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 |
| 1. Job | 顧客的職業 |
| 1. Marital | 顧客的婚姻狀態 |
| 1. Education | 顧客的教育程度 |
| 1. Default | 顧客的信用狀況 |
| 1. Housing | 顧客是否有房屋貸款 |
| 1. Loan | 顧客是否有個人貸款 |

|  |  |
| --- | --- |
| 與前一次行銷相關的特徵 | 描述 |
| 1. Contact | 使用電話或手機 |
| 1. Month | 月份(一至十二) |
| 1. Day\_of\_week | 星期(一至日) |
| 1. Duration | 通話時間 |

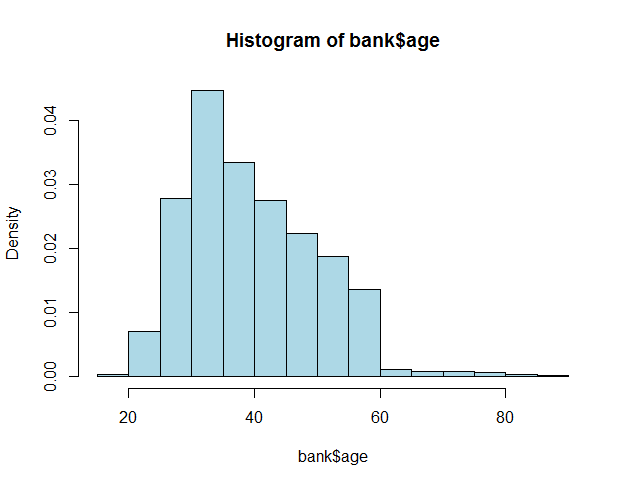
|  |  |
| --- | --- |
| 前一次行銷的特徵 | 描述 |
| 1. Campaign | 當前活動與顧客的聯絡次數 |
| 1. Pdays | 相隔前次聯絡的天數 |
| 1. Previous | 過去與顧客的聯絡次數 |
| 1. Poutcome | 過去的行銷結果 |

|  |  |
| --- | --- |
| 經濟條件 | 描述 |
| 1. emp.var.rate | 就業率(季) |
| 1. cons.price.idx | 消費者物價指數(月) |
| 1. cons.conf.idx | 消費者信心指數(月) |
| 1. euribor3m | 銀行同業拆放利率(日) |
| 1. nr.employed | 員工人數(季) |

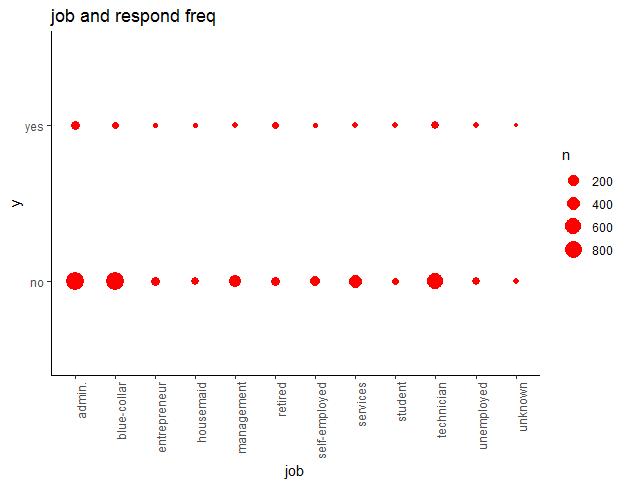
1. 資料前處理與初步分析Data preprocessing and analysis：
2. 電話行銷的結果總數。

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

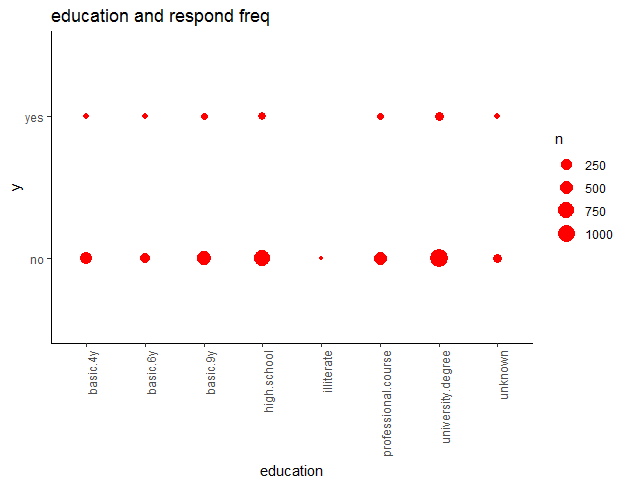
1. 顧客的年齡分佈：如圖所示大部分顧客的年齡為18至60歲。



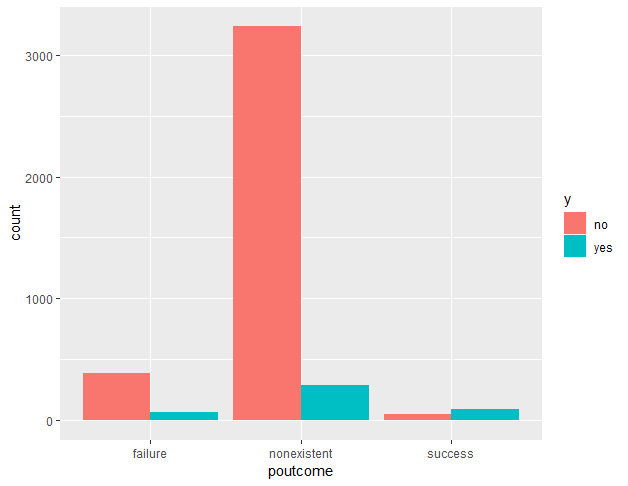
1. 工作與行銷成功的關係。



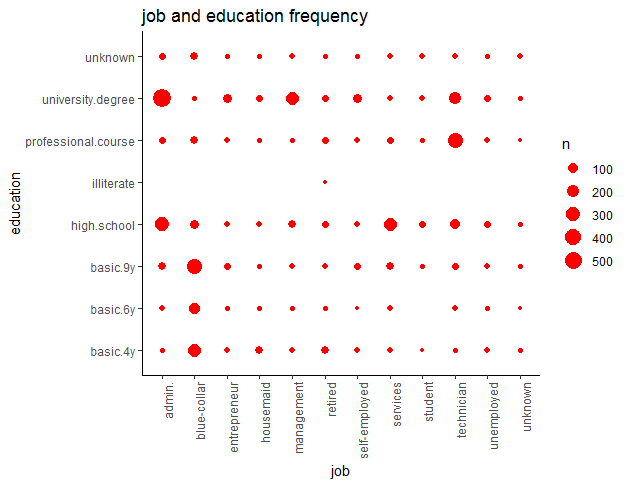
1. 教育程度與行銷成功的關係。



1. 前一次行銷結果與本次行銷結果的關係。

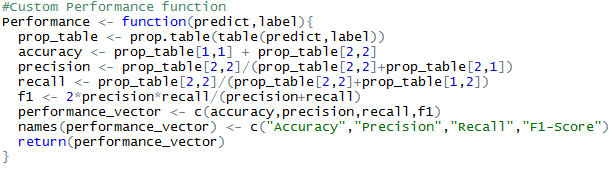


1. 工作與教育程度的關係。



1. 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。
2. 因Duration欄位在通話前無法得知，因此本研究將其捨棄。
3. 將資料分為訓練資料(80%)以及測試資料(20%)
4. 建立分類器(Classifier)模型 Modeling：

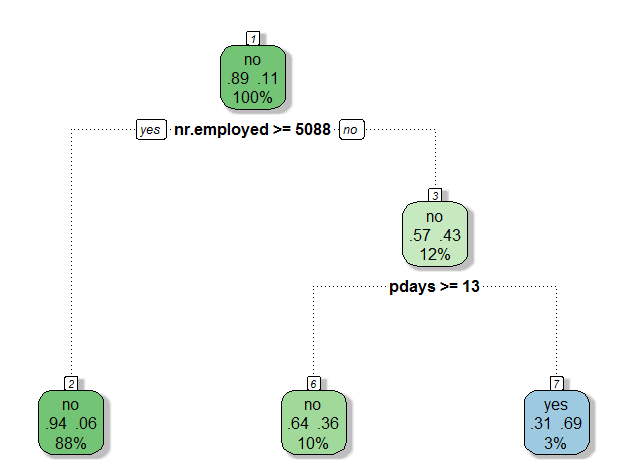
在建立各分類器前，本研究預先建立評估函式以便計算各模型之Accuracy, Precision, Recall及F1-Score，其代碼如下圖所示：



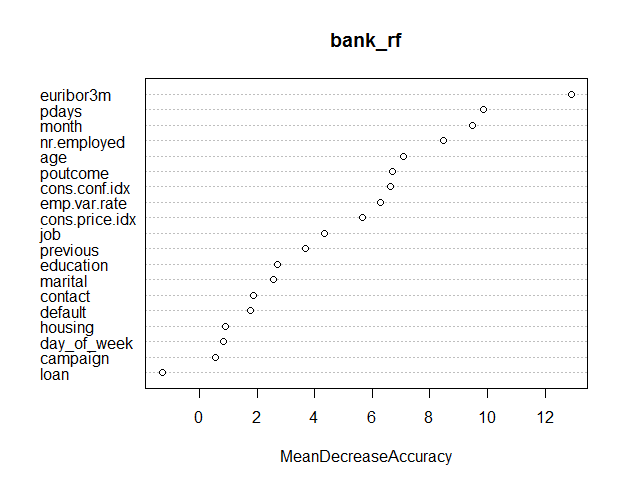
本研究總共建立了六種分類器，分別為：

* 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)

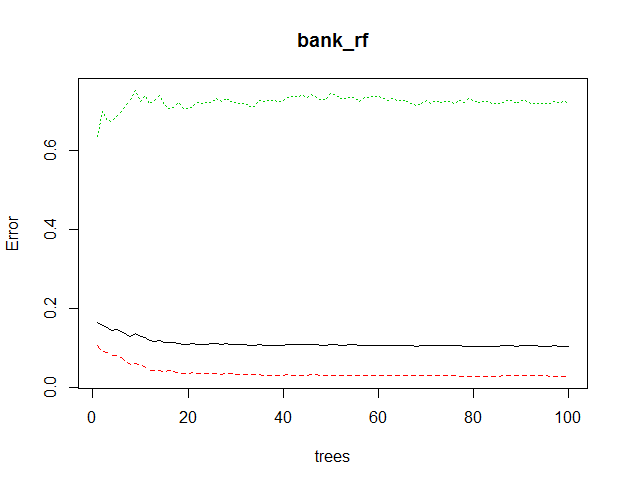
其中Decision Tree的規則可以較為直觀的呈現，如圖所示。



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。

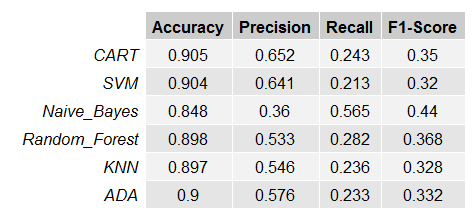


另外，值得一提的是在Random Forest中，雖然樹的數量越多整體正確率越高，但是相對犧牲了Label=Yes的正確率，如圖所示。



1. 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）：



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。

1. 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

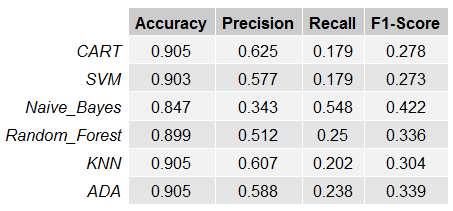
<https://www.kaggle.com/scotthorvath/implementation-of-c5-decision-tree>

(last access: 2019/06/05)

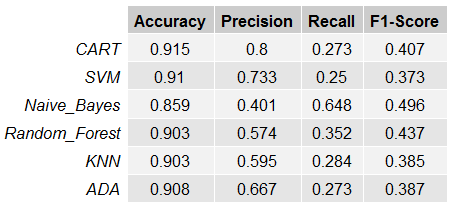
1. 附錄

以下為三次實驗的結果。

Seed = 100



Seed = 200



Seed = 300

