國立雲林科技大學

資訊管理所

資料探勘專案作業二

比較機器學習模型預測：
共享單車需求數量與人力薪資之研究

指導教授：許中川教授

學生：工業工程與管理學系 碩一

M11121019 劉驊維

M11121027 洪源廷

M11121036 杜佳容

M11121047 黃惠妤

摘要

世界上第一座共享單車系統是由法國在 2005 年 5 月建置完成，利用共享單車代替大眾運輸工具或自有車輛除了能紓解交通壅塞，還能減少空氣汙染、降低噪音等，在良好的規劃下建置共享單車能為城市帶來許多益處。隨著共享單車站點設置越多，共享單車不僅能用來通勤，也能延伸成為假日觀光或是日常運動的工具，相對於其他運輸工具自行車的需求也更容易受到天氣狀況影響，是否能及時滿足使用者需求成了問題，若能先預測出各個站點的需求，並提早在站點間調度腳踏車輛必能讓共享單車系統發揮更大效用。本研究數據來源為 UC Irvine Machine Learning Repository 中的 Seoul Bike Sharing Demand Data Set，數據包含 8760 筆實例 14 筆相關訊息，欲利用網格搜尋調整參數，比較 3 個機器學習模型分別 XGBoost、AdaBoost 與 RandomForest ，針對每小時自行車所需數量進行預測的能力，並利用 RMSE、MAPE、MAE 績效指標進行績效評比。

此外，在現今社會有許多貧困的家庭無法提供子女足夠的教育資源，造成階級複製正所謂的世襲貧窮，富裕的人更加富裕貧困的人生活卻是越來越平窮，M 型社會在全球化的趨勢下越來越極端，在財富分配如此不平均的社會裡，為了讓生活過得更好而努力工作，根據台灣勞工局 110 年勞工調查統計顯示有延長工時(加班)者占 46.3%，若長期工作超時必會造成身體的負擔。為了瞭解在努力工作賺錢與適當工作時數間取得平衡，本研究針對 UC Irvine Machine Learning Repository 中的 Adult Data Set 進行分析，比較 3 個機器學習模型分別為 XGBoost、AdaBoost 與 RandomForest，預測在不同條件因素下的每週工作時數能力，並利用 RMSE、MAPE、MAE 績效指標進行評比。

一、緒論

1.1 動機

Seoul Bike Data

城市發展必伴隨著空氣污染，據新聞媒體報導南韓於 2019 年空氣污染達到有史以來最高紀錄，且接連六天多達 15 個地區被迫實行管制政策，降低排放廢氣量，因此，韓國近年推行綠色環保的共享經濟，其一為共享單車。「共享單車」不僅可以降低空氣污染，還可以解決交通擁擠、油價上漲等問題，除此之外，還能豐富市民的休閒生活也能方便短程外出，提升市民的生活品質，進而打造健康的社會。而本研究為了使其預測準確率提高，將選擇較適用的模型，投入隨機森林、極限梯度提升及 Adaboost 進行模型間的比較，預測在不同站點的不同時間下單車需求量，且利用 RMSE、MAPE 與 MAE 進行模型績效的評估。

Adult Dataset

「M 型化社會」是由大前研一所提出的理論，指在全球化的趨勢下，富者財富快速攀升，而另一方面，隨著資源重新分配，中產階級因失去競爭力而淪落到中下階層，造成中間出現缺口，就像英文字母「M」。換言之，貧富差距擴大，因此人們為了爬上金字塔頂端必須更努力的工作，但在賺取金錢的同時必須同時兼顧身體健康，避免過度勞累，因此本研究透過比較三種機器學習模型選擇較適用、準確率較高的模型，預測在不同條件下的每週工作時數，並使用 RMSE、MAPE 與 MAE 進行績效的評估。

1.2 目的

Seoul Bike Data

在人工智慧盛行的時代，預測工具應用在產業上越來越廣泛，本研究採用 SeoulBikeData 資料集由 UCL 網站提供，收錄 2017 年至 2018 年首爾共享單車需求資料，投入三種模型預測共享單車租借需求數量，並利用均方根誤差、平均絕對誤差與平均絕對百分比誤差評比三種模型的績效狀況，而模型分別為隨機森林、極限梯度提升以及 Adaboost，其目的為評估此三種模型何者為較佳的模型，提高預測需求數量的準確率，進而對共享單車提高使用率，將運具的價值最大化，以利於共享業者在智慧化規劃有更佳的管理與決策。

Adult Dataset

時代日新月異，運用預測工具不只使用在產業智慧化，更加能使用工具探討人周遭事物，本研究採用 Barry Becker 於 1944 年從人口普查數據庫獲取的資料集，投入三種模型預測每週將要上班的小時數，模型分別為隨機森林、極限梯度提升以及 Adaboost，並利用多種績效指標評估三種模型的績效狀況，其目的為評比三種模型績效，選擇在此資料下較適用的模型，並期望能使其預測準確率提高，進而更加能準確探討，眾多因子的組合下需要花費多少上班時間。

二、方法

2.1 程式架構

本研究使用監督式學習方法極限梯度提升、隨機森林與自適應增強學習器進行預測回歸，研究流程如圖 2.1.1 所示。首先把 Seoul Bike Data 資料集隨機分為 9:1 的訓練集(Training Set)和測試集(Test Set)，再把訓練集(Training Set）隨機分為 8：2 的訓練集（Training Set）和驗證集（Validation Set），另外 Adult Dataset 資料集已有訓練集與測試集不需額外分訓練集與測試集，接著進行資料前處理包含透過特徵重要性篩選屬性、資料轉換與正規化，再將資料投入多種模型並利用網格搜尋調整參數並交叉驗證，最後進行績效評估。

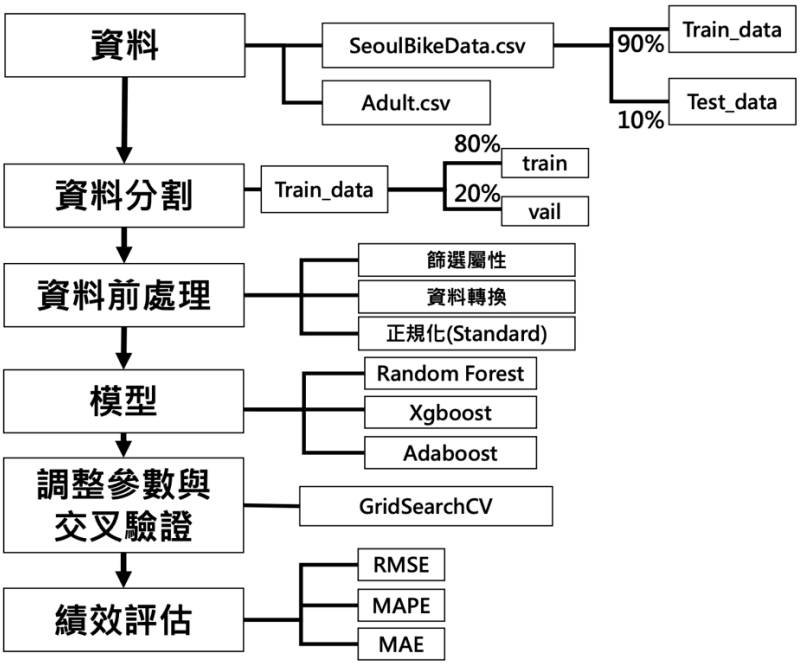


圖 2.1.1 研究流程

2.2 執行方法

必須在環境裡安裝套件

1. Sklearn
2. Pandas
3. Numpy
4. XGboost

兩者模型訓練集和測試集皆必須與程式在同一個資料夾。程式結束後訓練成果將輸出在 console 上，如圖 2.2.1 所示，第一行至第三行表示網格搜尋計算出的成績與最佳參數組合，第四行至第六行顯示分別為訓練績效與測試績效。

```
XGB最佳準確率: 0.8557507414011614，最佳參數組合：{'eta': 0.01, 'max_depth': 6, 'n_estimators': 1000}
RF最佳準確率: 0.8445242990527451，最佳參數組合：{'max_depth': 10, 'n_estimators': 100}
ADA最佳準確率: 0.6712840972353449，最佳參數組合：{'learning_rate': 0.1, 'n_estimators': 100}
Train：      XG_rmse      XG_mae     RF_rmse     RF_mae    ada_rmse    ada_mae
0  235.9162  149.741907  248.490243  157.15763  376.950581  274.80446
Train_time：        XG_time    RF_time   ada_time
0  229.719003  38.522425  58.667918
Test：       XG_rmse      XG_mae     RF_rmse      RF_mae    ada_rmse     ada_mae
0  236.721032  171.366436  245.017112  162.237294  367.816437  278.777942
```

圖 2.2.1 訓練成果

三、實驗

3.1 實驗架構

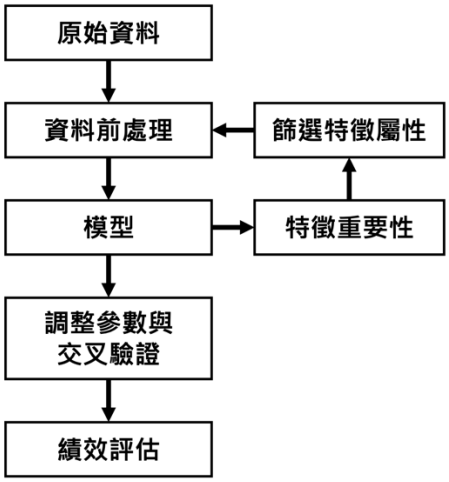


圖 3.1.1 實驗架構

3.2 資料集

3.2.1 Seoul Bike Sharing Demand Data Set

此研究資料集來自 UIC 機器學習儲存庫，名稱為 Seoul Bike Sharing Demand Data Set， 資料集內有 14 個欄位，共有 8760 筆資料，數據集包含數項天氣訊息、每小時租用的自行車數量和日期信息等，資料屬性如下表 3.1.1；資料部分內容如表 3.2.2、表 3.1.3。

表 3.2.1 Seoul Bike Sharing Demand Data Set 資料屬性

屬性名稱	屬性說明	屬性類別
Date	日期	Interval
Rented Bike Count	每小時租用的自行車計數	Interval
Hour	一天中第幾小時	Interval

表格接續下一頁……

Temperature	溫度	Ratio
Humidity(%)	濕度	Ratio
Wind speed (m/s)	風速	Ratio
Visibility (10m)	能見度	Ratio
Dew point temperature	露點溫度	Ratio
Solar Radiation (MJ/m2)	太陽輻射	Ratio
Rainfall(mm)	降雨量	Ratio
Snowfall (cm)	降雪量	Ratio
Seasons	季節	Nominal
Holiday	假日/非假日	Nominal
Functioning Day	工作日	Nominal

表 3.2.2 Seoul Bike Sharing Demand Data Set 資料集部分內容

Date	Rented Bike Count	Hour	Temperature	Humidity (%)	Wind speed (m/s)	Visibility (10m)
01/12/2017	254	0	-5.2	37	2.2	2000
01/12/2017	204	1	-5.5	38	0.8	2000
01/12/2017	173	2	-6	39	1	2000
01/12/2017	107	3	-6.2	40	0.9	2000

表 3.2.3 Seoul Bike Sharing Demand Data Set 資料集部分內容

Dew point temperature	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
-17.6	0	0	0	Winter	No Holiday	Yes
-17.6	0	0	0	Winter	No Holiday	Yes
-17.7	0	0	0	Winter	No Holiday	Yes
-17.6	0	0	0	Winter	No Holiday	Yes

3.2.2 Adult Dataset

Adult Dataset 資料集來自 UIC 機器學習儲存庫，總共有 48,8412 比實例，包含 15 個屬性欄位，資料屬性如表 3.2.4；資料部分內容表 3.2.5、表 3.2.6。

表 3.2.4 Adult Dataset 資料說明表

屬性名稱	屬性說明	屬性類別
age	年齡	Ratio
workclass	工作類別	Nominal
fnlwgt	人口普查員 ID	Nominal
education	教育程度	Ordinal
education-num	教育程度 -依數字排序	Ordinal
marital-status	婚姻狀況	Nominal
occupation	職業	Nominal
relationship	關係	Nominal
race	種族	Nominal
sex	性別	Nominal
capital-gain	資本收益	Ratio
capital-loss	資本損失	Ratio
hours-per-week	每週上班小時數	Nominal
native-country	國家	Nominal
target	薪資	Ratio

表 3.2.5 Adult Dataset 資料集部分內容

age	*workclass*	*fnlwgt*	*education*	*education*	*marital-status*
25	Private	226802	11th	7	Never-married
38	Private	89814	HS-grad	9	Married-civ-spouse
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse
44	Private	160323	Some-college	10	Married-civ-spouse

表 3.2.6 Adult Dataset 資料集部分內容

occupation	relationship	race	sex	capital-gain
Machine-op-inspct	Own-child	Black	Male	0
Farming-fishing	Husband	White	Male	0
Protective-serv	Husband	White	Male	0
Machine-op-inspct	Husband	Black	Male	7688

表 3.2.7 Adult Dataset 資料集部分內容

capital-loss	hours-per-week	native-country	target
0	40	United-States	<=50K.
0	50	United-States	<=50K.
0	40	United-States	>50K.
0	40	United-States	>50K.

3.3 前置處理

Seoul Bike Sharing Demand Data Set

- 篩選特徵屬性：利用特徵重要性篩選出較少被採用訓練的兩個特徵屬性，並刪除此屬性。
- 資料轉換-類別化：資料集內包含名目資料，因此先將這些資料利用 labelencoder 類別化；特徵屬性“Visibility (10m)”記錄可見度的距離詳細，但可見度有等級之分，所以將資料轉換成級別之分。
- 正規化：數值資料皆利用 StandardScaler 正規化，使資料分布更集中，有利於模型訓練。

Adult Dataset

- 篩選特徵屬性：利用特徵重要性篩選出較少被採用訓練的兩個特徵屬性，並刪除此屬性，除了用特徵重要性篩選之外，將不含有意義的欄位刪除，分別為 “capital-gain”與“capital-loss”。
- 資料轉換-類別化：資料集內包含名目資料，因此先將這些資料利用 labelencoder 類別化。
- 正規化：數值資料皆利用 StandardScaler 正規化，使資料分布更集中，有利於模型訓練。

3.4 實驗設計

本研究第一階段實驗皆使用預設參數訓練模型，計算出特徵屬性的重要性以及各模型績效，其中，查看重要性的屬性，並對欄位做篩選。

資料前處理後，第二階段將兩者資料模型皆使用網格搜尋套件，找尋參數 n_estimator、max_depth 與 eta/learning_rate 的最佳參數組合，但因為每個機器學習模型參數設置較不一致，所以會有參數的差異性，使其交叉驗證 5 次，使在訓練模型時能更加保守。

- 網格搜尋：

  參數 n_estimator：

  XGBoost 搜尋範圍 [500,625,750,875,1000]。

  RandomForest 搜尋範圍 [50,100,150,200,250]。

  AdaBoost 搜尋範圍 [10,20,30,40,50,60,70,80,90,100]。

  參數 max_depth：

  XGBoost 與 RandomForest 搜尋範圍 [2, 4, 6, 8, 10]。

  參數 eta/learning_rate：

  XGBoost 搜尋範圍：[0.01,0.05,0.1]

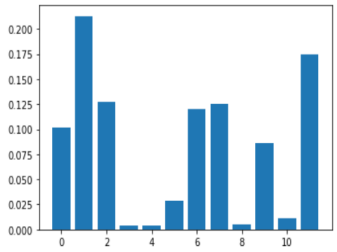
  AdaBoost 搜尋範圍：[0.1,0.325,0.55,0.775,1]

3.5 實驗結果

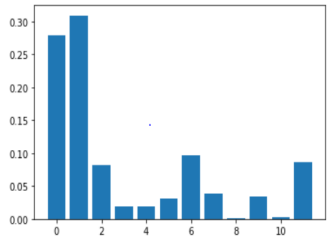
第一階段實驗-特徵屬性篩選前

Seoul Bike Sharing Demand Data Set

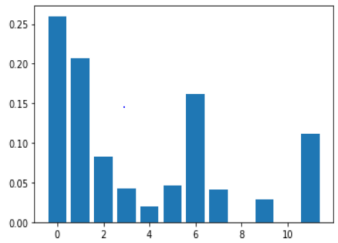
以下圖 3.5.1、3.5.2 與 3.5.3 分別為 XGBoost、RandomForest 與 AdaBoost 的特徵重要性狀況。



圖(3.5.1)



圖(3.5.2)



圖(3.5.3)

由上圖結果統整得知，欄位 8-Holiday 與 10-Snowfall(cm)在三個模型訓練之中並非必要，所以在第二階段實驗將此欄位刪除。下表 3.5.1 與 3.5.2 為第一階段各模型績效，此資料集的目標值含有多筆 0 的數值，所以將不使用 mape 作為此資料集的績效指標。由表 3.5.2 結果得知雖然 XGBoost 在 rmse 指標中是最佳的，但有過擬合的問題，並且在 mae 指標是沒有比 RandomForest 好的，所以若要求穩定性，選擇 RandomForest 較佳，但若要求最高準確率則 XGBoost 較佳。

表 3.5.1 各模型參數組合

	網格搜尋準確成績	最佳參數組合		
		n_estimator	max_depth	eta / learning_rate
XGBoost	0.8607	1000	6	0.01
RandomForest	0.8461	200	10	
AdaBoost	0.6683	100		0.1

表 3.5.2 在訓練集或測試集的各項績效指標

	XG_rmse	XG_mae	RF_rmse	RF_mae	ada_rmse	ada_mae
Train	229.8516	145.1168	247.9692	155.8884	378.8636	278.9693
Test	236.0565	170.6935	241.72	160.5947	365.9463	279.6555

Adult Dataset

以下圖 3.5.4、3.5.5 與 3.5.6 分別為 XGBoost、RandomForest 與 AdaBoost 的特徵重要性狀況。

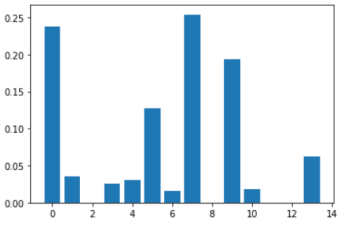


圖 3.5.4

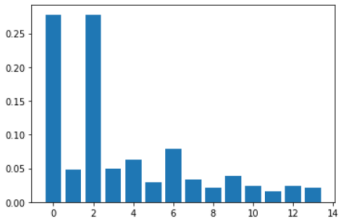


圖 3.5.5

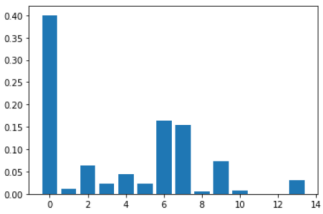


圖 3.5.6

由上圖結果統整得知，欄位 8-sex 與 12-native-country 在三個模型訓練之中並非必要，所以在第二階段實驗將此欄位刪除。下表 3.5.3 與 3.5.4 為第一階段各模型績效。由表 3.5.4 結果得知 RandomForest 的各個指標是三個模型中最佳的模型，且在測試集績效與訓練集績效的差距與其餘的兩者模型為最高，比預期評估的績效還更佳，所以若要在此研究三種模型挑選，選擇 RandomForest 為較佳的模型。

表 3.5.3 各模型參數組合

	網格搜尋準確成績	最佳參數組合		
		n_estimator	max_depth	eta / learning_rate
XGBoost	0.2406	1000	8	0.01
RandomForest	0.2428	250	10	
AdaBoost	0.1703	200		0.55

表 3.5.4 在訓練集或測試集的各項績效指標

	XG_rmse	XG_mae	XG_mape	RF_rmse	RF_mae	RF_mape	ada_rmse	ada_mae	ada_mape
train	10.59062	7.185999	0.2951	10.53181	7.147489	0.2935	11.53874	8.485104	0.3620
test	10.26659	6.977299	0.29.51	9.833566	6.709651	0.2935	11.51996	8.494135	0.3620

第二階段實驗-特徵屬性篩選後

Seoul Bike Sharing Demand Data Set

最佳參數組合、訓練時間以及預測的績效指標列為下表 3.5.5 與 3.5.6。

表 3.5.5 各模型參數組合

	網格搜尋準確成績	最佳參數組合		
		n_estimator	max_depth	eta / learning_rate
XGBoost	0.8558	1000	6	0.01
RandomForest	0.8445	100	10	
AdaBoost	0.6712	100		0.1

表 3.5.6 在訓練集或測試集的各項績效指標

	XG_rmse	XG_mae	RF_rmse	RF_mae	ada_rmse	ada_mae
Train	235.9162	149.7419	248.4902	157.1576	376.9506	274.8045
Test	236.721	171.3664	245.0171	162.2373	367.8164	278.7779

因為，此資料集的目標值含有多筆 0 的數值，所以不適用 mape 作為此資料集的績效指標。本研究在 SeoulBikeData 資料下從表 3.5.6 得知，XGBoost 在訓練集與測試集各項指標並未最佳，並且略一點點的過擬合，在 mae 指標顯示 RandomForest 比 XGBoost 更佳，以至於若要求穩定選擇 RandomForest 較佳，但若要求準確率則 XGBoost 較佳。

Adult Dataset

最佳參數組合、訓練時間以及預測的績效指標列為下表 3.5.7 與表 3.5.8。

表 3.5.7 各模型參數組合

	網格搜尋準確成績	最佳參數組合		
		n_estimator	max_depth	eta / learning_rate
XGBoost	0.2327	1000	8	0.01
RandomForest	0.2335	150	10	
AdaBoost	0.1621	200		0.55

表 3.5.8 在訓練集或測試集的各項績效指標

	XG_rmse	XG_mae	XG_mape	RF_rmse	RF_mae	RF_mape	ada_rmse	ada_mae	ada_mape
train	10.6619	7.1934	0.2945	10.5987	7.1548	0.2933	11.5256	8.399	0.3592
test	10.3319	6.9926	0.2945	9.8905	6.7258	0.2933	11.4793	8.4032	0.3592

從表 3.5.8 得知，本研究在 Adult Dataset 資料下，Random Forest 在訓練集與測試集各項指標皆為最佳，可觀察在訓練集與 XGBoost 相差性不高，但在測試集可發現 Random Forest 預測準確度更佳，以至於與 XGBoost 差距更大。

第一階段實驗與第二階段實驗在 SeoulBikeData 資料下，兩階段實驗皆指出，XGBoost 比起 RandomForest 績效較好但穩定度不佳，而比較結果未篩選特徵欄位績效比篩選刪除後的特徵欄位績效更佳，上述得知在 SeoulBikeData 資料下訓練模型，每個欄位都為重要的特徵屬性；在 Adult Dataset 資料下，兩階段實驗皆顯示出，RandomForest 在本研究三者模型中為最佳的模型，並且無過擬合的狀況，而兩階實驗比較結果無太大的差異，由此可知，在這個資料集下刪除不必要的欄位對於績效並無太大的幫助。

四、結論

在模型多元的人工智慧時代裡，選用適用的模型是提高準確率必要條件之一，當準確率越高，越能使使用者、管理者以及決策者能更有效的掌控未來的趨勢，以至於管理的更完善或做出更佳的決策，本研究使用 XGBoost、AdaBoost 與 RandomForest 三個機器學習模型，預測不同站點在不同時間下的單車需求數量，並使用 RMSE、MAPE、MAE 三個績效指標去評估績效，結果顯示若要要求預測準確率需使用 XGBoost，因其績效指標較佳，但如果要要求模型的穩定性則需使用 RandomForest，所以可以依照預測需求去選擇使用模型。

另一筆資料為預測在不同條件下的勞工工作時數，並比較三個機器模型所做出來的績效，且實驗中使用三種不同的績效做評估，分別為 RMSE、MAPE 與 MAE，結果皆顯示使用 RandomForest 為最佳，因此，如要預測勞工工作時數使用 RandomForest 的準確率為最佳。

參考文獻

https://www.mol.gov.tw/1607/1632/1633/48857/post

https://zh.m.wikipedia.org/zh-tw/%E5%85%AC%E5%85%B1%E8%87%AA%E8%A1%8C%E8%BB%8A

https://www.jendow.com.tw/wiki/M%E5%9E%8B%E7%A4%BE%E6%9C%83

https://www.storm.mg/article/1033782

https://blog.csdn.net/weixin_39037925/article/details/93628363

https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/learning-model-xgbregressor%E5%8F%83%E6%95%B8%E8%AA%BF%E6%95%B4-ca3dcebbe23