Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant

Kacper Sokol and **Peter Flach**
Department of Computer Science, University of Bristol
K.Sokol@bristol.ac.uk, Peter.Flach@bristol.ac.uk

Abstract

The prevalence of automated decision making, influencing important aspects of our lives – e.g., school admission, job market, insurance and banking – has resulted in increasing pressure from society and regulators to make this process more transparent and ensure its explainability, accountability and fairness. We demonstrate a prototype voice-enabled device, called Glass-Box, which users can question to understand automated decisions and identify the underlying model's biases and errors. Our system explains algorithmic predictions with class-contrastive counterfactual statements (e.g., "Had a number of conditions been different:... the prediction would change..."), which show a difference in a particular scenario that causes an algorithm to "change its mind". Such explanations do not require any prior technical knowledge to understand, hence are suitable for a lay audience, who interact with the system in a natural way – through an interactive dialogue. We demonstrate the capabilities of the device by allowing users to impersonate a loan applicant who can question the system to understand the automated decision that he received.

1 Introduction

In this paper we describe our demonstration of Glass-Box – a novel system that explains predictions of a Machine Learning model with class-contrastive counterfactual statements. These statements are provided to an explainee as answers to his or her requests in a chat- or voice-based interactive dialogue delivered by a virtual assistant. This interaction mode gives the process a natural feel suitable for a lay audience. Explanations provided through counterfactual statements are easy to understand even for individuals lacking technical knowledge and their explanatory powers are strongly grounded in social science research [Miller, 2017; Miller *et al.*, 2017]. They can be used to identify a model's biases and errors, and provide actionable prediction explanations, among others [Wachter *et al.*, 2017; Kusner *et al.*, 2017]. Furthermore, such explanations are parsimonious and, unlike many other approaches, our explanatory dialogues are interactive, hence allow the user to guide the explanation to suit his or her needs instead of being served a one-size-fits-all template.

In our demonstration we show how such a system can be built by combining recent advances in voice-enabled virtual digital assistants and counterfactual explainability approaches. To the best of our knowledge, our prototype is the first to utilise interaction through a (voice-enabled) dialogue to explain predictions of a Machine Learning model. The underlying counterfactual statements giving the explanatory power have, nonetheless, already found applications in Machine Learning [Wachter *et al.*, 2017; Tolomei *et al.*, 2017].

Our demo provides a hands-on experience, where users impersonate one of 10 loan applicants (to avoid a lengthy process of submitting personal details) and are able to interrogate and challenge an automated decision. The system then helps the users to understand the underlying decision process through a dialogue, such as the one in figure 1. Allowing users to interact with our system during the demonstration session will give us the opportunity to collect invaluable feedback from the artificial intelligence research community.

Ribeiro *et al.* and Smilkov *et al.* explain algorithmic predictions with interactive graphical interfaces and visualisations. We argue that such approaches require prior experience with this type of technology and domain-specific background knowledge to be fully appreciated. Moreover, the number of dimensions that can be visualised is limited due to the nature of the human visual system. The curse of dimensionality also plays a vital role when dealing with high-dimensional datasets. In such cases, exemplar-based explanations, showing similarities between data points are more natural for a lay audience [Kim *et al.*, 2014].

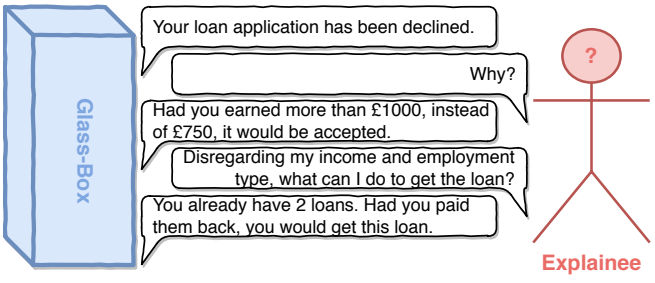

Figure 1: Example of an explanatory dialogue.

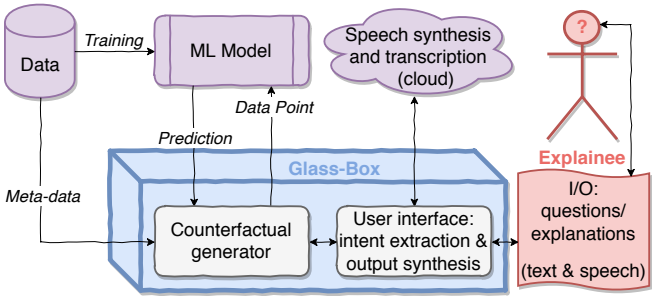

Figure 2: Glass-Box design.

2 Methodology

Counterfactuals, the backbone of our Glass-Box system, are generated with our novel approach (currently under review), specifically designed for decision trees, however it can be generalised for other logical models such as rule sets and lists. We extract logical conditions from the tree's splits and use them, as boolean meta-features, to describe every root-to-leaf path in the tree. When a new data point is classified, we compare the set of meta-features of its leaf with a set of meta-features for every other leaf of the opposite class. This allows us to find transformations of this data point under which the classification outcome changes and present them to the user as counterfactual statements. These can be efficiently retrieved by computing a leaf-to-leaf distance between their meta-features. We use a distance metric derived from the L1 distance, which favours sparsity – a desirable property when looking for the shortest possible counterfactual statement.

3 Implementation

Our prototype is based on Google's *DIY AI Voice Kit*[1], which provides a customisable hardware and software platform for development of voice interface-enabled systems. To improve the user interaction we have added a QR code scanner used as an alternative to voice input to enable quick loading of a data point to be classified. The QR codes encode data point features in JSON format and are printed on *profile cards*, which also display the feature values in a human readable format. The profile cards are used during the system demonstration to improve user experience, as described in the introduction.

The software counterpart of the system is written in Python. It uses digital assistant and speech services accessed through a cloud API, which provide it with natural language and speech processing capabilities. Therefore, the device can be interacted with via either voice commands or text-based chat. The modular design of the system – details are presented in figure 2 – allows it to be adapted to any data with human understandable features and any Machine Learning model for which we can efficiently generate counterfactuals.

4 Interacting With the System

The demo system first receives a data point to be classified by scanning a QR code or by asking questions to collect the necessary features. After that, it classifies the data point using the underlying Machine Learning model and outputs its decision. Then, the user can challenge the decision and request:

- a counterfactual explanation – the system returns the shortest possible class-contrastive counterfactual;
- a (partially-)fixed counterfactual explanation – the system returns a class-contrastive counterfactual that does not use a specified feature (and value) as its condition;
- a list of important factors – the system enumerates all the possible feature space perturbations, from the shortest to the longest, resulting in the prediction change.

Then, the *user interface* translates the user's question from natural language into a programmatic counterfactual query and the *counterfactual generator* composes an answer. After that, the *user interface* generates a natural language answer; if the interaction is voice-based the system transcribes and synthesises speech using cloud services. The system repeats these steps until the user is satisfied with the explanations received or it cannot generate any more counterfactuals.

5 System Demonstration

This particular demonstration uses a decision tree model trained with Scikit-learn[2] on the UCI German credit dataset[3], with its features annotated to improve their readability. The demonstration mimics a user trying to understand why a mock credit application has been rejected or accepted by interacting with our system. In this scenario the user can impersonate one of 10 loan applicants, each with particular personal characteristics (feature values). The interaction is initiated by the user selecting and scanning a loan application card. If the user does not agree with one of the profile characteristics, it can be altered through a dialogue with the system before asking for the result of the loan application. Then the user can interrogate the underlying Machine Learning model by asking one of the three question types outlined in section 4.

We will use the demonstration opportunity to carry out a user study among the population of IJCAI/ECAI attendees. The user study aims at comparing the quality and informativeness of our counterfactual explanations against just using the underlying decision tree structure to explain the same classification results. The participants will be asked to complete a questionnaire asking about the reasons and important factors of automated decisions identified when using both methods.

6 Concluding Remarks

We will demonstrate a system that explains automated predictions with class-contrastive counterfactual statements through a voice-enabled dialogue. In the future, we plan to extend the system with *narrative generation* capabilities to enable it to produce summaries of the dialogue that explains the rationale behind an automated decision. We also plan to add a display to the system to support the natural language explanations with appropriate plots and figures whenever this adds value.

[1] https://aiyprojects.withgoogle.com/voice

[2] http://scikit-learn.org/

[3] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

References

[Kim *et al.*, 2014] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, 2014.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[Miller *et al.*, 2017] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 36, 2017.

[Miller, 2017] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[Smilkov *et al.*, 2016] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016.

[Tolomei *et al.*, 2017] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 465–474. ACM, 2017.

[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*, (Forthcoming), 2017.