Statistical analysis for bigdata:
Home Work 2

Student: 610721204　陳克威

GitHub Repository URL:
https://github.com/D1034181036/Bank_Marketing_Analysis

1. 資料集 Data Set：

   此資料集為 UCI Machine Learning Repository 中的 Bank Marketing Data Set，是一葡萄牙銀行機構的電話行銷資料，其目標為分類(Classification)問題，預測顧客是否會定期存款。

   此資料集共有 20 個 Feature 與 1 個 Label(Yes/No)，樣本數為 4119 筆 (10%)，其特徵可大致分為四個部分：

顧客相關的特徵	描述
1. Age	顧客的年齡
2. Job	顧客的職業
3. Marital	顧客的婚姻狀態
4. Education	顧客的教育程度
5. Default	顧客的信用狀況
6. Housing	顧客是否有房屋貸款
7. Loan	顧客是否有個人貸款

與前一次行銷相關的特徵	描述
8. Contact	使用電話或手機
9. Month	月份(一至十二)
10. Day_of_week	星期(一至日)
11. Duration	通話時間

前一次行銷的特徵	描述
12. Campaign	當前活動與顧客的聯絡次數
13. Pdays	相隔前次聯絡的天數
14. Previous	過去與顧客的聯絡次數
15. Poutcome	過去的行銷結果

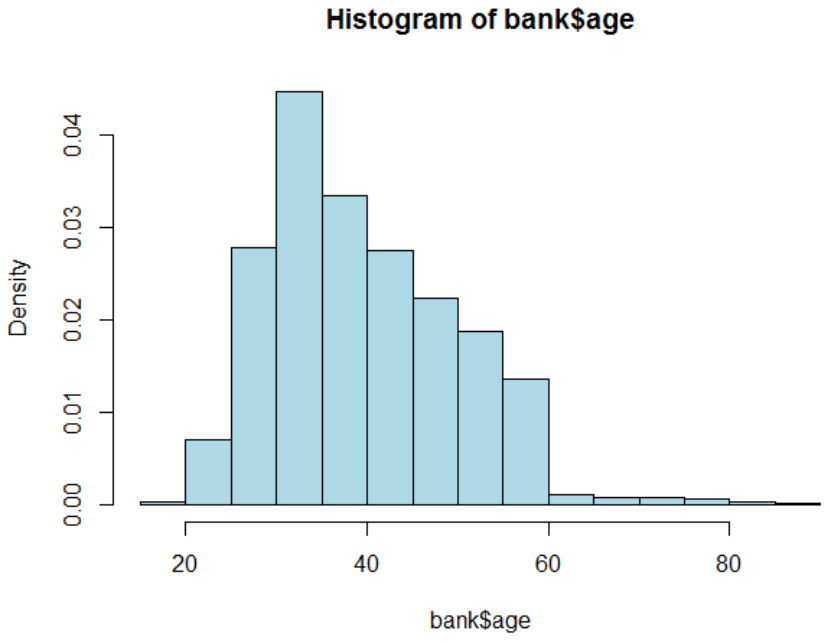
經濟條件	描述
16. emp.var.rate	就業率(季)
17. cons.price.idx	消費者物價指數(月)
18. cons.conf.idx	消費者信心指數(月)
19. euribor3m	銀行同業拆放利率(日)
20. nr.employed	員工人數(季)

2. 資料前處理與初步分析 Data preprocessing and analysis：

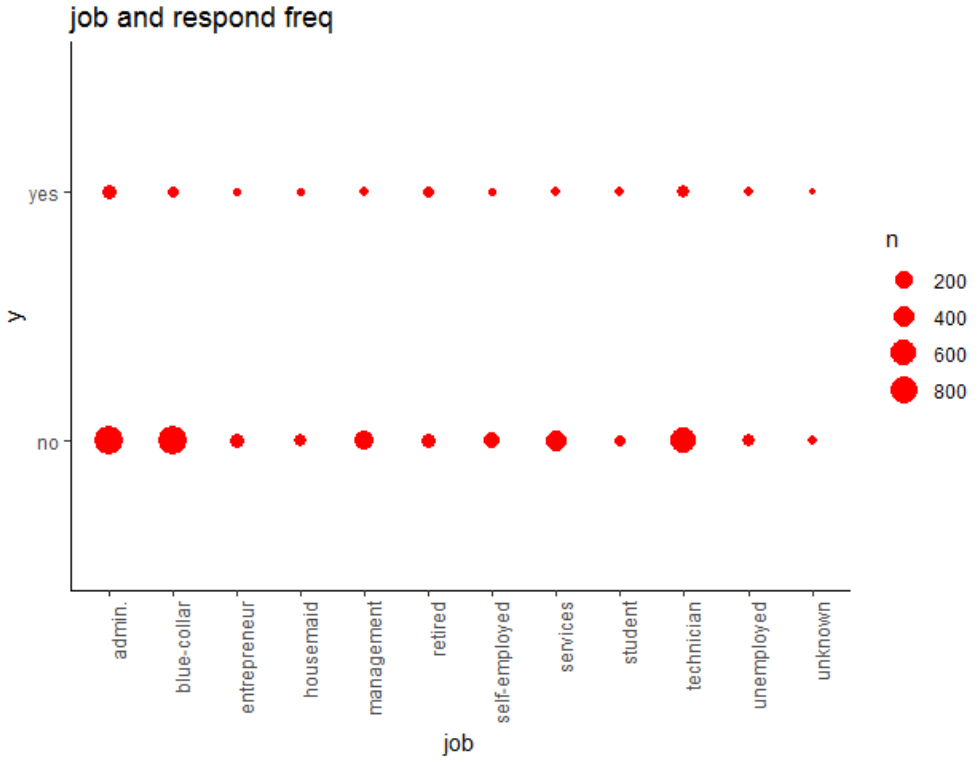
(1) 電話行銷的結果總數。

Yes	No
451 (11%)	3668 (89%)

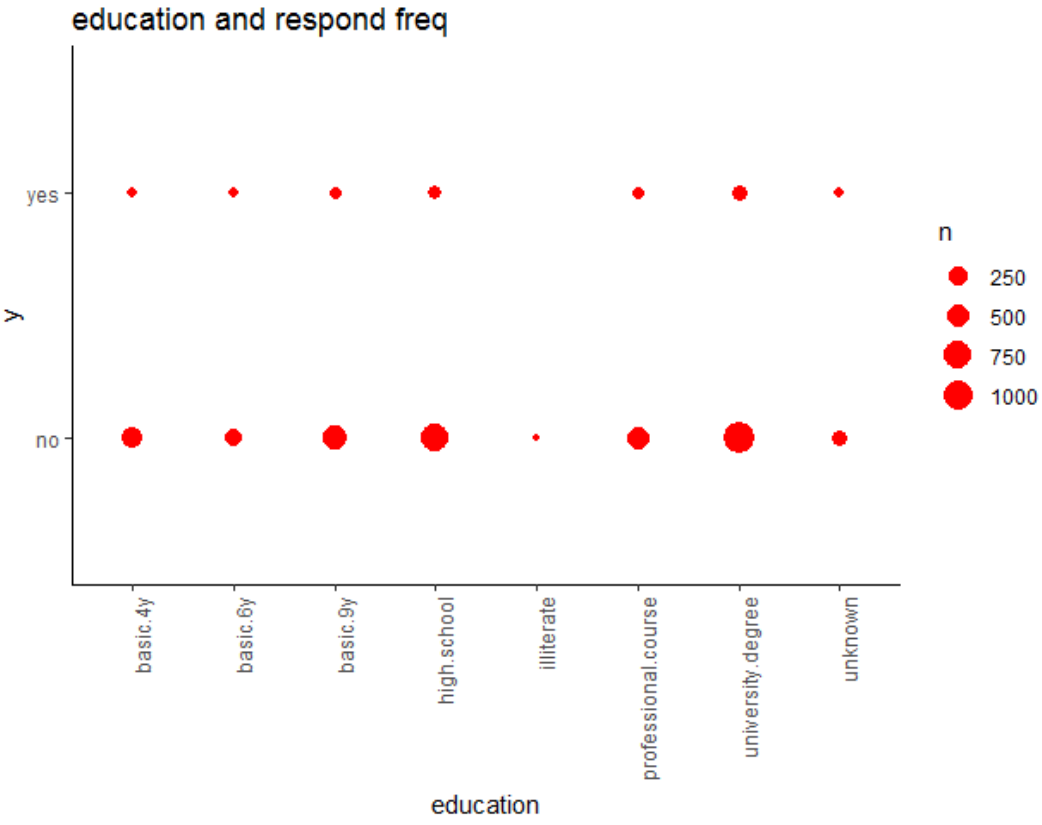
(2) 顧客的年齡分佈：如圖所示大部分顧客的年齡為 18 至 60 歲。



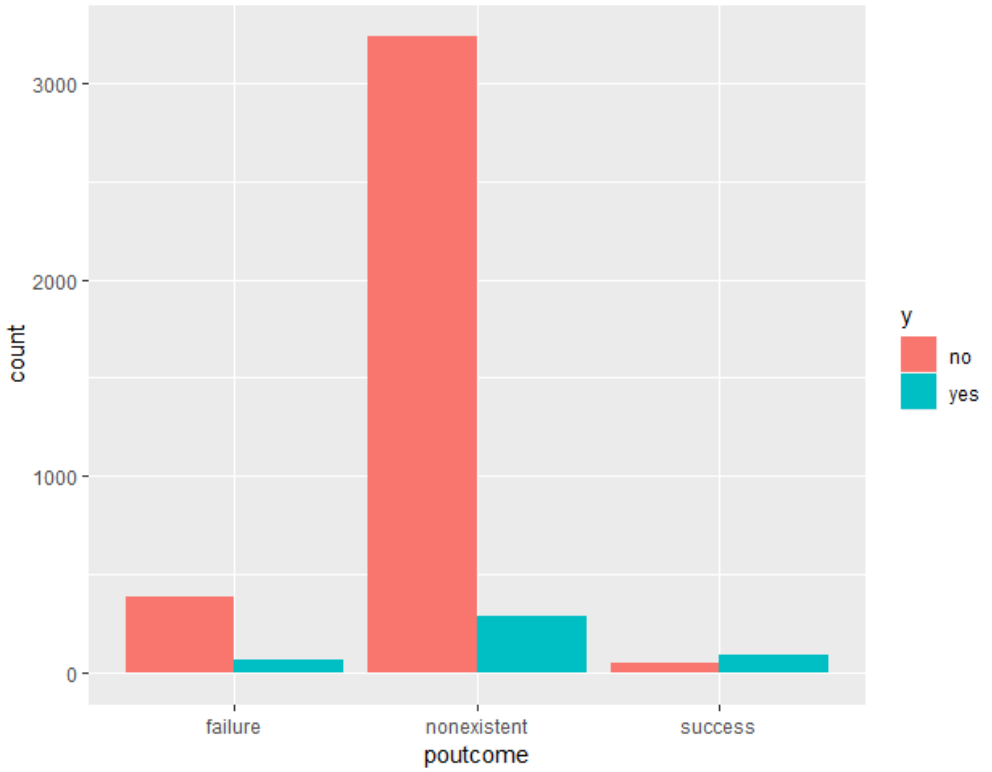
(3) 工作與行銷成功的關係。



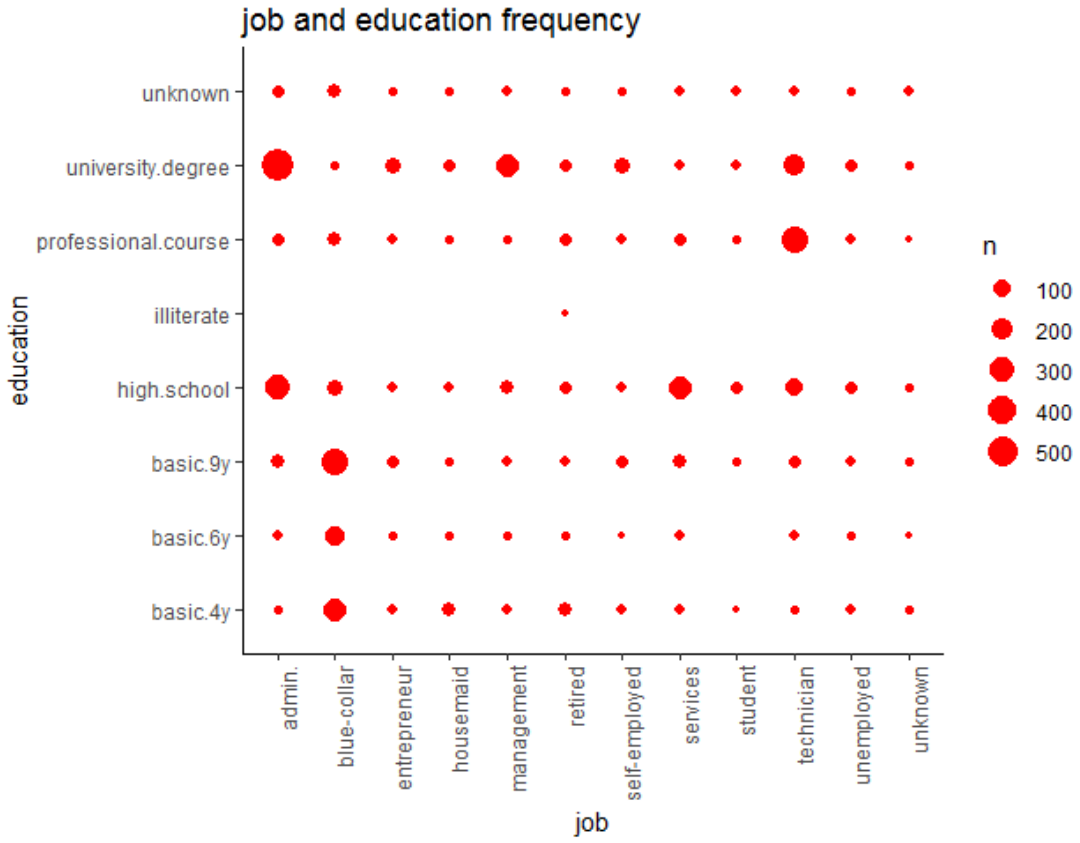
(4) 教育程度與行銷成功的關係。



(5) 前一次行銷結果與本次行銷結果的關係。



(6) 工作與教育程度的關係。



(7) 交易成功的資料裡，平均通話時間為 560 秒，最少的通話時間為 63 秒。
(8) 因 Duration 欄位在通話前無法得知，因此本研究將其捨棄。
(9) 將資料分為訓練資料(80%)以及測試資料(20%)

3. 建立分類器(Classifier)模型 Modeling：

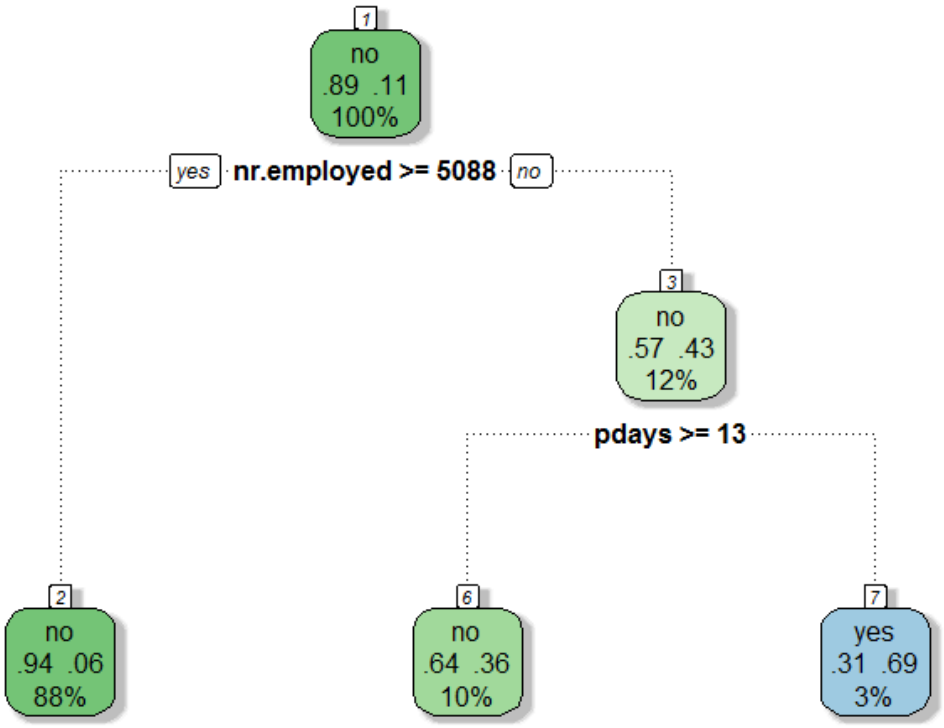
在建立各分類器前，本研究預先建立評估函式以便計算各模型之 Accuracy, Precision, Recall 及 F1-Score，其代碼如下圖所示：

```
#Custom Performance function
Performance <- function(predict,label){
  prop_table <- prop.table(table(predict,label))
  accuracy <- prop_table[1,1] + prop_table[2,2]
  precision <- prop_table[2,2]/(prop_table[2,2]+prop_table[2,1])
  recall <- prop_table[2,2]/(prop_table[2,2]+prop_table[1,2])
  f1 <- 2*precision*recall/(precision+recall)
  performance_vector <- c(accuracy,precision,recall,f1)
  names(performance_vector) <- c("Accuracy","Precision","Recall","F1-Score")
  return(performance_vector)
}
```

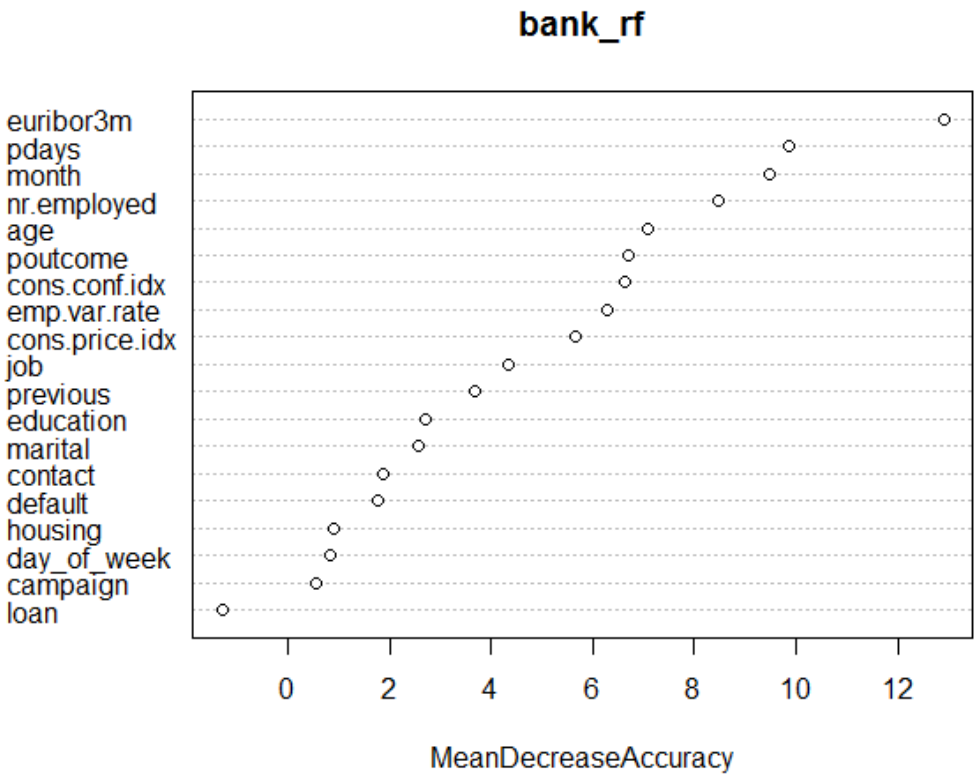
本研究總共建立了六種分類器，分別為：

(1) Decision Tree (CART)
(2) Support Vector Machine (963 Support Vectors)
(3) Naive Bayes
(4) Random Forest (n tree = 100, importance = True)
(5) K Nearest Neighbor (method = cv, k = 5)
(6) Ada Boost (loss = exponential)

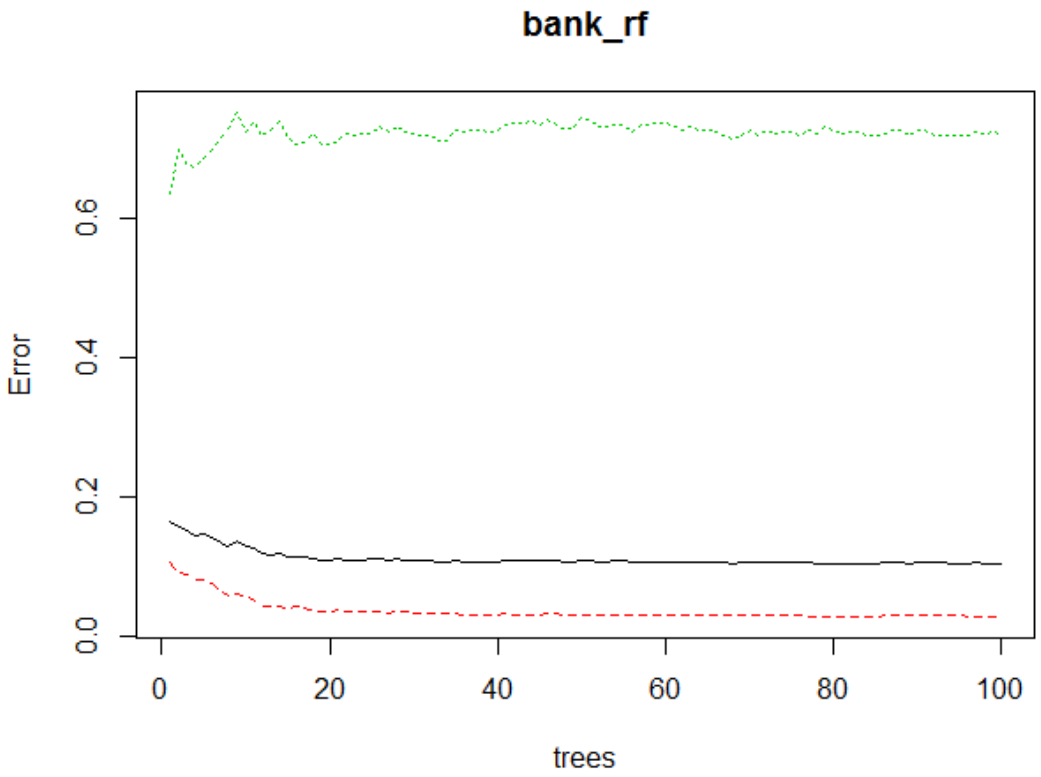
其中 Decision Tree 的規則可以較為直觀的呈現，如圖所示。



也可以從 Random Forest 中看出各屬性對於正確率的重要程度，如圖所示。



另外，值得一提的是在 Random Forest 中，雖然樹的數量越多整體正確率越高，但是相對犧牲了 Label=Yes 的正確率，如圖所示。



4. 實驗成果 Experiment

本研究總共做了三次實驗，分別設定 Seed 為 100, 200, 300，使用訓練資料建立分類器，再使用分類器預測測試資料，最後將其平均後的結果如下（詳細結果列於附錄中）：

	Accuracy	Precision	Recall	F1-Score
CART	0.905	0.652	0.243	0.35
SVM	0.904	0.641	0.213	0.32
Naive_Bayes	0.848	0.36	0.565	0.44
Random_Forest	0.898	0.533	0.282	0.368
KNN	0.897	0.546	0.236	0.328
ADA	0.9	0.576	0.233	0.332

在本研究中，Accuracy 的重要程度較低，即使分類器將所有樣本都分為 No 在 Accuracy 也可以有 89%的正確率，而 Precision 的損失代價也較小，本研究的目的是找出電話行銷成功率高的目標，因此提升 Recall 的分數會是我們的比較在乎的。

在表中可以發現，Naïve Bayes 雖然在 Accuracy 與 Precison 的分數較低，但是 Recall 及 F1-Score 的分數較高，可以理解為分類器較敢將目標分類為 Yes。

另外，Random Forest 整體的表現較好，F1-Score 也在第二名，是相當穩定的一種模型。

5. 參考資料

1. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

2. UCI Machine Learning Repository Bank Marketing Data Set : http://archive.ics.uci.edu/ml/datasets/Bank+Marketing (last access: 2019/06/05)

3. Arjun Reddy(19 June 2016) Building Machine Learning Models : https://rpubs.com/arjunreddyt/190610 (last access: 2019/06/05)

4. Pavan (14 July 2018) bank marketing analysis https://rpubs.com/pavan721/bank_marketing (last access: 2019/06/05)

5. Scott Horvath Implementation of C5 Decision Tree https://www.kaggle.com/scotthorvath/implementation-of-c5-decision-tree (last access: 2019/06/05)

6. 附錄

以下為三次實驗的結果。

Seed = 100

	Accuracy	Precision	Recall	F1-Score
CART	0.905	0.625	0.179	0.278
SVM	0.903	0.577	0.179	0.273
Naive_Bayes	0.847	0.343	0.548	0.422
Random_Forest	0.899	0.512	0.25	0.336
KNN	0.905	0.607	0.202	0.304
ADA	0.905	0.588	0.238	0.339

Seed = 200

	Accuracy	Precision	Recall	F1-Score
CART	0.915	0.8	0.273	0.407
SVM	0.91	0.733	0.25	0.373
Naive_Bayes	0.859	0.401	0.648	0.496
Random_Forest	0.903	0.574	0.352	0.437
KNN	0.903	0.595	0.284	0.385
ADA	0.908	0.667	0.273	0.387

Seed = 300

	Accuracy	Precision	Recall	F1-Score
CART	0.894	0.532	0.278	0.365
SVM	0.899	0.613	0.211	0.314
Naive_Bayes	0.837	0.336	0.5	0.402
Random_Forest	0.892	0.512	0.244	0.331
KNN	0.883	0.435	0.222	0.294
ADA	0.888	0.472	0.189	0.27