Assignment 7: Explainable Machine Learning with EBM

(60 Points Total)

Data available under Resources>Assignment Data.

In this exercise you will attempt to differentiate two species of rice grains based on geometric characteristics. These data were obtained from the University of California Irvine (UCI) Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/Rice+%28Cammeo+and+Osmancik%29

The name of the dataset is **rice_dataset.csv**. This is the description of the data provided on the website:

"Among the certified rice grown in Turkey, the Osmancik species, which has a large planting area since 1997 and the Cammeo species grown since 2014 have been selected for the study. When looking at the general characteristics of Osmancik species, they have a wide, long, glassy and dull appearance. When looking at the general characteristics of the Cammeo species, they have wide and long, glassy and dull in appearance. A total of 3810 rice grain's images were taken for the two species, processed and feature inferences were made. 7 morphological features were obtained for each grain of rice."

Here are the descriptions of the variables provided on the website:

AREA: *Returns the number of pixels within the boundaries of the rice grain.*
PERIMETER: *Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.*
MAJORAXIS: *The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.*
MINORAXIS: *The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.*
ECCENTRICITY: *It measures how round the ellipse, which has the same moments as the rice grain, is.*
CONVEX_AREA: *Returns the pixel count of the smallest convex shell of the region formed by the rice grain.*
EXTENT: *Returns the ratio of the region formed by the rice grain to the bounding box pixels.*
CLASS: *Cammeo and Osmancik rices*

These data are from the following publication:

Cinar, I. and Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. *International Journal of Intelligent Systems and Applications in Engineering*, vol.7, no.3 (Sep. 2019), pp.188-194.

You will train an Explainable Boosting Machine (EBM) model using the interpretML library to differentiate these rice grain types based on the provided geometric characteristics.

Your Tasks

T1: Read in the data table. Create a new column in which Cammeo species is coded to 1 and the Osmancik species is coded to 0. You will predict using this column as opposed to the original column ("CLASS"). Make sure the new column is defined as a string type as opposed to integer. This process can be completed using Pandas. **(6 Points)**

T2: Count the number of records for each species. How many samples are available for each species? **(6 Points)**

T3: Separate the dependent variable (your new 1 vs. 0 column) to a new data frame. Separate the predictor variables to a new data frame: AREA, PERIMETER, MAJORAXIS, MINORAXIS, ECCENTRICITY, CONVEX_AREA, and EXTENT. **(6 Points)**

T4: Use the **train_test_split()** function from **scikit-learn** to separate the data into training and testing sets. Roughly 33% of the data should be reserved for testing, and you should stratify on your new 1 vs. 0 column. **(6 Points)**

T5: Initialize then train the EBM classifier using the **ExplainableBoostingClassifier()** function from **interpretml**. You do not need to optimize hyperparameters. **(3 Points)**

T6: Obtain the global and local explanations using the **explain_global()** and **explain_local()** methods. **(3 Points)**

T7: Write a paragraph that discusses the following: **(12 Points)**

- Global importance of the predictor variables
- Importance of the predictor variables vs. the interaction terms
- What values (e.g., high, low, middle) were associated with the Cammeo class (code 1) for each of the predictor variables.

T8: Predict the withheld testing data and generate a **confusion matrix** along with **precision**, **recall**, and **f1-score** for each of the two classes. **(6 Points)**

T9: Write a paragraph explaining and discussing the confusion matrix and assessment metrics. **(12 Points)**

Deliverables

- Jupyter Notebook. Use Markdown to generate your short write ups within the Notebook.