迴歸期末報告

電信客戶是否流失

107354025　　統碩一　　林志軒

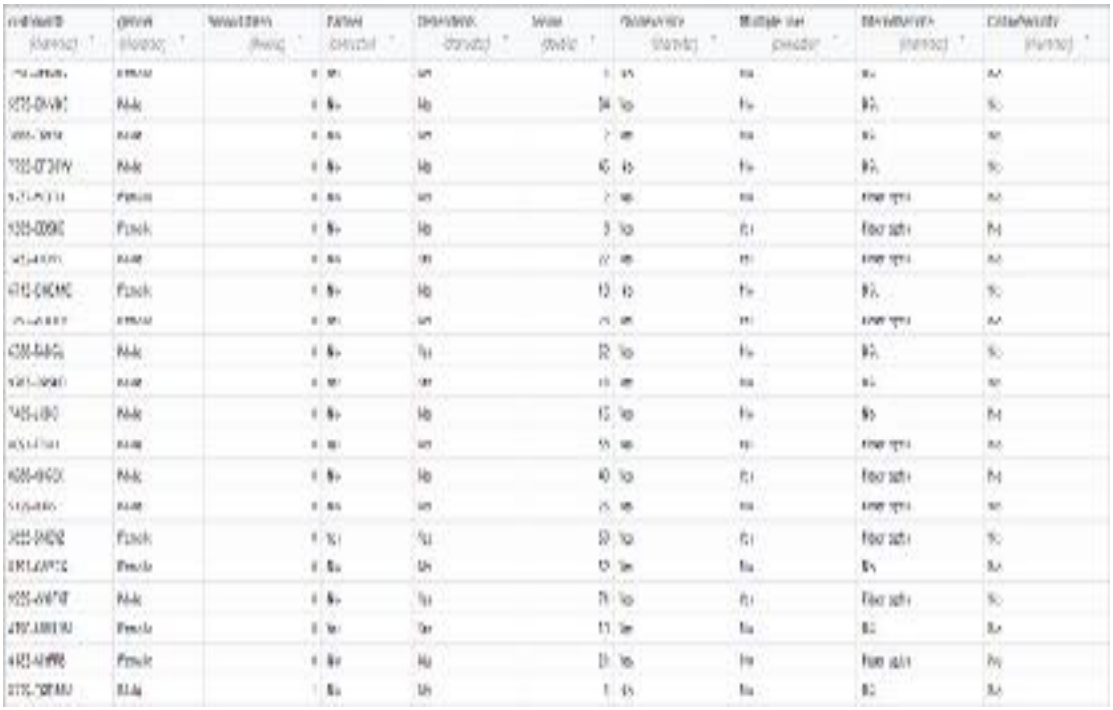
研究動機:

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



資料筆數:7043

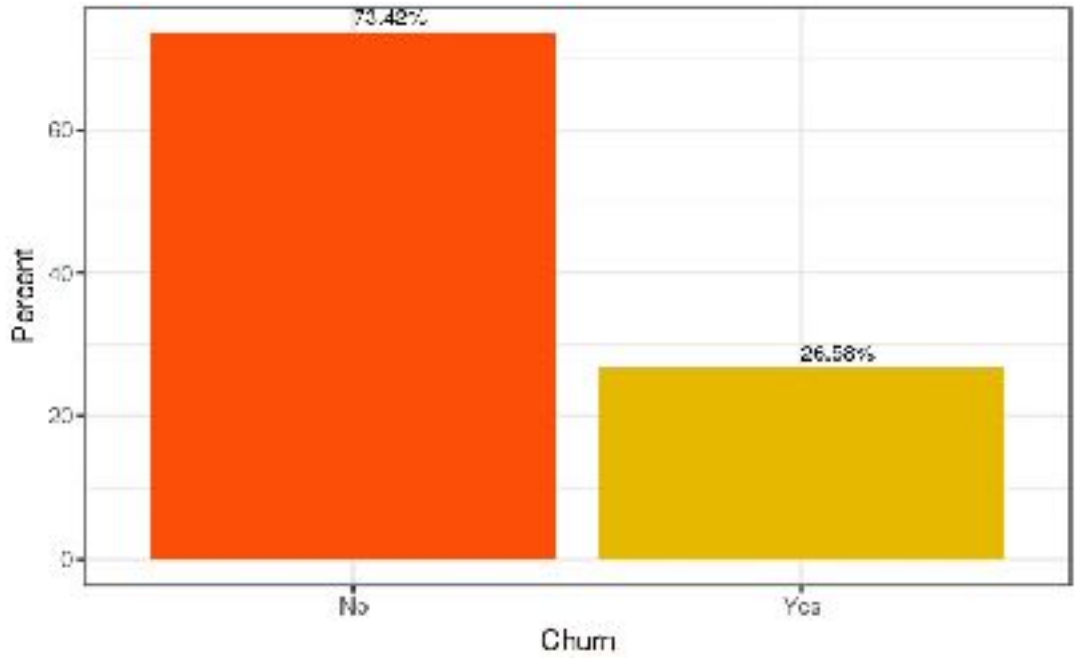
變數介紹:

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

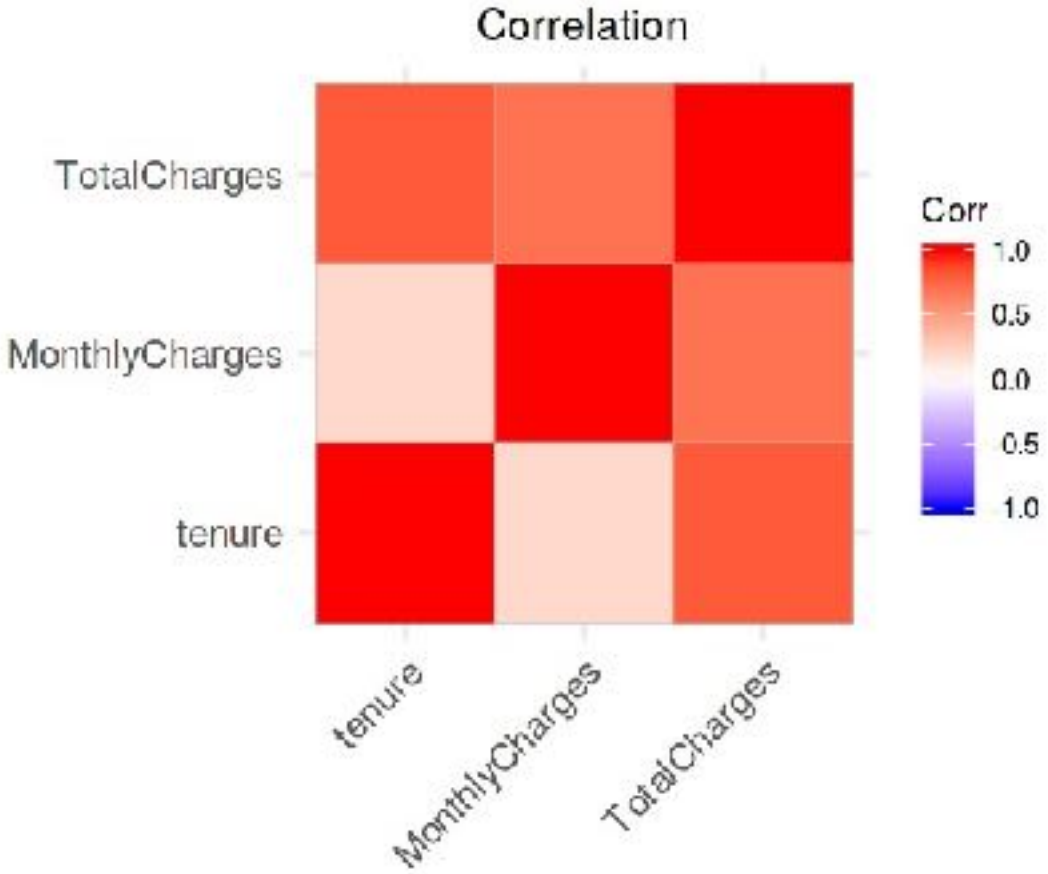
Dependents	有無家人使用	StreamingMovies	用手機看電影
tenure	目前使用月數	Contract	合約(月、年)
PhoneService	有無電話服務	PaperlessBilling	有無紙本帳單
MultipleLines	有無多個電信	PaymentMethod	付款方式
InternetService	有無網路服務	MonthlyCharges	每月金額
OnlineSecurity	有無網路安全性	TotalCharges	總金額
Churn	是否流失		

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:



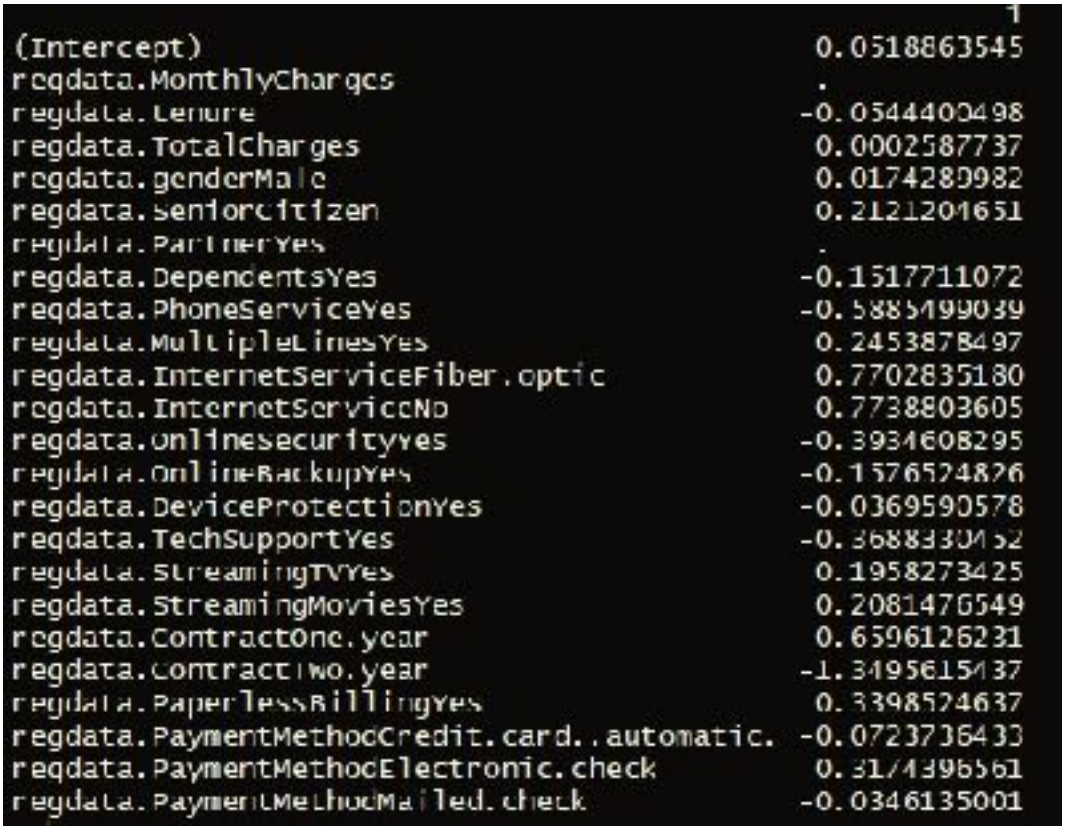
這個資料集的CHURN YES NO比例為26.6%,73.4%



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



	1
(Intercept)	0.0518863545
regdata.MonthlyCharges	.
regdata.Tenure	-0.0544400498
regdata.TotalCharges	0.0002587737
regdata.genderMale	0.0174289982
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	-0.3688330452
regdata.StreamingTVYes	0.1958273425
regdata.StreamingMoviesYes	0.2081476549
regdata.ContractOne.year	0.6596126231
regdata.ContractTwo.year	-1.3495615437
regdata.PaperlessBillingYes	0.3398524637
regdata.PaymentMethodCredit.card..automatic.	-0.0723736433
regdata.PaymentMethodElectronic.check	0.3174396561
regdata.PaymentMethodMailed.check	-0.0346135001

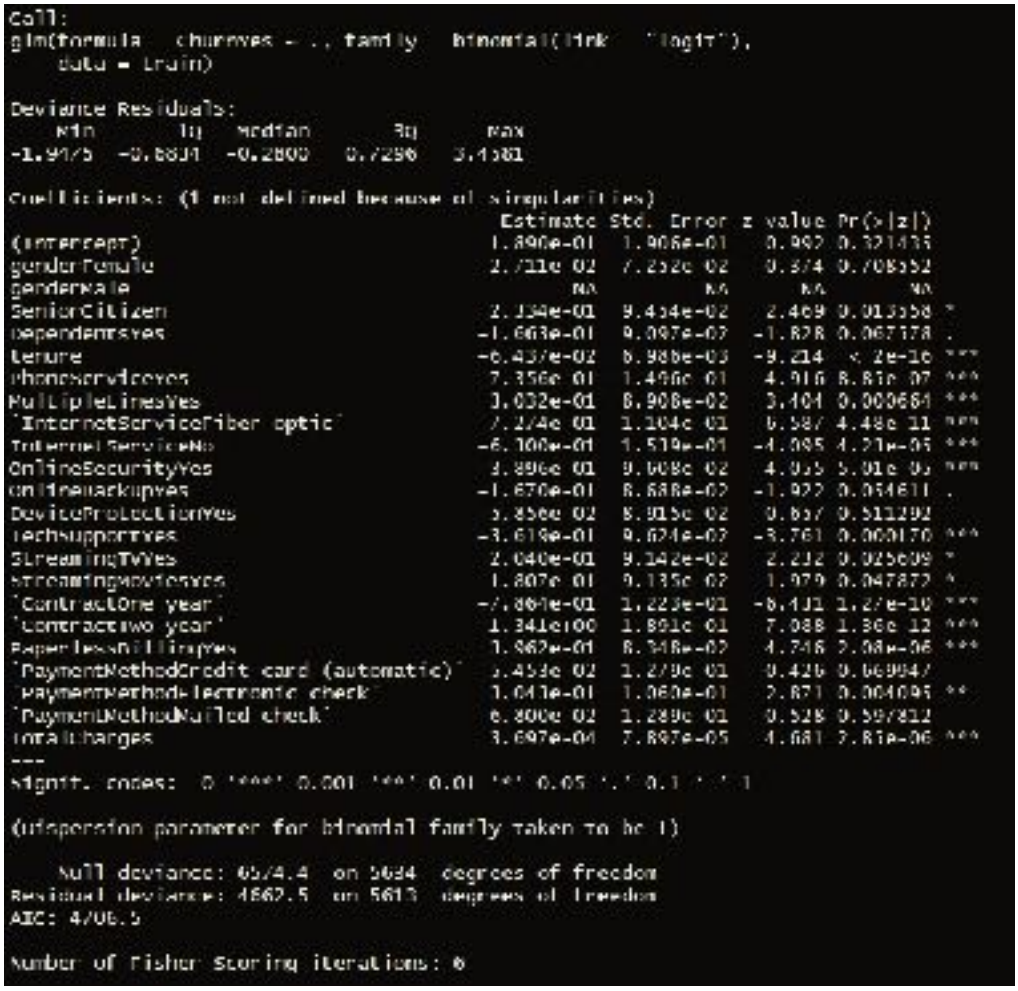
將 MonthlyCharges及 Partner刪掉

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

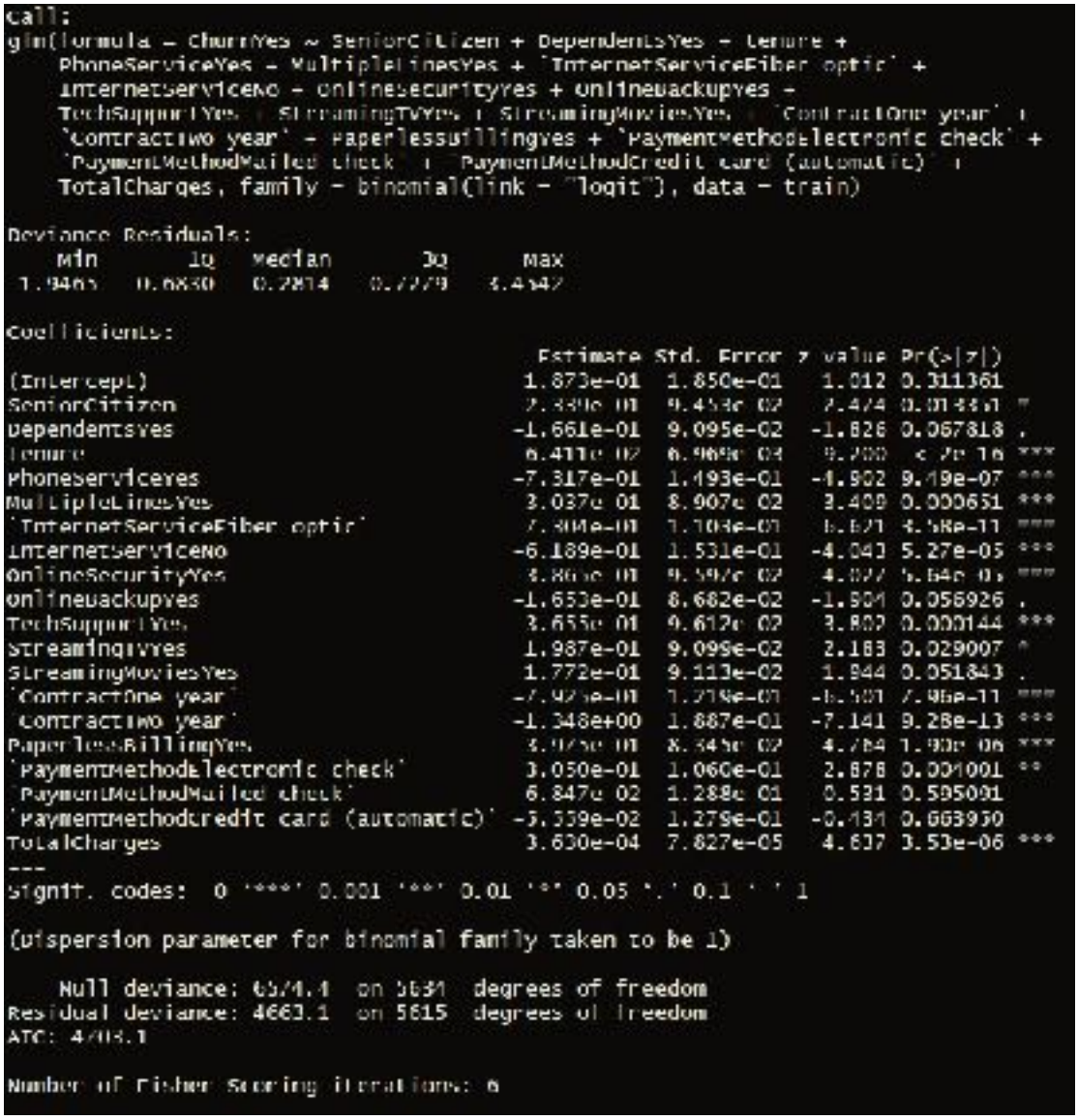
2.刪掉後的 model



```
Call:
glm(formula = churnYes ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9175  -0.8831  -0.2800   0.7290   3.1681

Coefficients: (1 not defined because of singularities)
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.890e-01  1.906e-01   0.992 0.321135
genderFemale                           2.711e-02  7.252e-02   0.374 0.708352
genderMale                                    NA         NA      NA       NA
SeniorCitizen                          2.134e-01  9.454e-02   2.469 0.013558 *
DependentsYes                         -1.663e-01  9.097e-02  -1.828 0.067578 .
tenure                                -6.437e-02  6.988e-03  -9.214  < 2e-16 ***
PhoneServiceYes                        7.356e-01  1.496e-01   4.916 8.81e-07 ***
MultipleLinesYes                       3.032e-01  8.908e-02   3.404 0.000664 ***
`InternetServiceFiber optic`           7.174e-01  1.104e-01   6.497 8.18e-11 ***
InternetServiceNo                     -6.100e-01  1.519e-01  -4.015 5.91e-05 ***
OnlineSecurityYes                      3.896e-01  9.608e-02   4.055 5.01e-05 ***
OnlineBackupYes                       -1.670e-01  8.688e-02  -1.922 0.054611 .
DeviceProtectionYes                    3.850e-02  8.915e-02   0.432 0.665792
TechSupportYes                        -3.619e-01  9.624e-02  -3.761 0.000170 ***
StreamingTVYes                         2.040e-01  9.142e-02   2.232 0.025609 *
StreamingMoviesYes                     1.807e-01  9.135e-02   1.979 0.047872 *
`Contract0ne year`                    -7.801e-01  1.223e-01  -6.381 1.76e-10 ***
`Contract2wo year`                    -1.341e+00  1.891e-01  -7.088 1.36e-12 ***
PaperlessBillingYes                    1.867e-01  8.348e-02   4.746 2.08e-06 ***
`PaymentMethodCredit card (automatic)` -3.459e-02  1.279e-01  -0.426 0.669847
`PaymentMethodElectronic check`        1.041e-01  1.060e-01   2.871 0.004095 **
`PaymentMethodMailed check`            6.800e-02  1.289e-01   0.528 0.597812
TotalCharges                           3.697e-04  7.897e-05   4.681 2.81e-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: 4662.5  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、
Onlinebakup
並且AIC = 4706.5

3.Stepwise後的最終模型



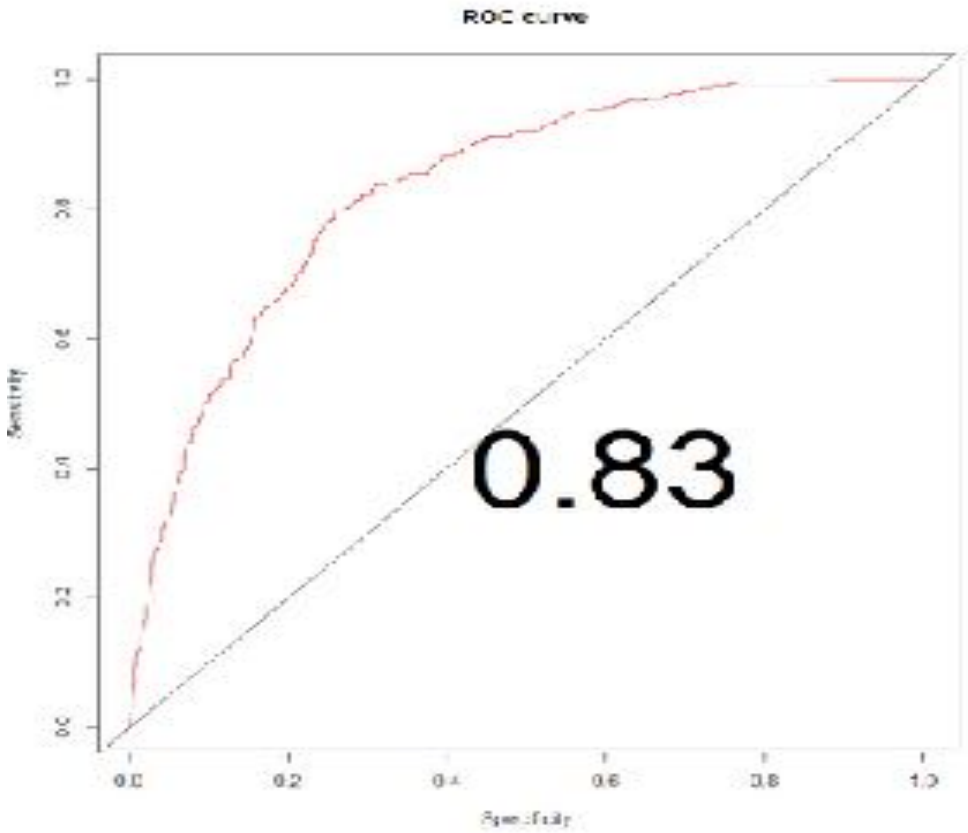
```
Call:
glm(formula = ChurnYes ~ SeniorCitizen + DependentsYes + tenure +
    PhoneServiceYes + MultipleLinesYes + `InternetServiceFiber optic` +
    InternetServiceNo + OnlineSecurityYes + OnlineBackupYes +
    TechSupportYes + StreamingTVYes + StreamingMoviesYes + `Contract0ne year` +
    `Contract2wo 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   0.7279   3.4542

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.873e-01  1.850e-01   1.012 0.311361
SeniorCitizen                          2.330e-01  9.458e-02   2.474 0.013361 *
DependentsYes                         -1.661e-01  9.095e-02  -1.826 0.067818 .
tenure                                -6.411e-02  6.906e-03  -9.283  < 2e-16 ***
PhoneServiceYes                       -7.317e-01  1.193e-01  -4.902 9.49e-07 ***
MultipleLinesYes                       3.037e-01  8.907e-02   3.409 0.000651 ***
`InternetServiceFiber optic`           7.804e-01  7.708e-02   10.171 < 2e-16 ***
InternetServiceNo                     -6.189e-01  1.531e-01  -4.043 5.27e-05 ***
OnlineSecurityYes                      4.866e-01  9.597e-02   5.072 3.94e-07 ***
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 .
`Contract0ne year`                    -7.923e-01  1.219e-01  -6.501 7.96e-11 ***
`Contract2wo year`                    -1.348e+00  1.887e-01  -7.141 9.28e-13 ***
PaperlessBillingYes                    4.070e-01  8.645e-02   4.708 2.50e-06 ***
`PaymentMethodElectronic check`        3.050e-01  1.060e-01   2.878 0.004001 **
`PaymentMethodMailed check`            6.847e-02  1.288e-01   0.531 0.595001
`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: 6514.4  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

$$\text{Log}(\frac{p}{1-p}) = 0.19 + 0.233\text{SeniorCitizen} - 0.166\text{*DependentsYes} - 0.064 * \text{tenure} - 0.73\text{*PhoneServiceYes} + 0.3 * \text{MultipleLinesYes} + 0.73 * \text{`InternetServiceFiber optic`} - 0.62 * \text{InternetServiceNo} - 0.39\text{*OnlineSecurityYes} - 0.17 * \text{OnlineBackupYes} - 0.37 * \text{TechSupportYes} + 0.2 * \text{StreamingTVYes} + 0.18 * \text{StreamingMoviesYes} - 0.79 * \text{`ContractOne year`} - 1.35 * \text{`ContractTwo year`} + 0.4 * \text{PaperlessBillingYes} + 0.31 * \text{`PaymentMethodElectronic check`} - 0.07 * \text{`PaymentMethodMailed check`} - 0.06 * \text{`PaymentMethodCredit card (automatic)`} + 0.0004 * \text{TotalCharges}$$

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