

DỰ ĐOÁN KHÁCH HÀNG NGƯỜNG SỬ DỤNG DỊCH VỤ



OVERVIEW

01

Mô tả dataset

02

Mô tả dữ liệu

03

Xử lý dữ liệu

04

EDA

05

Lựa chọn model

06

Kết luận

MÔ TẢ DATASET



Nguồn dataset từ Kaggle:

<https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset/data>



Dataset bao gồm :

400k dòng

12 cột



MÔ TẢ DATASET

	CustomerID	Age	Gender	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
0	2.0	30.0	Female	39.0	14.0	5.0	18.0	Standard	Annual	932.00	17.0	1.0
1	3.0	65.0	Female	49.0	1.0	10.0	8.0	Basic	Monthly	557.00	6.0	1.0
2	4.0	55.0	Female	14.0	4.0	6.0	18.0	Basic	Quarterly	185.00	3.0	1.0
3	5.0	58.0	Male	38.0	21.0	7.0	7.0	Standard	Monthly	396.00	29.0	1.0
4	6.0	23.0	Male	32.0	20.0	5.0	8.0	Basic	Monthly	617.00	20.0	1.0

CustomerID: Mã định danh duy nhất cho mỗi khách hàng

Age: Độ tuổi của khách hàng

Gender: Giới tính của khách hàng

Tenure: Khoảng thời gian tính bằng tháng mà khách hàng đã sử dụng sản phẩm hoặc dịch vụ của công ty

Usage Frequency: Số lần khách hàng đã sử dụng dịch vụ của công ty trong tháng qua

Support Calls: Số cuộc gọi mà khách hàng đã thực hiện tới bộ phận hỗ trợ khách hàng trong tháng qua

Payment Delay: Số ngày khách hàng chậm thanh toán trong tháng trước

Subscription Type: Loại thuê bao khách hàng đã chọn

Contract Length: Thời hạn hợp đồng mà khách hàng đã ký với công ty

Total Spend: Tổng số tiền khách hàng đã chi cho sản phẩm hoặc dịch vụ của công ty

Last Interaction: Số ngày kể từ lần tương tác cuối cùng mà khách hàng có với công ty

Churn: Nhãn nhị phân cho biết khách hàng đã rời bỏ (1) hay chưa (0)

XỬ LÝ DỮ LIỆU

Kiểm tra null

Drop null

Encoding

[] df.isna().sum()	
Age	1
Gender	1
Tenure	1
Usage Frequency	1
Support Calls	1
Payment Delay	1
Subscription Type	1
Contract Length	1
Total Spend	1
Last Interaction	1
Churn	1
dtype:	int64

XỬ LÝ DỮ LIỆU

Kiểm tra null

Drop null

Encoding

	Age	Gender	Tenure	Usage	Frequency	Support	Calls	Payment	Delay
199295	NaN	NaN	NaN			NaN		NaN	NaN

```
df = df.drop(199295)
df = df.reset_index(drop = True)
```

XỬ LÝ DỮ LIỆU

Kiểm tra null

```
[ ] df = pd.get_dummies(df)  
display(df.head(3))
```

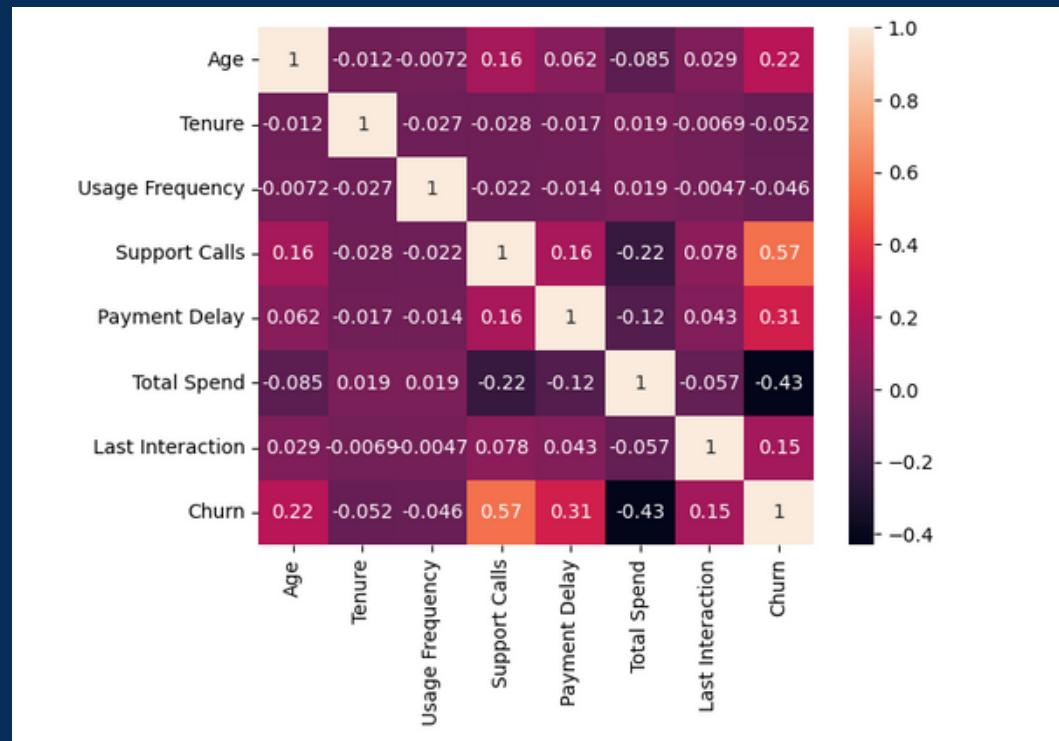
Drop null

	Age	Tenure	Usage Frequency	Support Calls	Payment Delay	Total Spend	Last Interaction	Churn	Gender_Female	Gender_Male
0	30.0	39.0	14.0	5.0	18.0	932.0	17.0	1.0	1	0
1	65.0	49.0	1.0	10.0	8.0	557.0	6.0	1.0	1	0
2	55.0	14.0	4.0	6.0	18.0	185.0	3.0	1.0	1	0

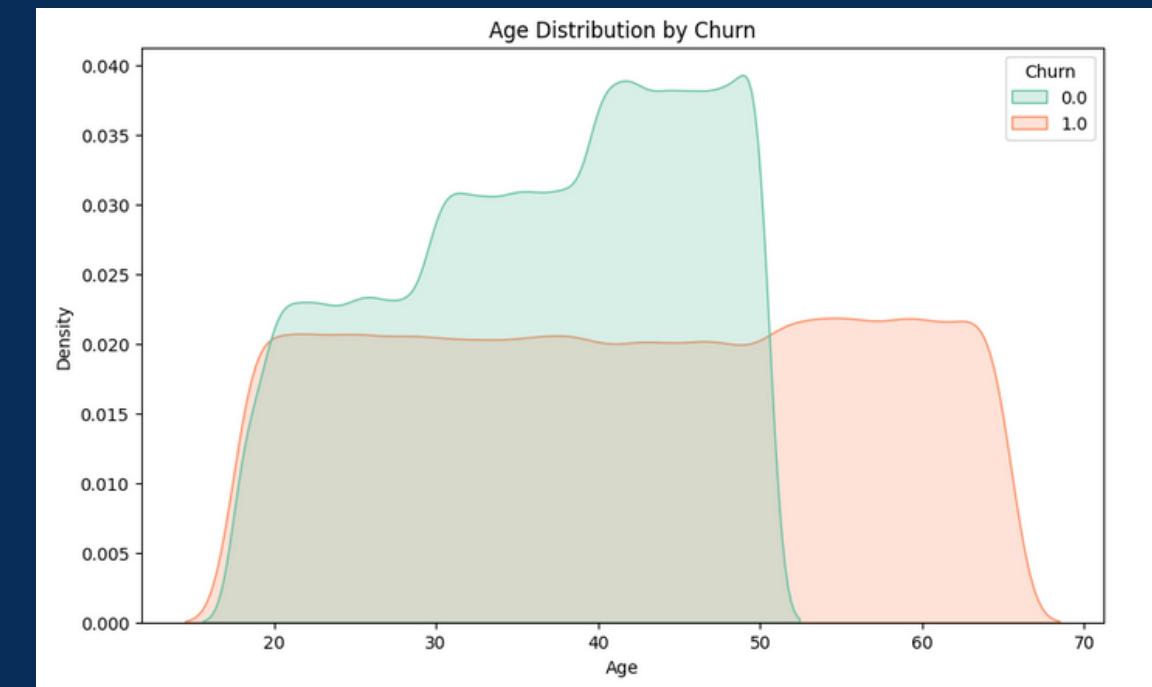
Encoding

EDA

Độ tuổi



Độ tuổi, cuộc gọi hỗ trợ, và thanh toán trễ là những yếu tố có liên quan nhất đến việc rời bỏ dịch vụ

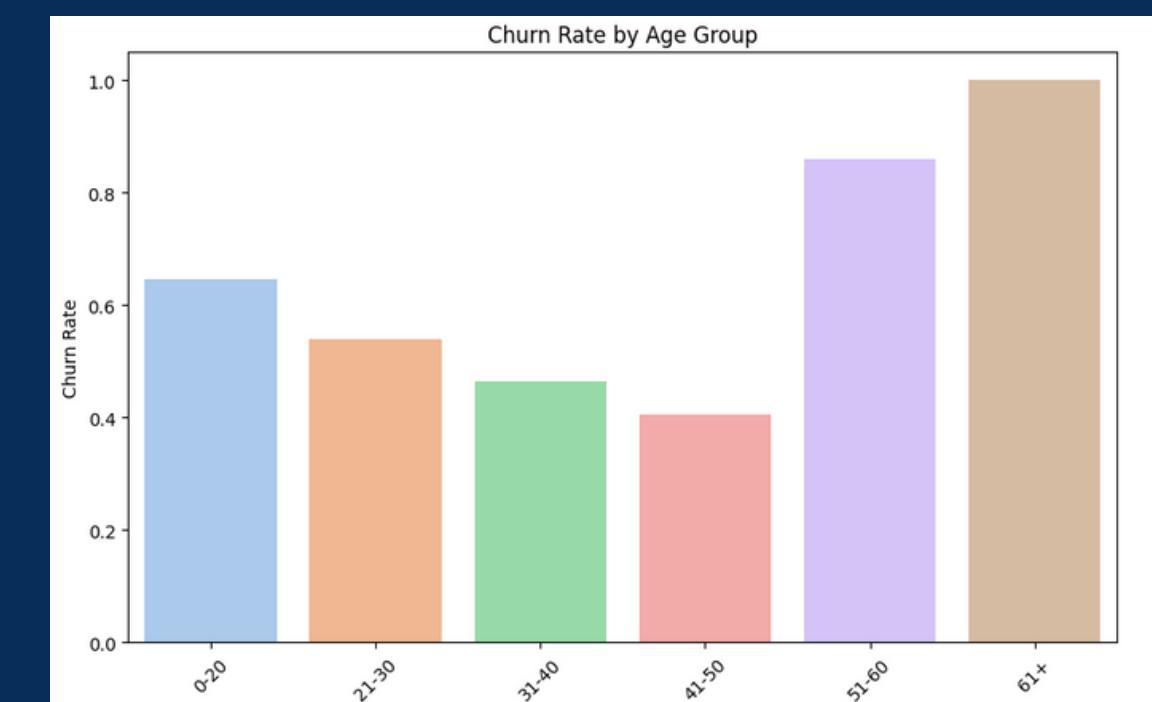


Độ tuổi có phân phối chuẩn nhưng hơi lệch bên trái → có nhiều khách hàng trẻ

Chúng tôi có một lượng lớn khách hàng ở độ tuổi 40-50, với một nhóm nhỏ hơn nhưng vẫn đáng kể ở độ tuổi 20-30

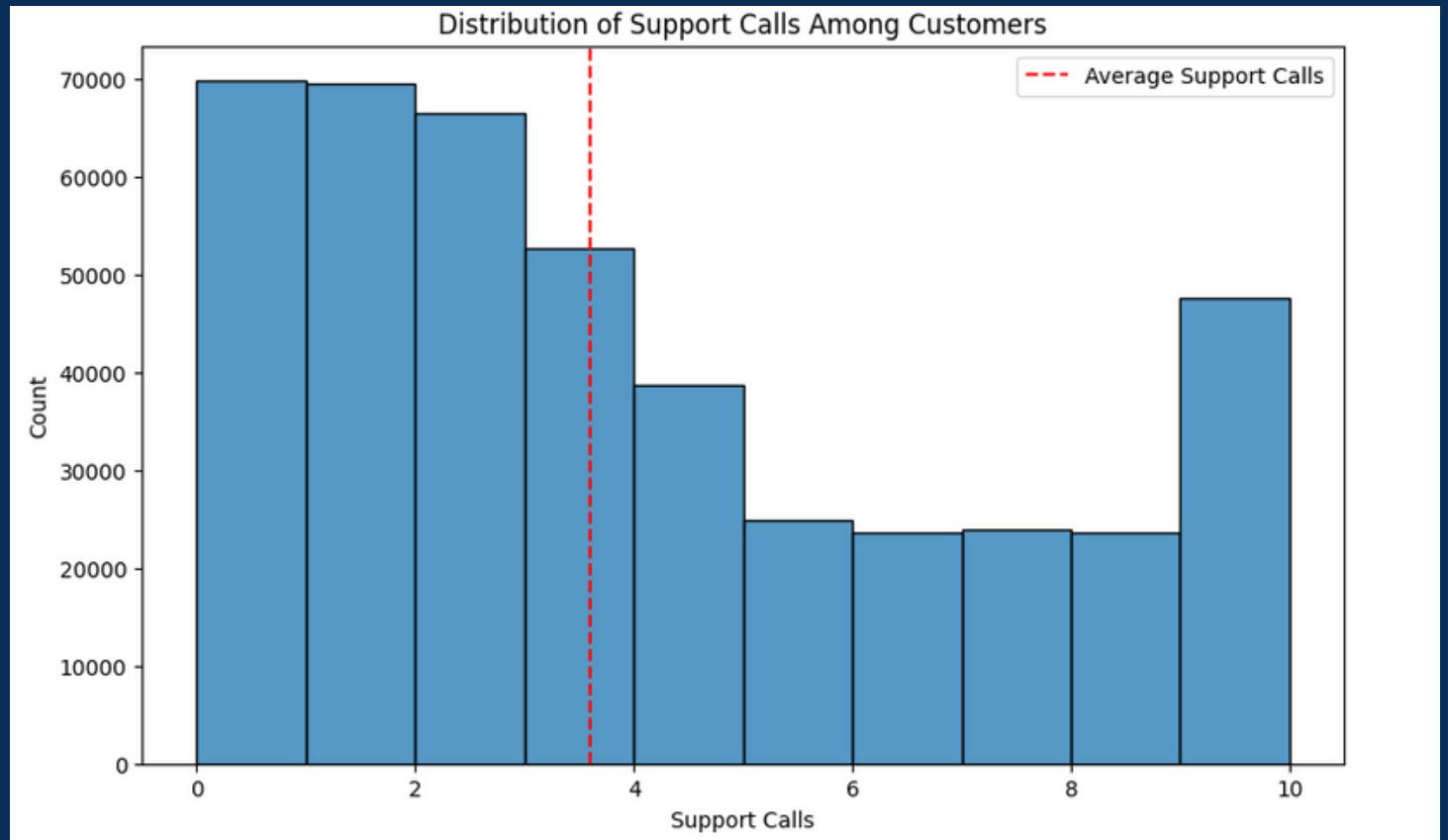
Gần một nửa số khách hàng ở nhóm 20-30 rời bỏ nhưng tỷ lệ rời bỏ giảm ở các nhóm tuổi tiếp theo cho đến nhóm tuổi 41-50

Từ độ tuổi 52 trở đi đều có xu hướng rời bỏ dịch vụ



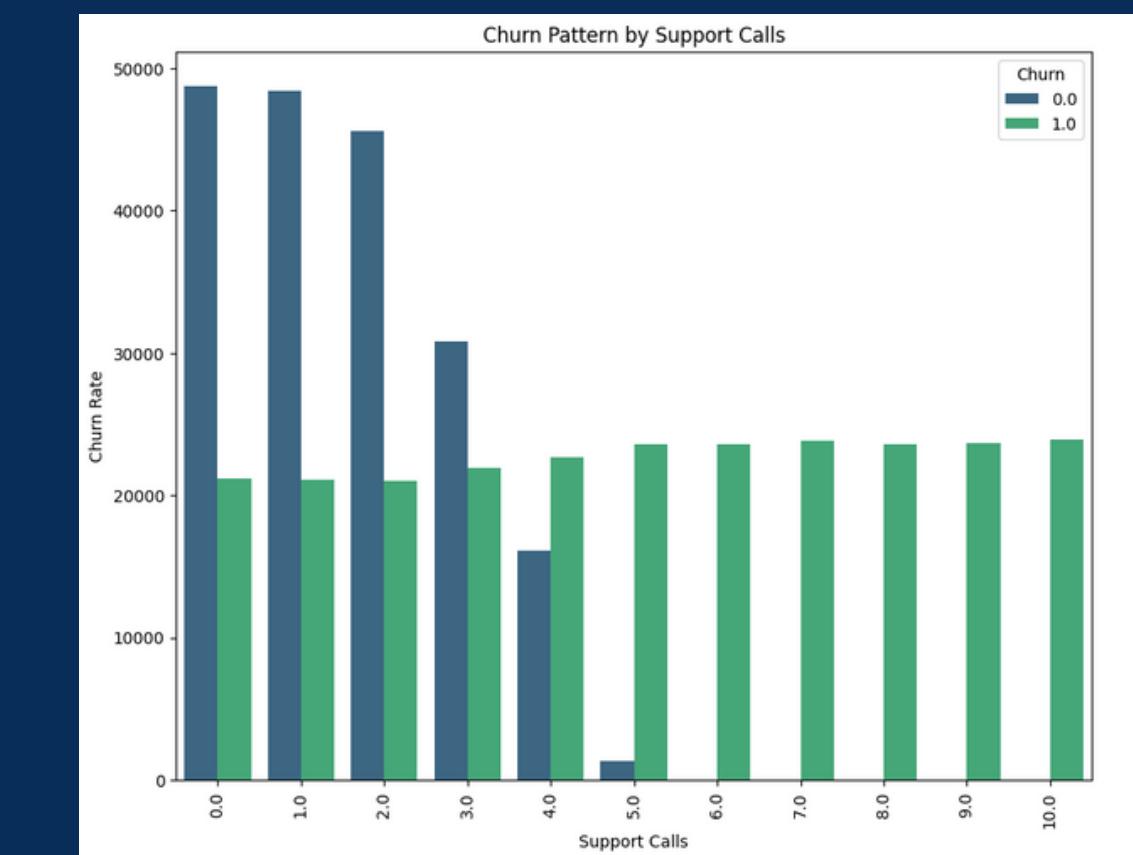
EDA

Phân tích sơ bộ



Trung bình khách hàng gọi hỗ trợ 3 lần

Khi các cuộc gọi hỗ trợ tăng cao thì tỷ lệ rời bỏ cũng cao hơn, đến mức trong hơn 5 cuộc gọi, không có khách hàng nào không rời bỏ và có vẻ như những người hỗ trợ không thành công trong việc giải quyết vấn đề của khách hàng



-> Gọi càng nhiều thì có nhiều vấn đề với dịch vụ nên khách hàng đã rời bỏ

LỰA CHỌN MÔ HÌNH PHÂN TÍCH VÀ KẾT QUẢ

Chọn mô hình: sử dụng các mô hình như sau

- `DecisionTreeClassifier`
- `RandomForestClassifier`
- `KNeighborsClassifier`
- `SVC`

Chia tập train-test theo tỷ lệ 80:20

```
X = df.drop(columns = ["Churn"])
y = df["Churn"]
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(x_train.shape, x_test.shape)
```

Kết quả:

Model	Accuracy	Precision		F1-score		Recall	
		Yes	No	Yes	No	Yes	No
KNN	96%	99%	92%	93%	95%	96%	99%
SVC	98%	100%	96%	98%	98%	97%	100%
RanForest	98%	100%	96%	98%	98%	97%	100%
DeTree	98%	100%	96%	98%	98%	97%	100%

Các model đều cho kết quả khá cao chỉ có model KNN thấp hơn 1 tí
→ có thể sử dụng

ĐÁNH GIÁ VÀ KẾT LUẬN

1.Với tập dữ liệu này thì mô hình sẽ đạt kết quả dự đoán tốt khi sử dụng tất cả các biến để dự đoán.

2. Kết hợp với việc phân tích dữ liệu ban đầu(EDA) để nhận diện được KH tiềm năng và nâng cao kết quả của dịch vụ trong tương lai:

- Tập trung vào nhóm KH có khả năng ở lại cao hơn:
 - > Dựa vào tt cá nhân: Nhóm KH trung niên có xu hướng ở lại dịch vụ cao
 - > Dựa vào tt thanh toán : Nhóm KH thanh toán sớm hoặc tổng chi của KH càng cao (>650\$) thì tỷ lệ ở lại cao
- Chất lượng cuộc gọi:
 - > Số lượng: dưới 5 cuộc có tỷ lệ ở lại cao



Thank's For Watching

