University of Science

Viet Nam National University - Ho Chi Minh city

TOÁN ỨNG DỤNG VÀ THỐNG KÊ
CHO CÔNG NGHỆ THÔNG TIN

Đồ án 3: Linear Regression

Giảng viên:
Vũ Quốc Hoàng
Lê Thanh Tùng
Phan Thị Phương Uyên
Nguyễn Văn Quang Huy

Lớp: 21CLC02
Sinh viên thực hiện:
Lê Hoàng Sang - 21127158

Mục lục

1 Tóm tắt

Đồ án nghiên cứu dự đoán mức lương của kỹ sư thông qua việc sử dụng mô hình hồi quy tuyến tính. Trong đó:

- Yêu cầu 1a: Sử dụng 11 đặc trưng đầu tiên gồm Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain để huấn luyện mô hình hồi quy tuyến tính. Công thức tính y theo 11 đặc trưng được thể hiện. Kết quả trên tập kiểm tra được báo cáo.
- Yêu cầu 1b: Đặc trưng tính cách gồm conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience được thử nghiệm để tìm ra đặc trưng tốt nhất dựa trên K-fold Cross Validation. Kết quả của 5 mô hình từ 5 lần Cross Validation được báo cáo.
- Yêu cầu 1c: Đặc trưng ngoại ngữ, lô-gic và định lượng gồm English, Logical, Quant được thử nghiệm để tìm ra đặc trưng tốt nhất dựa trên k-fold Cross Validation. Kết quả của 3 mô hình từ 3 lần Cross Validation được báo cáo.
- Yêu cầu 1d: Xây dựng và thử nghiệm m mô hình khác nhau, bao gồm việc kết hợp các đặc trưng, sử dụng biến đổi và tạo đặc trưng mới. Mô hình mới được lựa chọn dựa trên kết quả Cross Validation. Tất cả kết quả từ m mô hình được báo cáo.

Trong tóm tắt này, đồ án đã tiếp cận việc dự đoán mức lương thông qua mô hình hồi quy tuyến tính và phân tích ảnh hưởng của các đặc trưng khác nhau, giúp cung cấp thông tin hữu ích cho việc quyết định trong lĩnh vực tuyển dụng và phát triển nguồn nhân lực.

2 Các thư viện đã sử dụng

Các thư viện đã sử dụng trong đồ án này:

- Thư viện **pandas** được sử dụng để làm việc với dữ liệu dưới dạng bảng, gọi là DataFrame. Đây là một cách tiện lợi để tổ chức, xử lý và phân tích dữ liệu. Trong đoạn mã, thư viện này được nhập với tên viết tắt là pd.
- Thư viện **numpy** cung cấp nhiều chức năng liên quan đến tính toán số học và đại số đối với các mảng nhiều chiều. Nó thường được sử dụng để xử lý dữ liệu số học và khoa học. Trong đoạn mã, thư viện này được nhập với tên viết tắt là np.
- Thư viện **scikit-learn** (hay sklearn) cung cấp nhiều công cụ cho việc xây dựng và đánh giá mô hình học máy. Trong đoạn mã, thư viện này được sử dụng để import lớp

KFold từ module model_selection. KFold là một công cụ để thực hiện k-fold cross-validation, một phương pháp để đánh giá hiệu suất của mô hình bằng cách chia dữ liệu thành k phần (fold). Việc sử dụng KFold có thể chia dữ liệu thành các fold, sau đó sử dụng mỗi fold lần lượt làm tập kiểm tra và các fold còn lại làm tập huấn luyện để đánh giá hiệu suất của mô hình trên dữ liệu không nhìn thấy trong quá trình huấn luyện.

3 Mô tả các phương pháp

3.1 Cài đặt các hàm cần thiết

Hàm fit trong lớp OLSLinearRegression:

- Input: X: Ma trận dữ liệu huấn luyện, mỗi hàng là một điểm dữ liệu, mỗi cột là một đặc trưng; y: Mảng chứa giá trị mục tiêu tương ứng với mỗi điểm dữ liệu trong X.
- Output: Trả về chính đối tượng OLSLinearRegression đã được fit mô hình (không trả về giá trị).
- Ý tưởng: Tính ma trận pseudo-inverse (nghịch đảo giả) của ma trận X: X_pinv = np.linalg.inv(X.T @ X) @ X.T; Tính vector trọng số w của mô hình hồi quy: self.w = X_pinv @ y.

Hàm get_params trong lớp OLSLinearRegression:

- Input: Không có input.
- Output: Trả về vector trọng số w đã được học từ mô hình.

Hàm predict trong lớp OLSLinearRegression:

- Input: X: Ma trận dữ liệu kiểm tra hoặc dự đoán, tương tự như trong hàm fit.
- Output: Trả về mảng chứa các giá trị dự đoán tương ứng với mỗi điểm dữ liệu trong X.
- Ý tưởng: Tính dự đoán bằng cách nhân ma trận X với vector trọng số w đã học: np.sum(self.w.ravel() * X, axis=1).

Hàm model_mae:

- Input: y_test: Mảng chứa giá trị thực tế; y_pred: Mảng chứa giá trị dự đoán.

- Output: Trả về giá trị Mean Absolute Error (MAE) giữa các giá trị thực tế và dự đoán.
- Ý tưởng: Tính giá trị tuyệt đối của sai số tương ứng giữa giá trị thực tế và dự đoán: absolute_errors = [abs(actual - predicted) for actual, predicted in zip(y_test, y_pred)]. Tính MAE bằng cách lấy tổng các sai số tuyệt đối và chia cho số lượng điểm dữ liệu: mae = sum(absolute_errors) / len(y_test).

Phần này này triển khai một mô hình hồi quy tuyến tính cơ bản (OLSLinearRegression) cùng với hàm tính MAE để đánh giá mô hình.

3.2 Yêu cầu 1a

Sử dụng 11 đặc trưng đầu tiên đề bài cung cấp bao gồm: ‘Gender‘, ‘10percentage‘, ‘12percentage‘, ‘CollegeTier‘, ‘Degree‘, ‘collegeGPA‘, ‘CollegeCityTier‘, ‘English‘, ‘Logical‘, ‘Quant‘ và ‘Domain‘. Hướng giải quyết yêu cầu a như sau.

Trích xuất đặc trưng và dữ liệu: Dòng code đầu tiên trích xuất 11 đặc trưng cần thiết từ tập dữ liệu huấn luyện (X_train) và tập dữ liệu kiểm tra (X_test) để tạo X_train1a và X_test1a. Các đặc trưng này bao gồm Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, và Domain.

Huấn luyện mô hình Linear Regression cho yêu cầu 1a: Sử dụng hàm train_1a để huấn luyện mô hình dự đoán mức lương sử dụng các đặc trưng từ X_train1a và dữ liệu mục tiêu từ y_train. Hàm train_1a tạo một thể hiện của lớp OLSLinearRegression và gọi hàm fit trên thể hiện này, làm việc này để tính toán và lưu trọng số của mô hình.

Kiểm tra mô hình và tính toán MAE: Sử dụng hàm question_1a_MAE_check để kiểm tra hiệu suất của mô hình trên tập kiểm tra. Hàm này tính MAE bằng cách so sánh giá trị thực tế từ y_test và giá trị dự đoán từ mô hình y_pred.

In kết quả: Kết quả MAE và vector trọng số w được in ra để đánh giá hiệu suất của mô hình và trọng số tương ứng với các đặc trưng.

$$\text{MAE: } 104863.778 \tag{1}$$

$$\begin{aligned}\text{Salary} = (-22756.513) \cdot \text{Gender} + 804.503 \cdot \text{10percentage} & \quad (2)\\ + 1294.655 \cdot \text{12percentage} + (-91781.898) \cdot \text{CollegeTier} & \quad (3)\\ + 23182.389 \cdot \text{Degree} + 1437.549 \cdot \text{collegeGPA} & \quad (4)\\ + (-8570.662) \cdot \text{CollegeCityTier} + 147.858 \cdot \text{English} & \quad (5)\\ + 152.888 \cdot \text{Logical} + 117.222 \cdot \text{Quant} & \quad (6)\\ + 34552.286 \cdot \text{Domain} & \quad (7)\end{aligned}$$

3.3 Yêu cầu 1b

Phân tích ảnh hưởng của đặc trưng tính cách dựa trên điểm các bài kiểm tra của AMCAT gồm: ‘conscientiousness‘, ‘agreeableness‘, ‘extraversion‘, ‘nueroticism‘ và ‘openess_to_experience‘. Hướng giải quyết yêu cầu b như sau.

Chuẩn bị các dữ liệu và đặc trưng cần thiết: features_1b: Danh sách các đặc trưng bao gồm các yếu tố tính cách và Salary; train_1b: DataFrame chứa tất cả các đặc trưng cần thiết để thực hiện yêu cầu 1b.

Hàm question_1b_handler: thực hiện cross-validation để tìm ra đặc trưng tính cách tốt nhất dựa trên MAE trung bình của mô hình hồi quy tuyến tính với mỗi đặc trưng. shuffle_data là một đối tượng KFold được sử dụng để chia dữ liệu thành các fold và thực hiện cross-validation. Vòng lặp chạy qua mỗi fold và tính MAE trung bình cho mỗi đặc trưng, sau đó lưu vào mae_arr. Cuối cùng, trả về mảng MAE trung bình và đặc trưng tốt nhất dựa trên np.argmin(mae_arr).

Thực hiện cross-validation và tìm đặc trưng tốt nhất: train_1b được chia thành các fold và đặc trưng tốt nhất được tìm ra thông qua hàm question_1b_handler. Kết quả của đặc trưng tốt nhất được in ra.

Huấn luyện và đánh giá mô hình với đặc trưng tốt nhất: Dựa vào đặc trưng tốt nhất tìm thấy, dữ liệu huấn luyện và kiểm tra được chuẩn bị. Mô hình hồi quy tuyến tính train_1a được huấn luyện và MAE cùng trọng số của mô hình được tính toán. Kết quả trọng số và MAE của mô hình sử dụng đặc trưng tốt nhất được in ra.

Tóm lại, phần code này thực hiện tìm đặc trưng tính cách tốt nhất thông qua cross-validation và sử dụng đặc trưng này để huấn luyện và đánh giá mô hình hồi quy tuyến tính. Kết quả cuối cùng bao gồm trọng số của mô hình và MAE trên tập kiểm tra.

$$\text{MAE: } 291019.693 \tag{8}$$

$$\text{Salary} = (-56546.304) \cdot \text{nueroticism} \tag{9}$$

3.4 Yêu cầu 1c

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 gồm: 'English', 'Logical' và 'Quant'.

Tương tự như yêu cầu 1b, yêu cầu 1c tập trung vào việc xác định ảnh hưởng của một nhóm đặc trưng cụ thể đến mức lương của kỹ sư. Trong yêu cầu này, ta quan tâm đến các đặc trưng liên quan đến năng lực ngôn ngữ, logic và định lượng, bao gồm English, Logical, và Quant. Mục tiêu vẫn là xác định đặc trưng nào có ảnh hưởng lớn nhất thông qua k-fold Cross Validation.

Trong đoạn mã code, ta tiến hành tách các đặc trưng từ tập dữ liệu huấn luyện và kiểm tra để sử dụng cho mô hình hồi quy. Tiếp theo, ta sử dụng k-fold Cross Validation để tìm ra đặc trưng tốt nhất trong nhóm các đặc trưng đã chỉ định. Quá trình này tương tự như trong yêu cầu 1b. Mảng mae_arr_1c lưu trữ các giá trị MAE tương ứng với mỗi đặc trưng.

Sau khi xác định được đặc trưng tốt nhất, ta huấn luyện lại mô hình hồi quy trên toàn bộ tập dữ liệu huấn luyện bằng việc sử dụng đặc trưng tốt nhất này. Sau đó, ta thực hiện dự đoán trên tập kiểm tra và tính MAE trên tập kiểm tra để đánh giá hiệu suất của mô hình.

$$\text{MAE: } 106819.578 \tag{10}$$

$$\text{Salary} = 585.895 \cdot \text{Quant} \tag{11}$$

3.5 Yêu cầu 1d

3.5.1 Tìm hiểu các mô hình

Các outlier tham khảo trong bài Salary Prediction of Engineering Students:

	Outlier_percentage
Gender	23.882588
ComputerScience	23.348899
TelecomEngg	9.139426
Degree	8.038692
CollegeTier	7.538359
MechanicalEngg	6.237492
Domain	5.970647
ElectricalEngg	4.069380
Salary	2.601734
openess_to_experience	2.301534
agreeableness	1.634423
conscientiousness	1.367578
collegeGPA	1.067378
extraversion	1.000667
CivilEngg	0.867245
Quant	0.733823
10percentage	0.567045
Logical	0.500334
English	0.400267
nueroticism	0.400267
GraduationYear	0.066711
ComputerProgramming	0.033356
ElectronicsAndSemicon	0.033356
12percentage	0.033356
Specialization	0.000000
CollegeCityTier	0.000000
12board	0.000000

Trong hồi quy tuyến tính (linear regression), outlier là các điểm dữ liệu nằm xa khỏi phạm vi dự đoán được bởi mô hình hồi quy. Outlier có thể là những điểm dữ liệu đơn lẻ hoặc

nhóm các điểm dữ liệu có giá trị rất khác biệt so với các điểm dữ liệu khác trong tập dữ liệu. Outlier có thể ảnh hưởng đáng kể đến hiệu suất và độ tin cậy của mô hình hồi quy tuyến tính. Một số tác động của outlier bao gồm:

- Ảnh hưởng đến Đường Hồi Quy: Các điểm outlier có thể ảnh hưởng đến phương trình đường hồi quy tuyến tính, dẫn đến sự sai lệch trong dự đoán của mô hình.
- Dẫn Đến Không Ổn Định: Outlier có thể dẫn đến sự không ổn định trong các tham số của mô hình. Mô hình có thể trở nên quá nhạy cảm với các điểm dữ liệu nằm xa.
- Tác Động Lên Độ Chính Xác: Outlier có thể ảnh hưởng đến các độ đo đánh giá hiệu suất của mô hình như sai số bình phương trung bình (MSE) hoặc hệ số xác định (R-squared).

Trên thực tế, ta có thể thấy các thuộc tính về chuyên môn của ngành kĩ sư là các yếu tố quan trọng ảnh hưởng trực tiếp đến năng lực và phần nào dẫn đến sự chênh lệch lương giữa các kĩ sư.

Trong mô hình hồi quy tuyến tính, việc bình phương (quadratic) hoặc lấy căn bậc hai (square root) các đặc trưng (features) có thể được sử dụng để tạo ra mối quan hệ phi tuyến giữa các đặc trưng và biến mục tiêu (target variable). Mô hình tuyến tính trong trường hợp này được gọi là mô hình hồi quy đa thức (polynomial regression).

Việc áp dụng bình phương hoặc lấy căn bậc hai các đặc trưng có thể có tác dụng tạo ra các đặc trưng mới dựa trên đặc trưng gốc, mở rộng phạm vi của đa thức mô hình và cho phép nắm bắt được các mô hình không tuyến tính.

Cách này có thể giúp mô hình hồi quy tuyến tính bám sát hơn với các biểu đồ dữ liệu có dạng không tuyến tính, vì nó giúp mô hình biểu diễn được các biến đổi phi tuyến của đặc trưng và mục tiêu.

3.5.2 Triển khai các mô hình

Xây dựng mô hình, tìm mô hình cho kết quả tốt nhất từ m mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở a, b và c.
Ta xây dựng 6 mô hình dữ liệu sau bằng hàm data_constructor, mỗi là option truyền vào là một bộ dữ liệu.

- option = 0: Dựa trên điểm tốt nghiệp ở trường đại học và điểm các bài kiểm tra AMCAT liên quan đến chuyên môn các lĩnh vực kĩ sư, ta lấy các cột: $collegeGPA$, $ComputerScience$, $ComputerProgramming$, $ElectronicsAndSemicon$, $TelecomEngg$, $Salary$.

- option = 1: Dựa trên điểm AMCAT và căn bậc hai của một số thuộc tính liên quan đến kỹ năng chuyên môn. Lựa chọn các cột và áp dụng căn bậc hai: $English$, $Logical$, $Quant$, $\sqrt{Domain}$, $\sqrt{ComputerProgramming}$, $\sqrt{ElectronicsAndSemicon}$, $\sqrt{ComputerScience}$, $\sqrt{MechanicalEngg}$, $\sqrt{ElectricalEngg}$, $\sqrt{TelecomEngg}$, $\sqrt{CivilEngg}$, $conscientiousness$, $agreeableness$, $extraversion$, $nueroticism$, $openess\_to\_experience$, $Salary$.

- option = 2: Dựa trên điểm AMCAT và bình phương của các thuộc tính liên quan đến kỹ năng chuyên môn và điểm tính cách của bài kiểm tra AMCAT (lần lượt thử bình phương các tính cách AMCAT và extraversion cho kết quả tốt nhất). Lựa chọn các cột và bình phương: $English$, $Logical^2$, $Quant^2$, $Domain^2$, $ComputerProgramming^2$, $ElectronicsAndSemicon^2$, $ComputerScience^2$, $MechanicalEngg^2$, $ElectricalEngg^2$, $TelecomEngg^2$, $CivilEngg^2$, $conscientiousness$, $agreeableness$, $extraversion^2$, $nueroticism$, $openess\_to\_experience$, $Salary$.

- option = 3: Dựa trên điểm AMCAT chuyên ngành và bình phương của điểm bốn tham số đầu tiên (chủ yếu về logic/ tư duy). Lựa chọn các cột và bình phương: $English^2$, $Logical^2$, $Quant^2$, $Domain^2$, $ComputerProgramming$, $ElectronicsAndSemicon$, $ComputerScience$, $MechanicalEngg$, $ElectricalEngg$, $TelecomEngg$, $CivilEngg$, $Salary$.

- option = 4: Dựa trên điểm thi trung học và điểm trường đại học. Lựa chọn các cột: $10percentage$, $12percentage$, $collegeGPA$, $Salary$.

- option = 5: Chỉ dựa trên điểm trường đại học. Lựa chọn các cột: $collegeGPA$, $Salary$.

Đầu tiên, hàm question_1d_handler được định nghĩa để xử lý việc huấn luyện và đánh giá nhiều mô hình khác nhau. Hàm này nhận vào tập huấn luyện, số lượng mô hình cần thử nghiệm (number_of_models), và số lượng phân chia (k_fold) cho quá trình Cross Validation.

Trong vòng lặp chính, với mỗi lần phân chia, các mô hình khác nhau được xây dựng bằng cách gọi hàm data_constructor để chọn loại dữ liệu cần sử dụng cho mô hình. Sau đó, tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra.

Cho mỗi mô hình, các đặc trưng và nhãn của tập huấn luyện và kiểm tra được chuẩn bị. Mô hình được huấn luyện bằng cách gọi hàm train_1a và đánh giá trên tập kiểm tra bằng cách gọi hàm question_1a_MAE_check. Kết quả MAE được tích luỹ trong mảng mae_arr.

Sau khi hoàn thành vòng lặp, MAE của từng mô hình trong các lần Cross Validation được

tính trung bình và in ra. Hàm trả về mảng MAE và chỉ số của mô hình có MAE nhỏ nhất. Dưới phần code cho yêu cầu 1d, số lượng mô hình (m) được khai báo và hàm question_1d_handler được gọi để tìm mô hình tốt nhất dựa trên tập huấn luyện.

Kết quả mô hình tốt nhất và chỉ số của mô hình được in ra. Tiếp theo, mô hình tốt nhất được huấn luyện lại trên toàn bộ tập huấn luyện và tập kiểm tra. Cuối cùng, kết quả trọng số của mô hình và MAE trên tập kiểm tra với mô hình tốt nhất được in ra.

Tóm lại, code trong câu 1d tập trung vào việc xây dựng và đánh giá nhiều mô hình khác nhau để tìm ra mô hình có hiệu suất tốt nhất dựa trên kết quả Cross Validation.

$$\text{MAE: } 104739.987 \tag{12}$$

$$\begin{aligned}
\text{Salary} &= 229.513 \cdot \text{English} + 153.858 \cdot \text{Logical}^2 && (13)\\
&+ 200.945 \cdot \text{Quant} + 355.670 \cdot \text{Domain}^2 && (14)\\
&+ 249.029 \cdot \text{ComputerProgramming}^2 + 470.731 \cdot \text{ElectronicsAndSemicon}^2 && (15)\\
&- 362.753 \cdot \text{ComputerScience}^2 + 234.047 \cdot \text{MechanicalEngg}^2 && (16)\\
&- 174.574 \cdot \text{ElectricalEngg}^2 - 778.368 \cdot \text{TelecomEngg}^2 && (17)\\
&+ 560.902 \cdot \text{CivilEngg}^2 - 188.116 \cdot \text{conscientiousness} && (18)\\
&+ 172.035 \cdot \text{agreeableness} - 194.205 \cdot \text{extraversion}^2 && (19)\\
&- 876.389 \cdot \text{nueroticism} - 698.238 \cdot \text{openess_to_experience} && (20)
\end{aligned}$$

4 Nhận xét

4.1 Yêu cầu 1a

Gender: Hệ số hồi quy là -22756.513. Khi Gender tăng lên 1 đơn vị (từ nam sang nữ), dự đoán mức lương sẽ giảm đi 22756.513 đơn vị. Điều này cho thấy rằng mô hình dự đoán mức lương thấp hơn đối với nữ so với nam giới.

10percentage: Hệ số hồi quy là 804.503. Mỗi lần tăng 1 đơn vị trong tỉ lệ điểm số lớp 10 dự đoán mức lương tăng 804.503 đơn vị. Điều này cho thấy mối quan hệ dương đáng kể giữa điểm số kì thi lớp 10 và mức lương.

12percentage: Hệ số hồi quy là 1294.655. Tương tự như 10percentage, mức lương tăng 1294.655 đơn vị cho mỗi đơn vị tăng trong tỉ lệ điểm số kì thi lớp 12.

CollegeTier: Hệ số hồi quy là -91781.898. Điều này cho thấy rằng các trường đại học hạng cao hơn thường có mức lương dự đoán cao hơn.

Degree: Hệ số hồi quy là 23182.389. Các loại bằng cấp học vấn khác nhau có ảnh hưởng đến mức lương. Hệ số dương cho thấy mối quan hệ tích cực giữa loại bằng cấp và mức lương.

collegeGPA: Hệ số hồi quy là 1437.549. Mức lương tăng 1437.549 đơn vị cho mỗi đơn vị tăng trong collegeGPA. Điều này cho thấy mối quan hệ tích cực giữa điểm trung bình đại học và mức lương.

CollegeCityTier (Loại thành phố trường đại học): Hệ số hồi quy là -8570.662. Có sự phụ thuộc của loại thành phố trường đại học vào mức lương, với thành phố cấp cao có mức lương dự đoán cao hơn.

English (Điểm môn Anh): Hệ số hồi quy là 147.858. Mức lương tăng 147.858 đơn vị cho mỗi đơn vị tăng trong điểm môn Anh.

Logical (Điểm môn Logic): Hệ số hồi quy là 152.888. Mức lương tăng 152.888 đơn vị cho mỗi đơn vị tăng trong điểm môn Logic.

Quant (Điểm môn Toán): Hệ số hồi quy là 117.222. Mức lương tăng 117.222 đơn vị cho mỗi đơn vị tăng trong điểm môn Toán.

Domain (Chỉ số Domain): Hệ số hồi quy là 34552.286. Mức lương tăng 34552.286 đơn vị cho mỗi đơn vị tăng trong chỉ số Domain.

Phương trình hồi quy tuyến tính cho thấy mức lương có sự phụ thuộc mạnh mẽ vào nhiều yếu tố như điểm số học tập, loại học vấn, thành phố trường đại học, và các chỉ số khác.

4.2 Yêu cầu 1b

nueroticism (tạm dịch là chủ nghĩa về thần kinh): Hệ số hồi quy là -56546.304. Khi giá trị của nueroticism tăng lên 1 đơn vị, mức lương dự đoán sẽ giảm đi 56546.304 đơn vị. Hệ số âm cho thấy mức lương có xu hướng giảm khi tính cách về thần kinh (nueroticism) tăng lên.

Phương trình hồi quy tuyến tính này chỉ sử dụng một biến độc lập duy nhất là nueroticism để dự đoán mức lương. Dựa trên hệ số hồi quy, mức lương dự đoán có mối quan hệ âm với chủ nghĩa về thần kinh, tức mức lương giảm chủ nghĩa về thần kinh tăng lên.

4.3 Yêu cầu 1c

Quant (khả năng định lượng): Hệ số hồi quy là 585.895. Khi giá trị của Quant tăng lên 1 đơn vị, mức lương dự đoán sẽ tăng lên 585.895 đơn vị. Hệ số dương cho thấy mức lương có xu hướng tăng khi khả năng định lượng (Quant) tăng lên.

Phương trình hồi quy tuyến tính này chỉ sử dụng một biến độc lập duy nhất là Quant để dự đoán mức lương. Dựa trên hệ số hồi quy, mức lương dự đoán có mối quan hệ dương với điểm môn Toán, tức là mức lương tăng khi điểm môn Toán tăng lên.

4.4 Yêu cầu 1d

Các biến liên quan đến ngôn ngữ và kiến thức chuyên môn như English, Logical, Quant, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, và CivilEngg đều có hệ số hồi quy dương. Mức lương có xu hướng tăng khi các biến này tăng lên. Các biến Logical, ComputerProgramming, ElectronicsAndSemicon, và MechanicalEngg có mối quan hệ bậc hai, cho thấy mức lương tăng không đều khi giá trị của chúng tăng.

Các biến liên quan đến tính cách như conscientiousness, agreeableness, extraversion, nueroticism, và openess_to_experience đều có hệ số hồi quy, với các giá trị dương và âm khác nhau. Mức lương có thể tăng hoặc giảm tùy thuộc vào giá trị của từng biến tính cách. Ví dụ, mức lương có xu hướng tăng khi conscientiousness và agreeableness tăng, nhưng có thể giảm khi extraversion, nueroticism, và openess_to_experience tăng.

Phương trình hồi quy tuyến tính này là một mô hình phức tạp sử dụng nhiều biến độc lập để dự đoán mức lương. Mô hình này thể hiện mối quan hệ phức tạp giữa các biến độc lập và mức lương, bao gồm cả mối quan hệ tuyến tính và phi tuyến tính.

Số MAE ở câu d là 104739.987, thấp hơn so với các câu trước a, b và c. Điều này có ý nghĩa là mô hình tại câu d có hiệu suất tốt hơn trong việc dự đoán mức lương trên dữ liệu kiểm tra so với các mô hình ở câu a, b và c. MAE thấp hơn chỉ ra rằng mức sai số trung bình giữa giá trị dự đoán và giá trị thực tế là nhỏ hơn, tức là mô hình dự đoán tốt hơn.

Tuy nhiên, mặc dù MAE thấp hơn, việc đánh giá mô hình chỉ dựa trên giá trị MAE không đủ để đưa ra kết luận cuối cùng về hiệu suất của mô hình. Cần kiểm tra nhiều yếu tố khác như tương quan, phân phối của biến, độ tin cậy của hệ số hồi quy, kiểm tra tính linh hoạt của mô hình và có thể sử dụng các phương pháp đánh giá khác như R-squared để có cái nhìn tổng quan về hiệu suất của mô hình.

5 Kết luận

Tóm lại, đồ án này nghiên cứu về Linear Regression cho ta có cái nhìn tổng quan hơn về việc huấn luyện các bộ dữ liệu và dự đoán từ các mô hình có tác động lớn trong thực tế: dựa vào các chỉ số đặc trưng chung của các kĩ sư để dự đoán được mức lương của họ. Thông qua đó, ta có cái nhìn tổng quát hơn về Machine Learning và lĩnh vực Khoa học máy tính.

6 Đánh giá mức độ hoàn thành các yêu cầu

Bảng 1: Bảng thống kê mức độ hoàn thành các yêu cầu

STT	Chức năng	Đánh giá	Ghi chú
1	Yêu cầu 1a	100%	
2	Yêu cầu 1b	100%	
3	Yêu cầu 1c	100%	
4	Yêu cầu 1d	100%	Thử nghiệm 6 mô hình

7 Tham khảo

[1] Kiến thức môn Toán Ứng dụng và Thống kê - Khoa Công nghệ thông tin, Trường đại học Khoa học Tự nhiên TPHCM

[2] Mục 2. Exploratory Data Analysis - Enginering Graduate Salary Prediction Using Regression Model - Tác giả: Muh Asdar - RPubs

[3] Mục IV. DATA EXPLORATION - Salary Prediction for Computer Engineering Positions in India - Các tác giả: Ashty Kamal Mohamed Saeed, Pavel Abdullah, - ResearchGate

[4] Engineering Graduate Salary Prediction - Kaggle - Tác giả: MANISH KC

[5] File engineering-graduate-salary-prediction.ipynb - engineering-graduate-salary-analysis - Tác giả: satvikvirmani

[6] Phân tích Khám phá Dữ liệu EDA - machinelearningcoban - Tác giả: Tiep Vu

[7] Salary Prediction of Engineering Students - Kaggle - Tác giả: SUJITH K MANDALA

[8] numpy.ravel

[9] sklearn.model_selection.KFold