

Báo cáo đồ án

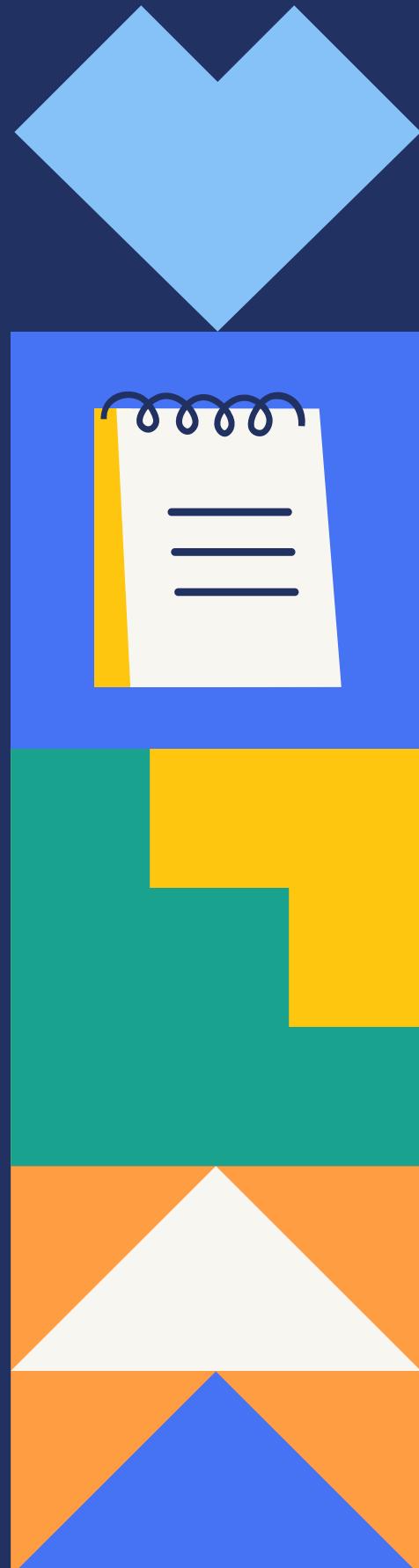
Phân tích dữ liệu

với R/python

GVHD: Nguyễn Quang Phúc

G_232IS2902_08





Nội dung báo cáo



Mô hình hồi quy tuyến tính



Phân tích dữ liệu chuỗi thời gian

- ARIMA
- Holt-Winter



Mô hình hồi quy logistic



Mô hình máy học

- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine



Mô hình hồi quy tuyến tính



Hồi quy tuyến tính

Nội dung

1 Mô tả bài toán

2 Tiền xử lý dữ liệu

3 Mô hình đơn biến

4 Mô hình đa biến

Dataset:
[BostonHousing.csv](#)

Một công ty bất động sản hiện tại đang đánh giá giá nhà đất dựa trên các yếu tố như diện tích, vị trí, tiện ích xung quanh và nhiều yếu tố khác.

Thuộc tính	Mô tả
crim	Tỷ lệ tội phạm bình quân đầu người theo thị trấn.
zn	Tỷ lệ đất ở được quy hoạch cho các lô có diện tích trên 25.000 m ² .
indus	Tỷ lệ mẫu đất kinh doanh phi bán lẻ trên mỗi thị trấn.
chas	Biến nhị phân cho biết nhà có ở gần sông Charles hay không (1 cho có, 0 cho không).
nox	Nồng độ oxit nitric (phần trên 10 triệu)



rm	Số phòng trung bình/căn hộ.
age	Tỷ lệ các căn hộ cũ do chủ sở hữu sử dụng xây dựng trước năm 1940.
dis	Khoảng cách có trọng số tới các trung tâm việc làm ở Boston.
rad	Chỉ số khả năng tiếp cận các đường cao tốc hướng tâm.
tax	Thuế suất tài sản trên 10.000 USD.
ptratio	Tỷ lệ học sinh-giáo viên trong thị trấn
b	Tỷ lệ người da đen theo thị trấn
lstat	Tỷ lệ dân số có địa vị thấp
medv	Giá trị trung bình của những ngôi nhà do chủ sở hữu sử dụng tính bằng \$1000

Làm thế nào xác định giá nhà đất một cách chính xác và hiệu quả hơn?



Hồi quy tuyến tính

Nội dung

1

Mô tả bài toán

2

Tiền xử lý dữ liệu

3

Mô hình đơn biến

4

Mô hình đa biến

Dataset:
BostonHousing.csv

Xử lý dữ liệu trùng lặp và không hợp lệ

- Không có dòng nào bị trùng lặp với nhau, đảm bảo các dòng là duy nhất

Xử lý dữ liệu và giá trị bị thiếu

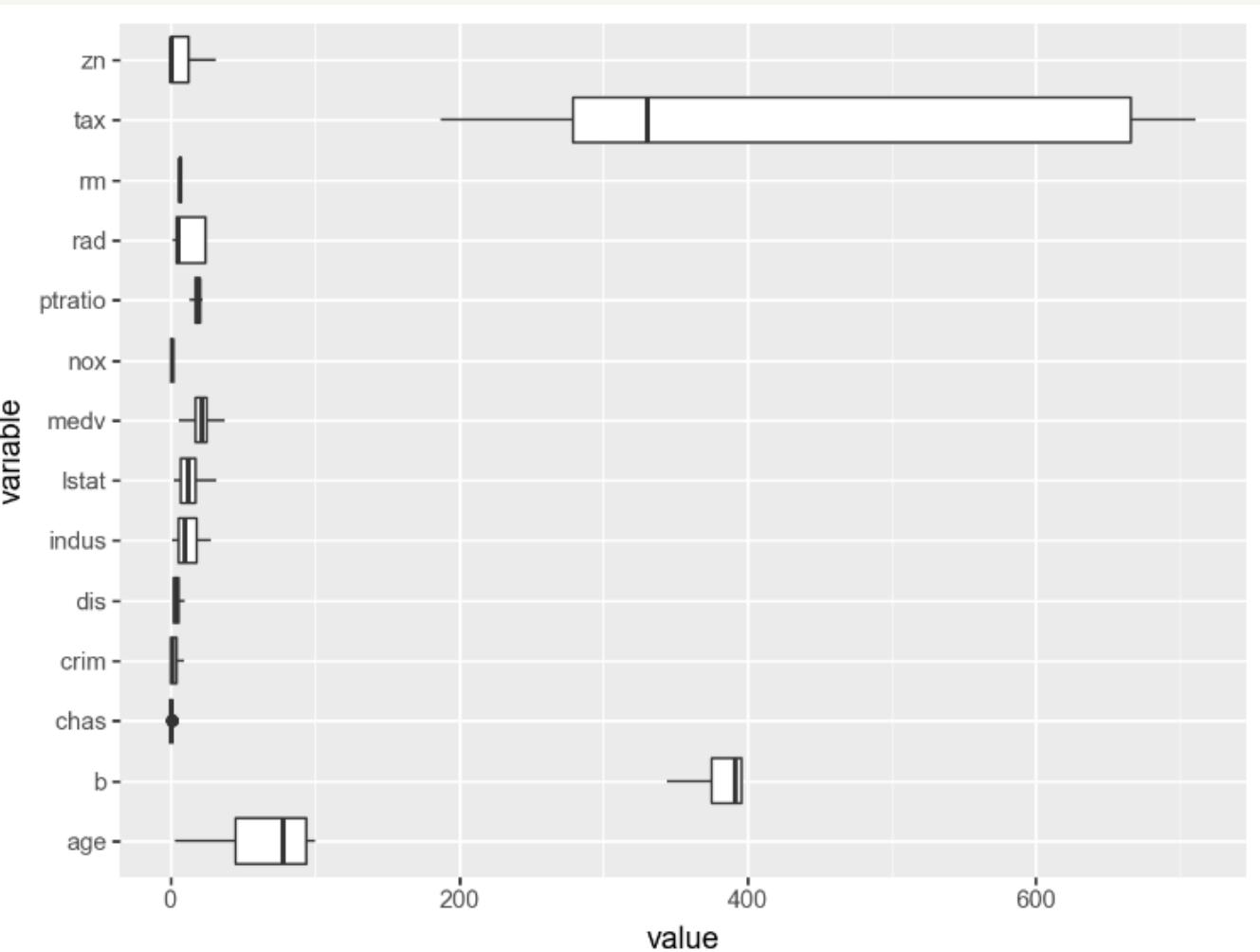
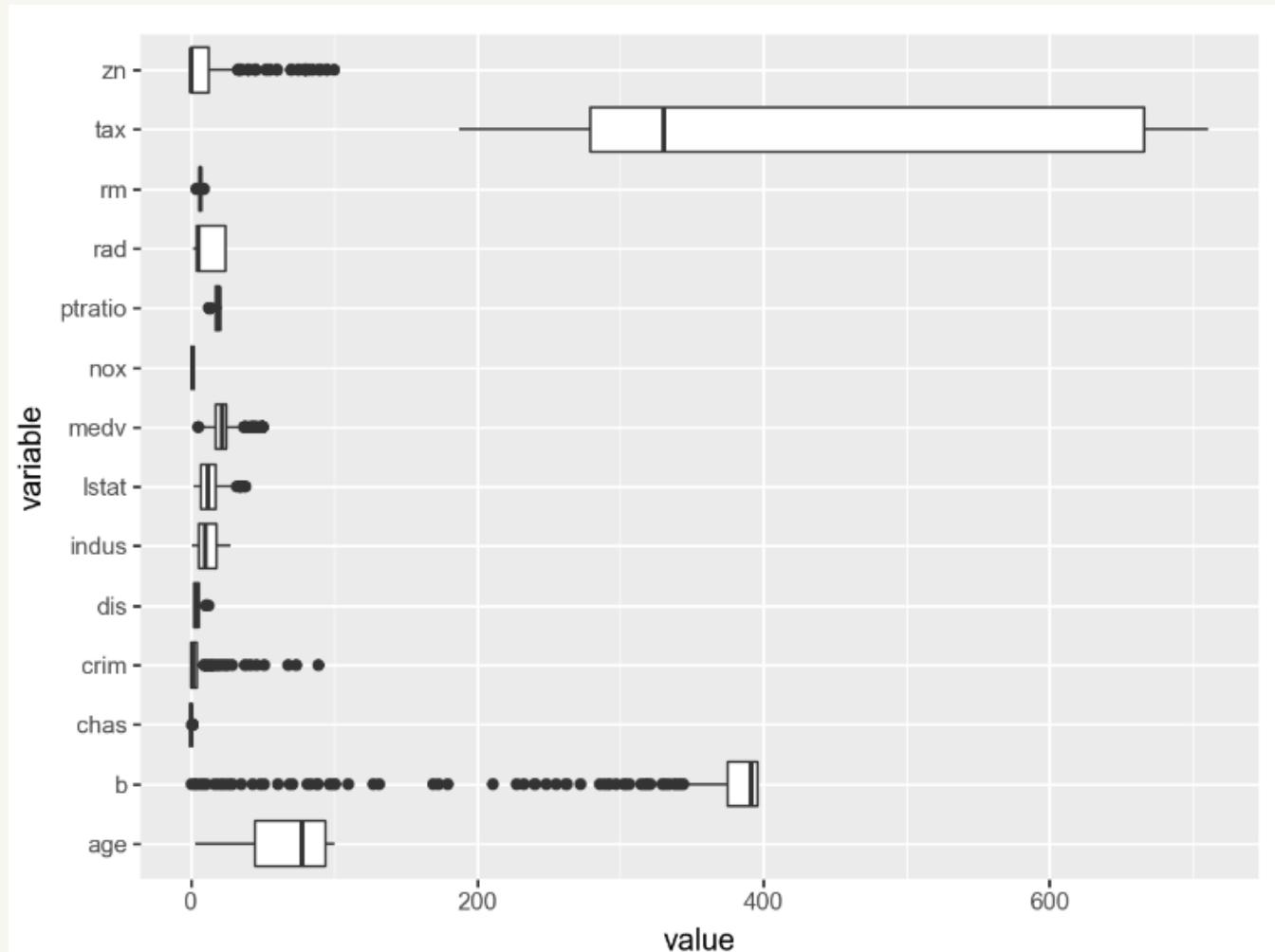
- Cột “rm” với 5 dữ liệu trống, chiếm 0.99% trên tổng bộ 506 dữ liệu của cột này thay thế bằng trung vị nhằm tránh ảnh hưởng tới phân phối ban đầu của dữ liệu.

Xử lý các ngoại lệ

- 5 cột không có giá trị ngoại lai bao gồm: “tax, rad, nox, indus, age” và các cột còn lại đều có các giá trị ngoại lai.
- Áp dụng phương pháp tính Độ trải giữa (IQR) để loại bỏ các giá trị ngoại lai và thay thế bằng các giá trị giới hạn trên và giới hạn dưới tương ứng.

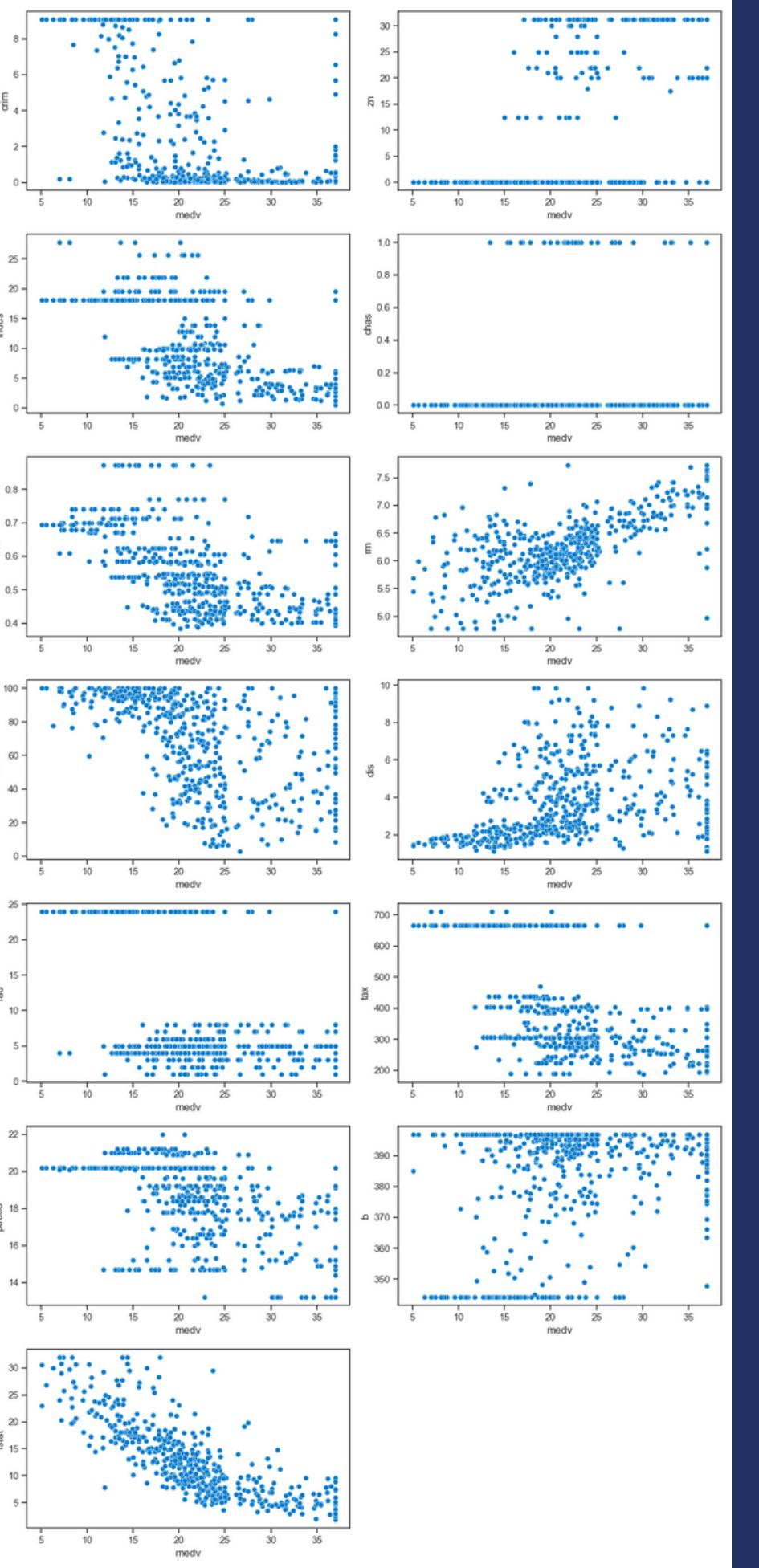
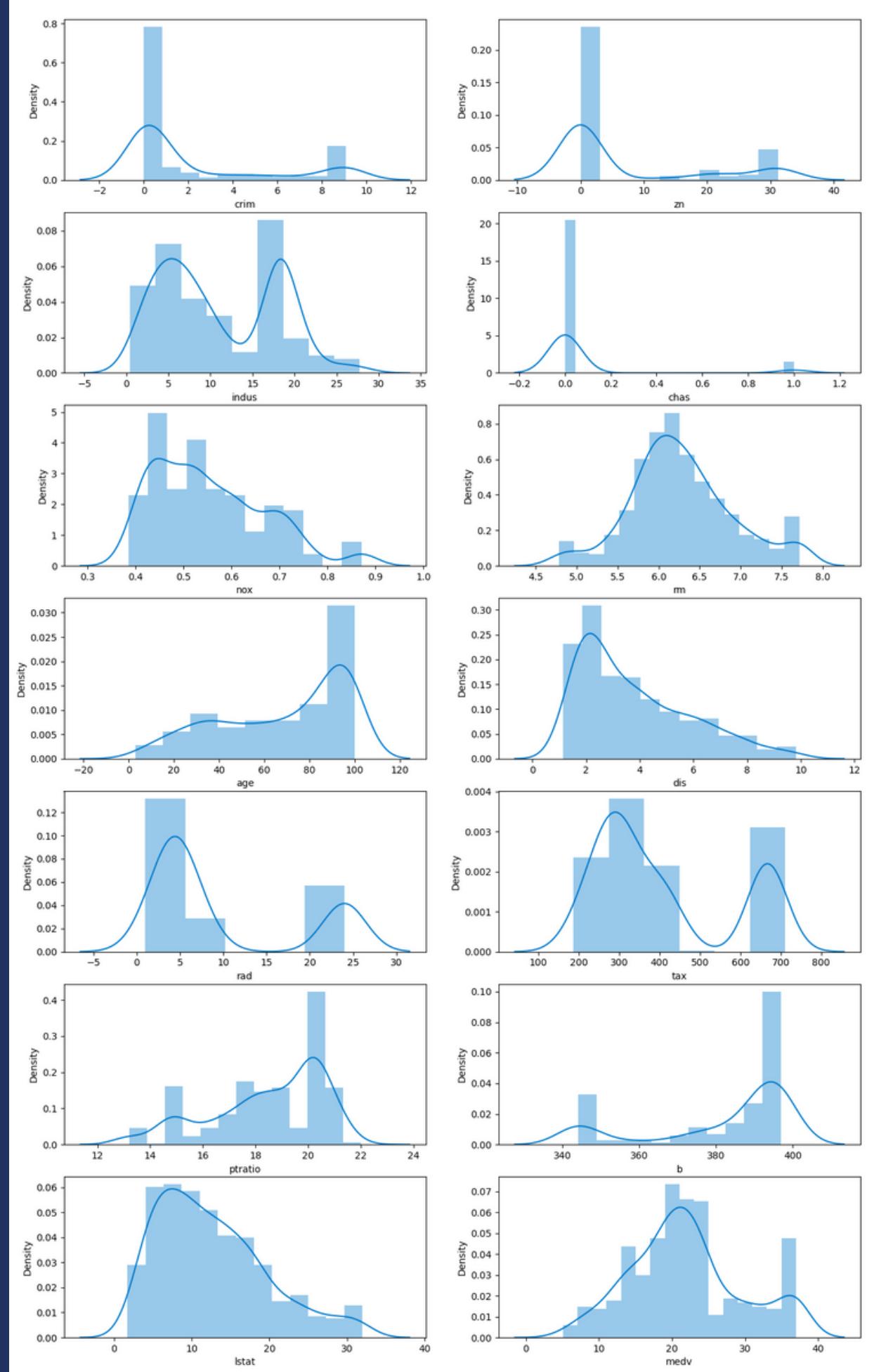
Phân tích khám phá dữ liệu

Xử lý ngoại lệ

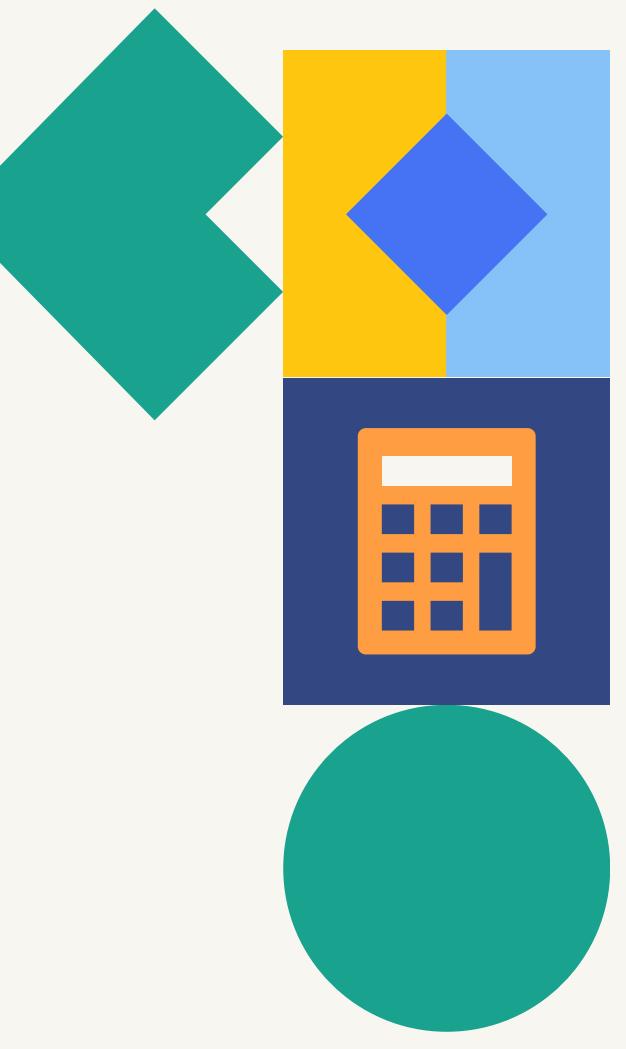


Hình: Biểu đồ boxplot các giá trị số

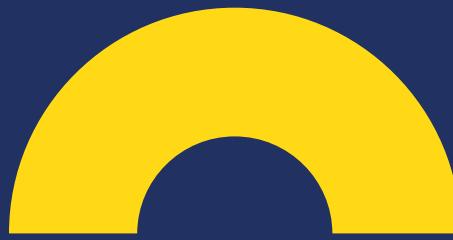
3 cột “zn, crim, b”
có nhiều giá trị
ngoại lai nhất



Tương quan

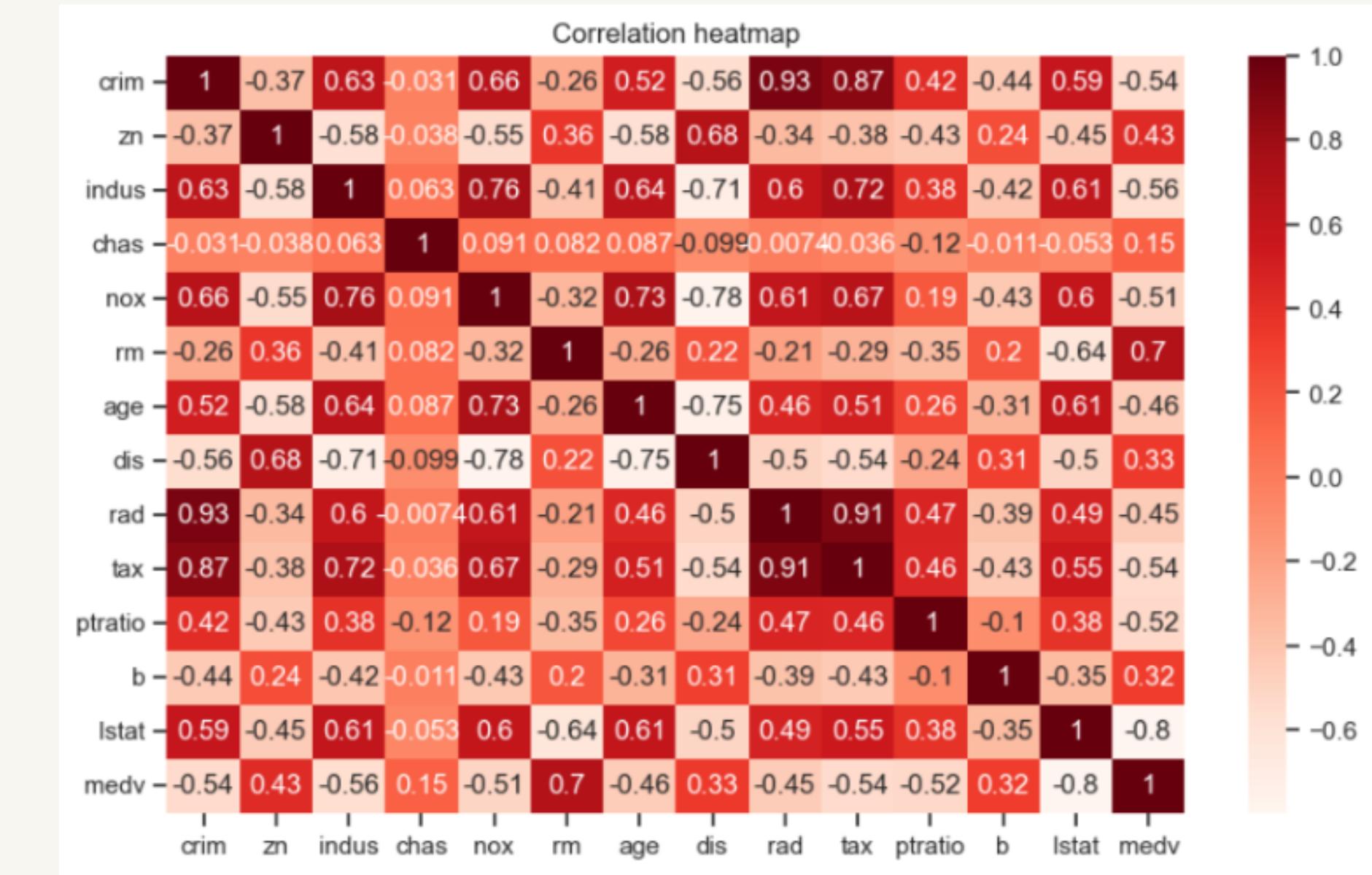


Nhận xét



Tương quan

Hai biến “rm” và “Istat” được đánh giá có **tương quan cao** bằng biểu đồ phân tán cũng ghi nhận lần lượt là biến có tương quan thuận cao nhất (**rm đạt 0.7**) và biến có tương quan nghịch cao nhất (**Istat đạt -0.8**)



Hồi quy tuyến tính

Nội dung

1

Mô tả bài toán

2

Tiền xử lý dữ liệu

3

Mô hình đơn biến

4

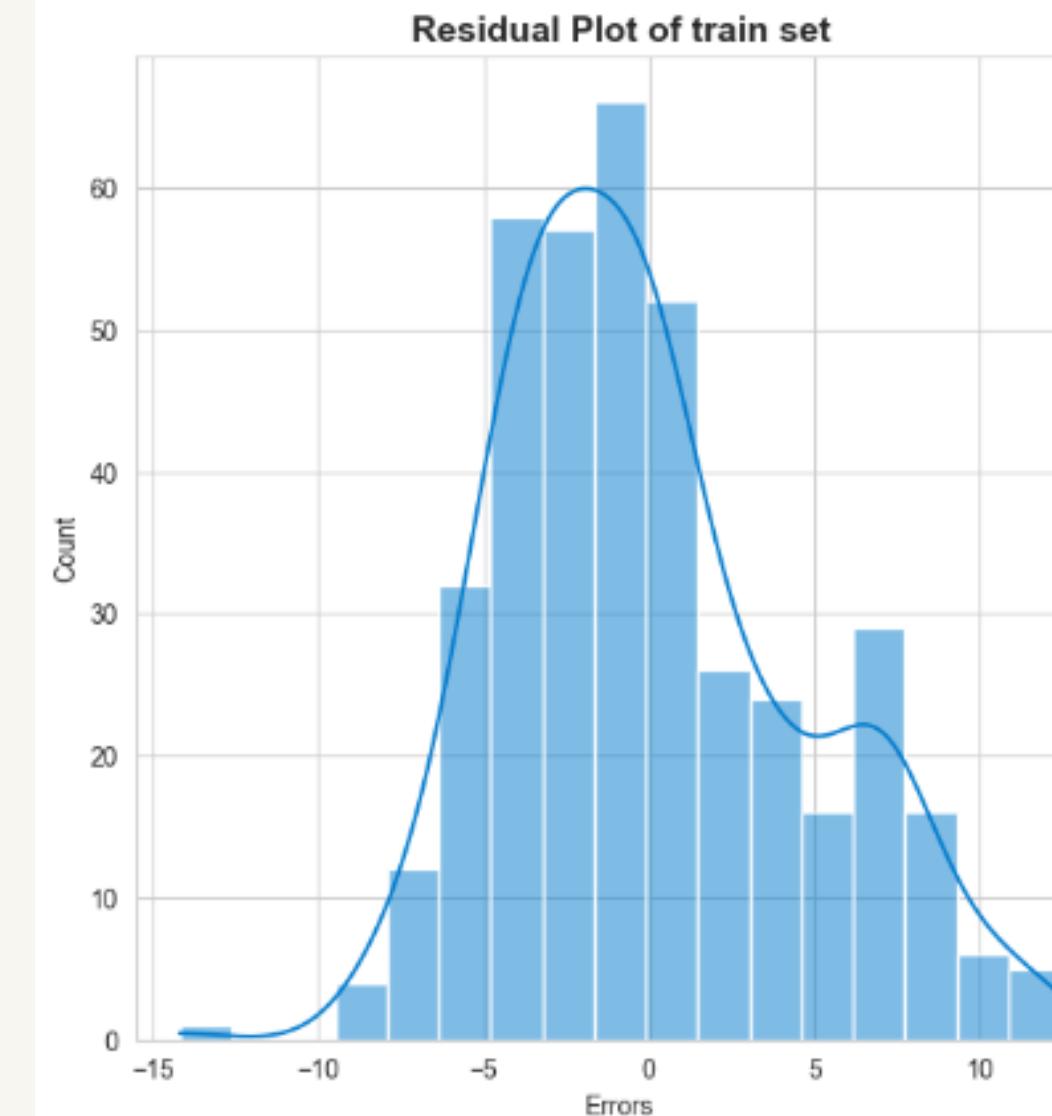
Mô hình đa biến

Dataset:

BostonHousing.csv

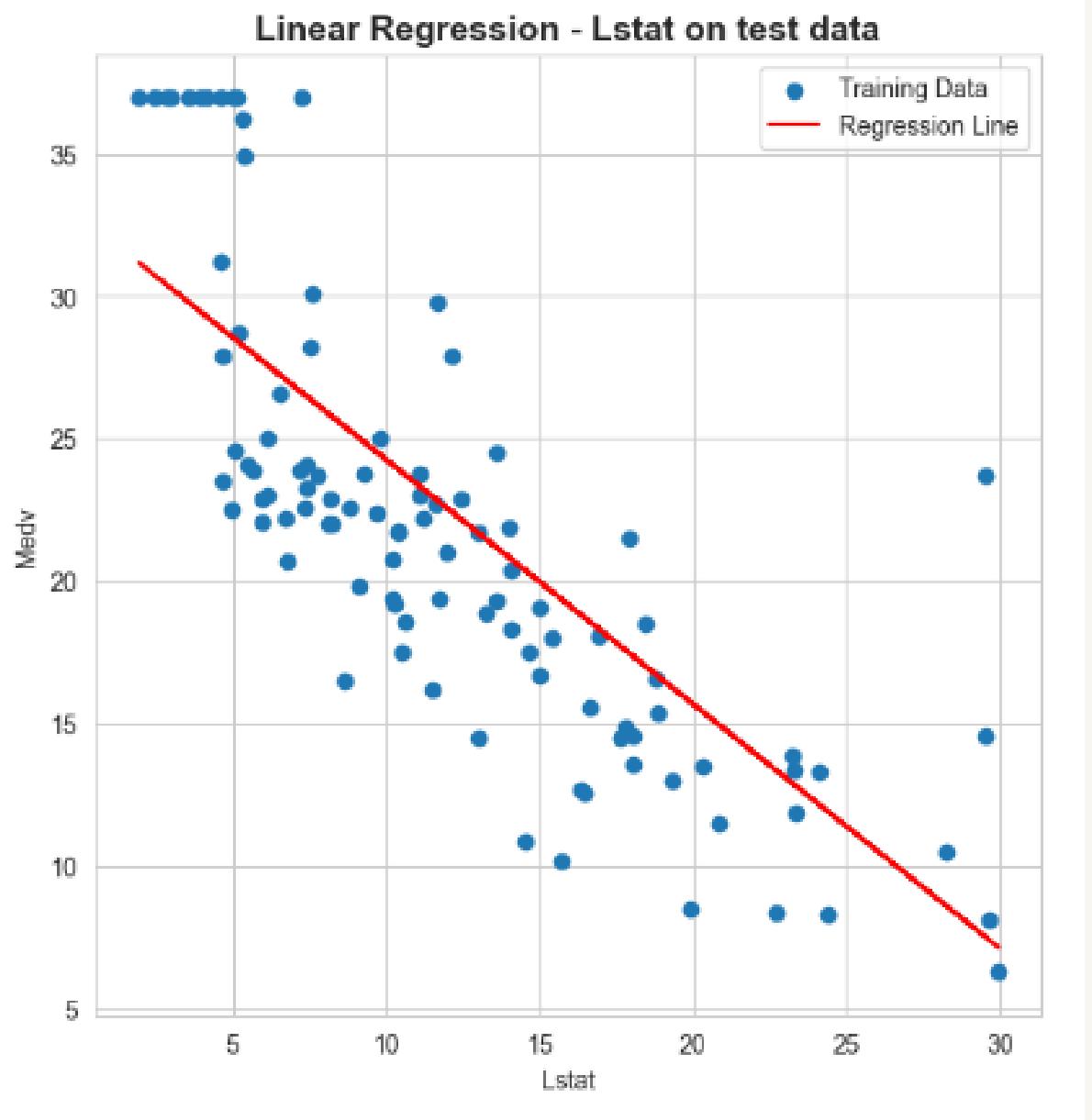
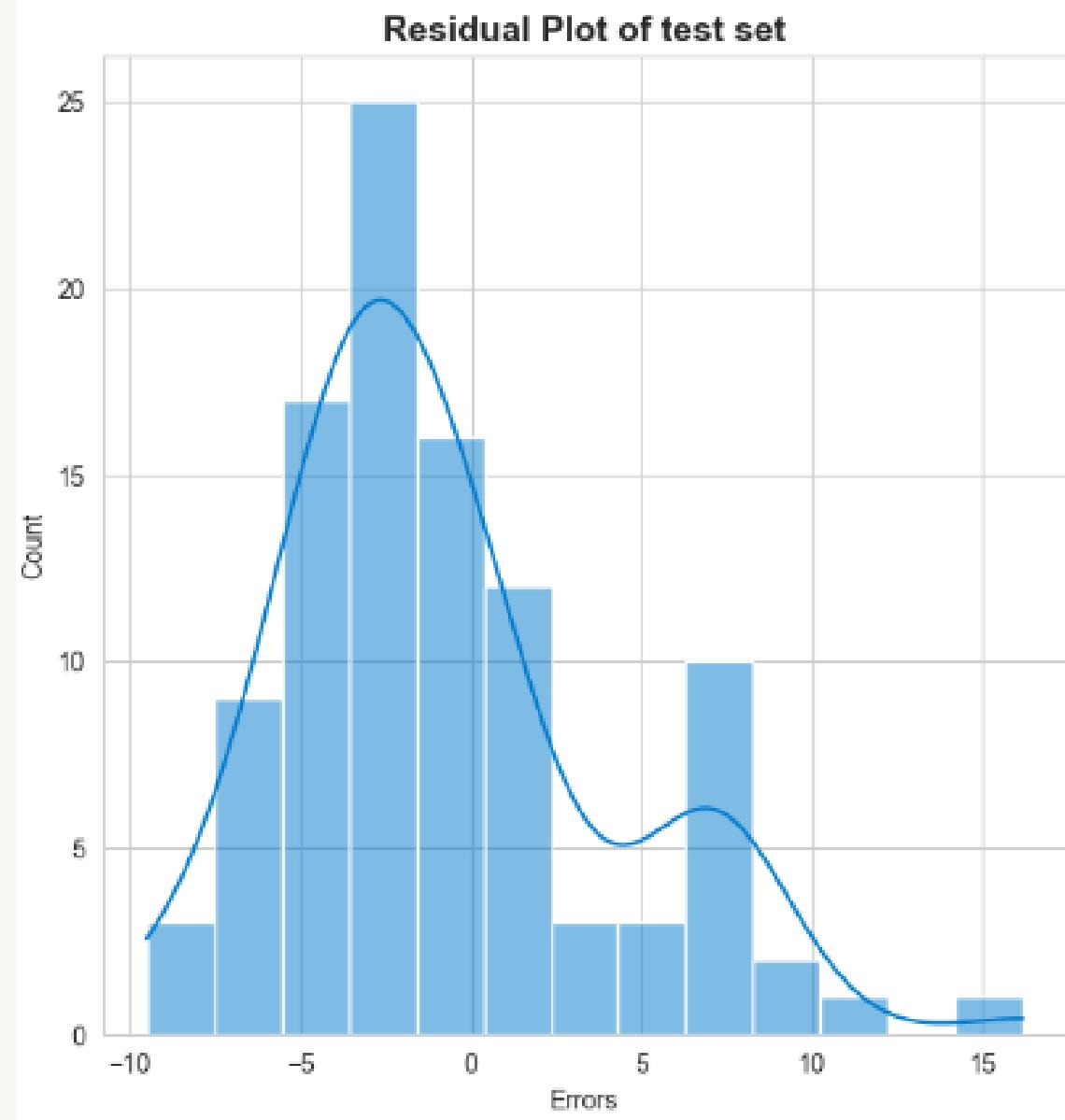
Đơn biến

$$\widehat{Medv} = 32,85 - 0,86 * Lstat \text{ (1000\$)}, \text{ với } R^2 = 0,64$$



Với tập huấn luyện dữ liệu này ngoài giá trị $R^2 = 0.64$, mô hình còn được đánh giá thêm các chỉ số $MSE = 20.5$, $MAE = 3.6$

Với tập thử nghiệm, mô hình ghi nhận $R^2 = 0.61$, $MSE = 23.2$ và $MAE = 3.9$



$$\begin{aligned}\widehat{Medv} &= 32,85 - 0,86 * Lstat \\ &= 32,85 - 0,86 * 10 \\ &= 24,25 \text{ (đơn vị: \$1000)}\end{aligned}$$

Đánh giá

Các chỉ số R^2 , MSE , MAE không chênh lệch quá nhiều so với mô hình huấn luyện, như vậy mô hình này có hiệu suất tương đối cao trên tập dữ liệu mới, không gặp tình trạng overfitting



Hồi quy tuyến tính

Nội dung

1

Mô tả bài toán

2

Tiền xử lý dữ liệu

3

Mô hình đơn biến

4

Mô hình đa biến

Dataset:
BostonHousing.csv

OLS Regression Results						
Dep. Variable:	medv	R-squared:	0.774			
Model:	OLS	Adj. R-squared:	0.770			
Method:	Least Squares	F-statistic:	212.9			
Date:	Wed, 17 Apr 2024	Prob (F-statistic):	3.34e-155			
Time:	01:56:00	Log-Likelihood:	-1367.5			
No. Observations:	506	AIC:	2753.			
Df Residuals:	497	BIC:	2791.			
Df Model:	8					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	35.1243	3.809	9.222	0.000	27.641	42.607
zn	0.0433	0.021	2.092	0.037	0.003	0.084
chas	2.0310	0.654	3.107	0.002	0.747	3.315
nox	-14.4602	2.736	-5.285	0.000	-19.836	-9.085
rm	3.0593	0.351	8.710	0.000	2.369	3.749
dis	-0.9933	0.146	-6.815	0.000	-1.280	-0.707
tax	-0.0037	0.002	-2.469	0.014	-0.007	-0.001
ptratio	-0.6998	0.100	-7.008	0.000	-0.896	-0.504
lstat	-0.5283	0.037	-14.329	0.000	-0.601	-0.456
Omnibus:	57.131	Durbin-Watson:	1.042			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	98.274			
Skew:	0.708	Prob(JB):	4.57e-22			
Kurtosis:	4.629	Cond. No.	1.15e+04			

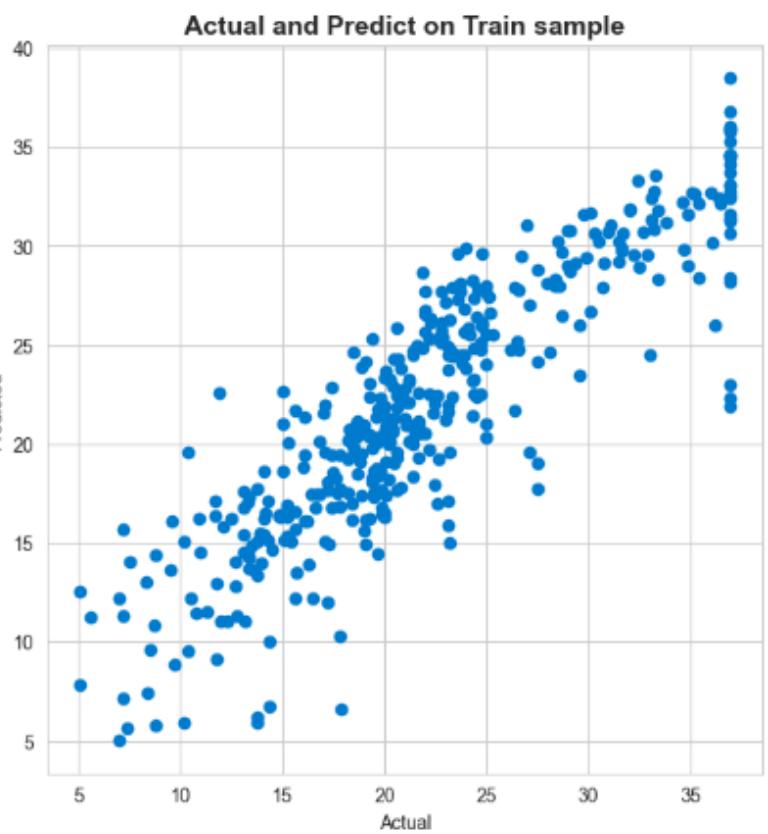
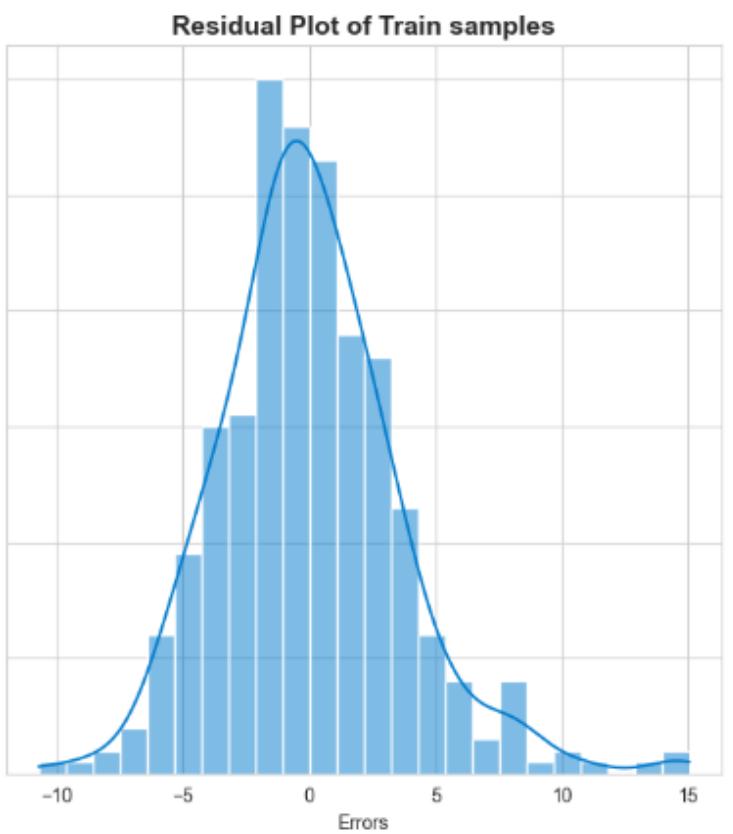
$$\widehat{Medv} = 37.56 + 0.04*Zn + 1.82*Chas - 15.52*Nox + 2.87*Rm - 0.94*Dis - 0.002*Tax - 0.77 *Pratio - 0.53*Lstat, \\ (\$1000)$$

Quá trình lựa chọn biến độc lập sẽ được xem xét lần lượt các giá trị của p-value và VIF (Variance inflation factor). Sau những lần kiểm tra mô hình cuối cùng xác định gồm 8 biến.

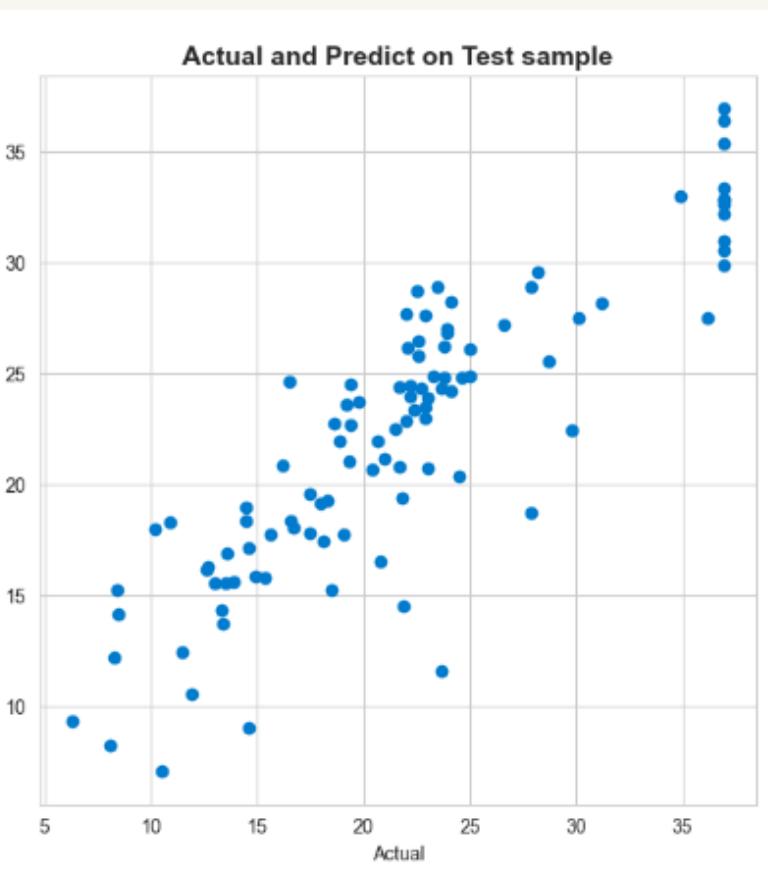
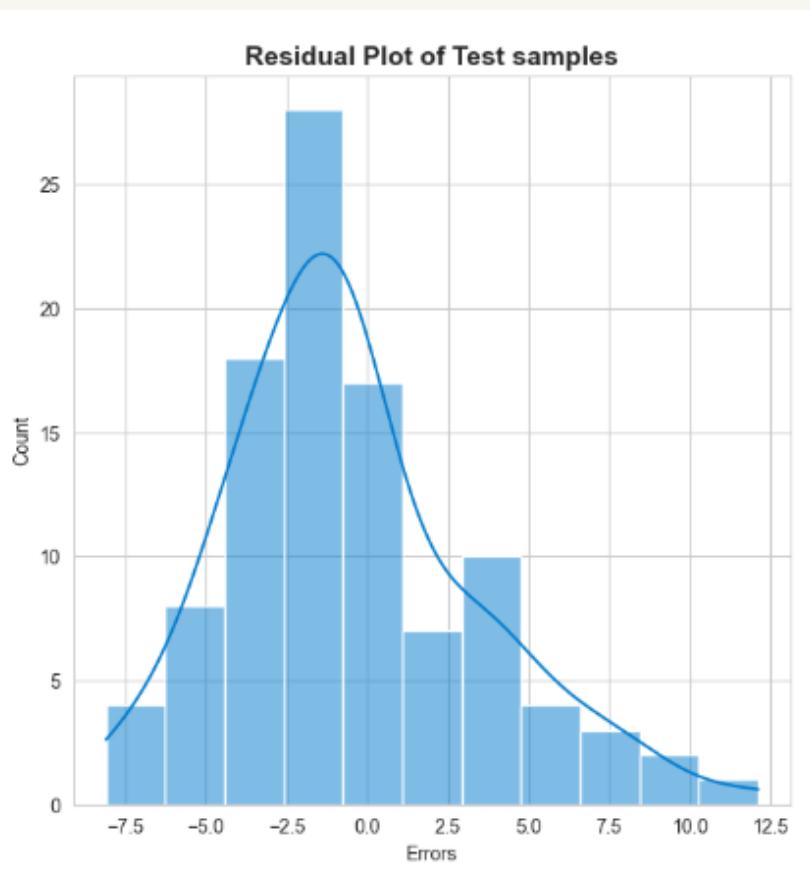
Dự đoán

Thực hiện dự đoán giá nhà với
giá trị của các biến độc lập là
 $Zn = 12$, $Chas = 1$, $Nox = 0.4$,
 $Rm = 5$, $Dis = 7.5$, $Tax = 617$,
 $Ptratio = 16.25$, $Lstat = 3.5$:

$$\begin{aligned} \text{Medv} &= 37.56 + 0.04*12 + \\ &1.82*1 - 15.52*0.4 + 2.87*5 - \\ &0.94*7.5 - 0.002*617 - 0.77 \\ &*16.25 - 0.53*3.5 \\ &= 25.7955 (\$1000) \end{aligned}$$

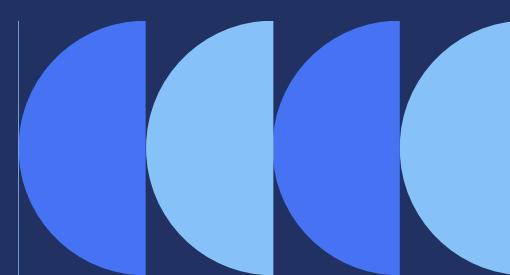


Với tập dữ liệu huấn
luyện ghi nhận giá trị
 $R^2 = 0.78$, $MSE = 12.6$,
 $MAE = 2.7$



Với tập thử nghiệm ghi
nhận giá trị $R^2 = 0.75$,
 $MSE = 14.9$ và $MAE =$
 3.03

Đánh giá





Mô hình hồi quy logistic

Hồi quy Logistic

Nội dung

1 Mô tả bài toán

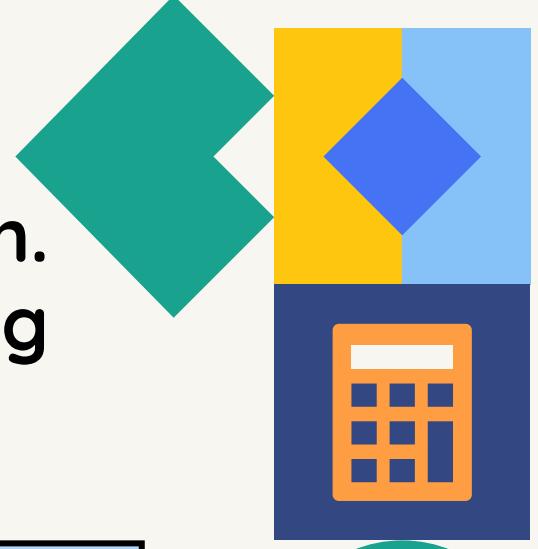
2 Làm sạch và khám phá dữ liệu

3 Chuẩn bị dữ liệu

4 Xây dựng mô hình

Dataset: Lead.csv

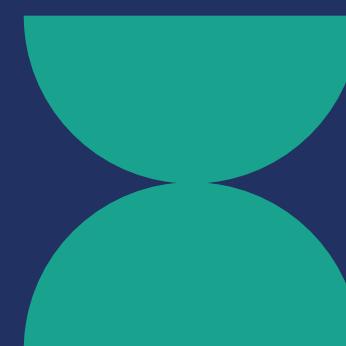
Một công ty giáo dục cung cấp các khóa học trực tuyến. Tỷ lệ chuyển đổi khách hàng tiềm năng điển hình ở công ty này là khoảng 30%.



Thuộc tính	Mô tả
Prospect ID	Một ID duy nhất xác định khách hàng cụ thể.
Lead Number	Một số hoặc mã được gán cho mỗi khách hàng.
Lead Origin	Mã nhận dạng nguồn mà khách hàng được xác định là khách hàng tiềm năng. Bao gồm API, Landing Page, v.v.
Lead Source	Nguồn truy cập. Bao gồm Google, tìm kiếm không phải trả tiền, Olark, v.v.
Do Not Email	Khách hàng lựa chọn có muốn nhận email về khóa học hay không.

Do Not Call	Khách hàng lựa chọn có muốn nhận cuộc gọi về khóa học hay không.
Converted	Khách hàng tiềm năng đã được chuyển đổi thành công hay chưa.
TotalVisits	Tổng số lượt truy cập của khách hàng trên trang web.
Total Time Spent on Website	Tổng thời gian khách hàng dành cho trang web.
Page Views Per Visit	Số trang trung bình trên trang web được xem trong các lượt truy cập.
Last Activity	Hoạt động cuối cùng được thực hiện bởi khách hàng.
Country	Quốc gia của khách hàng.
Specialization	Lĩnh vực ngành mà khách hàng đã làm việc trước đây.

Làm thế nào quá trình
chuyển đổi hiệu quả hơn ?



Hồi quy Logistic

Nội dung

1 Mô tả bài toán

2 Làm sạch và khám phá dữ liệu

3 Chuẩn bị dữ liệu

4 Xây dựng mô hình

Dataset: Lead.csv

Xử lý dữ liệu trùng lặp và không hợp lệ

- Chuyển giá trị “Select” thành dạng dữ liệu thiếu (NaN).
- Loại bỏ cột “Prospect ID” và “Lead Number”, “Last Notable Activity”.

Xử lý dữ liệu và giá trị bị thiếu

- Loại bỏ cột có giá trị thiếu trên 70% và xem xét loại bỏ các cột thiếu khoảng trên 40%.
- Xử lý giá trị thiếu từng cột.

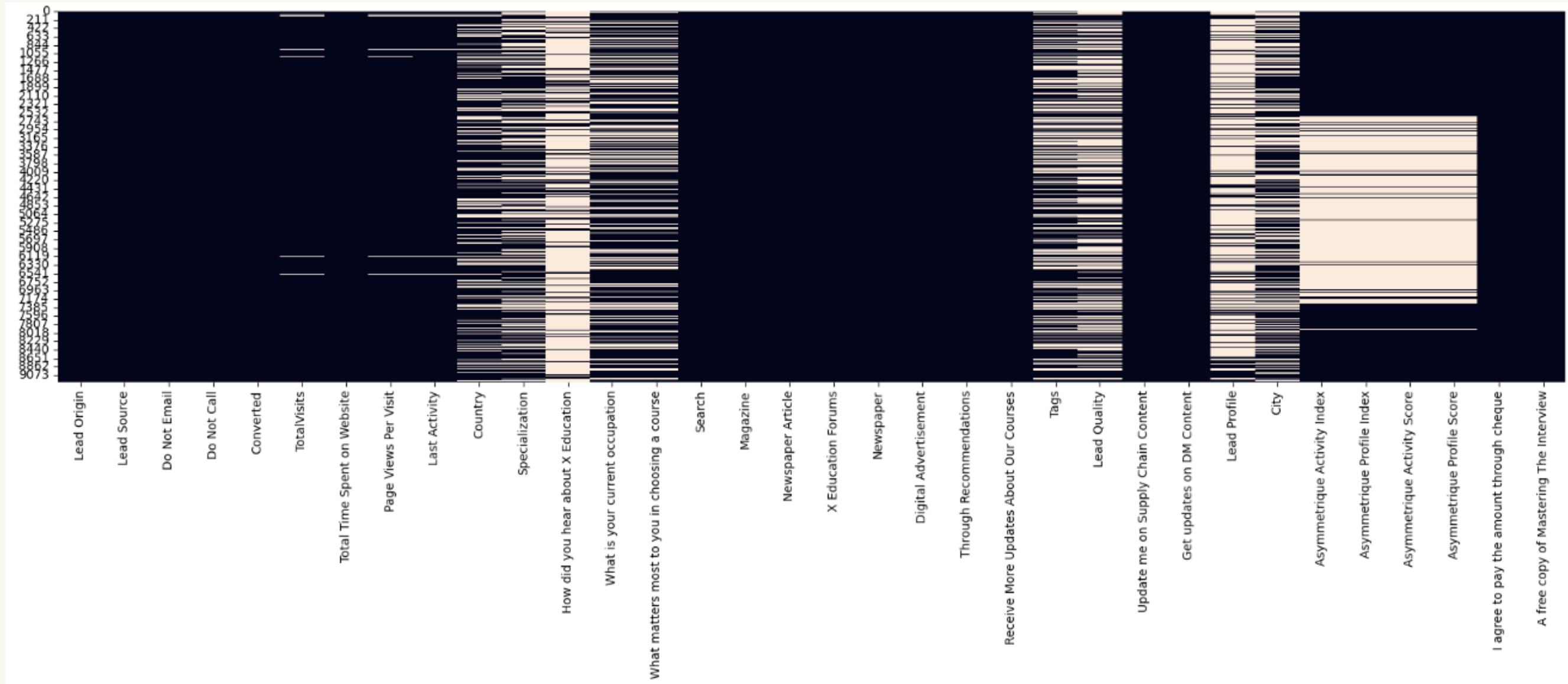
Xử lý các ngoại lệ

- Dữ liệu là số (int64) trừ cột “Converted” sẽ được kiểm tra ngoại lệ.
- Loại những mẫu có giá trị nằm trong 0,5% cao nhất ở mỗi cột.

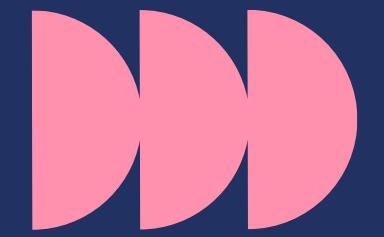
Phân tích khám phá dữ liệu

Trực quan các giá trị thiếu

20 cột không có giá trị thiếu (NaN) và 17 cột có phần trăm giá trị thiếu (NaN) trải dần từ 78,46% đến 0,39%.

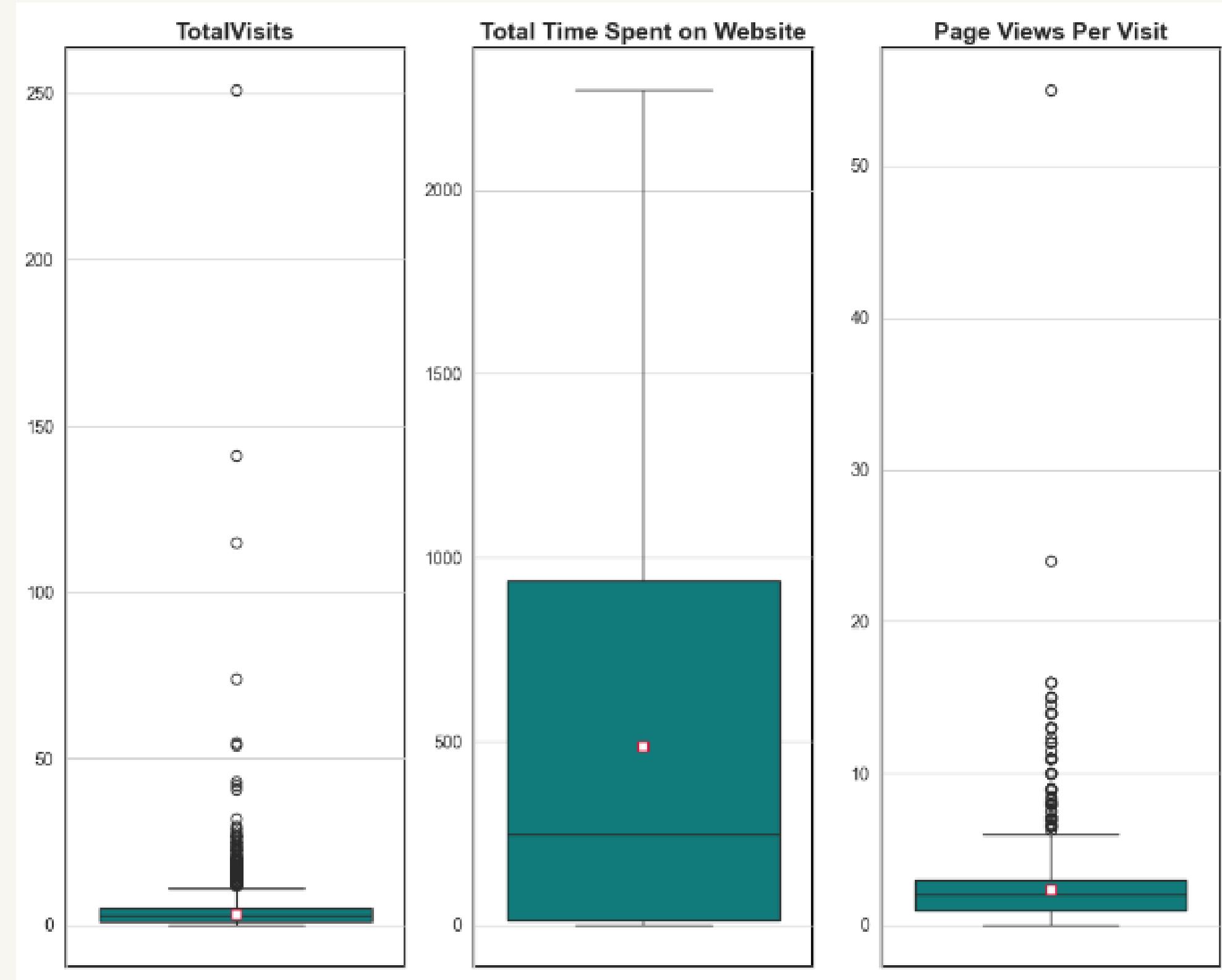
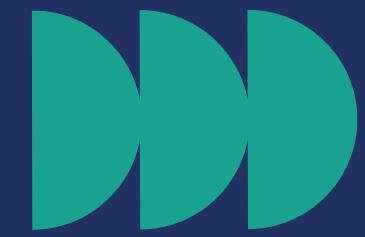


Hình: Biểu đồ trực quan các giá trị bị thiếu



- Trên 70% (“How did you hear about X Education”, “Lead Profile”).
- Từ 70% đến khoảng 40% (“Lead Quality”, “Asymmetrique Activity Index”, “Asymmetrique Profile Index”, “Asymmetrique Activity Score”, “Asymmetrique Profile Score”, “Tags”).

Xử lý ngoại lệ



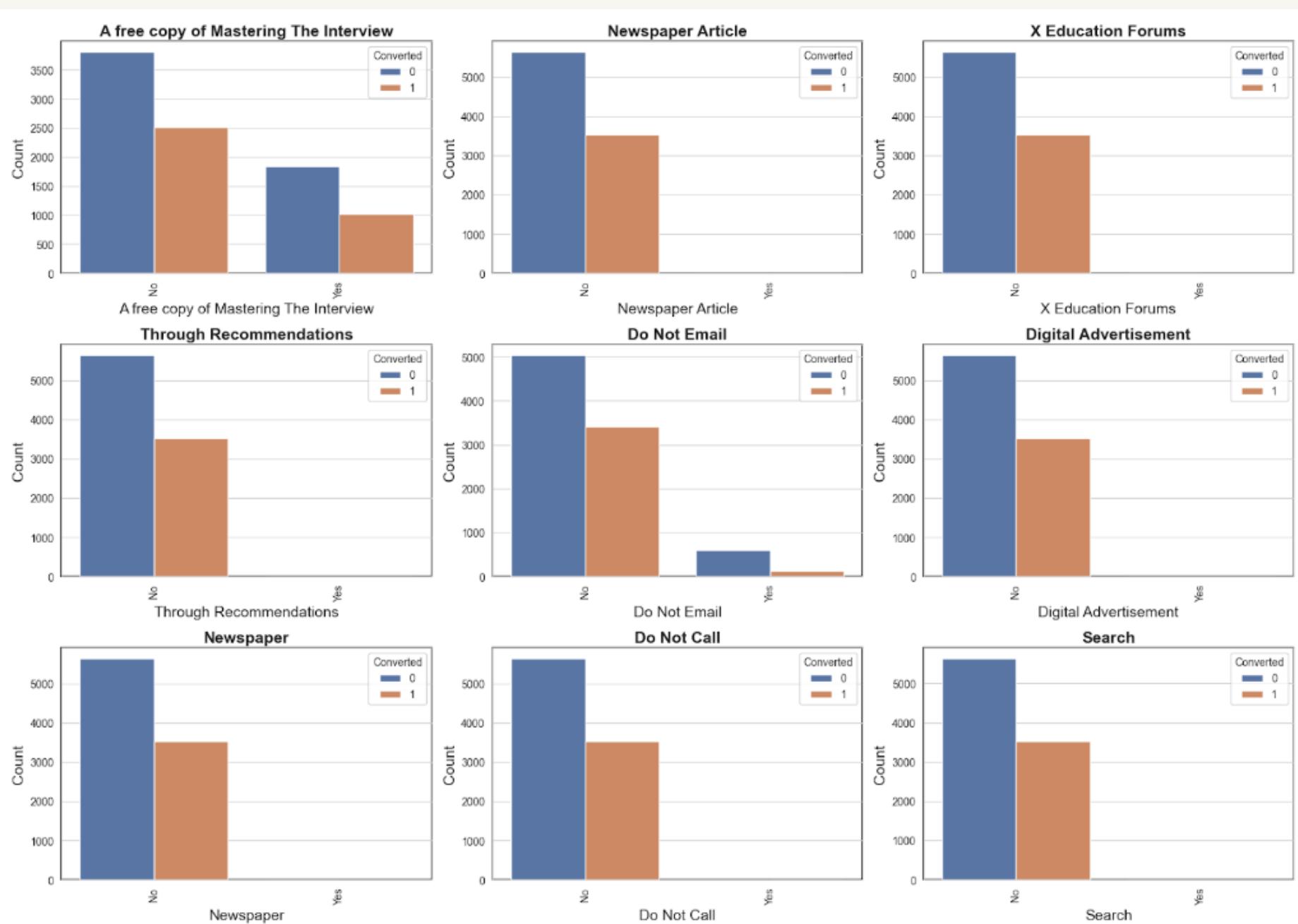
Hình: Biểu đồ boxplot các giá trị số

Hai cột “Total Visits”
và “Page Views Per
Visit” có dữ liệu với số
liệu lớn bất thường

Tổng quan

- Loại bỏ “I agree to pay the amount through cheque”, “Get updates on DM Content”, “Update me on Supply Chain Content”, “Receive More Updates About Our Courses”, “Magazine” vì không có tính đa dạng trong dữ liệu.
- Cột “A free copy of Mastering The Interview” và “Do Not Email” không bị mất cân đối nghiêm trọng và có giá trị dùng để dự đoán tỷ lệ chuyển đổi.

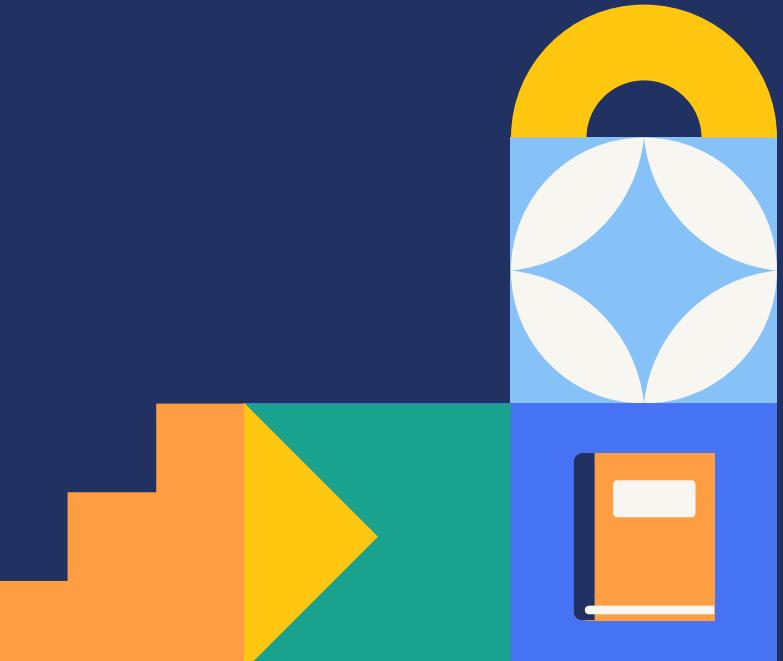
Phân tích khám phá dữ liệu



Nhận xét

- Có tỷ lệ **chuyển đổi khá cao** và có **số lượng khách hàng cao**.
 - Đến từ Mumbai, India.
 - Hoạt động cuối cùng thông thường là mở email.
 - Lý do lựa khóa học hầu hết mong muốn triển vọng nghề nghiệp tốt hơn.
- Có tỷ lệ chuyển đổi khách hàng khoảng 30-35% nhưng **chiếm lượng lớn khách hàng** so với các nguồn khác.
 - Nguồn: API, Landing Page.
 - Từ một số trang như Olark Chat, Organic Search, Direct Traffic, Google.
- Có tỷ lệ **chuyển đổi hơn cao** nhưng **số lượng khách hàng không cao**.
 - Nguồn Lead Add Form.
 - Các trang tiềm kiếm Reference và Welingak Website.
 - Hoạt động cuối cùng nhận tin nhắn.
 - Khách hàng đều có kinh nghiệm làm việc.

Khách hàng dành
nhiều thời gian
hơn trên web có
nhiều khả năng
được **chuyển đổi**
cao hơn.



Hồi quy Logistic

Nội dung

1 Mô tả bài toán

2 Làm sạch và khám phá dữ liệu

3 Chuẩn bị dữ liệu

4 Xây dựng mô hình

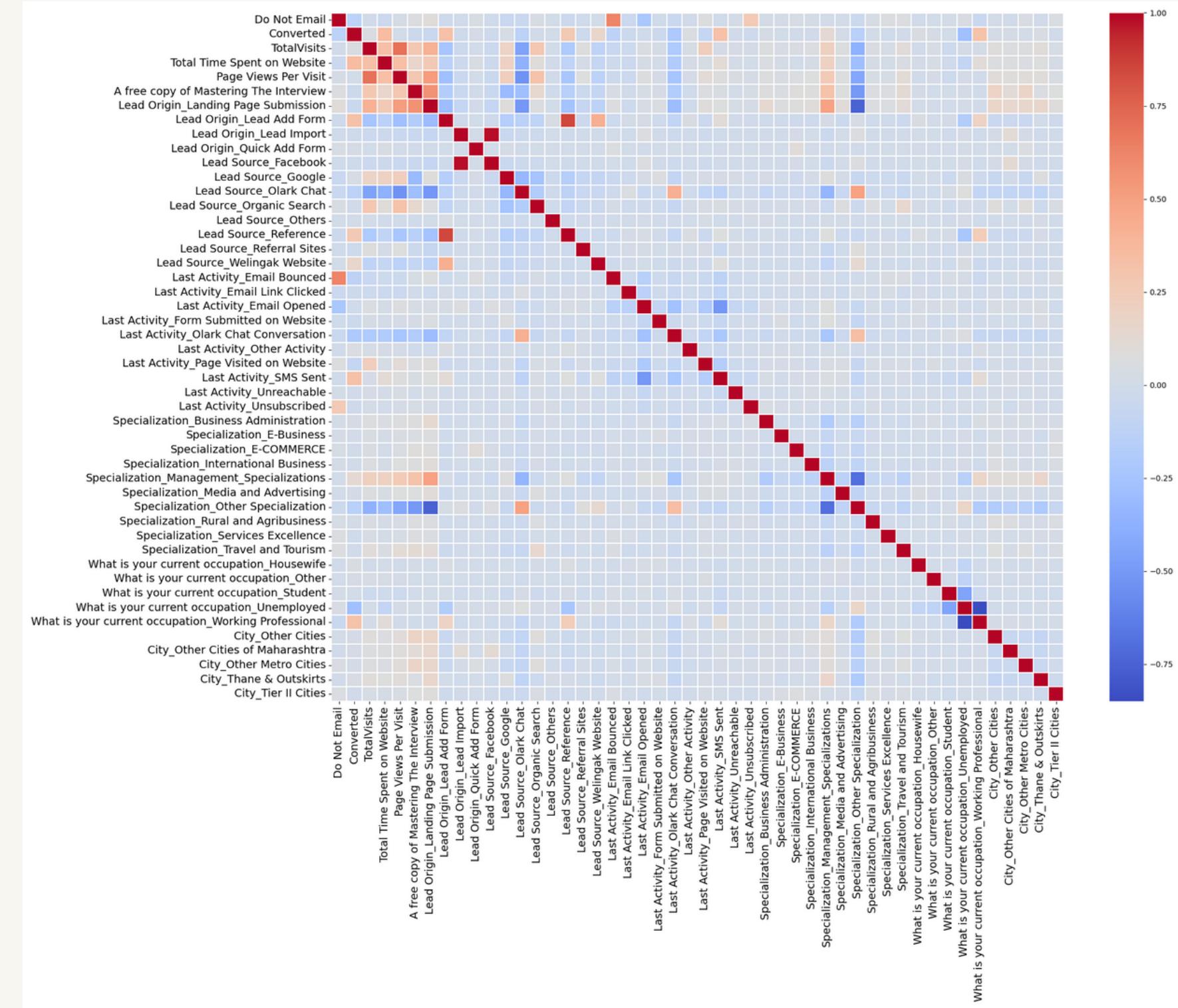
Dataset: Lead.csv

Chuyển đổi biến

1 Nhị phân

2 Danh định

3 Xác định tương quan



Biến độc lập
“Total Time Spent on Website” có mối tương quan cao nhất đối với biến phụ thuộc
“Converted” trong tất cả các biến khác.

Hồi quy Logistic

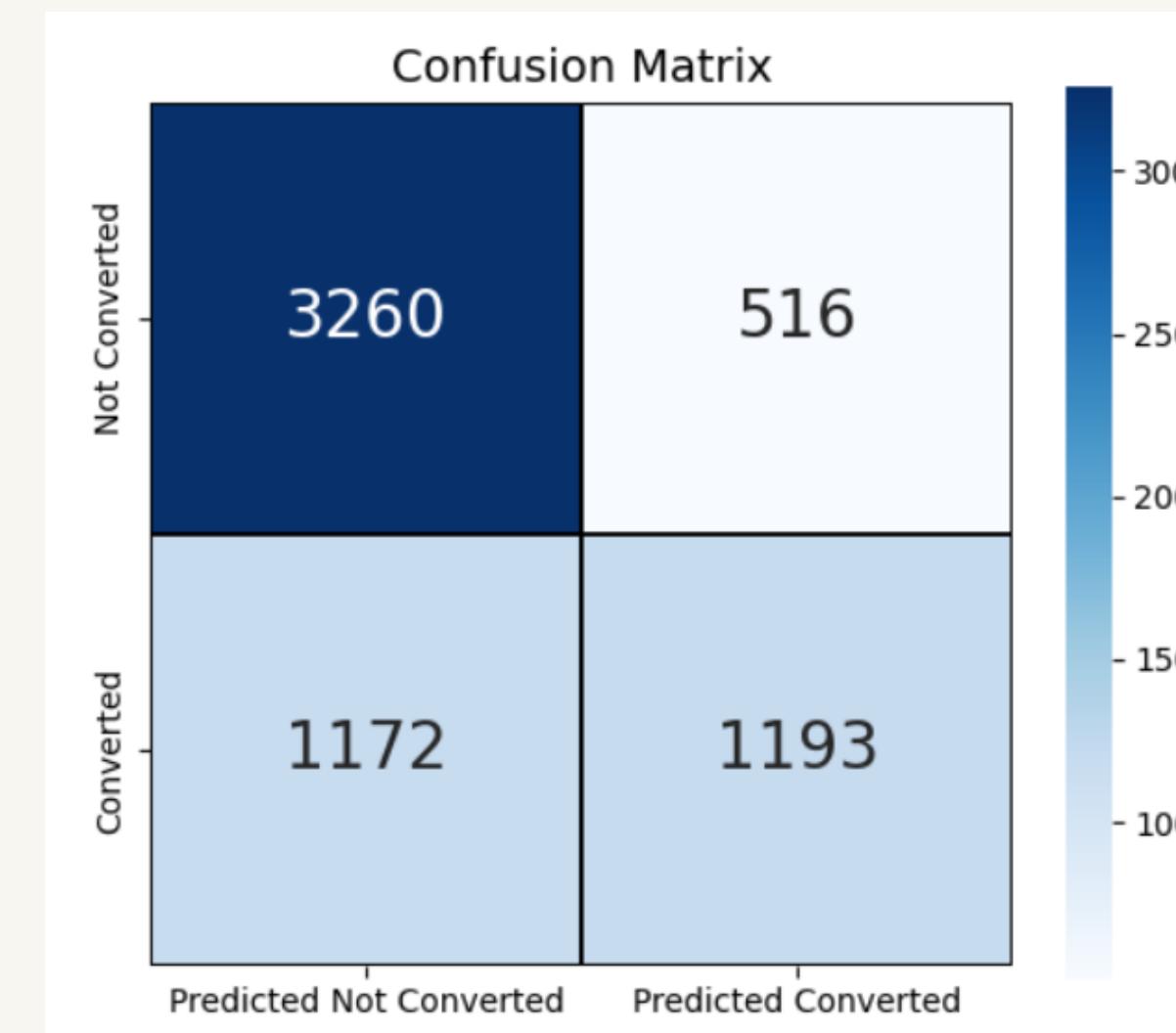
Nội dung

- 1 Mô tả bài toán
- 2 Làm sạch và khám phá dữ liệu
- 3 Chuẩn bị dữ liệu
- 4 Xây dựng mô hình

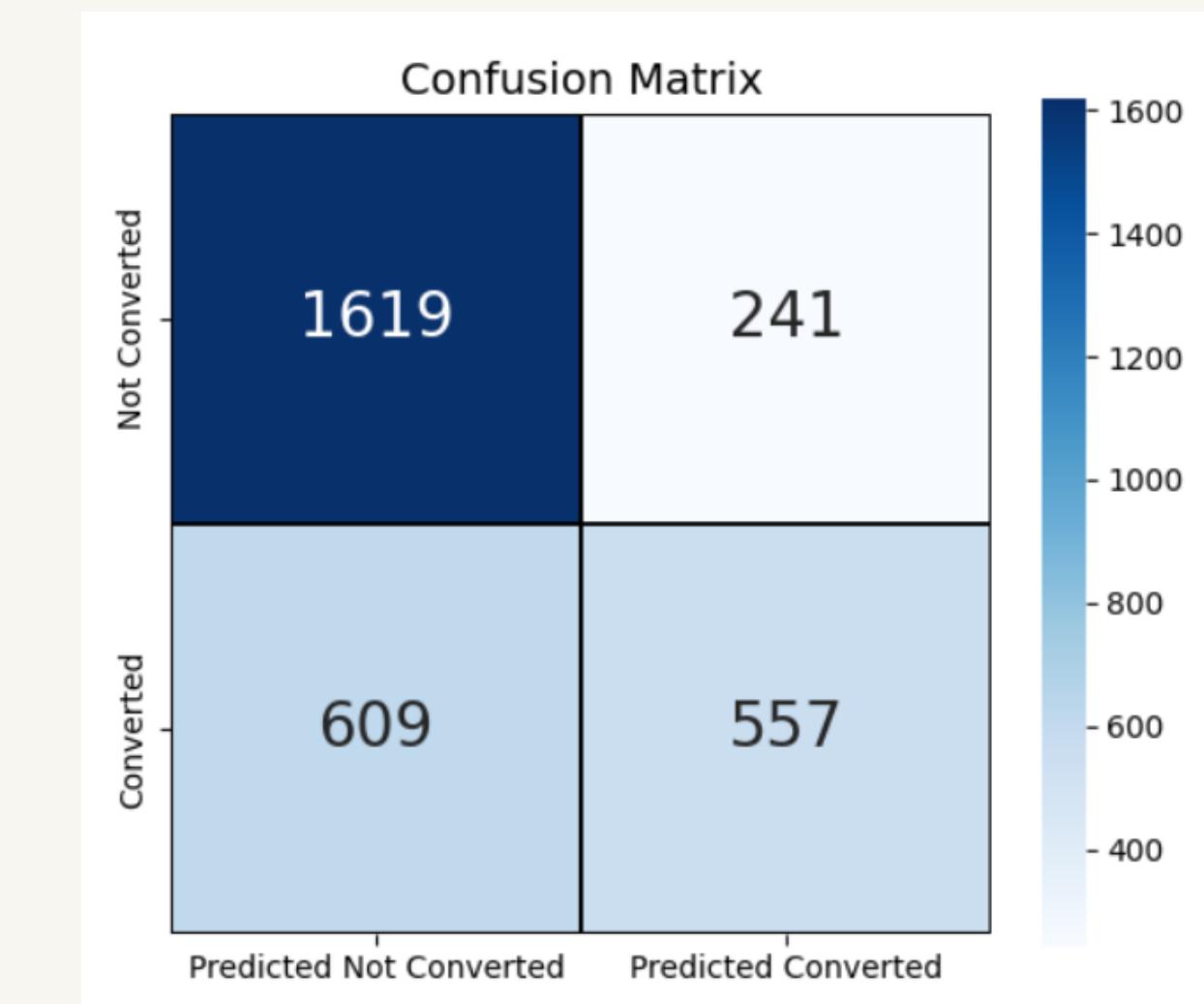
Dataset: Lead.csv

Đơn biến

$$\log(P(Y = 1)) = \log\left(\frac{P(Y=1)}{1 - P(Y=1)}\right) \approx -1.219 + 0.00145 * (\text{Total Time Spent On Website})$$



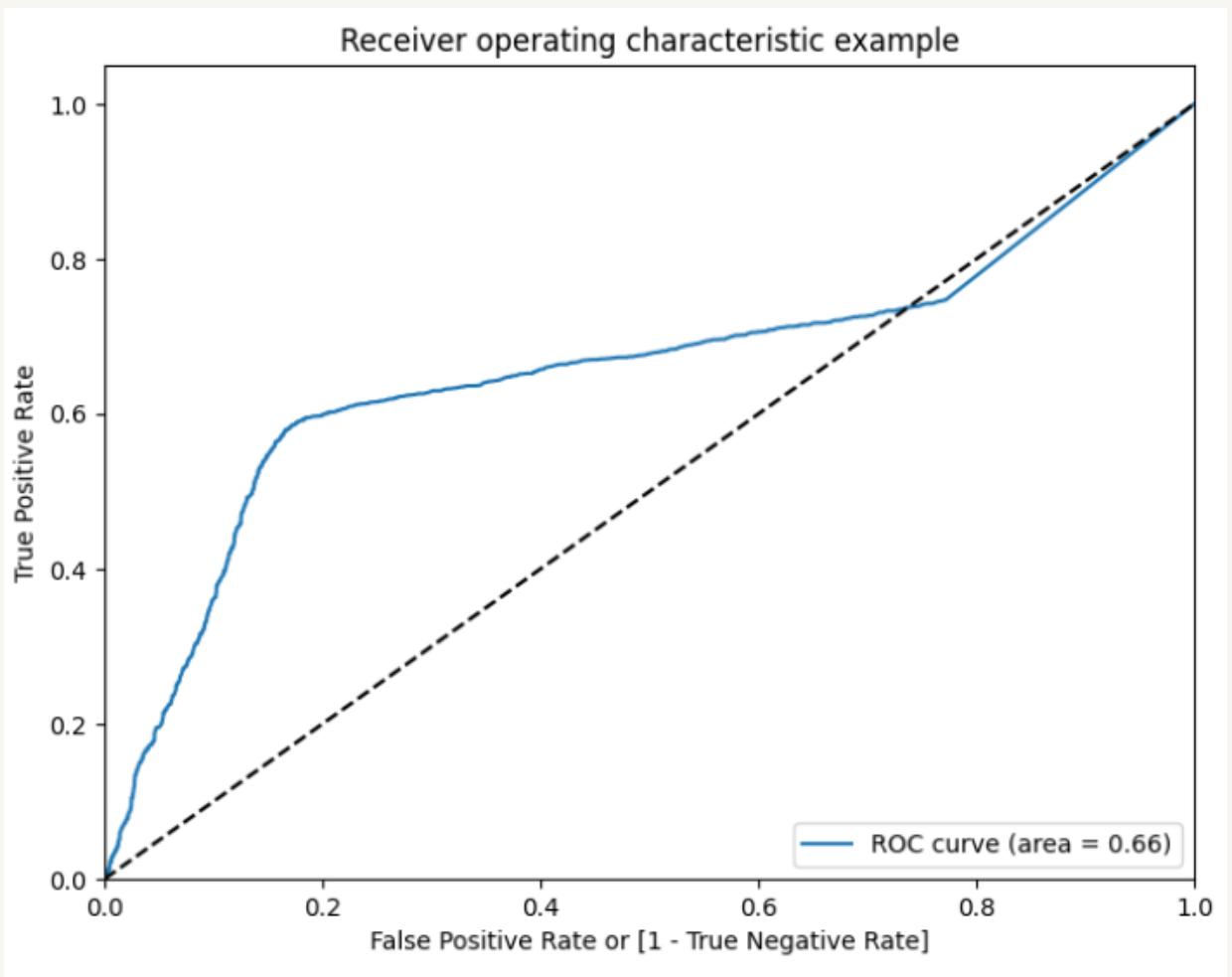
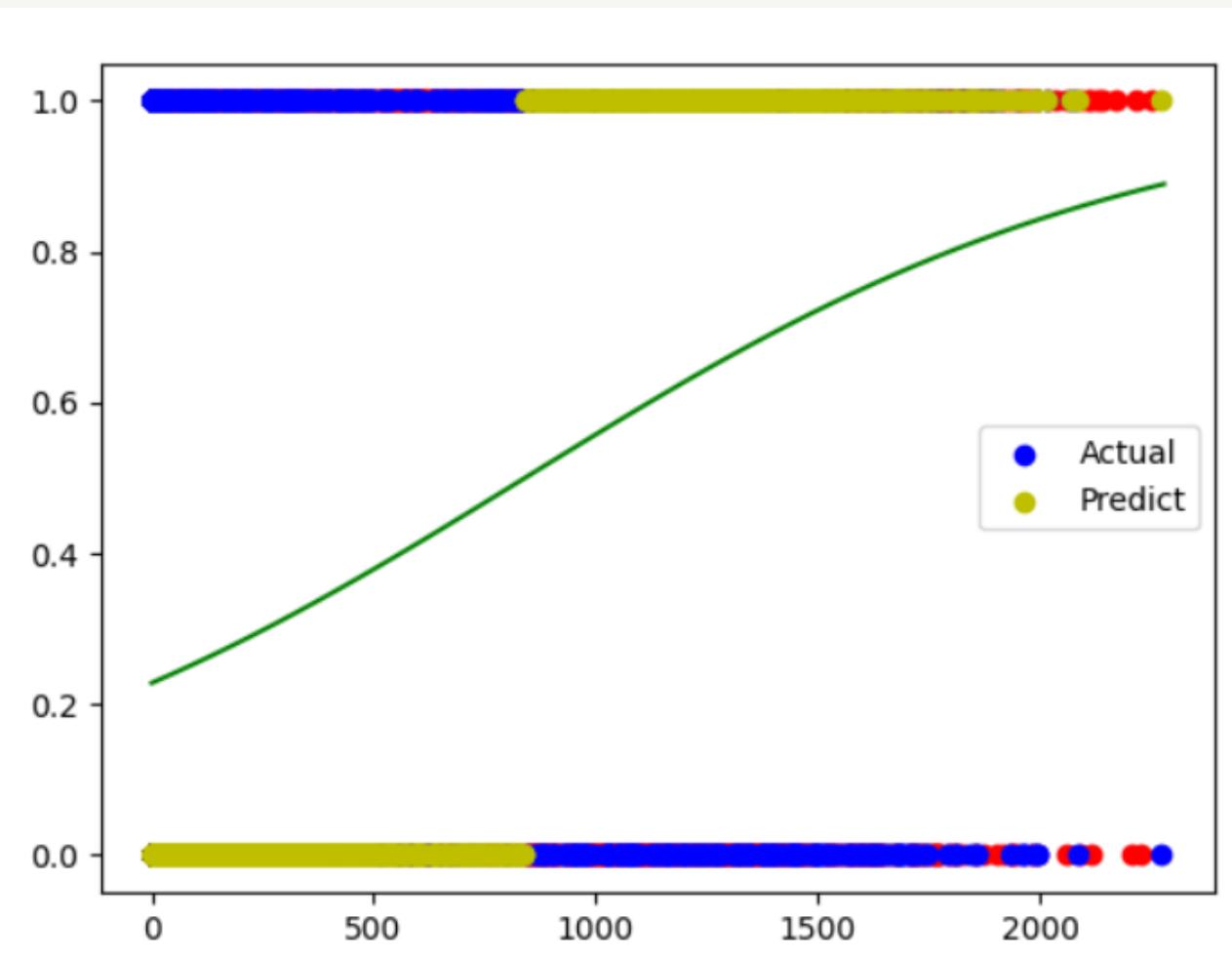
Dữ liệu huấn luyện



Dữ liệu kiểm thử

Việc chuyển đổi của khách hàng có thể bởi ảnh hưởng bởi các yếu tố khác không đặc biệt chỉ dựa vào thời gian trên web.

Đánh giá



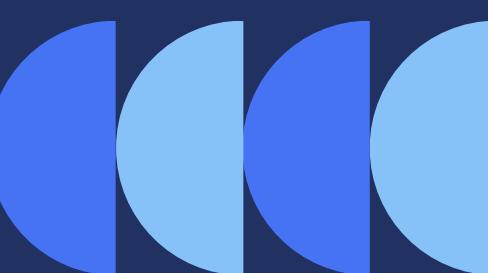
Score	Percent
Accuracy	72.51 %
Sensitivity	50.44 %
Specificity	86.33 %
Precision	69.81 %
Recall	50.44 %

$$\log(P(Y = 1)) = \log\left(\frac{P(Y=1)}{1 - P(Y=1)}\right)$$

$$\approx -1.219 + 0.00145 * (1000) \approx 0.23$$

$$\Rightarrow p = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}} = \frac{1}{1 + e^{-0.23}} \approx 0.56$$

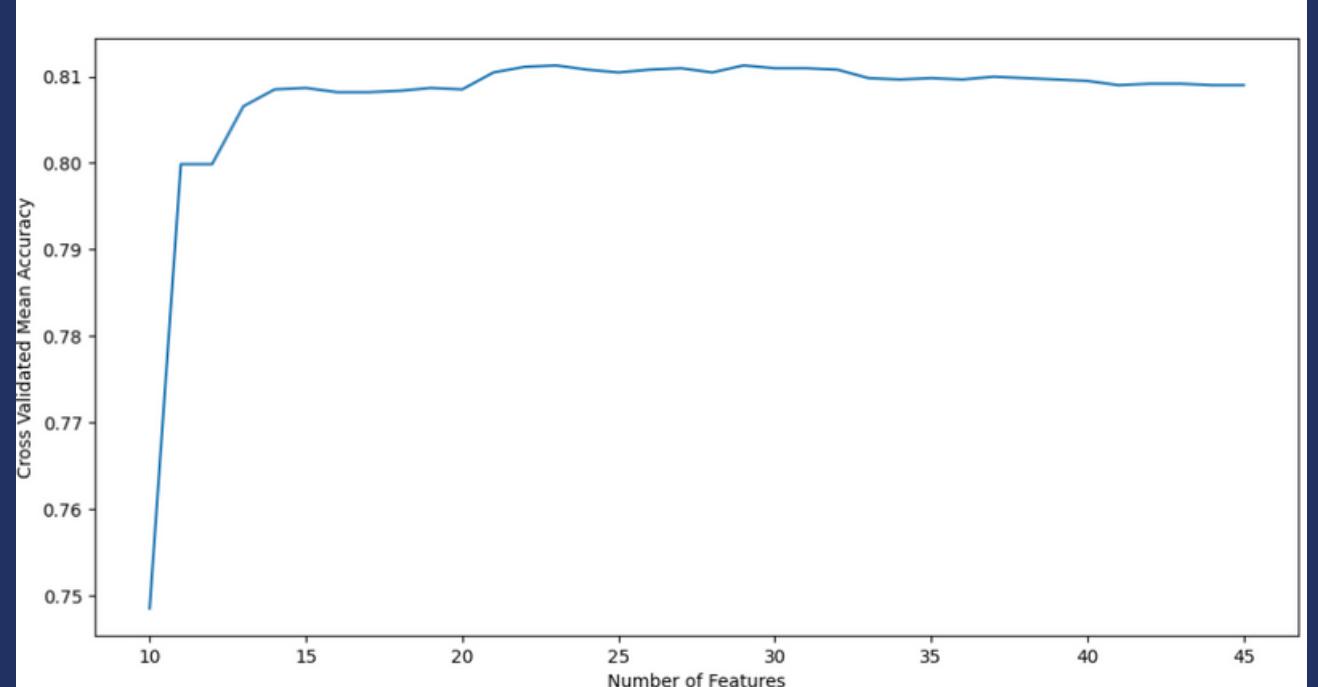
Predicted probability of conversion: 0.5574346526028935



Đa biến

Xây dựng mô hình

Lựa chọn thuộc tính bằng phương pháp Recursive feature elimination (RFE) và Cross Validation (Xác thực chéo)



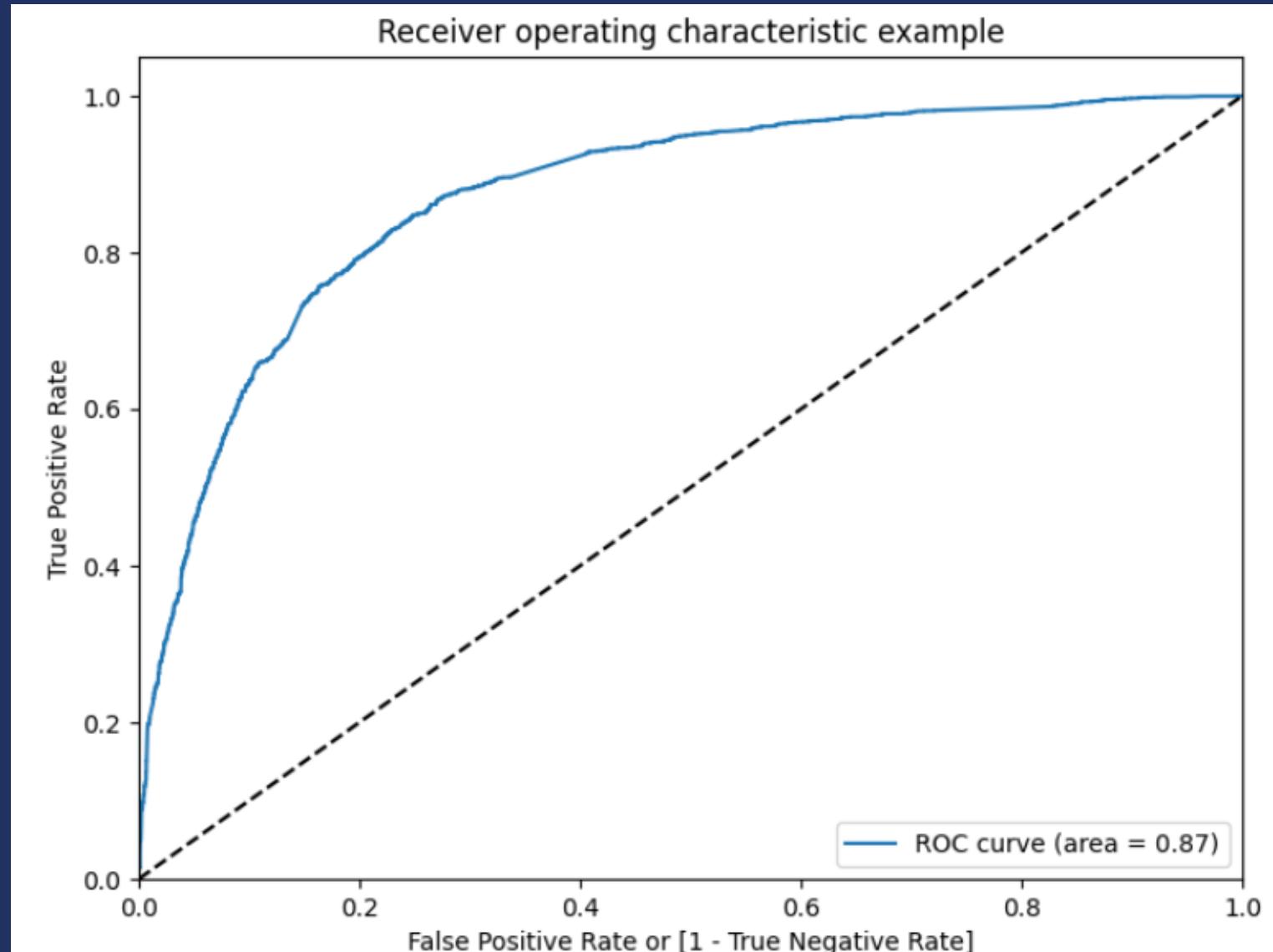
Generalized Linear Model Regression Results						
Dep. Variable:	Converted	No. Observations:	6141			
Model:	GLM	Df Residuals:	6111			
Model Family:	Binomial	Df Model:	29			
Link Function:	Logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-2539.8			
Date:	Wed, 17 Apr 2024	Deviance:	5079.6			
Time:	21:00:38	Pearson chi2:	9.28e+03			
No. Iterations:	21	Pseudo R-squ. (CS):	0.3970			
Covariance Type:	nonrobust					
	coef	std err	z	P> z	[0.025	0.975]
const	0.0887	0.718	0.124	0.902	-1.318	1.495
Do Not Email	-1.2072	0.193	-6.262	0.000	-1.585	-0.829
TotalVisits	0.2581	0.049	5.272	0.000	0.162	0.354
Total Time Spent on Website	1.0537	0.041	25.794	0.000	0.974	1.134
Page Views Per Visit	-0.2262	0.055	-4.095	0.000	-0.335	-0.118
A free copy of Mastering The Interview	-0.3430	0.091	-3.778	0.000	-0.521	-0.165
Lead Origin_Landing Page Submission	-0.9654	0.137	-7.049	0.000	-1.234	-0.697
Lead Origin_Lead Add Form	3.1983	0.235	13.610	0.000	2.738	3.659
Lead Origin_Lead Import	-0.3346	0.438	-0.765	0.444	-1.192	0.523
Lead Source_Olark Chat	0.9463	0.140	6.742	0.000	0.671	1.221
Lead Source_Referral Sites	0.3009	0.305	0.988	0.323	-0.296	0.898
Lead Source_Welingak Website	2.4494	0.760	3.222	0.001	0.959	3.940

Số lượng biến tối ưu là 29 với độ chính xác accuracy = 0,81

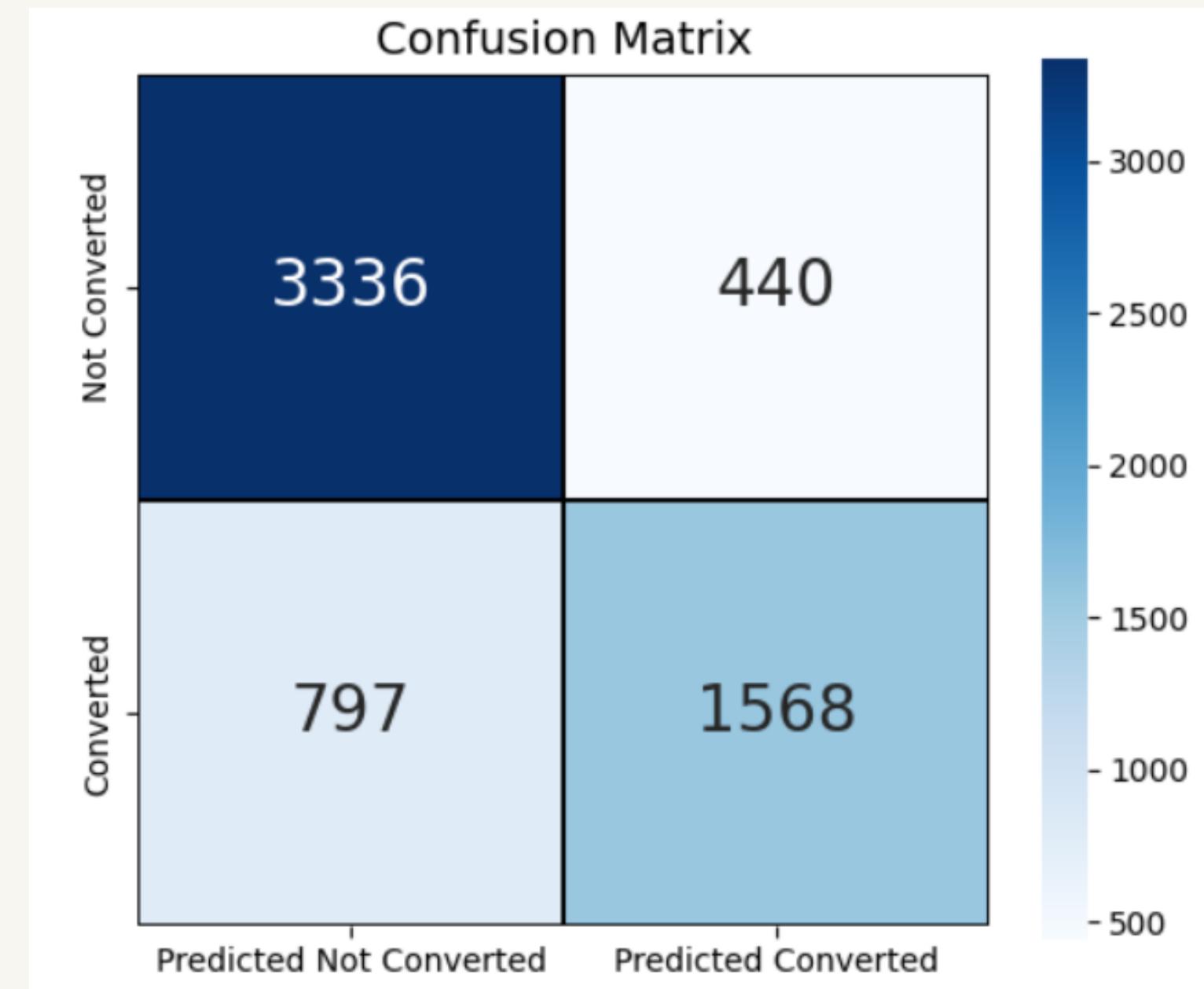
Giảm biến thủ công bằng các chỉ số đánh giá p-value < 0,05 và VIF < 5.

Xác định được 17 biến phù hợp để xây dựng mô hình hồi quy logistic

Nhận xét



Với điểm dự đoán $> 0,5$ sẽ có “chuyển đổi”

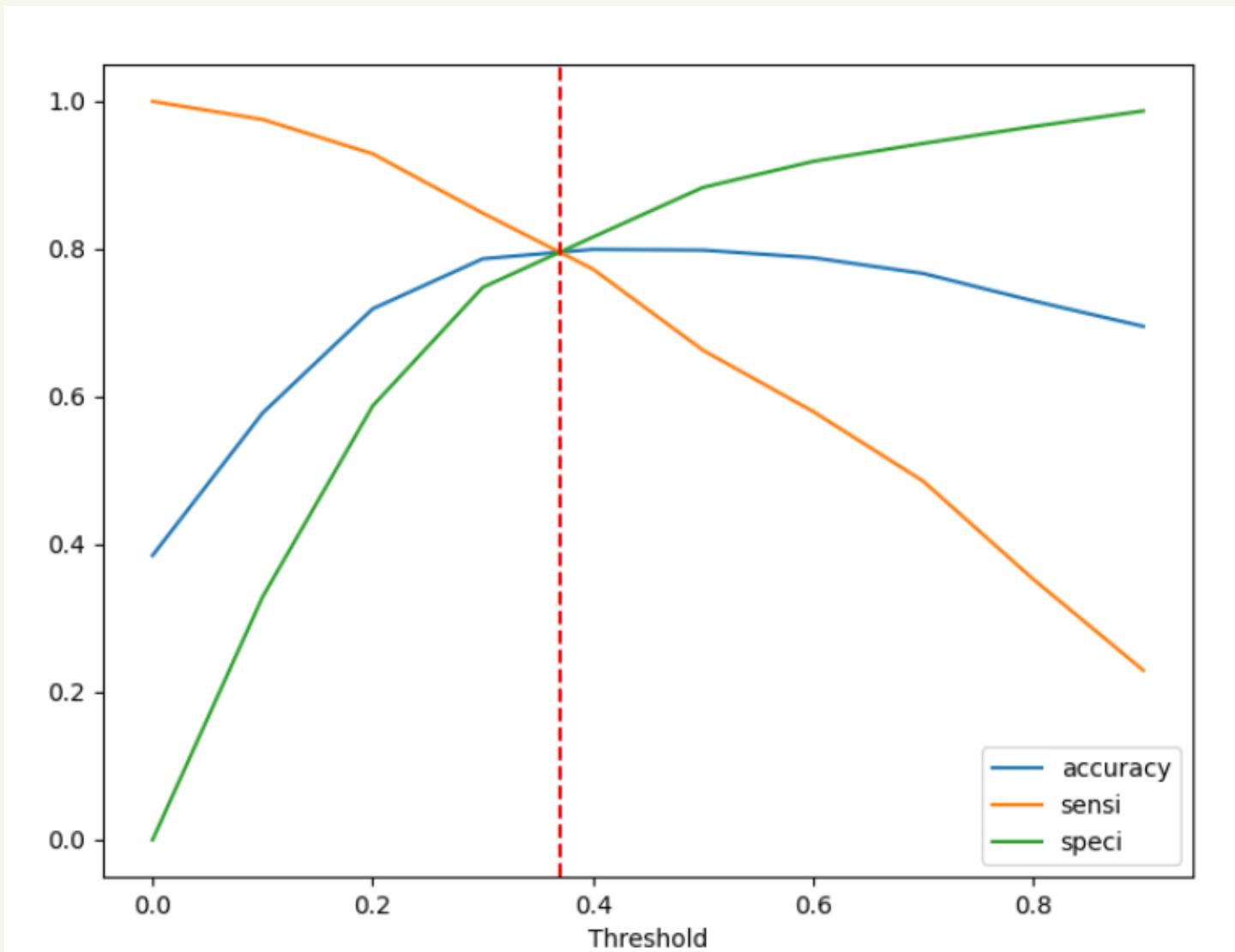


ROC

Đánh giá mô hình trên tập dữ liệu huấn luyện

Nhóm thực hiện xác định ngưỡng với các trường hợp khác nhầm tối ưu hơn mô hình dự đoán

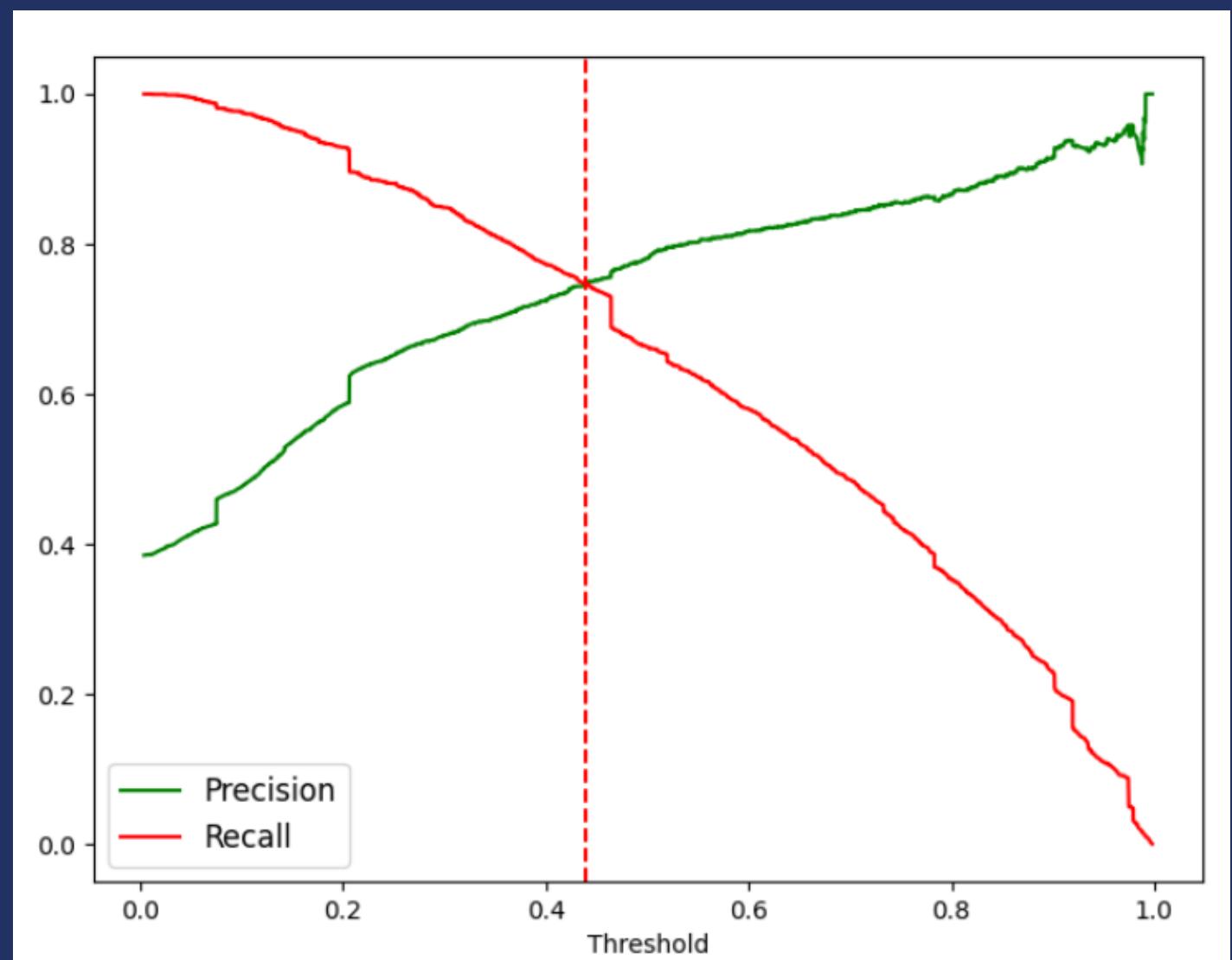
Xác định ngưỡng



Trade-off between Sensitivity & Specificity

Giả định với mục đích xác định chính xác những người sẽ chuyển đổi thành khách hàng tiềm năng.

Trade off between precision and recall



So với ngưỡng cắt 0,37, kết quả của mô hình đã thay đổi như sau

- True Positive đã giảm
- True Negative đã tăng
- False Negative đã tăng
- False Positive đã giảm



Kết luận

Confusion Matrix	
Not Converted	Converted
Predicted Not Converted	1482
Predicted Converted	947

Variable	Coefficient
Specialization_Other Specialization	-1.428037
Do Not Email	-1.153913
Lead Origin_Landing Page Submission	-1.079894
A free copy of Mastering The Interview	-0.322257
Page Views Per Visit	-0.216175
TotalVisits	0.231162
Last Activity_Page Visited on Website	0.496765
Last Activity_Email Link Clicked	0.714844
Lead Source_Olark Chat	0.838332
Last Activity_Unreachable	0.940433
Total Time Spent on Website	1.059700
Last Activity_Email Opened	1.163568
Last Activity_Unsubscribed	1.262775
Lead Source_Welingak Website	1.645910
Last Activity_Other Activity	1.964186
Last Activity_SMS Sent	2.369817

$$\Rightarrow \text{logit}(p) = 0.23 * (0) + 0.5 * (0) + 0.71 * (0) + 0.84 * (0) + 0.94 * (1) + 1.06 * (5) + 1.16 * (1) + 1.26 * (0) + 1.65 * (0) + 1.96 * (0) + 2.37 * (0) + 3.2 * (1) - 1.43 * (1) - 1.15 * (1) - 1.08 * (1) - 0.32 * (0) - 0.22 * (2) - 0.99 = 1,15$$

$$\Rightarrow p = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}} = \frac{1}{1 + e^{-1,15}} \approx 0,76$$

Do đó, xác suất rằng khách hàng sẽ chuyển đổi là khoảng 76%.

Predicted probability of conversion: [0.75954471]

Đánh giá

Score	Percent
Accuracy	80.56 %
Sensitivity	74.71 %
Specificity	84.22 %
Precision	74.78 %
Recall	74.71 %





Phân tích dữ liệu

chuỗi thời gian

1

ARIMA

2

Holt Winter



Dữ liệu thời gian

Nội dung

1 Thu thập dữ liệu
Mô tả bài toán

2 Phân tích khám phá
dữ liệu

3 Chuyển đổi dữ liệu

4 Xây dựng mô hình

5 Mô hình Holt-Winters

Dữ liệu được thu thập từ trang “Dữ liệu lịch sử” và thu thập các dữ liệu về giá chứng khoán của tập đoàn VinGroup với mã chứng khoán VIC.

Thuộc tính	Mô tả
Date	Ngày
Close_price	Giá đóng cửa (nghìn VNĐ)
Adjusted_price	Giá điều chỉnh (nghìn VNĐ)
Change	Tỷ lệ thay đổi
Auction_weight	Khối lượng giao dịch khớp lệnh
Auction_price	Giá trị giao dịch khớp lệnh (tỷ VNĐ)

Xây dựng một mô hình giúp dự đoán giá đóng cửa của mã cổ phiếu VIC trong tương lai thông qua dữ liệu lịch sử giao dịch của tập đoàn.

Put_through_weight	Khối lượng giao dịch thỏa thuận
Put_through_price	Giá trị giao dịch thỏa thuận (tỷ VNĐ)
Open_price	Giá mở cửa (nghìn VNĐ)
Highest_price	Giá cao nhất (nghìn VNĐ)
Lowest_price	Giá thấp nhất (nghìn VNĐ)

Trong 2 mô hình **ARIMA** và **Holt-Winters**, mô hình nào mang lại **hiệu quả dự đoán cao hơn** ?

Dữ liệu thời gian

Nội dung

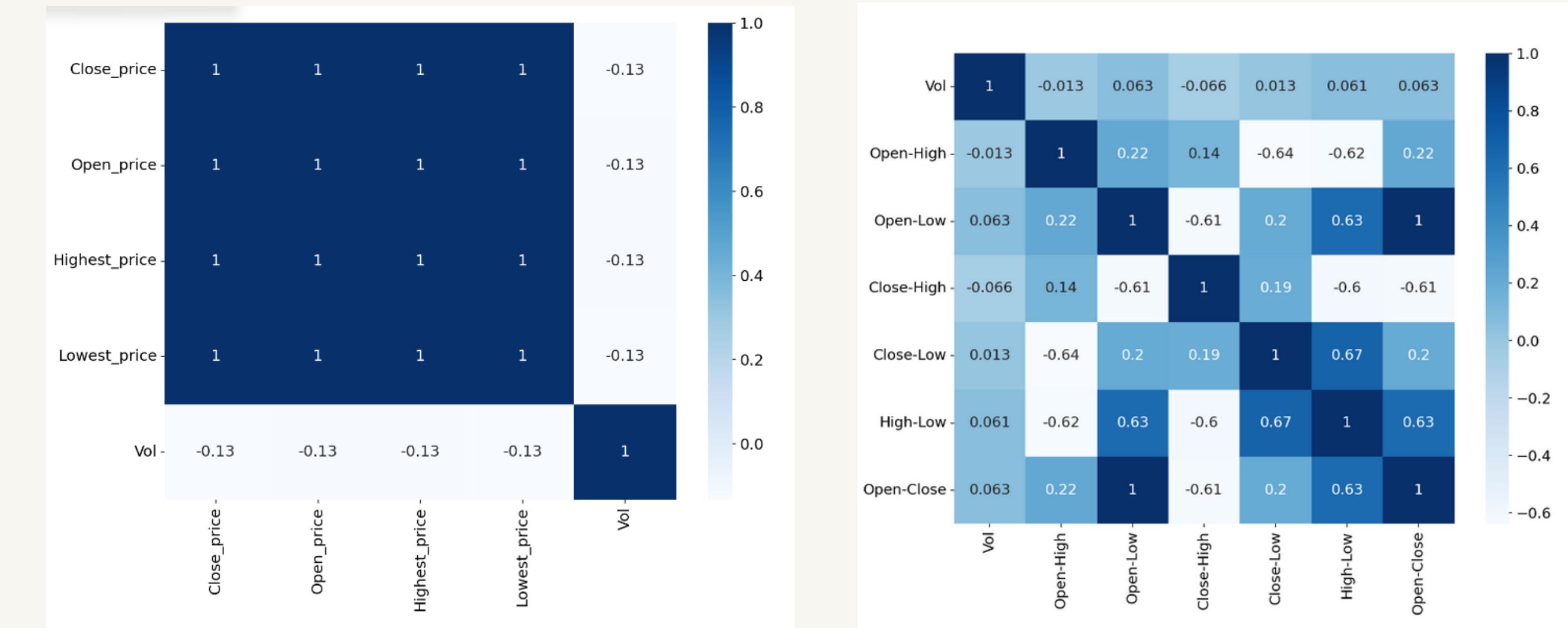
1 Thu thập dữ liệu
Mô tả bài toán

2 Phân tích khám phá
dữ liệu

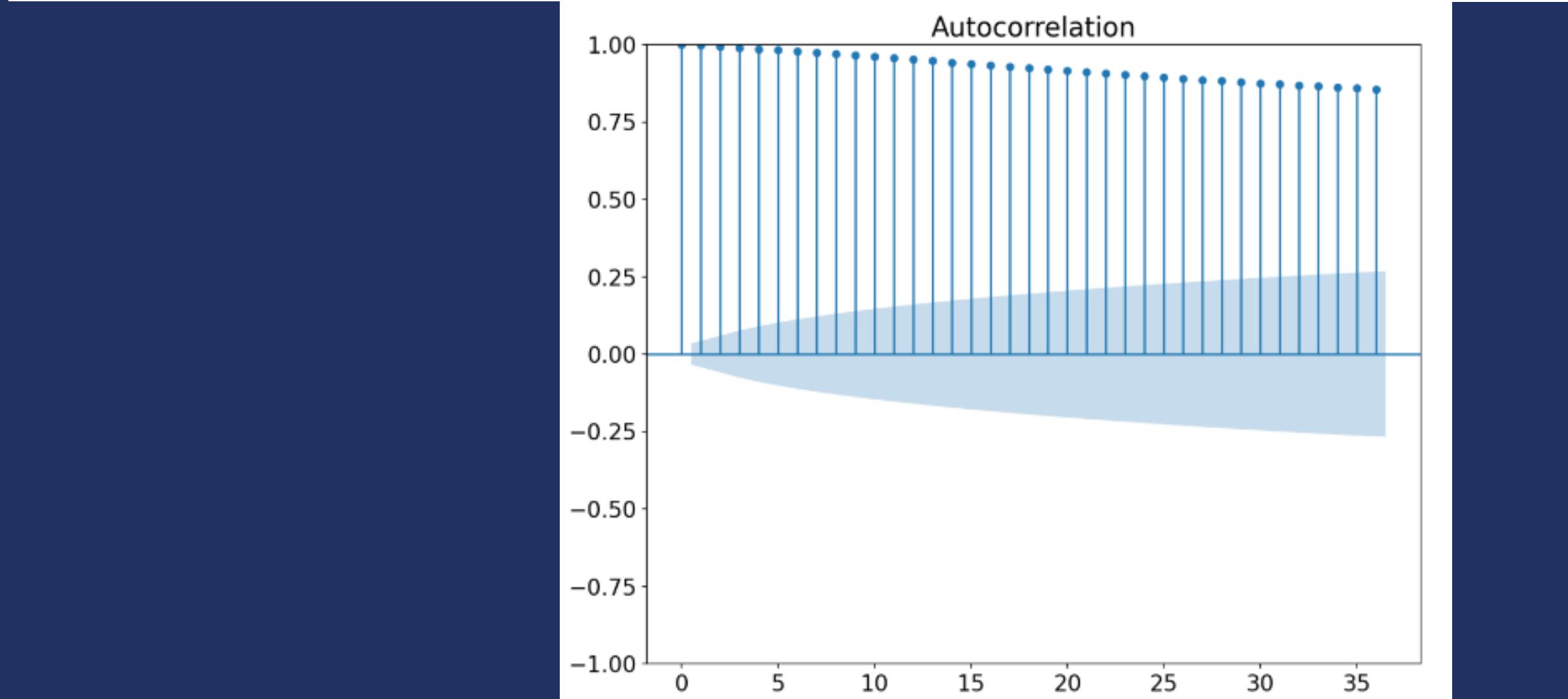
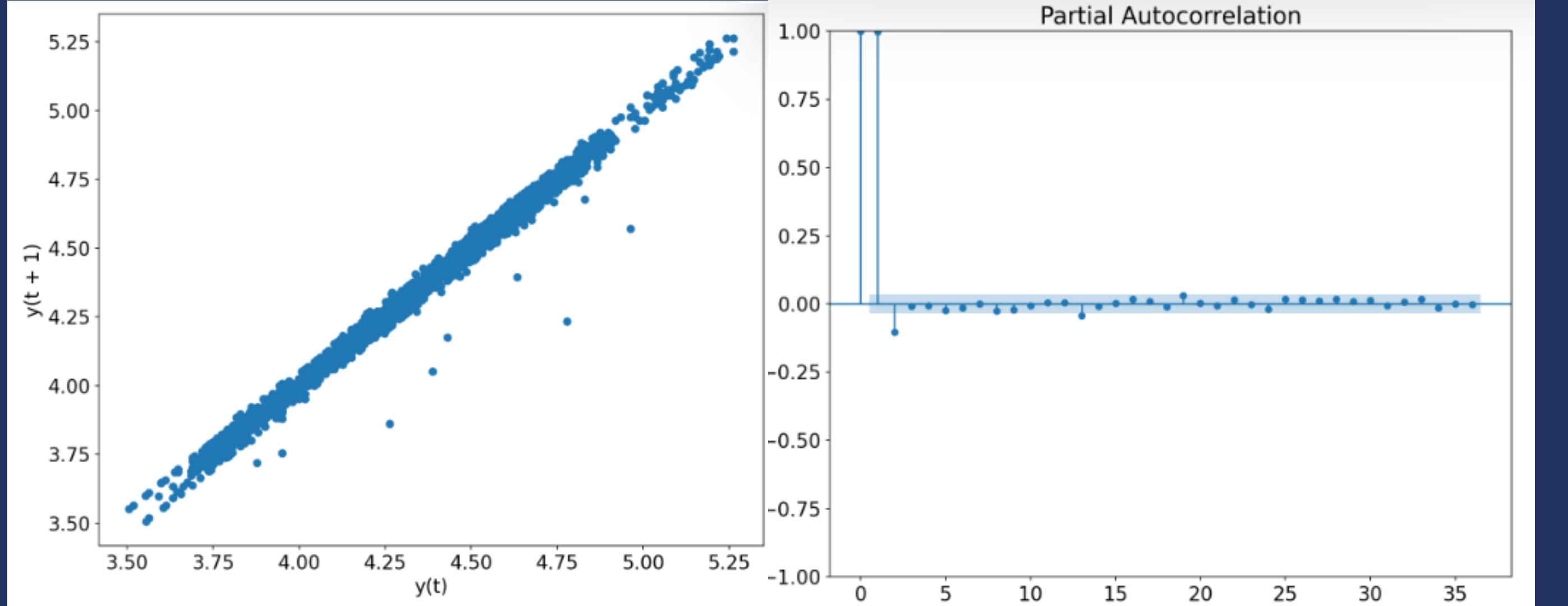
3 Chuyển đổi dữ liệu

4 Xây dựng mô hình

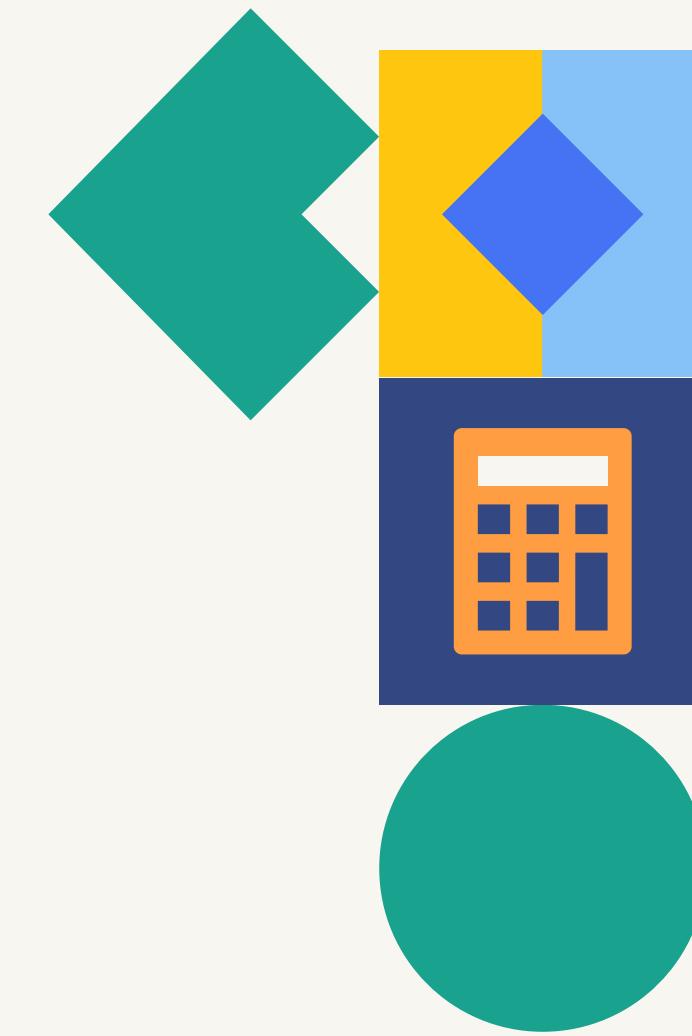
5 Mô hình Holt-Winters



- Tạo cột mới ‘Vol’ là số lượng giao dịch được tính dựa vào tổng ‘Auction_weight’ và ‘Put_through_weight’ và các cột mới đánh giá chênh lệch sử các giá từng thời điểm.
- Nhận xét: Các giá trị của hệ số tương quan không cao đến mức có thể kết luận rằng có mối tương quan mạnh giữa các biến.



Tương
quan



Dữ liệu thời gian

Nội dung

1 Thu thập dữ liệu
Mô tả bài toán

2 Phân tích khám phá
dữ liệu

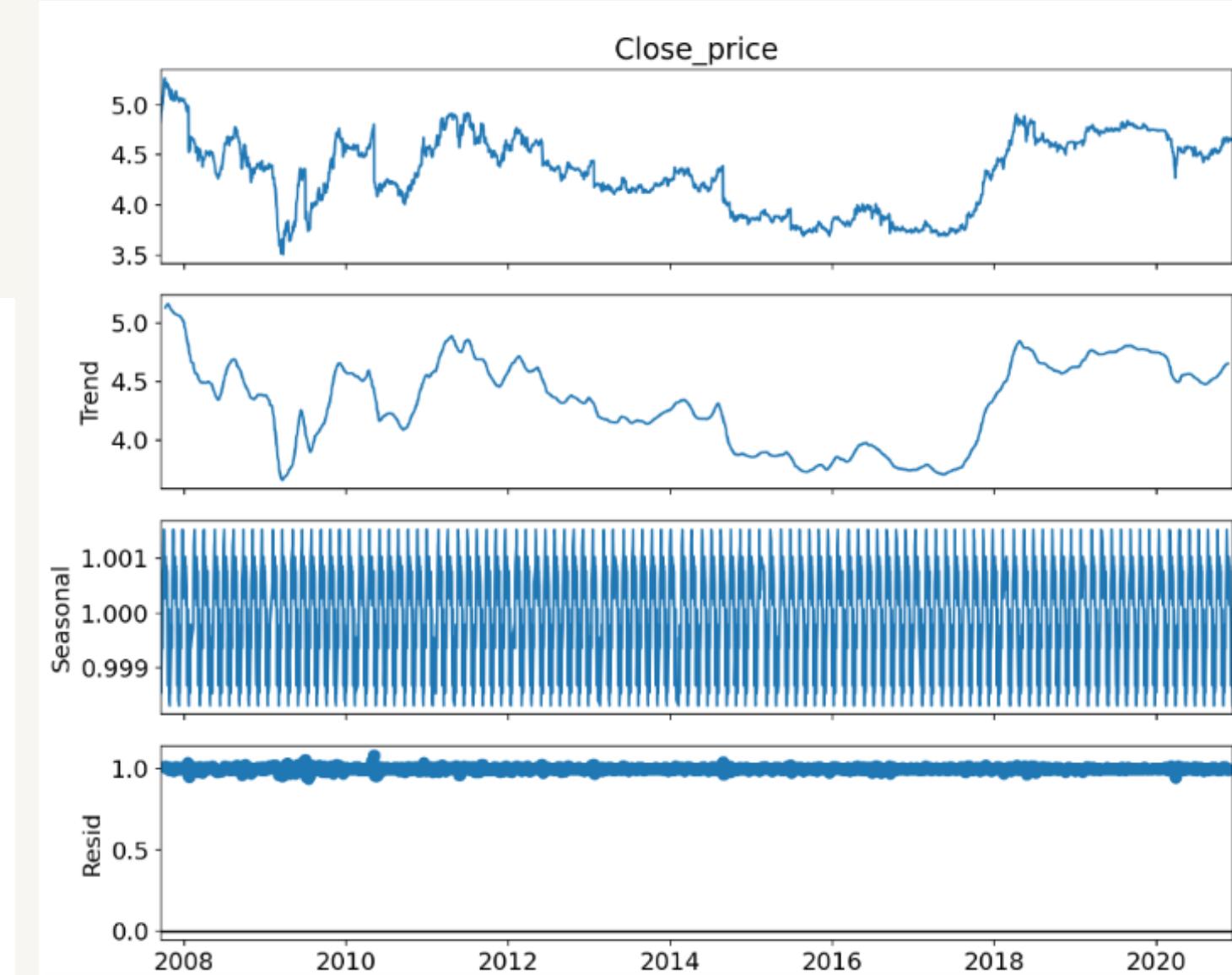
3 Chuyển đổi dữ liệu

4 Xây dựng mô hình

5 Mô hình Holt-Winters

1 Chia dữ liệu

2 Đánh giá



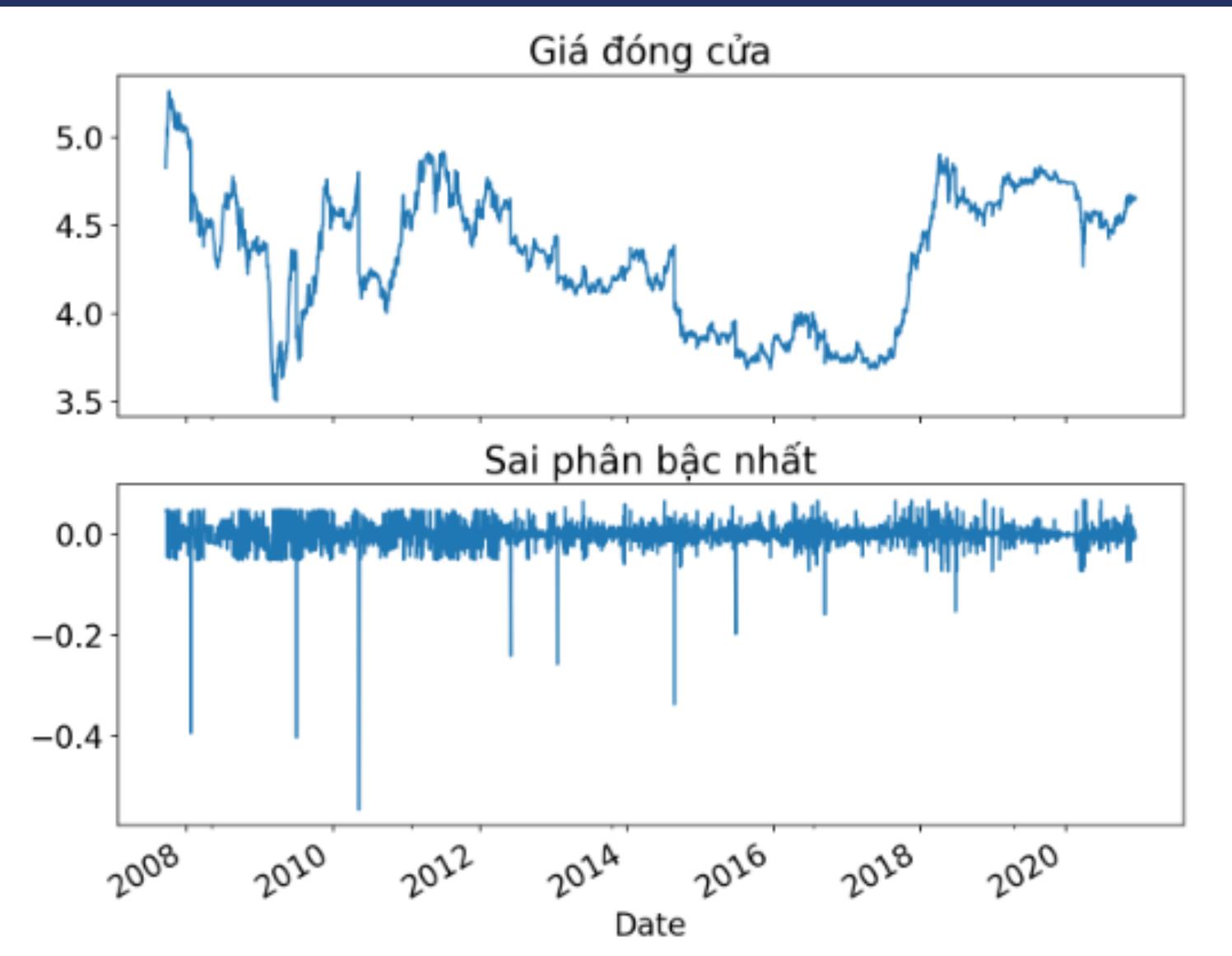
Nhận thấy rằng giá cổ phiếu đóng cửa không có hướng xu hướng cụ thể, không có xu hướng tăng hoặc giảm rõ rệt theo thời gian. Và có những năm biến động như trong giai đoạn năm 2008 đến năm 2010.

Xác định p,d,q

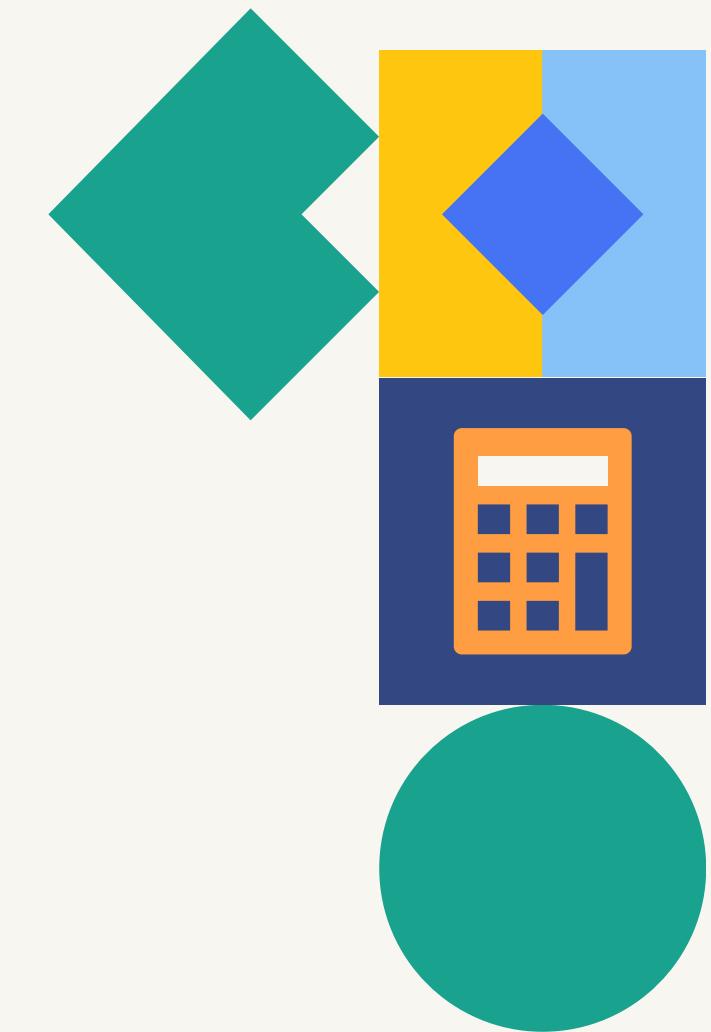
```
Fail to reject the null hypothesis (H0),  
the data is non-stationary  
ADF: Test statistic      -2.490281  
p value                  0.117857  
# of Lags                1.000000  
# of Observations        3297.000000  
Critical value (1%)      -3.432335  
Critical value (5%)      -2.862417  
Critical value (10%)     -2.567237  
dtype: float64  
-----  
Fail to reject the null hypothesis (H0),  
the data is non-stationary  
KPSS: Test statistic      1.045143  
p value                  0.010000  
# of Lags                37.000000  
Critical value (10%)     0.347000  
Critical value (5%)      0.463000  
Critical value (2.5%)    0.574000  
Critical value (1%)      0.739000
```

Dữ liệu chưa dừng

Chuyển dữ liệu về chuỗi dừng

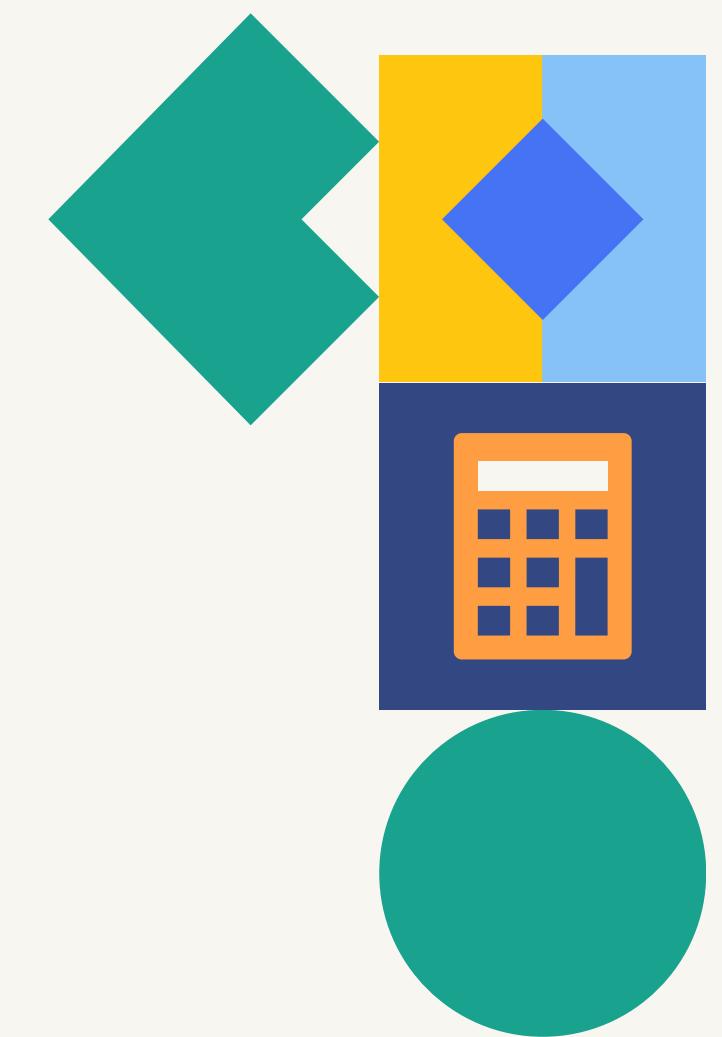
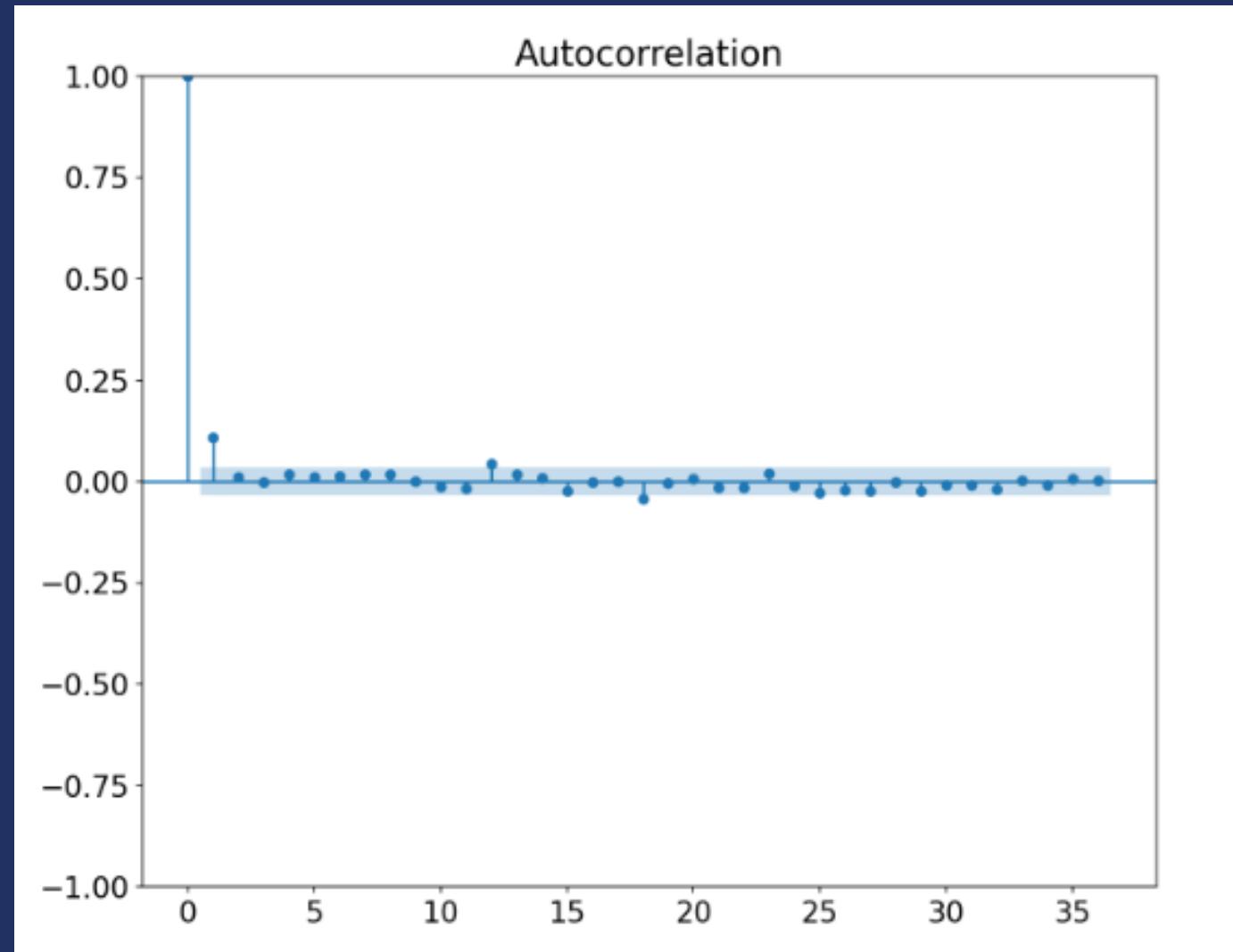
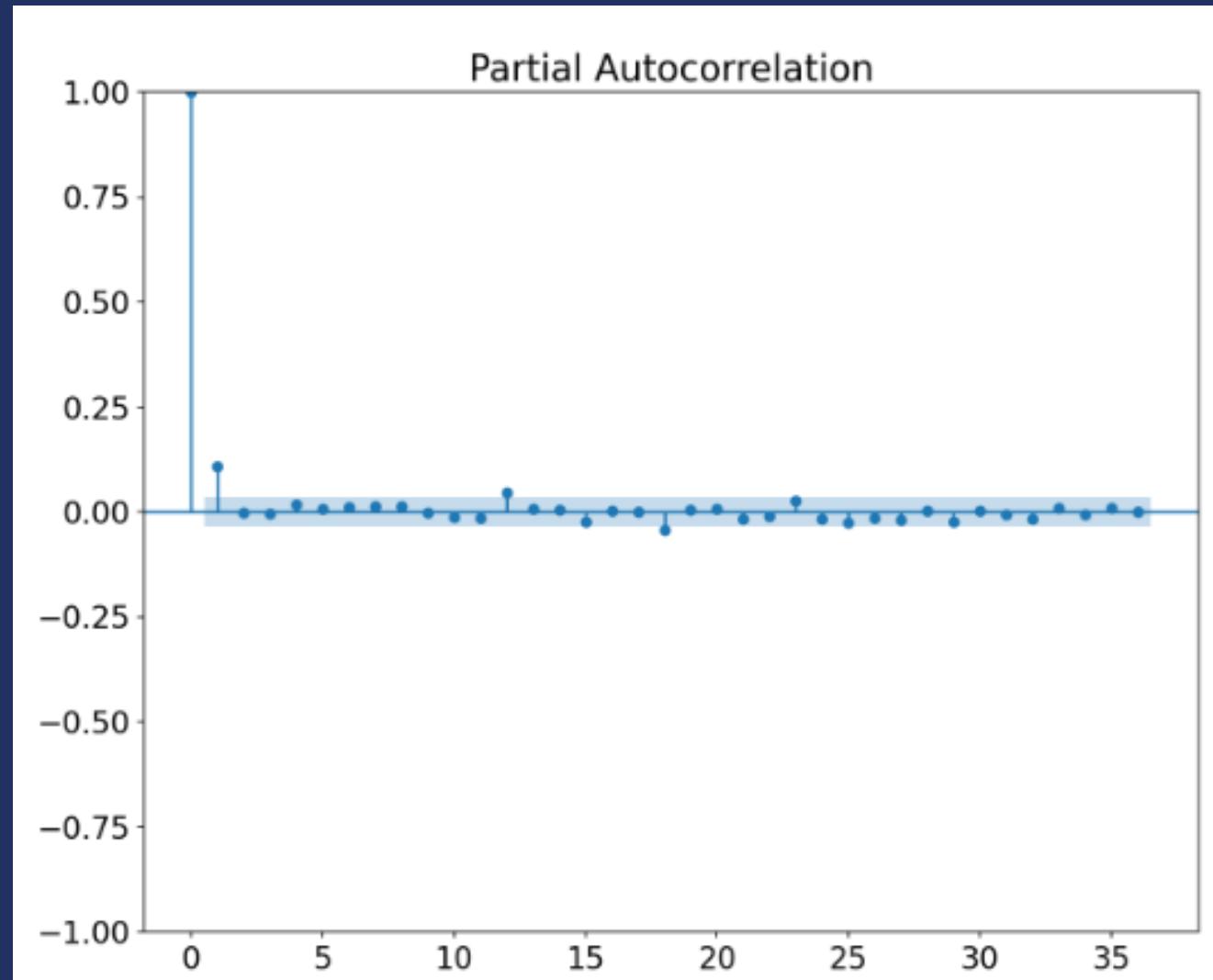


Dữ liệu đã dừng



d = 1

Xác định p,d,q



$p = 0,1$

$q = 0,1$

ARIMA

Kết hợp các kết quả từ các bậc của p, d và kết quả khi lấy sai phân $d = 1$, xác định các trường hợp để thực hiện kiểm định ARIMA:

	AIC
ARIMA(0,1,0)	-14487.383957742079
ARIMA(0,1,1)	-14523.496586429437
ARIMA(1,1,0)	-14523.864646668484
ARIMA(1,1,1)	-14521.861566201878



ARIMA tối ưu nhất là mô hình có chỉ số **AIC thấp nhất**. Do đó, nhóm nghiên cứu quyết định lựa chọn mô hình **ARIMA(1,1,0)**

Presenting Results

Performing stepwise search to minimize aic

```
ARIMA(2,1,2)(0,0,0)[0] intercept : AIC=-14517.826, Time=0.44 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=-14487.384, Time=0.19 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=-14523.865, Time=0.20 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=-14523.497, Time=0.32 sec
ARIMA(0,1,0)(0,0,0)[0]          : AIC=-14489.371, Time=0.11 sec
ARIMA(2,1,0)(0,0,0)[0] intercept : AIC=-14521.869, Time=0.47 sec
ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=-14521.869, Time=0.47 sec
ARIMA(2,1,1)(0,0,0)[0] intercept : AIC=-14519.869, Time=1.02 sec
ARIMA(1,1,0)(0,0,0)[0]          : AIC=-14525.855, Time=0.09 sec
ARIMA(2,1,0)(0,0,0)[0]          : AIC=-14523.860, Time=0.58 sec
ARIMA(1,1,1)(0,0,0)[0]          : AIC=-14523.852, Time=0.24 sec
ARIMA(0,1,1)(0,0,0)[0]          : AIC=-14525.487, Time=0.24 sec
ARIMA(2,1,1)(0,0,0)[0]          : AIC=-14521.860, Time=0.37 sec
```

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

Total fit time: 4.770 seconds

SARIMAX Results

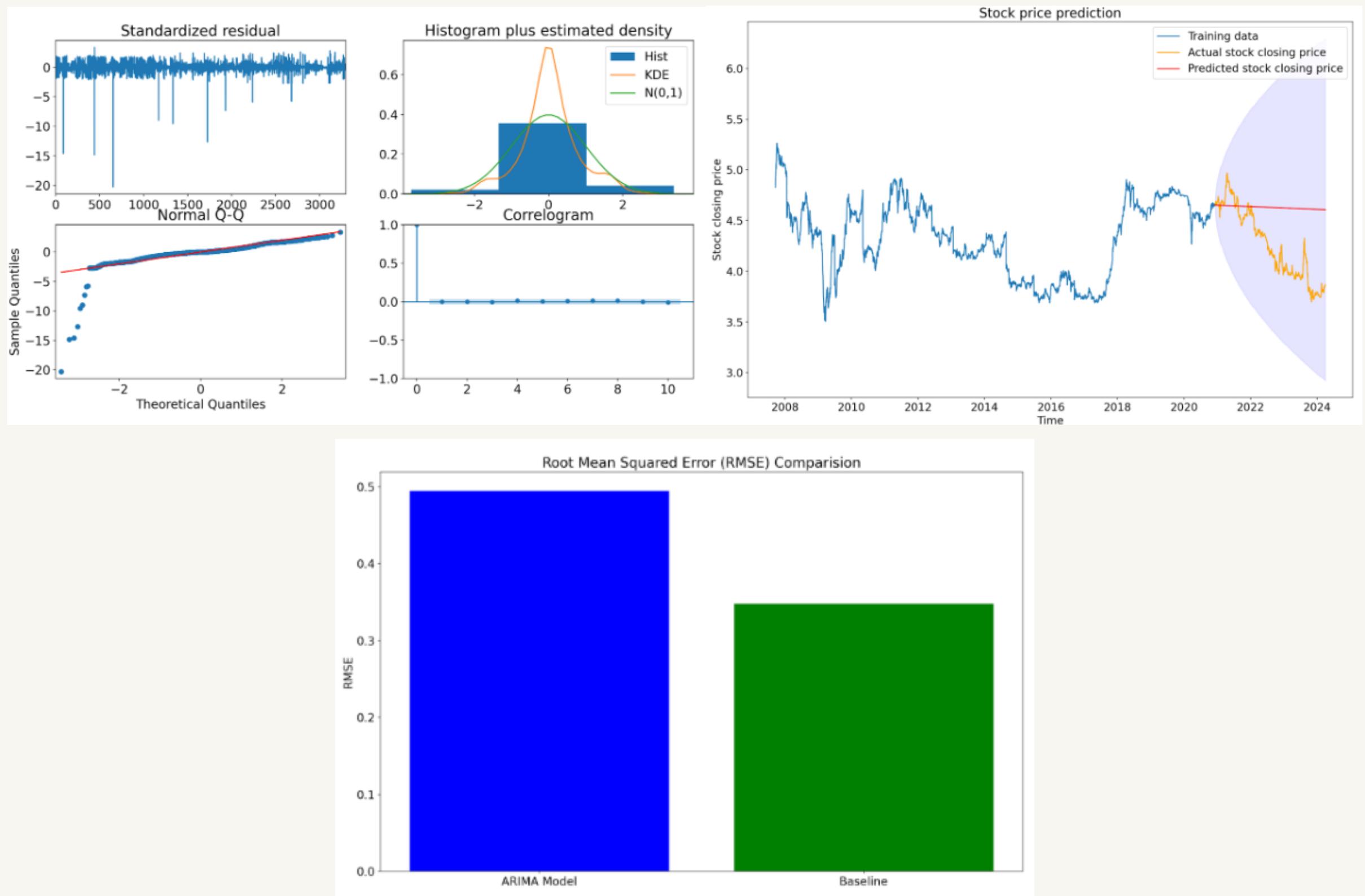
```
=====
Dep. Variable:                      y      No. Observations:             3299
Model:                 SARIMAX(1, 1, 0)   Log Likelihood:        7264.927
Date:                Sun, 14 Apr 2024   AIC:                  -14525.855
Time:                       23:22:14     BIC:                  -14513.653
Sample:                           0      HQIC:                  -14521.487
                                  - 3299
Covariance Type:                  opg
=====
```

	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.1077	0.011	9.441	0.000	0.085	0.130
sigma2	0.0007	2.69e-06	265.812	0.000	0.001	0.001

```
=====
Ljung-Box (L1) (Q):                   0.00    Jarque-Bera (JB):       1185193.90
Prob(Q):                               0.99    Prob(JB):            0.00
Heteroskedasticity (H):               0.28    Skew:                  -5.74
Prob(H) (two-sided):                  0.00    Kurtosis:             95.16
=====
```



Mô hình đã dự đoán đúng xu hướng của giá đóng cửa từ giai đoạn sau 2021 đến 2024 là giảm xuống. Đồng thời, chỉ số MSE và RMSE lần lượt là 0.245 và 0.495 là khá nhỏ chứng tỏ mô hình có độ chính xác cao.



Dữ liệu thời gian

Nội dung

- 1 Thu thập dữ liệu
Mô tả bài toán
- 2 Phân tích khám phá
dữ liệu
- 3 Chuyển đổi dữ liệu
- 4 Xây dựng mô hình
- 5 Mô hình Holt-Winters

Chuyển đổi dữ liệu

- Sử dụng phương thức resample để chuyển đổi dữ liệu trong DataFrame theo chu kỳ tháng (MS).
- Sử dụng phương thức mean() để tính trung các giá trị trong mỗi chu kỳ

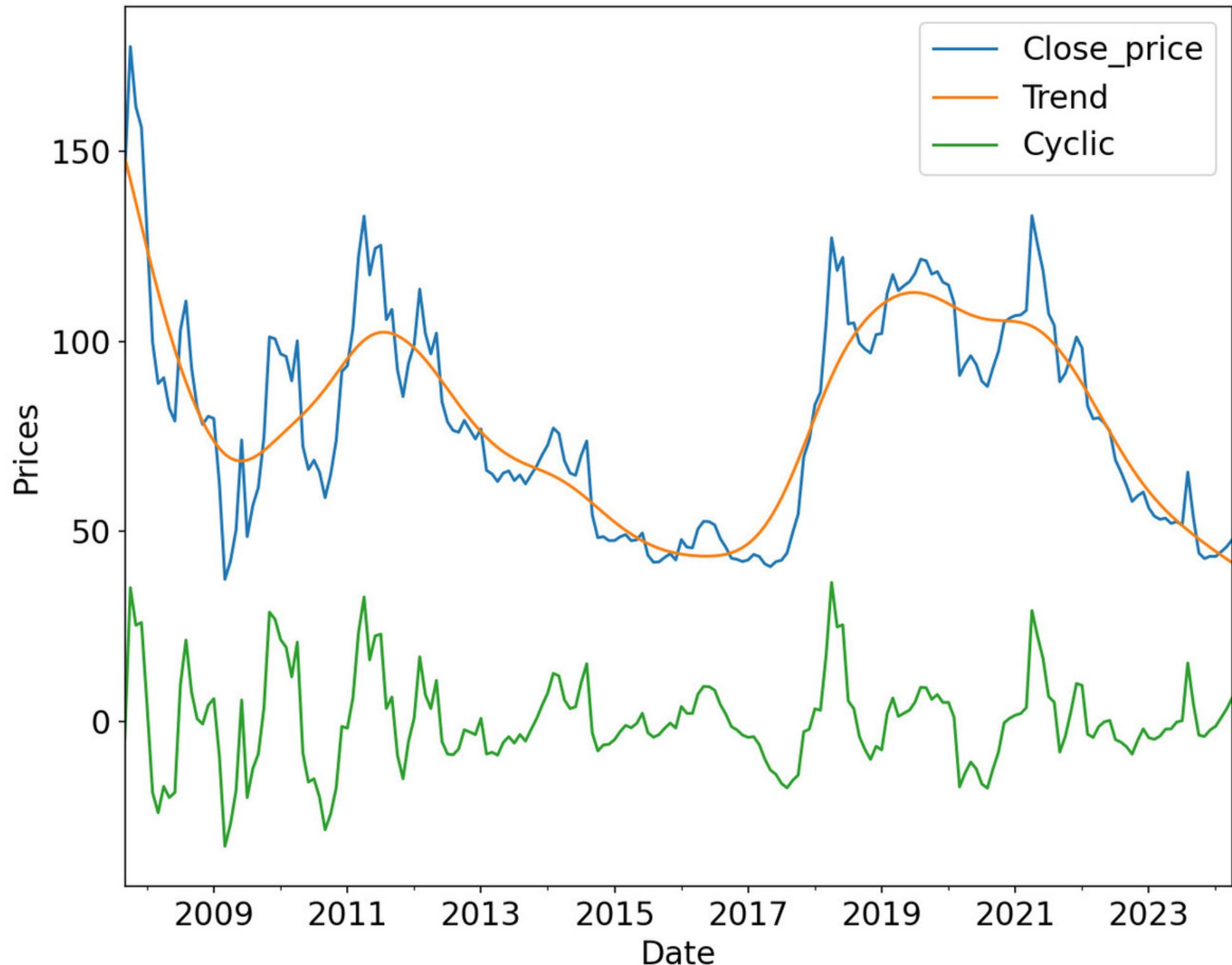
Kiểm tra tính xu hướng, chu kỳ

Kiểm tra mô hình phân rã chuỗi thời gian

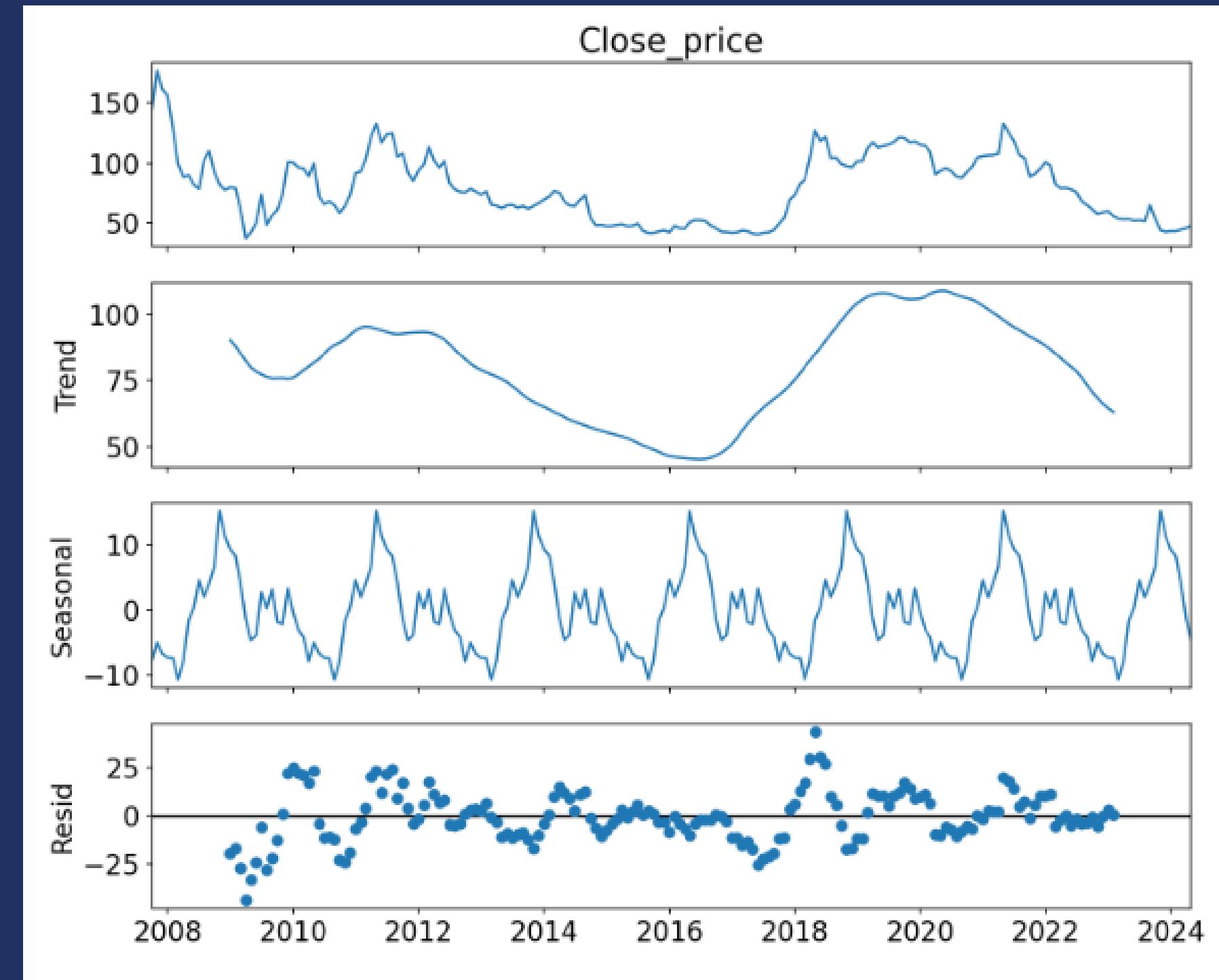
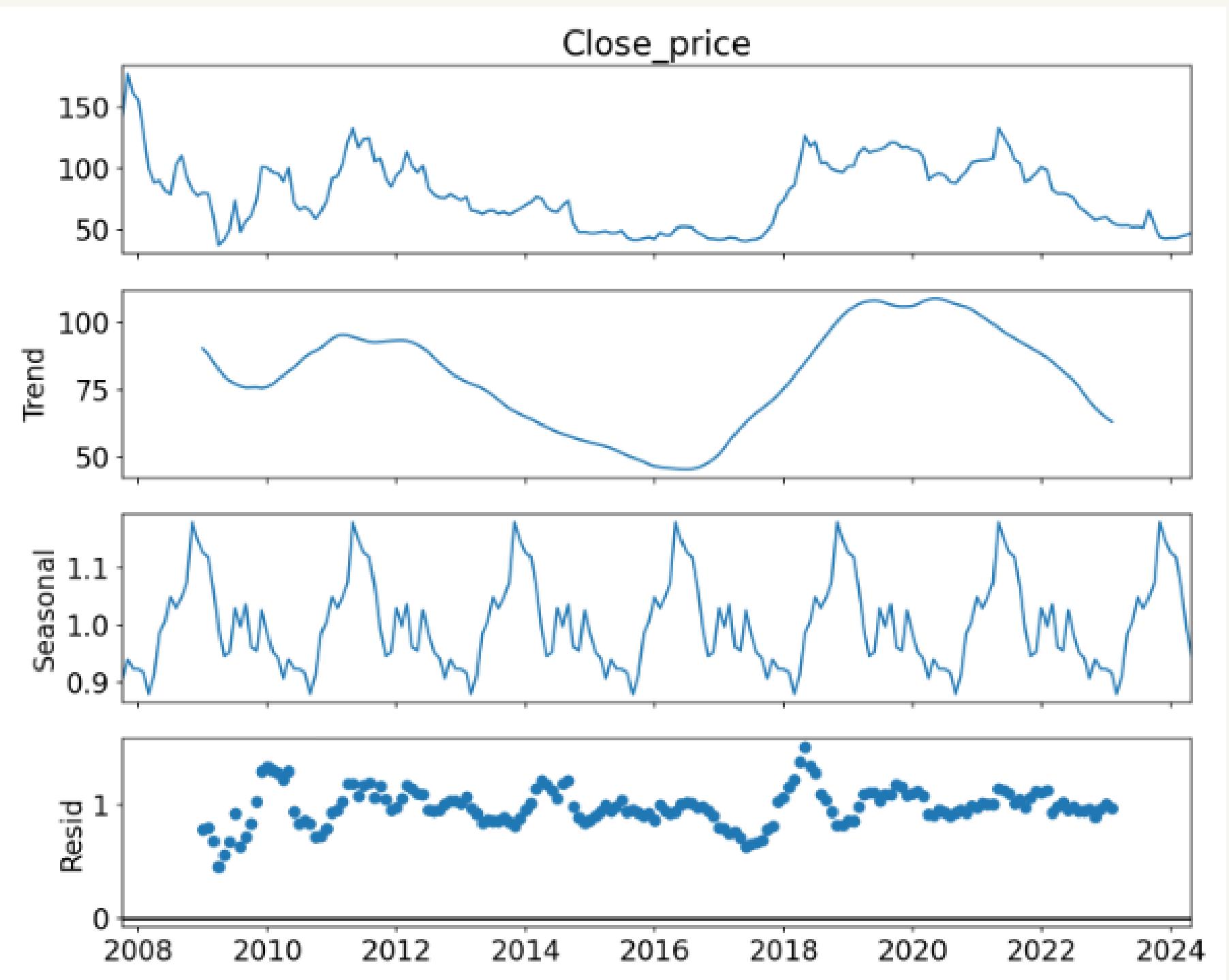
Tối ưu hóa siêu tham số (Hyperparameter Optimization)

Ước lượng chu kỳ và xu hướng (Cycle & Trend)

So sánh giá trị thật, chu
kỳ và kỳ vọng nhận
thấy giá trị thực tế là
tổng của 2 thành phần
xu hướng và chu kỳ.

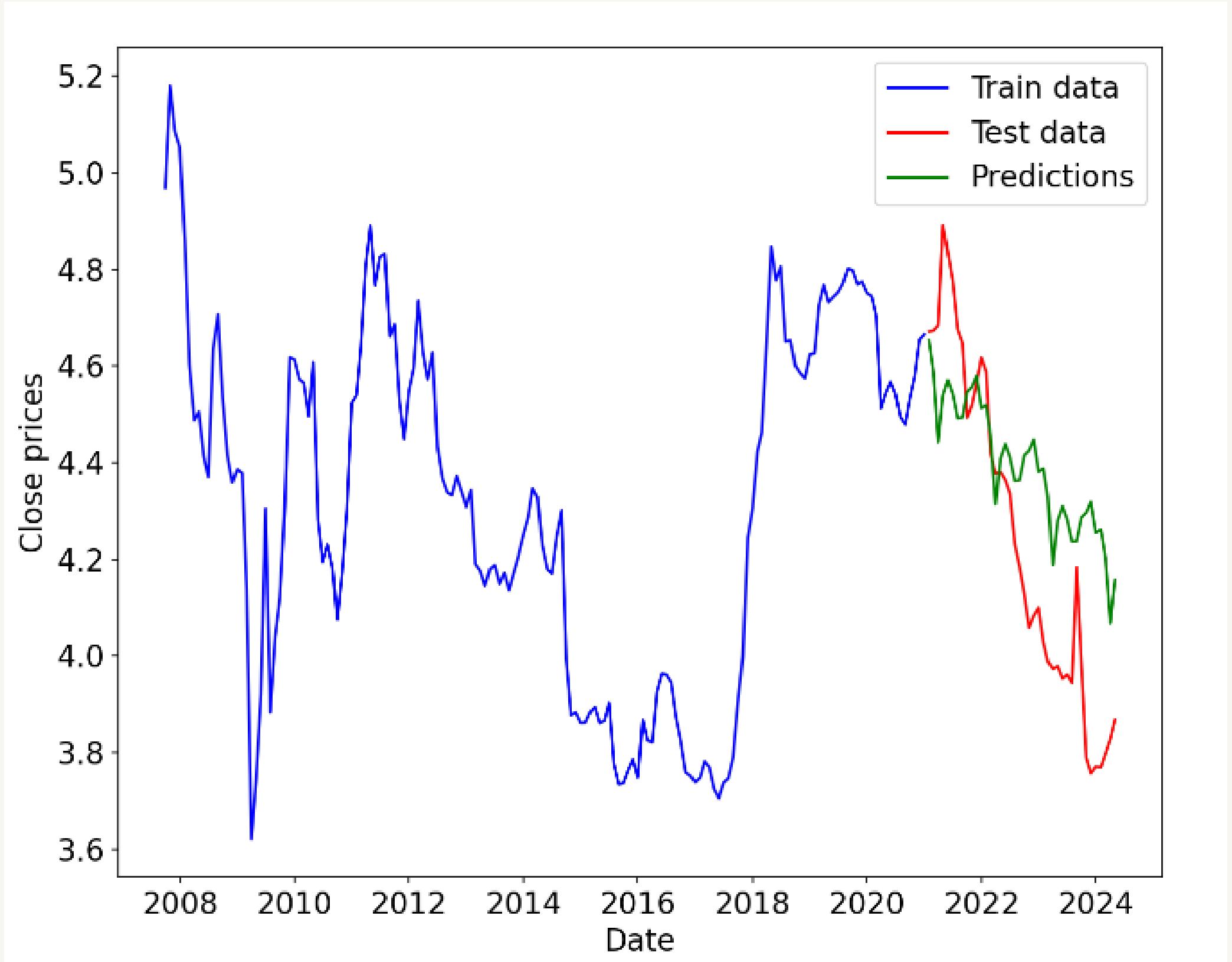


Mô hình nhân tính



Mô hình cộng tính

Holt-Winters



Xác định bộ siêu tham số, tối ưu nhất bằng cách thực hiện vòng lặp với các tổ hợp của 3 siêu tham số trên và tính chỉ số Mean Absolute Error (MAE) của từng tổ hợp. Tổ hợp tối ưu nhất sẽ là tổ hợp có chỉ số MAE thấp nhất.

$$\alpha = 0.6, \beta = 0, \gamma = 0.8.$$



Kết luận

Mô hình Holt-Winters mang lại kết quả với độ chính xác cao

- Cả 2 mô hình ARIMA và Holt-Winter đã dự đoán đúng xu thế chuyển động của test_data có xu hướng giảm.
- Đồng thời, kết quả của 3 chỉ số MSE, RMSE và MAPE chứng tỏ rằng mô hình đã hoạt động tốt trên bộ dữ liệu và đưa ra kết quả có sai số thấp.
- Bên cạnh đó, với 3 chỉ số của mô hình Holt-Winters tính được lần lượt là 0.076, 0.275, 0.057 nhỏ hơn so với 3 chỉ số của mô hình ARIMA là 0.245, 0.495, 0.100

Mô hình máy học

- 1 Decision Tree
- 2 Random Forest
- 3 Neural Network
- 4 Support Vector Machine

Mô hình máy học

Nội dung

1 Tổng quan

2 Decision Tree

3 Random Forest

4 Neural Network

5 Support Vector
Machine

Dữ liệu từ một ngân hàng về thông tin chi tiết về khách hàng liên quan đến việc bán tiền gửi có kỳ hạn. Mục tiêu của dự án là giúp nhóm tiếp thị xác định những khách hàng tiềm năng có nhiều khả năng đăng ký tiền gửi có kỳ hạn hơn và điều này làm tăng tỷ lệ trúng

Thuộc tính	Mô tả
Age	Tuổi của khách hàng
Job	Công việc của khách hàng
Martial	Tình trạng hôn nhân
Education	Trình độ học vấn
Default	Có thẻ tín dụng chưa
Balance	Số dư cá nhân
Housing	Có vay mua nhà không

Loan	Khoản vay cá nhân
Contact	Cách liên hệ
Month	Tháng liên lạc cuối cùng trong năm
Day	Ngày liên lạc cuối cùng trong tuần
Duration	Thời lượng liên lạc lần cuối (s)
Campaign	Số lượng liên hệ được thực hiện trong chiến dịch này và cho khách hàng này
Pdays	Số ngày đã trôi qua sau khi khách hàng được liên hệ lần cuối từ chiến dịch trước đó
Previous	Số lượng liên hệ được thực hiện trước chiến dịch này và cho khách hàng này
Poutcome	Kết quả của chiến dịch tiếp thị trước đó
Y	Khách hàng đăng ký tiền gửi có kỳ hạn không?

**Liệu khách hàng có quyết định
gửi tiền có kỳ hạn tại ngân
hàng hay không?**

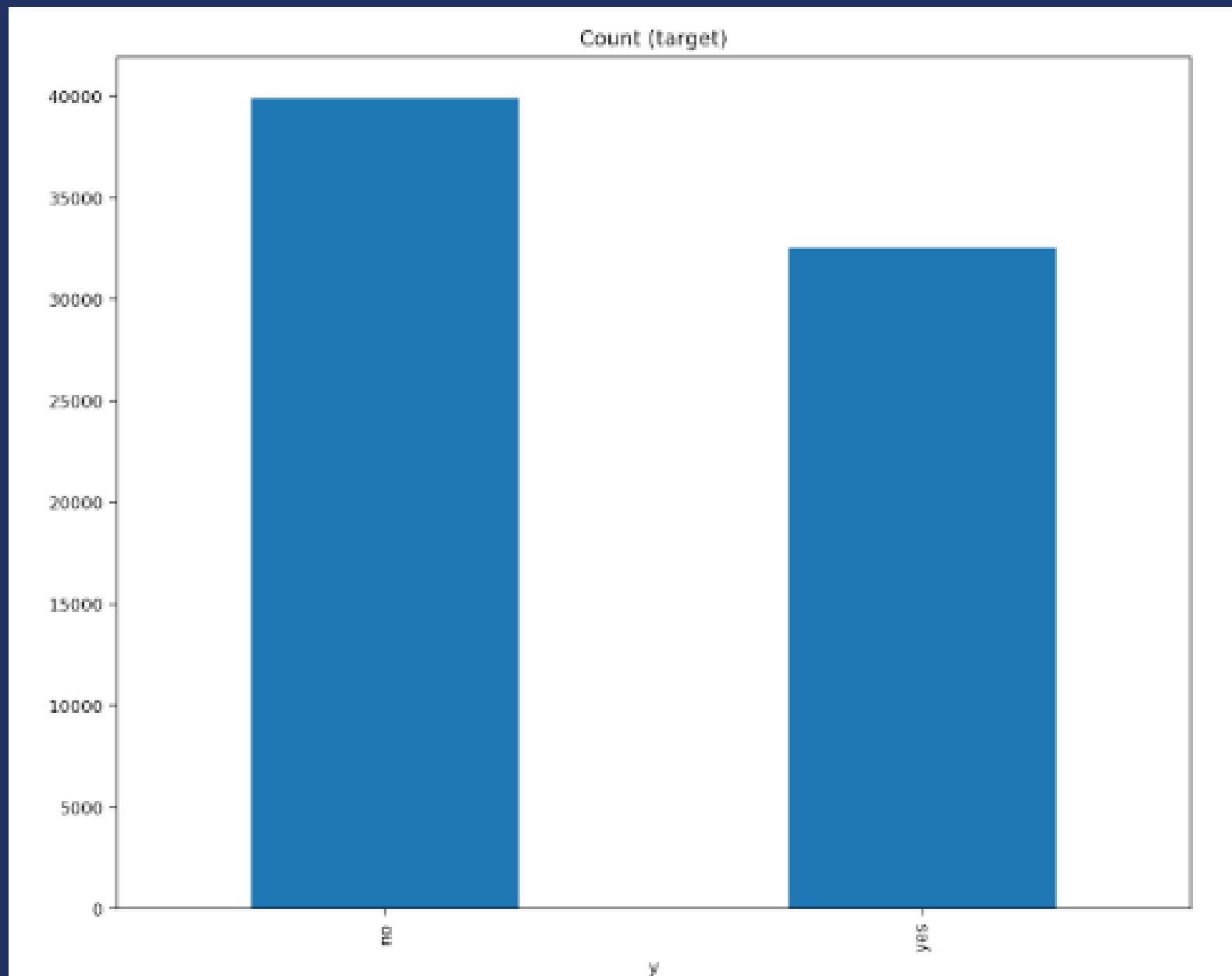


Xử lý mất cân bằng dữ liệu

Tỉ lệ mẫu trong lớp 0 chiếm khoảng 88.3%,
trong khi lớp 1 chỉ chiếm 11.7% ⇒ Điều này
cho thấy mất cân bằng lớn giữa hai lớp.

- Tiến hành Scale các chỉ số của các Feature và đưa về cùng 1 thang đo.
- Sau đó tiến hành SMOTEENN và ENN loại bỏ các mẫu nhiễu từ các lớp đa số dựa trên K-nearest neighbors.

Tỉ lệ mẫu trong lớp 0 giảm xuống khoảng 45.51 % trong khi lớp 1 tăng lên 54.49 %.
⇒ Dữ liệu đã được cân bằng



Mô hình máy học

Nội dung

1 Tiền xử lý dữ liệu

2 Decision Tree

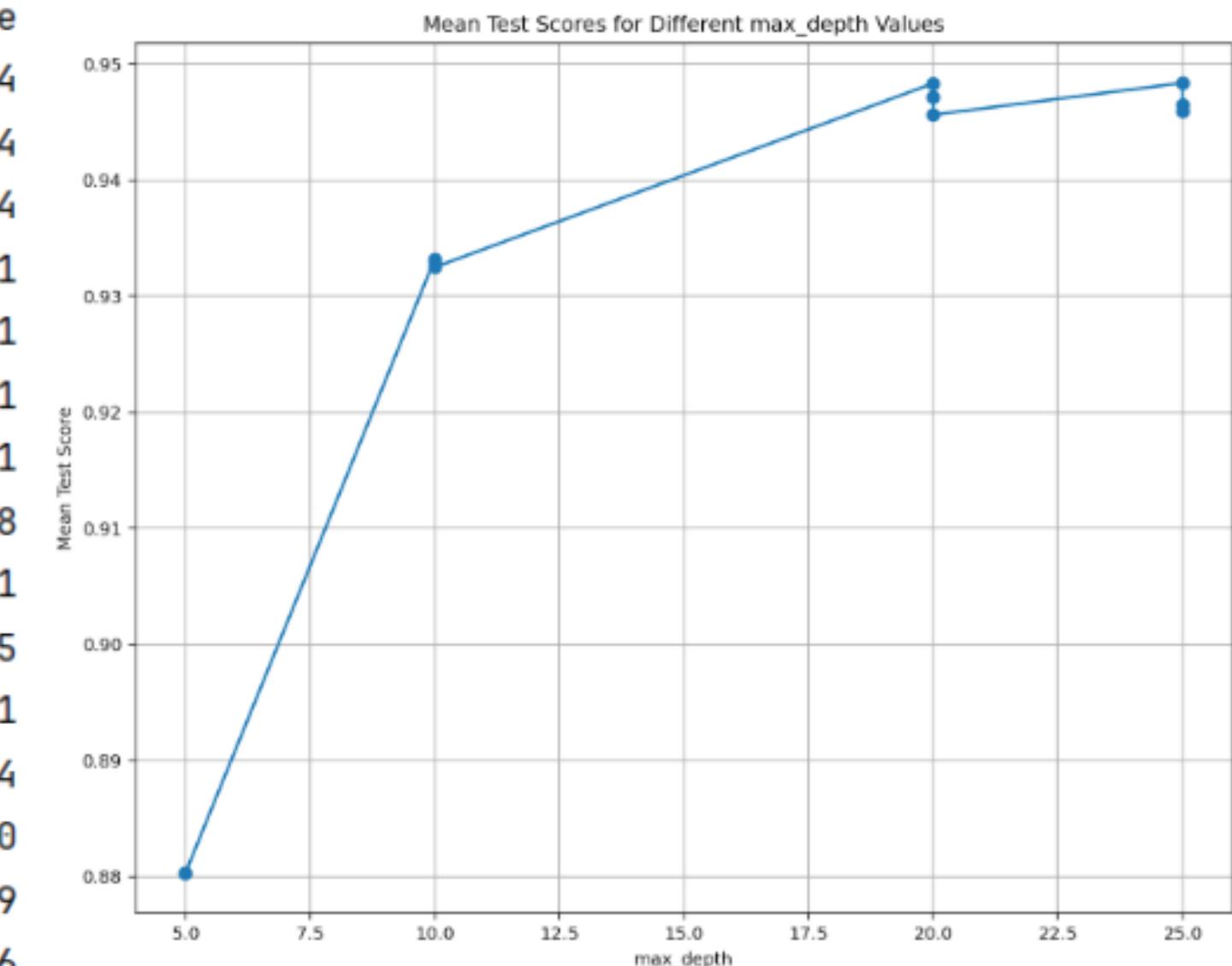
3 Random Forest

4 Neural Network

5 Support Vector
Machine

Tuning mô hình huấn luyện với các tham số ban đầu

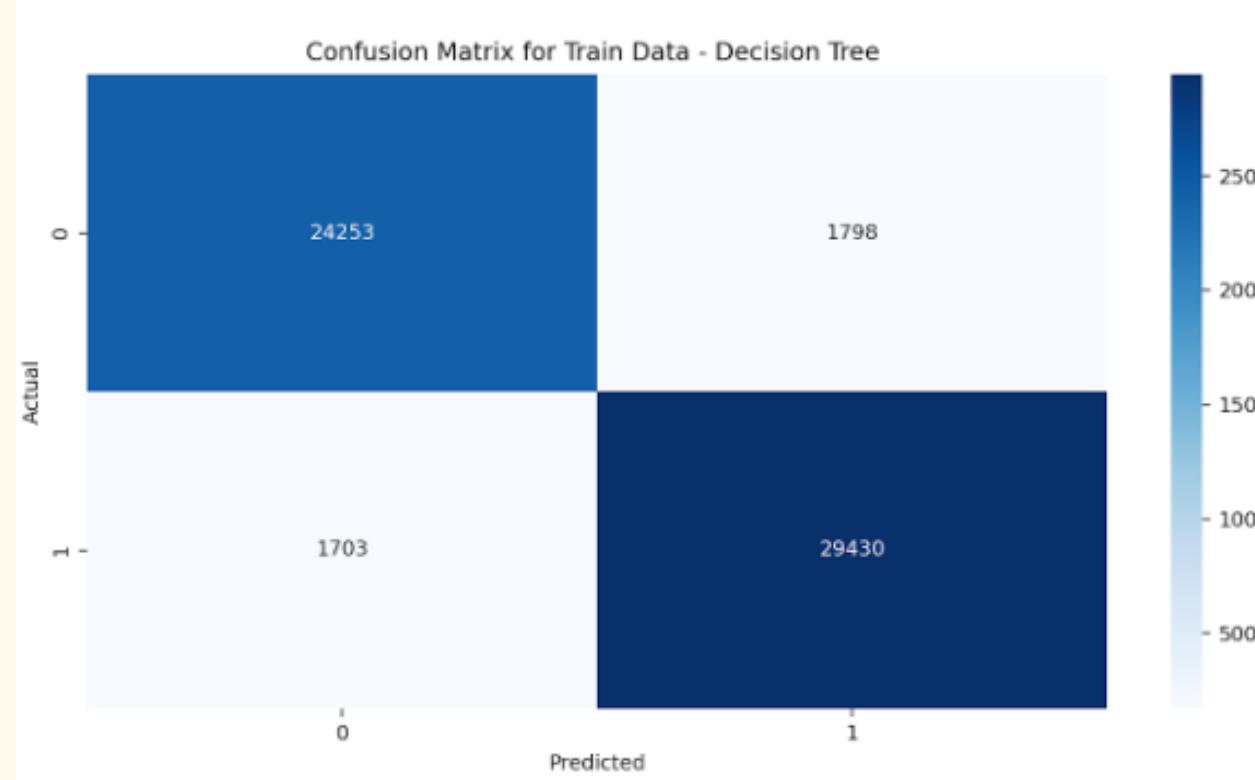
	max_depth	min_samples_split	mean_test_score
0	None	2	0.948744
1	None	5	0.946454
2	None	10	0.945964
3	5	2	0.880281
4	5	5	0.880281
5	5	10	0.880281
6	10	2	0.933181
7	10	5	0.933058
8	10	10	0.932481
9	20	2	0.948325
10	20	5	0.947171
11	20	10	0.945614
12	25	2	0.948360
13	25	5	0.946489
14	25	10	0.945946



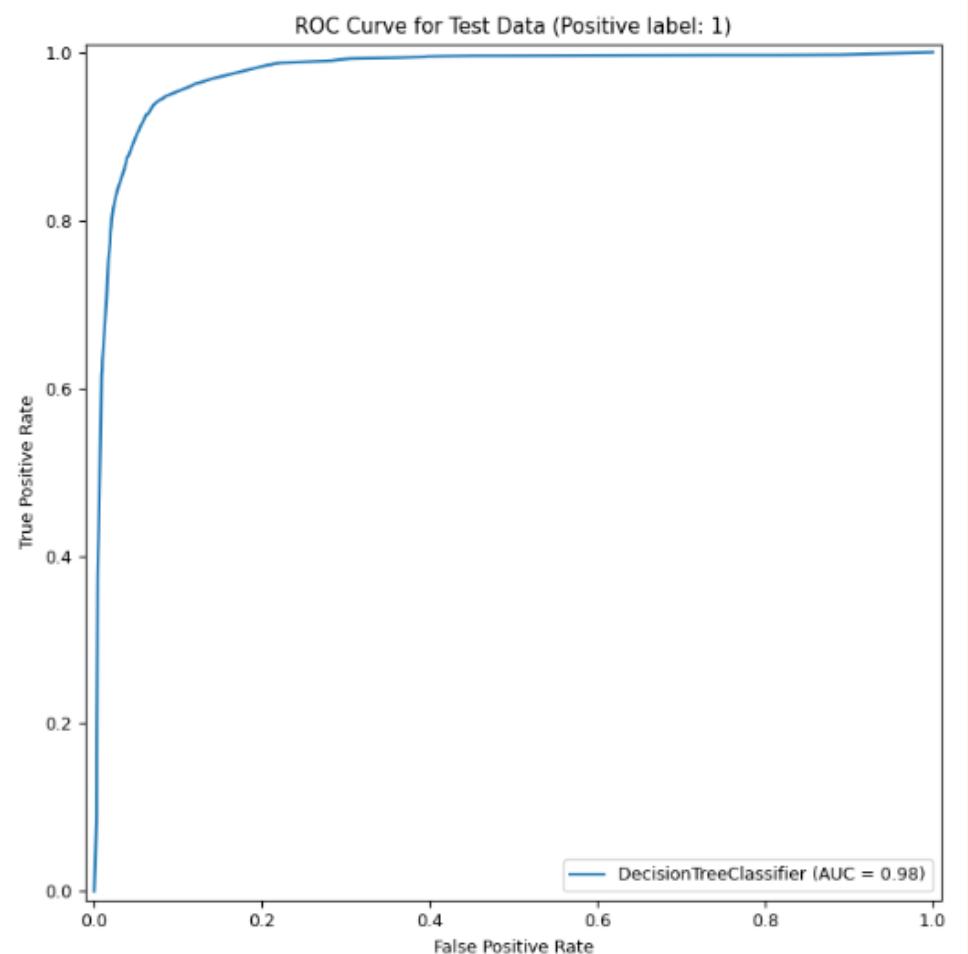
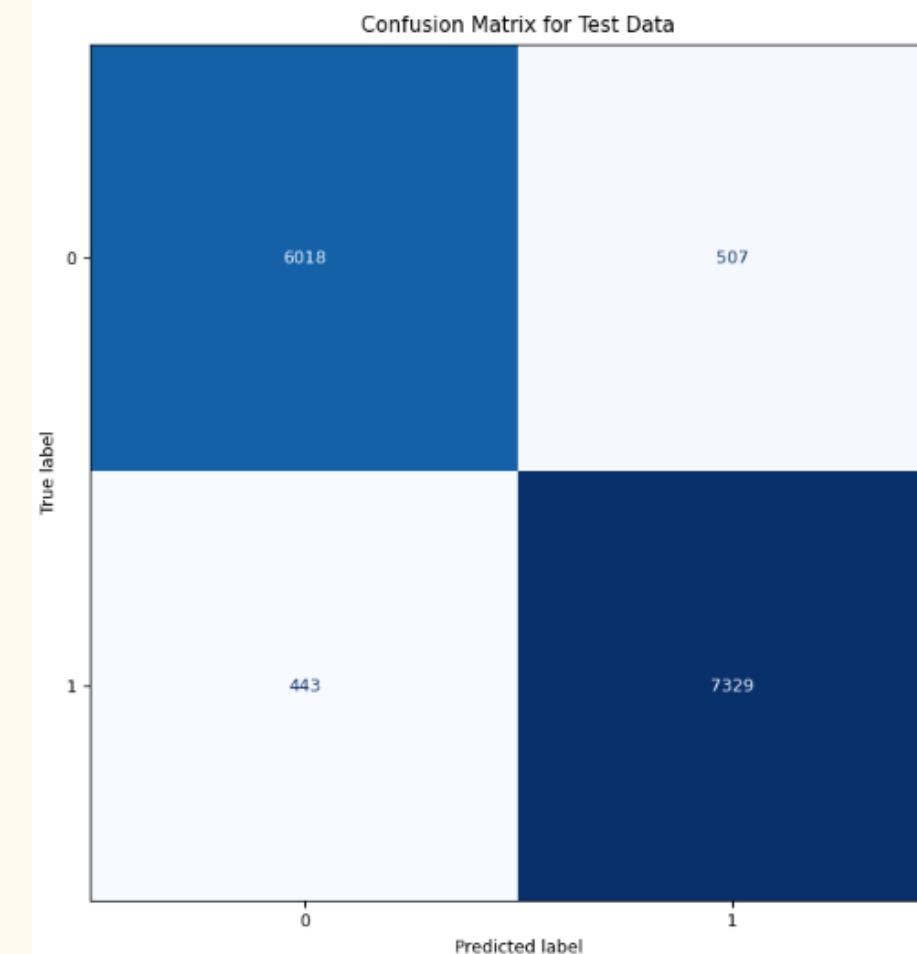
max_depth trong khoảng từ [5:10] và mean_samples_split là 2

Decision Tree Training Accuracy: 0.939
Decision Tree Test Accuracy: 0.934

Dữ liệu huấn luyện



Dữ liệu kiểm thử



Mô hình máy học

Nội dung

1 Tiền xử lý dữ liệu

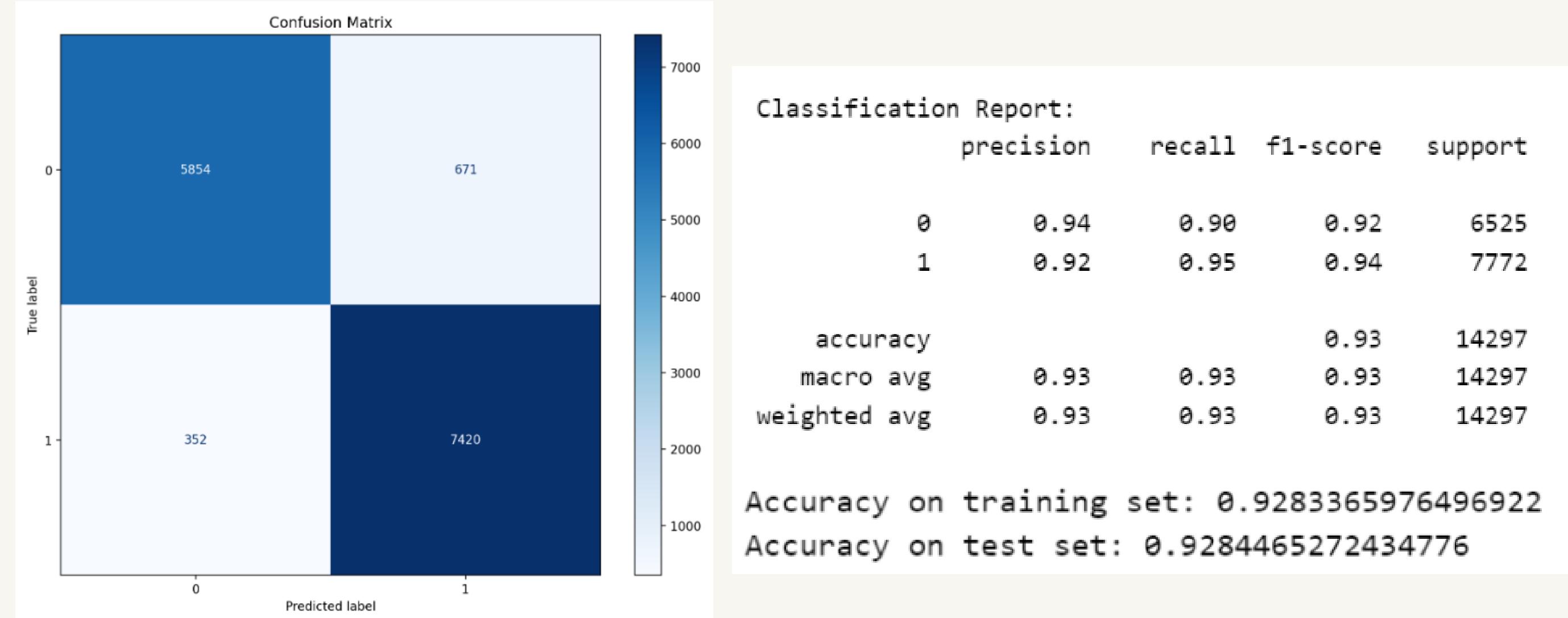
2 Decision Tree

3 Random Forest

4 Neural Network

5 Support Vector
Machine

max_depth = 9, min_samples_leaf=4,
min_samples_split=10, n_estimators=50



Mô hình thể hiện một hiệu suất tương đối cân bằng, với cả hai lớp đều đạt được F1-score khoảng 0.92 đến 0.94. Có khả năng phân loại tốt trên cả hai lớp, mặc dù vẫn còn một số cải thiện có thể được thực hiện để giảm thiểu các dự đoán sai.

Mô hình máy học

Nội dung

1 Tiềm xử lý dữ liệu

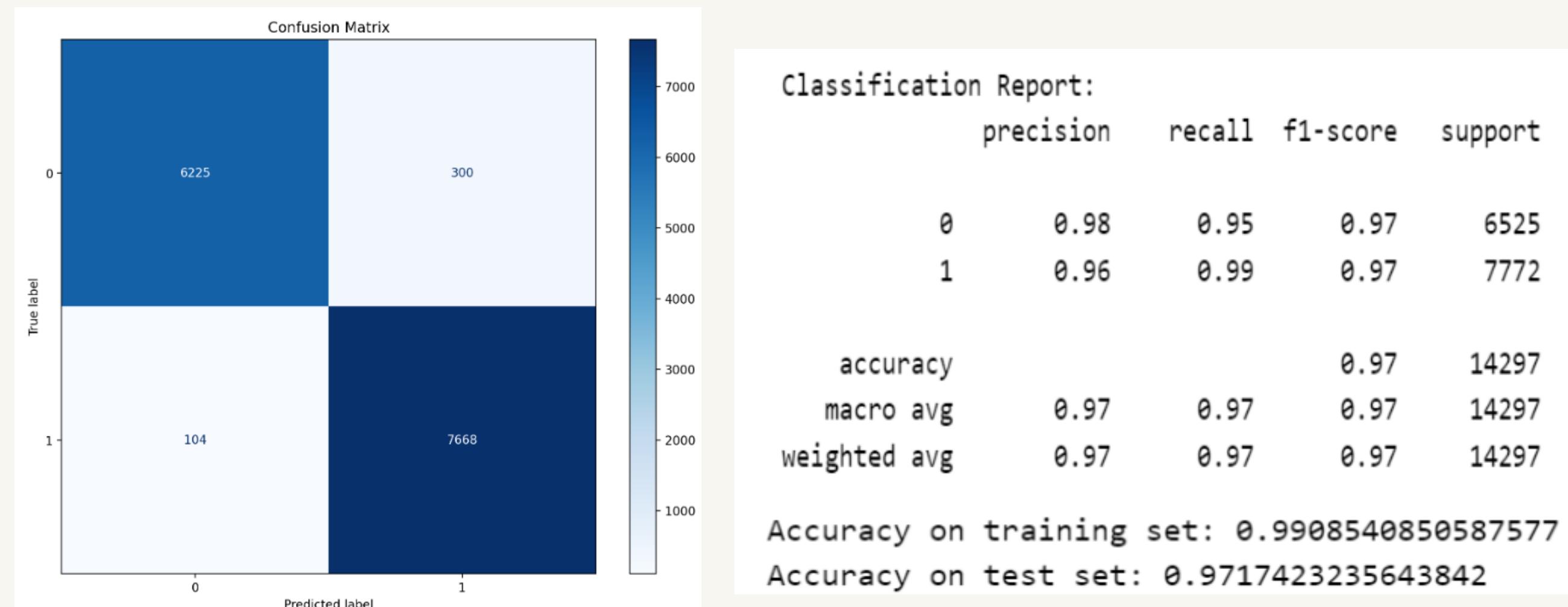
2 Decision Tree

3 Random Forest

4 Neural Network

5 Support Vector Machine

activation='tanh', solver='adam',
learning_rate='constant', hidden_layer_sizes=(50,50),
alpha=0.1



Mô hình đạt được một độ chính xác (accuracy) là 97%, cho thấy khả năng dự đoán đúng đối với hầu hết các trường hợp trong tập kiểm tra.

Mô hình máy học

Nội dung

1 Tiền xử lý dữ liệu

2 Decision Tree

3 Random Forest

4 Neural Network

5 Support Vector Machine

Tham số tối ưu

Quá trình tối ưu hóa tham số được thực hiện thông qua phương pháp tìm kiếm ngẫu nhiên (Randomized Search)

- 3 loại kernel là: 'rbf' , 'sigmoid' và 'polynomial' cho mô hình dự đoán
- 'gamma': [1, 0.1, 0.01, 0.001],
- 'C': [0.01, 0.1, 1, 10, 100],
- 'class_weight': class_weights_svm}

Kernel là 'rbf',

Gamma là 0.01,

Trọng số của các lớp là {0: 0.3639090909090909,

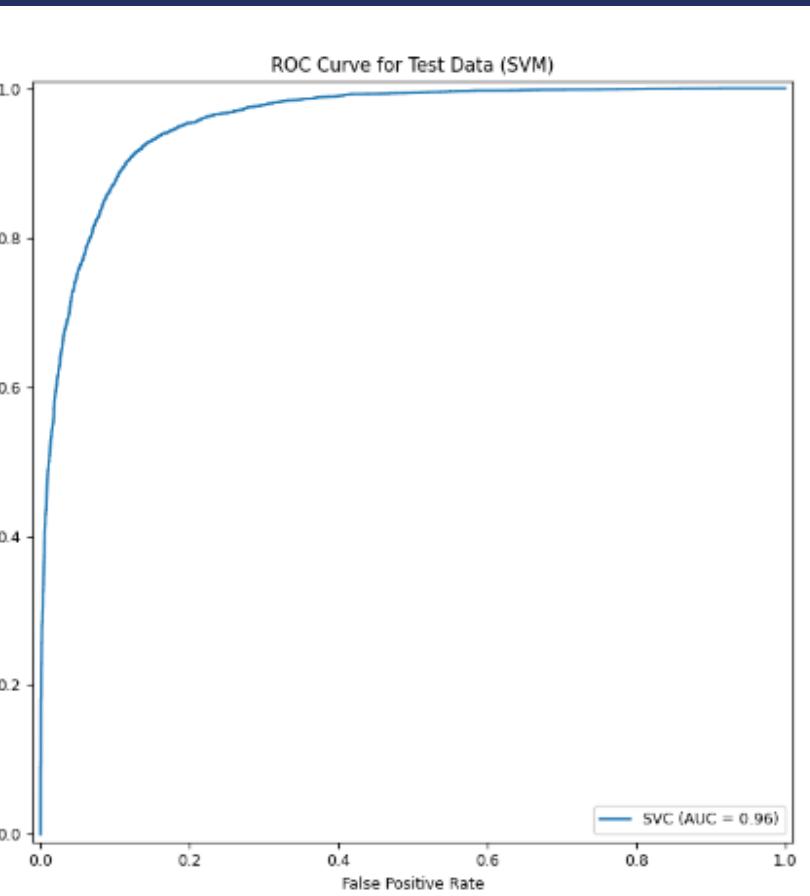
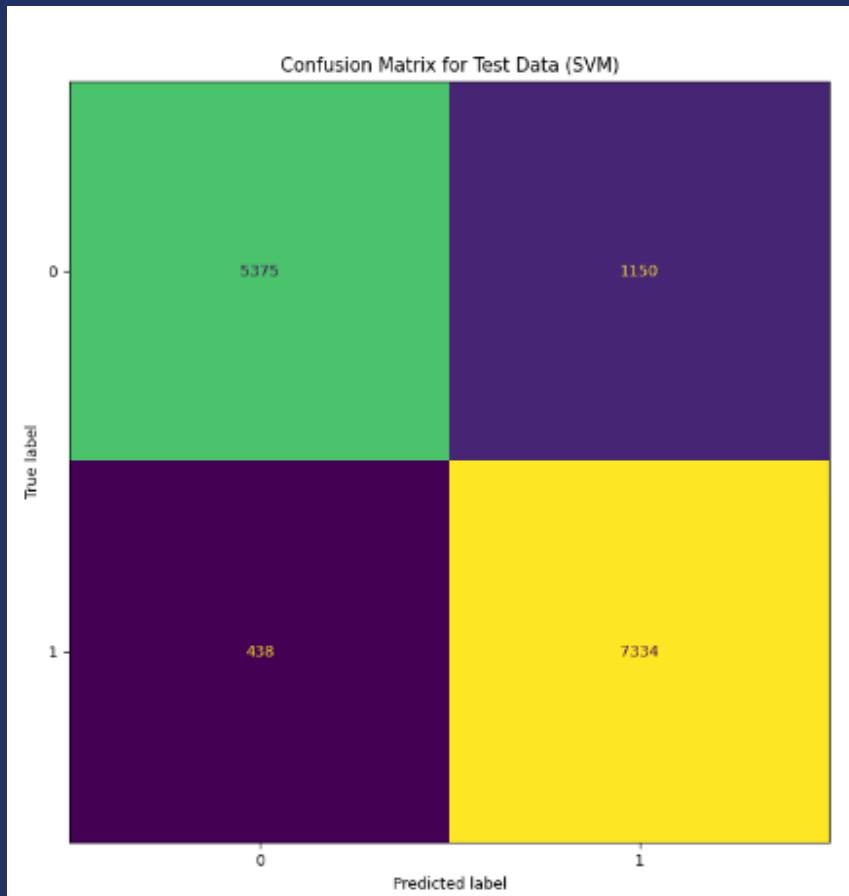
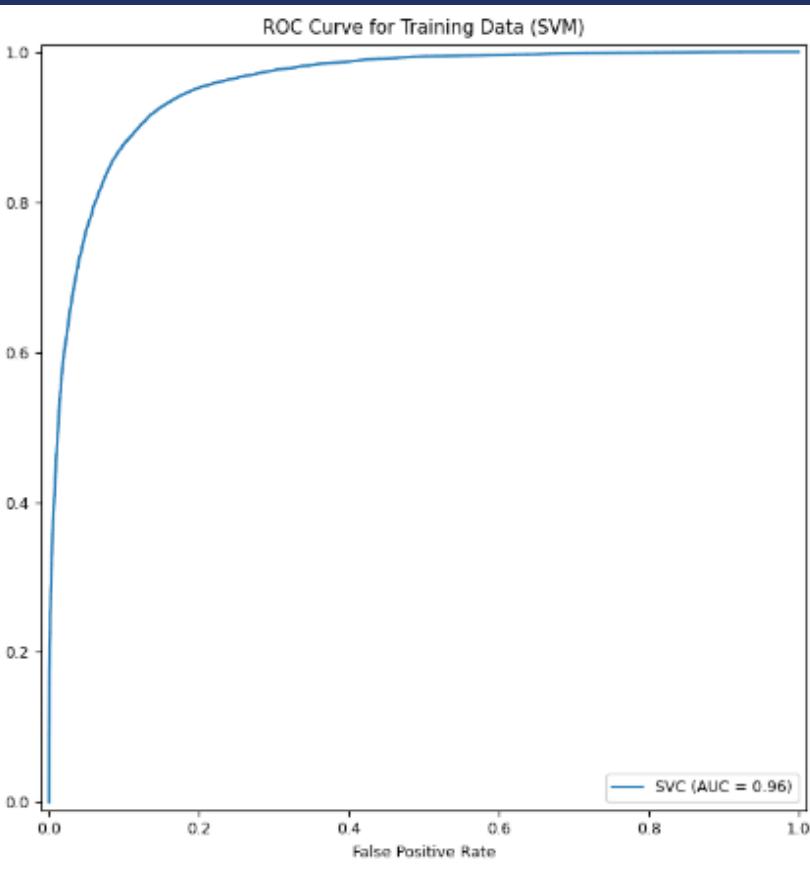
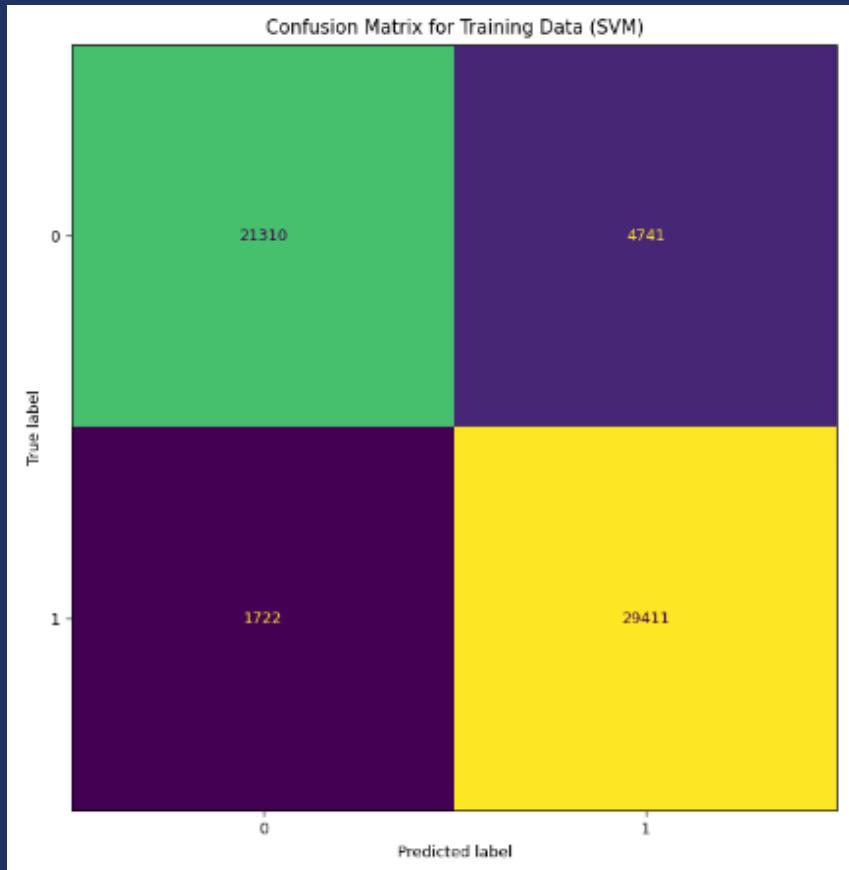
1: 0.636090909090909}

Tham số C là 1.0.

Độ chính xác tốt nhất trên tập huấn luyện là khoảng 0.8852.

Đánh giá

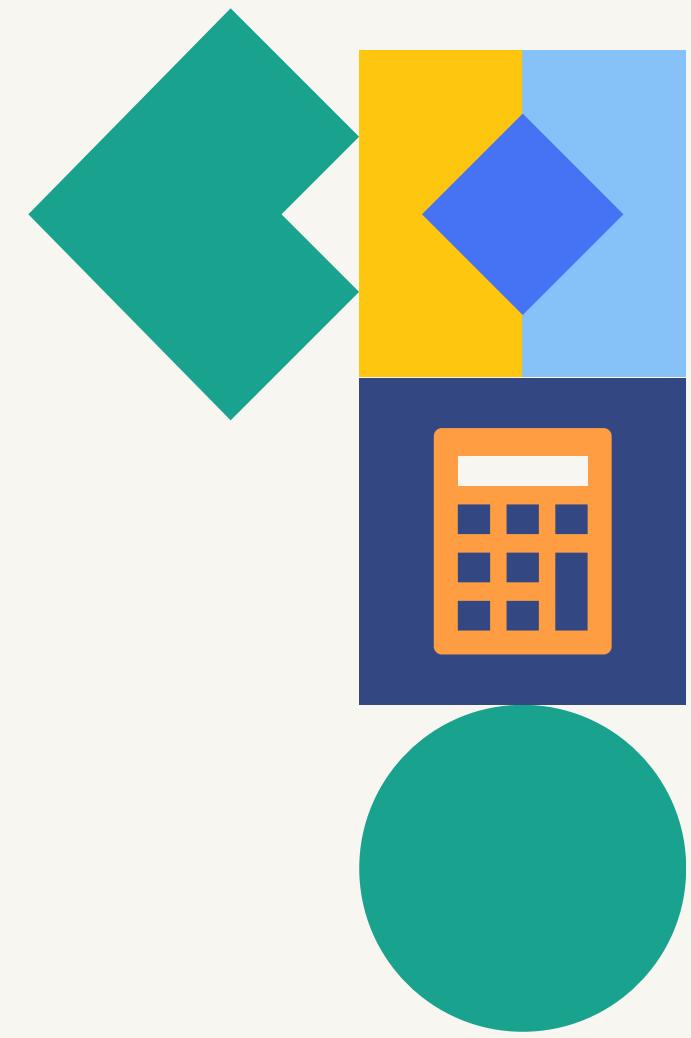
Dữ liệu huấn luyện



Dữ liệu kiểm thử

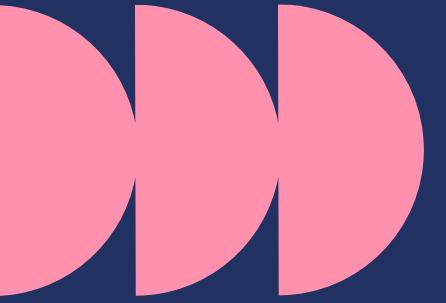
Accuracy on Training Data (SVM):
0.8869788752098489

Accuracy on Test Data (SVM):
0.8889277470798069



Đánh giá

Model	Accuracy	Precision	Recall	F1 - Score
Neural Network	95.24 %	94.98 %	96.35 %	95.66 %
Random Forest	94.27 %	93.53 %	96.11 %	94.8 %
Decision Tree	93.36 %	93.51%	94.34 %	93.92 %
Supporter Vetur Machine	88.89 %	86.45 %	94.36 %	90.23 %

Tổng quan, Neural Network và Random Forest có vẻ hiệu quả hơn so với Decision Tree và Support Vector Machine trong bài toán này, với Neural Network có thể là lựa chọn tốt nhất với độ chính xác cao nhất và F1-score tốt nhất.

**Cảm ơn thầy
đã lắng nghe
phần trình
bày**

G_232IS2902_08

