國立雲林科技大學資訊管理系

資料探勘-作業二

Department of Information Management

National Yunlin University of Science & Technology

Assignment

成人資料集與波士頓房價分析

Adult Dataset And Boston Housing Dataset Analysis

M11223018黃裕鳴

M11223020游棨翔

M11223043葉桔良

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國112年11月

October 2023

**摘要**

這份研究的主要目標在於比較不同演算法在相同資料集上的表現。經過充分的資料前置處理和模型訓練，結果顯示 XGBoost 演算法在測試集上取得了卓越的效果。該模型成功應用於每週工時的預測，為不同背景下的工時預測提供了實用性參考，可作為未來相關工作的指導方針。

此外，研究進一步深入探討了 XGBoost 演算法的效能。使用 K-fold cross validation 預測績效，並透過 XGBoost 的 feature importance 計算特徵重要性，移除多餘欄位以觀察前後績效的差異。

由於本研究所使用的資料集相對陳舊，未來的研究可能應考慮採用更現代的數據以提高預測模型的準確性。儘管如此，這項研究仍然為人們提供了深入了解每週工時的途徑，並演示了如何運用不同迴歸模型進行相關預測的實例。

**關鍵字**：每週工時、KNN、SVR、RandomForest、XGBoost、k-fold、feature importance

**一、緒論**

本研究旨在探討並比較KNN、SVR、RandomForest和XGBoost四種演算法在預測成人資料集中每週工時和波士頓房價資料集中自住房價格的表現。透過詳細的資料分析和模型選擇，本研究致力於找出最適合不同預測目標的演算法。此研究的成果將有助於選擇適當的預測模型，提升相關應用的效能。

1. **動機**

本研究的動機源於對KNN、SVR、RandomForest和XGBoost四種演算法在預測成人資料集的應用效能的好奇。透過比較不同演算法的表現，期望深入了解它們在不同領域的適用性，並為選擇最適當的預測模型提供更多想法。這有助於優化相關應用的效能，提高預測準確性。

隨後將深入研究XGBoost演算法在波士頓房價資料集的表現，使用k-fold進行預測績效評估，最終探討特徵的重要性。計劃透過移除部分欄位，觀察是否能夠進一步提升預測準確度。

1. **目的**

本研究主要探討不同機器學習演算法在預測成人資料集和波士頓房價資料集上的表現。具體而言，本研究將使用KNN、SVR、RandomForest和XGBoost四種演算法，分別對成人資料集中的每週工時和波士頓房價資料集中的自住房價格進行預測。透過比較不同演算法的效能，本研究旨在找出最適合不同預測目標的模型。

接著深入研究XGBoost模型，透過k-fold來預測此模型的績效，去觀察不同K值間的績效，最後使用特徵重要性將多餘欄位移除，來確認是否提高準確率，可讓研究者對於日後的研究模型有著更深的理解。

**二、方法**

在這項實驗中，本研究致力於預測成人資料集中每週工時和波士頓房價資料集中的自住房的平均價格。

首先對成人資料集進行了詳細的探索性資料分析，以了解數據的結構和特點。這包括檢查數據中的缺失值、探索性統計和視覺化分析。接著進行了必要的前置處理步驟。對於成人資料集，本研究進行了資料清理，處理了缺失值，並進行了標準化和獨熱編碼。

對於波士頓房價資料集，同樣進行了探索性資料分析，檢查數據的分佈和相關性。注意到了一些重要的特徵，這些特徵後來在模型訓練中被重點考慮。

在模型選擇方面，本研究選擇了KNN、SVR、RandomForest和XGBoost這四種演算法，這些演算法在預測不同目標上都有良好的表現。對於KNN，通過實驗調整K值，找到了最佳的K值。對於SVR，調整了正規化參數、最大迭代次數和內核緩存大小。對於RandomForest，測試了不同的深度值。而對於XGBoost，選擇了不使用超參數調整，因為默認參數已經在實驗中顯示出良好的性能。

最後，本研究評估了模型的性能，使用MAPE、RMSE和R2三種指標來量化預測的準確性。這些指標提供了對每個模型在預測上的不同優勢的觀察。

1. **實驗**
2. **資料集**
3. 成人資料集介紹

此資料集從美國1994年人口普查資料庫中抽取而來，共有48842筆資料，此資料已經先前被分為訓練資料（adult.data）與測試資料（adult.test），主要用途為什麼變項會影響到每週工時，以下表格將介紹Adult資料集欄位。

**表1**

Adult資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | age | Continuous |
| 1 | workplace | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never worked |
| 2 | fnlwt | continuous |
| 3 | 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 |
| 4 | education-num | continuous |
| 5 | marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |

（續）

|  |  |  |
| --- | --- | --- |
| 6 | 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 |
| 7 | relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| 8 | race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| 9 | sex | Female, Male |
| 10 | capital-gain | continuous |
| 11 | capital-loss | continuous |
| 12 | hours-per-week | continuous |
| 13 | 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 |
| 14 | income | <=50K,>50K |

1. 波士頓房價資料集介紹

波士頓房價資料集最初由Harrison, D. 和 Rubinfeld, D.L.並在1978年釋出，共有506筆資料，包含13個與波士頓地區不同方面相關的特徵，如犯罪率、住宅平均房間數等。其目標是預測自住房的平均房價。

**表2**

Boston Housing資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | CRIM | Continuous |
| 1 | ZN | Continuous |
| 2 | INDUS | Continuous |
| 3 | CHAS | Continuous |
| 4 | NOX | Continuous |
| 5 | RM | Continuous |
| 6 | AGE | Continuous |
| 7 | DIS | Continuous |
| 8 | RAD | Continuous |
| 9 | TAX | Continuous |
| 10 | PTRATIO | Continuous |
| 11 | B | Continuous |
| 12 | LSTAT | Continuous |
| 13 | MEDV | Continuous |

1. **成人資料集訓練資料**

檔案名稱：adult.data(訓練資料)

原始資料筆數：32560筆

**表3**

原始訓練資料部分資料

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **n** | **32559** | **32560** |
| **age** | 39 | 50 | … | 22 | 52 |
| **workclass** | State-gov | Self-emp-not-inc | … | Private | Self-emp-inc |
| **fnlwgt** | 77516 | 83311 | … | 201490 | 287927 |
| **education** | Bachelors | Bachelors | … | HS-grad | HS-grad |
| **education-num** | 13 | 13 | … | 9 | 9 |
| **marital-status** | Never-married | Married-civ-spouse | … | Never-married | Married-civ-spouse |
| **occupation** | Adm-clerical | Exec-managerial | … | Adm-clerical | Exec-managerial |
| **relationship** | Not-in-family | Husband | … | Own-child | Wife |
| **race** | White | White | … | White | White |
| **sex** | Male | Male | … | Male | Female |
| **capital-gain** | 2174 | 0 | … | 0 | 15024 |
| **capital-loss** | 0 | 0 | … | 0 | 0 |
| **hours-per-week** | 40 | 13 | … | 20 | 40 |
| **native-country** | United-States | United-States | … | United-States | United-States |
| **income** | <=50K | <=50K | … | <=50K | >50K |

1. **成人資料集測試資料**

檔案名稱：adult.test(測試資料)

原始資料筆數：16282筆

**表4**

原始成人資料集測試資料部分資料

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **n** | **16281** | **16282** |
| **age** | 25 | 38 | … | 44 | 35 |
| **workclass** | Private | Private | … | Private | Self-emp-inc |
| **fnlwgt** | 226802 | 89814 | … | 83891 | 182148 |
| **education** | 11th | HS-grad | … | Bachelors | Bachelors |
| **education-num** | 7 | 9 | … | 13 | 13 |
| **marital-status** | Never-married | Married-civ-spouse | … | Divorced | Married-civ-spouse |
| **occupation** | Machine-op-inspct | Farming-fishing | … | Adm-clerical | Exec-managerial |
| **relationship** | Own-child | Husband | … | Own-child | Husband |
| **race** | Black | White | … | Asian-Pac-Islander | White |
| **sex** | Male | Male | … | Male | Male |
| **capital-gain** | 0 | 0 | … | 5455 | 0 |
| **capital-loss** | 0 | 0 | … | 0 | 0 |
| **hours-per-week** | 40 | 50 | … | 40 | 60 |
| **native-country** | United-States | United-States | … | United-States | United-States |
| **income** | <=50K. | <=50K. | … | <=50K. | >50K. |

1. **波士頓房價資料集**

檔案名稱：HousingData.csv

原始資料筆數：507筆

**表5**

波士頓房價資料集部分資料

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **n** | **506** | **507** |
| **CRIM** | 0.00632 | 0.02731 | … | 0.10959 | 0.04741 |
| **ZN** | 18 | 0 | … | 0 | 0 |
| **INDUS** | 2.31 | 7.07 | … | 11.93 | 11.93 |
| **CHAS** | 0 | 0 | … | 0 | 0 |
| **NOX** | 0.538 | 0.469 | … | 0.573 | 0.573 |
| **RM** | 6.575 | 6.421 | … | 6.794 | 6.03 |
| **AGE** | 65.2 | 78.9 | … | 89.3 | NA |
| **DIS** | 4.09 | 4.9671 | … | 2.3889 | 2.505 |
| **RAD** | 1 | 2 | … | 1 | 1 |
| **TAX** | 296 | 242 | … | 273 | 273 |
| **PTRATIO** | 15.3 | 17.8 | … | 21 | 21 |
| **B** | 396.9 | 396.9 | … | 393.45 | 396.9 |
| **LSTAT** | 4.98 | 9.14 | … | 6.48 | 7.88 |
| **MEDV** | 24 | 21.6 | … | 22 | 11.9 |

1. **前置處理**

**3.3.1.** **前置處理-****Adult Dataset**

**圖1**

前置處理流程圖

1. 資料清理：將 fnlwgt 欄位與 education 欄位移除。
2. 處理缺失值：把資料內有"?"的資料取代為NaT。接下來將缺失值補齊，名目資料填眾數，數值資料填平均值。
3. 二值化：將income欄位大於50K更改為1，小於等於50K更改為0。
4. Z-Score：也稱為標準化值，各個樣本在經過轉換後，通常在正、負五到六之間不等。
5. One-hot Encoding：針對資料集中 'workclass', 'marital-status', 'occupation', 'relationship', 'race', 'sex', 'native-country' 欄位進行獨熱編碼。
6. 刪除重複列：將資料庫中冗餘數據做調整，增進數據的一致性。

**3.3.2.** **前置處理-** **Boston Housing Dataset**

**圖2**

前置處理流程圖

1. 處理缺失值：將資料內為NA的資料皆填上平均值。
2. Z-Score：也稱為標準化值，各個樣本在經過轉換後，通常在正、負五到六之間不等。但對於XGBoost演算法來說對數差距範圍不敏感，故是否使用影響不大。

**3.3.3. 未移除離群值之原因**

在應用四種演算法時，嘗試將離群值定義為1.5倍的四分位距（IQR）並移除，造成資料筆數減少約7000筆。這種處理方式對於模型的性能沒有正向效果，反而導致數值績效下降。因此，本研究最終選擇不使用這種離群值移除方法。

1. **實驗設計**

實驗設計的第一部分使用了KNN、SVR、RandomForest和XGBoost四種方式來預測成人資料集中的每週工時。第二部分使用XGBoost預測波士頓房價資料集中的自住房的平均房價。以下顯示本研究的實驗設計。

**3.4.1. 第一部分測試績效流程步驟**

下圖(圖3)顯示多種不同演算法的流程步驟，需測試不同演算法之績效。以下將解釋不同演算法之超參數數值推導步驟。

**圖3**

第一部分流程圖

載入模型

訓練模型

評估績效

KNN

比較差異

SVR

RandomForest

XGBoost

1. KNN的K值計算方式

如下表(表6)，透過逐步增加數值，K值的計算方式顯示部分標誌性數據，若K大於255或K小於255績效會明顯下降。最終得出K=255時呈現最佳績效。

**表6**

K值與績效指標相對應數值

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **MAPE** | **RMSE** | **R2** |
| 30 | 0.327 | 11.339 | 0.205 |
| 100 | 0.326 | 11.172 | 0.228 |
| 250 | 0.329 | 11.146 | 0.230 |
| 255 | 0.329 | 11.145 | 0.232 |
| 500 | 0.327 | 11.162 | 0.230 |

1. SVR設定參數指標

linear\_svr 是支援向量機回歸（SVR）模型的實例，使用 C=200 作為正規化參數，控制模型複雜度。max\_iter = 1000000 指定最大迭代次數，而 cache\_size = 1000 則是指定內核緩存大小，用於在訓練期間存儲計算結果以提高效能。

1. 隨機森林的深度計算方式

在下表(表7)中，透過逐步減少數值的方式，深度的計算方式展示了部分標誌性數據。若深度超過10或低於該值，績效將明顯下降。結論顯示，在深度等於10的情況下，模型呈現最佳績效。

**表7**

深度與績效指標相對應數值

|  |  |  |  |
| --- | --- | --- | --- |
| **深度** | **MAPE** | **RMSE** | **R2** |
| 5 | 0.325 | 11.203 | 0.224 |
| 10 | 0.315 | 10.959 | 0.258 |
| 15 | 0.314 | 11.025 | 0.249 |

1. XGBoost不使用超參數原因

不使用XGBoost超參數的原因為對於簡單問題或特定數據集而言，默認參數已足夠，超參數調整對性能提升影響不大，甚至會導致績效變差，故不考慮使用。

**3.4.2. 第二部分測試績效流程步驟**

下圖(圖4)顯示了預測波士頓房價資料集中MEDV欄位的流程步驟。該流程使用XGBoost演算法進行模型訓練，然後以K-fold中的K值為5來評估預測績效。接著，通過識別並選取重要特徵，將多餘且不重要的欄位從模型中刪除，輸出最終結果。

**圖4**

第二部分流程圖

載入模型

訓練模型

XGBoost

K設定5

評估績效

選取重要特徵

輸出結果

1. **實驗結果**

**3.5.1. 第一部分實驗結果**

下表(表8)共有四種模型對成人資料集的績效進行了評估。其中，RandomForest演算法表現最佳，但相對花費了18.746秒的時間。XGBoost演算法的績效僅次於RandomForest，兩者之間僅相差0.002，然而，XGBoost僅需1.939秒的時間，因此可視為最佳選擇。

**表8**

各種演算法間不同績效比較表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAPE | RMSE | R2 | time |
| KNN | 0.329 | 11.141 | 0.232 | 9.261s |
| SVR | 0.311 | 11.205 | 0.224 | 328.399s |
| RandomForest | 0.314 | 11.025 | 0.249 | 18.746s |
| XGBoost | 0.317 | 11.037 | 0.247 | 1.939s |

**3.5.2. 第二部分實驗結果**

下表(表9)呈現了針對波士頓房價資料集使用XGBoost演算法進行模型訓練的結果。評估採用K-fold中的K值為5，列出每一個fold的預測績效以及5個folds的平均績效。顯示平均績效達到88.4%，達到相當高水準。

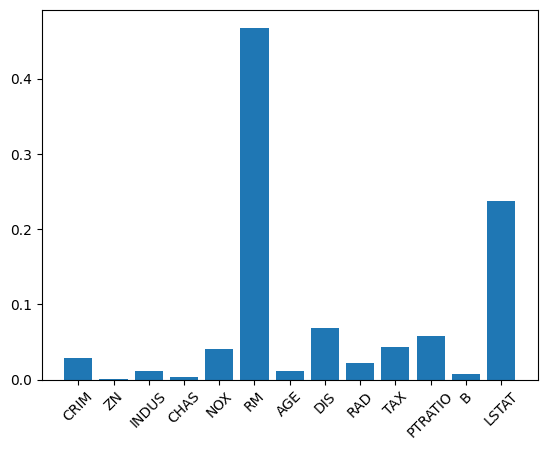
**表9**

各種演算法間不同績效比較表

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MAPE** | **RMSE** | **R2** |
| **1** | 0.107 | 3.083 | 0.868 |
| **2** | 0.096 | 2.915 | 0.917 |
| **3** | 0.127 | 3.429 | 0.871 |
| **4** | 0.117 | 2.936 | 0.902 |
| **5** | 0.111 | 3.041 | 0.864 |
| **avg** | 0.112 | 3.081 | 0.884 |

下圖(圖5)為計算特徵重要性的長條圖，很明顯可以看到前六特徵的重要性有’NOX’, ‘RM’, ‘DIS, ‘TAX’, PTRATIO, ‘LSTAT’，將此六項欄位留存，將其餘六項欄位做刪除。會發現結果(表10)內的R2值提升了0.1%，故刪除不重要欄位之後績效明顯提升。圖形化結果(圖6)可發現預測成功率是相當高的，都相當貼近中心紅虛線。

**圖5**

特徵重要性長條圖

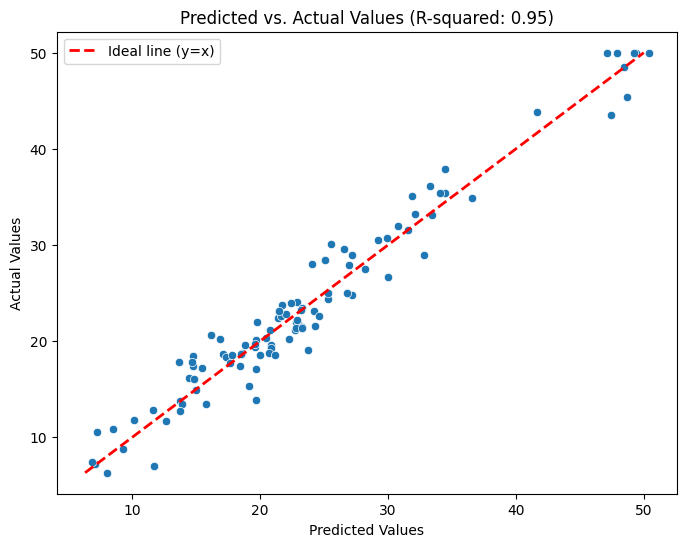
**表10**

特徵重要性長條圖

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAPE | RMSE | R2 |
| 刪除前 | 9.021 | 2.151 | 0.954 |
| 刪除後 | 8.711 | 2.148 | 0.955 |

**圖6**

刪除不重要特徵欄位之績效圖



**四、結論**

本實驗以成人資料集和波士頓房價資料集為基礎，採用KNN、SVR、RandomForest、XGBoost四種不同機器學習演算法進行預測與評估。

在成人資料集方面，RandomForest演算法表現最佳，但執行時間較長；而XGBoost演算法在績效上僅次於RandomForest，但執行時間明顯較短，因此在此實驗中被視為最佳選擇。超參數的調整對各演算法的表現具有顯著影響。

在波士頓房價資料集方面，使用XGBoost演算法進行模型訓練，以K-fold中的K值為5來評估預測績效，平均績效達到88.4%。特徵重要性分析顯示前六個重要特徵，並刪除不重要特徵後模型的R2值有所提升。

總體來說，實驗強調了模型選擇的重要性，並突顯了XGBoost在處理不同資料集上的優越性能。適當的資料前置處理和特徵處理對於模型性能至關重要。雖然嘗試過移除離群值，但在這次實驗中未帶來正向效果，反而導致模型績效下降，最終選擇不使用這種離群值移除方法。因此，實驗結果提供了在實際應用中指導模型選擇和調整的有價值見解。

**參考文獻**

Lenskit (2018年06月23日)。Python机器学习及实践——基础篇9（SVM回归）。https://blog.csdn.net/chenKFKevin/article/details/80786681

co1oratura (2019年12月16日)。利用K近邻（回归）KNeighborsRegressor进行回归训练并预测。https://doi.org/10.24432/C5XW20.