Trường Đại học Khoa học Tự nhiên - ĐHQG Thành phố Hồ Chí Minh

Toán Ứng Dụng Thống Kê

Báo Cáo Đồ Án 3

Chủ đề: Đồ Án 3 – Linear Regression

Lớp: 21CLC5

Giảng viên:

- Phan Thị Phương Uyên
- Nguyễn Văn Quang Huy

Thông tin sinh viên:

- 20127030 - Nguyễn Mạnh Hùng

Mục Lục

I. Giới thiệu

Ở đồ án này, ta sẽ lần lượt đi đọc dữ liệu từ các file train.csv, test.csv đã được tiền xử lý trước đó và xây dựng các mô hình máy học ở dạng hồi quy tuyến tính sao cho thỏa mãn các yêu cầu đề bài mà cô đã giao, cơ bản như:

- Xây dựng mô hình hồi quy tuyến tính dựa trên các thuộc tính đã cho
- Chọn ra một thuộc tính phù hợp nhất trong số các thuộc tính đã cho bằng phương pháp Cross-Validation và xây dựng mô hình hồi quy tuyến tính
- Tự xây dựng m mô hình hồi quy tuyến tính tùy ý và chọn ra mô hình tốt nhất trong số mô hình ấy.

Đồ án này được thực hiện trên file 20127030.ipynb – file jupyter notebook với phiên bản Python 3.10.11

II. Các thư viện đã sử dụng

Sau đây là toàn bộ các thư viện đã được sử dụng cũng như công năng của chúng trong đồ án này.

Bảng công năng của các thư viện

Thư viện	Công dụng
Numpy	Dùng để cast dữ liệu nhằm huấn luyện và fit dữ liệu cho mô hình, cũng như giúp mô hình dự đoán
Pandas	Dùng để đọc dữ liệu, hiển thị dữ liệu, biểu diễn thông số dữ liệu.
Seaborn và Matplotlib	Trực quan hóa dữ liệu giúp cho người dùng hiểu rõ hơn về phân bố và tầm ảnh hưởng của các thuộc tính với nhau.
Scikit-learn	Gọi mô hình hồi quy tuyến tính, huấn luyện và dự đoán kết quả. Đồng thời, gọi hàm tính độ sai số trung bình tuyệt đối và chuẩn hóa dữ liệu

Như vậy đó là toàn bộ các thư viện đã được cài đặt và sử dụng. Đồ án không còn sử dụng bất kỳ thư viện nào khác ngoài các thư viện đã được đề cập như trên.

III. Các hàm đã sử dụng.

Ở phần này, ta sẽ lần lượt đi qua các hàm thủ công cũng như các hàm thư viện đã được sử dụng trong phần đồ án này

Tên hàm	Input	Output	Công năng
MAE	predict: giá trị dự đoán của mô hình. trueval: giá trị thực tế của tập dữ liệu	Giá trị độ lỗi	Tính độ lỗi trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực**[1]**
make_new_Dataframe	cols: Danh sách các cột thuộc tính trong dataset. dataset: tập dataset truyền vào	Trả về dataframe mới chứa toàn bộ các dòng dữ liệu ứng với các cột đã truyền vào, các cột còn lại bị loại bỏ	Rút trích toàn bộ dữ liệu ứng với các thuộc tính của một dataset bất kỳ đã truyền vào hàm
Standard_Normalize	data: tập dataframe cần chuẩn hóa	Các dữ liệu trong dataframe được chuẩn hóa và trả về ở dạng mảng numpy	Dùng để chuẩn hóa dữ liệu theo phân phối chuẩn z-score thông qua hàm StandardScaler của sklearn**[2]**
shuffle	df: tập dataframe	Tập dataframe đã bị xáo trộn các dòng.	Dùng để xáo trộn các dòng của dataframe nhằm phục vụ cho Cross Validation**[3]**
CrossValidation	x và y: dữ liệu ở dạng mảng numpy. num_fold: số lượng batch. model: mô hình hồi quy tuyến tính	Giá MAE trung bình của mô hình sau khi fit và dự đoán kết quả tương ứng với các batch	Chia tập dataset ra làm num_fold phần bằng nhau và mỗi phần tính giá trị MAE và sau cùng tính MAE trung bình giữa các lần đó**[4]**
calculating	feature_list: danh sách các thuộc tính x và y: lần lượt là dữ liệu và target. model: mô hình hồi quy tuyến tính	Trả về mảng các mae ứng với từng thuộc tính trong danh sách.	Thực hiện tính Cross-Validation nhằm tìm ra thuộc tính tốt nhất cho từng mô hình

IV. Quá trình thực hiện.

Sau đây là toàn bộ quá trình thực hiện các yêu cầu bài toán.

1. ***Sử dụng toàn bộ 11 đặc trưng đầu tiên Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain***

Sau khi đọc và rút trích toàn bộ dữ liệu thành x_train, y_train, x_test và y_test ứng với các tập train và test. Ta lần lượt thực hiện như sau:

- ***Bước 1:*** Rút trích dữ liệu của 11 đặc trưng và mức lương (Salary) tương ứng của tập train và test và lưu vào các biến lần lượt là X_1a_train, X_1a_test, y_1a_train, y_1a_test.
- ***Bước 2:*** Gọi model Linear Regression**[5]** sau đó fit dữ liệu của tập train
- ***Bước 3:*** Dự đoán kết quả trên X_1a_test và tính MAE.

Kết quả thu được:

```
MAE =  105052.52978823145
```

Để biết được các hệ số của mô hình ta chỉ cần gọi hàm .coef_**[5]**

Ta có các hệ số sau:

```
[-23183.32950765    702.76679172   1259.0187879  -99570.60814074
  18369.9624496    1297.53200035  -8836.727123      141.75993906
    145.74234652    114.64331342  34955.75040521]
```

Sau cùng, ta có được công thức liên hệ giữa các thuộc tính với nhau là:

$$Salary = -23183.33 \times Gender + 702.767 \times 10percentage + 1259.019 \times 12percentage - 99570.608 \times CollegeTier + 18369.962 \times Degree + 1297.532 \times collegeGPA - 8836.727 \times CollegeCityTier + 141.759 \times English + 145.742 \times Logical + 114.643 \times Quant + 34955.750 \times Domain$$

2. ***Xây dựng mô hình sử dụng duy nhất 1 đặc trưng tính cách với các đặc trưng tính cách gồm conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience. Tìm mô hình cho kết quả tốt nhất***

Để thực hiện được yêu cầu trên ta cần phải thực hiện các yêu cầu sau:

- ***Bước 1***: thực hiện rút trích các thuộc tính sau từ tập train 'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism', 'openess_to_experience','Salary' từ tập train.csv
- ***Bước 2***: tiến hành xáo dữ liệu thông qua hàm shuffle.
- ***Bước 3***: tách cột Salary làm target.
- ***Bước 4***: ứng với mỗi thuộc tính mae trung bình thông qua cross validation**[3]**, toàn bộ quá trình được thực hiện ở hàm calculating.
- ***Bước 5***: chọn thuộc tính có mae trung bình nhỏ nhất và đi tìm mô hình giống câu đầu tiên bằng việc rút trích toàn bộ dữ liệu theo thuộc tính đã chọn.

Sau khi cross validation, ta thu được kết quả như sau:

```
AVG MAE result
['conscientiousness', 'agreeableness', 'extraversion', 'nueroticism', 'openess_to_experience']
[124045.34840412978, 123506.09835620275, 123781.07744693951, 123463.81615830222, 124015.80576072424]
Best features is  nueroticism
```

Như vậy, ta có thể thấy thuộc tính neuroticism và agreeableness lần lượt có avg mae nhỏ nhất nhì trong toàn bộ các thuộc tính. Vì thế, ta sẽ chọn nueroticism làm thuộc tính chính cho mô hình này.

Khi huấn luyện xong cho mô hình với thuộc tính nueroticism ta thu được kết quả về MAE và công thức như sau:

```
MAE =  119361.91739987816
[-16021.49366179]
```

$$\text{Salary} = -16021.494 \times \text{nueroticism}$$

Như vậy, ta có thể thấy rằng thuộc tính nueroticism tỷ lệ nghịch với mức lương.

Nhận xét: Sau khi tìm hiểu về Big 5 personality**[6][7][8]** thì có thể thấy rằng, ở mỗi quốc gia khác nhau thì các thuộc tính này có độ ảnh hưởng ở các mức độ khác nhau. Ví dụ như ở Đức, agreeableness sẽ khiến cho tiền lương giảm từ 2-5%, còn ở Vương quốc Anh, với môi trường làm việc bình đẳng, những nhân viên có đặc điểm agreeableness cao bị giảm lương khoảng 4-6% so với những người có agreeableness thấp**[8]**. Ngoài ra, đối với ở Vương Quốc Anh và ở Mỹ thì tính nueroticism tỷ lệ nghịch với khả năng lương tăng**[8][7]**. Ngoài ra, khi khảo sát ở Mỹ theo báo CNBC thì việc có chỉ số nueroticism cao sẽ khiến cho ta bị trì hoãn và không thăng tiến trong công việc và có chỉ số nueroticism thấp đi kèm với extraversion cao thì sẽ khiến cho ta dễ dàng thăng tiến trong công việc, đặc biệt là các lĩnh vực về sales, marketing**[7]**. Như vậy, việc lựa

chọn thuộc tính nueroticism làm mô hình và có hệ số âm đã phản ánh những gì mà 2 bài báo uy tín đã thể hiện và đề cập đến.

3. *Xây dựng mô hình sử dụng duy nhất 1 đặc trưng English, Logical, Quant. Tìm mô hình cho kết quả tốt nhất*

Để thực hiện được yêu cầu trên ta cần phải thực hiện các yêu cầu sau:

- ***Bước 1***: thực hiện rút trích các thuộc tính sau từ tập train 'English', 'Logical', 'Quant','Salary' từ tập train.csv
- ***Bước 2***: tiến hành xáo dữ liệu thông qua hàm shuffle.
- ***Bước 3***: tách cột Salary làm target.
- ***Bước 4***: ứng với mỗi thuộc tính mae trung bình thông qua cross validation**[3]**, toàn bộ quá trình được thực hiện ở hàm calculating.
- ***Bước 5***: chọn thuộc tính có mae trung bình nhỏ nhất và đi tìm mô hình giống câu đầu tiên bằng việc rút trích toàn bộ dữ liệu theo thuộc tính đã chọn.

Sau khi cross validation, ta thu được kết quả như sau:

```
AVG MAE result
['English', 'Logical', 'Quant']
[120975.63934186725, 120479.41733770589, 117297.11717656902]
Best features is  Quant
```

Như vậy, ta có thể thấy thuộc tính Quant có avg mae nhỏ nhất trong toàn bộ các thuộc tính. Vì thế, ta sẽ chọn Quant làm thuộc tính chính cho mô hình này.

Khi huấn luyện xong cho mô hình với thuộc tính Quant ta thu được kết quả về MAE và công thức như sau:

```
MAE =  8298522.218664632
[-16021.49366179]
```

$$\text{Salary} = -16021.494 \times \text{Quant}$$

Qua trên, ta có thể thấy được rằng thuộc tính Quant tỷ lệ nghịch với Salary.

Nhận xét: thông qua 3 kỹ năng cần khảo sát, ta có thể thấy các kỹ năng về định lượng và tư duy logic là các kỹ năng quan trọng nhất. Sau khi tìm hiểu các nguồn thông tin về các nhà tuyển dụng**[9][10]** thì các kỹ năng liên quan đến khả năng phân tích, tư duy

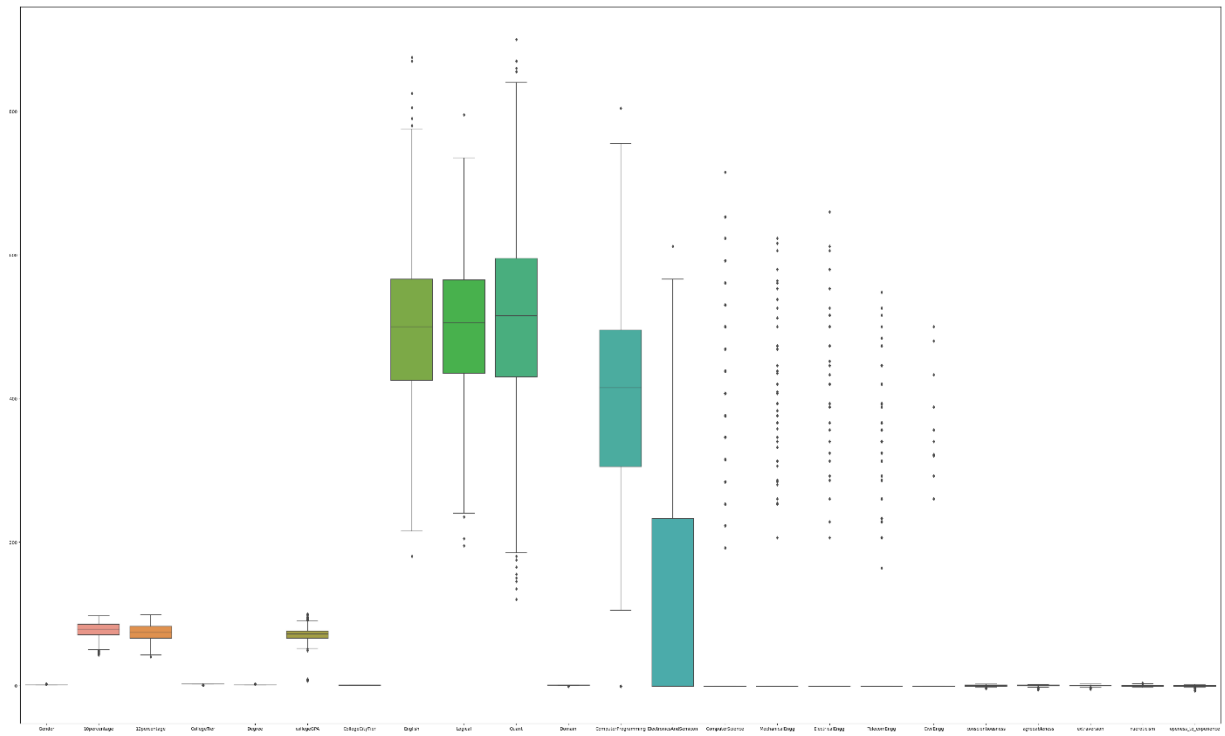
phản biện, quảng trị thời gian và tự học hỏi, khám phá là cực kỳ cần thiết đối với các nhà tuyển dụng. Qua các kỹ năng trên, ta phần nào thấy được khả năng định lượng và khả năng tư duy logic là các nền tảng quan trọng. Vì thế việc chọn yếu tố định lượng làm yếu tố chủ chốt trong 3 yếu tố trên là hoàn toàn hợp lệ.

4. Tự xây dựng mô hình

Phần này sẽ trình bày về quy trình chọn lọc dữ liệu và xây dựng mô hình tuyến tính hồi quy cho tập dữ liệu này.

a. Xử lý dữ liệu và chọn lọc thuộc tính

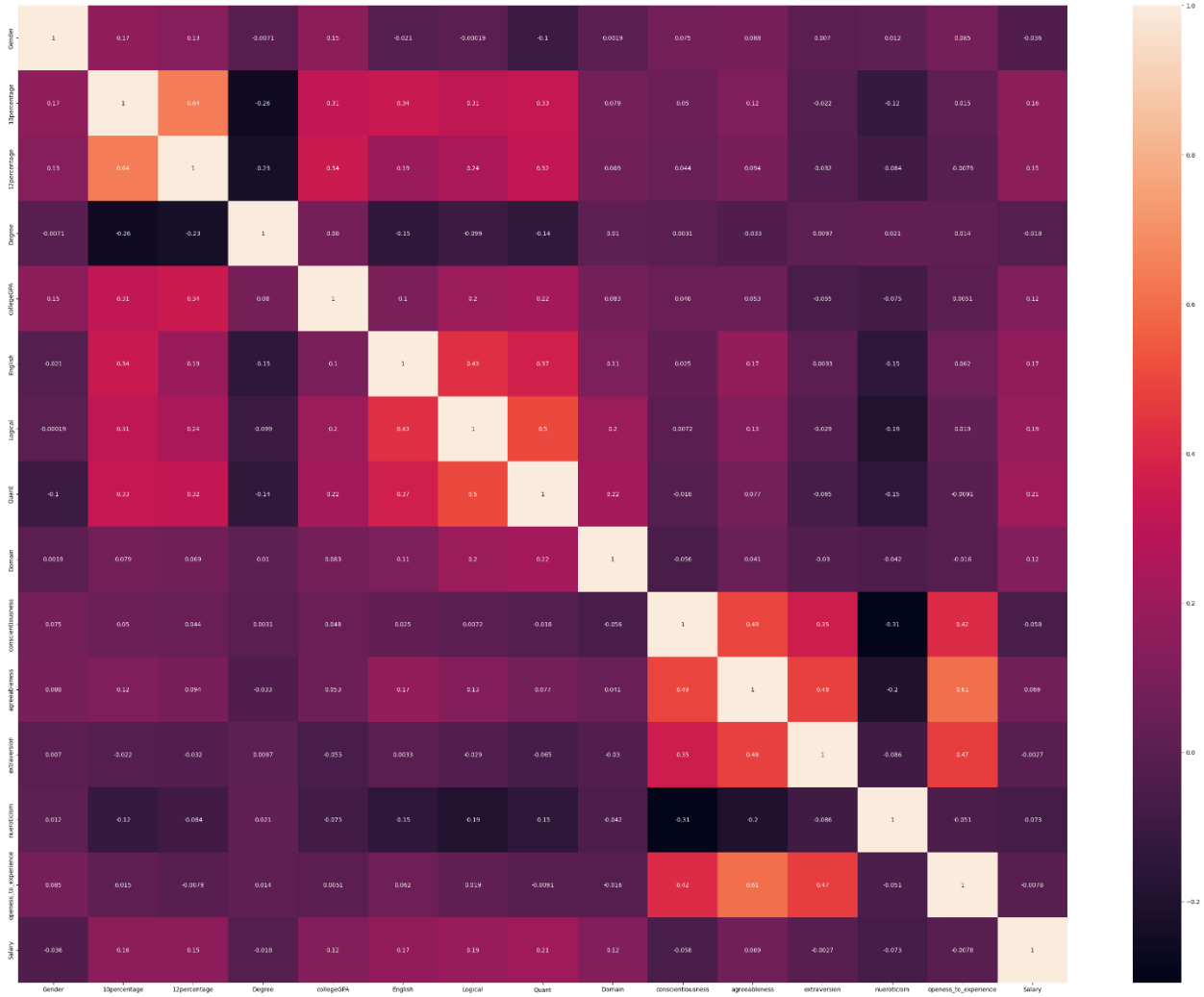
Đầu tiên, ta cần phải liệt kê toàn bộ thuộc tính có trong tập dữ liệu và vẽ boxplot để liệt kê toàn bộ các đặc trưng có chứa thông tin nhiễu và loại bỏ chúng**[11]**. Sau khi liệt kê và phát họa biểu đồ, ta được hình như sau:



Thông qua các hình trên, ta có thể thấy rằng các thuộc tính về chuyên ngành và cấp bậc của trường đại học lần lượt là 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'CollegeCityTier', 'CollegeTier' có số lượng thông tin quá rời rạc, thậm chí có một số

thuộc tính ghi nhận quá ít số lượng thông tin nên các thuộc tính trên sẽ bị loại bỏ ra khỏi tập dữ liệu để tránh tình trạng nhiễu (outliers) cho tập dữ liệu.

Sau khi loại bỏ toàn bộ các điểm nhiễu, ta sẽ vẽ ma trận tương quan (correlation matrix)**[12]** của tập dữ liệu để biết thêm về mối liên hệ giữa các thuộc tính với nhau trong tập dữ liệu. Ta có ma trận tương quan như sau:



Ta thấy, từng thuộc tính khác nhau sẽ có mức ảnh hưởng khác nhau về thuộc tính lương. Khi xét sự phụ thuộc giữa các thuộc tính với lương, ta thấy rằng, các hệ số > 0 thì các thuộc tính tương ứng sẽ tỷ lệ thuận với nhau và ngược lại nếu hệ số < 0 thì tỷ lệ nghịch với nhau. Trường hợp các hệ số = 0 thì 2 nhân tố đó không ảnh hưởng gì đến nhau.

Như vậy, thông qua toàn bộ quá trình phân tích ở trên, ta có thể thấy rằng các yếu tố như "Quant", "Logical", "English", "collegeGPA", "Domain", "10percentage",

"12percentage" đều có ảnh hưởng lớn đến mức lương. Chính vì thế, ta sẽ xây dựng mô hình có chứa những yếu tố này để dự đoán mức lương.

b. Xây dựng mô hình và chọn mô hình tốt nhất.

Ta lần lượt xây dựng 2 mô hình dự đoán mức lương có các thuộc tính lần lượt là "10percentage", "12percentage" và "Quant", "Logical", "English", "collegeGPA", "Domain". Cũng như các bước ở trên về quy trình thực hiện Cross-Validation[3], ta sẽ lần lượt đi rút trích các tập dữ liệu tương ứng với các cột, xáo dữ liệu và tính toán mae trung bình thông qua hàm calculating. Tuy nhiên, trước khi thực hiện cross-validation thì ta sẽ thực hiện chuẩn hóa dữ liệu theo z-score thông qua hàm Standard_Normalize**[2]** nhằm loại bỏ bớt các dữ liệu nhiễu (outliers) có sẵn trong các thuộc tính đã chọn.

Ta gọi model1 là mô hình dự đoán lương có các thuộc tính "Quant", "Logical", "English", "collegeGPA", "Domain" và model2 là mô hình dự đoán mức lương thông qua 2 thuộc tính lần lượt là "10percentage", "12percentage". Sau khi thực hiện cross-validation ta có được các thông số như sau:

Mô hình	Thông số
model1	`first model avg. mae =  0.5148199400936802`
model2	`second model avg. mae =  0.5337465031824032`

Như vậy, ta có thể thấy mô hình model1 có thông số lỗi nhỏ hơn nên ta sẽ chọn mô hình model1 làm mô hình tốt nhất và thử nghiệm trên tập test.

Sau khi rút trích và chuẩn hóa dữ liệu tương ứng ở tập test, ta tiến hành dự đoán dữ liệu và tính độ lỗi của chúng trên tập dữ liệu này, ta thu được kết quả về độ lỗi và các thông số của chúng như sau:

```
My best model MAE =  0.5847596331772111
[0.10814796 0.0693188  0.08517612 0.07046552 0.06871255]
```

Như vậy, sau cùng, ta có được công thức hồi quy tuyến tính tốt nhất để dự đoán được mức lương như sau:

$$\text{Salary} = 0.108 \times \text{Quant} + 0.069 \times \text{Logical} + 0.085 \times \text{English} + 0.070 \times \text{collegeGPA} + 0.069 \times \text{Domain}$$

V. Nhận xét và đánh giá kết quả

Bảng đánh giá kết quả

Yêu cầu được giao	Mức độ hoàn thành
Sử dụng toàn bộ 11 đặc trưng đầu tiên để xây dựng mô hình	100%
Xây dựng mô hình sử dụng duy nhất 1 đặc trưng tính cách với các đặc trưng tính cách và tìm mô hình cho kết quả tốt nhất	100%
Xây dựng mô hình sử dụng duy nhất 1 đặc trưng `English`, `Logical`, `Quant`, tìm mô hình cho kết quả tốt nhất	100%
Trình bày toàn bộ quy trình tự xây dựng m mô hình theo đánh giá của cá nhân (m=2)	100%
Tìm mô hình cá nhân tốt nhất trong m mô hình tự xây (m = 2)	100%

Nhận xét:

Sau khi chuẩn hóa toàn bộ dữ liệu thì em thấy độ lỗi mae giảm xuống rõ rệt trong quá trình huấn luyện và kiểm tra. Như vậy, có thể thấy, tuy được cô tiền xử lý dữ liệu nhưng dữ liệu vẫn còn tồn đọng khá nhiều điểm nhiễu (outliers) khiến cho độ lỗi chênh lệch ngày càng lớn.

Các thuộc tính mà cho ra kết quả, hoàn toàn phù hợp với các nghiên cứu của các bài báo uy tín và sự thật thực tế về các nhu cầu và tuyển dụng của các phòng nhân sự về các mức lương khác nhau dựa trên các kỹ năng và tính cách khi làm việc.

Lưu ý: Đôi lúc kết quả ở *câu 1b* sẽ cho ra thuộc tính tố nhất là agreeableness nhưng nhìn chung ở đa số các lần chạy, thuộc tính tốt nhất vẫn là neuroticism. Lý giải cho vụ này là do 2 thuộc tính này như đề cập ở trên có độ lỗi không chênh lệch nhau quá lớn, ngoài ra khi xáo ngẫu nhiên các vị trí và các trọng số khởi tạo mỗi mô hình khi gọi là hoàn toàn ngẫu nhiên, vì thế sẽ có trường hợp rơi vào agreeableness nhưng đó chỉ là trường hợp thiểu số còn toàn bộ chạy thì đều là neuroticism.

VI. Nguồn tham khảo

[1]: Mean Absolute Error https://www.ibm.com/docs/en/cloud-paks/cp-data/3.5.0?topic=overview-mean-absolute-error

[2]: z-score scaler: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[3]: Cách xáo dữ liệu trong dataframe: https://stackoverflow.com/questions/29576430/shuffle-dataframe-rows

[4]: k-fold Cross-Validation: https://en.wikipedia.org/wiki/Cross-validation_(statistics)#:~:text=Cross%2Dvalidation%20is%20a%20resampling,model%20will%20perform%20in%20practice

[5]: Mô hình hồi quy tuyến tính trong thư viện sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[6]: https://glints.com/vn/blog/mo-hinh-big-5-personality-la-gi/

[7]: https://www.cnbc.com/2022/04/24/report-personality-traits-correlate-with-promotions-high-salaries.html

[8]: https://kse.ua/kse-research/relationship-between-your-personality-and-your-salary-level/

[9]: https://tanca.io/blog/nhung-ky-nang-can-co-khi-di-lam-ma-nha-tuyen-dung-danh-gia-cao

[10]: https://www.linkedin.com/business/talent/blog/talent-strategy/linkedin-most-in-demand-hard-and-soft-skills

[11]: Sách Applied Multivariate Statistical Analysis, tác giả Richard A. Johnson chương số 4 về phân phối chuẩn trong tập dữ liệu và cách loại bỏ outliers

[12]: Sách Applied Multivariate Statistical Analysis, tác giả Richard A. Johnson chương 2.6 về công thức tính ma trận hệ số tương quan và ý nghĩa của ma trận hệ số tương quan