CMPSC497 Fall 2020 Programming Assignment 1.

Assigned: Friday, September 18, 2018

Due: Monday, October 4, 2020 (by midnight, submit a package of codes and report via Canvas)

Maximum: 100 point

Note: This assignment is to be done by an individual student, no teamwork 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 classes and feature values. 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. 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. Please only use built-in functions and functions in numpy, *pandas, sklearn* and *matplotlib* for this assignment.

Dataset

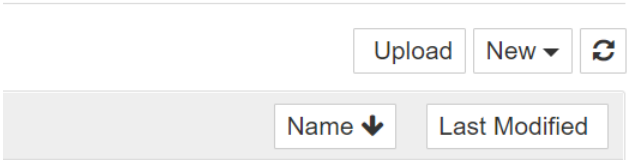
- Default of *Credit Card Clients Dataset* [1] contains information on default status. demographic factors, credit data, history of payment, and bill statement of credit card clients. The task for this dataset is to predict the default status of an unknown (new) client, given the information of this 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, 5,6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY	Payment status (-1=paid duly, n=paid late) for last 6 month
BILL_AMT	Amount of bill statement in dollar for last 6 month
PAY_AMT	Amount received in dollar for last 6 month
DEFAULT	Default status in 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.csv* is made available on Canvas under /Home/ Programming Assignments/. To load the dataset, you can use *pandas*, a python library. If you are on VMHost, you may need to install it on your own. To install, type the following command and run it in Jupyter Notebook.

```
In [ ]: import sys
        !sudo -H {sys.executable} -m pip install pandas
```

After installation, you can simply import pandas to load the dataset uploaded in the home directory using the upload button at Jupyter Home:




```
In [ ]: import pandas as pd
        data = pd.read_csv('credit_cards.csv')
```

Next, you can test if the data is loaded in successfully:

```
In [2]: data.head()
```

Out[2]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0
0	1	20000.0	2	2	1	24	2
1	2	120000.0	2	2	2	26	-1
2	3	90000.0	2	2	2	34	0
3	4	50000.0	2	2	1	37	0
4	5	50000.0	1	2	1	57	-1

5 rows × 25 columns

Data Analysis

- Task 1.
  a) 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 would do when there are missing values; and determine if there is any value missing in this dataset.
  b) Plot the distribution of values in the *class* attribute (DEFAULT) of the dataset using a bar chart. *Please describe what you observe*, *e.g.* whether the data distribution is imbalanced.

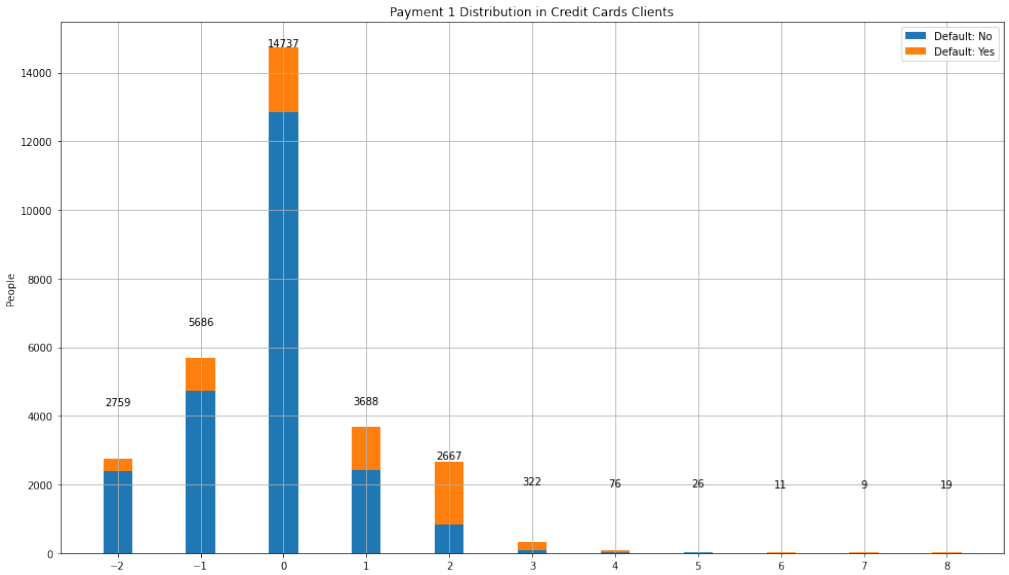
- Task 2. Follow the links below to read about Chi-Squared Test and Mutual information; and answer the following questions.
  a) Discuss the characteristics and differences of chi-square function (https://en.wikipedia.org/wiki/Chi-squared_test) and mutual information functions (https://en.wikipedia.org/wiki/Mutual_information) in your own words.
  b) Can we directly apply chi-square function and mutual information function on this dataset for feature selection? Please explain in accordance with the different attribute types in this dataset. (hint: the difference between categorical and numerical data)
  c) Employ chi-square or 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.
  Note: Please make two lists: one for chi-square and the other for mutual information. An attribute only belongs to one list.

- Task 3. Based on the two ranked lists obtained in Task 2, 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

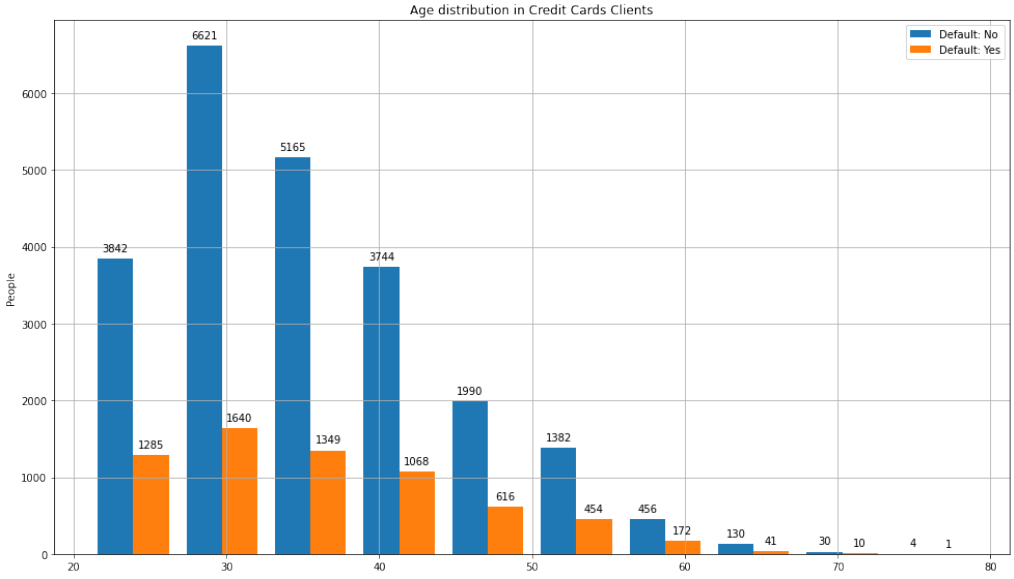
numerical features, and (iv) the lowest ranked three numerical features. *Describe what you observe from these value distributions and discuss whether the ranks are reasonable.*

Note: Please plot a Bar chart for the categorical feature and a Histogram for a numerical feature, 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 Histogram, please evenly divide the overall value range into 10 intervals.

Bar Chart



Histogram



Data preprocessing

- Task 3. In addition to features provided, new features may be generated for data mining. Please derive/generate at least one or more new features from existing ones in the dataset and explain why you believe they are useful.
- Task 4. 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 its original and normalized values.
- Task 5. 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 value 1, e.g., "state_is_PA": 15000, "state_is_NY": 20000 and "state_is_NJ": 10000.

Packages

- sklearn (http://scikit-learn.org/). A machine learning framework in Python
- matplotlib (https://matplotlib.org/). Website provides tutorials on how to plot bar chart and histogram in Python.
- NumPy (http://scikit-learn.org/). A fundamental package for scientific computing in Python.
- pandas (https://pandas.pydata.org/). A framework for easy-to-use data analysis and manipulation in Python

Deliverables

Please submit a report (.pdf) and the complete Jupyter Notebook (.ipynb) on Canvas. The report should contain your answers for the tasks along with the figures (plots).

Reference

[1] I-Cheng Yeh, Che-hui Lien, The comparison of data mining techniques for the predictive accuracy of probability of default card clients, Expert Systems with Applications, Volume 36 Issue 2 Part 1, 2009, Pages 2473-2480 https://www.sciencedirect.com/science/article/pii/S0957417407006719 (accessible on PSU VPN or PSU access)

[2] Please refer to http://scikit-learn.org/stable/modules/preprocessing.html#.