Report

F74066527 洪邵澤

[Online Shoppers Purchasing Intention Dataset Data Set]

classification problem: 預測 Revenue 是 True / False

- 因為是 imbalance data，所以用 f1 score 與 AUC 來評斷 model 的好壞

資料分析與處理

- 首先確認資料並沒有缺失
- 藉由畫圖可以看出 Revenue 滿不平均的，False 占了大部分
- 因為 Revenue 的不平均，以致我決定要用 f1 score 與 AUC 去當評量標準
- 再來可以畫出各種數值資料的分布情形，也畫出各種類別資料與 Revenue 的分布關係
- 我覺得不太容易直接從圖上找出太大的關聯性
- 所以我做了相關係數的 heatmap，也挑出與 Revenue 較具有關係的 data
  - 選出 Administrative、ProductRelated、ProductRelated_Duration、BounceRates、ExitRates、PageValues、VisitorType
- 再來就是套用不同的 dataset 與 model，並藉由調整參數查看與推斷結果

Original

- 首先我先對 類別(string) 性質的資料做 labelencoder，才能夠拿來使用
- 然後把全部的資料餵進去給 model
- 就這樣得出了 baseline 的結果
- DecisionTreeClassifier
  - f1_score: 0.5890603085553997
  - auc: 0.7463339564568817
- LogisticRegression
  - f1_score: 0.6136363636363635
  - auc: 0.7578050945070715

improve 1

- 再來我認為數值資料的範圍大小會影響到不同的結果，所以我將數值資料全部做 scale
- 然後一樣把所有的資料 (scale 過後)餵進去 model
- 結果分數比 original 還低，因此我認為將數值 scale 過後丟進去並不是好的做法
- DecisionTreeClassifier
  - f1_score: 0.5882352941176471
  - auc: 0.7460941483034045
- LogisticRegression
  - f1_score: 0.533106960950764
  - auc: 0.693806529579486

improve 2

- 於是我認為應該從放進去的資料下手
- 藉由我上面畫出與印出的相關係數去找應該用的資料
- 於是我用 abs(correlation) > 0.1 的資料餵進去給 model
- 我認為這樣的資料是與答案 (Revenue) 比較具有相關性的
- 從結果可以看出不論是 f1 或是 auc 的效果都比一開始的 baseline 好
- DecisionTreeClassifier
  - f1_score: 0.6068027210884355
  - auc: 0.761236050529655
- LogisticRegression
  - f1_score: 0.6394366197183099
  - auc: 0.7734398308125153

improve 3

- 再來我覺得使用較具有相關性的資料下去訓練是可行的，於是我想要換個 model 試試看
- 我選擇使用 ensemble 的 model
- 因為 ensemble 的 model 可以有效應付更多樣化的 data 以及更好的避免 overfitting
- 從結果上來看，ensemble model 的效果的確是最好的

- RandomForestClassifier
  - f1_score: 0.6528803545051699
  - auc: 0.7720406352083687
- AdaBoostClassifier
  - f1_score: 0.6477272727272728
  - auc: 0.7764308238448612

model 參數調整

- DecisionTreeClassifier
  - max_depth 控制深度避免 overfitting
  - max_features 控制 feature 數避免 overfitting
  - criterion 嘗試 gini 跟 entropy 的差別
- LogisticRegression
  - solver 選用 lbfgs
  - max_iter 控制跌代數
- RandomForestClassifier
  - n_estimators 增加 estimators 增加可能性
  - max_depth 控制深度避免 overfitting
  - min_samples_leaf 改變最小的 leaf 劃分，讓 feature 畫在一起
  - criterion 改為使用 entropy
  - oob_score 讓 model 自行驗證
- AdaBoostClassifier
  - n_estimators 改變 estimators 增加可能性

Summary

由於是 imbalance data，我所有 model 的 false negtive 都比 false postive 多，一開始 model 沒調整好便會全部都猜 false

- 其中影響最大的 feature 是 Page Value，因為他與 Revenue 的相關性最高
  - 查到的解釋是 Page Value is the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both).
  - 此頁面的價值與他是否能產生收入有相當大的關聯
- 話說使用者果然很多都是看一看沒有買 (我自己好像也是 XD)，Revenue 的 True 只有 15% 真的有夠少