DỰ ĐOÁN
Rủi Ro Tín Dụng

PRESENTED BY

Trần Đặng Mạnh An - Group 2

MỤC LỤC

1 Thông tin tập dữ liệu

2 Tìm hiểu và phân tích đánh giá sơ bộ (EDA)

3 Xử lý dữ liệu

4 Lựa chọn mô hình phân tích và kết quả

5 Đánh giá và kết luận

THÔNG TIN TẬP DỮ LIỆU

- **Đây là tập dữ liệu mô phỏng dữ liệu tín dụng.**
- **Mục tiêu: Dựa trên thông tin của dữ liệu dự đoán khách hàng có khả năng vỡ nợ hay không? Giúp ngân hàng có thể đưa ra một số giải pháp để giảm thiểu rủi ro tín dụng.**

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1	0.59	Y	3
1	21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0	0.10	N	2
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1	0.57	N	3
3	23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1	0.53	N	2
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1	0.55	Y	4

Dữ liệu bao gồm :

- 12 cột

- 32581 dòng

THÔNG TIN TẬP DỮ LIỆU

Cột	Miêu tả
person_age	Tuổi
person_income	Thu nhập hàng năm
person_home_ownership	Hình thức nhà sở hữu
person_emp_length	Thời gian làm việc (tính theo năm)
loan_intent	Mục đích vay
loan_grade	Xếp hạng tín dụng

Cột	Miêu tả
loan_amnt	Số tiền vay
loan_int_rate	Lãi suất vay
loan_status	Trạng thái cho vay (0 là không vỡ nợ, 1 là vỡ nợ)
loan_percent_income	Phần trăm khoản vay chiếm trên thu nhập
cb_person_default_on_file	Tiền sử thanh toán (Y: có, N: không)
cb_preson_cred_hist_length	Độ dài lịch sử tín dụng

TÌM HIỂU VÀ PHÂN TÍCH ĐÁNH GIÁ SƠ BỘ (EDA)

Kiểm tra dữ liệu

```
display(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32581 entries, 0 to 32580
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   person_age                  32581 non-null  int64  
 1   person_income               32581 non-null  int64  
 2   person_home_ownership       32581 non-null  object 
 3   person_emp_length           31686 non-null  float64
 4   loan_intent                 32581 non-null  object 
 5   loan_grade                  32581 non-null  object 
 6   loan_amnt                   32581 non-null  int64  
 7   loan_int_rate               29465 non-null  float64
 8   loan_status                 32581 non-null  int64  
 9   loan_percent_income         32581 non-null  float64
 10  cb_person_default_on_file   32581 non-null  object 
 11  cb_person_cred_hist_length  32581 non-null  int64  
dtypes: float64(3), int64(5), object(4)
memory usage: 3.0+ MB
None
```

```
df.isnull().sum()
```

```
person_age                      0
person_income                   0
person_home_ownership           0
person_emp_length             895
loan_intent                     0
loan_grade                      0
loan_amnt                       0
loan_int_rate                3116
loan_status                     0
loan_percent_income             0
cb_person_default_on_file       0
cb_person_cred_hist_length      0
dtype: int64
```

12 cột dữ liệu :

- 8 cột dạng số
- 4 cột dạng chữ

Dữ liệu bị null :

- 895 dữ liệu ở cột thời gian làm việc theo năm
- 3116 dữ liệu ở cột lãi suất vay

TÌM HIỂU VÀ PHÂN TÍCH ĐÁNH GIÁ SƠ BỘ (EDA)

Kiểm tra dữ liệu

```
df[df['person_age'] >= 100]
```

	person_age	person_income	person_home_ownership	person_emp_length
81	144	250000	RENT	4.0
183	144	200000	MORTGAGE	4.0
575	123	80004	RENT	2.0
32297	144	6000000	MORTGAGE	12.0

```
df[df['person_emp_length'] >= 70]
```

	person_age	person_income	person_home_ownership	person_emp_length
0	22	59000	RENT	123.0
210	21	192000	MORTGAGE	123.0

Các dữ liệu bị sai :

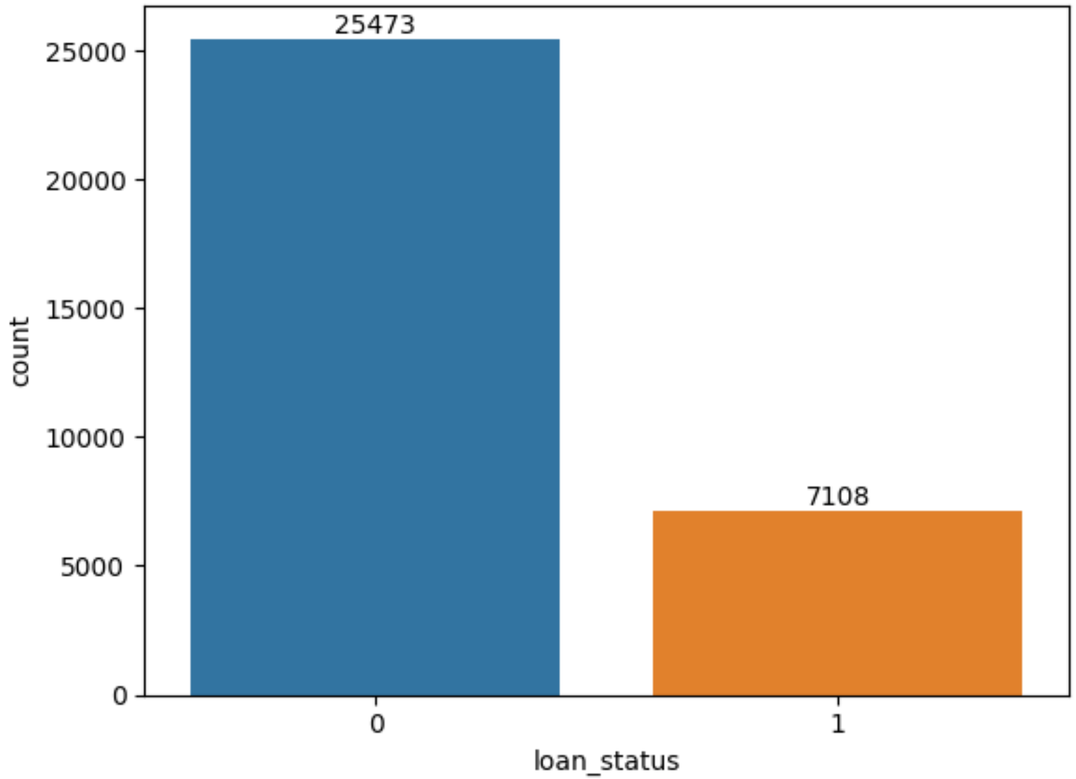
- **Tuổi 123,144**
- **Thời gian làm việc : 123 năm**

TÌM HIỂU VÀ PHÂN TÍCH ĐÁNH GIÁ SƠ BỘ (EDA)

Phân tích đánh giá sơ bộ

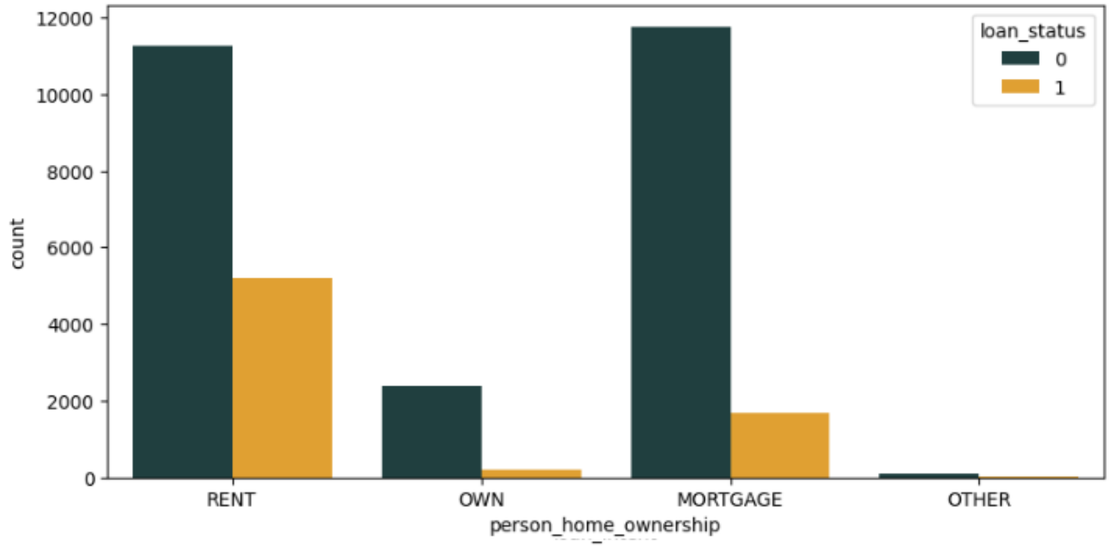
Số lượng người vỡ nợ tương đối thấp hơn số lượng người không vỡ nợ

Khoảng 21.87% trong tập dữ liệu vỡ nợ và 78.13% không vỡ nợ

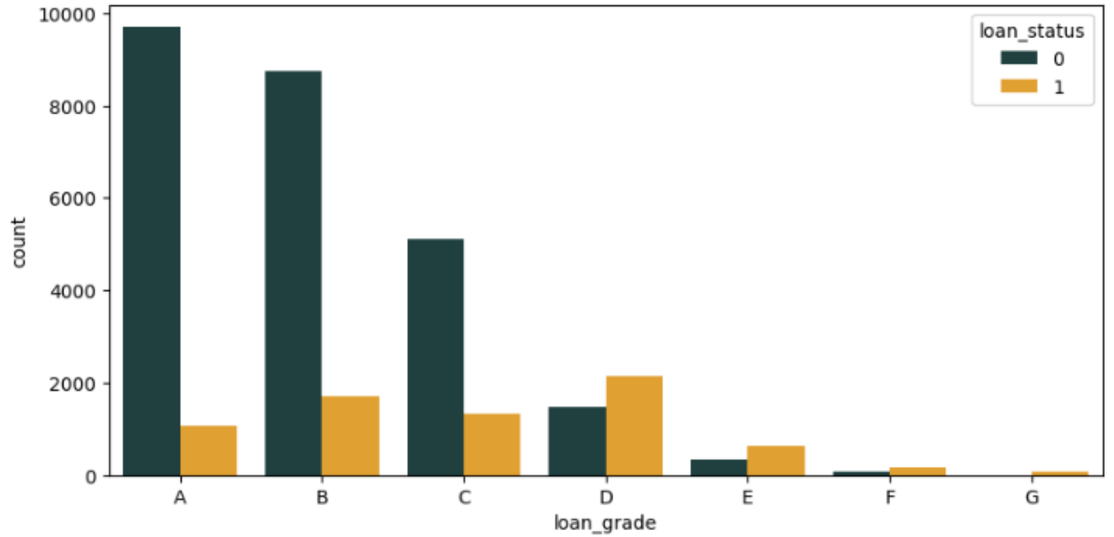


TÌM HIỂU VÀ PHÂN TÍCH ĐÁNH GIÁ SƠ BỘ (EDA)

Phân tích đánh giá sơ bộ



Người ở thuê có xu hướng bị vỡ nợ cao hơn là người có sỡ hữu nhà



Người có xếp hạng tín dụng càng thấp thì có khả năng vỡ nợ càng cao

XỬ LÝ DỮ LIỆU

Trở lại mục lục

```
df.dropna(inplace=True)
```

Xóa null

```
df.drop(df[df['person_age'] >= 100].index, inplace=True)
df.drop(df[df['person_emp_length'] >= 70].index, inplace=True)
```

Xóa những dữ liệu sai

```
#Ordinal Encoding
df["cb_person_default_on_file"].replace(['N' ,'Y'], [0,1], inplace = Tru
# Dùng thư viện one hot encoding
df = pd.get_dummies(df)
```

Encoding

	Trước	Sau
Dòng dữ liệu	32581	28632
Cột	12	26

XỬ LÝ DỮ LIỆU

Trở lại mục lục

Chia tập train - test:

- Tập test : 22905
- Tập train : 5727

Điều chỉnh tập dữ liệu imbalanced:

- Phương pháp oversampling SMOTE trên tập huấn luyện

Cân bằng dữ liệu:

- Phương pháp Min-Max Scaling giúp cải thiện hiệu suất của mô hình và tăng tốc quá trình học.

LỰA CHỌN MÔ HÌNH PHÂN TÍCH VÀ KẾT QUẢ

Sử dụng các mô hình:

- Logistic Regression
- Decision Tree
- KNeighborsClassifier
- Naive Bayes
- Random Forest

Model	Accuracy	Precision		Recall		F1-score		Confusion matrix
		0	1	0	1	0	1	
Logistic Regression	0.87	0.89	0.76	0.95	0.58	0.92	0.65	[[4260 228] [525 714]]
Decision Tree	0.88	0.93	0.71	0.91	0.77	0.92	0.74	[[4091 397] [286 953]]
KNeighborsClassifier	0.89	0.90	0.83	0.97	0.63	0.93	0.72	[[4333 155] [456 783]]
Naive Bayes	0.84	0.87	0.68	0.94	0.50	0.90	0.58	[[4201 287] [618 621]]
Random Forest	0.93	0.93	0.95	0.99	0.72	0.96	0.82	[[4439 49] [351 888]]

Trở lại mục lục

LỰA CHỌN MÔ HÌNH PHÂN TÍCH VÀ KẾT QUẢ

Sử dụng các mô hình:

- **Adaboost**
- **LightGBM**
- **XGBoost**
- **SVC**

Model	Accuracy	Precision		Recall		F1-score		Confusion matrix
		0	1	0	1	0	1	
Adaboost	0.88	0.92	0.73	0.93	0.70	0.92	0.72	[[4173 315] [368 871]]
LightGBM	0.94	0.93	0.96	0.99	0.73	0.96	0.83	[[4455 33] [337 902]]
XGBoost	0.93	0.93	0.94	0.99	0.74	0.96	0.83	[[4430 58] [323 916]]
SVC	0.90	0.91	0.89	0.98	0.63	0.94	0.74	[[4389 99] [453 786]]

Trở lại mục lục

ĐÁNH GIÁ VÀ KẾT LUẬN

KẾT LUẬN

Model	Accuracy	Precision		Recall		F1-score		Confusion matrix
		0	1	0	1	0	1	
Random Forest	0.93	0.93	0.95	0.99	0.72	0.96	0.82	[[4439 49] [351 888]]
LightGBM	0.94	0.93	0.96	0.99	0.73	0.96	0.83	[[4455 33] [337 902]]
XGBoost	0.93	0.93	0.94	0.99	0.74	0.96	0.83	[[4430 58] [323 916]]

-Các model cho kết quả tốt nhất.

- Random Forest có thể ổn định và dễ điều chỉnh
- LightGBM và XGBoost có thể đòi hỏi thời gian và tài nguyên tính toán cao hơn

ĐÁNH GIÁ VÀ KẾT LUẬN

KẾT LUẬN

Để giảm thiểu rủi ro tín dụng:

- **Tối ưu hóa quy trình phê duyệt vay:** Ngân hàng tăng cường quy trình phê duyệt vay cho những người thuê nhà hoặc có điểm tín dụng thấp để đảm bảo rằng họ có khả năng trả nợ.
- **Phát triển sản phẩm tài chính phù hợp:** Ngân hàng phát triển các sản phẩm tài chính phù hợp cho những nhóm có rủi ro cao để giúp họ quản lý tài chính hiệu quả hơn.

Thanks
for watching!