

Lending Club 대출 의사결정의 샤프 비율(Sharpe ratio) 기준 최적화: 내부 수익률(IRR)을 중심으로 4가지 모형의 비교

7조(김태완, 박예빈, 박정현, 이도아, 이상재, 이슬)

| | |
|--------------------------------------|----|
| 1. 서론 | 1 |
| 1.1 Lending Club 소개 | 1 |
| 1.2 연구 목적 | 1 |
| 1.3 기대효과 | 2 |
| 2. 본론 | 2 |
| 2.1 연구 방법 | 2 |
| 2.1.1 연구 모형 소개 | 2 |
| 2.1.2 핵심 개념 소개 | 6 |
| 2.2 연구 절차 | 10 |
| 2.2.1 EDA 및 전처리 | 10 |
| 2.2.2 1단계 분류모형 (부도 예측) | 13 |
| 2.2.3 2차 회귀모형(IRR 예측) | 15 |
| 2.2.4 3차 2단계 모델(결합 분류모형) | 17 |
| 2.2.5 LLM Feature Selection 모델 | 18 |
| 2.3 연구 결과 | 18 |
| 2.3.1 성능 비교 (샤프 비율 극대화 효과) | 18 |
| 2.3.2 최종모델 선정 | 21 |
| 2.3.3 모델 안정성 점검 및 추가 분석 | 22 |
| 3. 결론 | 23 |
| 3.1 결과 논의 및 시사점 | 23 |
| 3.2 한계점과 향후 과제 | 25 |
| 4. 참고문헌 | 26 |

1. 서론

1.1 Lending Club 소개

Lending Club은 미국의 대표적인 P2P대출 플랫폼으로, 개인 투자자와 차용자를 직접 연결해주는 중개형 온라인 금융 서비스이다. 전통적인 은행과 달리 Lending Club은 중개수수료로 수익을 창출한다. 이를 통해 차용자는 은행보다 낮은 금리로 자금을 조달하고 투자자는 여러 대출에 소액씩 분산 투자하여 이자 수익을 얻을 수 있다. 투자자는 하나의 대출 상품에 전액을 투자하는 것이 아니라 최소 25달러인 소액 단위로 분산 투자할 수 있으며, 각 대출은 36개월 또는 60개월 동안 매월 고정된 원리금 상환을 통해 정기적인 현금흐름을 생성한다. 이 구조는 투자자에게 채권 투자와 유사한 이자 수익 구조를 제공하여 차용자가 월 단위로 상환하는 원리금을 통해 투자자는 수익을 회수하게 된다.

Lending Club 플랫폼의 데이터는 개별 대출에 대한 다양한 신용 정보와 대출 이후의 상환 결과, 그리고 투자자가 회수한 현금흐름 정보까지 담겨 있어 신용위험 예측과 투자 성과 분석에 모두 유용한 특성을 지닌다.

1.2 연구 목적

본 연구의 목적은 P2P 대출 플랫폼 투자자 관점에서 위험 대비 수익률을 극대화하는 투자 전략을 개발하는 것이다. 기존의 대출 심사 모형은 주로 부도 발생 여부를 예측하여 원금 손실을 최소화하는 데 초점을 맞춘다. 그러나 투자자의 궁극적인 목표는 높은 수익을 올리면서도 그 변동성을 줄이는 것, 즉 위험 대비 높은 초과수익을 달성하는 것이다. 이를 계량적으로 평가하는 대표 지표가 샤프 비율(Sharpe Ratio)이다. 샤프 비율은 투자 포트폴리오의 초과수익률(투자수익률 - 무위험수익률)의 평균을 초과수익률의 표준편차로 나눈 값으로 정의되며, 단위 위험당 추가로 얻는 수익을 나타낸다.

그런데 왜 Lending Club의 단독 이익이 아닌, Lending Club과 투자자 모두의 이익을 극대화하는 방향으로 연구 목적을 설정하는 것이 타당한가? 코즈 정리(Coase theorem)에 따르면, 거래 비용이 최소화되고 당사자 간의 재산권이 명확하게 설정되어 있을 경우, 양측은 자원의 최적 배분을 위해 효율적으로 협상할 수 있다. Lending Club의 경우 투자자와 플랫폼 간 매칭 비용은 웹사이트 운영 비용 정도로 거의 0에 가깝고, 수익 배분 구조가 명확하게 설정되어 있다. 단기적으로 Lending Club은 수수료 수익만 많이 받으면 이익이기 때문에 대출 승인을 관대하게 할 수 있지만, 이는 결국 투자자의 손실로 이어져

장기적으로 플랫폼의 지속가능성을 해친다. 따라서 코즈의 정리에 의해 양측 모두의 이익을 극대화하는 최적 균형점이 존재한다.

이에 따라 본 연구는 샤프 비율을 목적함수로 설정하고, 각 대출의 내부수익률(IRR)과 무위험 수익률(대출 승인 시점 동일 만기의 미국 국채금리)를 바탕으로 초과수익률을 계산하여 이를 최대화하는 대출 선별 기준을 설계한다. 요약하면, 투자자에게 가장 유리한 대출 선별 기준을 찾아 위험 대비 수익을 극대화하는 것이 연구의 핵심 목표이다. 이를 통해 P2P 대출 플랫폼에서 단순 부실 회피를 넘어 수익과 위험을 균형 있게 고려한 투자전략을 제시하고자 한다.

1.3 기대효과

본 연구를 통해 투자자 중심의 새로운 신용평가 프레임워크를 제시함으로써 다음과 같은 기대효과를 얻을 수 있다. 첫째, 투자 수익률의 개선이다. 부도 가능성만 회피하는 기존 전략과 달리, 수익률까지 고려한 대출 선별로 포트폴리오의 기대수익을 높이면서도 변동성을 관리할 수 있다. 이는 실제 투자자의 재무성과 개선으로 이어질 수 있을 것이다. 둘째, 의사결정 지원 향상이다. 제안 모형은 샤프 비율의 극대화라는 명확한 평가 기준에 따라 만들어졌기 때문에 투자자나 금융기관이 위험-수익 균형을 정량적으로 고려하여 의사결정을 내리는 데 도움을 준다. 플랫폼 운영자 입장에서도 투자자 성향에 맞춘 상품 추천이나 맞춤형 포트폴리오 구성에 활용할 수 있을 것이다. 또한, 샤프 비율을 활용한 의사결정은 장기적 관점에서의 이윤 극대화라는 점에서 지속가능한 금융 플랫폼 생태계를 구축해나가는 데 기여한다. 셋째, 신규 평가체계 확립이다. 부도 가능성을 기준으로 위험을 평가하는 전통적 신용평가에서 나아가 투자성과 최적화 관점의 평가체계를 제시함으로써, P2P 금융뿐 아니라 일반 여신 심사 분야에서도 새로운 모델 개발의 계기를 마련할 수 있을 것으로 기대된다.

2. 본론

2.1 연구 방법

2.1.1 연구 모형 소개

본 연구에서는 4개의 모형을 활용했다. 이 중 3개의 모형은 서로 단계별로 연결되어 있다. 즉, 순차적으로 각 단계에서 서로 다른 모형과 의사결정 기준을 적용한 뒤, 샤프 비율 극대화 관점에서 성과를 비교하였다. 나머지 1개 모형은 LLM 기반 feature selection을

적용한 것으로 나머지 3개 모형과는 관련되어 있지 않다. 각 모형을 간략히 소개하면 아래와 같다.

① 1차 분류모형

우선 대출 부도 여부를 예측하는 이진 분류모형을 구축하였다. 종속변수는 [loan_status]를 기반으로 가공한 부도 여부(0=정상상환, 1=부도)이며, 투자 이전에 원금 손실을 최소화하기 위한 전통적인 접근이다. 이 단계에서는 로지스틱 회귀를 베이스라인으로 하고, 트리 기반 앙상블 등 다양한 분류 알고리즘을 적용하였다. 학습은 train 데이터로 수행하고 validation 데이터에 대해 예측을 산출한 후, 예측 확률에 대한 다양한 임계값(threshold)을 적용하여 대출 승인/거절을 결정하고 그에 따른 샤프 비율을 계산하였다. 각 모델에 대해 최적의 threshold를 탐색하여 샤프 비율을 최대화하고, 그 값을 1차 모형의 성능으로 삼았다. 1차 분류모형의 성능은 이후 다른 모델들의 성능을 평가할 때 기준점으로 설정되었다.

② 2차 회귀모형

다음으로 각 대출의 내부수익률(IRR)을 예측하는 회귀모형을 구축하였다. 종속변수를 IRR로 설정하여, 개별 대출의 예상 수익률 크기 자체를 예측하고자 한 것이다. 마찬가지로 train 데이터로 모델을 학습하고, validation 데이터에 대해 IRR 예측값을 산출하였다. 그런 다음 예측된 IRR 값에 기반하여 대출 승인 여부를 결정하는 임계값을 조정함으로써 샤프 비율을 계산하였다. IRR 예측모형의 성능은 1차 부도 예측모형과 직접 비교되었으며, 실제로 IRR을 타겟으로 할 경우 샤프 비율의 최대값이 얼마나 개선되는지 확인하였다. (IRR 및 샤프 비율 산출의 구체적 개념은 ‘2.1.2 핵심 개념 소개’에서 상세히 다룬다.)

③ 3차 2단계 결합모형

2차 회귀모형의 예측값(IRR_hat)을 활용하여, 투자 결정단계를 한 번 더 세분화한 2단계 모델을 설계하였다. 먼저 IRR_hat에서 해당 대출의 무위험 수익률(동일 만기의 국채 금리, treasury)를 차감하여 예상 초과수익률을 계산하였다. 이 예상 초과수익률이 양수인 경우 해당 대출을 ‘수익성 있음(profitable)’=1, 음수인 경우 ‘수익성 없음’=0으로 정의하여 새로운 종속변수 [profitable]을 생성했다. 이를 예측하는 이진 분류모형을 2단계로 학습함으로써, 모델이 직접 샤프 비율 극대화 관점에서 대출 승인 여부를 결정하도록

하였다. 다시 말해, 2차 회귀모형이 추정한 IRR_{hat} 을 입력으로 받아 국채 대비 수익성 여부를 판단하는 분류모델을 만들고, 이 모델의 예측확률에 대해 threshold를 적용하여 샤프 비율을 계산하였다. 이 과정에서도 다양한 threshold를 시험하여 샤프 비율이 최대가 되는 지점을 찾았으며, 해당 값을 3차 모형의 성과로 기록하였다.

④ LLM 기반 Feature Selection 모형

한편, IRR 예측을 더욱 향상시키기 위한 실험으로 대규모 언어모델(LLM)을 활용한 피처 선택을 시도하였다. 이는 각 변수의 의미와 중요도를 LLM이 분석하도록 하여, 모델에 투입할 중요 변수를 선별하는 접근이다. 예를 들어 변수 설명이나 데이터 디스커리를 LLM에 제공하고 IRR에 영향이 클 것으로 예상되는 요인을 도출한 뒤, 이렇게 선정된 변수를 가지고 IRR 예측 회귀모형을 다시 학습해보았다. 이 LLM 기반 변수선택 모형도 마찬가지로 validation 데이터에 대한 IRR_{hat} 을 산출하고 샤프 비율을 평가하여, 기존 모델과의 성능 차이를 비교하였다. LLM을 활용한 시도는 전통적 피처 선택과 비교하여 새로운 변수 발굴이나 중요도 확인에 도움이 될 수 있을지 탐색한 것이다.

위의 4 가지 모형들의 성능은 모두 동일하게 평가되었다. sklearn(Scikit-learn) 모듈의 train_test_split 함수를 활용해 학습 데이터(train set)와 검증 데이터(validation set)을 3대 1의 비율로 설정했다. 학습 데이터를 활용해 모델을 학습시킨 후, 검증 데이터를 활용해 평가 지표(샤프 비율)를 극대화하는 하이퍼파라미터를 구했다. 즉, 투자 여부의 기준이 되는 threshold를 조금씩 변화해나갈 때 포트폴리오 전체의 샤프 비율이 어떻게 변화하는지 관찰하고, 각 모델별 샤프 비율의 최대값을 비교하여 어떤 모델이 투자 선택에 가장 유리한지를 평가하였다. 샤프 비율을 기준으로 전통적인 부도예측 모형(1차)과, 수익률 예측 모형(2차)의 성과를 먼저 비교하고, 이어서 2단계 결합 모형(3차)과 LLM 활용 모형(4차)이 추가적으로 샤프 비율을 향상시킬 수 있는지 검증하였다.

2.1.2 연구 모형 설계 이유

Lending Club의 전통적인 부도 예측은 부도 확률이 높은 대출을 사전에 차단하여 원금 손실을 최소화하고 플랫폼의 건전성을 유지하는 데 목적이 있다. 하지만 이러한 방식은 샤프 비율과 같이 수익률과 변동성을 함께 고려하는 투자 성과 지표를 최적화하는 데 한계가 있다. 예를 들어, 부도 확률이 낮더라도 수익률이 매우 낮은 대출이 승인될 수

있으며, 반대로 부도 확률이 다소 높더라도 기대 수익률이 높아 샤프 비율 개선에 기여할 수 있는 대출이 배제될 수 있다. 본 연구는 이러한 한계를 보완하기 위해, 부도 여부와 무관하게 기대 수익률을 기준으로 대출 승인 여부를 결정하는 새로운 모형을 제안한다.

앞서 설명했듯 1차 분류모형은 부도 여부를 기준으로 대출의 리스크를 판단하는 전통적인 관점에 따라 종속변수를 `loan_status`로 설정하여 부도 여부를 예측하고, 해당 결과를 바탕으로 샤프 비율을 산출한다. 따라서 '`loan_status`'를 예측하는 1차 분류모형을 본 연구의 기준모형(baseline model)으로 삼고 향후 다른 모델들을 평가할 때의 기준점으로 삼았다.

그러나 샤프 비율은 수익률의 크기와 변동성(투자 대비 리스크)을 동시에 반영하는 지표이므로 단순히 부도 여부만을 예측하는 해당 1차 분류 모형은 본 연구의 목적함수와 직접적으로 일치하지 않는다. 따라서, 본 보고서에서는 목적함수와의 정합성을 높이기 위해, 샤프 비율에 직접적으로 영향을 미치는 IRR을 예측 대상으로 설정하였다. 이처럼 각 대출의 IRR을 예측하는 2차 회귀모형(`IRR_hat`)의 구체적인 설계와 절차는 추후 상세히 기술한다.

본 연구의 2단계 모델링은 단일 모델로 전체 과정을 학습하는 방식에 비해, 예측 목적의 분리와 의사결정 과정의 명확화라는 두 가지 측면에서 타당성을 가진다.

첫째, 목적변수의 차이이다. 1단계는 각 대출의 예상 수익률을 나타내는 IRR을 예측하는 회귀모형이며, 2단계는 1단계 예측 IRR에서 무위험 수익률을 차감하여 산출한 초과 수익률을 기준으로 `profitable` 여부를 분류하는 이진 분류모형이다. 이를 위해 먼저 예측 IRR 값(`IRR_hat`)에서 무위험 수익률(`treasury`)을 차감하여 초과 수익률을 계산하였다. 이후, 초과 수익률이 양수인 대출을 '수익성 있음(0)', 음수인 대출을 '수익성 없음(1)'으로 정의한 `profitable` 변수를 생성하여, 이를 종속변수로 하는 이진 분류모형을 설계하였다. 즉, 1단계의 모형은 연속형 수익률 예측값, 2단계의 모형은 투자 승인 여부 결정이라는 상이한 종속변수를 도출하므로, 두 단계가 분리되어 설계되는 것이 합리적이다.

둘째, 모형 해석과 정책 적용의 용이성이다. 1단계 IRR 예측 결과를 통해 각 대출의 수익률 분포를 파악할 수 있으며, 2단계 분류 결과를 통해 샤프 비율 최적화를 위한 투자대상 선별 과정을 명확하게 제시할 수 있다. 이는 단일 블랙박스 모델보다 의사결정 과정의 투명성을 높이고, 실무 적용 시 투자자 또는 기관의 위험 선호도에 따라 기준을 조정할 수 있는 유연성을 제공한다.

셋째, 모형적 안정성이다. 선형 회귀모형에서는 동일한 변수 집합을 연속적으로 사용하면 다중공선성으로 인해 계수 추정이 불가능하거나 불안정해질 수 있으나, 본 연구에서는

단계별로 입력변수의 역할과 사용 목적이 다르므로 이러한 문제가 완화된다. 또한 비선형 모델에서도 1단계와 2단계의 목적과 데이터 구조를 분리함으로써, 각 단계가 해당 목표에 최적화된 학습을 수행할 수 있다.

따라서, 본 연구의 2단계 설계는 샤프 비율이라는 목적함수 최적화를 위해 예측-판단 과정을 구조적으로 구분하고, 모형 해석 가능성과 실무 적용성을 높이는 측면에서 충분한 타당성을 가진다.

넷째, 단순 선형모형은 로 표현되며, 관측된 설명변수 만으로는 비선형성과 상호작용과 같은 잠재효과를 분리해 추정하기 어렵다. Two-stage 접근은 이러한 한계를 보완하기 위해 1단계에서 수익률 함수를 나타내는 잠재효과 를 먼저 학습하고(본 연구에서는 IRR 예측), 2단계에서는 이를 원천 특성 와 결합하여 형태로 최종 의사결정을 학습한다. 다시 말해 1단계는 고차원 특성을 수익률 중심의 요약표현으로 압축하고, 2단계는 그 요약정보와 원천 정보를 함께 사용해 “국채 대비 수익성 여부”라는 최종 결정경계를 학습한다. 승인 시점에 직접 관측할 수 없는 수익률 정보를 모형 내부에서 대리변수로 구성한 뒤 이를 근거로 투자 결정을 내리도록 설계한다는 점에서, 본 접근은 샤프 비율 극대화 목적과 정합적으로 연결된다.

마지막으로, 본 연구에서 LLM 기반 feature selection을 시도한 것은 기존 통계적 변수 선택 기법에 비해, 도메인 지식의 반영과 해석 가능성 강화라는 측면에서 타당성을 가진다.

첫째, 변수 선택 기준의 차별성이다. 전통적인 Lasso, Elastic Net 등은 데이터 내 상관관계에 기초해 변수 집합을 축소하지만, LLM 기반 접근법은 경제학·금융학의 도메인 지식을 텍스트 프롬프트 형태로 내재화하여, 이론적으로 설명 가능한 변수와 그렇지 않은 변수를 구분할 수 있다. 예컨대, 부채상환비율(DTI)의 음(-)의 방향성, 소득(log_annual_inc)의 양(+)의 방향성은 학부 수준의 경제학 논리에 부합하며, 이는 데이터 기반 자동화 기법만으로는 확보하기 어려운 직관적 타당성을 제공한다.

둘째, 모형 해석과 정책 적용의 용이성이다. LLM이 제시한 변수 중요도와 방향성은 단순한 수치적 결과를 넘어 “왜 이 변수가 IRR 예측에 중요하며, 어떤 방향으로 작용하는가”에 대한 서술적 근거를 동시에 제공한다. 이는 투자자나 정책 입안자가 모형의 판단 과정을 투명하게 이해하고, 위험 선호도나 규제 기준에 따라 가중치를 조정할 수 있는 유연성을 확보하게 한다.

셋째, 탐색적 분석에서의 효율성이다. LendingClub 데이터는 수백 개의 변수를 포함하나, 이 중 상당수는 예측 목적과 무관하거나 사후적 정보로 구성되어 있다. LLM 기반 선별

과정은 이러한 변수들을 조기 제거하고, 경제적으로 해석 가능한 후보군을 좁혀줌으로써, 후속 단계의 회귀·분류 모형 학습에서 계산 효율성을 높이고 ‘차원의 저주’와 마주할 가능성을 완화한다.

종합하여 정리하면, 샤프 비율의 두 구성 요소(수익률 평균과 표준편차)에 직접적으로 영향을 미치므로, 전통적인 부도 예측모델보다 샤프 비율 최적화에 더 적합하다. 결과적으로, 본 보고서에서 제안하는 모형은 단순 부도 확률에 대한 리스크 관리 중심의 평가 체계를 넘어, 장기적 관점에서 위험 대비 수익 극대화라는 투자자 중심의 의사결정 지원 도구로 기능할 수 있다.

2.1.2 핵심 개념 소개

① 수익률 지표 IRR

P2P 대출 플랫폼인 Lending Club의 투자 구조에서는 투자자가 개별 대출에 자금을 투자한 후 매월 일정한 원리금을 상환받는 방식으로 수익을 얻게 된다. 이러한 정기적 현금흐름을 전제로 한 투자 구조는 일반적인 주식 또는 채권 투자와는 본질적으로 상이하다. 특히 현금흐름의 시점이 고정되어 있기 때문에, 단순한 총수익률이나 연환산 수익률로는 이들 현금흐름의 시간적 가치를 충분히 반영하기 어렵다.

이에 따라 본 연구는 개별 대출의 수익률 지표로 내부수익률(Internal Rate of Return, IRR)을 채택하였다. IRR은 미래 현금흐름을 현재가치로 환산했을 때 순현재가치가 0이 되도록 하는 할인율로 정의된다. 즉, 각 월별 상환금에 대해 해당 시점까지의 할인율을 적용하여 현재가치를 계산함으로써, 해당 투자로부터 얻을 수 있는 실질 수익률을 나타낸다. IRR을 사용하면 단순한 원금 대비 회수 금액의 비율이 아닌, 현금흐름의 시간가치를 내재적으로 반영한 동적 수익률 측정이 가능하다. 이처럼 현금 흐름을 토대로 계산되는 IRR 수익률은 금융회사 입장에서 각 대출에 대한 수익성을 평가할 때 유용한 지표이다. 왜냐하면 각 대출 건에 대해 금융회사는 차입자로부터 매월 약정된 할부금을 받을 것으로 기대하며, 금융회사는 이러한 할부금을 그때그때 재투자하면서 수익을 극대화하려고 노력하기 때문이다.

IRR 기반의 수익률은 향후 분석에서 초과수익률(Excess Return) 계산에 활용되며 이는 동일 만기의 국채 수익률(무위험 수익률)을 기준으로 한 상대적인 성과 지표이다. 이후 이러한 초과수익률의 평균과 표준편차를 기반으로 투자자의 위험조정 성과를 나타내는 샤프 비율을 산출한다.

모든 대출건에 대해 IRR을 계산한 결과, 일부 사례에서 IRR이 음수로 나타나는 현상을 관찰되었다. 이는 해당 대출이 발생시키는 현금흐름의 현재가치가 초기 투자금보다 낮음을 의미하며, 투자자가 실질적으로 손실을 입은 경우에 해당한다. 이러한 음수 IRR은 주로 조기 상환 없이 장기간 연체되거나 상환금이 거의 없었던 부실 대출에서 발생하며, 대출 투자로 인한 실질적인 손실 수준을 정량화하는 데 유용한 지표로 기능한다. 따라서 샤프 비율 산출식에 이를 적용하여 손실비용을 포함하였다. 이에 대한 자세한 내용은 샤프 비율 산출법에서 후술하겠다.

한편, 본 연구에서는 조기상환이 이루어진 대출의 경우에도 IRR을 계산하여 분석에 활용하였다. 조기상환은 대출자가 만기 이전에 원리금을 모두 상환하는 경우로, 표면적으로는 연체나 부도와 달리 “정상 상환”에 속한다. 그러나 투자자 입장에서는 예정되었던 미래 이자수익을 온전히 얻지 못하게 되므로, 이는 수익 기회의 축소로 해석될 수 있다. 그러나, 본 연구에서는 조기상환 사례 역시 정상 상환과 동일하게 현금 흐름을 $-funded_amnt + installment \times term$ 방식으로 전처리하였다. 이는 LendingClub 데이터에서 대출별 월별 현금 흐름이 정확히 제공되지 않는 한계 속에서, IRR 계산을 가능하게 하기 위한 합리적 단순화이다. 일반적으로 금융회사는 대출 조기상환에 대해 수수료를 부과하여 기회비용을 보전하므로, 조기상환자의 현금 흐름 가치는 평균적으로 정상 상환자의 현금 흐름 가치와 동일하다고 가정하였다.

IRR 계산 과정에서 수학적으로 값이 산출되지 않는 경우(예: 초기 현금유출이 지나치게 커서 내부수익률 방정식에 실근이 존재하지 않는 상황)에는, IRR을 계산 불가능 값으로 두지 않고 최저값을 부여하는 방식으로 패널티를 적용하였다. 한편, 현금 흐름이 지나치게 큰 양수여서 IRR 계산 과정에서 내부수익률 방정식에 실근이 존재하지 않은 대출 사례는 없었다.

이와 같이 IRR은 단순한 회계적 수익률 지표가 아니라, 대출 현금흐름의 시계열 구조와 불확실성을 반영하는 지표라는 점에서 본 연구의 핵심적 의의를 가진다. 즉, 투자자는 매월 발생하는 상환금이 언제, 어떤 규모로 돌아오는지에 따라 동일한 총상환액을 받더라도 실제 수익률은 달라질 수 있으며, IRR은 이를 내재적으로 포착한다. 본 연구에서 조기상환과 계산 불능 사례에 대한 가정은 이러한 IRR의 특성을 현실적으로 보완하여, 이후 단계의 샤프 비율 최적화 모형을 구축할 때 보다 신뢰할 수 있는 입력값을 제공한다.

이와 같은 이유로 IRR을 수익률 지표로 채택하여 분석함으로써, 샤프 비율 산출 시 투자 손실과 조기상환으로 인한 효과까지 폭넓게 고려할 수 있었다. 결과적으로 IRR 예측을 통해

샤프 비율을 직접적으로 개선하고자 하는 본 연구의 목적에 부합하는 지표 체계를 구축하였다.

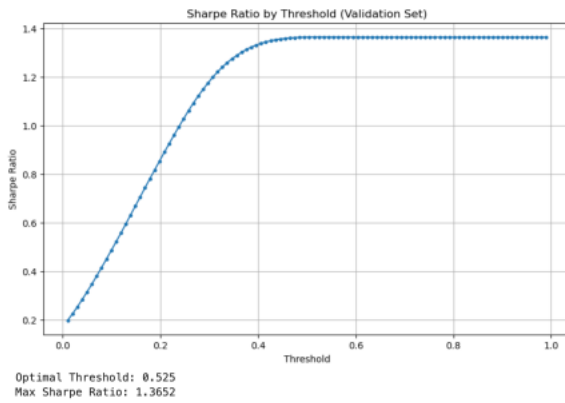
② IRR_hat

앞서 언급했듯 본 연구의 최종 목표는 샤프 비율의 극대화이다. 단순히 부도 여부만을 예측하는 이진 분류모형은 리스크 회피 전략에 그치기 쉽고, 각 대출의 수익률 크기와 분포의 이질성은 반영하지 못한다. 반면, IRR을 예측하면 대출 간 기대 수익의 차이를 정밀하게 추정할 수 있어, 결과적으로 포트폴리오 수익률의 분산(샤프 비율의 분모)에까지 영향을 미칠 수 있다. 따라서 IRR을 예측 대상으로 삼는 것이 샤프 비율 최적화에 더 유리한 방향이라고 판단하였다. 실제로 분석 결과에서도, 종속변수를 loan_status로 한 1차 분류모형 대비 IRR을 종속변수로 한 2차 회귀모형의 샤프 비율이 더 높게 나타났다. 이는 단순 부도 예측보다 직접적인 수익률 예측이 샤프 비율 개선에 유리함을 시사하는 대목이다.

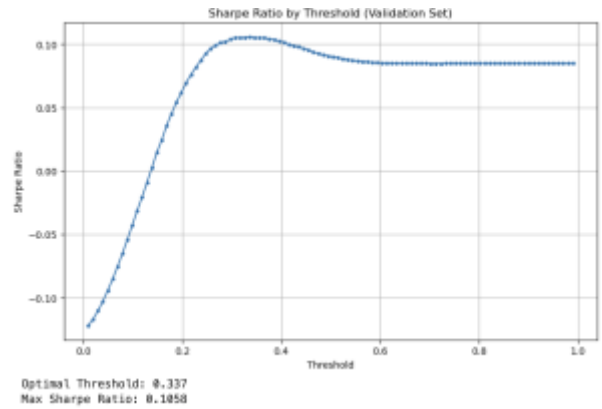
위와 같은 예측 IRR 값은 대출별로 정의되는 건별 수익률이므로, 모델이 개별 대출 단위로 미시적인 판단을 내릴 수 있는 기반을 제공한다. 즉, 각 대출의 예측 IRR 값에 기초한 전략적 투자 선별(예: 특정 예측 IRR 값 이상에 대해서만 대출 승인, 무위험 채권 수익률과 비교하여 수익성 있는 대출 선별 등) 등에 활용될 수 있다.

③ 샤프 비율 산출법

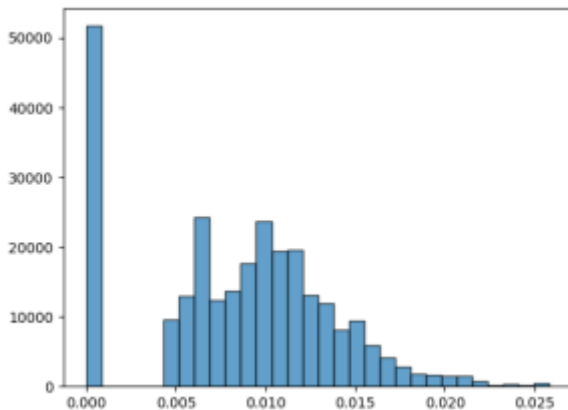
본 연구는 개별 대출에 대한 투자 여부를 판단하는 분류모형의 예측 확률을 활용하여 다양한 기준(threshold)에 따라 투자 여부를 결정하고, 이에 따른 샤프 비율을 산출하는 접근 방식을 채택하였다. 이때 대출의 수익률은 IRR을 사용하고 무위험 수익률은 동일 만기의 국채 수익률로 간주하여 투자에 따른 초과수익률(Excess Return)을 계산하였다.



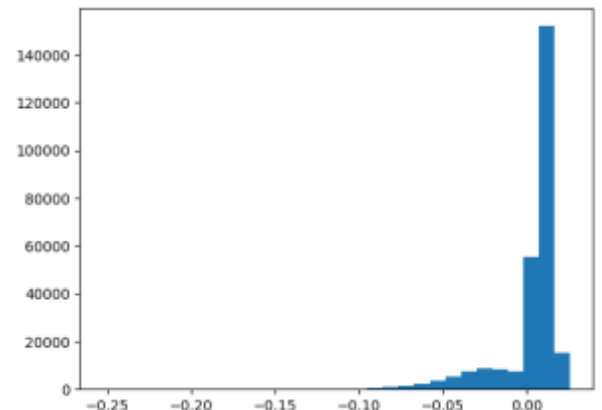
[그림1] 선형회귀 모델: 부도 건의 Return을 0으로
처리



[그림2] 선형회귀 모델: 부도 건의 Return을 irr로
처리



[그림3] 부도 건의 Return을 0으로 반영한 경우 IRR
분포



[그림4] 부도 건의 Return에 음수 IRR 반영한 분포

샤프 비율은 투자 전략의 성과를 평가하는 데 널리 사용되는 지표이지만, 부도 건의 Return을 0으로 처리할 경우 심각한 왜곡을 초래할 수 있다. Return을 0으로 통일하면 손실이 사라진 것처럼 분포가 왜곡되어, 변동성이 인위적으로 축소되고 샤프 비율이 높아진다. 이는 실제 투자자가 직면하는 부도 손실을 반영하지 못해 위험조정 성과를

과대평가하는 결과를 낳는다. 따라서 부도 건은 반드시 음수 IRR을 반영하여 손실 리스크를 현실적으로 포함시켜야 하며, 이를 통해서만 샤프 비율가 투자 전략의 성과를 올바르게 설명할 수 있다. 따라서 본 연구에서는 이와 같은 음의 IRR도 의사결정에 따른 수익률 계산 과정에 포함하여 실제 투자자의 관점에서 대출을 승인했을 경우 감내하게 될 손실까지 고려하였다.

구체적으로 검증 데이터셋에 대해 학습된 분류 모델의 예측 확률을 기반으로 임계값에 따라 다음과 같은 의사결정을 수행했다.

예측값(또는 확률)이 임계값 이하 → 투자 실행: 수익률 = 해당 대출의 IRR

예측값(또는 확률)이 임계값 초과 → 투자 거절: 수익률 = 국채수익률

임계값을 설정할 때, 회귀모형의 경우 예측값이, 분류모델의 경우 예측확률이 활용되었다. 개별 대출*i*에 대한 내부 수익률은 대출을 승인했을 때 정상상환을 하면 그 수익률은 해당 대출의 IRR이 되고, 정상상환이 되지 못할 경우 음수값을 갖는 IRR만큼 손실이 발생한다. 반대로 대출을 거절한 경우, 미국 국채 투자를 하여 무위험 채권에 대한 수익이 발생한다고 가정했다.

2.2 연구 절차

2.2.1 EDA 및 전처리

1) 데이터 개요 및 분할

Lending Club에서 공개된 대출 데이터셋을 기반으로 분석을 진행하였다. 데이터는 2007년부터 2020년까지의 총 292만 건의 대출 내역을 포함하며, 각 대출에 대해 신청 시점의 신용 정보, 대출 조건, 그리고 사후 상환 결과 등이 기록되어 있다. 그 중 전체 데이터의 약 75%를 학습용(train)으로, 25%를 검증용(validation)으로 분리하였다. 모델 학습과 성능 평가는 이 분할에 따라 실시하였다. 최종적으로 도출된 투자전략의 안정성을 확인하기 위해 추가로 별도의 테스트용 홀드아웃 데이터를 활용하여 결과를 검증하였다. 이에 관한 내용은 2.3.2 모델 안정성 점검에서 자세히 다룬다.

2) 종속변수 [loan_status]

대출 상환 상태를 나타내는 변수 loan_status는 'Fully Paid', 'Charged Off', 'Default',

'Does not meet the credit policy. Status: Fully Paid', 'Does not meet the credit policy. Status: Charged Off' 등 여러 범주값을 가진다. 본 연구에서는 이 값을 바탕으로 부도 여부를 나타내는 이진 분류용 타깃변수를 생성하였다. 구체적으로, 'Fully Paid' 및 'Does not meet the credit policy. Status: Fully Paid'는 정상 상환된 대출로 간주하여 0으로, 'Charged Off', 'Default', 'Does not meet the credit policy. Status: Charged Off'는 실질적 손실이 발생한 부실 대출로 간주하여 1로 라벨링하였다. 이러한 전처리를 통해 투자자의 입장에서 원금 손실 여부를 나타내는 변수를 정의하였고, 이는 1차 분류모형의 종속변수로 사용되었다.

3) 공동신청자 변수

데이터 중 약 3.7%는 공동 신청자(Joint Application)가 있는 대출 사례였다. 기초 분석 결과, 공동신청자가 있는 대출은 단독 신청자보다 부도율이 높게 나타나는 경향이 있었다. 이는 공동신청자의 경우 신청자 두 명의 소득 구조, 책임 분산, 신용정보 해석 등이 단독 신청자와 다르게 작용할 가능성이 있기 때문이다. 동일한 모델을 적용할 경우 예측 성능 저하나 편향이 발생할 수 있으며, 이들을 포함하려면 별도의 모델링 접근이 필요하다. 본 연구에서는 분석의 초점을 개인 단독 신청자에 맞추고자 공동신청자 관련 데이터를 제외하였다. 이를 통해 보다 동질적인 집단을 대상으로 모델을 학습시켜 예측 성능을 높이고자 하였다.

4) 사후변수

대출 실행 이후에 관측되는 변수들, 즉 사후변수는 인과적 왜곡이나 데이터 누설(leakage)을 초래할 수 있으므로 제거하였다. 예를 들어 대출 이후에 결정되는 최근 연체정보(mths_since_recent_delinq)나 최근 신용변동 사항(mths_since_recent_inq) 등의 변수는 대출 결과에 영향을 받아 산출된 값이므로, 모델 입력변수로 사용할 경우 부도여부를 미리 알아버리는 효과가 생길 수 있다. 본 연구의 목적이 대출 승인 시점 이전의 정보만으로 향후 결과를 예측하는 것이므로, 이러한 사후변수들은 분석 변수 목록에서 제외하여 타당한 예측모형을 구성하였다.

5) 결측치 처리

다수의 변수에서 결측치가 발견되었으며, 각 변수의 의미에 따라 적절한 대체 값을

부여하였다. 예를 들어 open_acc(개설된 신용계좌 수)나 mort_acc(주택담보대출 계좌 수) 등의 변수는 결측이 곧 해당 없음(0)으로 해석될 수 있어 0으로 대체하였다. total_bal_ex_mort(주택담보대출 제외 총 잔액), pct_tl_nvr_dlq(연체 없는 계좌 비율) 등 연속형 변수의 경우 중앙값으로 보완하여 극단값의 영향을 줄였다. 또한 특정 그룹의 변수가 공통으로 결측인 패턴이 있어, 예를 들어 Bankcard 관련 변수(bc_open_to_buy 등)와 리볼빙 관련 변수(revol_util 등)가 동시에 결측인 경우 이를 하나의 특징으로 인식할 수 있도록 결측 더미 변수(예: have_bc, have_rev)를 추가하여 해당 정보 부재 자체를 모형이 학습할 수 있게 하였다. 범주형 변수의 경우 결측치를 'Unknown' 등의 별도 범주로 처리하여 정보 손실을 최소화하였다.

6) 이상치 및 분포 변환

연속형 변수들 중 annual_inc(연소득), revol_bal(리볼빙 잔액) 등의 분포는 왜도(skewness)가 큰 특성이 있었다. 이러한 변수를 그대로 사용할 경우 일부 이상치나 치우친 분포로 인해 모델이 과도한 영향을 받을 수 있다. 따라서 annual_inc, open_acc, total_acc 등 심하게 오른쪽으로 꼬리가 긴 분포는 로그 변환을 적용하여 정규분포에 가까운 형태로 변환하였다. 이는 특히 선형모델이나 거리 기반 모형에서 예측 성능을 높이는 효과가 있다. (트리 기반 모델의 경우 이러한 변환 없이도 순서 정보만으로 분류가 가능하지만, 본 연구에서는 여러 모델을 공정히 비교하기 위해 주요 연속 변수들에 일괄적으로 로그 변환을 적용하였다.) 한편, 특이값 탐지를 통해 본 결과 극단적으로 큰 값들은 소수 존재했으나 해당 관측치를 제거하지는 않았다. 데이터가 충분히 크고 트리 기반 모델은 이상치에 비교적 강인하다고 판단하여, 이상치에 대해서는 로그 변환만으로 처리하고 추가적인 제거는 하지 않았다. 전반적으로 결측치 대체와 분포 변환 등 표준적인 전처리 기법을 적용하여 데이터 품질을 개선하였다.

7) 이진화 및 범주형 변수 처리

일부 희소한 범주 변수는 값을 단순화하였다. 예를 들어 tax_liens, pub_rec_bankruptcies, collections_12_mthbs_ex_med 등 대부분 값이 0값이고 극히 드물게 1 이상의 값이 나타나는 변수들은 0/1의 이진 변수로 변환하여 '있음/없음'만 표시하도록 하였다. home_ownership 변수의 경우 'MORTGAGE', 'RENT', 'OWN', 'ANY', 'OTHER', 'NONE' 등의 범주가 있는데, 이 중 'ANY', 'OTHER', 'NONE'은 명확한 주거 형태가 아니거나

신뢰성이 낮은 값이므로 하나의 그룹('OTHER')으로 묶어 정리하였다. term(대출기간) 변수는 36개월과 60개월 두 가지 값만 존재하므로 0/1 이진화하였다. inq_last_6mths(최근 6개월 내 조회 수) 변수는 값의 분포에 따라 0회, 1-2회, 3회 이상 등의 구간으로 나누어 범주화하였다. 또한 기간을 나타내는 mths_since_last_delinq, mths_since_recent_bc, mths_since_recent_inq 등은 연 단위 구간으로 변환하여 값의 범위를 줄였다.

8) 파생변수 생성

기존 변수들로부터 몇 가지 파생변수를 만들어 정보량을 늘렸다. FICO 신용 점수는 fico_range_high, fico_range_low 두 값으로 주어지는데, 항상 ± 4 정도 차이가 나는 동일 등급대 정보이므로 두 값의 평균을 취해 대표 신용점수 변수를 생성하였다. 이렇게 함으로써 중복 정보를 하나로 통합하고 모델 해석을 용이하게 했다. 또 하나는 개인의 신용이력 길이를 나타내는 변수로, 최초 신용계좌 개설일과 대출 신청일 간의 개월 차이를 계산하여 credit_hist_months 추가하였다. 일반적으로 신용 이력이 길수록 장기간 금융거래를 지속해왔음을 의미하므로 신용건전성 판단에 유의미한 지표다.

마지막으로, 본 연구는 무위험 수익률 변수(treasury)를 '월 단위'로 구성하여 각 대출의 IRR과 단위(월·복리 기준)를 정합시켰다. 구체적으로 대출 발행월과 대출기간에 대응하는 미국 국채(상수만기, CMT) 수익률을 매칭한 뒤, 이를 월 수익률로 변환하였다. 절차는 다음과 같다.

① 만기 매칭

36개월 대출 → 3년물 CMT(3Y), 60개월 대출 → 5년물 CMT(5Y)를 사용한다.

② 월별 기준치 산출

각 대출의 발행월 m에 대해, 해당 월에 월평균 CMT 연 수익률(연환산, %)을 구한다.

③ 연 → 월 수율 변환(복리 일관성 확보)

IRR을 월 단위 현금흐름으로 계산했으므로, 무위험 수익률도 같은 단위로 맞춘다.

$$treasury = (1 + r_{CMT, annual})^{\frac{1}{12}} - 1$$

요약하면, 본 연구의 treasury는 발행월 x 만기에 매칭된 CMT 연수익률을 월복리로

변환한 값이며, 이렇게 월 단위로 정렬된 무위험률을 사용해 IRR과 동일 단위에서 초과수익률과 샤프 비율을 계산했다. 이 설계는 현금흐름 기반(IRR) 투자 평가의 시간 단위와 무위험 기준의 시간 단위를 일치시켜, 의사결정과 성과평가의 정합성과 타당성을 높여준다.

이상의 전처리 과정을 통해 총 180개의 특성이 모델에 투입되었으며, 모든 특성은 대출 승인 시점에 획득 가능한 정보로만 구성되도록 관리하였다. 정제된 데이터를 바탕으로 다음 단계에서 다양한 예측 모형을 개발하고 샤프 비율 기반 성과를 평가하였다 .

2.2.2 1단계 분류모형 (부도 예측)

1단계에서는 대출 부도 여부를 사전 예측하는 이진 분류모형을 구축하였다. 이는 전통적인 신용평가 접근으로, 대출 실행 전에 부도 위험이 높은 신청자를 걸러내어 원금 손실을 피하는 데 목적이 있다. 모델링에는 앞서 전처리한 특성들을 모두 독립 변수로 사용하고, 종속변수는 가공된 부도 여부(0 or 1)로 하였다.

우선 로지스틱 회귀(Logistic Regression)를 베이스라인 모델로 설정하였다. 로지스틱 회귀는 해석이 용이하고 과적합 가능성이 낮다는 장점이 있어 기준모델로 적합하다. 이후 성능 향상을 위해 트리 기반 앙상블 모델들을 실험하였다. CatBoost, XGBoost, LightGBM 등의 그래디언트 부스팅 모델과 비교 목적으로 의사결정나무(Decision Tree) 모델도 구축하였다. 이러한 모델들은 다수의 특성들 간의 비선형 관계를 포착하고 상호작용 효과를 반영할 수 있어, 단순 선형인 로지스틱 회귀보다 뛰어난 예측 성능을 기대할 수 있다. 각 모델은 학습 데이터로 훈련한 뒤 검증 데이터에 대한 부도 발생 확률을 예측하였다.

모델 평가에 있어 일반적인 정확도(Accuracy)나 AUC 대신, 투자자의 실제 의사결정 상황을 반영한 샤프 비율을 사용한 것이 특징적이다. 구체적으로, 모델의 예측 확률을 기준으로 다양한 임계값(threshold)을 적용하여 대출 승인 여부를 결정하고, 그 결과로 형성된 포트폴리오의 샤프 비율을 계산하였다. 예를 들어 어떤 threshold를 0.2로 정하면, ‘예측부도확률 $\leq 20\%$ ’인 신청자들은 투자 승인하고 20% 초과인 경우는 거절한다. 승인된 대출들에 투자한 포트폴리오의 IRR 성과들과 거절된 경우는 국채 투자로 간주한 성과들을 종합하여 샤프 비율을 산출한다. 이러한 절차를 threshold 값을 0에서 1 사이에서 변화시키며 수행하면, threshold에 따른 샤프 비율 변화를 관찰할 수 있다. 본 연구에서는 검증 데이터셋에서 샤프 비율이 최대가 되는 임계값을 찾아 해당 모델의 최적 샤프 비율로 정의하였다.

2.2.3 2차 회귀모형(IRR 예측)

2단계에서는 대출의 예상 내부수익률(IRR)을 예측하는 회귀모형을 구축하였다. IRR은 각 대출에 투자했을 때 얻을 것으로 기대되는 월 수익률로서, 이를 사전에 예측하면 투자 결정에 있어 보다 정교한 판단이 가능해진다. 회귀모델의 독립변수는 1차 분류모형과 동일하게 대출 신청 시점의 정보들을 사용하였고, 종속변수는 훈련 데이터의 각 대출에 대해 사후 계산된 실제 IRR 값으로 하였다. (IRR 값은 대출이 완전히 상환되거나 부도 처리된 후 계산된 실현 수익률이다. 학습시에는 이 실제값을 사용하지만, 예측 시에는 해당 값이 없으므로 모델이 추정하게 된다. 이때 사후정보를 제거한 데이터로 학습했기 때문에 정보 누설 없이 IRR을 예측하도록 설계되었다.)

IRR 예측을 위해 회귀용 그래디언트 부스팅 기법을 활용하였다. 여러 회귀 알고리즘을 비교한 결과가 채택되었다. 해당 모델에 대해서도 하이퍼파라미터 튜닝을 수행하여 과적합을 방지하고 예측력을 높였다. 학습 완료 후 검증 데이터에 대한 IRR 예측치(IRR_hat)를 생성하였고, 이를 활용하여 투자 의사결정을 시뮬레이션하였다.

투자 판단은 IRR_hat 값이 높을수록 그 대출이 수익성이 좋을 것으로 기대된다고 보고, IRR_hat이 일정 수준 이상인 대출만 선택하는 전략으로 구현하였다. 구체적으로, IRR 예측모형의 출력에 대해 임계값을 변화시켜가며 ' $IRR_hat \geq \text{임계값}$ '인 대출들에 투자하고 그 미만은 거절하는 방식으로 포트폴리오를 구성하였다. (예를 들어 임계값을 10%로 하면 예측 IRR이 10%이상인 대출들만 투자하고, 나머지는 투자하지 않는 것으로 간주한다.) 이렇게 선택된 포트폴리오의 샤프 비율을 계산하여 임계값별로 비교한 뒤, 샤프 비율을 최대화하는 IRR_hat 기준값을 식별하였다.

2.2.4 3차 2단계 모델(결합 분류모형)

3단계에서는 2단계 IRR 회귀모형과 분류 아이디어를 2단계 모델링 접근을 구현하였다. 앞서 2.1.1에서 개략적으로 설명한 대로, 1차 IRR 예측모형의 출력인 IRR_hat을 활용하여 대출 승인 여부를 최종 판단하는 분류모형을 추가로 만든 것이다.

우선 각 대출의 수익성에 대해 라벨링을 진행하였다. 실제 IRR과 treasury를 뺀 값이 양수이면 profitable(0), 음수이면 non-profitable(1)로 라벨링을 한 이후, 실제 IRR은 사후변수이기에 제거하였다. 이후 1차 IRR 예측모델로 예측한 IRR_hat을 추가하여 독립변수로 사용하였다. 즉 1차 모델에서 사용한 변수 + 이를 사용해 예측한 IRR_hat 을

독립변수로 설정하고, [profitable] 을 종속변수로 두고 이를 예측하는 이진 분류모델을 학습하였다.

2단계 분류모델의 입력 특성으로는 IRR_hat 자체와 기타 주요 신청자 정보를 함께 활용하였다. IRR_hat은 이미 많은 정보를 응축한 예측치이지만, 혹시 IRR_hat의 오류나 편향이 존재할 경우 원천 특성들이 보완적 역할을 할 수 있기 때문이다. 예를 들어 IRR 예측모형이 특정 차용자 그룹에서 일관되게 IRR을 과대평가한다면, 2단계 모델은 그 그룹을 식별하여 해당 대출들을 ‘실제로는 profitable하지 않다’고 분류함으로써 예측 오차에 대한 교정 기능을 할 수 있다. 따라서 2단계 모델은 단순히 IRR_hat에 임계값을 적용하는 것보다 유연하고 자동화된 의사결정을 내릴 수 있다.

분류 알고리즘으로는 첫번째 모델과 마찬가지로 로지스틱 회귀와 트리 기반 앙상블을 시도하였다.

학습된 2단계 모델에 대해서도 검증 데이터에 적용하여 각 대출이 profitable일 확률을 예측하였다. 그리고 첫번째 모델과 유사하게 threshold를 변화시켜가며 어떤 수준에서 투자/비투자 결정을 내릴 때 샤프 비율이 극대화되는지 평가하였다. 예를 들어 threshold=0.5로 하면 모델이 50% 이상 확신하는 경우만 투자, 그 이하는 거절하는 전략이 되고, threshold를 높이면 더 엄선된 대출만 투자하게 된다.

2.2.5 LLM Feature Selection 모델

심사 시점에 관측 가능한 변수만 남기고 사후변수는 전량 제외한 뒤, LLM에 제공하여 IRR 예측 및 샤프 비율 극대화 관점에서 변수별 예측 관련성, 중요도 등을 서술하고 점수화하도록 했다. LLM이 제안한 순위를 기준으로 상위 50개 변수를 선정하여 최종 모델을 동일하게 재학습하여 예측하였다.

2.3 연구 결과

2.3.1 성능 비교 (샤프 비율 극대화 효과)

본 연구에서는 Lending Club 데이터를 활용하여 샤프 비율 극대화를 목적으로 하는 4가지 모델링 접근법을 구현하고 성능을 비교하였다. 각 모델의 결과를 단계별로 제시한다.

▶ 1단계 분류모형 (부도 예측)

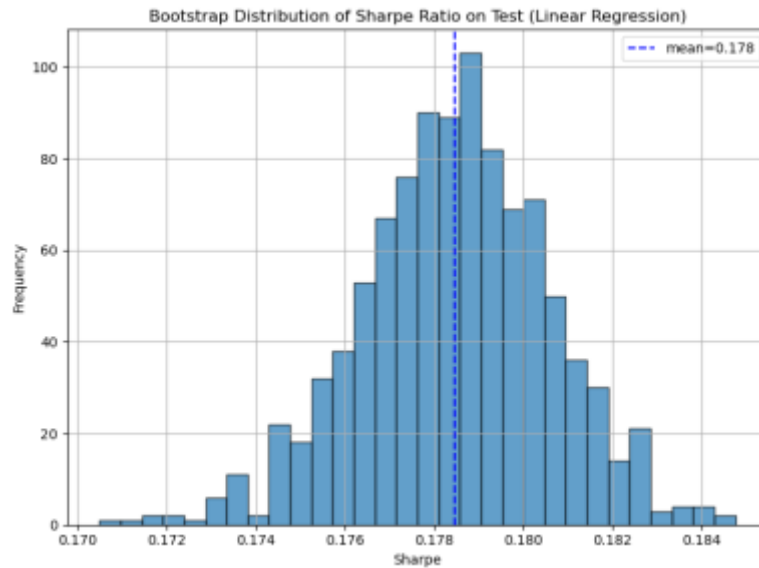
1차 부도 예측 분류모형에서는 7개의 머신러닝 알고리즘을 활용하여 부도 여부를

예측하고, 각 모델의 성능을 샤프 비율(Sharpe Ratio)을 기준으로 평가하였다. [표 1]은 각 모델의 성능 결과를 나타낸다. 분석 결과 CatBoost 모델이 샤프 비율 0.1392로 가장 우수한 성능을 보였으며, 이는 베이스라인 모델인 Logistic Regression의 0.1134보다 22.7% 높은 수치이다. CatBoost 모델의 최적 임계값(threshold)은 0.32로 설정되었다. [그림 5]에서 제시된 Bootstrap Distribution of Sharpe Ratio on Test는 CatBoost 모델의 안정성을 검증한 결과이다. 1,000회의 부트스트랩 샘플링을 통해 얻은 샤프 비율의 분포는 평균 0.178을 중심으로 정규분포에 가까운 형태를 보였으며, 이는 모델의 성능이 일관되고 안정적임을 시사한다. CatBoost 모델이 다른 앙상블 기법들(LGBM, Random Forest, XGBoost)보다 우수한 성능을 보인 이유는 다음과 같이 분석된다.

- ▶ **범주형 변수 처리의 우수성:** CatBoost는 범주형 변수를 자동으로 처리하는 기능이 뛰어나며, Lending Club 데이터에 포함된 다양한 범주형 변수들(신용등급, 대출목적 등)을 효과적으로 활용하였다.
- ▶ **과적합 방지:** CatBoost의 내장된 정규화 기법과 순서형 부스팅(Ordered Boosting) 알고리즘이 과적합을 효과적으로 방지하여 일반화 성능을 향상시켰다.
- ▶ **불균형 데이터 처리:** 부도 여부의 불균형 분포를 CatBoost의 클래스 가중치 조정 기능이 효과적으로 처리하였다.

| 성능평가 지표 | CatBoost | Logistic Regression | LGBM | Random Forest | XGBRF | XGBoost | Decision Tree |
|-----------|---------------|---------------------|--------|---------------|--------|---------|---------------|
| Threshold | 0.32 | 0.317 | 0.3070 | 0.346 | 0.65 | 0.42 | 0.5148 |
| 샤프 비율 | 0.1392 | 0.1134 | 0.1232 | 0.1113 | 0.0901 | 0.1047 | 0.0875 |

[표1] 1차 분류모델 결과



[그림5] 1차 모델 CatBoost Test set, 부트스트랩 (n=1000)

| 모델 | CatBoost |
|-------------|----------|
| Sharpe mean | 0.1785 |
| Sharpe Std | 0.0021 |

[표2] 1차 분류모델 Test set 부트스트래핑 결과(n=1000)

▶ 2차 회귀모형(IRR 예측)

2차 회귀모형에서는 IRR(Internal Rate of Return) 예측을 통해 대출의 수익성을 보다 정교하게 측정하고자 하였다. 4개의 머신러닝 알고리즘을 활용하여 IRR을 예측하고, 각 모델의 성능을 샤프 비율로 평가한 결과는 [표 2]와 같다. 분석 결과 XGBoost 모델이 샤프 비율 0.1985로 가장 우수한 성능을 보였으며, 이는 1차 분류모형의 최우수 모델인 CatBoost(0.1392)보다 42.6% 향상된 수치이다. XGBoost 모델의 최적 임계값은 0.119로 설정되었다. [그림 6],[그림 7]에서 제시된 부트스트랩 분석 결과는 XGBoost 모델의

안정성을 검증한다. 두 가지 전략에 대한 1,000회 부트스트랩 샘플링 결과는 다음과 같다.

음수 IRR로 처리: 샤프 평균 0.1936, 표준편차 0.0019

0으로 처리: 샤프 평균 1.3568, 표준편차 0.0017

XGBoost가 IRR 예측에서 우수한 성능을 보인 이유는 다음과 같이 분석된다.

- ▶ **연속형 변수 처리의 우수성:** IRR은 연속형 수치로, XGBoost의 그래디언트 부스팅 알고리즘이 연속형 타겟 변수 예측에 특화되어 있어 효과적인 성능을 발휘하였다.
- ▶ **복잡한 비선형 관계 모델링:** 대출 특성과 IRR 간의 복잡한 비선형 관계를 XGBoost의 트리 기반 앙상블이 효과적으로 포착하였다.
- ▶ **정규화 효과:** XGBoost의 L1, L2 정규화 기법이 과적합을 방지하여 일반화 성능을 향상시켰다.

부트스트랩 분석에서 확인된 두 가지 부도 Return 처리 전략의 성능 차이는 주목할 만하다.

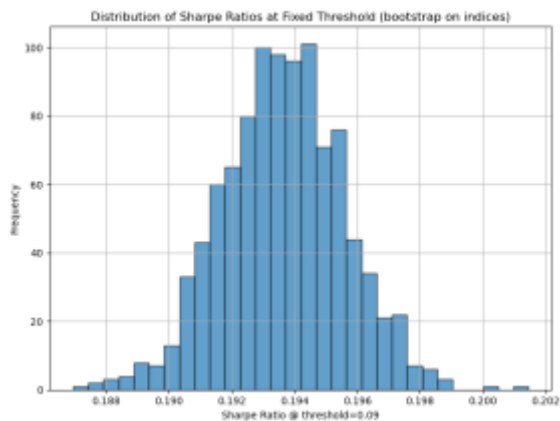
음수 IRR 전략에서는 샤프 비율이 0.1936으로 상대적으로 낮았으나, 분포가 0.188-0.202 범위에서 안정적으로 나타났다.

0으로 처리 전략에서는 샤프 비율이 1.3568로 현저히 높았으며, 1.352-1.362 범위에서 매우 안정적인 분포를 보였다.

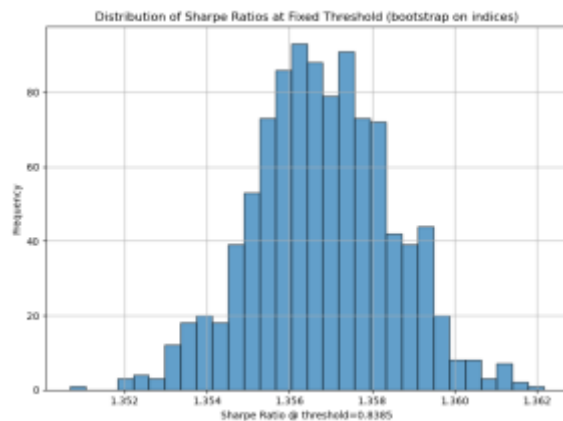
0으로 처리하는 전략이 우수한 성능을 보인 이유는 부도 대출의 리스크를 중립적으로 평가함으로써 포트폴리오 전체의 위험-수익 균형을 개선했기 때문으로 분석된다.

| 성능평가 지표 | CatBoost | Linear Regression | Random Forest | XGBoost |
|-------------|----------|-------------------|---------------|----------------|
| Threshold | 0.149 | 0.545 | 0.109 | 0.119 |
| Sharp ratio | 0.1565 | 0.1339 | 0.1330 | 0.1985 |

[표3] 2차 회귀모델(IRR예측) 결과



[그림6] XGBoost 부트스트랩(부도 Return 음수 IRR로 처리)



[그림7] XGBoost 부트스트랩(부도 Return 0으로 처리)

| 부도 Return | 음수 IRR로 처리 | 0으로 처리 |
|-------------|------------|--------|
| Sharpe Mean | 0.1936 | 1.3568 |
| Sharpe Std | 0.0019 | 0.0017 |

[표4] 2차 회귀모델 Test set 부트스트래핑 결과(n=1000)-부도 Return 처리에 따른 Sharpe 분포

▶ 3차 2단계 모델(결합 분류모형)

3차 2단계 수익성 분류모형은 1차 회귀모형에서 예측한 IRR 값을 활용하여 수익성 여부를 최종 분류하는 결합 모델이다. XGBoost 회귀모형과 3개의 서로 다른 분류모형을 결합한 하이브리드 접근법의 성능 결과는 [표 6]과 같다. 분석 결과 XGB-Cat(XGBoost 회귀 + CatBoost 분류) 결합 모델이 샤프 비율 0.1844로 가장 우수한 성능을 보였다. 흥미롭게도 세 모델 모두 동일한 최적 임계값(0.8415)을 가지며, 이는 2단계 모델 구조에서 최적 분류 기준점이 일관됨을 시사한다.

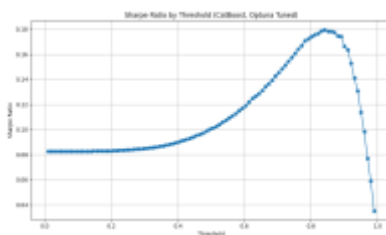
2단계 모델 구조의 핵심 아이디어는 회귀모형으로 연속적 수익성(IRR)을 예측한 후, 이를

기반으로 이산적 수익성 분류를 수행하는 것이다. 결과 분석을 통해 다음과 같은 특징을 확인하였다.

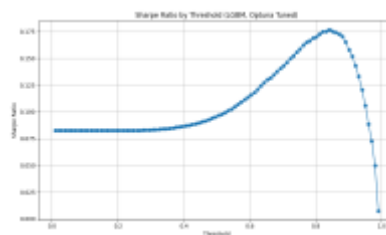
1. **일관된 임계값:** 세 결합 모델 모두 0.8415라는 동일한 최적 임계값을 보여, XGBoost 회귀모형의 IRR 예측이 안정적인 기준점을 제공함을 확인하였다.
2. **분류기별 성능 차이:** CatBoost 분류기가 LGBM과 XGBoost보다 미소하게 우수한 성능을 보였으나, 그 차이는 매우 작았다(0.1844 vs 0.1817 vs 0.1807).
3. **성능 수렴:** 세 모델의 샤프 비율이 0.18 구간에 수렴하는 현상은 2단계 구조에서 1차 회귀모형의 품질이 최종 성능을 크게 좌우함을 시사한다.

| 모델 | <i>XGB-Cat</i> | <i>XGB-LGBM</i> | <i>XGB-XGB</i> |
|--------------|----------------|-----------------|----------------|
| 최적 threshold | 0.8415 | 0.8415 | 0.8415 |
| 최대 샤프 비율 | 0.1844 | 0.1817 | 0.1807 |

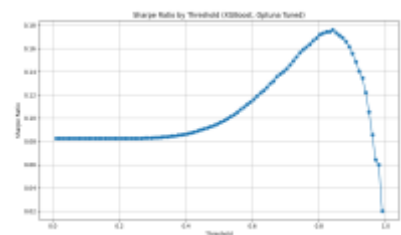
[표6] 3차 수익성 분류모델 결과



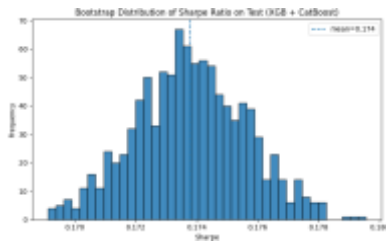
[그림8] XGB-Cat



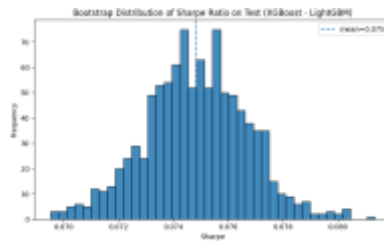
[그림9] XGB-LGBM



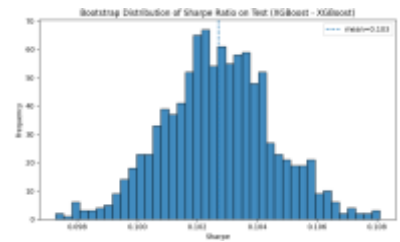
[그림10] XGB-XGB



[그림11] XGB-Cat



[그림12] XGB-LGBM



[그림13] XGB-XGB

| 모델 | XGB-Cat | XGB-LGBM | XGB-XGB |
|-------------|---------------|----------|---------|
| Sharpe Mean | 0.1738 | 0.0748 | 0.1027 |
| Sharpe Std | 0.0018 | 0.0019 | 0.0018 |

[표7] 3차 수익성 분류모델 Test set 부트스트래핑 결과(n=1000)

▶ LLM feature selection 모델

대규모 언어모델(ChatGPT)을 활용하여 Lending Club 데이터의 변수 중요도를 식견적으로 추출하고자 한 것으로, LLM이 제안한 주요 변수들은 신용등급(FICO), 부채비율(DTI), 소득, 신용이력 기간 등 도메인 상 합리적인 것들이었다.

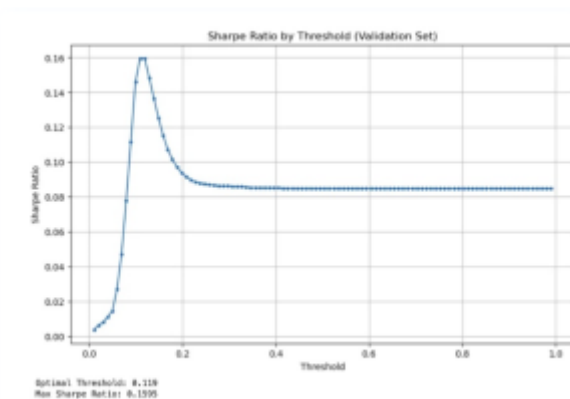
LLM이 높은 중요도로 뽑은 상위 특성들을 가지고 XGBoost 모형으로 IRR예측모델을 재학습한 결과, 최대 샤프 비율은 0.1595로 기존 XGBoost로 IRR예측모델에서 최대 샤프 비율이 0.1985를 기록한 것에 비해 오히려 성능이 하락했다.

그러나 LLM feature selection 모델은 기준모형으로 설정한 로지스틱 회귀 모형의 최대 샤프 비율 값(0.1134)보다 유의미한 성능 향상이 있었다. 이는 실제 머신러닝 모델이 데이터로부터 학습한 중요변수와 LLM이 사전 지식에 기반하여 추론한 중요 변수가 상당 부분 일치하고 있음을 의미한다. LLM이 제시한 몇 가지 변수(예: num_tl_90g_dpd_24m, num_tl_30dpd 등)는 기존에 간과했던 요인을 상기시켜주는 효과가 있었으나, 이미 전처리 단계에서 대부분 포함된 정보였다. 따라서 LLM 활용 피처 선택은 모델링에 즉각적인

이득을 주지는 않았지만, 도메인 지식 관점에서 모델을 해석하거나 추가적인 아이디어를 얻는 보조 도구로서 유용함을 확인하였다. 향후 LLM 기술이 발전함에 따라 데이터 의미에 기반한 피처 엔지니어링을 자동화하는 방향으로 응용 가능성을 엿볼 수 있었다.

| | 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 stabili... | 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_sld_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 ex... | -0.45 |
| 14 | num_tl_90d_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_tl_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 l... | 0.30 |
| 19 | num_actv_bc_tl | 0.28 | 1 | Active bankcard tradelines can indicate credit... | 0.28 |
| 20 | total_rev_h_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 |

[그림14] LLM Selected Features



[그림15] LLM_최대 샤프 비율

앞서 구축한 세 가지 주요 모델(1차 부도분류, 2차 IRR 회귀, 3차 profitable 여부 분류)에 대하여 샤프 비율 기준 성과를 비교하였다. 각 모델별로 검증 데이터셋에서 측정된 최고 샤프 비율의 평균을 정리하면 다음과 같다.

5가지 모델(Logistic, CatBoost, XGBoost, LightGBM, Decision Tree)에 대해 이러한 과정을 거쳐 샤프 비율을 비교한 결과, 앙상블 모형들이 베이스라인 로지스틱 모형보다 우수한 성과를 보였다. 특히 트리 기반 부스팅 모델들은 부도에 영향을 미치는 비선형 패턴을 잘 포착하여 더 높은 샤프 비율을 달성하였다.

| 모델 | 최대 샤프 비율 평균 | 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% |

[표5] 베이스라인 대비 성능 향상률(%)

IRR 예측모형을 통한 투자전략은 1차 부도 예측모형에 비해 더 높은 샤프 비율을 달성하였다. 이는 모델이 단순히 부도 확률이 낮은 안전한 대출만 고르는 것이 아니라, 다소 위험이 있더라도 보상(이자)이 충분히 큰 대출을 골라내 포트폴리오의 위험 대비 수익을 개선한 덕분이다. IRR 예측모형의 샤프 비율이 높게 나온 것은 예상 수익률의 절대적 크기를 고려하는 전략의 효용을 보여준다. 요약하면, IRR 회귀모형은 투자대상 선정을 수익률 기반으로 전환함으로써 샤프 비율을 높이는 데 효과적임을 확인하였다.

한편, 모델 간 비교에서는 부스팅 계열(특히 XGBoost, CatBoost)이 일관되게 가장 높은 샤프 비율을 기록했고, 선형 회귀와 단일 의사결정나무는 상대적으로 저조하였다. 본 연구의 실험에서도 단일 단계의 부스팅 회귀(XGBoost Regressor)가 Two-stage 결합 분류모형을 앞섰다. 이는 부스팅이 비선형과 고차 상호작용을 한 번의 단계에서 자연스럽게 학습해 수익률 신호를 손실 없이 보존하는 반면, Two-stage는 1단계 IRR_hat 추정오차, IRR_hat → profitable 라벨 변환에 따른 이산화·정보손실, 2단계 분류 학습목표(통상 로그로스)와 샤프 비율 극대화 간의 목적함수 불일치 등으로 오차 누적과 캘리브레이션 편차가 발생했기 때문으로 해석된다.

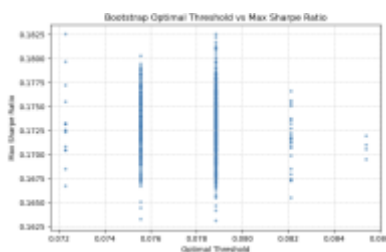
2.3.2 최종모델 선정

본 연구에서는 다양한 예측 모형을 비교한 결과, irr 예측 모형 중 XGBoost Regressor가 최대 샤프 비율 0.1985로 가장 우수한 성능을 보여 최종 모델로 선정하였다. XGBoost는 부스팅(Boosting) 기반의 비선형 회귀·분류 기법으로, 대규모 데이터에서의 계산 효율성과 복잡한 변수 간 상호작용을 포착하는 데 강점을 가진다. 특히 본 연구와 같이 다수의 연속형·범주형 변수가 혼재한 LendingClub 데이터셋에서, XGBoost는 다른 모형 대비

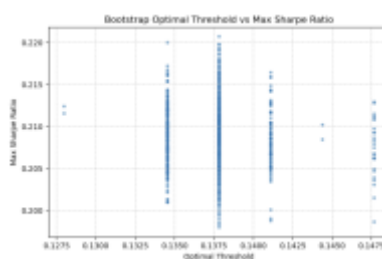
일관되게 높은 예측력을 보였다.

성능을 추가적으로 개선하기 위해 몇 가지 보완적 시도를 수행하였다. 우선, XGBoost 모델 구조에 대한 하이퍼파라미터 조정(parameter tuning)을 통해 모형 복잡도와 과적합 사이의 균형을 최적화하였다. 학습률(learning rate), 최대 깊이(max depth), 트리 수(n_estimators) 등 주요 매개변수를 체계적으로 탐색함으로써, 모형의 일반화 성능을 강화하였다. 기존 트리 개수(n_estimators) 200개, 최대 트리 깊이(max_depth) 5에서 트리 개수(n_estimators)는 500개, 최대 트리 깊이(max_depth)는 6으로 조정했다.

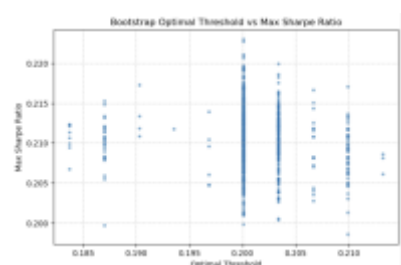
또한, LendingClub 데이터의 특성상 클래스 불균형(class imbalance) 문제가 존재하였다. 즉, ‘정상 상환’에 비해 ‘부도’ 사례가 상대적으로 적어, 단순 학습 시 모형이 다수 클래스에 편향될 우려가 있었다. 이를 해소하기 위해 ‘loan_status’ 불균형 클래스에 대해 SMOTE(Synthetic Minority Over-sampling Technique), 랜덤 언더샘플링(random undersampling), 랜덤 오버샘플링(random oversampling) 등 세 가지 방법을 시도하여 불균형 클래스 문제를 해결한 학습 데이터(train set)로 모형을 학습하여, 검증 데이터(Validation set)에서 1000회 부트스트래핑하여 최대 샤프 비율 값의 평균을 비교하였다. 최종적으로 오버샘플링을 활용하여 소수 클래스의 학습 기회를 확장했을 때, 최대 샤프 비율(max sharpe ratio)의 개선 효과가 가장 큰 것으로 드러났다. SMOTE로 클래스 불균형을 해소한 경우 최대 샤프 비율의 평균은 0.1731, 중앙값 0.1731, 표준편차가 0.0030, 랜덤 언더샘플링의 경우 최대 샤프 비율의 평균은 0.2090, 중앙값 0.2090, 표준편차 0.0037, 랜덤 오버샘플링의 경우 최대 샤프 비율의 평균은 0.2105, 중앙값 0.2106, 표준편차 0.0037로 나타났다.



[그림16] SMOTE_부트스트랩 산점도



[그림17] 랜덤 언더샘플링_부트스트랩 산점도

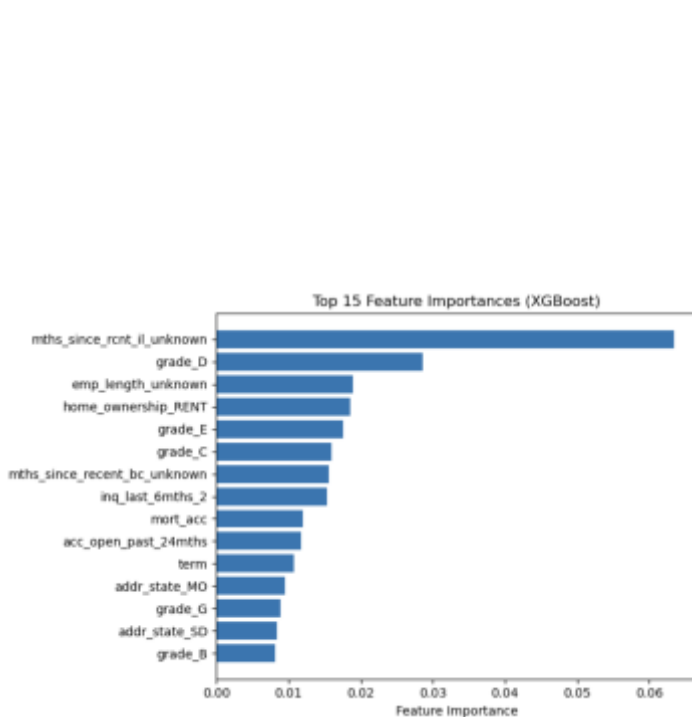


[그림18] 랜덤 오버샘플링_부트스트랩 산점도

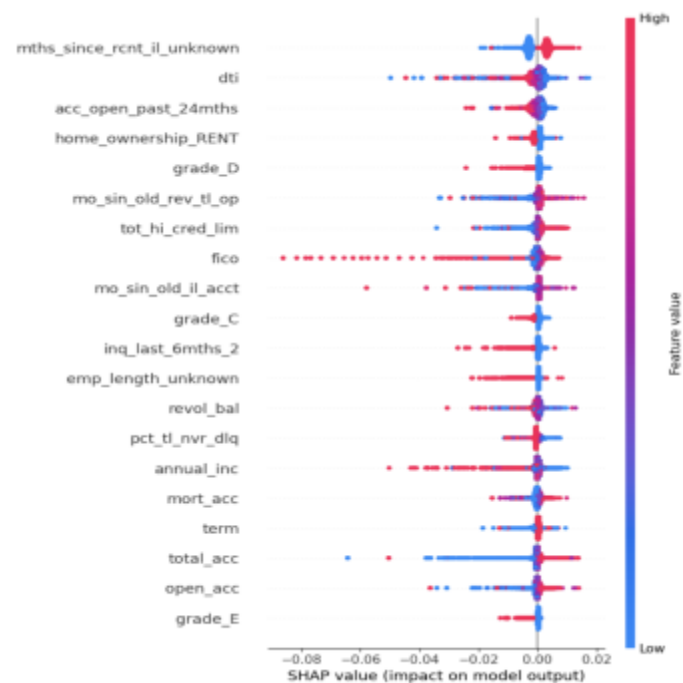
따라서, 본 연구의 최종 모델은 XGBoost를 기반으로 파라미터 최적화와 오버샘플링을 통한 ‘loan_status’ 클래스 불균형 조정을 병행하여 성능을 극대화한 형태로 확정되었다.

2.3.3 모델 안정성 점검 및 추가 분석

샤프 비율 개선이 일반적인지를 확인하기 위해 부트스트랩 기법을 활용하였다. 테스트셋의 대출들을 무작위로 재표본하여 포트폴리오 샤프 비율을 반복 계산함으로써, 각 전략에 대한 샤프 비율 분포와 신뢰구간을 추정하였다(반복 횟수 $n=1000$). 아래 그림은 세 가지 전략에 대한 부트스트랩 샤프 분포를 보여준다.



[그림19] XGB Regressor Feature Importance



[그림20] XGB Regressor SHAP value

최종 모델인 XGBoost Regressor의 중요도 분석 결과, [mths_since_rcnt_il_unknown](최근 할부계좌 정보 부재)이 최상위로 나타났으며, SHAP 분석에서는 해당 값이 1일 때 예측 IRR을 소폭 상향시키는 방향으로 작용했다. 이는 ‘최근 할부계좌가 없어 신규부채 누증이 크지 않다’는 신호가 디폴트 위험 감소를 통해 실현 IRR을 보전했을 가능성과 ‘결측·정의

혼재'에 따른 평균 효과가 반영된 결과로 해석된다. 한편 DTI가 높거나 RENT 거주, 고용연한 미기재는 예측 IRR을 낮추는 요인으로 확인되었고, 반대로 연체경험이 없는 계좌 비율이 높은 차입자는 IRR이 다소 높게 예측되었다. 소득·FICO가 높을수록 IRR이 낮게 예측된 것은 우량 차입자의 금리가 낮아지는 가격 효과가 실현 IRR에도 반영되었기 때문이다. 중요도와 방향성을 함께 고려할 때, '정보 부재의 신호성', '가격과 리스크의 상쇄', '비선형 상호작용'이 IRR 형성의 핵심 메커니즘임을 확인하였다.

3. 결론

3.1 결과 논의 및 시사점

본 연구는 P2P 대출에서 샤프 비율을 직접 최적화하는 관점으로 모형을 설계했고, 그 핵심은 부도 여부가 아니라 수익률(IRR) 자체를 예측해 의사결정에 반영한 데 있다. 실증 결과, IRR 예측 기반 전략(XGBoost 회귀)이 부도확률 예측 기반 전략(로지스틱 등)보다 일관되게 높은 샤프 비율을 보였고, 홀드아웃과 부트스트랩에서도 그 우위가 유지됐다. 이는 단순한 손실 회피가 아니라 가격-위험의 교환관계를 포트폴리오 차원에서 활용했기 때문으로 해석된다. 즉, 약간의 추가 위험이 수반되더라도 충분히 높은 금리(보상)를 가진 대출을 선별함으로써 평균 초과수익을 키우고, 분산은 관리하는 방식으로 위험 대비 수익을 개선했다.

여신 심사의 전통적 목표는 부도 회피지만, 투자자의 목표는 위험 대비 수익 극대화다. IRR은 매월 현금흐름의 시간가치를 반영하는 직접적인 수익률 지표이므로, 이를 예측하면 평균 초과수익을 높일 대출을 고를 수 있고, 선택된 대출군의 분산도 임계값 조정을 통해 관리할 수 있다. 특히 본 연구는 부도 건의 음(-)의 IRR을 그대로 수익률에 반영하여 비현실적으로 높은 샤프가 계산되는 오류를 방지했다는 점에서 평가의 현실성·보수성을 확보했다.

Two-stage 3차 결합분류 모형은 IRR 예측치를 입력으로 받아 무위험수익률 대비 초과수익이 양(+)인지를 직접 분류한다. 이 접근은 의사결정 규칙이 명료하고("국채 대비 이길 것인가?"), 규제·정책 환경에서 설명가능성이 높다는 장점이 있다. 다만 본 데이터에서는 강력한 비선형 회귀(XGBoost)가 이미 상당한 상호작용과 비선형을 학습해 단일 IRR 회귀가 가장 높은 샤프를 기록했다. 즉, 2단계가 항상 우월한 것은 아니며, 설명가능성 강화·정책 임계값 운영이 필요한 맥락에서 전략적으로 선택할 도구로 보는 것이 타당하다.

3.2 한계점과 향후 과제

본 연구의 한계로 먼저 데이터 및 환경 의존성을 들 수 있다. 사용된 데이터는 2010년대 후반까지의 Lending Club 대출 기록으로, 특정 시기의 경제 상황과 Lending Club의 신용정책이 반영되어 있다. 이후의 거시경제 변화(예: 금리 급변, 팬데믹 등)나 플랫폼 정책 수정이 있다면 모델의 예측 성능이 저하될 수 있다. 향후 연구에서는 최신 데이터로 모델을 재훈련하고, 다른 금융 데이터에도 적용해 봄으로써 모델의 범용성을 시험할 필요가 있다.

두 번째로, 공동신청자 대출을 제외한 점은 현실을 단순화한 처리였다. 공동신청자 대출은 본 연구에서 다루지 않았지만 전체 시장의 일부분을 차지하고 있으며, 이들의 특성과 위험도는 단독 신청자와 다를 수 있다. 향후 과제로 공동신청자에 특화된 샤프 최적화 모델을 개발하여 본 연구의 결과와 비교한다면 더욱 포괄적인 전략을 수립할 수 있을 것이다. 향후 연구에서 공동신청자 대출에 특화된 모델을 별도로 설계하여 비교하는 것도 유의미한 과제로 남아 있다.

세 번째로, 현 연구에서 IRR 계산을 위해 일정한 상환 흐름 가정을 사용한 점도 한계로 지적된다. IRR을 구할 때 중요하게 여겨지는 현금 흐름을 계산상 편의를 위해 만기까지 균등하게 원리금을 상환하였다고 가정하였다. 파산이나 상환을 하지 못한 경우나 조기상환을 한 경우에는 다른 정상 상환의 경우보다 현금 흐름이 불규칙하였을텐데, 이를 고려하지 못한 점이 IRR 예측이나 profitable 예측에 어느정도 영향을 주었을 것으로 생각된다. 이에 원리금 상환 데이터가 존재한다면 좀 더 정확한 IRR을 계산하여 좀 더 예측력 높은 모델을 만들 수 있을 것이다. 특히, 본 보고서에서 활용한 LC 데이터와는 다르게 실제 금융회사는 경우 실제 현금 흐름에 대한 정확한 날짜 및 금액 데이터를 활용할 수 있다. 따라서, 본 보고서의 IRR 예측 모형을 실제 금융회사에서 자체적으로 가진 데이터를 활용하여 만든다면 모델의 성능이 더 좋아질 것으로 기대한다.

네 번째, 샤프 비율 자체의 제약이다. 샤프 비율은 평균-분산 관점의 지표이므로 수익률 분포의 비대칭성이나 극단 손실(risk of ruin) 등을 반영하지 못한다. 예컨대 매우 소수의 대출이 큰 손실을 내는 케이스가 있다면, 샤프 비율은 평균과 표준편차에 묻혀 그 위험을 간과할 수 있다. 따라서 실제 투자 적용 시에는 VaR(Value at Risk), CVaR(Conditional VaR) 등 추가적인 위험관리 지표를 함께 고려하는 것이 바람직하다. 샤프 비율 최적화 모델이 제시하는 대출 중에서도 해당 지표들이 허용범위를 벗어나는 것은 제외하거나 한도를 두는 운영상의 보완이 필요하다.

마지막으로, 설명가능한 AI(XAI) 및 윤리적 측면의 고려가 필요하다. 투자 의사결정 자동화 모델은 그 결정 근거에 대한 설명을 요구받을 수 있고, 알고리즘으로 인해 특정 집단이 부당하게 차별받지는 않는지 검토해야 한다. 본 연구에서는 SHAP 분석 등을 통해 주요 결정 요소를 확인하는 작업을 했으나, 향후에는 공정성 지표 등을 함께 모니터링하여 모델이 편향되지 않도록 관리해야 할 것이다.

위의 한계에도 불구하고, 본 연구는 새로운 관점의 신용평가 모형이 투자 성과 향상에 기여할 수 있음을 실증적으로 보였다. 향후 과제들을 보완하고 더욱 정교한 모델링 기법을 도입한다면, P2P 대출뿐만 아니라 다양한 금융 투자 분야에서 활용될 수 있는 보편적인 위험-수익 최적화 모델로 발전할 수 있을 것으로 기대한다.

4. 참고문헌

- [1] Gutierrez, A., & Mathieson, D. (2016). *Optimizing investment strategy in peer-to-peer lending*. Stanford University, CS229 Final Report.
- [2] Li, P., & Han, G. (2018). *LendingClub loan default and profitability prediction*. Stanford University, CS229 Final Report.
- [3] Lee, H. J. (2019). *딥러닝 기법을 이용한 P2P 소셜 대출 채무자 부도 예측모델에 관한 연구*. 디지털콘텐츠학회논문지, 20(7), 1409-1416.
- [4] Sifrain, R. (2023). *Predictive analysis of default risk in peer-to-peer lending platforms: Empirical evidence from LendingClub*. Journal of Financial Risk Management, 12, 28-49.
- [5] Tsai, K., Ramiah, S., & Singh, S. (2014). *Peer lending risk predictor*. Stanford University, CS229 Final Report.
- [6] Chang, S., Kim, S. D., & Kondo, G. (2015). *Predicting default risk of Lending Club*.

Stanford University, CS229 Final Report.

[7] Bhatnagar, P., Chow, N., & Lai, M. (2015). *Demystifying the workings of Lending Club*. Stanford University, CS229 Final Report.

[8] 김민환, 김소희, 장수빈, 전민규, & 정명훈. (2024). *Lending Club data default Prediction*. 제9기 서울대학교 빅데이터 핀테크 전문가 과정 팀 프로젝트 발표자료.

[9] 김민범, 김민주, 김우진, 신주환, 양은정, & 최봉석. (2024). *Lending Club Annual General Meeting*. 제9기 서울대학교 빅데이터 핀테크 전문가 과정 팀 프로젝트 발표자료.

[10] 김태영, 류채환, 박준영, 신해술, & 홍성희. (2024). *Lending Club Project*. 제9기 서울대학교 빅데이터 핀테크 전문가 과정 팀 프로젝트 최종보고서.

[11] 구정모, 김지현, 류현우, 안상원, 우지수, & 정세영(2024). *머신러닝을 활용한 Lending Club 부도예측 모형의 ROE 향상방법 연구*. 제8기 서울대학교 빅데이터 핀테크 전문가 과정 팀프로젝트 최종보고서

[12] 강민채, 이선유, & 정인경(2025). *Lending Club Loan Risk Analysis*. 경제분석을 위한 데이터 사이언스(001) 팀 프로젝트 2차 발표자료.