Báo cáo đồ án cuối khóa

Bài toán phân tích dự đoán về tỉ lệ phân khúc khách hàng hủy sử dụng dịch vụ mạng

Contents

1. Giới thiệu về bộ dữ liệu và mục đích xây dựng mô hình để giải quyết vấn đề bài toán

- Bộ dữ liệu chứa các thông tin cơ bản cũng như các dịch vụ người dùng đăng ký (customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn) và ở đây ta sẽ sử dụng biến Churn là biến mục tiêu cho bài toán phân tích)
- Mục tiêu của project này là xây dựng một mô hình dự đoán tỉ lệ phân khúc khách hàng hủy sử dụng dịch vụ mạng và phân tích giải pháp cho việc giữ chân khách hàng. Chúng ta sẽ sử dụng các mô hình thuật toán để phân tích dữ liệu và dự đoán xác suất khả năng người dùng hủy dịch vụ. Từ đó có cái nhìn tổng quát về hiện trạng hủy dịch vụ mạng và thực hiện phương pháp để giảm việc khách hàng hủy dịch vụ cũng như giữ chân khách hàng.

2. Các phương pháp data processing và kết quả

- Do dữ liệu trong dataset đa số là kiểu object nên ta sẽ tiền xử lý nó bằng cách thay đổi sang dạng số để thuận tiện trong việc xử lý và phân tích

	customerI	gender	SeniorCitiz	Partner	Dependen	tenure	PhoneServ	MultipleLir	InternetSe	OnlineSecı	OnlineBac	DevicePro	TechSuppc	Streaming	StreamingI	Contract	PaperlessI	PaymentM	MonthlyCh	TotalCharg	Churn
2	7590-VHV	Female	0	Yes	No	1	No	No phone s	DSL	No	Yes	No	No	No	No	Month-to-r	Yes	Electronic	29.85	29.85	No
3	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed che	56.95	1889.5	No
4	3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-r	Yes	Mailed che	53.85	108.15	Yes
5	7795-CFO	Male	0	No	No	45	No	No phone s	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
6	9237-HQIT	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-r	Yes	Electronic	70.7	151.65	Yes
7	9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-r	Yes	Electronic	99.65	820.5	Yes
8	1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	No	Month-to-r	Yes	Credit carc	89.1	1949.4	No
9	6713-OKO	Female	0	No	No	10	No	No phone s	DSL	Yes	No	No	No	No	No	Month-to-r	No	Mailed che	29.75	301.9	No
10	7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-r	Yes	Electronic	104.8	3046.05	Yes
11	6388-TABC	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
12	9763-GRSI	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-r	Yes	Mailed che	49.95	587.45	No
13	7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Two year	No	Credit carc	18.95	326.8	No
14	8091-TTVA	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit carc	100.35	5681.1	No
15	0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-r	Yes	Bank trans	103.7	5036.3	Yes
16	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-r	Yes	Electronic	105.5	2686.05	No
17	3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit carc	113.25	7895.15	No
18	8191-XWS	Female	0	No	No	52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Mailed che	20.65	1022.95	No
19	9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
20	4190-MFLI	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-r	No	Credit carc	55.2	528.35	Yes
21	4183-MYFI	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-r	Yes	Electronic	90.05	1862.9	No
22	8779-QRD	Male	1	No	No	1	No	No phone s	DSL	No	No	Yes	No	No	Yes	Month-to-r	Yes	Electronic	39.65	39.65	Yes
23	1680-VDC	Male	0	Yes	No	12	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Bank trans	19.8	202.25	No
24	1066-JKSG	Male	0	No	No	1	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Month-to-r	No	Mailed che	20.15	20.15	Yes
25	3638-WEA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit carc	59.9	3505.1	No
26	6322-HRP	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to-r	No	Credit carc	59.6	2970.3	No
27	6865-JZNK	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-r	Yes	Bank trans	55.3	1530.6	No
28	6467-CHF	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes	Month-to-r	Yes	Electronic	99.35	4749.15	Yes
29	8665-UTDI	Male	0	Yes	Yes	1	No	No phone s	DSL	No	Yes	No	No	No	No	Month-to-r	No	Electronic	30.2	30.2	Yes
30	5248-YGIJI	Male	0	Yes	No	72	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	Yes	Credit carc	90.25	6369.45	No

Dataset trước khi preprocessing

	customerI	gender (Ma	SeniorCitiz	Partner (Ye	Dependen	tenure	PhoneServ	MultipleLir	InternetSe	OnlineSecı	OnlineBac	DevicePro	TechSuppc	Streaming	StreamingI	Contract (M	PaperlessI	PaymentM	MonthlyCh	TotalCharg	Churn (Yes = 1, No = 0)
2	7590-VHV	0	0	1	0	1	0	0	1	0	2	0	0	0	0	0	1	2	29.85	29.85	0
3	5575-GNV	1	0	0	0	34	1	1	1	2	0	2	0	0	0	1	0	3	56.95	1889.5	0
4	3668-QPYI	1	0	0	0	2	1	1	1	2	2	0	0	0	0	0	1	3	53.85	108.15	1
5	7795-CFO	1	0	0	0	45	0	0	1	2	0	2	2	0	0	1	0	0	42.3	1840.75	0
6	9237-HQIT	0	0	0	0	2	1	1	2	0	0	0	0	0	0	0	1	2	70.7	151.65	1
7	9305-CDSI	0	0	0	0	8	1	2	2	0	0	2	0	2	2	0	1	2	99.65	820.5	1
8	1452-KIOV	1	0	0	1	22	1	2	2	0	2	0	0	2	0	0	1	1	89.1	1949.4	0
9	6713-OKO	0	0	0	0	10	0	0	1	2	0	0	0	0	0	0	0	3	29.75	301.9	0
10	7892-POO	0	0	1	0	28	1	2	2	0	0	2	2	2	2	0	1	2	104.8	3046.05	1
11	6388-TABC	1	0	0	1	62	1	1	1	2	2	0	0	0	0	1	0	0	56.15	3487.95	0
12	9763-GRSI	1	0	1	1	13	1	1	1	2	0	0	0	0	0	0	1	3	49.95	587.45	0
13	7469-LKBC	1	0	0	0	16	1	1	0	1	1	1	1	1	1	2	0	1	18.95	326.8	0
14	8091-TTVA	1	0	1	0	58	1	2	2	0	0	2	0	2	2	1	0	1	100.35	5681.1	0
15	0280-XJGE	1	0	0	0	49	1	2	2	0	2	2	0	2	2	0	1	0	103.7	5036.3	1
16	5129-JLPIS	1	0	0	0	25	1	1	2	2	0	2	2	2	2	0	1	2	105.5	2686.05	0
17	3655-SNQ	0	0	1	1	69	1	2	2	2	2	2	2	2	2	2	0	1	113.25	7895.15	0
18	8191-XWS	0	0	0	0	52	1	1	0	1	1	1	1	1	1	1	0	3	20.65	1022.95	0
19	9959-WOF	1	0	0	1	71	1	2	2	2	0	2	0	2	2	2	0	0	106.7	7382.25	0
20	4190-MFLI	0	0	1	1	10	1	1	1	0	0	2	2	0	0	0	0	1	55.2	528.35	1
21	4183-MYFI	0	0	0	0	21	1	1	2	0	2	2	0	0	2	0	1	2	90.05	1862.9	0
22	8779-QRD	1	1	0	0	1	0	0	1	0	0	2	0	0	2	0	1	2	39.65	39.65	1
23	1680-VDC	1	0	1	0	12	1	1	0	1	1	1	1	1	1	1	0	0	19.8	202.25	0
24	1066-JKSG	1	0	0	0	1	1	1	0	1	1	1	1	1	1	0	0	3	20.15	20.15	1
25	3638-WEA	0	0	1	0	58	1	2	1	0	2	0	2	0	0	2	1	1	59.9	3505.1	0
26	6322-HRP	1	0	1	1	49	1	1	1	2	2	0	2	0	0	0	0	1	59.6	2970.3	0
27	6865-JZNK	0	0	0	0	30	1	1	1	2	2	0	0	0	0	0	1	0	55.3	1530.6	0
28	6467-CHF	1	0	1	1	47	1	2	2	0	2	0	0	2	2	0	1	2	99.35	4749.15	1
29	8665-UTDI	1	0	1	1	1	0	0	1	0	2	0	0	0	0	0	0	2	30.2	30.2	1
30	5248-YGIJI	1	0	1	0	72	1	2	1	2	2	2	2	2	2	2	1	1	90.25	6369.45	0

Dataset sau khi preprocessing

- Đọc file csv đã preprocessing

```
df = pd.read_csv("./Telco-Customer-Churn.csv")
```
✓ 0.0s Python

```
df
```
✓ 0.0s Python

	customerID	gender (Male = 1, Female = 0)	SeniorCitizen (Yes = 1, No = 0)	Partner (Yes = 1, No = 0)	Dependents (Yes = 1, No = 0)	tenure	PhoneService (Yes = 1, No = 0)	MultipleLines (0 = No phone service, 1 = No, 2 = Yes)	InternetService (0 = No, 1 = DSL, 2 = Fiber optic)	OnlineSecurity (No = 0, No internet service = 1, Yes = 2)	...	DeviceProtection (No = 0, No internet service = 1, Yes = 2)	TechSu (No = int servic Ye
0	7590-VHVEG	0	0	1	0	1	0	0	1	0	...	0	
1	5575-GNVDE	1	0	0	0	34	1	1	1	2	...	2	
2	3668-QPYBK	1	0	0	0	2	1	1	1	2	...	0	
3	7795-CFOCW	1	0	0	0	45	0	0	1	2	...	2	
4	9237-HQITU	0	0	0	0	2	1	1	2	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
7038	6840-RESVB	1	0	1	1	24	1	2	1	2	...	2	
7039	2234-XADUH	0	0	1	1	72	1	2	2	0	...	2	
7040	4801-JZAZL	0	0	1	1	11	0	0	1	2	...	0	

- Lấy số lượng cột và dòng, đồng thời lấy tên các cột

```
# Get the number of rows and columns
print(df.shape)
# get the column names
print(df.columns)
```
✓ 0.0s

```
(7043, 21)
Index(['customerID', 'gender (Male = 1, Female = 0)',
       'SeniorCitizen (Yes = 1, No = 0)', 'Partner (Yes = 1, No = 0)',
       'Dependents (Yes = 1, No = 0)', 'tenure',
       'PhoneService (Yes = 1, No = 0)',
       'MultipleLines (0 = No phone service, 1 = No, 2 = Yes) ',
       'InternetService (0 = No, 1 = DSL, 2 = Fiber optic)',
       'OnlineSecurity (No = 0, No internet service = 1, Yes = 2)',
       'OnlineBackup (No = 0, No internet service = 1, Yes = 2)',
       'DeviceProtection (No = 0, No internet service = 1, Yes = 2)',
```

- Tìm các kí tự đặc biệt trong giá trị từng cột và dòng

```
for i in df.columns:
    if is_numeric_dtype(df[i]) == False:
        list_char = []
        for j in range(len(df)):
            if type(df[i][j]) == str:
                list_char.extend(re.findall("[^A-Za-z0-9]", df[i][j]))
        print(i, list(dict.fromkeys(list_char)))
```

✓ 0.1s

```
customerID ['-']
```

- Tìm và xóa các giá trị trùng lặp

```
# show duplicate rows
df[df.duplicated()]
```

✓ 0.0s

customerID	gender (Male = 1, Female = 0)	SeniorCitizen (Yes = 1, No = 0)	Partner (Yes = 1, No = 0)	Dependents (Yes = 1, No = 0)	tenure	PhoneService (Yes = 1, No = 0)	MultipleLines (0 = No phone service, 1 = No, 2 = Yes)	InternetService (0 = No, 1 = DSL, 2 = Fiber optic)	OnlineSecurity (No = 0, No internet service = 1, Yes = 2)	...

0 rows × 21 columns

- Tìm và đếm số lượng giá trị riêng biệt trong tập dữ liệu

```
# Loop through each column and count the number of distinct values
for column in df.columns:
    num_distinct_values = len(df[column].unique())
    print(f"{column} -> {num_distinct_values} distinct values\n")
```

✓ 0.0s

```
customerID -> 7043 distinct values

gender (Male = 1, Female = 0) -> 2 distinct values

SeniorCitizen (Yes = 1, No = 0) -> 2 distinct values

Partner (Yes = 1, No = 0) -> 2 distinct values

Dependents (Yes = 1, No = 0) -> 2 distinct values
```

- Do tất cả dữ liệu trong tập dữ liệu đều đã thay đổi ở dạng number nên không cần thêm bước số hóa dữ liệu

3. Các thuật toán sử dụng

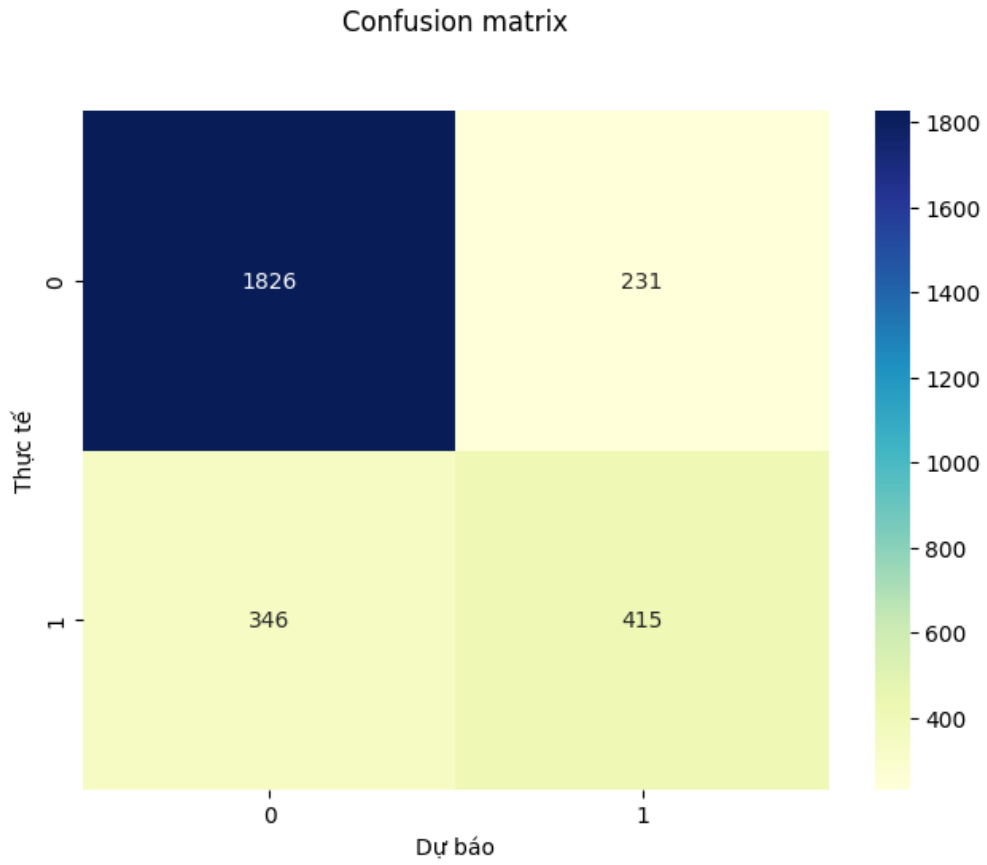
Chúng ta sẽ sử dụng biến Churn trong bộ dữ liệu để làm biến target và một loạt các thuật toán phân loại để dự đoán tỉ lệ tử vong, bao gồm Logistic Regression và Support Vector Machine.

- Logistic Regression
    - Chia các cột vào giá trị x và y (y là cột Churn) và chia tập dữ liệu thành train_size 70 và test_size 30
    - Khởi tạo mô hình Logistic Regression và huấn luyện dữ liệu để mô hình học từ dữ liệu và dự đoán, tính giá trị accuracy của mô hình thuật toán
    - Sử dụng thư viện Seaborn và scikit-learn để tính toán và hiển thị ma trận nhầm lẫn (confusion matrix) của mô hình Logistic Regression đã huấn luyện trước đó trên dữ liệu kiểm tra, sau đó tạo một heatmap (biểu đồ màu) cho ma trận nhầm lẫn (confusion matrix) đã tính toán trước đó và hiển thị các thông tin quan trọng liên quan đến hiệu suất của mô hình Logistic Regression trên dữ liệu kiểm tra.
    - Tính toán và in ra Classification Report gồm precision, recall, f1-score, support
- Support Vector Machine
    - Khởi tạo mô hình SVC cho SVM và thiết lập hàm kernel, tham số C và tham số Gamma. Hàm kernel là linear, tức là SVM sẽ thực hiện phân loại tuyến tính. Giá trị C lớn hơn (C=10) cũng có nghĩa là mô hình sẽ cố gắng phân loại đúng càng nhiều điểm dữ liệu trong tập huấn luyện mặc dù có thể phải chấp nhận sai số phân loại. Cuối cùng là huấn luyện dữ liệu
    - Tính toán và đưa ra accuracy của thuật toán
    - Đưa ra Classification Report và Confusion Matrix

4. Kết quả và so sánh

- Logistic Regression

- Confusion Matrix

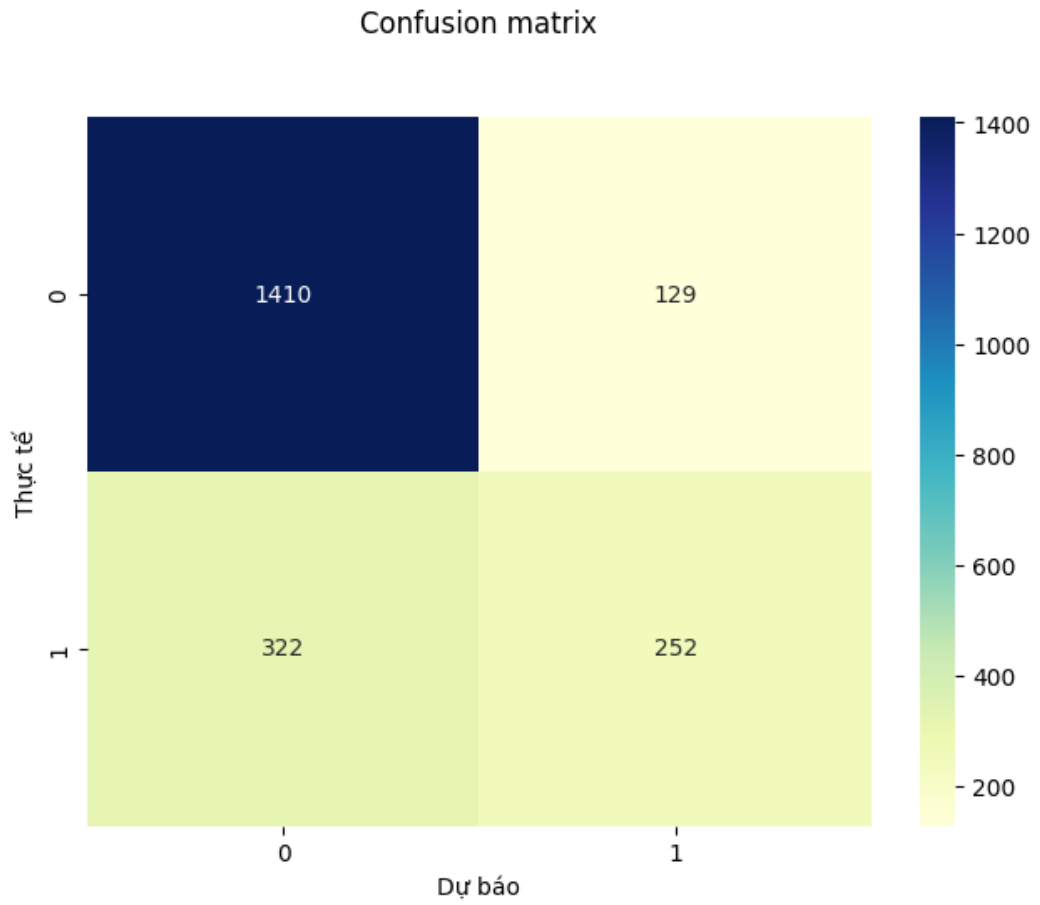


- Classification Report

	precision	recall	f1-score	support
0	0.84	0.89	0.86	2057
1	0.64	0.55	0.59	761
accuracy			0.80	2818
macro avg	0.74	0.72	0.73	2818
weighted avg	0.79	0.80	0.79	2818

- Support Vector Machine

- Confusion Matrix



- Classification Report

	precision	recall	f1-score	support
0	0.81	0.92	0.86	1539
1	0.66	0.44	0.53	574
accuracy			0.79	2113
macro avg	0.74	0.68	0.69	2113
weighted avg	0.77	0.79	0.77	2113

Dựa vào kết quả trên, ta đều có thể nhận định được rằng cả 2 mô hình đều cho ra độ chính xác khá cao (0.80 cho Logistic và 0.79 cho SVM).

Đối với chỉ số Precision của cả 2 mô hình cho trường hợp không hủy dịch vụ, cả 2 mô hình đều có khả năng dự đoán chính xác cao (theo lý thuyết, precision càng cao thì

khả năng dự đoán chính xác sẽ càng cao). Tuy nhiên precision cho trường hợp hủy dịch vụ lại tương đối thấp so với trường hợp không hủy, dẫn đến khả năng nhận diện những khách hàng có khả năng rời bỏ sẽ không chính xác.

Đối với chỉ số Recall của cả 2 mô hình, cả 2 đều cho chỉ số khá thấp ở lớp 1 (các trường hợp hủy dịch vụ), dẫn đến khả năng cả 2 mô hình sẽ bỏ qua 1 số trường hợp hủy.

Kết luận, cả 2 mô hình đều có khả năng dự đoán và phân loại khá tốt, cũng như độ chính xác cao cho các trường hợp không hủy dịch vụ (lớp 0), còn dự đoán về các trường hợp hủy dịch vụ cần phải được cải thiện hơn.

Giải pháp cho các trường hợp hủy dịch vụ thì ta có thể tăng chất lượng dịch vụ mạng như dịch vụ bảo mật hoặc streaming cũng như giảm thiểu chi phí dịch vụ hàng tháng.