TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM
KHOA CÔNG NGHỆ THÔNG TIN

---o0o---

BÁO CÁO ĐỒ ÁN THỰC HÀNH

LINEAR REGRESSION

Môn : **Toán ứng dụng và thống kê cho công nghệ thông tin**

Giảng Viên : **Phan Thị Phương Uyên**

Lớp : **21CLC05**

Tên : **Huỳnh Minh Quang**

MSSV : **21127149**

Thành phố Hồ Chí Minh, ngày 20 tháng 08 năm 2023

MỤC LỤC

I. Giới thiệu

1. Thông tin sinh viên

- Họ và tên: Huỳnh Minh Quang
- MSSV: 21127149
- Lớp: 21CLC05

2. Đồ án Linear Regression

- 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 [1].

II. Các thư viện sử dụng trong đồ án

1. pandas (import pandas as pd):

- **Chức năng:**

- Thư viện cho phân tích và xử lý dữ liệu.

- **Mục đích dùng:**

- Đọc dữ liệu từ file CSV (pd.read_csv), tạo DataFrame để lưu trữ dữ liệu, truy cập và chọn các dòng và cột của DataFrame (iloc), và thao tác với dữ liệu bảng.

2. numpy (import numpy as np):

- **Chức năng:**

- Thư viện cho tính toán số học và thao tác trên mảng đa chiều.

- **Mục đích dùng:**

- Chuyển đổi dữ liệu DataFrame thành mảng numpy để thao tác, tính toán bình phương dữ liệu (X_train_1d**2), và thực hiện các tính toán khác liên quan đến dữ liệu số.

3. seaborn (import seaborn as sns):

- **Chức năng:**

- Thư viện cho trực quan hóa dữ liệu.

- **Mục đích dùng:**

- Vẽ biểu đồ heatmap để hiển thị ma trận tương quan (sns.heatmap).

4. matplotlib.pyplot (import matplotlib.pyplot as plt):

- **Chức năng:**

- Thư viện cho tạo và hiển thị biểu đồ và đồ thị.

- **Mục đích dùng:**

- Vẽ biểu đồ tương quan heatmap (plt.figure, plt.show).

5. sklearn.linear_model.LinearRegression (from sklearn.linear_model import LinearRegression):

- **Chức năng:**

- Thư viện cho mô hình hồi quy tuyến tính.

- **Mục đích dùng:**

- Tạo và huấn luyện mô hình hồi quy tuyến tính (LinearRegression), truy xuất các trọng số và bias của mô hình sau huấn luyện (coef_, intercept_).

6. sklearn.metrics.mean_absolute_error (from sklearn.metrics import mean_absolute_error as MAE):

- **Chức năng:**

- Thư viện cho đánh giá mô hình bằng độ đo Mean Absolute Error (MAE).

- **Mục đích dùng:**

- Tính giá trị MAE giữa dự đoán và giá trị thực tế (MAE).

7. sklearn.model_selection.cross_val_score (from sklearn.model_selection import cross_val_score):

- **Chức năng:**

- Thư viện cho cross-validation của mô hình.

- **Mục đích dùng:**

- Thực hiện cross-validation trên mô hình để đánh giá hiệu suất (cross_val_score).

8. sklearn.model_selection.KFold (from sklearn.model_selection import KFold):

- **Chức năng:**

- Thư viện cho chia dữ liệu thành các tập con (folds) để thực hiện cross-validation.

- **Mục đích dùng:**

- Tạo đối tượng chia dữ liệu thành các fold (KFold).

III. Các hàm đã sử dụng

1. pd.read_csv('file_path'): [2]

- **Chức năng:**

- Hàm này dùng để đọc dữ liệu từ một file CSV và tạo một DataFrame từ dữ liệu đó.

- **Cú pháp:**

- pd.read_csv('file_path')

- **Tham số:**

- 'file_path': Đường dẫn tới tệp CSV chứa dữ liệu cần đọc.

- **Kết quả trả về:**

- Một DataFrame chứa dữ liệu từ tệp CSV.

2. DataFrame.iloc[row_indices, column_indices]: [2]

- **Chức năng:**

- Hàm này cho phép truy cập và trích xuất dữ liệu từ DataFrame bằng cách sử dụng chỉ số hàng và cột.

- **Cú pháp:**

- DataFrame.iloc[row_indices, column_indices]

- **Tham số:**

- row_indices: Chỉ số của các hàng mà bạn muốn truy cập.

- column_indices: Chỉ số của các cột mà bạn muốn truy cập.

- **Kết quả trả về:**

- Một DataFrame hoặc Series con chứa dữ liệu tương ứng với chỉ số hàng và cột được chỉ định.

3. LinearRegression(): [3]

- **Chức năng:**

- Hàm này tạo một đối tượng mô hình hồi quy tuyến tính, sẵn sàng để huấn luyện trên

dữ liệu.

- **Cú pháp:**

- LinearRegression()

- **Tham số:**

- Không có tham số đầu vào.

- **Kết quả trả về:**

- Một đối tượng mô hình hồi quy tuyến tính được khởi tạo.

4. cross_val_score(estimator, X, y, cv=kf, scoring='neg_mean_absolute_error'):

- **Chức năng:**

- Hàm này thực hiện cross-validation trên một mô hình cụ thể và đánh giá hiệu suất của mô hình đó.

- **Cú pháp:**

- cross_val_score(estimator, X, y, cv=kf, scoring='neg_mean_absolute_error')

- **Tham số:**

- estimator: Đối tượng mô hình đã được khởi tạo (ví dụ: LinearRegression).
- X: Dữ liệu đầu vào (đặc trưng).
- y: Dữ liệu mục tiêu.
- cv: Đối tượng chia dữ liệu thành các fold để thực hiện cross-validation.
- scoring: Độ đo để đánh giá hiệu suất mô hình.

- **Kết quả trả về:**

- Một mảng chứa các giá trị độ đo hiệu suất của mô hình trên các fold.

5. sns.heatmap(data, annot=True, cmap='coolwarm'):

- **Chức năng:**

- Hàm này tạo và hiển thị biểu đồ heatmap để hiển thị ma trận dữ liệu dưới dạng màu sắc.

- **Cú pháp:**

- sns.heatmap(data, annot=True, cmap='coolwarm')

- **Tham số:**

- data: Dữ liệu ma trận cần hiển thị.
- annot: Hiển thị giá trị trên từng ô của heatmap.
- cmap: Mã màu sắc được sử dụng cho biểu đồ.

- **Kết quả:**

- Hiển thị biểu đồ heatmap trong giao diện đồ họa.

6. mean_absolute_error(y_true, y_pred): [3]

- **Chức năng:**

- Hàm này tính độ đo Mean Absolute Error (MAE) giữa hai chuỗi giá trị.

- **Cú pháp:**

- mean_absolute_error(y_true, y_pred)

- **Tham số:**

- y_true: Chuỗi giá trị thực tế.
- y_pred: Chuỗi giá trị dự đoán.

- **Kết quả trả về:**

- Một giá trị số thực đại diện cho MAE giữa y_true và y_pred.

7. KFold(n_splits=k, shuffle=True, random_state=42) [4]:

- **Chức năng:**

- Hàm này tạo các tập con (folds) từ dữ liệu để thực hiện cross-validation.

- **Cú pháp:**

- KFold(n_splits=k, shuffle=True, random_state=42)

- **Tham số:**

- n_splits: Số lượng fold (tập con) mà dữ liệu sẽ được chia thành.
- shuffle: Xáo trộn dữ liệu trước khi chia thành các fold.
- random_state: Seed để đảm bảo kết quả tái lập được.

8. DataFrame.drop_duplicates(): [2]

- **Chức năng:**

- Hàm này loại bỏ các hàng trùng lặp trong DataFrame, trả về phiên bản duy nhất.

- **Cú pháp:**

- DataFrame.drop_duplicates()

- **Tham số:**

- Không có tham số.

- **Kết quả trả về:**

- Một phiên bản DataFrame mới chứa các hàng duy nhất.

9. DataFrame.sort_values(by='column_name', ascending=False): [2]

- **Chức năng:**

- Hàm này sắp xếp DataFrame dựa trên giá trị của một cột cụ thể.

- **Cú pháp:**

- DataFrame.sort_values(by='column_name', ascending=False)

- **Tham số:**

- by: Tên của cột mà bạn muốn sắp xếp theo.
- ascending: True để sắp xếp theo thứ tự tăng dần, False để sắp xếp theo thứ tự giảm dần.

- **Kết quả trả về:**

- Một phiên bản DataFrame mới đã được sắp xếp.

10. LinearRegression.coef_ và LinearRegression.intercept_: [3]

- **Chức năng:**

- Truy xuất trọng số và bias của mô hình hồi quy tuyến tính sau khi huấn luyện.

- **Cú pháp:**

- lr.coef_ (truy xuất trọng số), lr.intercept_ (truy xuất bias).

- **Kết quả trả về:**

- Một mảng chứa các trọng số của đặc trưng (coef_) và một số thực đại diện cho bias (intercept_).

IV. Kết quả và nhận xét

1. Yêu cầu 1a

- **Yêu cầu:**

- Sử dụng 11 đặc trưng đầu tiên đề bài cung cấp bao gồm: `Gender`, `10percentage`, `12percentage`, `CollegeTier`, `Degree`, `collegeGPA`, `CollegeCityTier`, `English`, `Logical`, `Quant`, `Domain`. Huấn luyện 1 lần duy nhất cho 11 đặc trưng nói trên cho toàn bộ tập huấn luyện (`train.csv`)
- Thể hiện công thức cho mô hình hồi quy (tính theo 11 đặc trưng trên)
- Báo cáo 1 kết quả trên tập kiểm tra (`test.csv`) cho mô hình vừa huấn luyện được

- **Kết quả:**

- Trọng số (weights) và bias của mô hình đã được truy xuất và hiển thị. Điều này cho biết độ ảnh hưởng của mỗi đặc trưng đến mức lương:

```
Weights:
          Feature        Weight
0          Gender -23183.329508
1    10percentage    702.766792
2    12percentage   1259.018788
3     CollegeTier -99570.608141
4          Degree  18369.962450
5      collegeGPA   1297.532000
6  CollegeCityTier  -8836.727123
7         English    141.759939
8         Logical    145.742347
9           Quant    114.643313
10         Domain  34955.750405
bias:  49248.089734813664
```

- Mô hình hồi quy được biểu diễn qua công thức, trong đó mức lương được dự đoán dựa trên các đặc trưng và trọng số tương ứng:

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

- Kết quả trên tập kiểm tra (test.csv):

```
MAE trên tập kiểm tra: 105052.52978823145
```

- MAE (Mean Absolute Error) trên tập kiểm tra là 105052.53. Điều này có nghĩa là giá trị dự đoán của mô hình trung bình sai lệch khoảng 105,052 đối với giá trị thực tế.

- **Nhận xét:**

- Mô hình hồi quy tuyến tính đã được áp dụng một cách cơ bản để dự đoán mức lương dựa trên 11 đặc trưng ban đầu.
- Giá trị MAE khá cao, cho thấy mô hình có khả năng sai số lớn trong việc dự đoán mức lương.
- Có thể cân nhắc tối ưu hóa mô hình bằng cách thử nghiệm các kỹ thuật tiền xử lý dữ liệu, chọn lọc đặc trưng hoặc sử dụng các mô hình phức tạp hơn để cải thiện hiệu suất dự đoán.

2. Yêu cầu 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`.
- Yêu cầu sử dụng k-fold Cross Validation (k tối thiểu là 5) để tìm ra đặc trưng tốt nhất trong các đặc trưng tính cách.

- **Kết quả:**

- Thử nghiệm các đặc trưng tính cách và sử dụng k-fold Cross Validation (k = 5) để đánh giá hiệu suất của mỗi mô hình.
- Kết quả từ Cross Validation giúp thể hiện độ sai lệch dự đoán trung bình, và tìm ra đặc trưng tốt nhất là 'nueroticism':

```
STT  | Mô hình với 1 đặc trưng tính cách | MAE
-------------------------------------------
 1   | conscientiousness                 | 124182.564
-------------------------------------------
 2   | agreeableness                     | 123706.055
-------------------------------------------
 3   | extraversion                      | 123809.926
-------------------------------------------
 4   | nueroticism                       | 123473.400
-------------------------------------------
 5   | openess_to_experience             | 123818.334
-------------------------------------------
Đặc trưng tốt nhất:  nueroticism
```

- Huấn luyện lại với đặc trưng tốt nhất là 'nueroticism', trọng số (weights) và bias của mô hình đã được truy xuất và hiển thị. Điều này cho biết độ ảnh hưởng của mỗi đặc trưng đến mức lương:

```
Weights:
       Feature        Weight
0  nueroticism -16021.493662
bias:  304647.55255226186
```

- Mô hình hồi quy được biểu diễn qua công thức, trong đó mức lương được dự đoán dựa trên các đặc trưng và trọng số tương ứng:

$$Salary = 304647.553 + (-16021.494 * nueroticism)$$

- Kết quả trên tập kiểm tra (test.csv):

```
Đặc trưng tốt nhất:  nueroticism
MAE trên tập kiểm tra: 119361.91739987815
```

- MAE trên tập kiểm tra khi sử dụng mô hình với đặc trưng nueroticism là 119361.92.
- Kết quả này cho thấy mức sai số trung bình giữa dự đoán và thực tế trên tập kiểm tra.

▪ **Nhận xét:**

- Dựa trên kết quả Cross Validation, tất cả các đặc trưng tính cách cho kết quả MAE tương tự nhau, không có sự khác biệt đáng kể.
- Mô hình dự đoán dựa trên nueroticism có MAE thấp nhất, tuy nhiên, giá trị này vẫn khá cao, cho thấy mô hình không thể dự đoán chính xác mức lương trên tập kiểm tra.
- Có thể cần thêm các đặc trưng hoặc xem xét các mô hình phức tạp hơn để cải thiện hiệu suất dự đoán.

3. Yêu cầu 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`
- Yêu cầu sử dụng k-fold Cross Validation (k tối thiểu là 5) để tìm ra đặc trưng tốt nhất.

▪ **Kết quả:**

- Thử nghiệm các đặc trưng tính cách và sử dụng k-fold Cross Validation (k = 5) để đánh giá hiệu suất của mỗi mô hình.
- Kết quả từ Cross Validation giúp thể hiện độ sai lệch dự đoán trung bình, và tìm ra đặc trưng tốt nhất là ‘Quant’:

```
STT  | Mô hình với 1 đặc trưng tính cách | MAE
-------------------------------------------
 1   | English                           | 120728.604
-------------------------------------------
 2   | Logical                           | 119932.504
-------------------------------------------
 3   | Quant                             | 117353.838
-------------------------------------------
Đặc trưng tốt nhất:  Quant
```

- Huấn luyện lại với đặc trưng tốt nhất là 'Quant', trọng số (weights) và bias của mô hình đã được truy xuất và hiển thị. Điều này cho biết độ ảnh hưởng của mỗi đặc trưng đến mức lương:

```
Weights:
  Feature      Weight
0   Quant  368.852464
bias:  117759.72931230717
```

- Mô hình hồi quy được biểu diễn qua công thức, trong đó mức lương được dự đoán dựa trên các đặc trưng và trọng số tương ứng:

$$\text{Salary} = 117759.729 + (368.852 * Quant)$$

- Kết quả trên tập kiểm tra (test.csv):

```
Đặc trưng tốt nhất:  Quant
MAE trên tập kiểm tra: 108814.05968837196
```

- MAE trên tập kiểm tra khi sử dụng mô hình với đặc trưng Quant là 108814.06.
- Kết quả này cho thấy mức sai số trung bình giữa dự đoán và thực tế trên tập kiểm tra.

▪ **Nhận xét:**

- Đặc trưng Quant có MAE thấp nhất, cho thấy đặc trưng này có khả năng dự đoán tốt hơn so với English và Logical.
- Mô hình dự đoán dựa trên đặc trưng Quant cho kết quả tốt hơn so với các đặc trưng khác, nhưng vẫn còn mức sai số khá cao.

4. Yêu cầu 1d

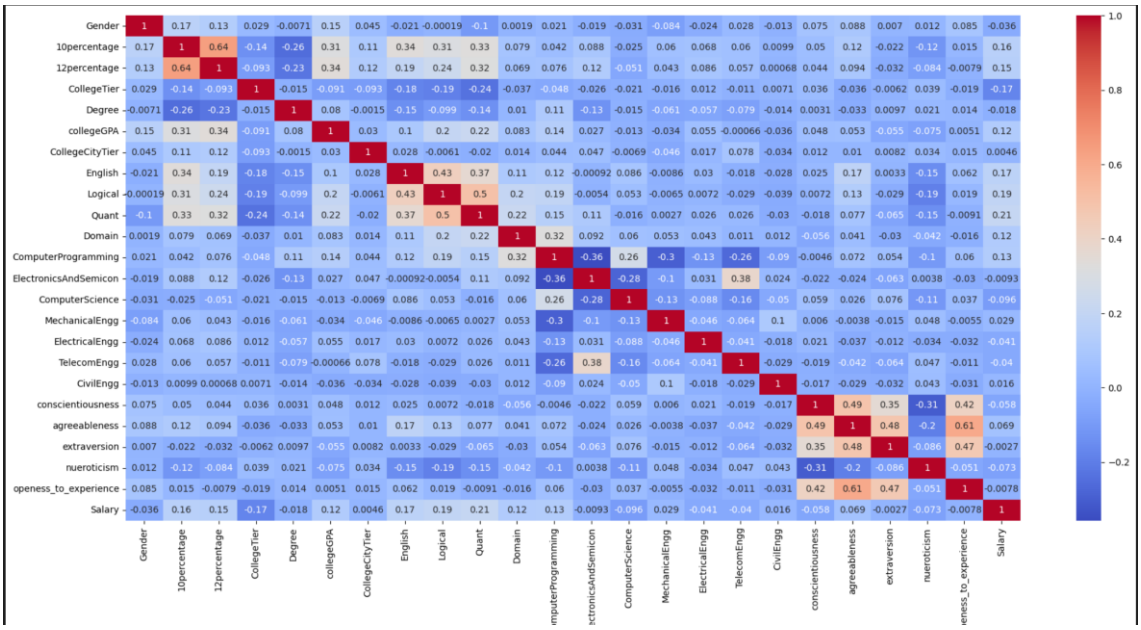
▪ **Yêu cầu:**

- Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất.
- Xây dựng `m` mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a, 1b và 1c
- Mô hình có thể là sự kết hợp của 2 hoặc nhiều đặc trưng
- Mô hình có thể sử dụng đặc trưng đã được chuẩn hóa hoặc biến đổi (bình phương, lập phương...)

- Mô hình có thể sử dụng đặc trưng được tạo ra từ 2 hoặc nhiều đặc trưng khác nhau (cộng 2 đặc trưng, nhân 2 đặc trưng...)

- **Xây dựng mô hình:**

- Sử dụng thư viện seaborn để vẽ biểu đồ tương quan heatmap dựa trên ma trận tương quan correlation_matrix [5].



- Các ô trong biểu đồ sẽ được đánh dấu bằng giá trị tương quan tương ứng.

- Mục đích là để hình dung mức độ tương quan giữa các cặp đặc trưng trong biến 'correlation_matrix'.

- Lấy các giá trị tương quan của biến 'Salary' với tất cả các đặc trưng từ correlation_matrix.

- Sắp xếp giá trị tương quan giảm dần.

- Chọn các đặc trưng có giá trị tuyệt đối của tương quan lớn hơn 0.1 và lưu chúng vào danh sách high_features và ta được các đặc trưng có độ tương quan cao so với 'Salary' là: 'Quant', 'Logical', 'English', '10percentage', '12percentage', 'ComputerProgramming', 'collegeGPA', 'Domain', 'CollegeTier'.

```
Các đặc trưng có tương quan cao với biến mục tiêu (Salary):
['Quant', 'Logical', 'English', '10percentage', '12percentage', 'ComputerProgramming', 'collegeGPA', 'Domain', 'CollegeTier']
```

- Xây dựng và đánh giá ba mô hình hồi quy tuyến tính khác nhau:

- **Mô hình 1:** Sử dụng các đặc trưng có tương quan cao với 'Salary' ban đầu.

- **Mô hình 2:** Sử dụng các đặc trưng có tương quan cao với 'Salary' và biến đổi bằng cách bình phương.

- **Mô hình 3:** Sử dụng các đặc trưng kết hợp 'Quant' và 'Logical'.

▪ **Kết quả:**

- Sử dụng k-fold Cross Validation (k = 5) để đánh giá hiệu suất của mỗi mô hình xây dựng.

```
-----------------------------------------------
STT         | Mô hình                                                  | MAE
-----------------------------------------------
1           | Sử dụng các đặc trưng có độ tương quan cao với salary       | 113147.568
-----------------------------------------------
2           | Sử dụng các đặc trưng có độ tương quan cao với salary và biến đổi (bình phương) | 113350.030
-----------------------------------------------
3           | Sử dụng các đặc trưng kết hợp (Quant) và (loical)           | 116679.675
-----------------------------------------------
Mô hình tốt nhất:  Sử dụng các đặc trưng có độ tương quan cao với salary
```

- Kết quả trên bảng tương ứng với mỗi mô hình, thể hiện độ sai lệch dự đoán trung bình.
- Mô hình có MAE thấp nhất là "Sử dụng các đặc trưng có độ tương quan cao với salary"
- Huấn luyện lại với mô hình tốt nhất là "Sử dụng các đặc trưng có độ tương quan cao với salary", trọng số (weights) và bias của mô hình đã được truy xuất và hiển thị. Điều này cho biết độ ảnh hưởng của mỗi đặc trưng đến mức lương:

```
Weights:
                Feature        Weight
0                 Quant    130.229185
1               Logical    127.518911
2               English    133.846284
3          10percentage    555.442391
4          12percentage   1056.192159
5   ComputerProgramming     69.645449
6             collegeGPA   1092.688515
7                Domain  26015.586709
8           CollegeTier -99626.911155
bias:  63725.571575201466
```

- Mô hình hồi quy được biểu diễn qua công thức, trong đó mức lương được dự đoán dựa trên các đặc trưng và trọng số của mô hình:

$$\begin{aligned}\text{Salary} = 63725.572 &+ (130.229 * Quant) + (127.519 * Logical) + (133.846 * English) \\ &+ (555.442 * 10percentage) + (1056.192 * 12percentage) + (69.645 * ComputerProgramming) \\ &+ (1092.689 * collegeGPA) + (26015.587 * Domain) + (-99626.911 * CollegeTier)\end{aligned}$$

- Kết quả trên tập kiểm tra (test.csv):

```
Mô hình tốt nhất:  Sử dụng các đặc trưng có độ tương quan cao với salary
MAE trên tập kiểm tra: 104201.44201243388
```

- Kết quả này cho thấy mức sai số trung bình giữa dự đoán và thực tế trên tập kiểm tra.

- **Nhận xét:**

- Mô hình "Sử dụng các đặc trưng có độ tương quan cao với salary" đã cho kết quả tốt nhất trong các mô hình đã xây dựng.
- Mô hình này dự đoán dựa trên một số đặc trưng có độ tương quan cao với biến mục tiêu, và cho thấy khả năng dự đoán tốt hơn so với các mô hình khác. Tuy nhiên, vẫn còn mức sai số khá cao cần cải thiện.

5. Tổng quan

- **Yêu cầu 1a:**

- Trong phần này, đã sử dụng 11 đặc trưng ban đầu để huấn luyện một mô hình hồi quy tuyến tính. Mô hình này đã cho kết quả MAE trên tập kiểm tra là 105,052.53. Điều này có thể chỉ ra rằng mô hình chưa tốt lắm trong việc dự đoán mức lương dựa trên các đặc trưng đã chọn. Có thể cần xem xét thêm các phương pháp khác để cải thiện độ chính xác của mô hình.

- **Yêu cầu 1b:**

- Trong phần này, đã phân tích ảnh hưởng của đặc trưng tính cách lên mức lương bằng cách thử nghiệm lần lượt trên các đặc trưng tính cách và sử dụng k-fold Cross Validation để tìm ra đặc trưng tốt nhất. Kết quả cho thấy đặc trưng "neuroticism" cho kết quả tốt nhất trong việc dự đoán mức lương, với MAE trên tập kiểm tra là 119,361.92. Điều này có thể chỉ ra rằng các đặc trưng tính cách có thể không có tác động lớn đến khả năng dự đoán mức lương trong trường hợp này.

- **Yêu cầu 1c:**

- Phần này là việc phân tích ảnh hưởng của các đặc trưng ngoại ngữ, logic và định lượng đến mức lương dựa trên điểm các bài kiểm tra của AMCAT. Kết quả cho thấy đặc trưng "Quant" cho kết quả tốt nhất trong việc dự đoán mức lương, với MAE trên tập kiểm tra là 108,814.06. Điều này có thể chỉ ra rằng khả năng của mô hình dự đoán mức lương có thể được cải thiện khi tập trung vào các đặc trưng định lượng.

- **Yêu cầu 1d:**

- Trong phần này, đã thực hiện việc xây dựng mô hình riêng của mình, thử nghiệm một loạt các mô hình khác nhau với các đặc trưng khác nhau và sử dụng k-fold Cross

Validation để tìm ra mô hình tốt nhất. Kết quả cho thấy mô hình sử dụng các đặc trưng có độ tương quan cao với mức lương cho kết quả tốt nhất, với MAE trên tập kiểm tra là 104,201.44. Điều này cho thấy tầm quan trọng của việc lựa chọn và kết hợp các đặc trưng quan trọng để cải thiện khả năng dự đoán mức lương.

V. Giả thuyết

1. Yêu cầu 1b

- **Giả thuyết:**

- Mô hình đạt kết quả tốt nhất trong yêu cầu 1b dựa trên đặc trưng "neuroticism", tức tính cách thể hiện mức độ cảm xúc và ổn định tinh thần [6].

- **Giải thích:**

- Một giả thuyết có thể là tính cách của một người ảnh hưởng đến cách họ tương tác với công việc và đồng nghiệp. Người có tính cách ổn định, ít biểu lộ cảm xúc mẫu mực (thấp về neuroticism) có thể dễ dàng làm việc cùng đồng nghiệp, giúp tạo ra môi trường làm việc hiệu quả. Điều này có thể dẫn đến hiệu suất công việc tốt hơn và mức lương cao hơn.

2. Yêu cầu 1c

- **Giả thuyết:**

- Mô hình đạt kết quả tốt nhất trong yêu cầu 1c dựa trên đặc trưng "Quant", tức điểm số trong kỳ thi định lượng [7].

- **Giải thích:**

- Điểm số trong kỳ thi định lượng có thể tương quan với khả năng này. Một kỹ sư có khả năng tốt trong việc sử dụng số liệu và định lượng có thể thực hiện công việc hiệu quả hơn, dẫn đến tăng mức lương.

3. Yêu cầu 1d

- **Giả thuyết:**

- Mô hình đạt kết quả tốt nhất là các đặc trưng có độ tương quan cao so với lương. Tức lương bị ảnh hưởng bởi nhiều yếu tố quan trọng [8].

- **Giải thích:**

- Có rất nhiều yếu tố có thể ảnh hưởng đến lương. Sự kết hợp giữa trình độ học vấn và kinh nghiệm làm việc thường là yếu tố quan trọng nhất trong việc xác định mức lương của nhân viên.

VI. Quá trình xây dựng 3 mô hình

- Tìm hiểu và lựa chọn các đặc trưng có độ tương quan cao: Để xây dựng mô hình dự

đoán mức lương, bắt đầu bằng việc tìm hiểu mối quan hệ giữa các đặc trưng và biến mục tiêu "Salary". Bằng cách tính toán hệ số tương quan Pearson giữa mỗi đặc trưng và biến mục tiêu, xác định được những đặc trưng có độ tương quan cao, cho thấy khả năng ảnh hưởng lớn đến mức lương [5].

- Lựa chọn các đặc trưng có độ tương quan cao: Dựa trên kết quả của phân tích tương quan, đã chọn ra những đặc trưng có độ tương quan cao nhất với biến mục tiêu gồm: 'Quant', 'Logical', 'English', '10percentage', '12percentage', 'ComputerProgramming', 'collegeGPA', 'Domain', và 'CollegeTier'.

- **Với mô hình 1:**

- Xây dựng từ các đực trưng tìm được (có độ tương quan cao so với lương) vì các đặc trung này rất quan trọng và độ ảnh hưởng đến lương cao.

- **Với mô hình 2:**

- Xây dựng mô hình bằng cách biến đổi các đặc trưng tìm được (có độ tương quan cao so với lương) và biến đổi bằng cách bình phương. Vì biến đổi bình phương các đặc trưng có thể mang lại sự khác biệt trong việc dự đoán mức lương. Việc này có thể giúp làm nổi bật mức độ ảnh hưởng của các đặc trưng.

Với mô hình 3:

- Xây dựng mô hình bằng cách kết hợp đặc trưng 'Quant' và 'Logical'. Đây là 2 đặc trưng có độ tương quan cao với lương và sự kết hợp tạo ra tương tác và thông tin mới để góp phần vào việc dự đoán lương

VII. Tài liệu tham khảo

References

[1] "Salary Prediction Classification," [Trực tuyến]. Available: https://www.kaggle.com/datasets/manishkc06/engineering-graduate-salary-prediction. [Đã truy cập 10 08 2023].

[2] "Thư viện pandas trong python," [Trực tuyến]. Available: https://blog.luyencode.net/thu-vien-pandas-python/. [Đã truy cập 11 08 2023].

[3] "Thư Viện Scikit-learn Trong Python Là Gì?," [Trực tuyến]. Available: https://codelearn.io/sharing/scikit-learn-trong-python-la-gi. [Đã truy cập 11 08 2023].

[4] "Kỹ thuật xác thực chéo K-Fold bằng Scikit-Learn trong Python," [Trực tuyến]. Available: https://tek4.vn/ky-thua-t-xac-thuc-cheo-k-fold-bang-scikit-learn-trong-python. [Đã truy cập 11 08 2023].

[5] "Python Data Visualization With Seaborn and Matplotlib," [Trực tuyến]. Available: https://builtin.com/data-science/data-visualization-tutorial. [Đã truy cập 15 08 2023].

[6] "Tính cách con người là gì? Và sự hình thành tính cách ở trẻ," [Trực tuyến]. Available: https://genetica.asia/blog/tinh-cach-la-gi.html. [Đã truy cập 14 08 2023].

[7] "Nghiên cứu định lượng," [Trực tuyến]. Available: https://vi.wikipedia.org/wiki/Nghi%C3%AAn_c%E1%BB%A9u_%C4%91%E1%BB%8Bnh_l%C6%B0%E1%BB%A3ng.

[8] "Các Yếu Tố Ảnh Hưởng Đến Tiền Lương Của Người Lao Động," [Trực tuyến]. Available: https://glints.com/vn/blog/cac-yeu-to-anh-huong-den-tien-luong/. [Đã truy cập 15 08 2023].

[9] "Giới thiệu về Numpy (một thư viện chủ yếu phục vụ cho khoa học máy tính của Python)," [Trực tuyến]. Available: https://viblo.asia/p/gioi-thieu-ve-numpy-mot-thu-vien-chu-yeu-phuc-vu-cho-khoa-hoc-may-tinh-cua-python-maGK7kz9Kj2. [Đã truy cập 11 08 2023].

[10] "sklearn.model_selection.KFold," [Trực tuyến]. Available: sklearn.model_selection.KFold. [Đã truy cập 11 08 2023].

[11] "Mean absolute error in sklearn," [Trực tuyến]. Available: https://www.educative.io/answers/mean-absolute-error-in-sklearn. [Đã truy cập 11 08 2023].

___Hết___