電子商務交易詐欺預測

蘇芷儀[1]、賴威博[2]、藍璟誠[3]、劉育佑[2]、趙駱翰[4]

[1]政大經濟 [2]政大資科 [3]政大資管 [4]政大地政

資料科學

Data Science

INTRODUCTION

- Background & Objective
  - 隨著電子支付的快速發展，交易詐欺也越來越普遍，不僅給個人帶來經濟損失，也對金融機構和市場造成嚴重影響。因此，我們希望能夠開發一個交易詐欺預測模型，以提高對交易詐欺行為的識別能力。
- Methodology Overview
  - 我們採用了多種機器學習模型來訓練和評估，包括Decision Tree、Random Forest、Logistic Regression、Support Vector Machine (SVM)、K-Nearest Neighbor (KNN)和XGBoost等。通過數據預處理、特徵工程和參數調整，最終選擇出最優模型進行預測。
- Key Steps
  - 數據預處理：包括處理缺失值、特徵縮放和類別變量編碼等。
  - EDA：通過可視化和統計分析，了解數據特徵及其分佈情況。
  - 模型訓練與評估：對多種模型進行訓練，並使用多種評估指標比較其性能。
  - 調整參數：利用交叉驗證和超參數調優技術進一步提升模型性能。

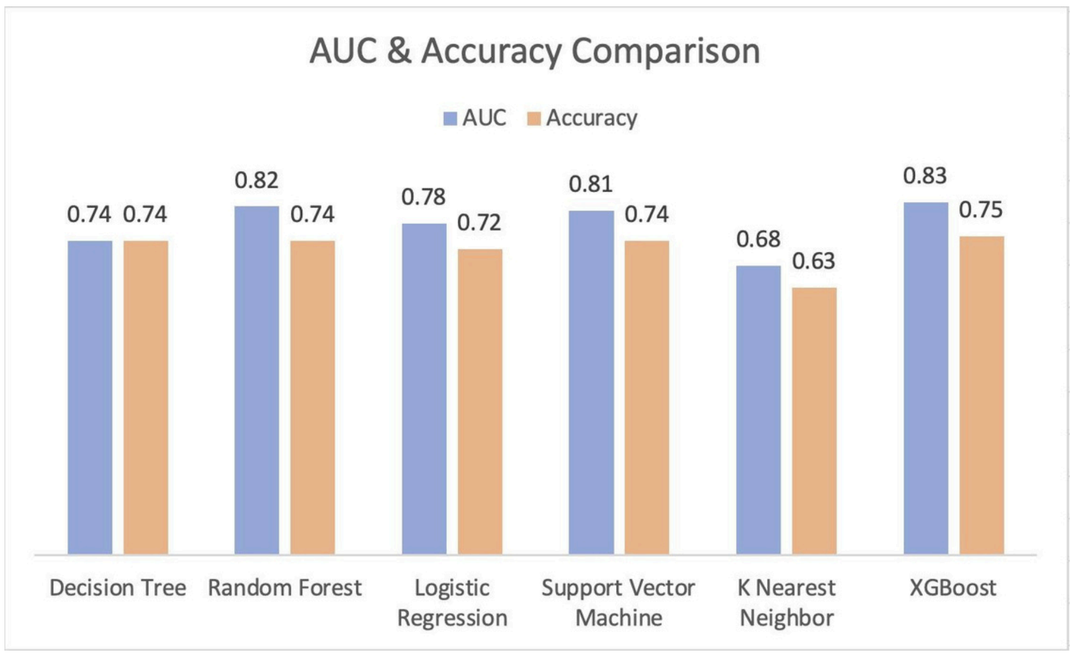
METHODS

- Datasets
  - 為了評估模型的效能，我們使用了兩組來自 Kaggle - Fraudulent E-Commerce Transactions 的資料集。第一組資料集包含1,472,952筆記錄，作為訓練數據（train data）；第二組資料集包含23,634筆記錄，作為測試數據（test data）。每筆資料包含15個特徵（features）和1個標籤（label），標籤0表示非電商詐欺，標籤1表示電商詐欺。
  - 初步的模型訓練使用清洗過後的訓練數據，在各模型中，我們皆達到了0.95以上的準確率（accuracy），但其特異性（specificity）極低。經過檢查，我們發現訓練數據的標籤分佈極不平衡。如果僅猜測所有樣本為非詐欺，仍能取得0.95以上的準確率。為了解決這個問題，我們採用了欠抽樣（undersampling）方法，使電商詐欺與非電商詐欺的樣本數量相等。
- Evaluation metrics
  - 模型選擇與理由： 我們選用了六種模型來進行電商詐欺預測分析，這些模型各有其優勢：
    i. Decision Tree：直觀易解釋，能夠捕捉數據中的非線性關係。
    ii. Random Forest：通過多個決策樹集成，提升模型的穩定性和準確性，減少過擬合。
    iii. Logistic Regression：適用於二元分類問題，結果具備概率解釋。
    iv. Support Vector Machine (SVM)：在高維空間中效果良好，適合處理特徵維度較高的數據
    v. K-Nearest Neighbor (KNN)：簡單易實現，對於特徵空間內相似的樣本效果好。
    vi. XGBoost：基於梯度提升的強大集成學習方法，具有高效性和準確性，適合處理大規模數據。
  - 指標 :
    - 利用Sensitivity、Specificity、Precision、Recall、F1、Accuracy、AUC等指標評估模型效能。

RESULTS

Testing Data

	Decision Tree	Random Forest	Logistic Regression	Support Vector Machine	K Nearest Neighbor	XGBoost
Sensitivity	0.90	0.78	0.72	0.79	0.66	0.78
Specificity	0.57	0.71	0.69	0.68	0.61	0.73
Precision	0.97	0.98	0.98	0.98	0.97	0.98
Recall	0.90	0.78	0.72	0.79	0.66	0.78
F1	0.94	0.87	0.83	0.87	0.78	0.87
Accuracy	0.74	0.74	0.72	0.74	0.63	0.75
AUC	0.74	0.82	0.78	0.81	0.68	0.83



CONCLUSIONS

- 根據六種模型的結果分析，我們發現XGBoost模型在電商詐欺預測分析中表現最佳。經過網格搜索調整參數後，XGBoost在測試數據上的預測結果顯示其準確率（Accuracy）達到0.75，AUC值為0.83。
- 在處理Positive case的策略上，我們考量到將標籤1（電商詐欺）作為Positive Case時，預測錯誤可能會導致顧客流失和經濟損失。因此，我們選擇將標籤0（非電商詐欺）設為Positive case，即寧可錯過一些電商詐欺的案例，也不願錯誤地將非電商詐欺的案例判定為詐欺。