國立雲林科技大學資訊管理系

資料探勘

專案作業二

組員

M11123037 許禾諭

M11123043 謝永盛

M11123050 李京樺

M11123053 葉家豪

**摘要**

現今許多城市皆有引入共享自行車的服務，為了避免車站出現沒有自行車的狀況，在正確時間提供自行車，減少市民等待時間成為主要問題。本研究採用首爾自行車共享資料集預測當地不同因素對於自行車租借次數的影響，為了符合台灣的環境，本研究將不採納降雪量因素。本研究使用RandomForest、KNN和SVR演算法建構預測模型，透過 MAE、RMSE和MAPE三種績效指標得知，使用RandomForest演算法的績效最佳。

關鍵字：共享自行車、RandomForest、KNN、SVR

1. 緒論
2. 動機

目前許多城市都有引入共享自行車的服務，提供市民作為短程的交通工具，以提高市民的外出便利性，但人流較多的自行車站點，通常是供不應求，導致熱門的站點出現沒有自行車的狀況，因此如何在正確時間提供自行車，減少市民等待時間成為主要問題。

1. 目的

本研究計畫利用首爾自行車共享需求資料集來預測不同變因對於自行車租借次數的影響，例如:不同天氣，以及是否為假期都可能會造成自行車租借次數的影響，讓自行車租借公司可以根據不同的變因，對應不同的情況提前部屬自行車，以減少民眾等待的時間，提高消費者的滿意度。

1. 方法
2. 程式架構

載入所需的函式庫，再進行資料的前處理，處理需正規化的欄位並移出不必要的欄位，再將資料集分割為80%的訓練資料和20%的測試資料，接下來分別以RandomForest、KNN和SVR 演算法訓練模型，並且在使用MAE、RMSE和MAPE等方法來評估績效。

1. 執行程式的方法

下載github上已經預先處理好的資料集以及程式，使用Jupyter開發環境讀取檔案的路徑，確認檔案可成功讀取後，即可執行程式。

1. 實驗
2. 資料集
3. Adult Data Set

表 1  
人口普查收入資料集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 資料集資訊表 | | | | | |
| Data Set Characteristics: | Multivariate | Number of Instances: | 48842 | Area: | Social |
| Attribute Characteristics: | Categorical, Integer | Number of Attributes: | 14 | Date Donated | 1996-05-01 |
| Associated Tasks: | Classification | Missing Values? | Yes | Number of Web Hits: | 727918 |

表 2  
人口普查收入資料集-屬性資料表

|  |  |
| --- | --- |
| 屬性名稱 | 屬性資訊 |
| Age | continuous |
| Workclass | Private、Self-emp-not-inc、Self-emp-inc、Federal-gov、Local-gov、State-gov、Without-pay、Never-worked |
| Fnlwgt | continuous |

續下表

承上表

|  |  |
| --- | --- |
| Education | Bachelors、Some-college、11th、HS-grad、Prof-school、Assoc-acdm、Assoc-voc、9th、7th-8th、12th、Masters、1st-4th、10th、Doctorate、5th-6th、Preschool |
| Education-num | continuous |
| Marital-status | Married-civ-spouse、Divorced、Never-married、Separated、Widowed、Married-spouse-absent、Married-AF-spouse. |
| Occupation | Tech-support、Craft-repair、Other-service、Sales、Exec-managerial、Prof-specialty、Handlers-cleaners、Machine-op-inspct、Adm-clerical、Farming-fishing、Transport-moving、Priv-house-serv、Protective-serv、Armed-Forces |
| Relationship | Wife、Own-child、Husband、Not-in-family、Other-relative、Unmarried |
| Race | White、Asian-Pac-Islander、Amer-Indian-Eskimo、Other、Black |
| Sex | Female、Male |
| Capital-gain | continuous |
| Capital-loss | continuous |
| Hours-per-week | continuous |
| Native-country | United-States、Cambodia、England、Puerto-Rico、Canada、Germany、Outlying-US(Guam-USVI-etc)、India、Japan、Greece、South、China、Cuba、Iran、Honduras、Philippines、Italy、Poland、Jamaica、Vietnam、Mexico、Portugal、Ireland、France、Dominican-Republic、Laos、Ecuador、Taiwan、Haiti、Columbia、Hungary、Guatemala、Nicaragua、Scotland、Thailand、Yugoslavia、El-Salvador、Trinadad&Tobago、Peru、Hong、Holand-Netherlands |

1. Seoul Bike Sharing Demand Data Set

表 3  
首爾自行車共享需求資料集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 資料集資訊表 | | | | | |
| Data Set Characteristics: | Multivariate | Number of Instances: | 8760 | Area: | Computer |
| Attribute Characteristics: | Integer, Real | Number of Attributes: | 14 | Date Donated | 2020-03-01 |
| Associated Tasks: | Regression | Missing Values? | N/A | Number of Web Hits: | 72142 |

表 4  
首爾自行車共享需求資料集-屬性資料表

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 屬性資訊 | 屬性簡介 |
| 日期(Date) | Day-month-year | 2017/12/01~2018/11/30 |
| 租用次數 (Rented Bike Count) | Count of bikes rented at each hour | 每小時自行車的租用次數 |
| 小時(Hour) | Hour of he day | 每小時紀錄一次 |
| 溫度(Temperature) | Temperature in Celsius | 首爾當天溫度 |
| 濕度(Humidity) | % | 首爾當天濕度 |
| 風速(Wind speed) | m/s | 首爾當天風速 |
| 能見度(Visibility) | 10m | 首爾當天能見度 |
| 露點溫度(Dew point temperature) | Celsius | 首爾當天露點溫度 |
| 太陽輻射 (Solar Radiation) | MJ/m2 | 首爾當天太陽輻射 |
| 降雨量(Rainfall) | mm | 首爾當天降雨量 |
| 降雪量(Snowfall) | cm | 首爾當天降雪量 |
| 季節(Seasons) | Winter、Spring、Summer、Autumn | 春、夏、秋、冬 |
| 假日(Holiday) | Holiday、No holiday | 假日、非假日 |
| 工作日(Functioning Day) | NoFunc(Non Functional Hours)、Fun(Functional hours) | 工作日、非工作日 |

1. 前置處理
2. Adult Data Set

本研究將前置處理分成兩個階段，首先使用Microsoft Excel試算表軟體檢查eclass所提供的訓練和測試資料集，訓練資料集的總筆數為32561，測試資料集為16281，查詢後發現訓練資料集缺失筆數(messing value)為2399，測試資料集為1221，皆佔總筆數約7%，本研究認為缺失值不影響後續分析，故在此階段將其排除。在檔案讀入程式後，因資料集本身無欄位名稱，故根據UCI Dataset上的資訊將其加入欄位中。此外，本研究透過LabelEncoder將資料集內的文字進行轉換。

1. Seoul Bike Sharing Demand Data Set

本研究首先在Microsoft Excel試算表軟體中，將原資料集的”日期(Date)”和名目屬性轉為數字方式呈現(如表5)，後續將excel檔案匯入程式時，使用最小值最大值正規化(Min-Max Normalization)統一數值過大的屬性資料以利後續分析(如表6)，並且考量到臺灣降雪機率非常低，故移除原資料集的”降雪量(Snowfall)”屬性。

表 5  
首爾自行車共享需求資料集-excel前置處理

|  |  |
| --- | --- |
| 屬性名稱 | EXCEL前置處理 |
| 日期 | 將日期轉成數值，dd/mm/yyyy轉為1~365 |
| 季節 | 將文字轉成數值，春、夏、秋、冬轉為0~3 |
|  | 續下表 |
| 承上表 |  |
| 假日 | 將文字轉成數值，假日、非假日轉為1跟0 |
| 工作日 | 將文字轉成數值，工作日、非工作日轉為1跟0 |

表 6  
首爾自行車共享需求資料集-程式前置處理

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 程式前置處理(正規化區間[0,1]) | |
| 最大值 | 最小值 |
| 租用次數 | 3556 | 0 |
| 溫度（攝氏度） | 39.4 | -17.8 |
| 濕度 | 98 | 0 |
| 風速 | 7.4 | 0 |
| 能見度 | 2000 | 27 |
| 露點溫度 | 27.2 | -30.6 |
| 太陽輻射 | 3.52 | 0 |
| 降雨量 | 35 | 0 |

1. 實驗設計

在經過資料的前置處理後，將本研究資料集以80%和20%分割出訓練資料和測試資料，本研究將使用RandomForest、SVR和KNN三種不同的演算法來訓練模型，並且透過MAE、RMSE和MAPE三個績效指標來計算出不同演算法訓練的模型其績效如何，並排列出特徵的重要性，試刪除高特徵重要性的欄位以觀察對模型績效的影響。Adult資料集採用的演算法與上述相同，其中SVR對於大型資料集的擬合時間過長，故將SVR改為XGBoost。

1. 實驗結果
2. Adult Data Set
3. RandomForest 績效：0.1899。

計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：7.5198

RMSE：10.8564

MAPE：0.2737

1. 刪除最高的特徵重要性屬性age後 RandomForest 績效：0.1158。

原特徵重要性屬性如圖1所示，刪除最高的特徵重要性屬性age後，如圖2所示。計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：7.9012

RMSE：11.3425

MAPE：0.2817

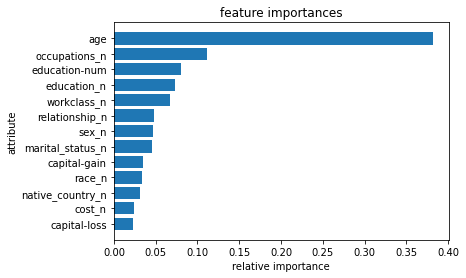


圖 1 RandomForest特徵重要性(前)

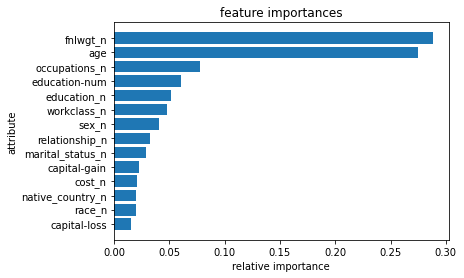


圖 2 RandomForest特徵重要性(後)

1. XGBoost 績效：0.2309。

計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：7.1758

RMSE：10.5783

MAPE：0.2552

1. 刪除最高的特徵重要性屬性sex後 XGBoost績效： 0.2438。

原特徵重要性屬性如圖3所示，刪除最高的特徵重要性屬性sex後，如圖4所示。計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：7.1565

RMSE：10.4892

MAPE：0.2648

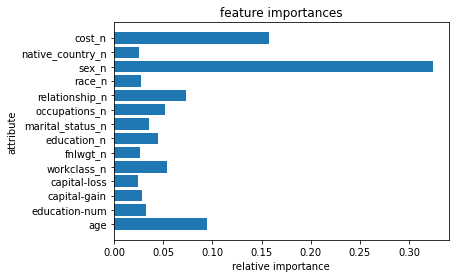


圖 3 XGBoost特徵重要性(前)

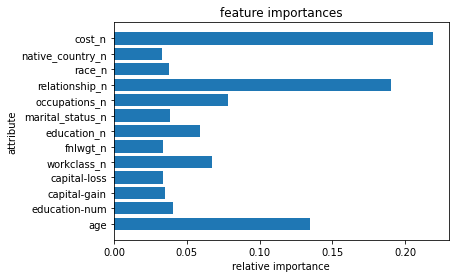


圖 4 XGBoost特徵重要性(前)

1. KNN 績效： -0.7606。

計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：10.9855

RMSE：16.0055

MAPE：0.3769

1. 刪除最高的特徵重要性屬性fnlwgt後 KNN績效： -0.4352。

原特徵重要性屬性如圖5所示，刪除最高的特徵重要性屬性fnlwgt後，如圖6所示。計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：9.5995

RMSE：14.4510

MAPE：0.3246

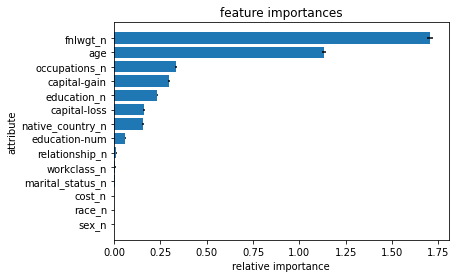


圖 5 KNN特徵重要性(前)

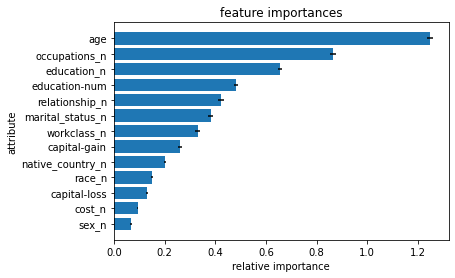


圖 6 KNN特徵重要性(後)

1. Seoul Bike Sharing Demand Data Set
2. RandomForest 績效：0.8804。

計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：0.0372

RMSE：0.0628

MAPE：1067276093566.5092

1. 刪除最高的特徵重要性屬性Temperature後 RandomForest績效： 0.8738。

原特徵重要性屬性如圖7所示，刪除最高的特徵重要性屬性Temperature後，如圖8所示。計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：0.0377

RMSE：0.0638

MAPE：454602378522.7820

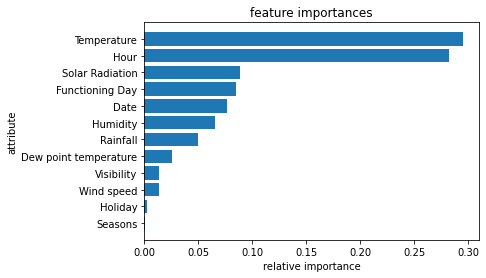


圖 7 RandomForest特徵重要性(前)

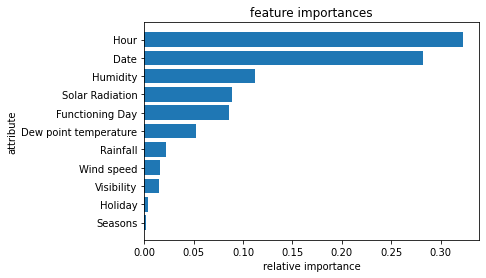


圖 8 RandomForest特徵重要性(後)

1. SVR 績效：0.5864。

計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：0.0984

RMSE：0.1162

MAPE：24876902542030.0350

1. 刪除最高的特徵重要性屬性Hour後 SVR績效： 0.4985。

原特徵重要性屬性如圖9所示，刪除最高的特徵重要性屬性Hour後，如圖10所示。計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：0.1079

RMSE：0.1280

MAPE：29195222333602.4700

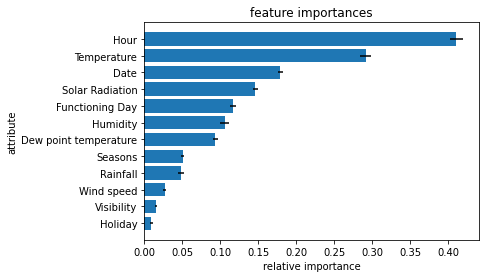


圖 9 SVR特徵重要性(前)

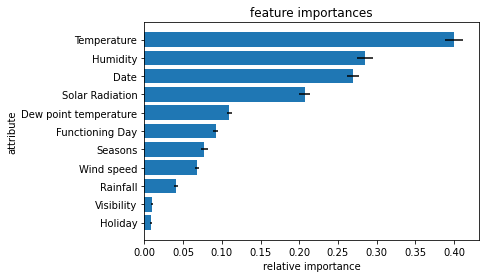


圖 10 SVR特徵重要性(後)

1. KNN 績效：0.7983。

計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：0.0497

RMSE：0.0811

MAPE：2338504475448.6055

1. 刪除最高的特徵重要性屬性Date後 KNN績效：0.6852。

原特徵重要性屬性如圖11所示，刪除最高的特徵重要性屬性Date後，如圖12所示。計算預測後的測試資料與真實的測試資料的績效，以MAE、RMSE和MAPE為績效指標，分別為：

MAE：0.0609

RMSE：0.1014

MAPE：1673458380422.0063



圖 11 KNN特徵重要性(前)

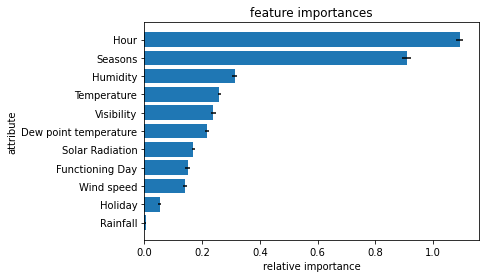


圖 12 KNN特徵重要性(後)

1. 結論

根據本研究實驗結果表明，利用RandomForest、KNN和SVR演算法來預測人口普查收入資料集中的每週時數和預測首爾共享自行車資料集中的租借數量，由 RandomForest演算法測試出來的 MAE、RMSE和MAPE績效是最佳的。我們刪除各個資料集中特徵重要性最高的欄位後發現，除了大多發生模型績效降低之外，各項評估的誤差指標大多有提高的趨勢。因此可以得出刪除重要的特徵會降低模型的表現。

1. 參考文獻
2. Mike Ku。2020年12月6日。[Pandas教學]5個實用的Pandas讀取Excel檔案資料技巧。Learncodewithmike。<https://www.learncodewithmike.com/2020/12/read-excel-file-using-pandas.html>
3. Tom。2018年12月31日。資料的正規化(Normalization)及標準化(Standardization)。Aifreeblog。<https://aifreeblog.herokuapp.com/posts/54/data_science_203/>
4. scikit-learn。RandomForestRegressor。 <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
5. scikit-learn。KNeighborsRegressor。<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>
6. scikit-learn。SVR 。<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>
7. scikit-learn。 mean\_absolute\_percentage\_error 。 <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html>
8. scikit-learn。mean\_squared\_error 。<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html>
9. scikit-learn。mean\_squared\_error 。 <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html>
10. Tango棒棒。2021年11月27日。Pandas讀取csv文件不要把第一行自動設置成表頭。CSDN-專業開發者社區。<https://reurl.cc/EX1vrk>
11. Scikit-learn course。Feature importance。<https://inria.github.io/scikit-learn-mooc/python_scripts/dev_features_importance.html#feature-importance-by-permutation>
12. [Jason Brownlee。2021年3月12日。XGBoost for Regression。 Machine Learning Mastery](https://machinelearningmastery.com/author/jasonb/) <https://machinelearningmastery.com/xgboost-for-regression/>
13. Piotr Płoński。2020 年 8 月 17 日。使用 Python 以 3 種方式計算 Xgboost 特徵重要性。MLJAR。 <https://mljar.com/blog/feature-importance-xgboost/>