SUMMARY REPORT

BINARY CLASSIFICATION MODEL FOR CREDIT RISK
ABC COMPANY PTE LTD
06 JUN 2023

AI VERIFY

AI GOVERNANCE TESTING FRAMEWORK AND TOOLKIT

INTRODUCTION

AIM OF AI VERIFY - AI GOVERNANCE TESTING FRAMEWORK AND TOOLKIT

AI Verify aims to help organisations validate the performance of their AI systems against a set of internationally recognised principles through standardised tests:

- the performance of their AI systems; and
- documentary evidence that their AI systems have been developed and deployed with processes designed to achieve the desired outcomes of these principles.

Companies can use the output from these tests to demonstrate their implementation of responsible AI and build trust with their stakeholders. Companies can also use the test results to identify potential gaps and take appropriate actions to address them, where applicable.

Please note that only reports generated by AI Verify Toolkit in accordance with the AI Verify Testing Framework, and without modification are AI Verify Reports.

USE CASE AND MODEL TESTED

Model Tested:	Binary Classification Model for Credit Risk
Purpose of Model:	This model is used to test if the applicant will default the loan

---

SCOPE OF CHECKS

This Summary Report provides an overview of how the AI model performs vis-à-vis the AI Verify testing framework. The framework covers 11 AI ethics principles, grouped into 5 focus areas.

These principles are assessed by a combination of technical tests and/or process checks.

TRANSPARENCY ON THE USE OF AI AND AI SYSTEMS Ensuring that individuals are aware and can make informed decisions			
TRANSPARENCY | Appropriate info is provided to individuals impacted by AI system			
UNDERSTANDING HOW AI MODELS REACH DECISION Ensuring AI operation/results are explainable, accurate and consistent	**SAFETY & RESILIENCE OF AI SYSTEM** Ensuring AI system is reliable and will not cause harm	**FAIRNESS / NO UNINTENDED DISCRIMINATION** Ensuring that use of AI does not unintentionally discriminate	**MANAGEMENT AND OVERSIGHT OF AI SYSTEM** Ensuring human accountability and control
EXPLAINABILITY⁺ Understand and interpret what the AI system is doing **REPEATABILITY / REPRODUCIBILITY** AI results are consistent: Be able to replicate an AI system's results by owner / 3rd-party.	**SAFETY** AI system safe: Conduct impact / risk assessment; Known risks have been identified/mitigated **SECURITY** AI system is protected from unauthorised access, disclosure, modification, destruction, or disruption **ROBUSTNESS⁺** AI system can still function despite unexpected inputs	**FAIRNESS⁺** No unintended bias: AI system makes same decision even if an attribute is changed; Data used to train model is representative **DATA GOVERNANCE** Good governance practices throughout data lifecycle	**ACCOUNTABILITY** Proper management oversight of AI system development **HUMAN AGENCY & OVERSIGHT** AI system designed in a way that will not decrease human ability to make decisions **INCLUSIVE GROWTH, SOCIETAL & ENVIRONMENTAL WELL-BEING** Beneficial outcomes for people and planet

⁺: Principles with technical tests

INTRODUCTION

AI VERIFY'S 11 PRINCIPLES

Area 1: Ensuring that individuals are aware and can make informed decisions

Transparency - Ability to provide responsible disclosure to those affected by AI systems to understand the outcome

Area 2: Ensuring AI operation/results are explainable, accurate and consistent

Explainability - Ability to assess the factors that led to the AI system's decision, its overall behaviour, outcomes, and implications

Repeatability / Reproducibility - The ability of a system to consistently perform its required functions under stated conditions for a specific period of time, and for an independent party to produce the same results given similar inputs

Area 3: Ensuring AI system is reliable and will not cause harm

Safety - AI should not result in harm to humans (particularly physical harm), and measures should be put in place to mitigate harm

Security - AI security is the protection of AI systems, their data, and the associated infrastructure from unauthorised access, disclosure, modification, destruction, or disruption. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure.

Robustness - AI system should be resilient against attacks and attempts at manipulation by third party malicious actors, and can still function despite unexpected input

Area 4: Ensuring that use of AI does not unintentionally discriminate

Fairness - AI should not result in unintended and inappropriate discrimination against individuals or groups

Data Governance - Governing data used in AI systems, including putting in place good governance practices for data quality, lineage, and compliance

Area 5: Ensuring human accountability and control

Accountability - AI systems should have organisational structures and actors accountable for the proper functioning of AI systems

Human Agency & Oversight - Ability to implement appropriate oversight and control measures with humans-in-the-loop at the appropriate juncture

Inclusive Growth, Societal & Environmental Well-being - This Principle highlights the potential for trustworthy AI to contribute to overall growth and prosperity for all – individuals, society, and the planet – and advance global development objectives

SUMMARY

This summary provides an overview of the AI model tested. The details of each principle and the interpretation can be found on the following pages.

AI MODEL INFORMATION

Name of Model Tested:	Binary Classification Model for Credit Risk
Model Type:	Classification
Model Filename:	binary_classification_mock_credit_risk_sklearn.linear_model._logistic.LogisticRegression.sav
Test Dataset:	pickle_pandas_mock_binary_classification_credit_risk_testing.sav
Report Completed:	06 Jun 2023, 12:03:62 PM

OVERALL COMPLETION STATUS

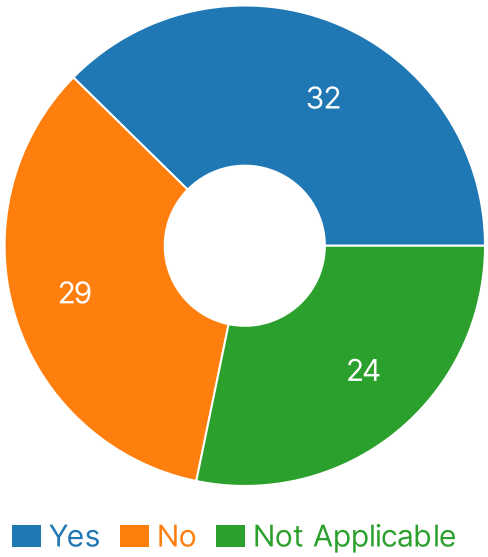
TECHNICAL TESTS

TESTS SUCCESSFULLY RUN	TESTS FAILED TO COMPLETE	TESTS SKIPPED BY USER
3 / 3	0 / 3	0 / 3

PROCESS CHECKS

The company has completed the process checklist of 85 process checks, of which:

- **32 process checks** are indicated as "Yes", meaning that there is documentary evidence for the implementation of these criteria.
- **29 process checks** are indicated as "No". As these process checks have not been implemented, there could be a potential risk that the company needs to assess and/or mitigate[1].
- **24 process checks** are indicated as "Not Applicable"[2].



[1]The company should periodically review that the reason(s) for not implementing the process checks remains valid and aligned with company's values, objectives and regulatory requirements.

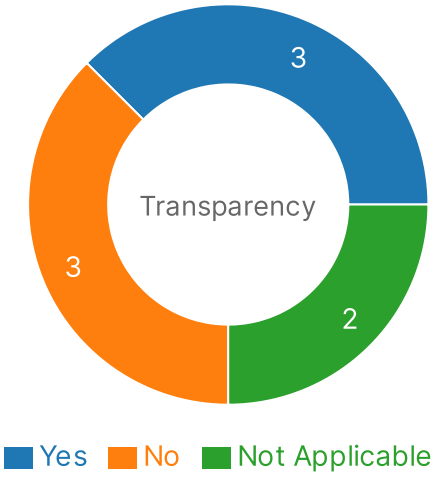
[2]If the operating environment or model changes, company should assess whether these process checks would become relevant.

01 / TRANSPARENCY ON THE USE OF AI AND AI SYSTEMS

Ensuring that individuals are aware and can make informed decisions

---

The principle of **Transparency** was assessed through 8 process checks.



What it means:
Company should review if the current communication mechanisms in place are sufficient to enable those using and/or affected by the AI system to understand how their data is collected and used, and the intended use and limitations of the AI system.

Recommendations(s):
Company can consider consulting the users of or individuals affected by the AI system to find out if the current level of information provided to them is adequate, and if not, to address the information gap accordingly.

Summary Justification
This is a sample summary justification for transparency process checks.

Company did not implement the following testable criteria fully:

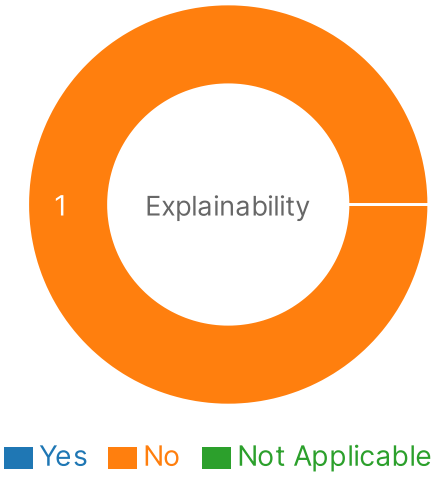
- Provide the necessary information to end users about the use of their personal data to ensure it is processed in a fair and transparent manner
- Where possible (e.g., not compromising IP, safety, or system integrity), identify appropriate junctures in the AI lifecycle to inform end users and/or subjects about the purpose, criteria, limitations, and risks of the decision(s) generated by the AI system in an accessible manner
- Provide information to guide end users on the proper use of the AI system in an accessible manner

02 / UNDERSTANDING HOW AI MODELS REACH DECISION

Ensuring AI operation/results are explainable, accurate and consistent

---

The principle of **Explainability** was assessed through 1 process check and technical test.



Summary Justification

The company did not provide any reason.

Company did not implement the following testable criteria fully:

- Demonstrate a preference for developing AI models that can explain their decisions or that are interpretable by default

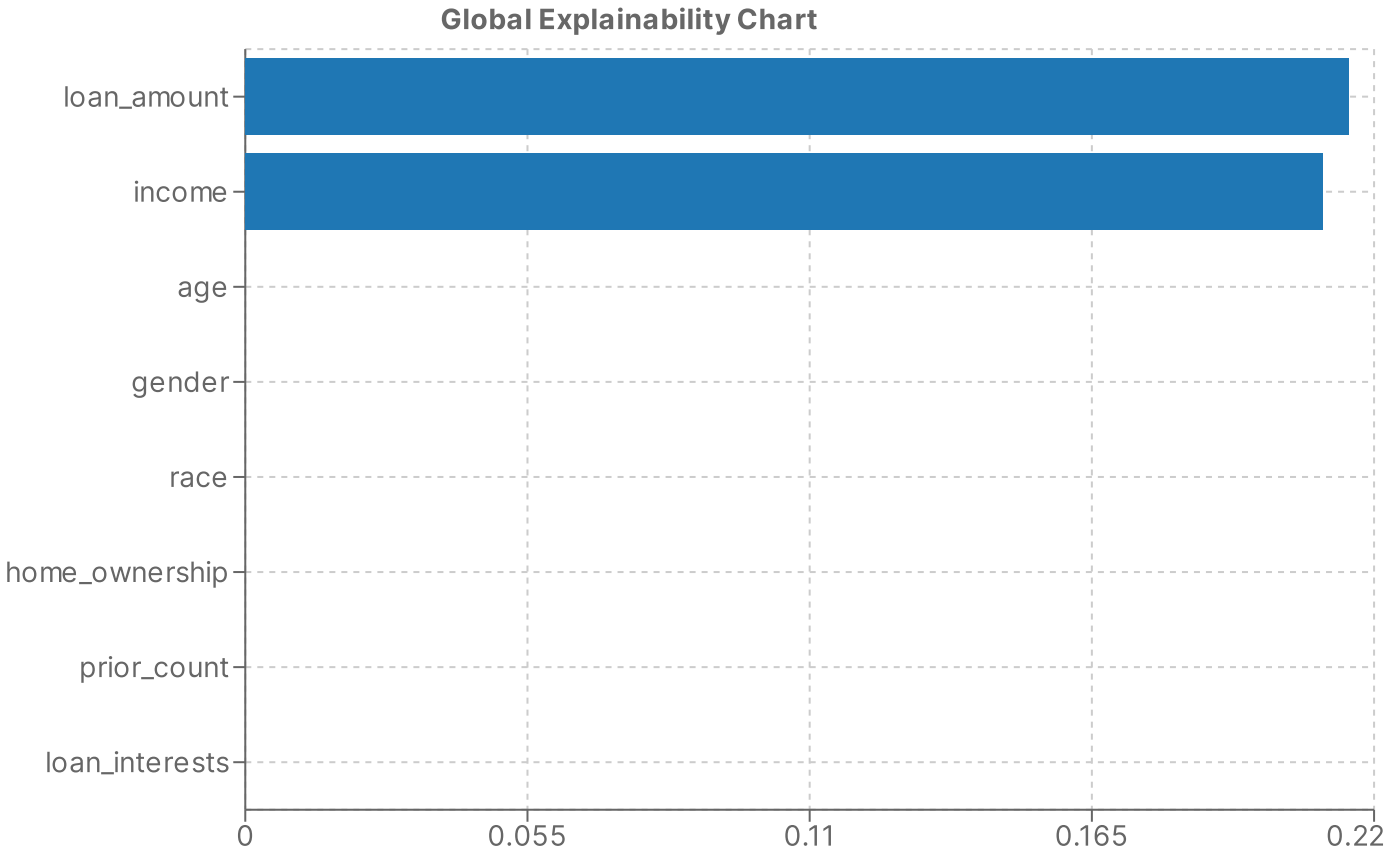
What it means:

When the performance of different models under consideration are similar, by not demonstrating a preference for the model that is more explainable or interpretable by default for deployment, Company runs the risk of not being able to communicate to its stakeholders how the AI model makes its recommendation and may lead to a lack of trust. Company should consider if such risk is acceptable, having considered regulatory requirements, company policies and the intended use of the AI model

Recommendations(s):

If Company chooses a less explainable modelling approach, Company should document its rationale for taking such a risk, having considered the prevailing regulatory requirements, its own internal policies, and the intended use of the AI model.

TECHNICAL TEST



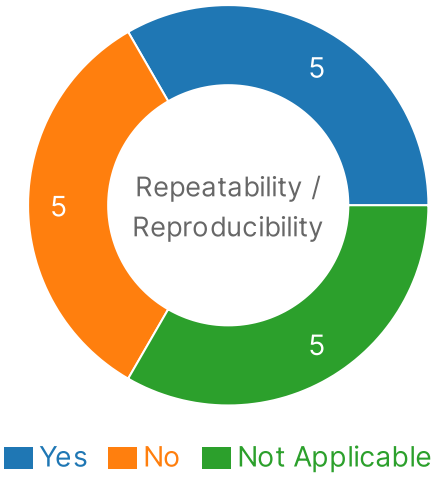
The global explainability test shows the top 8 features affecting the AI model's prediction.
Each bar represents a feature. They are ranked from the highest to the lowest contribution to the predictions. The length of the bar represents the absolute SHAP value across all predictions. A higher value means the feature had more importance on the predicitons, and vice-versa.

What it means:
The test results enable the Company to help its stakeholders understand key factors affecting the AI model's recommendation.

- These features contribute 100.00% towards the final predictions of the AI model.
- Company needs to consider the extent of which these features could be shared with stakeholders. If the company assess that these features should not be made public, company can consider aggregating them.

Recommendation(s)
Company can consider sharing these factors with its stakeholders so that they can better understand how the AI model makes a prediction. However, if the sharing of test results will compromise intellectual property, confidential information, safety and integrity of the system, Company may consider alternatives such as grouping the factors into more generic categories which are non-sensitive and share these categories with stakeholders.

The principle of **Repeatability / Reproducibility** was assessed through 15 process checks.



What it means:
Company may not be able to reproduce the same results and demonstrate consistency of the AI model's behavior under stated conditions. Company should consider if such risk is acceptable, having considered regulatory requirements, company policies and the intended use of the AI model.

Recommendations(s):
Company should consider putting in place processes and measures such as logging capabilities to enable reproducibility of the training process of a model. It is also recommended that Company trace the consistency of the data used by the AI system through the AI lifecycle.

Summary Justification
This is a sample summary justification for reproducibility process checks.

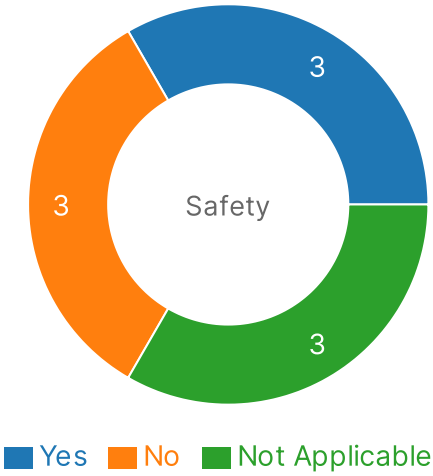
Company did not implement the following testable criteria fully:

- Put in place measures to ensure data quality over time
- Put in place measures to understand the lineage of data, including knowing where the data originally came from, how it was collected, curated, and moved within the organisation over time
- Trace the AI model or rules that led to the decision(s) or recommendation(s) of the AI system
- Put in place adequate logging practices to record the decision(s) or recommendation(s) of the AI system
- Assess for repeatability by reviewing if the model produces the same output based on the same input (Note: this is not relevant when it's time to the retrain model)
- Define the process for developing models and evaluate the process
- Establish a strategy for reproducing the input data used in the training process for every model
- Establish a strategy for ensuring that assumptions still hold across subsequent model retraining process on new input data
- If using a blackbox model or third party model, assess the vendor's claim on accuracy
- Establish a strategy to continuously assess the quality of the output(s) of the AI system and ensure that the operating conditions of a live AI system match the thesis under which it was originally developed

03 / SAFETY & RESILIENCE OF AI SYSTEM

Ensuring AI system is reliable and will not cause harm

The principle of **Safety** was assessed through 9 process checks.



What it means:
By not implementing all the testable criteria, the AI system may carry risk of harm to end users or individuals, which could have been mitigated. This could reduce the overall trust in the AI system.

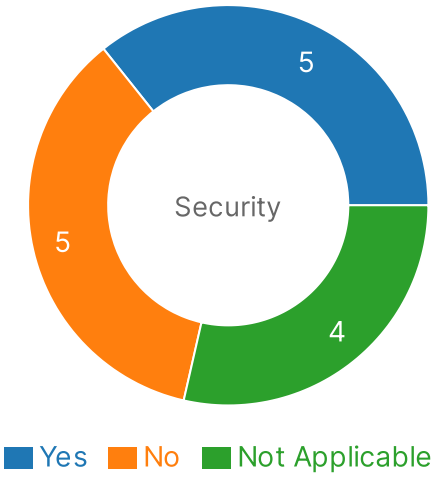
Recommendations(s):
Company should consider putting in place processes and measures to continuously assess, measure and monitor risks of the AI systems that may potentially cause harm. It is also recommended that Company performs risk assessment to demonstrate that sufficient mitigations have been taken to address potential harm.

Summary Justification
This is a sample summary justification for safety process checks.

Company did not implement the following testable criteria fully:

- Assess risks, risk metrics, and risk levels of the AI system in each specific use case, including the dependency of a critical AI system's decisions on its stable and reliable behaviour
- Put in place a process to continuously assess, measure and monitor risks, including the identification of new risks after deployment
- Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or 'conventional')
- Identify residual risk that cannot be mitigated and assess the organisation's tolerance for these risks

The principle of **Security** was assessed through 14 process checks.



What it means:
By not implementing all the testable criteria, Company's AI system may be vulnerable to exploitation by malicious actors, resulting in the compromise of its AI system's confidentiality, integrity and availability. This, in turn, could cause damage and harm to both the end users and the owner of the AI system, including privacy violations, fraud, reputational damage, and potential regulatory challenges.

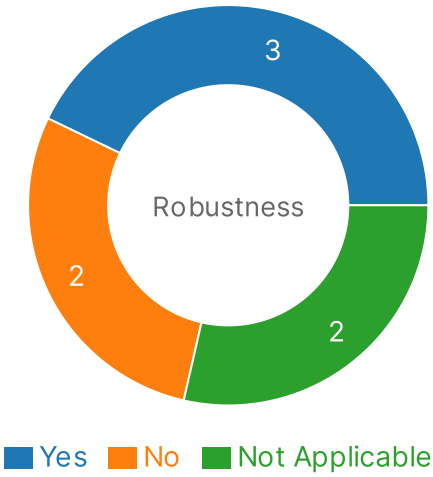
Recommendations(s):
Security is essential in building stakeholder trust in the AI system. Do review periodically the measures that company has chosen not to implement or has assessed to be not applicable to see if justifications for doing so remain valid. As security threats are fast evolving, it is recommended that company should periodically assess security risks and take appropriate actions to continually stay up-to-date.

Summary Justification
This is a sample summary justification for security process checks.

Company did not implement the following testable criteria fully:

- Conduct security risk assessment at the Inception of AI system development
- Put in place security measures during the Verification and Validation of AI system development
- Put in place security measures during the Design and Development of AI system development
- Put in place security measures during the Deployment and Monitoring of AI system development
- Put in place security measures for the Continual / Online Learning Model
- Put in place security measures for End of Life of AI System

The principle of **Robustness** was assessed through 7 process checks and technical test.



What it means:
Company may not be able to maintain AI model's level of performance under any circumstances, such as changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system. This may result in damaging consequences to Company's stakeholders.

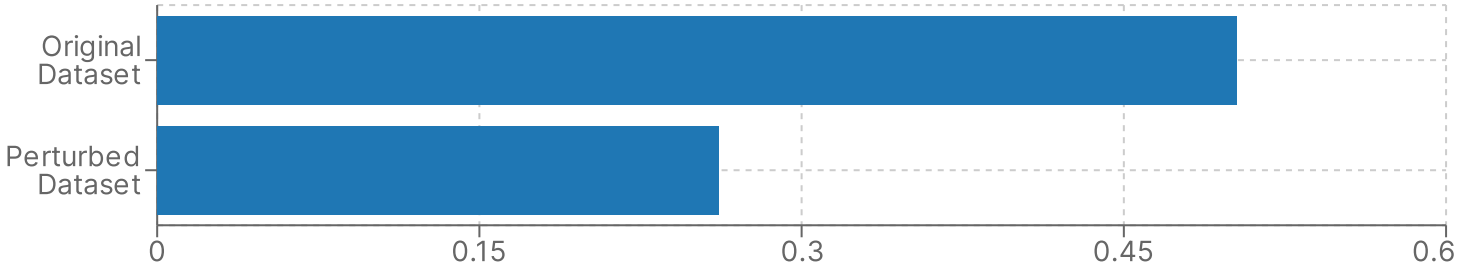
Recommendations(s):
Company should consider putting in measures and processes to monitor and assess the level of resilience against unexpected input that may happen under any circumstances.

Summary Justification
This is a sample summary justification for robustness process checks.

Company did not implement the following testable criteria fully:

- Review factors that may lead to a low level of accuracy of the AI system and assess if it can result in critical, adversarial, or damaging consequences
- Consider whether the AI system's operation can invalidate the data or assumptions it was trained on e.g., feedback loops, user adaptation, and adversarial attacks
- Establish a strategy to monitor and mitigate the risk of black box attacks on live AI systems

TECHNICAL TEST



The robustness test generates perturbed dataset based on your given test samples with the intention to cause your model to produce different predictions. Each bar represents the performance of the model. The longer the bar, the higher accuracy of the model. A robust model will achieve similar accuracy for both original dataset and perturbed dataset. If you model is not robust, the accuracy of the model will reduce with a perturbed dataset.

What it means:
The test results enable the Company to understand whether the model may be affected by dataset that might be perturbed incidentally or intentionally.

- The original and perturbed dataset achieved an accuracy of 50% and 26% respectively.
- The model may not be robust as there seems to have a 24.08% drop in accuracy.

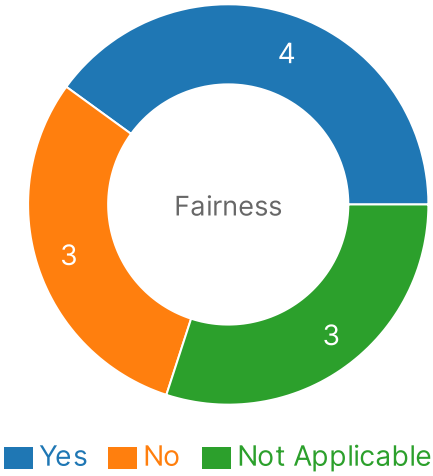
Recommendation(s):
As the magnitude of the drop is considered large, methods to improve the AI system's robustness can be explored. Some suggestions include adding noise and conducting data augmentation of the dataset during training. Additionally the user can consider to relook at the whole deployment and reevaluate the dataset.

04 / FAIRNESS / NO UNINTENDED DISCRIMINATION

Ensuring that use of AI does not unintentionally discriminate

---

The principle of **Fairness** was assessed through 10 process checks and technical test.



What it means:
By not implementing all the testable criteria, Company runs the risk of not being able to monitor and identify potential causes of bias and address them throughout the AI system's lifecycle. This may result in discriminatory outcomes for individuals affected by the AI system. This could also reduce overall trust in the system.

Recommendations(s):
Company should consider putting in place processes to identify and test for potential biases during the entire lifecycle of the AI system. It is also recommended that Company put in place mechanisms to perform mitigation where necessary and document possible limitations that may stem from the composition of the datasets.

Summary Justification
This is a sample summary justification for fairness process checks.

Company did not implement the following testable criteria fully:

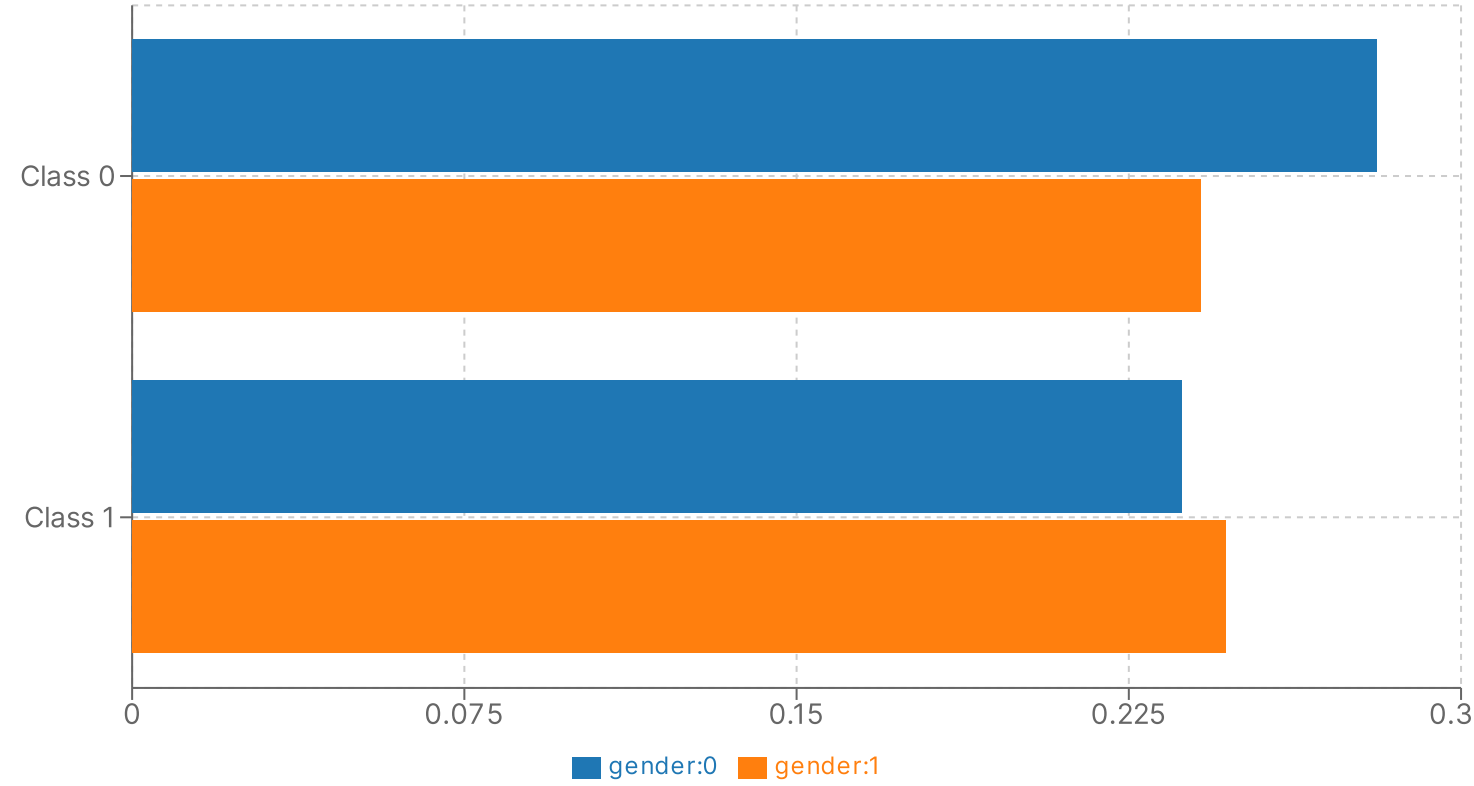
- Put in place processes to test for potential biases during the entire lifecycle of the AI system, so that practitioners can act to mitigate biases based on feedback (e.g., biases due to possible limitations stemming from the composition of the used data sets such as a lack of diversity and non-representativeness)
- Establish a strategy for the selection of fairness metrics that are aligned with the desired outcomes of the AI system's intended application
- Define sensitive features for the organisation that are consistent with the legislation and corporate values
- Establish a process for identifying and selecting sub-populations between which the AI system should produce fair outcomes
- Put in place a mechanism that allows for the flagging of issues related to bias, discrimination, or poor performance of the AI system
- Put in place appropriate mechanisms to ensure fairness in your AI system

TECHNICAL TEST

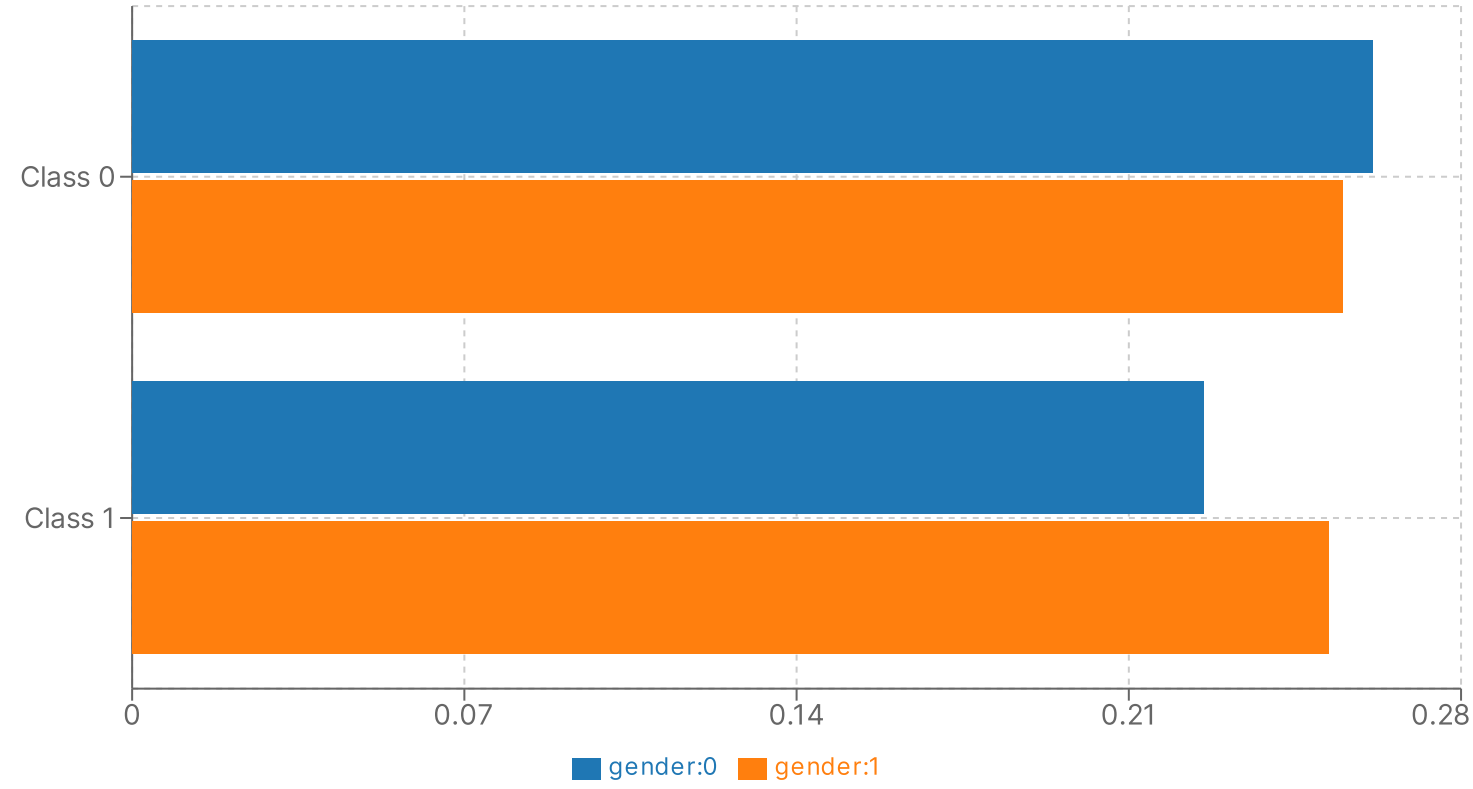
The fairness test shows how correctly your model has predicted the selected sensitive feature(s) (Selected: gender).

The displayed metric(s) are derived from the fairness decision tree's selection. Each bar corresponds to a unique combination of subgroups within the selected sensitive feature(s). The graph's length indicates the magnitude of accuracy/error made by your model while predicting the outcomes.

False Discovery Rate



Negative Predictive Value Parity



What it means:

The test results enable the Company to help its stakeholder understand if the model is able to predict the outcomes fairly among the demographic groups.

You have selected *False Discovery Rate* as an appropriate metric for your use case. In an ideal situation, the parity should be close to 0%.

- For Class 0, the parity between the two subgroups (gender:1 and gender:0) is 0.04
- For Class 1, the parity between the two subgroups (gender:0 and gender:1) is 0.01

You have selected *Negative Predictive Value Parity* as an appropriate metric for your use case. In an ideal situation, the parity should be close to 0%.

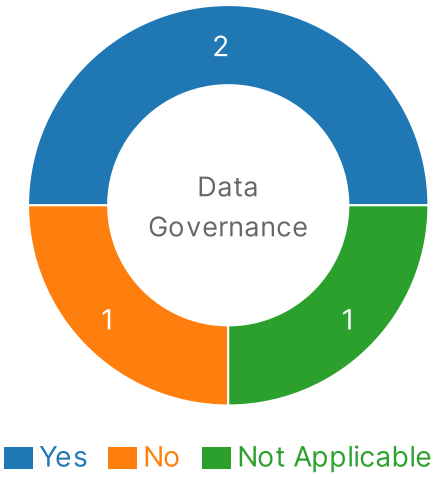
- For Class 0, the parity between the two subgroups (gender:1 and gender:0) is 0.006
- For Class 1, the parity between the two subgroups (gender:0 and gender:1) is 0.026

Recommendations:

Company can consider reviewing these fairness metrics with the relevant stakeholders so that they can better understand if the Al model has predicted outcome fairly among the sensitve features. If the parity is considered negligible and acceptable by the Company, there is no recommendation for further action. If the parity is not acceptable, consider doing the following:

1. Review your dataset to identify any inherent bias in the dataset
2. Review your model parameters and algorithms
3. Apply post-processing mitigation algorithms (See: A Reductions Approach to Fair Classification)

The principle of **Data Governance** was assessed through 4 process checks.



What it means:
By not implementing all the testable criteria, Company runs the risk of potential data quality issues affecting accuracy of the AI model, bias issues relating to unintended discrimination, data security risks resulting in unauthorized access, use or disclosure and/or compliance issues with data protection regulations and laws.

Recommendations(s):
It is recommended that Company implements all the testable criteria. Company should review the reasons for not implementing certain testable criteria and assess if these reasons are still valid. Company should review its data governance policy and explore putting in place relevant standards, guidelines and best practices.

Summary Justification
The company did not provide any reason.

Company did not implement the following testable criteria fully:

- Put in place measures to understand the lineage of data, including knowing where the data originally came from, how it was collected, curated, and moved within the organisation over time
- Ensure data practices comply with relevant regulatory requirements or industry standards

05 / MANAGEMENT AND OVERSIGHT OF AI SYSTEM

Ensuring human accountability and control

The principle of **Accountability** was assessed through 8 process checks.

What it means:
The current organisational structure and internal governance mechanism may not provide sufficient accountability and oversight of the AI system. This may have negative impact on the identification and mitigation of risks associated with this AI system.

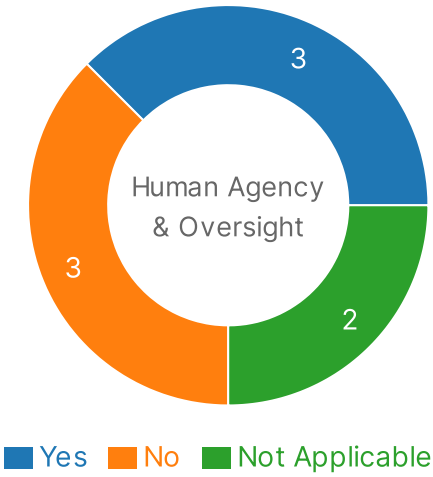
Recommendations(s):
Company should review the current organizational structure and internal governance mechanism to ensure clear accountability for those involved in Company's AI development and deployment.

Summary Justification
This is a sample summary justification for accountability process checks.

Company did not implement the following testable criteria fully:

- Establish clear internal governance mechanisms to ensure clear roles and responsibilities for the use of AI by the organisation
- Define the policy mechanism for enforcing access rights and permissions for the various roles of users
- Establish a strategy for maintaining independent oversight over the development and deployment of AI systems
- If you are using third-party 'black box' models, assess the suitability and limits of the model for your use case

The principle of **Human Agency & Oversight** was assessed through 8 process checks.



What it means:
Company may not have put in place adequate oversight and control measures for human to intervene should AI system fail to achieve its intended goal and result in a negative outcome. This may result in increase in risk of harm to end users of or individuals affected by the AI system.

Recommendations(s):
Company should review the current oversight and control measures to ensure that human is able to improve the operation of AI system or override it in a timely manner when system fails.

Summary Justification
This is a sample summary justification for Human Agency & Oversight process checks.

Company did not implement the following testable criteria fully:

- Ensure that the various parties involved in using, reviewing, and sponsoring the AI system are adequately trained and equipped with the necessary tools and information for proper oversight to:
  - Obtain the needed information to conduct inquiries into past decisions made and actions taken throughout the AI lifecycle
  - Record information on training and deploying models as part of the workflow process
- Ensure specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system
- Put in place a review process before AI models are put into production, where key features and properties of the AI model are shared and visualised in a way that is accessible to decision-makers within the organisation
- Ensure the appropriate parties who are accountable for the AI system (e.g., AI governance committee, AI system owner, and reviewers) have considered how the AI system is used to benefit humans in decision-making processes

The principle of **Inclusive Growth, Societal & Environmental Well-being** was assessed through 1 process check.

What it means:
Company has considered the broader implications of the AI system, i.e., its impact on society and environment, beyond its functional and commercial objectives.

Summary Justification
This is a sample summary justification for Inclusive Growth, Societal & Environmental Well-being process checks.

ANNEX A
PROCESS CHECKLISTS

AI VERIFY

AI GOVERNANCE TESTING FRAMEWORK AND TOOLKIT

TRANSPARENCY

Criteria 1.1 - *Provide the necessary information to end users about the use of their personal data to ensure it is processed in a fair and transparent manner*

1.1.1 Process Align with (1) the PDPC's Advisory Guidelines on Key Concepts in the PDPA; (2) Guide to Accountability; and (3) Guide to Data Protection Impact Assessments	**Process Checks** Documentary evidence of internal policy requiring alignment with existing data protection laws and regulations, which include: (in Singapore) - PDPC's Advisory Guidelines on Key Concepts in the PDPA; - Guide to Accountability; and - Guide to Data Protection Impact Assessments. (outside Singapore) - Applicable data protection laws/regulations	**Completed** Yes **Metric** Internal documentation (e.g., policy document)
Elaboration This is a sample elaboration.		

1.1.2 Process Publish a privacy policy on your organization's website to share information about the use of personal data in the AI system (e.g., data practices, and decision-making processes). The general disclosure notice could include: – Disclosure of third-party engagement – Definition of data ownership and portability – Depiction of the data flow and identify any leakages – Identification of standards the company is compliant with as assurance to customers	**Process Checks** Documentary evidence of a privacy policy on your organization's website to share information about the use of personal data in the AI system (e.g., data practices and decision-making processes). The general disclosure notice could include: – Disclosure of third-party engagement; – Definition of data ownership and portability; – Depiction of the data flow and identify any leakages; and – Identification of standards the company is compliant with as assurance to customers	**Completed** No **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

Criteria 1.2 - *Where possible (e.g., not compromising IP, safety, or system integrity), identify appropriate junctures in the AI lifecycle to inform end users and/or subjects about the purpose, criteria, limitations, and risks of the decision(s) generated by the AI system in an accessible manner*

1.2.1 Process Design an in-house policy on communication to consumers that articulates the principles for transparency, e.g., define the purpose and context of communication to determine how and what to communicate	**Process Checks** Documentary evidence of an in-house policy on communication to consumers that articulates the principles for transparency, e.g., define the purpose and context of communication to determine how and what to communicate	**Completed** Not Applicable **Metric** Internal documentation (e.g., policy document)
Elaboration This is a sample elaboration.		

1.2.2 Process Inform relevant stakeholders that AI is used in your products and/or services	**Process Checks** Documentary evidence of communication to relevant stakeholders that AI is used in the organisation's products and/or services	**Completed** Yes **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

1.2.3 Process For decisions made by the AI system, where possible, communicate to end users the factors leading to the decision e.g., "You are being shown this product because you bought this item."	**Process Checks** Documentary evidence of communicating to end users the factors that lead to decisions made by AI systems	**Completed** No **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

1.2.4 Process	Process Checks	Completed
Consult end users at the earliest stages of AI system development to communicate how the technology is used and how it will be deployed	Documentary evidence of communication with end users at early stages of AI system development concerning how the technology is used and how it will be deployed	Not Applicable **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

1.2.5 Process	Process Checks	Completed
Surface relevant information regarding accuracy, intended use cases, and limitations of the AI system, including the risk assessment, to end users	Documentary evidence of communication with end users concerning the AI system, which includes (where applicable): - accuracy; - confidence scores; - intended use cases; - limitations; and - risk assessment	Yes **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

Criteria 1.3 - *Provide information to guide end users on the proper use of the AI system in an accessible manner*

1.3.1 Process Provide information such as the purpose, intended use and intended response of the AI system to end users	**Process Checks** Documentary evidence of communication with end users concerning the intended use and intended response of the AI system (e.g., Model Card and Data Card)	**Completed** No **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

EXPLAINABILITY

Criteria 2.1 - *Demonstrate a preference for developing AI models that can explain their decisions or that are interpretable by default*

2.1.1 Process If choosing a less explainable modelling approach, document the rationale, risk assessments, and trade-offs of the AI model	**Process Checks** Documentary evidence of considerations for the choice of AI model Considerations include: - rationale; - risk assessment; and - trade-offs	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

REPEATABILITY / REPRODUCIBILITY

Criteria 3.1 - *Put in place methods to record the provenance of the AI model, including the various versions, configurations, data transformations, and underlying source code*

3.1.1 Process Implement version control of source code and frameworks used to develop the model. For each version of the model, track the code version, as well as the parameters, hyperparameters, and source data used	**Process Checks** Documentary evidence of version control of source code and frameworks used to develop the model, including considerations of how much version history is required Each version of the model should track the following: - code version; - parameters; - hyperparameters; and - source data	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 3.2 - *Put in place measures to ensure data quality over time*

3.2.1 Process Verify the quality of data used in the AI system. This may include the following: - accuracy in terms of how well the values in the dataset match the true characteristics of the entity described by the dataset - completeness in terms of attributes and items e.g., checking for missing values, duplicate records - veracity in terms of how credible the data is, including whether the data originated from a reliable source - How recently the dataset was compiled or updated - Relevance for the intended purpose - Integrity in terms of how well extraction and transformation have been performed if multiple datasets are joined; - Usability in terms of how the data are tracked and stored in a consistent, human-readable format - Providing distribution analysis e.g., feature distributions of input data	**Process Checks** Documentary evidence that proves due diligence has been done to ensure the quality of data. This can include the use of relevant processes or software that: - Conducts validation schema checks - Identifies possible errors and inconsistencies at the exploratory data analysis stage, before training the dataset - Assigns roles to the entire data pipeline to trace who manipulated data and by which rule - Allows for review before a change is made - Unit tests to validate that each data operation is performed correctly prior to deployment - Allow for periodic reviewing and update of datasets - Allow for continuous assessment of the quality of the input data to the AI system, including drift parameters and thresholds, where applicable	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.3 - *Put in place measures to understand the lineage of data, including knowing where the data originally came from, how it was collected, curated, and moved within the organisation over time*

3.3.1 Process Maintain a data provenance record to ascertain the quality of the data based on its origin and subsequent transformation. This could include the following: - Take steps to understand the meaning of and how data was collected - Document data usage and related concerns. - Ensure any data labeling is done by a representative group of labelers - Document the procedure for assessing labels for bias - Trace potential sources of errors -Update data - Attribute data to their sources	**Process Checks** Documentary evidence of a data provenance record that includes the following info, where applicable: - clear explanations of what data is used, how it is collected, and why - source of data and its labels - who the labelers were and whether bias tests were conducted to assess if the labelled data was biased (e.g., bias assessment) - how data is transformed over time - risk management if the origin of data is difficult to be established	**Completed** Not Applicable **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.4 - *Trace the data used by the AI system to make a certain decision(s) or recommendation(s)*

3.4.1 Process Log and capture clearly the data used to train a model version, as well as produce inference results using the model (batch scoring or API endpoint)	**Process Checks** Documentary evidence of data used. Data (raw and synthetic data) includes: - data used to train the AI model; - data used to produce inference results using the AI model (batch scoring or API endpoint)	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.5 - *Trace the AI model or rules that led to the decision(s) or recommendation(s) of the AI system*

3.5.1 Process Link the inference results of the model (batch scoring or API endpoint) back to the underlying model and source code	**Process Checks** Documentary evidence of linking the inference results of the model (batch scoring or API endpoint) back to the underlying model and source code	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 3.6 - *Put in place adequate logging practices to record the decision(s) or recommendation(s) of the AI system*

3.6.1 Process Log all inputs and inference outputs of the model, and store them for a reasonable duration so that they can be reviewed	**Process Checks** Documentary evidence of log records covering all inputs and inference outputs of the model. Log records would cover: - decisions(s) of AI system; and/or - recommendation(s) of the AI system - if a human accepted or rejected AI recommendations/decisions, especially when human-in-the-loop is required	**Completed** Not Applicable **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.7 - *Reproduce the training process for every evaluated model (except data)*

3.7.1 Process	**Process Checks**	**Completed**
Version control model artefacts by associating each artefact with the version of code, dependencies, and parameters used in training	Documentary evidence of version control model artefacts. Each artefact includes: - version of codef - dependencies; and - parameters used in training	Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.8 - *Assess for repeatability by reviewing if the model produces the same output based on the same input (Note: this is not relevant when it's time to the retrain model)*

3.8.1 Process	**Process Checks**	**Completed**
Calculate multiple inferences. If the data follows a normal distribution, the accepted limits of this difference (or 95% of it at least) are +/-1.96 times the standard deviation of the differences between the means of the two tests	Documentary evidence of assessment conducted to review if the model produces the same output based on the same input	No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 3.9 - *Define the process for developing models and evaluate the process*

3.9.1 Process Identify a combination of technical metrics and business metrics that AI models are designed to be assessed against	**Process Checks** Documentary evidence of metrics of AI models that are designed to be assessed against. Metrics include: - technical metrics; and/or - business metrics	**Completed** Not Applicable **Metric** Internal documentation.
Elaboration This is a sample elaboration.		

3.9.2 Process Keep track of experiments (e.g., hyperparameters and model performance) used to develop challenger models, document the rationale for developing these models, and how the final deployed model was arrived at	**Process Checks** Documentary evidence of the process in developing the AI model. The process includes: - hyperparameters, model performance, and other relevant aspects used to develop challenger models; - the rationale for developing these models; and - how the final deployed model was derived	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.10 - *Establish a strategy for reproducing the input data used in the training process for every model*

3.10.1 Process Version control the input data used for training where possible. If not possible, avoid changing the raw data at the source, and keep track of the various stages or transformation steps that are part of the data pipeline for AI model development, preferably as a directed acyclic graph (DAG)	**Process Checks** Documentary evidence of having implemented a strategy for reproducing the input data used in the training process for every model. This strategy includes: - data cleaning, data processing, and feature engineering - maintaining version control of the input data used for training the AI model, where possible; or - separating data manipulation process into extraction (data versioning) and processing; or - avoiding changes to the raw data at the source and keeping track of the various stages or transformation steps that are part of the data pipeline for AI model development, preferably as a directed acyclic graph (DAG).	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 3.11 - *Establish a strategy for ensuring that assumptions still hold across subsequent model retraining process on new input data*

3.11.1 Process Record the statistical distribution of input features and output results so that divergence during retraining can be flagged. Monitor input parameters and evaluation metrics for anomalies across retraining runs	**Process Checks** Documentary evidence of establishing a strategy for ensuring that assumptions still hold across subsequent model retraining process on new input data. For example: - K-L divergence and K-S test metrics can be used to compare the statistical distributions of inputs/outputs between two training runs - Moving average and standard deviations can be used to detect a significant change in model performance metrics	**Completed** Not Applicable **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 3.12 - *Reproduce outputs of the AI system*

3.12.1 Process Log audit trail of when and how each model was deployed, including the code used to serve the model, testing/validation data, and what version of the model artefact was used	**Process Checks** Documentary evidence of past outputs of deployed AI system, which can include: - when and how each model was deployed; - the code used to serve the model; and - the version of the model artefact used	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 3.13 - *If using a blackbox model or third party model, assess the vendor's claim on accuracy*

3.13.1 Process Curate the test set and apply the test set on the model to review performance	**Process Checks** Documentary evidence of assessment conducted concerning vendor's claim on the accuracy, if using a blackbox or third party model	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 3.14 - *Establish a strategy to continuously assess the quality of the output(s) of the AI system and ensure that the operating conditions of a live AI system match the thesis under which it was originally developed*

3.14.1 Process Continuous monitoring and periodic validation should be conducted even after models have gone live. This includes: - Model performance, e.g., monitor feature drift, inference drift, accuracy against ground truth - Application performance, e.g,, latency, throughput, error rates	**Process Checks** Documentary evidence of the conduct of continuous monitoring and periodic validation even after models have gone live. This can include: - Notifications to admins when a model/system exceeds some thresholds and the system is paused (if safe to do so) until the model can be improved. Any decisions that have been made/implemented while the AI was below a threshold should be flagged for reevaluation and potentially redress/remediation if harm occurred - Model performance (e.g.,monitor feature drift, inference drift, accuracy against ground truth) - Application performance (e.g., latency, throughput, error rates)	**Completed** Not Applicable **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 4.1 - *Carry out an assessment of materiality on key stakeholders*

4.1.1 Process Complete and submit the Assessment of Materiality to the appropriate parties who are accountable for the AI system (e.g., AI governance committee, AI system owner, and reviewers) and highlight the risks of the proposed AI solution. Document the justifications for decisions on materiality and the application of relevant governance and controls to demonstrate to regulators and auditors that sufficient responsibility has been taken by humans to address potential risks	**Process Checks** Documentary evidence of details of the assessment of materiality on key stakeholders, justifications for decisions on materiality, and the application of relevant governance/controls. The Assessment of Materiality includes the following impact dimensions (where applicable): - probability of harm; - severity of harm; - Likelihood of threat; - Extent of human involvement; - Complexity of AI model; - Extensiveness of impact on stakeholders; - Degree of Transparency; and - Impact on trust	**Completed** Yes **Metric** 1) Internal procedure manual 2) Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 4.2 - *Assess risks, risk metrics, and risk levels of the AI system in each specific use case, including the dependency of a critical AI system's decisions on its stable and reliable behaviour*

4.2.1 Process Document the intended use cases, capabilities, and limitations of AI models e.g., via model cards. This documentation should be stored and retrieved together with the model artefact, as well as surfaced during a review process before the model is deployed into production	**Process Checks** Documentary evidence of risk assessment done for specific use cases. This risk assessment includes documenting* the: - intended use cases, capabilities, and limitations of the AI model (e.g., via model cards) *Note: This documentation should be stored and retrieved together with the model artefact and surfaced during a review process before the model is deployed into production	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 4.3 - *Put in place a process to continuously assess, measure and monitor risks, including the identification of new risks after deployment*

4.3.1 Process Assign a reviewer who is familiar with the downstream use case of an AI model to review the model post-deployment. This process should include model cards/documentation to ensure alignment between intended use cases at modelling and post-deployment	**Process Checks** Documentary evidence of process for continuous risk monitoring for AI model. Process includes: - Assessing, measuring, and monitoring risks at modelling stage; and - identification of new risks after the post-deployment stage	**Completed** Not Applicable **Metric** Internal documentation (e.g., log, register or database)
Elaboration This is a sample elaboration.		

Criteria 4.4 - *Assess whether the AI system might fail by considering the input features and predicted outcomes to aid communication with stakeholders*

4.4.1 Process Where feasible, use AI models that can produce confidence score together with prediction. Low confidence scores may occur when the data contains values that are outside the range of the training data, or for data regions where there were insufficient training examples to make a robust estimate. Implement mechanisms to detect if model input represents an outlier in terms of training data, e.g., return some "data outlier score" with predictions	**Process Checks** Documentary evidence of assessment of whether the AI system might fail by considering the input features and predicted outcomes to aid communication to stakeholders	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 4.5 - *Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or 'conventional')*

4.5.1 Process Implement deployment strategies such as blue-green and canary deployments.	**Process Checks** Documentary evidence of: - implementation of deployment strategies such as blue-green and canary deployments - a plan for graceful failure or failover modes	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

4.5.2 Process Maintain backup model server in blue-green deployment mode.	**Process Checks** Documentary evidence of maintenance of the backup model server in blue-green deployment mode	**Completed** Not Applicable **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

4.5.3 Process Where feasible, use AI models that can produce a confidence score together with the prediction. Design the systems that are using the results of the AI model to handle cases where the model fails or has low confidence, falling back to backup model servers or sensible default behaviour.	**Process Checks** Documentary evidence of the use of AI models that can produce a confidence score together with the prediction, and that it can fall back to backup model servers or sensible default behaviour	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

4.5.4 Process Close the feedback loop by retraining models with ground truth obtained once models are in production.	**Process Checks** Documentary evidence of closing the feedback loop by retraining models with ground truth obtained once models are in production	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 4.6 - *Identify residual risk that cannot be mitigated and assess the organisation's tolerance for these risks*

4.6.1 Process Document the assessment of the residual risk and provide reasons for the tolerance level	**Process Checks** Documentary evidence of assessment of residual risk and the reasons for the organisation's tolerance for these risks	**Completed** Not Applicable **Metric** Internal documentation
Elaboration This is a sample elaboration.		

SECURITY

Criteria 5.1 - *Ensure Team Competency*

5.1.1 Process Ensure that relevant team members are knowledgeable about threats, vulnerabilities, impact, and mitigation measures relevant to securing AI systems and that their knowledge is up to date Relevant team members may include any employee that is involved in the model lifecycle	**Process Checks** Documentary evidence that team members have relevant security knowledge and training on threats, vulnerabilities, impact, and mitigation measures relevant to securing AI systems. This can include, where applicable: - Training records - Attendance records - Assessments - Certifications - Feedback forms	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 5.2 - *Conduct security risk assessment at the Inception of AI system development*

5.2.1 Process Ensure that proper risk assessment has been carried out, in accordance with the relevant industry standards. Risk mitigation steps have been taken	**Process Checks** Documentary evidence that risk assessment has been done in accordance with the relevant industry standards/guidelines/best practices, with risk mitigation steps and factors taken. This can include: - US NIST AI Risk Management Framework - UK NCSC guidance on secure development and deployment of software applications - OWASP Secure Software Development Lifecycle (SSDLC) - CIA triad	**Completed** No **Metric** Internal documentation (e.g., risk assessment)
Elaboration This is a sample elaboration.		

Criteria 5.3 - *Put in place security measures during the Verification and Validation of AI system development*

5.3.1 Process Ensure there is integrity in data and/or models and there is a chain of custody	**Process Checks** Documentary evidence that data and/or models have been obtained from a trusted source. If unable to obtain data from a trusted source, document the reason and process for using synthetic or limited data. This can include practices implemented according to: - UK NCSC supply chain security guidance - ETSI GR SAI 002 Securing AI Data Supply Chain Security - UK DSTL Machine Learning with Limited Data	**Completed** Not Applicable **Metric** Internal documentation
Elaboration This is a sample elaboration.		

5.3.2 Process Assess the integrity of acquired datasets with a robust validation and verification process	**Process Checks** Documentary evidence of assessment of the integrity of acquired datasets with a robust validation and verification process: - For internal labelled data: Have multiple labellers look at each data input and generate notification where labels differ - External procured/created data: Where possible, follow NCSC supply chain security guidance to find a trusted vendor - Randomized audits of data labels to assess error rates	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 5.4 - *Put in place security measures during the Design and Development of AI system development*

5.4.1 Process Ensure that the development environment has been secured, including trust access controls	**Process Checks** Documentary evidence that the development environment has been secured, including trust access controls. This can include: - Secure software development practices - Monitor Common Vulnerabilities and Exposures (CVEs) associated with the software used - Secure firmware and OS - Access controls following the principle of least privilege. - Access logging and monitoring	**Completed** No **Metric** Internal documentation (e.g., access control management document)
Elaboration This is a sample elaboration.		

5.4.2 Process Ensure that the digital assets have been secured, including data at rest and data in transit	**Process Checks** Documentary evidence that the digital assets have been secured, including data at rest and data in transit. This can include: - Implementation of recognised IT standards, such as ISO 27001	**Completed** Not Applicable **Metric** Internal documentation (e.g., asset management document)
Elaboration This is a sample elaboration.		

5.4.3 Process Ensure that changes to the model or data are tracked and stored in a consistent, human- readable format	**Process Checks** Documentary evidence that changes to the model or data are tracked and stored in a consistent, human-readable format. This can include the use of relevant software that: - Tracks which users have made changes - Allows for review before changes to an asset are made - Allows 'roll back' to a backup in case of a security incident	**Completed** Yes **Metric** Internal documentation (e.g., asset management document)
Elaboration This is a sample elaboration.		

5.4.4 Process Implement measures to mitigate attacks on the dataset (e.g., poisoning attacks) Where possible, conduct data sanitisation to remove suspicious or irrelevant data points. Augment the dataset with new data to diversify it and make it harder for attackers to inject poison data. Store the data set securely and ensure that sensitive data is protected and anonymised. Validate the performance of the machine learning model after training to ensure that it has not been poisoned	**Process Checks** Documentary evidence of details of relevant mitigation measures taken. This can include the following measures: - Data sanitisation - Dataset augmentation - Secure storage of dataset - Validation of model performance	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 5.5 - *Put in place security measures during the Deployment and Monitoring of AI system development*

5.5.1 Process Implement measures to mitigate Inference Attacks, Extraction Attacks, or equivalent	**Process Checks** Documentary evidence of relevant mitigation measures taken, including: - Monitoring for API calls and/or input queries - Internal limits on the number of queries allowed from the same IP or with similar inputs - Implementation of secure authentication and access controls to mitigate inference attacks	**Completed** Not Applicable **Metric** Internal documentation (e.g., log, register or database)
Elaboration This is a sample elaboration.		

5.5.2 Process Implement an alert system for anomalous behaviour (e.g., unathorised access)	**Process Checks** Documentary evidence of measures taken, including: - Following appropriate guidance when applying logging and auditing logs - Reporting to the relevant stakeholders and authority when an alert has been raised or an investigation has concluded that a cyber incident has occurred - Using human-in-the-loop to investigate what automated processes flag as unusual	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

5.5.3 Process Develop a vulnerability disclosure process for AI system and organisation. This will allow users to report vulnerabilities in a responsible way	**Process Checks** Documentary evidence that a vulnerability disclosure process for AI system and organisation is developed, such as using UK NCSC Vulnerability Disclosure Toolkit	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 5.6 - *Put in place security measures for the Continual / Online Learning Model*

5.6.1 Process Ensure that risks associated with continuous learning have been considered (e.g., poisoning attack, model/concept drift) Determine if continual learning is still justified with the proper risk mitigations implemented	**Process Checks** Documentary evidence of - Internal approval of pre-determined model performance targets - Continual learning model having achieved pre-determined performance targets before going into production	**Completed** Not Applicable **Metric** Internal documentation (e.g., risk management document)
Elaboration This is a sample elaboration.		

5.6.2 Process Ensure that approved, pre-determined performance targets are achieved before a newly updated continual learning model goes into production	**Process Checks** Documentary evidence of - Internal approval of pre-determined model performance targets - Continual learning model having achieved pre-determined performance targets before going into production	**Completed** Yes **Metric** Internal documentation (e.g., roadmap)
Elaboration This is a sample elaboration.		

Criteria 5.7 - *Put in place security measures for End of Life of AI System*

5.7.1 Process Ensure proper and secure disposal/disclosure/destruction of data and model in accordance with data privacy standards and/or relevant rules and regulations	**Process Checks** Documentary evidence of proper and secure disposal/disclosure/destruction of data and model. This can include adherence to relevant standards, guidelines, rules, and regulations	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

ROBUSTNESS

Criteria 6.1 - *Put in place measures to ensure the quality of data used to develop the AI system*

6.1.1 Process - Implement measures to ensure data is up-to-date, complete, and representative of the environment the system will be deployed in - Log training run metadata to do comparison in production, e.g., parameters, and version model to monitor model staleness - Monitor production versus training data characteristics at production stage e.g., statistical distribution, data types, and validation constraints, to detect data and concept drift	**Process Checks** Evidence of measures implemented that documents: - Performance metrics (e.g., accuracy, AUROC, AUPR) - Prediction confidence score, variation ratio for the original prediction, predictive entropy - That data is of high quality, up-to-date, complete, and representative of the environment the system will be deployed in	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 6.2 - *Review factors that may lead to a low level of accuracy of the AI system and assess if it can result in critical, adversarial, or damaging consequences*

6.2.1 Process Document intended use cases, risks, and limitations (e.g., in model cards)	**Process Checks** Documentary evidence of intended use cases, risks, and limitations in model cards	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 6.3 - *Consider whether the AI system's operation can invalidate the data or assumptions it was trained on e.g., feedback loops, user adaptation, and adversarial attacks*

6.3.1 Process Document intended use cases, risks, limitations (e.g., in model cards)	**Process Checks** Documentary evidence of intended use cases, risks, and limitations in model cards (e.g., in model cards)	**Completed** Not Applicable **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 6.4 - *Put in place a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness*

6.4.1 Process Implement a review process that highlights changes in code (e.g., training, serving), input data (e.g., raw data, features), and output data (e.g., inference results, performance metrics)	**Process Checks** Documentary evidence of mechanism to evaluate when an AI system has been changed to merit a new review of its technical robustness Mechanism should include a review process that highlights changes in: - code (training, serving); - input data (e.g., raw data, features); and - output data (e.g.,inference results, performance metrics)	**Completed** Yes **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 6.5 - *Establish a strategy to monitor and mitigate the risk of black box attacks on live AI systems*

6.5.1 Process Implement methods to mitigate known adversarial attacks at training time, including decisions whether to adopt / not adopt the methods. Note: It may not be possible for all models (e.g., if the model is deterministic or with a model with high level of interactivty with users)	**Process Checks** Documentary evidence of implementing methods to mitigate adversarial attacks at training time, including decisions on whether to adopt / not adopt the methods	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

6.5.2 Process Monitor requests made to live AI system, e.g., frequency and feature distribution of queries, in order to detect whether it is being used suspiciously	**Process Checks** Documentary evidence of monitoring requests made to live AI system, e.g, frequency and feature distribution of queries, in order to detect whether it is being used suspiciously	**Completed** Not Applicable **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

6.5.3 Process Take action on users who exhibit suspicious activity, e.g., flag for review, rate-limit or block further requests, suspend user accounts	**Process Checks** Documentary evidence of action taken on users who exhibit suspicious activity. Possible actions include to: - flag for review; - rate-limit or block further requests; and - suspend user accounts	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

FAIRNESS

Criteria 7.1 - *Assess within-group fairness (also known as individual fairness)*

7.1.1 Process Apply counterfactual fairness assessment	**Process Checks** Documentary evidence of counterfactual fairness assessment	**Completed** Yes **Metric** Internal Documentation
Elaboration This is a sample elaboration.		

Criteria 7.2 - *Put in place processes to test for potential biases during the entire lifecycle of the AI system, so that practitioners can act to mitigate biases based on feedback (e.g., biases due to possible limitations stemming from the composition of the used data sets such as a lack of diversity and non-representativeness)*

7.2.1 Process Monitor the changes in fairness metric values in the lifecycle of the AI system.	**Process Checks** Documentary evidence of implemented processes to test for potential biases during the entire lifecycle of the AI system	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 7.3 - *Establish a strategy for the selection of fairness metrics that are aligned with the desired outcomes of the AI system's intended application*

7.3.1 Process Consider using Fairness Decision Tree (e.g., AI Verify, Aequitas) to select the appropriate metric(s) for your application	**Process Checks** Documentary evidence of strategy/process undertaken to select fairness metrics that align with the desired outcomes of the AI system's intended application. For example, Binary and Multiclass Classification - Equal Parity - Disparate Impact - False Negative Rate Parity - False Positive Rate Parity - False Omission Rate Parity - False Discovery Rate Parity - True Positive Rate Parity - True Negative Rate Parity - Negative Predictive Value Parity - Positive Predictive Value Parity Regression - Mean Absolute Error Parity - Mean Square Error Parity	**Completed** Not Applicable **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 7.4 - *Define sensitive features for the organisation that are consistent with the legislation and corporate values*

7.4.1 Process Identify the sensitive features and their privileged and unprivileged groups where feasible.	**Process Checks** Documentary evidence of identification of sensitive features and its privileged and unprivileged groups. Examples of sensitive features could include religion, nationality, birthplace, gender, and race. Also refer to country-specific guidelines e.g., Singapore's Tripartite Guidelines on Fair Employment Practices and UK Equality Act	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

7.4.2 Process Where feasible, consult the impacted communities on the correct definition of fairness (e.g., representatives of elderly persons or persons with disabilities), values and considerations of those impacted (e.g., individual's preference)	**Process Checks** Documentary evidence of consultations conducted with impacted communities on the correct definition of fairness	**Completed** No **Metric** External / internal correspondence
Elaboration This is a sample elaboration.		

Criteria 7.5 - *Establish a process for identifying and selecting sub-populations between which the AI system should produce fair outcomes*

7.5.1 Process Define this partitioning in terms of sensitive features that models should be prohibited from being trained on, but are used in the evaluation of fairness outcomes.	**Process Checks** Documentary evidence of the establishment of a process for identifying and selecting sub-populations between which the AI system should produce fair outcomes	**Completed** Not Applicable **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 7.6 - *Establish a strategy or a set of procedures to check that the data used in the training of the AI model, is representative of the population who make up the end-users of the AI model*

7.6.1 Process Perform exploratory data analysis. For the sensitive feature, test the representation of each group in the data. Resample data or collect more data if a particular group is severely underrepresented.	**Process Checks** Documentary evidence of the establishment of a strategy or a set of procedures to check that the data used in the training of the AI model, is representative of the population who make up the end-users of the AI model	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 7.7 - *Put in place a mechanism that allows for the flagging of issues related to bias, discrimination, or poor performance of the AI system*

7.7.1 Process Monitor threshold violations of fairness metrics post-deployment and for actual harms	**Process Checks** Documentary evidence of - monitoring of threshold violations of fairness metrics - obtaining feedback from those impacted by the AI system, offering redress and remediation option if feasible	**Completed** No **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 7.8 - *Put in place appropriate mechanisms to ensure fairness in your AI system*

7.8.1 Process Monitor metrics for the latest set of data for the model currently being deployed on an ongoing basis.	**Process Checks** Documentary evidence of monitoring metrics for the latest set of data for the model currently being deployed on an ongoing basis	**Completed** Not Applicable **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 7.9 - *Address the risk of biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness), by applying appropriate adjustments on data samples of minorities*

7.9.1 Process Where possible, handle imbalanced training sets with minorities. Examples: - Oversample minority class - Undersample majority class - Generate synthetic samples (SMOTE)	**Process Checks** Documentary evidence of addressing the risk of biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness), by applying appropriate adjustments on data samples of minorities	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 8.1 - *Put in place measures to ensure data quality over time*

8.1.1 Process Verify the quality of data used in the AI system. This may include the following: - accuracy in terms of how well the values in the dataset match the true characteristics of the entity described by the dataset - completeness in terms of attributes and items e.g., checking for missing values, duplicate records - veracity in terms of how credible the data is, including whether the data originated from a reliable source - How recently the dataset was compiled or updated - Relevance for the intended purpose - Integrity in terms of how well extraction and transformation have been performed if multiple datasets are joined; - Usability in terms of how the data are tracked and stored in a consistent, human-readable format - Providing distribution analysis e.g., feature distributions of input data	**Process Checks** Documentary evidence that proves due diligence has been done to ensure the quality of data. This can include the use of relevant processes or software that: - Conducts validation schema checks - Identifies possible errors and inconsistencies at the exploratory data analysis stage before training the dataset - Assigns roles to the entire data pipeline to trace who manipulated data and by which rule - Allows for review before a change is made - Unit tests to validate that each data operation is performed correctly prior to deployment - Allow for periodic reviewing and update of datasets - Allow for continuous assessment of the quality of the input data to the AI system, including drift parameters and thresholds, where applicable	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 8.2 - *Put in place measures to understand the lineage of data, including knowing where the data originally came from, how it was collected, curated, and moved within the organisation over time*

8.2.1 Process Maintain a data provenance record to ascertain the quality of the data based on its origin and subsequent transformation. This could include the following: - Take steps to understand the meaning of and how data was collected - Document data usage and related concerns. - Ensure any data labeling is done by a representative group of labelers - Document the procedure for assessing labels for bias - Trace potential sources of errors -Update data - Attribute data to their sources	**Process Checks** Documentary evidence of a data provenance record that includes the following info, where applicable: - clear explanations of what data is used, how it is collected and why - source of data and its labels - who the labelers were and whether bias tests were conducted to assess if the labelled data was biased (e.g., bias assessment) - how data is transformed over time - risk management if the origin of data is difficult to be established	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 8.3 - *Ensure data practices comply with relevant regulatory requirements or industry standards*

8.3.1 Process Ensure that assessment has been carried out in accordance with the relevant regulatory requirements and/or industry standards. Mitigation steps have been taken.	**Process Checks** Documentary evidence that assessment has been done in accordance with the relevant data protection laws/ standards/guidelines/best practices. For example: - applicable data protection laws and regulations such as Singapore's Personal Data Protection Act, European Data Governance Act - Singapore's Data Protection Trustmark - Asia Pacific Economic Cooperation Cross Border Privacy Rules and Privacy Recognition for Processors - OECD Privacy Principles - Recognised data governance standards from international standard bodies (e.g., ISO, US NIST, IEEE)	**Completed** Not Applicable **Metric** 1) Internal documentation; 2) Assessment documentation or certification(s)
Elaboration This is a sample elaboration.		

Criteria 8.4 - *Ensure team competency in data governance*

8.4.1 Process Ensure that relevant team members are knowledgeable about their roles and responsibilities for data governance. Relevant team members include any employee that is involved in managing and using the data for the AI system. For example, having a data policy team to manage the tracking of data lineage with proper controls	**Process Checks** Documentary evidence that team members have relevant knowledge and training on data governance. This can include, where applicable: - Training records - Attendance records - Assessments - Certifications - Feedback forms	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

ACCOUNTABILITY

Criteria 9.1 - *Establish clear internal governance mechanisms to ensure clear roles and responsibilities for the use of AI by the organisation*

9.1.1 Process Adapt existing structures, communication lines, procedures, and rules (e.g., three lines of defense risk management model) or implement new ones	**Process Checks** Documentary evidence of adaptation or new implementation of structures, communication lines, procedures, and rules (e.g., three lines of defense risk management model)	**Completed** Yes **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

9.1.2 Process For organisations who are using AI across departments, establish an AI governance committee that comprises representatives from data science, technology, risk, and product to facilitate cross-departmental oversight for the lifecycle governance of AI systems	**Process Checks** Documentary evidence of the establishment of an AI governance committee. This committee should be sufficiently representative. One way to achieve this is by having representatives from: - data science; - technology; - legal and compliance; - risk and product; and - user experience research, ethics, and psychology	**Completed** No **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

9.1.3 Process Enable a process to report on actions or decisions that affect the AI system's outcome, and a corresponding process for the accountable party to respond to the consequences of such an outcome	**Process Checks** Documentary evidence that outlines roles, responsibilities, and key processes for - the reporting on actions or decisions that affect the AI system's outcome; - the corresponding process for the accountable party to respond to the consequences of such an outcome	**Completed** Not Applicable **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 9.2 - *Establish the appropriate process or governance-by-design technology to automate or facilitate the AI system's auditability throughout its lifecycle*

9.2.1 Process Process or technology should handle: - Version control of code and model - Version data or maintain immutable data - Audit trail of deployment history, log inputs/outputs, associate server predictions with the originating model	**Process Checks** Documentary evidence of the establishment of the appropriate process or governance-by-design technology to automate or facilitate the AI system's auditability throughout its lifecycle. The process or technology should handle: - Version control of code and model; - Version data or maintain immutable data; and - Audit trail of deployment history, log inputs/outputs, associate server predictions with the originating model	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 9.3 - *Define the policy mechanism for enforcing access rights and permissions for the various roles of users*

9.3.1 Process Implement fine-grained access control that aligns with various roles for users: - Access to code and data for training AI models - Access to code and data for deploying AI models - Access to different execution environments - Permission to perform various actions (e.g., launch training job, review model, deploy model server) - Permission to define access control rules and perform other administrative functions	**Process Checks** Documentary evidence of the implementation of fine-grained access control that aligns with various roles for users, which include: - Access to code and data for training AI models - Access to code and data for deploying AI models - Access to different execution environments - Permission to perform various actions (e.g., launch training job, review model, deploy model server) - Permission to define access control rules and perform other administrative functions	**Completed** No **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 9.4 - *Establish a strategy for maintaining independent oversight over the development and deployment of AI systems*

9.4.1 Process Reviewers should be distinct from those who are training and deploying models. However, it is acceptable to have the same individuals training and deploying models	**Process Checks** Documentary evidence of strategy for maintaining independent oversight over the development and deployment of AI systems	**Completed** Not Applicable **Metric** Internal documentation (e.g., log, register or database)
Elaboration This is a sample elaboration.		

Criteria 9.5 - *If you are using third-party 'black box' models, assess the suitability and limits of the model for your use case*

9.5.1 Process Evaluate the necessity of third-party models e.g., they are trained on data otherwise not accessible to your organisation ,or you do not have the requisite capability to build AI systems in-house	**Process Checks** Documentary evidence of evaluation completed regarding the necessity of third-party models	**Completed** Yes **Metric** Internal documentation
Elaboration This is a sample elaboration.		

9.5.2 Process Demonstrate effort to understand how the third-party models were built, including 1) what data was used to train the models, 2) how the models are assessed for effectiveness and explainability 3) under what circumstances does the AI system perform poorly	**Process Checks** Documentary evidence of effort undertaken to understand how the third-party models were built, which includes: - what data was used to train the models; - how the models are assessed for effectiveness and explainability; and - under what circumstances does the AI system perform poorly	**Completed** No **Metric** Internal documentation
Elaboration This is a sample elaboration.		

Criteria 10.1 - *Ensure that the various parties involved in using, reviewing, and sponsoring the AI system are adequately trained and equipped with the necessary tools and information for proper oversight to:*
- Obtain the needed information to conduct inquiries into past decisions made and actions taken throughout the AI lifecycle
- Record information on training and deploying models as part of the workflow process

10.1.1 Process Put in place guided flow for documenting (i) important info via model cards, forms, SDK library; and (ii) important processes that provide objective criteria for decision-making (e.g., fairness metrics selection)	**Process Checks** Documentary evidence of guided flow for documenting: - important info via model cards, forms, SDK library; and - important processes that provide objective criteria for decision-making (e.g., fairness metrics selection)	**Completed** Yes **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

10.1.2 Process Implement a data management system to gather and organise relevant information based on the needs of different user roles (e.g., reviewing models, and monitoring live systems)	**Process Checks** Documentary evidence of data management system to gather and organise relevant information based on the needs of different user roles	**Completed** No **Metric** Internal documentation (e.g., procedure manual, log, register, or database)
Elaboration This is a sample elaboration.		

Criteria 10.2 - *Ensure specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system*

10.2.1 Process Define the role of the human in its oversight and control of the AI system (e.g., human-in-the-loop, human-out-the-loop, human-over-the-loop)	**Process Checks** Documentary evidence of the definition of the role of human in oversight and control of the AI system	**Completed** Not Applicable **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

10.2.2 Process When the AI model is making a decision for which it is significantly unsure of the answer/prediction, consider designing the system to be able to flag these cases and triage them for a human to review.	**Process Checks** Documentary evidence of consideration made in the design of the AI system on its ability to flag instances when it is making a decision for which it is significantly unsure of the answer/prediction, in order that such cases be triaged for a human to review	**Completed** Yes **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

10.2.3 Process Implement mechanisms to detect if model input represents an outlier in terms of training data (e.g., return some "data outlier score" with predictions)	**Process Checks** Documentary evidence of implementation of mechanisms to detect if model input represents an outlier in terms of training data	**Completed** No **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 10.3 - *Put in place a review process before AI models are put into production, where key features and properties of the AI model are shared and visualised in a way that is accessible to decision-makers within the organisation*

10.3.1 Process Implement a systematic review process to present performance, explainability, and fairness metrics in a way that is understandable by data science, product, legal and risk, experience research, and ethics teams	**Process Checks** Documentary evidence of the implementation of a systematic review process to present performance, explainability, and fairness metrics in a way that is understandable by relevant teams (e.g., data science, product, legal and risk, experience research, and ethics teams)	**Completed** Not Applicable **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

Criteria 10.4 - *Establish a frequency and process for testing and re-evaluating AI systems*

10.4.1 Process After models are put into production, put in place mechanisms to review the performance of the models on an ongoing basis, either continuously or at regular intervals. Criteria could be time-based (e.g., every 2 years) or event-based (before the launch of a new AI product, after the introduction of new data, operating context has changed due to external circumstances), or when the AI system has undergone substantial modification.	**Process Checks** Documentary evidence of the establishment of a frequency and process for testing and re-evaluating AI systems	**Completed** Yes **Metric** Internal documentation of physical testing
Elaboration This is a sample elaboration.		

Criteria 10.5 - *Ensure the appropriate parties who are accountable for the AI system (e.g., AI governance committee, AI system owner, and reviewers) have considered how the AI system is used to benefit humans in decision-making processes*

10.5.1 Process Declaration of transparency on how and where in the decision-making process the AI system is used to complement or replace the human.	**Process Checks** Documentary evidence of the declaration of transparency on how and where in the decision-making process the AI system is used to complement or replace the human	**Completed** No **Metric** 1) Internal documentation (e.g., procedure manual) 2) External / internal correspondence
Elaboration This is a sample elaboration.		

INCLUSIVE GROWTH, SOCIETAL & ENVIRONMENTAL WELL-BEING

Criteria 11.1 - *Ensure that the development of AI system is for the beneficial outcomes for individuals, society, and the environment*

11.1.1 Process Put in place a process to determine that the development and deployment of the AI system is for the benefit of people, society, and the environment, where applicable	**Process Checks** Documentary evidence of consideration of AI system's impact on individuals, society, and environment, which may include (where applicable): - Human capabilities to learn and make decisions - Skills, jobs, and/or job quality - Creative economies - Discriminatory and/or exclusionary norms - Environmental concerns	**Completed** Yes **Metric** Internal documentation (e.g., procedure manual)
Elaboration This is a sample elaboration.		

ANNEX B
TECHNICAL TESTS

AI VERIFY

AI GOVERNANCE TESTING FRAMEWORK AND TOOLKIT

FAIRNESS TEST

Fairness is about designing AI systems that avoid creating or reinforcing unfair bias in the AI system, based on the intended definition of fairness for individuals or groups, that is aligned with the desired outcomes of the AI system.

In this technical test, the tool generates fairness metrics. Depending on the use case and type of model, users can select the relevant fairness metric(s) that are most appropriate.

Fairness for Classification

The fairness test shows how correctly your model has predicted the selected sensitive feature(s) (Selected: gender). These fairness metrics are calculated based on the performance measurement for classification models. The table shows a list of fairness metrics that are generated in this report.

Fairness Metrics	Description
False Negative Rate Parity	The difference between two groups based on the percentage of incorrect predictions among the actual negative values.
False Positive Rate Parity	The difference between two groups based on the percentage of incorrect predictions among the actual positive values.
False Discovery Rate Parity	The difference between two groups based on the percentage of incorrect predictions among those that are predicted as positive.
False Omission Rate Parity	The diffrence between two groups based on the percentage of incorrect predictions among those that are predicted as negative.
True Positive Rate Parity	The difference between two groups based on the percentage of correct predictions among the actual positive values.
True Negative Rate Parity	The difference between two groups based on the percentage of correct predictions among the actual negative values.
Positive Predictive Value Parity	The difference between two groups based on the percentage of correct predictions among the labels that are predicted as positive.
Negative Predictive Value Parity	The difference between two groups based on the percentage of correct predictions among the labels that are predicted as negative.

Fairness Metrics

The displayed metric(s) demonstrate the equity between two subgroups. In cases where the selected feature consists of more than two subgroups (such as race with multiple subgroups), the parity value is determined by comparing the subgroup with the smallest value to the subgroup with the largest value.

- *Disparate Impact*: The closer the value is to 1, the better it is.
- *Equal Selection Parity*: The smaller the value, the better it is.
- Other fairness metrics: The smaller the value, the better it is

False Negative Rate	False Positive Rate	False Discovery Rate
gender:0 vs gender:1 Class 0 0.013	gender:1 vs gender:0 Class 0 0.028	gender:1 vs gender:0 Class 0 0.04
gender:1 vs gender:0 Class 1 0.028	gender:0 vs gender:1 Class 1 0.013	gender:0 vs gender:1 Class 1 0.01

False Omission Rate	True Positive Rate	True Negative Rate
gender:0 vs gender:1 Class 0 0.01	gender:0 vs gender:1 Class 0 0.02	gender:1 vs gender:0 Class 0 0.008
gender:1 vs gender:0 Class 1 0.04	gender:1 vs gender:0 Class 1 0.008	gender:0 vs gender:1 Class 1 0.02

Positive Predictive Value Parity	Negative Predictive Value Parity	Equal Selection Parity
gender:0 vs gender:1 Class 0 0.026	gender:1 vs gender:0 Class 0 0.006	gender:0 and gender:1 Class 0 12
gender:1 vs gender:0 Class 1 0.006	gender:0 vs gender:1 Class 1 0.026	gender:0 and gender:1 Class 1 6

Disparate Impact
gender:0 and gender:1 Class 0 1.022
gender:0 and gender:1 Class 1 0.988

ROBUSTNESS TEST

Robustness requires that AI systems maintains its level of performance under any circumstances, including potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner.

In this technical test, the tool generates performance metrics when perturbed testing datasets were input to the model. The changes in performance give an idea of the model's robustness to potential changes in inputs. Depending on the use case and type of model, users can choose to investigate robustness to adversarial perturbations and/or natural corruptions

ROBUSTNESS TOOLBOX

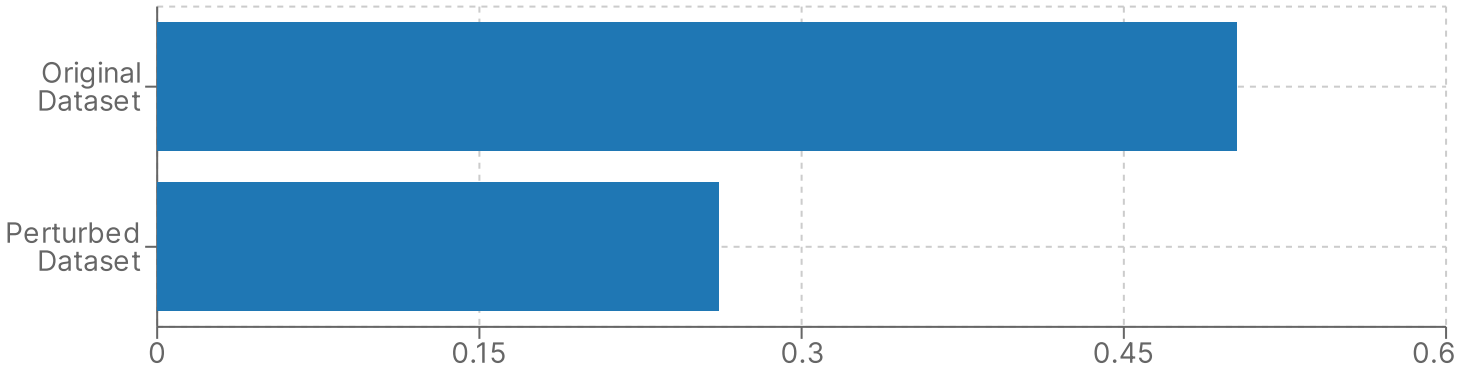
Robustness Toolbox uses Boundary Attack to perturb the test dataset. Boundary Attack is an attack that starts by adding a large amount of noise to a data point intentionally to cause a model it misclassified by the model. We use Salt-and-pepper noise to create the large amount of noise. Then, it will reduce the amount of noise added while maintaining misclassification. This algorithm does not depend on the underlying model's architecture or parameters.

This algorithm is developed for **image dataset** but can also be used to create noise on tabular dataset. However, it is to note that testing on tabular dataset may warrant caution when interpreting the results as this is not well-tested.

Results

Total Number of Samples 2500

Successful Perturbed Rate 100.00%



Each bar represents the performance of the model. The longer the bar, the higher accuracy of the model. A robust model will achieve similar accuracy for both original dataset and perturbed dataset. If you model is not robust, the accuracy of the model will reduce with a perturbed dataset.

What it means:
The test results enable the Company to understand whether the model may be affected by dataset that might be perturbed incidentally or intentionally.

- The original and perturbed dataset achieved an accuracy of 50% and 26% respectively.
- The performance for both datasets are the same.

Example of a perturbed sample and its predicted value
Note:

- The perturbed sample may not be successful in changing the prediction
- 5/8 features will be shown in the sample below

Feature Name	age	gender	income	race	Prediction
Original #0	86	1	64570	2	1
Perturbed #0	0	233600.629233494	154520.64634748254	37768.277248563085	0

EXPLAINABILITY TEST

Explainability is about ensuring AI driven decisions can be explained and understood by those directly using the system to enable or carry out a decision, to the extent possible.The degree to which explainability is needed also depends on the aims of the explanation, including the context, the needs of stakeholders, types of understanding sought, mode of explanation, as well as the the severity of the consequences of errorneous or inaccurate output on human beings.Explainability is an important component of a transparent AI system.

In this technical test, the tool generates feature contribution - based explanations from the given input testing data and model.The results determine if explanations can be generated for a given model, which is an indicator of explainability.

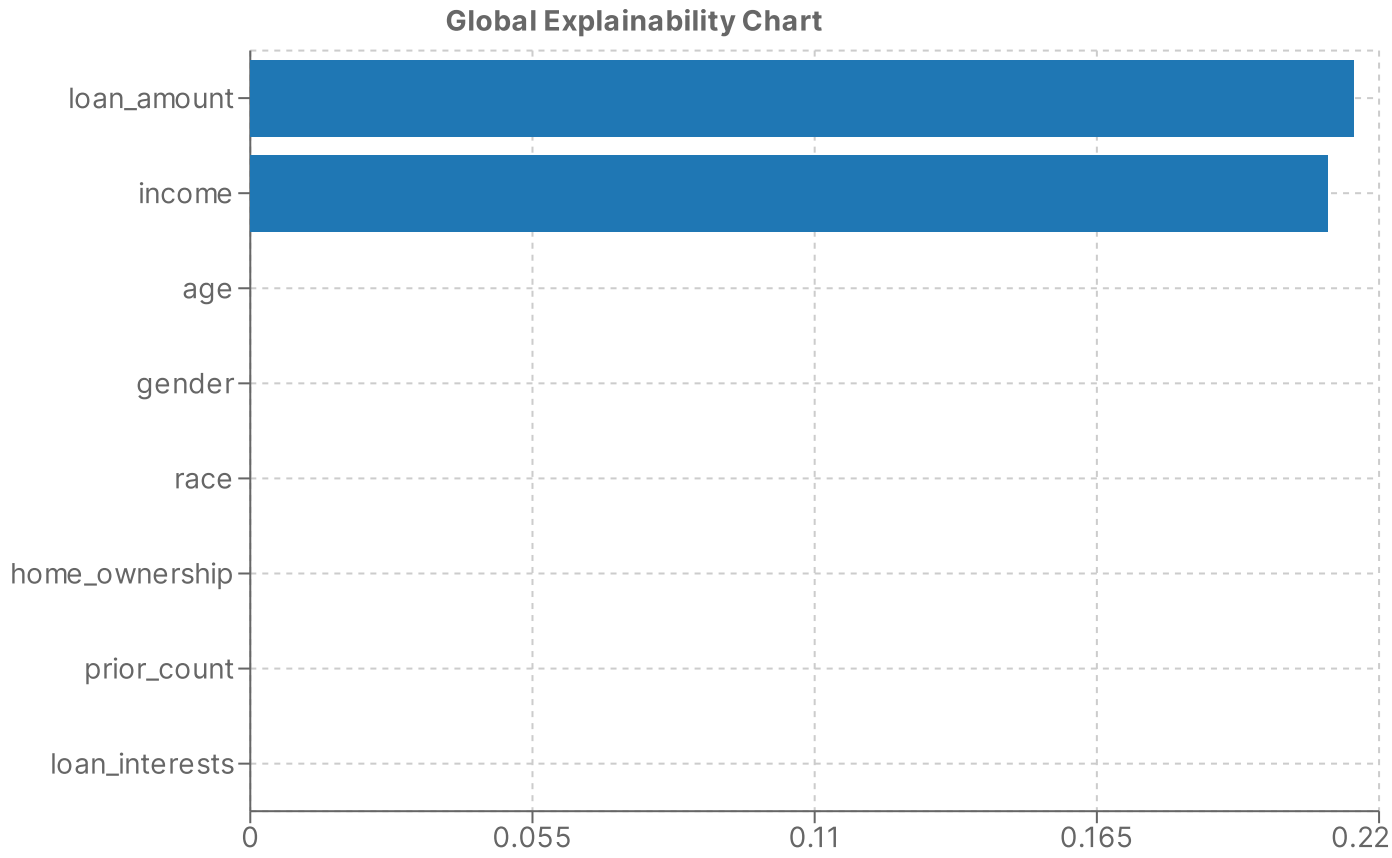
SHAP

SHAP explains how your features affect your overall predictions by using Shapley Values.

How to read?

The features are ranked according to their contributions to the overall predictions.

The y-axis represents the features. They are ranked from the highest to lowest contribution to the predictions. If the feature names are not given, they will be masked as Feature X (where X is a number) instead. The x-axis represents the absolute average SHAP values across all predictions. A higher value means that the feature had more influence on the predictions. The colours represent the output classes and the number of colours correspond to the number of unique output values in the predictions.



From the results, *loan_amount* contributed to the overall predictions the most as it has the highest SHAP value. This is useful for explaining that it is the most important factor influencing the model's predictions. A similar analysis can be done for the rest of the features.

Recommendation(s)

You may consider reviewing features with highest and lowest contribution to the predictions. Features with extremely high contribution might cause model overfits while features with extremely low contribution may cause an overhead to your model efficiency.