Manufacturing Data Science 製造數據科學

Assignment 1

Due Date: 5pm, Oct. 14, 2021

Please solve the following questions and justify your answer by using Python. Show all your analysis result including Python code in your report. **Upload your "zip" file including MS Word report (answering each question and its sub-questions) and Python code with 檔名: MDS_Assignment1_ID_Name.zip" to NTU COOL** by due. The late submission is not allowed.

1.* (35%) *Linear Regression Analysis for Wine Quality

For the attached metal furnace dataset (**MDS_Assignment1_furnace.csv**), please use "multiple regression" to find the potential linear pattern (i.e., linear regression equation) for 621 observations with 28 input variables (f0-f27) and 1 output variable (grade) (label variable is regarded as continuous variable 反應變數請視為連續變數). Please answer the following questions by using Python software and package:

(a) **(10%)** Show the results of regression analysis as follows.

variable	estimate	std. error	t-value	p-value
f0				
f1				
f2				
…				
f27				

R-squared: 0.xxxx, Adjusted R-squared: 0.xxxx

(b) **(5%)** The fitting of the linear regression is a good idea? If yes, why? If no, why? What's the possible reason of poor fitting?

(c) **(5%)** Based on the results, rank the independent variables by p-values and which one are statistically significant variables with p-values<0.01? (i.e. 重要變數挑選)

(d) **(15%)** Testify the underlying assumptions of regression (1) Normality, (2) Independence, and (3) Homogeneity of Variance with respect to residual.

Context

Manufacturing of any alloy is not a simple process. Many complicated factors are involved in the making of a perfect alloy, from the temperature at which various metals are melted to the presence of impurities to the cooling temperature set to cool down the alloy. Very minor changes in any of these factors can affect the quality or grade of the alloy produced.

Content

Given are 28 distinguishing factors in the manufacturing of an alloy, the objective is to build a Machine Learning model that can predict the grade of the product using these factors.

You are provided with 28 anonymized factors (f0 to f27) that influence the making of a perfect alloy that is to be used for various applications based on the grade/quality of the obtained product.

Source: Yash Ajgaonkar (2020), https://www.kaggle.com/esotericazzo/metal-furnace-dataset

2. (30%) Data Preprocessing and Generalized Linear Model (GLM)/Logistic Regression

This dataset can be used to predict the census income and it can be collected from the 1994 Census database. Data set is **MDS_Assignment1_census.csv** and data source is https://archive.ics.uci.edu/ml/datasets/Census+Income. The dataset includes 48842 observations, 14 attributes, and 1 response variable. The last attribute is the "Class" label.

Please answer the following questions:

(1) **(5%)** Provide the descriptive statistics. (i.e. exploratory data analysis, EDA) Eg. mean, variance, data distribution, # of missing value, # of outlier, etc.

(2) **(10%)** How to identify the outlier? How to impute the missing value?

(3) **(5%)** How to transform the categorical variable to dummy variable?

(4) **(5%)** How to "randomly" split the dataset into training dataset and testing dataset (eg. 80% vs. 20%)?

(5) **(5%)** Please use the Generalized Linear Model (GLM) (**OR** Logistic Regression) to predict the "Class" in the testing dataset.

3. (35%) Association Rule- Market Basket Analysis

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer's basket – and therefore 'Market Basket Analysis'.

That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item. You can see the Groceries data set (groceries.csv). Use "association rule" to find the potential patterns which satisfy the following criterion:

- Set the minimum support to 0.001
- Set the minimum confidence of 0.15

Please answer the following questions:

(1) **(10%)** How to handle the raw dataset via data preprocessing?

(2) **(10%)** What's the top 5 association rules? Show the support, confidence, and lift to each specific rule, respectively?

(3) **(5%)** Please provide/guess the "story" to interpret **one** of top-5 rules you are interested in.

(4) **(10%)** Give a visualization graph of your association rules.

You may follow the guideline in the linkage step-by-step:
https://pbpython.com/market-basket-analysis.html

Source: Salem Marafi,
http://www.salemmarafi.com/wp-content/uploads/2014/03/groceries.csv

Note

1. Show all your work in detail. **Innovative** idea is encouraged.
2. If your answer refers to any external source, please "must" give an academic citation. Any "plagiarism" is not allowed.