Homework5

競賽

一、小組各成員的姓名、系級與學號

1.統計113 朱庭暄 H24091215

2.統計113 黃纓婷 H24096168

3.外文111 劉倍嘉 B24071341

二、競賽敘述與目標

- **競賽敘述**

  由老師給我們的train.csv做資料前處理後再做訓練，預測test.csv的結果。最後使用三種評斷指標及不同的權重決定排名。

- **競賽目標**

  透過多種模型的分析，預測出與真實資料最相近的結果。

三、資料前處理

- 將非特徵('RowNumber','CustomerId','Surname')從dataframe中刪除
- 使用 StandardScaler() 將資料標準化

四、特徵處理與分析

- 使用onehot encoder將[“Gender”]的male跟female轉為0跟1
- 使用pd.get_dummies將[“Geography”]分成三個不同國家

五、預測訓練模型

- **SVC(linear)**

  -kernel=linear

  -C=0.5

- **Kmeans**

  -n_clusters=2

- **KNeighbors**

  -n_neighbors=5
- **xgboost (1)**

  -max_depth=7

  -min_child_weight =2

  -gamma=0.12

  -n_estimators=1800

  -learning_rate= 0.01
- **MLP**
- **SVC(rbf)**

  -kernel=rbf

  -gamma=0.04

  -C=0.85
- **SVC(poly)**

  -kernel=poly

  -gamma=0.04

  -C=0.85
- **SVC(sigmoid)**

  -kernel=sigmoid

  -gamma=0.04

  -c=0.87

- **decisiontree**
- **logistic regression**

  -沒有設hyperparameter
- **random forest**

  -n_estimators = 10,

  -min_samples_leaf = 9

  -max_features = "sqrt"
- **xgboost (2)**

  -verbosity = 0

  -subsample = 0.5

-min_child_weight = 5

-max_depth = 8

-gamma = 7

-eta = 0.8

六、預測結果分析

模型名稱	Accuracy	Precision	FScore
SVC(linear)	-	-	-
Kmeans	0.575	0.2372	0.3307
KNeighbors	0.845	0.7142	0.4464
xgboost (1)	0.885	0.7627	0.6617
MLP	0.8775	0.7187	0.6524
SVC(rbf)	0.8825	0.8750	0.5983
SVC(poly)	-	-	-
SVC(sigmoid)	0.855	0.9523	0.4081
decisiontree	0.78	0.4408	0.4823
logistic regression	0.815	0.615	0.17
random forest	0.8425	0.733	0.411
xgboost (2)	20.85	0.729	0.473

註:SVC(linear)第一次做完後發現linear不可行，因此後續就沒有上傳結果了。

註:SVC(poly)不小心把’exited’打成’excited’上傳，所以上傳的結果不準確(但最後一天次數用完了咩噗)

1.使用SVC(linear)、kmeans、kneighbors、xgboost模型做訓練

- SVC(linear)預測出來的結果都是0，將kernel從linear換成rbf後，結果滿準確的，因此確定不是資料處理的問題，初步推測可能是線性的方法不可行。跟第二位同學做對照後，發現SVC(rbf)準確率高很多，代表此data不適用SVC(linear)。
- Kmeans分析結果很差，猜測是因為kmeans模型為非監督式學習，沒有用到train中的exited資料結果，只是純粹將data分成兩群，並讓電腦分配是0或1，所以造成結果準確率不高。因為最後結果只有0與1，所以參數只能設定n_clusters=2。
- KNeighbors是我前三個模型中做的最準確的一個，嘗試到最後發現n_neighbors=5時是最準確的，accuracy/precision/f-score的最佳結果大約為0.84/0.62/0.5。
- 對上述三個模型的結果不大滿意，因此開始嘗試xgboost，結果比我做的其他模型好上許多。xgboost中我有調整了幾個參數，最後決定使用max_depth=7,min_child_weight =2,gamma=0.12,n_estimators=1800, learning_rate= 0.01這些參數設定，accuracy/precision/f-score的最佳結果大約為0.885/0.762/0.66，雖然最後不是所有模型中最佳的結果，但xgboost的表現還是相當不錯的。

2.使用MLP 、SVC(rbf)、SVC(poly)、SVC(sigmoid)、decisiontree 模型做訓練

- decision tree 分析稍有偏差，猜測原因是當要分析的類別太多時，預測的誤差率就會越大
- MLP透過多層感知訓練的結果優於decision tree，但比svc rbf 差，如果多花時間調整參數或權重可能會跟svc rbf 的結果並駕齊驅，表現還算行
- SVC(rbf) 一開始的表現就比之前兩個好
- 後來開始嘗試修改SVC(rbf)的gamma參數，最後試到gamma=0.04 C=0.87(中間有請教系上學長)得出最後結果，訓練跟預測的結果都比之前好了
- 最後有用SVC的其他方法做分析，poly & sigmoid 等，無奈礙於上傳次數不夠，無法再試其他參數結果
- 從中發現SVC算是蠻強大的訓練模型，但因為資料不是用於線性分析，所以第一位同學做出來的偏差很大。後來我有試RBF、SIGMOID 上傳結果都還不錯(SIGMOID 上傳的precision 在public 裡甚至高達了0.95)。

中間有因為怕上傳太多無用的資料，所以用上課所學的train-test split分割 train的資料，後來發現這個方法太耗時間，而且幾次分析都沒有比之前的準確率高，算是失敗的一次經驗。

3.使用logistic regression, random forest, xgboost模型做訓練

- xgboost使用random search估計subsample, min_child_weight, max_depth, gamma, eta等最佳參數值，但結果不如一開始只有自行指定gamma數值好
- 其他模型也做了random search，結果跟自行調整相差無幾
- 嘗試刪減變數做預測，結果不如全部使用來的精準
- accuracy, precision & fscore大小均是：xgboost > random forest > logistic regression
- 標準化可以降低outlier的影響，但測試結果是有沒有做都沒影響，推斷這次資料的outlier極少

七、感想與心得

朱庭暄

這次競賽對我來說是全新的體驗，不只是單純的寫一份作業，而是與組員共同分析一份資料。我們這組是每人分配幾個模型分開做，在最初選擇模型時，因為沒有相關的經驗，不大知道那些模型對這次的資料比較適合，只能一個一個慢慢嘗試。在嘗試的過程中，不斷地找尋網路範例，也在過程中大概地了解了每種模型的原理、計算方法以及何時適用。除了上課中學到的模型，我自己也去自學了一些其他的模型(因為我最初選的SVC(linear)、kmeans、kneighbors結果都不到非常理想)，最後也實作了xgboost。這次競賽過程中，對我來說最困難的就是調整參數。每種模型都有他們不同的參數設置(尤其是xgboost要調整的參數更是多)，每次調整參數前，都要先去找尋各個參數的解釋，也要去學習哪個參數影響最大、哪個參數需要先做調整...，而我也學會了利用做圖或寫迴圈去找尋最佳的參數大小。很喜歡這種競賽，而我也在競賽中對於機器學習漸漸有興趣也越來越熟悉了，希望期末報告與大學的未來兩年中，能夠學到更多關於這種分析與預測資料的知識。

黃纓婷:

- **所遇到的困難:**
    1. 一開始分析資料沒有標準化，雖然老師上課有講，但我可能天分不夠，沒有聽的很懂，所以有問同學跟查找網路。

2. 要做出來可上傳的檔案其實不難，我覺得難的地方是如何提高準確率、fscore，尤其是fscore，一方面是因為權重最高，另一方面是因為fscore很難提高，常常會發現前兩個指標很高，但fscore不大理想的狀況。

- **所學:**
    1. 老師上課教了很多模型，但資料是第一次接觸，所以要先觀察這筆資料的訊息，再整合之前所學的，做資料分析。
    2. 在這次小組合作的過程中，也有學到表達，因為平常日常生活中，不需要把程式的東西化成言語跟別人溝通，這次有試著把程式化成語言稍微溝通。
    3. 在程式上知道最多的是SVC這個強大的分析模型，有四種kernel，在此次的競賽中很適用，他不僅可以有效的處理高維度的數據，也可以透過改變kernel做不只是linear 的分析，像是poly,sigmoid。最後在rbf試到最高，如果有機會其實可以多試試看sigmoid。

這次競賽算是我第一次接觸到比較實際的資料，也結合之前很多次的所學的演算法，一同運用在這次的競賽中。之前大一的計概只有自己做小專題的經驗，這次的小組競賽算是我第一次跟組員分工合作比賽，雖然大家沒有一起聚在一起打程式上傳，但看的到大家努力把成績弄得更好，也會分享一些參數的調整，我覺得最後有超越我的預期了，希望之後可以多試試看類似的競賽，增強自己的實力。

劉倍嘉

接觸到sklearn後，第一個想法是，如果只要把資料丟進去套個模型，答案就出來的話，我何必花一整個學期去寫它背後的演算法呢？後來才明白，那些演算法才是精隨所在，也是在學習之後才更能對模型了解透徹。透過這次競賽，我學到了如何將資料轉換成能被讀懂的形式、挑選適當的變數及參數，以及耐心閱讀官方文件。然而找出適合的參數及其數值也是最困難且耗時的部分，因為適合訓練資料的參數值不一定適合測試資料。可惜由於上傳次數的限制，有些後來修改的檔案沒有機會跟正確答案做比對，希望之後能有無限次上傳而不列入排行榜的功能。

Github

repository網址: https://github.com/ying1027/yingdatascience

github html pages : https://ying1027.github.io/yingdatascience/HW3/hw3_H24096168.html