Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Demonstrations Track

Certifai: A Toolkit for Building Trust in AI Systems

Jette Henderson[1*], **Shubham Sharma**[2], **Alan Gee**[2], **Valeri Alexiev**[1], **Steve Draper**[1], **Carlos Marin**[1], **Yessel Hinojosa**[1], **Christine Draper**[1], **Michael Perng**[1], **Luis Aguirre**[1], **Michael Li**[2], **Sara Rouhani**[3], **Shorya Consul**[2], **Susan Michalski**[1], **Akarsh Prasad**[1], **Mayank Chutani**[1], **Aditya Kumar**[1], **Shahzad Alam**[1], **Prajna Kandarpa**[1], **Binnu Jesudasan**[1], **Colton Lee**[1], **Michael Criscolo**[1], **Sinead Williamson**[1,2], **Matt Sanchez**[1] and **Joydeep Ghosh**[1,2]
[1]CognitiveScale
[2]The University of Texas, Austin
[3]The University of Texas, Dallas
[*]jhenderson@cognitivescale.com

Abstract

As more companies and governments build and use machine learning models to automate decisions, there is an ever-growing need to monitor and evaluate these models' behavior once they are deployed. Our team at CognitiveScale has developed a toolkit called Cortex Certifai to answer this need. Cortex Certifai is a framework that assesses aspects of robustness, fairness, and interpretability of any classification or regression model trained on tabular data, without requiring access to its internal workings. Additionally, Cortex Certifai allows users to compare models along these different axes and only requires 1) query access to the model and 2) an "evaluation" dataset. At its foundation, Cortex Certifai generates counterfactual explanations, which are synthetic data points close to input data points but differing in terms of model prediction. The tool then harnesses characteristics of these counterfactual explanations to analyze different aspects of the supplied model and delivers evaluations relevant to a variety of different stakeholders (e.g., model developers, risk analysts, compliance officers). Cortex Certifai can be configured and executed using a command-line interface (CLI), within jupyter notebooks, or on the cloud, and the results are recorded in JSON files and can be visualized in an interactive console. Using these reports, stakeholders can understand, monitor, and build trust in their AI systems. In this paper, we provide a brief overview of a demonstration of Cortex Certifai's capabilities.

1 Introduction

As the world around us becomes more automated through the use of machine learning models to make decisions, it is becoming more necessary for model developers to ensure that the decisions their models are making can be trusted. Key aspects of building trust in models are ensuring that they are robust, fair, and understandable. Specifically, a model developer should know how easy it is to fool their model with corrupted or adversarial data, whether or not their model is fair to people with respect to certain attributes (e.g., legally protected groups), and which features are key indicators of a given model decision.

Extending into the deployment stage, a model developer and stakeholders within their institution may want to monitor the model over time. Additionally, a person impacted by a model's decision may want to understand how they can change their features in order to receive a different decision (e.g., a more favorable decision if their input features resulted in an unfavorable decision). Our demonstration will showcase the capabilities of Cortex Certifai, a product built to address these key concerns.

2 Cortex Certifai Background and Method

Cortex Certifai productionizes and improves upon the algorithm proposed in the recent conference article by [Sharma *et al.*, 2020]. Additionally, it expands upon the method to deliver scores along the axes of robustness, fairness, explainability, and performance.[1] The toolkit scans user-supplied classification (binary or multiclass) or regression models that accept tabular data as an input to the model.[2]

Briefly, Cortex Certifai's analysis is based on the generation of "counterfactual explanations". A counterfactual explanation[3] for a given data point is a minimally perturbed version of it, which is distinct enough to receive a different prediction than the original data point. Generating and using counterfactual explanations is a growing area of research since they offer an intuitive way for users to interact with model decisions (e.g., [Wachter *et al.*, 2017;

[*]Contact Author

[1]Sign up at https://www.cognitivescale.com/certifai/ to access the toolkit.

[2]Although the algorithm presented in [Sharma *et al.*, 2020] works for a variety of different data types, the Cortex Certifai toolkit only accepts tabular data at this time.

[3]A "counterfactual explanation" is different from a "counterfactual" from the causality literature [Pearl, 2000]. A good counterfactual explanation does not assume anything causal about the world (see page 9 of [Wachter *et al.*, 2017]).

Figure 1: Part of a sample report displayed in the Cortex Certifai console. The report compares four models, each represented by a different color bar along the axes of a performance metric (accuracy in this case), fairness, explainability, and robustness. The four measures are then combined for an overall trust score called the ATX. Users can click on any of the charts to get a detailed view of that axis.

Ustun *et al.*, 2019; Russell, 2019; Mothilal *et al.*, 2020]). However, [Sharma *et al.*, 2020] were the first to use these explanations to evaluate model characteristics across these three dimensions.

Given a model and a dataset (the "evaluation dataset") for which the model can make predictions, Cortex Certifai uses a genetic algorithm to generate counterfactual explanations. Using these explanations and their counterparts in the evaluation dataset, Cortex Certifai calculates robustness, fairness, and interpretability scores. Based on these scores and a performance score (e.g., accuracy), Cortex Certifai provides an overall trust score for the given model called the *ATX score*, which is a customizable, weighted average of the other scores. Additionally, Cortex Certifai can generate counterfactual explanations for the user to inspect if they provide an "explanation dataset."[4]

At a high level, the robustness score measures how resilient a model's predictions are to perturbed data inputs. The explainability score captures the complexity of the counterfactual explanations. Intuitively, an explanation that requires fewer changes to an input is more explainable to a user than one that requires all features to be changed. A fairness score can be calculated if there is a notion of a preferred model prediction (e.g., a loan being granted rather than denied). The fairness score evaluates how difficult it is for groups within a protected attribute (e.g., gender, age) to gain the preferred prediction from the model. Cortex Certifai does not require that the model use the protected attribute as a feature. The user can identify the attribute within the evaluation dataset or provide attribute information for each of the observations.

3 Workflow

The intended user of Cortex Certifai is a model developer who would like to compare a set of models in terms of performance and trust factors. The model developer may be trying to choose between models for a given task before deployment, or they may be interested in monitoring a deployed model over time. After initiating the scan, the model developer can make the results available to other stakeholders (e.g., risk management officers, individuals impacted by model decisions). The model developer can run a Cortex Certifai scan (i.e., an analysis) in a jupyter notebook environment or in a command-line interface (CLI) locally or in the cloud.

First, the model developer supplies a tabular evaluation dataset in csv format, model(s), and an optional explanation dataset. The models must be able to accept the evaluation dataset and therefore must include any data transformations necessary to make predictions. However, the model does not need to be in python. It only needs to be accessible via HTTP.

To run a scan in a jupyter notebook environment, the user initializes a Scanner object with parameters to be used in the scan. The user then has the choice to run the scan in a notebook environment, or they can save the Scanner definition as a YAML scan definition file that is then run in a CLI or on the cloud. Alternatively, a user can define the scan in a YAML file using a text editor and can then execute it in a CLI.

Each scan, regardless of where it is executed, generates one or more reports that are stored in JSON files. To visualize the report results, the user can run a console from the CLI. Figure 1 shows an example of part of the displays shown in the console. In the console, the user can interact with each score and compare the models along each score. Additionally, the user can explore the counterfactual explanations for a supplied set of observations in the explanation dataset.

To integrate Cortex Certifai into a pipeline delivering models into production, the model developer gives the scan definition YAML to an engineer. The engineer packages the model as a web service and sets up the scanner to run locally or in the cloud.

[4]The counterfactual explanations can be customized to the user's needs. The user can specify which features can change and how much they can change.

4 Contributions and Conclusion

Cortex Certifai is an innovative, interactive product to help organizations build trust in their models by analyzing robustness, fairness, and explainability. Unlike other available toolkits, Cortex Certifai only needs to be able to query the model with an input point and receive predictions and does not require access to the model internals (e.g., gradients). Since Cortex Certifai only requires queries, we have built the toolkit to be flexible enough to interact with different types of models developed in a variety of languages. Cortex Certifai can be integrated into existing machine learning pipelines allowing stakeholders to monitor models over time. In the future, we plan to continue to improve Cortex Certifai's capabilities to change and evolve with the needs of stakeholders.

References

[Mothilal *et al.*, 2020] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[Pearl, 2000] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[Russell, 2019] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM, 2019.

[Sharma *et al.*, 2020] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *3rd AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172, 2020.

[Ustun *et al.*, 2019] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM, 2019.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2), 2017.