Title

Explainable AI for Assisting Data Scientists

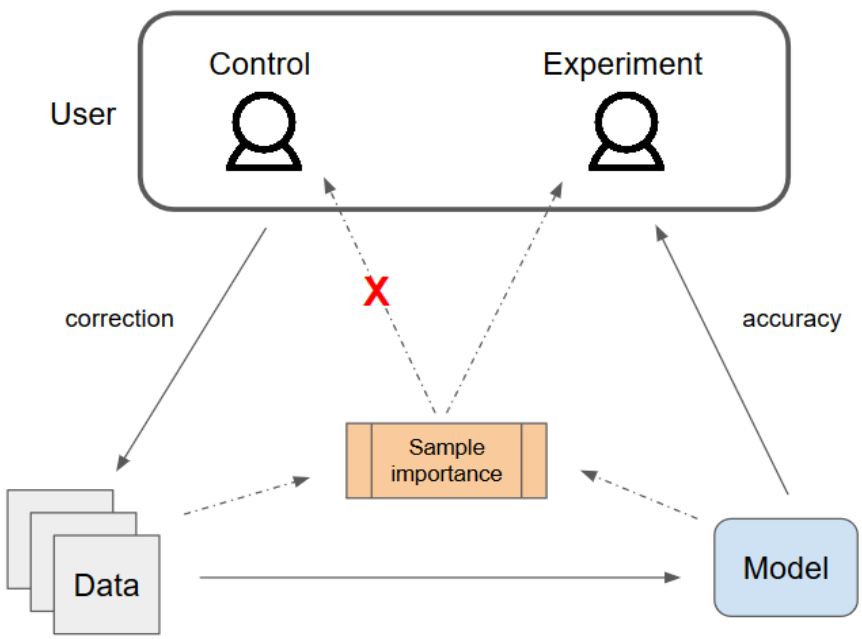
Team Composition

Carnegie Mellon University (CMU), J. Zico Kolter, Principal Investigator, zkolter@andrew.cmu.edu
CMU: Pradeep Ravikumar, Stephen Broomell, David Inouye, Joon Sik Kim, Leqi Liu, Edgar Xi, and George Haff

Problem Statement

CMU plans to give data scientists sample datasets and a model with some simple bugs, and ask them to debug with (experimental group) and without (control group) access to their XAI system (see Figure 1). The machine learning (ML) tasks can vary from supervised learning (e.g., some data is corrupted) to reinforcement learning (RL) (e.g., the rewards that are used are wrong). Users will be asked to debug the data and the model as much as they can, and the effectiveness of the XAI system in debugging ML models will be evaluated on several metrics based on the comparison between the user's performance in each group.

Figure 1. XAI for Assisting Data Scientists



Approach and Technique

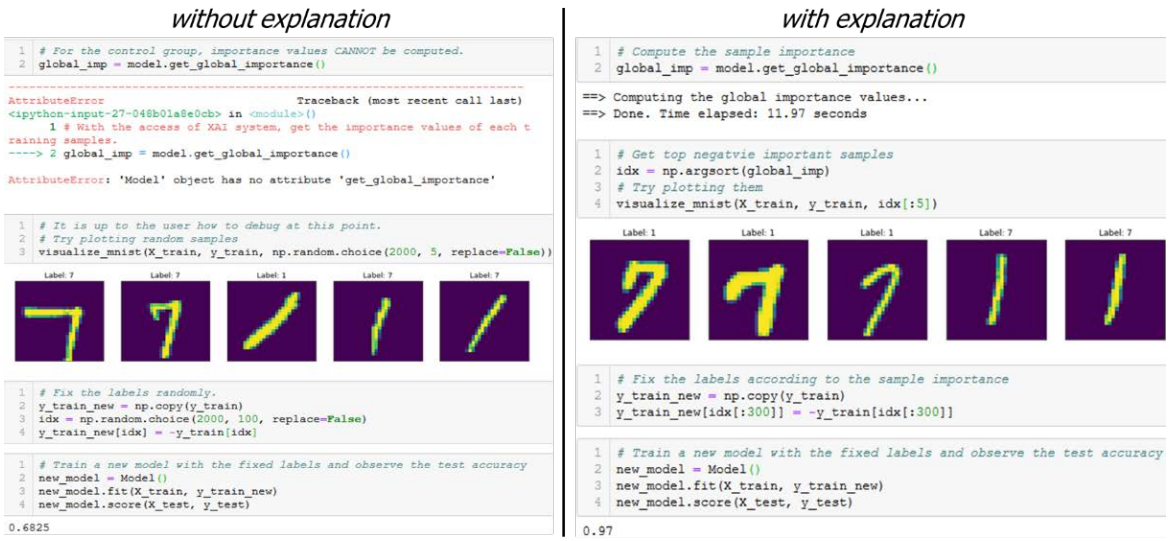
CMU plans to use students who are finishing introductory ML courses taught in the ML department (which number more than a thousand every year) as study participants. The task will be framed as a Data Science Competition. Some participants will be provided an XAI library, some will not. CMU will measure various quantitative metrics, as well as ask for responses to questionnaires for qualitative evaluations. The tasks will be presented to the users in a Jupyter Notebook format assuming that the users are familiar with standard Python libraries.

One example of the XAI system CMU uses is "sample importance" which tells the user what training samples are more/less important than others in influencing the model. Based on these values, users can

Approved for Public Release, Distribution Unlimited

correct the training data so that the produced model has improved accuracy. The control group does not have access to the method that can compute the sample importance values. Each individual gets their own Jupyter Notebook file to work on, and their work will be saved periodically for evaluation purposes. Figure 2 below depicts the user interface. The left side of the figure (without explanations) depicts an example where the users do not have access to the method computing the sample importance values, and therefore must rely on their own debugging scheme. The right side of the figure (with explanations) depicts an example where the users can compute the sample importance values and use them for debugging.

Figure 2. User interface view for sample importance



Key Objectives and Current Status

The study has gone through a small pilot phase and is currently waiting on the recruitment process to proceed with much larger groups. Initial evaluation of pilot data has shown to be promising in multiple ways. Users of the tool expressed overwhelmingly positive impressions of the tool, its usefulness, and confidence in it. Improvements in accuracy and recall of the models produced with the tool versus those without are apparent, despite the limited number of pilot users.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Perceived user confidence, reliability, efficiency, and satisfaction with the XAI tool	Determine value of the tool to users, using Likert scales for objectively comparable results	Overwhelmingly positive
Number of bugs identified	Compute precision and recall	Significant difference not yet observed due to limited sample size
Speed in identifying bugs	Identify tool impact on user time spent on the task	Unable to measure for the pilot study
Model accuracy on validation data	Determine tool impact on developing models that perform well on the data available to the user	Significant difference not yet observed due to limited sample size
Model accuracy on evaluation data	Determine tool impact on developing models that perform well on data not available to the user, demonstrating universality of results	Significant difference not yet observed due to limited sample size

Approved for Public Release, Distribution Unlimited

Test and Evaluation

Each user gets their own Jupyter Notebook file to work on, through which they get access to the data and the model. The data contains a certain portion of the training labels corrupted, therefore making the ML model achieve low test accuracy on the validation data. The user's goal is to discover the training data points that have corrupted labels and fix them to make the model achieve a higher accuracy on the validation data. While validation data is provided to the users to make sure that they know their fix is working, another held-out evaluation data set is prepared that is not available to the users for post-experiment analysis.

Once users finish their task, their fixed labels will be compared against ground truth labels to compute precision and recall for evaluating the number of bugs identified. The final model's performance on the evaluation data, as well as validation data, are measured. The speed in which the task was completed through the timestamp saved in each user's log file is measured. At the end of the task, the users are then asked to fill out a simple questionnaire regarding qualitative aspects (e.g., satisfaction, confidence) of the explanations.

Approved for Public Release, Distribution Unlimited