國立雲林科技大學  
資訊管理研究所

資料探勘專案作業2

指導教授：許中川教授  
學生：M11123039 廖顯億  
 M11123046 周子懷  
 M11123052 賴俊佑  
 M11123062 陳靖穎

Data Mining Homework: The Adult Dataset Mining

摘要

Covid-19疫情以及烏俄戰爭影響，已經造成全球經濟受到影響，供應鏈排程、隔離政策、通貨膨脹、消費緊縮、原料價格上漲等等影響，公司收入也受到嚴重影響。為了控制成本，透過利用KNN、Random Forest、SVR預測每週工作小時，發現Random Forest效果顯著。

**關鍵字**：Covid-19、烏俄戰爭、KNN、Random Forest、SVR

1. **緒論**

2020年Covid-19疫情席捲而來，除了對民眾健康造成影響外，對產業也造成衝擊，各種隔離政策及防疫措施，也改變了生活習慣。再加上2022年烏俄戰爭，全球供應鏈混亂，進而造成通貨膨脹問題，原料物價波動大，嚴重影響公司訂單排程。

* 1. **動機**

因通貨膨脹及市場緊縮問題，各家客戶都將交貨日延後，造成公司突然沒有訂單的情形發生，為了減緩公司支出，除了硬體成本之外，人事成本也是重大支出之一，因此需要瞭解每週工作小時。

* 1. **目的**

透過本公司的Adult資料集，瞭解人事成本，已達到公司最好的成本控制，透過計算預測每週工作小時，預測一週可以貢獻的時數作為調整人力成本支出之依據，以便推估人員時間稼動率。

1. **方法**

首先進行資料檔案性質轉換，再匯入資料。進行分析前，對資料進行前處理，因資料集中對於教育程度education與educational-num兩個屬性意思相同，所以將education欄位拋棄。再來也拋棄無用資料capital-gain、capital-loss。此外workclass以及native-country欄位會有缺失值，所以workclass跟及native-country欄位出現缺失值即拋棄。本研究針對workclass, marital-status, occupation, relationship, race, gender, native-country等欄位進行類別資料處理轉換成數值資料，最後進行KNN、Random Forest、SVR分析。

1. 實驗
2. 資料集

* 資料集名稱：Adult
* 資料筆數：48842
* 屬性數量：15

表1 Adult資料及屬性簡介

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 型態 | 尺度 |
| age | Int | 比例 |
| workclass | String | 名目 |
| fnlwgt | Int | 比例 |
| education | String | 順序 |
| educational-num | Int | 區間 |
| marital-status | String | 名目 |
| occupation | String | 名目 |
| relationship | String | 名目 |
| race | String | 名目 |
|  |  | （續下表） |
| gender | String | 名目 |
| capital-gain | Int | 比例 |
| capital-loss | Int | 比例 |
| hours-per-week | Int | 比例 |
| native-country | String | 名目 |
| income | String | 順序 |

1. 前置處理

首先因為education與educational-num兩個屬性意思相同，所以刪除education，僅使用educational-num進行預測，也拋棄無用資料capital-gain、capital-loss。再來因Workclass以及native-country欄位會有缺失值，所以先將資料集中Workclass及native-country缺失值的資料刪除避免訓練有問題。接著將workclass、marital-status、occupation、relationship、race、gender、native-country等類別型的特徵屬性進行標準化。

1. 實驗設計

一開始先進行預處理，將一些不必要之資訊進行刪除，並且利用KNN、SVR、Random forest 進行回歸預測，並計算出其特徵的重要性以及MAE、MAPE、RASE進行評估。

1. 實驗結果

表2 Random Forest之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | .386 |
| MAPE | .674 |
| RMSE | .961 |

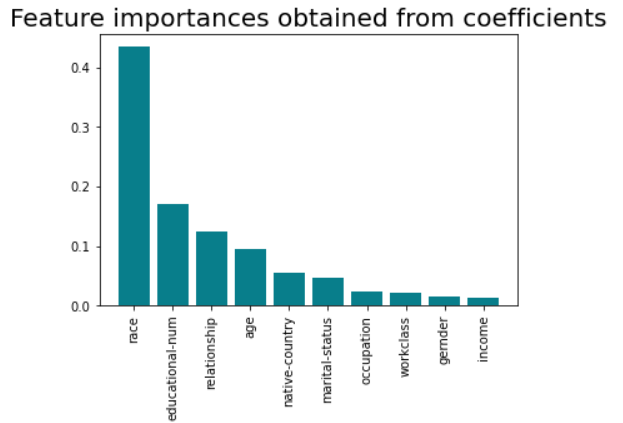


圖1 Random forest之特徵重要性

表3 去除掉高重要性之Random Forest之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | .086 |
| RMSE | .174 |
| MAPE | .421 |

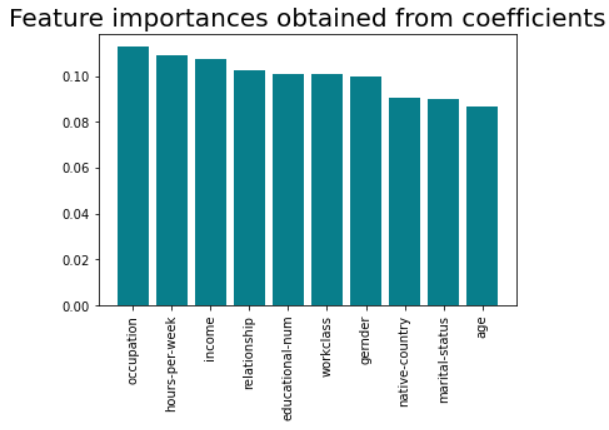


圖2 Random forest之特徵重要性

表4 KNN之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | .425 |
| RMSE | .664 |
| MAPE | .451 |

表5 SVR之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | .556 |
| RMSE | .849 |
| MAPE | .924 |

1. 結論

以KNN、SVR、Random forest回歸預測hours-per-week之數值，在其中Random forest去除掉Race(特徵重要值最高)後各數值皆有降低，而其他兩個沒有辦法計算特徵重要值，故單純進行回歸計算，各績效為Random forest最為顯著。

參考文獻

Imam Muhajir（2019年04月20日）。K-Neighbors Regression Analysis in Python。Medium。https://medium.com/analytics-vidhya/k-neighbors-regression-analysis-in-python-61532d56d8e4

Ronny Kohavi and Barry Becker (1996)。Adult。UCI Machine Learning Repository。<https://archive.ics.uci.edu/ml/datasets/Census+Income>

rosefunR（2017年12月18日）。随机森林random forest及python实现。CSDN。<http://t.csdn.cn/OQ95m>

李振麟（2022年07月01日）。全球性通膨危機大流行。聯合新聞網。<https://udn.com/news/story/6853/6429232>

中央社（2021年08月22）。疫情影響有多大，如何改變你的生活？3大數據揭露了。遠見雜誌。<https://www.gvm.com.tw/article/81845>

李明軒（2022年09月02日）。從烏俄戰爭看戰爭對經濟的影響。獨立評論。<https://opinion.cw.com.tw/blog/profile/52/article/12698>

关小羽（2020年04月19日）。SVR入门介绍（Python代码）。知呼。https://zhuanlan.zhihu.com/p/130974280

Data Mining Homework: Bias correction of numerical prediction model temperature forecast Data Set

摘要

隨著地球暖化的因素，天氣越來越不穩定，人們出門前挑選衣服時都會先查看天氣再決定今天的穿著，深怕沒注意到氣溫反而感冒。本研究預計以韓國氣象廳的數據作為資料的訓練集以及測試集，透過演算法建構數值預測模型，比較預測結果是否有越來越準確，幫助人們順利掌握氣候的變化。

**關鍵字**：極端氣候、氣象預報、KNN、SVR、Random Forest

1. **緒論**

在機器學習的領域中，常常利用演算法建構數值預測模型，比較調整前後績效的差異。本研究選用Bias correction of numerical prediction model temperature forecast資料集做預測模型，期望能比較出其中差異。

* 1. **動機**

現在這個極端氣候越來越嚴重的時代，早上還是大太陽但下午就下雪的現象越來越常見，人們需要提早預警因此需要借助氣象預報增添衣物。因此本研究利用資料集做預測比較天氣預報的預測結果是否越來越精準。

* 1. **目的**

為了解決上述動機發現的問題，本研究的目的為得知氣象預報預測結果的準確性是否有提高，透過本研究的預測，能夠幫助社會大眾更準確的穿衣，適時的增添衣物，尤其在疫情期間，更加需要照顧好自己的身體，避免因為免疫力下降染上感冒。

1. **方法**

首先進行資料檔案性質轉換並匯入資料後，對資料進行前處理，將資料做標準化處理、刪除遺失值，並針對Date資料類別進行數值化處理，最後進行調整前與調整後之MSE、RMSE和MAPE分析比較。

1. **實驗**
2. **資料集**

* 資料集名稱：Bias correction of numerical prediction model temperature forecast
* 資料筆數：7750
* 屬性數量：25

表1 Bias correction of numerical prediction model temperature forecast資料集屬性簡介

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 型態 | 尺度 |
| station | Int | 名目 |
| Date | String | 名目 |
| Present\_Tmax | Int | 名目 |
| Present\_Tmin | Int | 名目 |
| LDAPS\_RHmin | Int | 名目 |
| LDAPS\_RHmax | Int | 名目 |
| LDAPS\_Tmax\_lapse | Int | 名目 |
| LDAPS\_Tmin\_lapse | Int | 名目 |
| LDAPS\_WS | Int | 名目 |
| LDAPS\_LH | Int | 名目 |
| LDAPS\_CC1 | Int | 名目 |
| LDAPS\_CC2 | Int | 名目 |
|  |  | （續下表） |
| LDAPS\_CC3 | Int | 名目 |
| LDAPS\_CC4 | Int | 名目 |
| LDAPS\_PPT1 | Int | 名目 |
| LDAPS\_PPT2 | Int | 名目 |
| LDAPS\_PPT3 | Int | 名目 |
| LDAPS\_PPT4 | Int | 名目 |
| lat | Int | 名目 |
| lon | Int | 名目 |
| DEM | Int | 名目 |
| Slope | Int | 名目 |
| Solar radiation | Int | 名目 |
| Next\_Tmax | Int | 名目 |
| Next\_Tmin | Int | 名目 |

1. **前置處理**

因Date為類別型的特徵屬性，因此將其轉為數值型資料。

1. **實驗設計**

一開始進行調整前績效預測，接著刪除重要性最高之特徵，再分別檢視調整後之績效，比較各預測績效是否有提升。

1. **實驗結果**

表2 KNN之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | 8.728 |
| MAPE | 2.954 |
| RMSE | 0.072 |

表 3 去除最高重要性特徵之KNN正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | 9.215 |
| MAPE | 3.036 |
| RMSE | 0.073 |

表 4 SVR之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | 9.525 |
| MAPE | 3.086 |
| RMSE | 0.087 |

表 5 去除最高重要性特徵之SVR正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | 9.534 |
| MAPE | 3.088 |
| RMSE | 0.087 |

表 6 Random Forest之正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | .992 |
| MAPE | .996 |
| RMSE | .025 |

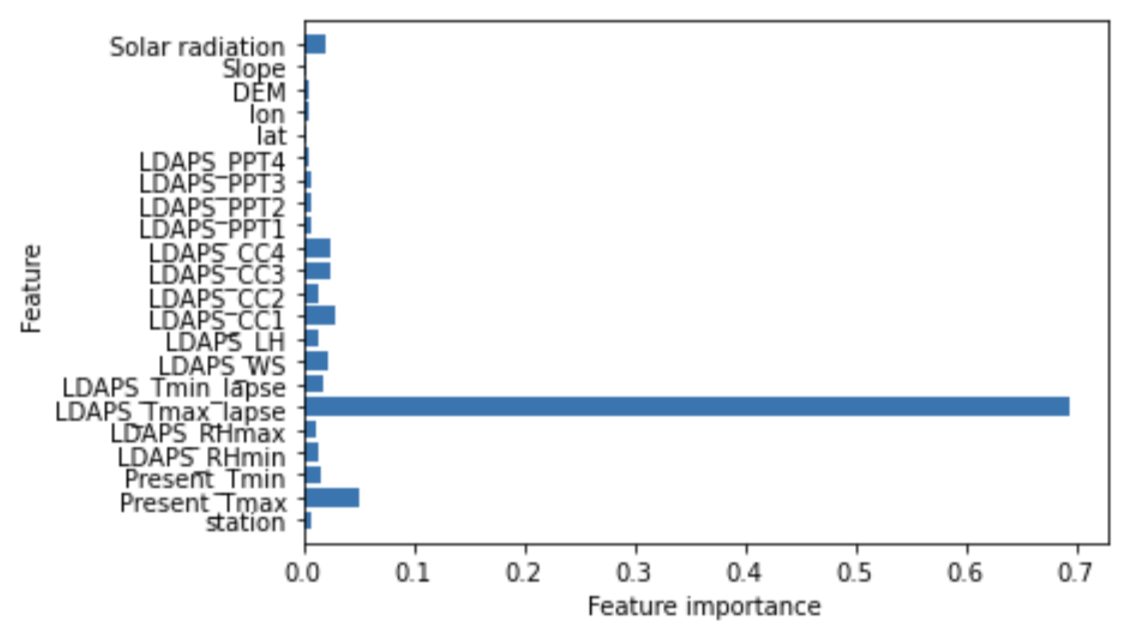
****

圖 1 Random forest之特徵重要性

表 7 去除最高重要性特徵之Random Forest正確率

|  |  |
| --- | --- |
|  | 訓練正確率 |
| MAE | .993 |
| MAPE | .997 |
| RMSE | .024 |

1. **結論**

本研究以KNN、SVR、Random forest回歸預測LDAPS\_Tmax\_lapse之數值，而三種演算法調整前與調整後無顯著差異，三者績效無顯著差異。

**參考文獻**

andy6804tw (2020, September 29). 12.KNN(Regression).ipynb. GitHub. [https://github.com/andy6804tw/2020-12th-ironman/blob/master/12.KNN(迴歸器)/12.KNN(Regression).ipynb](https://github.com/andy6804tw/2020-12th-ironman/blob/master/12.KNN(%E8%BF%B4%E6%AD%B8%E5%99%A8)/12.KNN(Regression).ipynb)

andy6804tw (2020, September 27). 14.SVR(Regression).ipynb. GitHub. <https://github.com/andy6804tw/2020-12th-ironman/blob/master/14.SVR(迴歸器)/14.SVR(Regression).ipynb>

andy6804tw (2020, October 4). 19.隨機森林(Regression).ipynb. GitHub. <https://github.com/andy6804tw/2020-12th-ironman/blob/master/19.隨機森林(迴歸器)/19.隨機森林(Regression).ipynb>

Dongjin Cho and Cheolhee Yoo (2020)。Bias correction of numerical prediction model temperature forecast。UCI Machine Learning Repository。<https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast#>

Huang Liz（2019年4月10日）。Python學習筆記#14：機器學習之KNN實作篇。Medium。[https://medium.com/@search.psop/python學習筆記-14-機器學習之knn實作篇-64071fbd0ac8](https://medium.com/@search.psop/python%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-14-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E4%B9%8Bknn%E5%AF%A6%E4%BD%9C%E7%AF%87-64071fbd0ac8)

小白掌柜（2022年6月16日）。調用sklearn模型遇到Unknown label type: continuous 的解決辦法。台部落。<https://www.twblogs.net/a/5ee7f5aae8d3b317573e453f>