BETAS - DATA 515 Project Component Specification

Joel Stremmel, Yiming Liu, Cathy Jia, Monique Bi, Arjun Singh

Software Components:

Binary Score Plot: A wrapper class for visualizing model performance plots for binary classification tasks given model scores and true labels. Includes commonly used performance plots such as the Receiver Operating Characteristic curve. Getter and setter methods support adding reusable plotting options.

Regression Analysis Plot: A wrapper class for visualizing model fit for regression tasks. Allows a user to asses regression model assumptions and judge the quality of the fit to the data use plots such as residuals v.s. fitted values and scatter plots with the fitted model.

Example Notebooks: Jupyter notebooks which will load example data and fit basic models from sklearn to obtain model scores. These notebooks will then step through how to initialize plotting objects, and use plotting methods from the classes above to analyze performance or fit of the example models.

Interactions:

One of our use cases is to allow users to visualize model performance, model fit, or model assumptions with one line of code by using an instance of a plotting class. The above software components interact to accomplish this task by allowing users to instantiate an instance of a plotting class and pass arguments to the object such as model scores and true labels. The plotting object has many methods which, when called, allow users to visualize model performance by looking at plots like the Receiver Operating Characteristic curve. If using the regression model class, the plotting object for that class has methods for analyzing properties like model fit or model assumptions such as constant variance. In this way, all interactions with this software will come by writing code using classes and methods from our package to quickly and efficiently analyze model performance and assumptions.

Preliminary Plan:

Our preliminary plan consists of accomplishing the following tasks in order of priority:

- Identifying example datasets and fitting basic classifiers from sklearn to these datasets to generate example model scores to be compared with true labels

- Creating plotting classes to enable the easy analysis of the various types of model performance we wish to analyze
- Create methods within each class to generate the most critical plots first, such as plots of precision and recall, as well as plots of the Receiver Operating Characteristic curve. We will repeat this for regression analysis, focusing on the most important plots for analyzing regression model performance and assumptions.
- Create and polish example notebooks to demonstrate the functionality of our package
- Make our package pip installable for easy use