clearbox AI

humans && machines

Technical whitepaper

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

We have created the **Clearbox AI Control Room**, an MLops solution to put existing Machine Learning (ML) models into production in accordance with the principles of **Trustworthy AI** [1]. Our solution is model agnostic and can be used to perform a model assessment and to create a production container for model deployment.

The **model assessment** step can help model owners to identify robustness issues, potential undesired behaviour, and explain errors and uncertainties regarding the model predictions.

The output of such an assessment is:

- A report containing an overview of the most important **validation** metrics, an analysis of the plausible **causes of error** and a **calibration** of the model output;
- An inference model converted into a production ready architecture;
- The possibility to **generate additional training** data to further improve the model performance.

[1] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Once a model has been validated, our Control Room can be used as an **MLOps** solution for deploying the validated model. The model is served by creating an **augmented API** for inference and a database for continuous monitoring. The packaging process is performed using open source and state of the art libraries such as ONNX and FastAPI which guarantee optimal computational performance.

The augmented API is generated by using our proprietary **eXplainable AI** libraries. This API is designed to:

- Enrich single queries with **local explanations** of the output and to enhance the interpretability of the model decisions. Explanations are presented in the form of historical examples, decision rules and counterfactual examples;
- Create an intuitive **feedback** mechanism to monitor the model performance over time;
- Perform robust **uncertainty quantification** of the model output to increase trust in each recommendation;
- Perform **anomaly detection** of live data and report **adversarial attacks** [2].

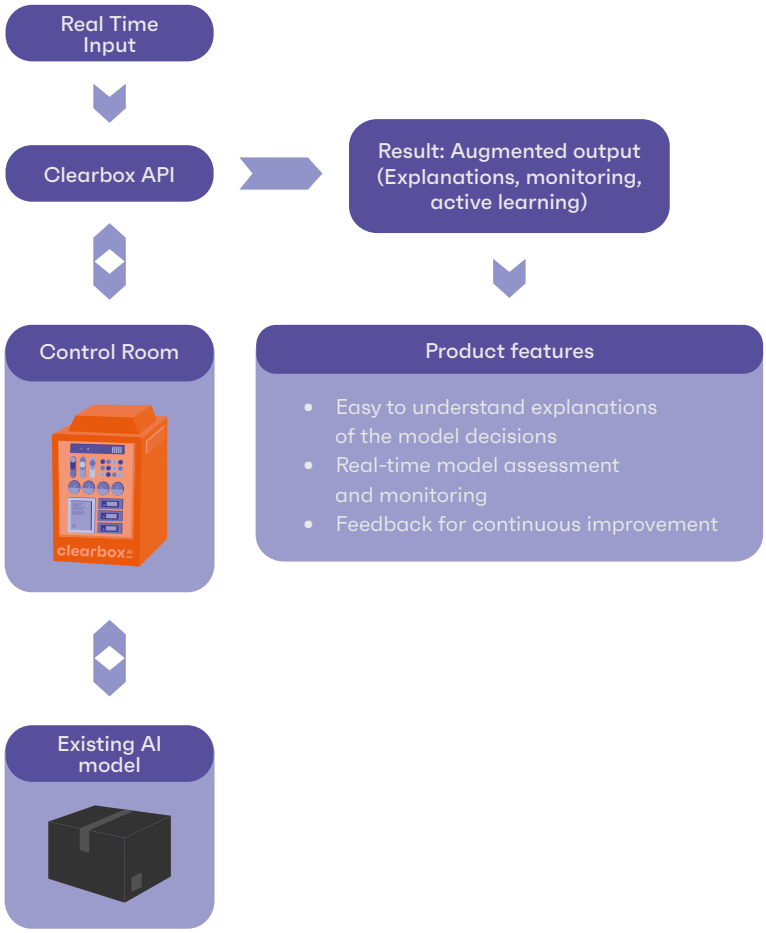
Our proprietary methods make use of a **generative model** to create meaningful perturbations that reduce the search space needed to find local explanations. The computational time required to build a single explanation is therefore an order of magnitude lower compared to common techniques such as SHAP or ANCHORS (a patent application is pending on our particular approach).

Another advantage of our solution is the possibility to aid data scientists with the generation of new training data for continuous model improvement. Our engine is able to select a subset of the past model queries for labeling, based on a combination of **active learning** and design of experiments. Our objective is to provide data scientists with a subset which is as small as possible as well as representative of real life data, thus minimizing labeling costs.

[2] https://openai.com/blog/adversarial-example-research/

We designed the API following the principles of **Human Centered Artificial Intelligence** [3]. Our aim is to improve the interaction between model decisions and end users by letting the latter 'converse' with the model through the API itself. This means that the API will not only be used to do model serving but it will also allow users to perform counterfactual reasoning and uncertainty analysis.

Clearbox AI Control Room:



[3] https://hai.stanford.edu/

humans && machines

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai