CMPSC497 Fall 2022 Programming Assignment 1.

Assigned: Monday, September 19, 2022

Due: Monday, October 17, 2022 (by midnight, submit a package of codes and report via Canvas)

Maximum: 102 points (out of 100 points)

Note: This assignment is to be done by an individual student. Teamwork is NOT allowed!

Data analysis and preprocessing is the very first and important step before applying data mining and machine learning models for tasks such as classification and clustering. Data analysis aims to study the datasets and get insights of the data and understand difficulty of the target tasks, e.g., observe the value distribution of features and classes, and it examines the correlation between feature values and classes (for the case of classification). Data preprocessing also aims to handle the potential noise in the data to make it suitable for the models.

In the Programming Assignment 1, you are asked to perform a number of data analysis and data preprocessing tasks using the following dataset, and then use the dataset to perform classification. For each task, there are questions that guide you to make observations. You need to answer those questions in your own words in the submitted assignment report. While built-in functions and functions in *numpy*, *pandas, sklearn* and *matplotlib* are most likely sufficient for completing this assignment, you may write your own code or use other packages (if they cannot be done using the above ones).

Dataset

- The *Credit Card Clients Dataset* contains information on default status, demographic factors, credit data, history of payment, and bill statement of credit card clients. The task for this dataset is to predict the default status of an unknown (new) client, given the information of this client.
- Each client, described by one data record, contains credit limit, sex, education, marriage, payment status, bill amount, and paid amount. Please treat column 1 to 24 as *features* (the explanatory variables) and column 25 as the *class* (the targeted variable for prediction).

- For specific information regarding the attributes, please refer to the following table:

ID	Identification of each client
LIMIT_BAL	Amount of given credit in dollar
SEX	Gender (1=male, 2=female)
EDUCATION	Education background (1=graduate school, 2=university, 3=high school, 4=others)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY	History of past payment status (for the last 6 months). Payment status: -1=paid duly, n= payment delayed for n months, 9 = payment delay for nine months or above.
BILL_AMT	Amount of bill statements in dollar for the last 6 months
PAY_AMT	Amount paid in dollar for the last 6 month
DEFAULT	Default status in the next month (1=yes, 0=no)

- Note that you are required to use the dataset provided (which contains only part of the original dataset). The dataset *credit_cards-2022-post.csv* is made available on Canvas under /Home/ Programming Assignments/. To load the dataset, you can use *pandas*, a python library.

Data Analysis

- Task 1. **[15 points]**
  a) It's critical to know the data! By examining the given dataset, do you find unclear or questionable attribute values? Describe what you found, how you identified them, and how you handle them (with justification) for the subsequent tasks. [5 points]
  b) Missing values are very common in real-world datasets as it is not easy to keep track on the data all the time. Discuss one way you could handle missing values; and determine if there is any value missing in this dataset (i.e., please identify and process them if there are missing values). [5 points]
  c) Show the summary statistics (including frequency and mode) of the target (class) attribute (i.e., DEFAULT). In addition, use matplotlib only to plot the distribution of values in the target (class) attribute using a bar chart. The chart should have multiple bars corresponding to the class labels, where each bar shows the count corresponding to each label. (Note: in this dataset, there are only two labels.) You may use different color for each

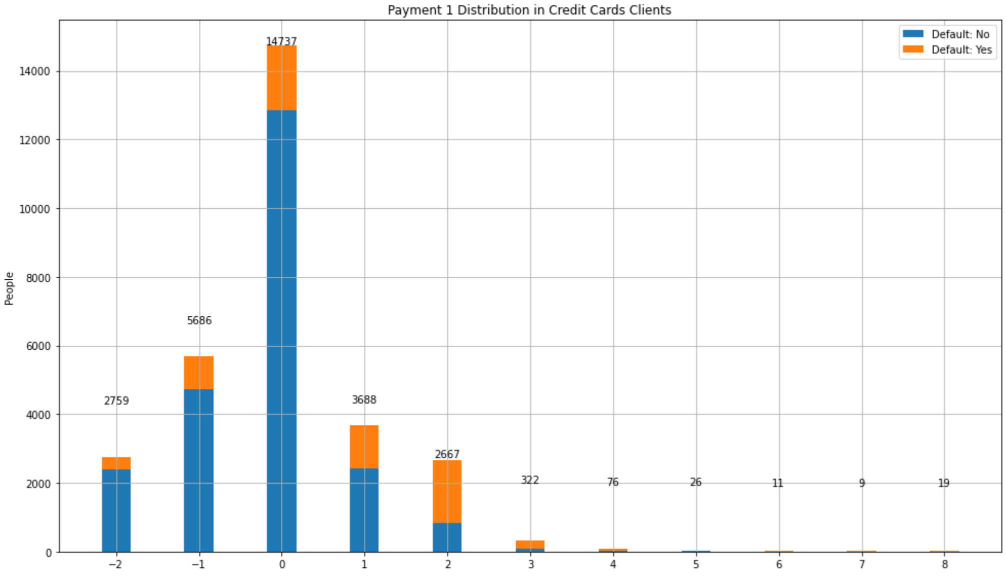
bar. *Please describe what you observed*, *e.g.,* whether the data distribution is imbalanced. [5 points]

- Task 2. [11 points]
  a) For each of the BILL_AMT attributes (6 in total), show the summary statistics (including minimum, maximum, frequency, mean, and standard deviation). [3 points]
  b) Plot a box plot that shows the 6 BILL_AMT attributes. [3 points]
  c) Describe your findings observed from the above questions (i.e., 2(a) and 2(b)). [5 points]

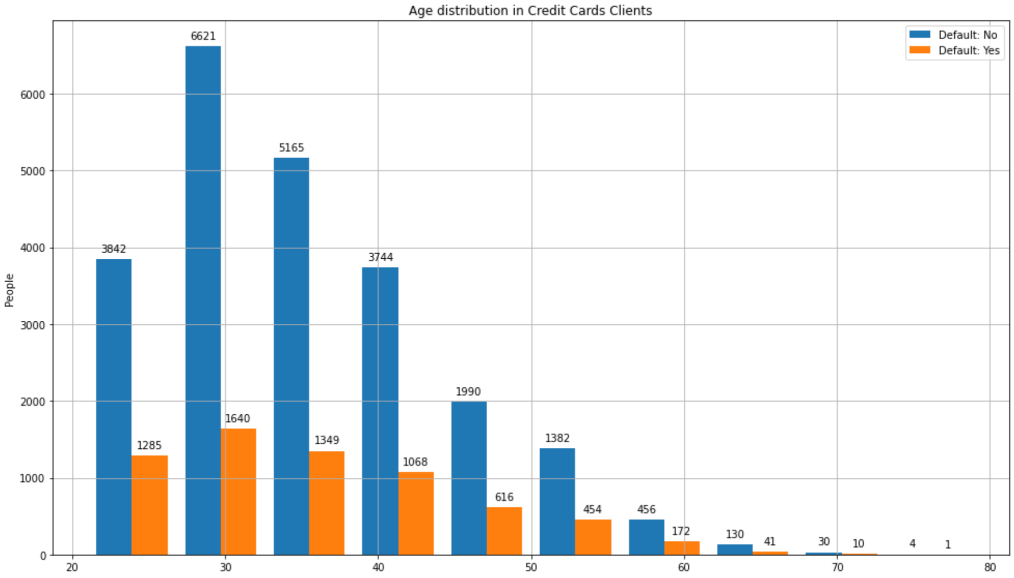
- Task 3. [28 points]
  Follow the links below to read about Chi-Squared Test and Mutual information; and answer the following questions.
  a) Discuss the characteristics and differences of chi-square functions (https://en.wikipedia.org/wiki/Chi-squared_test) and mutual information functions (https://en.wikipedia.org/wiki/Mutual_information) in your own words. [3 points]
  b) Can we directly apply a chi-square function and a mutual information function on this dataset for feature selection? Please explain in accordance with the different attribute types in this dataset. [5 points]
  c) Separately employ chi-square and mutual information as appropriate to obtain a measure between values of each feature and the class. Rank features by their measures of chi-square and mutual information, respectively. [10 points]
  Note: Please make at least two ranked lists: one for numerical features and the other for categorical features. ~~An attribute only belongs to one list.~~
  d) Based on the two ranked lists obtained in Task 3(c), use *matplotlib* only to plot the value distribution of (i) the highest ranked three categorical features, (ii) the lowest ranked three categorical features, (iii) the highest ranked three numerical features, and (iv) the lowest ranked three numerical features. *Describe what you observe from these value distributions and discuss whether the ranks are reasonable.* [10 points]

  Note: Please plot a Bar chart for the categorical features and a Histogram for the numerical features, correspondingly, with the *class* value. See below for examples. For each bar and interval, please color the portion of records/instances corresponding to different classes and show the overall count. For the Histogram, please evenly divide the overall value range into 10 intervals.

Bar Chart



Histogram



Data preprocessing

- Task 4. [20 points]
  a) In addition to the features provided, new features may be generated for data mining. Please derive/generate at least one new feature from existing ones in the dataset and explain why you believe it is useful. Please add the feature or features you propose (those in Assignment 1 and the new ones) to the dataset in accordance with the following format: Column 1 is the client ID and Column 2-24 are the features in the original dataset. Insert

your proposed feature in the subsequent columns, i.e., Column 25 and so on, and move the output class to the last column, e.g., Column 26 if you add only one new feature. [5 points]

b) Normalize the range of values of numerical features (including the generated ones as appropriate) into [0, 1], respectively. For each normalized numerical feature, show the ranges of both its original and normalized values. In your report, please specify (or describe) the adopted normalization scheme. [5 points]

c) Encode categorical features (including the generated ones as appropriate) using *one-hot representation* scheme. For example, assuming that there is a 'state' feature with three categorical values, 'PA', 'NY' and 'NJ'. Create three new binary features, namely 'state_is_PA', 'state_is_NY' and 'state_is_NJ' to replace 'state', where the feature values are either 0 or 1. For each new binary feature, count and report the number of times value 1 occurs, e.g., "state_is_PA": 15000, "state_is_NY": 20000 and "state_is_NJ": 10000. [10 points]

Classification Experiments

Based on the result of data preprocessing, you are asked to learn classification models that predict the *default status of clients in the next month*, and perform a number of tasks in order to evaluate the performance of models under different settings.

- Task 5. [15 points]

  In this task, you will train decision tree classifiers on the Credit Card Clients Dataset to predict whether a client will default in the next month.

  a) There are 14 numerical features and 9 categorical features. Please train **Decision Tree Model 1** based on normalized numerical features and one-hot encoded categorical features, and train **Decision Tree Model 2** based on unnormalized numerical feature and one-hot encoded categorical feature. Feel free to train additional models if you would like to compare the effect of encoding on categorial features (but please state your goal in the report). [5 points]

  b) Please use 5-Fold cross-validation for experiments. (See textbook and https://en.wikipedia.org/wiki/Cross-validation_(statistics)) Please summarize the definitions and mathematical formulae of **confusion matrix**, **precision metric**, **recall metric**, **f-measure metric,** and **accuracy metric.** Please compare the performance of **Decision Tree Model 1** and **Decision Tree Model 2** in terms of these four metrics and give your conclusion. [10 points]

  Note: In your report, please specify (or describe) the adopted normalization scheme.

- Task 6. [13 points]
  a) Please list the reasons why we perform feature selection. [3 points]
  b) Please perform feature selection based on the correlation results in Task 3 (using chi-square for categorical data and mutual information for numerical data). Generate partial datasets by only using top k (k =1, 3, 5) most correlated categorical features and numerical features for model training (i.e., k categorical features + k numerical features). [5 points]
  c) Follow the setup in Task 5 to compare the performance of Decision Tree Models trained using the partial datasets with the one using the original dataset (i.e., without including features generated in Task 4(a)). [5 points]

Deliverables

- Please submit a report (.pdf) and the complete Jupyter Notebook (.ipynb) on Canvas.
- The report should contain your answers to the tasks in the assignment along with the figures (plots) and/or experimental results.