BETAS - DATA 515 Functional Specification

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

Who are the users and what do they know:

Data scientists. Specifically data scientists who do not want to generate or implement visualizations for analyzing model scores on their own. Also caters to people who leverage machine learning models, without having a relevant background, or understanding how the models work.

What information users want from the system:

- Overall distribution of predicted probabilities
- Trade off between false positive and true positive rate for binary classifiers
- Optimal threshold for converting probabilities to binary response
- Information and statistics to analyze the goodness of fit for regression models
- Plots to assess model assumptions

Data sources. What data will be used and how it is structured:

The Titanic and House Price data on Kaggle are popular dataset for building classifier and regression models. Simple logistic regression and linear regression models can be fit on these datasets, and the predictions from models can be used as the data source for testing the tool's functionality of visualizing model performance and assessing model assumptions.

Use cases - how users interact with the system to get the information they want:

- The user will be able to install the python package and import it to make module functionality available
- Given predictions from a binary classifier or regression model, the user will be able to create an instance of a plotting class. Plotting classes will include functionality such as binary plots or regression plots. The user will be able to get and set attributes of that class.
- The user will be able to use and reuse instances of each plotting class by calling methods from that class to visualize model performance, model fit, or model assumptions with one line of code, in order to compare models and choose the one fitted the best.