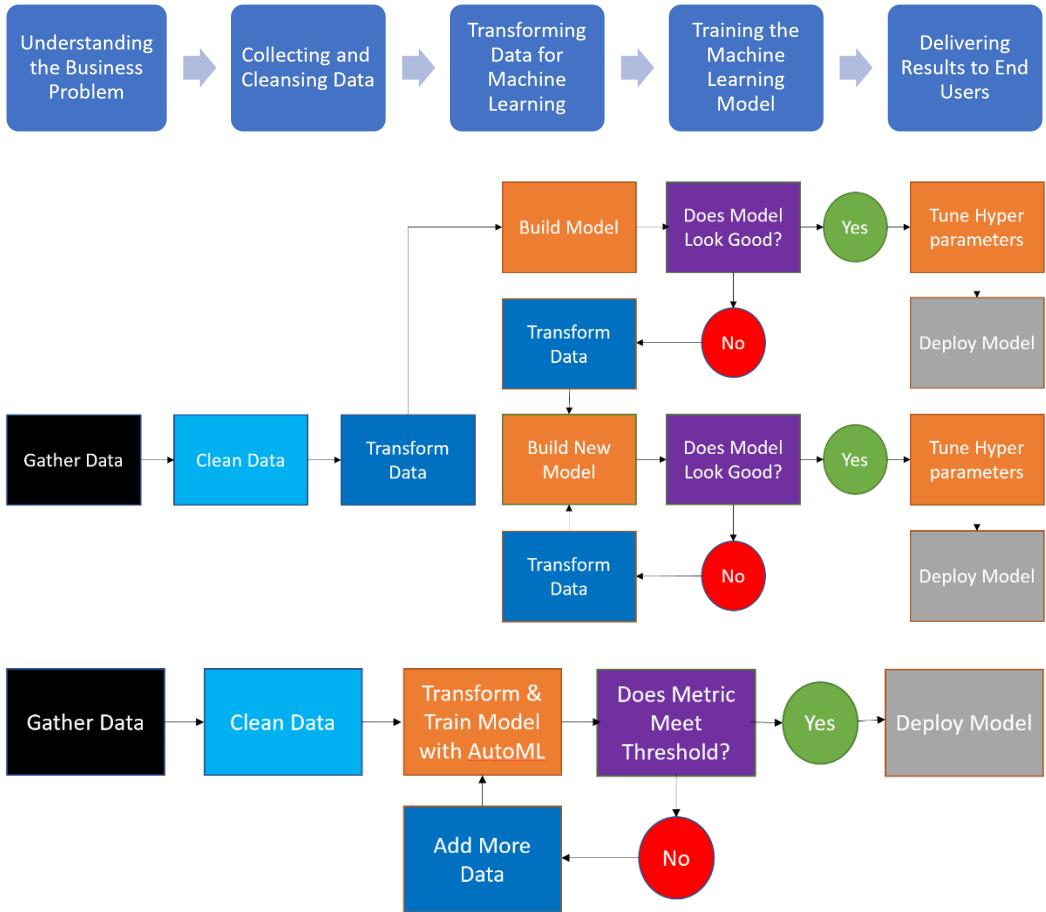
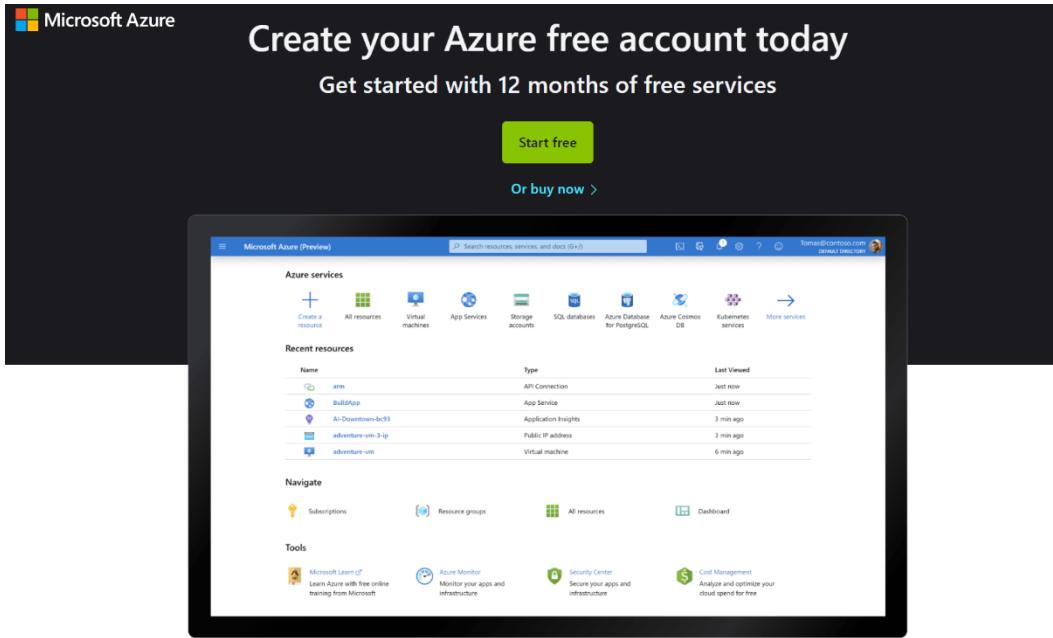


Chapter 1: Introducing AutoML



Chapter 2: Getting Started with Azure Machine Learning Service



Microsoft Azure (Preview)

Search resources, services, and docs

Create a resource

Azure services



Create a
resource

Machine
Learning

Resource
groups

Microsoft Azure (Preview)

Search resources, services, and docs

Home >

New



Machine Learning

Azure Marketplace

See all

Popular

Microsoft Azure (Preview)

Search resources, services, and docs

Home > New >

Machine Learning



Microsoft



Machine Learning

Save for later

Microsoft

Azure benefit eligible

Create

Basics Networking Advanced Tags Review + create

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ

Dennis Sawyers Internal MS Learning Account

Resource group * ⓘ

auto-ml-example-resource-group

[Create new](#)

Workspace details

Specify the name and region for the workspace.

Workspace name * ⓘ

automl-example-workspace

Region * ⓘ

North Central US

Storage account * ⓘ

(new) automlexamplew1005528502

[Create new](#)

Key vault * ⓘ

(new) automlexamplew3892816466

[Create new](#)

Application insights * ⓘ

(new) automlexamplew6936691182

[Create new](#)

Container registry * ⓘ

None

[Create new](#)

Microsoft Azure Machine Learning

ml-teaching-workspace > Home

Welcome to the studio!

Create new 

Notebooks 

Code with Python SDK and run sample experiments.

Start now

My recent resources

Runs				
Run	Run ID	Experiment	Status	Submit
Run 28	f5118fd2...	Titanic-Sc...	Completed	Sep 7, 2

Microsoft Azure Machine Learning

ml-teaching-workspace > Compute

Compute

[Compute instances](#) [Compute clusters](#) [Inference clusters](#) [Attached compute](#)

(i) In the wake of COVID-19, we are prioritizing maintaining service availability for first responders.

[New](#) [Refresh](#) [Start](#) [Stop](#) [Restart](#) [Delete](#) [View quota](#) [More](#)

Name	Status	Application URI
standard-compute3	Running	JupyterLab Jupyter RStudio
gurobi-test	Stopped	JupyterLab Jupyter RStudio

Select virtual machine

Select the virtual machine size you would like to use for your compute instance. Please note that only the creator can access the compute instance by default. Alternatively, you can assign the access to another user in the advanced settings section.

Region [?](#)
northcentralus

Virtual machine type [?](#)
 CPU GPU

Virtual machine size [?](#)
 Select from recommended options Select from all options

Total available quota: 600 cores [?](#)

Name ↑	Category	Workload types	Av...	Cost
<input type="radio"/> Standard_DS2_v2 2 cores, 7GB RAM, 14GB storage	General purpose	Development on Notebooks (or other IDE) and light weight testing	100 c...	\$0.15/hr
<input checked="" type="radio"/> Standard_DS3_v2 4 cores, 14GB RAM, 28GB storage	General purpose	Classical ML model training, AutoML runs, pipeline runs (default compute)	100 c...	\$0.29/hr
<input type="radio"/> Standard_DS12_v2 4 cores, 28GB RAM, 56GB storage	Memory optimized	Training on large datasets (>1GB) parallel run steps, batch inferencing	100 c...	\$0.37/hr
<input type="radio"/> Standard_F4s_v2 4 cores, 8GB RAM, 32GB storage	Compute optimized	Real-time inferencing and other latency-sensitive tasks	100 c...	\$0.17/hr

Configure Settings

Configure compute instance settings for your selected virtual machine size.

Name	Category	Cores	Available quota	RAM	Storage	Cost/Hour
Standard_DS3_v2	General purpose	4	100 cores	14 GB	28 GB	\$0.29/hr

Compute name * ⓘ

automl-example-compute

Enable SSH access ⓘ

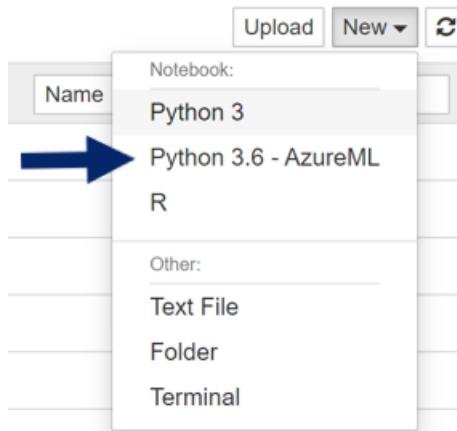
> Show advanced settings

Compute

Compute instances Compute clusters Inference clusters Attached compute

+ New ⏪ Refresh ⏴ Start ⏵ Stop ⏴ Restart ⏴ Delete ⏴ View quota Show all

Name	Status	Application URI
automl-example-comp...	Running	JupyterLab Jupyter VS Code RStudio Terminal



Compute

Compute instances

Compute clusters

Inference clusters

Attached compute



Scale your compute cluster from a single node to a multi node workload

Create a single or multi node compute cluster for your training, batch inferencing or reinforcement learning workloads. [Learn more](#)



[View Azure Machine Learning tutorials](#)

Select virtual machine

Select the virtual machine size you would like to use for your compute cluster.

Region i

northcentralus

Virtual machine priority i

Dedicated Low priority

Virtual machine type i

CPU GPU

Virtual machine size i

Select from recommended options Select from all options

Total available quota: 596 cores i

Name ↑	Category	Workload types	Av... i	Cost i
<input type="radio"/> Standard_DS2_v2 2 cores, 7GB RAM, 14GB storage	General purpose	Development on Notebooks (or other IDE) and light weight testing	96 co...	\$0.15/hr
<input checked="" type="radio"/> Standard_DS3_v2 4 cores, 14GB RAM, 28GB storage	General purpose	Classical ML model training, AutoML runs, pipeline runs (default compute)	96 co...	\$0.29/hr
<input type="radio"/> Standard_DS12_v2 4 cores, 28GB RAM, 56GB storage	Memory optimized	Training on large datasets (>1GB) parallel run steps, batch inferencing	96 co...	\$0.37/hr
<input type="radio"/> Standard_F4s_v2 4 cores, 8GB RAM, 32GB storage	Compute optimized	Real-time inferencing and other latency-sensitive tasks	100 c...	\$0.17/hr

Back

Next

Download a template for automation

Cancel

Configure Settings

Configure compute cluster settings for your selected virtual machine size.

Name	Category	Cores	Available quota	RAM	Storage	Cost/Hour
Standard_DS3_v2	General purpose	4	96 cores	14 GB	28 GB	\$0.29/hr

Compute name * ⓘ

Minimum number of nodes * ⓘ

Maximum number of nodes * ⓘ

Idle seconds before scale down * ⓘ

Enable SSH access ⓘ

[Advanced settings](#)

[Back](#)

[Create](#)

[Download a template for automation](#)

[Cancel](#)

The screenshot shows the left sidebar of the Azure Machine Learning Studio. It includes sections for 'Author' (Notebooks, Automated ML, Designer), 'Assets' (Datasets, Experiments, Pipelines, Models, Endpoints), and 'Manage' (Compute, Datastores, Data Labeling). The 'Datasets' option under 'Assets' is highlighted with a grey background.

Registered datasets

Dataset monitors (preview)



Register datasets to manage, machine learning workflows.

With Azure Machine Learning datasets, your storage referenced by datasets an training without worrying about conne

+ Create dataset ▾

- From local files
- From datastore
- From web files
- From Open Datasets

tutorials

Microsoft Azure Machine Learning

The screenshot shows the Microsoft Azure Machine Learning interface. On the left, there's a navigation sidebar with options like New, Home, Author, Notebooks, Automated ML (preview), Designer (preview), Assets, Datasets, Experiments, Pipelines, Models, Endpoints, Compute, Datastores, and Data Labeling. The 'Datasets' option is currently selected. In the main area, under 'Datasets', there's a section for 'Registered datasets' with a 'Create dataset' button. A success message 'Success: Diabetes_Sample data' is displayed above the dataset list. The list includes 'Diabetes_Sample', 'Titanic_To_Score_v2', 'Titanic_To_Score', 'Titanic_Training_Local', 'titanic_training_transformed', and 'titanic_training_raw'. To the right, a modal window titled 'Create dataset from Open Datasets' is open, with a sub-section 'Select Open Dataset'. It says 'Azure Open Datasets offers ML ready data from the open domain. open data in your experiments from a common storage location'. Below this is a search bar with 'Diabetes' typed in. A preview box for 'Sample: Diabetes' is shown, containing a checkmark icon, the dataset name, and a brief description: 'The Diabetes dataset has 442 samples with 10 features, making it ideal for getting started with machine learni...'. There's also a 'Learn more' link. At the bottom of the modal, there's a 'Dataset details' section with fields for 'Name *' (containing 'Diabetes Sample') and 'Dataset version' (containing '1').

The screenshot shows the Azure Machine Learning studio interface. On the left, there is a sidebar with the following navigation items:

- New
- Home
- Author
- Notebooks
- Automated ML (preview)
- Designer (preview)
- Assets
- Datasets
- Experiments
- Pipelines
- Models
- Endpoints
- Manage
- Compute
- Datastores

The main area is titled "Diabetes Sample" and shows "Version 1 (latest)". Below the title, there are tabs: Details, Consume (which is selected), Explore, and Models. There are also buttons for Refresh, Generate profile, Unregister, and New version.

In the "Sample usage" section, there is a code snippet:

```
# azureml-core of version 1.0.72 or higher is required
# azureml-datatransform[pandas] of version 1.1.34 or higher is required
from azureml.core import Workspace, Dataset

subscription_id = '47a7ec0c-37ad-428b-9114-b87ea1057632'
resource_group = 'ml-teaching'
workspace_name = 'ml-teaching-workspace'

workspace = Workspace(subscription_id, resource_group, workspace_name)

dataset = Dataset.get_by_name(workspace, name='Diabetes Sample')
dataset.to_pandas_dataframe()
```



Chapter 3: Training Your First AutoML Model

Register datasets to manage, share, and track data in your machine learning workflows.

With Azure Machine Learning datasets, you can keep a single copy of data in your storage referenced by datasets and seamlessly access data during model training without worrying about connection strings or data paths. [Learn more](#)

The screenshot shows the 'Create dataset' menu open, with two blue arrows pointing to the 'From local files' option. Below it, the 'Create dataset from local files' dialog is displayed. A vertical navigation bar on the left lists steps: 'Basic info' (selected), 'Datastore and file selection', 'Settings and preview', 'Schema', and 'Confirm details'. The 'Basic info' section contains fields: 'Name' (Titanic Training Data), 'Dataset type' (Tabular), and 'Description' (Titanic data containing passenger demographic information and ticket information. We will use this data to model who survived the Titanic voyage.).

Create dataset from local files

Basic info

Name * Dataset version

Dataset type *

Description

Basic info
Datastore and file selection
Settings and preview
Schema
Confirm details

Datastore and file selection

Select or create a datastore *

- Currently selected datastore: workspaceblobstore (Azure Blob Storage) (Default)
- Previously created datastore
- Create new datastore

Select files for your dataset *

After dataset creation, these files will be uploaded to your default Blob storage and made available in your workspace. Supported file types include: delimited (i.e. csv, tsv), Parquet, JSON Lines, and plain text.

Browse

1 files selected. Total size 0.05836 MiB. 0/1 files uploaded

File name	Size (MiB)	Upload %	Status
titanic.csv	0.05836		

Upload path

/titanic/train | Files will be uploaded to '\$(Upload path)/09-29-2020_123346_UTC'

Skip data validation ⓘ

Settings and preview

These settings were automatically detected. Please verify that the selections were made correctly or update

File format Delimited

Delimiter Comma Example Field1,Field2,Field3

Encoding UTF-8

Column headers Use headers from the first file

Skip rows None

Basic info

Datastore and file selection

Settings and preview

Schema

Confirm details

Path Not applicable to selected type String

PassengerId Not applicable to selected type Integer

Survived Not applicable to selected type Boolean

Pclass Not applicable to selected type Integer

Name Not applicable to selected type String

Sex Not applicable to selected type String

Confirm details

Basic info	Datastore and file selection
Name Titanic Training Data	Datastore workspaceblobstore
Dataset version 1	Selected files (1) titanic.csv
Dataset type Tabular	Path titanic/train/09-29-2020_123346.UTC/titanic.csv
Description Titanic data containing passenger demographic information and ticket information. We will use this data to model who survived the Titanic voyage.	

File settings
File format Delimited
Delimiter Comma
Encoding UTF-8
Column headers Use headers from the first file
Skip rows None

Profile this dataset after creation

The screenshot shows the Azure Machine Learning studio interface. On the left, a sidebar menu includes options like 'New', 'Home', 'Author', 'Notebooks', 'Automated ML' (which is selected and highlighted in grey), and 'Designer'. A blue arrow points from the 'Automated ML' menu item to the 'New Automated ML run' button. The main area is titled 'Automated ML' with the sub-instruction 'Let Automated ML train and find the best model'. Below this, a section titled 'Create a new Automated ML run' contains three steps: 'Select dataset' (which is completed, indicated by a solid blue circle), 'Configure run', and 'Task type and settings'. Another blue arrow points from the 'Task type and settings' step to the 'Dataset name' table. The table has two columns: 'Dataset name' and 'Dataset type'. It shows one entry: 'Titanic Training Data' under 'Dataset name' and 'Tabular' under 'Dataset type'. There is also a 'Create dataset' button and a 'Show supported datasets only' toggle switch.

Automated ML

Let Automated ML train and find the best model

New Automated ML run

Create a new Automated ML run

Select dataset

Select a dataset from the list below, or create a new dataset. Automated ML

+ Create dataset | Show supported datasets only

Dataset name	Dataset type
Titanic Training Data	Tabular

Select dataset

Configure run

Task type and settings

Configure run

Configure the experiment. Select from existing experiments or define a new name, select the target column and the training compute [experiment](#)

Dataset

Titanic Training Data ([View dataset](#))

Experiment name *



Select existing



Create new

New experiment name



Titanic-Training

Target column *

Survived



Select compute cluster *

compute-cluster



Create a new compute Refresh compute

Select task type

Select the machine learning task type for the experiment. Additional settings are available to fine tune the experiment if needed.

Classification

To predict one of several categories in the target column: yes/no, blue, red, green.



Enable deep learning

Regression

To predict continuous numeric values

Time series forecasting

To predict values based on time

 View additional configuration settings

 View featurization settings



Additional configurations

Primary metric

Accuracy

Explain best model

Blocked algorithms

A list of algorithms that Automated ML will not use during training.

Exit criterion

Training job time (hours)



0.25

Metric score threshold

Metric score threshold

Metric score threshold

Validation

Concurrency

 View featurization settings

Automated ML

Let Automated ML train and find the best model based on your data without writing code.

New Automated ML run

Recent Automated ML runs

Run	Run ID	Experiment
Run 1	AutoML_ca2d710d-6d3a-4f90-bcc8-25b...	Titanic-Training

Details Data guardrails **Models** Outputs + logs Child runs Snapshot

Deploy Download Explain model

Algorithm name	Explained	Accuracy ↓
VotingEnsemble	View explanation	0.83835
StandardScalerWrapper, XGBoostClassifier		0.82713
StandardScalerWrapper, XGBoostClassifier		0.82600

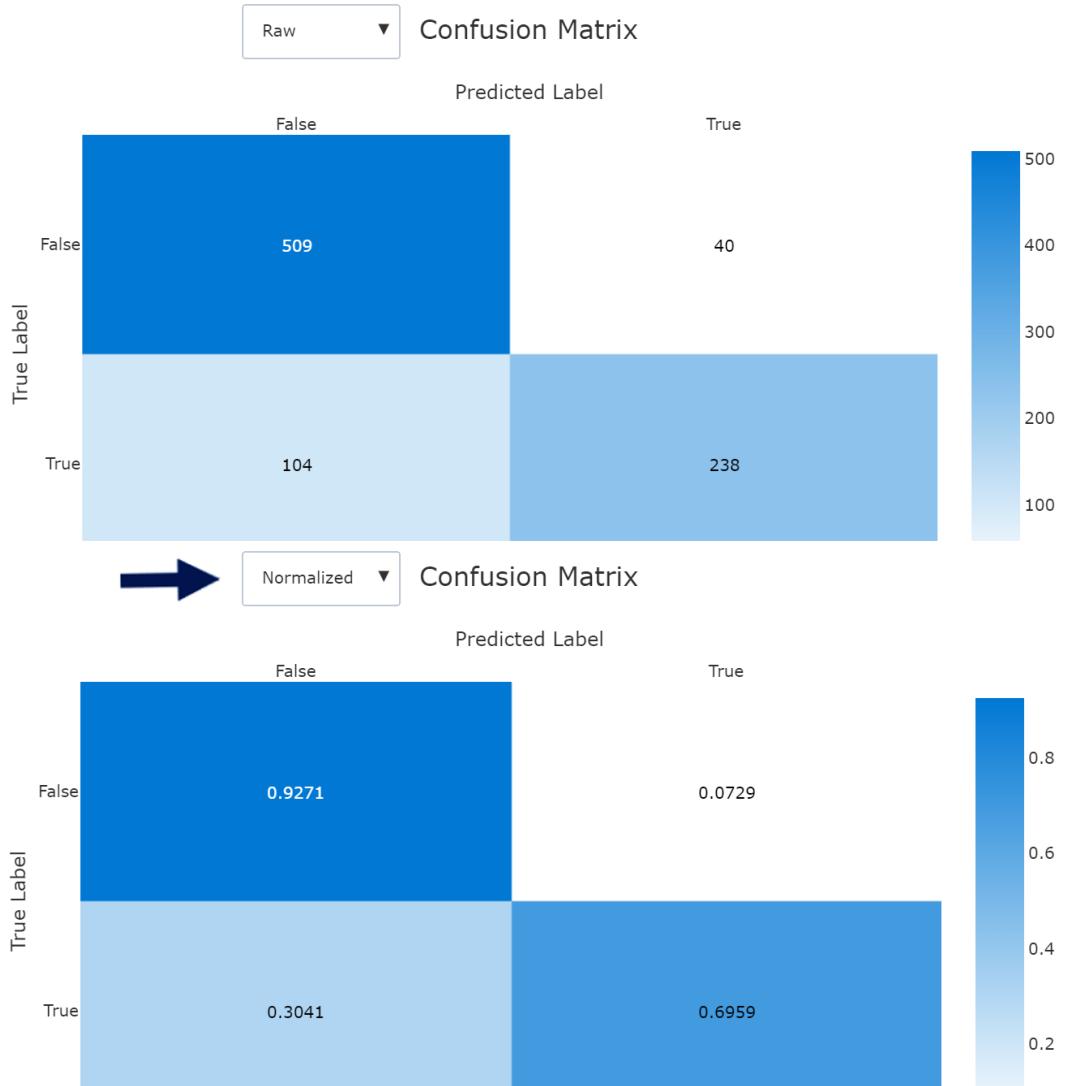
Details Model Explanations (preview) Metrics Outputs + logs Images Child runs Snapshot

Select a metric to see a visualization or table of the data.

View as: Chart Table

accuracy	matthews_correlation
0.8383520599250938	0.6540526267644078

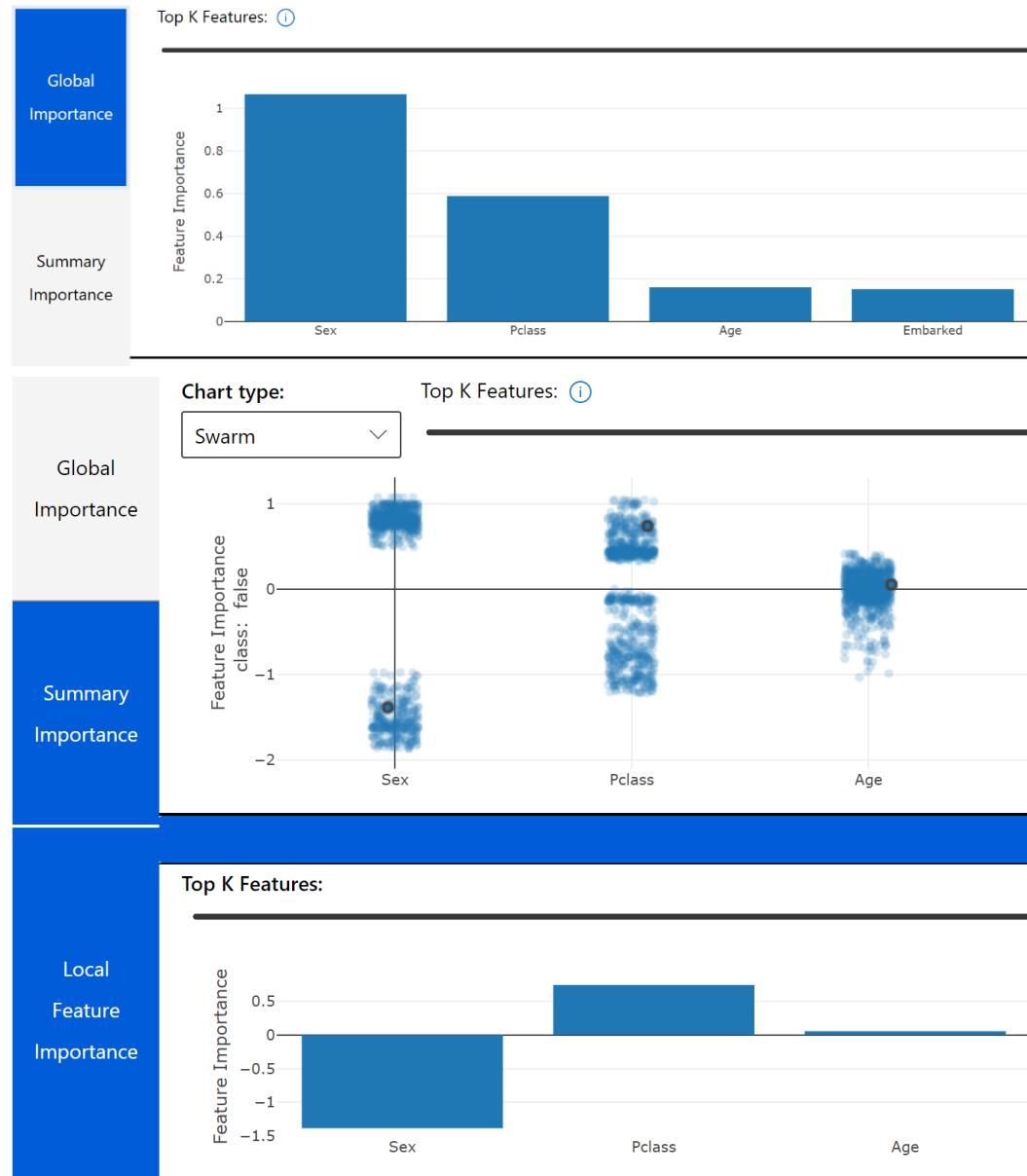
accuracy
 accuracy_table



Select Explanation

tabular | mimic.lightgbm | raw | classification | 4381fdb3-8165-4f04-8bcc-8e77d2e41987 | 9/29/2020, 10:39:44 PM

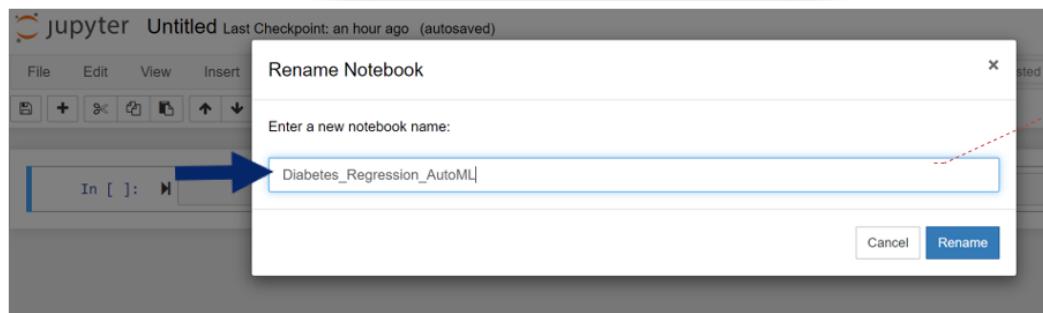
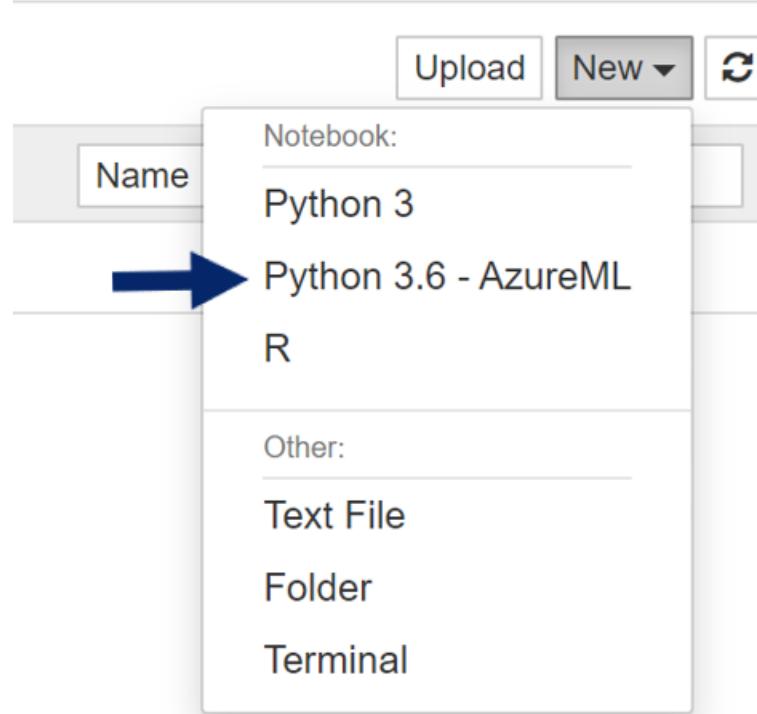
Explainer: mimic.lightgbm



Chapter 4: Building an AutoML Regression Solution

Standard regression algorithms	Tree algorithms	Gradient boosting algorithms	Nearest neighbor algorithms	Optimization algorithms
Elastic net	Decision tree	XGBoost	KNN	SGD
LARS Lasso	Random forest Extremely randomized trees	LightGBM Gradient boosting		Online gradient descent regressor Fast linear regressor

The screenshot shows the AutoML workspace interface. On the left is a sidebar with navigation links: New, Home, Author, Notebooks, Automated ML, Designer, Assets, Datasets, Experiments, Pipelines, Models, Endpoints, Manage, **Compute** (which is selected and highlighted in grey), Datastores, and Data Labeling. The main area is titled "automl-example-workspace > Compute" and contains a "Compute" section. It has tabs for Compute instances, Compute clusters, Inference clusters, and Attached compute. Under Compute instances, there is a table with one row. The table columns are Name, Status, and Application URI. The row shows "titanic-compute-instance" with a status of "Running". To the right of the table are buttons for Jupyter, RStudio, and SSH. A large blue arrow points from the "Running" status to the Jupyter button.



```
In [28]: # View your dataset by converting to pandas  
dataset.take(10).to_pandas_dataframe()
```

Out[28]:

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
0	59	2	32.1	101.0	157	93.2	38.0	4.00	4.8598	87	151
1	48	1	21.6	87.0	183	103.2	70.0	3.00	3.8918	69	75
2	72	2	30.5	93.0	156	93.6	41.0	4.00	4.6728	85	141
3	24	1	25.3	84.0	198	131.4	40.0	5.00	4.8903	89	206
4	50	1	23.0	101.0	192	125.4	52.0	4.00	4.2905	80	135
5	23	1	22.6	89.0	139	64.8	61.0	2.00	4.1897	68	97
6	36	2	22.0	90.0	160	99.6	50.0	3.00	3.9512	82	138
7	66	2	26.2	114.0	255	185.0	56.0	4.55	4.2485	92	63
8	60	2	32.1	83.0	179	119.4	42.0	4.00	4.4773	94	110
9	29	1	30.0	85.0	180	93.4	43.0	4.00	5.3845	88	310

DATA GUARDRAILS:

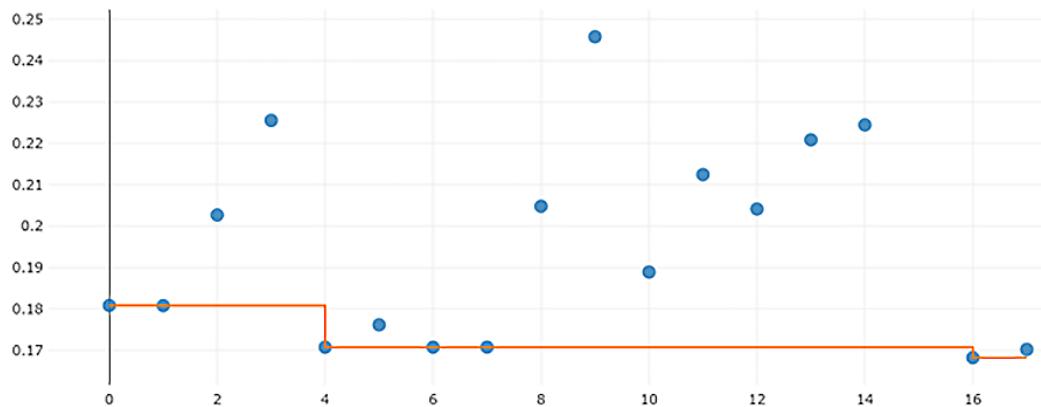
TYPE: Missing feature values imputation
STATUS: PASSED
DESCRIPTION: No feature missing values were detected in the training data.
Learn more about missing value imputation: <https://aka.ms/AutomatedMLFeaturization>

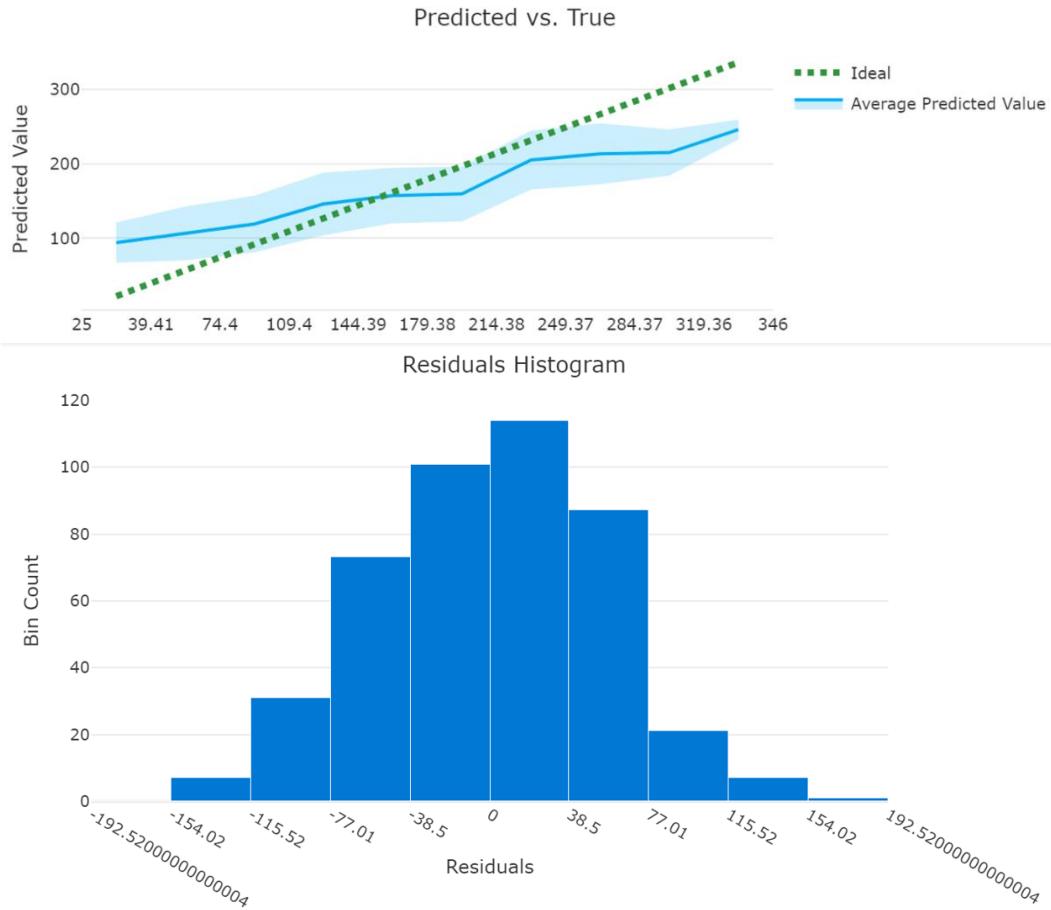
TYPE: High cardinality feature detection
STATUS: PASSED
DESCRIPTION: Your inputs were analyzed, and no high cardinality features were detected.
Learn more about high cardinality feature handling: <https://aka.ms/AutomatedMLFeaturization>

```
*****
ITERATION: The iteration being evaluated.
PIPELINE: A summary description of the pipeline being evaluated.
DURATION: Time taken for the current iteration.
METRIC: The result of computing score on the fitted pipeline.
BEST: The best observed score thus far.
*****
```

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:33	0.1808	0.1808
1	MaxAbsScaler XGBoostRegressor	0:00:25	0.1808	0.1808
2	MaxAbsScaler DecisionTree	0:00:30	0.2027	0.1808
3	MinMaxScaler DecisionTree	0:00:31	0.2255	0.1808
4	StandardScalerWrapper LassoLars	0:00:29	0.1708	0.1708
5	RobustScaler ElasticNet	0:00:35	0.1761	0.1708
6	StandardScalerWrapper LassoLars	0:00:32	0.1708	0.1708
7	RobustScaler ElasticNet	0:00:36	0.1708	0.1708
8	RobustScaler DecisionTree	0:00:31	0.2048	0.1708
9	MinMaxScaler DecisionTree	0:00:30	0.2458	0.1708
10	RobustScaler DecisionTree	0:00:31	0.1889	0.1708
11	MinMaxScaler DecisionTree	0:00:32	0.2125	0.1708
12	StandardScalerWrapper DecisionTree	0:00:29	0.2041	0.1708
13	StandardScalerWrapper DecisionTree	0:02:19	0.2209	0.1708
14	RobustScaler DecisionTree	0:00:31	0.2244	0.1708
15		0:00:08	nan	0.1708
16	VotingEnsemble	0:00:52	0.1682	0.1682
17	StackEnsemble	0:00:52	0.1702	0.1682

AutoML Run with metric : normalized_root_mean_squared_error





(25.0, 75.128]
 (75.128, 152.133]
 (152.133, 229.139]
 (229.139, 346.0]

Chapter 5: Building an AutoML Classification Solution

The screenshot shows the Azure Machine Learning Studio interface. On the left, a sidebar navigation menu includes options like New, Home, Notebooks, Automated ML, Designer, Datasets, Experiments, Pipelines, Models, Endpoints, Compute (which is selected), Datastores, and Data Labeling. The main workspace is titled "automl-example-workspace > Compute". The "Compute" section displays a table of compute instances:

Name	Status	Application URI
titanic-compute-instance	Running	Jupyter RStudio SSH

A blue arrow points from the "Compute" table to the "Application URI" column. Below the workspace, a Jupyter Notebook window is open with the title "jupyter Untitled Last Checkpoint: a few seconds ago (unsaved changes)". A "Rename Notebook" dialog box is overlaid on the notebook, containing a text input field with the value "Titanic_Classification_AutoML". A blue arrow points from the "Rename Notebook" dialog to the input field. The dialog also includes "Cancel" and "Rename" buttons.

```
In [7]: # View your dataset by converting to pandas  
dataset.take(10).to_pandas_dataframe()
```

Out[7]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	False	3	male	22.0	1	0	7.2500	S
1	True	1	female	38.0	1	0	71.2833	C
2	True	3	female	26.0	0	0	7.9250	S
3	True	1	female	35.0	1	0	53.1000	S
4	False	3	male	35.0	0	0	8.0500	S
5	False	3	male	NaN	0	0	8.4583	Q
6	False	1	male	54.0	0	0	51.8625	S
7	False	3	male	2.0	3	1	21.0750	S
8	True	3	female	27.0	0	2	11.1333	S
9	True	2	female	14.0	1	0	30.0708	C

DATA GUARDRAILS:

TYPE: Class balancing detection
STATUS: PASSED
DESCRIPTION: Your inputs were analyzed, and all classes are balanced in your training data.
Learn more about imbalanced data: <https://aka.ms/AutomatedMLImbalancedData>

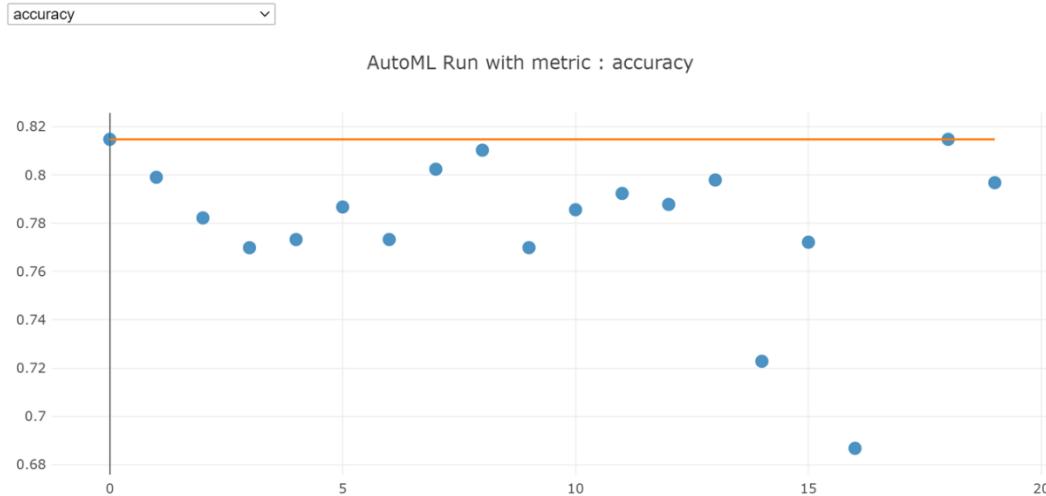
TYPE: Missing feature values imputation
STATUS: DONE
DESCRIPTION: If the missing values are expected, let the run complete. Otherwise cancel the current run and use a script to customize the handling of missing feature values that may be more appropriate based on the data type and business requirements.
Learn more about missing value imputation: <https://aka.ms/AutomatedMLFeaturization>

DETAILS:
+-----+-----+
|Column name |Missing value count |
+=====+=====+
|Embarked |2 |
+-----+-----+

TYPE: High cardinality feature detection
STATUS: PASSED
DESCRIPTION: Your inputs were analyzed, and no high cardinality features were detected.
Learn more about high cardinality feature handling: <https://aka.ms/AutomatedMLFeaturization>

```
*****
ITERATION: The iteration being evaluated.
PIPELINE: A summary description of the pipeline being evaluated.
DURATION: Time taken for the current iteration.
METRIC: The result of computing score on the fitted pipeline.
BEST: The best observed score thus far.
*****
```

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:33	0.8148	0.8148
1	MaxAbsScaler XGBoostClassifier	0:00:29	0.7991	0.8148
2	MaxAbsScaler RandomForest	0:00:28	0.7822	0.8148
3	MaxAbsScaler RandomForest	0:00:30	0.7699	0.8148
4	MaxAbsScaler SGD	0:00:29	0.7733	0.8148
5	MaxAbsScaler SGD	0:00:33	0.7868	0.8148
6	MaxAbsScaler ExtremeRandomTrees	0:00:34	0.7733	0.8148
7	MaxAbsScaler ExtremeRandomTrees	0:00:29	0.8024	0.8148
8	MaxAbsScaler ExtremeRandomTrees	0:00:29	0.8183	0.8148
9	MaxAbsScaler ExtremeRandomTrees	0:00:31	0.7699	0.8148
10	MaxAbsScaler SGD	0:00:32	0.7856	0.8148
11	MaxAbsScaler SGD	0:00:33	0.7924	0.8148
12	MaxAbsScaler RandomForest	0:00:32	0.7878	0.8148
13	StandardScalerWrapper ExtremeRandomTrees	0:00:31	0.7979	0.8148
14	MaxAbsScaler RandomForest	0:00:29	0.7228	0.8148
15	MaxAbsScaler SGD	0:00:32	0.7722	0.8148
16	MaxAbsScaler RandomForest	0:00:33	0.6868	0.8148
17		0:00:08	nan	0.8148
18	VotingEnsemble	0:00:53	0.8148	0.8148
19	StackEnsemble	0:00:56	0.7968	0.8148



```
In [8]: # View your dataset by converting to pandas  
dataset.take(10).to_pandas_dataframe()
```

Out[8]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
ITERATION	PIPELINE		DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM		0:00:42	0.9467	0.9467
1	MaxAbsScaler XGBoostClassifier		0:00:49	0.9467	0.9467
2	MinMaxScaler RandomForest		0:00:55	0.9733	0.9733
3	MinMaxScaler RandomForest		0:00:40	0.9467	0.9733
4	MinMaxScaler RandomForest		0:00:49	0.9400	0.9733
5	MinMaxScaler SVM		0:00:52	0.9467	0.9733
6	MaxAbsScaler GradientBoosting		0:00:42	0.9533	0.9733
7	SparseNormalizer XGBoostClassifier		0:00:43	0.9200	0.9733
8	StandardScalerWrapper RandomForest		0:00:49	0.9533	0.9733
9	SparseNormalizer LightGBM		0:00:45	0.9733	0.9733
10	StandardScalerWrapper LightGBM		0:00:39	0.9400	0.9733
11	RobustScaler ExtremeRandomTrees		0:00:39	0.9533	0.9733
12	SparseNormalizer XGBoostClassifier		0:00:42	0.9667	0.9733
13	SparseNormalizer XGBoostClassifier		0:00:45	0.9067	0.9733
14	SparseNormalizer XGBoostClassifier		0:00:43	0.9133	0.9733
15	RobustScaler KNN		0:00:51	0.9533	0.9733
16	VotingEnsemble		0:00:57	0.9867	0.9867
17	StackEnsemble		0:01:05	0.9867	0.9867

Iteration	Pipeline	Iteration metric	Best metric	Status	Duration	Started
16	VotingEnsemble	0.98666667	0.98666667	Completed	0:01:07	Jan 18, 2021 9:18 PM
17	StackEnsemble	0.98666667	0.98666667	Completed	0:01:16	Jan 18, 2021 9:19 PM
2	MinMaxScaler, RandomForest	0.97333333	0.97333333	Completed	0:01:07	Jan 18, 2021 9:04 PM
9	SparseNormalizer, LightGBM	0.97333333	0.97333333	Completed	0:00:55	Jan 18, 2021 9:11 PM
12	SparseNormalizer, XGBoostClassifier	0.96666667	0.97333333	Completed	0:00:52	Jan 18, 2021 9:14 PM



Standard Classification Algorithms	Tree Algorithms	Gradient Boosting Algorithms	Support Vector Algorithms	Optimization Algorithms
Logistic regression	Decision tree	XGBoost	SVC	SGD
Naïve Bayes	Random forest	LightGBM	Linear SVC	Averaged perceptron classifier
K-nearest neighbors	Extremely randomized trees	Gradient boosting	Linear SVM classifier	

Chapter 6: Building an AutoML Forecasting Solution

In [47]: df.head(10)

Out[47]:

	WeekStarting	Store	Brand	Quantity	Advert	Price	Revenue
0	1990-06-14	1000	dominicks	12003	1	2.59	31087.77
1	1990-06-21	1000	dominicks	10239	1	2.39	24471.21
2	1990-06-28	1000	dominicks	17917	1	2.48	44434.16
3	1990-07-05	1000	dominicks	14218	1	2.33	33127.94
4	1990-07-12	1000	dominicks	15925	1	2.01	32009.25
5	1990-07-19	1000	dominicks	17850	1	2.17	38734.50
6	1990-07-26	1000	dominicks	10576	1	1.97	20834.72
7	1990-08-02	1000	dominicks	9912	1	2.26	22401.12
8	1990-08-09	1000	dominicks	9571	1	2.11	20194.81
9	1990-08-16	1000	dominicks	15748	1	2.42	38110.16

DATA GUARDRAILS:

TYPE: Frequency detection
STATUS: PASSED
DESCRIPTION: The time series was analyzed, all data points are aligned with detected frequency.

TYPE: Missing feature values imputation
STATUS: PASSED
DESCRIPTION: No feature missing values were detected in the training data.
Learn more about missing value imputation: <https://aka.ms/AutomatedMLFeaturization>

TYPE: Short series handling
STATUS: PASSED
DESCRIPTION: Automated ML detected enough data points for each series in the input data to continue with training.

ITERATION: The iteration being evaluated.
PIPELINE: A summary description of the pipeline being evaluated.
DURATION: Time taken for the current iteration.
METRIC: The result of computing score on the fitted pipeline.
BEST: The best observed score thus far.

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler DecisionTree	0:00:33	0.0608	0.0608
1	MinMaxScaler DecisionTree	0:00:27	0.0761	0.0608
2	StandardScalerWrapper LassoLars	0:00:33	0.0244	0.0244
3	RobustScaler ElasticNet	0:00:31	0.1121	0.0244
4	StandardScalerWrapper LassoLars	0:00:33	0.0244	0.0244
5	RobustScaler ElasticNet	0:00:29	0.0241	0.0241
6	RobustScaler DecisionTree	0:00:33	0.0648	0.0241
7	MinMaxScaler DecisionTree	0:00:31	0.0525	0.0241
8	RobustScaler DecisionTree	0:00:36	0.0890	0.0241
9	MinMaxScaler DecisionTree	0:00:31	0.0847	0.0241
10	StandardScalerWrapper DecisionTree	0:00:34	0.0621	0.0241
11	StandardScalerWrapper DecisionTree	0:00:28	0.0395	0.0241
12	RobustScaler DecisionTree	0:00:34	0.0410	0.0241
13	RobustScaler ElasticNet	0:00:31	0.0762	0.0241
14	StandardScalerWrapper DecisionTree	0:00:37	0.0563	0.0241
15	MinMaxScaler DecisionTree	0:00:40	0.0458	0.0241
16	MinMaxScaler DecisionTree	0:00:28	0.0586	0.0241
17		0:00:29	nan	0.0241
18	VotingEnsemble	0:00:41	0.0164	0.0164
ITERATION	PIPELINE	DURATION	METRIC	BEST
0	AutoArima	0:01:31	0.2910	0.2910
1	ProphetModel	0:00:58	0.0352	0.0352
2	MaxAbsScaler DecisionTree	0:00:31	0.0466	0.0352
3	MinMaxScaler DecisionTree	0:00:35	0.0702	0.0352
4	StandardScalerWrapper LassoLars	0:00:41	0.0244	0.0244
5	RobustScaler ElasticNet	0:00:33	0.1121	0.0244
6	StandardScalerWrapper LassoLars	0:00:33	0.0244	0.0244
7	RobustScaler ElasticNet	0:00:32	0.0241	0.0241
8	RobustScaler DecisionTree	0:00:28	0.0603	0.0241
9	MinMaxScaler DecisionTree	0:00:44	0.0415	0.0241
10	RobustScaler DecisionTree	0:00:34	0.0717	0.0241
11	MinMaxScaler DecisionTree	0:00:30	0.0716	0.0241
12	StandardScalerWrapper DecisionTree	0:00:42	0.0486	0.0241
13	StandardScalerWrapper DecisionTree	0:00:38	0.0590	0.0241
14	VotingEnsemble	0:00:49	0.0174	0.0174

Standard regression algorithms	Tree algorithms	Gradient boosting algorithms	Other algorithms shared with regression	Forecasting-specific algorithms
Elastic Net LARS Lasso	Decision Tree Random Forest Extremely Randomized Trees	XGBoost LightGBM Gradient Boosting	Stochastic Gradient Descent (SGD) KNN	Prophet ARIMA ForecastTCN

Chapter 7: Using the Many Models Solution Accelerator

jupyter Accelerator_Installation Last Checkpoint: Yesterday at 8:44 PM (autosaved)

In [2]: `git clone https://github.com/microsoft/solution-accelerator-many-models`

```
Cloning into 'solution-accelerator-many-models'...
remote: Enumerating objects: 181, done.
remote: Counting objects: 100% (181/181), done.
remote: Compressing objects: 100% (137/137), done.
remote: Total 2403 (delta 71), reused 94 (delta 27), pack-reused 2222
Receiving objects: 100% (2403/2403), 6.04 MiB | 3.70 MiB/s, done.
Resolving deltas: 100% (1468/1468), done.
Checking connectivity... done.
Checking out files: 100% (39/39), done.
```

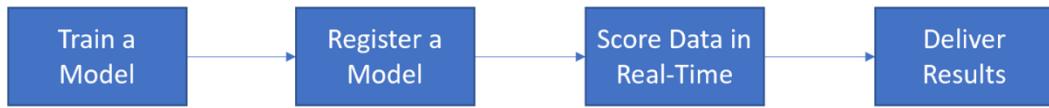
The screenshot shows a Jupyter Notebook interface with a code cell containing the command to clone the repository. Below the notebook is a flowchart titled "many-models-train". The flowchart starts with a box labeled "oj_data_small_train" which outputs to a node labeled "Dataset output". This output feeds into a node labeled "train_10_models", which then outputs to a box labeled "many-models-train". A status indicator in this box says "Running". Finally, the output from "train_10_models" leads to a box labeled "many_models_training_outpu...".

	Week Starting	Store	Brand	Quantity	Advert	Price	Revenue	Predicted
0	1992-05-28	1001	dominicks	12736	1	2.37	30184.32	12655.926953
1	1992-06-04	1001	dominicks	17512	1	2.28	39927.36	17589.279167
2	1992-06-11	1001	dominicks	17769	1	2.18	38736.42	17616.802905
3	1992-06-18	1001	dominicks	17825	1	2.64	47058.00	18321.770558
4	1992-06-25	1001	dominicks	11105	1	2.02	22432.10	11428.825570
5	1992-07-02	1001	dominicks	15763	1	1.97	31053.11	15358.724956
6	1992-07-09	1001	dominicks	10904	1	2.33	25406.32	10827.578098
7	1992-07-16	1001	dominicks	16153	1	2.27	36667.31	15773.192383
8	1992-07-23	1001	dominicks	14260	1	2.03	28947.80	14397.971780
9	1992-07-30	1001	dominicks	9284	1	2.67	24788.28	8715.212884

Prediction has 124 rows. Here the first 10.

	Date	Store	Sales	Predicted
0	2021-03-01	San Francisco	83	54.902630
1	2021-03-02	San Francisco	37	49.947213
2	2021-03-03	San Francisco	44	49.690911
3	2021-03-04	San Francisco	50	54.322844
4	2021-03-05	San Francisco	73	53.245385
5	2021-03-06	San Francisco	91	50.389134
6	2021-03-07	San Francisco	55	53.390174
7	2021-03-08	San Francisco	9	50.627610
8	2021-03-09	San Francisco	98	48.155031
9	2021-03-10	San Francisco	8	49.585355

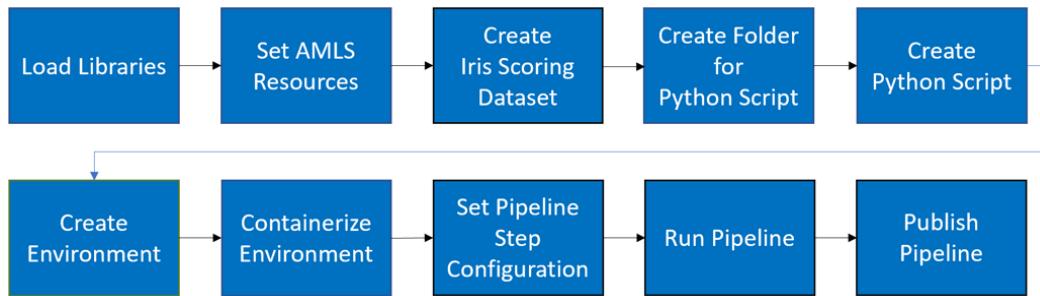
Chapter 8: Choosing Real-Time versus Batch Scoring



	Advantages	Disadvantages	When to use
Batch scoring	Cheap Easy to replicate Simple	Slow	When end users can wait for predictions with new data
Real-time scoring	Fast	Costly Complex Unique	When end users require predictions immediately

Scenario	Type of solution
Fraud detection	Real-time
Predictive maintenance	Dependent on when machines are serviced
Once a day, week, or month scoring	Batch
Recommendation engines	Real-time
Forecasting for the long term	Batch
Forecasting minute by minute	Real-time
Web-based applications	Dependent on user expectations

Chapter 9: Implementing a Batch Scoring Solution



iris-scoring-step - Finished

```
published_pipeline = pipeline_run.publish_pipeline(  
    name='Iris-Scoring-Pipeline',  
    description='Pipeline that Scores Iris Data', version= '1.0')  
  
published_pipeline
```

Name	Id	Status	Endpoint
Iris-Scoring-Pipeline	466fc3a2-cfcf-4a31-8fa5-21762e3c4b6e	Active	REST Endpoint

Home > automlexamplew1005528502 >

azureml-blobstore-4e7d29d8-4f58-466c-90bf-5c3a9f2f3162 ...

Container

Search (Ctrl+/)

Upload Change access level Refresh Delete Change tier

Overview Access Control (IAM)

Authentication method: Access key ([Switch to Azure AD User Account](#))
Location: [azureml-blobstore-4e7d29d8-4f58-466c-90bf-5c3a9f2f3162](#) / Output_Folder

Search blobs by prefix (case-sensitive)

Name

[.] Iris_Parallel_Predictions.csv Iris_Predictions.csv

Settings

Shared access signature
Access policy
Properties
Metadata
Editor (preview)

Load Libraries → Set AMLS Resources → Create Iris Scoring Dataset → Create Folder for two Python Scripts → Create two Python Scripts

Set Pipeline Data → Retrieve & Containerize Environments → Configure two Pipeline Steps → Run two-step Pipeline → Publish Pipeline

```
graph LR; subgraph TopFlow [Top Flow]; A[Load Libraries] --> B[Set AMLS Resources]; B --> C[Create Iris Scoring Dataset]; C --> D[Create Folder for two Python Scripts]; D --> E[Create two Python Scripts]; end; subgraph BottomFlow [Bottom Flow]; F[Set Pipeline Data] --> G[Retrieve & Containerize Environments]; G --> H[Configure two Pipeline Steps]; H --> I[Run two-step Pipeline]; I --> J[Publish Pipeline]; end; A -.-> F
```

Iris Parallel Scoring

```
graph TD; data --> Iris_parallel_scoring[Iris_parallel_scoring]
```

iris-parallel-scoring-step - Finished

```
graph TD; parallel_predictions --> parallel_predictions
```

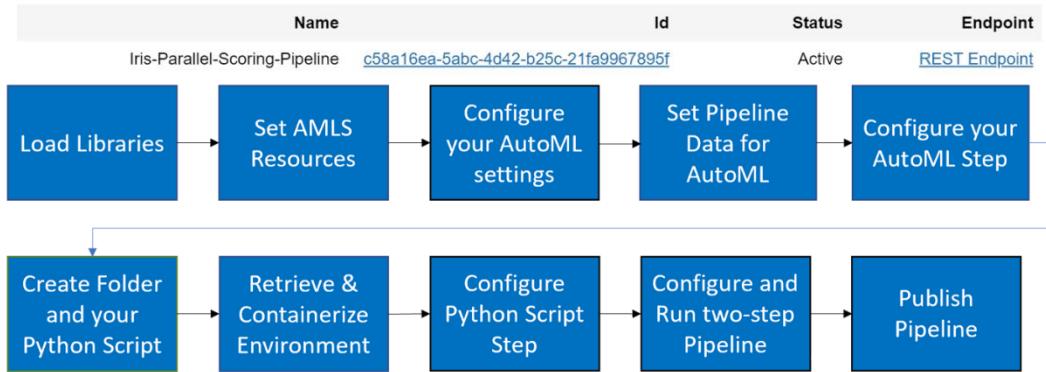
iris-output-step - Finished

```

published_pipeline = pipeline_run.publish_pipeline(
    name='Iris-Parallel-Scoring-Pipeline',\
    description='Pipeline that Scores Iris Data in Parallel', version= '1.0')

published_pipeline

```



Iris Training

data
training_data

Multiclass_AutoML_Step - Finished

metrics_data
model_data

Model-Registration-Step - Finished

Name	Id	Status	Endpoint
Iris-AutoML-Training-Pipeline	bdafc5af-4964-40ac-bb01-50dd49b6dcb5	Active	REST Endpoint

Chapter 10: Creating End-to-End AutoML Solutions

Create Data Factory ...

⚠ Changes on this step may reset later selections you have made. Review all options prior to deployment.

Basics Git configuration Networking Advanced Tags Review + create

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ Dennis Sawyers Internal MS Learning Account

Resource group * ⓘ auto-ml-example-resource-group

Create new

Instance details

Region * ⓘ North Central US

Name * ⓘ automl-adf713

Version * ⓘ V2

Search resources, services, and docs (G+)

Azure services

Create a resource Monitor Logic apps Data factories Subscriptions Azure Active Directory Kubernetes services Machine learning Resource groups More services

adf-service-principal | Certificates & secrets ✖️

 Search (Ctrl+ /)

«

 Got feedback?

 Overview

 Quickstart

 Integration assistant

Manage

 Branding

 Authentication

 Certificates & secrets

 Token configuration

Azure services

 Create a resource

 Azure Active Directory

 Subscriptions

 Data factories

 Machine Learning

 Monitor

 Resource groups

 Event Hubs

 Azure Databricks

 More services

Add a client secret

Description

ADF-Secret

Expires

In 1 year

In 2 years

Never

Add

Cancel

Add role assignment

X

Role ⓘ

Contributor ⓘ



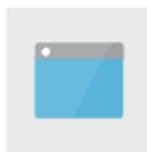
Assign access to ⓘ

User, group, or service principal



Select ⓘ

adf-service-principal



adf-service-principal

Microsoft Azure | Data Factory > automl-adf713

Azure Data Factory

Let's get started

The screenshot shows the Azure Data Factory interface. At the top, there's a navigation bar with icons for Home, Create, Monitor, and Help. Below it, a banner says "Let's get started" with six circular icons representing different operations: "Create pipeline" (blue valve), "Create data flow" (two green cylinders with arrows), "Create pipeline from template" (green squares with arrows), "Copy data" (two blue cylinders with yellow stars), "Configure SSIS Integration" (blue cylinder with a gear), and "Set up code repository" (red GitHub-like icon). The main area has a dark blue background with a grid pattern.

Create pipeline Create data flow Create pipeline from template Copy data Configure SSIS Integration Set up code repository

New linked service

Data store Compute

 Search



Azure Batch



Azure Data Lake Analytics



Azure Databricks



Azure Function



Azure HDInsight



Azure Machine Learning

Microsoft Azure | Data Factory > automl-adf713

» Data Factory Validate all Publish all Refresh

Factory Resources

Filter resources by name +

- ▶ Pipelines → Pipeline
- ▶ Datasets Pipeline from template
- ▶ Data flows Dataset
- ▶ Power Query (Preview) Data flow
- Power Query
- Copy Data tool

pipeline1

Activities

Save as template Validate Debug Add trigger

Search activities

Move & transform

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Machine Learning Batch...

Machine Learning Up...

Machine Learning Execute Pipeline

General Settings User properties

Name * Machine Learning Execute Pipeline1

Data Factory Validate all Publish all Refresh Discard all

Factory Resources

Filter resources by name

Pipelines 1

Iris Scoring Pipeline

Datasets 0

Data flows 0

Power Query (Preview) 0

Machine Learning Execute Pipeline

Execute Iris Scoring Pipeline

The screenshot displays the Azure Data Factory pipeline editor interface. On the left, the 'Factory Resources' pane shows a single pipeline named 'Iris Scoring Pipeline'. The main workspace shows the creation of a new pipeline named 'Machine Learning Execute Pipeline1'. This pipeline contains a single activity named 'Execute Iris Scoring Pipeline'. The pipeline is currently in the 'General' tab, with other tabs like 'Settings' and 'User properties' available. The pipeline editor includes standard tools for saving, validating, and debugging.

Integration runtime setup

Integration Runtime is the native compute used to execute or dispatch activities. Choose what integration runtime to create based on required capabilities. [Learn more](#)



Azure, Self-Hosted

Perform data flows, data movement and dispatch activities to external compute.



Azure-SSIS

Lift-and-shift existing SSIS packages to execute in Azure.

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities

[+ New](#)

[↻ Refresh](#)

[Filter by name](#)

Showing 1 - 2 of 2 items

Name ↑↓

Type ↑↓

Sub-type ↑↓

Status ↑↓

AutoResolveIntegra...	Azure	Public	Running
IntegrationRuntime	Self-Hosted	---	Running

Microsoft Azure | automl-adf713

» Data Factory ✓ Validate all ⚡ Publish all ⌂ Refresh

Factory Resources

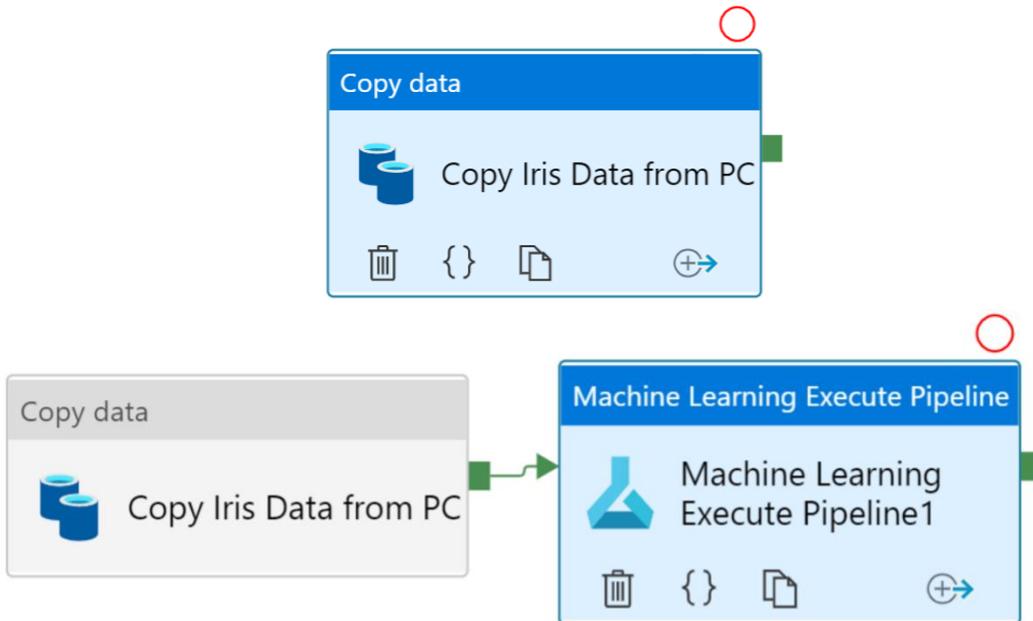
Filter resources by name

- ▶ Pipelines
- ▶ Datasets
- ▶ Data flows
- ▶ Power Query (Preview)

+ Pipeline
Pipeline from template
Dataset
Data flow
Power Query
Copy Data tool

A screenshot of the Microsoft Azure Data Factory interface. The top navigation bar shows 'Microsoft Azure' and the resource group 'automl-adf713'. Below the navigation bar is a toolbar with icons for Data Factory, Validate all, Publish all, and Refresh. The main area is titled 'Factory Resources' and contains a sidebar with icons for Home, Edit, and Delete. A search bar labeled 'Filter resources by name' is present. On the left, there's a list of resource types: Pipelines, Datasets, Data flows, and Power Query (Preview). To the right of this list is a context menu with several options: Pipeline, Pipeline from template, Dataset, Data flow, Power Query, and Copy Data tool. The 'Copy Data tool' option is highlighted with a large blue arrow pointing towards it.

Save as template Validate Validate copy runtime Debug



Set properties

Name

ScoringResults

Linked service *

AMLSDatastoreLink

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime

File path

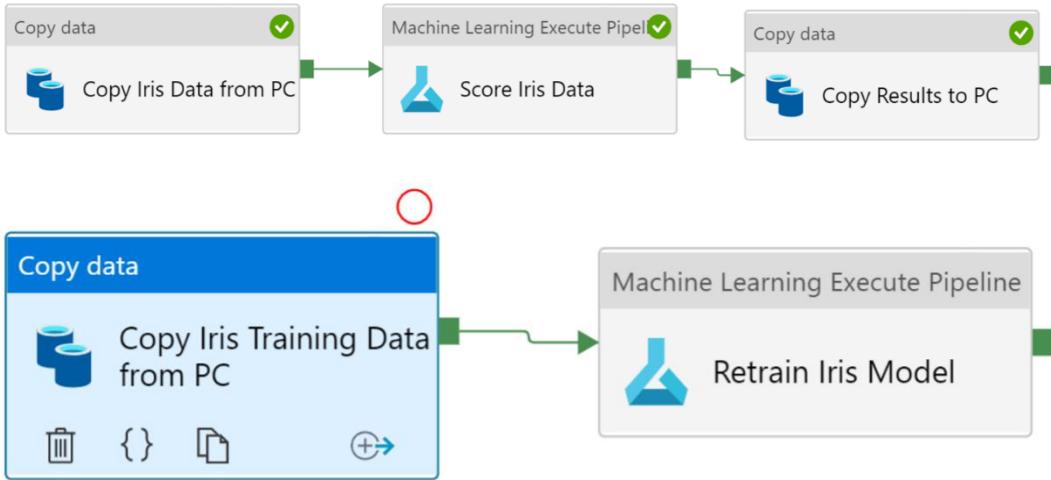
azureml-blobstore-4e7d2 / Output_Folder / Iris_Predictions.csv

First row as header

Import schema

From connection/store From sample file None

▷ Advanced



Chapter 11: Implementing a Real-Time Scoring Solution

Deploy a model X

Name * i eye

diabetes-aci-gui

Description i

Compute type * i *

Azure Container Instance

Models: Diabetes-AllData-Regression-AutoML-R2:1

Enable authentication (On)

Configure Settings

Configure compute cluster settings for your selected virtual machine size.

Name	Category	Cores	Available quota	RAM	Storage
Standard_DS3_v2	General purpose	4	88 cores	14 GB	28 GB

Compute name * i (Edit)

aks-amls-cluster

Cluster purpose

Production Dev-test

Number of nodes * i

- 3

Network configuration i

Basic Advanced

Deploy a model

X

Name * ⓘ



Description ⓘ

Compute type * ⓘ

 *

Compute name * ⓘ

 ▾

Models: Diabetes-AllData-Regression-AutoML-R2:1

Enable authentication

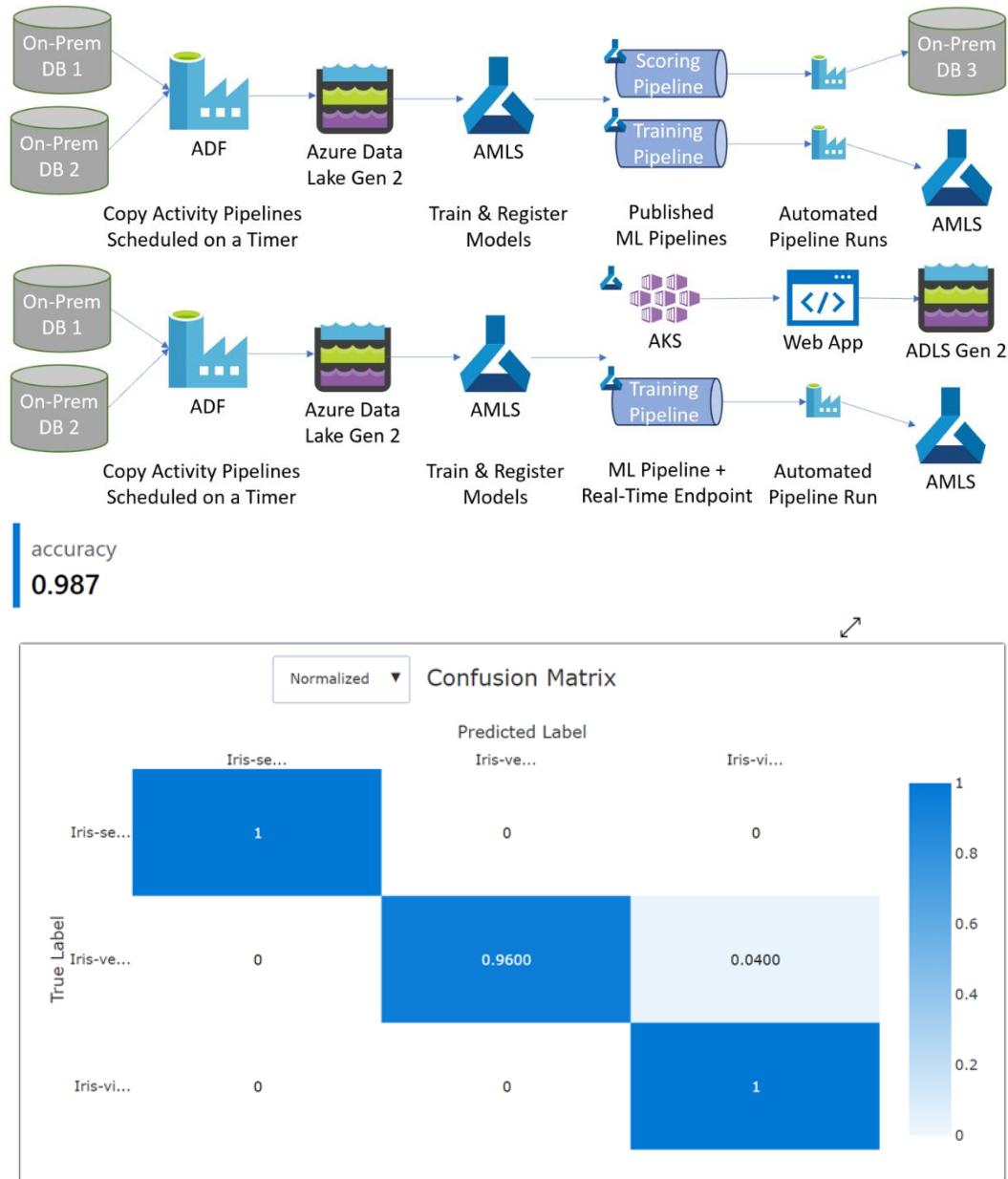


Type

 ▾

```
'{"data": [{"AGE": 68, "SEX": 1, "BMI": 41.3, "BP": 119, "S1": 145, "S2": 65.7, "S3": 27, "S4": 8.51, "S5": 4.823, "S6": 94}, {"AGE": 59, "SEX": 1, "BMI": 32.1, "BP": 104, "S1": 104, "S2": 187.6, "S3": 97, "S4": 6.87, "S5": 3.438, "S6": 120}, {"AGE": 64, "SEX": 1, "BMI": 30.4, "BP": 67, "S1": 20, "S2": 169.0, "S3": 86, "S4": 4.79, "S5": 5.704, "S6": 86}, {"AGE": 57, "SEX": 1, "BMI": 32.3, "BP": 113, "S1": 290, "S2": 116.5, "S3": 58, "S4": 5.35, "S5": 3.872, "S6": 101}, {"AGE": 36, "SEX": 1, "BMI": 26.5, "BP": 131, "S1": 131, "S2": 122.5, "S3": 57, "S4": 6.9, "S5": 5.634, "S6": 111}]}'
```

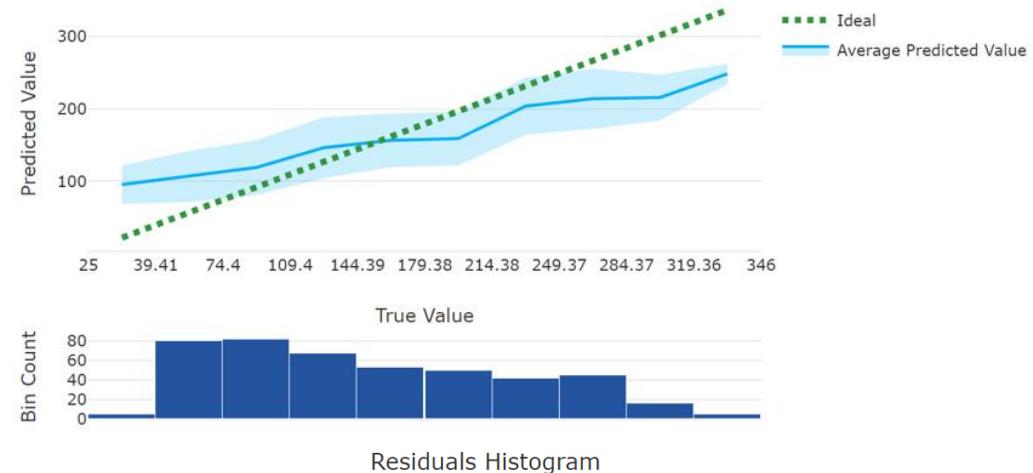
Chapter 12: Realizing Business Value with AutoML



mean_absolute_percentage_e...

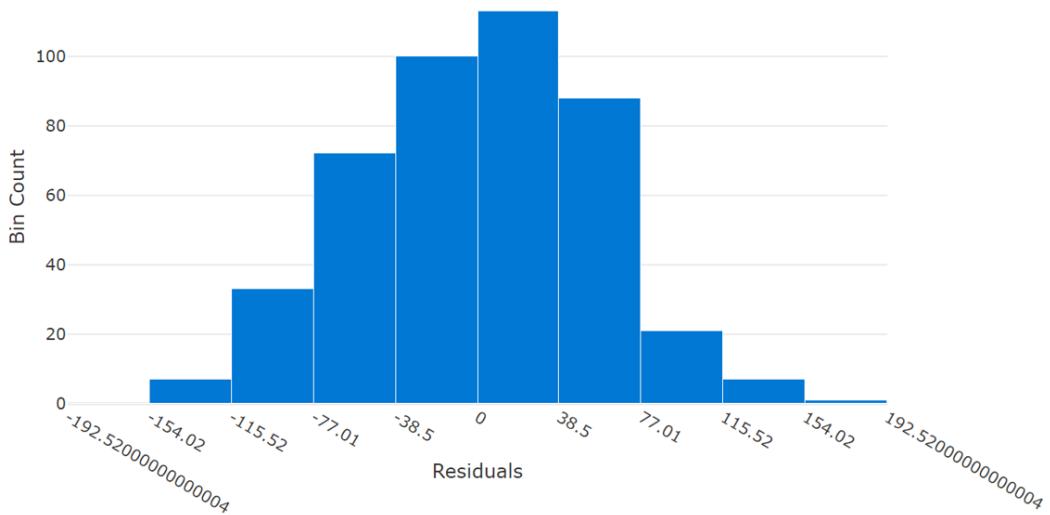
38.941

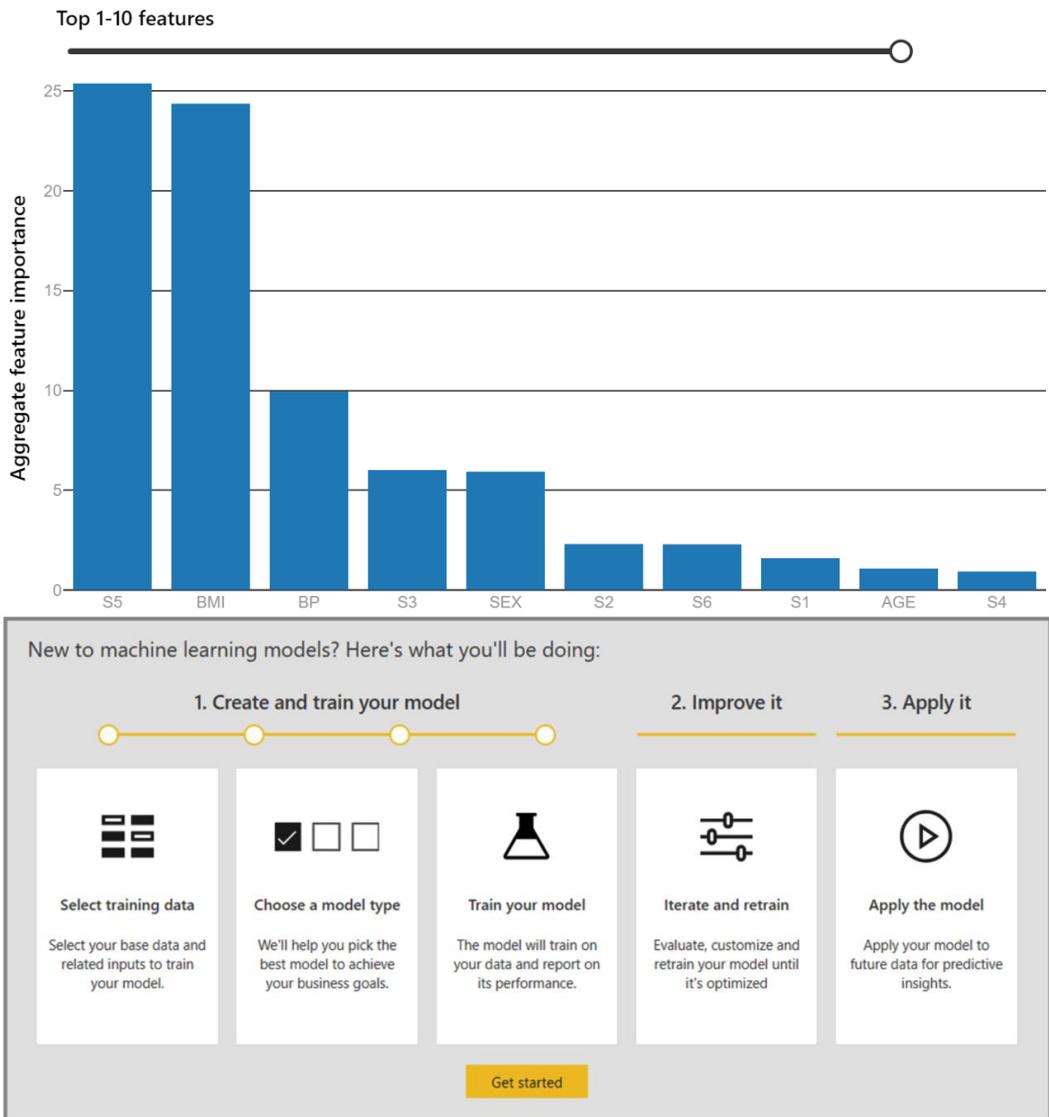
Predicted vs. True

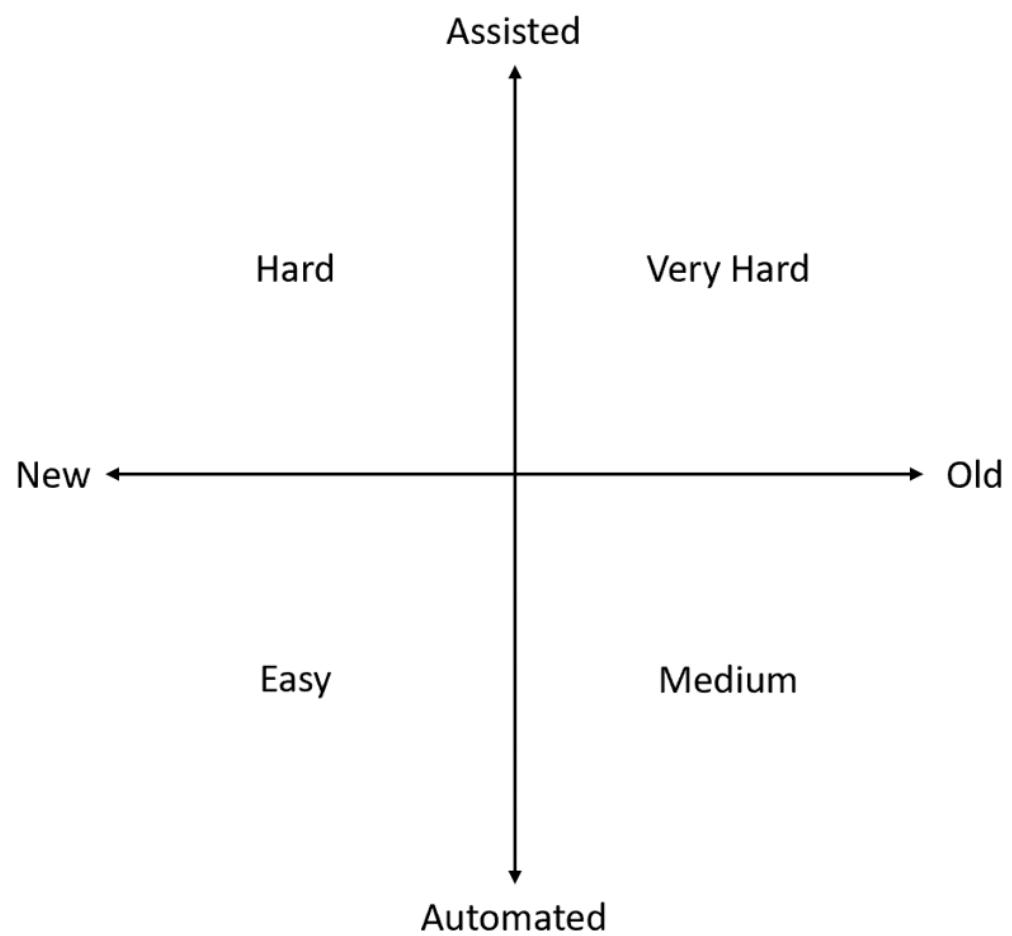


True Value

Residuals Histogram







Key questions to ask	Common answers to key questions
Where is my input data coming from?	On-premise database, cloud database
How often should my solution score?	On certain days at certain times
How do end users access results?	Report generated by an on-premise database
When should I retrain my model?	Once a month as new data becomes available
How should I orchestrate my solution?	Azure Data Factory

Key questions to ask	Common answers to key questions
Where is my input data coming from?	On-premise database, cloud database
How often should I retrain my model?	Whenever model performance degrades
Where is my endpoint located?	A user-facing web app, an Azure function
How many requests are coming in at once?	A maximum of 100 requests per minute
How fast do end users expect a response?	Within 10 seconds of submitting a request
How large should my AKS cluster be?	Large VM size for faster performance

	Difficulty of gaining trust	Who is most likely to offer resistance	How to overcome resistance
Completely new automated process	Easy as the process is being created with ML in mind	Resistance is unlikely if your ML model performs as expected	Build an ML model that exceeds expectations and works well
Replacing an old, automated process	Medium as you are replacing an older process	People who built or oversee the output of the older process	Compare results of the old and new process side-by-side
Completely new assisted decision making process	Hard as many people may not trust AI generated output	People who feel their decision making is threatened by AI	Stress that AI exists as an advisor to enhance their own decisions
Replacing an old, assisted decision making tool	Extremely difficult as users will fondly recall the old tool	People who cannot understand the tech behind the new tool	Emphasize the explainability of your AutoML-built solution