ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN


# ĐỒ ÁN 3 - LINEAR REGRESSION

## TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CÔNG NGHỆ THÔNG TIN

LỚP 21CLC08

GIẢNG VIÊN: PHAN THỊ PHƯƠNG UYÊN

Nguyễn Hồng Hạnh 21127503

# Mục lục

# I Thư viện và Hàm

## 1 Thư viện

- **pandas**: dùng để thao tác với bộ dữ liệu và phân tích dữ liệu: đọc dữ liệu, tìm kiếm sự tương quan giữa các dữ liệu,...
- **numpy**: thao tác trên ma trận
- **seaborn**: tạo ra các biểu đồ trực quan trên bộ dữ liệu
- **matplotlib.pyplot**: hiển thị biểu đồ
- **itertools**: thao tác với vòng lặp
- **scipy.stats**: thực hiện các phép tính thống kê
- **sklearn**: dùng cho các thuật toán linear regression, mean absolute error,...

## 2 Hàm

**loc và iloc (pandas)**

- Được dùng để truy cập và lấy dữ liệu từ dataframe bằng nhãn (loc) hoặc chỉ mục (iloc)

**read_data()**

- Sử dụng hàm pd_read để đọc dữ liệu từ file và trả về các đặc trưng X và giá trị mục tiêu y cho các tập huấn luyện (train) và kiểm tra (test)

**detect_outliers**

- Dùng hàm quantile của thư viện pandas để tính phân vị trên và dưới, từ đó tính giới hạn trên và dưới và hàm sẽ trả về vị trí của các giá trị ngoại lệ [1]

**LinearRegression().fit()**

- Hàm sẽ giúp tính toán các hệ số và tạo mô hình hồi quy tuyến tính phù hợp với dữ liệu

**LinearRegression().predict()**

- Dùng để thực hiện dự đoán giá trị biến mục tiêu dựa trên mô hình hồi quy tuyến tính đã được tạo trước đó

**mean_absolute_error**

- Được dùng để tính giá trị sai số tuyệt đối trung bình giữa các giá trị thực tế và các giá trị dự đoán, từ đó đánh giá hiệu suất của mô hình

**sample(frac=1)**

- Dùng để xáo trộn bộ dữ liệu trước khi tiến hành cross validation

**KFold**

- Thực hiện phân chia dữ liệu bằng phương pháp K-Fold Cross Validation

**histplot, boxplot, heatmap**

- Tạo biểu đồ histogram, boxplot, heatmap từ dữ liệu truyền vào

**pd.crosstab**

- Tạo bảng tần suất chéo liên kết giữa các biến trong dữ liệu

**stats.chi2_contingency**

- Thực hiện kiểm định dựa trên thống kê chi-square trên bảng tần suất chéo từ đó kết luận xem hai biến phân loại có tương quan hay không [2]

**corr()**

- Tính ma trận tương quan giữa các cột trong một DataFrame

# II Kết quả và nhận xét

## 1 1a

- Công thức hồi quy:

$$\begin{aligned}\text{Salary} = &-23183.330 \times \text{Gender} + 702.767 \times \text{10percentage} + 1259.019 \times \text{12percentage} \\ &- 99570.608 \times \text{CollegeTier} + 18369.962 \times \text{Degree} + 1297.532 \times \text{collegeGPA} - 8836.727 \times \text{CollegeCityTier} \\ &+ 141.760 \times \text{English} + 145.742 \times \text{Logical} + 114.643 \times \text{Quant} + 34955.750 \times \text{Domain} + 49248.090\end{aligned}$$

- MAE: 105052.52978823156

- Nhận xét: Với một mô hình có quá nhiều thuộc tính (11 thuộc tính) thì kết quả đo không quá khả quan và gần như không kết luận được sự liên kết giữa các thuộc từ từ các hệ số

## 2 1b

- Kết quả tương ứng cho 5 mô hình từ k-fold Cross Validation (k = 8) (kết quả có thể thay đổi sau mỗi lần chạy do xáo trộn dữ liệu):

| STT | Đặc trưng | MAE |
|---|---|---|
| 1 | conscientiousness | 119446.90136011844 |
| 2 | agreeableness | 118177.47568442518 |
| 3 | extraversion | 118943.33570124724 |
| 4 | nueroticism | 119395.95220026761 |
| 5 | openess_to_experience | 118951.05658691078 |

- Đặc trưng tốt nhất: agreeableness

- Công thức hồi quy theo đặc trưng tốt nhất:

$$\text{Salary} = 15834.939 \times \text{agreeableness} + 305037.280$$

- MAE: 118153.16335110964

- Nhận xét: MAE tương ứng cho 5 mô hình không có sự chênh lệch quá lớn, nhìn chung kết quả không mấy khả quan

- Giải thích: Khi tìm kiếm việc làm thì những yếu tố về tính cách không được để ý tới quá nhiều và thường không phải yếu tố quyết định cho việc nhận được việc làm hay tăng lương

## 3 1c

- Kết quả tương ứng cho 3 mô hình từ k-fold Cross Validation (k = 8) (kết quả có thể thay đổi sau mỗi lần chạy do xáo trộn dữ liệu):

| STT | Đặc trưng | MAE |
|---|---|---|
| 1 | English | 115692.84492739601 |
| 2 | Logical | 114838.00915836754 |
| 3 | Quant | 108832.02818478004 |

- Đặc trưng tốt nhất: Quant

- Công thức hồi quy theo đặc trưng tốt nhất:

$$textSalary = 368.852 \times \text{Quant} + 117759.729$$

- MAE: 108814.05968837194

- Nhận xét: Kết quả đo được của đặc trưng Quant có sự chênh lệch rõ ràng hơn so với 2 đặc trưng còn lại là English và Logical và kết quả đo được của các đặc trưng kỹ năng tốt hơn kết quả đo được của các đặc trưng tính cách

- Giải thích: Mức lương và việc làm của kỹ sư ngay sau khi tốt nghiệp sẽ ảnh hưởng nhiều bởi các kỹ năng của kỹ sư và khi đi ứng tuyển việc làm thì tiêu chí về kỹ năng sẽ được ưu tiên hơn là các tiêu chí về tính cách

## 4 1d

- Tiền xử lí dữ liệu:

+ Loại bỏ các cột có missing value $\geq 50\%$ (ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg)

+ Phân loại dữ liệu thành 2 kiểu dữ liệu chính: categorical data và numerical data

- Categorical: Gender, CollegeTier, Degree, CollegeCityTier
- Numerical: 10percentage, 12percentage, collegeGPA, English, Logical, Quant, Domain, ComputerProgramming, conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience

+ Đối với numerical data, kiểm tra và loại bỏ outliers

- Tìm kiếm sự tương quan giữa các đặc trưng:

+ Với categorical data: dùng kiểm định chi bình phương với từng cặp đặc trưng để kiểm tra sự tương quan giữa các đặc trưng $\rightarrow$ Kết quả có được: có cặp đặc trưng Gender - CollegeCityTier tương quan với nhau (p-value = $0.038 < 0.05$)

+ Với numerical data: tính hệ số tương quan giữa từng cặp đặc trưng (corr()) $\rightarrow$ Kết quả có được: 2 cặp đặc trưng 12percentage - 10percentage (0.65) và ComputerProgramming - Domain (0.58) tương quan thuận mạnh

- Xây dựng mô hình:

+ model_1 - Gender$\times$ 2 + CollegeCityTier: dựa trên sự tương quan có được từ kiểm định chi bình phương (Gender $\times$ 2 để đảm bảo không bị trùng giá trị với Gender và CollegeCityTier khác nhau: $2 + 0 = 1 + 1$)

+ model_2 - 12percentage, 10percentage: cặp đặc trưng có hệ số tương quan cao nhất trong các cặp numerical data

+ model_3 - Quant, agreeableness, $\frac{12percentage+10percentage}{2}$, ComputerProgramming + Domain, Gender$\times$ 2 + CollegeCityTier: kết hợp với đặc trưng agreeableness, Quant ở câu b, c, cặp đặc trưng tương quan mạnh nhất thể hiện tổng điểm đạt được ở trung học 12percentage - 10percentage, cặp đặc trưng tương quan mạnh thứ hai thể hiện điểm ở phần chuyên ngành ComputerProgramming - Domain và cặp đặc trưng ở categorical data phụ thuộc nhau Gender - CollegeCityTier

- Kết quả tương ứng cho các mô hình từ k-fold Cross Validation (k = 8) (kết quả có thể thay đổi sau mỗi lần chạy do xáo trộn dữ liệu):

| STT | Mô hình | MAE |
|---|---|---|
| 1 | model_1 | 119318.12766466783 |
| 2 | model_2 | 110938.4162553937 |
| 3 | model_3 | 104681.90101242208 |

- Mô hình tốt nhất: model_3

- Công thức hồi quy theo đặc trưng tốt nhất:

$$\begin{aligned}\text{Salary} =& 255.973 \times \text{Quant} + 9480.813 \times \text{agreeableness} + 2668.116 \times \frac{10\text{percentage} + 12\text{percentage}}{2} \\ &+ 104.774 \times (\text{ComputerProgramming} + \text{Domain}) - 10124.612 \times (\text{Gender} \times 2 + \text{CollegeCityTier})\end{aligned}$$

- MAE: 104601.4384174429

- Nhận xét: Mô hình bao gồm 8 đặc trưng (Quant, agreeableness, 12percentage, 10percentage, ComputerProgramming, Domain, Gender, CollegeCityTier) nhưng cho được kết quả đo tốt hơn mô hình 11 đặc trưng ở câu a do ở mô hình này kết hợp các đặc trưng nổi bật ở câu b, c và ở hai mô hình còn lại

# III Tài liệu tham khảo

[1] Detect and remove outliers

[2] Pearson's Chi-Square Test