**兆豐銀行大數據分析人員上機實作**

**【第一題】**

資料一：白葡萄酒分級資料(資料檔名：winequality-white.csv 內含4,898筆資料)

**說明：**

本資料中共含有12個變數，其中有11個變數是葡萄酒檢測值(如：pH 值)，其中”葡萄酒分級變數”(quality)為一主觀的變數，其數值是由至少三個專家的評分取中位數而來，每個專家可以對一種酒給0~10的評分，其中0是非常差、10是非常好。

請分別以下列二種方法來建立，可預測某一種酒之等級的模型：

1. 多元線性回歸：(15分)

多元線性回歸模型的部分，可以全部的資料作為訓練樣本，最終模型的調整後判定係數(Adjusted R-square)需大於0.33。

1. 決策樹：(15分)

決策樹的部分，請抽樣80%的資料作為訓練樣本、20%的資料作為測試樣本。在模型建置完成後，以此模型對測試樣本做預測，並產出混淆矩陣與預測準確度，其中預測準確度需大於0.65。

|  |
| --- |
| 1. 資料分析流程 2. 資料分析過程中使用到的分析工具 3. 資料分析時使用到的演算法、以及選用該演算法的原因 4. 模型調校的過程與步驟、最終調校好的結果(回歸模型公式與ANOVA報表、決策樹與混淆矩陣) |

資料(winequality-white.csv)中的變數資訊如下：

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

**【第二題】**

資料二：經去識別化的帳務交易資料(txhist.csv，共50萬筆資料)

資料三：匯率資料(FXRATE.csv，各種幣別對台幣的匯率資料)

**說明：**

資料二為一段區間內客戶的交易**帳務資料**，包含4個欄位，資訊如下表：

|  |  |  |
| --- | --- | --- |
| ***Column Name*** | ***Column Type*** | ***Description*** |
| ID | String | 客戶ID |
| DR | String | 帳務借貸方  D:借方 C:貸方 |
| CUR\_CODE | String | 2位幣別代碼(例：’00’, ‘01’) |
| AMT | float | 交易金額 |

資料三**匯率資料**包含2個欄位，資訊如下表：

|  |  |  |
| --- | --- | --- |
| ***Column Name*** | ***Column Type*** | ***Description*** |
| CUR\_CODE | String | 2位幣別代碼(例：’00’, ‘01’) |
| FXR\_TWD | Float | 該幣別兌換台幣的匯率，例：幣別‘01’ = 30.71 |

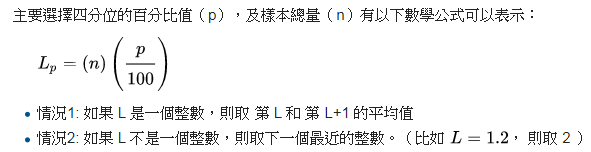
由帳務交易資料，計算各客戶ID借方與貸方的等值台幣總金額，並提供下列資料：  
(1) 各客戶借方總金額中的第一四分位數(Q1)、中位數(Q2)的金額與對應客戶ID：(15分)

1. [借方總金額 – 貸方總金額]的最大值與對應客戶ID： (15分)

註1：金額皆保留小數位數，不作四捨五入

註2：若轉換後金額為Na(或Nan)，以0取代

註3：四分位數公式參考如下：



**計算範例：**

帳務交易資料：

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **DR** | **CUR\_CODE** | **AMT** |
| @123456789 | D | ‘00’ | 1000 |
| @123456789 | D | ‘01’ | 100 |
| @123456789 | C | ‘00’ | 500 |
| B666123777 | D | ‘39’ | 200 |
| B666123777 | C | ‘00’ | 100 |
| C543210888 | C | ‘00’ | 600 |

匯率資料

|  |  |
| --- | --- |
| **CUR\_CODE** | **FXR\_TWD** |
| 00 | 1.0 |
| 01 | 30.91 |
| 39 | 4.48888 |

@123456789 的借方台幣總金額 = (1000 \* 1.0) + (100\*30.91) = 3091

@123456789 的貸方台幣總金額 = (500 \* 1.0) + (100\*30.91) = 500

@123456789 的借貸台幣差額 = 3091 – 500 = 2591

B666123777的借方台幣總金額 = (200\*4.4888) = 897.76

B666123777的貸方台幣總金額 = (100\*1) = 100

B666123777的借貸台幣差額 = 897.76 – 100 = 797.76

C543210888的借方台幣總金額 = 600\*1.0 = 600

C543210888的貸方台幣總金額 = 0

C543210888的借貸台幣差額 = 600 – 0 = 600

計算後資料

|  |  |  |
| --- | --- | --- |
| **ID** | **借方總金額** | **借貸差額** |
| @123456789 | 3091 | 2591 |
| B666123777 | 897.76 | 797.76 |
| C543210888 | 600 | 600 |

**故借方總金額Q2=897.76， 對應ID= B666123777**

**MAX(借貸差額) = 2591，對應ID=@123456789**

**投影片製作：**

請準備一份投影片並以十分鐘長度簡單說明以下幾點：

(1)題目一：白葡萄酒分級資料**(20%)**

1. 資料分析流程
2. 資料分析過程中使用到的分析工具
3. 資料分析時使用到的演算法、以及選用該演算法的原因
4. 模型調校的過程與步驟、最終調校好的結果(回歸模型公式與ANOVA報表、決策樹與混淆矩陣)

(2)題目二：**(20%)**

a. 列出兩小題題目之答案、簡述得到兩小題題目之答案所用的技巧

