中 央 大 學

 機 器 學 習 期 末 專 題

第 三 組

使用機器學習偵測洗錢行為

Money Laundering Forecast using Machine Learning

學生：張凱東111522071、陳丕中 111527002、劉洮語 111525025

中華民國111年12月04日

1. **Motivation and Goal**

對於金融業而言，洗錢防制是一個必然面對的難題與挑戰。犯罪者利用各種洗錢管道將非法資金漂白，以逃避司法機關的查緝與追訴，甚至將該犯罪所得再次利用於其他不法行為當中。金融機構若不積極審查各種由其經手的交易行為，則將淪為犯罪集團的漂白管道，除了損及自身商譽之外，又因為金融機構具有集合廣大人民資金的特性，而擾亂了金融市場秩序。除此之外，金融犯罪者不斷以新興科技或渠道從事不法所得的掩飾或隱匿行為，洗錢態樣推陳出新，致使金融業單靠人力顯然不足以辨識出可疑的犯罪活動。

在現代金融系統的飛速發展和複雜的金融交易，已經導致各類詐騙和洗錢活動更加嚴重。為了有效防止洗錢活動，政府和銀行們正在尋求更有效的技術，以便檢測和監控可疑交易。機器學習技術可以提供一種更有效的方式，通過採用大數據技術和複雜的分析算法，可以更加有效地檢測和檢查可疑的洗錢活動。另外，機器學習技術可以提供更多的精準度和穩定性，更好地支持銀行對可疑交易的檢查和監控。因此，我們希望能透過機器學習的方式實作出疑似洗錢交易預測模型，減少人力成本同時增加預測效率。

1. **Related Work / Market Survey**

我們總結了洗錢防制預測領域的現有工作，洗錢防制開始於1970年，當時金融機構開始向政府報告大額交易。一些統計方法像Bayesian models跟時間序列的匹配載1990年後期被用於檢測洗錢活動，而在2004年之後，機器學習方法開始應用時洗錢預測的模式裡。

我們這次的研究提出了使用機器學習技術，來偵測洗錢活動的可行性。在根據上述對市場上的商業解決方案進行了分析後，我們發現它們集中在金融業務處理，並且在偵測洗錢活動方面存在許多局限性，例如低準確度和無法檢測新的洗錢模式。為了解決上述問題，我們提出使用機器學習技術來偵測洗錢行為，深度學習技術可以比其他技術更有效地檢測洗錢活動，並且可以更快速地偵測新的洗錢模式。

1. **Target Customers**

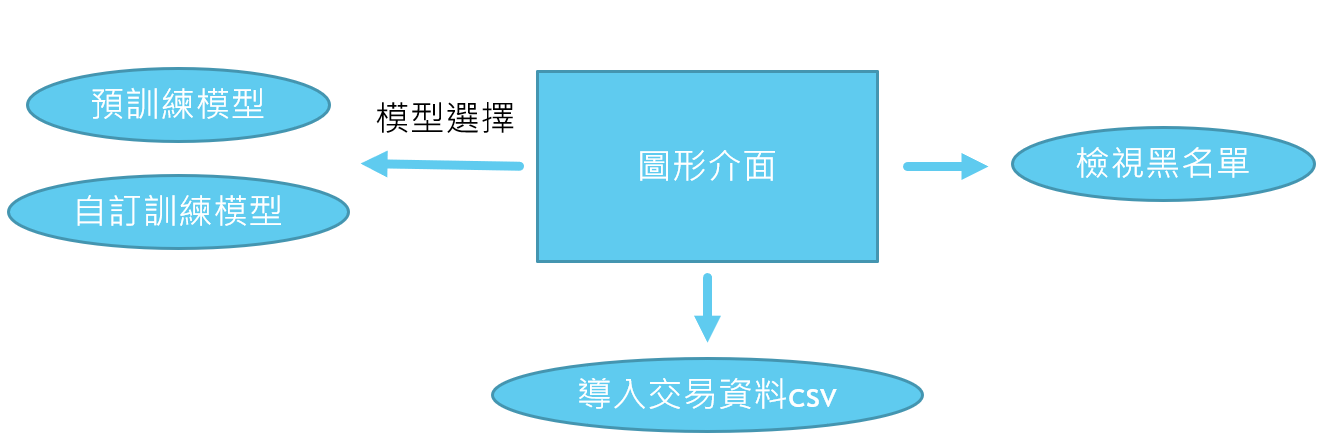
隨著新一代投資產品的普及，以及COVID-19所造成的遠距、無接觸金融服務持續增加，使得金融機構的洗錢預測能力日漸增強。為此我們研發的這一個模型將可以用來提供給金融機構，使其可以更有效率的檢測洗錢的行為，並提高檢測的精準度。此外，我們的模型也可以用於檢測加密貨幣及虛擬資產等新一代的投資新寵，從而為金融機構帶來更多的安全保證。

1. **Explanation of Product Features**

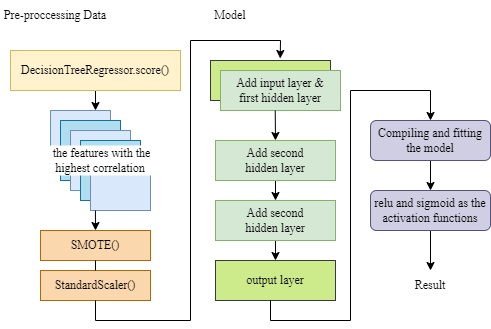
洗錢防制是一個高度勞力密集的產業。在沒有相關系統的輔助下，一些較大的銀行可能每天面臨數以萬計的洗錢相關案例，導致時間及人力成本極其高昂。系統篩檢後，仍需要客服人員實際了解狀況，特別是一些網路交易中每天會有多筆收入進帳的情況，多次下來的誤會可能會造成客戶的困擾。要避免誤會，就需要仰賴機器學習來達到更高準確率的洗錢預測，以大幅減少人力及時間成本。

1. **The Solution Architecture**

**5.1 初步架構：**

我們預計使用圖形化介面設計，讓使用者可以簡單上手，透過導入交易資料的csv檔，進而判斷哪幾筆交易是疑似洗錢，並提供黑名單紀錄。

**5.2 詳細程式架構：**



**5.2.1 資料前處理(Data Preprocessing) :**

在計畫資料前處理的程序，以確保模型的準確性和優化使用效率。

1. 由於確定係數`R^2`介於0至1之間，1代表完美擬合，負`R^2`意味著模型無法擬合該數據，因此，在決定特徵相關性時，會使用DecisionTreeRegressor.score()來取得特徵與目標變量之間的相關性，並使用最高相關性的特徵來訓練模型。

2. 由於數據不均衡，因此，將使用SMOTE()來平衡數據。

3. 將使用StandardScaler()來對所有特徵進行平均和變異數標準化，以將數據標準化為高斯分佈，使其平均值為0、變異數為1，以加快模型收斂。

**5.2.2 訓練架構(Training process) :**

這次project我們將使用一個名為Keras 的deep learning library，來將每次的交易都分類為fraudulent(1) and non-fraudulent(0)。

1. Adding the input layer and the first hidden layer.
2. Adding the second hidden layer
3. Adding the output layer
4. Compiling and fitting the model
5. Plot the training curves so that we can see how well the training process go
6. As the output of the model is binary (0 or 1), we have used. They are considered to be more effective to deal with binary classification problems. In the graphs it can be seen that the model is learning at every epoch and minimizing the loss. These graphs make it easier to choose the number of epochs as after a certain level the loss will start moving up again. It is also used to tackle over fitting while training the models.