IBM Watson OpenScale

Model Risk Evaluation - P4 GradientBoostingClassifierEstimator - Test Evaluation Report

June 06, 2020

Overview

Total red breaches

2

Deployed model:

P4 GradientBoostingClassifierEstimator - Test - Deployment

Report Details

Evaluated by:	admin (admin)
Report generated by:	admin (admin)
Report generated on:	June 06, 2020 15:03:28 UTC

Model Details

Deployment ID:	f5c737f2-28cc-484c-9fb5-c5ea5dcc5eea
Model name:	Model Risk Evaluation - P4 GradientBoostingClassifierEstimator - Test
Model ID:	bf2ebb06-136c-4132-8cc8-a917679188cb
Data type:	Numeric/Categorical
Algorithm type:	Binary classification
Number of explanations:	2

Training data details

Storage location:	db2
Database:	BLUDB
IP address:	dashdb-txn-sbox-yp-dal09-08.services.dal.bluemix.net
Port:	50000
Username:	nmx87075
Table:	GERMAN_CREDIT_RISK_DATA
Label column:	Risk
Deployment prediction:	prediction
Training features:	Age, CheckingStatus, CreditHistory, CurrentResidenceDuration, Dependents, EmploymentDuration, ExistingCreditsCount, ExistingSavings, ForeignWorker, Housing, InstallmentPercent, InstallmentPlans, Job, LoanAmount, LoanDuration, LoanPurpose, OthersOnLoan, OwnsProperty, Sex, Telephone

Metrics

Metric details

Summary

Deployed model	Model ID	Test data set
P4 GradientBoostingClassifierEstimator - Test - Deployment	bf2ebb06-136c-4132-8cc8-a917679188cb	german_credit_data_biased_test_2.csv

Metric	Score
Drift	8% RED BREACH

Summary

Base accuracy:	81%
Drift threshold:	5%
Drop in accuracy:	8%
Drop in data consistency:	6%
Estimated accuracy:	73%
Threshold violation:	3%
Minimum sample size:	100

Metric	Score
Fairness	78% RED BREACH

Summary

Fairness score:	78%
Fairness threshold:	98%
Favorable outcome:	No Risk
Threshold violation:	20%
Unfavorable outcome:	Risk
Minimum sample size:	100

Sex

Fairness score:	82%
Fairness threshold:	98%
Monitored group:	female
Reference group:	male

Age

Fairness score:	78%

Metrics

Fairness threshold:	98%
Monitored group:	44-67
Reference group:	19-43

Metric

Quality

Summary

Quality score:	0.78
Quality threshold:	0.7
Threshold violation:	N/A
Minimum sample size:	100

Statistics

True positive rate (TPR):	0.64
Area under ROC:	0.78
Precision:	0.81
F1-Measure:	0.72
Accuracy:	0.83
Logarithmic loss:	0.38
False positive rate (FPR):	0.08
Area under PR:	0.73
Recall:	0.64

Test summary

Tests passed

1

Tests failed

2

Number of evaluated records

200

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
Performance measures	Throughput

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

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

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

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

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

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

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.

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

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

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

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

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

Appendix

Quality measures

F1-Measure

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

Formula

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

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:

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

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

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

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.

Appendix

Performance measures

Throughput

Throughput measures the average scoring requests per minute.

Formula

$$\frac{\text{Number of transactions received in 1 hour}}{\text{60 minutes}}$$