IBM Watson OpenScale

Veritas_AutoAI_Best_Model Evaluation Report

July 24, 2023

Overview

Deployed model:

Veritas_Deployment_Preprod_Tradeoff

Total red breaches

3

Report Details

Evaluated by:	admin (admin)
Report generated by:	admin (admin)
Report generated on:	July 24, 2023 01:47:26 UTC

Model Details

Deployment ID:	cae64b39-0b7e-4c2b-9e8f-52b2edd85d14
Model name:	Veritas_AutoAI_Best_Model
Model ID:	3d81ad44-c0be-46a5-b5d1-df0775b123fe
Data type:	Numeric
Algorithm type:	Binary classification

Training data details

Storage location:	Cloud Object Storage
Url:	https://s3.au-syd.cloud-object-storage.appdomain.cloud
Resource instance id:	crn:v1:bluemix:public:cloud-object-storage:global:a/d190f3df28cc47629db4057098f6407f:8279ce79-41a8-4a89-a025-f97996871a14::
Filename:	Training_Data_Credit_Risk_v3.csv
Bucket:	uob-rfp-mlops-initiation
Label column:	y_train
Deployment prediction:	prediction
Training features:	LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6

Metrics

Metric details

Summary

Deployed model	Model ID	Test data set
Veritas_Deployment_Preprod_Tradeoff	3d81ad44-c0be-46a5-b5d1-df0775b123fe	Test_Data_Credit_Risk_v3.csv

Metric

Drift

Alerts

1

RED BREACH

Summary

Number of metrics:	4
Alerts:	1

Statistics

Drop in data consistency

Status:	RED BREACH
Score:	100%
Threshold:	10%

Drop in accuracy

Status:	GREEN
Score:	0%
Threshold:	10%

Estimated accuracy

Status:	GREEN
Score:	83%
Threshold:	- -

Base accuracy

Status:	GREEN
Score:	82%
Threshold:	- -

Properties

Minimum sample size:	100
Evaluated records count:	750

Metrics

Metric

Fairness

Alerts
1
RED BREACH

Summary

Number of metrics:	1
Monitored features:	1
Favorable outcome:	No Risk
Unfavorable outcome:	Risk
Alerts:	1

SEX

Monitored groups:	1-1
Reference groups:	2-2
Alerts:	1

Disparate impact

Status:	RED BREACH
Score:	123%
Threshold:	120%

Properties

Minimum sample size:	100
Evaluated records count:	100

Metric

Quality

Alerts
1
RED BREACH

Summary

Number of metrics:	9
Alerts:	1

Statistics

True positive rate (TPR)

Status:	GREEN
Score:	0.94
Threshold:	0.80

Area under ROC

Status:	RED BREACH
Score:	0.65
Threshold:	0.80

Precision

Status:	GREEN
Score:	0.86
Threshold:	0.80

Metrics

F1-Measure

Status:	GREEN
Score:	0.90
Threshold:	0.80

Accuracy

Status:	GREEN
Score:	0.83
Threshold:	0.80

Logarithmic loss

Status:	GREEN
Score:	0.46
Threshold:	0.80

False positive rate (FPR)

Status:	GREEN
Score:	0.64
Threshold:	0.80

Area under PR

Status:	GREEN
Score:	0.86
Threshold:	0.80

Recall

Status:	GREEN
Score:	0.94
Threshold:	0.80

Properties

Minimum sample size:	100
Evaluated records count:	1500

Metric

Veritas toolkit metrics perf fair tradeoff sex

Alerts
0
GREEN

Summary

Number of metrics:	9
Alerts:	0

Statistics

Max perf neutral fair priv th

Status:	GREEN
Score:	0.70

Metrics

Threshold:	- -

Max perf neutral fair unpriv th

Status:	GREEN
Score:	0.42
Threshold:	- -

Max perf single th priv th

Status:	GREEN
Score:	0.30
Threshold:	- -

Max perf separated th best bal acc

Status:	GREEN
Score:	0.34
Threshold:	- -

Max perf neutral fair best bal acc

Status:	GREEN
Score:	0.32
Threshold:	- -

Max perf separated th unpriv th

Status:	GREEN
Score:	0.70
Threshold:	- -

Max perf separated th priv th

Status:	GREEN
Score:	0.70
Threshold:	- -

Max perf single th best bal acc

Status:	GREEN
Score:	0.34
Threshold:	- -

Max perf single th unpriv th

Status:	GREEN
Score:	0.30
Threshold:	- -

Metric

Veritas toolkit metrics performance

Alerts

0

GREEN

Metrics

Summary

Number of metrics:	11
Alerts:	0

Statistics

F1 score

Status:	GREEN
Score:	0.91
Threshold:	- -

Tnr

Status:	GREEN
Score:	0.50
Threshold:	- -

Roc auc

Status:	GREEN
Score:	0.22
Threshold:	- -

Precision

Status:	GREEN
Score:	0.86
Threshold:	- -

Balanced acc

Status:	GREEN
Score:	0.73
Threshold:	- -

Fnr

Status:	GREEN
Score:	0.04
Threshold:	- -

Selection rate

Status:	GREEN
Score:	0.85
Threshold:	- -

Accuracy

Status:	GREEN
Score:	0.85
Threshold:	- -

Log loss

Status:	GREEN

Score: 1.63
Threshold: - -

Npv
Status: GREEN
Score: 0.80
Threshold: - -

Recall
Status: GREEN
Score: 0.96
Threshold: - -

Metric

Veritas toolkit metrics fairness sex

Alerts
0
GREEN

Summary
Number of metrics: 17
Alerts: 0

Statistics

Log loss parity
Status: GREEN
Score: -0.09
Threshold: -0.43

Auc parity
Status: GREEN
Score: -0.11
Threshold: -0.43

Fpr parity
Status: GREEN
Score: -0.17
Threshold: -0.43

Demographic parity
Status: GREEN
Score: -0.12
Threshold: -0.43

Calibration by group
Status: GREEN
Score: 0.18
Threshold: -0.43

Metrics

Equal opportunity

Status:	GREEN
Score:	-0.10
Threshold:	-0.43

Mi independence

Status:	GREEN
Score:	0.02
Threshold:	-0.43

Equal odds

Status:	GREEN
Score:	-0.14
Threshold:	-0.43

Fdr parity

Status:	GREEN
Score:	-0.02
Threshold:	-0.43

Ppv parity

Status:	GREEN
Score:	0.02
Threshold:	-0.43

Mi separation

Status:	GREEN
Score:	0.05
Threshold:	-0.43

For parity

Status:	GREEN
Score:	0.33
Threshold:	-0.43

Tnr parity

Status:	GREEN
Score:	0.17
Threshold:	-0.43

Fnr parity

Status:	GREEN
Score:	0.10
Threshold:	-0.43

Neg equal odds

Status:	GREEN

Metrics

Score:	0.14
Threshold:	-0.43

Npv parity

Status:	GREEN
Score:	-0.33
Threshold:	-0.43

Mi sufficiency

Status:	GREEN
Score:	0.03
Threshold:	-0.43

Test summary

Tests passed

3

Tests failed

3

Number of evaluated records

750

Appendix

Quality Measures	Area under ROC Area under PR Accuracy True positive rate (TPR) False positive rate (FPR) Recall Precision F1-measure Logarithmic loss
Fairness measures	Fairness
Drift measures	Drop in accuracy Drop in data consistency Estimated accuracy Base accuracy

Appendix

Quality measures

Area under ROC

The Area under ROC is plotted parametrically as the True positive rate versus the False positive rate with respect to a threshold T.

Area under PR

Area under Precision Recall gives the total for both Precision + Recall. Precision (P) is defined as the number of true positives (Tp) over the number of true positives plus the number of false positives (Fp)

Formula

$$\textbf{Precision} = \frac{\text{number of true positives}}{(\text{number of true positives} + \text{number of false positives})}$$

Recall (R) is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

$$\textbf{Recall} = \frac{\text{number of true positives}}{(\text{number of true positives} + \text{number of false negatives})}$$

Appendix

Quality measures

Accuracy

Base accuracy is calculated from the training data. It is the percentage of predictions that the model got correct when tested against the training data.

True positive rate (TPR)

The True positive rate is calculated by the following formula:

Formula

$$\textbf{TPR} = \frac{\text{number of true positives}}{(\text{number of true positives} + \text{number of false negatives})}$$

False positive rate (FPR)

The false positive rate is calculated as the total number of false positives divided by the number of false positives and the number of true negatives.

$$\textbf{FPR} = \frac{\text{number of false positives}}{(\text{number of false positives} + \text{number of true negatives})}$$

Appendix

Quality measures

Recall

Recall (R) is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

Formula

$$\textbf{Recall} = \frac{\text{number of true positives}}{(\text{number of true positives} + \text{number of false negatives})}$$

Precision

Precision (P) is defined as the number of true positives (Tp) over the number of true positives plus the number of false positives (Fp).

Formula

$$\textbf{Precision} = \frac{\text{number of true positives}}{(\text{number of true positives} + \text{number of false positives})}$$

Appendix

Quality measures

F1-Measure

The F1-Measure is the weighted harmonic average, or mean, of precision and recall.

Formula

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

Logarithmic loss

For a binary model, Logarithmic loss is calculated by using the following formula:

Formula

$$-(y \log (p) + (1-y) \log (1-p))$$

where p = true label and y = predicted probability

For a multi-class model, Logarithmic loss is calculated by using the following formula:

$$-\sum_{c=1}^{M} Yo,c \log(Po,c)$$

where M > 2, p = true label, and y = predicted probability

Appendix

Fairness measures

Fairness

The fairness metric used in Watson OpenScale is disparate impact, which is a measure of how the rate at which an unprivileged group receives a certain outcome or result compares with the rate at which a privileged group receives that same outcome or result.

Formula

$$\textbf{Disparate impact} = \frac{(\text{num\_positives(privileged=False)}/\text{num\_instance(privileged=False)})}{(\text{num\_positives(privileged=True)}/\text{num\_instance(privileged=True)})}$$

Appendix

Drift measures

Drop in accuracy

Watson OpenScale analyzes each transaction to estimate if the model prediction is accurate. If the model prediction is inaccurate, the transaction is marked as drifted. The Estimated accuracy is then calculated as the fraction of non-drifted transactions to the total number of transactions analyzed. The Base accuracy is the accuracy of the model on the test data. Watson OpenScale calculates the extent of the drift in accuracy as the difference between Base accuracy and Estimated accuracy. Further, Watson OpenScale analyzes all the drifted transactions; and then, groups transactions based on the similarity of each feature's contribution to the drift in accuracy. In each cluster, Watson OpenScale also estimates the important features that played a major role in the drift in accuracy and classifies their feature impact as large, some, and small.

Drop in data consistency

Watson OpenScale analyzes each transaction for data inconsistency, by comparing the transaction content with the training data patterns. If a transaction violates one or more of the training data patterns, the transaction is marked as drifted. Watson OpenScale then estimates the magnitude of data inconsistency as the fraction of drifted transactions to the total number of transactions analyzed. Further, Watson OpenScale analyzes all the drifted transactions; and then, groups transactions that violate similar training data patterns into different clusters. In each cluster, Watson OpenScale also estimates the important features that played a major role in the data inconsistency and classifies their feature impact as large, some, and small.

Appendix

Drift measures

Estimated accuracy

Estimated accuracy is the accuracy score at runtime estimated by Watson OpenScale. As part of drift monitor configuration, Watson OpenScale trains a drift detection model that identifies when the original model is likely to provide an incorrect response to a transaction. As the original model receives a new transaction, the transaction is evaluated by the drift model. If the drift model believes that the model likely provided an incorrect response, the transaction is identified as a drifted transaction. The Estimated accuracy is then calculated as the fraction of non-drifted transactions to the total number of transactions analyzed.

Formula

$$\textbf{Estimated Accuracy} = \frac{\text{Number of non-drifted transactions*}}{\text{Total number of transactions}}$$

*determined by the Watson OpenScale drift model

Base Accuracy

This is calculated from the training data. It is the percentage of predictions that the model got correct when tested against the training data.