

資料探勘

## 專案作業二

組員：黃凱翔、吳俊園、賀鈞嗣、吳鈞達

2022 年 11 月

# 摘要

此次作業使用三種演算法:KNN、RandomForest、XGBoost 來建構數值預測模型，並使用 scikit-learn 的 feature\_importance 計算特徵重要性，對欄位特徵進行篩選，比較特徵欄位刪減前後績效之差異，針對 adult 資料集和 bikeshare 資料集進行資料預處理，藉由各別資料集中的屬性之間的相關性進行預測，例如:透過 adult 資料集的 workclass、income、education 等屬性間的關聯性推測出工時、透過 bike-sharing 的假日和天氣的關聯性推測出 bike 的單日租用數量。

透過訓練完成的模型處理測試資料，通過計算得到訓練與測試時的準確率、MAE(平均絕對誤差)、RMSE(均方根誤差)及 MAPE(平均絕對百分比誤差)等資料並進行分析。

關鍵字:python、KNN、RandomForest、XGBoost

# 第一章、緒論

## 1.1 動機

透過對課堂上所教授的 KNN、SVR、RandomForest、XGBoost 演算法，以前項作業所使用之 Adult 資料集進行實作練習，以了解學習的成果和累積實作之經驗。

完成後，再行自選一資料集 Bike Sharing，用同樣的方法運算一遍，觀察結果差異。

## 1.2 目的

在 Adult 資料集中預測其 hours-per-week，意即每周工時之欄位，觀察在各項其他資料(例如：學歷、收入等等)對工時的影響性，並預測測試資料集中其他對象的工時準確度。

在 bikesharing 資料集中預測其腳踏車的總使用數，觀察天氣和是否假日等對總使用數的影響和預測的準確度。

## 第二章、方法

首先引入 pandas、sklearn、numpy、matplotlib、seaborn 等函式庫，讀取資料，進行資料預處理，將屬性數值化，設定預測目標，針對實驗所需要使用的三種演算法(KNN、RandomForest、XGBoost)分別安裝和匯入模組。

程式輸出結果如計算特徵重要性、比較特徵欄位刪減的績效差異，三種績效指標(MAE、RMSE、MAPE)。

## 第三章、實驗

### 3.1 資料集

#### 3.1.1 Adult 資料集

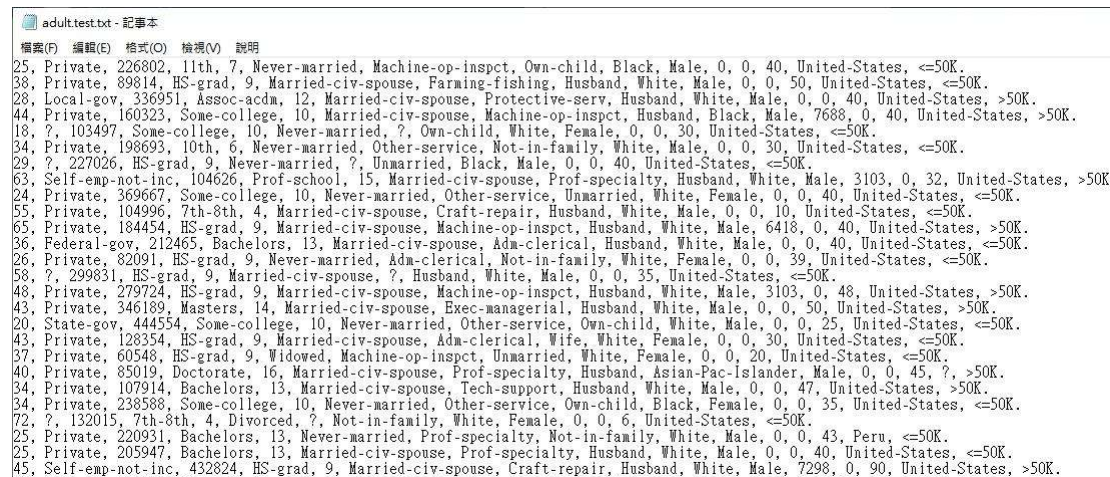


adult.train.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

89, 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, 284509, 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  
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K  
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K  
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K  
32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K  
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K  
34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K  
25, Self-emp-not-inc, 176756, HS-grad, 9, Never-married, Farming-fishing, Own-child, White, Male, 0, 0, 35, United-States, <=50K  
32, Private, 186824, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, United-States, <=50K  
38, Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 50, United-States, <=50K  
43, Self-emp-not-inc, 232175, Masters, 14, Divorced, Exec-managerial, Unmarried, White, Female, 0, 0, 45, United-States, >50K  
40, Private, 193524, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 60, United-States, >50K  
54, Private, 302146, HS-grad, 9, Separated, Other-service, Unmarried, Black, Female, 0, 0, 20, United-States, <=50K  
35, Federal-gov, 76845, 9th, 5, Married-civ-spouse, Farming-fishing, Husband, Black, Male, 0, 0, 40, United-States, <=50K  
43, Private, 117037, 11th, 7, Married-civ-spouse, Transport-moving, Husband, White, Male, 0, 2042, 40, United-States, <=50K  
59, Private, 109015, HS-grad, 9, Divorced, Tech-support, Unmarried, White, Female, 0, 0, 40, United-States, <=50K  
56, Local-gov, 216851, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 40, United-States, >50K

Figure 1 adult.train 部分資料



adult.test.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

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.  
65, Private, 184454, HS-grad, 9, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 6418, 0, 40, United-States, >50K.  
36, Federal-gov, 212465, Bachelors, 13, Married-civ-spouse, Adm-clerical, Husband, White, Male, 0, 0, 40, United-States, <=50K.  
26, Private, 82091, HS-grad, 9, Never-married, Adm-clerical, Not-in-family, White, Female, 0, 0, 39, United-States, <=50K.  
58, ?, 299831, HS-grad, 9, Married-civ-spouse, ?, Husband, White, Male, 0, 0, 35, United-States, <=50K.  
48, Private, 279724, HS-grad, 9, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 3103, 0, 48, United-States, >50K.  
43, Private, 346189, Masters, 14, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 50, United-States, >50K.  
20, State-gov, 444554, Some-college, 10, Never-married, Other-service, Own-child, White, Male, 0, 0, 25, United-States, <=50K.  
43, Private, 128354, HS-grad, 9, Married-civ-spouse, Adm-clerical, Wife, White, Female, 0, 0, 30, United-States, <=50K.  
37, Private, 60548, HS-grad, 9, Widowed, Machine-op-inspct, Unmarried, White, Female, 0, 0, 20, United-States, <=50K.  
40, Private, 85019, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 45, ?, >50K.  
34, Private, 107914, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 47, United-States, >50K.  
34, Private, 238588, Some-college, 10, Never-married, Other-service, Own-child, Black, Female, 0, 0, 35, United-States, <=50K.  
72, ?, 132015, 7th-8th, 4, Divorced, ?, Not-in-family, White, Female, 0, 0, 6, United-States, <=50K.  
25, Private, 220931, Bachelors, 13, Never-married, Prof-specialty, Not-in-family, White, Male, 0, 0, 43, Peru, <=50K.  
25, Private, 205947, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 40, United-States, <=50K.  
45, Self-emp-not-inc, 432824, HS-grad, 9, Married-civ-spouse, Craft-repair, Husband, White, Male, 7298, 0, 90, United-States, >50K.

Figure 2 adult.test 部分資料

**Adult 資料如下：**

- (a) 資料筆數：48842。
- (b) 資料屬性欄位數：14。
- (c) 資料型態：整數資料、分類型別資料。
- (d) 缺失值：有。

## 其屬性的資訊為

- (a) 收入：大於 50K，小於 50K。
- (b) 工作：私人公司、自營公司、政府雇員或從未工作等。
- (c) 教育程度：小學、中學、高中、大學、碩士、博士等。
- (d) 婚姻狀況：未婚、已婚、離婚、喪偶等。
- (e) 職業：各種不同職業。
- (f) 關係：親人、妻子、丈夫、非親屬、其他親屬。
- (g) 種族：白人、印地安人、亞裔等。
- (h) 投資收益、投資損失：正收益或副收益數值。
- (i) 性別：男、女。
- (j) 每周工作小時數：0 至 144 小時之間。
- (k) 出生國家：各個國家。

## 3.1.2 Bike Sharing 資料集

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	instant	datetime	season	yr	mnth	holiday	weekday	workingday	weather	temp	atemp	hum	windspeed	casual	registered	cnt
2	1	2011/1/1	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
3	2	2011/1/2	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
4	3	2011/1/3	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
5	4	2011/1/4	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
6	5	2011/1/5	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
7	6	2011/1/6	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
8	7	2011/1/7	1	0	1	0	5	1	2	0.196522	0.208039	0.490696	0.168726	148	1362	1510
9	8	2011/1/8	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
10	9	2011/1/9	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
11	10	2011/1/10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
12	11	2011/1/11	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263
13	12	2011/1/12	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25	1137	1162
14	13	2011/1/13	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38	1368	1406
15	14	2011/1/14	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54	1367	1421
16	15	2011/1/15	1	0	1	0	6	0	2	0.233333	0.248112	0.49875	0.157963	222	1026	1248
17	16	2011/1/16	1	0	1	0	0	0	1	0.231667	0.234217	0.48375	0.188433	251	953	1204
18	17	2011/1/17	1	0	1	1	1	0	2	0.175833	0.176771	0.5375	0.194017	117	883	1000
19	18	2011/1/18	1	0	1	0	2	1	2	0.216667	0.232333	0.861667	0.146775	9	674	683
20	19	2011/1/19	1	0	1	0	3	1	2	0.292174	0.298422	0.741739	0.208317	78	1572	1650
21	20	2011/1/20	1	0	1	0	4	1	2	0.261667	0.25505	0.538333	0.195904	83	1844	1927

Figure 3 bike sharing 部分資料

**Bike 資料如下：**

- (a) 資料筆數：17379。
- (b) 資料屬性欄位數：18。
- (c) 資料型態：數值正規化資料。
- (d) 缺失值：無。

**其屬性的資訊為：**

- (a) 編號：紀錄的編號，從 1 開始。
- (b) 日期：年、月、日。
- (c) 季節：春夏秋冬(1-4)。
- (d) 年：從 2011 年起為第 0 年。
- (e) 月：月份。
- (f) 小時：當日幾時。
- (g) 假日：是否為假期(不含周休)。
- (h) 星期：為星期幾。
- (i) 工作日：非周末且非假期為 1，否則為 0。
- (j) 氣象站：1 為晴、2 為多雲、3 為小雪、4 為大雨。
- (k) 溫度：經過正規化運算後的溫度。

- (l) 體感溫度：經過正規化運算後的體感溫度。
- (m) 溫度與體感溫度的差異百分比：溫度除以體感溫度。
- (n) 風速：經過正規化運算後的風速。
- (o) 臨時用戶數：非會員的用戶數量。
- (p) 註冊會員數：有註冊的會員使用數量。
- (q) 總腳踏車使用數：會員以及非會員的腳踏車總租用量。

### 3.2 前置處理

在 Adult 的預處理中，利用 `panda` 內的 `replace` 函式將工作的缺失值取代為 `Private`，職業的缺失值取代為 `Prof-specialty`，出生國家的缺失值取代為 `United-States`。

接著使用同一函式，針對教育程度欄位將離散的文字資料 `Preschool`, `'1st-12th'` 轉換為統一的 `school`，`'HS-grad'` 轉換為 `'high school'`，將 `'Assoc-voc'`, `'Assoc-acdm'`, `'Prof-school'`, `'Some-college'`, `'Bachelor'` 轉換為 `'higher'`，將 `'Masters'` 轉換為 `'grad'`，將 `'Doctorate'` 轉換為 `'Doc'`，將 `'Masters'` 轉換為 `'grad'`。

將結婚欄位的資料中 `'Married-civ-spouse'`, `'Married-AF-spouse'` 轉換成 `'married'`，`'Never-married'` 轉換為 `'not-married'`，以及 `'Divorced'`, `'Separated'`, `'Widowed'`, `'Married-spouse-absent'` 轉換為 `'other'`。



最後將資料集要判別的收入欄位，`income` 中大於 50K 的轉換為 1，小於等於 50K 的轉換為 0。

預處理完之後，將所有除了預測目標以外的數值標準化，使數據的平均值為 0，方差為 1。

### 3.3 實驗設計

下載課堂上提供的 `Adult` 資料集，裡面已經包含了訓練和測試資料，使用 `Panda` 直接讀取，並進行預處理和正規化。

對三種演算法分別進行運算，並輸出預測績效，逐次調整參數使其績效提高，也就是績效值越小、表現越好。

`Random Forest` 中，`n_estimators`(子樹的數量)為 100，`n_jobs`(能使用處理器的數量)設定為 1，`oob_score`(是否驗證子樹)為 `true`。

`KNN` 中，`n_neighbors`(選取最鄰近的點數量)為 101，`weights`(權重)選取為 `distance`，`p`(距離度量)為 2。

`XGBoost` 中，按照模型給的預設參數下去運算，並未調整。

在 `Bike-sharing` 中，由於先套用 `Adult` 中調整好的參數後準確率非常高，因此使用同一參數設定。

### 3.4 實驗結果

以 `python` 輸出資料格式表示資料集在三種演算法得到結果。如 `Figure 4`、`Figure 5`、`Figure 6`、`Figure 7`、`Figure 8`、`Figure 9` 所示

```

RandomForestregression
Acc on training data: 0.886
Acc on test data: 0.202
MAE :
7.745468616646339
MAPE :
0.4335125455009968
RMSE
11.147444130794238
特徵重要性:
      featureimportant
0          0.282436
1          0.044127
2          0.301549
3          0.030731
4          0.072582
5          0.021008
6          0.077679
7          0.030619
8          0.021137
9          0.038313
10         0.021883
11         0.014958
12         0.020789
13         0.022189

```

Figure 4 Adult 資料集 RandomForest 結果

```

knn回歸器
MAE :
7.519929566717921
MAPE :
0.3218385464024124
RMSE
11.019684886162558
Acc on training data: 0.999
Acc on test data: 0.220

```

Figure 5 Adult 資料集 KNN 結果

```

XGBoost
訓練集: 0.4514131113853884
測試集: 0.25817565786631635
特徵重要性:
      featureimportant
0          0.117882
1          0.052318
2          0.028113
3          0.029402
4          0.052162
5          0.027048
6          0.056362
7          0.069941
8          0.028131
9          0.231735
10         0.029463
11         0.026573
12         0.024894
13         0.225977

```

Figure 6 Adult 資料集 XGBoost 結果

```

RandomForestregression
Acc on training data: 1.000
Acc on test data: 1.000
MAE :
0.9899904104334485
MAPE :
8.064775602548908
RMSE
3.142682561505943
特徵重要性:
      featureimportant
0          0.000015
1          0.000004
2          0.000028
3          0.000034
4          0.000001
5          0.000025
6          0.000004
7          0.000007
8          0.000031
9          0.000031
10         0.000047
11         0.000039
12         0.051198
13         0.948536

```

Figure 7 Bike 資料集 RandomForest 結果

```

knn回歸器
MAE :
2.0942608383482386
MAPE :
0.037638942979833015
RMSE
5.421562460666625
Acc on training data: 1.000
Acc on test data: 0.999

```

Figure 8 Bike 資料集 KNN 結果

```

XGBoost
訓練集: 0.9999290406693654
測試集: 0.9996305983861592
特徵重要性:
      featureimportant
0          0.000025
1          0.001337
2          0.000076
3          0.000066
4          0.000055
5          0.000067
6          0.000096
7          0.000044
8          0.000062
9          0.000125
10         0.000099
11         0.000073
12         0.061983
13         0.935893

```

Figure 9 Bike 資料集 XGBoost 結果

## 第四章、結論

在兩個資料集中，Adult 花費特別多時間調整參數，準確率一直無法大幅提升，用最初始的參數值時測試資料集的準確率只有 10% 至 20%之間，後續針對不同演算法調整其參數才有一定的提升，但在測試資料集中依舊表現未達滿意標準。

考量測試時間，直接用同一批參數下對 Bike 資料集運算，所得的準確率近乎 100%，反覆比對後，推測是在資料前處理的部分的差別，可能在前處理時，要針對各欄位中的離群值做前處理，也可能是 Adult 中資訊欄位較為複雜以及缺失值較多造成的。

因此，若要再次進行實驗，應選擇較完整的資料集，或者是資料欄位設計較相關的資料集，或能表現較好。

## 第五章、參考文獻

- [1] *Adult*. (n.d.). UCIdataset. <https://archive.ics.uci.edu/ml/datasets/adult>
- [2] *Bike-Sharing*. (n.d.). UCIdataset.  
<https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>
- [3] *RandomForestClassifier*. (n.d.). Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [4] Hallows, S. (n.d.). *Using XGBoost with Scikit-Learn*. Kaggle.  
<https://www.kaggle.com/code/stuarthallows/using-xgboost-with-scikit-learn/notebook>
- [5] *KNeighborsClassifier*. (n.d.). Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>