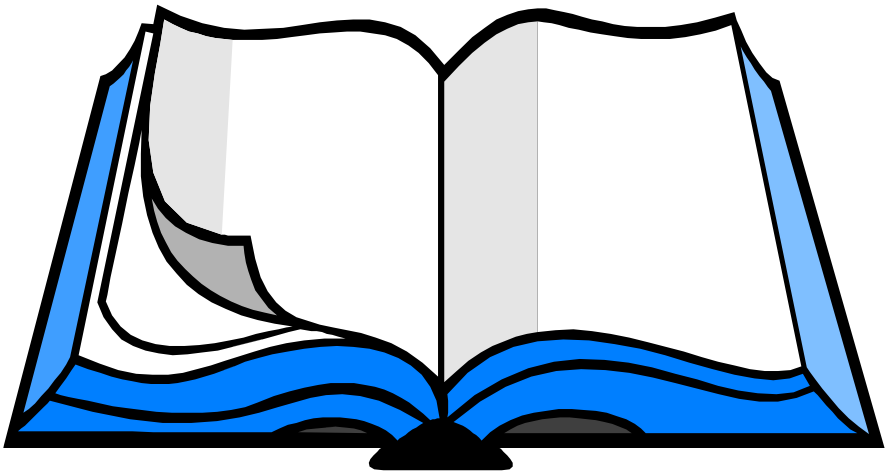
University of Science

Information Technology Department

PROJECT 03: OLS LINEAR REGRESSION



Giảng viên:

- Nguyễn Văn Quang Huy
- Ngô Đình Hy
- Nguyễn Đình Thúc

Sinh viên: 21127667 – Trương Công Gia Phát

MỤC LỤC

I. Thư viện sử dụng

pandas: sử dụng để có thể đọc dữ liệu và thao tác trên dataframe.

numpy: sử dụng thư viện để có thể chuyển hoá dataframe dưới dạng ma trận và tính toán.

seaborn: sử dụng seaborn để có thể trực quan hoá dữ liệu bằng heatmap.

LinearRegression của sklear.linear_model: sử dụng thư viên để có thể tính toán hệ số góc (coef_) và hệ số tự do(intercept_) của phương trình hồi quy tuyến tính.

mean_absolute_error của sklearn.metrics: sử dụng để có thể tính toán MAE.

II. Hàm sử dụng

Em cài đặt toàn bộ hàm từ class **OLSLinearRegression()** của lab04.

fit(self, X, y) dùng để huấn luyện tập X với kết quả y để có thể ra phương trình hồi quy.

get_params() dùng để lấy ra các hệ số của phương trình hồi quy.

predict(self, X) nhận vào một tập dữ liệu X sẽ trả về kết quả y_hat dựa trên phương trình hồi quy đã được huấn luyện.

mae(y, y_hat) dùng để tính sai số mae giữa tập kết quả y và y_hat là kết quả trả về của predict().

select_attribute (threshold, corr_matrix) nhận vào một threshold và độ tương quan giữa các thuộc tính với một thuộc tính nào đó. Hàm sẽ trả về danh sách các thuộc tính mà trị tuyệt đối giá trị tương quan của nó lớn hơn threshold.

III. Báo cáo kết quả

1a.

Công thức hồi quy:

Salary = -22756.513*X1 + 804.503*X2 + 1294.654*X3 - 91781.897*X4 + 23182.389*X5 + 1437.549*X6 - 8570.662*X7 + 147.858*X8 + 152.888*X9 + 117.222*X10 + 34552.286*X11

Chỉ số MAE: 104863.77754032993

1b.

Sau khi chạy k-fold cross validation trên từng thuộc tính ra được kết quả:

	Thuộc tính	MAE
0	nueroticism	124529.853028
1	agreeableness	125087.697550
2	extraversion	125170.086999
3	conscientiousness	125523.337223
4	openess_to_experience	125565.728015

Do đó thuộc tính tốt nhất là neuroticism vì có sai số MAE là nhỏ nhất.

Huấn luyện lại mô hình bằng thuộc tính neuroticism thì thu được
Phương trình hồi quy:

Salary = -16021.494*X + 304647.552

Sai số MAE: 119361.91739987815

1c.

Sau khi chạy k-fold cross validation trên từng thuộc tính ra được kết quả:

	Thuộc tính	MAE
0	Quant	118279.364517
1	Logical	120686.840058
2	English	121522.381223

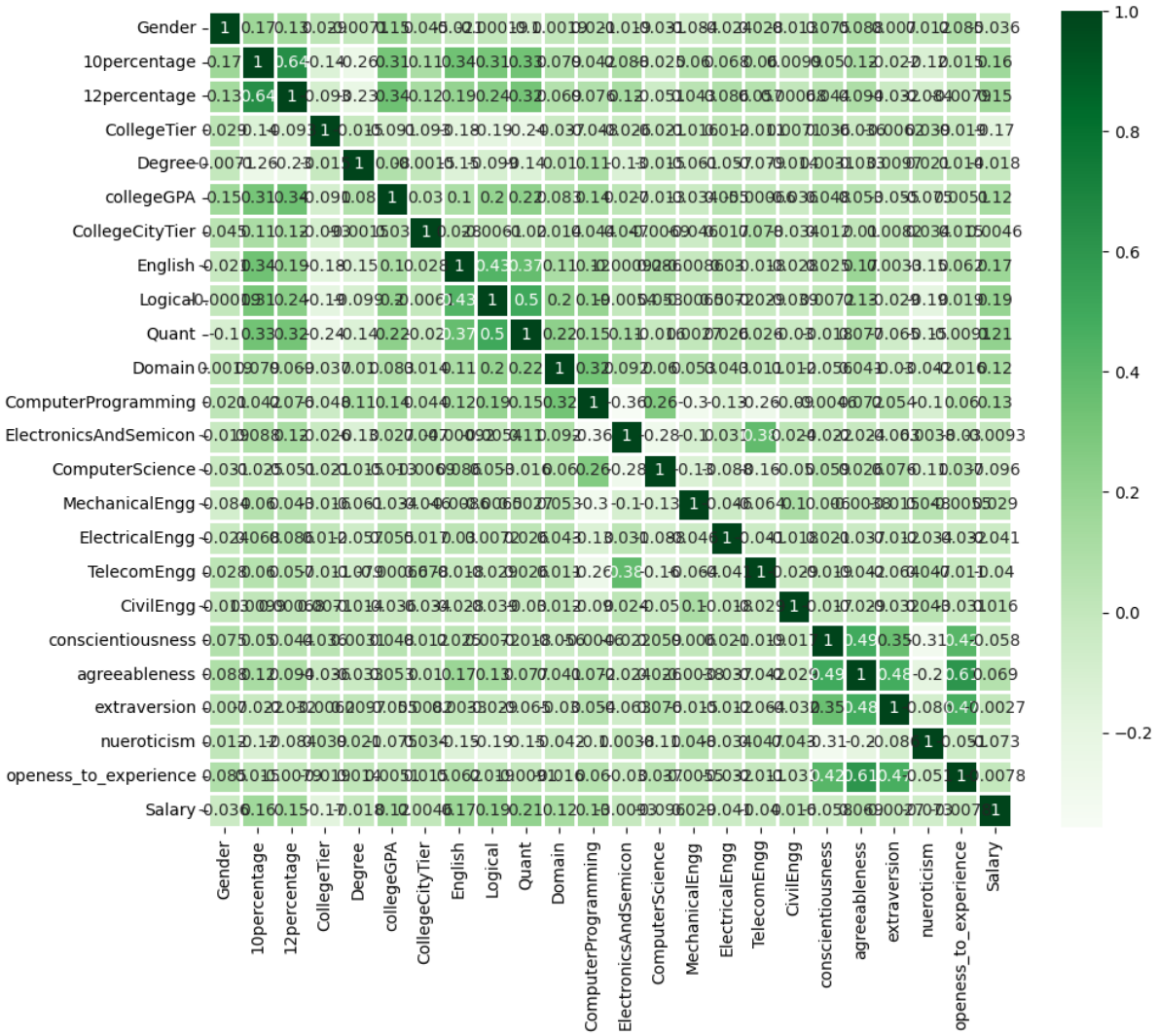
Do đó thuộc tính tốt nhất là Quant vì có sai số MAE là nhỏ nhất.

Huấn luyện lại mô hình bằng thuộc tính neuroticism thì thu được
Phương trình hồi quy:

Salary = 368.852*X + 117759.729

Sai số MAE: 108814.05968837196

1d.



Sử dụng heatmap để đánh giá độ tương quan giữa các thuộc tính.

Em lựa chọn các thuộc tính Quant (vì nó tương quan với nhiều thuộc tính khác), neuroticism (vì nó ít tương quan với các thuộc tính khác) và CompputerProgramming (vì nó tương quan vừa phải với các thuộc tính khác). Em tạo 3 dataframe, mỗi dataframe là top 5 thuộc tính có tương quan với 3 thuộc tính trên nhất.

Sử dụng k-fold cross validation tính toán sai số MAE của 3 dataframe trên:

	Dataframe	MAE
0	Nueroticism	114299.261779
1	ComputerProgramming	115780.413385
2	Quant	116102.195231

Do đó em lựa chọn dataframe 2 là mô hình cho ra kết quả tốt nhất.

Huấn luyện lại dataframe 2 và thu được phương trình hồi quy:

Salary = -9448.587*X1 + 139.608*X2 -163.243*X3 + 1790.680*X4 + 259.252*X5

Sai số MAE: 104029.88954523273

IV. Nguồn tham khảo

https://www.geeksforgeeks.org/how-to-extract-the-intercept-from-a-linear-regression-model-in-r/ sử dụng để tính hệ số góc và hệ số tự do trong câu 1b, 1c

https://www.geeksforgeeks.org/python-pandas-dataframe-corr/ sử dụng để tính hệ số tương quan trong câu 1d.

https://www.geeksforgeeks.org/pandas-concat-function-in-python/ sử dụng để hợp 2 dataframe trong câu 1b, 1c, 1d trong k-fold cross validation