Gradient Boosting (梯度提升) 概念簡介及應用解析

編撰：屏東大學 周國華老師 (與 Google Gemini 共筆) 2025/11/24

Gradient Boosting 是一個非常強大的算法，通常是數據競賽（如 Kaggle）中的常勝軍，因為它在準確度上表現極佳。以下分兩個部分說明：首先是概念解析，接著是一個具體的審計/稅務操作範例。

---

1. 什麼是 Gradient Boosting (梯度提升)？

簡單來說，Gradient Boosting 是一種「三個臭皮匠，勝過一個諸葛亮」的策略，但在這裡，每一個新的「臭皮匠」都是為了修正前一個人的錯誤而生的。

核心運作邏輯

它是 **Ensemble Learning (集成學習:** 是將不同的演算法結合起來，最後得出一個效果更好模型的方法。**)** 的一種。不同於 Random Forest (隨機森林) 是平行建立多棵樹來投票，Gradient Boosting 是**「序列式」(Sequential)** 地建立模型。

1. **第一步**：建立一個簡單的模型（通常是決策樹），做初步預測。
2. **第二步**：計算第一步的**殘差 (Residuals)**，也就是「預測值」與「真實值」之間的差距（錯誤的部分）。
3. **第三步**：建立第二個模型，**專門用來預測這些「錯誤」**，而不是預測原本的目標。
4. **第四步**：將修正後的結果加總。
5. **重複**：不斷重複這個過程，每一次的新模型都在修正前一次的不足，直到誤差降到最低。

💡 審計直觀比喻

想像你在查核一家公司的帳目：

- **初級審計員 A (第一棵樹)** 檢查了一遍，發現了 70% 的明顯錯誤。
- **中級審計員 B (第二棵樹)** 不會重頭看，而是**專注於檢查 A 沒抓出來的那些複雜錯誤**。
- **高級經理 C (第三棵樹)** 再進一步，專注於 **B 還是沒看懂的極端案例**。
- **最終結果**：結合 A + B + C 的發現，形成一份極高準確度的查核報告。

---

2. Orange 操作範例：企業逃漏稅風險預測

我們將以「稅務查核 (Tax Audit)」**為場景。假設你是稅務局或事務所的查核人員，你想預測哪些公司具有**「高逃漏稅風險」。

準備工作

假設你有一個 Excel/CSV 檔案 (Tax_Risk_TrainingData.csv)，包含以下欄位：

- **特徵 (Features/Inputs):**
  - Revenue (營業收入)
  - Gross_Margin (毛利率)
  - Effective_Tax_Rate (有效稅率)
  - Cash_Transactions_Ratio (現金交易比例 - 這是高風險指標)
  - Industry (行業別)
- **目標 (Target):**
  - Audit_Result (過去查核結果：0=正常, 1=有逃漏稅)

Orange 操作步驟 (Step-by-Step)

請依照以下流程在 Orange 畫布上拖拉 Widget (元件)：

第一步：載入與設定數據

1. **File**: 點選並載入你的 Tax_Risk_TrainingData.csv。
2. **Select Columns**: 連接到 File。
   - 將 Audit_Result 拖拉到 **Target** 欄位。
   - 將其餘財務指標拖拉到 **Features** 欄位。No.放到 Meta 欄位。

第二步：建立模型 (使用 Orange 的 Gradient Boosting 模型)

1. 在左側選單找到 **Model** 區塊。
2. 拖拉 **Gradient Boosting** 元件到畫布上。
3. **設定參數 (雙擊元件)**：
   - *Number of trees (樹的數量)*: 預設通常是 100。越多通常越準，但計算越久且可能過擬合 (Overfitting)。建議設為 100-500。
   - *Learning rate (學習率)*: 控制每棵樹修正錯誤的幅度。通常設小一點 (如 0.1 或 0.05) 效果較好，雖然跑得慢但更穩健。
   - *Max depth (樹的深度)*: Gradient Boosting 通常使用淺樹 (例如深度 3-5)，不像隨機森林用深樹。

第三步：訓練與評估

1. 拖拉 **Test & Score** 元件。
2. 將 **Select Columns** 連線到 **Test & Score** (作為 Data)。

3. 將 **Gradient Boosting** 連線到 **Test & Score** (作為 Learner)。
4. (選做) 為了比較，你可以同時拖拉一個 **Logistic Regression** 或 **Tree** 也連到 Test & Score。

第四步：解讀結果

1. 雙擊 **Test & Score** 查看評分：
    - **AUC**: 這是最重要的指標。如果 AUC > 0.85，代表模型區分「正常」與「逃漏稅」的能力很強。
    - **F1-Score**: 在審計中，我們更在意「沒抓到壞人」(False Negative)，所以 F1 分數很有參考價值。

第五步：找出關鍵風險因子 (審計最重要的一環)

Gradient Boosting 是**黑盒模型**(亦即：準確率很高，但難以解釋原因，容易變成「電腦說了算」)，但在審計中我們需要知道「為什麼這家公司被標記為高風險？」。

1. 在 Model 區塊下方找到 **Explain** 或 **Rank** 相關的元件（Orange 較新版本有 Feature Importance 或可透過 Rank 查看）。
2. 連接 Data 到 **Rank**，選擇 Gradient Boosting 作為評分依據。
3. **結果解讀**：系統可能會告訴你，影響預測結果最大的因子是 Cash_Transactions_Ratio (現金交易比例) 和 Effective_Tax_Rate (有效稅率)。這能直接指導審計人員：「先去查那些現金交易多且稅率異常低的公司」。

---

3. 為什麼這對財務/審計很有用？

特性	對財務/審計的優勢
處理非線性關係	財務比率與風險通常不是直線關係（例如：毛利率太低有風險，太高也有造假風險）。Gradient Boosting 能完美捕捉這種複雜曲線。
高準確度	在詐欺偵測 (Fraud Detection) 這類「大海撈針」的任務中，它比傳統的 Logistic Regression 更能抓出微小的異常模式。
處理缺失值	財務數據常有缺漏，Gradient Boosting 內建處理缺失值的機制，不需要太繁瑣的清理。

總結

Gradient Boosting 就像是一個「極致追求完美的審計團隊」，透過不斷修正前人的錯誤來達到最高的查核準確率。在 Orange 中，你只需要簡單的三個元件連線 (File -> Test & Score <- Gradient Boosting) 就能完成強大的風險預測模型。