BÁO CÁO ĐỒ ÁN

Đồ án:

PREDICT TOI (TOTAL OPERATING INCOME) OF RETAIL BANKING CUSTOMER

Học viên thực hiện:

1. **TRẦN NAM PHONG – 0976614585**

TP. Hồ Chí Minh, ngày … tháng … năm …

GIỚI THIỆU ĐỒ ÁN

- TOI (Total operating income) là khái niệm kế toán để đo lường tổng lợi nhuận ngân hàng thu được từ tất cả các hoạt động kinh doanh của mình hàng năm sau khi trừ đi chi phí hoạt động và các chi phí phát sinh khác.

- Đồ án nhằm xây dựng công cụ/báo cáo dự đoán tổng lợi nhuận thu được của các khách hàng cá nhân sử dụng sản phẩm dịch vụ (SPDV) của ngân hàng như: Tiền gửi, tiền vay, dịch vụ thẻ tín dụng, phí dịch vụ và và hoạt động ngoại hối trong tương lai gần 6 tháng- 1 năm.

- Việc dự đoán chỉ số TOI là rất quan trọng để xác định lợi nhuận hàng năm phục vụ kế hoạch tăng trưởng của ngân hàng. Ngoài ra việc xác định các nhóm khách hàng tiềm năng thường xuyên mang lại TOI cao trong tương lai từ các nhóm SPDV cũng rất quan trọng để lập ra chiến lược tiếp cận bán hàng,chăm sóc khách hàng và phát triển kinh doanh cho ngân hàng trong tương lai.

MỤC LỤC

1 TỔNG QUAN 4

1.1 Giới thiệu 4
1.2 Thực trạng và giải pháp 5

2 CHUẨN BỊ VÀ LÀM SẠCH DỮ LIỆU (DATA ACQUISITION AND CLEANING) 7

2.1 Phân tích nghiệp vụ (Business understanding) 7
2.2 Lựa chọn dữ liệu (Data Source) 8
2.3 Làm sạch dữ liệu & Chuẩn hóa dữ liệu 10
2.4 Trích chọn thuộc tính (Feature Selection) 11

3 TÌM HIỂU DỮ LIỆU (EXPLORATORY DATA ANALYSIS) 15

3.1 Thống kê mô tả dữ liệu tiền gửi 15
3.2 Thống kê mô tả dữ liệu tiền vay 16
3.3 Mối quan hệ giữa số lượng KH tham gia đóng góp TOI 17
3.4 Mối quan hệ giữa tuổi KH , giới tính KH và TOI đóng góp 18
3.5 Mối quan hệ giữa BANK RELATION và TOI 21
3.6 Mối quan hệ giữa CUS_TARGET_CDE và TOI 22
3.7 Mối quan hệ giữa PROVINCE và TOI 23
3.8 Mối quan hệ giữa MARITAL_STATUS và TOI 24
3.9 Phân tích các thuộc tính của sản phẩm tiền gửi 24
3.10 Phân tích các thuộc tính của sản phẩm tiền vay 25

4 XÂY DỰNG MÔ HÌNH DỰ ĐOÁN TOI 26

4.1 Biến đổi dữ liệu (Data Transformation): 26
4.2 Lựa chọn mô hình dự đoán (Model Selection) 28

5 TRIỂN KHAI MÔ HÌNH 40

6 KẾT LUẬN (CONCLUSIONS) 41

7 ĐỀ XUẤT CẢI TIẾN (FUTURE DIRECTIONS) 41

8 TÀI LIỆU THAM KHẢO 42

1 TỔNG QUAN

1.1 Giới thiệu

**Giới thiệu đồ án:*

- *Hệ thống hỗ trợ xác định khách hàng tiềm năng khi khách hàng đến giao dịch/khi CVBH tiếp xúc với khách hàng, dự đoán được TOI khách hàng mang lại, đồng thời chào bán các sản phẩm phù hợp với nhu cầu khách hàng, nâng cao cơ hội bán hàng. Dựa trên dữ liệu về khách hàng và sản phẩm dịch vụ trên DW.*
- *Hệ thống tích hợp với app mWork hiện đang triển khai cho đội ngũ CVBH để tăng năng suất lao động cũng như nâng cao hiệu quả bán hàng, chăm sóc khách hàng tốt hơn.*

**Mục tiêu đồ án:*

- *Hỗ trợ phòng KHCN xây dựng công cụ/ báo cáo dự đoán giá trị TOI của khách hàng cá nhân mang lại cho Ngân hàng trong tương lai (6 tháng – 1 năm tới). Tích hợp kết quả dự đoán TOI vào màn hình mWork để hỗ trợ nhân viên tư vấn chăm sóc khách hàng tốt hơn.*
- *Phân nhóm các khách hàng tiềm năng có khả năng mang lại TOI cao trong từng nhóm sản phẩm dịch vụ (Deposit, Loan, Card, FX, FEE) cho ngân hàng trong tương lai, có cơ chế linh hoạt giá sản phẩm dịch vụ đối với các nhóm khách hàng tiềm năng này.*
- *Hỗ trợ dự phóng tổng nguồn thu (TOI) mang lại từ các nhóm sản phẩm dịch vụ (Deposit, Loan, Card, FX, FEE) của khác hàng cá nhân để dự báo, lập kế hoạch tăng trưởng lợi nhuận trong tương lai.*

**Phạm vi đồ án:*

- *Đối với sản phẩm tiền gửi (Deposit), đồ án tập trung phân tích các tài khoản trong vòng 4 năm (từ 2016-2019) cho các mã sản phẩm tiền gửi sau:*
    - *'-1--1003',*
    - *'10011--1003',*
    - *'10013--1003',*
    - *'10015--1003',*
    - *'10020--1003',*
    - *'10032--1003',*
    - *'10073--1003',*
    - *'10075--1003',*
    - *'10079--1003',*
    - *'10100--1003',*
    - *'11011--1003',*

- *'11015--1003',*
- *'11026--1003',*
- *'11032--1003'*

- *Đối với sản phẩm tiền vay (Loan), đồ án tập trung phân tích các tài khoản trong vòng 4 năm (từ 2016-2019) cho các mã sản phẩm tiền vay sau:*
  - *'60172--21060'*
  - *'60127--21060'*
  - *'60126--21050'*
  - *'60036--21060'*
  - *'60058--21050'*
  - *'60052--21060'*

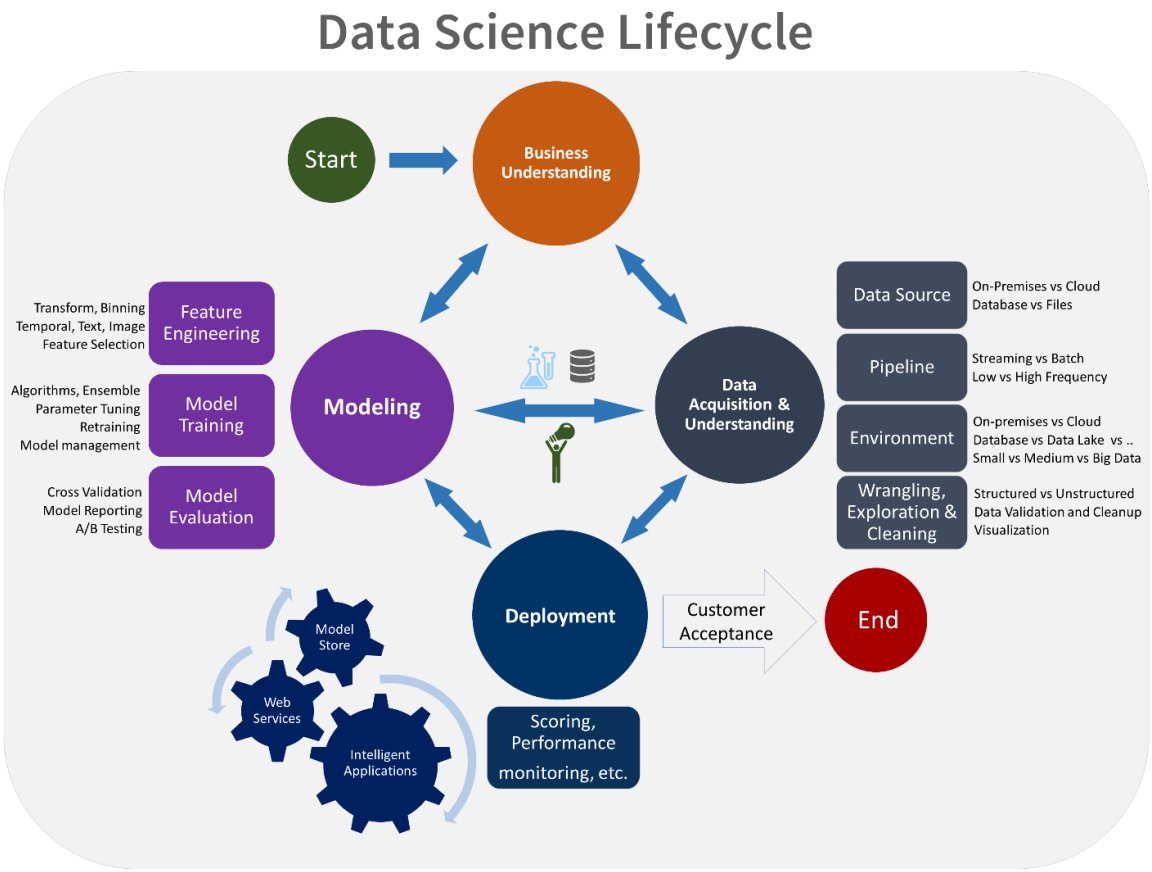
1.2 Thực trạng và giải pháp

- Thực trạng:

- Hiện tại phòng KHCN chưa có 1 công cụ chính thức để hỗ trợ dự báo giá trị TOI mang lại của các khách hàng cá nhân trong tương lai. Nhằm phục vụ cho chiến lược kinh doanh, chăm sóc và hỗ trợ cho các khách hàng tiềm năng.
- Bộ phận Phát Triển MIS của Trung tâm Phát Triển Ứng Dụng đã xây dựng và lưu trữ được 1 nguồn dữ liệu liên qua đến thông tin khách hàng (Customer Demographic), thông tin giao dịch khách hàng (Interactive Transactions), thông tin TOI quá khứ của khác hàng, cũng như thông tin sản phẩm (Product master) trên Data Warehouse từ năm 2012 đến nay. Đây là nguồn thông tin quý giá hỗ trợ cho việc phân tích/ xây dựng mô hình dự báo TOI trong tương lai cho khách hàng cá nhân.

- Giải pháp:

- Giải pháp tổng thể của mô hình dự đoán TOI cho khách hàng cá nhân tuân theo qui trình của 1 dự án phân tích dữ liệu tổng quát (Data Science Process) là **1 vòng lặp** các bước chi tiết như sau:

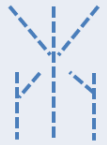


(Nguồn: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle-modeling)

- Áp dụng kỹ thuật học máy (Machine Learning) thuộc nhóm Supervised Learning (học có giám sát) để dự đoán giá trị TOI mang lại đối tượng Khách hàng cá nhân. Cụ thể là sử dụng nhóm giải thuật hồi quy (Regression models).

aisoma.de

Top Machine Learning Algorithms for Predictions

Name	Type	Description	Advantages	Disadvantages
Linear Regression		-The best fit line through all data points	-Easy to understand -you can clearly see what the biggest drivers odf the model are.	-sometimes to simple to capture cpmöex relationships between variables, -Tendency für the model to overfit.
Logistic Regression		-The adoption for linear regression to problembs of classification	-Easy to understand	-sometimes to simple to capture cpmöex relationships between variables, -Tendency für the model to overfit.
Decision Tree		-A graph that uses branching method to match all possible outcomes of a decision	-Easy to understand and implement.	-Not often use of ist own for prediction because it's also often too simple and not powerful enough for complex data.
Random Forest		- Takes the average of many decision trees. Each tree is weaker than the full decision tree, but combining them we get better overall performance.	-A sort of „wisdom of the crowd", Tend to result in very high quality results. -Fast to train	-Can be slow to output predictions relative to other algorithms. -Not easy to understand predictions.
Gradient Boosting		-Uses even weaker decision trees that increasingly focused on „hard examples"	-High-performing	-A small change in the future set or training set can create radical changes in the model. -Not easy to understand predictions.
Neural Networks		-Mimics the behaviour of the brain. NNs are interconnected Neurons that pass messages to each other. Deep Learning uses severak layers of NNs to put one after the other.	-Can handle extremely complex tasks. No other alsgorithm comes close in image recognition.	-very very slow to train. Because they have so many layers. Require a lot of power. -Almost impossible to understand predictions.

Hình: Các mô hình giải thuật hồi quy (Regression models).

2 CHUẨN BỊ VÀ LÀM SẠCH DỮ LIỆU (DATA ACQUISITION AND CLEANING)

2.1 Phân tích nghiệp vụ (Business understanding)

- Làm việc với các đơn vị có liên quan để xác định vấn đề và làm rõ yêu cầu phía nghiệp vụ:
  - ➢ Xác định các nhóm sản phẩm mang lại TOI bao gồm 2 nhóm:
    - Sản phẩm tiền gửi (Deposit)
    - Sản phẩm tiền vay (Loan)

- Xác định đối tương Khách hàng cho mô hình dự đoán:
    - Khách hàng trong mô hình dự đoán TOI thành phần là khách hàng đã hoặc mới tham gia sử dụng ít nhất 1 hoặc nhiều sản phẩm tại thời điểm dự đoán.
- Mục tiêu:
    - Dự đoán TOI đóng góp trong tương lai của KH trong vòng 6 tháng/ 1 năm tới gồm:
        - TOI tổng = $\sum_1^2 TOI(i)$
        - Các TOI thành phần:
            - NII huy động.
            - NII Cho vay.
- Tìm hiểu hiện trạng các ứng dụng, công thức dự phóng TOI hiện tại P. KHCN đang sử dụng.
- Phối hợp với các BP. MIS, BP. TVGP và P. KHCN để đưa ra các kịch bản giả định, thuộc tính tiềm năng cho mô hình dự đoán (model), đưa ra các tiêu chí kiểm định và đánh giá kết quả (vd: A/B Testing).

2.2 Lựa chọn dữ liệu (Data Source)

- Xác định các nguồn dữ liệu hiện có trên DWH phục vụ cho bài toán phân tích dự đoán TOI gồm (Data Source):
    - Các bảng dữ liệu liên quan sao kê tiền gửi.
    - Các bảng dữ liệu liên quan sao kê tiền vay.
    - Các bảng dữ liệu liên sao kê quan thẻ.
    - Các bảng dữ liệu giao dịch liên quan phí dịch vụ.
    - Các bảng dữ liệu liên quan sao kê ngoại hối.
    - Các bảng dữ liệu liên quan thông tin KH.
    - Các bảng dữ liệu chứa TOI thành phần quá khứ của KH.
- Xây dựng STORE tạo bảng dữ liệu input phục vụ cho việc xây dựng model (Pipeline).
- Tạo các bảng output cho mô hình dự đoán:
    - DW_DA_TOI_SCORING_OUTPUT_FCT: Chứa thông tin kết quả dự đoán của model.

Stt	Tên cột	Kiểu dữ liệu	Mô tả
1	COMPANY_KEY	NUMBER (4)	
2	COMPANY_CDE	VARCHAR2 (15 Byte)	
3	DAY_KEY	NUMBER (7)	
4	SUB_BRANCH_KEY	NUMBER (6)	
5	SUB_BRANCH_CDE	VARCHAR2 (15 Byte)	
6	SUB_INDUSTRY_CDE	VARCHAR2 (15 Byte)	
7	CUSTOMER_CDE	VARCHAR2 (15 Byte)	
8	SUB_SECTOR_CDE	VARCHAR2 (15 Byte)	
9	SUB_SEGMENT_CDE	VARCHAR2 (15 Byte)	
10	CUS_TARGET_CDE	VARCHAR2 (15 Byte)	

11	MODEL_CDE	VARCHAR2 (50 Byte)	
12	PREDICT_YEAR	VARCHAR2 (15 Byte)	
13	TOI LOAN	NUMBER (38,6)	
14	TOI_DP	NUMBER (38,6)	
15	TOI_CARD	NUMBER (38,6)	
16	TOI_FX	NUMBER (38,6)	
17	TOI_FEE	NUMBER (38,6)	
18	TOI	NUMBER (38,6)	
19	PROCESS_DT	DATE	
20	REC_CREATE_DT	DATE	
21	REC_UPDATE_DT	DATE	

- DW_DA_EVALUATION_FCT: Chứa thông tin kết quả đánh giá của mô hình

Stt	Tên cột	Kiểu dữ liệu	Mô tả
5	EVALUATE_DT	DATE	Ngày thực hiện đánh giá
6	MODEL_CDE	VARCHAR2 (50 Byte)	Mã mô hình
8	PROCESS_DT	DATE	Ngày xử lý
9	R_SQUARED	FLOAT (126)	Độ phù hợp dữ liệu của mô hình dự đoán
10	REC_CREATE_DT	DATE	Ngày tạo
11	REC_UPDATE_DT	DATE	Ngày cập nhật
12	RECALL	FLOAT (126)	Độ nhầm lẫn khi phân loại
13	RMSE	FLOAT (126)	Độ sai lệch trung bình của mô hình dự đoán
14	SCORE_CDE	VARCHAR2 (50 Byte)	Mã ghi nhận

- DW_DA_MODEL_DIM: Chứa thông tin mô hình.

Stt	Tên cột	Kiểu dữ liệu	Mô tả
2	ACTIVE	INTEGER	Tình trạng hoạt động của mô hình
4	DELETED	VARCHAR2 (5 Byte)	Tình trạng của mô hình (1: Xóa/ 0: Không xóa)
5	DESCRIPTION	VARCHAR2 (4000 Byte)	Mô tả mô hình
8	MODEL_CDE	VARCHAR2 (50 Byte)	Mã mô hình
9	MODEL_NAME	VARCHAR2 (200 Byte)	Tên mô hình
10	PATH_FILE	VARCHAR2 (4000 Byte)	Đường dẫn đến file mô hình
12	R_SQUARED	FLOAT (126)	Độ phù hợp
13	REC_CREATE_DT	DATE	Ngày tạo
14	REC_UPDATE_DT	DATE	Ngày cập nhật
16	RMSE	FLOAT (126)	Độ lỗi của mô hình dự đoán
17	VERSION	FLOAT (126)	Thế hệ mô hình

- Cái đặt môi trường phát triển (Environment):
  - Cài đặt môi trường DB chứa dữ liệu input và output cho Model: Oracle
  - Môi trường DEV, UAT, Production các model dự đoán:

- Operating System: Microsoft Windows Server 2018
- Tool & Packages: Anaconda (Python 3.7 or later)

2.3 Làm sạch dữ liệu & Chuẩn hóa dữ liệu

- Sử dụng 1 số kỹ thuật làm sạch dữ liệu cho các bảng input như sau:
  - Xử lý dữ liệu bị thiếu (missing data)

```
RangeIndex: 1563702 entries, 0 to 1563701
Data columns (total 16 columns):
CUSTOMER_CDE     1563702 non-null object
PRODUCT_CDE      1563702 non-null object
LD_ID            1563702 non-null object
AMT_INIT         1563702 non-null float64
AMT_CUR          1563702 non-null float64
INTEREST_RATE    1563702 non-null float64
RATE_FTP         1563702 non-null float64
KYHAN            1563702 non-null int64
LOAITRAGOP       1563702 non-null object
MUCDICHVAY       1563702 non-null object
LMV              1563702 non-null float64
LBV              1563702 non-null float64
ACCT_USE_DAYS    1563702 non-null int64
TOI              1563702 non-null float64
PROCESS_MONTH    1563702 non-null object
PROCESS_YEAR     1563702 non-null object
dtypes: float64(7), int64(2), object(7)
memory usage: 190.9+ MB
```

Hình: Thông tin các thuộc tính trong bảng input LOAN

```
Out[8]: CUSTOMER_CDE     0
        PRODUCT_CDE      0
        LD_ID            0
        AMT_INIT         0
        AMT_CUR          0
        INTEREST_RATE    0
        RATE_FTP         0
        KYHAN            0
        LOAITRAGOP       0
        MUCDICHVAY       0
        LMV              0
        LBV              0
        ACCT_USE_DAYS    0
        TOI              0
        PROCESS_MONTH    0
        PROCESS_YEAR     0
        dtype: int64
```

Hình : Kiểm tra dữ liệu null trong bảng input LOAN

```
Out[7]: CUSTOMER_CDE          0
        PRODUCT_CDE           0
        ACCT_ID               0
        AMT_INIT              0
        AMT_CUR               0
        INTEREST_RATE     76892
        RATE_FTP              0
        SUB_TERM_ID           0
        LMV                   0
        LBV                   0
        ACCT_USE_DAYS         0
        TOI                   0
        PROCESS_MONTH         0
        PROCESS_YEAR          0
        dtype: int64
```

Hình: Kiểm tra dữ liệu null trong bảng input DEPOSIT

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
- Xử lý dữ liệu không nhất quán (inconsistent data)
- Dùng 1 số kỹ thuật xử lý làm sạch dữ liệu trên:
  - PL/SQL Query
  - Python

2.4 Trích chọn thuộc tính (Feature Selection)

- DW_DA_CUSTOMER_DIM: Chứa thông tin tổng hợp của Khách hàng phục vụ cho việc phân tích dữ liệu

Stt	Tên cột	Kiểu dữ liệu	Mô tả
1	CUSTOMER_CDE	VARCHAR2 (15 Byte)	Mã Khách hàng
2	COMPANY_KEY	NUMBER (4)	
3	COMPANY_CDE	VARCHAR2 (15 Byte)	
4	SUB_SECTOR_CDE	VARCHAR2 (15 Byte)	Mã khu vực
5	DAO_CDE	VARCHAR2 (15 Byte)	Mã người giới thiệu
6	OTHER_OFFICER_CDE	VARCHAR2 (15 Byte)	
7	SUB_INDUSTRY_CDE	VARCHAR2 (15 Byte)	Mã ngành nghề
8	SUB_SEGMENT_CDE	VARCHAR2 (15 Byte)	Mã vùng
9	CUS_TARGET_CDE	VARCHAR2 (15 Byte)	Mã khu vực
10	CUSTOMER_STATUS_CDE	VARCHAR2 (15 Byte)	Tình trạng khách hàng
11	SUB_BRANCH_KEY	NUMBER (6)	Mã chi nhánh
12	GENDER_CDE	VARCHAR2 (15 Byte)	Giới tính
13	CIF_CDE	VARCHAR2 (15 Byte)	Mã T24
14	SHORT_NAME	VARCHAR2 (250 Byte)	Tên rút gọn
15	FULL_NAME	VARCHAR2 (250 Byte)	Tên đầy đủ
16	STREET	VARCHAR2 (250 Byte)	Đường

17	ADDRESS	VARCHAR2 (250 Byte)	Địa chỉ
18	TOWN_COUNTRY	VARCHAR2 (200 Byte)	Vùng
19	POST_CDE	VARCHAR2 (15 Byte)	Mã bưu chính
20	COUNTRY	VARCHAR2 (100 Byte)	Quốc gia
21	NATIONALITY	VARCHAR2 (350 Byte)	Quốc tịch
22	RELATION_CDE	VARCHAR2 (550 Byte)	Mã quan hệ
23	REL_CUSTOMER	VARCHAR2 (1400 Byte)	#
24	RESIDENCE	VARCHAR2 (25 Byte)	#
25	CONTACT_DT	DATE	#
26	INTRODUCER	VARCHAR2 (25 Byte)	Người giới thiệu
27	LEGAL_ID	VARCHAR2 (100 Byte)	#
28	LEGAL_DOC_NAME	VARCHAR2 (100 Byte)	#
29	LEGAL_HOLDER_NAME	VARCHAR2 (100 Byte)	#
30	LEGAL_ISS_AUTH	VARCHAR2 (100 Byte)	#
31	LEGAL_ISS_DT	VARCHAR2 (4000 Byte)	#
32	LEGAL_EXP_DT	VARCHAR2 (4000 Byte)	#
33	OFFICE_PHONE	VARCHAR2 (255 Byte)	#
34	REVIEW_FREQUENCY	VARCHAR2 (25 Byte)	#
35	BIRTHDAY	DATE	Ngày sinh
36	ISSUE_CHEQUES	VARCHAR2 (25 Byte)	#
37	MARITAL_STATUS	VARCHAR2 (25 Byte)	Tình trạng hôn nhân
38	NO_OF_DEPENDENTS	NUMBER	
39	HOME_PHONE	VARCHAR2 (25 Byte)	Số điện thoại cá nhân
40	SMS_PHONE	VARCHAR2 (255 Byte)	Số điện thoại nhận sms
41	EMAIL	VARCHAR2 (100 Byte)	Địa chỉ thư điện tử
42	EMP_STATUS	VARCHAR2 (50 Byte)	Tình trạng công việc
43	OCCUPATION	VARCHAR2 (50 Byte)	Nghề nghiệp
44	JOB_TITLE	VARCHAR2 (50 Byte)	Chức danh
45	EMP_COMPANY_NAME	VARCHAR2 (100 Byte)	Tên công ty khách hàng
46	SALARY	NUMBER	Thu nhập khách hàng
47	FAX	VARCHAR2 (25 Byte)	Số fax
48	LOCAL_REF	VARCHAR2 (4000 Byte)	#
49	CURR_NO	NUMBER	#
50	INPUTER	VARCHAR2 (500 Byte)	#
51	CUS_OPEN_DT	DATE	Ngày mở tài khoản thanh toán
52	INPUT_DT	DATE	#
53	AUTH_DT	DATE	#
54	AUTHORISER	VARCHAR2 (50 Byte)	#
55	DEPT_CDE	VARCHAR2 (25 Byte)	#
56	MAIN_CLASS	VARCHAR2 (15 Byte)	#
57	BANK_RELATION	VARCHAR2 (15 Byte)	Quan hệ ngân hàng
58	CONTACT_NAME	VARCHAR2 (250 Byte)	

59	CONTACT_POSITION	VARCHAR2 (25 Byte)	#
60	CONTACT_PHONE	VARCHAR2 (100 Byte)	#
61	CONTACT_EMAIL	VARCHAR2 (100 Byte)	#
62	CONTACT_REMARKS	VARCHAR2 (250 Byte)	#
63	TOTAL_CAPITAL	VARCHAR2 (250 Byte)	#
64	CUS_REFERRED_EMP	NUMBER	#
65	NO_OF_EMP	NUMBER	#
66	COMPANY_BOOK	VARCHAR2 (250 Byte)	#
67	ATTR2	VARCHAR2 (250 Byte)	#
68	COMPANY_USER	VARCHAR2 (250 Byte)	#
69	CUS_EMPLOYERS_ADD	VARCHAR2 (500 Byte)	#
70	ATTR5	VARCHAR2 (250 Byte)	#
71	OTHER_NATIONAL	VARCHAR2 (1000 Byte)	#
72	SEGMENT_TYPE	VARCHAR2 (250 Byte)	#
73	LIABILITY	VARCHAR2 (250 Byte)	#
74	GB_NAME	VARCHAR2 (250 Byte)	#
75	POSTING_RESTRICT	VARCHAR2 (250 Byte)	#
76	ACTIVE	INTEGER	#
77	CREATE_DT	DATE	Ngày tạo
78	UPDATE_DT	DATE	Ngày cập nhật
79	CLOSE_DT	DATE	Ngày đóng
80	REC_CREATE_DT	DATE	#
81	REC_UPDATE_DT	DATE	#
82	POSTING_RESTRICT_1	VARCHAR2 (10 Byte)	#

- DW_DA_TOI_LOAN_INPUT_FCT: Chứa thông tin các thành phần cần thiết (Dữ liệu kết quả sau bước phân tích nghiệp vụ) phục vụ cho việc xây dựng mô hình dự đoán cho Loan:

Stt	Tên cột	Kiểu dữ liệu	Mô tả
1	CUSTOMER_CDE	VARCHAR2 (15 Byte)	Mã khách hàng
2	AGE	NUMBER	Tuổi
3	PRODUCT_CDE	VARCHAR2(15 Byte)	Mã sản phẩm
4	LD_ID	VARCHAR2(15 Byte)	Số tài khoản vay
5	AMT_INIT	NUMBER	Số tiền vay ban đầu
6	AMT_CUR	NUMBER	Dư nợ
7	ACCT_USE_DAYS	NUMBER	Số ngày phát sinh TOI
8	INTEREST_RATE	NUMBER	Lãi suất vay cố định
9	RATE_FTP	NUMBER	Lãi suất FTP
10	LMV	NUMBER	Lãi mua vốn
11	LBV	NUMBER	Lãi bán vốn
12	KYHAN	NUMBER	Kỳ hạn khoản vay
13	LOAITRAGOP	VARCHAR2(15 Byte)	Loại hình trả góp

14	MUCDICHVAY	VARCHAR2(15 Byte)	Mục đích khoản vay
15	PROCESS_MONTH	NUMBER	Tháng quan sát
16	PROCESS_YEAR	NUMBER	Năm quan sát
17	PROCESS_DT	DATE	Ngày xử lý
18	REC_CREATE_DT	DATE	Ngày tạo
19	REC_UPDATE_DT	DATE	Ngày cập nhật

- DW_DA_TOI_DEPOSIT_INPUT_FCT: Chứa thông tin các thành phần cần thiết phục vụcho việc xây dựng mô hình dự đoán cho Loan (Dữ liệu kết quả sau bước phân tích nghiệp vụ):.

Stt	Tên cột	Kiểu dữ liệu	Mô tả
1	CUSTOMER_CDE	VARCHAR2 (15 Byte)	Mã khách hàng
2	AGE	NUMBER	Tuổi
3	PRODUCT_CDE	VARCHAR2(15 Byte)	Mã sản phẩm
4	ACCT_ID	VARCHAR2(15 Byte)	Số tài khoản tiền gửi
5	AMT_INIT	NUMBER	Số tiền gửi ban đầu
6	AMT_CUR	NUMBER	Số dư tiền gửi hiện tại
7	ACCT_USE_DAYS	NUMBER	Số ngày phát sinh TOI
8	INTEREST_RATE	NUMBER	Lãi suất tiền gửi
9	RATE_FTP	NUMBER	Lãi suất FTP
10	LMV	NUMBER	Lãi mua vốn
11	LBV	NUMBER	Lãi bán vốn
12	KYHAN	NUMBER	Kỳ hạn khoản vay
13	PROCESS_MONTH	NUMBER	Tháng quan sát
14	PROCESS_YEAR	NUMBER	Năm quan sát
15	PROCESS_DT	DATE	Ngày xử lý
16	REC_CREATE_DT	DATE	Ngày tạo
17	REC_UPDATE_DT	DATE	Ngày cập nhật

3 TÌM HIỂU DỮ LIỆU (EXPLORATORY DATA ANALYSIS)

3.1 Thống kê mô tả dữ liệu tiền gửi

	PRODUCT_CDE	count
10	11011--1003	161365
13	11032--1003	152390
3	10015--1003	49076
0	-1--1003	27652
6	10073--1003	4978
2	10013--1003	4140
7	10075--1003	2853
8	10079--1003	1173
4	10020--1003	418
5	10032--1003	52
1	10011--1003	42
9	10100--1003	35
11	11015--1003	25
12	11026--1003	16

Bảng: Thống kê số lượng tài khoản theo nhóm SPDV Tiền gửi từ năm 2016-2019

*Nhận xét:

- Sản phẩm 11011-1003 và 11032-1003 có số lượng tài khoản chiếm nhiều nhất lần lượt là 161.365 và 152.390.
- Nhóm sản phẩm 10032-1003,10011-1003,13100-1003,11015-1003,11026-1003 có số lượng tài khoản ít nhất là 16.
- Tỉ lệ tài khoản mở trong các nhóm SPDV tiền gửi là không cân bằng nhau.

Out[113]:

	AMT_INIT	AMT_CUR	ACCT_USE_DAYS	INTEREST_RATE	RATE_FTP	TOI
count	7.517330e+05	7.517330e+05	751733.000000	674841.000000	751733.000000	7.517330e+05
mean	1.205472e+07	1.239948e+07	30.287702	0.193662	5.560265	5.495521e+04
std	1.726676e+08	1.807178e+08	1.908243	0.143505	0.232131	8.113396e+05
min	0.000000e+00	0.000000e+00	1.000000	0.000000	0.850000	0.000000e+00
25%	5.000000e+04	5.000000e+04	30.000000	0.000000	5.400000	2.219178e+02
50%	5.566400e+04	5.740100e+04	31.000000	0.300000	5.500000	2.689178e+02
75%	6.162500e+05	1.341604e+06	31.000000	0.300000	5.500000	5.977213e+03
max	3.397118e+10	5.879732e+10	31.000000	0.300000	6.000000	2.885968e+08

Hình: Thống kê các thuộc tính numerical trong bảng DEPOSIT

3.2 Thống kê mô tả dữ liệu tiền vay

	LOAITRAGOP	count
1	E	816171
0	B	580896
2	S	166635

Bảng: Thống kê số lượng tài khoản tiền vay theo loại hình trả góp (từ năm 2016-2019)

*Ghi chú: S: Loại hình vay thẻ dư nợ góp đều.
E, B: Loại hình vay theo dư nợ giảm dần.

	PRODUCT_CDE	count
5	60172--21060	314881
4	60127--21060	302696
3	60126--21050	259144
0	60036--21060	248730
2	60058--21050	225779
1	60052--21060	212472

Bảng: Thống kê số lượng tài khoản tiền vay theo sản phẩm tiền vay (từ năm 2016-2019)

*Nhận xét:

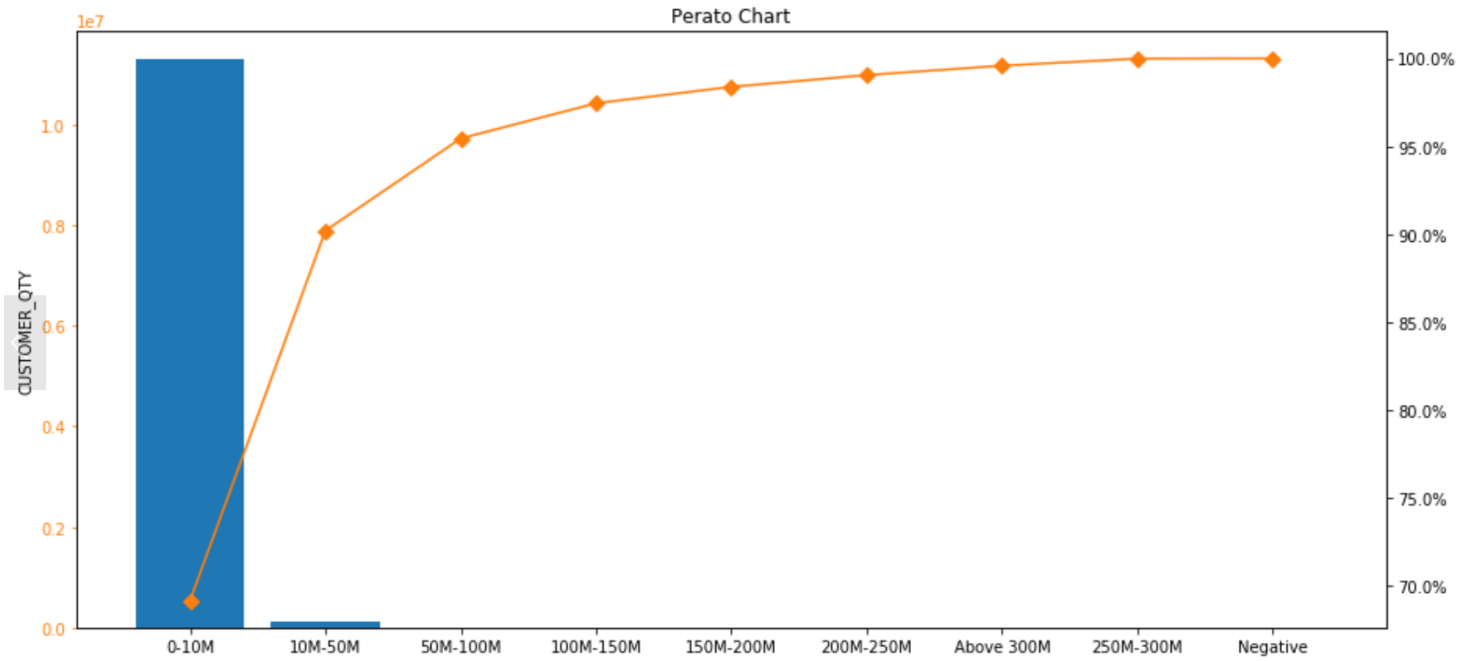
- Sản phẩm 60172-21060 có số lượng tài khoản chiếm nhiều nhất là 314881.
- Sản phẩm 60052-21060 có số lượng tài khoản ít nhất là 212472.
  Tỉ lệ tài khoản mở trong các nhóm SPDV tiền vay là cân bằng nhau.

Out[14]:

	AMT_INIT	AMT_CUR	ACCT_USE_DAYS	INTEREST_RATE	RATE_FTP	KYHAN	LMV	LBV	TOI
count	1.563702e+06	1.563702e+06	1.563702e+06	1.563702e+06	1.563702e+06	1.563702e+06	1.563702e+06	1.563702e+06	1.563702e+06
mean	2.992100e+08	2.136678e+08	2.657697e+01	1.266215e+01	7.681331e+00	3.542651e+01	1.899740e+06	1.408624e+06	4.911161e+05
std	3.408625e+08	2.850449e+08	3.454984e+01	4.370297e+00	1.467862e+00	2.748858e+01	4.722458e+06	1.761666e+06	4.236072e+06
min	2.702000e+03	0.000000e+00	-1.771000e+03	3.000000e-01	2.930000e+00	1.000000e+00	0.000000e+00	-9.542947e+06	-1.620469e+07
25%	8.000000e+07	3.000000e+07	3.000000e+01	1.050000e+01	7.066806e+00	1.200000e+01	5.197808e+05	2.899726e+05	1.874658e+05
50%	2.000000e+08	1.300000e+08	3.000000e+01	1.150000e+01	7.540000e+00	3.600000e+01	1.245808e+06	9.002917e+05	3.484055e+05
75%	4.000000e+08	3.000000e+08	3.100000e+01	1.255000e+01	9.070000e+00	6.000000e+01	2.589041e+06	1.912611e+06	6.320377e+05
max	2.998200e+10	2.400000e+10	3.100000e+01	3.640000e+01	1.379000e+01	1.082000e+03	5.262330e+09	1.272921e+08	5.260677e+09

Hình: Thống kê các thuộc tính numerical trong bảng LOAN

3.3 Mối quan hệ giữa số lượng KH tham gia đóng góp TOI



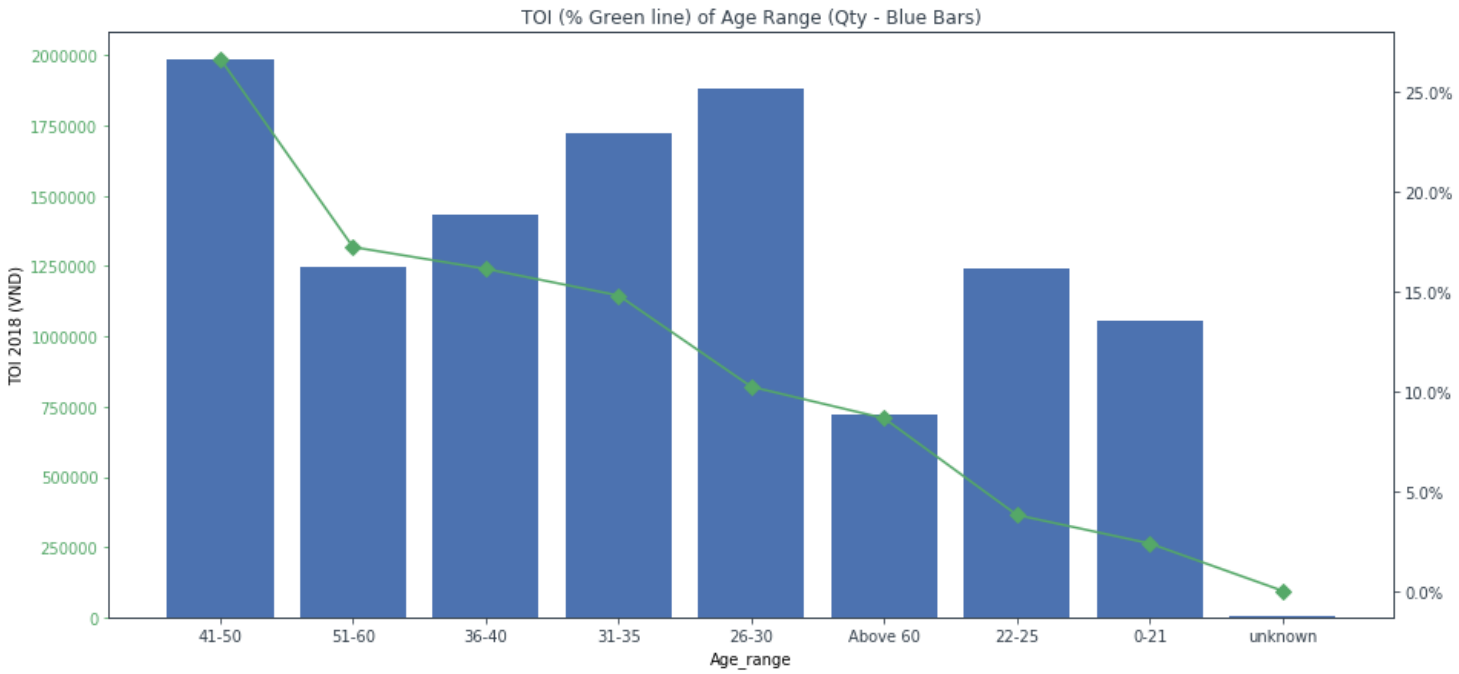
Hình: Biểu đồ biểu diễn quan hệ giữa lượng KH sử dụng SPDV và TOI (năm 2018)
Ghi chú: Cột xanh đại diện cho số lượng KH ở từng nhóm , đường màu cam đại diện cho mức đóng góp TOI của KH ở từng level đóng góp TOI.

*Nhận xét:

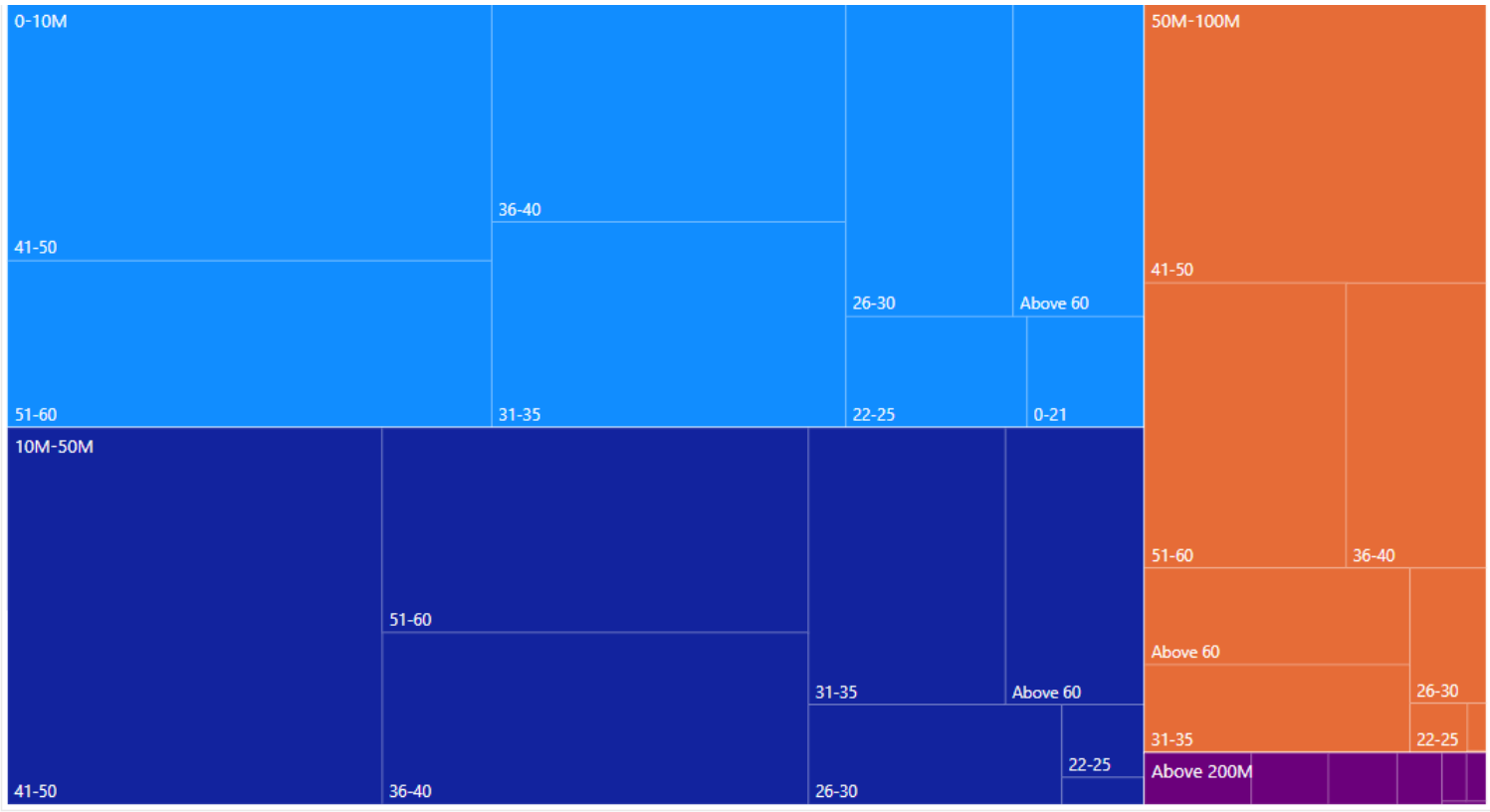
- KH đóng góp TOI trong khoảng từ 0-10 triệu chiếm 98% lượng KH mang lại TOI cho NH năm 2018.

- KH đóng góp TOI trong khoảng từ 10 triệu chiếm 2% lượng KH mang lại TOI cho NH năm 2019.
- Lượng KH đóng góp TOI càng cao (trên 50 triệu) thì có tỉ lệ càng thấp.
- Nhóm KH có đóng góp TOI trên 10 triệu (một năm) tuy có số lượng KH thấp (ít hơn 10% trên tổng lượng) nhưng đóng góp TOI trên tổng KH ~ 50% tổng giá trị TOI của NH.

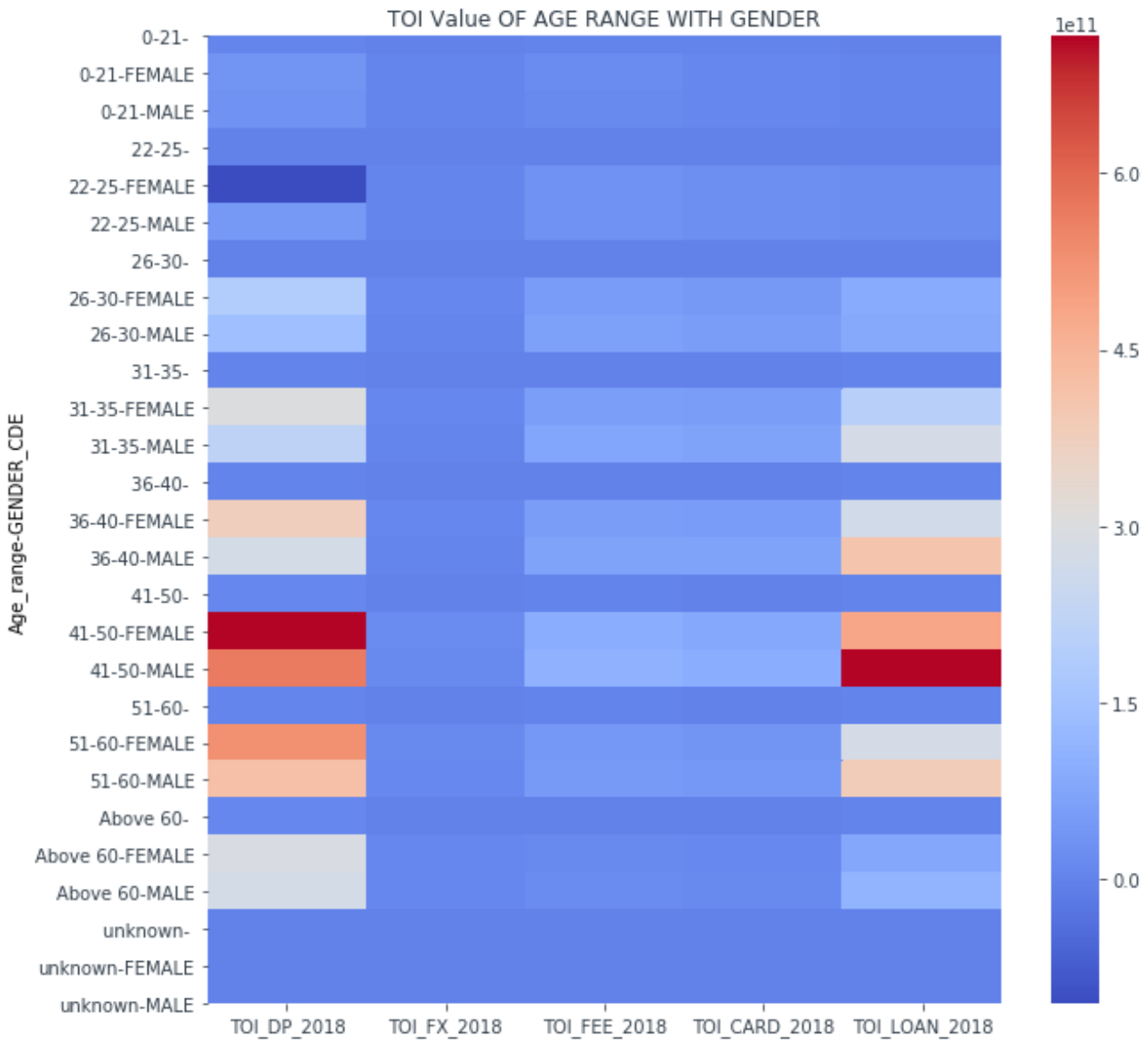
3.4 Mối quan hệ giữa tuổi KH , giới tính KH và TOI đóng góp



Hình: Biểu đồ biểu diễn quan hệ giữa AGE_RANGE và TOI (năm 2018)



Hình: Biểu đồ phân bố TOI theo nhóm tuổi KH (năm 2018)

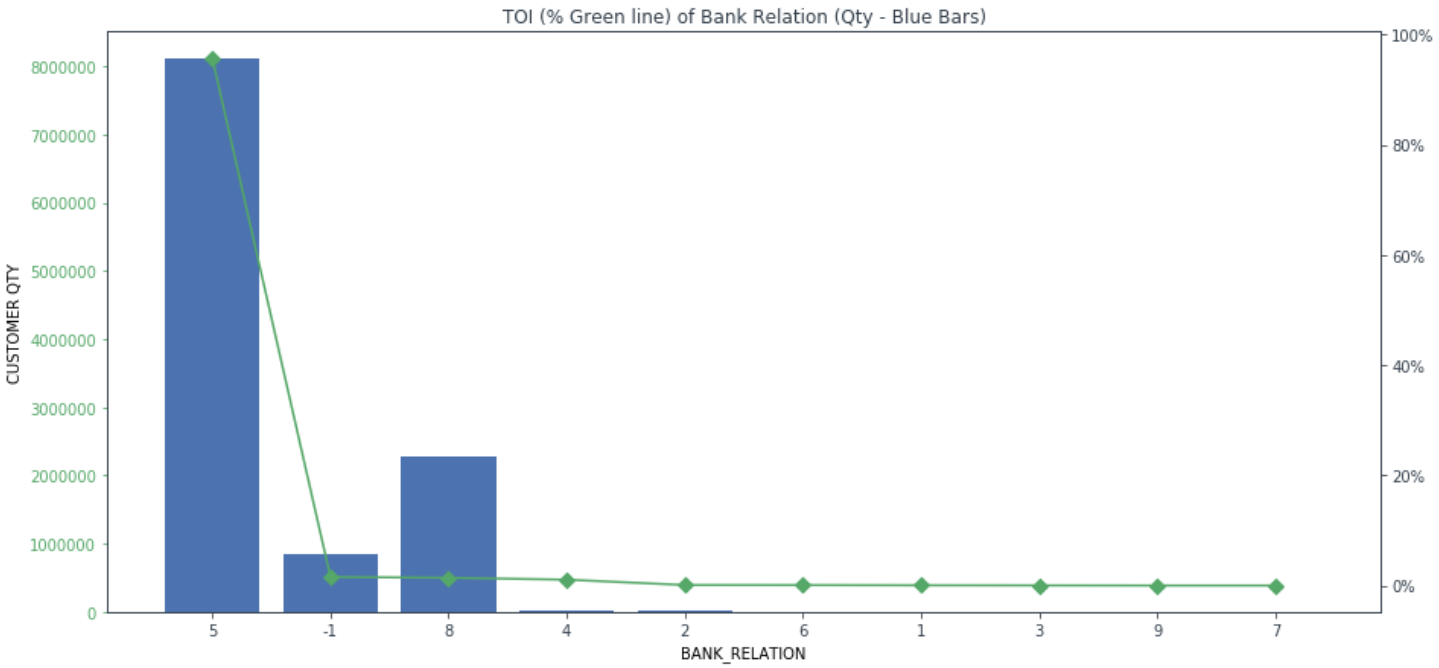


Hình: Biểu đồ thể hiện quan hệ giữa tuổi, giới tính và TOI đóng góp của KH (năm 2018)

*Nhận xét: Từ biểu đồ ta thấy

- Nhóm KH từ 41-50 tuổi mang lại tổng giá trị TOI nhiều hơn cho NH hơn.
- KH nữ trong nhóm này có xu hướng sử dụng tiền gửi nhiều hơn so với nam giới, ngược lại nam giới có xu hướng sử dụng tiền vay nhiều hơn.

3.5 Mối quan hệ giữa BANK RELATION và TOI



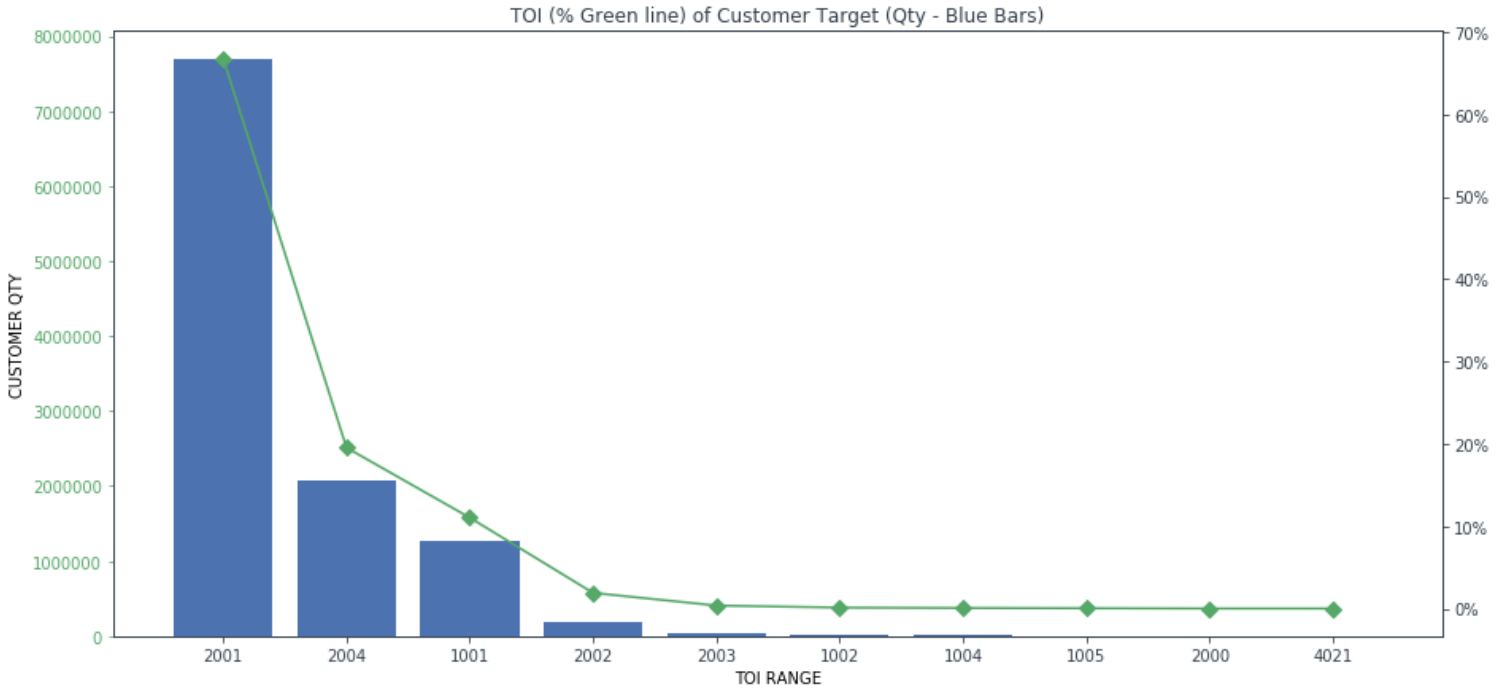
Hình: Biểu đồ biểu diễn quan hệ giữa BANK RELATION và TOI (năm 2018)

*Ghi chú : Cột xanh đại diện cho số lượng khách hàng , đường màu xanh đại diện cho mức đóng góp TOI của KH theo BANK_RELATION.

*Nhận xét:

- KH có BANK_RELATION = 5 chiếm số lượng (~ 8 triệu) mức đóng góp TOI cao nhất > 90%.
- KH có BANK_RELATION = 8 chiếm số lượng đứng thứ 2 (~2 triệu).

3.6 Mối quan hệ giữa CUS_TARGET_CDE và TOI



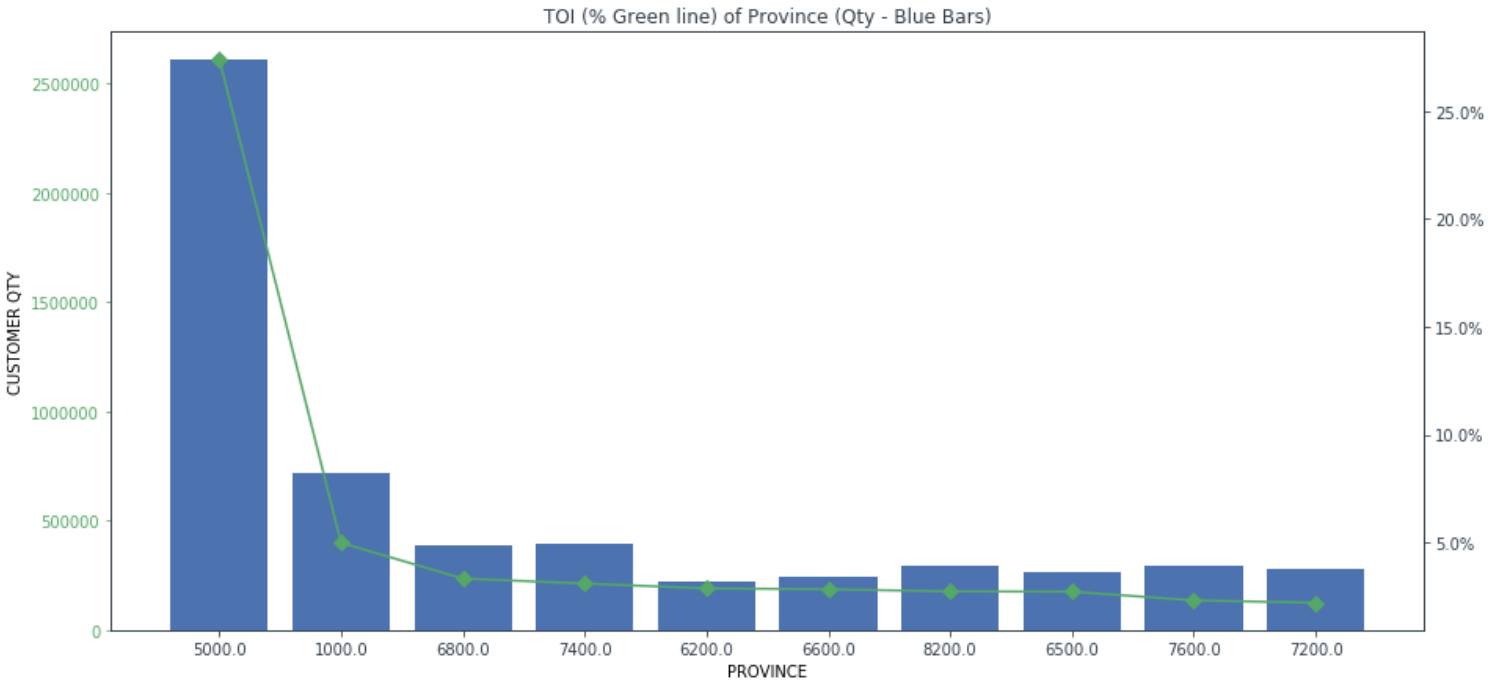
Hình: Biểu đồ biểu diễn quan hệ giữa CUS_TARGET_CDE và TOI (năm 2018)

*Ghi chú : Cột xanh đại diện cho số lượng khách hàng , đường màu xanh đại diện cho mức đóng góp TOI của KH theo CUS_TARGET_CDE (ngành nghề khách hàng).

*Nhận xét:

- Khách hàng có CUS_TARGET_CDE =2001 chiếm số lượng nhiều nhất (>7 triệu) và có đóng góp TOI cao nhất chiếm 65% tổng TOI của ngân hàng.
- Các khách hàng có CUS_TARGET_CDE = 2004, 1001 chiếm số lượng cao (> 1triệu) và có lượng TOi đóng góp chiếm tỷ lệ xấp xỉ trên dưới 10% tổng TOI.

3.7 Mối quan hệ giữa PROVINCE và TOI



Hình: Biểu đồ biểu diễn quan hệ giữa PROVINCE và TOI (năm 2018)

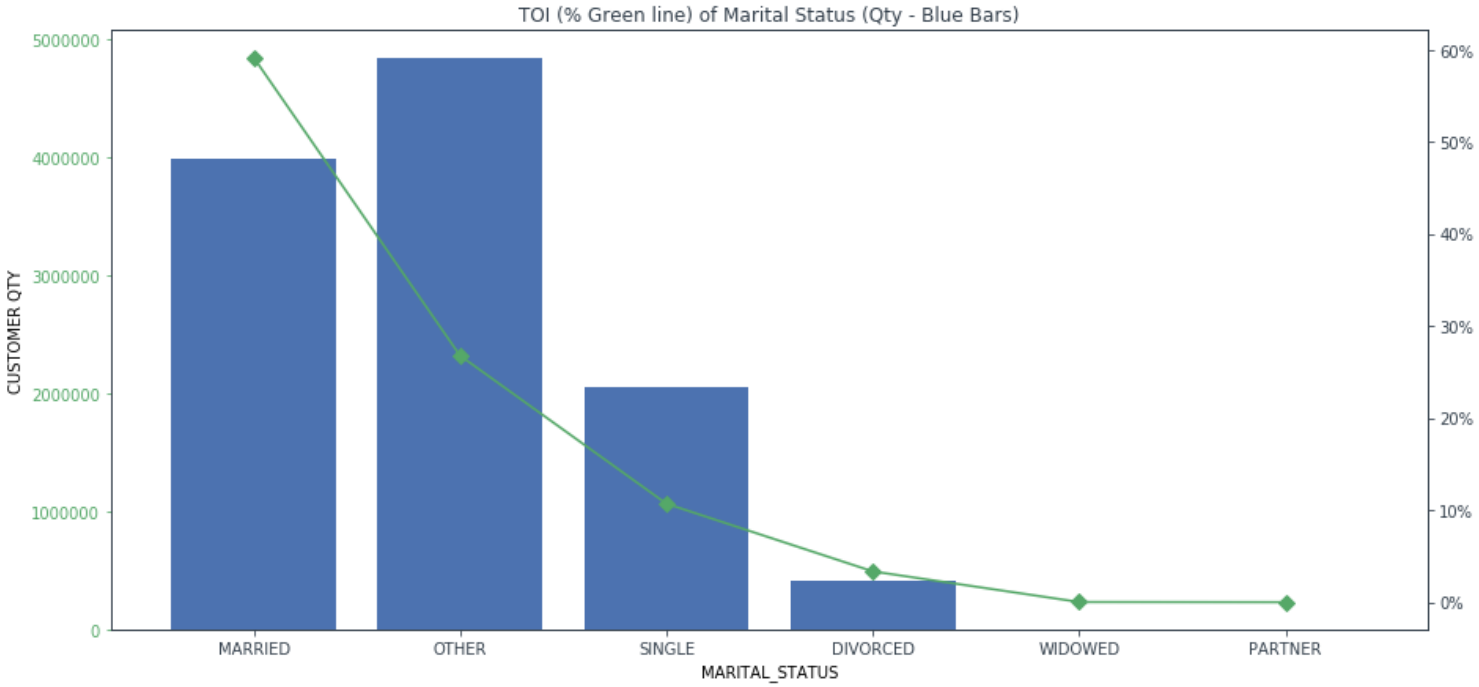
*Ghi chú :

Cột xanh đại diện cho số lượng khách hàng , đường màu xanh đại diện cho mức đóng góp TOI của KH theo PROVINCE.

*Nhận xét:

- Khách hàng có PROVINCE =5000 (tp.Hồ Chí Mình) chiếm số lượng nhiều nhất (>2.5 triệu) và có đóng góp TOI cao nhất chiếm trên 25% tổng TOI của ngân hàng.
- Các khách hàng có CUS_TARGET_CDE = 1000 (Hà Nội) chiếm số lượng thứ 2 (> 500.000) và có lượng TOi đóng góp chiếm tỷ lệ gần 10% tổng TOI.

3.8 Mối quan hệ giữa MARITAL_STATUS và TOI



Hình: Biểu đồ biểu diễn quan hệ giữa MARITAL_STATUS và TOI (năm 2018)

*Ghi chú :

Cột xanh đại diện cho số lượng khách hàng , đường màu xanh đại diện cho mức đóng góp TOI của KH theo MARITAL_STATUS (tình trạng hôn nhân).

*Nhận xét:

- Khách hàng có tình trạng hôn nhân Married chiếm số lượng cao (~4 triệu) và có đóng góp TOI cao nhất chiếm trên 50% tổng TOI của ngân hàng.
- Các khách hàng có tình hôn nhân là Partner và widowed chiếm số lượng thấp nhất (<1000) và có lượng TOi đóng góp chiếm tỷ lệ thấp nhất (~1%).

3.9 Phân tích các thuộc tính của sản phẩm tiền gửi

Hình: Ma trận tương quan các thuộc tính của sản phẩm tiền gửi (DEPOSIT)

*Nhận xét: trong 4 năm từ 2016-2019 ta thấy

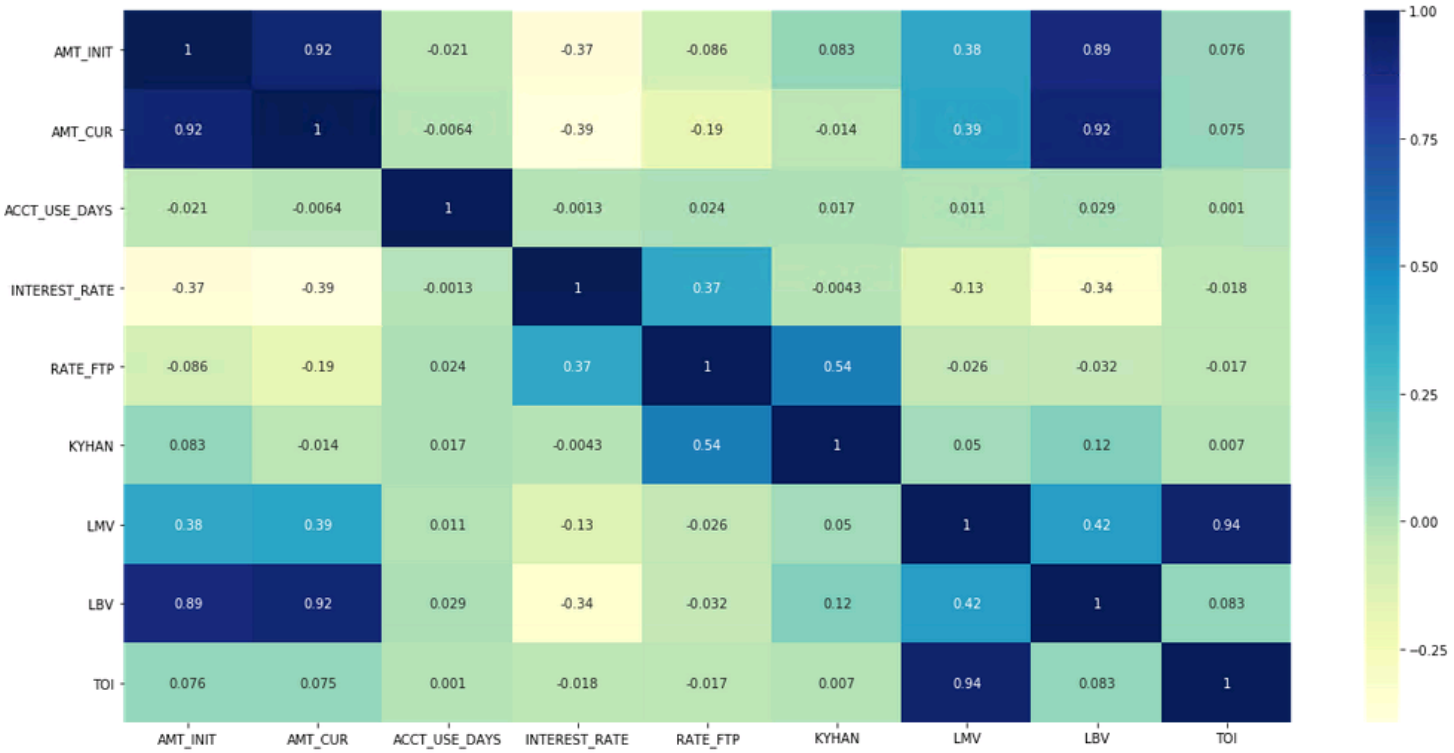
- Số dư tiền gửi hiện tại (AMT_CUR) tương quan thuận mạnh với TOI đóng góp (P ~ 1).

=> Kết hợp giữa kết quả ma trận tương quan trên và Domain Knowledge của nghiệp vụ. Ta xác định tập các thuộc tính sau sẽ được sử dụng để xây dựng model

X = ['AMT_CUR_LOG','ACCT_USE_DAYS','INTEREST_RATE','RATE_FTP']

y = ['TOI_LOG']

3.10 Phân tích các thuộc tính của sản phẩm tiền vay



Hình: Ma trận tương quan các thuộc tính của sản phẩm tiền vay (LOAN)

*Nhận xét:

- Lãi mua vốn (LMV) tương quan thuận mạnh với TOI đóng góp (P ~ 0.94).
- Dư nợ vay hiện tại (AMT_CUR) và dư nợ vay ban đầu (AMT_INT) có sự tương quan thuận mạnh với Lãi bán vốn lần lượt là 0.92 và 0.89.

=> Kết hợp giữa kết quả ma trận tương quan trên và Domain Knowledge của nghiệp vụ. Ta xác định tập các thuộc tính sau sẽ được sử dụng để xây dựng model

X = ['AMT_CUR_LOG',
'ACCT_USE_DAYS','RATE_FTP','INTEREST_RATE','KYHAN']
y = ['TOI_LOG']

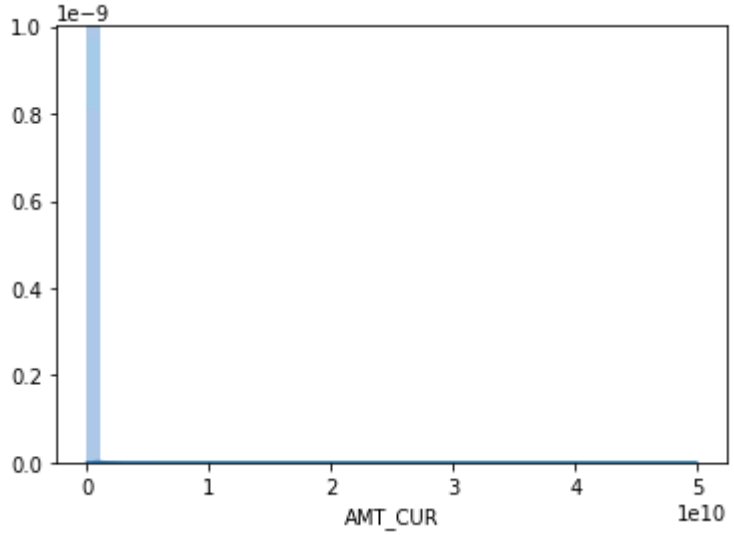
4 XÂY DỰNG MÔ HÌNH DỰ ĐOÁN TOI

4.1 Biến đổi dữ liệu (Data Transformation):

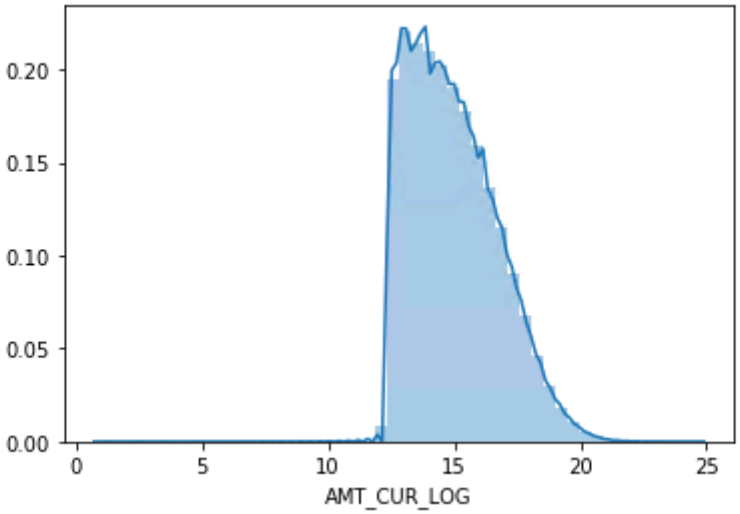
4.1.1 Log Normalization

- Các thuộc tính AMT CUR, AMT_INIT, TOI có phương sai lớn so với các thuộc tính còn lại và phân bố thuộc tính bị lệch phải (right long tail).

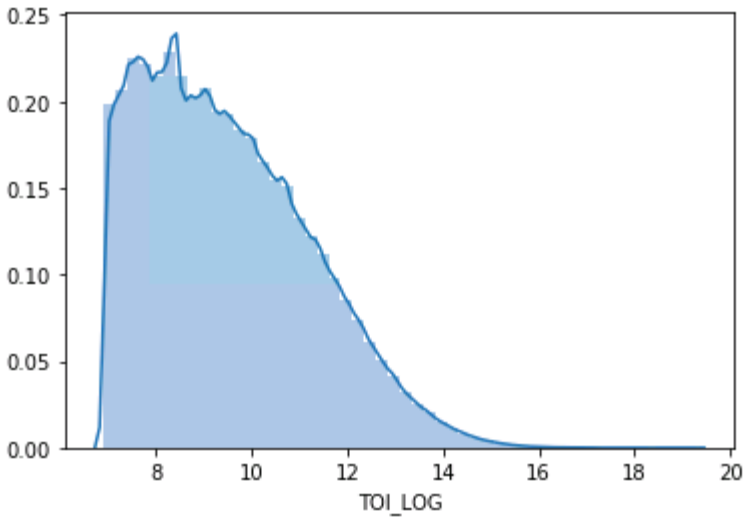
- Sử dụng Log normalization để biến đổi 2 thuộc tính AMT_CUR và TOI để giảm ảnh hưởng của các thuộc tính có giá trị ngoại lai đồn thời tăng tốc độ huấn luyện mô hình.



Hình: Đồ thị phân bố AMT_CUR trước khi thực hiện Log normalization



Hình: Đồ thị phân bố AMT_CUR sau khi thực hiện Log normalization



Hình: Phân bố TOI sau khi thực hiện Log normalization

4.1.2 Scaling Data

4.1.2.1 Sử dụng Robust Scaler của thư viện Sklearn

- Các thuộc tính cần scale:
  'AMT_CUR_LOG','ACCT_USE_DAYS','INTEREST_RATE','RATE_FTP'

4.1.2.2 Sử dụng One hot encoding

Out[52]:

	PRODUCT_-1--1003	PRODUCT_10011--1003	PRODUCT_10013--1003	PRODUCT_10015--1003	PRODUCT_10020--1003	PRODUCT_10032--1003	PRODUCT_10073--1003
0	0	0	0	0	0	0	1
1	1	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1
4	1	0	0	0	0	0	0

In [54]:
```
PRODUCT_CDE_df.columns
```

Out[54]:
```
Index(['PRODUCT_-1--1003', 'PRODUCT_10011--1003', 'PRODUCT_10013--1003',
       'PRODUCT_10015--1003', 'PRODUCT_10020--1003', 'PRODUCT_10032--1003',
       'PRODUCT_10073--1003', 'PRODUCT_10075--1003', 'PRODUCT_10079--1003',
       'PRODUCT_10100--1003', 'PRODUCT_11011--1003', 'PRODUCT_11015--1003',
       'PRODUCT_11026--1003', 'PRODUCT_11032--1003'],
      dtype='object')
```

4.2 Lựa chọn mô hình dự đoán (Model Selection)

4.2.1 So sánh độ chính xác của các mô hình dự đoán (Models Comparision)

1 ==========

2 DecisionTreeRegressor

```
Result-test
RMSE: 0.14184685233846345
Result-train
RMSE: 0.011218016273881366
The Training R^2 is:  99.9895317082189 %
The Testing R^2 is:  98.19982687104144 %
==================================================
==================================================
AdaBoostRegressor
Result-test
RMSE: 0.37631639644546533
Result-train
RMSE: 0.3655600077636232
The Training R^2 is:  88.88367991275814 %
The Testing R^2 is:  87.32986765194266 %
==================================================
==================================================
GradientBoostingRegressor
Result-test
RMSE: 0.08544478761836384
Result-train
RMSE: 0.09326856487910923
The Training R^2 is:  99.27637381056533 %
The Testing R^2 is:  99.3468002377361 %
==================================================
==================================================
ExtraTreesRegressor
Result-test
RMSE: 0.0769375064308784
Result-train
RMSE: 0.011217987328615708
The Training R^2 is:  99.98953176224042 %
The Testing R^2 is:  99.47039621304143 %
==================================================
==================================================
RandomForestRegressor
Result-test
RMSE: 0.1021348882183774
Result-train
RMSE: 0.05673443320571948
The Training R^2 is:  99.73224528407914 %
The Testing R^2 is:  99.0666960485192 %
```

```
==================================================
==================================================
Ridge
Result-test
RMSE: 0.27418009645615465
Result-train
RMSE: 0.3385056240201417
The Training R^2 is:  90.46818806806274 %
The Testing R^2 is:  93.27415710776957 %
==================================================
==================================================
Lasso
Result-test
RMSE: 1.0572387903506781
Result-train
RMSE: 1.0964230475386305
The Training R^2 is:  0.0 %
The Testing R^2 is:  -0.004846363934185582 %
==================================================
==================================================
LinearRegression
Result-test
RMSE: 0.27420188386716915
Result-train
RMSE: 0.3385044628571185
The Training R^2 is:  90.46825346117761 %
The Testing R^2 is:  93.27308814231458 %
==================================================
==================================================
ElasticNet
Result-test
RMSE: 0.8899179772616381
Result-train
RMSE: 0.9278200679894724
The Training R^2 is:  28.390406040912975 %
The Testing R^2 is:  29.14429413352363 %
==================================================
==================================================
Lars
Result-test
RMSE: 0.2742018838671695
Result-train
```

```
RMSE: 0.33850446285711855
The Training R^2 is:  90.46825346117761 %
The Testing R^2 is:  93.27308814231456 %
==================================================
==================================================
SVR
Result-test
RMSE: 0.18024392601608072
Result-train
RMSE: 0.14079190238306713
The Training R^2 is:  98.35108206380812 %
The Testing R^2 is:  97.09332720609768 %
==================================================
==================================================
KNeighborsRegressor
Result-test
RMSE: 0.2283842155701889
Result-train
RMSE: 0.19581229774702769
The Training R^2 is:  96.81049259299583 %
The Testing R^2 is:  95.33333056332881 %
==================================================
==================================================
XGBRegressor
Result-test
RMSE: 0.06389594629365607
Result-train
RMSE: 0.016994098812199772
The Training R^2 is:  99.97597629612129 %
The Testing R^2 is:  99.63472371139727 %
==================================================
```

	regressors	MSE
0	XGBRegressor	0.063896
0	ExtraTreesRegressor	0.076938
0	GradientBoostingRegressor	0.085445
0	RandomForestRegressor	0.102135
0	DecisionTreeRegressor	0.141847
0	SVR	0.180244
0	KNeighborsRegressor	0.228384
0	Ridge	0.274180
0	LinearRegression	0.274202
0	Lars	0.274202
0	AdaBoostRegressor	0.376316
0	ElasticNet	0.889918
0	Lasso	1.057239

Models	R_squared	
	Train	Test
DecisionTreeRegressor	99.9895	98.1998
AdaBoostRegressor	88.8836	87.3298
GradientBoostingRegressor	99.2763	99.3468
ExtraTreesRegressor	99.9895	99.4703
RandomForestRegressor	99.7322	99.0666
Ridge	90.4681	93.2741
Lasso	0.00%	-0.0048
LinearRegression	90.4682	93.273
ElasticNet	28.3904	29.1442
Lars	90.4682	93.273
SVR	98.351	97.0933
KNeighborsRegressor	96.8104	95.3333
XGBRegressor	99.9759	99.6347

*Nhận xét: Qua kết quả trên ta thấy mô hình huấn luyện dùng giải thuật XGBoost Regressor cho độ chính xác trên tập train và test cao nhất lần lượt là 99.976% và 99.635%. Đồng thời cho độ lỗi LOG thấp nhất là 0.063896.
=> Ta lựa chọn giải thuật XGBoost Regressor này để sử dụng cho mô hình dự đoán TOI.

4.2.2 Cải tiến các tham số (Hyper params tunning)

- Sử dụng GridSearchCV để tìm kiếm bộ tham số tối ưu cho mô hình XGboost.

In [47]:
```
from sklearn.model_selection import train_test_split,ShuffleSplit
def XGBRegressor_cross(param_grid, n_jobs):
    estimator = XGBRegressor()
    cv = ShuffleSplit(n_splits = 10, test_size = 0.2, random_state = 42)
    rgmodel = GridSearchCV(estimator=estimator, cv=cv, param_grid=param_grid, n_jobs=n_jobs)
    rgmodel.fit(X_train, y_train)
    print("Best Estimator learned through GridSearch")
    print (rgmodel.best_estimator_)
    return cv, rgmodel.best_estimator_
```

In [48]:
```
param_grid={'n_estimators':[100,500,1000],
            'learning_rate': [0.1,0.05, 0.01],
            'max_depth':[4,5,6]
#             'min_samples_leaf':[3,5,9,17],
#             'max_features':[1.0,0.3,0.1]
            }
n_jobs=4
cv,best_est=XGBRegressor_cross(param_grid, n_jobs)
```

```
Best Estimator learned through GridSearch
XGBRegressor(base_score=0.5, booster=None, colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints=None,
             learning_rate=0.05, max_delta_step=0, max_depth=4,
             min_child_weight=1, missing=nan, monotone_constraints=None,
             n_estimators=1000, n_jobs=0, num_parallel_tree=1,
             objective='reg:squarederror', random_state=0, reg_alpha=0,
             reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method=None,
             validate_parameters=False, verbosity=None)
```

*Nhận xét: Qua kết quả trên ta thấy GridSearchCV cho kết quả bộ tham số tốt nhất là:
N_estimators=1000
Learning_rate = 0.05
Max_depth = 4

4.2.3 Đánh giá mô hình (Model Evaluation)

4.2.3.1 Đối với mô hình huấn luyện sản phẩm tiền vay (LOAN)

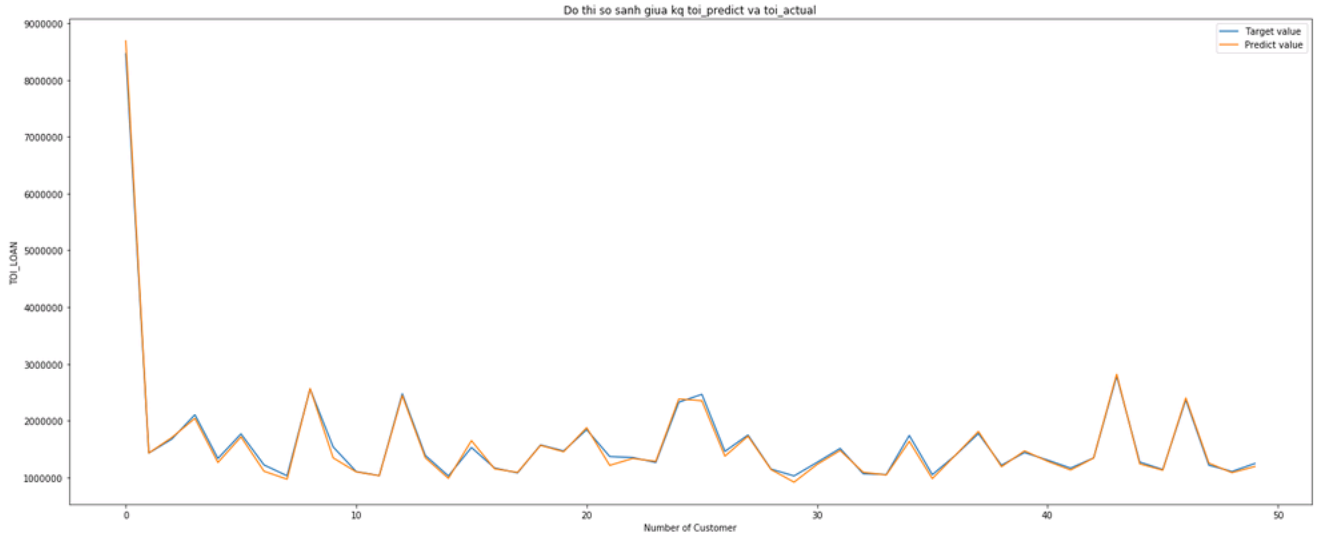
- Áp dụng bộ tham số trên để huấn luyện lại mô hình dự đoán, ta có kết quả độ chính xác như sau:
- Đối với mô hình huấn luyện sản phẩm LOAN:

```
MSE         = 0.0098
R-squared   = 0.9895
```

- Tạo bảng so sánh giữa giá tri TOI dự đoán và TOI thực tế sau:
    - TOI_LOAN_PREDICT : Giá trị TOI dự đoán từ mô hình
    - TOI_LOAN_ACTUAL: Giá trị TOI thực tế
    - TOI_STD: Độ chênh lệch TOI giữa 2 giá trị trên

	TOI_LOAN_PREDICT	TOI_LOAN_ACTUAL	TOI_STD
100036	1.427867e+07	1.442733e+07	1.486609e+05
34782	1.385315e+07	1.426849e+07	4.153441e+05
184166	1.401826e+07	1.394484e+07	7.342515e+04
138299	1.406261e+07	1.249399e+07	1.568619e+06
32364	1.138065e+07	1.167583e+07	2.951809e+05
61393	1.071800e+07	1.149467e+07	7.766664e+05
178518	9.200840e+06	9.653146e+06	4.523062e+05
72930	9.080561e+06	9.633633e+06	5.530726e+05
32215	8.426948e+06	8.944207e+06	5.172592e+05
18	8.690701e+06	8.464361e+06	2.263396e+05
66757	8.132430e+06	8.154702e+06	2.227201e+04
61311	8.247509e+06	8.150000e+06	9.750852e+04
71119	7.040314e+06	7.964489e+06	9.241753e+05
104293	7.895342e+06	7.945914e+06	5.057219e+04
190953	7.377320e+06	7.846998e+06	4.696783e+05
151422	7.052779e+06	7.783357e+06	7.305780e+05

Hình: Bảng kết quả thống kê độ chênh lệch giữa giá trị TOI dự đoán và thực tế



Hình: Đồ thị thể hiện giá trị TOI tiền vay dự đoán và thực tế của 50 khách hàng có TOI tiền vay thực tế từ 1 triệu VNĐ.

* Nhận xét: Từ kết quả huấn luyện trên ta thấy

- Mô hình huấn luyện có khả năng giải thích 98.95% dữ liệu input.
- Mô hình cho độ lỗi LOG thấp khoảng 0.0098.
- Dùng model dự đoán thử kết quả cho dữ liệu từ tháng 01 đến tháng 05-2020:

```
RangeIndex: 237302 entries, 0 to 237301
Data columns (total 16 columns):
CUSTOMER_CDE     237302 non-null int64
PRODUCT_CDE      237302 non-null object
LD_ID            237302 non-null object
AMT_INIT         237302 non-null int64
AMT_CUR          237302 non-null int64
INTEREST_RATE    237302 non-null float64
RATE_FTP         237302 non-null float64
KYHAN            237302 non-null int64
LOAITRAGOP       237302 non-null object
MUCDICHVAY       237302 non-null int64
LMV              237302 non-null float64
LBV              237302 non-null float64
ACCT_USE_DAYS    237302 non-null int64
TOI              237302 non-null float64
PROCESS_MONTH    237302 non-null int64
PROCESS_YEAR     237302 non-null int64
dtypes: float64(5), int64(8), object(3)
memory usage: 29.0+ MB
```

Hình: Format của dữ liệu test

- Đánh giá model dự đoán cho dữ liệu từ tháng 01 đến tháng 05-2020:

```
Tong sai lech            = 49976.2625
Do phu hop cua mo hinh   = 99.0813 %
```

)4]:

	TOI_LOAN_PREDICT	TOI_LOAN_ACTUAL	TOI_STD
44224	12753018.0	1.412329e+07	1.370270e+06
92473	11120410.0	1.189093e+07	7.705207e+05
92113	11104429.0	1.161282e+07	5.083885e+05
92245	11104429.0	1.131031e+07	2.058795e+05
92472	10457743.0	1.110537e+07	6.476280e+05
92246	10194624.0	1.102241e+07	8.277889e+05
59264	11042250.0	1.094937e+07	9.288014e+04
59186	10486686.0	1.024296e+07	2.437271e+05
59263	4284778.0	1.009370e+07	5.808921e+06
59718	11042250.0	9.864438e+06	1.177812e+06
28822	8231345.0	9.061612e+06	8.302665e+05
60981	6142650.5	9.060493e+06	2.917843e+06
4136	8018283.5	8.975342e+06	9.570590e+05
3877	7706836.5	8.772717e+06	1.065880e+06
3532	7629248.0	8.616438e+06	9.871904e+05
26524	8988937.0	8.611638e+06	3.772990e+05
27111	8886721.0	8.470723e+06	4.159975e+05
27109	9053979.0	8.411111e+06	6.428681e+05

Hình: Bảng kết quả thống kê độ chênh lệch giữa giá trị TOI dự đoán và thực tế

*Nhận xét kết quả test:

- Mô hình cho kết quả dự đoán tốt với việc giải thích được 99% kết quả test và cho độ lỗi (MSE) thấp khoảng 49976.

4.2.3.2 Đối với mô hình huấn luyện sản phẩm tiền gửi (DEPOSIT)

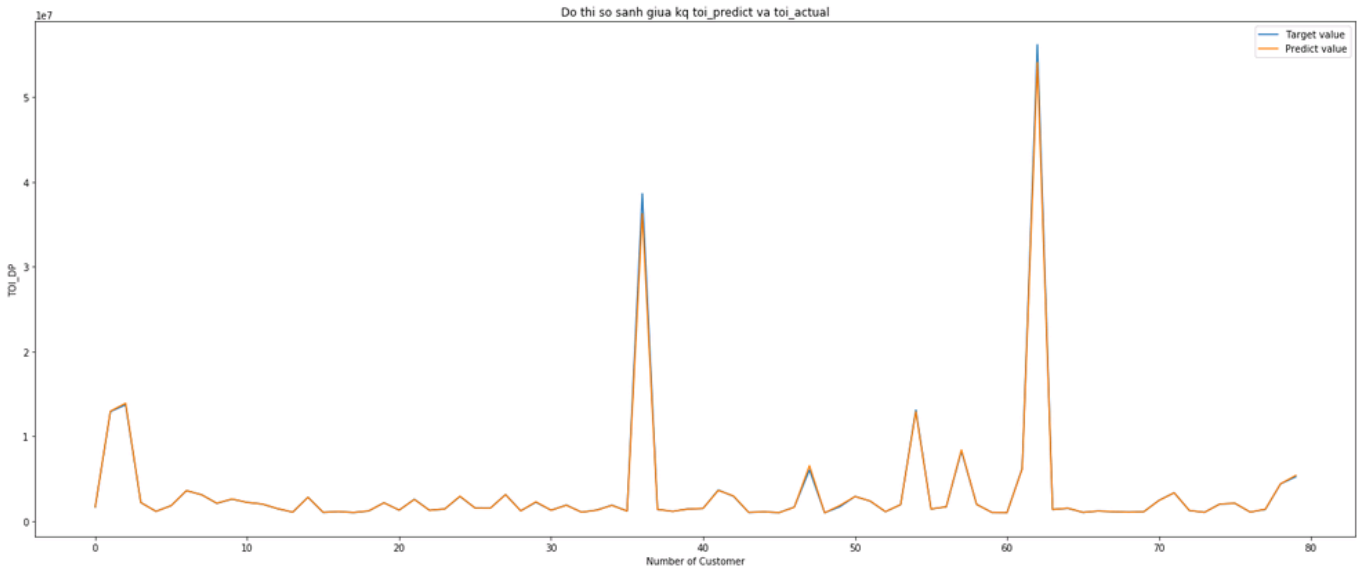
- Áp dụng bộ tham số trên để huấn luyện lại mô hình dự đoán, ta có kết quả độ chính xác như sau:
- Kết quả đánh giá độ chính xác mô hình huấn luyện sản phẩm DEPOSIT:

```
RMSE        = 0.0634
R-squared   = 0.9995
```

Out[83]:

	TOI_DP_PREDICT	TOI_DP_ACTUAL	TOI_STD
55061	1.213678e+08	1.244869e+08	3.119081e+06
2497	3.627202e+07	3.863261e+07	2.360595e+06
4277	5.408887e+07	5.619868e+07	2.109810e+06
57015	3.026729e+07	2.849106e+07	1.776238e+06
70343	2.674770e+07	2.509556e+07	1.652149e+06
25118	2.916234e+07	2.761614e+07	1.546199e+06
5463	2.367127e+07	2.220475e+07	1.466519e+06
65344	6.138982e+07	6.007897e+07	1.310842e+06
10881	6.192313e+07	6.069496e+07	1.228168e+06
41864	1.522878e+06	2.728965e+06	1.206086e+06
24554	4.781454e+07	4.894371e+07	1.129171e+06
18597	2.951235e+07	3.052452e+07	1.012174e+06
8897	5.922772e+07	5.827872e+07	9.489938e+05
66757	2.115732e+07	2.028714e+07	8.701810e+05
78486	1.095754e+08	1.087074e+08	8.679683e+05
46278	2.017383e+02	8.516041e+05	8.514024e+05
71457	4.291604e+06	3.465868e+06	8.257351e+05
55598	2.423872e+07	2.504036e+07	8.016402e+05
61600	3.859529e+06	3.058819e+06	8.007094e+05
56692	6.008206e+04	8.561432e+05	7.960612e+05

Hình: Bảng kết quả thống kê độ chênh lệch giữa giá trị TOI tiền gửi dự đoán và thực tế



Hình: Đồ thị thể hiện giá trị TOI tiền gửi dự đoán và thực tế của 50 khách hàng có TOI tiền gửi thực tế từ 1 triệu VNĐ

* Nhận xét: Từ kết quả huấn luyện trên ta thấy

- Mô hình huấn luyện có khả năng giải thích 99.95% dữ liệu input.
- Mô hình cho độ lỗi LOG thấp khoảng 0.0634.

- Dùng model dự đoán thử kết quả cho dữ liệu từ tháng 01 đến tháng 05-2020:

Out[93]:

	PRODUCT_CDE	count
11	11032--1003	33685
9	11011--1003	29440
3	10015--1003	10715
0	-1--1003	10363
6	10073--1003	4955
8	10079--1003	1708
2	10013--1003	1420
7	10075--1003	309
4	10020--1003	307
5	10032--1003	10
10	11015--1003	7
1	10011--1003	6

Hình: Bảng thống kê dữ liệu test input theo các sản phẩm tiền gửi

- Dữ liệu test gồm 92925 mẫu , kết quả đánh giá mô hình dự đoán trên dữ liệu test tiền gửi như sau:

```
RMSE        = 41635.2377
R-squared   = 0.9958
```

:

	TOI_DP_PREDICT	TOI_DP_ACTUAL	TOI_STD
50915	82344272.0	8.827117e+07	5.926896e+06
23228	76679848.0	8.259035e+07	5.910504e+06
23227	69252800.0	6.549140e+07	3.761399e+06
12803	17759996.0	1.458632e+07	3.173678e+06
4750	17949616.0	1.519101e+07	2.758601e+06
45838	31302384.0	2.899035e+07	2.312039e+06
13021	18584258.0	1.636399e+07	2.220270e+06
46218	62184584.0	6.034473e+07	1.839849e+06
12802	17379188.0	1.556531e+07	1.813881e+06
24572	17954752.0	1.625010e+07	1.704655e+06
5300	15465982.0	1.387613e+07	1.589854e+06
6654	12986554.0	1.144209e+07	1.544462e+06
6092	10138854.0	8.915863e+06	1.222991e+06
46874	8717479.0	7.579718e+06	1.137761e+06
85731	6139857.0	5.072182e+06	1.067675e+06
24973	7539983.5	6.519106e+06	1.020878e+06
8679	13286764.0	1.235787e+07	9.288920e+05
3328	7581552.0	6.673561e+06	9.079909e+05
46219	10143690.0	9.235781e+06	9.079091e+05
25616	8526217.0	7.710327e+06	8.158895e+05

Hình: Bảng kết quả thống kê độ chênh lệch giữa giá trị TOI tiền gửi dự đoán và thực tế trên tập test

*Nhận xét kết quả test:

- Mô hình cho kết quả dự đoán tốt với việc giải thích được 99.58% kết quả test và cho độ lỗi (MSE) thấp khoảng 41635.
- Như vậy mô hình dự đoán TOI tiền gửi dùng giải thuật XGboost là mô hình ứng viên phù hợp.

5 TRIỂN KHAI MÔ HÌNH

- Sử dụng Streamlit tạo UI Dashboard có khả năng:
  - Đọc file input danh sách các khách hàng cá nhân và sản phẩm sử dụng.
  - Tiến hành dự đoán TOI của khác hàng cho 6 tháng, 1 năm tới.
  - Kết quả dự đoán theo tháng hoặc theo từng sản phẩm
  - Xuất kết quả dự đoán ra file.

MÔ HÌNH DỰ ĐOÁN TOI TƯƠNG LAI CHO KHÁCH HÀNG CÁ NHÂN TRONG LĨNH VỰC NGÂN HÀNG

Chọn File

inference_toiloan_group01.csv

Tổng sai lệch (RMSE) = 49976.2625

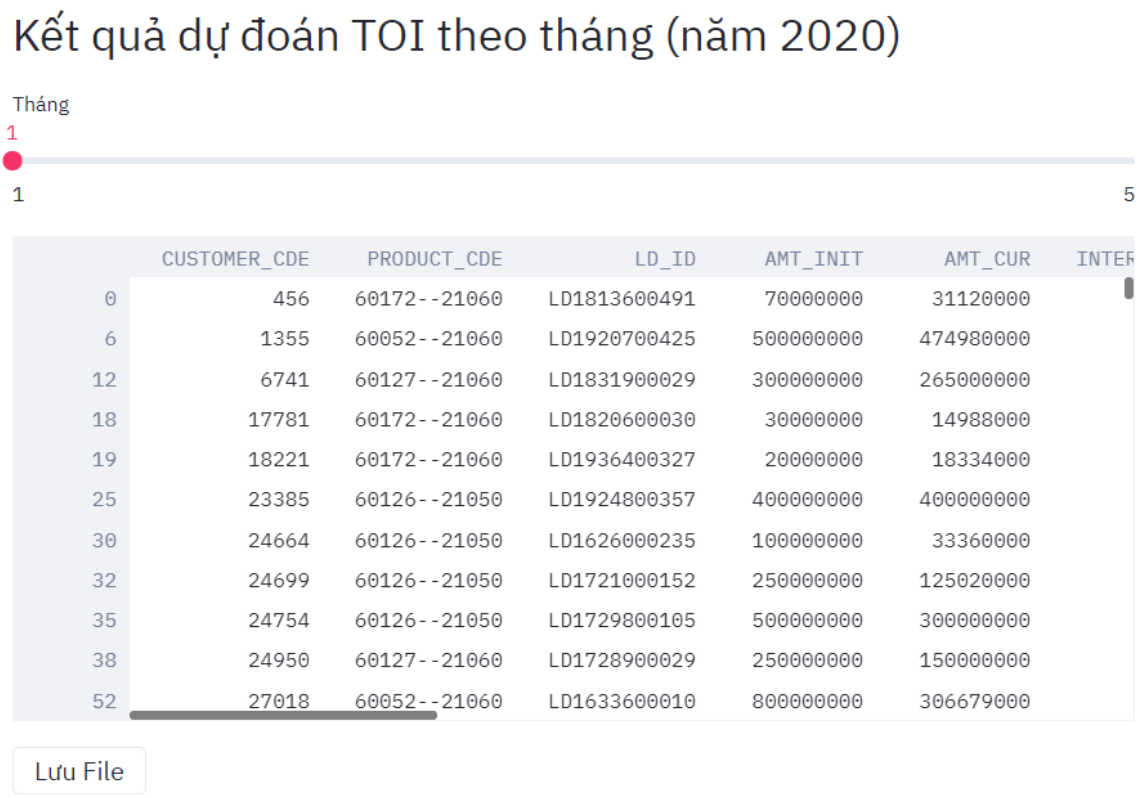
Độ phù hợp của mô hình (R_SQUARED) = 99.0813 %

Kết quả dự đoán TOI theo sản phẩm

Sản phẩm

60172--21060

	OCESS_MONTH	PROCESS_YEAR	TOI_PREDICT	TOI_6M_NEXT	TOI_ONEYEAR_NEXT
0	1	2020	295,507.5938	1,715,850.5444	3,479,363.6038
1	5	2020	221,474.4531	1,285,980.6956	2,607,683.0771
2	3	2020	253,382.2813	1,471,251.9556	2,983,372.0212
3	3	2020	476,786.1563	2,768,435.7460	5,613,772.4849
4	2	2020	468,621.8438	2,908,687.3060	5,898,171.4817
7	3	2020	179,733.2031	1,043,612.1472	2,116,213.5207
8	3	2020	185,322.2188	1,076,064.4960	2,182,019.6724
9	2	2020	186,171.6250	1,155,548.0172	2,343,194.5905
10	3	2020	357,363.0938	2,075,011.5121	4,207,662.2329
11	4	2020	326,116.2500	1,956,697.5000	3,967,747.7083
18	1	2020	137,407.6875	797,851.0887	1,617,864.7077



Hình: Giao diện Dashboard kết quả dự đoán TOI sử dụng Streamlit

6 KẾT LUẬN (CONCLUSIONS)

- Mô hình XGboost có khả năng giải thích 98.95% cho dữ liệu train sản phẩm tiền vay (Loan).
- Mô hình XGboost có khả năng giải thích 99.95% cho dữ liệu train sản phẩm tiền gửi (Deposit).
- Độ sai lệch của mô hình dự đoán XGboost là 49976 cho dữ liệu test input sản phẩm tiền vay và 41116 cho dữ liệu input sản phẩm tiền gửi.
- Các khách hàng trong độ tuổi trung niên từ 41-50 tuổi có đóng góp TOI cao nhất cho NH. Khách hàng nữ trong nhóm này có xu hướng sử dụng tiền gửi nhiều hơn so với nam giới, ngược lại nam giới có xu hướng sử dụng tiền vay nhiều hơn.

7 ĐỀ XUẤT CẢI TIẾN (FUTURE DIRECTIONS)

- Cần triển khai kết hợp với các bài toán dự đoán lãi suất (dùng Time series), dự đoán hành vi mua hàng, dự đoán hành vi rời bỏ sản phẩm của khách hàng để hoàn thiện mô hình dự đoán TOI.

8 TÀI LIỆU THAM KHẢO

[1] Max Kuhn, "Applied Predictive Modeling". Springer 1st ed. 2013, Corr. 2nd printing 2018 Edition.
[2] Thomas W. Miller, "Marketing Data Science: Modeling Techniques in Predictive Analytics with R and Python". ISBN-13: 978-0133886559 Pearson FT Press; 1 edition (May 22, 2015).
[3] C. E. Shannon, (1948) 'A mathematical theory of communication' Bell System Technical Journal, 27:379-423,623-656.
[4] V.Vapnik. The Nature of Statistical Learning Theory. Springer, NewYork – 1995.
[5] Xgboost, https://ongxuanhong.wordpress.com/2017/12/21/xgboost-thuat-toan-gianh-chien-thang-tai-nhieu-cuoc-thi-kaggle/,30.06.2020.