TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN
KHOA TOÁN KINH TẾ

----------**

BÀI TẬP LỚN

MÔN DATA-DRIVEN MARKETING

***Đề tài*: Phân tích chiến dịch quảng cáo mời khách hàng mở tài khoản tiết kiệm không kỳ hạn.**

Sinh viên: Đào Thị Hồng Nhung
Nguyễn Thị Hoài Linh
Bùi Thị Mai Lương
Nguyễn Bá Đăng Khôi
Lớp : DSEB 61

Giới thiệu

Mở tài khoản tiết kiệm là một dịch vụ truyền thống của ngân hàng. Ngân hàng có những chiến dịch marketing thu hút khách hàng mở tài khoản tiết kiệm. Tuy nhiên trong chiến dịch marketing vừa rồi tỷ lệ khách hàng đồng ý mở tài khoản tiết kiệm rất thấp. Bài báo cáo này sẽ phân tích những yếu tố của khách hàng ảnh hưởng đến việc quyết định mở tài khoản tiết kiệm, đồng thời xây dựng model dự đoán khả năng khách hàng có đồng ý mở tài khoản.

Ở chiến dịch marketing trước, công ty đã đạt được một số kết quả nhất định với 41188 mẫu khách hàng:

- Tổng số khách hàng tham gia 41176.
- Số lượng khách hàng đồng ý mở tài khoản 4639.
- Tỷ lệ chuyển đổi của chiến dịch là 11.27%.

- Vấn đề (C): Tỷ lệ chuyển đổi từ khách hàng tiềm năng thành khách hàng thấp.
- Câu hỏi (Q): Làm sao để có thể tăng tỷ lệ chuyển đổi.
- Trả lời (A):
    + Tìm ra yếu tố tác động đến quyết định mở tài khoản tiết kiệm của khách hàng.
    + Tạo ra những chiến dịch marketing hiệu quả.

Outline

Nghiên cứu khách hàng: Phân tích khách hàng mở tài khoản tiết kiệm.

Phần 1: Thông tin nghiên cứu

Phần 2: Mục tiêu nghiên cứu

Phần 3: Phân tích dữ liệu

Phần 4: Xây dựng model

Phần 5: Kiến nghị

Phần 1: Thông tin nghiên cứu

1. Thông tin khảo sát.

- Tổng mẫu nghiên cứu: 41188.
- Độ tuổi: 17 – 98.
- Điều kiện nghiên cứu: Dữ liệu khách hàng có trong bank.
- Mục tiêu nghiên cứu: Phân tích dữ liệu khách hàng đưa ra các insight và dự đoán khách hàng có mở tài khoản tiết kiệm.

2. Bộ dữ liệu

Khách hàng

- Age: Số tuổi khách hàng.
- Job: Công việc hiện tại của khách hàng.
- Marital: Tình trạng hôn nhân của khách hàng.
- Education: Trình độ học vấn của khách hàng.
- Default: Khách hàng có tín dụng trong tình trạng vỡ nợ.
- Housing: Khách hàng có khoản vay mua nhà không.
- Loan: Khách hàng có khoản vay cá nhân không.

Liên quan tới khách hàng

- Contact: Phương thức liên lạc với khách hàng.
- Month: Tháng liên hệ với khách hàng.
- Day_of_week: Ngày trong tuần liên hệ với khách hàng.
- Duration: Thời lượng liên hệ với khách hàng.

Xã hội và kinh tế

- Emp.var.rate: Tỷ lệ thay đổi việc làm (báo cáo hàng quý).
- Cons.price.idx: Chỉ số giá tiêu dùng (báo cáo hàng tháng).
- Cons.conf.idx: Chỉ số niềm tin của người tiêu dung (báo cáo hàng tháng).
- Euribor3em: lãi suất 3 tháng euribor (báo cáo hàng ngày).
- Nr.employed: Số lượng nhân viên (báo cáo hàng quý).

Khác

- Campaign: số lượng liên hệ với khách hàng trong chiến dịch.
- Pdays: Số ngày kể từ khi lần cuối liên hệ với khách hàng từ một chiến dịch trước .
- Previous: Số lượng địa chỉ liên hệ được thực hiện trước chiến dịch này cho khách hàng.
- Poutcome: kết quả của chiến dịch marketing trước đó.

Y: Tình trạng khách hàng đã đăng ký mở tài khoản.

Phần 2: Mục tiêu nghiên cứu

- Phân tích dữ liệu và chỉ ra insight khách hàng.
- Xây dựng mô hình dự đoán quyết định khách hàng.
- Một số khuyến nghị cần chú ý.

Phần 3: Phân tích dữ liệu

1. Mô tả dataset

- Không có điểm dữ liệu nào bị trống.
- Có 12 dữ liệu bị lặp lại. Xóa dữ liệu bị lặp trong dataframe.
- Có 21 features:
    - Categorical features gồm 11 features: contact, day_of_week, default, education, housing, job, loan, marital, month, poutcome, y.
    - Numeric features gồm 10 features: age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, eurbor3m, nr.employed.

1.1. Đổi các giá trị 'unknown' trong mỗi feature thành np.nan. Số lượng và phần trăm null của mỗi feature.

	Features	Nbr_Null	Pct_Null
0	age	0	0.00
1	job	330	0.80
2	marital	80	0.19
3	education	1730	4.20
4	default	8596	20.88
5	housing	990	2.40
6	loan	990	2.40
7	contact	0	0.00
8	month	0	0.00
9	day_of_week	0	0.00
10	duration	0	0.00
11	campaign	0	0.00
12	pdays	0	0.00
13	previous	0	0.00
14	poutcome	0	0.00

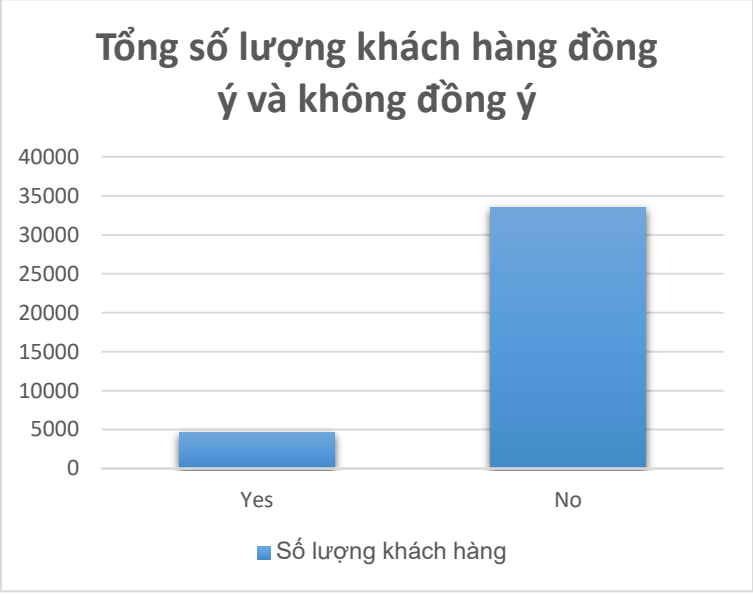
15	emp.var.rate	0	0.00
16	cons.price.idx	0	0.00
17	cons.conf.idx	0	0.00
18	euribor3m	0	0.00
19	nr.employed	0	0.00
20	y	0	0.00

- ⇨ Sau khi chuyển đổi giá trị 'unknown' thành 'np.nan' thì có 6 features xuất hiện giá trị null: job, marital, education, default, housing, loan.
- ⇨ Trước khi xử lý các biến bị null đó, chúng tôi sẽ nghiên cứu từng biến. Chúng tôi không thể xóa các giá trị null được vì những giá trị đó chiếm hơn 20% trong bộ dữ liệu.

2. Phân tích dữ liệu và tìm insight

2.1. Conversion rate

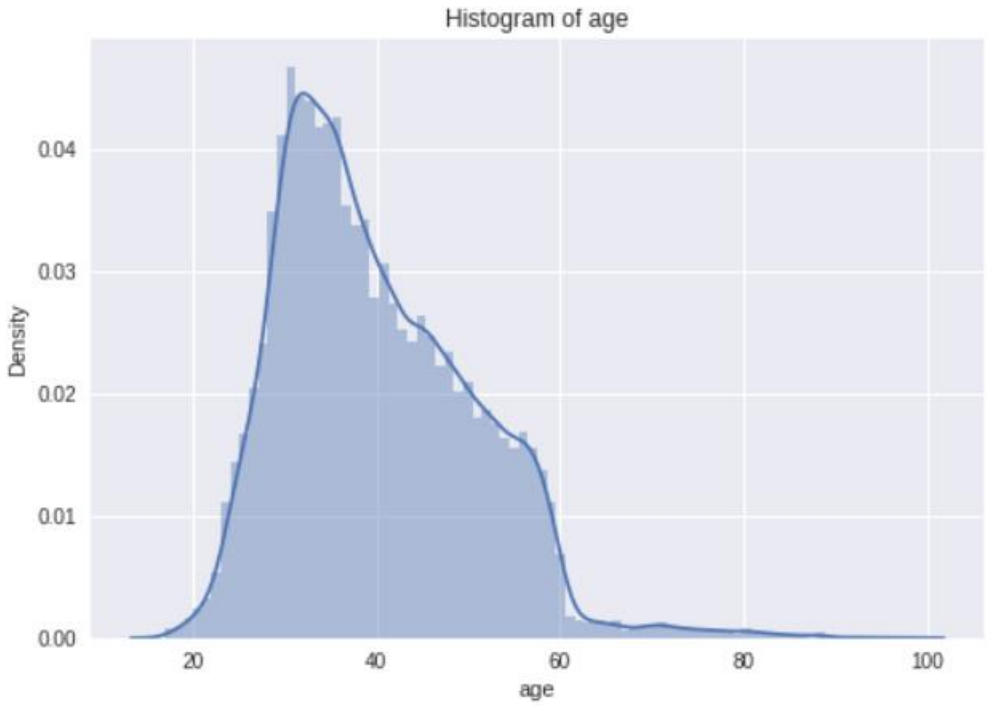
2.1.1. Tính tỷ lệ chuyển đổi



- ⇨ Tỷ lệ chuyển đổi (tỷ lệ phần trăm khách hàng đồng ý mở tài khoản tiết kiệm): 11.27%.
- ⇨ Số khách hàng đồng ý mở tài khoản rất ít => Tỷ lệ chuyển đổi thấp.

2.1.2. Tính tỷ lệ chuyển đổi theo tuổi.

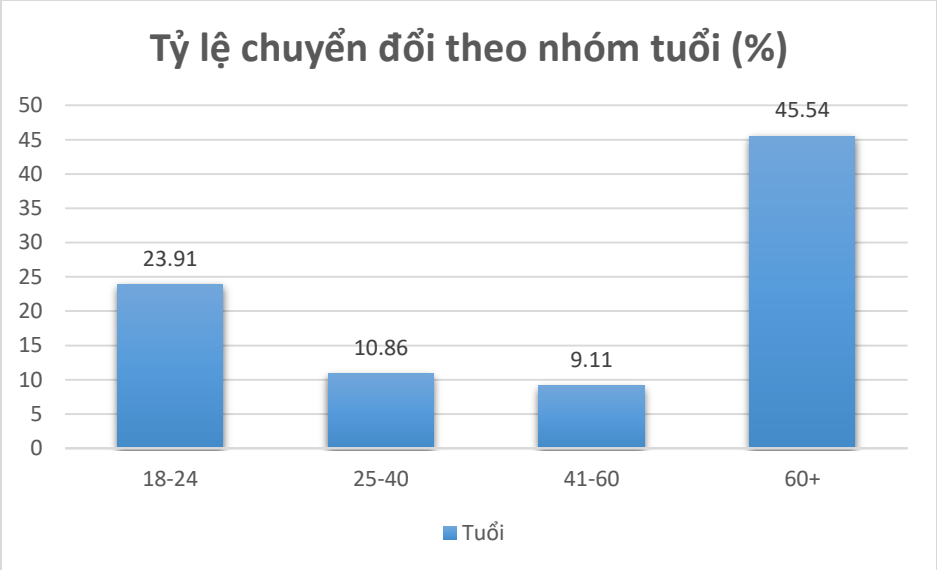
a) Phân phối của tuổi



➔ Khách hàng trong tập dữ liệu có độ tuổi phân phối lệch sang trái chứng tỏ khách hàng mà ngân hàng đang liên hệ chủ yếu thuộc nhóm người trẻ và trung niên, trong khoảng từ 25-60 tuổi. Cụ thể là:

- Nhóm 1 (từ 18 - 24 tuổi): 1062 khách hàng.
- Nhóm 2 (từ 25 – 40 tuổi): 22694 khách hàng.
- Nhóm 3 (từ 41 – 60 tuổi): 16506 khách hàng.
- Nhóm 4 (60 tuổi trở lên): 909 khách hàng.

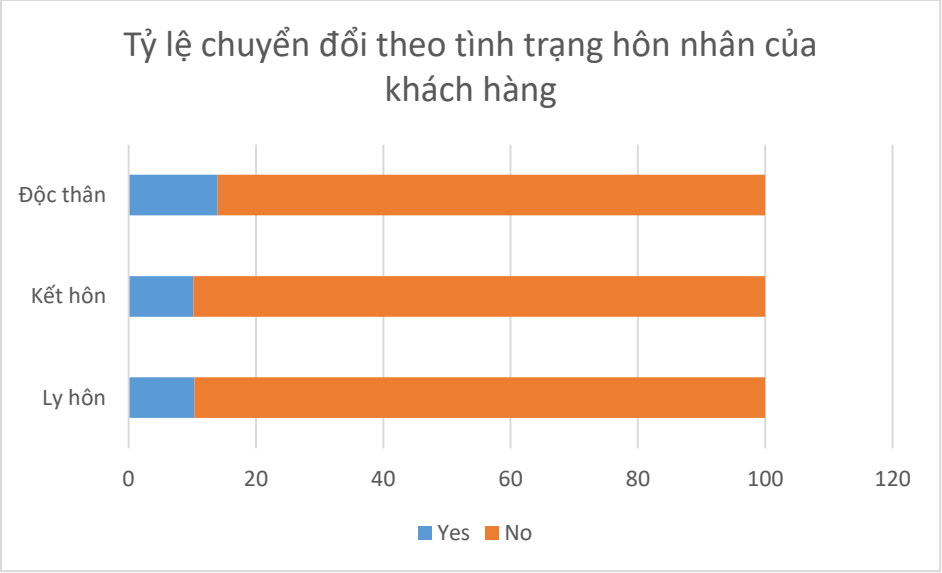
b) Tỷ lệ chuyển đổi



- ⇨ Tỷ lệ chuyển đổi của nhóm người từ 18-24 tuổi và trên 60 tuổi khá cao. Tuy nhiên số lượng khách hàng thuộc 2 nhóm này không nhiều so với số lượng khách hàng từ 25 – 60 tuổi.
- ⇨ Tỷ lệ chuyển đổi của nhóm người từ 25 – 60 tuổi khá thấp so với nhóm tuổi từ 18 – 24 và trên 60+ mặc dù số lượng khách hàng của nhóm tuổi này chiếm 95.2% tổng số khách hàng.
- ⇨ Cần xây dựng chiến dịch quảng cáo phù hợp để thu hút khách hàng trong nhóm tuổi từ 25 – 60.

2.1.3. Tỷ lệ chuyển đổi so với tỷ lệ không chuyển đổi trong tình trạng mối quan hệ.

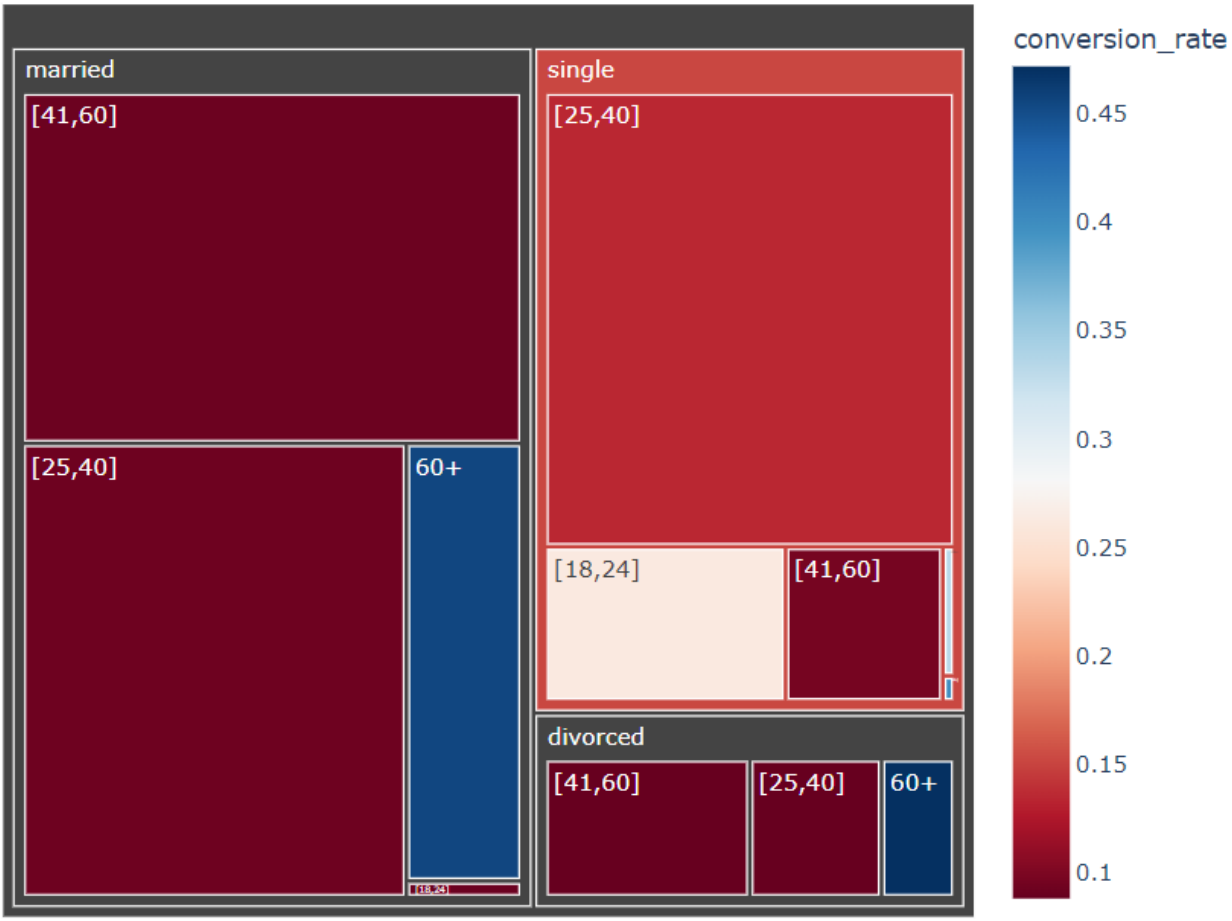
	No	Yes	Tổng	Tỷ lệ chuyển đổi
Ly hôn	4135	476	4611	10.32%
Kết hôn	22390	2531	24921	10.16%
Độc thân	9944	1620	11564	14%
Tổng	36469	4887	41096	11.89%



- Dù khách hàng đang ở tình trạng hôn nhân là độc thân, kết hôn hay li dị thì tỷ lệ khách hàng đăng kí mở tài khoản tiết kiệm đều tương đối thấp.
- Nhóm khách hàng đã kết hôn có số lượng khá lớn (chiếm 60.65% tổng số khách hàng) nhưng tỷ lệ chuyển đổi lại thấp nhất trong 3 nhóm khách hàng (10.16%).
- Tổng số khách hàng độc thân nhưng tỷ lệ chuyển đổi (14%) có phần cao hơn so với tỷ lệ chuyển đổi của nhóm khách hàng đã kết hôn.
    - ⇨ Cần tìm hiểu nguyên nhân và tập trung tạo chiến dịch marketing phù hợp với lượng lớn người đã kết hơn để tang số lượng khách mở tài khoản.

2.1.4. Tỷ lệ chuyển đổi theo nhóm tuổi và tình trạng hôn nhân.

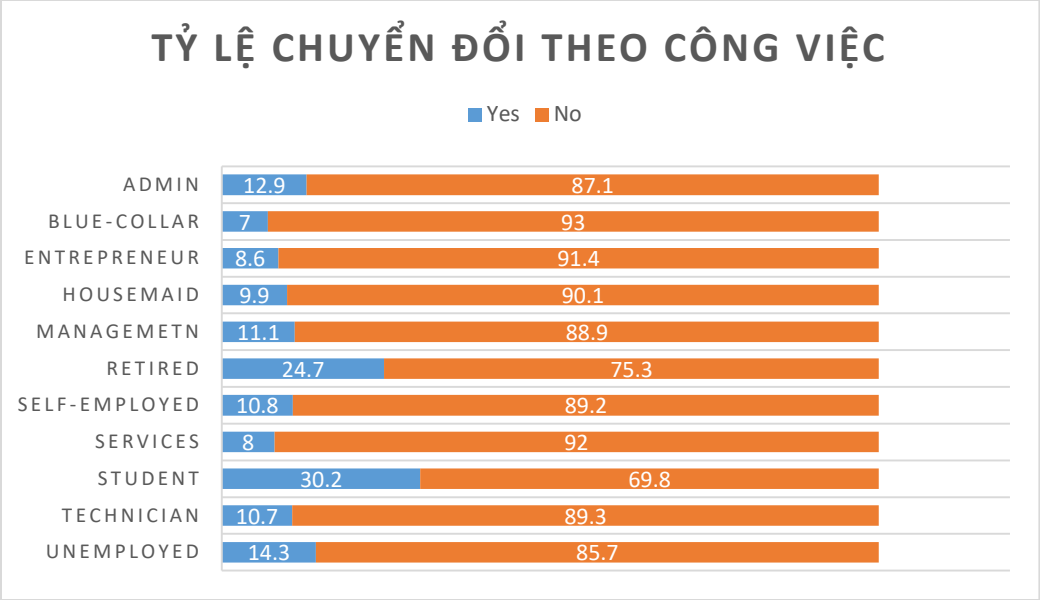
	Ly hôn	Kết hôn	Độc thân
18-	NaN	NaN	2.0
18-24	0.0	12.0	242.0
25-40	153.0	1097.0	1208.0
41-60	238.0	1104.0	158.0
60+	85.0	318.0	10.0



- Tỷ lệ chuyển đổi ở nhóm khách hàng phân chia theo tình trạng hôn nhân và độ tuổi ở độ tuổi từ 60 trở lên đều cao. Tuy nhiên nhóm khách hàng này có số lượng rất ít.
- Ngược lại ở các độ tuổi từ 18-60 chiếm đa số, đặc biệt là độ tuổi từ 25-40 trong nhóm độc thân hay kết hôn đều có số lượng khách hàng tiềm năng là rất lớn nhưng tỷ lệ chuyển đổi lại rất thấp.
    - ⇨ Cần chú trọng mở rộng khách hàng trên 60 tuổi, vì nhu cầu mở tài khoản tiết kiệm của nhóm khách hàng rất cao.
    - ⇨ Tìm hiểu nguyên nhân và xây dựng chiến dịch marketing để tăng tỷ lệ chuyển đổi nhóm khách hàng từ 25-60.

2.2. Các yếu tố ảnh hưởng tới tỷ lệ chuyển đổi.

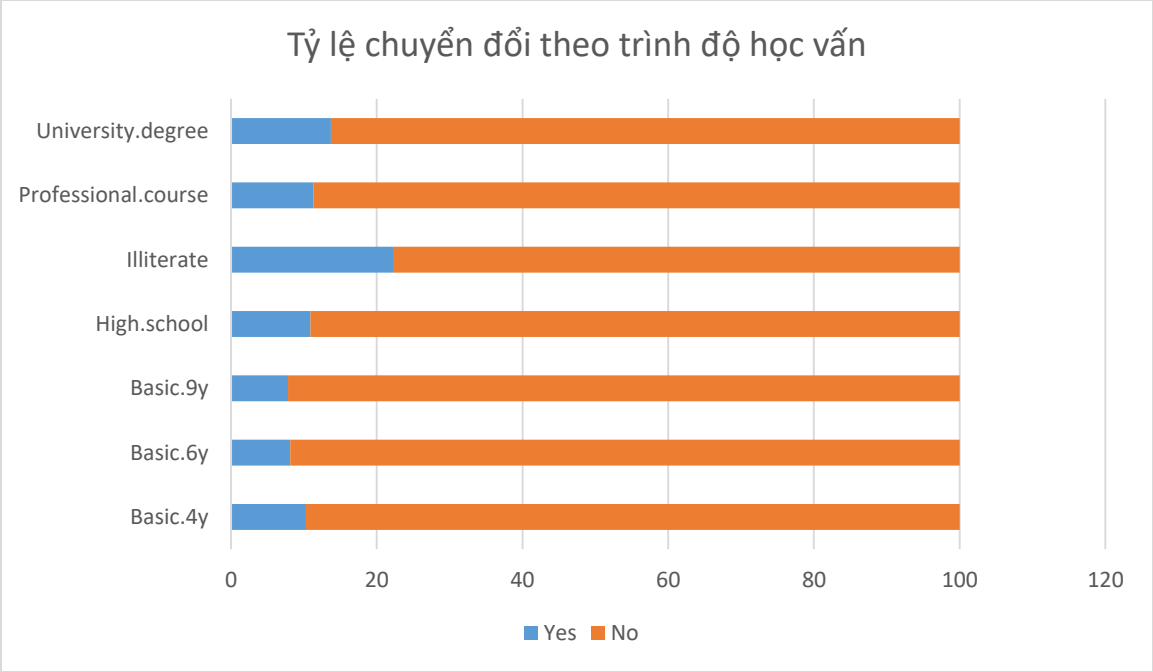
2.2.1.Công việc.



- Tỷ lệ đồng ý đăng ký gửi tiền tiết kiệm ở tất cả các nghề đều thấp hơn so với lượng người từ chối không đăng ký.
- Tuy nhiên đối với những khách hàng là học sinh- sinh viên và nghỉ hưu có khả năng đồng ý đăng ký mở sổ tiết kiệm khá cao (30.2% và 24.7%).
- Nguyên nhân:
  + Đối với người già thì mục tiêu chính của họ là gửi tiết kiệm để an hưởng tuổi già hoặc dành dụm cho con cháu; gửi tiền không kì hạn tuy lãi thấp nhưng lại là khoản đầu tư an toàn nhất.
  + Đối với học sinh - sinh viên thì thường không có đủ tiền hoặc kiến thức chuyên môn để đầu tư vào những khoản đầu tư phức tạp. Do vậy gửi tiết kiệm cũng là giải pháp đầu tư an toàn nhất.

2.2.2. Trình độ học vấn.

	Yes	No	Tổng	Tỷ lệ chuyển đổi (%)
Basic.4y	413	3594	4007	10.3
Basic.6y	180	2029	2209	8.15
Basic.9y	459	5405	5864	7.83
High.school	1008	8247	9255	10.89
Illiterate	4	14	18	22.22
Professional.course	578	4525	5103	11.32
University.degree	1624	10220	11844	13.71



- Qua biểu đồ ta thấy, tỷ lệ chuyển đổi của các nhóm khách hàng đều xấp xỉ nhau và có tỷ lệ chuyển đổi không cao.
- Tỷ lệ đồng ý đăng ký của nhóm khách hàng illiterate cao hơn so với những nhóm khác, tuy nhiên số lượng khách hàng ở nhóm này quá ít (18 khách hàng).
    - ⇨ Trình độ học vấn của khách hàng không có ảnh hưởng đến khách hàng quyết định mở tài khoản tiết kiệm.

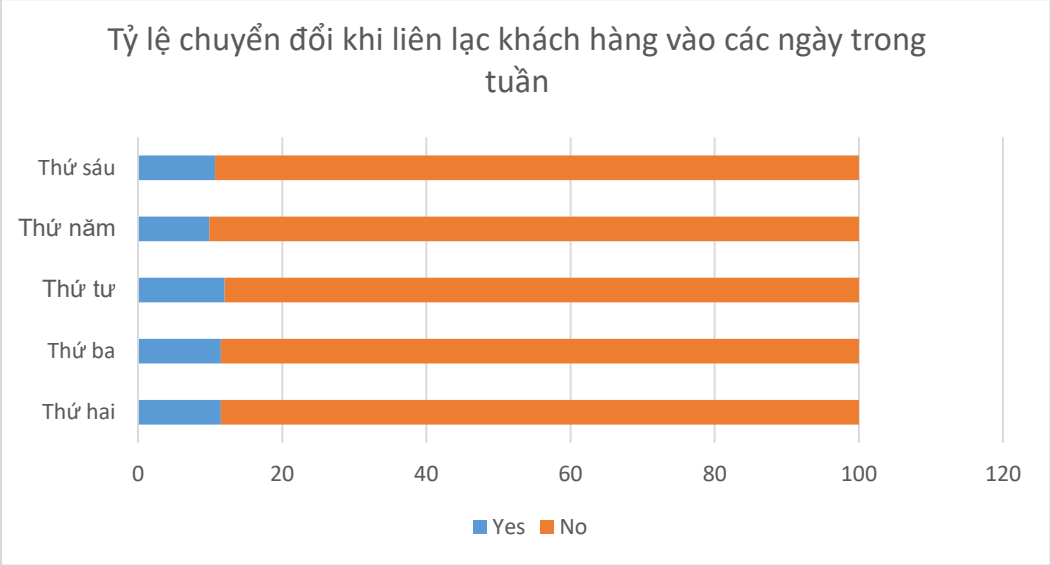
2.2.3. Nợ

	No	Yes	Tổng
Nợ	5328	639	5967
Không nợ	28706	3627	32333

- Tỷ lệ đăng ký mở tài khoản của khách hàng có khoản nợ và không có khoản nợ xấp xỉ nhau và đều rất thấp.
    - ⇨ Tình trạng khoản nợ của khách hàng không ảnh hưởng đến quyết định đăng ký mở tài khoản tiết kiệm của khách hàng.

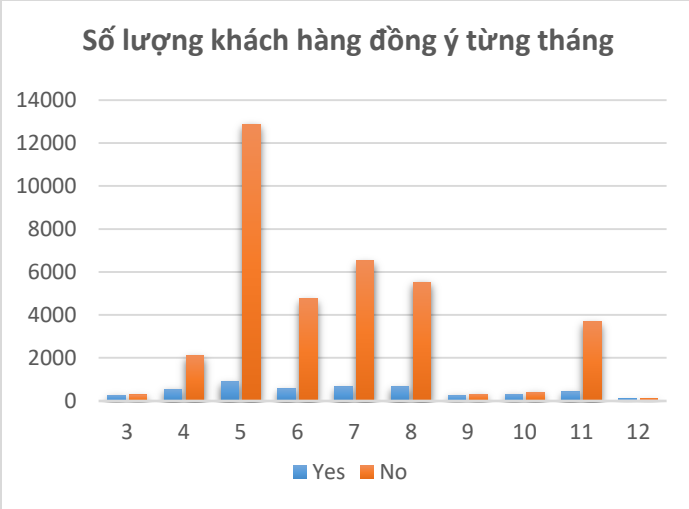
2.2.4. Thời gian liên hệ với khách hàng.

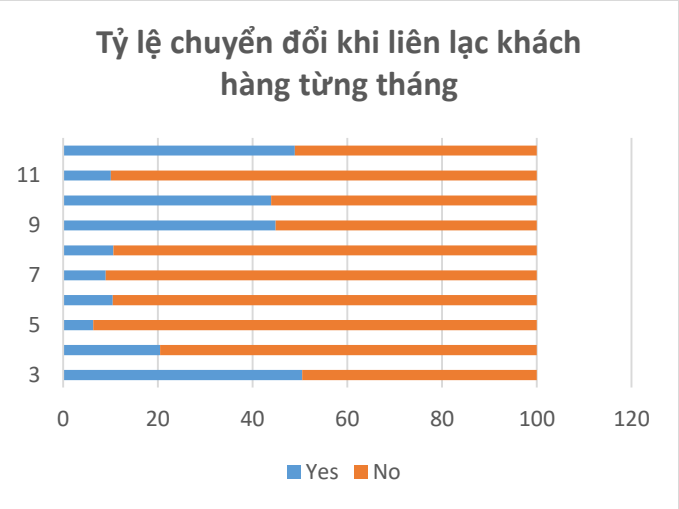
2.2.4.1. Ngày.



- Tỷ lệ chuyển đổi khi liên lạc vào các ngày trong tuần không chênh lệch nhau nhiều và các tỷ lệ đó đều khá thấp.
    - ⇨ Có thể liên lạc với khách hàng bất kỳ ngày nào trong tuần vì điều đó không ảnh hưởng lớn đến quyết định mở tài khoản tiết kiệm của họ.

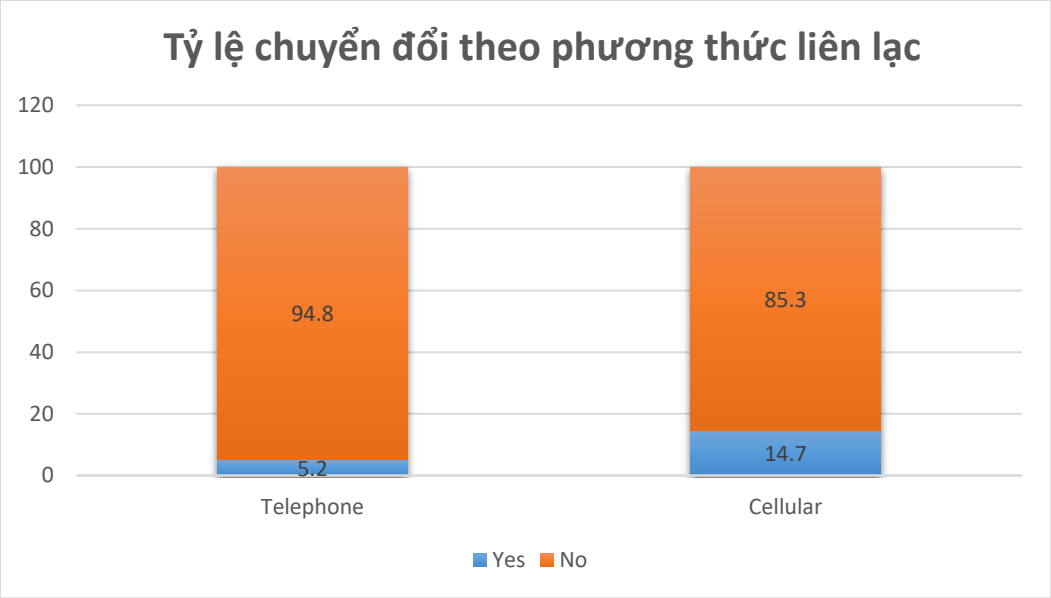
2.2.4.2. Tháng





- Qua biểu đồ số lượng khách hàng đồng ý từng tháng, ngân hàng chạy chiến dịch vào tháng 5, 6, 7, 8 và 11 với số lượng khá lớn. Tuy nhiên số lượng khách hàng từ chối mở tài khoản nhiều, dẫn đến tỷ lệ chuyển đổi rất thấp, đặc biệt vào tháng 5 với số lượng khách hàng rất nhiều (13767 khách hàng) nhưng chỉ có 6% số lượng khách hàng đồng ý.
    - ⇨ Cần tìm hiểu nguyên nhân tại sao xảy ra tình trạng như vậy vào các tháng đẩy mạnh chiến dịch đến khách hàng nhưng khách hàng lại không tiếp nhận.
- Tháng 3, 9, 10 và 12 tuy chiến dịch lan tỏa không đến nhiều khách hàng nhưng hiệu quả mang lại vượt qua mong đợi. Đặc biệt vào tháng 3 tuy chỉ có 546 khách hàng nhưng có đến 276 (50.5% số lượng khách hàng) đồng ý.
    - ⇨ Khảo sát thị trường, tìm hiểu nguyên nhân dẫn đến chiến dịch hiệu quả vào những tháng 3, 9, 10 và 12.

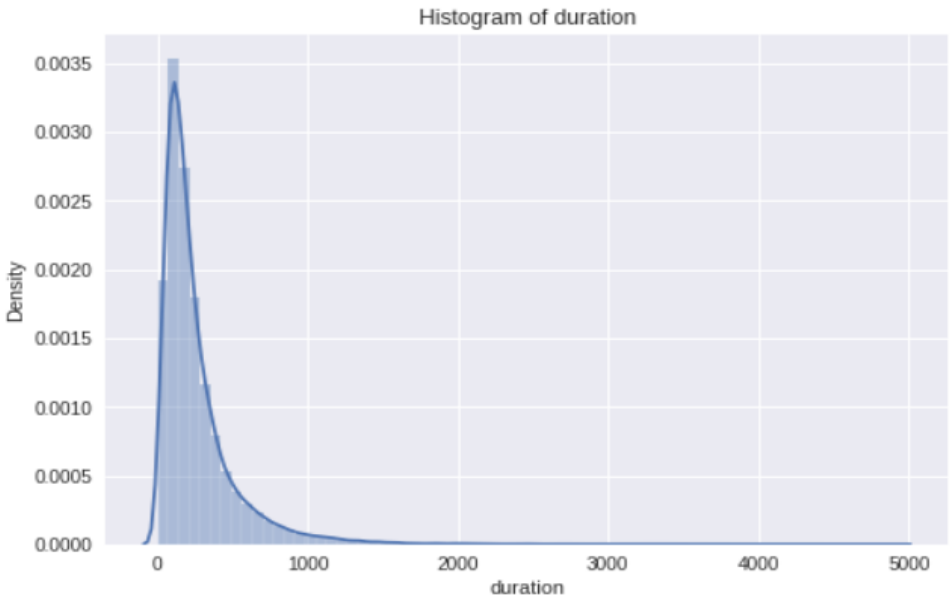
2.2.5. Phương thức liên lạc khách hàng.



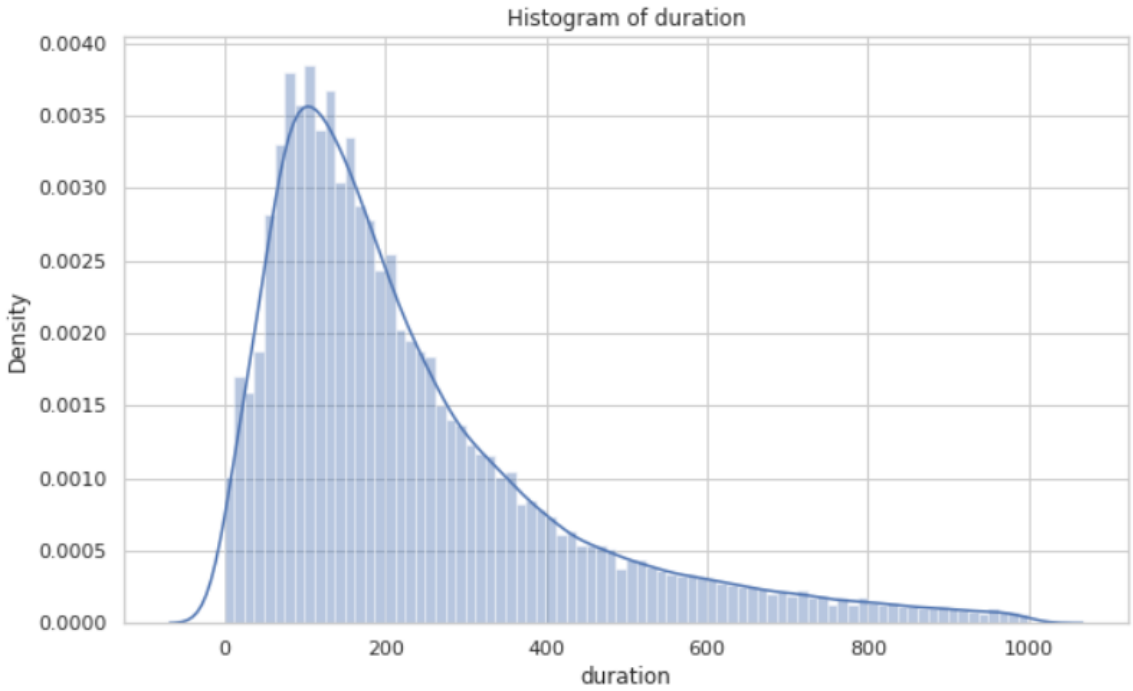
- Phương thức liên lạc ‘cellular’ có tỷ lệ chuyển đổi cao gấp 3 lần tỷ lệ chuyển đổi khi liên lạc với khách hàng bằng ‘telephone’.
- Công nghệ ngày càng phát triển, mạng di động càng phổ biến và khách hàng tiếp cận mạng di động nhiều hơn.
    - ⇨ Khi liên lạc với khách hàng chú ý phương thức liên lạc ‘cellular’, xây dựng những quảng cáo, dịch vụ qua phương thức này.

2.2.6. Thời lượng mỗi khi liên lạc với khách hàng.

a) Đồ thị phân phối của ‘duration’



- Theo đồ thị phân phối ‘duration’ ta thấy rằng dữ liệu tập trung chủ yếu dưới mốc thời gian 1000. Nên tạm thời ta sẽ bỏ các dữ liệu có giá trị > 1000 để nhìn rõ hơn về phân phối của các điểm dữ liệu.

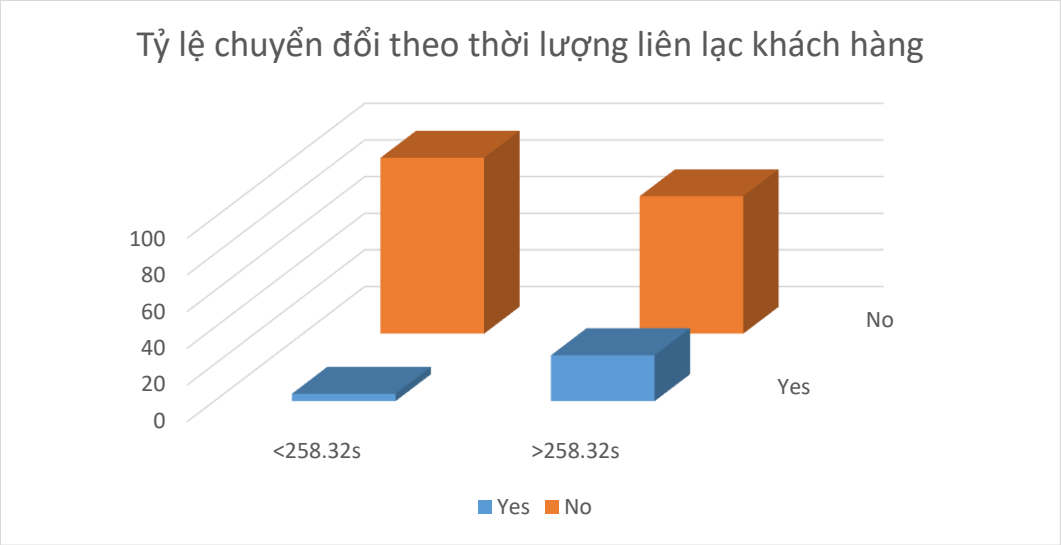


➔ Dễ thấy, dữ liệu tập trung chủ yếu từ 90 – 150s.

b) Nhận xét

- Thời lượng trung bình khi liên lạc với khách hàng (dựa theo bộ dữ liệu) là 258.32s.
- Chia khách hàng thành hai nhóm: nhóm có thời lượng liên lạc <258.32s và nhóm có thời lượng liên lạc >258.32s.

	Yes	No	Tổng
>258.32s	3441	10222	13663
<258.32s	1198	26315	27513



- Tổng số khách hàng có thời lượng liên lạc < 258.32s (27513 người) nhiều hơn gấp 2 lần so với tổng số khách hàng có thời lượng liên lạc > 258.32 (13663 người).
- Qua biểu đồ, khách hàng chấp nhận mở tài khoản tiết kiệm khi liên lạc với thời lượng trên 258.32s khá cao (tỷ lệ chuyển đổi lên đến 25%); cao gấp gần 6 lần so với khách hàng chấp nhận mở tài khoản khi có thời gian liên lạc dưới 258.32s.
- Khi thời gian tiếp xúc với khách hàng lâu => tăng độ tin tưởng và có nhiều thời gian thuyết phục khách hàng hơn.
    - ⇨ Thời lượng liên lạc khách hàng có ảnh hưởng lớn đến tỷ lệ chuyển đổi.

Phần 4: Xây dựng model.

1. Xử lý dữ liệu.

1.1. Loại bỏ và số hóa các feature.

- Loại bỏ features: “month”, “day_of_week”, “contact”.
- Số hóa các feature:

  + Education: “basic.4y”, “high.school”, “basic.6y”, “basic.9y”, “professional.course”, “university.degree”, “illiterate”, “unknown” thay bằng [0,1,2,3,4,5,6,np.nan].

  + Housing và Loan: thay ‘no’ thành 0, ‘yes’ thành 1, ‘unknown’ thành np.nan.
  Nhận thấy feature Housing và Loan đều là khoản nợ, nên ta gộp giá trị biến Housing và Loan.

```
df["loan"]=df["loan"]+df["housing"]
df = df.drop("housing",axis=1)
```

  + Số hóa các biến “job”, “marital”, “default”, “poutcome”, “campaign”.

```
o =['job','marital','default','poutcome','campaign']
labelencoder = LabelEncoder()
for c in o:
    X[c]      = labelencoder.fit_transform(X[c])
X.head()
```

+ "y" : thay 'no' thành 0 và 'yes' thành 1.

```
df["y"]=df["y"].replace("no",0)
df["y"]=df["y"].replace("yes",1)
```

1.2. Outlier và loại bỏ các outlier.

- Duration:

```
Q1 = df['duration'].quantile(.25)
Q3 = df['duration'].quantile(.75)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR
df = df[df['duration'] >= lower]
df = df[df['duration'] <=upper]
```

- Age:

```
Q1 = df['age'].quantile(.20)
Q3 = df['age'].quantile(.80)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR
df = df[df['age'] >= lower]
df = df[df['age'] <=upper]
```

- Previous:

```
Q1 = df['previous'].quantile(.20)
Q3 = df['previous'].quantile(.80)
IQR = Q3 - Q1
lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR
df = df[df['previous'] >= lower]
df = df[df['previous'] <=upper]
```

1.3. Thay thế những giá trị null thành các giá trị lân cận.

```
from sklearn.impute import KNNImputer
imputer = KNNImputer(n_neighbors=9)
X=pd.DataFrame(imputer.fit_transform(X),columns=X.columns)
```

2. Modeling

Chia model thành 2 tập dữ liệu là tập train và tập test với tỉ lệ 8:2.

```
from sklearn.model_selection import train_test_split
test_size = 0.2
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size, random_state=42)
```

2.1. Model logistic regression.

```
from sklearn.linear_model import LogisticRegression
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
logpred = logmodel.predict(X_test)
```

```
accuracy_log = accuracy_score(y_test, logpred)
accuracy_log
```

```
0.9499543656829936
Mean Absolute Error      :  0.05004563431700639
Mean Squared  Error      :  0.05004563431700639
Root Mean Squared  Error :  0.22370881591257505
R Squared Error          :  0.04069337570018161
```

⇨ Model logistic regression có chỉ số accuracy 95% và MSE là 22.37%.

2.2. Model Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(criterion='gini')
dtree.fit(X_train, y_train)
dtreepred = dtree.predict(X_test)
```

```
accuracy_dtree = accuracy_score(y_test, dtreepred)
accuracy_dtree
```

```
0.9353513842409492
Mean Absolute Error      :  0.0646486157590508
Mean Squared  Error      :  0.0646486157590508
Root Mean Squared  Error :  0.2542609206288902
R Squared Error          :  -0.23922588245417264
```

⇨ Model Decision Tree có chỉ số accuracy 93.53% và MSE là 25.43%

⇨ **Model Logistic Regression đưa ra dự đoán tốt hơn Model Decision Tree.**

Phần 5: Kiến nghị

- Số lượng khách hàng tiềm năm từ 18 – 60 tuổi rất lớn, nhưng tỷ lệ chuyển đổi rất thấp so với các nhóm còn lại => Cần tạo chiến dịch dành cho khách hàng tiềm năng ở nhóm tuổi này để tăng tỷ lệ chuyển đổi.
- Phương thức liên lạc ảnh hưởng rất nhiều đến tỷ lệ chuyển đổi: tỷ lệ chuyển đổi khi liên lạc với khách hàng bằng cellular cao hơn rất nhiều so với telephone => triển khai phương thức cellular nhiều hơn, tập trung đầu tư cho các dịch vụ của cellular, tiếp cận với khách hàng không chỉ chủ động liên lạc mà còn thông qua các bài viết (SEO).
- Thời gian liên lạc với khách hàng có tỷ lệ chuyển đổi cao là tháng 3, 9, 10, 12. Có thể thời gian này nhu cầu mở tài khoản tiết kiệm của khách hàng tăng => tìm hiểu nguyên nhân tại sao chạy chiến dịch marketing đến ít khách hàng nhưng hiệu quả và những tháng 4, 5, 6, 7, 8 và 11 lại có tỷ lệ chuyển đổi thấp.
- Thời lượng liên lạc với khách hàng là yếu tố quan trọng đến tỷ lệ chuyển đổi. Khi tư vấn cho khách hàng qua bất cứ phương thức nào nhân viên nên cố gắng làm quen, trò chuyện với khách hàng nhiều hơn, để hiểu rõ nhu cầu của họ.

Phần kết

Cảm ơn cô Nguyễn Thị Quỳnh Giang đã cho nhóm 6 cơ hội hoàn thành dự án phân tích khách hàng mở tài khoản tiết kiệm.

- Link code:
  - Analyst: Data Analysis
  - Science: Modeling
- Link data: Bank Customer

Đóng góp:

1. Đào Thị Hồng Nhung (leader) (25%): xử lý dữ liệu, xây dựng model, thuyết trình.
2. Nguyễn Thị Hoài Linh (25%): phân tích dữ liệu, tìm insight, làm slide.
3. Bùi Thị Mai Lương (30%): xử lý dữ liệu, xây dựng model, viết báo cáo.
4. Nguyễn Bá Đăng Khôi (20%): phân tích dữ liệu, tìm insight.