電信用戶流失

流失風險預測與因素分析

資料集來源:Telco Customer Churn

Linkedin

104人力銀行

陳勁瑋　製作

AGENDA

1. 專案目標
2. 數據洞察
3. 數據處理
4. 建造模型
5. 總結
6. 附錄

AGENDA

1. 專案目標
2. 數據洞察
3. 策略建議
4. 建造模型
5. 總結
6. 附錄

假設我任職於一 電信公司，今日主管針對 顧客流失率上升 的問題，期望我提出改善方案

情境	任職於一電信公司 內，近期 顧客流失率 顯著上升
	主管好奇顧客流失背後的 原因，並期望我提出 解方
	使用顧客數據進行分析，識別流失 /未流失顧客的差異
專案脈絡	先理解顧客流失背後可能的原因，並建立 假說
	針對假 說進行驗證，並依據不同客群制定 專屬 的解方

建立一個 預測模型，量化各因素對於客 戶流失的影響程度，並預測 客戶的流失率

現有痛點	電信客 戶逐漸流失，傳統分析方法無法揭示數據背後的深層意義
改善方式	使用 機器學習 分析客 戶行為數據，識別高風險流失族群
	針對不同流失族群採取專屬的商業策略
專案目標	建立有效的預測模型，分析新顧客的 未來流失風險
	找出顯著影響顧客是否 流失的關鍵特徵

AGENDA

1. 專案目標
2. **數據洞察**
3. 數據處理
4. 建造模型
5. 總結
6. 附錄

資料集介紹

類別變數

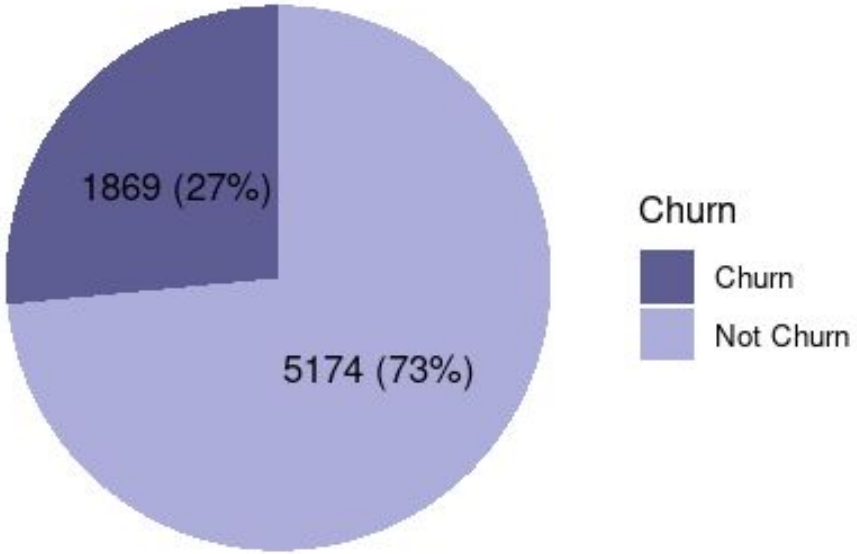
- Gender
- SeniorCitizen
- Partner
- Dependents
- Contract
- PhoneService
- StreamingTV
- MutipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- TechSupport
- PaymentMethod
- PaperlessBilling
- DeviceProtection
- StreamingMovies

連續變數

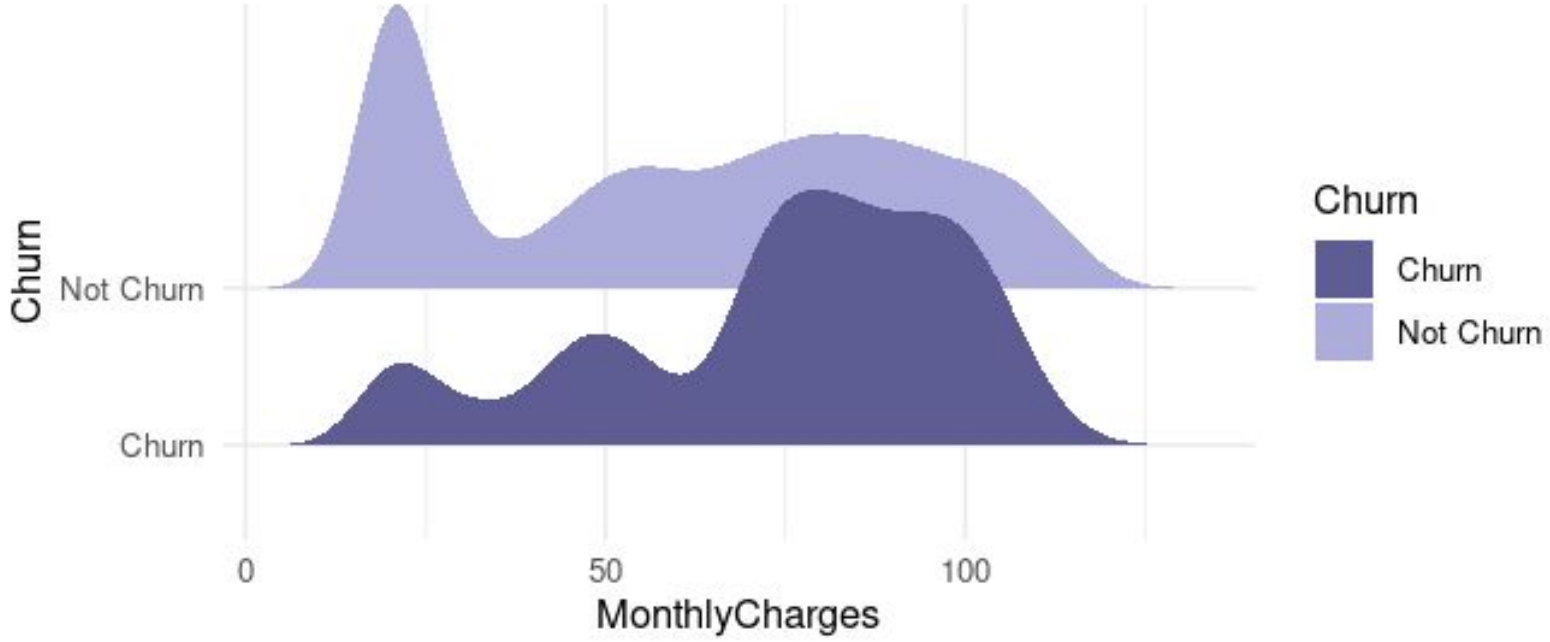
- Tenure
- MonthlyCharges
- TotalCharges

→ **Churn**

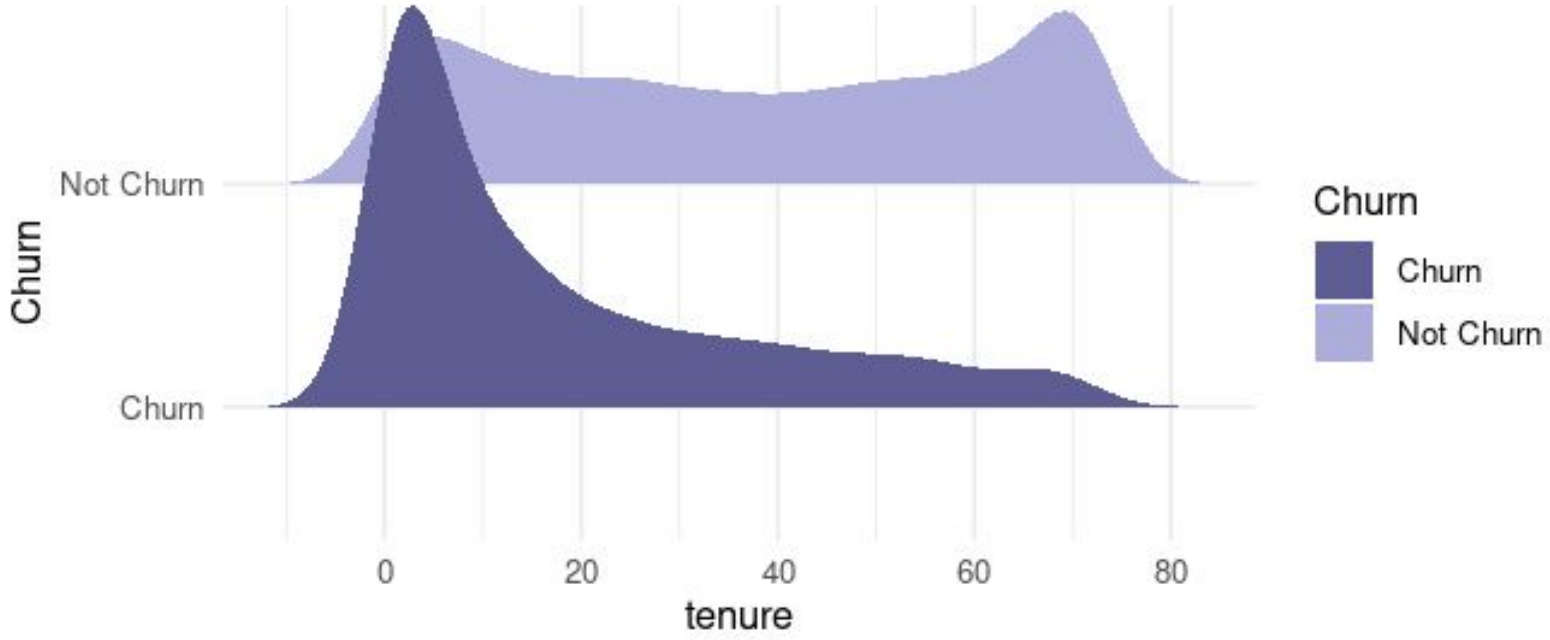
由下圖發現流失:未流失顧客約為 1:3



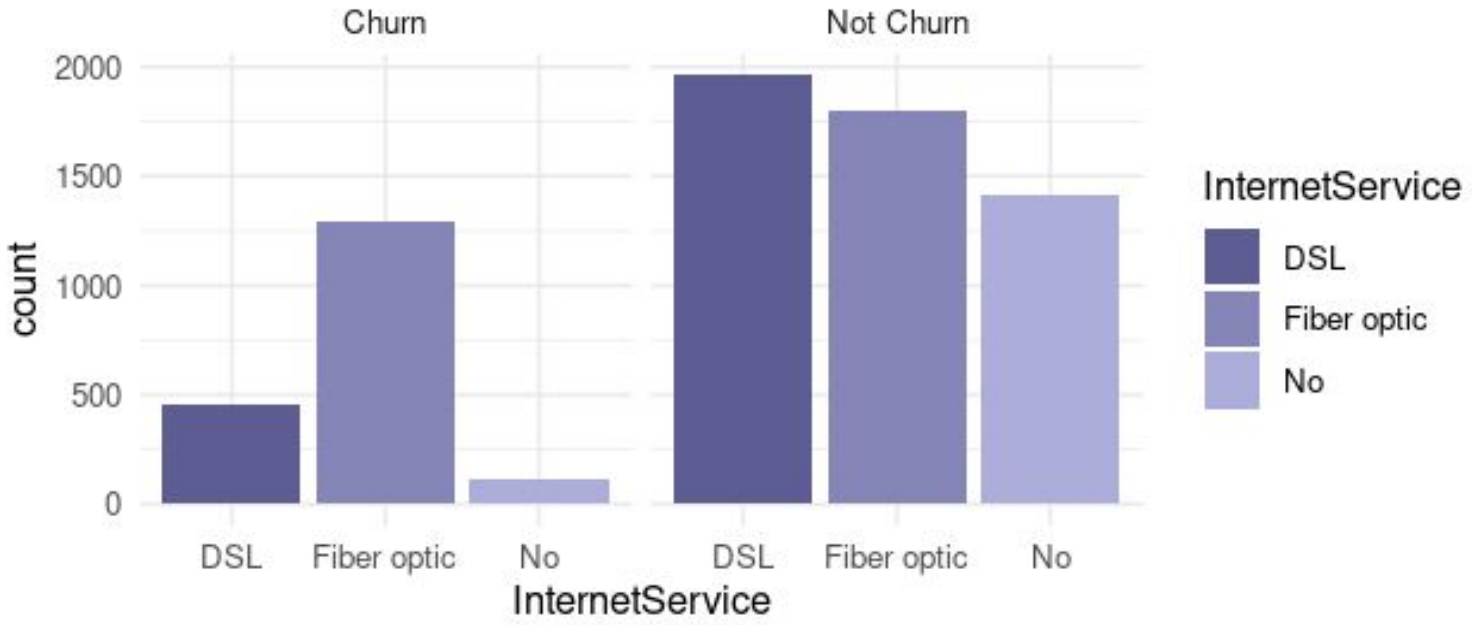
流失、未流失顧客皆呈現 左偏，但未流失顧客在較低月支付有 較高的峰值



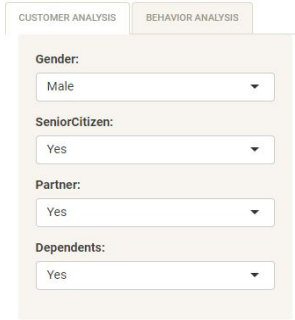
Tenure中的未流失顧客為 雙峰分布，流失顧客則呈現 右偏

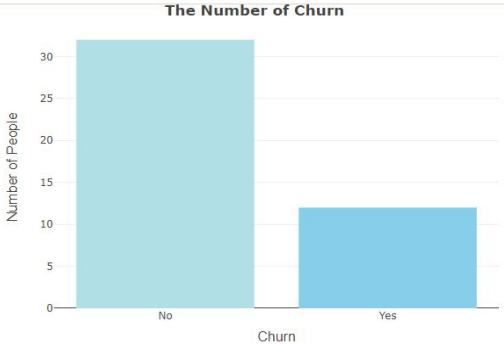


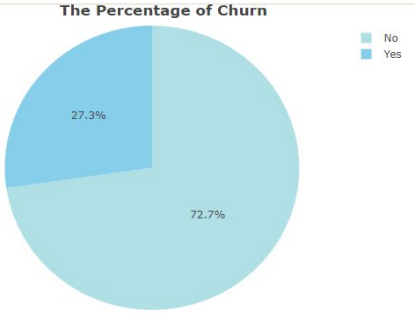
在InternetService的流失顧客中，以 Fiber optic最顯著

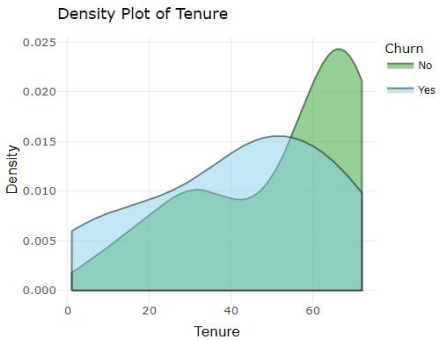


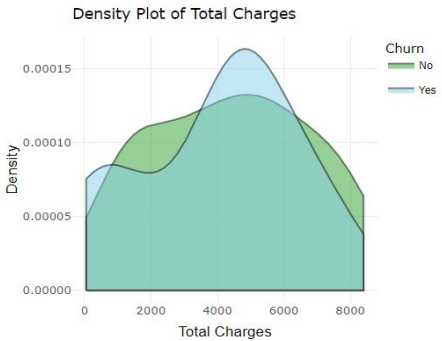
更多交互式 EDA，請點擊我或掃下方 QR code

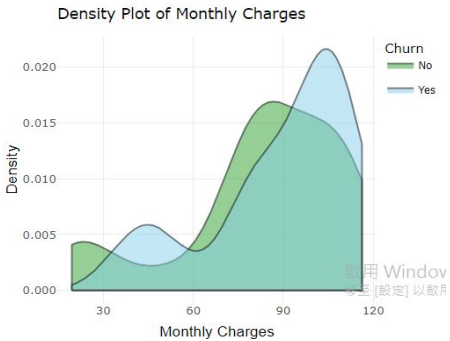










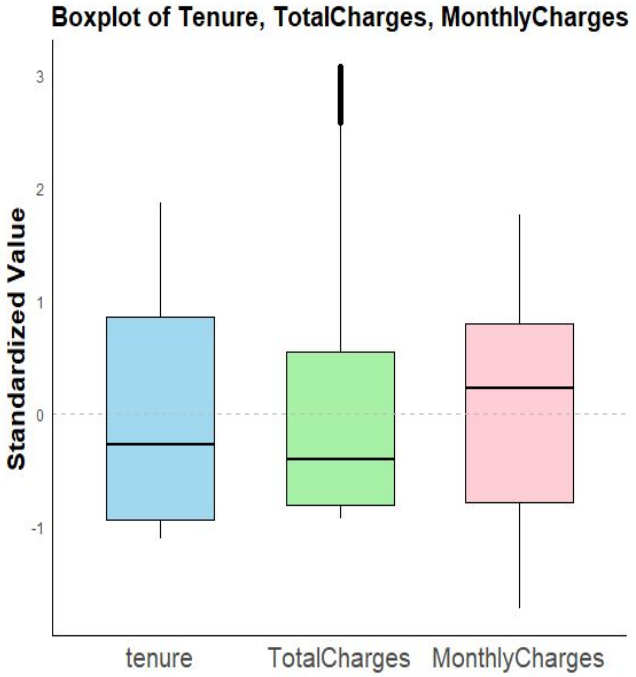


AGENDA

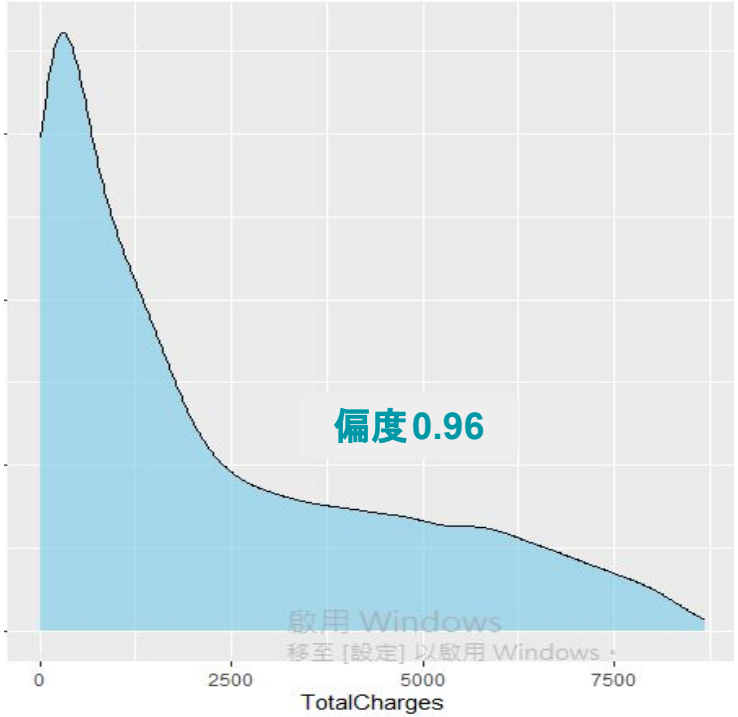
1. 專案目標
2. 數據洞察
3. **數據處理**
4. 建造模型
5. 總結
6. 附錄

步驟一:去除極端 值

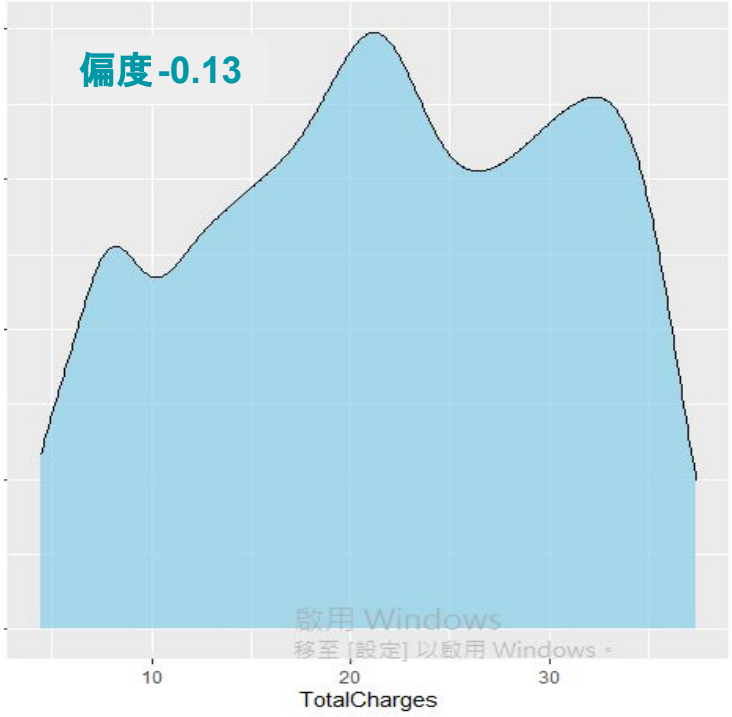
Numeric	針對連續變數的分析，需考慮是否為極端 值
去除方法	使用IQR法清理資料。當有數 值在第1、3四分位距外的1.5 X IQR距離時，視為極端 值
結果	透過IQR法檢驗後，並未發現任何極端 值



步驟二:改善 TotalCharges偏態問題



Boxcox轉換



步驟三:類別變數處理

現有問題	有部分變數會顯示顧客是否有使用該公司的某項附加服務
	例如在MultipleLines中，共有Yes、No、No PhoneService三類
隱患	某些變數可由另一個變數完全表示，引發 共線性、線性相依問題
改善方式	將上述MultipleLines中的No及No PhoneService均視為同一類
預期結果	降低共線性問題，使各個變數間彼此獨立，提升預測準確率

- MultipleLines
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV/Movies

步驟四:解決目標變數樣本不平衡

現有問題

客戶流失比(Churn)為1:3，流失的客戶佔比較少

可能使模型**隨機猜測**仍保有高準確率， 但Recall值低

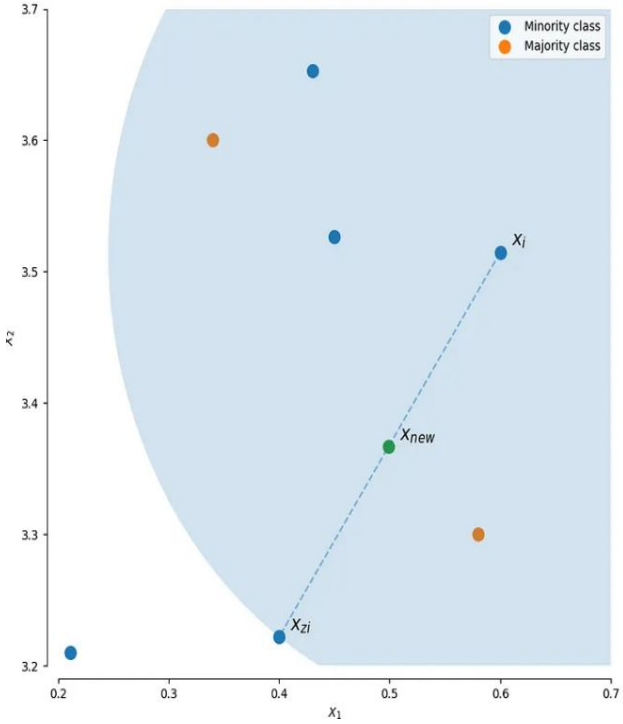
改善方式

因數據量有限，採取 OVERSAMPLE方式，降低模型 Bias

使用**SMOTE-NC**處理包含連續、類別變數的資料集

預期結果

透過OVERSAMPLE解決樣本不平衡可能造成的隱患，
使模型的預測效率以及 Recall值提高



AGENDA

1. 專案目標
2. 數據洞察
3. 數據處理
4. **建造模型**
5. 總結
6. 附錄

使用三種特徵工程方法，對比結果後進行特徵選擇

法一：逐步迴歸

- 依據模型 AIC值的變化，逐步添加或者移除特徵
- 對於**線性**的模型有良好的解釋力

法二：RFE

- 依據模型特徵重要性的評估，遞迴地增加或移除變數
- 廣泛應用於任何 **基於樹** 的模型，但相當耗時

法三：Boruta

- 創立影子特徵，並與原始特徵相比，篩選出對模型有幫助的特徵
- 專為**隨機森林** 所設立的特徵篩選方式

特徵選擇

- 比較三種特徵工程結果的 **異同**，再進行最終特徵選擇

根據羅吉斯迴歸的 VIF值，移除產生共線性的特徵

自變數	GVIF
MonthlyCharges	85.9
InternetServices	50.0
TotalCharges	20.7
Tenure	15.5
Others	<10

Monthly Charges

自變數	GVIF
TotalCharges	20.2
Tenure	15.2
Others	<10

Total Charges

模型不存在共線性

根據前一步驟遺留下來的變數，進行羅吉斯逐步迴歸

全模型開始

將前一步驟遺留下來的 **全部變數** 放進模型中

後項消除

若 **移除** 某個變數可以 **降低** 模型的AIC值，則將其移除

前項選擇

若 **增加** 某個變數可以 **降低** 模型的AIC值，則將其加入

OnlineBackup | MutipleLines | Dependents

Tenure | TechSupport | Contract

PhoneService | StreamingTV | InternetService

StreamingMovies | PaymentMethod

OnlineSecurity | PaperlessBilling

特徵工程方法二 (Recursive Feature Elimination)

依據隨機森林模型特徵的 重要性，遞迴的移除不重要的變數

特徵重要性	使用所有特徵訓練模型，依據特徵重要性進行排序
遞迴移除	刪除不重要的一個或多個特徵，並重新建立模型，遞迴重複此步驟直到達到預設的特徵數量
模型評估	通過交叉驗證來評估不同特徵子集下模型的預測效果，並尋找出最佳特徵數量及特徵子集
最終結果	RFE方法傾向於保留 全數特徵

特徵工程方法三 (Boruta)

在隨機森林模型中以創造 影子特徵 的方式，與原始特徵進行比較並保留重要的特徵

影子特徵	將每一個特徵的數據順序隨機打亂，創建影子特徵作為新變量
特徵比較	比較每個原始特徵與其對應的影子特徵的重要性，選取影子特徵中最高的重要性值作為閾值。若原始特徵的重要性小於該閾值，則移除該特徵
重複進行	重複進行上述步驟多次，以增加特徵選擇的穩定性，防止數據隨機打亂順序對結果的影響
最終結果	Boruta方法傾向於保留 全數特徵

使用逐步迴歸的結果選擇特徵，將特徵數量由 19降至13

方法異同	逐步迴歸：在移除不顯著的變數後，模型保留了 13個特徵
	RFE及Boruta：均傾向於保留全部的特徵
可能隱患	複雜的模型可能導致 過度擬合 問題，這使得模型在面對新數據時預測能力可能變差
	保留所有變數會增加模型計算的時間成本，尤其對大數據集來 說極其耗時
最終抉擇	選擇逐步迴歸的特徵選擇方法，因為它在統計上保持 顯著性，同時計算時間較為節省
	儘管RFE降至13個變數僅損失約 1%的準確度，考慮到以上問題，因此決定以此作為平衡的選擇

建造隨機森林模型，並結合交叉驗證尋找最佳超參數

訓練模型	結果
ntree	500
mtry	6
splitrule	gini
min.node.size	1
Precision	89.5%
Recall	97.3%

	Yes	No
Yes	4020	111
No	474	3657

測試模型	結果
Accuracy	82.5%
Precision	80.0%
Recall	86.6%
Sensitivity	86.6%
Specificity	78.3%
F1 Score	83.2%

	Yes	No
Yes	894	138
No	224	808

建造XGboost模型，並結合交叉驗證尋找最佳超參數

訓練模型	結果
nrounds	150
max_depth	6
eta	0.3
gamma	0.1
colsample bytree	0.8
Precision	91.2%
Recall	90.0%

	Yes	No
Yes	3700	431
No	297	3834

測試模型	結果
Accuracy	83.9%
Precision	84.2%
Recall	83.3%
Sensitivity	83.6%
Specificity	84.3%
F1 Score	83.9%

	Yes	No
Yes	863	169
No	162	870

AGENDA

1. 專案目標
2. 數據洞察
3. 數據處理
4. 建造模型
5. **總結**
6. 附錄

隨機森林與 XGboost皆是優良模型 – 提升原始模型的 精確率 及 召回率

Null Model

	Yes	No
Yes	216	157
No	111	921

Precision	67.1%
Recall	57.9%
F1 Score	62.2%
Accuracy	80.9%

RandomForest

	Yes	No
Yes	894	138
No	224	808

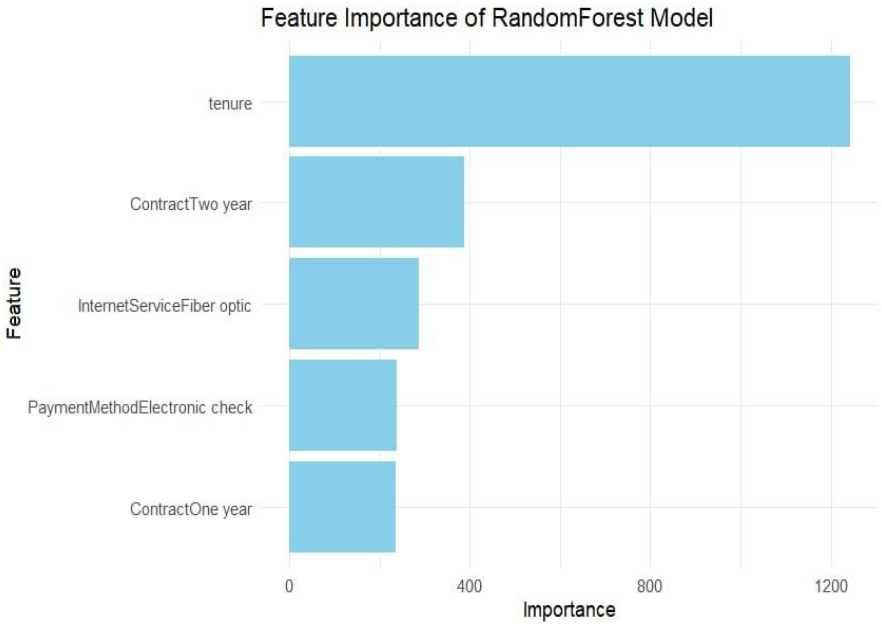
Precision	80.0%
Recall	86.6%
F1 Score	83.2%
Accuracy	82.5%

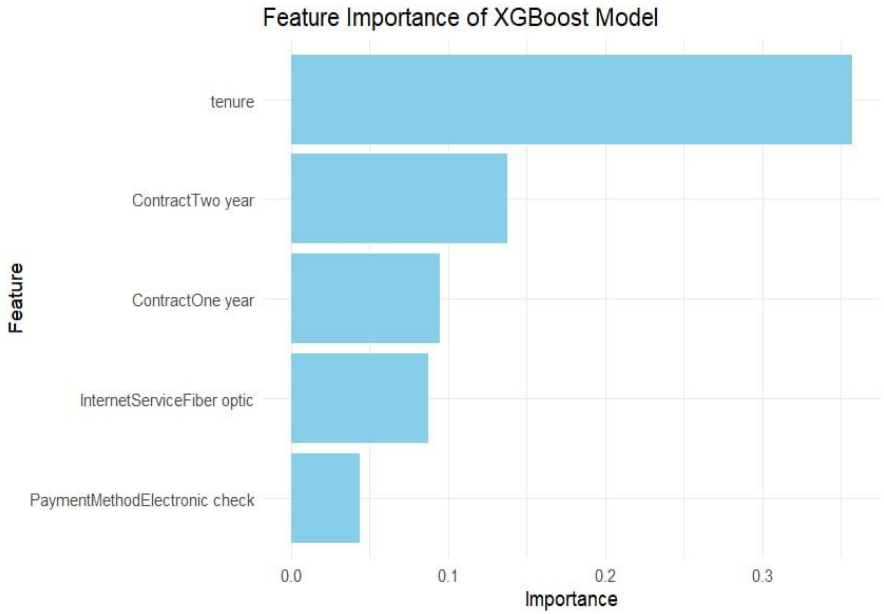
XGboost

	Yes	No
Yes	863	169
No	162	870

Precision	84.2%
Recall	83.6%
F1 Score	83.9%
Accuracy	83.9%

預測顧客是否流失的重要指標





預測顧客是否流失的重要指標

相關性	Random Forest	XGboost
負/負	Tenure	Tenure
負/負	Contract Two Year	Contract Two Year
正/負	InternetService Fiber optic	Contract One Year
正/正	PaymentMethod Electronic check	InternetService Fiber optic
負/正	Contract One Year	PaymentMethod Electronic check

羅吉斯逐步回歸幫助鑑定變數相關性

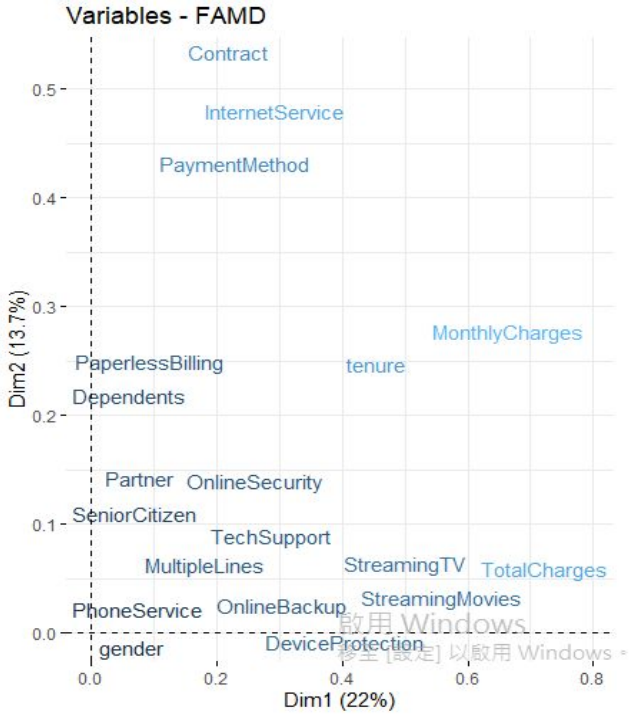
↓

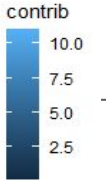
若係數為正則代表正相關

↓

若係數為負則代表負相關

預測顧客是否流失的重要指標 –FAMD分析





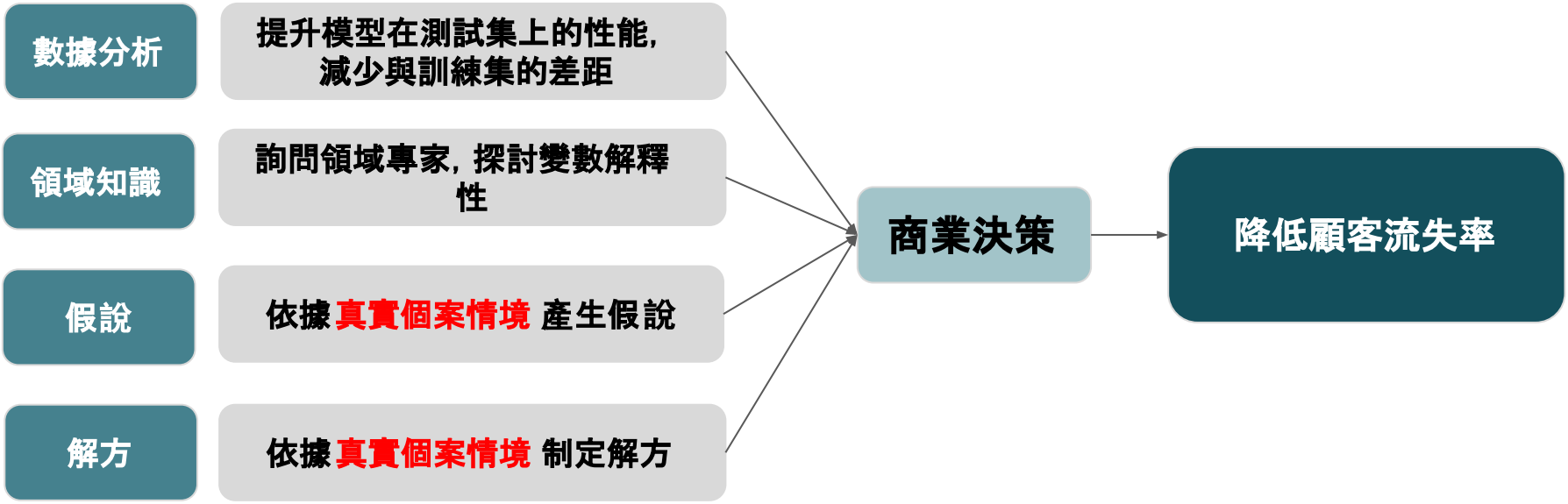
重要變數

- TotalCharges
- MonthlyCharges
- Tenure
- Contract
- InternetService
- PaymentMethod

分析總結

預測效率	預測未來用戶流失的準確率約為83%，且能夠以86%的準確度識別出流失顧客
	對於高風險流失顧客，可 提前採取專業策略以防止顧客流失
Tenure/Charges	使用期間長/總花費越高的顧客流失率較低，可能與顧客忠誠度、轉換成本、顧客慣性有關
Contract	簽訂兩年或一年合約的顧客流失率較低，可能與契約約束、優惠折扣及服務穩定性有關
InternetService	使用光纖網路的顧客流失率較高，可能與服務問題、成本高昂及市場競爭激烈有關
PaymentMethod	使用電子支付的顧客流失率較高，可能與支付過程不流暢及代收費用高昂有關

Next Step - 結合領域知識及商業決策



AGENDA

1. 專案目標
2. 數據洞察
3. 策略建議
4. 成效分析
5. 總結
6. **附錄**

羅吉斯逐步迴歸係數係數

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.570976	0.128203	4.454	8.44e-06	***
DependentsYes	-0.373902	0.070747	-5.285	1.26e-07	***
tenure	-0.032374	0.002024	-15.998	< 2e-16	***
PhoneServiceYes	-0.347918	0.110425	-3.151	0.001629	**
MultipleLinesYes	0.252633	0.072303	3.494	0.000476	***
`InternetServiceFiber optic`	0.938935	0.082032	11.446	< 2e-16	***
InternetServiceNo	-0.766592	0.107698	-7.118	1.10e-12	***
OnlineSecurityYes	-0.524045	0.075367	-6.953	3.57e-12	***
OnlineBackupYes	-0.134981	0.068992	-1.956	0.050409	.
TechSupportYes	-0.467485	0.076298	-6.127	8.95e-10	***
StreamingTVYes	0.239726	0.074792	3.205	0.001349	**
StreamingMoviesYes	0.331994	0.074307	4.468	7.90e-06	***
`Contract One year`	-0.803014	0.090763	-8.847	< 2e-16	***
`Contract Two year`	-1.463869	0.140579	-10.413	< 2e-16	***
PaperlessBillingYes	0.393372	0.065184	6.035	1.59e-09	***
`PaymentMethodCredit card (automatic)`	0.075733	0.099828	0.759	0.448066	
`PaymentMethodElectronic check`	0.498962	0.083925	5.945	2.76e-09	***
`PaymentMethodMailed check`	0.073753	0.099276	0.743	0.457538	

SMOTE-NC原理及參考文章

6.1 SMOTE-NC

While our SMOTE approach currently does not handle data sets with all nominal features, it was generalized to handle mixed datasets of continuous and nominal features. We call this approach Synthetic Minority Over-sampling TEchnique-Nominal Continuous [SMOTE-NC]. We tested this approach on the Adult dataset from the UCI repository. The SMOTE-NC algorithm is described below.

1. Median computation: Compute the median of standard deviations of all continuous features for the minority class. If the nominal features differ between a sample and its potential nearest neighbors, then this median is included in the Euclidean distance computation. We use median to penalize the difference of nominal features by an amount that is related to the typical difference in continuous feature values.

2. Nearest neighbor computation: Compute the Euclidean distance between the feature vector for which k-nearest neighbors are being identified (minority class sample) and the other feature vectors (minority class samples) using the continuous feature space. For every differing nominal feature between the considered feature vector and its potential nearest-neighbor, include the median of the standard deviations previously computed, in the Euclidean distance computation. Table 2 demonstrates an example.

F1 = 1 2 3 A B C [Let this be the sample for which we are computing nearest neighbors]
F2 = 4 6 5 A D E
F3 = 3 5 6 A B K
So, Euclidean Distance between F2 and F1 would be:
Eucl = sqrt[(4-1)^2 + (6-2)^2 + (5-3)^2 + Med^2 + Med^2]
Med is the median of the standard deviations of continuous features of the minority class.
The median term is included twice for feature numbers 5: B→D and 6: C→E, which differ for the two feature vectors: F1 and F2.

3. Populate the synthetic sample: The continuous features of the new synthetic minority class sample are created using the same approach of SMOTE as described earlier. The nominal feature is given the value occuring in the majority of the k-nearest neighbors.

The SMOTE-NC experiments reported here are set up the same as those with SMOTE, except for the fact that we examine one dataset only. SMOTE-NC with the Adult dataset differs from our typical result: it performs worse than plain under-sampling based on AUC, as shown in Figures 26 and 27. We extracted only continuous features to separate the effect of SMOTE and SMOTE-NC on this dataset, and to determine whether this oddity was due to our handling of nominal features. As shown in Figure 28, even SMOTE with only continuous features applied to the Adult dataset, does not achieve any better performance than plain under-sampling. Some of the minority class continuous features have a very high variance, so, the synthetic generation of minority class samples could be overlapping with the majority class space, thus leading to more false positives than plain under-sampling. This hypothesis is also supported by the decreased AUC measure as we SMOTE at degrees greater than 50%. The higher degrees of SMOTE lead to more minority class samples in the dataset, and thus a greater overlap with the majority class decision space.

https://www3.nd.edu/~dial/publications/chawla2002smote.p

Thank you for watching