Lending club 의
Sharpe Ratio 극대화 모델 수립 연구

5조
(김예빈, 류혜린, 박정원, 윤태영, 이창재, 정혜주)

목 차

Ⅰ. 서론 1

1.1 연구배경 1

1.2 연구주제 2

1.3 연구단계 수립 3

Ⅱ. 본론 5

2.1 탐색적 데이터 분석 (Exploratory Data Analysis, EDA) 5

2.1.1 대출 기본 조건 관련 변수 (loan_amnt, term, grade) 6

2.1.2 신청자 특성 관련 변수 (acc_open_past_24mths, home ownership) 8

2.1.3 변수 이해 및 전처리 진행 10

2.2 연구절차 12

2.2.1 데이터 전처리 12

2.2.2 IRR 계산 13

2.2.3 무위험 수익률 데이터 매칭 14

2.3 모델수립 14

2.3.1 데이터셋 분할 14

2.3.2 종속 변수 설정 15

2.3.3 초과 수익률(Excess Return) 회귀 모델의 필요성 16

2.3.4 부도 예측 모델과의 차별성 17

2.3.5 모델 설정 17

2.4 연구결과 18

2.4.1 모델결과 18

2.4.2 결과분석 23

III. 결론 25

3.1 연구 시사점 25

3.1.1 연구의의 25

3.1.2 연구한계 26

3.2 연구제언 26

참고문헌

I. 서론

1.1 연구배경

최근 디지털 전환과 함께 금융산업은 빠르게 변화하고 있으며, 그 중심에는 투자자와 차입자를 직접 연결하는 P2P(Peer-to-Peer) 금융 플랫폼이 있다. 미국의 Lending Club은 이러한 P2P 대출 시장을 대표하는 플랫폼으로, 2007년 설립 이후 온라인 기반 대출 중개 방식을 통해 대출의 접근성과 효율성을 크게 개선해왔다. 특히 전통 금융기관이 내부 심사 및 자금 운용을 중심으로 한 폐쇄적인 구조를 갖고 있던 반면, Lending Club은 대출 신청, 심사, 투자, 상환까지의 전 과정을 자동화하여 중개 비용을 절감하고, 보다 넓은 계층에게 자금 접근 기회를 제공하였다.

Lending Club은 신용점수 600점대 후반부터 신청이 가능하고, 1,000달러에서 40,000달러까지의 무담보 대출을 최대 5년까지 제공하며, 고정 이율과 고정 월 상환금 구조로 운영된다. 신청 절차는 전적으로 온라인으로 진행되며, 승인까지 걸리는 시간은 빠르면 24시간 이내로 처리될 수 있어 신속한 자금 확보가 가능하다 [1, 2]. 차입자는 신용점수, 소득, 고용상태, 부채비율(DTI) 등을 바탕으로 심사를 받으며, 승인 시 최대 약 6% 수준의 origination fee(기원 수수료)와 연체 시 부과되는 late fee가 적용된다[4]. 이와 같이 수수료 체계는 차입자에게 부담이 될 수 있으나, 절차가 빠르고 접근성이 높으며, 비교적 신용 점수가 낮은 이용자도 대출 신청이 가능하다는 점에서 포용적 금융 서비스로 기능하고 있다.

2020년을 기점으로 Lending Club은 기존의 순수 P2P 중개 모델을 종료하고, 미국 연방은행 인가를 받은 Radius Bank를 인수하며 디지털 뱅크(Neo-bank) 형태로 사업 모델을 전환하였다. 이에 따라 개인 투자자 대상의 직접 투자 기능은 중단되었으나, 플랫폼에서 축적된 대규모 대출 데이터를 바탕으로 수익성과 리스크를 분석하려는 연구는 지속적으로 이루어지고 있다. 최근에는 인공지능 기반 자동화 심사 시스템과 함께, 기관투자자의 비중이 높아지며 Lending Club은 안정적인 수익 창출과 리스크 관리를 병행하는 방향으로 전략을 조정하고 있다. 따라서 과거 P2P 성격의 데이터를 바탕으로 하되, 금융기관화된 구조 속에서 이를 재해석할 필요성이 높아진 시점이다.

Lending Club의 대출은 일반적으로 원리금 균등상환 방식으로 이루어지며, 월별 현금 흐름이 일정하게 구성되기 때문에 투자자는 IRR(Internal Rate of Return, 내부수익률)이나 Sharpe Ratio와 같은 수익률 기반 지표를 활용해 투자 성과를 분석할 수 있다. 또한 플랫폼은 각 대출 건에 대해 연체, 부도, 회수 여부를 포함한 세부 데이터를 제공하고 있어, 실질적인 리스크 분석과 수익

예측에 적합한 기반 자료로 기능하고 있다. 실제로 Lending Club은 투자자에게 다양한 등급의 대출 상품을 제공하고, 투자자는 이를 선별하여 분산 투자함으로써 자신의 리스크-수익 선호에 맞는 포트폴리오를 구성할 수 있다.

이러한 구조적 특성과 데이터의 투명성으로 인해, Lending Club 데이터를 활용한 학술 연구는 꾸준히 이루어지고 있다. 최근 연구는 단순히 부도 확률을 예측하는 데 그치지 않고, 수익률 자체를 최대화하는 전략 설계에 집중되는 경향을 보인다. 예를 들어, Li와 Xu는 Sharpe Ratio를 목적함수로 설정해 투자 성과를 개선하는 머신러닝 기반 모델의 우수성을 입증하였으며, Misheva 등은 SHAP과 같은 해석 기법을 활용해 변수 중요도를 설명 가능한 형태로 제시하였다 [3, 4]. 이처럼 수익률 기반 분석과 모델 해석력 강화가 동시에 요구되는 최근의 연구 흐름은, 본 연구의 방향성과도 밀접하게 연결된다.

1.2 연구주제

Lending Club의 2022년까지의 자료를 이용해 부도예측모형을 구축하는 것이 주제이다. 이 주제에는 다음과 같은 요구사항이 있다.

첫째, 자료의 excess_return을 적절히 가공하여 부도 여부에 해당하는 종속변수 혹은 라벨로 코딩한 뒤 이를 예측한다.

둘째, 목적함수는 위험대비 초과수익률인 Sharpe Ratio이고 이를 극대화하는 방향으로 모형을 구축한다.

셋째, 부도가 예측되어 대출을 승인하지 않았더라면, 해당 투자금은 투자 결정 당시의 3년/5년 만기 미국 국채에 투자하였다고 가정한다.

넷째, 월별 cash flow를 계산하고, 이 cash flow를 현재 투자 시점으로 할인하여 내부수익률을 구하는 것을 기본 골자로 한다.

본 연구는 위 요구사항에서 부도예측 정확도가 목적이 아니라 Sharpe Ratio 극대화를 연구의 핵심 목표로 삼았다.

1.3 연구단계 수립

본 연구는 Lending Club의 대출 데이터를 바탕으로 IRR 극대화를 통해 궁극적으로 포트폴리오 수익률인 Sharpe Ratio 극대화를 추구한다. 이를 위해서 각 대출 건에 대한 수익률에 대한 평가 지표로 IRR을 계산하였고, 대출 포트폴리오 수익성을 평가하기 위해서 수익성과 리스크를 모두 고려한 Sharpe Ratio를 사용하였다.

1.3.1 IRR(Internal Rate of Return, 내부수익률)

Lending Club의 대출은 원리금 균등상환 방식으로 설계되어 있다. 이는 대출자가 대출 기간 동안 매월 동일한 금액(원금 + 이자)을 상환하는 구조이다. 이러한 특성으로 인해 각 대출에 대한 월별 현금흐름을 기반으로 한 IRR 계산이 가능하며 표준화된 대출별 수익률 비교가 가능하다.

$$0 = \sum_{t=1}^{T} \frac{C_t}{(1+r)^t} - C_0$$

- C_0: 대출 원금, C_t: t기 시점의 현금 흐름(cash flow)
- r: IRR (내부수익률), T: 총 상환 기간

1.3.2 Sharpe Ratio

포트폴리오 수익률을 평가하기 위해서 Sharpe Ratio를 적용하였다. 수익률을 평가하는 지표 중 Sharpe Ratio는 위험을 고려한 수익률 계산이다. σ_r 은 수익률의 표준편차로 초과 수익률($\mu_r - r_f$)의 리스크를 조정해서 포트폴리오의 성과를 나타낸다.

$$Sharpe\ Ratio = \frac{\mu_r - r_f}{\sigma_r}$$

- μ_r : 포트폴리오 수익률(IRR)의 평균
- r_f : 무위험 수익률 (3년/5년 만기 미국 국채 수익률)
- σ_r : 포트폴리오 수익률의 표준편차

결국 본 연구는 IRR 예측을 통해서 개별 대출의 수익률을 예측하고, 이를 통해 대출 승인 결정을 하였다. 예측한 수익률이 미국 국채보다 작을 시 대출 승인을 거부하고 대출 신청 금액을 안전 자산인 미국 국채에 투자한다. 만약에 예측한 수익률이 미국 국채보다 클 경우 대출을 승인하여 수익을 내는 전략을 취하였다. 이후 Sharpe Ratio를 통해 대출 포트폴리오의 종합

성과를 산출하였다. 최종적으로 Sharpe Ratio를 극대화하기 위해 IRR 예측 모델 조정 및 임계값 조정하는 방식으로 진행하였다.

1.3.3 연구 방향성

본 연구는 IRR을 예측하는 회귀 모델을 구축하여, 각 대출 건에 대한 예상 IRR을 산출하였다. IRR 예측 정확도를 통해서 종합적인 IRR 극대화를 통해 최종적으로는 가장 높은 Sharpe Ratio를 찾는 것을 목표로 하였다.

IRR 예측을 위해서 대출 신용 평가 이전의 사전 변수들만 사용하여서 예측하였다. 실제 상환 기록을 바탕으로 개별 대출의 IRR 값을 계산한 후, (IRR-무위험 수익률 = 초과수익률)을 종속변수(target)로 설정하여 회귀 모델을 학습하였다. 이후 테스트 데이터에 대해 예측된 초과수익률 값을 0 기준으로 비교하여, 예측 IRR이 무위험 수익률을 초과하는 대출을 투자 대상으로 선별하였다.

이와 같은 전략은 수익률 그 자체를 중심으로 투자 여부를 판단하기 때문에, 분류 기반 접근보다 더 정밀하고 연속적인 투자 판단이 가능하다는 장점을 갖는다. 최종적으로는, 이와 같이 선별된 포트폴리오의 예측 수익률 분포를 바탕으로 Sharpe Ratio를 계산하고 이를 성과 평가 지표로 활용하였다.

IRR을 예측하는 회귀모델을 구축하기 전 모델로는 부도 분류 모델과 부도 예측 확률 모델을 시도하였다. 부도 분류 모델은 개별 대출을 부도 또는 대출 상환으로 분류하는 모델이었다. 이때 부도로 예측된 대출은 국채에 투자하고, 상환으로 예측된 모델은 대출을 승인해서 개별 대출들의 IRR을 계산하고, 계산된 IRR로 Sharpe Ratio를 산출하는 방식으로 진행하였다. 부도 예측 확률을 기반으로 특정 임계값을 설정하고, 해당 임계값 이하의 대출을 선별하여 투자 포트폴리오를 구성하는 방식이었다.

부도 확률 예측 모델이 부도 분류 모델보다 부도 대출을 맞추는 성능은 높았지만 두 모델 모두 한계가 존재하였다. 부도 여부를 종속변수로 사용하는 위 모델들은 투자 대상 선정 기준이 이진적인 분류 결과에 종속된다는 점, 그리고 수익률 자체에 대한 정보를 직접 반영하지 못한다는 점에서 제약이 있었다. 따라서, 본 연구는 개별 대출 수익성을 고려하는 IRR 예측 모델을 활용해서 Sharpe Ratio 최적화 접근 방법을 선택하였다.

II. 본론

2.1 탐색적 데이터 분석 (Exploratory Data Analysis, EDA)

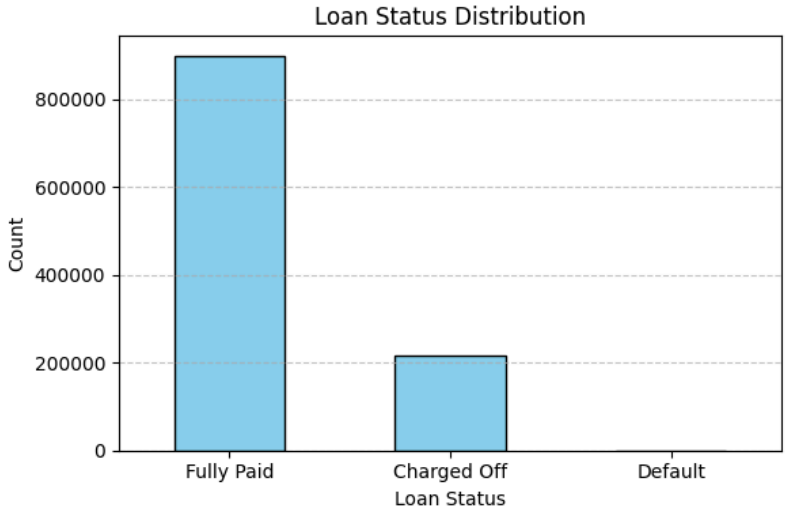
본 연구에서는 미국의 대표적인 P2P 대출 플랫폼인 Lending Club에서 제공하는 2020년 대출 데이터를 활용하였다. 해당 데이터는 총 1,116,156건의 대출 기록과 141개의 변수로 구성되어 있으며, 각 대출에 대한 신청 정보, 신용등급, 상환 상태 등이 포함되어 있다. Sharpe Ratio 극대화를 목표로 하는 본 연구의 목적에 따라, 분석에는 채무 이행 여부가 명확히 결정된 대출 건들만을 사용하였다.

Lending Club에서는 채무 상태를 연체정도와 대출상환완료 여부에 따라 구분하고 있다. "loan_status" 변수에 총 10개의 범주로 채무 상태를 분류하고 있다. 채무 상태를 나타내는 변수 중 채무 이행 여부가 결정된 Fully Paid, Default, Charged Off인 되는 대출들만 분석에 사용하였다. Fully Paid는 상환 완료한 대출, Default는 채무 불이행 상태 대출, Charged Off는 대출 상환이 불가하다고 평가한 대출이다. 채무 이행 또는 불이행의 명확하지 않은 데이터(Current, Late (16-30 days), Late (31-120 days), In Grace Period, Issued) 그리고 내부 정책에 미충족되었지만 대출이 되었던 데이터(Does not meet the credit policy. Status:Fully Paid, Does not meet the credit policy. Status:Charged Off)를 제외하였다.

〈표 1〉 Loan Status and Number of Loans

Loan Status	Number of Loans
Fully Paid	898,522
Charged Off	217,366
Default	268
Current	618,688
Late (31-120 days)	9,840
Late (16-30 days)	1,620
In Grace Period	6,049
Issued	1,258
Does not meet the credit policy. Status:Fully Paid	1,223
Does not meet the credit policy. Status:Charged Off	460

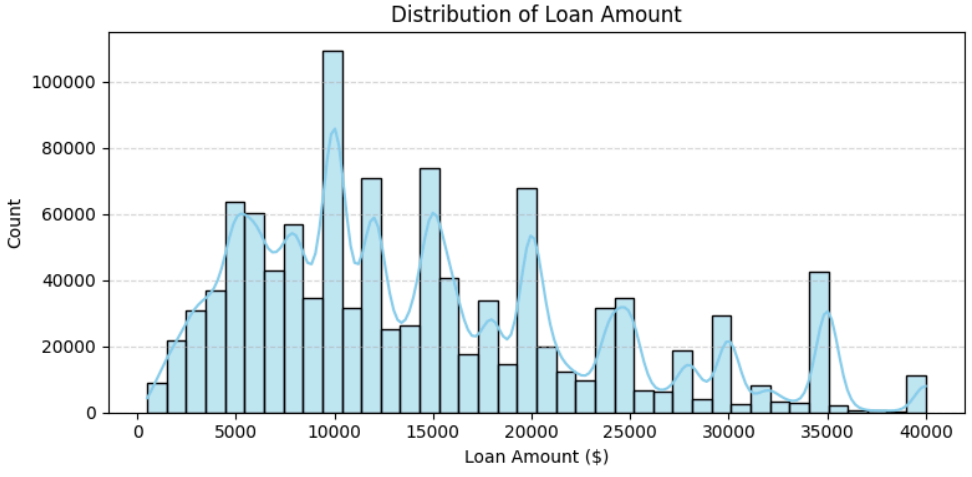
〈그림 1〉은 대출 상환 여부가 결정된 Fully Paid, Charged Off, Default 대출 상태 분포를 나타낸다. 이를 통해, Lending Club 데이터셋이 정상 상환(Fully Paid) 대출 비중이 압도적으로 높고, 대손 처리(Charged Off), 부도(Default) 건수가 상대적으로 매우 적은 심각한 클래스 불균형 데이터셋이라는 것을 알 수 있다.



〈그림 1〉 Loan Status Distribution

2.1.1 대출 기본 조건 관련 변수 (loan_amnt, term, grade)

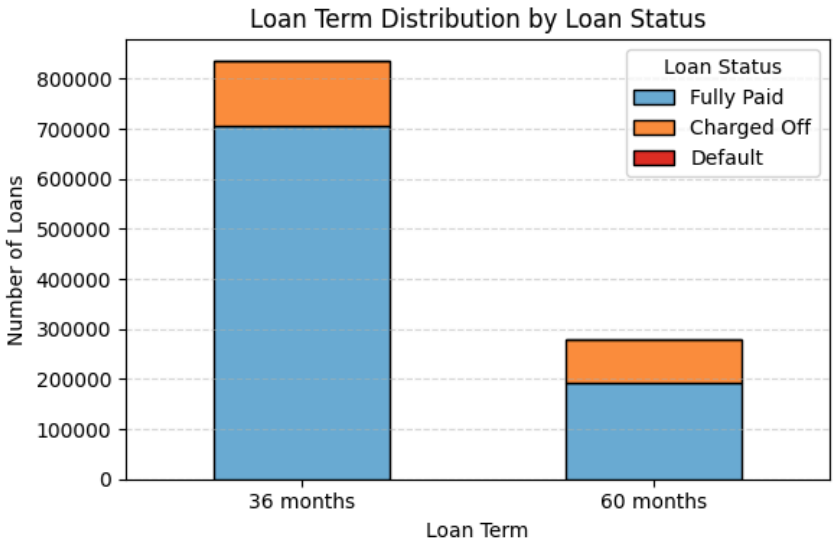
대출 금액(loan_amnt)은 최소 USD $500에서 최대 USD $40,000까지 분포하며, 평균은 약 USD $14,578이다. 전체 대출 중 약 25%는 USD $7,925 이하, 75%는 USD $20,000 이하로, 대출 금액은 USD $10,000 ~ $20,000 구간에 가장 집중되어 있다.



〈그림 2〉 Loan Amount Distribution

전체 대출 중 약 75%는 36개월 대출, 25%는 60개월 대출로 구성되어 있다. 상환 상태를 비교한 결과, 36개월 대출의 상환 완료 비율은 84.3%, 반면 60개월 대출의 상환 완료 비율은 69.0%로

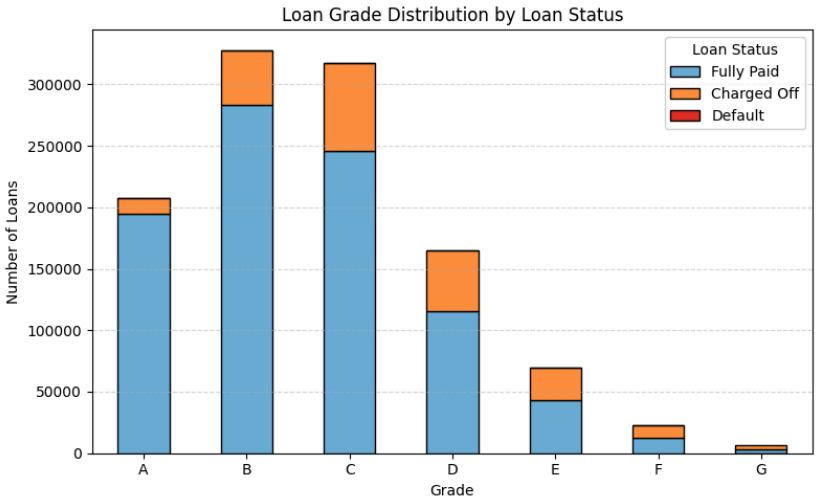
나타났다. 60개월 대출은 36개월 대출에 비해 상환 실패(Default 또는 Charged Off) 비율이 약 두 배 이상 높게 나타났다.



〈그림 3〉 Loan Term Distribution by Loan Status

Lending Club은 대출 신청자에게 신용등급(grade)을 부여하며, 이는 A부터 G까지 총 7단계로 구성된다. B 등급과 C 등급 대출이 전체 대출 승인 건수 중 가장 큰 비중을 차지하였으며, A 등급은 그 뒤를 이었다. 상환 결과를 살펴보면, A 등급 대출의 약 93%가 정상적으로 상환(Fully Paid) 되었으며, B 등급은 약 87%, C 등급은 약 78% 수준으로 상환 완료 비율이 점차 감소하였다.이와 같은 감소 추세는 하위 등급에서도 동일하게 나타났으며, D 등급은 약 70%, E 등급은 63%, F 등급은 56%, G 등급은 50% 수준의 상환 완료율을 보였다. 이는 신용등급이 낮아질수록 상환 실패 위험이 명확히 증가하는 경향을 확인할 수 있다. B , C 등급에서는 A 등급에 비해 채무 불이행 대출 승인 비율이 급격히 증가하는 것이 관찰된다.

이러한 상환 성과는 실제 대출별 IRR에도 반영되었다. A 등급 IRR은 약 -0.003으로 손익 분기점인 0에 가장 가깝고, 등급이 낮아질수록 평균 IRR은 점차 하락하여 G 등급에서는 -0.057을 기록하였다. 또한 표준편차 역시 A 등급의 0.056에서 G등급의 0.149로 점차 커지며 변동성이 증가가 관찰되었다. 이는 신용등급이 하락할수록 단순히 상환성공률이 낮아질 뿐 아니라, 수익률 측면에서도 평균적인 손실 위험이 증가를 의미한다. 등급 변수는 대출 승인 과정에서 Lending Club 측에서 채무 이행률을 고려해서 만든 변수이기 때문에 사후변수로 취급하여 모델 학습과정에서는 제외하였다.



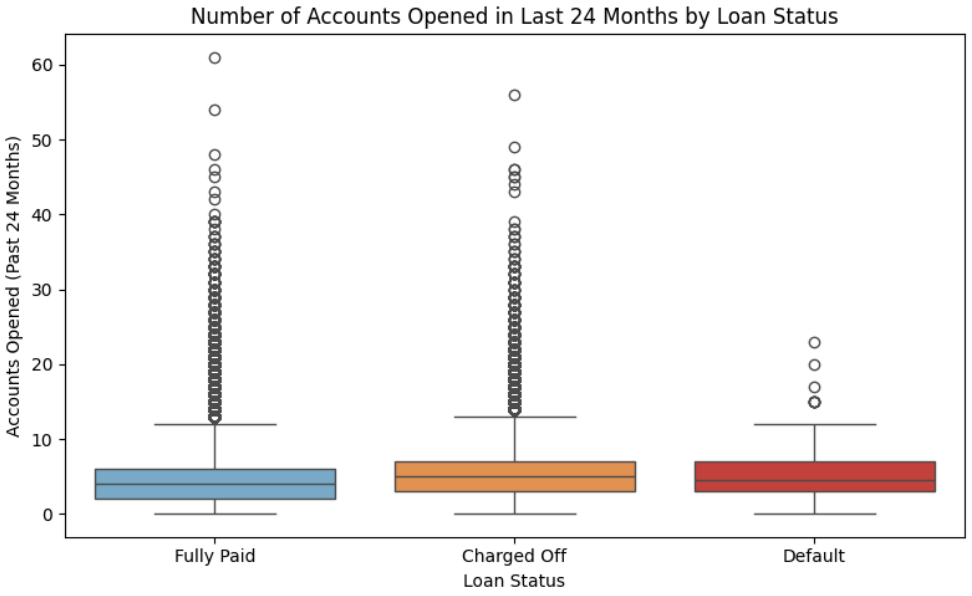
〈그림 4〉 Loan Grade Distribution by Loan Status

〈표 2〉 등급별 내부수익률(IRR) 평균과 표준편차

Grade	Mean	Std
A	−0.003	0.056
B	−0.007	0.072
C	−0.016	0.089
D	−0.027	0.113
E	−0.032	0.116
F	−0.041	0.128
G	−0.057	0.149

2.1.2 신청자 특성 관련 변수 (acc_open_past_24mths, home ownership)

acc_open_past_24mths 변수는 최근 2년간 개설된 금융 계좌 수를 나타낸다. 정상 상환(Fully Paid: 4.50건) 이 상환 실패 그룹(Charged Off: 5.26건, Default: 5.02건) 비해 평균적으로 더 적게 계좌를 개설하였다. 최근 2년간 계좌를 더 많이 개설한 대출자일수록 상환실패를 할 가능성이 아주 약하게 있을 수 있음을 나타낸다.



〈그림 5〉 Loan Status 에 따른 acc_open_past_24mths

Home Ownership(주거 형태) 변수는 대출자의 신용 특성을 반영하는 주요 변수 중 하나이다. MORTGAGE(주택담보 보유자)는 Fully Paid 대출의 절반 이상(50.99%)을 차지하며, 상환 실패 그룹인 Charged Off(42.59%)와 Default(45.15%)보다 높은 비율을 보인다. 이는 주택을 담보로 보유한 대출자의 상환 가능성이 상대적 높다는 점을 시사한다. RENT(임차인)는 Charged Off에서 가장 높은 비율(45.98%)을 기록했으며, Default에서도 41.79%를 차지하여 임차인의 경우 상대적으로 부도 위험이 큰 경향을 보인다.

〈표 3〉 Home Ownership 데이터 현황

Home Ownership	Fully Paid	Charged Off	Default
MORTGAGE	50.99	42.59	45.15
RENT	38.00	45.98	41.79
OWN	10.94	11.35	12.69
기타 (ANY/NONE/OTHER)	0.07	0.08	0.37

2.1.3 변수 이해 및 전처리 진행

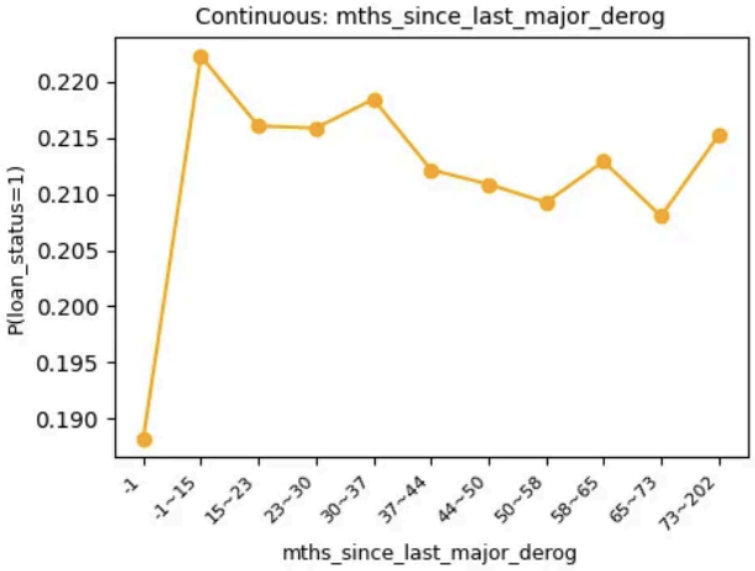
이 외에도 데이터셋에는 채무수준 및 상환능력과 관련된 변수 (dti,dti_joint, loan_income_ratio, percent_bc_gt_75), 신용 점수 및 조회 관련 변수 (fico_mean, weighted_inq, mths_since_recent_inq) 등 다양한 변수가 존재하였다. 변수에 대한 충분한 이해를 거친 후, 불필요한 데이터를 처리하고 각 데이터의 특성에 맞는 전처리 방식을 적용하였다.

먼저, 모델 학습에 앞서 불필요하거나 예측 성능에 기여하지 않을 것으로 판단되는 변수를 제거하였다. 신용평가 시점 이후에만 관측 가능한 사후변수를 제외하였으며, 결측치 비율이 90% 이상으로 과도하게 높은 변수 또한 배제하였다. 이후 변수들에 대해 개별적으로 시각화를 수행하여, 부도율과 뚜렷한 연관성이 나타나지 않는 변수들을 추가로 삭제하였다 〈표 4〉.

예를 들어 〈그림 #〉은 최근 중대한 연체 발생 이후 경과 개월 수(mths_since_last_major_derog)와 부도율의 관계를 나타낸다. 그림에서 확인할 수 있듯이, 경과 개월 수와 부도율 사이에 뚜렷한 패턴이나 추세가 관찰되지 않았다. 따라서 해당 변수는 부도 여부를 설명하거나 예측하는 데 기여도가 낮다고 판단하여 최종 변수에서 제외하였다.

〈표 4〉 신용평가 시점 이후 변수

신용평가 시점 이후 변수
• acc_now_delinq , tot_coll_amt ,*_cur_bal , * chargedoff_within_12_mths , num_tl _*_dpd *, num_tl_op_past_12m , revol_bal_joint , hardship_* , dept_settlement_flag



〈그림 6〉 mths_since_last_major_derog과 loan_status와의 관계

정규화가 필요한 경우, min-max 스케일링을 적용하고 추가적으로 각 칼럼의 왜도를 계산한 후 왜도가 큰 칼럼에 대해서는 로그 변환을 진행하였다. 또한 정확한 예측을 위해 기존 변수에 기반한 파생변수를 생성하였고 〈표 7〉, 전처리 과정을 거쳐 총 63개의 변수로 데이터셋을 완성하였다 〈표 5, 6〉.

〈표 5〉 데이터 전처리 진행방식

<table>
<tr><th>변수명</th><th>결측치 처리</th><th>전처리 방법</th><th>결측치 대체</th><th>스케일링</th></tr>
<tr><td rowspan="3">수치형 변수</td><td rowspan="5">결측치 행 제거</td><td>그대로 유지</td><td rowspan="5">중앙값, 평균값,
파생변수로 대체</td><td rowspan="3">Log transformation,
Min-Max Scaling</td></tr>
<tr><td>파생변수 생성</td></tr>
<tr><td>이진 변수화</td></tr>
<tr><td rowspan="2">범주형 변수</td><td>원핫 인코딩</td><td rowspan="2">X</td></tr>
<tr><td>순서형 인코딩</td></tr>
</table>

〈표 6〉 데이터 전처리 진행방식 (종합)

	진행방식
1차	● loan_status(Fully paid, Charged off, Default)를 제외한 140개 변수들 필터링 ● 크게 4가지 방법(사후변수, 부도율과 연관성 낮은 변수, 유사변수 통합, 무관한 변수)으로 진행 ● 140 → 61개 변수로 정리 후, 각 데이터 특성에 맞게 전처리 진행
2차	● 1차적으로 불필요한 열을 제외하고, 2차적으로 필요한 열들의 행 결측치 및 전처리 진행 ● 61 → 63개 변수로 정리

〈표 7〉 파생변수 의미

변수명	의미	설명
credit_utilization_ratio	신용카드 한도 대비 사용률	total_bal_ex_mort / (total_bc_limit + 1)
loan_per_month	월별 상환금액 수준	loan_amnt / (term_months +1)
loan_income_ratio	소득 대비 대출비율	loan_amnt / (annual_inc +1)
fico_mean	평균 신용점수	(fico_range_low + fico_range_high) / 2
loan_to_credit_limit	대출금 대비 신용한도	loan_amnt / (tot_hi_cred_lim + 1)
mortgage_ratio	모기지 계좌 비중	mort_acc / (total_acc + 1)
high_utilization_score	고위험 카드 사용 지표	(percent_bc_gt_75 * bc_util) / 100

2.2 연구절차

본 연구의 목적은 Sharpe Ratio를 극대화하는 것이다. 따라서 Lending Club의 대출 원데이터를 기반으로 투자성과 지표인 IRR을 개별 대출 단위로 산출하였다. IRR은 각 대출의 상환 패턴에 따라 달라지므로, 데이터 전처리 단계에서 대출 상태와 상환 기간을 체계적으로 정리한 후, 현금흐름(cash flow)을 정의하여 계산하였다. 또한, Sharpe Ratio 산출을 위해 무위험 수익률(risk-free rate)을 추가로 매칭하였다.

2.2.1 데이터 전처리

데이터셋에 변수인 대출 개시일(issue_d)과 마지막 상환일(last_pymnt_d)은 문자열(예: *Jan-2020*)로 제공되었다. 이를 datetime 객체로 변환한 뒤, 두 날짜 차이를 개월 수 단위로 환산하여 n_months 변수를 생성하였다. 이 변수는 IRR 산출 시 실제 상환 기간을 반영하기 위해 반드시 필요한 요소이다.

- 예를 들어, 2020년 1월에 발급된 대출이 2022년 6월에 마지막 상환을 기록했다면, 총 상환 개월 수(n_months)는 29개월로 계산된다.
- 단, 계산 결과가 0개월이 되는 경우를 방지하기 위해 최소값을 1로 설정하였다.

대출 상태(loan_status)는 투자 회수 여부를 기준으로 이진화하였다. 이 과정을 통해 이후 IRR 계산에서 조건문을 활용할 수 있도록 데이터가 정리하였다.

- 정상 상환(Fully Paid)은 0으로 코딩하여 부도 없음(non-default)을 의미한다.
- 부도(Default, Charged Off)는 1로 코딩하여 손실 발생(default)을 의미한다.

2.2.2 IRR 계산

본 연구에서는 대출 상태에 따라 다른 상환 패턴을 고려하여 IRR을 계산하였다. 이를 위해 각 대출별 현금흐름(cash flow)을 구성한 뒤, numpy-financial 패키지의 irr() 함수를 활용하였다.

1) 정상 상환(loan_status = 0)

(1) 조기 상환(elapsed < term)

- 대출 계약 기간보다 짧은 시점에서 상환이 종료된 경우, 총 상환금액(total_pymnt)과 마지막 상환액(last_pymnt_amnt)을 활용하여 실제 월별 상환액을 계산하였다.
- 상환 기간이 단 1개월인 경우에는 분모가 0이 되는 문제를 방지하기 위해, 최초 투자금(-loan_amnt)과 전액 상환금(total_pymnt)만을 현금흐름에 포함하였다.

(2) 정상 만기 상환(elapsed = term)

- 계약된 상환 개월 수(term) 동안 매월 동일한 금액(installment)이 지급되는 것으로 가정하였다.

2) 부도 발생(loan_status =1)

- 최초 투자금(-loan_amnt)이 지출된 이후, 실제 상환 개월 수(n_months) 동안 일부 상환금(installment)이 유입된다.
- 마지막 달에는 실제 납부된 금액(last_pymnt_amnt)과 회수금(recoveries)이 추가로 반영되어 최종 현금흐름을 구성하였다.

이와 같이 정의된 cashflow를 기반으로 IRR을 산출하였다. 최종적으로 모든 대출 건에 대해 IRR을 계산하여 irr 변수에 저장하였다.

2.2.3 무위험 수익률 데이터 매칭

본 연구에서는 미국 국채 수익률(U.S. Treasury Yield) 데이터를 사용하여 Sharpe Ratio 산출에 활용하였다. 미국 국채는 세계에서 가장 신뢰할 수 있는 무위험 자산으로 간주되며, Lending Club 대출이 미국 금융시장에서 발행되어 기준 통화와 시장이 일치하기 때문이다. [5]

무위험 수익률 데이터는 FRED(Federal Reserve Economic Data)에서 제공하는 3년·5년 만기 국채 금리를 불러와 다음과 같이 활용하였다.

- 대출의 계약기간(term)이 36개월인 경우, 3년물 국채 금리를 매칭하였다.
- 대출기간이 60개월인 경우, 5년물 국채 금리를 매칭하였다.

이 과정에서 대출 발급일(issue_d) 기준으로 해당 시점에 가장 근접한 국채 금리를 매칭하여, 시계열적 정합성을 확보하였다.

위와 같이 전처리를 진행한 이후 초과 수익률(IRR − 무위험 수익률)의 평균을 표준편차로 나눈 값을 Sharpe Ratio로 정의하여 모델을 수립하였다.

2.3 모델수립

2.3.1 데이터셋 분할

데이터셋은 학습용(train), 검증용(validation), 시험용(test)으로 분리하여 활용하였다 〈표 8〉. 데이터 분할 과정에서 시간적 순서를 고려하여, 모델 학습에 미래 정보가 포함되지 않도록 하였다. 또한 IRR 계산이 불가능하거나 결측치가 존재하는 샘플(1개)은 제거하였다.

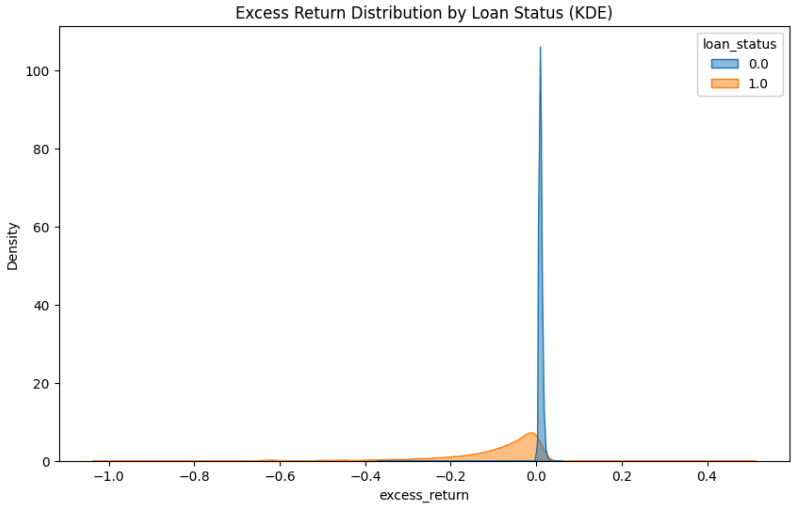
〈표 8〉 데이터셋 분할

구분	비율	용도
Train Set	40%	모델 학습
Validation Set	20%	하이퍼파라미터 탐색, 검증
Test Set	40%	최종 성능평가

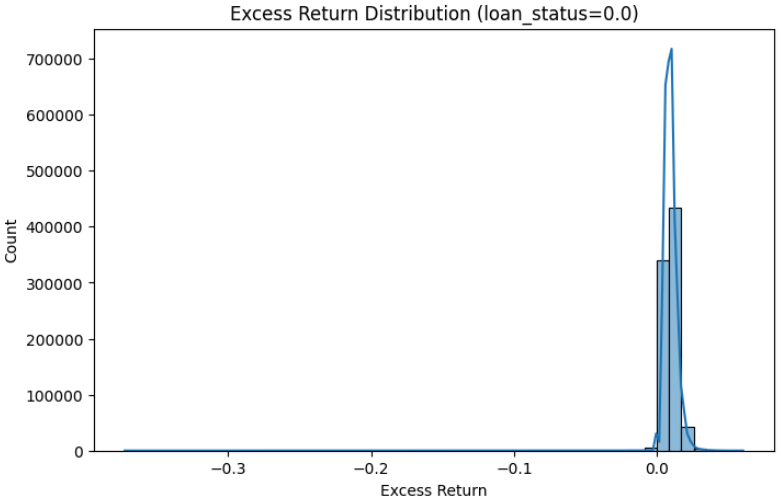
2.3.2 종속 변수 설정

기존의 대출 부도 예측 연구에서는 대출 상환 여부(부도 = 1, 정상상환 = 0)를 종속 변수로 설정하여 분류 문제로 접근하는 경우가 많다. 그러나 이러한 방식은 단순히 부도 여부를 예측하는 것에 그치며, 투자 성과를 직접적으로 극대화하는 목적과는 괴리가 존재한다. 예를 들어, 동일하게 부도(=1)로 분류된 사례라도 실제 회수 금액이나 조기상환 여부에 따라 투자자의 수익률은 크게 달라질 수 있다.

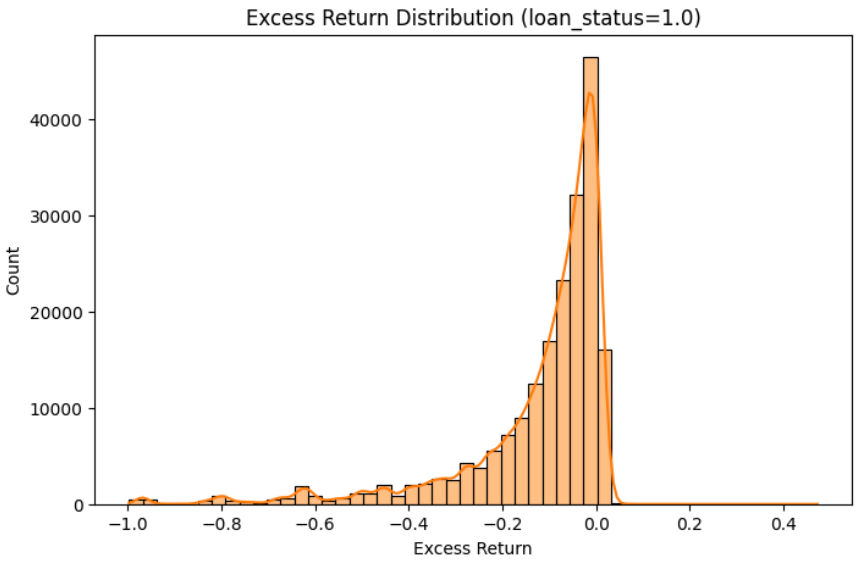
본 연구 또한 초기에는 대출 원금 회수율을 기반으로 y = 1−funded_amnt / total_rec_prncp 형태로 타깃을 설정하여 분류 모델을 구축하였다. 그러나 이는 직관성이 부족하며, 투자자의 최종 목표인 수익률 극대화와 일치하는 지표로서 IRR을 직접 타깃으로 설정하는 것이 보다 타당하다.



〈그림 7〉 부도 여부별 초과수익률의 분포



〈그림 8〉 정상 대출자의 초과 수익률 분포, 그림 1 일부분



〈그림 9〉 부도 대출자의 초과 수익률 분포, 그림1 일부분

위 그림에서 확인할 수 있듯이, 부도 여부와 수익 발생 여부는 유사하지만 완전히 동일하지 않다. 정상 대출자의 경우 대부분 초과수익률이 양수로 나타났으나, 일부 부도 대출자 역시 양수의 초과수익률을 기록하였다. 이는 부도가 발생했더라도 원금의 상당 부분을 상환한 사례에 해당한다. 따라서 모든 부도 대출자를 동일하게 부정적으로 간주하는 것은 적절하지 않으며, 본 연구에서는 이러한 차이를 반영하기 위해 타깃값을 차등화하여 학습을 진행하였다. 다만, 전반적으로 부도 대출자의 초과수익률은 음수, 정상 대출자는 양수로 나타났기 때문에, 이는 기존의 부도 여부 예측과 완전히 다른 접근이라기보다는 개선된 방식으로 이해할 수 있다.

2.3.3 초과 수익률(Excess Return) 회귀 모델의 필요성

단순 IRR만을 예측할 경우, 거시적 금리 환경의 변동을 충분히 반영하지 못한다는 한계가 있다. 동일한 IRR이라도 투자 시점의 무위험 수익률(risk-free rate)에 따라 초과 성과 여부가 달라지기 때문이다. 예를 들어 연 8%의 IRR은 무위험 금리가 1%일 때는 매력적이지만, 무위험 금리가 7%일 때는 초과 성과가 사실상 1%에 불과하다. 따라서 본 연구에서는 IRR 대신 초과 수익률(Excess Return = IRR − risk-free)을 종속 변수로 설정하였다. 이렇게 함으로써, 모델은 투자 성과를 절대적 수익률이 아니라 시장 대비 초과 성과(알파) 관점에서 평가·예측할 수 있으며, 이는 실제 투자 의사결정에 더 직결되는 정보라 할 수 있다.

2.3.4 부도 예측 모델과의 차별성

본 연구의 접근은 전통적인 부도 예측 모델의 궤를 같이하면서도, 몇 가지 측면에서 더 나은 장점을 제공한다.

첫째, 부도 여부는 단순한 이진 결과지만, 초과 수익률은 연속형 지표로서 투자 성과를 보다 정밀하게 반영한다. 이는 같은 "정상 상환"이라도 조기상환, 부분상환 등 다양한 시나리오를 수치적으로 구분할 수 있게 한다.

둘째, 초과 수익률 기반 회귀는 부도율 예측이 갖는 한계(0/1 분류의 불확실성)를 넘어, 실제 투자자의 효용과 직결된 "얼마나 버는가"라는 질문에 답할 수 있다.

셋째, 부도 예측 모델은 부정적 사건을 피하는 데 초점을 두는 반면, 초과 수익률 예측 모델은 긍정적 사건(수익 극대화)까지 동시에 고려할 수 있어, 투자 전략 수립에 있어 보다 포괄적인 의사결정 지원이 가능하다.

2.3.5 모델 설정

본 연구에서는 회귀 예측 모델로 LightGBM (Light Gradient Boosting Machine) Regressor를 사용하였다. LightGBM은 Microsoft Research에서 제안한 그래디언트 부스팅(Gradient Boosting) 계열 알고리즘으로, 의사결정나무(Decision Tree)를 약한 학습기(weak learner)로 사용하여 성능을 점진적으로 향상시키는 부스팅(Boosting) 방법론을 따른다. 특히 LightGBM은 다음과 같은 장점을 갖는다.

- 효율성: Histogram 기반 분할 알고리즘과 Gradient-based One-Side Sampling(GOSS), Exclusive Feature Bundling(EFB) 기법을 적용하여, 기존의 XGBoost 대비 계산 속도와 메모리 사용량이 크게 개선되었다. 대규모 데이터셋에서도 빠른 학습이 가능하다.

- 예측 성능: 전통적인 부도 예측이나 신용위험 모델에서 중요한 복잡한 비선형 관계와 변수 간 상호작용을 효과적으로 포착할 수 있다. 이는 대출 특성과 투자 성과 사이의 복잡한 관계를 모델링하는 데 유리하다.

- 정규화 기능: max depth, min child samples, learning rate 등 다양한 하이퍼파라미터를 통해 과적합을 방지하고 모델의 일반화 성능을 확보할 수 있다.

하이퍼파라미터 최적화는 Bayesian Optimization을 통해 수행하였다. Bayesian Optimization은 임의 탐색이나 그리드 탐색에 비해 적은 탐색 횟수로도 효율적으로 최적값에 수렴할 수 있는 장점이 있다. 각 후보 파라미터에 대해 5-Fold 교차검증(K-Fold Cross Validation)을 실시하여 검증 데이터셋에서의 평균 성능을 평가하였고, 평균 절대 오차(Mean Absolute Error, MAE)를 주요 평가 지표로 활용하였다. MAE는 실제값과 예측값 간 절대 오차의 평균으로, 모델이 초과 수익률을 얼마나 근접하게 예측했는지 직관적으로 보여준다. 단, 하이퍼파라미터 최적화 이후 모델의 성능을 평가할때는 Sharpe ratio를 사용하여 모델을 평가하였다.

또한 실제 투자 환경을 모사하기 위하여, 예측된 초과 수익률이 양수(>0)인 대출만 선별하여 투자하고, 음수인 대출의 투자금은 투자 결정 당시의 3년/5년만기 미 국채에 투자하여 성과를 분석하였다. 이는 음수 초과 수익률이 예상되는 대출은 투자자가 실제로 선택하지 않는다는 합리적 가정을 반영한 것이다. 따라서 본 연구의 모델은 단순한 수익률 예측을 넘어, 실제 투자 의사결정과 직결되는 투자 대상 선별 과정까지 포함한다는 점에서 의의가 있다.

2.4 연구결과

2.4.1 모델결과

본 연구에서는 랜덤 포레스트, LightGBM, CatBoost, Lasso, Ridge 등 다양한 단일 모델을 적용하고 그 결과를 비교하였다. 각 모델의 성능은 투자자 관점에서의 Sharpe Ratio를 통해 평가하였다. 보고서에서는 먼저 각 모델별 주요 성능 결과를 요약하고, 이후 최종적으로 선택한 모델의 세부 결과와 해석을 제시한다. 마지막으로, 모델 성능 차이에 대한 원인을 분석하고, 결과가 가지는 의미를 투자적 관점에서 논의한다.

본 연구에서 최종 모델의 선택 기준은 Sharpe Ratio로 설정하였다. 이는 투자자의 관점에서 위험 대비 초과 수익을 평가할 수 있는 핵심 지표이기 때문이다. 그러나 Sharpe Ratio만으로는 평가가 제한적일 수 있어, 그 계산이 적절한지를 판단하기 위해 보조 평가지표를 함께 활용하였다. 참고로, test set 기준 모든 건을 대출해준 상황의 sharpe ratio는 -0.176이기때문에 해당 값을 벤치마크로 사용하였다.

우선, Hit Ratio는 초과 수익률의 부호 적중 여부를 평가하는 지표로 사용하였다. 이는 대출이 올바른 대상자에게 이뤄졌는지를 판단하는 데 유용하다. 즉, 부도자가 초과 수익률을 양수로 예측하거나 반대의 경우를 방지하는 기능을 한다.

또한, R^2는 회귀 모델의 일반적인 성능을 나타내는 지표로, 대출 여부와 무관하게 초과 수익률을 얼마나 잘 예측하는지를 보여준다. 다만, 본 연구에서는 이를 참고용 보조 지표로만 활용하였다.

마지막으로, 대출률을 함께 고려하였다. 이는 test set 내 대출 신청자 중 실제로 대출이 진행된 비율을 의미한다. 대출률을 극도로 낮추면 Sharpe Ratio를 인위적으로 극대화할 수 있으나, 이는 실질적으로 의미가 없는 결과를 낳는다. 따라서 본 연구에서는 모델이 일정 수준 이상의 대출률을 유지하면서도 Sharpe Ratio를 극대화하는 방향을 중시하였다.

〈표 9〉 모델 별 성능

	LGBM	XGB	CAT BOOST	Random Forest	Lasso	Ridge	MLP
Sharpe Ratio	**1.390**	1.305	1.355	1.296	1.289	1.219	1.172
Hit Ratio	**0.9667**	0.9240	0.9232	0.9123	0.9053	0.8915	0.8512
R^2	**0.3829**	0.3902	0.3201	0.3516	0.3444	0.3433	0.3211
대출률	**79.66%**	80.36%	80.53%	80.21%	80.51%	71.85%	76.7%

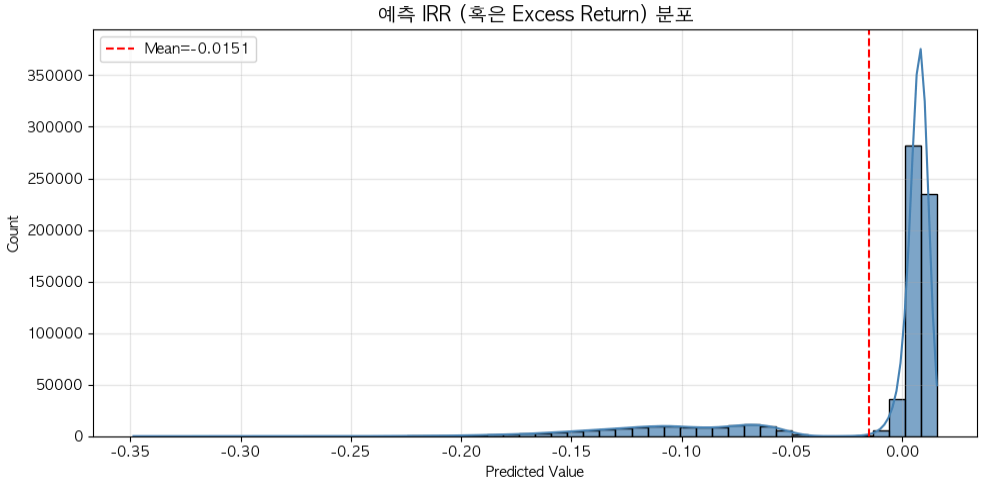
〈표 10〉 최종모델(LGBM)의 하이퍼 파라미터 탐색 결과

파라미터	값	설명
boosting_type	gdbt	부스팅 방식 (Gradient Boosting)
num_leaves	988	트리의 최대 leaf 수 (복잡도 제어)
max_depth	17	트리 최대 깊이 (과적합 방지)
learning_rate	0.0109	학습률 (step size)
n_estimators	291	트리 개수 (학습 반복 횟수)
min_child_samples	49	리프 노드가 가질 최소 샘플 수
subsample	0.6	데이터 샘플링 비율 (bagging)
subsample	0.806	트리 생성 시 컬럼 샘플링 비율
reg_alpha	0.146	L1 정규화 계수
reg_lambda	0.192	L2 정규화 계수

모델들의 R^2 값은 모두 0.4를 넘기지 못할 정도로 낮은 수준에 머물렀다. 이는 개별 고객의 초과 수익률을 정밀하게 예측하는 능력이 제한적이었음을 의미한다. 그럼에도 불구하고 본 연구에서 Sharpe Ratio가 높게 산출된 이유는 Hit Ratio의 우수한 성과 때문이다. 실제로 Hit Ratio는 약 93% 수준으로, 모델이 개별 수익률의 정확한 크기를 맞추지는 못했더라도, 수익이 발생할 고객과 손실이 발생할 고객을 구분하는 데 매우 높은 정확도를 보였다.

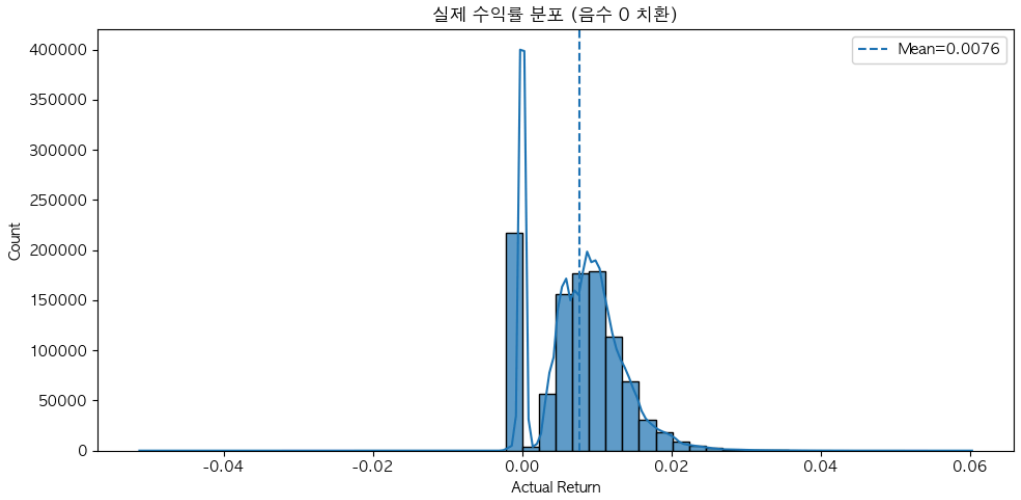
따라서 높은 Sharpe Ratio는 모델이 초과 수익률 자체의 정밀도보다는 투자 의사결정의 방향성(수익/손실 구분)을 효과적으로 제시했기 때문으로 해석할 수 있다. 이는 금융 투자 환경에서 모델의 활용 가능성을 높여주는 중요한 결과라 할 수 있다.

〈그림 10〉 Excess Return의 분포

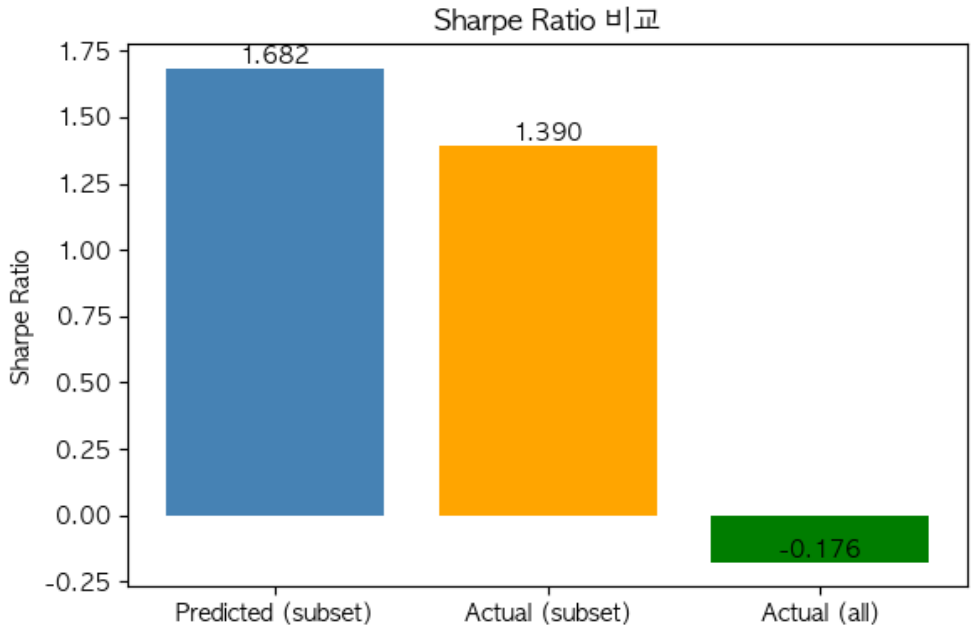


〈그림 11〉 최종 모델로 예측한 초과 수익률의 분포

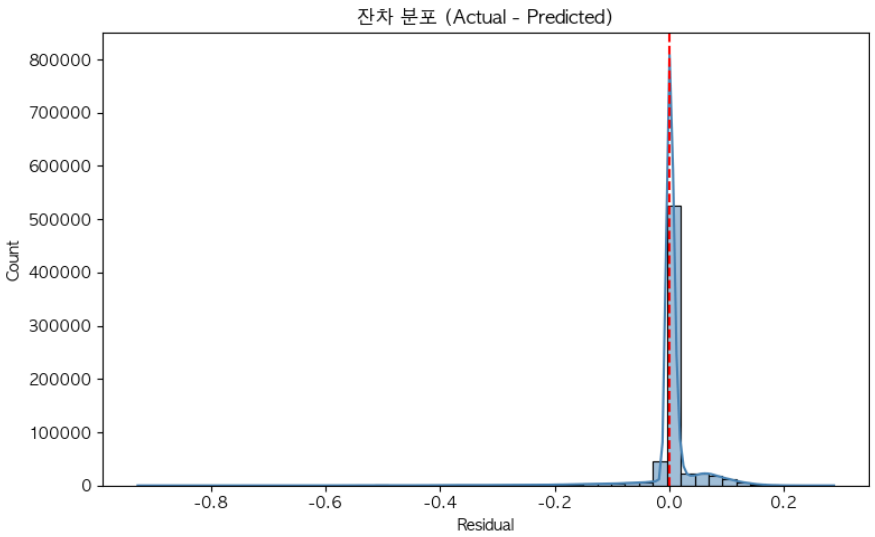
최종 모델의 예측 초과 수익률 분포를 살펴보면, 전체적으로 0 부근에 밀집되어 있으며 좌측으로 긴 꼬리를 가지는 왼쪽으로 치우친 분포 형태를 보인다. 평균 예측값은 약 -0.0151로 음수에 위치하고 있다. 실제 초과수익률의 분포보다 더욱 좌측으로 치우쳐진 보수적인 모델이라는 것을 알 수있다. 오대출시의 손실이 더욱 크기에 적합한 모델의 형태로 보인다.



〈그림 12〉 모델의 예측값 중 0미만인 값 0으로 대체한 이후 분포



〈그림 13〉 예측값의 srape ratio(좌), 실제 값의 sharpe ratio(중, 실제 성능),
모든건 대출시 sharpe ratio(우, 벤치마크)



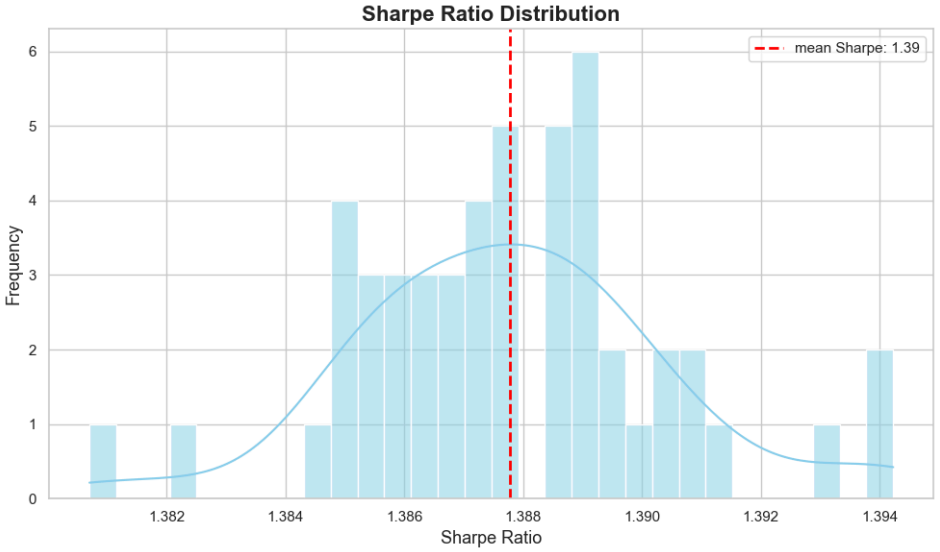
〈그림 14〉 예측값과 실제값 간 잔차의 분포

예측 초과 수익률이 음수인 경우(즉, 무위험 수익률보다 낮은 경우)에는 해당 값을 0으로 대체하여 Sharpe Ratio를 계산하였다. 이는 곧 무위험 수익률보다 낮은 투자 성과는 무위험 자산에 투자한 것과 동일하게 처리한다는 의미를 갖는다.

또한 본 연구에서는 타깃값을 초과 수익률로 설정하였기 때문에 threshold는 자동으로 0으로 결정되었다. 이론적으로 threshold 값을 더 높게 설정할 경우 Sharpe Ratio는 상승하지만, 동시에 대출률이 급격히 하락하여 소수의 인원에게만 대출이 이뤄지는 결과를 초래한다. 이는 실제 금융 의사결정의 현실성을 훼손할 수 있기 때문에, 본 연구에서는 투자 성과와 대출률 간 균형을 고려하여 threshold = 0을 최종적으로 채택하였다.

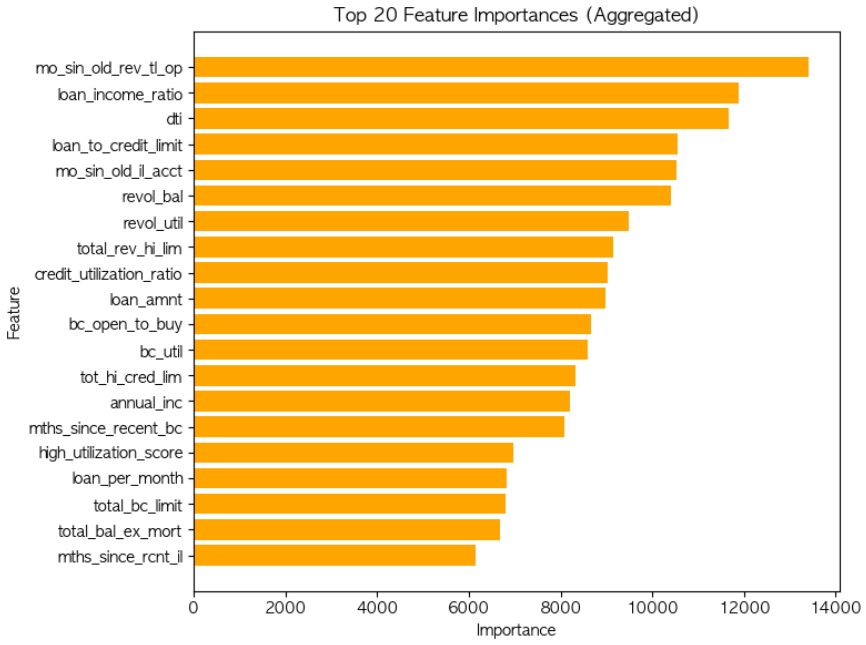
이 방식으로 산출된 결과, 최종 모델의 Sharpe Ratio는 1.390으로 나타났다. 동시에, 전체 신청자 중 약 79.66%가 대출 승인을 받는 수준의 대출률을 기록하였다. 이는 곧 모델이 전체 신청자의 약 80%에게 대출을 허용하면서도, 결과적으로 무위험 수익률 대비 1.39배의 초과 이익을 달성했음을 의미한다.

따라서 본 모델은 대출 승인률과 위험 대비 수익성을 동시에 고려했을 때, 투자자에게 현실적이고 합리적인 의사결정 기준을 제시할 수 있음을 보여준다. 최종 모델의 안정성을 검증하기 위해 50회의 무작위 test set 추출을 수행하고 각 경우에 대해 Sharpe Ratio를 계산하였다. 그 결과, Sharpe Ratio의 평균값은 **1.39**로 이전에 보고된 수치와 유사하게 나타났으며, 전반적으로 성능의 분포가 큰 변동 없이 안정적으로 유지됨을 확인할 수 있었다. 이는 본 모델이 특정 데이터 분할에 의존하지 않고, 다양한 상황에서도 일관된 투자 성과를 제공할 수 있음을 보여준다.



〈그림 15〉 무작위 Test set 반복에 따른 sharpe ratio 분포

2.4.2 결과분석



〈그림 16〉 최종모델의 변수중요도

최종 모델의 변수 중요도를 살펴본 결과, 대출자의 신용 이력과 소득 대비 부채 수준이 가장 핵심적인 요인으로 나타났다. 특히, mo_sin_old_rev_tl_op(가장 오래된 리볼빙 계좌 개설 이후 경과 개월 수)와 loan_income_ratio(대출금 대비 소득 비율), dti(부채 대비 소득 비율)가 상위 3개

변수로 나타나, 차주의 장기적인 신용 거래 이력과 상환 능력이 대출 성과 예측에 중요한 역할을 하고 있음을 알 수 있다.

또한, loan_to_credit_limit(대출금 대비 신용한도 비율)와 mo_sin_old_il_acct(가장 오래된 분할상환 계좌 이후 경과 개월 수) 역시 높은 중요도를 보였다. 이는 차주의 신용카드 및 할부 거래 이력의 안정성이 부도 가능성 판단에 기여하고 있음을 의미한다. 이와 함께 신용카드 사용 패턴 관련 지표들이 다수 포함되어, 단기적인 신용 사용 행태 또한 예측에 큰 영향을 미치는 것으로 나타났다.

그 외에도, loan_amnt(대출 금액), annual_inc(연소득), bc_open_to_buy(신용카드 가용 한도)와 같은 기본적인 대출 규모 및 소득 관련 지표가 중요한 변수로 확인되었다. 이는 대출 규모와 소득 수준의 불균형이 발생할 경우 부실 위험이 커진다는 기존 금융 이론과도 일치한다.

본 연구에서는 원천 데이터에서 직접 제공되지 않는 정보를 반영하기 위해 다양한 파생변수를 생성하였다. 그 중에서도 credit_utilization_ratio(총 신용한도 대비 사용률), loan_per_month(월별 상환금액), loan_to_credit_limit(대출금 대비 신용한도 비율), high_utilization_score(고위험 사용 패턴 점수) 등이 높은 중요도를 기록하였다. 이는 모델 학습 과정에서 파생변수가 핵심적인 설명력을 제공했음을 보여준다.

특히 이러한 파생변수들은 기존 원천 변수만으로는 포착하기 어려운 차주의 신용 행태와 상환 능력 간의 복합적 관계를 반영한다. 예를 들어, 단순한 대출 금액이나 소득 수준만으로는 설명하기 어려운 소득 대비 신용카드 사용 수준이나 대출금 규모 대비 신용 여력과 같은 정보를 모델이 학습할 수 있도록 하였다. 그 결과, 파생변수들은 모델이 잠재적 부실 위험을 보다 정교하게 구분하도록 도와 Sharpe Ratio 개선에 기여하였다.

III. 결론

3.1 연구 시사점

3.1.1 연구의의

본 연구는 기존의 P2P 대출 데이터를 활용한 연구들이 대부분 부도율 예측에 초점을 맞추고, AUC, F1-score 등 분류 모델의 성능 지표를 중심으로 평가해온 것과 달리, Sharpe Ratio라는 수익성 중심의 지표를 직접 최적화 대상으로 설정했다는 점에서 중요한 차별성을 지닌다. 이는 단순히 “누가 부도가 날지를 예측”하는 접근을 넘어, “어떤 대출에 투자했을 때 수익성과 리스크 간 균형이 가장 우수한가”라는 실질적인 투자 의사결정의 본질에 보다 가까운 문제를 다루고자 한 시도로 평가될 수 있다.

Sharpe Ratio를 목적함수로 설정한 분석은 금융 투자 전략 최적화라는 실무적 맥락에서 매우 유효하며, 머신러닝 기법을 통해 실제 수익률을 향상시킬 수 있는 전략을 구성했다는 점에서 전통적인 신용평가 모델의 틀을 넘어서는 응용 가능성을 제시하였다. 특히 이와 같은 실험적 접근은, 향후 머신러닝 기반 금융 의사결정 모델링에서 단순 예측 정확도를 넘어 실질 성과 중심으로 패러다임을 전환하는 흐름에 기여할 수 있다는 점에서 학술적 의의도 크다.

또한 본 연구는 단순히 가상의 조건이나 모형 기반 시뮬레이션이 아닌, 미국 내 대표 P2P 금융 플랫폼인 LendingClub에서 실제로 발생한 대규모 대출 데이터를 기반으로 분석을 수행하였다는 점에서도 높은 실증적 가치를 지닌다. 이러한 현실 데이터는 대출 조건, 상환 구조, 연체 여부, 회수 금액 등 다양한 요소가 실제 상황에서 어떻게 작동하는지를 반영하고 있으며, 본 연구는 이를 바탕으로 현실 적용 가능성이 높은 투자 전략을 도출하였다. 특히 IRR, 수익률, 회수율 등 정량적 지표를 통해 투자성과를 평가하고, 모델을 통해 직접 투자 대상을 선별하는 구조는 실제 투자운용 시스템이나 자동화 포트폴리오 구성 알고리즘으로 확장 가능성도 내포한다.

마지막으로, 본 연구는 설명 가능한 머신러닝 기법과 결합하여 모델 해석 가능성까지 확보할 수 있는 방향성을 열어두었다. 이는 향후 금융 규제 환경이나 신뢰 기반 모델링이 요구되는 상황에서도 활용될 수 있는 기반을 마련했다는 점에서, 이론적·실무적·기술적 측면 모두에서 의의가 있다고 할 수 있다.

3.1.2 연구한계

본 연구에서 다음과 같은 다섯가지 한계가 있었다.

첫째, Sharpe ratio라는 목적함수 자체의 한계가 있었다. Sharpe ratio는 정규분포를 전제로 하지만, 실제 금융 데이터는 fat-tail 등의 비정규 특성을 가져서 채무불이행등의 극단적 손실 리스크를 제대로 반영하지 못한다.

둘째, 하방위험이 무시된다는 점이 있다. 변동성을 상·하방을 모두 동일하게 취급하여 실제 투자자는 손실 변동성만 신경 쓰는데, Sharpe ratio는 이익의 변동성도 리스크로 포함해버려 왜곡이 생긴다.

셋째, Sharpe ratio는 단순히 "평균 수익 / 표준편차"라서 현금 유동성 위험을 반영하지 못한다. 즉, 같은 Sharpe ratio라도 유동성이 나쁜 포트폴리오일 수 있다.

넷째, 기간 선택의 민감성이 있다. 대출의 IRR은 기간에 종속적인데, Sharpe ratio 계산 시 월간과 연동 기준 변동성 추정이 크게 달라질 수 있고, 조기상환과 연체 등에 민감해서 목적함수로는 불안정하다.

마지막으로 변수 설정의 한계가 있었다. 주어진 변수만으로는 극단적인 부도 상황을 예측하기 어려워, 그것을 보완하는 파생변수 설정에 한계가 있었다.

3.2 연구 제언

본 연구는 Lending Club 데이터를 활용하여 대출 부도 여부를 예측하고, 각 대출의 IRR을 산출한 뒤 Sharpe Ratio를 극대화하는 투자 전략을 제시하였다. 분석 과정에서 상환 기간(n_months)을 반영하여 개별 대출의 실제 상환 패턴을 정밀하게 모델링하고, 미국 국채 수익률을 무위험 수익률로 매칭함으로써 보다 현실성 있는 성과 측정을 가능하게 하였다. 이러한 접근은 P2P 금융 투자자의 위험 조정 수익률을 정량적으로 평가할 수 있다는 점에서 의미가 크다.

그러나 본 연구에는 몇 가지 한계가 존재하며, 이를 보완하기 위한 추가 연구가 필요하다. 첫째, 데이터 측면에서 Lending Club 단일 플랫폼에 국한되었기 때문에 결과의 일반화에는 제약이 따른다. 향후 연구에서는 Prosper, Upstart 등 다른 P2P 금융 플랫폼 데이터를 비교 분석하거나, 실업률·금리 수준 등 거시경제 변수를 결합하여 외부 환경이 부도율과 수익률에 미치는 영향을 검증할 필요가 있다. 또한 대출 목적이나 설명 문구와 같은 비정형 텍스트 데이터를 자연어처리 기법을 통해 정량화하면 예측 성능을 한층 높일 수 있을 것이다.

둘째, 방법론적 확장이 요구된다. 본 연구는 Sharpe Ratio 최적화와 전통적 머신러닝 분류 기법을 중심으로 분석을 진행했으나, 향후에는 생존분석(Survival Analysis)을 적용하여 부도 발생 시점까지의 시간을 모형화하거나, 베이지안 접근을 통해 불확실성을 고려한 추정치를 제시하는 것도 유용할 것이다. 또한 최신 딥러닝 기반 시계열 모델이나 Explainable AI(XAI) 기법을 활용하여 예측 성능과 해석 가능성을 동시에 강화할 수 있다.

셋째, 투자 전략 차원에서 현실적인 제약을 반영할 필요가 있다. 본 연구는 고정된 포트폴리오 Sharpe Ratio를 기준으로 최적화했으나, 실제 투자 환경에서는 대출 상태가 지속적으로 업데이트되므로 동태적 리밸런싱 전략을 고려해야 한다. 더불어 투자자의 위험 성향에 따라 Sortino ratio, Omega ratio 등 대체 성과 지표를 활용하는 것도 합리적이다. 나아가 플랫폼 수수료, 세금, 거래 비용 등을 반영하여 순수익률(Net Return)을 기반으로 한 성과 측정이 이루어진다면 투자 의사결정의 실효성이 더욱 높아질 것이다.

마지막으로, 본 연구의 방법론은 P2P 금융을 넘어 다른 대체 금융 자산으로 확장될 수 있다. 예를 들어 부동산 크라우드펀딩, 마이크로파이낸스, 나아가 탈중앙화 금융(DeFi) 대출에도 동일한 분석 틀을 적용할 수 있다. 또한 차입자와 투자자 간 연결망을 네트워크 분석으로 확장하면, 대출 위험이 시스템적으로 확산되는 과정을 파악할 수 있는 새로운 연구 주제가 될 것이다.

종합적으로, 본 연구는 Sharpe Ratio 극대화라는 투자 성과 최적화 관점에서 P2P 대출 데이터를 분석하였다는 점에서 기초적 성과를 제시하였다. 향후 연구에서는 데이터의 외연을 확장하고, 방법론을 고도화하며, 실제 투자 제약을 반영함으로써 보다 현실적이고 일반화 가능한 신용평가 및 투자 전략을 제시할 수 있을 것으로 기대된다.

참고문헌

[1] Debt.org. (2024). *LendingClub review: Pros and cons of personal loans*. https://www.debt.org/credit/loans/personal/lending-club-review/

[2] FinanceBuzz. (2024). *LendingClub personal loan review: What to know*. https://financebuzz.com/lendingclub-personal-loan-review

[3] Li, Y., & Xu, Q. (2022). Predicting loan default and profitability: Random forest or logistic regression? *Financial Innovation*, 8(15). https://doi.org/10.1186/s40854-022-00338-5

[4] Misheva, B., Geng, X., & Treleaven, P. (2021). Explainable credit risk prediction using machine learning: Evidence from peer-to-peer lending. *arXiv preprint arXiv:2103.00949*. https://arxiv.org/abs/2103.00949

[5] Van Binsbergen, J. H., Diamond, W. F., & Grotteria, M. (2022). Risk-free interest rates. *Journal of Financial Economics*, *143*(1), 1-29.