Brief Write-up for Machine Learning Homework 2

Tommy Yu

The description of this machine learning pipeline is best presented by the function headers. Only several key steps will be re-illustrated in this brief write-up.

In data exploration, a few interesting facts emerged. Notice that the people who experienced 90 days past due delinquency or worse (i.e., Y=1) are referred to as those who are in financial distress.

- Those who are in financial distress are on average 7 years younger, earn $1,000 less per month, and support 0.2 more dependents than those who are not.
- It is substantially more likely for those in financial distress to have their bill payments past due.
- They also have a lower debt ratio, fewer open loans & credit lines, and lower credit balance on average.
- Check the box charts for a sense of the distribution of the selected independent variables.
- Those with over 8 dependents are never in financial distress.
- The more often “past due” occurs, the more likely it’s the case for a person in financial distress.

The mean of existing data is used to fill in the missing values of monthly income and number of dependents.

“Age” variable is used to create categorical dummy variables. All ages are split into 4 groups corresponding to the four quartiles. A dummy variable is created for each quartile.

The standard logistics regression is adopted as the classifier. All independent variables (age replaced by the 4 dummy variables), except for zip code, are used in the regression.

80% of the available data are used as the training set, while the remaining 20% as testing data.

Accuracy is high at 93.26%, and the AUC of the ROC curve is at 0.68 – both indicating the model adopted is fairly decent.