Assignment 5

Assignment on Azure Cloud Platform

PartB

1.

We aim to develop a robust machine learning model to identify and classify spam comments on YouTube—a growing concern for content creators and viewers alike. Utilizing the YouTube Spam Collection dataset, which consists of 1,956 comments from five popular videos, our goal is to accurately distinguish between class 1 (irrelevant or inappropriate messages) and class 0 (relevant, appropriate comments).

This classification is crucial for maintaining a healthy digital environment on YouTube, as spam comments can be disruptive, misleading, or even harmful. By addressing this issue, we not only seek to enhance the user experience on YouTube but also to gain valuable insights into the nature of spam comments on social media platforms. Ultimately, our model will contribute to creating a safer and more enjoyable space for users to engage with content online.

Solution:

We propose a two-fold approach utilizing Logistic Regression and Random Forest classifiers:

Logistic Regression:

Why: Logistic Regression is a simple yet effective linear model for binary classification problems. It provides a clear interpretation of feature importance and performs well on linearly separable data.
How: We will use Logistic Regression to create a baseline model, focusing on its ability to identify relationships between features and the binary outcome (spam vs. non-spam).

Random Forest:

Why: Random Forest is an ensemble learning method that builds multiple decision trees and merges them to achieve higher accuracy and robustness. It excels at handling complex, non-linear relationships in data and is less prone to overfitting compared to individual decision trees.
How: We will train a Random Forest model to capture the intricate patterns in the data, leveraging its ability to manage feature interactions and provide feature importance insights.

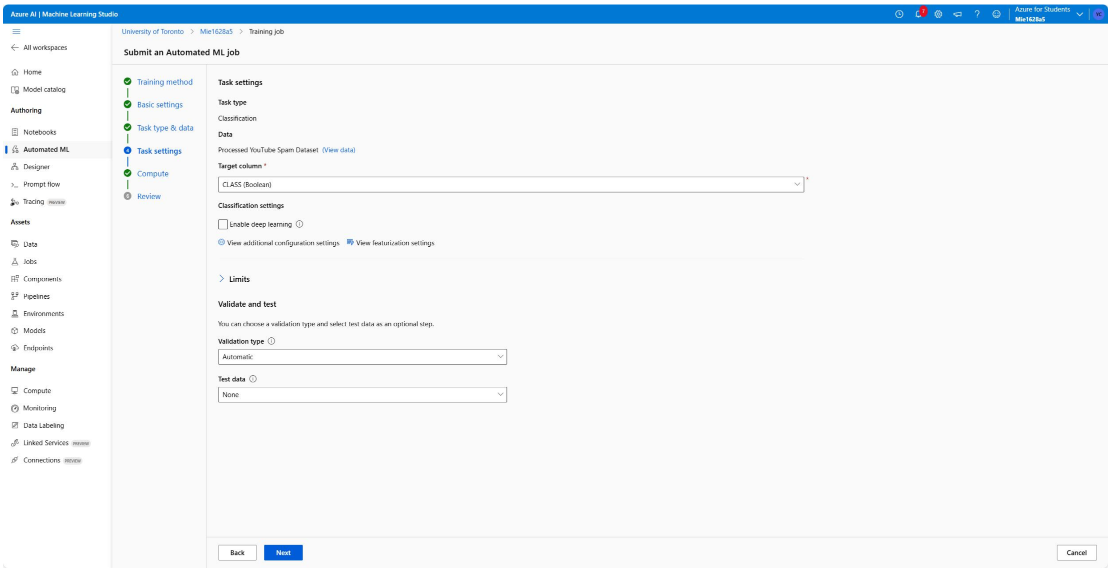
I have written the answers for Q2,Q3,Q4 in the Markdown block of the notebook and exported the codes along with the explanations. See A5_Notebook.pdf for details.

Q5 is answered on the following pages

Q5:

I will use Automated ML for my data set and explain the best model results.

Configure Automated ML job seetings



University of Toronto > Mie1628a5 > Training job

Submit an Automated ML job

Task settings

Task type

Classification

Data

Processed YouTube Spam Dataset (View data)

Target column *

CLASS (Boolean)

Classification settings

□ Enable deep learning

View additional configuration settings View featurization settings

Limits

Validate and test

You can choose a validation type and select test data as an optional step.

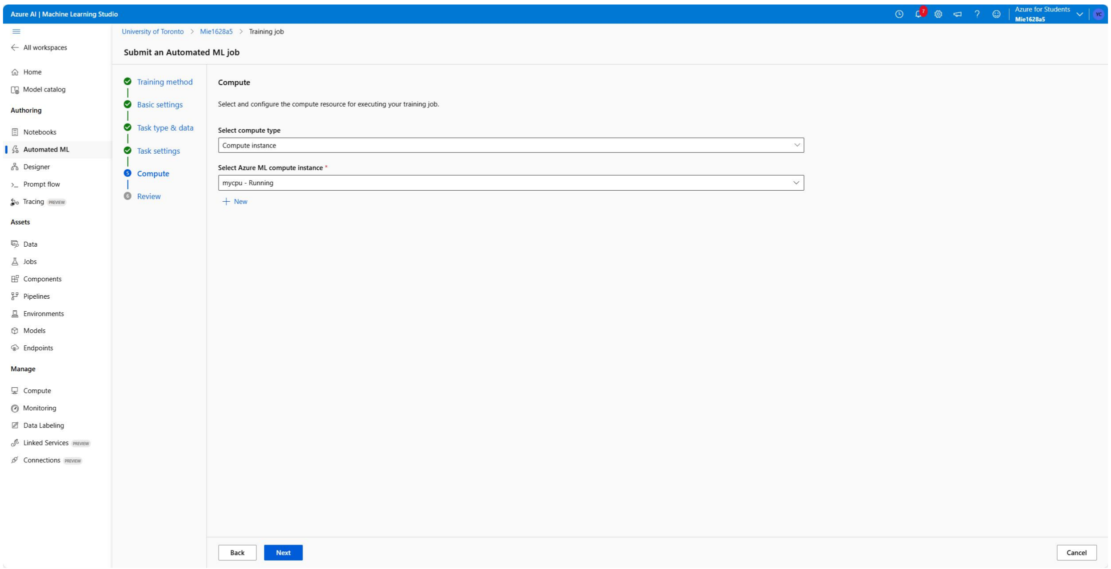
Validation type

Automatic

Test data

None

Choose my compute instance



University of Toronto > Mie1628a5 > Training job

Submit an Automated ML job

Compute

Select and configure the compute resource for executing your training job.

Select compute type

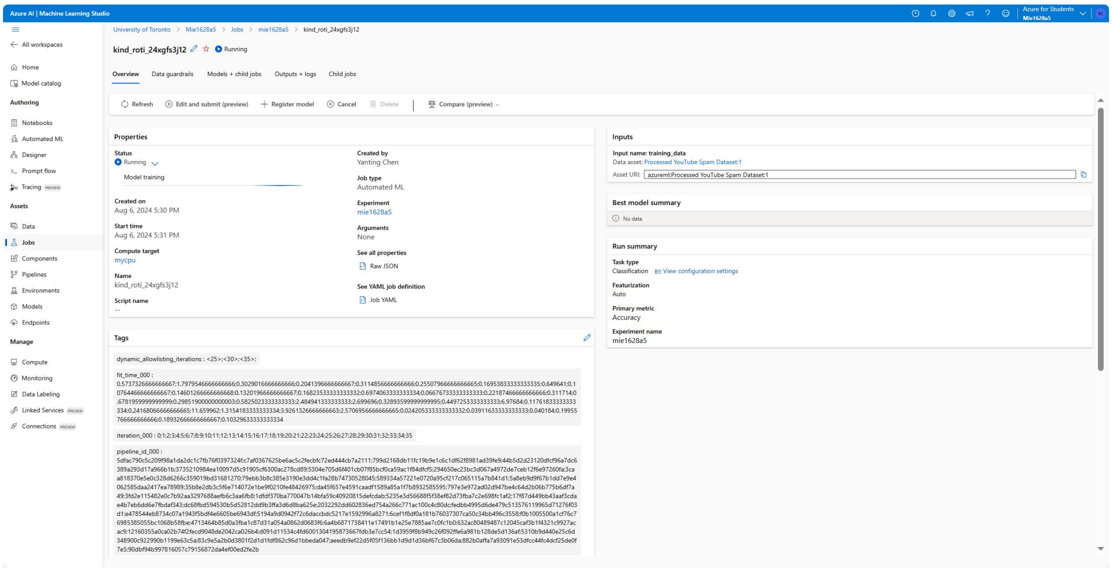
Compute instance

Select Azure ML compute instance *

mycpu - Running

+ New

Start Running:



kind_roti_24xgfs3j12 Running

Properties

Status
Running

Model training

Created on
Aug 6, 2024 5:30 PM

Start time
Aug 6, 2024 5:31 PM

Compute target
mycpu

Name
kind_roti_24xgfs3j12

Script name
--

Created by
Yanting Chen

Job type
Automated ML

Experiment
mie1628a5

Arguments
None

See all properties
Raw JSON

See YAML job definition
Job YAML

Tags

dynamic_allowlisting_iterations : <25>;<30>;<35>;

fit_time_000 :
0.5737326666666667;1.7979546666666666;0.3029016666666666;0.2041396666666667;0.3114856666666666;0.2550796666666665;0.1695383333333335;0.649641;0.1076446666666667;0.1460126666666668;0.1320196666666667;0.1682353333333332;0.6974063333333334;0.0667673333333333;0.2218746666666666;0.311714;0.6781959999999999;0.2985190000000003;0.5825023333333333;2.484941333333333;2.699696;0.3289359999999995;0.4497253333333333;6.97684;0.1176183333333334;0.2416806666666665;11.659962;1.3154183333333334;3.9261326666666663;2.5706956666666665;0.0242053333333332;0.0391163333333333;0.040184;0.19955766666666666;0.18932666666666667;0.10329633333333334

iteration_000 : 0;1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;23;24;25;26;27;28;29;30;31;32;33;34;35

pipeline_id_000 :
5dfac790c5c209f98a1da2dc1c7fb76f0397324f;c7af0367625be6ac5c2fecbfc72ed444cb7a2111;799d2168db11fc19b9e1c6c1df62f8981ad39fe9;44b5d2d23120dfcf96a7dc6389a293d17a966b1b;3735210984ea10097d5c91905cf6300ac278cd89;5304e705d6f401cb07f85bcf0ca59ac1f84dfcf5;294650ec23bc3d067a4972de7ceb12f6e97260fa;3caa818370e5e0c328d6266c359019bd31681270;79ebb3b8c385e3190e3dd4c1fa28b74730528045;589334a57221e0720a95cf217c065115a7b841d1;5a8eb9d9f67b1dd7e9e4062585daa2417ea78989;35b8e2db3c5f6e714072e1be9f0210fe48426975;da45f657e4591caadf1589a85a1f7b8932585595;797e3e972ad02d947be4c64d2b06b775b6df7a49;3fd2e115482e0c7b92aa3297688aefb6c3aa6fb8;1dfdf370ba770047b14bfa59c40920815defcdab;5235e3d56688f5f38ef82d73fba7c2e698fc1af2;17f87d449bb43aaf3cdae4b7eb6dd6e7fbdaf343;dc68fbd594530b5d52812dd9b3ffa3d6d8ba625e;2032292dd602836ed754a266c771ac100c4c80dc;fedbb4995d6de479c513576119965d71276f03d1;e478544eb8734c07a1943f5bdf4e6605be6943df;5194a9d0942f72c6daccbdc5217e1592996a8271;6cef1f8df0a181b76037307ca50c34bb496c3558;f0b1005500a1cf76c76985385055bc1068b58fbe;4713464b85d0a3fba1c87d31a054a0862d0683f6;6a4b6871738411e17491b1e25e7885ae7c0fc1b0;632ac80489487c12045caf3b1f4321c9927acac9;12160355a0ca02b74f2fecd9048de2042ca026b4;d091d11534c4fd6001304195873667fdb3e7cc54;1d3959f8b949c26f092ffe6a981b128de5d136af;5310b9d440e25c6d348900c922990b1199e63c5a;83c9e5a2b0d3801f2d1d1fdf862c96d1bbeda047;aeedb9ef22d5f05f136bb1d9d1d36bf67c3b06da;882b0affa7a93091e53dfcc44fc4dcf25de0f7e5;90dbf94b997816057c79156872da4ef00ed2fe2b

Inputs

Input name: training_data
Data asset: Processed YouTube Spam Dataset:1
Asset URI: azureml:Processed YouTube Spam Dataset:1

Best model summary

No data

Run summary

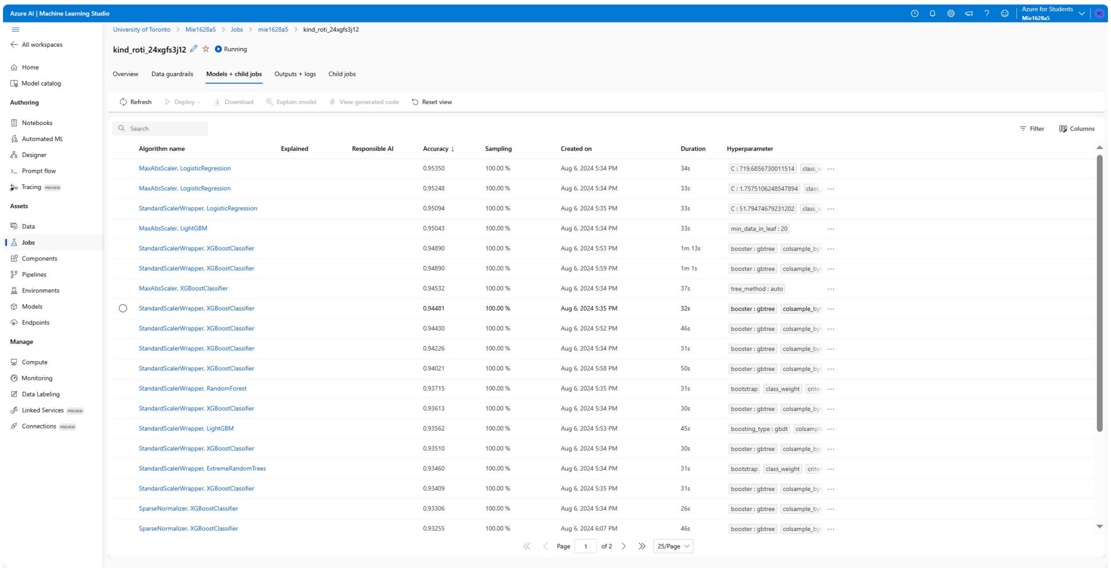
Task type
Classification View configuration settings

Featurization
Auto

Primary metric
Accuracy

Experiment name
mie1628a5

It will run mutilple models to find the best one

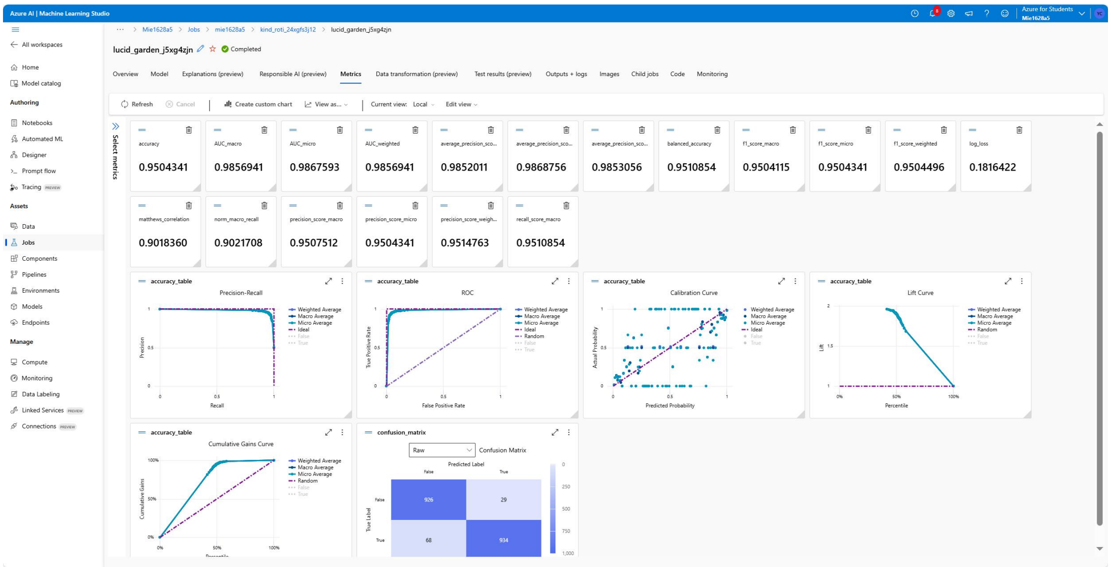


kind_roti_24xgfs3j12 Running

	Algorithm name	Explained	Responsible AI	Accuracy ↓	Sampling	Created on	Duration	Hyperparameter
	MaxAbsScaler, LogisticRegression			0.95350	100.00 %	Aug 6, 2024 5:34 PM	34s	C : 719.6856730011514 class_w
	MaxAbsScaler, LogisticRegression			0.95248	100.00 %	Aug 6, 2024 5:34 PM	33s	C : 1.7575106248547894 class_
	StandardScalerWrapper, LogisticRegression			0.95094	100.00 %	Aug 6, 2024 5:35 PM	33s	C : 51.79474679231202 class_w
	MaxAbsScaler, LightGBM			0.95043	100.00 %	Aug 6, 2024 5:34 PM	33s	min_data_in_leaf : 20
	StandardScalerWrapper, XGBoostClassifier			0.94890	100.00 %	Aug 6, 2024 5:53 PM	1m 13s	booster : gbtree colsample_by
	StandardScalerWrapper, XGBoostClassifier			0.94890	100.00 %	Aug 6, 2024 5:59 PM	1m 1s	booster : gbtree colsample_by
	MaxAbsScaler, XGBoostClassifier			0.94532	100.00 %	Aug 6, 2024 5:34 PM	37s	tree_method : auto
○	StandardScalerWrapper, XGBoostClassifier			0.94481	100.00 %	Aug 6, 2024 5:35 PM	32s	booster : gbtree colsample_by
	StandardScalerWrapper, XGBoostClassifier			0.94430	100.00 %	Aug 6, 2024 5:52 PM	46s	booster : gbtree colsample_by
	StandardScalerWrapper, XGBoostClassifier			0.94226	100.00 %	Aug 6, 2024 5:34 PM	31s	booster : gbtree colsample_by
	StandardScalerWrapper, XGBoostClassifier			0.94021	100.00 %	Aug 6, 2024 5:58 PM	50s	booster : gbtree colsample_by
	StandardScalerWrapper, RandomForest			0.93715	100.00 %	Aug 6, 2024 5:35 PM	31s	bootstrap class_weight criter
	StandardScalerWrapper, XGBoostClassifier			0.93613	100.00 %	Aug 6, 2024 5:34 PM	30s	booster : gbtree colsample_by
	StandardScalerWrapper, LightGBM			0.93562	100.00 %	Aug 6, 2024 5:53 PM	45s	boosting_type : gbdt colsample
	StandardScalerWrapper, XGBoostClassifier			0.93510	100.00 %	Aug 6, 2024 5:34 PM	30s	booster : gbtree colsample_by
	StandardScalerWrapper, ExtremeRandomTrees			0.93460	100.00 %	Aug 6, 2024 5:34 PM	31s	bootstrap class_weight criter
	StandardScalerWrapper, XGBoostClassifier			0.93409	100.00 %	Aug 6, 2024 5:35 PM	31s	booster : gbtree colsample_by
	SparseNormalizer, XGBoostClassifier			0.93306	100.00 %	Aug 6, 2024 5:34 PM	26s	booster : gbtree colsample_by
	SparseNormalizer, XGBoostClassifier			0.93255	100.00 %	Aug 6, 2024 6:07 PM	46s	booster : gbtree colsample_by

Page 1 of 2 25/Page

Choose one we can see the training details of the model:



… > Mie1628a5 > Jobs > mie1628a5 > kind_roti_24xgfs3j12 > lucid_garden_j5xg4zjn

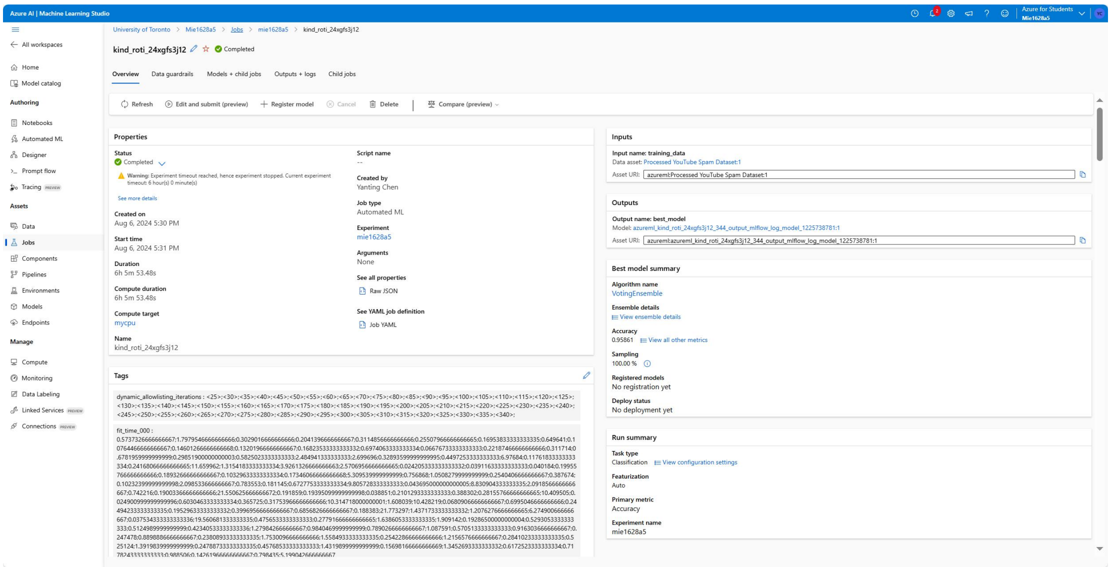
lucid_garden_j5xg4zjn ✓ Completed

accuracy	AUC_macro	AUC_micro	AUC_weighted	average_precision_sco...	average_precision_sco...	average_precision_sco...	balanced_accuracy	f1_score_macro	f1_score_micro	f1_score_weighted	log_loss
0.9504341	0.9856941	0.9867593	0.9856941	0.9852011	0.9868756	0.9853056	0.9510854	0.9504115	0.9504341	0.9504496	0.1816422

matthews_correlation	norm_macro_recall	precision_score_macro	precision_score_micro	precision_score_weigh...	recall_score_macro
0.9018360	0.9021708	0.9507512	0.9504341	0.9514763	0.9510854

	False	True
False	926	29
True	68	934

Job completed



University of Toronto > Mie1628a5 > Jobs > mie1628a5 > kind_roti_24xgfs3j12

kind_roti_24xgfs3j12 ✓ Completed

Properties

Status
✓ Completed

⚠ Warning: Experiment timeout reached, hence experiment stopped. Current experiment timeout: 6 hour(s) 0 minute(s)

See more details

Created on
Aug 6, 2024 5:30 PM

Start time
Aug 6, 2024 5:31 PM

Duration
6h 5m 53.48s

Compute duration
6h 5m 53.48s

Compute target
mycpu

Name
kind_roti_24xgfs3j12

Script name
--

Created by
Yanting Chen

Job type
Automated ML

Experiment
mie1628a5

Arguments
None

See all properties
Raw JSON

See YAML job definition
Job YAML

Tags

```
dynamic_allowlisting_iterations : <25>;<30>;<35>;<40>;<45>;<50>;<55>;<60>;<65>;<70>;<75>;<80>;<85>;<90>;<95>;<100>;<105>;<110>;<115>;<120>;<125>;
<130>;<135>;<140>;<145>;<150>;<155>;<160>;<165>;<170>;<175>;<180>;<185>;<190>;<195>;<200>;<205>;<210>;<215>;<220>;<225>;<230>;<235>;<240>;
<245>;<250>;<255>;<260>;<265>;<270>;<275>;<280>;<285>;<290>;<295>;<300>;<305>;<310>;<315>;<320>;<325>;<330>;<335>;<340>;
```

```
fit_time_000 :
0.5737326666666667;1.7979546666666666;0.3029016666666666;0.20413966666666667;0.3114856666666666;0.25507966666666665;0.16953833333333335;0.649641;0.1
0764466666666667;0.14601266666666668;0.13201966666666667;0.16823533333333332;0.6974063333333334;0.06676733333333333;0.22187466666666666;0.311714;0
.6781959999999999;0.29851900000000003;0.5825023333333333;2.484941333333333;2.699696;0.32893599999999995;0.4497253333333333;6.97684;0.11761833333333
334;0.24168066666666665;11.659962;1.3154183333333334;3.9261326666666663;2.5706956666666665;0.024205333333333332;0.03911633333333333;0.040184;0.19955
766666666666;0.18932666666666667;0.10329633333333334;0.17346066666666668;5.309539999999999;0.756868;1.0508279999999999;0.2540406666666667;0.387674;
0.10232399999999998;2.098533666666667;0.783553;0.181145;0.6727753333333334;9.805728333333333;0.043695000000000005;8.830904333333335;2.0918566666666
667;0.742216;0.19003366666666666;21.550625666666672;0.191859;0.19395099999999998;0.038851;0.2101293333333333;0.388302;0.28155766666666665;10.409505;0.
024900999999999996;0.6030463333333334;0.365725;0.31753966666666666;10.314718000000001;1.608039;10.428219;0.06809066666666667;0.6995046666666666;0.24
494233333333335;0.19529633333333332;0.39969566666666667;0.6856826666666667;0.188383;21.773297;1.4371733333333332;1.2076276666666665;6.274900666666
667;0.037534333333333336;19.560681333333335;0.4756533333333333;0.2779166666666665;1.6386053333333335;1.909142;0.19286500000000004;0.5293053333333
333;0.5124989999999999;0.42340533333333336;1.279842666666667;0.9840469999999999;0.7890266666666667;1.087591;0.5705133333333333;0.9163036666666667;0.
247478;0.8898886666666667;0.23808933333333335;1.7530096666666666;1.5584933333333335;0.2542286666666666;1.2156576666666667;0.28410233333333335;0.5
25124;1.3919839999999999;0.24788733333333335;0.4576853333333333;1.4319899999999999;0.15698166666666669;1.3452693333333332;0.6172523333333334;0.71
78243333333333;0.988506;0.14261966666666667;0.798435;5.199042666666667
```

Inputs

Input name: training_data
Data asset: Processed YouTube Spam Dataset:1
Asset URI: azureml:Processed YouTube Spam Dataset:1

Outputs

Output name: best_model
Model: azureml_kind_roti_24xgfs3j12_344_output_mlflow_log_model_1225738781:1
Asset URI: azureml:azureml_kind_roti_24xgfs3j12_344_output_mlflow_log_model_1225738781:1

Best model summary

Algorithm name
VotingEnsemble

Ensemble details
View ensemble details

Accuracy
0.95861 View all other metrics

Sampling
100.00 %

Registered models
No registration yet

Deploy status
No deployment yet

Run summary

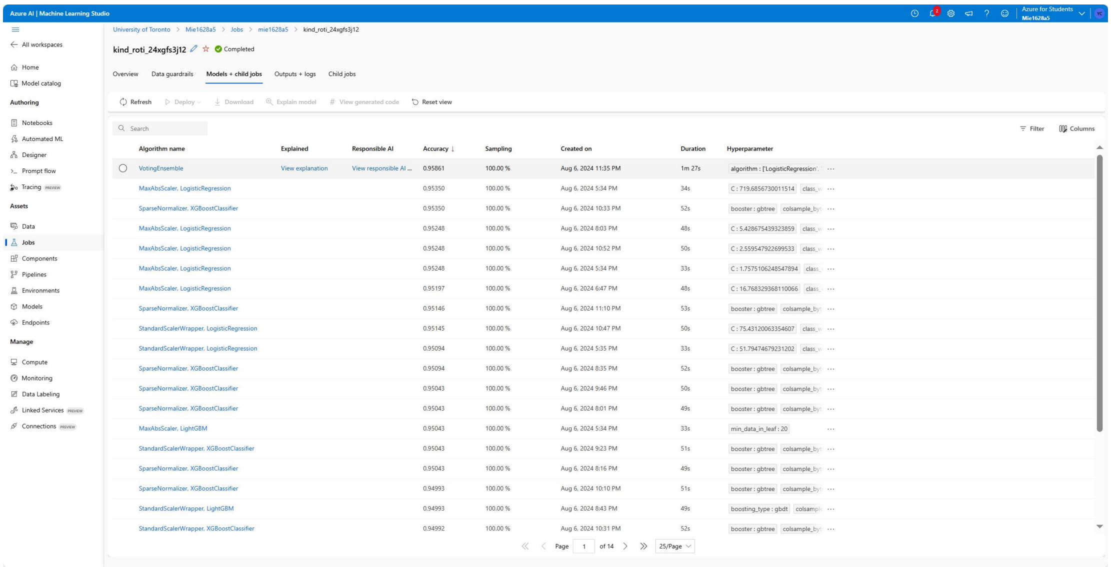
Task type
Classification View configuration settings

Featurization
Auto

Primary metric
Accuracy

Experiment name
mie1628a5

We can expolre the best model



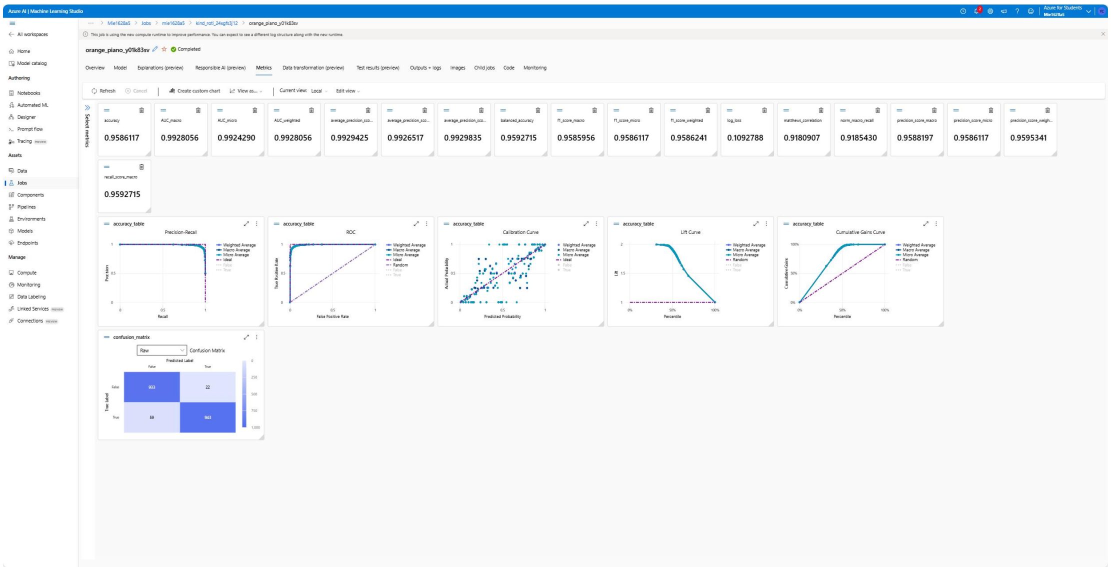
University of Toronto > Mie1628a5 > Jobs > mie1628a5 > kind_roti_24xgfs3j12

kind_roti_24xgfs3j12 ☆ Completed

	Algorithm name	Explained	Responsible AI	Accuracy ↓	Sampling	Created on	Duration	Hyperparameter
○	VotingEnsemble	View explanation	View responsible AI ...	0.95861	100.00 %	Aug 6, 2024 11:35 PM	1m 27s	algorithm : ['LogisticRegression', ···
	MaxAbsScaler, LogisticRegression			0.95350	100.00 %	Aug 6, 2024 5:34 PM	34s	C : 719.6856730011514 class_w ···
	SparseNormalizer, XGBoostClassifier			0.95350	100.00 %	Aug 6, 2024 10:33 PM	52s	booster : gbtree colsample_byt ···
	MaxAbsScaler, LogisticRegression			0.95248	100.00 %	Aug 6, 2024 8:03 PM	48s	C : 5.428675439323859 class_w ···
	MaxAbsScaler, LogisticRegression			0.95248	100.00 %	Aug 6, 2024 10:52 PM	50s	C : 2.559547922699533 class_w ···
	MaxAbsScaler, LogisticRegression			0.95248	100.00 %	Aug 6, 2024 5:34 PM	33s	C : 1.7575106248547894 class_ ···
	MaxAbsScaler, LogisticRegression			0.95197	100.00 %	Aug 6, 2024 6:47 PM	48s	C : 16.768329368110066 class_ ···
	SparseNormalizer, XGBoostClassifier			0.95146	100.00 %	Aug 6, 2024 11:10 PM	53s	booster : gbtree colsample_byt ···
	StandardScalerWrapper, LogisticRegression			0.95145	100.00 %	Aug 6, 2024 10:47 PM	50s	C : 75.43120063354607 class_w ···
	StandardScalerWrapper, LogisticRegression			0.95094	100.00 %	Aug 6, 2024 5:35 PM	33s	C : 51.79474679231202 class_w ···
	SparseNormalizer, XGBoostClassifier			0.95094	100.00 %	Aug 6, 2024 8:35 PM	52s	booster : gbtree colsample_byt ···
	SparseNormalizer, XGBoostClassifier			0.95043	100.00 %	Aug 6, 2024 9:46 PM	50s	booster : gbtree colsample_byt ···
	SparseNormalizer, XGBoostClassifier			0.95043	100.00 %	Aug 6, 2024 8:01 PM	49s	booster : gbtree colsample_byt ···
	MaxAbsScaler, LightGBM			0.95043	100.00 %	Aug 6, 2024 5:34 PM	33s	min_data_in_leaf : 20 ···
	StandardScalerWrapper, XGBoostClassifier			0.95043	100.00 %	Aug 6, 2024 9:23 PM	51s	booster : gbtree colsample_byt ···
	SparseNormalizer, XGBoostClassifier			0.95043	100.00 %	Aug 6, 2024 8:16 PM	49s	booster : gbtree colsample_byt ···
	SparseNormalizer, XGBoostClassifier			0.94993	100.00 %	Aug 6, 2024 10:10 PM	51s	booster : gbtree colsample_byt ···
	StandardScalerWrapper, LightGBM			0.94993	100.00 %	Aug 6, 2024 8:43 PM	49s	boosting_type : gbdt colsample ···
	StandardScalerWrapper, XGBoostClassifier			0.94992	100.00 %	Aug 6, 2024 10:31 PM	52s	booster : gbtree colsample_byt ···

Page 1 of 14 25/Page

Choose the best model “VotingEnsemble” which achieved the highest accuracy



··· > Mie1628a5 > Jobs > mie1628a5 > kind_roti_24xgfs3j12 > orange_piano_y01k83sv

This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

orange_piano_y01k83sv ☆ Completed

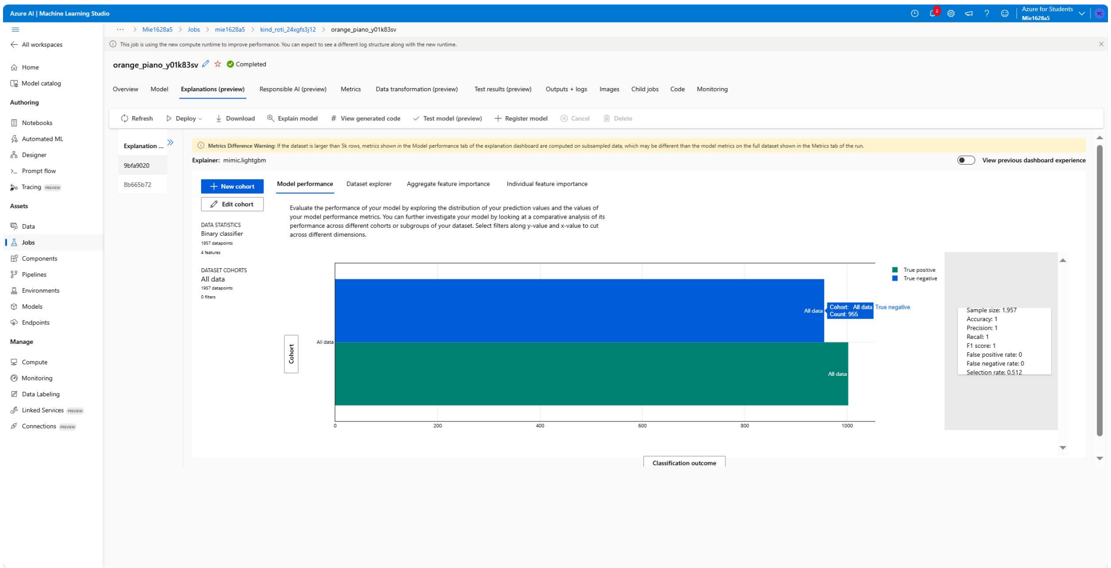
Current view: Local

accuracy	AUC_macro	AUC_micro	AUC_weighted	average_precision_sco...	average_precision_sco...	average_precision_sco...	balanced_accuracy	f1_score_macro	f1_score_micro	f1_score_weighted	log_loss	matthews_correlation	norm_macro_recall	precision_score_macro	precision_score_micro	precision_score_weigh...	recall_score_macro
0.9586117	0.9928056	0.9924290	0.9928056	0.9929425	0.9926517	0.9929835	0.9592715	0.9585956	0.9586117	0.9586241	0.1092788	0.9180907	0.9185430	0.9588197	0.9586117	0.9595341	0.9592715

	False	True
False	933	22
True	59	943

From the **Explanations** dashboard, we can get insight into this trained model:

Model performance:



This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

orange_piano_y01k83sv ✓ Completed

Metrics Difference Warning: If the dataset is larger than 5k rows, metrics shown in the Model performance tab of the explanation dashboard are computed on subsampled data, which may be different than the model metrics on the full dataset shown in the Metrics tab of the run.

Explainer: mimic.lightgbm

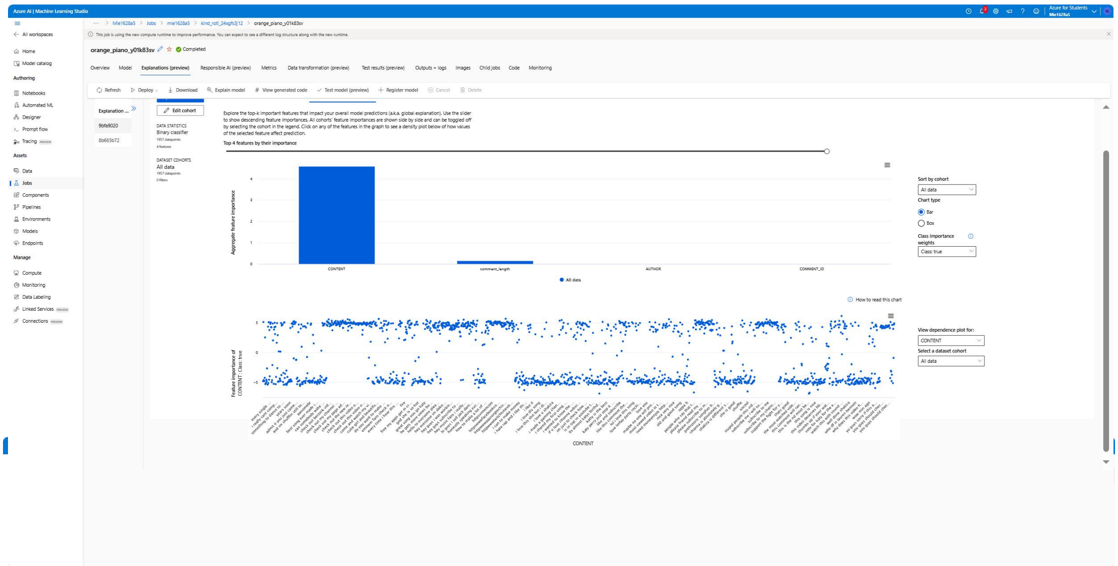
DATA STATISTICS
Binary classifier
1957 datapoints
4 features

DATASET COHORTS
All data
1957 datapoints
0 filters

Evaluate the performance of your model by exploring the distribution of your prediction values and the values of your model performance metrics. You can further investigate your model by looking at a comparative analysis of its performance across different cohorts or subgroups of your dataset. Select filters along y-value and x-value to cut across different dimensions.

Sample size: 1,957
Accuracy: 1
Precision: 1
Recall: 1
F1 score: 1
False positive rate: 0
False negative rate: 0
Selection rate: 0.512

Explore the top-k important features that impact the overall model predictions:



orange_piano_y01k83sv ✓ Completed

DATA STATISTICS
Binary classifier
1957 datapoints
4 features

DATASET COHORTS
All data
1957 datapoints
0 filters

Explore the top-k important features that impact your overall model predictions (a.k.a. global explanation). Use the slider to show descending feature importances. All cohorts' feature importances are shown side by side and can be toggled off by selecting the cohort in the legend. Click on any of the features in the graph to see a density plot below of how values of the selected feature affect prediction.

Top 4 features by their importance

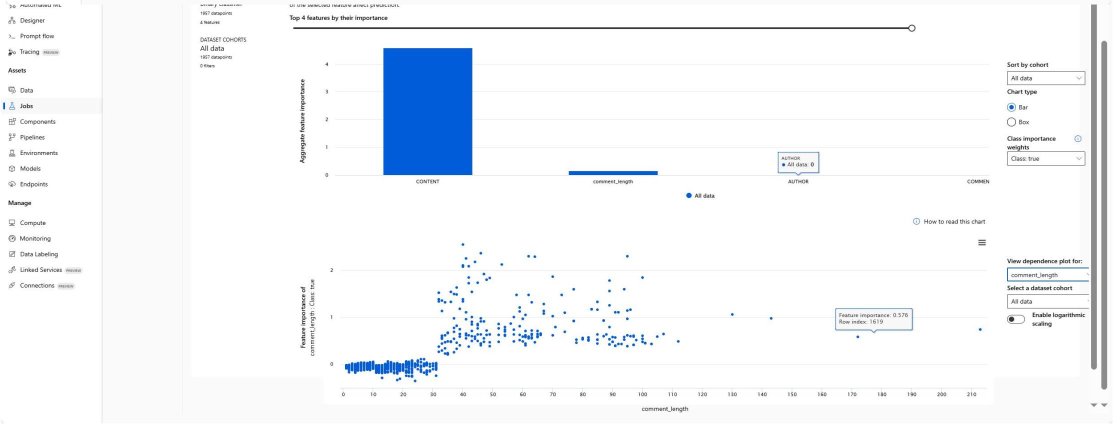
Sort by cohort: All data

Chart type: ● Bar ○ Box

Class importance weights: Class: true

View dependence plot for: CONTENT

Select a dataset cohort: All data



1957 datapoints
4 features

DATASET COHORTS
All data
1957 datapoints
0 filters

of the selected feature affect prediction.

Top 4 features by their importance

Sort by cohort: All data

Chart type: ● Bar ○ Box

Class importance weights: Class: true

View dependence plot for: comment_length

Select a dataset cohort: All data

○ Enable logarithmic scaling