

ĐẠI HỌC QUỐC GIA TP.HCM  
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT



BÁO CÁO ĐỒ ÁN MÔN HỌC

PHÂN TÍCH DỮ LIỆU VỚI R/PYTHON

LINK NỘI DUNG: [G\\_232IS2902\\_08 - Google Drive](#)

Giảng viên: Nguyễn Quang Phúc

Nhóm 8

Mã học phần 232IS2902

Thành phố Hồ Chí Minh, tháng 4, 2024

## LỜI CẢM ƠN

Nhóm chúng em xin gửi lời cảm ơn chân thành đến thầy Nguyễn Quang Phúc đã nhiệt tình hỗ trợ nhóm chúng em hoàn thành đồ án môn học Phân tích dữ liệu với R/Python thông qua sự hướng dẫn, mở rộng kiến thức và giải đáp thắc mắc trong suốt quá trình thực hiện. Do kiến thức và khả năng nghiên cứu còn hạn chế nên trong quá trình thực hiện có thể sẽ mắc phải một số sai sót nhỏ. Nhóm chúng em rất mong nhận được những phản hồi mang tính xây dựng từ thầy để hoàn thiện đồ án này hơn. Những nhận xét và đánh giá của thầy không chỉ hữu ích cho nhóm chúng em trong môn học này mà còn góp phần vào việc học tập của chúng em trong các môn học sau này.

Tp. Hồ Chí Minh, tháng 4, 2024

Chữ ký

Nhóm 8

## **LỜI CAM KẾT**

Nhóm chúng em xin cam đoan kết quả đồ án báo cáo do chính các thành viên trong nhóm nghiên cứu và thực hiện với sự hướng dẫn của thầy Nguyễn Quang Phúc và không sao chép bất kỳ tài liệu nào khác. Trong toàn bộ nội dung được trình bày trong đồ án nghiên cứu là do nhóm tổng hợp, phân tích và nghiên cứu. Các số liệu, kết quả trong nghiên cứu, các tài liệu nghiên cứu liên quan, tài liệu tham khảo đều và đã được dẫn cụ thể, đầy đủ.

## THÀNH VIÊN NHÓM

STT	Họ và tên	MSSV	Đóng góp
1	Phạm Thanh Thảo (Nhóm trưởng)	K214110852	100%
2	Trần Thị Minh Hân	K214110831	100%
3	Lê Nhật Thành	K214110851	100%
4	Lê Bảo Minh	K214110840	100%
5	Lê Quang Thành Tài	K214111962	100%

# MỤC LỤC

<b>LỜI CẢM ƠN .....</b>	<b>ii</b>
<b>LỜI CAM KẾT .....</b>	<b>iii</b>
<b>THÀNH VIÊN NHÓM .....</b>	<b>iv</b>
<b>MỤC LỤC .....</b>	<b>v</b>
<b>DANH MỤC HÌNH ẢNH .....</b>	<b>xii</b>
<b>DANH MỤC BẢNG .....</b>	<b>xviii</b>
<b>CHƯƠNG 1: CƠ SỞ LÝ THUYẾT .....</b>	<b>1</b>
<b>1.1. Cơ sở chọn biến .....</b>	<b>1</b>
1.1.1. Lựa chọn biến giám sát .....	2
1.1.2. Lựa chọn biến không giám sát .....	2
1.1.3. Đa cộng tuyến .....	2
<b>1.2. Mô hình hồi quy tuyến tính .....</b>	<b>3</b>
<b>1.3. Mô hình hồi quy logistic .....</b>	<b>5</b>
1.3.1. Giá trị ngưỡng .....	6
1.3.2. Diễn giải tham số của mô hình hồi quy Logistic .....	7
1.3.3. Biến phụ thuộc là biến thứ tự .....	7
<b>1.4. Phân tích dữ liệu chuỗi thời gian .....</b>	<b>9</b>
1.4.1. Mô hình ARIMA .....	9
1.4.1.1. Cấu trúc mô hình ARIMA .....	9
1.4.1.2. Công thức mô hình ARIMA (p, d, q) .....	10
1.4.2. Mô hình Holt Winter .....	11
1.4.3. Nhiễu trắng .....	13
<b>1.5. Mô hình máy học .....</b>	<b>14</b>
1.5.1. Decision tree .....	14

1.5.1.1. Cấu trúc của cây quyết định .....	14
1.5.1.2. Phân loại .....	15
1.5.1.3. Xây dựng Decision Tree .....	15
1.5.1.4. Cắt tỉa (Pruning) .....	17
1.5.2. Random Forest.....	18
1.5.3. Neural Network.....	19
1.5.3.1. Perceptron cơ bản .....	19
1.5.3.2. Kiến trúc mạng Neural Network .....	19
1.5.4. Support Vector Machine.....	19
1.5.5. So sánh các mô hình .....	20
<b>1.6. Đánh giá mô hình .....</b>	<b>21</b>
1.6.1. Bài toán phân loại .....	21
1.6.1.1. Độ chính xác (Accuracy) .....	21
1.6.1.2. Precision .....	21
1.6.1.3. Recall .....	22
1.6.1.4. F1 Score .....	22
1.6.2. Bài toán hồi quy .....	23
1.6.2.1. MSE .....	23
1.6.2.2. MAE .....	23
1.6.2.3. RMSE .....	24
<b>1.7. Một số vấn đề về xây dựng mô hình .....</b>	<b>25</b>
1.7.1. Độ lệch và phương sai .....	25
1.7.1.1. Độ lệch (Bias) .....	25
1.7.1.2. Phương sai (Variance) .....	25
1.7.2. Overfitting và Underfitting .....	26

1.7.2.1. Overfitting .....	26
1.7.2.1. Underfitting .....	28
<b>CHƯƠNG 2: MÔ HÌNH HỒI QUY TUYẾN TÍNH .....</b>	<b>30</b>
<b>2.1. Mô tả bài toán .....</b>	<b>30</b>
2.1.1. Đặt vấn đề .....	30
2.1.2. Bộ dữ liệu .....	30
2.1.3. Xây dựng câu hỏi nghiên cứu .....	31
<b>2.2. Tiền xử lý dữ liệu .....</b>	<b>31</b>
2.2.1. Tổng quan cấu trúc bộ dữ liệu .....	31
2.2.2. Làm sạch dữ liệu .....	34
2.2.2.1. Xử lý dữ liệu trùng lặp và không hợp lệ .....	34
2.2.2.2. Xử lý dữ liệu thiếu (NaN) .....	34
2.2.2.3. Xử lý các ngoại lệ .....	35
2.2.3. Phân tích khám phá dữ liệu (EDA) .....	39
2.2.3.1. Phân tích phân phối dữ liệu .....	39
2.2.3.2. Phân tích tương quan bằng biểu đồ phân tán .....	41
2.2.3.3. Phân tích tương quan bằng biểu đồ nhiệt .....	43
<b>2.3. Xây dựng mô hình hồi quy tuyến tính đơn biến .....</b>	<b>45</b>
2.3.1. Xây dựng mô hình .....	45
2.3.2. Đánh giá mô hình .....	46
2.3.2.1. Đánh giá mô hình trên tập huấn luyện .....	46
2.3.2.2. Đánh giá mô hình trên tập thử nghiệm .....	47
2.3.2.3. Kết luận và dự đoán .....	48
<b>2.4. Xây dựng mô hình hồi quy tuyến tính đa biến .....</b>	<b>49</b>
2.4.1. Xác định biến độc lập .....	49

2.4.2. Xây dựng mô hình .....	52
2.4.3. Đánh giá mô hình .....	53
2.4.3.1. Đánh giá mô hình trên tập huấn luyện .....	53
2.4.3.2. Đánh giá mô hình trên tập thử nghiệm .....	54
2.4.3.3. Kết luận và dự đoán .....	56
<b>CHƯƠNG 3: MÔ HÌNH HỎI QUY LOGISTIC .....</b>	<b>57</b>
<b>    3.1. Mô tả bài toán .....</b>	<b>57</b>
3.1.1. Đặt vấn đề .....	57
3.1.2. Bộ dữ liệu .....	57
3.1.3. Xây dựng câu hỏi nghiên cứu .....	60
<b>    3.2. Làm sạch và khám phá dữ liệu .....</b>	<b>61</b>
3.2.1. Tổng quan cấu trúc bộ dữ liệu .....	61
3.2.2. Làm sạch dữ liệu .....	65
3.2.2.1. Xử lý dữ liệu trùng lặp và không hợp lệ .....	65
3.2.2.2. Xử lý dữ liệu thiếu (NaN) .....	66
3.2.2.3. Xử lý giá trị thiếu (NaN) từng cột .....	69
3.2.2.4. Xử lý ngoại lệ .....	79
3.2.3 Phân tích khám phá dữ liệu .....	81
3.2.3.1. Tổng quan .....	81
3.2.3.2. Phân tích từng biến .....	84
<b>    3.3. Chuẩn bị dữ liệu .....</b>	<b>93</b>
3.3.1. Chuyển đổi biến sang dạng nhị phân: .....	93
3.3.2. Chuyển đổi biến nhiều cấp độ (dummies): .....	94
3.3.3. Xác định tương quan giữa các biến .....	94
3.3.4. Chia dữ liệu huấn luyện (train) và kiểm chứng (test) .....	96

<b>3.4. Xây dựng mô hình hồi quy Logistic .....</b>	<b>97</b>
3.4.1. Đơn biến .....	97
3.4.1.1. Biểu đồ phân tán và xây dựng mô hình .....	97
3.4.1.2. Đánh giá mô hình với dữ liệu tập huấn .....	98
3.4.1.3. Đánh giá mô hình dựa trên dữ liệu kiểm thử .....	101
3.4.1.4. Kết luận .....	102
3.4.2. Đa biến .....	103
3.4.2.1. Lựa chọn biến độc lập .....	103
3.4.2.2. Đánh giá mô hình trên tập dữ liệu huấn luyện .....	108
3.4.2.3. Đánh giá mô hình với dữ liệu kiểm tra .....	113
3.4.2.4. Kết luận .....	114
<b>CHƯƠNG 4: MÔ HÌNH CHUỖI THỜI GIAN .....</b>	<b>119</b>
<b>4.1. Thu thập dữ liệu .....</b>	<b>119</b>
<b>4.2. Mô tả bài toán .....</b>	<b>120</b>
4.2.1. Đặt vấn đề .....	120
4.2.2. Bộ dữ liệu .....	120
4.2.3. Câu hỏi nghiên cứu .....	121
<b>4.3. Phân tích khám phá dữ liệu .....</b>	<b>121</b>
4.3.1. Kiểm tra dữ liệu và tổng quan số liệu .....	121
4.3.2. Phân tích dữ liệu .....	124
4.3.3. Xu hướng .....	128
4.3.4. Kiểm tra tính dừng .....	130
<b>4.4. Chuyển đổi dữ liệu thành chuỗi dừng và tự tương quan .....</b>	<b>132</b>
4.4.1. Chuyển đổi dữ liệu thành chuỗi dừng .....	132
4.4.2. Kiểm tra tính dừng của chuỗi dữ liệu đã lấy sai phân .....	134

4.4.3. Kiểm định tính tự tương quan .....	135
<b>4.5. Xây dựng mô hình .....</b>	<b>136</b>
4.5.1. Mô hình ARIMA .....	136
4.5.2. Mô hình SARIMAX .....	137
<b>4.6. Xây dựng mô hình Holt-Winters .....</b>	<b>140</b>
4.6.1. Chuyển đổi dữ liệu .....	140
4.6.2. Kiểm tra tính xu hướng, mùa vụ, chu kỳ .....	142
4.6.3. Kiểm tra mô hình phân rã chuỗi thời gian .....	144
4.6.4. Tối ưu hóa siêu tham số (Hyperparameter Optimization) .....	145
4.6.5. Dự báo bằng mô hình Holt-Winters .....	146
<b>4.7. Kết luận .....</b>	<b>147</b>
<b>CHƯƠNG 5: MÔ HÌNH HỌC MÁY .....</b>	<b>149</b>
<b>5.1. Mô tả bài toán .....</b>	<b>149</b>
5.1.1. Đặt vấn đề .....	149
5.1.2. Bộ dữ liệu .....	149
5.1.3. Câu hỏi nghiên cứu .....	150
<b>5.2. Tiền xử lý dữ liệu .....</b>	<b>151</b>
5.2.1. Tổng quan cấu trúc bộ dữ liệu .....	151
5.2.2. Làm sạch dữ liệu .....	153
5.2.2.1. Xử lý dữ liệu trùng lặp và không hợp lệ .....	153
5.2.2.2. Xử lý dữ liệu thiếu (NaN) .....	153
5.2.2.3. Xử lý ngoại lệ .....	154
5.2.3. Phân tích khám phá dữ liệu .....	156
5.2.3.1. Biến liên tục .....	156
5.2.3.2. Biến phân loại và biến nhị phân .....	158

<b>5.3. Chuẩn bị dữ liệu .....</b>	<b>165</b>
5.3.1. Chuyển đổi dữ liệu .....	165
5.3.2. Xử lý mất cân bằng dữ liệu .....	166
5.3.3. Phân dữ liệu thành các phần huấn luyện và kiểm thử .....	167
<b>5.4. Thực hiện các mô hình học máy .....</b>	<b>167</b>
5.4.1. Decision Tree .....	167
5.4.1.1. Tuning mô hình huấn luyện với các tham số ban đầu .....	167
5.4.1.2. Đánh giá mô hình với dữ liệu huấn luyện .....	170
5.4.1.3. Đánh giá mô hình với dữ liệu kiểm tra .....	172
5.4.1.4. Tổng kết .....	174
5.4.2 Random Forest .....	174
5.4.2.1. Tìm hiểu tham số và huấn luyện .....	174
5.4.2.2. Đánh giá mô hình .....	176
5.4.3. Neural Network .....	178
5.4.3.1. Tìm hiểu tham số và huấn luyện .....	178
5.4.3.2. Đánh giá mô hình .....	180
5.4.4. Support Vector Machine .....	181
5.4.4.1. Tạo mô hình phân loại SVM .....	181
5.4.4.2. Đánh giá mô hình .....	183
5.4.5. Đánh giá và so sánh hiệu quả giữa các mô hình .....	186
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>188</b>

## **DANH MỤC HÌNH ẢNH**

Hình 1.1 Hai mô hình lựa chọn biến .....	1
Hình 1.2 Minh họa hồi quy tuyến tính .....	4
Hình 1.3 Biểu đồ Sigmoid .....	6
Hình 1.4 Ví dụ biến phụ thuộc là biến thứ tự .....	8
Hình 1.5 Mối quan hệ giữa độ lệch và phương sai .....	26
Hình 1.6 Dừng sớm (early stopping) .....	27
Hình 1.7 K-fold cross-validation .....	28
Hình 2.1 Biểu đồ Boxplot mô tả dữ liệu trong dataset trước khi clean .....	36
Hình 2.2 Biểu đồ phân phối dữ liệu #1 .....	39
Hình 2.3 Biểu đồ phân phối dữ liệu #2 .....	40
Hình 2.4 Biểu đồ phân tán dữ liệu #1 .....	41
Hình 2.5 Biểu đồ phân tán dữ liệu #2 .....	42
Hình 2.6 Biểu đồ Heatmap thể hiện tương quan của dữ liệu .....	43
Hình 2.7 Biểu đồ đánh giá mô hình hồi quy đơn biến trên tập huấn luyện .....	46
Hình 2.8 Biểu đồ đánh giá mô hình hồi quy đơn biến trên tập thử nghiệm .....	47
Hình 2.9 Kết quả tất cả p-value đạt điều kiện (lần 1) .....	50
Hình 2.10 Kết quả tất cả p-value đạt điều kiện (lần 2) .....	51
Hình 2.11 Biểu đồ đánh giá mô hình hồi quy đa biến trên tập huấn luyện .....	53
Hình 2.12 Biểu đồ đánh giá mô hình hồi quy đa biến trên tập thử nghiệm .....	55
Hình 3.1 Trực quan dữ liệu thiếu .....	69
Hình 3.2 Trực quan dữ liệu “City” .....	70
Hình 3.3 Trực quan dữ liệu “Specialization” .....	71
Hình 3.4 Trực quan dữ liệu “What matters most to you in choosing a course” .....	72
Hình 3.5 Trực quan dữ liệu “What is your current occupation” .....	73

Hình 3.6 Trực quan dữ liệu “Country”.....	74
Hình 3.7 Trực quan dữ liệu “TotalVisits” .....	75
Hình 3.8 Trực quan dữ liệu “Page Views Per Visit” .....	75
Hình 3.9 Trực quan dữ liệu “Last Activity” .....	76
Hình 3.10 Trực quan dữ liệu “Lead Source” .....	77
Hình 3.11 Trực quan các giá trị ngoại lai .....	80
Hình 3.12 Trực quan hóa dữ liệu nhị phân .....	83
Hình 3.13 Giá trị của cột “Lead Origin” và tỷ lệ chuyển đổi .....	85
Hình 3.14 Giá trị của cột “Lead Source” và tỷ lệ chuyển đổi .....	86
Hình 3.15 Giá trị của cột “Lead Source” và tỷ lệ chuyển đổi sau khi gộp các giá trị ..	87
Hình 3.16 Giá trị của cột “Last Activity” và tỷ lệ chuyển đổi .....	87
Hình 3.17 Giá trị của cột “Last Activity” và tỷ lệ chuyển đổi sau khi gộp các giá trị ..	88
Hình 3.18 Giá trị của cột “Country” và tỷ lệ chuyển đổi .....	88
Hình 3.19 Giá trị của cột “Specialization” và tỷ lệ chuyển đổi .....	89
Hình 3.20 Giá trị của cột “Specialization” và tỷ lệ chuyển đổi sau khi gộp các giá trị	90
Hình 3.21 Giá trị của cột “What is your current occupation” và tỷ lệ chuyển đổi .....	90
Hình 3.22 Giá trị của cột “What matters most to you in choosing a course” và tỷ lệ chuyển đổi .....	91
Hình 3.23 Giá trị của cột “City” và tỷ lệ chuyển đổi .....	92
Hình 3.24 Giá trị của cột “Total Visits”, “Total time spent on website” và “Page views per visit” và tỷ lệ chuyển đổi .....	93
Hình 3.25 Biểu đồ Heat map thể hiện ương quan giữa các biến .....	95
Hình 3.26 Mối quan hệ giữa “Total Time Spent on Website” và “Conversion” .....	97
Hình 3.27 Biểu đồ ma trận thống kê giá trị thực tế và giá trị dự đoán của mô hình đơn biến trên tập huấn luyện .....	98
Hình 3.28 Kết quả phân loại của mô hình trên tập huấn luyện .....	99

Hình 3.29 Đường cong ROC .....	100
Hình 3.30 Trục quan hàm Sigmoid .....	100
Hình 3.31 Biểu đồ ma trận thống kê giá trị thực tế và giá trị dự đoán của mô hình đơn biến trên tập huấn luyện .....	101
Hình 3.32 Giá trị Accuracr kiểm chéo với số lượng biến tăng dần .....	104
Hình 3.33 Chạy mô hình với 29 biến .....	104
Hình 3.34 Chạy mô hình với 19 biến .....	105
Hình 3.35 Kết quả kiểm tra VIF lần 1 .....	106
Hình 3.36 Chạy mô hình với 17 biến .....	107
Hình 3.37 Kết quả kiểm tra VIF lần 2 .....	107
Hình 3.38 Biểu đồ ma trận thống kê giá trị thực tế và giá trị dự đoán của mô hình đơn biến trên tập huấn luyện .....	108
Hình 3.39 Đường cong ROC .....	108
Hình 3.40 Trade off giữa sensitivity và specificity .....	109
Hình 3.41 Ma trận hỗn loạn ở điểm cắt 0.37 .....	110
Hình 3.42 Trade off giữa Precision và Recall .....	111
Hình 3.43 Ma trận hỗn loạn ở điểm cắt 0.44 .....	112
Hình 3.44 Ma trận hỗn loạn với dữ liệu kiểm thử .....	113
Hình 3.45 Trục quan hệ số coef của các biến độc lập .....	114
Hình 4.1 Biểu đồ nhiệt tương quan giữa các biến #1 .....	125
Hình 4.2 Biểu đồ nhiệt tương quan giữa các biến #2 .....	126
Hình 4.3 Tập huấn luyện và tập kiểm tra của bài nghiên cứu .....	127
Hình 4.4 Biểu đồ trực quan sau khi rolling dữ liệu .....	128
Hình 4.5 . Biểu đồ phân rã chuỗi dữ liệu .....	129
Hình 4.6 Kết quả kiểm định ADF và KPSS .....	130
Hình 4.7 . Biểu đồ thể hiện tính tự tương quan .....	131

Hình 4.8 Tương quan PACF .....	131
Hình 4.9 Tương quan ACF .....	132
Hình 4.10 Biểu đồ thể hiện dữ liệu ban đầu và sau khi lấy sai phân bậc 1 .....	133
Hình 4.11 Kết quả kiểm định ADF và KPSS sau khi lấy sai phân bậc 1 .....	134
Hình 4.12 Tương quan PACF sau khi lấy sai phân bậc 1 .....	135
Hình 4.13 . Tương quan ACF sau khi lấy sai phân bậc 1 .....	136
Hình 4.14 Kết quả chạy mô hình Auto ARIMA #1 .....	137
Hình 4.15 Kết quả chạy mô hình Auto ARIMA #2 .....	138
Hình 4.16 . Biểu đồ đánh giá tính hợp lệ của các giả định được đưa ra bởi mô hình Auto ARIMA .....	138
Hình 4.17 Kết quả mô hình .....	139
Hình 4.18 Tính toán chỉ số MSE, RMSE và MAPE .....	139
Hình 4.19 Kết quả dự đoán giá đóng cửa .....	140
Hình 4.20 Biểu đồ so sánh giữa RMSE và Baseline RMSE .....	140
Hình 4.21 Mô tả dữ liệu sau khi chuyển đổi theo chu kỳ tháng và 4 dòng đầu của bộ dữ liệu .....	141
Hình 4.22 Bộ dữ liệu sau khi chuyển đổi theo chu kỳ tháng .....	142
Hình 4.24 Kết quả ước lượng chu kỳ .....	142
Hình 4.25 Kết quả ước lượng xu hướng .....	143
Hình 4.26 Biểu đồ so sánh giá trị thật, chu kỳ và kỳ vọng .....	143
Hình 4.27 . Biểu đồ phân rã chuỗi dữ liệu theo mô hình nhân tính .....	144
Hình 4.28 Biểu đồ phân rã chuỗi dữ liệu theo mô hình cộng tính .....	145
Hình 4.29 Kết quả ước lượng tổ hợp siêu tham số tối ưu nhất .....	145
Hình 4.30 Tập huấn luyện và tập kiểm tra của dataset mới .....	146
Hình 4.31 Kết quả dự đoán giá đóng cửa chứng khoán bằng mô hình Holt-Winters	147
Hình 4.32 Tính toán chỉ số MSE, RMSE và MAPE .....	147

Hình 5.1 Biểu đồ Boxplot thể hiện các giá trị ngoại lai .....	156
Hình 5.2 Biểu đồ thể hiện phân bố của các biến liên tục .....	157
Hình 5.3 Biểu đồ thể hiện phân bố của biến “Y” .....	159
Hình 5.4 Biểu đồ thể hiện phân bố của biến “Month” .....	160
Hình 5.5 Biểu đồ thể hiện phân bố của biến “Contact” .....	160
Hình 5.6 Biểu đồ thể hiện phân bố của biến “Marital” .....	161
Hình 5.7 Biểu đồ thể hiện phân bố của biến “Housing” .....	162
Hình 5.8 Biểu đồ thể hiện phân bố của biến “Loan” .....	163
Hình 5.9 Biểu đồ thể hiện phân bố của biến “Job” .....	163
Hình 5.10 Biểu đồ thể hiện phân bố của biến “Poutcome” .....	164
Hình 5.11 Biểu đồ thể hiện phân bố của biến “Default” .....	164
Hình 5.12 Biểu đồ thể hiện phân bố của biến “Education” .....	165
Hình 5.13 Số lượng lớp 0 và 1 của bộ dữ liệu trước khi Smoteen .....	166
Hình 5.14 Số lượng lớp 0 và 1 của bộ dữ liệu sau khi Smoteen .....	167
Hình 5.15 Đồ thị biểu diễn sự hội tụ của mô hình theo các giá trị max_depth khác nhau .....	169
Hình 5.16 Báo cáo số liệu của Decision Tree trên dữ liệu Train .....	170
Hình 5.17 Confusion Matrix của Decision Tree cho dữ liệu Train .....	171
Hình 5.18 Báo cáo số liệu của Decision Tree trên dữ liệu Test .....	172
Hình 5.19 Confusion Matrix và Biểu đồ ROC Curve Của Decision Tree cho dữ liệu Test .....	173
Hình 5.20 Trực quan cây quyết định với max_depth = 9 .....	174
Hình 5.21 Confusion Matrix của Random Forest cho dữ liệu Test .....	176
Hình 5.22 Báo cáo số liệu của Random Forest trên dữ liệu Test .....	177
Hình 5.23 Confusion Matrix của Neural Network cho dữ liệu Test .....	180
Hình 5.24 Báo cáo số liệu của Neural Network trên dữ liệu Test .....	181

Hình 5.25 Kết quả mô hình .....	182
Hình 5.26 Báo cáo số liệu của Support Vector Machine trên dữ liệu Train .....	183
Hình 5.27 Confusion Matrix và Roc Curve cho dữ liệu Train .....	184
Hình 5.28 Báo cáo số liệu của Support Vector Machine trên dữ liệu Test .....	184
Hình 5.29 Confusion Matrix và Roc Curve cho Test Data .....	185

## DANH MỤC BẢNG

Bảng 1.1 So sánh các mô hình .....	21
Bảng 2.1 . Mô tả ý nghĩa các biến thuộc bộ dữ liệu .....	31
Bảng 2.2 . Tổng quan về các biến thuộc bộ dữ liệu .....	32
Bảng 2.3 Mô tả thống kê lần 1 #1 .....	33
Bảng 2.4 Mô tả thống kê lần 1 #2 .....	34
Bảng 2.5 Mô tả thống kê các giá trị thiếu .....	35
Bảng 2.6 Mô tả thống kê lần 2 #1 .....	38
Bảng 2.7 Mô tả thống kê lần 2 #1 .....	38
Bảng 2.8 Thống kê tương quan của các biến với “Medv” .....	44
Bảng 2.9 Kết quả tất cả VIF đạt điều kiện (lần 1) .....	51
Bảng 2.10 Kết quả tất cả VIF đạt điều kiện (lần 2) .....	52
Bảng 3.1 Mô tả dữ liệu .....	60
Bảng 3.2 . Tổng quan về các biến thuộc bộ dữ liệu .....	63
Bảng 3.3 Mô tả thống kê #1 .....	64
Bảng 3.4 Mô tả thống kê #2 .....	65
Bảng 3.5 Mô tả thống kê các giá trị thiếu .....	68
Bảng 3.6 Kiểm tra giá trị thiếu .....	79
Bảng 3.7 Thống kê phân phối các cột có giá trị ngoại lai .....	80
Bảng 3.8 Số giá trị từng biến .....	82
Bảng 3.9 Sự chênh lệch giữa số lượng giá trị biến “Yes” và “No” .....	84
Bảng 3.10 Thống kê các biến có độ tương quan cao .....	96
Bảng 3.11 Hiệu suất của mô hình phân loại .....	99
Bảng 3.12 Các chỉ số đánh giá mô hình ở điểm cắt 0.37 .....	110
Bảng 3.13 Các chỉ số đánh giá mô hình ở điểm cắt 0.44 .....	112

Bảng 3.14 Các chỉ số đánh giá mô hình với dữ liệu kiểm thử.....	114
Bảng 3.15 Thống kê giá trị coef của các biến độc lập .....	115
Bảng 4.1 Mô tả tiêu đề các cột của bộ dữ liệu VIC_2007_2024.....	121
Bảng 4.2 Tổng quan về các biến thuộc bộ dữ liệu VIC_2007_2024 .....	122
Bảng 4.3 Mô tả thống kê bộ dữ liệu VIC_2007_2024 #1 .....	123
Bảng 4.4 Mô tả thống kê bộ dữ liệu VIC_2007_2024 #2 .....	123
Bảng 4.5 Thống kê số giá trị bị thiếu của từng biến trong dataset .....	124
Bảng 4.6 Tính chỉ số AIC để xác định mô hình ARIMA tối ưu nhất.....	137
Bảng 5.1 Mô tả tiêu đề các cột của bộ dữ liệu .....	150
Bảng 5.2 Tổng quan về các biến thuộc bộ dữ liệu .....	152
Bảng 5.3 Mô tả thống kê .....	153
Bảng 5.4 Mô tả thống kê các giá trị thiếu .....	154
Bảng 5.5 Thống kê các giá trị duy nhất ở mỗi cột .....	158
Bảng 5.6 So sánh kết quả giữa các max_depth và min_samples_split tương ứng .....	169
Bảng 5.7 Thống kê các chỉ số đánh giá của các mô hình .....	186

# CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

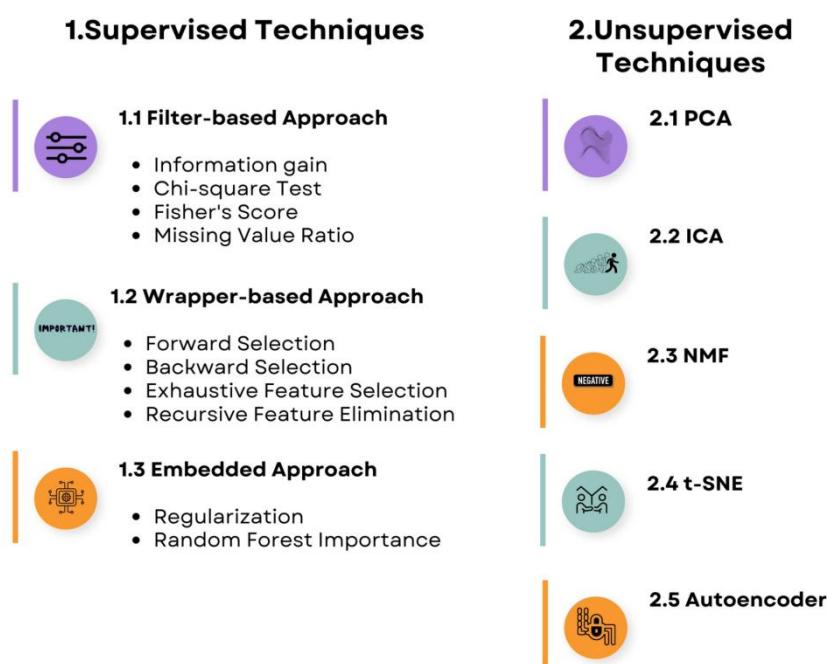
## 1.1. Cơ sở chọn biến

Một số vấn đề về mô hình dự đoán có số lượng lớn các biến có thể làm chậm quá trình phát triển và huấn luyện mô hình cũng như yêu cầu một lượng lớn bộ nhớ hệ thống. Ngoài ra, hiệu suất của một số mô hình có thể suy giảm khi đưa vào các biến đầu vào không liên quan đến biến mục tiêu.

Có hai loại lựa chọn biến chính là kỹ thuật học máy có giám sát và không giám sát:

- *Mô hình được giám sát*: Sử dụng lớp nhãn đầu ra để lựa chọn đối tượng, sử dụng các biến mục tiêu để chọn ra các biến có thể tăng tính hiệu quả cho mô hình. Sẽ sử dụng các độ đo trong thống kê như spearman, pearson để loại bỏ các biến dư thừa. Ví dụ khi 2 biến có độ tương quan cao thì chúng ta sẽ chỉ giữ lại 1 trong 2 biến đó.
- *Mô hình không giám sát*: Không cần lớp nhãn đầu ra để lựa chọn đối tượng.

## Advanced Feature Selection Techniques



Hình 1.1 Hai mô hình lựa chọn biến

### **1.1.1. Lựa chọn biến giám sát**

Một số phương pháp phổ biến là:

- *Bộ lọc (Filter Method)*: các biến được loại bỏ dựa trên mức độ tương quan của chúng với đầu ra. Đánh giá xem các tính năng có tương quan tích cực hoặc tiêu cực với nhãn đầu ra không, sau đó loại bỏ các tính năng không phù hợp. Để xác định mối quan hệ này, chúng ta có thể sử dụng một số phương pháp thống kê khác nhau, đây là cách thức được sử dụng rất rộng rãi.
- *Phương pháp Wrapper*: quá trình thêm và bớt các biến được thực hiện một cách lặp đi lặp lại, thông qua việc đánh giá hiệu suất của mô hình sau mỗi lần điều chỉnh. Quá trình này tiếp tục cho đến khi đạt được một tiêu chí dừng được xác định trước, chẳng hạn như đạt được độ chính xác mong muốn hoặc khi không còn cải thiện đáng kể trong hiệu suất của mô hình. Điều này giúp tối ưu hóa số lượng và chất lượng của các biến được chọn để cải thiện hiệu suất dự đoán của mô hình.

### **1.1.2. Lựa chọn biến không giám sát**

Một số phương pháp phổ biến là:

- *Principal Component Analysis (PCA)*: Phương pháp này giảm chiều dữ liệu bằng cách tìm kiếm các thành phần chính (principal components) của dữ liệu, tức là các hướng có phương sai lớn nhất. Các thành phần này thường tóm tắt các đặc tính quan trọng nhất của dữ liệu.
- *Independent Component Analysis (ICA)*: Tương tự như PCA, ICA cũng là một phương pháp giảm chiều dữ liệu nhưng tập trung vào việc tách các thành phần độc lập từ dữ liệu gốc.

### **1.1.3. Đa cộng tuyến**

Đa cộng tuyến là hiện tượng mà các biến độc lập trong mô hình có mối quan hệ tuyến tính mạnh với nhau. Có hai trường hợp đa cộng tuyến: hoàn hảo và không hoàn hảo. Trong trường hợp hoàn hảo, các biến độc lập liên quan một cách hoàn hảo và không thể ước lượng đồng thời và thường loại bỏ một trong hai biến, chỉ ước lượng hệ số hồi quy cho biến còn lại mà thôi. Dấu hiệu nhận biết hiện tượng đa cộng tuyến:

- Dựa vào ma trận tương quan Pearson.
- Dựa vào giá trị VIF ở hồi quy: Nếu  $VIF < 10$  thì không có hiện tượng đa cộng tuyến xảy ra. Tuy nhiên, mốc đánh giá ở mức 10 sẽ phù hợp với những đề tài về kỹ thuật, vật lý không sử dụng thang đo Likert. Về kinh tế, xã hội, các nhà nghiên cứu cho rằng  $VIF > 2$  sẽ có hiện tượng đa cộng tuyến xảy ra.
- Hệ số  $R^2$  lớn nhưng tỷ số t nhỏ.
- Tương quan cặp giữa các biến giải thích cao.

Đa cộng tuyến sẽ gây ra một số hậu quả như tỉ số t không có ý nghĩa khi kiểm định giả thuyết  $H_0$ , thường sử dụng tỷ số t và so sánh giá trị ước lượng của t với giá trị bảng tới hạn của t. Trong trường hợp đa cộng tuyến cao thì sai số chuẩn sẽ rất lớn và do đó làm giá trị t nhỏ đi, sẽ làm tăng chấp nhận giả thuyết  $H_0$ . Hoặc một số hậu quả khác như khoảng tin cậy rộng hơn,  $R^2$  cao nhưng tỉ số t ít có ý nghĩa, ... Một số cách khác phục hiện tượng đa cộng biến là:

- Dùng thông tin tiên nghiệm.

$$\text{Ví dụ } Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \text{ biết } \beta_3 = 0.1\beta_2$$

Thực hiện biến đổi

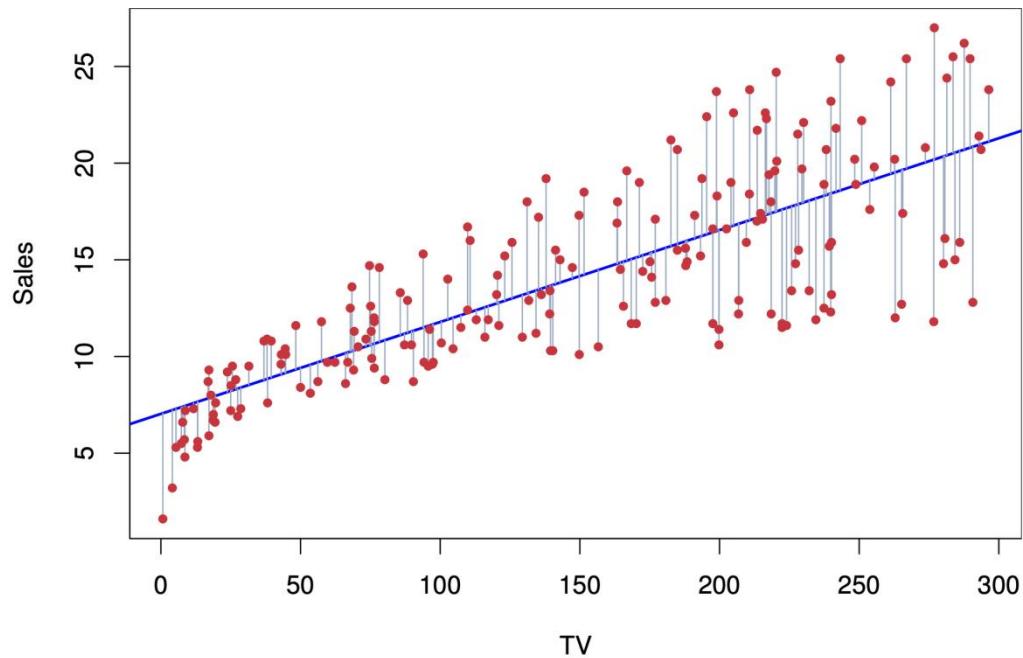
$$Y_i = \beta_1 + \beta_2 X_{2i} + 0.1\beta_2 X_{3i} + u_i$$

$$Y_i = \beta_1 + \beta_2 X_i + u_i \text{ với } X_i = X_{2i} + 0.1 X_{3i}$$

- Loại trừ một biến giải thích ra khỏi mô hình.
- Bổ sung thêm dữ liệu hoặc chọn mẫu mới.

## 1.2. Mô hình hồi quy tuyến tính

Hồi quy là các phương pháp để xây dựng mối quan hệ giữa điểm dữ liệu x và mục tiêu với giá trị số thực y. Bài toán hồi quy thường sử dụng dự đoán một giá trị số. Các ví dụ phổ biến bao gồm dự đoán giá cả (nhà, cổ phiếu, ...), thời gian bệnh nhân nằm viện, nhu cầu trong ngành bán lẻ,... Được sử dụng để tìm một đường thẳng (hàm tuyến tính) mô tả quan hệ tuyến tính giữa biến phụ thuộc (biến mục tiêu) và biến độc lập (biến dự báo).



*Hình 1.2 Minh họa hồi quy tuyến tính*

### Mô hình hồi quy tổng thể

$$Y = \alpha + \beta X + \varepsilon$$

- $Y$  : biến ngẫu nhiên đại diện cho biến trả lời
- $X$  : biến ngẫu nhiên đại diện cho biến dự đoán (dự đoán, yếu tố rủi ro)
- Cả  $Y$  và  $X$  có thể là biến gián đoạn (ví dụ, có/không) hoặc biến liên tục (tuổi, cân nặng,...). Nếu  $Y$  là biến gián đoạn thì mô hình là hồi quy logistic; nếu  $Y$  là biến liên tục thì mô hình hồi quy tuyến tính đơn giản.

Với các giả định:

- Mỗi liên hệ là tuyến tính đối với các tham số khảo sát;
- $X$  được đo lường không có sai số;
- Các giá trị của  $Y$  độc lập với nhau (e.g.,  $Y_1$  không tương quan với  $Y_2$ );
- Sai số ngẫu nhiên tuân theo phân bố chuẩn với giá trị trung bình mean = 0 và phương sai là hằng số.

### Mô hình cho hồi quy mẫu

$$Y = a + bX + e$$

- $a$  : intercept (điểm cắt của đường thẳng hồi quy và trục  $Y$  khi  $X = 0$ )

- $b$  : slope / gradient (hệ số góc)
- $e$ : random error (sai số ngẫu nhiên - sự thay đổi giữa các đối tượng trong  $y$  ngay cả khi  $x$  là không đổi, ví dụ: sự thay đổi cholesterol ở bệnh nhân cùng độ tuổi.)

## Ưu điểm

- *Đơn giản và dễ triển khai:* Hồi quy tuyến tính không yêu cầu quá nhiều giả định phức tạp và dễ dàng triển khai trên các bài toán đơn giản.
- *Điễn giải:* Kết quả của mô hình hồi quy tuyến tính có thể dễ dàng hiểu và diễn giải, giúp người dùng hiểu rõ hơn về mối quan hệ giữa các biến.

## Nhược điểm

- *Giả định về tuyến tính:* Phương pháp này dựa trên giả định rằng mối quan hệ giữa các biến là tuyến tính, điều này có thể không phù hợp với một số bài toán có mối quan hệ phức tạp hơn.
- *Nhạy cảm với nhiễu:* Nếu dữ liệu chứa nhiễu hoặc giá trị ngoại lai, hiệu suất của mô hình hồi quy tuyến tính có thể bị ảnh hưởng nghiêm trọng.

### 1.3. Mô hình hồi quy logistic

Logistic Regression (Hồi quy logistic) là một kỹ thuật thống kê xem xét mối liên hệ giữa biến độc lập (biến liên tục hoặc nhị phân) và biến phụ thuộc (biến nhị phân). Đây là mô hình giúp chúng ta phải trả lời câu hỏi “có hay không” dựa trên những cân nhắc được/mất của việc ra quyết định, cụ thể đó là xác suất xảy ra khả năng “Có” và xác suất xảy ra khả năng “Không” dựa trên những yếu tố có thể ảnh hưởng đến việc ra quyết định. Trong mô hình hồi quy Logistic, hàm logit được sử dụng để biểu diễn mối quan hệ giữa các biến độc lập và biến phụ thuộc. Hàm logit chuyển đổi giá trị của biến phụ thuộc từ một dải giá trị liên tục sang một dải giá trị giữa 0 và 1.

Hàm logit được biểu diễn như sau:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta x + \varepsilon$$

Trong đó:

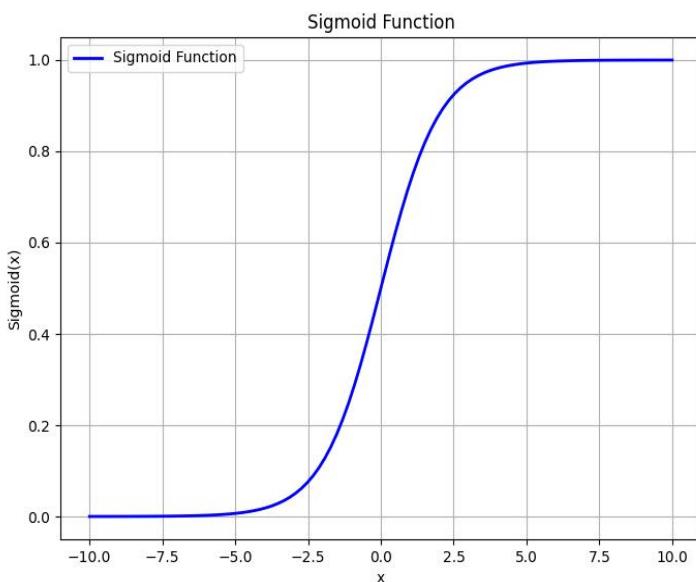
- $y$ : Biến phụ thuộc với 2 trạng thái (0/1; true/false; yes/no)
- $x$ : Biến độc lập

- $p$ : Xác suất của biến phụ thuộc
- $\varepsilon$ : Sai số ngẫu nhiên trong mô hình

### 1.3.1. Giá trị ngưỡng

Hồi quy logistic sử dụng hàm sigmoid để ánh xạ các dự đoán và xác suất tương ứng của chúng. Hàm sigmoid chuyển đổi giá trị của biến đầu vào từ một dải giá trị bất kỳ thành một phạm vi giữa 0 và 1, phản ánh xác suất của một sự kiện xảy ra.

$$\text{sigmoid}(x) = p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$



*Hình 1.3 Biểu đồ Sigmoid*

Từ xác suất vừa thu được, ta cần xác định một giá trị ngưỡng  $t$  ( $0 < t < 1$ ) để dự báo kết quả của mô hình. Thông thường, ta chọn  $t = 0.5$ .

- Nếu  $p \geq t$ : Dự báo  $y = 1$
- Nếu  $p < t$ : Dự báo  $y = 0$

Mô hình hồi quy Logistic có thể được viết theo một cách khác là dùng hệ số Odds. Odds là xác suất biến có xảy ra chia cho xác suất biến không xảy ra. Khi Odds càng lớn thì xác suất để  $y = 1$  càng lớn.

$$Odds = \frac{p}{1 - p} = e^{\alpha + \beta x}$$

$$\log(Odds) = \alpha + \beta x$$

Trong đó

- $p$ : Xác suất biến có xảy ra.
- $1 - p$ : Xác suất biến có không xảy ra.

### 1.3.2. Diễn giải tham số của mô hình hồi quy Logistic

Điễn giải của hệ số  $\beta$  ứng với biến x

- Khi biến x tăng 1 đơn vị thì giá trị  $\log(Odds)$  tăng  $\beta$  đơn vị.
- Khi biến x tăng 1 đơn vị thì tỷ số Odds tăng lên  $\exp(\beta)$  lần.
- $\beta > 0$ : x càng lớn, xác suất để  $y = 1$  càng lớn.
- $\beta < 0$ : x càng lớn, xác suất để  $y = 1$  càng nhỏ.

### 1.3.3. Biến phụ thuộc là biến thứ tự

Biến thứ tự là biến phân loại trong đó các cấp độ có thứ tự tự nhiên (ví dụ: trầm cảm được phân loại là Tối thiểu, Nhẹ, Trung bình, Nặng vừa phải và Nặng). Hồi quy logistic thứ tự có thể được sử dụng để đánh giá mối liên quan giữa các yếu tố dự đoán và kết quả thứ tự. Trong hồi quy logistic nhị phân, kết quả Y có hai cấp độ. Nếu hai mức này được gọi là 0 và 1, và lập mô hình xác suất với  $Y = 1$ . Mô hình hồi quy logistic thứ tự, cho kết quả Y với các mức  $\ell = 1, 2, \dots, L$ .

$$\ln\left(\frac{P(Y \leq \ell)}{P(Y > \ell)}\right) = \zeta_\ell - \eta_1 X_1 - \eta_2 X_2 - \dots - \eta_k X_k$$

Với mỗi mức  $\ell = 1, 2, \dots, L - 1$ . Phương trình chỉ đúng với các mức lên đến  $L - 1$  vì nếu lên tới  $L$  thì chúng ta sẽ có  $P(Y > L) = 0$  trong mẫu số. Cũng giống như hồi quy logistic nhị phân, vé trái là log-tỷ lệ cược của một xác suất, nhưng thay vì xác suất của một kết quả bằng một mức độ, đó là xác suất tích lũy của một kết quả ở bất kỳ mức độ nào cho đến và bao gồm một mức độ xác định. Giải thích các thuật ngữ hồi quy:

- $\zeta_l$  là phần chặn và biểu thị log-odds của  $Y_\ell$  khi tất cả các yếu tố dự đoán ở mức 0 hoặc mức tham chiếu của chúng. Do đó,  $P(Y < \ell)$  là logit nghịch đảo của  $\zeta_l$ . Không giống như các dạng hồi quy khác, mô hình hồi

quy logistic thứ tự có nhiều điểm chặn, một điểm chặn cho mỗi cấp độ Y từ 1 đến L - 1.

- Mỗi  $\eta_k$  với  $k = 1, 2, \dots, K$  là log của tỷ số chênh lệch so sánh chênh lệch  $Y < 1$  giữa các cá thể khác nhau 1 đơn vị ở X (hoặc so sánh các cá thể ở mức  $X_k$  và mức tham chiếu).  $e^{-\eta_k}$  là tỷ lệ chênh lệch (odds ratio) so sánh tỷ lệ chênh lệch của  $Y < 1$  giữa những chênh lệch 1 đơn vị trong  $X_k$ .

### Ví dụ

##	Minimal	Mild Moderate to Severe
##	0.7963	0.9370 1.0000

Hình 1.4 Ví dụ biến phụ thuộc là biến thứ tự

Trong ví dụ biến phụ thuộc có 3 cấp độ, nên chỉ có 2 hệ số chặn. Nói cách khác, với 3 cấp số, logit nghịch đảo của điểm chặn đầu tiên P ( $Y \leq 1$ ) và logit điểm chặn thứ 2 là P ( $Y \leq 2$ ). Không có điểm chặn thứ 3 vì chỉ có 3 mức nên P ( $Y = 3$ ) luôn bằng 1 nên không cần ước tính.

- Hệ số chặn đầu tiên đại diện cho log odds của trạng thái "Minimal depression". Nhìn "Minimal|Mild" cho biết rằng tỷ lệ này tương ứng với xác suất của chán nản ở mức "Minimal" so với mức "Mild" hoặc cao hơn. Nói cách khác, cho biết khả năng một cá nhân có chán nản ở mức "Minimal" so với mức "Mild" hoặc cao hơn.

$$\frac{P(\text{Depression} \leq \text{Minimal})}{P(\text{Depression} > \text{Minimal})} \Leftrightarrow \frac{P(\text{Depression} \leq \text{Minimal})}{P(\text{Depression} \geq \text{Mild})}$$

- Tương tự, hệ số chặn thứ hai đại diện cho log odds của trạng thái "up to Mild depression" (bao gồm cả "Minimal" và "Mild"). Nhìn "Mild|Moderate to Severe" biểu thị rằng tỷ lệ này tương ứng với xác suất của chán nản ở mức "Mild" so với mức "Moderate to Severe" hoặc cao hơn. Điều này giúp hiểu rằng mô hình đang đo lường khả năng của một cá nhân có chán nản ở mức "Mild" so với mức "Moderate to Severe" hoặc cao hơn. Odds tương ứng

$$\frac{P(\text{Depression} \leq \text{Mild})}{P(\text{Depression} \geq \text{Moderate to Severe})}$$

⇒ Kết luận từ dữ liệu là tỷ lệ ước lượng của cá nhân có chứng tỏ mức độ "Minimal depression" là 0.796 và tỷ lệ ước lượng của cá nhân có chứng tỏ mức độ "Minimal or Mild depression" là 0.939.

## 1.4. Phân tích dữ liệu chuỗi thời gian

### 1.4.1. Mô hình ARIMA

Mô hình Arima là một loại mô hình được sử dụng phổ biến trong kinh tế lượng. Có thể hiểu, Arima là mô hình được sử dụng để dự đoán và khai phá các dữ liệu trong ngành tài chính và chứng khoán. Đây là một phương pháp nghiên cứu độc lập thông qua việc dự đoán theo các chuỗi thời gian. Sau đó, các nhà nghiên cứu sẽ sử dụng các thuật toán dự báo độ trễ để đưa ra mô hình phù hợp.

#### 1.4.1.1. Cấu trúc mô hình ARIMA

ARIMA model là viết tắt của cụm từ Autoregressive Intergrated Moving Average. Mô hình sẽ biểu diễn phương trình hồi quy tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) là 3 thành phần chính:

- *Autoregressive (AR)*: Kí hiệu là AR. Đây là thành phần tự hồi quy bao gồm tlop hợp các độ trễ của biến hiện tại. Độ trễ bậc  $\rho$  chính là giá trị lùi về quá khứ  $\rho$  bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ  $\rho$ . Cụ thể, quá trình AR ( $\rho$ ) của chuỗi được biểu diễn như bên dưới:

$$AR(p) = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p}$$

- *Integrated (I)*: Thành phần này xử lý vấn đề tính dừng của chuỗi thời gian. Nếu chuỗi thời gian không dừng, cần thực hiện phép lấy sai phân (differencing) d lần để biến đổi nó thành chuỗi dừng. Tham số d đại diện cho số lần lấy sai phân.
- *Moving Average (MA)*: Quá trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung

bình thường như là một chuỗi nhiễu trắng. Tham số q đại diện cho số lượng giá trị trễ của sai số ngẫu nhiên được sử dụng trong mô hình.

#### 1.4.1.2. Công thức mô hình ARIMA (p, d, q)

Mô hình ARIMA(p, d, q) được biểu diễn bằng công thức sau:

$$\Delta x_t = \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \dots + \phi_p \Delta x_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

Trong đó  $\Delta x_t$  là giá trị sai phân bậc d và  $\epsilon_t$  là các chuỗi nhiễu trắng.

#### *Ưu điểm*

- Đơn giản và dễ hiểu: Mô hình ARIMA có cấu trúc tương đối đơn giản, dễ dàng được giải thích và triển khai.
- Tính linh hoạt: ARIMA có thể áp dụng cho nhiều loại chuỗi thời gian khác nhau, từ đơn giản đến phức tạp.
- Ít nhạy cảm với nhiễu: Mô hình ARIMA ít bị ảnh hưởng bởi các giá trị ngoại lệ hoặc dữ liệu thiếu trong chuỗi thời gian.

#### *Nhược điểm*

- Lựa chọn tham số: Việc lựa chọn các tham số p (số độ trễ tự hồi quy), d (số lần lấy sai phân) và q (số độ trễ trung bình trượt) đóng vai trò quan trọng trong hiệu quả dự báo của mô hình. Tuy nhiên, việc lựa chọn tham số phù hợp có thể gặp nhiều khó khăn và phụ thuộc vào đặc điểm của từng chuỗi thời gian.
- Tính nhạy cảm với thay đổi: ARIMA có thể không phù hợp với những chuỗi thời gian có cấu trúc thay đổi theo thời gian hoặc có các yếu tố theo mùa.
- Giới hạn trong mô tả: Mô hình ARIMA chỉ tập trung vào mối quan hệ tuyến tính giữa các giá trị trong chuỗi thời gian, và không thể mô tả rõ ràng các yếu tố phi tuyến tính hoặc các mẫu không dừng khác.

#### *Vấn đề liên quan đến mô hình ARIMA*

- Tính dừng của chuỗi thời gian: ARIMA yêu cầu chuỗi thời gian phải là chuỗi dừng, nghĩa là giá trị trung bình và phương sai của chuỗi không thay đổi theo thời gian. Nếu chuỗi thời gian không dừng, cần

thực hiện phép lấy sai phân để biến đổi thành chuỗi dừng trước khi áp dụng mô hình ARIMA.

- Tự tương quan: Sai số dự báo của mô hình ARIMA cần phải không có hiện tượng tự tương quan, nghĩa là các sai số không phụ thuộc vào nhau. Nếu tồn tại tự tương quan, cần điều chỉnh mô hình hoặc sử dụng các phương pháp khác để xử lý.
- Phân phối của sai số: Sai số dự báo của mô hình ARIMA cần phải tuân theo phân phối nhiễu trắng, nghĩa là có trung bình bằng 0 và phương sai không đổi. Nếu phân phối sai số không phù hợp, cần điều chỉnh mô hình hoặc sử dụng các phương pháp khác để mô tả sai số.

#### **1.4.2. Mô hình Holt Winter**

Mô hình Holt-Winter là một phương pháp dự báo chuỗi thời gian phổ biến được sử dụng để dự đoán giá trị tương lai dựa trên dữ liệu quá khứ, đặc biệt hữu ích cho các chuỗi thời gian có xu hướng theo mùa. Mô hình này được phát triển dựa trên hai yếu tố chính:

- *Mức độ tron tru (level)*: Biểu thị giá trị trung bình của chuỗi thời gian tại thời điểm dự báo.
- *Xu hướng (trend)*: Biểu thị xu hướng tăng hoặc giảm của chuỗi thời gian theo thời gian.

#### **Phiên bản chính**

- *Holt-Winters (additive)*: Giả định rằng xu hướng và mùa vụ cộng dồn vào mức độ tron tru.
- *Holt-Winters (multiplicative)*: Giả định rằng xu hướng và mùa vụ nhân với mức độ tron tru.
- *Holt-Winters (exponential)*: Giả định rằng xu hướng thay đổi theo thời gian theo hàm mũ.

#### **Công thức mô hình Holt-Winter**

- Mức độ tron tru (level):

$$L_t = \alpha \left( \frac{Y_t}{S_{t-s}} \right) + (1 - \alpha)(L_{t-1} + b_{t-1})$$

- Xu hướng (trend):

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta) * b_{t-1}$$

- Mùa vụ (seasonality):

$$S_t = \gamma \left( \frac{Y_t}{L_t} \right) + (1 - \gamma) * S_{t-s}$$

- Dự báo (forecast):

$$F_{t+m} = (L_t + m * b_t) * S_{t-s+m}$$

*Trong đó:*

$L_t$ : Ước lượng mức độ trung bình của chuỗi thời gian tại thời điểm t

$b_t$ : Dự báo xu hướng tại thời điểm t

$S_t$ : Dự báo mùa vụ tại thời điểm t

$F_{t+m}$ : Dự báo cho m chu kỳ trong tương lai

$\alpha, \beta, \gamma$ : Tham số điều chỉnh

m: Số chu kỳ trong một mùa

s: là số giai đoạn trong một vòng thời vụ

### ***Ưu điểm***

- Đơn giản và dễ sử dụng: Mô hình có cấu trúc tương đối đơn giản, dễ dàng được giải thích và triển khai.
- Hiệu quả với chuỗi thời gian theo mùa: Mô hình có khả năng mô tả tốt các chuỗi thời gian có xu hướng theo mùa.
- Ít tham số: Mô hình chỉ sử dụng một số ít tham số, giúp việc điều chỉnh và tối ưu hóa trở nên dễ dàng hơn.

### ***Nhược điểm***

- Giả định về xu hướng: Mô hình giả định rằng xu hướng là tuyến tính hoặc hàm mũ, điều này có thể không phù hợp với tất cả các chuỗi thời gian.
- Giả định về mùa vụ: Mô hình giả định rằng chu kỳ mùa vụ có độ dài cố định và hình dạng lặp lại, điều này có thể không đúng với một số chuỗi thời gian.

- Nhạy cảm với giá trị ngoại lệ: Mô hình có thể bị ảnh hưởng bởi các giá trị ngoại lệ trong chuỗi thời gian.

### **Vấn đề liên quan đến mô hình Holt-Winter**

- Lựa chọn tham số: Việc lựa chọn các tham số  $\alpha$ ,  $\beta$ ,  $\gamma$  phù hợp đóng vai trò quan trọng trong hiệu quả dự báo của mô hình. Tuy nhiên, việc lựa chọn tham số phù hợp có thể gặp nhiều khó khăn và phụ thuộc vào đặc điểm của từng chuỗi thời gian.
- Kiểm tra tính phù hợp: Cần kiểm tra xem mô hình có phù hợp với chuỗi thời gian hay không bằng cách đánh giá các tiêu chí như độ chính xác dự báo, sự hiện diện của tự tương quan trong sai số dự báo, v.v.
- Điều chỉnh mô hình: Nếu mô hình không phù hợp, cần điều chỉnh mô hình bằng cách thay đổi cấu trúc hoặc sử dụng các phương pháp khác để mô tả xu hướng hoặc mùa vụ.

#### **1.4.3. Nhiễu trắng**

##### **Định nghĩa**

Nhiễu trắng là một chuỗi dừng, trong đó các giá trị của nó được giả sử là độc lập và có cùng phân phối xác suất. Tính đặc trưng của nhiễu trắng là giá trị kỳ vọng (mean) và phương sai (variance) không thay đổi theo thời gian.

Trong trường hợp của nhiễu trắng, mỗi mẫu được sinh ra độc lập và không liên quan đến nhau, do đó không có mối liên hệ hoặc biến động giữa các mẫu.

Nhiễu trắng có thể được xem là sai số ngẫu nhiên trong mô hình hồi quy do nó không thể xác định được trong 1 chuỗi thời gian. Nghiên cứu sẽ xuất hiện ở cuối cùng ở các mô hình AR hay MA. Nếu một chuỗi thời gian là nhiễu trắng, nó là một chuỗi số ngẫu nhiên và không thể được dự đoán. Nghiên cứu là một khái niệm quan trọng trong phân tích và dự báo chuỗi thời gian vì hai lý do chính:

- Dự đoán: Nếu chuỗi thời gian là nhiễu trắng, theo định nghĩa, nó là ngẫu nhiên. Không thể mô hình hóa nó một cách hợp lý và đưa ra dự đoán.
- Dự đoán mô hình: Chuỗi lõi từ một mô hình dự báo chuỗi thời gian nên lý tưởng là nhiễu trắng.

## Điều kiện chuỗi nhiễu trắng

- *Giá trị trung bình (mean)*: Trong trường hợp của nhiễu trắng, giá trị trung bình là 0. Điều này không có nghĩa là tổng của tất cả các giá trị trong chuỗi bằng 0, mà là giá trị trung bình của chuỗi không thay đổi theo thời gian.
- *Kỳ vọng*: Một chuỗi sẽ được gọi là nhiễu trắng nếu thỏa mãn kỳ vọng bằng 0. Trong trường hợp của nhiễu trắng, giá trị trung bình là 0. Điều này không có nghĩa là tổng của tất cả các giá trị trong chuỗi bằng 0, mà là giá trị trung bình của chuỗi không thay đổi theo thời gian.
- *Hiệp phương sai (covariance)*: Hiệp phương sai giữa hai thời điểm trong chuỗi nhiễu trắng là 0. Không có sự phụ thuộc tuyến tính giữa các giá trị tại các thời điểm khác nhau trong chuỗi.  
⇒ Hiệp phương sai đo lường mức độ biến động giữa X và Y. Nếu giá trị hiệp phương sai dương, X và Y có xu hướng di chuyển cùng nhau. Nếu giá trị hiệp phương sai âm, chúng di chuyển ngược lại → Nếu bằng 0 thì X và Y không có mối liên hệ → Mà mô hình thời gian dự đoán tương lai dựa vào tương quan → Do đó mô hình là nhiễu trắng → Không phân tích

## 1.5. Mô hình máy học

### 1.5.1. Decision tree

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Là một mô hình học máy có giám sát được sử dụng để giải quyết các vấn đề phân loại và dự đoán. 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).

#### 1.5.1.1. Cấu trúc của cây quyết định

- *Nút gốc (root node/the root)*: Đây là nút đầu tiên của cây và không có nút cha. Nút gốc đóng vai trò là điểm bắt đầu của cây.
- *Nút lá (leaf node)*: Đây là các nút cuối cùng của cây, không có nút con. Mỗi nút lá thường chứa một giá trị hoặc thông tin cụ thể.

- *Nút trong (internal node)*: Đây là các nút nằm giữa nút gốc và các nút lá. Mỗi nút trong chứa một giá trị và có hai nút con.
- *Nhánh và Cạnh (Branches and Edges)*: Là các mũi tên, hiển thị các kết quả có thể xảy ra của một điều kiện. Chúng dẫn đến các nút hoặc lá con.

#### **1.5.1.2. Phân loại**

- Cây phân loại (Classification Tree): Dự báo giá trị của biến phụ thuộc bằng giá trị có tỷ lệ cao nhất trong nhánh cuối cùng (lá) của cây nhị phân, với biến phụ thuộc là biến phân loại (ví dụ dự đoán giới tính, kết quả của một trận đấu).
- Cây hồi quy (Regression Tree): Dự báo biến phụ thuộc bằng trung bình của biến phụ thuộc tại lá của cây nhị phân, với biến phụ thuộc là biến liên tục (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện).

#### **1.5.1.3. Xây dựng Decision Tree**

Các chỉ số đánh giá lựa chọn thuộc tính: Nếu tập dữ liệu bao gồm N thuộc tính thì việc quyết định đặt thuộc tính nào ở *Root Node* hoặc *Internal Node* là một bước phức tạp. Chọn ngẫu nhiên bất kì thuộc tính làm *Root Node* sẽ khó khăn để phân chia các mẫu vào cùng một lớp. Để giải quyết vấn đề này, chỉ số ASM được sử dụng để đánh giá độ hiệu quả của việc phân tách các nhánh trong Decision Tree. Bằng việc phân tích dữ liệu đầu vào, ASM đưa ra xếp hạng cho các thuộc tính, sau đó lựa chọn thuộc tính tốt nhất làm điểm để chia nhánh dữ liệu. Các chỉ số ASM thường được sử dụng là Entropy, Information gain, Gini index...

#### ***Entropy***

Entropy là được sử dụng để đánh giá sự phân phối của các lớp trong một nhánh. Entropy sẽ cho giá trị thấp nhất nếu tất cả các mẫu trong đó nhánh đó thuộc vào cùng một lớp (purity), và cho giá trị cao nếu các mẫu trong nhánh đó thuộc nhiều lớp khác nhau (impurity).

Với một phân phối xác suất của một biến rời rạc  $x$  có thể nhận  $n$  giá trị khác nhau  $x_1, x_2, \dots, x_n$ . Giả sử, xác suất để  $x$  nhận các giá trị này là  $p_i = p(x=x_i)$ . Entropy  $H(X)$  của  $x$  được tính theo công thức:

$$H(p) = - \sum_{i=1}^n p_i \log(p_i)$$

### ***Information gain***

Information Gain đo lường sự giảm không chắc chắn (uncertainty) trong biến phụ thuộc sau khi biết giá trị của một thuộc tính cụ thể. Information Gain được tính bằng độ giảm giữa Entropy trước và sau khi phân nhánh. Giá trị Information Gain càng cao, khi dùng thuộc tính đó để phân chia dữ liệu thì các nhánh con sẽ có Entropy thấp hơn so với trước khi phân chia, tức là khả năng các mẫu trong nhánh con đó thuộc cùng một lớp cao hơn.

Để xác định các nút trong mô hình cây quyết định (Decision tree), ta thực hiện tính Information Gain tại mỗi nút theo trình tự sau:

- *Bước 1:* Tính toán hệ số Entropy của biến mục tiêu  $S$  có  $N$  phần tử với  $N_c$  phần tử thuộc lớp  $c$  cho trước:

$$H(S) = - \sum_{c=1}^C \frac{N_c}{N} \log\left(\frac{N_c}{N}\right)$$

- *Bước 2:* Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính  $x$ , các điểm dữ liệu trong  $S$  được chia ra  $K$  child node  $S_1, S_2, \dots, S_K$  với số điểm trong mỗi child node lần lượt là  $m_1, m_2, \dots, m_K$ , ta có:

$$H(x, S) = \sum_{k=1}^K \left( \frac{m_k}{N} \right) H(S_k)$$

- *Bước 3:* Chỉ số Gain Information của thuộc tính  $x$  được tính bằng:

$$G(x, S) = H(S) - H(x, S)$$

### ***Gini index***

Chỉ số Gini, được sử dụng phổ biến nhất, đo lường mức độ tất cả mẫu trong đó nhánh đó thuộc vào cùng một lớp. Trong ngữ cảnh của cây quyết định,

Gini Index đo lường xác suất khi lấy ngẫu nhiên một mẫu từ một lớp và phân loại sai lớp của mẫu đó. Thuộc tính có Gini Index thấp hơn khi phân chia dữ liệu sẽ tạo ra các mẫu thuộc vào cùng một lớp cao hơn.

Công thức tính chỉ số Gini:

$$GI = 1 - \sum_{i=1}^n (p_i)^2$$

với ‘pi’ là xác suất để một đối tượng được phân loại vào một lớp cụ thể.

### **Gain Ratio**

*Gain Ratio (Tỷ lệ lợi ích):* Gain Ratio là một biến thể của Information Gain, được sử dụng để điều chỉnh số lượng giá trị của thuộc tính và kích thước của các nhóm con. Gain Ratio đo lường sự tăng trưởng thông tin của một thuộc tính chia đều cho số lượng lớp khác nhau của biến phụ thuộc. Điều này giúp tránh tình trạng ưu tiên các thuộc tính có nhiều giá trị nhưng mỗi giá trị chỉ chia ra thành các lớp con nhỏ.

Công thức tính Gain Ratio:

$$Gain\ Ratio = \frac{Information\ Gain}{Split\ Info}$$

$$\text{với } Split\ Info = - \sum_{i=1}^n D_i \log_2 D_i$$

Giả sử chúng ta phân chia biến thành n nút con và  $D_i$  đại diện cho số lượng giá trị thuộc nút đó.

#### **1.5.1.4. Cắt tỉa (Pruning)**

Trong các thuật toán xây dựng Decision Tree, có trường hợp nếu cứ liên tục lặp lại việc phân chia các node chưa tinh khiết (chưa nằm cùng một lớp), kết quả sẽ thu được một Decision tree mà mọi điểm trong tập huấn luyện đều được dự đoán đúng. Dẫn tới, Decision tree trở nên quá phức tạp và quá nhớ các điểm dữ liệu trong tập huấn luyện. Điều này dẫn đến dữ liệu trong tập huấn luyện được phân loại chính xác, nhưng mô hình không thể tổng quát hóa được cho các tập dữ liệu khác. Để tránh tình trạng Overfitting, có hai kỹ thuật phổ biến để giảm bớt tình trạng “Overfitting” là pre-pruning và post-pruning.

- *Pre-pruning* là kỹ thuật dừng quá trình xây dựng Decision Tree bằng cách kiểm tra một số điều kiện dừng (stopping conditions) trước khi phân chia, nếu điều kiện dừng được thỏa mãn thì cây sẽ không tiếp tục phân chia node đó và coi là một leaf node.
- *Post – pruning* là kỹ thuật loại bỏ những leaf node không cần thiết sau khi xây dựng xong mô hình Decision Tree đến khi không thể tối ưu hơn thì ngừng việc loại bỏ.

### 1.5.2. Random Forest

Random Forest là một phương pháp thống kê mô hình hóa bằng máy (machine learning statistic) có thể sử dụng cho cả bài toán Classification và Regression. Random Forest được xây dựng bằng cách kết hợp nhiều cây quyết định (decision trees). Mỗi cây quyết định trong mô hình Random Forest được xây dựng bằng cách sử dụng một tập dữ liệu con ngẫu nhiên và một số thuộc tính ngẫu nhiên từ tập dữ liệu ban đầu. Kết quả cuối cùng của Random Forest được tính dựa trên việc lấy trung bình hoặc kết hợp kết quả từ tất cả các cây. Đồng thời kết quả dự báo từ nhiều cây sẽ có phuơng sai nhỏ hơn so với chỉ một cây. Điều này giúp cho mô hình khắc phục được hiện tượng Overfitting.

Nguyên lý hoạt động của Random Forest:

- *Tập hợp Cây Quyết Định:* Rừng Ngẫu Nhiên xây dựng một tập hợp các Cây Quyết Định. Mỗi cây được xây dựng độc lập và có thể sử dụng các phương pháp khác nhau như CART hoặc ID3. Trong quá trình này, phương pháp Bootstrap sẽ cung cấp ngẫu nhiên một phần dữ liệu con khác nhau trong tập huấn luyện để tạo cây. Sự ngẫu nhiên này giúp mỗi cây tập trung vào các thuộc tính khác nhau của dữ liệu, tạo ra sự đa dạng trong Random Forest và giảm nguy cơ overfitting.
- *Độ quan trọng của thuộc tính:* Do mỗi cây quyết định tập trung vào các thuộc tính khác nhau của dữ liệu. Độ quan trọng của mỗi đặc tính được tính toán bằng cách sử dụng các phương pháp như Gini Importance hoặc Mean Decrease Impurity để hiểu được ảnh hưởng của mỗi đặc tính đối với quyết định cuối cùng của mô hình.
- *Kết hợp các dự đoán:* Do mỗi cây quyết định được xây dựng độc lập, việc huấn luyện Random Forest có thể được thực hiện song song trên nhiều nút, giúp tăng

tốc quá trình huấn luyện. Khi đưa ra dự đoán, mỗi cây trong Rừng Ngẫu Nhiên sẽ cho ra một kết quả dự đoán riêng. Kết quả từ tất cả các cây quyết định được kết hợp lại thông qua phương pháp Bagging. Dự đoán cuối cùng sẽ được xác định bằng cách chọn dự đoán thường xuyên nhất trong bài toán Classification hoặc lấy giá trị trung bình trong bài toán Regression. Cơ chế độc lập này đảm bảo sự cân bằng và sự đa dạng trong quyết định cuối cùng.

### 1.5.3. Neural Network

#### 1.5.3.1. Perceptron cơ bản

Một mạng nơ-ron được cấu thành bởi các nơ-ron đơn lẻ được gọi là các perceptron. Nơ-ron nhân tạo được lấy cảm hứng từ nơ-ron sinh học một nơ-ron có thể nhận nhiều đầu vào và cho ra một kết quả duy nhất. Mô hình của perceptron cũng tương tự như vậy. Một perceptron sẽ nhận một hoặc nhiều đầu x vào dạng nhị phân và cho ra một kết quả o dạng nhị phân duy nhất. Các đầu vào được điều phối tầm ảnh hưởng bởi các tham số trọng lượng tương ứng  $w$  của nó, còn kết quả đầu ra được quyết định dựa vào một ngưỡng quyết định b nào đó.

#### 1.5.3.2. Kiến trúc mạng Neural Network

Mạng Neural Network là sự kết hợp của các tầng perceptron hay còn được gọi là perceptron đa tầng (multilayer perceptron):

- Lớp đầu vào đại diện cho các dữ liệu đầu vào
- Lớp ẩn đại diện cho các nút trung gian phân chia không gian đầu vào thành các vùng có ranh giới. Nó nhận vào một tập hợp các đầu vào có trọng số và kết quả đầu ra thông qua một chức năng kích hoạt
- Lớp đầu ra đại diện cho đầu ra của mạng neural

Một Neural Network chỉ có 1 tầng vào và 1 tầng ra nhưng có thể có nhiều tầng ẩn.

### 1.5.4. Support Vector Machine

Máy vector hỗ trợ (Support Vector Machine - SVM) là một mô hình học máy được sử dụng chủ yếu cho việc phân loại dữ liệu, tuy nhiên cũng có thể được áp dụng

cho bài toán hồi quy. Trong thuật toán SVM, chúng ta đặt dữ liệu vào không gian nhiều chiều ( $n$  chiều, trong đó  $n$  là số lượng các tính năng), trong đó mỗi điểm dữ liệu được biểu diễn bởi một vector và mỗi chiều tương ứng với một tính năng. Mục tiêu của SVM là tìm ra một "đường phân chia" (hyperplane) trong không gian nhiều chiều đó để phân tách các điểm dữ liệu thuộc vào các lớp khác nhau. Đường phân chia này được gọi là hyperplane và đơn giản là một đường thẳng (trong không gian hai chiều) hoặc một siêu phẳng (trong không gian ba chiều hoặc cao hơn) có khả năng phân chia dữ liệu thành hai phần riêng biệt, mỗi phần đại diện cho một lớp. Mục tiêu khi tìm kiếm hyperplane là tìm ra đường phân chia mà có khoảng cách lớn nhất giữa nó và các điểm dữ liệu gần nhất của cả hai lớp. Khoảng cách này được gọi là margin, và việc tối đa hóa margin là một trong những nguyên tắc cơ bản của SVM.

### 1.5.5. So sánh các mô hình

Mô hình	Ưu điểm	Nhược điểm
<b>Decision Tree</b>	<ul style="list-style-type: none"> <li>- Dễ hiểu và diễn giải.</li> <li>- Không yêu cầu chuẩn hóa dữ liệu.</li> </ul>	<ul style="list-style-type: none"> <li>- Dễ bị overfitting nếu không được cắt tỉa (prune) hoặc sử dụng kỹ thuật ensemble.</li> <li>- Không phù hợp cho các bài toán có dữ liệu lớn và phức tạp.</li> </ul>
<b>Random Forest</b>	<ul style="list-style-type: none"> <li>- Giảm thiểu overfitting bằng cách kết hợp nhiều cây quyết định.</li> <li>- Hiệu suất tốt trên dữ liệu có nhiều biến và có nhiễu</li> </ul>	<ul style="list-style-type: none"> <li>- Cần nhiều thời gian để huấn luyện với dữ liệu lớn và số lượng cây (estimators) lớn.</li> <li>- Khó diễn giải so với một cây quyết định đơn lẻ.</li> </ul>
<b>Neural Network</b>	<ul style="list-style-type: none"> <li>- Có khả năng học được các mô hình phức tạp và phân loại phi tuyến tính.</li> <li>- Hiệu suất tốt trên các bài toán có dữ liệu lớn và phức</li> </ul>	<ul style="list-style-type: none"> <li>- Đòi hỏi nhiều dữ liệu huấn luyện.</li> <li>- Cần nhiều thời gian và tài nguyên để huấn luyện, đặc biệt là với mạng lớn.</li> </ul>

	tập.	
<b>Support Vector Machine (SVM)</b>	<ul style="list-style-type: none"> <li>- Hiệu suất tốt trên các bài toán phân loại tuyến tính và phi tuyến tính.</li> </ul>	<ul style="list-style-type: none"> <li>- Khó mở rộng cho các bài toán có dữ liệu lớn vì đòi hỏi nhiều tài nguyên tính toán.</li> </ul>

Bảng 1.1 So sánh các mô hình

## 1.6. Đánh giá mô hình

### 1.6.1. Bài toán phân loại

Khi thực hiện bài toán phân loại, có 4 trường hợp của dự đoán có thể xảy ra:

- True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)
- True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)
- False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error
- False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error

Trong thực tế có ba độ đo chủ yếu để đánh giá một mô hình phân loại là Accuracy, Precision and Recall.

#### 1.6.1.1. Độ chính xác (Accuracy)

Độ chính xác là thước đo phổ biến nhất được sử dụng trong cuộc nói chuyện hàng ngày. Độ chính xác trả lời câu hỏi "Trong số tất cả những dự đoán đã được đưa ra, có bao nhiêu dự đoán là đúng?"

$$Accuracy = \frac{correct predictions}{all predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

#### 1.6.1.2. Precision

Precision là thước đo cung cấp cho tỷ lệ số dương thực sự trên tổng số số dương mà mô hình dự đoán. Trả lời câu hỏi “Trong số tất cả những dự đoán

tích cực mà đã đưa ra, có bao nhiêu dự đoán là đúng?". Sử dụng khi quan tâm nhiều đến FP một số ngành như ecommerce website, high quality product. Ví dụ email không phải spam nhưng lại làm spam như vậy sẽ chuyển một số email quan trọng vào mục spam vô tình mất một số email quan trọng. Thể hiện sự chuẩn xác của việc phát hiện các điểm positive. Số này càng cao thì mô hình nhận các điểm Positive càng chuẩn.

$$Precision = \frac{TP}{TP + FP}$$

#### 1.6.1.3. Recall

Tập trung vào mức độ hiệu quả của mô hình trong việc tìm ra tất cả các điểm tích cực. Việc thu hồi còn được gọi là tỷ lệ dương tính thực sự và trả lời câu hỏi "Trong số tất cả các điểm dữ liệu được dự đoán là đúng, có bao nhiêu điểm đã dự đoán đúng là đúng ? Sử dụng khi âm tính giả FN quan trọng hơn tối thiểu hóa nó trong ngành y học - nguy hiểm cho bệnh nhân càng giảm càng tốt  
 $\Rightarrow$  Recall càng lớn càng đúng bỏ sót ít trường hợp.

$$Recall = \frac{TP}{TP + FN}$$

#### 1.6.1.4. F1 Score

F1 Score kết hợp Precision và Recall thành một số liệu duy nhất, cung cấp đánh giá cân bằng về hiệu suất của mô hình. Mặc dù Precision và Recall là quan trọng nhưng chỉ sử dụng chúng có thể không mang lại đánh giá toàn diện. Điểm F1 tính đến cả số liệu và giá trị trung bình hài hòa của chúng, để đạt được sự cân bằng giữa việc giảm thiểu kết quả dương tính giả và âm tính giả. Điểm F1 dao động từ 0 đến 1, trong đó 1 biểu thị độ Precision và Recall hoàn hảo, còn 0 biểu thị hiệu suất kém. Một số ưu điểm so với việc chỉ dựa vào Precision và Recall:

- Cân bằng giữa Precision và Recall: Điểm F1 xem xét cả độ chính xác và thu hồi và coi trọng cả hai chỉ số như nhau. Điều này đảm bảo rằng mô hình không chỉ tối ưu hóa về độ chính xác hoặc thu hồi mà còn tạo ra sự cân bằng giữa việc dự đoán chính xác các trường hợp tích cực và giảm thiểu kết quả dương tính giả và âm tính giả.

- Đối với các tập dữ liệu không cân bằng: Trong các trường hợp mà tập dữ liệu mất cân bằng, một lớp có ít phiên bản hơn đáng kể so với lớp kia, việc sử dụng điểm F1 sẽ giúp giảm thiểu các đánh giá sai lệch. Vì điểm F1 đánh giá cả độ Precision và Recall nên cung cấp đánh giá công bằng về hiệu suất của mô hình trong những trường hợp như vậy.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

### 1.6.2. Bài toán hồi quy

#### 1.6.2.1. MSE

Sai số bình phương trung bình (MSE) đo lường mức độ sai sót trong các mô hình thống kê là số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Đánh giá sự khác biệt bình phương trung bình giữa các giá trị được quan sát và dự đoán. Khi một mô hình không có lỗi, MSE bằng 0. Khi lỗi mô hình tăng lên, giá trị của nó cũng tăng lên. MSE càng thấp thì dự báo càng tốt.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Trong đó:

- $y_i$  là giá trị quan sát thứ i.
- $\hat{y}_i$  giá trị dự đoán tương ứng.
- n là số lượng quan sát.

MSE bình phương các sai số, làm tăng đáng kể ảnh hưởng của các sai số lớn hơn. Điều này giúp mô hình phản ánh chính xác hơn sự chênh lệch giữa dự đoán và giá trị thực tế, đặc biệt là khi các sai số lớn có ý nghĩa đặc biệt. MSE gán trọng số lớn hơn cho các sai số lớn hơn so với MAE. Trong một số trường hợp khi sai số nhỏ không quan trọng và muốn tập trung vào việc giảm thiểu các sai số lớn hơn.

#### 1.6.2.2. MAE

Sai số trung bình tuyệt đối (Mean absolute error). MAE đo lường mức độ trung bình của các lỗi trong một tập hợp các dự đoán, mà không xem xét hướng của chúng. Đó là trung bình trên mẫu thử nghiệm về sự khác biệt tuyệt

đối giữa dự đoán và lượng quan sát thực tế, trong đó tất cả các khác biệt với trọng số bằng nhau.

$$MAE = \frac{\sum |y_i - \hat{y}_i|}{n}$$

*Trong đó:*

- $y_i$  là giá trị quan sát thứ i.
- $\hat{y}_i$  giá trị dự đoán tương ứng.
- n là số lượng quan sát.

Một điểm tối ưu của MAE là không quan tâm đến hướng của sai số. Điều này có nghĩa là nếu mô hình dự đoán một căn nhà với giá trị là \$190,000 trong khi giá thực tế là \$200,000, sai số là  $|200,000 - 190,000| = \$10,000$ . So với MSE, MAE thường được ưu tiên khi có nhiều giá trị ngoại lai trong dữ liệu. MSE bình phương các sai số, khiến cho các giá trị ngoại lai (có sai số lớn hơn) có ảnh hưởng lớn hơn đến kết quả cuối cùng. Trong khi đó, MAE không bị ảnh hưởng nhiều bởi các giá trị ngoại lai, vì nó không bình phương sai số. Điều này làm cho MAE trở nên mạnh mẽ hơn trong một số trường hợp, đặc biệt là khi muốn mô hình đối phó tốt với các giá trị ngoại lai.

### 1.6.2.3. RMSE

Root Mean Squared Error (RMSE) là một biến thể của Mean Squared Error (MSE), được tính bằng cách lấy căn bậc hai của giá trị MSE nhằm biết mức độ sai số trung bình giữa dự đoán và giá trị thực tế trong các đơn vị của biến đang dự đoán. Ví dụ, nếu MSE của mô hình là \$100,000\$, thì RMSE sẽ là căn bậc hai của \$100,000\$, tức là khoảng \$316.23\$. Điều này có nghĩa là sai số trung bình giữa dự đoán và giá trị thực tế là khoảng \$316.23\$ đơn vị của giá nhà.

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

## **1.7. Một số vấn đề về xây dựng mô hình**

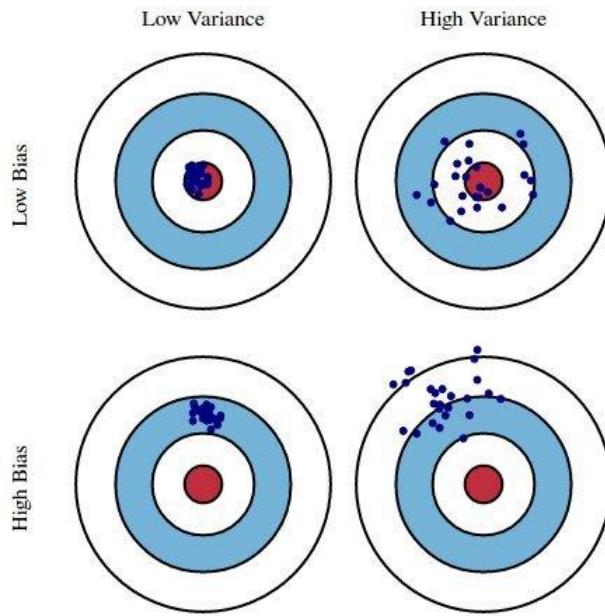
### **1.7.1. Độ lệch và phương sai**

#### **1.7.1.1. Độ lệch (Bias)**

Độ chêch là sai khác giữa giá trị dự báo và giá trị ground truth của một mô hình. Nếu mô hình có bias cao, nghĩa là các dự đoán của mô hình sẽ rất khác biệt so với giá trị thực tế, điều này thường xảy ra khi mô hình quá đơn giản hoặc không đủ phức tạp để biểu diễn đầy đủ các đặc điểm của dữ liệu. giả định rằng mối quan hệ giữa diện tích của căn nhà và giá của nó là tuyến tính, tức là giá nhà tăng một cách đồng đều với diện tích. Nếu chỉ sử dụng mô hình đơn giản này để dự đoán giá nhà, có thể dẫn đến bias cao. Bởi vì mô hình chỉ dựa vào một biến duy nhất (diện tích), không thể hiện đầy đủ các yếu tố khác có thể ảnh hưởng đến giá nhà. Một mô hình có bias cao sẽ không thể hiệu quả trong việc dự đoán dữ liệu mới hoặc dữ liệu mà chưa từng nhìn thấy trước đó, điều này thường được gọi là tình trạng "underfitting".

#### **1.7.1.2. Phương sai (Variance)**

Khái niệm quan trọng trong việc đánh giá sự phân tán của các dự đoán từ mô hình hoặc độ nhạy của mô hình đối với biến động trong dữ liệu. Phương sai cho biết mức độ dự đoán của mô hình sẽ thay đổi ra sao trên các tập dữ liệu khác nhau. Một mô hình có phương sai cao sẽ tạo ra các đưa ra các dự đoán rất khác nhau cho các bộ dữ liệu khác nhau (thậm chí chỉ là một sự thay đổi nhỏ trong dữ liệu). Phương sai cao xảy ra khi mô hình quá phức tạp với nhiều tham số, dẫn đến việc có thể phù hợp với dữ liệu huấn luyện rất tốt. Khi đối mặt với dữ liệu mới mà không có các đặc điểm giống như các dữ liệu cũ mà nó đã "ghi nhớ", mô hình sẽ không thể dự đoán một cách chính xác. Trong trường hợp này, mô hình đang bị Overfitting.



*Hình 1.5 Mối quan hệ giữa độ lệch và phương sai*

### 1.7.2. Overfitting và Underfitting

#### 1.7.2.1. Overfitting

##### ***Khái niệm***

Overfitting là hiện tượng khi mô hình xây dựng thể hiện được chi tiết bộ dữ liệu huấn luyện. Bao gồm dữ liệu nhiều, hoặc dữ liệu bất thường trong tập huấn luyện đều được chọn và học để đưa ra quy luật mô hình. Những quy luật này sẽ không có ý nghĩa nhiều khi áp dụng với bộ dữ liệu mới có thể có dạng dữ liệu nhiều khác. Khi đó, nó ảnh hưởng tiêu cực tới độ chính xác của mô hình nói chung.

Hiện tượng Overfitting thường xảy ra trong các mô hình phi tham số hoặc phi tuyến, những mô hình có sự linh hoạt cao trong xây dựng hàm mục tiêu.

##### ***Nhận biết***

Độ lệch thấp (Low Bias) và phương sai cao (High Variance).

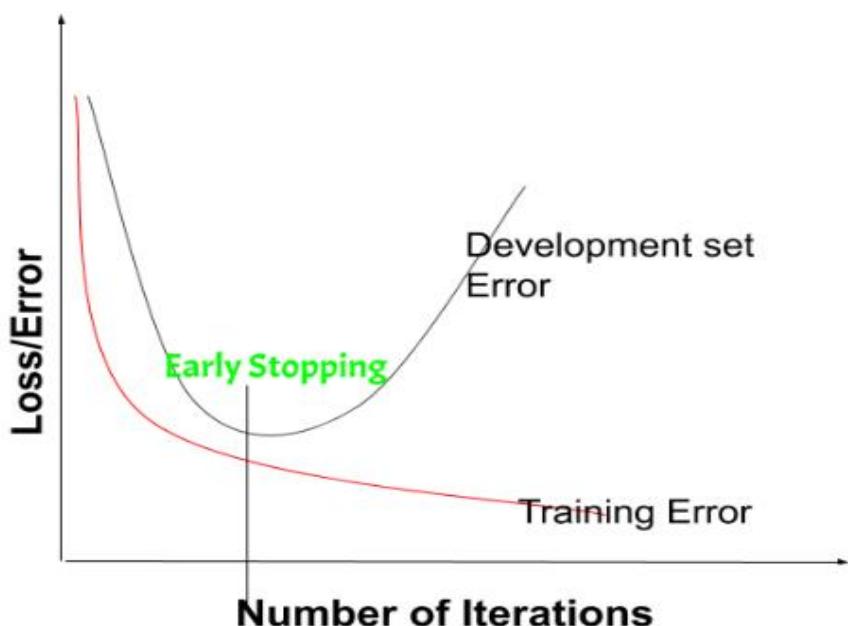
##### ***Lý do dẫn đến Overfitting***

- Kích thước dữ liệu training quá nhỏ và không chứa đủ mẫu dữ liệu
- Dữ liệu training chứa một lượng lớn thông tin bị nhiễu, thông tin không liên quan.

- Mô hình sử dụng quá phức tạp cho một nhiệm vụ đơn giản.

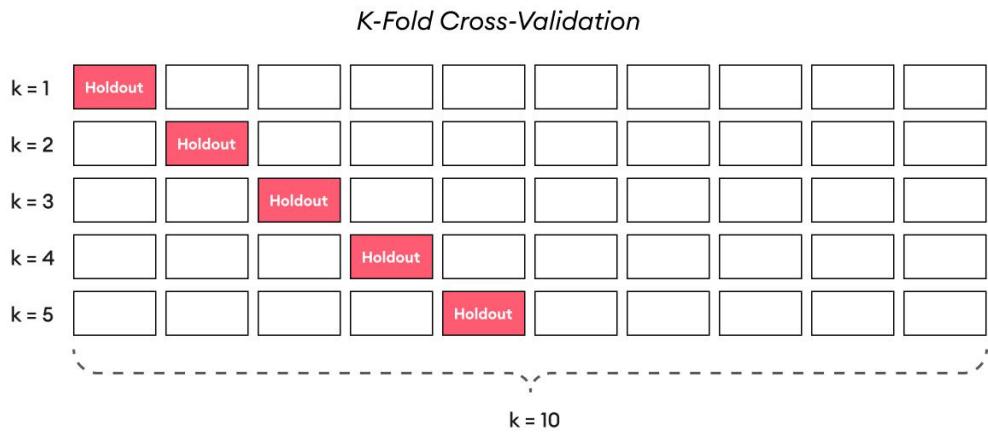
### **Làm thế nào để tránh Overfitting**

- *Cải thiện chất lượng dữ liệu training:* nhằm tập trung vào các mẫu có ý nghĩa, giảm thiểu rủi ro khớp các tính năng nhiễu hoặc không liên quan.
- *Tăng dữ liệu huấn luyện:* điều này giúp cải thiện khả năng khái quát hóa của mô hình đối với dữ liệu chưa nhìn thấy và giảm khả năng khớp quá mức.
- *Giảm độ phức tạp* của mô hình.
- *Dừng sớm (early stopping):* dừng trong giai đoạn training, trước khi mô hình máy học học cả phần nhiễu trong dữ liệu nhằm tăng độ chính xác của mô hình ( thông qua sử dụng hàm mất mát trong giai đoạn training).



*Hình 1.6 Dừng sớm (early stopping)*

- *Dropout dữ liệu:* thông qua đó, một phần các đơn vị trong lớp được loại bỏ ngẫu nhiên trong quá trình huấn luyện.
- *Lấy lại mẫu (resampling methods):* được gọi là k-fold cross validation, bằng cách chia tập dữ liệu thành k tập con. Điều này cho phép training trên các tập dữ liệu khác nhau k lần, và từ đó, xây dựng ước lượng độ chính xác của mô hình học máy với dữ liệu mới.



*Hình 1.7 K-fold cross-validation*

- *Sử dụng Cross-validation:* giúp đánh giá hiệu suất mô hình bằng cách chia dữ liệu thành các tập huấn luyện và kiểm tra. Quá trình này giúp đánh giá khả năng tổng quát hóa của mô hình trên các tập dữ liệu độc lập.

### 1.7.2.1. Underfitting

#### *Khái niệm*

Underfitting (chưa khớp) là hiện tượng mô hình Machine Learning hoặc Deep Learning không học được đủ kiến thức từ dữ liệu huấn luyện và không đạt được hiệu suất tốt trên cả tập huấn luyện và tập kiểm tra.

#### *Nhận biết*

Khi mô hình cho kết quả độ lệch lớn nhưng phương sai nhỏ (Low variance nhưng High bias).

#### *Lý do dẫn đến Underfitting*

- Kích thước dữ liệu training quá nhỏ và không chứa đủ mẫu dữ liệu.
- Dữ liệu training chứa một lượng lớn thông tin bị nhiễu, thông tin không liên quan.
- Mô hình sử dụng quá phức tạp cho một nhiệm vụ đơn giản.

#### *Làm thế nào để tránh Underfitting:*

- *Huấn luyện một mô hình phức tạp hơn:* Sự đơn giản của bộ dữ liệu là một trong những lý do dẫn đến mô hình không phù hợp.

- *Thêm thời gian training:* Việc tăng thêm thời gian sẽ giúp cho mô hình học được nhiều hơn tránh tình trạng underfitting. Do đó, chúng ta có thể tăng số lượng epoch hoặc tăng thời gian training để có kết quả tốt hơn.
- *Loại bỏ nhiễu khỏi dữ liệu:* Bởi sự tồn tại của các giá trị ngoại lệ và giá trị không chính xác trong tập dữ liệu. Kỹ thuật làm sạch dữ liệu có thể giúp giải quyết vấn đề này.
- *Điều chỉnh các tham số chính quy (regularization parameters):* Hệ số chính quy có thể gây ra cả mô hình quá khớp và thiếu khớp.
- *Thử một mô hình khác:* Để nâng cao hiệu quả cần thử một mô hình mới có bản chất phức tạp hơn. Ví dụ: Thay thế mô hình tuyến tính bằng mô hình đa thức bậc cao hơn.

## CHƯƠNG 2: MÔ HÌNH HỒI QUY TUYẾN TÍNH

### 2.1. Mô tả bài toán

#### 2.1.1. Đặt vấn đề

Một công ty bất động sản địa phương muốn hiệu chỉnh quá trình xác định giá nhà đất để tăng cường hiệu suất bán hàng. Hiện tại, công ty này đánh giá giá nhà đất dựa trên các yếu tố như diện tích, vị trí, tiện ích xung quanh và nhiều yếu tố khác. Tuy nhiên, quá trình này không hiệu quả vì không phản ánh chính xác giá trị thực của căn nhà.

Do đó họ muốn xác định giá nhà đất một cách chính xác và hiệu quả hơn để đảm bảo rằng giá đề xuất cho khách hàng là hợp lý và cạnh tranh trên thị trường. Nếu cải thiện được điều này, công ty sẽ tăng khả năng cạnh tranh trên thị trường cũng như thu hút thêm được nhiều khách hàng hơn.

Bộ dữ liệu về giá nhà ở ở Boston là một bộ dữ liệu ghi nhận nhiều thông tin về các yếu tố như diện tích đất, vị trí địa lý, tiện ích xung quanh, thông tin về cơ sở hạ tầng, thuế, môi trường xã hội, và một số yếu tố khác có thể ảnh hưởng đến giá nhà ở tại khu vực Boston, Massachusetts. Công ty có thể căn cứ vào nguồn tài nguyên cực kỳ hữu ích này để phân tích và xây dựng riêng một mô hình đánh giá những yếu tố ảnh hưởng đến thị trường nhà và đưa ra những dự đoán giá nhà gần với thực tế nhất.

#### 2.1.2. Bộ dữ liệu

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 <sup>2</sup> .
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

Bảng 2.1. Mô tả ý nghĩa các biến thuộc bộ dữ liệu

### 2.1.3. Xây dựng câu hỏi nghiên cứu

- Các yếu tố ảnh hưởng đến giá nhà ở (biến Medv) như thế nào?
- Yếu tố nào có ảnh hưởng lớn nhất đến giá nhà?

## 2.2. Tiền xử lý dữ liệu

### 2.2.1. Tổng quan cấu trúc bộ dữ liệu

Dữ liệu bao gồm 14 cột với 506 dòng với các kiểu dữ liệu là float64 (11) và int64 (3)

Index	Column	Non-Null Count	Data Type
1	crim	506	float64
2	zn	506	float64
3	indus	506	float64

4	chas	506	int64
5	nox	506	float64
6	rm	501	float64
7	age	506	float64
8	dis	506	float64
9	rad	506	int64
10	tax	506	int64
11	ptratio	506	float64
12	b	506	float64
13	lstat	506	float64
14	medv	506	float64

Bảng 2.2. Tổng quan về các biến thuộc bộ dữ liệu

Bảng bên dưới mô tả thống kê của các cột trong bộ dữ liệu. Các thông số này cung cấp những thông tin cơ bản về giá trị nhỏ nhất, giá trị lớn nhất, trung vị,... để tiến hành phân tích và đánh giá chất lượng của bộ dữ liệu.

	crim	zn	indus	chas	nox	rm	age
count	506	506	506	506	506	501	506
mean	3.61352	11.36363	11.13677	0.06917	0.55469	6.28434	68.574901
std	8.33548	7.92525	7.10551	0.08520	0.09520	1.39534	12.50140
min	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
25%	0.38500	1.78500	2.31500	0.02620	0.45500	4.98500	45.00000
50%	0.63245	2.36000	3.21000	0.04630	0.50000	5.42500	54.93750
75%	1.00000	3.67500	4.98000	0.08600	0.65000	6.98000	66.18750
max	12.87500	39.69000	48.87500	8.00000	1.00000	10.00000	99.90000

<b>std</b>	8.60154 5	23.32245 3	6.860353	0.25399 4	0.11587 8	0.70558 7	28.148861
<b>min</b>	0.00632	0	0.46	0	0.385	3.561	2.9
<b>25%</b>	0.08204 5	0	5.19	0	0.449	5.884	45.025
<b>50%</b>	0.25651	0	9.69	0	0.538	6.208	77.5
<b>75%</b>	3.67708 3	12.5	18.1	0	0.624	6.625	94.075
<b>max</b>	88.9762	100	27.74	1	0.871	8.78	100

Bảng 2.3 Mô tả thống kê lần 1 #1

	dis	rad	tax	ptratio	b	lstat	medv
<b>count</b>	506	506	506	506	506	506	506
<b>mean</b>	3.79504 3	9.54940 7	408.23715 4	18.45553 4	356.6740 32	12.653 063	22.53280 6
<b>std</b>	2.10571	8.70725 9	168.53711 6	2.164946	91.29486 4	7.1410 62	9.197104
<b>min</b>	1.1296	1	187	12.6	0.32	1.73	5
<b>25%</b>	2.10017 5	4	279	17.4	375.3775	6.95	17.025
<b>50%</b>	3.20745	5	330	19.05	391.44	11.36	21.2

<b>75%</b>	5.18842 5	24	666	20.2	396.225	16.955	25
<b>max</b>	12.1265	24	711	22	396.9	37.97	50

Bảng 2.4 Mô tả thống kê lần 1 #2

## 2.2.2. Làm sạch dữ liệu

### 2.2.2.1. Xử lý dữ liệu trùng lặp và không hợp lệ

Dữ liệu trùng lặp và không hợp lệ là giá trị không phù hợp với ngữ cảnh hoặc không thể tham khảo. Kết quả sau khi kiểm tra bằng Python đã cho ra bộ dữ liệu này không có dòng nào bị trùng lặp với nhau, đảm bảo các dòng là duy nhất.

### 2.2.2.2. Xử lý dữ liệu thiếu (NaN)

Dữ liệu thiếu (NaN) là những dữ liệu sẽ làm sai lệch kết quả phân tích và tạo ra thông tin thừa nên cần được kiểm tra và xử lý loại bỏ phù hợp. Tổng số lượng các giá trị thiếu (NaN) và phần trăm của mỗi cột của bộ dữ liệu được thể hiện lần lượt qua các cột “Total”, “Percent” ở bảng bên dưới.

Column	Total	Percent
rm	5	0.99
crim	0	0
zn	0	0
indus	0	0
chas	0	0
nox	0	0

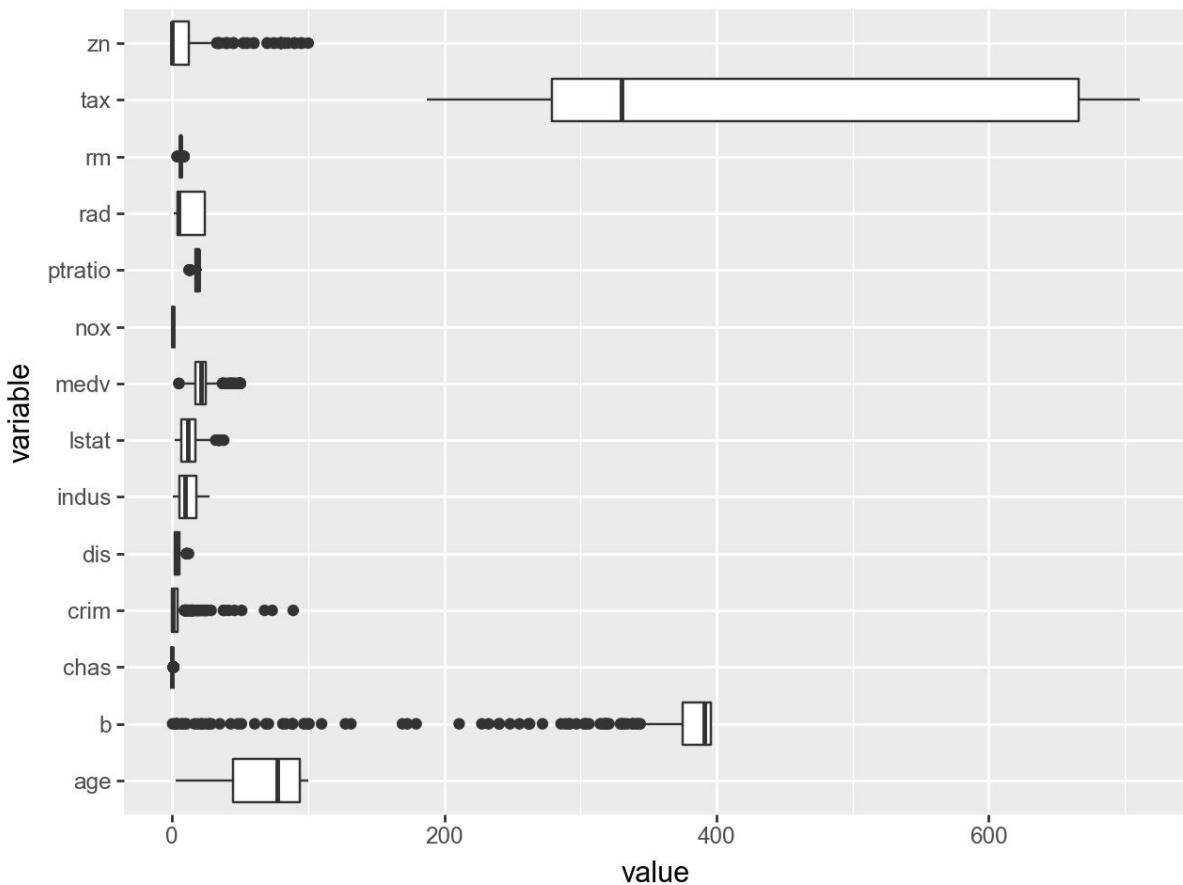
age	0	0
dis	0	0
rad	0	0
tax	0	0
ptratio	0	0
b	0	0
lstat	0	0
medv	0	0

Bảng 2.5 Mô tả thống kê các giá trị thiếu

Nhìn vào bảng thống kê, có thể thấy bộ dữ liệu này chỉ có 1 cột có dữ liệu thiếu (NaN) là cột “rm” với 5 dữ liệu trống, chiếm 0.99% trên tổng bộ 506 dữ liệu của cột này. Do số lượng không đáng kể nên các dữ liệu thiếu trên sẽ được thay thế bằng trung vị nhằm tránh ảnh hưởng tới phân phối ban đầu của dữ liệu.

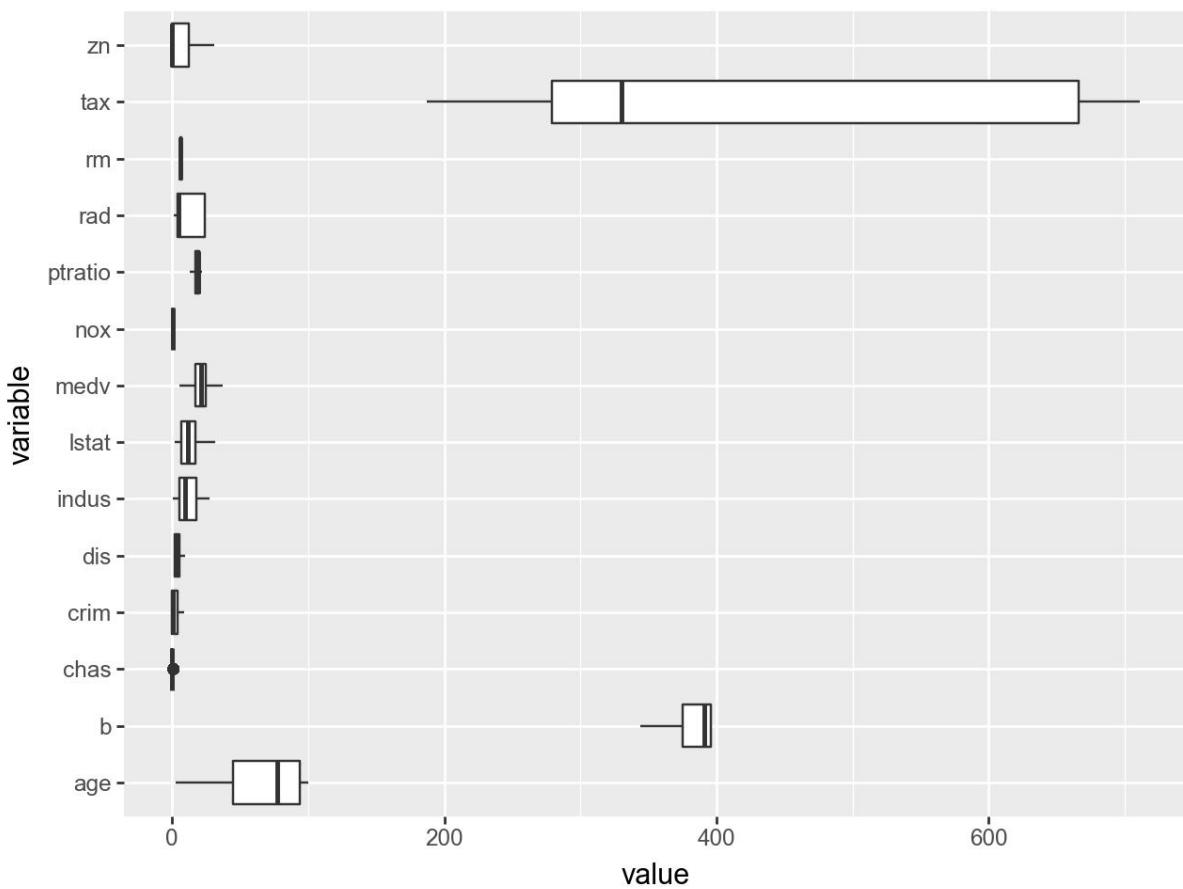
#### 2.2.2.3. Xử lý các ngoại lệ

Ở bước này, các cột có dữ liệu số sẽ được kiểm tra ngoại lệ thông qua biểu đồ Boxplot. Biểu đồ Boxplot là một biểu đồ rất có ích để thể hiện các chỉ số thống kê của các biến liên tục, biểu đồ này sẽ hiển thị các dữ liệu ngoại lệ thông qua các chấm tròn.



Hình 2.1 Biểu đồ Boxplot mô tả dữ liệu trong dataset trước khi clean

Dựa vào biểu đồ, có thể thấy có 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 với tiêu biểu là 3 cột “*zn, crim, b*” có nhiều giá trị ngoại lai nhất. Các giá trị ngoại lai có thể làm thay đổi mức độ tương quan giữa các biến, ảnh hưởng lớn đến việc đánh giá mô hình. Do đó, nhóm sẽ áp dụng phương pháp tính *Độ trai 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. Phương pháp này được sử dụng do nó sẽ đảm bảo tính toàn vẹn của dữ liệu và mức độ tương quan của các biến sẽ ít bị ảnh hưởng nhất.



Hình 2.7. Biểu đồ Boxplot mô tả các dòng dữ liệu trong dataset sau khi clean

Bởi vì những giá trị ngoại lệ bị loại bỏ nên phân phối của bộ dữ liệu sẽ có những thay đổi về giá trị, bảng bên dưới sẽ tổng quát lại cấu trúc dữ liệu sau khi thay đổi.

	<b>crim</b>	<b>zn</b>	<b>indus</b>	<b>chas</b>	<b>nox</b>	<b>rm</b>	<b>age</b>
<b>count</b>	506	506	506	506	506	506	506
<b>mean</b>	2.225032	6.963439	11.136779	0.06917	0.554695	6.273861	68.574901
<b>std</b>	3.313353	12.028788	6.860353	0.253994	0.115878	0.628218	28.148861
<b>min</b>	0.00632	0	0.46	0	0.385	4.785625	2.9

<b>25%</b>	0.082045	0	5.19	0	0.449	5.8855	45.025
<b>50%</b>	0.25651	0	9.69	0	0.538	6.208	77.5
<b>75%</b>	3.677083	12.5	18.1	0	0.624	6.61875	94.075
<b>max</b>	9.069639	31.25	27.74	1	0.871	7.718625	100

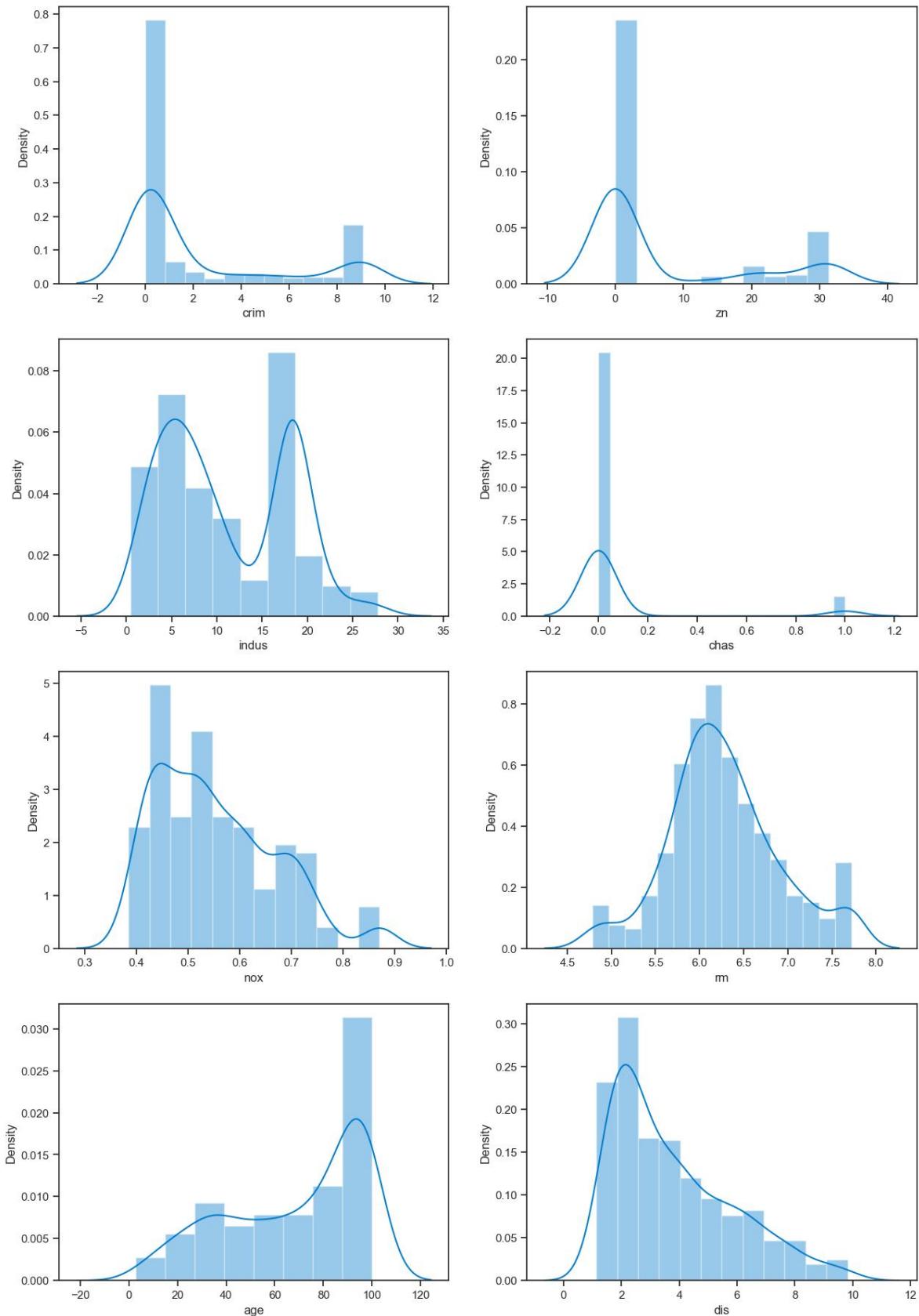
Bảng 2.6 Mô tả thống kê lần 2 #1

	dis	rad	tax	ptratio	b	lstat	medv
<b>count</b>	506	506	506	506	506	506	506
<b>mean</b>	3.783947	9.549407	408.237154	18.463834	381.918836	12.612011	21.877075
<b>std</b>	2.069765	8.707259	168.537116	2.143924	19.054913	7.016829	7.602976
<b>min</b>	1.1296	1	187	13.2	344.10625	1.73	5.0625
<b>25%</b>	2.100175	4	279	17.4	375.3775	6.95	17.025
<b>50%</b>	3.20745	5	330	19.05	391.44	11.36	21.2
<b>75%</b>	5.188425	24	666	20.2	396.225	16.955	25
<b>max</b>	9.8208	24	711	22	396.9	31.9625	36.9625

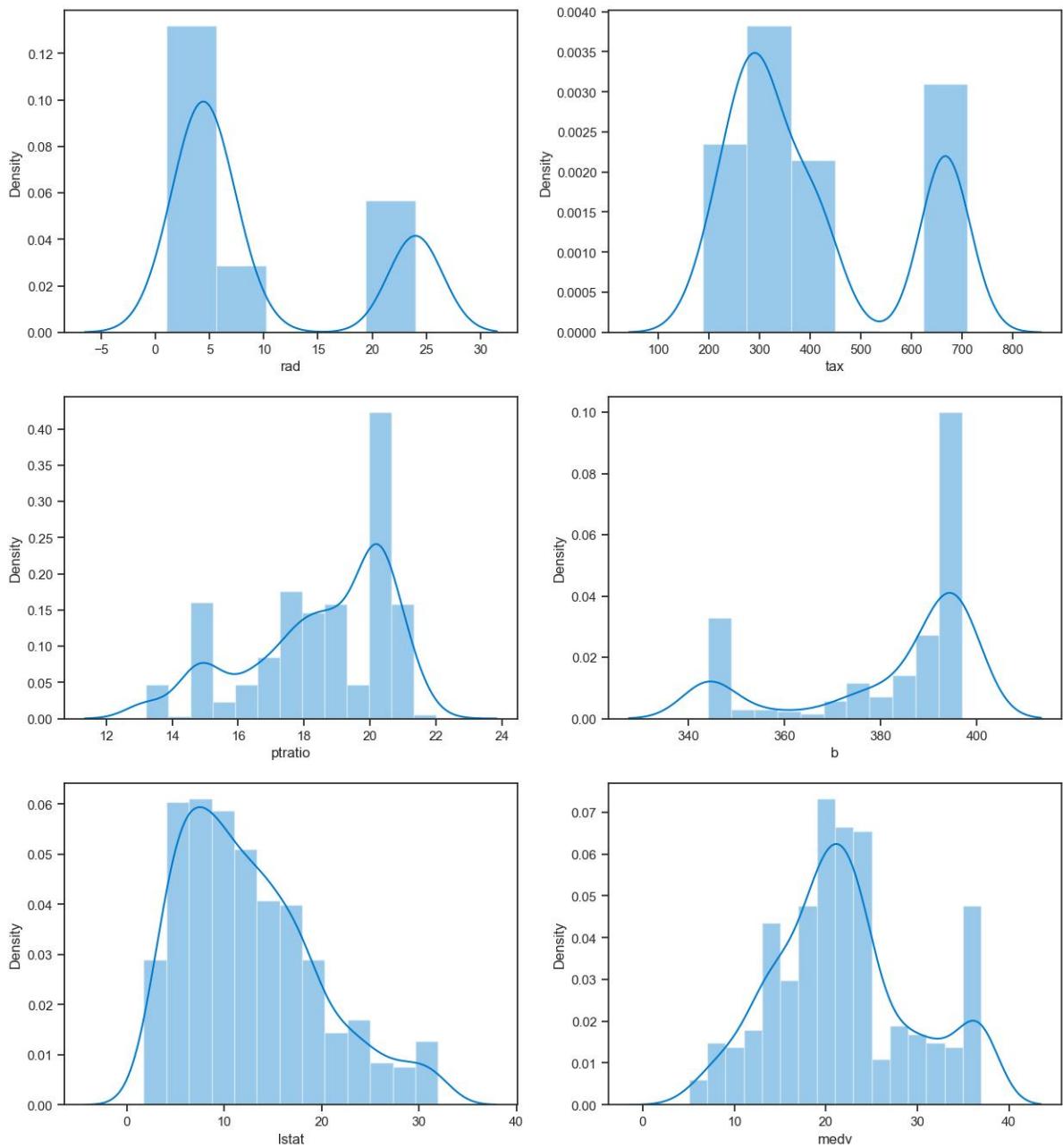
Bảng 2.7 Mô tả thống kê lần 2 #1

## 2.2.3. Phân tích khám phá dữ liệu (EDA)

### 2.2.3.1. Phân tích phân phối dữ liệu



Hình 2.2 Biểu đồ phân phối dữ liệu #1



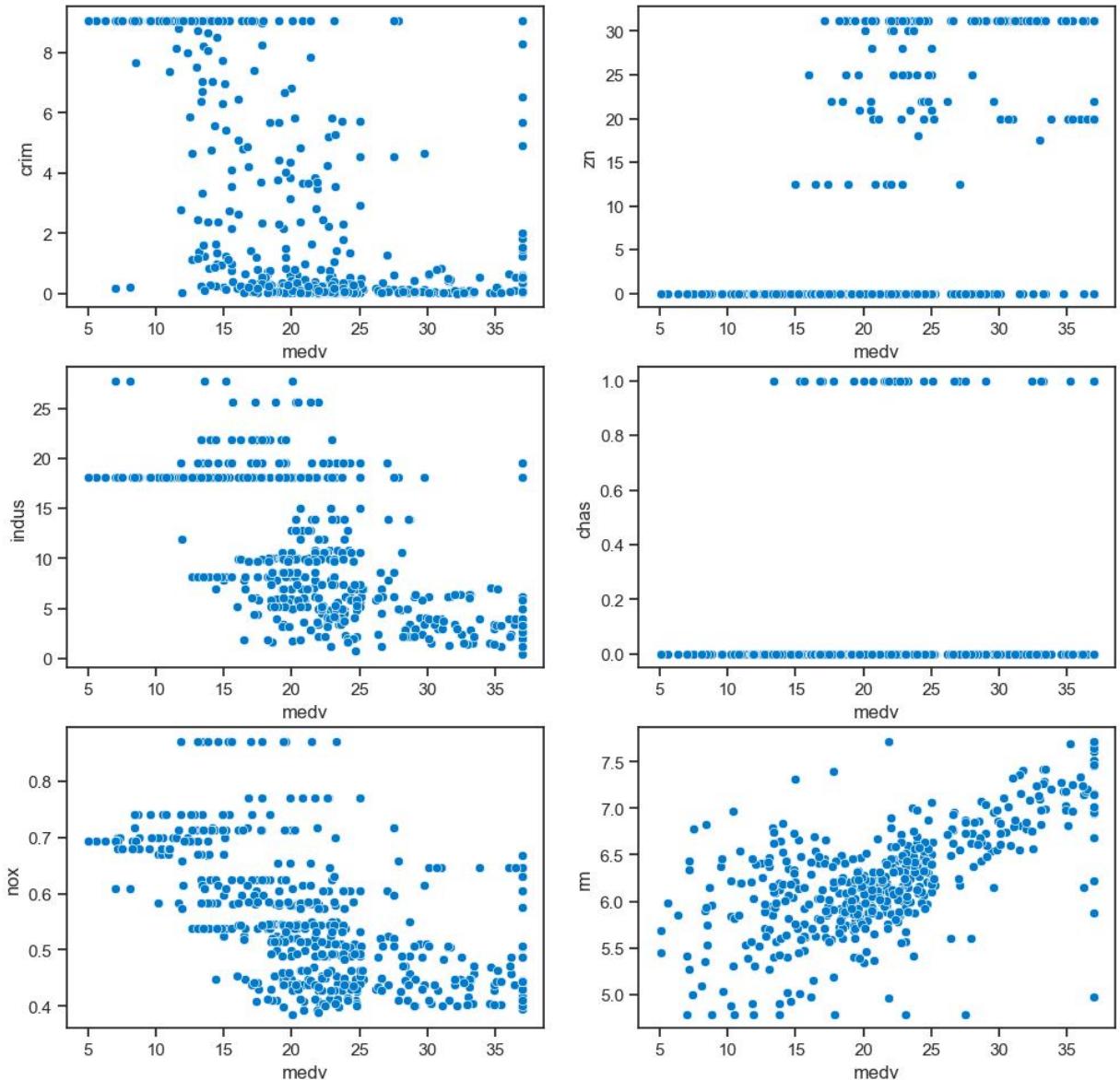
Hình 2.3 Biểu đồ phân phối dữ liệu #2

Dựa vào các biểu đồ, ta có thể nhận xét được về phân phối của các cột như sau:

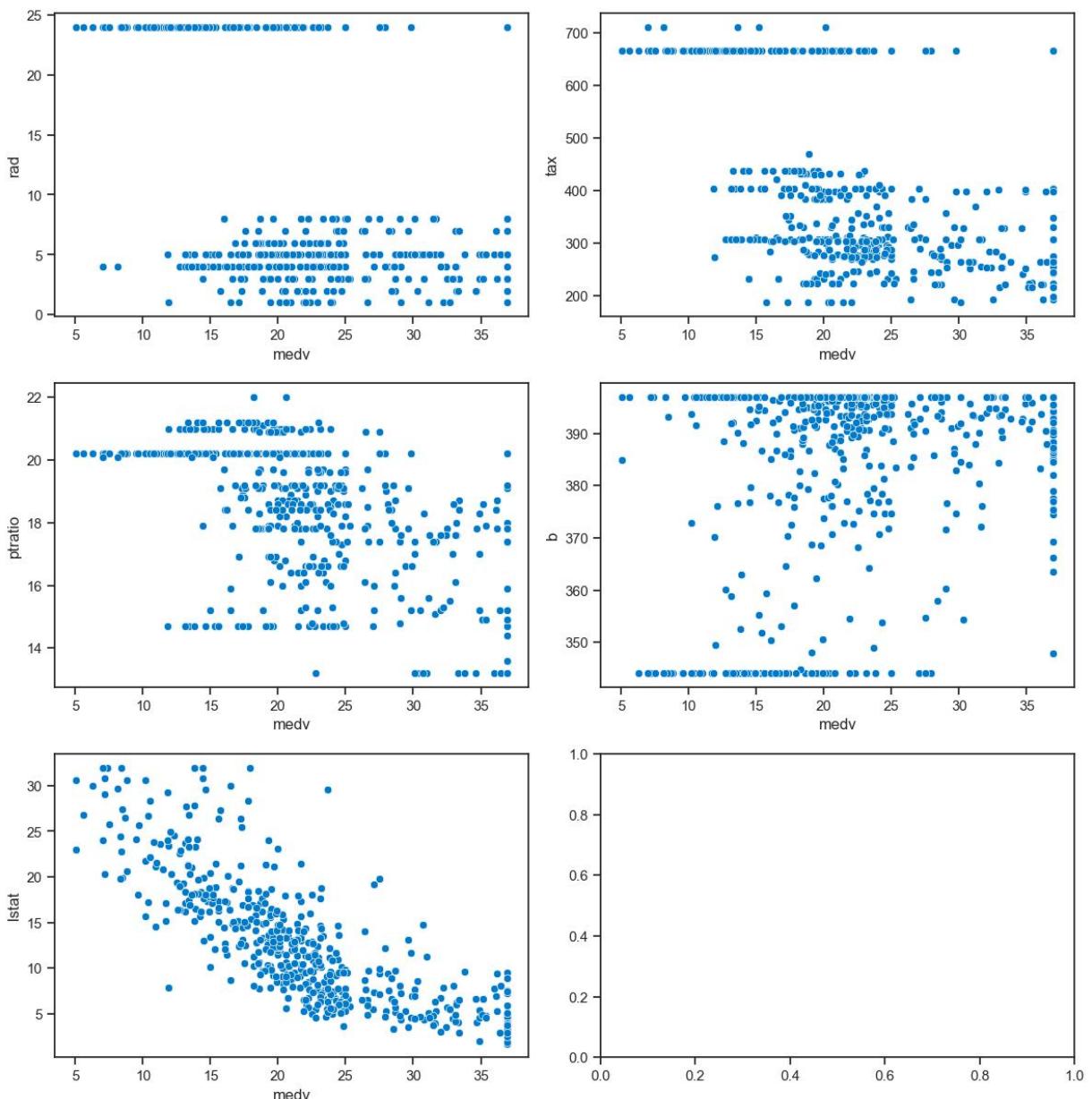
- Một số biến có phân phối đồ thị lệch phải như: `crim`, `nox`, `dis`, `lstat`.
- Một số biến có phân phối đồ thị lệch trái như: `age`, `b`, `pratio`.
- Một số biến có phân phối đồ thị tương đối đối xứng như: `indus`, `rm`, `medv`.
- Ngoài ra, 1 số biến có đồ thị phân phối gồm 2 phần như:
  - + `zn`: Đa số giá trị là 0% và các giá trị còn lại trải dài từ ~10 đến ~31%
  - + `chas`: Gồm 2 phần là 0: “Không gần sông” và 1: “Gần sông” và đa số nhà được bán không gần sông.

- + rad: Có 2 khu vực mà các căn nhà đang được bán tọa lạc là nhà ở khu vực nông thôn và nhà ở khu vực gần thành thị.
- + tax: Tương tự gồm 2 phần là giá thuê tài sản thấp có thể là đến từ các căn nhà ở khu vực nông thôn và giá thuê tài sản cao đến từ các căn nhà ở khu vực gần thành thị.

### 2.2.3.2. Phân tích tương quan bằng biểu đồ phân tán



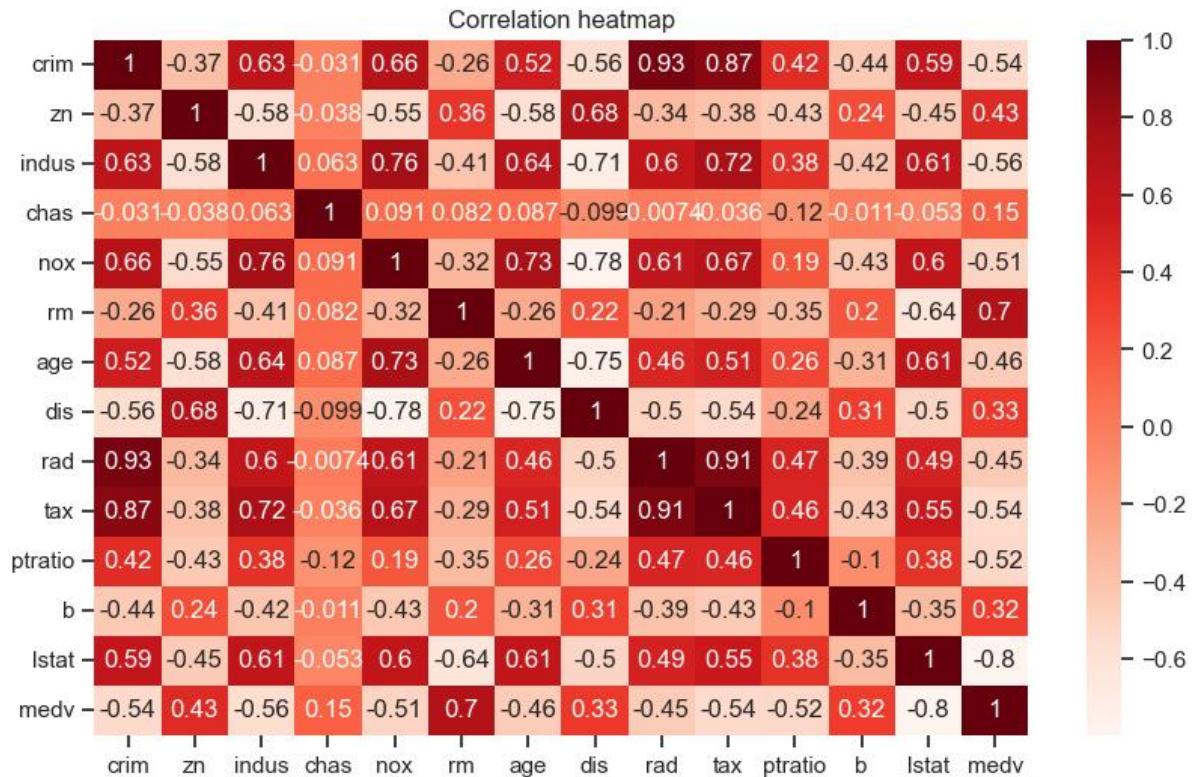
Hình 2.4 Biểu đồ phân tán dữ liệu #1



*Hình 2.5 Biểu đồ phân tán dữ liệu #2*

Qua biểu đồ phân tán cho thấy giữa biến “Medv” với “rm” và “lstat” có mối quan hệ tuyến tính. Điều này cho thấy khi số phòng trung bình/căn hộ hoặc tỷ lệ dân số có địa vị thấp thay đổi thì giá nhà có xu hướng tăng hoặc giảm theo. Các biến còn lại sẽ được xem xét thêm bằng biểu đồ nhiệt (Heat map) để tìm hiểu rõ hơn về mối tương quan giữa các biến này với biến “Medv”

### 2.2.3.3. Phân tích tương quan bằng biểu đồ nhiệt



Hình 2.6 Biểu đồ Heatmap thể hiện tương quan của dữ liệu

Cột	Tương quan
lstat	-0.7972
indus	-0.561174
tax	-0.543545
crim	-0.538589
ptratio	-0.523993
nox	-0.506505
age	-0.458662

rad	-0.452679
chas	0.146061
b	0.32125
dis	0.333079
zn	0.428126
rm	0.697787
medv	1

Bảng 2.8 Thông kê tương quan của các biến với “Medv”

Kết quả từ biểu đồ nhiệt và bảng số liệu cụ thể, ta có thể thấy rằng hầu hết các biến độc lập có mức độ tương quan tương đối đến biến phụ thuộc là *medv*. Trong đó có 8 biến độc lập có tương quan nghịch và 5 biến có tương quan thuận với biến phụ thuộc. Đặc biệt là hai biến “rm” và “lstat” được đánh giá có tương quan cao bằng biểu đồ phân tán cũng nhận lần lượt là biến có tương quan thuận cao nhất (*rm* đạt 0.7) và biến có tương quan nghịch cao nhất (*lstat* đạt -0.8). Điều này có thể đưa ra một vài nhận định về các yếu tố ảnh hưởng tới giá nhà như sau:

- Số phòng ở và tỷ lệ dân số có địa vị thấp trong khu vực là hai yếu tố ảnh hưởng đáng kể tới giá nhà.
- Mối quan hệ giữa số lượng phòng và giá nhà là tỷ lệ thuận, số lượng phòng trong nhà tăng thì giá nhà cũng tăng theo. Điều này được lý giải vì người mua nhà thường có nhu cầu tìm kiếm những ngôi nhà có nhiều diện tích sử dụng và cung cấp được nhiều tiện ích cho gia đình, do đó một ngôi nhà với nhiều phòng ở hơn sẽ có sức hấp với người mua nhiều hơn và từ đó làm tăng giá trị của ngôi nhà lên đáng kể.

- Mỗi quan hệ giữa tỷ lệ dân số có địa vị thấp và giá nhà là tỷ lệ nghịch, khi tỷ lệ dân số có địa vị thấp trong khu vực tăng thì giá nhà trong khu vực sẽ giảm. Điều này lý giải là khi tỷ lệ dân số có địa vị thấp trong khu vực cao, thì các chủ sở hữu bất động sản sẽ không quyết định xây dựng các dự án nhà ở đắt tiền của họ tại đây vì phần lớn người dân sẽ không thể chi trả được cũng như môi trường sống thiếu nhiều tiện nghi cần thiết, khó thu hút khách hàng. Vì vậy, trung bình giá của các căn nhà trong khu vực đó sẽ rẻ hơn.
- Mối tương quan âm giữa tỷ lệ tội phạm trong khu vực và nồng độ oxit nitơ với giá nhà có thể chỉ ra rằng những khu vực có môi trường sống không an toàn và ô nhiễm thường có giá nhà thấp hơn. Điều này có thể làm giảm sự hấp dẫn của khu vực đó và ảnh hưởng đến giá nhà.

## 2.3. Xây dựng mô hình hồi quy tuyến tính đơn biến

### 2.3.1. Xây dựng mô hình

Mô hình hồi quy tuyến tính đơn biến là mô hình chỉ có một biến độc lập. Mô hình này có thể dễ dàng xác định hiệu ứng của mỗi biến độc lập lên biến phụ thuộc mà không cần lo lắng về sự phức tạp của tương tác giữa các biến.

Để xây dựng và đánh giá về mô hình, tập dữ liệu sẽ được chia thành 80% để tạo thành tập huấn luyện, 20% để tạo thành tập kiểm tra. Với biến phụ thuộc: *medv* và biến độc lập được chọn là biến có mức độ tương quan cao nhất với biến phụ thuộc là: *lstat*.

Mô hình được xây dựng có phương trình:

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

**Nhận xét:** Trong điều kiện các yếu tố khác không đổi, khi tỷ lệ dân số có địa vị thấp trong khu vực tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà sẽ giảm 860\$. Với mô hình này, biến độc lập giải thích được 64% sự biến thiên của biến phụ thuộc.

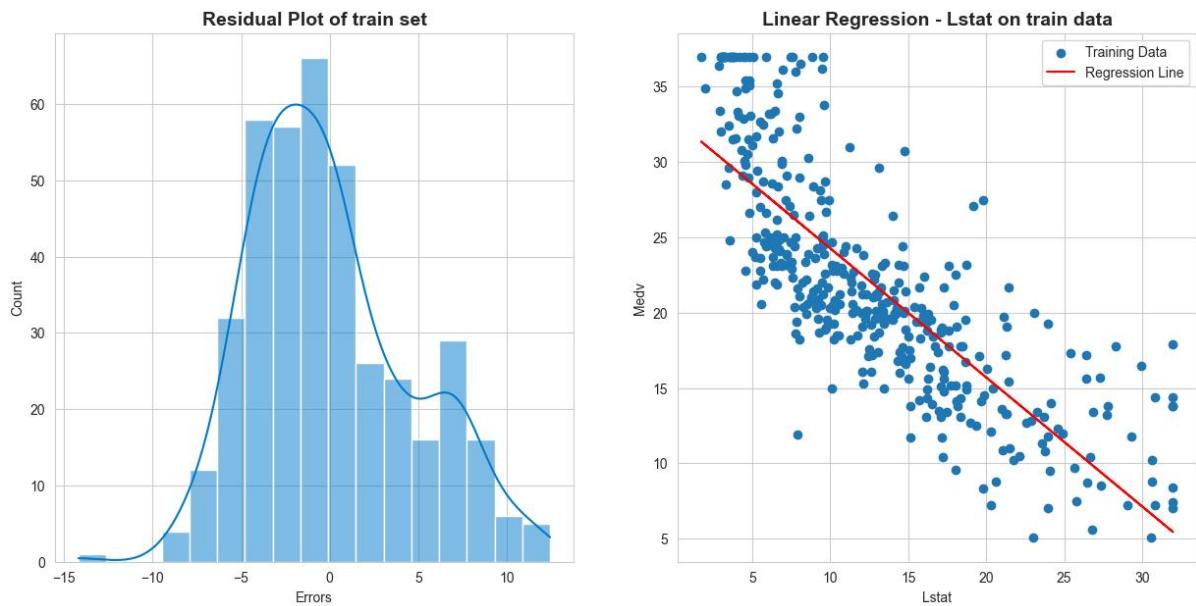
⇒ Như vậy, biến “Lstat” là ảnh hưởng đáng kể đến biến phụ thuộc. Các khu vực dân số có địa vị thấp với tỷ lệ cao có thể có các yếu tố xã hội hoặc kinh tế không tốt hơn, dẫn đến giá nhà thấp hơn.

### 2.3.2. Đánh giá mô hình

Mô hình hồi quy tuyến tính đơn biến “Lstat” sẽ được kiểm nghiệm thêm trên tập huấn luyện và tập thử nghiệm để có nhiều cơ sở đánh giá mô hình.

#### 2.3.2.1. Đánh giá mô hình trên tập huấn luyện

Với tập 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, biểu đồ phân phối của các sai số giữa giá trị thực tế và giá trị dự đoán và biểu đồ phân tán giữa giá trị Lstat và Medv.



Hình 2.7 Biểu đồ đánh giá mô hình hồi quy đơn biến trên tập huấn luyện

#### Nhận xét

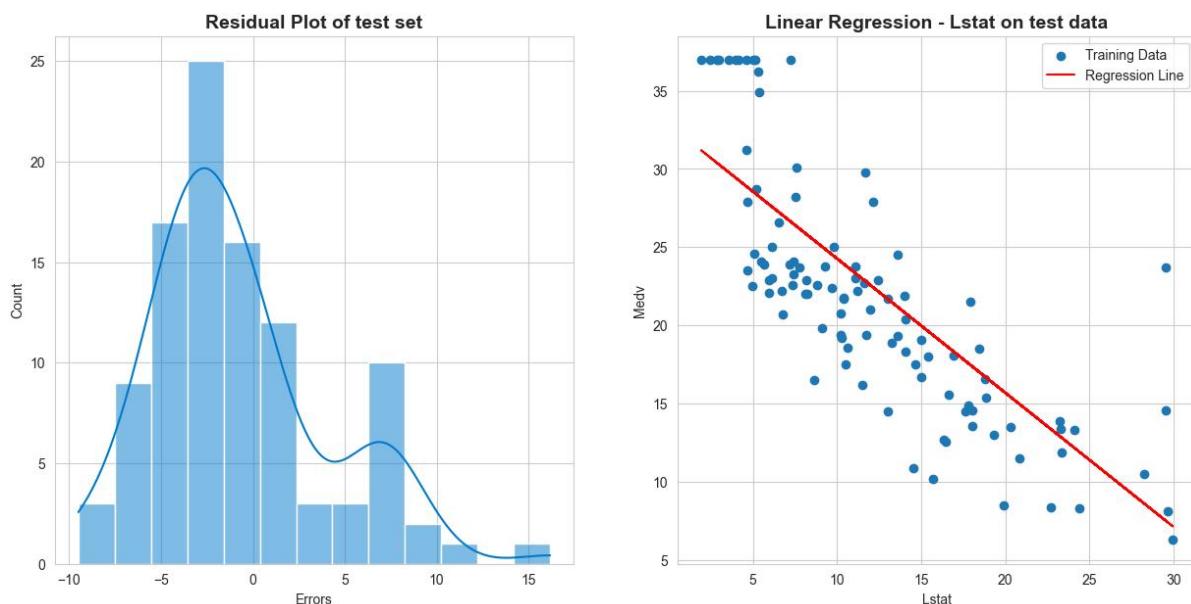
- Trung bình sai số của giữa giá trị dự đoán và giá trị thực tế MSE là 20,5. MSE tăng trưởng nhanh hơn khi có các sai số lớn.
- Trung bình giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực tế MAE là 3,6. MAE không bị ảnh hưởng bởi các sai số lớn.
- Phân phối các sai số của mô hình trải dài từ [-15: 10] trong đó tập trung chủ yếu ở khoảng [-5:5]. Như vậy sai số của mô hình tập trung ở mức độ tương đối nhỏ và giá trị sai số lớn chiếm một phần không quá lớn.
- Biểu đồ phân tán cho thấy các điểm dữ liệu đa phần có xu hướng được dự đoán tốt vì nằm cùng hướng với đường hồi quy. Tuy nhiên cũng có những điểm dữ liệu có xu hướng khác với xu hướng chung.

## Kết luận

- Mô hình có xu hướng dự đoán tốt ở một số điểm dữ liệu vì sai số rất ít và ngược lại có một số điểm dữ liệu dự đoán có sai lệch cao so với thực tế, điều này có thể do ảnh hưởng của các giá trị ngoại lai hoặc các điểm dữ liệu đặc biệt.
- Mô hình đưa ra dự đoán dựa trên tập huấn luyện khá tốt khi giá nhà dao động từ 5-36 (đơn vị \$1000, dựa trên thống kê giá trị medv) và dự đoán được mô hình đưa ra phần lớn chỉ chênh lệch khoảng 5 so với giá trị thực.

### 2.3.2.2. Đánh giá mô hình trên tập thử nghiệm

Với tập thử nghiệm, mô hình ghi nhận  $R^2 = 0.61$ , MSE = 23.2 và MAE = 3.9 và biểu đồ phân phối của các sai số giữa giá trị thực tế và giá trị dự đoán và biểu đồ phân tán giữa giá trị Lstat và Medv.



Hình 2.8 Biểu đồ đánh giá mô hình hồi quy đơn biến trên tập thử nghiệm

## Nhận xét

- Với mô hình này, biến độc lập giải thích được 61% sự biến thiên của biến phụ thuộc → giảm 3% so với tập dữ liệu huấn luyện.
- Trung bình sai số của giữa giá trị dự đoán và giá trị thực tế MSE là 23.2 → tăng 2,7 so với tập dữ liệu huấn luyện.

- Trung bình giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực té MAE là 3,9 → tăng 0,3 so với tập dữ liệu huấn luyện.
- Phân phối sai số của mô hình trải dài từ [-10: 15] trong đó tập trung chủ yếu ở khoảng [-5: 0]. Như vậy sai số của mô hình tập trung ở mức độ tương đối nhỏ và giá trị sai số lớn chiếm một phần không quá lớn.
- Biểu đồ phân tán cho thấy các điểm dữ liệu đa phần có xu hướng được dự đoán tốt vì nằm cùng hướng đường hồi quy. Tuy nhiên cũng có những điểm giá trị có xu hướng khác với xu hướng chung.

### **Kết luận**

- 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.
- Tuy tự trên tập huấn luyện, mô hình có xu hướng dự đoán tốt ở một số điểm dữ liệu vì sai số rất ít và ngược lại có một số điểm dữ liệu dự đoán có sai lệch cao so với thực tế, điều này có thể do ảnh hưởng của các giá trị ngoại lai hoặc các điểm dữ liệu đặc biệt.
- Mô hình đưa ra dự đoán dựa trên tập mới khá tốt khi giá nhà dao động từ 5-36 (đơn vị \$1000, dựa trên thống kê giá trị medv) và dự đoán được mô hình đưa ra phần lớn chỉ chênh lệch khoảng 5 so với giá trị thực.

#### **2.3.2.3. Kết luận và dự đoán**

Như vậy, mô hình dự đoán giá nhà được xây dựng dựa trên biến “lstat” có hiệu suất khá tốt trên tập huấn luyện và thử nghiệm. Thực hiện dự đoán với giá trị lstat = 10:

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

## 2.4. Xây dựng mô hình hồi quy tuyến tính đa biến

### 2.4.1. Xác định biến độc lập

Để thực hiện xác định các biến độc lập nào có giá trị khảo sát và tác động đến giá nhả, nhóm thực hiện chạy mô hình hồi quy tuyến tính đa bội đối với 13 biến độc lập và biến phụ thuộc là Medv.

Quá trình lựa chọn biến độc lập sẽ được xem xét lần lượt các giá trị của p-value và VIF (Variance inflation factor). Đối với mô hình hồi quy tuyến tính đa bội, để xác định một biến độc lập có tác động đến biến phụ thuộc hay không, sẽ nhìn vào giá trị p-value của biến đó. Nếu biến A có p-value  $\leq 0.05$ , có thể bác bỏ giả thiết “Biến độc lập A không tác động đến biến phụ thuộc”, từ đó khẳng định rằng biến độc lập A có ảnh hưởng đến biến phụ thuộc là biến Medv. Bên cạnh đó mô hình hồi quy tuyến tính đa biến còn có khả năng gặp vấn đề đa cộng tuyến giữa các biến độc lập, điều này có thể làm giảm độ chính xác của các kết luận từ mô hình phân tích. Vậy nên, giá trị VIF được xem xét nhằm xác định mức độ tương quan giữa các biến độc lập. Các biến được coi là không có đa cộng tuyến nếu giá trị VIF  $< 5$ . Do đó, các biến độc lập phù hợp phải thỏa mãn điều kiện là p-value  $\leq 0.05$  và VIF  $< 5$ .

#### ***Quy trình của quá trình xác định biến độc lập sẽ diễn ra theo trình tự***

- Bước 1 (Kiểm tra p-value): Chạy mô hình và kiểm tra giá trị p-values, sau đó loại biến có p-value cao nhất trong tất cả các biến có p-value  $> 0.05$  và lặp lại bước 1 tới khi tất cả các biến có p-value  $\leq 0.05$ .
- Bước 2 (Kiểm tra VIF): Tính giá trị VIF của các biến đã xác định được và tiến hành kiểm tra, nếu có VIF  $> 5$  thì sẽ loại biến có VIF cao nhất trong tất cả các biến có VIF  $> 5$  và lặp lại bước 2 tới khi tất cả các biến đều có VIF  $< 5$ .
- Bước 3: Nếu ở bước 2 có tiến hành loại bỏ một biến thì sẽ thực hiện lại bước 1, ngược lại sẽ kết thúc quá trình xác định biến độc lập.

#### ***Kết quả thực hiện quá trình xác định biến độc lập***

- Kiểm tra p-value lần 1, loại bỏ “indus” có p-value = 0.7, cao nhất trong tất cả các biến có p-value  $> 0.05$ .

- Kiểm tra p-value lần 2, loại bỏ biến “b” có p-value = 0.454, cao nhất trong tất cả các biến có p-value > 0.05.
- Kiểm tra p-value lần 3, loại bỏ biến “age” có p-value = 0.395, cao nhất trong tất cả các biến có p-value > 0.05.
- Kiểm tra p-value lần 4, xác nhận tất cả các biến đều có p-value  $\leq 0.05$

OLS Regression Results						
Dep. Variable:	medv	R-squared:	0.787			
Model:	OLS	Adj. R-squared:	0.783			
Method:	Least Squares	F-statistic:	183.1			
Date:	Wed, 17 Apr 2024	Prob (F-statistic):	2.93e-159			
Time:	01:46:55	Log-Likelihood:	-1352.4			
No. Observations:	506	AIC:	2727.			
Df Residuals:	495	BIC:	2773.			
Df Model:	10					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	38.0127	3.937	9.654	0.000	30.277	45.749
crim	-0.6231	0.145	-4.296	0.000	-0.908	-0.338
zn	0.0439	0.020	2.172	0.030	0.004	0.084
chas	1.6459	0.640	2.573	0.010	0.389	2.903
nox	-14.1920	2.673	-5.309	0.000	-19.444	-8.940
rm	2.9390	0.349	8.420	0.000	2.253	3.625
dis	-1.0470	0.142	-7.353	0.000	-1.327	-0.767
rad	0.3440	0.063	5.473	0.000	0.221	0.468
tax	-0.0103	0.002	-4.154	0.000	-0.015	-0.005
ptratio	-0.7998	0.099	-8.053	0.000	-0.995	-0.605
lstat	-0.4829	0.038	-12.705	0.000	-0.558	-0.408
Omnibus:	71.181	Durbin-Watson:	1.066			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	133.879			
Skew:	0.819	Prob(JB):	8.48e-30			
Kurtosis:	4.915	Cond. No.	1.21e+04			

Hình 2.9 Kết quả tất cả p-value đạt điều kiện (lần 1)

- Kiểm tra VIF lần 1, loại bỏ biến “rad” = 0.395 có VIF = 12.1, cao nhất trong tất cả các biến có VIF > 5.
- Kiểm tra VIF lần 2, xác nhận tất cả các biến đều có VIF < 5.

Features	crim	zn	chas	nox	rm	dis	tax	ptratio	lstat
VIF Factor	4.8	2.4	1	3.9	1.9	3.5	4.8	1.8	2.8

Bảng 2.9 Kết quả tất cả VIF đạt điều kiện (lần 1)

- Kiểm tra p-value lần 5, loại bỏ biến “crim” có p-value = 0.497, cao nhất trong tất cả các biến có p-value > 0.05.
- Kiểm tra p-value lần 6, xác nhận tất cả các biến đều có p-value  $\leq 0.05$ .

OLS Regression Results						
<hr/>						
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				
<hr/>						
	coef	std err	t	P> t	[0.025	0.975]
<hr/>						
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
<hr/>						
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	
<hr/>						

Hình 2.10 Kết quả tất cả p-value đạt điều kiện (lần 2)

- Kiểm tra VIF lần 3, xác nhận tất cả các biến đều có  $VIF < 5$ . Lựa chọn tập biến này để thực hiện xây dựng mô hình hồi quy.

Features	zn	chas	nox	rm	dis	tax	ptratio	lstat
VIF Factor	2.4	1	3.8	1.9	3.5	2.5	1.7	2.5

Bảng 2.10 Kết quả tất cả VIF đạt điều kiện (lần 2)

#### 2.4.2. Xây dựng mô hình

Mô hình của dữ liệu này sẽ được huấn luyện thông qua 80% dữ liệu gốc. Việc chia dữ liệu này giúp việc đánh giá mô hình sau khi được xây dựng một cách khách quan tránh được trường hợp mô hình được xây dựng quá khớp với dữ liệu huấn luyện và không thể tổng quát hóa tốt trên dữ liệu mới.

Mô hình hồi quy tuyến tính đa biến này sẽ có dạng:

$$\widehat{Medv} = 37.56 + 0.04*Zn + 1.82*Chas - 15.52*Nox + 2.87*Rm - 0.94*Dis - 0.002*Tax - 0.77 *Ptratio - 0.53*Lstat, (\$/1000), \text{ với } R^2 = 0.78$$

#### Nhận xét

- Trong điều kiện các yếu tố khác không đổi, khi tỷ lệ đất ở được quy hoạch cho các lô có diện tích trên 25.000 m<sup>2</sup> tăng 1 đơn vị thì giá nhà trung bình của những ngôi nhà tăng \$4.
- Trong điều kiện các yếu tố khác không đổi, khi ngôi nhà gần sông Charles thì giá trị trung bình của những ngôi nhà tăng \$1820.
- Trong điều kiện các yếu tố khác không đổi, khi nồng độ oxit nitric tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà giảm \$15520.
- Trong điều kiện các yếu tố khác không đổi, khi số phòng trung bình/căn hộ tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà tăng \$2870.
- Trong điều kiện các yếu tố khác không đổi, khi khoảng cách tới các trung tâm việc làm ở Boston tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà giảm \$940.

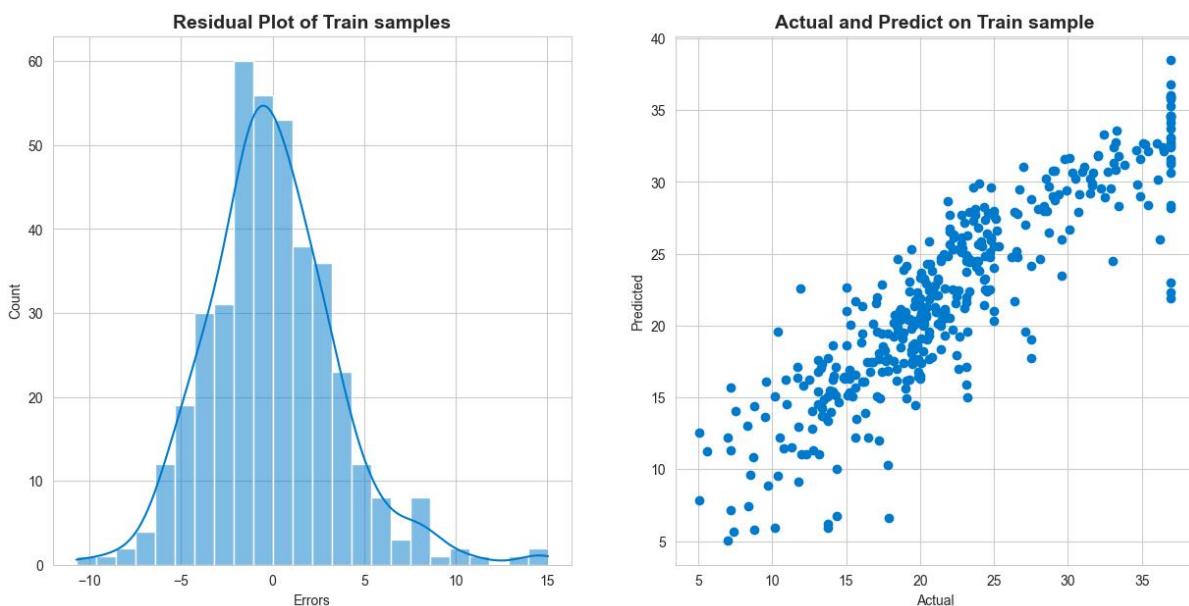
- Trong điều kiện các yếu tố khác không đổi, khi thuế suất tài sản trên 10.000 USD tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà giảm \$0,2.
- Trong điều kiện các yếu tố khác không đổi, khi tỷ lệ học sinh-giáo viên trong thị trấn tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà giảm \$770.
- Trong điều kiện các yếu tố khác không đổi, khi tỷ lệ dân số có địa vị thấp tăng lên 1 đơn vị thì giá trị trung bình của những ngôi nhà giảm \$530.
- Với mô hình này, các biến độc lập giải thích được 78% sự biến thiên của biến phụ thuộc.

### 2.4.3. Đánh giá mô hình

Mô hình hồi quy tuyến tính đa biến sẽ được kiểm nghiệm thêm trên tập huấn luyện và tập thử nghiệm để có nhiều cơ sở đánh giá mô hình

#### 2.4.3.1. Đánh giá mô hình trên tập huấn luyện

Với tập dữ liệu này ngoài giá trị  $R^2 = 0.78$ , mô hình còn được đánh giá thêm các chỉ số MSE = 12.6, MAE = 2.7, biểu đồ phân phối của các sai số giữa giá trị thực tế và giá trị dự đoán và biểu đồ phân tán giữa giá trị thực tế và giá trị dự đoán.



Hình 2.11 Biểu đồ đánh giá mô hình hồi quy đa biến trên tập huấn luyện

## Nhận xét

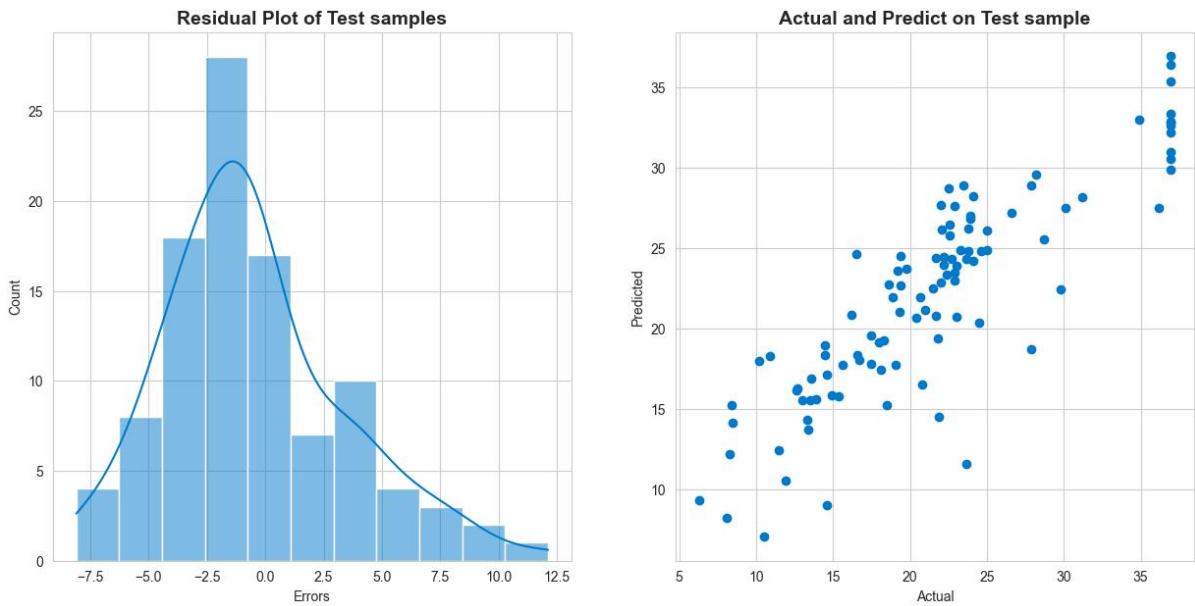
- Trung bình sai số của giữa giá trị dự đoán và giá trị thực té MSE là 12.6. MSE tăng trưởng nhanh hơn khi có các sai số lớn.
- Trung bình giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực té MAE là 2.7. MAE không bị ảnh hưởng bởi các sai số lớn.
- Phân phối các sai số của mô hình trải dài từ [-10: 15] trong đó tập trung chủ yếu ở khoảng [-5:5]. Như vậy sai số của mô hình tập trung ở mức độ tương đối nhỏ và giá trị sai số lớn chiếm một phần không quá lớn.
- Biểu đồ phân tán cho thấy các điểm dữ liệu đa phần có xu hướng được dự đoán tốt vì nằm thành một đường thẳng. Tuy nhiên cũng có những điểm giá trị có xu hướng khác với xu hướng chung.

## Kết luận

- Mô hình có xu hướng dự đoán tốt ở một số điểm dữ liệu vì sai số rất ít và ngược lại có một số ít điểm dữ liệu dự đoán sai lệch cao so với thực tế, điều này có thể do ảnh hưởng của các giá trị ngoại lai hoặc các điểm dữ liệu đặc biệt.
- Mô hình đưa ra dự đoán dựa trên tập huấn luyện khá tốt khi giá nhà dao động từ 5-36 (đơn vị \$1000, dựa trên thống kê giá trị medv) và dự đoán được mô hình đưa ra phần lớn chỉ chênh lệch khoảng 5 so với giá trị thực.

### 2.4.3.2. Đánh giá mô hình trên tập thử nghiệm

Với tập thử nghiệm, mô hình ghi nhận  $R^2 = 0.75$ , MSE = 14.9 và MAE = 3.03 và biểu đồ phân phối của các sai số giữa giá trị thực té và giá trị dự đoán và biểu đồ phân tán giữa giá trị thực té và giá trị dự đoán.



Hình 2.12 Biểu đồ đánh giá mô hình hồi quy đa biến trên tập thử nghiệm

### Nhận xét

- Với mô hình này, biến độc lập giải thích được 75% sự biến thiên của biến phụ thuộc → giảm 3% so với tập dữ liệu huấn luyện.
- Trung bình sai số của giữa giá trị dự đoán và giá trị thực té MSE là 14.9 → tăng 2,3% so với tập dữ liệu huấn luyện
- Trung bình giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực té MAE là 3.03 → tăng 0,33% so với tập dữ liệu huấn luyện.
- Phân phối sai số của mô hình trải dài từ [-7.5: 12.5] trong đó tập trung chủ yếu ở khoảng [-5: 2.5]. Như vậy sai số của mô hình tập trung ở mức độ tương đối nhỏ và giá trị sai số lớn chiếm một phần không quá lớn.
- Biểu đồ phân tán cho thấy các điểm dữ liệu đa phần có xu hướng được dự đoán tốt vì nằm thành một đường thẳng. Tuy nhiên cũng có những điểm giá trị có xu hướng khác với xu hướng chung.

### Kết luận

- 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.

- Tuy tự trên tập huấn luyện, mô hình có xu hướng dự đoán tốt ở một số điểm dữ liệu và ngược lại có một số điểm dữ liệu dự đoán bị sai lệch cao so với thực tế, điều này có thể do ảnh hưởng của các giá trị ngoại lai hoặc các điểm dữ liệu đặc biệt.
- Mô hình đưa ra dự đoán dựa trên tập mới khá tốt khi giá nhà dao động từ 5-36 (đơn vị \$1000 dựa trên thông kê giá trị medv) và dự đoán được mô hình đưa ra phần lớn chỉ chênh lệch khoảng 5 so với giá trị thực.

#### **2.4.3.3. Kết luận và dự đoán**

Như vậy, mô hình hồi quy tuyến tính đa biến dùng để dự đoán giá có hiệu suất khá tốt trên cả hai tập huấn luyện và thử nghiệm. Mô hình này đã cải thiện đáng kể các chỉ số so với mô hình hồi quy đơn biến và cho ra sai số kết quả tốt hơn. Công ty có thể sử dụng mô hình để dự đoán giá nhà dựa trên những yếu tố được liệt kê trong mô hình cũng như tập trung vào các biện pháp cải thiện hoặc phát triển những yếu tố này để tăng giá trị của ngôi nhà.

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

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

## CHƯƠNG 3: MÔ HÌNH HỒI QUY LOGISTIC

### 3.1. Mô tả bài toán

#### 3.1.1. Đặt vấn đề

Một công ty giáo dục cung cấp các khóa học trực tuyến và tiếp thị các khóa học của mình trên một số trang web và công cụ tìm kiếm. Khi những người này truy cập trang web, họ có thể xem qua các khóa học hoặc điền vào biểu mẫu cho khóa học hoặc xem một số video. Khi những người này điền vào biểu mẫu cung cấp địa chỉ email hoặc số điện thoại, họ được phân loại là khách hàng tiềm năng.

Sau khi có được những khách hàng tiềm năng này, nhân viên từ nhóm bán hàng bắt đầu thực hiện cuộc gọi, viết email, v.v. Thông qua quá trình này, một số khách hàng tiềm năng sẽ được chuyển đổi trong khi hầu hết thì không. Tỷ lệ chuyển đổi khách hàng tiềm năng điển hình ở công ty này là khoảng 30%. Có thể nhận thấy rằng, tỷ lệ chuyển đổi của họ khá thấp trong khoảng 100 khách hàng tiềm năng thì chỉ có khoảng 30 khách hàng tiềm năng trong số đó chuyển đổi.

Do đó để làm cho quá trình này hiệu quả hơn, công ty mong muốn xác định những khách hàng tiềm năng. Nếu họ xác định thành công nhóm khách hàng tiềm năng này, tỷ lệ chuyển đổi khách hàng tiềm năng sẽ tăng lên vì đội ngũ bán hàng sẽ tập trung nhiều hơn vào việc giao tiếp với khách hàng tiềm năng đặc biệt thay vì thực hiện cuộc gọi cho tất cả mọi người.

#### 3.1.2. Bộ dữ liệu

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.
How did you hear about X Education	Khách hàng biết đến X Education từ đâu.
What is your current occupation	Cho biết khách hàng là sinh viên, đang thất nghiệp hay đã đi làm.
What matters most to you in choosing this	Lý do chọn khóa học

course	
Search	
Magazine	
Newspaper Article	Khách hàng đã xem quảng cáo trong bất kỳ mục nào được liệt kê hay chưa.
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	Khách hàng biết đến thông qua giới thiệu.
Receive More Updates About Our Courses	Khách hàng có chọn nhận thêm thông tin cập nhật về các khóa học hay không.
Tags	Trạng thái khách hàng.
Lead Quality	Đánh giá chất lượng khách hàng.
Update me on Supply Chain Content	Khách hàng có muốn nhận nội dung cập nhật về chuỗi cung ứng.
Get updates on DM Content	Khách hàng có muốn nhận về cập nhật nội dung thông qua tin nhắn.
Lead Profile	Cấp độ khách hàng tiềm năng được đánh giá cho mỗi

	khách hàng dựa trên hồ sơ của họ.
City	Thành phố của khách hàng
Asymmetrique Activity Index	
Asymmetrique Profile Index	Chỉ số và điểm số (liên quan đến hệ thống đánh giá khách hàng trong Marketing) được gán cho từng khách hàng dựa trên hoạt động và hồ sơ của họ.
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Khách hàng đã đồng ý thanh toán số tiền qua séc hay không.
a free copy of Mastering The Interview	Khách hàng muốn một bản sao miễn phí của 'Mastering the Interview' hay không
Last Notable Activity	Hoạt động đáng chú ý cuối cùng.

Bảng 3.1 Mô tả dữ liệu

### 3.1.3. Xây dựng câu hỏi nghiên cứu

- Có yếu tố nào có thể tác động mạnh đến tỷ lệ chuyển đổi khách hàng tiềm năng cho công ty giáo dục?
- Những đặc điểm chung của những khách hàng tiềm năng đặc biệt?

⇒ Có thể tận dụng những yếu tố này để phát triển các kế hoạch hoặc xây dựng mô hình chuyển đổi hiệu quả giúp công ty kinh doanh hiệu quả.

### 3.2. Làm sạch và khám phá dữ liệu

#### 3.2.1. Tổng quan cấu trúc bộ dữ liệu

Trước tiên, bộ dữ liệu sẽ được phân tích tổng quan để hiểu cấu trúc cơ bản của dữ liệu, bao gồm kích thước bộ dữ liệu, các loại dữ liệu và giá trị thiếu (NaN).

Sau khi tiến hành kiểm tra, bộ dữ liệu này bao gồm 37 cột và 9240 mẫu với các kiểu dữ liệu là Float64 (4), Int64 (3) và Object (30). Bảng bên dưới mô tả chi tiết dữ liệu từng cột:

#	Column	Non-Null Count	Data type
1	Prospect ID	9240	Object
2	Lead Number	9240	Int64
3	Lead Origin	9240	Object
4	Lead Source	9204	Object
5	Do Not Email	9240	Object
6	Do Not Call	9240	Object
7	Converted	9240	Int64
8	TotalVisits	9103	Float64
9	Total Time Spent on Website	9240	Int64

10	Page Views Per Visit	9103	Float64
11	Last Activity	9137	Object
12	Country	6779	Object
13	Specialization	7802	Object
14	How did you hear about X Education	7033	Object
15	What is your current occupation	6550	Object
16	What matters most to you in choosing a course	6531	Object
17	Search	9240	Object
18	Magazine	9240	Object
19	Newspaper Article	9240	Object
20	X Education Forums	9240	Object
21	Newspaper	9240	Object
22	Digital Advertisement	9240	Object
23	Through Recommendations	9240	Object
24	Receive More Updates About Our Courses	9240	Object

25	Tags	5887	Object
26	Lead Quality	4473	Object
27	Update me on Supply Chain Content	9240	Object
28	Get updates on DM Content	9240	Object
29	Lead Profile	6531	Object
30	City	7820	Object
31	Asymmetrique Activity Index	5022	Object
32	Asymmetrique Profile Index	5022	Object
33	Asymmetrique Activity Score	5022	Float64
34	Asymmetrique Profile Score	5022	Float64
35	I agree to pay the amount through cheque	9240	Object
36	A free copy of Mastering The Interview	9240	Object
37	Last Notable Activity	9240	Object

Bảng 3.2. Tổng quan về các biến thuộc bộ dữ liệu

Bảng bên dưới mô tả thông kê của các cột có kiểu dữ liệu số. Các thông số này cung cấp những thông tin cơ bản để tiến hành phân tích và đánh giá chất lượng của bộ dữ liệu.

	<b>Lead Number</b>	<b>Converted</b>	<b>Total Visits</b>	<b>Total Time Spent on Website</b>
<b>count</b>	9240	9240	9103	9240
<b>mean</b>	617188.4356	0.38539	3.445238	487.698268
<b>std</b>	23405.9957	0.486714	4.854853	548.021466
<b>min</b>	579533	0	0	0
<b>25%</b>	596484.5	0	1	12
<b>50%</b>	615479	0	3	248
<b>75%</b>	637387.25	1	5	936
<b>max</b>	660737	1	251	2272

Bảng 3.3 Mô tả thống kê #1

	<b>Page Views Per Visit</b>	<b>Asymmetrique Activity Score</b>	<b>Asymmetrique Profile Score</b>
<b>count</b>	9103	5022	5022
<b>mean</b>	2.36282	14.306252	16.344883
<b>std</b>	2.161418	1.386694	1.811395

<b>min</b>	0	7	11
<b>25%</b>	1	14	15
<b>50%</b>	2	14	16
<b>75%</b>	3	15	18
<b>max</b>	55	18	20

Bảng 3.4 Mô tả thống kê #2

### 3.2.2. Làm sạch dữ liệu

#### 3.2.2.1. Xử lý dữ liệu trùng lặp và không hợp lệ

Dữ liệu trùng lặp và không hợp lệ là giá trị không phù hợp với ngữ cảnh hoặc không thể tham khảo. Kết quả sau khi kiểm tra bằng Python đã cho ra bộ dữ liệu này không có dòng nào bị trùng lặp với nhau, đảm bảo các dòng là duy nhất.

Bên cạnh đó, vì bộ dữ liệu này được thu thập từ dữ liệu khách hàng điền vào các biểu mẫu nên bộ dữ liệu này ghi nhận giá trị “Select” ở một số cột (Specialization, How did you hear about X Education, Lead Profile, City,...). Việc này được suy ra là do người được khảo sát không đưa ra lựa chọn ở một vài câu hỏi dropdown/checkbox và giá trị “Select” được ghi nhận lại thay thế câu trả lời. Tuy dữ liệu này được ghi nhận nhưng chúng không có giá trị tham khảo nên sẽ được chuyển thành dạng dữ liệu thiêu (NaN).

Ngoài ra, dữ liệu ở hai cột “Prospect ID” và “Lead Number” là những dữ liệu cố định, không thể hiện mối quan hệ với các biến khác và không có tính đóng góp vào mục tiêu nghiên cứu nên hai cột này sẽ được loại bỏ để quá trình phân tích phía sau được thuận lợi hơn. Bên cạnh đó xem qua dữ liệu có thể thấy, cột “Last Notable Activity” là dữ liệu được trích xuất từ cột “Last Activity” nên cột “Last Notable Activity” sẽ được loại bỏ để tránh trùng lặp dữ liệu.

### **3.2.2.2. Xử lý dữ liệu thiếu (NaN)**

Dữ liệu dữ liệu thiếu (NaN) là những dữ liệu sẽ làm sai lệch kết quả phân tích và tạo ra thông tin thừa nên trước khi phân tích cần được kiểm tra và loại bỏ phù hợp.

Bảng bên dưới tổng hợp số lượng giá trị thiếu (NaN) và phần trăm của chúng ở mỗi cột lần lượt qua các cột “Total” và “Percent”.

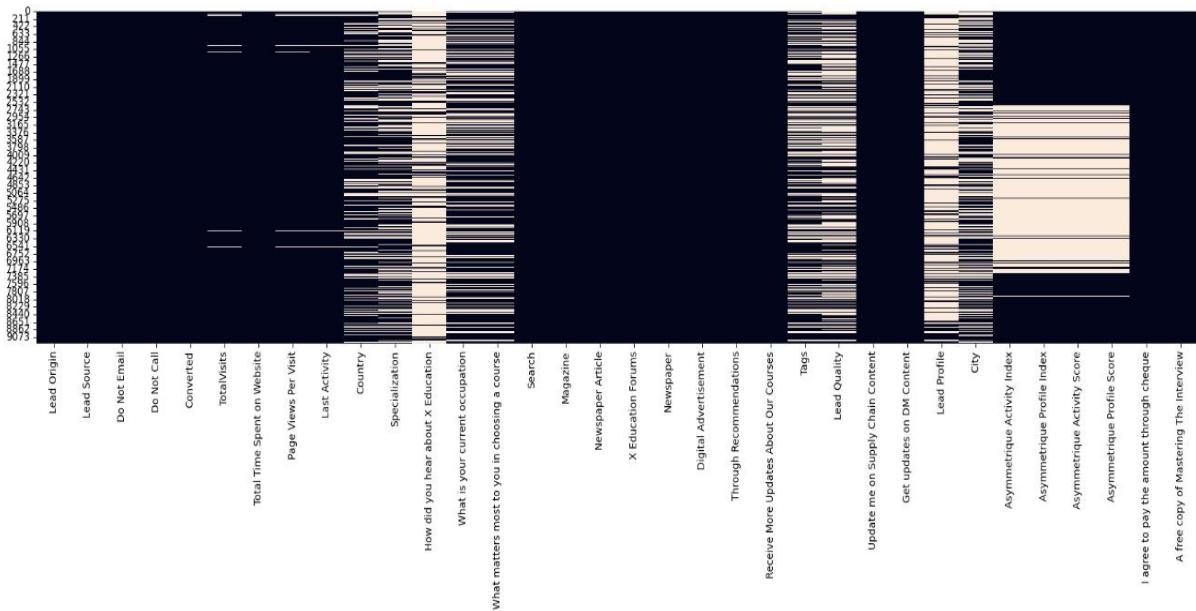
<b>Index</b>	<b>Column</b>	<b>Total</b>	<b>Percent</b>
1	How did you hear about X Education	7250	78.46
2	Lead Profile	6855	74.19
3	Lead Quality	4767	51.59
4	Asymmetrique Profile Score	4218	45.65
5	Asymmetrique Activity Score	4218	45.65
6	Asymmetrique Activity Index	4218	45.65
7	Asymmetrique Profile Index	4218	45.65
8	City	3669	39.71
9	Specialization	3380	36.58
10	Tags	3353	36.29
11	What matters most to you in choosing a course	2709	29.32

12	What is your current occupation	2690	29.11
13	Country	2461	26.63
14	Page Views Per Visit	137	1.48
15	TotalVisits	137	1.48
16	Last Activity	103	1.11
17	Lead Source	36	0.39
18	Receive More Updates About Our Courses	0	0
19	I agree to pay the amount through cheque	0	0
20	Get updates on DM Content	0	0
21	Update me on Supply Chain Content	0	0
22	A free copy of Mastering The Interview	0	0
23	Newspaper Article	0	0
24	Through Recommendations	0	0
25	Digital Advertisement	0	0
26	Newspaper	0	0

27	X Education Forums	0	0
28	Magazine	0	0
29	Search	0	0
30	Total Time Spent on Website	0	0
31	Converted	0	0
32	Do Not Call	0	0
33	Do Not Email	0	0
34	Lead Origin	0	0

*Bảng 3.5 Mô tả thông kê các giá trị thiếu*

Nhìn vào bảng thống kê, bộ dữ liệu này bao gồm 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%. Để có thêm căn cứ quyết định xử lý dữ liệu thiếu (NaN) mà vẫn đảm bảo tính chính xác của bộ dữ liệu, biểu đồ ma trận sẽ được dùng để trực quan hóa cấu trúc của dữ liệu bị thiếu. Hình bên dưới mô tả ma trận, với dòng màu trắng biểu thị với giá trị NaN và dòng màu tối biểu thị dữ liệu có sẵn.



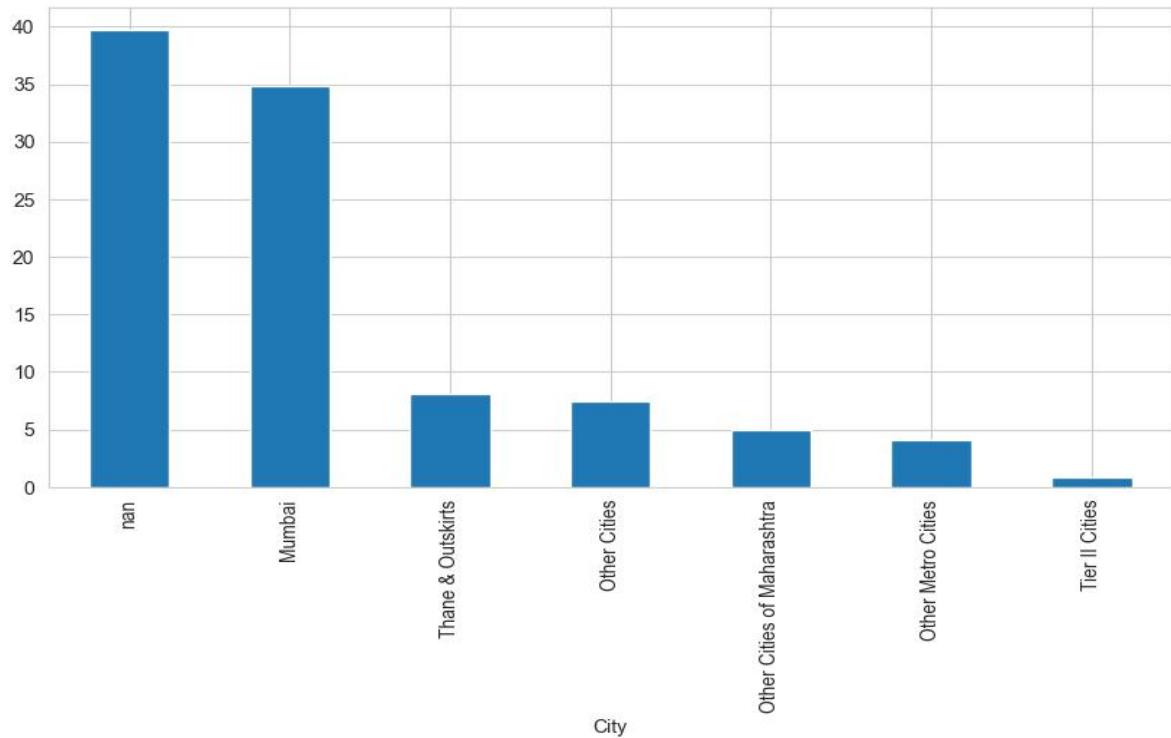
*Hình 3.1 Trực quan dữ liệu thiếu*

Nhìn vào ma trận, mỗi hàng sẽ có số lượng cột có dữ liệu thiếu (NaN) khác nhau. Tình trạng thiếu không đồng đều giữa các cột có thể dẫn đến lệc kết quả phân tích, đặc biệt nếu các hàng bị thiếu hụt chứa thông tin quan trọng. Với những cột có nhiều giá trị thiếu (NaN) trên 70% (“How did you hear about X Education”, “Lead Profile”) sẽ không có đủ thông tin để tìm ra mối quan hệ giữa các biến, nếu đưa vào sử dụng thì kết quả phân tích sẽ bị lệch nghiêm trọng. Bên cạnh đó, những cột có giá trị thiếu (NaN) từ 70% đến 40% (“Lead Quality”, “Asymmetrique Activity Index”, “Asymmetrique Profile Index” “Asymmetrique Activity Score” “Asymmetrique Profile Score”) theo mô tả thì đây là chỉ số và điểm số được gán cho từng khách hàng dựa trên hoạt động và hồ sơ của họ, do đó những cột này không phản ánh bản chất đặc điểm của khách hàng nên không có giá trị tham khảo. Từ đó, những cột có giá trị thiếu (NaN) trên 40% sẽ được loại bỏ. Cột “Tags” có không có mối liên hệ rõ ràng giữa các biến để dự đoán giá trị bị thiếu và không có dữ liệu để đánh giá cột này có khả năng được thực hiện thủ công và có giá trị thiếu khá cao nên cần được loại bỏ khỏi mô hình.

### 3.2.2.3. Xử lý giá trị thiếu (NaN) từng cột

#### Cột “City”

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “City”.

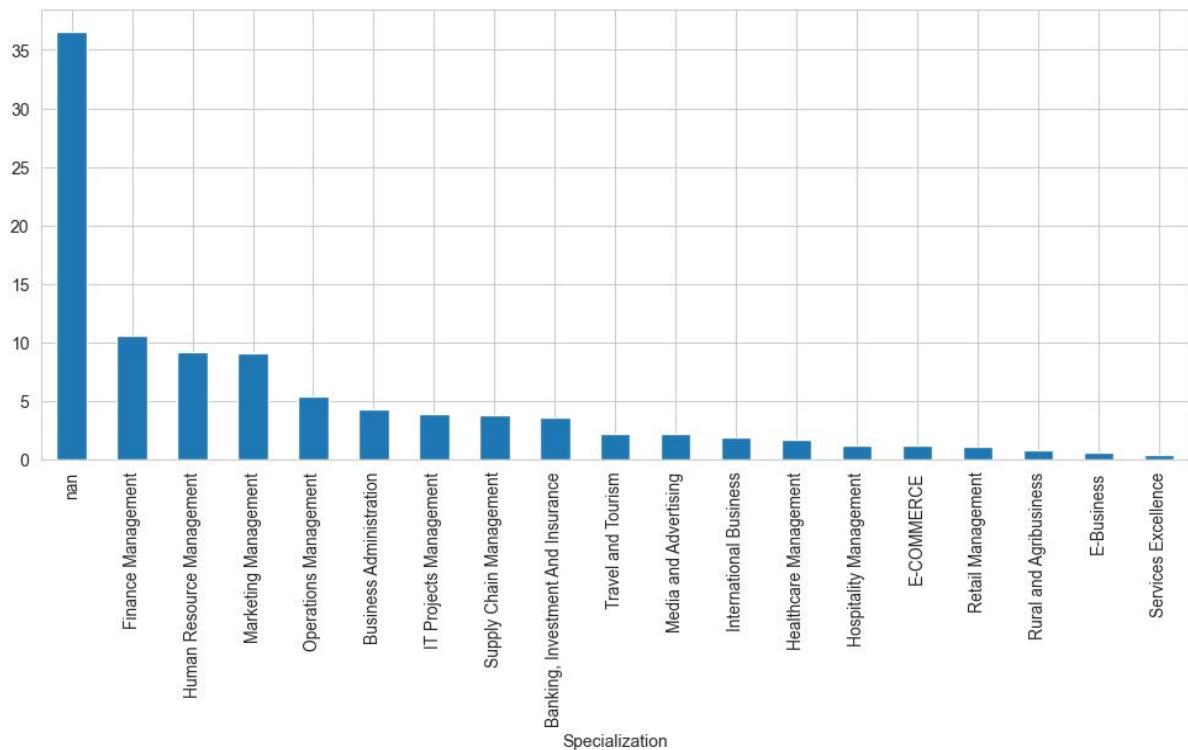


*Hình 3.2 Trực quan dữ liệu “City”*

Cột “City” có 6 giá trị khác nhau thể hiện các thành phố. Giá trị thiếu (NaN) là giá trị phổ biến nhất và không cung cấp thông tin hữu ích; giá trị Mumbai có số lượng lớn đáng kể thứ hai. Do đó thay thế giá trị thiếu (NaN) bằng giá trị “Mumbai sẽ không làm biến đổi lớn xu hướng trong dữ liệu, giảm thiểu nhiều khi phân tích sau đó.

### **Cột “Specialization”**

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “Specialization”.

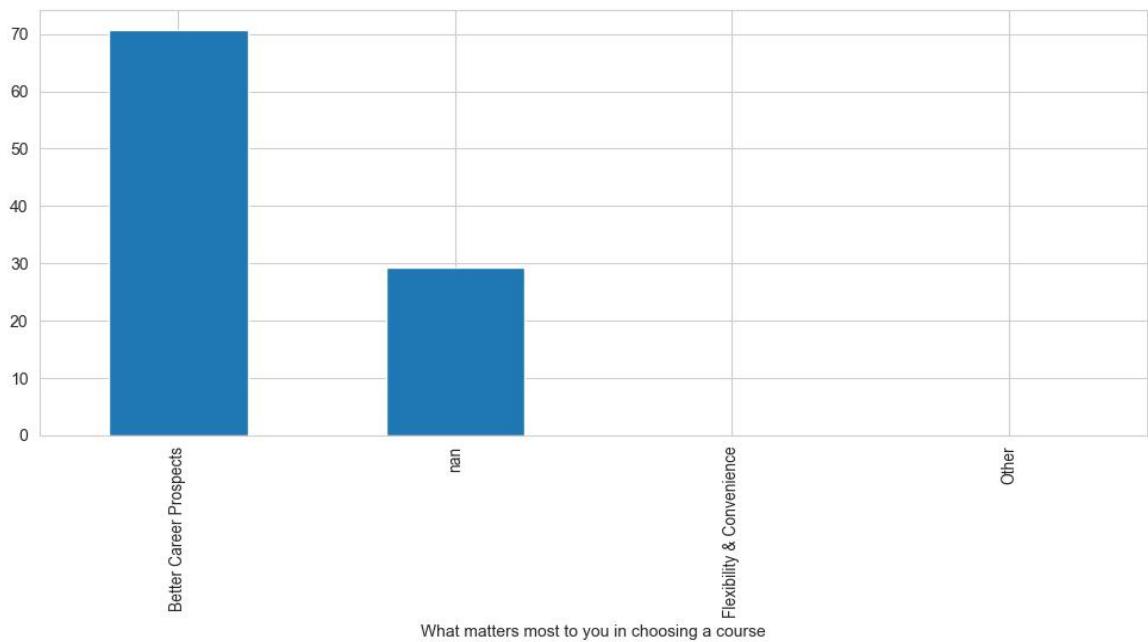


*Hình 3.3 Trực quan dữ liệu “Specialization”*

Cột “Specialization” có 18 giá trị khác nhau và không có mối liên hệ rõ ràng giữa các biến để dự đoán giá trị bị thiếu. Giá trị thiếu (NaN) là giá trị phổ biến nhất, có thể đại diện cho một ý nghĩa cụ thể. Do đó, những giá trị thiếu (NaN) sẽ được gán thành giá trị “Other Specialization” để phân biệt rõ ràng các giá trị thiếu và giá trị ban đầu, tránh hiểu lầm khi tiến hành phân tích các bước tiếp theo. Ngoài ra, việc này giữ cho dữ liệu gốc được bảo toàn mà không làm biến đổi quá mức.

#### ***Cột “What matters most to you in choosing a course”***

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “What matters most to you in choosing a course”.

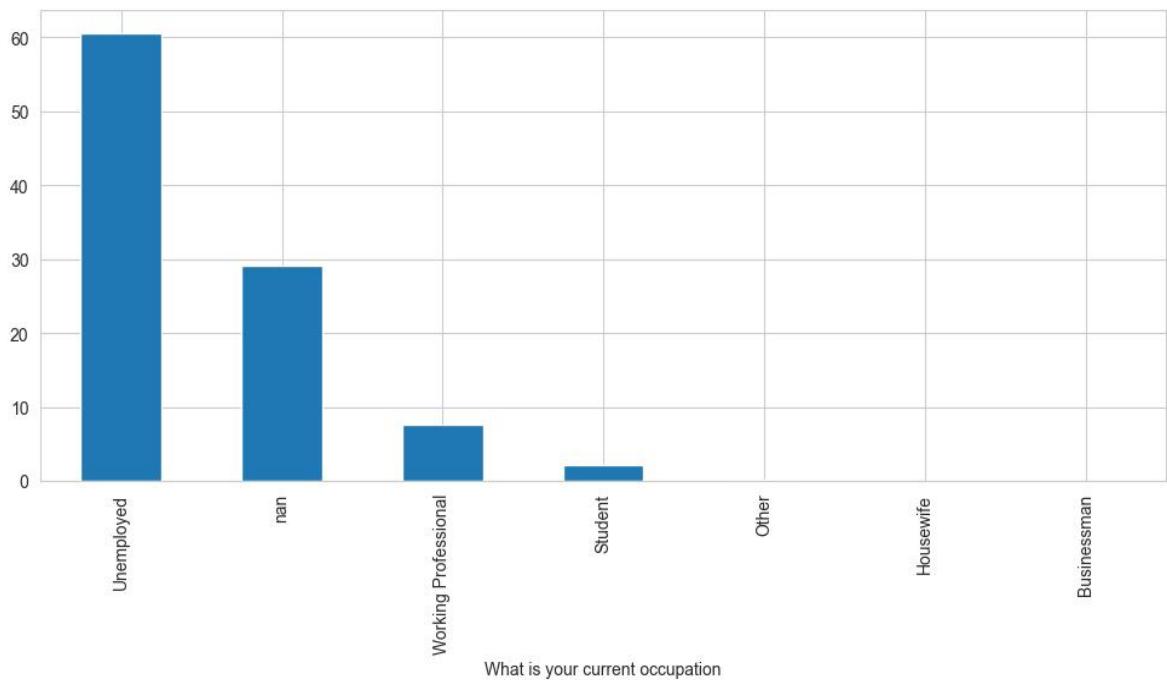


*Hình 3.4 Trực quan dữ liệu “What matters most to you in choosing a course”*

Cột “What matters most to you in choosing a course” có 3 giá trị khác nhau và không có mối liên hệ rõ ràng giữa các biến để dự đoán giá trị bị thiếu. Giá trị “Better Career Prospects” là giá trị phổ biến nhất và phản ánh xu hướng chung; giá trị thiếu (NaN) có số lượng đáng kể thứ hai và các biến còn lại cực ít. Do đó thay thế giá trị thiếu (NaN) bằng giá trị “Better Career Prospects” sẽ không làm biến đổi lớn xu hướng trong dữ liệu, giảm thiểu nhiều khi phân tích sau đó.

#### **Cột “What is your current occupation”**

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “What is your current occupation”.

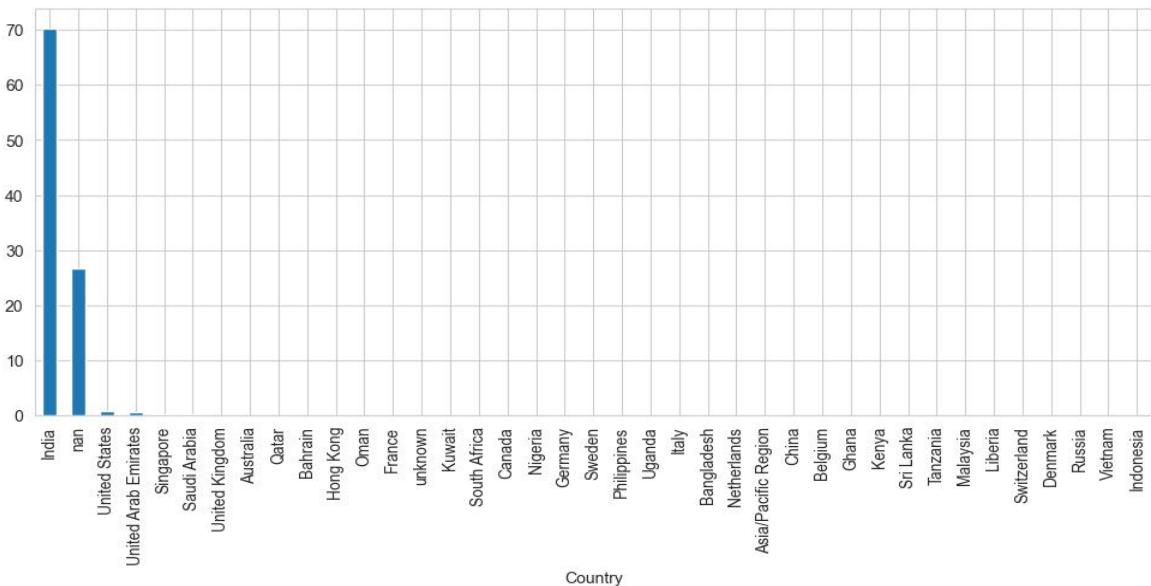


*Hình 3.5 Trục quan dữ liệu “What is your current occupation”*

Cột “What is your current occupation” có 6 giá trị khác nhau. Giá trị “Unemployed” là giá trị phổ biến nhất và phản ánh xu hướng chung; giá trị thiểu (NaN) có số lượng đáng kể thứ hai và các biến còn lại rất ít. Do đó thay thế giá trị thiểu (NaN) bằng giá trị “Unemployed” sẽ không làm biến đổi lớn xu hướng trong dữ liệu, giảm thiểu nhiễu khi phân tích sau đó.

### **Cột “Country”**

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “Country”.

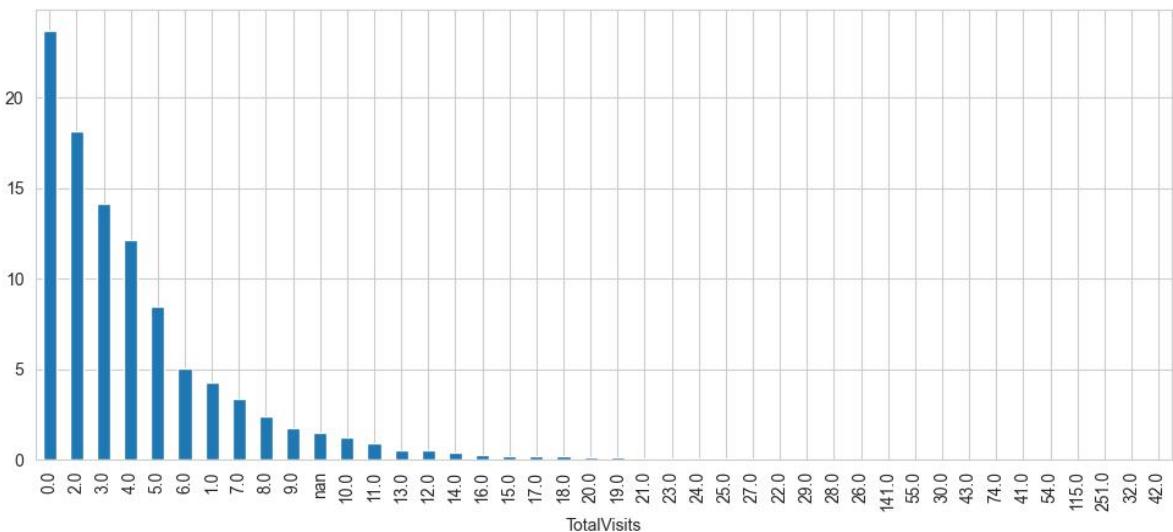


Hình 3.6 Trực quan dữ liệu “Country”

Cột “Country” có 38 giá trị khác nhau và không có mối liên hệ rõ ràng giữa các biến để dự đoán giá trị bị thiếu. Giá trị “India” là giá trị phổ biến nhất và phản ánh xu hướng chung; giá trị thiếu (NaN) có số lượng đáng kể thứ hai và các biến còn lại cực kỳ ít. Do đó thay thế giá trị thiếu (NaN) bằng giá trị “India” sẽ không làm biến đổi lớn xu hướng trong dữ liệu, giảm thiểu nhiều khi phân tích sau đó.

### Cột “TotalVisits”

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “TotalVisits”

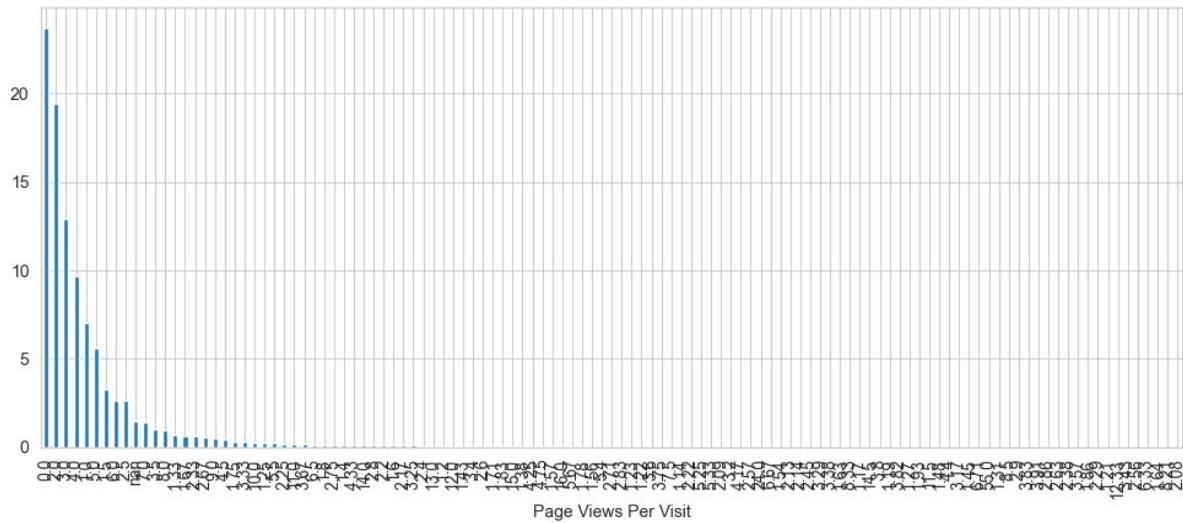


*Hình 3.7 Trực quan dữ liệu “TotalVisits”*

Vì tính chất của kiểu dữ liệu trong cột này nên để tránh việc thay thế một dữ liệu mới sẽ trở thành giá trị ngoại lệ. Giá trị thiếu (NaN) ở cột này sẽ thay thế bằng dữ liệu trung vị của cột.

#### ***“Page Views Per Visit”***

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “Page Views Per Visit”.

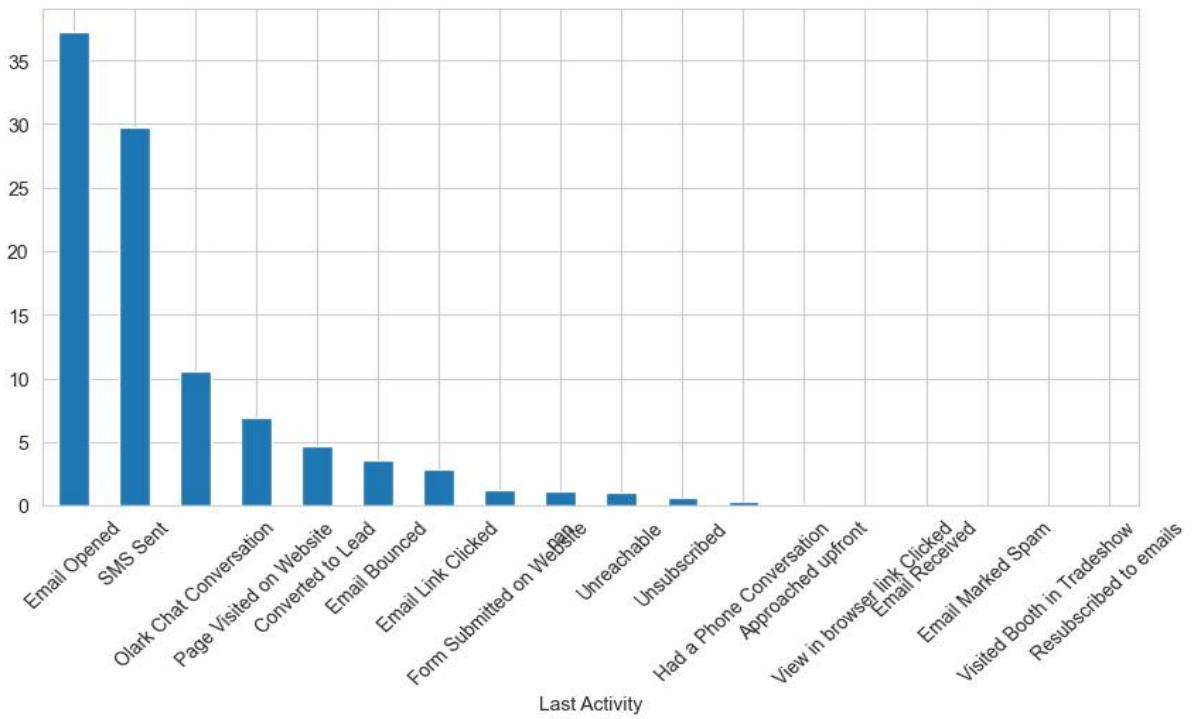


*Hình 3.8 Trực quan dữ liệu “Page Views Per Visit”*

Vì tính chất của kiểu dữ liệu trong cột này nên để tránh việc thay thế một dữ liệu mới là một giá trị ngoại lệ. Giá trị thiếu (NaN) ở cột này sẽ thay thế bằng dữ liệu trung vị của cột.

#### ***Cột “Last Activity”***

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “Last Activity”.

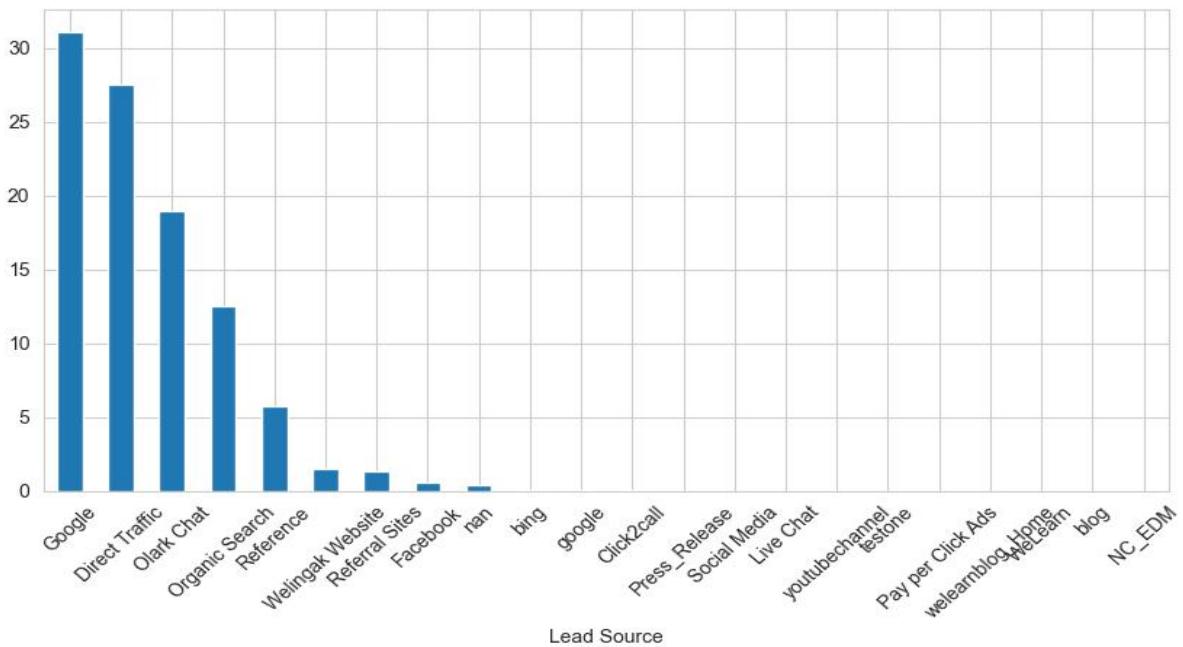


Hình 3.9 Trực quan dữ liệu “Last Activity”

Cột “Last Activity” có 17 giá trị khác nhau và không có mối liên hệ rõ ràng giữa các biến để dự đoán giá trị bị thiếu. Giá trị “Email Opened” là giá trị phổ biến nhất và phản ánh xu hướng chung; giá trị thiếu (NaN) có số lượng cực kỳ ít. Do đó thay thế giá trị thiếu (NaN) bằng giá trị “Email Opened” sẽ không làm biến đổi lớn xu hướng trong dữ liệu, giảm thiểu nhiều khi phân tích sau đó.

#### Cột “Lead Source”

Biểu đồ bên dưới mô tả số lượng mỗi giá trị duy nhất trong cột “Lead Source”.



Hình 3.10 Trực quan dữ liệu “Lead Source”

Cột “Lead Source” có 21 giá trị khác nhau và không có mối liên hệ rõ ràng giữa các biến để dự đoán giá trị bị thiếu. Giá trị “Google” là giá trị phổ biến nhất, phản ánh xu hướng chung và giá trị thiếu (NaN) có số lượng cực kỳ ít. Do đó thay thế giá trị thiếu (NaN) bằng giá trị “Google” sẽ không làm biến đổi lớn xu hướng trong dữ liệu, giảm thiểu nhiễu khi phân tích sau đó. Bên cạnh đó, tồn tại giá trị “google” tương tự giá trị “Google” nên hai giá trị này sẽ được gộp lại với nhau.

Kết quả sau khi xử lý giá trị thiếu (NaN), tất cả giá trị bị thiếu (NaN) ở các cột được thay thế bằng các giá trị phù hợp để đảm bảo quá trình phân tích phía sau. Bảng dưới thể hiện kết quả kiểm tra giá trị thiếu (NaN) ở mỗi cột.

Column	Non - Null Count
Lead Origin	FALSE
Lead Source	FALSE
Do Not Email	FALSE

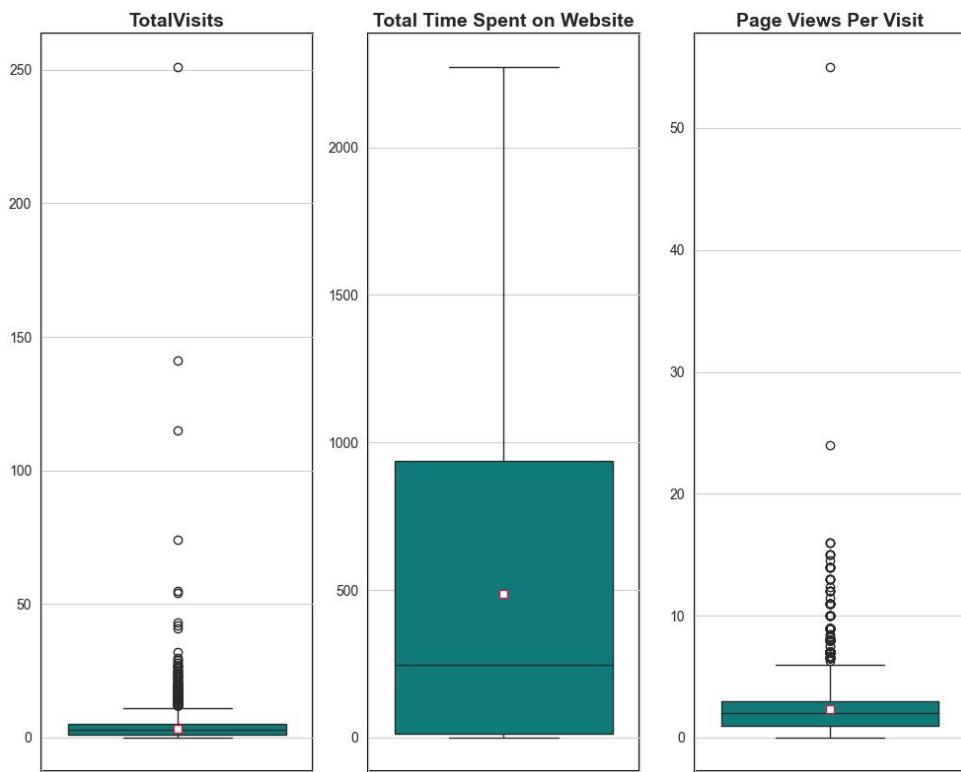
Do Not Call	FALSE
Converted	FALSE
TotalVisits	FALSE
Total Time Spent on Website	FALSE
Page Views Per Visit	FALSE
Last Activity	FALSE
Country	FALSE
Specialization	FALSE
What is your current occupation	FALSE
What matters most to you in choosing a course	FALSE
Search	FALSE
Magazine	FALSE
Newspaper Article	FALSE
X Education Forums	FALSE
Newspaper	FALSE

Digital Advertisement	FALSE
Through Recommendations	FALSE
Receive More Updates About Our Courses	FALSE
Tags	FALSE
Update me on Supply Chain Content	FALSE
Get updates on DM Content	FALSE
City	FALSE
I agree to pay the amount through cheque	FALSE
A free copy of Mastering The Interview	FALSE

Bảng 3.6 Kiểm tra giá trị thiếu

#### 3.2.2.4. Xử lý ngoại lệ

Ở bước này, các cột dữ liệu số (trừ “Convert”) sẽ được kiểm tra ngoại lệ thông qua biểu đồ Boxplot. Biểu đồ Boxplot là một biểu đồ rất có ích để thể hiện các chỉ số thống kê của các biến liên tục, biểu đồ này sẽ hiển thị các dữ liệu ngoại lệ thông qua các chấm tròn.



*Hình 3.11 Trực quan các giá trị ngoại lai*

Dựa vào biểu đồ, ta thấy hai cột “Total Visits” và “Page Views Per Visit” có dữ liệu với số liệu lớn bất thường. Do đó, dữ liệu sẽ được kiểm tra kỹ hơn để phát hiện phân phối của các giá trị ngoại lệ.

	min	5%	25%	50%	75%	90%	95%	99%	99.5%	max
<b>Total Visits</b>	0	0	1	3	5	7	10	17	21	251
<b>Page Views Per Visit</b>	0	0	1	2	3	5	6	9	11	55

*Bảng 3.7 Thống kê phân phối các cột có giá trị ngoại lai*

Dựa vào bảng thống kê, có thể thấy số liệu có sự tăng cao đột ngột ở các giá nằm trong 0.5% cao nhất ở hai cột. Do đó, dữ liệu ở phạm vi này sẽ được

loại bỏ để giảm thiểu ảnh hưởng của các giá trị ngoại lai tới kết quả phân tích mà vẫn giữ lại sự đa dạng và tính đầy đủ của dữ liệu. Cách làm này được áp dụng vì theo mô tả cột dữ liệu này ghi lại thời gian thực hiện hoạt động của khách hàng, những giá trị này là hợp lý. Theo đó, dữ liệu vẫn sẽ giữ lại một phần giá trị ngoại lai (theo biểu đồ Boxplot) nhưng sẽ được điều chỉnh để nằm trong khoảng xác định.

### **3.2.3 Phân tích khám phá dữ liệu**

#### **3.2.3.1. Tổng quan**

Bảng bên dưới sẽ tổng hợp số lượng giá trị của mỗi cột không phải kiểu dữ liệu số để có cái nhìn tổng quan về tính đầy đủ và chất lượng của dữ liệu.

<b>Columns</b>	<b>Number value</b>
I agree to pay the amount through cheque	1
Get updates on DM Content	1
Update me on Supply Chain Content	1
Receive More Updates About Our Courses	1
Magazine	1
X Education Forums	2
Do Not Email	2
Do Not Call	2
Search	2

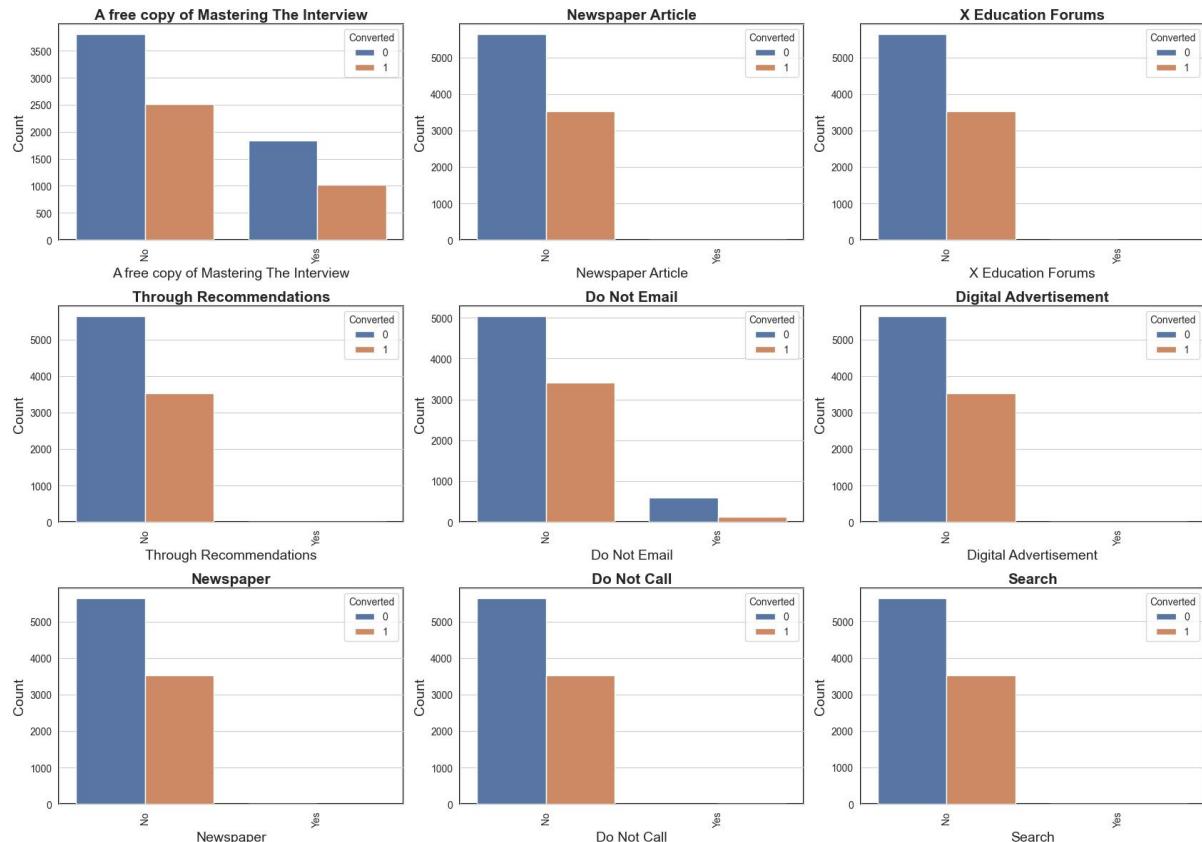
Newspaper Article	2
A free copy of Mastering The Interview	2
Newspaper	2
Digital Advertisement	2
Through Recommendations	2
What matters most to you in choosing a course	3
Lead Origin	5
City	6
What is your current occupation	6
Last Activity	8
Lead Source	8
Specialization	19
Tags	27
Country	38

Bảng 3.8 Số giá trị từng biến

Các cột có kiểu dữ liệu là biến nhị phân chỉ có một giá trị cho tất cả các mẫu, cột này không cung cấp bất kỳ thông tin phân biệt và không thể rút ra bất

kỳ kết luận ý nghĩa nào từ cột đó. Do đó, năm cột “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” sẽ được loại bỏ vì không có tính đa dạng trong dữ liệu.

Những cột nhị phân còn lại sẽ được phân tích qua biểu đồ cột đôi để xem xét tỷ lệ chuyển đổi và giá trị trong cột để hiểu rõ hơn về mối quan hệ giữa cột này và kết quả chuyển đổi.



Hình 3.12 Trực quan hóa dữ liệu nhị phân

Qua biểu đồ, ở các cột có sự chênh lệch lớn giữa số lượng giá trị “Yes” và giá trị “No”. Để xem xét kỹ số lượng “Yes” và “No”, ta có bảng bên dưới

Column	Yes	No
Do Not Call	2	9165
Search	14	9153

Newspaper	1	9166
Newspaper Article	2	9165
Through Recommendations	7	9160
X Education Forums	1	9166
A free copy of Mastering The Interview	2846	6321
Digital Advertisement	4	9163
Do Not Email	719	8448

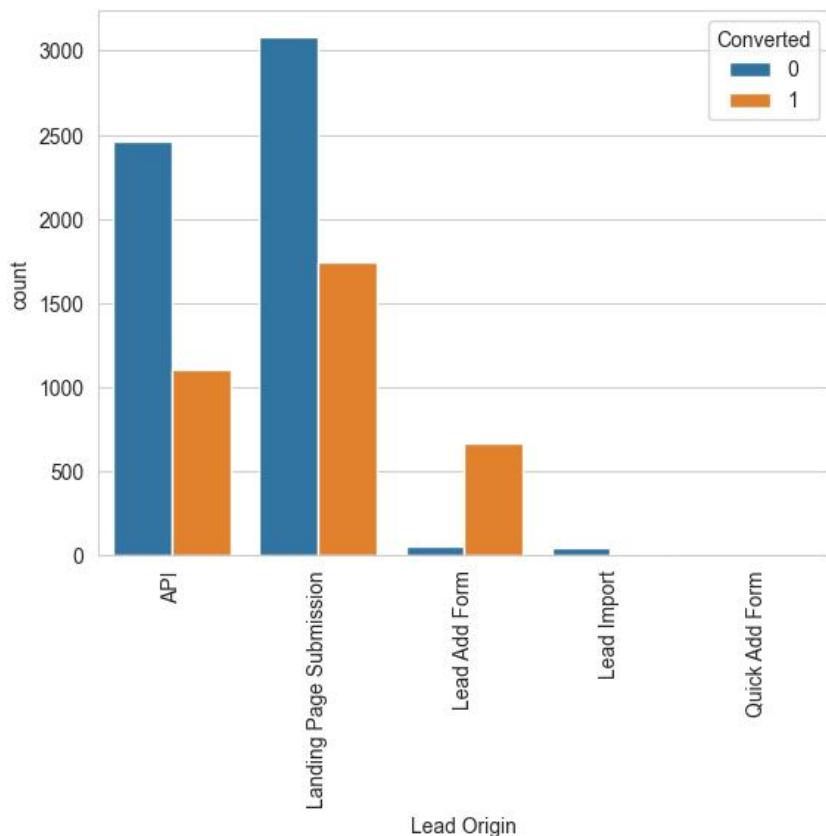
Bảng 3.9 Sự chênh lệch giữa số lượng giá trị biến “Yes” và “No”

Dựa theo thống kê của bản bảng, một số cột bị mất cân đối nghiêm trọng. Dữ liệu không có sự phân biệt và không thể cung cấp mức độ ảnh hưởng đến kết quả chuyển đổi của khách hàng.

Riêng cột “A free copy of Mastering The Interview” và “Do Not Email” không bị mất cân đối nghiêm trọng và có sự phân biệt. Điều này có thể chỉ ra rằng giá trị hai cột này có thể dùng để dự đoán tỷ lệ chuyển đổi.

### 3.2.3.2. Phân tích từng biến

#### Cột “Lead Origin”

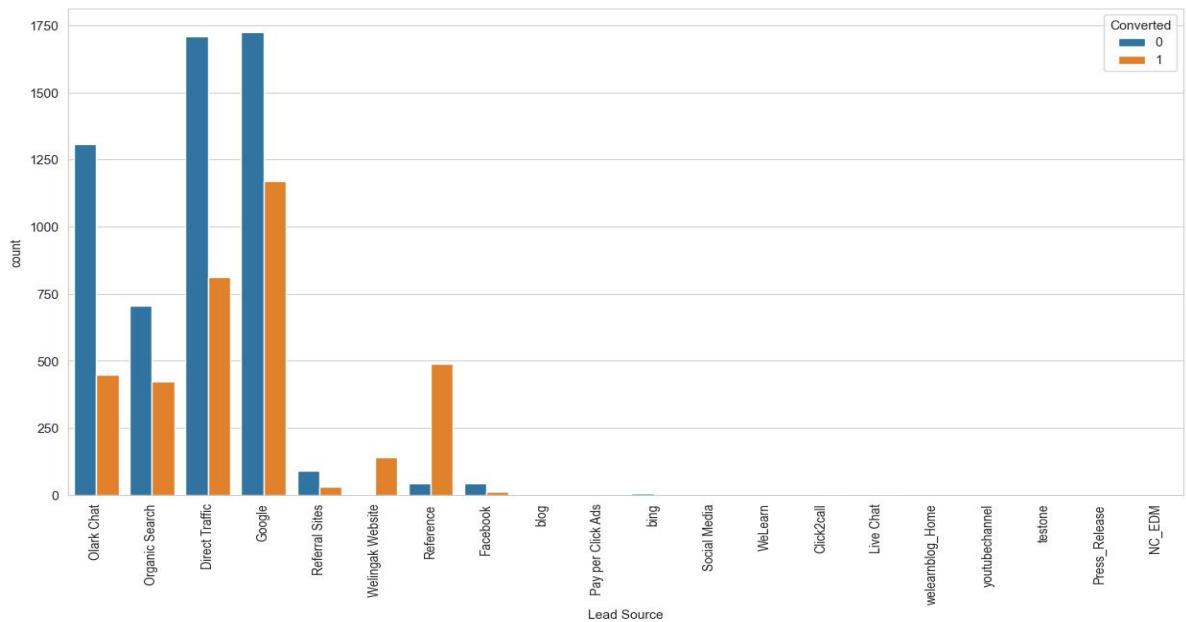


*Hình 3.13 Giá trị của cột “Lead Origin” và tỷ lệ chuyển đổi*

Thông qua biểu đồ phân tích được ảnh hưởng của các mã nguồn khác nhau tới tỷ lệ chuyển đổi khách hành. Cụ thể, ở hai mã nguồn là API và Landing Page Submission là hai mã nguồn có số lượng khách hàng cao nhất với tỷ lệ chuyển đổi khách hàng khoảng 30-35%. Mã nguồn Lead Add Form có tỷ lệ chuyển đổi hơn 90% nhưng số lượng khách hàng không cao. Hai mã nguồn còn lại có số lượng rất ít.

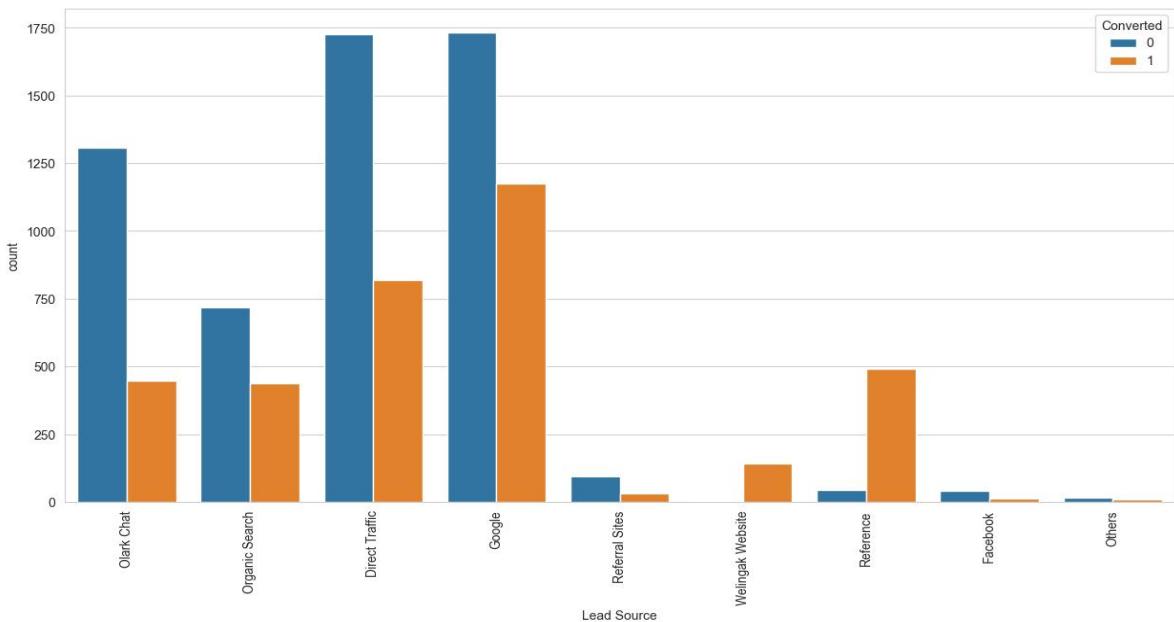
*Kết luận:* Một vài giá trị trong cột “Lead Origin” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

### **Cột “Lead Source”**



*Hình 3.14 Giá trị của cột “Lead Source” và tỷ lệ chuyển đổi*

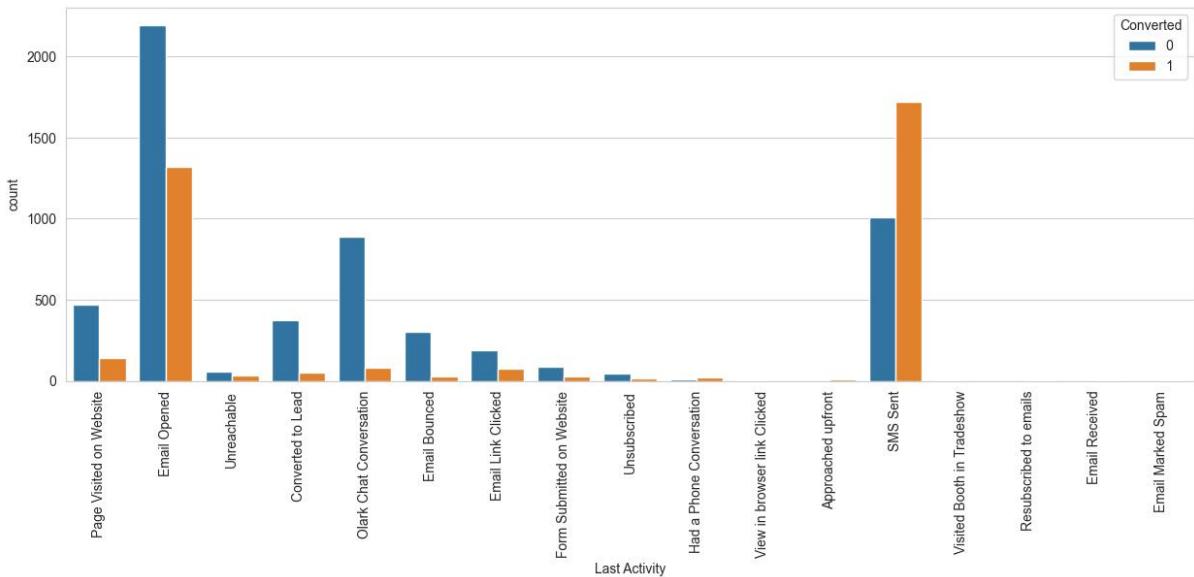
Thông qua biểu đồ ta phân tích được ảnh hưởng của các nguồn truy cập khác nhau tới tỷ lệ chuyển đổi khách hành. Cụ thể, ở các nguồn truy cập là Olark Chat, Organic Search, Direct Traffic, Google có tỷ lệ chuyển đổi khách hàng khoảng 30-35% nhưng số lượng khách hàng có nguồn gốc từ chúng là đáng kể. Nguồn truy cập Reference và Welingak Website có tỷ lệ chuyển đổi cao nhưng số lượng khách hàng không cao lắm. Bên cạnh đó, có thể rõ ràng quan sát thấy số lượng khách hàng tiềm năng từ các nguồn khác gần như không đáng kể và do đó có thể gộp chúng thành nguồn "Other" để có một cách hiển thị và phân tích tốt hơn.



Hình 3.15 Giá trị của cột “Lead Source” và tỷ lệ chuyển đổi sau khi gộp các giá trị

*Kết luận* Một vài giá trị trong cột “Lead Source” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

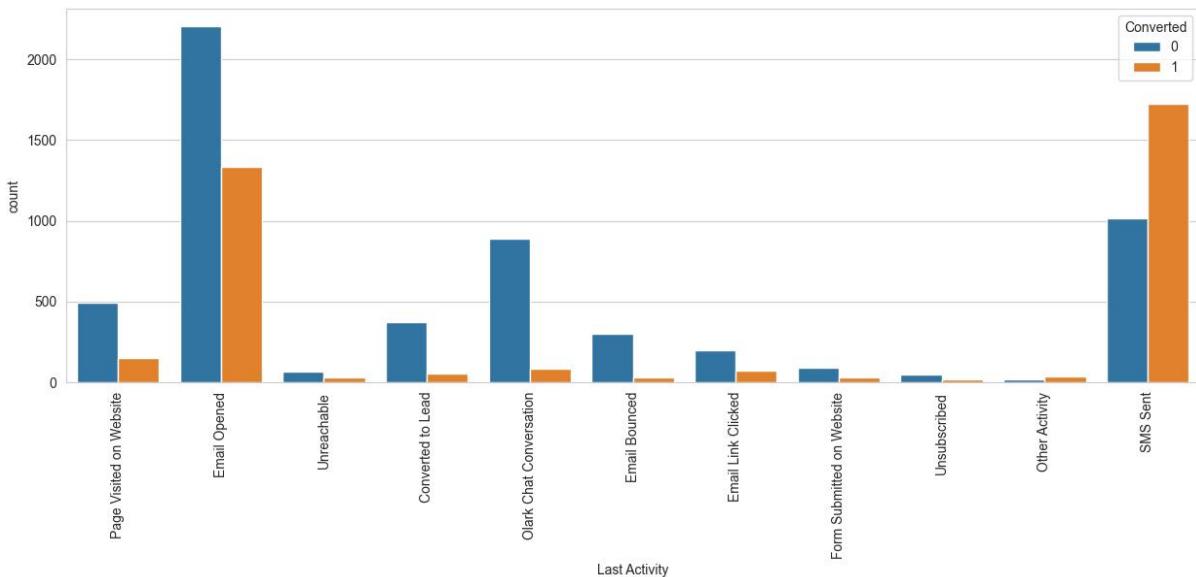
### Cột “Last Activity”



Hình 3.16 Giá trị của cột “Last Activity” và tỷ lệ chuyển đổi

Thông qua biểu đồ ta phân tích được ảnh hưởng của các hoạt động cuối cùng được thực hiện bởi khách hàng tới tỷ lệ chuyển đổi khách hàng. Cụ thể, ở hoạt động như Email Opened có tỷ lệ chuyển đổi khách hàng ổn định và số

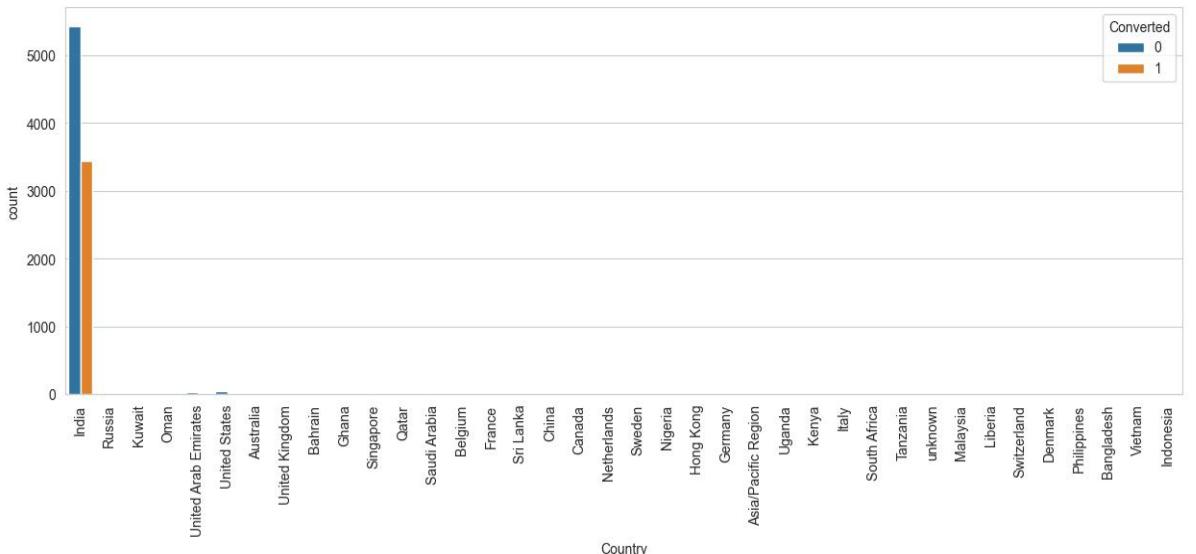
lượng khách hàng có nguồn gốc từ đây là cao nhất. Hoạt động cuối cùng là SMS Sent có tỷ lệ chuyển đổi cao nhưng số lượng khách hàng không cao. Các nguồn truy cập còn lại có số lượng không đáng kể.



Hình 3.17 Giá trị của cột “Last Activity” và tỷ lệ chuyển đổi sau khi gộp các giá trị

*Kết luận:* Một vài giá trị trong cột “Last Activity” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

### Cột “Country”

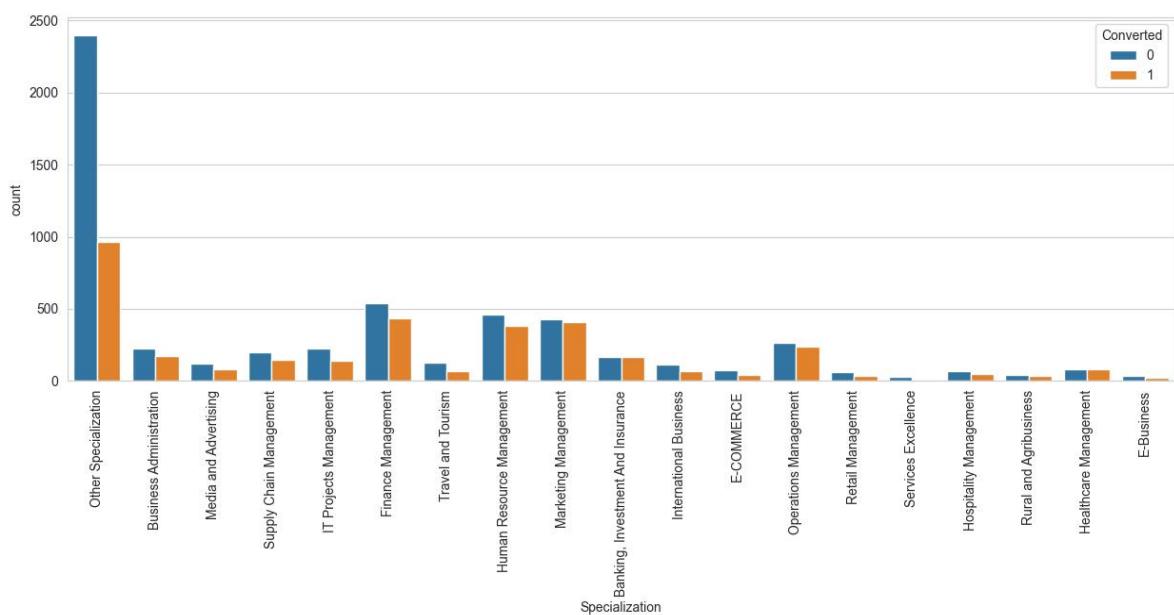


Hình 3.18 Giá trị của cột “Country” và tỷ lệ chuyển đổi.

Thông qua biểu đồ ta phân tích được ảnh hưởng của vị trí quốc gia tới tỷ lệ chuyển đổi khách hành. Cụ thể, khách hàng hầu như có nguồn gốc từ India và khách hàng chuyển đổi cũng hầu như là ở đây. Các quốc gia còn lại có số lượng rất ít không có tính so sánh.

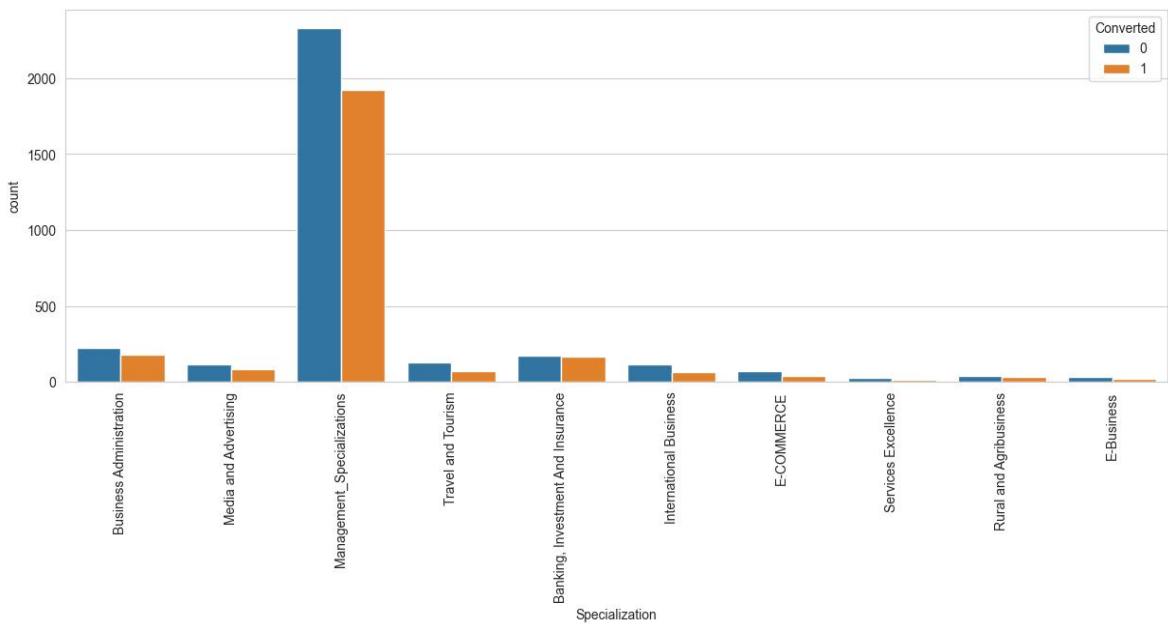
*Kết luận* Các giá trị trong cột “Country” không có tính phân loại cao để cung cấp thông tin so sánh. Do đó, cột này cần được loại bỏ để không làm ảnh hưởng tới kết quả phân tích.

### Cột “Specialization”



Hình 3.19 Giá trị của cột “Specialization” và tỷ lệ chuyển đổi.

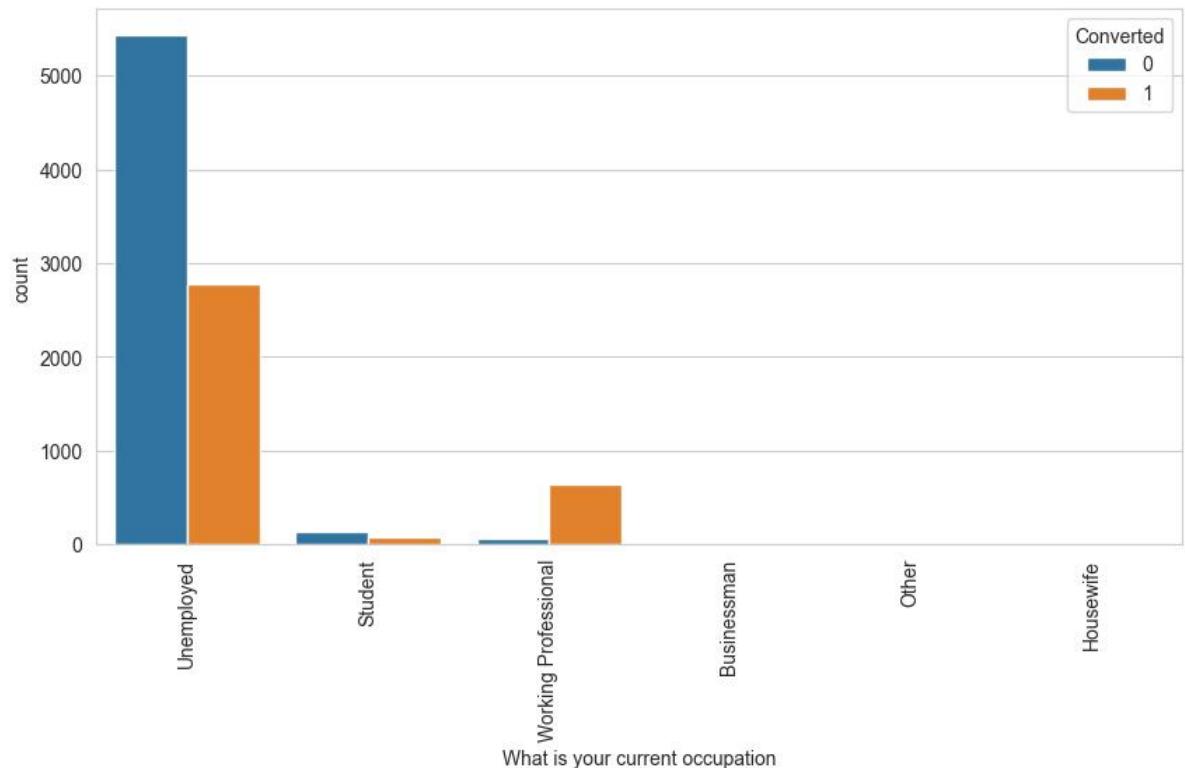
Thông qua biểu đồ ta phân tích được ảnh hưởng của lĩnh vực làm việc tới tỷ lệ chuyển đổi khách hành. Cụ thể, các nhóm khách hàng làm trong lĩnh vực quản lý (“Finance Management”, “Human Resource Management”, “Marketing Management”, “Operations Management”, “IT Projects Management”, “Supply Chain Management”, “Healthcare Management”, “Hospitality Management”, “Retail Management”) đều có tỷ lệ chuyển đổi khá cao nên sẽ được gộp thành một giá trị là Management Specializations để thuận tiện phân tích. Bên cạnh đó, tỷ lệ chuyển đổi của những khách hàng chưa rõ lĩnh vực cũng khá cao. Các lĩnh vực còn lại có số lượng khá ít.



Hình 3.20 Giá trị của cột “Specialization” và tỷ lệ chuyển đổi sau khi gộp các giá trị

**Kết luận:** Một vài giá trị trong cột “Specialization” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

#### Cột “What is your current occupation”

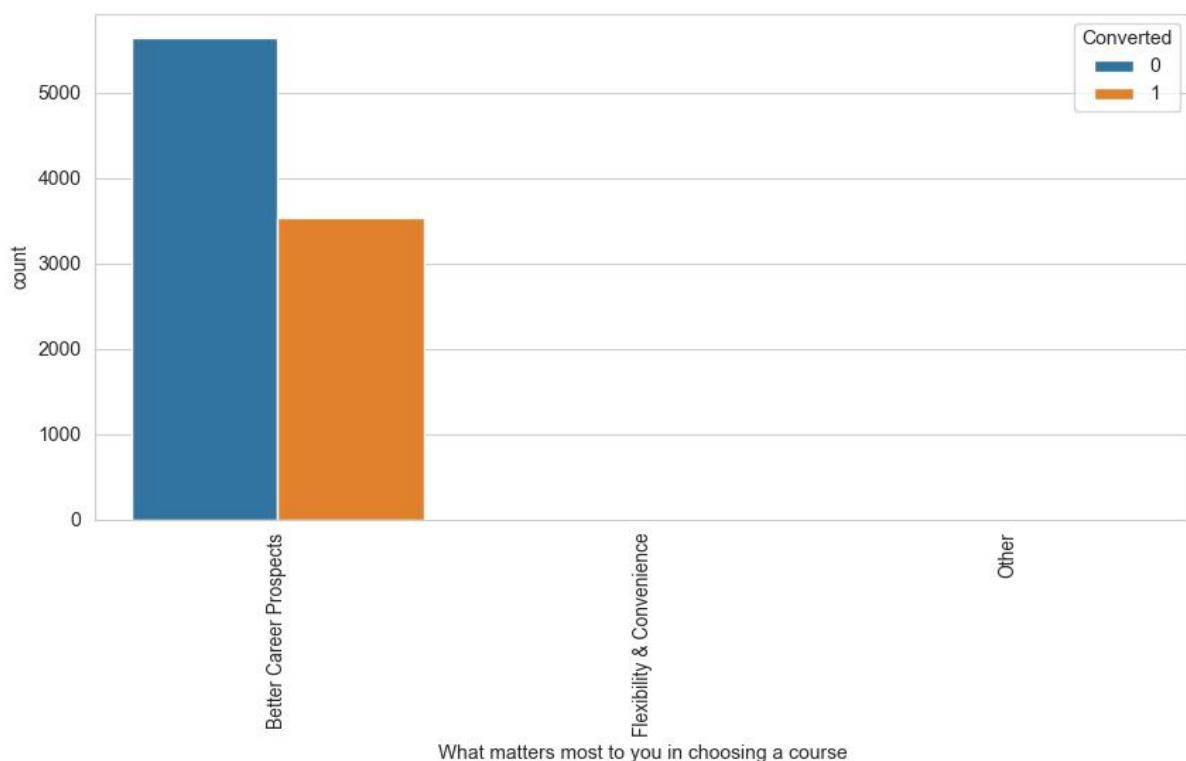


Hình 3.21 Giá trị của cột “What is your current occupation” và tỷ lệ chuyển đổi.

Thông qua biểu đồ ta phân tích được ảnh hưởng tình trạng việc làm tới tỷ lệ chuyển đổi khách hành. Cụ thể, các nhóm “Unemployed” đều có tỷ lệ chuyển đổi khá cao và có số lượng khách hàng cao nhất. Nhóm “Working Professional” có tỷ lệ chuyển đổi cao nhưng số lượng khách hàng không cao lắm. Các nhóm còn lại có số lượng không đáng kể.

*Kết luận:* Một vài giá trị trong cột “What is your current occupation” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

#### Cột “*What matters most to you in choosing a course*”

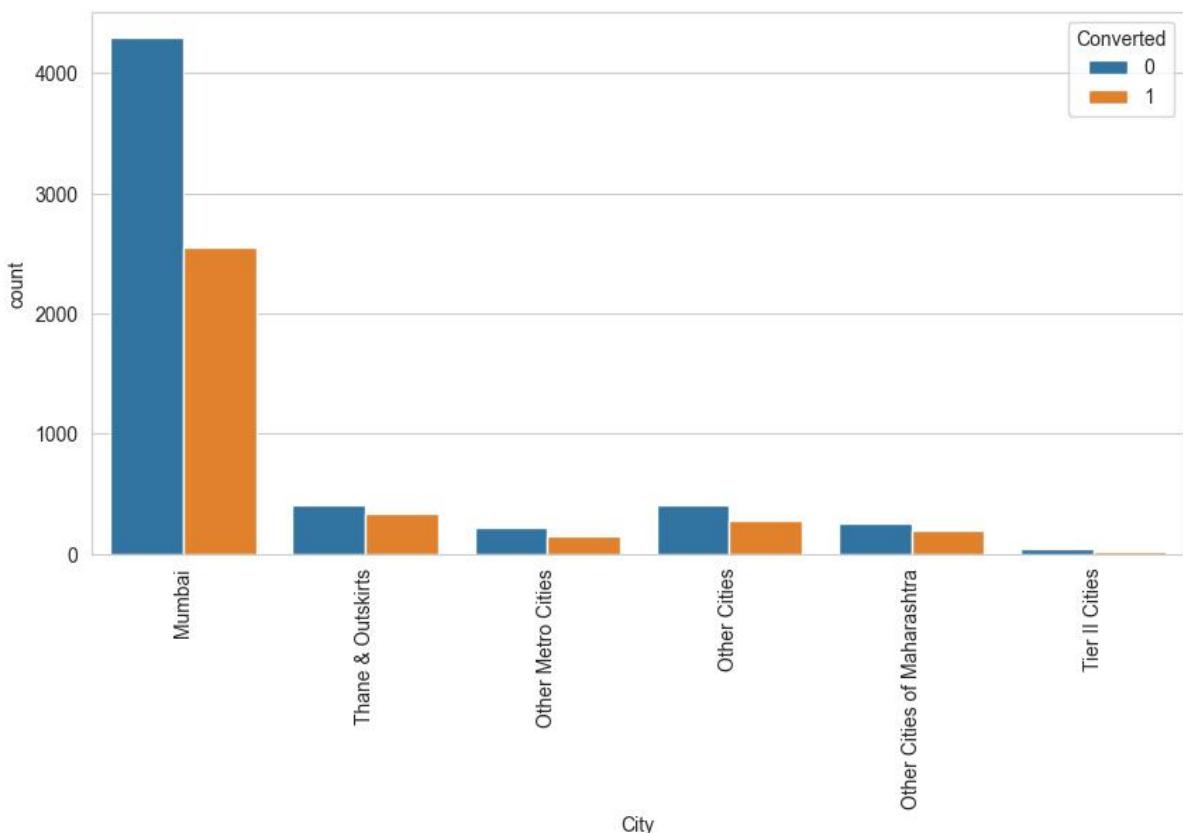


*Hình 3.22 Giá trị của cột “*What matters most to you in choosing a course*” và tỷ lệ chuyển đổi.*

Thông qua biểu đồ ta phân tích được ảnh hưởng của lý do chọn khóa học của khách hàng tới tỷ lệ chuyển đổi khách hàng. Cụ thể, khách hàng hầu như có lý do là “Better Career Prospect” và khách hàng chuyển đổi cũng hầu như là ở đây. Các lý do còn lại có số lượng rất ít không có tính so sánh.

*Kết luận:* Các giá trị trong cột “What matters most to you in choosing a course” bị lệch, không có tính phân loại cao để cung cấp thông tin so sánh. Do đó, cột này cần được loại bỏ để không làm ảnh hưởng tới kết quả phân tích.

### Cột “City”

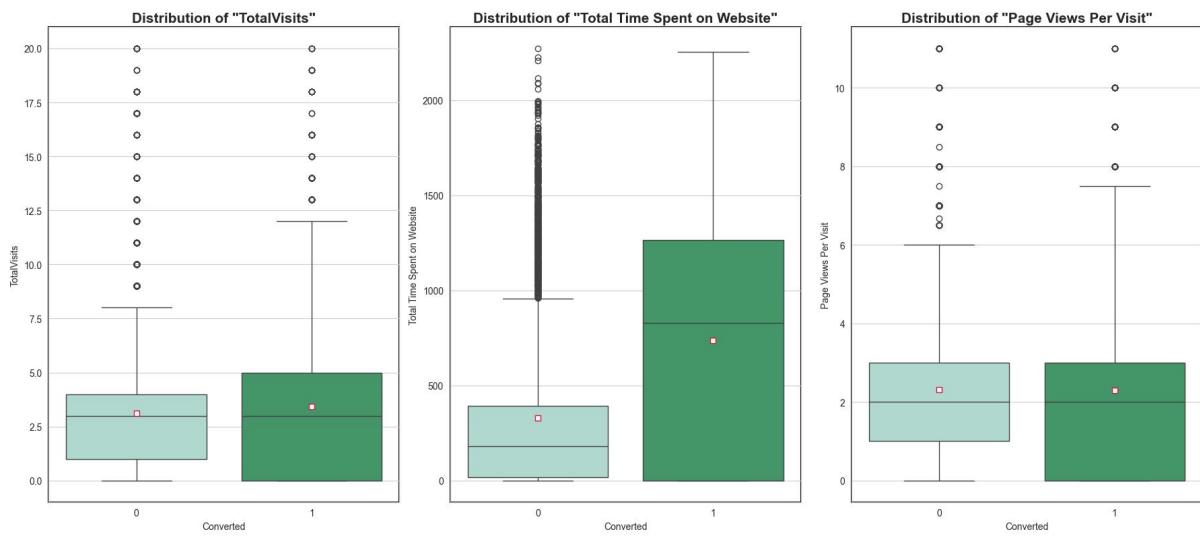


Hình 3.23 Giá trị của cột “City” và tỷ lệ chuyển đổi.

Thông qua biểu đồ ta phân tích được ảnh hưởng của vị trí thành phố tới tỷ lệ chuyển đổi khách hành. Cụ thể, Mumbai có tỷ lệ chuyển đổi khách hàng ổn định và số lượng khách hàng có nguồn gốc từ đây là cao nhất. Các nguồn truy cập còn lại có tỷ lệ chuyển đổi cao nhưng số lượng khách hàng không cao lắm.

*Kết luận:* Một vài giá trị trong cột “City” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

### Cột “Total Visits”, “Total time spent on website” và “Page views per visit”



*Hình 3.24 Giá trị của cột “Total Visits”, “Total time spent on website” và “Page views per visit” và tỷ lệ chuyển đổi.*

Thông qua biểu đồ ta phân tích được ảnh hưởng các giá trị tới tỷ lệ chuyển đổi khách hành. Đối với Total Visits và Page Views Per Visit giá trị trung bình cho khách hàng tiềm năng đã chuyển đổi và không chuyển đổi là như nhau. Cột Total time spent on website, khách hàng dành nhiều thời gian hơn trên web có nhiều khả năng được chuyển đổi hơn.

*Kết luận:* Cột “Total time spent on website” có ảnh hưởng với tỷ lệ chuyển đổi. Do đó, cột này cung cấp thông tin giá trị cho việc dự đoán kết quả chuyển đổi.

### 3.3. Chuẩn bị dữ liệu

Sau khi thực hiện phân tích khám phá dữ liệu cần thực hiện các bước chuẩn bị dữ liệu để có thể thực hiện mô hình hồi quy logistic.

#### 3.3.1. Chuyển đổi biến sang dạng nhị phân:

Các cột “Do Not Email”, “A free copy of Mastering The Interview” có giá trị Yes/ No nên cần chuyển về dạng biến nhị phân 0/1 để thực hiện phân tích mô hình.

#%% - Chuyển biến nhị phân

```
varlist = ['Do Not Email','A free copy of Mastering The Interview']
```

```
def binary_map(x):
```

```
return x.map({'Yes': 1, "No": 0})  
  
df[varlist] = df[varlist].apply(binary_map)
```

### 3.3.2. Chuyển đổi biến nhiều cấp độ (dummies):

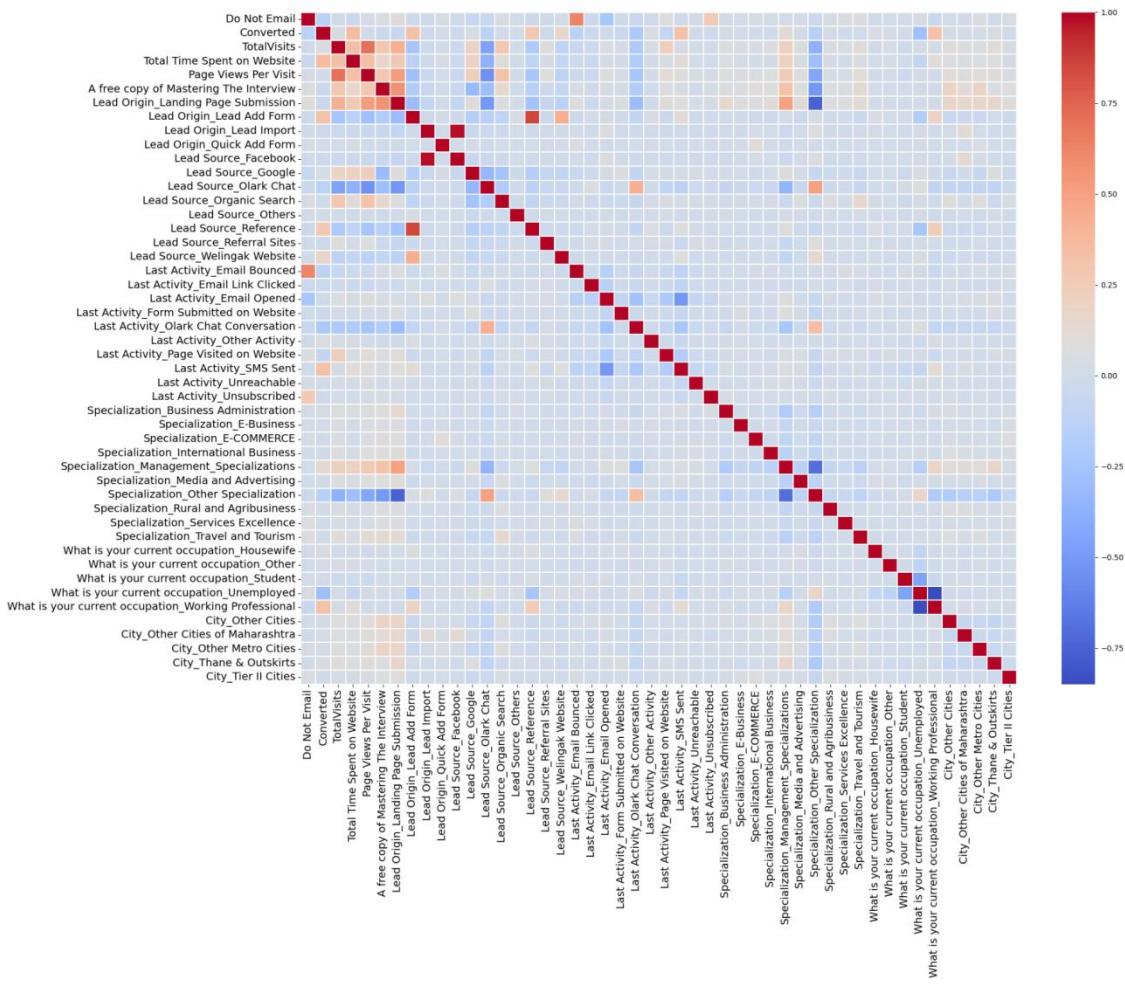
Đối với các biến phân loại cần chuyển đổi thành các giá trị số, bằng cách tạo các biến giả thực hiện theo cách này nhằm để trực quan hóa sự tương quan của các biến phân loại với mục tiêu.

#%% - Chuyển biến phân loại

```
dummy = pd.get_dummies(df[['Lead Origin', 'Lead Source', 'Last Activity',  
'Specialization','What is your current occupation',  
'City']], drop_first=True,dtype=int)  
  
df.drop(['Lead Origin', 'Lead Source', 'Last Activity', 'Specialization','What is  
your current occupation',  
'City'], axis=1, inplace = True)  
  
df = pd.concat([df, dummy], axis=1)
```

### 3.3.3. Xác định tương quan giữa các biến

Sau khi chuyển đổi tất cả các cột được chuyển đổi thành dữ liệu số, tiếp đến cần xác định mối tương quan giữa các biến dựa vào biểu đồ nhiệt.



Hình 3.25 Biểu đồ Heat map thể hiện ương quan giữa các biến

Mối tương quan các biến được hiển thị ở trên bằng bản đồ nhiệt, có sự tương quan cao giữa một số biến, dữ liệu có hiện tượng đa cộng tuyến. Xem xét các biến có mối tương quan lẫn nhau cao.

Cột 1	Cột 2	Tương quan
Lead Origin_Lead Import	Lead Source_Facebook	0,981708
Lead Origin_Lead Add Form	Lead Source_Reference	0,852955
TotalVisits	Page Views Per Visit	0,703643
Do Not Email	Last Activity_Email Bounced	0,625356
A free copy of Mastering The	Lead Origin_Landing Page	0,568449

Interview	Submission	
Page Views Per Visit	Lead Origin_Landing Page Submission	0,520416
Lead Source_Olark Chat	Specialization_Other Specialization	0,499642
Lead Origin_Landing Page Submission	Specialization_Management_Specializations	0,498047
Lead Origin_Lead Add Form	Lead Source_Welingak Website	0,430615
Lead Source_Olark Chat	Last Activity_Olark Chat Conversation	0,425906

Bảng 3.10 Thống kê các biến có độ tương quan cao

Loại bỏ các biến có đa cộng tuyến cao, và xác định được 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.

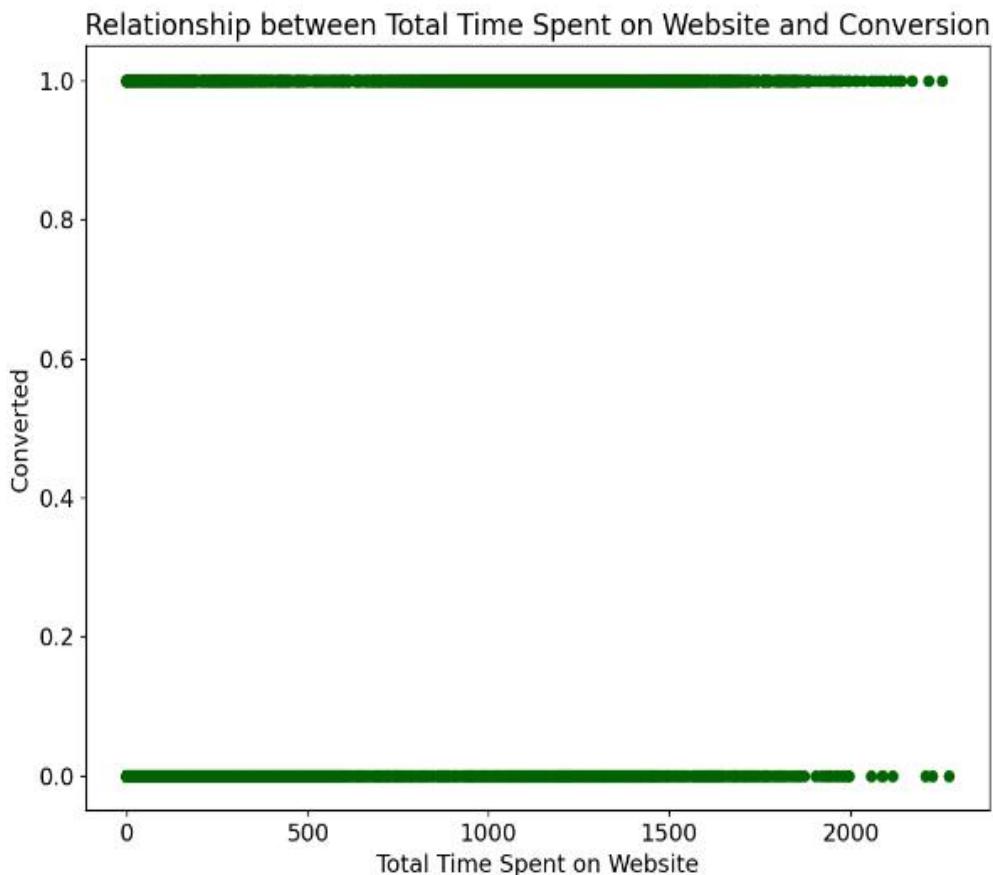
### 3.3.4. Chia dữ liệu huấn luyện (train) và kiểm chứng (test)

Nhóm thực hiện phân chia phân tầng để ngăn chặn ảnh hưởng của sự mất cân bằng lớp của biến mục tiêu, do đó ngăn ngừa sai lệch. Tiếp đến thực hiện chuẩn hóa bằng cách Scale dữ liệu các đặc trưng số trong dữ liệu để đảm bảo rằng chúng có cùng phạm vi giá trị. Quá trình này giúp cải thiện hiệu suất của mô hình học máy.

## 3.4. Xây dựng mô hình hồi quy Logistic

### 3.4.1. Đơn biến

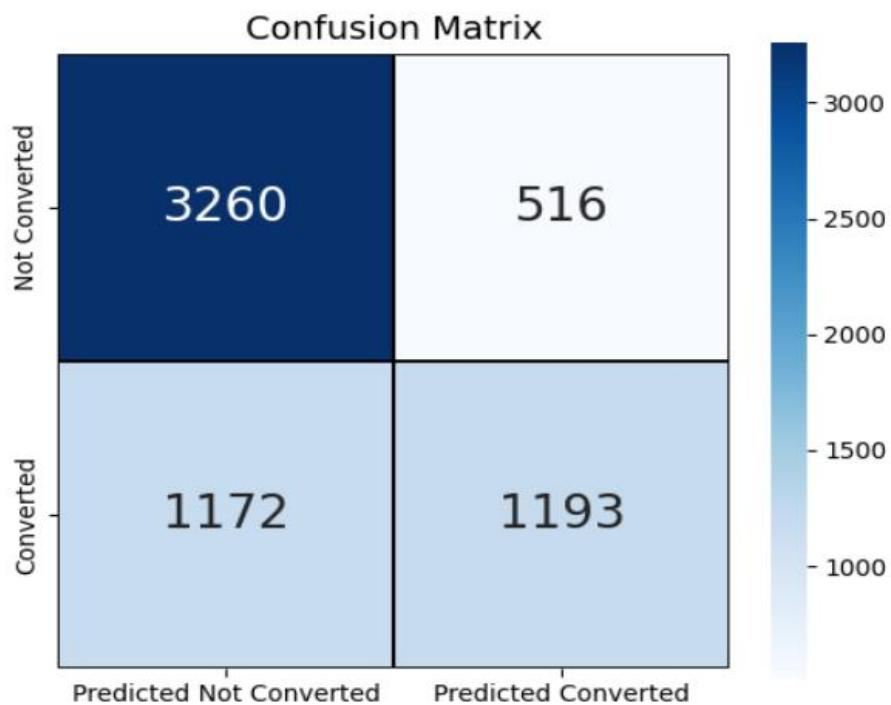
#### 3.4.1.1. Biểu đồ phân tán và xây dựng mô hình



Hình 3.26 Mối quan hệ giữa “Total Time Spent on Website” và “Conversion”

Từ biểu đồ trên có thể đưa ra nhận định, nếu số thời gian trên trang web cao hơn 1800 thì khả năng khách hàng chuyển đổi cao hơn. Tuy nhiên khó đánh giá khoảng thời gian khách hàng có thể chuyển đổi, thực hiện xây dựng mô hình hồi quy logistic.

### 3.4.1.2. Đánh giá mô hình với dữ liệu tập huấn

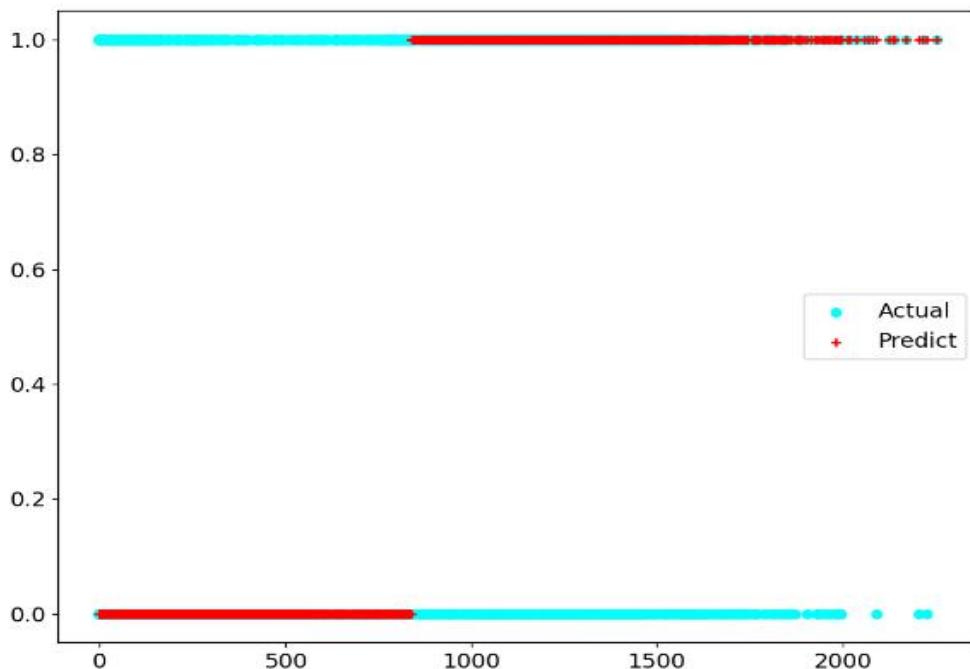


*Hình 3.27 Biểu đồ ma trận thống kê giá trị thực tế và giá trị dự đoán của mô hình đơn biến trên tập huấn luyện*

Chỉ số	Giá trị
Model Accuracy value	72.51 %
Model Sensitivity value	50.44 %
Model Specificity value	86.33 %
Model Precision value	69.81 %
Model Recall value	50.44 %
Model True Positive Rate (TPR)	50.44 %
Model False Positive Rate (FPR)	13.67 %

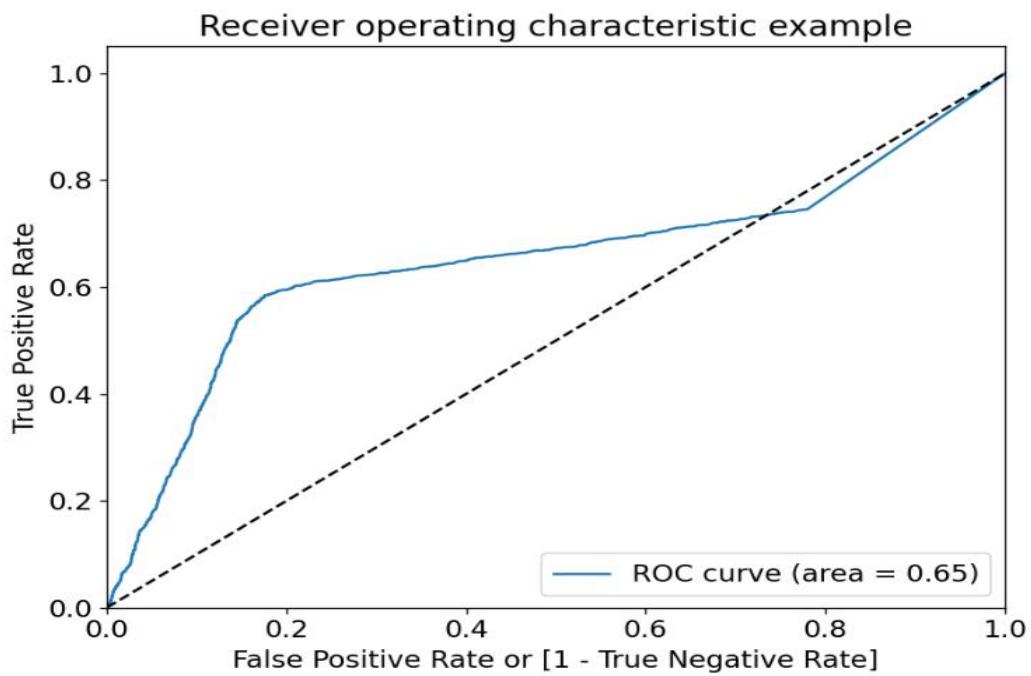
Model Poitive Prediction Value	69.81 %
Model Negative Prediction value	73.56 %

Bảng 3.11 Hiệu suất của mô hình phân loại



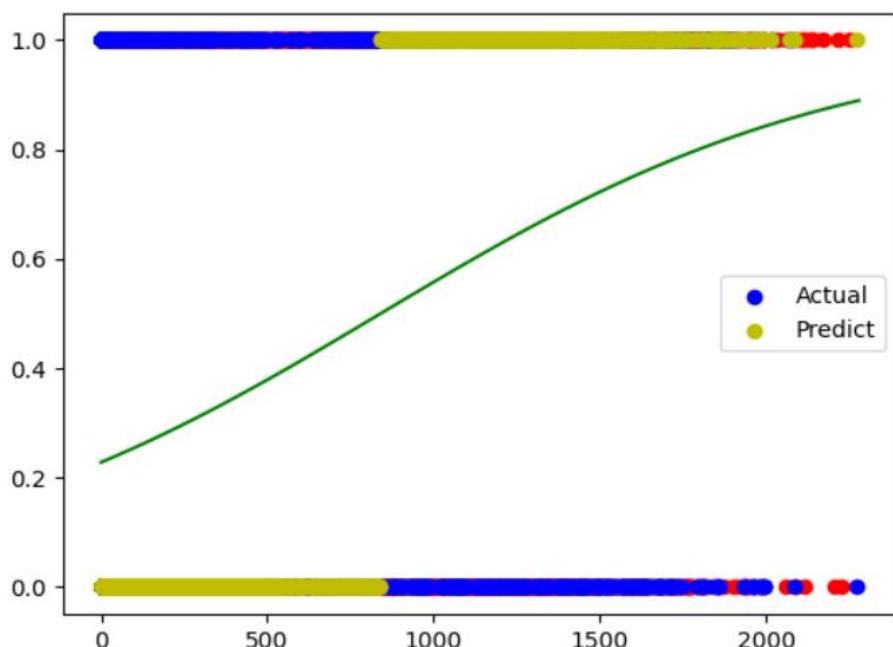
Hình 3.28 Kết quả phân loại của mô hình trên tập huấn luyện

Dựa vào hình có thể thấy mô hình chưa thực sự phân loại tốt, cũng giống như trong thực tế 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 để dự đoán được.



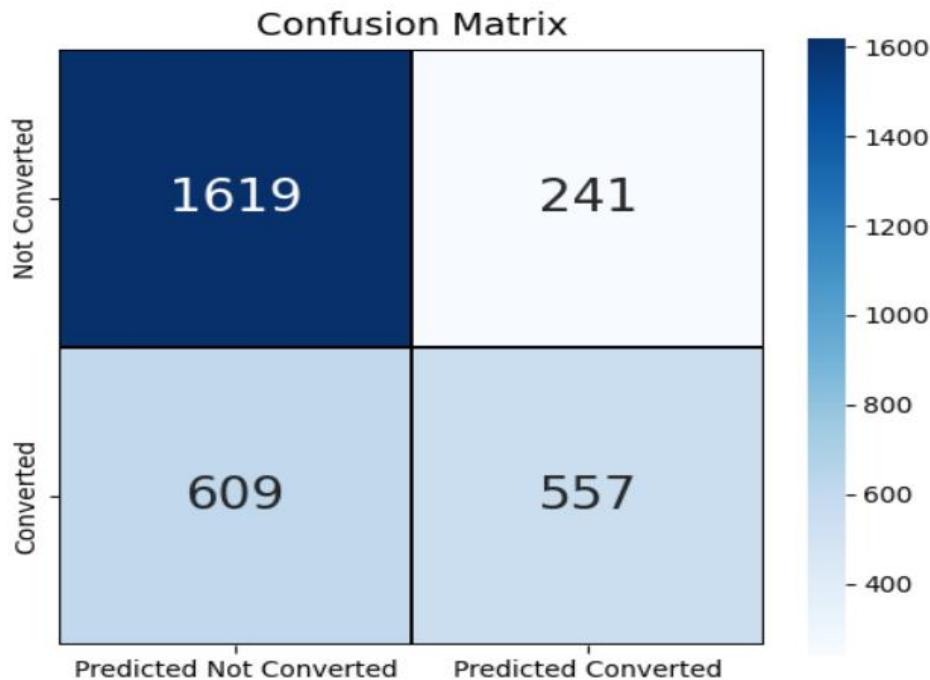
*Hình 3.29 Đường cong ROC*

Chỉ số diện tích dưới đường cong ROC (AUC - Area Under Curve) bằng 0,65 chỉ ra rằng mô hình có hiệu suất tương đối tốt hơn so với dự đoán ngẫu nhiên, nhưng nó vẫn còn một khoảng cách xa so với một mô hình lý tưởng ( $AUC = 1$ ) hoặc một mô hình hoàn hảo. Mô hình được cải thiện để đạt được hiệu suất tốt hơn.



*Hình 3.30 Trục quan hàm Sigmoid*

### 3.4.1.3. Đánh giá mô hình dựa trên dữ liệu kiểm thử



Hình 3.31 Biểu đồ ma trận thống kê giá trị thực tế và giá trị dự đoán của mô hình đơn biến trên tập huấn luyện

Trên cả tập huấn luyện và tập kiểm tra:

- *Độ chính xác (Accuracy)*: độ chính xác của mô hình khá gần nhau, lần lượt là 72.51% trên tập huấn luyện và 71.91% trên tập kiểm tra. Mô hình không gặp phải tình trạng quá mức khớp (overfitting) hoặc không đủ dữ liệu để học.
- *Độ nhạy (Sensitivity/Recall/True Positive Rate)*: độ nhạy của mô hình không cao, lần lượt là 50.44% trên tập huấn luyện và 47.77% trên tập kiểm tra có xu hướng bỏ sót một phần các trường hợp dương thực sự.
- *Độ đặc hiệu (Specificity)*: lần lượt là 86.33% trên tập huấn luyện và 87.04% trên tập kiểm tra mô hình hiệu quả trong việc loại bỏ các trường hợp âm.
- *Độ chính xác dự đoán dương (Precision/Positive Prediction Value)*: độ chính xác dự đoán dương của mô hình khá gần nhau và cao, lần lượt là 69.81% trên tập huấn luyện và 69.8% trên tập kiểm tra có khả năng đưa ra các dự đoán dương chính xác.

### 3.4.1.4. Kết luận

[ -1.21915254 ]	index	Variable	Coefficient
0	Total Time Spent on Website		0.00145

**Phương trình**

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

$$\approx -1.219 + 0.00145 * (\text{Total Time Spent On Website})$$

Trong trường hợp này:

- **Hệ số chặn của mô hình là -1.219** là logit của một khách hàng không dành thời gian trên trang web (Total Time Spent On Website = 0), xác suất chuyển đổi sẽ bằng -1,219. Hay có nghĩa là, giá trị odds của một khách hàng sẽ chuyển đổi khi thời gian dành trên trang web là bằng 0 là  $\exp(-1.219) = 0.2955$  tức gần 30% so với xác suất không chuyển đổi là  $1 - 0.2955 = 0.705$  tức gần 70%.
- **Hệ số dốc của mô hình là 0.00145** mỗi đơn vị tăng trong thời gian đã dành trên trang web sẽ làm tăng xác suất của sự kiện xảy ra lên 0.00145 lần.

Ví dụ:

+ Khi Total Time Spent On Website = 1000 thì logit cho việc chuyển đổi là

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

$$\approx -1.219 + 0.00145 * 1000$$

+ Khi Total Time Spent On Website = 1001 thì logit cho việc chuyển đổi là

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

$$\approx -1.219 + 0.00145 * 1001$$

⇒ Ứng với việc gia tăng 1 đơn vị tổng thời gian thì chuyển đổi sẽ thay đổi một lượng:  $0.23245 - 0.231 = 0.00145$

Hoặc có thể diễn giải thành dạng tỉ lệ odds có nghĩa khi tăng một đơn vị trong tổng thời gian trên web sẽ làm tăng khoảng 0.14%.

## Dự đoán

Thực hiện dự đoán với thời gian khách hàng ở website là 1000, ta có:

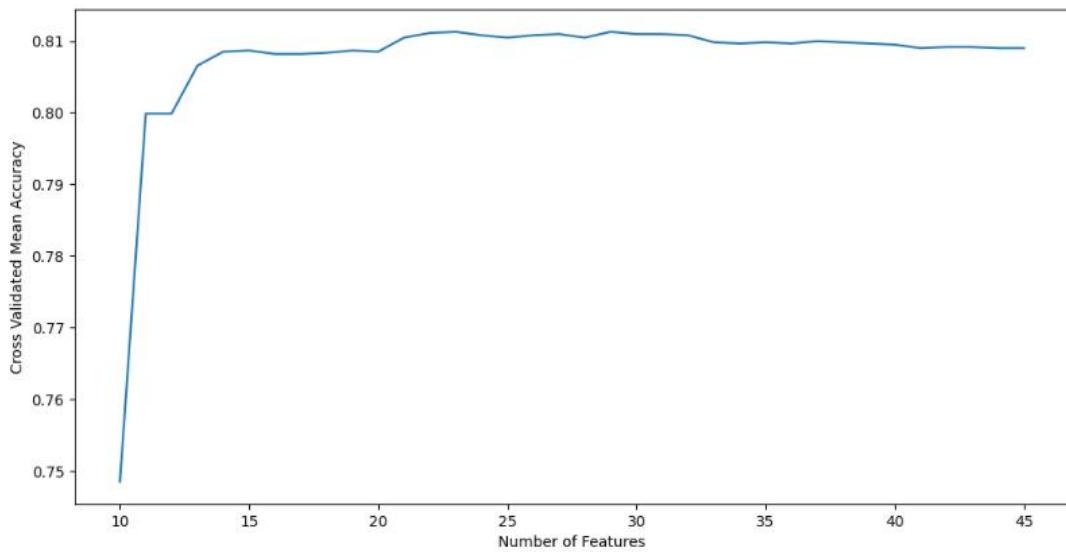
$$\log(P(Y=1)) = \log\left(\frac{P(Y=1)}{1 - P(Y=1)}\right)$$
$$\approx -1.219 + 0.00145 * (1000) \approx 0.23$$
$$\Rightarrow p = \frac{e^{\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_n \cdot X_n}}{1 + e^{\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_n \cdot X_n}} = \frac{1}{1 + e^{-0.23}} \approx 0.56$$

Predicted probability of conversion: 0.5574346526028935

### 3.4.2. Đa biến

#### 3.4.2.1. Lựa chọn biến độc lập

Dựa trên tập dữ liệu hiện có rất nhiều biến độc lập, vấn đề đặt ra là cần chọn lọc ra những biến có ảnh hưởng tốt đến mô hình và những biến nào cần loại bỏ. Lựa chọn thuộc tính bằng phương pháp Recursive feature elimination (RFE) và Cross Validation (Xác thực chéo). Bằng cách lặp qua số lượng tính năng, và đưa ra mô hình nào cho kết quả cao nhất.



Hình 3.32 Giá trị Accuracr kiểm chéo với số lượng biến tăng dần

Số lượng biến tối ưu để sử dụng là 29 với độ chính xác accuracy = 0,8112689270957294. Xây dựng mô hình lần lượt các mô hình bằng Statsmodel và giảm biến thủ công bằng các chỉ số đánh giá p-value < 0,05 và VIF < 5.

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:	Sun, 07 Apr 2024	Deviance:	5079.6						
Time:	22:49:54	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			
Last Activity_Email Link Clicked		0.6918	0.265	2.610	0.009	0.172 1.211			
Last Activity_Email Opened		1.1036	0.171	6.470	0.000	0.769 1.438			

Hình 3.33 Chạy mô hình với 29 biến

Chạy mô hình 1 với 29 biến, biến “What is your current occupation\_Housewife” có giá trị p-value = 0,999 cần được loại bỏ khỏi mô hình. Sau khi loại bỏ cần tiếp tục thực hiện xây dựng mô hình tiếp tục.

Generalized Linear Model Regression Results						
Dep. Variable:	Converted	No. Observations:	6141			
Model:	GLM	Df Residuals:	6121			
Model Family:	Binomial	Df Model:	19			
Link Function:	Logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-2548.1			
Date:	Sun, 07 Apr 2024	Deviance:	5096.3			
Time:	22:56:37	Pearson chi2:	9.17e+03			
No. Iterations:	7	Pseudo R-squ. (CS):	0.3954			
Covariance Type:	nonrobust					
		coef	std err	z	P> z	[0.025 0.975]
const		1.1971	0.241	4.965	0.000	0.725 1.670
Do Not Email		-1.1682	0.192	-6.087	0.000	-1.544 -0.792
TotalVisits		0.2479	0.049	5.075	0.000	0.152 0.344
Total Time Spent on Website		1.0522	0.041	25.851	0.000	0.972 1.132
Page Views Per Visit		-0.2203	0.055	-4.016	0.000	-0.328 -0.113
A free copy of Mastering The Interview		-0.3132	0.090	-3.480	0.001	-0.490 -0.137
Lead Origin_Landing Page Submission		-0.9525	0.134	-7.096	0.000	-1.216 -0.689
Lead Origin_Lead Add Form		3.1823	0.233	13.664	0.000	2.726 3.639
Lead Source_Olark Chat		0.8974	0.135	6.669	0.000	0.634 1.161
Lead Source_Welingak Website		2.4503	0.762	3.214	0.001	0.956 3.944
Last Activity_Email Link Clicked		0.8721	0.237	3.683	0.000	0.408 1.336
Last Activity_Email Opened		1.2879	0.124	10.394	0.000	1.045 1.531
Last Activity_Other Activity		2.3497	0.456	5.155	0.000	1.456 3.243
Last Activity_Page Visited on Website		0.6445	0.184	3.503	0.000	0.284 1.005

Hình 3.34 Chạy mô hình với 19 biến

		Features	VIF
18	What is your current occupation_Unemployed		9.00
5	Lead Origin_Landing Page Submission		6.05
16	Specialization_Other Specialization		3.91
10	Last Activity_Email Opened		2.87
3	Page Views Per Visit		2.61
7	Lead Source_Olark Chat		2.44
13	Last Activity_SMS Sent		2.37
1	TotalVisits		2.23
4	A free copy of Mastering The Interview		2.22
6	Lead Origin_Lead Add Form		1.84
12	Last Activity_Page Visited on Website		1.40
0	Do Not Email		1.28
8	Lead Source_Welingak Website		1.28
2	Total Time Spent on Website		1.27
17	What is your current occupation_Student		1.23
9	Last Activity_Email Link Clicked		1.14
15	Last Activity_Unsubscribed		1.10
14	Last Activity_Unreachable		1.05
11	Last Activity_Other Activity		1.04

Hình 3.35 Kết quả kiểm tra VIF lần 1

Đối với lần chạy mô hình thứ 11 các giá trị p-value đều thấp hơn 0,05 với tất cả 19 biến. Tuy nhiên chỉ số VIF của biến “What is your current occupation\_Unemployed” cao trên nên cần loại bỏ. Tiếp tục thực hiện xây dựng mô hình loại bỏ các biến không hợp lệ.

Generalized Linear Model Regression Results						
Dep. Variable:	Converted	No. Observations:	6141			
Model:	GLM	Df Residuals:	6123			
Model Family:	Binomial	Df Model:	17			
Link Function:	Logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-2681.9			
Date:	Sun, 07 Apr 2024	Deviance:	5363.9			
Time:	23:03:45	Pearson chi2:	6.71e+03			
No. Iterations:	7	Pseudo R-squ. (CS):	0.3684			
Covariance Type:	nonrobust					
	coef	std err	z	P> z	[0.025	0.975]
const	-1.0839	0.155	-6.997	0.000	-1.388	-0.780
Do Not Email	-1.2339	0.185	-6.655	0.000	-1.597	-0.871
TotalVisits	0.2320	0.048	4.880	0.000	0.139	0.325
Total Time Spent on Website	1.0702	0.040	26.896	0.000	0.992	1.148
Page Views Per Visit	-0.2135	0.053	-4.041	0.000	-0.317	-0.110
A free copy of Mastering The Interview	-0.3231	0.087	-3.730	0.000	-0.493	-0.153
Lead Origin_Landing Page Submission	-1.1051	0.130	-8.522	0.000	-1.359	-0.851
Lead Origin_Lead Add Form	3.2640	0.229	14.266	0.000	2.816	3.712
Lead Source_Olark Chat	0.8911	0.132	6.759	0.000	0.633	1.150
Lead Source_Welingak Website	2.3402	0.763	3.069	0.002	0.846	3.835
Last Activity_Email Link Clicked	0.8306	0.230	3.607	0.000	0.379	1.282
Last Activity_Email Opened	1.2679	0.121	10.505	0.000	1.031	1.504
Last Activity_Other Activity	2.4095	0.458	5.256	0.000	1.511	3.308

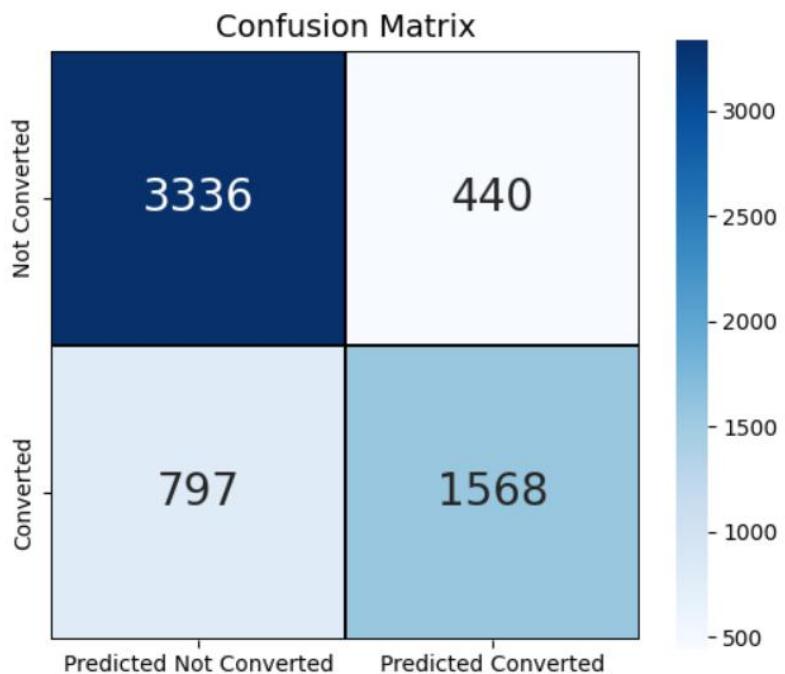
*Hình 3.36 Chạy mô hình với 17 biến*

	Features	VIF
5	Lead Origin_Landing Page Submission	4.05
3	Page Views Per Visit	2.61
10	Last Activity_Email Opened	2.55
16	Specialization_Other Specialization	2.34
7	Lead Source_Olark Chat	2.33
1	TotalVisits	2.23
13	Last Activity_SMS Sent	2.21
4	A free copy of Mastering The Interview	2.20
6	Lead Origin_Lead Add Form	1.81
12	Last Activity_Page Visited on Website	1.35
8	Lead Source_Welingak Website	1.28
2	Total Time Spent on Website	1.27
0	Do Not Email	1.26
9	Last Activity_Email Link Clicked	1.11
15	Last Activity_Unsubscribed	1.10
14	Last Activity_Unreachable	1.04
11	Last Activity_Other Activity	1.03

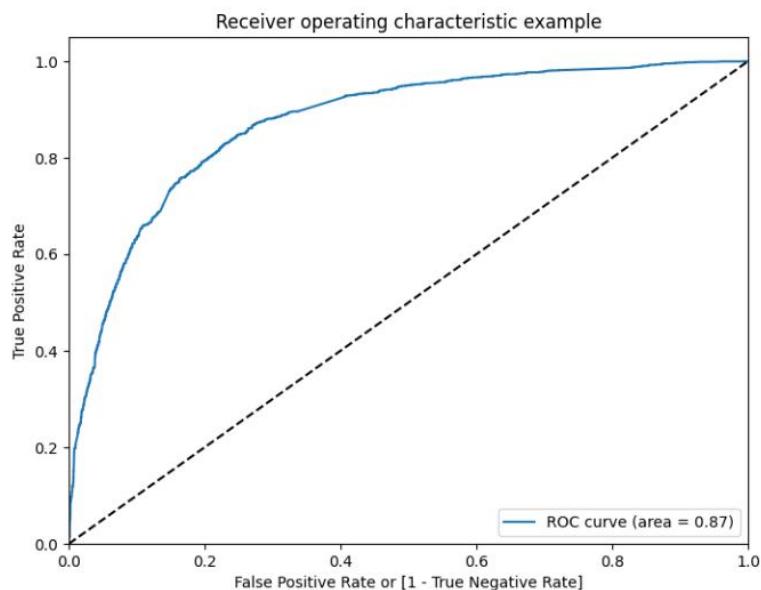
*Hình 3.37 Kết quả kiểm tra VIF lần 2*

Với lần chạy thứ 13, tất cả các biến đều có giá trị p-value và VIF đều phù hợp với yêu cầu. Vì vậy nhóm sử dụng 17 biến 'Do Not Email', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'A free copy of Mastering The Interview', 'Lead Origin\_Landing Page Submission', 'Lead Origin\_Lead Add Form', 'Lead Source\_Olark Chat', 'Lead Source\_Welingak Website', 'Last Activity\_Email Link Clicked', 'Last Activity\_Email Opened', 'Last Activity\_Other Activity', 'Last Activity\_Page Visited on Website', 'Last Activity\_SMS Sent', 'Last Activity\_Unreachable', 'Last Activity\_Unsubscribed', 'Specialization\_Other Specialization' để xây dựng mô hình hồi quy logistic và đánh giá mô hình trên tập dữ liệu huấn luyện và kiểm tra.

### 3.4.2.2. Đánh giá mô hình trên tập dữ liệu huấn luyện



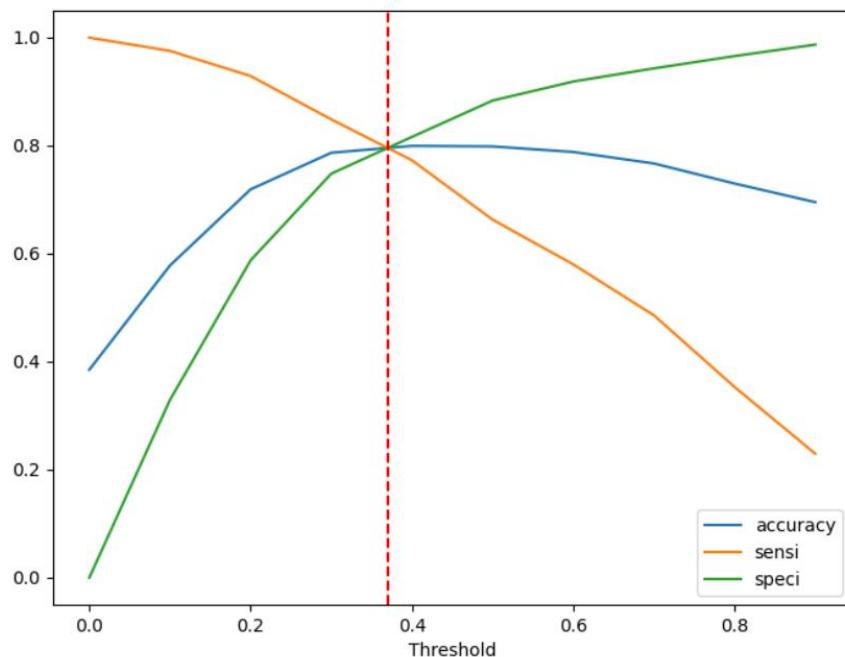
Hình 3.38 Biểu đồ ma trận thống kê giá trị thực tế và giá trị dự đoán của mô hình đơn biến trên tập huấn luyện



Hình 3.39 Đường cong ROC

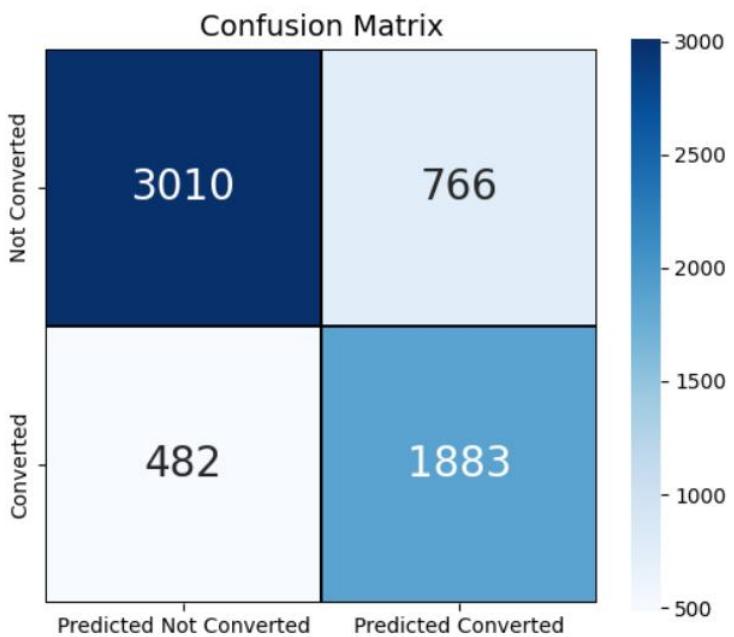
Đường cong ROC gần ranh giới bên trái hơn, cho thấy độ chính xác của mô hình khá cao. Diện tích dưới đường cong (AUC) phải gần với mức tối đa (tức là 1) và ở đây là 0,87 cho thấy một mô hình dự đoán tốt. Với điểm dự đoán  $> 0,5$  sẽ có “chuyển đổi” theo mặc định trong python chỉ là để kiểm tra

hiệu suất của mô hình một cách tổng quát chung. Trong trường hợp, quan tâm đến việc đảm bảo rằng mọi khách hàng chuyển đổi đều được phát hiện (sensitivity cao), nhưng đồng thời cũng muốn giảm thiểu số lượng khách hàng chuyển đổi phân loại nhầm là không chuyển đổi (specificity cao).



*Hình 3.40 Trade off giữa sensitivity và specificity*

Phù hợp với thực tế việc phát hiện chính xác càng nhiều tích cực (khách hàng tiềm năng chuyển đổi) càng tốt và do đó, cần tập trung nhiều vào độ nhạy. Đồng thời, họ cũng không muốn tốn nhiều nguồn lực cho những điều tiêu cực (dẫn đến những người sẽ không chuyển đổi). Dựa vào biểu đồ có thể thấy điểm cắt khoảng 0,37 sẽ tốt nhất.

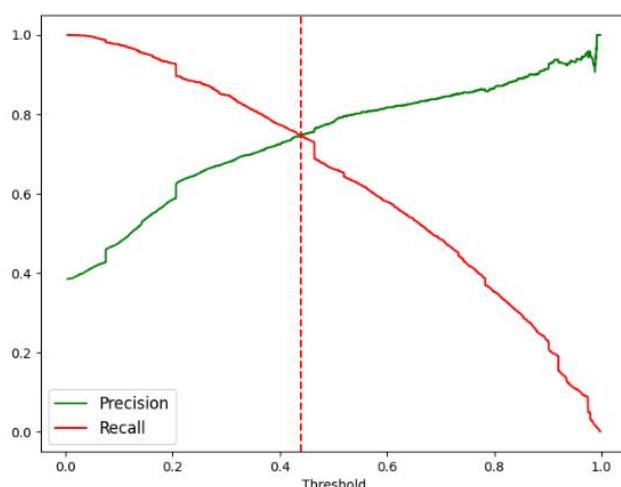


*Hình 3.41 Ma trận hỗn loạn ở điểm cắt 0.37*

Chỉ số	Giá trị
Model Accuracy value	79.68 %
Model Sensitivity value	79.62 %
Model Specificity value	79.71 %
Model Precision value	71.08 %
Model Recall value	79.62 %
Model True Positive Rate (TPR)	79.62 %
Model False Positive Rate (FPR)	20.29 %
Model Poitive Prediction Value	71.08 %
Model Negative Prediction value	86.2 %

*Bảng 3.12 Các chỉ số đánh giá mô hình ở điểm cắt 0.37*

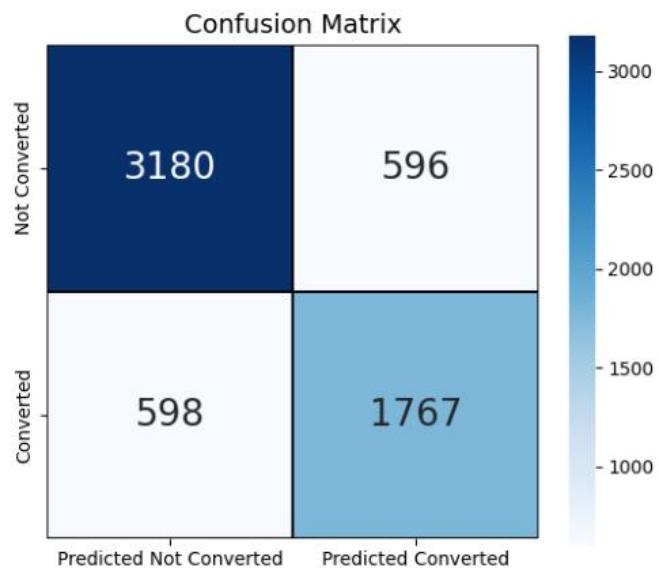
Trong trường hợp, quan tâm đến cả việc đảm bảo rằng khách hàng chuyển đổi được phát hiện (recall cao) và đồng thời đảm bảo rằng phần lớn các khách hàng không chuyển đổi thì thực sự là không chuyển đổi (precision cao). Nếu đặt ngưỡng quyết định cao, precision có thể tăng lên nhưng recall sẽ giảm, và ngược lại. Khi đánh giá đường cong này, có thể chọn ngưỡng quyết định phù hợp dựa trên mục tiêu cụ thể. Ví dụ, nếu muốn đảm bảo rằng hầu hết khách hàng chuyển đổi đều được phát hiện mà không cần quá lo lắng về việc phát hiện nhầm, có thể chọn một ngưỡng quyết định mà precision vẫn cao và recall là một giá trị chấp nhận được.



Hình 3.42 Trade off giữa Precision và Recall

Bằng cách sử dụng các điểm cắt biểu đồ đánh đổi Chính xác - Thu hồi (Precision- Recall Trade off curve) dựa vào biểu đồ ngưỡng cắt phù hợp là 0.44, 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



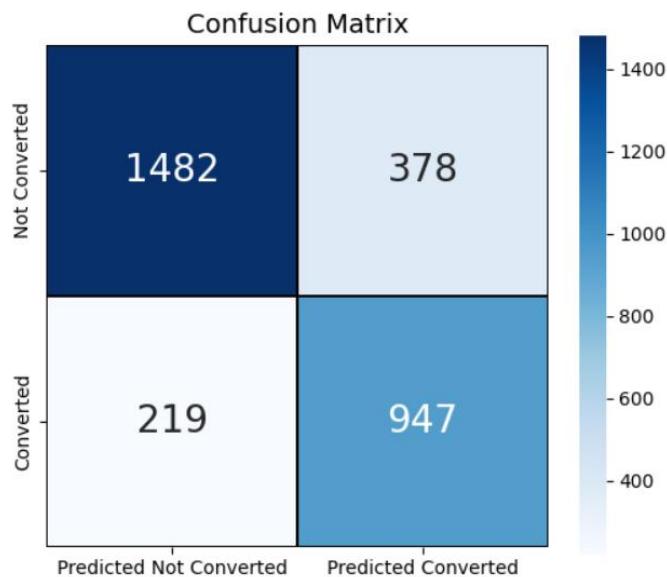
*Hình 3.43 Ma trận hỗn loạn ở điểm cắt 0.44*

Chỉ số	Giá trị
Model Accuracy value is	80.56 %
Model Sensitivity value is	74.71 %
Model Specificity value is	84.22 %
Model Precision value is	74.78 %
Model Recall value is	74.71 %
Model True Positive Rate (TPR)	74.71 %
Model False Positive Rate (FPR)	15.78 %
Model Poitive Prediction Value is	74.78 %
Model Negative Prediction value is	84.17 %

*Bảng 3.13 Các chỉ số đánh giá mô hình ở điểm cắt 0.44*

Giả định với mục đích xác định chính xác những người sẽ chuyển đổi thành khách hàng tiềm năng. Nên không thể sử dụng phương pháp đánh đổi Precision-Recall vì nó làm giảm True Positive. Vì vậy, nhóm sẽ sử dụng 0,37 làm điểm giới hạn.

### 3.4.2.3. Đánh giá mô hình với dữ liệu kiểm tra



Hình 3.44 Ma trận hỗn loạn với dữ liệu kiểm thử

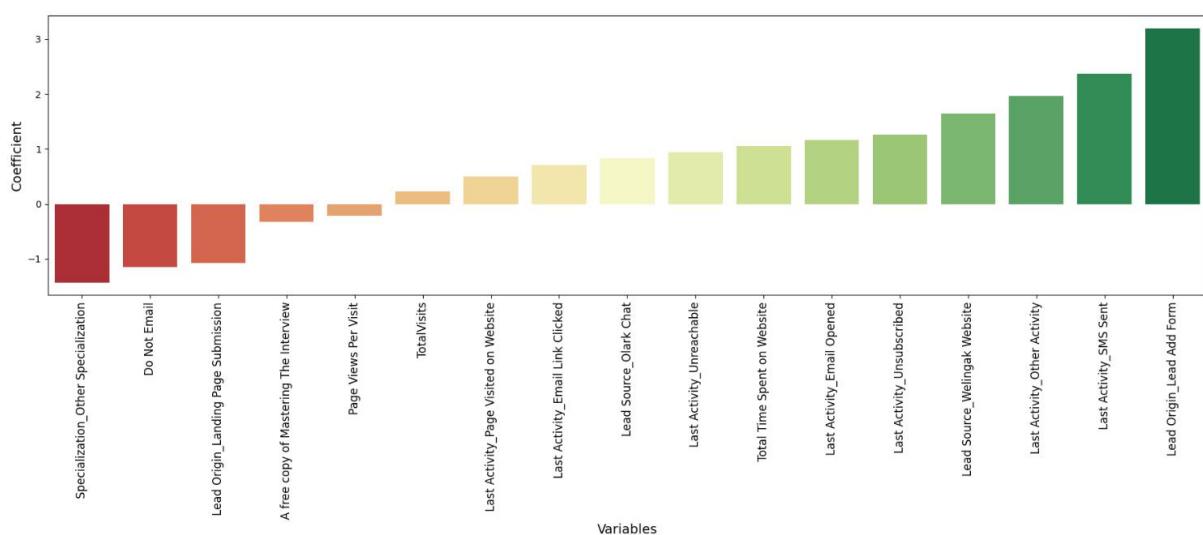
Chỉ số	Giá trị
Model Accuracy value	80.27%
Model Sensitivity value	81.22%
Model Specificity value	79.68 %
Model Precision value	71.47 %
Model Recall value	81.22 %
Model True Positive Rate (TPR)	81.22 %
Model False Positive Rate (FPR)	20.32 %

Model Poitive Prediction Value	71.47 %
Model Negative Prediction value	87.13 %

Bảng 3.14 Các chỉ số đánh giá mô hình với dữ liệu kiểm thử

Giá trị độ nhạy trên dữ liệu kiểm tra là 81,22% so với 79.62 % trong dữ liệu huấn luyện. Giá trị chính xác là 80.27 %. Cho thấy mô hình cũng hoạt động tốt trong tập dữ liệu thử nghiệm.

#### 3.4.2.4. Kết luận



Hình 3.45 Trục quan hệ số coef của các biến độc lập

Cột	Giá trị 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	0.949433
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	2.36917

Bảng 3.15 Thống kê giá trị coef của các biến độc lập

Điền giải các hệ số của biến, trong trường hợp các điều kiện khác không xảy ra (bằng 0) thì:

- Specialization\_Other Specialization (0.23): Khi khách hàng thuộc vào chuyên ngành không được phân loại cụ thể, tỷ lệ cơ hội của sự chuyển

đổi tăng thêm 0.23 lần so với khi khách hàng thuộc vào các chuyên ngành khác.

- Last Activity\_Page Visited on Website (0.5): Khi hoạt động cuối cùng của khách hàng là "Page Visited on Website", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 0.5 lần so với các hoạt động cuối cùng khác.
- Last Activity\_Email Link Clicked (0.71): Khi hoạt động cuối cùng của khách hàng là "Email Link Clicked", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 0.71 lần so với các hoạt động cuối cùng khác.
- Lead Source\_Olark Chat (0.84): Khi nguồn dẫn đến trang web là "Olark Chat", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 0.84 lần so với các nguồn dẫn khác.
- Last Activity\_Unreachable (0.94): Khi hoạt động cuối cùng của khách hàng là "Unreachable", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 0.94 lần so với các hoạt động cuối cùng khác.
- Total Time Spent on Website (1.06): Khi tổng thời gian trên trang web tăng thêm 1 đơn vị, tỷ lệ cơ hội của sự chuyển đổi tăng thêm 1.06 lần.
- Last Activity\_Email Opened (1.16): Khi hoạt động cuối cùng của khách hàng là "Email Opened", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 1.16 lần so với các hoạt động cuối cùng khác.
- Last Activity\_Unsubscribed (1.26): Khi hoạt động cuối cùng của khách hàng là "Unsubscribed", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 1.26 lần so với các hoạt động cuối cùng khác.
- Lead Source\_Welingak Website (1.65): Khi nguồn dẫn đến trang web là "Welingak Website", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 1.65 lần so với các nguồn dẫn khác.
- Last Activity\_Other Activity (1.96): Khi hoạt động cuối cùng của khách hàng là "Other Activity", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 1.96 lần so với các hoạt động cuối cùng khác.

- Last Activity\_SMS Sent (2.37): Khi hoạt động cuối cùng của khách hàng là "SMS Sent", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 2.37 lần so với các hoạt động cuối cùng khác.
- Lead Origin\_Lead Add Form (3.2): Khi nguồn dẫn đến trang web là "Lead Add Form", tỷ lệ cơ hội của sự chuyển đổi tăng thêm 3.2 lần so với các nguồn dẫn khác.
- Specialization\_Other Specialization (-1.43): Tương tự như biến đầu tiên, đây là hệ số cho chuyên ngành "Other Specialization", nhưng ở đây là khi nó không có mặt.
- Do Not Email (-1.15): Khi khách hàng không muốn nhận email, tỷ lệ cơ hội của sự chuyển đổi giảm đi 1.15 lần.
- Lead Origin\_Landing Page Submission (-1.08): Khi nguồn dẫn đến trang web là "Landing Page Submission", tỷ lệ cơ hội của sự chuyển đổi giảm đi 1.08 lần so với các nguồn dẫn khác.
- A free copy of Mastering The Interview (-0.32): Khi khách hàng không chọn nhận bản sao miễn phí của "Mastering The Interview", tỷ lệ cơ hội của sự chuyển đổi giảm đi 0.32 lần.
- Page Views Per Visit (-0.22): Khi số lượt xem trung bình trên mỗi lượt truy cập tăng thêm 1 đơn vị, tỷ lệ cơ hội của sự chuyển đổi giảm đi 0.22 lần.
- Intercept (-0.99): Đây là hệ số chặn (intercept), đại diện cho giá trị log-odds khi tất cả các biến độc lập đều bằng 0.

Dựa vào đó công thức tổng quát, ta có:

$$\begin{aligned}
 \text{logit}(p) &= \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_n \cdot X_n \\
 &= 0.23 * (\text{Specialization_Other Specialization}) + 0.5 * (\text{Last Activity_Page Visited on Website}) + 0.71 * (\text{Last Activity_Email Link Clicked}) \\
 &\quad + 0.84 * (\text{Lead Source_Olark Chat}) + 0.94 * (\text{Last Activity_Unreachable}) + \\
 &\quad 1.06 * (\text{Total Time Spent on Website}) + 1.16 * (\text{Last Activity_Email Opened}) +
 \end{aligned}$$

$1.26 * (\text{Last Activity\_Unsubscribed}) + 1.65 * (\text{Lead Source\_Welingak Website})$   
 $+ 1.96 * (\text{Last Activity\_Other Activity}) + 2.37 * (\text{Last Activity\_SMS Sent}) +$   
 $3.2 * (\text{Lead Origin\_Lead Add Form}) - 1.43 * (\text{Specialization\_Other Specialization}) - 1.15 * (\text{Do Not Email}) - 1.08 * (\text{Lead Origin\_Landing Page Submission}) - 0.32 * (\text{A free copy of Mastering The Interview}) - 0.22 * (\text{Page Views Per Visit}) - 0.99$

- Giả sử có một khách hàng mới đến trang web có thông tin sau về họ như sau chuyên ngành thương mại điện tử (E-COMMERCE). Khách hàng lựa chọn không muốn nhận email về khóa học và không thể truy cập hoạt động cuối, mã nguồn truy cập từ Landing Page Submission với thời gian hoạt động trên website là 5 và Số lượt xem trang trên mỗi lượt truy cập 2.

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

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

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

Predicted probability of conversion: [0.75954471]

## CHƯƠNG 4: MÔ HÌNH CHUỖI THỜI GIAN

### 4.1. Thu thập dữ liệu

Quá trình thu thập dữ liệu được thực hiện bằng cách sử dụng thư viện Selenium và ChromeDriver để truy cập vào URL của website [s.cafef.vn](http://s.cafef.vn). 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. Quá trình thu thập dữ liệu diễn ra như sau:

Đầu tiên, import một số thư viện vào Python phục vụ cho việc Crawl data như sau:

- *selenium.webdriver*: Đây là thư viện chính của Selenium, cung cấp các chức năng để tương tác với trình duyệt web.
- *selenium.common.NoSuchElementException*: Exception này được sử dụng khi không tìm thấy phần tử trên trang web, thường được sử dụng để xử lý các trường hợp lỗi khi trình duyệt không thể tìm thấy phần tử cần thiết.
- *selenium.webdriver.common.by.By*: Định nghĩa các phương thức để tìm kiếm phần tử trên trang web, như tìm kiếm theo ID, tên lớp, hoặc xpath.
- *selenium.webdriver.common.keys.Keys*: Chứa các hằng số để mô phỏng các phím trên bàn phím, được sử dụng để gửi các tương tác như nhập liệu hoặc di chuyển trong các trường hợp tự động hóa.
- *csv*: Thư viện dùng để đọc và ghi dữ liệu vào các file CSV, thường được sử dụng để lưu trữ dữ liệu hoặc kết quả từ việc tự động hóa trình duyệt web.
- *time.sleep*: Sử dụng để tạm ngừng chương trình trong một khoảng thời gian nhất định, thường được sử dụng để đợi cho trang web hoàn tất việc tải xong hoặc xử lý các tương tác khác.
- *re*: Thư viện regex (Regular Expressions) của Python, được sử dụng để thực hiện các thao tác tìm kiếm và thay thế trên chuỗi văn bản. Thường được sử dụng trong việc trích xuất dữ liệu từ trang web hoặc xử lý các chuỗi văn bản khác.

Tiếp theo, sử dụng các thư viện vừa import vào để truy cập vào trang cần lấy dữ liệu thông qua ID hoặc XPath. Sau đó, sử dụng vòng lặp để lần lượt lấy được dữ liệu của tất cả các trang và lưu vào file csv.

## 4.2. Mô tả bài toán

### 4.2.1. Đặt vấn đề

Tập đoàn Vingroup là một tập đoàn đa ngành của Việt Nam. Bắt đầu từ năm 2007, Tập đoàn Vingroup bắt đầu niêm yết trên Sở giao dịch Chứng khoán Thành phố Hồ Chí Minh với mã cổ phiếu VIC. Cho đến nay, tập đoàn Vingroup đã gia nhập thị trường chứng khoán được 16 năm và giá cổ phiếu đã trải qua nhiều giai đoạn thăng trầm.

Giá đóng cửa trong chứng khoán thường được coi là một chỉ số quan trọng vì nó cho biết giá cuối cùng của một cổ phiếu hoặc chỉ số chứng khoán trong một phiên giao dịch. Điều này có thể cung cấp thông tin quan trọng về xu hướng của thị trường trong ngày đó, đôi khi cũng là một phần quyết định xu hướng giá vào ngày hôm sau, dĩ nhiên là còn phải tùy thuộc vào các yếu tố khác như biến động lực tăng giá và nhiều yếu tố ngoại biên khác.

Hiện tại, tập đoàn VinGroup mong muốn xây dựng một mô hình giúp dự đoán giá đóng cửa của mã cổ phiếu VIC trong tương lai thông qua dữ lịch sử giao dịch của tập đoàn. Thông qua đó, kết quả thu được sẽ giúp tập đoàn đưa ra quyết định đầu tư sáng suốt hơn, tăng cơ hội sinh lời, giảm thiểu rủi ro, ...

### 4.2.2. Bộ dữ liệu

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)
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)

Bảng 4.1 Mô tả tiêu đề các cột của bộ dữ liệu VIC\_2007\_2024

#### 4.2.3. Câu hỏi nghiên cứu

- Xây dựng mô hình ARIMA để dự đoán giá đóng cửa và so sánh với kết quả thực tế?
- Xây dựng mô hình Holt-Winters để dự đoán giá đóng cửa và so sánh với kết quả thực tế?
- Trong 2 mô hình ARIMA và Holt-Winters, mô hình nào mang lại hiệu quả dự đoán cao hơn?

#### 4.3. Phân tích khám phá dữ liệu

##### 4.3.1. Kiểm tra dữ liệu và tổng quan số liệu

Sau khi tiến hành kiểm tra, bộ dữ liệu này bao gồm 11 cột và 4124 mẫu với các kiểu dữ liệu là datetime64[ns] (1), float64 (6), int64 (2), object (2). Bảng bên dưới mô tả chi tiết dữ liệu từng cột:

#	Column	Non-Null Count	Dtype
0	Date	4124 non-null	datetime64[ns]

1	Close_price	4124 non-null	float64
2	Adjusted_price	4124 non-null	object
3	Change	4124 non-null	object
4	Auction_weight	4124 non-null	int64
5	Auction_price	4124 non-null	float64
6	Put_through_weight	4124 non-null	int64
7	Put_through_price	4124 non-null	float64
8	Open_price	4124 non-null	float64
9	Highest_price	4124 non-null	float64
10	Lowest_price	4124 non-null	float64

Bảng 4.2 Tổng quan về các biến thuộc bộ dữ liệu VIC\_2007\_2024

Bảng bên dưới mô tả thống kê của các cột có kiểu dữ liệu số. Các thông số này cung cấp những thông tin cơ bản để tiến hành phân tích và đánh giá chất lượng của bộ dữ liệu.

	Date	Close_price	Auction_weight	Auction_price	Put_through_weight
count	4124	4124	4124	4124	4124
mean	2015-12-27 22:59:15	79,44838749	1109464,5	80,79593 841	392733,371
min	2007-09-19 00:00:00	33,3	0	0	0

<b>25%</b>	2011-11-07 18:00:00	53	152915	11,7375	0
<b>50%</b>	2015-12-28 12:00:00	77	479015	32,77	0
<b>75%</b>	2020-02-18 06:00:00	101	1349095	93,635	100000
<b>max</b>	2024-04-01 00:00:00	193	28220800	1809,9	51518053
<b>std</b>		28,58061318	2070922,7 71	141,8949 285	1709616,195

Bảng 4.3 Mô tả thống kê bộ dữ liệu VIC\_2007\_2024 #1

	<b>Put_through_price</b>	<b>Open_price</b>	<b>Highest_price</b>	<b>Lowest_price</b>
<b>count</b>	4124	4124	4124	4124
<b>mean</b>	27,93763094	79,39359845	80,37322987	78,35378274
<b>min</b>	0	33,3	33,3	33
<b>25%</b>	0	53	53,6	52,4
<b>50%</b>	0	77	78	76
<b>75%</b>	6,7975	100,425	102	99
<b>max</b>	5821,9	198	198	189
<b>std</b>	140,3759594	28,57446398	28,92069555	28,19519352

Bảng 4.4 Mô tả thống kê bộ dữ liệu VIC\_2007\_2024 #2

Kiểm tra giá trị bị thiếu của bảng dữ liệu, bảng bên dưới tổng hợp số lượng giá trị thiếu (NaN). Và dự vào bảng dữ liệu có thể thấy không có giá trị nào bị thiếu. Tuy

nhiên quá trình kiểm thử công nhận thấy dữ liệu cột ‘*Adjusted\_price*’ chưa có dữ liệu và thay thế bằng ‘--’ tuy nhiên cột không có ảnh hưởng đến quá trình thực hiện bài toán.

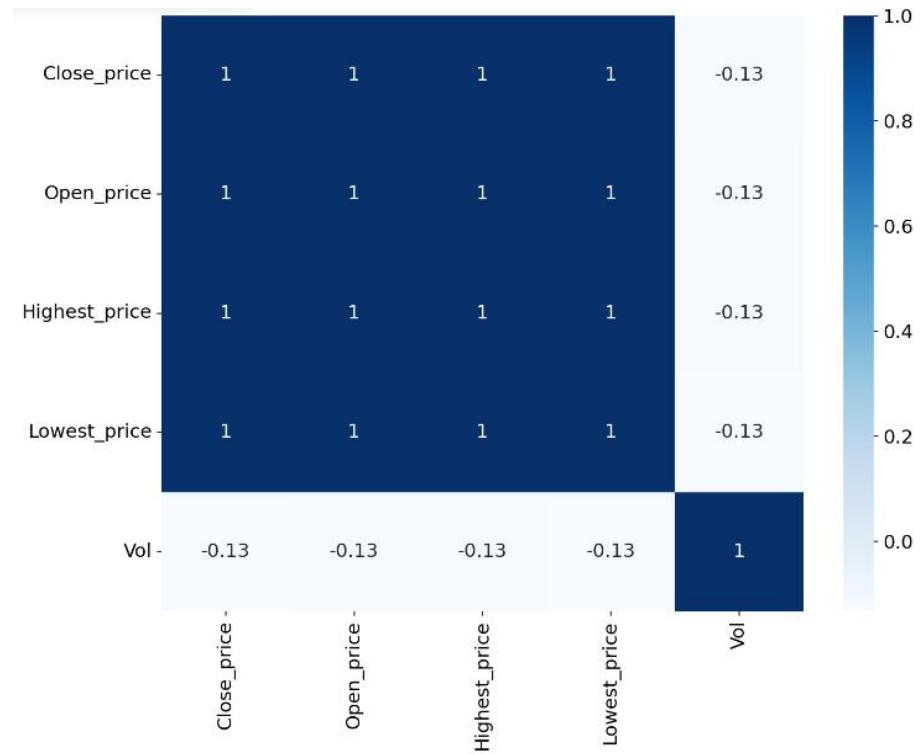
Column	Total
Date	0
Close_price	0
Adjusted_price	0
Change	0
Auction_weight	0
Auction_price	0
Put_through_weight	0
Put_through_price	0
Open_price	0
Highest_price	0
Lowest_price	0

Bảng 4.5 Thống kê số giá trị bị thiếu của từng biến trong dataset

#### 4.3.2. Phân tích dữ liệu

Dựa vào yêu cầu bài toán, nhóm nghiên cứu thực hiện loại bỏ một số cột không cần thiết (Adjusted\_price, Change, Auction\_weight, Auction\_price, Put\_through\_weight, Put\_through\_price) và tạo cột mới ‘Vol’ là số lượng giao dịch

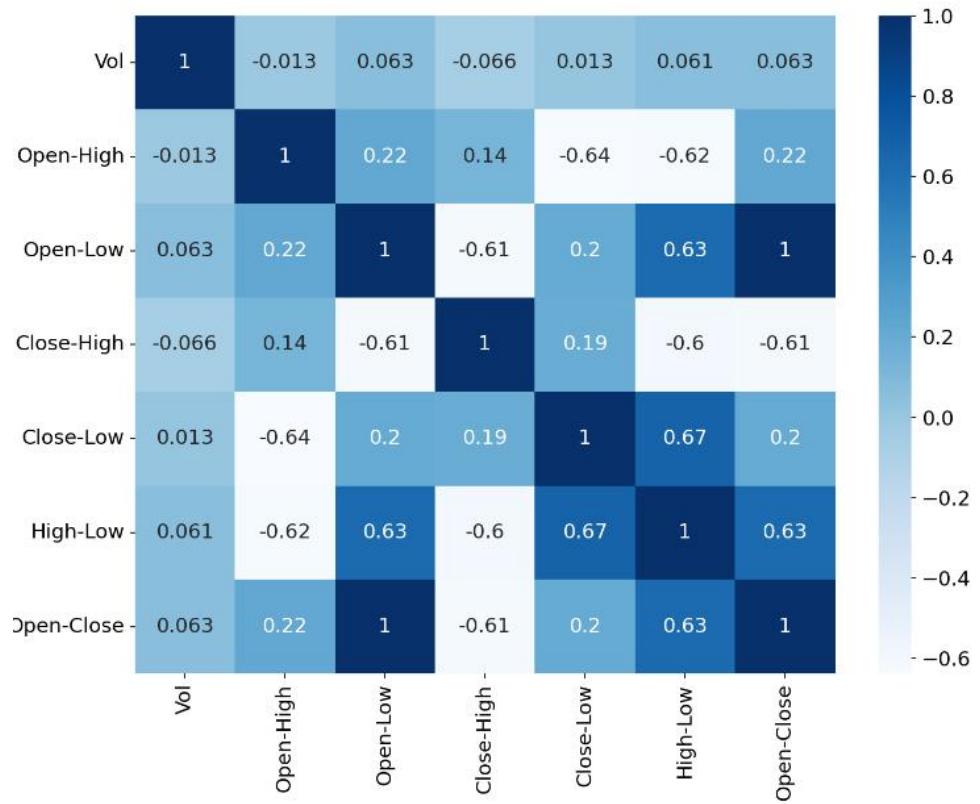
được tính dựa vào tổng ‘Auction\_weight’ và ‘Put\_through\_weight’. Sau đó tiếp tục thực hiện phân tích dữ liệu ảnh hưởng giá cổ phiếu như thế nào đối với các giao dịch.



*Hình 4.1 Biểu đồ nhiệt tương quan giữa các biến #1*

Từ biểu đồ tương quan trên, có thể thấy một số lượng lớn các số 1. Điều này có nghĩa là các biến có mối tương quan dương cao và có liên quan với nhau. Điều này có thể xảy ra vì sự khác biệt rất nhỏ giữa các giá trị đó. Tuy nhiên, trên thị trường chứng khoán, giá trị nhỏ này lại là điều tạo nên sự khác biệt. Vì vậy, nhóm sẽ làm thêm các biến mới vào tập dữ liệu thể hiện sự khác biệt giữa các giá trị đó.

- df['Open-High']: Sự khác biệt giữa giá mở cửa và giá cao nhất trong ngày.
- df['Open-Low']: Sự khác biệt giữa giá mở cửa và giá thấp nhất trong ngày.
- df['Close-High']: Sự khác biệt giữa giá đóng cửa và giá cao nhất trong ngày.
- df['Close-Low']: Sự khác biệt giữa giá đóng cửa và giá thấp nhất trong ngày.
- df['High-Low']: Sự khác biệt giữa giá cao nhất và giá thấp nhất trong ngày.
- df['Open-'Close]: Sự khác biệt giữa giá mở cửa và giá đóng cửa.



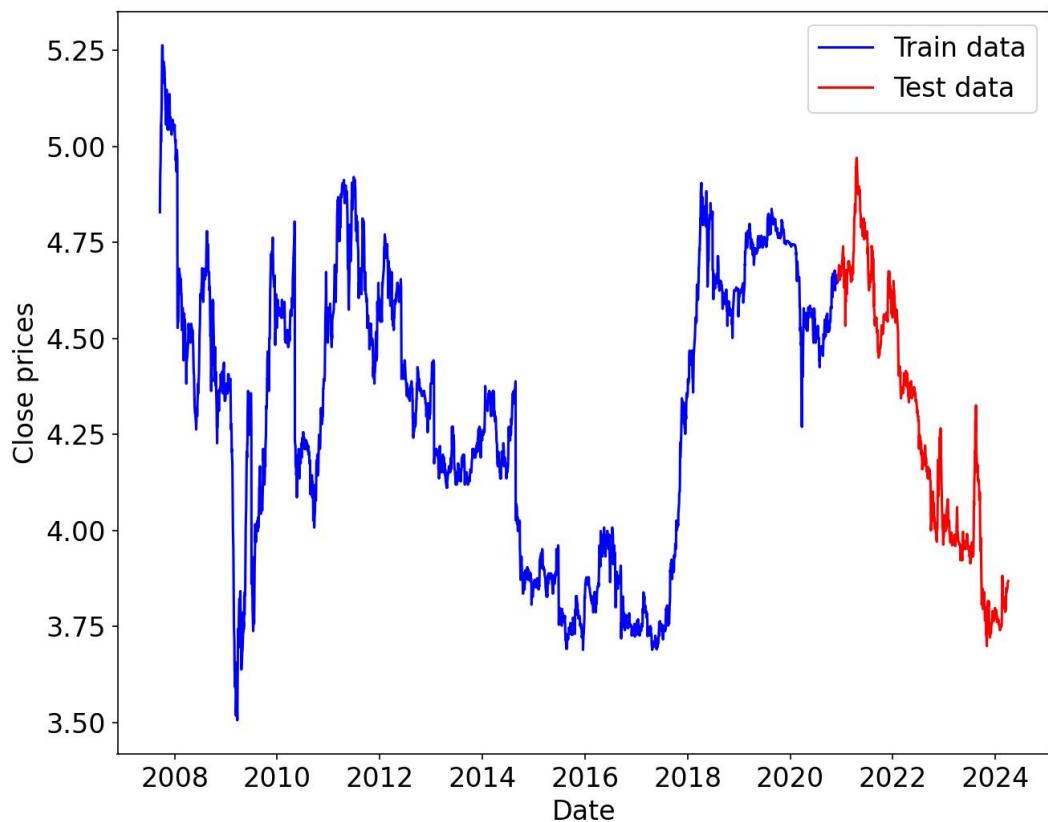
Hình 4.2 Biểu đồ nhiệt tương quan giữa các biến #2

Từ biểu đồ nhiệt trước đó, có thể thấy rằng những giá trị đó không rất hữu ích. Vì vậy, để rõ ràng hơn sẽ loại bỏ những giá trị đó và chỉ sử dụng các giá trị mới để tạo biểu đồ nhiệt. Dựa vào biểu đồ trên, có một số phân tích ảnh hưởng của giá so với giao dịch phù hợp với thực tế như sau:

- Sự chênh lệch giữa giá đóng cửa và giá cao nhất có hệ số tương quan nghịch tương đối cao so với còn lại là 0.066 có thể thấy là khi giá đóng cửa cao hơn nhiều so với giá thấp nhất thì số lượng cổ phiếu giao dịch càng thấp.
- Sự chênh lệch giữa giá mở cửa và giá thấp nhất có hệ số tương quan thuận tương đối cao so với còn lại là 0.063 có thể thấy là khi giá mở cửa cao hơn nhiều so với giá thấp nhất thì số lượng cổ phiếu giao dịch càng nhiều.

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. Thay vào đó, nó chỉ cho thấy một mối tương quan nhỏ giữa chúng. Điều này có thể do ảnh hưởng của các yếu tố khác trong thị trường chứng khoán mà không phản ánh mối quan hệ trực tiếp giữa giá đóng cửa và giá cao nhất. Dựa vào đặt vấn đề nhóm sẽ thực hiện ứng dụng mô hình chuỗi thời gian để dự đoán giá đóng cửa.

Chia tập dữ liệu thành tập huấn luyện và tập kiểm tra. Trong đó, tập dữ liệu huấn luyện là 80% của dataset và tập dữ liệu kiểm tra là 20% còn lại của dataset.



Hình 4.3 Tập huấn luyện và tập kiểm tra của bài nghiên cứu

#### 4.3.3. Xu hướng



Hình 4.4 Biểu đồ trực quan sau khi rolling dữ liệu

Dựa vào biểu đồ, không thể đưa ra kết luận cụ thể giá đóng cửa luôn tăng hoặc luôn giảm qua các năm. Đúng như tình hình thực tế giá cổ phiếu sẽ có những biến động liên tục ở từng thời gian khác nhau do ảnh hưởng của nhiều yếu tố khác tác động không thể luôn kinh doanh thuận lợi mà luôn tăng mà sẽ có những lúc biến động. Và đặc biệt có những năm biến động như trong giai đoạn năm 2008 đến năm 2010. Thực tế giai đoạn này đang có nhiều biến động đặc biệt trong tình hình kinh tế năm 2008 đánh dấu sự sụt giảm của nền kinh tế nói chung và thị trường bất động sản nói riêng. Điều này khiến cho các doanh nghiệp xây dựng và bất động sản gặp nhiều khó khăn. Tuy nhiên trong các năm sau, bắt đầu từ năm 2011 khi các dự án này hoàn thành, cùng với khả năng thị trường bất động sản khởi sắc trở lại.



*Hình 4.5. Biểu đồ phân rã chuỗi dữ liệu*

Sau khi phân tích dữ liệu, 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. Cũng như các nhận định trước đó đặc biệt hơn VinGroup đầu tư trong nhiều lĩnh vực khác nhau như ngành bất động sản cao cấp, giá cổ phiếu có thể không theo một hướng cụ thể do ảnh hưởng của nhiều yếu tố như thị trường bất động sản, tình hình kinh tế, chính sách quản lý nhà nước và các yếu tố khác. Tuy nhiên, có thể thấy một sự biến động theo chu kỳ mùa vụ, giá cổ phiếu có xu hướng biến động theo một mô hình lặp lại trong một khoảng thời gian nhất định.

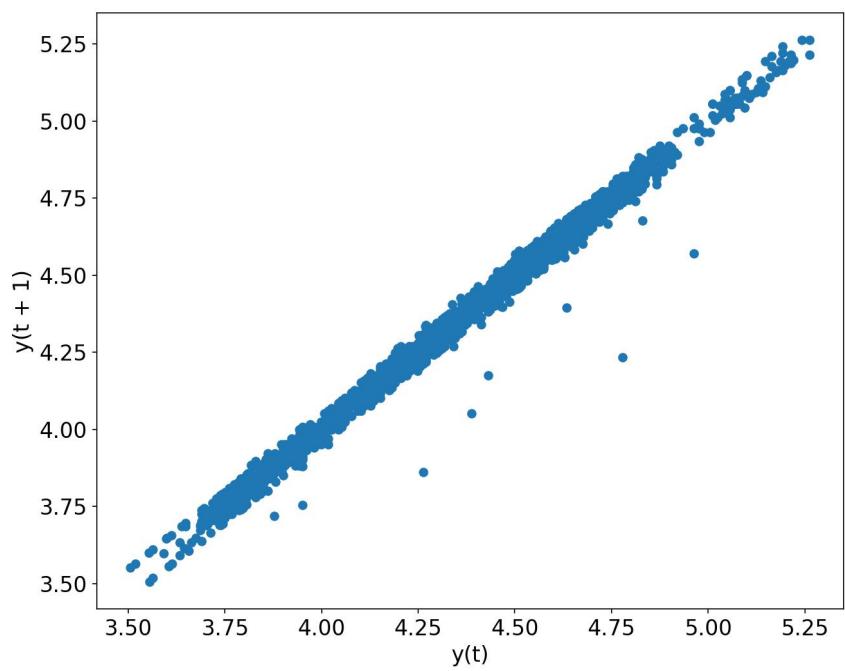
#### 4.3.4. Kiểm tra tính dừng

```
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
```

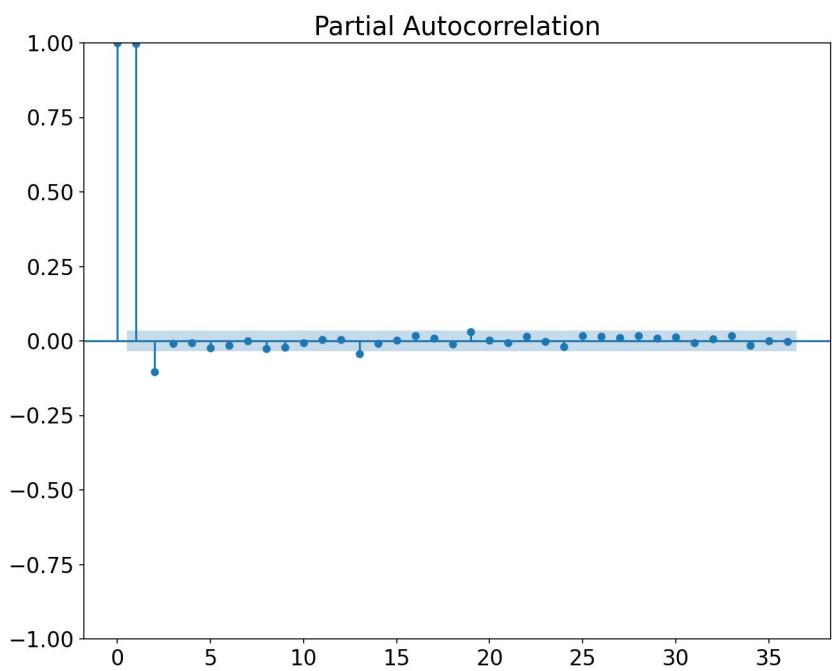
Hình 4.6 Kết quả kiểm định ADF và KPSS

Để thực hiện kiểm tra tính dừng của dữ liệu nhóm sử dụng cả hai phương pháp kiểm định ADF và KPSS.

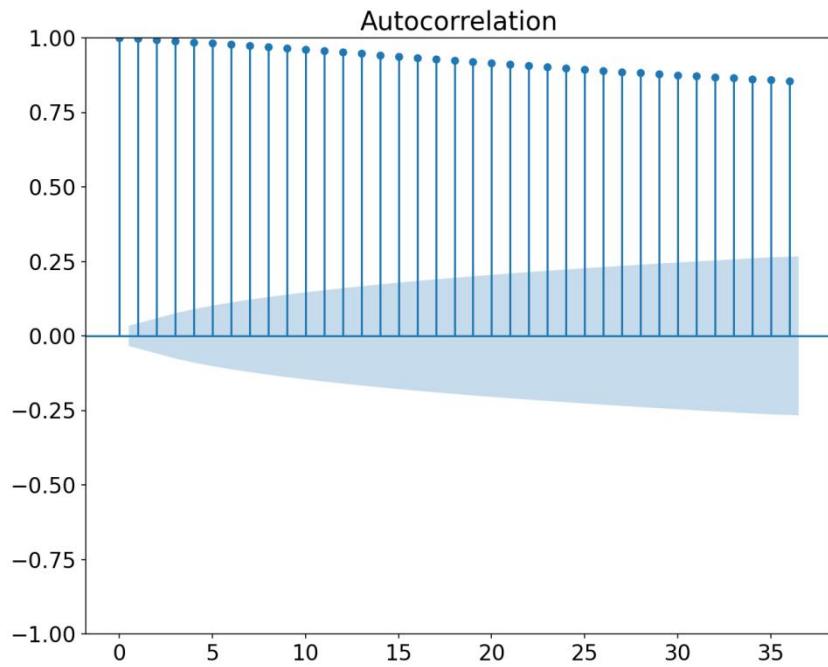
- ADF Test (Augmented Dickey-Fuller Test):
  - + Giá trị p-value của ADF Test là 0.117857, lớn hơn mức ý nghĩa 0.05.
  - + Bên cạnh đó, trị tuyệt đối giá trị thống kê (Test statistic) của ADF Test là 2.490281, nhỏ hơn các ngưỡng tối hạn tương ứng với mức ý nghĩa 1%, 5%, và 10%.  
⇒ Không thể bác bỏ giả thuyết  $H_0$ , dữ liệu không có tính dừng.
- KPSS Test (Kwiatkowski-Phillips-Schmidt-Shin Test):
  - + Giá trị p-value của KPSS Test là 0.01, nhỏ hơn mức ý nghĩa 0.05.
  - + Trị tuyệt đối của giá trị thống kê (Test statistic) của KPSS Test là 1.045143, lớn hơn các ngưỡng tối hạn tương ứng với mức ý nghĩa 10%, 5%, 2.5%, và 1%.  
⇒ Bác bỏ giả thuyết:  $H_0$ , dữ liệu không có tính dừng.



Hình 4.7. Biểu đồ thể hiện tính tự tương quan



Hình 4.8 Tương quan PACF

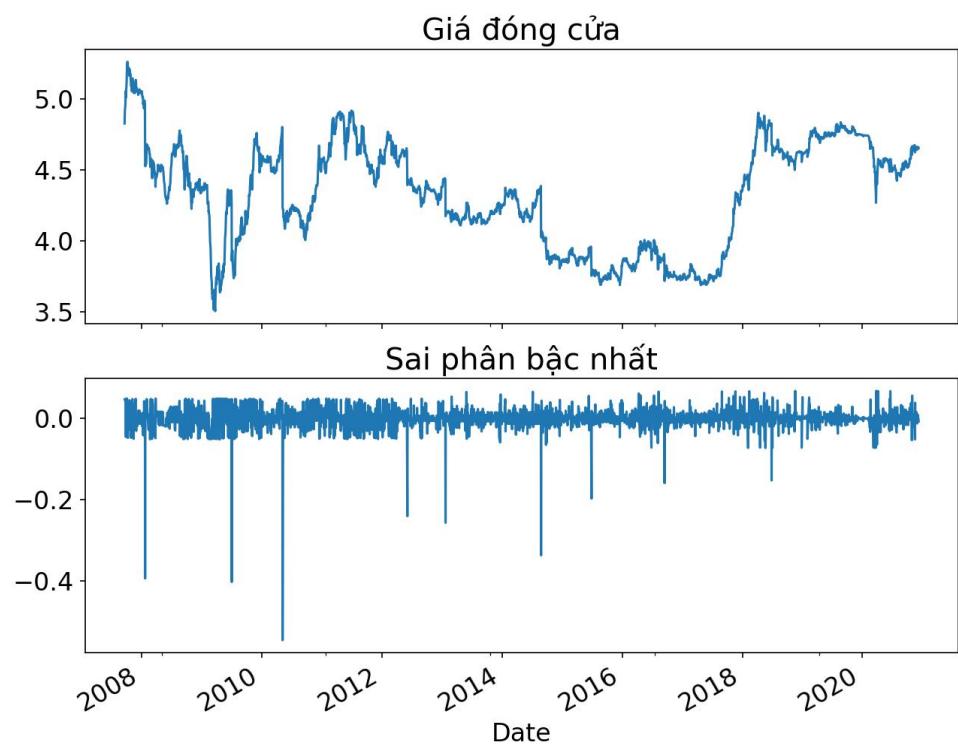


*Hình 4.9 Tương quan ACF*

#### **4.4. Chuyển đổi dữ liệu thành chuỗi dừng và tự tương quan**

##### **4.4.1. Chuyển đổi dữ liệu thành chuỗi dừng**

Khi dữ liệu chưa dừng (non-stationary), việc phân tích trên bộ dữ liệu này sẽ ảnh hưởng đến mô hình và kết quả dự báo. Vì vậy, nhóm chuyển đổi dữ liệu sang tính dừng để đảm bảo không ảnh hưởng đến mô hình và kết quả dự báo bằng cách sử dụng sai phân.



Hình 4.10 Biểu đồ thể hiện dữ liệu ban đầu và sau khi lấy sai phân bậc 1

#### 4.4.2. Kiểm tra tính dừng của chuỗi dữ liệu đã lấy sai phân

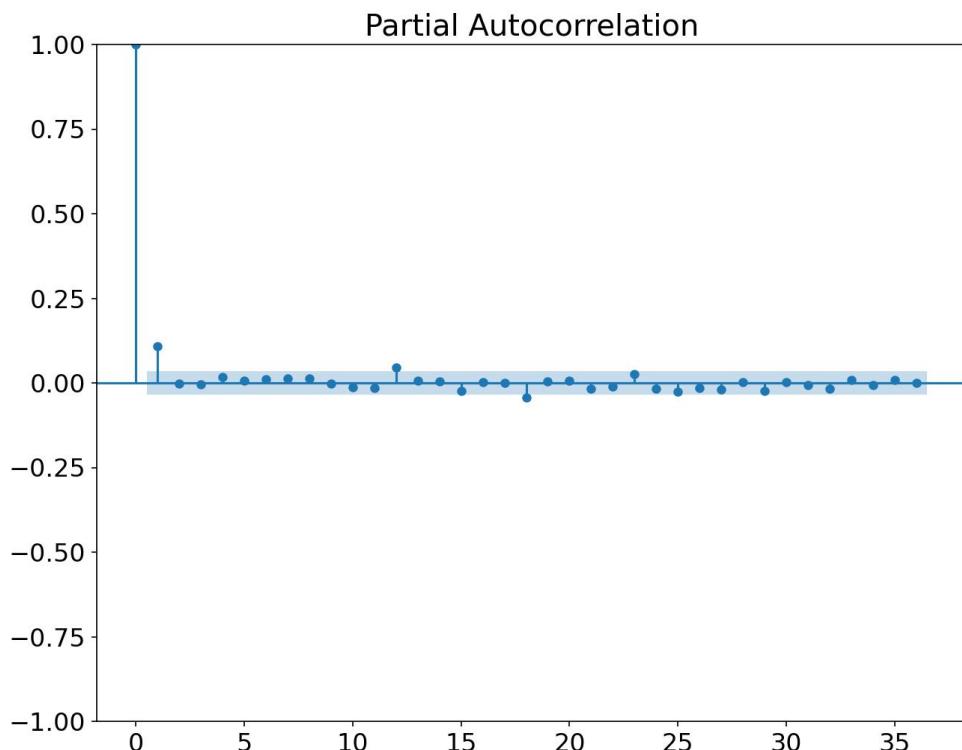
```
Reject the null hypothesis (H0),  
the data is stationary  
ADF: Test statistic      -51.544520  
p value                  0.000000  
# of Lags                 0.000000  
# of Observations        3297.000000  
Critical Value (1%)       -3.432335  
Critical Value (5%)       -2.862417  
Critical Value (10%)      -2.567237  
dtype: float64  
-----  
Fail to reject the null hypothesis (H0),  
the data is stationary  
KPSS: Test statistic      0.103201  
p value                   0.100000  
# of Lags                 10.000000  
Critical Value (10%)      0.347000  
Critical Value (5%)       0.463000  
Critical Value (2.5%)     0.574000  
Critical Value (1%)       0.739000  
dtype: float64
```

Hình 4.11 Kết quả kiểm định ADF và KPSS sau khi lấy sai phân bậc 1

- ADF Test (Augmented Dickey-Fuller Test):
  - + Giá trị p-value của ADF Test là 0, nhỏ hơn mức ý nghĩa 0.05.
  - + Bên cạnh đó, trị tuyệt đối giá trị thống kê (Test statistic) của ADF Test là 51.544520, lớn hơn các ngưỡng tối hạn tương ứng với mức ý nghĩa 1%, 5%, và 10%.
- ⇒ Bác bỏ giả thuyết  $H_0$ , dữ liệu đã dừng.
- KPSS Test (Kwiatkowski-Phillips-Schmidt-Shin Test):
  - + Giá trị p-value của KPSS Test là 0.11, nhỏ hơn mức ý nghĩa 0.05.

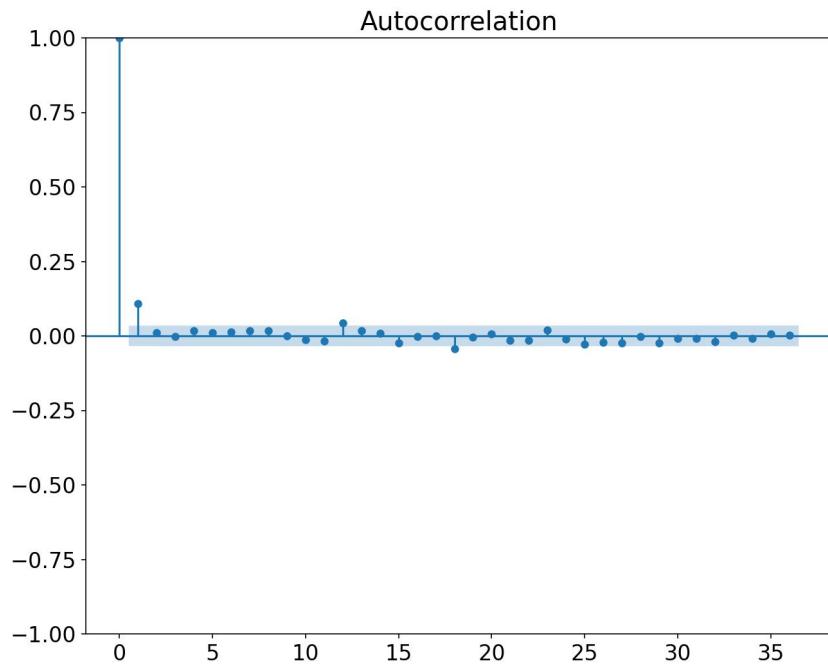
- + Giá trị thống kê (Test statistic) của KPSS Test là 0.103201, nhỏ hơn các ngưỡng tối hạn tương ứng với mức ý nghĩa 10%, 5%, 2.5%, và 1%.
- ⇒ Bác bỏ giả thuyết  $H_0$ , dữ liệu đã dừng.

#### 4.4.3. Kiểm định tính tự tương quan



Hình 4.12 Tương quan PACF sau khi lấy sai phân bậc 1

Từ biểu đồ tương quan PACF, nhóm nghiên cứu xác định p có thể bằng 1 trong 2 giá trị 0 và 1.



*Hình 4.13. Tương quan ACF sau khi lấy sai phân bậc 1*

Từ biểu đồ tương quan ACF, nhóm nghiên cứu xác định q có thể bằng 1 trong 2 giá trị 0 và 1.

#### 4.5. Xây dựng mô hình

##### 4.5.1. Mô hình ARIMA

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

Mô hình	Giá trị
ARIMA(0,1,0)	AIC = -14487.383957742079
ARIMA(0,1,1)	AIC = -14523.496586429437
ARIMA(1,1,0)	AIC = -14523.864646668484
ARIMA(1,1,1)	AIC = -14521.861566201878

*Bảng 4.6 Tính chỉ số AIC để xác định mô hình ARIMA tối ưu nhất*

Mô hình 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) với AIC = -14523.86464666849 là thấp nhất trong 4 mô hình ARIMA được liệt kê.

#### 4.5.2. Mô hình SARIMAX

Từ kết quả của biểu đồ phân rã chuỗi dữ liệu, nhóm nghiên cứu thấy rằng dữ liệu có tính mùa vụ. Qua đó, áp dụng mô hình SARIMAX để dự báo.

- ❖ Sử dụng Auto ARIMA để tự động xác định tham số p, d, q cho mô hình ARIMA

```
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
```

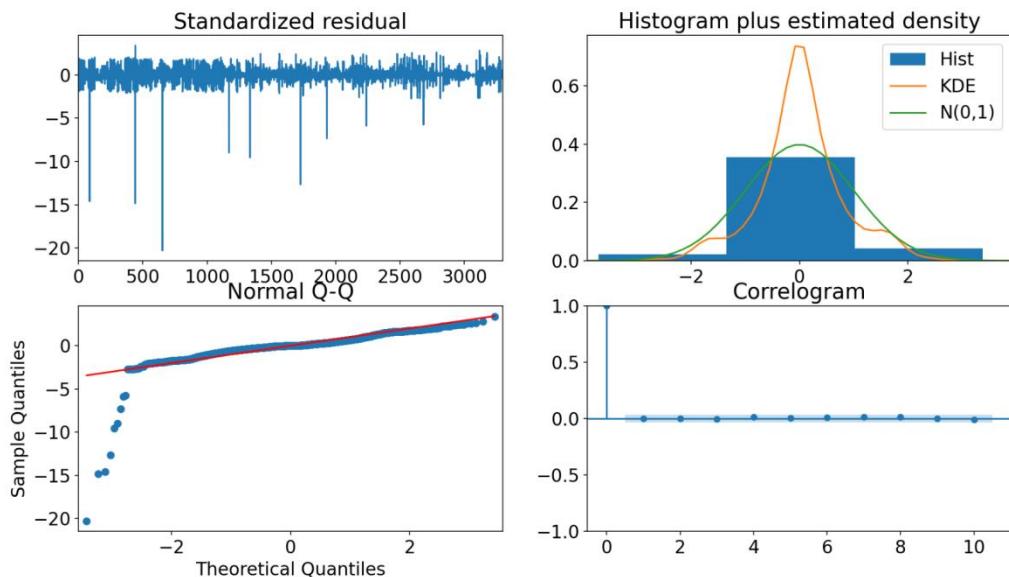
*Hình 4.14 Kết quả chạy mô hình Auto ARIMA #1*

```

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 - 3299         HQIC:                -14521.487
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
=====
```

Hình 4.15 Kết quả chạy mô hình Auto ARIMA #2



Hình 4.16. Biểu đồ đánh giá tính hợp lệ của các giả định được đưa ra bởi mô hình Auto ARIMA

Từ kết quả trên, nhóm nghiên cứu thấy mô hình phù hợp nhất là ARIMA (1,1,0).

- ❖ Tạo model với ARIMA (1,1,0) và dự báo

```

SARIMAX Results
=====
Dep. Variable: Close_price   No. Observations: 3299
Model: ARIMA(1, 1, 0)   Log Likelihood 7264.932
Date: Mon, 15 Apr 2024   AIC -14523.865
Time: 01:13:15   BIC -14505.561
Sample: 0   HQIC -14517.313
                - 3299
Covariance Type: opg
=====
            coef    std err        z      P>|z|      [0.025      0.975]
-----
x1     -5.45e-05    0.001   -0.084      0.933     -0.001      0.001
ar.L1      0.1077    0.011     9.409      0.000      0.085      0.130
sigma2     0.0007  3.24e-06   220.285      0.000      0.001      0.001
=====
Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 1185194.98
Prob(Q): 0.99 Prob(JB): 0.00
Heteroskedasticity (H): 0.28 Skew: -5.74
Prob(H) (two-sided): 0.00 Kurtosis: 95.16

```

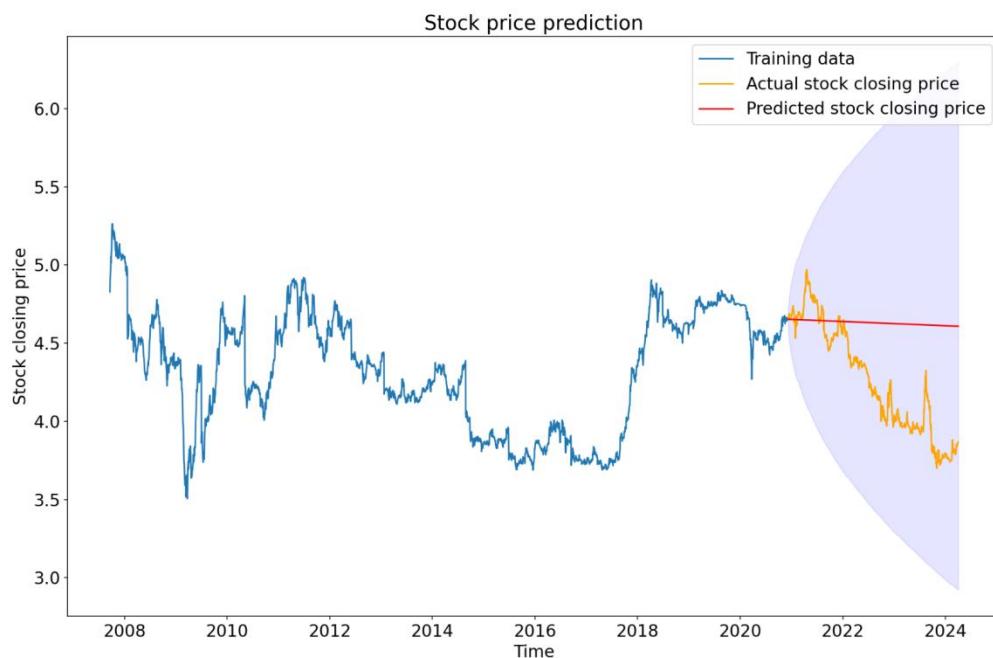
Hình 4.17 Kết quả mô hình

**Test MSE: 0.245**

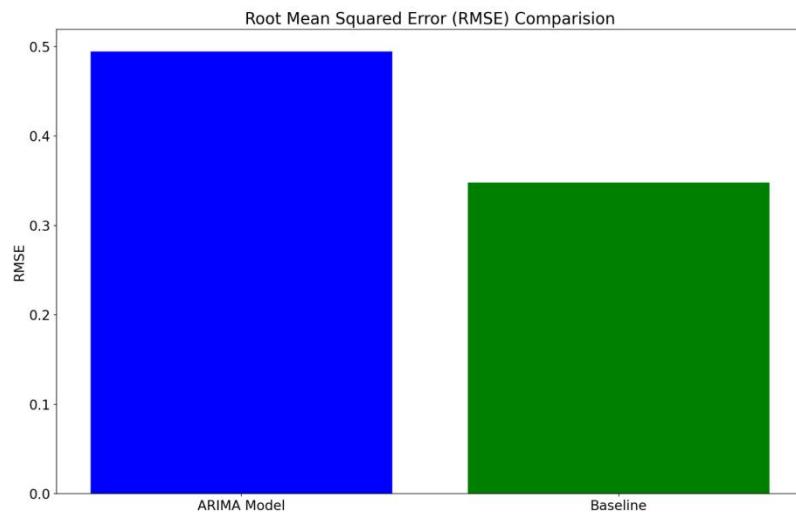
**Test RMSE: 0.495**

**MAPE: 0.10077488055744872**

Hình 4.18 Tính toán chỉ số MSE, RMSE và MAPE



*Hình 4.19 Kết quả dự đoán giá đóng cửa*



*Hình 4.20 Biểu đồ so sánh giữa RMSE và Baseline RMSE*

Từ kết quả trên, 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.

## **4.6. Xây dựng mô hình Holt-Winters**

### **4.6.1. Chuyển đổi dữ liệu**

Từ dataset VIC\_2007\_2024, lựa chọn giữ lại cột đang cần phân tích là “Close\_price” và xóa đi các cột còn lại không phục vụ cho mô hình dự đoán. Sau đó, sử dụng phương thức resample để chuyển đổi dữ liệu trong DataFrame theo chu kỳ tháng (MS). Đồng thời, do số phiên giao dịch trong mỗi tháng là không đều nhau cho nên nhóm nghiên cứu sẽ sử dụng phương thức mean() được áp dụng để tính trung các giá trị trong mỗi chu kỳ vì điều này sẽ đặc trưng được cho giá đóng cửa của các phiên giao dịch trong 1 tháng.

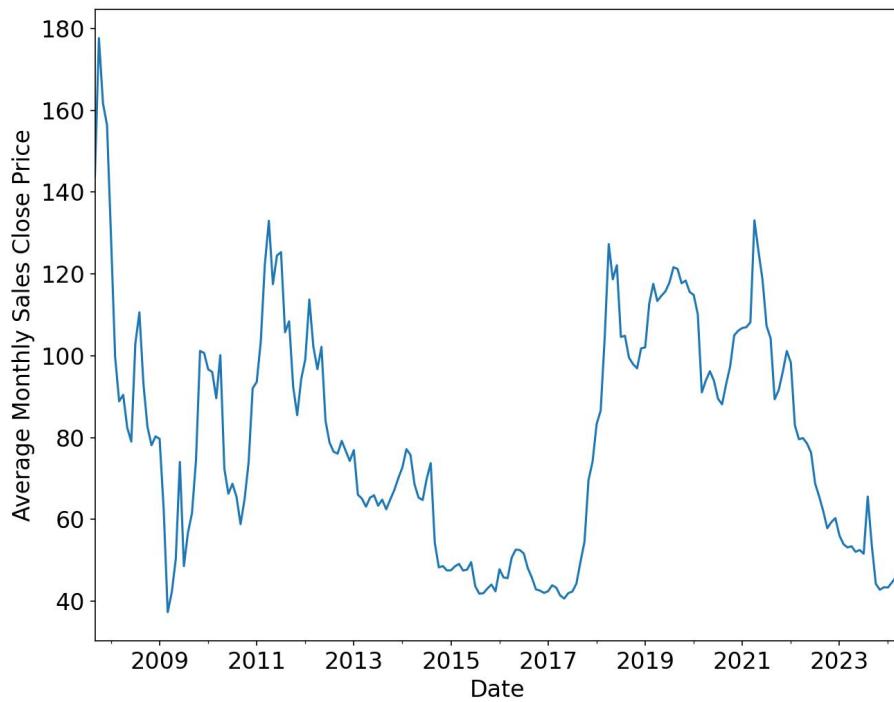
```

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 200 entries, 2007-09-30 to 2024-04-30
Freq: ME
Data columns (total 1 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Close_price  200 non-null    float64
dtypes: float64(1)
memory usage: 3.1 KB
None
Close_price
Date
2007-09-30  143.750000
2007-10-31  177.652174
2007-11-30  161.727273
2007-12-31  156.400000

```

*Hình 4.21 Mô tả dữ liệu sau khi chuyển đổi theo chu kỳ tháng và 4 dòng đầu của bộ dữ liệu*

Sau khi chuyển đổi dữ liệu, kết quả thu được bộ dữ liệu mới với 200 dòng, không có giá trị null với kiểu dữ liệu float64.



Hình 4.22 Bộ dữ liệu sau khi chuyển đổi theo chu kỳ tháng

#### 4.6.2. Kiểm tra tính xu hướng, mùa vụ, chu kỳ

Áp dụng bộ lọc HP (Hodrick-Prescott), lọc chuỗi thời gian thành 2 thành phần là thành phần xu hướng  $\tau_t$  và thành phần chu kỳ  $c_t$ .

Date	
2007-09-30	-4.832248
2007-10-31	35.125316
2007-11-30	25.259831
2007-12-31	25.970758
2008-01-31	4.314173
	...
2023-12-31	-2.353901
2024-01-31	-1.360180
2024-02-29	0.912871
2024-03-31	3.246496
2024-04-30	6.172447
Freq:	ME
Name:	Close_price_cycle
Length:	200
dtype:	float64

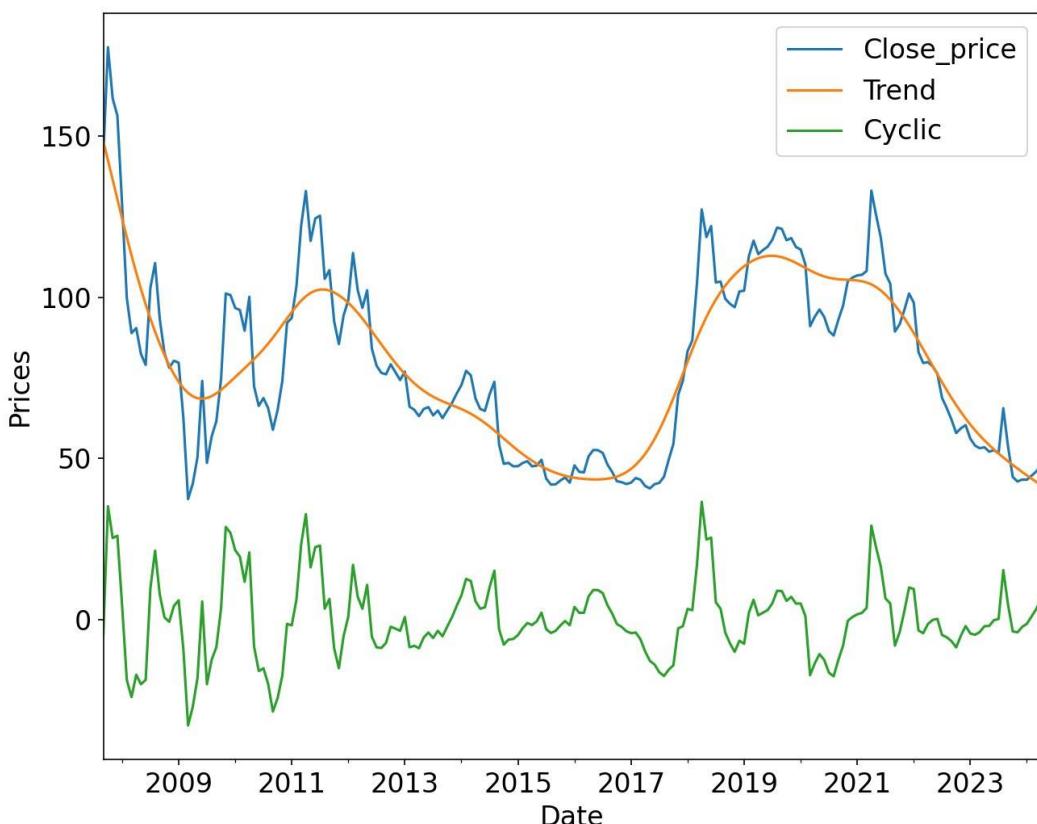
Hình 4.24 Kết quả ước lượng chu kỳ

```

Date
2007-09-30    148.582248
2007-10-31    142.526858
2007-11-30    136.467441
2007-12-31    130.429242
2008-01-31    124.458554
...
2023-12-31    45.768187
2024-01-31    44.721544
2024-02-29    43.696504
2024-03-31    42.684457
2024-04-30    41.677553
Freq: ME, Name: Close_price_trend, Length: 200, dtype: float64

```

*Hình 4.25 Kết quả ước lượng xu hướng*

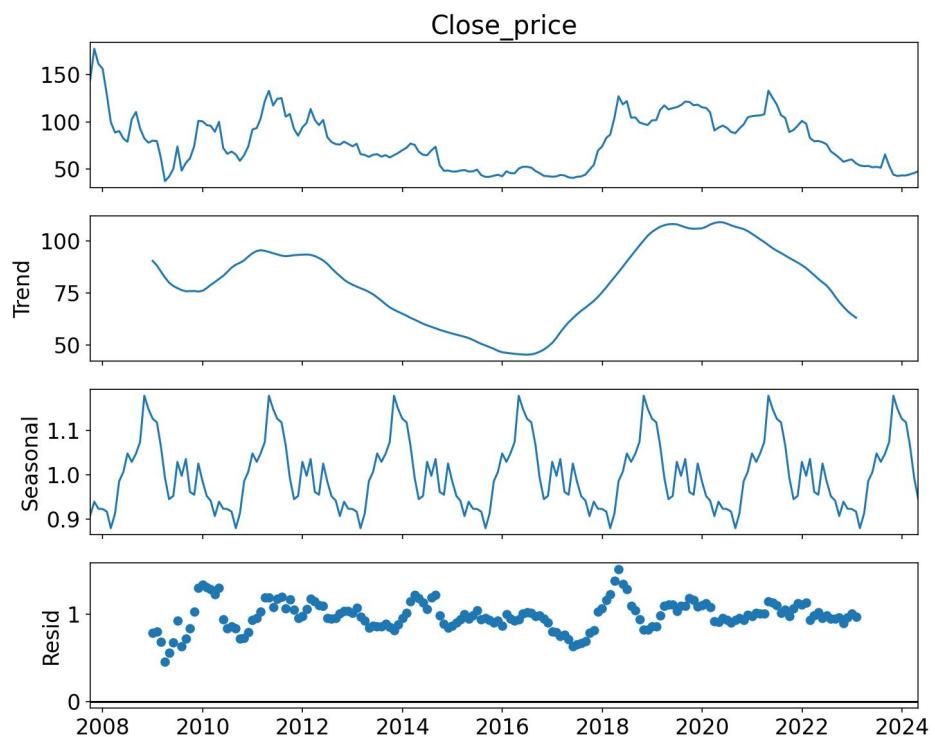


*Hình 4.26 Biểu đồ so sánh giá trị thật, chu kỳ và kỳ vọng*

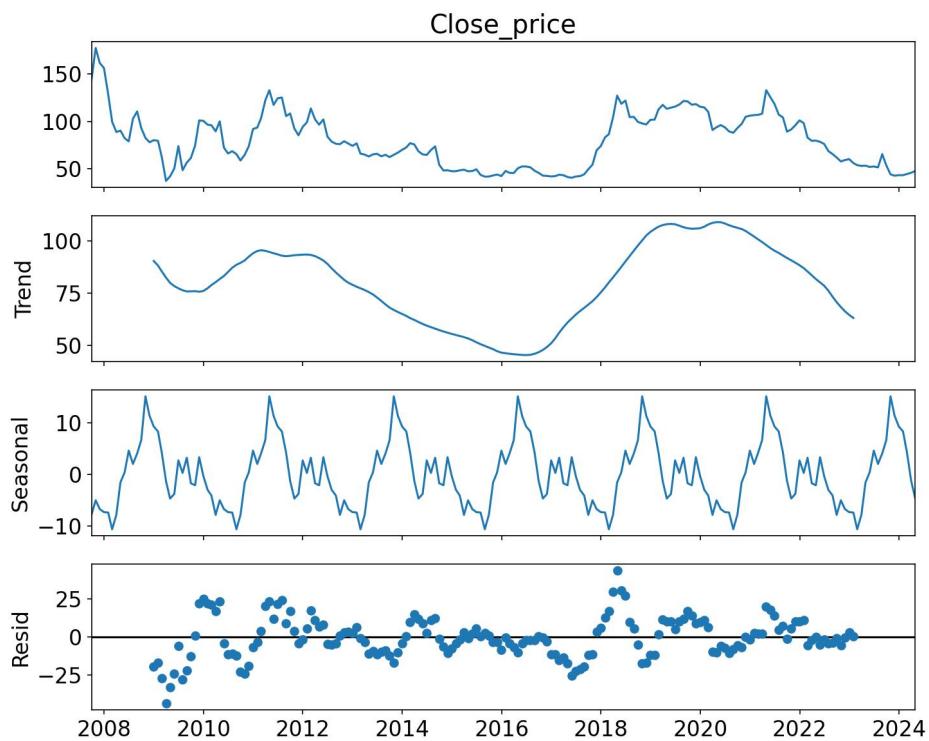
Từ biểu đồ so sánh giá trị thật, chu kỳ và kỳ vọng, nhóm nghiên cứu nhận thấy giá trị thực tế là tổng của 2 thành phần xu hướng và chu kỳ.

Nhìn chung, xu hướng của dữ liệu vẫn tương tự như trước khi chuyển đổi theo chu kỳ tháng là không thể đưa ra kết luận cụ thể giá đóng cửa luôn tăng hoặc luôn giảm qua các năm. Đồng thời, thành phần chu kỳ thể hiện rằng dữ liệu có tính chu kỳ tương đối nhưng không liên tục.

#### 4.6.3. Kiểm tra mô hình phân rã chuỗi thời gian



Hình 4.27. Biểu đồ phân rã chuỗi dữ liệu theo mô hình nhân tính



Hình 4.28 Biểu đồ phân rã chuỗi dữ liệu theo mô hình cộng tính

Dựa trên 2 biểu đồ phân rã chuỗi dữ liệu, bộ dữ liệu có tính mùa vụ do các biến động có biên độ giống nhau. Đồng thời, mô hình nhân tính sẽ phù hợp hơn trong việc phân tích bộ dữ liệu do các giá trị biến động nhỏ và ổn định hơn.

#### 4.6.4. Tối ưu hóa siêu tham số (Hyperparameter Optimization)

Trong mô hình Holt-Winters, 3 siêu tham số  $\alpha$ ,  $\beta$ ,  $\gamma$  rất quan trọng vì chúng được sử dụng để điều chỉnh mức độ ảnh hưởng của 3 yếu tố level, trend và seasonal lên kết quả dự báo chuỗi thời gian. Do đó, nhóm nghiên cứu sử dụng hàm `test_optimizer()` để xác định bộ siêu tham số  $\alpha$ ,  $\beta$ ,  $\gamma$  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.

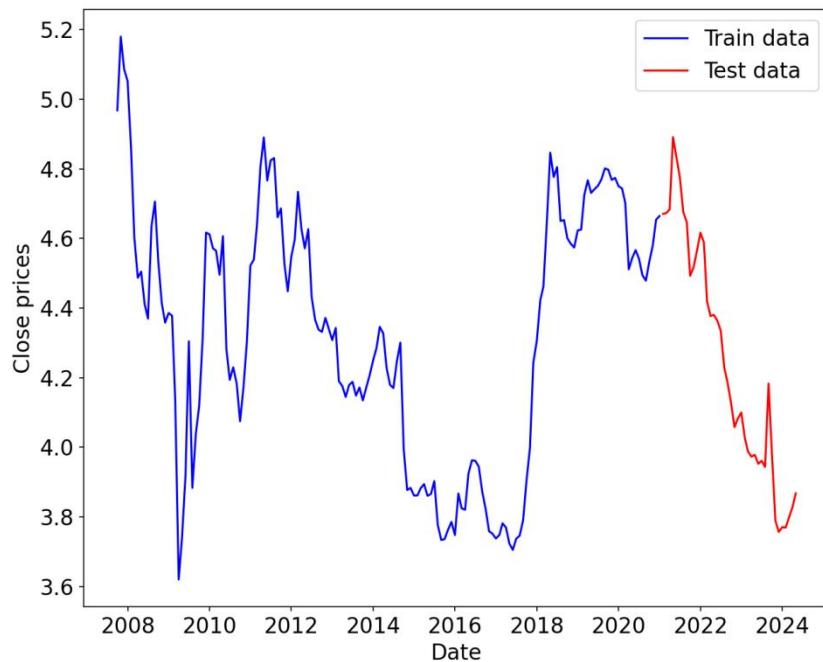
```
best_alpha: 0.6 best_beta: 0.0 best_gamma: 0.8 best_mae: 0.2304
```

Hình 4.29 Kết quả ước lượng tổ hợp siêu tham số tối ưu nhất

Kết quả thu được bộ siêu tham số giúp mô hình mang lại hiệu quả tối ưu nhất là:  
 $\alpha = 0.6$ ,  $\beta = 0$ ,  $\gamma = 0.8$ .

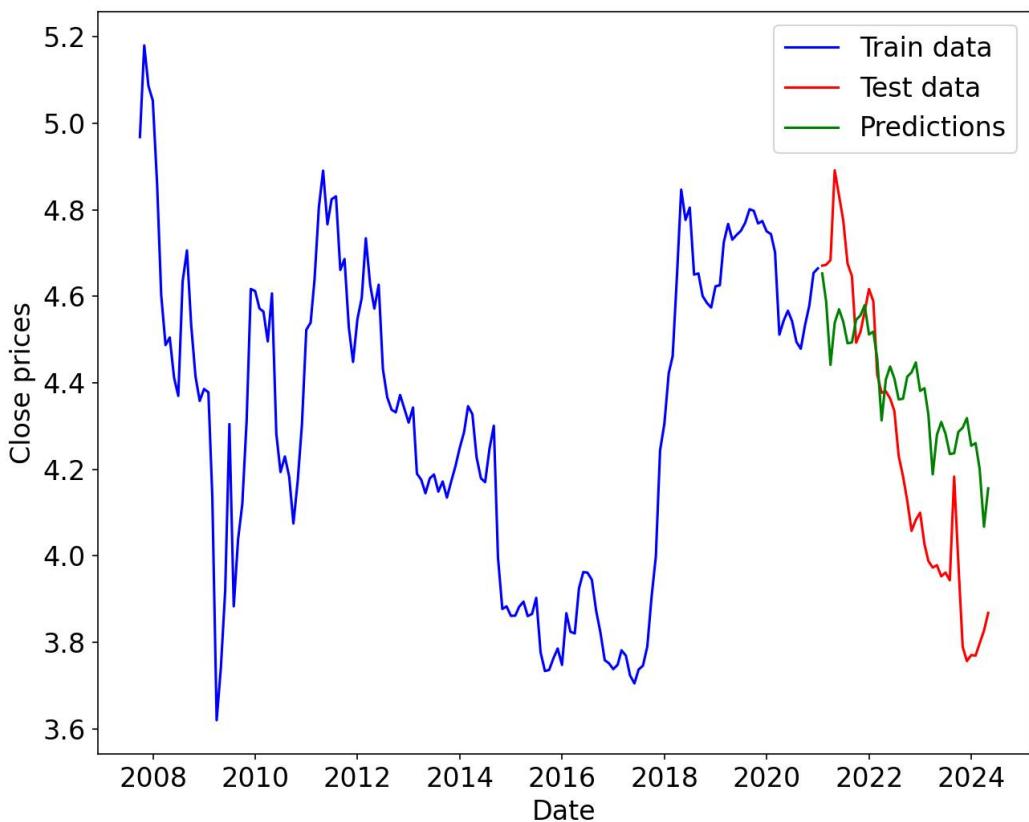
#### 4.6.5. Dự báo bằng mô hình Holt-Winters

Chia tập dữ liệu sau khi chuyển đổi theo chu kỳ thành tập huấn luyện và tập kiểm tra. Trong đó, tập dữ liệu huấn luyện là 80% dataset, tập dữ liệu kiểm tra là 20% dataset.



Hình 4.30 Tập huấn luyện và tập kiểm tra của dataset mới

Tiếp theo, sử dụng phương pháp Triple Exponential Smoothing (TES) để fit model vì dataset không có tính mùa vụ rõ ràng. Đồng thời, phương pháp TES phù hợp khi dữ liệu có tồn tại thành phần mùa vụ mà không có xu hướng tuyến tính rõ ràng. Sau đó, dự báo giá đóng cửa của VIC và so sánh với tập kiểm tra.



Hình 4.31 Kết quả dự đoán giá đóng cửa chứng khoán bằng mô hình Holt-Winters

**Test MSE: 0.076**  
**Test RMSE: 0.275**  
**MAPE: 0.056532617068493815**

Hình 4.32 Tính toán chỉ số MSE, RMSE và MAPE

Từ kết quả trên, mô hình Holt-Winters 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, RMSE và MAPE lần lượt là 0.076, 0.275 và xấp xỉ 0.057 là khá nhỏ chứng tỏ mô hình có độ chính xác cao bởi 3 chỉ số trên càng gần với giá trị 0 thì mô hình càng chính xác.

#### 4.7. Kết luận

Cả 2 mô hình ARIMA và Holt-Winter đã dự đoán đúng xu thế chuyển động của test\_data trong bộ dữ liệu là 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. Đồng thời, 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 thể hiện rằng mô hình Holt-Winters mang lại kết quả với độ chính xác cao hơn so với mô hình ARIMA hay SARIMA.

## CHƯƠNG 5: MÔ HÌNH HỌC MÁY

### 5.1. Mô tả bài toán

#### 5.1.1. Đặt vấn đề

Trong ngành ngân hàng, việc thu hút khách hàng để gửi tiền có kỳ hạn là một trong những yếu tố quan trọng nhằm tăng cường nguồn vốn và tạo lợi nhuận cho ngân hàng. Đặc biệt là trong năm 2023 khi mà các ngân hàng đồng thời tăng lãi suất khuyến khích khách hàng gửi tiền năm gia tăng nguồn cung tiền, từ đó cho các doanh nghiệp khác vay. Tuy nhiên, đối với mỗi khách hàng, quyết định gửi tiền (có kỳ hạn) không chỉ phụ thuộc vào những yếu tố tài chính mà còn phụ thuộc vào nhiều yếu tố khác như tuổi tác, nghề nghiệp, tình trạng hôn nhân, trình độ học vấn và một số yếu tố khác.

Trước thách thức của việc dự đoán xác suất một khách hàng sẽ quyết định gửi tiền (có kỳ hạn) tại ngân hàng, chúng ta cần tìm ra các yếu tố quan trọng nhất ảnh hưởng đến quyết định này. Do đó cần một công cụ hay một chuyên gia có thể phân tích được các vấn đề khó khăn của ngân hàng. Việc này có thể giúp ngân hàng tối ưu hóa chiến lược tiếp thị và tài chính, từ đó tăng cường hiệu quả kinh doanh và cung cấp giải pháp tài chính phù hợp với nhu cầu của từng khách hàng.

#### 5.1.2. Bộ dữ liệu

Thuộc tính	Mô tả
Age	Tuổi của khách hàng
Job	Nghề nghiệp của khách hàng
Martial	Tình trạng hôn nhân của khách hàng
Education	Trình độ học vấn của khách hàng
Default	Biến nhị phân cho biết khách hàng đã có tín dụng mặc định hay chưa

Housing	Biến nhị phân cho biết khách hàng có vay mua nhà hay không
Loan	Biến nhị phân cho biết khách hàng có khoản vay cá nhân hay không
Balance	Số dư cá nhân của khách hàng
Contact	Kiểu liên lạc
Month	Tháng cuối cùng trong năm liên lạc với khách hàng
Day	Ngày cuối cùng trong tuần liên lạc với khách hàng
Duration	Thời lượng trong lần liên lạc lần cuối (đơn vị giây)
Campaign	Số lần liên hệ cho khách hàng trong chiến dịch
Pdays	Số ngày trôi qua kể từ lần liên hệ cuối cùng trong chiến dịch trước
Previous	Số lần liên hệ được thực hiện trước chiến dịch
Poutcome	Kết quả của chiến dịch tiếp thị trước đó
Y	Khách hàng đăng ký tiền gửi có kỳ hạn không?

Bảng 5.1 Mô tả tiêu đề các cột của bộ dữ liệu

### 5.1.3. Câu hỏi nghiên cứu

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

## 5.2. Tiền xử lý dữ liệu

### 5.2.1. Tổng quan cấu trúc bộ dữ liệu

Dữ liệu bao gồm 17 cột với 45211 dòng với các kiểu dữ liệu là int64(7) và object(10).

Index	Column	Non-Null Count	Data type
1	age	45211 non-null	int64
2	job	45211 non-null	object
3	marital	45211 non-null	object
4	education	45211 non-null	object
5	default	45211 non-null	object
6	balance	45211 non-null	int64
7	housing	45211 non-null	object
8	loan	45211 non-null	object
9	contact	45211 non-null	object
10	day	45211 non-null	int64
11	month	45211 non-null	object
12	duration	45211 non-null	int64

13	campaign	45211 non-null	int64
14	pdays	45211 non-null	int64
15	previous	45211 non-null	int64
16	poutcome	45211 non-null	object
17	y	45211 non-null	object

Bảng 5.2 Tổng quan về các biến thuộc bộ dữ liệu

Bảng bên dưới mô tả thống kê của các cột trong bộ dữ liệu. Các thông số này cung cấp những thông tin cơ bản về giá trị nhỏ nhất, giá trị lớn nhất, trung vị,... để tiến hành phân tích và đánh giá chất lượng của bộ dữ liệu

	age	balance	day	duration	campaign	pdays	previous
<b>count</b>	45211	45211	45211	45211	45211	45211	45211
<b>mean</b>	40.93621	1362.272058	15.806419	258.16308	2.763841	40.197828	0.580323
<b>std</b>	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
<b>min</b>	18	-8019	1	0	1	-1	0
<b>0.25</b>	33	72	8	103	1	-1	0
<b>0.5</b>	39	448	16	180	2	-1	0
<b>0.75</b>	48	1428	21	319	3	-1	0

max	95	102127	31	4918	63	871	275
-----	----	--------	----	------	----	-----	-----

Bảng 5.3 Mô tả thống kê

### 5.2.2. Làm sạch dữ liệu

#### 5.2.2.1. Xử lý dữ liệu trùng lặp và không hợp lệ

Dữ liệu trùng lặp và không hợp lệ là giá trị không phù hợp với ngữ cảnh hoặc không thể tham khảo. Kết quả sau khi kiểm tra bằng Python đã cho ra bộ dữ liệu này không có dòng nào bị trùng lặp với nhau, đảm bảo các dòng là duy nhất.

#### 5.2.2.2. Xử lý dữ liệu thiếu (NaN)

Dữ liệu thiếu (NaN) là những dữ liệu sẽ làm sai lệch kết quả phân tích và tạo ra thông tin thừa nên cần được kiểm tra và xử lý loại bỏ phù hợp. Tổng số lượng các giá trị thiếu (NaN) và phần trăm của mỗi cột của bộ dữ liệu được thể hiện lần lượt qua các cột “Total”, “Percent” ở bảng bên dưới.

Column	Total	Percent
age	0	0
day	0	0
poutcome	0	0
previous	0	0
pdays	0	0
campaign	0	0
duration	0	0

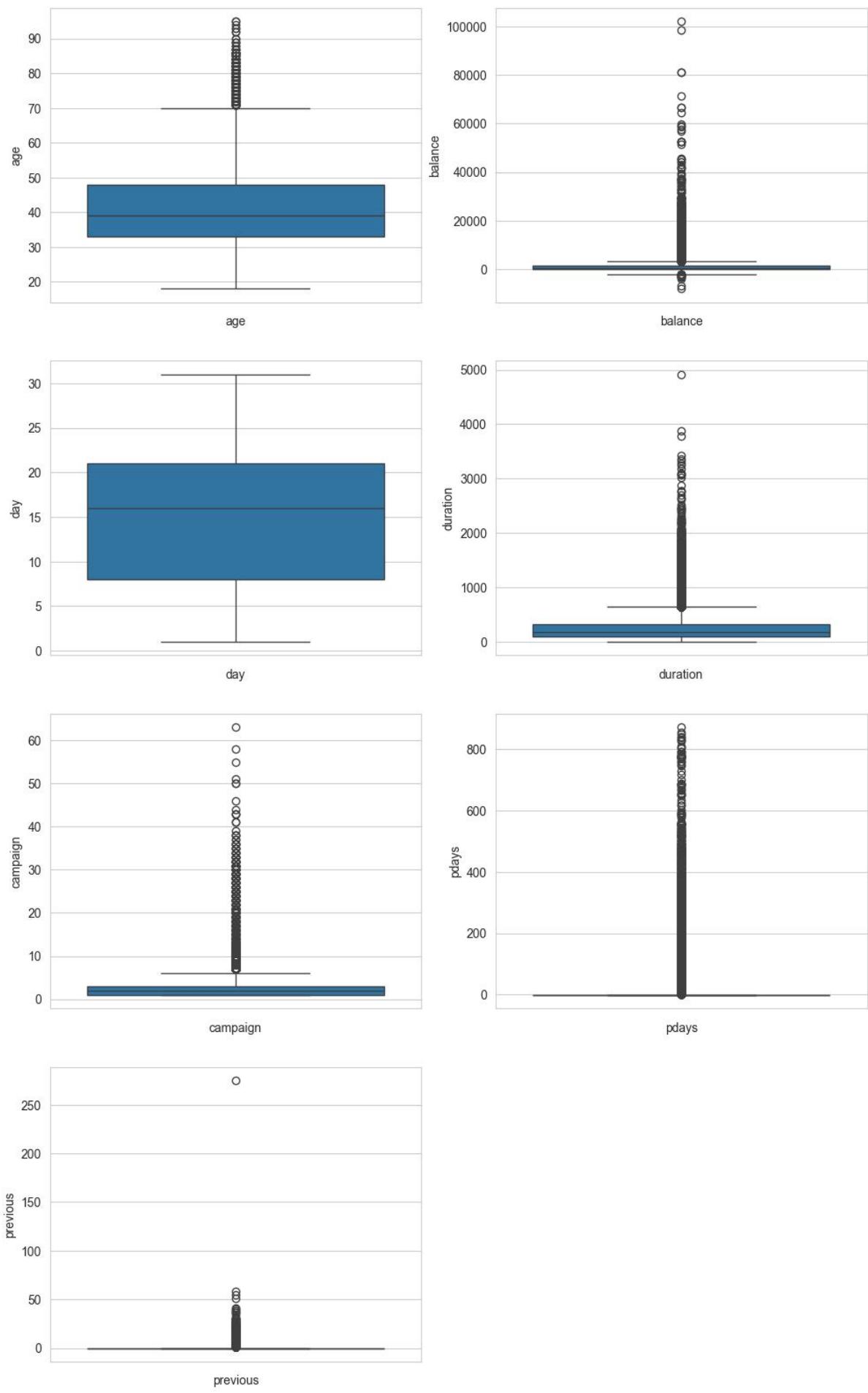
month	0	0
contact	0	0
job	0	0
loan	0	0
housing	0	0
balance	0	0
default	0	0
education	0	0
marital	0	0
y	0	0

Bảng 5.4 Mô tả thống kê các giá trị thiếu

Theo kết quả thu được, ta có thể thấy được bộ dữ liệu này không có dữ liệu thiếu ở bất kỳ cột nào, đảm bảo kết quả phân tích được tạo ra từ bộ dữ liệu là đáng tin cậy.

#### 5.2.2.3. Xử lý ngoại lệ

Tiếp theo, 7 cột có kiểu dữ liệu số sẽ được kiểm tra những giá trị là ngoại lệ thông qua biểu đồ Boxplot. Biểu đồ Boxplot sẽ hiển thị các dữ liệu ngoại lệ thông qua các chấm tròn.



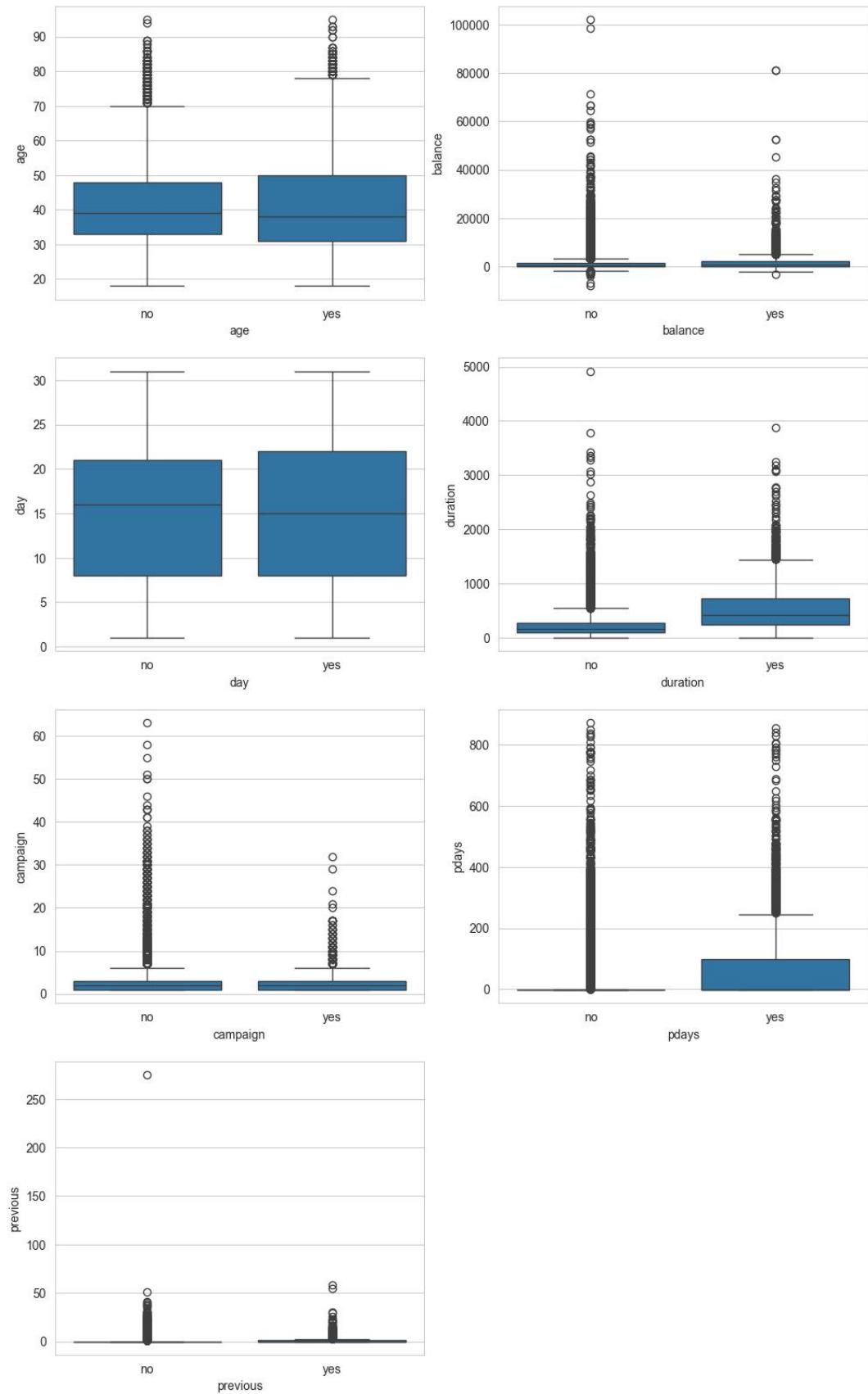
### *Hình 5.1 Biểu đồ Boxplot thể hiện các giá trị ngoại lai*

Nhìn vào biểu đồ, ta có thể thấy hầu hết các cột đều có giá trị ngoại lệ, tuy nhiên vì mục đích của bài toán là phân loại nên vẫn chấp nhận những giá ngoại lệ này để tìm hiểu sâu hơn về đặc điểm chính xác của nhóm đối tượng.

#### **5.2.3. Phân tích khám phá dữ liệu**

##### **5.2.3.1. Biến liên tục**

Các biểu đồ bên dưới thể hiện giá trị của mỗi cột so với giá trị của biến “Y”



Hình 5.2 Biểu đồ thể hiện phân bố của các biến liên tục

### 5.2.3.2. Biến phân loại và biến nhị phân

Bảng bên dưới thống kê số lượng các biến trong một cột

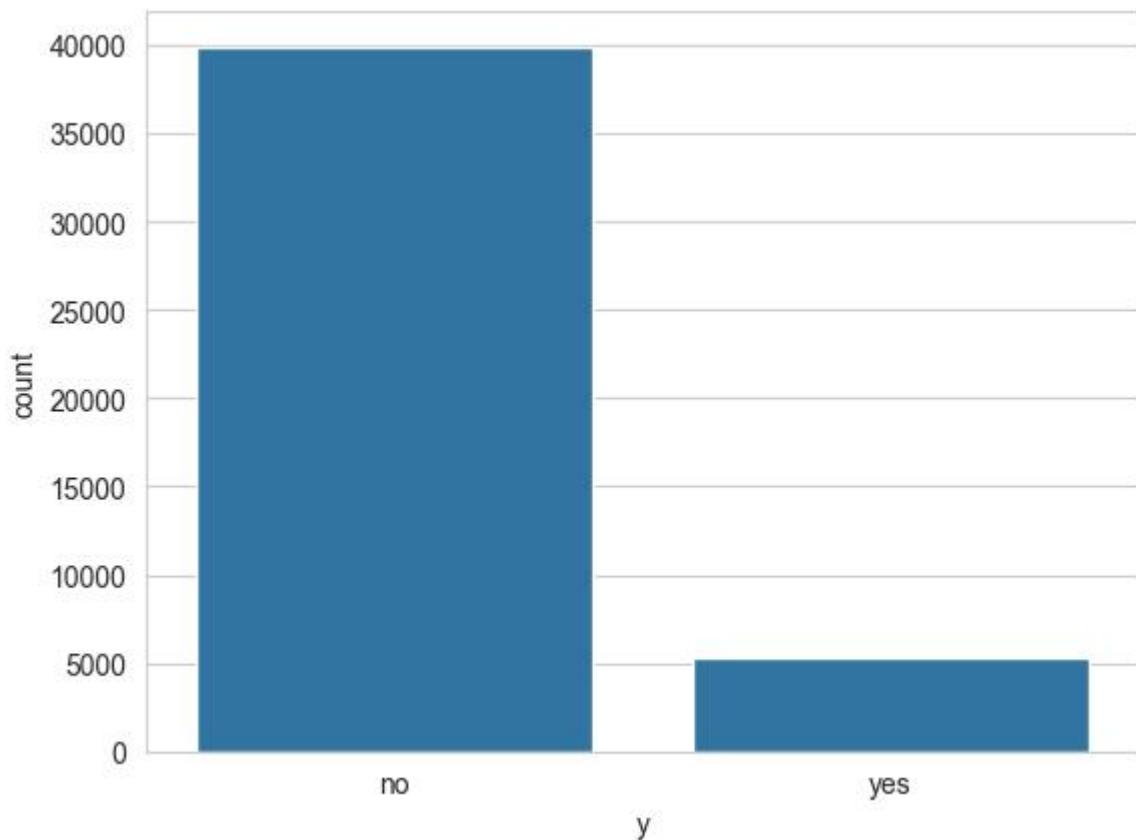
Column	Unique Values	Total Unique Count
job	management, technician, entrepreneur, blue-collar, unknown, retired, admin., services, self-employed, unemployed, housemaid, student	12
marital	married, single, divorced	3
education	tertiary, secondary, unknown, primary	4
default	no, yes	2
housing	yes, no	2
loan	no, yes	2
contact	unknown, cellular, telephone	3
month	may, jun, jul, aug, oct, nov, dec, jan, feb, mar, apr, sep	12
poutcome	unknown, failure, other, success	4
y	no, yes	2

Bảng 5.5 Thống kê các giá trị duy nhất ở mỗi cột

Dựa vào bảng thống kê thì các giá trị trong mỗi cột đều có ít nhất hai giá trị khác nhau và số lượng các giá trị này đều từ 2 trở lên. Điều này đảm bảo dữ liệu trong các cột không chỉ đảm bảo tính đa dạng mà còn tăng cường tính đáng tin cậy của dữ liệu, tạo cơ sở vững chắc cho phân tích dữ liệu sâu hơn.

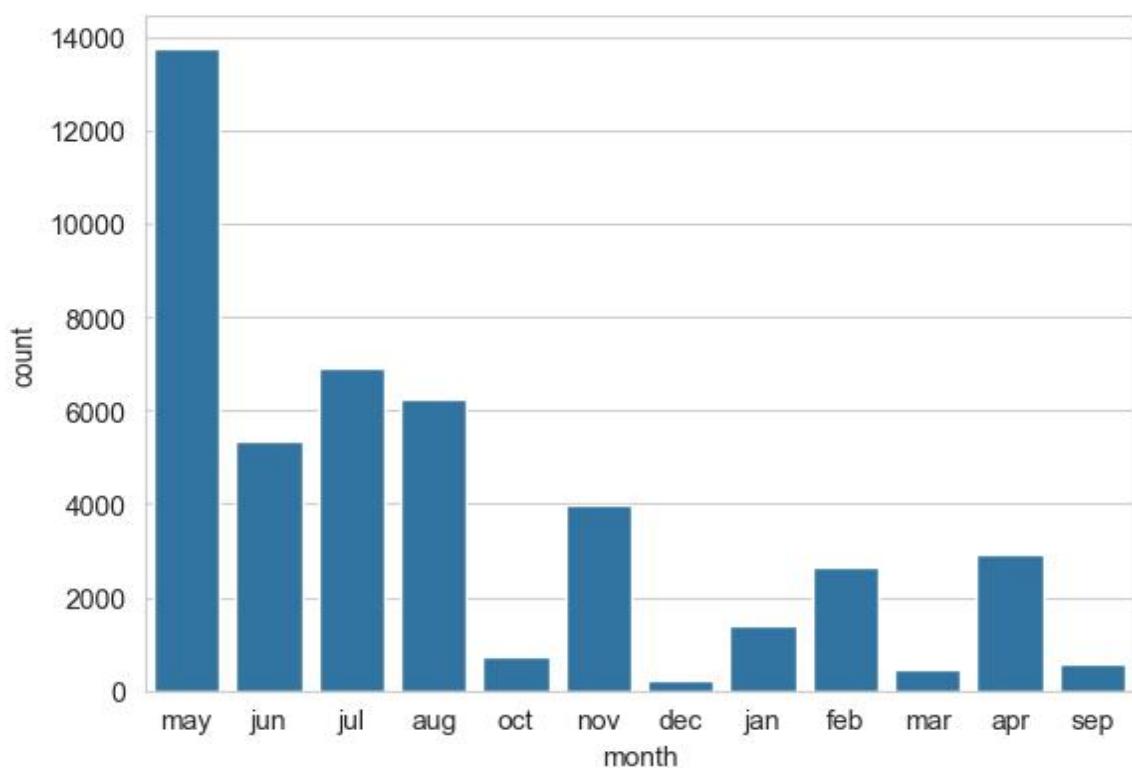
Các biểu đồ bên dưới thể hiện chi tiết số lượng giá trị trong mỗi cột, từ đó giúp hiểu rõ hơn về phân phối và cung cấp cái nhìn chi tiết về mỗi nhóm trong mỗi cột.

### **Biến “Y”**



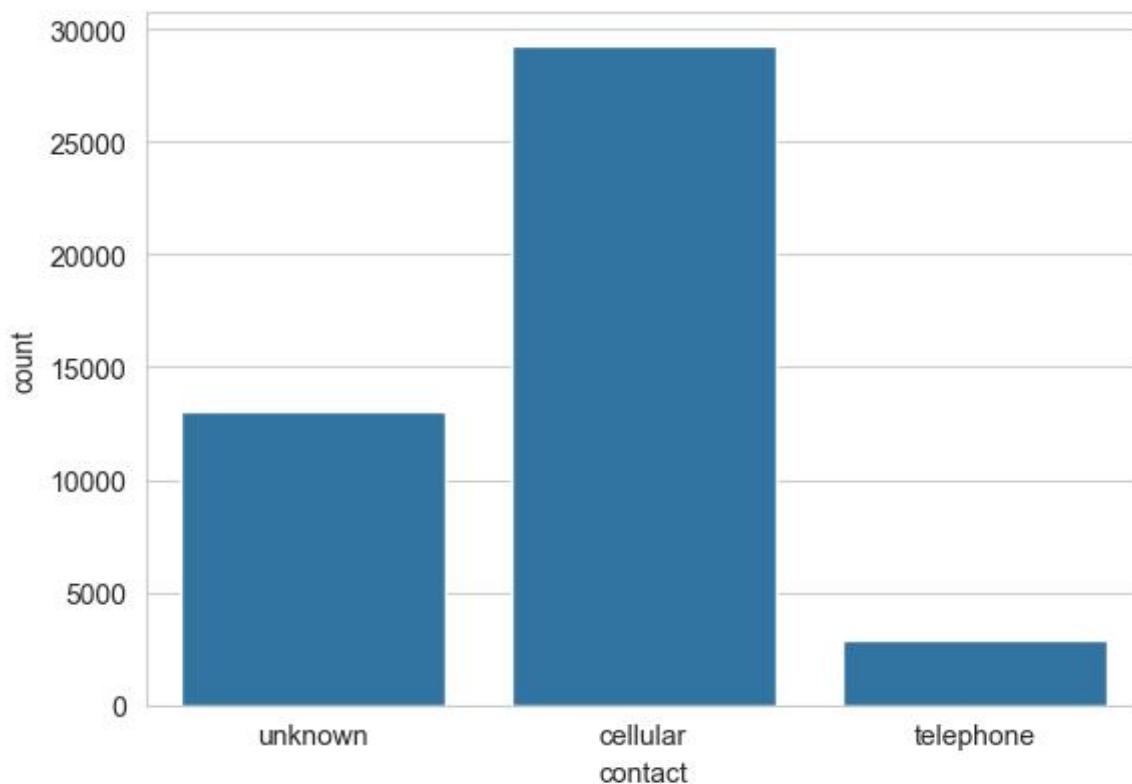
*Hình 5.3 Biểu đồ thể hiện phân bố của biến “Y”*

### **Biến “Month”**



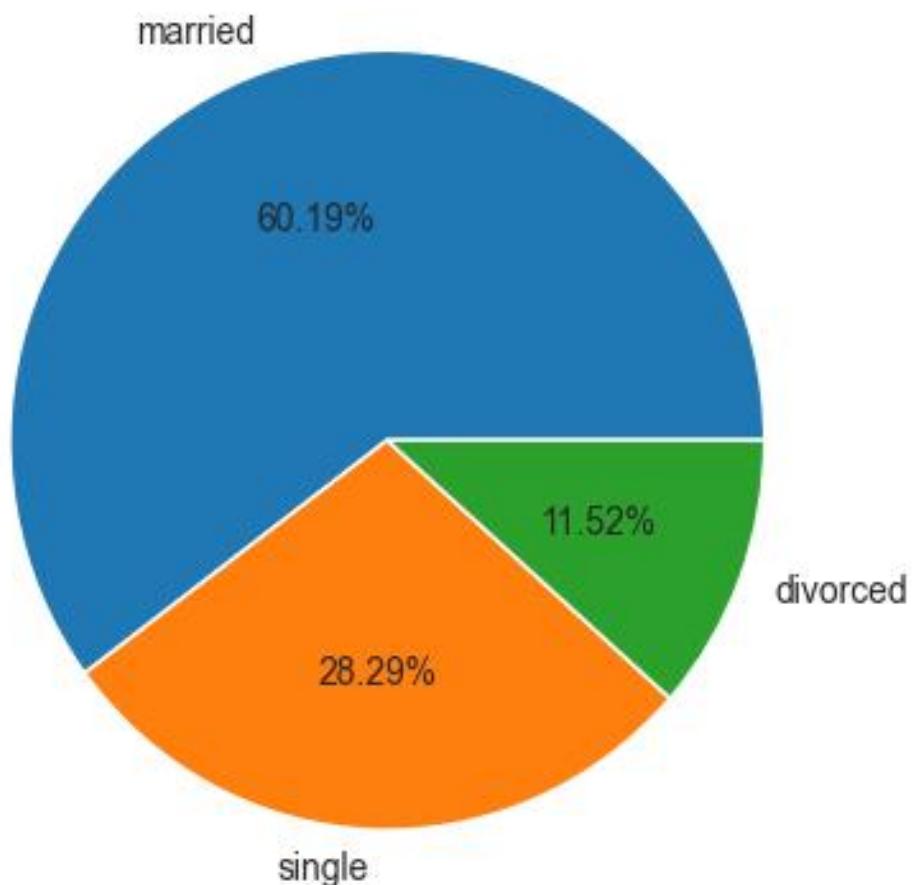
Hình 5.4 Biểu đồ thể hiện phân bố của biến “Month”

### Biến “Contact”



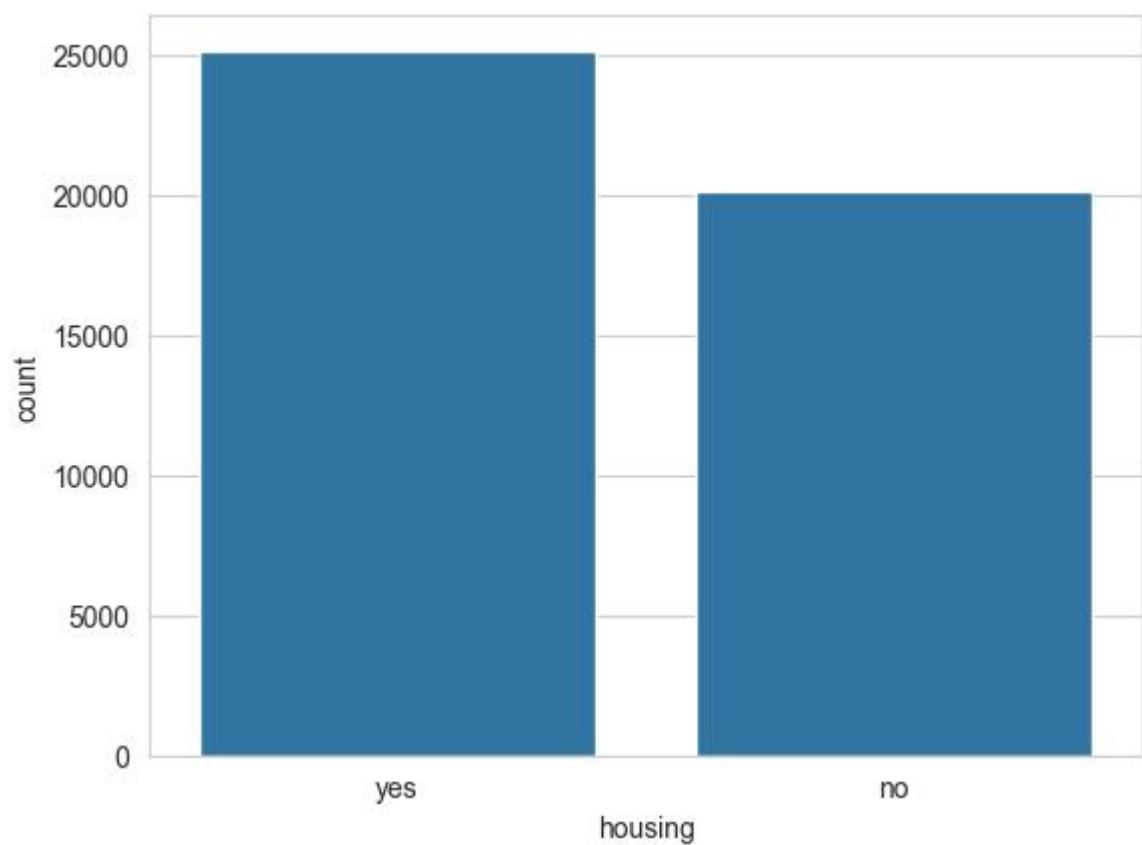
Hình 5.5 Biểu đồ thể hiện phân bố của biến “Contact”

### **Biến “Marital”**



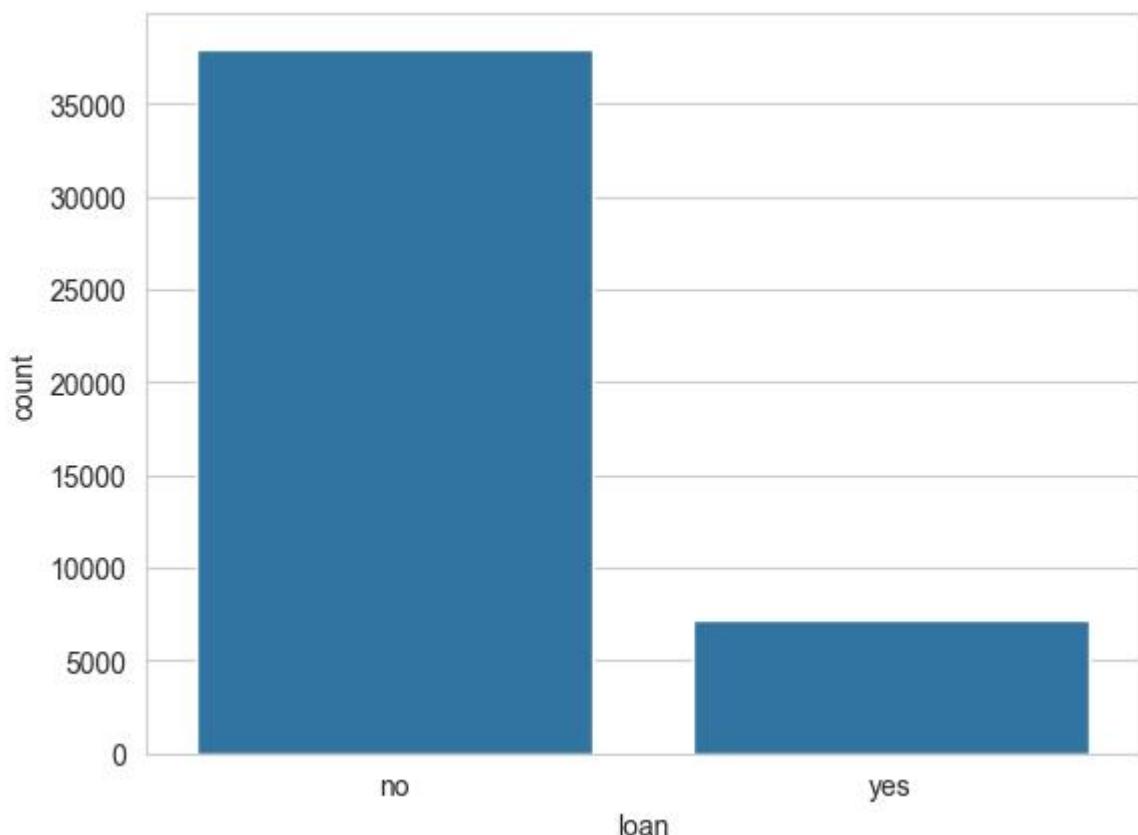
*Hình 5.6 Biểu đồ thể hiện phân bố của biến “Marital”*

### **Biến “Housing”**



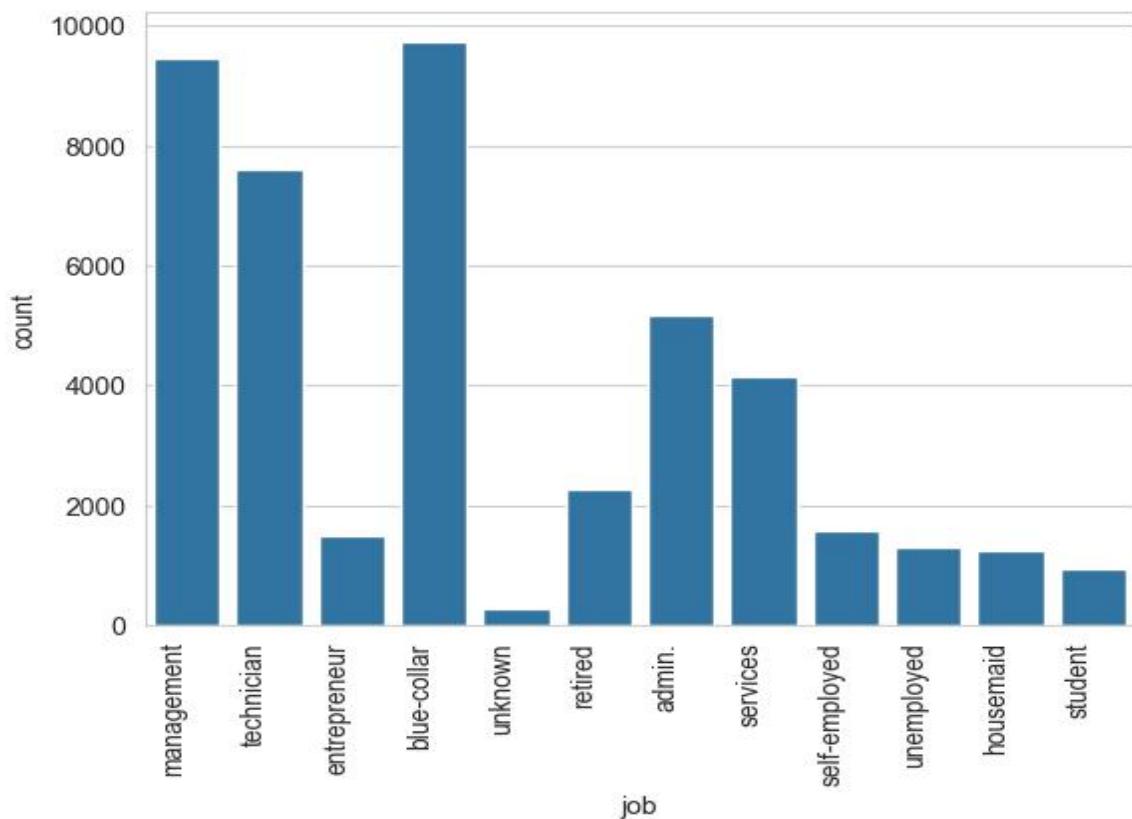
Hình 5.7 Biểu đồ thể hiện phân bố của biến “Housing”

### Biến “Loan”



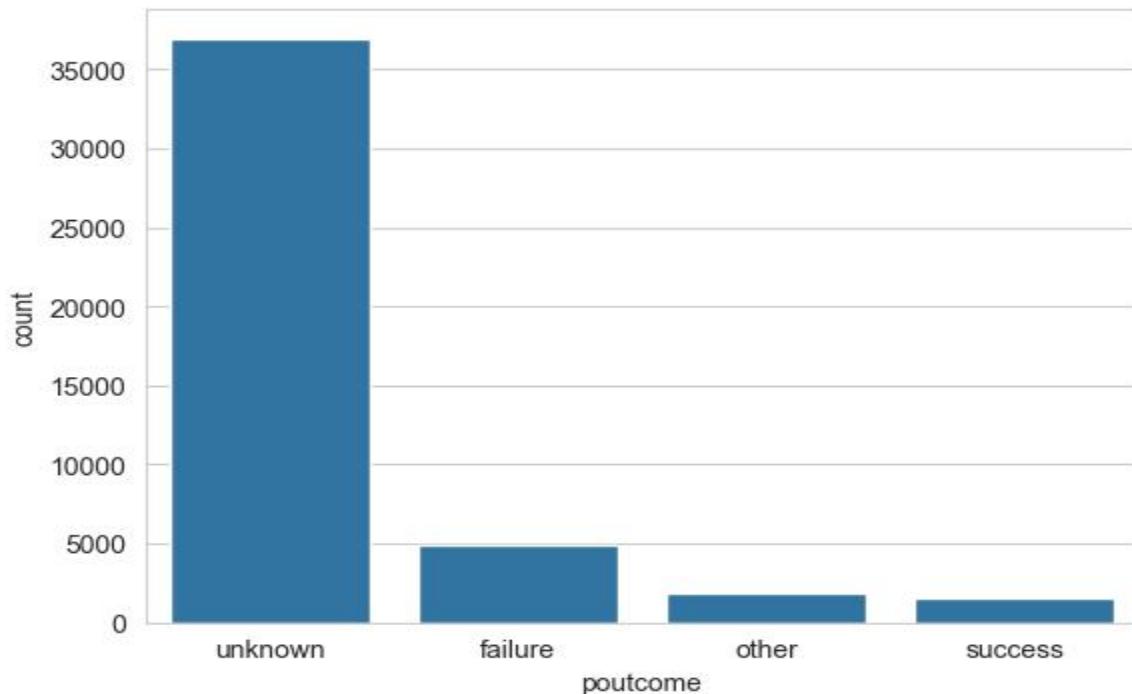
Hình 5.8 Biểu đồ thể hiện phân bố của biến “Loan”

### Biến “Job”



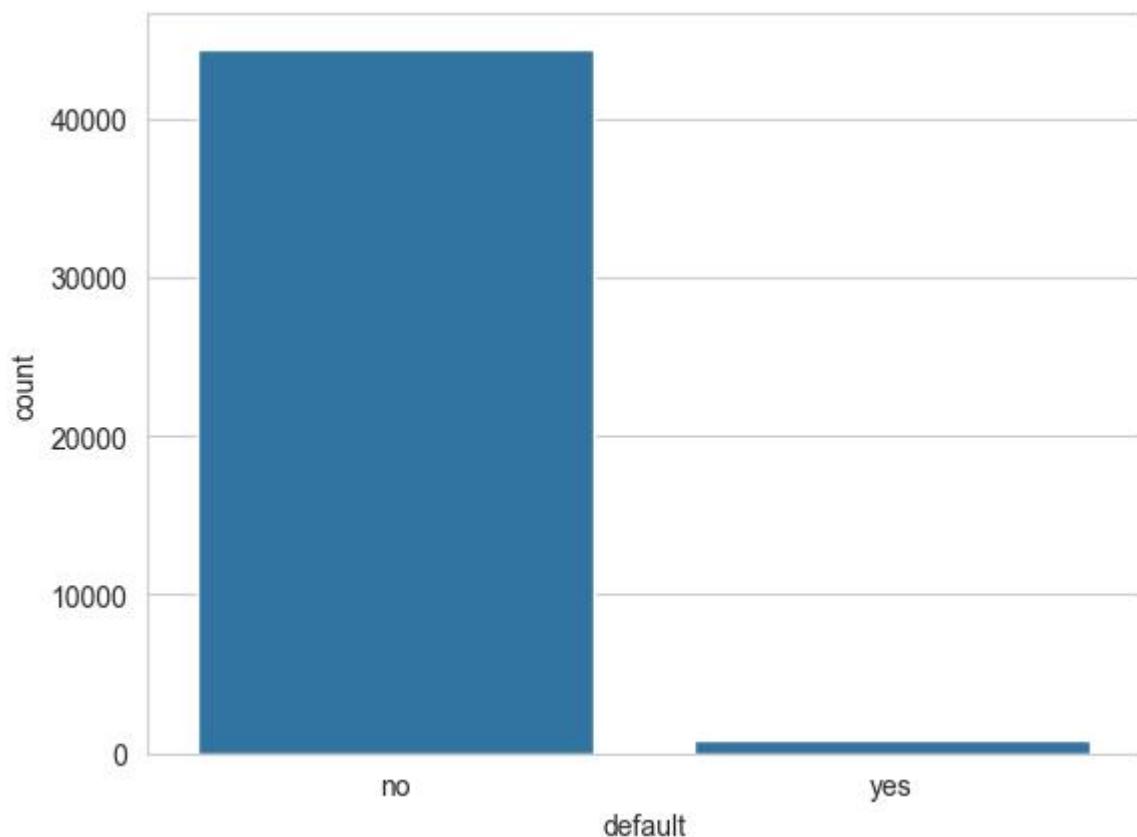
Hình 5.9 Biểu đồ thể hiện phân bố của biến “Job”

### Biến “Poutcome”



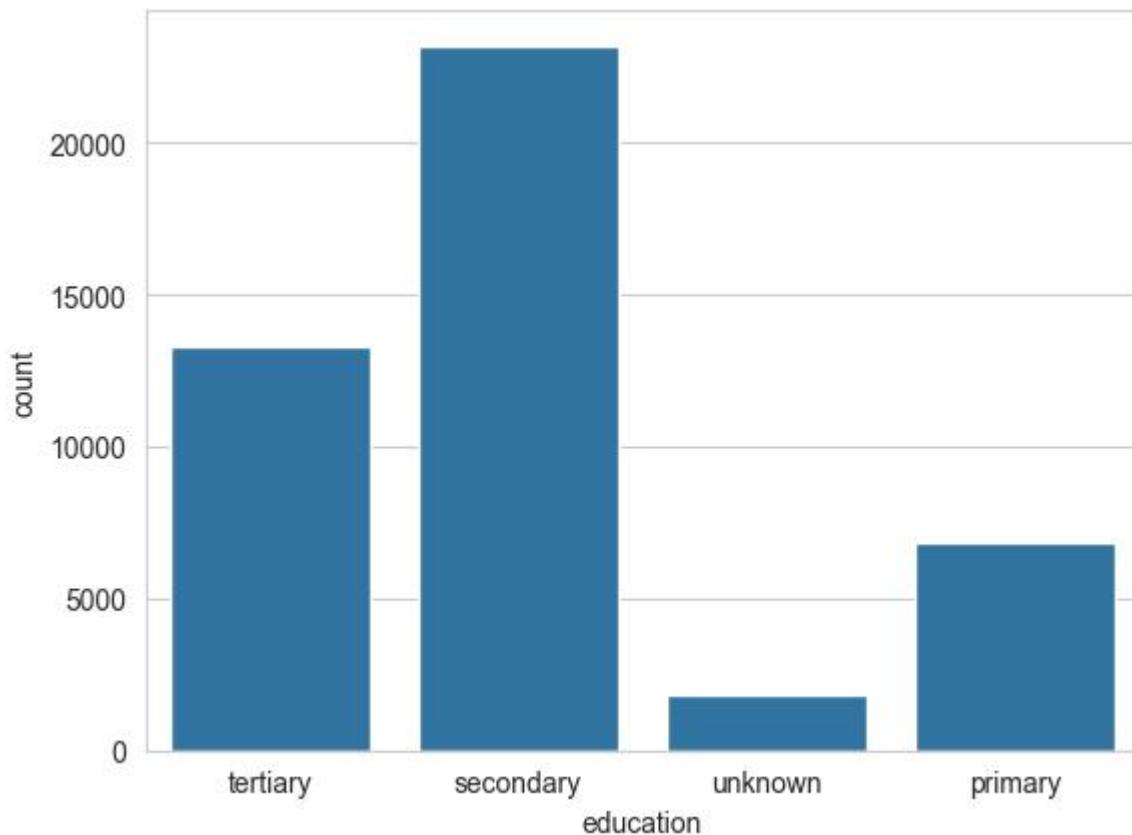
*Hình 5.10 Biểu đồ thể hiện phân bố của biến “Poutcome”*

**Biến “Default”**



*Hình 5.11 Biểu đồ thể hiện phân bố của biến “Default”*

**Biến “Education”**



*Hình 5.12 Biểu đồ thê hiện phân bố của biến “Education”*

### 5.3. Chuẩn bị dữ liệu

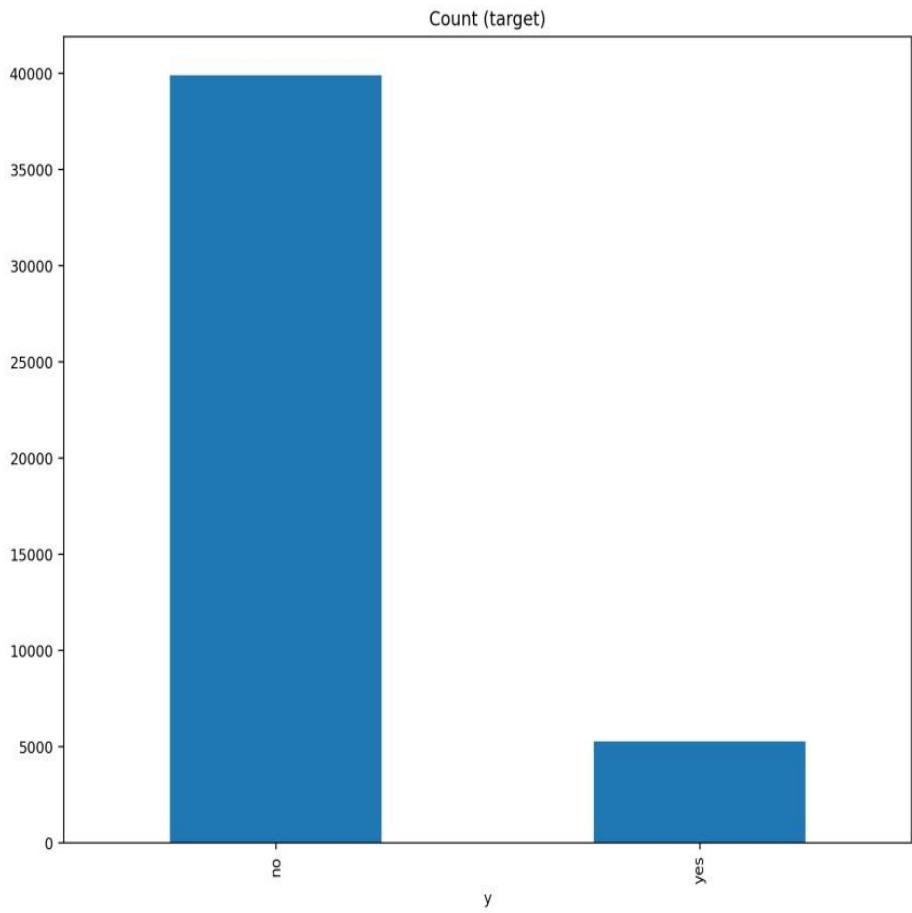
#### 5.3.1. Chuyển đổi dữ liệu

Khi chạy bộ dữ liệu **train.csv** và **test.csv** ta nhận thấy rằng dữ liệu còn một số cột đang ở dạng object (chuỗi) , do đó ta thực hiện chuyển đổi dữ liệu dạng object sang dữ liệu kiểu int (số nguyên). Điều này được thực hiện bằng cách ánh xạ (mapping) các giá trị chuỗi sang các giá trị số nguyên tương ứng.

Sau khi thực hiện mapping ta đếm được số lượng trong cột giá trị “ y ” *khiến khách hàng đăng ký tiền gửi có kỳ hạn:*

- Tổng số mẫu trong lớp 0 (không đăng ký) là 39922 và trong lớp 1 (đăng ký) là 5289.
- Tỉ lệ mẫu trong lớp 0 chiếm khoảng 88.3%, trong khi lớp 1 chỉ chiếm 11.7%.

⇒ Điều này cho thấy mâu cân bằng lớn giữa hai lớp.



Hình 5.13 Số lượng lớp 0 và 1 của bộ dữ liệu trước khi Smoteen

### 5.3.2. Xử lý mất cân bằng dữ liệu

- Tiến hành Scale các chỉ số của các Feature và đưa về cùng 1 thang đo cho việc huấn luyện hiệu quả hơn và hội tụ nhanh hơn.
- Sau đó tiến hành SMOTEENN dữ liệu để vừa kết hợp giữa SMOTE giúp tạo ra các mẫu cho lớp thiểu số (tăng cường mẫu dữ liệu) và ENN loại bỏ các mẫu nhiễu từ các lớp đa số dựa trên K-nearest neighbors.

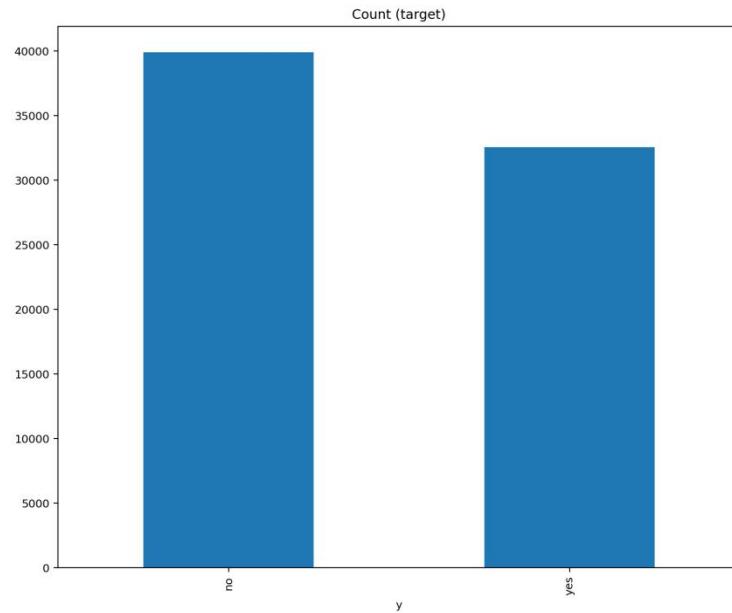
```
smt = SMOTEENN(sampling_strategy='all')
X_resampled, y_resampled= smt.fit_resample(X_scaled, y)
```

#### Kết quả

- Số lượng mẫu trong lớp 0 tăng lên thành 32556 và trong lớp 1 tăng lên thành 38973.

- 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 hơn sau khi sử dụng kỹ thuật SMOTEEN.



*Hình 5.14 Số lượng lớp 0 và 1 của bộ dữ liệu sau khi Smoteen*

### 5.3.3. Phân dữ liệu thành các phần huấn luyện và kiểm thử

Chia dữ liệu thành tập huấn luyện và kiểm tra. 80% dữ liệu sẽ được chọn để tạo thành tập huấn luyện, 20 % còn lại cho tập kiểm tra. Ngoài ra sử dụng tham số random\_state để đảm bảo kết quả chia dữ liệu luôn nhất quán cho việc chạy code ở các lần tiếp theo. Sau đó thực hiện lưu dữ liệu thành 2 file: “*train.csv*” và “*test.csv*”.

## 5.4. Thực hiện các mô hình học máy

### 5.4.1. Decision Tree

#### 5.4.1.1. Tuning mô hình huấn luyện với các tham số ban đầu

Đầu tiên nhằm giúp cho gia tăng, tối ưu được tốc độ xử lý, nhóm đã tìm ra một vài tham số tiêu biểu nhất phù hợp cho việc dự đoán của mô hình, nhóm đã thực hiện tạo kết quả cho các khoảng siêu tham số và rà soát xem sự biến động của chúng cho ra các điểm như thế nào. Dưới đây là tham số mà nhóm đã dùng:

```

param_grid_dt_0 = {'max_depth': [None, 5, 10, 20, 25],
                   'min_samples_split': [2, 5, 10]}

dt_model_0 = DecisionTreeClassifier()

dt_grid = GridSearchCV(dt_model_0,
                       param_grid_dt_0,
                       cv=5)

dt_grid.fit(X_train, y_train)

```

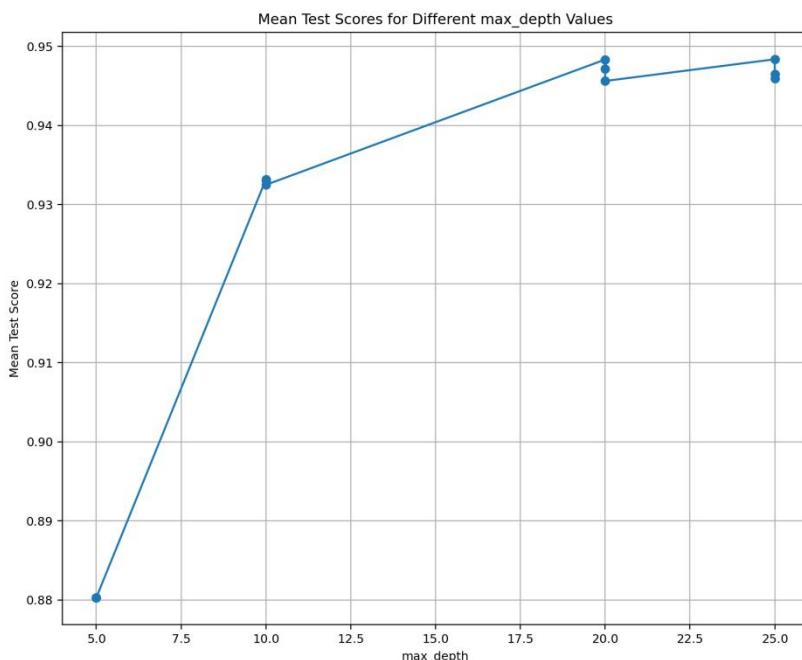
Bảng dưới đây cho thấy sự biến đổi của mean test score khi max\_depth và mean\_samples\_split thay đổi, xu hướng chung là max\_depth nằm ở khoảng nhỏ và mean\_samples\_split càng nhỏ thì điểm càng cao và sẽ càng tối ưu hơn.

<b>Max_depth</b>	<b>Min_samples_split</b>	<b>Mean_test_score</b>
None	2	0.948744
None	5	0.946454
None	10	0.945964
5	2	0.880281
5	5	0.880281
5	10	0.880281
10	2	0.933181
10	5	0.933058
10	10	0.932481
20	2	0.948325
20	5	0.947171

20	10	0.945614
25	2	0.948360
25	5	0.946489
25	10	0.945946

Bảng 5.6 So sánh kết quả giữa các `max_depth` và `min_samples_split` tương ứng

Khi quan sát bảng trên ta thấy, khi cùng một `max_depth`, thì với số lượng mẫu tối thiểu cần thiết để phân chia một nút (`mean_samples_split`) là 2 sẽ cho ra kết quả khả quan nhất.



Hình 5.15 Đồ thị biểu diễn sự hội tụ của mô hình theo các giá trị `max_depth` khác nhau

Sau khi nhìn vào *Đồ thị biểu diễn sự hội tụ của mô hình theo các giá trị `max_depth` khác nhau* ta nhận thấy rằng Score tăng lên đáng kể khi độ sâu của cây nằm trong khoảng từ [5,10] từ 0.88 - 0.93. Và giảm dần độ dốc khi độ sâu tăng dần từ [10, 25]. Đối với `max_depth` từ 10-25 kết quả có sự biến động nhưng kết quả không có sự chênh lệch quá lớn. Do đó để tiết kiệm được tài nguyên và tối ưu được tốc độ xử lý của mô hình. Nhóm quyết định chỉ xét

những trường hợp có ***max\_depth*** trong khoảng từ [5:10] và ***mean\_samples\_split*** là 2:

```
param_grid_dt = {'criterion': ['gini',  
                                'entropy',  
                                'log_loss'],  
  
                 'max_depth': np.arange(5, 10),  
  
                 'min_samples_split': [2],  
  
                 'min_samples_leaf': np.arange(1, 10)  
                }
```

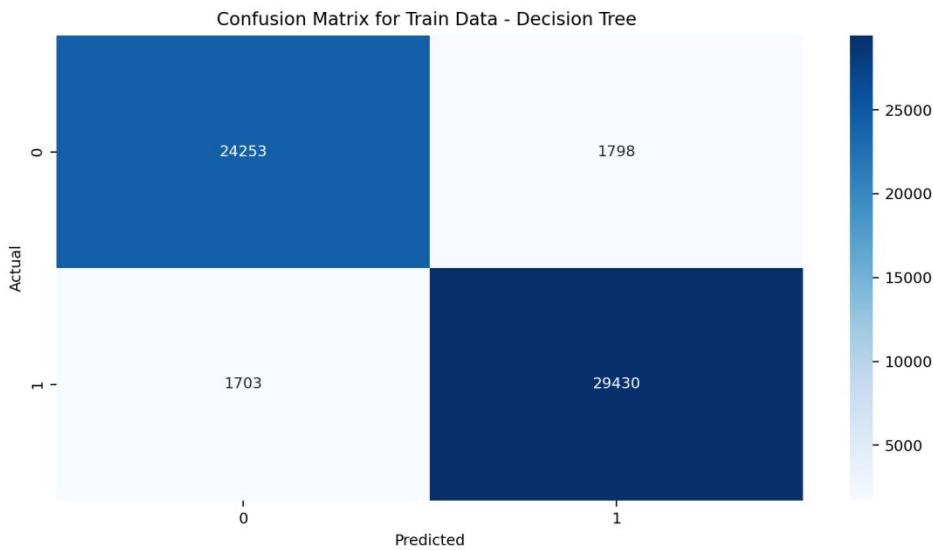
#### 5.4.1.2. Đánh giá mô hình với dữ liệu huấn luyện

##### Số liệu

Classification report for training set				
	precision	recall	f1-score	support
0	0.93	0.93	0.93	26051
1	0.94	0.95	0.94	31133
accuracy			0.94	57184
macro avg	0.94	0.94	0.94	57184
weighted avg	0.94	0.94	0.94	57184

Hình 5.16 Báo cáo số liệu của Decision Tree trên dữ liệu Train

- Mô hình có độ chính xác trên tập Training là khoảng gần 94 %.
- Precision, Recall và F1-score cho cả hai lớp (0 và 1) đều cao, đạt trên 0.93 cho lớp 0 và trên 0.94 cho lớp 1.



*Hình 5.17 Confusion Matrix của Decision Tree cho dữ liệu Train*

Confusion matrix cho thấy có ở dữ liệu huấn luyện có 24253 trường hợp được dự đoán là lớp 0 và thực sự là lớp 0 (True Negative), 29430 trường hợp được dự đoán là lớp 1 và thực sự là lớp 1 (True Positive), có 1703 trường hợp bị dự đoán sai (False Negative) và 1798 trường hợp bị dự đoán sai (False Positive).

### **Nhận xét**

- **Precision**
  - + Đối với lớp 0 (không đăng ký), mô hình đạt được độ chính xác là 93%. Điều này có nghĩa là trong số tất cả các trường hợp mà mô hình dự đoán là không đăng ký, thì có 93% thực sự là không đăng ký.
  - + Trong khi cho lớp 1 (đăng ký) là 94%. Cụ thể, trong số tất cả các trường hợp mà mô hình dự đoán là đăng ký, thì có 94% thực sự là đăng ký. Điều này chỉ ra rằng mô hình đạt được tỷ lệ cao hơn trong việc dự đoán các trường hợp đăng ký gửi tiền so với các trường hợp không đăng ký gửi tiền.
- **Recall:** Tỷ lệ các trường hợp thực sự thuộc một lớp mà mô hình dự đoán đúng. Recall cho lớp 0 và 1 lần lượt là 93 % và 95%, điều này chỉ ra rằng mô hình hiệu quả trong việc bắt trọn những người không đăng ký gửi tiền và những người đăng ký gửi tiền. Tuy nhiên việc dự đoán những người đăng ký gửi tiền có phần tốt hơn.

- **F1-score:** Trung bình điều hòa giữa precision và recall. F1-score cho lớp 0 là 93%, và cho lớp 1 chỉ là 94%. Điều này cho thấy mô hình đạt được sự cân bằng tốt giữa Precision và Recall trên cả hai lớp.

#### 5.4.1.3. Đánh giá mô hình với dữ liệu kiểm tra

##### Số liệu

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.92	0.93	6525
1	0.94	0.94	0.94	7772
accuracy			0.93	14297
macro avg	0.93	0.93	0.93	14297
weighted avg	0.93	0.93	0.93	14297

Confusion Matrix:

```
[[6018 507]
 [ 443 7329]]
```

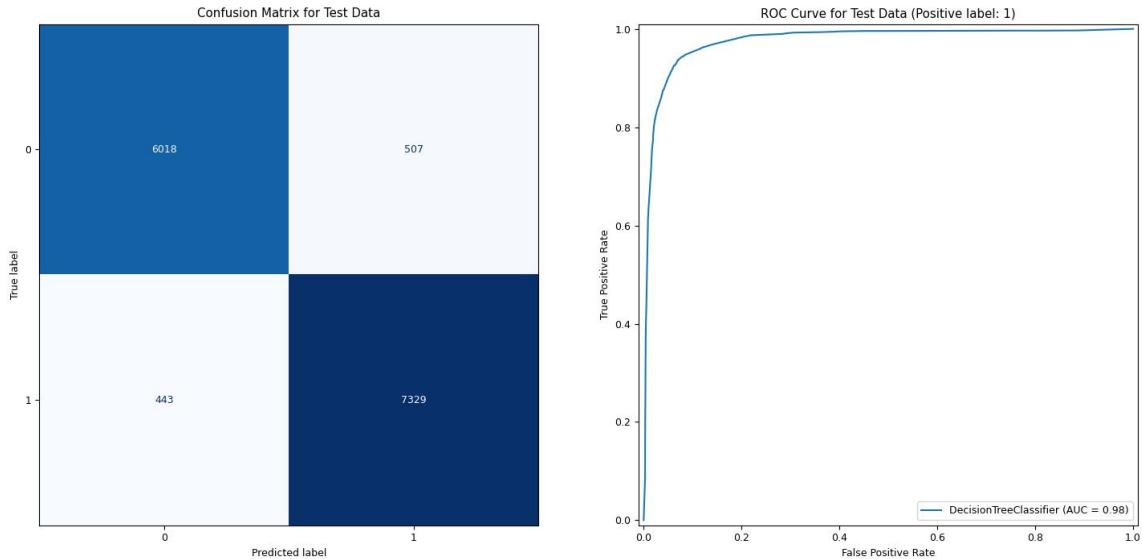
ROC AUC Score: 0.9761791912415725

Decision Tree Training Accuracy: 0.939

Decision Tree Test Accuracy: 0.934

Hình 5.18 Báo cáo số liệu của Decision Tree trên dữ liệu Test

- Mô hình có độ chính xác trên tập Testing là khoảng gần 93%.
- Precision, Recall và F1-score cho cả hai lớp (0 và 1) đều cao, đạt trên 92% cho lớp 0 và trên 94% cho lớp 1.



*Hình 5.19 Confusion Matrix và Biểu đồ ROC Curve Của Decision Tree cho dữ liệu Test*

Confusion matrix cho thấy có ở dữ liệu huấn luyện có 6018 trường hợp được dự đoán là lớp 0 và thực sự là lớp 0 (True Negative), 7329 trường hợp được dự đoán là lớp 1 và thực sự là lớp 1 (True Positive), có 443 trường hợp bị dự đoán sai (False Negative) và 507 trường hợp bị dự đoán sai (False Positive).

Đường cong ROC được vẽ để đánh giá khả năng phân loại của mô hình, và ROC AUC Score đạt khoảng 0.98, cho thấy mô hình **có khả năng phân loại tốt**.

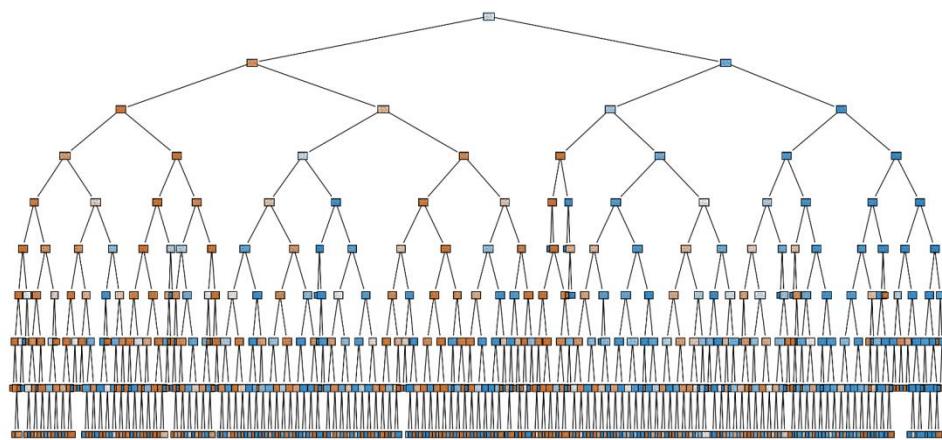
#### Nhận xét:

- **Precision:**
  - + Đối với lớp 0 (không đăng ký), mô hình đạt được độ chính xác là 93%. Điều này có nghĩa là trong số tất cả các trường hợp mà mô hình dự đoán là không đăng ký, thì có 93% thực sự là không đăng ký.
  - + Trong khi cho lớp 1 (đăng ký) là 94%. Cụ thể, trong số tất cả các trường hợp mà mô hình dự đoán là đăng ký, thì có 94% thực sự là đăng ký. Điều này chỉ ra rằng mô hình đạt được tỷ lệ cao hơn trong việc dự đoán các trường hợp đăng ký gửi tiền so với các trường hợp không đăng ký gửi tiền.
- **Recall:** Tỷ lệ các trường hợp thực sự thuộc một lớp mà mô hình dự đoán đúng. Recall cho lớp 0 và 1 lần lượt là 92 % và 94%, điều này chỉ ra rằng mô hình hiệu quả trong việc bắt trọn những người không đăng ký gửi tiền và những

người đăng ký gửi tiền. Tuy nhiên việc dự đoán những người đăng ký gửi tiền có phần tốt hơn.

- **F1-score:** Trung bình điêu hòa giữa precision và recall. F1-score cho lớp 0 là 93%, và cho lớp 1 chỉ là 94%. Điều này cho thấy mô hình đạt được sự cân bằng tốt giữa Precision và Recall trên cả hai lớp.

Dưới đây là hình cây quyết định được vẽ ra từ dữ liệu test thông qua mô hình đã được xác định ở trên:



Hình 5.20 Trực quan cây quyết định với  $\text{max\_depth} = 9$

#### 5.4.1.4. Tổng kết

*Decision Tree Training Accuracy: 0.939*

*Decision Tree Test Accuracy: 0.934*

Mô hình đã đạt được độ chính xác khoảng 93.4 % trên tập kiểm tra, và cũng có được kết quả gần như tương tự với độ chính xác trên tập huấn luyện là 93.9 %. Điều này cho thấy mô hình không bị quá mức (**overfitting**) và **có khả năng tổng quát hóa tốt** trên dữ liệu mới.

### 5.4.2 Random Forest

#### 5.4.2.1. Tìm hiểu tham số và huấn luyện

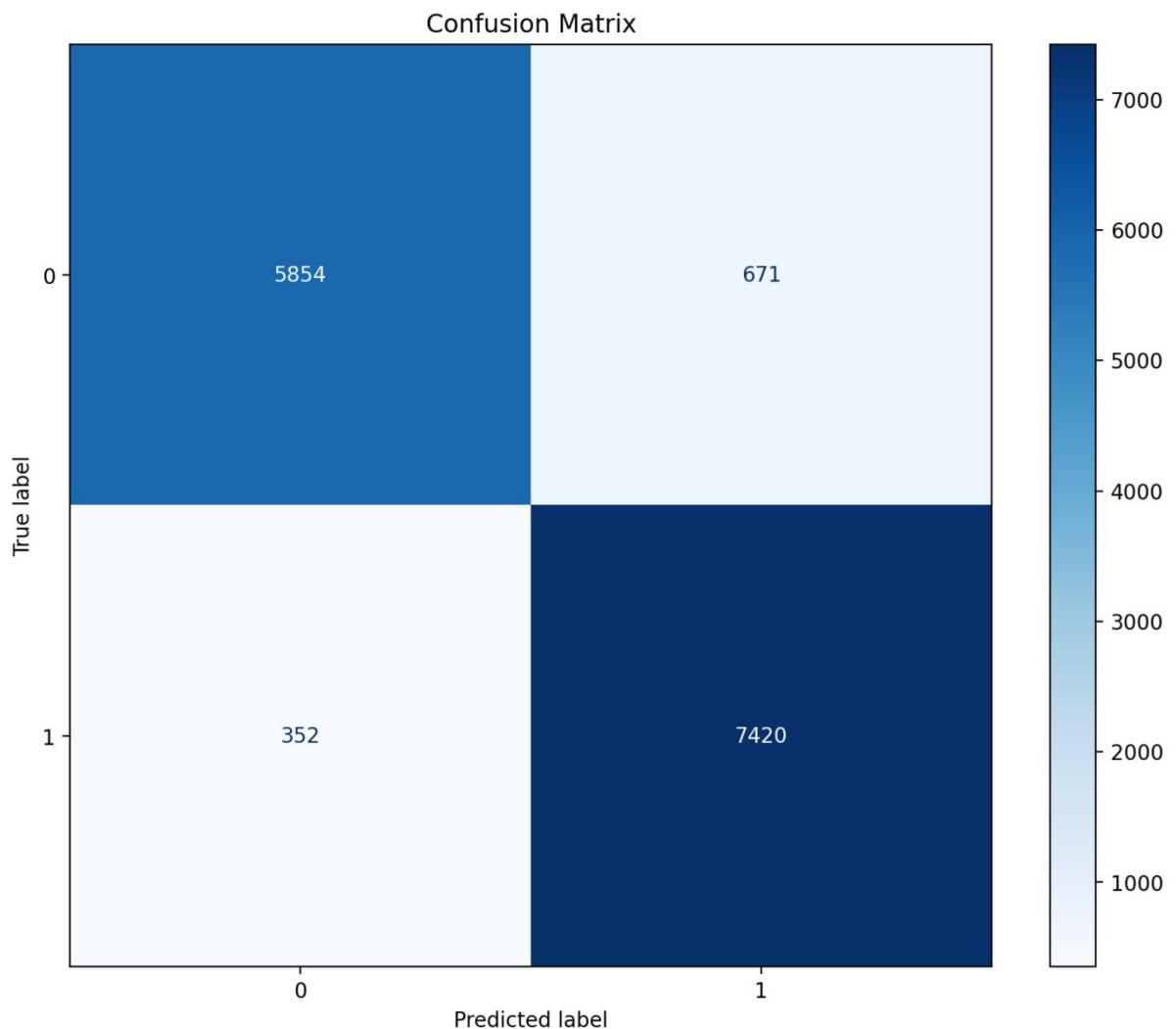
Để phù hợp và tiện so sánh, đánh giá với các mô hình khác, trong Random Forest nhóm quyết định sử dụng tham số  $\text{max\_depth} = 9$ . Lý do cho việc lựa chọn tham số  $\text{max\_depth} = 9$  là vì trong mô hình Decision Tree, nhóm

đã đưa ra đánh giá và cho rằng tham số đó phù hợp với bộ dữ liệu này. Tiếp theo đó, nhóm sử dụng GridSearchCV để tìm ra các tham số cho mô hình phù hợp với bộ dữ liệu.

```
param_grid = {'n_estimators': [50, 100, 150],  
             'min_samples_split': [2, 5, 10],  
             'min_samples_leaf': [1, 2, 4],  
             'max_depth': [7]}  
  
rf_model = RandomForestClassifier(random_state=42,  
                                   max_depth=7)  
  
grid_search = GridSearchCV(estimator=rf_model,  
                           param_grid=param_grid,  
                           cv=5)  
  
grid_search.fit(X_train, y_train)
```

Nhờ đó, nhóm đã xác định được các tham số khác phù hợp với bộ dữ liệu như: min\_samples\_leaf=4, min\_samples\_split=10, n\_estimators=50. Tham số max\_depth xác định độ sâu tối đa của mỗi cây trong rừng, và giá trị 9 cho thấy mỗi cây có độ sâu tối đa là 9, giá trị này nhằm giới hạn độ sâu của cây, tránh việc xảy ra overfitting và cũng tiết kiệm được thời gian trong quá trình training. Đồng thời, min\_samples\_leaf là số lượng mẫu tối thiểu cần có trong mỗi lá của cây, và giá trị 4 cho biết mỗi lá phải chứa ít nhất 4 mẫu, tham số này cũng giới hạn số lượng tối thiểu mỗi mẫu, tránh việc trong lá đó có quá ít mẫu. Tham số min\_samples\_split xác định số lượng mẫu tối thiểu cần có trong mỗi nút để chia nút thành các nút con, và giá trị 10 ý nghĩa là mỗi nút phải có ít nhất 10 mẫu để chia. Cuối cùng, n\_estimators là số lượng cây trong rừng, và giá trị 50 cho biết mô hình sẽ sử dụng một rừng gồm 50 cây để thực hiện dự đoán.

#### 5.4.2.2. Đánh giá mô hình



Hình 5.21 Confusion Matrix của Random Forest cho dữ liệu Test

Confusion matrix cho thấy có 5854 trường hợp được dự đoán là lớp 0 và thực sự là lớp 0 (True Negative), 7420 trường hợp được dự đoán là lớp 1 và thực sự là lớp 1 (True Positive), có 362 trường hợp bị dự đoán là lớp 0 trong khi thực sự là lớp 1 (False Negative) và 671 trường hợp bị dự đoán là lớp 1 khi thực tế nằm ở lớp 0 (False Positive).

Classification Report:					
	precision	recall	f1-score	support	
0	0.94	0.90	0.92	6525	
1	0.92	0.95	0.94	7772	
accuracy			0.93	14297	
macro avg	0.93	0.93	0.93	14297	
weighted avg	0.93	0.93	0.93	14297	

Accuracy on training set: 0.9283365976496922

Accuracy on test set: 0.9284465272434776

*Hình 5.22 Báo cáo số liệu của Random Forest trên dữ liệu Test*

Kết quả của mô hình phân loại trên tập kiểm tra cho thấy một hiệu suất đáng chú ý và cân nhắc. Với mức độ chính xác (accuracy) đạt khoảng 92.8%, mô hình đã chứng minh khả năng dự đoán đúng đối với hầu hết các trường hợp trong tập kiểm tra. Tuy nhiên, khi phân tích chi tiết hơn, ta thấy tỷ lệ phát hiện (recall) đối với cả hai lớp đều khá ấn tượng, đạt 90% cho lớp negative (0) và 95% cho lớp positive (1). Điều này cho thấy mô hình có khả năng phát hiện các trường hợp tích cực với hiệu suất cao.

Mặc dù vậy, khi xem xét tỷ lệ precision, nhận thấy mô hình đôi khi dự đoán sai lớp của một số mẫu. Độ chính xác của lớp negative (0) là khoảng 94%, trong khi đó chỉ là khoảng 92% đối với lớp positive (1). Điều này ngụ ý rằng, một số trường hợp dương tích cực được dự đoán sai là âm tích cực và ngược lại. Tuy nhiên, chỉ số precision nằm ở mức khá ổn, cho thấy mô hình vẫn hoạt động hiệu quả trong việc phân loại.

Tổng hợp, 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. Điều này cho thấy mô hình 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.

### 5.4.3. Neural Network

#### 5.4.3.1. Tìm hiểu tham số và huấn luyện

Nhóm sử dụng RandomizedSearchCV nhằm tìm ra tổ hợp siêu tham số tốt nhất. Một số tham số được nhóm chọn là hidden\_layer\_sizes: kích cỡ của lớp ẩn nằm giữa lớp đầu vào và đầu ra, activation.

```
param_dist = {'hidden_layer_sizes': [(50, 50),  
                                      (100,),  
                                      (50,100),  
                                      (100,50)],  
  
              'activation': ['tanh'],  
  
              'solver': ['adam'],  
  
              'alpha': [0.001, 0.01, 0.1],  
  
              'learning_rate': ['constant',  
                                'adaptive'],  
  
              }  
  
nn_model = MLPClassifier(random_state=42)  
  
random_search = RandomizedSearchCV(  
                                      nn_model,  
                                      param_distributions=param_dist,  
                                      n_iter=25,  
                                      cv=5,  
                                      scoring='accuracy',  
                                      random_state=42)  
  
random_search.fit(X_train, y_train)
```

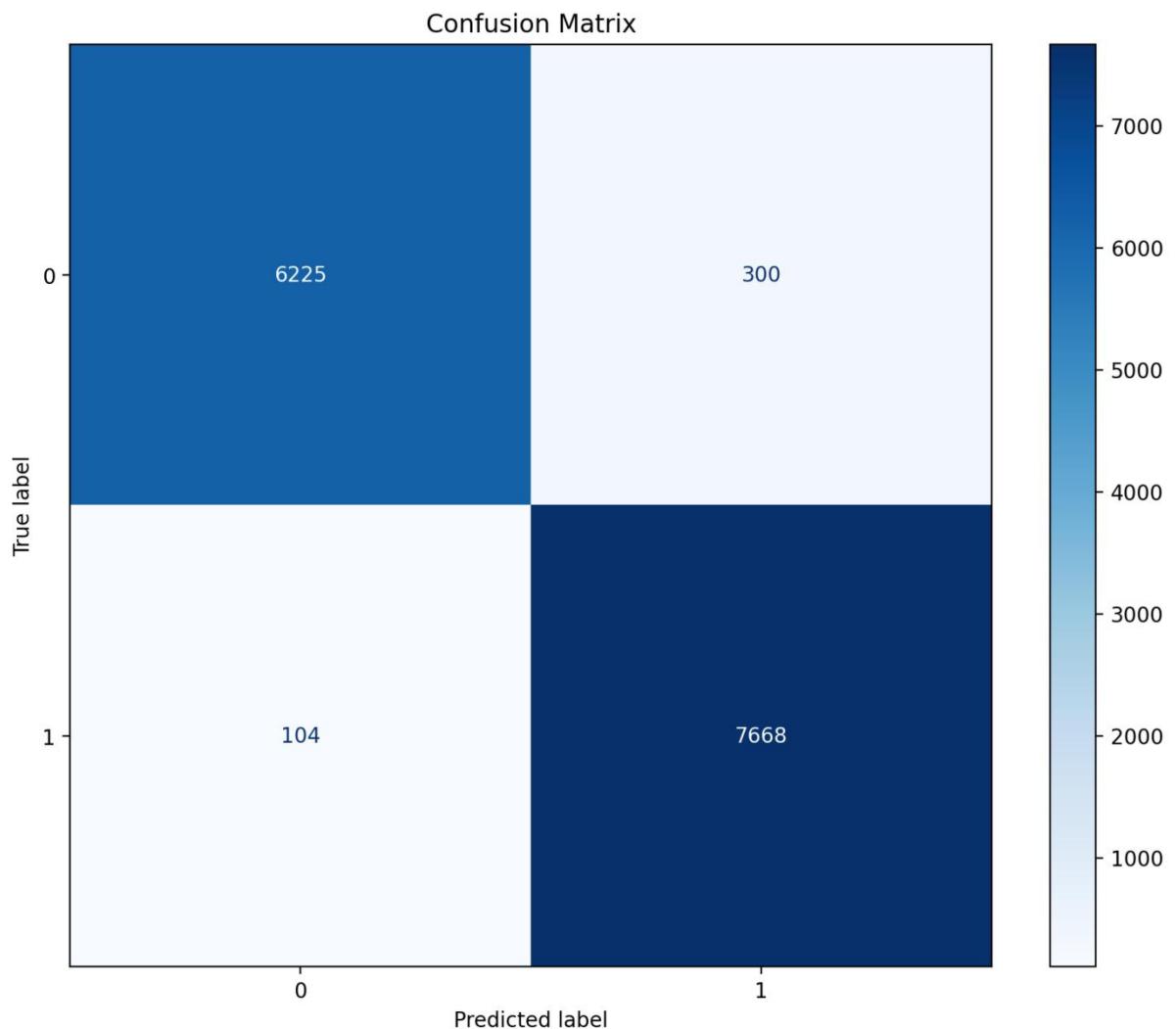
Mô hình MLPClassifier được định cấu hình với các tham số được chọn cẩn thận để tối ưu hiệu suất của nó trong quá trình huấn luyện và dự đoán. Cụ

thể, việc sử dụng hàm kích hoạt tanh (`activation='tanh'`) là một điểm mạnh, vì hàm tanh là một hàm phi tuyến rất phổ biến trong mạng nơ-ron nhân tạo. Hàm tanh giúp mô hình học được các biểu diễn phi tuyến tính và giảm thiểu các giá trị đầu ra trong khoảng từ -1 đến 1, điều này có thể giúp tăng khả năng học của mô hình trên các dữ liệu phức tạp.

Thêm vào đó, việc sử dụng tham số điều chuẩn (`alpha=0.01`) cũng là một điểm mạnh của mô hình. Tham số alpha giúp kiểm soát overfitting trong quá trình huấn luyện bằng cách giảm thiểu sự phụ thuộc quá mức vào dữ liệu huấn luyện. Điều này giúp cải thiện khả năng tổng quát hóa của mô hình và tránh hiện tượng overfitting trên dữ liệu kiểm tra.

Cuối cùng, việc chọn kiến trúc mạng với hai lớp ẩn, mỗi lớp có 50 nơ-ron (hidden\_layer\_sizes=(50, 50)) cũng là một lựa chọn phù hợp với bộ dữ liệu. Kiến trúc này có độ phức tạp trung bình, đủ để mô hình có thể học được các biểu diễn phức tạp hơn so với một mạng nơ-ron với ít lớp ẩn hoặc ít nơ-ron, nhưng không quá phức tạp để gây ra hiện tượng overfitting. Điều này giúp mô hình cân bằng giữa việc học từ dữ liệu và khả năng tổng quát hóa trên dữ liệu kiểm tra.

#### 5.4.3.2. Đánh giá mô hình



Hình 5.23 Confusion Matrix của Neural Network cho dữ liệu Test

Confusion matrix cho thấy có 6225 trường hợp được dự đoán là lớp 0 và thực sự là lớp 0 (True Negative), 7668 trường hợp được dự đoán là lớp 1 và thực sự là lớp 1 (True Positive), có 104 trường hợp bị dự đoán là lớp 0 trong khi thực sự là lớp 1 (False Negative) và 300 trường hợp bị dự đoán là lớp 1 khi thực tế nằm ở lớp 0 (False Positive).

```

Classification Report:
precision    recall   f1-score   support

          0       0.98      0.95      0.97      6525
          1       0.96      0.99      0.97      7772

   accuracy                           0.97      14297
macro avg       0.97      0.97      0.97      14297
weighted avg    0.97      0.97      0.97      14297

Accuracy on training set: 0.9908540850587577
Accuracy on test set: 0.9717423235643842

```

*Hình 5.24 Báo cáo số liệu của Neural Network trên dữ liệu Test*

Sau khi xem xét Classification Report, ta thấy các chỉ số precision, recall, và f1-score cho cả hai lớp đều ở mức rất cao. Đối với lớp âm tích cực (0), tỷ lệ precision là 98% và tỷ lệ recall là 95%. Điều này ngụ ý rằng, trong số các mẫu được dự đoán là âm tích cực, 98% thực sự là đúng và mô hình có thể phát hiện được 95% của tổng số mẫu âm tích cực. Đối với lớp dương tích cực (1), tỷ lệ precision là 96% và tỷ lệ recall là 99%. Điều này ngụ ý rằng, trong số các mẫu được dự đoán là dương tích cực, 96% thực sự là đúng và mô hình có khả năng phát hiện được 99% của tổng số mẫu dương tích cực.

Tổng thể, 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. Mặc dù còn một số mẫu bị dự đoán sai lớp, nhưng tỷ lệ các chỉ số precision, recall, và f1-score cho cả hai lớp đều ở mức cao, cho thấy mô hình có khả năng phân loại hiệu quả trên cả hai lớp.

#### 5.4.4. Support Vector Machine

##### 5.4.4.1. Tạo mô hình phân loại SVM

Nhằm giúp mô hình nâng cao độ chính xác và đáng tin cậy hơn nhóm đã thực hiện việc tối ưu hóa tham số cho mô hình SVM. Quá trình tối ưu hóa tham số được thực hiện thông qua phương pháp tìm kiếm ngẫu nhiên (*Randomized*

*Search*). Tuy nhiên để giảm thời gian chạy, nhóm chỉ thử nghiệm với số vòng lặp là  $n\_iter=15$ .

Trong phạm vi thử nghiệm, nhóm đã sử dụng 2 loại kernel là: '***rbf***' , '***sigmoid***' và '***polynomial***' cho mô hình dự đoán:

- RBF (Radial Basis Function) Kernel: Kernel RBF là một kernel phi tuyến có thể ánh xạ dữ liệu vào một không gian có số chiều vô hạn với các hàm sóng hình cầu. Kernel này thích hợp cho các bài toán có dữ liệu không tuyến tính và không xác định rõ ràng như trong dữ liệu của nhóm
- Sigmoid Kernel: Kernel sigmoid là một kernel phi tuyến giúp xử lý các bài toán phân loại phi tuyến và tương tự như mạng neural có một lớp ẩn với hàm sigmoid.
- Polynomial Kernel: Kernel đa thức được sử dụng để ánh xạ dữ liệu vào một không gian đặc trưng có số chiều cao hơn, cho phép SVM xử lý các quyết định phân loại phi tuyến.

Ngoài ra nhóm không sử dụng dụng Linear Kernel vì có thể không đạt được kết quả tốt với bộ dữ liệu của nhóm khi trong một đặc tính (biến độc lập) có nhiều biến khác nhau, điều này có thể gây ra sự không tuyến tính trong dữ liệu. Trường hợp này làm cho việc phân loại bằng một siêu phẳng tuyến tính trở nên khó khăn hoặc không hiệu quả.

### **Sau khi tạo mô hình ta tìm được**

```
Best Parameters (SVM): {'kernel': 'rbf', 'gamma': 0.01, 'class_weight': {0: 0.3639090909090909, 1: 0.636090909090909}, 'C': 1}
Best Accuracy (SVM): 0.8852475916010892
```

*Hình 5.25 Kết quả mô hình*

### **Các thông số tốt nhất của mô hình**

- 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.

#### 5.4.4.2. Đánh giá mô hình

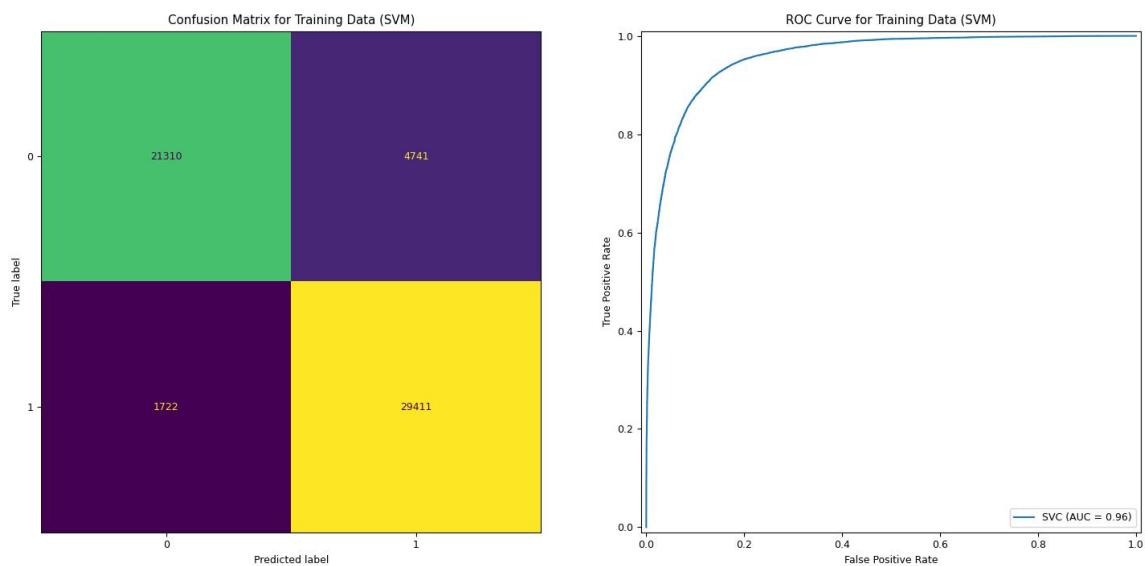
##### Dữ liệu huấn luyện

Classification report for training set (SVM)				
	precision	recall	f1-score	support
0	0.93	0.82	0.87	26051
1	0.86	0.94	0.90	31133
accuracy			0.89	57184
macro avg	0.89	0.88	0.88	57184
weighted avg	0.89	0.89	0.89	57184

Hình 5.26 Báo cáo số liệu của Support Vector Machine trên dữ liệu Train

##### Nhận xét

- *Precision*: Độ chính xác của mô hình trong việc dự đoán mỗi lớp. Precision cho lớp 0 (không đăng ký) là 93%, trong khi cho lớp 1 (đăng ký) là 86%. Điều này chỉ ra rằng mô hình đạt được tỷ lệ cao hơn trong việc dự đoán các trường hợp không đăng ký gửi tiền so với các trường hợp đăng ký gửi tiền.
- *Recall*: Tỷ lệ các trường hợp thực sự thuộc một lớp mà mô hình dự đoán đúng. Recall cho lớp 0 là 82%, chỉ ra rằng mô hình bỏ sót một vài trường hợp không đăng ký gửi tiền. Recall cho lớp 1 là 94%, điều này có nghĩa là mô hình gần như không bỏ sót các trường hợp đăng ký gửi tiền.
- *F1-score*: Trung bình điều hòa giữa precision và recall. F1-score cho lớp 0 là 87 %, trong khi cho lớp 1 chỉ là 90%. Điều này cho thấy sự cân bằng trong hiệu suất giữa hai lớp.



*Hình 5.27 Consusion Maxtrix và Roc Curve cho dữ liệu Train*

Confusion matrix cho thấy có ở dữ liệu huấn luyện có 21310 trường hợp được dự đoán là lớp 0 và thực sự là lớp 0 (True Negative), 29411 trường hợp được dự đoán là lớp 1 và thực sự là lớp 1 (True Positive), có 1722 trường hợp bị dự đoán sai (False Negative) và 4741 trường hợp bị dự đoán sai (False Positive).

### Dữ liệu kiểm thử

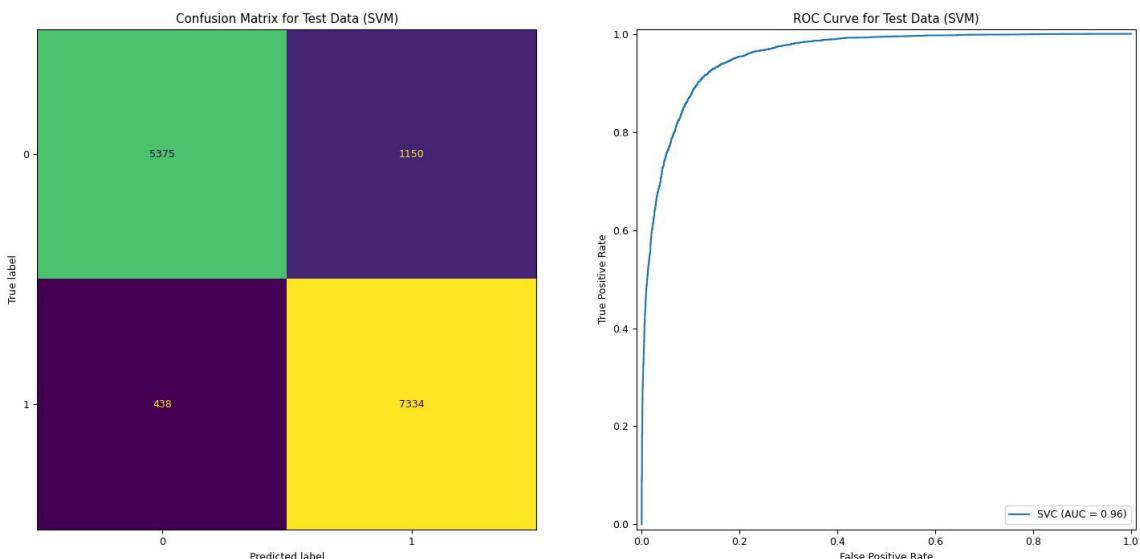
Classification report for Test set (SVM)				
	precision	recall	f1-score	support
0	0.92	0.82	0.87	6525
1	0.86	0.94	0.90	7772
accuracy			0.89	14297
macro avg	0.89	0.88	0.89	14297
weighted avg	0.89	0.89	0.89	14297

*Hình 5.28 Báo cáo số liệu của Support Vector Machine trên dữ liệu Test*

Nhận xét

- *Precision*: Độ chính xác của mô hình trong việc dự đoán mỗi lớp. Precision cho lớp 0 (không đăng ký) là 93%, trong khi cho lớp 1 (đăng ký) là 86%. Điều này chỉ ra rằng mô hình đạt được tỷ lệ cao hơn trong việc dự đoán các trường hợp không đăng ký gửi tiền so với các trường hợp đăng ký gửi tiền.
- *Recall*: Tỷ lệ các trường hợp thực sự thuộc một lớp mà mô hình dự đoán đúng. Recall cho lớp 0 là 82%, chỉ ra rằng mô hình bỏ sót một vài trường hợp không đăng ký gửi tiền. Recall cho lớp 1 là 94%, điều này có nghĩa là mô hình gần như không bỏ sót các trường hợp đăng ký gửi tiền.
- *F1-score*: Trung bình điều hòa giữa precision và recall. F1-score cho lớp 0 là 87 %, trong khi cho lớp 1 chỉ là 90%. Điều này cho thấy sự cân bằng trong hiệu suất giữa hai lớp.

Mô hình đã đạt được độ chính xác khoảng 88.70 % trên tập kiểm tra, chỉ hơi thấp hơn so với độ chính xác trên tập huấn luyện là 88.90 %. Điều này cho thấy mô hình không bị quá mức (**overfitting**) và **có khả năng tổng quát hóa tốt** trên dữ liệu mới.



Hình 5.29 Confusion Matrix và Roc Curve cho Test Data

Confusion matrix cho thấy có ở dữ liệu huấn luyện có 5375 trường hợp được dự đoán là lớp 0 và thực sự là lớp 0 (True Negative), 7334 trường hợp được dự đoán là lớp 1 và thực sự là lớp 1 (True Positive), có 438 trường hợp bị dự đoán sai (False Negative) và 1150 trường hợp bị dự đoán sai (False Positive).

## Tổng kết

- Accuracy on Training Data (SVM): 0.8869788752098489
- Accuracy on Test Data (SVM): 0.8889277470798069

### 5.4.5. Đánh giá và so sánh hiệu quả giữa các mô hình

	Accuracy	Precision	Recall	F1-Score
<b>Neural Network</b>	95.24%	94.98%	96.35%	95.66%
<b>Random Forest</b>	94.27%	93.53%	96.11%	94.8%
<b>Decision Tree</b>	93.36%	93.51%	94.34%	93.92%
<b>Support Vector Machine</b>	88.89%	86.45%	94.36%	90.23%

Bảng 5.7 Thống kê các chỉ số đánh giá của các mô hình

#### **Neural Network (Mạng nơ-ron)**

- Độ chính xác (accuracy) cao nhất trong số các mô hình, đạt 95.24%.
- Độ precision và recall gần như tương đương
- F1-score của mô hình này cũng cao nhất trong số các mô hình, đạt 95.66%.

#### **Random Forest (Rừng ngẫu nhiên)**

- Đạt được độ chính xác cao, gần bằng với Neural Network với 94.27%.
- Precision và recall cũng khá cao, tương đương với Neural Network là 93.53%
- F1-score của Random Forest là 94.8%, chỉ thấp hơn một chút so với Neural Network.

#### **Decision Tree (Cây quyết định)**

- Độ chính xác của Decision Tree là 93.36%, thấp hơn một chút so với Random Forest và Neural Network.
- Precision và recall gần như bằng nhau và tương đối cao khoảng 94 %
- F1-score của Decision Tree là 93.92%, cũng cao nhưng thấp hơn so với hai mô hình trên.

### ***Support Vector Machine (Máy vector hỗ trợ)***

- Đây là mô hình có độ chính xác thấp nhất trong số các mô hình với chỉ 88.89%.
- Precision và recall thấp hơn đáng kể so với các mô hình khác.
- F1-score cũng thấp nhất trong số các mô hình, chỉ đạt 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.

## TÀI LIỆU THAM KHẢO

1. Wikipedia. (n.d.). Truy cập ngày 18 tháng 4 năm 2024, từ <https://bookdown.org/rwnahhas/RMPH/blr-ordinal.html>
2. Jain, S. (2024, Tháng 3 11). ML | Underfitting và Overfitting. GeeksforGeeks. Truy cập ngày 18 tháng 4 năm 2024, từ <https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/>
3. Kanade, V. (2022, Tháng 4 18). Logistic Regression: Phương trình, Giả định, Loại, và Thực hành tốt nhất. Spiceworks. Truy cập ngày 18 tháng 4 năm 2024, từ <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>
4. Karabiber, F. (n.d.). Dummy Variable Trap – LearnDataSci. LearnDataSci. Truy cập ngày 18 tháng 4 năm 2024, từ <https://www.learndatasci.com/glossary/dummy-variable-trap/>
5. Khoa học dữ liệu. (2019, Tháng 12 12). Khoa học dữ liệu. Truy cập ngày 18 tháng 4 năm 2024, từ <https://phamdinhkhanh.github.io/2019/12/12/ARIMAmode.html>
6. Machine Learning cơ bản. (2017, Tháng 1 16). Machine Learning cơ bản. Truy cập ngày 18 tháng 4 năm 2024, từ <https://machinelearningcoban.com/2017/01/16/gradientdescent2/>
7. [NN] Mạng nơ-ron nhân tạo - Neural Networks. (2018, Tháng 4 23). Hai's Blog. Truy cập ngày 18 tháng 4 năm 2024, từ <https://dominhhai.github.io/vi/2018/04/nn-intro/>
8. Overfitting and underfitting in machine learning. (2022, Tháng 10 17). SuperAnnotate. Truy cập ngày 18 tháng 4 năm 2024, từ <https://www.superannotate.com/blog/overfitting-and-underfitting-in-machine-learning>
9. Rosidi, N. (2023, Tháng 9 4). Các Kỹ thuật Lựa chọn Đặc trưng trong Machine Learning. StrataScratch. Truy cập ngày 18 tháng 4 năm 2024, từ

<https://www.stratascratch.com/blog/feature-selection-techniques-in-machine-learning/>

10. Vấn đề Overfitting & Underfitting trong Machine Learning. (2019, Tháng 4 2).  
Trí tuệ nhân tạo. Truy cập ngày 18 tháng 4 năm 2024, từ  
<https://trituenhantao.io/kien-thuc/van-de-overfitting-underfitting-trong-machine-learning/>