TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

----------o0o----------

BÁO CÁO ĐỒ ÁN

TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CNTT

Đồ án 3: Linear Regression

Tên sinh viên : Châu Tấn Kiệt
Lớp : 21CLC04
MSSV : 21127329
Giảng viên hướng dẫn : Nguyễn Đình Thúc
Nguyễn Văn Quang Huy
Ngô Đình Hy

Thành phố Hồ Chí Minh, Tháng 8 năm 2023

Mục lục

I. 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.

II. Tiến độ hoàn thành

STT	Mục tiêu	Tỉ lệ hoàn thành
1	Yêu cầu 1a	100%
2	Yêu cầu 1b	100%
3	Yêu cầu 1c	100%
4	Yêu cầu 1d	100%

III. Các thư viện sử dụng

- **pandas**: Đọc csv, xử lí, tính toán trên Dataframe
- **sci-kit learn:** Sử dụng mô hình hồi quy tuyến tính (Linear Regression) và thuật toán K-Fold Cross Validation
- **numpy**: Tính toán trên ma trận nhiều chiều.
- **matplotlib.pyplot**: Vẽ các đồ thị, ma trận.
- **seaborn**: Vẽ ma trận tương quan.
- **statsmodels.stats.outliers_influence:** Tính toán loại bỏ các dữ liệu ngoại lai (outlier)

IV. Các hàm sử dụng

1. Hàm LinearRegression

Hàm LinearRegression được dùng để tìm hồi quy tuyến tính của 1 mô hình.

- linear_model.LinearRegression().fit(a, b): Tìm hệ số cho phương trình hồi quy tuyến tính, sử dụng công thức

$$w = (X^T X)^{-1} X^T y$$

Input:
- a: Mảng giá trị các thuộc tính
- b: Giá trị của kết quả cuối cùng

Output:
- Intercept: Giá trị dự đoán của biến phụ thuộc khi các biến độc lập bằng 0
- Coefficients: Mảng giá trị hệ số của các biến phụ thuộc

2. Hàm findFormula

Dùng để xuất ra các giá trị của Intercept và Coefficient

3. Hàm tính MAE

Input:
- y_test: kết quả thực tế
- y_prediction: kết quả được dự đoán từ mô hình hồi quy

Output:
- mae: sai số tuyệt đối trung bình của 2 kết quả

Ý tưởng cài đặt:

- Áp dụng công thức sai số tuyệt đối trung bình [1]

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

- n: Số lượng phần tử trong bộ dữ liệu
- y_i: Giá trị thực tế
- $\hat{y}_i$: Giá trị dự đoán

4. Hàm CrossValidation

Input:
- a: Mảng giá trị các thuộc tính
- b: Giá trị của kết quả cuối cùng
- k: số tập dữ liệu (số fold) để chia dữ liệu.

Output:
- error_list: mảng chứa MAE của từng đặc trưng có trong dữ liệu nhập vào

Ý tưởng cài đặt:

- K-Fold Cross-validation là một kỹ thuật trong Machine Learning để đánh giá hiệu suất của một mô hình dự đoán. Mục tiêu của nó là đo lường khả năng tổng quát hóa của mô hình trên dữ liệu mới mà nó chưa từng thấy.

- Cách hoạt động chia dữ liệu thành k phần bằng nhau, gọi là "fold". Trong mỗi lần thực hiện, một phần trong số k fold được chọn làm tập kiểm tra (testing set), trong khi các folds còn lại được kết hợp thành tập huấn luyện (training set). Mô hình được huấn luyện trên tập huấn luyện và sau đó được đánh giá trên tập kiểm tra. Quá trình này được thực hiện K lần, trong đó mỗi lần một fold khác nhau được chọn làm tập kiểm tra. Kết quả đánh giá từ K lần thực hiện này được tổng hợp để đưa ra đánh giá cuối cùng về hiệu suất của mô hình.

- Cuối cùng lấy tổng MAES của các mô hình rồi chia trung bình.

5. Hàm chooseBestProperty

Input:
- a: Mảng giá trị các thuộc tính
- b: Giá trị của kết quả cuối cùng

Output:
- result: mảng chứa MAE của từng đặc trưng có trong dữ liệu nhập vào
- np.argmin(result): MAE nhỏ nhất trong mảng, điều này cho thấy đặc trưng tốt nhất trong tập đặc trưng đang xét

6. Hàm replace_outliers_with_iqr

Input:
- data_frame: khung dữ liệu cần chỉnh sửa
- column_name: giá trị của đặc trưng cần chỉnh sửa

Output:
- DataFrame đã được thay thế bằng trung vị (median) của bộ dữ liệu thuộc đặc trưng đó

Ý tưởng cài đặt:

- IQR (Interquartile Range) là một khái niệm thống kê được sử dụng để đo lường phạm vi giữa các phần tư phân vị của một tập dữ liệu số học. IQR thường được sử dụng để đánh giá sự biến đổi và phân phối của dữ liệu mà bỏ qua những giá trị ngoại lai (outliers) có thể gây nhiễu hoặc ảnh hưởng mạnh đến tính chất thống kê của tập dữ liệu.

- Xác định phân vị: Tính toán phân vị thứ nhất (Q1) và phân vị thứ ba (Q3) của tập dữ liệu. Phân vị thứ nhất là giá trị chia tập dữ liệu thành 25% dưới và 75% trên, trong khi phân vị thứ ba chia tập dữ liệu thành 75% dưới và 25% trên.

- Tính IQR bằng cách lấy hiệu giữa phân vị thứ ba và phân vị thứ nhất:

$$IQR = Q3 - Q1.$$

Xác định giá trị biên dưới và biên trên: Để xác định các giá trị ngoại lai trong tập dữ liệu, ta sử dụng giới hạn biên dưới (lower bound) và biên trên (upper bound)

$$\text{Biên dưới: } Q1 - 1.5 * IQR$$

$$\text{Biên trên: } Q3 + 1.5 * IQR$$

Loại bỏ ngoại lai: Tất cả các giá trị dữ liệu nằm ngoài khoảng giới hạn giữa biên dưới và biên trên được coi là ngoại lai và có thể bị loại bỏ khỏi tập dữ liệu.

V. Kết quả và nhận xét

1. Yêu cầu 1a

Yêu cầu:

- Huấn luyện 1 lần duy nhất cho 11 đặc trưng "*Gender*", "*10percentage*", "*12percentage*", "*CollegeTier*", "*Degree*", "*collegeGPA*", "*CollegeCityTier*", "*English*", "*Logical*", "*Quant*", "*Domain*" trên toàn bộ tập huấn luyện.

- Thể hiện công thức cho mô hình hồi quy.

Kết quả:

- Hệ số của các đặc trưng sau khí huấn luyện mô hình:

Đặc trưng	Hệ số
Gender	−23183.330
10percentage	702.767
12percentage	1259.018
CollegeTier	−99570.608
Degree	18369.962
collegeGPA	1297.532
CollegeCityTier	−8836.727
English	141.760
Logical	145.742
Quant	114.643
Domain	34955.750

- Công thức hồi quy (tính Salary theo 11 đặc trưng trên):

$$\text{Salary} = 49248.090 + (-23183.330) * \textbf{Gender} + 702.767 * \textbf{10percentage} + 1259.018 * \textbf{12percentage} + (-99570.608) * \textbf{CollegeTier} + 18369.962 * \textbf{Degree} + 1297.532 * \textbf{collegeGPA} + (-8836.727) * \textbf{CollegeCityTier} + 141.760 * \textbf{English} + 145.742 * \textbf{Logical} + 114.643 * \textbf{Quant} + 34955.750 * \textbf{Domain}$$

- MAE của mô hình trên là: **105052.52978823145**

Nhận xét:

So với các mô hình khác thì mô hình này đạt kết quả tốt hơn (MAE nhỏ hơn). Tuy nhiên trong thực tế mô hình này vì MAE trung bình vẫn là **105052.529** rupee, nếu đổi ra tiền VND thì khoảng hơn 30 triệu đồng, là 1 chênh lệch rất đáng kể

2. Yêu cầu 1b

Yêu cầu:

- Phân tích ảnh hưởng của đặc trưng tính cách dựa trên điểm các bài kiểm tra của AMCAT. Các đặc trưng sử dụng: "*conscientiousness*", "*agreeableness*", "*extraversion*", "*nueroticism*", "*openess_to_experience*".

- Yêu cầu sử dụng k-fold Cross Validation (k tối thiểu là 5) để tìm ra đặc trưng tốt nhất trong các đặc trưng tính cách.

Kết quả:

- 5 kết quả tương ứng cho 5 mô hình từ k-fold Cross Validation (với k = 10):

Mô hình với 1 đặc trưng	MAE
conscientiousness	124142.8455714931
agreeableness	123542.65308903123
extraversion	123910.43408286237
neuroticism	123433.98575751542
openness_to_experience	123865.91570689075

- Đặc trưng tốt nhất (MAE nhỏ nhất) là neuroticism với hệ số sau khi huấn luyện với toàn bộ dữ liệu trong *train.csv*: -16021.494

- Công thức hồi quy (tính Salary theo đặc trưng tốt nhất):

$$\text{Salary} = 304647.553 - 16021.494 * \textbf{Nueroticism}$$

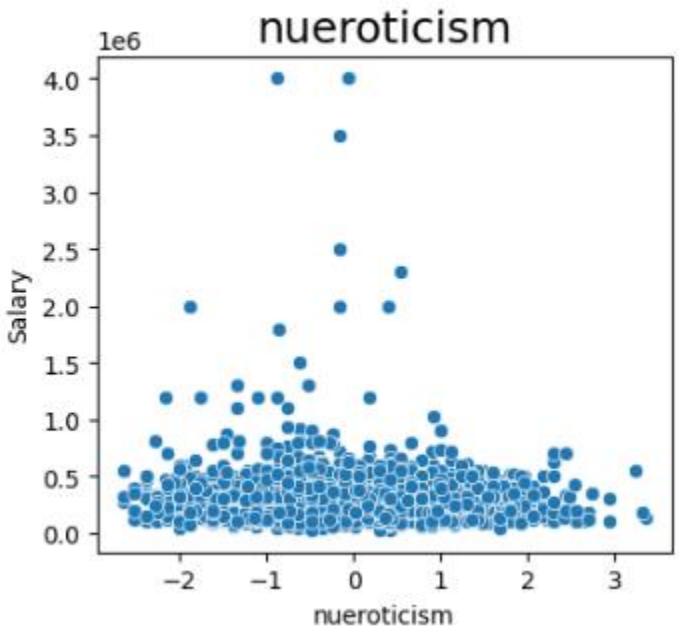
- MAE của mô hình trên là: **120007.64898995696**

Nhận xét:

Mô hình với một đặc trưng tính cách có MAE khá cao, đồng nghĩa với việc dự đoán mô hình sẽ có sự sai lệch với giá trị thực tế khá lớn. Độ chênh lệch trung bình lên đến khoảng 35 triệu VND.

Giả thuyết:

Trong số các đặc trưng tính cách, Neuroticism là đặc trưng có ảnh hưởng nhất đến Salary. Neuroticism - còn gọi là sự nhạy cảm hay sự bất ổn trong cảm xúc của một cá nhân. Đặc điểm tính cách này được thể hiện qua các khía cạnh như: sự lo lắng, tức giận, trầm cảm, sự tự ý thức, nhạy cảm và mức độ dễ bị tổn thương. Giá trị Neuroticism cao cũng đồng nghĩa với việc Salary cũng giảm theo, điều này đã được thể hiện sau khi vẽ mối quan hệ giữa Neuroticism và Salary



3. Yêu cầu 1c

Yêu cầu:

- Phân tích ảnh hưởng của đặc trưng ngoại ngữ, lô-gic, định lượng đến mức lương của các kỹ sư dựa trên điểm các bài kiểm tra của AMCAT. Các đặc trưng sử dụng: "*English*", "*Logical*", "*Quant*".

- Yêu cầu sử dụng k-fold Cross Validation (k tối thiểu là 5) để tìm ra đặc trưng tốt nhất trong các đặc trưng tính cách.

Kết quả:

- 3 kết quả tương ứng cho 3 mô hình từ k-fold Cross Validation (với k = 10):

Mô hình với 1 đặc trưng	MAE
English	120703.21583891594
Logical	119967.51482126901
Quant	117282.01041750479

- Đặc trưng tốt nhất (MAE nhỏ nhất) là Quant với hệ số sau khi huấn luyện với toàn bộ dữ liệu trong *train.csv*: 368.852

- Công thức hồi quy (tính Salary theo đặc trưng tốt nhất):

$$\text{Salary} = 117759.729 + 368.852 * \textbf{Quant}$$

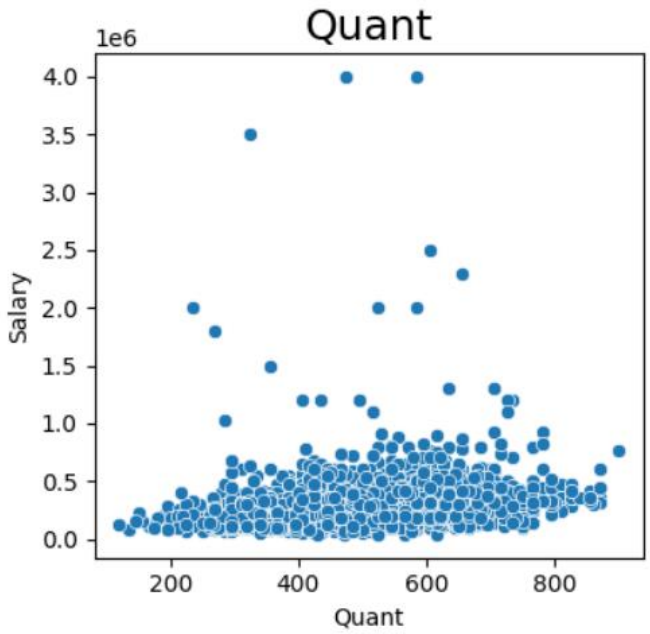
- MAE của mô hình trên là: **125338.42789216392**

Nhận xét:

Mô hình với một đặc trưng định lượng có độ lỗi cao hơn khá nhiều so với 2 mô hình 1 tính cách và mô hình sử dụng 11 đặc trưng đầu tiên.

Giả thuyết

Trong số các đặc trưng kỹ năng, Quant là đặc trưng có ảnh hưởng nhất đến Salary. Neuroticism - còn gọi là khả năng áp dụng Toán học và Thống kê để giải quyết các vấn đề về giải quyết tài chính hoặc quản lí. Giá trị Quant cao cũng đồng nghĩa với việc Salary cũng tăng theo, điều này đã được thể hiện sau khi vẽ mối quan hệ giữa Quant và Salary



4. Yêu cầu 1d

Yêu cầu:

- Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất.

- Xây dựng m mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a, 1b và 1c.

- Sử dụng k-fold Cross Validation (k tối thiểu là 5) để tìm ra đặc trưng tốt nhất trong các đặc trưng tính cách.

Kết quả:

- 4 mô hình tìm được:

+ Mô hình 1: Model 11 đặc trưng nhưng loại bỏ các đặc trưng có hệ số tương quan cao gồm các đặc trưng "*Gender*", "*12percentage*", "*CollegeTier*", "*Degree*", "*collegeGPA*", "*English*", "*Logical*", "*Quant*", "*Domain*"

+ Mô hình 2: Ảnh hưởng của điểm số các môn học kỹ thuật của AMCAT gồm các đặc trưng *'Quant', '10percentage', 'Logical', '12percentage', 'English', 'collegeGPA', 'Domain', 'ComputerProgramming','CollegeTier','nueroticism'*

+ Mô hình 3 : Những thuộc tính có VIF (Variance Inflation Factor) nhỏ hơn 20 gồm 15 đặc trưng sau:

"Gender","Degree","CollegeCityTier","Domain","ComputerProgramming","ElectronicsAnd Semicon","ComputerScience","MechanicalEngg","ElectricalEngg","TelecomEngg",

"CivilEngg","conscientiousness","agreeableness","extraversion","nueroticism", "openess_to_experience"

+ Mô hình 4: Top 10 những đặc trưng có Cross-validation thấp nhất gồm 10 đặc trưng *'Quant', '10percentage', 'Logical', '12percentage', 'English', 'collegeGPA', 'Domain', 'ComputerProgramming', 'CollegeTier', 'nueroticism'*

- 4 kết quả tương ứng cho 4 mô hình từ k-fold Cross Validation (với k = 10):

Mô hình	MAE
Mô hình 1	105236.591630
Mô hình 2	114805.548019
Mô hình 3	111924.334440
Mô hình 4	104324.433231

- Mô hình tốt nhất (mô hình với MAE nhỏ nhất) là Mô hình 4 với hệ số của các đặc trưng sau khi huấn luyện với toàn bộ dữ liệu trong *train.csv* là:

Đặc trưng	Hệ số
Quant	128.601
10percentage	529.78
Logical	120.956
12percentage	1062.286
English	131.316
CollegeGPA	1083.743
Domain	26282.433
ComputerProgramming	67.969
CollegeTier	-99901.372
Nueroticism	-4558.527

- Công thức hồi quy (tính Salary theo 10 đặc trưng trên):

Salary = 71626.094 + 128.601 * **Quant** + 529.780 *** 10percentage** + 120.956 ***Logical** + 1062.287***12percentage** + 131.316*** English** + 1083.743 *** collegeGPA** +

(-26282.433) *** Domain** + 67.969 * **ComputerProgramming** + (-99901.372) ***CollegeTier** (-4588.527) ***Nueroticism**

Nhận xét:

Mô hình 4 có độ lỗi thấp hơn so với các mô hình ở yêu cầu 1a, 1b, 1c. nhưng sự sai lệch khi dự đoán vẫn khá lớn so với thực tế.

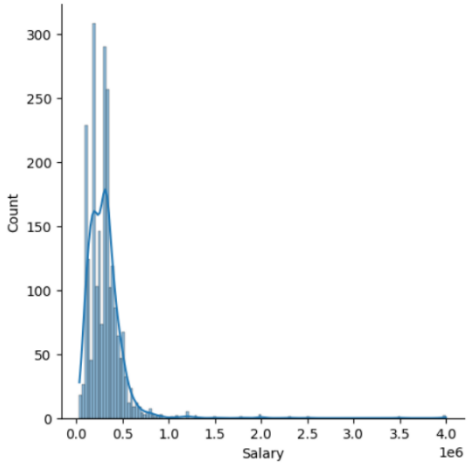
Giả thuyết

Theo kết quả ở yêu cầu 1d thì mô hình tốt nhất tìm được chính là mô hình 4, mô hình sử dụng đặc trưng đã được chọn lọc là có những thuộc tính có Cross-Validation nhỏ nhất. Lý do mô hình này tốt hơn các mô hình còn lại có thể nằm ở việc đây là những thuộc tính có ảnh hưởng nhất dựa trên kết quả Cross-Validation, dẫn đến việc Salary của họ bị ảnh hưởng nhiều hơn. Tuy nhiên, điều này chỉ vẫn là 1 giả thuyết

VI. Quá trình xây dựng mô hình

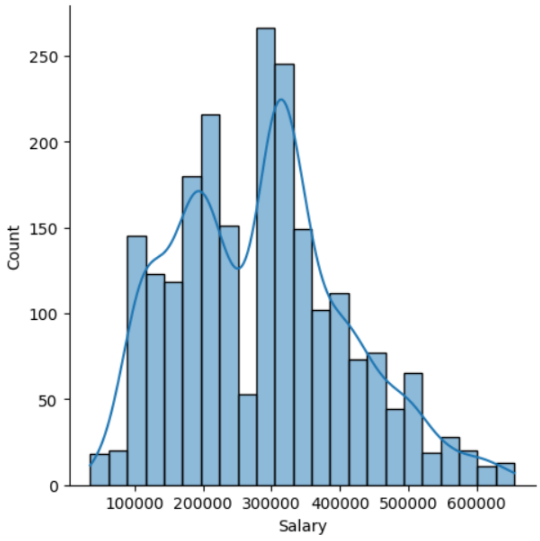
1. Tái xử lý dữ liệu

Trong dữ liệu gốc, có thể thấy dữ liệu đang bị dồn vào 1 chỗ, có khả năng điều này là do các dữ liệu ngoại lai (outlier)



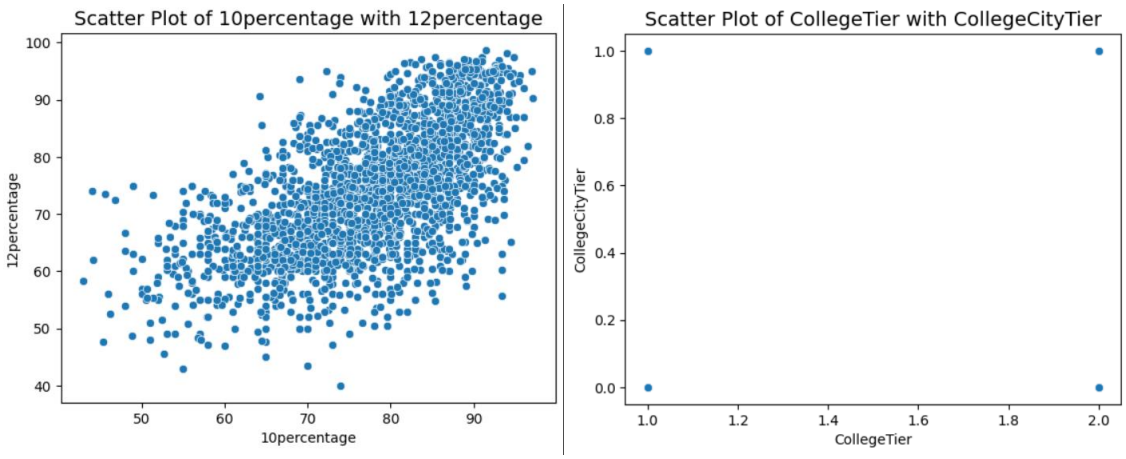
Có rất nhiều cách để xử lí dữ liệu ngoại lai, trong số đó có phương pháp độ trải giữa (IQR)

Đây là dữ liệu sau khi xử lí:



2. Mô hình 1

Trong mô hình 1, tồn tại những đặc trưng có hệ số tương quan lẫn nhau khá cao, dẫn đến việc kết quả cuối sẽ ảnh hưởng không nhiều đến mô hình thực tế, như phân tích ở dưới



Ta thấy rằng có vài đặc trưng có tương quan giữa 10percentage và 12percentage khá là cao nên sẽ loại 10percentage khỏi mô hình (12percentage có vẻ sẽ gần đây hơn) và CollegeCityTier lại không có sự liên quan đến lương nhiều hơn so với CollegeTier, nên chúng ta cũng loại bỏ CollegeCityTier, dẫn đến còn 9 đặc trưng gồm: "*Gender*", "*12percentage*", "*CollegeTier*", "*Degree*", "*collegeGPA*", "*English*", "*Logical*", "*Quant*", "*Domain*"

3. Mô hình 2

Đối với mô hình 2, sử dụng các đặc trưng về điểm thi các môn học kỹ thuật của AMCAT để có thể thấy được ảnh hưởng của chúng đối với lương

4. Mô hình 3

Đối với mô hình 3, lựa chọn các đặc trưng bằng cách lấy ra những đặc trưng có Hệ số lạm phát phương sai nhỏ nhất (VIF), hệ số này đo lường mức độ nghiêm trọng của đa cộng tuyến trong phân tích hồi quy. Nó có thể được sử dụng để đánh giá sức mạnh của mối quan hệ giữa các biến và để mô hình hóa mối quan hệ trong tương lai giữa chúng.

Sau khi sàng lọc dữ liệu, ta có được những đặc trưng có VIF nhỏ hơn 20 gồm 15 thuộc tính:

Gender","Degree","CollegeCityTier","Domain","ComputerProgramming",

"ElectronicsAndSemicon","ComputerScience","MechanicalEngg","ElectricalEngg",

"TelecomEngg","CivilEngg","conscientiousness","agreeableness","extraversion",

"nueroticism", "openess_to_experience"

5. Mô hình 4

Tương tự với yêu cầu và 1b và 1c, chúng ta sẽ lấy những đặc trưng tốt nhất từ trên xuống, dựa trên sai số cross-validation

VII. Tổng kết

Thông qua đồ án này, ta có thể hiểu cách xây dựng một mô hình hồi quy tuyến tính nhằm dự đoán mức lương của cá nhân dựa trên các đặc trưng liên quan như tính cách, tư duy, chuyên môn học thuật, và nhiều yếu tố khác. Tuy nhiên, các mô hình này hiện vẫn chưa đạt được độ chính xác cao (sai số thậm chí có thể lên đến hàng chục triệu VND nếu chuyển đổi sang đơn vị tiền tệ Ấn Độ). Nguyên nhân có thể là do số lượng dữ liệu hiện còn hạn chế, hoặc dữ liệu chưa được làm sạch và tiền xử lý đầy đủ, cũng có thể là do sự phức tạp và mức độ tương quan thấp giữa dữ liệu và đặc trưng cần dự đoán.

Để xây dựng một mô hình có độ chính xác cao, như yêu cầu trong dự án, chỉ số sai số trung bình (MAE) cần phải được duy trì ở mức rất thấp. Điều này sẽ giúp đưa sai số dự đoán lương xuống còn vài trăm nghìn đồng. Tuy nhiên, việc dự đoán mức lương là một nhiệm vụ khó khăn bởi vì có quá nhiều yếu tố thực tế ảnh hưởng đến mức lương và công việc của một người, không giới hạn chỉ trong 24 đặc trưng có sẵn trong bộ dữ liệu. Một mô hình hiệu quả để dự đoán mức lương sẽ phải phức tạp hơn nhiều so với mô hình tuyến tính thông thường.

Như vậy, để đạt được một mô hình dự đoán lương chính xác, chúng ta cần xem xét đến sự phức tạp và đa dạng của các yếu tố ảnh hưởng, và có thể cần áp dụng các mô hình phức tạp hơn, không giới hạn bởi tính tuyến tính, để có khả năng mô phỏng sát hơn với sự phức tạp của thực tế.

VIII. Tài liệu tham khảo

Mean absolute error - Wikipedia

machine learning - Why is 10 considered the default value for k-fold cross-validation? - Data Science Stack Exchange

Interquartile range - Wikipedia

How does a Quant’s salary compare to other employees salary at hedge funds? - Quora

Correlation - Wikipedia

Neurotic Personalities Earn Lower Salaries | Psychology Today