玉山人工智慧

公開挑戰賽

隊伍：Brainchild

成員：徐正憲、李坤瑋、劉家達、楊閔程、莊子逵

★摘要

經分析與研究後，發現資料集 dp 產製的統計量特徵較其餘資料有效偵測洗錢活動，且與不同時間粒度結合，將有助於提升偵測犯罪者洗錢活動的準確度，例如：當日交易差額比率、當時交易筆數與交易時間差等。

模型演算法採用常見的樹狀分類器(e.g. XGBoost, LightGBM, ngboost, CatBoost, GradientBoosting, HistGradientBoosting)與集成學習(Ensemble Learning)，並透過網格搜尋法(Grid Search)最佳化模型參數。

★環境

系統平台	Jupyter Notebook/Google Colab
程式語言	Python
函式庫	import pandas as pd import numpy as np from xgboost import XGBClassifier from sklearn.model_selection import train_test_split from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score from sklearn.ensemble import RandomForestClassifier from sklearn.ensemble import HistGradientBoostingClassifier import catboost as cb import ngboost as nb from sklearn.tree import DecisionTreeRegressor, DecisionTreeClassifier from sklearn.ensemble import GradientBoostingClassifier import lightgbm as lgb from sklearn.ensemble import StackingClassifier from sklearn.neural_network import MLPClassifier from sklearn.linear_model import SGDClassifier from sklearn.svm import LinearSVC, SVC from sklearn.naive_bayes import GaussianNB

★特徵

使用資料集 dp 產製以下共 12 個特徵供 model1 使用。

Name	Description
session_amt_diff_ratio	session 交易差額比率
date_time_amt_diff_ratio	當時交易差額比率
tx_cnt_date_time	當時交易筆數
txbranch_day_time_cnt	當時總分行數
day_time_atm_txn_ratio	當時 ATM 佔交易數比例
day_time_cross_bank_ratio	當時跨行 佔交易數比例
date_amt_diff_ratio	當日 交易差額比率
tx_cnt_date	當時交易筆數
txbranch_day_cnt	單日總分行數
day_atm_txn_ratio	當日 ATM 佔交易數比例
day_cross_bank_ratio	當日跨行 佔交易數比例
time_diff	交易時間差

使用 model1 output 及 dp 產製以下共 12 個特徵供 model2 使用。

Name	Description
1	風險等級 1
2	風險等級 2
3	風險等級 3
4	風險等級 4
5	風險等級 5
6	風險等級 6
7	風險等級 7
8	風險等級 8
9	風險等級 9
10	風險等級 10
db_cr_ratio	Db 占比
all_txn_cnt	交易次數

★訓練模型

使用模型與對應參數請參考下表。

模型名稱	模型參數
XGBoost	XGBClassifier(base_score= 0.5, booster= 'gbtree', colsample_bylevel= 1, colsample_bynode= 1, colsample_bytree= 1, gamma= 0, learning_rate= 0.05, max_delta_step= 0, max_depth= 3, min_child_weight= 1, n_estimators=200, nthread= 16, objective= 'binary:logistic', reg_alpha= 0.1, reg_lambda= 2, scale_pos_weight= 1, seed= 0, subsample= 1, verbosity= 1)
LightGBM	lgb.LGBMClassifier(learning_rate = 0.1 , max_depth=3 , reg_lambda=0 , n_estimators=100 , reg_alpha=0.01)
nboost	nb.NGBClassifier(n_estimators = 500 , Base = DecisionTreeRegressor(criterion='friedman_mse', max_depth=4)

	, learning_rate = 0.1)
CatBoost	cb.CatBoostClassifier(learning_rate = 0.5 , max_depth=3 , reg_lambda=2 , n_estimators=150 , subsample = 1, verbose= 0)
GradientBoosting	GradientBoostingClassifier(learning_rate= 0.2, max_depth= 2, n_estimators=100, subsample=1, random_state = 0)
HistGradientBoosting	HistGradientBoostingClassifier(learning_rate= 0.05, max_depth= 3, max_iter=200)

★訓練方式及原始碼

[請說明本次比賽答案的產出方式並提供有效之原始碼(連結亦可)]

請參考 github 連結:

https://github.com/jasonliu1990/esun_winter_competition_2022

★結論

[請簡易說明本次競賽後所得的結論]

不用複雜的神經網路/深度學習模型，用簡單的特徵及一般機器學習演算法一樣可以達到相同的效果。