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，及為什麼選擇這個方法？為什麼不選擇其他方法嗎？

Answer:
一開始先利用不同種類的模型進行基本的預測，使用 Logistic Regresstion、Randomforest、DicisonTree 以及 Gradient Boosting 算法，經過測試發現 Gradient Boosting 和 Logistic 兩種模型能夠達到比其他種模型更好同時更快的結果，結合作業的提示我優先選擇這兩種算法，而在經過超參數調整後發現 Gradient Boosting 中的 Catboost 算法雖然需要耗最久的時間卻可以達到最好的 AUC 分數，因此在沒有時間限制的情況下選擇 Gradient Boosting 中的 Catboost 做為方法。

2. 特徵工程（Feature Engineering）8 分

• 你有處理缺失值（Missing Values）嗎？

Answer:
在利用 featuretool 套件生成不同特徵組合後有利用 numpy 套件將新特徵的缺失值或是無限大值轉換原數據的中位數。

• 你有做特徵篩選（Feature Selection）嗎？使用了什麼方法？

Answer:
在這次作業中原來有採用三種不同的特徵篩選或壓縮方法，分別是利用相關係數選擇、LFDA 特徵壓縮、CatBoost 嵌入式特徵篩選，那在經過測試後前二者不僅無法提高模型準確度，甚至造成執行時間增長因此最後決定放棄採用，改用模型內嵌的特徵篩選做為主要方法。

• 是否有進行特徵轉換（例如標準化、偏度處理）？

Answer:
由於選用的是樹模型，根據它的計算原理是不需要做標準化的，不過在寫的過程中也有測試，但結果並沒有讓模型表現變得更好，偏度處理也是一樣的情況。因此為了減少運算時間就不採用了。

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.）？

Answer:

使用 stratified k-fold，讓樣本平均並輪流做為驗證集，最後取最好的 AUC 作為結果。

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

Answer:

選擇超參數的方法是利用 grid search，但為避免執行時間過長，採用一次一個或兩個參數為一組進行最佳化。

• 請附上 learning curve 或 tuning 過程的圖表。

Answer:

1. iteration tuning:

```
Best Parameters: {'iterations': 500}
Best CV AUC: 0.8976086185719915
CatBoost Test log_loss: 0.3745
CatBoost Test AUC: 0.9101
```

2. depth tuning:

```
Best Parameters: {'depth': 4}
Best CV AUC: 0.8986158942636626
CatBoost Test log_loss: 0.3748
CatBoost Test AUC: 0.9092
```

3. l2_leaf_reg:

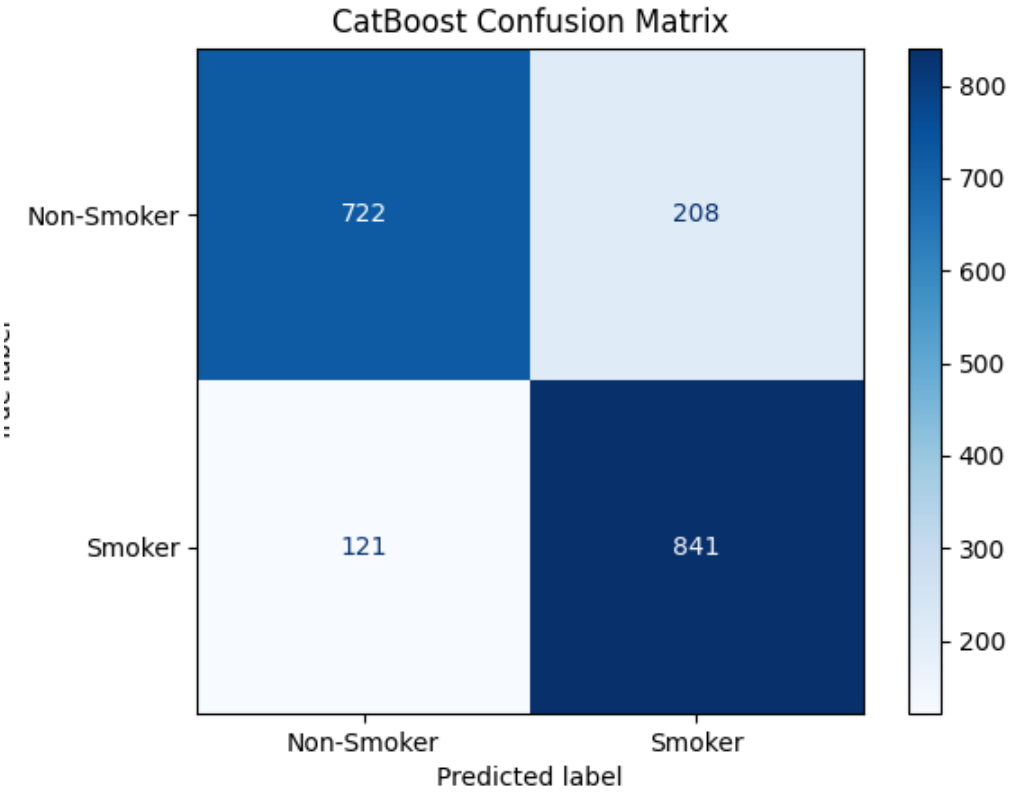
```
Best Parameters: {'l2_leaf_reg': 4}
Best CV AUC: 0.8977185855135298
CatBoost Test log_loss: 0.3750
CatBoost Test AUC: 0.9089
```

4. learning_rate:

```
Best Parameters: {'learning_rate': 0.05}
Best CV AUC: 0.8975708501800144
CatBoost Test log_loss: 0.3745
CatBoost Test AUC: 0.9093
```

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.）？

Answer:

Validation AUC(選最好的):0.8964

Test AUC:0.9101

•（可選）你有比較不同方法的表現嗎？如果有，結果如何？

Answer:

如第一題所提到的同樣以 AUC 分數作為評斷標準的話，Gradient Boosting 和 Logistic 兩種模型能夠達到比其他種模型更好同時更快的結果，而在經超參數最佳化後則選用 Gradient Boosting 中的 CatBoost 算法。

5. 結論與反思（Conclusion & Reflection）

• 這次作業中遇到的主要挑戰是什麼？

Answer:

我認為這一次所遇到的最大挑戰在於如何製作出有用的特徵，從相關係數的量測結果來看幾乎沒有一個特徵對是否有抽菸有大於 0.5 的相關程度，因此特徵重製勢必非常重要，但可惜的是我並沒有相關領域的知識可以輔助我自行建立有效的特徵，只能依靠自動生成，但結果依舊不如預期，雖然在驗證集表現有所增長 ，但需要耗費數倍的時間，若是作業有時間上的限制將會變成一大挑

戰。再者，模型過擬合的問題也很嚴重，雖然從資料集中切出訓練集、驗證集和測試集，甚至在測試集和驗證集的 AUC 都得到非常好的結果，但在上傳到 Kaggle 平台檢驗時卻少了將近 10%的 AUC 分數，因此或許需要考量到實際測試的結果不要一味追求本地端的結果，而是在兩者之間取得平衡，這是我在這次作業遇到的兩個挑戰。

• 你認為你的模型還可以如何改進？

Answer:

可以觀察到在 Kaggle 平台上有將近 9 成的人 AUC 分數是介於 88 到 91 這個區間但我的確只有 82 左右，因此我想在方法的選用以及特徵處理上似乎就有很大方向的不同，或許可以試著找更多不同模型甚至使用 ensemble 的方式，也需要找到不同的生成特徵方法以及篩選形式，當然不排除我一開始為了樣本分布平均有去原資料集採集樣本導致訓練的資料分布與測試分布有所不同，這點可能需要等到隱藏排行榜出來才能確定。