ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ

Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Nguyễn Minh Hiếu	20521326	HTTT
2	Nguyễn Thành Nhân	20521701	KHDL

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Khách hàng rời bỏ (Customer Churn) là việc khách hàng ngừng sử dụng sản phẩm hoặc dịch vụ của công ty hoặc doanh nghiệp. Nói cách khác là việc mất đi khách hàng vì bất kì lí do gì trong một khung thời gian nhất định. Bài toán **Dự đoán khách hàng rời bỏ** *(Customer Churn Prediction)* là việc dựa trên dữ liệu về lịch sử mua hàng, lịch sử sử dụng dịch vụ, các dữ liệu thống kê về khách hàng cũng như những thông tin được cung cấp từ khách hàng để phân tích, đề xuất và xây dựng phương pháp – mô hình dự đoán. Kết quả khi thu được sẽ trả lời cho câu hỏi khách hàng này là khách hàng tiềm năng hay có nguy cơ rời bỏ.

Qua quá trình thực hiện, nhờ việc áp dụng các kiến thức đã học kết hợp sử dụng các công cụ - thư viện hỗ trợ như Pandas, Numpy, Matplotlib, Seaborn,… chúng em đã hoàn thành việc phân tích, trực quan và tìm ra các yếu tố tiềm ẩn trong bộ dữ liệu. Đồng thời cũng xây dựng phương pháp – mô hình dự đoán khách hàng rời bỏ cho kết quả cao. Kết quả cao nhất đạt được trên mô hình Random Forest với độ đo đánh giá accuracy: 88.25% và f1-score: 85.00% và mô hình Gradient Boosting với accuracy: 88.11% và f1-score: 84.75%.

Bộ dữ liệu dùng để phân tích là bộ dữ liệu thống kê về khách hàng của công ty bán lẻ trang sức là Pandora [1] dành cho thực tập sinh (đây là bộ dữ liệu không công khai và cho phép thực tập sinh sử dụng để nghiên cứu). Tập dữ liệu lưu trữ thông tin về các thông tin mua hàng của khách hàng và loại bỏ các thông tin cá nhân quan trọng. Bộ dữ liệu thể hiện thông tin về việc khách hàng nào đã *không quay lại mua hàng trong 12 tháng*.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu đã được loại bỏ các thông tin cá nhân của khách hàng và chỉ dùng để nghiên cứu phục vụ đồ án môn học chứ không đưa vào hay mang bất kỳ mục đích thương mại nào khác. Đây là bộ dữ liệu cho biết những thông tin về khách hàng của một công ty trang sức, các thuộc tính như bảng sau:

Bảng 2.1. Mô tả các thuộc tính của bộ dữ liệu

Tên thuộc tính	Loại thuộc tính	Kiểu giá trị	Giá trị minh họa	Ý nghĩa
Member No.	Numerical	object	“0000098000030”,…	Mã khách hàng
Total_Lifetime_Spending	Numerical	int64	20655000, 27830000	Tổng số tiền khách hàng đã mua
Correct_Category_Member	Categorical	object	“Bronze”, “Silver”, “Gold”	Hạng khách hàng
Num_of_Win	Numerical	int64	1, 2, 3, ...	Số lần khách hàng thăng hạng
Num_of_Drop	Numerical	int64	1, 2, 3, …	Số lần khách hàng xuống hạng
Number_of_purchases	Numerical	int64	1, 2, 3, …	Số lần mua hàng
Discount_code_usage_count	Numerical	int64	1, 2, 3, …	Số lần sử dụng chương trình giảm giá
Avg_Days_Between_Transactions	Numerical	float64	38.468750, 46.428571,...	Trung bình số ngày khách hàng quay lại mua hàng
Gender	Categorical	object	“Male”, “Female”	Giới tính
Age_Group	Categorical	object	“18-24 Years”,…	Nhóm tuổi

Store_No.	Categorical	object	"HCMC", "HN", "ONLINE"	Nơi khách hàng mua hàng lần cuối
Purchase_Type	Categorical	object	"Gift to Others", "For Self"	Mua hàng nhằm mục đích gì
Customer_Type	Categorical	object	"Individual", "Company"	Phân biệt loại khách hàng
Churn	Binary	int64	0, 1	Kiểm tra khách hàng còn mua hàng không

Bảng 2.2. Mô tả ý nghĩa các giá trị của thuộc tính

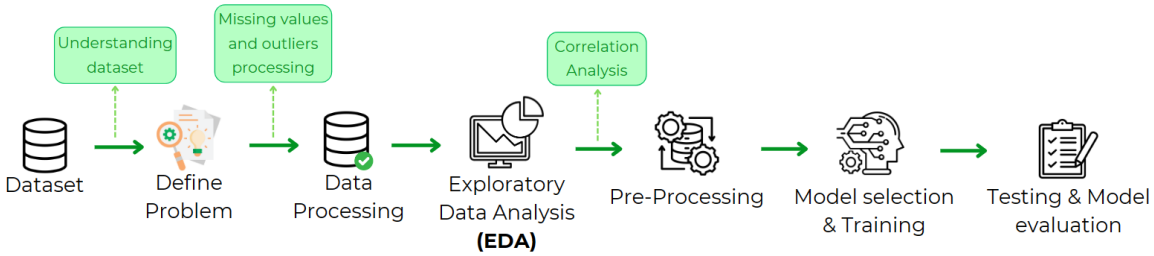
Tên thuộc tính	Tên giá trị	Ý nghĩa
Correct_Category_Member	Bronze	Khách hàng Đồng
	Silver	Khách hàng Bạc
	Gold	Khách hàng Vàng
Gender	Male	Giới tính Nam
	Female	Giới tính Nữ
Age_Group	18-24 Years	Nhóm tuổi từ 18 – 24 tuổi
	25-34 Years	Nhóm tuổi từ 25 – 34 tuổi
	35-44 Years	Nhóm tuổi từ 35 – 44 tuổi
	45-54 Years	Nhóm tuổi từ 45 – 54 tuổi
	More than 55	Nhóm tuổi hơn 55 tuổi
Store_No.	HCMC	Nơi mua hàng là Tp HCM
	HN	Nơi mua hàng là là Tp HN
	ONLINE	Nơi mua hàng là Website
Purchase_Type	For Self	Mua cho bản thân
	Gift to Others	Mua để tặng
Customer_Type	Individual	Khách hàng Cá nhân
	Company	Khách hàng là doanh nghiệp
Churn	0	Khách hàng vẫn mua hàng
	1	**Khách hàng không còn mua hàng**

Bộ dữ liệu có:

- 14 cột thuộc tính
- 5138 mẫu dữ liệu
- Số giá trị khuyết:
  - "Gender": 73
  - "Age_Group": 553
  - "Purchase Type": 122
  - "Customer Type": 119

3. PHƯƠNG PHÁP PHÂN TÍCH

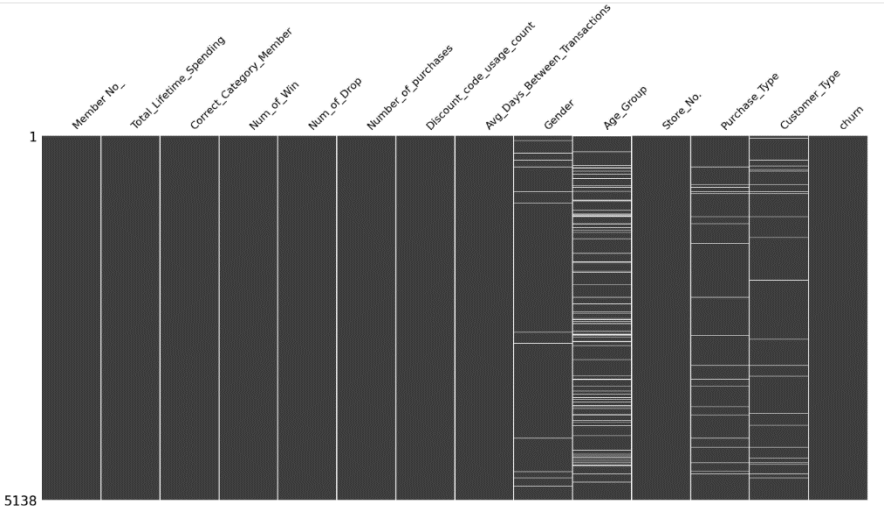
Sau khi có được bộ dữ liệu, chúng em xây dựng quy trình phân tích gồm các bước: Xác định vấn đề, làm sạch bộ dữ liệu, phân tích thăm dò và trực quan, tiền xử lý dữ liệu, chọn và huấn luyện mô hình và cuối cùng là đánh giá.



Hình 3.1. Quy trình phân tích dữ liệu

3.1. Xác định vấn đề

- Biến mục tiêu (target) là 'Churn' với 2 giá trị là:
  - "0": Khách hàng vẫn mua hàng
  - "**1**": **Khách hàng rời bỏ - Khách hàng không còn mua hàng trong 12 tháng**
- Kiểm tra kiểu dữ liệu: các biến đều có kiểu dữ liệu đúng với mong đợi
- Kiểm tra giá trị khuyết (Missing value): Có tổng 873 giá trị khuyết và tập trung ở các biến "Gender", "Age_Group", "Purchase_Type", "Customer_Type". Điểm chung ở đây đều là các thuộc tính chứa thông tin được cung cấp từ khách hàng.



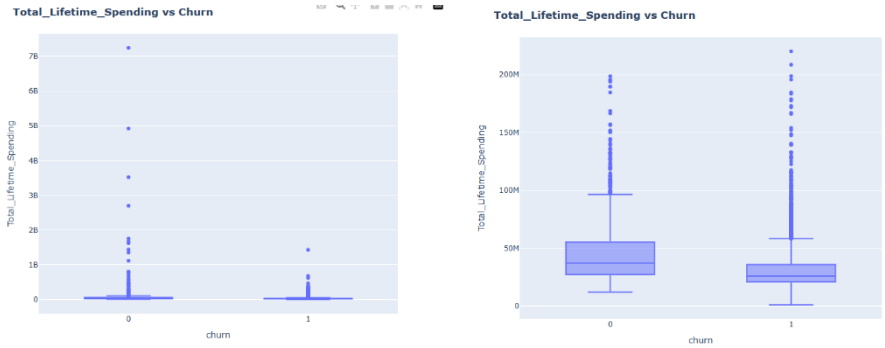
Hình 3.2. Trực quan giá trị khuyết sử dụng Missingno[2]

- Kiểm tra giá trị ngoại lệ (Outliers): Ngoại trừ Churn, còn lại còn lại 6 biến số (Numerical). Trong đó có 2 biến là "Num of Win" và "Num of Drop" chứa ít giá trị (chỉ trong khoảng từ 0 - 5), 4 biến còn lại là "Total_Lifetime_Spending", "Number_of_Purchases", "Discount_code_usage_count", "Avg_Days_Between_Transactions", chứa nhiều giá trị ngoại lệ quá lớn khiến cho miền giá trị khó quan sát.

3.2. Làm sạch bộ dữ liệu

- Xử lý giá trị khuyết: Vì đây là các thuộc tính phân loại *(Categorical)* chứa thông tin được cung cấp từ khách hàng. Do đó chúng em chọn giá trị "Unknown" để đại diện cho chúng.

– Xử lý giá trị ngoại lệ: Qua quan sát và phân tích, chúng em nhận thấy có đến hơn 1000 điểm dữ liệu nằm ngoài phạm vi tứ phân vị đối với 4 biến nêu trên. Vì trong thực tế sẽ có rất nhiều trường hợp gây nhiễu như: Khách hàng là một doanh nghiệp, họ mua một lượng lớn trang sức với số lượng lên tới hàng ngàn và tổng số tiền lên đến hàng tỷ. Nhưng sau đó hơn 12 tháng họ không quay lại mua hàng,… qua phân tích, chỉ có khoảng ít hơn 5% trong tổng số dữ liệu có giá trị vượt quá mức. Các giá trị ngoài tứ phân vị khác thì lại có mật độ không nhỏ. Do đó chúng em đã quyết định chọn xử lý các giá trị ngoại lệ ở mức **5%.**

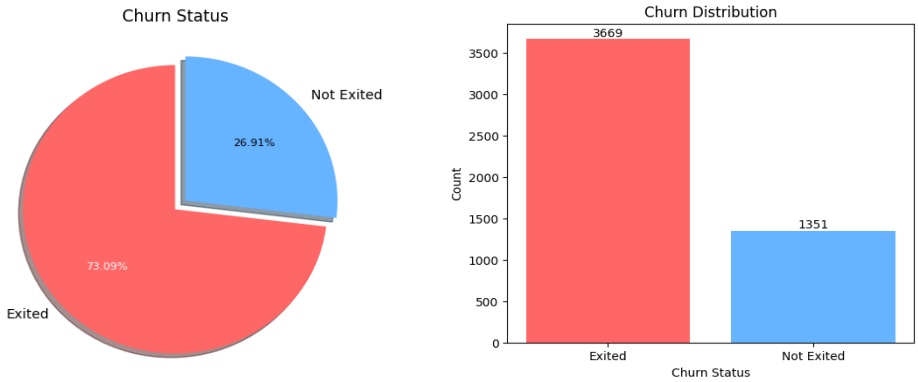


Hình 3.3. Biểu đồ minh họa trước (trái) và sau (phải) xử lý giá trị ngoại lệ

4. PHÂN TÍCH THĂM DÒ

4.1. Biến mục tiêu “Churn”

Sau khi làm sạch, bộ dữ liệu còn 5020 giá trị.



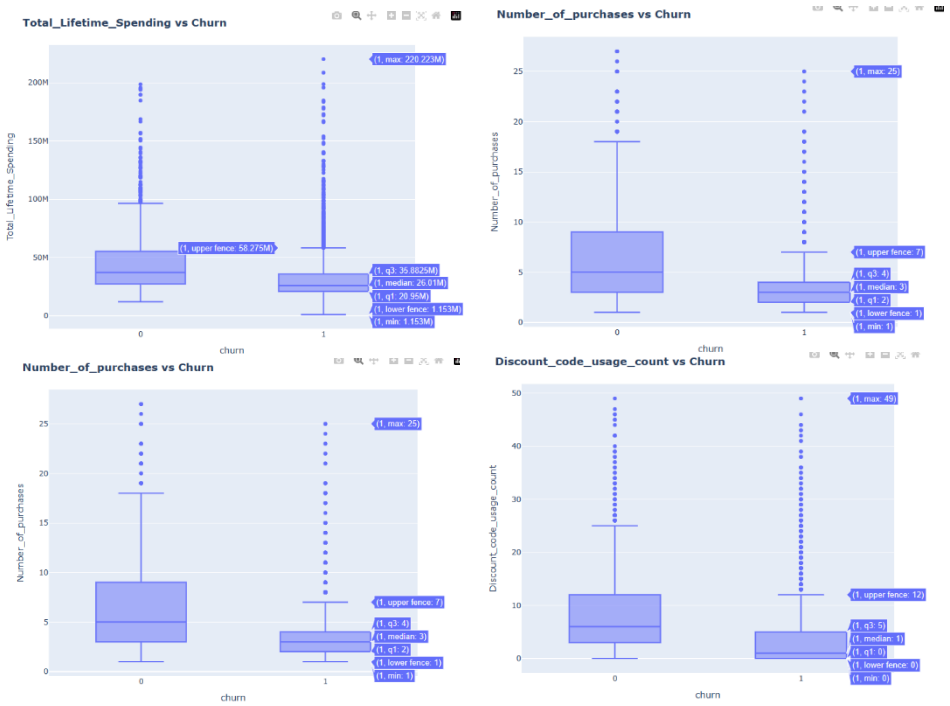
*Hình 4.1. Biểu đồ thể hiện số lượng nhãn của biến mục tiêu “**Churn**”*

Đối với biến mục tiêu “Churn”. Nhãn khách hàng rời bỏ (**1 – Exited**) nhiều hơn gần gấp 3 lần nhãn khách hàng còn tiếp tục mua hàng (**0 – Not Exited)**. Như vậy, bộ dữ liệu bị mất cân bằng.

4.2. Tương quan của các biến đối với biến mục tiêu “Churn”

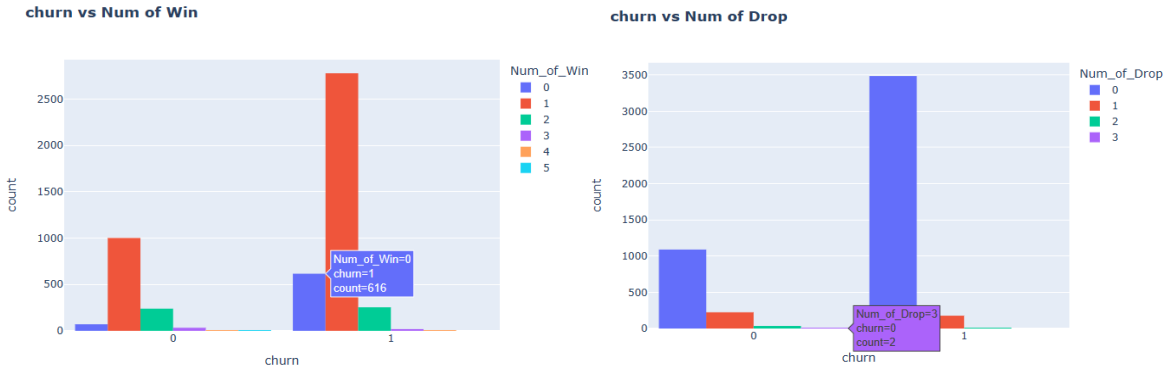
Các biến số hầu như đều có cùng chiều dữ liệu tương tự nhau.

- Thông qua biểu đồ Boxplot (Hình 4.2), có thể thấy đối với nhóm khách hàng vẫn còn mua hàng luôn có giá trị tổng số tiền mua hàng cao hơn hẳn so với nhóm khách hàng rời bỏ.



Hình 4.2. Các biểu đồ thể hiện tương quan giữa "Churn" và các biến số

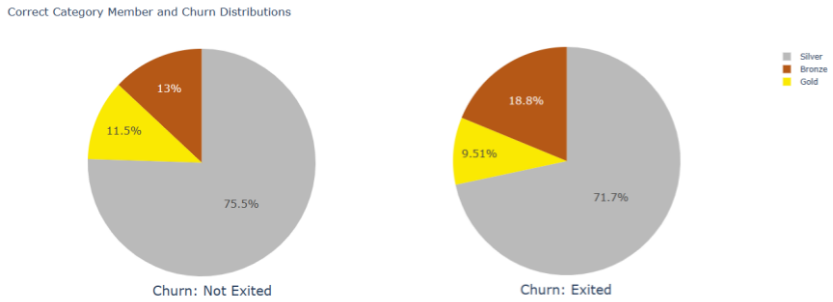
Tuy nhiên xu hướng chiều dữ liệu của các biến đều tập trung phía dưới giá trị trung bình. Bên cạnh đó, biểu đồ của các biến cho thấy cái Box không quá chồng lấp nhau → có ý nghĩa thống kê.



Hình 4.3. Biểu đồ thể hiện tương quan giữa "Churn" và 2 biến "Num of Win/Drop"

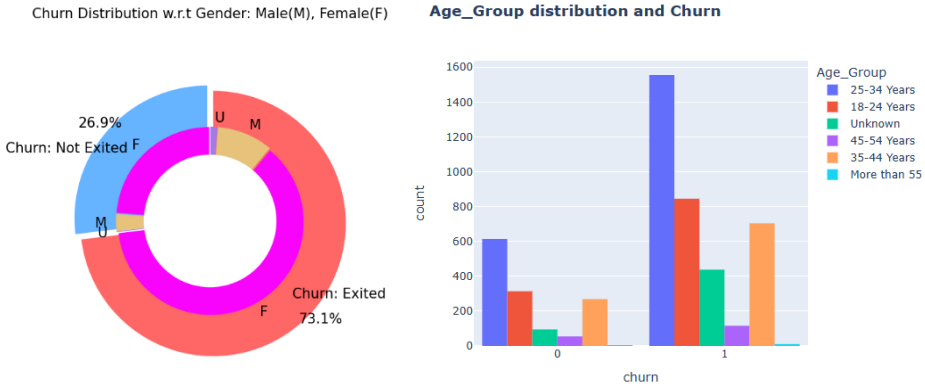
- Việc thống kê số lần nâng rank "Num of Win" và bị hạ rank "Num of Drop" cũng sẽ là một biến tiềm năng để dự đoán khách hàng rời bỏ. Chúng ta thấy nhóm khách hàng rời bỏ chưa được nâng hạng lần nào có tỉ lệ cao vượt trội so với khách hàng còn mua hàng. Nhưng nhóm khách hàng còn mua hàng thì có số lần nâng hạng nhiều hơn 2 lần có tỉ lệ cao hơn. Điều này cho thấy một số ít khách hàng có thể vì không có nhu cầu mua trang sức thường xuyên, hoặc là họ không tiếp tục mua hàng đạt mức để duy trì mức rank sẽ làm giảm rank của họ, và khi họ mua hàng lại và đạt mức thì họ lại được nâng rank. Vì thế không quá ngạc nhiên khi số lần hạ rank từ 1 lần trở lên của nhóm khách hàng còn mua hàng có số lượng cao hơn khách hàng rời bỏ.

- Sự tương quan giữa biến "Correct Category Member" và biến mục tiêu (Hình 4.4) cho thấy lượng khách hàng rời bỏ có mức rank 'Bronze' cao hơn khách hàng còn mua hàng.



Hình 4.4. Biểu đồ thể hiện tương quan giữa "Churn" và Member Rank

Ngược lại thì rank 'Gold' và 'Silver' của nhóm này thì thấp hơn. Điều này cho thấy khách hàng tiếp tục mua hàng (ngoại trừ những người mới mua 1 lần trong 12 tháng thống kê gần nhất) thì có tỉ lệ ở mức rank cao sẽ cao hơn.

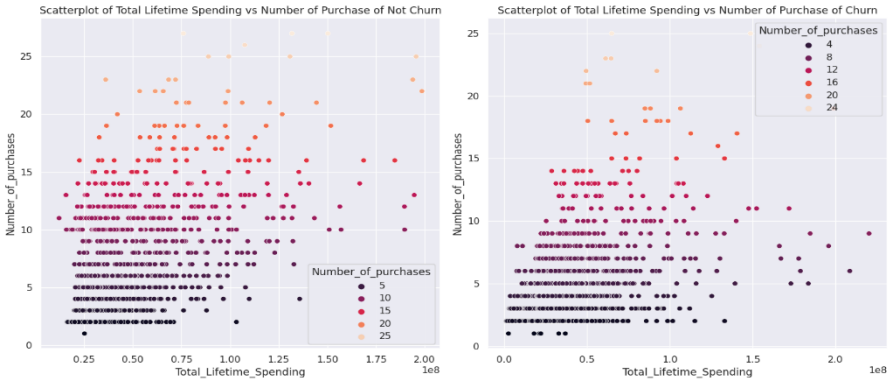


Hình 4.5. Các biểu đồ thể hiện tương quan giữa "Churn" với giới tính và nhóm tuổi

Thông tin về giới tính và nhóm tuổi được khách hàng cung cấp.

- Vì mặt hàng là trang sức, không quá khó hiểu khi khách hàng là nữ chiếm đa số và số lượng khách hàng còn mua hàng hay rời bỏ là nữ cũng chiếm đa số. Bên cạnh đó, lượng khách hàng không cho biết thông tin giới tính của mình có tỉ lệ rời bỏ cao hơn tỉ lệ còn tiếp tục mua hàng.

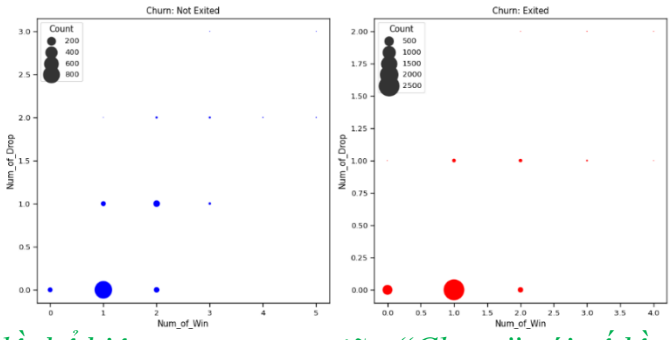
- Tỉ lệ nhóm tuổi của mỗi nhóm khách hàng khá đồng đều. Tuy nhiên ta có thể thấy lượng khách hàng thuộc nhóm trên 55 tuổi và không cung cấp thông tin nhóm tuổi có tỉ lệ rời bỏ cao hơn.



Hình 4.6. Biểu đồ thể hiện tương quan giữa "Churn" với tổng số tiền và số lượng hàng

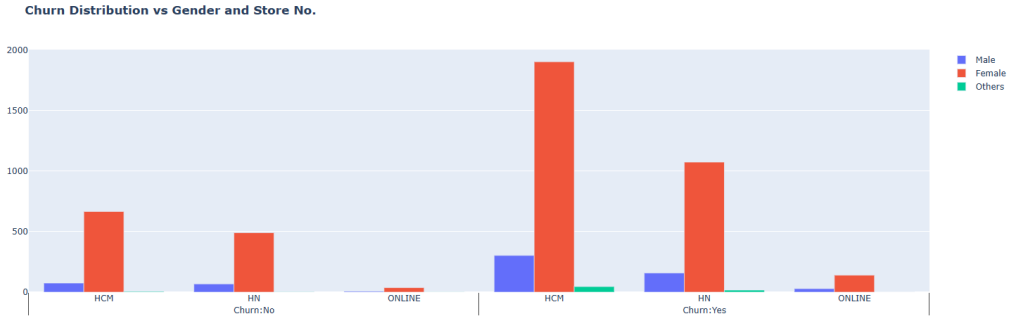
- Hình trên cho thấy ở cả 2 nhóm khách hàng mua tổng số lượng món hàng trong khoảng 0 – 10 món và có tổng tiền từ 25 – 75 triệu có mật độ rất dày.

+ Ở nhóm khách hàng còn mua hàng, các khách hàng đều mua từ 2 món trở lên và tổng tiền ít nhất cũng từ 10 triệu trở lên. Tuy nhiên biểu đồ phân tán còn cho thấy khách hàng mua hàng từ 10 món trở lên và có tổng tiền từ 125 triệu trở xuống rất cao. Ngoài ra cũng có một bộ phận khách hàng mua hàng càng nhiều thì tổng tiền cũng càng cao.
+ Ở nhóm khách hàng rời bỏ, các khách hàng có tổng tiền ít nhất từ khoảng 1 triệu và có số lượng rất nhiều trong khoảng từ 1-10 món với khoảng tiền từ 1 – 75 triệu. Số lượng khách hàng mau từ 15 – 25 món chỉ có tổng tiền trong khoảng từ 50 – 125 triệu. Ngoài ra có xuất hiện các khách hàng mua trong khoảng 10 món nhưng lại có tổng tiền tới hơn 150 triệu.



Hình 4.7. Biểu đồ thể hiện tương quan giữa "Churn" với số lần tăng/giảm hạng

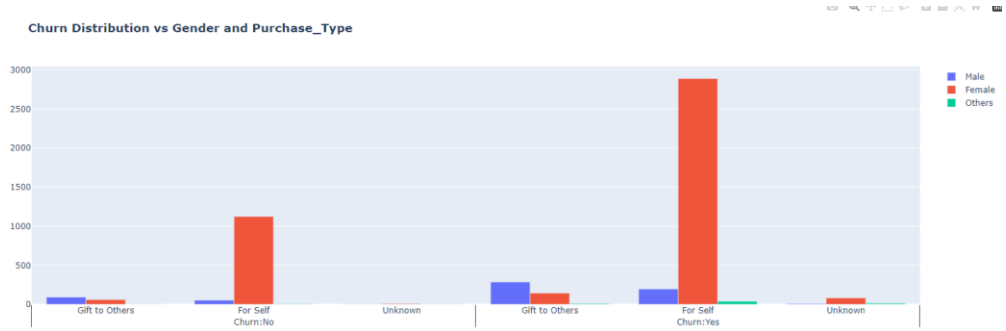
- Xét mối quan hệ giữa việc tăng và giảm hạng so với churn, theo độ to điểm dữ liệu của biểu đồ, các khách hàng được tăng hạng 1 lần và chưa bị giảm hạng có tỉ lệ cao nhất ở cả 2 nhóm khách hàng. Khách hàng rời bỏ tập trung nhiều ở điểm thể hiện chưa tăng cũng như chưa giảm. Ngược lại khách hàng còn tiếp tục mua hàng tập trung nhiều ở các điểm 'tăng 2-giảm 1', 'tăng 2-giảm 0', 'tăng 1-giảm 1'.



Hình 4.8. Biểu đồ thể hiện tương quan giữa "Churn" với giới tính và khu vực

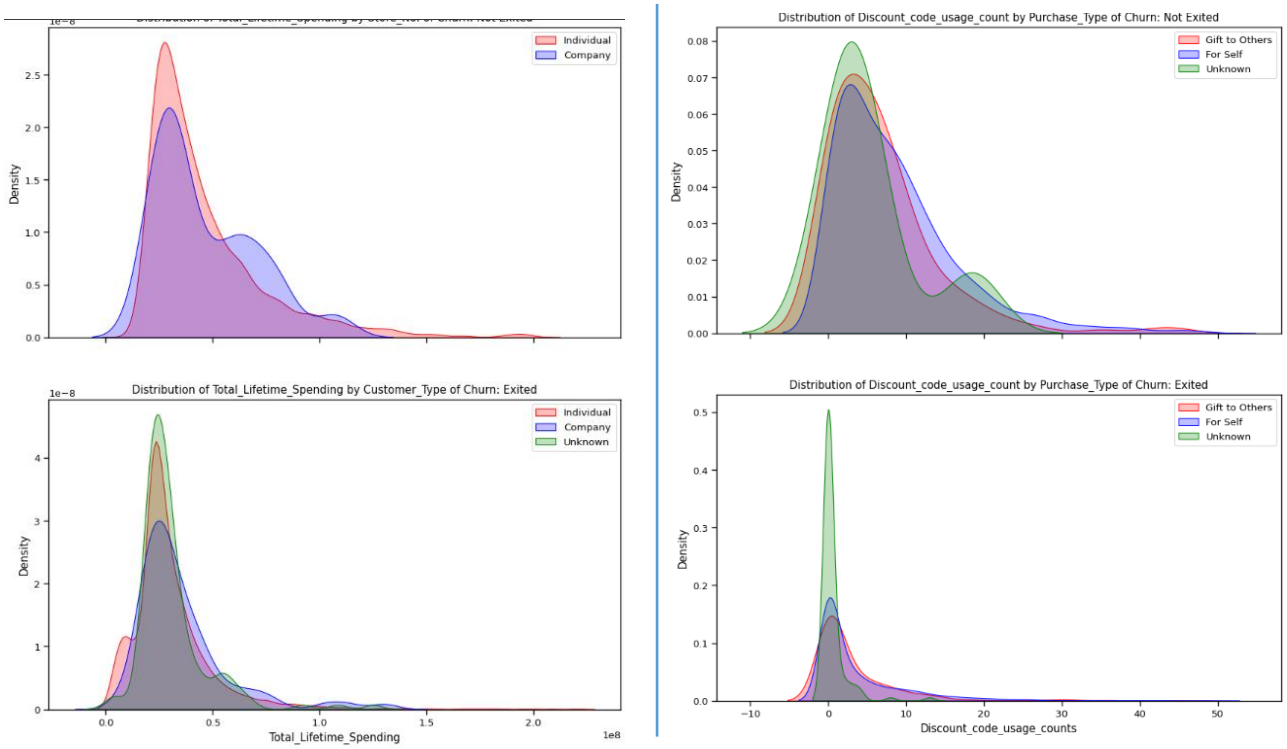
- Lượng khách hàng thuộc nhóm khách hàng rời bỏ và không cho biết giới tính chiếm tỉ lệ cao hơn sao với khách hàng còn mua hàng, và có nhiều ở Tp.HCM và HN.

- Ở biểu đồ Hình 4.9, ta thấy khách hàng nữ mua trang sức để làm quà tặng rất ít, mua cho bản thân chiếm tỉ lệ cao. Bên cạnh đó, giới tính nam nhưng mua để làm quà hay mua cho bản thân có tỉ lệ rời bỏ cao vượt trội.



Hình 4.9. Biểu đồ thể hiện tương quan giữa "Churn" với giới tính và mục đích

Về các khách hàng không cung cấp thông tin giới tính, hầu hết đều thuộc nhóm rời bỏ.



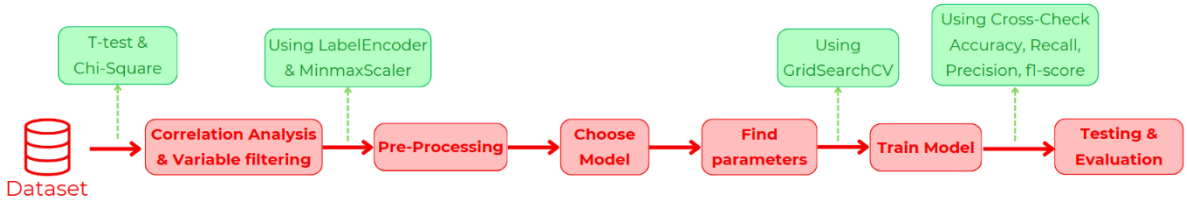
Hình 4.10. Biểu đồ thể hiện tương quan giữa "Churn" với tổng số tiền và loại khách hàng (trái); với số lần sử dụng chương trình giảm giá và mục đích (phải)

- Ở nhóm khách hàng rời bỏ chứa toàn bộ lượng khách hàng không rõ thông tin của thuộc tính loại khách hàng. Giá trị tổng số tiền của nhóm này tập trung nhiều nhất ở khoảng 25 triệu trở xuống. Ở nhóm khách hàng còn lại, khách hàng cá nhân tập trung nhiều ở khoảng tổng tiền mua hàng ở khoảng hơn 25 triệu trở lên, riêng khách hàng là doanh nghiệp thì có một số tập trung ở khoảng tổng tiền cao hơn từ 50 triệu trở lên.

- Ở nhóm khách hàng rời bỏ, những khách hàng không rõ mục đích ở khoảng 0 có số lượng rất nhiều, tức chưa sử dụng chương trình giảm giá lần nào. Các khách hàng có mục đích thì ở khoảng có sử dụng từ 1 – 10 chương trình giảm giá. Ở nhóm khách hàng tiếp tục mua hàng thì số lượt sử dụng chương trình giá cao vượt trội hơn so với khách hàng rời bỏ. Khách hàng có lượt sử dụng từ 1 – 10 và từ 15 – 25 chương trình giảm giá lại là các khách hàng không cung cấp mục đích Riêng các khách hàng sử dụng từ 10 – 30 chương trình thì với mục đích mua cho bản thân vượt trội hơn.

5. PHƯƠNG PHÁP THỰC NGHIỆM

Quy trình thiết kế thực nghiệm được chúng em thể hiện cụ thể như hình dưới.



Hình 5.1. Quy trình thiết kế thực nghiệm

5.1. Kiểm tra độ tương quan và lọc thuộc tính:

Để kiểm tra mức độ ảnh hưởng của các biến đối với biến mục tiêu "Churn", chúng em sử dụng:

- *Tính độ tương quan Correlation*: Các biến số đều cho tương quan tiêu cực mức thấp.
- *Kiểm định T-test* đối với các biến số: Các biến số đều cho P-value < 0.05, tức là *có sự khác biệt giữa nhóm khách hàng tiếp tục mua hàng và rời bỏ*. Có ý nghĩa thông kê đối với bài toán phân loại nhị phân.
- *Kiểm định Chi-Square* đối với các biến phân loại: Các biến phân loại đều cho P-value < 0.05, tức là *có đủ bằng chứng để bác bỏ giả thuyết các biến này không có mối quan hệ với biến mục tiêu*. Có ý nghĩa thống kê là toàn bộ các biến phân loại đều có mối quan hệ đối với biến mục tiêu "Churn".

Vì vậy giữ lại toàn bộ thuộc tính để để triển khai huấn luyện mô hình.

5.2. Tiền xử lý dữ liệu:

Đầu tiên sẽ loại bỏ biến mã khách hàng "Member No.".

- Đối với các biến số, dùng *MinmaxScaler* để chuẩn hóa về cùng phạm vi các biến.
- Đối với các biến phân loại, dùng *LabelEncoder* để map các giá trị chuỗi (str) thành số (int). Ngoài ra đối với các biến có chứa giá trị đại diện giá trị khuyết "Unknown", chúng em chuẩn hóa thành giá trị ***-1*** để máy học các giá trị này mang hướng tiêu cực.
- Data sau khi xử lý sẽ được chia thành 2 tập Train – Test với tỉ lệ 7 – 3.

5.3. Mô hình máy học:

Các mô hình máy học chúng em sử dụng bao gồm:

- LogisticRegression
- LogisticRegression kết hợp chuẩn hóa PolynomialFeatures bậc 2
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting

5.4. Training:

- Đầu tiên, chúng em dùng *GridSearchCV* để train các model đã chọn để tìm ra các tham số phù hợp nhất (Best Parameters) cho mô hình để đánh giá bộ dữ liệu.
- Thực hiện huấn luyện model với tham số đã tìm được.

5.5. Đánh giá mô hình:

Chúng em dùng các phương pháp sau để đánh giá mô hình:

- Cross-check: Kiểm tra xem mô hình có học tốt trên toàn bộ bộ dữ liệu hay không.
- Các độ đo đánh giá: Accuracy, Precision, Recall, F1-Score.
- ROC Curve Plot: Đánh giá hiệu suất và khả năng phân loại của mô hình.

6. KẾT QUẢ THỰC NGHIỆM

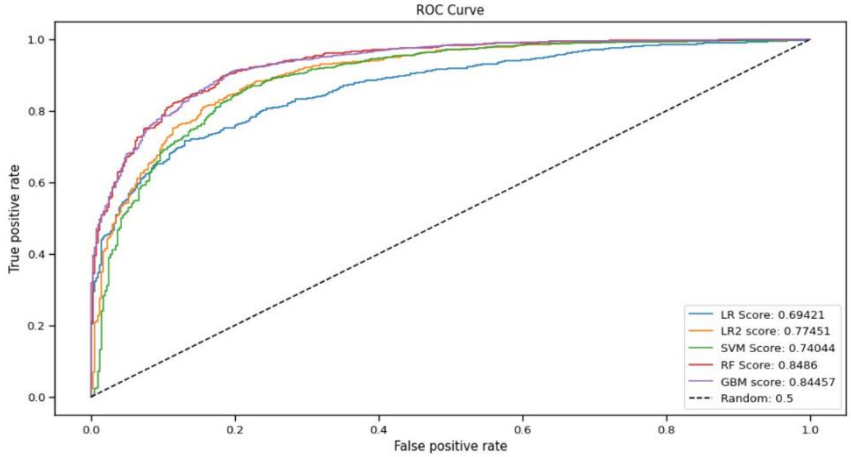
6.1. Kết quả các mô hình

- Các mô hình khi thực hiện kiểm tra chéo Cross-check đều cho kết quả tốt.

Bảng 6.1. Kết quả các mô hình

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegression	0.8028	0.7641	0.6942	0.7146
LogisticRegression + poly(degree = 2)	0.8506	0.8270	0.7745	0.7945
SVM	0.8453	0.8482	0.7404	0.7712
Random Forest	**0.8825**	**0.8514**	**0.8486**	**0.8500**
Gradient Boosting	0.8811	0.8507	0.8446	0.8475

Bảng kết quả cho thấy hai mô hình **Random Forest** và **Gradient Boosting** cho kết quả cao nhất.

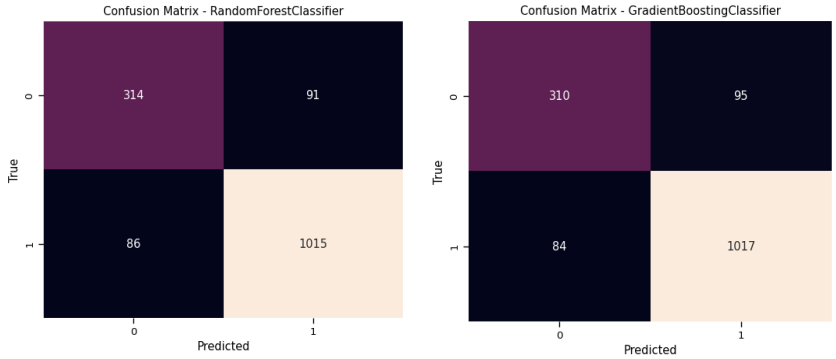


Hình 6.1. Đồ thị đường cong ROC

ROC[3] (Receiver Operating Characteristic) là một đồ thị biểu diễn khả năng phân loại của một mô hình nhị phân. Đường cong ROC biểu thị mức độ tách biệt giữa các lớp dự đoán và được tính toán bằng cách vẽ đường cong dựa trên hai yếu tố: tỷ lệ chính xác đúng dương và tỷ lệ sai dương. Diện tích dưới đường cong ROC (AUC - Area Under the Curve) thường được sử dụng để đánh giá độ chính xác của mô hình dự đoán. Điều này được thể hiện qua tỷ lệ diện tích dưới đường cong ROC với diện tích toàn bộ không gian được biểu diễn bởi đường cong. Qua đó, giá trị AUC càng lớn, mô hình dự đoán càng chính xác.

Kết luận: hai mô hình Random Forest và Gradient Boosting cho đường cong mở rộng về phía trên cùng bên trái nhiều nhất → diện tích lớn nhất → hai mô hình cho kết quả dự đoán trên tập dữ liệu khách hàng rời bỏ tốt nhất trong số các mô hình thử nghiệm.

6.2. Phân tích lỗi:



Hình 6.2. Ma trận nhầm lẫn của hai mô hình Random Forest và Gradient Boosting

Xét theo tỉ lệ, nhãn 0 bị dự đoán thành nhãn 1 chiếm tỉ lệ cao nhất, có 2 lí do giải thích cho điều này:

- Dữ liệu bị mất cân bằng, nhãn 0 chỉ bằng 1/3 nhãn 1
- Thể hiện sai số trong thực tế, sẽ có nhiều trường hợp bất thường và không theo quy luật, thể hiện thông qua các giá trị nhiễu có trong bộ dữ liệu.

7. KẾT LUẬN

Qua đồ án này, chúng em thành công trong việc phân tích và tìm các vấn đề tìm ẩn có trong một bộ dữ liệu, và cụ thể ở đề tài này là dữ liệu về khách hàng rời bỏ. Nhờ việc áp dụng các kiến thức đã học và sử dụng các công cụ hỗ trợ, chúng em đã hiểu rõ, phân tích, và trực quan được các yếu tố có trong bộ dữ liệu có ảnh hưởng đến mục tiêu của đồ án. Sau khi thiết kế và chạy thực nghiệm, chúng em đã thu được kết quả đúng mong đợi với kết quả tốt nhất đạt được trên trên mô hình Random Forest với độ đo đánh giá accuracy: 88.25% và f1-score: 85.00% và mô hình Gradient Boosting với accuracy: 88.11% và f1-score: 84.75%. Trong quá trình thực hiện đồ án, chúng em đã hiểu rõ hơn về phân tích và trực một bộ dữ liệu, thiết kế thực nghiệm và các kiến thức về máy học. Trong tương lai, chúng em có thể sẽ thực hiện thêm các thử nghiệm trên bộ dữ liệu này như: thử nghiệm thêm các mô hình máy học, áp dụng các thuật toán để xử lý mất cân bằng dữ liệu, các thuật toán giảm chiều dữ liệu cho bài toán phân lớp.

TÀI LIỆU THAM KHẢO

Chú ý: Đây là cách viết TLTK không đúng chuẩn. KHÔNG dùng định dạng này vào khóa luận tốt nghiệp và môn học khác.

[1] Pandora, Link: https://pandora.norbreeze.vn/

[2] Geeksforgeeks – Python | Visualize missing values (NaN) values using Missingno Library, Link: Geeksforgeeks, (04/07/2019)

[3] Geeksforgeeks – Guide to AUC ROC Curve in Machine Learning, Link: Geeksforgeeks, (10/06/2023)

[4] scikit-learn – Machine Learning in Python, Link: scikit-learn

[5] Thư viện SciPy, Link: SciPy

[6] Thư viện Pandas, Link: Pandas

[7] Thư viện NumPy, Link: NumPy

[8] Thư viện Matplotlib, Link: Matplotlib

[9] Thư viện Seaborn, Link: Seaborn

[10] Thư viện Plotly, Link: Plotly

[11] Geeksforgeeks – How to Conduct a Two Sample T-Test in Python, Link: Geeksforgeeks, (17/10/2022)

[12] Geeksforgeeks – Python – Pearson's Chi-Square Test, Link: Geeksforgeeks

[13] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, 10/03/2017.

[14] Daniel Nelson, Data Visualization in Python, 09/2020

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Minh Hiếu	- Thu thập dữ liệu - Phân tích EDA - Train model - Viết báo cáo + Slide *Đánh giá: 45%
2	Nguyễn Thành Nhân	- Phân tích dữ liệu - Phân tích EDA - Tính tương quan + Grid tìm param - Chọn và Train model - Đánh giá model - Viết báo cáo + Slide - Dashboard *Đánh giá: 55%