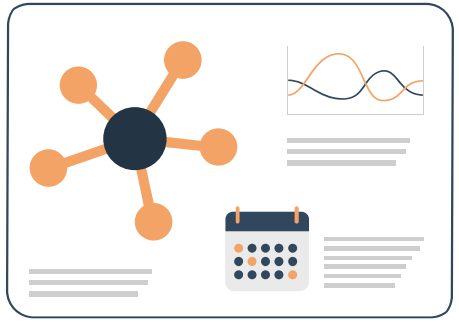
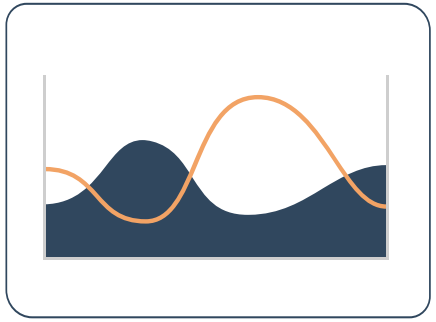
Báo cáo đồ án
Ứng dụng dữ liệu lớn





20120055 Nguyễn Thế Đạt

20120084 Nguyễn Văn Hiếu

20120085 Trần Xuân Hòa

20120113 Lê Nguyên Khang

Mô tả

Mô tả

Mục tiêu:
Phân tích nhu cầu, hành vi của khách hàng.

- Thói quen mua sắm, sở thích, ...

Xem xét các yếu tố ảnh hưởng đến sự hài lòng của khách hàng.

- Chất lượng sản phẩm, giá tiền, ...

Tìm ra những khách hàng **có khả năng cao** sẽ tiếp tục mua sản phẩm.
⟶ Xây dựng model dự đoán khách hàng tiềm năng.

Mô tả

Dữ liệu được thu thập về từ **tiki** trong mục **nhà sách tiki.**
Gồm hai bộ dữ liệu: **products** và **reviews**.

- Dữ liệu sách tiki từ **2012 - 14/12/2023**

products: thông tin về sản phẩm sách trong nhà sách tiki.
Dữ liệu có **2030** dòng và **13** cột.

reviews: đánh giá về sản phẩm được đề cập trong products.
Dữ liệu có **605259** dòng và **10** cột.

Mô tả

reviews	
id	mã đánh giá
product_id	mã sản phẩm
rating	đánh giá
content	nội dung đánh giá
author_name	tên tác giả
title	tiêu đề
created_at	thời gian đánh giá
customer_id	mã khách hàng
customer_name	tên khách hàng
thank_count	số lượng lượt thích đánh giá
thank_count	mã sản phẩm của người bán

Mô tả

product	
id	mã sản phẩm
name	tên sản phẩm
price	giá sản phẩm
original_price	giá niêm yết
discount_rate	tỉ lệ giảm giá
quantity_sold	số lượng đã bán
rating_average	đánh giá trung bình
review_count	số lượng đánh giá
seller_id	mã người bán
category	thể loại
seller_name	tên người bán
author_name	tác giả
spid	tên người bán

Tiền xử lý dữ liệu

Tiền xử lý dữ liệu

Duplicated:

- Qua kiểm tra thì có 21 dòng trùng trong tập products. Các dòng này có thể trùng trong quá trình thu thập dữ liệu.

```
products.duplicated().sum()
```

```
21
```

- Và ở tập dữ liệu **products** có cột: **spid** và **id** đều là mã sản phẩm nên nhóm xóa cột **spid**.

Tiền xử lý dữ liệu

Duplicated:

- Ngoài ra sau khi kiểm tra thì nhóm phát hiện 1 số dòng chỉ khác nhau ở 1 cột duy nhất có thể là thu thập.

	id	created_at	rating	title	content	thank_count	customer_name	customer_id	product_id
561250	14779771	1643853610	5	Cực kì hài lòng	NaN	3	Nguyễn Long	12536047	146223395
561252	14779771	1643853610	5	Cực kì hài lòng	NaN	2	Nguyễn Long	12536047	146223395
529064	16392632	1653012913	5	Cực kì hài lòng	NaN	0	Trần Dương Minh Quang	7603930	67991600
529065	16392632	1653012913	5	Cực kì hài lòng	NaN	0	Khách Hàng	7603930	67991600

Nhóm sẽ xóa các dòng có index là 561250 và 529065.

Tiền xử lý dữ liệu

Missing Values:

- Cột **seller_name** là bị thiếu dữ liệu.

seller_name ⟶ **seller_name** (nếu đã từng mua).

seller_name ⟶ **seller_id** (nếu chưa từng mua).

author_name ⟶ **'Unknown'**

```
products.isnull().sum()
```

```
id                0
name              0
price             0
original_price    0
discount_rate     0
quantity_sold     0
rating_average    0
review_count      0
seller_id         0
category          0
seller_name       2
author_name     374
dtype: int64
```

Tiền xử lý dữ liệu

Data Type:

- Chuyển đổi dữ liệu các cột, thêm các cột mới.

```
products.apply(open_object_dtype)
```

```
id                              {<class 'int'>}
name                            {<class 'str'>}
price                           {<class 'int'>}
original_price                  {<class 'int'>}
discount_rate                   {<class 'int'>}
quantity_sold                   {<class 'int'>}
rating_average                {<class 'float'>}
review_count                    {<class 'int'>}
seller_id                       {<class 'int'>}
category                        {<class 'str'>}
seller_name       {<class 'str'>, <class 'int'>}
author_name                     {<class 'str'>}
dtype: object
```

Đổi tên → **Product_id**

kiểu dữ liệu → **str**

Tiền xử lý dữ liệu

```
products['author_name'] = products['author_name'].str.replace('Choi Kwanghuyn', 'Choi Kwanghyun')
products['author_name'] = products['author_name'].str.replace('Song Hong Binh', 'Song Hong Bing')
products['author_name'] = products['author_name'].str.replace('Rhowa Byrne', 'Rhonda Byrne')
products['author_name'] = products['author_name'].str.replace('Khailed Hosseini', 'Khaled Hosseini')
```

```
products['author_name'] = products['author_name'].str.replace('Baird TSpalding', 'Baird T Spalding')
products['author_name'] = products['author_name'].str.replace('Briad LWeiss', 'Briad L Weiss')
products['author_name'] = products['author_name'].str.replace('Robert TKiyosaki', 'Robert T Kiyosaki')
products['author_name'] = products['author_name'].str.replace('Stephen MRCovey', 'Stephen RCovey')
```

Sửa các lỗi chính tả ở cột **author_name** có thể là lỗi do quá trình thu thập.

Tiền xử lý dữ liệu

Duplicated:

- Tương tự cho tập dữ liệu **reviews** cũng có **14774** dòng dữ liệu bị trùng chiếm khoảng **2,4%** dữ liệu nên nhóm xóa các cột này đi.

```
reviews.duplicated().sum()
```

```
14774
```

- và ở tập dữ liệu **reviews** có cột: **seller_product_id** và **product_id** đều là mã sản phẩm nên nhóm xóa cột **seller_product_id**.

Tiền xử lý dữ liệu

Missing Values:

Tập **reviews**,

title đối với các dòng bị thiếu đều có **rating 5** mà

rating 5 đều có title là 'Cực kỳ hài lòng'.

title → **'Cực kỳ hài lòng'.**

Content → **'No comment'.**

```
reviews.isnull().sum()
```

```
id                0
created_at        0
rating            0
title             5
content      392619
thank_count       0
customer_name  6879
customer_id       0
product_id        0
dtype: int64
```

Tiền xử lý dữ liệu

Missing Values:

- Đối với **customer_name** nhóm sẽ xử lý như cột **seller_name** của tập **products**.

```
reviews.isnull().sum()
```

```
id                0
created_at        0
rating            0
title             5
content      392619
thank_count       0
customer_name  6879
customer_id       0
product_id        0
dtype: int64
```

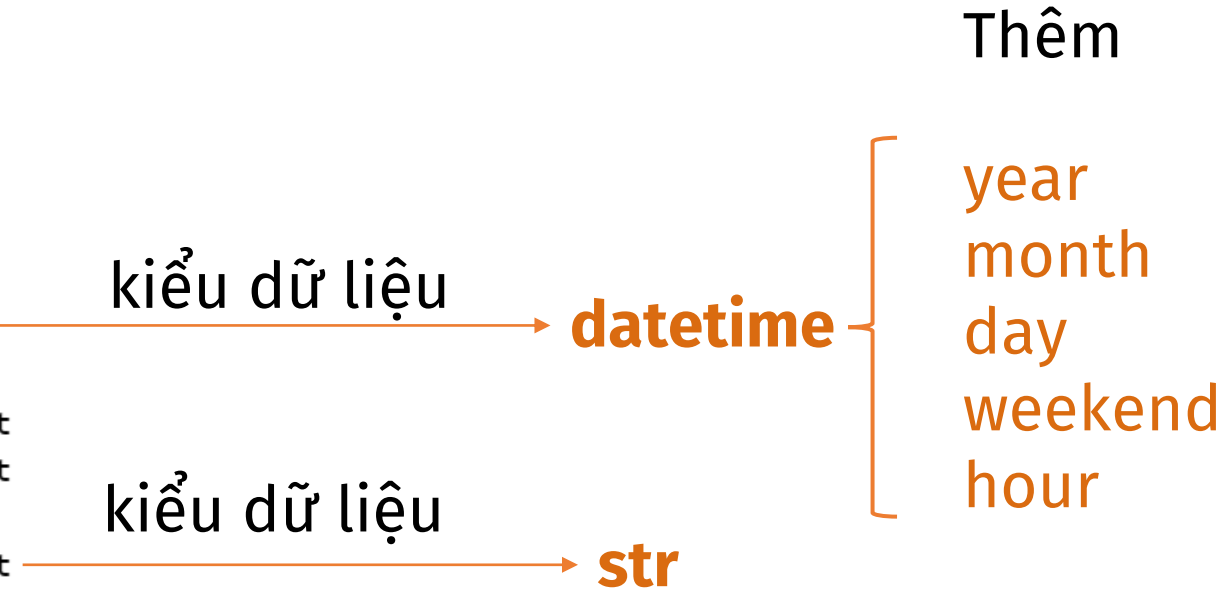
Tiền xử lý dữ liệu

Data Type:

Chuyển đổi dữ liệu các cột, thêm các cột mới.

```
reviews.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 590483 entries, 0 to 605258
Data columns (total 9 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   id             590483 non-null  int64
 1   created_at     590483 non-null  int64
 2   rating         590483 non-null  int64
 3   title          590483 non-null  object
 4   content        590483 non-null  object
 5   thank_count    590483 non-null  int64
 6   customer_name  590483 non-null  object
 7   customer_id    590483 non-null  int64
 8   product_id     590483 non-null  int64
dtypes: int64(6), object(3)
memory usage: 45.1+ MB
```

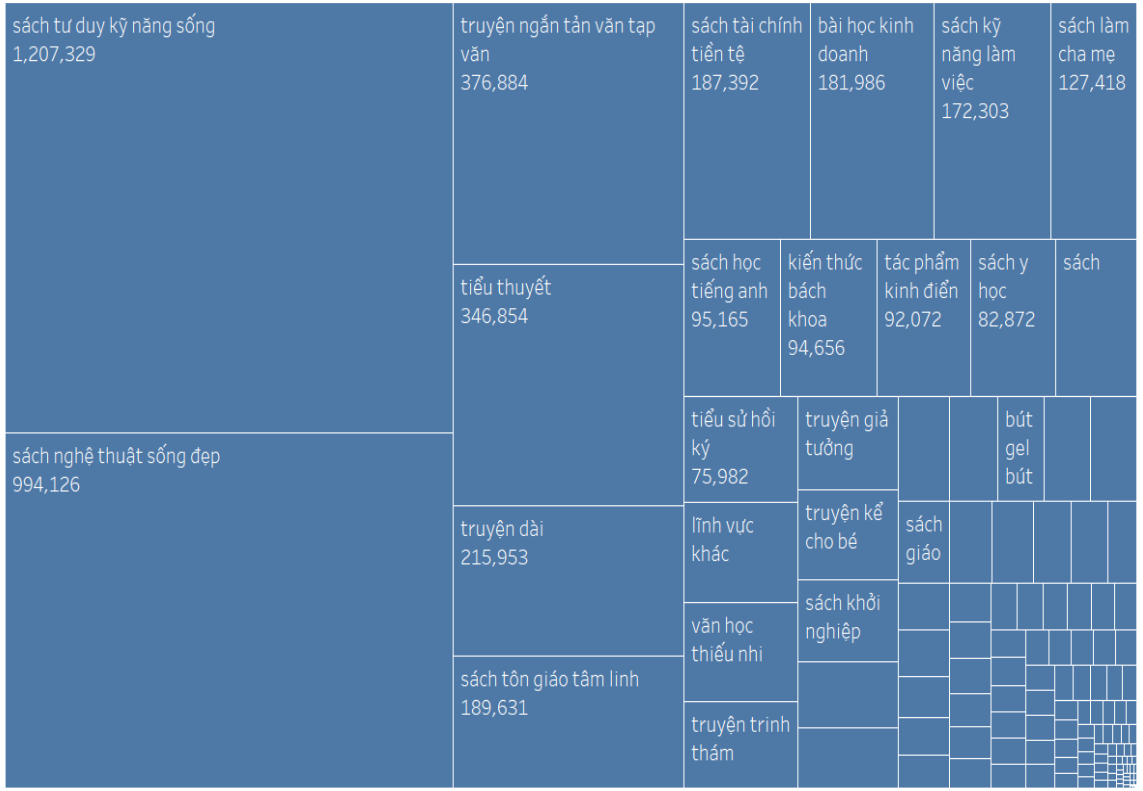


Khám phá & Phân tích dữ liệu

Khám phá dữ liệu

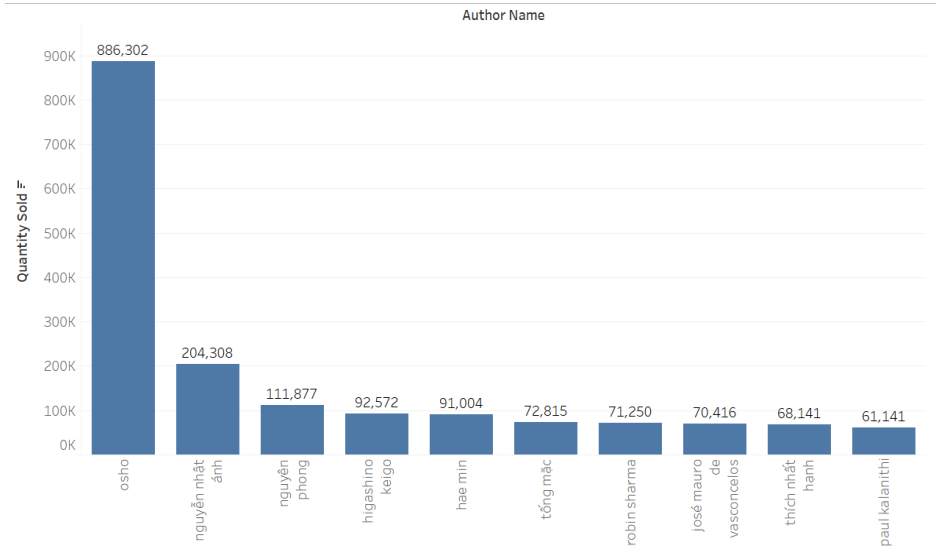
→ Có nhu cầu mua các sách phát triển bản than, nâng cao **kỹ năng sống** và các **đề tài thực tế** như tài chính, kinh doanh.

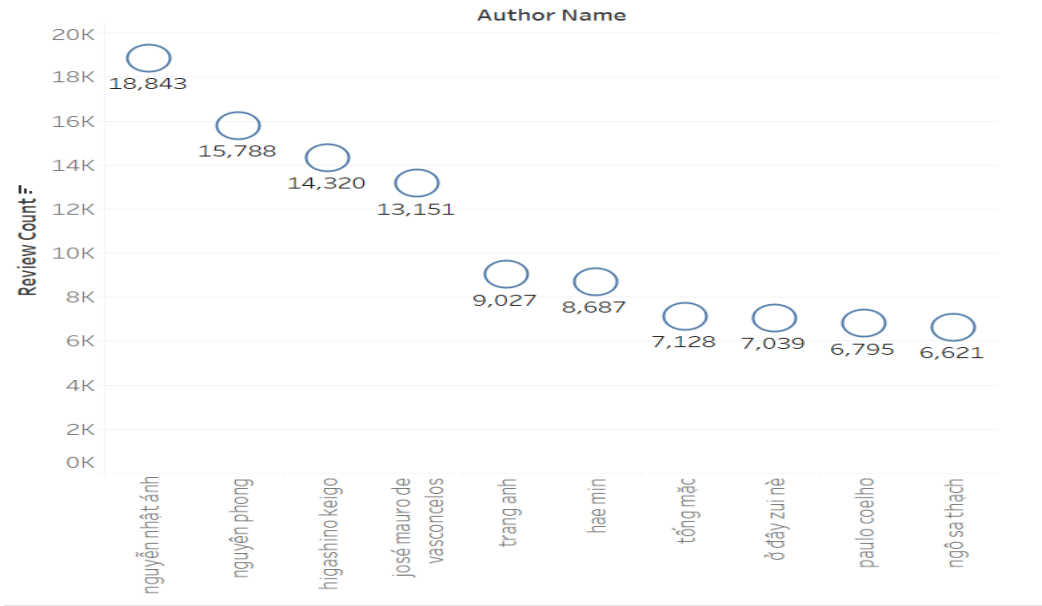
Số lượng bán theo thể loại:



Khám phá dữ liệu

Số lượng bán, số lượng review theo tác giả:





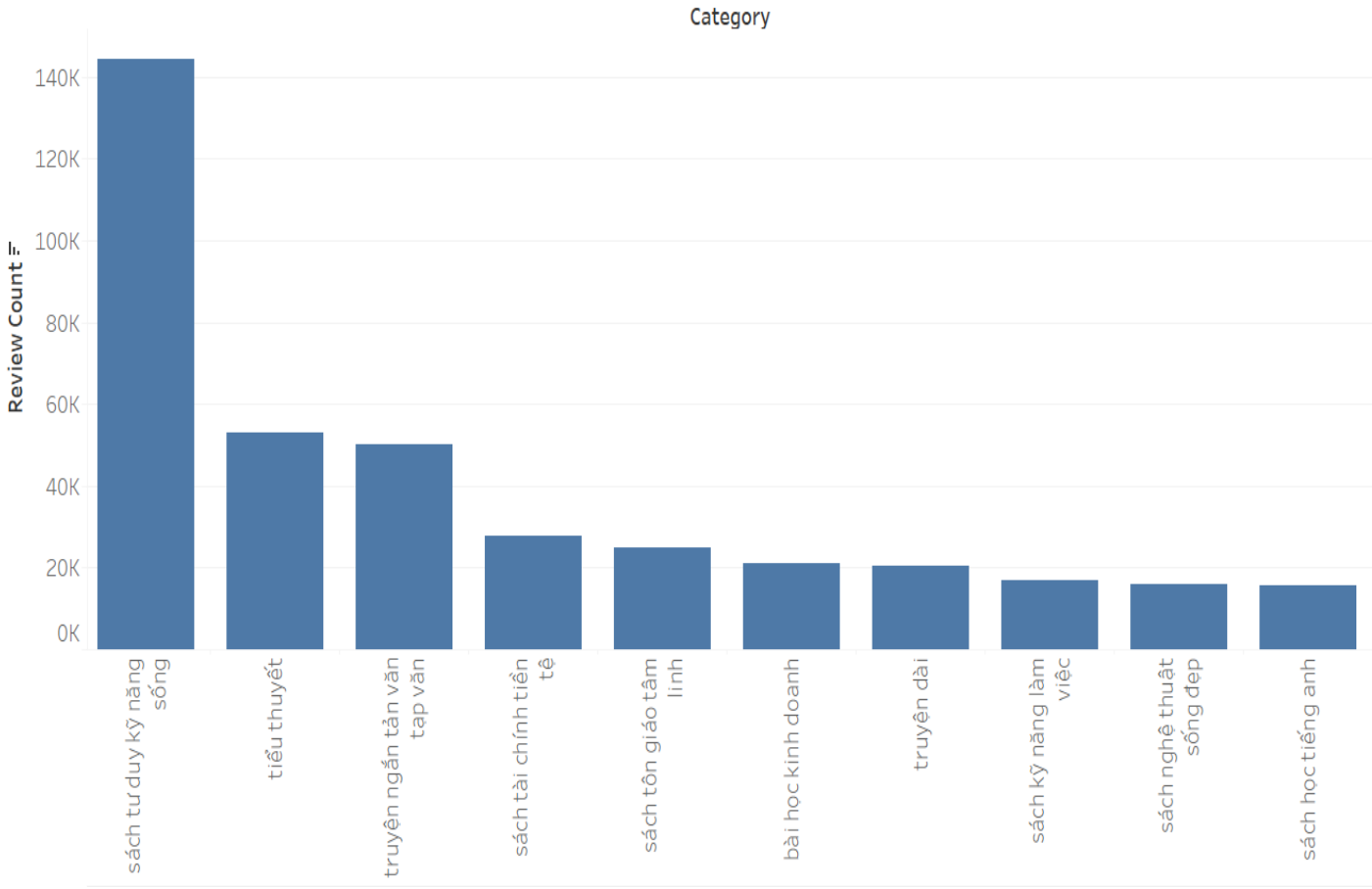
→ Có xu hướng tìm đọc những tác giả nổi tiếng, có uy tín trong các lĩnh vực của mình. (Osho chiếm hơn **80%** nghệ thuật sống đẹp, Nguyễn Nhật Ánh chiếm hơn **50%** truyện dài).

Khám phá dữ liệu

sách học tiếng Anh thuộc top 10 thể loại có review cao nhất.

→ Điều này cho thấy nhu cầu học tiếng Anh của khách hàng khá cao.

Số lượng review theo thể loại:

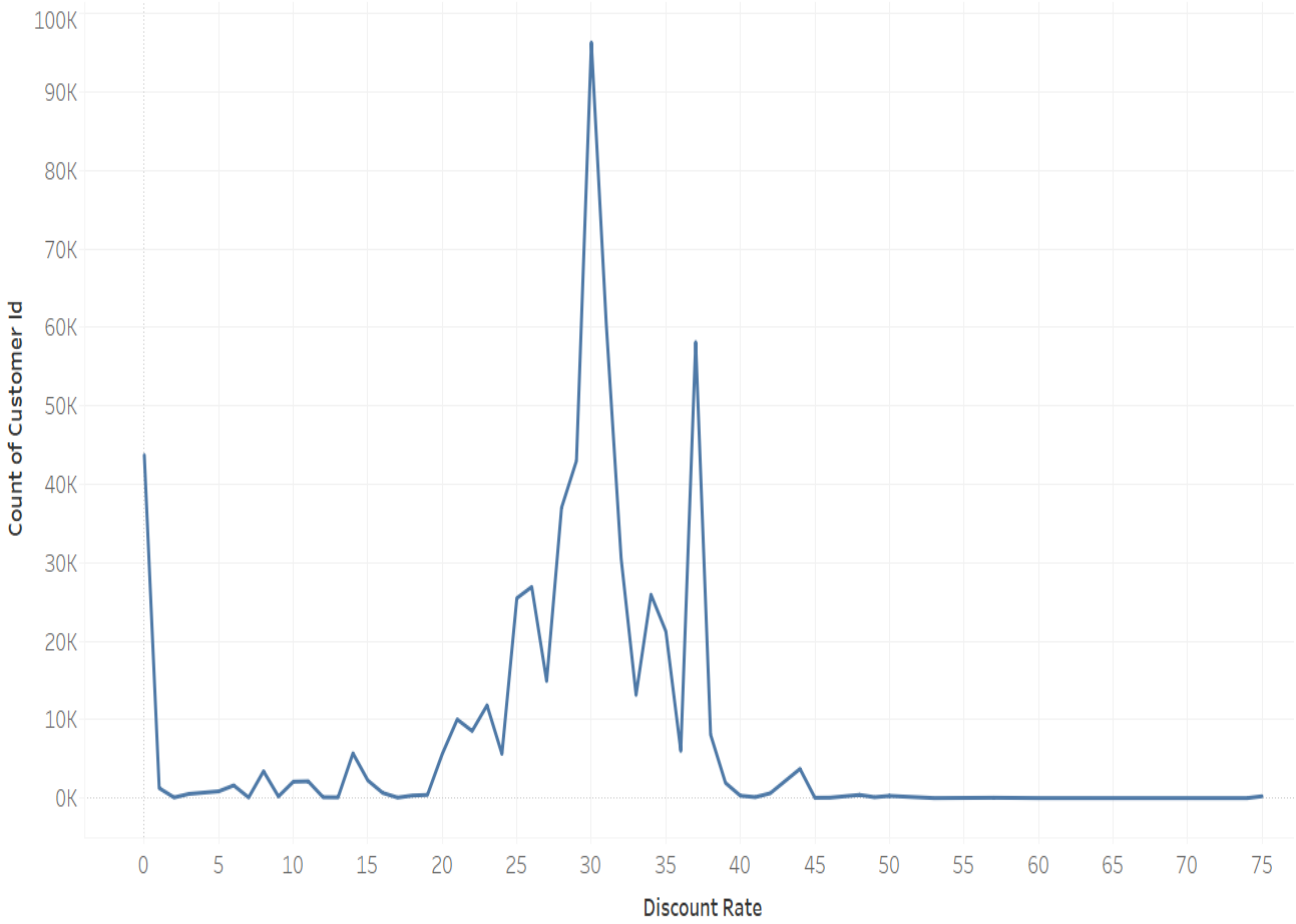


Khám phá dữ liệu

Sản phẩm ở khoảng giảm giá **25%->37%** có lượng khách hàng lớn.

Sản phẩm không được giảm giá cũng có lượt mua khá cao.

Số khách hàng mua theo giảm giá:

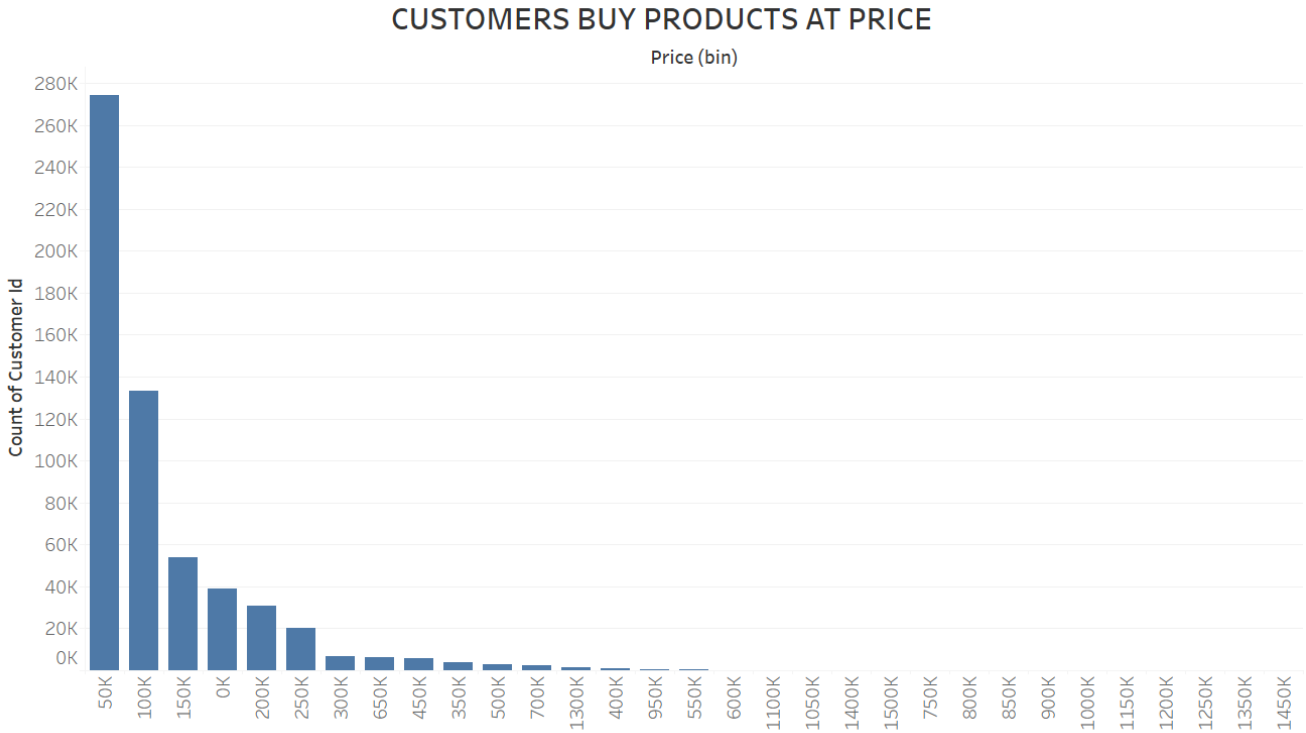


Khám phá dữ liệu

Khách hàng mua theo giá sản phẩm:

Sản phẩm < 200.000 chiếm gần **90%** lượng khách hàng mua sản phẩm.

→ Khách hàng có xu hướng mua hàng với giá trong khoản dưới 200.000.



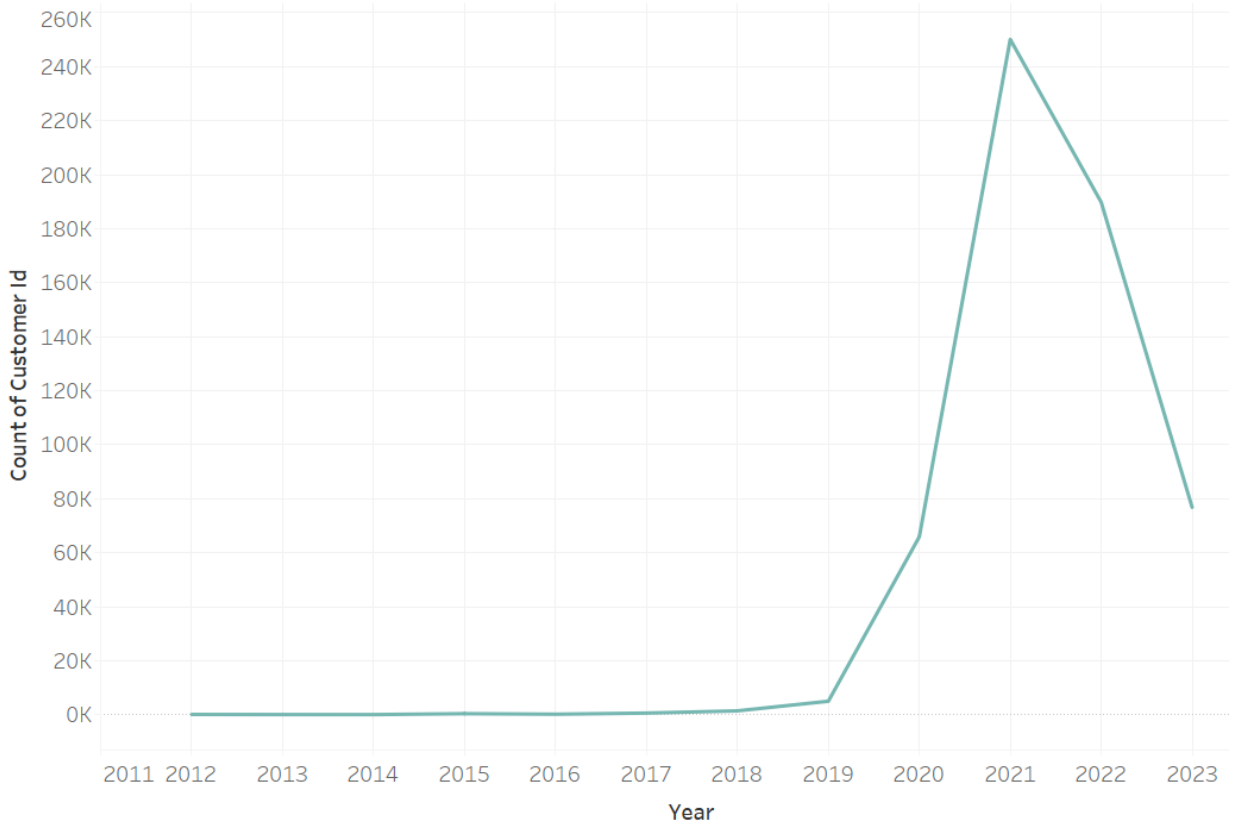
Khám phá dữ liệu

Năm 2020 Thương mại điện tử

Điều kiện mức sống của khách hàng

Năm 2021 → Tăng trưởng mạnh mẽ về khách hàng lẫn số sản phẩm bán được.

Khách hàng theo năm:



Khám phá dữ liệu

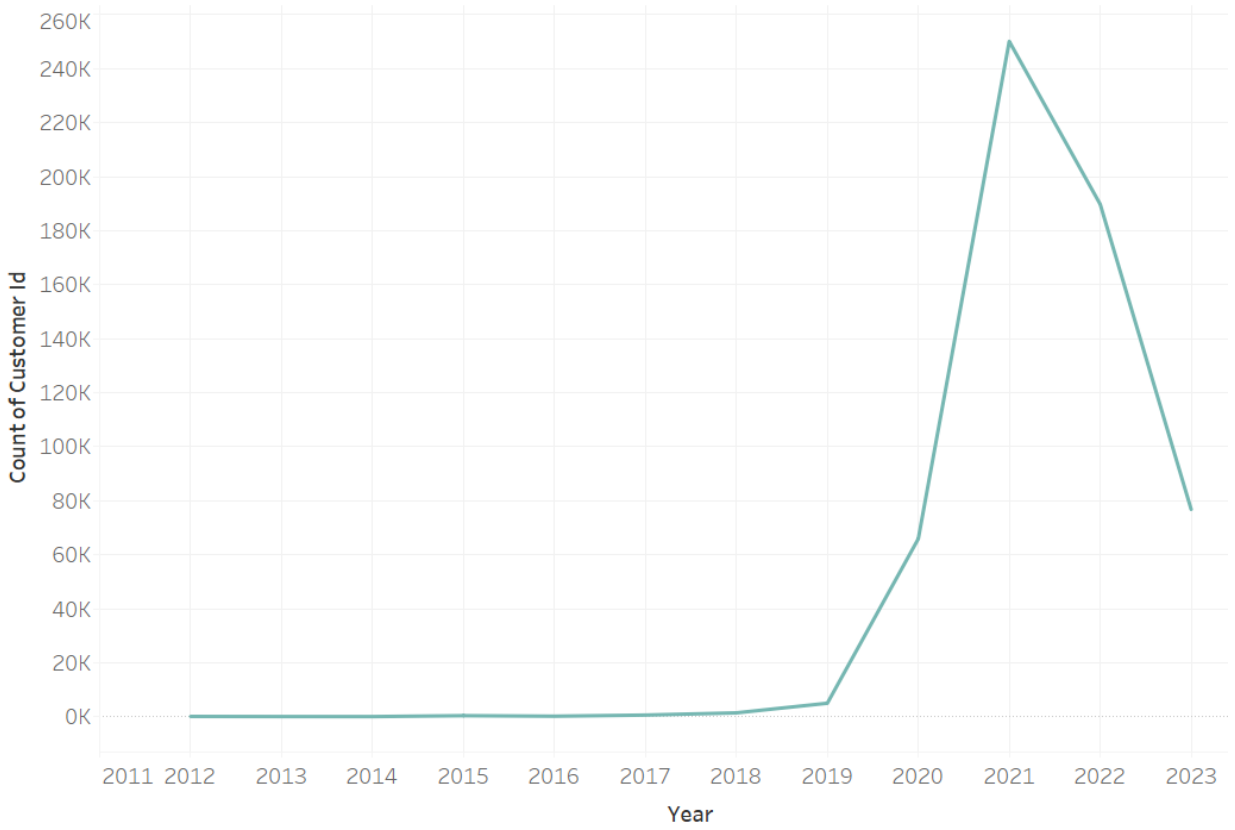
Năm 2022, 2023

Số lượng khách hàng bắt đầu giảm xuống

Thời điểm này đã hết dịch

Một số ứng dụng bán hàng online

Khách hàng theo năm:



Phân tích dữ liệu

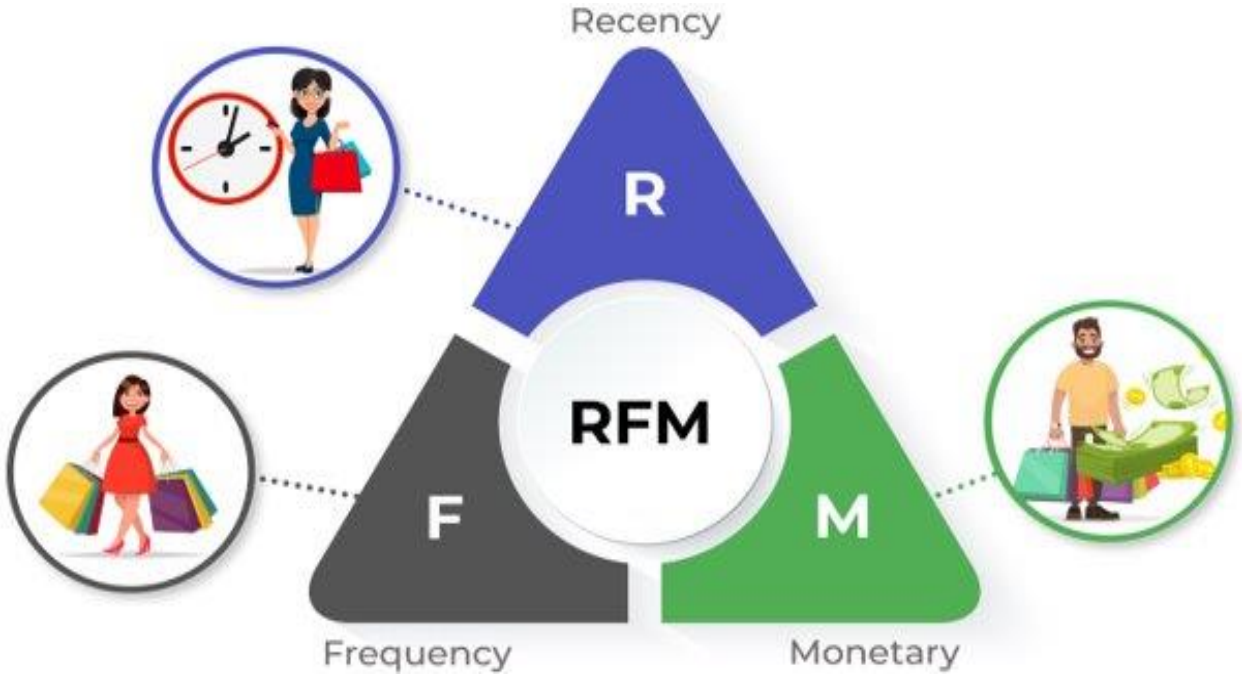
dữ liệu → **insight** → hành vi khách hàng

Sử dụng **RFM** → phân nhóm khách hàng

→ Tùy vào từng nhóm khách hàng đưa ra chiến lược phù hợp.

Phân tích dữ liệu

Mô hình RFM:



- Mô hình phân tích **RFM** (Recency, Frequency, Monetary)

Phân tích dữ liệu

- **Recency**: lần mua hàng gần nhất.
- **Frequency**: Tổng số lần mua hàng.
- **Monetary**: Chi phí mua hàng của khách hàng.

Phân tích dữ liệu

Áp dụng thời gian **1/1/2020 - 14/12/2023.**

Recency: Dùng ngày created_at.

Frequency: Đếm số lượng review.

Monetary: Dựa trên giá của sản phẩm.

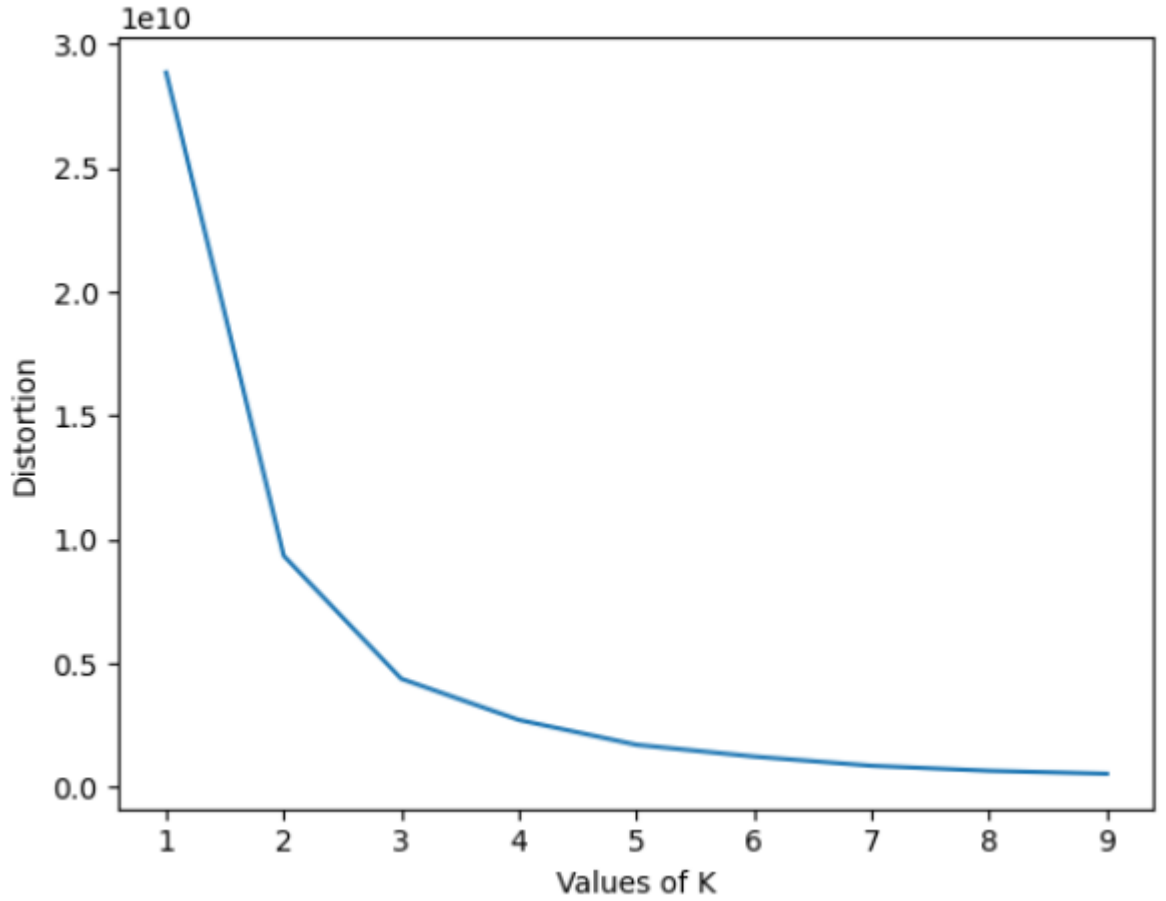
Phân tích dữ liệu

Dùng **Kmean** → phân cụm.

Ta chọn **K = 5**.

Để cho đồng nhất ta lấy **K = 5** cho **Frequency, Monetary.**

Recency



Phân tích dữ liệu

Nối nhãn cụm của **R, F, M** ta được **chuỗi RFM.**

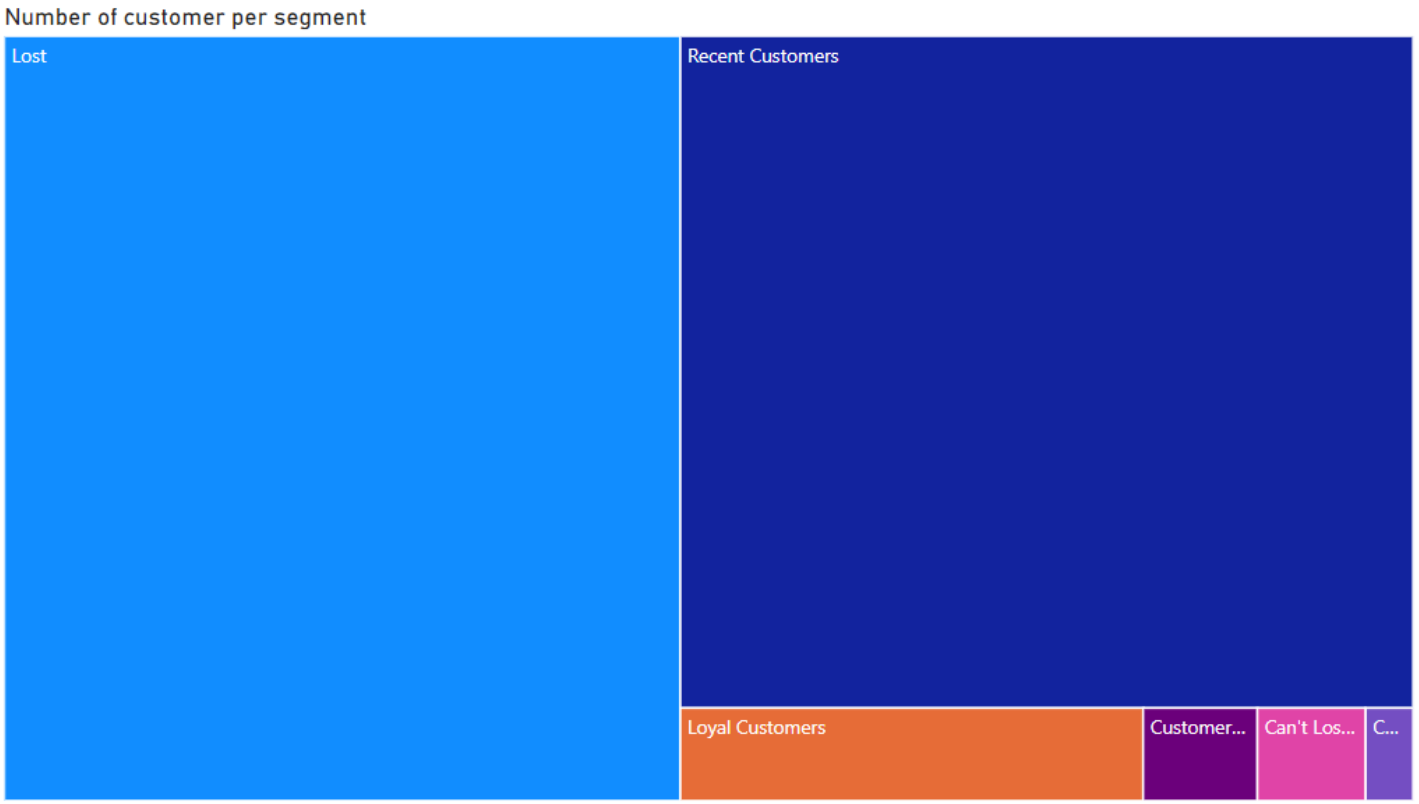
- **555**: có mua hàng gần đây, tầng suất lớn, chi phí lớn.
- **253**: đã khá lâu không mua, mua hàng nhiều lần, giá trị đơn khá lớn.

Phân tích dữ liệu

Đặc điểm	Champions	Loyal Customers	customer Needing Attention	Recent Customers	Can't Lose Them	Lost
Recency	Mua gần đây	Mua gần đây	Không mua gần đây	Mua gần đây	Không mua gần đây	Không mua gần đây
Frequency	Thường xuyên	Thường xuyên	Thường xuyên	Không mua thường xuyên	Thường xuyên	Hiếm khi mua
Monetary	Giá trị lớn	Trung bình	Giá trị lớn	Giá trị thấp	Giá trị lớn	Giá trị thấp

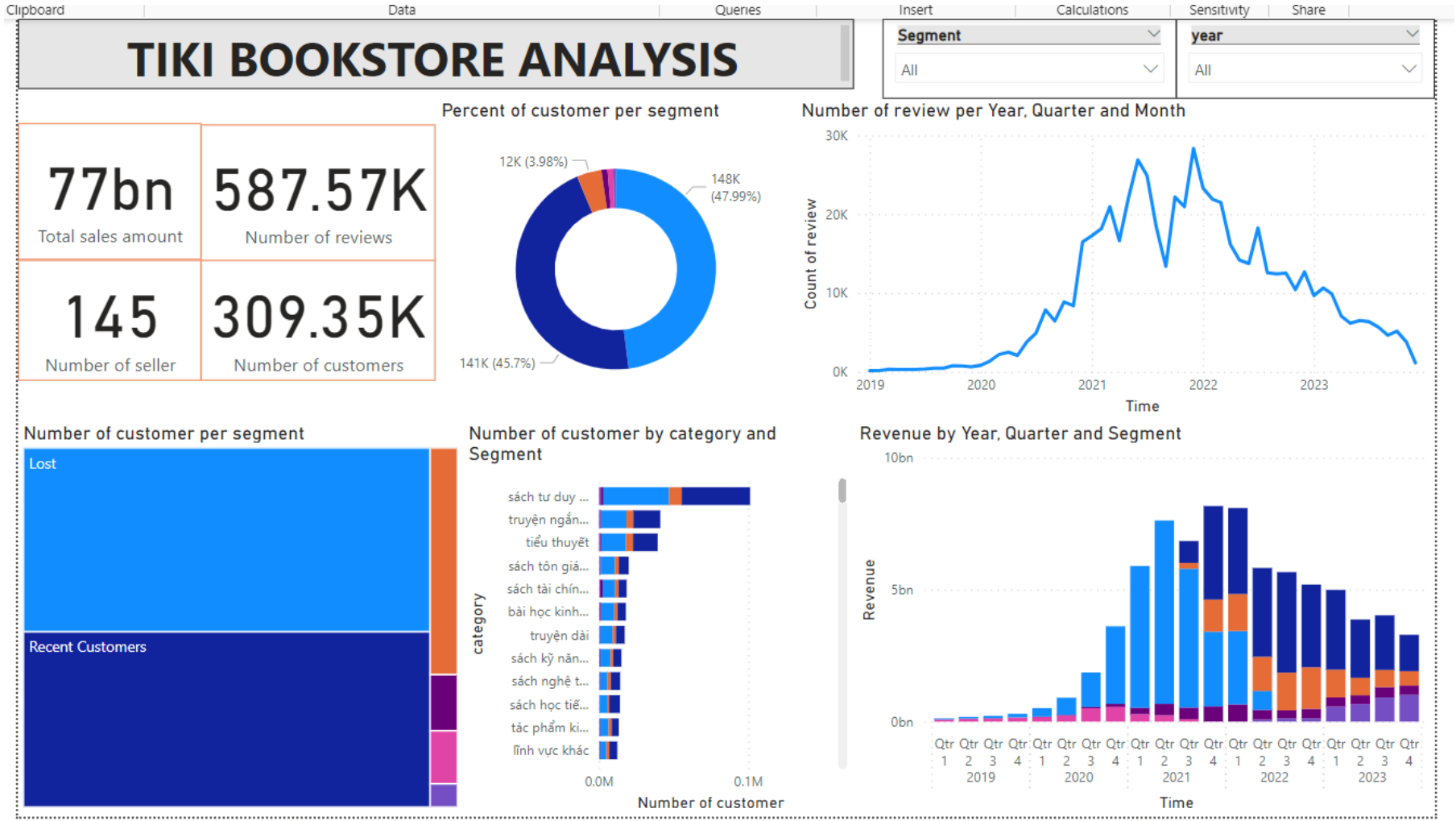
Phân tích dữ liệu

- Dùng **Treemap** trong Power BI để thể hiện kết quả.



Phân tích dữ liệu

Dashboard



Khám phá dữ liệu

Nhóm khách hàng	Cách xử lý	Mục đích
Lost, Can't Lose Them	Gửi mail, gọi điện thoại, ...	Kích thích mua
Recent Customers	Lấy ý kiến, tặng voucher, ...	Tăng sự hài lòng
Loyal Customers	Cho nhận các khuyến mãi đặc biệt, ...	Nâng giá trị giỏ hàng
customer Needing Attention	Tìm nguyên nhân ít mua → đề xuất các chương trình khuyến mãi	Tăng tần suất mua hàng
Champion	VIP, tích điểm, ...	Giữ chân họ tiếp tục mua

Xây dựng mô hình

Gán nhãn

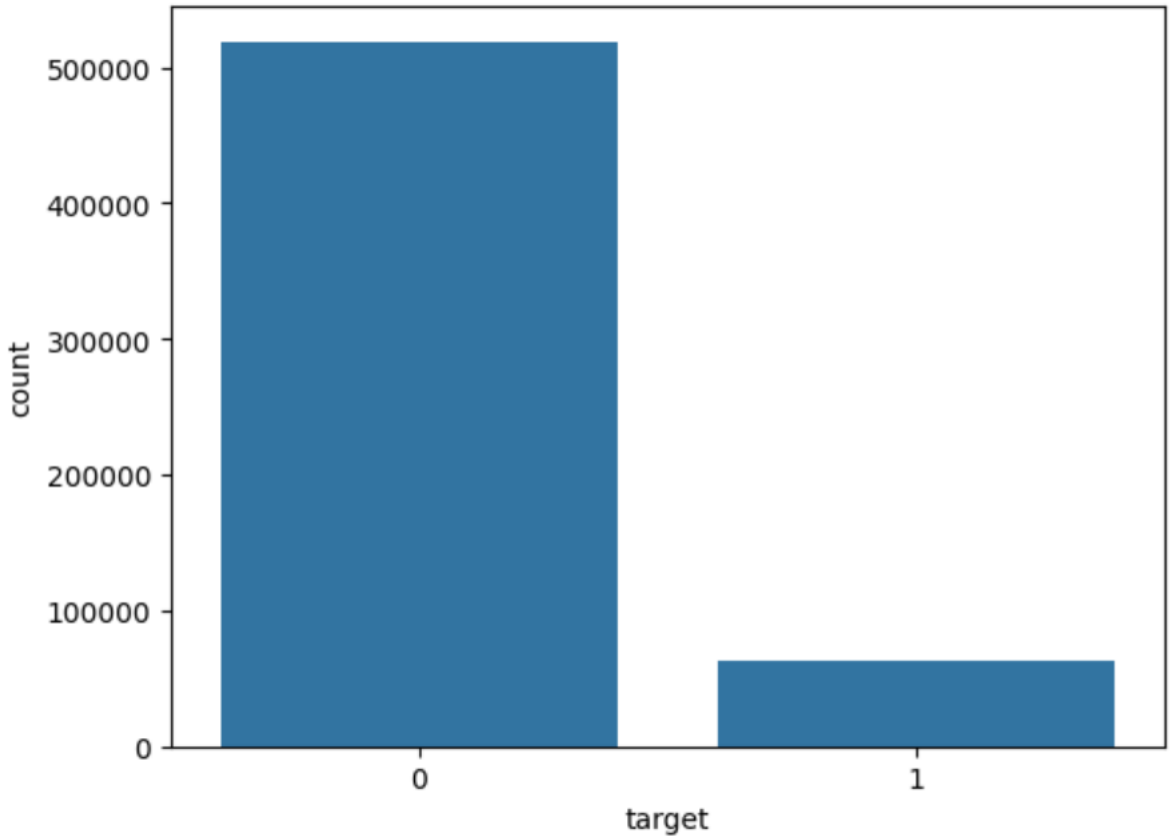
Đánh nhãn khách hàng tiềm năng dựa trên:

- Lần mua gần nhất
- Tần suất mua
- Sở thích(theo tác giả, theo thể loại)
- Theo giá tiền

→ Thỏa 2 trong các điều kiện trên đánh dấu là 1 trong cột target

Phân bố của target

Số khách hàng được đánh nhãn 1 trong bộ dữ liệu chiếm hơn 10%.



Split, GirdSearch

Train test splitting

↓

Split theo tỉ lệ 8/2

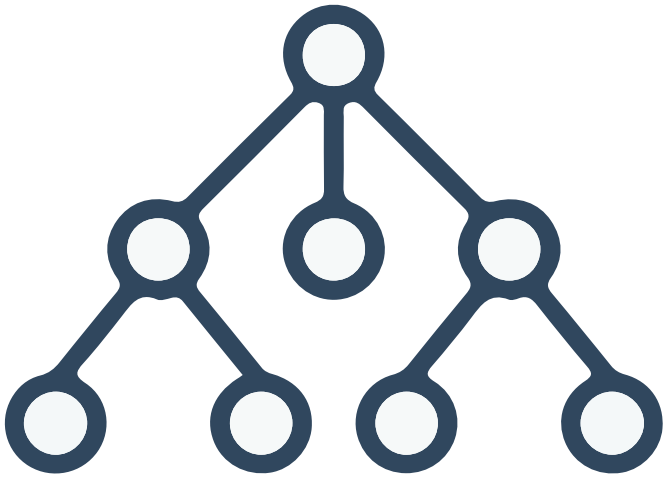
Gridsearch

↓

Cung cấp cho từng model các bộ param

Các loại mô hình

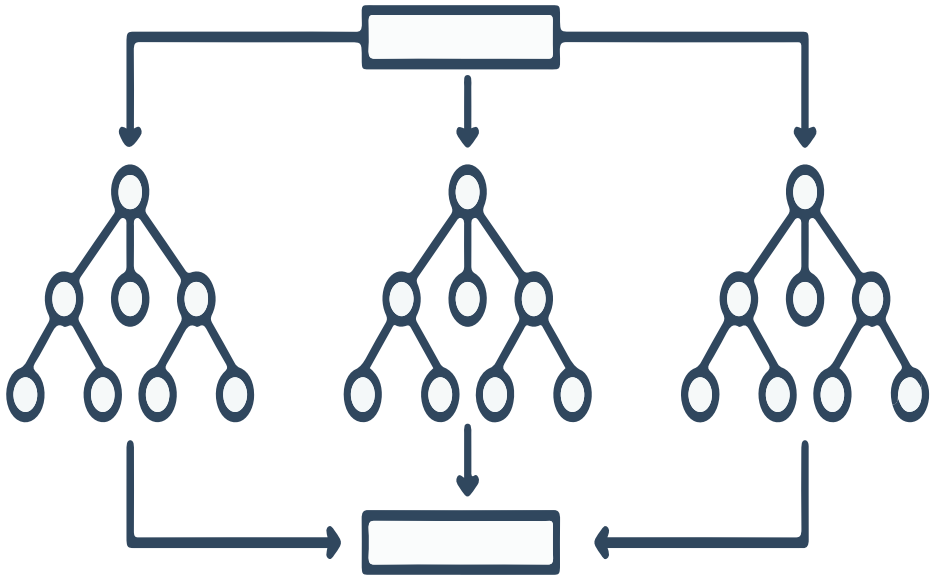
Decision	Random	Gradient	Multinomial
Tree	Forest	Boosting	Naïve
Classifier	Classifier	Classifier	Bayes



Có khoảng 76,84% khách hàng tiềm năng thực sự.

Các loại mô hình

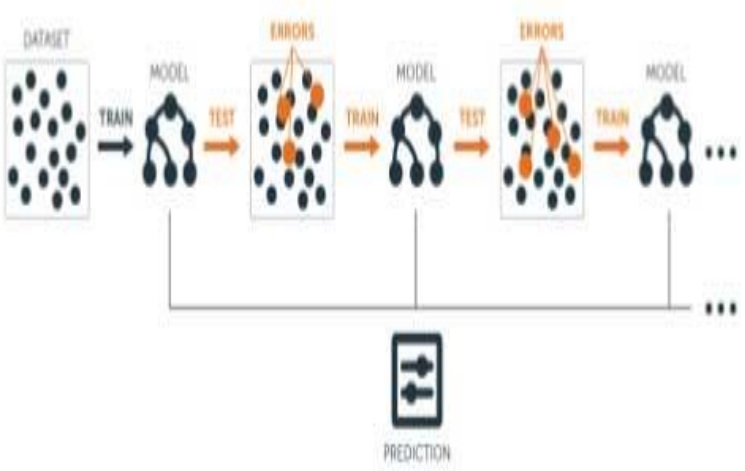
Decision	Random	Gradient	Multinomial
Tree	Forest	Boosting	Naïve
Classifier	Classifier	Classifier	Bayes



Có khoảng 76,54% khách hàng tiềm năng thực sự.

Các loại mô hình

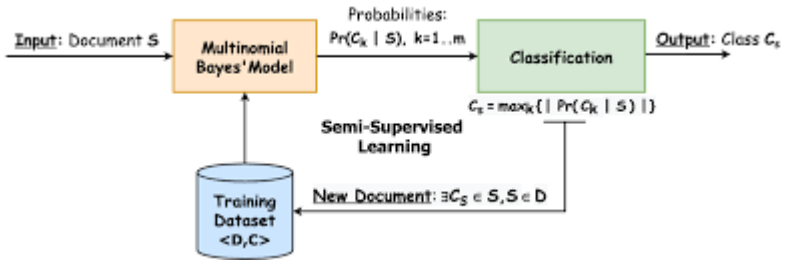
Decision Tree Classifier	Random Forest Classifier	Gradient Boosting Classifier	Multinomial Naïve Bayes



Có khoảng 76,89% khách hàng tiềm năng thực sự.

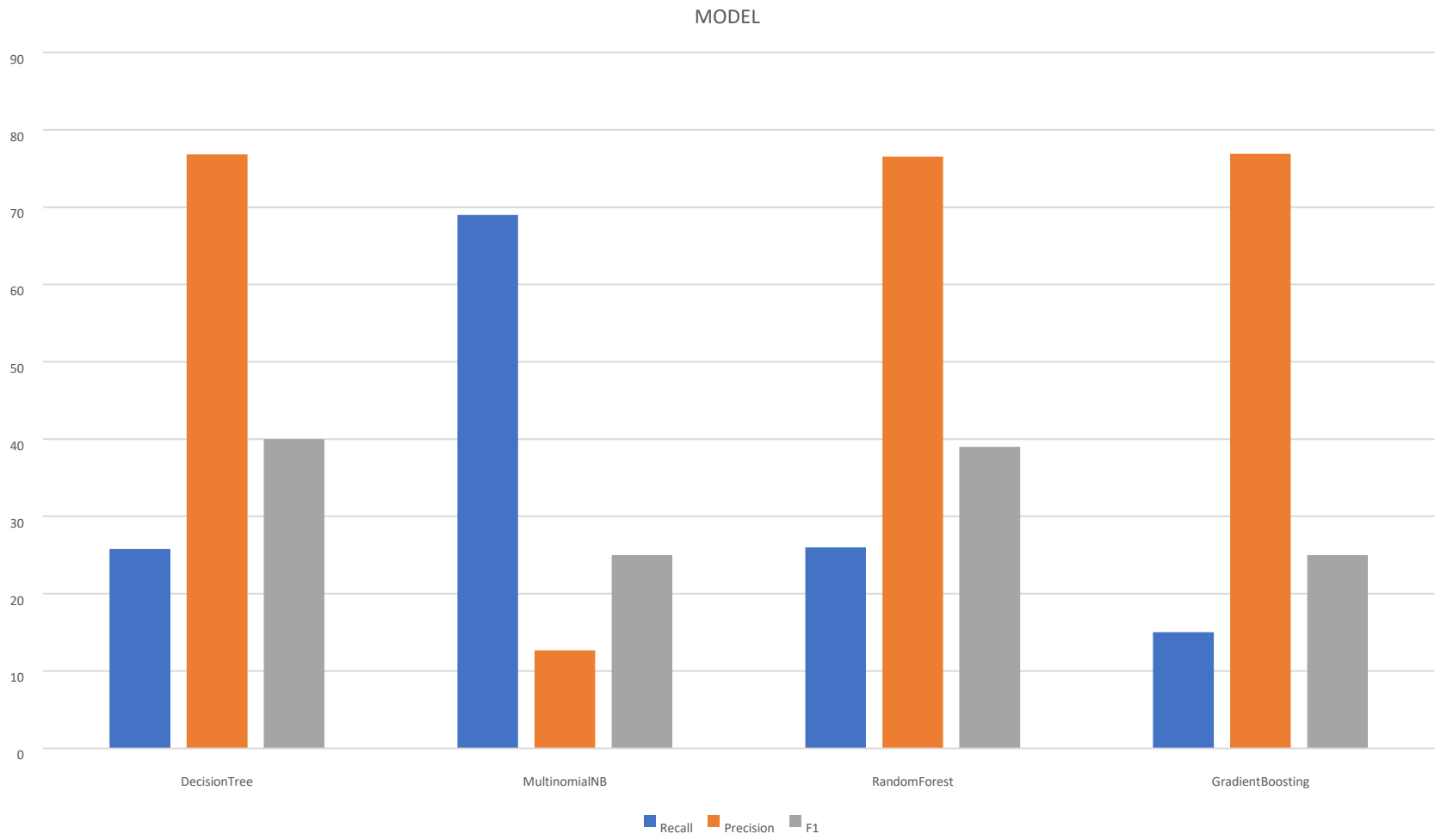
Các loại mô hình

Decision	Random	Gradient	Multinomial
Tree	Forest	Boosting	Naïve
Classifier	Classifier	Classifier	Bayes



Có khoảng 68,9% khách hàng tiềm năng thực sự.

Kết quả


MODEL
90
80
70
60
50
40
30
20
10
0
DecisionTree
MultinomialNB
RandomForest
GradientBoosting
Recall
Precision
F1

The end