

**TRƯỜNG ĐẠI HỌC NGOẠI THƯƠNG**

**TRUNG TÂM DATAPOT & MB BANK**

**⯎⯎⯎⯎⯎⯎⯎**



**PROJECT**

**CHALLENGE 1: Profile customers of the bank and predict who**

**are likely to churn. As part of the requirements, students**

**have to define “What is a churn customer” by themselves.**

**Hà Nội, tháng 07 năm 2022**

**MỤC LỤC**

[LỜI MỞ ĐẦU 1](#_Toc109155775)

[CHƯƠNG 1 CƠ SỞ LÝ THUYẾT 2](#_Toc109155776)

[1.1 Cơ sở lý thuyết về các thuật toán được sử dụng trong bài 2](#_Toc109155777)

[1.1.1 Logistic Regression 2](#_Toc109155778)

[1.1.2 K-nearest Neighbors 2](#_Toc109155779)

[1.1.3 Decision Tree 3](#_Toc109155780)

[1.1.4 XGBoost 3](#_Toc109155781)

[1.1.5 Random Forest 3](#_Toc109155782)

[1.2 Tổng quan nghiên cứu về “khách hàng từ bỏ” 3](#_Toc109155783)

[1.3 Phương pháp nghiên cứu 4](#_Toc109155784)

[1.3.1 Dữ liệu 4](#_Toc109155785)

[1.3.2 Chỉ số đánh giá model 6](#_Toc109155786)

[CHƯƠNG 2 QUÁ TRÌNH NGHIÊN CỨU 7](#_Toc109155787)

[2.1 Load library and data 7](#_Toc109155788)

[2.2 Clean & Transform data 8](#_Toc109155789)

[2.2.1 Clean missing values 8](#_Toc109155790)

[2.2.2 Transform data 10](#_Toc109155791)

[2.3 EDA 11](#_Toc109155792)

[2.3.1 Demographics 11](#_Toc109155793)

[2.3.2 Customer’s Transaction 17](#_Toc109155794)

[2.4 Xây dựng và đánh giá mô hình 25](#_Toc109155795)

[2.4.1 Logistic Regression 25](#_Toc109155796)

[2.4.2 K-nearest Neighbors 26](#_Toc109155797)

[2.4.3 Decision Tree 27](#_Toc109155798)

[2.4.4 XGBoost 28](#_Toc109155799)

[2.4.5 Random Forest 29](#_Toc109155800)

[2.5 Phát triển và tối ưu hóa mô hình 29](#_Toc109155801)

[2.5.1 Mô hình XGBoost 30](#_Toc109155802)

[2.5.2 Mô hình Random Forest 31](#_Toc109155803)

[2.6 Kiểm tra mô hình trên test set 31](#_Toc109155804)

[2.6.1 Mô hình XGBoost 32](#_Toc109155805)

[2.6.2 Mô hình Random Forest 32](#_Toc109155806)

[2.7 Đánh giá chung 33](#_Toc109155807)

[CHƯƠNG 3 KẾT QUẢ NGHIÊN CỨU VÀ KIẾN NGHỊ 35](#_Toc109155808)

[3.1 Kết luận chung 35](#_Toc109155809)

[3.2 Kết quả nghiên cứu và kiến nghị 35](#_Toc109155810)

[LỜI KẾT 37](#_Toc109155811)

[TÀI LIỆU THAM KHẢO 38](#_Toc109155812)

# LỜI MỞ ĐẦU

Vấn đề về tỷ lệ khách hàng rời bỏ, giữ chân khách hàng từ trước đến nay vẫn luôn là một trong những vấn đề quan trọng và nhận được nhiều sự chú ý của doanh nghiệp. Điều này vô cùng quan trọng khi xây dựng giá trị lâu dài của doanh nghiệp. Tỷ lệ rời bỏ lớn có thể gây ra ảnh hưởng xấu đến tăng trưởng và lợi nhuận của doanh nghiệp. Doanh nghiệp cần xác định được những nhóm khách hàng rời bỏ để có thể đưa ra những giải pháp phù hợp với từng tệp khách hàng để giữ chân khách hàng và thuyết phục họ ở lại. Phân tích sự rời bỏ của khách hàng thông qua những dữ liệu có sẵn có thể góp phần giúp doanh nghiệp có được cái nhìn khái quát và chính xác hơn về tình hình rời bỏ của khách hàng và có được những dự đoán cũng như những giải pháp phù hợp để giữ chân họ.

Nghiên cứu này sẽ trình bày một phân tích về sự biến động của khách hàng trong ngân hàng MB - Ngân hàng quân đội Việt Nam. Bài nghiên cứu bao gồm hai mục tiêu lớn. Mục tiêu đầu tiên là xác định ai là những khách hàng đang rời bỏ ngân hàng MB. Mục tiêu thứ hai là dự báo về khách hàng rời bỏ dựa trên 5 thuật toán, từ đó lựa chọn ra thuật toán phù hợp nhất và đưa ra các giải pháp xây dựng cho tương lai.

Ngoài Lời mở đầu, Lời kết và Tài liệu tham khảo, bài nghiên cứu bao gồm 3 chương. Chương 1 nêu ra cơ sở lý thuyết về các mô hình được sử dụng cùng tổng quan nghiên cứu đi trước và phương pháp nghiên cứu trong bài. Chương 2 trình bày chi tiết quá trình nghiên cứu của nhóm để đưa ra các insight có giá trị. Từ đó, chương 3 tổng kết lại các kết quả và đưa ra các giải pháp tương ứng cho doanh nghiệp. Phần phân tích của bài báo cáo này được tiến hành bằng công cụ lập trình Python - một ngôn ngữ lập trình phổ biến dành cho Data Science.

# CƠ SỞ LÝ THUYẾT

## Cơ sở lý thuyết về các thuật toán được sử dụng trong bài

### Logistic Regression

Mô hình Logistic Regression là một trong những mô hình cơ bản trong lớp bài toán classification. Ưu điểm của mô hình này là thuật toán đơn giản, dễ hiểu, dễ sử dụng và có thể áp dụng cho gần như mọi bài toán classification 2 lớp giá trị, thậm chí là nhiều lớp giá trị. Chính vì vậy, mỗi khi gặp các bài toán về phân lớp giá trị, chúng ta luôn nghĩ về mô hình Logistic Regression đầu tiên. Ngược lại, chính vì thuật toán đơn giản và dễ sử dụng, nên mô hình này thường cho ra những dự đoán với kết quả không quá chính xác.

Để cải thiện việc dự đoán, từ mô hình Logistic Regression làm cơ sở, người ta đã xây dựng các mô hình Machine Learning với thuật toán cao cấp, phức tạp hơn và Neural Network trong Deep Learning. Tuy nhiên, mô hình Logistic Regression vẫn có thể dùng được, thậm chí cho ra kết quả dự đoán với độ chính xác cao, nếu ta có thể tối ưu hóa, chọn lựa đúng các trường dữ liệu đầu vào quan trọng thông qua quá trình phân tích và vận dụng Domain Knowledge.

### K-nearest Neighbors

K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này *không học* một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning)), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. Thay vì dự đoán ra giá trị label, nó sẽ tìm những điểm dữ liệu gần nhất với điểm dữ liệu cần dự đoán và đưa ra kết quả dựa vào cách voting giữa các điểm tìm được.

K- nearest neighbor có thể áp dụng được vào cả 2 loại bài toán là [Classification](https://machinelearningcoban.com/2016/12/27/categories/#classification-phan-loai) và [Regression](https://machinelearningcoban.com/2016/12/27/categories/#regression-hoi-quy). KNN còn được gọi là một thuật toán [Instance-based hay](https://en.wikipedia.org/wiki/Instance-based_learning) [Memory-based learning.](https://en.wikipedia.org/wiki/Instance-based_learning)

### Decision Tree

Mô hình Decision Tree cũng là một mô hình được sử dụng khá phổ biến và hiệu quả trong cả 2 lớp bài toán phân loại và dự báo. Khác với những thuật toán khác, mô hình Decision Tree không tồn tại phương trình dự báo. Thay vào đó, thuật toán này sẽ đưa các điểm gia trị về đúng lớp label của nó, bằng cách phân loại chúng dựa vào các trường dữ liệu đầu vào.

### XGBoost

Mô hình XGBoostlà một mô hình thuộc nhóm Ensemble Methode (Học kết hợp), có nghĩa là thay vì xây dựng một mô hình đơn giản duy nhất để dự báo kết quả label, thuật toán sẽ xây dựng các mô hình học yếu (weak learner), từ đó kết hợp chúng lại thành một mô hình vượt trội nhất.

Mô hình XGBoost sẽ được tiến hành dựa trên phương pháp boosting, nghĩa là các mô hình weak learner sẽ được update liên tục trong quá trình học để hạn chế những dự đoán sai, từ đó cho ra mô hình hoàn thiện cuối cùng. XGBoost là một thuật toán rất mạnh, cho ra dự đoán có độ chính xác rất cao.

### Random Forest

Random Forests là thuật toán Supervised Learning. Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng, không yêu cầu cao về các biến đầu vào. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Thật vậy, Random forests tạo ra các Decision Tree nhỏ trên các mẫu dữ liệu và trường dữ liệu được chọn ngẫu nhiên, mỗi cây sẽ cho ra dự đoán của mình để từ đó tổng hợp, voting để cho ra dự đoán cuối cùng. Thuật toán cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của các biến đầu vào. Random forest là một thuật toán khá mạnh, và được ứng dụng nhiều trong thực tế.

## Tổng quan nghiên cứu về “khách hàng từ bỏ”

Các nghiên cứu đi trước đã cho thấy có nhiều định nghĩa khác nhau về “khách hàng từ bỏ”. Trong nghiên cứu của (Lu, 2002), khách hàng từ bỏ là những người chuyển sang loại hình dịch vụ khác. Trong nghiên cứu (Neslin et al., 2006), khách hàng từ bỏ là những người không sử dụng dịch vụ trong một khoảng thời gian ấn định.

Trong các bài nghiên cứu về lĩnh vực ngân hàng cũng có nhiều định nghĩa khách hàng từ bỏ khác nhau. (Glady et al., 2008) cho rằng khách hàng từ bỏ là những người khối lượng giao dịch nhỏ hơn một ngưỡng nhất định, cụ thể, họ là những người có tài sản trong ngân hàng nhỏ hơn 2500 Euro (bao gồm tiền tiết kiệm, chứng khoán và các loại sản phẩm tài chính khác).

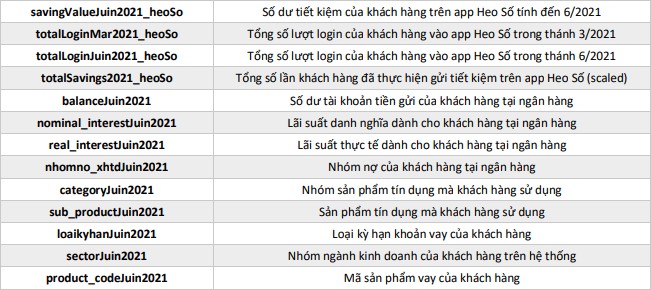
Van der Poel and Larivière (2004) định nghĩa những khách hàng này là những người đã đóng tài khoản. Guangli Nie và cộng sự (2011) định nghĩa khách hàng từ bỏ là những người không thực hiện giao dịch qua ngân hàng đó trong một khoảng thời gian ấn định. Kết hợp định nghĩa trong nghiên cứu của Guangli Nie (2011), nhóm định nghĩa khách hàng rời bỏ là khách hàng không có phát sinh giao dịch nào (tổng số giao dịch bằng 0) trong tháng 3 và tháng 6 năm 2021.

## Phương pháp nghiên cứu

### Dữ liệu

Bộ dữ liệu được lấy từ cơ sở dữ liệu thực tế của ngân hàng thương mại MB tại Việt Nam, và chứa thông tin về lịch sử thanh toán, giao dịch, các yếu tố nhân khẩu học và sở thích của khách hàng từ năm 2020 đến năm 2021. Dưới đây là bảng mô tả biến:





### Chỉ số đánh giá model

Khi xây dựng một mô hình Machine Learning, chúng ta cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh khả năng của các mô hình.

Đối với bài toán Classification 2 giá trị, đặc biệt là khi bộ dữ liệu bị mất cân bằng (giá trị đích 'churn' chiếm thiểu số), ta đánh giá mô hình dựa trên các metrics ưu tiên là: Recall, F1score và AUC ROC.

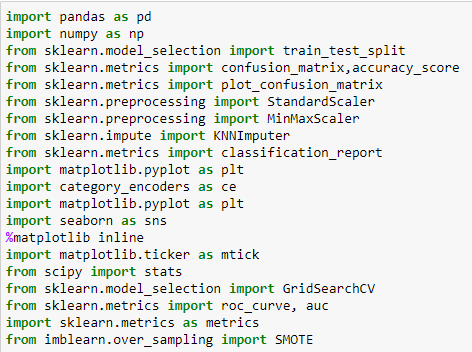
Recall được ưu tiên hơn Precision, bởi ta mong muốn tìm được các điểm giá trị Positive thực tế nhiều nhất có thể để có thể đề ra các biện pháp xử lý kịp thời nhằm giữ chân khách hàng, và hạn chế các điểm False Negative (khách hàng churn nhưng lại dự đoán là không churn). F1 Score được sử dụng để cân bằng Recall và Precision.

# QUÁ TRÌNH NGHIÊN CỨU

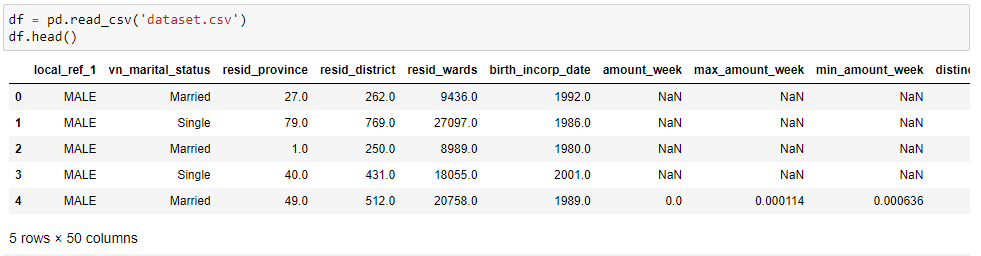
Dưới đây là kết quả quá trình nghiên cứu của nhóm để đạt được mục tiêu nghiên cứu, sử dụng ngôn ngữ Python. Toàn bộ quá trình cụ thể và chi tiết sẽ được nhóm trình bày trong file .ipynb đính kèm cùng report này.

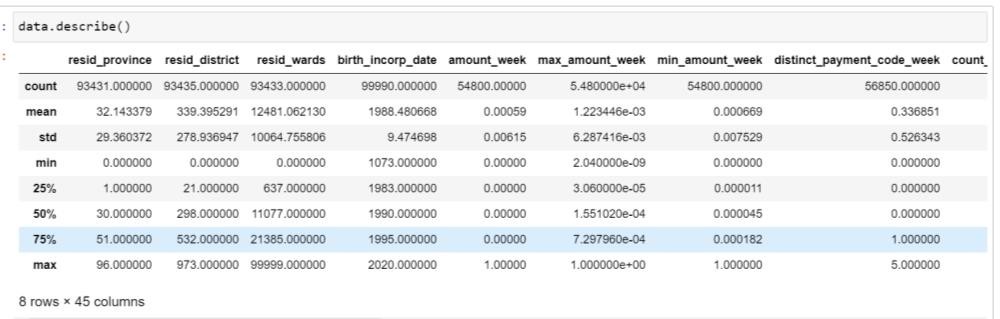
## Load library and data

Dữ liệu được lưu ở file dataset.csv. Pandas là thư viện được sử dụng trong việc load và transform dữ liệu.



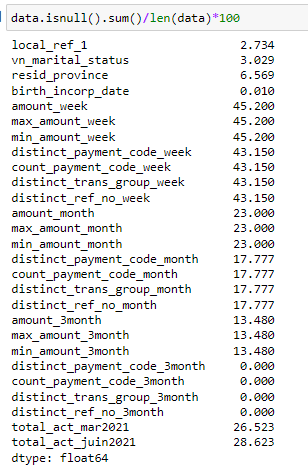
Xem nhanh qua bộ dữ liệu, ta có thể thấy nhanh các trường thông tin cơ bản như giới tính, tình trạng hôn nhân, tỉnh thành sinh sống, tổng số tiền giao dịch của khách hàng trong 1 tuần trước đó …., đồng thời ta cũng thấy ngay bộ liệu đã được normalize trong khoảng [0,1] trước đó đối với những trường liên quan đến số tiền. Các trường Tỉnh thành khách hàng sinh sống, Quận/huyện khách hàng sinh sống, Phường/xã khách hàng sinh sống cũng được thay thế tên các địa điểm thành các số nguyên. Nhìn chung, 50 biến được thu thập từ cơ sở dữ liệu khách hàng. Các biến này có liên quan đến các chủ đề như sau: (1) giao dịch tài khoản, (2) chỉ báo dịch vụ, (4) thông tin hồ sơ cá nhân, và (5) thông tin app Heo số, (6) vay tín dụng. Cơ sở dữ liệu tổng thể có 100 000 khách hàng.

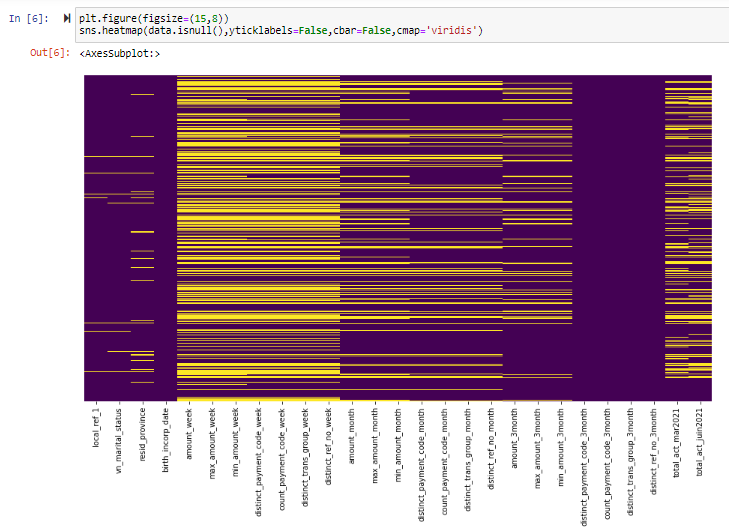


Ta xem qua thống kê cơ bản của bộ dữ liệu:

## Clean & Transform data

### Clean missing values

******

******

Sau khi tính toán tỷ lệ missing values trong từng trường dữ liệu và nhìn vào biểu đồ, ta nhận thấy rằng:

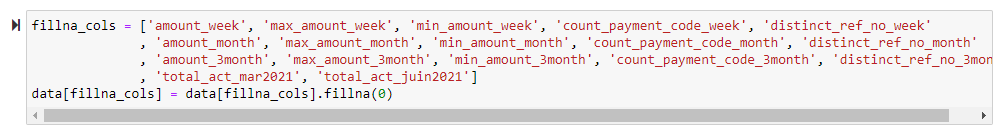
Đối với các trường dữ liệu về giao dịch của khách hàng, các missing values có xu hướng tăng dần theo thời gian. Cụ thể:

* Vào thời điểm 3 tháng trước, chỉ có 13.48% khách hàng không có dữ liệu về số tiền giao dịch, trong khi dữ liệu về loại dịch vụ và nhóm giao dịch là không có missing values nào.
* Vào thời điểm 1 tháng trước, 23% khách hàng không có dữ liệu về số tiền giao dịch, và 17.77% khách hàng không có dữ liệu về loại dịch vụ và nhóm giao dịch.
* Vào thời điểm 1 tuần trước, tỷ lệ khách hàng không có dữ liệu về số tiền giao dịch lên tới 45.2%, và tỷ lệ khách hàng không có dữ liệu về loại dịch vụ và nhóm giao dịch là 43.15%

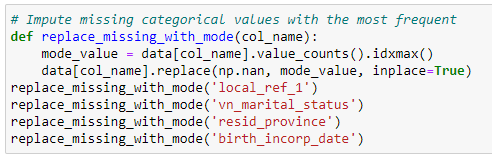
Từ những quan sát trên, ta rút ra những kết luận rằng:

* Một số khách hàng không có ghi nhận dữ liệu về số tiền giao dịch, những vẫn còn dữ liệu về loại dịch vụ và nhóm giao dịch. Chứng tỏ rằng họ tạm thời không phát sinh giao dịch, tuy nhiên vẫn sử dụng dịch vụ của ngân hàng.
* Tỷ lệ khách hàng không có ghi nhận dữ liệu về số tiền giao dịch và loại dịch vụ, nhóm giao dịch tăng lên theo thời gian cho thấy việc khách hàng đang tạm dừng/rời bỏ sử dụng dịch vụ của ngân hàng.

Vì vậy, các missing values trong nhóm trường dữ liệu về giao dịch của khách hàng có rất nhiều ý nghĩa để xác định đặc trưng mẫu dữ liệu và xây dựng mô hình dự đoán khả năng rời bỏ dịch vụ của khách hàng. Do đó, ta quyết định không drop những trường dữ liệu này hoặc impute các missing values bằng mean, median, mode mà thay thế chúng bằng giá trị 0 (với ý nghĩa là giao dịch = 0), nhằm giữ lại đặc trưng của mẫu mặc dù điều này có thể gây ra sự mất cân bằng dữ liệu.



Đối với các trường dữ liệu về nhân khẩu học, ta tiến hành impute missing values bằng các giá trị có tần suất xuất hiện nhiều nhất.



### Transform data

Đầu tiên, ta đưa trường dữ liệu 'birth\_incorp\_date' về dạng 'age' để thuận tiên hóa phân tích và áp dụng vào mô hình dự đoán.



Sau đó, ta tính tổng số lượng dịch vụ sử dụng và số lượng đối tác từ các giao dịch của khách hàng trong khoảng thời gian quan sát để tránh đa cộng tuyến và đơn giản hóa việc phân tích.



Tiếp theo, ta tính tổng số giao dịch của khách hàng trong tháng 3 và tháng 6 năm 2021.



Cuối cùng, ta tiến hành xác định biến Y label “Churn”

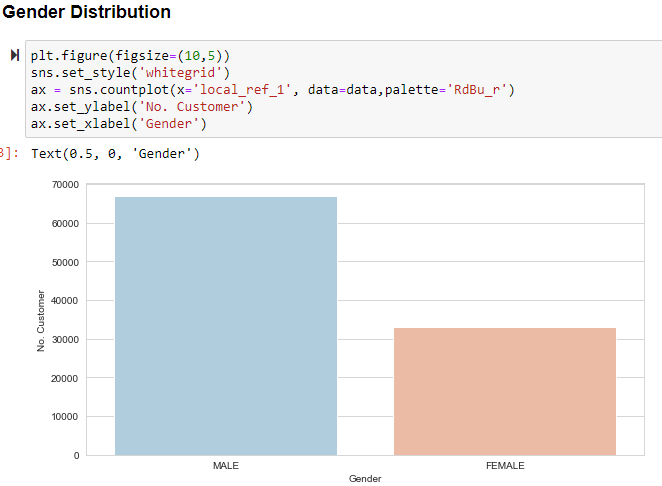


## EDA

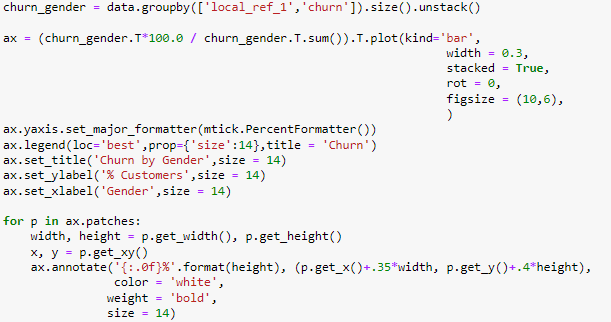
### Demographics

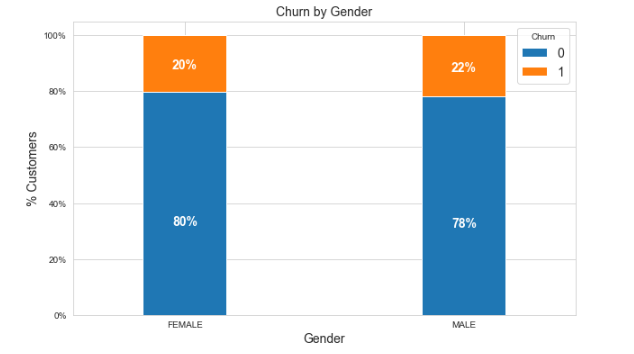
#### Phân bổ theo giới tính

Đầu tiên, ta xem xét tỉ lệ khách hàng giữa các giới tính.



Từ biểu đồ phía trên ta có thể thấy rõ do số lượng khách hàng Nam gấp đôi số lượng khách hàng Nữ. Vì biến này chỉ có 2 giá trị duy nhất là Nam và Nữ, nên ta đưa biến giới tính từ dạng rời rạc về dạng liên tục bằng get\_dummies.



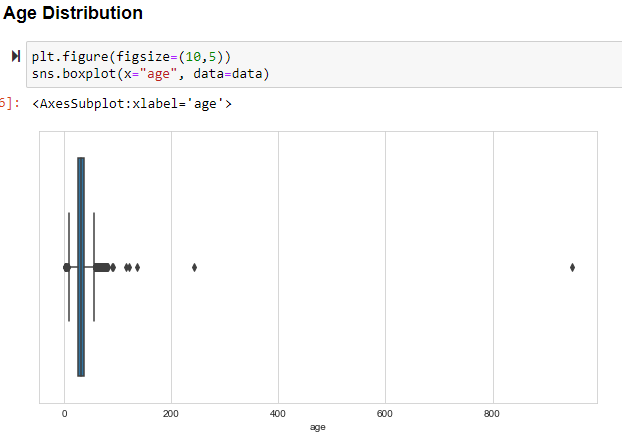




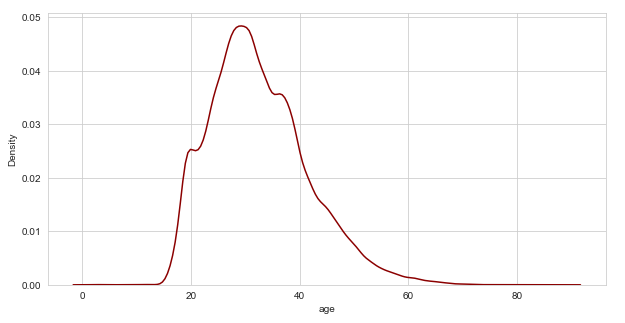
Mặc dù có sự chênh lệnh rõ rệt giữa nhóm khách hàng theo giới tính, tuy nhiên điều này ảnh hưởng không nhiều tới tỷ lệ Churn giữa 2 nhóm khách hàng này.

#### Phân bổ theo độ tuổi

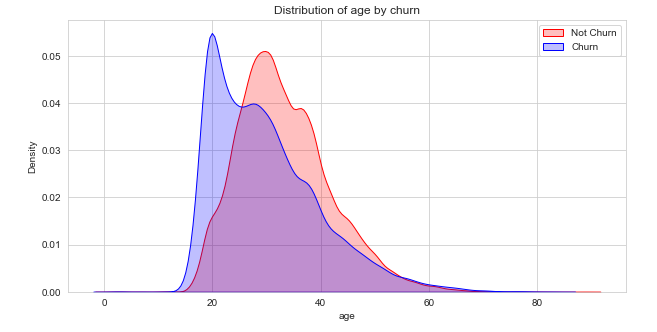
Phân bố khách hàng theo độ tuổi như sau:



Trong trường dữ liệu về độ tuổi khách hàng có xuất hiện những outliers không hợp lý so với nhân khẩu học người Việt Nam (tuổi > 200 và > 800), cho nên ta tiến hành xử lý các outliers này bằng cách thay chúng bằng median. Sau khi xử lý các outliers, độ tuổi khách hàng quay về phân bố tương đối chuẩn, có thể áp dụng vào mô hình.

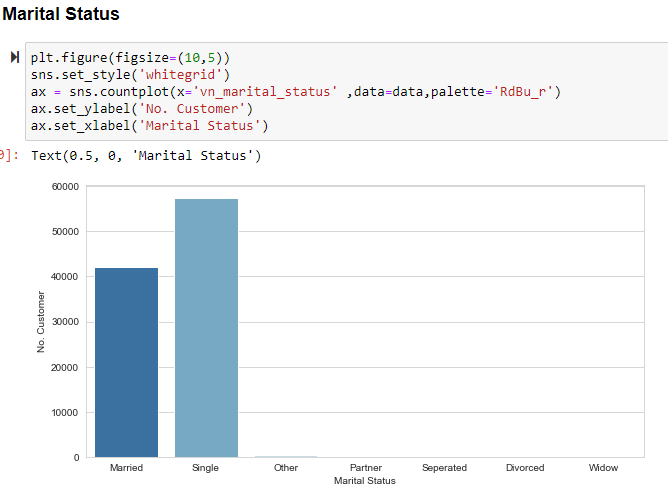


Ta thấy được độ tuổi khách hàng càng cao, khả năng Churn giảm xuống. Đặc biệt, nhóm khách hàng Churn tập trung chủ yếu vào độ tuổi quanh 20. Đây là những khách hàng thuộc tập học sinh/sinh viên, có tài chính thấp và nhạy cảm về biểu giá các loại dịch vụ và ưa thích thay đổi giữa các dịch vụ ngân hàng.

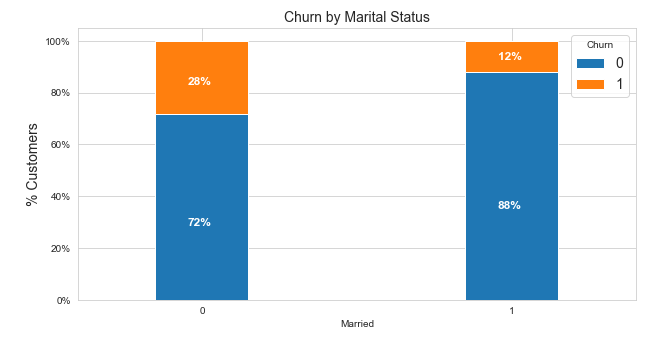


#### Phân bổ theo tình trạng hôn nhân

Tiếp theo, ta xem xét số lượng khách hàng phân bổ theo tình trạng hôn nhân



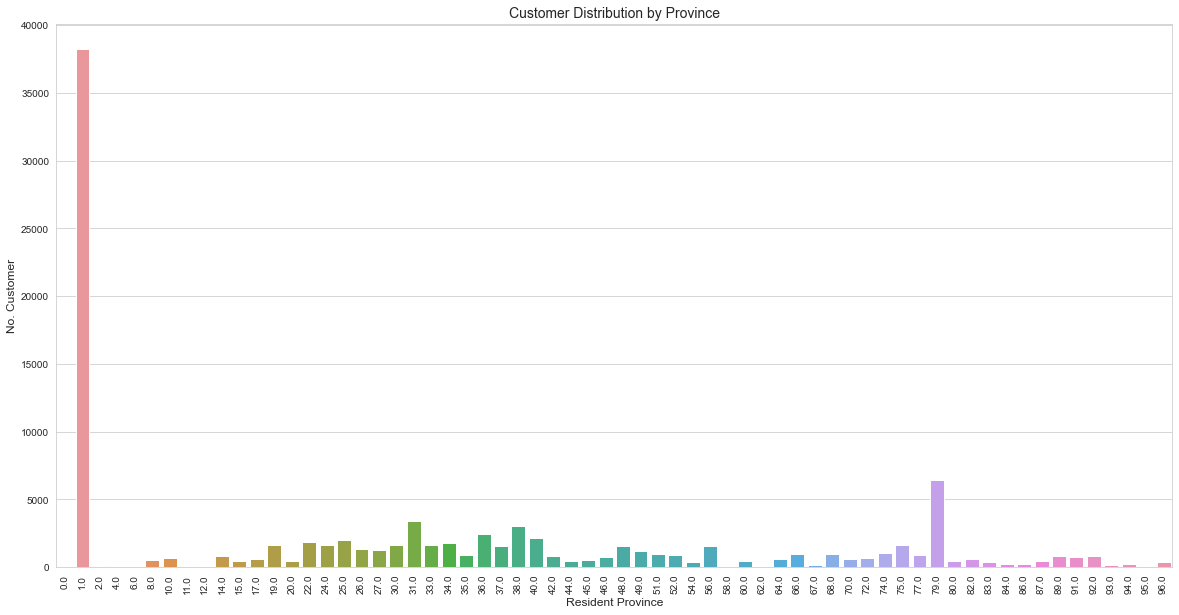
Trong bộ dữ liệu, gần như toàn bộ khách hàng có tình trạng hôn nhân là Married hoặc Single, nên ta sẽ tập trung chủ yếu vào 2 tập khách hàng này và bỏ qua những tập còn lại. Ta tiến hành get\_dummies để đưa về dạng biến liên tục 'Married'.



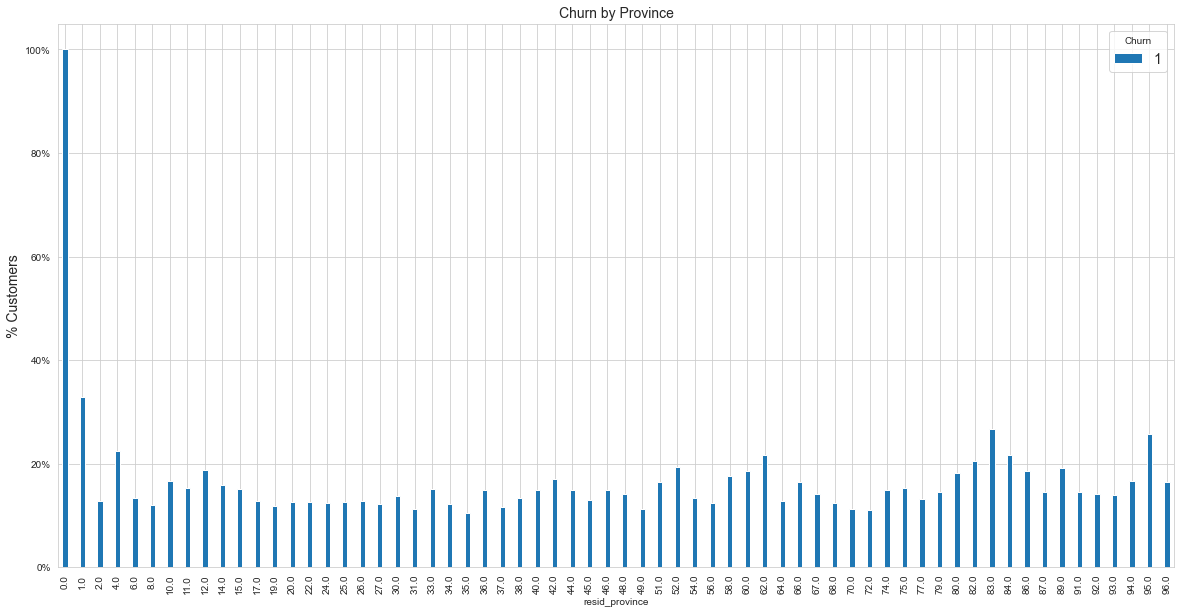
Từ biểu đồ trên, ta có thể thấy được khách hàng đã kết hôn có ít khả năng rời bỏ đi hơn những người độc thân hay những người gặp vấn đề với hôn nhân. Có thể những người đã kết hôn sẽ nghiêm túc và quan tâm đến tài chính cá nhân hơn những người độc thân.

#### Phân bổ theo tỉnh thành

Số lượng khách hàng phân bổ theo tỉnh thành được thể hiện qua biểu đồ dưới đây:



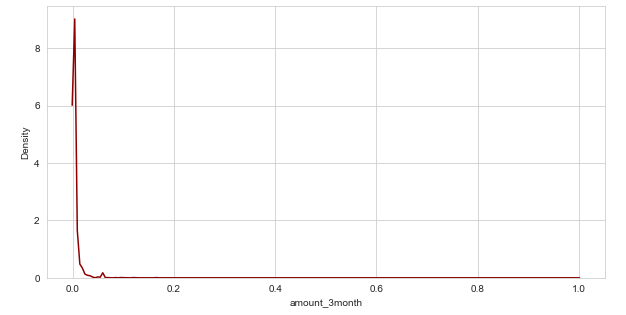
Tỉnh 1 là nơi tập trung lượng khách hàng đông nhất, chiếm gần 40% tổng số khách hàng được ghi nhận trong bộ dữ liệu.

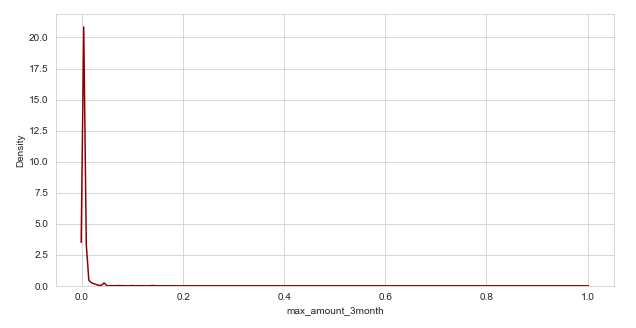


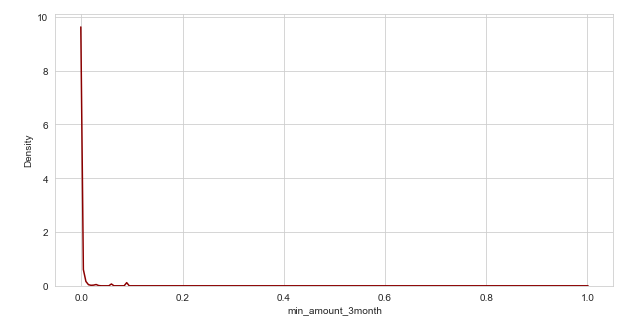
Ta thấy rằng tỷ lệ khách hàng churn giữa các tỉnh thành không có nhiều khác biệt, dao động trong khoảng 15% - 22%. Đối với tỉnh 1 là nơi có tập trung khách hàng lớn nhất (~ 40% tổng số khách hàng), thì tỷ lệ khách hàng churn cũng là lớn nhất, gấp đôi mức tỷ lệ churn theo tỉnh thành trung bình (tỷ lệ churn trung bình theo tỉnh thành là khoảng 16,8%). Các tỉnh có tỷ lệ churn cao vượt trội tiếp theo là tỉnh 0, 83, 95. Tuy nhiên các tỉnh này lại có lượng khách hàng rất ít. Cá biệt là tỉnh 0 có 100% tỷ lên churn, nhưng điều này xảy ra là do ở tình thành này chỉ có 1 khách hàng duy nhất và là khách hàng churn.

### Customer’s Transaction

#### Transaction Amount

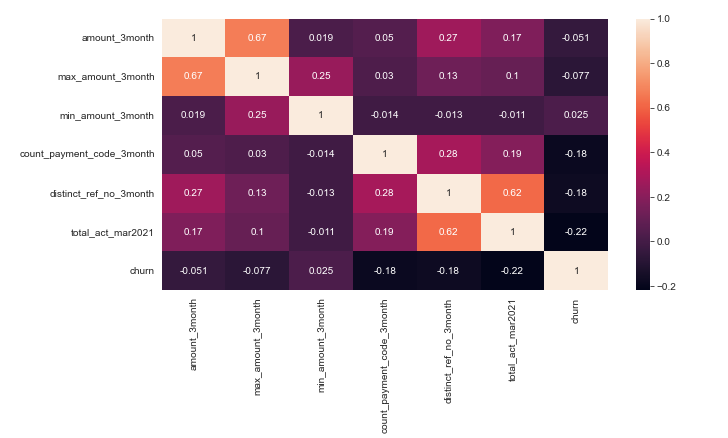




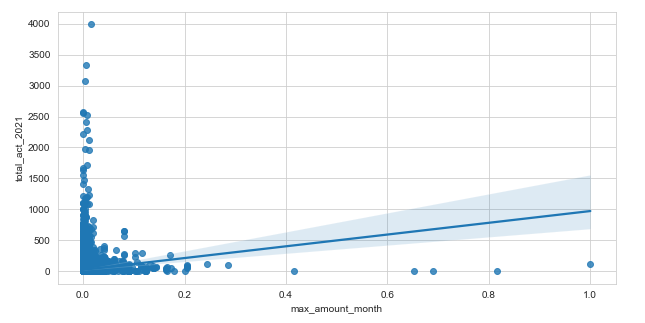


Từ 3 biểu đồ trên, có thể thấy các trường dữ liệu về số tiền giao dịch của khách hàng nằm trong miền giá trị [0; 1], cho thấy các trường này đã được normalize MinMax.

Ta có heatmap tương quan sau:



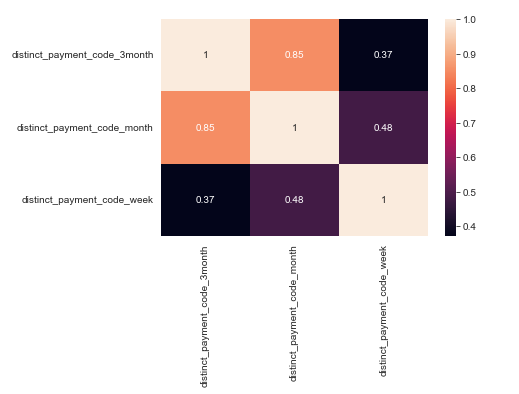
Từ heatmap, có thể thấy tổng số tiền giao dịch và số tiền giao dịch lớn nhất trong 1 khoảng thời gian có độ tương quan cao (số tiền giao dịch lớn nhất càng lớn thì tổng số tiền giao dịch càng lớn). Vì vậy, để tránh đa cộng tuyến, ta chỉ giữ lại biến 'max\_amount'.



Từ biểu đồ Simple Linear Regression, ta thấy có sự tương quan dương giữa số tiền giao dịch lớn nhất trong khoảng thời gian quan sát với tổng số giao dịch thực hiện trong khoảng thời gian target là T3 và T6/2021. Điều đó cho thấy những khách hàng có hành vi giao dịch với số tiền lớn sẽ ít có khả năng churn hơn.

#### Phân bổ theo Payment Code

Với biến ‘Payment code’, ta có heatmap tương quan như sau:



Có thể thấy các biến 'distinct\_payment\_code\_3month', 'distinct\_payment\_code\_month', 'distinct\_payment\_code\_week' có tương quan lớn với nhau. Giải thích bằng việc khách hàng có tiếp tục sử dụng dịch vụ mà mình đã sử dụng trong thời gian trước hay không. Giữ lại 'distinct\_payment\_code\_3month' và thay các variables 'distinct\_payment\_code\_month', 'distinct\_payment\_code\_week' thành 'continue\_use\_month' và 'continue\_use\_week' là các binary varibles trả về (0: dừng sử dụng dịch vụ) và (1: tiếp tục sử dụng).

Ta có tỷ lệ phân bổ khách hàng theo Payment Code như sau:

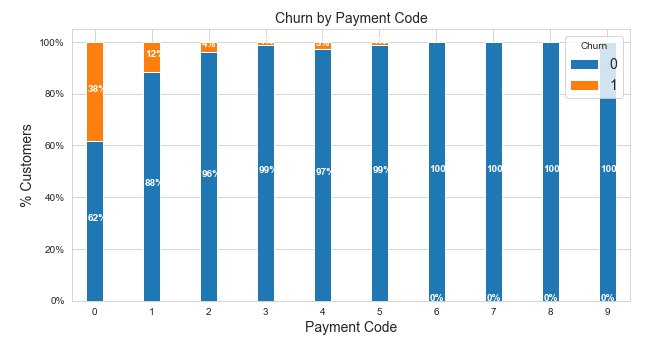


Số lượng khách theo Payment Code



Tần suất giao dịch theo Payment Code

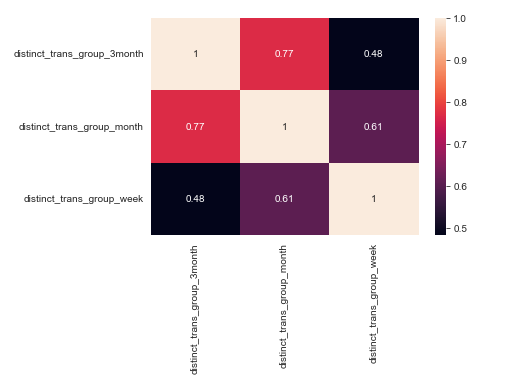
Ta thấy rằng số lượng khách hàng phân bố không đều giữa các loại dịch vụ (lệch phải), tập trung lớn nhất vào loại 0 và 1. Tuy nhiên, tần suất giao dịch trung bình lại phân bố ngược lại.



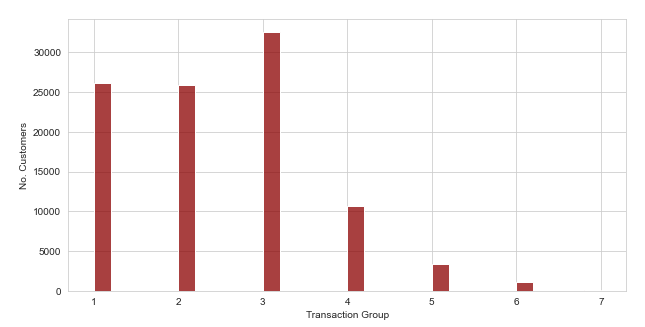
Như đã xem xét ở trên, mặc dù số khách hàng sử dụng dịch vụ 0, 1, 2 là rất lớn, tuy nhiên tần suất giao dịch thấp, và họ cũng là những khách hàng có tỷ lệ churn cao.

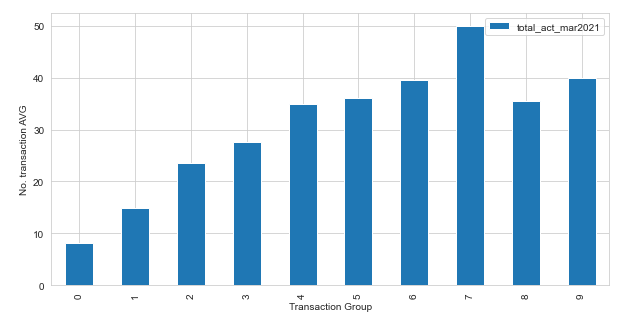
#### Phân bổ theo Transaction Group

Với biến ‘Transaction Group’:

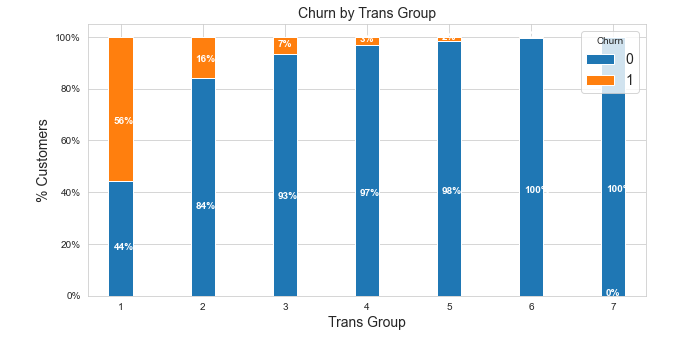


Các biến 'distinct\_trans\_group\_3month', 'distinct\_trans\_group\_\_month', 'distinct\_trans\_group\_week' có tương quan lớn với nhau. Giải thích bằng việc khách hàng có tiếp tục sử dụng nhóm giao dịch mà mình đã sử dụng trong thời gian trước hay không. Tương tự như với Payment Code, ta giữ lại 'distinct\_trans\_group\_3month' và thay các variables 'distinct\_trans\_group\_month', 'distinct\_trans\_group\_week' thành 'continue\_trans\_month' và 'continue\_trans\_week' là các binary varibles trả về (0: dừng sử dụng nhóm giao dịch) và (1: tiếp tục sử dụng).



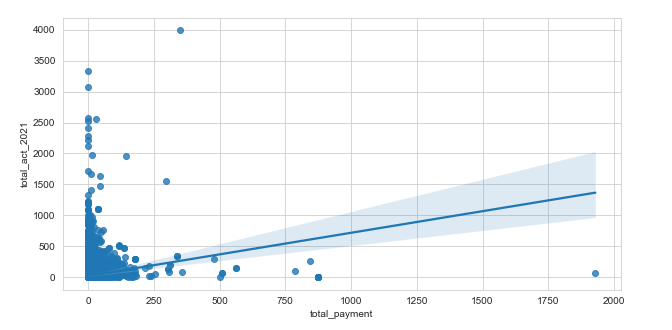


Ta thấy rằng số lượng khách hàng phân bố không đều giữa các nhóm giao dịch (lệnh phải), tập trung lớn nhất vào nhóm 1, 2, 3. Tuy nhiên, tần suất giao dịch trung bình lại phân bố ngược lại. Sự phân bố ngược lại này khá giống với số lượng khách hàng và tấn suất giao dịch chia theo Payment Code.



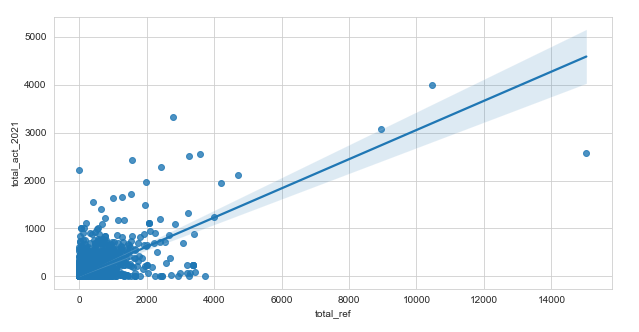
Như đã xem xét ở trên, mặc dù số khách hàng sử dụng nhóm giao dịch 1, 2, 3 là rất lớn, tuy nhiên tần suất giao dịch thấp, và họ cũng là những khách hàng có tỷ lệ churn cao.

Với biến ‘Number of services’:

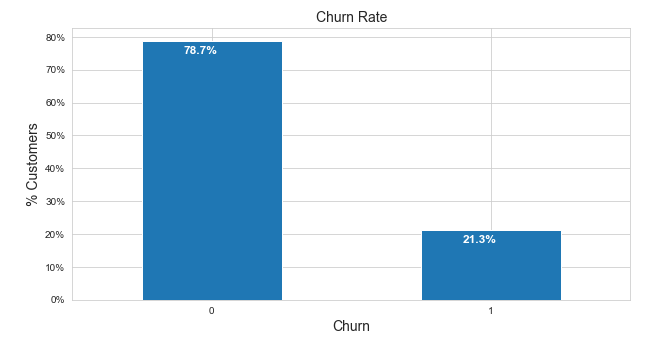


Từ biểu đồ Simple Linear Regression, ta thấy có sự tương quan dương giữa tổng số dịch vụ sử dụng trong khoảng thời gian quan sát với tổng số giao dịch thực hiện trong khoảng thời gian target là tháng 3 và tháng 6 năm 2021. Điều đó cho thấy những khách hàng sử dụng nhiều loại dịch vụ sẽ ít có khả năng churn hơn.

Với biến ‘Number of Partners’:



Từ biểu đồ Simple Linear Regression, ta thấy có sự tương quan dương giữa tổng số đối tác giao dịch của khách hàng trong khoảng thời gian quan sát với tổng số giao dịch thực hiện trong khoảng thời gian target là tháng 3 và tháng 6 năm 2021. Điều đó cho thấy những khách hàng thường xuyên giao dịch với nhiều đối tác khách nhau sẽ ít có khả năng churn hơn.



78,7% không rời bỏ dịch vụ của ngân hàng, có thể thấy rõ ràng rằng bộ dữ liệu này đang imbalance. Điều này có thể dẫn đến những sai lệch trong mô hình dự đoán sau này, nhất là trong việc dự đoán giá trị thiểu số. Tuy nhiên, điều này lại là tín hiệu tốt đối với hoạt động của doanh nghiệp bởi tỷ lệ khách hàng trung thành là cao.

## Xây dựng trên Train set

Bài nghiên cứu sử dụng 5 loại Model là: Logistic Regression, Decision Tree, Random Forest, XGBoost và K-Nearest Neighbors để thử nghiệm và chọn ra Model đem lại kết quả cao nhất.

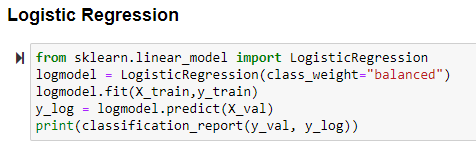
Trong đó:

* Biến Y là biến Churn với định nghĩa khách hàng rời bỏ (“Churn”) là những người không có giao dịch nào trong tháng 3/2021 và tháng 6/2021.
* Biến X bao gồm: col\_0, col\_1, col\_2, col\_3, col\_4, col\_5, col\_6, col\_7, max\_amount\_week, min\_amount\_week, max\_amount\_month, min\_amount\_month, max\_amount\_3month, min\_amount\_3month, age, total\_payment, total\_ref, churn, MALE, Married, continue\_use\_month, continue\_use\_week, pay\_0, pay\_1, pay\_2, pay\_3, pay\_4, continue\_trans\_month, continue\_trans\_week, trans\_1, trans\_2, trans\_3, trans\_4.
* Chia bộ dữ liệu thành 3 tập: train/validation/test set với training size là 70%, validation size là 20% và test size là 10%.
* Trước khi xây dựng mô hình, nhóm tiến hành Normalize train set bằng MinMaxScaler để đưa tất cả các trường dữ liệu về miền [0; 1].
* Để xử lý vấn đề mất cân bằng dữ liệu, nhóm tiến hành Oversampling train set bằng SMOTE (dựa trên thuật toán K-nearest Neighbors, tạo ra các điểm dữ liệu ảo nhãn thiểu số để cân bằng với nhãn đa số).

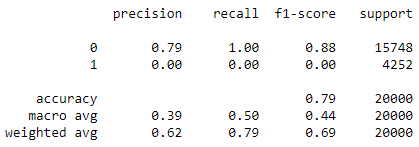
Nhóm đánh giá mô hình dựa trên các chỉ số AUC-ROC và Recall/F1-score

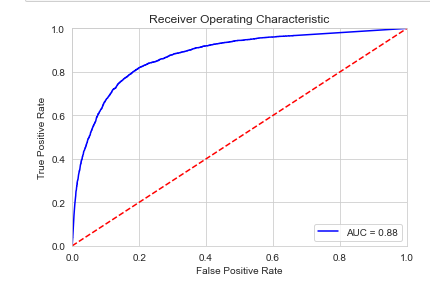
**Xây dựng mô hình trên Train set và kết quả nhận về như sau:**

### Logistic Regression

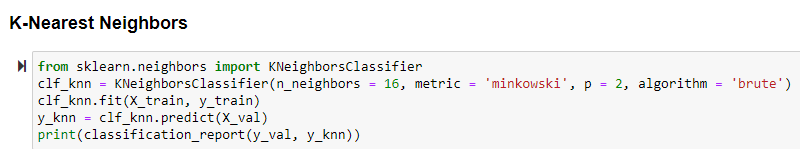
**

Kết quả các mô hình Logistic Regression như sau:

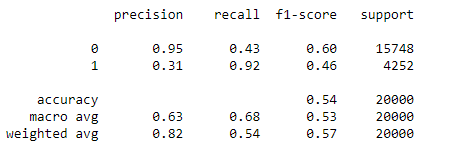


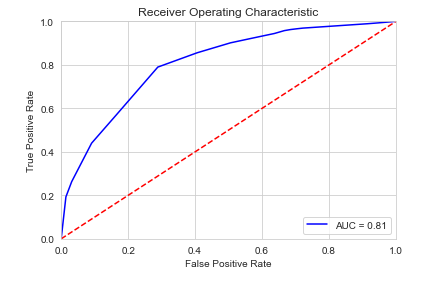


### K-nearest Neighbors

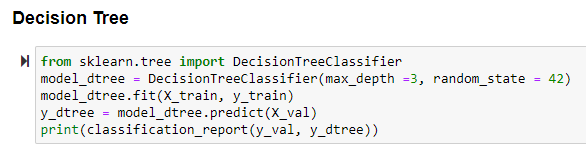


Kết quả nhận về như sau:

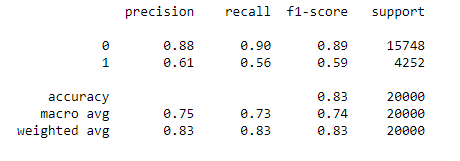


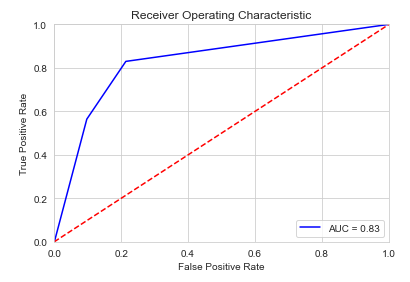


### Decision Tree

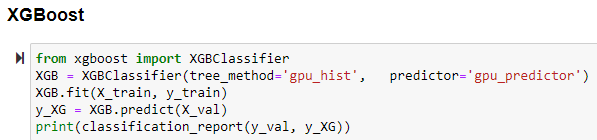


Kết quả nhận về như sau:

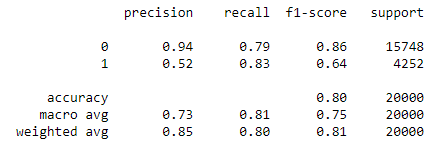


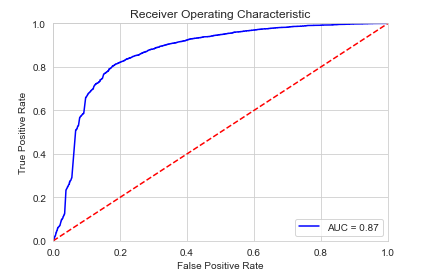


### XGBoost

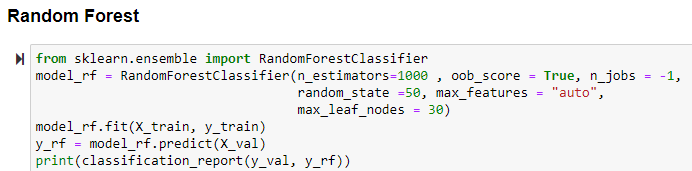
**

Kết quả nhận về như sau:

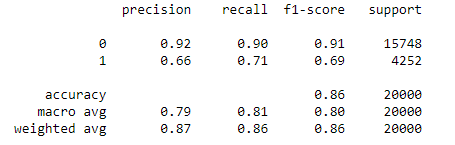


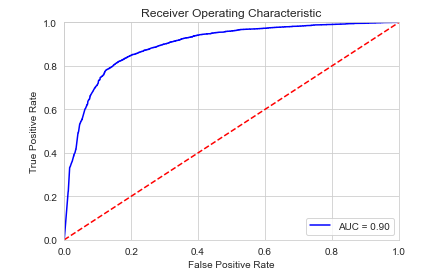


### Random Forest

**

Kết quả của mô hình như sau:





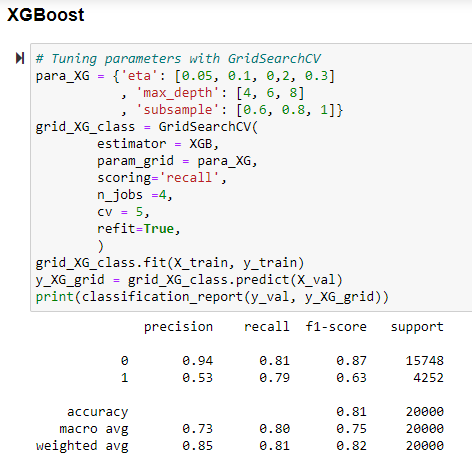
## Phát triển và tối ưu hóa mô hình trên Validation set

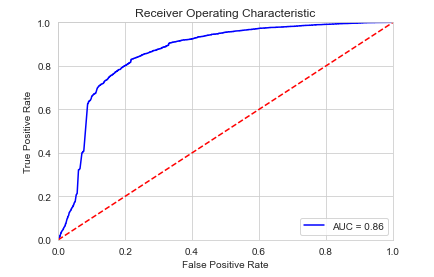
Sau khi thử nghiệm các mô hình dự đoán khác nhau, cho ra được 2 mô hình có độ chính xác cao nhất là XGBoost và Random Forest.

Ta tiến hành sử dụng GridSearchCV để tối ưu hóa các hyper parameters và kiểm tra lại mô hình bằng tính năng kiểm tra chéo Cross Validation.

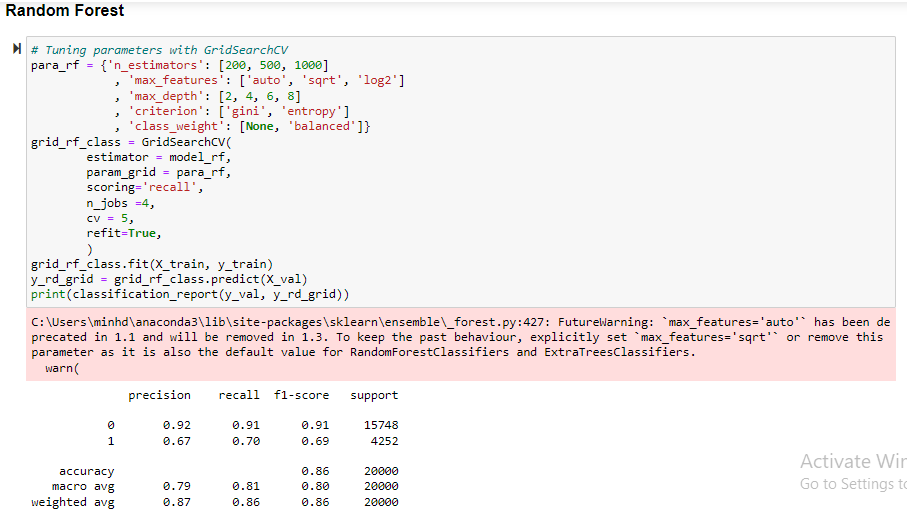
Kết quả nhận về như sau:

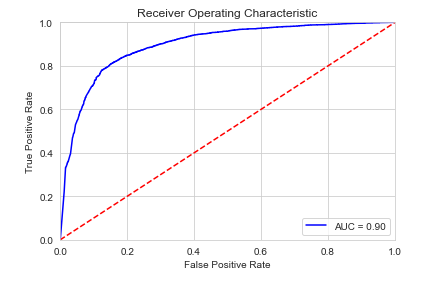
### Mô hình XGBoost





### Mô hình Random Forest

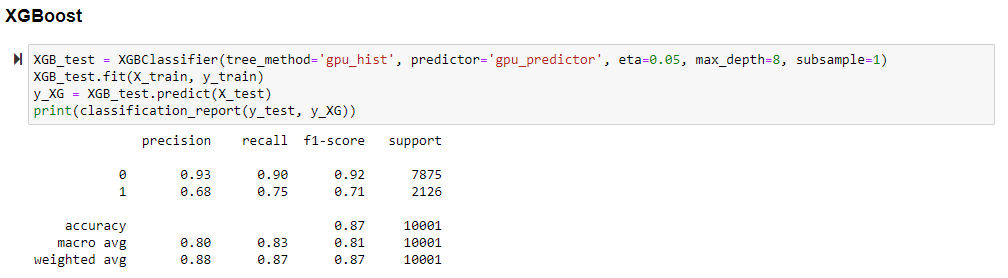


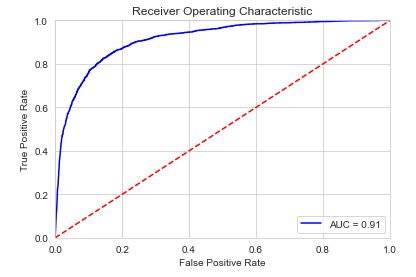


## Kiểm tra mô hình trên Test set

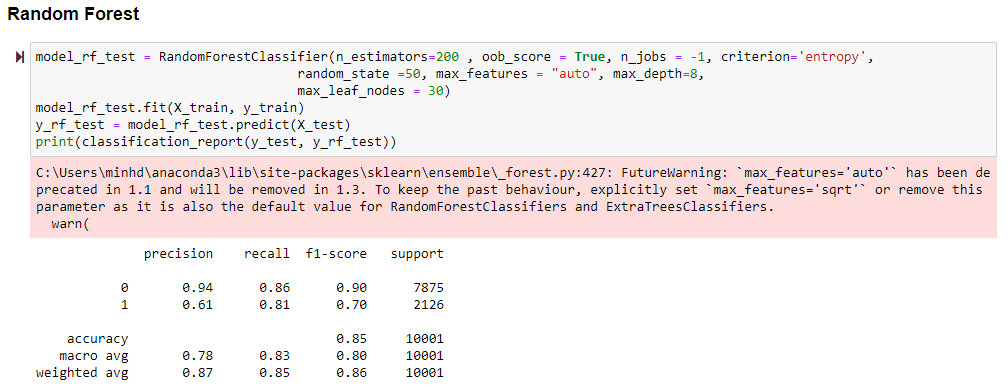
Kiểm tra mô hình trên test set, ta nhận về kết quả cuối cùng như sau:

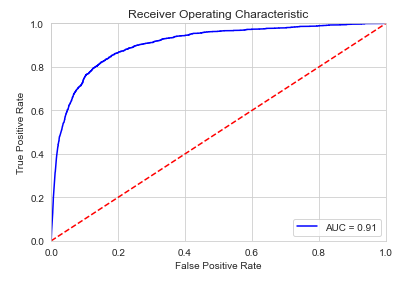
### Mô hình XGBoost





### Mô hình Random Forest





## Đánh giá chung

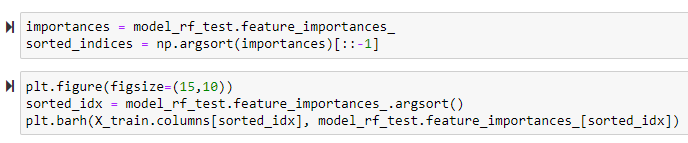
Có thể thấy, trong 5 mô hình đưa vào thử nghiệm, Random Forest là mô hình có kết quả tốt nhất với các chỉ số như sau:

Recall = 0.81

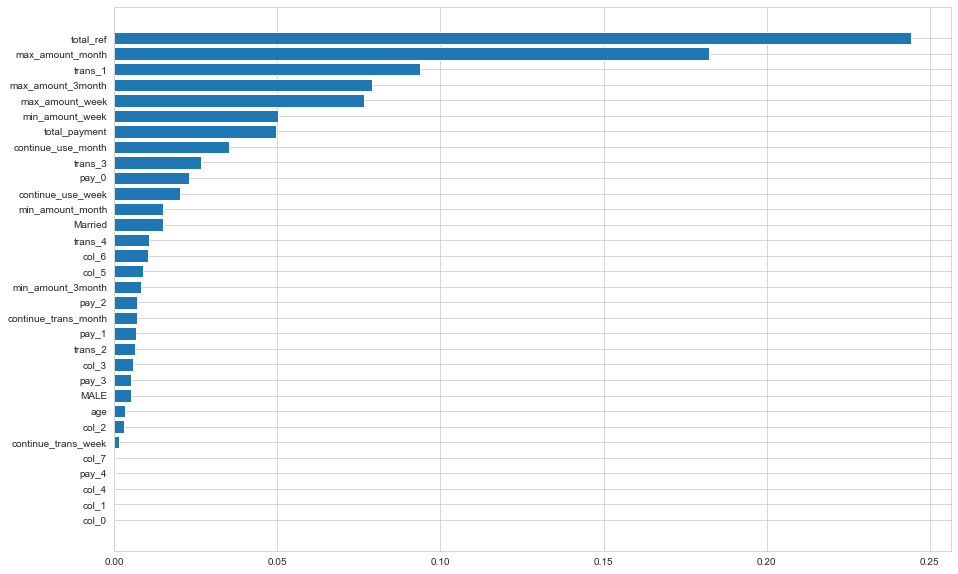
F1 score = 0.7

AUC = 0.91

Như vậy, dựa trên số liệu có được trong khoảng thời gian nghiên cứu, có thể kết luận Random Forest Model là Model tốt nhất dành cho việc dự đoán tỷ lệ khách hàng có khả năng rời bỏ của ngân hàng MB Bank. Ngoài ra, ta cũng có thể tìm được những Features quan trọng dựa vào thuật toán Random Forest Classification:



Kết quả nhận về như sau:



Như vậy, các biến quan trọng ảnh hưởng đến khả năng rời bỏ của khách hàng có thể kể đến là:

(1) Số tiền giao dịch lớn nhất trong 3 tháng/1 tháng/1 tuần trước T12/2020

(2) Số tiền giao dịch nhỏ nhất trong 3 tháng/1 tháng/1 tuần trước T12/2020

(3) Số lượng đối tác trong các giao dịch trong 3 tháng/1 tháng/1 tuần trước T12/2020

# KẾT QUẢ NGHIÊN CỨU VÀ KIẾN NGHỊ

## Kết luận chung

Trong khoảng thời gian nghiên cứu, có 21,3% trong tổng số khách hàng là khách hàng rời bỏ. Tại báo cáo này, nhóm nghiên cứu ưu tiên sử dụng Recall hơn với những lý do đã nếu ở phần 1.3.2. Như vậy, Random Forest Model được kết luận là Model tốt nhất dành cho việc dự đoán tỷ lệ khách hàng có khả năng rời bỏ của ngân hàng MB Bank.

Nhóm nghiên cứu cũng chỉ ra những tập khách hàng có khả năng rời bỏ cao, cụ thể như sau:

* Về nhân khẩu học: Nhóm khách hàng có độ tuổi khoảng 20, chưa có gia đình hoặc gặp các vấn đề về hôn nhân.
* Về hành vi giao dịch: Những khách hàng sử dụng loại dịch vụ 0, 1, 2 và nhóm giao dịch 1, 2, 3 có khả năng rời bỏ cao.
* Ngoài ra, những khách hàng sử dụng nhiều loại dịch vụ, và có nhiều đối tác có ít khả năng rời bỏ.

## Kết quả nghiên cứu và kiến nghị

Nhóm nghiên cứu sau quá trình nghiên cứu đã rút ra một số kết luận chi tiết cùng các kiến nghị đi kèm như sau:

Thứ nhất, mặc dù số khách hàng Nam chiếm đa số, tuy nhiên điều này không ảnh hưởng quá nhiều tới tỷ lệ churn giữa 2 nhóm khách hàng này. Như vậy, chúng ta chưa thể kết luận được ảnh hưởng của giới tính đến khả năng rời bỏ của khách hàng, nên cần tiếp tục theo dõi mối liên hệ giữa giới tính và xu hướng rời bỏ của khách hàng.

Thứ hai, khách hàng có độ tuổi cao hơn thì có tỷ lệ rời bỏ thấp hơn. MB Bank cần có các chính sách nhằm tiếp tục giữ chân những người lớn tuổi bằng cách duy trì ổn định chất lượng phục vụ, và cũng cần tập trung vào các khách hàng trẻ tuổi để giảm bớt tỷ lệ rời bỏ như Để cạnh tranh với các loại ví điện tử, cần miễn phí chuyển tiền trên app Mobile Banking, tuy nhiên có thể xem xét cắt giảm các loại phí như: Phí phát hành thẻ ATM, phí rút tiền tại ATM và đặc biệt là cải thiện chất lượng dịch vụ, tốc độ giao dịch trên mobile app,…

Thứ ba, những khách hàng chưa có gia đình hoặc đang độc thân sẽ có tỷ lệ rời bỏ cao hơn. Nguyên nhân có thể do sau khi lập gia đình khách hàng có xu hướng ổn định và nghiêm túc hơn với các vấn đề tài chính. Ngân hàng cần lưu ý đến vấn đề này và tiếp tục theo dõi xu hướng và mối liên hệ giữa tình trạng hôn nhân và mức độ rời bỏ của khách hàng.

Thứ tư, mặc dù số khách hàng sử dụng dịch vụ 0, 1, 2 là rất lớn, tuy nhiên tần suất giao dịch thấp, và cũng là những dịch vụ có tỷ lệ khách hàng rời bỏ cao. Ta cần xem xét lại về hiệu quả hoạt động và chăm sóc khách hàng của hai dịch vụ này để đưa ra giải pháp cụ thể để giảm tỷ lệ rời bỏ, phát huy tốt nhất lợi thế số người dùng lớn của các dịch vụ này.

Thứ năm, tỷ lệ rời bỏ có xu hướng giảm dần theo nhóm giao dịch mà khách hàng sử dụng. Đặc biệt là nhóm 1 có tỷ lệ khách hàng rời bỏ rất cao (64%). Tương tự như dịch vụ 0, 1, 2, ta cần phải nghiên cứu lại nguyên nhân và tình hình hoạt động của nhóm giao dịch này để đưa ra giải pháp phù hợp. Chúng ta có thể nghiên cứu thêm về thói quen và loại hình giao dịch của tệp khách hàng theo độ tuổi, tình trạng kết hôn, … từ đó phân chia được rõ ràng và cụ thể hơn xu hướng, nhu cầu của khách hàng và vạch ra những kế hoạch phù hợp cho từng tệp khách hàng cụ thể.

Thứ sáu, cần khảo sát, kiểm tra thường xuyên để nâng cao chất lượng dịch vụ ở các cơ sở ngân hàng, đặc biệt là những nơi có khách hàng đông đảo - cũng là những nơi có tỷ lệ rời bỏ cao.

Cuối cùng, ngân hàng nên thực hiện các khảo sát nguyên nhân tại sao khách hàng rời bỏ để kịp thời có biện pháp giải quyết. Dựa trên phân tích về số liệu, các mô hình dự báo và kết quả khảo sát đó, ngân hàng sẽ tập trung nhắm đến những đối tượng có nguy cơ ngừng dịch vụ để đưa ra các gói tiện ích, như các đợt ưu đãi hay nâng cấp một số tính năng cụ thể và nâng cao chất lượng dịch vụ để thuyết phục khách hàng ở lại.

# LỜI KẾT

Việc phân tích và tìm ra tỷ lệ khách hàng rời bỏ là hoạt động vô cùng quan trọng, nhất là đối với một doanh nghiệp lớn như MB Bank, bởi tỷ lệ khách hàng rời bỏ sẽ có tác động mạnh mẽ và lâu dài đến giá trị của doanh nghiệp. Qua nghiên cứu của nhóm, với định nghĩa khách hàng rời bỏ là những khách hàng không có giao dịch nào trong tháng 3 và tháng 6 năm 2021 của doanh nghiệp, nhóm nghiên cứu đã đưa ra kết luận rằng tỷ lệ khách hàng rời bỏ trong suốt quãng thời gian nghiên cứu là 21,3% và Random Forest Model là mô hình phù hợp nhất để tiếp tục dự đoán tỷ lệ khách hàng rời bỏ của MB Bank.

Nhìn chung, việc phân tích dữ liệu và xây dựng mô hình để dự đoán về tỷ lệ rời bỏ của khách hàng tại MB cần trải qua nhiều bước, sử dụng nhiều kỹ thuật và cũng cần thêm nhiều thời gian để quan sát và đưa ra kết luận một cách chính xác nhất.

Trong quá trình nghiên cứu, do sự hạn chế về thời gian cũng như về kinh nghiệm, khả năng, nghiên cứu của nhóm không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được những phản hồi và đóng góp từ Giảng viên và anh chị chuyên viên đến từ Quý công ty để bài nghiên cứu có thể hoàn thiện hơn.

Chúng em xin trân trọng cảm ơn!

# TÀI LIỆU THAM KHẢO

1. Belém, N. and Rosa, C. (n.d.). Gauging and Foreseeing Customer Churn in the Banking Industry: A Neural Network Approach Project Work report presented as partial requirement for obtaining the Masters degree in Information Management. [online] Available at: <https://run.unl.pt/bitstream/10362/71584/1/TGI0223.pdf> [Accessed 16 Mar. 2022].
2. Guangli Nie , Wei Rowe , Lingling Zhang , Yingjie Tian , Yong Shi. (2011). Credit card churn forecasting by logistic regression and decision tree.
3. Glady, N., Baesens, B. and Croux, C. (2009). Modeling churn using customer lifetime value. European Journal of Operational Research, 197(1), pp.402–411 [Accessed 16 Mar. 2022].
4. Lu,. (2002). Predicting Customer Churn in the Telecommunications Industry –– An Application of Survival Analysis Modeling Using SAS, P 114-27 [Accessed 16 Mar. 2022].
5. Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43(2), pp.204–211 [Accessed 16 Mar. 2022].
6. Nie, G., Rowe, W., Zhang, L., Tian, Y. and Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 38(12), pp.15273–15285 [Accessed 16 Mar. 2022].
7. Scikit-learn - Machine Learning in Python. [online]. Available at: <https://scikit-learn.org/stable/>[Accessed 16 Mar. 2022]
8. Van den Poel, D. and Larivière, B. (2004) "Customer attrition analysis for financial services using proportional hazard models", European Journal of Operational Research, 157(1), pp. 196-217. doi: 10.1016/s0377-2217(03)00069-9 [Accessed 16 Mar. 2022].