Assignment 1: HELOC Data Analysis

Date: September 22, 2020

Sumbit (in the ipynb format) via Moodle before 11:59pm October 10, 2020.

Consider an anonymized data of Home Equity Line of Credit (HELOC) loans, with the dataset (`HelocData.csv` and `HelocDataDict.xlsx`) hosted on STAT3612 Github site. The target variable *RiskFlag* indicates whether the loan is ever 90-day delinquent over a two-year period. The feature variables pulled the credit bureau are described in the data dictionary file. Some special values are also described in the data dictionary.

In this assignment you are required to perform the exploratory data analysis and generalized linear modeling. You will need to submit your works in the Python notebook that includes reproducible Python code and adequate description. Indicate your name and UID in the first cell of the notebook.

(1). (20%) The negative values (-7, -8, -9) for each variable can be be regarded as missing information. Calculate the missing value frequencies separately for each feature, then visualize the result by a bar chart.

(2). (20%) Use `np.random.seed` to set your UID as the random seed, then split the data into training (80%) and testing (20%) sets. For each feature variable in the training data, impute the missing values by the mean of observed values.

(3). (20%) For the training data, draw the boxplot for each feature variable as grouped by RiskFlag. Lay out all the plots appropriately and annotate each box plot with the corresponding feature name.

(4). (20%) Select top-5 features based on the boxplots and elaborate your reasons of such choices. Fit a logistic model on the training data with selected features. Rank-order the variable importance based on the Wald tests.

(5). (20%) Interpret the fitted logistic model in (4). Then, test its performance on the testing data and report the prediction accuracy.