TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

KHOA TOÁN KINH TẾ

BÀI TẬP NHÓM

Môn học: DATA DRIVEN MARKETING

Chủ đề: Phân tích chiến dịch quảng cáo mời khách hàng mở tài khoản tiết kiệm có kỳ hạn

Lớp: DSEB K61

Nhóm: 9

Giảng viên: Nguyễn Quỳnh Giang

Thành viên: Nguyễn Thùy Linh

Phạm Ngọc Anh

Lê Thu Hiền

Hồ Đức Duy

Hà Nội, 2022

Phần I: Giới thiệu

Đây là dữ liệu có liên quan đến các chiến dịch tiếp thị trực tiếp (gọi điện thoại) của một ngân hàng.

Mục tiêu:

+ Phân tích dữ liệu tìm ra insight và xác định tập khách hàng tiềm năng. Từ đó, hỗ trợ cho các chiến dịch tiếp theo đạt hiệu quả tốt hơn.

+ Xây dựng model để dự đoán một khách hàng có chấp nhận đăng ký mở tài khoản tiền gửi có kỳ hạn hay không (biến y).

1. Dữ liệu gốc gồm 41176 hàng và 21 cột. Chia làm ba bộ dữ liệu:

- **Thông tin khách hàng**
- Age: Tuổi của khách hàng
- Job: Công việc của khách hàng (phân loại: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- Marital: Tình trạng hôn nhân (phân loại: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
- Education: Trình độ học vấn (phân loại: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- Default: Tình trạng thiếu nợ (phân loại: 'no', 'yes', 'unknown')
- Housing: Tình trạng vay mua nhà (phân loại: 'no', 'yes', 'unknown')
- Loan: Tình trạng vay nợ (categorical: 'no', 'yes', 'unknown')

- **Thông tin về chiến dịch**
- Contact: Phương thức liên lạc khách hàng (phân loại: 'cellular','telephone')
- Month: Tháng gần nhất liên lạc với khách hàng (phân loại: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- Pdays: Số ngày đã trôi qua kể từ khi khách hàng đã được liên lạc từ chiến dịch trước (số; 999 nghĩa là khách hàng chưa được liên lạc)
- Day of Week: Ngày gần nhất liên lạc
- Previous: Số lần liên lạc trước chiến dịch
- Duration: Thời lượng liên lạc (theo giây)
- Poutcome: Kết quả của chiến dịch lần trước (phân loại: 'failure', 'nonexistent', 'success')
- Campaign: Số lần liên lạc trong chiến dịch này

- **Các yếu tố kinh tế - xã hội khác**

- Emp.var.rate: tỷ lệ thay đổi việc làm
- Cons.price.idx: Chỉ số giá tiêu dùng
- Cons.conf.idx: Chỉ số niềm tin người tiêu dùng
- Euribor3m: Tỉ giá gửi tín dụng đồng euro trong 3 tháng
- Nr.employe: Số nhân viên

- **Biến đầu ra**

- Y: Tình trạng đăng ký dịch vụ (nhị phân: 'yes', 'no')

	age	job	marital	education	default	housing	loan	contact	month	day_of_week
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon
1	57	services	married	high.school	unknown	no	no	telephone	may	mon
2	37	services	married	high.school	no	yes	no	telephone	may	mon
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon
4	56	services	married	high.school	no	no	yes	telephone	may	mon

campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no

2. Thống kê mô tả:

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41176.00000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	40.02380	258.315815	2.567879	962.464810	0.173013	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	10.42068	259.305321	2.770318	186.937102	0.494964	1.570883	0.578839	4.627860	1.734437	72.251364
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

3. Xử lý missing value, duplicate

- Dữ liệu không thiếu nhưng có 12 dòng bị lặp lại

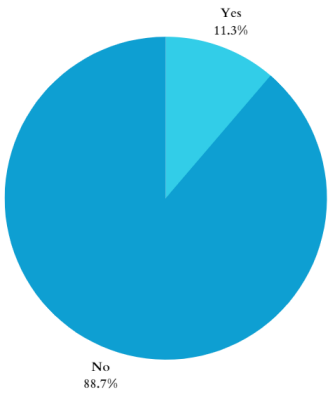
=> Tiến hành xóa các dòng bị lặp

Phần II: Phân tích dữ liệu

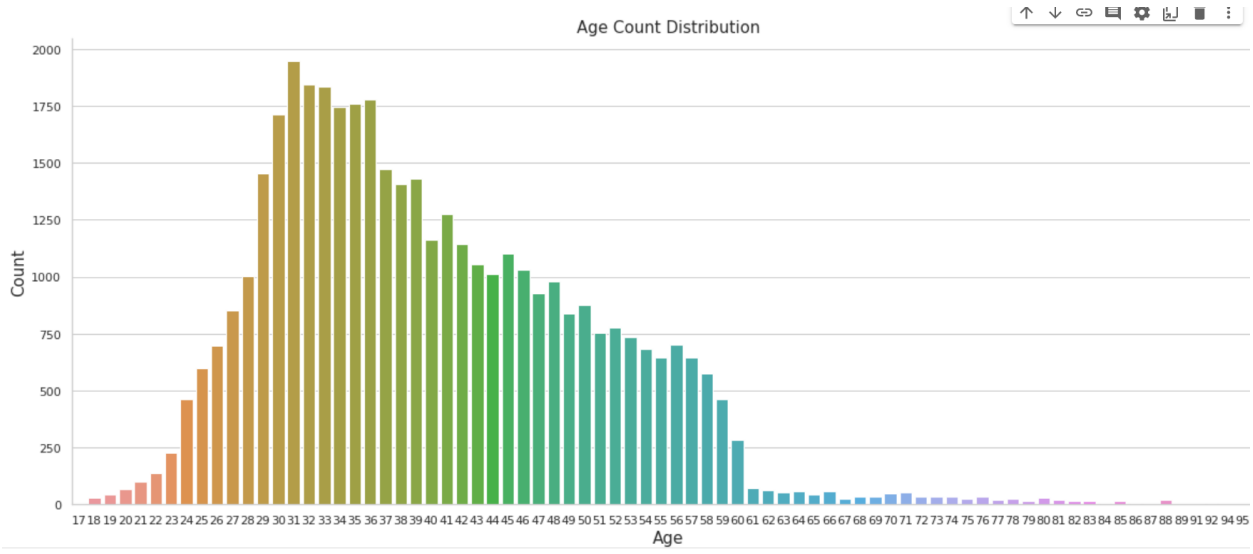
- ***Tỉ lệ chuyển đổi chung của chiến dịch***:

Tổng số khách hàng tham gia chiến dịch là 41176 người với 4639 khách hàng đồng ý mở tài khoản

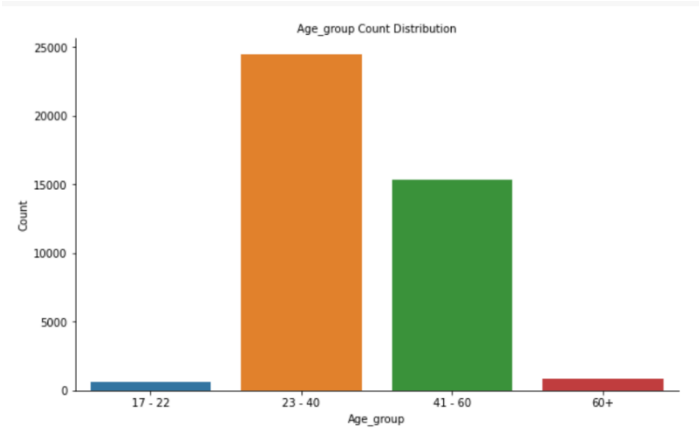
⇨ Tỉ lệ chuyển đổi chung là 11.27%



- ***Độ tuổi tham gia chiến dịch(age):***

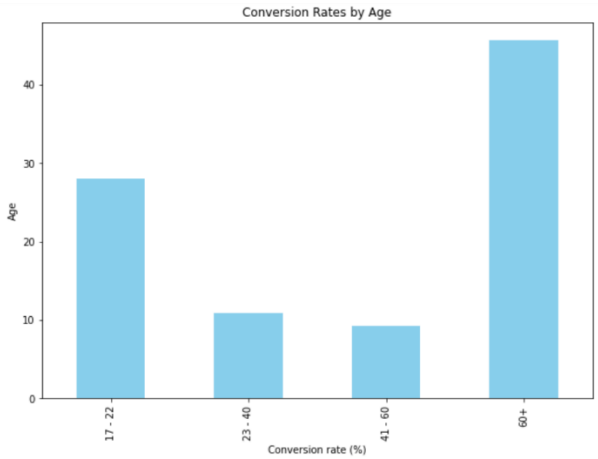


- Khách hàng tham gia chiến dịch có độ tuổi từ 17 – 98 tuổi. Biểu đồ này đang lệch trái cho ta thấy phần lớn khách hàng tham gia chiến dịch là người trẻ.



age_group	sum	count	conversion_rate
17 - 22	168	600	28.000000
23 - 40	2664	24433	10.903287
41 - 60	1423	15302	9.299438
60+	382	836	45.693780

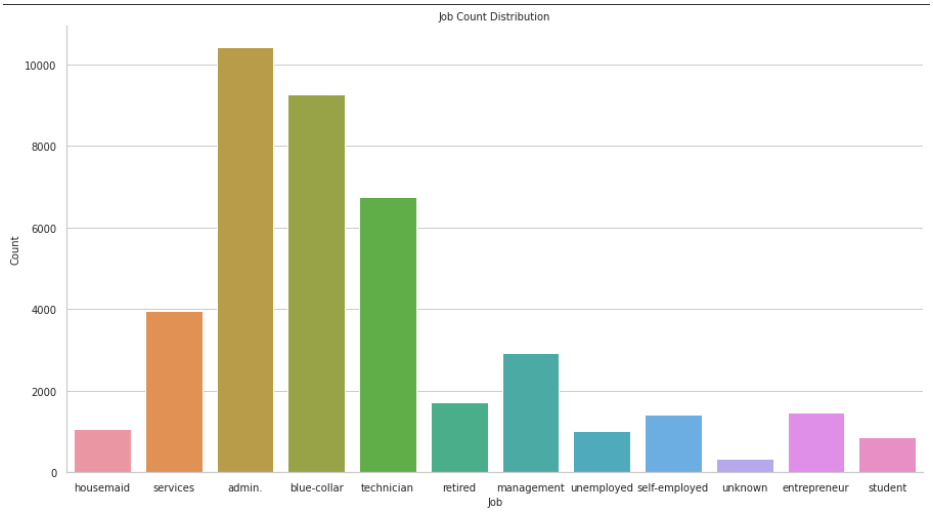
- Vì khách hàng tham gia chiến dịch trải dài từ 17 – 98 nên ta sẽ tiến hành nhóm các tuổi lại và chia thành 4 nhóm chính:
  + 17 – 22: nhóm khách hàng trong độ tuổi đi học
  + 23 – 40: nhóm khách hàng trẻ
  + 41 – 60: nhóm khách hàng trung niên
  + 60+: nhóm khách hàng đã nghỉ hưu
- Nhóm khách hàng trẻ 23 - 40 là nhóm khách hàng tham gia chiến dịch nhiều nhất, tiếp đến là nhóm khách hàng từ 41 - 60 và nhóm khách hàng cao tuổi 60+ và nhóm khách hàng 17 - 22 là nhóm có tỉ lệ tham gia chiến dịch thấp nhất.



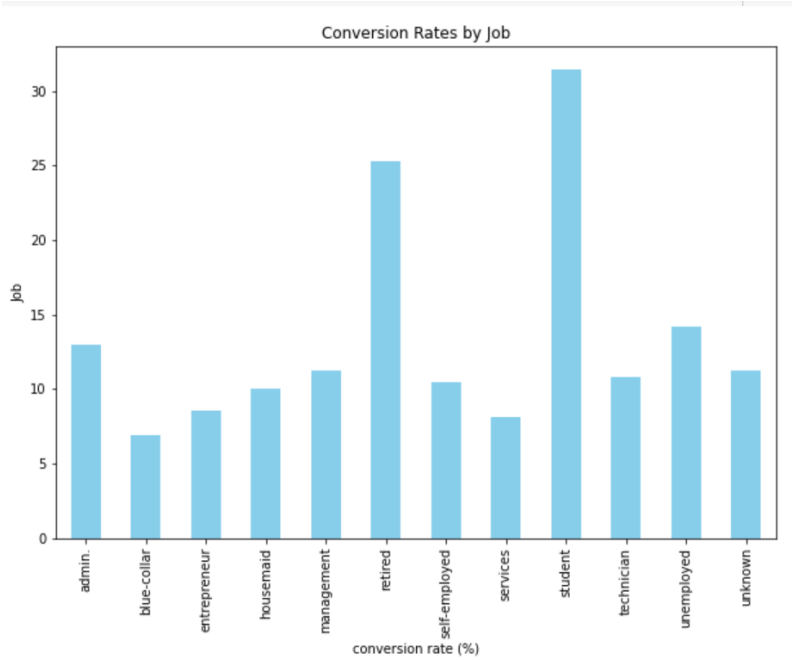
- Nhóm khách hàng trong độ tuổi 60+ có tỉ lệ chuyển đổi cao. Có thể khách hàng nhóm này chủ yếu trong độ tuổi nghỉ hưu, họ có tiền và không muốn đầu tư mạo hiểm (chứng khoán, …) nên có xu hướng gửi tiết kiệm nhiều hơn. Tiếp đến là nhóm khách hàng nhóm khách hàng từ 17 - 22. Đây là nhóm khách hàng trong độ tuổi đi học.

⇨ Action: Nên tập trung giới thiệu các dịch vụ đến 2 nhóm khách hàng này (nhóm người trẻ và nhóm người già)

- ***Nghề nghiệp (job):***



- Phần lớn khách hàng tham gia chiến dịch có nghề nghiệp là quản lý, công nhân và kỹ thuật viên

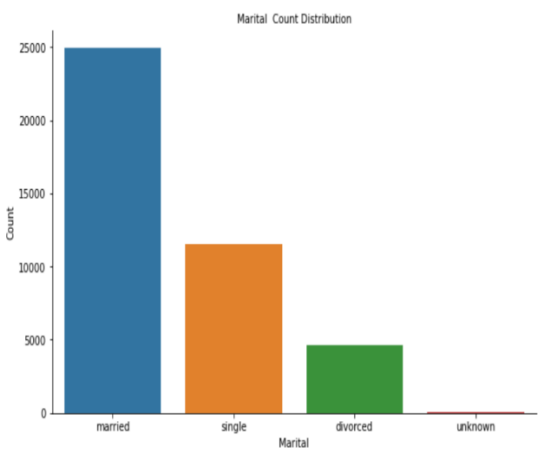


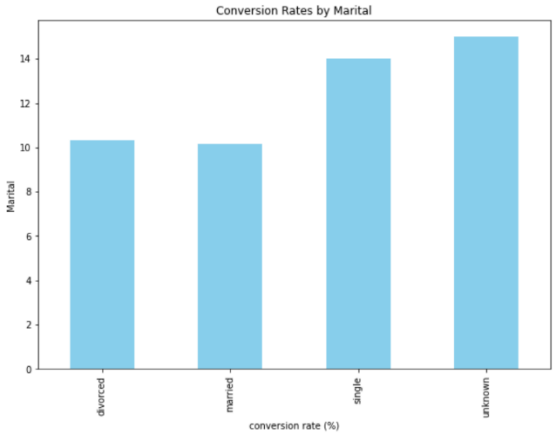
- Tỉ lệ chuyển đổi ở 3 nhóm khách hàng Học sinh, Người đã nghỉ hưu, Nhà quản lý là cao nhất trong khi đó họ không phải nhóm khách hàng tham gia chiến dịch lớn nhất. Với nhóm khách hàng công nhân và người làm dịch vụ có tỉ lệ chuyển đổi thấp nhất.

⇨ Nghề nghiệp không ảnh hưởng nhiều đến Tỉ lệ chuyển đổi

⇨ Action: Tập trung quảng cáo các dịch vụ của ngân hàng đối với các nhóm khách hàng là Học sinh, Người đã nghỉ hưu, Nhà quản lý

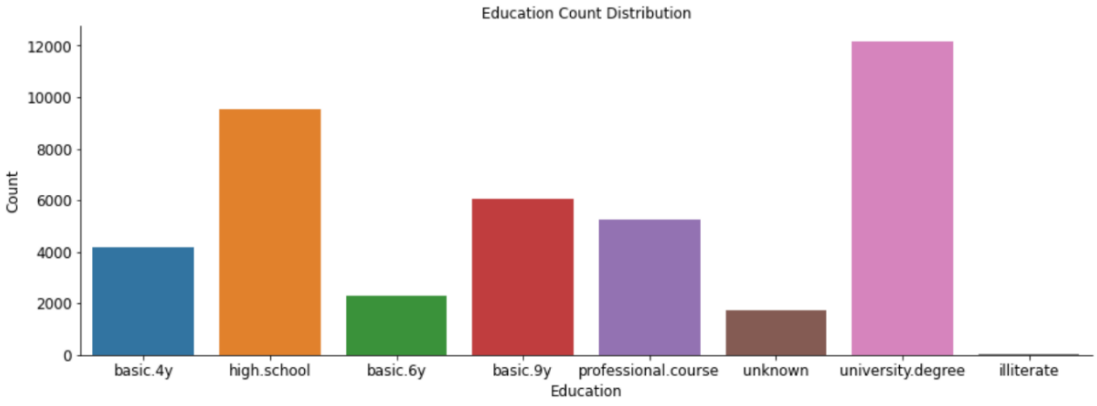
- ***Tình trạng hôn nhân (martial):***

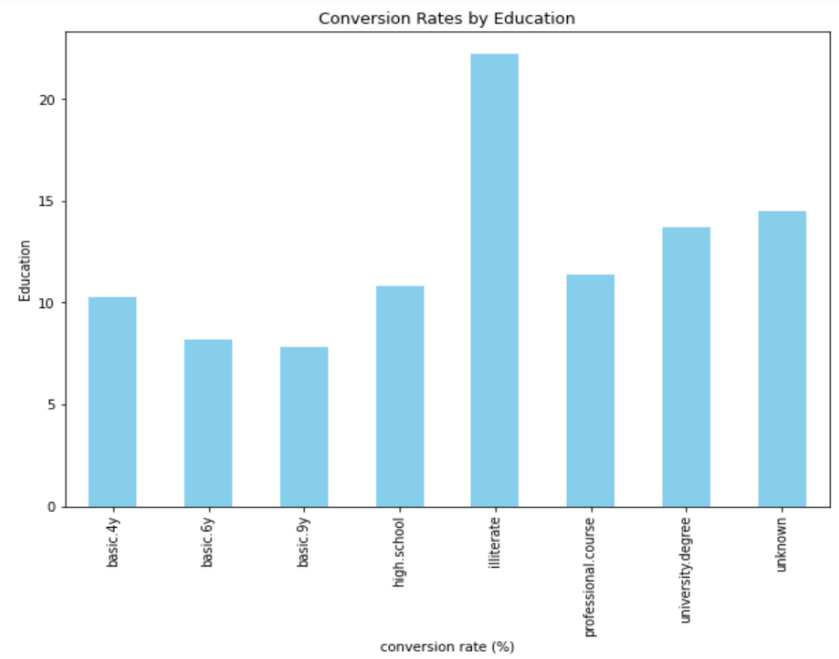




- Khách hàng tham gia chiến dịch với tình trạng hôn nhân "Đã kết hôn" chiếm số lượng lớn nhất nhưng tỉ lệ chuyển đổi thấp nhất. Điều này có thể do họ đã kết hôn và có con cái nên có nhiều thứ cần chi tiêu không có nhiều tiền nhàn rỗi để gửi tiết kiệm.
- Nhóm khách hàng trong tình trạng "Độc thân" cũng có tỉ lệ chuyển đổi cao.

- ***Trình độ học vấn (education):***



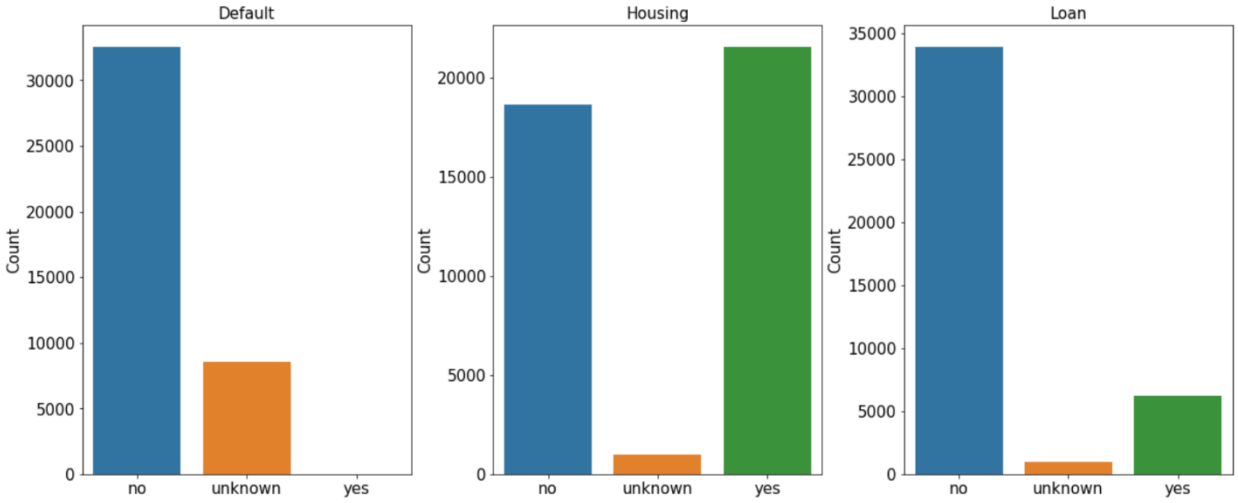


- Nhóm khách hàng có bằng đại học tham gia chiến dịch nhiều nhất và cũng là nhóm có tỉ lệ chuyển đổi cao nhất.
- Nhóm khách hàng Illiterate cũng có tỉ lệ chuyển đổi cao nhưng số lượng khách hàng tham gia quá ít. Tiếp đến là nhóm khách hàng có bằng đại học có tỉ lệ chuyển đổi cao.Với nhóm khách hàng có bằng đại học thì hầu hết họ sẽ làm các công việc có mức thu nhập cao nên có nhiều tiền để có thể mở tài khoản tiết kiệm.

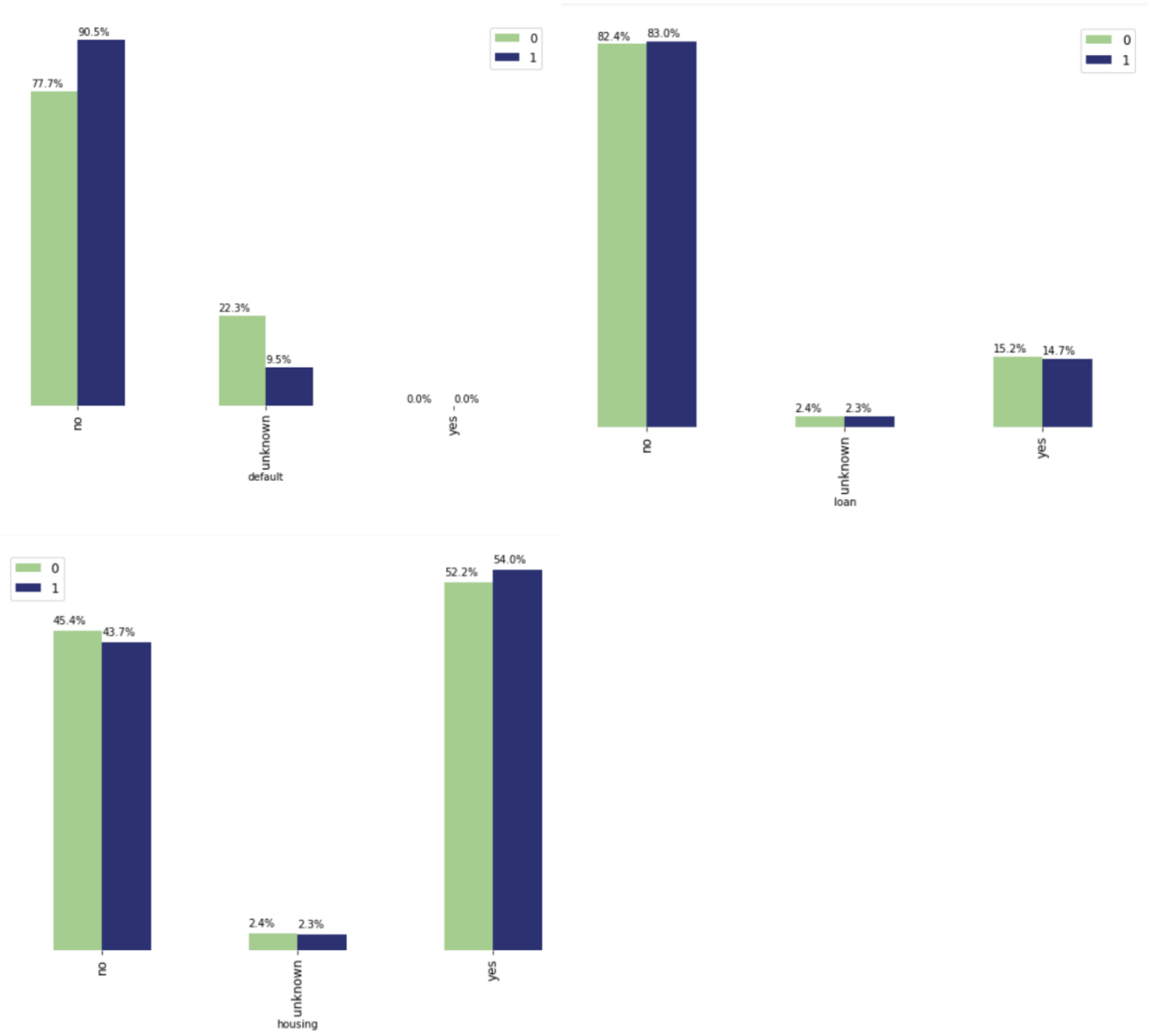
⇨ Trình độ học vấn có ảnh hưởng đến tỉ lệ chuyển đổi

⇨ Action: Nên tập trung giới thiệu sản phẩm gửi tiết kiệm đến nhóm khách hàng có bằng đại học

- ***Tình trạng vỡ nợ, vay mua nhà và vay cá nhân (Default, Housing, Loan):***

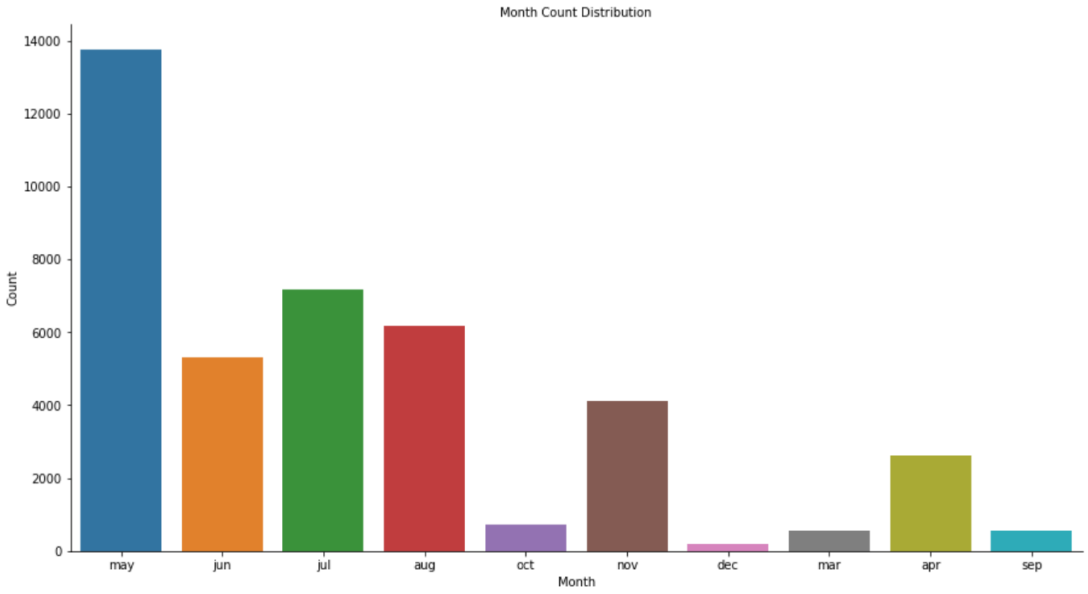


- Nhìn chung, khách hàng tham gia chiến dịch đều không trong tình trạng vỡ nợ hay có khoản vay cá nhân. Về khoản vay mua nhà có hơn một nửa khách hàng tham gia chiến dịch đang có khoản vay mua nhà.

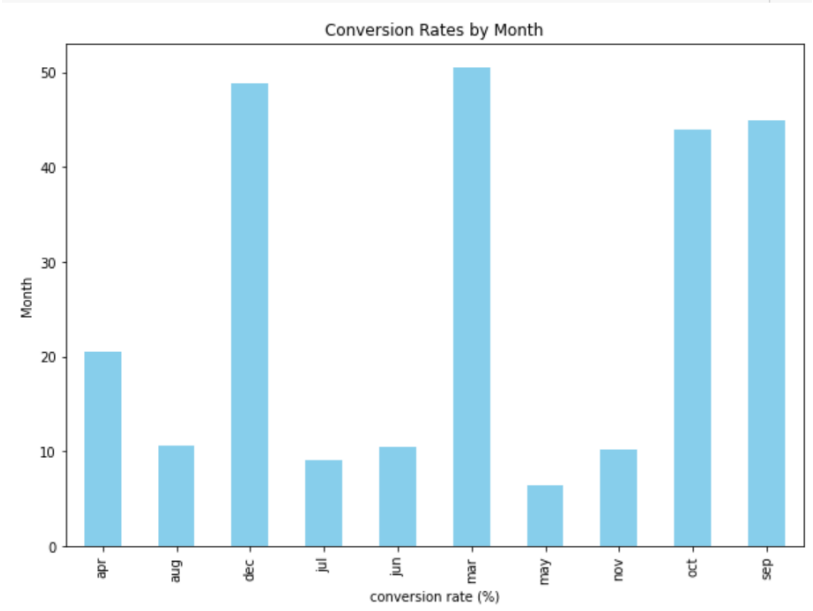


- Khách hàng mà không trong tình trạng vỡ nợ hay có khoản vay cá nhân sẽ có tỉ lệ chuyển đổi cao hơn.

- ***Tháng có cuộc gọi cuối cùng với khách hàng (month):***



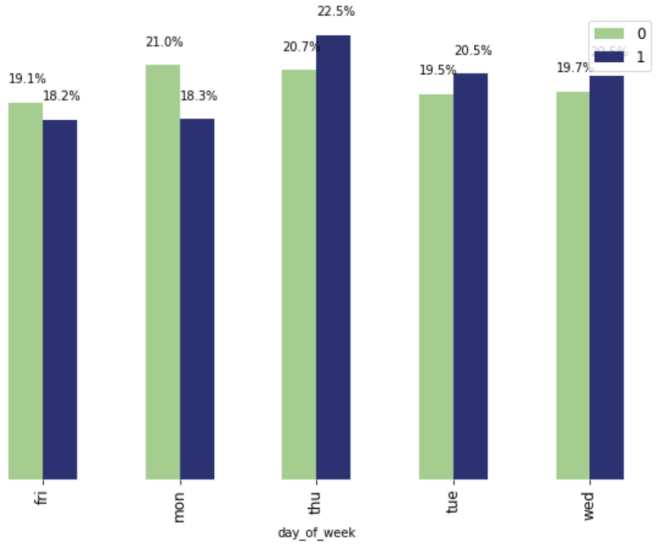
- Các tháng 5, 6, 7, 8 có số lượng cuộc gọi lần cuối nhiều nhất. Đặc biệt là tháng 5 với số lượng khách hàng liên lạc lần cuối cao gần gấp đôi tháng cao thứ 2 (tháng 7)



- Các tháng 3,9,12 có tỉ lệ chuyển đổi cao. Mặc dù số lượng liên lạc ở các tháng này là khá thấp.

⇨ Nên tập trung liên lạc vào các tháng cuối quý để thu hút được nhiều khách hàng hơn

- ***Ngày gần nhất liên lạc (day of week):***

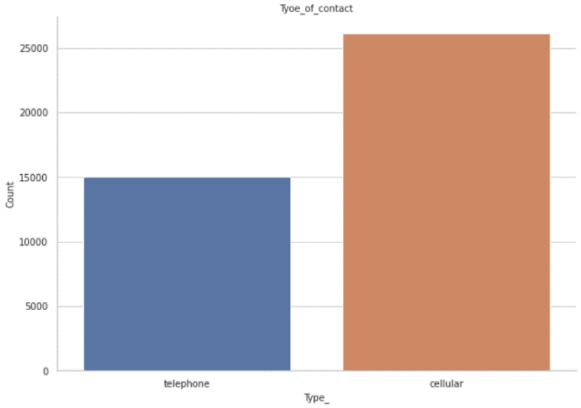


- Chiến dịch diễn ra vào các ngày trong tuần là các ngày làm việc của ngân hàng (từ thứ hai đến thứ sáu)
- Nhìn chung, các ngày trong tuần có số lượng liên lạc đều nhau
- Từ biểu đồ thể hiện tử lệ chuyển đổi, ta thấy các ngày giữa tuần (thứ 3 – thứ 5) có tỉ lệ chuyển đổi cao hơn 2 ngày đầu tuần và cuối tuần

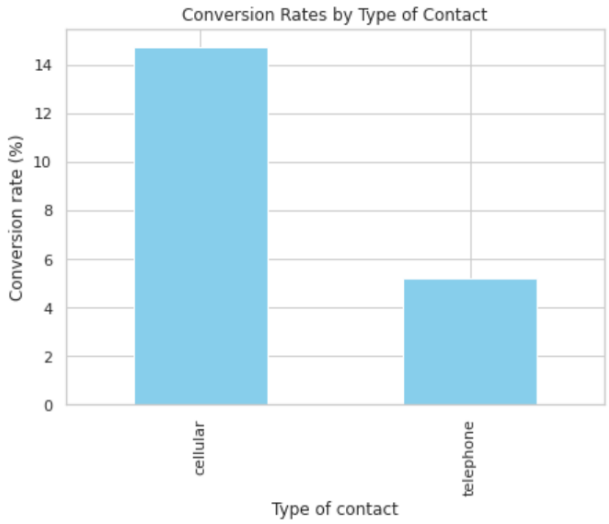
⇨ Action: Nên liên lạc với khách hàng vào các ngày giữa tuần

- ***Phương thức khách hàng liên lạc***

- Vì đây là chiến dịch gọi điện để tiếp thị trực tiếp nên khách hàng đều sử dụng phương thức liên lạc là điện thoại bàn hoặc điện thoại di động. Trong đó:



- Số lượng khách hàng sử dụng điện thoại di động cao gần gấp đôi so với khách hàng sử dụng điện thoại bàn

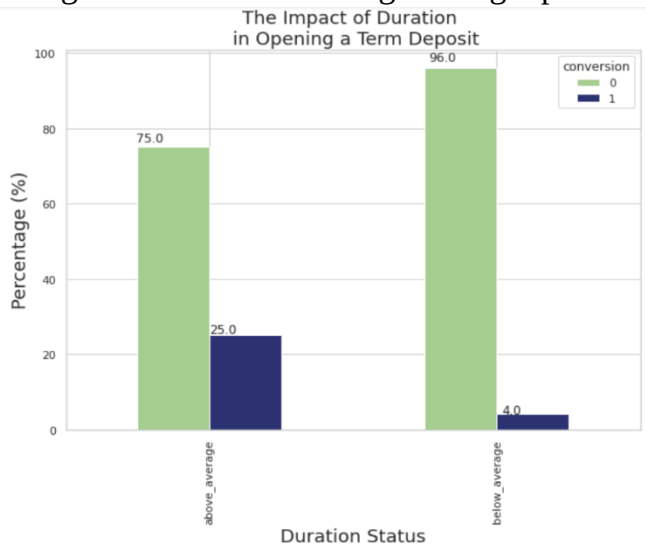


- Tỉ lệ chuyển đổi ở nhóm khách hàng có phương thức liên lạc bằng điện thoại di động là hơn 14% cao gần gấp 3 lần so với nhóm khách hàng sử dụng phương thức liên lạc là điện thoại bàn.

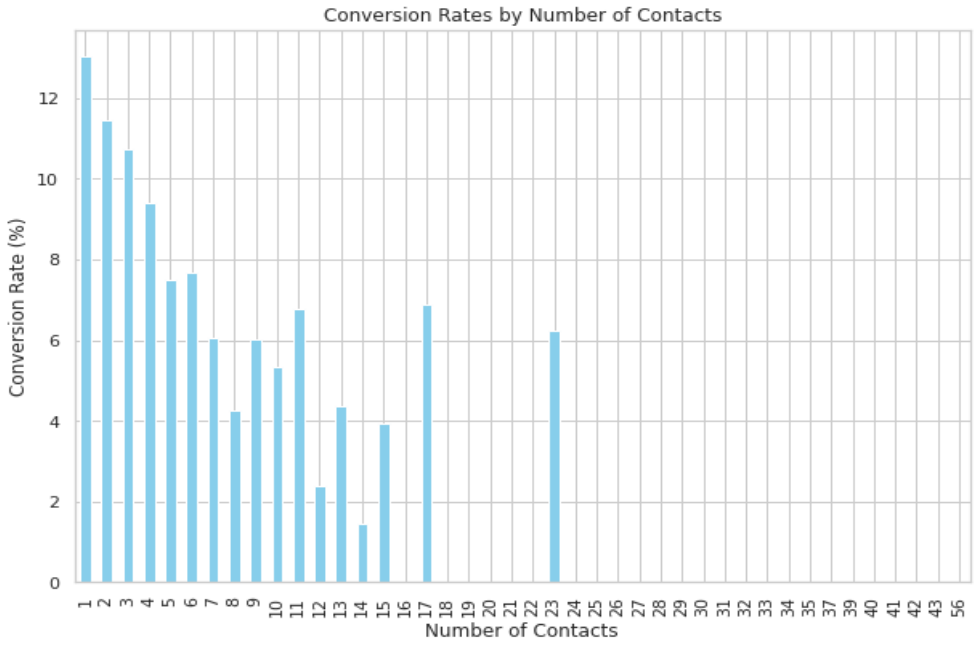
⇨ Action: Nên tăng cường liên lạc với nhóm khách hàng sử sụng phương thức liên lạc là điện thoại di động.

- ***Thời lượng liên lạc:***

- Thời gian liên lạc trung bình với 1 khách hàng khoảng 4 phút 18 giây.



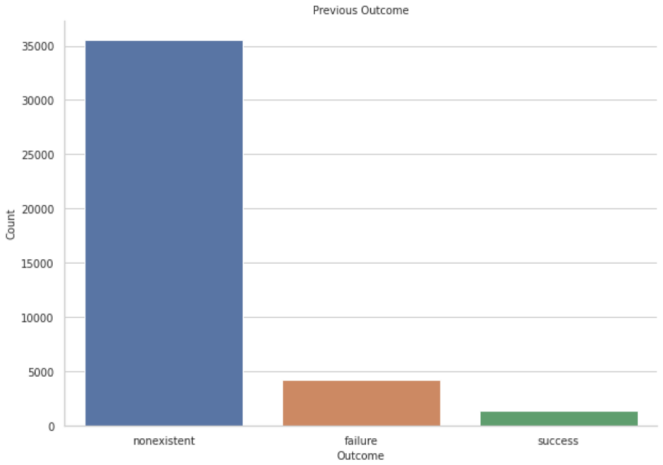
- Khách hàng có thời lượng liên lạc lâu hơn thời gian trung bình có tỉ lệ chuyển đổi cao hơn so với các khách hàng có thời gian liên lạc thấp hơn thời gian trung bình
  ⇨ Thời gian liên lạc với khách hàng càng lâu thì tỉ lệ chuyển đổi càng cao
  ⇨ Action: Ngân hàng nên kéo dài thời gian liên lạc với khách hàng bằng cách tăng tương tác với họ thông qua các câu hỏi hay. Có thể xây dựng các bảng hỏi cho các chiến dịch sau đó.



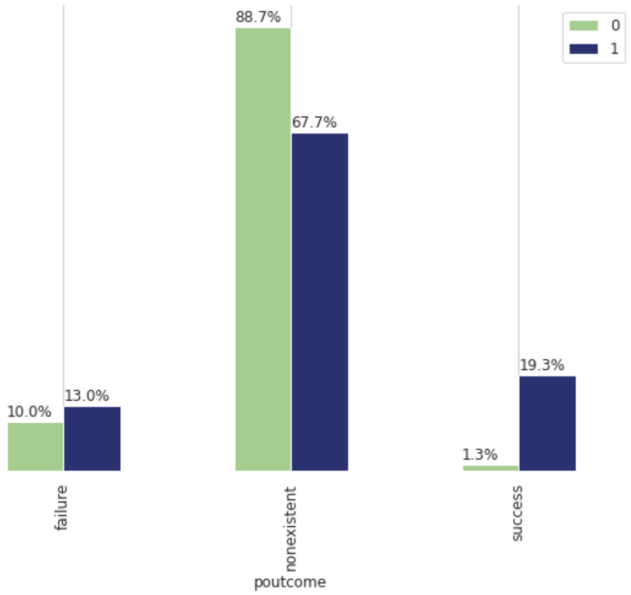
- Tỉ lệ chuyển đổi cao nhất ở nhóm khách hàng với 1 lần liên lạc và có xu hướng giảm dần khi khách hàng được liên lạc thưởng xuyên trong chiến dịch đó. Điều này có thể do việc ngân hàng gọi nhiều lần sẽ khiến họ cảm thấy bị làm phiền và không còn hứng thú với các dịch vụ mà ngân hàng giới thiệu trong cuộc gọi.

⇨ Hạn chế việc liên lạc thường xuyên với khách hàng. Có thể áp dụng chính sách không gọi quá 4 cuộc gọi với 1 khách hàng.

- **Previous outcome**



- Khách hàng tham gia chiến dịch lần này chủ yếu là khách hàng mới chỉ có một số ít là khách hàng của các chiến dịch trước.



- Từ biểu đồ ta thấy, với nhóm khách hàng tham gia từ chiến dịch trước có tỉ lệ đăng ký cao ở chiến dịch lần này, đặc biệt là nhóm khách hàng đồng ý trong chiến dịch trước.

=> Action: Tăng cường tiếp thị đến nhóm khách hàng đã tham gia ở các chiến dịch trước.

Phần III. Modeling

1. Kỹ thuật xử lý dữ liệu:

Ta chia một phần của bộ dữ liệu ban đầu ra nhiều bộ dữ liệu khác nhau để dễ dàng xử lý

```
bank_client = bank.iloc[: , 0:7]
bank_client.head()
```

	age	job	marital	education	default	housing	loan
0	56	housemaid	married	basic.4y	no	no	no
1	57	services	married	high.school	unknown	no	no
2	37	services	married	high.school	no	yes	no
3	40	admin.	married	basic.6y	no	no	no
4	56	services	married	high.school	no	no	yes

```
bank_related = bank.iloc[: , 7:11]
bank_related.head()
```

	contact	month	day_of_week	duration
0	telephone	may	mon	261
1	telephone	may	mon	149
2	telephone	may	mon	226
3	telephone	may	mon	151
4	telephone	may	mon	307

```
bank_se = bank.loc[: , ['emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed']]
bank_se.head()
```

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
0	1.1	93.994	-36.4	4.857	5191.0
1	1.1	93.994	-36.4	4.857	5191.0
2	1.1	93.994	-36.4	4.857	5191.0
3	1.1	93.994	-36.4	4.857	5191.0
4	1.1	93.994	-36.4	4.857	5191.0

```
bank_o = bank.loc[: , ['campaign', 'pdays','previous', 'poutcome']]
bank_o.head()
```

	campaign	pdays	previous	poutcome
0	1	999	0	nonexistent
1	1	999	0	nonexistent
2	1	999	0	nonexistent
3	1	999	0	nonexistent
4	1	999	0	nonexistent

- **Xóa các Unknown**

```
def dropna(self):
    self = self.replace('Unknown', np.nan)
    self = self.dropna()
```

- **Encoder** đối với:
  - Biến ‘Age’, ‘duration’: phân loại theo quartile và outlier sau đo dán nhãn bằng số từ 1 đến 5

```
def age(dataframe):
    dataframe.loc[dataframe['age'] <= self['age'].quantile(q = 0.25), 'age'] = 1
    dataframe.loc[(dataframe['age'] > self['age'].quantile(q = 0.25)) & (dataframe['age'] <= self['age'].quantile(q = 0.50)), 'age'] = 2
    dataframe.loc[(dataframe['age'] > self['age'].quantile(q = 0.50)) & (dataframe['age'] <= self['age'].quantile(q = 0.75)), 'age'] = 3
    dataframe.loc[(dataframe['age'] > self['age'].quantile(q = 0.75)) & (dataframe['age'] <= (self['age'].quantile(q = 0.75) + 1.5*(self['age'].quantile(q =
    dataframe.loc[(dataframe['age'] > (self['age'].quantile(q = 0.75)
                    + 1.5*(self['age'].quantile(q = 0.75)
                           - self['age'].quantile(q = 0.25)))) & (dataframe['age'] <= self['age'].quantile(q = 1.00)), 'age'] = 5
    return dataframe
age(self);
self['poutcome'].replace(['nonexistent', 'failure', 'success'], [1,2,3], inplace  = True)
def duration(data):

    #data.loc[(data['duration']) < self['duration'].quantile(q = 0.25), 'duration'] = 1

    data.loc[(data['duration']) <= self['duration'].quantile(q = 0.25), 'duration'] = 1
    data.loc[(data['duration'] > self['duration'].quantile(q = 0.25)) & (data['duration'] <= self['duration'].quantile(q = 0.50))  , 'duration']    = 2
    data.loc[(data['duration'] > self['duration'].quantile(q = 0.50)) & (data['duration'] <= self['duration'].quantile(q = 0.75))  , 'duration']   = 3
    data.loc[(data['duration'] > self['duration'].quantile(q = 0.75)) & (data['duration'] <= self['duration'].quantile(q = 0.75)
                                        + 1.5*(self['duration'].quantile(q = 0.75)
                                               - self['duration'].quantile(q = 0.25))), 'duration'] = 4
    data.loc[data['duration']  > self['duration'].quantile(q = 0.75) + 1.5*(self['duration'].quantile(q = 0.75)
                                                                            - self['duration'].quantile(q = 0.25)), 'duration'] = 5


    return data
duration(self);
```

- Biến ‘poutcome’: dán nhãn 'nonexistent', 'failure', 'success' lần lượt bằng (1,2,3)

- **Label encoder** đối với các biến phân loại còn lại:

```
def CatEncode(self):
#    bool_feat=['y']
    categorical_feats=['job','marital','education','default','housing','loan','contact','month','day_of_week','poutcome']
    #Converting dependent variable categorical to dummy
#    for x in bool_feat:
#        df[x] = pd.get_dummies(df[x], columns = [x], prefix = [x], drop_first = True)
    self.target = pd.get_dummies(self['y'], columns = ['y'], prefix = ['y'], drop_first = True)
```

2. Xây dựng model cho bài toán

- Bài toán này, ta sử dụng 2 thuật toán là Logistics regression và Decision Tree

❖ **Đánh giá độ chính xác của hai model logistic và decision tree**

Dựa trên accurancy score:

Đối với logistic model:

```
Accuracy on Testing set of logistic model:  0.9047098810390871
Accuracy on Traning set of logistic model:  0.9062215477996965
```

Ta có thể thấy rằng độ chính xác giữa hai tập dữ liệu khá tương đồng với nhau, điều này tốt bởi vì nó cho thấy rằng không có hoặc ít dấu hiệu underfitting và overfitting

Đối với decision tree model:

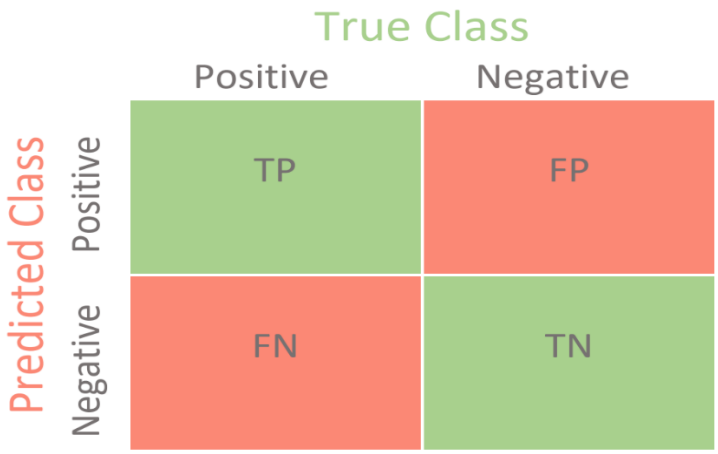
```
Accuracy on Traning set of decision tree model:  0.997298937784522
Accuracy on Testing set of decision tree model:  0.8695071619325079
```

Ta có thể thấy rằng sự khác nhau giữa độ chính xác của model với bộ dữ liệu train và bộ dữ liệu test quá chênh lệch nhau. Điều này cho thấy rằng model đang bị overfitting.

- **Dựa trên Confusion matrix và ROC Curve**

Confusion Matrix

Là một bố cục bảng cụ thể cho phép hình dung hiệu suất của một thuật toán



Ở trên ta có thể thấy được 4 giá trị trong bảng:

TP: Chỉ số những dự đoán đúng đúng với với thực tế

FN: Chỉ những dự đoán sai sai với thực tế

FP: Chỉ những dự đoán đúng sai với thực tế

TN: Chỉ những dự đoán sai đúng với thực tế

Và ta tìm độ chính xác của model bằng cách áp dụng công thức sau: (TP + TN)/(TP + TN + FP + FN)

Logistic Regression:

```
[[7136  167]
 [ 618  317]]
0.9060394537177542
```

Với những thông số trong bảng như trên, ta thấy được độ chính xác của model LogisticRegression là 90%

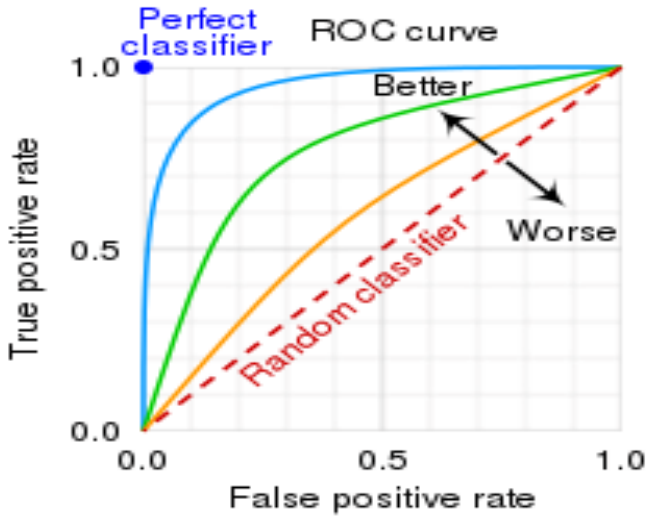
Decision Tree:

```
[[6751  552]
 [ 523  412]]
0.8710773899848254
```

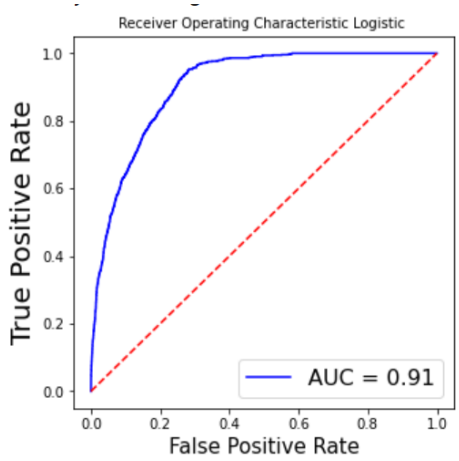
Với những thông số trong bảng như trên, ta thấy được độ chính xác của model LogisticRegression là 87%

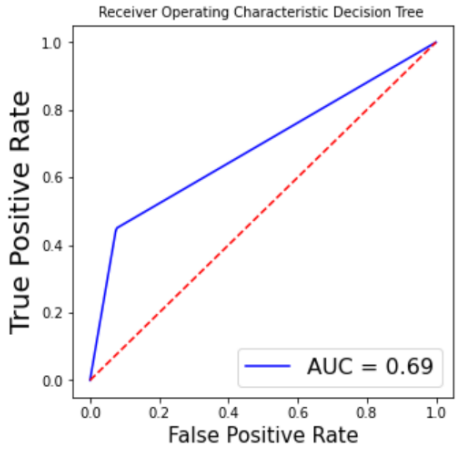
*** ROC Curve**

Là một đồ thị cho ta thấy khả năng dự đoán của một hệ thống phân cấp nhị phân,



Khoảng nằm dưới ROC curve được gọi là AUC (area under the ROC curve). Phần diện tích này càng lớn, mô hình càng hiệu quả trong việc phân loại dữ liệu. Và ROC curve càng nghiêng về phía bên trái càng tốt





KẾT LUẬN:

Qua những biểu đồ và phân tích ở trên, xét theo độ chính xác trong việc dự báo dữ liệu, model Logistic Regression vượt trội hơn hẳn so với Decision Tree trên bộ dữ liệu.

- ⇨ Lựa chọn thuật toán Logistics regression cho bài toán này.