머신러닝을 활용한 Lending Club 부도 예측 모델의 샤프비율 극대화 방법 연구

2025년 8월

서울대학교 빅데이터 AI 핀테크 고급 전문가 과정 11기 (2조)

강수정, 배기태, 심준선, 이강산, 이선유, 전상언, 황정현

초록

본 연구는 미국 최대 P2P 대출 플랫폼인 Lending Club의 2007~2020년 대출 데이터를 활용하여, 머신러닝 기반 부도 예측 모델을 통해 샤프 비율(Sharpe Ratio)을 극대화하는 투자 전략을 제시한다. 기존 연구가 예측 정확도나 단순 수익률에 집중한 것과 달리, 본 연구는 위험조정수익률을 핵심 지표로 삼아 투자자의 실질적 효용에 초점을 두었다. 이를 위해 약 175만 건의 대출 데이터를 전처리하고, 내생성이 존재하는 변수를 제거하여 현실적 의사결정 환경에 부합하는 독립변수 체계를 구축하였다. 이후 Random Forest, XGBoost, LightGBM 세 가지 트리 기반 앙상블 모델을 적용하여 부도 확률을 추정하고, 임계값 조정을 통해 승인 대출군을 선별하였다. 승인 건은 내부수익률(IRR)을 현금흐름 시뮬레이션으로 산출하여 초과수익률을 계산하고, 거절 건은 동기간 미국 국채수익률을 무위험 수익률로 적용하였다. 100번의 반복 학습 결과, XGBoost가 가장 높은 샤프 비율을 달성하며 위험 대비 수익 극대화에 유리함을 보였다. 또한, SHAP 분석을 통해 신용등급, 부채비율, 최근 신용활동 등 단기·장기적 특성이 모두 성과에 기여함을 확인하였다. 본 연구는 단순한 분류 정확도를 넘어 투자 전략 최적화의 관점에서 P2P 금융의 위험–수익 구조를 체계적으로 분석했다는 점에서 학문적·실무적 의의를 갖는다. 향후 연구에서는 정상 대출의 오분류 최소화, 설명변수 확장, 임계값 탐색 정밀화 등을 통해 보다 정교한 샤프 비율 최적화가 가능할 것이다.

키워드: Lending Club, 머신러닝, 부도 예측, 샤프 비율, 투자 전략, XGBoost

목 차

제 1 장 서 론

제 1 절 연구 배경 및 목적

최근 10여 년간 P2P (Peer-to-Peer) 대출 시장은 급격히 성장하며 전통 금융기관과 더불어 중요한 대체 투자 수단으로 자리매김하였다. 특히 미국의 대표적 P2P 플랫폼인 Lending Club은 2007년 설립 이후 2020년까지 약 175만 건 이상의 대출 데이터를 축적하며, P2P 대출 시장의 선도적 역할을 담당해왔다. Lending Club 플랫폼은 개인 간 직접 대출을 중개함으로써 새로운 금융 생태계를 구축했다는 점에서 혁신적인 비즈니스 모델로 평가 받는다. 이에 따라 Lending Club이 공개한 데이터는 학계와 산업계 모두에서 신용위험 분석, 머신러닝 기반 신용평가 모델 개발, 그리고 투자 전략 연구의 핵심 자료로 활용되고 있다.

P2P 대출 투자의 성과를 자기자본이익률(Return on Equity, ROE), 내부수익률(Internal Rate of Return, IRR) 중심으로 평가할 경우, 동일한 기대 수익률을 보이는 두 투자안에 대해, 수익의 변동성 수준이 크게 상이한 경우의 차이를 제대로 반영하지 못한다. 특히 P2P 대출은 개별 대출의 부도 위험, 유동성 제약, 플랫폼 운영위험 등 다층적인 리스크 구조로 인해 단순한 수익률 지표만으로 투자 성과를 적절히 평가하기 어렵다.

본 연구는 재무이론의 표준적 성과 지표인 샤프 비율(Sharpe Ratio)을 Lending Club의 대출 투자 성과 평가의 중심축으로 설정한다. William F. Sharpe가 고안한 샤프 비율은 위험조정수익률을 나타내는 대표적 지표로서, 투자 포트폴리오의 초과수익률을 변동성으로 나눈 값으로 정의된다. 이는 투자자가 위험 한 단위를 추가적으로 감수할 경우 얻는 초과 수익의 크기를 측정하여, 수익률과 리스크를 모두 반영한 투자 성과 평가를 가능케 한다.

샤프 비율의 도입은 표준화된 지표를 P2P 대출 영역에 적용함으로써, 일관성 있는 성과 비교 기준을 마련하는 데에 기여한다. 더불어 위험과 수익의 상충관계(trade-off)를 고려함으로써 P2P 대출이 가지는 위험 특성을 투자 의사결정에 체계적으로 반영할 수 있다. 이에 따라 자산 배분 최적화, 포트폴리오 구성, 투자전략 수립 등의 의사결정 과정에서 활용 가능한 정량적 기준을 제시하므로, 실무적 가치를 가진다고 할 수 있다.

제 2절 연구 개요

본 연구는 Lending Club의 대규모 데이터를 활용하여 샤프 비율을 극대화하는 투자전략을 탐색하는 것을 목적으로 한다. 수익률 또는 예측 정확도 향상에만 중점을 두는 것이 아니라, 투자자의 실질적인 효용 극대화 관점에서 위험과 수익률의 균형을 추구하는 최적화 전략을 수립하고자 한다.

이러한 목표 달성을 위해 본 연구는 다음과 같은 방법론을 채택한다. 먼저 개별 대출건에 대해 현금 흐름 시뮬레이션을 수행하여 이를 토대로 내부수익률을 계산한다. 계산된 내부수익률은 연환산 수익률로 표준화한다. 다음으로는 머신러닝 기반 부도 예측 모델을 구축한다. 이 과정에서 트리 기반 앙상블 모델인 XGBoost, LightGBM, Random Forest를 활용한 머신러닝 기반 부도 예측 모델을 구축한다. 2007년부터 2020년까지의 데이터를 분석함으로써 개별 차입자의 부도 확률을 추정하며, 부도 확률에 따른 모델의 임계값(threshold) 조정을 통해 임계값에 따른 승인 대출군의 구성 변화가 샤프 비율에 미치는 영향을 정량적으로 분석한다.

승인된 대출에 대해서는 현금 흐름을 바탕으로 계산한 내부수익률을 수익률로 사용하며, 거절된 대출에 대해서는 대출 발행 시기의 3년 만기, 5년 만기 미국 국채 수익률을 무위험 수익률로 가정하여 적용한다. 연구의 통계적 안정성과 일반화 가능성 확보를 위해 학습 및 검증 데이터 분리, 모델 학습, 샤프 비율 극대화를 위한 임계값 탐색 과정을 100번 반복한다.

제 3절 기대 효과

본 연구가 제시하는 샤프 비율 기반의 대출 투자 전략은 학문적, 실무적, 정책적 차원에서 다각적인 효과를 창출할 것으로 기대된다.

학문적 측면에서 본 연구는 P2P 대출 데이터를 활용한 신용 위험 예측 연구의 패러다임을 확장한다는 점에서 의의를 가진다. 예측 정확도 등의 단일 성과 지표 개선에 집중하기보다, 본 연구는 샤프 비율을 대출 투자 분석에 접목한다. 더불어 머신러닝 기반의 예측 모델과 금융 이론의 통합적 활용을 통해 실용적인 분석 프레임워크 개발에 기여한다.

실무적 측면에서는 금융 기관과 개인 투자자 모두에게 실질적 가치를 제공한다. 금융 기관의 경우, 본 연구에서 제시된 방법론을 바탕으로 단순 내부수익률 극대화 전략을 넘어선, 위험관리형 투자 전략을 효과적으로 수립할 수 있다. 특히 포트폴리오 차원에서의 샤프 비율 분석 결과는 기존의 리스크 관리 체계 및 신용 평가 정책과 유기적으로 연계하여 실무에 적용될 수 있다. 또한, 대형 금융기관만이 보유하던 리스크 모델링 기법을 본 연구가 제시함으로써, 개인 투자자들 역시 단편적 정보에 의존하지 않는 과학적인 투자 의사결정을 내리도록 돕는다.

정책적 측면에서는 P2P 금융 시장의 건전성 제고와 투자자 보호에 기여할 수 있는 시사점을 제공한다. P2P 대출 시장에서는 투자자들이 높은 명목 수익률 뿐만 아니라 위험 수준 역시 간과하지 않도록 하는 정책이 필요하다. 본 연구가 제시하는 위험대비 성과 중심의 평가 체계는 투명하고 균형 잡힌 P2P 대출 환경 조성에 기여할 수 있다. 금융 당국이 P2P 플랫폼의 건전성 평가, 투자자 보호를 위한 정책 설계, 시장 모니터링 체계 구축 과정에서 활용 가능한 정량적 기준과 분석 도구를 제공한다.

제 2 장 데이터 및 전처리

제 1 절 데이터 개요

본 연구는 미국 P2P 대출 플랫폼인 Lending Club이 공개한 2007년부터 2020년까지 약 175만 건의 대출 데이터를 활용한다. 데이터는 차입자의 인구통계학적, 재무적 특성, 대출 조건, 상환 결과라는 세가지 범주로 구분된다. 차입자 특성과 관련된 변수에는 신용 등급을 포함한 신용정보, 연체기록, 연소득, 부채액 및 비율, 고용기간 등이 포함된다. 대출 조건 관련 변수로는 대출금액, 이자율, 상환기간 등을 담고 있다. 마지막으로 상환 결과와 관련된 변수로는 상환 여부, 연체 및 부도 여부 등이 해당된다.

이와 같은 데이터 구조는 대출 승인 이전 시점에 차입자의 신용위험을 정교하게 평가할 수 있는 기초 근거를 제공한다. 동시에 투자자 입장에서는 포트폴리오의 기대수익과 위험을 관리하기 위한 핵심 자료가 된다. 특히 P2P 금융의 경우 차입자 간 이질성이 크기 때문에 방대한 데이터를 기반으로 변수 선택 및 리스크 관리 전략을 수립하는 것이 부도 확률 예측 모델 수립과 샤프 비율 극대화 연구의 출발점이라 할 수 있다.

제 2절 변수 선택

Lending Club이 공개한 데이터에는 약 150여 개의 변수가 존재한다. 이를 모두 모델 분석에 적용할 경우, 모델 복잡도 증가에 따른 훈련시간 및 메모리 사용량이 증가하게 된다. 공개된 데이터에는 부도 여부, 최종 납부 일자와 같이 대출 승인 이후 시점에서야 파악이 가능한 변수들이 다수 포함되어 있다. 이렇듯 내생성이 있는 변수들이 독립변수에 포함될 경우 미래 정보를 이용한 과거 예측으로 성능 및 성과가 과대평가될 위험이 있다. 따라서 본 연구에서는 금융공학적 이론을 바탕으로 변수의 설명력을 고려하는 동시에 투자 의사결정 시점에서의 활용 가능성을 기준으로 독립변수를 선정하였다. 구체적인 기준은 아래와 같다.

첫째, 대출 승인 이후 시점에서 알 수 있는 변수를 제외하였다. Lending Club이 승인한 대출건에 대한 이자율, 최종 승인 금액, 연체 발생 시의 연체 기록·총상환금액·채무 조정 계획 등이 내생성이 있는 변수들에 해당한다. 둘째, 차입자의 상환 능력과 관련된 변수들을 포함하였다. 차입자의 연 소득, 부채비율, 고용기간 등은 차입자의 경제적 능력을 평가할 수 있는 기초적인 정보이므로, 독립변수에 포함하였다. 특히 전통적 신용평가 이론에서 중요하게 다루어진 연체이력, 부도계좌수, 리볼빙 계좌수 등은 차입자의 리스크 프로파일링을 강화하기 위한 요소로 반영하였다. 셋째, 결측률이 70% 이상인 변수들은 정보량이 부족하다고 판단하여 분석에서 제외하였다. 특히 공동차입 여부와 관련된 변수는 2017년 이후 공동대출 축소라는 배경과 맞물려 결측률이 매우 크게 나타나 제외하였다. 한편, 대출

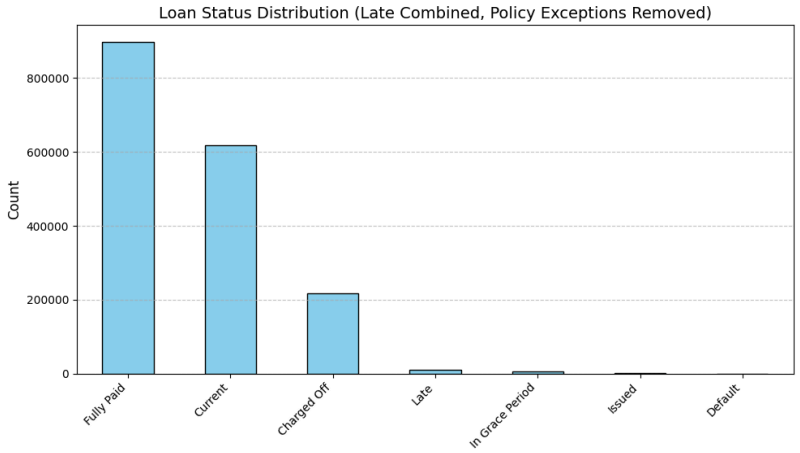
기간, 이자율, 대출 발행일, 최종 납부일, 등은 내생성이 존재한다고 판단하여 독립변수에서는 제외하였으나, 수익률 추정 및 국채수익률 연동을 위한 변수로서 사용하였다.

위와 같은 변수 선택을 통해, 단순히 데이터 차원을 축소하는 것을 넘어, 투자자 및 금융기관의 실무적 활용 가능성을 반영한 리스크 지표 체계 구축을 지향하고자 하였다.

제 3절 데이터 전처리

본 연구의 전처리 과정은 데이터 품질 확보와 모델의 해석력 제고를 위해 설계되었다. 주요 내용은 아래와 같다.

첫째, 대출 상태(loan_status) 변수를 활용하여 새로운 종속변수인 default 변수를 생성하였다. loan_status 변수는 Lending Club이 승인한 대출건의 상환 상태를 보여주는 변수로, 값의 분포는 [그림 1]과 같다. "Fully Paid"대출이 만기까지 정상 상환된 경우를 의미하며, "Current"는 현재 정상 상환중인 경우를 말한다. "Charged Off"와 "Default"의 경우 채무 불이행으로 상각 처리되거나 대출 계약 위반 상태로, 사실상 부도가 난 경우를 말한다. "Late"는 16~120일 동안 연체된 경우를, "In Grace Period"는 상환 유예기간에 있는 경우를 의미하며, "Issued"는 대출이 발행되었으나 아직 상환 기록을 확인할 수 없는 경우를 말한다. 부도 여부를 파악할 수 없는 "Current", "Late", "In Grace Period", "Issued" 값은 분석에서 제외하였다. 새로 정의된 default 변수에서는 "Fully Paid"인 경우를 0으로, "Charged Off"인 경우와 "Default"인 경우를 1로 매핑하였다.



[그림 1] "loan_status" 변수의 고유값 분포

둘째, 문자형 변수에 대한 수치화 및 one-hot encoding 과정을 수행하였다. 대출 기간을 의미하는 term 변수와 리볼빙 한도 대비 사용률을 나타내는 revol_util 변수의 경우, 문자열을 단순 제거하여 수치형 변수로 변환하였다. 근속연수를 나타내는 emp_length 변수의 경우, "n years" 형태의 고유값을 가지므로, 문자열을 제거하고 "< 1years" 값을 0.5로 처리하였다. Lending Club이 부여한 세부 신용 등급인 sub_grade 변수는 A1~G5의 35개 범주값을 가지고 있으므로, 1~35의 수치형 등급으로 변환하였다. 차입자의 거주지가 등록된 주(state) 정보를 담은 addr_state, 주택 소유 형태를

보여주는 home_ownership, 대출 목적을 나타내는 purpose, 소득 원천을 Lending Club이 검증했는지의 여부를 나타내는 verfication_status 등의 변수는 One-hot encoding 방식으로 더미화하였다.

셋째, 결측치 처리에서는 변수별 결측률을 기준으로 이원화 전략을 적용하였다. 먼저 결측률이 10% 이상인 변수에 대해서는 결측치를 0으로 대체한 후, 결측 여부를 나타내는 더미변수를 변수별로 추가하였다. 이로써 데이터 손실을 최소화하는 동시에, 결측 여부 자체가 가지는 정보와 신호를 모델이 학습할 수 있도록 하였다. 결측률이 10% 이하인 경우, 결측 여부 자체가 가지는 정보량이 크지 않다는 판단 하에, 메모리 사용량을 줄이기 위해 결측 여부를 나타내는 더미변수를 추가하지 않았다. 대신 변수의 분포와 기초 통계량을 고려하여 대체값을 설정하였다. 이 중 근속연수를 나타내는 emp_length와 신용한도 대비 사용 비율의 75% 초과 여부를 나타내는 percent_bc_75는 결측치를 0으로 채운 후 별도의 missing label을 부여하였다. 변수의 고유값 분포에서 0이 대부분을 차지하는 변수들은 결측치를 0으로 대체하였다. 그 외 변수들의 경우, 변수들의 비대칭 분포로 인해 평균과 중앙값에 큰 차이가 있다는 점을 고려하여 중앙값으로 결측치를 대체하였다.

넷째, 로그 변환을 적용하여 분포의 왜도가 큰 연속형 변수의 정규성을 확보하고자 하였다. 이때, 금융 데이터에서 이상치가 갖는 부도 예측 신호로서의 가치를 고려하여, 이상치를 제거하는 클리핑은 진행하지 않았다. 또한 과거 부도 기록, 세금 체납 기록 등 발생 빈도보다 발생 여부가 중요한 변수들은 바이너리 변환을 실시하여 해당 사건이 발생한 경우 1, 발생하지 않은 경우 0으로 변환하였다. 이로써 정보 손실을 최소화하는 동시에 변수 분포의 안정성을 균형 있게 달성하고자 하였다.

제 3 장 연구 방법

제 1 절 샤프 비율 기반 목적함수 설계

샤프 비율은 위험(표준편차) 대비 초과수익을 측정하는 대표적인 위험조정 수익률 지표이다. William F. Sharpe(1966)가 제안한 원래의 샤프 비율 공식은 다음과 같다.

$$S_p = \frac{E(R_p) - R_f}{\sigma_p}$$

여기서 $E(R_p)$ 는 포트폴리오의 기대수익률, R_f 는 무위험수익률, σ_p 는 수익률의 표준편차이다. 본 연구는 대출 현금흐름 기반 내부수익률을 활용함으로써 샤프 비율을 다음과 같이 재정의했다. 다음의 정의는 1994년 Sharpe의 개정 공식과 유사한 접근법이다.

$$Sharpe\ Ratio = \frac{\overline{R - R_f}}{\sigma\left(R - R_f\right)}$$

여기서 R 은 대출별 내부수익률, R_f 는 무위험수익률, σ 는 초과수익률의 표준편차이다. 개별 대출 수준에서 내부수익률을 산출한 후 초과수익률을 계산하고 전체 분포를 집계하는 방식을 채택하였다. 즉, 개별 차입자의 상환 여부와 상환 패턴을 반영한 내부수익률을 사용함으로써, 단순 대출 금리보다 더 정교한 위험조정 수익 평가가 가능하다. 또한, 변동성은 포트폴리오 수준에서의 분산을 반영하기 위해 표본 표준편차(ddof=1)로 계산했다. 무위험 수익률은 FRED API를 활용하여 대출 발행 시점에 대응하는 3년/5년 만기 미국 국채 수익률로 설정했다.

샤프 비율을 목적함수로 활용함으로써, 단순히 예측 정확도를 높이는 것에 그치지 않고 수익성과 변동성을 동시에 고려한 모델 선택 전략을 구현할 수 있다. 이는 P2P 투자 환경에서 투자자가 실제 의사결정 과정에서 직면하는 '위험 대비 수익 극대화' 문제와 연결된다.

제 2 절 샤프 비율 산출 구조

본 연구는 다음의 절차를 통해 샤프 비율을 산출하였다. 먼저 머신러닝 기반 예측 모델을 통해 산출된 개별 부도 확률이 임계값 이하인 대출만 승인하여 투자 대상 대출군을 선별한다. 이는 부도 가능성이 낮은 우량 대출군에만 투자하는 전략을 반영한다.

선별된 대출군에 대해서는 상환 계획에 따른 개별 현금흐름을 시뮬레이션한다. 투자자 입장에서 대출 승인 시점에는 Lending Club이 승인한 대출 금액(funded_amnt)만큼의 현금이 유출되며, 이후에는 매 달 대출자가 정기적으로 납부하는 금액(installment) 만큼의 현금이 유입된다고 가정한다. 이때, 부도 여부에 따라 현금 흐름 패턴이 달라진다. 정상 상환건의 경우 만기까지 매달 월 납입액(installment) 만큼의 현금이 유입되는 반면, 부도 대출 건의 경우 마지막 납입 시점(last_pymnt_d)까지만 (insalllment) 만큼의 현금이 유입되고 이후 대출 기관의 추심 절차가 진행된다. 추심 과정에서는 부도가 발생한 시점 기준 다음 달에 추심액(recoveries)에서 추심수수료(collection_recovery_fee)를 차감한 순추심액이 일시금 형태로 유입된다.

이렇게 구성된 개별 현금흐름으로부터 내부수익률을 계산하고, 이를 연환산 수익률로 변환한다. 승인된 대출 건의 경우, 계산된 내부수익률에서 동기간 무위험 수익률에 해당하는 미국 3년물 및 5년물의 국채수익률을 차감하여 초과수익률을 도출한다. 거절된 대출건의 경우 수익률은 무위험 수익률로 가정한다. 이렇듯 초과수익률을 집계하여 평균과 표준편차를 도출하고, 최종적으로 샤프 비율을 계산한다.

제 3 절 모델링 전략

본 연구에서는 다양한 머신러닝 알고리즘 중 트리 기반 앙상블 모델인 Random Forest, XGBoost, LightGBM을 중심으로 분석을 수행하였다. 이러한 모델 선택은 Lending Club 데이터의 고유 특성에 기반한다. 첫째, 약 175만 건에 달하는 대규모 관측치를 효율적으로 처리할 수 있으며, 둘째, 범주형 변수와 연속형 변수가 혼재하는 데이터 구조에 강건하고, 셋째, 변수 간 복잡한 비선형 관계 및 상호작용 효과를 적절히 포착한다는 장점을 가진다.

모델링 절차는 다음과 같이 진행하였다. 먼저 분석 대상이 되는 전체 데이터를 학습 데이터, 검증 데이터, 평가 데이터를 6:2:2의 비율로 분할하였다. 이후 각 모델에서 산출된 부도 예측확률을 기반으로 임계값을 조정하여 대출 승인군과 비승인군을 구분하였다. 임계값의 경우, 단순히 AUC 극대화가 아닌 샤프 비율 극대화를 목표로 탐색함으로써 예측 정확도보다 실제 투자성과 최적화에 중점을 두고자 하였다. 승인군이 확정되면 해당 대출의 현금흐름을 시뮬레이션하여 내부수익률을 계산하고, 이를 집계하여 샤프 비율로 환산하였다. 마지막으로, 동일 과정을 100회 반복 실행하면서 단일 시행에서 발생할 수 있는 우연성을 배제하고, 모델 기반 투자 전략의 재현 가능성과 신뢰성을 제고하고자 하였다.

제 4 절 하이퍼파라미터 탐색

각 모델의 예측 성능을 극대화하기 위하여, 체계적인 하이퍼파라미터 탐색을 수행하였다. Random Forest의 경우 트리 개수(n_estimators), 최대 깊이(max_depth), 최소 분할 샘플 수(min_samples_split), 최소 리프 노드 샘플 수(min_samples_leaf) 등의 주요 하이퍼 파라미터를 조정하였다. XGBoost와 LightGBM의 경우 학습률(learning_rate), 부스팅 반복 횟수(n_estimators), 정규화 계수(reg_alpha, reg_lambda), 최대 깊이(max_depth), 서브 샘플링 비율(subsample) 등을 탐색하였다.

이때, 계산 자원의 제약을 고려하여 샤프 비율 최적화를 목적함수로 둔 전수탐색(Grid Search) 대신 RandomizedSearchCV 방식을 채택하였다. 튜닝의 평가 지표로는 AUC(Area Under Curve)를 사용하였다. 이는 분류 성능, 특히 재현율(recall) 향상이 부도 예측의 정확성을 높이고, 궁극적으로는 샤프 비율 극대화에 기여한다는 이론적 가정에 근거한 것이다.

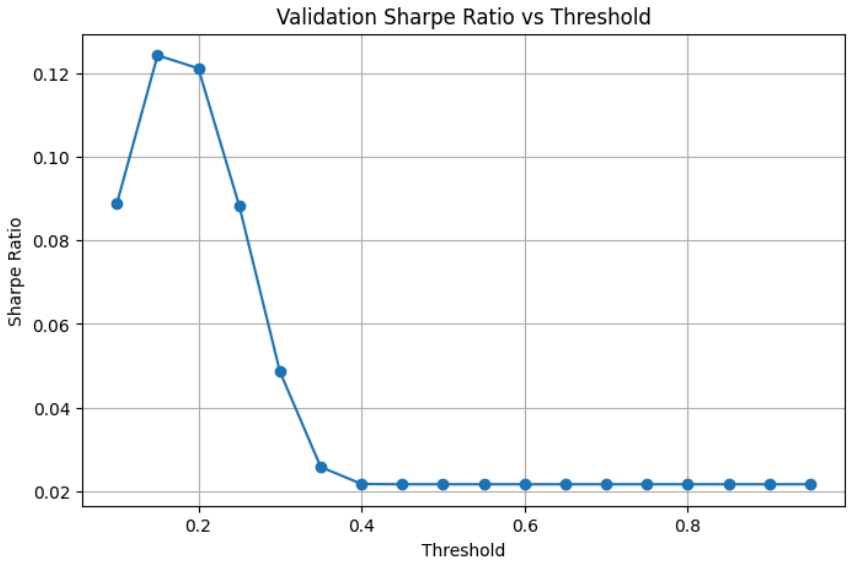
제 4 장 연구 결과

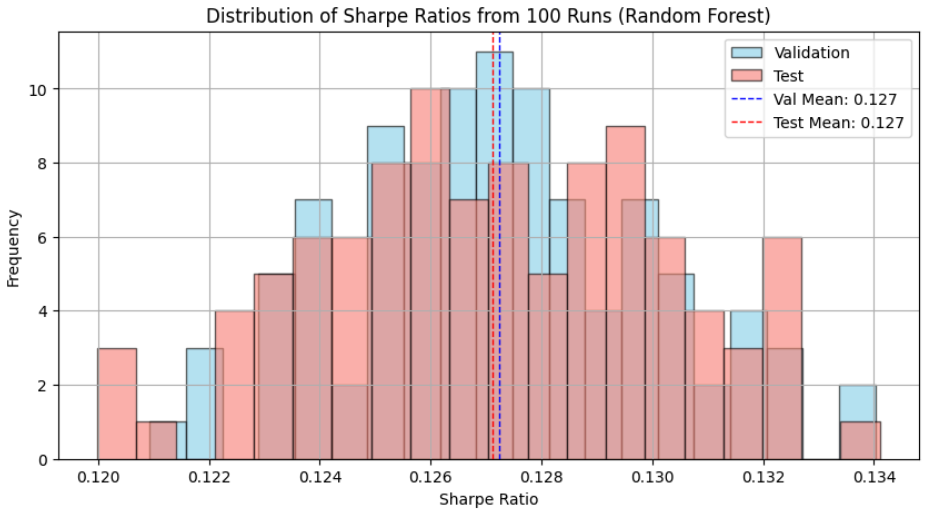
제 1 절 임계값(Threshold) 분석

본 연구에서는 각 머신러닝 모델별로 예측 확률에 따른 승인 여부의 임계값를 조정하여 샤프 비율 성과를 측정하였다. 분석 결과, 세 모델 모두 0.15 수준에서 최적 임계치가 도출되었다(그림 2-4 참조). 이에 따른 각 모델의 최적 Validation / Test Sharpe Ratio는 Random Forest 0.127 / 0.134, XGBoost 0.107 / 0.115, LightGBM 0.110 / 0.115로 나타났다.

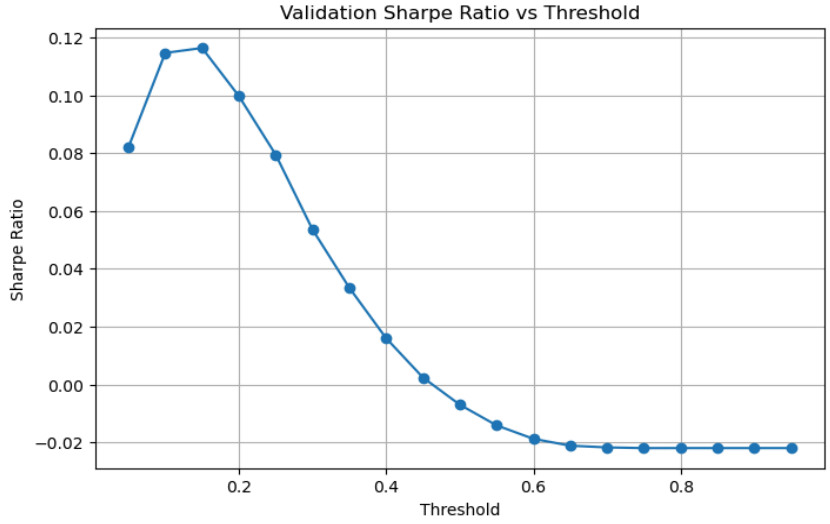
그림에서 확인할 수 있듯이, 임계값이 지나치게 낮거나 높은 경우에는 승인 대출의 분산 효과가 충분히 발휘되지 못해 샤프 비율이 급격히 하락하는 양상이 관찰되었다. 반면, 0.15 수준에서는 Validation과 Test 모두에서 샤프 비율이 상대적으로 안정적으로 유지되었으며, 100회 반복 실험 결과의 분포 또한 평균값을 중심으로 좁은 범위에 수렴하는 모습을 보였다. 이는 보수적인 승인 기준을 설정하는 것이 위험 대비 수익률을 극대화하는 데 기여함을 실증적으로 뒷받침한다.

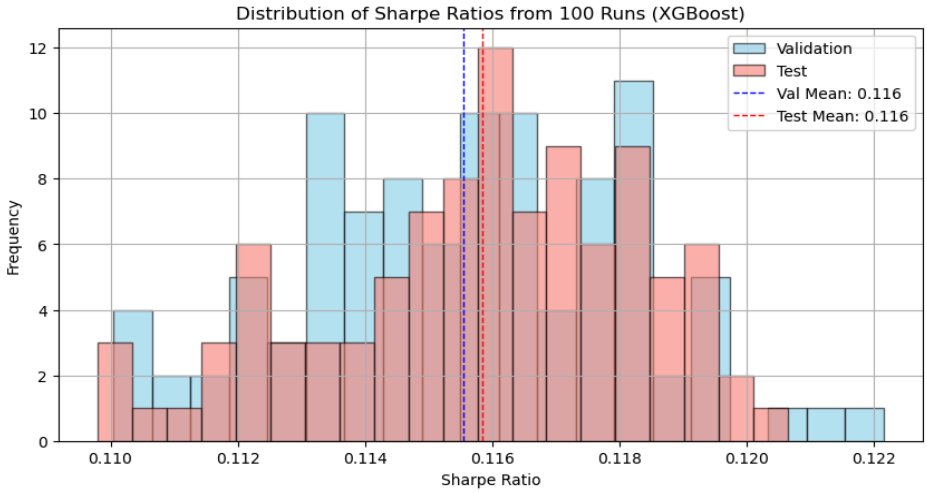
또한 세 모델이 공통적으로 최적 임계치(0.15)를 도출하였다는 사실은, 이후 Out-of-sample 성과 비교가 동등한 조건하에서 수행될 수 있음을 의미한다. 제 2절에서는 동일한 임계값 하에서 Random Forest, XGBoost, LightGBM의 상대적 성과 차이를 심층적으로 분석한다.



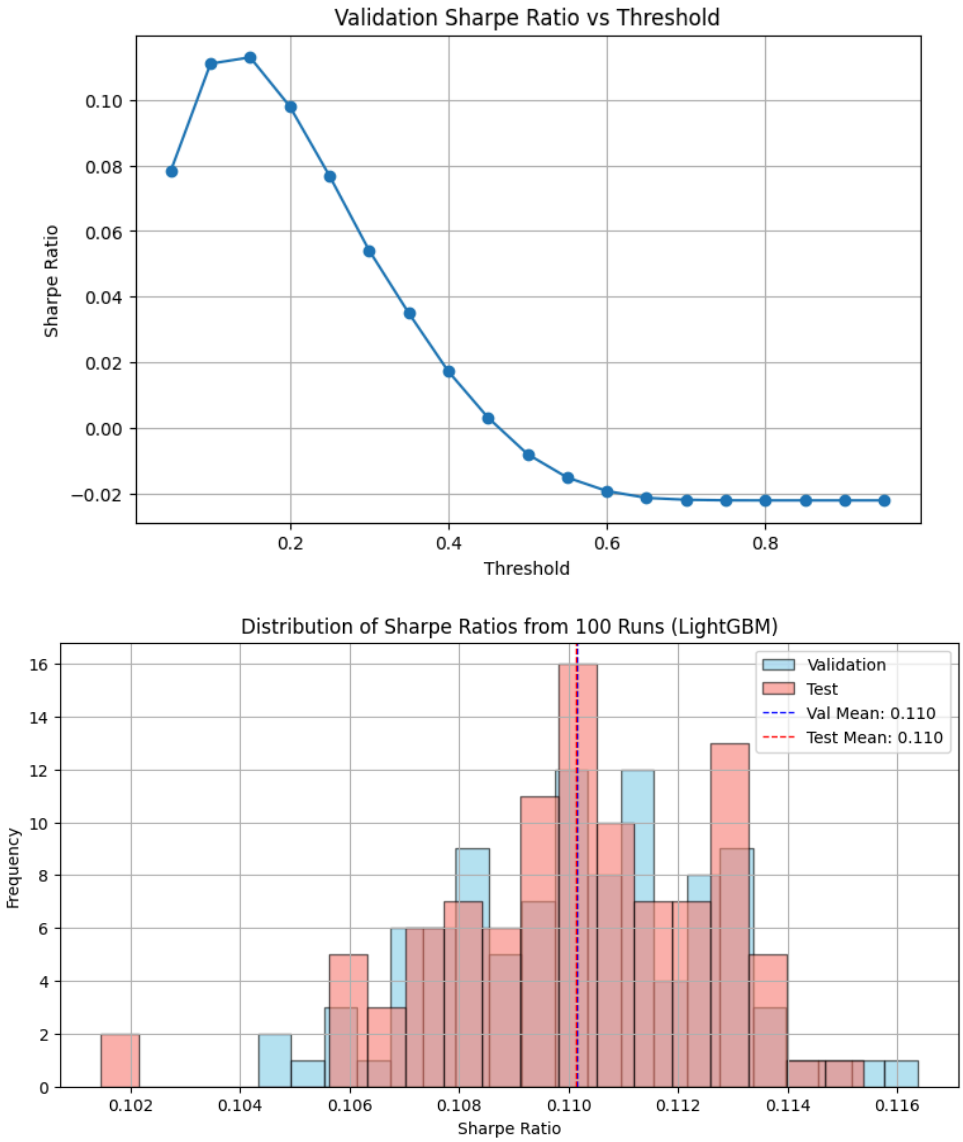


[그림 2] Random Forest 기반 임계값별 Validation Sharpe Ratio 및 반복 실험 결과 분포





[그림 3] XGBoost 기반 임계값별 Validation Sharpe Ratio 및 반복 실험 결과 분포



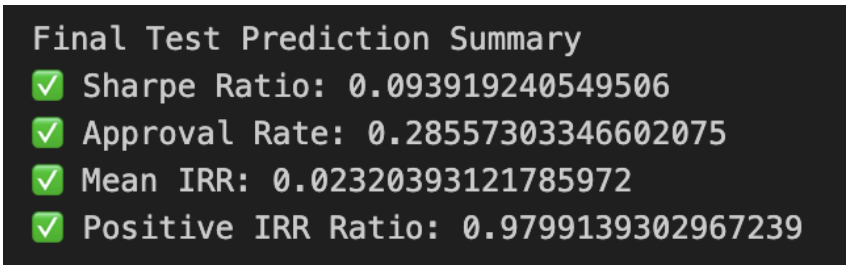
[그림 4] LightGBM 기반 임계값별 Validation Sharpe Ratio 및 반복 실험 결과 분포

제 2 절 모델별 성과 비교

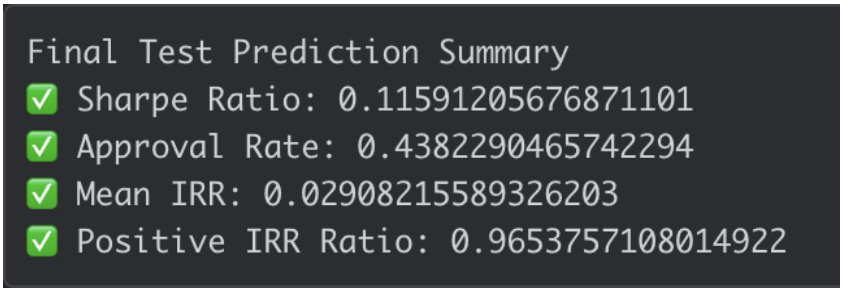
세 가지 모델의 최적 임계치(0.15)에서 도출된 Out-of-sample 성과를 비교한 결과, Boosting 계열(XGBoost, LightGBM)이 Random Forest 대비 샤프 비율 측면에서 일관되게 우수한 성과를 나타냈다(그림 5-7). 구체적으로 Random Forest의 최종 테스트 샤프 비율은 0.094로, XGBoost(0.116)와 LightGBM(0.111)에 비해 낮게 나타났다.

Random Forest는 승인 기준을 엄격히 적용하여 승인률이 약 29% 수준에 불과했으며, 그 결과 승인된 대출의 대부분이 양의 내부수익률을 기록하여 Positive IRR Ratio가 97.9%로 가장 높게 도출되었다.

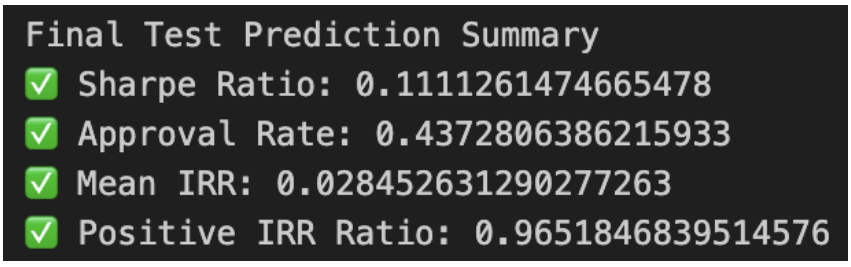
반면, XGBoost와 LightGBM은 승인률이 약 44% 수준으로 상대적으로 더 많은 대출을 승인하였다. 이에 따라 Positive IRR Ratio는 약 9.65%로 다소 낮아졌으나, 승인 표본 확대에 따른 분산 효과가 위험을 상쇄하며 결과적으로 샤프 비율이 개선되었다. XGBoost는 샤프 비율 0.116으로 세 모델 중 가장 높은 성과를 보였으며, LightGBM 또한 0.111을 기록하며 XGBoost에 근소하게 차이가 났으나 Random Forest보다는 우수한 성과를 나타냈다. 이러한 결과는 Boosting 계열 모델이 샘플의 다양성과 분산 효과를 보다 효과적으로 활용하여 Random Forest 대비 샤프비율 극대화에 구조적으로 유리함을 시사한다.



[그림 5] Random Forest 모델의 최종 테스트 예측 요약



[그림 6] XGBoost 모델의 최종 테스트 예측 요약



[그림 7] LightGBM 모델의 최종 테스트 예측 요약

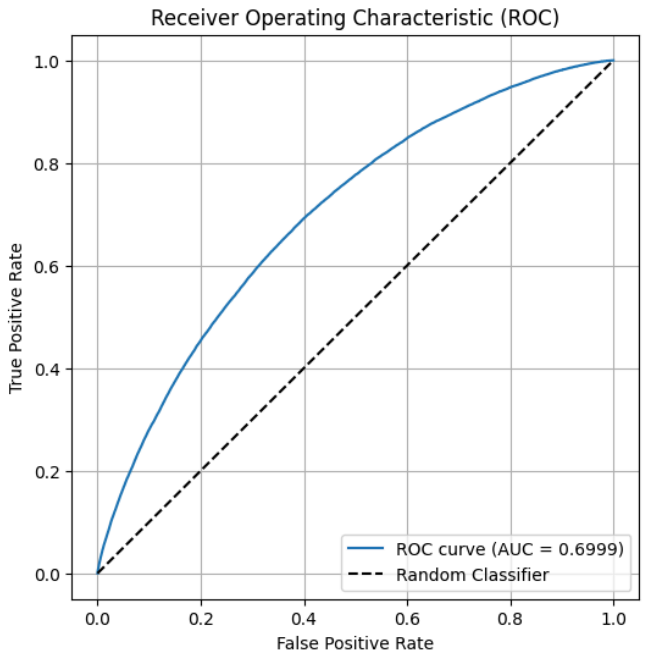
제 3 절 보조적 지표 분석

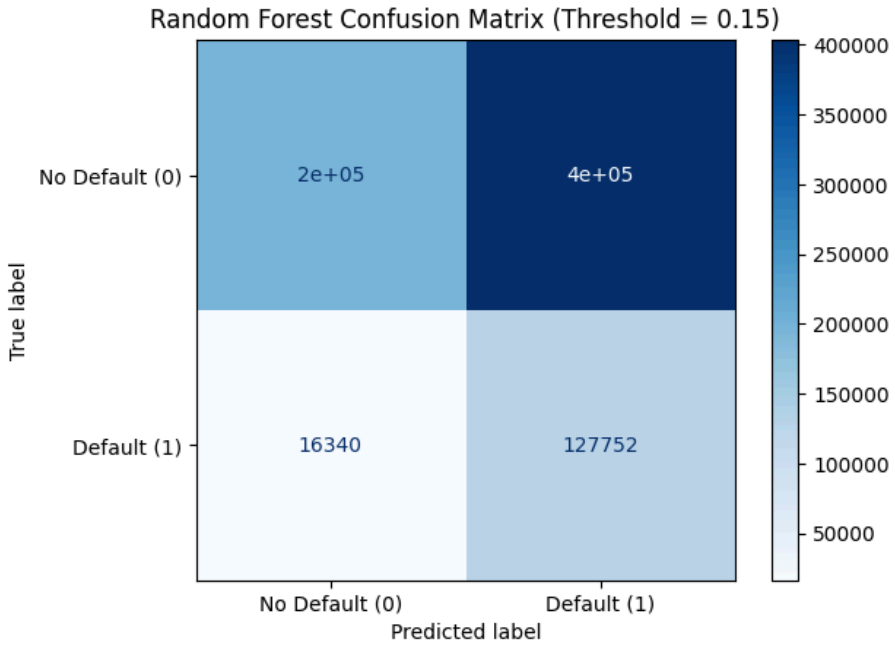
샤프 비율을 중심으로 한 본 연구의 목적을 보완하기 위하여, 모델의 분류 성능을 ROC 곡선(AUC)과 혼동 행렬(Confusion Matrix)을 통해 추가적으로 점검하였다.

AUC 값은 Random Forest 0.700, XGBoost 0.720, LightGBM 0.719로 나타나 제안된 모델들이 대출 상환 여부를 예측하는 데 있어 비교적 높은 정확도를 보유함을 확인하였다(그림 8-10 참조). 그러나 이러한 예측 성능이 반드시 높은 샤프 비율로 이어지지는 않았으며, 이는 단순한 예측 성능과 위험조정 성과 간의 괴리를 보여주는 중요한 결과라 할 수 있다.

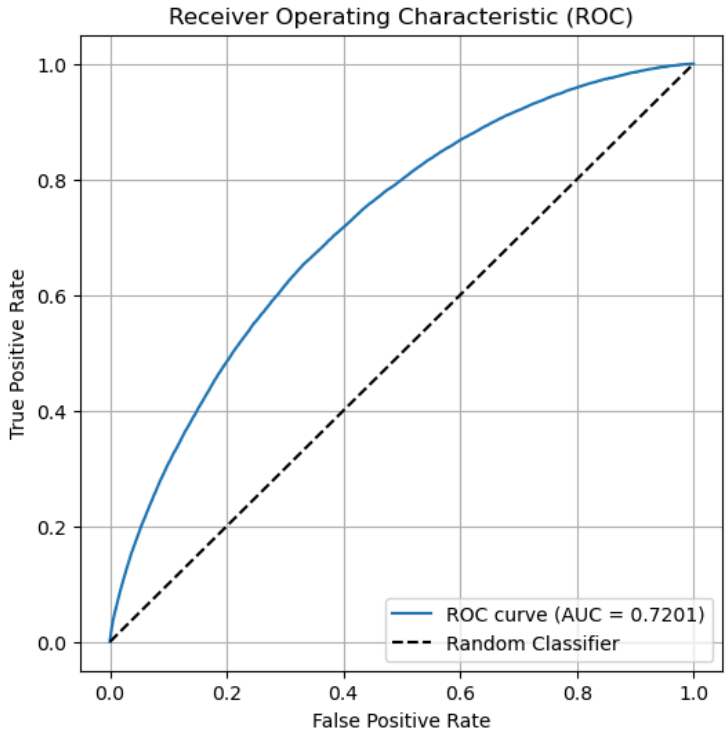
혼동 행렬 분석 결과, 정밀도와 재현율의 균형은 모델별로 상이하게 나타났다. Random Forest는 부도 사례 탐지 비율이 상대적으로 높아 재현율 측면에서 강점을 보였으나, 정상 대출을 부도로 잘못 분류하는 비율이 증가함에 따라 정밀도는 낮게 도출되었다. 반면, XGBoost와 LightGBM은 일정 수준의 부도 탐지율을 유지하면서도 정상 대출 식별 성능이 개선되어 정밀도과 재현율 간 균형이 더 적절하게 확보되었다.

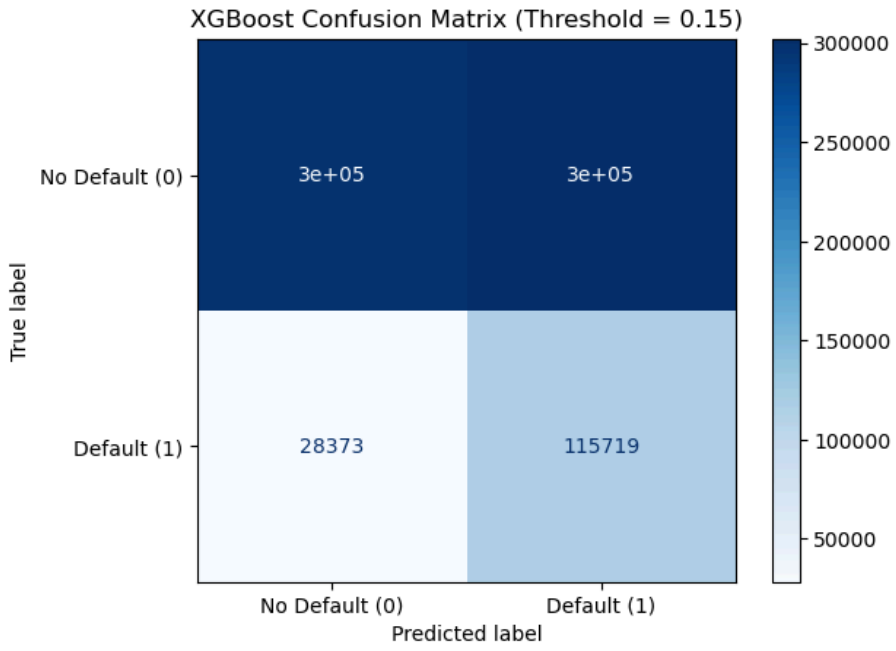
이러한 결과는 본 연구의 핵심 성과지표인 샤프 비율와 직접적인 연관성은 낮지만, 투자자의 위험 선호도와 전략적 의사결정을 보완하는 참고 지표로 활용될 수 있다. 즉, 부도 탐지를 극대화하여 손실 위험을 최소화하는 전략(재현율 중시)과 정상 대출을 보다 정확히 선별하여 안정적 수익을 추구하는 전략(정밀도 중시) 사이에서의 선택은 투자자 관점의 전략적 고려사항으로 제시될 수 있다.



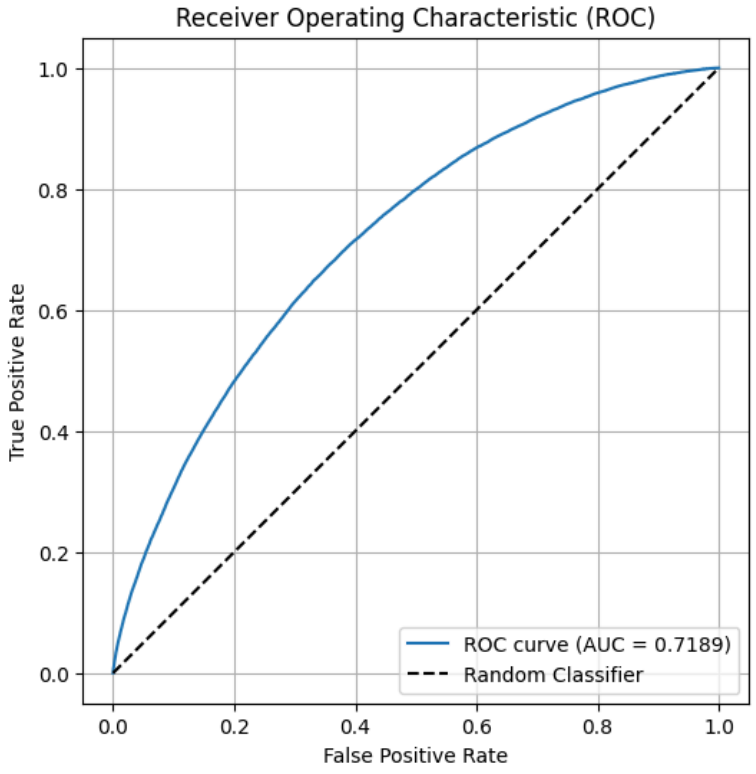


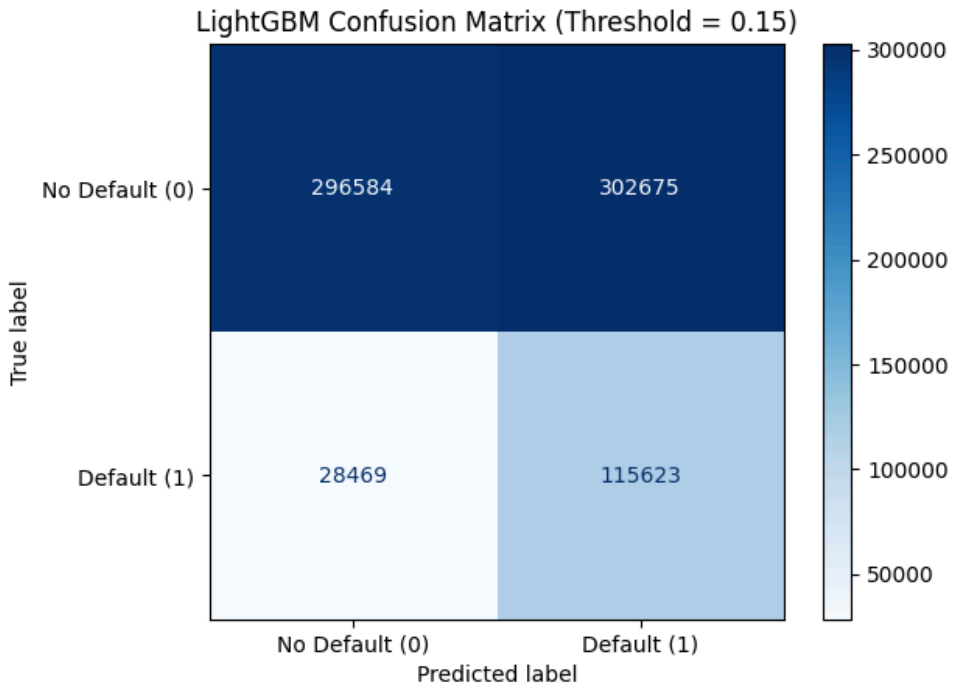
[그림 8] Random Forest 모델의 ROC 곡선 및 혼동 행렬 비교





[그림 9] XGBoost 모델의 ROC 곡선 및 혼동 행렬 비교





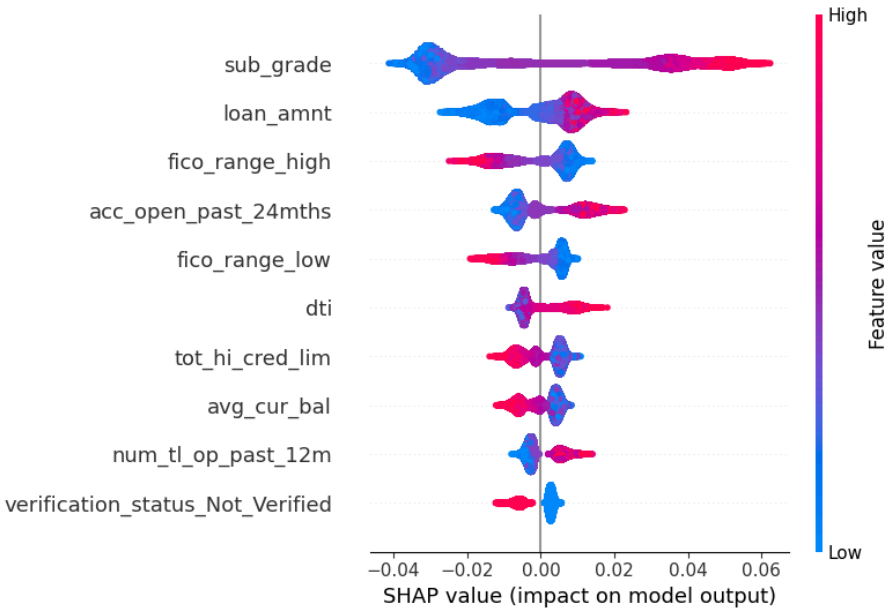
[그림 10] LightGBM 모델의 ROC 곡선 및 혼동 행렬 비교

제 4 절 특성 중요도 (SHAP)

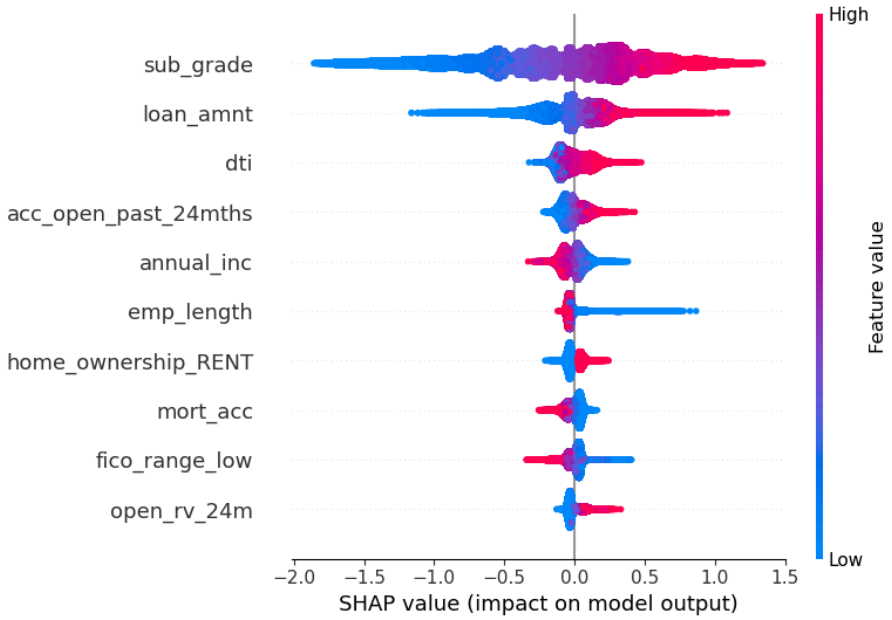
모델 해석 가능성을 높이기 위해 SHAP(Shapley Additive Explanations) 값을 산출하였다. 분석 결과, 세 모델 모두 신용등급(sub_grade)과 대출금액(loan_amnt), 부채소득비율(dti), 그리고 24개월 내 계좌 개설 수(acc_open_past_24mths) 변수가 공통적으로 중요한 요인으로 나타나 샤프 비율 성과에 핵심적으로 기여함을 확인하였다. 신용등급 변수는 대출자의 상환 가능성을 직접적으로 반영할 뿐만 아니라, 임계값 설정 과정에서 승인 건수와 샤프 비율 간의 균형을 결정하는 주요 요인으로 작용하였다. 이는 금융기관 실무에서 변수 해석 가능성을 기반으로 한 리스크 관리의 중요성을 뒷받침한다.

세부적으로는 모델 간 차이가 존재하였다. Random Forest는 fico_range_high, fico_range_low 등 전통적인 신용점수 변수와 tot_hi_cred_lim, avg_cur_bal, verification_status_Not Verified와 같은 전통적 신용검증 지표에 크게 의존하는 경향을 보였다. 반면, Boosting 계열(XGBoost, LightGBM)은 전통적 신용점수에 대한 의존도가 낮고, 대신 dti, acc_open_past_24mths 등 단기적 신용활동 변수인 부채관리 및 최근 신용활동 변수를 보다 중시하였다. 특히 XGBoost는 소득(annual_inc), 고용 안정성(emp_length), 주거 형태(home_ownership_RENT), 장기 채무(mort_acc) 관련 변수를 중요하게 반영하는 특징이 있었다. LightGBM은 추가적으로 num_actv_rev_tl과 같은 리볼빙 계좌 관련 단기 신용 변수까지 함께 고려하는 경향을 보였다.

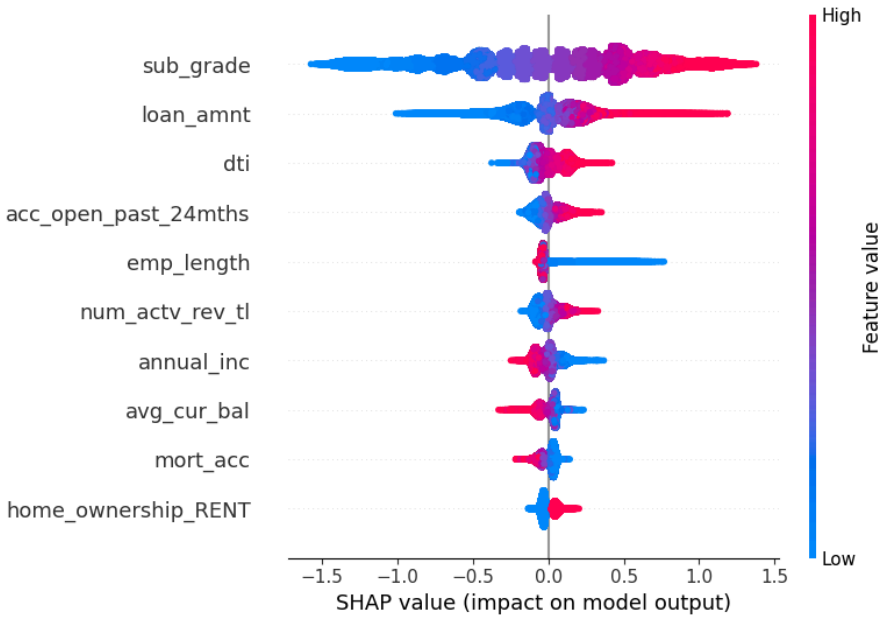
이러한 결과는 금융 실무 측면에서 중요한 시사점을 제공한다. Boosting 계열 모델은 단기적 지표(dti, acc_open_past_24mths)와 장기적 특성(emp_length, home_ownership_RENT, mort_acc)을 동시에 반영함으로써 차입자의 상환능력을 보다 입체적으로 평가할 수 있다. 이는 기존 신용등급 중심 평가 방식을 보완하여 리스크 관리 및 투자 전략 고도화에 기여할 수 있음을 시사한다.



[그림 11] Random Forest 모델의 SHAP 기반 특성 중요도



[그림 12] XGBoost 모델의 SHAP 기반 특성 중요도



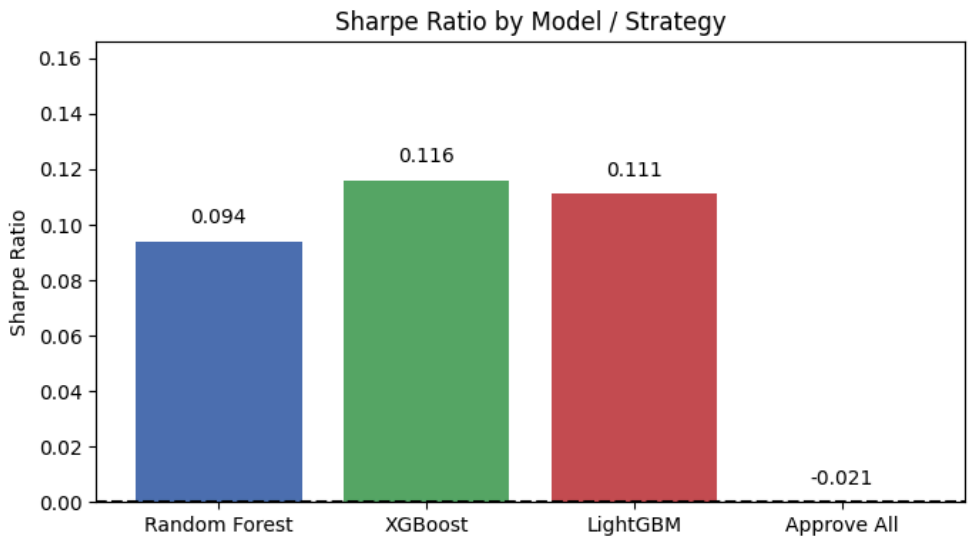
[그림 13] LightGBM 모델의 SHAP 기반 특성 중요도

제 5 절 벤치마크(Benchmark)와의 비교

본 연구의 머신러닝 기반 전략을 단순 벤치마크 전략과 비교하였다. 벤치마크로 설정한 단순 전략(Approve All)의 경우 샤프 비율이 −0.021로 나타나, 모든 대출을 승인하는 방식은 위험 대비 성과가 낮았다(그림 14). 이는 대출 승인 과정에서 위험 차등화를 고려하지 않을 경우, 안정적인 위험조정 성과를 달성하기 어렵다는 점을 보여준다.

반면, 머신러닝 기반 전략은 예측 확률을 활용하여 승인 대상을 선별함으로써 모든 모델에서 샤프 비율이 벤치마크 대비 유의하게 개선되었다. 특히 XGBoost(0.116)와 LightGBM(0.111)은 Random Forest(0.094)를 상회하며 샤프 비율 측면에서 벤치마크보다 개선된 성과를 기록하였다.

이와 같은 결과는 제 1절에서 확인한 공통 최적 임계값(0.15)와 제 2절*제 3절에서의 성과 비교 분석, 그리고 제 4절의 변수 중요도 해석과 일관된 흐름을 보인다. 나아가 Validation과 Test set 모두에서 유사한 결과가 관찰됨에 따라 본 연구의 머신러닝 기반 전략이 단순 과적합을 넘어 일정 수준의 일반화 가능성을 확보했음을 시사한다. 따라서 모델 적용 가능성의 당위성을 보여준다고 할 수 있다.



[그림 14] 모델별 샤프 비율과 벤치마크 비교

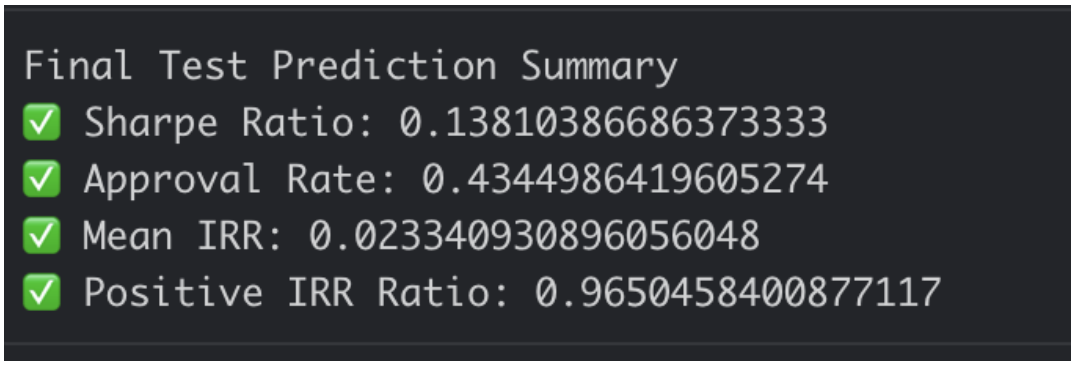
제 6 절 샤프 비율 계산 보완

제 2절에서 XGBoost의 성능이 가장 좋음을 확인했다. 본 절에서는 무위험수익률 계산시 고려하지 않았던 조기상환과 기타 고려해볼 사항에 대해 논의하고자 한다. 벤치마크의 샤프비율은 모델을 적용하지 않았을 때의 모든 사후적인 현금흐름을 고려하여 계산된 값이다. 따라서 비교 목적 상 모델을 적용하여 벤치마크와 비교하는 경우, 모델의 부도 확률 예측이 임계값을 넘지 않아 부도라고 예측되지 않았지만 실제 부도였던 경우의 현금흐름을 고려하는 것이 적절하다. 또한, 제대로 실행된 대출도 36개월, 60개월로 정확히 종료되는 것이 아니라 조기상환되는 경우가 있다. 이 경우 샤프비율 계산에 들어가는 무위험수익률을 36개월, 60개월로 정한다면 해당 무위험수익률은 과대 평가될 가능성이 있다. 따라서 이를 조정해주는 방법이 필요하다. 본 절에서는 각 대출 건의 만기를 실질적으로 고려하는 듀레이션(Duration) 개념을 사용하여 무위험수익률 계산에 필요한 기간을 계산한다. 본 연구에서 각 대출 건은 원리금균등상환을 가정하였기 때문에 듀레이션이 본래의 만기보다 짧다. 무이표채가 아닌 대출상품의 경우 주식과는 달리 이자가 고정적으로 나오기 때문에 원금을 회수하는 기간이 정해진 만기보다 짧아진다. 이를 고려하여 듀레이션을 실질적인 만기라고 볼 수 있다. 듀레이션은 다음과 같이 정의되는 맥컬리 듀레이션을 사용한다.

$$DM = \frac{\sum_{t=1}^{T} \left(\frac{t \cdot CFt}{(1+y)^t} \right)}{\sum_{t=1}^{T} \left(\frac{CFt}{(1+y)^t} \right)}$$

- CFt : 시점 t에서의 현금흐름 (이자 + 원금상환)
- y : 만기수익률(Yield to Maturity, YTM)
- T : 만기까지의 기간

즉, 듀레이션은 현금흐름의 현재가치 가중평균으로 모든 대출 건별에 대해 듀레이션을 계산할 수 있다. 각 듀레이션 기간을 1월물, 3월물, 6월물, 1년물, 3년물, 5년물 미국 국채 범위에서 보간한 값과 대응해 무위험수익률을 계산하여 샤프 비율을 계산했다. 모델의 계산결과 나오는 부도확률이 임계값을 넘어서 대출이 이루어지지 않는 경우는 대출약정에 따라 36개월, 60개월 무위험수익률을 배정했다. 이는 해당 국채가 무이표채라고 가정하고 듀레이션과 만기가 같다는 가정하에 계산한 것이다. 계산결과는 다음과 같다.



[**그림 15**] Best Model(XGBoost)의 조기상환 적용 최종 테스트 결과

듀레이션과 조기상환을 고려한 벤치마크의 샤프 비율은 0.007299였다. 기존과 달리 비율의 상승이 있는 이유는 대출 기간을 가중평균 실질만기로 설정하면서 그에 해당하는 무위험국채의 만기가 줄어들었기 때문이다. 비교목적상이라면 어느 샤프 비율을 사용해도 괜찮겠지만, 더 타당한 접근을 위한 시도로 조기상환과 듀레이션을 고려한 샤프 비율을 계산했다. 제 2절과 본 절에서의 결과를 바탕으로 모델을 적용하는 것이 적용하지 않는 것보다 샤프 비율 면에서 성과가 좋음을 알 수 있다. 향후 모델 적용 시에는 거시 경제 변수, 소득, 정치 변수, 문화적 변수, 개인 변수 등을 고려한 대출 상환여부 확률분포를 보여줄 수 있는 다른 모델이 있다는 가정 하에 기대수익률과 표준편차를 계산하여 본 모델에 넣어 기대 샤프 비율을 계산할 수 있다. 이를 통해, 사업의 확장과 철수, 마케팅에 직간접적으로 도움을 줄 수 있을 것이다.

제 5 장 결론 및 시사점

제 1 절 핵심 결과 요약

본 연구는 Lending Club 데이터를 활용하여 샤프 비율을 극대화하는 머신러닝 기반 대출 승인 모델을 구축하였다. 분석 결과, XGBoost가 위험조정 성과에서 우위를 보였다. 이는 단순한 분류 성능 지표(AUC)가 높다고 해서 투자자의 관점에서 반드시 바람직한 결과로 이어지지 않음을 보여주며, 위험조정 수익률을 고려한 의사결정의 중요성을 강조한다.

제 2 절 인사이트

본 연구에서는 샤프 비율을 극대화하기 위해 다각도의 접근을 시도하였다. 임계값을 0부터 1까지 0.05 단위로 체계적으로 조정하며 각 구간에서의 샤프 비율 변화를 분석함으로써 최적의 임계값을 도출하였다. 샤프 비율이 0.5 이하로 나타난 점은 데이터와 시장 구조의 제약으로 인해 고위험, 고수익 투자 특성이 반영된 결과로 해석할 수 있다. 즉, 부도 대출 건으로 인한 손실에 조금 더 비중을 두어 전략을 수립하였으며, P2P 금융 시장의 불안요소를 줄이는 방향으로 모델링한 결과다. 이는 단순 통계적 정확도보다는 실제 투자 상황을 고려해 현실적인 전략을 취했다는 점에서 의미가 있다.

샤프 비율이 1이하로 나타난 점 역시 고위험, 고수익 투자 특성이 반영된 결과로 해석할 수 있다. 그러나 샤프 비율을 핵심 성과 지표로 도입함으로써 투자자는 절대적인 수익률 극대화가 아닌 위험 대비 성과 최적화 전략을 수립할 수 있음을 확인하였다. 이는 P2P 금융시장과 같이 부도 위험이 내재된 환경에서 특히 중요한 의미를 지닌다.

제 3 절 한계점 및 향후 연구

본 연구의 의의는 예측 정확도보다는 샤프 비율의 극대화에 초점을 두어 실질적인 투자 전략 수립에 기여했다는 데에 있다. 하지만 모델의 예측 정확도 역시 배제할 수는 없는 요소이며, 혼동 행렬에 대한 추가적인 분석을 통해 모델의 예측 성능을 보완하는 동시에 샤프 비율 극대화라는 목표 역시 효과적으로 달성할 방법을 모색할 수 있다. 더불어 임계값 탐색 과정에서 구간을 0.05 단위로 설정했으나, 구간 범위를 더욱 조밀하게 설정할 경우 더욱 정밀한 방식으로 샤프 비율 극대화를 달성할 수 있다.

(1) 혼동 행렬을 통한 최적화 전략

제4장 제3절의 혼동 행렬에 따르면, 본 연구에서 사용된 모든 세 모델에서 정상 대출을 부도로 예측하는 경우가 상당히 많았다. 부도 대출을 정상 대출로 예측해 승인하는 경우의 손실이 정상

대출을 부도 대출로 예측해 거절하는 경우에 발생하는 손실보다 Lending Club과 투자자 모두에게 부정적인 영향을 줄 수 있다는 점에서, 보수적인 예측 전략은 타당하다고 할 수 있다. 즉, 현재의 모델에서는 부도 대출을 부도 대출로 적절히 예측하여 높은 재현율을 보여준다.

하지만 정상 대출을 부도로 예측하는 경우, 추가적인 대출 승인을 통해 얻을 수 있는 잠재적 수익을 잃게 된다. 그러므로 부도 대출에 대한 예측 정확도를 유지하는 동시에 정상 상환이 가능한 대출군을 더욱 많이 포함하면서 초과 수익률을 개선하고, 궁극적으로는 샤프 비율을 개선하기 위한 전략 역시 고려할 수 있다. 이를 위해서는 모델링 과정에서 sklearn 패키지의 class_weight 파라미터를 사용해, 정상 대출을 부도 대출로 예측하는 경우에 대한 패널티를 부과하는 방식 등을 적용할 수 있다.

(2) 설명 변수 보완

예측력을 개선함으로써 샤프 비율을 극대화하기 위한 또 다른 보완책으로 설명 변수를 언급해볼 수 있다. 정상 대출을 부도 대출로 잘못 예측한 경우를 줄이기 위해서는 조금 더 많은 정보가 필요하다. 그 중 하나가 차입자의 주거 주 정보(addr_state)로만 들어있는 주거 지역에 대한 설명을 보완하는 것이다. 본 연구에서는 우편번호(zip_code) 변수를 분석에서 제외하였으나, 현재 공개되어 있는 Zillow 주택 가격 정보를 해당 변수를 바탕으로 매칭함으로써 주거 상태에 대한 조금 더 세부적인 정보를 보완할 수 있다. 이로써 예측력 개선과 샤프 비율 극대화라는 목표를 더욱 효과적으로 달성할 수 있을 것이다.

(3) 임계값 구간 탐색 세분화

본 연구는 임계값 구간 탐색에 있어, 0~1 사이의 범위를 0.05 단위로 분할하여 진행했다. 본 연구에서 사용한 세 가지 모델을 적용한 결과, 임계값이 모두 0.15로 도출된 것은 0.05의 탐색 단위가 다소 느슨했다는 결과로도 해석 가능하다. 때문에 이를 0.01단위로 조절해 탐색한다면, 모델별로 더욱 세밀한 임계값을 찾아 진정한 의미의 샤프 비율 극대화를 달성할 수 있을 것이다.

정리하자면, 향후 연구에서는 (1) 정상 대출을 부도 대출로 예측하는 경우를 줄임으로써 높은 재현율 수준 유지와 정밀도 개선, (2) 임계값 탐색 단위의 세분화, (3) 설명 변수 확대를 고려할 수 있다. 이는 샤프 비율에 직간접적으로 관련된 요소들로, 투자 전략을 더욱 정교하게 설계할 수 있는 가능성을 제시한다. 이를 통해 위험과 기회 간의 균형을 실질적으로 극대화하는 P2P 금융 투자 전략을 제시할 수 있을 것이다.

참고 문헌

[단행본]

- Fabozzi, F. J. (2016). *Bond markets, analysis, and strategies* (9th ed.). Pearson.
- Gutierrez, A., & Mathieson, D. (2017). *Optimizing investment strategy in peer-to-peer lending* (CS229 Final Report). Stanford University.

[논문]

- Mo, L., & Yae, J. (2022). Lending Club meets Zillow: local housing prices and default risk of peer-to-peer loans. *Applied Economics*, *54*(35), 4101–4112. https://doi.org/10.1080/00036846.2021.2022089
- Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business, 39*(1), 119–138. https://doi.org/10.1086/294846
- Sharpe, W. F. (1994). The Sharpe ratio. *The Journal of Portfolio Management, 21*(1), 49–58. https://doi.org/10.3905/jpm.1994.409501