Summary for Homework 2

Xuan Bu

The followings illustrate the process of building my pipeline:

1. In the first step, I wrote the function read_data to import the raw data (csv) into python. Meanwhile, I dropped the unused column 'zipcode', and set the 'PersonID' as index.
2. In the second step, first, I wrote two functions (summary_continuous_vars, summary_categorical_vars) to do descriptive statistics for both continuous variables and categorical variables; second, I wrote two functions (generate_graph, generate_corr_graph) to generate graphs of variables; third, the function count_outliers is for counting the outliers of different variables.
3. In the third step, there is only one function (fill_missing_with_median) to pre-process the data, i.e., replacing the missing values with median.
4. In the fourth step, to generate features, first, I wrote function discretize_continuous_var to discretize two continuous variables (in this case, I chosen 'MonthlyIncome' and 'age'); second, I used the function create_binary_var to create dummy variables for both variables I chosen.
5. In the fifth step, with using the package sklearn, I first split the data into training set and testing set (split_data), then use the function build_classifier to build three classifiers (Logistic Regression, K-Nearest Neighbors and Decision Tree).
6. In the last step, to evaluate the three classifiers, I chosen the accuracy score and precision score as the criterions, then use the function evaluate_classifier to do the evaluation.

The detailed analysis of the results by running the pipeline is written in the file "write-up.ipynb".