AI Skunkworks Project

AI-Addin

Project Manager: Ziwei Fan (fan.ziw@husky.neu.edu)

Co-Project Manager: Grace Gu (gu.ru@husky.neu.edu)

Expected Start Date: 20th March 2019

Expected Project Duration: 6 months (For Spring 2019 students who need to submit a project, the deadline of given tasks will be course project deadline)

Abstract

Even if a machine learning model performs well, it is often crucial to understanding how an algorithm makes a certain decision. Trust in the model is enhanced when the logic is exposed. Machine learning models pick up biases from the training data, and exposing the internals of these models reveals their bias.

Model interpretability algorithms can reveal the logic and bias of these models, exposing the reasons in their predictions. Often the reasons for the predictions are as important as the predictions themselves. For example, accurately predicting someone's weight may not be as interesting as the reasons for the weight loss. The reasons provide the guidelines on how to change behavior. Understanding the reasons and logic can provide human-friendly explanations. Human-friendly explanations allow non-experts to understand complicated algorithms.

AIaddin(AI-addin) is artificial intelligence software that decides when and how to automatically apply model interpretability algorithms to any data set that a user uploads for analysis, including: global surrogate models; word embeddings; individual conditional expectation (ICE) plots; k-local interpretable model-agnostic explanations (K-LIME); leave-one-covariance (LOCO); local feature importance; partial dependency plots; random forest feature importance; standardized coefficient importance; visualization of neural network layers; generalized low rank estimators; feature extraction and ranking; accumulated local effects (ALE) plots; and Shapley values.

We evaluate the system qualitatively, asking users if the human-friendly explanations are understandable and make logical sense given their domain expertise.

Deliverables

Phase 1.1: Predict, Interpret in finance: Whether & Why this customer will default on the credit card bill with the visible interpretable plots.

- UCI credit card default data prepared
- Model trained on H2O
- Generate descriptions for model
- Generate the ranked predictions plot

- Generate the reason codes.

Prerequisites

Please contact Project Manager

Milestones

No	Tasks	Due Date
1	Acquire, explore, and prepare UCI credit card default data	03-20-2019
2	Train an H2O GBM classifier	03-25-2019
3	Generate descriptions for a local region with a perturbed sample	03-29-2019
4	Generate reason codes	04-02-2019
5	generate descriptions for a local region with a practical sample	04-09-2019
6	Generate a ranked predictions plot to assess validity of local explanatory model	04-16-2019
7	Generate reason codes using a practical sample	04-20-2019