AI Explainability WG

May 2024

May 2024 updates

- TrustyAI core / service
  - 0.14.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.14.0
  - quay.io/trustyai/trustyai-service:v0.14.0
- TrustyAI operator
  - 1.20.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.20.0
  - quay.io/trustyai/trustyai-service-operator:v1.20.0
- TrustyAI KServe explainer
  - 0.1.0 release
  - https://github.com/trustyai-explainability/trustyai-kserve-explainer/releases/tag/v0.1.0
  - quay.io/trustyai/trustyai-kserve-explainer:v0.1.0

What’s new?

TrustyAI - What's new?

- **TrustyAI core/service 0.14.0**
  - *Explainers*
    - Add KServe v1 HTTP prediction provider [#544]
    - Add KServe v1 HTTP parser optional output size [#552]
    - Consistent default naming for feature parsing [#553]
  - *Feature flags*
    - Separate download+upload endpoints, add feature flag to /download endpoint [#549]
    - Add enable/disable flags to feature endpoints [#546]
  - *Fixes*
    - Resolve import errors in integration test module [#550]
    - Add support for listing metric request subsets [#555]
    - Add more detail to invalid feature/output name errors [#556]
  - *Security*
    - Upgrade Quarkus from 3.2.11.Final to 3.2.12.Final (CVE-2024-2700) [#547]

[1]: https://github.com/trustyai-explainability/trustyai-explainability-python-examples/blob/main/examples/DataDrift.ipynb

TrustyAI - What's new?

- **TrustyAI operator 1.20.0**
  - Add ODH and RHOAI overlays [#237]
  - Fixes
    - Change route termination to reencrypt [#234]
    - Remove requeueing from finalizer [#224]
  - Security
    - Upgrade golang.org/x/net 0.18.0 -> 0.23.0 [#232]
- **TrustyAI Kserve explainer 0.1.0**
  - Add more robust CLI parsing with picocli [#1]
  - Add quay image build [#3]
  - Add missing Maven from image builder [#4]
  - Add missing explainer builder [#5]
  - Add LIME parameters [#6]

[1]: https://github.com/trustyai-explainability/trustyai-explainability-python-examples/blob/main/examples/DataDrift.ipynb

Current work

TrustyAI - current work

- **TrustyAI core/service**
  - Support for database storage
  - New Feature type: Tensor
- **TrustyAI operator**
  - TLS-enabled Kubernetes `Service` for payload consumer
  - KServe readiness
- **TrustyAI KServe explainer**
  - SHAP support

KServe explainer demo

Predictor only

```
apiVersion: "serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "explainer-test"
  annotations:
       sidecar.istio.io/inject: "true"
       sidecar.istio.io/rewriteAppHTTPProbers: "true"
       serving.knative.openshift.io/enablePassthrough:
"true"
spec:
  predictor:
       model:
        modelFormat:
        name: sklearn
       protocolVersion: v2
       runtime: kserve-sklearnserver
       storageUri:
https://github.com/ruivieira/model-collection/raw/main/c
redit-score/model.joblib
```

Predictor and explainer

```
apiVersion: "serving.kserve.io/v1beta1"
kind: "InferenceService"
metadata:
  name: "explainer-test"
  annotations:
       sidecar.istio.io/inject: "true"
       sidecar.istio.io/rewriteAppHTTPProbers: "true"
       serving.knative.openshift.io/enablePassthrough: "true"
spec:
  predictor:
       model:
        modelFormat:
        name: sklearn
       protocolVersion: v2
       runtime: kserve-sklearnserver
       storageUri:
https://github.com/ruivieira/model-collection/raw/main/credit-score/
model.joblib
  explainer:
       containers:
       name: explainer
       image: quay.io/trustyai/trustyai-kserve-explainer:latest
```

KServe explainer demo

- **Tutorial**
  - https://trustyai-explainability.github.io/trustyai-site/main/saliency-explanations-with-kserve.html
- **Architecture**
  - https://trustyai-explainability.github.io/trustyai-site/main/component-kserve-explainer.html

Detoxification / SFT

- Detoxifying LLMs during training
  - Using TrustyAI Detoxify to rephrases toxic text
  - HuggingFace's SFT Trainer (Supervised Fine-Tuning Trainer).
- **Demo repository**
  - https://github.com/trustyai-explainability/trustyai-detoxify-sft

Roadmap

TrustyAI 2024 roadmap

- **December 2023 - March 2024 (proposal / discussion)**
  - Out-of-distribution (OOD) metrics completion
    - Finalise OOD metrics, namely data uploading aspect
      - Provide data connection for data upload
  - Explainability service endpoints completion
    - Formalise explainability payloads schema
    - Refining handling of synthetic payloads to avoid interference with metrics
    - Handle potential large computational times in a service setting
  - Detoxification at the library and service level
    - Integrate detoxification with Python TrustyAI (Jupyter main target)
    - Token scoring at the service level
  - Database backend
    - Address scalability
    - Replace PVC with DB?
  - Expand supported types (*eg* image data)
    - Metrics and explainability for non-tabular data
  - Model drift/data drift/anomaly detection
  - Improve handling of unsupported model serving runtimes

Legend

Not started

In progress

Completed

TrustyAI 2024 roadmap

- **March 2024 - May 2024 (proposal / discussion)**
  - KServe explainer integration
  - Detoxification fine-tuning

Legend
Not started
In progress
Completed