

# Lending Club: IRR기반 수익성 평가모델 설계

발표일 2025.08.08

7조 이상재 김태완 이솔 박예빈 박정현 이도아



# IRR기반 수익성 평가모델

01



02



03



04



## 연구 소개

- 연구 목적
- 연구 차별점
- 1단계&2단계 모델
- 모델설계 프로세스

## 전처리 및 EDA

- 데이터셋 개요
- 결측치 및 이상치처리
- 범주형 변수 인코딩
- 스케일링

## 모델설계 및 학습

- 1차 분류모델
- 2차 회귀모델
- 3차 분류모델: 수익률 예측

## 결론 및 해석

- 주요 시사점
- 해석 가능성 확보
- 한계점 및 향후 과제

# 01

---

## 연구 소개

- 연구 목적
- 연구 차별점
- 1단계&2단계 모델
- 모델설계 프로세스

# 연구 목적

## 투자자의 수익을 극대화하는 신용평가 모델 개발



### Lending Club 실데이터 기반 신용평가 모델 개발

2007–2020년 약 175만 건의 P2P 대출 데이터를 활용하여  
부도 여부를 예측하는 모델 구축



### Sharpe Ratio 극대화를 목표로 한 성능 평가

원리금 균등상환 기반 IRR 계산 후, 대출 승인/거절 전략에  
따라 초과수익률과 위험을 반영한 Sharpe ratio를 극대화



### Train–Validation–Test 반복 실험

학습-검증-평가 데이터 분할을 수백 번 반복하여  
Sharpe ratio의 전반적인 분포 평가

연구목적 데이터 전처리 모델링 결론 및 해석

# Lending Club

Lending Club은 대출 신청자와 투자자를 직접 연결  
하는 P2P 금융 플랫폼이다. 대출심사는 LC가 진행하  
고, 투자자는 승인된 대출 중 원하는 건에 투자를 한다.



## 연구 차별점

# 수익률 극대화하는 모델을 구축하고 싶어요

우리가 구축한 모델은 말이져

- ▶ 이러한 점에서 의의가 있습니다



### 1. IRR 정의

부도 시에도 매달 같은 금액을 받았다고  
가정하여 IRR 계산



### 2. 초과수익률 정의

대출 여부와 상환 여부에 따라 4가지 경우로  
나누어 수익률 계산



### 3. 2단계 모델링

1단계 모델링과 2단계 모델링을 같이  
진행하여 두 모델의 성능 비교



# 1단계 & 2단계 모델

## 3가지 모델



부도 여부 예측:  
Classification



IRR 예측:  
Regression



수익성 여부 예측:  
Classification

## 분석한 모델 종류

CatBoost LGBM  
XGBRF XGB

Random Forest  
Decision Tree

## 타겟과 목적

**Loan\_status (부도 여부)**

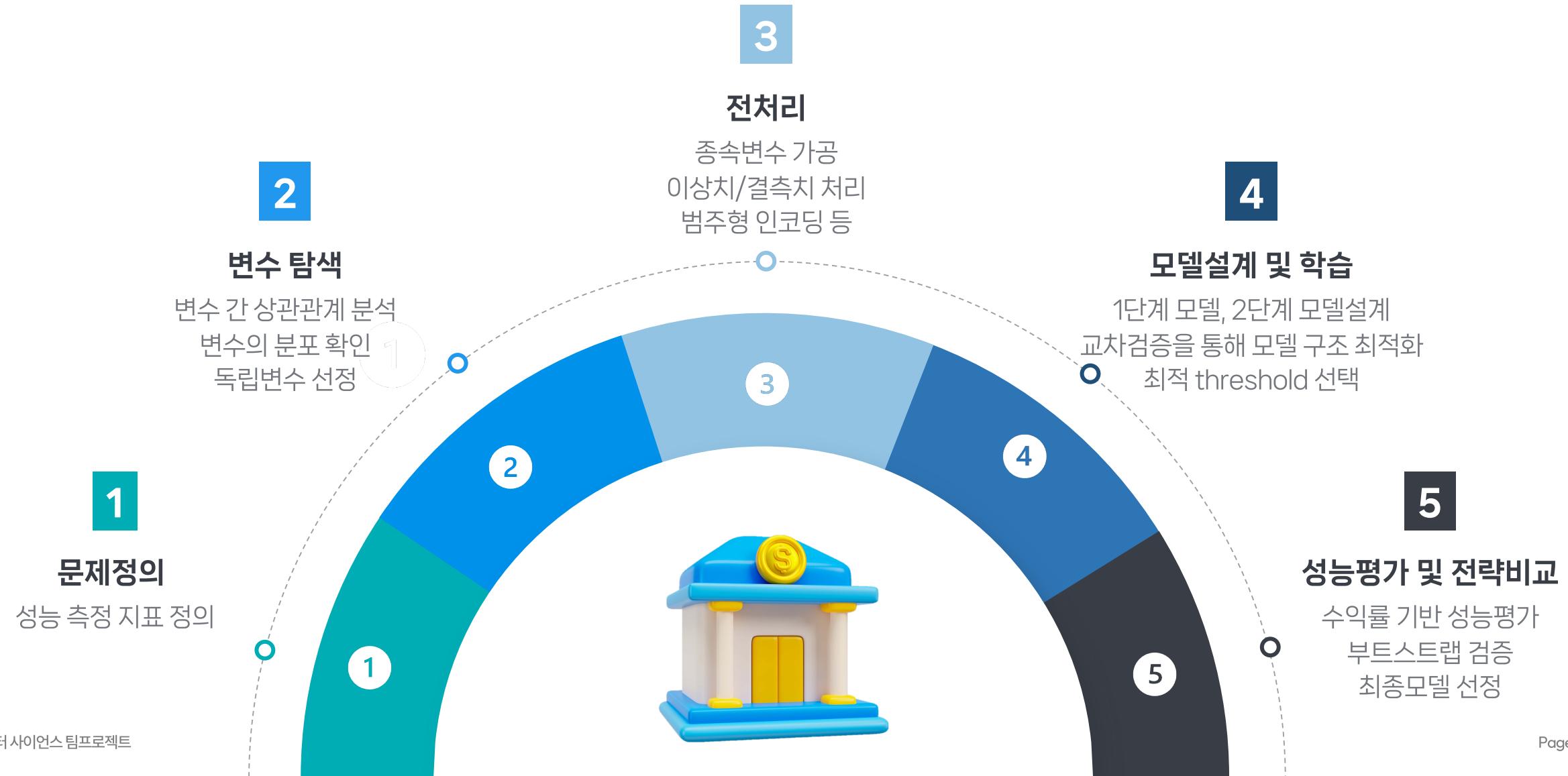
CatBoost Linear Regression  
XGBoost Random Forest

**IRR (수익률)**

XGB-CAT: 1단계 : XGBoost, 2단계 : CatBoost  
XGB-LGBM: 1단계 : XGBoost, 2단계 : LGBM  
XGB-XGB: 1단계 : XGBoost, 2단계 : XGBoost

**Profitable (수익성 여부)**

# 모델설계 프로세스



# 02

---

## 전처리 및 EDA

- 독립/종속 변수 선정 기준
- 데이터 전처리 방법
- 최종 선택한 변수

# | 대출 시기 구분



## 1. 대출절차 시작

대출신청액, 연간소득, DTI, 연체기록, 신용평가점수 등

**t<sub>initial</sub>**

**t<sub>investment</sub>**

**t<sub>post</sub>**

**t<sub>final</sub>**

## 2. 대출완료

대출금액



## 3. 대출상환중

받은 원금 총액, 받은 이자 총액, 남은 원금,  
계좌잔액, 연체금액 등



## 4. 기한 만료

파산판정



# 데이터 전처리 방법

모델학습하기 전에, 이상치, 결측치등을 적절히 처리하고, 범주형 변수를 인코딩하는 단계



# 01. 독립변수 선정

모델학습에 적절하지 않은 변수들을 제거

1



## 공동대출(join) 변수 제거

application\_type = 'joint app'  
컬럼

주대출과 비교했을 때 가중치  
를 부여하기 애매함

2



## 사후변수 제거

예: hardship, settlement  
관련 컬럼 등

대출 심사할 때 사용할 수 있  
는 정보만 사용하기 위함

3



## 결측치가 50% 이상인 변수 제거

예: hardship, settlement  
관련 컬럼 등

4



## 그 외 부적절한 변수 제거

예: 식별자, 아예 관계없는 변  
수(title 등), 상관관계 높은 변  
수들

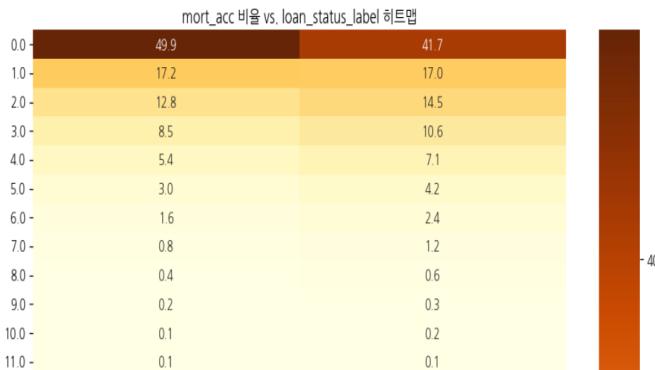
# 02. 이상치, 결측치 처리

## 1 이상치 처리 방법

### Clipping으로 극단값 처리

이상치는 EDA 결과 보면서 변수에 따라 다른 처리방법

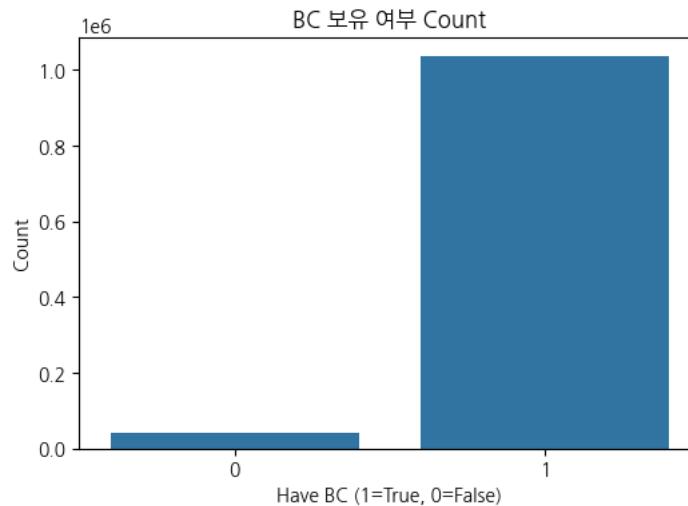
ex) mort\_acc의 경우 6 이상의 값들은 클리핑 처리



## 2 결측치 처리 방법

### 1) 0으로 대체

ex) BC 카드 보유 계좌 수의 경우



### 2) 중앙값으로 대체

### 3) 결측치 가진 행 drop

#### BC 카드/리볼빙 카드 관련 컬럼

(예: bc\_open\_to\_buy, num\_bc\_tl 등)

- 공통된 결측치는 해당 계좌가 없음으로 가정하여 0으로 대체

- have\_bc, have\_rev와 같이 더미 변수화 처리

#### 범주형 컬럼(예: emp\_length)

- 결측치는 원-핫 인코딩 적용

- 적용 예시: emp\_length\_unknown

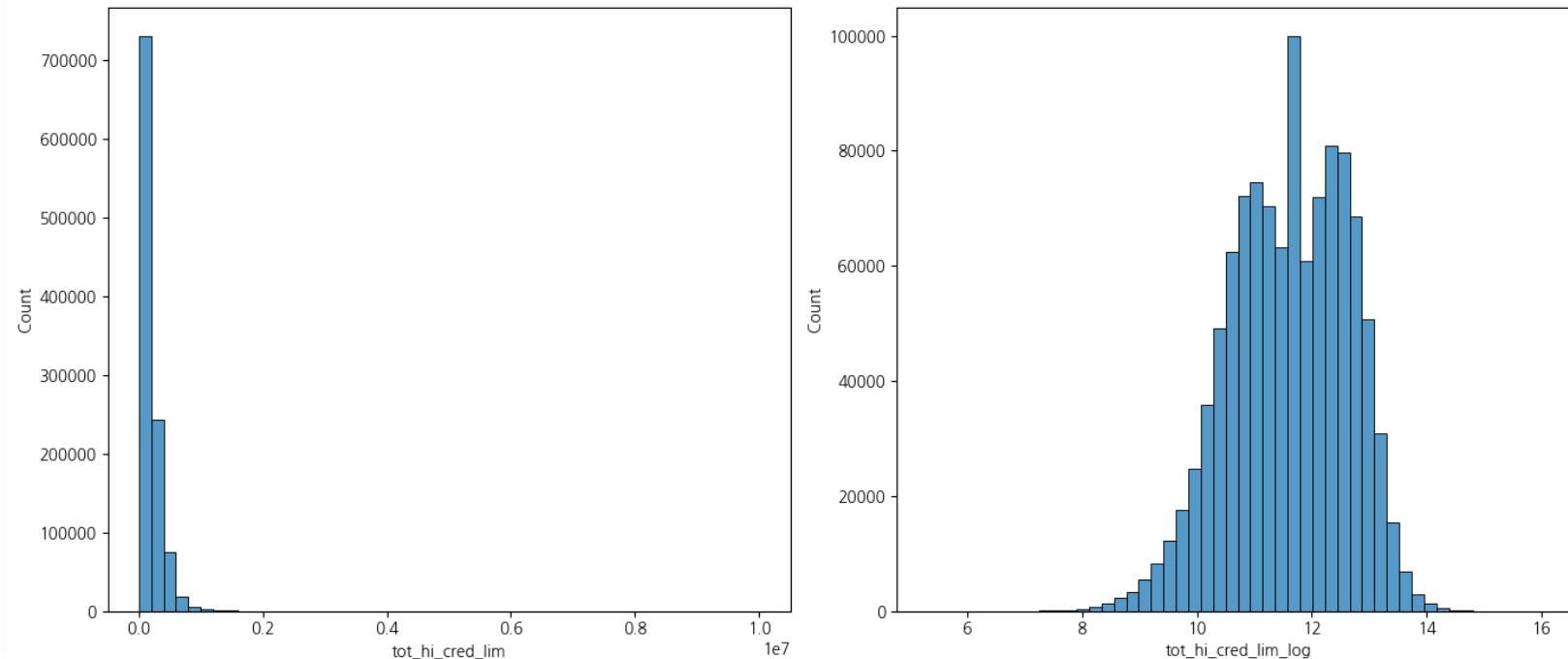
# 03. 로그변환

분포 치우친 변수 정규화

3

## 로그변환

'tot\_hi\_cred\_lim'처럼 분포 치우친 변수 로그변환

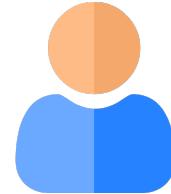


## 1 대출관련 변수

대출조건 및 승인 당시의 기본 정보

term emp\_length home\_ownership purpose  
verification\_status installment loan\_amnt  
funded\_amnt

총 7개



## 2 신용관련 변수

개인신용점수, 과거 연체이력 등

fico\_range\_high, fico\_range\_low, revol\_util  
delinq\_2yrs, mths\_since\_last\_delinq,  
mths\_since\_last\_major\_derog  
mths\_since\_last\_record, mths\_since\_recent\_bc,  
mths\_since\_recent\_bc\_dlq  
mths\_since\_recent\_inq,  
mths\_since\_recent\_revol\_delinq, inq\_last\_6mths  
num\_rev\_tl\_bal\_gt\_0, num\_rev\_accts, num\_bc\_tl  
num\_il\_tl, num\_actv\_bc\_tl, num\_actv\_rev\_tl  
num\_bc\_sats, num\_op\_rev\_tl, mo\_sin\_rcnt\_rev\_tl\_op  
mo\_sin\_rcnt\_tl, mo\_sin\_old\_rev\_tl\_op,  
mo\_sin\_old\_il\_acct  
num\_tl\_op\_past\_12m, num\_tl\_90g\_dpd\_24m,  
num\_accts\_ever\_120\_pd  
chargeoff\_within\_12\_mths,  
collections\_12\_mths\_ex\_med, tot\_coll\_amt  
total\_rev\_hi\_lim, acc\_now\_delinq, bc\_open\_to\_buy  
bc\_util, tot\_hi\_cred\_lim, pct\_tl\_nvr\_dlq

총 36개



## 3 재무상태 및 소득변수

개인 재무상태

annual\_inc, dti, total\_acc  
total\_bal\_ex\_mort, credit\_hist\_months, emp\_title  
tax\_liens, total\_rev\_hi\_lim, total\_il\_high\_credit\_limit  
bc\_ratio, num\_op\_rev\_tl, mo\_sin\_rcnt\_tl  
num\_bc\_sats, num\_bc\_tl, mo\_sin\_rcnt\_rev\_tl\_op  
num\_il\_tl, installment, funded\_amnt

총 18개

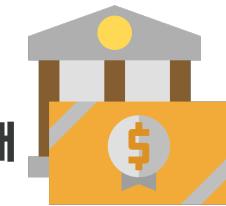


## 4 외부정보

미국 국채수익률

treasury

총 1개



# 종속변수 전처리

1



## loan\_status(종속변수)

'Fully Paid', 'Current', 'Charged Off', 'Late (16-30 days)',  
'Late (31-120 days)', 'In Grace Period', 'Issued', 'Does  
not meet the credit policy. Status:Fully Paid', 'Does not  
meet the credit policy. Status:Charged Off', 'Default'

2

## 전처리 코드 입력

```
raw_df = raw_df[raw_df['loan_status'].isin( ['Fully Paid', 'Charged Off', 'Does not meet the credit policy. Status:Fully Paid', 'Does not meet the credit policy. Status:Charged Off', 'Default'])]
```

2

## 5개만 선택 후

정상상환 (0)  
연체/부실 (1)로 라벨링



# 03

---

## 모델설계 및 학습

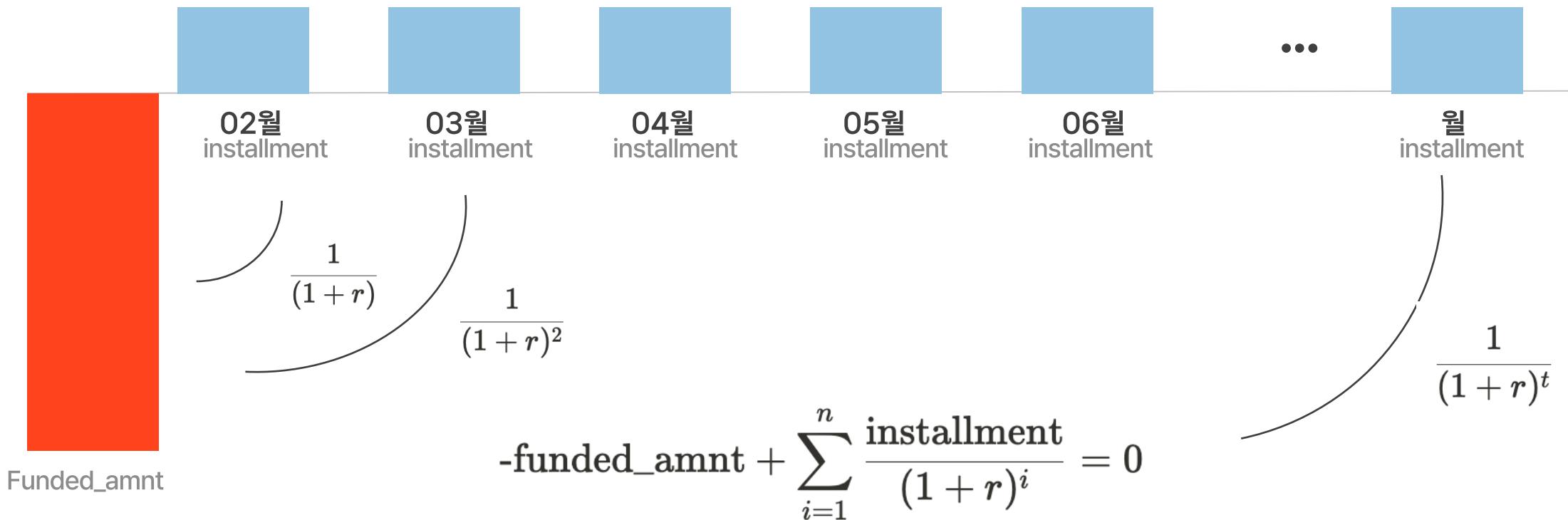
- IRR, Sharp ratio 개념 정의
- 베이스라인 모델
- 1번 분류모델: 부도예측
- 2번 회귀모델: IRR 예측
- 3번 분류모델: 수익성 여부

## 수익률 계산

## IRR 개념 정의

## 정상상환한 버전

매월 원금+이자 균등상환

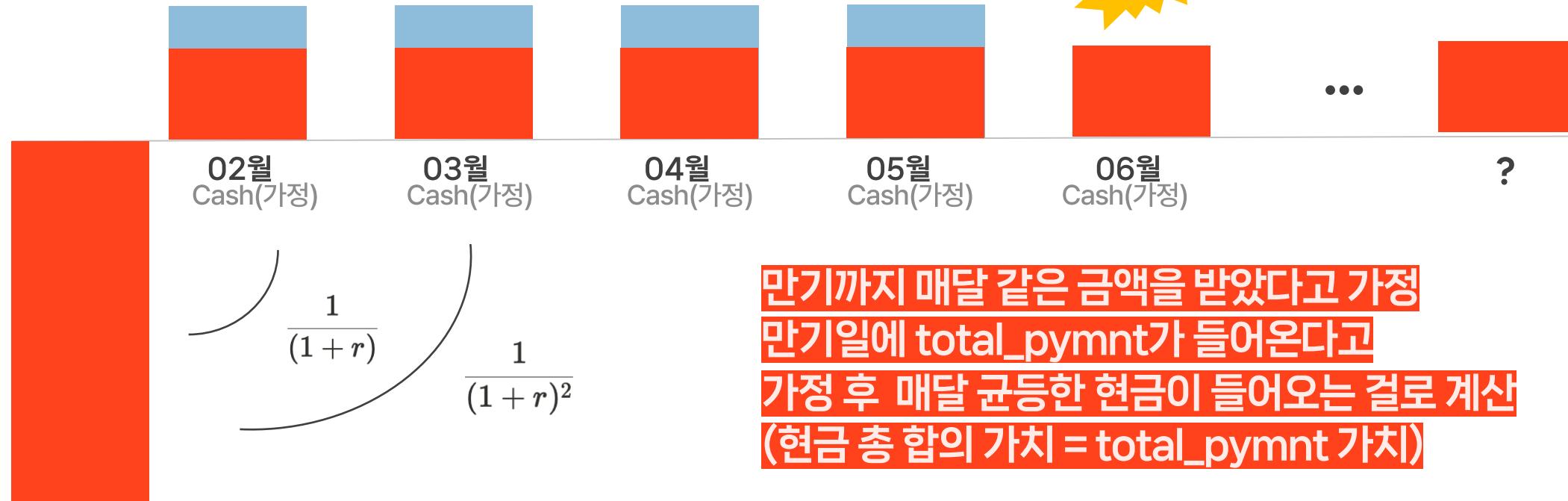


## 수익률 계산

## IRR 개념 정의

## 상환 못한 버전

매월 원금+이자 균등상환



만기까지 매달 같은 금액을 받았다고 가정  
 만기일에 total\_pymnt가 들어온다고  
 가정 후 매달 균등한 현금이 들어오는 걸로 계산  
 (현금 총 합의 가치 = total\_pymnt 가치)

그 이후 IRR 구하는 과정은 정상상환과 동일

# Sharp Ratio 개념 정의

연구목적 데이터 전처리 모델링 결론 및 해석

대출 & 상환결과 분류

1

대출 O -> 상환 O

2

대출 O -> 상환 X

3

대출 X -> 상환 O

4

대출 X -> 상환X

return

irr

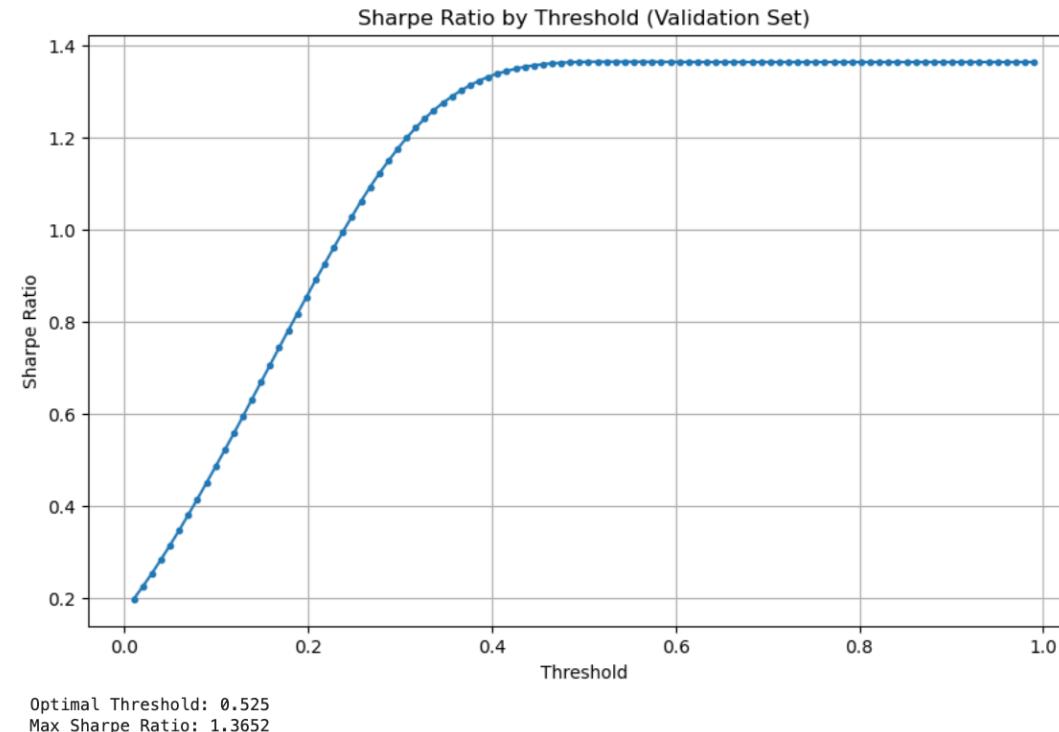
0

treasury(미국 국채금리)

treasury(미국 국채금리)

$$Sharpe = \frac{R_p - R_{treasury}}{\text{std}(R_p - R_{treasury})}$$

# 선형회귀 모델



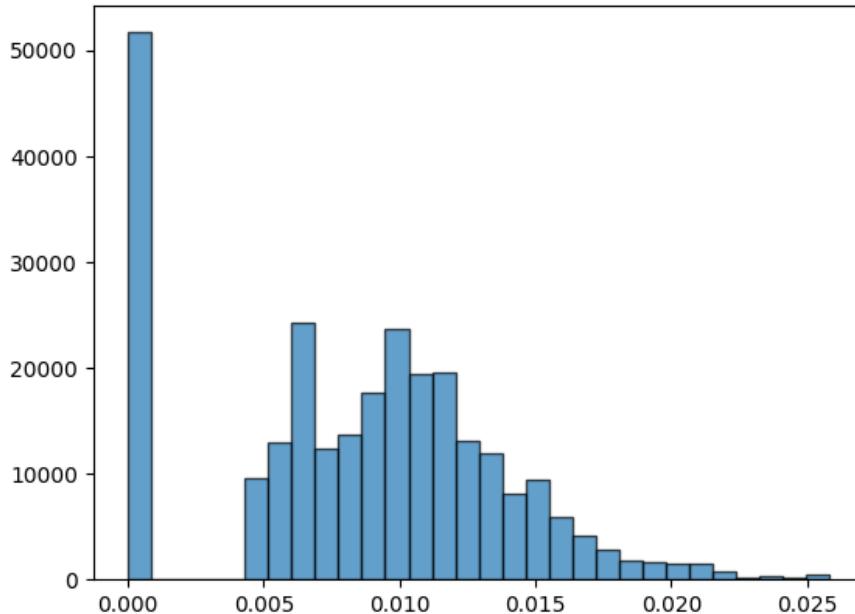
Optimal Threshold: 0.525

Max Sharp Ratio: 1.3652

# 부도 처리 방법에 따른 Return 분포 비교

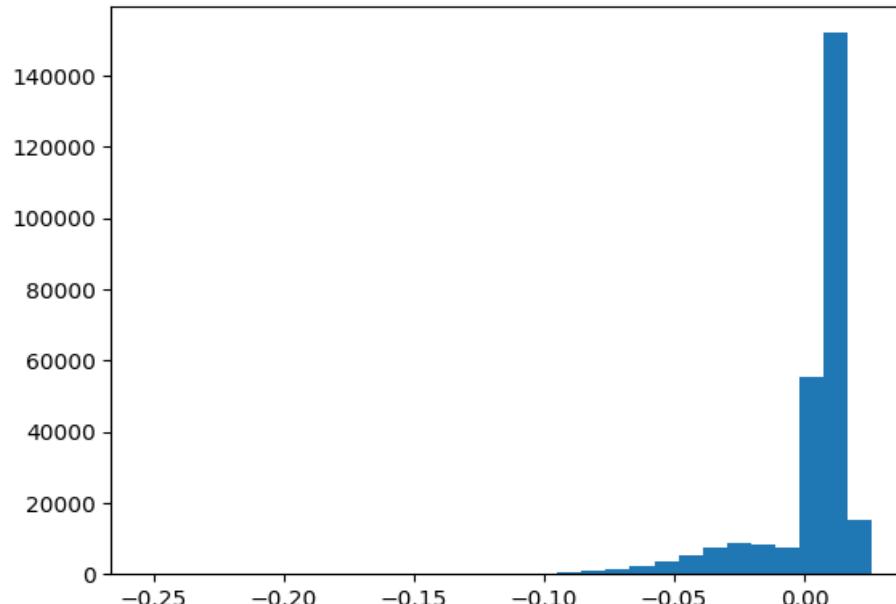
연구목적 데이터 전처리 모델링 결론 및 해석

## 01 부도 건의 Return을 0으로 처리



대출했으나 상환 못함 → Return을 0으로 반영

## 02 부도 건에도 IRR 반영



상환 못했을 때에도 음수 IRR을 그대로 반영

# Sharp Ratio 개념 정의

연구목적 데이터 전처리 모델링 결론 및 해석

## 대출 & 상환결과 분류

1

대출 O -> 상환 O

2

대출 O -> 상환 X

3

대출 X -> 상환 O

4

대출 X -> 상환X

## return

irr

irr

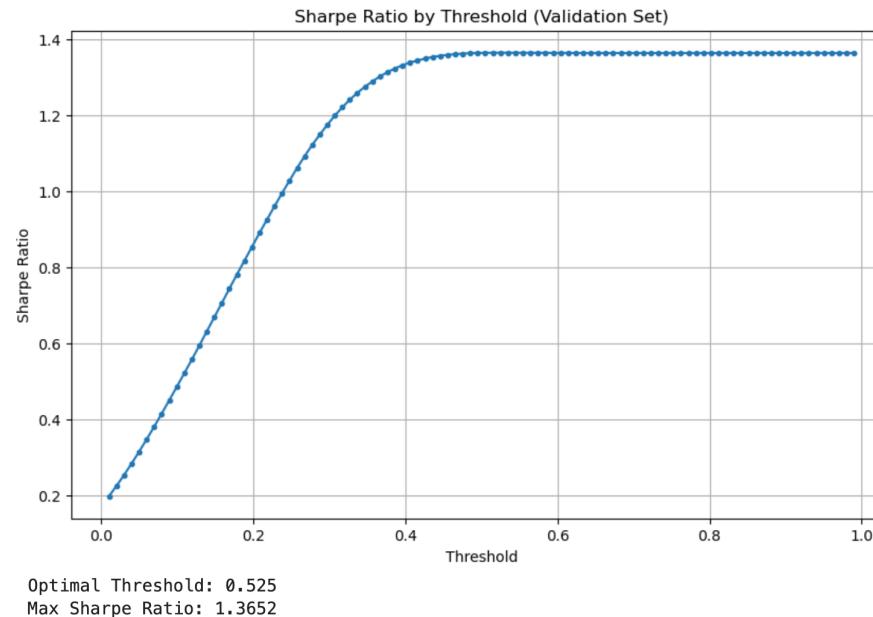
treasury

treasury

$$Sharpe = \frac{R_p - R_{treasury}}{std(R_p - R_{treasury})}$$

# 00. 베이스라인 모델 : loan\_status 예측

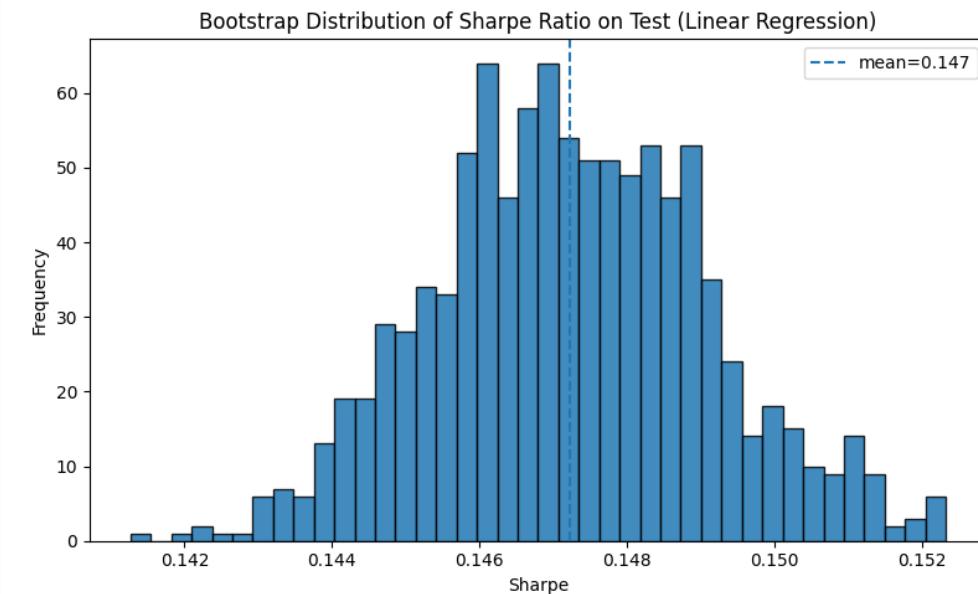
## 01 선형분류모델



Optimal Threshold: 0.337

Max Sharp Ratio: 0.1058

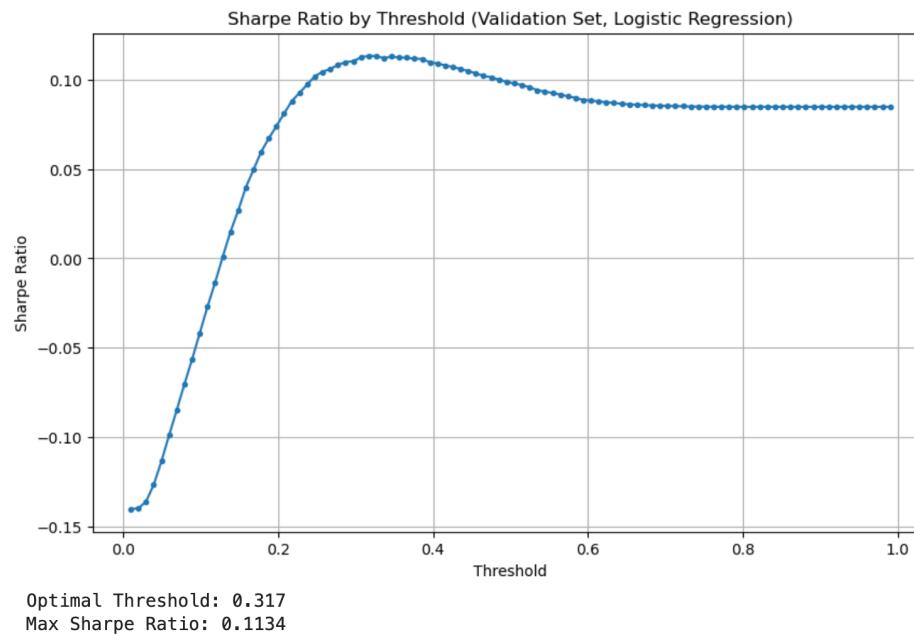
## 01 선형분류모델



평균 Sharp Ratio: 0.1037 S.D Sharp Ratio: 0.0020

# 00. 베이스라인 모델 : loan\_status 예측

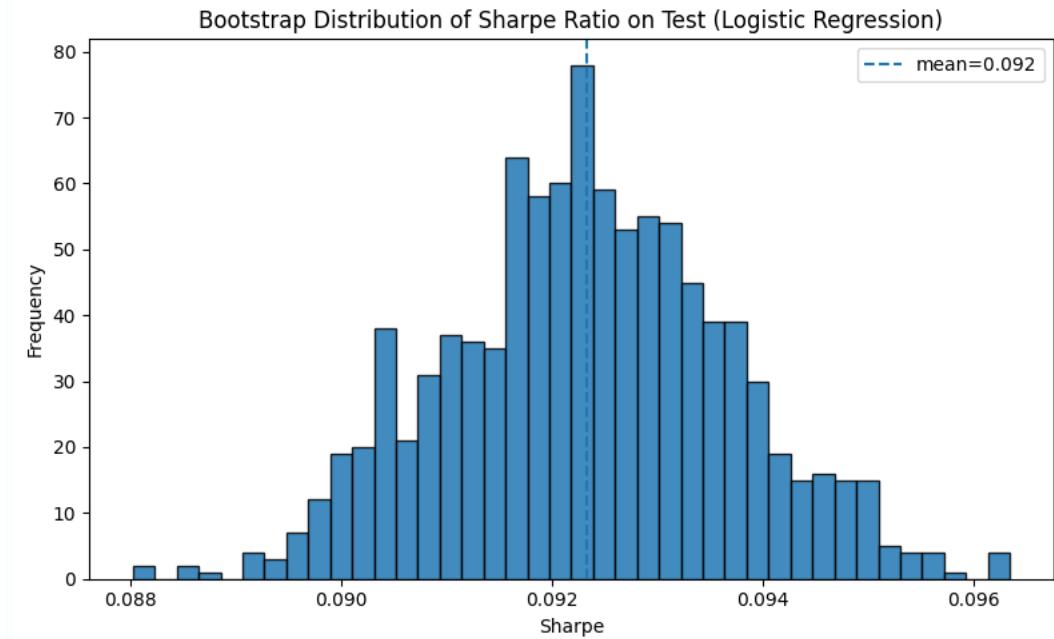
## 02 로지스틱 회귀모델



Optimal Threshold: 0.317

Max Sharp Ratio: 0.1134

## 02 로지스틱 회귀모델



평균 Sharp Ratio: 0.1112      S.D Sharp Ratio: 0.0019

# 1차 분류모델 (Classification) 결과

[이진분류모델성능 평가 결과]

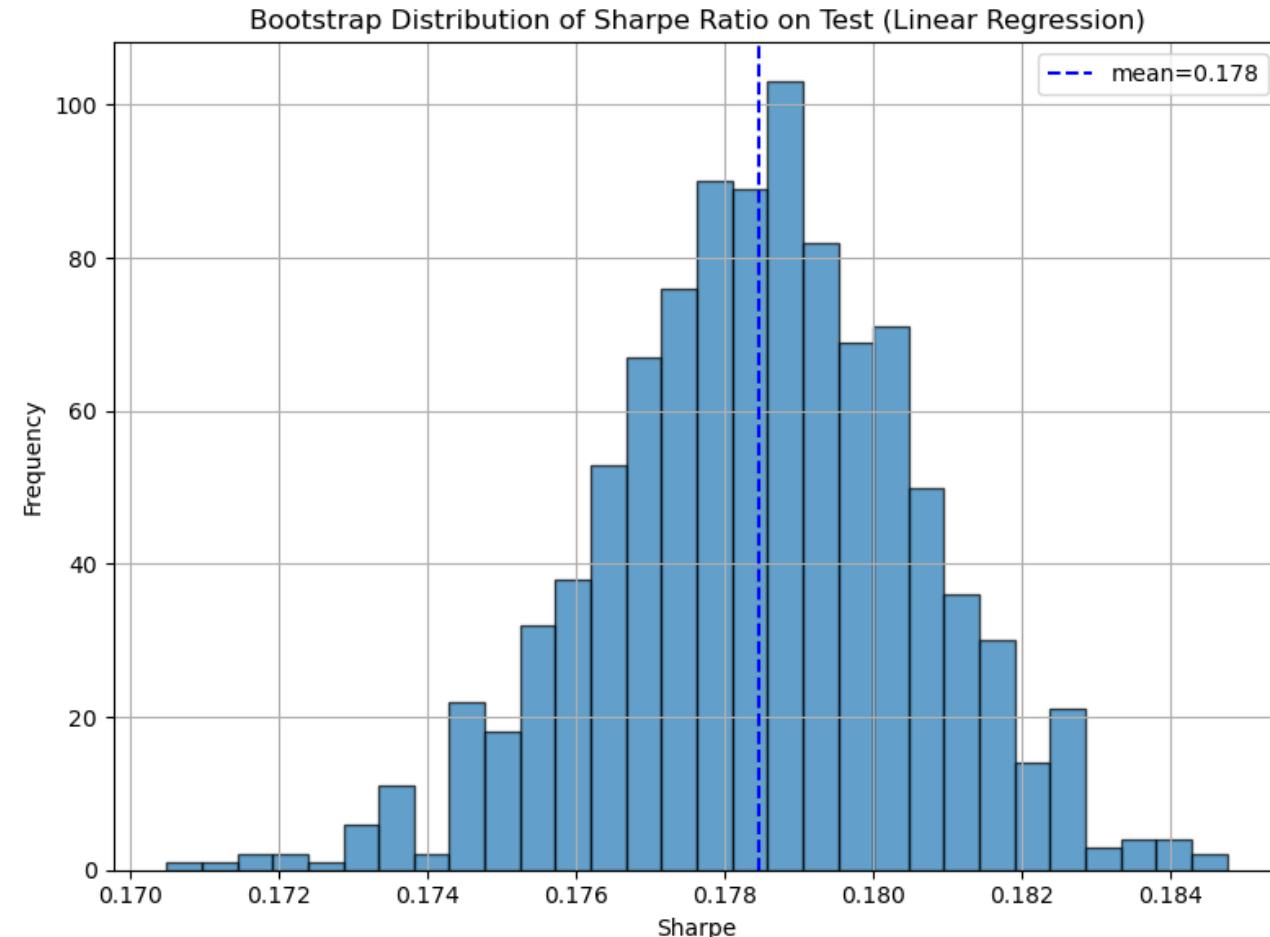
성능평가 지표	CatBoost	LGMB	Random Forest	XGBRF	XGB	Decision Tree
Threshold	<b>0.32</b>	0.3070	0.346	0.65	0.42	0.5148
Sharp ratio	<b>0.1392</b>	0.1232	0.1113	0.0901	0.1047	0.0875

CatBoost 모델이 Sharpe Ratio 0.139로, 위험 대비 수익률이 가장 우수한 것으로 나타남. 해당 모델이 예측한 포트폴리오가 안정적인 수익을 낼 가능성이 높다는 의미

# 1차 모델: CatBoost: Test set

부트스트랩 (n=1000)

## Sharp Ratio 분포



mean : 0.1785

Std : 0.0021

## 2차 회귀모델 결과: IRR 예측

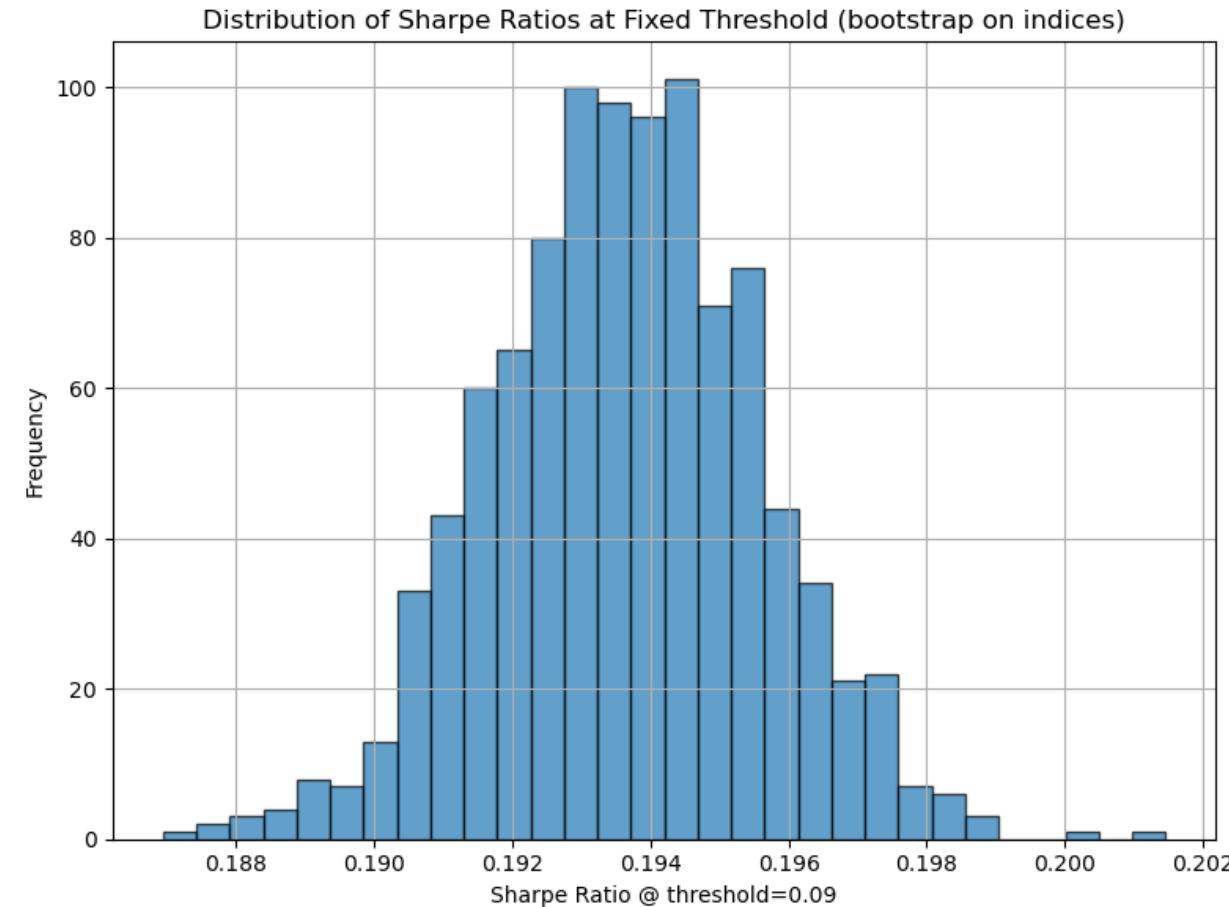
[회귀모델 성능 평가 결과]

성능평가 지표	Cat Boost	Linear Regression	Random Forest	XGBoost
Threshold	0.149	0.545	0.109	0.119
Sharp ratio	0.1565	0.1339	0.1330	0.1985

XGBoost 모델이 Sharpe Ratio 0.1985로, 위험 대비 수익률을 가장 우수한 것으로 나타남. 해당 모델이 예측한 포트폴리오의 초과수익률의 변화에 민감하게 반응한다(Sharp ratio의 분모가 큼)

# 2차모델 XGBoost: Test set 부트스트랩 (n=1000)

## Sharp Ratio 분포



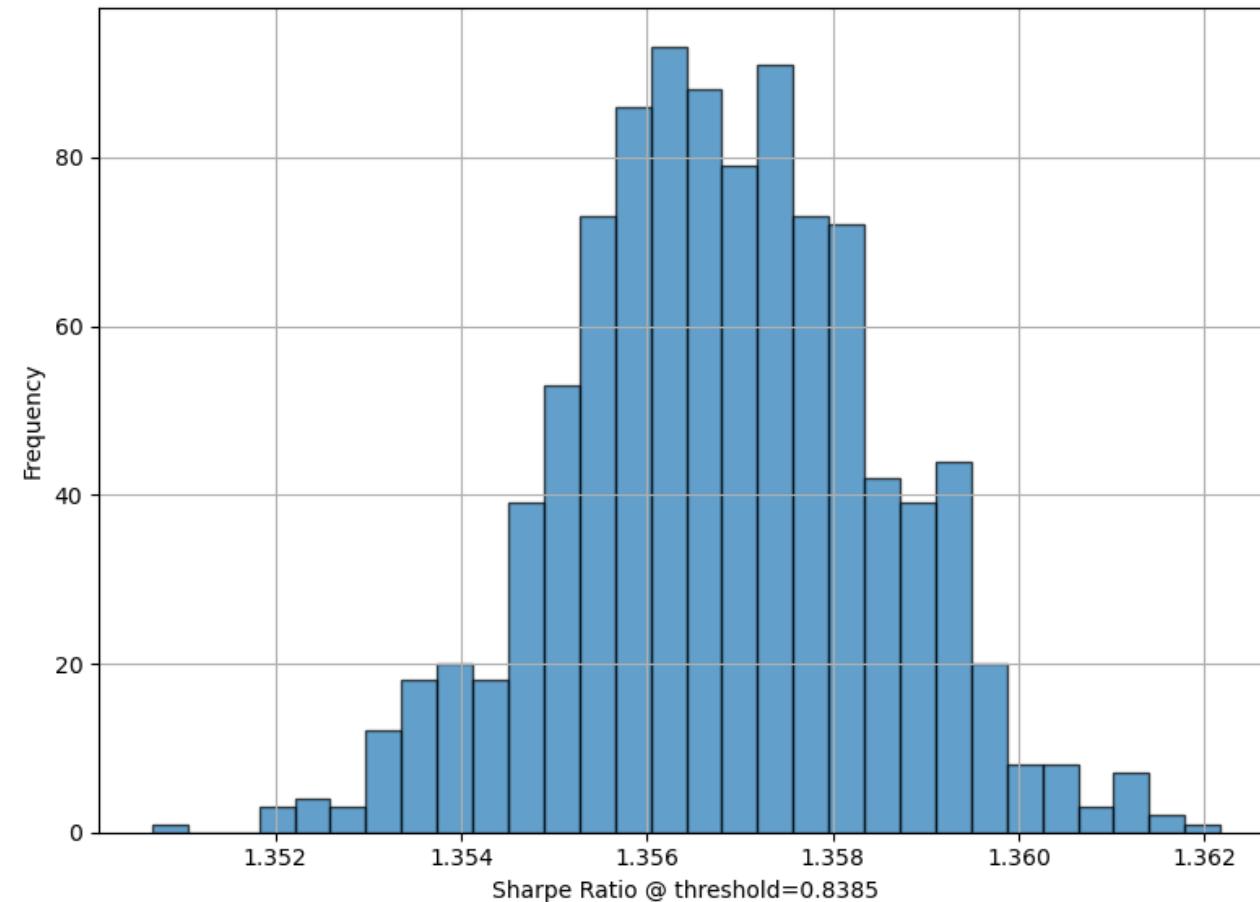
# 2차모델 XGBoost: Test set

부트스트랩 (n=1000)

대출 이후 상환X 0으로 처리한 경우

## Sharp Ratio 분포

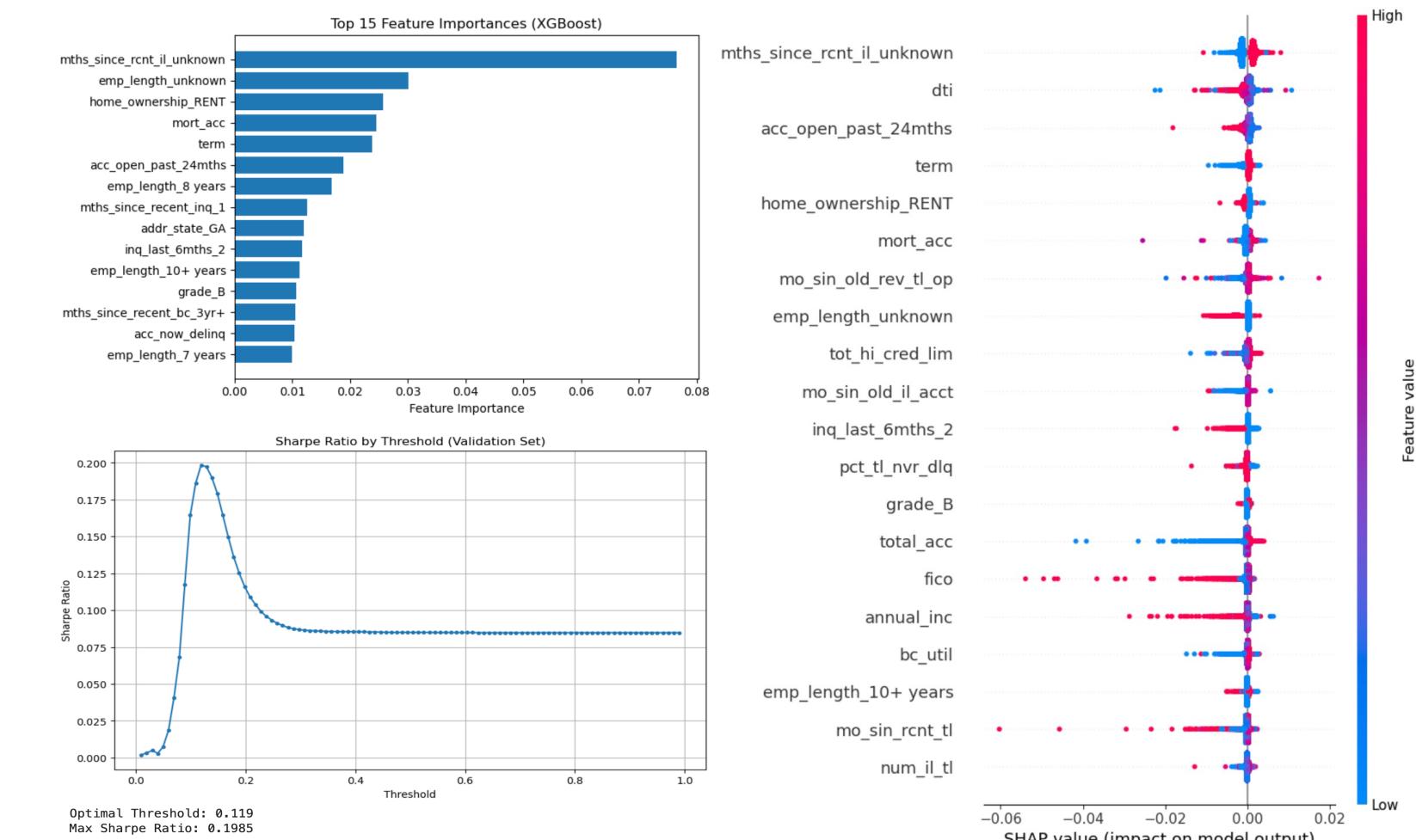
Distribution of Sharpe Ratios at Fixed Threshold (bootstrap on indices)



mean : 1.3568

Std : 0.0017

## 2차모델 XGBoost :



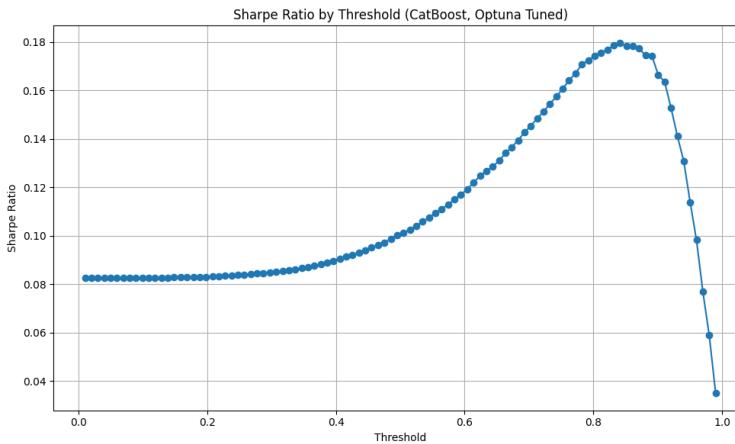
이 XGBoost 회귀모델에서는 '최근 대출 관련 정보의 부재(`mths_since_rcnt_il_unknown`)'가 수익률 예측에 가장 큰 영향을 미치며, `dti`가 높을수록 IRR이 낮아짐.

# 3차 수익성 분류모델 결과

## 1 XGB-CAT

최적 threshold: 0.8415

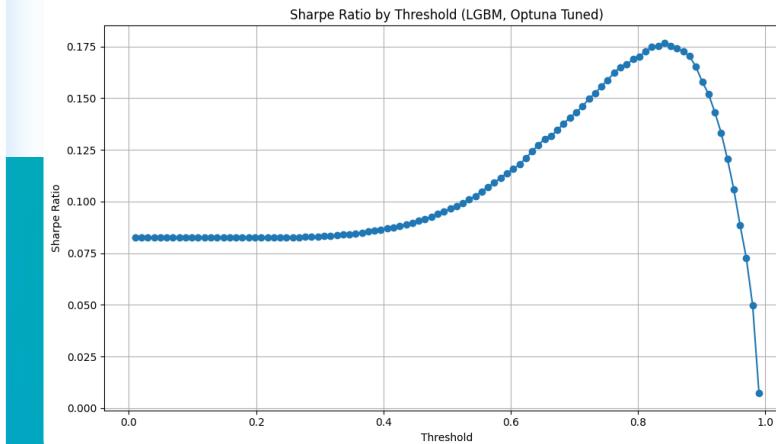
최대 Sharp Ratio: 0.1844



## 2 XGB-LGBM

최적 threshold: 0.8415

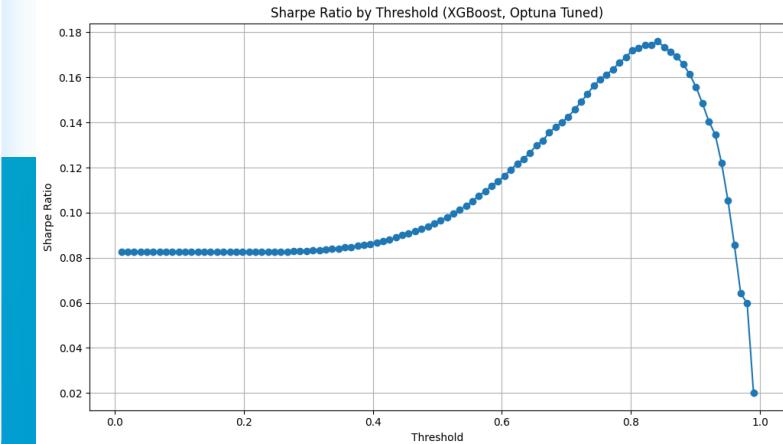
최대 Sharp Ratio: 0.1817



## 3 XGB-XGB

최적 threshold: 0.8415

최대 Sharp Ratio: 0.1807



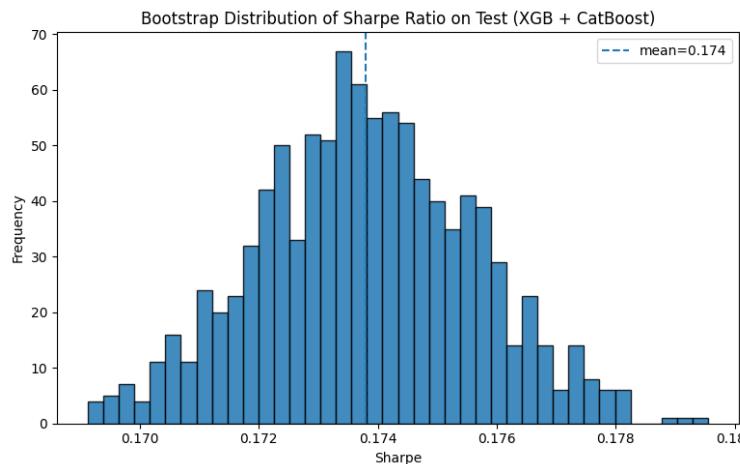
# 3차 수익성 분류모델 Test set

## 부트스트래핑 결과 (n=1000)

### 1 XGB-CAT

Sharp Mean: 0.1738

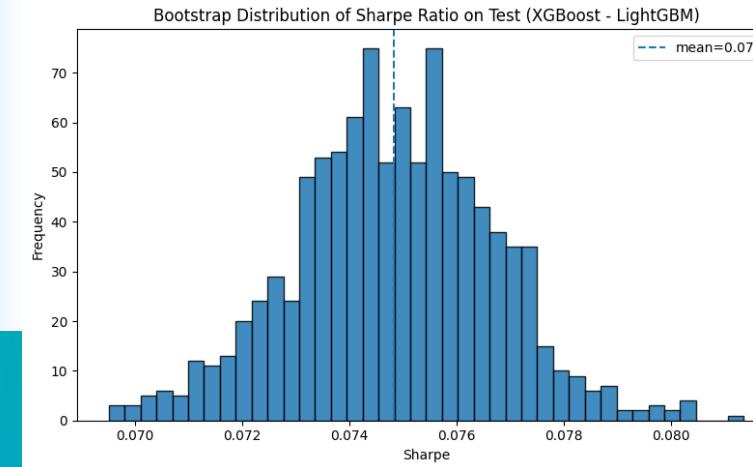
Sharp Std: 0.0018



### 2 XGB-LGBM

Sharp Mean: 0.0748

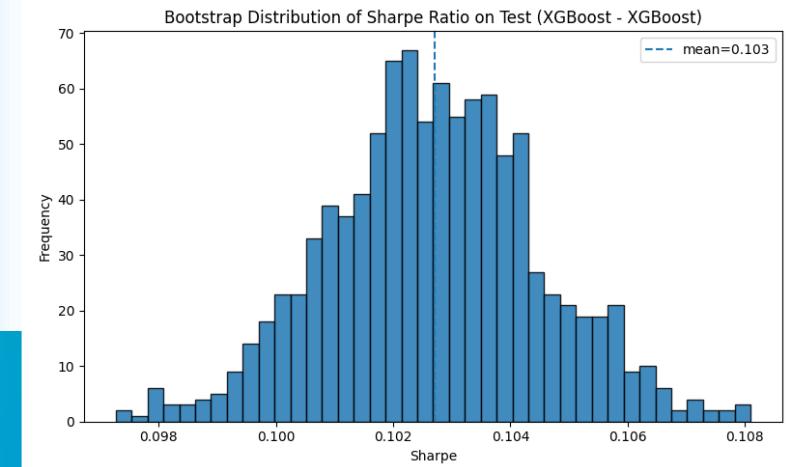
Sharp Std: 0.0019



### 3 XGB-XGB

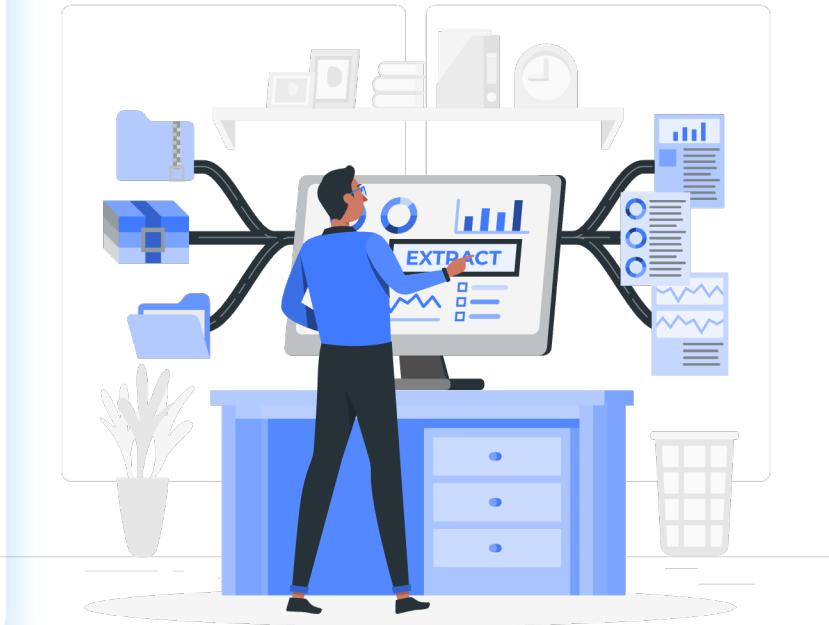
Sharp Mean: 0.1027

Sharp Std: 0.0018



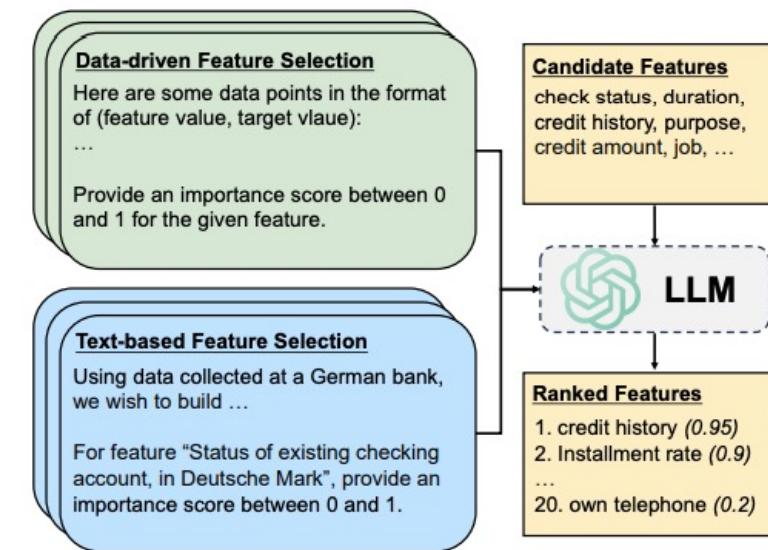
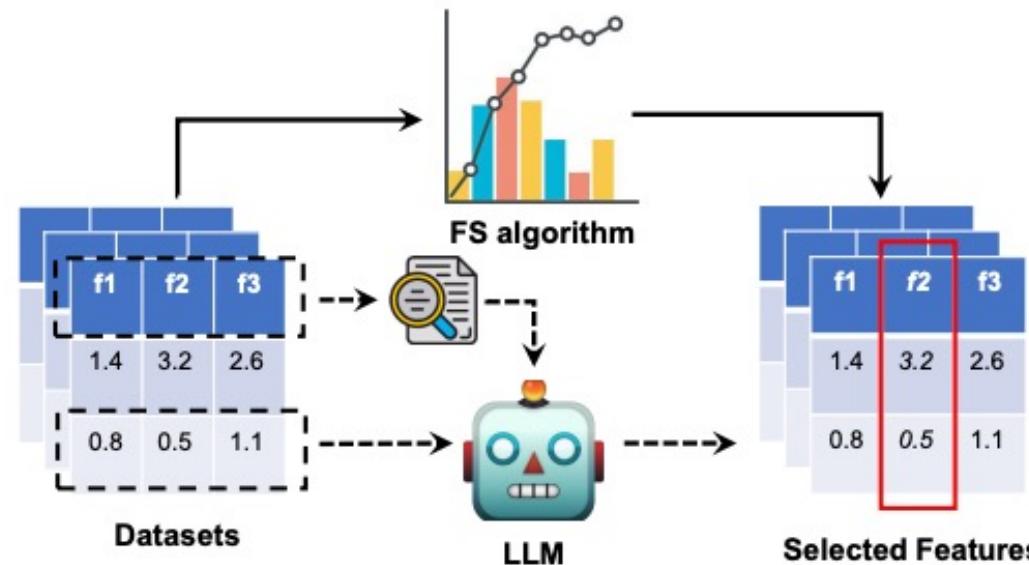
# 베이스라인 대비 성능 향상률(%)

모델	평균 Sharp Ratio	향상률 (%)
Linear Regression (Baseline)	0.0923	-
1차 분류모델: CatBoost	0.1390	50.5%
2차 회귀모델: XGBoost	0.1985	115.1%
3차 분류모델: XGB-Cat	0.1738	88.3%



# LLM feature selection

## Exploring Large Language Models for Feature Selection: A Data-centric Perspective



# LLM feature selection

```

# ----- 1) 시스템 프롬프트 -----
SYSTEM_PROMPT = """
/* Dataset Context [ Lending Club */
We work with Lending Club consumer-loan application data.
Goal: at origination time, predict the loan's internal rate of return (IRR) as accurately as possible.

/* Main Task */
1. Start from the provided feature set; freely create economically meaningful derived features (e.g. log_annual_inc,
2. Apply a Lasso / Elastic-Net style penalty to drop redundant or weak features; aim for ≤ 25 final features.
3. Output a ranked list with: feature, importance (0-1), direction, reason.

/* Output Format [ CSV ONLY (single string) */
feature,importance,direction,reason
dti_pct,0.97,positive,"Higher DTI reduces free cash flow, increasing expected losses and lowering IRR."
log_annual_inc,0.85,negative,"Income (log-scaled) proxies repayment capacity; higher income generally boosts IRR."
...

*Hard constraints*
- Return exactly one CSV string, header included—no markdown, no code fences.
- importance ∈ [0,1] (2 decimal places); reason in double quotes, single line.
"""

# ----- 2) 전체 feature 목록 -----
features = [
    "addr_state", "inq_last_6mths", "emp_length", "home_ownership", "purpose",
    "mths_since_last_delinq", "mths_since_last_major_derog", "mths_since_last_record",
    "mths_since_rcnt_il", "mths_since_recent_bc", "mths_since_recent_bc_dlq",
    "mths_since_recent_inq", "mths_since_recent_revol_delinq", "verification_status",
    "grade", "mo_sin_old_rev_tl_op", "open_acc", "num_tl_30dpd", "mo_sin_rcnt_tl",
    "acc_open_past_24mths", "mort_acct", "collections_12_mths_ex_med", "tot_hi_cred_lim",
    "total_rev_hi_lim", "num_rev_accts", "mo_sin_rcnt_rev_tl_op", "mo_sin_old_il_acct",
    "num_accts_ever_120_pd", "revol_bal", "num_actv_rev_tl", "term", "num_bc_tl",
    "acc_now_delinq", "dti", "fico", "num_actv_bc_tl", "pct_tl_nvr_dlq", "num_il_tl",
    "bc_open_to_buy", "pub_rec", "bc_util", "revol_util", "num_bc_sats", "delinq_2yrs",
    "total_bal_ex_mort", "total_acc", "num_tl_op_past_12m", "have_bc", "credit_hist_months",
    "have_rev", "tot_coll_amt", "annual_inc", "pub_rec_bankruptcies", "num_op_rev_tl",
    "num_tl_120dpd_2m", "tax_liens", "chargeoff_within_12_mths", "num_tl_90g_dpd_24m"
]

```

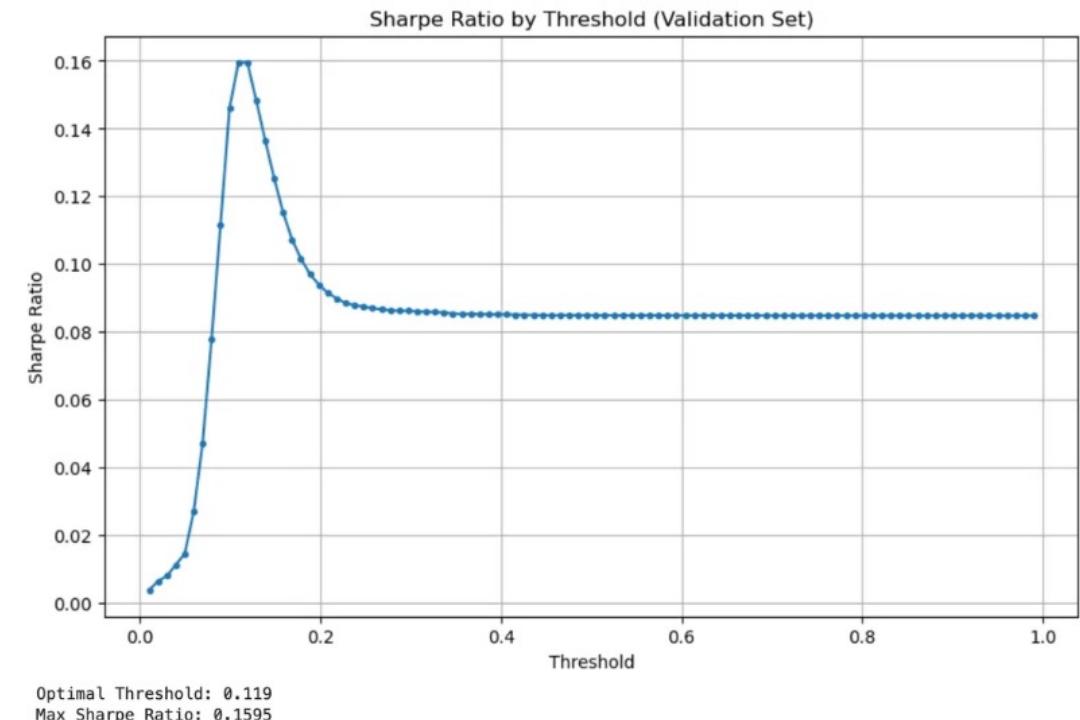
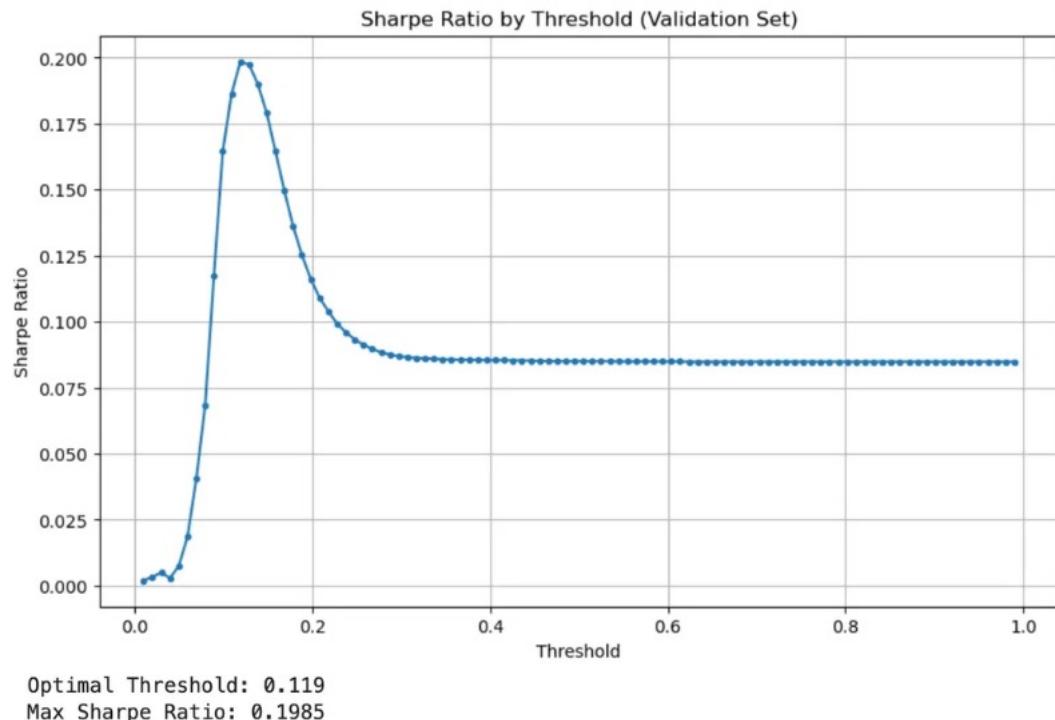
## Main Prompt



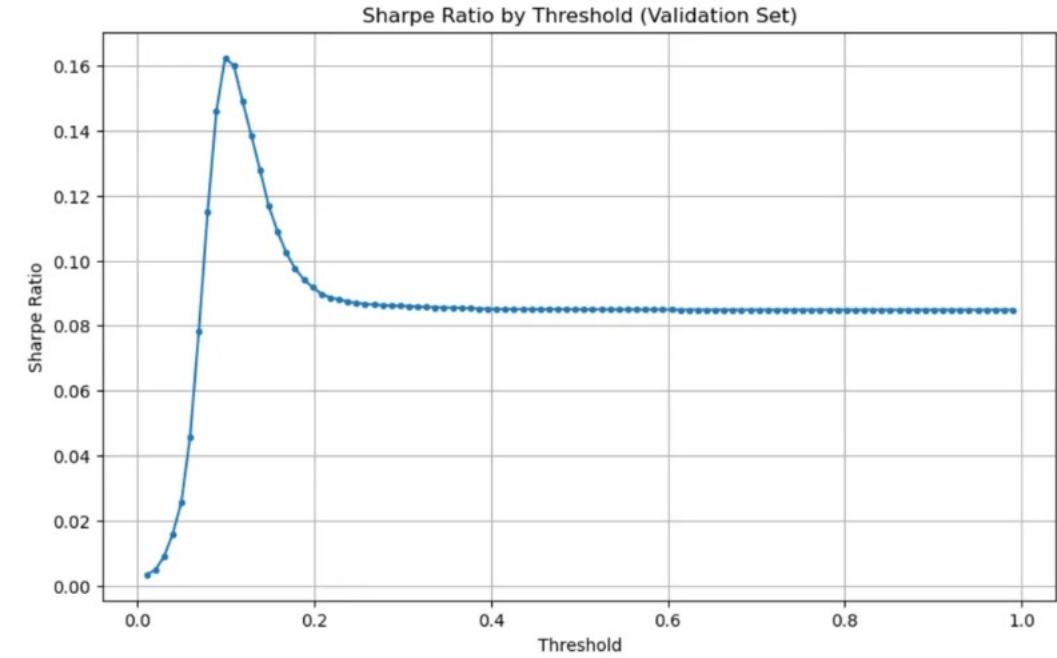
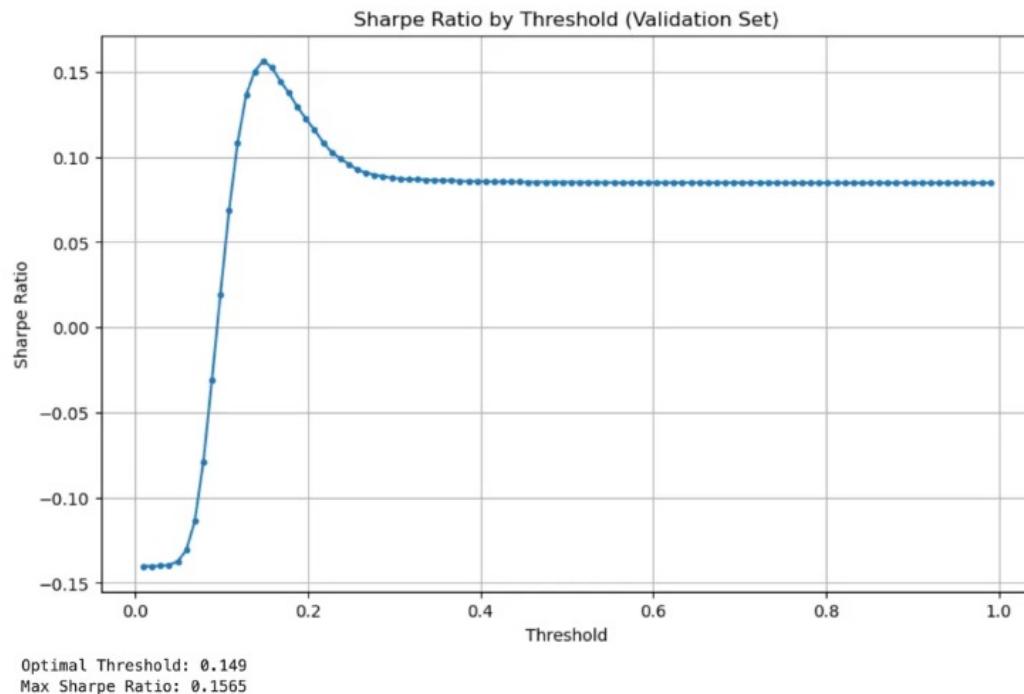
	feature	importance	direction	reason	wieght
0	dti	0.97	-1	Higher DTI reduces free cash flow, increasing ...	-0.97
1	log_annual_inc	0.85	1	Income (log-scaled) proxies repayment capacity...	0.85
2	fico	0.92	1	Higher FICO scores indicate better creditworth...	0.92
3	revol_util	0.88	-1	Higher revolving utilization suggests higher c...	-0.88
4	credit_hist_months	0.80	1	Longer credit history indicates more experienc...	0.80
5	bc_util	0.75	-1	High bankcard utilization can signal financial...	-0.75
6	home_ownership	0.70	1	Home ownership can indicate financial stabilit...	0.70
7	verification_status	0.65	1	Verified income increases confidence in repaym...	0.65
8	term	0.60	-1	Longer loan terms can increase risk of default...	-0.60
9	grade	0.78	1	Higher grades reflect lower risk profiles, imp...	0.78
10	num_actv_rev_tl	0.55	-1	More active revolving accounts can indicate hi...	-0.55
11	num_bc_tl	0.50	1	A higher number of bankcard tradelines can ind...	0.50
12	mo_sin_old_rev_tl_op	0.48	1	Older revolving accounts suggest stable credit...	0.48
13	open_acc	0.45	-1	A high number of open accounts can indicate ov...	-0.45
14	num_tl_90g_dpd_24m	0.40	-1	Recent delinquencies indicate higher risk, pot...	-0.40
15	mths_since_recent_bc	0.38	1	Recent bankcard activity can indicate active c...	0.38
16	num_il_tl	0.35	1	A higher number of installment loans can indic...	0.35
17	delinq_2yrs	0.33	-1	Recent delinquencies suggest higher risk, pote...	-0.33
18	mths_since_last_delinq	0.30	1	Longer time since last delinquency indicates i...	0.30
19	num_actv_bc_tl	0.28	1	Active bankcard tradelines can indicate credit...	0.28
20	total_rev_hi_lim	0.25	1	Higher credit limits suggest financial stabili...	0.25
21	acc_open_past_24mths	0.22	-1	Frequent account openings can indicate credit...	-0.22
22	mths_since_recent_inq	0.20	1	Longer time since last inquiry suggests reduce...	0.20
23	collections_12_mths_ex_med	0.18	-1	Recent collections indicate higher risk, poten...	-0.18
24	pub_rec	0.15	-1	Public records of derogatory events indicate h...	-0.15

## Selected Feature

# LLM feature selection (XGBOOST)



# LLM feature selection (CATBOOST)



# 04

---

## 결론 및 해석

- 주요 시사점
- 해석 가능성 확보
- 한계점 및 향후과제

주요 시사점

# 샤프지수 극대화하는 모델

1 stage

2 stage

LLM feature selection



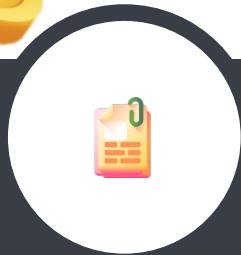
## 1. 샤프지수 극대화

부도예측의 목적이 아닌  
샤프 지수 최대화를 위한 모델 구축  
이를 위한 최적의 threshold 설정



## 2. 2 stage model

IRR 을 예측해  $IRR_{hat0}$ 라는 변수를 만들고,  
이를 독립변수로 사용하여 profitable 분류



## 3. LLM feature selection

LLM을 사용하여 변수를 선정하는 연구 진행  
유의미한 성능 향상은 없었지만, 변수를 1차  
적으로 선별하기엔 유의미하다는 것을 확인

# 한계점 및 향후과제

1

## 공동대출정보 활용 한계

Co-borrower 데이터가 존재했으나, 단독 대출과의 차이를 비교·분석하는데 충분히 반영하지 못함  
향후 공동 대출 여부에 따른 상환 패턴 및 부도율 차 이를 심층 분석하는 연구 필요



2

## 결측치 처리

가장 성능이 좋았던 XGBOOST의 Shapley value를 확인해본 결과 최근 대출 관련 정보의 부재가 가장 유의미(mths\_since\_rcnt\_il\_unknown) 결측치가 중요해보여 unknown으로 처리하였지만 이게 가장 중요한 변수가 되어 버림



3

## IRR 계산 제약

매월 어느 시점에 얼마가 상환되었는지에 대한 세부 데이터가 부재. 실제 현금흐름을 반영하지 못해, 이상적인 가정에 기반한 IRR 계산에 의존할 수밖에 없었음  
→ 향후 월별 실제 상환 일정 및 금액 데이터를 확보하여, 보다 현실적인 IRR 및 수익성 분석 가능



통계 데이터 사이언스

# 감사합니다 😊

발표일 2025. 08.08

7조 이상재 김태완 이솔 박예빈 박정현 이도아

통계 데이터 사이언스 팀프로젝트 발표

