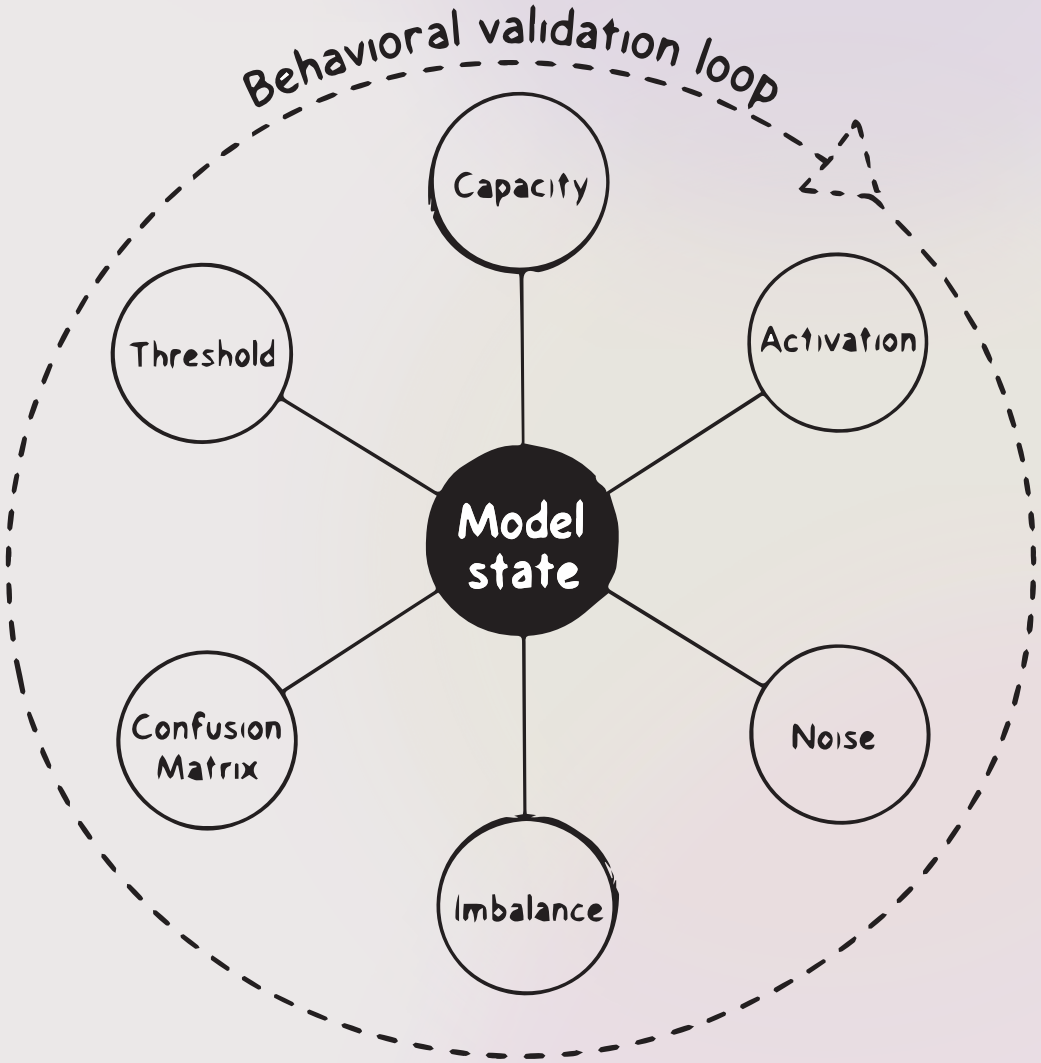
Training Behavior of a Titanic Survival Model



Validation loop

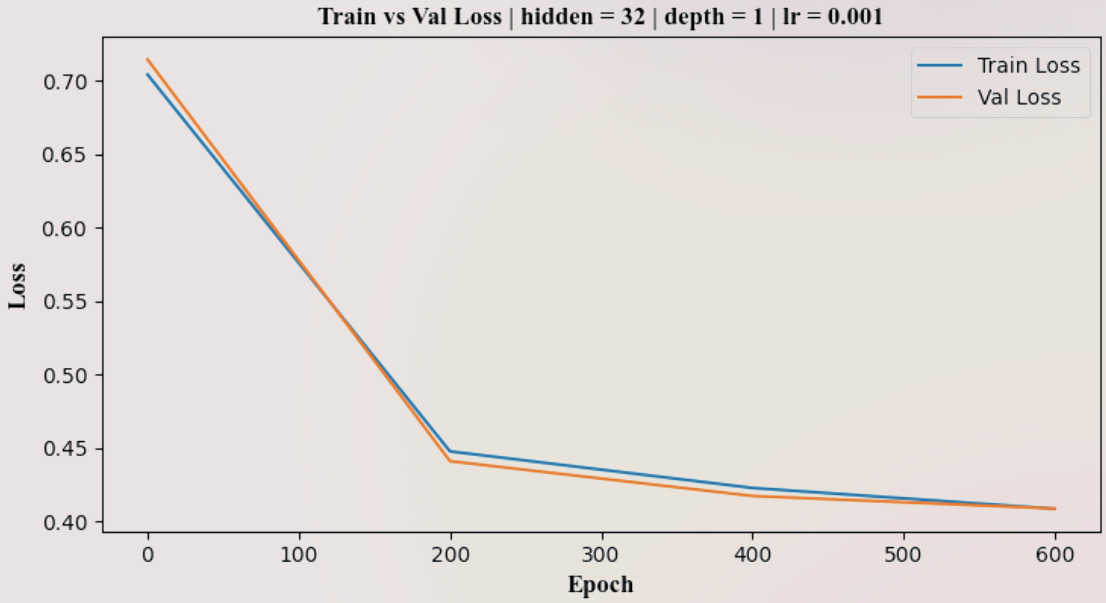
Global settings

SEED = 33
HIDDEN = 32
EPOCHS = 600
TEST_SIZE = 0.2

LR = 1e-3
DEPTH = 1
WEIGHT_DECAY = 1e-4
THRESHOLDS = (0.3, 0.5, 0.8)

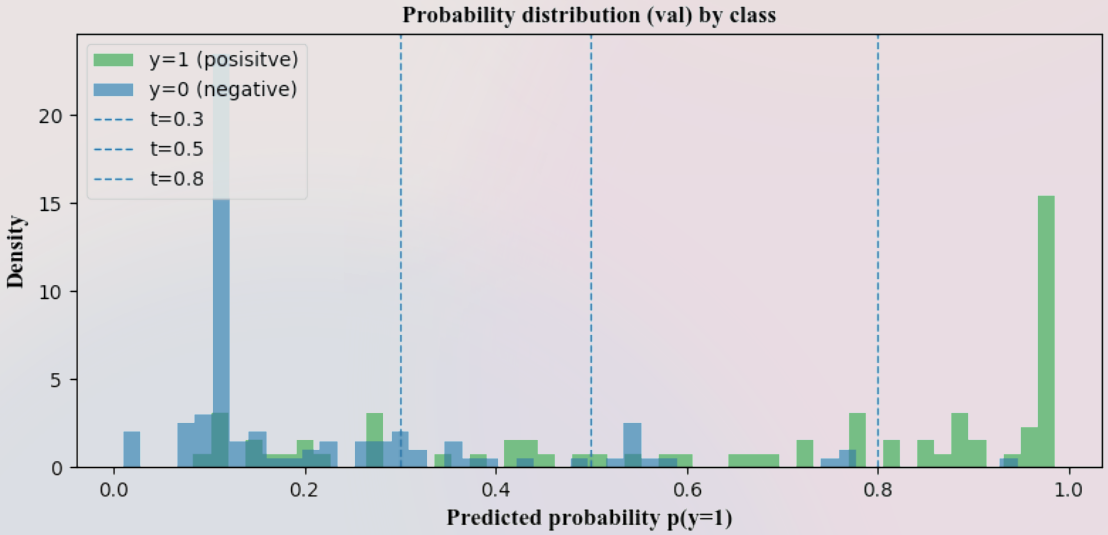
找合適的capacity與depth與epoch，透過probability檢查loss與activation的一致性，檢查logit分佈狀態是否屬於當前sigmoid合適區間段(確保saturation的狀況不會發生)。

Capacity & Depth sweep



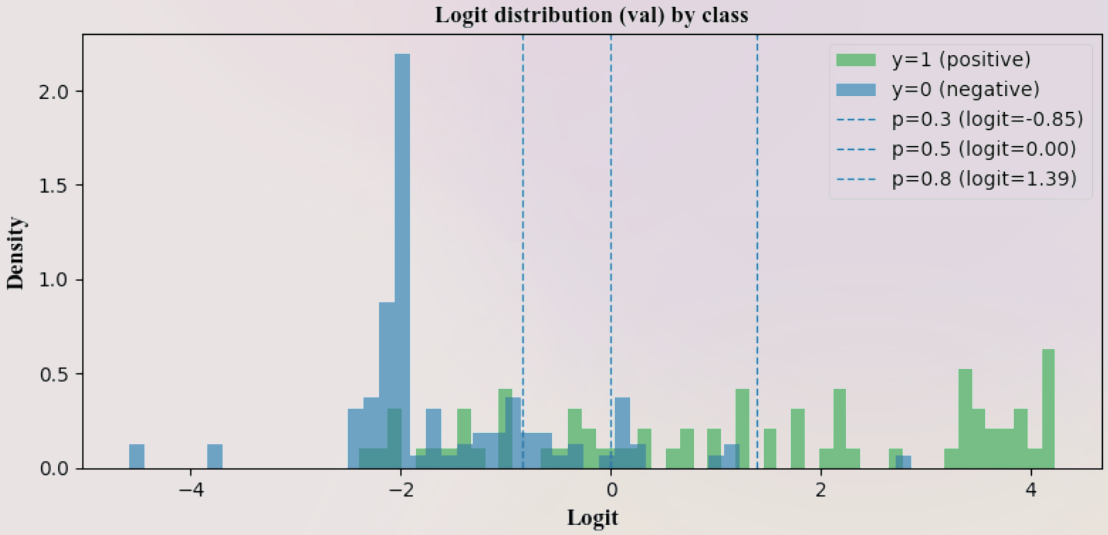
從目前的train loss與validation loss比較，目前的capacity與depth組合呈現收斂狀態。

Probability



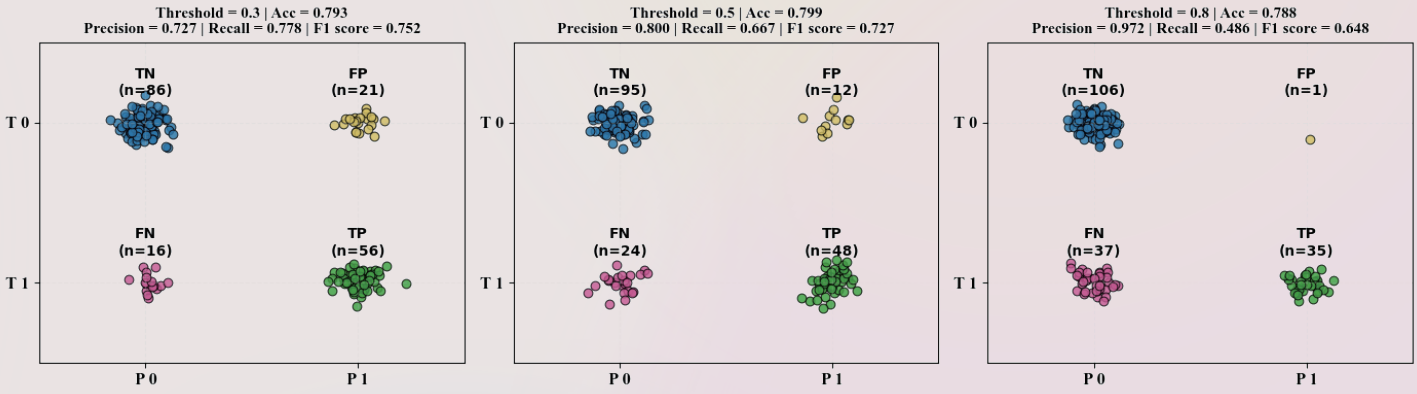
從機率分佈可以觀察，除了明確的positive與negative區域之外，仍存在一段類別重疊的灰色區間（代表目前現有特徵對類別分類能力有限）。透過預先設定的多個threshold可以進一步觀察模型在不同決策門檻下，對於正負類別的信心分級與風險取捨方式。

Logit



同樣從logit分佈可以觀察到，當threshold = 0.3時，分類決策規則改變，使模型傾向於將更多樣本判為positive。且目前有梯度平緩飽和區風險存在

Confusion matrix



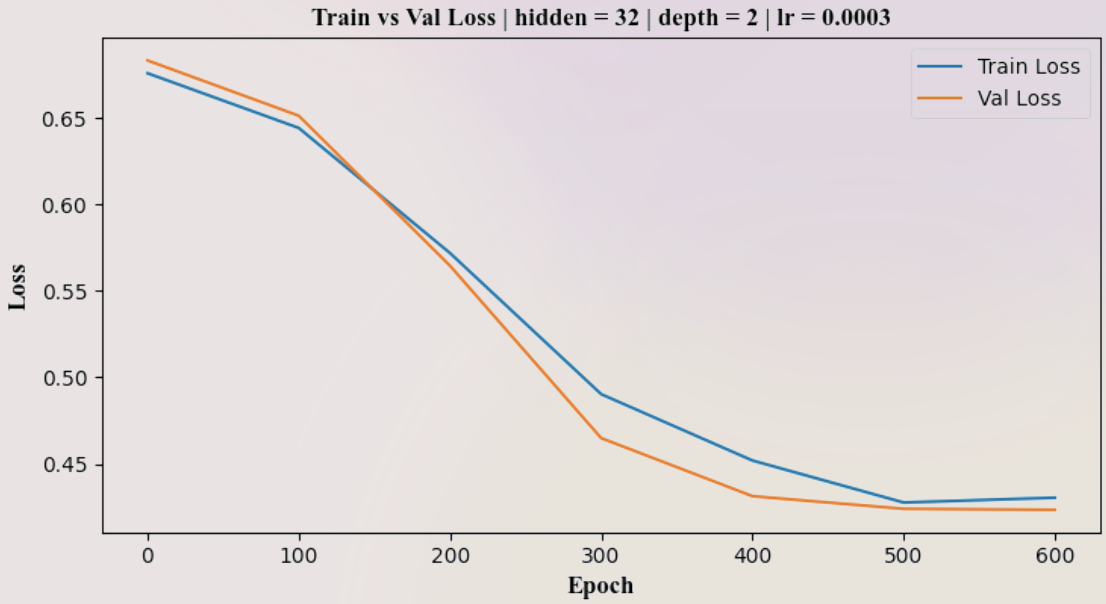
進一步透過confusion matrix得到驗證，在threshold = 0.3的情況下，False negative的數量明顯少於其他決策門檻，顯示模型在此設定下更積極嘗試抓取positive樣本。

然而整體結果仍未達到理想的風險取捨狀態，代表目前特徵組合下，模型在正負類別的可分性仍受限制。因此下一步將嘗試調整feature的組合，以改善logit分佈的分離程度，並重新評估不同threshold下的決策行為。

Global settings(updated)

LR = 3e-4
DEPTH = 2
DROPOUT = 0.3

Feature



透過interaction feature加強關聯性欄位有助於模型提升非線性學習關聯，目前在此條件設定下有良好的學習結構曲線(顯示模型capacity被更有效的利用，而非單純增加複雜度）

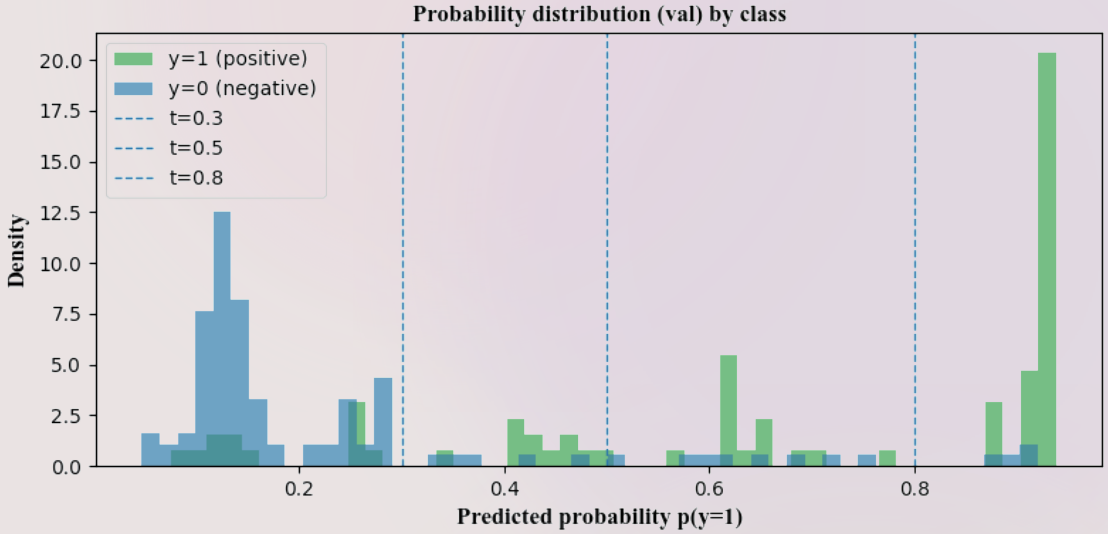
Feature engineering

```
# sex + pclass----------------------------------------------------------------
X_train['Sex_Pclass'] = X_train['Sex'] + '_P' + X_train['Pclass'].astype(str)
X_val['Sex_Pclass'] = X_val['Sex'] + '_P' + X_val['Pclass'].astype(str)
# --Ticket-------------------------------------------------------------------------------
ticket_counts = X_train['Ticket'].value_counts()
X_train['TicketGroupSize'] = X_train['Ticket'].map(ticket_counts).fillna(1).astype(int)
X_val['TicketGroupSize']   = X_val['Ticket'].map(ticket_counts).fillna(1).astype(int)
# ticket group size-----------------------------------------------------------------
bins = [0, 2, 4, float('inf')]
labels = ['solo', 'small', 'medium']
X_train['IsGroupTicket'] = pd.cut(X_train['TicketGroupSize'], bins=bins, labels=labels, include_lowest=True)
X_val['IsGroupTicket'] = pd.cut(X_val['TicketGroupSize'], bins=bins, labels=labels, include_lowest=True)
# fare and group ticket compare-----------------------------------------------------------+
fare_median = X_train['Fare'].median()
X_train['Fare'] = X_train['Fare'].fillna(fare_median)
X_val['Fare'] = X_val['Fare'].fillna(fare_median)
q = np.quantile(X_train['Fare'], [0, 0.25, 0.5, 0.75, 1.0])
q = np.unique(q)
labels = [f"Q{i}" for i in range(len(q)-1)]
X_train['FareBin'] = pd.cut(X_train['Fare'], bins=q, labels=labels, include_lowest=True)
X_val['FareBin']   = pd.cut(X_val['Fare'],   bins=q, labels=labels, include_lowest=True)
# -- compose-----------------------------------------------------------------------
X_train['GroupTicket_Fare'] = 'G' + X_train['IsGroupTicket'].astype(str) + '_' + X_train['FareBin'].astype(str)
X_val['GroupTicket_Fare']   = 'G' + X_val['IsGroupTicket'].astype(str)   + '_' + X_val['FareBin'].astype(str)
# -- input-----------------------------------------------------
num_columns = ['Age']
cat_columns = ['Sex_Pclass', 'GroupTicket_Fare']
```

Feature engineering設計思路

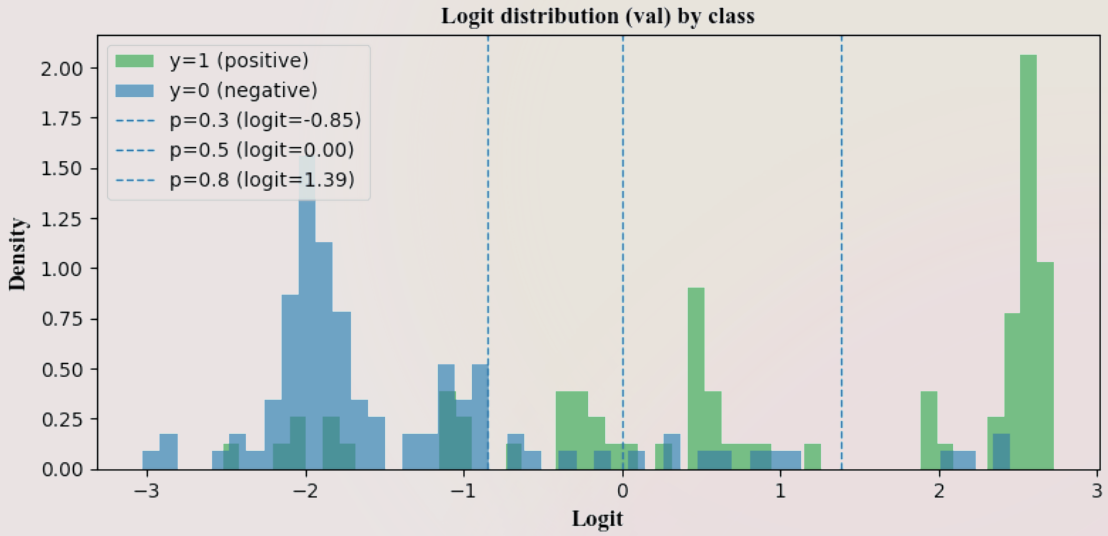
將Sex與Pclass交互，由Fare與Ticket團體票價交互，並將團體總類分為單人、2~4小團、5人以上大團。

Feature × Probability



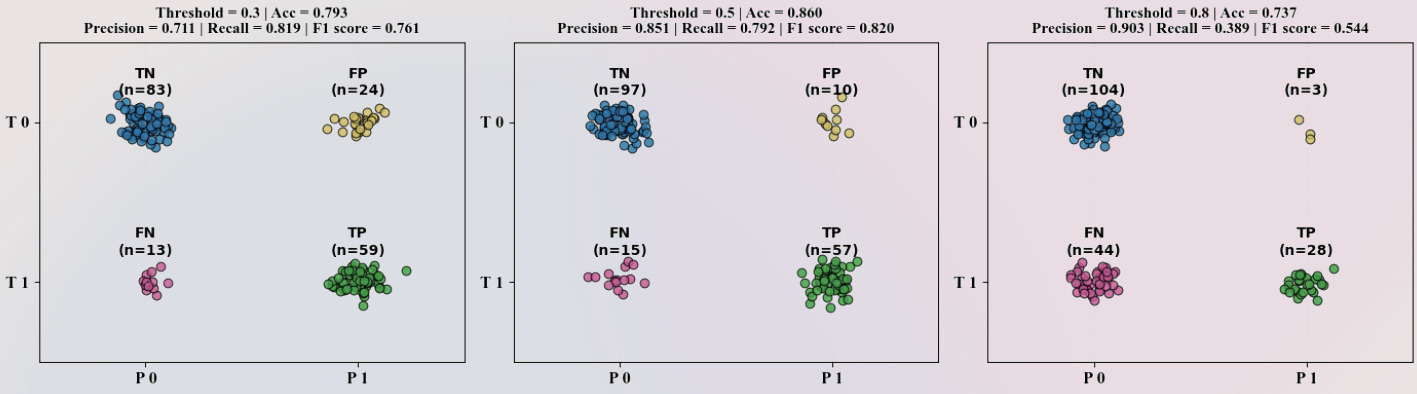
Probability density的集中顯示模型在引入語意特徵後，能將原本分散的樣本投影到更一致的決策區域，反映出關聯結構被更有效的抓取。

Feature × Logit



透過logit空間可以觀察到positive與negative呈現更穩定的可分趨勢。

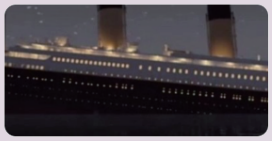
Feature × Confusion matrix



confusion matrix有著明顯的改變，模型針對誤判positive與negative有著顯著的下降，雖然保守但是結果較為穩定。

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

All Successful Errors Recent

Submission and Description	Public Score
submission.csv Complete · now	0.77511

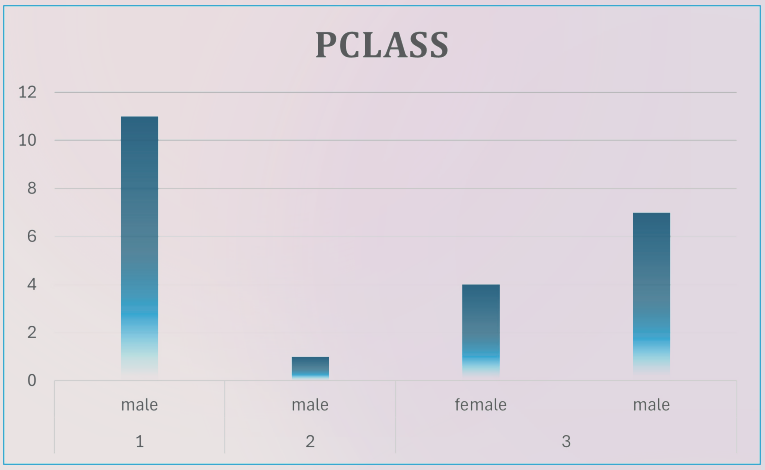
Model architecture

```
class Titanic(nn.Module):
    def __init__(self, input_dim, hidden=16, depth=2, dropout=0.3):
        super().__init__()
        self.depth = depth
        self.dropout = nn.Dropout(dropout)
        self.fc1 = nn.Linear(input_dim, hidden)

        if depth == 2:
            self.fc2 = nn.Linear(hidden, hidden // 2)
            self.fc3 = nn.Linear(hidden // 2, 1)
        else:
            self.fc2 = nn.Linear(hidden, 1)

    def forward(self, x):
        x = F.relu(self.fc1(x))
        x = self.dropout(x)
        if self.depth == 2:
            x = F.relu(self.fc2(x))
            x = self.dropout(x)
            logits = self.fc3(x)
        else:
            logits = self.fc2(x)
        return logits
```

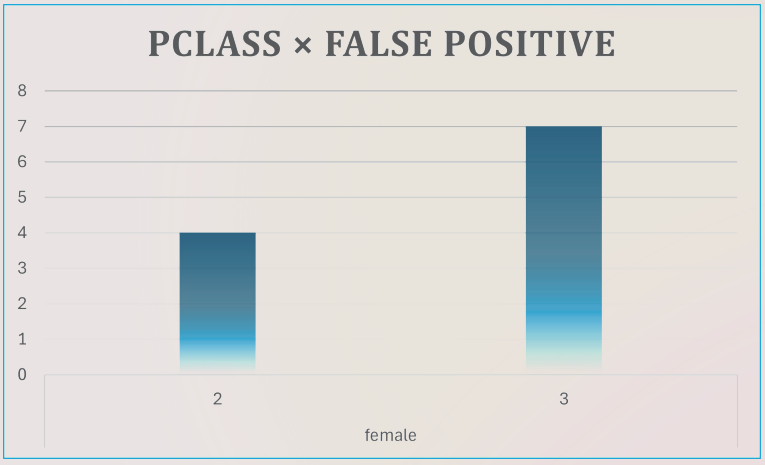
從前置feature engineering的重要性包含到interaction feature的設置與test當中最後使用ensemble測試seed的穩定性。資料結構本身可詮釋上限，以及model architecture選擇的重要。

結果:或許?還有尚未找齊的特徵未納入，嘗試把FN與FP輸出尋找關聯性。

False negative

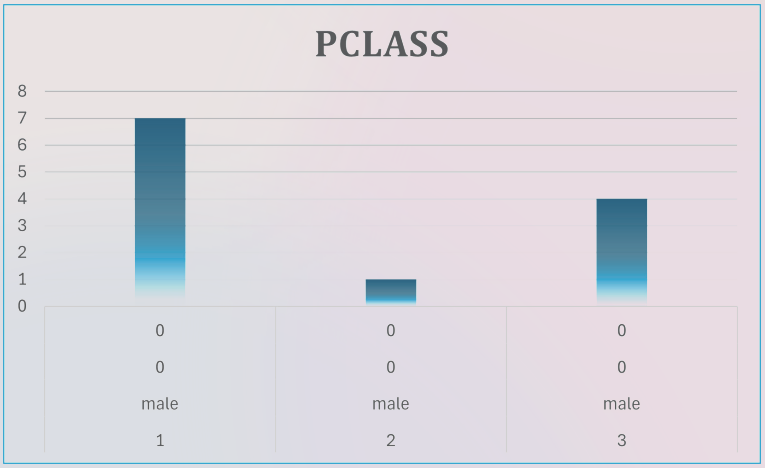


樞紐圖False negative的大部分是男性，且pcalss = 1的佔比非常多，根據我們設定的模型世界觀，資料顯示，也許當中並未與其他人發生組合抑或是Sex特徵過於強，導致模型無法模擬其他交互關係。



檢查False positive觀察，確實Sex特徵具有強烈暗示模型行為。

SibSp × Parch (FN)



篩選出SibSp與Parch的為0的欄位統計，驗證結果，在目前的特徵設計下模型傾向將solo類別視為缺乏可分的訊號樣本，導致預測信心降低（尤其是針對Sex主要特徵）。

Conditional modeling

[Female BEST RESULT]

Threshold : 0.50
Accuracy : 0.873
Precision : 0.882
Recall   : 0.957
F1-score : 0.918

TN: 10 | FP: 6
FN: 2 | TP: 45

[Male BEST RESULT]

Threshold : 0.15
Accuracy : 0.750
Precision : 0.410
Recall   : 0.727
F1-score : 0.525

TN: 71 | FP: 23
FN: 6 | TP: 16

Feature

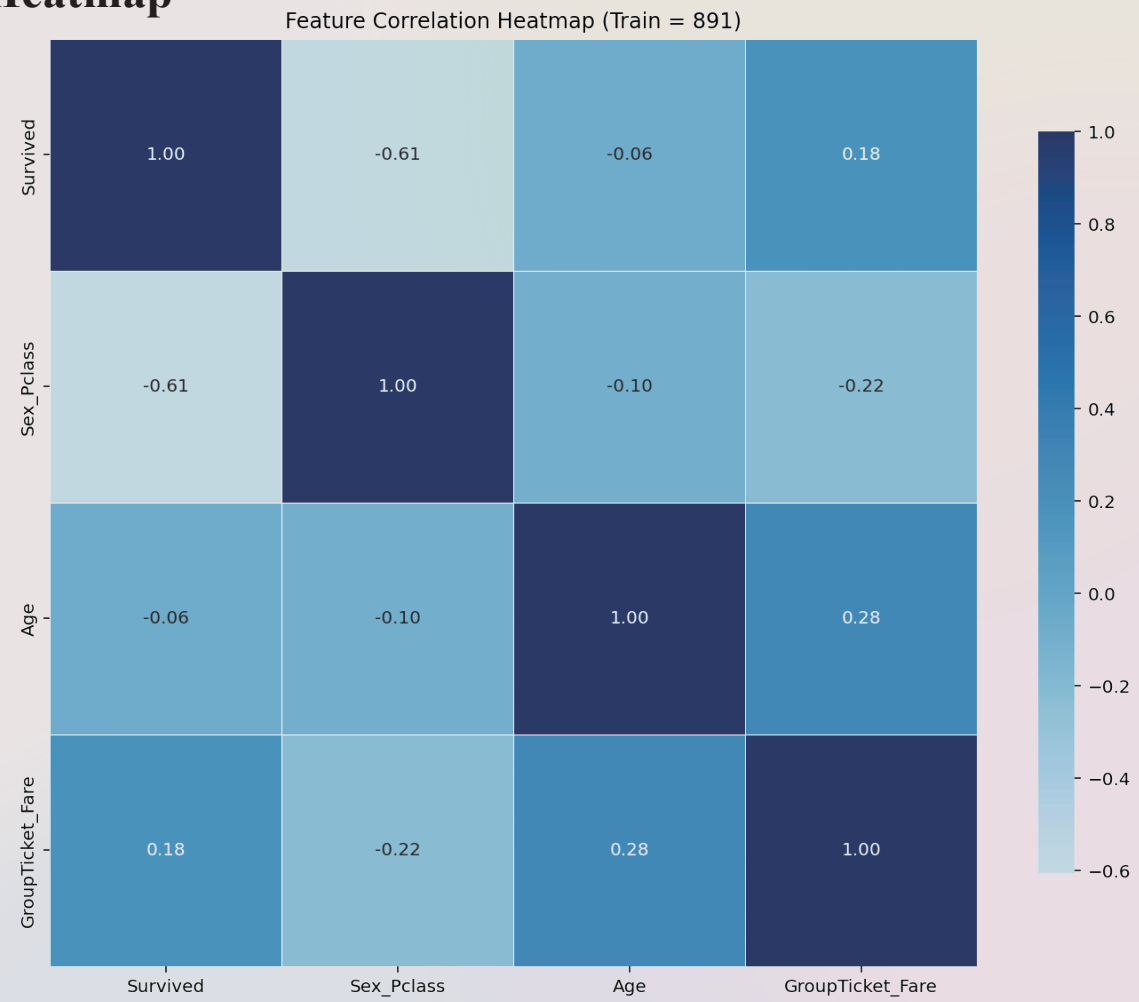
num_columns = ['Age', 'SibSp', 'Parch', 'Fare']

cat_columns = ['Pclass', 'Embarked']

本次表格採用最基本的feature
進行Sex個別測試。

嘗試引入條件訓練，將Sex進行分流檢驗結果，該結果顯示模型針對male判斷有些激進需要依賴調低threshold才能增加recall從目前基礎特徵發現尚不足以達成穩定true positive與negative的判斷，可能需要更具體條件的特徵交互補強。

Heatmap



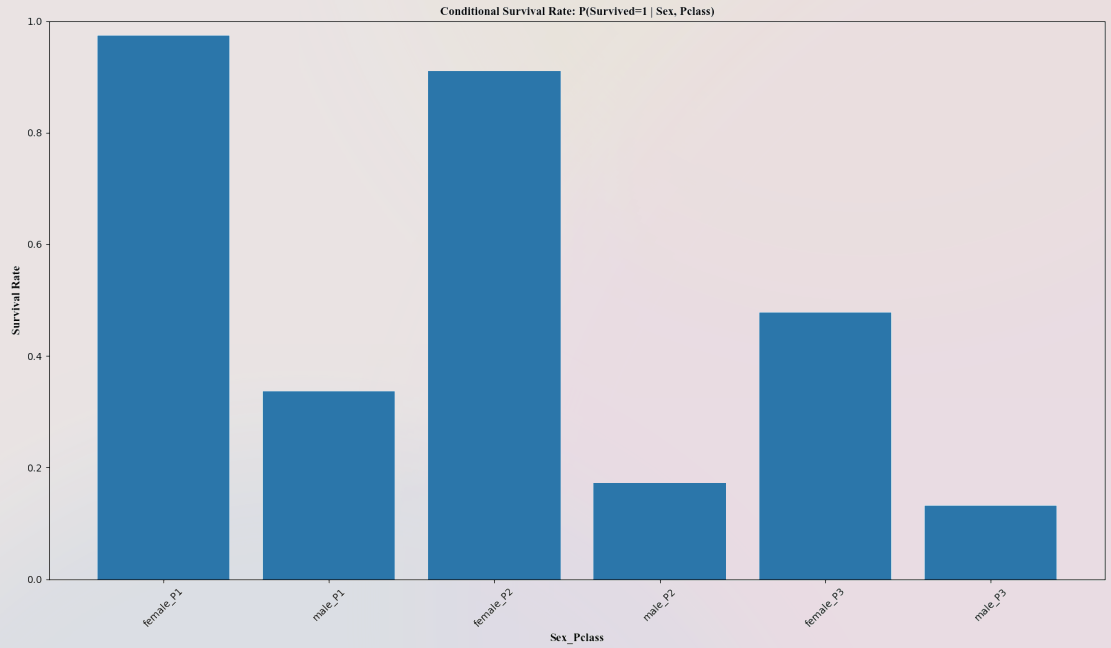
透過heatmap查看feature engineering結構是否存在相互矛盾，目前結果顯示與之設定條件吻合。
主要特徵sex_pclass，輔助特徵groupticket_fare, 次要特徵age。

Cross validation

```
CV ONLY | K=5 | HIDDEN=32 | DEPTH=2 | LR=0.0003 | DROPOUT=0.3

================ Fold 1/5 ================
[VAL @ threshold=0.30] | ACC = 0.7821 | Precision= 0.6923 | Recall= 0.7826 | F1_balance= 0.7347 | TP=54  FP=24  FN=15  TN=86
[VAL @ threshold=0.50] | ACC = 0.8045 | Precision= 0.8036 | Recall= 0.6522 | F1_balance= 0.7200 | TP=45  FP=11  FN=24  TN=99
[VAL @ threshold=0.80] | ACC = 0.7709 | Precision= 0.9667 | Recall= 0.4203 | F1_balance= 0.5859 | TP=29  FP=1  FN=40  TN=109

================ Fold 2/5 ================
[VAL @ threshold=0.30] | ACC = 0.8034 | Precision= 0.7143 | Recall= 0.8088 | F1_balance= 0.7586 | TP=55  FP=22  FN=13  TN=88
[VAL @ threshold=0.50] | ACC = 0.8034 | Precision= 0.7619 | Recall= 0.7059 | F1_balance= 0.7328 | TP=48  FP=15  FN=20  TN=95
[VAL @ threshold=0.80] | ACC = 0.7697 | Precision= 0.9091 | Recall= 0.4412 | F1_balance= 0.5941 | TP=30  FP=3  FN=38  TN=107

================ Fold 3/5 ================
[VAL @ threshold=0.30] | ACC = 0.7584 | Precision= 0.6374 | Recall= 0.8529 | F1_balance= 0.7296 | TP=58  FP=33  FN=10  TN=77
[VAL @ threshold=0.50] | ACC = 0.8483 | Precision= 0.8475 | Recall= 0.7353 | F1_balance= 0.7874 | TP=50  FP=9  FN=18  TN=101
[VAL @ threshold=0.80] | ACC = 0.8202 | Precision= 1.0000 | Recall= 0.5294 | F1_balance= 0.6923 | TP=36  FP=0  FN=32  TN=110

================ Fold 4/5 ================
[VAL @ threshold=0.30] | ACC = 0.7697 | Precision= 0.6709 | Recall= 0.7794 | F1_balance= 0.7211 | TP=53  FP=26  FN=15  TN=84
[VAL @ threshold=0.50] | ACC = 0.7865 | Precision= 0.7778 | Recall= 0.6176 | F1_balance= 0.6885 | TP=42  FP=12  FN=26  TN=98
[VAL @ threshold=0.80] | ACC = 0.6180 | Precision= 0.0000 | Recall= 0.0000 | F1_balance= 0.0000 | TP=0  FP=0  FN=68  TN=110

================ Fold 5/5 ================
[VAL @ threshold=0.30] | ACC = 0.7697 | Precision= 0.6458 | Recall= 0.8986 | F1_balance= 0.7515 | TP=62  FP=34  FN=7  TN=75
[VAL @ threshold=0.50] | ACC = 0.8708 | Precision= 0.8286 | Recall= 0.8406 | F1_balance= 0.8345 | TP=58  FP=12  FN=11  TN=97
[VAL @ threshold=0.80] | ACC = 0.8034 | Precision= 0.9250 | Recall= 0.5362 | F1_balance= 0.6789 | TP=37  FP=3  FN=32  TN=106

================ CV Summary (mean ± std) ================

[t=0.30] ACC 0.7766±0.0171 | P 0.6721±0.0320 | R 0.8245±0.0508 | F1 0.7391±0.0156 || TP~56.4 FP~27.8 FN~12.0 TN~82.0
[t=0.50] ACC 0.8227±0.0353 | P 0.8039±0.0352 | R 0.7103±0.0860 | F1 0.7527±0.0581 || TP~48.6 FP~11.8 FN~19.8 TN~98.0
[t=0.80] ACC 0.7564±0.0804 | P 0.7602±0.4264 | R 0.3854±0.2216 | F1 0.5102±0.2893 || TP~26.4 FP~1.4 FN~42.0 TN~108.4
```

cross validation揭示了在Fold 4資料當中存在大量非典型樣本，導致在threshold = 0.8時產生異常數據。precision, recall, f1 = 0.000 (TP = 0, FP = 0)，模型同樣都對positive無法產生足夠的信心判斷，顯示出標籤與特徵的矛盾。

Conditional Survival Rate



確認原本的sex與pclass這條特徵組合對於survived的資料客觀允許機率，可以看見該組合具備分類性，雖然存在模糊帶，但其可達信心上限仍受樣本數、模糊子群、正則化與模型校準影響。
我所採用的validation loop驗證了在目前feature space與模型族群下，性能達到上限，在反覆的驗證與推翻之後，目前沒有更好的答案，也或許沒有最好的答案，只有合適的答案。