資料探勘專案作業二

訓練模型預測數值

指導教授：

許中川 教授

成員：

M10921002 宋沂芸

M10921032 林師弘

M10921036 童湘庭

M10921038 張珮柔

日期：

2020 年 11 月 18 日

摘要

共享自行車在台灣一直都是一個很成功的代步工具,不僅方便也對環境相當友善，我們使用的數據集分別蒐集了年、月、日、季節、天氣等資料，並對於公共自行車出租總數進行預測，我們分別使用類神經網路、SVR、RandomForest、XGBoost 進行預測，利用 MAPE 及 RMSE 來進行評估，研究結果顯示 SVR 方法績效最高，其次是隨機森林。

據統計資料來看全球收入評比，發現年均所得高的國家幾乎為已開發國家，而以台灣來看，擁有科學園區的新竹其收入也為全台之冠，因此，本研究想了解造成高收入之影響因素，故選取 UCI 資料庫的 Adult 資料，使用類神經網路、XGBoost、Random Forest 與 SVR 進行分析預測及比較，藉此了解高的收入的背景因素，透過 MAPE 和 RMSE 績效評估，研究結果顯示 SVR 方法績效最高，其次是隨機森林。

關鍵字：類神經網路、SVM、XGBoost、隨機森林。

一、 緒論

1.1 動機

1.1.1 Bike Sharing Dataset

共享自行車在台灣的數量日益攀升，根據臺北市政府交通局統計室統計民國 105 年到 106 年間就從原有的 288 站(9442 台)上升到 344 站(11290 台)[4]，面對地狹人稠的台灣交通一直是我們所探討的問題，根據 Fan et al., 2019，中提出共享自行車乞今可以說是最有效的接駁工具，可以針對大眾運輸工具出入不便的區域提供替代方案，且至 2017 年全球已有 20 多個國家 304 個城市使用共享自行車[5]，Cao and Shen, 20193 指出共享自行車帶來許多經濟、環保、社會等效益，公共自行車不僅可以有效降低碳排放量也可以舒緩交通以及汙染問題，至今各國推行共享自行車可說是越來越風行，因此我們針對共享自行車的數量做進一步的預測。

1.1.2 Adult Dataset

2018 年時，全球最高年均所得最高國家為瑞士，人均國民總收入為 83,580 美元，第二名為挪威收入為 80,790 美元，台灣在世界排名 30，收入為 25,360[1]，可發現排名全世界前幾名的薪水幾乎都屬於已開發國家。以台灣為例，新竹的平均月薪全台最高，為台幣 47,523 元，第二名則為苗栗，為台幣 41,167 元，其餘的六都排名卻都在苗栗後面[2]，而造成此排名之主要原因正是苗栗地理位置處於新竹周邊，而新竹科學園區帶動了周邊產業的發展，半導體產業也幫助周圍城市提高平均薪資，能發現產業會影響薪資的表現，又根據新竹科學園區統計，竹科員工的教育程度多為碩士及學士，年齡平均為 38.27[3]，可以發現產業、教育程度與年齡可能都會影響收入高低的原因，故本研究想了解其他會對於全球收入造成影響之因素。

1.2 研究目的

1.2.1 Bike Sharing Dataset

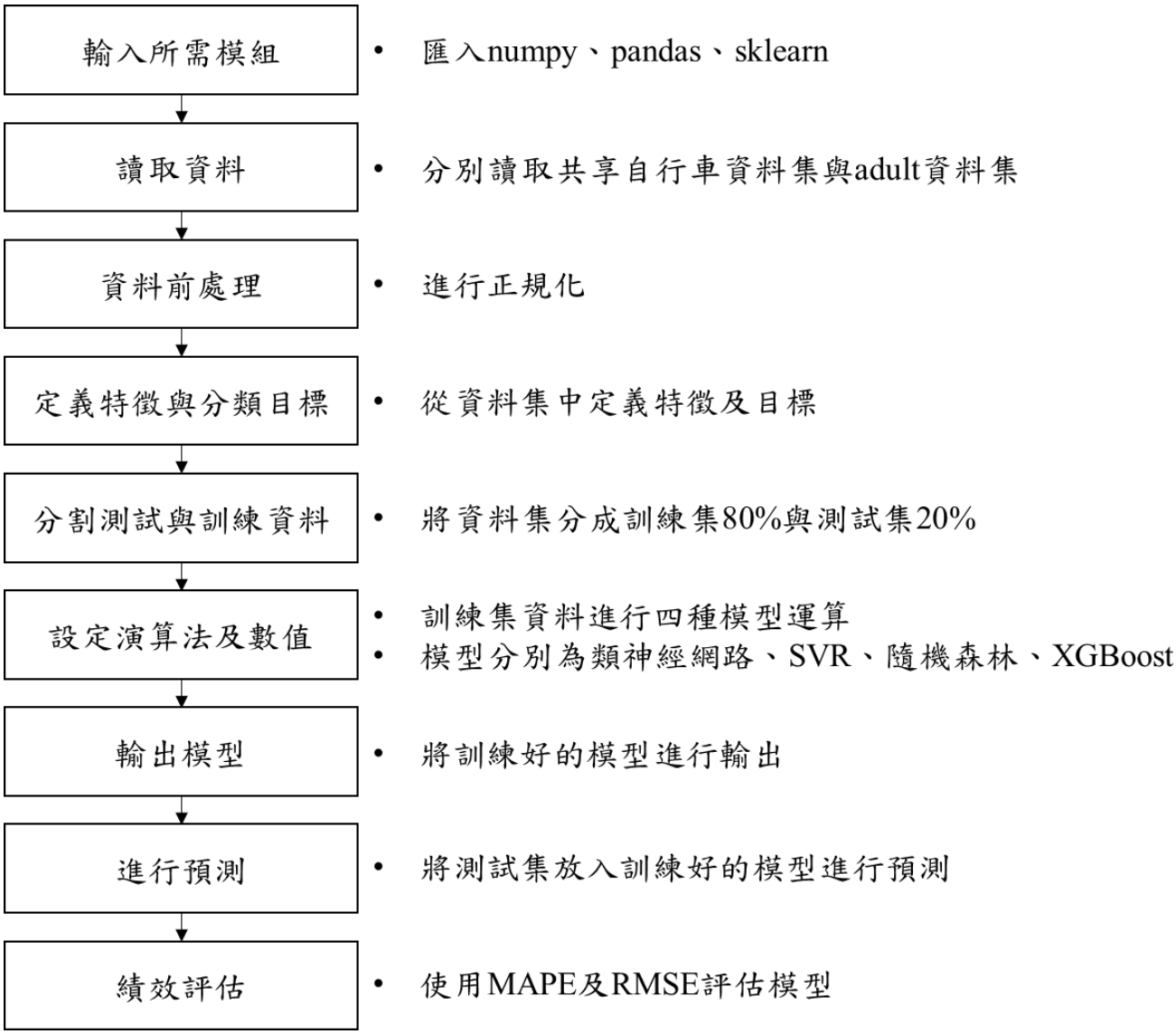
隨著共享自行車以及使用人數不斷地增加，如何有效的調度並確保每站都有足夠的自行車可以使用成為一大問題，為確保民眾都可以有充足的數量可以使用共享自行車，因此對於各站分配可說極為重要，我們目的在於預測各種天氣或是季節情況下自行車的出租量，可以藉此當參考從適當地時機來調度各站的共享自行車。

1.2.2 Adult Dataset

本研究欲探討影響收入高低之因素，故選取 Adult 資料集，並使用類神經網路、XGBoost、Random Forest 與 SVR 進行分析預測及比較，分析年齡、工作類別、教育程度、職業、種族等 14 種屬性資料分析，預測出一個人的年收入是否超過 50K，藉此更加了解什麼樣背景的人能夠獲取更高的收入。

二、 方法

2.1 程式架構



圖一 程式架構流程圖及說明

2.2 執行方法

2.2.1 SVR

支持向量回歸，在樣本空間中找到一個回歸平面，讓所有數據和特定平面的距離最近[12]。

2.2.2 隨機森林

隨機森林將多棵使用 Gini 係數的決策數結合在一起，並加入隨機分配的訓練資料，此方法結合多個較弱的學習器來建構較強的模型同時也增加了模型的多樣性。

2.2.3 XGBoost

每一次保留原來的模型不變，並且加入一個新的函數至模型中，糾正上一棵樹的殘差，以提升整體的模型[14]。

2.2.4 類神經網路

當對神經元進行輸入後，經過激發函數與內部迴歸模型對輸入的權重加乘，再加入偏誤後，便完成了該節點的輸出，之後在傳給下一個神經元，如此一層層傳遞直到最後的輸出層並產生預測結果，類神經網路會由預測結果與真實結果之間的差距，對整個神經網路進行更新[13]。

三、 實驗

3.1 資料集

3.1.1 Bike Sharing Dataset 說明

此資料集建立於 2013 年 12 月 20 日共有 17,389 筆，23 個欄位。

表一 Bike Sharing Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
instant	資料索引。
dteday	資料紀錄日期。
season	季節
yr	年份
mnth	月份
hr	小時
holiday	當天是否為假日
weekday	當天星期幾
workingday	當天是否為工作天
weathersit	天氣
temp	攝氏溫度
atemp	標準化後的攝氏體感溫度
hum	標準化濕度

（續下頁）

（承上頁）

欄位名稱	欄位說明
windspeed	標準化風速
casual	休閒用戶數量
registered	註冊用戶
cnt	出租自行車總數

3.1.2 adult dataset 說明

此資料集建立於 1996 年 5 月 1 日共有 48,842 筆，14 個欄位。

表二 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	教育年級
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	每周多少小時
native-country	祖國
annual salary	年收入

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	annual salary
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K

圖二 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	annual salary
25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K.
38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K.
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K.
44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K.
18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K.
34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male	0	0	30	United-States	<=50K.
29	?	227026	HS-grad	9	Never-married	?	Unmarried	Black	Male	0	0	40	United-States	<=50K.
63	Self-emp-not-inc	104626	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	3103	0	32	United-States	>50K.
24	Private	369667	Some-college	10	Never-married	Other-service	Unmarried	White	Female	0	0	40	United-States	<=50K.
55	Private	104996	7th-8th	4	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	10	United-States	<=50K.

圖三 adult_test 資料集

(3)bike_sharing 資料集

instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011/1/1	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	2011/1/1	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	2011/1/1	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	2011/1/1	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	2011/1/1	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	2011/1/1	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	2011/1/1	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	2011/1/1	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
9	2011/1/1	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
10	2011/1/1	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14

圖四 bike_sharing 資料集

3.2 前置處理

3.2.1 Bike Sharing Dataset

將 csv 檔案中的特徵「dteday」內的日期利用 dataframe 拆解成年、月、日三個數字，並另外新增三個欄位「year」、「month」、「date」儲存拆解出的三種數值。SVR 方法將資料常態分佈化，平均值為 0，標準差為 1，使離群值降低。

3.2.2 adult dataset

將 adult.train.txt 和 adult.test.txt 轉為 csv 檔案，將檔案內的字串進行編碼正規化而後刪除帶有空白值的資料列，使資料更具有分析效益。SVR 方法將資料常態分佈化，平均值為 0，標準差為 1，使離群值降低。

3.3 實驗設計

3.3.1 Bike Sharing Dataset

設定 x 為不包含「cnt」的 data，y 為特徵「cnt」，再將 data 分成 80%的訓練資料，20%的測試資料。SVR 懲罰參數設為 1、隨機森林限制決策數最大深度為 2、XGBoost 設定最大深度為 3。

3.3.2 adult dataset

將 x_train 設定為資料集 adult.train 去除特徵「hours_per_week」的 data，x_test 設定為資料集 adult.test 的特徵「hours_per_week」進行測試。SVR 懲罰參數設為 1、隨機森林限制決策數最大深度為 2、XGBoost 設定最大深度為 3。

3.4 實驗結果

3.4.1 Bike Sharing Dataset

本研究使用 SVR、隨機森林、XGBoost 和類神經網路方法，由下表可得知 Bike Sharing Dataset 中類神經網路績效最高，其次是 XGBoost。

表三 Bike Sharing Dataset 績效評估彙總表

方法	MAPE	RMSE
SVR	114%	54.85
隨機森林	115%	59.87
XGBoost	53%	21.8
類神經網路	26.56%	0.43

3.4.2 adult dataset

本研究使用 SVR、隨機森林、XGBoost 和類神經網路方法，由下表可得知 adult dataset 中 SVR 方法績效最高，其次是隨機森林。

表四 adult dataset 績效評估彙總表

方法	MAPE	RMSE
SVR	18.38%	0.33
隨機森林	35.89%	0.36
XGBoost	67.22%	12.09
類神經網路	43.45%	25.34

四、 結論

4.1 Bike Sharing Dataset

本研究利用四種方法，經過 Bike Sharing Dataset 發現類神經網路對於時間序列資料預測績效最高，其次則是 XGBoost，未來若有研究預測連續時間相關數據可優先選擇類神經演算法。

4.2 Adult Dataset

本研究利用 SVR、隨機森林、XGBoost、類神經網路以上四種方法，對於 adult 資料集進行分析，研究結果顯示 SVR 績效最高，其次是隨機森林，多種屬性資料預測適合 SVR 演算法。

五、 參考文獻

[1]各國人均國民總收入列表
https://is.gd/KApaG7
[2]2020 全台薪資薪水比較
https://salary.tw/map
[3]新竹科學園區-園區歷年就業員工數之成長-依教育程度區分
https://is.gd/G3bH3g
[4]台北市公共自行車使用特性
https://is.gd/3ZC8HD
[5] How have travelers changed mode choices for first/last mile trips after the introduction of bicycle-sharing systems: An empirical study in Beijing, China)
https://www.hindawi.com/journals/jat/2019/5426080/
[6] Contribution of shared bikes to carbon dioxide emission reduction and the economy in Beijing
https://www.sciencedirect.com/science/article/pii/S2210670719308765
[7]雲端基礎教學(2) colab 基本操作與建議
https://ithelp.ithome.com.tw/articles/10217962
[8] XGBoost – A Scalable Tree Boosting System：Kaggle 競賽最常被使用的演算法之一
https://is.gd/td7vlr
[9]python 將 txt 文件轉為 csv 檔案
https://blog.csdn.net/kanon122500000/article/details/56844893
[10]pandas 外部處理缺失值
https://blog.csdn.net/dss_dssssd/article/details/82814673
[11]資料前處理-標籤編碼
https://is.gd/9O66vm
[12]SVR 介紹
https://is.gd/bCApfW
[13]類神經網路介紹
https://is.gd/bw2OvQ
[14]XGBoost 介紹
https://is.gd/Mf411F