**Fundamentals of Data Analytics HW3-2 Report**

**F74062028 歐子杰**

**Choose a dataset**

* **選擇 Online Shoppers Purschasing**

所有的特徵欄位都有先命名比較好處理

**Analyze the data (statistics, correlation…)**

* **所有的特徵都沒有缺失值需要處理**
* **Administrative\_Duration、Informational\_Duration、ProductRelated\_Duration 的數值分佈範圍較廣，推測對結果較有影響**
* **Month, VisitorType 為 String，需做 label encoder**

**Data Preprocessing**

* **Month, VisitorType 做 Label Encoder**
* **其他項目皆無需preprocessing**

**Define a reasonable problem**

* **我認為 Revenue 適合當作 y，表示是否獲利**
* **以其他特徵作為影響的因素 x**
* **問題的意義：哪些特徵對獲利最有影響**
* **Revenue 的值為 true or false，不需要再做 Label**

**Train model and predict, then try to Improvement Result**

* **決定仿照作業二的訓練方式及計算準確度的方式**

寫成function方便做improve後的比較

首先使用全部的資料進行訓練

* **Original Resault**

一張含有 瓶, 相片, 黑色, 標誌 的圖片

自動產生的描述

* **Improved Reasons and Approaches 1**

結果明顯overfitting，推測試使用的特徵太多，決定drop掉差異範圍較小的幾個特徵

* **Improved Results 1**

一張含有 相片, 標誌 的圖片

自動產生的描述

* **Improved Reasons and Appraoches 2**

依然overfitting，繼續drop，最後剩下範圍差異較大的幾個特徵

Administrative\_Duration、Informational\_Duration、ProducRelated\_Duration、PageValues

* **Improved Results 2**

一張含有 相片, 標誌 的圖片

自動產生的描述

* **Improved Reasons and Appraoches 3**

前面兩次嘗試對結果都沒有影響，依據前面作業經驗，可能是RandomForest的max\_depth、n\_estimators有影響

降低depth，減少n\_estimators

* **Improved Results 3**

一張含有 相片, 黑色, 白色, 標誌 的圖片

自動產生的描述

* **Final Result**

Improvement Results 3結果train valid誤差小於0.05沒有overfitting

**Discussion**

* **How did you preprocess this dataset ?**

這次的preprocess只有Month、VisitorsType的數值轉換

以及刪除沒必要的特徵

* **How did you improve your classifiers ?**

主要是調整RandomForest中tree的max\_depth，以及n\_estimators

避免過度的分類導致overfitting