資料介紹

名稱：Telcom Customer Churn

簡介：本資料為某家電信公司下的客戶各項數據，內容包含客戶的基本資料、訂購的各項服務以及費用，還有客戶是否離開電信公司。

研究目的

1. Supervised Learning：
   將其中一變數 Contract(合約長度)設為 label class，然後藉由各個變數的數據來對客戶的下一個合約長度分類做預測。
2. Unsupervised Learning
   我們藉由分群，試著找出同群內客戶之間的共通點，藉此來制定銷售策略。

切割資料

我們發現合約長度為一個月的樣本有 3875 個，長度為一年的有 1473 個，長度為 2 年的有 1695 個。

Contract Type	#Sample
Month-to-month	3875
One year	1473
Two year	1695

我們希望讓訓練集跟測試集樣本個數比例為 7:3，因此我們根據不同長度的合約類型分三層，進行分層抽樣得到的測試集和訓練集個數如下

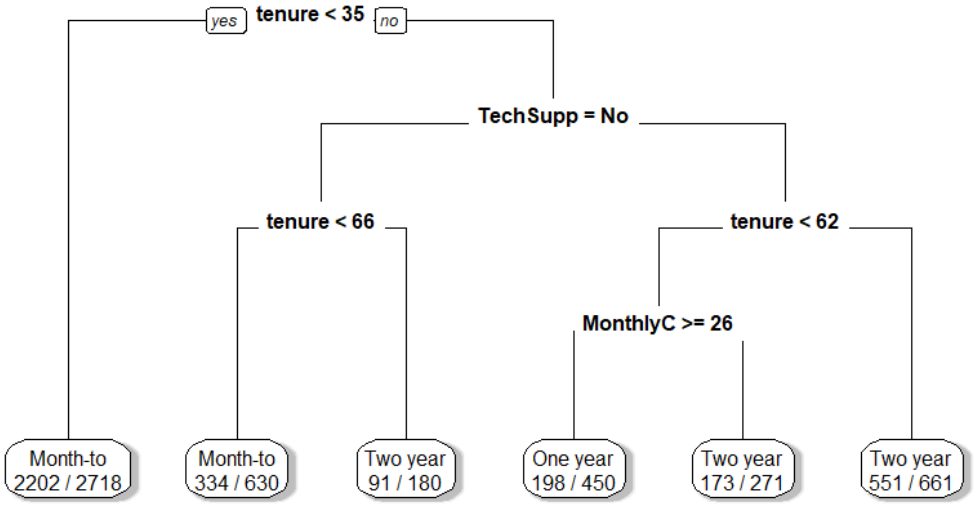
Contract Type	Training set sample size	Testing set sample size
Month-to-month	2700	1175
One year	1030	443
Two year	1180	515

我們就使用這些訓練集和測試集進行不同的分類，包括 Decision Tree, Naïve Bayes, Bagging 三種方法。

分類

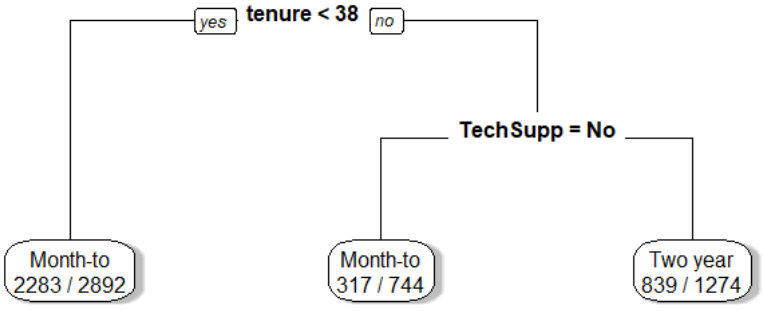
1. Decision Tree

我們拿訓練集生成的 decision tree 如下，其訓練集準確度到達 72%



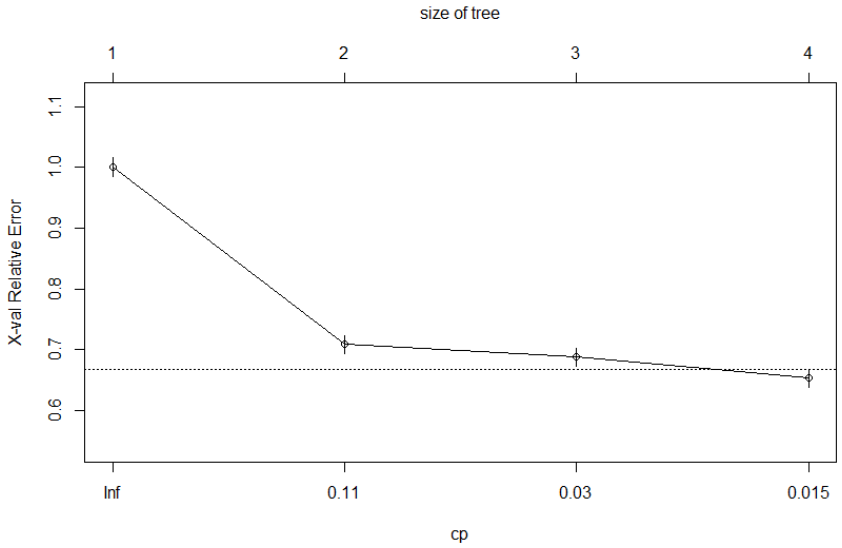

```
Classification tree:
rpart(formula = Contract ~ ., data = training_set, method = "class",
    control = telcom_data.control)

Variables actually used in tree construction:
[1] TechSupport tenure

Root node error: 2210/4910 = 0.4501

n= 4910

        CP nsplit rel error
1 0.273303      0   1.00000
2 0.055204      1   0.72670
3 0.030769      2   0.67149
4 0.010000      3   0.64072
```

接下來，我們選取 cp = 0.03 來做剪枝，剪枝完後的樹如下



最後我們用 10-fold CV 得到的準確度一樣為 1 - 0.63*0.45 = 72.1%，另外，我們根據 1-SE rule，即可找出 best tree size，即是 0.639+0.01 = 0.649 對應到的 n-split = 3，所以我們保留剪枝完的那棵樹，做為測試 testing set 的樹，訓練集的準確度為 1-0.643*0.45 = 71.0%

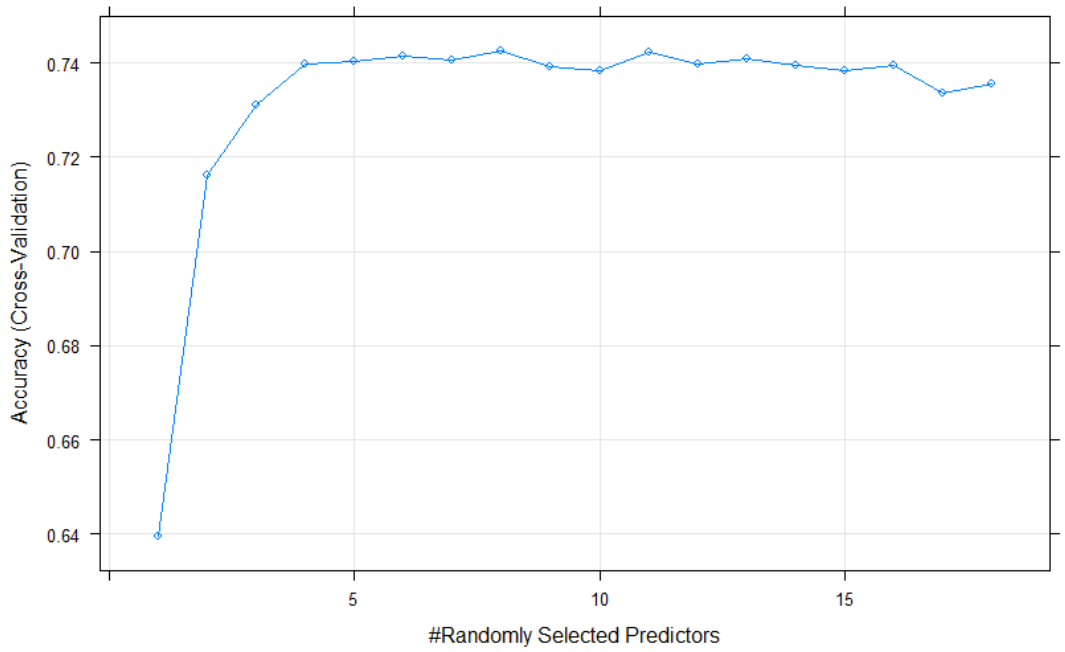
```
Classification tree:
rpart(formula = Contract ~ ., data = training_set, method = "class",
    parms = list(split = "information"), control = telcom_data.control)

Variables actually used in tree construction:
[1] TechSupport tenure

Root node error: 2210/4910 = 0.4501

n= 4910

        CP nsplit rel error  xerror     xstd
1 0.260181      0   1.00000 1.00000 0.015774
2 0.074208      1   0.73982 0.73982 0.014943
3 0.022624      2   0.66561 0.68100 0.014618
4 0.014480      3   0.64299 0.64977 0.014423
5 0.010000      4   0.62851 0.63801 0.014345
```

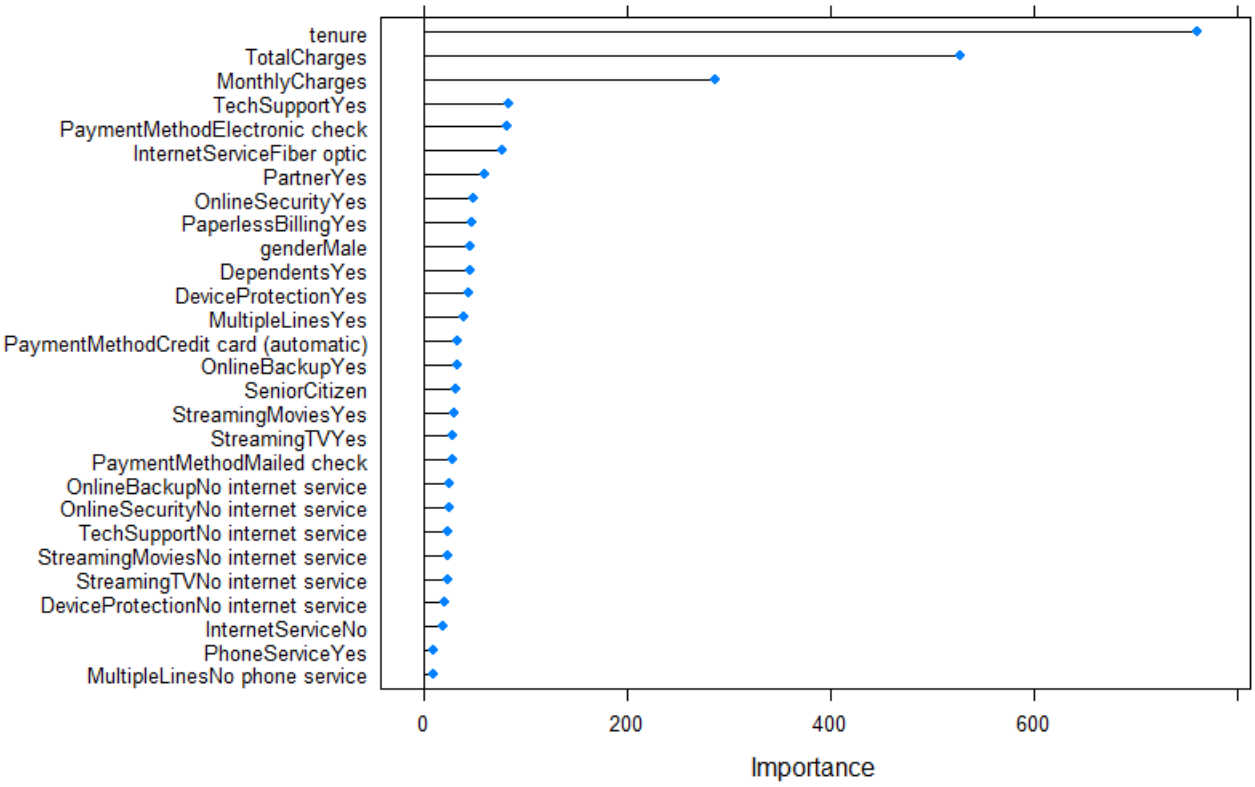


2. Random Forest

我們使用資料集裡面所有 18 個變數，產生 n = 500 的隨機森林的 CV 模型結果如下，最高的準確度大約為 74%，我們也可以知道，當放入第 5 個變數進去時，準確度已經不會再提升了
在測試集中的 Apparent accuracy 也有 90.33%，是所有我們試過的模型中表現最好的

```
                predictor
                 Month-to-month One year Two year
Month-to-month             2618       65       17
One year                    160      772       98
Two year                     48       87     1045
>
```


Accuracy (Cross-Validation)
0.74
0.72
0.70
0.68
0.66
0.64
5
10
15
#Randomly Selected Predictors


tenure
TotalCharges
MonthlyCharges
TechSupportYes
PaymentMethodElectronic check
InternetServiceFiber optic
PartnerYes
OnlineSecurityYes
PaperlessBillingYes
genderMale
DependentsYes
DeviceProtectionYes
MultipleLinesYes
PaymentMethodCredit card (automatic)
OnlineBackupYes
SeniorCitizen
StreamingMoviesYes
StreamingTVYes
PaymentMethodMailed check
OnlineBackupNo internet service
OnlineSecurityNo internet service
TechSupportNo internet service
StreamingMoviesNo internet service
StreamingTVNo internet service
DeviceProtectionNo internet service
InternetServiceNo
PhoneServiceYes
MultipleLinesNo phone service
0
200
400
600
Importance

3. Naïve Bayes

Naïve Bayes 的結果如下，apparent accuracy 為 68.66%

```
                 Telcom_data_NB_CV_pred
                  Month-to-month One year Two year
Month-to-month               2084      327      289
One year                      262      417      351
Two year                       46      264      870
>
```

測試集

我們使用表現最好的 Random Forest 做預測，以下是 Random Forest 在測試集的表現，準確度為 90%

```
                predictor
                 Month-to-month One year Two year
Month-to-month             1131       38        6
One year                     78      320       45
Two year                     19       29      467
>
```

分群

方法

在上課時有教過許多分群法，例如 ：k-means、Spectral Clustering、Kernel K-means、Mini Batch K-means 等…，然而由於這筆電信客戶資料 Telcom Customer Churn 的變數多半是類別型 (17 個類別變數，3 個數值變數)，無法計算歐式距離 (Euclidean distance)。

因此在使用上述方法時，要先定義距離矩陣，這次我們用上課教過的 Gower's coefficient 定義距離/不相似度。若是類別變數會看是否相同 ： 不同則定義距離為 1、相同則定義距離為 0，即$d_{r,s}^{f} = \begin{cases} 1; x_r^f \neq x_s^f \\ 0; x_r^f = x_s^f \end{cases}$為物件 r 及物件 j 在類別變數 f 的距離。而連續型變數則是直接取距離後，再除上該變數的全距，即 $d_{r,s}^{f} = \frac{|x_r^f - x_s^f|}{R^f}$，最後再將各變數的距離做加權平均 $d_{r,s} = \frac{\sum_f d_{r,s}^f}{\#\ f}$ 就是物件 r 與物件 j 的距離/不相似度。

算出距離矩陣後，透過 MDS (Multidimensional Scaling)將原本的資料(變數包含類別及連續)投影成皆為連續型變數的資料，投影過後各點距離矩陣會與原本的距離矩陣相似，雖然經 MDS 投影的資料代表性沒有原始資料好，但幫助我們解決類別變數無法計算歐式距離的問題，因此我們可以對投影後的資料做上述提到的分群法。

實際操作

首先要挑選進行分群的變數，由於我們的目的是要將客戶分群後制定適當的行銷策略，因此我們傾向挑選更接近客戶本質的變數 (例如 ： 性 別、已/未婚、當月帳單金額…)，而捨棄太細太雜的變數 (是否購買線上備份、網路電視等…)，最後選擇的變數有 ：**Gender：客戶性別(男、女)(類別變數) 、 Senior Citizens：是否為老年人(類別變數)、Partner：是否有伴侶(類別變數)、Dependence：是否有依附親屬(類別變數)、Tenure：客戶在公司下待了幾個月(數值變數)、Phone Service：是否有申請電話服務(類別變數)、Internet Service：是否有申請網路服務(DSL、光纖、無網路服務) (類別變數)、Contract： 合約長度(單月、一年、二年到期) (類別變數)、Paperless Billing： 是否使用無紙化帳單(類別變數)、Payment Method： 付款方式(銀行轉帳、信用卡、電子支票、郵寄支票) (類別變數)、Monthly Charges： 當月帳單金額(數值變數)**

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
1	Female	0	Yes	No	1	No
2	Male	0	No	No	34	Yes
3	Male	0	No	No	2	Yes
4	Male	0	No	No	45	No
5	Female	0	No	No	2	Yes

InternetService	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges
DSL	Month-to-month	Yes	Electronic check	29.85
DSL	One year	No	Mailed check	56.95
DSL	Month-to-month	Yes	Mailed check	53.85
DSL	One year	No	Bank transfer (automatic)	42.30
Fiber optic	Month-to-month	Yes	Electronic check	70.70

挑選變數後即可計算距離矩陣，從圖中可看出距離/不相似度介於 0 與 1 之間

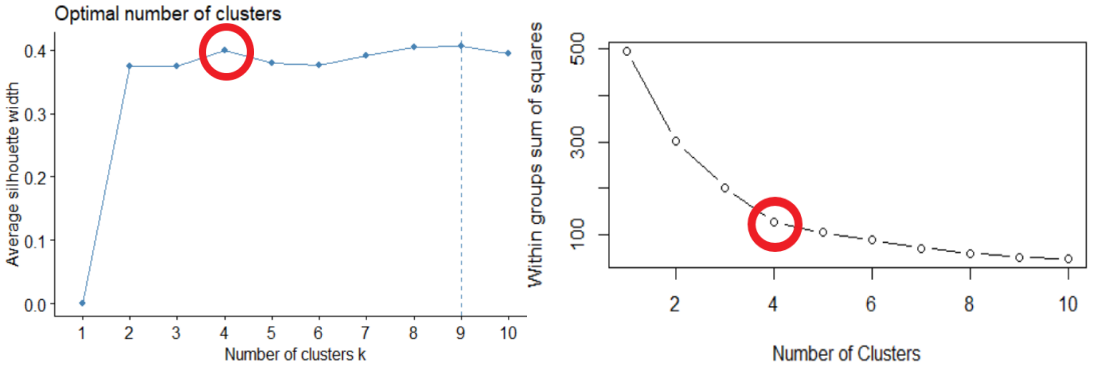
	1	2	3	4	5	6	7	8	9	10
1	0.0000000	0.6116350	0.38660862	0.5213629	0.310941505	0.34470451	0.6255654	0.2841814	0.28370647	0.7371740
2	0.6116350	0.0000000	0.22502638	0.2089590	0.507387306	0.52599879	0.4987788	0.3276346	0.59631389	0.2178954
3	0.3866086	0.2250264	0.00000000	0.4283771	0.287969245	0.32173225	0.3298658	0.3046284	0.44255239	0.4414744
4	0.5213629	0.2089590	0.42837705	0.0000000	0.625437208	0.64404870	0.6168287	0.3282715	0.71436379	0.2158111
5	0.3109415	0.5073873	0.28796924	0.6254372	0.000000000	0.03376300	0.3146239	0.4107794	0.15458314	0.6343736
6	0.3447045	0.5259988	0.32173225	0.6440487	0.033763003	0.00000000	0.2999472	0.4293909	0.12082014	0.6529851
7	0.6255654	0.4987788	0.32986582	0.6168287	0.314623850	0.29994723	0.0000000	0.6142922	0.38541384	0.4439469
8	0.2841814	0.3276346	0.30462837	0.3282715	0.410779436	0.42939092	0.6142922	0.0000000	0.54516056	0.5440826
9	0.2837065	0.5963139	0.44255239	0.7143638	0.154583145	0.12082014	0.3854138	0.5451606	0.00000000	0.7233002
10	0.7371740	0.2178954	0.44147445	0.2158111	0.634373587	0.65298507	0.4439469	0.5440826	0.72330017	0.0000000

做 MDS 投影(選擇維度為 2 維)，以這筆投影後的資料做分群

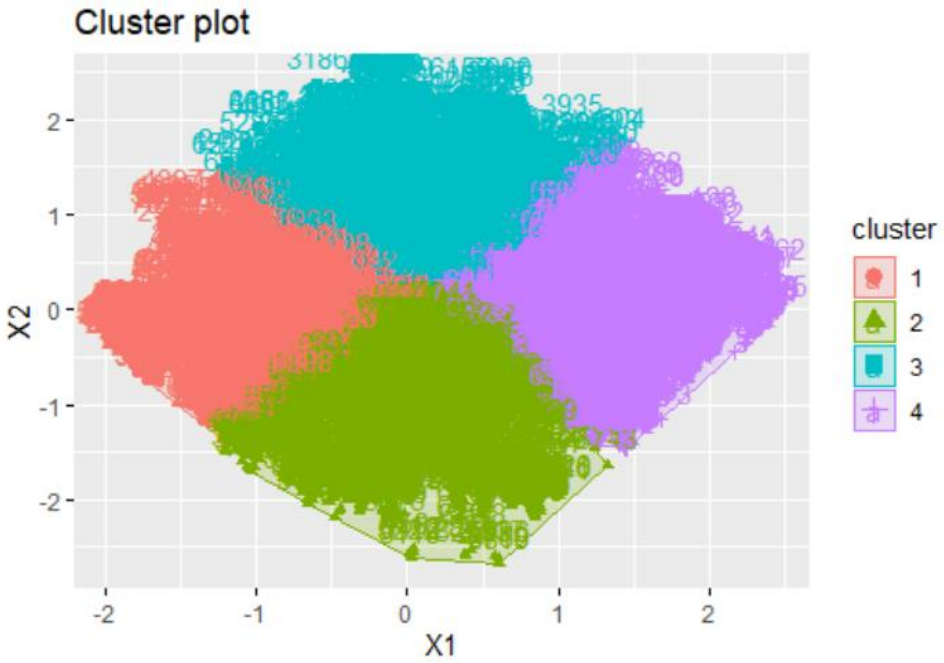
```
mds$points
                [,1]          [,2]
 [1,] -8.525588e-02  4.928924e-02
 [2,]  6.904241e-02  2.344584e-01
 [3,] -1.357045e-01  1.825215e-01
 [4,]  1.105383e-01  2.443426e-01
 [5,] -2.658369e-01  1.536546e-02
 [6,] -2.810697e-01 -1.947659e-02
 [7,] -1.010437e-01 -7.067405e-02
 [8,] -2.605881e-02  4.021969e-01
 [9,] -1.534968e-01 -1.744332e-01
[10,]  2.380770e-01  1.050623e-01
```

K-Means

首先以 Silhouette Coefficient 及組內變異決定群數 k，以圖可看出在 k=4 時 SC 為第二高(僅次 k=9)，組內變異在 4 之後就沒有顯著下降，因此選擇 k=4



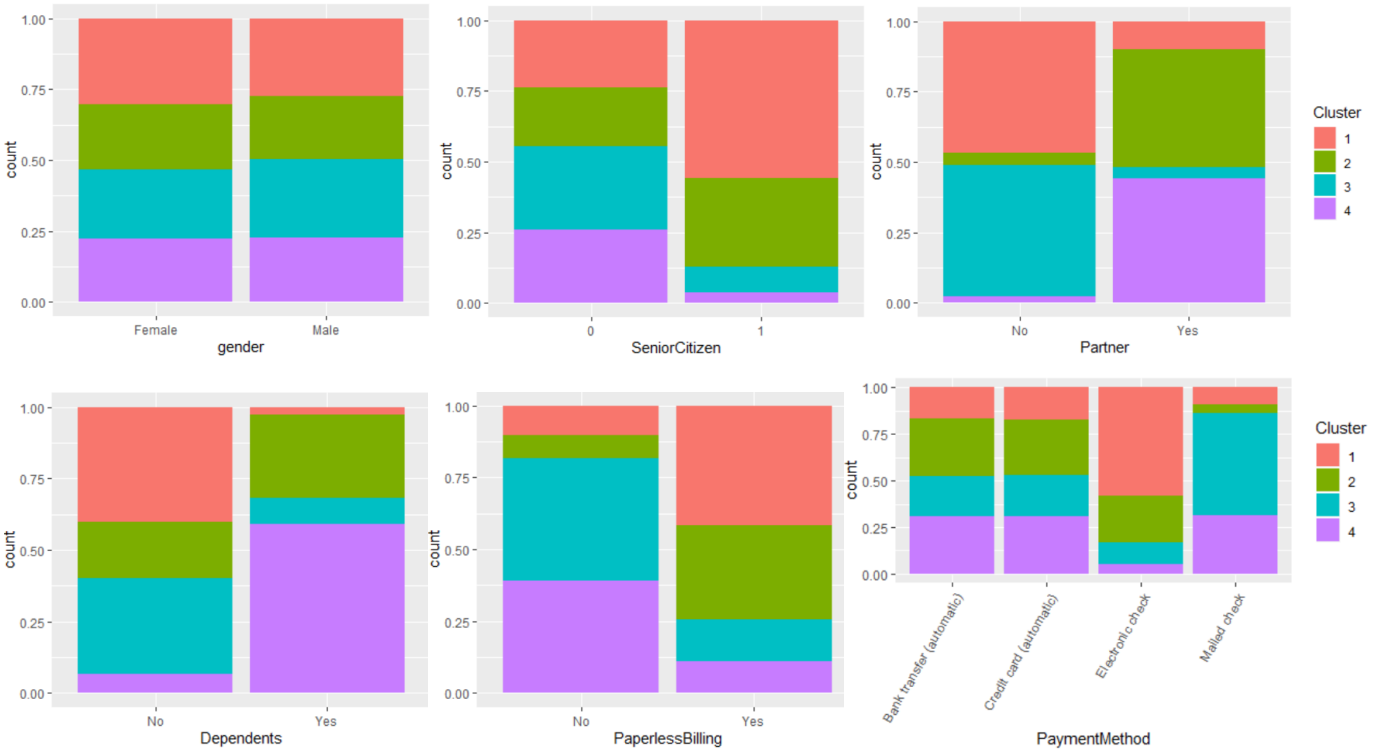
以下的圖為 K-Means 在 MDS 投影空間的分群結果，可看出 k-Means 對 MDS 投影後的資料切得不錯，且數量分佈均勻，但不保證這樣的分群法對原本的資料是好的，我們觀察此分群對原先定義的距離矩陣所算出的 Silhouette Coefficient = 0.1853294，並不能算是個很好的分群，但再試過其他方法 ：Spectral Clustering、Kernel K-means、Mini Batch K-means…，在任何群數的情況下 SC 值皆沒有比 0.18 高，因此最後以此分群作為分析。

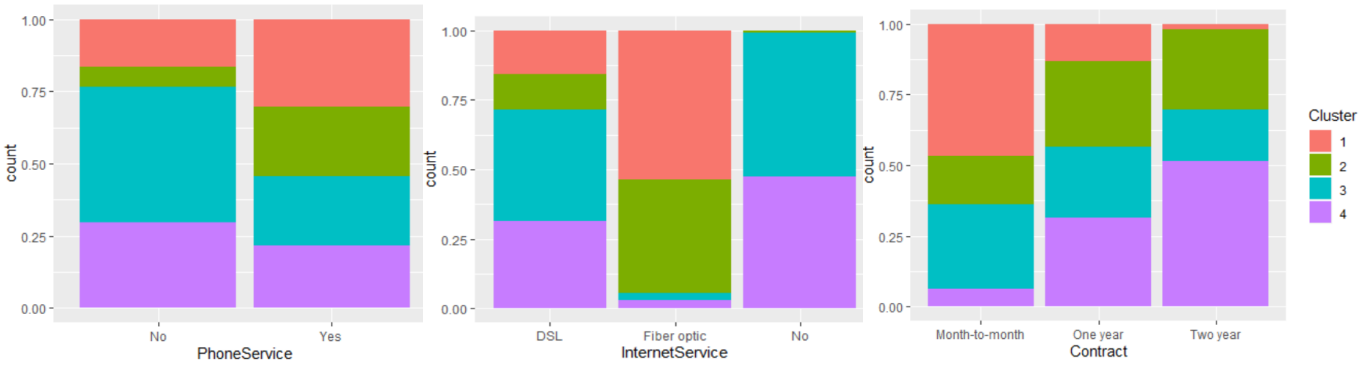


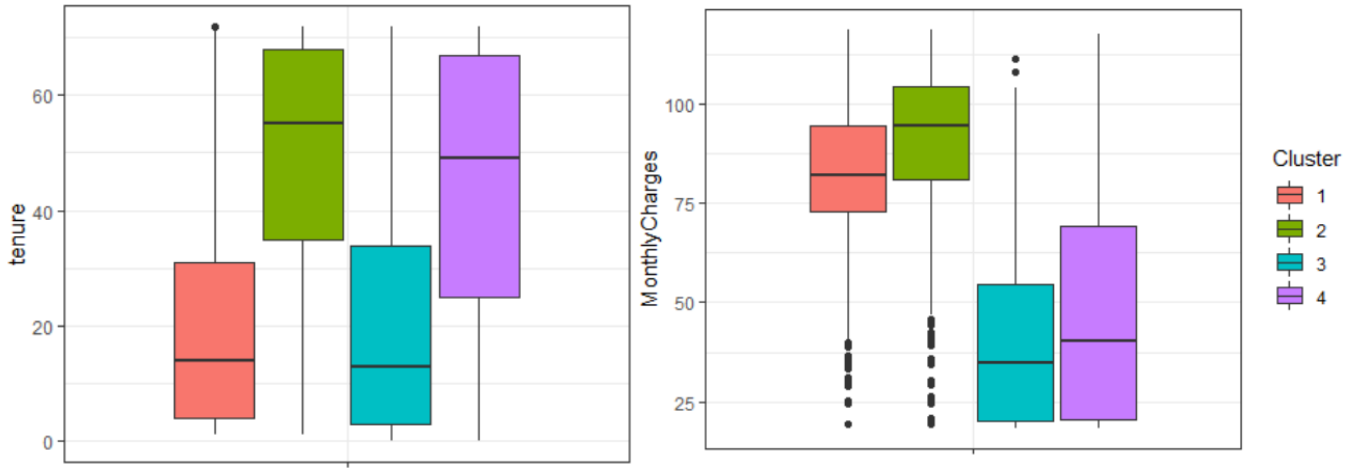

```
> km$size
[1] 2040 1586 1837 1580
```

分群後的分析

以下各圖是分群後各變數於各群中的分佈圖，類別型變數的長條圖是代表各群在各類別中的比重，而連續型變數的箱型圖則代表各群的四分位數。從下圖可看出第一群及第二群客戶月費較高，多數皆有申請電話及網路，並且較傾向於使用無紙化帳單，差別在於第一群客戶是單身為主，大多數沒有配偶及依附親屬，傾向於簽訂單月的合約，平均資歷(tenure)也較短。第三群及第四群客戶月費較低，多數沒有申請電話或網路服務，而兩群間的差異同樣是家庭組成，第三群客戶大多數沒有配偶及依附親屬，而第四群的客戶明顯地傾向簽訂更長的合約。







未放入分群模型變數的分析

在分群完成後分析未放入分群模型的變數，看它們在各群間的表現為何

從網路安全防護這個變數可看出來，擁有配偶及依附親屬卻較少申請網路服務的第四群客戶，反而很在乎網路安全。從是否解約這個變數來看，第四群客戶在解約的客戶中比例較少，幾乎不太會解約，到目前為主幾乎可判斷第四群客戶為對電信市場較消極的客戶，在電話及網路消費不高，但卻不容易解約投入其他電信公司的懷抱，平均合約的時限也較長。而月費高且多數單身的第一群客戶在解約的客戶中佔大多數，幾乎都是簽單月合約，估計屬於對電信市場較積極的客戶，願意消費且會關心市場是否有其他公司提出更適合自己的電信方案，應屬於電信公司需鎖定的客群。