NCCU DS 第四組

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

專案成員

企管三 陳勁瑋
企管三 張育維
資科三 簡佑成
資科三 李典陽
資科三 徐宏宇
統計三 周幼臻

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

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，傳統的預測方法無法知曉數據背後的意涵
改善方式	透過分析客戶行為數據，識別高風險的流失客戶
	從分析的結果，採取針對性的商業策略
專案目標	找出顯著影響顧客是否流失的特徵
	建立有效的預測模型，分析新顧客未來流向

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

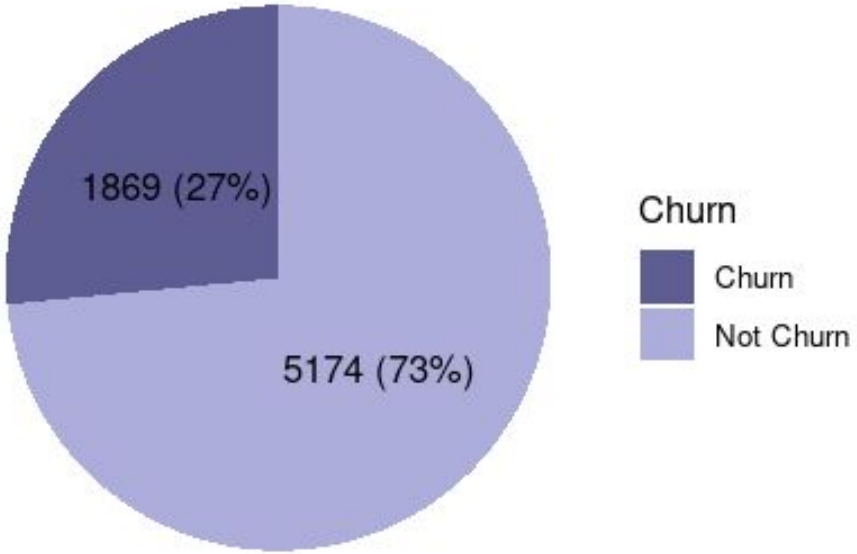
- Gender
- SeniorCitizen
- Partner
- Dependents
- Contract
- PhoneService
- StreamingTV
- MutipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- TechSupport
- PaymentMethod
- PaperlessBilling
- DeviceProtection
- StreamingMovies

連續變數

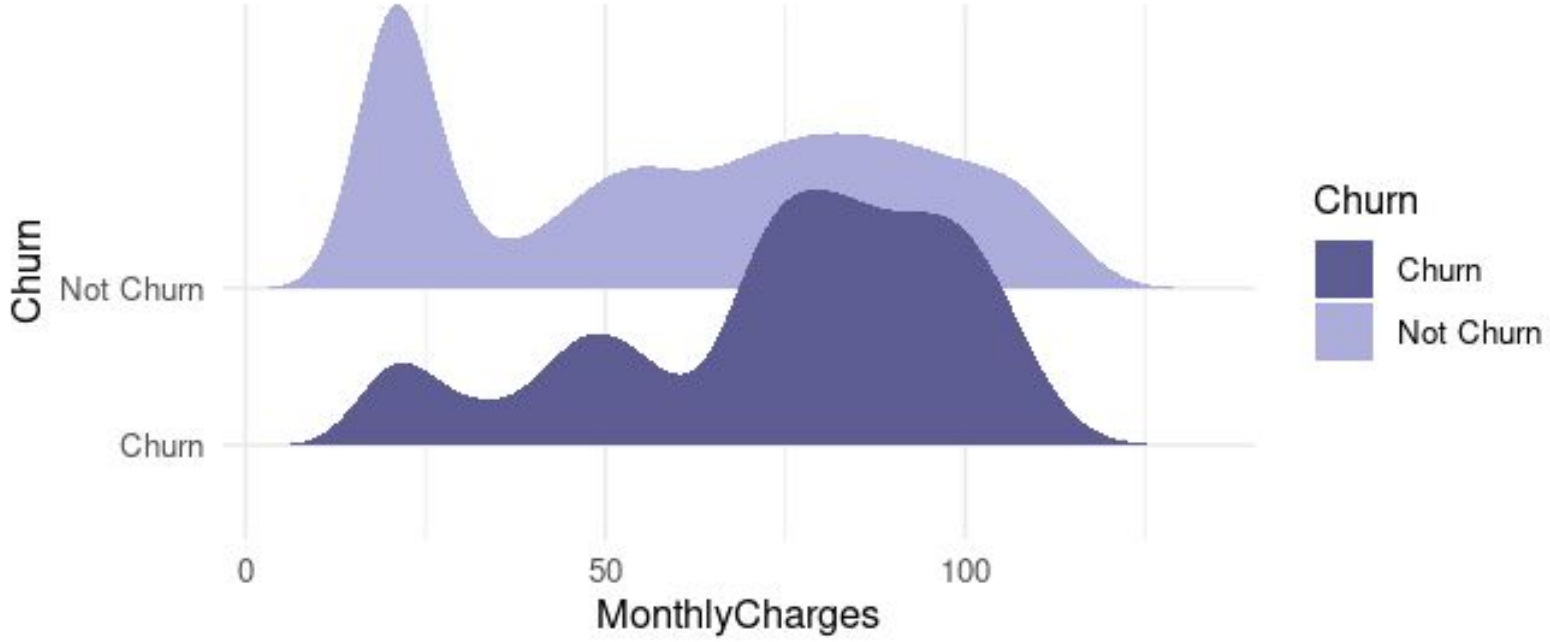
- Tenure
- MonthlyCharges
- TotalCharges

→ **Churn**

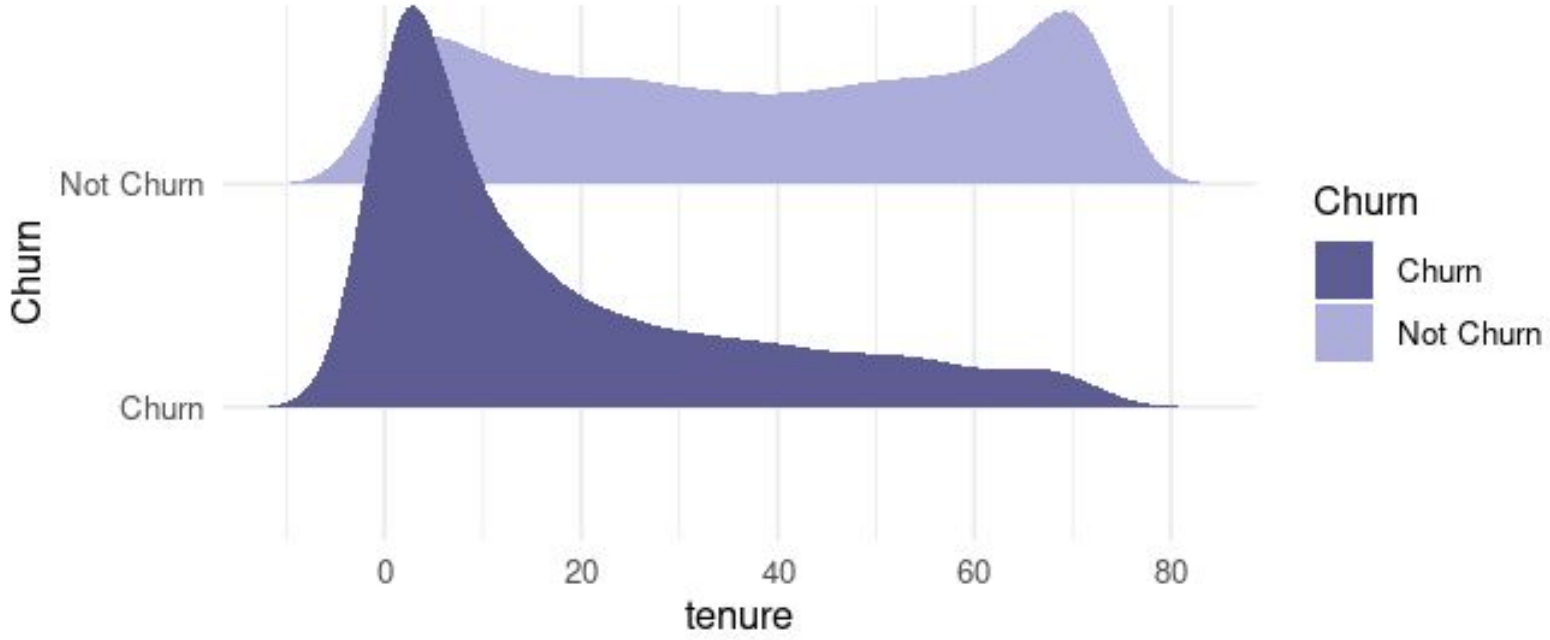
由下圖發現流失:未流失顧客約為1:3



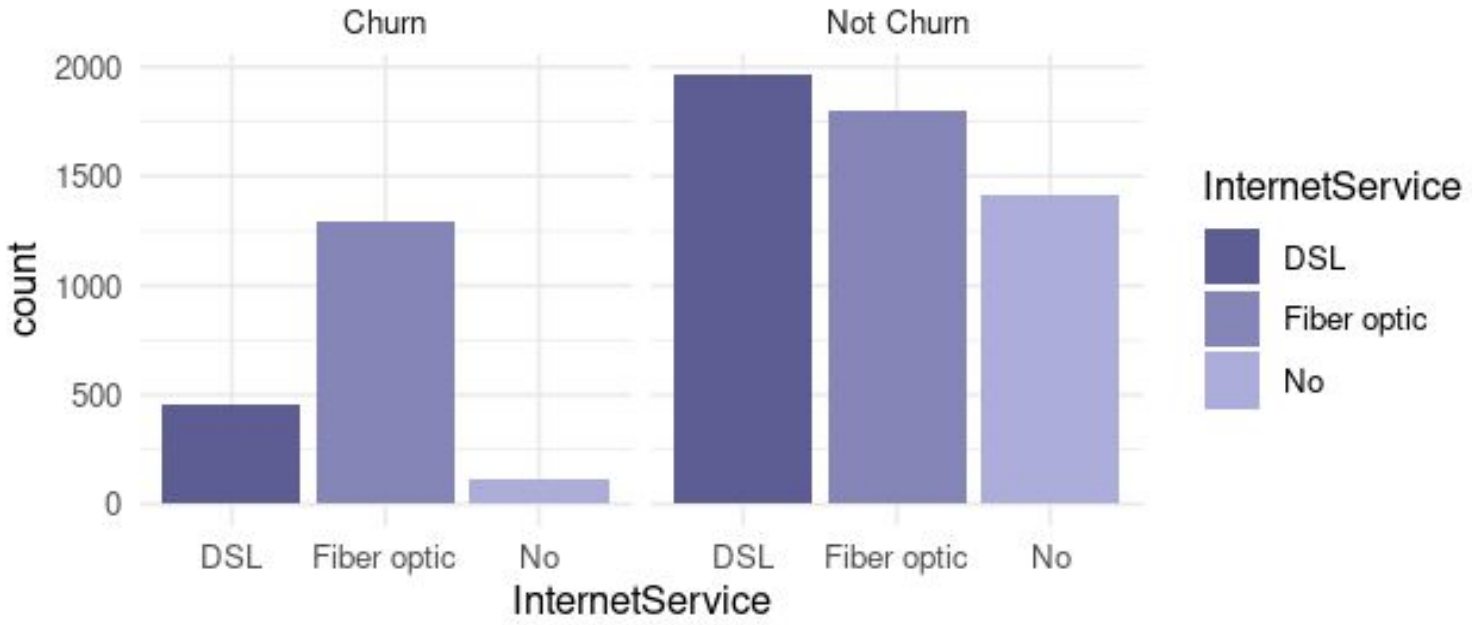
流失、未流失顧客皆呈現左偏，但未流失顧客在較低月支付有較高的峰值



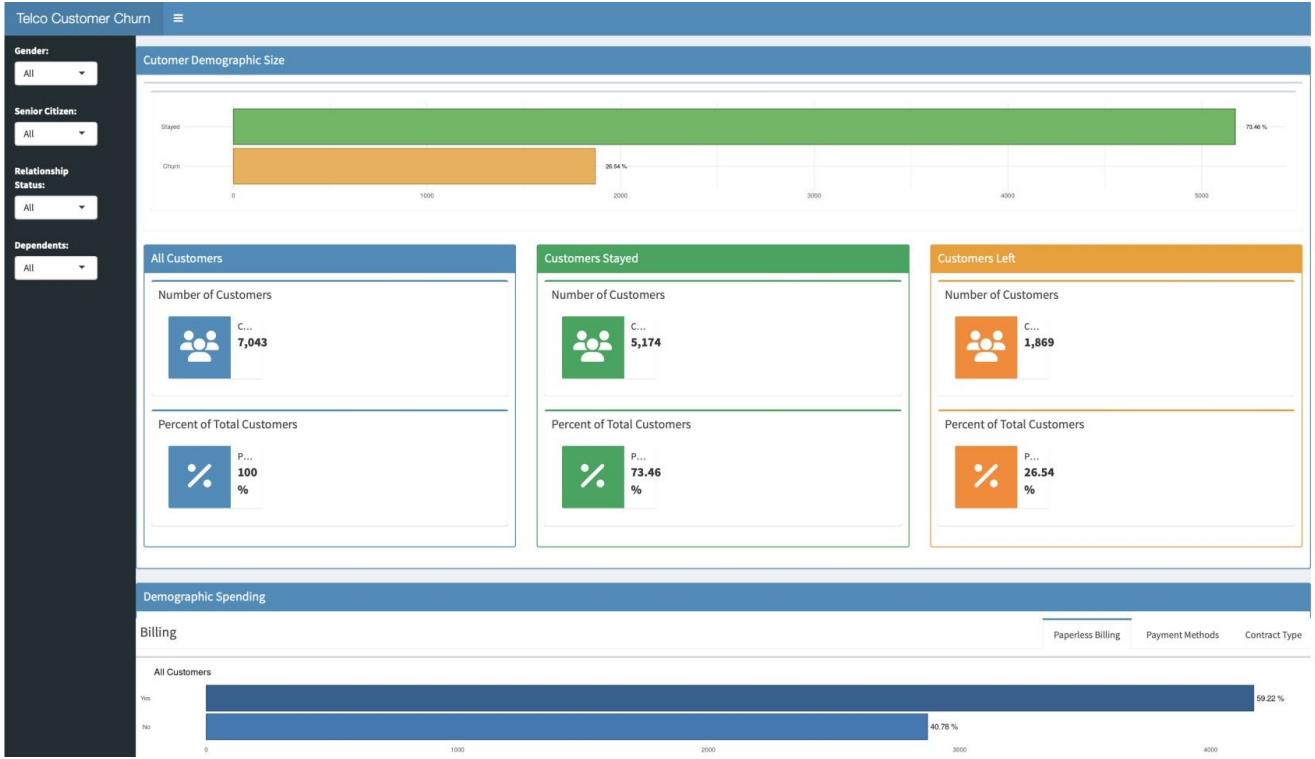
Tenure中的未流失顧客為雙峰分布，流失顧客則呈現右偏



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

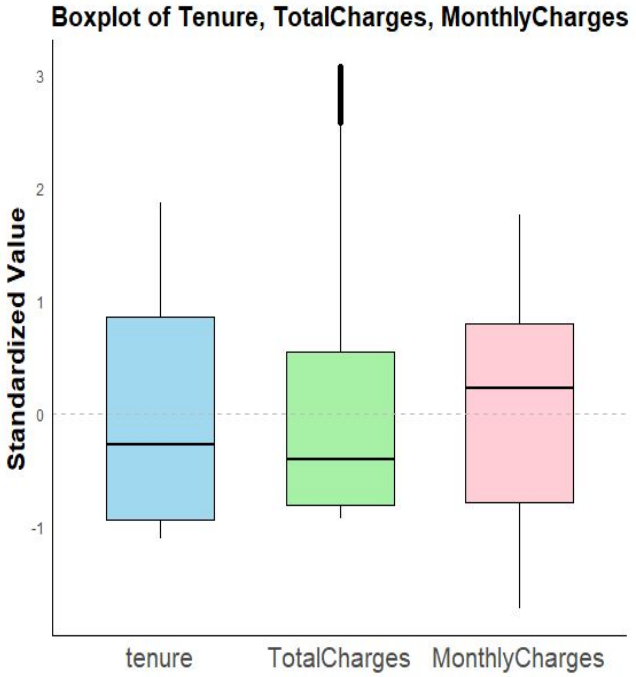


AGENDA

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



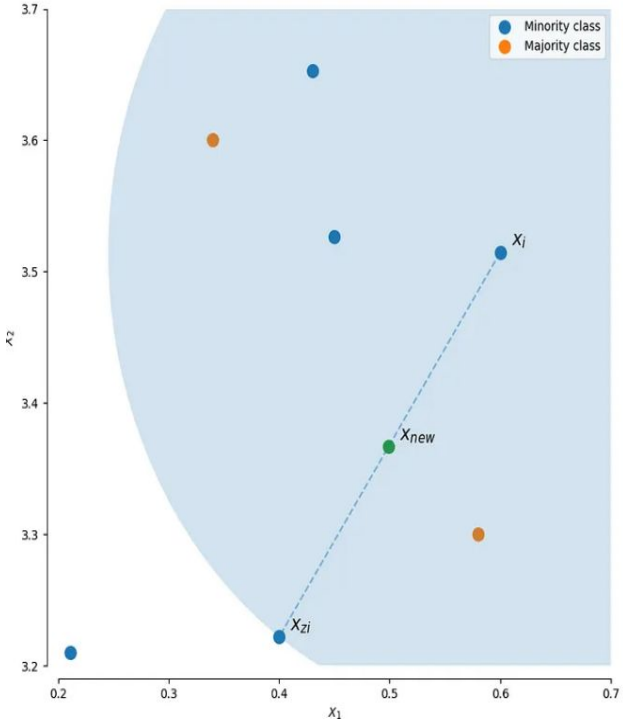
步驟二:類別變數處理

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

步驟三: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

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

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

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

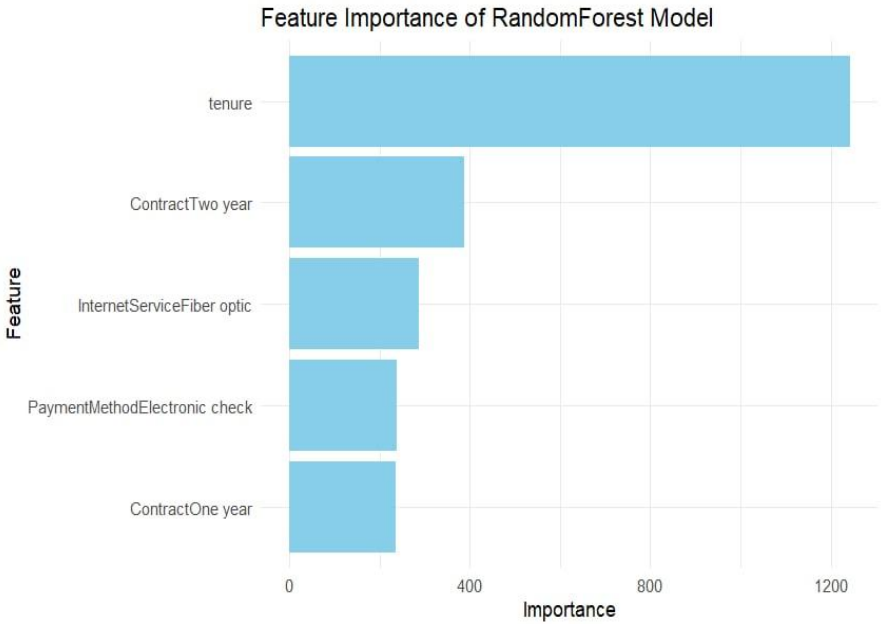
Precision	80.0%
Recall	86.6%
F1 Score	83.2%
Accuracy	82.5%

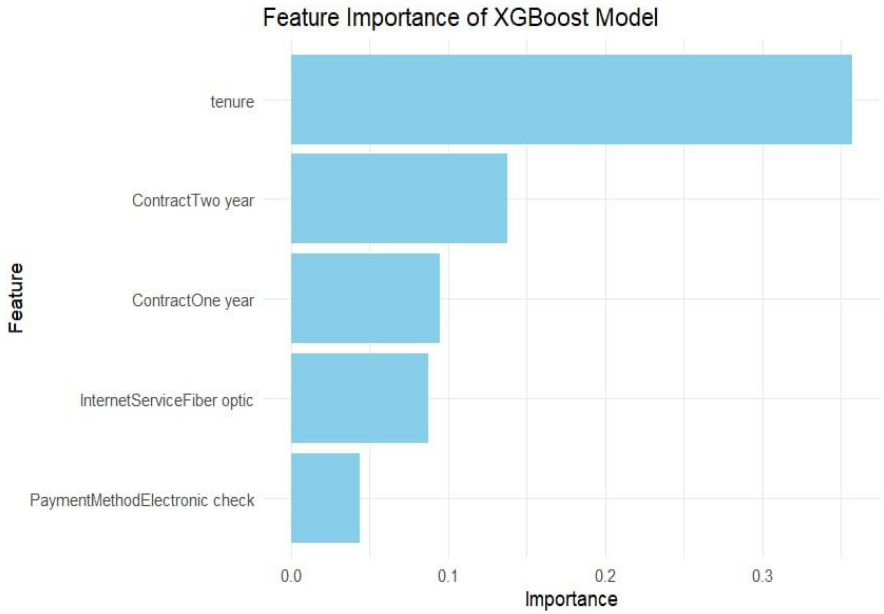
XGboost

	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

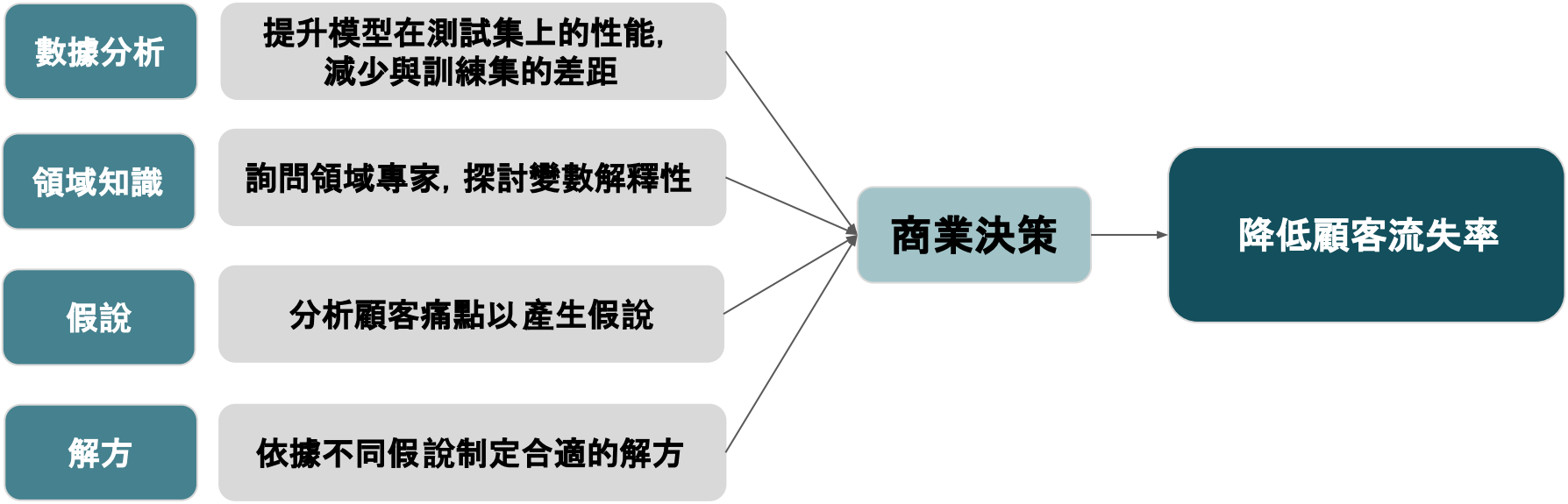
↓

若係數為正則代表正相關

↓

若係數為負則代表負相關

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**

數據處理使用程式碼

```
library(tidyverse)
library(data.table)
library(randomForest)
library(themis)
library(MASS)
library(caret)
library(car)
library(ranger)
```

```
data <- data %>%
  drop_na() %>%
  distinct()
data <- data[, -1]

# 以下的type轉成no
data <- data %>%
  mutate(
    MultipleLines = case_when(
      MultipleLines == "No phone service" ~ "No",
      TRUE ~ as.character(MultipleLines)
    ),
.......
```

(數據處理處有提到的變數皆需轉換)

數據處理使用程式碼

```
# as factor

data <- data %>%
  mutate(gender = factor(gender, levels =
c("Female", "Male"))) %>%

  mutate(SeniorCitizen = factor(SeniorCitizen,
levels = c(0, 1), labels = c("notsenior",
"senior"))) %>%

  mutate(Partner = factor(Partner, levels =
c("No", "Yes"))) %>%
.......
```

(數據處理處有提到的類別變數皆需轉換)

```
# smotenc oversample

set.seed(111)
new_data <- smotenc(data,
"Churn", k =5,  over_ratio =
1)
```

(使用themis包的smotenc進行oversample)

數據處理使用程式碼

```
# train and test

set.seed(111)
step_train_index <- createDataPartition(new_data$Churn, p = 0.8, list =
FALSE)

step_train_data <- new_data[step_train_index,]

step_test_data <- new_data[-step_train_index,]
```

（切分訓練集與測試集）

特徵選擇使用程式碼

```
## 2.選擇特徵

# 建造羅吉斯迴歸full model
model_test <- glm(Churn ~., step_train_data, family = binomial)

# full model的vif
summary(model_test)
vif(model_test)    #    把最大的拿掉

formula1 <- formula(model_test)
new_1 <- update(formula1, . ~ . - MonthlyCharges)
final_one<- update(model_test, formula = new_1)
vif(final_one)
```

（重複此步驟以確保模型中的變數不具有共線性 ）

特徵選擇使用程式碼

```
#   使用逐步迴歸進行變數篩選

final_model <- stepAIC(final_two, direction = "both")
vif(final_model)
summary(final_model)
```

隨機森林交叉驗證使用程式碼

```
## 設置交叉驗證使用的超參數
random_control <- trainControl(method = "cv", number = 5)
grid <- expand.grid(
  mtry = c(2, 3, 4, 5, 6),
  splitrule = c("gini"),
  min.node.size = c(1, 3, 5, 7, 9)
)
```

（設置隨機森林使用的交叉驗證超參數範圍 ）

隨機森林交叉驗證使用程式碼

```
# Train the model

set.seed(111)
rf_model <- train(
  Churn ~ Dependents + tenure + PhoneService + MultipleLines +
    InternetService + OnlineSecurity + OnlineBackup + TechSupport +
StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod,
  data = step_train_data,
  method = "ranger",
  trControl = random_control,
  tuneGrid = grid,
  importance = 'impurity'
)
```

（訓練隨機森林模型，變數使用逐步迴歸的結果 ）

XGboost交叉驗證使用程式碼

```
## 設置交叉驗證使用的超參數

set.seed(111)
xg_control <- trainControl(method = "cv", number = 10)
xg_grid <- expand.grid(
  nrounds = c(100, 150),
  max_depth = c(6, 8),
  eta = c(0.1, 0.3),
  gamma = c(0, 0.1),
  colsample_bytree = c(0.8, 1.0),
  subsample = c(0.8, 1.0),
  min_child_weight = c(1, 3)
)
```

（設置XGboost使用的交叉驗證超參數範圍）

XGboost交叉驗證使用程式碼

```
# Train the model

set.seed(111)
xgb_model <- train(
  Churn ~ Dependents + tenure + PhoneService + MultipleLines +
  InternetService + OnlineSecurity + OnlineBackup + TechSupport +
StreamingTV +
  StreamingMovies + Contract + PaperlessBilling + PaymentMethod,
  data = step_train_data,
  method = "xgbTree",
  trControl = xg_control,
  tuneGrid = xg_grid
)
```

（訓練XGboost模型，變數使用逐步迴歸的結果）

隨機森林混淆矩陣

```
# train data
predictions <- predict(rf_model, newdata = step_train_data)
conf_matrix <- confusionMatrix(predictions, step_train_data$Churn)
print(conf_matrix)


# test data
rf_prediction <- predict(rf_model, step_test_data, type = "raw")
rf_confusion_matrix <- confusionMatrix(rf_prediction, step_test_data$Churn)
print(rf_confusion_matrix)
```

（查看隨機森林訓練集及測試集的混淆矩陣 ）

XGboost混淆矩陣

```
## train data
train_predictions <- predict(xgb_model, newdata = step_train_data)
train_conf_matrix <- confusionMatrix(train_predictions,
step_train_data$Churn)
print(train_conf_matrix)


# test data
xg_prediction <- predict(xgb_model, step_test_data, type = "raw")
xg_confusion_matrix <- confusionMatrix(xg_prediction, step_test_data$Churn)
print(xg_confusion_matrix)
```

（查看Xgboost訓練集及測試集的混淆矩陣）

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 listening