摘要

本實驗基於 UCI datasets 中的 Adult Data Set 資料集以及 Bike Sharing Data Set 資料集作為實驗樣本，前者為 1994 年美國成年人收入的相關調查資料，後者為 Capital 共享單車系統中 2011 至 2012 年的共享單車每日租賃數量及相關數據。

本實驗將對於 hours-per-week 屬性欄位做預測；後者為 CNT 屬性欄位做預測。本實驗使用線性回歸、隨機森林、KNN 以及 XGBoost，分別透過 MAE (平均絕對誤差)、RMSE (均方根誤差)和 MAPE (平均絕對百分比誤差) 等三個衡量指標評估各個模型對於數值的預測績效，並參考特徵重要性提取出重要程度較高的屬性將之保留，再次利用模型預測及評估，比較此資料集在三種演算法下的績效改善率。

關鍵字：線性迴歸；隨機森林；XGBoost；KNN；機器學習

一、緒論

1.1 動機

在早期的環境下，可能因為教育程度普遍不高，許多人會依賴拉長工時，賺取額外的薪水。本實驗選擇了 UCI Datasets 的 Adult Data Set 資料集進行實驗，教育程度與薪水等等因素是否對於一個人的每周工時有無影響，在這項實驗中將探究出真正因素。

共享單車系統是近期因為交通、環境、健康的問題，讓人們對他感到興趣，使得有快速的發展，實驗選擇了 Bike Sharing Data Set 資料集進行實驗，探討在甚麼條件下容易影響到單車的租借數量。

1.2 目的

本實驗將利用資料分析技術建構出模型，檢視影響每週工時的因素，從中理解主要原因是否為教育程度及薪水所導致，抑或是其他因素造成。透過此實驗將得知主要因素，且能夠以結果端視政府以及人民素養和生活品質水準，利用此實驗結果向政府提供建議及反饋。

由於共享單車在發展上其實一直存在許多問題，藉由實驗分析後，大概可以得知在甚麼情況下，單車的日租使用量，可以更好的讓機構控管單車放置數量，盡可能地減少不必要的維護、商品成本等等。

二、方法

本實驗所使用的方法分為四個階段，第一階段選擇資料集，並將資料集引入進行後續實驗。第二階段做前置處理，將缺失值做刪減，調整維度大小，將離散型特徵屬性及連續型特徵屬性做處理，最後將轉換後的數字做 MinMax 尺度標準化。第三階段將資料分成訓練集和測試集。第四階段建構出模型使用線性迴歸、隨機森林、KNN 和 XGBoost，分別計算特徵重要性，對特徵屬性進行篩選，比較欄位刪減前後的績效差異。更詳細的內容如實驗步驟所說明。

三、實驗

3.1 資料集

本實驗的資料來源於 UCI Datasets 的 Adult Data Set 資料集以及 Bike Sharing Data Set 資料集。Adult Data Set 資料集是 Barry Becker 從 1994 年的人口普查資料庫中提取出來的。本實驗將前者的 adult.data (檔案更名為 adult.train) 的資料作為訓練集，有 32561 筆資料，adult.test 資料作為測試集，有 16281 筆資料。資料集有 15 個屬性欄位，其中 9 個屬性屬於離散型，另外 6 個屬於連續型屬性。由於欄位過多無法全部呈現，部分資料內容如表 1 所示；後者的是 Bike Sharing Data Set 的資料集是 2011 年到 2012 年共享單車系統中取得的。Bike Sharing Data Set 的資料集裡共有 17389 筆資料，有 17 個屬性欄位，其中 13 個屬性屬於離散型，另外 4 個屬性屬於連續行。由於欄位過多無法全部呈現部分資料內容如表 2 所示。

表 1

Adult Data Set 資料集部分內容

age	workclass	fnlwgt	education	education-num	marital-status	…
39	State-gov	77516	Bachelors	13	Never-married	…
…	…	…	…	…	…	…

表 2

Bike Sharing 資料集部分內容

dteday	season	yr	mnth	hr	hoilday	…
2011/1/1	1	0	1	0	0	…
…	…	…	…	…	…	…

3.1.1 準備資料集

下載 Adult Data Set 的兩個資料集 adult.data 和 adult.test，及 Bike Sharing Data Set 的 hours 資料集。

3.1.2 導入相關套件

使用 numpy、pandas、xgboost、matplotlib.pyplot、sklearn 的 RandomForestRegressor、LinearRegression、KNeighborsRegressor、sklearn.preprocessing 的 MinMaxScaler、sklearn.metrics 的 mean_squared_error、mean_absolute_error 和 mean_absolute_percentage_error。

3.1.3 導入資料集

利用 pandas 的 read_table() 導入 Adult 資料集與 read_csv()導入 Bike_sharing 資料集。

3.2 前置處理

3.2.1 缺失值處理

由於 Adult Data Set 資料集的訓練集具有相當的資料筆數，缺失值的數量佔比相對整體資料集並不足以影響結果預測，因此本實驗採用直接刪除含有缺失值的個案。刪除所有缺失值後，剩餘的資料筆數為 30162 筆；Bike Sharing Data Set 資料集不具有缺失值。

3.2.2 降維處理

Adult 資料集中分析 15 個屬性欄位，其中 fnlwgt、capital-gain、capital-loss 等三個屬性與本實驗的分析較無相關，education-num 與 education 屬性的意義重複了，relationship 則可藉由 sex 及 marital-status 判斷出，因此本實驗將 fnlwgt、capital-gain、capital-loss、education-num、relationship 等五個屬性納入了不考量的欄位，剩餘的屬性數量為 10 個。

Bike_sharing 資料集中分析了 17 個屬性欄位，其中 instant、dteday 屬性與本實驗的分析較無相關，而 cnt 屬性為 casual、registered 兩個屬性的加總，因此本實驗納入 cnt 欄位而刪除 causal、regustered 欄位，剩餘的屬性數量為 13 個。

3.2.3 離散型屬性處理

剩餘的 10 個屬性中，有 8 個屬於離散型屬性，下列為分別的處理方式。

(1) workclass：具有 8 個類別，將各類型進行調整合併為 5 大類別，雖不具有順序，若利用 pd.get_dummies() 會導致 sparsity 過高，因此本實驗利用 pd.factorize() 轉換為數值。

(2) education：具有 16 個類別，將各類型進行調整合併為 7 大類別，此為順序型屬性，本實驗按照順序做數值轉換。

(3) marital-status：具有 7 個類別，將各類型進行調整合併為 5 大類別，雖不具有順序，若利用 pd.get_dummies() 會導致 sparsity 過高，因此本實驗利用 pd.factorize() 轉換為數值。

(4) occupation：具有 14 個類別，將各類型進行調整合併為 6 大類別，雖不具有順序，若利用 pd.get_dummies() 會導致 sparsity 過高，因此本實驗利用 pd.factorize() 轉換為數值。

(5) race：具有 5 個類別，雖不具有順序，若利用 pd.get_dummies() 會導致 sparsity 過高，因此本實驗利用 pd.factorize() 轉換為數值。

(6) sex：具有 2 個類別，雖不具有順序，但屬性本身屬於二元類別，因此本實驗利用 pd.factorize() 轉換為數值。

(7) native-country：具有 41 個類別，將各類型進行調整合併為 2 大類別，雖不具有順序，由於剩餘 2 個類別，因此本實驗利用 pd.factorize() 轉換為數值。

(8) income：具有 2 個類別，雖不具有順序，但屬於二元類別，因此本實驗利用 pd.factorize() 轉換為數值。

3.2.4 連續型屬性處理

剩餘的 10 個屬性中，有 2 個屬於連續型屬性，下列為分別的處理方式。

(1) age：本實驗利用 pd.factorize() 轉換為 0 到 1 之間的數值。

(2) hours-per-week：本實驗利用 pd.factorize() 轉換為 0 到 1 之間的數值。

3.2.5 資料標準化

本實驗使用 sklearn.preprocessing 提供的 MinMaxScaler 作為標準化工具。

3.3 分割資料集

本實驗將 Adult Data Set 作者提供的 adult.data (檔案更名為 adult.train)作為訓練集，adult.test 作為測試集；Bike Sharing 資料集透過 sklearn 所提供的 train_test_split() 分成訓練集和測試集、訓練集共包含 12165 筆資料，占總資料集的 70%、測試集則含有 5214 筆，占總資料集的 30%。

3.4 實驗設計

在 Adult Data Set 中，實驗建構出三個模型，分別為線性迴歸、隨機森林以及 XGBoost 模型，將對於特徵篩選前後的 MAE、RMSE、MAPE 指標績效表現做比較。其中，計算 MAPE 的真實值是放在分母當除數，此資料集當中因為有 0 的資料存在，因此得出的數值非常高。線性迴歸模型無特別調整參數，透過 LassoCV 得出權重 (ssswill，2019)，並將權重可視化得到重要度，篩選出 age、education、marital-status、income 等四個欄位其餘刪除，刪除前後的績效如表 3 所示；隨機森林模型的參數 n_estimators 經實驗發現設為 100 時有較佳績效，此參數的實驗比較如表 4 所示。透過 feature_importance_ 得出特徵重要度，篩選出 age、workclass、education、marital-status、occupation、sex 等六個欄位其餘刪除，刪除前後的績效如表 5 所示；XGBoost 模型參數無調整，皆為預設值，透過 feature_importance_得出特徵重要度，篩選出 age、workclass、marital-status、occupation、sex、income 等 6 個欄位其餘刪除，刪除前後的績效如表 6 所示。綜合以上實驗結果統整出預測 Adult Data Set 的三個模型績效前後差異如表 7 所示。

在 Bike_Sharing Data Set 中，實驗建構出三個模型，分別為隨機森林、KNN 及 XGBoost 模型，將對於特徵篩選前後的 MAE、RMSE、MAPE 指標績效表現做比較。隨機森林模型的參數 n_estimators 經實驗發現設為 150 時有較佳績效，此參數的實驗比較如表 8 所示。，並透過 feature_importance_ 得出特徵重要度，將重要度較低的欄位進行刪除，最後篩選出 yr、hr、workingday、temp 等 4 個欄位其餘刪除，刪除前後的績效如表 9 所示。在 KNN 的模型中使用迴圈去測試當 n_neighbor 從 2 至 30 時的績效，並從中取出 n_neighbor 的最佳值，迴圈呈現如圖 1 所示。KNN 並無特別調整參數，透過 LassoCV 得出權重，並將權重可視化得到重要度，保留特徵權重最高的前 4 個欄位 atemp、hr、temp、yr 其餘欄位皆刪除。刪除前後的績效如表 10 所示。XGBoost 模型的參數 n_estimators 經實驗發現設為 150 時有較佳績效，此參數的實驗比較如表 11 所示。並透過 feature_importance_ 得出特徵重要度，將重要度較低的欄位進行刪除，最後篩選出 season、yr、hr、workingday、weathersit、temp 和 atemp 共計 7 個欄位其餘欄位皆刪除，刪除前後的績效如表 12 所示。綜合以上實驗結果統整出預測 Bike_Sharing Data Set 的三個模型績效前後差異如表 13 所示。

3. 5 實驗結果

表 3

Adult Data Set 線性迴歸特徵篩選前後的績效表現

	篩選前	篩選後
MAE	0.079	0.076
RMSE	0.117	0.119
MAPE	625992815449.615	423354833355.006

由表 3 可得知，本實驗對於 Adult Data Set 資料集所探討之篩選特徵前後的績效在線性迴歸模型上較無明顯差異。

表 4

Adult Data Set 隨機森林 n_estimators 參數不同時的績效表現

	n_estimators=50	n_estimators=100	n_estimators=150
MAE	0.090	0.089	0.089
RMSE	0.128	0.126	0.126
MAPE	548504954284.201	554746179039.942	553816669259.833

由表 4 可得知，參數值為 100 時較佳，若設為 150 也具有同樣績效。

表 5

Adult Data Set 隨機森林特徵篩選前後的績效表現

	篩選前	篩選後
MAE	0.089	0.060
RMSE	0.126	0.087
MAPE	554746179039.942	298372664301.463

由表 5 可得知，本實驗對於 Adult Data Set 資料集使用隨機森林模型，在篩選特徵前後的績效有明顯差異。

表 6

Adult Data Set XGBoost 特徵篩選前後的績效表現

	篩選前	篩選後
MAE	0.089	0.068
RMSE	0.127	0.100
MAPE	480360244312.169	314918555160.189

由表 6 可得知，本實驗對於 Adult Data Set 資料集使用 XGBoost 模型，在篩選特徵前後的績效也有明顯差異。

表 7

Adult Data Set 在不同模型下篩選特徵前後的績效下降程度

	線性迴歸	隨機森林	XGBoost
MAE	0.003	0.029	0.021
RMSE	-0.002	0.039	0.027
MAPE	202637982094.609	256373514738.479	165441689151.980

由表 7 可得知，篩選特徵後線性迴歸的績效表現普通，在 RMSE 指標中甚至得出了反效果，因此較不適於此實驗；篩選特徵後隨機森林的績效表現最有成效，三個指標中所下降的幅度最大；篩選特徵後 XGBoost 的績效表現也不亞於隨機森林。

表 8

Bike Sharing Data Set 隨機森林 n_estimators 參數不同時的績效表現

	n_estimators=50	n_estimators=100	n_estimators=150
MAE	26.213	26.137	26.033
RMSE	44.086	43.625	43.275
MAPE	0.332	0.332	0.332

由表 8 可得知，比較隨機森林各 n_estimators 底下的最佳績效，實驗後 n_estimators = 150 時具有最佳績效。

表 9

Bike Sharing Data Set 隨機森林特徵篩選前後的績效表現

	篩選前	篩選後
MAE	26.033	41.990
RMSE	43.275	68.376
MAPE	0.332	0.479

由表 9 可得知，隨機森林模型經過刪除特徵重要性較低的欄位後，篩選前的績效比篩選後績效差。

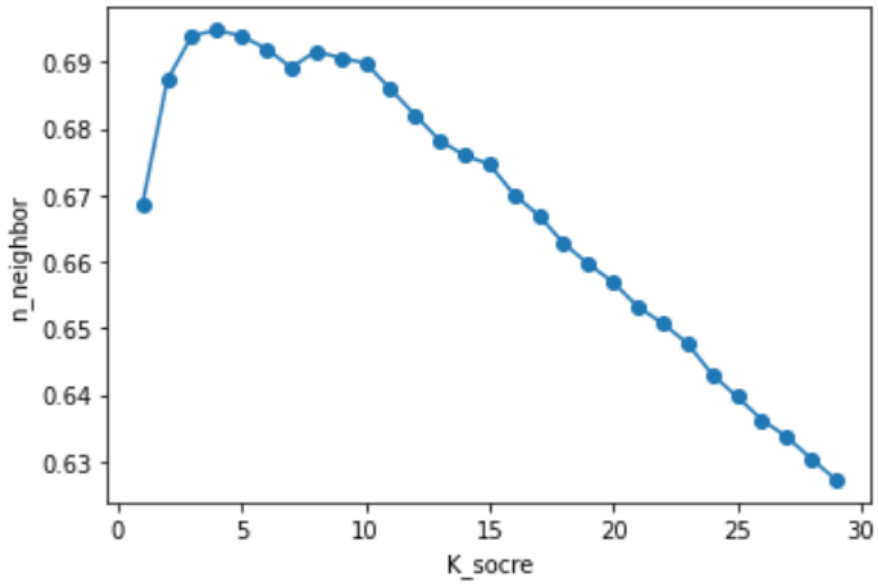

圖 1 Bike Sharing Data Set KNN 不同績效下 n_neighbor 變化圖

由圖 1 得知，n_neighbor 在 5 的時候有最佳績效，之後逐漸遞減

表 10

Bike Sharing Data Set KNN 特徵篩選前後的績效表現

	篩選前	篩選後
MAE	65.049	43.545
RMSE	100.819	69.945
MAPE	1.495	0.515

由表 10 可得知，KNN 模型經過刪除特徵重要性較低的欄位後，篩選後績效比篩選前績效高。

表 11

Bike Sharing Data Set XGBoost n_estimators 參數不同時的績效表現

	n_estimators=50	n_estimators=100	n_estimators=150
MAE	26.801	26.113	25.838
RMSE	41.976	40.970	40.616
MAPE	0.432	0.432	0.428

由表 11 可得知，比較 XGBoost 各 n_estimators 底下的最佳績效，實驗後 n_estimators = 150 時具有最佳績效。

表 12

Bike Sharing Data Set XGBoost 特徵篩選前後的績效表現

	篩選前	篩選後
MAE	25.838	30.542
RMSE	40.161	49.234
MAPE	0.428	0.401

由表 12 可得知，XGBoost 模型經過刪除特徵重要性較低的欄位後，篩選後績效比篩選前績效差。

表 13

Bike Sharing Data Set 在不同模型下篩選特徵前後的績效下降程度

	隨機森林	KNN	XGBoost
MAE	-15.957	21.504	-4.704
RMSE	-25.102	30.874	-9.071
MAPE	-0.147	0.98	0.024

由表 13 可得知，在隨機森林與 XGBoost 模型下，經過特徵欄位篩選後得到的績效都比篩選前有些微的下降。而 KNN 模型經過特徵欄位篩選後所得到的績效呈現明顯的上升。因此在三種模型對比之下，KNN 模型特徵篩選後有最佳績效。

四、結論

本實驗中，特徵重要度的選擇決定了實驗結果，由於排除了一些影響相對比較小的特徵屬性後，績效能夠獲得改善。第一個實驗的 Adult Data Set 資料集在線性迴歸中呈現的績效在特徵篩選前後的差異較無法凸顯出來，而在隨機森林及 XGBoost 等兩種演算法下，績效有明顯的改善。線性迴歸容易實現且構建迅速，對於小資料量、簡單的關係很有效，但是此資料集的筆數屬於大量資料，因此成效較不理想；隨機森林及 XGBoost 的特性是能夠處理大量的資料並且有良好的準確度，此資料集屬於大量資料，因此在實驗中也證明出這兩種演算法確實呈現了較佳的表現。第二個實驗的 Bike Sharing Data Set 資料集在隨機森林中與 XGBoost 兩種演算法下篩選後的績效比篩選前有著些微的下降，而在 KNN 中篩選欄位後比篩選前的績效有明顯的提升，因為 KNN 的方法主要是靠周圍有限的樣本，而不是依靠判別類域的方法來確定所屬類別，因此對於類域較多的交叉或重疊的資料及來說 KNN 較其他方法更為適合。

參考文獻

Chipeyown（2019 年 12 月 18 日）。*Pandas 重新設置索引*。Chinese Software Developer Network。https://blog.csdn.net/weixin_43745169/article/details/103593673

媽噠好氣哦（2020 年 7 月 9 日）。*基於人口普查數據的收入預測模型構建及比較分析（Python 數據分析分類器模型實踐）*。Chinese Software Developer Network。https://blog.csdn.net/weixin_39858881/article/details/107235037

機器學習與統計學（2020 年 7 月 4 日）。*Python 數據預處理：徹底理解標準化和歸一化 - 掘金*。稀土掘金。https://juejin.cn/post/6847902216196620295

Pyinvest（2019 年 11 月 3 日）。*[Python 實作] 隨機森林模型 Random Forest*。PyInvest https://pyecontech.com/2019/11/03/python_random_forest/

手撕機（2019 年 2 月 21 日）。*預測評價指標 RMSE、MSE、MAE、MAPE、SMAPE*。Chinese Software Developer Network。https://blog.csdn.net/guolindonggld/article/details/87856780

不停下腳步的烏龜（2020 年 3 月 3 日）。*【代碼模版】以 MAPE 為指標的評估函數模版*。Chinese Software Developer Network. https://blog.csdn.net/weixin_44680262/article/details/104630323

皮卡丘打排球（2021 年 10 月 3 日）。*Day 20 : 線性迴歸與羅吉斯迴歸*。IT 邦幫忙。https://ithelp.ithome.com.tw/articles/10272968

ssswill（2019 年 1 月 13 日）。*線性迴歸 RidgeCV,LassoCV 及迴歸權重重要性可視化*。Chinese Software Developer Network。https://blog.csdn.net/ssswill/article/details/86411009

10 程式中 (2021 年 9 月 22 日) 。*[Day 10]近朱者赤，近墨者白 - KNN* 。it 邦幫忙。https://ithelp.ithome.com.tw/articles/10272968

Chwang (2021 年 8 月 1 日) 。*Machine Learning- 交叉驗證(Cross Validation)-找到 KNN 中適合的 K 值 - Scikit Learn 一步一步實作教學* 。Medium。https://chwang12341.medium.com/machine-learning-%E4%BA%A4%E5%8F%89%E9%A9%97%E8%AD%89-cross-validation-%E6%89%BE%E5%88%B0knn%E4%B8%AD%E9%81%A9%E5%90%88%E7%9A%84k%E5%80%BC-scikit-

learn%E4%B8%80%E6%AD%A5%E4%B8%80%E6%AD%A5%E5%AF%A6%E4%BD%9C%E6%95%99%E5%AD%B8-4109bf470340