Assignment 2

CAFe

111423016 陳敬元 111423017 王駿豪 111423057 李宣緯

Preprocessing

1. 將 id 欄位 drop 掉，進行 datatype 轉換(yt 轉為 float，selt_intro 轉為 str，time_stamp 轉為日期格式再對 timestamp 做排序)。

2. 根據 timestamp 將作答時間萃取成 time_delta，關聯性很浮動，並且表現並不突出，再加上老師認證 timestamp 可以直接砍掉，因此最後捨棄。

3. 新增 BMI feature。

4. 使用 spacy 和 nltk 對 self_intro 做情感分析，新增 Polarity(spacy)、Subjectivity(spacy)、sentiment(NLTK)、word_count(算 self_intro 的字數)等 feature，做完後將 self_intro 刪除

5. 對 phone_os、star_sign 做 One_hot encoding，將 phone_os_JohnCena 刪除(出現 phone_os_Windows Phone 大小寫問題)

6. 建立男、女性用詞(名詞、形容詞)辭典，並對照 self_intro，建立 male_word 及 female_word，若有出現則為 1，但是在做完 Rf_feature_importance 及 ANOVA 後，都獲得很低的關聯度，因此後面就並未採用。

7. 將 'height', 'weight', 'iq', 'bmi','sleepiness', 'fb_friends', 'yt', 'word_count' 八項 features 做 MinMax Scaling，以增進模型的效能，並降低 outliers 對訓練的影響。

8. 將資料重複值 drop 掉。

做完資料前處理後共有 26 個 feature，除了 gender 之外，包含了 height, weight, sleepiness, iq, fb_friends, yt, bmi, Polarity, Subjectivity, sentiment, word_count, phone_os_Android, phone_os_Apple, phone_os_Windows Phone, star_sign_天秤座, star_sign_天蠍座,star_sign_射手座,star_sign_巨蟹座,star_sign_摩羯座 star_sign_雙魚座,star_sign_水瓶座,star_sign_牡羊座,star_sign_獅子座,star_sign_處女座,star_sign_金牛座,star_sign_雙子座。

Expirenment design

Feature selection

Feature selection 我們採用 Sequential Feature selection(SFS)，SFS 會考慮所有可能的 feature 組合，透過比較 scoring 選出最佳組合，由於資料集並不平衡，我們的目標訂為提升模型的鑑別

能力， 因此我們的 scoring 採用了 **roc_auc**，而非 kaggle 上的衡量標準 accuracy。另外我們測試了三個分類器，包含 Lightgbm、SVM 和 XGBoost，選用這三個模型的原因是 Lightbgm 和 XGBoost 在 kaggle 競賽上為最常被使用的模型之一，且此兩模型作為 tree-based 的代表; 同時再加入 SVM 作為非 tree-based 模型進行對照。

Oversampling

我們使用了 ProWSyn 作為方法，原因為此方法的優勢在於能夠較好的保存 minority class 的重要 feature，我們認為這對於此種小樣本不平衡資料集能夠有較好的表現。

K-fold cross validation

我們採用 5-fold cross valiation，以降低結果的 bias。

我們總共進行了三個實驗，以下進行說明：

Experiment 1

控制變因：train_test_split = 0.9, SFS: scoring = roc_auc & cv =5, K-fold cv 次數: 5
操作變因：Backword ＆ Forward
Backward

```
SvcModel test
acc:0.47619047619047616
auc:0.415625
Xgboost test
acc:0.8095238095238095
auc:0.8062500000000001
LgbModel test
acc:0.9047619047619048
auc:0.86875
```

accuracy & AUC 最好的 model 的 ouput: {1:139, 2:104}
vs.

Forward

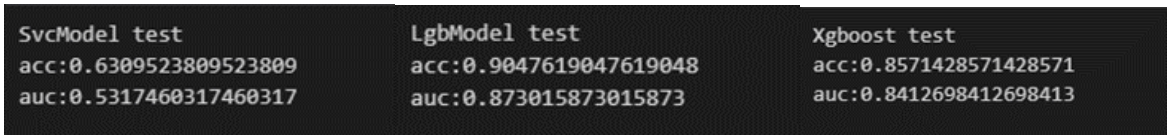
```
SvcModel test
acc:0.23809523809523808
auc:0.5
LgbModel test
acc:0.8809523809523809
auc:0.8875
Xgboost test
acc:0.8333333333333334
auc:0.821875
```

accuracy & AUC 最好的 model 的 ouput: {1:206, 2:37}

Experiment 2

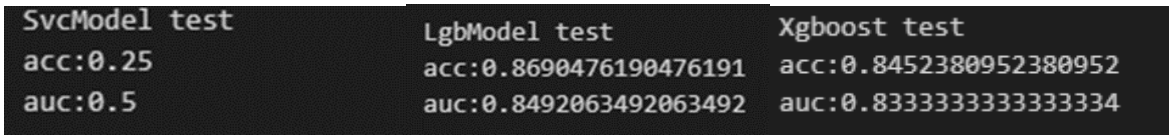
與實驗一做對照：
控制變因：Backward/Forward, SFS: scoring = roc_auc, cv =5, K-fold cv 次數: 5
操作變因：train_test_split = 0.8

Backward



accuracy & AUC 最好的 model 的 ouput: {1:226, 2:17}
vs.

Forward

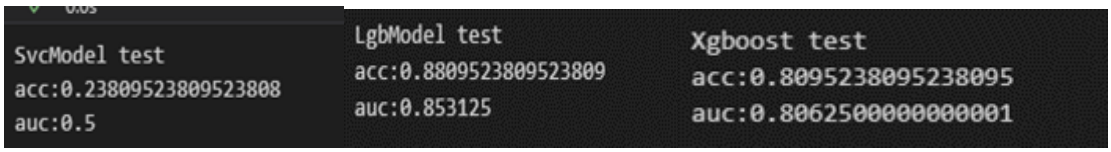


accuracy & AUC 最好的 model 的 ouput: {1:233, 2:10}

Experiment 3

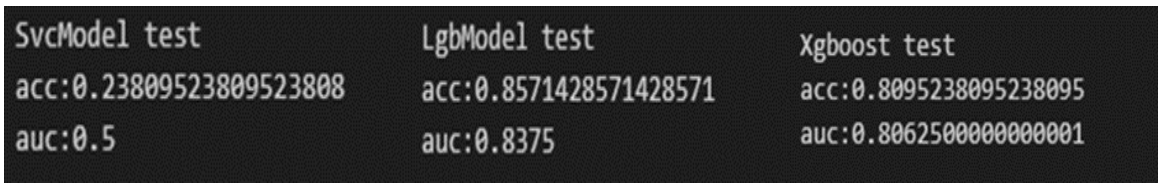
與實驗一做對照：
控制變因：train_test_split = 0.9, SFS: scoring = roc_auc, Forward/Backward , K-fold cv 次數: 5
操作變因：SFS: cv=10

Backward



accuracy & AUC 最好的 model 的 ouput: {1:216, 2:27}
vs.

Forward



accuracy & AUC 最好的 model 的 ouput: {1:228, 2:15}

判斷實驗成果的依據為：在全部實驗模型中，Accuracy 及 AUC 最高的可獲得一分，output 結果最貼近 test 資料分布的獲得兩分。

實驗結果

Expirenment 1 {train_test_split = 0.9, SFS: Forward, scoring = roc_auc, sfs cv =5, 5 fold cv, output: {1:206, 2:37} }、Expirenment2 {train_test_split = 0.8, SFS: Backward, scoring = roc_auc, sfs cv =5, 5 fold cv, {1:226, 2:17} }皆獲得兩分，因此我們採用組員投票並選定 Expirenment1: {train_test_split = 0.9, SFS: Forward, scoring = roc_auc, sfs cv =5, 5 fold cv, output: {1:206, 2:37} }。

Conclusion and Discussion

我們最高 submission accuracy 的做法是首先將空值直接捨棄，接著使用 SFS 做 feature selection 選擇出 star_sign、phone_os、height、weight、sleepiness、iq、fb_friends 共 6 個 feature，並對 phone_os 和 star_sign 做 label encoding，最後使用 RandomForestClassifier 做預測。

我們的最終 submission accuracy 僅有 Public: 0.63636 & Private: 0.60655, 我們認為導致此結果的最終原因可能是 overfittting，會有此結論的原因是我們實驗所得出的 output 變動量相當大，並且在 kaggle 上面所反映出的 accuracy 也相當低。

我們在做作業時採取的策略是先盡量搜集可能會對模型有幫助的 features，並做用了許多方法利用原有的 feature 產生出新的且有幫助的 feature，再透過 feature selection 的方法幫助我們篩選出不適合的 features，並透過 cross validation 降低 bias。在實驗中途，我們就發現到前段所描述的問題，我們對此進行的處理是

1. 調整 cross validation 的次數
2. 使用不同的 oversampling 方法
3. 針對選用的 SFS 方法進行各項參數的嘗試
4. 針對離群值做調整，例如將不在[0.25,97.75]區段內的值以眾數、以第 0.25、97.75 百分位數取代，最後我們選擇使用 MinMax Scaler 保留原始資料分佈，並直接進行訓練
5. 將重複值 drop 掉以避免 overfitting

，但是其實效果都沒有顯著的提升。

未來可能解決方法：

可能需要使用不一樣的 Feature selection 的方法對資料進行處理，抑或是再針對離群值做更妥善地處理。

分工

分工	陳敬元	王駿豪	李宣緯
蒐集資料	V	V	V
文獻探討	V	V	
實驗設計			V
進行實驗	V	V	V
實驗報告	V	V	V