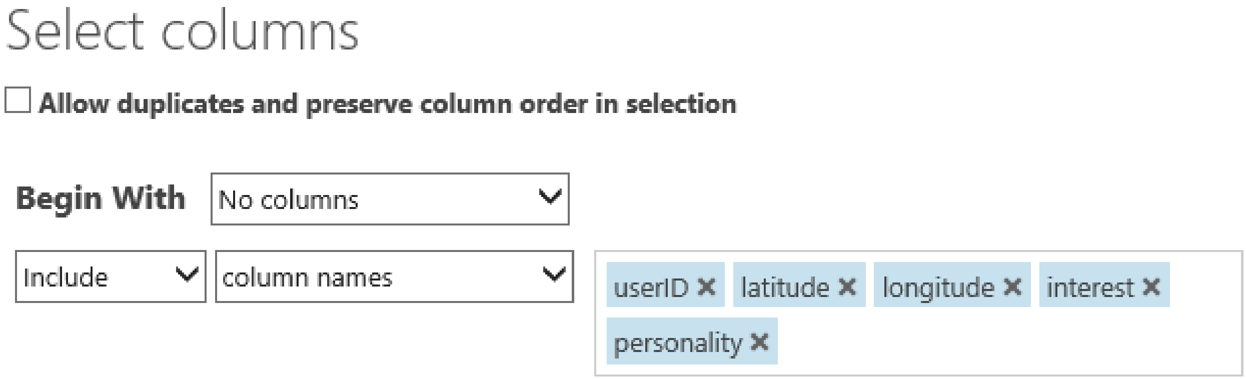


Select columns

☐ Allow duplicates and preserve column order in selection

Begin With: No columns

Include · column names · placeID ✕ latitude ✕ longitude ✕ price ✕

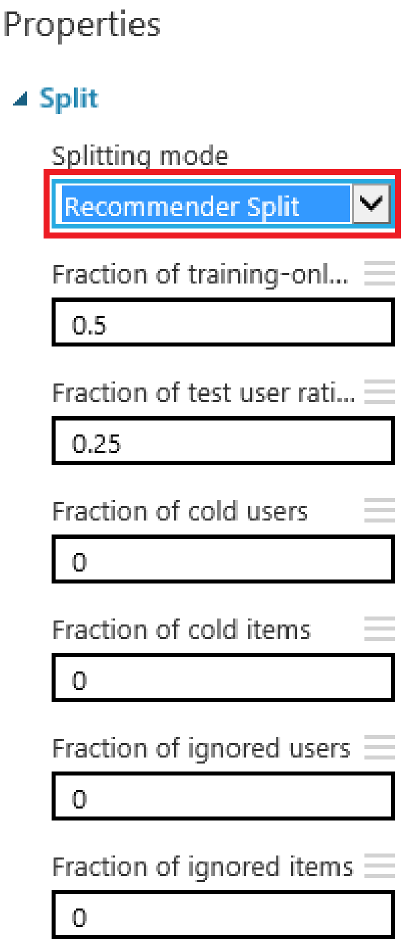


Select columns

☐ Allow duplicates and preserve column order in selection

Begin With: No columns

Include · column names · userID ✕ latitude ✕ longitude ✕ interest ✕ personality ✕



Properties

Split

Splitting mode: Recommender Split

Fraction of training-onl...: 0.5

Fraction of test user rati...: 0.25

Fraction of cold users: 0

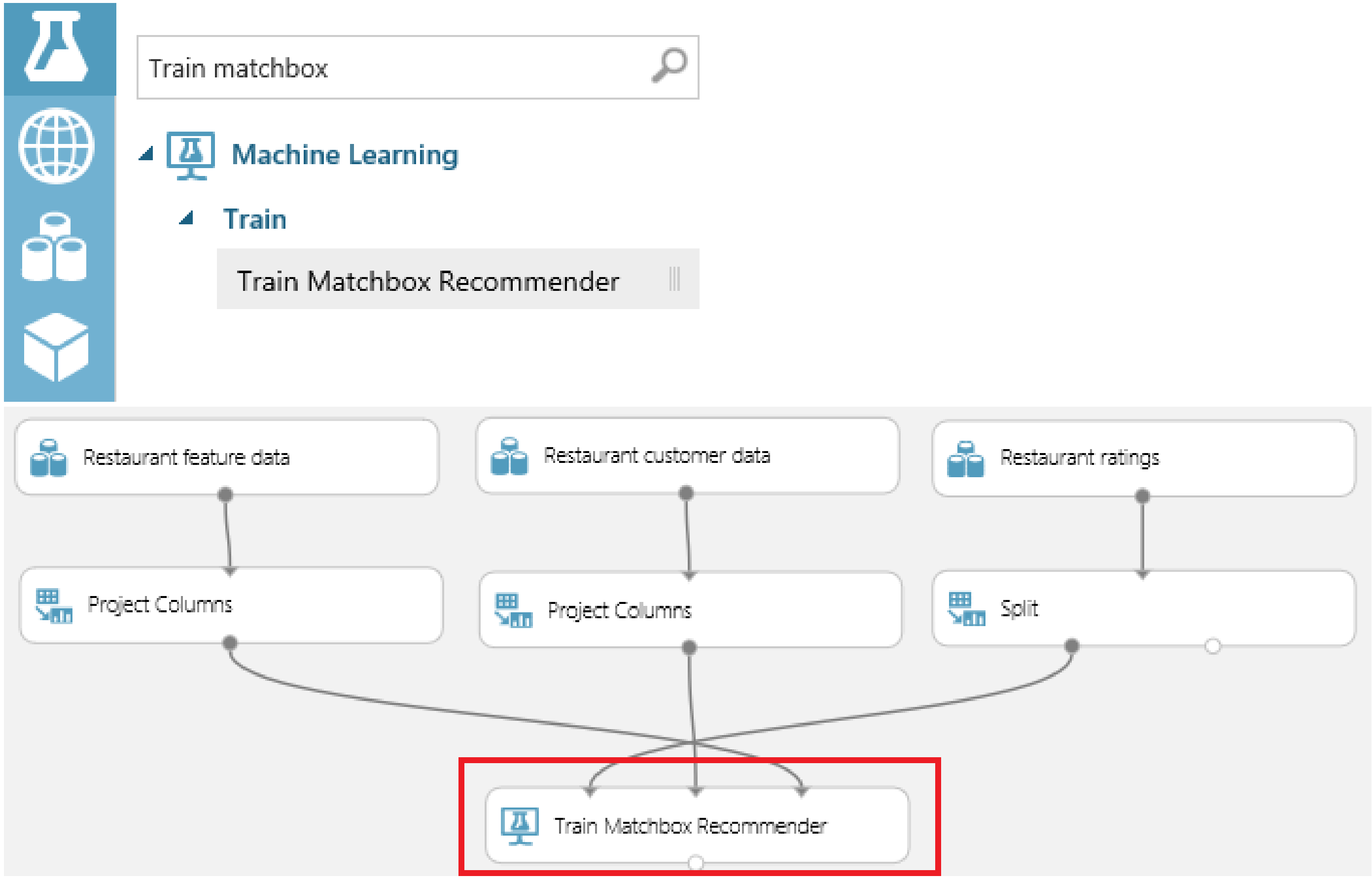
Fraction of cold items: 0

Fraction of ignored users: 0

Fraction of ignored items: 0

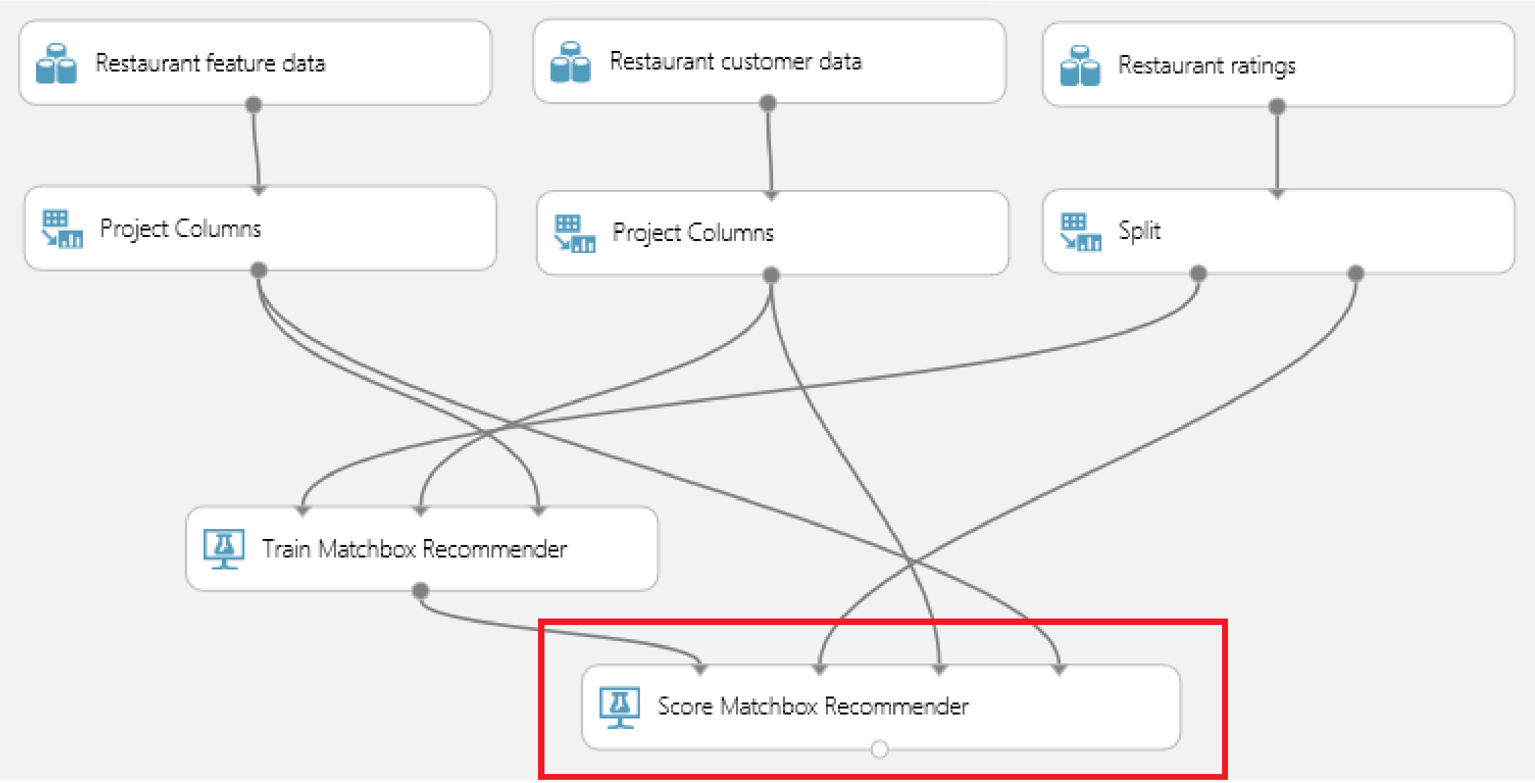
Part III Azure ML: Building the restaurant ratings recommender

4. Train Matchbox Recommender module
    a. Training dataset of user-item-rating triples: Ratings of items by users, expressed as a triple (Use, Item, Rating).
    b. Training dataset of user features: Dataset containing features that describe users.
    c. Training dataset of item features: Dataset containing features that describe items.



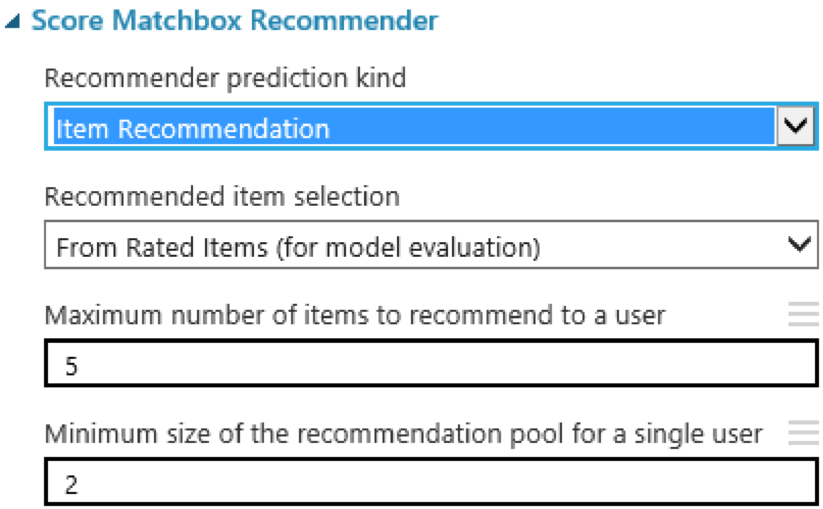
Part III Azure ML: Building the restaurant ratings recommender

5. Score Matchbox
    a. Trained Matchbox recommender.
    b. Dataset to score.
    c. User features: Dataset containing features that describe users.
    d. Item features: Dataset containing features that describe items.



Part III Azure ML: Building the restaurant ratings recommender

5. Score Matchbox
   a. Rating Prediction: Predict the rating that a customer will give a particular restaurant.
   b. Item Recommendation: Predict which restaurants will be most highly rated by the user.
   c. Related Users: Predict which customers (users) are most like this customer.
   d. Related Items: Predict which restaurants (items) are most like this restaurant.



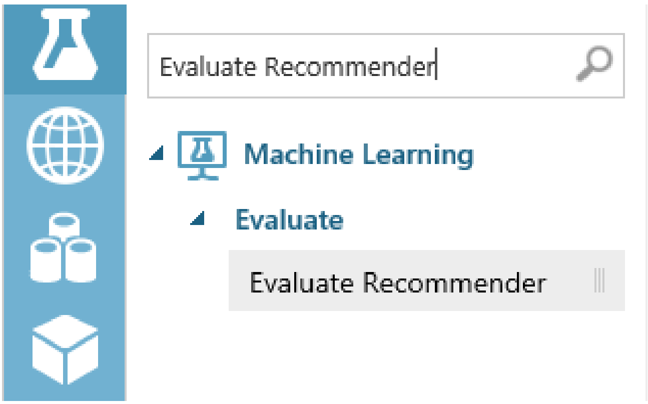
◢ Score Matchbox Recommender

Recommender prediction kind: Item Recommendation

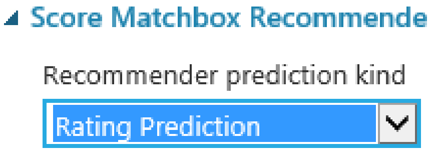
Recommended item selection: From Rated Items (for model evaluation)

Maximum number of items to recommend to a user: 5

Minimum size of the recommendation pool for a single user: 2



Evaluate Recommender

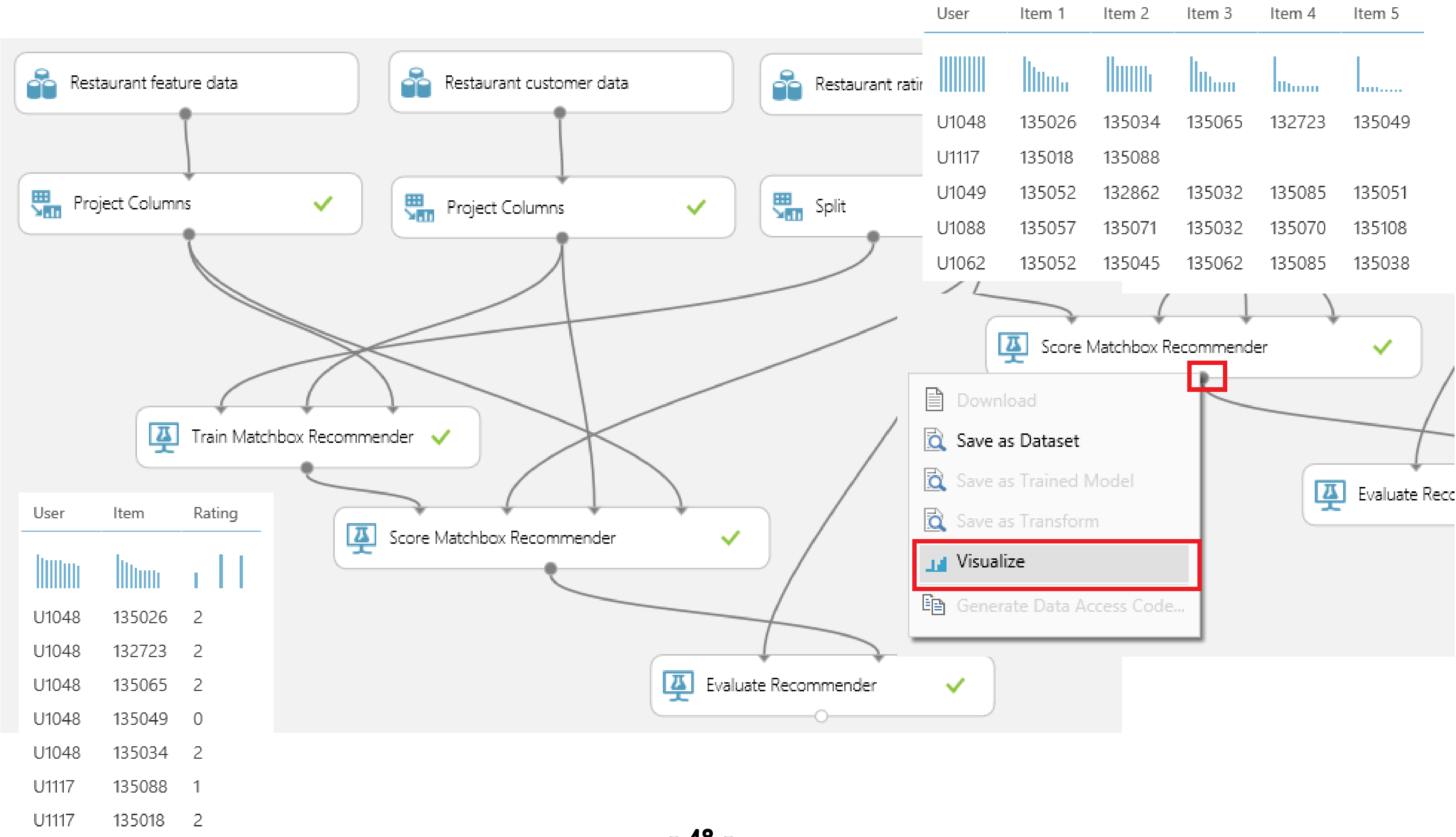


◢ Score Matchbox Recommender

Recommender prediction kind: Rating Prediction

Part III Azure ML: Building the restaurant ratings recommender

5. Evaluate Recommender



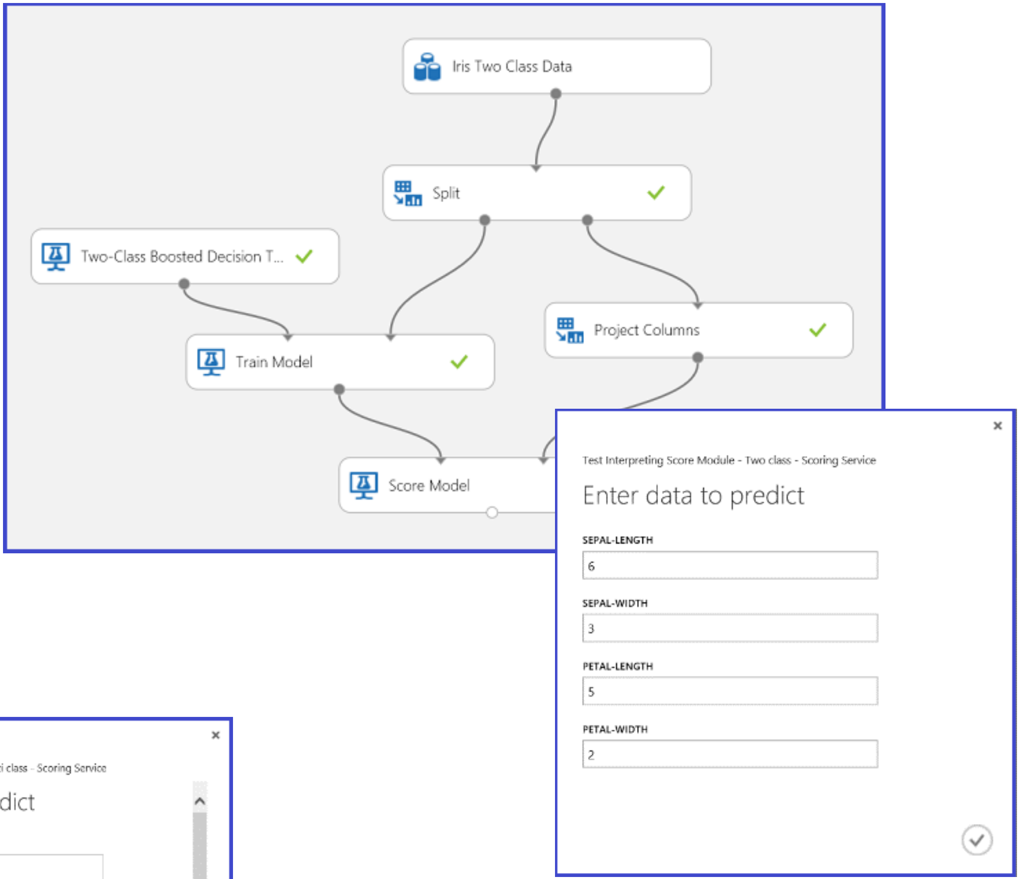
User	Item 1	Item 2	Item 3	Item 4	Item 5
U1048	135026	135034	135065	132723	135049
U1117	135018	135088			
U1049	135052	132862	135032	135085	135051
U1088	135057	135071	135032	135070	135108
U1062	135052	135045	135062	135085	135038

User	Item	Rating
U1048	135026	2
U1048	132723	2
U1048	135065	2
U1048	135049	0
U1048	135034	2
U1117	135088	1
U1117	135018	2

Part III Azure ML Exercises: Classification

1. Two-class classification
   a. To classify Iris flowers based on their features.
   b. Two flower species (classes 0 and 1).
   c. Four features for each flower (sepal length, sepal width, petal length, and petal width)

2. Multi-class classification
   1. To classify a letter (class), given some attribute values extracted from the hand-written letter images.
   2. Twenty-six letters form twenty-six classes.



Test Interpreting Score Module - Two class - Scoring Service

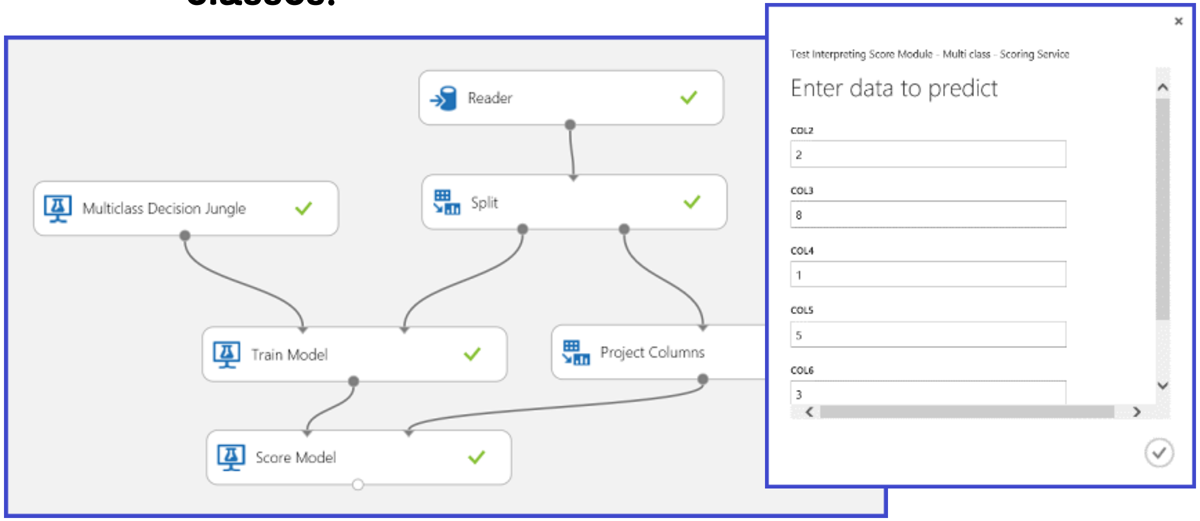
Enter data to predict

SEPAL-LENGTH: 6

SEPAL-WIDTH: 3

PETAL-LENGTH: 5

PETAL-WIDTH: 2



Test Interpreting Score Module - Multi class - Scoring Service

Enter data to predict

COL2: 2

COL3: 8

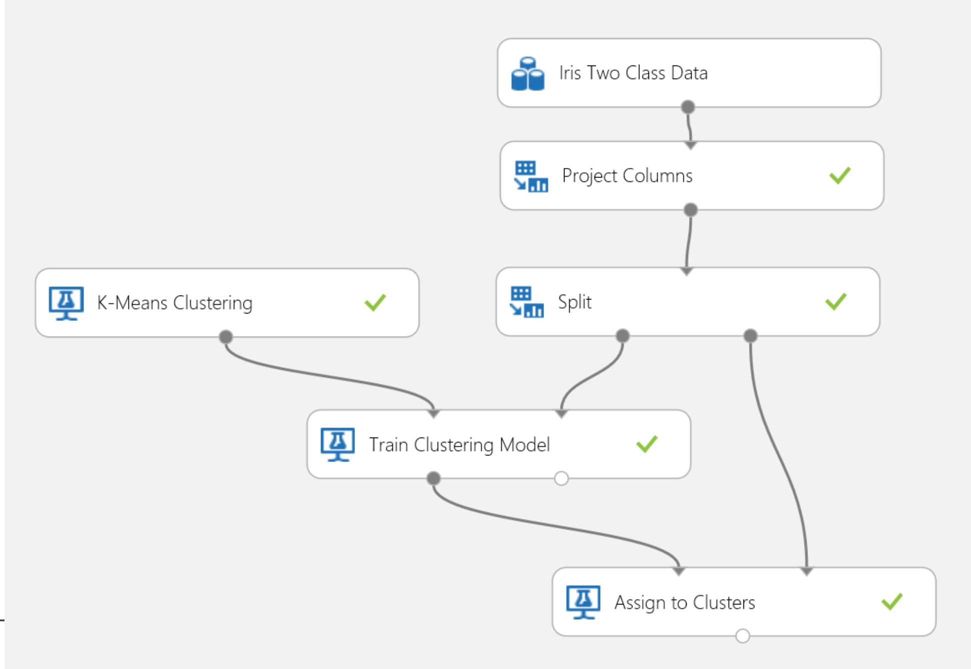
COL4: 1

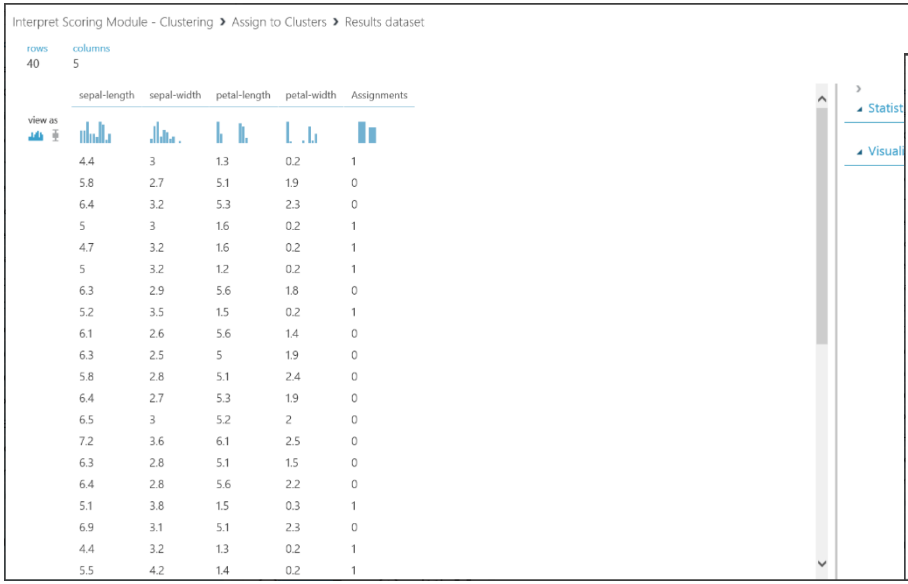
COL5: 5

COL6: 3

Part III Azure ML Exercises: Clustering

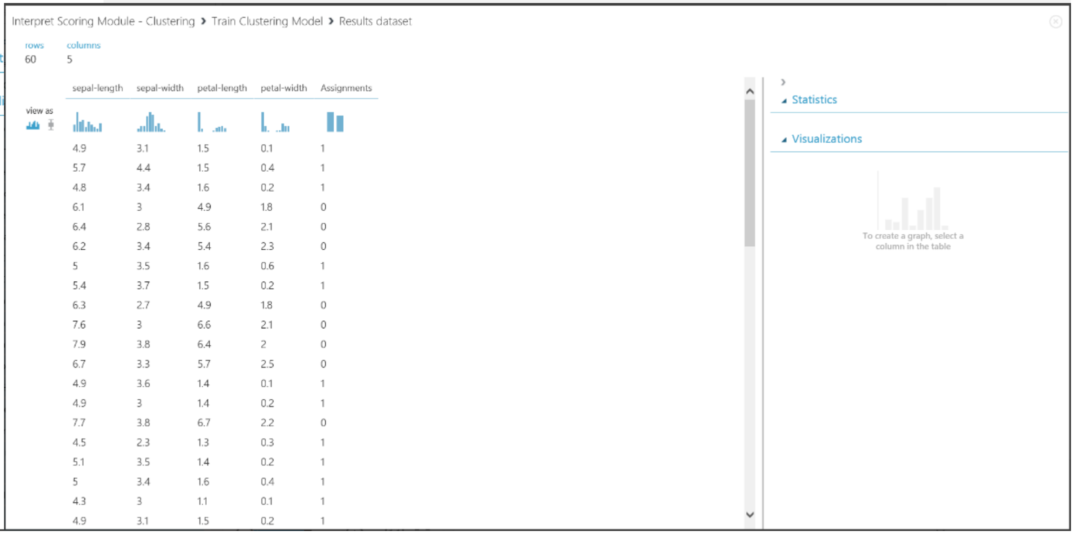
- Clustering differs from classification in that the training dataset does not have ground-truth labels by itself.
- Group the training dataset instances into distinct clusters.
- During the training process, the model labels the entries by leaning the differences between their features.


Iris Two Class Data
Project Columns ✓
K-Means Clustering ✓
Split ✓
Train Clustering Model ✓
Assign to Clusters ✓


Interpret Scoring Module - Clustering › Assign to Clusters › Results dataset

rows 40
columns 5

sepal-length	sepal-width	petal-length	petal-width	Assignments
4.4	3	1.3	0.2	1
5.8	2.7	5.1	1.9	0
6.4	3.2	5.3	2.3	0
5	3	1.6	0.2	1
4.7	3.2	1.6	0.2	1
5	3.2	1.2	0.2	1
6.3	2.9	5.6	1.8	0
5.2	3.5	1.5	0.2	1
6.1	2.6	5.6	1.4	0
6.3	2.5	5	1.9	0
5.8	2.8	5.1	2.4	0
6.4	2.7	5.3	1.9	0
6.5	3	5.2	2	0
7.2	3.6	6.1	2.5	0
6.3	2.8	5.1	1.5	0
6.4	2.8	5.6	2.2	0
5.1	3.8	1.5	0.3	1
6.9	3.1	5.1	2.3	0
4.4	3.2	1.3	0.2	1
5.5	4.2	1.4	0.2	1

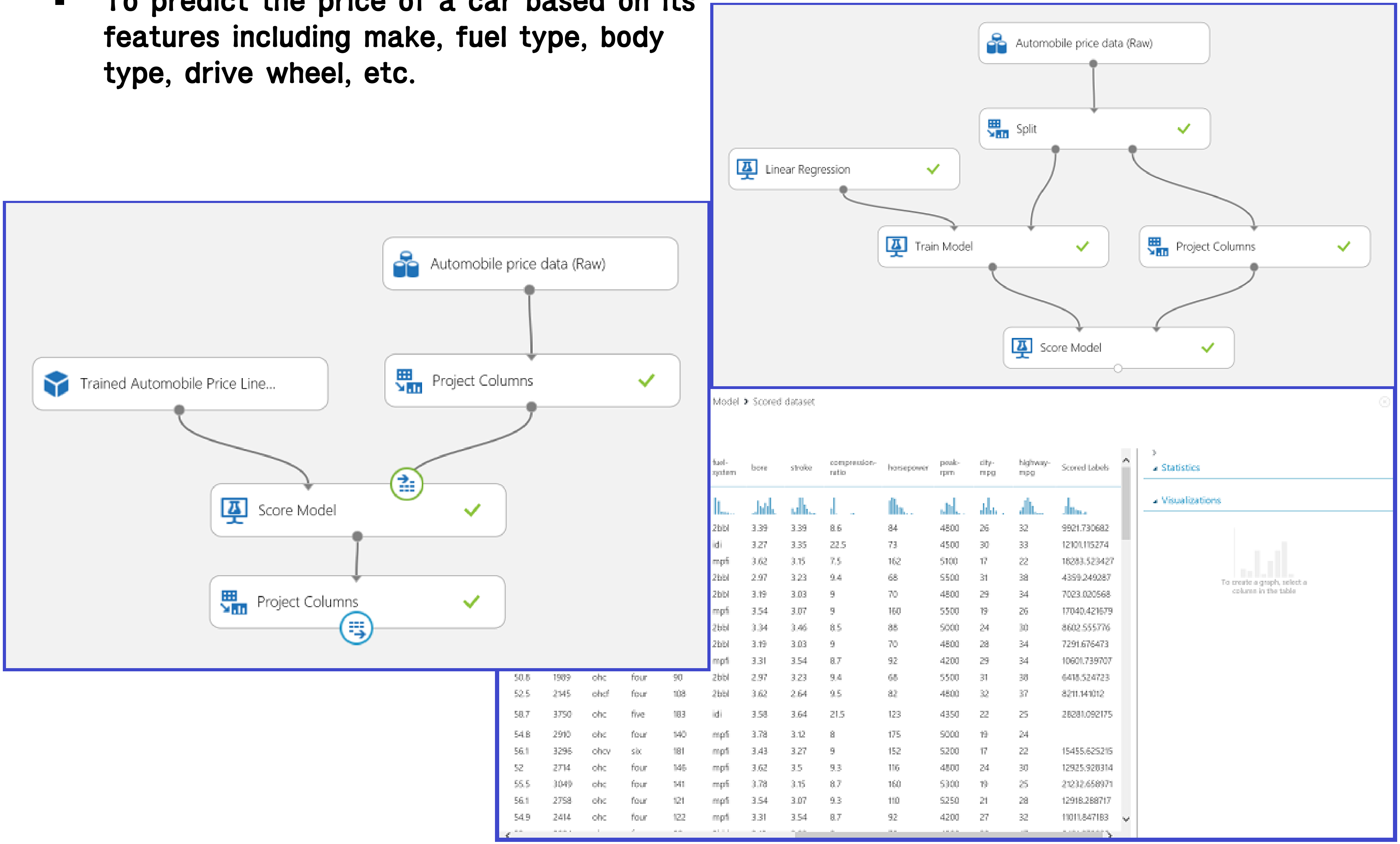

Interpret Scoring Module - Clustering › Train Clustering Model › Results dataset

rows 60
columns 5

sepal-length	sepal-width	petal-length	petal-width	Assignments
4.9	3.1	1.5	0.1	1
5.7	4.4	1.5	0.4	1
4.8	3.4	1.6	0.2	1
6.1	3	4.9	1.8	0
6.4	2.8	5.6	2.1	0
6.2	3.4	5.4	2.3	0
5	3.5	1.6	0.6	1
5.4	3.7	1.5	0.2	1
6.3	2.7	4.9	1.8	0
7.6	3	6.6	2.1	0
7.9	3.8	6.4	2	0
6.7	3.3	5.7	2.5	0
4.9	3.6	1.4	0.1	1
4.9	3	1.4	0.2	1
7.7	3.8	6.7	2.2	0
4.5	2.3	1.3	0.3	1
5.1	3.5	1.4	0.2	1
5	3.4	1.6	0.4	1
4.3	3	1.1	0.1	1
4.9	3.1	1.5	0.2	1

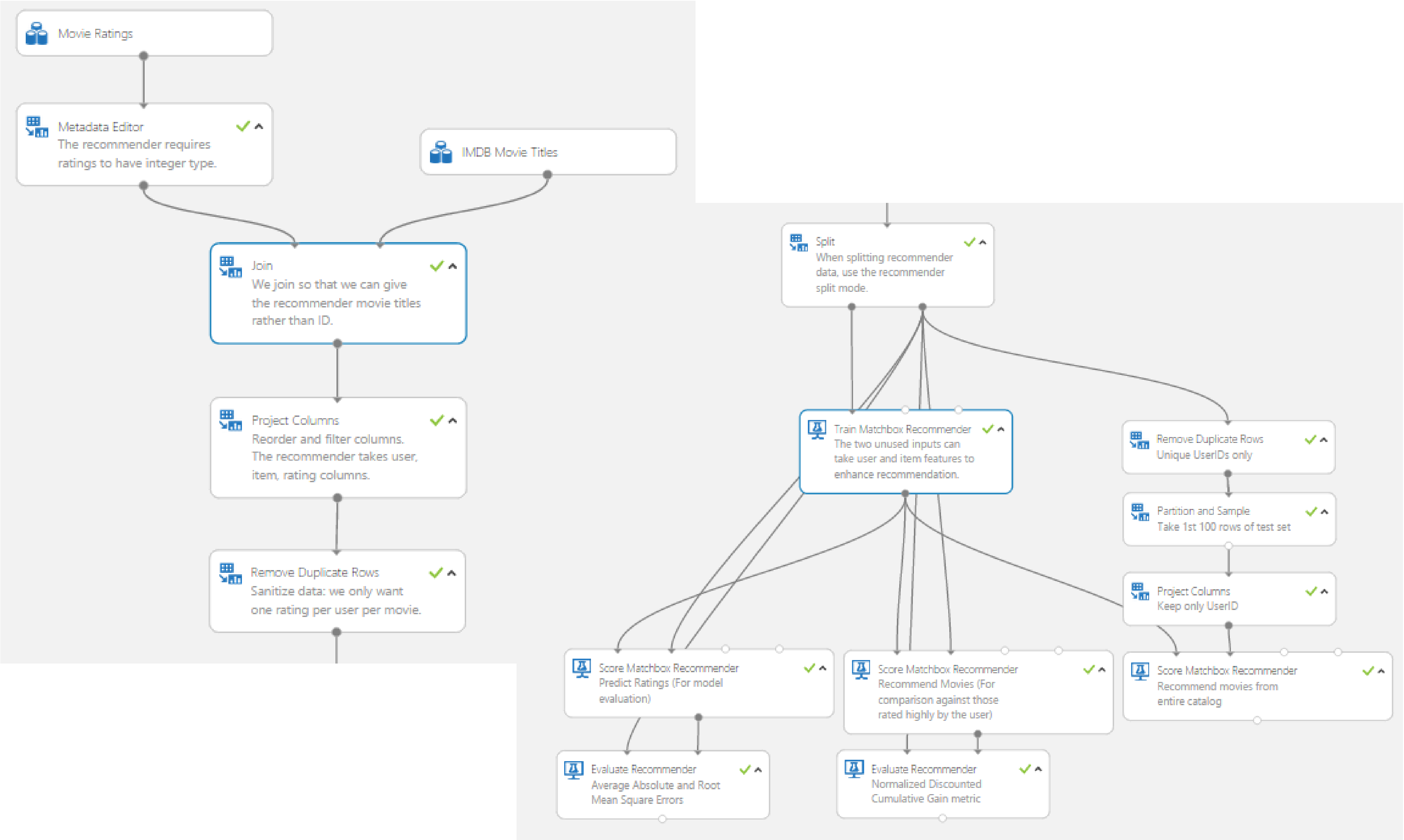
Part III Azure ML Exercises: Regression

- To predict the price of a car based on its features including make, fuel type, body type, drive wheel, etc.



					fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	Scored Labels
					2bbl	3.39	3.39	8.6	84	4800	26	32	9921.730682
					idi	3.27	3.35	22.5	73	4500	30	33	12101.115274
					mpfi	3.62	3.15	7.5	162	5100	17	22	18283.523427
					2bbl	2.97	3.23	9.4	68	5500	31	38	4359.249287
					2bbl	3.19	3.03	9	70	4800	29	34	7023.020568
					mpfi	3.54	3.07	9	160	5500	19	26	17040.421679
					2bbl	3.34	3.46	8.5	88	5000	24	30	8602.555776
					2bbl	3.19	3.03	9	70	4800	28	34	7291.676473
					mpfi	3.31	3.54	8.7	92	4200	29	34	10601.739707
50.8	1989	ohc	four	90	2bbl	2.97	3.23	9.4	68	5500	31	38	6418.524723
52.5	2145	ohcf	four	108	2bbl	3.62	2.64	9.5	82	4800	32	37	8211.141012
58.7	3750	ohc	five	183	idi	3.58	3.64	21.5	123	4350	22	25	28281.092175
54.8	2910	ohc	four	140	mpfi	3.78	3.12	8	175	5000	19	24	
56.1	3296	ohcv	six	181	mpfi	3.43	3.27	9	152	5200	17	22	15455.625215
52	2714	ohc	four	146	mpfi	3.62	3.5	9.3	116	4800	24	30	12925.928314
55.5	3049	ohc	four	141	mpfi	3.78	3.15	8.7	160	5300	19	25	21232.658971
56.1	2758	ohc	four	121	mpfi	3.54	3.07	9.3	110	5250	21	28	12918.288717
54.9	2414	ohc	four	122	mpfi	3.31	3.54	8.7	92	4200	27	32	11011.847183

Part III Azure ML Exercises: Recommendation



Movie Ratings

Metadata Editor
The recommender requires ratings to have integer type.

IMDB Movie Titles

Join
We join so that we can give the recommender movie titles rather than ID.

Project Columns
Reorder and filter columns. The recommender takes user, item, rating columns.

Remove Duplicate Rows
Sanitize data: we only want one rating per user per movie.

Split
When splitting recommender data, use the recommender split mode.

Train Matchbox Recommender
The two unused inputs can take user and item features to enhance recommendation.

Remove Duplicate Rows
Unique UserIDs only

Partition and Sample
Take 1st 100 rows of test set

Project Columns
Keep only UserID

Score Matchbox Recommender
Predict Ratings (For model evaluation)

Score Matchbox Recommender
Recommend Movies (For comparison against those rated highly by the user)

Score Matchbox Recommender
Recommend movies from entire catalog

Evaluate Recommender
Average Absolute and Root Mean Square Errors

Evaluate Recommender
Normalized Discounted Cumulative Gain metric

Part IV Lessons and Resources

1. Data wrangling is important
   a. More time is spent on data wrangling than model building.
   b. Different sources, formats, schemas, missing values and noisy data.
   c. Data manipulation modules are very popular: Execute R script, SQL Transform, etc.

2. Azure ML Modeling
   a. Modeling depends on the Business Application Domain and Data.
   b. Feature Engineering is essential.
   c. Parameter Tuning is needed.
   d. Learn R or Python Script.

3. Resources
   a. Getting Started: https://studio.azureml.net
   b. Gallery: http://gallery.azureml.net/
   c. Site/ML Studio/Docs: http://azure.microsoft.com/en-us/services/machine-learning/
   d. Blog: http://blogs.technet.com/b/machinelearning/
   e. edX: Microsoft DAT203X Data Science and Machine Learning Essentials https://courses.edx.org/courses/course-v1:Microsoft+DAT203x+3T2015/info
   f. Microsoft Virtual Academy: https://mva.microsoft.com/

Q and A