

Đề tài: Dự đoán khả năng trả
nợ của khoản vay




Cấu trúc bài thuyết trình

- 
- I.** Giới thiệu đề tài
 - II.** Lựa chọn hướng pháp
 - III.** Thu thập dữ liệu
 - IV.** Thấu hiểu dữ liệu
 - V.** Chuẩn bị dữ liệu
 - VI.** Xây dựng mô hình
 - VII.** Đánh giá mô hình

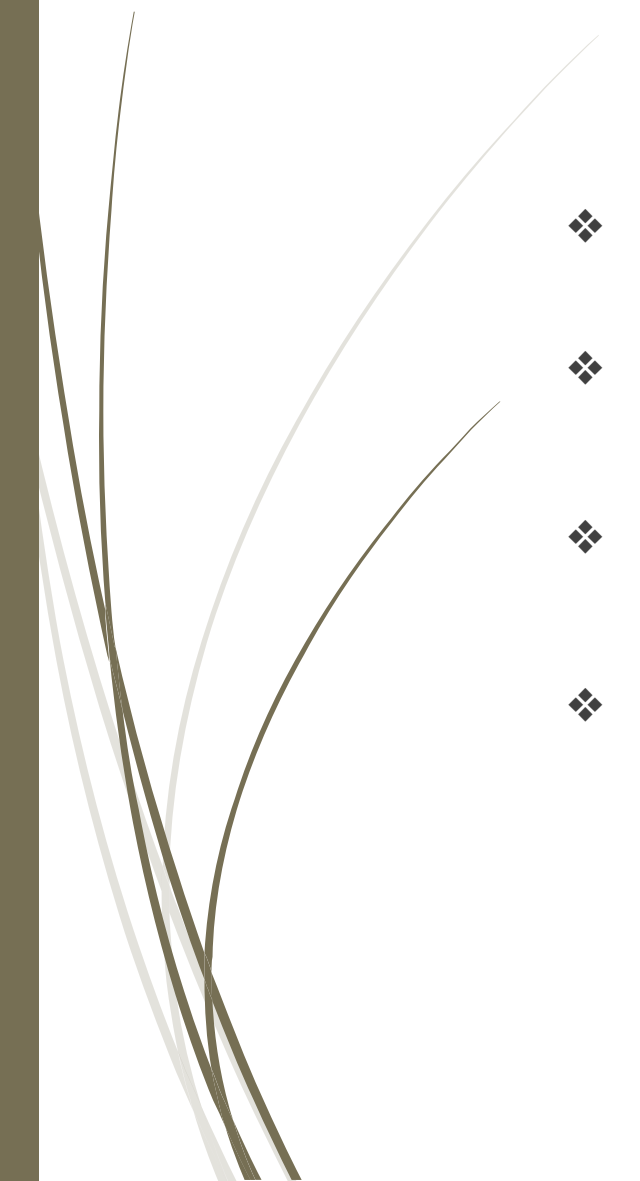


Làm rõ yêu cầu

- ❖ Một khoản vay có nên được phê duyệt hay không? Các ngân hàng có thể đánh giá rủi ro của một đơn xin vay mới được không?
 - ❖ Mục tiêu của việc xây dựng mô hình là dựa vào dữ liệu quá khứ để dự đoán một khoản vay có trả được nợ không trong tương lai. Từ việc dự đoán này ngân hàng có thể ra quyết định từ chối cho vay.
- 



Lựa chọn phương pháp

- ❖ Mô hình Baseline: Logistic Regression
 - ❖ Mô hình nâng cao: Gradient Boosting
 - ❖ Phép đo hiệu quả của mô hình: F1 score
 - ❖ Mục tiêu: F1 score $\geq 75\%$
- 

Thu thập dữ liệu

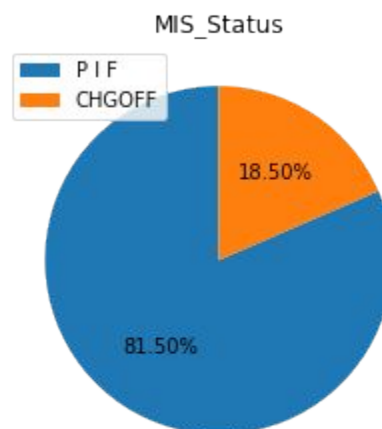
- ❖ Nguồn: <https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied>
- ❖ Điều chỉnh sai sót:
 - Có ChgOffDate hoặc có ChgOffPrinGr nhưng gắn nhãn PIF
 - Loại bỏ trường hợp thiếu nhãn
- ❖ Lọc dữ liệu: Đa số các khoản vay có thời hạn 5 năm. Dữ liệu có các được thu thập đến 2014 do đó từ 2009 - 2014 thiếu dữ liệu các khoản vay chưa kết thúc. Do đó ta loại bỏ các khoản vay có ngày giải ngân trong khoảng thời gian này.
- ❖ Phân tách tập huấn luyện, tập kiểm tra:

	Số quan sát	%
Tập huấn luyện	637,255	75
Tập kiểm tra	212,419	25

Thấu hiểu dữ liệu

Biến mục tiêu (MIS_Status)

Nhãn	Encode	Positive / Negative	Ý nghĩa
P I F	0	Negative	Trả được nợ (Paid In Full)
CHGOFF	1	Positive	Không trả được (Charge Off)





Thấu hiểu dữ liệu

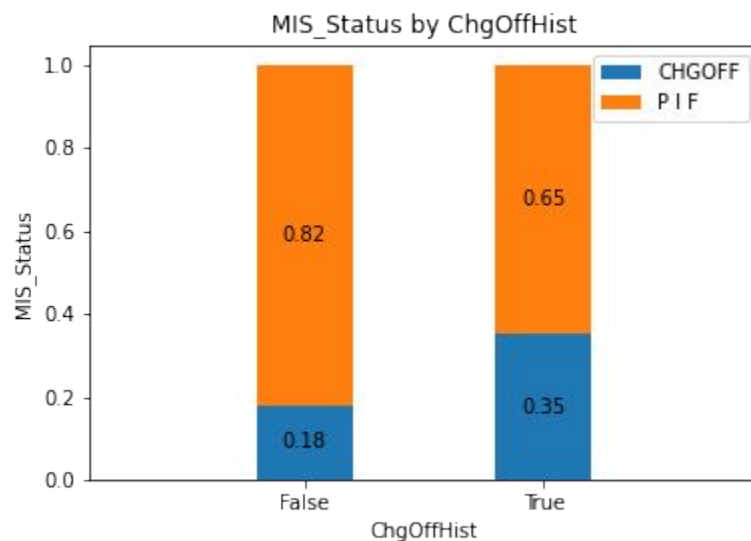
26 thuộc tính

- ❖ Record ID (**1**): LoanNr_ChkDgt
- ❖ Định lượng (**9**): Term, NoEmp, CreateJob, RetainJob, DisbursementGross, BalanceGross, GhgOffPrinGr, GrAppr, SBA_Apprv
- ❖ Phân loại (**11**): City, State, Zip, Bank, BankState, NAICS, FranchiseCode, UrbanRural, RevLineCr, LowDoc, Name
- ❖ Thời gian (**5**): ApprovalDate, ApprovalFY, ChgOffDate, DisbursementDate

Thấu hiểu dữ liệu

ChgOffHist

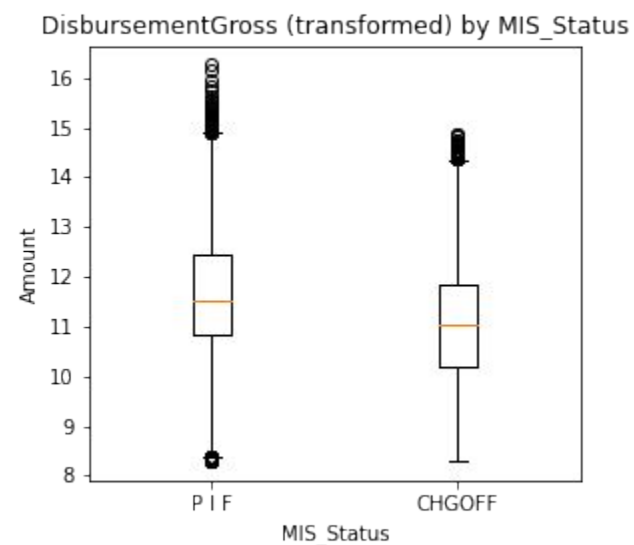
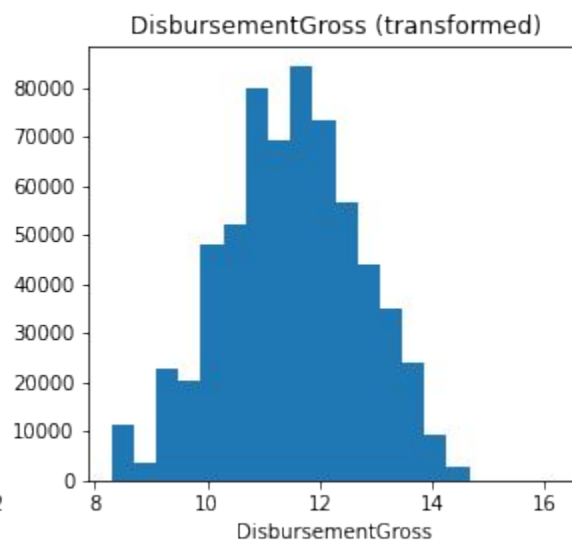
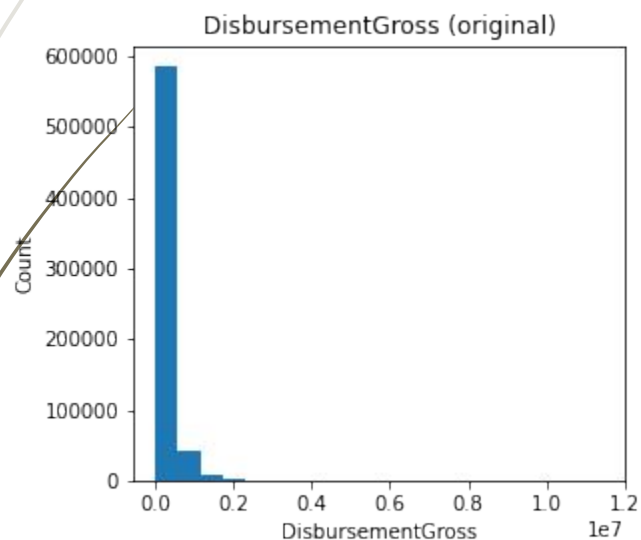
- ❖ Một khách hàng có thể có nhiều khoản vay.
- ❖ Nếu khoản vay trước đó không trả được nợ thì các khoản vay sau đó cũng có xu hướng không trả được nợ



Thấu hiểu dữ liệu

DisbursementGross

- Thực hiện biến đổi log, qua biểu đồ trực quan hoá có thể thấy các khoản vay PIF có số tiền lớn hơn khoản vay CHGOFF.

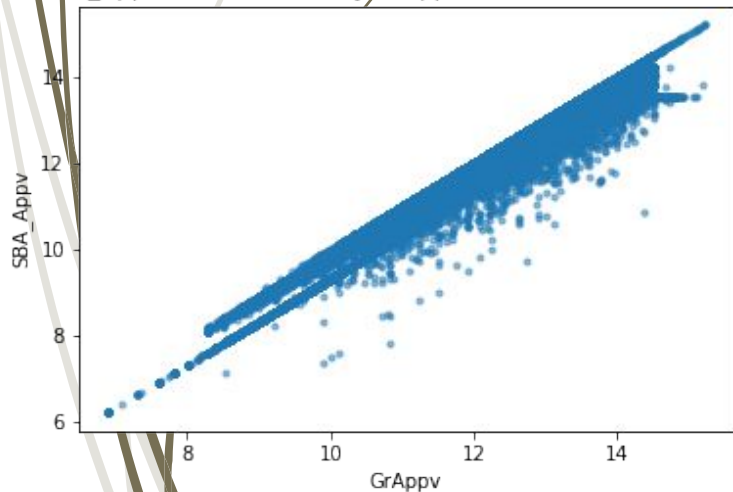


Thấu hiểu dữ liệu

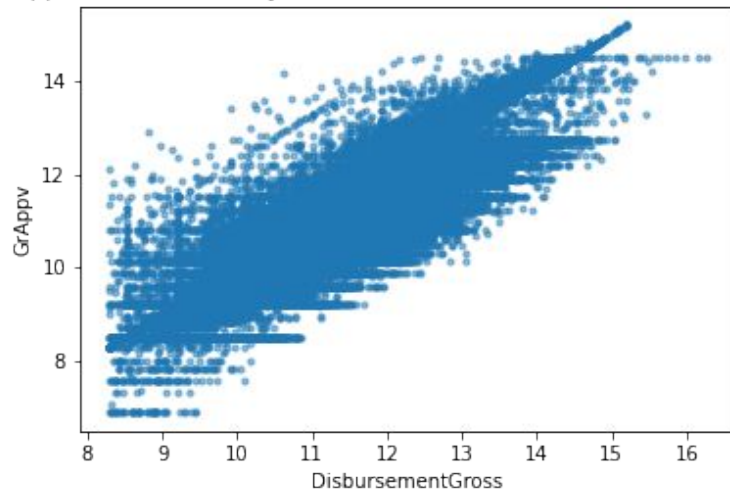
GrAppv, SBA_Appv

- DisbursementGross, GrAppv, SBA_Appv có quan hệ tuyến tính rõ với nhau. Do đó ta chỉ chọn 1 trong 3 biến để đưa vào huấn luyện.

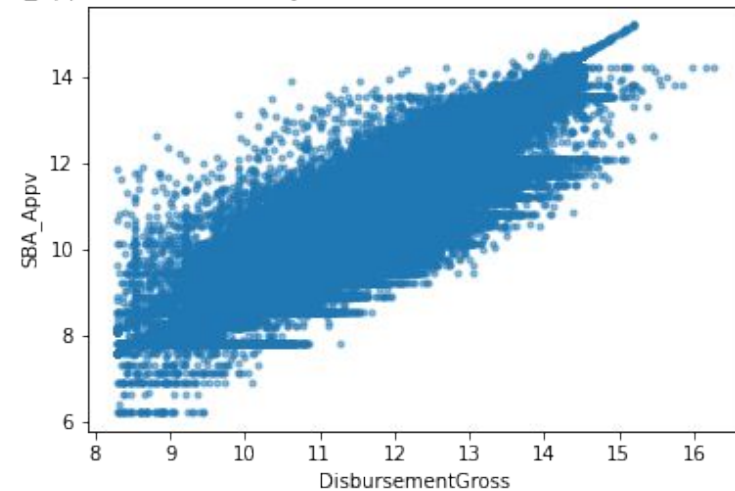
SBA_Appv (transformed) by GrAppv (transformed) - R2: 0.987



GrAppv (transformed) by DisbursementGross (transformed) - R2: 0.965



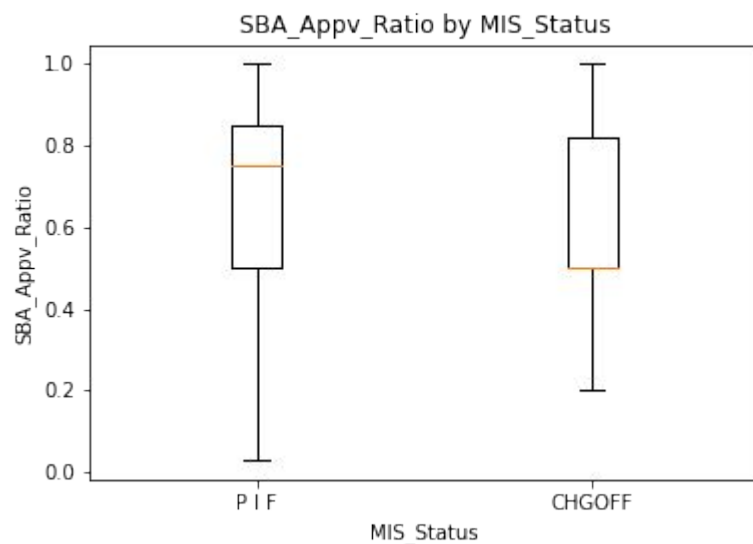
SBA_Appv (transformed) by DisbursementGross (transformed) - R2: 0.94



Thấu hiểu dữ liệu

SBA_Appv_Ratio

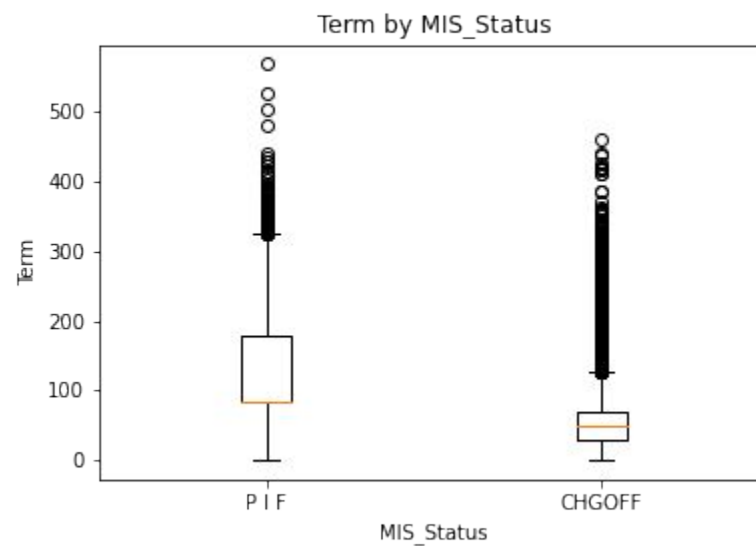
- Khi loại bỏ GrAppv và SBA_Appv và thay thế bằng tỷ lệ $SBA_Appv / GrAppv$ (tỷ lệ bảo lãnh) thì biến mới này có khả năng giải thích kết quả CHGOFF / PIF. (Khoản vay được đánh giá tốt thì ngân hàng chấp nhận tỷ lệ bảo lãnh cao).



Thấu hiểu dữ liệu

Term

- Có khoảng 600 trường hợp có kỳ hạn bằng 0 có thể gán giá trị mode.
- Các khoản vay CHGOFF thường có kỳ hạn thấp hơn khoản vay PIF.



Thấu hiểu dữ liệu

Recession


- Giai đoạn suy thoái kinh tế từ tháng 12/2007 đến tháng 6/2009
https://en.wikipedia.org/wiki/Great_Recession
- Qua phân tích các khoản vay trong giai đoạn này (ngày hết hạn từ 12/2007 - 6/2009) có tỷ lệ CHGOFF cao hơn bình thường
- Ngày hết hạn = ngày giải ngân + kỳ hạn





Thấu hiểu dữ liệu

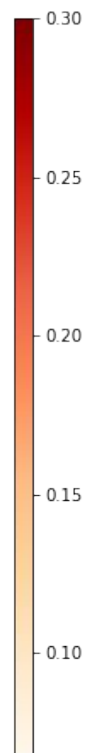
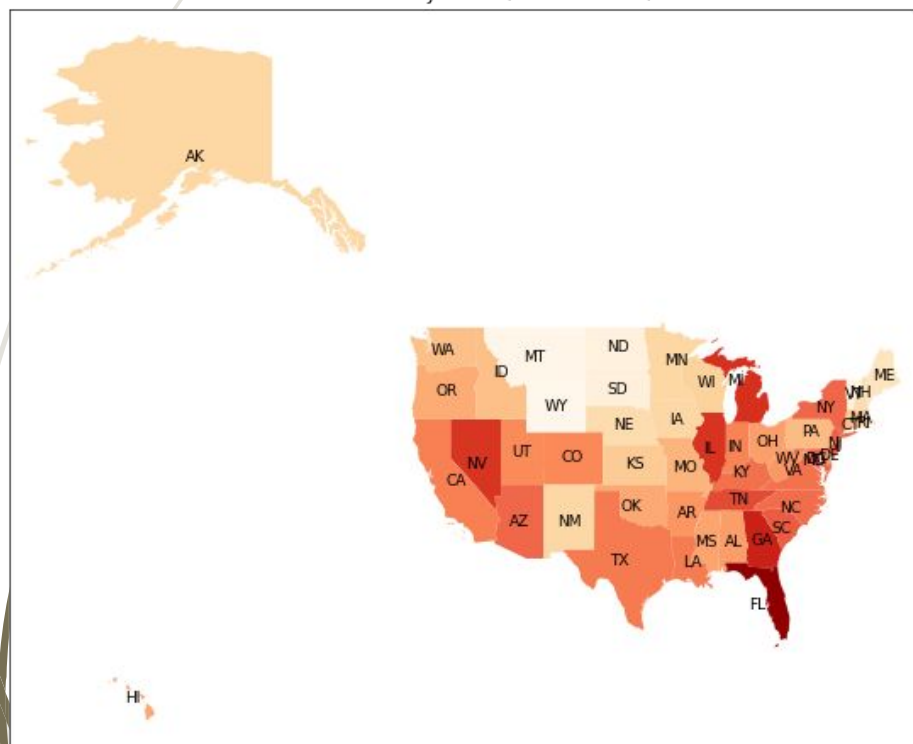
State, BankState

- Các bang khác nhau có tỷ lệ CHGOFF khác nhau
 - Khi có kèm yếu tố Recession thì tỷ lệ CHGOFF cũng có sự thay đổi khác nhau giữa các bang
 - Các trường hợp thiếu dữ liệu có thể Impute căn cứ vào ZIP (State), giá trị mode (BankState)
- 

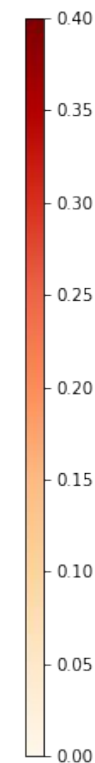
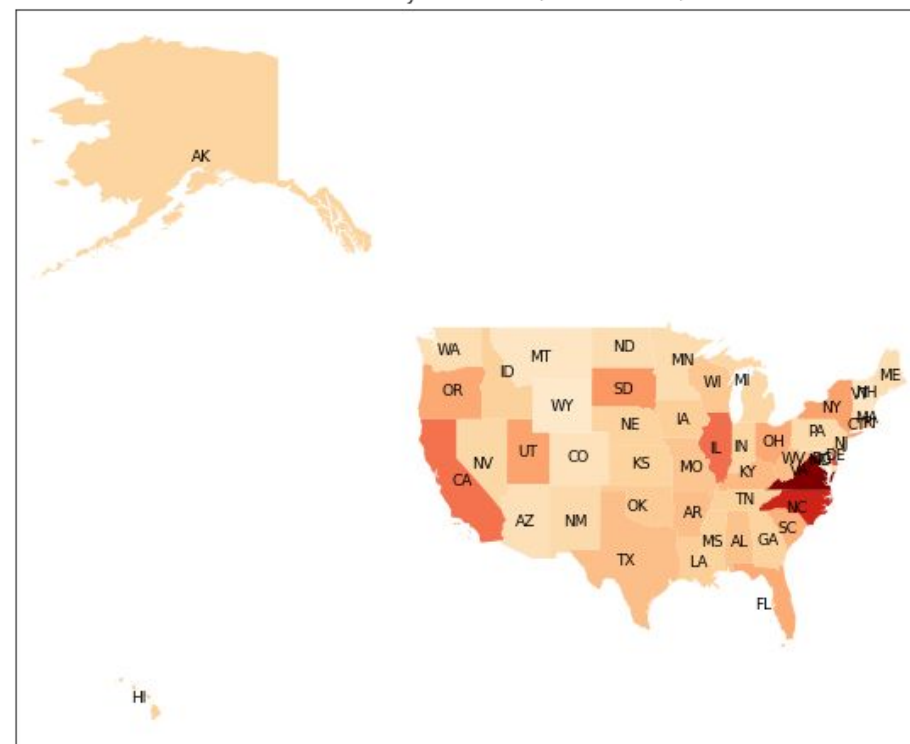
Thấu hiểu dữ liệu

State, BankState (cont)

CHGOFF rate by State (no Recession)



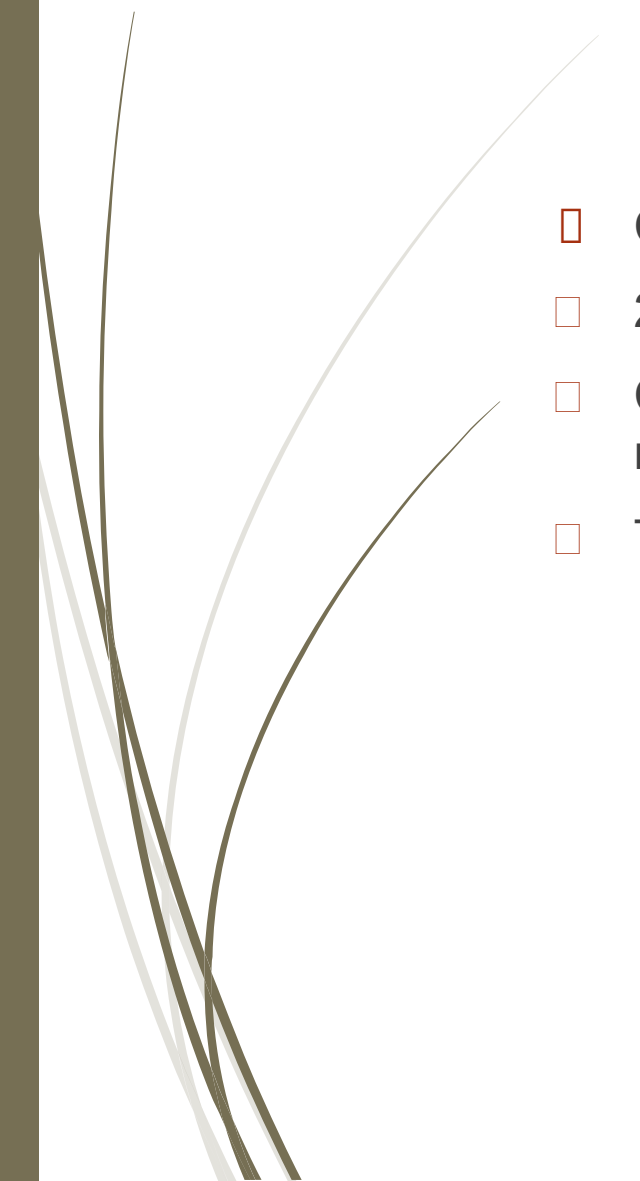
CHGOFF rate by BankState (no Recession)





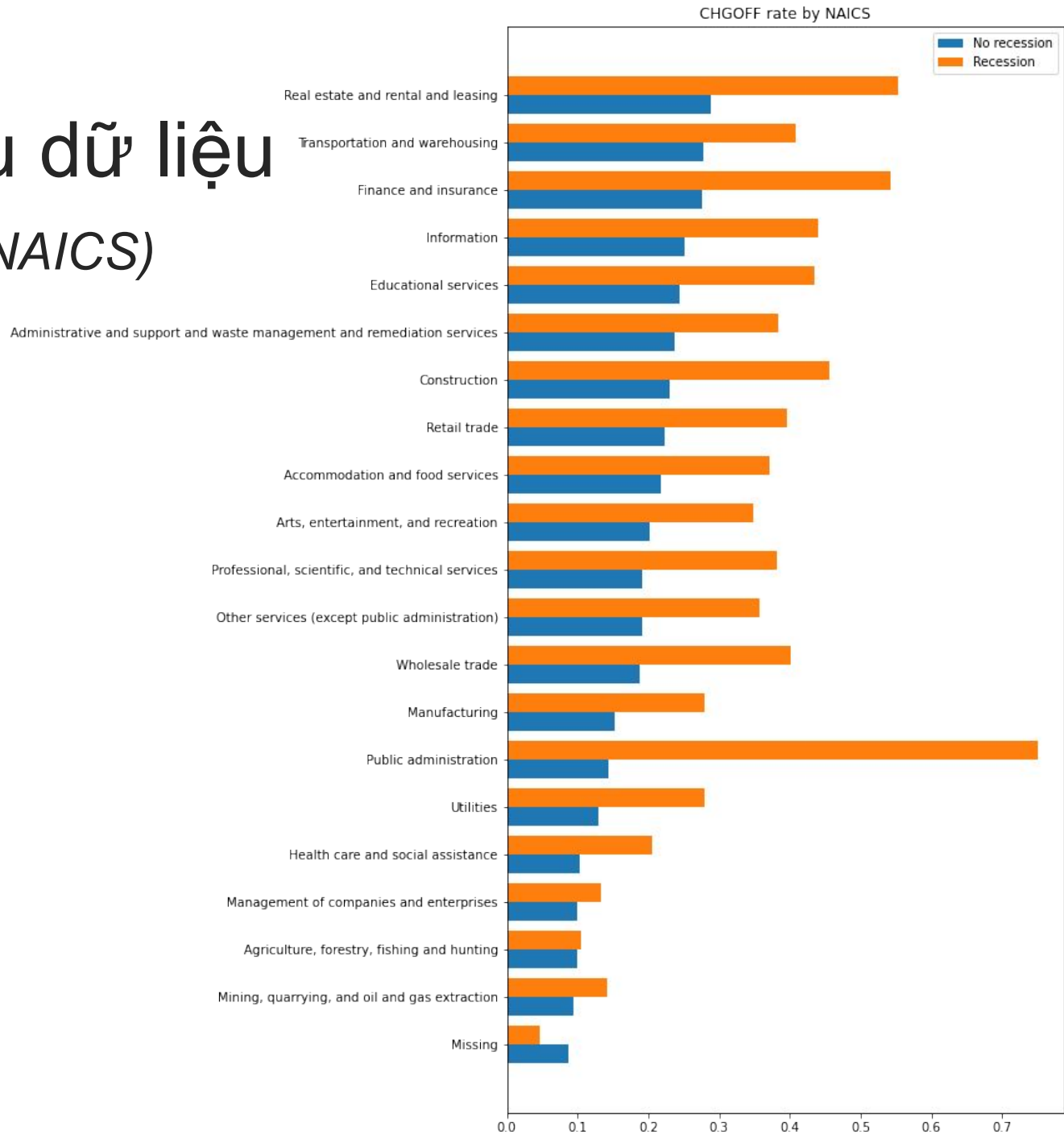
Thấu hiểu dữ liệu

Industry (NAICS)

- Có khoảng 150.000 trường hợp thiếu dữ liệu mã ngành gán category missing.
 - 2 ký tự đầu cột này được code từ 11 - 92 cho mỗi ngành
 - Các nhóm ngành khác nhau có mức độ rủi ro khác nhau có tỷ lệ CHGOFF sẽ khác nhau.
 - Trong thời kỳ suy thoái, sự thay đổi tỷ lệ CHGOFF là khác nhau giữa các ngành
- 

Thấu hiểu dữ liệu

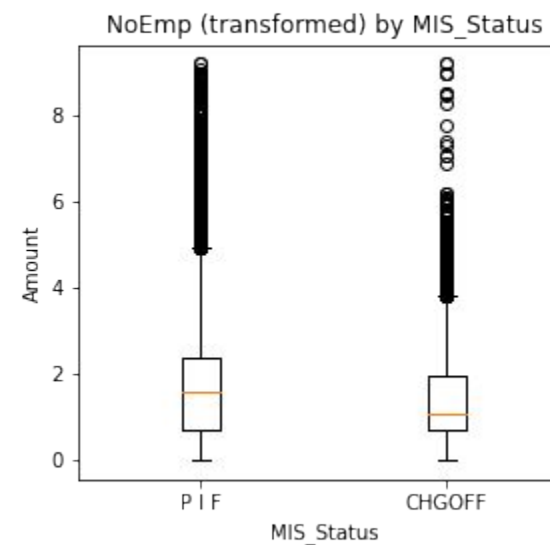
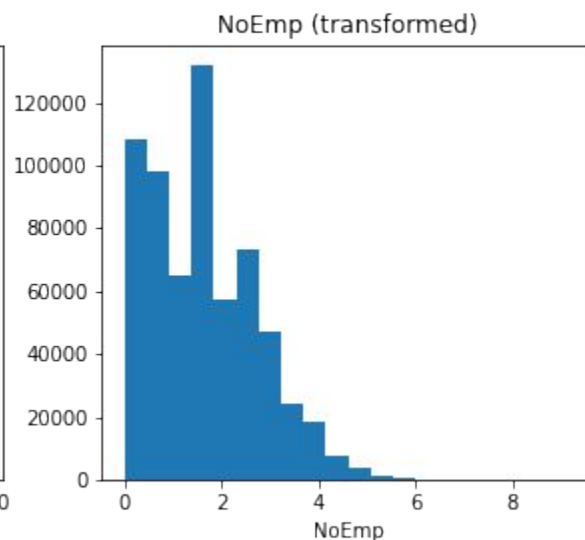
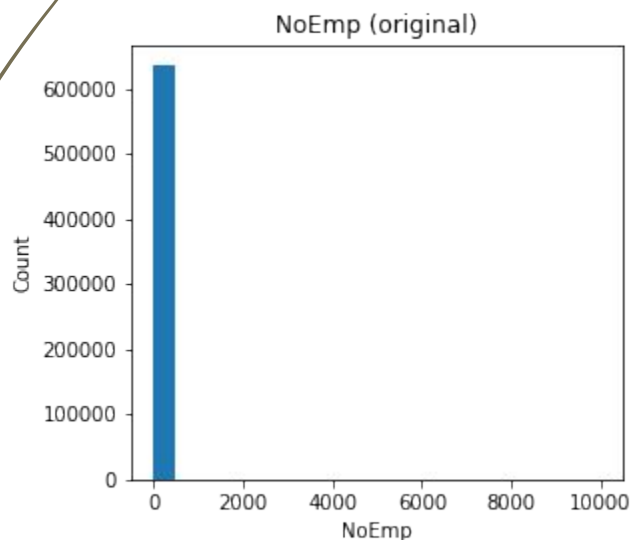
Industry (NAICS)



Thấu hiểu dữ liệu

NoEmp

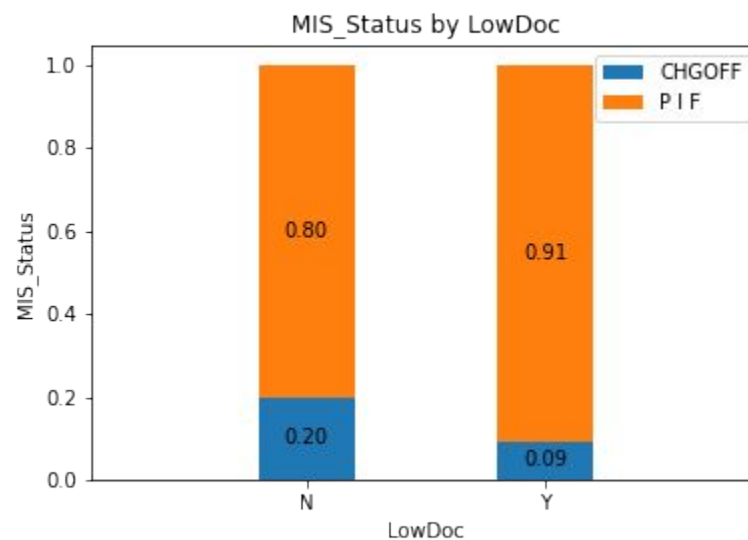
- Thực hiện biến đổi log, có thể thấy các khoản vay PIF thì Khách hàng có nhiều lao động hơn khoản vay CHGOFF (Quy mô lớn hơn thì có khả năng trả nợ tốt hơn).
- Có khoảng 3300 trường hợp số lao động bằng không có thể thay thế bằng giá trị median



Thấu hiểu dữ liệu

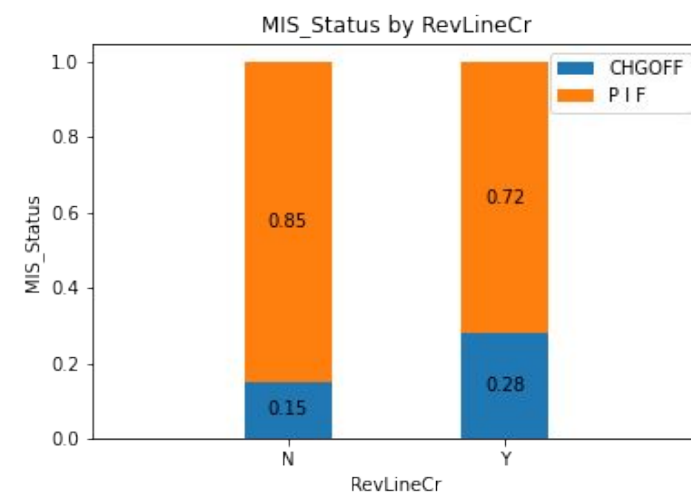
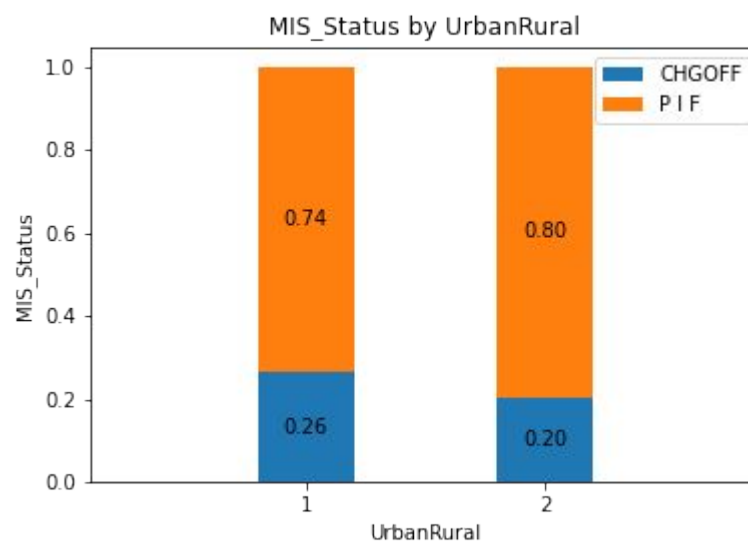
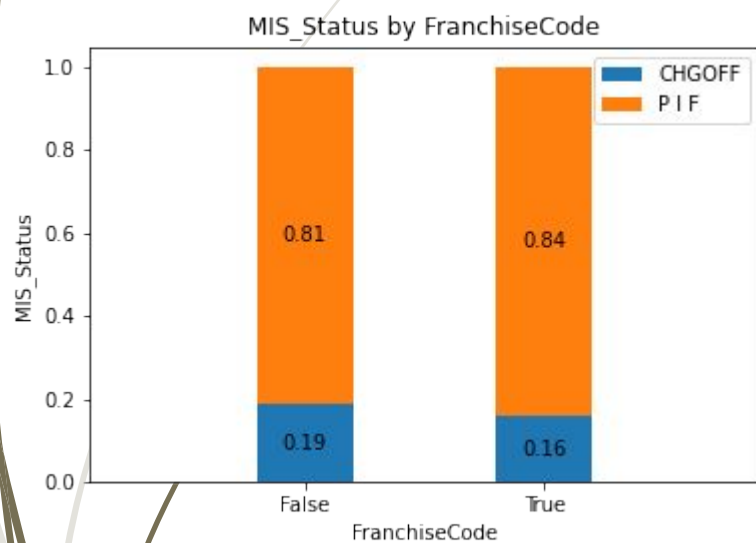
LowDoc

- Các trường hợp vay LowDoc có tỷ lệ CHGOFF thấp hơn vay thông thường
- Các trường hợp lỗi, thiếu dữ liệu có thể gán bằng giá trị mode.



Thấu hiểu dữ liệu

FranchiseCode, UrbanRural, RevLineCr



Thấu hiểu dữ liệu

Các thuộc tính loại bỏ

- ❖ LoanNr_ChkDgt: Record ID
- ❖ Name: Tạo ra thuộc tính ChgOffHist
- ❖ GrAppv, SBA_Appv: Tạo ra thuộc tính SBA_Appv_Ratio
- ❖ GhgOffPrinGr, ChgOffDate: Cùng ý nghĩa với MIS_Status
- ❖ City, Zip, Bank: Nhiều category (>27k), sử dụng State thay thế
- ❖ ApprovalDate, ApprovalFY, DisbursementDate: Tạo ra thuộc tính Recession
- ❖ CreateJob, RetainedJob: Tương tự NoEmp

Chuẩn bị dữ liệu

Data Pipeline

- Impute dữ liệu
- Dữ liệu số
 - + Chuyển đổi string \square float
 - + Chuyển đổi log
- Thời gian: Chuyển đổi string \square datetime
- Tạo feature (SBA_Appv_Ratio, Recession, ChgOffHist)



- Scale dữ liệu [0, 1]
- One-hot encode



RFECV
(Recursive Feature
Elimination with Cross
Validation)



X_train shape
(637255, 120)

- ❖ Tổng số thuộc tính 143
- ❖ Thuộc tính bị loại bỏ: 23
 - State: 10 state
 - BankState: 5 state
 - Industry: 8 industry
- ❖ Số thuộc tính còn lại: 120

Xây dựng mô hình

Logistic Regression

```
param_grid = {  
    'C': [1e-4, 1e-2, 1.0, 1e2, 1e4],  
    'fit_intercept': [False, True],  
    'class_weight': [None, 'balanced'],  
}  
lr_model = GridSearchCV(LogisticRegression(solver='lbfgs', max_iter=400,  
                                           n_jobs=-1, tol=0.0002),  
                        param_grid,  
                        cv=kfold,  
                        return_train_score=True,  
                        scoring='f1',  
                        verbose=1)
```

Xây dựng mô hình

Gradient Boosting

```
param_grid = {  
    'max_depth': [1, 3, 6, 9],  
    'learning_rate': [0.1, 0.25, 0.5],  
    'gamma': [0, 0.001],  
    'scale_pos_weight': [1.0, WEIGHT]  
}  
  
gb_model = GridSearchCV(xgb.XGBClassifier(use_label_encoder=False,  
                                          objective='binary:logistic',  
                                          booster='gbtree',  
                                          n_estimators=40,  
                                          n_jobs=-1),  
                        param_grid,  
                        cv=kfold,  
                        return_train_score=True,  
                        scoring='f1',  
                        verbose=1)
```


Đánh giá mô hình

F1-Score

Logistic model:

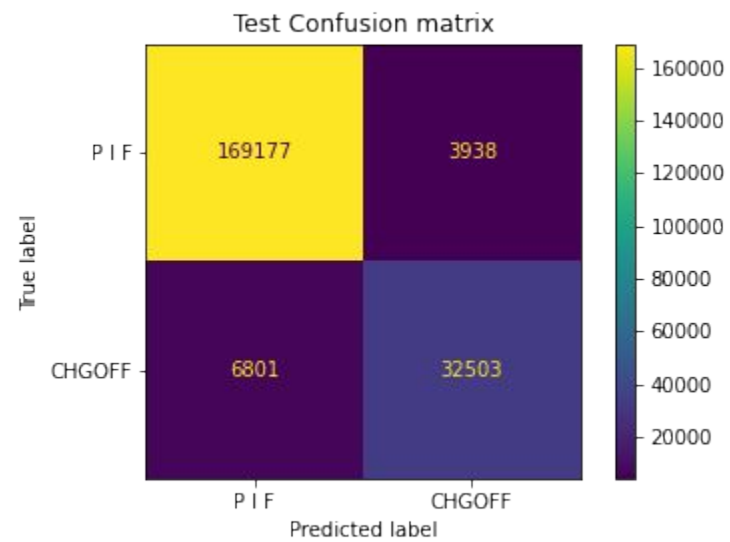
	precision	recall	f1-score	support
0	0.95	0.73	0.82	173115
1	0.41	0.82	0.54	39304
accuracy			0.74	212419
macro avg	0.68	0.78	0.68	212419
weighted avg	0.85	0.74	0.77	212419

Gradient Boosting model:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	173115
1	0.89	0.83	0.86	39304
accuracy			0.95	212419
macro avg	0.93	0.90	0.91	212419
weighted avg	0.95	0.95	0.95	212419

Đánh giá mô hình

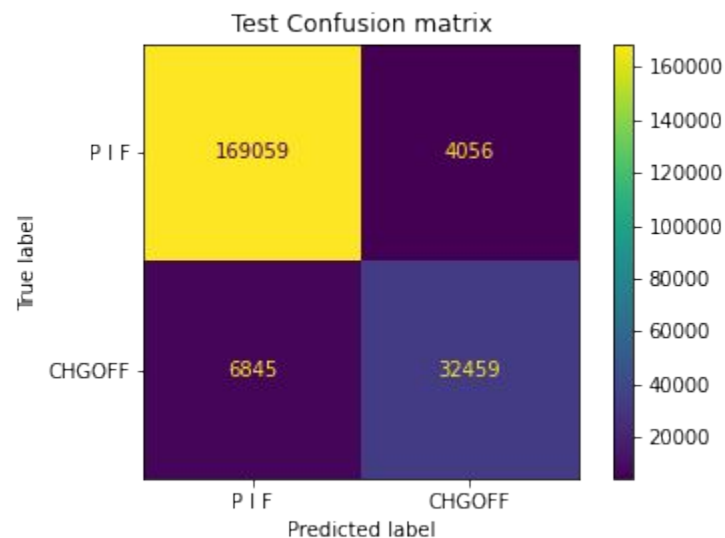
Confusion Matrix



Đánh giá mô hình

Phân tích lỗi

- Các trường hợp thiếu dữ liệu RevLineCr, UrbanRural chiếm tỷ lệ cao trong các trường hợp lỗi (34,4%, 35,3%)
- Thử impute các trường hợp này bằng giá trị mode, kết quả không được cải thiện so với việc cho trường hợp lỗi là 1 category.





Đánh giá mô hình

Kết luận

- ❖ Căn cứ F1 score có thể thấy mô hình Gradient Boosting cho kết quả tốt nhất.
 - ❖ Các thuộc tính RevLineCr, UrbanRural có số quan sát thiếu dữ liệu là lớn (>200k) so với mẫu ảnh hưởng tới kết quả
 - ❖ Phương hướng cải thiện mô hình: Xem xét lại các thu thập dữ liệu RevLineCr, UrbanRural
- 