ĐẠI HỌC QUỐC GIA TPHCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN CÔNG NGHỆ TRI THỨC

Báo cáo Đồ án Thực hành

Đề tài: Linear Regression

Môn học: Toán Ứng dụng và Thống kê cho Công nghệ Thông tin

Sinh viên thực hiện:
Trà, Hoàng Anh (21127453)

Giáo viên hướng dẫn:
ThS. Vũ Quốc Hoàng
ThS. Nguyễn Văn Quang Huy
CN. Ngô Đình Hy
ThS. Phan Thị Phương Uyên

Ngày 22 tháng 8 năm 2023

LỜI MỞ ĐẦU

Hồi quy tuyến tính (linear regression) là một kỹ thuật phân tích dữ liệu dựa trên mối quan hệ tuyến tính giữa các biến. Hồi quy tuyến tính được ứng dụng rộng rãi trong nhiều lĩnh vực như kinh tế, y học, kỹ thuật, khoa học xã hội, v.v. Mục tiêu của hồi quy tuyến tính là tìm ra một đường thẳng (hoặc một mặt phẳng, một siêu phẳng) sao cho khoảng cách từ các điểm dữ liệu đến đường thẳng (hoặc mặt phẳng, siêu phẳng) là nhỏ nhất. Đường thẳng (hoặc mặt phẳng, siêu phẳng) này được gọi là đường hồi quy (regression line) và có thể được sử dụng để dự đoán giá trị của biến phụ thuộc (output) khi biết giá trị của biến độc lập (input).

Trong báo cáo này, em sẽ thực hành các bước cơ bản để xây dựng một mô hình dự đoán sử dụng mô hình hồi quy tuyến tính đơn giản (simple linear regression) theo nhiều trường hợp khác nhau.

Do thời gian và trình độ chuyên môn còn hạn chế nên đồ án này không thể tránh khỏi những thiếu sót. Em rất mong nhận được nhiều ý kiến đóng góp từ các thầy, cô giáo và bạn bè để đồ án ngày càng hoàn thiện hơn!

Sinh viên thực hiện:
Trà, Hoàng Anh (21127453)

Mục lục

1 Khái quát nội dung đồ án

"Mục tiêu của đồ án là tìm hiểu các yếu tố quyết định mức lương và việc làm của các kỹ sư ngay sau khi tốt nghiệp. Các yếu tố như điểm số ở các cấp/trường đại học, kỹ năng của ứng viên, sự liên kết giữa trường đại học và các khu công nghiệp/công ty công nghệ, bằng cấp của sinh viên và điều kiện thị trường cho các ngành công nghiệp cụ thể sẽ ảnh hưởng đến điều này.

Bộ dữ liệu được sử dụng trong đồ án này thu thập tại Ấn Độ, nơi có hơn 6000 cơ sở đào tạo kỹ thuật công nghệ với khoảng 2,9 triệu sinh viên đang học tập. Mỗi năm, trung bình có 1,5 triệu sinh viên tốt nghiệp chuyên ngành Công nghệ/Kỹ thuật, tuy nhiên do thiếu kỹ năng cần thiết, ít hơn 20 % trong số họ có việc làm phù hợp với chuyên môn của mình. Bộ dữ liệu này không chỉ giúp xây dựng công cụ dự đoán mức lương mà còn cung cấp thông tin về các yếu tố ảnh hưởng đến mức lương và chức danh công việc trên thị trường lao động. Sinh viên sẽ được khám phá những thông tin này trong phạm vi đồ án." [6]

2 Nội dung lý thuyết và xây dựng ý tưởng

2.1 Hồi quy tuyến tính sử dụng phương pháp bình phương tối thiểu - Ordinary Least Squares Linear Regression

Cho ma trận **A** có kích thước m × n và vector b có kích thước m.[5]
Ta cần tìm nghiệm của phương trình $Ax \approx b$
Chuẩn Euclidean của bình phương phần dư r của $Ax - b$ là:

$$r = ||Ax - b||^2 \quad (1)$$

Để giải được nghiệm x cho hệ phương trình, ta thực hiện tối thiểu hóa công thức (1) được nghiệm x của hệ phương trình được tính như sau:

$$x = (A^T A)^{-1} A^T b$$

Và để kiểm tra sai số đối với nghiệm tìm được, MAE được dùng để ước lượng **trung bình của sai số** (độ lỗi), được tính bằng công thức:

$$MAE = \frac{1}{n} \sum_{n}^{i=1} |y_i - \hat{y}_i|$$

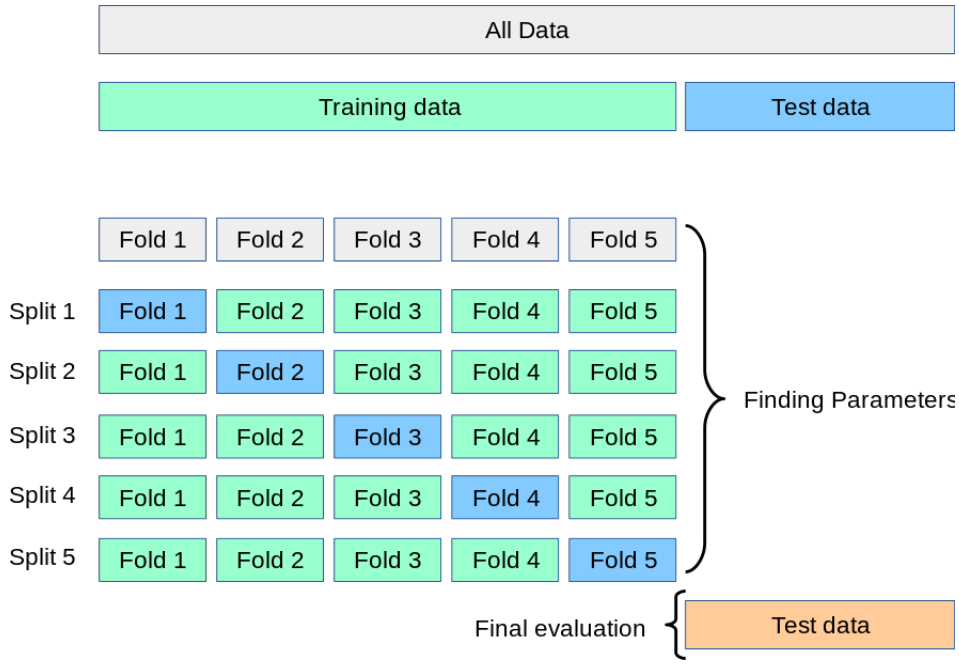
Trong đó: - n: số lượng mẫu quan sát - y_i: giá trị mục tiêu của mẫu thứ i - $\hat{y}_i$: giá trị mục tiêu của mẫu thứ i được dự đoán từ mô hình hồi quy tuyến tính
Như vậy, với phương pháp trên, ta có thể xây dựng ý tưởng như sau:

- Với bộ dữ liệu đầu vào `X_train` và `y_train` cùng với một mô hình M_i nhất định, ta đi tìm nghiệm của phương trình $X\_train \times x \approx y\_train$.
- Sau khi tìm ra nghiệm x thì sử dụng x để tính toán MAE bằng bộ dữ liệu kiểm tra `X_test` và `y_test` của mô hình M_i.
- Sau khi tính toán thì so sánh độ lỗi MAE của các mô hình M_i với nhau để chọn ra mô hình tốt nhất với MAE nhỏ nhất.

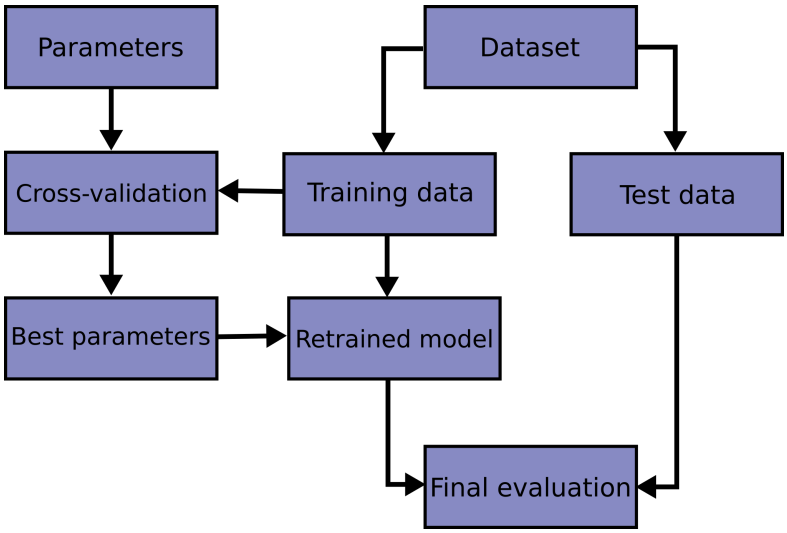
2.2 K-Fold Cross Validation

Cross validation là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp bộ dữ liệu không đủ lớn. [3] [4]
Tham số quan trọng trong kỹ thuật này là k, đại diện cho số nhóm mà dữ liệu sẽ được chia ra. Vì lý do đó, nó được mang tên k-fold cross-validation. Khi giá trị của k được lựa chọn, người ta sử dụng trực tiếp giá trị đó trong tên của phương pháp đánh giá. Ví dụ với k=10, phương pháp sẽ mang tên 10-fold cross-validation.
Kỹ thuật này thường bao gồm các bước như sau:

- Xáo trộn dataset một cách ngẫu nhiên
- Chia dataset thành k nhóm
- Với mỗi nhóm:
  - Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
  - Các nhóm còn lại được sử dụng để huấn luyện mô hình
  - Huấn luyện mô hình
  - Đánh giá và sau đó hủy mô hình
- Tổng hợp hiệu quả của mô hình dựa từ các số liệu đánh giá



Hình 1: Minh họa KFold



Hình 2: Quá trình thực hiện Cross Validation tổng quát

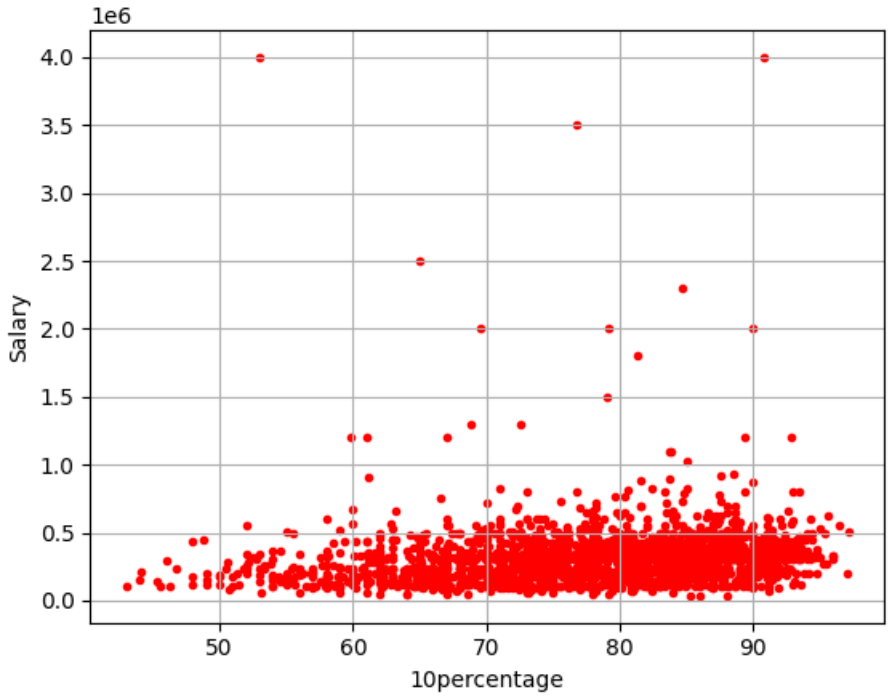
2.3 Xây dựng các mô hình để tìm ra mô hình cho kết quả tốt nhất từ bộ dữ liệu "Mức lương kỹ sư tốt nghiệp đại học"

2.3.1 Ý tưởng ban đầu

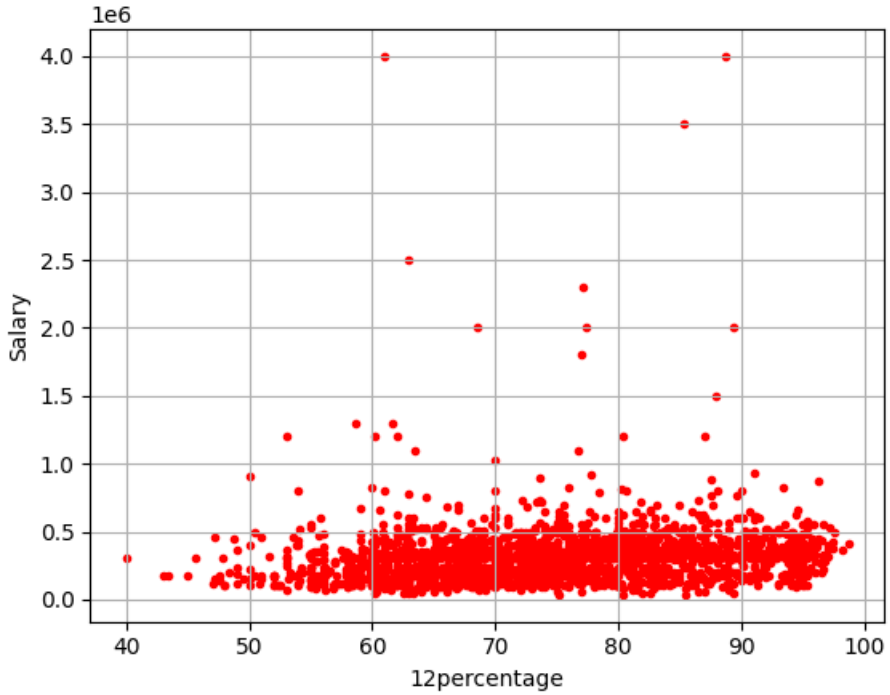
Ở các yêu cầu khác, đồ án đã xây dựng mô hình từ tập nhiều đặc trưng hay các đặc trưng cụ thể như tính cách, ngoại ngữ, logic, định lượng, nên để không đi theo lối mòn cũ hay xây dựng dựa trên toàn bộ đặc trưng một cách mất phương hướng, chúng ta tập trung vào các điểm số và đặc trưng có liên quan. [2]
Hơn nữa vì các điểm số này đều cùng cho ra một ngưỡng giá trị là như nhau và sự phân bố của chúng cũng khá tương đồng nên việc xây dựng mô hình xoay quanh những đặc trưng này cũng đưa ra các sự đánh giá có giá trị sử dụng.

2.3.2 Biểu đồ phân bố các đặc trưng cần thiết theo `Salary`

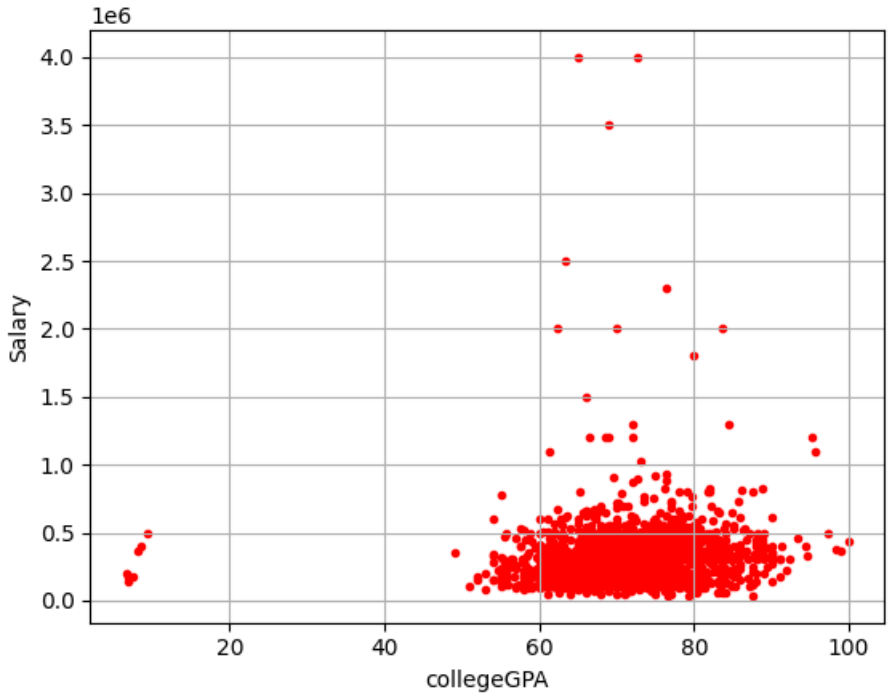
Các đặc trưng cần quan tâm là: `10percentage`, `12percentage`, `collegeGPA` và các đặc trưng liên quan là `Degree` và `CollegeTier`.



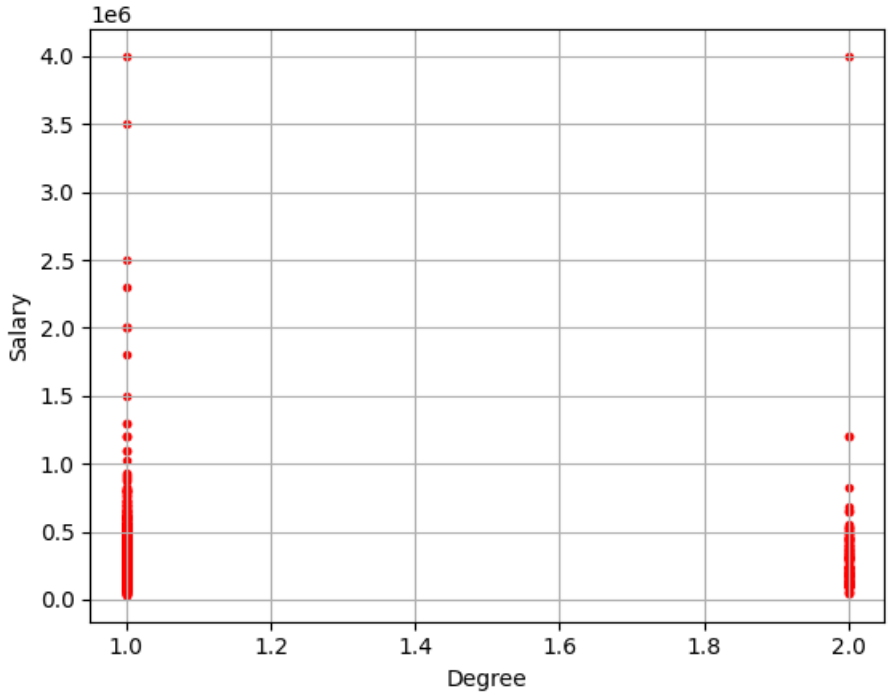
Hình 3: Phân bố giá trị đặc trưng 10percentage theo Salary



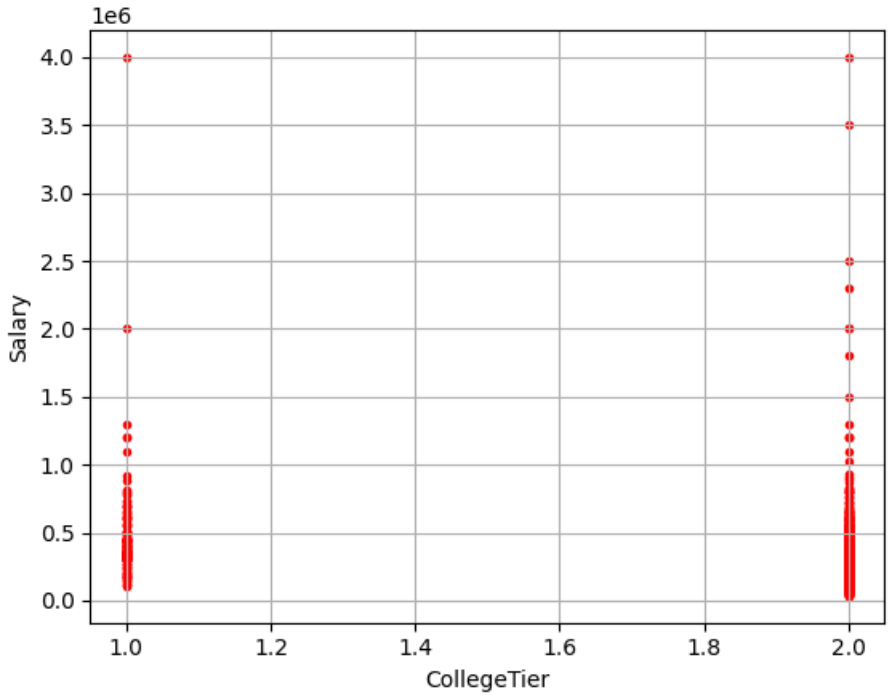
Hình 4: Phân bố giá trị đặc trưng 12percentage theo Salary



Hình 5: Phân bố giá trị đặc trưng collegeGPA theo Salary



Hình 6: Phân bố giá trị đặc trưng degree theo Salary



Hình 7: Phân bố giá trị đặc trưng collegeTier theo Salary

2.3.3 Xây dựng mô hình

Phân tích:

- Sự phân bố giữa 2 biểu đồ 10percentage và 12percentage gần như là tương đồng.
- Xét về ý nghĩa, đối với `Degree` thì giá trị càng cao chứng tỏ bằng càng có giá trị và ngược lại, đối với `CollegeTier` giá trị càng thấp chứng tỏ Trường theo học càng có giá trị.
- Sự phân bố của collegeGPA tập trung nhiều và rất nhiều về phía bên phải chứng tỏ số lượng điểm cao khi tốt nghiệp là rất lớn. Nên nếu cần để so sánh thì nên gia tăng khoảng cách giữa `10percentage` và `12percentage` với `collegeGPA` bằng cách bình phương.

Như vậy đối với sự phân tích trên, ta xây dựng được 3 mô hình:

- Mô hình 1: Đánh giá sức ảnh hưởng của Tổng điểm đạt được trong kỳ thi lớp 10, lớp 12 tới Salary.

$$Salary = a \times 10percentage + b \times 12percentage$$

- Mô hình 2: Đánh giá sức ảnh hưởng của GPA tại thời điểm tốt nghiệp và Loại bằng cấp đã đạt tới Salary.

$$Salary = a \times CollegeGPA \times Degree$$

- Mô hình 3: Đánh giá sức ảnh hưởng của Tổng điểm đạt được trong kỳ thi lớp 10, lớp 12, GPA tại thời điểm tốt nghiệp và Loại bằng cấp đã đạt, hạng của trường tới Salary.

$$Salary = a \times 10percentage^2 + b \times 12percentage^2 + c \times (\frac{CollegeGPA \times Degree}{CollegeTier})^2$$

3 Mô tả hàm và thuật toán

3.1 Các thư viện sử dụng

- `Pandas`: Đọc và tính toán trên dữ liệu.
- `NumPy`: Tính toán ma trận.
- `matplotlib`: Xây dựng và hiển thị biểu đồ.

3.2 Các hàm, class và thao tác hỗ trợ

3.2.1 Đọc dữ liệu bằng `Pandas`

`Thực hiện:` [6]

1. Sử dụng phương thức `read_csv()` của `Pandas` để đọc dữ liệu từ 2 file **train.csv** và **test.csv**.
2. Sử dụng phương thức `iloc()` của thư viện `Pandas` [1] để lấy các đặc trưng X và giá trị mục tiêu y cho các tập huấn luyện (train) và kiểm tra (test).

3.2.2 Class `OLSLinearRegression`

Tham khảo từ [5]
Phương thức `fit()` tìm nghiệm của $Ax \approx b$ `Input:`

- `X`: tương ứng với ma trận A
- `y`: tương ứng với vector b

`Output:`

- `w` là nghiệm của $Ax \approx b$

`Thực hiện:`

1. Dùng phương thức `inv()` trong `Numpy` để tính $X\_pinv = (A^TA)^{-1}A^T$.
2. Nghiệm của phương trình là $w = X\_inv \times b$ và truyền w vào đối tượng class.

Phương thức `get_params()` trả về nghiệm của $Ax \approx b$.
Phương thức `predict()` trả về vector $\hat{y}$
`Input:`

- `X`: là ma trận kiểm tra.

`Thực hiện:`

1. Dùng phương thức `ravel()` để tạo một mảng phẳng liền kề cho nghiệm w rồi đem nhân với `X` để tính $\hat{y}$

3.2.3 Hàm tính MAE `mae()`

`Input:`

- Hai vector y và $\hat{y}$

`Output:`

- Giá trị MAE là trung bình sai số.

`Thực hiện:`

1. Làm phẳng 2 vector thành mảng bằng phương thức `ravel()`.
2. Sử dụng công thức đã nêu ở trên kèm các phương thức trong thư viên `Numpy` để tính ra giá trị cuối cùng.

3.2.4 Hàm thực hiện `KFoldCrossValidation()`

`Input:`

- `X`: Ma trận chứa các đặc trưng huấn luyện.
- `Y`: Vector chứa 1 giá trị mục tiêu kiểm tra.
- `k`: Số lượng Fold, mặc định là 5
- `i`: Số thứ tự Fold, mặc định là 0

`Output:`

- 2 Ma trận và 2 vector để test và train sau khi chia fold.

`Thực hiện:`

1. Tính toán số lượng giá trị, vị trí bắt đầu và kết thúc của Fold thứ `i`.
2. Dùng phương thức `iloc()` để lấy các giá trị trong Fold ra để được ma trận và vector test và `drop()` các giá trị đã lấy để được ma trận và vector train.

3.2.5 Hàm tiền xử lý `preprocess`

`Input:`

- `main_df`: ma trận chứa dữ liệu train.

`Output:`

- Ma trận kết quả sau khi trích dữ liệu và xử lý.

`Thực hiện:`

1. Lấy các cột cần sử dụng trong mô hình.
2. Dùng `combine()` và `pow()` [1] để xử lý các trường hợp cần nhân giá trị hay bình phương.

4 Kết quả thực hiện và nhận xét

4.1 Yêu cầu 1a: Sử dụng 11 đặc trưng đầu tiên Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain

Sơ lược về cách thực hiện:

- Trích 11 cột đầu tiên của `X_train` và `X_test`
- Huấn luyện cho mô hình với dữ liệu đã trích.
- Lấy kết quả và tính giá trị MAE.

Kết quả:
Vector `w`:

0	-22756.512821
1	804.503156
2	1294.654565
3	-91781.897531
4	23182.388679
5	1437.548672
6	-8570.661985
7	147.858299
8	152.888476
9	117.221846
10	34552.286221

Giá trị MAE: 104863.7775403321
Công thức hồi quy: Salary $= -22756.513X_1 + 804.503X_2 + 1294.655X_3 + (-91781.898)X_4 + 23182.389X_5 + 1437.549X_6 + (-8570.662)X_7 + 147.858X_8 + 152.888X_9 + 117.222X_{10} + 34552.286X_{11}$
Nhận xét:

- Trong phần này, 11 đặc trưng của dữ liệu được sử dụng để xây dựng mô hình nên mô hình thu được không có nhiều thông tin để ứng dụng phân tích cụ thể và chưa hoàn toàn đảm bảo về độ chính xác tuy nhiên đây lại là mô hình có giá trị MAE nhỏ nhất trong toàn báo cáo.

4.2 Yêu cầu 1b: 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

Sơ lược về cách thực hiện:

- Trích 5 cột tương ứng với 5 đặc trưng cần sử dụng từ `X_train` và `X_test`

- Với mỗi đặc trưng, huấn luyện cho mô hình với đặc trưng đó k lần bằng `K-Fold Cross Validation`.
- Tính giá trị MAE trung bình của đặc trưng đó.
- So sánh các giá trị MAE trung bình của 5 đặc trưng, mô hình sử dụng đặc trưng nào tốt nhất thì có giá trị nhỏ nhất.
- Huấn luyện lại mô hình đó với toàn bộ dữ liệu và tính giá trị MAE cuối cùng.

`Kết quả`:
Với k = 10, mô hình tốt nhất có đặc trưng là `nueroticism`.
Bảng các giá trị MAE:

	Feature	MAE
0	conscientiousness	306397.949158
1	agreeableness	301037.130601
2	extraversion	307415.843129
3	nueroticism	299750.683490
4	openess_to_experience	303469.783744

Giá trị MAE cuối cùng: 291019.693226953
Công thức hồi quy:

$$\text{Salary} = -56546.304 \times \textit{nueroticism}$$

Nhận xét:

- So với giá trị MAE ở yêu cầu 1a, MAE ở yêu cầu 1b có giá trị lớn hơn nhiều khi sử dụng mô hình của đặc trưng nueroticism.
- Số lượng đặc trưng chúng ta thêm vào tùy thuộc vào những gì chúng ta muốn mô hình của mình lưu trữ. Nếu đặc trưng này không phù hợp hay liên quan với giá trị mục tiêu và chúng ta thêm nó vào mô hình của mình thì nó sẽ khiến cho mô hình lệch xa khỏi giá trị ban đầu.

4.3 Yêu cầu 1c: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng English, Logical, Quant, tìm mô hình cho kết quả tốt nhất

Sơ lược về cách thực hiện:

- Trích 3 cột tương ứng với 3 đặc trưng cần sử dụng từ `X_train` và `X_test`
- Với mỗi đặc trưng, huấn luyện cho mô hình với đặc trưng đó k lần bằng `K-Fold Cross Validation`.
- Tính giá trị MAE trung bình của đặc trưng đó.
- So sánh các giá trị MAE trung bình của 3 đặc trưng, mô hình sử dụng đặc trưng nào tốt nhất thì có giá trị nhỏ nhất.

- Huấn luyện lại mô hình đó với toàn bộ dữ liệu và tính giá trị MAE cuối cùng.

`Kết quả`:
Với k = 10, mô hình tốt nhất có đặc trưng là `Quant`.
Bảng các giá trị MAE:

	Feature	MAE
0	English	121957.533801
1	Logical	120361.177717
2	Quant	118102.771490

Giá trị MAE cuối cùng: 106819.57761989674
Công thức hồi quy:

$$\text{Salary} = 585.895 \times Quant$$

Nhận xét:

- So với giá trị MAE ở yêu cầu 1a và 1b, MAE ở yêu cầu 1c có giá trị lớn hơn 1a nhưng lại nhỏ hơn 1b khi sử dụng mô hình của đặc trưng Quant. Điều đó cho thấy khả năng định lượng cũng ảnh hương đáng kể tới lương.

4.4 Yêu cầu 1d: Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất

Sơ lược về cách thực hiện:

- Trích các cột cần sử dụng và dùng `preprocess()` để xử lý nếu cần từ `X_train` và `X_test`
- Với mỗi mô hình, huấn luyện cho mô hình đó k lần bằng `K-Fold Cross Validation`.
- Tính giá trị MAE trung bình của mô hình đó.
- So sánh các giá trị MAE trung bình của 3 mô hình, mô hình nào tốt nhất thì có giá trị nhỏ nhất.
- Huấn luyện lại mô hình đó với toàn bộ dữ liệu và tính giá trị MAE cuối cùng.

3 mô hình lần lượt là:

- Mô hình 1: Đánh giá sức ảnh hưởng của Tổng điểm đạt được trong kỳ thi lớp 10, lớp 12 tới Salary.

$$Salary = a \times 10percentage + b \times 12percentage$$

- Mô hình 2: Đánh giá sức ảnh hưởng của GPA tại thời điểm tốt nghiệp và Loại bằng cấp đã đạt tới Salary.

$$Salary = a \times CollegeGPA \times Degree$$

- Mô hình 3: Đánh giá sức ảnh hưởng của Tổng điểm đạt được trong kỳ thi lớp 10, lớp 12, GPA tại thời điểm tốt nghiệp và Loại bằng cấp đã đạt, hạng của trường tới Salary.

$$Salary = a \times 10percentage^2 + b \times 12percentage^2 + c \times (\frac{CollegeGPA \times Degree}{CollegeTier})^2$$

`Kết quả`:
Với k = 10, mô hình tốt nhất có đặc trưng là `Mô hình thứ 3`.
Bảng các giá trị MAE:

	MAE
First Model	120608.770112
Second Model	132722.232001
Third Model	115715.879659

Giá trị MAE cuối cùng: 108453.72292049666
Công thức hồi quy:

$$Salary = 19.117 \times 10percentage^2 + 9.033 \times 12percentage^2 + 26.315 \times (\frac{CollegeGPA \times Degree}{CollegeTier})^2$$

Nhận xét:

- So với giá trị MAE ở yêu cầu 1a, 1b và 1c, MAE ở mô hình thứ 3 nhỏ hơn nhiều so với 1b nhưng vẫn lớn hơn 1a và 1c.
- Khi so sánh 3 mô hình với nhau, dễ dàng thấy được cả 3 điểm số lớp 10, 12 và khi tốt nghiệp đều ảnh hưởng tới Lương và đặc biệt là cũng liên quan tới thứ hạng trường theo học và loại bằng. Hơn nữa như đã phân tích ở trên thì ảnh hưởng GPA khi tốt nghiệp là đáng kể hơn so với điểm lớp 10 và 12.
- Tuy nhiên sai số vẫn là quá lớn khiến cho chúng ta khó tin rằng các yếu tố kể trên thực sự có ảnh hưởng.

Tài liệu

[1] Pandas Developers. Pandas reference. 2023.

[2] satvikvirmani. engineering-graduate-salary-analysis. *github.com*.

[3] ScikitLearn. Cross-validation: evaluating estimator performance.

[4] Trí tuệ nhân tạo. Giới thiệu về k-fold cross-validation. *trituenhantao.io*.

[5] ThS. Phan Thị Phương Uyên. Lab 4: Ols linear regression. 2023.

[6] ThS. Phan Thị Phương Uyên. Nội dung Đồ án 3: Linear regression. 2023.