T Brain AI實戰吧 玉山人工智慧公開挑戰賽 - 金融商品交易預測

分析流程說明

1. 資料清理與預處理

這次比賽提供的資料區分成八份資料，故我們將所有資料進行個別處理後才丟入模型，以下以條列式的方式說明處理流程。另外此次處理皆以Ｒ語言作為主要分析工具。

1. 資料的期間為過去120天，我們透過程式將期間區分為Ｍ1、M2、M3 與M4（ M4為最近期的30天），並將Ｍ4是否購買金融商品作為模型中的預測目標建立模型，另先以Ｍ1~M3 作為預測時使用之資料。
2. 比賽中有一份資料為顧客曾經瀏覽之網址，我們透過程式找出使用率較高的七種網址，並將顧客在M1~M3瀏覽這七種網址的次數作為我們的解釋變數丟入模型。
3. 在TBN\_CIF資料中，提供了顧客的基本屬性資料，然而基本屬性資料可能因為顧客不願填寫而產生缺失值，若資料型態為數值形態，我們採用的是以平均數填補的方法，若資料型態為類別變數，我們則以眾數填補，而非直接刪除有缺失值的資料，以保證資料的完整性。
4. 在借貸的資料中，有一項變數為借貸用途，然而顧客可能在同一個期間借貸了兩次，因而產生不一樣的借貸用途，在此我們先挑選出有在同一期間借貸數次的顧客，並以最後一次借貸行為之用途作為此一名顧客的借貸用途。
5. 借貸與信託資料中，皆有一個交易金額之變數，同樣也會遇到在同一期間發生數次借貸或購買信託之情況，因此我們將在同一期間重複之顧客挑出，並以平均數作為這些顧客之交易金額。
6. 最後以CUST\_NO顧客代碼作為ＫＥＹ，將上述資料合併在一起，完成資料清理與預處理的步驟。
7. 建立模型（以決策樹為主要建模策略）

在建立模型的過程中，我們發現目標變數（ＥＸ：是否在M4借貸）有資料不平衡之問題，若是直接建立模型並預測，則預測出來的結果會是全部顧客預測為不進行借貸，因此在建模之前必須要處理此一問題。

最後我們透過Synthetic Minority Over-Sampling Technique (SMOTE)，對資料進行修正，並以決策樹作為我們主要的預測模型，模型建立完成後，為避免overfitting問題，我們亦對模型進行剪枝。

1. 改進策略

雖然我們最終並沒有獲得頂尖的排名，但經由組內檢討以及前六名之經驗分享，我們仍總結了以下幾點，並期望能以此經驗為底並用以精進自身。

1. 將資料處理的過程導入相關係數分析，以此排除預測效果不利之變數。
2. 解析瀏覽網址的過程中，考量時間趨勢性，增強模型預測能力。
3. 利用關聯式規則分析商品間的關係。
4. 利用Ensemble 的方式結合多種預測模型來提升整體預測能力。
5. 最終比賽結果

在這次比賽中，我們最後獲得了第153名（共1121隊）。透過這次比賽，我們不僅更加熟悉資料清洗與預處理的技巧，也研究了諸如決策樹、ＳＶＭ、隨機森林等預測模型之演算規則以及模型修正技巧，雖然並沒有在比賽中拔得頭籌，但我相信，透過這次經驗，我們在資料分析能力上已有一定的提升。