本研究所使用之垃圾郵件（Spam）資料集來自UCI機器學習資料庫中。該資料集是由 Hewlett-Packard Laboratories 的 Mark Hopkins、Erik Reeber、George Forman 及 Jaap Suermondt 等人於 1999 年整理提供。此資料集的建立，部分內容參考了 Cranor 與 LaMacchia（1998）於 Communications of the ACM 期刊中發表的「Spam!」一文。

表5.1為垃圾郵件資料集的變數說明，此資料集共包含 58 個變數，涵蓋了單字出現比例、特殊字元出現比例、與大寫字母使用情形相關的統計指標，以及是否為垃圾郵件的分類標籤。資料集中共有 4,601 筆觀察值，其中標記為非垃圾郵件（記作 0）者共有 2,788 筆，占整體資料的約 60.6%；標記為垃圾郵件（記作 1）者共有 1,813 筆，占整體資料的約 39.4%。

表5.1 垃圾郵件資料集變數說明

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 變數型態 | 備註 |
| 1 | V1 | 單字匹配的百分比 | 數值型 | 單字為「make」 |
| 2 | V2 | 單字匹配的百分比 | 數值型 | 單字為「address」 |
| 3 | V3 | 單字匹配的百分比 | 數值型 | 單字為「all」 |
| 4 | V4 | 單字匹配的百分比 | 數值型 | 單字為「3d」 |
| 5 | V5 | 單字匹配的百分比 | 數值型 | 單字為「our」 |
| 6 | V6 | 單字匹配的百分比 | 數值型 | 單字為「over」 |
| 7 | V7 | 單字匹配的百分比 | 數值型 | 單字為「remove」 |
| 8 | V8 | 單字匹配的百分比 | 數值型 | 單字為「internet」 |
| 9 | V9 | 單字匹配的百分比 | 數值型 | 單字為「order」 |
| 10 | V10 | 單字匹配的百分比 | 數值型 | 單字為「mail」 |
| 11 | V11 | 單字匹配的百分比 | 數值型 | 單字為「receive」 |
| 12 | V12 | 單字匹配的百分比 | 數值型 | 單字為「will」 |
| 13 | V13 | 單字匹配的百分比 | 數值型 | 單字為「people」 |
| 14 | V14 | 單字匹配的百分比 | 數值型 | 單字為「report」 |
| 15 | V15 | 單字匹配的百分比 | 數值型 | 單字為「addresses」 |
| 16 | V16 | 單字匹配的百分比 | 數值型 | 單字為「free」 |
| 17 | V17 | 單字匹配的百分比 | 數值型 | 單字為「business」 |
| 18 | V18 | 單字匹配的百分比 | 數值型 | 單字為「email」 |
| 19 | V19 | 單字匹配的百分比 | 數值型 | 單字為「you」 |
| 20 | V20 | 單字匹配的百分比 | 數值型 | 單字為「credit」 |
| 21 | V21 | 單字匹配的百分比 | 數值型 | 單字為「your」 |
| 22 | V22 | 單字匹配的百分比 | 數值型 | 單字為「font」 |
| 23 | V23 | 單字匹配的百分比 | 數值型 | 單字為「000」 |
| 24 | V24 | 單字匹配的百分比 | 數值型 | 單字為「money」 |
| 25 | V25 | 單字匹配的百分比 | 數值型 | 單字為「hp」 |
| 26 | V26 | 單字匹配的百分比 | 數值型 | 單字為「hpl」 |
| 27 | V27 | 單字匹配的百分比 | 數值型 | 單字為「george」 |
| 28 | V28 | 單字匹配的百分比 | 數值型 | 單字為「650」 |
| 29 | V29 | 單字匹配的百分比 | 數值型 | 單字為「lab」 |
| 30 | V30 | 單字匹配的百分比 | 數值型 | 單字為「labs」 |
| 31 | V31 | 單字匹配的百分比 | 數值型 | 單字為「telnet」 |
| 32 | V32 | 單字匹配的百分比 | 數值型 | 單字為「857」 |
| 33 | V33 | 單字匹配的百分比 | 數值型 | 單字為「data」 |
| 34 | V34 | 單字匹配的百分比 | 數值型 | 單字為「415」 |
| 35 | V35 | 單字匹配的百分比 | 數值型 | 單字為「85」 |
| 36 | V36 | 單字匹配的百分比 | 數值型 | 單字為「technology」 |
| 37 | V37 | 單字匹配的百分比 | 數值型 | 單字為「1999」 |
| 38 | V38 | 單字匹配的百分比 | 數值型 | 單字為「parts」 |
| 39 | V39 | 單字匹配的百分比 | 數值型 | 單字為「pm」 |
| 40 | V40 | 單字匹配的百分比 | 數值型 | 單字為「direct」 |
| 41 | V41 | 單字匹配的百分比 | 數值型 | 單字為「cs」 |
| 42 | V42 | 單字匹配的百分比 | 數值型 | 單字為「meeting」 |
| 43 | V43 | 單字匹配的百分比 | 數值型 | 單字為「original」 |
| 44 | V44 | 單字匹配的百分比 | 數值型 | 單字為「project」 |
| 45 | V45 | 單字匹配的百分比 | 數值型 | 單字為「re」 |
| 46 | V46 | 單字匹配的百分比 | 數值型 | 單字為「edu」 |
| 47 | V47 | 單字匹配的百分比 | 數值型 | 單字為「table」 |
| 48 | V48 | 單字匹配的百分比 | 數值型 | 單字為「conference」 |
| 49 | V49 | 符號匹配的百分比 | 數值型 | 符號為「;」 |
| 50 | V50 | 符號匹配的百分比 | 數值型 | 符號為「(」 |
| 51 | V51 | 符號匹配的百分比 | 數值型 | 符號為「[」 |
| 52 | V52 | 符號匹配的百分比 | 數值型 | 符號為「!」 |
| 53 | V53 | 符號匹配的百分比 | 數值型 | 符號為「$」 |
| 54 | V54 | 符號匹配的百分比 | 數值型 | 符號為「#」 |
| 55 | V55 | 連續大寫字母序列的平均長度 | 數值型 |  |
| 56 | V56 | 最長的連續大寫字母序列的長度 | 數值型 |  |
| 57 | V57 | 所有連續大寫字母序列長度的總和 | 數值型 |  |
| 58 | y | 垃圾郵件判斷 | 數值型 | 1=是；0=否 |

表5.2為垃圾郵件資料集各變數變異數膨脹因子（Variance inflation factor），為了避免變數共線性的問題，使用邏輯斯迴歸計算各變數的變異數膨脹因子，由於連續大寫字母序列的平均長度（V55）與最長的連續大寫字母序列的長度（V56）具有較高的變異數膨脹因子，分別為為 6.0892 和 5.3020 ，這表明這兩個變數之間可能存在共線性問題。因此，決定將連續大寫字母序列的平均長度變數（V55）刪除，並重新計算。結果顯示，刪除 連續大寫字母序列的平均長度（V55） 後，最長的連續大寫字母序列的長度（V56）的降低變異數膨脹因子至 2.5347 ，已回到正常範圍，顯示出共線性問題得以改善。

表5.2 垃圾郵件資料集各變數變異數膨脹因子

|  |  |  |  |
| --- | --- | --- | --- |
| 編號 | 變數名稱 | 變異數膨脹因子 | 變異數膨脹因子（移除V55） |
| 1 | V1 | 1.3687 | 1.3586 |
| 2 | V2 | 1.0236 | 1.0221 |
| 3 | V3 | 1.0531 | 1.0511 |
| 4 | V4 | 1.0116 | 1.0116 |
| 5 | V5 | 1.0546 | 1.0544 |
| 6 | V6 | 1.0578 | 1.0474 |
| 7 | V7 | 1.0586 | 1.0584 |
| 8 | V8 | 1.0652 | 1.0654 |
| 9 | V9 | 1.1345 | 1.1042 |
| 10 | V10 | 1.0388 | 1.0371 |
| 11 | V11 | 1.4760 | 1.4697 |
| 12 | V12 | 1.105 | 1.1024 |
| 13 | V13 | 1.0535 | 1.0531 |
| 14 | V14 | 1.0359 | 1.0347 |
| 15 | V15 | 1.0608 | 1.0604 |
| 16 | V16 | 1.1127 | 1.1123 |
| 17 | V17 | 1.0911 | 1.0886 |
| 18 | V18 | 1.0957 | 1.0902 |
| 19 | V19 | 1.2165 | 1.2161 |
| 20 | V20 | 1.1866 | 1.1781 |
| 21 | V21 | 1.4014 | 1.3980 |
| 22 | V22 | 2.8747 | 2.8932 |
| 23 | V23 | 1.0708 | 1.0693 |
| 24 | V24 | 1.056 | 1.0563 |
| 25 | V25 | 1.5359 | 1.5317 |
| 26 | V26 | 1.4926 | 1.4903 |
| 27 | V27 | 3.1748 | 2.1941 |
| 28 | V28 | 1.1207 | 1.1196 |
| 29 | V29 | 1.0127 | 1.0127 |
| 30 | V30 | 1.0411 | 1.0393 |
| 31 | V31 | 1.0134 | 1.0146 |
| 32 | V32 | 1.0248 | 1.0283 |
| 33 | V33 | 1.1155 | 1.1122 |
| 34 | V34 | 1.0162 | 1.0185 |
| 35 | V35 | 1.0930 | 1.0915 |
| 36 | V36 | 1.1512 | 1.1506 |
| 37 | V37 | 1.0673 | 1.0661 |
| 38 | V38 | 1.0295 | 1.0290 |
| 39 | V39 | 1.0660 | 1.0660 |
| 40 | V40 | 1.0615 | 1.0610 |
| 41 | V41 | 1.0433 | 1.0409 |
| 42 | V42 | 1.0469 | 1.0481 |
| 43 | V43 | 1.0178 | 1.0177 |
| 44 | V44 | 1.1539 | 1.1484 |
| 45 | V45 | 1.1481 | 1.1482 |
| 46 | V46 | 1.0609 | 1.0598 |
| 47 | V47 | 1.0313 | 1.0305 |
| 48 | V48 | 1.0684 | 1.0659 |
| 49 | V49 | 2.4571 | 2.4688 |
| 50 | V50 | 1.0808 | 1.0798 |
| 51 | V51 | 1.0227 | 1.0219 |
| 52 | V52 | 1.0627 | 1.0625 |
| 53 | V53 | 1.1985 | 1.2015 |
| 54 | V54 | 1.4730 | 1.4776 |
| 55 | V55 | 6.0892 |  |
| 56 | V56 | 5.3020 | 2.5347 |
| 57 | V57 | 1.6845 | 1.5274 |

為了避免共線性的問題，對資料集做主成分分析，表5.3垃圾郵件資料集主成分解釋變異表，圖5.1及圖5.2是將表5.3進行視覺化。從圖5.1可以看到，解釋變異在第4個主成分時已經趨於平緩；而根據表5.3與圖5.2顯示，累積解釋變異在第39個主成分時達到了85%以上。因此，選擇了前39個主成分作為分析的基礎。

表5.3 垃圾郵件資料集主成分解釋變異表

|  |  |  |  |
| --- | --- | --- | --- |
| 主成分 | 標準差 | 解釋變異 | 累積解釋變異 |
| PC1 | 2.5675 | 0.1157 | 0.1157 |
| PC2 | 1.8076 | 0.0573 | 0.1730 |
| PC3 | 1.4153 | 0.0351 | 0.2081 |
| PC4 | 1.2701 | 0.0283 | 0.2364 |
| PC5 | 1.2435 | 0.0271 | 0.2635 |
| PC6 | 1.2094 | 0.0257 | 0.2892 |
| PC7 | 1.1891 | 0.0248 | 0.3140 |
| PC8 | 1.1726 | 0.0241 | 0.3381 |
| PC9 | 1.1381 | 0.0227 | 0.3608 |
| PC10 | 1.1300 | 0.0224 | 0.3833 |
| PC11 | 1.1031 | 0.0214 | 0.4046 |
| PC12 | 1.0631 | 0.0198 | 0.4244 |
| PC13 | 1.0544 | 0.0195 | 0.4439 |
| PC14 | 1.0465 | 0.0192 | 0.4632 |
| PC15 | 1.0426 | 0.0191 | 0.4822 |
| PC16 | 1.0312 | 0.0187 | 0.5009 |
| PC17 | 1.0240 | 0.0184 | 0.5193 |
| PC18 | 1.0116 | 0.0180 | 0.5372 |
| PC19 | 1.0063 | 0.0178 | 0.5550 |
| PC20 | 1.0014 | 0.0176 | 0.5726 |
| PC21 | 0.9980 | 0.0175 | 0.5901 |
| PC22 | 0.9891 | 0.0172 | 0.6072 |
| PC23 | 0.9821 | 0.0169 | 0.6241 |
| PC24 | 0.9702 | 0.0165 | 0.6407 |
| PC25 | 0.9678 | 0.0164 | 0.6571 |
| PC26 | 0.9614 | 0.0162 | 0.6733 |
| PC27 | 0.9567 | 0.0161 | 0.6894 |
| PC28 | 0.9511 | 0.0159 | 0.7052 |
| PC29 | 0.9346 | 0.0153 | 0.7206 |
| PC30 | 0.9306 | 0.0152 | 0.7358 |
| PC31 | 0.9146 | 0.0147 | 0.7504 |
| PC32 | 0.9092 | 0.0145 | 0.7649 |
| PC33 | 0.8931 | 0.0140 | 0.7789 |
| PC34 | 0.8841 | 0.0137 | 0.7926 |
| PC35 | 0.8813 | 0.0136 | 0.8063 |
| PC36 | 0.8692 | 0.0132 | 0.8195 |
| PC37 | 0.8568 | 0.0129 | 0.8324 |
| PC38 | 0.8503 | 0.0127 | 0.8451 |
| PC39 | 0.8394 | 0.0124 | 0.8574 |
| PC40 | 0.8306 | 0.0121 | 0.8696 |
| PC41 | 0.8213 | 0.0118 | 0.8814 |
| PC42 | 0.8161 | 0.0117 | 0.8931 |
| PC43 | 0.7868 | 0.0109 | 0.9039 |
| PC44 | 0.7800 | 0.0107 | 0.9146 |
| PC45 | 0.7628 | 0.0102 | 0.9248 |
| PC46 | 0.7597 | 0.0101 | 0.9349 |
| PC47 | 0.7242 | 0.0092 | 0.9441 |
| PC48 | 0.6991 | 0.0086 | 0.9527 |
| PC49 | 0.6711 | 0.0079 | 0.9606 |
| PC50 | 0.6396 | 0.0072 | 0.9678 |
| PC51 | 0.6129 | 0.0066 | 0.9744 |
| PC52 | 0.6047 | 0.0064 | 0.9808 |
| PC53 | 0.5787 | 0.0059 | 0.9867 |
| PC54 | 0.5526 | 0.0054 | 0.992 |
| PC55 | 0.5103 | 0.0046 | 0.9966 |
| PC56 | 0.4362 | 0.0033 | 0.9999 |
| PC57 | 0.0621 | 0.0001 | 1.0000 |

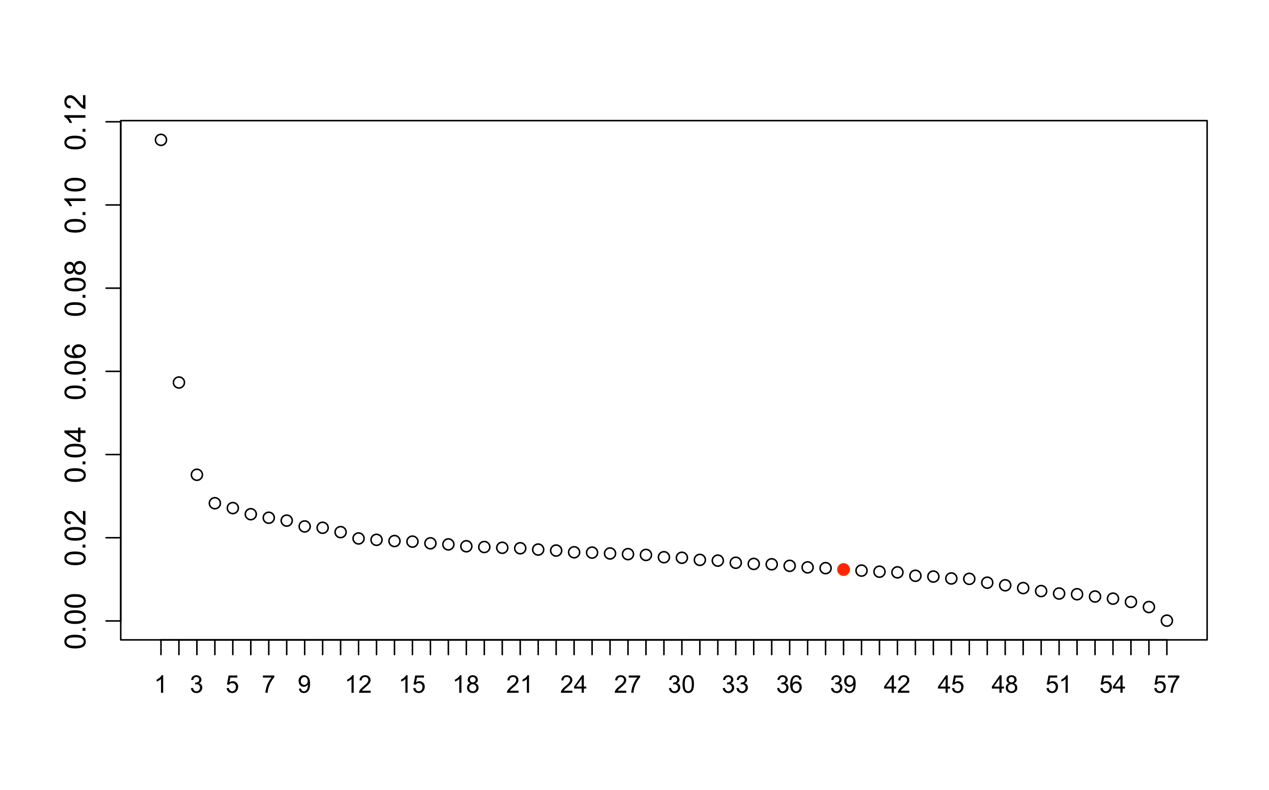


圖5.1 垃圾郵件資料集主成分解釋變異（x軸爲主成分，y軸爲解釋變異）

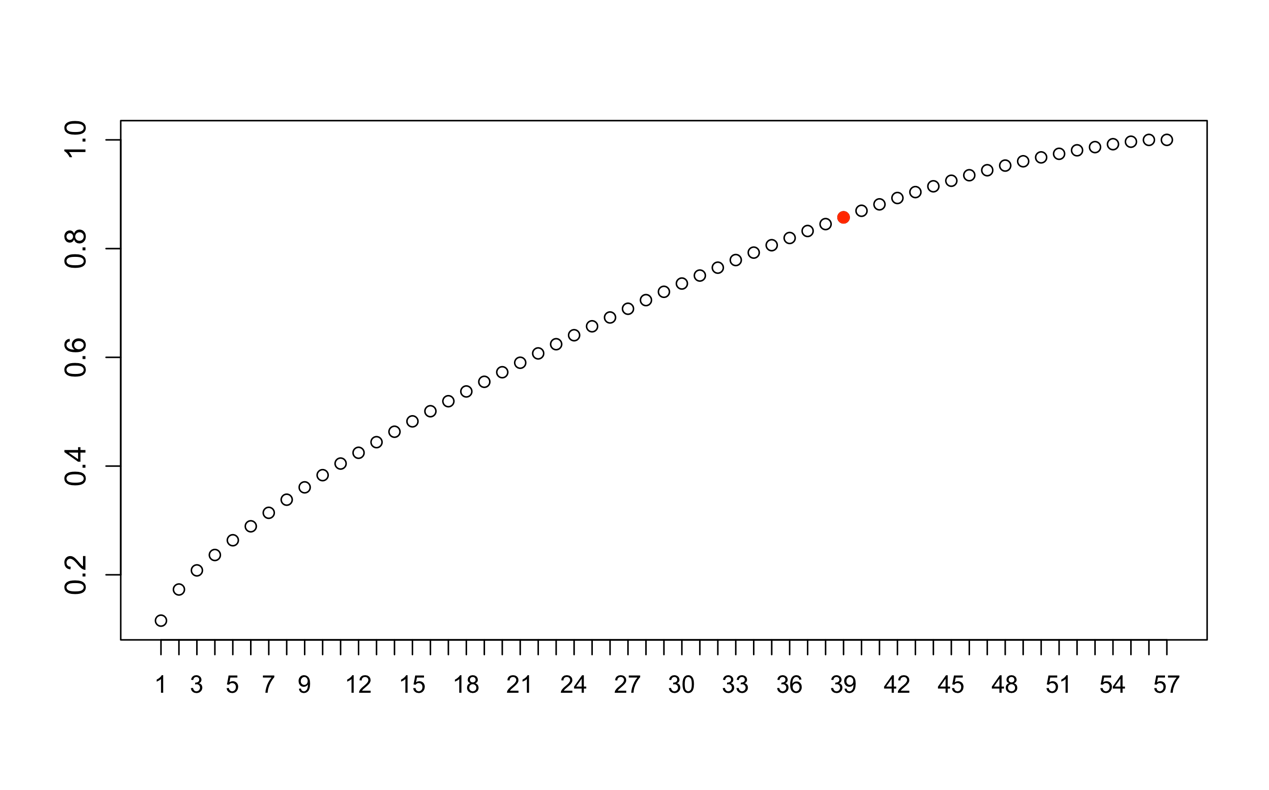


圖5.2 垃圾郵件資料集主成分累積解釋變異（x軸爲主成分，  
y軸爲累積解釋變異）

第二組資料為銀行定期存款資料集，同樣來源自 UCI 機器學習資料庫，由葡萄牙中央銀行（Banco de Portugal）提供。資料時間範圍為 2008 年 5 月至 2010 年 11 月，原始資料共有 41,188 筆觀察值，並從中隨機抽取 10% 作為本次分析所使用的樣本，以利於機器學習模型之訓練與測試。

表5.4為銀行定期存款資料集的變數說明，此資料集共包含 20 個變數，其中有 10 個數值型變數與 11 個類別型變數。資料集中共計 4,119 筆觀察值，其中標記為未訂閱定期存款（記作 0）者有 3,668 筆，占整體資料約 89.1%；標記為訂閱定期存款（記作 1）者有 451 筆，占整體資料約 10.9%。

由於職業、婚姻狀況等類別型變數中存在遺失值，考量到資料集中觀察值數量充足，因此採取刪除含有遺失值的觀察值作為處理方式。刪除1,029筆含有遺失值的觀察值後，資料集剩餘 3,090 筆觀察值，其中未訂閱定期存款（記作 0）者共有 2,720 筆，占整體資料約 88.0%；訂閱定期存款（記作 1）者共有 370 筆，占整體資料約 12.0%。

表5.4 銀行定期存款資料集變數說明

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 變數型態 | 備註 |
| 1 | age | 年齡 | 數值型 |  |
| 2 | job | 職業 | 類別型 | 參照表5.5 |
| 3 | marital | 婚姻狀況 | 類別型 | 參照表5.6 |
| 4 | education | 教育程度 | 類別型 | 參照表5.7 |
| 5 | default | 信用違約 | 二元型 | 參照表5.8 |
| 6 | housing | 房貸 | 類別型 | 參照表5.9 |
| 7 | loan | 個人貸款 | 二元型 | 參照表5.10 |
| 8 | contact | 聯絡方式 | 類別型 | 參照表5.11 |
| 9 | month | 最後連絡日期(月) | 類別型 | 參照表5.12 |
| 10 | day\_of\_week | 最後聯絡為星期幾 | 類別型 | 參照表5.13 |
| 11 | duration | 最後連絡持續時間 | 數值型 | 以秒為單位 |
| 12 | campaign | 活動期間聯絡次數 | 數值型 | 包括最後一次聯絡 |
| 13 | pdays | 自上次活動聯絡客戶以來經過的天數 | 數值型 | 999：表示之前未聯絡客戶 |
| 14 | previous | 本次活動之前為該客戶進行的聯繫次數 | 數值型 |  |
| 15 | poutcome | 先前行銷結果 | 類別型 | 參照表5.14 |
| 16 | emp.var.rate | 就業變動率 | 數值型 | 每季指標 |
| 17 | cons.price.idx | 消費者物價指數 | 數值型 | 每月指標 |
| 18 | cons.conf.idx | 消費者信心指數 | 數值型 | 每月指標 |
| 19 | euribor | 三個月期歐洲銀行間同業拆借利率 | 數值型 | 每日指標 |
| 20 | nr.employed | 就業人數 | 數值型 | 每季指標 |
| 21 | y | 訂閱定期存款 | 二元型 | 是否訂閱？ |

表5.5 職業（job）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 |
| 1 | admin | 行政人員 | 854 | 736 | 118 |
| 2 | blue-collar | 藍領工人 | 554 | 507 | 47 |
| 3 | entrepreneur | 企業家 | 107 | 100 | 7 |
| 4 | housemaid | 家政人員 | 75 | 66 | 9 |
| 5 | management | 管理人員 | 265 | 241 | 24 |
| 6 | retired | 退休人員 | 119 | 90 | 29 |
| 7 | self-employed | 自營業者 | 126 | 113 | 13 |
| 8 | services | 服務業 | 276 | 252 | 24 |
| 9 | student | 學生 | 55 | 41 | 14 |
| 10 | technician | 技術人員 | 573 | 504 | 69 |
| 11 | unemployed | 失業 | 86 | 70 | 16 |

表5.6 婚姻狀況（marital）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | divorced | 離婚（包含喪偶） | 348 | 310 | 38 | |
| 2 | married | 已婚 | 1,791 | 1,592 | 199 | |
| 3 | single | 未婚 | 951 | 818 | 133 | |

表5.7 教育程度（education）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | basic.4y | 四年基礎教育 | 243 | 218 | 25 | |
| 2 | basic.6y | 六年基礎教育 | 150 | 138 | 12 | |
| 3 | basic.9y | 九年基礎教育 | 407 | 371 | 36 | |
| 4 | high.school | 高中 | 728 | 641 | 87 | |
| 5 | illiterate | 文盲 | 1 | 1 | 0 | |
| 6 | professional.course | 技職教育 | 454 | 397 | 57 | |
| 7 | university.degree | 大學教育 | 1,107 | 954 | 153 | |

表5.8 信用違約（default）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | no | 無信用違約 | 3,089 | 2,719 | 370 | |
| 2 | yes | 有信用違約 | 1 | 1 | 0 | |

表5.9 房貸（housing）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | no | 無房屋貸款 | 1,402 | 1,239 | 163 | |
| 2 | yes | 有房屋貸款 | 1,688 | 1,481 | 207 | |

表5.10 個人貸款（loan）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | no | 無個人貸款 | 2,269 | 2,269 | 314 | |
| 2 | yes | 有個人貸款 | 507 | 451 | 56 | |

表5.11 聯絡方式（contract）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | cellular | 行動電話 | 2,108 | 1,787 | 321 | |
| 2 | telephone | 固定電話 | 982 | 933 | 49 | |

表5.12 最後連絡日期(月)（month）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | mar | 三月 | 42 | 18 | 24 | |
| 2 | apr | 四月 | 169 | 138 | 31 | |
| 3 | may | 五月 | 981 | 911 | 70 | |
| 4 | jun | 六月 | 365 | 314 | 51 | |
| 5 | jul | 七月 | 514 | 471 | 43 | |
| 6 | aug | 八月 | 495 | 441 | 54 | |
| 7 | sep | 九月 | 57 | 35 | 22 | |
| 8 | oct | 十月 | 61 | 37 | 24 | |
| 9 | nov | 十一月 | 387 | 345 | 42 | |
| 10 | dec | 十二月 | 19 | 10 | 9 | |

表5.13 最後聯絡為星期幾（day\_of\_week）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | mon | 星期一 | 642 | 558 | 84 | |
| 2 | tue | 星期二 | 613 | 543 | 70 | |
| 3 | wed | 星期三 | 625 | 560 | 65 | |
| 4 | thu | 星期四 | 630 | 548 | 82 | |
| 5 | fri | 星期五 | 580 | 511 | 69 | |

表5.14 先前行銷結果（poutcome）變數說明

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 編號 | 變數名稱 | 意義 | 筆數 | 反應變數筆數 | |
| 0 | 1 | |
| 1 | failure | 失敗 | 360 | 304 | 56 | |
| 2 | success | 成功 | 128 | 46 | 82 | |
| 3 | nonexistent | 不存在 | 2,602 | 2,370 | 232 | |

進一步處理資料集，首先處理數值型變數，「聯絡持續時間變數（duration）」對目標變數有高度影響，但在實際應用情境中，通話持續時間在通話結束前是未知的。因此，在建立符合真實情境的預測模型時，應避免使用此變數，故決定將其移除；另外，另外，「自上次活動聯絡客戶以來經過的天數（pdays）」中，將表示之前未聯絡客戶的 999 改成 -1，原因是 999 並非實際天數，若直接保留，會使模型誤認為是極大值，影響數值型變數的分布與解讀。改為 -1，可明確表示「未曾聯絡」的情況，且在後續進行標準化時，避免因極端異常值而造成平均數和標準差嚴重偏移，進而影響模型表現。

為避免共線性問題，本研究針對連續型變數進行主成分分析，包括：「年齡（age）」、「活動期間聯絡次數（campaign）」、「自上次活動聯絡客戶以來經過的天數（pdays）」、「本次活動之前對該客戶進行的聯繫次數（previous）」、「就業變動率（emp.var.rate）」、「消費者物價指數（cons.price.idx）」、「消費者信心指數（cons.conf.idx）」、「三個月期歐洲銀行間同業拆借利率（euribor）」與「就業人數（nr.employed）」。本研究將累積解釋變異設定為約90%。表5.15為銀行定期存款資料集的主成分解釋變異表，圖5.3與圖5.4則將該表加以視覺化。

從圖5.3可觀察到，自第四個主成分之後，解釋變異的變化趨於平緩。然而，根據表5.15與圖5.3顯示，在第四個主成分時，累積解釋變異僅達約0.7923，尚未達到理想的資訊保留比例；直到第五個主成分時，累積解釋變異才達到約0.8936。因此，本研究保留前五個主成分作為後續分析之用。

表5.15 銀行定期存款資料集主成分解釋變異表

|  |  |  |  |
| --- | --- | --- | --- |
| 主成分 | 標準差 | 解釋變異 | 累積解釋變異 |
| PC1 | 1.9411 | 0.4186 | 0.4186 |
| PC2 | 1.1580 | 0.1490 | 0.5676 |
| PC3 | 1.0375 | 0.1196 | 0.6872 |
| PC4 | 0.9726 | 0.1051 | 0.7923 |
| PC5 | 0.9549 | 0.1013 | 0.8936 |
| PC6 | 0.6867 | 0.0524 | 0.9460 |
| PC7 | 0.6706 | 0.0500 | 0.9960 |
| PC8 | 0.1582 | 0.0028 | 0.9988 |
| PC9 | 0.1050 | 0.0012 | 1.0000 |

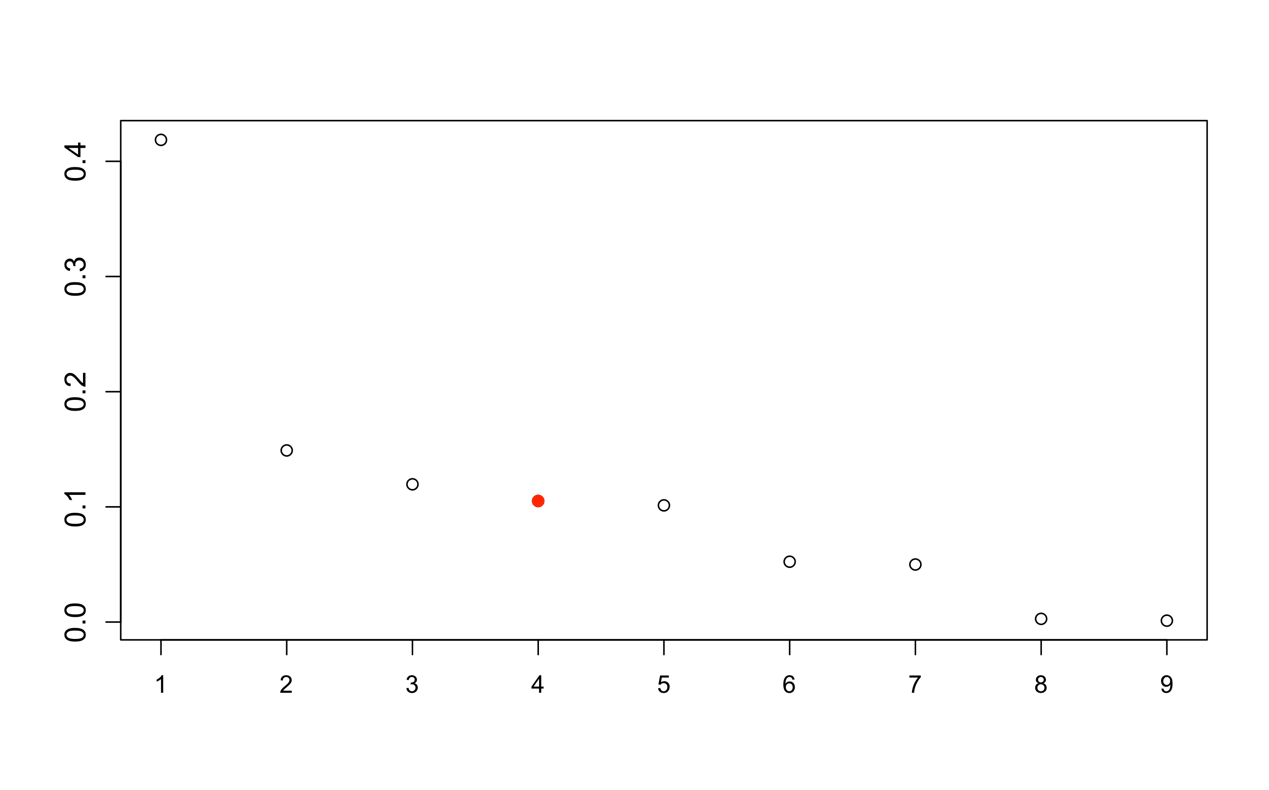


圖5.3 銀行定期存款資料集主成分解釋變異（x軸爲主成分，y軸爲解釋變異）

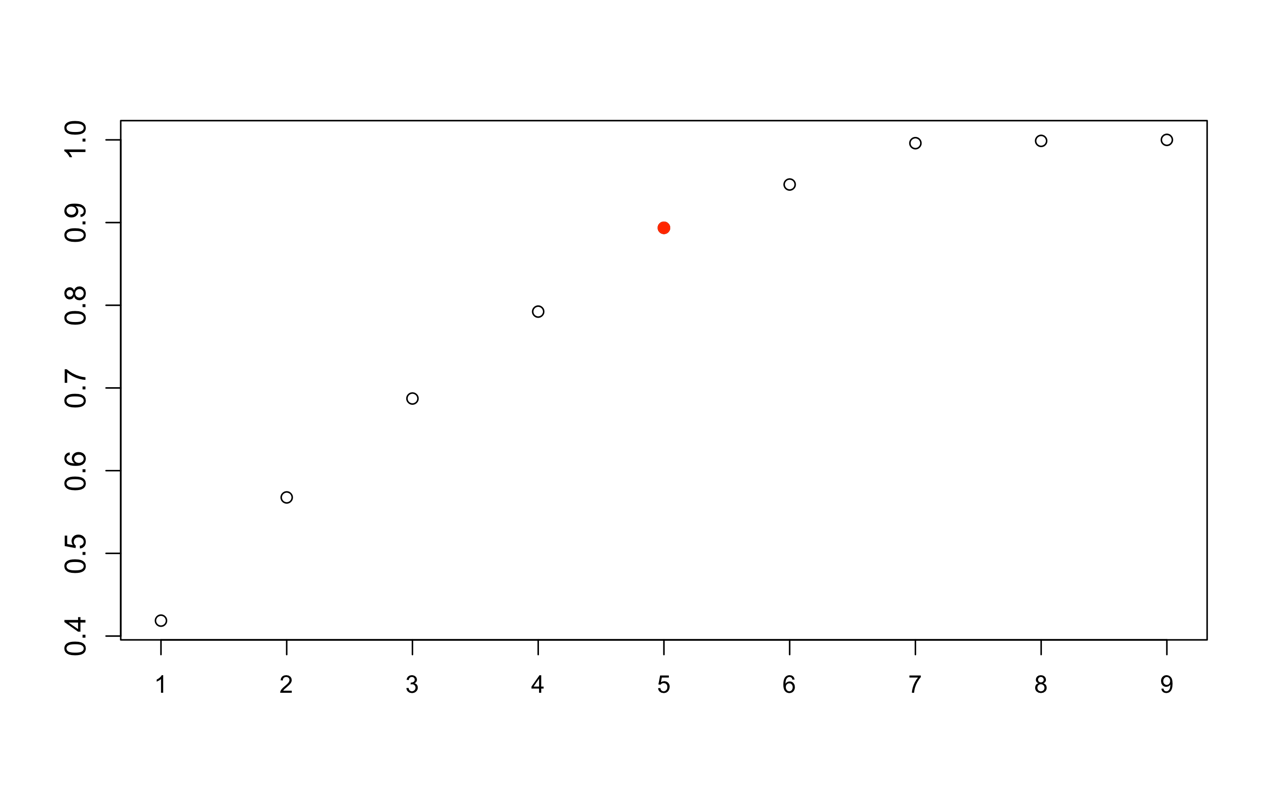


圖5.4 銀行定期存款資料集主成分累積解釋變異（x軸爲主成分，  
y軸爲累積解釋變異）

針對銀行定期存款資料集中的連續型變數進行主成分分析，各主成分之變數負荷量如表5.16所示。

由結果可知，第一主成分主要由「三個月期歐洲銀行間同業拆借利率（euribor3m）」、「就業變動率（emp.var.rate）」、「就業人數（nr.employed）」以及「消費者物價指數（cons.price.idx）」等變數構成，反映整體經濟環境與就業市場的變動，因此命名為「整體經濟與就業條件」。

第二主成分則主要由「自上次活動聯絡客戶以來經過的天數（pdays）」、「本次活動之前對該客戶進行的聯繫次數（previous）」、「消費者物價指數（cons.price.idx）」及「消費者信心指數（cons.conf.idx）」所組成，綜合描述客戶聯絡經驗與市場信心等因素，故命名為「客戶聯絡經驗與市場信心」。

第三主成分受到「年齡（age）」與「消費者信心指數（cons.conf.idx）」的強烈影響，顯示此成分與客戶的年齡結構及其對經濟環境的信心有關，因此命名為「年齡與消費信心」。

第四主成分幾乎完全由「活動期間聯絡次數（campaign）」所主導，反映出行銷聯絡頻率對該主成分的影響，故命名為「行銷聯絡強度」。

第五主成分則由「年齡（age）」與「消費者信心指數（cons.conf.idx）」呈現相反方向的高負荷與高正負荷，加上「消費者物價指數（cons.price.idx）」也有一定影響，顯示該主成分與年齡結構差異與消費心理反應的交互作用有關，因此命名為「年齡差異下的市場反應」。

表5.16 銀行定期存款資料集各主成分各變數負荷量

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 變數 | PC1 | PC2 | PC3 | PC4 | PC5 |
| age | 0.0435 | 0.1735 | 0.6503 | -0.2403 | 0.6954 |
| campaign | -0.1239 | 0.0926 | -0.2487 | -0.9495 | -0.1108 |
| pdays | 0.2073 | 0.6526 | -0.1647 | 0.0811 | -0.0524 |
| previous | 0.3019 | 0.5053 | -0.1619 | 0.0577 | 0.0609 |
| emp.var.rate | -0.4955 | 0.177 | -0.0506 | 0.0845 | 0.0605 |
| cons.price.idx | -0.3453 | 0.3594 | -0.265 | 0.1102 | 0.2611 |
| cons.conf.idx | -0.0900 | 0.3183 | 0.6212 | -0.0468 | -0.6519 |
| euribor3m | -0.4992 | 0.1282 | 0.0536 | 0.0795 | -0.0285 |
| nr.employed | -0.4760 | -0.0431 | 0.0210 | 0.0547 | 0.009 |

為檢查資料的共線性問題，將經過主成分分析轉換得到的主成分與類別型變數結合後，一併進行共線性檢查。由於資料中包含類別型變數，因此採用「廣義變異數膨脹因子（Generalized Variance Inflation Factor, GVIF）」作為檢查指標，其結果如表5.17所示。從表5.17可以看出，各變數的GVIF值皆接近1，顯示各變數間並無嚴重共線性問題。

表5.17 銀行定期存款資料集各變數廣義變異數膨脹因子

|  |  |  |  |
| --- | --- | --- | --- |
| 變數 | 廣義變異數膨脹因子 （GVIF） | 自由度 | GVIF^(1/(2\*Df)) |
| PC1 | 2.9119 | 1 | 1.7064 |
| PC2 | 3.1541 | 1 | 1.7760 |
| PC3 | 2.1938 | 1 | 1.4812 |
| PC4 | 1.2165 | 1 | 1.1029 |
| job | 4.9307 | 10 | 1.0830 |
| marital | 1.3424 | 2 | 1.0764 |
| education | 3.1778 | 6 | 1.1011 |
| default | 1.0000 | 1 | 1.0000 |
| housing | 1.0375 | 1 | 1.0186 |
| loan | 1.0279 | 1 | 1.0139 |
| contact | 1.6207 | 1 | 1.2730 |
| month | 4.7585 | 9 | 1.0905 |
| day\_of\_week | 1.1193 | 4 | 1.0142 |
| poutcome | 3.7788 | 2 | 1.3942 |

接著處理類別型變數。由於類神經網路與零膨脹伯努利模型（需將類別型變數轉換為虛擬變數（dummy variables），然而直接轉換可能導致維度過高，影響模型的穩定性與解釋性。因此，將根據主觀判斷進行適度合併，並透過似然比檢定進行檢驗，以確認合併後的變數組成是否合適。

根據表5.8，資料集中僅有一位樣本出現信用違約紀錄。由於該類別樣本數過少，無法作為有效的預測依據，因此將此觀察值排除。然而，「信用違約（default）」屬於二元變數，刪除該觀察值後，該變數僅剩單一類別，亦失去分析意義，因此一併將「信用違約（default）」變數從模型中刪除。

根據表5.7，依照葡萄牙對於教育程度的分類，將「四年基礎教育（basic4y）」、「六年基礎教育（basic6y）」、「九年基礎教育（basic9y）」及「文盲（illiterate）」歸類為「初等教育（primary）」；將「高中教育（high.school）」及「技職教育（professional.course）」歸類為「中等教育（secondary）」；而「大學教育（tertiary）」則歸類為「高等教育（tertiary）」。

為檢驗重新分類後的教育變數是否適合，配合邏輯斯迴歸模型進行似然比檢定（Likelihood Ratio Test），結果得到 p 值為 0.8416，遠大於 0.05，顯示合併後的教育層級變數與原變數相比並未顯著降低模型的適配性，因此重新分類是合理且可接受的。