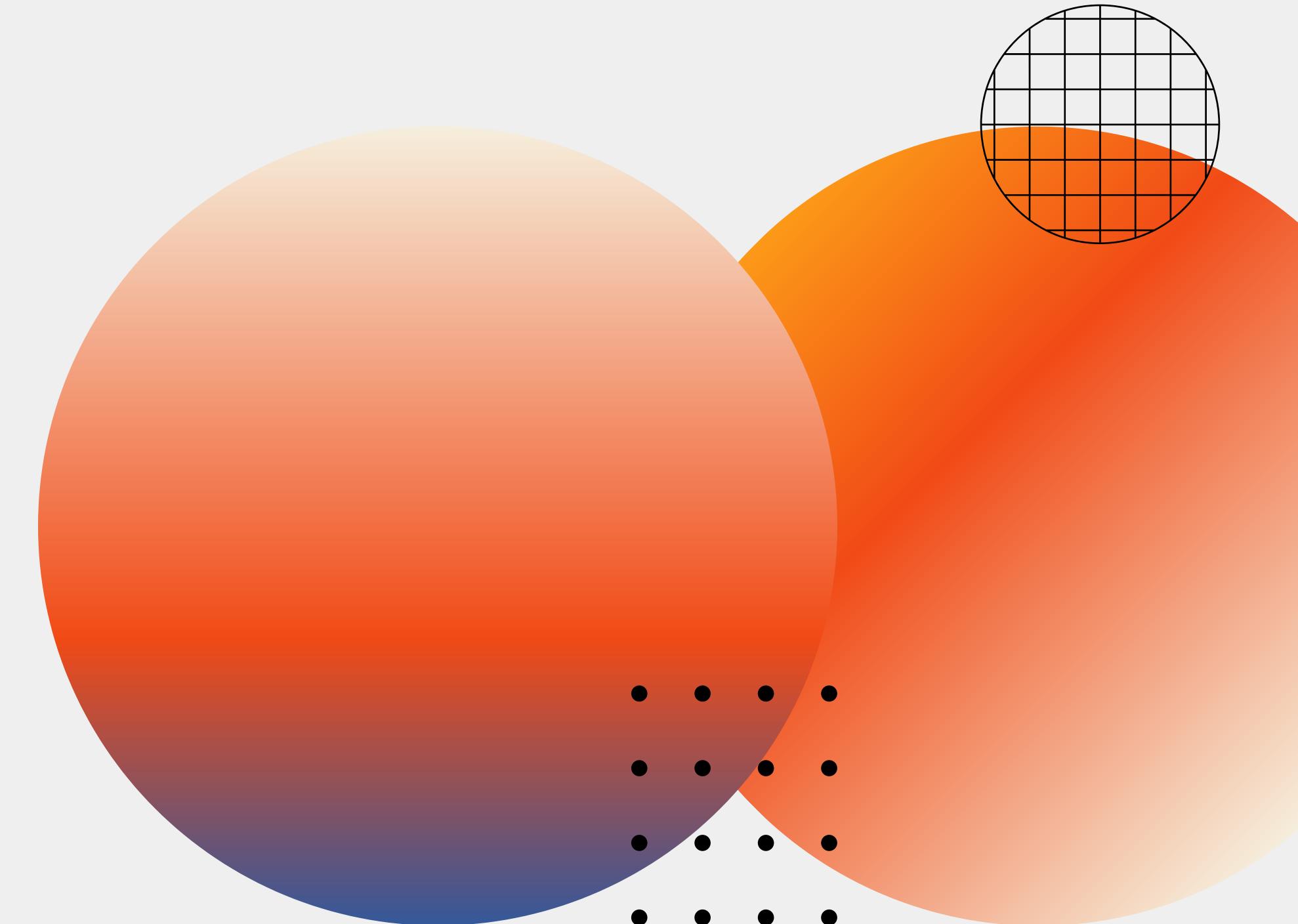


Let's Start

Telecom Customer Churn Prediction

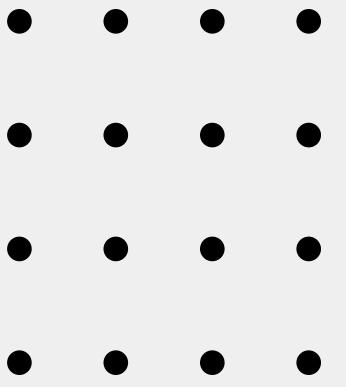
GAD251 第二組 - 林靖軒 陳晏綾 孫敏嘉 吳建儒



Agenda

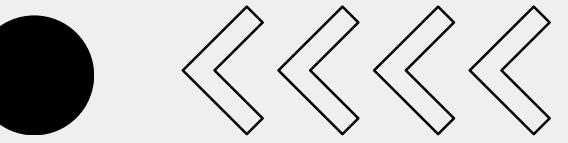


專案流程圖與分工
專案進度時程表
資料集介紹 & 前處理
模型建構 & 優化歷程
SHAP洞察與策略發想
介面設計與部署
問題與解決方案
結論



Project Flowchart

顧客流失預測



吳建儒



陳晏綾



孫敏嘉



林靖軒 (組長)

資料前處理

模型訓練

SHAP洞察與策略發想

介面設計與部署

資料探索 (EDA)

建立基線模型 (邏輯回歸)

分析特徵重要性

介面設計 (Streamlit)

資料清理

多模型訓練 + 調參
(隨機森林、*XGBoost*、*LightGBM*)

流失顧客留存策略

開發與部署

特徵工程

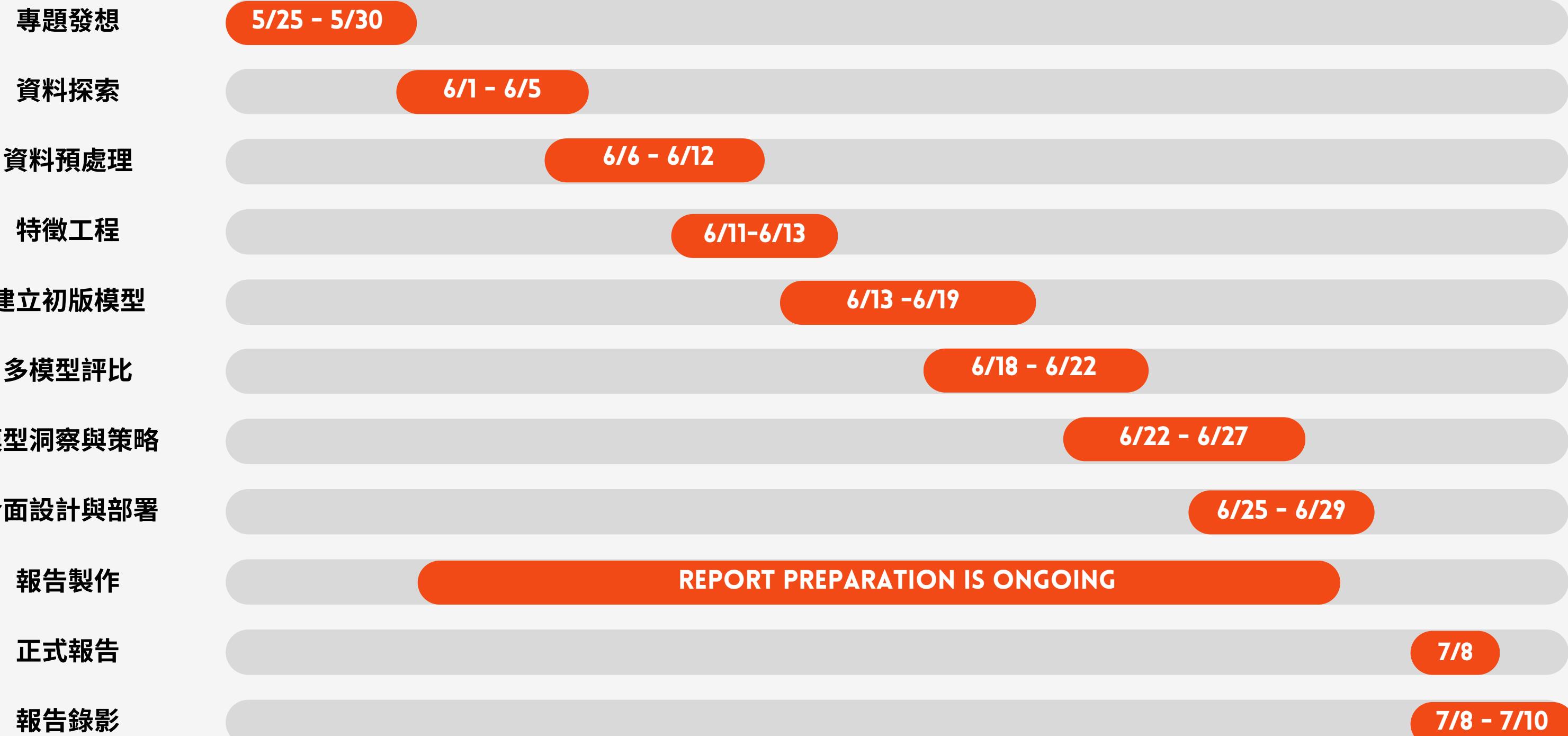
評比選出最終模型

每階段分派一位主負責人，全專案組員皆共同參與協作。

PROJECT TIMELINE

MAY W4

JUL W2

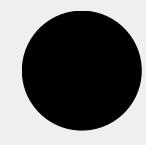
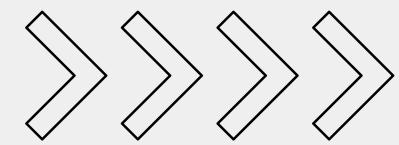




吳建儒

資料集介紹 & 前處理

Dataset Introduction & Preprocessing



資料集介紹



吳建儒

資料來源

kaggle - Telecom Customer Churn Prediction
(Maven Telecom Churn Challenge)

資料筆數

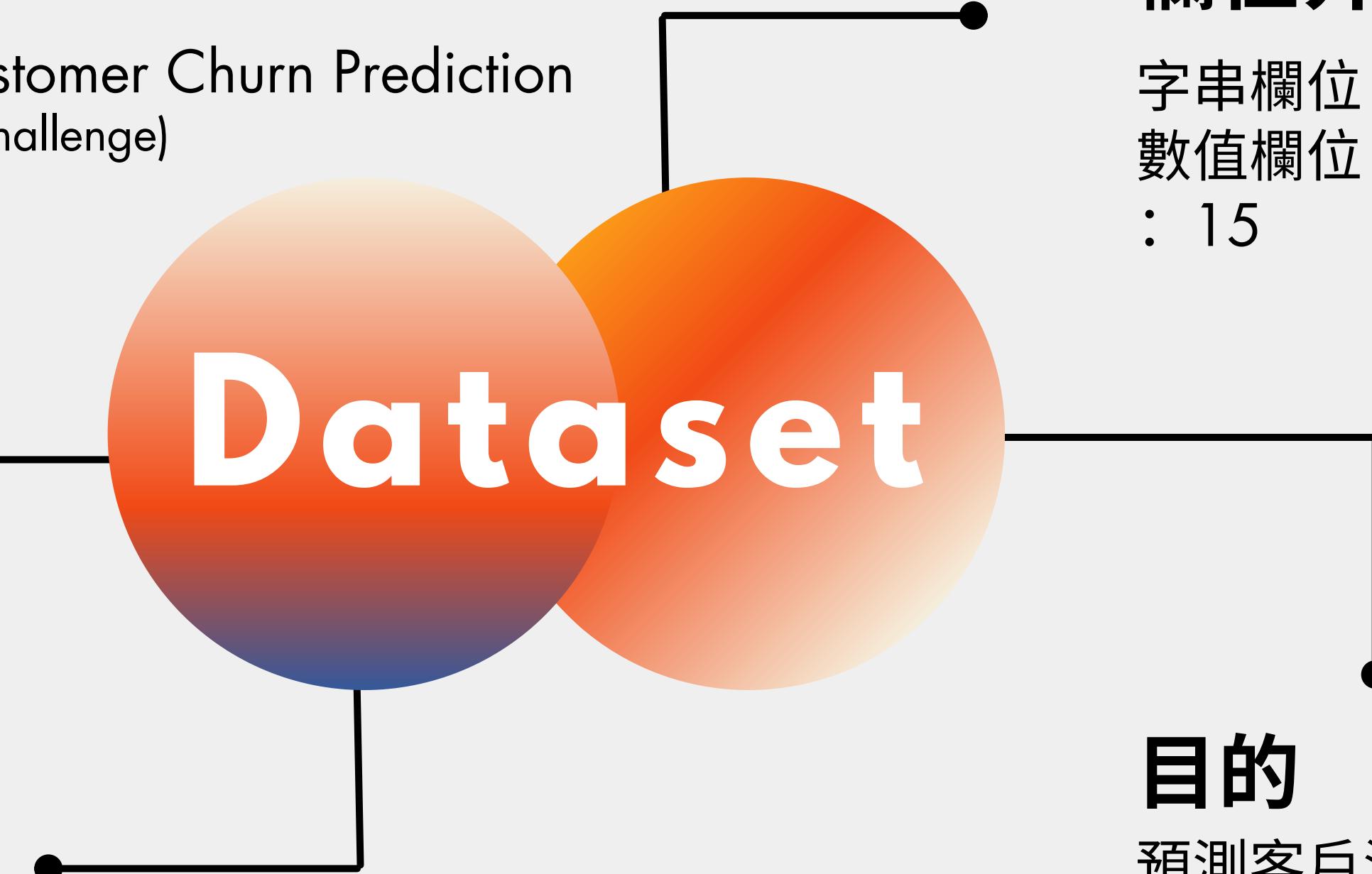
7043筆

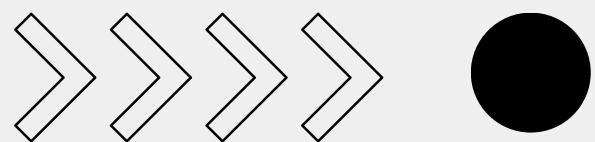
欄位介紹

字串欄位 (object) : 23
數值欄位 (int64/float64) : 15

目的

預測客戶流失機率，
提供企業有效留客策略





資料集介紹



吳建儒

Telecom Customer Churn Prediction

Why customers churn? How can you improve customer retention?

Data Card Code (54) Discussion (1) Suggestions (0)

About Dataset

Contents

This dataset contains 2 tables, in CSV format:

- The Customer Churn table contains information on all 7,043 customers from a Telecommunications company.
- Each record represents one customer, and contains details about their demographics, location, tenure, subscription quarter (joined, stayed, or churned), and more!
- The Zip Code Population table contains complimentary information on the estimated populations for the Customers in the Customer Churn table.

Collection Methodology

The public dataset is completely available on the Maven Analytics website platform where it stores and consolidates analysis in the Data Playground. The specific telecom customer churn dataset at hand can be obtained in this link: <https://www.mavenanalytics.io/blog/maven-churn-challenge>

This score is calculated by Kaggle.

Completeness · 100%

- ✓ Subtitle
- ✓ Tag
- ✓ Description
- ✓ Cover Image

Credibility · 100%

- ✓ Source/Provenance
- ✓ Public Notebook
- ✓ Update Frequency

Compatibility · 100%

- ✓ License
- ✓ File Format
- ✓ File Description
- ✓ Column Description

Jun 23, 2022 / Data Challenges

Introducing the Maven Churn Challenge

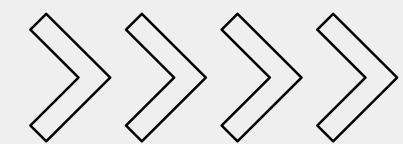
3 min read

Enrique Ruiz
Sr. Learning Experience Designer

✓ 資料品質佳 ✓ 實務導向

✓ 高共鳴性

• • •
• • •
• • •
• 7 •
• • •



資料前處理



吳建儒

空值

1526: 補 "No"

Internet Service (Yes) → Online Security (Yes / No)

Internet Service (No) → Online Security (空值)

5174: 補 "Not Churned"

Customer Status → 4720(Stayed) + 1869(Churned) + 454(Joined)

3877: 補 "No"

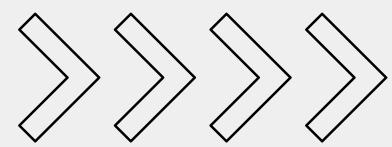
占比55%，邏輯判斷非空值，此為接受的行銷方案種類，故可能沒有接受任何行銷方案

682: 補 "No"

Phone Service (Yes) → Multiple Lines (Yes / No)

Phone Service (No) → Multiple Lines (空值)

	類別型欄位名稱	類別數量	非空值數	空值數	空值比例
0	Customer ID	7043	7043	0	0.000000
1	Gender	2	7043	0	0.000000
2	Married	2	7043	0	0.000000
3	City	1106	7043	0	0.000000
4	Offer	5	3166	3877	0.550476
5	Phone Service	2	7043	0	0.000000
6	Multiple Lines	2	6361	682	0.096834
7	Internet Service	2	7043	0	0.000000
8	Internet Type	3	5517	1526	0.216669
9	Online Security	2	5517	1526	0.216669
10	Online Backup	2	5517	1526	0.216669
11	Device Protection Plan	2	5517	1526	0.216669
12	Premium Tech Support	2	5517	1526	0.216669
13	Streaming TV	2	5517	1526	0.216669
14	Streaming Movies	2	5517	1526	0.216669
15	Streaming Music	2	5517	1526	0.216669
16	Unlimited Data	2	5517	1526	0.216669
17	Contract	3	7043	0	0.000000
18	Paperless Billing	2	7043	0	0.000000
19	Payment Method	3	7043	0	0.000000
20	Customer Status	3	7043	0	0.000000
21	Churn Category	5	1869	5174	0.734630



資料前處理



吳建儒

空值

1526: 補"0"

Internet Service (Yes) → Avg Monthly GB Download (有數值)

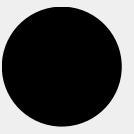
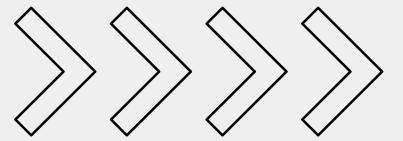
Internet Service (No) → Avg Monthly GB Download (空值)

682: 補"0"

Phone Service (Yes) → Avg Monthly Long Distance Charges (有數值)

Phone Service (No) → Avg Monthly Long Distance Charges (空值)

	數值型欄位名稱	非空值數	空值數	空值比例
0	Age	7043	0	0.000000
1	Number of Dependents	7043	0	0.000000
2	Zip Code	7043	0	0.000000
3	Latitude	7043	0	0.000000
4	Longitude	7043	0	0.000000
5	Number of Referrals	7043	0	0.000000
6	Tenure in Months	7043	0	0.000000
7	Avg Monthly Long Distance Charges	6361	682	0.096834
8	Avg Monthly GB Download	5517	1526	0.216669
9	Monthly Charge	7043	0	0.000000
10	Total Charges	7043	0	0.000000
11	Total Refunds	7043	0	0.000000
12	Total Extra Data Charges	7043	0	0.000000
13	Total Long Distance Charges	7043	0	0.000000
14	Total Revenue	7043	0	0.000000



資料前處理



吳建儒

類別合併

Customer Status (Y) → 4720(Stayed) + 1869(Churned) + 454(Joined)

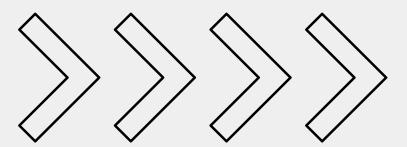
Joined: 使用月份數3個月以下 (**新進客戶**)

Stayed: 使用月份數4個月以上 (**舊客戶**)

Churned: 使用月份數3個月以下 & 4個月以上 (**流失客戶**)

Stayed = Stayed + Joined (**先刪除1個月的新進客戶**)

Churned = Churned



資料前處理



吳建儒

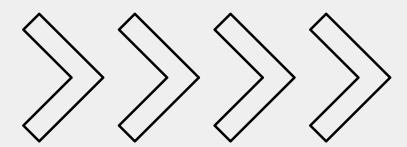
洩漏欄位（會暴露流失結果）

發現：

屬於事後資訊，若保留會使模型提早知曉結果。

做法：直接刪除。

- Churn Category (流失分類結果)
- Churn Reason (流失原因，屬於事後資訊)



資料前處理



吳建儒

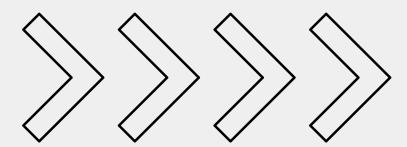
洩漏欄位（會暴露流失結果）

發現：

屬於**累積型**特徵，若保留有機會使模型提早知曉結果。

做法：直接刪除。

- Total Revenue
- Total Charges
- Total Refunds
- Total Extra Data Charges
- Total Long Distance Charges



資料前處理



吳建儒

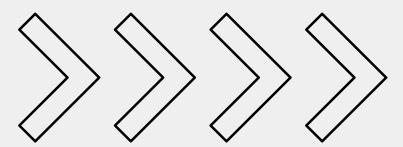
無實質預測意義的欄位

發現：

內容固定或不可操作，無法帶入策略應用。

做法：刪除地理與不可解讀欄位。

→ City、Latitude、Longitude、Zip Code、Customer ID
(皆為加州，預計不會讓使用者於介面上操作)



資料前處理



吳建儒

不合實務考量的資料

發現：

出現異常值，與實際情境不符。

→ Monthly Charge 為負值的資料共 120 筆（約1.7%）。

檢查：

與使用月份數相乘 → 未得到合理結論

猜測與退費有關 → 有退費為0，Monthly Charge為負值的資料

做法：刪除異常值資料



陳晏綾

模型建構 & 優化歷程

Model Development and Optimization Process

資料切分與編碼策略



陳晏綾

資料切分邏輯

資料切分比例：Train 64%、Val 16%、Test 20%

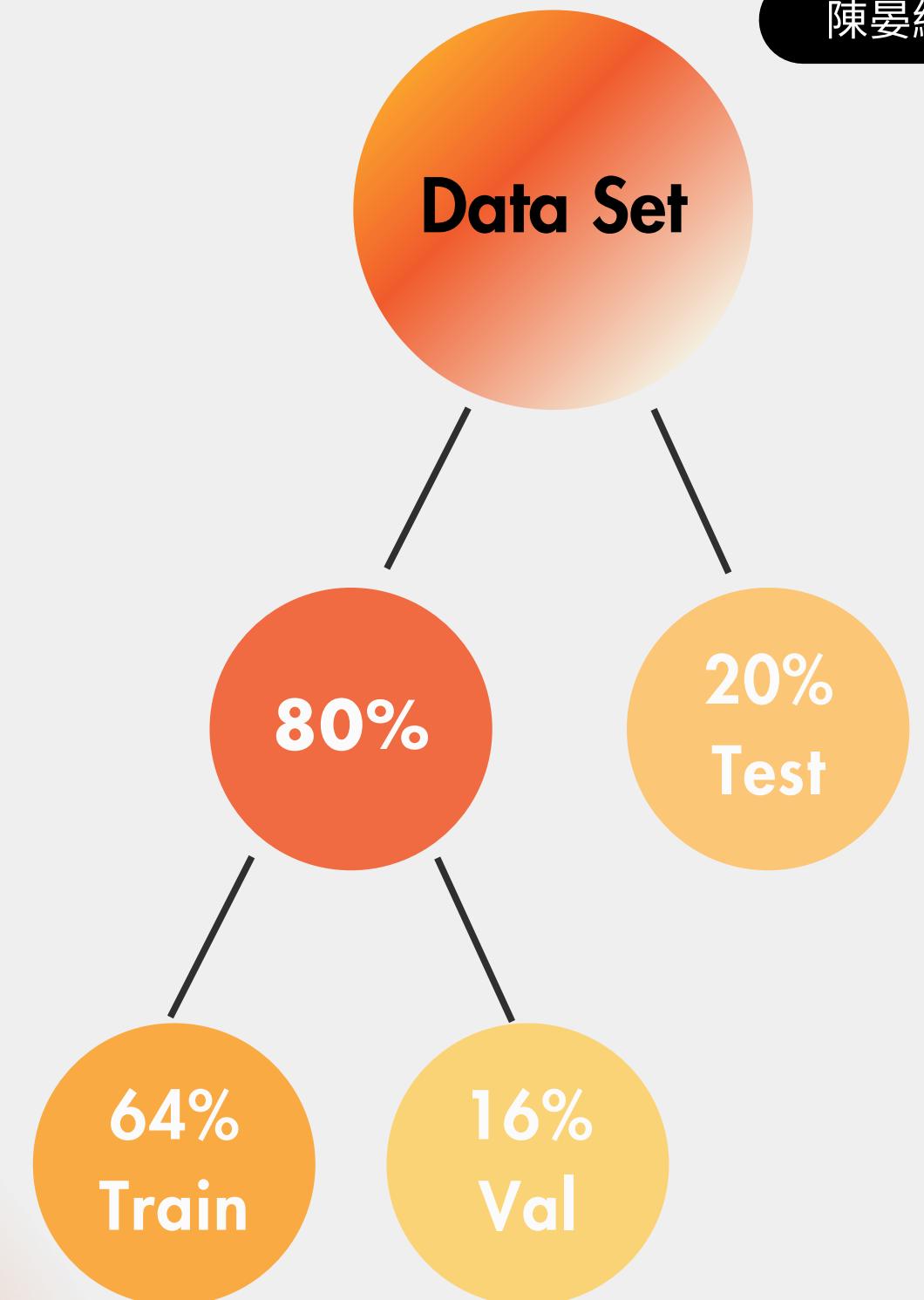
編碼方式選擇

根據模型性質調整編碼方式：

- Logistic Regression → One-hot + MinMaxScaler
- Tree-based → Label

統一轉換邏輯

為了防止資料洩漏，編碼器僅對訓練集進行擬合，
並套用至驗證集與測試集





模型初選與基準比較

使用模型	Accuracy	AUC	Precision (1)	Recall (1)	F1-score (1)
邏輯回歸	0.83	0.90	0.71	0.68	0.69
隨機森林	0.84	0.91	0.77	0.64	0.70
XGBoost	0.86	0.92	0.77	0.69	0.73
LightGBM	0.86	0.92	0.77	0.70	0.74

- 指標計算基於 Validation Set、分類對象為正類 (class=1)

✓ 我們選用邏輯迴歸作為基準模型，其簡單性與可解釋性提供了良好起點。



陳晏綾

精選模型優化 PK

✓ 調參策略：

- GridSearchCV
- RandomizedSearchCV
- 手動調參
- 交叉驗證

✓ 特徵篩選：

- Feature Importance
- Gain 累積貢獻
- Permutation Importance
- RFE

使用模型	特徵版本	Accuracy	AUC	Precision (1)	Recall (1)	F1-score (1)	備註
LightGBM	PI 取 TOP10	0.74	0.92	0.51	0.95	0.67	Grid Search
XGBoost	Gain 80% (10 欄)	0.82	0.91	0.63	0.86	0.73	手動調參
XGBoost	RFE Top 13	0.82	0.92	0.63	0.84	0.72	Random Search
XGBoost	Gain Top 7	0.69	0.92	0.47	0.99	0.63	Random Search
XGBoost	Gain 95% (10 欄)	0.73	0.90	0.51	0.93	0.66	Random Search

► 指標計算基於 Validation Set、Threshold = 0.5

最終模型樣貌

XGBoost



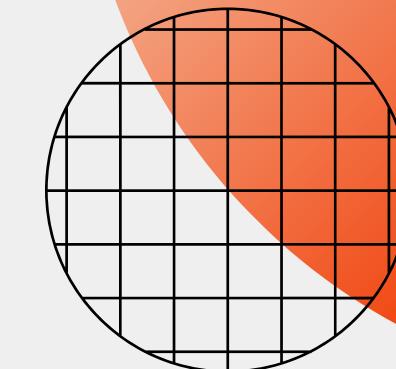
陳晏綾

✓ 參數設定：

參數名稱	設定值	備註
objective	'binary:logistic'	二元分類問題
eval_metric	'aucpr'	以 Precision-Recall AUC 為評估指標
random_state	42	固定隨機種子，確保可重現性
n_estimators	100	樹的數量
learning_rate	0.1	學習率
max_depth	5	模型複雜度
scale_pos_weight	3	處理類別不平衡問題

最終模型樣貌

XGBoost



陳晏綾

✓ 特徵版本：

Contract、Internet Service、Number of Referrals、Monthly Charge、Number of Dependents、Married、Streaming Movies、Streaming TV、Tenure in Months、Online Security

Gain 累積貢獻	欄位數	Accuracy	AUC	Precision (1)	Recall (1)	F1-score (1)
78%	9	0.82	0.91	0.62	0.86	0.72
79-81%	10	0.83	0.91	0.63	0.86	0.73
82-83%	11	0.82	0.92	0.63	0.86	0.73
84-85%	12	0.83	0.92	0.64	0.85	0.73
86-87%	13	0.83	0.92	0.64	0.84	0.73



陳晏綾

模型測試結果

- ▶ Threshold = 0.5 此版本將用於後續部署

使用模型	特徵版本	Accuracy	AUC	Precision (1)	Recall (1)	F1-score (1)
XGBoost	Gain 80% (10 欄)	0.84	0.93	0.65	0.88	0.75

- ▶ Threshold = 0.31 (Recall + Precision 最大) 可依實際狀況考量再做調整

使用模型	特徵版本	Accuracy	AUC	Precision (1)	Recall (1)	F1-score (1)
XGBoost	Gain 80% (10 欄)	0.79	0.93	0.57	0.93	0.71





孫敏嘉

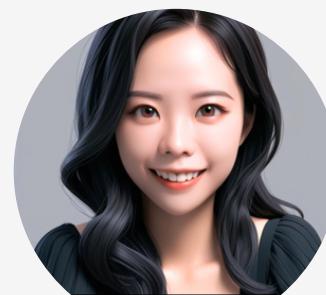
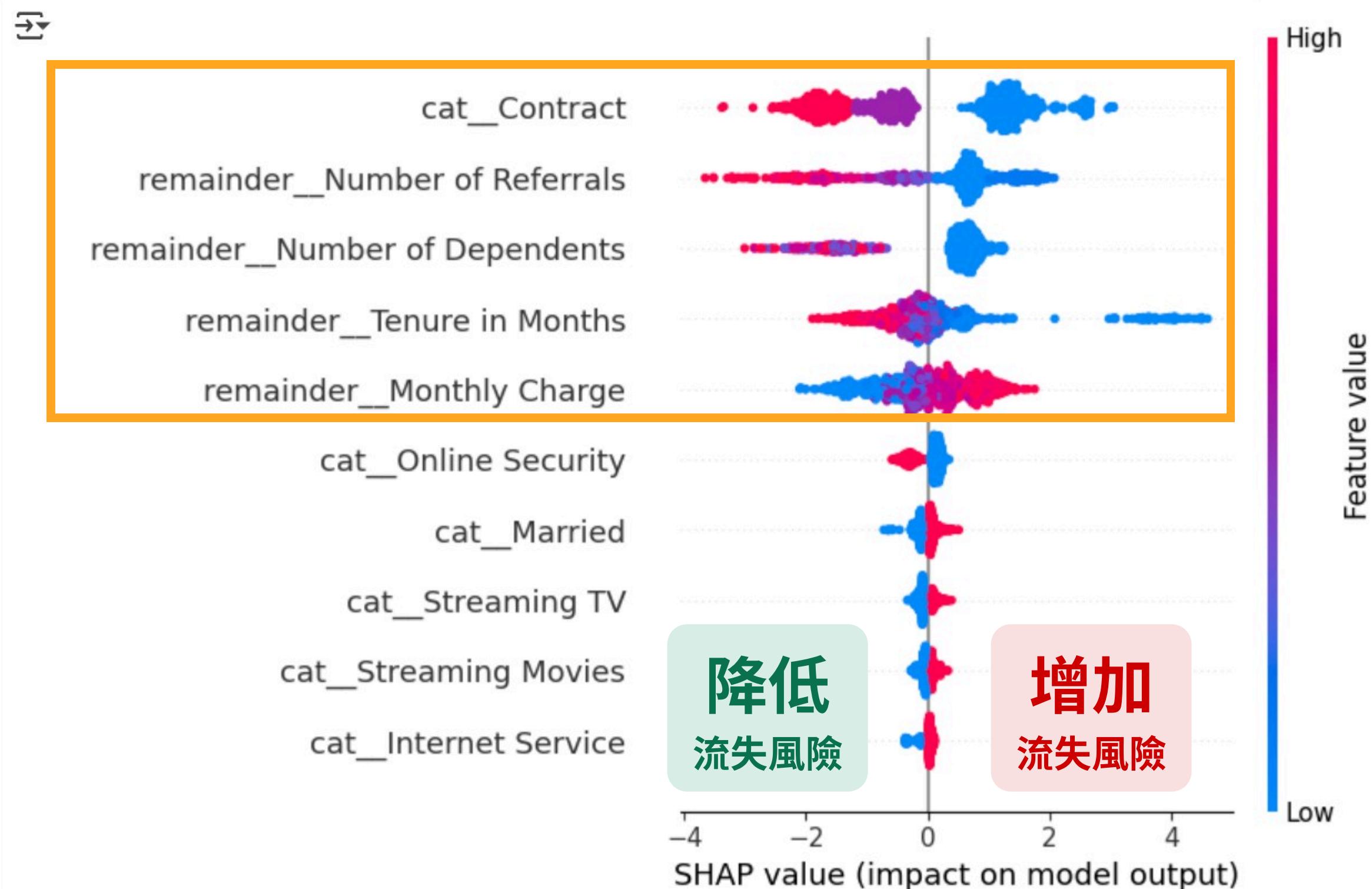
SHAP洞察與策略發想

SHAP Explain and Strategic Recommendations

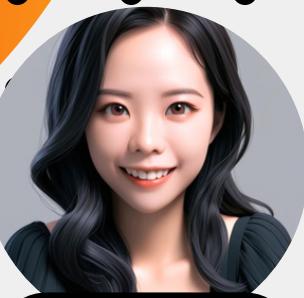
SHAP分析: Summary

後續方案建議將著重
影響力較大的欄位

```
1 feature_names = pipeline_80.named_steps['preprocess'].get_feature_names_out()  
2  
3 shap.summary_plot(  
4     shap_values,                      # 直接傳 Explanation 物件  
5     features=X_val_processed,  
6     feature_names=feature_names  
7 )  
8
```



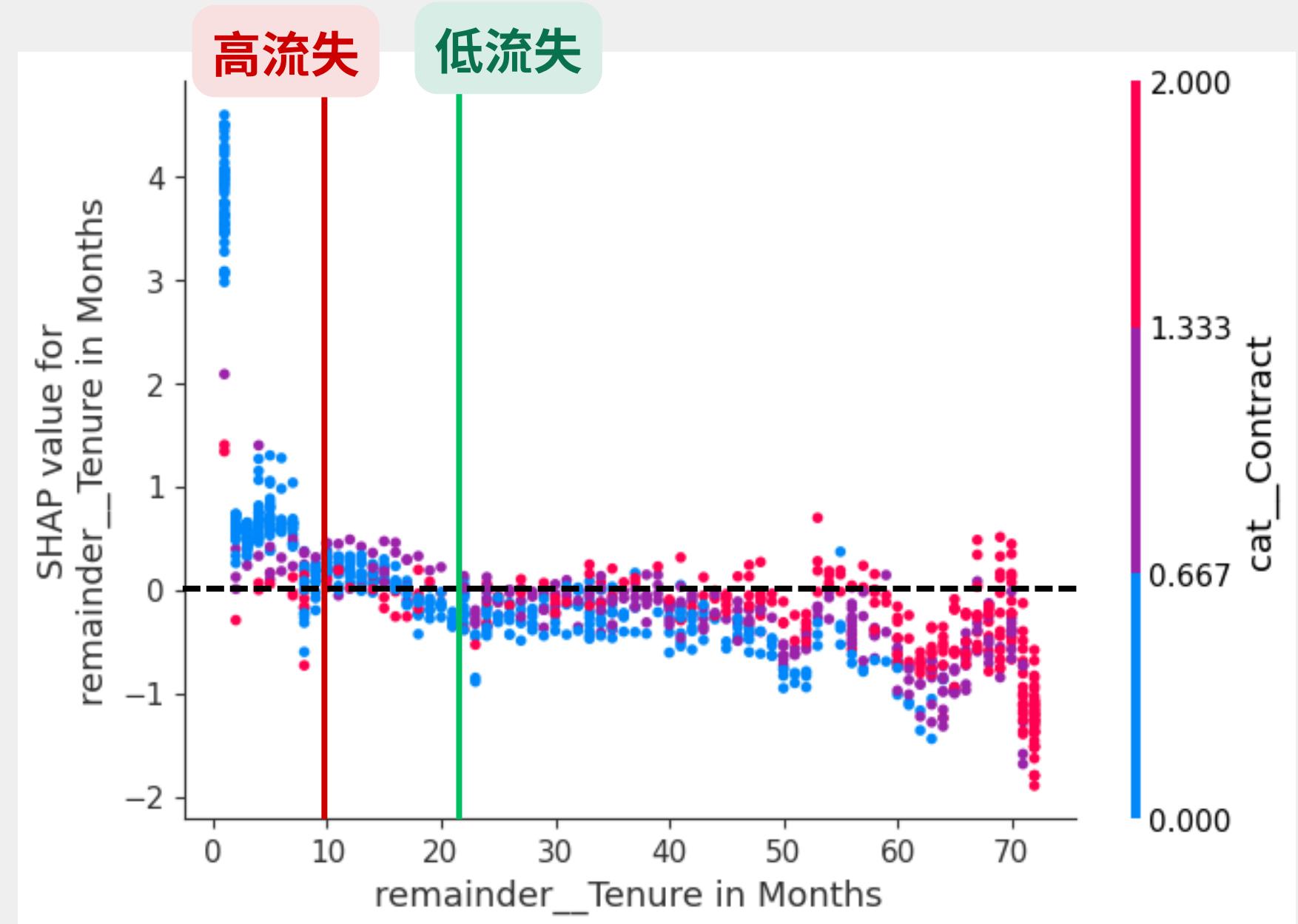
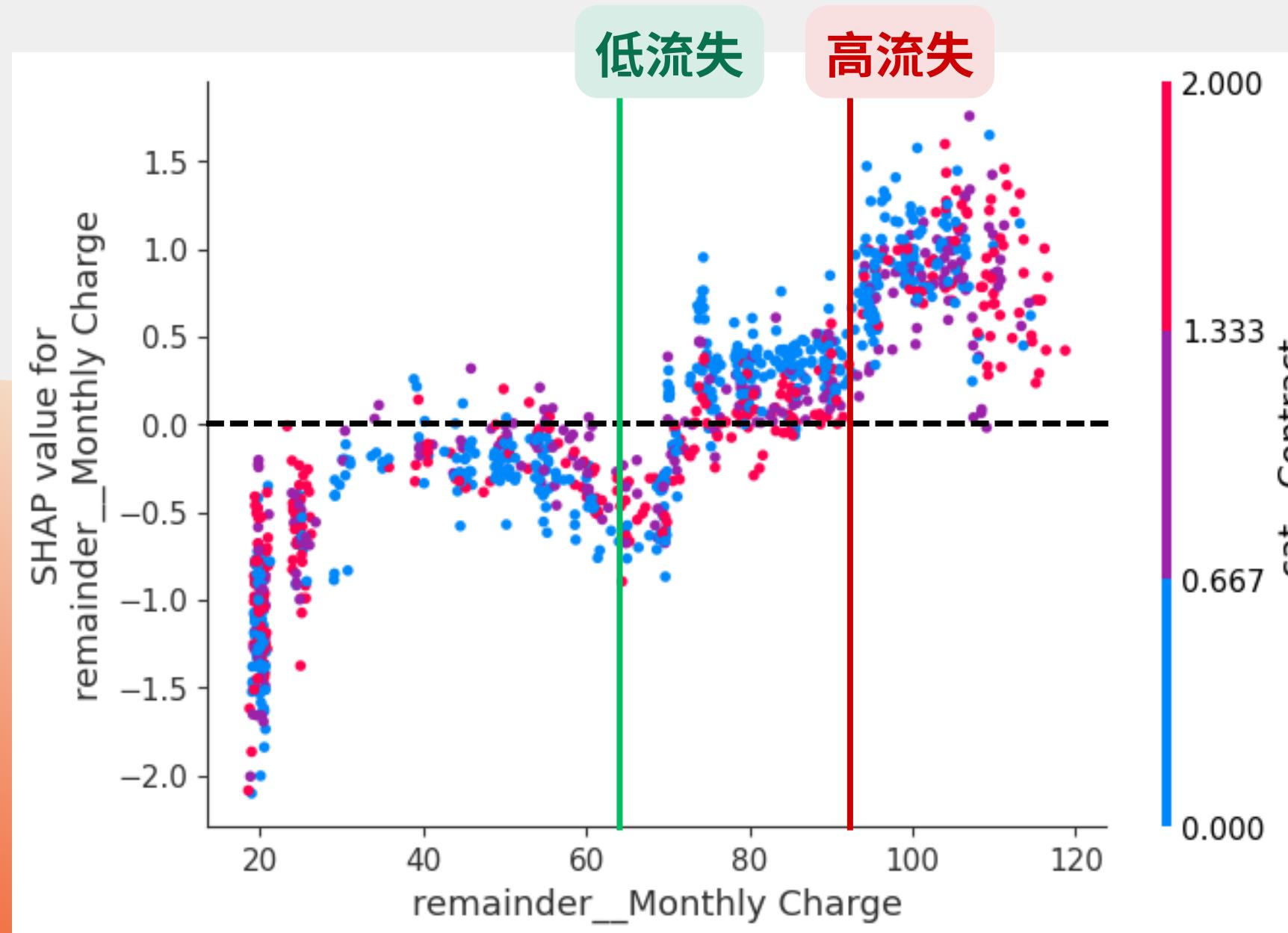
孫敏嘉

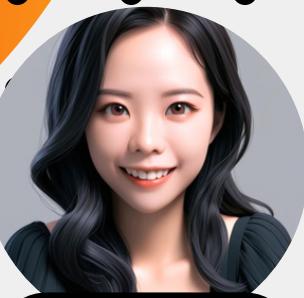


孫敏嘉

SHAP分析 - Dependence

利用圖表進一步觀察，了解特徵之間的交互作用以及**特徵數值**分布





孫敏嘉

流失風險判斷

風險等級	合約類型	推薦人數	扶養人口	累計合約期	月費
高	月租型	少	少	< 10個月	> \$100

主動推出對應方案，降低流失風險

低	年約型	多	多	> 24個月	< \$65
---	-----	---	---	--------	--------

行銷策略與方案



孫敏嘉

合約類型

推薦人數

扶養人口

累計合約期間

月費

目標: 延長用戶的合約期

[合約等級說明]

一等 | 月租制

二等 | 一年約

三等 | 兩年約

四等 | 三年以上合約

[優惠內容]

凡升級合約等級之用戶，依升級級數享月費折扣：

升級一級 → 月費 95 折

升級二級 → 月費 9 折

升級三級（即由月租升至三年以上）→ 成為終生會員，在服務內容不變的情況下，享永久固定資費續約

行銷策略與方案



孫敏嘉

合約類型

推薦人數

扶養人口

累計合約期間

月費

目標: 提升用戶推薦人數達 2 人以上

[適用對象]

所有現有用戶皆可參與推薦計畫

[獎勵內容]

每成功推薦 2 位新用戶，獲得一張 \$5 現金抵用券

每多推薦 1 人，額外加贈一張 (例如推薦 3 位 → 獲得 2 張，以此類推)

抵用券與推薦人數均無可上限累積

每人每月限使用 1 張抵用券

抵用券不得兌換現金



孫敏嘉

行銷策略與方案

合約類型

推薦人數

扶養人口

累計合約期間

月費

目標: 強化用戶黏著力、增強使用綁定性

[適用對象]

無登記扶養人口但有寵物，且屬高流失風險用戶

[方案內容]

申辦「寵物友善－加值流量方案」，可選擇：

- (a) 享優惠價購買寵物攝影機，或
- (b) 綁約 30 個月以上享免費租用攝影機

此方案可搭配原主方案使用，以情感連結為切入點，藉由觀看寵物功能延長用戶關係週期

行銷策略與方案



孫敏嘉

合約類型

推薦人數

扶養人口

累計合約期間

月費

目標: 鼓勵用戶累積合約期達12個月以上，以增加轉續年約機會

[適用對象]

累積使用合約未滿一年的月租用戶，可預繳補足升級福利

[優惠內容]

若合約累積使用時間未滿 12 個月，用戶可預繳剩餘月份，即刻享有年約福利。例如：已使用 8 個月，預繳 4 個月，即同等享「一年約」福利

福利包含但不限於：

專屬客服服務、加值服務折扣、限量活動邀約（如新品體驗、VIP日等）

行銷策略與方案



孫敏嘉

合約類型

推薦人數

扶養人口

累計合約期

月費

目標: 利用一次性預付金額換取長期折扣，以降低價格敏感用戶流失

[適用對象]

月費偏高、價格敏感、高流失風險之用戶

[優惠內容]

用戶預儲值優惠：一次性儲值 \$350 元(或以上)，即享月費 85 折優惠。

折扣自儲值當月起生效，持續至儲值金額扣抵完畢為止。

每位會員限參加一次，且儲值金額不可退費

此為適用任何用戶的大眾方案

可同時提高企業短期收益並延長用戶使用時間



孫敏嘉

預測結果與方案推薦機制

預測流失機率**50%**為分水嶺：

- 以下為低風險用戶，穩定客情維護
- 以上為高風險用戶，主動推薦對應留客方案

預測結果

顧客流失機率為：**35.45%**

穩定用戶，流失風險低

預測結果

顧客流失機率為：**73.39%**

高風險用戶，建議主動聯繫留客

預測結果與方案推薦機制



孫敏嘉

根據會「提升流失風險增加」的特徵欄位，推薦3項對應挽留方案。

以右方SHAP Decision Plot為例，將優先推薦：

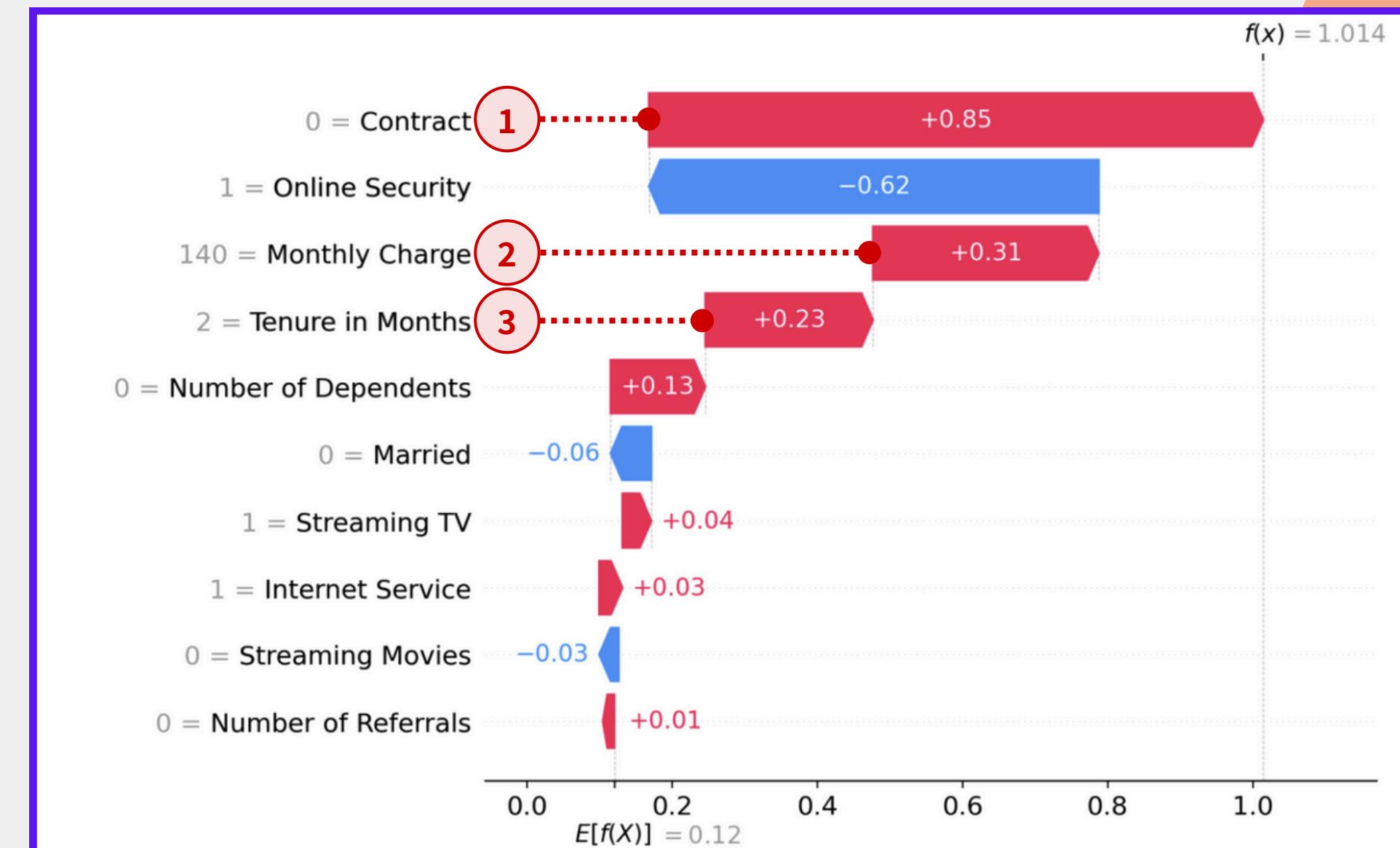
(1) 合約類型 → 延長合約優惠

(2) 月費 → 用戶預儲值優惠

(3) 累積合約期 → 預繳月費升級福利

顧客流失機率為：73.39%

⚠ 高風險用戶，建議主動聯繫留客



圖例說明：

● 紅色：提升流失風險 / ● 藍色：降低流失風險



林靖軒

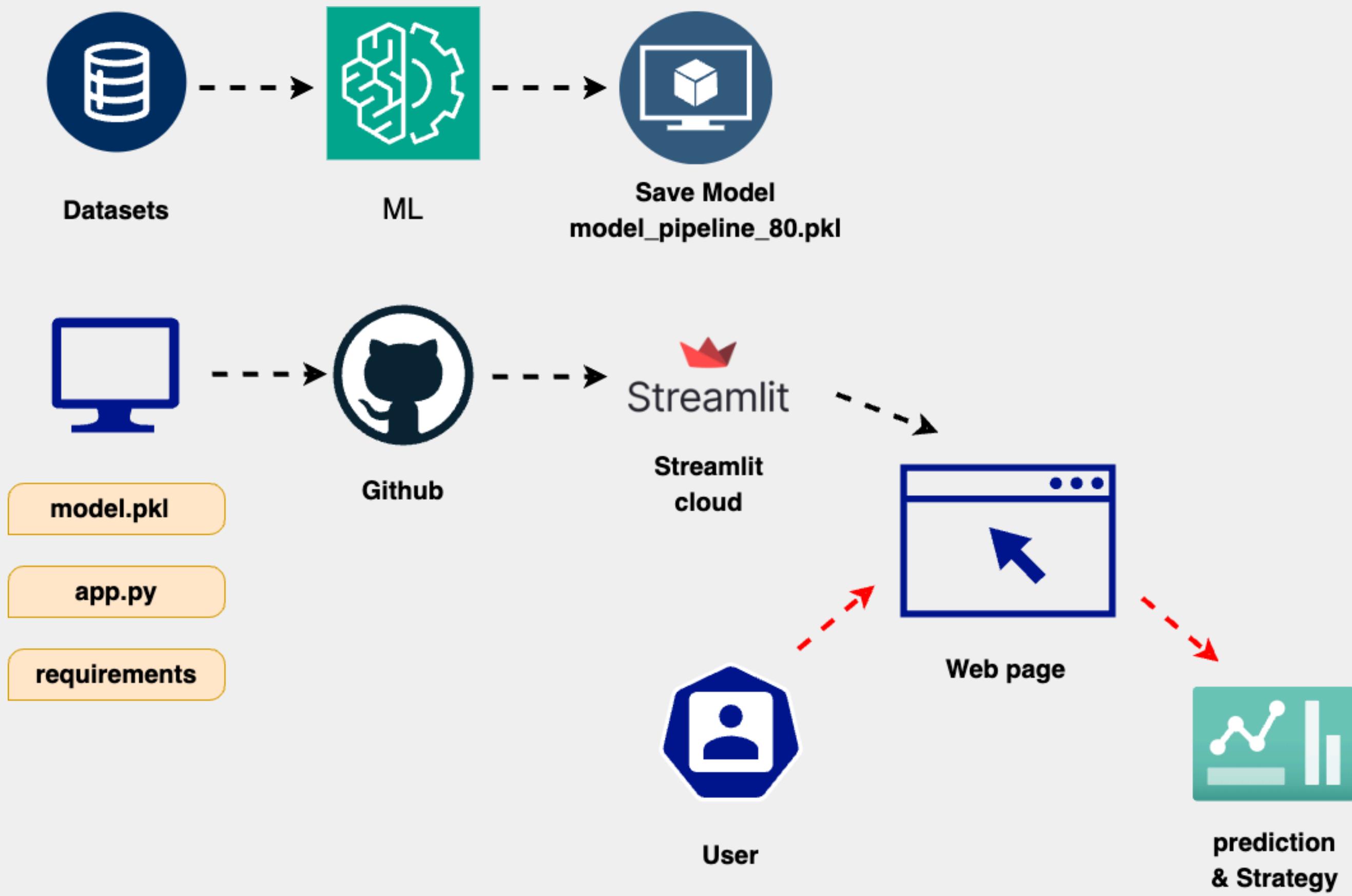
介面設計與部署

UI Design and Deployment

Deployment Workflow



林靖軒





林靖軒

撰寫 APP.PY

```
10 st.title("顧客流失預測器")
11
12 # — 載入模型 Pipeline —
13 with open('model_pipeline_80.pkl', 'rb') as f:
14     data = pickle.load(f)
15     pipeline = data['pipeline']
16     features = data['selected_features']

41 # — 使用者輸入 —
42 with st.expander("合約與帳務 ", expanded=True):
43     contract = st.radio("合約類型", ['Month-to-Month', 'One Year', 'Two Year'])
44     tenure = st.number_input("使用月數", 0, 100, 10)
45     monthly_charge = st.number_input("每月費用", 0.0, 500.0, 70.0)
46
47 with st.expander("服務項目 (有使用請勾選)", expanded=True):
48     col1, col2, col3, col4 = st.columns(4)
49     with col1:
50         internet = st.checkbox("上網服務")
51     with col2:
52         online_security = st.checkbox("網路安全服務")
53     with col3:
54         streaming_movies = st.checkbox("串流電影服務")
55     with col4:
56         streaming_tv = st.checkbox("串流電視服務")
57
58 with st.expander("個人狀況", expanded=True):
59     referrals = st.number_input("推薦人數", 0, 100, 1)
60     dependents = st.number_input("扶養人數", 0, 10, 0)
61     married = st.radio("是否已婚", ['Yes', 'No'])

62
```

- 使用 `pickle` 讀取已訓練並儲存的模型檔案
- 提供多種輸入元件讓使用者填寫客戶資訊



林靖軒

撰寫 APP.PY

```
111 # — 預測與顯示結果 —
112 if st.button("🔮 預測是否流失"):
113     # 將布林值轉成模型訓練用的字串
114     internet_str = "Yes" if internet else "No"
115     online_security_str = "Yes" if online_security else "No"
116     streaming_movies_str = "Yes" if streaming_movies else "No"
117     streaming_tv_str = "Yes" if streaming_tv else "No"
118
119     # 組成輸入 DataFrame
120     input_dict = {
121         'Contract': contract,
122         'Internet Service': internet_str,
123         'Number of Referrals': referrals,
124         'Number of Dependents': dependents,
125         'Married': married,
126         'Streaming Movies': streaming_movies_str,
127         'Streaming TV': streaming_tv_str,
128         'Tenure in Months': tenure,
129         'Online Security': online_security_str,
130         'Monthly Charge': monthly_charge
131     }
132     input_df = pd.DataFrame([input_dict])[features]
133
134     # 模型機率預測
135     prob = pipeline.predict_proba(input_df)[0][1]
136     st.subheader("📊 預測結果")
137     st.markdown(
138         f"<div style='font-size:24px; font-weight:bold;'>顧客流失機率為: {prob:.2%}</div>",
139         unsafe_allow_html=True
140     )
141
```

- 按下「預測是否流失」按鈕後，將輸入資料轉成 DataFrame。
- 使用載入的模型計算流失機率。



林靖軒

撰寫 APP.PY

```
if prob > 0.5:  
    st.warning("⚠ 高風險用戶，建議主動聯繫留客")  
  
feature_names = get_feature_names(pipeline.named_steps['preprocess'])  
X_trans = pipeline.named_steps['preprocess'].transform(input_df)  
X_trans_df = pd.DataFrame(X_trans, columns=feature_names)  
  
explainer = shap.Explainer(pipeline.named_steps['clf'])  
shap_values = explainer(X_trans_df)  
  
vals = shap_values.values[0]  
pos_vals = np.where(vals > 0, vals, -np.inf)  
top3_idx = np.argsort(pos_vals)[-3:][::-1]  
top3_feats = [feature_names[i] for i in top3_idx]  
  
st.subheader("💡 建議的留客策略")  
  
default_strategy = (  
    "<div style='font-size:18px; font-weight:bold;'>用戶預儲值優惠</div>"  
    " [適用對象] 月費偏高、價格敏感、高流失風險之用戶<br>"  
    " [優惠內容] <br>"  
    "一次性儲值 $350 元（或以上）即日起享 月費 85 折優惠。折扣自儲值當月起生效，持  
    "每位會員限參加一次，且儲值金額不可退費</div>"  
)  
  
for feat in top3_feats:  
    txt = strategy_map.get(feat, default_strategy)  
    st.markdown(txt, unsafe_allow_html=True)  
    st.markdown('<hr style="border:1px solid #ccc;">', unsafe_allow_html=True)  
  
fig, ax = plt.subplots()  
shap.plots.waterfall(shap_values[0], max_display=10, show=False)  
st.pyplot(fig)  
st.markdown("")  
    **圖例說明：**  
    ● 紅色：提升流失風險 / 藍色：降低流失風險  
    "")  
else:  
    st.success("✅ 穩定用戶，流失風險低")
```

- 判斷預測流失機率是否超過 50%
- 若是，則利用 SHAP 解釋模型找出影響流失風險的前三大特徵，並根據這些特徵顯示對應的留客策略與視覺化圖表，幫助理解用戶流失原因。



林靖軒

建立虛擬環境

目的：避免套件版本衝突，確保專案穩定運作



```
● venvritalin@RitadeMacBook-Air churn-app % source venv/bin/activate
○ (venv) venvritalin@RitadeMacBook-Air churn-app % █
```



林靖軒

建立requirements.txt

```
numpy           2.3.1
packaging      25.0
pandas          2.3.0
pillow          11.2.1
pip             24.3.1
protobuf        6.31.1
pyarrow         20.0.0
pydeck          0.9.1
python-dateutil 2.9.0.post0
pytz            2025.2
referencing     0.36.2
requests         2.32.4
rpds-py          0.25.1
scikit-learn    1.7.0
scipy            1.16.0
six              1.17.0
smmap            5.0.2
streamlit        1.46.0
tenacity         9.1.2
threadpoolctl   3.6.0
toml             0.10.2
tornado          6.5.1
typing_extensions 4.14.0
tzdata           2025.2
urllib3          2.5.0
xgboost          3.0.2
```

- 使用 `pip list` 查看目前虛擬環境已安裝的套件及版本

```
☰ requirements.txt × app.py
☰ requirements.txt
1 matplotlib==3.10.3
2 numpy==2.2.6
3 pandas==2.3.0
4 shap==0.48.0
5 streamlit==1.46.0
6 shap==0.48.0
7 xgboost==3.0.2
```

- 將套件資訊輸出並寫入 `requirements.txt`, 方便環境重現



林靖軒

上傳至 GitHub

Streamlit Cloud 主要是與 GitHub 連動

The screenshot shows a GitHub repository page for a public repository named "Telecom-Customer-Churn-Prediction-App". The repository has one branch ("main") and no tags. The interface includes a search bar labeled "Go to file". Below the repository name, there is a note from a user named "ritatalin" suggesting to "Add files via upload". The repository contains four files: ".gitignore", "app.py", "model_pipeline_80.pkl", and "requirements.txt". Each file has an "Add files via upload" button next to it.

File	Action
.gitignore	Create .gitignore
app.py	Add files via upload
model_pipeline_80.pkl	Add files via upload
requirements.txt	Add files via upload



林靖軒

部署到 Streamlit Cloud

Deploy a public app from GitHub

My code is ready on a GitHub repo, and it is totally awesome.

[Deploy now](#)

Deploy an app

Repository [Paste GitHub URL](#)
ritatalin/Telecom-Customer-Churn-Prediction-App

Branch
main

Main file path
app.py

App URL (optional)
telecom-customer-churn-prediction-app-cptrmtlp4386mxiewywnm .streamlit.app

Domain is available

✓ 快速且便利完成部署

DEMO

最終介面呈現

顧客流失預測器

合約與帳務

合約類型

Month-to-Month

One Year

Two Year

使用月數

10

每月費用

70.00

服務項目 (有使用請勾選)

上網服務 網路安全服務 串流電影服務 串流電視服務

個人狀況

推薦人數

1

扶養人數

林靖軒



SCAN ME



林靖軒

問題與解決方案

Problem and solution



林靖軒

問題與解方 - 資料前處理

Customer Status 欄位歸類標準不同

→ 刪除 Joined 且 Contract 為 Month-to-month 且 Tenure = 1 的樣本，
其餘視為 Stayed。

面對大量欄位，應該先在介面篩選還是讓模型決定重要特徵？

→ 欄位多且影響介面設計，團隊曾考慮先剔除欄位，但擔心影響模型表現。
最後依老師建議，先全部放進模型選出重要特徵後，再思考介面的包裝呈現。

條件缺失欄位的「No Internet」值處理

→ 機器學習模型本身無法主動理解欄位間的業務邏輯關係，
若欄位包含過多類別或特殊狀態（如 No Internet），可能造成編碼複雜且影響模型學習穩
定性，因此將其統一為 No，有助於簡化欄位結構與提升模型表現。



林靖軒

問題與解方 - 模型訓練

調參後的模型表現相差不大，難以選出最終模型

→ 當以 Recall 最大進行模型優化時，組員們的模型成績都很接近
最終以 $\text{threshold} = 0.5$ 的預設設定作為比較基準，
在這個基礎下觀察各模型的整體表現差異，並選出最終模型。

原資料集流失樣本比例不均

→ 整體樣本數，非流失與流失比例大約是 7:3，
以調整 `scale_pos_weight` 參數的方式，解決樣本數不平衡的問題。。



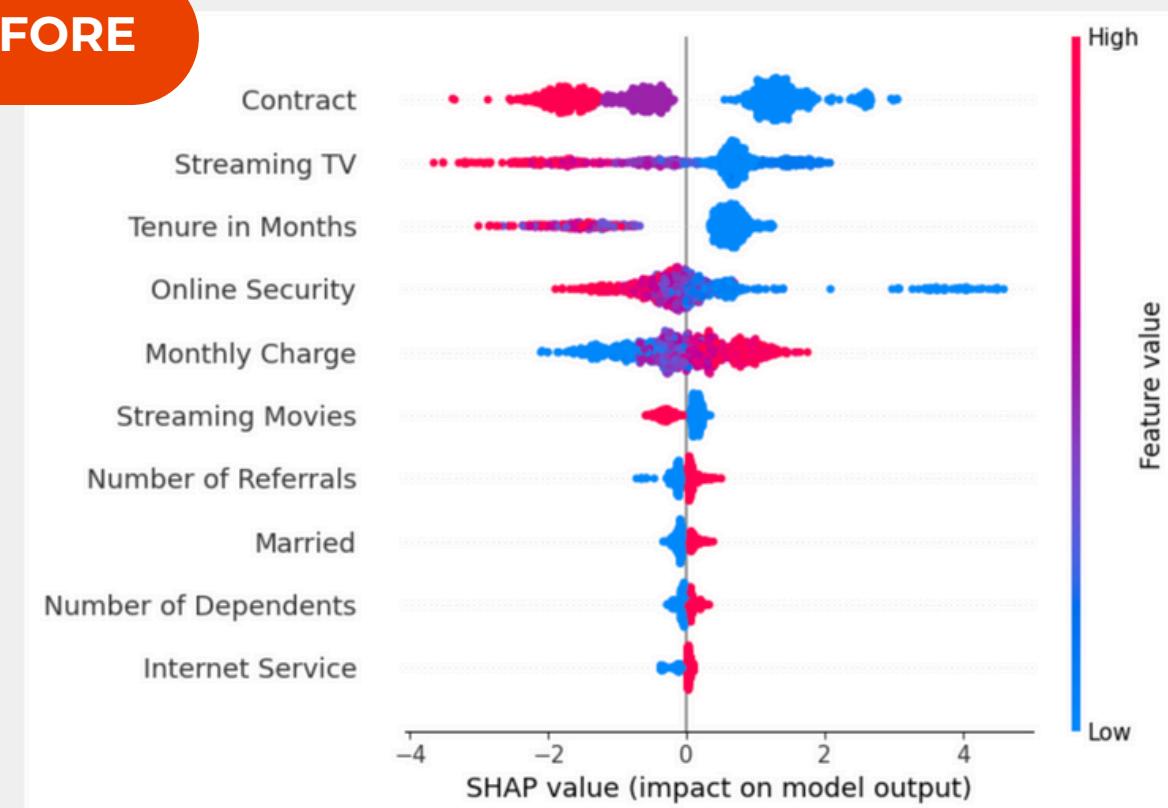
林靖軒

問題與解方 - SHAP 洞察

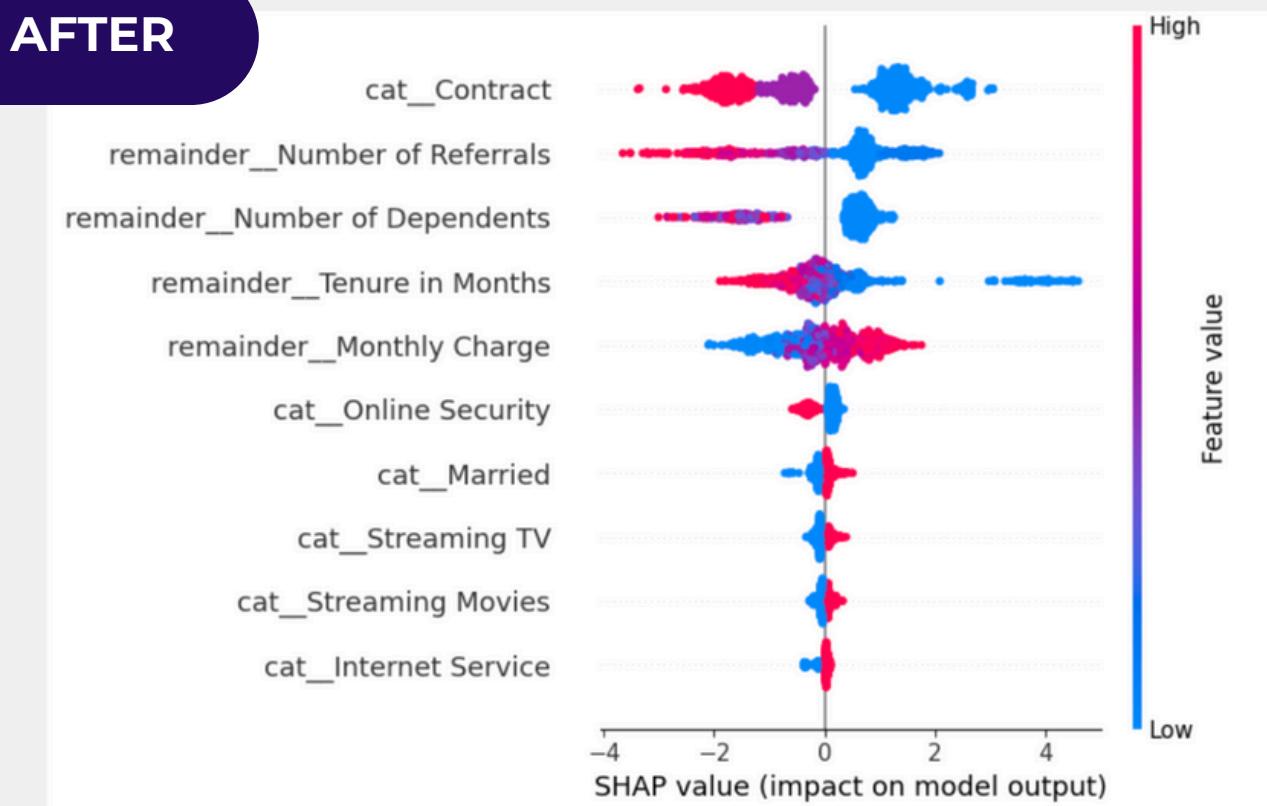
SHAP summary 與 dependence 圖表的解釋不一致

→ 經重新檢視後發現，summary 圖表中的數據圖所使用的特徵欄位有誤，導致數據圖呈現的特徵順序和Feature value與 dependence 不同，進而引起解讀上的誤差。
修正變數後，summary 與 dependence 圖表已能一致呈現正確的特徵影響關係。

BEFORE



AFTER



FEATURE_NAMES=
PIPELINE_80.NAMED_STEPS['PREPROCESS'].
TRANSFORM(X_VAL_SEL)

FEATURE_NAMES=
PIPELINE_80.NAMED_STEPS['PREPROCESS'].
GET_FEATURE_NAMES_OUT()



林靖軒

問題與解方 - 介面設計與部署

Colab 執行 Streamlit 複雜不穩定

→ Colab 需搭配 ngrok 等工具做端口轉發，
後改為本地 VSCode 環境開發，提升穩定性與使用便利。

模型預測中 Contract 欄位被編成 -1，對預測結果也沒影響

→ 因介面輸入的 Contract 字串大小寫和訓練時不符，Encoder 把它當作未知類別編成 -1。
需留意輸入時大小寫與訓練類別一致，才能正確反映該欄位對預測和 SHAP 的影響。

合約與帳務 Contract & Billing

Contract 合約類型

- Month-to-month
- One year
- Two year

BEFORE

合約與帳務 Contract & Billing

Contract 合約類型

- Month-to-Month
- One Year
- Two Year

AFTER



林靖軒

結論

Conclusion

結論



林靖軒

專案淬鍊與成長

核心學習與能力提升

資料洞察力

精準處理複雜數據（如「No Internet」值），強化數據品質控管。

模型優化實戰

累積不平衡資料處理與調參經驗， 聚焦提升召回率。

價值傳遞力

運用SHAP解釋模型，轉化技術為商業洞察， 成功部署可視化介面。

- A 5x5 grid of black dots, arranged in five rows and five columns, centered on a light gray background.

驅動未來商業價值

未來實務擴展潛力

持續迭代優化

定期追蹤關鍵指標，利用最新客戶數據更新並再訓練模型，提升預測準確度與適應性。

策略成效驗證

行銷端可透過A/B測試評估策略成效，並拓展至服務升級、產品偏好等行為預測，支持更全面的數據驅動決策。

深化客戶洞察

預測多元客戶行為，挖掘深入數據洞察，
助力企業制定完整且系統化的數據驅動策略，
提升經營效益與競爭力。

End

Thank you

Do you have any questions?

