

資料科學導論HW5報告

組員：工資111 H34076144 王彥評
工資111 H34076178 吳羽泰

1. 競賽敘述及目標

本次競賽之主題為銀行用戶之客戶流失預測，目的為探討影響客戶流失之重大因素，藉此建立流失預測之模型，以期能降低客戶流失。

2. 資料前處理與特徵處理

- 2.1. 資料數值化 :利用pandas.get_dummies將Geography以及Gender欄位的文字資料數值化，轉換為one-hot encode的形式。
- 2.2. 資料標準化 :利用Z-Score Standardization將CreditScore, Balance 以及 EstimatedSalary欄位之資料轉換為Z-Score，目的為減少資料之間的差距，平衡各個特徵對於預測結果的貢獻，以增加模型的準確度。
- 2.3. 特徵處理:在觀察過資料集之後，我們決定把CustomerID以及Surname這兩個欄位排除在外，原因為這兩個特徵的值都是用來識別用戶的，對於預測訓練模型的建立沒有影響。

3. 預測訓練模型

- 3.1. XGBClassifier (n_estimators=200, max_depth= 3, use_label_encoder =False, eval_metric="error")
- 3.2. MLPClassifier(hidden_layer_sizes=(17),random_state=0)
- 3.3. svm.SVC(kernel='linear')
- 3.4. KNeighborsClassifier(n_neighbors=5)
- 3.5. LogisticRegression()

4. 預測結果分析

- 4.1. XGBClassifier: 當XGBClassifier參數設定成 n_estimators=200, max_depth= 3能得到我們在本次競賽中最適合的模型。

Parameters	Accuracy	Precision	F-Score
n_estimators=100, max_depth= 3	0.8825	0.7500	0.6569
n_estimators=100, max_depth= 4	0.87	0.6923	0.6338
n_estimators=160, max_depth= 3	0.8825	0.7500	0.6569
n_estimators=200, max_depth= 3	0.885	0.7627	0.6618
n_estimators=300, max_depth= 3	0.875	0.7213	0.6377

4.2. MLPClassifier

Parameters	Accuracy	Precision	F-Score
hidden_layer_sizes = (11,)	0.8575	0.6667	0.5839
hidden_layer_sizes = (16,)	0.8625	0.6719	0.6099
hidden_layer_sizes = (17,)	0.8725	0.7407	0.6107
hidden_layer_sizes = (23,)	0.8675	0.7069	0.6074
hidden_layer_sizes = (5,5)	0.86	0.6780	0.5882
hidden_layer_sizes = (20,10,2)	0.86	0.6721	0.5942

4.3. svm.SVC

Parameters	Accuracy	Precision	F-Score
kernel='linear'	0.8225	0.6667	0.2526

4.4. KNeighborsClassifier

Parameters	Accuracy	Precision	F-Score
n_neighbor=2	0.7975	0.4286	0.2286
n_neighbor=5	0.835	0.6122	0.4762

4.5. LogisticRegression

Parameters	Accuracy	Precision	F-Score
Default	0.82	0.5862	0.3208

4.6. 最終上傳之模型與相關建議：從上述數據來看，binary classification比較適合利用decision tree的分類方式，所以利用 XGBoost或是Random Forest會得比較好的效果。另外，關於資料前處理部分，因為本次競賽中feature數量並不多，因此利用PCA或是Feature Selection都不會有太好的成效。此外，我們有試著從各個特徵值中挑選出離群值(標準化後絕對值大於3的)，這Neural Network 或是 KNN中都要讓準確率上升，但當此篩選離群值的方法運用在XGBoost時，反而會造成部分資訊流失，而得到較差的效果。

5. 感想與心得

王彥評：這是我第一次參加資料分析相關的競賽，而我其實在這一學期因為修習了這堂課才對於資料分析領域有所接觸。透過這個競賽，我更完整的學習了整個預測模型的建立，其中包括了一開始的資料前處理，像是將類別型文字資料轉換為0和1的one-hot encoding，以及改善每個特徵間資料差異的Z-score和MinMaxScaler；以及接下來選擇建立模型的方法，除了上課所教的像是Logistic regression和SVM等等方法以外，透過自學也了解到像是Multilayer Perceptron以及XGBOOST等等較為進階的分類預測方法。競賽中最麻煩的部分我想應該就是測試各方法中每個參數要如何設置才能達到最好的預測效果，而有遺憾的部分是還有其他課要顧慮所以沒辦法全心全意地投入競賽，依舊有許多方法尚未做學習及嘗試。總結來說，我認為這樣的競賽對於課程的有趣程度以及學生的學習是很有幫助的，尤其是對才剛接觸資料分析的我，透過競賽能使我學到更多，不只學到了知識也學到了實作。最後，在這邊謝謝我的組員吳羽泰，也謝謝李教授以及助教們，備課辛苦了！這是堂好課！

Github: <https://github.com/WYP2189114/DataScience>

Github_page:

https://wyp2189114.github.io/DataScience/HW3/H34076144_hw3.html

吳羽泰：透過競賽的參與，讓我更加熟悉課堂上老師所教的Scikit-Learn的套件以及各個資料分析工具，比如說KNN和SVM等方法，也透過自學的方式學習到機器學習常勝軍XGBoost 套件的原理。在整個競賽期間，花費最多時間的部分為去理解每個資料分析工具的用處，從資料前處理的Z-Score Standardization、PCA以及SMOTE，到資料分析階段的XGBoost、SVM，必須真正理解其中的原理，才可以依據資料集特性來選取適合資料分析工具，以獲得最佳的分析效果。最後，李政德老師及助教們辛苦了！在這堂課學到很多！

Github: <https://github.com/wyt890301/Introduction-to-Data-Science>

Github_page:

https://wyt890301.github.io/Introduction-to-Data-Science/hw3/hw3_web/home_page.html