

PHÂN TÍCH VÀ DỰ ĐOÁN KHẢ NĂNG RỜI BỎ CỦA KHÁCH HÀNG

Đoàn Ngọc Tuấn

NỘI DUNG

01

TỔNG QUAN

02

PHÂN TÍCH

03

**XÂY DỰNG MÔ
HÌNH DỰ ĐOÁN**

04

KẾT LUẬN



NHÓM 10

TỔNG QUAN

PROBLEM STATEMENT

- Các ngân hàng đều muốn giữ chân khách hàng của mình để duy trì hoạt động kinh doanh và ngân hàng Đa quốc gia ABC cũng muốn điều đó. Dưới đây là dữ liệu khách hàng của các khách hàng tại Ngân hàng Đa quốc gia ABC có phát sinh giao dịch và mục đích của dữ liệu sẽ là dự đoán **Tỷ lệ khách hàng rời bỏ**.
- Giả sử bạn là Data Analyst cho ngân hàng ABC. BOD đang cố gắng tìm hiểu xem tại sao lại xảy ra vấn đề trên và liệu người dùng các dịch vụ có rời bỏ ABC hay không (hủy sử dụng dịch vụ) trong vài ngày tới.



DATASET

- Dữ liệu về thông tin của khách hàng có phát sinh giao dịch tại ngân hàng.
- Dữ liệu gồm 1000 dòng, 12 thuộc tính
- Không có giá trị duplicate và null

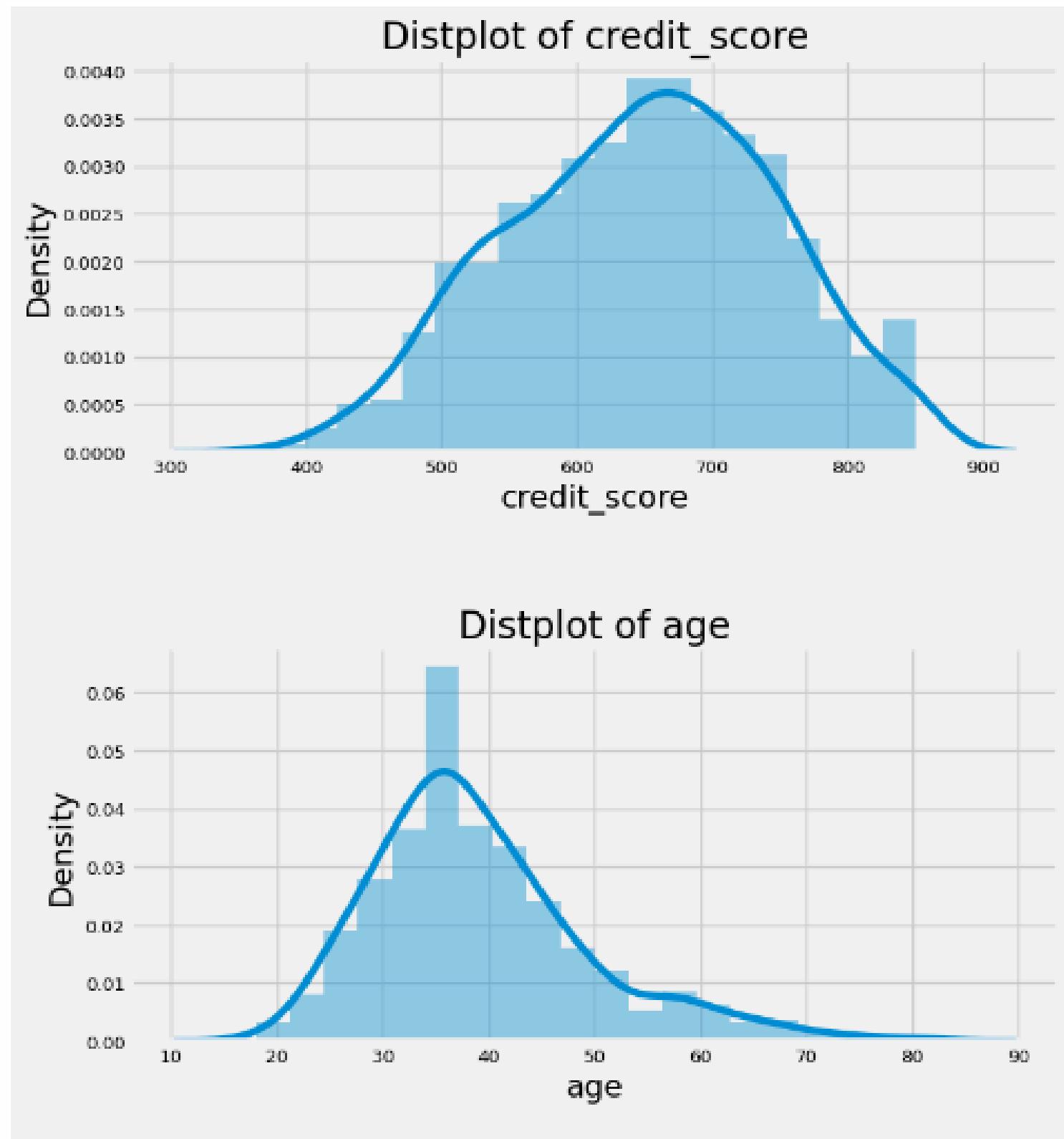
customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
15634602	619	France	Female	42	2	0		1	1	1	10134888
15647311	608	Spain	Female	41	1	8380786		1	0	1	11254258
15619304	502	France	Female	42	8	1596608		3	1	0	11393157
15701354	699	France	Female	39	1	0		2	0	0	9382663
15737888	850	Spain	Female	43	2	12551082		1	1	1	790841
15574012	645	Spain	Male	44	8	11375578		2	1	0	14975671
15592531	822	France	Male	50	7	0		2	1	1	100628
15656148	376	Germany	Female	29	4	11504674		4	1	0	11934688
15792365	501	France	Male	44	4	14205107		2	0	1	749405
15592389	684	France	Male	27	2	13460388		1	1	1	7172573
15767821	528	France	Male	31	6	10201672		2	0	0	8018112
15737173	497	Spain	Male	24	3	0		2	1	0	7639001
15632264	476	France	Female	34	10	0		2	1	0	2626098
15691483	549	France	Female	25	5	0		2	0	0	19085779
15600882	635	Spain	Female	35	7	0		2	1	1	6595165
15643966	616	Germany	Male	45	3	14312941		2	0	1	6432726
15737452	653	Germany	Male	58	1	13260288		1	1	0	509767
15788218	549	Spain	Female	24	9	0		2	1	1	1440641
15661507	587	Spain	Male	45	6	0		1	0	0	15868481



NHÓM 10

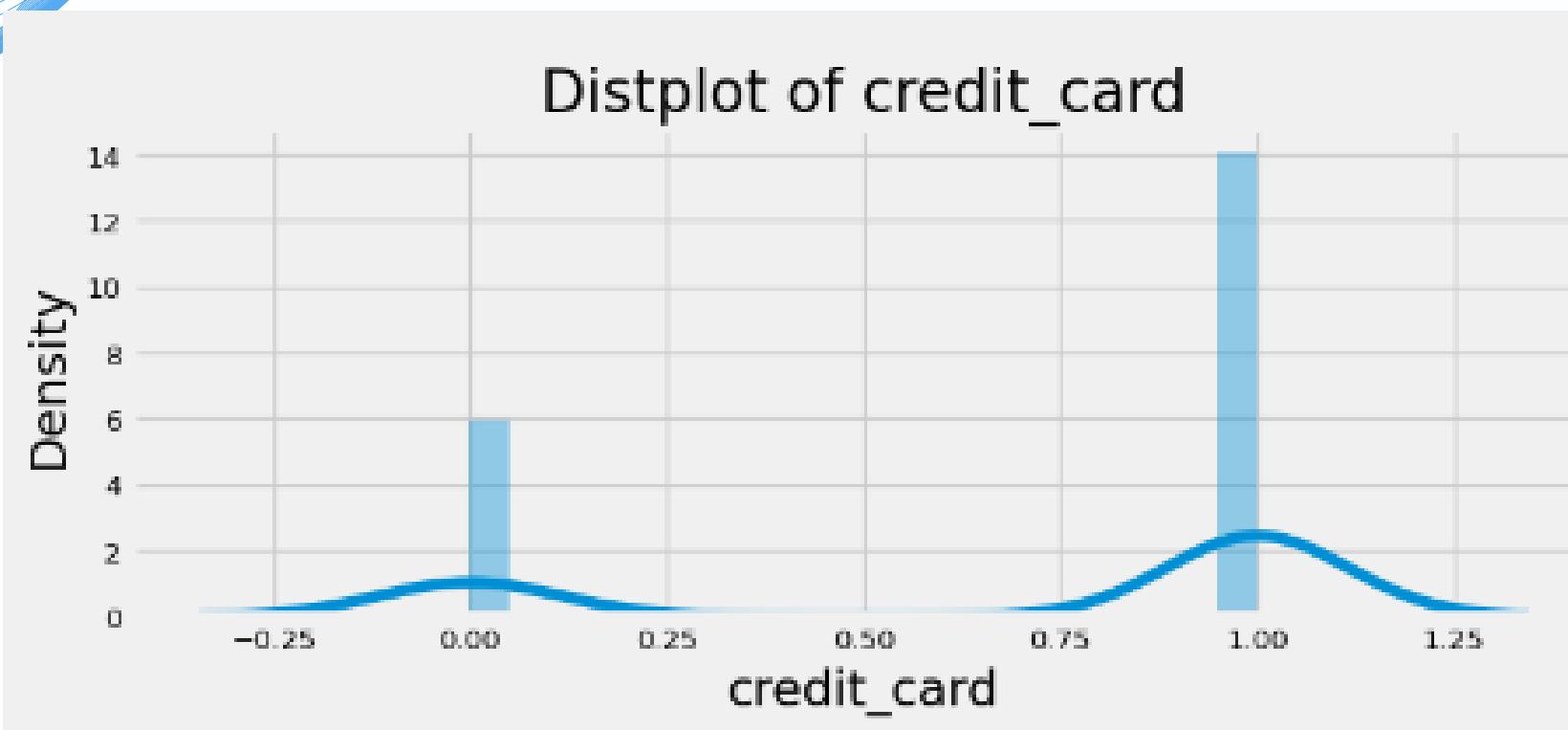
PHÂN TÍCH

PHÂN TÍCH TỔNG QUAN

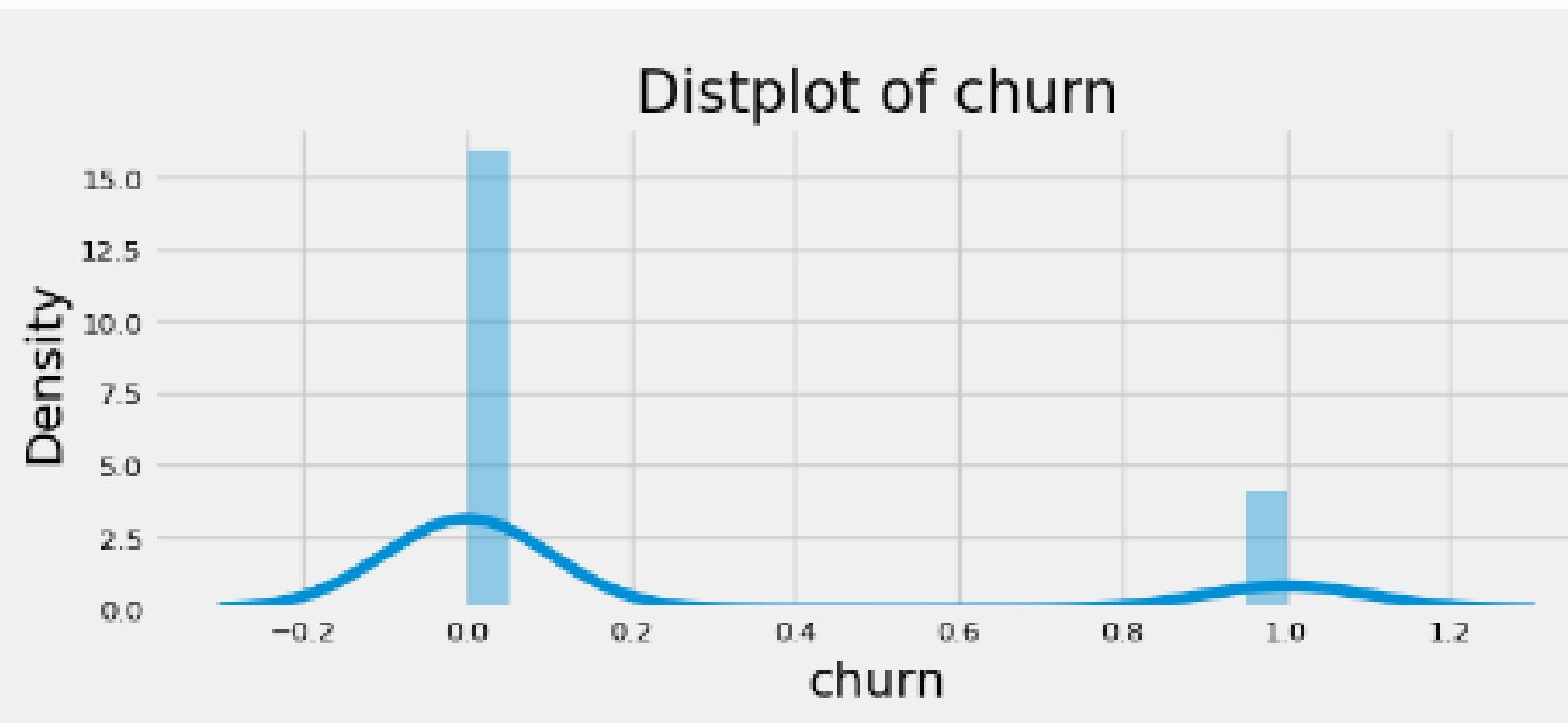


- **Credit Score** của khách hàng tập trung chủ yếu trong khoảng **550-750**.
- **Tuổi** của khách hàng phân bố chủ yếu khoảng từ **30 đến 45**. Và khách hàng có độ tuổi quanh **35** có số lượng rất lớn.

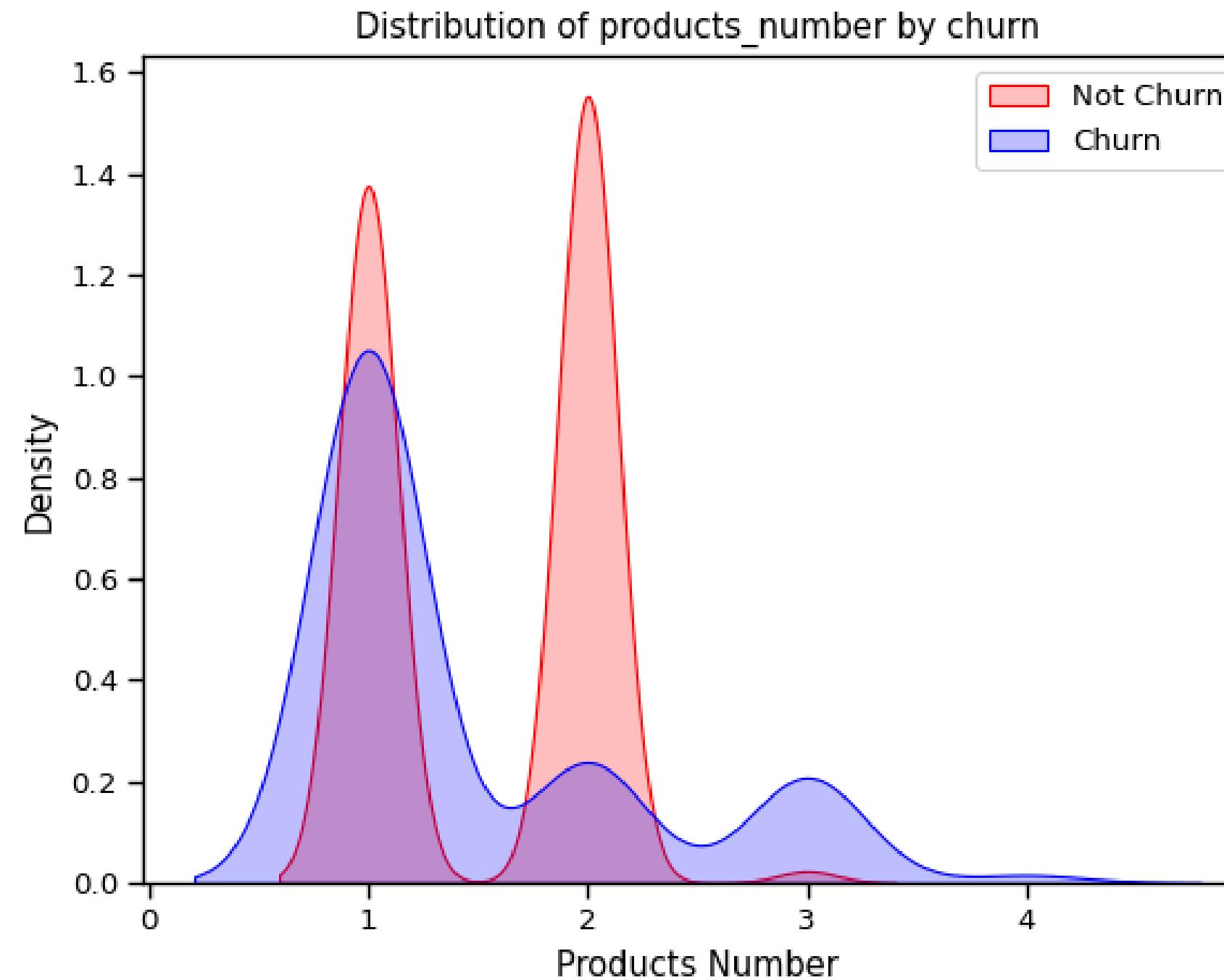
PHÂN TÍCH TỔNG QUAN



- **Đa phần** khách hàng đã có thẻ credit.
- Lượng khách hàng không rời bỏ cao **gần gấp 4** lần lượng khách hàng rời bỏ.



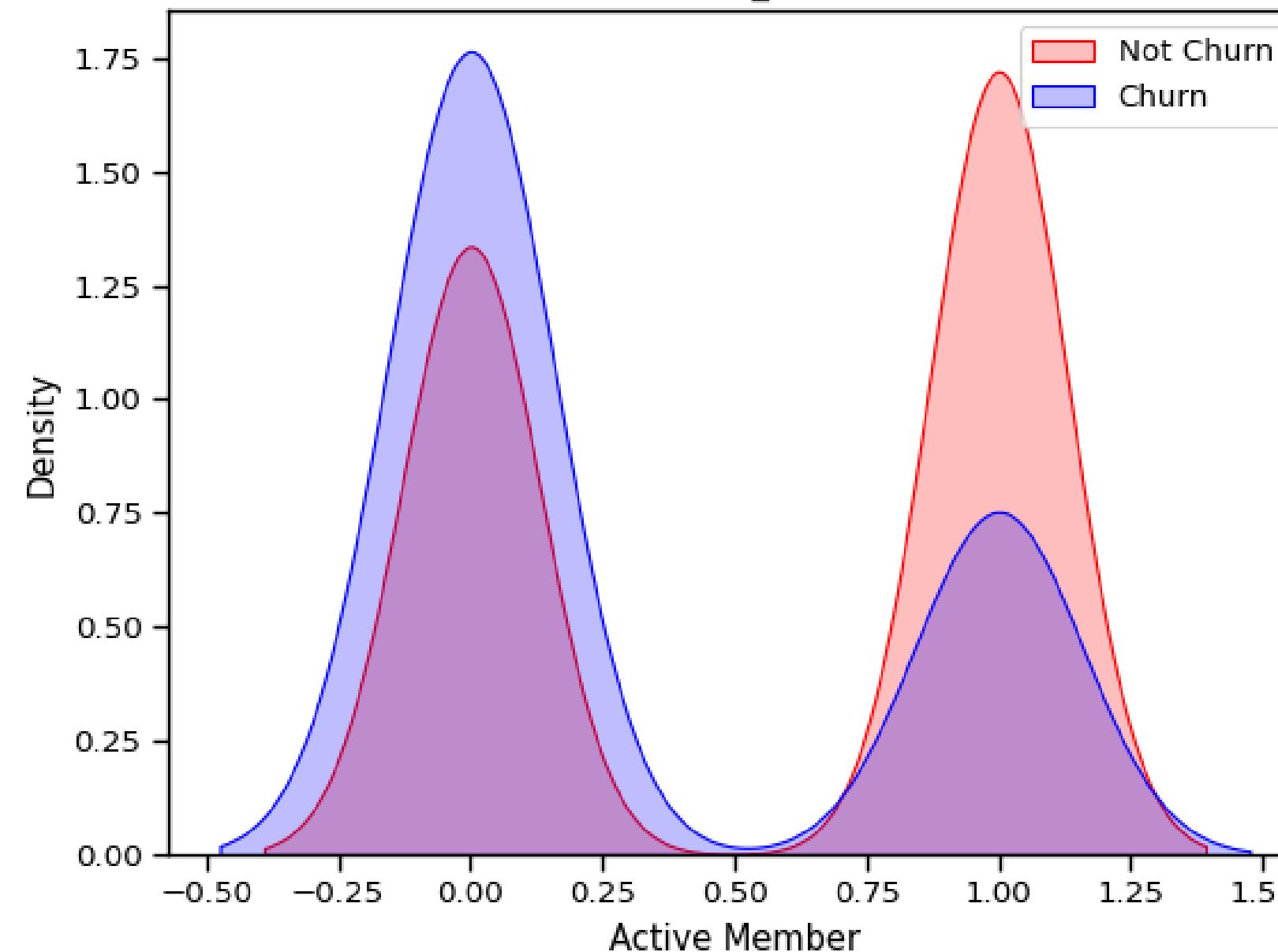
PHÂN TÍCH CHI TIẾT



- Những khách hàng có **products_number = 2** thì **tỉ lệ rời bỏ cực kì thấp**, cao gấp 8 lần lượng khách hàng rời bỏ.
- Đối với những khách hàng có **products number = 1** thì **tỉ lệ rời bỏ cực kì cao**, gần bằng lượng khách hàng không rời bỏ.
- Với khách hàng có **product_numbers = 3** thì tỉ lệ khách hàng rời bỏ **tương đối cao**, cao gấp nhiều lần so với lượng khách hàng không rời bỏ.

PHÂN TÍCH CHI TIẾT

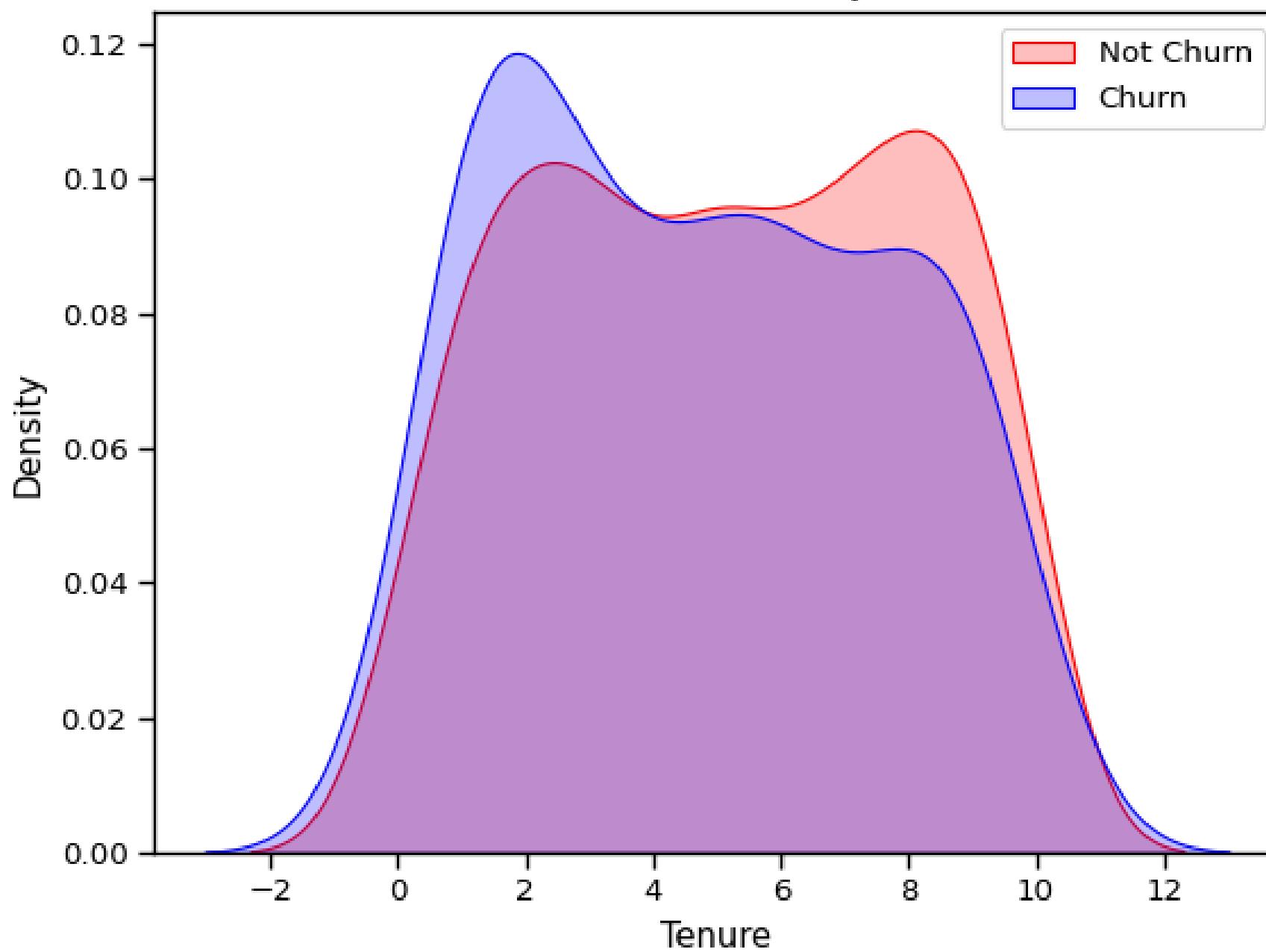
Distribution of active_member by churn



- Những khách hàng **là Member** ($active_member = 1$) thì khả năng rất cao sẽ **không rời bỏ**, tỉ lệ rời bỏ của nhóm khách hàng này rất thấp, chỉ bằng $1/2$ tỉ lệ khách hàng không rời bỏ.
- Những khách hàng **không phải Member** ($active_member = 0$) thì có **tỉ lệ rời bỏ rất cao**, cao hơn hẳn tỉ lệ khách hàng không rời bỏ.

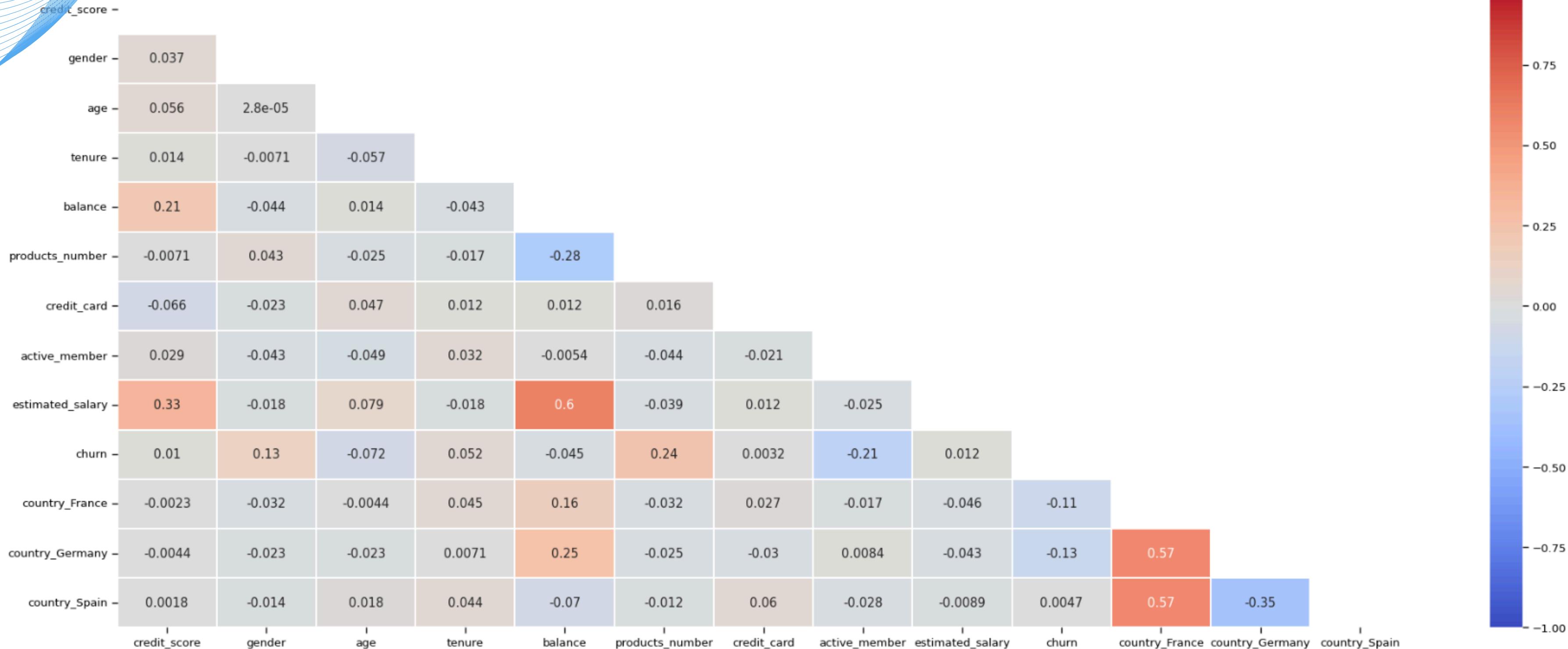
PHÂN TÍCH CHI TIẾT

Distribution of tenure by churn



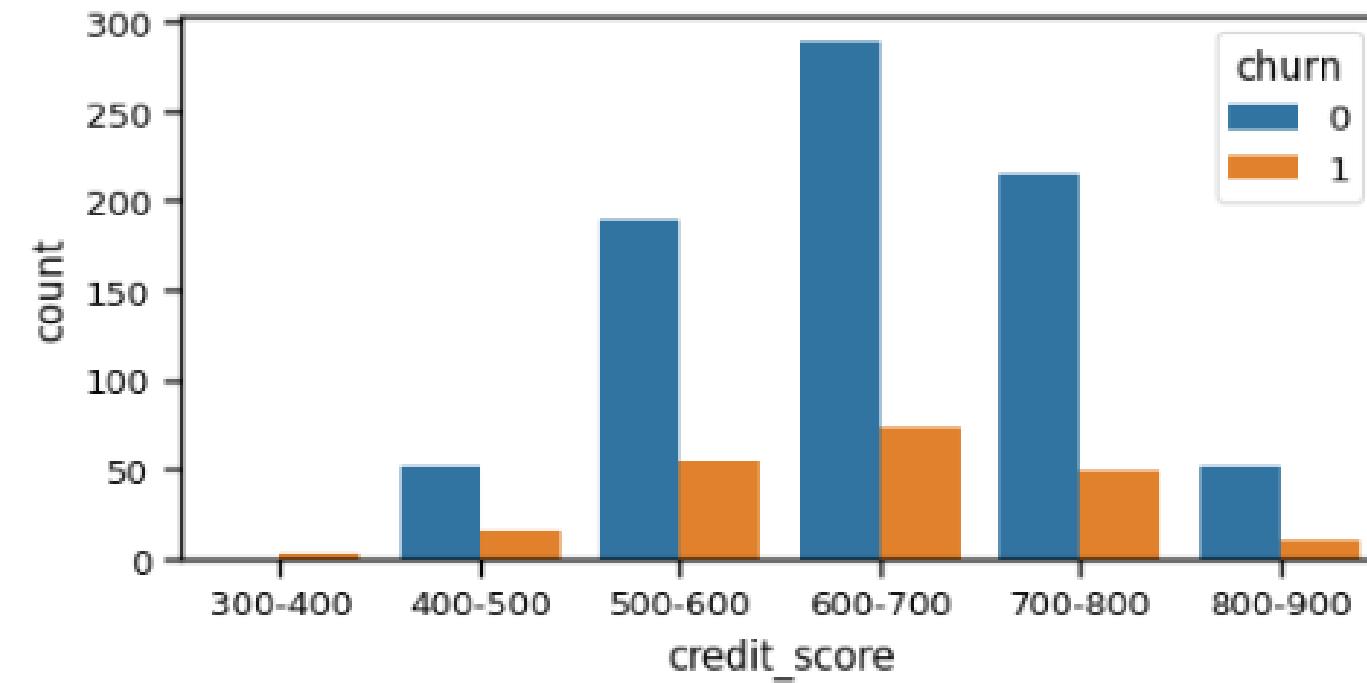
- Những khách hàng mới sử dụng ngân hàng thì có tỉ lệ rời bỏ cao hơn
- Đối với những khách hàng đã sử dụng ngân hàng từ lâu thì khả năng khách hàng rời bỏ sẽ thấp hơn.

PHÂN TÍCH CHI TIẾT

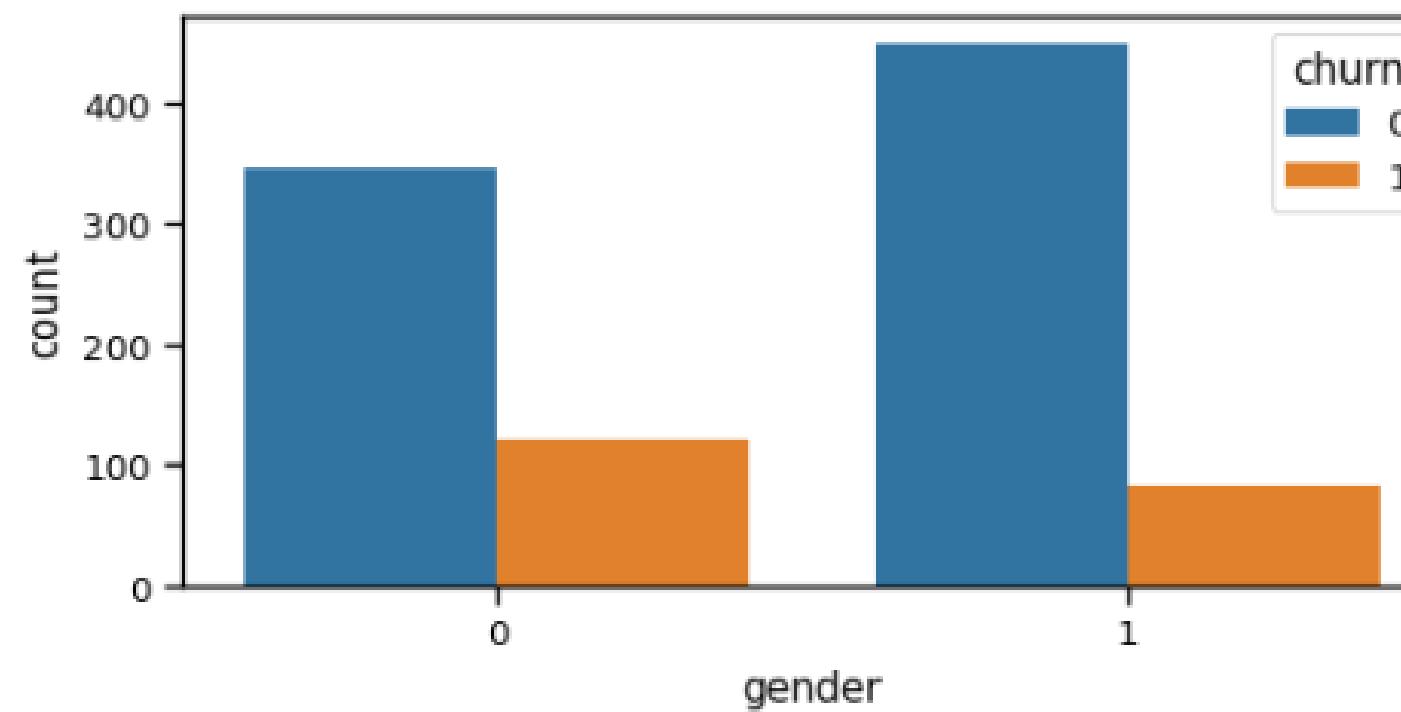


=> Đa phần các thuộc tính có tương quan thấp. Chỉ có balance phụ thuộc tương đối nhiều vào estimated_salary.

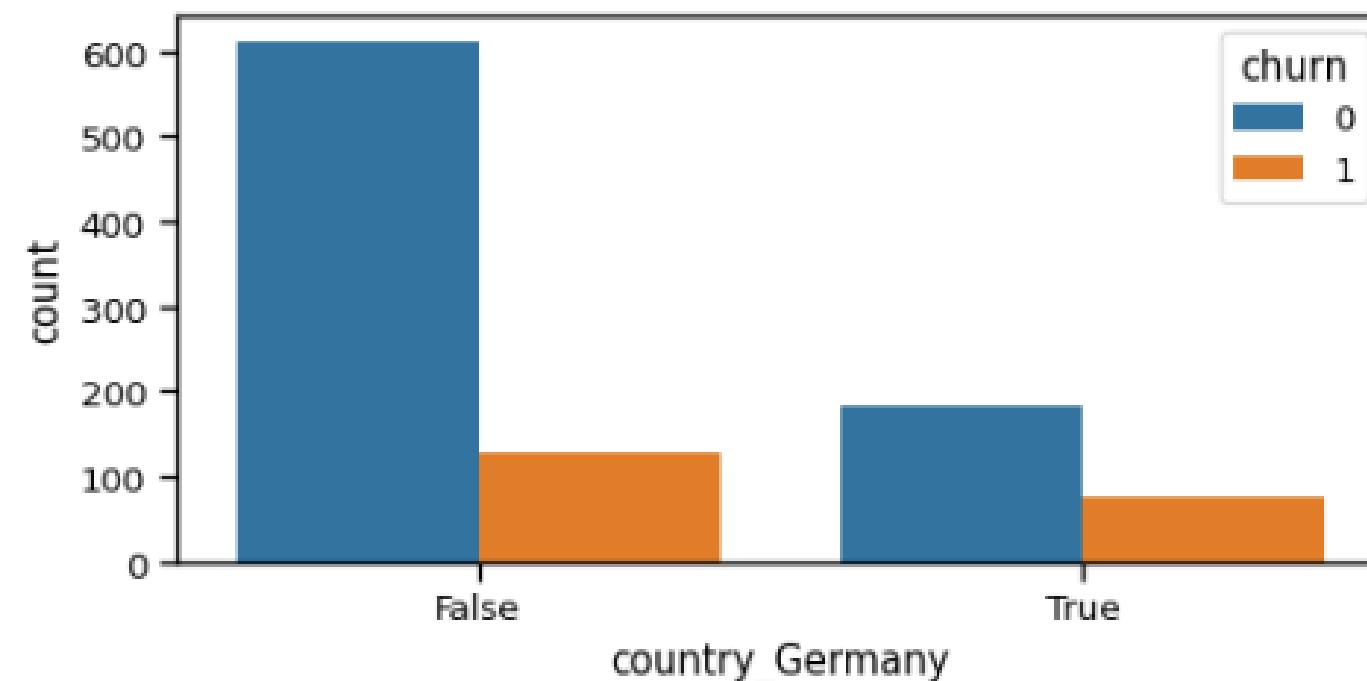
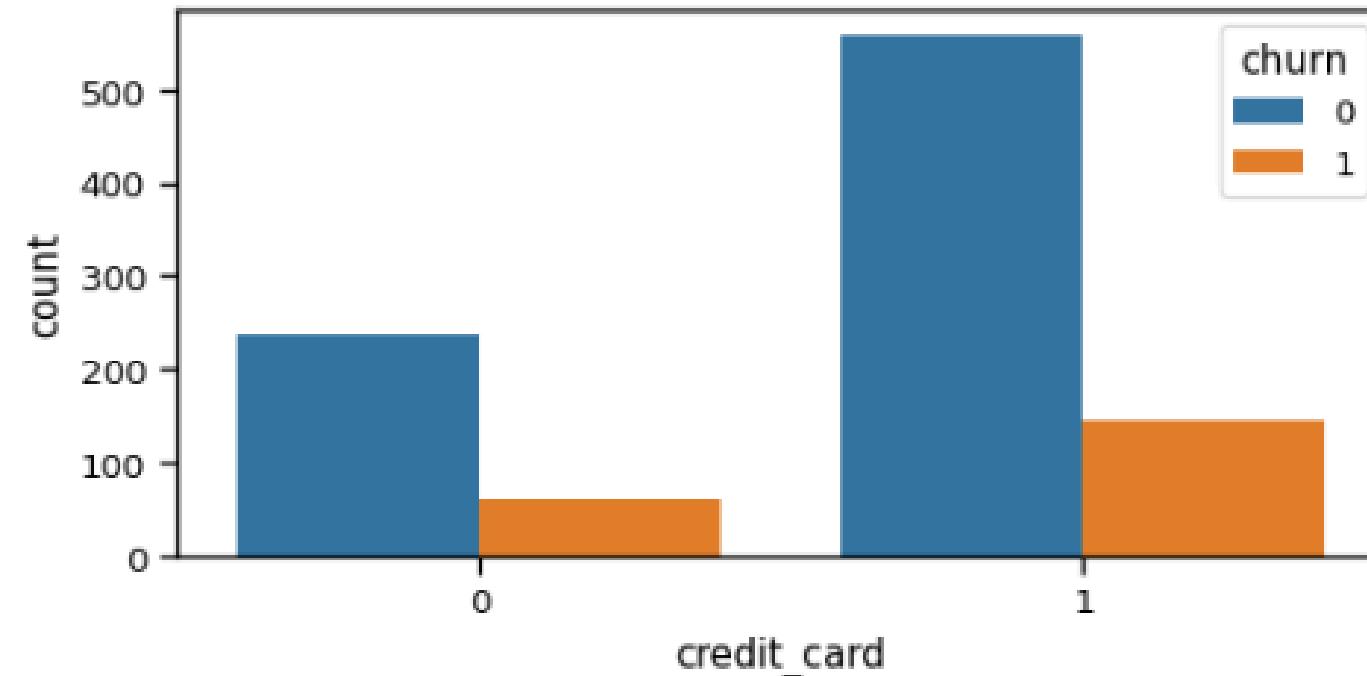
PHÂN TÍCH CHI TIẾT



- Những khách hàng có độ tuổi từ 40-70, có tỉ lệ rời bỏ cực kì cao, gần với tỉ lệ không rời bỏ của khách hàng.
- Những khách hàng nữ có tỉ lệ rời bỏ cao hơn so với những khách hàng nam.



PHÂN TÍCH CHI TIẾT



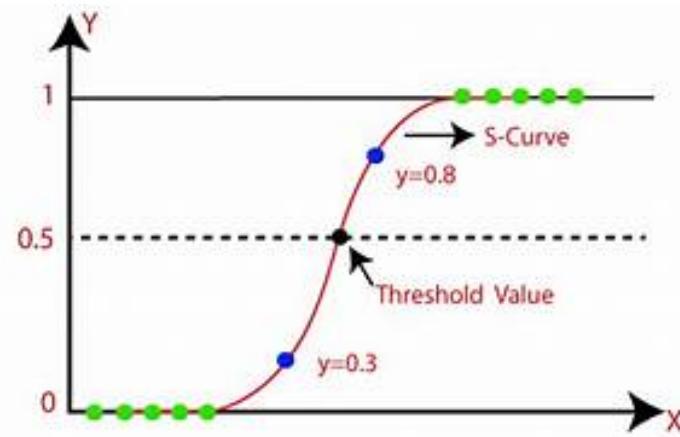
- Những khách hàng cư trú ở Germany có tỉ lệ rời bỏ khá cao, gần bằng 1/2 so với lượng khách hàng không rời bỏ.
- Những khách hàng có credit score thấp thì khả năng rất cao sẽ rời bỏ.



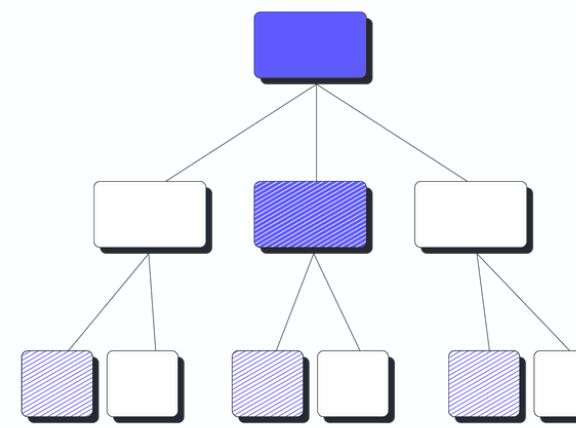
NHÓM 10

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN

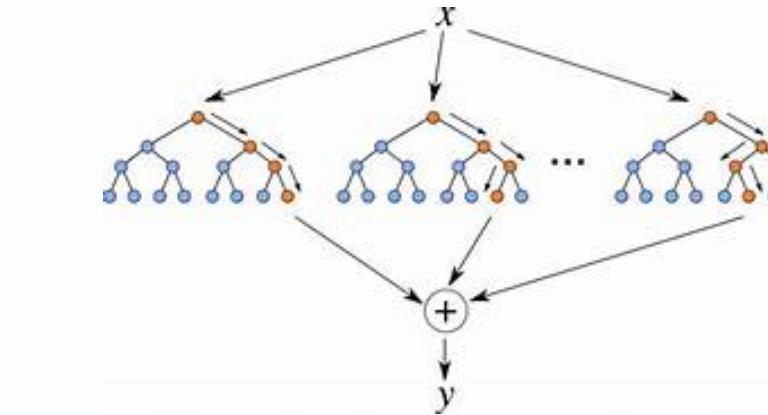
THUẬT TOÁN SỬ DỤNG



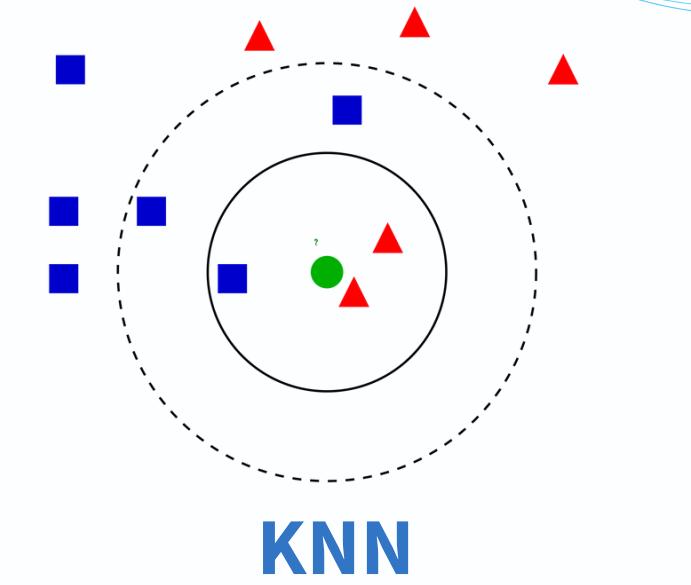
Logistic Regression



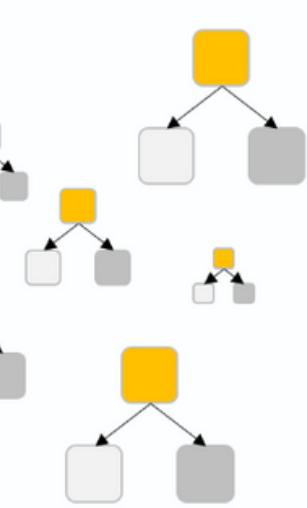
Decision Tree



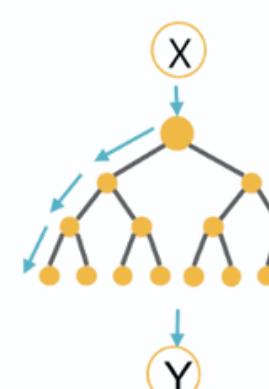
Random Forest



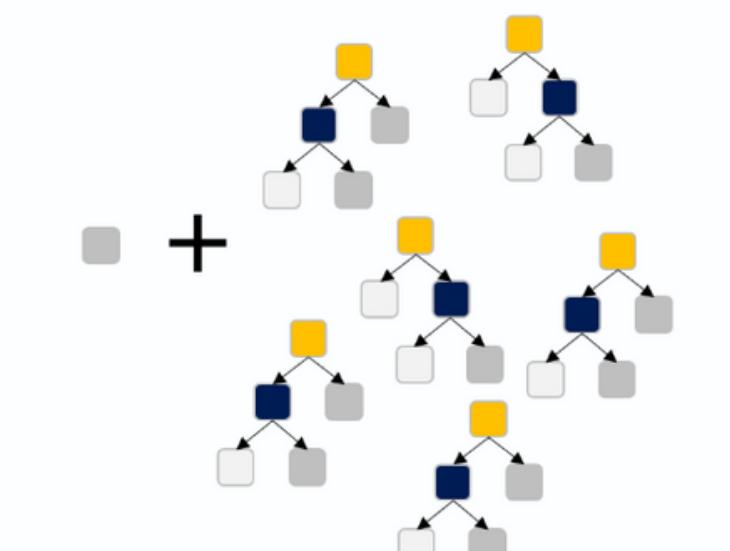
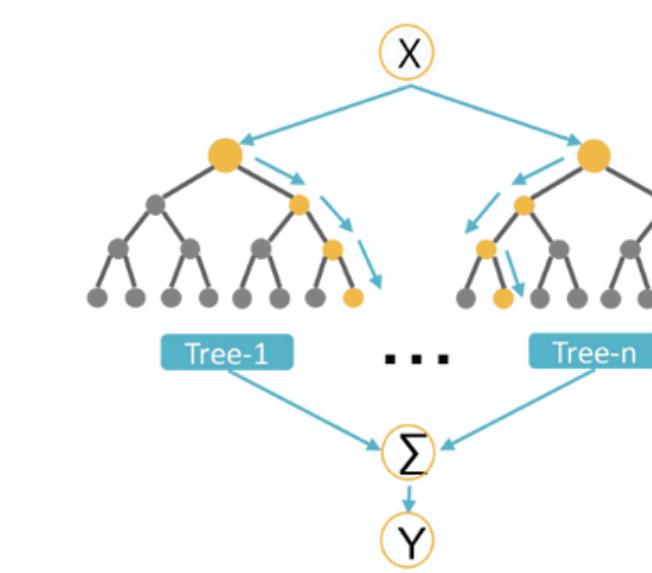
KNN



Ada Boost



XGBoost



Gradient Boosting

THUẬT TOÁN SỬ DỤNG

```
models = [
    LogisticRegression(),
    DecisionTreeClassifier(),
    GridSearchCV(RandomForestClassifier(), param_grid={
        'n_estimators': [10 ,50, 100, 200],
    }),
    GridSearchCV(KNeighborsClassifier(), param_grid={
        'n_neighbors': [10 ,50, 100, 200],
    }),
    GridSearchCV(AdaBoostClassifier(), param_grid={
        'n_estimators': [10 ,50, 100, 200],
    }),
    GridSearchCV(GradientBoostingClassifier(), param_grid={
        'n_estimators': [10 ,50, 100, 200],
    }),
    GridSearchCV(XGBClassifier(), param_grid={
        'n_estimators': [10 ,50, 100, 200],
    }),
]
```

TIỀN XỬ LÝ DỮ LIỆU

Label Encoder và OneHotEncoder những cột dữ liệu dạng chuỗi.

country	gender
France	Female
Spain	Female
France	Female
France	Female
Spain	Female



gender	country_France	country_Germany	country_Spain
0	True	False	False
0	False	False	True
0	True	False	False
0	True	False	False
0	False	False	True

KẾT QUẢ MÔ HÌNH

	precision	recall	f1-score	support
0	0.78	0.94	0.85	156
1	0.25	0.07	0.11	44
accuracy			0.75	200
macro avg	0.52	0.51	0.48	200
weighted avg	0.66	0.75	0.69	200

Logistic Regression

	precision	recall	f1-score	support
0	0.85	0.81	0.83	156
1	0.43	0.50	0.46	44
accuracy			0.74	200
macro avg	0.64	0.66	0.65	200
weighted avg	0.76	0.74	0.75	200

Decision Tree

	precision	recall	f1-score	support
0	0.84	0.99	0.91	156
1	0.88	0.32	0.47	44
accuracy			0.84	200
macro avg	0.86	0.65	0.69	200
weighted avg	0.85	0.84	0.81	200

Random Forest

	precision	recall	f1-score	support
0	0.78	1.00	0.88	156
1	0.00	0.00	0.00	44
accuracy			0.78	200
macro avg	0.39	0.50	0.44	200
weighted avg	0.61	0.78	0.68	200

KNN

KẾT QUẢ MÔ HÌNH

	precision	recall	f1-score	support
0	0.86	0.96	0.91	156
1	0.77	0.45	0.57	44
accuracy			0.85	200
macro avg	0.82	0.71	0.74	200
weighted avg	0.84	0.85	0.83	200

Ada Boost

	precision	recall	f1-score	support
0	0.85	0.97	0.91	156
1	0.82	0.41	0.55	44
accuracy			0.85	200
macro avg	0.84	0.69	0.73	200
weighted avg	0.85	0.85	0.83	200

Gradient Boosting

0	0.85	0.96	0.90	156
1	0.75	0.41	0.53	44
accuracy			0.84	200
macro avg	0.80	0.69	0.72	200
weighted avg	0.83	0.84	0.82	200

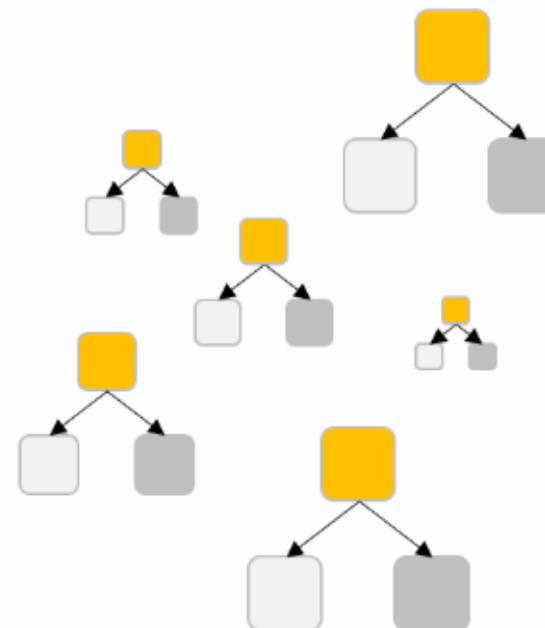
XGBoost

SO SÁNH METRICS

Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
LogisticRegression	0.75	0.78	0.25	0.94	0.07	0.85	0.11
Decision Tree	0.74	0.85	0.43	0.81	0.50	0.83	0.46
Random Forest	0.84	0.84	0.88	0.99	0.32	0.91	0.47
KNN	0.78	0.78	0.00	1.00	0.00	0.88	0.00
AdaBoost	0.85	0.86	0.77	0.96	0.45	0.91	0.57
GradientBoosting	0.85	0.85	0.82	0.97	0.41	0.91	0.55
XGBoost	0.84	0.85	0.75	0.96	0.41	0.90	0.53

KẾT QUẢ MÔ HÌNH

Dựa vào giá trị precision, recall trên lớp 0, 1 và accuracy. => Mô hình được sử dụng cho việc dự đoán những khách hàng có khả năng cao sẽ rời đi với độ chính xác cao nhất là AdaBoostClassifier



	precision	recall	f1-score	support
0	0.86	0.96	0.91	156
1	0.77	0.45	0.57	44
accuracy			0.85	200
macro avg	0.82	0.71	0.74	200
weighted avg	0.84	0.85	0.83	200

Ada Boost

OVERSAMPLING

Sử dụng SMOTE để oversampling dữ liệu.

```
from imblearn.over_sampling import SMOTE  
smote = SMOTE(random_state=42)  
X_train, y_train = smote.fit_resample(X_train, y_train)
```

```
y_train.value_counts()
```

```
churn  
0    640  
1    640  
Name: count, dtype: int64
```

KẾT QUẢ MÔ HÌNH

	precision	recall	f1-score	support
0	0.87	0.60	0.71	156
1	0.33	0.68	0.44	44
accuracy			0.62	200
macro avg	0.60	0.64	0.58	200
weighted avg	0.75	0.62	0.65	200

Logistic Regression

	precision	recall	f1-score	support
0	0.86	0.83	0.84	156
1	0.46	0.52	0.49	44
accuracy			0.76	200
macro avg	0.66	0.67	0.67	200
weighted avg	0.77	0.76	0.77	200

Decision Tree

	precision	recall	f1-score	support
0	0.89	0.89	0.89	156
1	0.61	0.61	0.61	44
accuracy			0.83	200
macro avg	0.75	0.75	0.75	200
weighted avg	0.83	0.83	0.83	200

Random Forest

	precision	recall	f1-score	support
0	0.80	0.73	0.77	156
1	0.28	0.36	0.31	44
accuracy			0.65	200
macro avg	0.54	0.55	0.54	200
weighted avg	0.69	0.65	0.67	200

KNN

KẾT QUẢ MÔ HÌNH

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.89	0.90	0.90	156		0	0.90	0.90	156
1	0.64	0.61	0.63	44		1	0.65	0.64	44
accuracy			0.84	200	accuracy			0.84	200
macro avg	0.77	0.76	0.76	200	macro avg	0.77	0.77	0.77	200
weighted avg	0.84	0.84	0.84	200	weighted avg	0.84	0.84	0.84	200

Ada Boost

	precision	recall	f1-score	support
0	0.89	0.90	0.89	156
1	0.63	0.61	0.62	44
accuracy			0.83	200
macro avg	0.76	0.76	0.76	200
weighted avg	0.83	0.83	0.83	200

Gradient Boosting

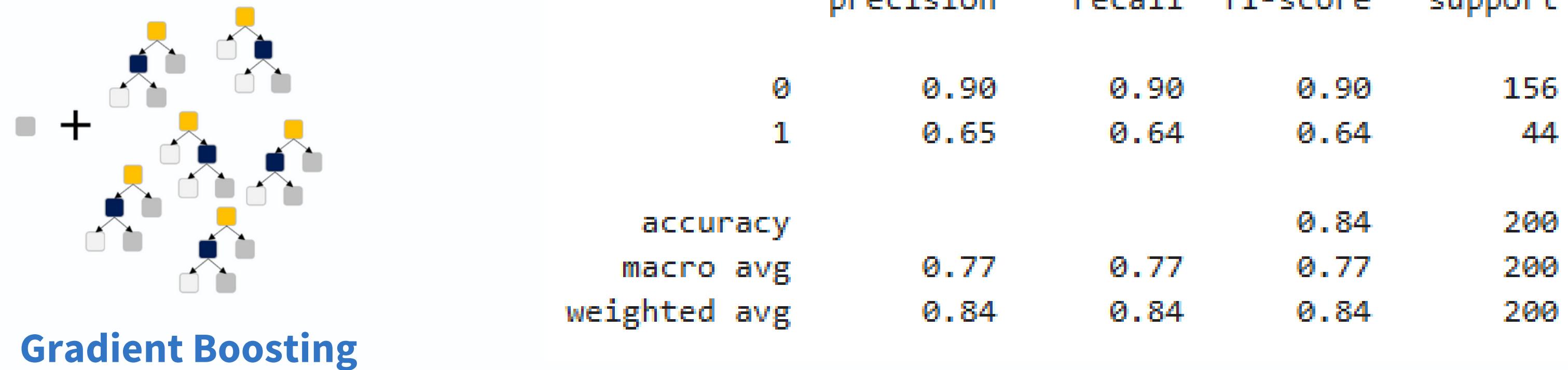
XGBoost

SO SÁNH METRICS

Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
LogisticRegression	0.62	0.87	0.33	0.6	0.68	0.71	0.44
Decision Tree	0.76	0.86	0.46	0.83	0.52	0.84	0.49
Random Forest	0.83	0.89	0.61	0.89	0.61	0.89	0.61
KNN	0.65	0.80	0.28	0.73	0.26	0.77	0.31
AdaBoost	0.84	0.89	0.64	0.90	0.61	0.90	0.63
GradientBoosting	0.84	0.90	0.65	0.90	0.64	0.90	0.64
XGBoost	0.83	0.89	0.64	0.90	0.61	0.89	0.62

KẾT QUẢ MÔ HÌNH

- Khi thực hiện UnderSampling bộ dữ liệu thì ta thấy các mô hình có sự cải thiện rõ rệt khi dự đoán trên lớp 1.
- Dựa vào giá trị precision, recall trên lớp 0, 1 và accuracy. => Mô hình được sử dụng cho việc dự đoán những khách hàng có khả năng cao sẽ rời đi với độ chính xác cao nhất là Gradient Boosting.





NHÓM 10

KẾT LUẬN

NGUYÊN NHÂN KHÁCH HÀNG RỜI BỎ

- Những khách hàng không là thành viên (Member) thì không có những đãi ngộ tốt hơn về các dịch vụ từ phía ngân hàng, nên khả năng rời bỏ rất cao.**
- Khách hàng sử dụng ít sản phẩm, dịch vụ từ ngân hàng thì khả năng rời bỏ cực kì cao. Có thể do còn nhiều sản phẩm, dịch vụ chưa tốt nên khách hàng chưa muốn trải nghiệm.**
- Những khách hàng có credit score thấp hẵu như sẽ rời bỏ dịch vụ, có thể do credit score ảnh hưởng tới khả năng vay vốn và lãi suất.**
- Nhóm khách hàng có độ tuổi từ 40-70 có tỉ lệ rời bỏ khách hàng cao hơn nhiều so với những nhóm khác.**

- **Tạo ra những chương trình đãi ngộ, khuyến khích khách hàng trở thành thành viên Member.**
- **Nâng cao chất lượng chăm sóc khách hàng, đặc biệt đối với tệp khách hàng có độ tuổi từ 40-70.**
- **Khuyến khích khách hàng sử dụng thêm sản phẩm, dịch vụ của ngân hàng.**



**THANK YOU FOR
YOUR ATTENTION!**