

Part 0 : Preparation

Starting by entering the deployment space to the overview page

The screenshot shows the IBM WatsonX interface. At the top, there's a navigation bar with tabs for Overview, Assets, Deployments, Jobs, and Manage. The Overview tab is selected. Below the navigation bar, there's a section titled "Jump back in" with a link to "final-P2_hpo_d_output" (Deployed 6 minutes ago). To the right, there are two summary boxes: "Deployments" (1 Deployed, 0 Failed) and "Space history" (No notifications). Further down, there are sections for "Job runs" (0 Active, 0 Failed last 24 hours) and "AI governance".

Then, go to the deployments tab and click on the deployment which you have deployed earlier

The screenshot shows the IBM WatsonX interface, specifically the Deployments Factsheet for a deployment named "final-P2_hpo_d_output". The left side of the screen displays a table with columns for Name, Type, Status, Tags, and Last modified. The table shows one row for "Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8" with status "Online" and "Deployed". The right side of the screen contains detailed information about the asset, including its name, description (No description provided), asset details (Type: wml-hybrid_01, Model ID: 822a356c-baea-4f..., Software specification: hybrid_0.1), tags (parent_run_0b6cad9b-f2f5-41fe-9070-2c137c18), and metadata like last modified (1 minute ago by Service) and created on (Oct 7, 2024 by Chananop Wannatipaporn).

Once you clicked on the deployment, you should be entering this page

The screenshot shows the deployment details for a model named 'Deployment for model 822a356c-baea-4fcf-a098-12d32ba84ef8'. The status is 'Deployed' and 'Online'. The 'API reference' tab is selected, showing endpoints for inferencing:

- Private endpoint: <https://private.us-south.ml.cloud.ibm.com/v4/deployments/d1b9903c-9284-4c3d-ae19-37e8b9f63bfa/predictions?version=2021-05-01>
- Public endpoint: <https://us-south.ml.cloud.ibm.com/v4/deployments/d1b9903c-9284-4c3d-ae19-37e8b9f63bfa/predictions?version=2021-05-01>

Code snippets section:

- cURL
- Java
- JavaScript
- Python
- Scala

```
# NOTE: you must set $API_KEY below using information retrieved from your IBM Cloud account (https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/ml-authentication.html#api-key)
curl --insecure -X POST -H "Content-Type: application/x-www-form-urlencoded" -H "Accept: application/json" -H "data-urlencode grant_type=urn:ibm:params:oauth:grant-type:apikey" -H "data-urlencode apikey=$API_KEY" "https://iam.cloud.ibm.com/identity/token"

# The above CURL request will return an auth token that you will use as $IAM_TOKEN in the scoring request below
# TODO: manually define and pass values to be scored below
curl -X POST -H "Content-Type: application/json" -H "Accept: application/json" -H "Authorization: Bearer $IAM_TOKEN" -d '[{"input_data": [{"fields": [{"array": "INPUT_FIELDS"}, {"values": [{"array": "VALUES_TO_BE_SCORED"}, {"array": "$ANOTHER_ARRAY_OF_VALUES_TO_BE_SCORED"}]}]}, {"url": "https://private.us-south.ml.cloud.ibm.com/v4/deployments/d1b9903c-9284-4c3d-ae19-37e8b9f63bfa/predictions?version=2021-05-01"}]]'
```

About this deployment:

- Name: Deployment for model 822a356c-baea-4fcf-a098-12d32ba84ef8
- Description: No description provided.
- Deployment Details:
 - Deployment ID: d1b9903c-9284-4c3d-ae19-37e8b9f63bfa
 - Serving name: No serving name.
 - Software specification: hybrid_0.1
 - Hybrid pipeline software specifications: autoai-kb_r24.1-py3.11
 - Copies: 1
- Tags: Add tags to make assets easier to find.
- Associated asset: final-P2_hpo_d_output
- Last modified: 1 minute ago
- Created on: Oct 7, 2024

Part 1 : Configuring model evaluations

Click on Evaluations tab then click on Configure OpenScale evaluation settings

The screenshot shows the deployment details for the same model. The 'Evaluations' tab is selected. The page displays the following content:

Evaluation results
View Watson OpenScale evaluation results to gain insights about your model performance

Start configuring model evaluations
To view evaluation results, configure evaluations to monitor the deployment for fairness or performance drift.

Configure OpenScale evaluation settings

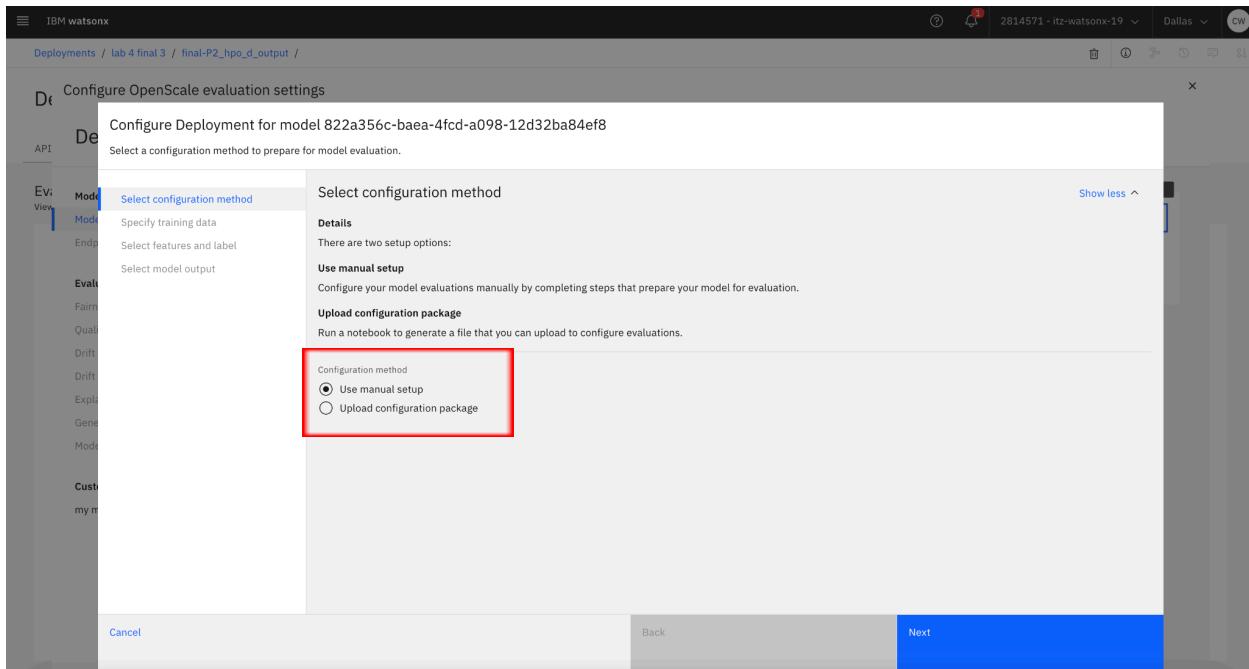
Once you click on the Configure OpenScale evaluation settings, you should be entering this page. Set the data type and algorithm type according to your data asset. When done click view summary.

The screenshot shows the 'Prepare for evaluation' configuration page. It includes sections for 'Provide model information' and 'Details'. A red box highlights the 'Data type' and 'Algorithm type' dropdowns. The 'Data type' dropdown is set to 'Numeric/categorical' and the 'Algorithm type' dropdown is set to 'Binary classification'. At the bottom right, there is a blue button labeled 'View summary'.

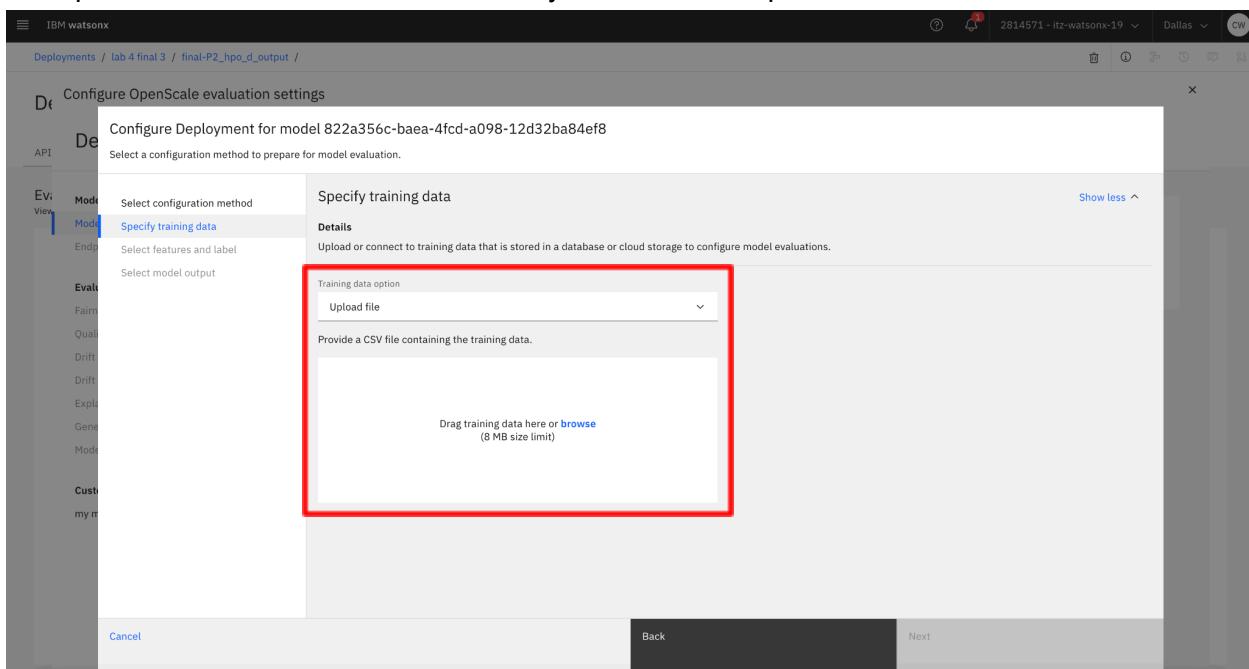
After clicking view summary, check your deployed model, model input data type, algorithm type and feature columns whether it's correct or not. Once it's all correct click save and continue.

The screenshot shows the 'Review setup for Deployment' summary page. It lists four items with radio buttons: 'Deployed model' (selected), 'Model input data type' (set to 'Numeric/categorical'), 'Algorithm type' (set to 'Binary classification'), and 'Feature columns' (list of columns: ApplicantIncome, CoapplicantIncome, Credit_History, Dependents, Education, Gender, Loan_Amount_Term, LoanAmount, LoanAmount_Term, Married, Property_Area, Self_Employed). A red box highlights the 'Save and continue' button at the bottom right.

Once you click save and continue, you should now be at this page. Select your configuration method as use manual setup then click next.



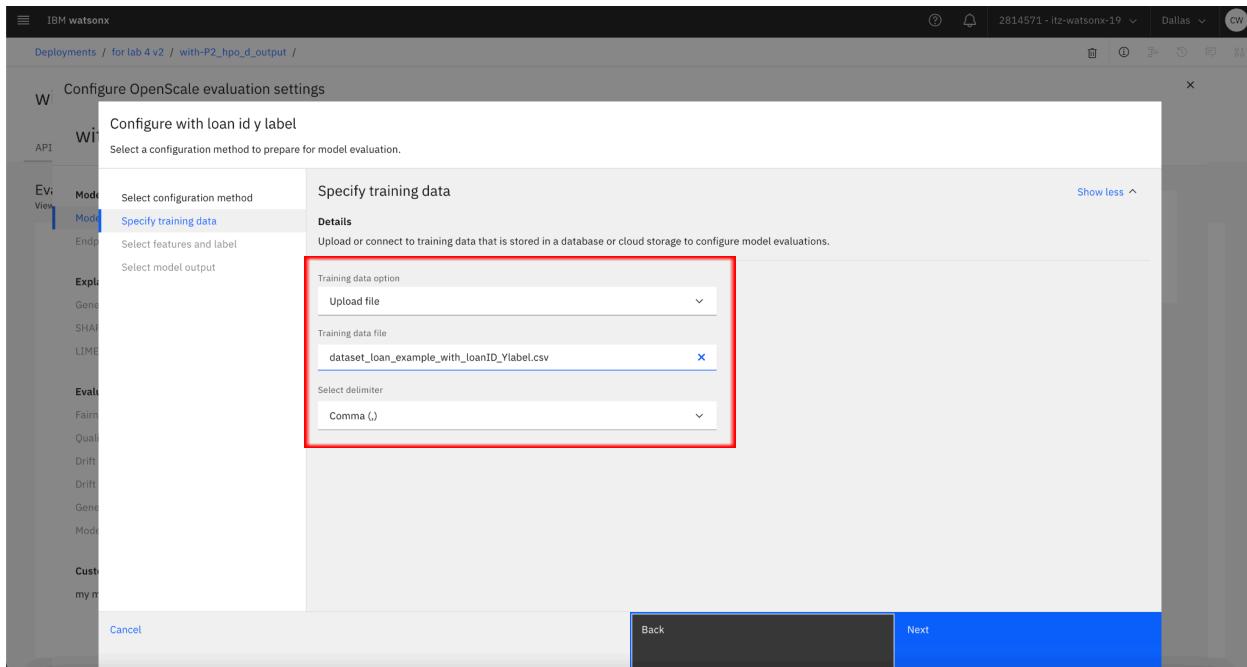
After that, upload your training data. However, the label column of the data file must not include a unique value and need to be named as 'y' as in the example data format below



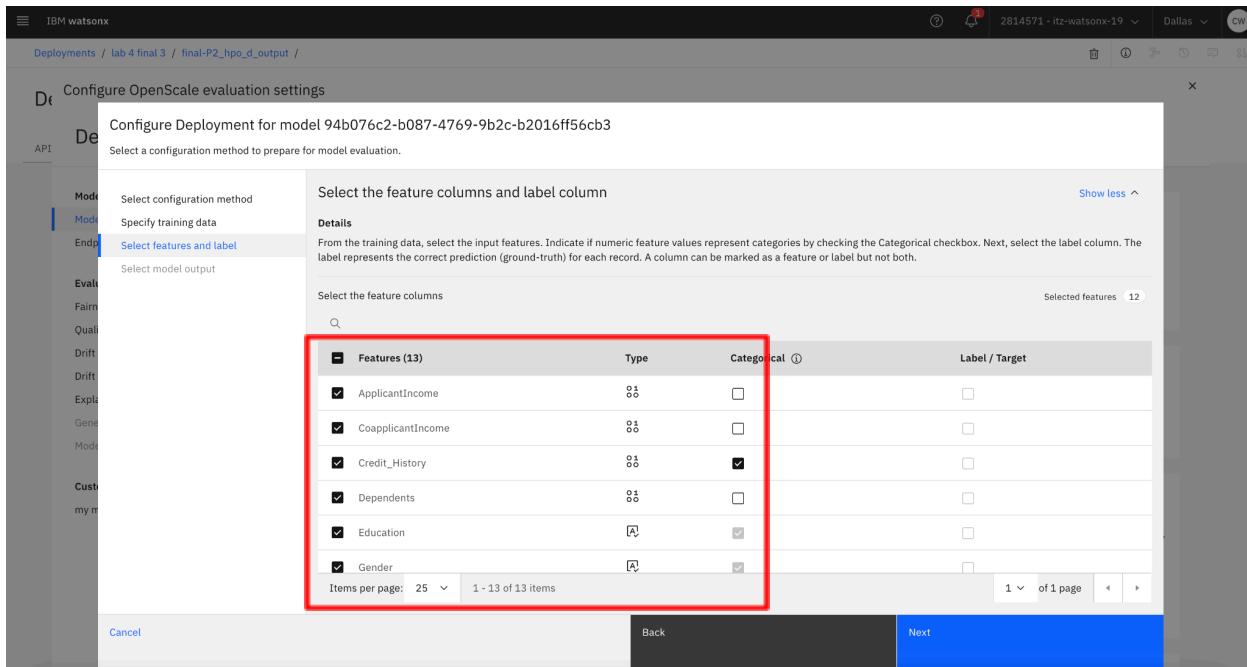
Data format should have no unique value and required the label column to be named as 'y'

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	LoanAmount_Term	y
Male	Yes	1	Graduate	No	4583	1508	128	360.0	1.0	Rural	360	N
Male	Yes	0	Graduate	Yes	3000	0	66	360.0	1.0	Urban	360	Y
Male	Yes	0	Not Graduate	No	2583	2358	120	360.0	1.0	Urban	360	Y
Male	No	0	Graduate	No	6000	0	141	360.0	1.0	Urban	360	Y
Male	Yes	2	Graduate	Yes	5417	4196	267	360.0	1.0	Urban	360	Y
Male	Yes	0	Not Graduate	No	2333	1516	95	360.0	1.0	Urban	360	Y
Male	Yes	3	Graduate	No	3036	2504	158	360.0	0.0	Semiurban	360	N
Male	Yes	2	Graduate	No	4006	1526	168	360.0	1.0	Urban	360	Y
Male	Yes	1	Graduate	No	12841	10968	349	360.0	1.0	Semiurban	360	N
Male	Yes	2	Graduate	No	3200	700	70	360.0	1.0	Urban	360	Y
Male	Yes	2	Graduate		2500	1840	109	360.0	1.0	Urban	360	Y
Male	Yes	2	Graduate	No	3073	8106	200	360.0	1.0	Urban	360	Y
Male	No	0	Graduate	No	1853	2840	114	360.0	1.0	Rural	360	N
Male	Yes	2	Graduate	No	1299	1086	17	120.0	1.0	Urban	120	Y
Male	No	0	Graduate	No	4950	0	125	360.0	1.0	Urban	360	Y
Male	No	1	Not Graduate	No	3596	0	100	240.0		Urban	240	Y
Female	No	0	Graduate	No	3510	0	76	360.0	0.0	Urban	360	N
Male	Yes	0	Not Graduate	No	4887	0	133	360.0	1.0	Rural	360	N
Male	Yes	0	Not Graduate	No	7660	0	104	360.0	0.0	Urban	360	N
Male	Yes	1	Graduate	No	5955	5625	315	360.0	1.0	Urban	360	Y
Male	Yes	0	Not Graduate	No	2600	1911	116	360.0	0.0	Semiurban	360	N
	Yes	2	Not Graduate	No	3365	1917	112	360.0	0.0	Rural	360	N
Male	Yes	1	Graduate		3717	2925	151	360.0		Semiurban	360	N
Male	Yes	0	Graduate	Yes	9560	0	191	360.0	1.0	Semiurban	360	Y
Male	Yes	0	Graduate	No	2799	2253	122	360.0	1.0	Semiurban	360	Y
Male	Yes	2	Not Graduate	No	4226	1040	110	360.0	1.0	Urban	360	Y
Male	No	0	Not Graduate	No	1442	0	35	360.0	1.0	Urban	360	N
Female	No	2	Graduate		3750	2083	120	360.0	1.0	Semiurban	360	Y
Male	Yes	1	Graduate		4166	3369	201	360.0		Urban	360	N
Male	No	0	Graduate	No	3167	0	74	360.0	1.0	Urban	360	N
Male	No	1	Graduate	Yes	4692	0	106	360.0	1.0	Rural	360	N
Male	Yes	0	Graduate	No	3500	1667	114	360.0	1.0	Semiurban	360	Y
Male	No	3	Graduate	No	12500	3000	320	360.0	1.0	Rural	360	N

After you are done uploading the data file, select the delimiter according to your file format. Which for .csv is Comma (,). Then click next to continue onto the next page.



Once you are done with the previous step, now you should be landing on this page. Select the column type for feature column (categorical/non-categorical)



Then select label column which in this place is 'y' by ticking in this area

Configure Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Select a configuration method to prepare for model evaluation.

Select the feature columns and label column

Details

From the training data, select the input features. Indicate if numeric feature values represent categories by checking the Categorical checkbox. Next, select the label column. The label represents the correct prediction (ground-truth) for each record. A column can be marked as a feature or label but not both.

Select the feature columns

Column	Type	Categorical	Selected
ApplicantIncome	0.1	<input type="checkbox"/>	<input type="checkbox"/>
CoapplicantIncome	0.1	<input type="checkbox"/>	<input type="checkbox"/>
LoanAmount	0.0	<input type="checkbox"/>	<input type="checkbox"/>
Loan_Amount_Term	0.1	<input type="checkbox"/>	<input type="checkbox"/>
Credit_History	0.0	<input type="checkbox"/>	<input type="checkbox"/>
Property_Area	A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
LoanAmount_Term	0.1	<input type="checkbox"/>	<input type="checkbox"/>
y	A	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Items per page: 25 | 1 - 13 of 13 items

Selected features: 12

Cancel Back Next

Once you done marking all the features and label column to its type, click next.

Configure Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Select a configuration method to prepare for model evaluation.

Select the feature columns and label column

Details

From the training data, select the input features. Indicate if numeric feature values represent categories by checking the Categorical checkbox. Next, select the label column. The label represents the correct prediction (ground-truth) for each record. A column can be marked as a feature or label but not both.

Select the feature columns

Column	Type	Categorical	Selected
ApplicantIncome	0.0	<input type="checkbox"/>	<input type="checkbox"/>
CoapplicantIncome	0.1	<input type="checkbox"/>	<input type="checkbox"/>
LoanAmount	0.0	<input type="checkbox"/>	<input type="checkbox"/>
Loan_Amount_Term	0.1	<input type="checkbox"/>	<input type="checkbox"/>
Credit_History	0.0	<input type="checkbox"/>	<input type="checkbox"/>
Property_Area	A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
LoanAmount_Term	0.0	<input type="checkbox"/>	<input type="checkbox"/>
y	A	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Items per page: 25 | 1 - 13 of 13 items

Selected features: 12

Cancel Back Next

Then continue with selecting the model output by marking prediction as 'Prediction' and probability as 'Probability'.

Configure Deployment for model 822a356c-baea-4fcf-a098-12d32ba84ef8

Select a configuration method to prepare for model evaluation.

Select model output

Details

From the model output data, select the column that contains the prediction generated by the deployed model. Select the prediction probability column which contains the model's confidence in the prediction.

Select the prediction and probability column(s).

Features (2)	Type	Prediction	Probability
prediction	A	<input checked="" type="checkbox"/>	
probability	B		<input checked="" type="checkbox"/>

No probability column selected
To support explainability, drift detection, and debiasing for binary classification you must select one or two prediction probability columns

Cancel Back View summary

Once done marking, click 'View Summary'.

Configure Deployment for model 822a356c-baea-4fcf-a098-12d32ba84ef8

Select a configuration method to prepare for model evaluation.

Select model output

Details

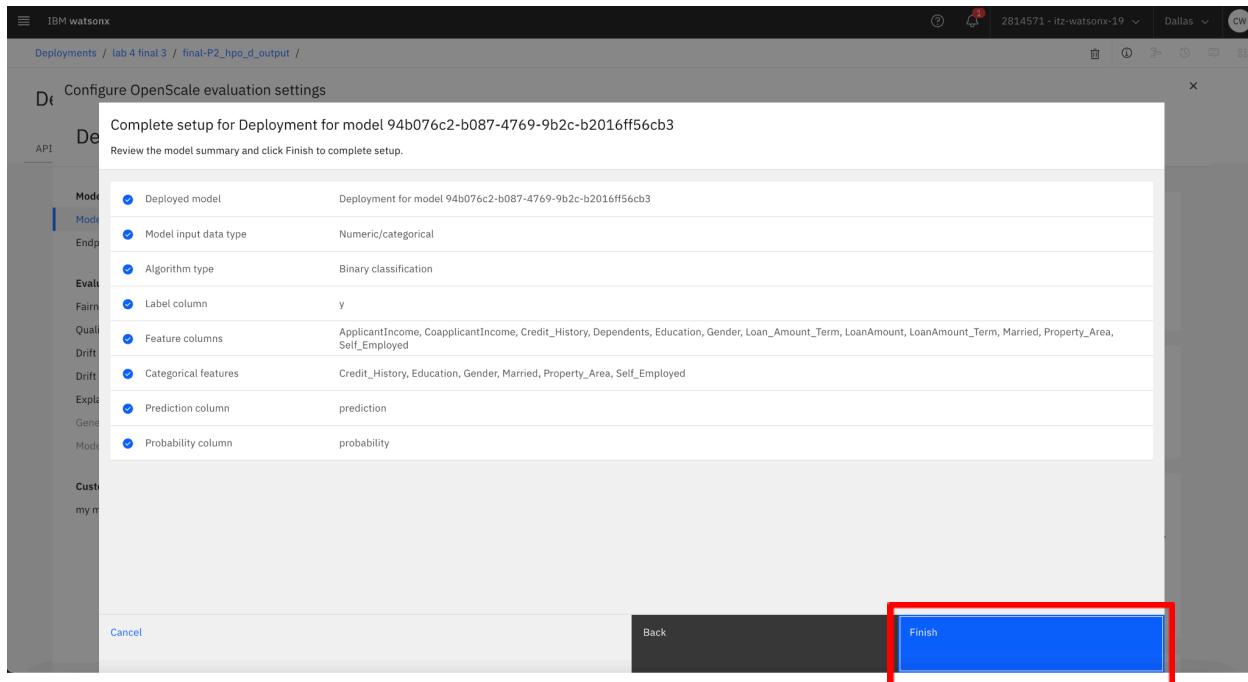
From the model output data, select the column that contains the prediction generated by the deployed model. Select the prediction probability column which contains the model's confidence in the prediction.

Select the prediction and probability column(s).

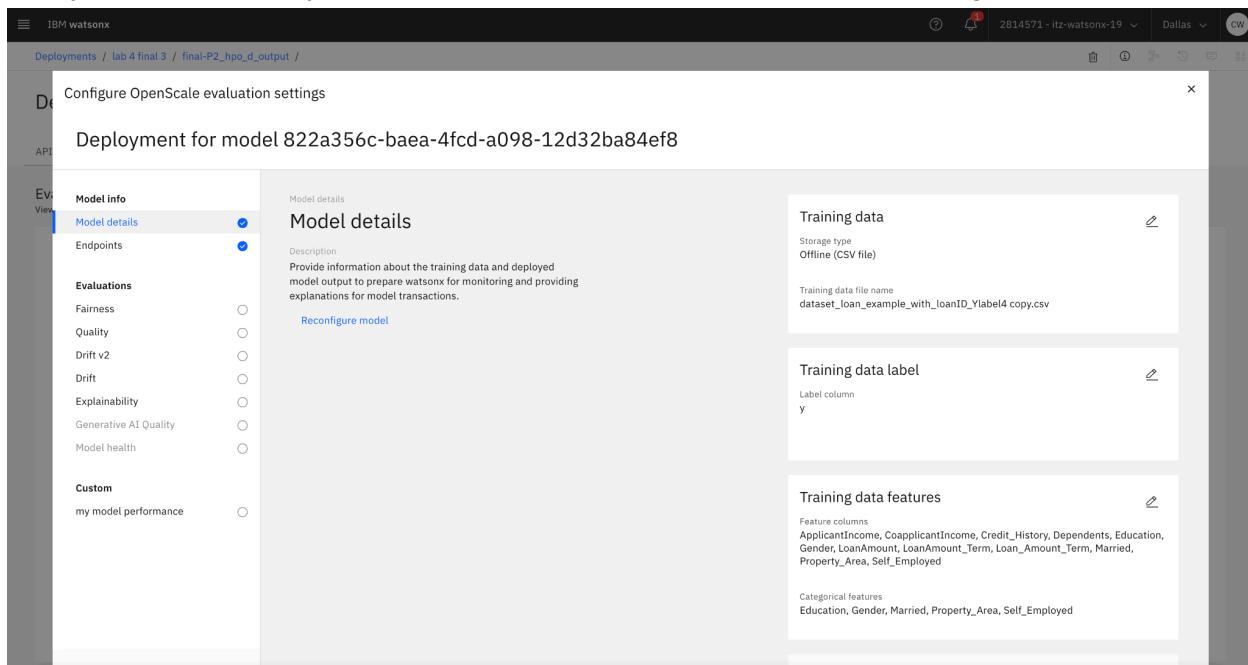
Features (2)	Type	Prediction	Probability
prediction	A	<input checked="" type="checkbox"/>	
probability	B		<input checked="" type="checkbox"/>

View summary

Check all your setup and make sure that what you choose aligns with your dataset. Once done, click finish.



Now you have finished your model setup and should be able to see this page.



Part 2 : Configuring model explainability

Go to general settings under the explainability topic and you should be able to see this page.

After that click on the edit icon of the explanation method to start configuring the model explanation method.

The screenshot shows the 'Configure OpenScale evaluation settings' page for a deployment. On the left, there's a sidebar with sections like Model info, Endpoints, Evaluations, and Custom. Under Evaluations, 'Explainability' is selected and highlighted with a red box and the number '1.'. In the main content area, there's a 'Parameters' section with an edit icon (highlighted with a red box and the number '2.') and other sections for Controllable features and Language support.

Then choose 'LIME (enhanced)' which is a local explanation as an explanation method and click next to continue onto the next page.

This screenshot shows the same configuration page as before, but with the 'Explainability' method selected in the sidebar. In the 'Parameters' section, there's a dropdown for 'Local explanation method' with two options: 'SHAP (Kernel Explainer)' and 'LIME (enhanced)'. The 'LIME (enhanced)' option is selected and highlighted with a red box. At the bottom right, there are 'Cancel' and 'Next' buttons, with 'Next' being highlighted in blue.

Set number of perturbations per record, it is a measure used to indicate how frequently each data record has been subject to modifications during the perturbation-based analysis.

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

LIME settings

Number of perturbations per record

Set this to 0 to automatically choose a suitable count

Select the features that are controllable.

Feature	Type	Controllable
Gender	A	<input checked="" type="checkbox"/> On
Married	A	<input checked="" type="checkbox"/> On
Dependents	BB	<input checked="" type="checkbox"/> On
Education	A	<input checked="" type="checkbox"/> On
Self_Employed	A	<input checked="" type="checkbox"/> On
ApplicantIncome	BB	<input checked="" type="checkbox"/> On
CoapplicantIncome	BB	<input checked="" type="checkbox"/> On
LoanAmount	BB	<input checked="" type="checkbox"/> On
Loan_Amount_Term	BB	<input checked="" type="checkbox"/> On

Once done choosing click save

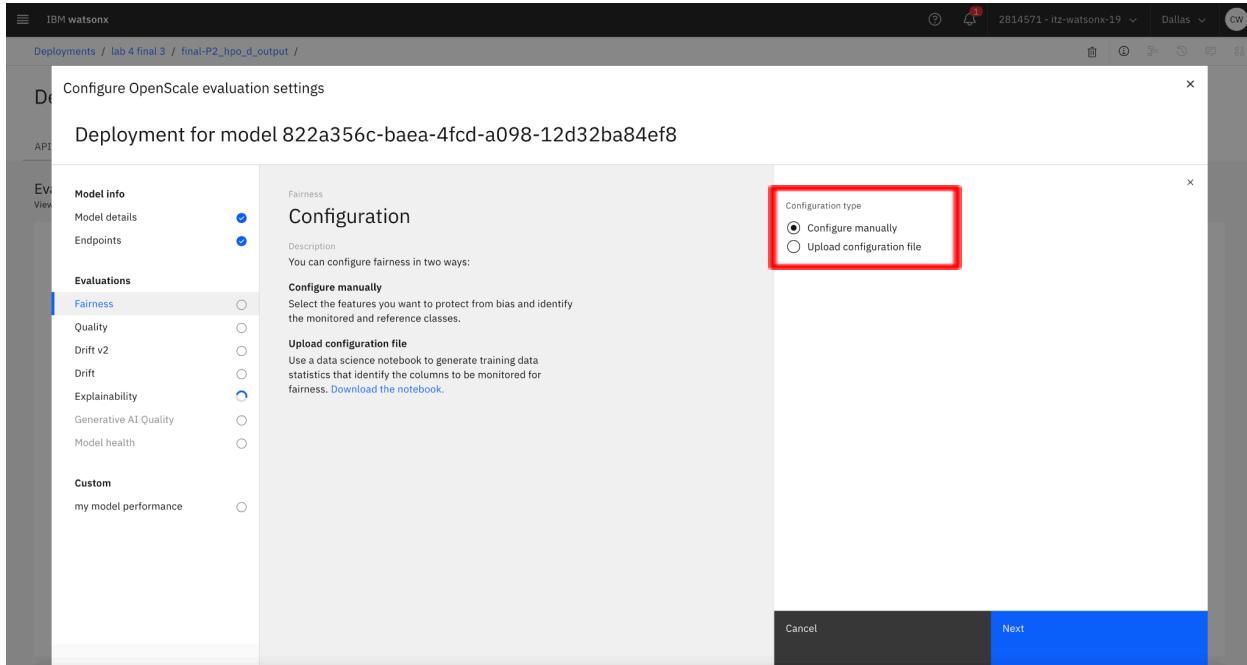
The screenshot shows the 'Configure OpenScale evaluation settings' interface for a deployment. On the left, there's a sidebar with 'Evaluations' selected. The main area shows 'Controllable features' with a list of various model inputs like Gender, Married, Dependents, etc., each with a toggle switch. Most switches are off, except for 'LoanAmount' and 'Loan_Amount_Term' which are set to 'On'. At the bottom right of the main panel is a blue 'Save' button, which is highlighted with a red box.

Part 3 : Configuring model fairness monitor

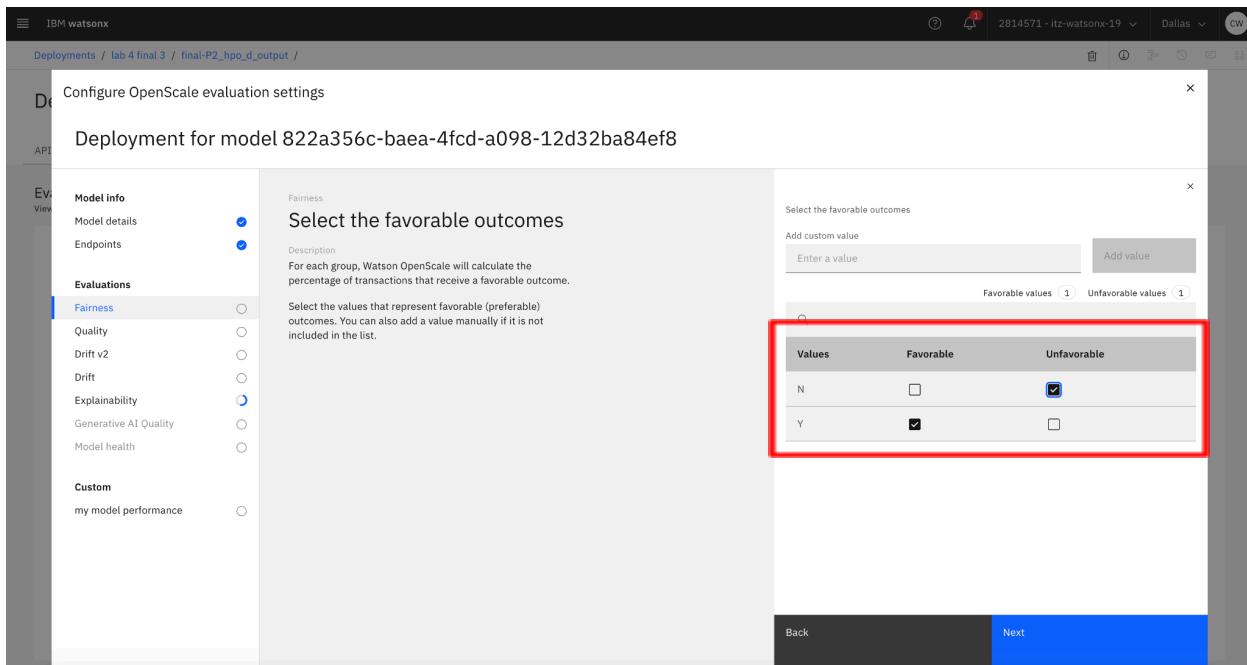
Go to Fairness under the Evaluations topic and you should be able to see this page. After that click on the edit icon of the configuration to start configuring the model fairness monitor.

The screenshot shows the 'Configure OpenScale evaluation settings' interface again, but this time the 'Fairness' section in the sidebar is highlighted with a red box. The main panel displays the 'Fairness' configuration page. It includes sections for 'Description', 'Configuration' (with an edit icon highlighted by a red box), 'Favorable outcomes' (with an edit icon), 'Sample size' (with an edit icon), and 'Metrics' (with an edit icon). The 'Fairness' section also contains a detailed description of what the monitor does.

Then choose configure manually under configuration type and click next to continue onto the next page.



Once you land on this page, choose the value of the outcomes according to the dataset ('Y' is favorable and 'N' is unfavorable)



Once you are done selecting, click next to continue.

The screenshot shows the 'Configure OpenScale evaluation settings' dialog for a deployment. On the left, there's a sidebar with 'Model info', 'Endpoints', and 'Evaluations' sections. Under 'Evaluations', 'Fairness' is selected. The main panel shows a 'Fairness' section with a sub-section titled 'Select the favorable outcomes'. It includes a description and a table for selecting favorable outcomes. The table has columns for 'Values', 'Favorable', and 'Unfavorable'. It lists two rows: 'N' (Favorable checked, Unfavorable unchecked) and 'Y' (Favorable unchecked, Unfavorable checked). At the bottom of the dialog are 'Back' and 'Next' buttons, with 'Next' being highlighted by a red box.

Set the minimum sample size to be whatever number which is smaller than the prepared dataset. Once done, click next to continue onto the next page.

The screenshot shows the 'Configure OpenScale evaluation settings' dialog for a deployment. The 'Evaluations' section on the left has 'Fairness' selected. The main panel shows a 'Sample size' section with a description. To the right is a configuration panel with fields for 'Evaluate using balanced data set', 'Minimum sample size' (set to 100), and 'Maximum sample size (optional)' (set to 10000). A red box highlights the 'Evaluate using balanced data set' checkbox and the sample size input fields. At the bottom are 'Back' and 'Next' buttons, with 'Next' being highlighted by a red box.

Select the metrics which you want to monitor which in this case we only select Disparate impact. Once done, click next to continue.

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Metrics

The Fairness monitor tracks multiple fairness metrics. You can select the metrics that will be monitored across the features in the future, and set up some default thresholds. You can always overwrite the default thresholds when setting up the monitored features. Disparate impact is pre-selected as the default fairness metric.

Select monitored metrics

Metrics generated with payload data, training data, and balanced data ⓘ

- Disparate impact
- Statistical parity difference
- Impact score

Metrics generated with feedback data ⓘ

- Average odds difference
- Average absolute odds difference
- False negative rate difference
- False positive rate difference
- False discovery rate difference
- False omission rate difference
- Error rate difference

Back Next

Set the threshold for the metrics we selected previously which in this place is Disparate impact. Once done, click next to continue.

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Metrics

The Fairness monitor tracks multiple fairness metrics. You can select the metrics that will be monitored across the features in the future, and set up some default thresholds. You can always overwrite the default thresholds when setting up the monitored features. Disparate impact is pre-selected as the default fairness metric.

Disparate impact ⓘ

Lower threshold for fairness value
80

Upper threshold for fairness value
120

Other metrics ⓘ
(Difference metrics are not selected.)

Back Next

Select field for tendency monitoring to make sure that model provides output which have acceptable fairness level. Once done, click next to continue.

Configure OpenScale evaluation settings

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Fairness

Description: For each field you select, Watson OpenScale will monitor the deployed model's tendency to provide a favorable (preferred) outcome for one group over another.

If you select a field that is not a feature column, called an added field, Watson OpenScale will look for indirect bias by finding associated values in the feature columns. For example, the profession "student" may imply a younger individual even though the Age field was excluded from model training.

Recommended features: Watson OpenScale analyzed your training data to recommend which features should be monitored for fairness. These features are identified in the Recommended column.

Fields	Recommended	Type
<input checked="" type="checkbox"/> Gender	<input checked="" type="checkbox"/>	
<input type="checkbox"/> Married	<input checked="" type="checkbox"/>	
<input type="checkbox"/> Dependents	<input checked="" type="checkbox"/>	
<input type="checkbox"/> Education	<input checked="" type="checkbox"/>	
<input type="checkbox"/> Self_Employed	<input checked="" type="checkbox"/>	
<input type="checkbox"/> ApplicantIncome	<input checked="" type="checkbox"/>	
<input type="checkbox"/> CoapplicantIncome	<input checked="" type="checkbox"/>	
<input type="checkbox"/> LoanAmount	<input checked="" type="checkbox"/>	
<input type="checkbox"/> Loan_Amount_Term	<input checked="" type="checkbox"/>	

Items per page: 25 | 1 - 12 of 12 items | 1 of 1 page | Back | Next

Select value from the selected field to be monitored or to be a reference which the output value from those two groups will be compared in order to check for potential bias. Once done, click next to continue.

Configure OpenScale evaluation settings

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Fairness

Specify the monitored groups for [Gender]

Description: Select the groups to monitor.

The percentage of favorable outcomes delivered to the monitored groups will be compared to the percentage of favorable outcomes delivered to the remaining groups (the reference groups) to check for potential bias.

A fairness score of 100% implies that the monitored group and reference group received an equal number of favorable values. Likewise, a fairness score of 50% implies that the monitored group received half as many favorable outcomes as the reference group.

Set the fairness alert threshold to track when the fairness value falls below an acceptable level.

Recommended groups: Watson OpenScale analyzed this feature to provide recommendations for group assignments. Less frequently occurring values in the training data may be suggested as monitored groups. Always review these recommended selections before proceeding.

Values	Monitored	Reference	Recommended
Female	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Monitored
Male	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Reference

Back | Next

Select threshold for the monitored field which in this place we will use the default threshold. Once done, click next to continue.

The screenshot shows the IBM Watsonx interface for configuring OpenScale evaluation settings. On the left, a sidebar lists 'Model info', 'Endpoints', and 'Evaluations'. Under 'Evaluations', 'Fairness' is selected. The main panel displays a 'Specify the monitored groups for [Gender]' section with a 'Description' and a 'Select the groups to monitor' area. To the right, a 'Set up threshold for feature [Gender]' box is highlighted with a red border, containing two radio button options: 'Use default thresholds' (selected) and 'Set up different thresholds'. Below this is a note about 'Disparate impact 80 - default'. At the bottom are 'Back' and 'Save' buttons.

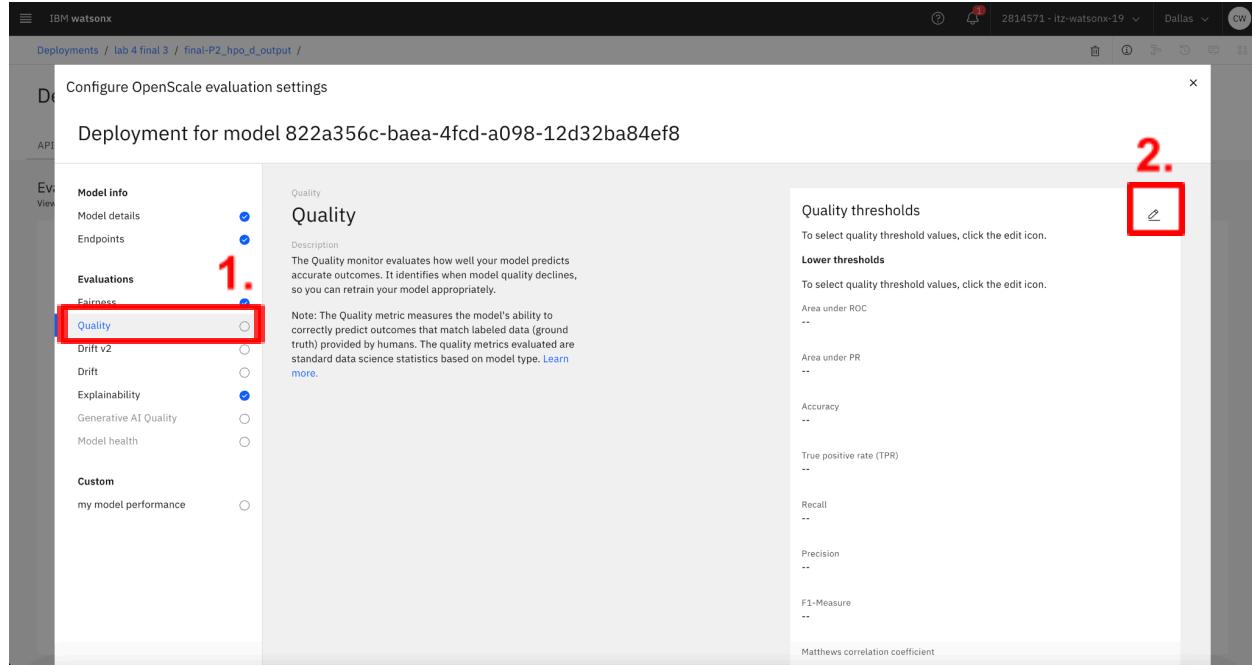
Now you have finished your model fairness monitor setup and should be able to see this page.

The screenshot shows the completed configuration of the Fairness monitor. The 'Fairness' section is now fully visible, displaying detailed configuration options. These include:

- Configuration:** Configuration file: --
- Favorable outcomes:** Favorable outcomes: Y
Unfavorable outcomes: N
- Sample size:** Minimum sample size: 100
Maximum sample size (optional): --
- Metrics:** (This section is partially visible at the bottom)

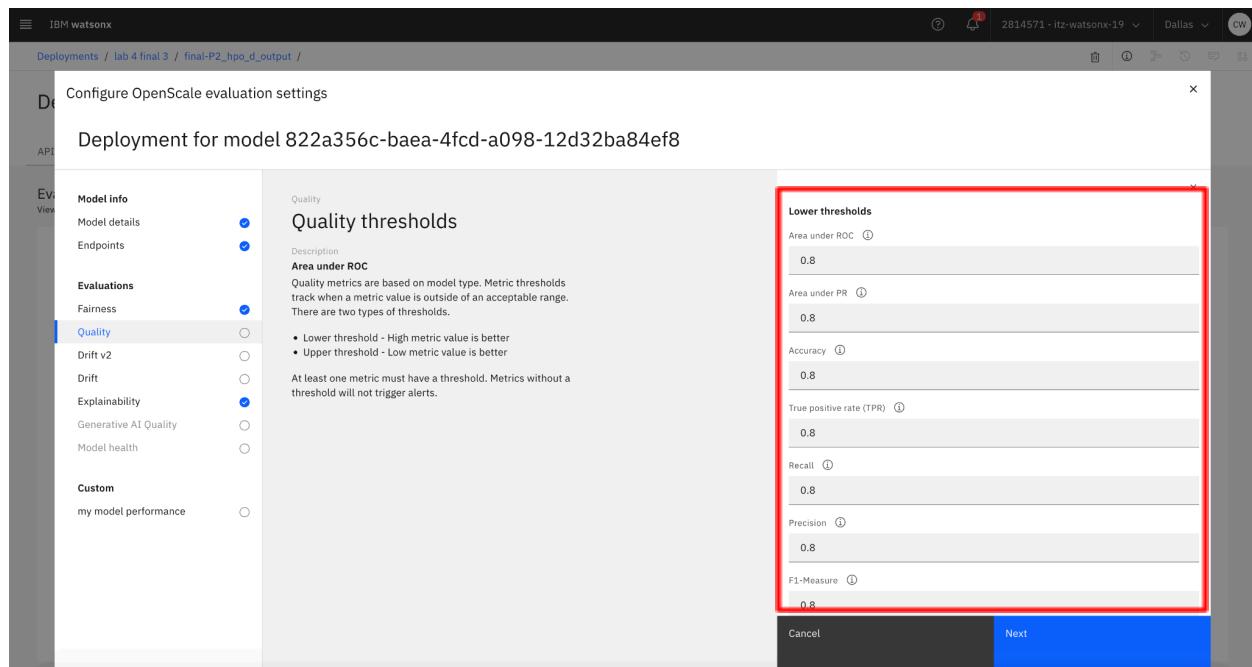
Part 4 : Configuring model quality monitor

Go to Quality under the Evaluations topic and you should be able to see this page. After that click on the edit icon of the configuration to start configuring the model quality monitor.



The screenshot shows the IBM WatsonX interface for configuring a model quality monitor. On the left, a sidebar lists various evaluation types: Model info, Endpoints, Evaluations (Fairness, Quality, Drift v2, Drift, Explainability, Generative AI Quality, Model health), and Custom (my model performance). The 'Quality' option is selected and highlighted with a red box and labeled '1.'. In the main content area, the 'Quality' tab is active, displaying a description of the Quality monitor's function. To the right, a 'Quality thresholds' section is shown, which includes a heading 'Lower thresholds' and a list of metrics with their current threshold values set to 0.8. An edit icon in the top right corner of this section is highlighted with a red box and labeled '2.'

Once you click on the configuration icon, you should enter this page. In this page you can adjust the thresholds for model quality as you prefer but for now, we will leave it as it is. Click next to continue.



This screenshot shows the 'Quality thresholds' configuration page. The left sidebar is identical to the previous one, with 'Quality' selected. The main content area is titled 'Quality thresholds' and contains a 'Lower thresholds' section. This section lists several metrics with their current threshold values set to 0.8: Area under ROC, Area under PR, Accuracy, True positive rate (TPR), Recall, Precision, F1-Measure, and Matthews correlation coefficient. A large red box highlights the entire configuration panel on the right side of the screen.

Set the minimum sample size to be whatever number which is smaller than the prepared dataset. Once done, click next to continue onto the next page.

The screenshot shows the IBM Watsonx interface with the following details:

- Deployment Path:** Deployments / lab 4 final 3 / final-P2_hpo_d_output /
- Section:** Configure OpenScale evaluation settings
- Evaluations:** Quality (selected), Fairness, Drift v2, Drift, Explainability, Generative AI Quality, Model health, Custom (my model performance)
- Quality Tab:** Sample size
Description: Ensure that your minimum sample size is large enough to accurately represent the variety of requests the deployment receives.
Minimum sample size: 100
Maximum sample size (optional): 1000
- Buttons:** Back, Save

Now you have finished your model quality monitor setup and should be able to see this page.

The screenshot shows the IBM Watsonx interface with the following details:

- Deployment Path:** Deployments / lab 4 final 3 / final-P2_hpo_d_output /
- Section:** Configure OpenScale evaluation settings
- Evaluations:** Quality (selected), Fairness, Drift v2, Drift, Explainability, Generative AI Quality, Model health, Custom (my model performance)
- Quality Tab:** Quality
Description: The Quality monitor evaluates how well your model predicts accurate outcomes. It identifies when model quality declines, so you can retrain your model appropriately.
Note: The Quality metric measures the model's ability to correctly predict outcomes that match labeled data (ground truth) provided by humans. The quality metrics evaluated are standard data science statistics based on model type. [Learn more.](#)
- Quality thresholds:** Lower thresholds
Area under ROC: 0.8
Area under PR: 0.8
Accuracy: 0.8
True positive rate (TPR): 0.8
Recall: 0.8
Precision: 0.8
F1-Measure: 0.8
Matthews correlation coefficient: 0.8
Label skew: 0.8

Part 5 : Configuring model Drift v2 monitor

Go to Drift v2 under the Evaluations topic and you should be able to see this page. After that click on the edit icon of the configuration to start configuring the model drift v2 monitor.

The screenshot shows the Watson OpenScale configuration interface. On the left, there's a sidebar with sections like Model info, Endpoints, Evaluations, Custom, and a navigation bar with Deployments, API, and View. Under Evaluations, the 'Drift v2' option is highlighted with a red box and labeled '1.'. The main content area is titled 'Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8'. It contains a 'Drift v2' section with a description: 'The Drift monitor checks if your deployments are up-to-date and behaving consistently. Model input/output data is analyzed in relation to the training/baseline data.' To the right, there are several configuration panels: 'Compute the drift archive' (with an edit icon labeled '2.'), 'Drift thresholds' (with an edit icon), 'Important features' (with an edit icon), and 'Most important features' (with an edit icon). The entire configuration window has a light gray background.

Select the compute option to be within Watson OpenScale. Once done, click next to continue.

This screenshot shows the 'Compute option' dialog box overlaid on the configuration interface. The dialog has a title 'Compute the drift archive' and a description: 'Watson OpenScale analyzes your training data to determine the data distributions of the features. You can perform this analysis in Watson OpenScale or you can do it yourself using a custom notebook.' Below this, there are two radio button options: 'Compute in Watson OpenScale' (selected) and 'Compute in a data science notebook'. The entire dialog is highlighted with a red box. At the bottom of the dialog are 'Cancel' and 'Next' buttons. The background of the configuration interface is visible, showing the same layout as the previous screenshot.

Drift upper thresholds track the degree of change of model input and output from its training data. Set it as you prefer but in this case we will leave it as it is. Click next to continue.

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Drift thresholds

Description

Drift metrics track the degree of change in model input and output in relation to the training or baseline data. Classification model metrics include:

Output drift

Measures the change in the model confidence distribution. [Read a sample scenario.](#)

Model quality drift

Measures the drop in accuracy by comparing the estimated runtime accuracy to the training accuracy. Ground truth labels are not required. [Read a sample scenario.](#)

Feature drift

Measures the change in value distribution for important features. [Read a sample scenario.](#)

Prediction drift

Measures the change in model prediction classes distribution. [Read a sample scenario.](#)

Upper thresholds

Output drift ⓘ	0.05
Model quality drift ⓘ	0.05
Feature drift ⓘ	0.05
Prediction drift ⓘ	0.05

Back Next

Select the important features which impact the model from the list.

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Important features ⓘ

Description

The system uses feature importance to determine the impact of feature drift on the model. For example, a small amount of drift in an important feature may have a bigger impact on the model than a moderate amount of drift in a less important feature.

Note: When SHAP is configured, the important features are automatically detected using the model's global explanation. As SHAP is not configured, you must indicate the important features manually.

Select from list

Select up to 100 important features.

Upload list

Generate a global explanation and upload the list of important features as a json file. Some snippets/examples to extract this information for popular ML frameworks has been provided on [this wiki](#).

Selection options

- Select from list
- Upload list
- Select important features

Features (12)	Type
ApplicantIncome	01
CoapplicantIncome	01
Credit_History	01
Dependents	00
Education	A
Gender	A
LoanAmount	01
LoanAmount_Term	01

Back Next

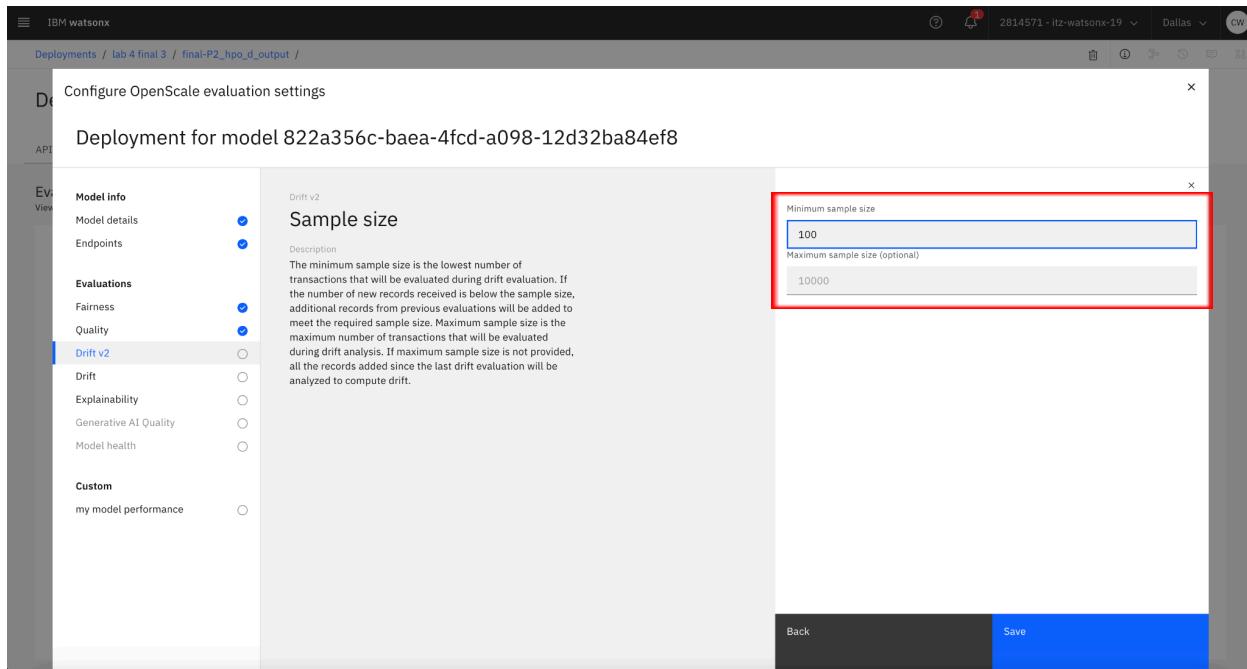
Once done click next to continue.

The screenshot shows the 'Configure OpenScale evaluation settings' dialog for 'Drift v2'. On the left, there's a sidebar with various evaluation options like Model info, Endpoints, Fairness, Quality, Drift v2 (selected), Drift, Explainability, Generative AI Quality, Model health, and Custom. The main area has sections for 'Important features' (with a note about drift impact) and 'Select from list' (checkboxes for ApplicantIncome, Credit_History, LoanAmount, and LoanAmount_Term). A 'Next' button at the bottom is highlighted with a red box.

Select the most important features which have high impact on the model from the list.

The screenshot shows the 'Configure OpenScale evaluation settings' dialog for 'Drift v2'. The sidebar and main area are similar to the previous screenshot, but the 'Select from list' section is replaced by a 'Select most important features' section. This section contains a table with four selected features: ApplicantIncome, Credit_History, LoanAmount, and LoanAmount_Term. A 'Next' button at the bottom is highlighted with a red box.

Set the minimum sample size to be whatever number which is smaller than the prepared dataset. Once done, click save to continue onto the next page.



Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Model info

Evaluations

- Fairness
- Quality
- Drift v2**
- Drift
- Explainability
- Generative AI Quality
- Model health

Custom

my model performance

Sample size

Description

The minimum sample size is the lowest number of transactions that will be evaluated during drift evaluation. If the number of new records received is below the sample size, additional records from previous evaluations will be added to meet the required sample size. Maximum sample size is the maximum number of transactions that will be evaluated during drift analysis. If maximum sample size is not provided, all the records added since the last drift evaluation will be analyzed to compute drift.

Minimum sample size

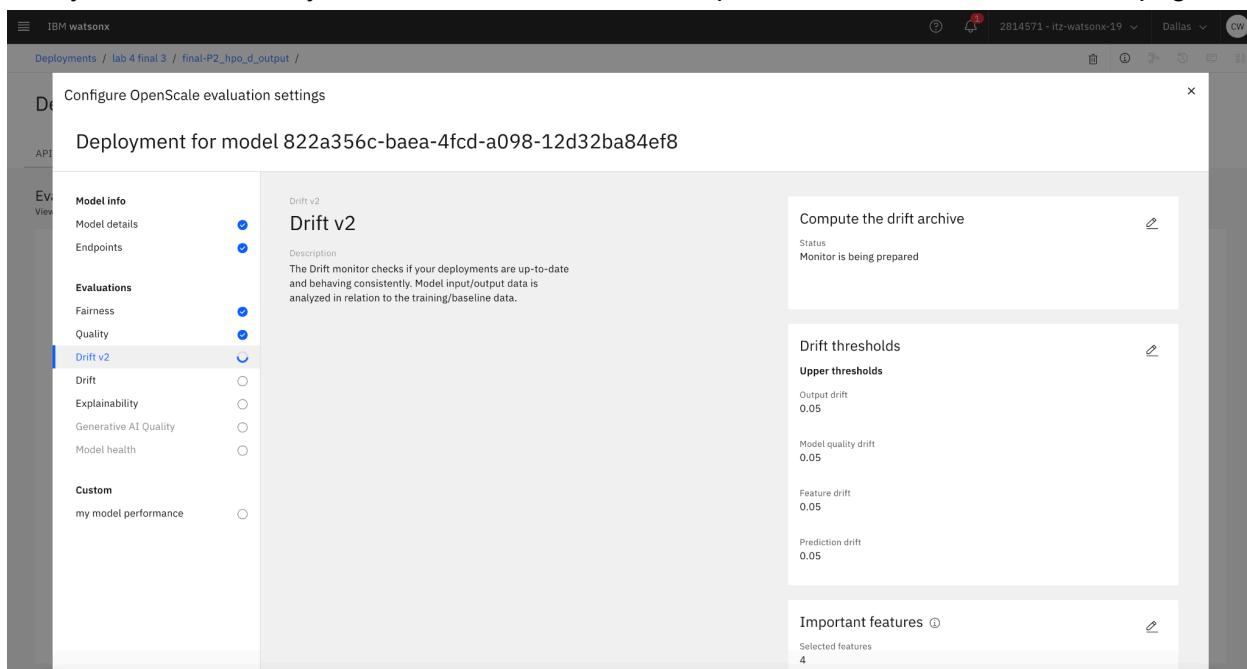
100

Maximum sample size (optional)

10000

Back Save

Now you have finished your model drift v2 monitor setup and should be able to see this page.



Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8

Model info

Evaluations

- Fairness
- Quality
- Drift v2**
- Drift
- Explainability
- Generative AI Quality
- Model health

Custom

my model performance

Drift v2

Description

The Drift monitor checks if your deployments are up-to-date and behaving consistently. Model input/output data is analyzed in relation to the training/baseline data.

Compute the drift archive

Status

Monitor is being prepared

Drift thresholds

Upper thresholds

Output drift
0.05

Model quality drift
0.05

Feature drift
0.05

Prediction drift
0.05

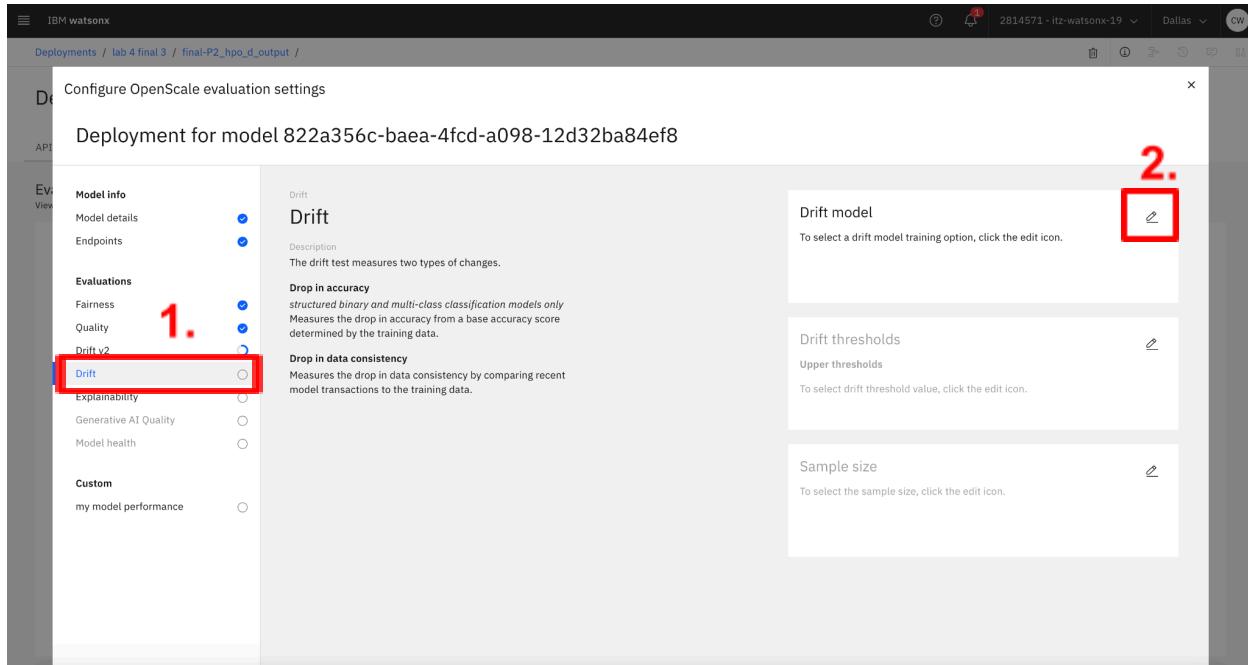
Important features

Selected features

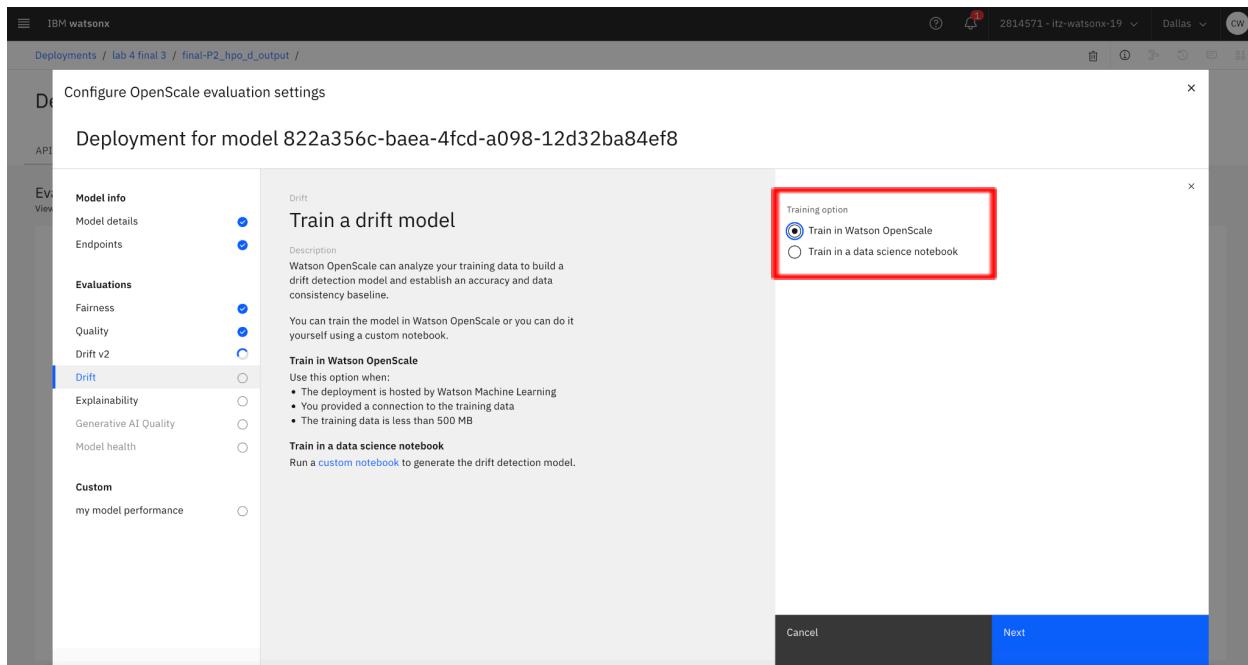
4

Part 6 : Configuring model Drift monitor

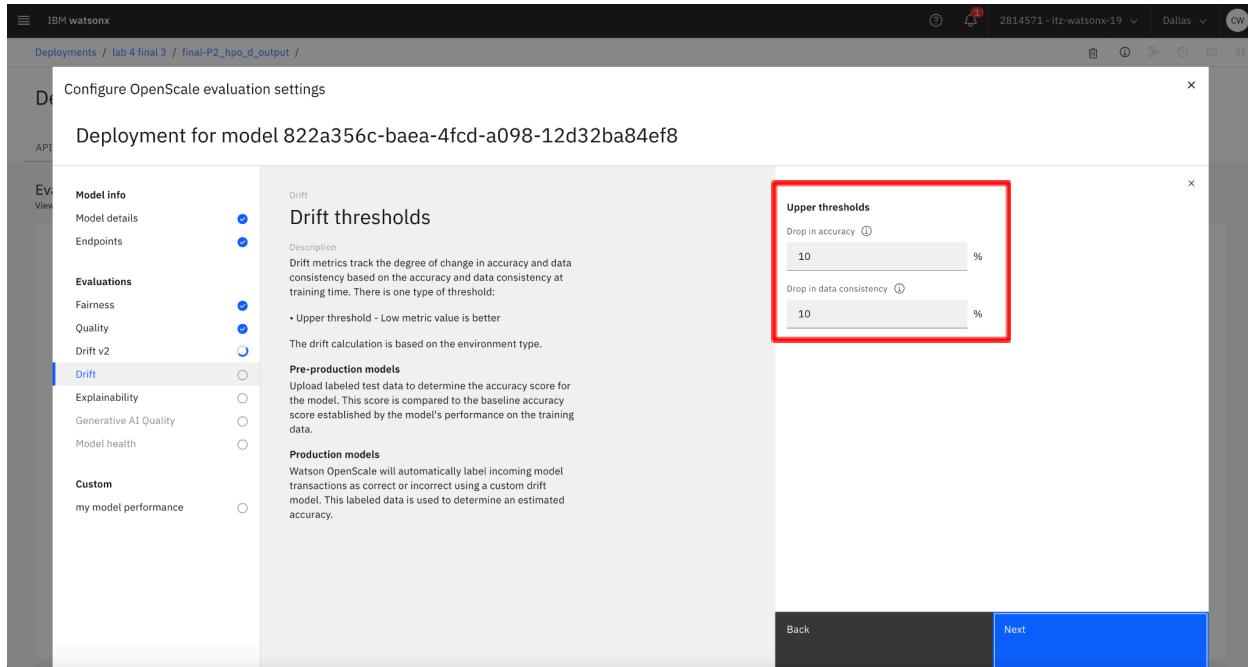
Go to Drift under the Evaluations topic and you should be able to see this page. After that click on the edit icon of the configuration to start configuring the model drift monitor.



Select the training option to be within Watson OpenScale. Once done, click next to continue.

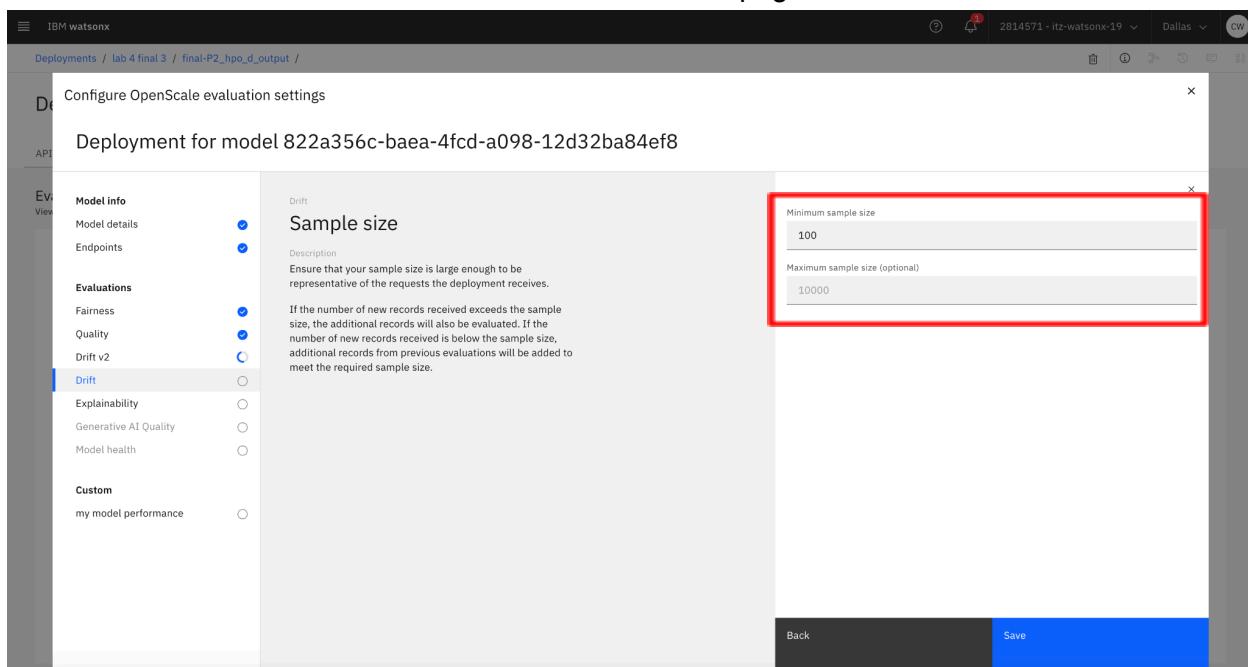


Setup drift upper thresholds which “drop in accuracy” means to measure the drop in accuracy from a base accuracy score determined by the training data and “drop in data consistency” means to measure the drop in data consistency by comparing recent model transactions to the training data.



The screenshot shows the IBM Watsonx interface for configuring OpenScale evaluation settings. The left sidebar lists various evaluations: Model info, Endpoints, Fairness, Quality, Drift v2, Drift (selected), Explainability, Generative AI Quality, Model health, and Custom. The main panel displays the 'Drift' section with the title 'Drift thresholds'. It includes a description of drift metrics and a note that the drift calculation is based on the environment type. Two input fields are shown under 'Upper thresholds': 'Drop in accuracy' with a value of 10% and 'Drop in data consistency' with a value of 10%. A red box highlights these two input fields. At the bottom are 'Back' and 'Next' buttons, with 'Next' being highlighted in blue.

Set the minimum sample size to be whatever number which is smaller than the prepared dataset. Once done, click save to continue onto the next page.



The screenshot shows the IBM Watsonx interface for configuring OpenScale evaluation settings. The left sidebar lists various evaluations: Model info, Endpoints, Fairness, Quality, Drift v2, Drift (selected), Explainability, Generative AI Quality, Model health, and Custom. The main panel displays the 'Drift' section with the title 'Sample size'. It includes a description stating that the sample size should be large enough to be representative of the requests the deployment receives. Two input fields are shown: 'Minimum sample size' with a value of 100 and 'Maximum sample size (optional)' with a value of 10000. A red box highlights these two input fields. At the bottom are 'Back' and 'Save' buttons, with 'Save' being highlighted in blue.

Now you have finished your model drift monitor setup and should be able to see this page.

The screenshot shows the IBM WatsonX interface with the following details:

- Deployment Path:** Deployments / lab 4 final 3 / final-P2_hpo_d_output /
- Section:** Configure OpenScale evaluation settings
- Deployment ID:** Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8
- Model Info:** Model details, Endpoints
- Evaluations:** Fairness, Quality, Drift v2, Drift (selected), Explainability, Generative AI Quality, Model health
- Custom:** my model performance
- Drift Test Configuration:**
 - Drift:** Description: "The drift test measures two types of changes. structured binary and multi-class classification models only".
 - Drop in accuracy:** Measures the drop in accuracy from a base accuracy score determined by the training data.
 - Drop in data consistency:** Measures the drop in data consistency by comparing recent model transactions to the training data.
 - Drift thresholds:** Upper thresholds: Drop in accuracy 10%, Drop in data consistency 10%
 - Sample size:** Minimum sample size 100, Maximum sample size (optional) --

Part 7 : Start evaluation

Once you are done configuring the evaluation monitor you should be on this page. Click on evaluate now under actions dropdown.

The screenshot shows the IBM WatsonX interface with the following details:

- Deployment Path:** Deployments / lab 4 final 3 / final-P2_hpo_d_output /
- Deployment Status:** Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8 (Deployed, Online)
- Actions Dropdown:** Actions
- Evaluations Tab:** Selected
- Transactions Tab:** Available
- AI Factsheet Tab:** Available
- Last Evaluation:** Last evaluation: --
- Test Data Set:** Test data set --
- Number of Explanations:** Number of explanations 0
- Test Results:** 0 Tests run, 0 Tests passed, 0 Tests failed
- Note:** Awaiting test data to evaluate. Select the Actions button to evaluate now.

Once you click evaluate now, click import from CSV file to import the test data using .csv file.

The screenshot shows the IBM WatsonX interface with the 'Evaluate' dialog open. The 'Import test data' section is highlighted with a red box. The 'Choose an option' dropdown is expanded, showing 'from database or cloud storage' and 'from CSV file'. The 'from CSV file' option is selected.

upload the test data into the red area.

The screenshot shows the IBM WatsonX interface with the 'Evaluate' dialog open. The 'Import test data' section is highlighted with a red box. The 'from CSV file' input field is selected. Below it is a large red box containing the 'Drag test data file here or browse' button and the 'Test data includes model output' checkbox.

Once you upload the data, your file name should appear on the test data set. Click upload and evaluate to evaluate.

Part 8 : Evaluation result

Once done with evaluate, you should be able to get the evaluation result as shown below. You can click on the arrow button inside each metrics monitor to find further information for that monitor type.

IBM watsonx

Deployments / lab 4 final 3 / final-P2_hpo_d_output /

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8 Deployed Online

API reference Test Evaluations Transactions AI Factsheet

Associated group: Female
84 records evaluated

Evaluation Metric	Score	Violation
Area under PR	0.95	none
Recall	0.98	none
Brier score	0.21	none

566 records evaluated

Drift

Alerts triggered: 1

Metric	Score	Violation
Drop in data consistency	100.00%	90.00%
Drop in accuracy	1.63%	none
Evaluated accuracy	96.29%	--
Base accuracy	97.92%	--

566 records evaluated

Here is the example of further information for each evaluation metrics.

IBM watsonx

Deployments / lab 4 final 3 / final-P2_hpo_d_output /

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8 Deployed Online

API reference Test Evaluations Transactions AI Factsheet

Fairness Quality Drift Back to summary

Fairness

Monitored attribute: Gender | Fairness metric: Disparate impact | Data set: Balanced | Date and time: 10/7/2024 1:11 PM | View payload transactions

Disparate impact: 100% | Favorable outcomes: Y

How the disparate impact score was determined (balanced data set):
The monitored group Female received favorable outcomes 67.857% of the time. The perfect equality is 67.857%. The fairness score for Gender is 100.0% ($67.857/67.857$). [View calculation](#).

% favorable outcomes: 20, 40, 60, 80

View percentage (radio button selected) View count

IBM Watsonx

Deployments / lab 4 final 3 / final-P2_hpo_d_output /

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8 Deployed Online

Evaluations Test Fairness Quality Drift AI Factsheet Back to summary

Quality

Quality monitor calculates a list of metrics to compare an automatically produced classification against a reference (human-produced) classification.

Mon, Oct 7, 2024, 1:10 PM +07

Area under ROC	Area under PR	Accuracy	True positive rate (TPR)	False positive rate (FPR)	Recall	Precision	F1-Measure	Logarithmic loss	Brier score	Matthews correlation coefficient	Label skew
0.94	0.95	0.96	0.98	0.11	0.98	0.96	0.97	0.32	0.21	0.9	-0.88 <small>0.38 lower threshold violation</small>

Confusion matrix

Prediction		Predicted		Percent correct
Actual		N	Y	
N		151	18	89.3%
Y		6	391	98.5%

IBM Watsonx

Deployments / lab 4 final 3 / final-P2_hpo_d_output /

Deployment for model 822a356c-baea-4fcd-a098-12d32ba84ef8 Deployed Online

Evaluations Test Fairness Quality Drift AI Factsheet Back to summary

Drift

View the transactions with incorrect predictions, inconsistent data, or both.

10/7/2024 1:10 PM

Incorrect predictions (21) Inconsistent data (566) Both (21)

Transactions with incorrect predictions

Drop in accuracy ⓘ 1.6%

Number of transactions	Number of transactions	Number of transactions			
13	3	2			
Grouped by	Dominant values	Grouped by	Dominant values	Grouped by	Dominant values
Credit_History	1	Dependents	0	Dependents	3
Loan_Amount_Term	378.46	CoapplicantIncome	0	Loan_Amount_Term	360
LoanAmount_Term	378.46	Loan_Amount_Term	360	LoanAmount_Term	360
Self_Employed	No	Education	Graduate	Education	Graduate
Gender	Male	Gender	Male	Gender	Male

For model explainability, click on transactions tab in the red area

The screenshot shows the 'Transactions' tab highlighted with a red box. The page displays the last evaluation date (Mon, Oct 7, 2024, 1:10 PM +07) and a summary of tests run (4 total, 1 passed, 3 failed). Below this, there are three sections: Fairness, Quality, and Drift v2, each with specific metrics and alert counts.

Metric	Score	Violation
Disparate impact	100.00%	none

Metric	Score	Violation
True positive rate (TPR)	0.98	none
Area under ROC	0.94	none

Drift v2 section:

- Alerts triggered: 1
- Drift computation failed. Argument 'value' has incorrect type (expected int, got numpy.int64)

Once you click on transactions you should see this page. Then, click on the bar chart to view its detail

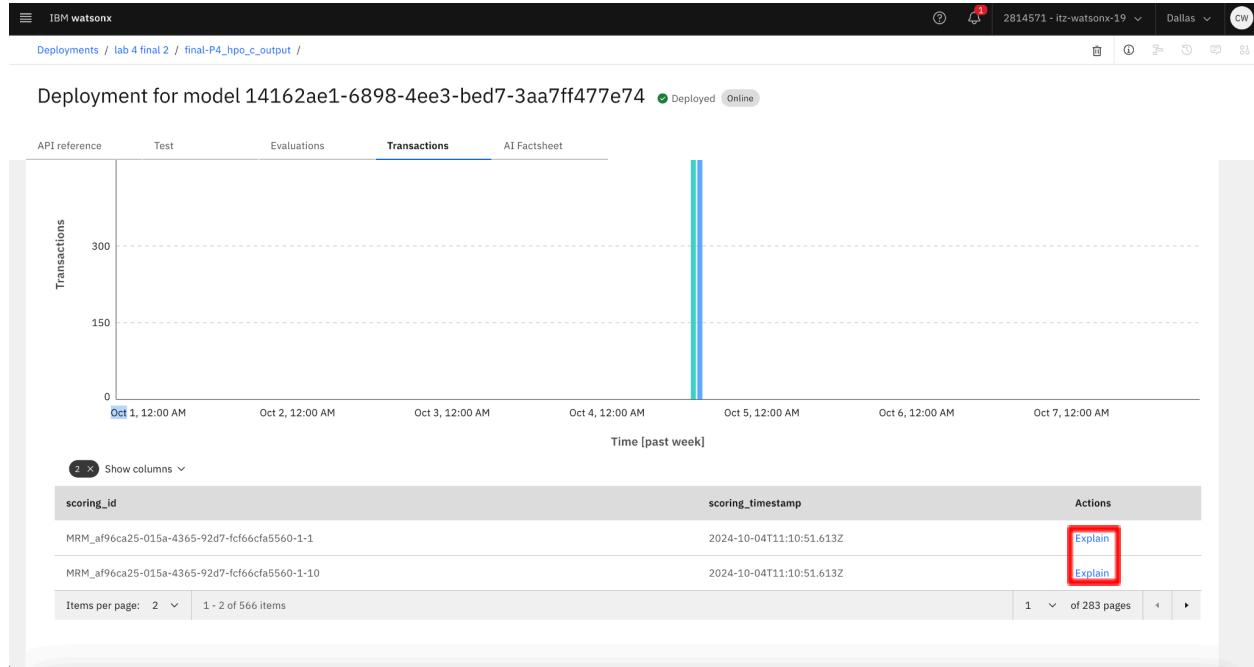
The screenshot shows the 'Transactions' tab highlighted with a red box. The page displays an evaluation history with a search bar and filters for time frame (Hourly, Past week) and date range. A bar chart shows the number of transactions over time, with a prominent peak on October 5th, which is highlighted with a red box.

Time frame	Date range
Hourly	Past week

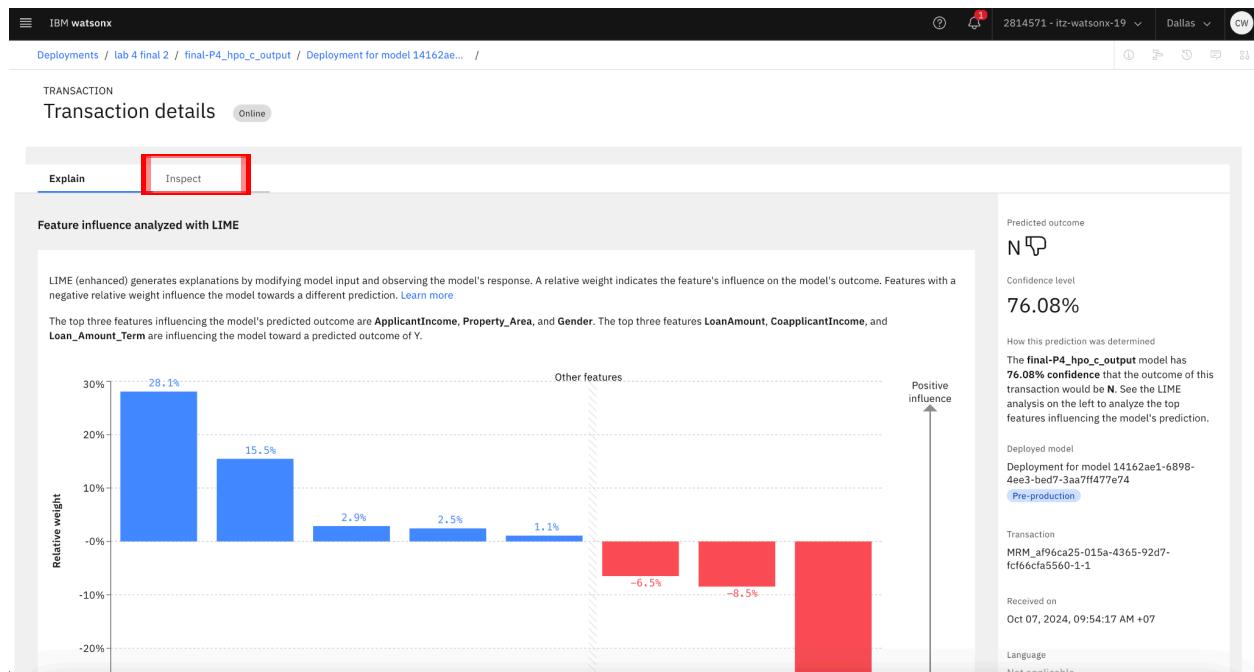
Transactions chart:

- Y-axis: Transactions (0 to 567)
- X-axis: Date (Oct 1 to Oct 7, 2024)
- A vertical red box highlights the peak on Oct 5, 2024, reaching approximately 567 transactions.

After you click on the bar chart the list underneath should appear. Click explain to view model explainability.



Now you are able to observe the model explainability through feature importance. If you want to see how the change of the feature affects the model you can use the 'inspect' tab to observe.



Once you click inspect, you should be entering this page. In this page you can config the feature values to predict a different outcome by adjusting the feature value in the red area.

For the model to have predicted a different outcome for this transaction, the value of all listed features would need to change to the indicated minimum value. Note that changing a feature value by more than the minimum value may affect the minimum change of other features for the model to predict a different outcome. Higher feature importance numbers indicate a greater likelihood of changing the prediction.

Analyze controllable features only ⓘ

Feature	Original value	New value	Value for a different outcome ⓘ	Importance
Gender	Male	Male		
Married	Yes	Yes		
Dependents	1	1 ⓘ		
Education	Graduate	Graduate		
Self_Employed	No	No		
ApplicantIncome	4583	4583 ⓘ		
CoapplicantIncome	1508	1508 ⓘ		
LoanAmount	128	128 ⓘ		
Predicted outcome N	Confidence 79.34%	Predicted outcome N	Confidence 79.34%	

Want to know what feature values will result in a different outcome? Analyze the model to get a suggestion. This analysis will take a few minutes.

Run analysis

Once you are done configuring the value, click “Score new values” and then click “Run analysis”.

For the model to have predicted a different outcome for this transaction, the value of all listed features would need to change to the indicated minimum value. Note that changing a feature value by more than the minimum value may affect the minimum change of other features for the model to predict a different outcome. Higher feature importance numbers indicate a greater likelihood of changing the prediction.

Analyze controllable features only ⓘ

Feature	Original value	New value	Value for a different outcome ⓘ	Importance
Gender	Male	Male		
Married	Yes	Yes		
Dependents	1	1 ⓘ		
Education	Graduate	Graduate		
Self_Employed	No	Yes ⓘ		
ApplicantIncome	4583	4583 ⓘ		
CoapplicantIncome	1508	1508 ⓘ		
LoanAmount	128	128 ⓘ		
Predicted outcome N	Confidence 79.34%	Score new values	Reset	

Want to know what feature values will result in a different outcome? Analyze the model to get a suggestion. This analysis will take a few minutes.

Run analysis

After you click run analysis, you should be able to see the outcome in the red area.

The screenshot shows the 'Inspect' tab of the IBM Watson AI Model Explainability interface. It displays a table titled 'Reaching a different predicted outcome' with columns for Feature, Original value, New value, and Importance. A red box highlights the 'Value for a different outcome' table, which shows the predicted outcome changing from N (Confidence 79.34%) to Y (Confidence 80.24%).

Feature	Original value	New value	Importance
ApplicantIncome	4583	4583	1.00
Gender	Male	Male	0.00
Married	Yes	Yes	0.00
Dependents	1	1	0.00
Education	Graduate	Graduate	0.00
Self_Employed	No	No	0.00
CoplicantIncome	1508	1508	0.00
LoanAmount	128	128	0.00
Predicted outcome N	Confidence 79.34%	Predicted outcome Y	Confidence 80.24%

Furthermore, you can reconfigure the evaluation metrics monitor by clicking on the configure button in the red area.

The screenshot shows the 'Evaluations' tab of the IBM Watson AI Model Monitoring interface. It displays various evaluation metrics and their status. A red box highlights the 'Fairness' section, which shows 'No alerts' for gender.

Metric	Score	Violation
Disparate impact	98.15%	none

Metric	Score	Violation
True positive rate (TPR)	0.84	none
Area under ROC	0.70	0.10
Precision	0.81	none
Matthews correlation coefficient	0.42	0.38

Which once you clicked on the reconfigure button in the red area. Now you can reconfigure OpenScale evaluation settings.

The screenshot shows the IBM Watson AI interface with the following details:

- Deployment Path:** Deployments / for lab 4 v2 / with-P2_hpo_d_output /
- Section:** Configure OpenScale evaluation settings
- Context:** with loan id y label 3
- Left Sidebar (Model Info):** Model details, Endpoints, Explainability (General settings, SHAP, LIME (enhanced)), Evaluations (Fairness, Quality, Drift v2, Drift, Generative AI Quality, Model health), Custom (my model performance).
- Selected Evaluation Type:** Fairness
- Fairness Configuration:**
 - Description:** Fairness
 - Description:** The Fairness monitor checks your deployments for biases. It tracks when the model shows a tendency to provide a favorable (preferable) outcome more often for one group over another. You will specify which values represent favorable outcomes, select the features to monitor for bias (for example, Age or Sex), and specify the groups to monitor for each selected feature.
 - Configuration:** Configuration file (empty)
 - Favorable outcomes:** Favorable outcomes: Y, Unfavorable outcomes: N
 - Sample size:** Minimum sample size: 100, Maximum sample size (optional): empty
 - Metrics:** Metrics (empty)