

**ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT
KHOA HỆ THỐNG THÔNG TIN**



**ĐỒ ÁN MÔN HỌC
PHÂN TÍCH DỮ LIỆU VỚI
R/THON**

Giảng viên bộ môn : Ths. Nguyễn Quang Phúc

Tp. Hồ Chí Minh, ngày 11 tháng 8 năm 2022

DANH MỤC HÌNH ẢNH

Hình 1. Kiểm tra dữ liệu.....	2
Hình 2. Trực quan hoá sự tương quan giữa các biến.....	4
Hình 3. Kết quả thu được từ mô hình hồi quy.....	5
Hình 4. Trực quan hoá tương quan giữa Sales và Youtube spent.	6
Hình 5. Kết quả thu được từ tập dữ liệu training.....	7
Hình 6. Tương quan giữa Sales và Facebook spent.	8
Hình 7. Kết quả thu được từ mô hình hồi quy tuyến tính đa biến.	9
Hình 8. Tiên lượng với dữ liệu mới trong tương lai.....	11
Hình 9. Nhập tất cả các thư viện cần thiết.....	12
Hình 10. Đọc và kiểm tra dữ liệu.	13
Hình 11. Biểu đồ Histogram nhận định dữ liệu ban đầu.	13
Hình 12. Gán biến độc lập và biến phụ thuộc.	14
Hình 13. Tạo mô hình hồi quy với train data.	14
Hình 14. Dự đoán kết quả với tập dữ liệu test data.	14
Hình 15. Ma trận nhầm lẫn kết hợp biểu đồ nhiệt.....	14
Hình 16. Xem các giá trị xác suất xảy ra.....	15
Hình 17. Classification Report.	15
Hình 18. ROC Curve.	16
Hình 19. Kiểm tra tính toàn vẹn.	18
Hình 20. Biểu đồ pairplot.	19
Hình 21. Biểu đồ nhiệt kiểm tra đa cộng tuyến.....	20
Hình 22. Chỉ số intercept và giá trị coef.....	21
Hình 23. Độ chính xác của mô hình là 80%.....	21
Hình 24. Kết quả dự báo trên dữ liệu test data.	21
Hình 25. Đường hồi quy của mô hình dự báo.	22
Hình 26. Kiểm tra toàn vẹn dữ liệu.	24
Hình 27. Mã hoá cột Salary.....	24
Hình 28. Chuyển về giá trị từ 0 đến 1.	25

Hình 29. Biểu đồ nhiệt heatmap.	25
Hình 30. Giá trị Intercept và Coef.	25
Hình 31. Độ chính xác của mô hình là 78%.	26
Hình 32. Kết quả tiên lượng thu được với test data.	26
Hình 33. Đường hồi quy của mô hình	27
Hình 34. Dữ liệu thực tế của chỉ số S&P 500.	28
Hình 35. Dữ liệu phân tách train test.	29
Hình 36. Phân rã chuỗi thời gian.	30
Hình 37. Trực quan hoá các đường trung bình.	31
Hình 38. Kiểm định tự tương quan.	32
Hình 39. Xây dựng mô hình.	34
Hình 40. Biến động của chỉ số S&P với mức độ tương quan tại thời điểm t và $t-1$	35
Hình 41. Dự báo giá cổ phiếu.	36
Hình 42. Xu hướng biến động của chỉ số giai đoạn 2020 - 2022.	37
Hình 43. Chia tập dữ liệu train test theo tỷ lệ 9:1.	38
Hình 44. Phân rã chuỗi dữ liệu.	39
Hình 45. Kiểm định mô hình với ADF và KPSS.	40
Hình 46. Kết quả lấy sai phân bậc nhất.	41
Hình 47. Kiểm định tương quan tổng thể.	42
Hình 48. Kiểm định tương quan riêng.	43
Hình 49. Xây dựng mô hình ARIMA.	44
Hình 50. Xây dựng mô hình ARIMA theo (2,1,2).	45
Hình 51. Dự đoán tỷ lệ thất nghiệp dựa trên mô hình.	46
Hình 52. Kiểm tra thông tin chung của dữ liệu.	48
Hình 53. Chuyển đổi biến phân loại thành các biến chỉ số.	49
Hình 54. Chia tập dữ liệu train test.	49
Hình 55. Tạo mô hình cây quyết định với tập dữ liệu test data.	50
Hình 56. Độ chính xác của mô hình cây quyết định là 95%.	50
Hình 57. Cây quyết định với tập test data.	50

Hình 58. Sử dụng phương thức <code>tree.plot tree</code> để hiển thị cây.....	51
Hình 59. Trực quan hoá cây quyết định.	51
Hình 60. Mô hình được xây dựng với 20 cây (<code>n_estimators = 20</code>).	52
Hình 61. Mô hình với 20 cây được xây dựng.....	52
Hình 62. Mô hình với 200 cây được xây dựng.....	53
Hình 63. Xây dựng Neural Network.	54
Hình 64. Ma trận nhầm lẫn với train data.....	54
Hình 65. Ma trận nhầm lẫn với test data.	54
Hình 66. Quản lý và phân công công việc với Jira.	Error! Bookmark not defined.

MỤC LỤC

PHẦN A. NỘI DUNG BÁO CÁO KẾT QUẢ PHÂN TÍCH	1
I. Hồi quy tuyến tính	1
I.1. Đặt vấn đề:	1
I.2. Tổng quan bộ dữ liệu	1
I.3. Quá trình xây dựng mô hình.....	2
I.3.1. Thông các tin cơ bản về bộ dữ liệu.....	2
I.3.2. Cơ sở lựa chọn biến độc lập:	2
I.3.3. Xây dựng mô hình và nhận xét.....	5
I.3.3.1. Chia dữ liệu training và testing:	5
I.3.3.2. Xây dựng mô hình hồi quy đơn biến.....	5
I.3.3.2.1. Mô hình hồi quy đơn biến với biến độc lập là youtube	5
I.3.3.2.2. Mô hình hồi quy đơn biến với biến độc lập là facebook	7
I.3.3.3. Xây dựng mô hình hồi quy đa biến	9
I.4. Kết quả phân tích và dự đoán	10
II. Hồi quy logistic	11
II.1. Hồi quy logistic đơn biến.....	11
II.1.1. Đặt vấn đề.....	11
II.1.2. Tổng quan bộ dữ liệu.....	11
II.1.3. Quá trình xây dựng mô hình.....	12
II.1.4. Kết quả phân tích và dự đoán	16
II.2. Hồi quy logistic đa biến với biến độc lập là biến tuyến tính	16
II.2.1. Đặt vấn đề.....	16
II.2.2. Tổng quan bộ dữ liệu.....	17

II.2.3. Quá trình xây dựng mô hình.....	17
II.2.4 Kết quả phân tích và dự đoán	22
II.3. Hồi quy logistic đa biến với biến độc lập là biến thứ bậc và biến nhị phân.....	23
II.3.1. Đặt vấn đề.....	23
II.3.2. Tổng quan bộ dữ liệu.....	23
II.3.3. Quá trình xây dựng mô hình.....	23
II.3.4 Kết quả phân tích và dự đoán	27
III. Mô Hình Chuỗi Thời Gian	27
III.1. Đặt vấn đề	27
III.2. Mô hình trung bình trượt (MA).....	28
III.2.1 Tổng quan bộ dữ liệu	28
III.2.2 Xây dựng mô hình.....	28
III.2.3 Kết quả phân tích và dự đoán.....	36
III.3. Mô hình ARIMA	38
III.3.1 Tổng quan bộ dữ liệu	38
III.3.2 Xây dựng mô hình.....	38
III.3.3 Kết quả phân tích và dự đoán.....	46
IV. Các mô hình máy học.....	47
IV.1. Đặt vấn đề:	47
IV.2. Tổng quan bộ dữ liệu:	47
IV.3. Quá trình xây dựng mô hình với các giải thuật.....	48
IV.3.1. Thông tin dữ liệu:.....	48
IV.3.3. Xây dựng Decision Tree	49
IV.3.4. Xây dựng Random Forest:	51

IV.3.5. Xây dựng Neural Network:.....	53
IV.4. Kết quả phân tích:	55
PHẦN B. TỔNG KẾT VÀ ĐỊNH HƯỚNG PHÁT TRIỂN:	56
I. Những điều đã làm được:	56
II. Những mặt hạn chế:	56
IV. Định hướng tương lai:	566

PHẦN A. NỘI DUNG BÁO CÁO KẾT QUẢ PHÂN TÍCH

I. Hồi quy tuyến tính

I.1. Đặt vấn đề:

Quảng cáo trên các phương tiện truyền thông mạng xã hội đóng một vai trò rất quan trọng

trong hoạt động kinh doanh của các công ty ngày nay. Nó giúp cho các nhà tiếp thị xây dựng mối quan hệ với khách hàng của họ và tăng doanh số bán hàng.

Các nhà tiếp thị sử dụng phương tiện truyền thông xã hội để quảng cáo sản phẩm của họ và tạo ra doanh số bán hàng. Các trang mạng xã hội như Youtube, Facebook, TikTok, Instagram... rất cần thiết trong các hoạt động kinh doanh cạnh tranh ngày nay để giúp thúc đẩy doanh số bán hàng của công ty.

Như vậy câu hỏi đặt ra ở đây là làm thế nào để dự đoán tác động của quảng cáo trên mạng xã hội đối với doanh số bán hàng của một công ty. Chúng ta có thể xây dựng một mô hình hồi quy tuyến tính để dự đoán doanh số bán hàng dựa trên số tiền chi tiêu cho các hoạt động quảng cáo trên các phương tiện thông tin đại chúng (Social Media Budget) để làm rõ điều này. Để đạt được những mục tiêu trên nhóm thực hiện đề án tiến hành:

- Phân tích mối quan hệ giữa ngân sách quảng cáo trên phương tiện truyền thông và doanh số bán hàng
- Xây dựng mô hình hồi quy tuyến tính bằng cách sử dụng tập dữ liệu training dataset.
- Đưa ra dự đoán doanh số bán hàng dựa trên số tiền chi tiêu quảng cáo trên các kênh Social Media với test data

I.2. Tổng quan bộ dữ liệu

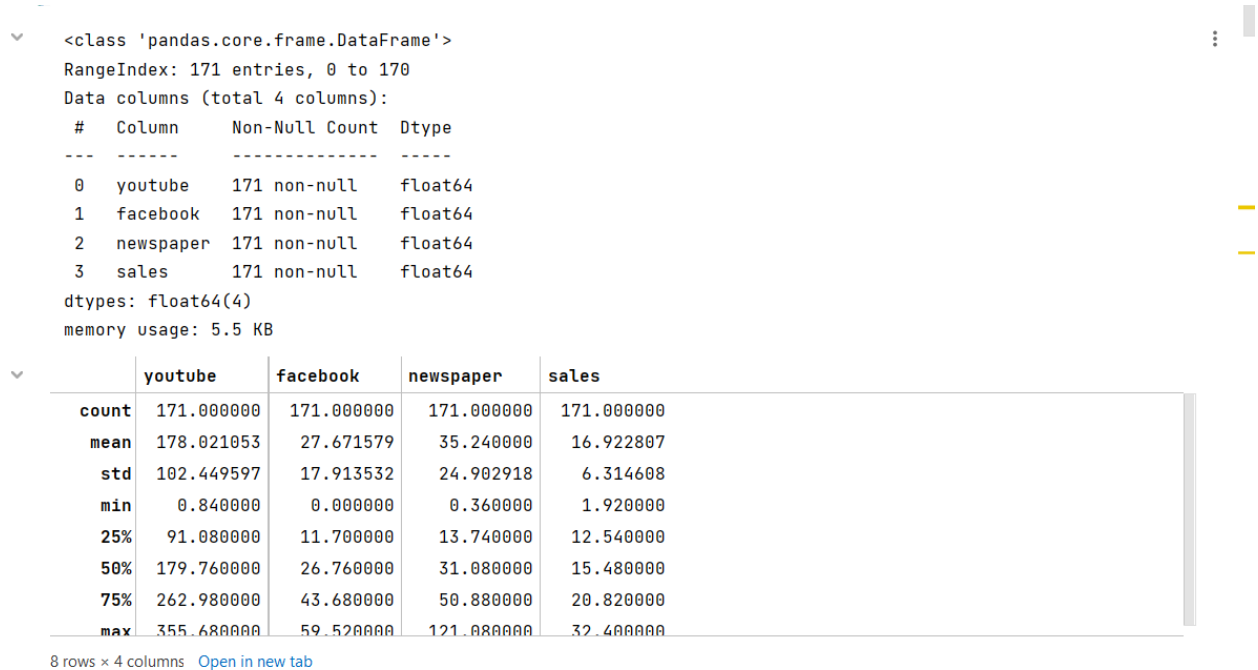
- Tập dữ liệu: Marketing Linear Regression (Marketing_Data.csv) from Kaggle.
- Chuẩn bị dữ liệu:

Tập dữ liệu tiếp thị này bao gồm ảnh hưởng của ba phương tiện quảng cáo (Youtube, Facebook, Newspaper) và doanh số tương ứng (sales)

I.3. Quá trình xây dựng mô hình

I.3.1. Thông các tin cơ bản về bộ dữ liệu

Bắt đầu quá trình phân tích, chúng ta tiến hành xem xét bộ dữ liệu với các thông số cơ bản và kiểm tra dữ liệu đã “sạch” chưa.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 171 entries, 0 to 170
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  -
0   youtube     171 non-null    float64
1   facebook    171 non-null    float64
2   newspaper   171 non-null    float64
3   sales       171 non-null    float64
dtypes: float64(4)
memory usage: 5.5 KB
```

	youtube	facebook	newspaper	sales
count	171.000000	171.000000	171.000000	171.000000
mean	178.021053	27.671579	35.240000	16.922807
std	102.449597	17.913532	24.902918	6.314608
min	0.840000	0.000000	0.360000	1.920000
25%	91.080000	11.700000	13.740000	12.540000
50%	179.760000	26.760000	31.080000	15.480000
75%	262.980000	43.680000	50.880000	20.820000
max	355.680000	59.520000	121.080000	32.400000

8 rows x 4 columns [Open in new tab](#)

Hình 1. Kiểm tra dữ liệu.

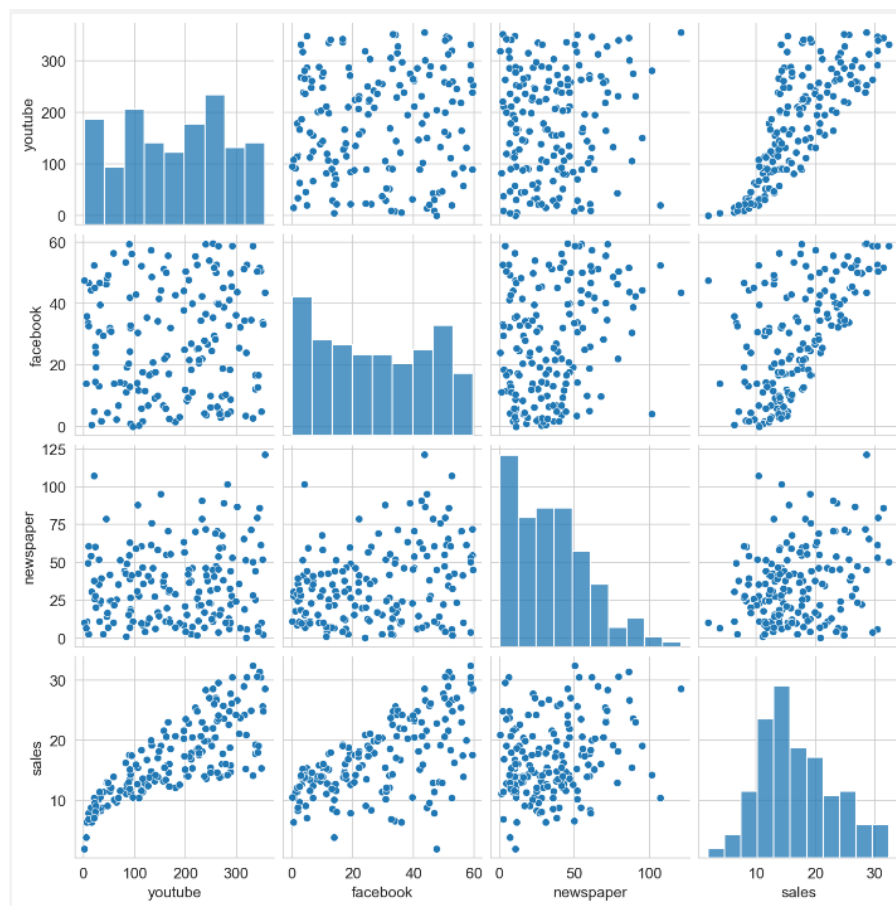
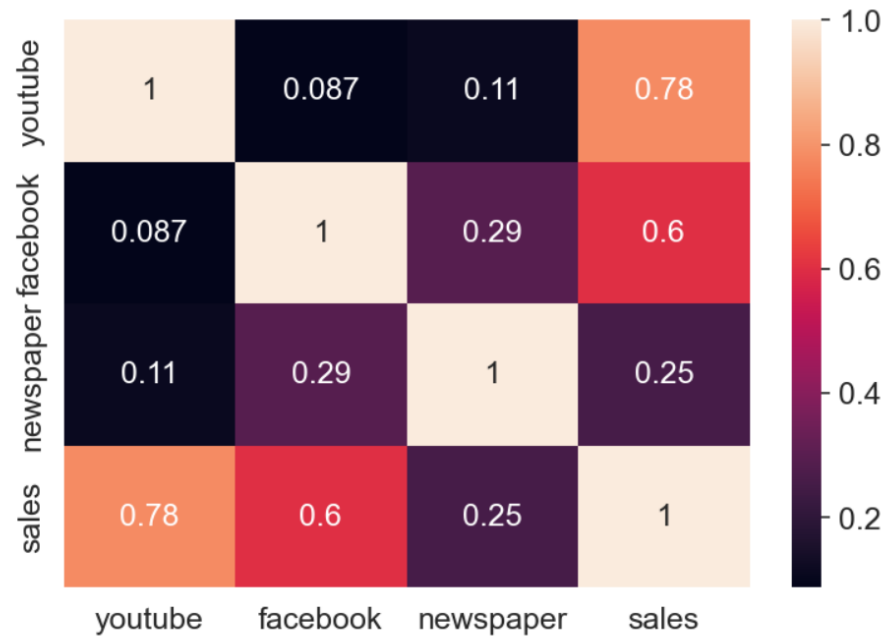
Kết quả cho thấy bộ dữ liệu với 171 dòng và không có dữ liệu rỗng.

I.3.2. Cơ sở lựa chọn biến độc lập:

Sau đó với kết quả từ biểu đồ phân tán và biểu đồ nhiệt (heatmap) của các biến, ta có thể nhận xét được rằng:

- Mức độ tương quan mạnh giữa 2 biến youtube và facebook đối với biến doanh thu (sales) (lần lượt là 0.78, 0.6) và bên cạnh đó biến newspaper có độ tương quan thấp đối với biến doanh thu (0.25). Bên cạnh đó 3 biến facebook, youtube và newspaper thể hiện mức độ tương quan yếu lẫn nhau (các cặp hệ số tương quan đều ở mức thấp, không vượt quá 0.5)
- Khi trực quan hóa những mối tương quan này bằng biểu đồ pairplot cũng nhận thấy những điểm tương tự

- Dựa vào mức độ tương qua giữa các biến trong bộ dữ liệu. Nhóm tiến hành dựa vào cơ sở lý thuyết của mô hình hồi quy tuyến tính đơn biến và đa biến để lựa chọn biến độc lập và biến phụ thuộc của mô hình hồi quy như sau:
 - + Biến phụ thuộc: Doanh thu (sales)
 - + Biến độc lập: youtube và facebook



Hình 2. Trực quan hoá sự tương quan giữa các biến.

I.3.3. Xây dựng mô hình và nhận xét

I.3.3.1. Chia dữ liệu training và testing:

Dữ liệu training và testing sẽ được chia theo tỉ lệ:

Training data: testing data = 80: 20

I.3.3.2. Xây dựng mô hình hồi quy đơn biến

I.3.3.2.1. Mô hình hồi quy đơn biến với biến độc lập là youtube

Xây dựng mô hình hồi quy tuyến tính đơn biến với thư viện statsmodels trên tệp dữ liệu training và thu được kết quả của mô hình hồi quy như sau:

```
OLS Regression Results
=====
Dep. Variable:          y      R-squared:          0.608
Model:                  OLS    Adj. R-squared:       0.605
Method:                 Least Squares    F-statistic:       208.1
Date:                   Wed, 10 Aug 2022    Prob (F-statistic): 4.70e-29
Time:                   16:50:41    Log-Likelihood:    -380.18
No. Observations:       136    AIC:              764.4
Df Residuals:           134    BIC:              770.2
Df Model:                1
Covariance Type:        nonrobust
=====
                    coef    std err          t      P>|t|      [0.025      0.975]
-----
const                8.3196     0.688     12.088     0.000     6.958     9.681
x1                   0.0479     0.003     14.425     0.000     0.041     0.054
=====
Omnibus:                0.035    Durbin-Watson:       2.000
Prob(Omnibus):          0.983    Jarque-Bera (JB):    0.132
Skew:                   -0.033    Prob(JB):            0.936
Kurtosis:                2.863    Cond. No.             417.
=====
```

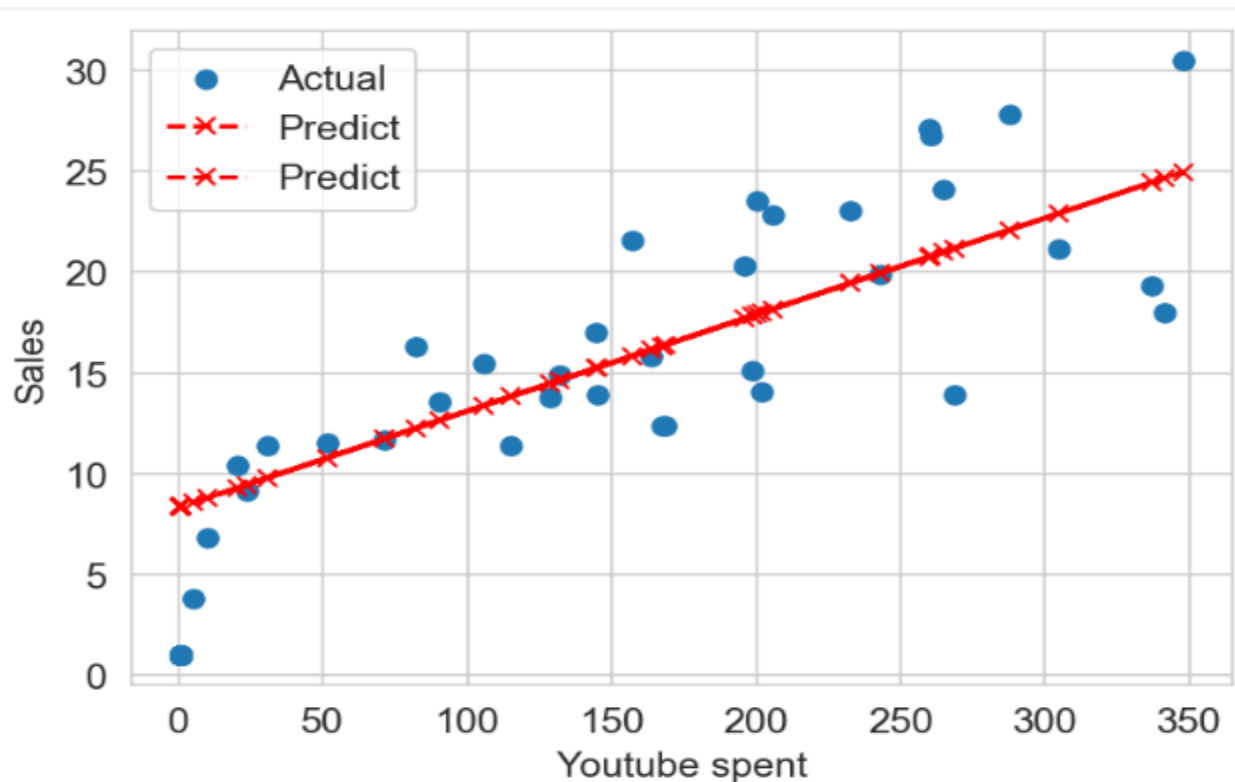
Hình 3. Kết quả thu được từ mô hình hồi quy.

Đối với mô hình hồi quy tuyến tính đơn biến, nhóm tập trung chủ yếu vào các trị thống kê quan trọng và cơ bản nhất.

- Mô hình cho ra kết quả R-squared = 0.608, điều này có nghĩa là biến độc lập youtube giải thích được 60,8% độ phụ thuộc của doanh số (sales) vào ngân sách chi cho quảng cáo trên youtube (youtube).
- Const = 8.3196 và x1 = 0.0479 chính là các hệ số trong phương trình hồi quy tuyến tính. Như vậy ta có thể suy ra phương trình hồi quy như sau:

$$\text{Sales} = 8.3196 + 0.0479 * \text{youtube}$$
- Phương trình này mang ý nghĩa rằng
 - + Khi không có đầu tư vào quảng cáo trên youtube thì mức doanh thu sẽ giữ nguyên ở mức 8.3196
 - + Mỗi khi tăng mức đầu tư vào quảng cáo trên youtube lên 1 đơn vị thì doanh thu sẽ tăng lên 0.0479 đơn vị tiền tệ

Mô hình cho ra R-squared = 0.608 là một kết quả tích cực và có thể tin cậy được để tiến hành dự đoán với tập dữ liệu test và kết quả được trực quan hóa như sau:



Hình 4. Trục quan hoá tương quan giữa Sales và Youtube spent.

I.3.3.2.2. Mô hình hồi quy đơn biến với biến độc lập là facebook

Xây dựng mô hình hồi quy tuyến tính đơn biến với thư viện statsmodels trên tập dữ liệu training và thu được kết quả của mô hình hồi quy như sau:

```

                        OLS Regression Results
=====
Dep. Variable:          y      R-squared:          0.371
Model:                  OLS    Adj. R-squared:      0.366
Method:                 Least Squares    F-statistic:      78.87
Date:                   Wed, 10 Aug 2022    Prob (F-statistic): 3.80e-15
Time:                   11:58:54    Log-Likelihood:    -412.43
No. Observations:      136    AIC:              828.9
Df Residuals:          134    BIC:              834.7
Df Model:               1
Covariance Type:        nonrobust
=====
                        coef    std err          t      P>|t|      [0.025    0.975]
-----
const          11.0880      0.788     14.065     0.000      9.529     12.647
x1              0.2136      0.024      8.881     0.000      0.166      0.261
=====
Omnibus:          19.498    Durbin-Watson:      2.073
Prob(Omnibus):    0.000    Jarque-Bera (JB):    23.576
Skew:             -0.877    Prob(JB):            7.60e-06
Kurtosis:         4.042    Cond. No.            59.6
=====
```

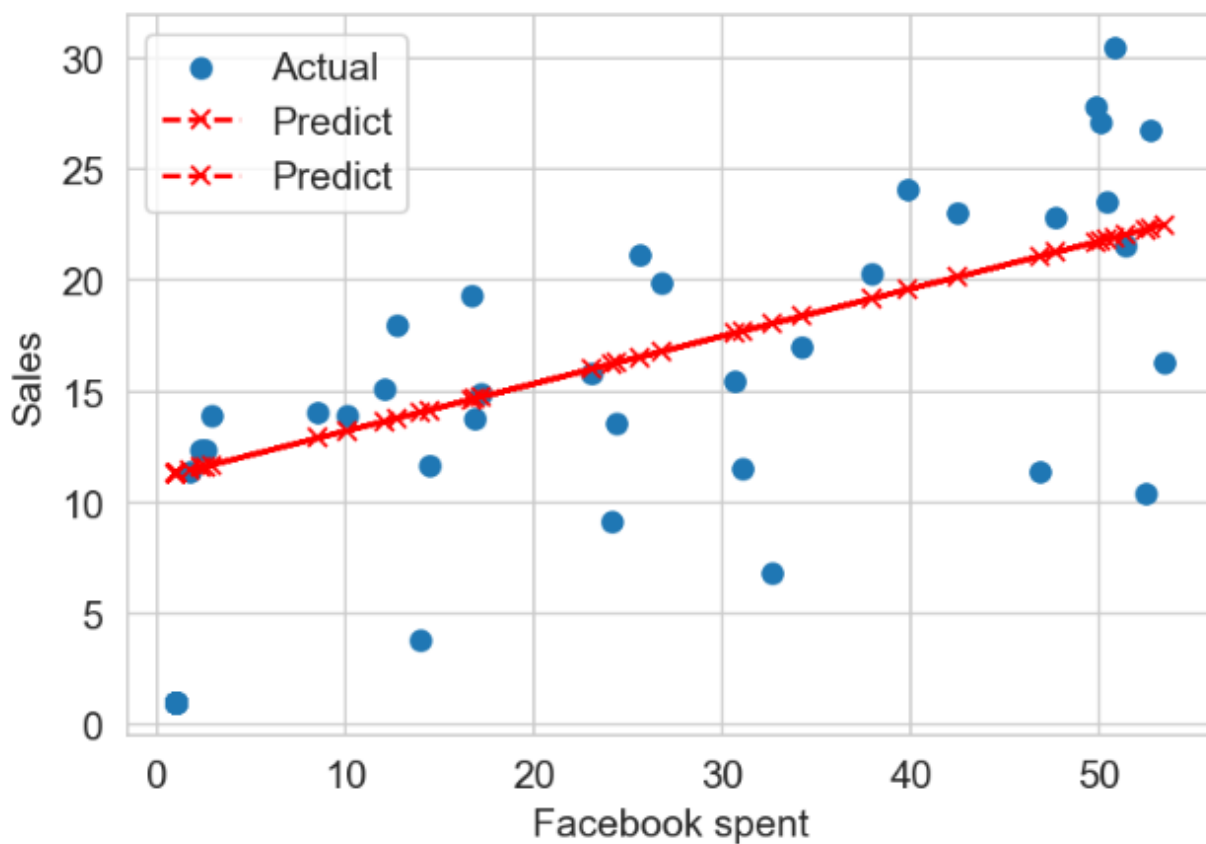
Hình 5. Kết quả thu được từ tập dữ liệu training.

Đối với mô hình hồi quy tuyến tính đơn biến, nhóm tập trung chủ yếu vào các trị thống kê quan trọng và cơ bản nhất.

- Mô hình cho ra kết quả R-squared = 0.371, điều này có nghĩa là biến độc lập facebook giải thích được 37.1% độ phụ thuộc của doanh số (sales) vào ngân sách chi cho quảng cáo trên youtube (youtube).
- Const = 11.0880 và x1 = 0.2136 chính là các hệ số trong phương trình hồi quy tuyến tính. Như vậy ta có thể suy ra phương trình hồi quy như sau:
$$\text{Sales} = 11.0880 + 0.2136 * \text{facebook}$$

- Phương trình này mang ý nghĩa rằng
 - + Khi không có đầu tư vào quảng cáo trên youtube thì mức doanh thu sẽ giữ nguyên ở mức 11.0880
 - + Mỗi khi tăng mức đầu tư vào quảng cáo trên youtube lên 1 đơn vị thì doanh thu sẽ tăng lên 0.2136 đơn vị tiền tệ

Đối với mô hình hồi quy này, $R_squared$ chỉ đạt mức dưới trung bình, mô hình trực quan hóa dưới đây khi áp dụng mô hình với tập dữ liệu test sẽ làm rõ hơn mức độ ảnh hưởng có phần kém hơn so với youtube của biến facebook lên biến phụ thuộc là sales.



Hình 6. Tương quan giữa Sales và Facebook spent.

I.3.3.3. Xây dựng mô hình hồi quy đa biến

Dựa vào nhận định ban đầu rằng hai biến độc lập trong mô hình hồi quy là youtube và facebook thể hiện một độ tương quan lẫn nhau rất thấp (0.087) và cả hai đều có những ảnh hưởng nhất định đến biến doanh thu (sales), nhóm sẽ tiến hành xây dựng mô hình hồi quy tuyến tính đa biến để đánh giá tiếp tục. Kết quả của mô hình hồi quy được diễn tả như trong hình:

OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.896			
Model:	OLS	Adj. R-squared:	0.895			
Method:	Least Squares	F-statistic:	573.5			
Date:	Wed, 10 Aug 2022	Prob (F-statistic):	4.05e-66			
Time:	12:10:27	Log-Likelihood:	-289.94			
No. Observations:	136	AIC:	585.9			
Df Residuals:	133	BIC:	594.6			
Df Model:	2					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	3.7136	0.429	8.653	0.000	2.865	4.562
x1	0.0447	0.002	25.937	0.000	0.041	0.048
x2	0.1891	0.010	19.193	0.000	0.170	0.209
=====						
Omnibus:	55.538	Durbin-Watson:	2.121			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	182.583			
Skew:	-1.523	Prob(JB):	2.25e-40			
Kurtosis:	7.790	Cond. No.	507.			
=====						

Notes:

Hình 7. Kết quả thu được từ mô hình hồi quy tuyến tính đa biến.

Cũng như các mô hình hồi quy tuyến tính đơn biến đã thực hiện ở trên, những chỉ số của mô hình cần đánh giá là:

- Kết quả R-squared = 0.896, điều này có nghĩa là hai biến độc lập youtube, facebook giải thích được 89.6% độ phụ thuộc của doanh số (sales) vào ngân sách chi cho quảng cáo trên youtube và facebook

- $Const = 3.7136$, $x_1 = 0.0447$ và $x_2 = 0.1891$ chính là các hệ số trong phương trình hồi quy tuyến tính. Như vậy ta có thể suy ra phương trình hồi quy như sau:

$$Sales = 3.7136 + 0.0447 * youtube + 0.1891 * facebook$$

Bên cạnh đó đối với mô hình hồi quy tuyến tính đa biến, chỉ số Adjusted R_Squared đóng vai trò quan trọng trong việc xác định mô hình có đang bị tình trạng đa cộng tuyến hay không. Đối với kết quả từ mô hình Adj. R_squared=0.895 cao hơn so với chỉ số này ở các mô hình đơn biến đã thực hiện trước đó, kết hợp với yếu tố ảnh hưởng yếu dựa vào độ tương quan giữa 2 biến độc lập, ta có thể kết luận rằng mô hình không có tình trạng đa cộng tuyến

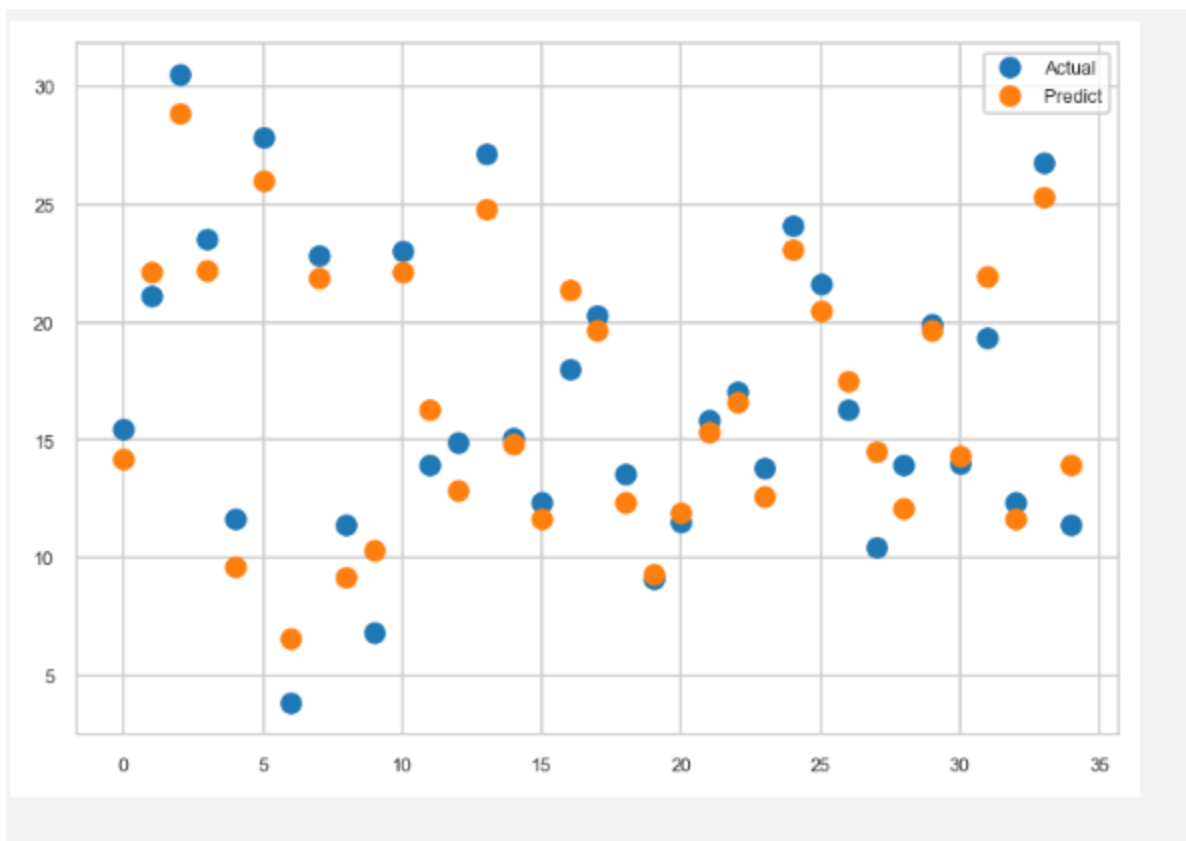
Hơn thế nữa chỉ số Prob (F-statistic) rất nhỏ và tiến về 0 cũng chứng minh rằng cả hai biến độc lập có tác động mạnh mẽ đến doanh thu (sales). Một yếu tố để hỗ trợ ý kiến này đó chính là p-value của cả 2 biến đều là 0, chỉ số này nói lên rằng 0% cơ hội ngân sách cho quảng cáo trên youtube và facebook không có tác động đến doanh thu.

I.4. Kết quả phân tích và dự đoán

Sau khi thực hiện quá trình phân tích với tệp dữ liệu, nhóm nhận thấy rằng trong thời đại chuyển đổi số ngày nay thì các phương thức tiếp thị trên các mạng xã hội đang ngày càng chứng minh vị thế của mình và bỏ xa các loại báo giấy truyền thống.

Hơn thế nữa, mô hình hồi quy tuyến tính đa biến có chỉ số R_squared = 0.896, vượt trội hơn hẳn so với chỉ số này ở mô hình hồi quy tuyến tính đơn biến với các điều kiện khác không đổi và không xảy ra đa cộng tuyến, chứng tỏ rằng khi kết hợp quảng bá truyền thông trên cùng lúc hai nền tảng mạng xã hội lớn nhất hiện nay là Facebook và Youtube thì sẽ có tác động tích cực đến doanh thu.

Và cuối cùng, với mức R_squared cao như vậy, có thể tiến hành tiên lượng với các dữ liệu mới trong tương lai.



Hình 8. Tiên lượng với dữ liệu mới trong tương lai.

Sau đó nhóm đồ án thực hiện lại quy trình phân tích trên RStudio với ngôn ngữ R và cho ra kết quả tương tự.

II. Hồi quy logistic

II.1. Hồi quy logistic đơn biến

II.1.1. Đặt vấn đề

Mỗi năm có hàng trăm, hàng ngàn học sinh thi tuyển vào các trường đại học và tỷ lệ đậu - rớt của học sinh cũng đánh giá được một phần về chất lượng giáo dục. Nếu xem xét mối tương quan giữa điểm thi của Bài kiểm tra tiêu chuẩn của học sinh và khả năng được nhận vào trường, chúng ta có thể dựa vào cơ sở đó xây dựng một mô hình hồi quy Logistic để dự đoán cơ hội đậu vào của một ứng viên tiềm năng khi đăng ký xét tuyển.

II.1.2. Tổng quan bộ dữ liệu

Tập dữ liệu được sử dụng ở đây là MS Admission dataset. Tập dữ liệu có 2 biến trong đó có một biến độc lập (gre) và một biến phụ thuộc (admitted). Bộ dữ liệu này sẽ cho biết

liệu ứng viên có được nhận vào một Trường đại học danh tiếng hay không dựa trên điểm của bài kiểm tra tiêu chuẩn (gre - Graduate Record Examination) của họ. Từ đây chúng ta có thể xây dựng một mô hình hồi quy Logistic với một biến độc lập để tiên lượng/ dự đoán khả năng trúng tuyển của ứng viên.

gre: Điểm bài kiểm tra tiêu chuẩn của học sinh (the student's Graduate Record Examination (GRE) score)

admitted: Biến nhị phân cho biết học sinh có được nhận vào trường hay không (a binary variable describing if the student was admitted into grad school or not)

II.1.3. Quá trình xây dựng mô hình

Mô hình được xây dựng dựa trên ngôn ngữ Python và được tổ chức dưới dạng file Jupyter Notebook.

Bước 1: Import thư viện

```
In [2]: # import thư viện
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Hình 9. Nhập tất cả các thư viện cần thiết.

Bước 2: Đọc dữ liệu, kiểm tra dữ liệu

```
In [2]: # import thư viện
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [3]: # đọc dữ liệu
df = pd.read_csv("../data/ms_admission.csv")
```

```
In [4]: # kiểm tra giá trị null
print(df.isna().sum())
```

```
gre      0
admitted 0
dtype: int64
```

```
In [5]: # mô tả các giá trị mean min max std
print(df.describe())
```

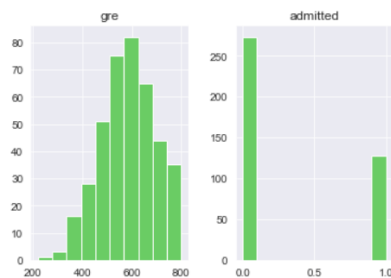
	gre	admitted
count	400.000000	400.000000
mean	587.700000	0.317500
std	115.516536	0.466087
min	220.000000	0.000000
25%	520.000000	0.000000
50%	580.000000	0.000000
75%	660.000000	1.000000
max	800.000000	1.000000

Hình 10. Đọc và kiểm tra dữ liệu.

Kiểm tra dữ liệu có bị null không và kiểm tra tính toàn vẹn của dữ liệu.

Bước 3: Nhận định dữ liệu sơ bộ thông qua biểu đồ histogram.

```
In [21]: # Generate histograms
sns.set_color_codes('muted')
df.hist(color='g')
plt.show()
```



Hình 11. Biểu đồ Histogram nhận định dữ liệu ban đầu.

Với biểu đồ này, có thể thấy điểm số Kiểm tra tiêu chuẩn của học sinh dao động ở mức 500 - 600 điểm. Trong khi số lượng không được nhận vào vượt trội so với số lượng các bạn học sinh được nhận vào trường đại học.

Bước 4: Gán biến độc lập 'gre' cho biến X và biến phụ thuộc 'admitted' cho y đồng thời chia tập dữ liệu huấn luyện (train) và kiểm thử (test) sử dụng phương thức `train_test_split` của Sklearn. Với tỷ lệ 50% tập dữ liệu cho train và 50% tập dữ liệu cho test.

```
In [22]: # gán biến độc lập cho X và biến phụ thuộc cho y
x = df[['gre']]
y = df['admitted']

In [23]: # chia tập dữ liệu huấn luyện và kiểm thử
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=5)
```

Hình 12. Gán biến độc lập và biến phụ thuộc.

Bước 5: Tạo mô hình hồi quy Logistic và fit với Train data

```
In [24]: # tạo mô hình hồi quy
model = LogisticRegression()
model.fit(X_train, y_train)
```

Hình 13. Tạo mô hình hồi quy với train data.

Bước 6: Tiên lượng dự đoán dựa trên mô hình hồi quy Logistic đã xây dựng với tập dữ liệu Test data (X_test) để nhận về kết quả. Kết quả cho thấy các giá trị trả về là 0 (not_admitted) và 1 (admitted).

```
In [25]: # tiên lượng dựa trên tập dữ liệu kiểm thử với 0 là not_admitted và 1 là admitted
y_predictions = model.predict(X_test)
print(y_predictions)

[0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0
 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 1 1 0 0 0 0 1 0 0 0 1 0 0 0 0]
```

Hình 14. Dự đoán kết quả với tập dữ liệu test data.

Bước 7: Vẽ ma trận nhầm lẫn (Confusion Matrix)

```
In [56]: # vẽ ma trận nhầm lẫn sử dụng biểu đồ nhiệt heatmap
print("Confusion Matrix:\n", metrics.confusion_matrix(y_test, y_predictions))
confusion_matrix = confusion_matrix(y_test, y_predictions)
sns.heatmap(confusion_matrix, annot=True, xticklabels=['not_admitted', 'admitted'],
            yticklabels=['not_admitted', 'admitted'])
# sns.heatmap(confusion_matrix, annot=True)
plt.figure(figsize=(3, 3))
plt.show()

Confusion Matrix:
[[134  7]
 [ 49 10]]

<Figure size 216x216 with 0 Axes>
```



Hình 15. Ma trận nhầm lẫn kết hợp biểu đồ nhiệt.

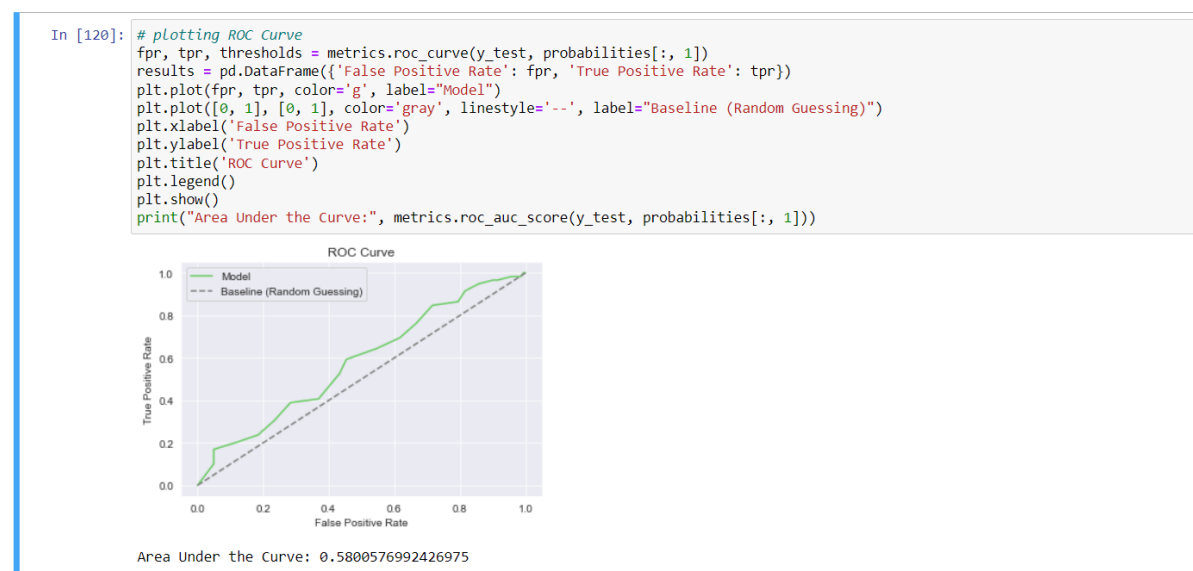
Kết quả cho thấy dựa trên 200 dòng dữ liệu đem đi dự đoán cho ra các giá trị TP = 134, FP = 49, FN = 7 và TN = 10. Theo đó mô hình dự đoán chính xác có 134 học sinh không được nhận vào trường so với 141 học sinh không được nhận vào trong thực tế, tương tự

Độ chính xác của mô hình là 72%, tốt hơn so với mô hình hồi quy Logistic không có phân tách train/test data.

Hình 16. Xem các giá trị xác suất xảy ra.

Hình 17. Classification Report.

Sử dụng đường cong ROC để hình dung rõ ràng hơn về hiệu quả của mô hình.



Hình 18. ROC Curve.

Như biểu đồ trên cho thấy, mặc dù mô hình hồi quy Logistic này không thực sự tốt tuy nhiên nó vẫn tốt hơn so với dự đoán ngẫu nhiên (random guessing) khi nằm trên phía trên bên trái đường Baseline. Phần diện tích bên dưới đường biểu diễn (AUC) được tính là 0.5800577, điều này cho thấy ý nghĩa về mặt giải thích trên thực tế của mô hình là thấp nên không thể áp dụng cho dự đoán thực tế được.

II.1.4. Kết quả phân tích và dự đoán

Có thể thấy mô hình đưa ra dự đoán tỷ lệ được nhận vào là thấp do đó không có tính ứng dụng vào thực tế.

II.2. Hồi quy logistic đa biến với biến độc lập là biến tuyến tính

II.2.1. Đặt vấn đề

Nhận thấy được tiềm năng của hình thức mua hàng trước thanh toán sau đồng thời với đó là việc lãi suất và thủ tục cho vay lại vô cùng đơn giản như thanh toán bằng credit card nên đi kèm với đó là tỉ lệ vỡ nợ của người vay cũng rất tiềm tàng nên để có được những cơ sở dự đoán tốt hơn về việc liệu có nên cho một khách hàng cụ thể nào đó vay với số tiền lớn hơn độ tin cậy mà họ đem lại hay không. Đó là lý do để mô hình dự báo vỡ nợ dưới đây ra đời.

II.2.2. Tổng quan bộ dữ liệu

person_age: Tuổi một người

person_income: Thu nhập của một người

person_home_ownership: Tài sản nhà đất

person_emp_length: Số năm đi làm

loan_intent: Ý định cho vay

loan_grade: cấp bậc cho vay

loan_amnt: Số lượng vay

loan_int_rate: Lãi suất

loan_status: tình trạng vỡ nợ

loan_percent_income: Phần trăm thu nhập

cb_person_default_on_file: Lịch sử tín dụng

cb_person_cred_hist_length: Lịch sử độ dài cho vay

II.2.3. Quá trình xây dựng mô hình

Mô hình được xây dựng trên ngôn ngữ Python và được tổ chức dưới dạng file Jupyter Notebook.

Bước 1: Import thư viện (Tại bước này các thư viện có thể không được import đồng thời tại một thời điểm nhưng các thư viện sẽ được lưu lại trong một block để dễ hơn cho việc cập nhật và thêm sửa)

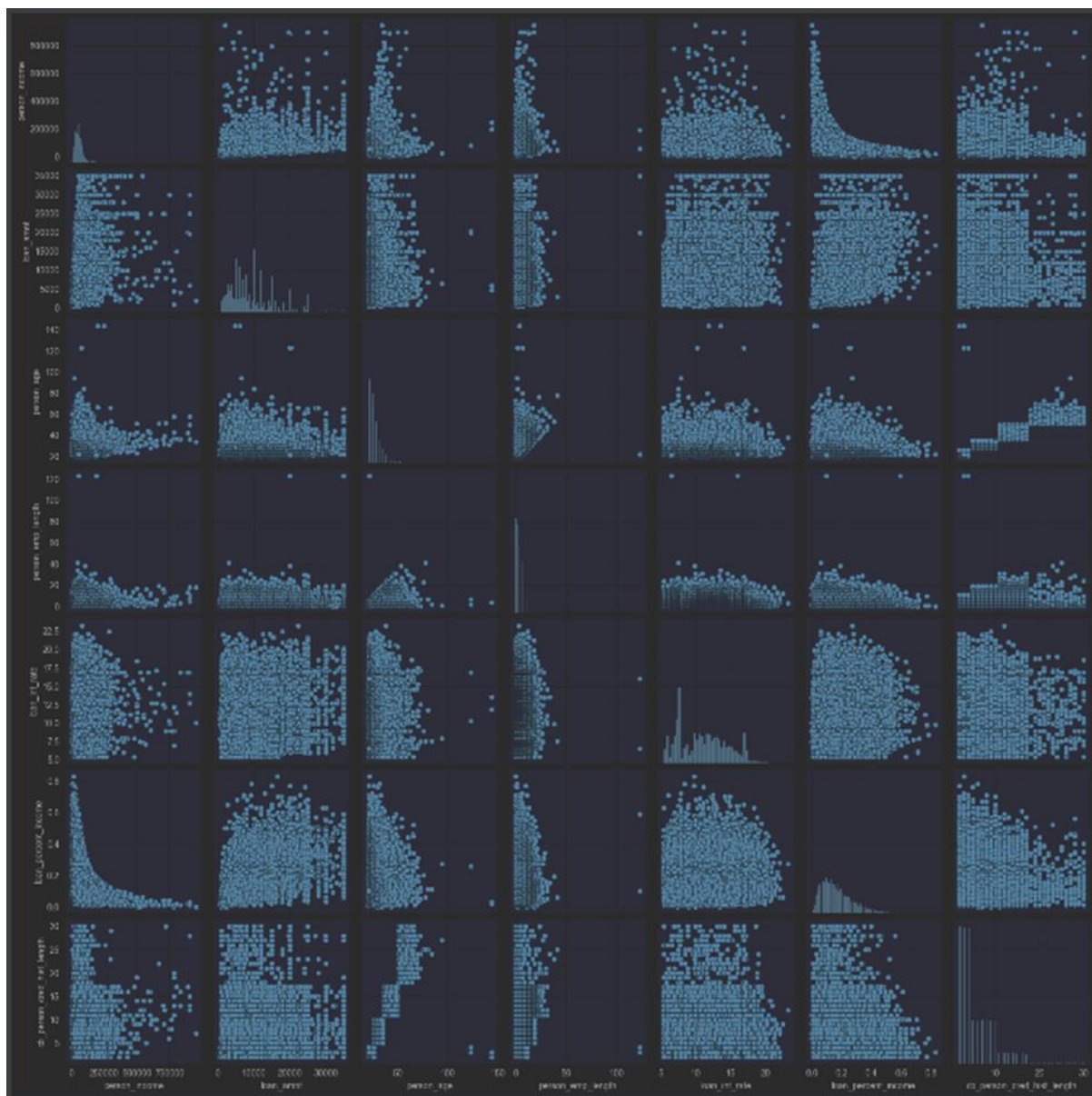
Bước 2: Đọc dữ liệu, kiểm tra dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32572 entries, 0 to 32571
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   person_age                            32572 non-null  int64
1   person_income                         32572 non-null  int64
2   person_home_ownership                 32572 non-null  object
3   person_emp_length                     32572 non-null  int64
4   loan_intent                           32572 non-null  object
5   loan_grade                           32572 non-null  object
6   loan_amnt                             32572 non-null  int64
7   loan_int_rate                         32572 non-null  float64
8   loan_status                           32572 non-null  int64
9   loan_percent_income                   32572 non-null  float64
10  cb_person_default_on_file             32572 non-null  object
11  cb_person_cred_hist_length            32572 non-null  int64
dtypes: float64(2), int64(6), object(4)
```

Hình 19. Kiểm tra tính toàn vẹn.

Kết quả thu được cho thấy dữ liệu toàn vẹn không xảy ra hiện tượng thiếu hụt dữ liệu

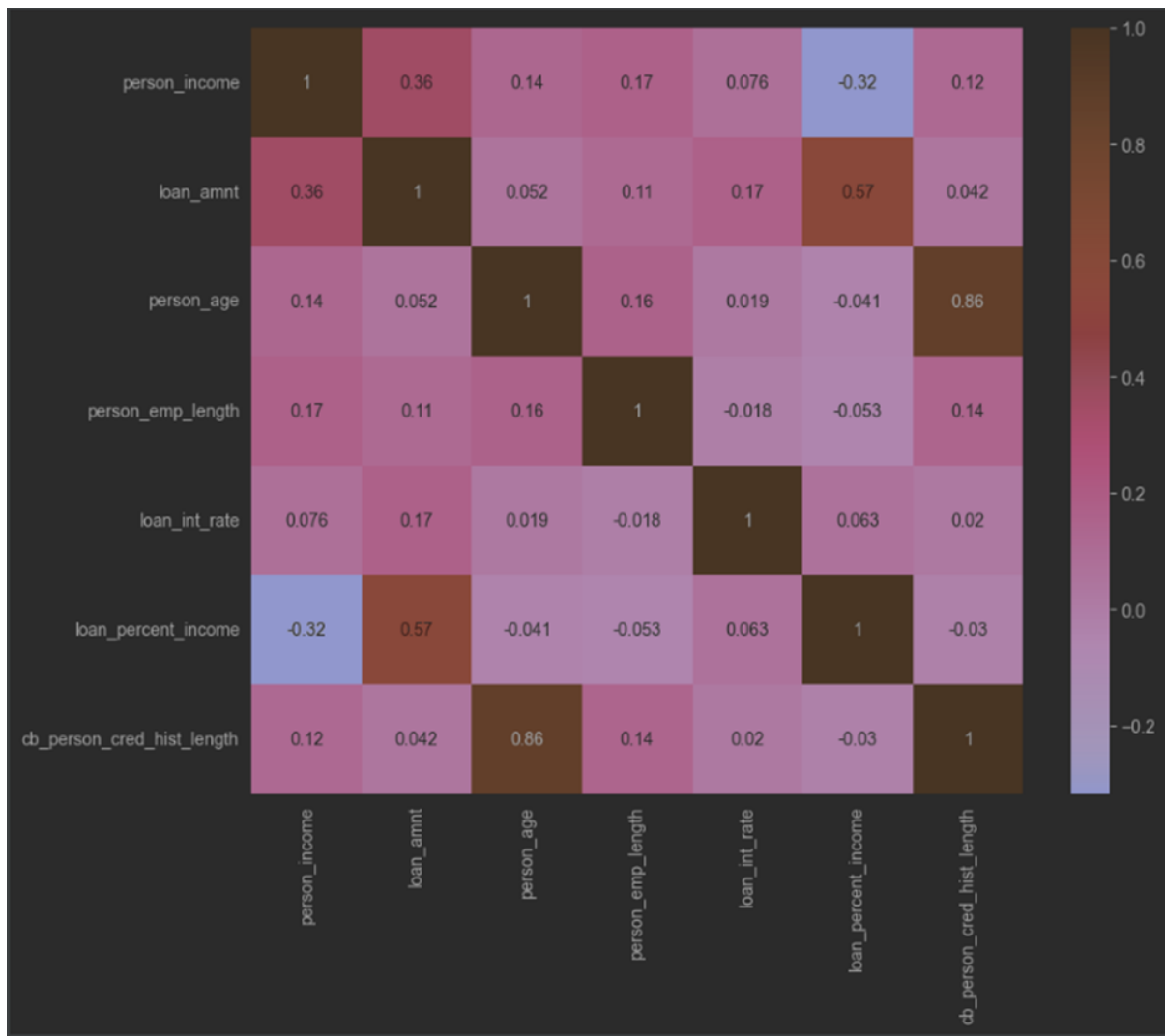
Bước 3: Trực quan hóa dữ liệu, nhận định dữ liệu sơ bộ. Nhóm đã dùng thư viện seaborn và biểu đồ pairplot để biểu thị sự tương quan giữa các cột với nhau.



Hình 20. Biểu đồ pairplot.

Qua biểu đồ trên có thể nhìn thấy các cột hầu như không có sự tương quan nhiều với nhau chỉ có cột `person_age` và `cb_person_cred_hist_length` có sự ảnh hưởng lẫn nhau theo một tỉ lệ nhất định.

Bước 4: Xác minh đa cộng tuyến. Bước này nhằm xác định lại một lần nữa xem có các biến xảy ra tình trạng đa cộng tuyến với nhau hay không



Hình 21. Biểu đồ nhiệt kiểm tra đa cộng tuyến.

Qua kết quả thu được có thể thấy cột person_age và cb_person_cred_hist_length có chỉ số tương quan cao, kết hợp với dự đoán từ biểu đồ trực quan hóa pairplot có thể kết luận 2 cột trên xảy ra hiện tượng đa cộng tuyến. Vì vậy loại bỏ 2 cột này.

Bước 5: Chia train, test data. Đây là công đoạn chia dữ liệu thành tập xây dựng mô hình và tập kiểm thử. Tỷ lệ 9:1 (train: test)

Bước 6: Tính chỉ số intercept và các giá trị coef

```
0.8044620317936823
[[-4.09802950e-05  1.08009487e-04 -8.92013229e-08  2.14009960e-07
  3.03064961e-09]]
```

Hình 22. Chỉ số intercept và giá trị coef.

Tính được các chỉ số intercept và coef như trên giá trị cho từng biến lần lượt là person_income, loan_amnt, person_emp_length, loan_int_rate, loan_percent_income

Bước 7: Tính confusion matrix để đánh giá độ hiệu quả của mô hình

	precision	recall	f1-score	support
0	0.81	0.98	0.89	22910
1	0.73	0.17	0.27	6404
accuracy			0.80	29314
macro avg	0.77	0.57	0.58	29314
weighted avg	0.79	0.80	0.75	29314

Hình 23. Độ chính xác của mô hình là 80%.

Kết quả từ bảng confusion matrix thu được như trên hình, có thể thấy tỉ lệ Recall của việc dự báo khả năng không vỡ nợ và vỡ nợ có sự chênh lệch rất lớn. Trong khi độ chính xác khi xác định tỉ lệ không vỡ nợ lên đến 0.98 thì dự báo khả năng vỡ nợ lại chỉ đạt được 0.17 là thông số có sự tin cậy cực kỳ thấp trong dự báo. Điều này dẫn đến việc dự báo khả năng vỡ nợ đối với mô hình này gần như là không thể.

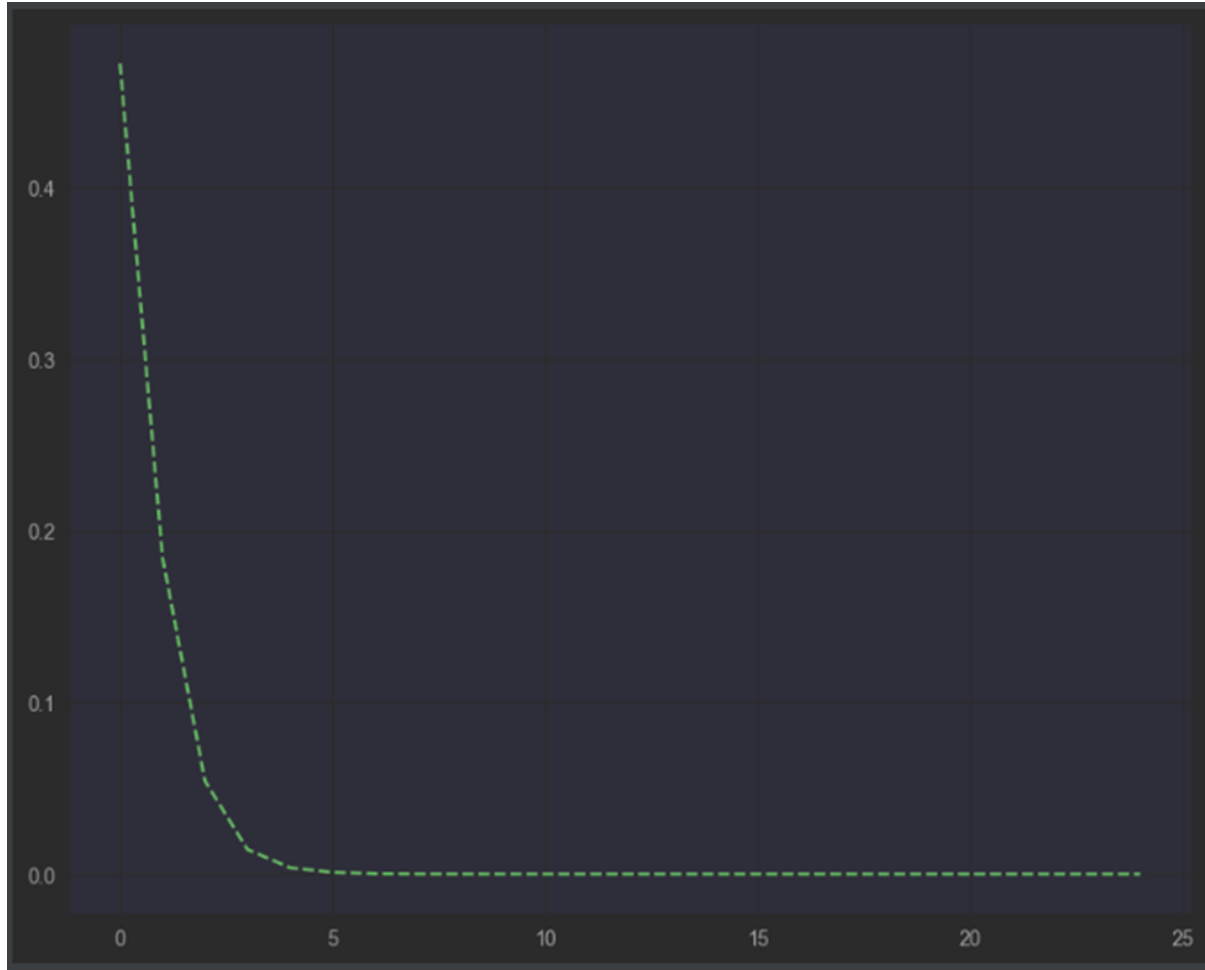
Bước 8: Tiến hành thực nghiệm dự báo trên bộ test data

```
[[0.63624785 0.36375215]
 [0.59033008 0.40966992]
 [0.34271376 0.65728624]
 ...
 [0.91211112 0.08788888]
 [0.83542887 0.16457113]
 [0.87031496 0.12968504]]
```

Hình 24. Kết quả dự báo trên dữ liệu test data.

Kết quả trên có thể thấy như dòng đầu tiên tỉ lệ vỡ nợ của người này là 36,4% và có 63,6% đảm bảo người này sẽ trả nợ đúng hạn.

Bước 9: Tiến hành tạo hàm sigmoid và trực quan hóa đường hồi quy của mô hình



Hình 25. Đường hồi quy của mô hình dự báo.

II.2.4 Kết quả phân tích và dự đoán

Vì chỉ số trung bình điều hòa F1-score khi dự báo tỷ lệ vỡ nợ là chưa cao nên mô hình gần như không có khả năng ứng dụng vào thực tế. Cần có những phương pháp kiểm thử để xác định về việc cần thêm biến hay loại bỏ bớt biến độc lập trên phương trình hồi quy từ đó xây dựng mô hình có độ tin cậy cao hơn trong quá trình phát triển tương lai.

II.3. Hồi quy logistic đa biến với biến độc lập là biến thứ bậc và biến nhị phân

II.3.1. Đặt vấn đề

Trong nền kinh tế hiện nay, việc tìm kiếm được những nhân viên giỏi là rất quan trọng, không chỉ vậy khi đã chiêu mộ được nhân viên tốt công ty cần phải làm thêm những gì để họ có thể phát huy hết tiềm năng của mình và những yếu tố nào ảnh hưởng đến năng suất làm việc của một nhân viên cũng là điều hết sức quan trọng. Vì vậy, nhóm đã tiến hành thực hiện một mô hình dự báo về khả năng nghỉ việc của nhân viên dựa trên các yếu tố về mức độ hài lòng của họ cũng như tình trạng thăng tiến và mức lương thì sẽ ảnh hưởng thế nào đến với nhân viên. Từ đó tạo nên mô hình dự báo giúp cho các nhà quản trị có thể dựa vào đó để phát huy được hết tiềm lực của một nhân viên cũng như tránh khỏi tình trạng nhân viên giỏi nghỉ việc ngoài mong muốn của công ty.

II.3.2. Tổng quan bộ dữ liệu

satisfacion_level: Mô tả về mức độ hài lòng của nhân viên

last_evaluation: Mức độ hài lòng cuối cùng của nhân viên

number_project: Số dự án đã thực hiện

average_monthly_hours: Số giờ làm trong tháng

time_spend_company: Số năm làm trong công ty

Work_accident: Có gặp tai nạn lao động hay không (0: Không; 1: Có)

Left: Nhân viên đã nghỉ việc hay chưa (0: Chưa nghỉ; 1: Nghỉ việc)

promotion_last_5years: Có thăng tiến trong 5 năm qua hay không (0: Không; 1: Có)

dept: Phòng ban làm việc

salary: Mức lương (Low; Medium, High)

II.3.3. Quá trình xây dựng mô hình

Mô hình được xây dựng trên ngôn ngữ Python và được tổ chức dưới dạng file Jupyter Notebook.

Bước 1: Import thư viện (Tại bước này các thư viện có thể không được import đồng thời tại một thời điểm nhưng các thư viện sẽ được lưu lại trong một block để dễ hơn cho việc cập nhật và thêm sửa)

Bước 2: Đọc dữ liệu, kiểm tra dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   satisfaction_level      14999 non-null  float64
1   last_evaluation         14999 non-null  float64
2   number_project          14999 non-null  int64
3   average_monthly_hours  14999 non-null  int64
4   time_spend_company      14999 non-null  int64
5   Work_accident           14999 non-null  int64
6   left                   14999 non-null  int64
7   promotion_last_5years  14999 non-null  int64
8   sales                   14999 non-null  object
9   salary                  14999 non-null  object
dtypes: float64(2), int64(6), object(2)
```

Hình 26. Kiểm tra toàn vẹn dữ liệu.

Kết quả thu được cho thấy dữ liệu toàn vẹn không xảy ra hiện tượng thiếu hụt dữ liệu

Bước 3: Chọn biến.

Vì mô hình chỉ được xây dựng trên mức cơ bản về độ hài lòng, lương bổng và khả năng thăng tiến nên sẽ chỉ sử dụng các cột `satisfaction_level`, `last_evaluation`, `promotion_last_5years` và `salary` để phục vụ cho việc xây dựng mô hình dự báo.

Bước 4: Mã hóa cột salary.

Vì dữ liệu cột `salary` ở dạng chuỗi ký tự nên cần phải mã hóa về dạng số để xây dựng mô hình trực quan. Các giá trị theo thứ bậc tăng dần từ `low`, `medium`, `high` sẽ được mã hóa thành 1, 2, 3

```
0      1
1      2
2      2
3      1
4      1
..
14994  1
14995  1
14996  1
14997  1
14998  1
```

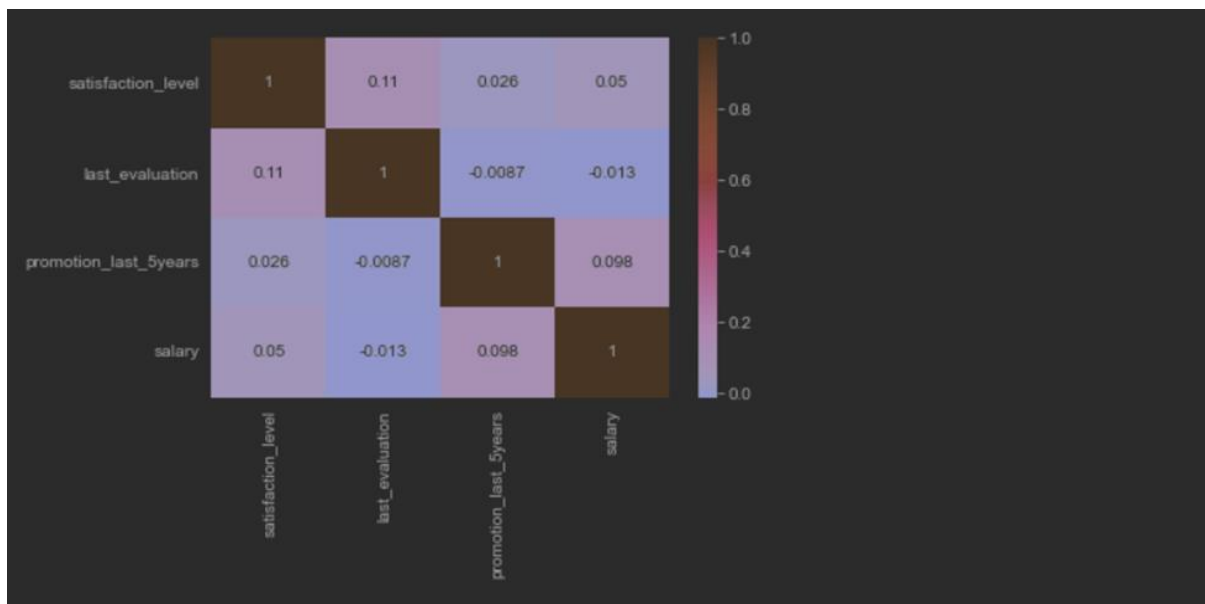
Hình 27. Mã hoá cột Salary.

Bước 5: Chuyển những cột chứa dữ liệu có giá trị quá lớn về chuỗi giá trị từ 0 đến 1

```
0      0.0
1      0.5
2      0.5
3      0.0
4      0.0
...
14994  0.0
14995  0.0
14996  0.0
14997  0.0
14998  0.0
```

Hình 28. Chuyển về giá trị từ 0 đến 1.

Bước 6: Kiểm tra hiện tượng đa cộng tuyến



Hình 29. Biểu đồ nhiệt heatmap.

Từ kết quả thu được có thể thấy các biến độc lập đã chọn không xảy ra hiện tượng đa cộng tuyến.

Bước 7: Chia train, test data. Đây là công đoạn chia dữ liệu thành tập xây dựng mô hình và tập kiểm thử. Tỷ lệ 9:1 (train: test)

Bước 8: Tính các chỉ số intercept, coef

```
0.7795392251277873
[[-3.89708223  0.51987487 -1.19474258 -1.28141713]]
```

Hình 30. Giá trị Intercept và Coef.

Khi ra được kết quả intercept và coef, có thể thấy được các giá trị coef lần lượt ứng với các biến `satisfaction_level`, `last_evaluation`, `promotion_last_5years` và `salary`.

Bước 9: Xây dựng confusion matrix

	precision	recall	f1-score	support
0	0.80	0.94	0.87	10285
1	0.58	0.27	0.37	3214
accuracy			0.78	13499
macro avg	0.69	0.60	0.62	13499
weighted avg	0.75	0.78	0.75	13499

Hình 31. Độ chính xác của mô hình là 78%.

Tương tự như mô hình hồi quy logistic đa biến có thể thấy được độ hiệu quả của mô hình là chưa cao

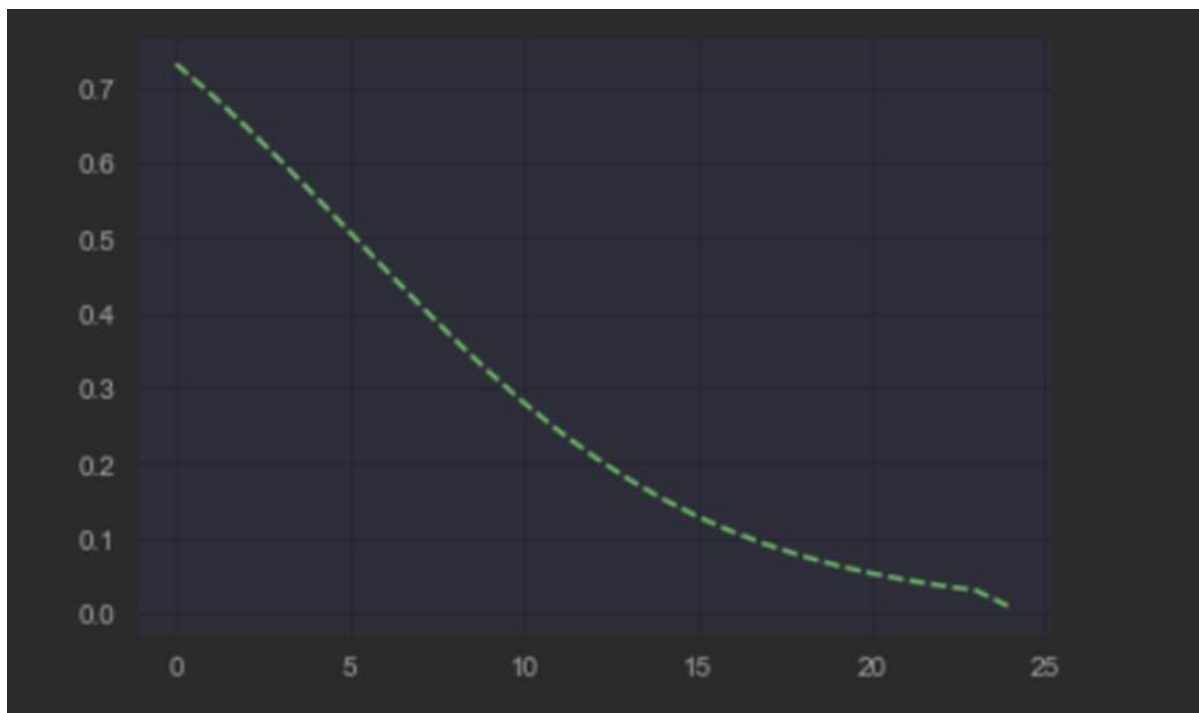
Bước 10: Tiến hành thực nghiệm dự báo trên bộ test data

```
[[0.32304015 0.67695985]
 [0.82046654 0.17953346]
 [0.77825637 0.22174363]
 ...
 [0.25792133 0.74207867]
 [0.66054715 0.33945285]
 [0.71138865 0.28861135]]
```

Hình 32. Kết quả tiên lượng thu được với test data.

Kết quả thu được như hàng đầu tiên cho thấy chỉ số tác động đến nhân viên thứ nhất sẽ khiến nhân viên này có khả nghỉ việc rất cao là 67,7% và tỉ lệ tiếp tục làm việc với công ty chỉ đạt mức dự báo 32,3%. Vì vậy cần có biện pháp để cải thiện mong muốn làm việc của nhân viên này.

Bước 11: Tạo hàm sigmoid và trực quan hóa đường hồi quy



Hình 33. Đường hồi quy của mô hình .

II.3.4 Kết quả phân tích và dự đoán

Mô hình hồi quy trên chưa đạt độ chính xác cao chưa đủ cơ sở để có thể đưa vào dự báo thực tế. Tuy nhiên mô hình mới chỉ ở bước cơ bản có thể tiếp tục phát triển nghiên cứu dựa trên những yếu tố còn lại chưa khai thác trong bộ dữ liệu để kiểm chứng xem những yếu tố còn lại có giúp làm tăng độ chính xác của mô hình lên không từ đây giúp mô hình có tính ứng dụng cao hơn.

III. Mô Hình Chuỗi Thời Gian

III.1. Đặt vấn đề

2019 – 2020 – 2021, 3 năm biến động do sự bùng phát của dịch Covid 19, kéo theo đó là thách thức chưa từng có đến sự phát triển kinh tế toàn cầu. Giống như một vụ nổ BIGBANG, dịch bệnh xảy đến vượt qua mọi suy đoán của nhiều người. Tác động tiêu cực đến kinh tế toàn cầu. Tuy nhiên, để nói 2020 là năm khủng hoảng kinh tế, không hẳn là vậy, đánh giá trên mức độ nào đó, 2020 là một năm suy thoái (chưa đến mức khủng hoảng) đối với kinh tế thế giới bởi những dự đoán về chu kỳ khủng hoảng đã được phân tích trước đó, những biện pháp phòng tránh khủng hoảng cũng được tổ chức kinh tế thế giới lên kế hoạch thông qua kinh nghiệm từ những cuộc khủng hoảng trước.

Vậy để dự báo khủng hoảng, những yếu tố nào được đem ra làm tiêu chí để đánh giá và dự đoán. Lãi suất, mức độ chi tiêu, chính sách tiền tệ,... Tất cả đều đúng cả, tuy nhiên đánh giá vĩ mô và trực quan nhất thì nhóm lựa chọn 2 tiêu chí để dự báo khủng hoảng: Chỉ số S&P 500 và Tỷ lệ toàn dụng lao động từ năm 1986 đến 2020

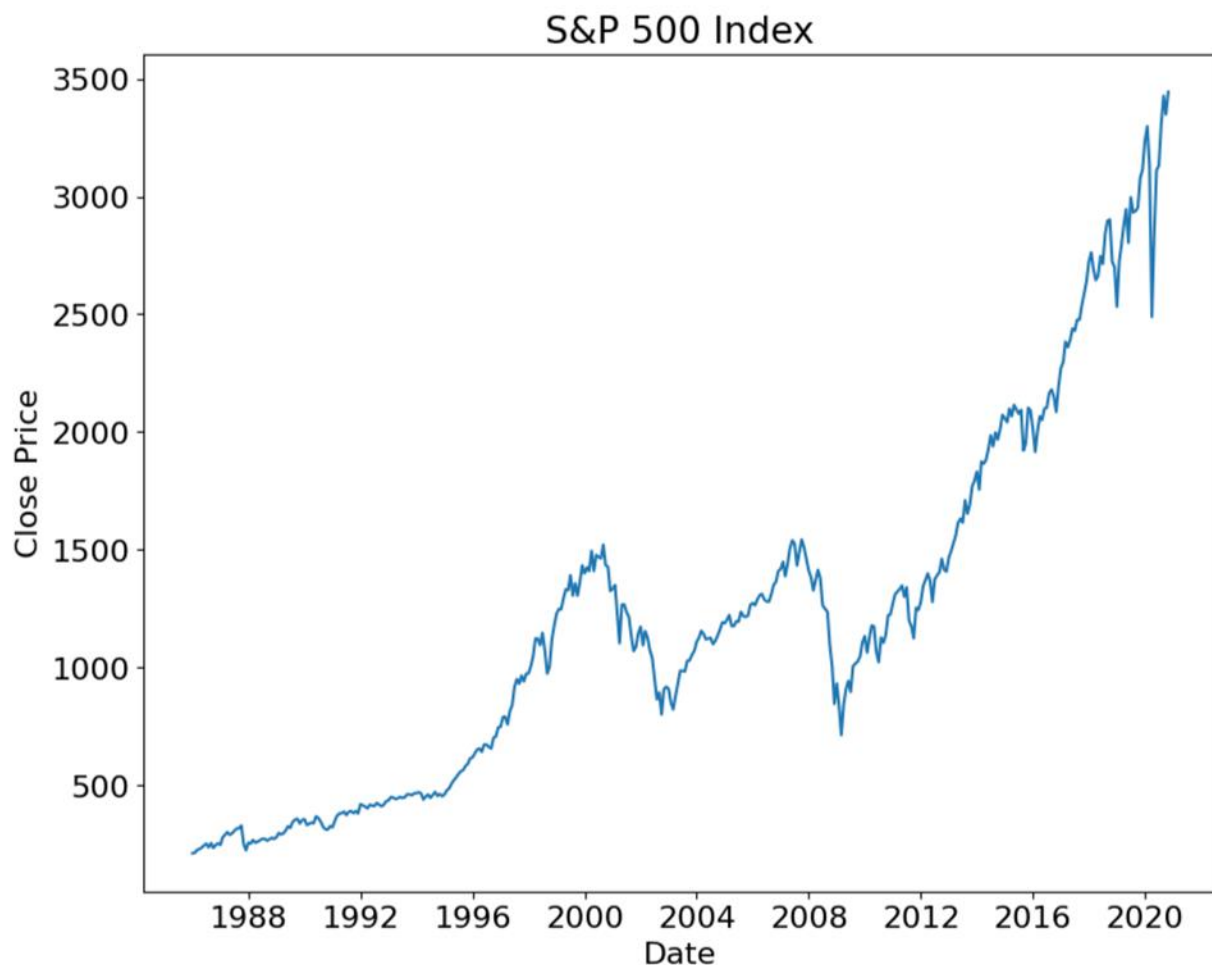
III.2. Mô hình trung bình trượt (MA)

III.2.1 Tổng quan bộ dữ liệu

Dataset gồm 2 cột: thời gian và chỉ số S&P 500 được ghi nhận theo mỗi tháng, tính từ tháng 1/1986 đến tháng 11/2020

III.2.2 Xây dựng mô hình

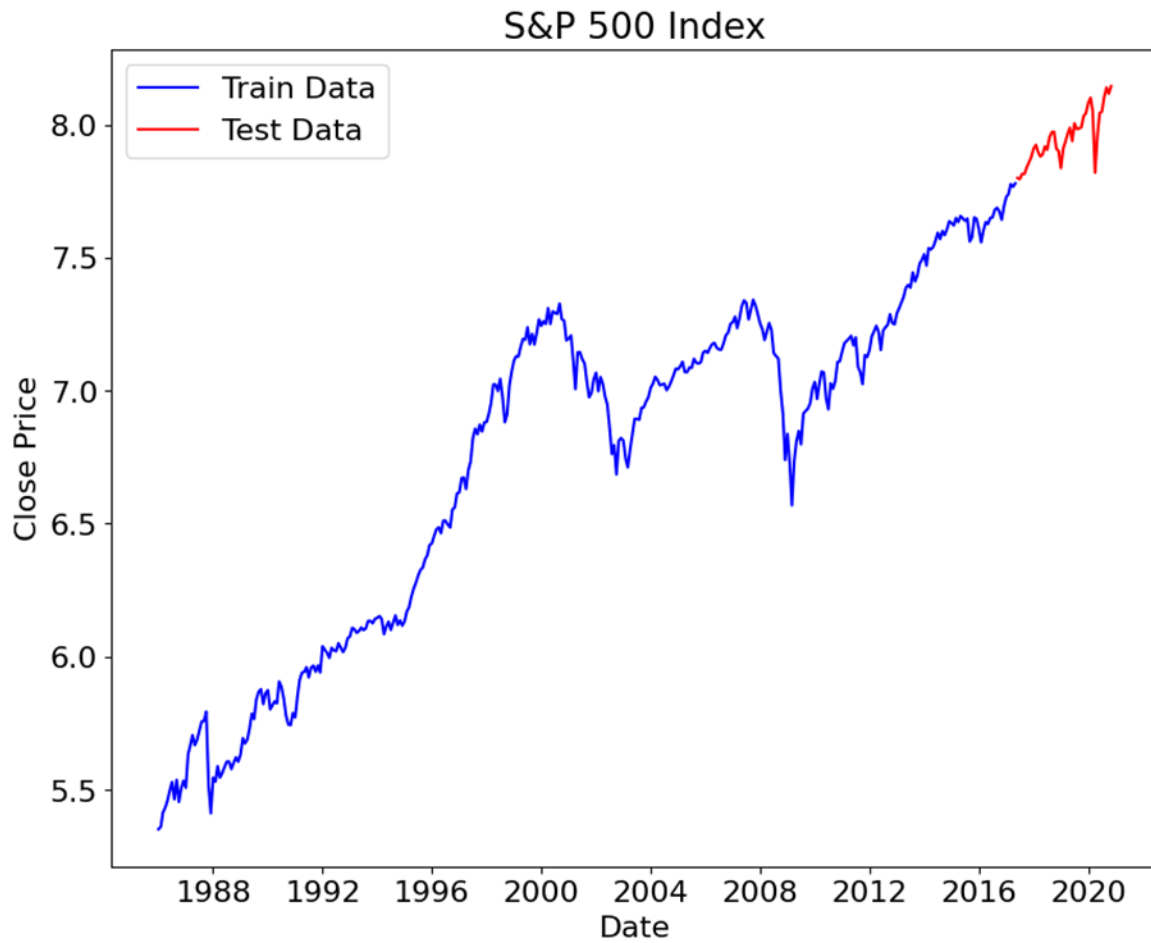
- Tách dữ liệu train và test
- Dữ liệu thực tế



Hình 34. Dữ liệu thực tế của chỉ số S&P 500.

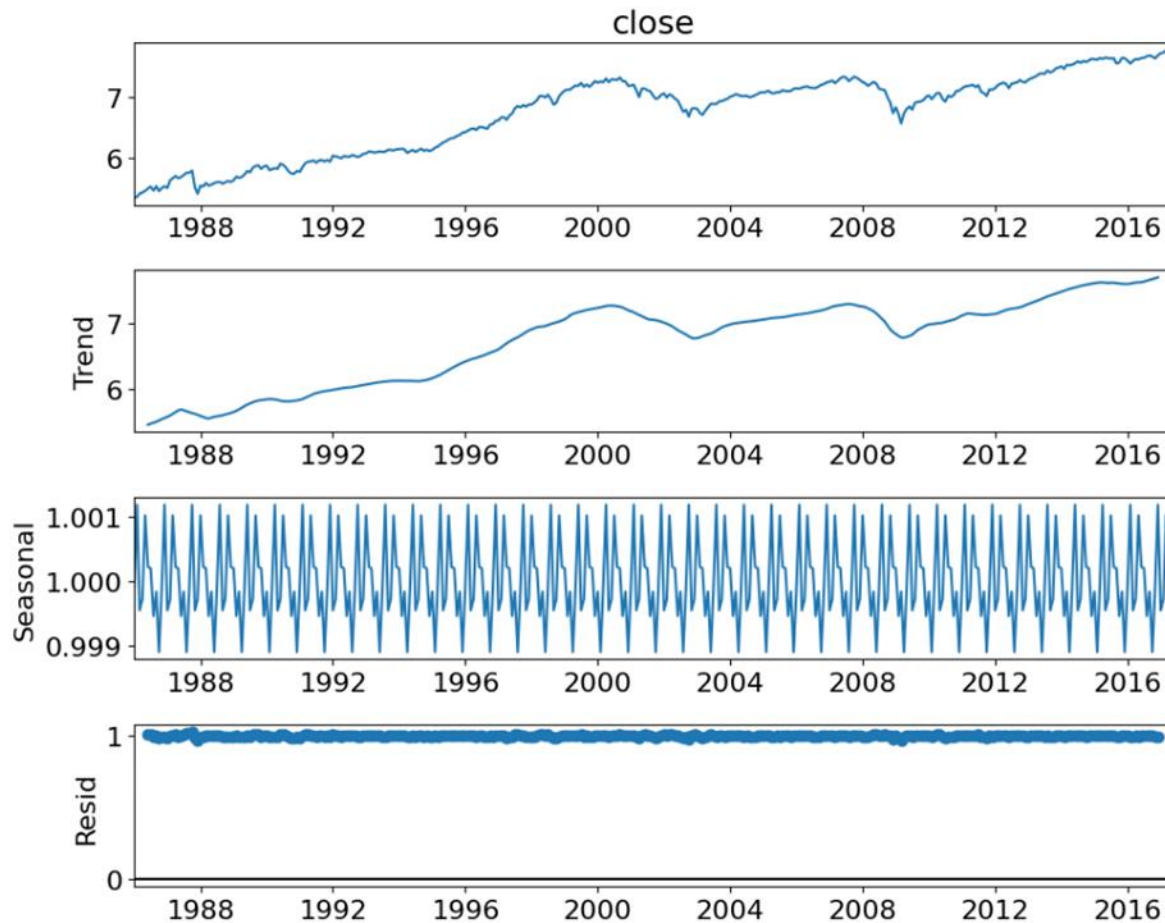
- Dữ liệu sau khi chia train và test data theo tỷ lệ train:test = 9:1

```
train_data, test_data = df_close[:int(len(df_close)*0.9)], df_close[int(len(df_close)*0.9):]
```



Hình 35. Dữ liệu phân tách train test.

- Phân rã chuỗi thời gian



Hình 36. Phân rã chuỗi thời gian.

- Xét trên tập dữ liệu train, nhìn chung dữ liệu có xu hướng tăng, từ năm 2020 đến 2009, dữ liệu có xu hướng biến động theo phương ngang (dao động sideways) và lại trở lại xu hướng tăng từ năm 2010 trở đi
- Xu hướng di chuyển của các đường trung bình 10, 20, 50

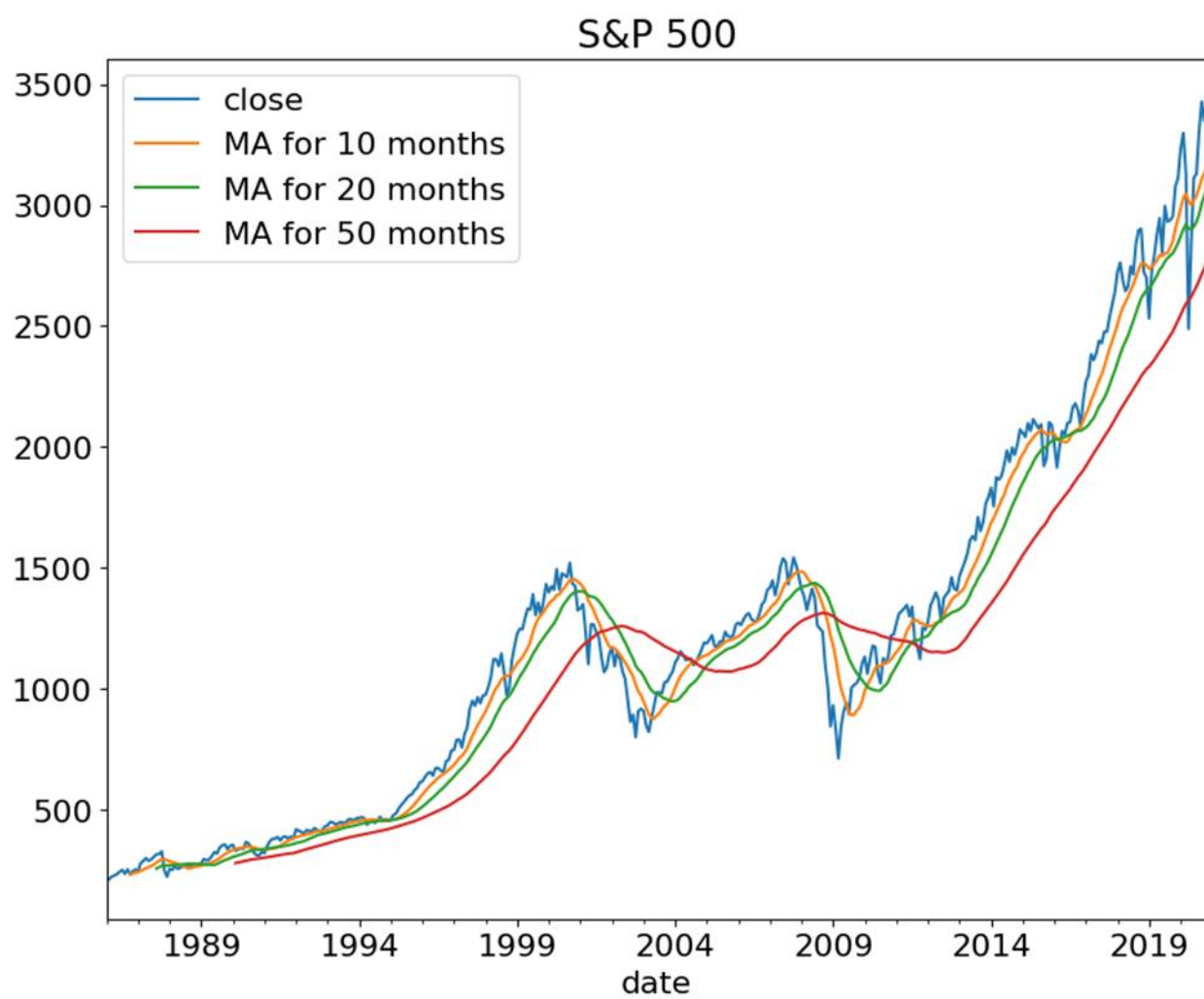
```
ma_month = [10, 20, 50]
```

```
for ma in ma_month:
```

```
    column_name = f"MA for {ma} months"
```

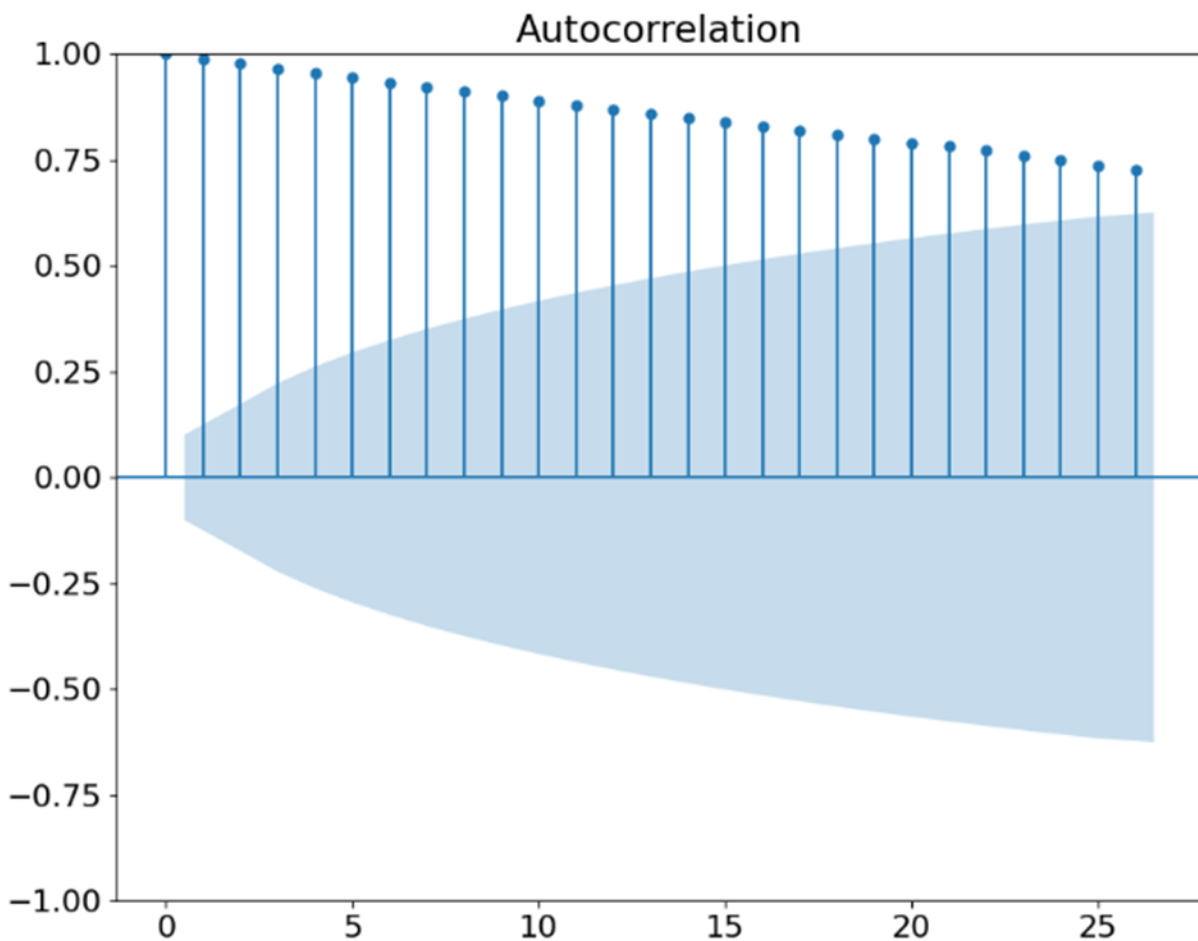
```
    df[column_name] = df['close'].rolling(ma).mean()
```

```
df[['close', 'MA for 10 months', 'MA for 20 months', 'MA for 50 months']].plot().set_title('S&P 500')
```



Hình 37. Trực quan hoá các đường trung bình.

- Kiểm định mô hình



Hình 38. Kiểm định tự tương quan.

- Kiểm định mô hình ACF

Nhóm muốn kiểm định sự tương quan của chỉ số index hiện tại so với index tại các thời điểm trong quá khứ và ACF là mô hình đáp ứng điều đó.

Dựa vào mô hình ACF, ta thấy tại hầu hết các thời điểm trong quá khứ đều có sự tương quan mạnh (tương quan dương) so với chỉ số index hiện tại, nhưng có xu hướng giảm dần.

- Kiểm định ADF

```
print(adf_test(train_data))
```

Giả thuyết kiểm định ADF, H_0 là chuỗi dữ liệu không dừng. Trị tuyệt đối giá trị kiểm định (1,408) nhỏ hơn giá trị tuyệt đối các giá trị tới hạn nên chấp nhận giả thuyết H_0 , chuỗi không dừng

- Xây dựng mô hình MA

Chọn hệ số $q = 1$ tương ứng mô hình MA(1) làm ví dụ

- Xây dựng hàm tìm theta

```
def lag_view(x, order):  
    y = x.copy()  
    x = np.array([y[-(i + order):][:order] for i in range(y.shape[0])])  
    x = np.stack(x)[::-1][order - 1: -1]  
    y = y[order:]  
  
    return x, y
```

```
def ma_process(eps, theta):  
    theta = np.array([1] + list(theta))[::-1][:, None]  
    eps_q, _ = lag_view(eps, len(theta))  
    return eps_q @ theta
```


- Xây dựng mô hình

SARIMAX Results						
=====						
Dep. Variable:	close	No. Observations:	377			
Model:	ARIMA(0, 0, 1)	Log Likelihood	-133.452			
Date:	Wed, 10 Aug 2022	AIC	272.903			
Time:	18:31:08	BIC	284.700			
Sample:	01-01-1986	HQIC	277.585			
	- 05-01-2017					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

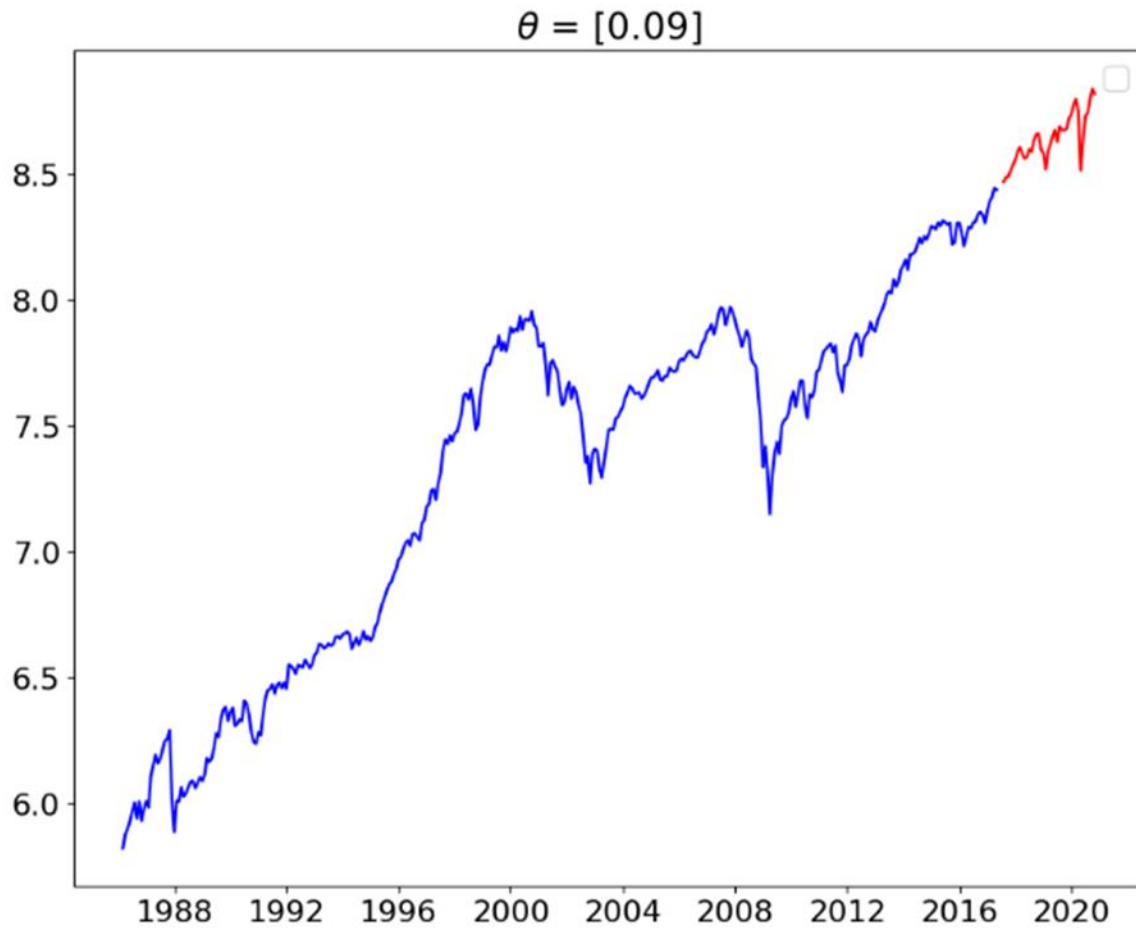
const	6.7388	0.040	168.780	0.000	6.661	6.817
ma.L1	0.9564	0.019	49.979	0.000	0.919	0.994
sigma2	0.1181	0.013	8.932	0.000	0.092	0.144
=====						
Ljung-Box (L1) (Q):	306.68	Jarque-Bera (JB):	27.43			
Prob(Q):	0.00	Prob(JB):	0.00			
Heteroskedasticity (H):	0.52	Skew:	-0.49			
Prob(H) (two-sided):	0.00	Kurtosis:	2.12			
=====						

Hình 39. Xây dựng mô hình.

Mean value xấp xỉ 6.73

Ví dụ theta = 0.09

```
i = 0
theta = np.random.uniform(0, 1, size=i + 1)
# plt.subplot(a)
plt.title(f'$\\theta$ = {theta.round(2)}')
plt.plot(ser, color='blue')
plt.plot(ser1, color='red')
plt.legend()
```

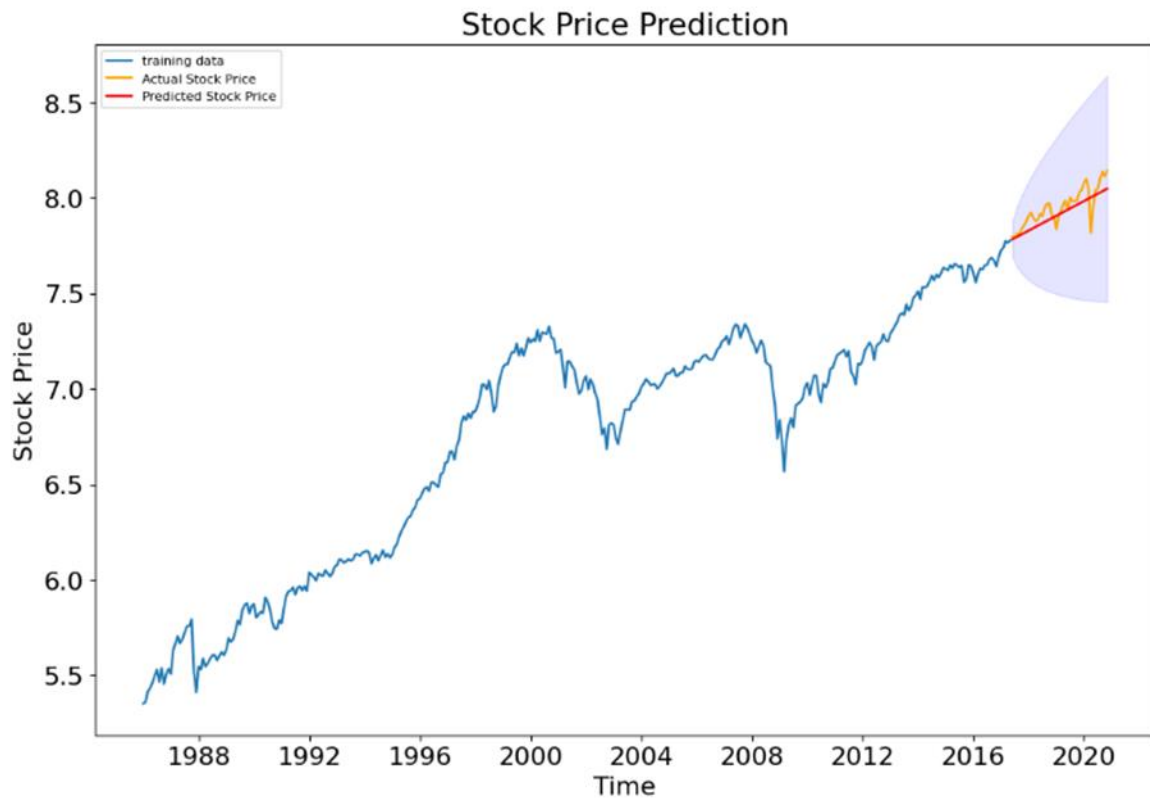


Hình 40. Biến động của chỉ số S&P với mức độ tương quan tại thời điểm t và $t-1$.

Biến động của chỉ số S&P với mức độ tương quan tại thời điểm t và $t-1$ là 9% ($\theta = 0.09$)

Nhận xét: Nếu mô hình chạy N lần thì các giá trị có thể sẽ khác nhau sau mỗi lần chạy nhưng nếu N càng lớn thì ta thấy các giá trị đều giao động xung quanh vị trí trung bình

III.2.3 Kết quả phân tích và dự đoán



Hình 41. Dự báo giá cổ phiếu.

Đối với thị trường tài chính, xét trên dữ liệu train, so với các giai đoạn trong lịch sử, ở những mốc thời gian thế giới phải hứng chịu những cuộc khủng hoảng toàn cầu 1986 (Dầu mỏ, vàng, tiền tệ) , 2000(Dotcom) , 2008 (Nhà đất). Sau đó khoảng 12- 18 tháng, chỉ số S&P 500 hay thị trường tài chính Mỹ phải hứng chịu những giai đoạn downtrend tương đối mạnh.

Xét trên dữ liệu test, tính đến tháng 8/2021 đã có một giai đoạn downtrend ngắn hạn, tuy nhiên, tiềm tàng 1 giai đoạn downtrend dài hạn vẫn là có. 2020 là năm suy thoái. Có thể đến đầu 2022 sẽ là giai đoạn downtrend mạnh cho thị trường kinh tế Mỹ



Hình 42. Xu hướng biến động của chỉ số giai đoạn 2020 - 2022.

Dữ liệu thực tế của chỉ số S&P 500 tại thời điểm hiện tại. 1/2022 đến 6/2022 là giai đoạn downtrend của Mỹ. Tuy nhiên giai đoạn downtrend này nhiều tổ chức lại kiếm được tiền hoặc không bị mất tiền nhiều. Phải chăng họ đã dự đoán được khủng hoảng. Chỉ số dự báo tiếp theo sẽ giải đáp câu hỏi này.

III.3. Mô hình ARIMA

III.3.1 Tổng quan bộ dữ liệu

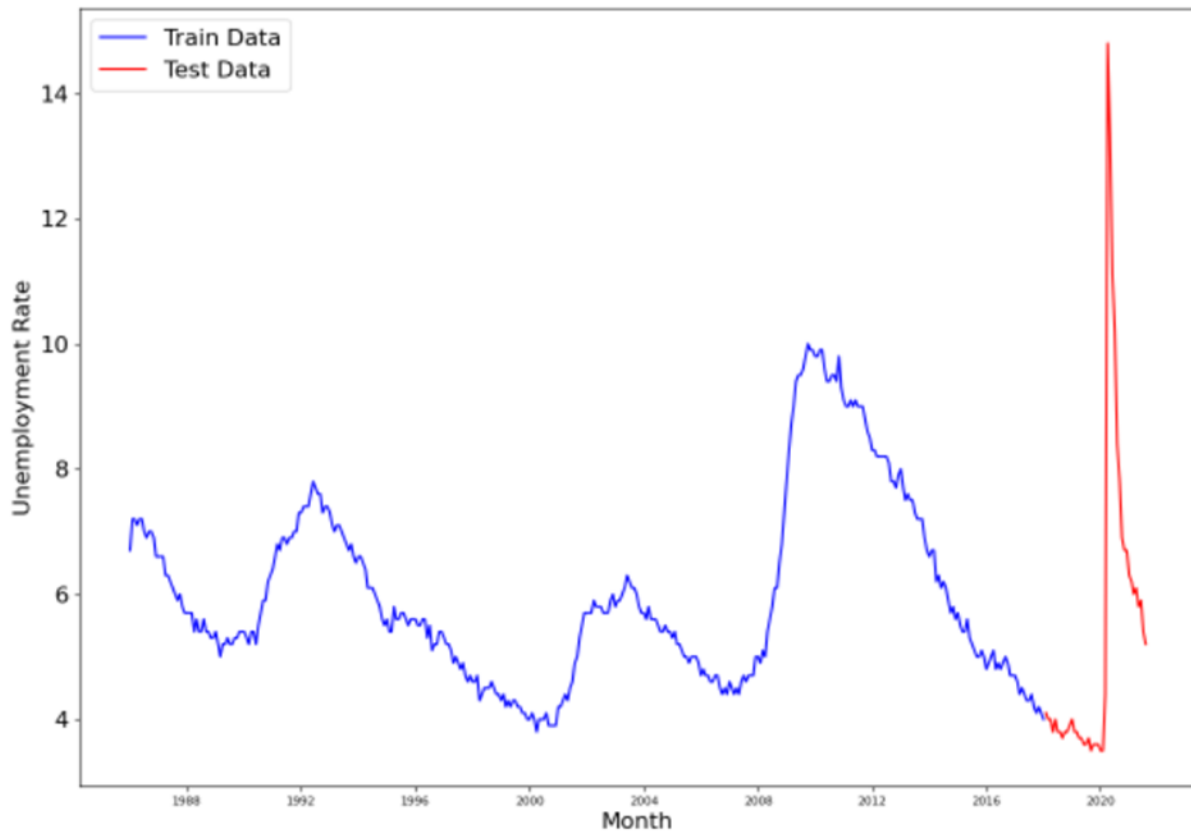
Bộ dữ liệu là tỷ lệ thất nghiệp tại Mỹ được ghi nhận nhận từ năm 1986 đến tháng 8/2021

III.3.2 Xây dựng mô hình

- Tách dữ liệu train, test

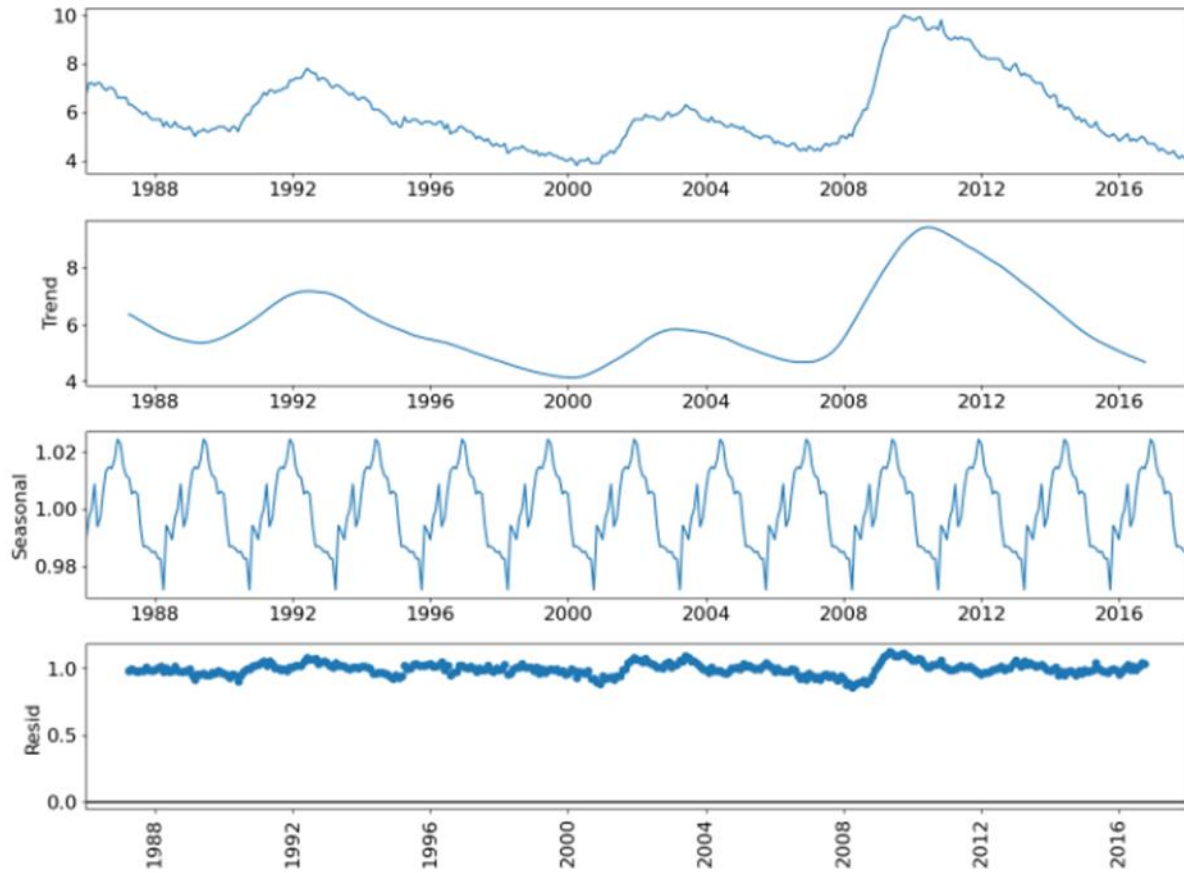
Chia tập dữ liệu train, test theo tỉ lệ tương ứng 9:1

```
train_data, test_data = df[:int(len(df)*0.9)], df[int(len(df)*0.9):]
```



Hình 43. Chia tập dữ liệu train test theo tỷ lệ 9:1.

- Phân rã mô hình



Hình 44. Phân rã chuỗi dữ liệu.

Tỷ lệ thất nghiệp có xu hướng biến thiên theo chu kỳ 10 năm, tăng mạnh vào những năm 1992, 2003, 2009.

- Kiểm định mô hình

ADF: Test statistic	-2.695189
p values	0.074875
# of lags	6.000000
# of Observations	378.000000
Critical value: (1%)	-3.447769
Critical value: (5%)	-2.869217
Critical value: (10%)	-2.570860
dtype: float64	

KPSS: Test statistic	0.282551
p values	0.100000
# of lags	11.000000
Critical value: (10%)	0.347000
Critical value: (5%)	0.463000
Critical value: (2.5%)	0.574000
Critical value: (1%)	0.739000
dtype: float64	

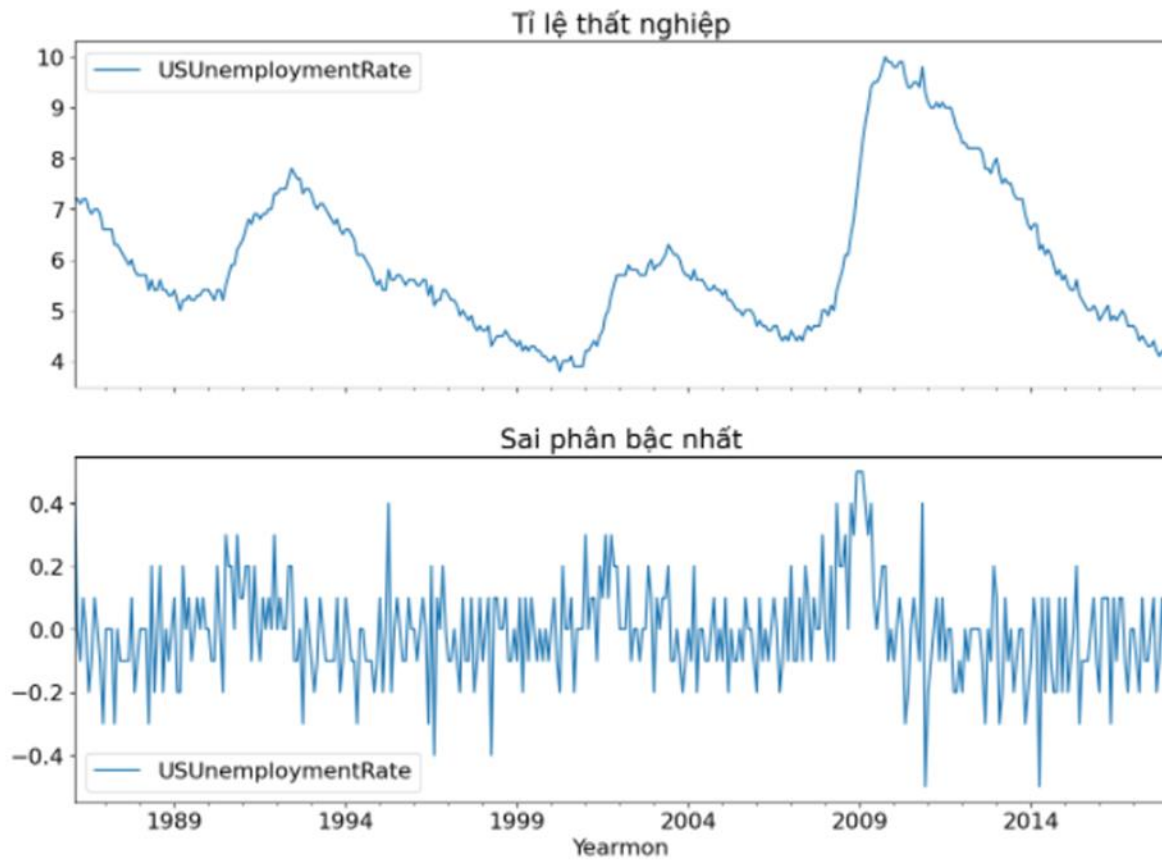
Hình 45. Kiểm định mô hình với ADF và KPSS.

Giả thuyết kiểm định KPSS, H_0 là chuỗi dữ liệu dừng. Giá trị kiểm định (0.282) nhỏ các giá trị tới hạn nên bác bỏ giả thuyết H_0 , chuỗi không dừng

Giả thuyết kiểm định ADF, H_0 là chuỗi dữ liệu dừng. Trị tuyệt đối giá trị kiểm định (2.695) lớn hơn tuyệt đối 1 giá trị tới hạn nên bác bỏ giả thuyết H_0 , chuỗi dừng ->

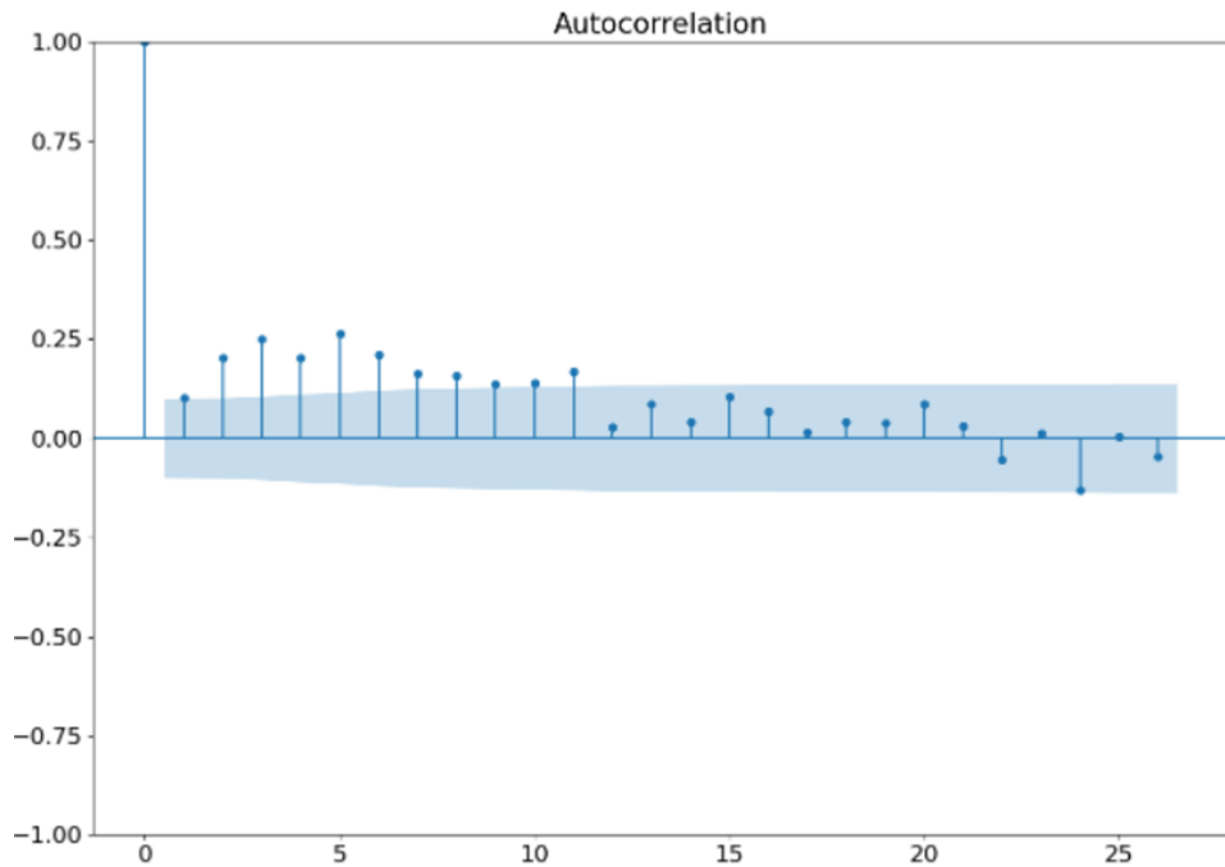
Tiến hành sai phân bậc 1 để dữ liệu thành dữ liệu dừng

- Sai phân bậc 1



Hình 46. Kết quả lấy sai phân bậc nhất.

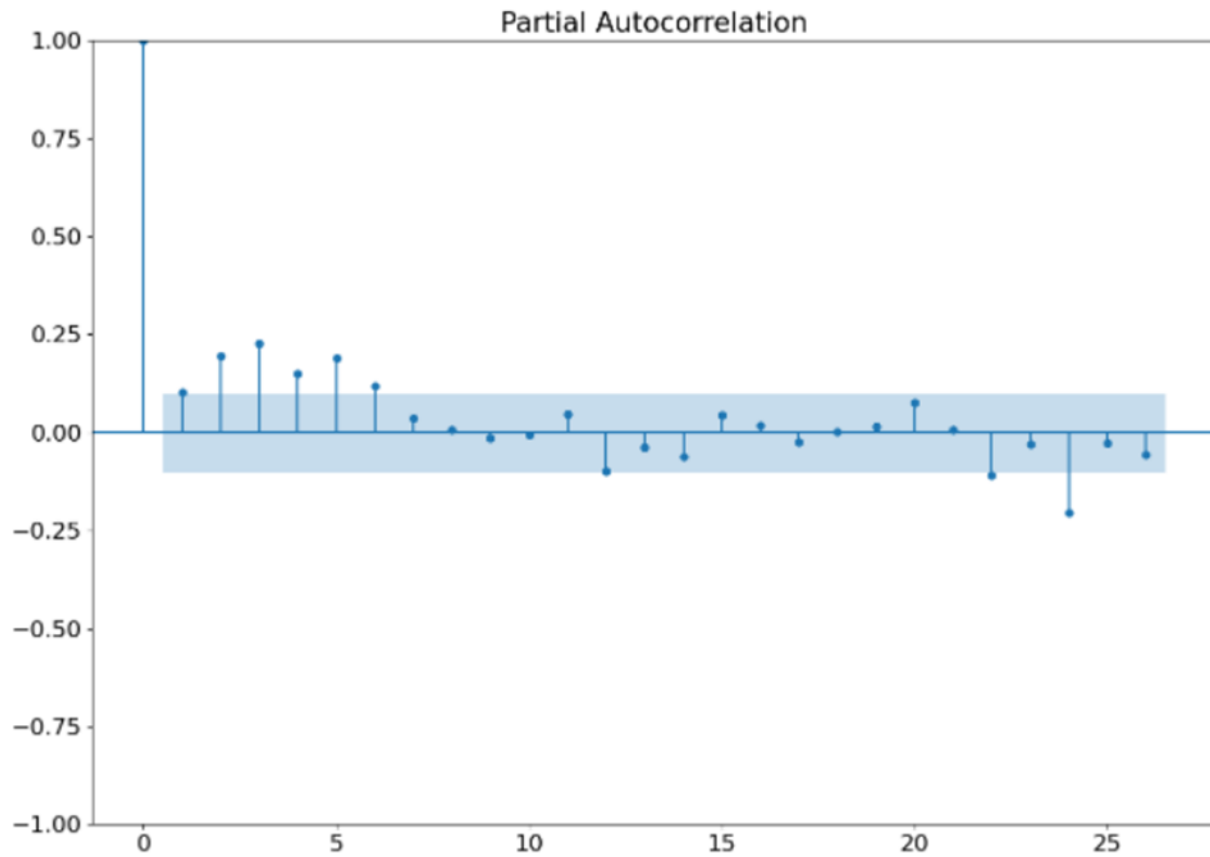
- Kiểm định ACF



Hình 47. Kiểm định tương quan tổng thể.

Kiểm định ACF (Xác định hệ số q), qua kiểm định, k -lag, $k < 12$ có sự tương quan mạnh, hay nói cách khác, tỷ lệ thất nghiệp tại thời điểm từ $t-11$, $t-10$, ..., $t-1$ có sự tương quan mạnh với tỷ lệ thất nghiệp tại thời điểm hiện tại (thời gian cuối cùng trong chuỗi dữ liệu train).

- Kiểm định PACF



Hình 48. Kiểm định tương quan riêng.

Kiểm định PACF (Xác định hệ số p), mức độ toàn dụng lao động có xu hướng di chuyển theo đồ thị hình sin và có biên độ giao động giảm dần.

- Xây dựng mô hình

```

ARIMA(2,1,2)(0,0,0)[0] intercept : AIC=-395.632, Time=0.88 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=-336.175, Time=0.04 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=-338.321, Time=0.12 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=-337.195, Time=0.12 sec
ARIMA(0,1,0)(0,0,0)[0] : AIC=-337.390, Time=0.03 sec
ARIMA(1,1,2)(0,0,0)[0] intercept : AIC=inf, Time=0.74 sec
ARIMA(2,1,1)(0,0,0)[0] intercept : AIC=-387.895, Time=0.54 sec
ARIMA(3,1,2)(0,0,0)[0] intercept : AIC=-386.881, Time=1.07 sec
ARIMA(2,1,3)(0,0,0)[0] intercept : AIC=-393.559, Time=1.27 sec
ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=-378.515, Time=0.51 sec
ARIMA(1,1,3)(0,0,0)[0] intercept : AIC=-393.996, Time=0.58 sec
ARIMA(3,1,1)(0,0,0)[0] intercept : AIC=-390.572, Time=0.58 sec
ARIMA(3,1,3)(0,0,0)[0] intercept : AIC=-391.906, Time=1.17 sec
ARIMA(2,1,2)(0,0,0)[0] : AIC=-397.547, Time=0.48 sec
ARIMA(1,1,2)(0,0,0)[0] : AIC=-393.514, Time=0.86 sec
ARIMA(2,1,1)(0,0,0)[0] : AIC=-389.816, Time=0.23 sec
ARIMA(3,1,2)(0,0,0)[0] : AIC=-389.034, Time=0.64 sec
ARIMA(2,1,3)(0,0,0)[0] : AIC=-395.559, Time=0.62 sec
ARIMA(1,1,1)(0,0,0)[0] : AIC=-380.426, Time=0.27 sec
ARIMA(1,1,3)(0,0,0)[0] : AIC=-395.945, Time=0.38 sec
ARIMA(3,1,1)(0,0,0)[0] : AIC=-392.499, Time=0.31 sec
ARIMA(3,1,3)(0,0,0)[0] : AIC=-396.969, Time=1.03 sec

```

Best model: ARIMA(2,1,2)(0,0,0)[0]

Hình 49. Xây dựng mô hình ARIMA.

```

model = ARIMA(train_data, order=(2,1,2))
fitted = model.fit()
print(fitted.summary())

```

- Xây dựng mô hình theo ARIMA(2, 1, 2)

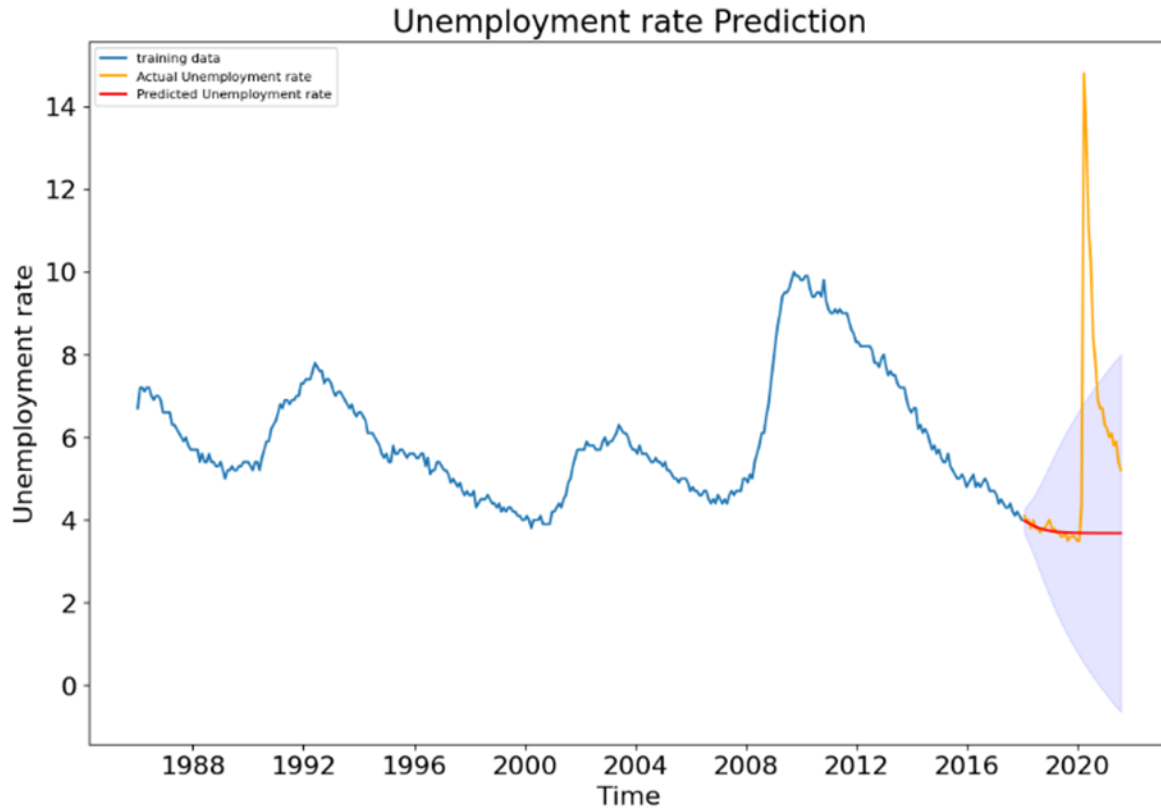
SARIMAX Results						
=====						
Dep. Variable:	USUnemploymentRate	No. Observations:	385			
Model:	ARIMA(2, 1, 2)	Log Likelihood	203.773			
Date:	Wed, 10 Aug 2022	AIC	-397.547			
Time:	21:13:50	BIC	-377.793			
Sample:	01-01-1986	HQIC	-389.712			
	- 01-01-2018					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

ar.L1	1.4208	0.148	9.623	0.000	1.131	1.710
ar.L2	-0.4867	0.144	-3.382	0.001	-0.769	-0.205
ma.L1	-1.4822	0.131	-11.355	0.000	-1.738	-1.226
ma.L2	0.6580	0.113	5.816	0.000	0.436	0.880
sigma2	0.0202	0.001	15.061	0.000	0.018	0.023
=====						
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	4.13			
Prob(Q):	0.93	Prob(JB):	0.13			
Heteroskedasticity (H):	1.29	Skew:	0.13			
Prob(H) (two-sided):	0.15	Kurtosis:	3.43			
=====						

Hình 50. Xây dựng mô hình ARIMA theo (2,1,2).

Hệ số P value xấp xỉ 0, nhỏ hơn 5% (độ tin cậy 95%) nên mô hình thống kê có ý nghĩa

III.3.3 Kết quả phân tích và dự đoán



Hình 51. Dự đoán tỷ lệ thất nghiệp dựa trên mô hình.

Với dữ liệu train, đỉnh của tỷ lệ thất nghiệp nằm ở các năm 1986, 1993, 2003, 2009, trùng với khoảng thời gian 1-3 năm sau những khủng hoảng toàn cầu, 1983, khủng hoảng hàng hóa (dầu mỏ, vàng, tiền tệ,...), 2001, khủng hoảng Dotcom, 2008 khủng hoảng nhà đất.

Hầu hết những tổ chức chính phủ hay tổ chức kinh tế trên toàn thế giới đều xem tỉ lệ thất nghiệp như một chỉ báo quan trọng và tỷ lệ thất nghiệp gần chạm đáy vào năm 2018 là một dấu hiệu cho thấy sắp có cuộc suy thoái

Với dữ liệu test, covid19 đã đẩy nhanh quá trình suy thoái kinh tế thế giới, gây ra nhiều hậu quả nghiêm trọng, tỷ lệ thất nghiệp cao nhất từng được ghi nhận. Tuy nhiên để gọi là khủng hoảng thì 2019 chưa hẳn là một năm khủng hoảng với thế giới, suy thoái đã được dự đoán từ trước và tổ chức tài chính, tổ chức chính phủ đã có những kế hoạch chuẩn bị và phòng ngừa rủi ro

IV. Các mô hình máy học

IV.1. Đặt vấn đề:

Trong nền kinh tế hiện nay, việc tìm kiếm được những nhân viên giỏi là rất quan trọng, không chỉ vậy khi đã chiêu mộ được nhân viên tốt công ty cần phải làm thêm những gì để họ có thể phát huy hết tiềm năng của mình và những yếu tố nào ảnh hưởng đến năng suất làm việc của một nhân viên cũng là điều hết sức quan trọng. Việc phân loại các yếu tố để xác định khả năng nhân viên có nghỉ việc hay không là rất quan trọng. Nhóm đã tiến hành xây dựng ba mô hình máy học Decision Tree, Random Forest, Neural Network nhằm trả lời câu hỏi: Liệu tình trạng nghỉ làm của nhân viên bị tác động bởi các yếu tố khác như thế nào và mô hình máy học nào có độ chính xác cao nhất để phục vụ cho việc ứng dụng vào thực tiễn.

IV.2. Tổng quan bộ dữ liệu:

Bộ dữ liệu được cung cấp bởi Kaggle là bộ dữ liệu gồm 14999 dòng và được cấu hình bởi các thuộc tính sau:

satisfacion_level: Mô tả về mức độ hài lòng của nhân viên

last_evaluation: Mức độ hài lòng cuối cùng của nhân viên

number_project: Số dự án đã thực hiện

average_monthly_hours: Số giờ làm trong tháng

time_spend_company: Số năm làm trong công ty

Work_accident: Có gặp tai nạn lao động hay không (0: Không; 1: Có)

Left: Nhân viên đã nghỉ việc hay chưa (0: Chưa nghỉ; 1: Nghỉ việc)

promotion_last_5years: Có thăng tiến trong 5 năm qua hay không (0: Không; 1: Có)

dept: Phòng ban làm việc

salary: Mức lương (Low; Medium, High)

IV.3. Quá trình xây dựng mô hình với các giải thuật

IV.3.1. Thông tin dữ liệu:

```
> <class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
#   Column                      Non-Null Count  Dtype
---  -
0   satisfaction_level          14999 non-null  float64
1   last_evaluation             14999 non-null  float64
2   number_project              14999 non-null  int64
3   average_monthly_hours      14999 non-null  int64
4   time_spend_company          14999 non-null  int64
5   Work_accident               14999 non-null  int64
6   left                        14999 non-null  int64
7   promotion_last_5years       14999 non-null  int64
8   sales                       14999 non-null  object
9   salary                     14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

Hình 52. Kiểm tra thông tin chung của dữ liệu.

Từ kết quả kiểm tra sơ bộ trên có thể thấy dữ liệu có tình toàn vẹn. Các biến định lượng và định tính dưới dạng nhị phân được định dạng dưới dạng dữ liệu int và float và bên cạnh đó các biến định tính và thứ bậc được định dạng dưới dạng dữ liệu object.

Khác với mô hình hồi quy logistic ở trên chỉ tập trung vào độ hài lòng của của nhân viên thì khi áp dụng các mô hình máy học, nhóm tiến hành sử dụng tất cả các biến độc lập để phân loại dự đoán xem nhân viên đó có khả năng nghỉ việc hay không. Các giải thuật được sử dụng đối với bài toán này sẽ bao gồm Decision Tree, Random Forest và Neural Network. Đối với từng mô hình sẽ có cách xử lý chuẩn hóa dữ liệu khác nhau.

IV.3.3. Xây dựng Decision Tree

Mô hình cây quyết định là một dạng máy học có giám sát (Supervised Learning) trong các mô hình máy học Machine Learning (ML). Cây quyết định là một dạng cây phân cấp có cấu trúc và được dùng để phân lớp các đối tượng dựa vào các dãy thuật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary), Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi thuộc tính phân loại phải có kiểu dữ liệu là Binary hoặc Ordinal.

Với tập dữ liệu trên, nhóm áp dụng mô hình cây quyết định để dự đoán khả năng rời công ty của nhân viên.

Tiến hành thiết lập dữ liệu bằng cách tạo ra các phần tử chứa chuỗi 'salary' sau đó chuyển đổi dữ liệu phân loại thành các biến chỉ số.

```
In [28]: # tiến hành thiết lập dữ liệu bằng cách tạo ra các phần tử chứa chuỗi 'salary'
# sau đó chuyển đổi dữ liệu phân loại thành các biến chỉ số
Elements = ['salary']
final_data = pd.get_dummies(df, columns=Elements, drop_first=True)
final_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   satisfaction_level     14999 non-null  int64
1   last_evaluation        14999 non-null  int64
2   number_project         14999 non-null  int64
3   average_monthly_hours  14999 non-null  int64
4   time_spend_company     14999 non-null  int64
5   work_accident          14999 non-null  int64
6   left                  14999 non-null  int64
7   promotion_last_5years  14999 non-null  int64
8   sales                  14999 non-null  int32
9   salary_1              14999 non-null  uint8
10  salary_2              14999 non-null  uint8
dtypes: int32(1), int64(8), uint8(2)
memory usage: 1.0 MB
```

Hình 53. Chuyển đổi biến phân loại thành các biến chỉ số.

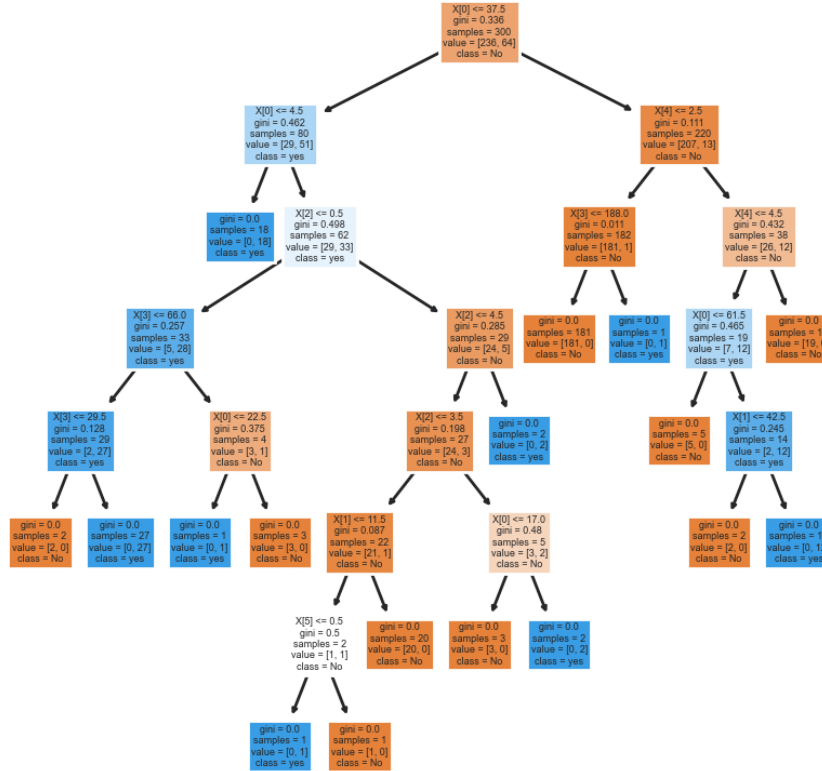
Sau đó, chia tập dữ liệu huấn luyện (train) và kiểm thử (test) sử dụng phương thức `train_test_split` trong Sklearn. Do tập dữ liệu quá lớn (15.000 dòng) do đó sử dụng 2% cho tập dữ liệu Test data và 98% còn lại cho tập Train data.

```
In [29]: # chia tập dữ liệu train test
from sklearn.model_selection import train_test_split
X = final_data.drop('left', axis=1)
y = final_data['left']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.02, random_state=6)
```

Hình 54. Chia tập dữ liệu train test.


```
In [33]: # plot tree
plt.figure(figsize=(5, 5), dpi=200)
t = tree.plot_tree(dtree, class_names=["No", "yes"], filled=True)
plt.show()
```

Hình 58. Sử dụng phương thức `tree.plot tree` để hiển thị cây.



Hình 59. Trực quan hoá cây quyết định.

F1-score cao (89%), đối với nhóm nhân viên sẽ rời công ty - True Negative (TN) với 3261 nhân viên được dự đoán đúng sẽ rời công ty trên tổng số 3507 số nhân viên nghỉ làm thực tế. Điều này cho thấy mô hình Cây quyết định áp dụng với tập dữ liệu là một mô hình phân loại tốt và có ý nghĩa thống kê, tiên lượng.

IV.3.4. Xây dựng Random Forest:

Nhóm tiến hành xây dựng mô hình dựa trên Random Forest cho Classifier. Vì dữ liệu phân loại 0 hoặc 1 nên nhóm chọn RandomForestClassifier thay vì RandomForestRegressor

```
from sklearn.ensemble import RandomForestClassifier
```

```
classifier = RandomForestClassifier(n_estimators=20, random_state=0)
```

```

classifier.fit(x_train, y_train)
y_pred = classifier.predict(x_test)
print(confusion_matrix(y_test,y_pred.round()))
print(classification_report(y_test,y_pred.round()))
print(accuracy_score(y_test, y_pred.round()))

```

Hình 60. Mô hình được xây dựng với 20 cây ($n_estimators = 20$).

```

[[2294   5]
 [  21 680]]

```

	precision	recall	f1-score	support
0	0.99	1.00	0.99	2299
1	0.99	0.97	0.98	701
accuracy			0.99	3000
macro avg	0.99	0.98	0.99	3000
weighted avg	0.99	0.99	0.99	3000

```

0.9913333333333333

```

Hình 61. Mô hình với 20 cây được xây dựng.

Độ chính xác của mô hình rất cao 99,133%

Thay đổi mô hình, thực hiện với 200 cây ($n_estimators = 200$)

[[2294 5]					
[17 684]]					
		precision	recall	f1-score	support
0	0.99	1.00	1.00	2299	
1	0.99	0.98	0.98	701	
accuracy				0.99	3000
macro avg		0.99	0.99	0.99	3000
weighted avg		0.99	0.99	0.99	3000
0.9926666666666667					

Hình 62. Mô hình với 200 cây được xây dựng.

Độ chính xác đối của random forest classifier với 200 cây là 99,267%, không khác lần thực hiện với 20 cây. Nên thay đổi số lượng estimators cho trường hợp này không thay đổi đáng kể đến kết quả. 99,267% là con số đẹp và không nhất thiết phải thay đổi số cây để cải thiện để nâng cao độ chính xác (accuracy).

RandomForest đã cho thấy độ chính xác cao khi thực hiện mô hình và phản ánh chính xác nhân viên đó có khả năng nghỉ việc hay không

IV.3.5. Xây dựng Neural Network:

Đối với giải thuật Neural Network, nhóm đồ án thực hiện xây dựng mô hình neural network với 3 hidden layers với số neuron bằng với các thuộc tính dùng để phân loại (8). Và Activation Function được sử dụng với mô hình là relu - rectified linear unit.

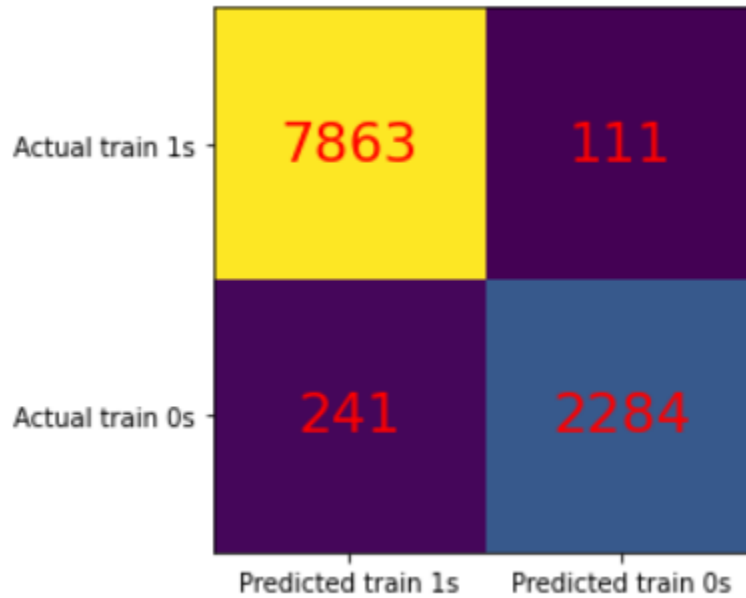
```
#Create model neural network
model = MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu', solver='adam',
max_iter=500)
model.fit(x_train,y_train)
```

MLPClassifier

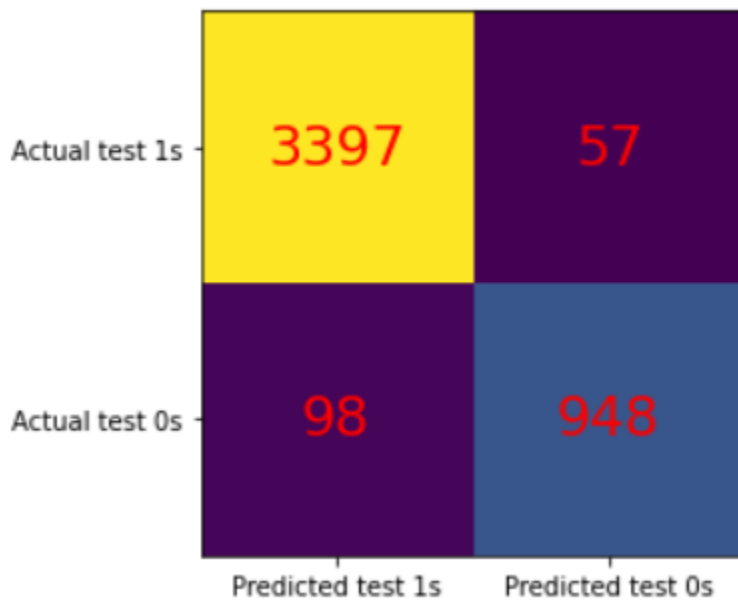
MLPClassifier(hidden_layer_sizes=(8, 8, 8), max_iter=500)

Hình 63. Xây dựng Neural Network.

Kết quả dự đoán của mô hình với 2 tập dữ liệu train và test lần lượt được thể hiện qua ma trận nhầm lẫn sau:



Hình 64. Ma trận nhầm lẫn với train data.



Hình 65. Ma trận nhầm lẫn với test data.

Mô hình dự đoán có kết quả tốt khi tỷ lệ True Positive chiếm tỷ trọng lớn chứng tỏ mô hình Neural network có khả năng phân loại được người đó đã nghỉ việc hay chưa.

IV.4. Kết quả phân tích:

Ba mô hình Decision Tree, Random Forest và Neural Network đều cho ra độ chính xác rất cao (đều từ 90% trở lên). Tập dữ liệu ra kết quả có xu hướng bias và overfitting. Tuy nhiên độ chính xác cao nên tập dữ liệu dự đoán khả năng nhân viên nào có khuynh hướng sẽ nghỉ trong tương lai cũng rất chính xác. Điều này giúp doanh nghiệp đưa ra hoặc phát triển thêm những yếu tố tác động mạnh mẽ đến việc nhân viên có muốn làm tại công ty hay không, từ đó có chính sách cải thiện và tạo môi trường làm việc phù hợp. Ngoài ra, doanh nghiệp còn biết được các yếu tố tưởng chừng có tác động nhưng thực tế lại tác động ít, từ đó cắt giảm và đầu tư vào phát triển những yếu tố thật sự ảnh hưởng

PHẦN B. TỔNG KẾT VÀ ĐỊNH HƯỚNG PHÁT TRIỂN:

I.Những điều đã làm được:

Áp dụng các kiến thức nền tảng của môn học Phân tích dữ liệu với R/Python, đề án của nhóm đã thực hiện được các mục tiêu sau :

- Phân tích, dự báo với mô hình hồi quy tuyến tính trường hợp đơn biến, đa biến và nêu rõ cơ sở chọn biến độc lập.
- Phân tích, dự báo với mô hình hồi quy Logistic trường hợp đơn biến, Logistic đa biến với biến độc lập là biến tuyến tính, Logistic đa biến với biến độc lập là biến nhị phân và biến thứ bậc.
- Phân tích dữ liệu chuỗi thời gian với mô hình ARIMA
- Ứng dụng các mô hình máy học Machine Learning trong phân tích dự báo sử dụng các giải thuật : Decision Tree, Random Forest, Neural Network

II.Những mặt hạn chế:

Do lần đầu tiên nhóm tiếp cận với các kiến thức liên quan đến ngôn ngữ R và Python cũng như các nội dung của môn học Kinh tế lượng nên không thể tránh khỏi các thiếu sót khi thực hiện đề án môn học.

Do thời gian có phần hạn chế, nhóm không thể tìm hiểu và ứng dụng được tất cả các mô hình máy học trong phân tích dự báo theo như yêu cầu của đề án môn học.

III. Định hướng tương lai:

Về định hướng phát triển trong tương lai, nhóm tiếp tục nghiên cứu sâu hơn về ngôn ngữ R và Python, các mô hình trong phân tích dự báo để tiếp tục hoàn thiện những mặt còn thiếu sót của đề án.

Có thể áp dụng thêm các mô hình máy học khác (Support Vector Machine, K-nearest Neighbor, Naive Bayes..) để có cái nhìn tổng quan và chính xác hơn khi áp dụng các mô hình này trong tiên lượng dự đoán với cùng một tập dữ liệu dataset.