Responsible Machine Learning
Assignment 2
10 points

May 30, 2021

In Assignment 2, you will work with your group to analyze and explain interpretable machine learning (ML) models following the instructions below. A template has been provided as an example of how to explain and compare a few different interpretable models. For those of you who use Python virtual environments, a basic `requirements.txt` file is also available for the template.

For each section below, you should be trying to think through whether the explanatory results make sense from a domain knowledge perspective, as well as the differences between your models and if those are logical and/or informative.

Please let me know immediately if you find typos or mistakes in this assignment or related materials.

1 Calculate and Plot Global Feature Importance.

Use regression coefficients, Shapley values, or other reputable techniques to calculate global feature importance for your models. The template uses coefficients from the elastic net GLM – extracted in cell 12, Shapley values for the monotonic GBM – calculated in cell 19–20, and local feature scores for the EBM – extracted in cells 27–28, to create global feature importance. (Depending on your package version, `interpret` can calculate these quantities much more directly using the `predict_and_explain()` function.) Plot these values as bar charts, ideally comparing how your models treat input features differently, as in cell 30 of the template.

2 Calculate and Plot Local Feature Importance.

Using approaches similar to those in Section 1, calculate local feature importance for your models at three percentiles of predicted probability. Cell 10 of the template provides a simple function for calculating percentiles, and percentiles for each models' predictions are found in cells 11, 18, and 26.

Calculate local feature importance for the individuals at the 10^{th}, 50^{th}, and 90^{th} percentiles of predicted probability. Local feature importance for these individuals is calculated in cells 13, 21, and 29 of the template. Once you have local feature importance values, plot them to compare how your models treat similar individuals. Cell 31 of the template shows a comparison of local feature importance values across the three percentiles and three models.

3 Calculate and Plot Feature Behavior.

To enable analysis of feature behavior under each model, the template calculates partial dependence for each main effect feature for each model. (You may optionally plot interaction contour or surface plots for any interaction features, likely discovered by EBM.) You may follow this example or use additional techniques,

such as ALE, ICE, or feature scores in EBM. Cell 32 of the template defines a function for partial dependence and cells 33-34 calculate and display the main effect feature behavior across the three models.

4 Submit Code Results.

Your deliverable for this assignment is to update your group's GitHub repository to reflect this explanatory analysis. Global and local feature importance plots for all models and features are worth 3 pts. and feature behavior plots for all models and features are worth 4 pts., for a total of 10 pts.

Your deliverables are due Sunday, June 6th, at 11:59:59 PM ET.

Note that you may also improve Assignment 1 scores throughout the Summer I Session to improve your ranking, your Assignment 1 grade, and your final project grade.