Data Mining Howork2 作業模板

作業回答共 20 分，未繳交或未做答 0 分計算。

1. 方法選擇（Method Selection）4 分

・請說明你選擇的模型：

本次作業選用了三種常見的梯度提升模型：XGBoost、LightGBM 與 CatBoost，並將預測結果以三者預測取平均。

・為什麼選擇這個方法？

有參考原始資料的開放程式碼，發現高分的作品(大於 0.89 的)都有使用這三個模型進行訓練，實際跑我們的資料也有不錯的成績，起始公開成績就有 0.90129，按照模型的特性都是滿適合分類的樹模型，在比賽中也常得獎。

・為什麼不選擇其他方法嗎？

針對分類的問題，就優先想選擇跟樹模型相關的模型，雖沒有嘗試其他模

型，但有比較這三個模型在訓練上的彼此關係，auc 分數都滿接近的。

2. 特徵工程（Feature Engineering）8 分

・你有處理缺失值（Missing Values）嗎？

使用 SimpleImputer(strategy='mean') 來補全數值欄位的缺失值，但資料本身沒有缺失值，所以其實不太需要。有部分有異常值，有使用 winsorize 處理，但成績並沒有提升，故仍保留讓模型進行訓練。

・你有做特徵篩選（Feature Selection）嗎？使用了什麼方法？

使用 SHAP 值判斷模型中影響預測的重要特徵，並根據重要性刪除低權重特徵。共刪除了 5 個特徵，主要是類別特徵，吸煙與不吸煙沒有明顯差異的。

此外，也有手動創造特徵如 BMI 指數（BMI = weight / (height/100)^2）。雖然在後續訓練時，發現對於成績沒有提升效果，後續又刪除特徵。

・是否有進行特徵轉換（例如標準化、偏度處理）？

所有數值型欄位先經 PowerTransformer(method='yeo-johnson') 轉換偏態分佈，接著進行 MinMaxScaler 標準化處理。類別型特徵則透過 OneHotEncoder 做轉換。

但後來看書發現 boost 的模型不一定需要標準化，所以有試過沒標準化的版本，但成績沒有更好，所以就維持標準化的版本繼續分析。

3. 交叉驗證與模型調整（Cross-validation & Model Tuning）5 分

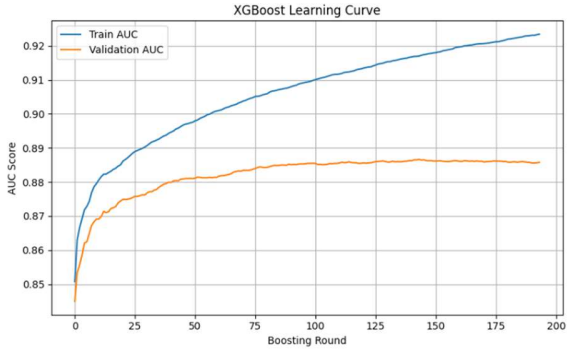
・你使用了哪種交叉驗證方法（k-fold, stratified k-fold, train-test split, etc.）？

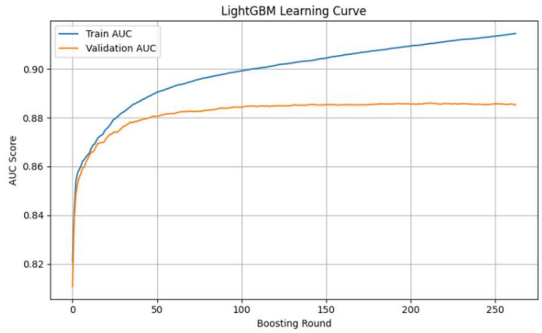
採用 StratifiedKFold(n_splits=5)，確保每一折中的類別分布與整體一致，避免偏差。

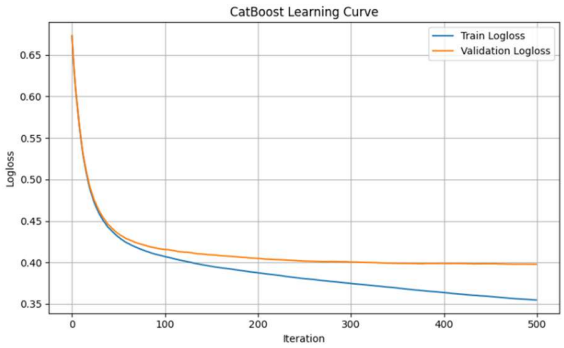
・如何選擇最佳超參數？是否使用 learning curve / grid search / random search？

使用 Optuna 自動化調參，目標最大化 validation AUC。並有記錄調參歷程與最佳組合。

- 請附上 learning curve 或 tuning 過程的圖表。

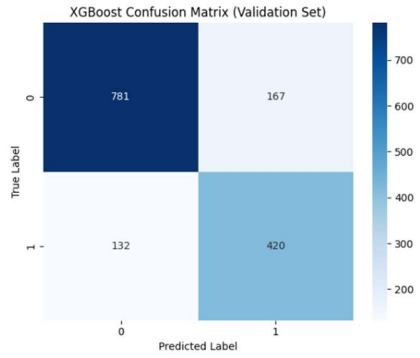


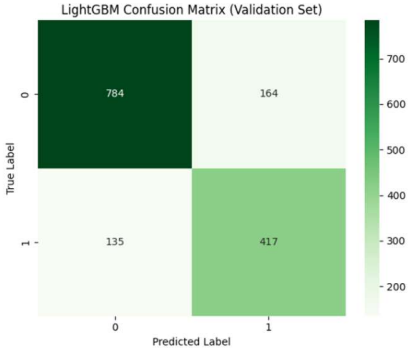


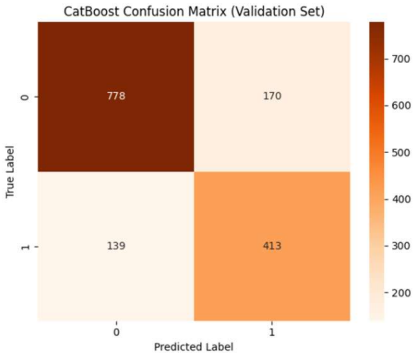


4. 模型表現與分析（Model Performance & Analysis）3 分

・請畫出混淆矩陣（confusion matrix）







・你的最終模型表現如何（Test AUC, Valid AUC etc.）？

Test AUC Public score:0.90219

三個模型個別的 Valid AUC：

XGBoost Best AUC: 0.8839

LightGBM 0.886047

CatBoost 0.8835

・（可選）你有比較不同方法的表現嗎？如果有，結果如何？
沒有，以上面三個模型為主。

5. 結論與反思（Conclusion & Reflection）

・這次作業中遇到的主要挑戰是什麼？

資料的特徵增加與減少、或是異常、偏態的處理，有時以為增加了這些處理會帶來更好的提升，但不盡然，覺得還是沒有找到逐步讓模型提升的方法，不太有一個處理的順序，一堆變因更改，但組合起來可能成績還是一樣不好。

・你認為你的模型還可以如何改進？

滿希望可以跟前幾名的同學交流，想瞭解自己處理數據上是不是有漏掉或盲點的地方。