資料科學導論

撰寫者:

統計系 111 級 H24074019 黃科銓

Introduction

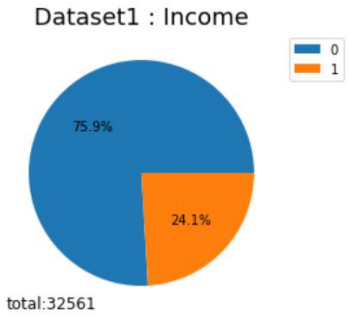
分類問題（Classification）一直是很熱門的預測議題，在金融商業、社群網路或是科技業等許多領域皆可以看到，而對於運用機器學習方法來預測問題，除了資料前處理及資料視覺化是不可或缺的步驟外，「如何評估模型好壞」也是一個重要的議題。因此此次報告書將特別針對「二元分類問題」為背景，希望透過五種不同資料集(不同樣本數、不同欄位數)，以整體而言來看，試圖普遍性評估八種機器學習，藉此來了解方法的好壞及適合情境的使用。

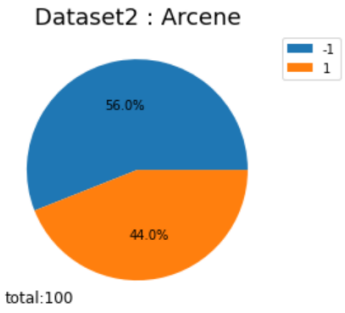
分類問題 # 機器學習 # 模型評估

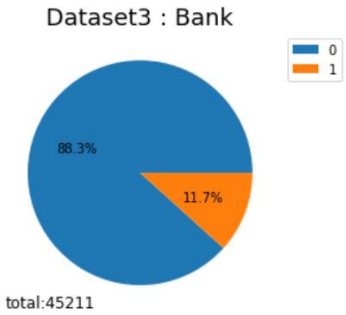
Methodology

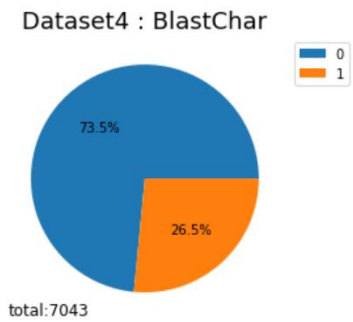
一、 資料說明

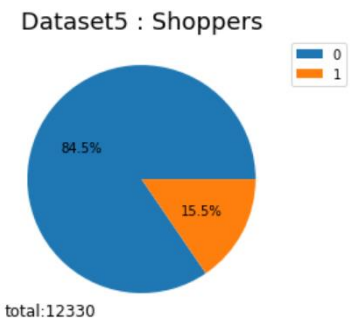
此次分析總共有六個資料集，分別為 Income、Arcene、Bank、BlastChar、Shopper、SHrutime，皆為二元分類資料，各資料 labels 比例如下：

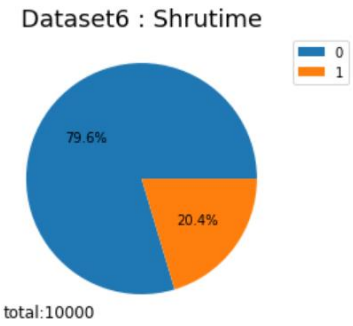












【小結論】

樣本數：Bank>Income>Shoppers>Shrutime>BlastChar>Arcene

資料平衡狀況(%)：Arcene>BlastChar>Income>SHrutime>Shoppers>Bank

二、 資料前處理

為了確保預測差異來自模型選擇，不管是哪一個資料集，皆用同一個方式

處理，而此次選擇使用 sklearn 的 Label encoding 方法，用於轉換非數值資料，而不是使用 One hot encoding 轉換的原因有兩個，第一個是為避免造成 column 數過多，以使模型跑不動，第二個是 Label encoding 便可以達到區分類別資料的效果，故此用 Label encoding。

	age	workclass
0	39	State-gov
1	50	Self-emp-not-inc
2	38	Private
3	53	Private
4	28	Private

	age	workclass
0	39	7
1	50	6
2	38	4
3	53	4
4	28	4

三、 模型調整

- KNN
  KNeighborsClassifier(n_neighbors=5)
- Logistic Regression
  LogisticRegression(solver="lbfgs")
- Random Forest
  RandomForestClassifier()
- MLP (2-layer NN)
  MLPClassifier()
- SVM
  svm.SVC()
- XGBoost
  XGBClassifier()
- CatBoost
  CatBoostClassifier(iterations=10,learning_rate=1,depth=2,loss_function='MultiClass')
- LightGBM
  lgb.LGBMClassifier(application='multiclass', boosting='gbdt', learning_rate=0.1, max_depth=-5, feature_fraction=0.5, random_state=42)

四、 自訂模型

本次模型將使用 vote 方式，基於在 supervised learning 的表現上，「Random Forest、XGBoost、CatBoost、LightGBM」都呈現不錯的預測，因此透過這四個模型進行投票，並預測為 1 的票數大於 2 票者，皆是為 1，否則皆為 0。

Random Forest	XGBoost	CatBoost	LightGBM	Total	Final

1	0	1	0	2	1

Experiments

一、supervised learning

資料切割：Train / Validation / Test = 65% / 10% / 25%

重複次數：5 次平均 Accuracy/AUC

	Income	Arcene	Bank	BlastChar	Shopper	SHrutime
KNN	0.773197	0.820	0.883889	0.687223	0.863574	0.75304
LM	0.792654	0.884	0.926055	0.984668	0.897632	0.79088
RF	1.000000	0.828	1.000000	1.000000	1.000000	1.00000
MLP	0.709471	0.716	0.986057	0.912890	0.989296	0.78056
SVM	0.795676	0.740	0.884615	0.736854	0.847875	0.79088
XGB	1.000000	1.000	1.000000	1.000000	1.000000	1.00000
CAT	1.000000	1.000	1.000000	1.000000	1.000000	1.00000
LGBM	1.000000	1.000	1.000000	1.000000	1.000000	1.00000

	Income	Arcene	Bank	BlastChar	Shopper	SHrutime
KNN	0.618813	0.826716	0.617853	0.513771	0.637509	0.508255
LM	0.613590	0.873575	0.720366	0.995860	0.707023	0.500000
RF	1.000000	0.818296	1.000000	1.000000	1.000000	1.000000
MLP	0.578485	0.655335	0.962084	0.965355	0.972971	0.502518
SVM	0.574467	0.696554	0.504169	0.500000	0.507535	0.500000
XGB	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
CAT	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LGBM	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

左圖為 accuracy，右圖為 AUC

針對上圖可知，可以整理歸納出以下資訊：

在 Accuracy 或是 AUC 上，不管哪一個資料集，RandForst、XGboost、CatBoost、LGBM 皆達到 100%。

二、semi-supervised learning

資料切割：50 samples are Training set, the remaining is esting set

重複次數：5 次平均 Accuracy/AUC

	Income	Arcene	Bank	BlastChar	Shopper	SHrutime
KNN	0.725533	0.723765	0.881996	0.656256	0.802378	0.719558
LM	0.771487	0.782100	0.854436	0.753382	0.863925	0.784462
RF	0.799969	0.701677	0.884161	0.763564	0.877932	0.802653
MLP	0.666181	0.566866	0.822364	0.720378	0.861059	0.625065
SVM	0.759208	0.648636	0.882328	0.734878	0.845114	0.796382
XGB	0.772028	0.658815	0.875973	0.738138	0.884723	0.767819
CAT	0.790723	0.639422	0.876849	0.755098	0.886238	0.790352
LGBM	0.774489	0.665412	0.869308	0.747776	0.824577	0.781447

	Income	Arcene	Bank	BlastChar	Shopper	SHrutime
KNN	0.500468	0.746517	0.519897	0.498083	0.518865	0.502759
LM	0.549988	0.755728	0.623730	0.659290	0.742515	0.501737
RF	0.615543	0.693016	0.521645	0.628382	0.695214	0.515256
MLP	0.508906	0.643691	0.552040	0.511875	0.636406	0.497585
SVM	0.500000	0.667297	0.500000	0.499821	0.501457	0.500000
XGB	0.640517	0.651529	0.550840	0.646659	0.739125	0.567113
CAT	0.611889	0.621531	0.556333	0.643084	0.751858	0.541641
LGBM	0.567142	0.640726	0.531130	0.652167	0.520294	0.557033

左圖為 accuracy，右圖為 AUC

針對上圖可知，可以整理歸納出以下資訊：

【資料集 1 – Income】

Accuracy：Randforst > Catboost > LightGBM

AUC：Xgboost> Randforst>Catboost

【資料集 2 – Arcene】

Accuracy：LM > KNN > Randforst

AUC：LM > KNN > Randforst

【資料集 3 – Bank】

Accuracy：Randforst > SVM >KNN
AUC：LM > Catboost >Xgboost
【資料集 4 – BlastChar】
Accuracy：Randforst > Catboost >LM
AUC：LM > LGBM> XGboost
【資料集 5 – Shopper】
Accuracy：Catboost > XGboost > Randforst
AUC：Catboost >LM >XGboost
【資料集 6 – SHrutime】
Accuracy：Randforst > SVM > Catboost
AUC：Catboost >LightGBM > Catboost

【小結論】

1. 即使在 semi-supervised learning 內，雖然樣本數變少，但 RandForst、XGboost、CatBoost、LGBM 四種皆為常勝軍，尤其是 RandForst 向來在 accuracy 的效果都不錯。
2. Logistic Regression 似乎在資料標籤較平均的資料集，表現也不錯，像是 Arcene、BlastChar。
3. 在 Bank 資料上，不管使用哪一種機器學習方法，Accuracy 似乎都差不多，然而在 AUC 就相對比較有差異，可能是因為資料較為不平均。
4. 整體來說，Accuracy 效果最差的為 BlastChar，AUC 效果最差的為 SHrutime。

Conclusions and Thoughts

此次作業除了讓我複習許多機器學習的套件使用外，最重要的是，可以去洞察不同模型對於資料處理的效果評估，也讓我去深刻了解資料不均衡對於預測的效果影響。而在制定模型過程中，總是要去思考該如何預測才會做好，好在還記得「Vote 方法」，果然多數決仍有一定的效果。相信這次作業只是在暖身，期待未來的課程可以更加豐富多元，自己也可以有所成長！