BI Final Project

客戶對銀行電話行銷的反應預測

資管碩一 黃翌

資管碩一 徐亦華

背景與動機

由於新冠肺炎疫情的影響，許多人轉為在家工作，或是因政策因素需要長時間在家中，無法任意出門，有些人則基於防疫的要求，盡量待在家中，減少出門頻率，在這樣的情況下，銀行和金融機構要向客戶進行服務的銷售時，便需要更加依賴電話、電子郵件等不用直接接觸的管道進行行銷，才能增加營銷機會。

然而，對所有人進行電話行銷顯然是不太實際的做法，銀行應該找出較有可能接受電話行銷的人群再進行推銷，才是較有效率的做法。不過，要如何識別那些有可能訂閱其產品、優惠和其他服務的客戶對於銀行及金融機構來說將是個挑戰，因此，我們在此Project中，想要嘗試透過使用幾種分類方法，來將客戶的資料進行分類，並辨別出有較高機率訂閱銀行服務的客戶，藉以預測客戶訂閱服務的機率，以做為銀行及金融機構的決策輔助，以提高行銷的效率，節省銀行人力與時間成本。

文獻回顧

許多研究人員和數據挖掘人員對通過電話進行營銷活動的銀行直接營銷現象進行了幾項研究。廣泛探索了決策支持系統以及其他數據驅動的方法。在Asare-Frempong, J., & Jayabalan, M. (2017, September)的研究中提到，直接營銷使銀行和其他金融機構能夠專注於那些有可能訂閱其產品、優惠和其他服務的客戶。但在大多數情況下，識別這些客戶群對金融機構來說是個挑戰。此研究考慮了具有兩個主要目標的銀行直銷活動數據集的典型案例。並通過應用四種分類方法來預測客戶對銀行直銷的反應，方法為以下：多層感知器神經網絡（MLPNN）、決策樹（C4.5）、邏輯回歸和隨機森林（RF）。研究結果，準確率為 87% 的隨機森林是準確率最高的分類方法。當使用分類準確度和ROC對分類方法進行評估，隨機森林的準確率分別為86.80%和92.7%，也是表現最佳的。第二個目標是進行聚類分析以確定訂閱並最有可能隨後訂閱定期存款的客戶的主要特徵。結果顯示，通話時間越長的客戶訂閱定期存款的可能性越高。結果進一步表明，具有最低中學教育程度的客戶是銀行定期存款認購的良好目標。

在Alexandra, J., & Sinaga, K. P. (2021, October)的研究中指出銀行可以利用有助於決策的信息技術空間進行市場分析。通過分析銀行營銷數據，可以用來選擇營銷類型。營銷活動可以通過電子郵件、電話和直接電子郵件發送給潛在客戶，讓潛在客戶決定是否接受所提供的產品。隨著時間的增加，傳入的數據量繼續增長。隨著數據的增加，其中一家銀行機構發現很難預測他們的客戶是否會訂閱定期存款。此研究透過數據挖掘過程使用分類（決策樹、貝氏分類器和隨機森林）和聚類（K-Means、K-Medoids 和 DBSCAN）方法來預測客戶是否會訂閱一個定期存款。

根據以上研究，我們發現可以使用決策樹、隨機森林與單純貝氏分類器等分類方法來進行我們的Project，並且這些方法在其研究中皆有不錯的表現。

研究方法

本Project將使用決策樹(Decision Tree)、隨機森林(Random Forest)以及貝氏分類器(Bayes classifier)這些分類方法，來對客戶資料及進行分類與預測。

決策樹(Decision Tree)是一個預測模型；他代表的是對象屬性與對象值之間的一種映射關係。樹中每個節點表示某個對象，而每個分叉路徑則代表某個可能的屬性值，而每個葉節點則對應從根節點到該葉節點所經歷的路徑所表示的對象的值。決策樹僅有單一輸出，若欲有複數輸出，可以建立獨立的決策樹以處理不同輸出。 數據挖掘中決策樹是一種經常要用到的技術，可以用於分析數據，同樣也可以用來作預測。決策樹演算法可以使用不同的指標來評估分枝的好壞，常見的決策亂度評估指標有Information gain、Gain ratio、Gini index。我們目標是從訓練資料中找出一套決策規則，讓每一個決策能夠使訊息增益最大化。Information Gain其算法主要是計算熵，因此經由決策樹分割後的資訊量要越小越好。而 Gini 的數值越大代表序列中的資料越亂，數值皆為 0~1 之間，其中 0 代表該特徵在序列中是完美的分類。常見的資訊量評估方法有兩種：資訊獲利（Information Gain）以及 Gini 不純度（Gini Impurity）。

$$Entropy = -\sum_j p_j \log_2 p_j \qquad Gini = 1 - \sum_j p_j^2$$

隨機森林(Random Forest)是一個包含多個決策樹的分類器，並且其輸出的類別是由個別樹輸出的類別的眾數而定。要形成多顆具差異性的樹以進行Ensemble Method，就要產生不同的數據集，才能產生多顆具差異性的決策樹，其作法有兩種方式：

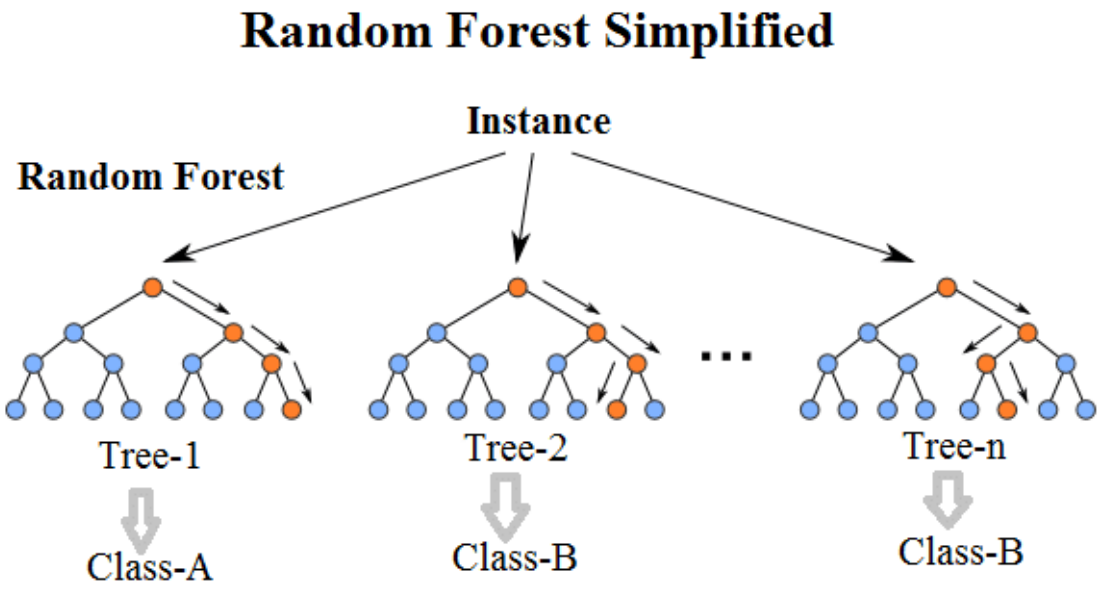
1.Bagging(Bootstrap Aggregation):

Bootstrap指的是「重新取樣原有資料產生新的資料，取樣的過程是均勻且可以重複取樣的」，使用Bootstrap我們就可以從一組資料中生出多組資料集。

此種方法會從訓練資料集中取出K個樣本，再從這K個樣本訓練出K個分類器（在此為樹）。每次取出的K個樣本皆會再放回母體，因此這個K個樣本之間會有部份資料重複，不過由於每顆樹的樣本還是不同，因此訓練出的分類器（樹）之間是具有差異性的，而每個分類器的權重一致最後用投票方式(Majority vote)得到最終結果。

2. Boosting:

與Bagging類似，但更強調對錯誤部份加強學習以提升整體的效率。是透過將舊分類器的錯誤資料權重提高，加重對錯誤部分的練習，訓練出新的分類器，這樣新的分類器就會學習到錯誤分類資料(misclassified data)的特性，進而提升分類結果。



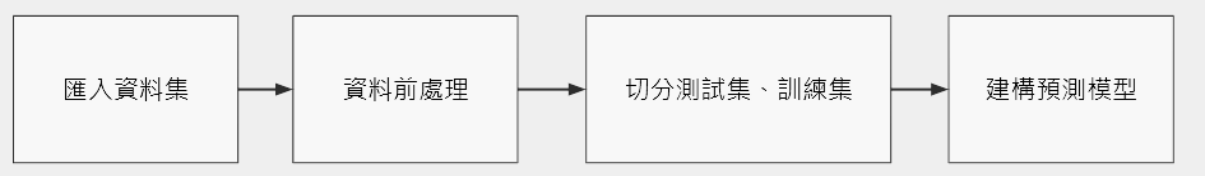
▲隨機森林概念圖

貝式分類器（Bayesian Classifier）是一種基於機率模型的機器學習模型。其根據貝氏定理（Bayes’ theorem）為基礎，透過機率統計來判斷未知的資料類別。貝氏定理描述在一些已知的一些條件下，某件事發生的機率。公式如下：

$$P(A|B)=\frac{P(B|A)*P(A)}{P(B)}$$

貝式分類器就是一個機率模型分類器。因此所有的模型參數都可以通過訓練集的相關頻率來估計。

實驗細節



▲實驗架構圖

此Project所使用的資料為UCI上的Bank Marketing Data Set，來源為Moro, S., Cortez, P., & Rita, P. (2014)的研究，內容為一家葡萄牙銀行機構的電話直接營銷活動相關的資訊，其中總共有41188筆資料，並有17個屬性。

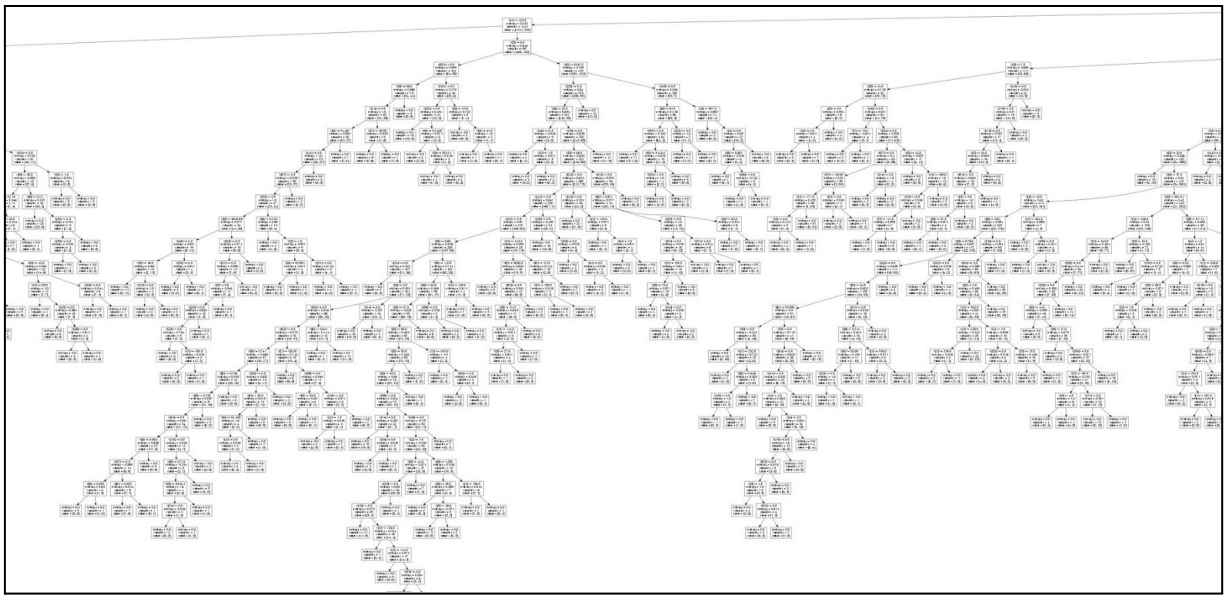
首先是資料前處理，將類別屬性(如job、marital、education等)，透過 One-Hot Encoding 方法轉為虛擬變數。

將欄位轉換完成後，將整體資料集切分成訓練集、測試集。本實驗採取二八切分，將兩成的資料選為測試集；八成的資料選為訓練集。

模型設定部分，決策樹與隨機森林，都採用 Entropy（熵）作為決策樹建構的算法；貝式分類器則採用具代表性的高斯分類器。

實驗結果中，三種模型準確度如下:

- 隨機森林擁有最高的準確度，高達 91.15%
- 單決策樹的準確度為 88.91%
- 貝式分類器的準確度為 85.90%



▲部分的視覺化決策樹

討論

我們使用決策樹、隨機森林以及貝氏分類器對資料集進行訓練與預測，得到的準確度依序為，決策樹 88.91%、隨機森林 91.15%、 貝式分類器 85.90%。隨機森林的準確度最高，推測原因為，當客戶資料集裡擁有許多離散欄位時，決策樹等針對離散資料為主的分類法會表現較佳，而隨機森林為決策樹的改良版，因此準確度會更好。

銀行過往進行電話行銷時，需要花費大量時間與人力一一撥打電話，而透過使用這些非類預測方法，銀行可以對客戶資料進行預測，並挑選預測結果為成功的客戶進行電話推銷，以節省人力與時間成本。

參考文獻

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Asare-Frempong, J., & Jayabalan, M. (2017, September). Predicting customer response to bank direct telemarketing campaign. In 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T) (pp. 1-4). IEEE.

Alexandra, J., & Sinaga, K. P. (2021, October). Machine Learning Approaches for Marketing Campaign in Portuguese Banks. In 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-6). IEEE.