資料探勘專案作業二

開發演算法建模與預測

指導教授：

許中川 教授

成員：

M11112075 徐紹鈞

M11121028 陳韋諠

日期：

2022 年 11 月 24 日

摘要

目前在分類預測上，會運用許多機器學習技術，可透過不同的演算法，從大量的數據中挖掘數據，稱為資料探勘，本研究將會利用 Adult 資料集以及 Bank 資料集，Adult 資料集將會去預測一週的工作時長，Bank 資料集將會預測是否會認購定期存款，本研究將會建立 KNN、SVR 以及 XGBoost 三種不同模型，並且去比較三種的 MAE、MAPE、RMSE 等，去找出兩個資料集中分別績效最好的模型

關鍵字：KNN、SVR、XGBoost

一、 緒論

1.1 動機

1.1.1 Adult Dataset

工作時長一直是大家關心的重點，為此政府也制定了很多相關政策，例如勞工正常工作時間每日不得超過 8 小時，每週不得超過 40 小時，若可以利用人員的一些資料，例:國籍、職業、收入等，可預測出一週的工作時長，便可以幫助政府制定更加完善的制度。

1.1.2 Bank Dataset

在金融相關產業中為了尋求在直接營銷活動中客戶是否會認購定期存款，若可以利用客戶基本資料、當前活動的最後一次聯絡人有關資料以及社會和經濟背景屬性來預測是否會認購定期存款，將可以為銀行抓取重點民眾進行推銷

1.2 研究目的

1.2.1 Adult Dataset

目前的演算法有相當多種，本研究利用 KNN、XGBoost 以及 SVR 三種來建構預測模型，投入工作類別、薪資等 15 種屬性資料，希望透過三種演算法找出預測最佳的績效，並投入實際情況來做使用

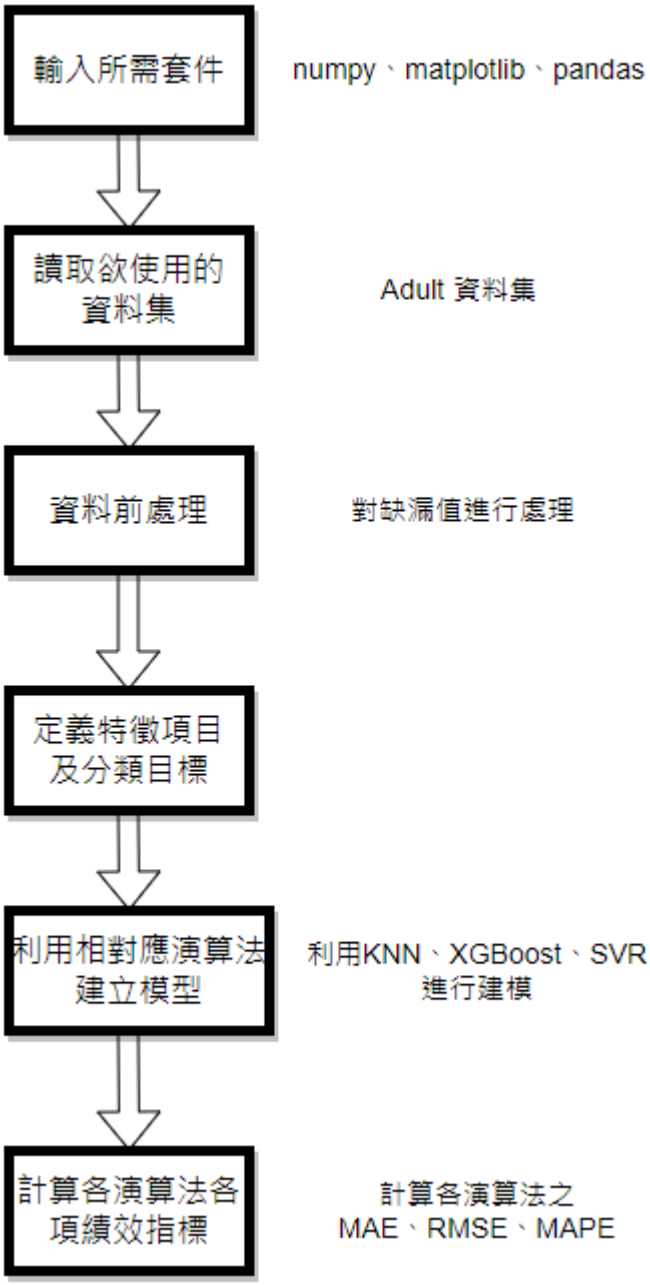
1.2.2 Bank Dataset

為了為銀行抓取重要客戶，本研究將會投入 Bank 資料集中的年齡、工作、教育程度等 16 種屬性，利用 KNN、XGBoost 以及 SVR 三種來建構預測模型，希望透過三種演算法找出最佳的績效，並投入實際情況做使用

二、 方法

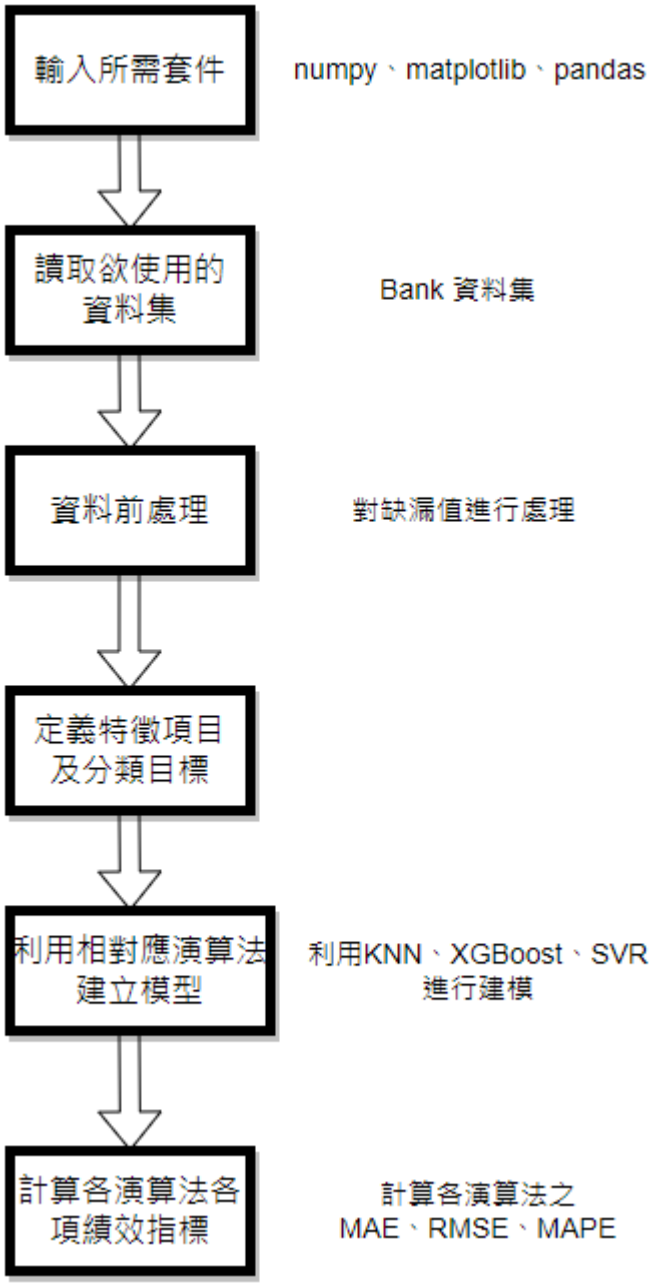
2.1 程式架構

2.1.1 Adult Dataset



圖一 Adult Dataset 之程式架構流程圖及說明

2.1.2 Bank Dataset



圖二 Bank Dataset 之程式架構流程圖及說明

2.2 程式執行方法

2.2.1 **KNN（K nearest neighbor）**

KNN（K nearest neighbor）是一種用於分類及回歸的無母數統計方法，KNN 需要量化相似度，其中最常用的是歐幾里得距離，而 KNN 不僅可以作為分類器，還可以做為回歸連續性的數值預測，分類的標準是由鄰居「多數表決」決定的，而回歸的數值預測則是最近鄰居值的平均值，KNN 也是所有機器學習演算法最簡單之一。

2.2.2 **SVR（Support Vector Regression/Machine）**

SVR（support vector Regression/Machine）支持向量機，可以處理分類以及回歸問題，同時，SVR 也是 SVM 的延伸，SVR 也能夠在資料樣本少的情況下獲得良好的結果。

2.2.3 **XGBoost（eXtreme Gradient Boosting)**

XGBoost（eXtreme Gradient Boosting)，可以做分類也可以進行回歸連續值的預測，通常效果不會太差，並且會利用 Boosting 技巧將許多弱決策樹集成在一起形成一個強的預測模型。

三、 實驗

3.1 資料集

3.1.1 Adult Dataset 說明

此資料集建立於 1996 年 5 月 1 日共有 48,842 筆，15 個欄位。

表一 Adult Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Age	年齡
Workclass	工作類別
Fnlwgt	連續數值
Education	教育程度
Education-num	教育人數
Marital-status	婚姻狀況
Occupation	職業
Relationship	關係
Race	種族
Sex	性別
Capital-gain	資本收益
Capital-loss	資本損失
Hours-per-week	小時/周
Native-country	所屬國家
Salary	年收入

3.1.2 Bank Dataset 說明

此資料集建立於 2018 年 9 月 7 日共有 4899 筆，17 個欄位。

表二 Bank Dataset 欄位資料說明彙總表

age	年齡
job	工作
marital	婚姻
education	教育
default	NA
housing	房產
loan	貸款
contact	合約
month	月份
day_of_week	星期
duration	期間
campaign	活動
pdays	工作日
previous	先前
poutcome	結果
emp.var.rate	就業變動率
cons.price.idx	消費者價格指數
cons.conf.idx	消費者信心指數

3.1.3 實驗數據

(1) Adult Train 資料集

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	output
0	39	7	77516	9	13	4	1	1	4	1	2174	0	40	39	0
1	50	6	83311	9	13	2	4	0	4	1	0	0	13	39	0
2	38	4	215646	11	9	0	6	1	4	1	0	0	40	39	0
3	53	4	234721	1	7	2	6	0	2	1	0	0	40	39	0
4	28	4	338409	9	13	2	10	5	2	0	0	0	40	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	4	257302	7	12	2	13	5	4	0	0	0	38	39	0
32557	40	4	154374	11	9	2	7	0	4	1	0	0	40	39	1
32558	58	4	151910	11	9	6	1	4	4	0	0	0	40	39	0
32559	22	4	201490	11	9	4	1	3	4	1	0	0	20	39	0
32560	52	5	287927	11	9	2	4	5	4	0	15024	0	40	39	1

32561 rows × 15 columns

圖三 Adult Train 資料集

(2) Adult Test 資料集

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	output
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K.
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K.
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K.
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K.
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16276	39	Private	215419	Bachelors	13	Divorced	Prof-specialty	Not-in-family	White	Female	0	0	36	United-States	<=50K.
16277	64	?	321403	HS-grad	9	Widowed	?	Other-relative	Black	Male	0	0	40	United-States	<=50K.
16278	38	Private	374983	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	50	United-States	<=50K.
16279	44	Private	83891	Bachelors	13	Divorced	Adm-clerical	Own-child	Asian-Pac-Islander	Male	5455	0	40	United-States	<=50K.
16280	35	Self-emp-inc	182148	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	60	United-States	>50K.

16281 rows × 15 columns

圖四 Adult Test 資料集

(3) Bank Dataset 資料集

bank-full

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
2	**58**	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
3	**44**	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
4	**33**	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
5	**47**	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
6	**33**	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
7	**35**	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
8	**28**	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
9	**42**	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
10	**58**	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
11	**43**	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
12	**41**	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
13	**29**	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
14	**53**	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
15	**58**	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
16	**57**	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
17	**51**	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
18	**45**	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
19	**57**	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
20	**60**	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
21	**33**	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
22	**28**	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
23	**56**	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no

圖一

圖五 Bank Dataset 資料集

3.2 前置處理

3.2.1 Adult Train Dataset

在 Adult Train 資料集中，首先將原先的 txt 檔轉檔轉為 csv 檔，接著在觀察檔案時發現有許多資料有缺值的現象產生，我們將有缺值的資料進行刪除，完成資料清洗。

3.2.2 Adult Test Dataset

在 Adult Test 資料集中，首先將原先的 txt 檔轉檔轉為 csv 檔，接著在觀察檔案時發現有許多資料有缺值的現象產生，我們將有缺值的資料進行刪除，完成資料清洗。

3.2.3 Bank Dataset

將 Bank 資料集，接著在觀察檔案時發現有許多資料有缺值的現象產生，我們將有缺值的資料進行刪除，完成資料清洗，最後載入進 jupyter Notebook。

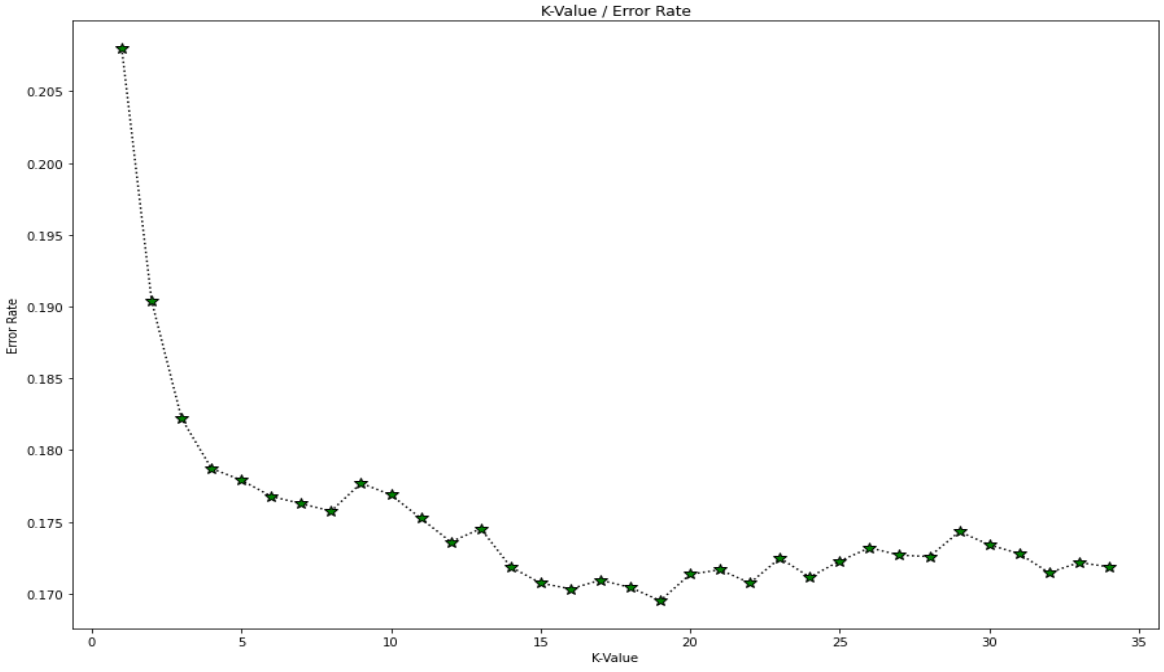
3.3 實驗設計

3.3.1 Adult Train/Test Dataset

(1)正規化：利用 Sklearn.preprocessing 中的 StandardScaler 套件進行正規化。

(2)套件載入：利用 Sklearn 載入 KNN 建模(KNeighborsRegressor)、SVR 建模(SVM)以及 XGBoost(XGBRegressor)以利後續建模。

(3)KNN：首先先利用訓練資料集訓練 KNN 模型，從 K=1 開始接著找尋鄰近 1-2 點，最近利用 for-loop 迴圈，將 K=1~35 的錯誤率畫出，如下圖(六)所示最後丟入測試資料，利用剛建立好的 KNN 模型進行預測。



圖(六)

(4)SVR：限制模型的複雜度，防止過度擬合,kernel 採用 poly

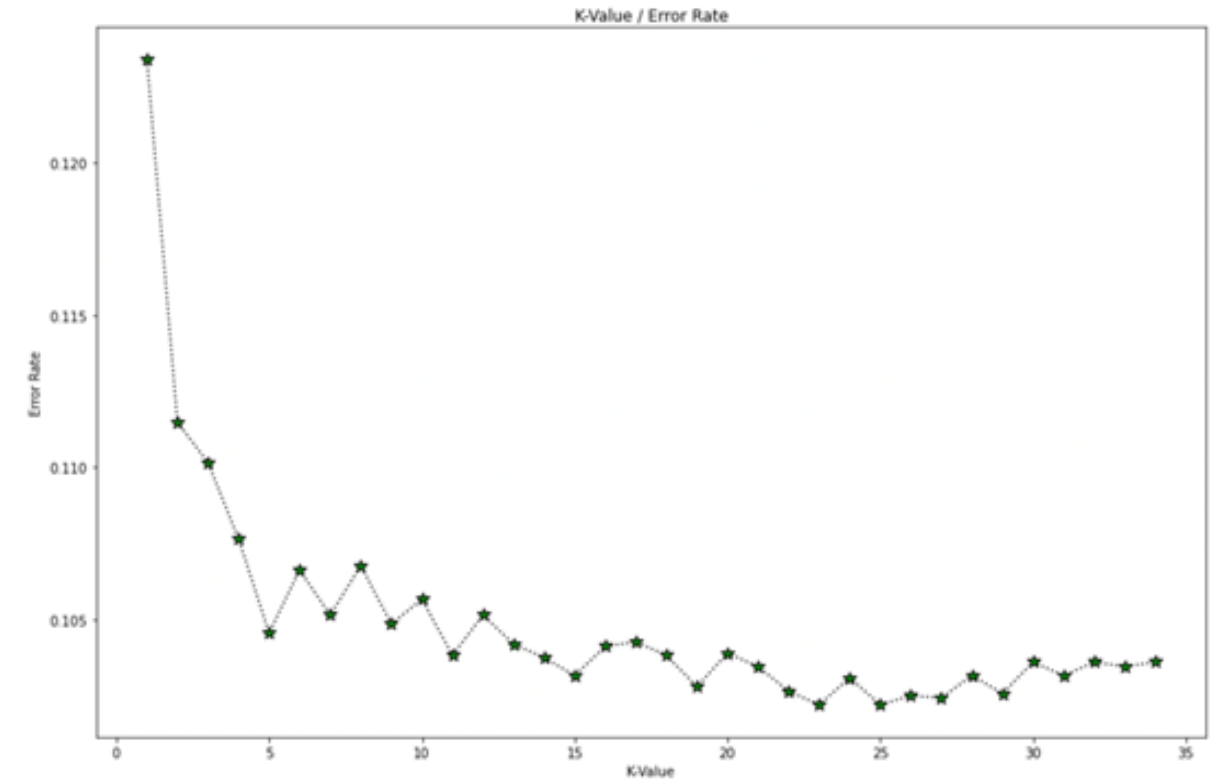
(5)XGBoost：將模型樣本數(n_estimators)設為 100，學習率(learning_rate)設為 0.3，完成參數設定後進行模型建立。

3.3.2 Bank Dataset

(1)正規化：利用 Sklearn.preprocessing 中的 StandardScaler 套件進行正規化。

(2)套件載入：利用 Sklearn 載入 KNN 建模(KNeighborsRegressor)、SVR 建模(SVM)以及 XGBoost(XGBRegressor)以利後續建模。

(3)KNN：首先先利用訓練資料集訓練 KNN 模型，從 K=1 開始接著找尋鄰近 1-2 點，最近利用 for-loop 迴圈，將 K=1~35 的錯誤率畫出，如下圖(七)所示，最後丟入測試資料，利用剛建立好的 KNN 模型進行預測。



圖(七)

(4)SVR：限制模型的複雜度，防止過度擬合,kernel 採用線性

(5)XGBoost：將模型樣本數(n_estimators)設為 100，學習率(learning_rate)設為 0.3，完成參數設定後進行模型建立。

3.4 實驗結果

3.4.1 Adult 績效評估

	KNN	XGBoost	SVR
MAE	10.03	7.44	7.48
RMSE	14.95	10.89	12.46
MAPE	37.8	29.6	35.24

表一 Adult 績效評估

3.4.2 Bank 績效評估

	KNN	XGBoost	SVR
MAE	0.123	0.094	0.230
RMSE	0.351	0.30	0.364
MAPE	12.3	12.73	22.992

表二 Bank 績效評估

四、 結論

Adult Dataset 資料集以 XGBoost 有較好的績效。

Bank Dataset 資料集以 XGBoost 有較好的績效。

總結以上，XGBoost 是一個最佳的分類器。

五、 參考文獻

[1] 後疫情時期下金融業財務及財報之影響與因應：
https://www2.deloitte.com/content/dam/Deloitte/tw/Documents/about-deloitte/tw-Covid19/tw-covid19-newsletters-fsi2-pdf.pdf
[2] 機器學習演算法-KNN
https://aiec.nccu.edu.tw/ai-column/26
[3] KNN
https://pyecontech.com/2020/05/03/python_knn/

[4] SVR 演算法
https://www.796t.com/content/1546145477.html
[5] 核模型-支持向量機（ＳＶＭ）
https://ithelp.ithome.com.tw/m/articles/10270447
[6] Jupyter notebook xgboost import
https://stackoverflow.com/questions/44856105/jupyter-notebook-xgboost-import
[7] 機器學習常勝軍 XGboost
https://ithelp.ithome.com.tw/articles/10273094?sc=iThomeR
[8] bank Dataset
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing