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

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

BÁO CÁO ĐỒ ÁN TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CÔNG NGHỆ THÔNG TIN

Project 3: Linear Regression

Tên sinh viên : Đặng Hà Huy
Lớp : 21CLC05
Mã số sinh viên : 21127296
Giảng viên : Phan Thị Phương Uyên

Thành phố Hồ Chí Minh, tháng 8 năm 2023

MỤC LỤC

GIỚI THIỆU CHUNG

I/ Giới thiệu

- **Họ tên sinh viên:** Đặng Hà Huy
- **MSSV:** 21127296
- **Lớp:** 21CLC05

II/ Thống kê mức độ hoàn thành

STT	Các hàm chức năng	Mức độ hoàn thành	Đánh giá cá nhân
1	Câu 1a	100%	Mô hình ở câu 1a đạt kết quả khá tốt so với các mô hình khác trong đồ án
2	Câu 1b	100%	Các mô hình 1 đặc trưng tính cách có hiệu suất thấp hơn so với các mô hình 1 đặc trưng ở 1c
3	Câu 1c	100%	Các mô hình 1 đặc trưng kĩ năng có hiệu suất khá cao so với các mô hình 1 đặc trưng ở câu 1b
4	Câu 1d	70%	Các mô hình đưa ra vẫn chưa hoàn toàn tốt hơn nhiều so với các mô hình khác trong đồ án này

III/ Giới thiệu về đồ án

1/ Nội dung đồ án

- **Tên đồ án:** Linear Regression (Hồi quy tuyến tính)

- 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.

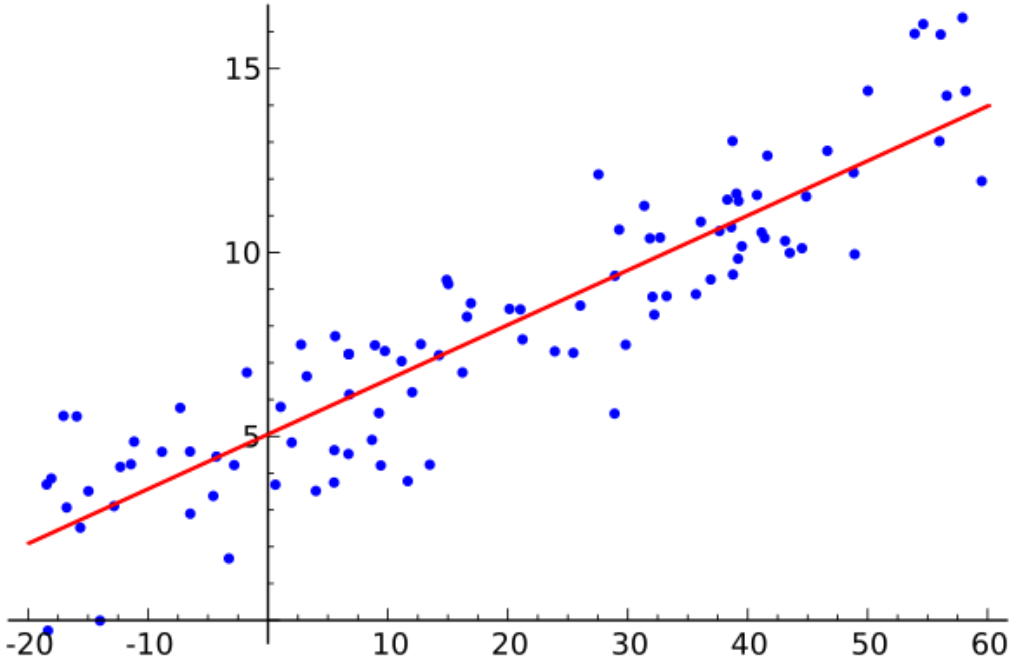
- Bộ dữ liệu được sử dụng trong đồ án này thu thập tại Ấn Độ, nơi có hơn 6000 cơ sở đào tạo kỹ thuật công nghệ với khoảng 2,9 triệu sinh viên đang học tập. Mỗi năm, trung bình có 1,5 triệu sinh viên tốt nghiệp chuyên ngành Công nghệ/Kỹ

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.

- Xem thêm về dữ liệu tại đây[5]

2/ Giới thiệu về hồi quy tuyến tính[1][2]

Linear Regression hay hồi quy tuyến tính là một thuật toán học máy dựa trên Supervised learning (Học có giám sát). Mô hình hồi quy một giá trị mục tiêu dự đoán dựa trên các biến độc lập. Nó chủ yếu được sử dụng để tìm ra mối quan hệ giữa các biến và dùng để dự báo. Các mô hình hồi quy khác nhau khác nhau dựa trên loại mối quan hệ giữa các biến phụ thuộc và biến độc lập mà chúng đang xem xét và số lượng biến độc lập được sử dụng.



Hồi quy tuyến tính thực hiện nhiệm vụ dự đoán giá trị của biến phụ thuộc (y) dựa trên biến độc lập (x) cho trước. Vì vậy, kỹ thuật hồi quy này tìm ra mối quan hệ tuyến tính giữa x (đầu vào) và y (đầu ra).

Trong khi sử dụng hồi quy tuyến tính, mục tiêu của chúng ta là để làm sao một đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó. Nếu chúng ta

có một biến phụ thuộc Y và một biến độc lập X - mối quan hệ giữa X và Y có thể được biểu diễn dưới dạng phương trình sau:

$$y = \theta_1 + \theta_2 x$$

Trong đó:

- y: là biến phụ thuộc
- x: là biến độc lập
- θ_1: là hằng số
- θ_2: là hệ số quan hệ giữa x và y

Tính chất:

- Đường hồi quy luôn luôn đi qua trung bình của biến độc lập (x) cũng như trung bình của biến phụ thuộc (y)

- Đường hồi quy tối thiểu hóa tổng của "Diện tích các sai số". Đó là lý do tại sao phương pháp hồi quy tuyến tính được gọi là "Ordinary Least Square (OLS)"

- θ_2 giải thích sự thay đổi trong y với sự thay đổi X bằng một đơn vị. Nói cách khác, nếu chúng ta tăng giá trị của X bởi một đơn vị thì nó sẽ là sự thay đổi giá trị của y

Hiệu suất mô hình:

- Mean Absolute Error (MAE) hay sau số trung bình tuyệt đối là một phép đo khác để đánh giá sự sai lệch giữa các giá trị dự đoán và giá trị thực tế trong các bài toán dự đoán. MAE đo độ lớn trung bình của sai số tuyệt đối giữa các dự đoán và giá trị thực tế. Dưới đây là công thức của MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{predict_i}|$$

Trong đó:

- n: là số lượng điểm dữ liệu trong tập dữ liệu mà bạn muốn đánh giá
- y: là giá trị thực tế tương ứng với điểm dữ liệu thứ i
- y_predict: là giá trị dự đoán tương ứng với điểm dữ liệu thứ i

- Công thức tính MAE tính trung bình cộng của các giá trị tuyệt đối của sai số. Điều này giúp đánh giá sự sai lệch trung bình giữa các dự đoán và giá trị thực tế.

3/ K-fold cross valiadation[3][4]

- Cross validation là một phương pháp thống kê được sử dụng để ước lượng hiệu quả của các mô hình học máy. Thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán

- Cross validation là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp dữ liệu không được dồi dào cho lắm.

- 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.

- 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
- Ví dụ, ta có một dataset S = $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ gồm 6 mẫu và ta muốn thực hiện 3-cross-validation
  + Đầu tiên chúng ta chia S ra làm 3 subset riêng
  - $S_1 = \{x_1, x_2\}$
  - $S_2 = \{x_3, x_4\}$
  - $S_3 = \{x_5, x_6\}$
- Sau đó, chúng ta huấn luyện và đánh giá mô hình 3 lần. Mỗi lần, hai tập con tạo thành tập huấn luyện, trong khi tập còn lại đóng vai trò là tập kiểm tra. Trong ví dụ này:

- Cuối cùng, ta tính số điểm trung bình của mô hình dựa trên 3 số điểm mà chúng ta vừa tìm được

$$overall\ score = \frac{score_1 + score_2 + score_3}{3}$$

NỘI DUNG CHÍNH

I/ Các hàm thư viện

- pandas: đọc các file dữ liệu .csv và cung cấp các cấu trúc để xử lý dữ liệu

- numpy: sử dụng các hàm tính toán cơ bản và tạo ma trận

- seaborn: sử dụng để tạo ra các hình ảnh trực quan, heatmap

- matplotlib: sử dụng các hàm hỗ trợ vẽ đồ thị

II/ Mô tả các hàm hỗ trợ

1/ Lớp hồi quy tuyến tính (OLSLinearRegression)

a/ Hàm đào tạo mô hình (fit)

- Ý tưởng: Thực hiện quá trình đào tạo mô hình hồi quy tuyến tính với đầu vào **X** dạng array hoặc dataframe và **y** dạng array hoặc dataframe

- Tên hàm: fit(**self**, **X** : np.array hoặc dataframe, **y** : np.array hoặc dataframe)

- Đầu vào:

+ **X** là ma trận chứa các đặc trưng đầu vào của dữ liệu huấn luyện Mỗi hàng đại diện cho một điểm dữ liệu và mỗi cột ứng với một đặc trưng. Kích thước của ma trận là (số lượng điểm dữ liệu) x (số lượng đặc trưng).

+ **y** là vector chứa các giá trị mục tiêu ứng với các điểm dữ liệu trong ma trận **X**. Kích thước của vector y phải phù hợp với số lượng điểm dữ liệu trong **X**.

- Đầu ra: Trả về biến self kiểu dữ liệu tương tự với lớp

- Mô tả thuật toán[6]:

+ Bước 1: Tính ma trận nghịch đảo giả (nghịch đảo Moore-Penrose) sử dụng hàm nghịch đảo ma trận **np.linalg.pinv()** trong thư viện numpy

$$X\_pinv = (X^T X)^{-1} X^T$$

+ Bước 2: Tính vector trọng số bằng cách nhân ma trận nghịch đảo với giá trị mục tiêu **y**

$$self.w = X\_pinv * y$$

b/ Hàm trả về vector trọng số (get_params)

- Ý tưởng: Trả về trọng số sau khi đào tạo mô hình

- Tên hàm: get_params()

- Đầu ra: Là vector trọng số **self.w**

- Mô tả thuật toán: Hàm trả về trọng số **w** sau khi đã đào tạo mô hình

c/ Hàm dự đoán giá trị mục tiêu (predict)

- Ý tưởng: Dự đoán các giá trị mục tiêu dựa vào các đặc trưng đầu vào

- Tên hàm: predict(**self**, **X** : np.array hoặc dataframe)

- Đầu vào: **X** là ma trận chứa các đặc trưng đầu vào của dữ liệu muốn dự đoán

- Đầu ra: một mảng chứa các giá trị dự đoán dựa trên các đặc trưng đầu vào

- Mô tả thuật toán: Nhận vào một tập các đặc trưng mới (**X**) và dự đoán các giá trị mục tiêu dựa vào công thức:

$$y\_predict = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Với mỗi phần tử trong mảng đầu ra là tổng của tích vô hướng giữa mảng trọng số đã làm phẳng **self.w** với hàm **ravel()** và dòng tương ứng trong **X**

2/ Hàm tính sai số trung bình tuyệt đối (MAE)

- Ý tưởng: Tính sai số trung bình tuyệt đối của giá trị mục tiêu (**y**) và giá trị dự đoán (**y_hat**)

- Tên hàm: MAE(**y** : np.array hoặc dataframe , **y_hat** : np.array hoặc dataframe)

- Đầu vào:

+ **y** là vector chứa các giá trị mục tiêu

+ **y_hat** là vector chứa các giá trị dự đoán tương ứng (sử dụng hàm predict của lớp OLSLinearRegression để dự đoán)

- Đầu ra: Giá trị MAE (Mean Absolute Error) hay sai số trung bình tuyệt đối giữa các giá trị mục tiêu (**y**) và các giá trị dự đoán tương ứng với y (**y_hat**)

- Mô tả thuật toán:

+ Bước 1: Hàm nhận vào 2 đối số là **y** chứa các giá trị mục tiêu và **y_hat** chứa các giá trị dự đoán

+ Bước 2: Sau đó hàm thực hiện tính toán giá trị MAE theo công thức:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{predict_i}|$$

3/ Hàm xáo trộn dữ liệu (shuffle)

- Ý tưởng: Xáo trộn dữ liệu đầu vào một cách ngẫu nhiên với thư viện numpy

- Tên hàm: shuffle(**data** : np.array hoặc dataframe)

- Đầu vào: **data** một dataset dạng array hoặc dataframe cần được xáo trộn

- Đầu ra: **data_shuffled** một dataset dạng array hoặc dataframe đã được xáo trộn ngẫu nhiên

- Mô tả thuật toán: Sử dụng hàm shuffle của numpy để xáo trộn dữ liệu một cách ngẫu nhiên

4/ Hàm kiểm chứng chéo (cross_validation)

- Ý tưởng: Chia dataset đầu vào thành k tập khác nhau và lần lượt lựa chọn từng tập làm tập thử nghiệm và các tập còn lại sẽ kết hợp lại để làm tập huấn luyện

- Tên hàm: cross_validate(**dataset** : np.array hoặc dataframe, **k** : int)

- Đầu vào:

+ **dataset** là một ma trận dữ liệu với cột cuối cùng là giá trị mục tiêu (**y**) còn các cột còn lại là các đặc trưng (**X**)

+ **k** là số lượng fold cho quá trình kiểm chứng chéo (mặc định là 5)

- Đầu ra: **maes_avg** là một mảng giá trị MAE trung bình ứng với mỗi đặc trưng

- Mô tả thuật toán:

+ Bước 1: Chia ma trận dữ liệu **dataset** thành **k** fold

+ Bước 2: Chọn 1 fold để làm tập kiểm tra, tất cả các fold còn lại kết hợp với nhau để làm tập huấn luyện, với mỗi vòng lặp thì đổi sang 1 fold khác để làm tập thử nghiệm

+ Bước 3: Huấn luyện mô hình hồi quy tuyến tính trên tập huấn luyện và dự đoán giá trị trên mô hình đó

+ Bước 4: Tính toán giá trị MAE giữa giá trị thực tế và giá trị dự đoán trên tập thử nghiệm, lưu vào danh sách và cộng các giá trị đó lại với nhau

+ Bước 5: Trả về giá trị MAE đã chia cho **k** để có được giá trị MAE trung bình cho từng đặc trưng trong tập dữ liệu

5/ Hàm lọc đặc trưng dựa trên ngưỡng tương quan giữa các đặc trưng (correlation_filter)[7]

- Ý tưởng: Lọc bỏ các đặc trưng có độ tương quan cao với nhau dựa trên một ngưỡng bất kỳ mà người dùng nhập vào (Khi 2 đặc trưng có độ tương quan cao với nhau thì ta có thể bỏ 1 trong 2 bởi vì chúng sẽ không có ảnh hưởng quá lớn đối với mô hình)

- Tên hàm: correlation_filter(**dataset** : np.array hoặc dataframe, **threshold** : float)

- Đầu vào:

+ **dataset** là một ma trận dữ liệu với cột cuối cùng là giá trị mục tiêu (**y**) còn các cột còn lại là các đặc trưng (**X**)

+ **threshold** là ngưỡng tương quan giữa các đặc trưng. Các đặc trưng có tương quan cao hơn ngưỡng này sẽ bị loại bỏ

- Đầu ra: **filter_features** là một mảng chứa tên các đặc trưng bị loại bỏ dựa trên ngưỡng tương quan đầu vào

- Mô tả thuật toán:

+ Bước 1: Tính toán các ma trận tương quan của tất cả các thuộc tính bên trong **dataset.** Ma trận tương quan là một bảng chứa tất cả các giá trị tương quan giữa từng cặp thuộc tính trong dữ liệu

+ Bước 2: Duyệt qua từng dòng và cột trong ma trận tương quan và kiểm tra giá trị tương quan ở dòng và cột đó. Nếu giá trị tương quan tuyệt đối vượt quá ngưỡng tương qua đầu vào (**threshold**) nên sẽ thêm tên đặc trưng ở cột đó vào danh sách **filter_features**

+ Bước 3: Sau khi duyệt qua hết tất cả các dòng và cột, hàm sẽ trả về danh sách các đặc trưng đã vượt quá ngưỡng đầu vào để có thể lọc bỏ nếu cần

III/ Mô tả nội dung bài làm

1/ Các bước tiền xử lý

- Cài đặt các hàm hỗ trợ như trên

- Đọc dữ liệu dữ liệu từ các file cho sẵn "**train.csv**" và "**test.csv**" và tách các đặc trưng và giá trị mục tiêu

2/ Câu 1a

- Yêu cầu: Sử dụng toàn bộ 11 đặc trưng đầu tiên Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain

- Các bước thực hiện:

+ Bước 1: Huấn luyện mô hình hồi quy tuyến tính dựa trên 2 tập đặc trưng và giá trị mục tiêu của 11 đặc trưng đầu tiên

+ Bước 2: Kết quả hệ số của mô hình trả về:

STT	Đặc trưng	Hệ số
1	Gender	-22756.512821
2	10percentage	804.503156
3	12percentage	1294.654565
4	CollegeTier	-91781.897531
5	Degree	23182.388679
6	collegeGPA	1437.548672
7	CollegeCityTier	-8570.661985
8	English	147.858299
9	Logical	152.888476
10	Quant	117.221846
11	Domain	34552.286221

- Thể hiện công thức hồi quy dựa trên 11 đặc trưng đầu tiên là:

$$\begin{aligned} Salary \;=\; & -22756.513 * \boldsymbol{Gender} + 804.503 * \boldsymbol{10percentage} \\ & +1294.655 * \boldsymbol{12percentage} - 91781.898 * \boldsymbol{CollegeTier} \\ & +23182.389 * \boldsymbol{Degree} + 1437.549 * \boldsymbol{collegeGPA} \\ & -8570.662 * \boldsymbol{CollegeCityTier} + 147.858 * \boldsymbol{English} \\ & +152.888 * \boldsymbol{Logical} + 117.222 * \boldsymbol{Quant} + 34552.286 * \boldsymbol{Domain} \end{aligned}$$

+ Bước 3: Dùng mô hình để dự đoán giá trị trên tập thử nghiệm. Kết quả giá trị MAE của mô hình với 11 đăc trưng

MAE = 104863.77754033303

- Nhận xét: Mô hình với 11 đặc trưng đầu đạt được kết quả khá tốt so với các mô hình khác trong đồ án. Nhưng khi áp dụng vào thực tế thì sai này vẫn quá lớn và cần được cải thiện

3/ Câu 1b

- Yêu cầu: Phân tích ảnh hưởng của đặc trưng tính cách dựa trên điểm các bài kiểm tra của AMCAT. Thử nghiệm lần lượt trên các đặc trưng tính cách gồm: conscientiousness, agreeableness, extraversion, nueroticism, openess to experience

- Giả thuyết:

+ Cả 5 đặc trưng trên "conscientiousness", "agreeableness", "extraversion", "nueroticism" và "openess_to_experience" đều đóng một vai trò nhất định trong việc chọn ra mức lương phù hợp tuỳ vào tính cách của mỗi người.

+ Theo các bài báo này[8][9], các tính cách như "conscientiousness" và "extraversion" có xu hướng tăng mức lương lên lần lượt là 3-5% và 5-6%. Bên cạnh đó tính cách như "neuroticism" lại có xu hướng giảm mức lương đi 5-9%. Tính cách như "agreeableness" lại ảnh hưởng xấu đến mức lương trong khi đó "extraversion" lại có ảnh hưởng tích cực đến mức lương

+ Chính vì thế, dựa vào các bài báo trên ta có thể kết luận rằng "nueroticism" là tính cách có thể ảnh hưởng đến mức lương cáo theo hướng tiêu cực, làm giảm mức lương hằng năm

- Các bước thực hiện:

+ Bước 1: Xáo trộn tập dữ liệu và thực hiện kiểm chứng chéo (cross valiadation) với số lượng k-fold = 5

+ Bước 2: Sau khi kiểm chứng chéo, hàmsẽ trả về một mảng các giá trị MAE trung bình của từng mô hình. Để tìm được đặc trưng tốt nhất ta cần tìm đặc trưng với giá trị MAE trung bình nhỏ nhất

- Các giá trị MAE trung bình trên tập huấn luyện sau mỗi lần chạy sẽ khác nhau do xáo trộn dữ liệu trước khi kiểm chứng chéo nhưng vẫn luôn giữ đúng thứ tự về độ lớn giữa mỗi đặc trưng

STT	Mô hình với 1 đặc trưng	MAE
1	conscientiousness	306098.118860
2	agreeableness	300676.303963
3	extraversion	306907.749873
4	nueroticism	299256.778298
5	openess_to_experience	302943.270945

- Ta có thể nhận thấy rằng đặc trưng "**nueroticism**" có giá trị MAE nhỏ nhất là **299256.778298**

+ Bước 3: Sau đó ta huấn luyện lại mô hình trên tập thử nghiệm với kết quả tham số là **-56546.303753.** Công thức của mô hình:

$$Salary = -56546.304 * \boldsymbol{nueroticism}$$

- Giá trị MAE của mô hình là:

MAE = 291019.693226953

- Nhận xét: Mô hình với 1 đặc trưng tốt nhất là tính cách không đạt được kết quả tốt bởi vì với mỗi đặc trưng thì giá trị MAE trên tập thử nghiệm đều khá cao so với tất cả các mô hình trong đồ án này

4/ Câu 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. Thử nghiệm trên các đặc trưng gồm: English, Logical, Quant

- Giả thuyết: Ba kĩ năng trên "English", "Logical" và "Quant" đều cần thiết trong thời đại hiện nay để có thể kiếm được việc làm với mức lương cao.

+ Với kĩ năng ngoại ngữ "English"[10], mức lương dự kiến có thể cao hơn đến 18% hay thậm chí là 30% đối với những người thuần thục với ngoại ngữ

+ Với kĩ năng "Logical" và "Quant" cũng có ảnh hưởng không kém đối với mức lương bởi vì đây đều là các kĩ năng cần thiết mà một người lao động chất lượng cao cần nên có

- Chỉ với việc nhìn qua 3 kĩ năng trên ta cũng khó có thể đoán được rằng kĩ năng nào có thể có ảnh hưởng lớn nhất đối với mức lương. Chính vì thế ta cần phải thử nghiệm thực tế cũng như căn cứ vào các con số để có thể đưa ra kết luận chính xác

- Các bước thực hiện:

+ Bước 1: Xáo trộn tập dữ liệu vừa thu thập được và thực hiện kiểm chứng chéo (cross valiadation) với số lượng k-fold = 5

+ Bước 2: Sau khi kiểm chứng chéo, hàmsẽ trả về một mảng các giá trị MAE trung bình của từng mô hình. Để tìm được đặc trưng tốt nhất ta cần tìm đặc trưng với giá trị MAE trung bình nhỏ nhất

- Các giá trị MAE trung bình trên tập huấn luyện sau mỗi lần chạy sẽ khác nhau do xáo trộn dữ liệu trước khi kiểm chứng chéo nhưng vẫn luôn giữ đúng thứ tự về độ lớn giữa mỗi đặc trưng

STT	Mô hình với 1 đặc trưng	MAE
1	English	121884.613321
2	Logical	120276.695391
3	Quant	118137.055407

- Ta có thể nhận thấy rằng đặc trưng "**Quant**" có giá trị MAE nhỏ nhất là **118137.055407**

+ Bước 3: Sau đó ta huấn luyện lại mô hình trên tập thử nghiệm với kết quả tham số là **585.895381.** Công thức của mô hình:

$$Salary = -585.895 * \boldsymbol{Quant}$$

- Giá trị MAE của mô hình 1 đặc trưng kĩ năng tốt nhất là:

MAE = 106819.57761989674

- Nhận xét: Trái với mô hình với 1 đặc trưng tính cách tốt nhất thì mô hình với 1 đặc trưng kĩ năng tốt nhất lại đạt được kết quả khả quan hơn. Tuy vậy mô hình này vẫn chưa phải là mô hình tốt nhất kể cả trong đồ án này lẫn thực tiễn và cần cải thiện thêm

5/ Câu 1d

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

- Mô hình 1: Sử dụng đặc trưng có giá trị MAE nhỏ hơn 200000

- Mô hình 1 gồm các đặc trưng: '**Quant**', '**12percentage**', '**10percentage**', '**collegeGPA**', '**Logical**', '**English**', '**CollegeTier**', '**Degree**', '**Gender**', '**ComputerProgramming**', '**Domain**'

- Các bước thực hiện:

+ Bước 1: Huấn luyện mô hình dựa trên tập huấn luyện của từng đặc trưng, dự đoán một tập giá trị mới và tính giá trị MAE

+ Bước 2: Sau khi tính giá trị MAE, thêm chúng vào một danh sách, sắp xếp lại theo thứ tự từ bé đến lớn và bỏ đi những đặc trưng có giá trị MAE > 200000

- Bảng các đặc trưng có giá trị MAE < 200000

STT	Đặc trưng	MAE
1	Quant	106819.577620
2	12percentage	111427.175482
3	10percentage	111941.803536
4	CollegeGPA	114600.224618
5	Logical	115082.021387
6	English	117213.928965
7	CollegeTier	126802.148183
8	Degree	132669.852864
9	Gender	143583.739602
10	ComputerProgramming	153100.424596
11	Domain	165468.188876

+ Bước 3: Khi đã tìm được những đặc trưng cho mô hình, ta tiến hành huấn luyện chúng và thu được các hệ số của từng đặc trưng:

STT	Đặc trưng	Hệ số
1	Quant	118.216153
2	12percentage	1182.287207
3	10percentage	895.933118
4	CollegeGPA	1333.176343
5	Logical	139.036379
6	English	139.311223
7	CollegeTier	-89585.622331
8	Degree	17267.986790
9	Gender	-23440.703979
10	ComputerProgramming	68.023963
11	Domain	25926.671077

- Công thức hồi quy của mô hình 1 là:

$$Salary = 118.216 * \boldsymbol{Quant} + 1182.287 * \boldsymbol{12percentage} + 895.933 * \boldsymbol{10percentage} + 1333.176 * \boldsymbol{collegeGPA} + 139.036 * \boldsymbol{Logical} + 139.311 * \boldsymbol{English} - 89585.622 * \boldsymbol{CollegeTier} + 17267.987 * \boldsymbol{Degree} - 23440.704 * \boldsymbol{Gender} + 68.024 * \boldsymbol{ComputerProgramming} + 25926.671 * \boldsymbol{Domain}$$

- <u>Mô hình 2:</u> Sử dụng đặc trưng có độ tương quan với Lương lớn hơn 0.05

- Mô hình 2 gồm các đặc trưng: '**Quant**', '**Logical**', '**CollegeTier**', '**English**', '**10percentage**', '**12percentage**', '**ComputerProgramming**', '**collegeGPA**', '**Domain**', '**ComputerScience**', '**nueroticism**', '**agreeableness**', '**conscientiousness**'

- <u>Các bước thực hiện:</u>

+ Bước 1: Chia tập huấn luyện thành 2 tập đặc trưng và giá trị mục tiêu . Tính giá trị tuyệt đối giữa độ tương quan giữa các đặc trưng với giá trị mục tiêu

+ Bước 2: Duyệt qua từng giá trị tương quan giữa các đặc trưng và bỏ đi các đặc trưng có độ tương quan với lương cao hơn 0.05

+ Bước 3: Huấn luyện lại mô hình với các đặc trưng còn lại và cho ra kết quả các hệ số của từng đặc trưng như sau:

STT	Đặc trưng	Hệ số
1	Quant	110.103254
2	Logical	128.359541
3	CollegeTier	-81441.646930
4	English	150.449833
5	10percentage	705.996554
6	12percentage	951.608458
7	ComputerProgramming	102.502612
8	CollegeGPA	1479.717445
9	Domain	21946.267633
10	ComputerScience	-161.448398
11	nueroticism	-10623.850936
12	agreeableness	14638.231791
13	conscientiousness	-21020.657245

- Công thức hồi quy của mô hình 2 là:

$$\begin{aligned} Salary \;=\; & 110.103 * \boldsymbol{Quant} + 128.360 * \boldsymbol{Logical} \\ & -81441.647 * \boldsymbol{CollegeTier} + 150.450 * \boldsymbol{English} \\ & +705.997 * \boldsymbol{10percentage} + 951.608 * \boldsymbol{12percentage} \\ & +102.503 * \boldsymbol{ComputerProgramming} + 1479.717 * \boldsymbol{CollegeGPA} \\ & +21946.268 * \boldsymbol{Domain} - 161.448 * \boldsymbol{ComputerScience} \\ & -10623.851 * \boldsymbol{nueroticism} + 14638.231791 * \boldsymbol{agreeableness} \\ & -21020.657 * \boldsymbol{conscientiousness} \end{aligned}$$

- <u>Mô hình 3:</u> Sử dụng đặc trưng có độ tương quan với nhau dưới 0.25

- Mô hình 3 gồm các đặc trưng: '**Gender**', '**10percentage**', '**CollegeTier**', '**CollegeCityTier**', '**Domain**', '**ElectricalEngg**', '**CivilEngg**', '**conscientiousness**'

- <u>Các bước thực hiện:</u>

+ Bước 1: Sử dụng hàm correlation_filter() để chọn ra các đặc trưng có độ tương quan với nhau vượt quá 0.25 và loại bỏ chúng ra khỏi mô hình

+ Bước 2: Dùng các đặc trưng còn lại để huấn luyện mô hình và cho ra kết quả các hệ số của từng đặc trưng như sau:

STT	Đặc trưng	Hệ số
1	Gender	-20235.195558
2	10percentage	5195.985168
3	CollegeTier	-49844.149673
4	CollegeCityTier	-8208.577458
5	Domain	53210.067457
6	ElectricalEngg	-153.711750
7	CivilEngg	70.228367
8	conscientiousness	-12315.204220

- Công thức hồi quy của mô hình 3 là:

$$\begin{aligned} Salary \;=\; & -20235.196 * \boldsymbol{Gender} + 5195.985 * \boldsymbol{10percentage} \\ & -49844.150 * \boldsymbol{CollegeTier} - 8208.577 * \boldsymbol{CollegeCityTier} \\ & +53210.067 * \boldsymbol{Domain} - 153.712 * \boldsymbol{ElectricalEngg} \\ & +70.228 * \boldsymbol{CivilEngg} - 12315.204 * \boldsymbol{conscientiousness} \end{aligned}$$

- **<u>Đánh giá cả 3 mô hình:</u>**

- Giá trị MAE sau mỗi lần chạy sẽ khác nhau do xáo trộn dữ liệu nhưng giá trị MAE của mô hình tốt nhất sẽ luôn luôn thấp nhất

Mô hình	MAE
Mô hình 1	113190.439784
Mô hình 2	111280.711070
Mô hình 3	116609.272082

- Từ bảng giá trị MAE phía trên ta nhận thấy rằng mô hình thứ 2 có giá trị MAE thấp nhất gồm các đặc trưng '**Quant**', '**Logical**', '**CollegeTier**', '**English**', '**10percentage**', '**12percentage**', '**ComputerProgramming**', '**collegeGPA**', '**Domain**', '**ComputerScience**', '**nueroticism**', '**agreeableness**', '**conscientiousness**'

- Giá trị MAE của mô hình tốt nhất là

MAE = 102366.6017832477

- **Giả thuyết:**

- Mô hình tốt nhất trong số 3 mô hình trên là mô hình số 2. Được tìm thấy bằng cách tìm ra các đặc trưng có tương quan đến mức lương cao nhất.

- Điều này có thể là do mức lương chỉ cần được tính những đặc trưng ảnh hưởng lớn hoặc có tầm quan trọng cao và thật sự cần thiết để có thể thăng tiến dễ dàng

- Nhận xét:

- Với 3 mô hình trên, giá trị MAE của mỗi mô hình trên tập thử nghiệm lần lượt là:
  + Model 1: MAE = 104441.2976452526
  + Model 2: MAE = 102366.6017832477
  + Model 3: MAE = 107927.44224926866

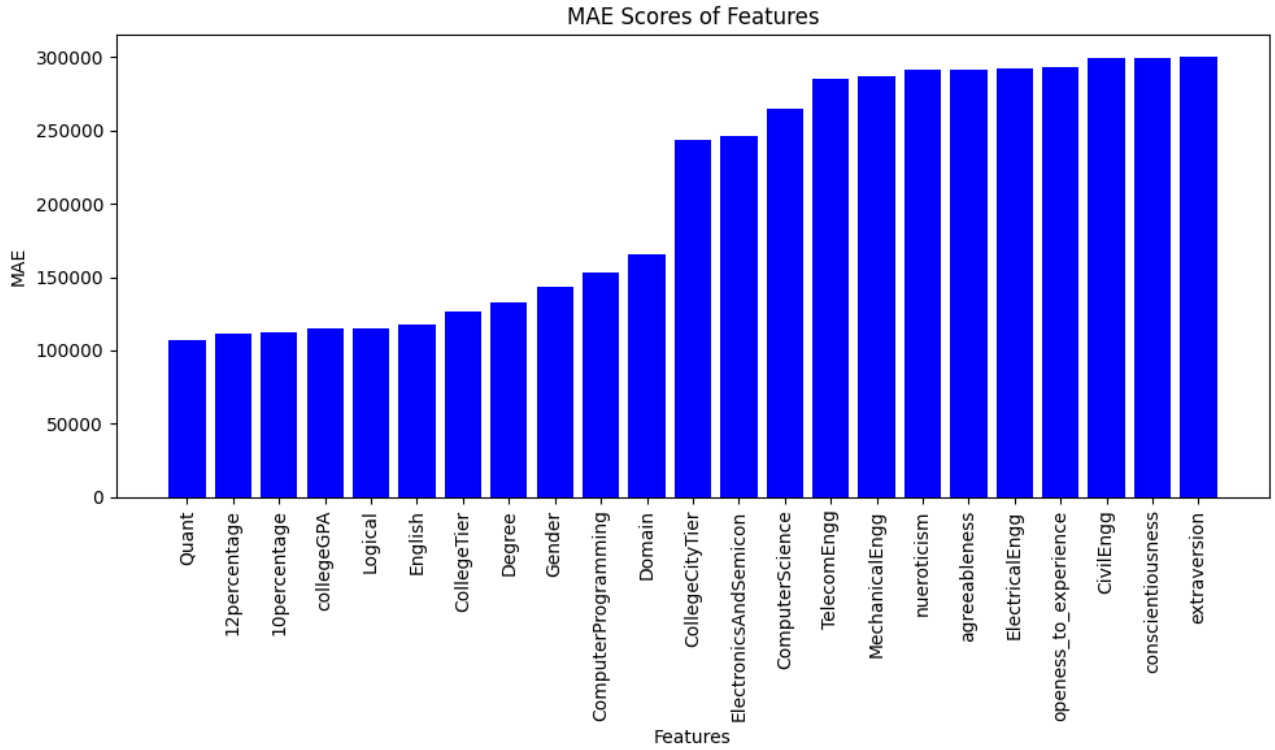
- Có thể nhận thấy cả 3 mô hình có giá trị MAE khá thấp so với các mô hình của đề bài. Ngoại trừ mô hình 3 có giá trị MAE cao hơn mô hình ở câu 1a thì tất cả các mô hình còn lại đều thấp hơn so với các mô hình đề bài.

- Trong đó mô hình 2 có giá trị MAE thấp nhất được tìm thấy bằng phương pháp tìm đặc trưng với độ tương quan với Lương lớn nhất có thể bởi các mô hình đó có ảnh hưởng đến lương cao nhất.

6/ Quá trình xây dựng m mô hình

a/ Mô hình 1

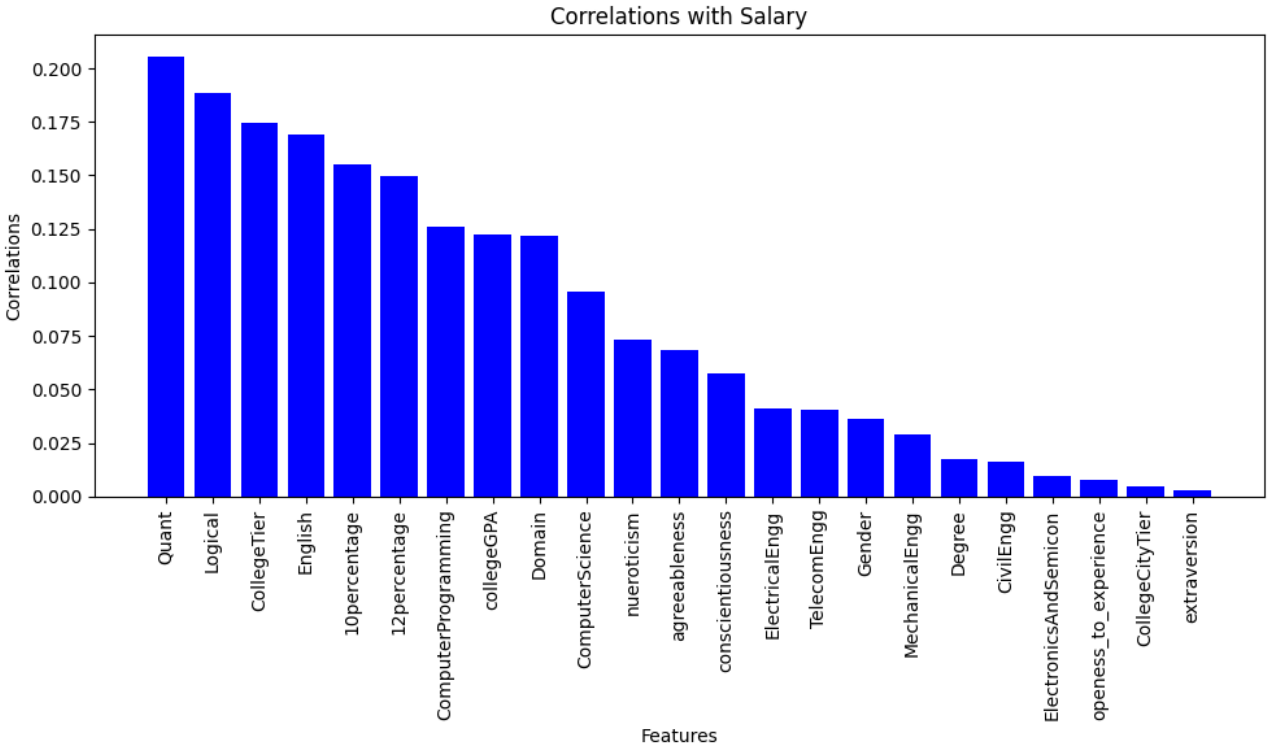
- Đối với mô hình 1, từ các câu 1b và 1c trong đồ án, ta có thể suy ra rằng dựa vào giá trị MAE ta có thể tìm kiếm được các đặc trưng tốt nhất cho mô hình
- Sau khi tính toán giá trị MAE và sắp xếp theo thứ tự ta được biểu đồ sau



- Có thể thấy rằng có đặc trưng **'Quant', '12percentage', '10percentage', 'collegeGPA', 'Logical', 'English', 'CollegeTier', 'Degree', 'Gender', 'ComputerProgramming', 'Domain'** có giá trị MAE thấp hơn hẳn so với các đặc trưng còn lại

b/ Mô hình 2

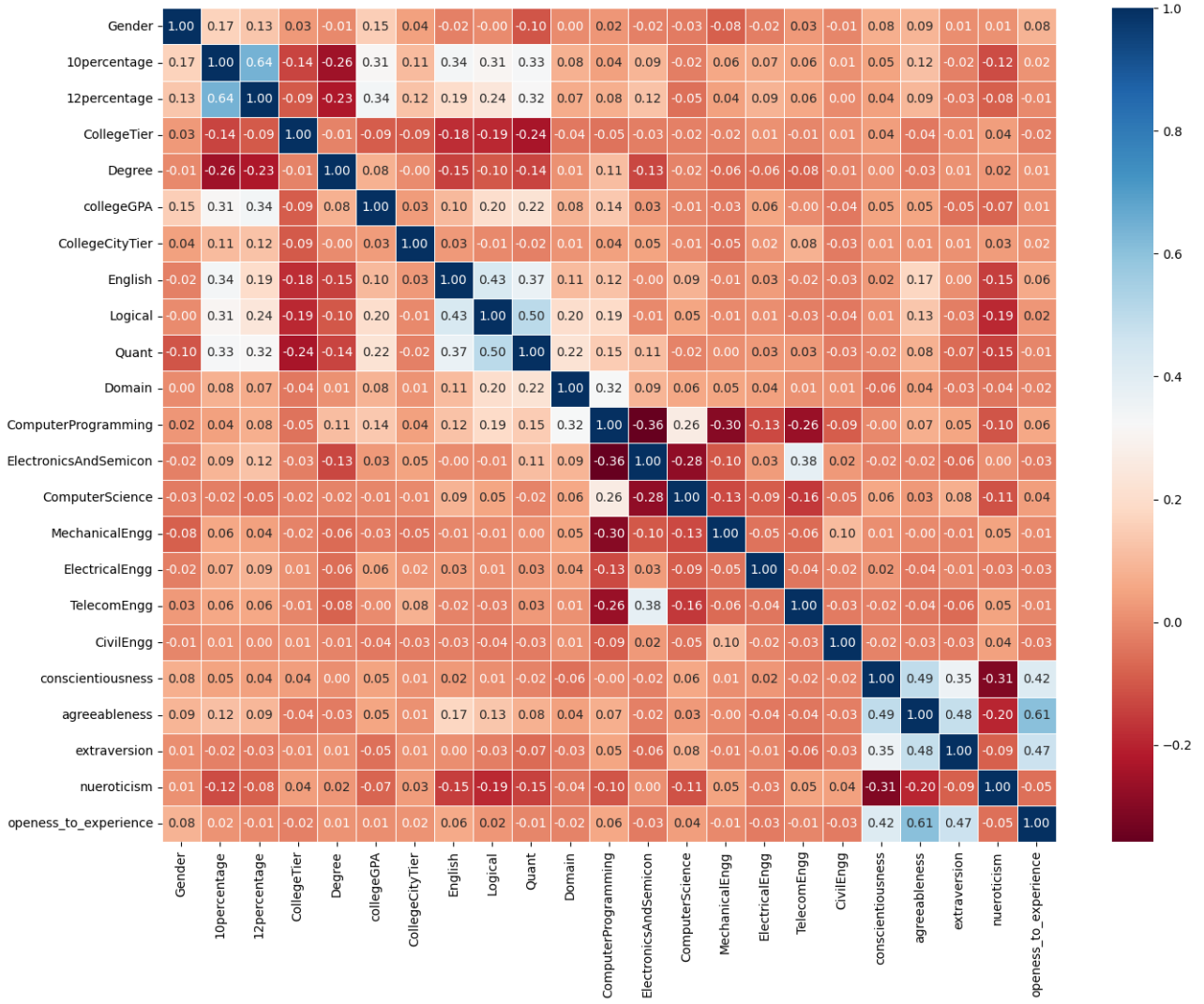
- Đối với mô hình 2, ta sử dụng một các tính mới là độ tương quan “correlation” sẽ tính được xem đặc trưng đó ảnh hưởng đến giá trị mục tiêu nhiều hay ít
- Những đặc trưng nào có độ tương quan với giá trị mục tiêu càng lớn thì ta sẽ chọn ra chúng để có thể tính toán giá trị mục tiêu một các chính xác hơn
- Độ tương quan của các đặc trưng đối với mức lương sau khi tính toán và sắp xếp ta sẽ được biểu đồ sau



- Với mô hình 2 này, các đặc trưng có độ tương quan cao (ở mức >0.05) sẽ được thêm vào mô hình. Việc chọn mức 0.05 là để ta có thể có được nhiều đặc trưng hơn để có thể tính toán chính xác hơn

c/ Mô hình 3

- Ở mô hình 3, ta loại bỏ bắt các mô hình có độ tương quan cao với nhau bởi vì khi 2 đặc trưng có độ tương quan cao với, chúng gần như tương đồng và có thể bỏ đi 1 trong 2. Ta có biểu đồ heatmap sau:



- Căn cứ vào biểu đồ heatmap, ta có thể thấy các đặc trưng có độ tương quan với nhau cao như **"12percentage"** với **"10percentage"** với độ tương quan 0.64, **"Quant"** và **"English"** với độ tương quan 0.43, **"Quant"** và **"Logical"** với độ tương quan 0.5,... Ta chọn ngưỡng tương quan là 0.25 để có thể loại bỏ bớt các đặc trưng có tương quan với nhau quá 0.25 và giữ lại các đặc trưng khác

IV/ Tổng kết

- Qua đồ án này, ta đã học được cách để có thể dựa vào các đặc trưng như tính cách, kĩ năng cá nhân, chuyên ngành mà người đó học để có thể dự đoán được mức lương hợp lý cho từng cá nhân.
- Tuy vậy nhưng những mô hình trong đồ án trong đồ án này vẫn có sai số quá cao dẫn đến không phù hợp để có thể sử dụng trong thực tế bởi lẽ với giá trị MAE vượt

quá 100000 dẫn đến việc lương có thể chênh lệch với thực tế đến xấp xỉ 29 triệu VND (dữ liệu vào tháng 8/2023 với 1 đồng rupee bằng với 286.92 VND)
- Để có thể xây dựng một mô hình dự đoán mức lương hoàn hảo là gần như không thể bởi vì mức lương bị ảnh hưởng bởi rất nhiều yếu tố không thể kiểm soát được. Chính vì thế việc tạo mô hình tốt nhất là việc khiến cho mô hình đó có độ lệch nhỏ nhất có thể

TÀI LIỆU THAM KHẢO

[1] Linear Regression - Hồi quy tuyến tính trong Machine Learning - Viblo

[2] Linear Regression in Machine Learning - GeeksforGeeks

[3] Giới thiệu về k-fold cross-validation - trituenhantao

[4] Cross-Validation: K-Fold vs. Leave-One-Out - baeldung

[5] Engineering Graduate Salary Prediction – kaggle

[6] Data fitting và phương pháp OLS - phitran

[7] Feature Selection Correlation - Github

[8] The Personality Traits That Increase Your Salary - LinkedIn

[9] Here's How Your Personality Type May Affect Your Income - fool

[10] Improve your salary with English fluency - LinkedIn