Giáo viên:

Vũ Quốc Hoàng – Nguyễn Văn Quang Huy – Ngô Đình Hy – Phan Thị Phương Uyên

21127428	Phạm Nguyễn Quốc Thanh

Giới thiệu 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.

Các thư viện sử dụng:

[1] numpy
[2] pandas
[3] IPython
[4] sklearn

Các hàm và lớp được sử dụng:

def mae(y, y_hat): Tính độ đo MAE. Sử dụng công thức ở phía dưới để tính toán. Hàm ravel được dùng để đưa giá trị y và giá trị y mũ về mảng 1 chiều và mean được dùng để tính trung bình.

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

Lớp OLSLinearRegression: lớp được dùng để sử dụng mô hình hồi quy tuyến tính. Trong đấy, có 3 hàm chính:

def fit(self, X, y): Được sử dụng để tính toán các trọng số cho mô hình dựa trên giá trị mục tiêu. Sử dụng công thức ở phía dưới để tính toán. Trong đấy, X_pinv tương đương với $(X^T X)^{-1} X^T$, hàm linalg.inv được dùng để lấy nghịch đảo của $(X^T X)$. Dấu "@" được dùng như một phép nhân giữa các ma trận.

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

def get_params(self): Được sử dụng để lấy các trọng số w sau khi khớp dữ liệu cho mô hình.

def predict(self, X): Được sử dụng để tính các giá trị mục tiêu mới có thể có sau khi khớp dữ liệu, sử dụng các trọng số đã được tính ở phía trên và một bộ dữ liệu X.

def format_float(x): Được sử dụng để định dạng lại các giá trị MAE làm tròn đến 3 chữ số.

Kết quả và nhận xét:

a. Sử dụng 11 đặc trưng đầu tiên:

- Giải thích:
  - lr_1a là biến được dùng để gọi lớp OLSLinearRegression và gọi hàm fit để khớp dữ liệu giữa 11 đặc trưng đầu tiên và tập huấn luyện.
  - y_hat là các giá trị mục tiêu mới sau khi đã khớp mô hình dữ liệu sử dụng hàm predict ở lớp OLSLinearRegression với tham số X là first_11_data_test.
  - Gọi hàm mae để tính độ đo MAE giữa giá trị mục tiêu của tập kiểm tra và giá trị mục tiêu của mô hình.
- Kết quả:

w	Đặc trưng	Giá trị
w_1	Gender	-22756.513
w_2	10percentage	804.503
w_3	12percentage	1294.655
w_4	CollegeTier	-91781.898
w_5	Degree	23182.389
w_5	collegeGPA	1437.549
w_7	CollegeCityTier	-8570.662
w_8	English	147.858
w_9	Logical	152.888
w_{10}	Quant	117.222
w_{11}	Domain	34552.286
MAE		104863.778

- Công thức:

$$Salary = -22756.513 \times Gender + 804.503 \times 10percentage + 1294.655 \times 12percentage + -91781.898 \times CollegeTier + 23182.389 \times Degree + 1437.549 \times collegeGPA + -8570.662 \times CollegeCityTier + 147.858 \times English + 152.888 \times Logical + 117.222 \times Quant + 34552.286 \times Domain$$

b. Sử dụng các đặc trưng tính cách để tìm mô hình tốt nhất:

- Ý tưởng: Dựa vào [1], ta muốn xem xét rằng liệu tính cách nào là tốt nhất cho các kỹ sư. Có 5 tính cách chính: 'conscientiousness', 'agreeableness', 'extraversion', 'neuroticism' và 'openess_to_experience'. Dựa vào kỹ thuật k-fold cross validation để tính độ đo MAE tốt nhất cho mỗi đặc trưng. Với mỗi tính cách, ta đều chia ra k_fold bộ dữ liệu dựa trên bộ dữ liệu đã được xáo trộn. Thực hiện xáo trộn trên toàn bộ tập huấn luyện để mỗi bộ dữ liệu đều được khớp với giá trị mục tiêu đã có. Sử dụng 1 bộ dữ liệu để huấn luyện và k – 1 bộ còn lại để tính sai số.
- Giải thích: train_1b là bộ dữ liệu đã được xáo trộn. y_train_1b được dùng để lưu lại các giá trị mục tiêu sau khi xáo trộn. models_train là danh sách bao gồm các mô hình sẽ được sử dụng. Ở đây là 'conscientiousness', 'agreeableness', 'extraversion', 'neuroticism' và 'openess_to_experience'. Tất cả mô hình đều được reshape lại thành 1 cột. models_test cũng

tương tự nhưng là dành cho sau khi đã kiếm được mô hình tốt nhất. average_maes là danh sách các kết quả MAE của mỗi mô hình. Sử dụng hàm cross_val_score [] và mean để tính toán giá trị trung bình của mỗi mô hình trong k_fold lần. best_model_index là biến được dùng để lấy index của mô hình có giá trị MAE thấp nhất. df là bảng so sánh các mô hình và giá trị MAE.

- Kết quả:

Mô hình với 1 đặc trưng	MAE
neuroticism	299277.037
agreeableness	300788.191
openess_to_experience	302950.151
conscientiousness	306141.932
extraversion	306920.779

- Nhận xét: Có thể thấy rằng, 'neuroticism' có kết quả MAE tốt nhất trong 5 đặc trưng. Để lý giải cho điều này, ta vốn biết rằng đặc thù công việc của 1 kỹ sư vốn rất áp lực. Về neuroticism, một người sở hữu chỉ số cao về tính nhạy cảm(neuroticism) có các đặc điểm như: lo lắng nhiều thứ, hay overthinking, thường xuyên cảm thấy buồn phiền, tâm trạng bồn chồn, bất an. Ngược lại, những người có chỉ số thấp lại có các đặc điểm như: cảm xúc ở mức ổn định, biết cách đối mặt với sự căng thẳng… Chính vì vậy, ta có thể thấy rằng neuroticism có thể phù hợp với kỹ sư và trọng số của neuroticism cũng là một số âm, (được thể hiện ở phần sau) cho rằng chỉ số thấp sẽ ảnh hưởng tới mức lương nhiều hơn. Tuy nhiên, ta cũng có thể thấy rằng độ đo MAE của cả 5 đặc trưng đều khá lớn so với độ đo MAE ở câu a. Điều này có nghĩa là mặc dù tính cách có ảnh hưởng tới mức lương của kỹ sư nhưng lại không quá nhiều.
- Mô hình tính cách tốt nhất: neuroticism.
    - Sau khi tính toán và so sánh các đặc trưng tính cách, ta lấy ra index của đặc trưng tốt nhất bằng cách sử dụng hàm index(min(average_maes)) và lưu vào biến best_model_index. my_best_personality_feature_model sẽ là mô hình sử dụng tính cách tốt nhất và khớp dữ liệu với bộ dữ liệu y_train_1b đã được xáo trộn. X_Para là bộ dữ liệu sẽ được sử dụng để tính toán độ đo MAE và được dựa vào best_model_index để chọn. y_hat là các giá trị mục tiêu mới sau khi đã khớp mô hình dữ liệu sử dụng hàm predict ở lớp OLSLinearRegression với tham số X là X_Para. Sau đấy, ta gọi hàm mae để tính độ đo MAE giữa giá trị mục tiêu của tập kiểm tra và giá trị mục tiêu của mô hình.
    - Kết quả:

w	Đặc trưng	MAE
-56546.304	neuroticism	291019.693226953

c. Sử dụng các đặc trưng về khả năng cá nhân như logic, ngoại ngữ và định lượng:

- Ý tưởng: Ta muốn xem xét rằng liệu khả năng cá nhân của mỗi người có ảnh hưởng nhiều tới mức lương của họ hay không. Cách thức làm hoàn toàn giống câu b, ta cũng sẽ sử dụng kỹ thuật k-fold cross validation.
- Kết quả:

Mô hình với 1 đặc trưng	MAE
Quant	118070.330
Logical	120272.797
English	121879.658

- Nhận xét: Có thể thấy rằng, Quant (khả năng định lượng) có kết quả MAE tốt nhất trong 3 khả năng và cả 3 độ đo MAE chênh lệch với nhau không quá lớn và với câu a cũng không quá lớn. Quant (Quantitative) (khả năng định lượng) là khả năng của 1 người để xử lý và giải

quyết dữ liệu "numerical" và "categorical". Logical là khả năng suy luận dựa trên lý lẽ. Chính vì vậy, 2 đặc trưng này được xem là 1 yếu tố cần thiết cho 1 kỹ sư. Tiếp theo, về khả năng ngoại ngữ, ngôn ngữ thường được sử dụng bởi các kỹ sư để có thể trao đổi, giải thích các ý tưởng cho nhau là tiếng Anh. Hơn nữa, hầu hết các bài báo cáo học thuật đều được viết bằng tiếng Anh. Thứ hai, tiếng Anh còn giúp bản thân ta giao tiếp với những người đến từ những đất nước khác, hoặc những khách hàng tiềm năng, những người ít nhất cũng kì vọng rằng bản thân ta biết nói tiếng Anh một cách chuẩn xác và lưu loát. Chính vì vậy, đây chính là ngôn ngữ quan trọng nhất mà mọi kỹ sư đều cần phải biết. Thế nên, việc cả 3 khả năng này đều có độ đo MAE tương đối ổn là điều khá dễ hiểu.

- Mô hình khả năng cá nhân tốt nhất: Quant
  - Cách thức thực hiện tương tự ở đặc trưng neuroticism.
  - Kết quả:

w	Đặc trưng	MAE
585.895	Quant	106819.57761989674

  - Nhận xét: Có thể thấy, trọng số của đặc trưng Quant là số dương, có nghĩa là điểm Quant càng cao sẽ ảnh hưởng tới mức lương của 1 kỹ sư và độ đo MAE của Quant cũng không chênh lệch quá nhiều so với câu a.

d. Xây dựng mô hình riêng:

[1] Tìm mô hình: Thực hiện k-fold cross validation trên 24 mô hình tổng cộng. Với 23 mô hình đầu tiên là mỗi đặc trưng và 1 mô hình cuối cùng là tất cả các đặc trưng.

a. Kết quả:

Mô hình	MAE
Tất cả	110697.934043
Quant	118099.615094
10percentage	118789.005650
12percentage	120015.789205
Logical	120305.273224
collegeGPA	121454.057704
English	121917.981052
CollegeTier	133412.897312
Degree	137512.266873
Gender	150308.742888
ComputerProgramming	155968.072055
Domain	175653.306866
CollegeCityTier	251435.500577
ElectronicsAndSemicon	257435.849861
ComputerScience	268462.421655
TelecomEngg	290911.645603
MechanicalEngg	297357.786953
nueroticism	299328.749851
agreeableness	300763.478723
electricalEngg	301169.971288
openess_to_experience	303058.869685
CivilEngg	306176.407508
conscientiousness	306229.087198
extraversion	306929.502341

b. Nhận xét: Có thể thấy mô hình sử dụng tất cả các đặc trưng có độ đo MAE tốt nhất, khả năng định lượng (Quant) là đặc trưng tốt nhất nếu xét các mô hình chỉ sử dụng 1

đặc trưng và các đặc trưng tính cách đều thuộc vào nhóm dưới, nhóm có các độ đo MAE tệ nhất.