

# 科技金融導論期末專題報告

## 第 21 組：個人投資風險管家

徐蘇緯  
r08921a30  
台大電機所

劉羽忻  
b06902126  
台大資工系

駱皓正  
d08227104  
台大心理所

## 第一部分 問題定義

本次專題將透過玉山證券所提供的最近九季（2019 Q1 至 2021 Q1）的投資人歷史交易資料（總數約 548 萬筆）、投資標的（公司）的相關產業資料以及投資人的基本資料中，試圖透過大量的交易資料達到以下兩個目標：

1. **適用於投資人的即時風險評估：**對於投資人而言，本專案試圖積極的從源頭（下單）前的行為，「即時」且合理的替客戶評估風險值，並在此專題中透過合理的分析資料來幫助投資人，避免下單的金額造成未來的違約。
2. **適用於券商的交易風險評估：**對於證券商而言，本專案試圖透過大量的客戶交易資料以機器學習的方法訓練出一個可靠的模型，針對想要投資的客戶，透過該客戶的基本資料與其交易資料以預測這個人當次交易的違約風險值。

## 第二部分 即時風險評估—適用於投資人

### 2.1 資料來源與特性

- 資料來源：證券交易所每日股票交易 API。
- 資料數量：近 n 天交易的開盤(open)、最高(high)、最低(low)、收盤(close)價格、以及交易量(quantity)。
- 資料特性與假設：
  1. 股票的波動其實在短期內的「波動率」是可預測的。  
假設以前十天為例，將最前十天的交易資料最高漲幅(%)及最大跌幅(%)取平均，最近 1 日的當天最大漲幅、最大跌幅也幾乎都會落在前十天的平均內→**一隻股票的波動率其實是有跡可尋的。**
  2. 「當沖的交易量」必須列入風險考慮。  
俗話說得好，新手看價、老手看量、高手看勢、傻瓜看電視。是的，誠如俗話所說，老手看量，每天股票的成交量也是能否獲利的重點，若成交量太低，就算有價差也可以瞬間風雲變色，所以在風險評估中成交量也必須列作風險考量的因子之一。一般來說，成交量越低代表在市場中的買賣熱門程度越差，如果遇到持有大量股票的狀況，很有可能會有賣不掉的風險，因此風險值對應的權重較高。
  3. 「越接近跌停價」當日買入的風險值越低。  
除了跌停鎖死不看，只要越接近跌停的價格，當天能跌的幅度就越有限相對來說風險值就越低。

## 2.2 程式與操作介面

- 輸入：股票代碼（或名稱）、欲下單的金額。
- 流程：即時呼叫爬蟲程式去抓取近 n 日的交易資料，步驟如下：
  - Step 1.** 計算出最前 n 天的交易資料的最大漲幅平均(%)與最大跌幅平均(%)。
  - Step 2.** 統計最近 10 天的平均交易量，將門檻劃分成低、中、高分別對應風險值權重高到低(1.2/1/0.8)。
  - Step 3.** 依照以下公式計算輸入的金額對應前一天收盤價格的漲跌幅對應到 Step 1 的範圍。假設 Step 1 算出範圍是+8%至-8%輸入金額為-2%，其對應的風險值為  $(-2) - (-8) / 8 - (-8) = 37.5\%$  屬於中低風險區。由此公式可合理的判斷，輸入的金額越低風險值就越低。

$$\text{即時風險評估公式} = \frac{\text{輸入金額漲跌幅}(\%) - \text{最大跌幅平均}(\%)}{\text{最大漲幅平均}(\%) - \text{最大跌幅平均}(\%)} \times 100\% \times \text{交易量權重}$$

- 示意圖：



- 輸出：
  - 最前 n 天的交易資料最大漲幅平均(%)及最大跌幅平均(%)、平均交易量。
  - 針對輸入的股票代碼及價格「即時」的風險高低評估。
- 程式介面



## 第三部分 交易風險評估—適用於券商

### 3.1 大綱與實驗路徑圖

本專案認為「高風險投資行為」等價於「投資人違約交割」之行為。在這個前提之下，本專案將以機器學習取向建構預測模型，以回答「給定某次交易資料，預測該次交易是否會發生違約交割（或是發生違約交割的機率）」，並找出「預測違約交割的關鍵因子」。具體而言，本專案包含以下幾個部分。首先，針對資料集的描述與觀察，透過此階段了解資料特性並找出適切的預測模型。其次，機器學習的方法介紹，有鑑於本資料的不平衡與分類特性，本專案將以異常檢測(one-class SVM anomaly detection)與基於樹的 boosting 分類法(XGboost)兩類取徑建構預測違約交割的模型。在此部分中，會詳細介紹各自取徑中的資料前處理、演算法、以及實驗情境。接著，本專案會以各類指標(precision, recall, lift chart)進行兩類方法的比較優劣，並給予實務上的建議。最後，我們的預測模型會成為前端的後端模型讀取前端資料，反饋前端風險值。介面如下圖所示。



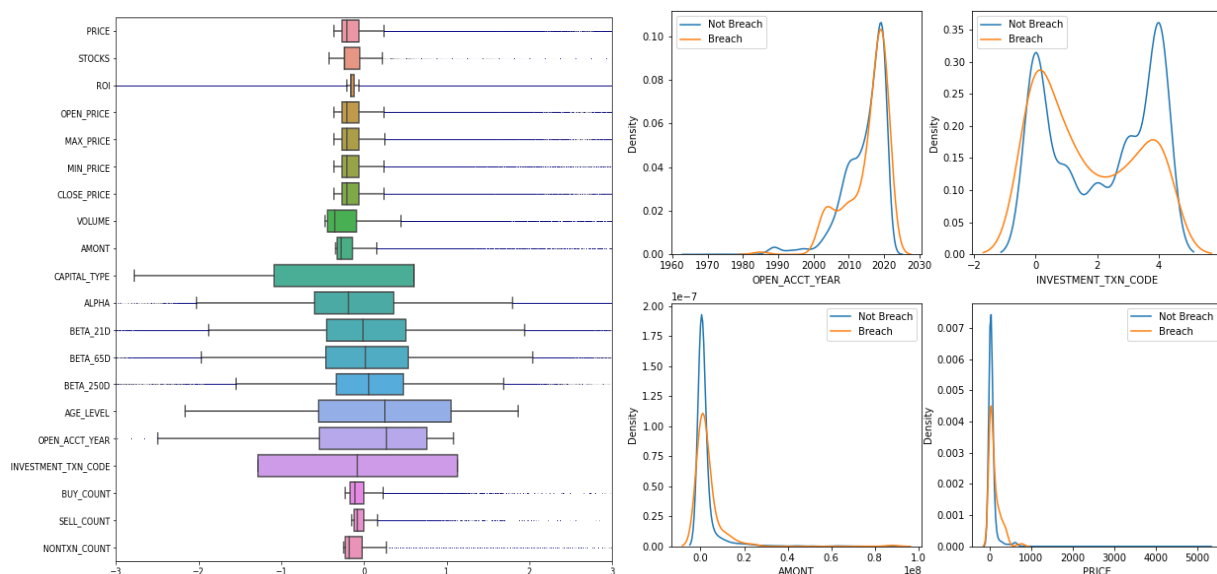
### 3.2 資料集描述統計與觀察

#### 資料來源與特性

- 資料來源：玉山證券 2019Q1 至 2021 年 Q1 共九季的交易資料。
- 資料數量：約五百萬至六百萬筆交易資料與約十萬筆用戶資料。
- 資料特性：
  - ☞ 資料的失衡性：本資料集在違約交割方面是極度失衡，在上述數百萬筆交易資料中，只有約兩百筆違約交割的資料（約 0.003%）。
  - ☞ 本資料集有許多資料缺失：大約有一百萬筆資料有缺值（包含部分缺值或完全缺值，共占約 12%）。究其原因，可能是部份股票資料在計算 beta 值時，會因股票下市等問題而發生缺值。針對缺值的處理，本專案採直接刪除該資料的作法。

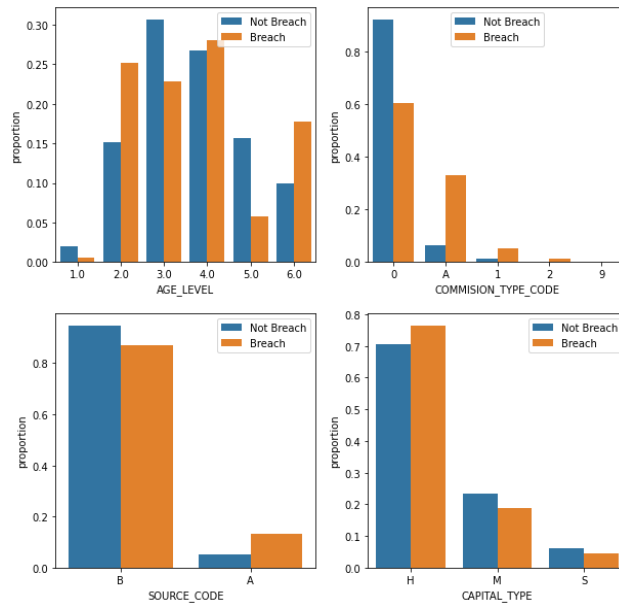
## 特徵分布與描述統計

- 連續型特徵分布：**首先，本專案將以盒鬚圖(boxplot)檢視連續型特徵的分布情形。為了可比較性，本專案首先將各連續型特徵進行標準化，據此，再繪製盒鬚圖。由下左圖個變項的盒鬚圖與離群值(outliers)分布可知，各個連續型變項的分布都呈現相當的偏度(high skewness)且非常態，因此不適用傳統常態假設的統計模型作為預測模型。



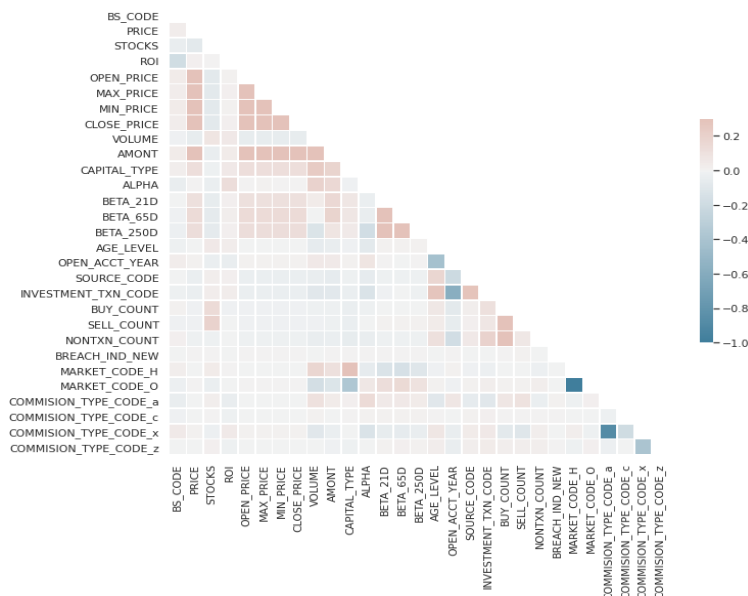
針對連續型的特徵，我們進一步檢視「投資人開戶年」、「交易經驗」、「交易時的價」、以及「交易時的量」，前兩特徵與投資人的投資經驗有關，而後兩特徵是投資人常參考的投資指標。本專案認為此四特徵可能與違約交割有關聯性，故將檢視此四特徵分別在非違約交割與違約交割的情境下的分布情形。值得注意的是，有鑑於資料的不平衡性，因此我們以機率密度圖(density plot)來表示。結果顯示（上右圖）：就「投資人開戶年」與「交易經驗」來說，可以發現到相較於非違約交割的資料、違約交割的資料分布較偏向於「開戶年度較近」或「交易經驗較少」，這可能是由於因為股市新手較容易判斷錯誤而導致。就「交易時的價」與「交易時的量」來說，可以發現到相較於非違約交割的資料、違約交割的資料分布較偏向於「交易量大時進行交易」或「交易價格較高」，量大時交易是常見的進場捷思法(heuristics)，導致風險的可能性也較高，而以較高的價格進行交易，也是極具風險的，因此這樣的行為與違約交割較相關實屬合理。

- 離散型特徵分布：**針對離散型的特徵，我們以長條圖(barplot)特別檢視「交易者年齡（等級）」、「投資平台類型」、「交易類型」、以及「標的公司規模」。前兩者與交易經驗有關（例如低年齡層、或年輕族群較喜愛的富果平台），而「交易類型」中不同型態的交易本身即對應不同投資風險，最後「標的公司的規模」對應投資者使用某類投資的捷思法，因此我們認為這些變項可能在非違約交割的交易與違約交割的交易上會有不同的型態。值得注意的是，有鑑於資料的不平衡性，因此我們的 y 軸為比率，以利比較。



在「交易者年齡（等級）」方面，即使分布與非違約資料類型類似，但亦可以看出違約資料的交易者年齡大多落在 20 至 50 歲年齡區間。在「投資平台類型」方面，可以發現相較於非違約交割的資料、違約交割的資料的投資人多偏好使用富果平台，這可能是因為富果平台的投資人多是年輕的剛進入股市的投資人。在「投資類型」方面，可以發現比起非違約交割的資料、違約交割的資料多發生在現股交易之外的現股當沖及當沖交易，事實上，這些交易類型的風險較大，故違約交割發生常發生在這些交易類型實屬合理。

- 特徵相關性與維度縮減：**本專題將玉山證券所提供的投資交易紀錄、投資客戶、產業基本資料等不同特徵(features)計算相關性矩陣(Pearson correlation matrix)，如下圖所示。值得注意的是，由於Pearson相關有資料連續的假設，故針對類別資料本專案先進行one-hot vector的前處理。透過相關矩陣，我們可以發現特徵兩兩間的相關並不高，相關係數多介於-0.2 至 0.2 之間。有鑑於變項間相關皆低，意味著變項間並無指涉至一或多個潛在特徵。因此，沒有為維度縮減之必要。

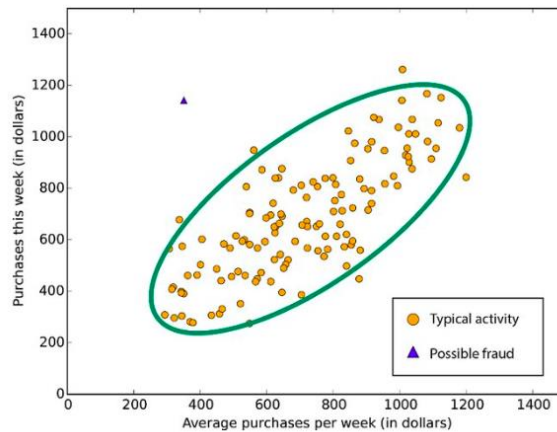




### 3.3 實驗設計

#### 取徑一：One Class SVM

- One Class SVM 介紹：OneClass SVM 是一個 unsupervised 的算法，顧名思義訓練數據只有一個分類。透過這些正常樣本的特徵去學習一個決策邊界，再透過這個邊界去判別新的資料點是否與訓練數據類似，超出邊界即視為異常，並得到如下圖所示的結果。

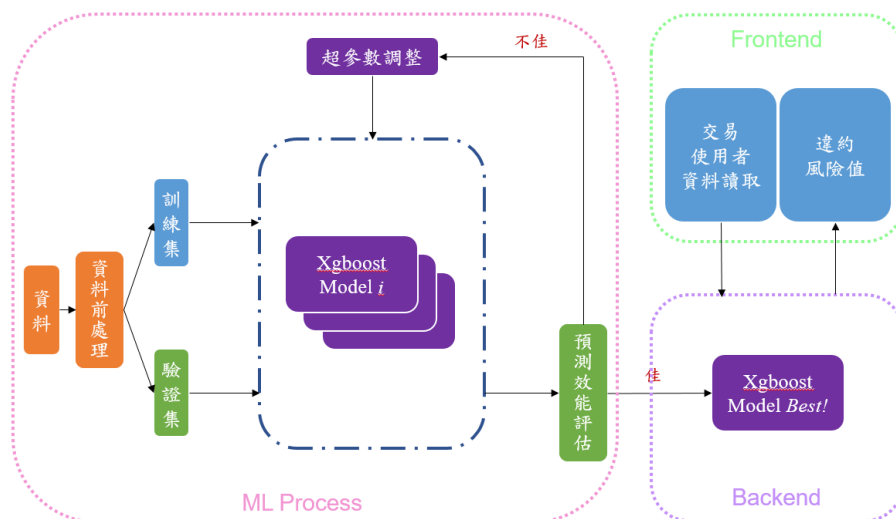


[Source from: Pattern Recognition “Anomaly Detection Challenges”](#)

- 資料前處理：把資料集中有 missing value 的資料剔除，留下的完整的資料集分類成有違約交割及無違約交割的資料分開，因完整的資料及過大，所以我們用挑了 5 萬多筆沒有違約的資料出來以降低訓練時長，再把無違約交割的資料以 0.99:0.01 的比例切成 training set 與 testing set(normal)，違約交割的資料則為 testing set(abnormal)。
- 模型訓練：在 scikit-learn 提供 OneClass SVM 的說明頁面有提到：預測返回的 y 值為 1 代表正常 (inlier)，返回值為 -1 代表異常 (outlier)。OneClass SVM 透過 unsupervised learning 的方式極小化正常樣本與邊界的距離，其中有兩個相當重要的 parameter，第一個是 nu，他控制了邊界的定義，舉例來說， $\nu=0.1$  即為正常樣本卻誤判為異常的最多不超過 10%。第二個是 gamma，gamma 代表了 RBF Kernel 將樣本投影到高維空間的縮放比例，gamma 值設定的越小代表樣本在高維空間的越分散，在訓練時的作用可能會造成準確率較低，但預測未知樣本的泛化能力強；反之 gamma 值設定的越大，樣本在高維空間會擠在一起，所獲得的 support vectors 就會較少，訓練時準確率較高，但預測未知的泛化能力弱。本次實驗中，我們設定的  $\nu=0.05$  而  $\gamma=0.1$ 。

## 取徑二：Extreme gradient boosting (XGBoost)

- 建模流程：下圖是，XGBoost 的建模流程。首先，先將資料進行資料前處理。爾後，將資料隨機分成訓練集與驗證集。下一步，使用訓練集建模，然後以驗證集做驗證，如果驗證的預測效能不如理想，則調整參數繼續下一輪的訓練與驗證。如果驗證的預測效能很符合標準，則停止訓練。爾後，報告預測效能並做為我們專案的後端模型。當前端給予交易投資人的資料時，模型會反饋給前端投資風險值，也就是其可能違約的機率。



- 資料前處理：針對資料前處理，主要可分為以下幾個部分：第一、預測變項：本取向納入除了身分標記的特徵作為預測變項，以預測是否違約（二元分類問題）。這些預測變項包含當次股市資料（開盤價、最高價、最低價、收盤價、交易股數、交易量、公司規模、Alpha 值、Beta 值）、投資人資料（年齡區間、開戶年、開戶別、玉證交易經驗代碼、交易次數）、以及當次交易資料（買賣別、市場別、交易別、交易價格、交易股數、報酬率）。值得注意的是，針對類別資料的特徵，我以 one-hot vector 進行處理。共計 28 個預測特徵。第二、因應資料的不平衡：雖然過去文獻建議，針對不平衡的資料可以做 oversampling 或是 undersampling (García et al., 2018)。然而在 XGboost 相關文獻中，除了上述兩種作法外，亦可以針對稀少類別的資料(minority)做相對應比率的加權（權重等於主要類別樣本數除稀少類別樣本數）以因應不平衡的資料集。另外，由於 XGBoost 算法中會同時衡量驗證集，根據 Saito & Rehmsmeier (2015)建議，針對不平衡的資料集應以 aucpr 作為指標。有鑑於此，本取徑將採取加權與 aucpr 的因應措施。
- 資料集切分：由於本資料的違約資料稀少，故在切分訓練集與驗證集時，必須考量稀少類別在兩類資料集有類似的比率，故我以 8 比 2 的隨機分層抽樣，確保兩類資料集的隨機性及具有類似的分布。此外，在考量過大的資料量會戕害訓練的時間與硬體資源，故本研究共使用訓練集 700140 筆資料，而驗證集 175035 筆資料。



- 方法概述：XGBoost 是一種基於樹(tree-based)的前瞻機器學習模型，有鑑於其在 Kaggle 和 KDDCup 等無數比賽中能夠達到前所未有的預測效能，因此受到學術界和務實界的廣泛關注(Chen & Guestrin, 2016)。因此，我們應用 XGBoost 來解決當前的任務。XGBoost 大體的流程是：首先，先建構一棵初始的決策樹，其次，然後遍歷所有樣本找出分類不正確的部分（殘差），爾後，針對再建構一棵決策樹並且與先前建構好的決策樹合併，最後，重複上述步驟直到無法再建構為止。
- 實驗：超參數搜尋策略：貝氏優化與早停技術：貝氏優化這是一種基於(1)目標函數的先驗分佈和(2)不同模型之驗證效能序列的損失函數的有效超參數調整模式，用於選擇最佳模型(Bobak Shahriari & de Freitas, 2015)。本實驗將據此在各類有關 XGboost 的參數上進行搜尋：
  1. 在一般參數的部分：eta 定義了每次迭代時的步長，同時朝著損失函數的最小值移動，其搜索範圍  $\in \{0.1, 0.01\}$ 。
  2. 在特定於樹的參數部分：min\_child\_weight 定義子節點所需的所有觀察的最小權重總和，其搜索範圍  $\in \{20, 21, \dots, 30\}$ 。max\_depth 定義了樹的最大深度，其搜索範圍  $\in \{30, 31, \dots, 50\}$ 。gamma 指定進行拆分所需的最小損失減少，其搜索範圍  $\in \{0.1, 0.2, 0.3\}$ 。subsample 表示每棵樹的隨機樣本觀察值的比例，其搜索範圍  $\in \{0.6, 0.7, \dots, 1\}$ 。colsample\_bytree 表示每棵樹隨機採樣的列的比例，其搜索範圍  $\in \{0.5, 0.8, \dots, 0.9\}$ 。
  3. 正則化參數。lambda 在權重上定義 L2 正則化項，其搜索範圍  $\in \{100, 10E-1, \dots, 10E-6\}$ 。alpha 定義了權重的 L1 正則化項，設置為 0.1。

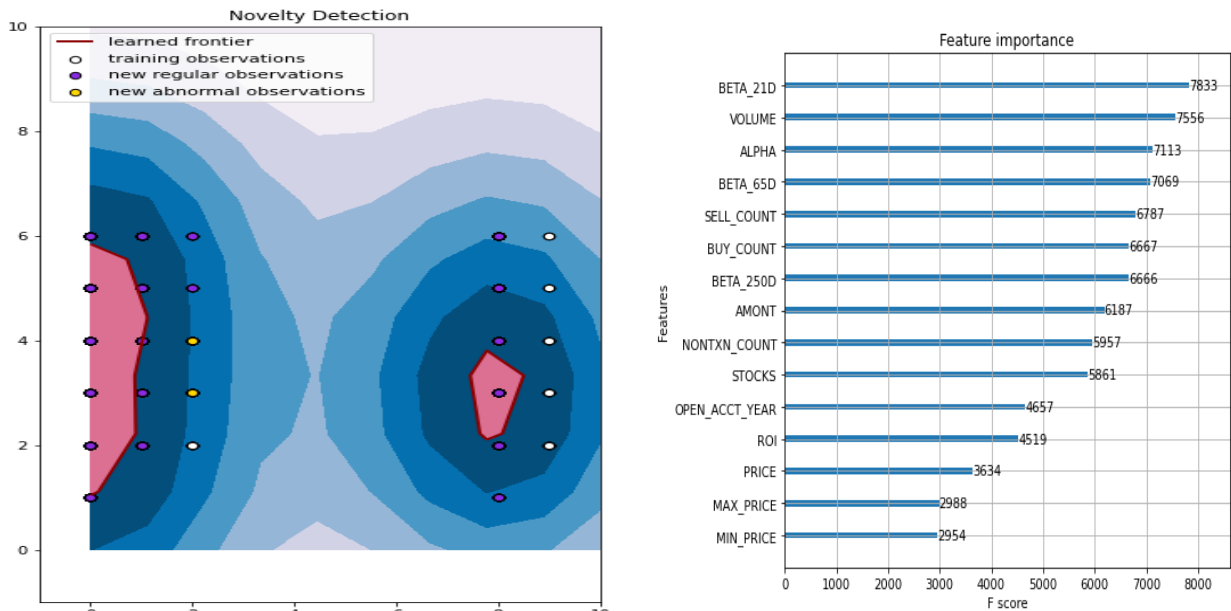
在訓練 XGBoost 模型時，早停技術是一種常見的避免過度擬合的策略(Zhang & Yu, 2005)。具體來說，early\_stopping\_rounds 指定驗證錯誤沒有顯著減少的最大回數。透過貝葉斯優化選擇最佳模型後，我們在最佳模型上考慮並試驗了四個 early\_stopping\_rounds 候選值  $\in \{1, 2, 5, 10\}$ 。
- 實驗環境：硬體資源、CPU：Intel(R) Xeon(R) CPU @ 2.30GHz；RAM：12GB。編程環境、Python 3.7.10。套件、sklearn，xgboost，hyperopt。

### 3.4 實驗結果及討論

#### 取徑一：One Class SVM 實驗結果

下左圖的 X 軸為投資種類，Y 軸為年齡分布，紫色的點為正常的 testing data，黃色的點為異常的 testing data，而紅色線框住的粉紅色區域則為我們的 model 找出來的正常資料的範圍，而藍色顏色越深的部分則代表他是異常的案例中比較接近正常的。

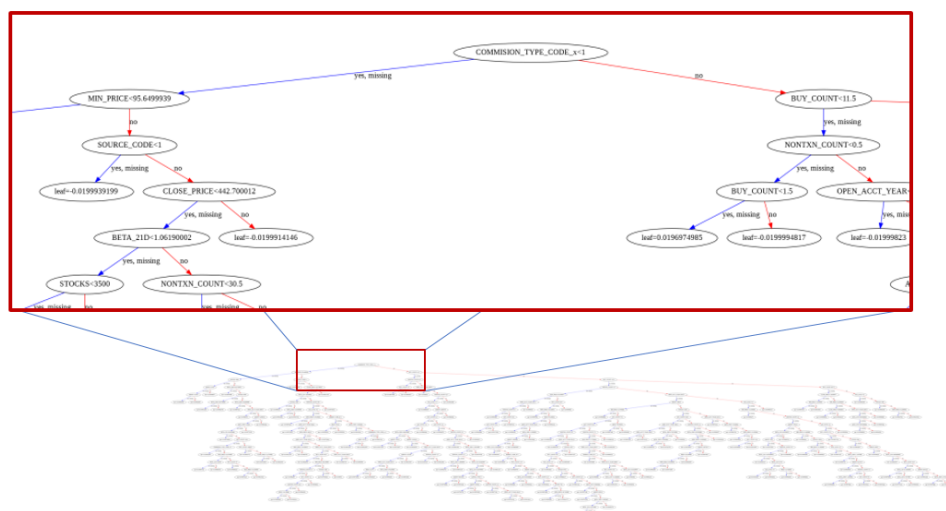
我們可以看到在投資種類為 0(現股交易)時，幾乎所有的資料都會被認為是正常的，而當交易種類為 8(現股當沖)時，只有年齡為 30~39 這個區間的人會被視為安全的，而其他的資料都被認為有違約風險。若以正確率來分，正常的資料約有 85% 會被視為是正常的，而異常的資料只有 27% 會被視為是異常的。



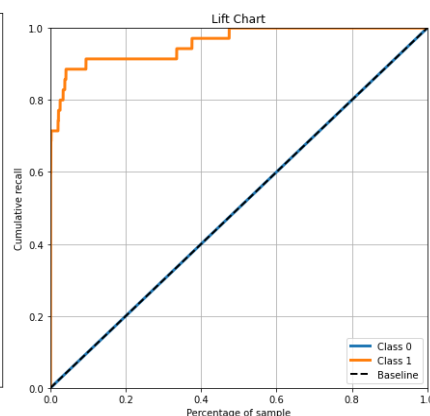
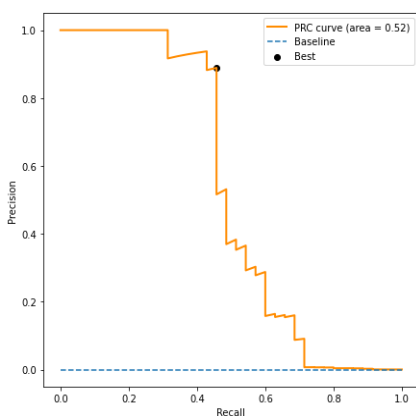
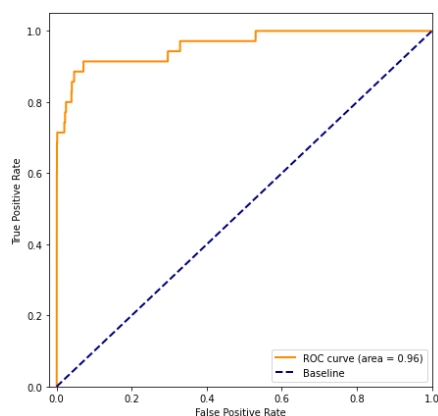
#### 取徑二：Extreme gradient boosting (XGBoost)

- 最佳實驗參數值：根據貝氏優化與早停實驗的結果，我們可以得出的最佳預測模型在驗證集上的預測力為 99.98%，其對應的最佳的參數為：'max\_depth': 20, 'n\_estimators': 1000, 'subsample': 1, 'colsample\_bytree': 0.7, 'learning\_rate': 0.01, 'reg\_alpha': 0.1, 'reg\_lambda': 1e-05, 'min\_child\_weight': 23, 'gamma': 0.2, 'early\_stopping': 10。
- 特徵重要性：我們可以由決策樹上結點的分布可以了解到有哪些特徵對於預測違約交割是非常重要的，如上右圖所展示前十五大重要的特徵，我們可以發現當日的交易量和投資者長期的交易量和經驗都是重要的預測因子。

- Xgboost 決策樹結果：



- 分類錯誤的案例：仔細檢查分類錯誤的案例發現，當交易類型不是現股交易的資料常常分錯，推論其原因可能是因為不是現股（如借券、當沖、現股當沖）的資料較少 (<10000)。舉例來說，由上圖的決策樹可以得知，非現股交易為根節點的左子樹，而現股交易為根節點的右子樹，左子樹相對於右子樹的節點是相對稀少的，因此是個不平衡的決策樹。這類不平衡的結果，可以歸咎於非現股的資料量大幅少於現股的資料量，從而導致非現股的資料在此部分有較大的誤差，因而導致分類錯誤。一個可能的解決辦法是，增加非現股的資料量。
- ROC 與 PRC：本方法在驗證集上，依照不同閾值所繪製的 ROC 如下圖一所示，可以發現其 AUC 高達 0.96。儘管如此，有鑑於本資料集特性極度不平衡，過去研究發現 ROC 的表現會過於樂觀，建議應以 PRC 及其 AUC 檢視預測效能較為有意義(Saito & Rehmsmeier, 2015)。據此，本研究繪製 PRC 及其 AUC 如下圖二，依照 F1（同時考量 recall 與 precision）找出最佳的閾值。結果為：最佳閾值為 0.95，precision 為 0.88，recall 為 0.42，F1 為 0.57。



- Lift Chart：依照上圖三 Lift Chart 的結果，我們僅需抓出前百分之十的預測機率最高的樣本即可獲得超過 80% 的 recall。

## 討論：方法之間的比較

就本實驗的結果來看，我們可以發現 XGBoost 在本資料集以及本作業上的表現相對於 One-Class SVM 的結果略勝一籌，無論在 recall、precision、以及 F1 相關的預測效能指標上皆有相對比較的好表現，在訓練時間與資源上 XGBoost 也表現較佳。然而，整體而言，本專案的效能依舊不足以令人信服，特別在 recall 上，因為本作業是個一旦發生後果就較為嚴重的事件。建議實務者可以犧牲 precision 以提高 recall 的比率。

	Recall	Precision	F1	訓練資源 同資料量 /同特徵
One-Class SVM	27%	39%	33%	5hr
XGboost	42%	88%	57%	0.5hr

## 第四部分 結論及展望












- 對於證券公司而言，本專案提供兩種人工智慧預測機制
  - 有鑑於敏感度不高，未來可以蒐集更多違約交割的資料，以便讓機器更了解違約交割的資料型態。
  - 未來可以使用更多樣化的模型，由於本專案的資料是時間序列且巢套資料，或許可以使用考量到隨機效果的統計模型或是時間序列相關的機器學習模式
  - 未來可以增加更多變項作為預測變項，例如使用者的風險性格或是情緒穩定性跟人類決策相關的心理指標。
  - 未來可以考量更多樣化的投資風險類型。
- 對於一般投資人言而，本專案提供了一個即時的交易風險評估程式
  - 未來可以試圖發展 Web 或 Mobile App 版本的應用。

## 第五部分 附錄

### 5.1 參考文獻

- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 11). Berlin: Springer.
- Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (1999). Estimating the support of a high-dimensional distribution. *Technical Report MSR-T R-99-87, Microsoft Research (MSR)*.
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), e0118432.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1), 148-175.
- Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4), 1538-1579.

### 5.2 組員執掌

徐蘇緯	劉羽忻	駱皓正
 使用者介面程式		 資料集與資料前處理
 即時風險評估	 資料集與資料前處理	 XGBoost 方法與結果
 短片製作	 OCSVM 方法與結果	 方法比較與結論展望
 期末報告撰寫	 長片製作	 長片製作
 長片短片合併及上字幕	 期末報告撰寫	 期末報告撰寫
 開會紀錄		 彙整原始碼與定稿
 檔案上傳與教務聯繫		

### 5.3 資料、紀錄及相關連結

1. 原始碼：<https://github.com/hc-psy/fintech-ntu>
2. 小組開會紀錄：<https://hackmd.io/X3o3TH7STuqYkpzq5N8o1A>
3. 短片連結：<https://www.youtube.com/watch?v=W6A30BSuzHc>
4. 長片連結：<https://www.youtube.com/watch?v=mr7RZNAoTKs>