Fair AI Platform

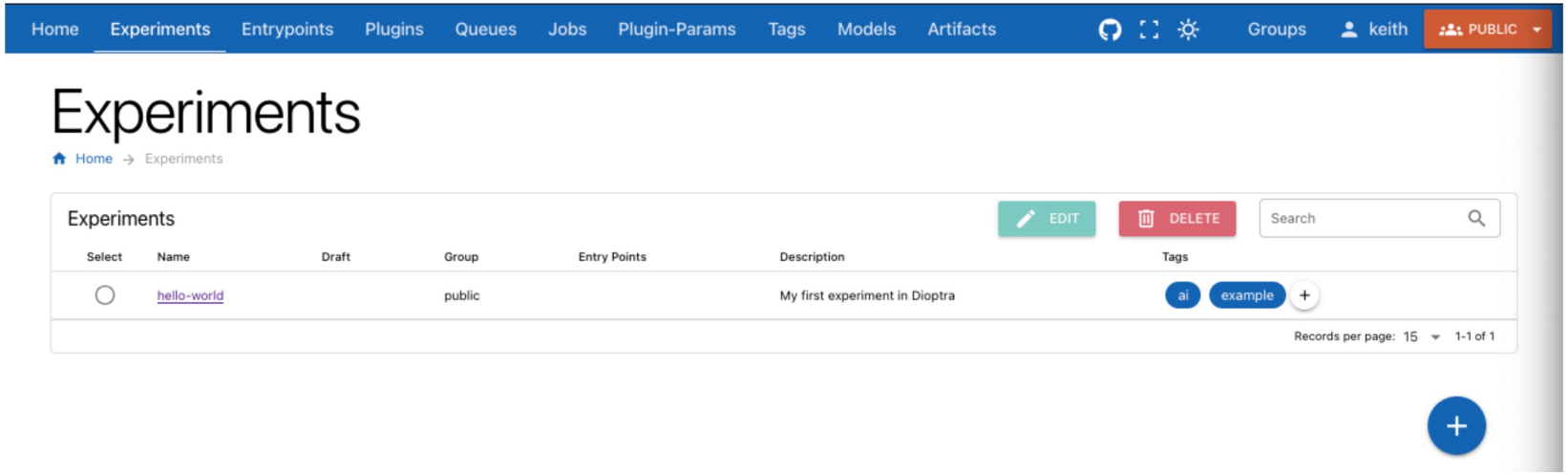
Dioptra, AI360, AI Verify

김초은

Dioptra

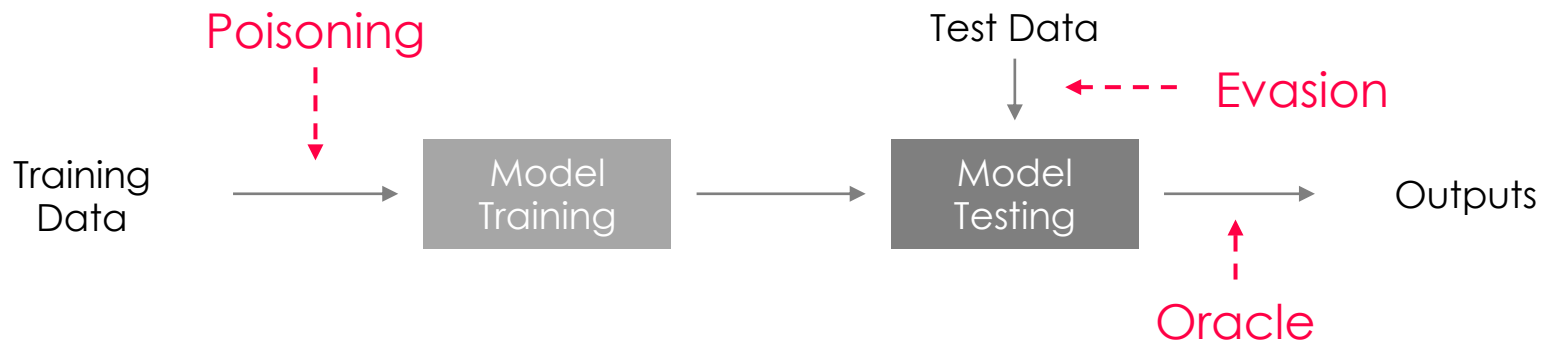
Dioptra

ML 알고리즘이 다양한 적대적 공격에 얼마나 잘 견딜 수 있는지 판단하는 데
도움을 주는 테스트베드



- 모듈형 구조를 가져 다양한 데이터셋, 모델, 적대적 공격 그리고 방어 등 모델 환경에 대해 쉽게 사용자화하여 실험할 수 있음
- attack, defense 와 관련된 내장 플러그인이 많이 구현되어 있어 빠른 실험 세팅과 모니터링이 가능함

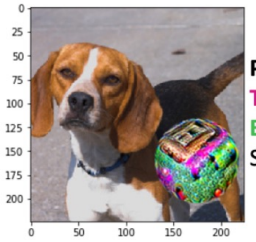
Attack



Evasion Attack	model이 잘못 작동하도록 test data에 조작을 가하는 방법 Patch evasion, Noise evasion
Poisoning Attack	model이 잘못 학습되도록 train data에 조작을 가하는 방법 Clean Label Poisoning
Oracle Attack	모델을 복제할 목적으로 특정 parameter 또는 사용된 train set에 대한 상세 정보 등을 알아내는 방법

Attack

Patch Evasion

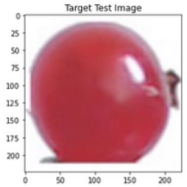


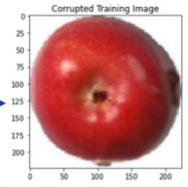
Predictions:
Toaster 30%
Beagle 11%
Spatula 8%

Clean Label Poisoning

Original: Apple

Target: Currant

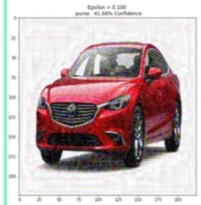




Poisoned Apple

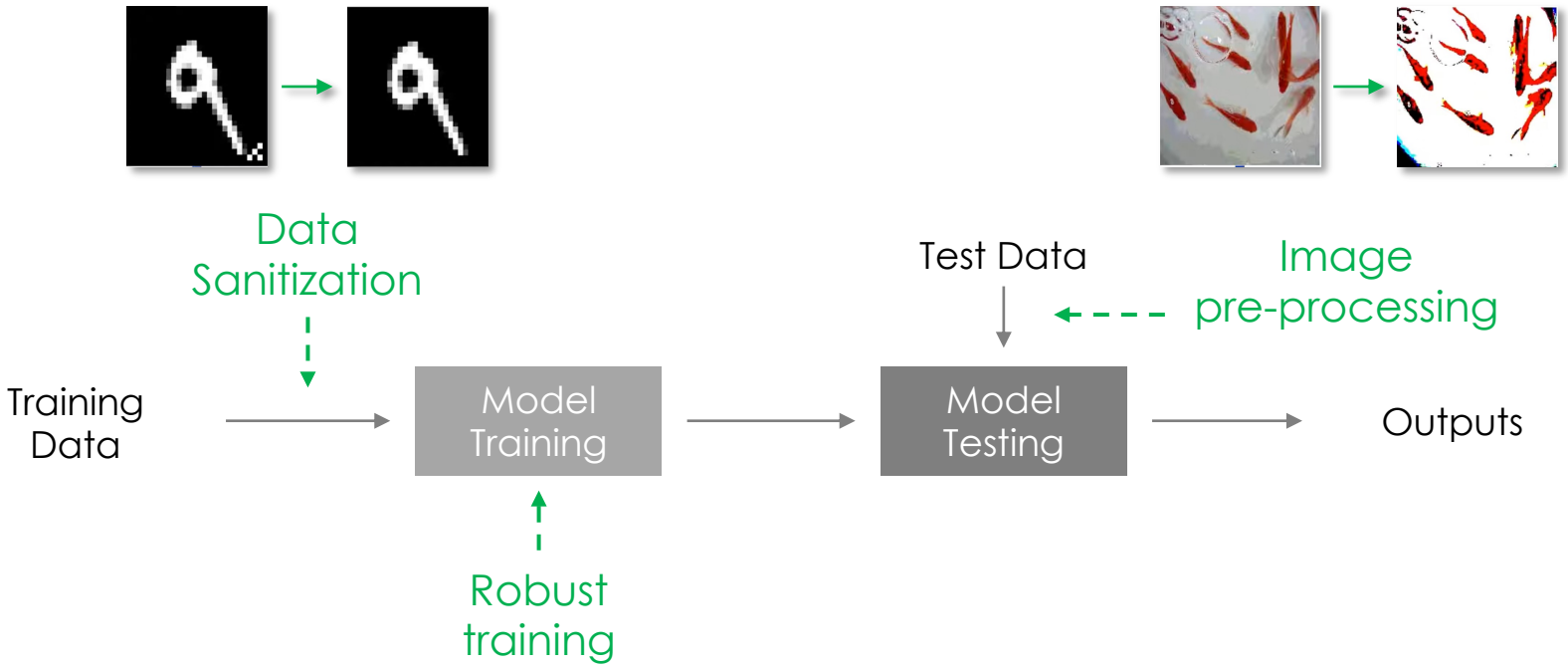
Noise Evasion

sports car: 75%



purse: 41%

Defense



외에도 다양한 defense 방법이 연구되고 있음

Entry point

training architecture

- ShallowNet
- AlexNet
- LeNet
- ResNet50
- VGG16
- …

data augmentation

- patch augmentation
- poison frogs
- adversarial training
- …

inference pre-processing

- spatial smoothing
- defensive distillation
- …

dataset

- MNIST
- Fruits360
- ImageNet
- …

attack on trained model

- fast gradient method
- patch
- membership inference

metric

- clean accuracy
- adversarial accuracy
- robustness radius

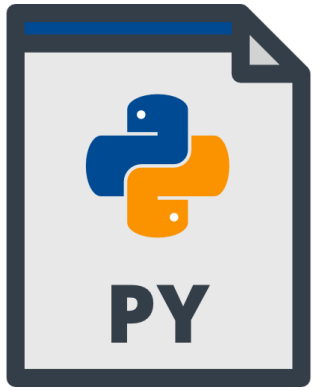
테스트베드 모듈화로 다양한 환경에서 평가가 용이하도록 함

Entry point

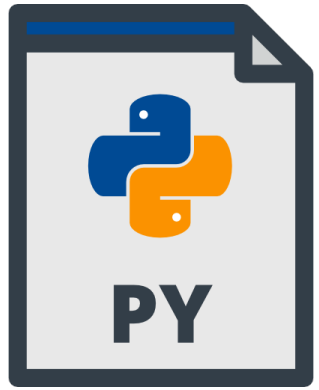
yaml formatted MLproject file

```yaml
name: My Project

entry_points:
  train:
    parameters:
      data_dir: { type: path, default: "/nfs/data" }
      image_size: { type: string, default: "28,28,1" }
    command: >
      python src/train.py
      --data-dir {data_dir}
      --image-size {image_size}
  infer:
    parameters:
      run_id: { type: string }
      image_size: { type: string, default: "28,28,1" }
    command: >
      python src/infer.py
      --run-id {run_id}
      --image-size {image_size}
```

executable Python script



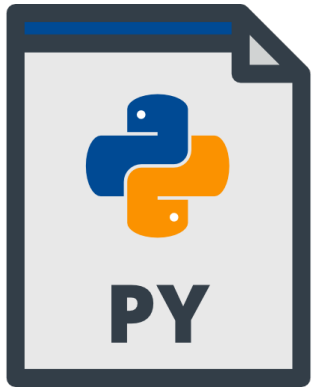
`src/train.py`



`src/infer.py`

모델을 학습하고 평가하는 `train`, `infer` 외에도 `gaussian_augmentation`, `spatial_smoothing`과 같은 데이터와 관련된 entry point도 정의할 수 있다.

Entry point



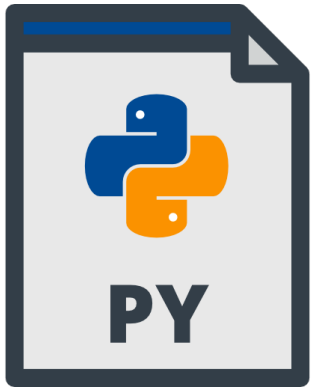
src/train.py

모델을 학습하고 평가하는 모듈

```
import os

import click
from dioptra.sdk.utilities.contexts import plugin_dirs
from dioptra.sdk.utilities.logging import (
    StderrLogStream,
    StdoutLogStream,
    attach_stdout_stream_handler,
    clear_logger_handlers,
    configure_structlog,
    set_logging_level,
)


def _coerce_comma_separated_ints(ctx, param, value):
    return tuple(int(x.strip()) for x in   value.split(","))

@click.command()
@click.option(
    "--data-dir",
    type=click.Path(
        exists=True, file_okay=False, dir_okay=True,
        resolve_path=True, readable=True
    ),
    help="Root directory for shared datasets",
)
@click.option(
    "--image-size",
    type=click.STRING,
    callback=_coerce_comma_separated_ints,
    help="Dimensions for the input images",
)
```

환경 세팅

Entry point



src/train.py

모델을 학습하고 평가하는 모듈

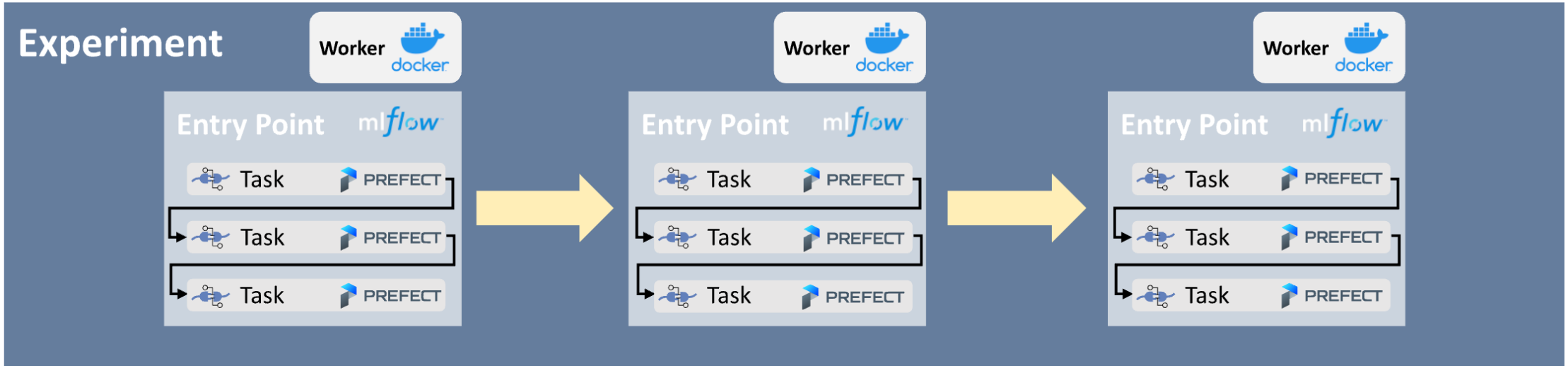
```python
def train(data_dir, image_size):
    # Only use this when training a model
    mlflow.autolog()
    # Start the active run context for MLFlow
    with mlflow.start_run() as active_run:
        flow = init_flow()
        state = flow.run(parameters = dict(data_dir=data_dir, image_size=image_size))
    return state

from prefect import Flow, Parameter
from dioptra import pyplugs

_PLUGINS_IMPORT_PATH: str = "dioptra_builtins"


def init_flow() -> Flow:
    with Flow("Image Resizer") as flow:
        data_dir, image_size = Parameter("data_dir"), Parameter("image_size")
        resize_output u= pyplgs.call_task(
            f"{_PLUGINS_IMPORT_PATH}.data",
            "images",
            "resize",
            data_dir=training_dir,
            image_size=image_size,
        )
    ...

if __name__ == "__main__":
    log_level = os.getenv("DIOPTRA_JOB_LOG_LEVEL", default="INFO")
    as_json = True if os.getenv("DIOPTRA_JOB_LOG_AS_JSON") else False

    clear_logger_handlers(get_prefect_logger())
    attach_stdout_stream_handler(as_json)
    set_logging_level(log_level)
    configure_structlog()

    with plugin_dirs(), StdoutLogStream(as_json), StderrLogStream(as_json):
        _ = train()
```

metric, parameter, model 등에 대한 로그

dioptra 내장 플러그인
https://pages.nist.gov/dioptra/user-guide/task-plugins-collection.html

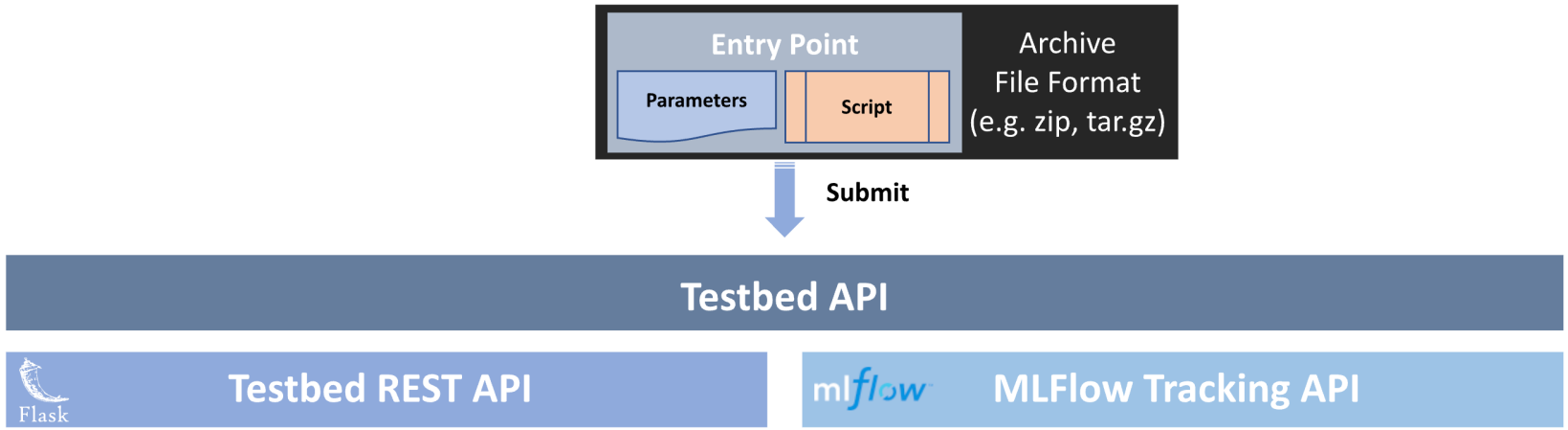
'train' entry point의 flow 정의

Entry point


Experiment
Worker
docker
Entry Point
mlflow
Task
PREFECT
Task
PREFECT
Task
PREFECT
Worker
docker
Entry Point
mlflow
Task
PREFECT
Task
PREFECT
Task
PREFECT
Worker
docker
Entry Point
mlflow
Task
PREFECT
Task
PREFECT
Task
PREFECT

Job submit

다양한 attack, defense, model architecture 등을 entry point로 정의하여 모듈화



Testbed API를 통해 웹브라우저에서 모니터링할 수 있다.

(MLFlow 대시보드에 접근해서 확인하는 것도 가능)

Use case

Newcomer

테스트베드에 대한 경험이 없는 사용자
제공된 데모의 매개변수를 변경하여 기존 실험에 약간의 변형을 만들 수 있음

Analyst

더 다양한 시나리오에서 분석하고자 하는 사용자
내장 플러그인을 활용하여 사용자화 가능
다양한 attack에 대해 제품을 테스트하여 어떤 유형의 attack이 가장 해로운지 이해할 수 있음

Researcher

새로운 metric, 알고리즘 및 분석 기술을 사용하여 실험을 하고자 하는 사용자
자체 플러그인을 구현하여 사용자화 가능
광범위한 attack에 대해 평가할 수 있음

Developer

배포에 기여하여 테스트베드의 핵심 기능을 확장할 수 있는 사용자

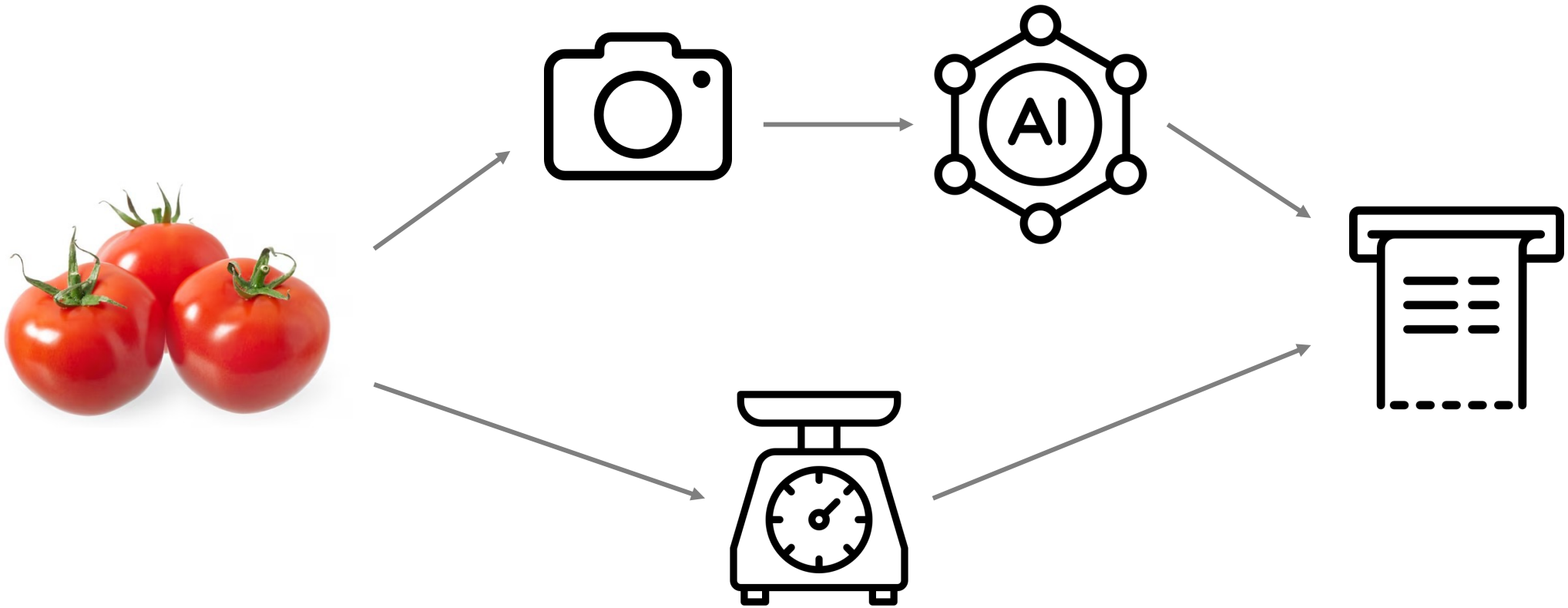
Use case

셀프 계산대에 사용할 이미지 기반 물품 식별 모델을 구매할 의향이 있는 CTO

<리스크 평가 과정>

1. 태스크 파악
2. AI가 필요한가?
3. 어떤 attack이 있을 수 있는가?
4. 최우선으로 고려해야 할 리스크는 무엇인가?
5. 이를 평가하기 위한 지표로 어떤 것이 있는가?
6. 실험 구성 및 결과 분석

Use case



Use case

training architecture

- Vendor 1 model (VGG16)
- Vendor 2 model (defended VGG16)

data augmentation

- adversarial training

inference pre-processing

- None

dataset

- Fruits360

attack on trained model

- patch

metric

- clean accuracy
- adversarial accuracy
- cost per misclassification
- ….

Use case

training architecture

Vendor 1

metric

attack

Use case

training architecture

Vendor 2

metric

…
Peach
Banana
Tomato
Apple

attack

…
Peach
Banana
Tomato
Apple

Demo

https://github.com/usnistgov/dioptra/blob/more-readme-updates/examples/tensorflow-adversarial-patches/demo-fruits360-patches.ipynb

```
response_vgg16_train = restapi_client.submit_job(
    workflows_file=WORKFLOWS_TAR_GZ,
    experiment_name=EXPERIMENT_NAME,
    entry_point="train",
    entry_point_kwargs=" ".join([
        "-P batch_size=20",
        f"-P register_model_name={EXPERIMENT_NAME}_vgg16",
        "-P image_size=224,224,3",
        "-P model_architecture=vgg16",
        f"-P data_dir_training={DATASET_DIR}/training",
        f"-P data_dir_testing={DATASET_DIR}/testing",
    ]),
    queue="tensorflow_gpu",
    timeout="1h",
)

print("Training job for VGG16 neural network submitted")
print("")
pprint.pprint(response_vgg16_train)
```

```
response_patches_adv_training = restapi_client.submit_job(
    workflows_file=WORKFLOWS_TAR_GZ,
    experiment_name=EXPERIMENT_NAME,
    entry_point="train_on_Fruits360_patched",
    entry_point_kwargs=" ".join(
        [
            f"-P dataset_run_id_testing={response_deploy_vgg16_patches_testing['mlflowRunId']}",
            f"-P dataset_run_id_training={response_deploy_vgg16_patches_training['mlflowRunId']}",
            "-P batch_size=256",
            f"-P register_model_name={EXPERIMENT_NAME}_adversarial_patch_vgg16",
            "-P image_size=224,224,3",
            "-P epochs=10",
            f"-P data_dir_testing={DATASET_DIR}/testing",
        ]
    ),
    queue="tensorflow_gpu",
    depends_on=response_deploy_vgg16_patches_training["jobId"],
)

print("Patch adversarial training (VGG16 architecture) job submitted")
print("")
pprint.pprint(response_patches_adv_training)
print("")

response_patches_adv_training = get_run_id(response_patches_adv_training)
```

AIF360

AIF360

ML model의 bias를 식별하고 완화하는 데 도움을 주는 라이브러리

Supported bias mitigation algorithms

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al., 2018)
- Rich Subgroup Fairness (Kearns, Neel, Roth, Wu, 2018)
- Exponentiated Gradient Reduction (Agarwal et al., 2018)
- Grid Search Reduction (Agarwal et al., 2018, Agarwal et al., 2019)
- Fair Data Adaptation (Plečko and Meinshausen, 2020, Plečko et al., 2021)
- Sensitive Set Invariance/Sensitive Subspace Robustness (Yurochkin and Sun, 2020, Yurochkin et al., 2019)

Supported fairness metrics

- Comprehensive set of group fairness metrics derived from selection rates and error rates including rich subgroup fairness
- Comprehensive set of sample distortion metrics
- Generalized Entropy Index (Speicher et al., 2018)
- Differential Fairness and Bias Amplification (Foulds et al., 2018)
- Bias Scan with Multi-Dimensional Subset Scan (Zhang, Neill, 2017)

- 패키지로 구현되어 있으며 Python, R에서 사용 가능
- 다양한 fairness metric과 bias mitigation 알고리즘들을 포함

Demo

- German Credit 사용 → 모델은 각 개인의 신용을 평가 (binary classification)
- 나이(25세 미만/이상) 를 sensitive attribute으로 사용
- training dataset 내에 있는 bias를 식별하고 완화하고자 함

Step1. Compute fairness metric on original training set

```
metric_orig_train = BinaryLabelDatasetMetric(dataset_orig_train,
                                             unprivileged_groups=unprivileged_groups,
                                             privileged_groups=privileged_groups)
display(Markdown("#### Original training dataset"))
print("Difference in mean outcomes between unprivileged and privileged groups = %f" % metric_orig_train.mean_difference())
```

Original training dataset

```
Difference in mean outcomes between unprivileged and privileged groups = -0.169905
```

Step2. Mitigate bias by transforming the original dataset

```
RW = Reweighing(unprivileged_groups=unprivileged_groups,
                privileged_groups=privileged_groups)
dataset_transf_train = RW.fit_transform(dataset_orig_train)
```

Demo

- German Credit 사용 → 모델은 각 개인의 신용을 평가 (binary classification)
- 나이(25세 미만/이상) 를 sensitive attribute으로 사용
- training dataset 내에 있는 bias를 식별하고 완화하고자 함

Step3. Compute fairness metric on transformed dataset

```
metric_transf_train = BinaryLabelDatasetMetric(dataset_transf_train,
                                               unprivileged_groups=unprivileged_groups,
                                               privileged_groups=privileged_groups)
display(Markdown("#### Transformed training dataset"))
print("Difference in mean outcomes between unprivileged and privileged groups = %f" % metric_transf_train.mean_difference())
```

Transformed training dataset

```
Difference in mean outcomes between unprivileged and privileged groups = 0.000000
```

Demo

https://github.com/Trusted-AI/AIF360/blob/main/examples/README.md

AI Verify

AI Verify

AI VERIFY FOUNDATION

AI 거버넌스 검증 프레임워크 및 툴킷

AI 거버넌스 : 인공지능 시스템의 개발, 배포, 운영 과정에서 윤리적·법적·기술적 기준을 정하고 관리하는 체계

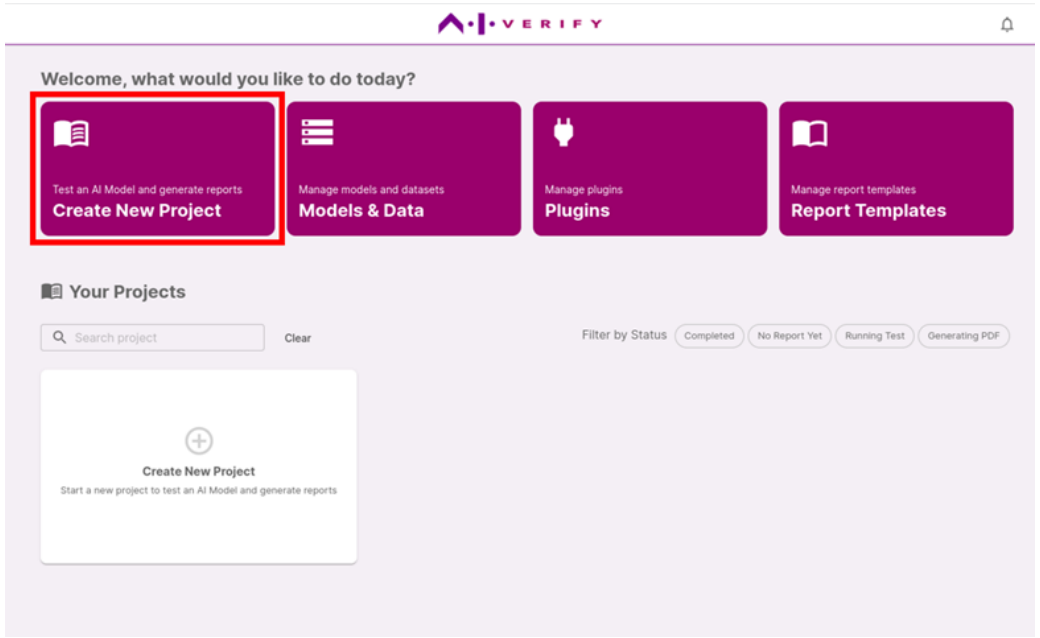
AI가 안전하고 공정하게 사용되도록 보장하는 것이 목표

Test Framework

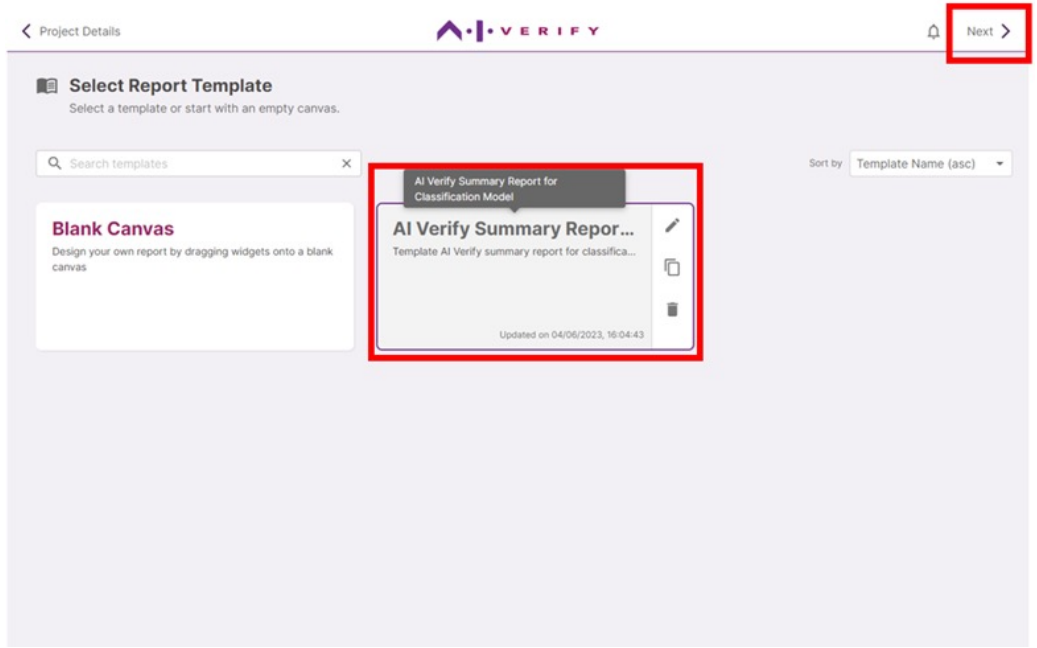
TRANSPARENCY ON THE USE OF AI AND AI SYSTEMS Ensuring that individuals are aware and can make informed decisions	UNDERSTANDING HOW AI MODELS REACH DECISION Ensuring AI operation/results are explainable, accurate and consistent	SAFETY & RESILIENCE OF AI SYSTEM Ensuring AI system is reliable and will not 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
TRANSPARENCY Appropriate information is provided to individuals impacted by AI system	**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

- 위 11개의 AI 윤리원칙을 검증하는 약 90개의 질문들로 구성
- fairness, robustness, explainability에 대해서는 추가적인 technical test 진행

Toolkit

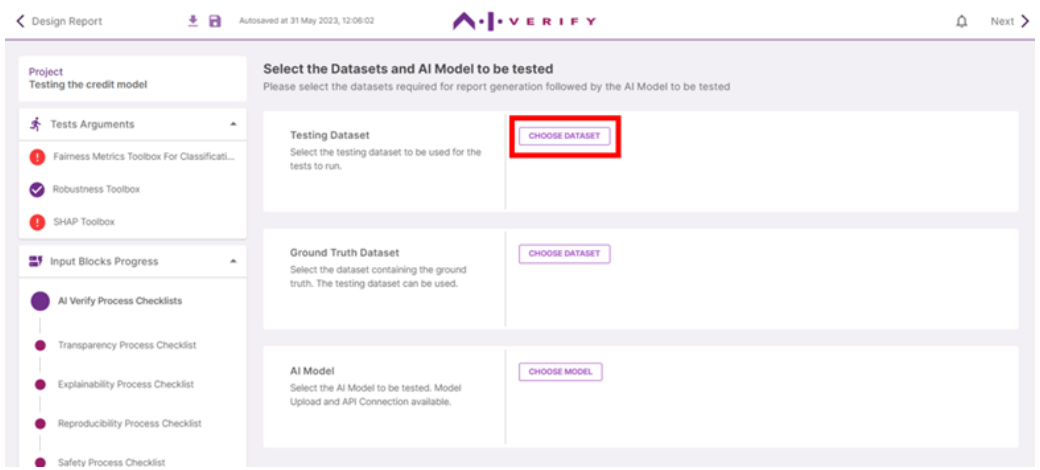


1. 새 프로젝트 생성

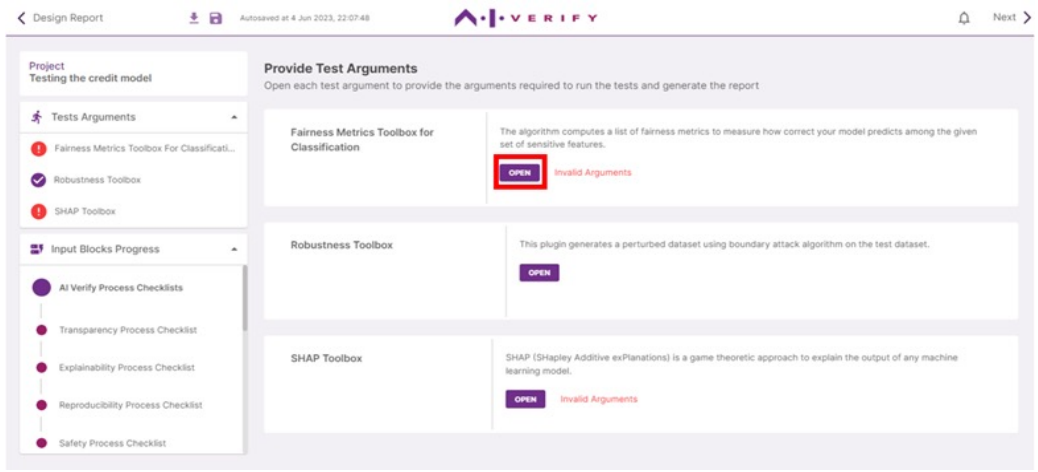


2. 보고서 템플릿 선택

Toolkit

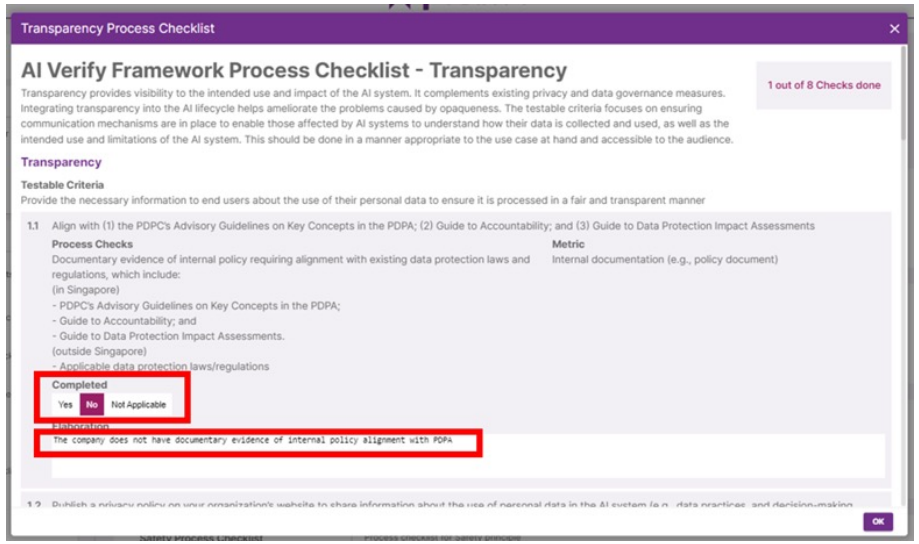


3. 데이터셋과 모델 업로드

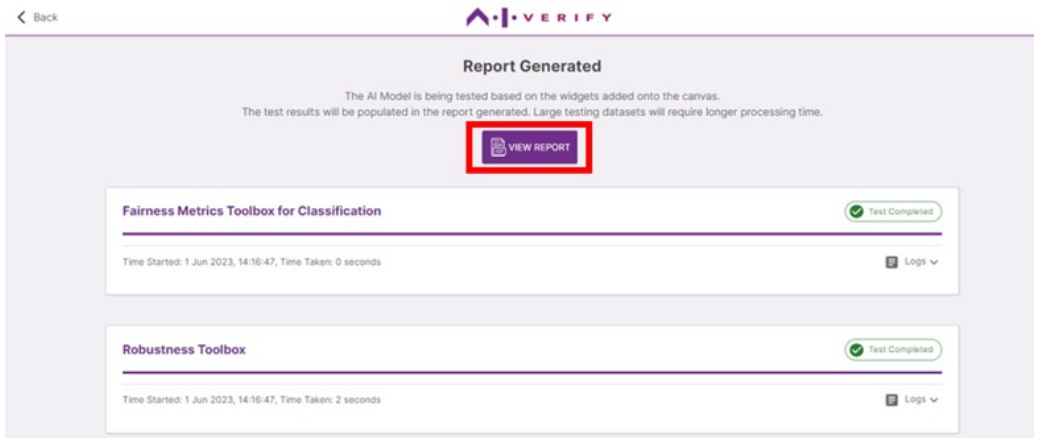


4. technical test 관련 config

Toolkit



5. 프레임워크 체크리스트 응답



6. 보고서 생성

Report

SUMMARY REPORT

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

AI GOVERNANCE TESTING FRAMEWORK AND TOOLKIT

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

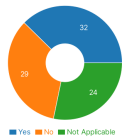
TESTS SUCCESSFULLY RUN	TESTS FAILED TO COMPLETE	TESTS SKIPPED BY USER
3 / 3	0 / 3	0 / 3

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].

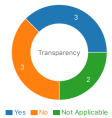
[1]The company should periodically review that the reason(s) for not implementing the process checks remains valid and aligned with company's values, objectives and regulatory requirements.
[2]If the operating environment or model changes, company should assess whether these process checks would become 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.



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

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.

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.

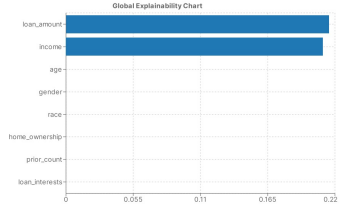
Company did not implement the following testable criteria fully:

- 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 predictions, 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.

ANNEX A
PROCESS CHECKLISTS

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.		

ANNEX B
TECHNICAL TESTS

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.