EDA

月底資訊變數：

|  |  |  |
| --- | --- | --- |
| 月底資訊 | 月底專案實收總資費級距(7階) | RATE\_PLAN\_RNG |
| 月底資訊 | 月底專案原訂語音數據資費 | MONTHLY\_RATE\_S |
| 月底資訊 | 月底專案原訂數據資費 | MONTHLY\_RATE\_I |
| 月底資訊 | 月底專案原訂語音資費 | MONTHLY\_RATE\_V |
| 月底資訊 | 月底專案原定總資費 | TOTAL\_MONTHLY\_RATE |
| 月底資訊 | 月底專案實收語音數據資費 | MONTHLY\_FEE\_S |
| 月底資訊 | 月底專案實收數據資費 | MONTHLY\_FEE\_I |
| 月底資訊 | 月底專案實收語音資費 | MONTHLY\_FEE\_V |
| 月底資訊 | 月底專案實收總資費 | TOTAL\_MONTHLY\_FEE |
| 月底資訊 | 月底專案語音吃到飽 | VOICE\_UNLIM |

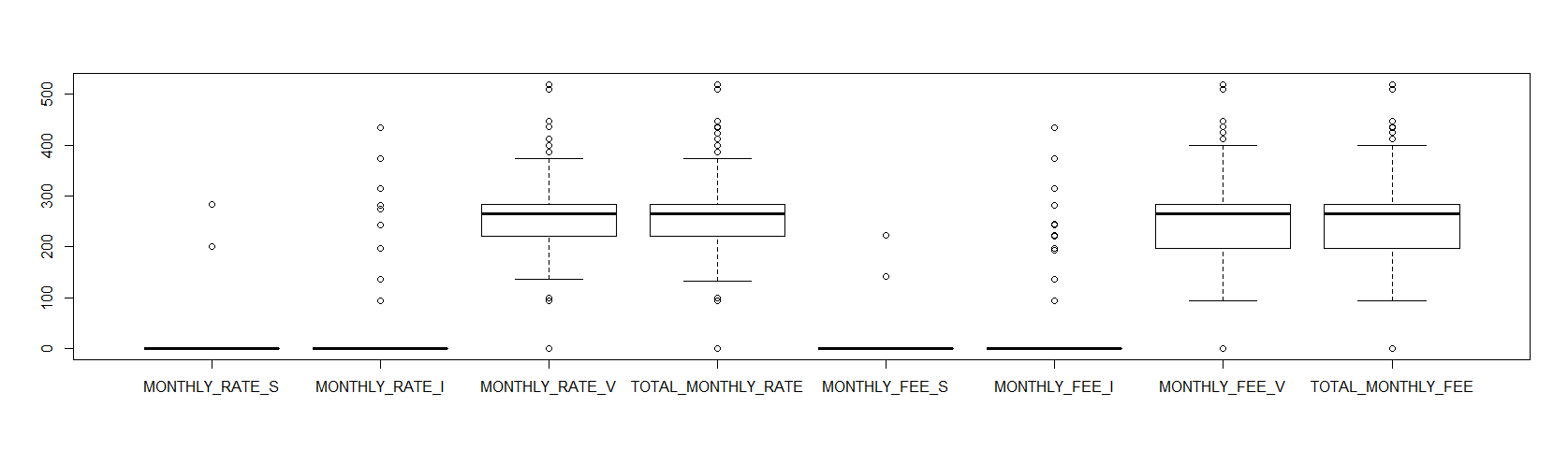
分群變數：

|  |  |  |
| --- | --- | --- |
| 預測目標 | 是否為壞用戶 | FLAG |
| 繳款資訊 | 近一個月是否曾欠費 | N1\_OWE\_MNY\_FLAG |
| 繳款資訊 | 近一個月是否曾溢繳 | N1\_OVER\_PAY\_FLAG |
| DCB用戶資訊 | 信用評級 | CURRENT\_LEVEL |

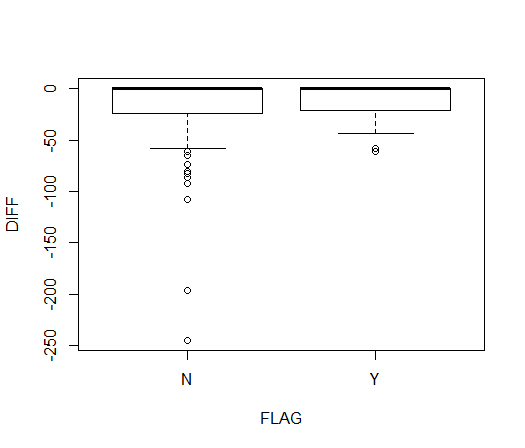
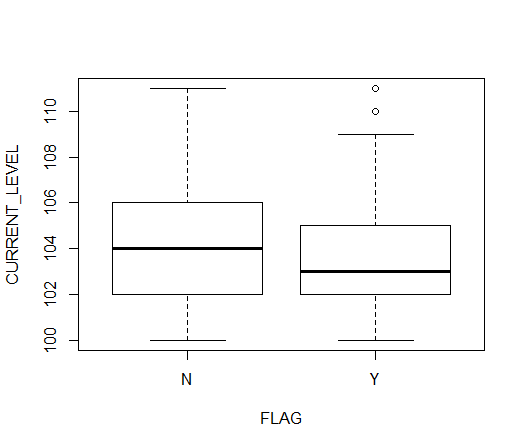
衍生變數：

|  |  |  |
| --- | --- | --- |
| 月底資訊 | 月底專案實收總資費 - 月底專案原定總資費 | DIFF |

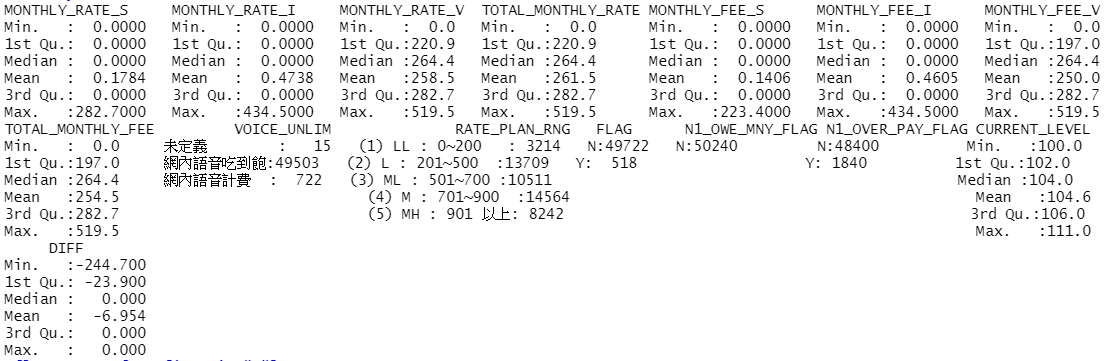
1. 使用「原訂」和「實收」的差距作為變數的原因是我認為最後實際收到的錢和預計要收的錢的差距會影響用戶分群。
2. 針對不同資費數據和「原訂」和「實收」兩類比較，繪製boxplot。
3. *MONTHLY\_RATE\_S*, *MONTHLY\_RATE\_I*, *MONTHLY\_FEE\_S*和*MONTHLY\_FEE\_I* 大部分數據都為0，只有少部分數據大於0。
4. *MONTHLY\_RATE\_V, TOTAL\_MONTHLY\_RATE, MONTHLY\_FEE\_V*和*TOTAL\_MONTHLY\_FEE*大部分數據集中在150和350之間。

**

1. 分別針對*FLAG*和*CURRENT\_LEVEL*, *DIFF*做Boxplot
2. 壞用戶(*FLAG*=Y)的信用評級較低。
3. 「原訂」和「實收」金額差異對於好壞用戶差異不大，因為差距大部分是0。



1. 看summary report
2. *MONTHLY\_RATE\_S*, *MONTHLY\_RATE\_I*, *MONTHLY\_FEE\_S*和*MONTHLY\_FEE\_I*有超過3/4的數據為0。
3. *MONTHLY\_RATE\_V, TOTAL\_MONTHLY\_RATE, MONTHLY\_FEE\_V*和*TOTAL\_MONTHLY\_FEE*大部分數據集中在150和350之間。
4. *VOICE\_UNLIM* 數據超過98%都是網內語音吃到飽。
5. *N1\_OWE\_MNY\_FLAG* 數據全是沒有(N)。
6. *N1\_OVER\_PAY\_FLAG* 數據超過96%都是沒有(N)。
7. *CURRENT\_LEVEL*分布在100到111之間。
8. *DIFF*有超過一半的數據都是0，也就是「原訂」和「實收」金額一樣。
9. *FLAG*為Y的比例約為1%，也就是壞用戶比例約為1%。



1. 結論
2. 因為*N1\_OWE\_MNY\_FLAG*數據全是沒有(N)，所以未來分析要去掉這個變數。
3. 因為*MONTHLY\_RATE\_S*, *MONTHLY\_RATE\_I*, *MONTHLY\_FEE\_S*和*MONTHLY\_FEE\_I* 大部分數據都為0，所以未來分析可能可以不用列入考慮。
4. *DIFF*分布對於*FLAG*的影響有限。