**Tỷ lệ khách hàng rời bỏ: Nghiên cứu các nhân tố**

**quyết định tỷ lệ khách hàng rút tiền tại Ngân hàng sử dụng các thuật toán Machine Learning**.

# **TÓM TẮT**

Tỷ lệ khách hàng rời bỏ là một trong những vấn đề hàng đầu được các doanh nghiệp quan tâm, đặc biệt là ngành Ngân hàng. Tỷ lệ này ảnh ảnh hưởng trực tiếp đến lợi nhuận của doanh nghiệp do đó việc xác định các nhân tố ảnh hưởng đến tỷ lệ khách hàng rời bỏ là rất cần thiết và tối quan trọng. Trên cơ sở đó, ta có thể ước tính xác suất rời bỏ của từng khách hàng nhằm đưa ra các phương án tối ưu hóa chi phí – lợi nhuận của Ngân hàng.

Ngân hàng có rất nhiều mảng kinh doanh trong đó có ba mảng mang lại lợi nhuận chính đó là: cho vay vốn, huy động vốn và dịch vụ. Trong bài này, tác giả tập trung phân tích các yếu tố tác động đến hành vi rút tiền của khách hàng tiền gửi. Số liệu của tác giả sử dụng trong bài được cung cấp bởi một Ngân hàng tại thành phố Hồ Chí Minh.

Số liệu đượ cung cấp là dữ liệu thô chưa qua các bước tiền xử lý nên tác giả sử dụng PySpark để biến đổi, làm sạch, và sàng lọc các thuộc tính cho việc phân tích. Sau khi có bộ dữ liệu các thuộc tính, tác giả chọn năm 2018 làm năm cơ sở để gán nhãn khách hàng gửi – rút tiền. Bên cạnh đó, tác giả đề xuất việc gán nhãn khách hàng theo hai phương án: (1) So sánh và đối chiếu bộ dữ liệu trong năm 2018 với bộ dữ liệu khách hàng vào 31/12/2019; (2) Sử dụng thuật toán bán giám sát Propagation Labeling, được huấn luyện theo một bộ mẫu 10% các khách hàng đã được tác giả gán nhãn tại phương án (1). Qua đó, thuật toán bán giám sát cho kết quả gán nhãn sát đến 89% so với bộ dữ liệu gốc.

Các thuật toán Machine Learning được tác giả sử dụng và phân tích đó là: *Logistics Regression, Decision Tree, Random Forest và Extreme Gradient Boosting (XGBoost)*. Trong đó, thuật toán *XGBoost* cho hiệu suất với recall-score đạt 94%. Bộ dữ liệu được chia thành hai bộ riêng để huấn luyện (train) và đánh giá (test) mô hình. Tác giả sử dụng f1-score và recall-score để đo lường hiệu quả của mô hình, accuracy-score để kiểm tra mức độ quá khớp (over-fitting).

Ngoài ra, để kiểm tra chất lượng và tính nhất quán của mô hình tác giả sử dụng phương pháp xác thực chéo (cross validating), qua đó chia dữ liệu thành n lần nhằm tính trung bình các kết quả. Cuối cùng, tác giả ứng dụng phương pháp LIME (local interpretable model-agnostic explanations) nhằm cố gắng giải thích cách thức mô hình hoạt động.

**Từ khóa:** Tỷ lệ khách hàng rời bỏ, rút tiền, ngân hàng, tiền gửi, machine learning.

# **DANH MỤC HÌNH - BẢNG BIỂU – BIỂU ĐỒ**

Hình 1.1. Các bước thực hiện mô hình Data Science trong bài nghiên cứu.

Hình 2.1. Cấu trúc dữ liệu gốc.

Hình 2.2 Diễn giải cục bộ trên một quan sát ví dụ 1.

Hình 2.3 Diễn giải cục bộ trên một quan sát ví dụ 2.

Bảng 1.1. Các các giả và các thuật toán sử dụng.

Bảng 2.1 Các thuộc tính được lựa chọn từ dữ liệu Statements.

Bảng 2.3 Giải thích ý nghĩa các cột thuộc tính.

Bảng 2.4 p-value của các biến liên tục với lable.

Biểu đồ 2.1. Tỷ trọng kỳ hạn gửi tiền.

Biểu đồ 2.2 Tỷ trọng các loại tiền gửi.

Biểu đồ 2.3 Tỷ trọng khách hàng tại các địa điểm gửi tiền.

Biểu đồ 2.4. Đồ thị phân phối của max\_VND, avg\_VND với các cách biến đổi biến liên tục khác nhau.

Biểu đồ 2.5 Trung bình tiền gửi – giới tính, lãi suất – giới tính với nhãn.

Biểu đồ 2.6 Thước đo khoảng cách Cook với biến max\_VND.

Biểu đồ 2.7 Trung bình các hệ số tương quan của LIME trong tập đại diện

**MỤC LỤC**

[**TÓM TẮT** ii](#_Toc56932397)

[**DANH MỤC HÌNH - BẢNG BIỂU – BIỂU ĐỒ** iv](#_Toc56932398)

[**PHẦN 1. GIỚI THIỆU** 1](#_Toc56932399)

[**1.1 Cơ sở đề tài** 1](#_Toc56932400)

[**1.2 Các nghiên cứu liên quan** 2](#_Toc56932401)

[**PHẦN 2. NGHIÊN CỨU TÌNH HUỐNG** 4](#_Toc56932402)

[**2.1 Mô tả dữ liệu** 4](#_Toc56932403)

[**2.1.1 Nguồn gốc dữ liệu** 4](#_Toc56932404)

[**2.1.1 Sao kê hằng ngày (Statements):** 5](#_Toc56932405)

[**2.1.2 Thông tin khách hàng (INFO):** 7](#_Toc56932406)

[**2.1.3 Gán nhãn dữ liệu** 7](#_Toc56932407)

[**2.2 Explore Data Analysis:** 8](#_Toc56932408)

[**2.2.1 Phân tích đơn biến** 9](#_Toc56932409)

[**2.2.2 Phân tích hai biến** 11](#_Toc56932410)

[**2.2.3 Các ngoại lệ (outliers)** 12](#_Toc56932411)

[**2.2.4 Xử lý mất cân bằng dữ liệu** 14](#_Toc56932412)

[**2.3 Thực hiện mô hình.** 14](#_Toc56932413)

[**2.3.1 Mô hình dự đoán nhãn gán bằng Cách 1** 14](#_Toc56932414)

[**2.3.2 Mô hình dự đoán nhãn gán bằng Cách 2** 15](#_Toc56932415)

[**2.3.3 Diễn giải mô hình bằng LIME** 15](#_Toc56932416)

[**PHẦN 3: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 17](#_Toc56932417)

[**3.1 Kết luận mô hình** 17](#_Toc56932418)

[**3.2 Hướng phát triển** 17](#_Toc56932419)

[**3.2.1 Về độ chính xác của mô hình** 17](#_Toc56932420)

[**3.2.2 Về mở rộng khả năng phát triển của mô hình** 17](#_Toc56932421)

[PHỤ LỤC 1. TÊN CỘT VÀ GIẢI THÍCH STATEMENT VÀ INFO 19](#_Toc56932422)

[PHỤ LỤC 2. GIẢI THÍCH Ý NGHĨA CỦA MAGD 21](#_Toc56932423)

[PHỤ LỤC 3. KIỂM ĐỊNH GIẢ THUYẾT CHI BÌNH PHƯƠNG GIỮA LABEL VÀ CÁC BIẾN PHÂN LOẠI 22](#_Toc56932424)

[PHỤ LỤC 4: CODE PYTHON GIẢI NÉN VÀ ĐƯA FILE VÀO ĐƯỜNG DẪN ĐƯỢC CHỈ ĐỊNH 23](#_Toc56932425)

[**TÀI LIỆU THAM KHẢO** 25](#_Toc56932426)

# **PHẦN 1. GIỚI THIỆU**

## **1.1 Cơ sở đề tài**

Hiện nay, có rất nhiều nghiên cứu về lý thuyết về hành vi chuyển đổi và các nhân tố ảnh hưởng đến quyết định rời bỏ của khách hàng. Mặc dù có rất nhiều các yếu tố ảnh hưởng, nhiều nghiên cứu đã đi đến một kết luận rằng sự không hài lòng về sản phẩm dịch vụ là nhân tố then chốt khiến khách hàng bỏ đi (*Bansal và Taylor 1999; Bansal, Taylor, và St. James 2005; Gustafsson, Johnson, và Roos 2005*). Theo đó, khi sử dụng sản phẩm dịch vụ, khách hàng thường sẽ đánh giá về chất lượng, sau đó so sánh với mức độ kỳ vọng trước đó và lựa chọn sẽ tiếp tục hay ngưng sử dụng, dựa trên các cảm quan và cảm tính của khách hàng (*Dechant A, Spann và Becker, 2019*).

Các ngành thường quan tâm đến tỷ lệ khách hàng rời bỏ thông thường là những ngành dịch vụ và có sự cạnh tranh cao như viễn thông, điện tử (Kriti, 2019; Ahmad et al, 2019)… Nhằm tồn tại trong môi trường cạnh tranh, nhiều doanh nghiệp lớn đã đưa ra nhiều chiến lược khác nhau. Ba chiến lược chính để tăng doanh thu (*Wei CP, Chiu IT, 2002*) là: (1) thu hút khách hàng mới, (2) bán chéo sản phẩm, bán hàng gia tăng (Recommender System) (3) gia tăng vòng đời sử dụng sản phẩm của một khách hàng. Trong đó, chiến lược thứ ba mang lại tỷ suất lợi nhuận cao hơn cả (*Ahmad et al, 2019*), vì chi phí thu hút khách hàng mới ở ngành này rất cao.

Ngân hàng vốn là một ngành đặc thù, là một đầu mối trung gian để điều hòa vốn của nền kinh tế. Bên cạnh đó, ngân hàng cũng là một ngành kinh doanh thuần túy, có các loại sản phẩm dịch vụ, tiện ích đa dạng để đáp ứng cho nhiều phân khúc khách hàng, cũng như thu hút khách hàng mới. Tuy nhiên không phải vì thế mà ngân hàng thiếu sự cạnh tranh. Ngược lại, đây một trong những ngành có nhiều sự cạnh tranh khốc liệt (*TS. Nguyễn Đức Trường, ThS. Hà Tú Anh, ThS. Nguyễn Thị Thanh Bình 2019*), đặc biệt ở các thành phố lớn. Chi phí của khách hàng huy động vốn tại một Ngân hàng là sẽ gắn với vòng đời sản phẩm mà khách hàng đó sử dụng, cụ thể là lãi suất loại tiền gửi của khách hàng. Do đó, các ngân hàng thường quan tâm đến xác suất khách hàng đó rút tiền để đưa vào bài toán quy hoạch tối ưu hóa chi phí, lợi nhuận. Trong bài nghiên cứu này, tác giả chỉ tập trung phân tích các khách hàng cá nhân, vì hành vi của khách hàng cá nhân dễ dự đoán hơn hành vi của một khách hàng tổ chức.

## **1.2 Các nghiên cứu liên quan**

Có nhiều nghiên cứu đã khai thác các thuật toán khác nhau để phân tích và dự đoán khách hàng rời bỏ trong hệ thống tài chính ngân hàng. Logistics regression được vận dụng trong nghiên cứu của *Buckinx và đồng sự* (2005) vì: (1) có thể diễn giải xác suất có điều kiện, (2) dễ dàng sử dụng, có thể chạy ra kết quả nhanh và đáng tin cậy.

*Nie và đồng sự* (2011) đã đưa thuật toán Logistic Regression và Decision Tree để xây dựng một mô hình dự đoán trong bối cảnh ngành ngân hàng.

*Yaya và đồng sự* (2009) đề xuất mô hình Improved Balanced Random Forests (IBRF) để giải quyết các vấn đề về dữ liệu như bất cân bằng dữ liệu, các nhiễu (noise) trong mô hình.

*Anil Kumar và Ravi* (2008) sử dụng các mô hình Multilayer perceptron, Logistic Regression, Decision Tree, Random Forest, radial basis function network, Support Vector Machine để dự đoán khách hàng rời bỏ dịch vụ thẻ tín dụng.

*Lin và đồng sự* (2011) đã sử dụng lý thuyết tập thô, các kỹ thuật dựa trên đặc trưng cơ sở để trích xuất các đặc trưng có liên quan đến các tài khoản thẻ của khách hàng căn cứ phân tích mạng lưới của đồ thị.

*Farquad và đồng sự* (2014) sử dụng SVM để dự đoán khách hàng rời bỏ khi sử dụng thẻ tín dụng. Mô hình của họ bao gồm ba giai đoạn: 1) đệ quy SVM để loại bỏ dần các đặc tính không có ý nghĩa; 2) bộ dữ liệu giảm thiểu được sử dụng để dự đoán dựa vào SVM, và 3) tạo ra cây đặc trưng bằng mô hình Naïve Bayes.

Các mô hình trên được tóm tắt vào Bảng 1.

Bảng 1.1. Các các giả và các thuật toán sử dụng.

|  |  |
| --- | --- |
| Tác giả | Thuật toán sử dụng |
| Buckinx và đồng sự (2005) | Logistics regression |
| Anil Kumar and Ravi (2008) | Multilayer perceptron, logistic regression, decision tree, random forest, radial basis function network, support vector machine techniques |
| Nie và đồng sự (2011) | Logistic regression, decision tree |
| Lin và đồng sự (2011) | Rough set theory, rule-based decision-making technique, flow network graph |
| Farquad và đồng sự (2014) | Support vector machine, naïve Bayes, tree rules |

**1.3 Phương pháp nghiên cứu**

Bài nghiên cứu được tác giả thực hiện theo các bước mô tả như Hình 1.1 bao gồm các giai đoạn:

1. Tổng hợp dữ liệu, chuẩn bị dữ liệu.

2. Làm sạch dữ liệu, lựa chọn thuộc tính.

3. Chuyên đổi dữ liệu, lựa chọn thuộc tính gán nhãn.

4. Ứng dụng các thuật toán.

5. Đánh giá mô hình, lặp lại bước 2 để có các thuật toán cho hệ số cao hơn.

Hình 1.1. Các bước thực hiện mô hình Data Science trong bài nghiên cứu.

**Graphical user interface

Description automatically generated**

# **PHẦN 2. NGHIÊN CỨU TÌNH HUỐNG**

## **2.1 Mô tả dữ liệu**

## **2.1.1 Nguồn gốc dữ liệu**

Trong bài nghiên cứu này, tác giả sử dụng bộ dữ liệu do một chi nhánh ngân hàng cung cấp. Bộ dữ liệu được lấy trực tiếp từ cơ sở dữ liệu của trụ sở chính phân về cho chi nhánh và có cấu trúc cây như vế bên trái của Hình 2.1.

Dữ liệu cần thu thập nằm trong các tập tin con theo năm tháng ngày, định dạng nén (.rar) và có đường dẫn như sau: “…/201\*/T\*/N\*/Statement\*.rar” và “…/201\*/T\*/N\*/info\*.rar”. Tập tin Statement\*.rar có chứa sao kê số dư hằng ngày của khách hàng và info\*.rar chứa thông tin của khách hàng mới phát sinh trong từng giao dịch. Lý do có 2 tập tin statement khác nhau là vì chi nhánh quản lý hai địa bàn giao dịch khác nhau, vì vậy cũng phát sinh tập tin info1, info2 theo từng địa bàn đó. Như vậy để có thể dễ dàng kiến trúc và phân tích dữ liệu, tác giả đã thực hiện giải nén các tập tin theo định dạng trên vào một thư mục chung như vế bên phải của Hình 2.1. Vì tính chất khác nhau như trên, tác giả đã soạn một đoạn code Python[[1]](#footnote-1) để thực hiện hai nhiệm vụ: (1) giải nén tệp tin, (2) đưa các tệp tin đã giải được giải nén ở (1) lần lượt vào hai thư mục khác nhau.:

* Thư mục Statements: chứa 1316 tập tin sau khi giải nén,
* Thư mục INFO: chứa 1083 tập tin sau khi giải nén.

Sau khi đã có hai thư mục chứa tập tin statement và info, tác giả sẽ tổng hợp, làm sạch, lựa chọn và kiến trúc dữ liệu lần lượt từng thư mục khác nhau.

Hình 2.1 Cấu trúc dữ liệu gốc.

A screenshot of a cell phone

Description automatically generated

### **2.1.1 Sao kê hằng ngày (Statements)**

Là một bộ sao kê hằng ngày của từng khách hàng trong khoảng thời gian 01/01/2018 – 31/12/2019. Tổng cộng trong bộ dữ liệu có 74 cột[[2]](#footnote-2), trong đó có một số thông tin cơ bản để định danh khách hàng như ngày giao dịch, mã khách hàng, loại khách hàng.

Trong tập tin của một ngày có xấp xỉ 60.000 dòng tương đương với 60.000 tài khoản tiền gửi và dư nợ. Một khách hàng có thể có nhiều tài khoản nhưng chỉ có duy nhất một mã khách hàng để định danh. Vì số lượng tập tin theo từng ngày chứa lượng thông tin rất lớn nên tác giả sẽ sử dụng thư viện PySpark để thực hiên các bước tiền xử lý dữ liệu.

Bộ dữ liệu Statements của khách hàng có tổng cộng 39.965.359 dòng, cần 24 phút để đọc dữ liệu. Vì bộ dữ liệu chứa cả nghiệp vụ huy động vốn và cho vay của khách hàng, do đó tác giả chỉ lựa chọn các thuộc tính có liên quan đến nguồn vốn, loại bỏ các thuộc tính của cho vay và không liên quan khác.

Sau khi loại bỏ thì lại các cột (biến) không mang ý nghĩa trong mô hình, tác giả thu được các thuộc tính được chọn để thực hiện biến đổi như bảng 2.1.

Bảng 2.1 Các thuộc tính được lựa chọn từ dữ liệu Statements.

|  |  |
| --- | --- |
| Tên cột | Giải thích |
| NgayGD | Ngày giao dịch (20171231, 20180102, …) |
| LoaiTien | Các loại tiền gửi (AUD, EUR, USD, VND). |
| MaKH | Mã khách hàng (mã định danh một khách hàng) |
| SoDuCuoi | Số dư tiền gửi cuối ngày |
| MaGD | Loại chi tiết tiền gửi (01, 02, 03…) |
| NghiepVuChiTiet | Nghiệp vụ chi tiết  (SA: tiền gửi có kỳ hạn, DA: Tiền gửi không kỳ hạn) |
| LaiSuat | Lãi suất hợp đồng tiền gửi (%) |
| CachTinhLai | Cách tính lãi hằng tháng hoặc năm |
| KyHan | Kỳ hạn tiền gửi |
| PGD | Phòng giao dịch (00, 01, 02, 03, 05, 06, 07, 08) |
| TPKinhTe | Thành phần kinh tế |
| LoaiKH | Loại khách hàng (100-300 là cá nhân, >300 là doanh nghiệp) |
| LoaiKHChiTiet | Loại khách hàng chi tiết |

Tiếp theo, tác giả tạo ra một bộ dữ liệu tổng hợp với mỗi dòng đại diện cho một khách hàng với các thuộc tính dựa vào các cột trong bảng 2.1:

* Cột LoaiTien: chỉ giữ hai loại tiền tệ là VND và USD (vì đây là hai loại tiền được giao dịch nhiều nhất).
* Cột SoDuCuoi: tạo thành các cột avg, min, max của từng loại tiền cho từng khách hàng.
* Cột LaiSuat: mỗi loại tiền gửi sẽ ứng với mức lãi suất khác nhau. Theo thống kê, tiền gửi không kỳ hạn có lãi suất bình quân là 0.2%, từ 1 – 5 tháng có lãi suất bình quân từ 4,12% – 4,97%, lãi suất kỳ hạn 7 tháng 4,15%, 9 – 12 tháng từ 5,69 – 6,64%, từ 13 – 24 tháng lãi suất dao động quanh mức 6,29% - 7,04%, kỳ hạn 3 năm 5 năm và 10 năm lãi suất lần lượt là 9,86%.

Bảng 2.2. Trung bình lãi suất của các kỳ hạn khác nhau.

|  |  |  |  |
| --- | --- | --- | --- |
| Kỳ hạn  (Tháng) | Lãi suất bình quân  (%) | Kỳ hạn  (Tháng) | Lãi suất bình quân (%) |
| 0 | 0,2 | 12 | 6,64 |
| 1 | 4,12 | 13 | 7,04 |
| 2 | 4,04 | 15 | 6,89 |
| 3 | 4,71 | 18 | 6,29 |
| 4 | 4,81 | 24 | 6,64 |
| 5 | 4,97 | 36 | 9,3 |
| 6 | 5,31 | 60 | 6,36 |
| 7 | 4,15 | 120 | 9,86 |
| 9 | 5,69 | - | - |

Vì từng khách hàng có nhiều tài khoản tiền gửi với mức lãi suất, kỳ hạn gửi, số tiền gửi khác nhau nên tác giả sẽ tính lãi suất bình quân của một khách hàng bằng bình quân gia quyền trên số tiền gửi.

Trong đó:

+ ravg : Lãi suất bình quân của một khác hàng,

+ STi: Số tiền gửi của khách hàng tại kỳ thứ i,

+ ri: Lãi suất tương ứng với loại tiền gửi và kỳ hạn tại kỳ thứ i.

* Cột NghiepVuChiTiet, MaGD, PGD: là các thuộc tính chứa thông tin lần lượt về: tính chất khái quát tiền gửi (có kỳ hạn hay không kỳ hạn), tính chất chi tiết tiền gửi (kỳ hạn gửi, cách tính lãi, thành phần gửi - được giải thích cụ thể ở phần phụ lục) và nơi giao dịch (nơi gửi). từng loại sẽ được tạo thành các thuộc tính riêng biệt để theo dõi hành vi gửi tiền khách hàng.

### **2.1.2 Thông tin khách hàng (INFO)**

Là các khách hàng mới phát sinh giao dịch hằng ngày. Vì vậy lượng thông tin của khách hàng đã giao dịch sẽ được lấy từ hệ thống truy vấn cũ của Ngân hàng.

Tuy có nhiều thuộc tính nhưng số lượng thuộc tính có thể sử dụng và có ý nghĩa rất ít nên tác giả chỉ chọn các cột như GIOITINH (giới tính), (DIACHINR) địa chỉ nhà riêng, và các khách hàng được chăm sóc thường xuyên (Account\_Management) để khảo sát và đưa vào mô hình.

### **2.1.3 Gán nhãn dữ liệu**

**Cách 1. Phương pháp loại suy theo kinh nghiệm**

**Nguyên tắc**

Theo V. Avon (2016) một trong các nguyên tắc thông thường để gán nhãn là:

Cố định một thời khoảng thời gian theo tháng hoặc năm.

Đếm số lượng khách hàng rời bỏ trong kỳ này.

Chia số đó cho số lượng khách hàng đầu kỳ.

Tác giả tổng hợp các khách hàng gửi/rút tiền trong năm 2018. Sau đó, so sánh năm 2018 với danh sách khách hàng tại thời điểm 31/12/2019. Các khách hàng mới trong năm 2019 không được tính vì sẽ ảnh hưởng đến các tính chất đã được quan sát trong một kỳ (V. Avon, 2016). Nếu khách hàng nào không xuất hiện, nhãn được gán là 1, ngược lại là 0. Dựa vào kinh nghiệm, có hai lý do để gán nhãn theo cách trên:

i. Đa số khách hàng cá nhân gửi tiền với kỳ hạn ngắn. Thật vây, biểu đồ 2.1 cho thấy các loại tiền gửi không kỳ hạn chiếm tỷ trọng đến 87%, sau đó lần lượt kỳ hạn 1-6 tháng, 7-12 tháng và trên 12 tháng.

ii. Một số khách hàng gửi trong năm nay nhưng thời điểm kết thúc hợp đồng tiền gửi vào năm tiếp theo.

Biểu đồ 2.1. Tỷ trọng kỳ hạn gửi tiền.

Chart, pie chart

Description automatically generated

**Cách 2. Sử dụng kết hợp giữa suy luận loại suy và học bán giám sát**

Cách 1 là một phương pháp dựa vào kinh nghiệm và các quy tắc do chủ thể gán nhãn đặt ra nên ít dựa vào các phương pháp khoa học. Do đó, tác giả đề xuất một phương pháp gán nhãn kết hợp của loại suy và học bán giám sát.

Propagate labeling là một thuật toán vòng lặp ứng dụng lý thuyết đồ thị và được sử dụng để gán nhãn theo phương thức lan truyền trong tập dữ liệu. Propagate labeling được Xiaojin Zhu và Zoubin Ghahramani công bố vào năm 2002. Vì thuật toán này cần một tập dữ liệu có nhãn để học nên tác giả sẽ lấy ngẫu nhiên 10% số lượng mẫu từ tập dữ liệu được gán ở phần Cách 1 và gán cho những phần tử mẫu còn lại của tập dữ liệu.

## **2.2 Explore Data Analysis:**

Các biến đã lọc ở phần 2.1 được tập hợp thành một bảng dữ liệu hoàn chỉnh gồm 62.536 dòng và 52 cột (bao gồm nhãn), mỗi dòng của dữ liệu đại diện cho một khách hàng trong năm 2018 chứa hành vi, thông tin gửi, loại tiền tệ gửi, trung bình số dư tiền gửi, mã tiền gửi, mã nơi gửi[[3]](#footnote-3), địa chỉ nhà, giới tính, người quản lý. Các cột thuộc tính trong bộ dữ liệu có ý nghĩa và giá trị như bảng 2.3.

Bảng 2.3. Giải thích ý nghĩa các cột thuộc tính.

|  |  |
| --- | --- |
| Tên cột | Giải thích |
| avg\_VND | Trung bình tiền gửi VND (ngàn VND) |
| avg\_USD | Trung bình tiền gửi USD (USD) |
| max\_VND | Tiền gửi VND tối đa (ngàn VND) |
| min\_VND | Tiền gửi VND đối thiểu (ngàn VND) |
| max\_USD | Tiền gửi USD tối đa (USD) |
| min\_USD | Tiền gửi USD đối thiểu (USD) |
| avg\_interest | Bình quân lãi suất tiền gửi (%) |
| 00\_MaGD – 32\_MaGD | Mã tiền gửi từ 00\_MaGD – 32\_MaGD  (0: không có tài khoản, 1: Có tài khoản) |
| 00\_PGD - 08\_PGD | Mã nơi gửi 00 – 08  (0: Không mở tài khoản, 1: Có mở tài khoản) |
| DA\_NghiepVuChiTiet | Tiền gửi không kỳ hạn (0: không gửi, 1: Có gửi) |
| SA\_NghiepVuChiTiet | Tiền gửi có kỳ hạn (0: không gửi, 1: Có gửi) |
| GIOITINH | Giới tính khách hàng (0: Nữ, 1: Nam) |
| Account\_Management | Có người quản lý thường xuyên (0: không, 1: Có) |
| label | Nhãn khách hàng (0: Đã rút tiền, 1: Còn gửi tiền) |

### **2.2.1 Phân tích đơn biến**

**A. Biến phân loại.**

Mô hình có 45 biến phân loại, trong đó các biến có tần số xuất hiện thấp và không có khách hàng gửi được loại bỏ: 22\_MaGD, 24\_MaGD, 25\_MaGD, 26\_MaGD, 27\_MaGD, 28\_MaGD, 29\_MaGD, 30\_MaGD, 31\_MaGD, 32\_MaGD.

Theo biểu đồ 2.2, mã tiền gửi 0\_MaGD chiếm tỷ trọng lớn nhất với 68%, kế đến là 1\_MaGD với tỷ trọng 18%, lần lượt thứ 4, 5, 6 là các loại 3\_MaGD, 4\_MaGD, 2\_MaGD. Các loại tiền gửi có tần suất xuất hiện thấp được gộp lại thành ‘other’. Tuy nhiên, trong mô hình các loại tiền gửi vẫn được tách biệt để nghiên cứu hành vi khách hàng.

Biểu đồ 2.2. Tỷ trọng các loại tiền gửi.

Chart, pie chart

Description automatically generated

Với nơi gửi tiền ta có các cột với tỷ trọng khách hàng được biểu diễn ở biểu đồ 2.3.

Biểu đồ 2.3 Tỷ trọng khách hàng tại các địa điểm gửi tiền.

Chart, pie chart

Description automatically generated

56% khách gửi tại 00\_PGD, 21% gửi tại 03\_PGD, 7% chọn 02\_PGD, 5% gửi tại 08\_PGD, 4% khách gửi có tiền gửi tại 07\_PGD và 05\_PGD, còn lại là 06\_PGD.

**B. Biến liên tục.**

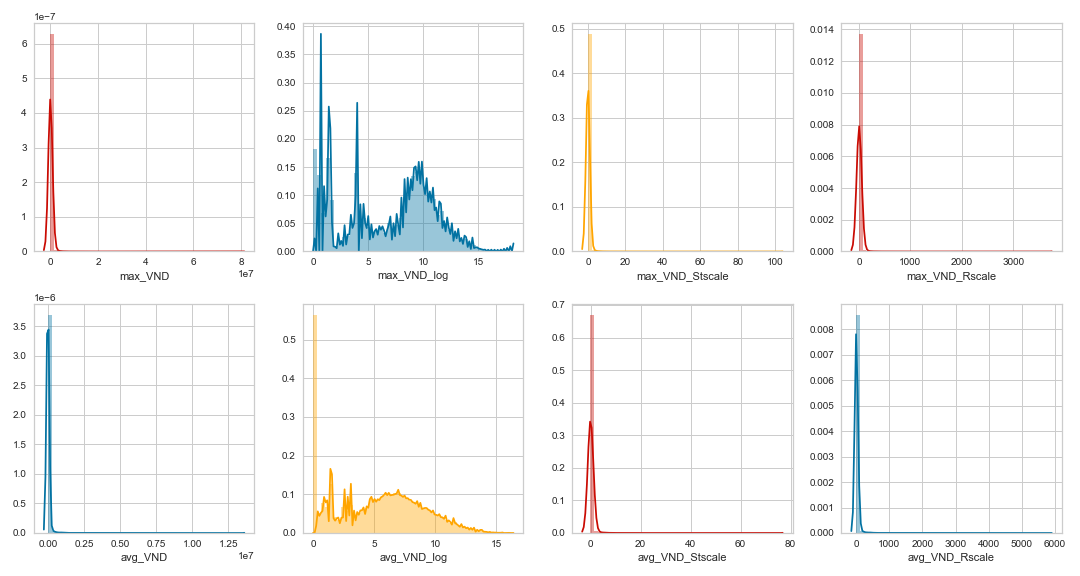
Mô hình có 07 biến liên tục là avg\_VND, avg\_USD, max\_VND, min\_VND, max\_USD, min\_USD, avg\_interest. Thực hiện thống kê mô tả cho ta các thông số về min, max, độ lệch chuẩn, các tứ phân vị như Bảng 2.4 thể hiện.

Bảng 2.4 Thống kê mô tả các biến liên tục.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | avg\_VND | min\_VND | max\_VND | avg\_USD | min\_USD | max\_USD | avg\_interest |
| mean | 23.921,28 | 2.489,88 | 99.255,30 | 95,01 | 31,64 | 127,99 | 0,60 |
| std | 192.986,80 | 68.797,23 | 799.276,50 | 4.289,36 | 2.567,24 | 5.645,14 | 1,36 |
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 4,30 | 0 | 25,50 | 0 | 0 | 0 | 0,20 |
| 50% | 202,22 | 0 | 2.963,20 | 0 | 0 | 0 | 0,21 |
| 75% | 2.523,31 | 41,75 | 215.19,14 | 0 | 0 | 0 | 0,24 |
| max | 13.348.670 | 9.239.000 | 78.606.410 | 619.318 | 515.965 | 700.000 | 11,01 |

Vì các biến trên có cận trên và cận dưới rất cách biệt và đơn vị của các biến khác nhau (VND, USD, %), nên tác giả sẽ sử dụng các phép biến đổi log, Robust Scaler, và Standard Scaler.

Biểu đồ 2.4. Đồ thị phân phối của max\_VND, avg\_VND với các cách biến đổi khác nhau.



Các thuộc tính tiền gửi trước và sau khi biến đổi có sự thay đổi về khoảng giá trị để loại bỏ sự khác biệt về độ lớn, cũng như đơn vị định lượng.

### **2.2.2 Phân tích hai biến**

**A. Giữa label các biến phân loại**

Tác giả thực hiện xác định mối quan hệ giữa các biến phân loại bằng kiểm định Chi bình phương[[4]](#footnote-4). Theo đó, giả thiết H0 giả định hai biến độc lập, đối giả thuyết H1 giả định hai biến phụ thuộc.

Với mức ý nghĩa 5%, có căn cứ và bằng chứng thống kê cho thấy có mối quan hệ phụ thuộc giữa label và các biến 0\_MaGD, 11\_MaGD, 15\_MaGD, 1\_MaGD, 2\_MaGD, 3\_MaGD, 4\_MaGD, 6\_MaGD, 7\_MaGD, 8\_MaGD, 00\_PGD, 02\_PGD, 03\_PGD, 05\_PGD, 06\_PGD, 07\_PGD, 08\_PGD, DA\_NghiepVuChiTiet, SA\_NghiepVuChiTiet, GIOITINH, Account\_Management.

**B. Giữa label các biến liên tục**

Tác giả thực hiện xác định mối quan hệ thống kê bằng kiểm định z-test thông qua mô hình hồi quy tuyến tính. Qua đó, với mức ý nghĩa 5% ta xác định được các p-value của mô hình trong bảng 2.4.

Với mức ý nghĩa 5%, có căn cứ kết luận và bằng chứng thống kê cho thấy có mối quan hệ có ý nghĩa giữa label và các biến avg\_VND, max\_VND, min\_VND, avg\_interest. Các thuộc tính thuộc về USD chưa cung cấp đủ bằng chứng để đánh giá hành vi rút tiền của khách hàng, do số lượng nhỏ và chủ yếu tập trung ở các khách hàng tổ chức, doanh nghiệp xuất nhập khẩu tại ngân hàng.

Bảng 2.4 p-value của các biến liên tục với lable.

|  |  |
| --- | --- |
|  | P > |t| |
| avg\_VND | **0.000** |
| avg\_USD | **0.393** |
| max\_VND | **0.000** |
| min\_VND | **0.011** |
| max\_USD | **0.724** |
| min\_USD | **0.149** |
| avg\_interest | **0.000** |

Bên cạnh đó, biểu đồ 2.5 cho thấy trung bình tiền gửi và lãi suất càng cao thì tỷ lệ rút tiền càng lớn. Phát biểu trên đúng với cả giới tính nam và nữ.

Biểu đồ 2.5 Trung bình tiền gửi – giới tính, lãi suất – giới tính với nhãn

Chart, bar chart

Description automatically generated

### **2.2.3 Các ngoại lệ (outliers)**

Ngoài các phương pháp được sử dụng thường xuyên nhằm phát hiện các ngoại lệ như vẽ boxplot, histogram hoặc tính bằng trị số z-score nếu thuộc tính có phân phối chuẩn. Trong bài này, tác giả đề xuất thêm một phương pháp nhằm phát hiện các quan sát có ảnh hưởng lớn, đó là thước đo khoảng cách Cook[[5]](#footnote-5).

Biểu đồ 2.6 Thước đo khoảng cách Cook với biến max\_VND.

Chart, histogram

Description automatically generated

Biểu đồ 2.6 cho thấy có một quan sát D > 1.4 cho thấy có ảnh hưởng lớn đến biến label, với giá trị max\_VND = 78.606.414 ở quan sát thứ 44878. Đây là một khách hàng tại nơi giao dịch số 06\_PGD, loại gửi tiền 0\_MaGD, giới tính Nam, lãi suất bình quân 2.0%, là một khách hàng ưu tiên, nhãn không rút tiền. Với các thuộc tính trên, tác giả không thấy có bất thường nên quyết định giữ lại quan sát này.

### **2.2.4 Xử lý mất cân bằng dữ liệu**

Dữ liệu quan sát có tỷ lệ rút tiền là 1/5, nên xảy ra hiện tượng mất cân bằng dữ liệu. Trong bài, hai thuật toán được sử dụng để cân bằng dữ liệu là Smote và thư viện utils của sklearn. Nếu Smote là một thuật toán ứng dụng k giá trị lân cận thì utils cân bằng bộ dữ liệu bằng bằng phép gán giá trị ngẫu nhiên. Việc so sánh các phương thức trên sẽ được tác giả thực hiện đồng thời khi thực hiện mô hình.

## **2.3 Thực hiện mô hình.**

Các bài toán dự đoán các khách hàng rời bỏ (rút tiền) đã có rất nhiều cách tiếp cận và ứng dụng các thuật toán đa dạng. Trong bài này, các thuật toán được sử dụng bao gồm Logistics Regression, Decision Tree, Random Forest, và Extreme Gradient Boosting (XGBoost). Các thuật toán trên sẽ được huấn luyện với các bộ dữ liệu có cách biến đổi biến liên tục (i) và cách xử lý cân bằng dữ liệu (j) khác nhau. Trong đó:

i = (Không xử lý, Logarithm, Standard Scaler, Robust Scaler, Logarithm & Standard Scaler, Logarithm & Robust Scaler).

j = (Không xử lý, Cân bằng Smote, Cân bằng Utils)

Sau đó các hệ số f1 và hệ số recall sẽ được so sánh để tìm ra mô hình có kết quả tốt nhất. Ngoài ra, các mô hình cũng được đánh giá hiệu quả qua việc so sánh thời gian xử lý dữ liệu.

### **2.3.1 Mô hình dự đoán nhãn gán bằng Cách 1**[[6]](#footnote-6)

Với các thông số tốt nhất ta có kết quả:

* hệ số f1 lớn nhất (~95%) thuộc về mô hình: i = Standard Scaler, j = không xử lý, thuật toán XGBoost.
* hệ số recall lớn nhất (~95%) thuộc về mô hình i = không xử lý, j = cân bằng Utils, thuật toán XGBoost.
* thời gian xử lý nhanh nhất (~0.8 giây) thuộc về thuật toán Logistic Regression với mô hình i = không xử lý, j = cân bằng Utils.

Với các thông số xấu nhất ta có kết quả:

* hệ số f1 nhỏ nhất (~4%) thuộc về mô hình: i = không xử lý, j = không xử lý, thuật toán Logistic Regression.
* hệ số recall nhỏ nhất (~2.2%) thuộc về mô hình i = không xử lý, j = không xử lý, thuật toán Logistic Regression.
* thời gian xử lý chậm nhất (~18 giây) thuộc về thuật toán Random Forest với mô hình i = Logarithm, j = không xử lý.

### **2.3.2 Mô hình dự đoán nhãn gán bằng Cách 2**

Với các thông số tốt nhất ta có kết quả:

* hệ số f1 lớn nhất (~96.7%) thuộc về mô hình i = không xử lý, j = không xử lý, thuật toán Random Forest.
* hệ số recall lớn nhất (~98%) thuộc về mô hình i = Logarithm, j = Cân bằng Utils, thuật toán XGBoost.
* thời gian xử lý nhanh nhất (~0.22 giây) thuộc về thuật toán Decision Tree với mô hình i = không xử lý, j = không xử lý.

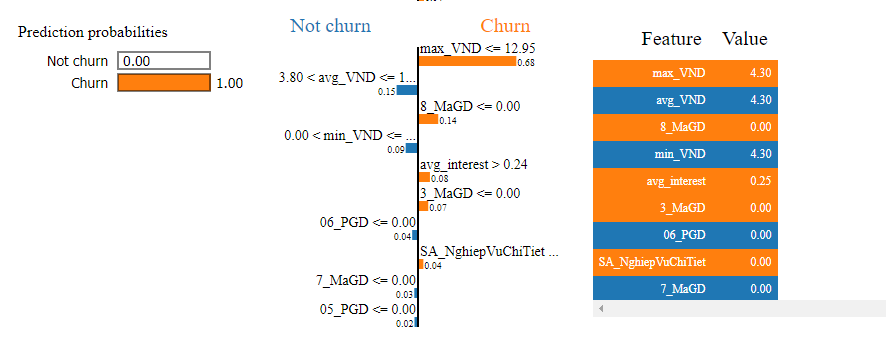
Với các thông số xấu nhất ta có kết quả:

* hệ số f1 nhỏ nhất (~2,3%) thuộc về mô hình i = không xử lý, j = không xử lý, thuật toán Logistic Regression.
* hệ số recall nhỏ nhất (~61%) thuộc về mô hình i = không xử lý, j = không xử lý, thuật toán Logistic Regression.
* thời gian xử lý chậm nhất (~5 giây) thuộc về thuật toán Random Forest với mô hình i = Robust Scaler, j = Cân bằng Smote.

### **2.3.3 Diễn giải mô hình bằng LIME**

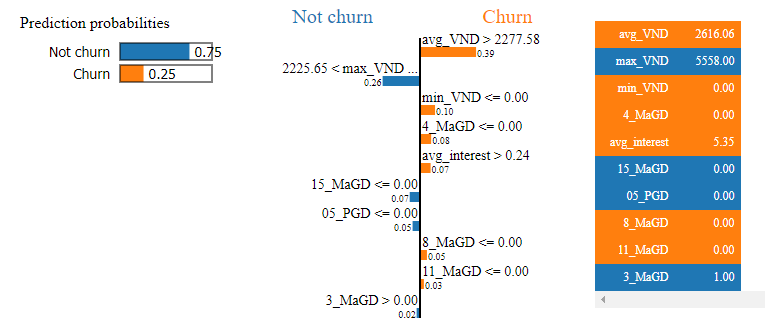
LIME là viết tắt của Local Interpretable Model-agnostic Explanations (Tạm dịch: phép diễn giải cục bộ cho mô hình bất khả tri). Đây một kỹ thuật của vận dụng cách thức diễn giải cục bộ nhằm giải thích cho những mô hình phân loại có tính phức tạp cao. LIME có hai khả năng diễn giải mô hình:

1. Diễn giải mô hình cục bộ.
2. Chức năng lựa chọn một bộ dữ liệu mang tính đại diện nhằm đánh giá tính nhất quán của mô hình khi mô phỏng hành vi của con người. Tập đại diện này sẽ cung cấp một góc nhìn toàn cảnh để người sử dụng dễ hiểu mô hình hơn.

Hình 2.2 Diễn giải cục bộ trên một quan sát ví dụ 1

Theo hình 2.2, khách hàng A gửi tại có số tiền gửi nhỏ hơn 12.95, không gửi tài khoản loại 3\_MaGD, 8\_MaGD, lãi suất bình quân lớn hơn 0,24% Thuộc tính Max\_VND đã đẩy xác xuất rút tiền của A lên rất cao.

Hình 2.3 Diễn giải cục bộ trên một quan sát ví dụ 2



Xem hình 2.3, ta quan sát cách mô hình quyết định, ta có thể thấy được có lực đẩy và kéo tại nơi mô hình dự đoán. Một mặt, yếu tố avg\_VND > 2277.58 làm tăng khả năng rút tiền, nhưng yếu tố max\_VND lại kéo khả năng khách hàng ở lại. Cuối cùng lực kéo đã chiến thắng và mô hình đã dự đoán xác suất khách hàng ở lại đến 75%.

Bên cạnh đó, như tác giả đã đề cập, LIME cung cấp một bộ dữ liệu con có thể chứa các tính chất đại diện cho toàn bộ dữ liệu. Xem xét biểu đồ 2.7, nếu số dư tối đa trong tài khoản VND lớn hơn 21.596,27 thì xác suất khách hàng ở lại cao hơn các khách hàng rút tiền. Hoặc, nếu khách hàng có max\_VND nhỏ hơn 12,95 sẽ thì nhiều khả năng sẽ rút tiền.

Biểu đồ 2.7 Trung bình các hệ số tương quan của LIME trong tập đại diệnA picture containing chart

Description automatically generated

# **PHẦN 3: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **3.1 Kết luận mô hình**

Thứ nhất, qua việc phân tích kết quả các mô hình và thuật toán, ta có thể kết luận rằng có sự đánh đổi giữa hiệu suất và thời gian xử lý. Tuy nhiên thuật toán XGBoost mang lại sự tin cậy tương đối về cả hiệu suất và thời gian xử lý mô hình. Theo đó, với XGBoost thời gian xử lý thấp nhất là 1.23 giây, với chỉ số f1 và recall tương ứng là ~91% và ~91%. Ngược lại, đối với hệ số f1 và recall lớn nhất cho thời gian xử lý lần lượt là 1.45 giây và 2.1 giây.

Thứ hai, thuật toán propagate labeling tỏ ra hiệu quả hơn so với việc gán nhãn bằng kinh nghiệm và ngoại suy. Tuy nhiên, do thời gian ngắn cũng như xác suất gán nhãn đúng chỉ xấp xỉ 84% nên tác giả cần thêm dữ liệu để đánh giá đầy đủ tính khả thi và khát quát của thuật toán này để đưa mô hình ứng dụng vào thực tế.

Thứ ba, mô hình đã phát hiện rằng nếu số tiền nhỏ hơn 12.000 VND thì xác suất rút tiền của khách hàng rất cao. Thật vậy, đối với các khách hàng chỉ có tài khoản thanh toán mà không thực hiện giao dịch, số tiền sẽ tiến về 0. Điều này hàm ý rằng có sự tương quan dương giữa số tiền gửi và khả năng rút tiền.

Thứ tư, việc dự đoán được khách hàng rút tiền có các ứng dụng thực tế:

1. Dự đoán các khách hàng rút tiền để đề xuất các chiến thuật giữ chân khách hàng nhằm nâng cao tỷ lệ khách hàng ở lại.

2. Tính toán và dự báo kỳ vọng lợi nhuận của ngân hàng trong năm dựa vào xác suất rời bỏ của khách hàng.

## **3.2 Hướng phát triển**

## **3.2.1 Về độ chính xác của mô hình**

Tuy đạt được độ chính xác tương đối cao, nhưng mô hình chưa tuyệt đối. Nguyên nhân có thể nằm tại nhiễu do đưa quá nhiều biến, hoặc do một số thuộc tính tác giả chưa đưa vào mô hình như nơi ở, nơi mở tài khoản gốc, gốc thời gian bắt đầu sử dụng dịch vụ ngân hàng, độ tuổi, tình trạng hôn nhân… vì nguyên nhân khách quan là cơ sở dữ liệu của ngân hàng chưa đầy đủ.

Ngoài ra, mô hình là một đánh giá trong ngắn hạn, chưa thực hiện quan sát trong tầm trung hạn (1-5 năm) và dài hạn (từ 5 năm). Do đó, ngân hàng cần tiếp tục thu thập, lưu trữ thêm thông tin về các thuộc tính trên nhằm tiếp tục theo dõi hành vi, đánh giá, cập nhật, cải thiện mô hình. Trên cơ sở đó, cung cấp một thước đo đúng đắn nhất về khách hàng, nâng cao năng lực cạnh tranh của ngân hàng trong một thời kỳ có nhiều biến đổi và bùng nổ về thông tin như hiện nay.

## **3.2.2 Về mở rộng khả năng phát triển của mô hình**

Đề tài đã khai thác một mảng nghiệp vụ lớn của ngân hàng là huy động vốn. Tuy nhiên, để đánh giá tổng quát một khách hàng thì phải xem xét tất cả các dịch vụ mà khách hàng sử dụng tại ngân hàng đó, bao gồm vay vốn và dịch vụ (sử dụng thẻ, các dịch vụ chuyển tiền, Internetbanking, E-Mobile Banking – nếu có).

Do đó để có thể khai thác, tìm kiếm các mô thức ẩn trong hệ thống dữ liệu lớn mang lại hiệu quả kinh doanh cao, tác giả khuyến nghị:

1. Lập ra một bộ phận chuyên trách bao gồm: các thành viên từ ban chuyên môn, hiểu biết về nghiệp vụ ngân hàng; các thành viên từ ban công nghệ thông tin và hiểu biết về hệ thống dữ liệu cơ sở; các thành viên có hiểu biết về triển khai, xây dựng, đánh mô hình thuật toán trí tuệ nhân tạo, machine learning.

2. Nhiệm vụ của bộ phận trên là tổng hợp, lựa chọn các thông tin của khách hàng từ các module nghiệp vụ khác nhau, từ đó xây dựng nên một mô hình chung cho toàn bộ khách hàng và áp dụng các bài toán khác nhau dựa vào mô hình đó. Trong đó có mô hình khách hàng rời bỏ, mô hình đánh giá rủi ro tín dụng, mô hình marketing dựa vào phân cụm khách hàng, mô hình phát hiện gian lận và các ứng dụng khác của bộ môn khoa học dữ liệu.

# PHỤ LỤC 1. TÊN CỘT VÀ GIẢI THÍCH STATEMENT VÀ INFO

| **STATEMENT COLUMNS** | | **INFO COLUMNS** | |
| --- | --- | --- | --- |
| **Tên cột** | **Giải thích** | **Tên cột** | **Giải thích** |
| NGAYGD \*\* | Ngày giao dịch | NGAYGD | Ngày giao dịch |
| MACN | Mã chi nhánh (1, 2) | STT | Số thứ tự |
| LOAITIEN \*\* | Loại tiền gửi (VND, USD) | MACN | Mã chi nhánh |
| MATK \*\* | Mã tài khoản | MAKH | Mã khách hàng |
| MAKH | Mã khách hàng | MATT | Mã thông tin |
| SOTT | Số thứ tự | NGAYTAO | Ngày khởi tạo |
| SODUDAU | Số dư đầu ngày | MSNSD | Mã số nhân viên khởi tạo |
| SODUCUOI \*\* | Số dư cuối ngày | MANQL | Mã số người phê duyệt |
| MAGD \*\* | Mã giao dịch | TTKH | Thông tin khách hàng |
| SOGD | Số giao dịch | TENQG | Tên quốc gia |
| NGHIEPVU | Nghiệp vụ (huy động vốn, cho vay, bảo lãnh…) | LOAIKH | Loại khách hàng |
| NGHIEPVUCHITIET \*\* | Nghiệp vụ chi tiết | LOAIKHCT | Loại khách hàng chi tiết |
| MALSCOBAN | Mã lãi suất cơ bản | MANH | Mã người nhận |
| LSCOBAN | Lãi suất cơ bản | TENTATKH | Tên tắt khách hàng |
| CHENHLECH | Chênh lệch với lãi suất tiền gửi | TENKH | Tên khách hàng |
| LAISUAT \*\* | Lãi suất hợp đồng tiền gửi | EMAIL | Email |
| DUTHU/DUCHI | Dự thu/Dự chi | MANKT | Mã người khởi tạo |
| NGAYMO | Ngày mở hợp đồng tiền gửi | VIPTT | VIP thông thường |
| NGAYDH | Ngày đến hạn hợp đồng tiền gửi | MST | Mã số thuế |
| NGAYDONG | Ngày đóng hợp đồng tiền gửi | CNGOC | Chi nhánh gốc |
| NGAYTRALAITIEP | Ngày trả lãi tiếp theo | DSDEN | Doanh số đến |
| NGAYTRAGOCTIEP | Ngày trả gốc tiếp theo | CMT | Chứng minh thư |
| SOTRAGOCTIEP | Số tiền gốc phải trả | CMT\_NOICAP | Nơi cấp chứng minh thư |
| CACHTINHLAI \*\* | Cách tính lãi | CMT\_NGAYCAP | Ngày cấp chứng minh thư |
| CACHTRALAI | Cách trả lãi | CMT\_THOIHAN | Thời hạn chứng minh thư |
| CACHTINHKYHANTRA | Cách tính kỳ hạn trả | HC\_SO | Số Hộ chiếu |
| UNNAMED: 26 | Cột trống | HC\_NOICAP | Nơi cấp hộ chiếu |
| SONGAYTRONGNAM | Số ngày trong năm | HC\_NGAYCAP | Ngày cấp hộ chiếu |
| COCAUNO | Cơ cấu nợ | HC\_THOIHAN | Thời hạn hộ chiếu |
| SOTKTHAMCHIEU | Số tài khoản tham chiếu | MST\_SO | Số mã số thuế |
| LOAIKYHANNGAYVAYLAI | Loại kỳ hạn hợp đồng vay | MST\_NOICAP | Nơi cấp mã số thuế |
| DSTHUNOGOCXLRR | Doanh số thu nợ gốc xử lý rủi ro | MST\_NGAYCAP | Ngày cấp mã số thuế |
| DSNOGOC | Doanh số nợ gốc | MST\_THOIHAN | Thời hạn mã số thuế |
| LAIDUTHU/CHICUOITHANG | Lãi dự thu/chi cuối tháng | DKKD\_SO | Số đăng ký kinh doanh |
| KYHAN \*\* | Kỳ hạn hợp đồng tiền gửi | DKKD\_NOICAP | Nơi cấp giấy đăng ký kinh doanh |
| LAIQUAHAN | Lãi cho vay quá hạn | DKKD\_NGAYCAP | Ngày cấp giấy đăng ký kinh doanh |
| GHICHU | Ghi chú | DKKD\_THOIHAN | Thời hạn giấy đăng ký kinh doanh |
| LOAIHACHTOAN | Loại hạch toán | MAQG | Mã quốc gia |
| SOTIENNO | Số tiền nợ | MABG | Mã tỉnh |
| SOTIENCO | Số tiền có | MATH | Mã thành phố |
| PGD \*\* | Phòng giao dịch | MAQH | Mã quận huyện |
| MACNGD | Mã chi nhánh giao dịch | MAXA | Mã xã |
| COQUAHAN | Nợ quá hạn | DIACHI\_NR | Địa chỉ nhà riêng |
| TKMOI | Tài khoản hạch toán mới | SODT\_NR | Số điện thoại nhà riêng |
| TIGIACB | Tỷ giá cơ bản | SOFAX\_NR | Số fax nhà riêng |
| TPKINHTE \*\* | Thành phần kinh tế | DIACHI\_CQ | Địa chỉ cơ quan |
| LOAIKH \*\* | Loại khách hàng | SODT\_CQ | Số điện thoại cơ quan |
| MAQG | Mã quốc gia | SOFAX\_CQ | Số fax cơ quan |
| TIGIACUOITHANG | Tỷ giá cuối tháng | GIOITINH | Giới tính |
| MAPHONGBAN | Mã phòng ban | SMSDK | Đăng ký SMS |
| TKHOANNHNN | Tài khoản NHNN |  |  |
| LOAINGUOCGOCVAY | Loại nguồn gốc vay |  |  |
| COLAPLAILICHVAY | Có lập lãi lịch vay |  |  |
| LIDONOXAU | Lý do nợ xấu |  |  |
| TINHTRANGKHOANVAY | Tình trạng khoản vay |  |  |
| SOTIENVAYTRONGHOPDONG | Số tiền vay trong hợp đồng |  |  |
| LOAIHINHKTE | Loại hình kinh tế |  |  |
| KYHANVAY | Kỳ hạn cho vay |  |  |
| TENKH | Tên khách hàng |  |  |
| CODAMBAOTSAN | Có tài sản đảm bảo |  |  |
| LYDOBAODAM | Lý do bảo đảm |  |  |
| MANGUONVON | Mã nguồn vốn |  |  |
| PHANLOAINO | Phân loại nợ |  |  |
| LOAIVAY | Loại vay |  |  |
| MAHOPDONG | Người quản lý |  |  |
| SOHOPDONG | Số hợp đồng cho vay |  |  |
| NGAYTRANOCUOI | Ngày trả nợ cuối |  |  |
| NGAYGIAHAN | Ngày gia hạn nợ |  |  |
| LAITUNGAYDENNGAY | Lãi từ ngày đến ngày |  |  |
| LOAIKHCHITIET \*\* | Loại khách hàng chi tiết |  |  |
| LOAIDN | Loại Doanh nghiệp |  |  |
| THULAITRONGHAN | Thu lãi trong hạn |  |  |
| THULAIQUAHAN | Thu lãi quá hạn |  |  |
| UNUSED | Cột trống |  |  |

# PHỤ LỤC 2. GIẢI THÍCH Ý NGHĨA CỦA MAGD

| **Tên Mã** | **Ý nghĩa** | **Kỳ gửi** |
| --- | --- | --- |
| 0\_MaGD | TGKKH cá nhân mã mới | Trả lãi hằng tháng |
| 1\_MaGD | TGKKH sinh viên | Trả lãi hằng tháng |
| 2\_MaGD | TGKKH hưu trí hưởng lương NSNN | Trả lãi hằng tháng |
| 3\_MaGD | Tiết kiệm linh hoạt dưới 12T | Trả lãi khi đến hạn |
| 4\_MaGD | TGTK có kỳ hạn dưới 12 tháng | Trả lãi khi đến hạn |
| 5\_MaGD | TGTK có kỳ hạn dưới 24 tháng | Trả lãi khi đến hạn |
| 6\_MaGD | Tiết kiệm linh hoạt từ 12 đến dưới 24T | Trả lãi khi đến hạn |
| 7\_MaGD | TGKKH cán bộ nhân viên | Trả lãi hằng tháng |
| 8\_MaGD | TGTK dự thưởng từ 12 đến 24 tháng | Trả lãi khi đến hạn |
| 9\_MaGD | TGTK có kỳ hạn dưới 24 tháng | Trả lãi sau định kỳ |
| 10\_MaGD | Tiền gửi tiết kiệm (TGTK) không kỳ hạn | Trả lãi hằng tháng |
| 11\_MaGD | Tiết kiệm an sinh từ 12 tháng đến dưới 24 tháng | Gửi góp không theo định kỳ hàng tháng |
| 12\_MaGD | TGKKH cá nhân mã cũ | Trả lãi hằng tháng |
| 13\_MaGD | TGTK dự thưởng TW có kỳ hạn dưới 24T | Trả lãi khi đến hạn |
| 14\_MaGD | TGTK dự thưởng TW có kỳ hạn dưới 12T | Trả lãi khi đến hạn |
| 15\_MaGD | TG trực tuyến < 12T | Trả lãi khi đến hạn |
| 16\_MaGD | Trái phiếu 10 năm | Trả lãi khi đến hạn |
| 17\_MaGD | TGTK có kỳ hạn dưới 12 tháng | Trả lãi tháng |
| 18\_MaGD | TGTK có kỳ hạn từ 24 tháng | Trả lãi khi đến hạn |
| 19\_MaGD | Tiết kiệm an sinh từ 24 tháng trở lên | Gửi góp không theo định kỳ hàng tháng |
| 20\_MaGD | TGTK trả lãi bậc thang (đã hết sử dụng) | Trả lãi định kỳ |
| 21\_MaGD | Tiết kiệm gửi góp dưới 12 tháng | Gửi góp không theo định kỳ hàng tháng |
| 22\_MaGD | Tiết kiệm linh hoạt từ 24T | Trả lãi khi đến hạn |
| 23\_MaGD | TK Dự thưởng TW từ 24T lãi sau | Trả lãi khi đến hạn |
| 24\_MaGD | Tiết kiệm gửi góp học đường từ 12 đến dưới 24 tháng | Thanh toán hàng tháng |
| 25\_MaGD | TK dự thưởng 12T | Trả lãi khi đến hạn |
| 26\_MaGD | TG trực tuyến từ 12 đến dưới 24 tháng | Trả lãi khi đến hạn |
| 27\_MaGD | TGTK Hưu trí từ 12 đến dưới 24 tháng | Trả lãi khi đến hạn |
| 28\_MaGD | Kỳ phiếu trả lãi trước | Trả lãi trước |
| 29\_MaGD | TGTK Hưu trí từ trên 24 tháng | Trả lãi khi đến hạn |
| 30\_MaGD | TG CKH dưới 12 tháng | Trả lãi khi đến hạn |
| 31\_MaGD | TGTK CKH từ 24 tháng | Trả lãi tháng |
| 32\_MaGD | Tiền gửi không kỳ hạn (TGKKH) cá nhân | Trả lãi hằng tháng |

# PHỤ LỤC 3. KIỂM ĐỊNH GIẢ THUYẾT CHI BÌNH PHƯƠNG GIỮA LABEL VÀ CÁC BIẾN PHÂN LOẠI

|  |  |  |
| --- | --- | --- |
|  | p\_value | Kết luận |
| 0\_MaGD | 0 | Bác bỏ H0 |
| 10\_MaGD | 0.53226 | Chấp nhận H0 |
| 11\_MaGD | 2.77309e-10 | Bác bỏ H0 |
| 13\_MaGD | 0.385971 | Chấp nhận H0 |
| 14\_MaGD | 0.809838 | Chấp nhận H0 |
| 15\_MaGD | 0.000436746 | Bác bỏ H0 |
| 16\_MaGD | 0.275589 | Chấp nhận H0 |
| 17\_MaGD | 0.837599 | Chấp nhận H0 |
| 18\_MaGD | 0.870493 | Chấp nhận H0 |
| 19\_MaGD | 0.695033 | Chấp nhận H0 |
| 1\_MaGD | 0 | Bác bỏ H0 |
| 20\_MaGD | 0.765473 | Chấp nhận H0 |
| 21\_MaGD | 0.799892 | Chấp nhận H0 |
| 23\_MaGD | 0.439718 | Chấp nhận H0 |
| 2\_MaGD | 4.66017e-35 | Bác bỏ H0 |
| 3\_MaGD | 1.02996e-44 | Bác bỏ H0 |
| 4\_MaGD | 2.0125e-08 | Bác bỏ H0 |
| 5\_MaGD | 0.212275 | Chấp nhận H0 |
| 6\_MaGD | 3.74876e-05 | Bác bỏ H0 |
| 7\_MaGD | 1.21401e-10 | Bác bỏ H0 |
| 8\_MaGD | 8.35811e-17 | Bác bỏ H0 |
| 9\_MaGD | 0.896609 | Chấp nhận H0 |
| 00\_PGD | 7.54731e-05 | Bác bỏ H0 |
| 02\_PGD | 0.0373125 | Bác bỏ H0 |
| 03\_PGD | 1.55404e-07 | Bác bỏ H0 |
| 05\_PGD | 0.0459202 | Bác bỏ H0 |
| 06\_PGD | 0.0147659 | Bác bỏ H0 |
| 07\_PGD | 8.34329e-10 | Bác bỏ H0 |
| 08\_PGD | 2.55066e-25 | Bác bỏ H0 |
| DA\_NghiepVuChiTiet | 0 | Bác bỏ H0 |
| SA\_NghiepVuChiTiet | 4.99782e-128 | Bác bỏ H0 |
| GIOITINH | 8.83121e-25 | Bác bỏ H0 |
| Account\_Management | 0.00416977 | Bác bỏ H0 |

# PHỤ LỤC 4: CODE PYTHON GIẢI NÉN VÀ ĐƯA FILE VÀO ĐƯỜNG DẪN ĐƯỢC CHỈ ĐỊNH

'''

This script is used to find zipped files, which have .gzip formatted,

in a folder and unzipped to a destined directory for further data analysing.

'''

# To get system path name

**import** **os**

# To parse regular expression to find file’s name

**import** **re**

# Read and unzip \*.RAR, \*.GRAR files

**import** **gzip**, **shutil**

# Run threading

**import** **concurrent.futures**

# Calculate time

**import** **time**

t1 = time.perf\_counter()

# Directory to paste Statement files

DESTINED\_PATH\_MAIN = os.environ.get('DESTINED\_PATH\_MAIN')

# Directory to paste Info files

DESTINED\_PATH\_INFO = os.environ.get('DESTINED\_PATH\_INFO')

# Directory to fetch unzipped files

GET\_PATH = os.environ.get('GET\_PATH')

# Get customers statement:

dirpaths\_main = []

# Get customers information:

dirpaths\_info = []

# Define pattern to recognize string liked that has 8 digits:

pattern = '\d{8}'

# Finding customer statement files in a destined folder:

**for** dirpath, dirname, filenames **in** os.walk(GET\_PATH):

* # Find out if the direction path has a pattern like '\d{8}'

**if** re.findall(pattern, dirpath) != []:

**for** filename **in** filenames:

# Split elements of a file name:

head = os.path.splitext(filename)[**0**]

tail = os.path.splitext(filename)[**1**]

# Customer statement file has string liked 'SAOKE'

**if** tail == '.zip' **and** re.findall('[A-Z0-9\_]\*SAOKE', head):

path\_in = dirpath + '/' + filename

path\_out = DESTINED\_PATH\_MAIN + head + '\_' + \

re.findall(pattern, dirpath)[**0**] + '.csv'

dirpaths\_main.append((path\_in, path\_out))

# Finding customer information files in a destined folder:

**for** dirpath, dirname, filenames **in** os.walk(GET\_PATH):

**if** re.findall(pattern, dirpath) != []:

**for** filename **in** filenames:

head = os.path.splitext(filename)[**0**]

tail = os.path.splitext(filename)[**1**]

# Customer information file has string liked 'TTKH'

**if** tail == '.zip' **and** re.findall('[A-Z0-9\_]\*TTKH', head):

path\_in = dirpath + '/' + filename

path\_out = DESTINED\_PATH\_INFO + head + '\_' + \

re.findall(pattern, dirpath)[**0**] + '.csv'

dirpaths\_info.append((path\_in, path\_out))

**def** **unzipped\_file**(path):

''' To unzip a gzip-formated file to a destined directory.'''

**with** gzip.open(path[**0**], 'r') **as** f\_in,\

open(path[**1**], 'wb') **as** f\_out:

shutil.copyfileobj(f\_in, f\_out)

# Run concurrent with defined user function unzipped\_file:

**with** concurrent.futures.ThreadPoolExecutor() **as** executor:

executor.map(unzipped\_file, dirpaths\_info)

**with** concurrent.futures.ThreadPoolExecutor() **as** executor:

executor.map(unzipped\_file, dirpaths\_main)

t2 = time.perf\_counter()

**print**(f'It takes {t2 - t1} seconds for completing the task.')

**print**('Unzipping completed.')

# **TÀI LIỆU THAM KHẢO**

1. Ahmad et al (2019), *Customer churn prediction in telecom using machine learning in big data platform*, J Big Data, pages 6:28.

2. Buckinx, W., Baesens, B., den Poel, D., van Kenhove, P., Vanthienen, J. (2002), *Using Machine Learning Techniques to Predict Defection of Top Clients. In: Proc. 3rd International Conference on Data Mining Methods and Databases*, pages 509–517.

3. Buckinx W., den Poel. (2005), *Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting,* European Journal of Operational Research 164, pages 252–268.

4. Kriti (2019), *Customer churn: A study of factors affecting customer churn using machine learning*, Iowa State University.

5. Ribeiro M, Singh S., Guestrin C, *Why Should I Trust You?: Explaining the Predictions of Any Classifier* , arXiv:1602.04938 [cs.LG]

5. Tchaleu B.Y. (2019), *Effective Algorithms to Predict Customer Churn in Financial Services*, University of Johannesburg.

6. Tsai C.F, Lu Y.H (2009), *Customer churn prediction by hybrid neural networks*, Expert Systems with Applications 36, pages 12547–12553.

7. V. Avon (2016), *Machine learning techniques for customer churn prediction in banking environments*, University of Padua.

8. Yaya X. et al (2009), *Customer churn prediction using improved balanced random forests*, Expert Systems with Applications Volume 36, Issue 3, Part 1, pages 5445-5449.

1. Xem Phụ lục 4. [↑](#footnote-ref-1)
2. Xem Phụ lục 1. [↑](#footnote-ref-2)
3. Mã tiền gửi quy định các thuộc tính của khách hàng, loại khách hàng, loại tiền gửi, cách thức trả lãi, kỳ hạn gửi (xem Phụ lục 2). [↑](#footnote-ref-3)
4. Xem chi tiết Phụ lục 3. [↑](#footnote-ref-4)
5. Thước đo khoảng cách Cook (Di) sử dụng đòn bẩy của quan sát thứ i và phần dư của quan sát thứ i, để xác định quan sát nào có ảnh hưởng lớn (theo *Statics for Business and Economics*). Theo kinh nghiệm giá trị D*i* > 1 biểu thị rằng quan sát thứ i có ảnh hưởng lớn và nên được nghiên cứu sâu hơn. [↑](#footnote-ref-5)
6. Xem 2.1.3. [↑](#footnote-ref-6)