

迴歸期末報告

電信客戶是否流失



107354025
軒

統碩一

林志

研究動機:

現在的電信業蓬勃發展，想要了解如何減少客戶的流失率，將其重要的變數也就是哪些是客戶需要的服務項目找出來，鞏固自己的市占率。而這筆資料提供了一些變數給我參考，藉此分

析一下那些變數是其客戶重視的。

資料來源:kaggle

<https://www.kaggle.com/blastchar/telco-customer-churn>

Product ID (Product)	Product Name (Product)	Manufacturer (Product)	Material (Product)	Dimensions (Product)	Weight (Product)	Capacity (Product)	Usage (Product)	Availability (Product)	Comments (Product)
PROD-001	Item A	Manufacturer X	Material Y	10 x 5 x 2	1.5 kg	50 L	Industrial	In Stock	Good quality
PROD-002	Item B	Manufacturer Y	Material Z	15 x 8 x 3	2.5 kg	75 L	Commercial	Low Stock	Check price
PROD-003	Item C	Manufacturer Z	Material A	20 x 10 x 4	3.5 kg	100 L	Residential	Out of Stock	Find alternative
PROD-004	Item D	Manufacturer A	Material B	25 x 12 x 5	4.5 kg	125 L	Industrial	Back Order	Lead time 4 weeks
PROD-005	Item E	Manufacturer B	Material C	30 x 15 x 6	5.5 kg	150 L	Commercial	Discontinued	Find replacement
PROD-006	Item F	Manufacturer C	Material D	35 x 18 x 7	6.5 kg	175 L	Residential	Low Stock	Check price
PROD-007	Item G	Manufacturer D	Material E	40 x 20 x 8	7.5 kg	200 L	Industrial	Back Order	Lead time 6 weeks
PROD-008	Item H	Manufacturer E	Material F	45 x 22 x 9	8.5 kg	225 L	Commercial	Discontinued	Find replacement
PROD-009	Item I	Manufacturer F	Material G	50 x 25 x 10	9.5 kg	250 L	Residential	Low Stock	Check price
PROD-010	Item J	Manufacturer G	Material H	55 x 28 x 11	10.5 kg	275 L	Industrial	Back Order	Lead time 8 weeks
PROD-011	Item K	Manufacturer H	Material I	60 x 30 x 12	11.5 kg	300 L	Commercial	Discontinued	Find replacement
PROD-012	Item L	Manufacturer I	Material J	65 x 32 x 13	12.5 kg	325 L	Residential	Low Stock	Check price
PROD-013	Item M	Manufacturer J	Material K	70 x 35 x 14	13.5 kg	350 L	Industrial	Back Order	Lead time 10 weeks
PROD-014	Item N	Manufacturer K	Material L	75 x 38 x 15	14.5 kg	375 L	Commercial	Discontinued	Find replacement
PROD-015	Item O	Manufacturer L	Material M	80 x 40 x 16	15.5 kg	400 L	Residential	Low Stock	Check price

資料筆數:7043

變數介紹:

customerID	客戶名	OnlineBackup	有無在線備份
gender	性別	DeviceProtection	有無服務保護
SeniorCitizen	是否為老年人	TechSupport	有無技術支援
Partner	有無朋友使用	StreamingTV	用手機看電視

Dependents	有無家人使用	StreamingMovies	用手機看電影
tenure	目前使用月數	Contract	合約(月、年)
PhoneService	有無電話服務	PaperlessBilling	有無紙本帳單
MultipleLines	有無多個電信	PaymentMethod	付款方式
InternetService	有無網路服務	MonthlyCharges	每月金額
OnlineSecurity	有無網路安全性	TotalCharges	總金額
Churn	是否流失		

tenure(目前使用月數): 0-72個月

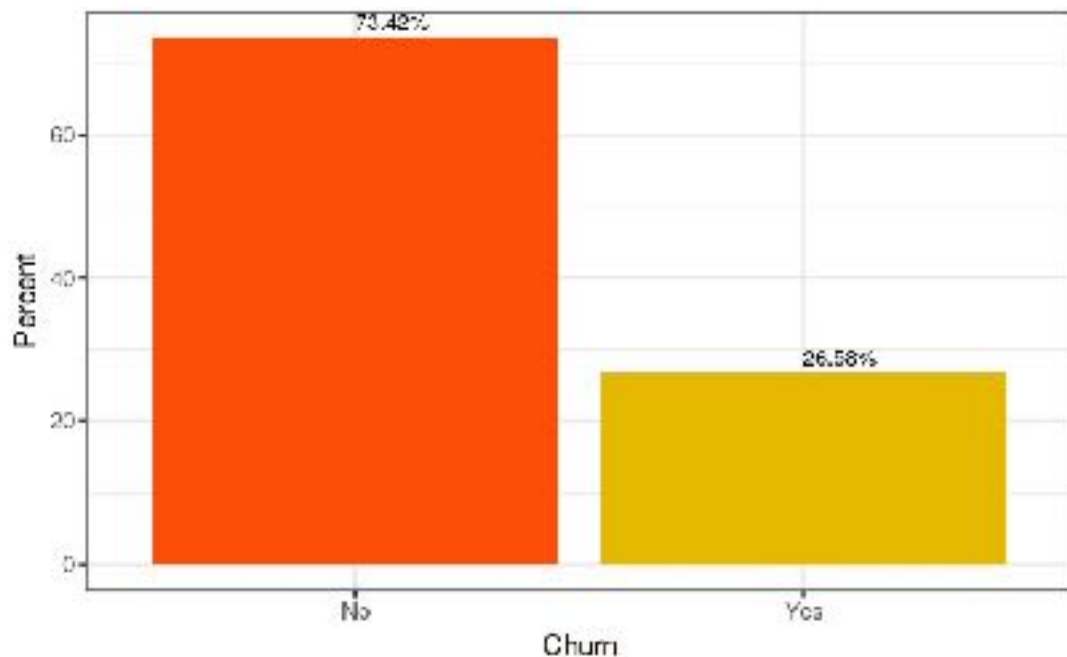
InternetService(網路服務):Fiber optic(光纖)、DSL(電話撥接)、NO

Contract(合約): month-to-month、1年、2年

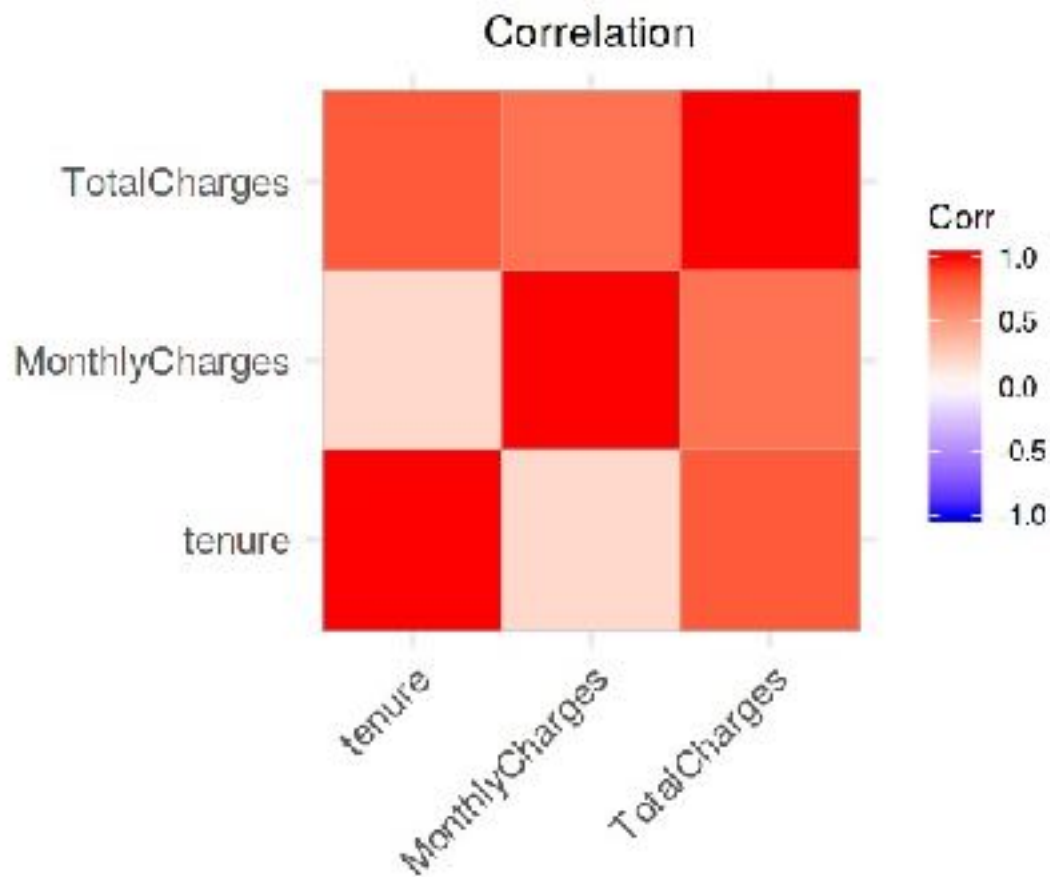
PaymentMethod:

bank transfer(銀行轉帳)、Credit card(信用卡)、
electric check(電子支票)、Mail check(郵寄支票)

EDA:



這個資料集的CHURN YES NO比例為26.6%,73.4%



迴歸模型:

利用全部變數去跑其所有變數加交互作用項後，發現其交互作用項不顯著，故不放交互作用項。

1.Lasso

(Intercept)	0.0518863545
regdata.MonthlyCharges	.
regdata.L tenure	-0.0344400498
regdata.TotalCharges	0.0002587737
regdata.genderMale	0.0174280982
regdata.SeniorCitizen	0.2121201651
regdata.PartnerYes	.
regdata.DependentsYes	-0.1517711072
regdata.PhoneServiceYes	-0.5885499039
regdata.MultipleLinesYes	0.2453878497
regdata.InternetServiceFiber, optic	0.7702835180
regdata.InternetServiceNo	0.7738803605
regdata.onlineSecurityYes	-0.3934608295
regdata.OnlineBackupYes	-0.1576524826
regdata.DeviceProtectionYes	-0.0369590578
regdata.TechSupportYes	-0.3688330152
regdata.StreamingTVYes	0.1958273425
regdata.StreamingMoviesYes	0.2081470549
regdata.ContractOne, year	0.6596126231
regdata.ContractTwo, year	-1.3495615137
regdata.PaperlessBillingYes	0.3398524637
regdata.PaymentMethodCredit, card, automatic	-0.0723736433
regdata.PaymentMethodElectronic, check	0.3174396561
regdata.PaymentMethodMailed, check	-0.0346135001

將 MonthlyCharges及 Partner刪掉

$\text{Lambda} = 0.0004035769$

將資料按照 8 比 2 分為 train 及 test
並讓剩下的變數去跑full model

2.刪掉後的 model

```

Call:
glm(formula = churnyes ~ ., family = binomial(link = "logit"),
     data = Train)

Deviance Residuals:
    min       1q   median       3q      max
-1.9175  -0.8891  -0.2800   0.7296   3.4381

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.890e-01  1.906e-01   0.992  0.321435
genderFemale  2.721e-02  7.252e-02   0.374  0.708352
genderMale    NA             NA      NA      NA
SeniorCitizen 2.134e-01  9.434e-02   2.469  0.013558 *
dependentyes  -1.663e-01  9.097e-02  -1.828  0.067178 .
tenure        -6.437e-02  8.988e-03  -9.214  < 2e-16 ***
phonoserviceyes 7.356e-01  1.496e-01   4.916  8.81e-07 ***
MultipleLinesYes 3.032e-01  8.908e-02   3.404  0.000684 ***
"InternetServiceFiber optic" 7.274e-01  1.204e-01   6.087  4.48e-11 ***
"InternetServiceNo" -6.106e-01  1.519e-01  -4.025  4.21e-05 ***
OnlineSecurityYes 3.896e-01  9.608e-02   4.055  5.01e-05 ***
OnlineBackupyes -1.676e-01  8.688e-02  -1.922  0.054611 .
DeviceProtectionyes 3.356e-02  8.915e-02   0.357  0.511292
TechSupportyes -3.619e-01  9.624e-02  -3.761  0.000170 ***
StreamingTVyes 2.040e-01  9.142e-02   2.212  0.025609 *
StreamingMoviesyes 1.807e-01  9.135e-02   1.979  0.047872 .
"ContractOne_year" -7.869e-01  1.223e-01  -6.431  1.27e-10 ***
"ContractTwo_year" 1.341e+00  1.891e-01   7.088  1.36e-12 ***
PaperlessBillsyes 1.862e-01  8.348e-02   2.248  0.024096 ***
"PaymentMethodCredit card (automatic)" 3.453e-02  1.279e-01   0.426  0.669947
"PaymentMethodElectronic check" 1.041e-01  1.060e-01   0.871  0.004095 **
"PaymentMethodMailled check" 6.800e-02  1.289e-01   0.528  0.597812
installcharges 3.697e-01  7.897e-02   4.681  7.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6574.4 on 5634 degrees of freedom
Residual deviance: 4562.5 on 5613 degrees of freedom
AIC: 4706.5

Number of Fisher Scoring iterations: 6

```

可得不顯著的變數有：

gender、

Dependent、

deviceprotection、

Onlinebackup

並且AIC = 4706.5

3.Stepwise後的最終模型


```

Call:
glm(formula = ChurnYes ~ SeniorCitizen + DependentsYes + Tenure +
    PhoneServiceYes + MultipleLinesYes + 'InternetServiceFiber optic' +
    InternetServiceNo + onlineSecurityYes + onlineBackupYes +
    TechSupportYes + StreamingTVYes + StreamingMoviesYes + 'ContractOne year' +
    'ContractTwo year' + PaperlessBillingYes + 'PaymentMethodElectronic check' +
    'PaymentMethodMailed check' + 'PaymentMethodCredit card (automatic)' +
    TotalCharges, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    min       1q   median       3q      Max
 1.9465  0.6830  0.2814  0.7779  3.4547

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.873e-01  1.850e-01   1.012  0.311361
SeniorCitizen  -2.880e-01  9.453e-02  -3.047  0.002331 *
dependentsyes  -1.661e-01  9.095e-02  -1.826  0.067818 .
Tenure         6.411e-02  6.969e-03   9.200  < 2e-16 ***
phoneserviceyes  -7.317e-01  1.493e-01  -4.902  9.19e-07 ***
MultipleLinesYes  3.037e-01  8.907e-02  3.408  0.000651 ***
'InternetServiceFiber optic'  7.304e-01  1.103e-01  6.621  4.58e-11 ***
InternetServiceNo -6.189e-01  1.531e-01  -4.041  5.27e-05 ***
onlineSecurityYes  3.865e-01  9.597e-02  4.027  5.64e-05 ***
onlineBackupYes  -1.653e-01  8.682e-02  -1.904  0.056926 .
TechSupportYes    3.655e-01  9.612e-02  3.802  0.000144 ***
StreamingTVYes    1.987e-01  9.099e-02  2.183  0.029007 *
StreamingMoviesYes  1.772e-01  9.113e-02  1.944  0.051843 .
'ContractOne year' -7.925e-01  1.278e-01  -6.197  7.96e-11 ***
'ContractTwo year' -1.348e+00  1.887e-01  -7.141  9.28e-13 ***
PaperlessBillingYes  3.075e-01  8.345e-02  3.684  1.93e-06 ***
'PaymentMethodElectronic check'  3.050e-01  1.060e-01  2.878  0.004001 **
'PaymentMethodMailed check'    6.847e-02  1.288e-01  0.531  0.595091
'PaymentMethodCredit card (automatic)' -5.559e-02  1.279e-01  -0.434  0.663950
TotalCharges     3.630e-04  7.827e-05  4.637  3.53e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6574.1  on 5631  degrees of freedom
Residual deviance: 4663.1  on 5615  degrees of freedom
AIC: 4703.1

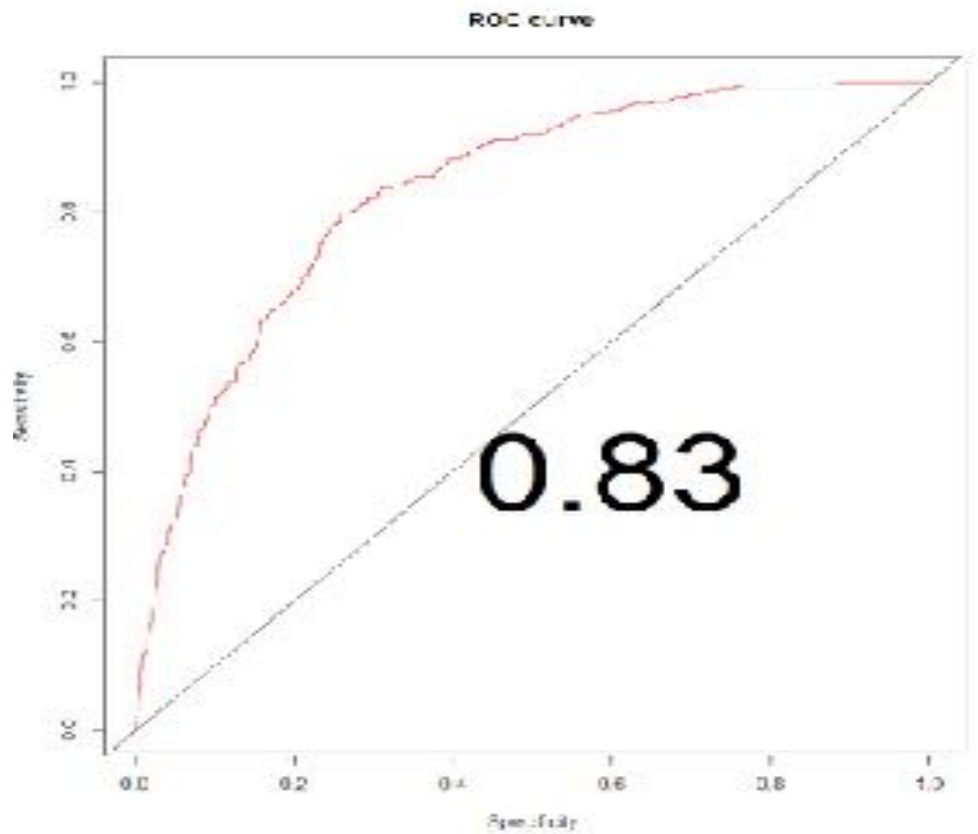
Number of Fisher Scoring iterations: 6

```

AIC:4703.1

AIC相較於full model小且變數幾乎都顯著
將此模型當作最終模型

ROC曲線:



Confusion matrix:

	流失	沒流失
流失	942	164
沒流失	119	183
Accuracy	0.799	
Sensitivity	0.89	
Specificity	0.53	

AUC為0.83且正確率達到0.799表示為一個不差的模型。

結論：

經由lasso及逐步迴歸的篩選後，最終的模型為

$$\text{Log}\left(\frac{p}{1-p}\right) = 0.19 + 0.233\text{SeniorCitizen} -$$

0.166*DependentsYes

- 0.064 * tenure - 0.73*PhoneServiceYes

+ 0.3 * MultipleLinesYes + 0.73 * `InternetServiceFiber
optic` - 0.62 * InternetServiceNo - 0.39

*OnlineSecurityYes

-0.17 * OnlineBackupYes - 0.37 * TechSupportYes

+ 0.2 * StreamingTVYes + 0.18 * StreamingMoviesYes

- 0.79 * `ContractOne year` -1.35 * `ContractTwo

year`

+ 0.4 * PaperlessBillingYes

+ 0.31 * `PaymentMethodElectronic check`

-0.07 * `PaymentMethodMailed check`

- 0.06 * `PaymentMethodCredit card (automatic)`

+ 0.0004 * TotalCharges

舉例來說：

固定其他變數下，每增加一單位的tenure，則*odds*
會降低 $e^{-0.064}$ 倍(老客戶)

固定其他變數下，有MultipleLines會比無
MultipleLines 的*odds*會提高 $e^{0.3}$ 倍