LendSure(LS) Model
: Model with “The Right” Threshold

통계 데이터 사이언스 4조
문현빈, 신주영, 이수빈, 이정민, 이지호, 조정욱, 현민재

Table of Contents

1. Introduction

1.1. P2P Lending and Lending Club

Lending Club(이하 LC)은 2006년에 설립된 미국의 금융 서비스 기업이며 최초의 개인 간(Peer-to-Peer, P2P) 대출 플랫폼으로, 개인 대출자와 투자자 간의 직접적인 연결을 통해 중간자 없이 대출을 제공하는 전통적 금융 시스템과는 차별화된 금융 서비스를 제공한다. 또한 미국 증권거래위원회(SEC)에 보안을 등록하고 제2 시장에서 대출 거래를 제공한 최초의 회사이기도 하다표[1]. LC는 개인 대출, 자동차 Re-financing, 장학 및 의료 대출, 상업 대출 등의 여러 금융 상품을 제공하였으며, 2020년대에는 Radius Bank를 인수한 후 디지털 은행으로 전환하여, P2P 대출 플랫폼 운영을 중단하고 은행 서비스를 메인으로 제공하는 방향으로 사업 모델을 변경하였다,

본 과제는 LC의 개인 대출상품의 데이터를 기반으로 sharpe ratio 극대화를 주요 목표로 설정하고 있기에, LC의 개인 대출상품을 자세하게 소개하고자 한다. LC의 개인 대출은 $1,000에서 $40,000까지 가능하며, 일반적으로 36개월(3년) 또는 60개월(5년)로 제공된다. 대출 자격 요건으로는 최소 신용 점수 600점 이상이 요구된다. 이를 충족한 대출 희망자에 한하여 이자율은 신용 등급에 따라 차등적으로 고정금리가 부과되며, 대출 관련 기타 수수료를 포함한 APR(Annual Percentage Rate)이 적용된다. 대출금 상환은 고정금리로 대출 이자가 부과됨에 따라 원리금 균등상환의 방식이 적용되며, 대출자는 대출금 조기 상환 시 조기 상환 수수료가 부과되지 않는다. 대출 신청 시의 대출 목적에 따라 대출이 진행되며, 대출은 부채 통합, 비상 자금, 주택 마련 및 개조의 유형으로 분류된다 [2].

LC의 대출 프로세스는 크게 5단계로 나누어 진다.

1. 대출자의 대출 목적, 대출 금액, 신용 등급 등의 정보를 바탕으로 사전 심사 진행
2. 대출자가 대출 조건에 적합한 경우 여러 대출을 제안
   (공동 대출자를 추가 시, 더 나은 금리나 더 큰 대출금 제공)
3. 대출자는 온라인으로 필요한 문서를 제출한 후, LC의 서류 검토 진행
4. 대출 심사 시 Hard Credit Inquiry가 수행되며,
   대출 승인 시 대출 계약서 서명이 진행됨
5. 계약 완료 후 대출금이 대출자의 계좌로 직접 입금되며, 대출자는 대출금 상환해 나감

투자자는 LC 플랫폼을 통해 대출 목록을 검토하고 투자할 수 있다. 대출 목록을 검토 시, 대출자의 신용등급이나 대출 목적 등을 확인할 수 있다. 투자자의 최소 투자 금액은 $25이며, 투자자는 대출자가 상환하는 이자를 통하여 수익을 획득한다.

LC는 위와 같은 개인 대출 방식을 통하여 대출자와 투자자 모두에게 수익이 창출되는 Business Model을 구축하였다. 대출자에게는 대출자의 신용 등급에 따라 차등적으로 부과되

는 발급 수수료가 부과된다. 투자자에게는 대출자가 상환하는 금액의 일정 비율로 책정되는 서비스 이용 수수료가 부과된다 [3], [4].

LC의 개인 대출 사업은 여러 의의를 지닌다. 대출 희망자의 여러 고금리 신용카드 부채를 하나의 저금리 대출로 통합하여 월간 상환 금액을 줄임으로써, 상환 부담을 줄이는 데 일조했다. 부채 통합을 통해 대출자는 신용 점수를 개선할 수 있으며, 장기적인 금융 안정성 확보의 측면에서도 긍정적인 효과를 창출했다. 또한 온라인 기반 대출이 진행되었기 때문에, 간단한 신청과 빠른 대출금 지급이 이루어지는 시스템으로 구축되었다.

위와 같은 효과를 창출하는 효율적인 개인 대출 사업을 통해 LC는 전 세계에서 가장 큰 P2P 대출 플랫폼 중 하나로 성장할 수 있었으며, 2015년 말까지 플랫폼을 활용하여 $15.98B 이상의 대출이 이루어졌다. P2P 대출 시장에서 압도적인 성장을 보여줌에 따라 2014년에 뉴욕 증권거래소(NYSE)에 IPO를 진행하였고, 약 $900M를 모금하면서 당 해 미국에서 가장 큰 Tech IPO가 되었다.

1.2. Problem Finding

앞서 언급한 것처럼 P2P 대출은 은행 및 대부업에서 진행되는 전통적인 대출과는 다른 방식으로 운용되는 특수한 대출이다. 왜냐하면 은행에서 주관되는 전통적인 대부업과 달리, 대부업을 운용하는 주체의 수익이 대부 상품 자체에서 이루어지는 것이 아니라, 그것의 거래 수수료를 통해서 이루어지는 구조이기 때문이다. 즉, P2P 대출은 대출건수에 비례하여 운영 주체의 수익이 증가하는 구조이기에, 이는 운영 주체의 moral hazard로 인해 초래되는 agency problem의 발생 가능성을 높이는 구조다. 실제로 LC도 agency problem을 경험했다. 이는 P2P 대출의 본래 목적[1]에 부합하는 방식으로 회사를 사업을 운용하지 않음과 동시에, 거래건수를 늘려 단기적인 이윤 극대화를 추구했기에 발생했다.

P2P 대출의 특성을 토대로 해당 사업 모델에 내재된 2가지 위험 요소를 도출해보았다. 첫째는 채무자의 파산 및 기한 내 상환액 미납으로 인한 투자자의 수익 감소이고, 둘째는 운영 사업체의 자금 운용 실패로 인한 재정난이다. 그리고 해당 위험은 주로 전자에서 후자로 발전하는 경향이 있다. LC가 경험한 문제도 이와 유사하다. 채무자들의 파산으로 인해 투자자들의 수익률 감소하여 투자자 유치가 어려워졌고, 그로 인해 투자 건수가 감소하여 회사의 자금 운용의 측면에서 문제가 발생하였다. 이를 통해, 채무자의 파산여부에 대한 정확한 판단이 P2P 대출업의 내재된 리스크 해결의 핵심적인 요소임을 알 수 있다.

P2P 사업체들은 투자자 유치에서 큰 어려움을 경험하고 있으며, 이는 LC가 P2P 대출을 그만두게 된 핵심적인 사유 중 하나이기도 하다. 본 연구팀은 해당 현상의 원인을 크게 2가지로 분석하였다. 첫째는 충분한 위험고지의 실패이다. 주식과 채권, 그리고 기타 파생상품들과 달리 P2P 대출은 투자자 유치에서 유난히 큰 어려움을 겪고 있다. 이는 투자자들이 P2P 대출에서 발생하는 손해에 대해 더 민감하게 반응하고 있음을 방증한다. 이러한 현상이 발생하는 이

1) 본 연구팀은 P2P 대출의 본래 목적이 ‘대안데이터를 활용하여 전통적인 신용평가모델이 발견하지 못하는 invisible prime을 발견하는 것’과 ‘thin filer에 대한 신용평가를 진행하는 것’이라 생각한다.

유는 여러 가지로 추론될 수 있으나, 본 연구팀은 P2P 대출에 내재되어 있는 리스크에 대한 명확한 인식이 투자자들 사이에 존재하지 않음이 가장 큰 원인이라고 생각한다. 그러기에, 모델의 sharpe ratio를 공시함으로써 투자자들에게 리스크를 명확히 인식시킨다면, 투자자 유치의 문제를 부분적으로 해결할 것으로 기대된다. 둘째, 데이터로부터 유의미한 정보를 충분히 획득하지 못하였다. 데이터가 충분히 많으면 object oriented inference가 가능하기에, 빅데이터에 기반을 둔 부도 예측 모델의 제작은 이론적으로 가능하다. 허나, 그 데이터를 토대로 유의미한 모델을 만드는 것은 다른 차원의 일이다. 적절한 feature의 선택부터 원하는 결과값이 도출될 수 있는 모델을 선택하고 튜닝시키는 과정까지 면멸히 진행되고 검토되어야, 현실에서도 유용한 모델이 만들어질 수 있다. 그러므로 대부분의 P2P 대출 사업체들이 투자자 유치에 어려움을 겪고 있음은 핀테크 업체들이 빅데이터로부터 유의미한 정보를 충분히 추출해내지 못함의 결과로 해석될 수 있다. 본 연구팀은 economic significance를 가지는 모델 선택, domain knowledge까지 활용한 feature selection 실행, 목적함수를 기준으로 적절한 model selection 진행하는 것으로 이것이 해결될 수 있다고 생각한다.

본 연구팀은 LC의 P2P업 재개를 위해서는 투자자와 운영주체의 joint profit을 최대화하는 모델이 제시되어야 한다고 생각한다. 그러므로 LC의 agency problem 해결과 joint profit maximization을 위해 부도 예측 모델을 개발하였다. 이 모델명은 LendSure(이하 LS)이며, 이는 LC의 모델에 새로운 SurePoint를 주겠다는 의미를 담고 있다.

2. Method 1 : Ideas Before Modeling

2.1. Base Objectives of Modeling

(1) Make Economic Significant Model

의미있는 data scientific한 task를 수행하기 위해서는 'right or wrong'을 다루기보다는 'whether useful or not'에 집중하는 것이 필요하다. 그러므로 금융과 관련된 데이터를 다루는 본 과제에서는 economically significant한 모델을 구축하는 것이 핵심 목표가 되어야 한다. 데이터 사이언스는 통계학 기반 모델을 중점적으로 연구하는 학문이지만 실생활에서의 유의미한 운용도 중요하게 여겨지고 있기에, 통계학적 모델은 statistically significance만을 추구해서는 안된다. 왜냐하면 통계적으로 유의미한 것과 실질적으로 돈을 만들어내는 선택은 다르기 때문이다. 그러므로 본 연구팀에서는 economically significant한 부도 예측 모델을 구축하고자 한다.

비부도보다 부도의 리스크 정도가 더 크기 때문에, 전통적으로 부도 예측 모델을 개발할 때에는 recall 값이 높은 모델을 선택하곤 한다. 하지만, 이러한 선택은 economically significant한 판단이 아니다. 왜냐하면 실질적으로 P2P 대출 사업체에게 경제적으로 효과적인 해답을 제공하기 위해서는 초과 수익률까지 고려해야 하기 때문이다. P2P 대출에서의 이자율은 수익률로 해석되며, 초과 수익률은 "이자율 - 무위험수익률"로 정의될 수 있다. 앞서 언급했듯 P2P 대출업의 존속을 위해서는 안정적인 수익률이 보장되어야 하기 때문에, 이를

고려하여 모델 평가기준 및 threshold 설정이 진행될 필요가 있다.

(2) Big Data and Feature Selection

Bayesian approach란 “data를 prior knowledge로 간주하여 prior probability를 도출하고, 이를 기반으로 more likely한 선택을 하는 것이 통계적으로 유의미하다”고 판단하는 접근법이다. 이는 머신러닝에 기초가 되는 이론이므로, 머신러닝의 궁극적인 목적은 정확한 data를 토대로 최대한 정확한 prior knowledge and probability를 구하는 것으로 이해될 수 있다. 정확한 prior knowledge and probability를 도출하기 위해 빅데이터가 활용되며, 데이터가 커질수록 더 복잡한 모델링이 가능하다.

데이터가 클수록 object oriented inference가 용이해지기 때문에, 더 적은 prior knowledge가 필요하다. 허나, 모델 자체에 온전히 의지하여 feature를 선택하는 것은 많은 위험요소를 내포하고 있는 행위이다. 그러므로 모델의 feature selection 과정에서는 data science와 domain knowledge가 동시에 작용해야한다. 그러기에 LS model은 feature selection의 과정에서 domain knowledge와 모델에서 제공하는 feature selection 및 managing method를 활용하여 feature를 선정하였다 (Table 1).

Feature	Reason
dti	● Serrano-Cinca et al. (2015) : 이 연구에서는 부채상환비율(DTI)을 차주의 부채 수준을 나타내는 변수로 사용하였으며, 부도 확률과의 관계 분석. "The accuracy of the model is improved by adding other information, especially the borrower's debt level."이라고 언급하며, 차주의 부채 수준이 모델의 예측 정확도를 향상시킨다고 주장 [5]. ● Emekter et al. (2015) : 이 연구에서는 DTI 가 부도 확률에 유의미한 영향을 미친다고 주장. "We find that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults." DTI 가 부도에 중요한 역할을 한다고 밝힘 [6].
revol_util	● Serrano-Cinca et al. (2015) : 이 연구에서는 신용카드 한도 소진율(revol_util)을 차주의 부채 수준을 나타내는 변수 중 하나로 포함하였으며, 부도 확률과의 관계를 분석. "The accuracy of the model is improved by adding other information, especially the borrower's debt level."이라고 언급, 차주의 부채 수준이 모델의 예측 정확도를 향상시킨다고 주장. ● Emekter et al. (2015) : 신용카드 한도 소진율(revol_util)이 부도 확률에 유의미한 영향을 미친다고 주장. "We find that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults."라고 언급, revol_util 이 부도에 중요한 역할을 한다고 주장.
fico_avg	● Emekter et al. (2015) : 이 연구에서는 FICO 점수가 부도 확률에 유의미한 영향을 미친다고 보고. "We find that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults."라고 언급하며, FICO 점수가 부도에 중요한 역할을 한다고 주장.

Table 1. Feature selected based on domain knowledge

2.2. Reflection and Implementation of Midterm Presentation Feedback

중간 발표에 대한 교수님과 조교님의 피드백을 본 연구팀은 본 모델에 다음과 같은 방식을 통해 반영하였다.

(1) Delete Int_rate Model and Recovery Model

본 연구팀은 중간발표에서 구축한 이자율 모델, 회수율 모델, 부도 예측 모델을 소개하였다. 하지만 교수님께서 이자율과 회수율을 예측하여 이를 부도 예측 모델의 input variable로 사용한다면, 한 행에서의 값이 달라지게 되어 인과적으로 부도확률도 달라질 수 없다고 말씀하셨다. 본 연구팀도 해당 지적사안에 대해 깊이 동의했다. 그러기에 중간발표 이후 이자율과 회수율 모델 제작에 힘쓰기보다는 부도 예측 모델의 논리 및 성능 향상에 집중하였다.

(2) Redefine Sharpe Ratio

중간발표 때에는 sharpe ratio의 본 의미를 명확히 파악하지 못했기 때문에, 직접 sharpe ratio의 수식을 산정하는 큰 오류를 범하였다. Sharpe ratio의 개념에 대한 명확한 이해가 부족했기에, sharpe ratio의 의미인 '위험대비 초과수익률'이라는 개념을 차용하여 새로운 수식을 제작하였다 (Equation 1-3). 이는 개개인의 대출건수에 대한 위험대비 초과수익률을 구하려는 시도였으나, 교수님과 조교님의 피드백을 토대로 숙고해본 결과, 해당 수식이 sharpe ratio의 의미를 담고 있지 않음을 깨달았다.

$$\frac{\mathbb{E}[\text{기대수익률} - \text{무위험수익률}]}{\sqrt{\text{위험도}}}$$

Equation 1. Sharpe ratio of midterm presentation

$$\text{기대수익률} = P(\text{비부도}) \times (\text{비부도 수익률}) + P(\text{부도}) \times (\text{부도 수익률})$$

$$\text{무위험수익률} = \text{대출 발생 당시의 미국 국채 수익률}$$

$$\text{위험도} = (\text{비부도 수익률}) - (\text{부도 수익률}) = 1 - \text{회수율}$$

Equation 2. Supplementary equations for calculating sharpe ratio

$$\text{비부도 수익률} = \frac{\text{상환기간} \times \text{월 상환금액} - \text{현재까지 승인된 총 대출 금액}}{\text{현재까지 승인된 총 대출 금액}}$$

$$\text{부도 수익률} = \frac{\text{상환기간} \times \text{월 상환금액} \times \text{회수율} - \text{현재까지 승인된 총 대출 금액}}{\text{현재까지 승인된 총 대출 금액}}$$

Equation 3. Earning rate depending on default probability

그래서 최종보고서에서는 sharpe ratio의 본래 의미에 따라 다음과 같이 정의하였다 (Equation 4).

$$\text{Sharpe Ratio} = \frac{E[\text{기대수익률} - \text{무위험수익률}]}{\sqrt{\text{위험}}} = \frac{E[\text{초과수익률}]}{\sqrt{\text{위험}}}$$
$$= \frac{\frac{1}{N}\sum_{i=1}^{N}((\text{이자율})_i - (\text{T-bill})_i)}{\sigma_{re}}$$

$$N = \text{test data의 행 수}$$
$$\text{T-bill} = \text{해당 대출의 issue date의 DTB3}$$
$$\sigma_{re} = \text{초과수익률 histogram의 std}$$

Equation 4. Sharpe ratio of final report

Sharpe ratio는 위험대비 초과수익을 수치화한 지표이다. 이를 최대화하는 모델을 만들기 위해서는 분자인 초과수익을 최대화하고 분모인 위험을 최소화하는 것이 필요하다. P2P 대출은 채권과 유사한 성격을 가지므로[2] 위험을 최소화하는 것이 최우선적으로 고려되어야 한다.

그러므로 본 연구팀은 위험을 최소화하는 것을 최우선적으로 고려함과 동시에 수익을 최대화하는 방향으로 sharpe ratio를 최대화하였다. 분모인 위험을 최소화는 정확한 부도예측을 통해 실현될 수 있다. 왜냐하면 정확한 부도 예측을 통해 2종 오류에 속하는 상품에 대한 대출을 최소화하는 것이 곧 부도의 위험을 최소화하는 행위이기 때문이다. 분자인 초과수익률을 최대화하는 방법은 투자 거래 건수를 최대화와 동치 개념이다. 왜냐하면 P2P 대출회사의 수익률은 거래 건수에 비례하는 수수료 수익 증가로 결정되기 때문이다.[3]

---

2) 채권의 경우 기준금리가 변동하면서 채권 가격의 변동성이 존재하기에, 수익 정산 시 기초 약정되어 있던 수익률과의 차이가 존재하므로 P2P대출과 완전히 동일하지는 않다. 또한 P2P 대출은 대출이 진행될 때 채무관계가 발생할 뿐, 채권 발행은 진행되지 않아 이를 직접 거래하는 시장이 부재하므로, 유동성 측면에서는 채권과의 차이를 보인다. 하지만 P2P 대출과 채권은 모두 최고 수익률이 정해져 있다는 점, 채무관계가 형성된다는 점과 부도여부가 중요하다는 공통점을 가지므로, P2P 대출이 채권과 유사한 성격을 가진다고 간주할 수 있다. 채권은 상품 특성 상 위험 관리가 굉장히 중요하므로 본 연구팀 역시 위험 관리에 초점을 맞춰 표준편차를 최소화하는 것을 최우선 목표로 삼아 LS 모델을 개발하였다.

3) 수익률 극대화를 위해 IRR을 고려하여 대출이자 산정한다면, IRR보다 낮을 경우 대출 승인을 하지 않는 방법의 도입으로 투자자의 수익률을 보장해줄 수 있다. 하지만, LC가 P2P 기반 플랫폼이라는 점에 주목한다면, 대출 승인 건수가 늘어날수록 중개 수수료 등을 통해 LC의 수익이 극대화될 것이다. LS 모델은 부도를 정확히 탐지하여 위험이 최소화된 투자상품을 선별하는 모델이다. 그러므로 해당 모델에서 비부도로 탐지된 투자 상품의 경우 부도 확률이 현저히 낮다. 그러므로 LS모델의 경우에는 IRR과 무관하게 대출 희망자에 한하여 대출을 해줌으로써, 운영주체의 수익률을 극대화하는 대출 승인 모델을 계획했다.

(3) Make Default Prediction Results Meaningful

중간발표 이후 조교님께서 따로 찾아오셔서 부도 여부를 알기 위해 조정한 threshold가 sharpe ratio를 구하는 데에 전혀 사용되지 않은 점을 지적하셨다. 이는 본 연구팀도 중간발표를 준비하면서 인지하고 있던 사안이었지만, 마땅한 해답을 발겨하지 못하여 발표자료를 수정하지 못한 상황이었다. 교수님과 조교님의 피드백을 바탕으로 모델의 전반적인 구조를 고민하다보니, 비부도인 사람에게만 대출을 해주는 것이 P2P 대출의 위험을 최소화하는 방향임을 깨달았고 이를 바탕으로 수익률을 최대화하는 방향으로 부도 예측 모델의 threshold를 선정했다.

3. Method 2 : Modeling

3.1. Data Purification

(1) Train and Test Data

본 보고서에는 업로드된 train과 test data를 어떠한 방식으로 전처리를 했는지에 대해서 간략하게 서술해두었다. 자세한 내용은 Supplementary Materal에 수록되어 있으므로 참고부탁한다.

(1)-1. Data Selection

모델에 사용되는 변수는 실제 사용되는 상황을 상정하고 선정되어야 한다. 모든 데이터는 시간 순으로 발생하기 때문에, 모델의 target variable에 따라 예측에 사용할 수 있는 독립변수가 결정된다. 따라서 실제 P2P 대출 프로세스를 시점에 따라 구분하는 것이 필요했다. 본 연구는 데이터 발생을 3개의 시점으로 구분하였다. 부도예측이 필요한 시점이 대출을 승인하기 직전이라 판단을 하였기에, LC 방문 시 발생한 데이터까지를 사용하여 부도 예측 모델의 데이터셋을 구성하였다.

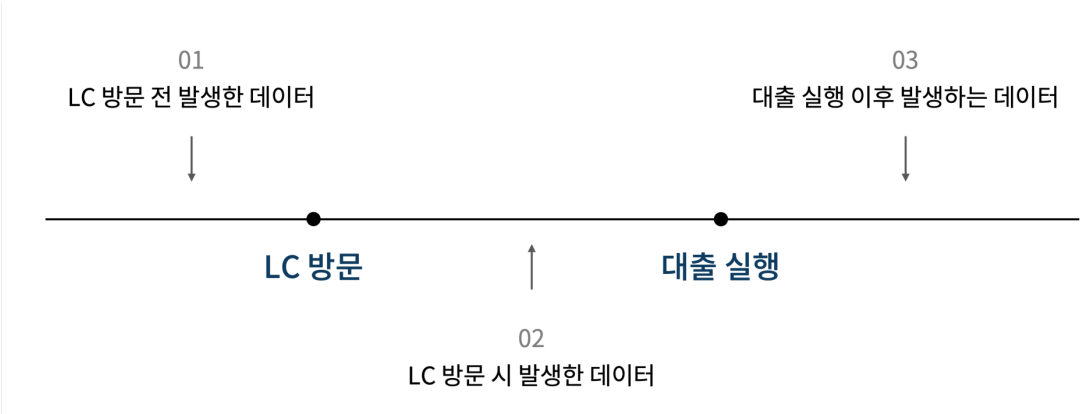


Figure 1. Data timeline

(1)-2. Data Preprocessing

LC raw data의 feature들 중 부도 예측에 불필요하다고 여겨지는 feature는 크게 두가지 경우 중 하나에 해당한다. 첫째는 id, title, policy_code처럼 메타데이터에 가까운 경우이고, 둘째는 emp_title, purpose, addr_state, zip_code처럼 데이터 정제를 하면 유의미한 데이터로 이용할 수 있을 것으로 추정되지만, 전처리에 어려움이 있거나 현재 가설과 로직에 부합하지 않는 경우이다. 본 연구에서는 두 경우에 해당하는 feature를 모두 삭제하였다. 또한 실제 모델의 운용을 생각해보았을 때 채무 불이행에 대한 예측은 채무 불이행 발생 이전의 데이터로만 예측되어야 하기 때문에, 채무 불이행 이후에 발생하는 열은 삭제하였다.

순서가 있는 범주형 변수의 경우 순서를 반영하여 전처리를 진행하였고, 순서가 없는 범주형 변수는 binary encoding과 one-hot encoding을 통해 전처리하였다. 수치형 변수는 pandas data frame에서 숫자로 인식될 수 있도록 “%”같은 문자열을 제거한 후 숫자 데이터로 변환해주었고, percentage를 나타내는 변수의 경우 소수점 실수값(e. 10% -> 0.10)의 형태로 전처리하였다.

마지막으로 결측치는 train과 test dataset 각각의 결측치 개수와 결측치 비율을 구하여 공결측 feature를 찾아냈다. 공통결측칼럼의 결측치 비율을 내림차순 해보면, 약 45%의 결측치 비율을 가지는 칼럼을 마지막으로, 결측치 비율이 10% 미만으로 현저하게 줄어드는 것이 확인된다. 그러기에 본 연구팀은 결측치 비율이 40% 이상인 칼럼을 제거하였다. 그외의 공통결측 feature에 대해서는 결측치가 있는 행을 제거하는 방식으로 전처리하였다.

(2) T-Bill Data

T-Bill(Treasury Bill)은 미국 재무부(Department of the Treasury)가 발행하는 만기 1년 이하의 단기 국채로, 대표적인 무위험 수익률 지표이다. 대출거래가 활발히 이루어지는 LC에서 대출상품의 초과수익률을 정확히 산출하기 위해서는 T-Bill 중 발행주기가 짧은 4주, 13주(3달), 26주(반년) 만기 T-Bill 중 하나를 선택하여 사용하는 것이 필요하다. 그 중 13주, 26주 만기의 T-Bill이 일반적으로 많이 거래되는 채권이기 때문에, 둘 중 만기가 상대적으로 짧은 DTB3(DiscounT Basis 3-month)의 값을 무위험 수익률로 선정하였다.

DTB3의 값은 fred.stlouisfed.org에서 2007.06~2020.09.의 기간의 값을 추출해왔다[4]. LC dataset의 issue_d(issue date, 대출 실행 일자)가 월단위까지 나와있기에, DTB3 데이터의 경우 각 월의 대표값을 선정하는 과정이 반드시 수반되어야 한다. 기준금리를 결정하는 회의인 FOMC 회의의 일정이 연 8회 중 5~6번 정도로 주로 15일 전후로 분포되어 있기에, FOMC 회의 직후의 금리 변동이 반영된 T-Bill 이자율을 활용하기 위하여 15일의 DTB3 값을 각 월의 대푯값으로 간주하였다. 단, 15일의 값이 없는 경우, 해당 달의 중앙값을 대푯값으로 사용하였다.

4) 자료 출처 링크 : https://fred.stlouisfed.org/series/DTB3

3.2. Modeling

(1) Modeling Objectives

(1)-1. Threshold Objective Function

분류 모델의 경우 threshold가 그 모델의 performance에 지대한 영향을 끼친다. LS model은 sharpe ratio를 maximize하는 방향으로 구축되어야 하기에, 해당 모델의 threshold를 결정하는 경우에도 위의 목적함수가 반드시 고려되어야 한다. Sharpe ratio의 분자는 각 대출상품에 내재된 수익률을 기반으로 결정되며, 그것의 분모는 분자의 std를 통해 종속적으로 결정된다. Sharpe ratio의 분모는 분자보다 선행되어 구해질 수 없으므로, 모델링을 할 때 가장 주요하게 고려해야 하는 것은 분자의 maximization이다. 그러므로 본 연구팀은 초과수익률을 가장 최대화할 수 있는 threshold를 결정하는 것이 필요하다고 판단하였다.

본 연구에서는 loan_status=default인 경우를 1로 매핑하였고, loan_status=fully_paid인 경우를 0으로 매핑하였다. 그러므로 TN은 실제로 fully_paid인 경우에 대해 fully_paid라고 예측한 경우이고, FN은 실제로 default이나 fully_paid로 잘못 예측한 경우이다. 부도난 채무자의 투자상품의 경우, 기대수익률에 양의 영향을 미치지 못한다. 본 연구팀은 Equation 5에서 보이는 것과 같이 TN의 추가수익률과 FN의 추가수익률의 차를 maxmize함으로써, 모델의 초과수익률 최대화를 유도했다. 그후 초과수익률의 argmax 값을 threshold로 선정함을 통해, 최대 초과수익률을 도출하는 training model을 구축하였다.[5)]

$$\arg\max_{t} \text{초과수익률} = \sum_{i=1}^{N} (\text{TNer}_i - \text{FNer}_i)$$

$$i\text{th 상품의 TN 추가수익률} = \text{TNer}_i = \begin{cases} \text{초과수익률}_i, & \text{if TN} \\ 0, & \text{if not} \end{cases}$$

$$i\text{th 상품의 FN 추가수익률} = \text{FNer}_i = \begin{cases} \text{초과수익률}_i, & \text{if FN} \\ 0, & \text{if not} \end{cases}$$

$$t = \text{threshold of default classification model}$$

Equation 5. Threshold objective function

5) 실제 모델링을 실행할 때에는 argmax가 아니라 -초과수익률의 argmin을 구하는 방식으로 진행되었다. 이는 최적화 알고리즘 수행 시, computing time을 고려했을 때 maximization보다 minimization problem을 solving하는 것이 더 수월하기 때문에, maximization problem을 의미적으로 동일한 minimization problem으로 산정한 후 계산을 수행하였다.

(1)-2. Loan Approval Policy

앞서 언급한 것처럼, 채무자의 부도는 기대수익률의 실현을 좌절시키기 때문에, P2P 대출업에서 가장 피해야 하는 risk이다. 그러므로 LS 모델은 부도날 것으로 예상이 되는 대출희망자의 대출을 거부하고 부도를 하지 않을 것으로 예상되는 대출 희망자에게는 대출을 해주는 방향으로 대출 승인 모델을 구상하였다 (Equation 6).

We will only approve loans for the ones who have (대출수락)i = 1

$$(\text{대출 수락})_i = \begin{cases} 1, & \text{if } f(x_i) < t \\ 0, & \text{if } f(x_i) \geq t \end{cases}$$

$$x_i = i\text{th investment product}$$
$$f(x) = \text{default probability prediction model}$$
$$t = \text{LS model threshold}$$

Equation 6. Loan approval function

(2) Type Selection of the Model

LS model의 목적은 채무자의 부도 여부를 정확히 판단하는 분류모델이기에, 분류 모델 중에서 tree model 기반 boosting 계열 모델인 XGBoost와 LightGBM을 base model로 선택하였다. 선택한 이유는 3가지로 정리된다. 첫째, tree model 기반 boosting 계열 모델의 경우 데이터 불균형에 강건한 특성을 보인다. 왜냐하면 해당 알고리즘들에 활용하는 트리 분할 방식은 impurity 감소를 최대화하는 방향으로 데이터를 분할하기에, 소수의 데이터의 경우에도 중요하게 고려될 수 있기 때문이다. 또한 boosting 계열 알고리즘이기에, 오분류된 데이터에 대한 가중치가 부여된다. 불균형 데이터의 경우 소수 데이터에 가중치가 더 많이 부과되는 경향성을 보이기에, 소수 데이터가 겪을 것으로 예상되었던 패널티의 상당부분은 가중치 부과를 통해 경감될 수 있다. 그러기에 fully paid 행과 default 행의 비율이 4:1이 되는 현재의 데이터셋의 경우, tree model 기반 boosting 계열의 모델들은 좋은 분류능을 선보일 것으로 기대된다.

둘째, 학습속도가 빠르다. 본 연구팀이 중요하게 여기는 것은 목표함수에 대한 최적의 모델 개발이다. 최적의 모델을 exploit하기 위해서는 많은 exploration을 선행되어야 하기 때문에, 모델을 학습시키는 데에 요구되는 시간적인 cost를 최소화하는 것이 중요하다. XGBoost와 LightGBM의 경우 학습속도가 빠르기 때문에, 본 연구팀이 요구하는 모델의 조건에 부합한다.

셋째, 각 feature의 가중치 조정이 자동적으로 이루어져, feature selection에 상대적으로 덜 민감하다. 데이터 사이언스란 기존의 도메인 지식을 넘어서서 데이터 자체가 주는 정보를 포착하여 모델을 구축하는 것, 즉 데이터를 통한 새로운 가능성을 발견하는데 목적을 두는 학문이다. 많은 선행연구들을 통해 부도 여부에 영향을 주는 변수들은 발견되었지만, 이러한 변

수들과 알려진 가중치를 사용하여 모델링을 하는 것은 전통적인 부도 예측 모델과의 차별점이 없는 모델의 구축으로 이어지게 된다. 본 연구팀의 LS model은 전통적인 부도 예측 모델의 한계를 빅데이터를 통해 넘어서는 것을 목표로 한다. 해당 목표를 성취하기 위해서는 금융학적인 이론으로부터의 접근이 아니라 데이터로부터의 접근이 필요하다. 그러므로 feature selection과 그것의 파라미터에 대한 요구되는 주관적인 판단의 정도가 낮은 tree model 기반 boosting 계열 모델을 선택하였다.

(3) Modeling Procedure

Figure 2은 본 연구의 전체적인 모델링 flow chart이다. 모델링 과정은 크게 model developing – model evaluation – model comparison으로 진행된다. 하부 장에서는 각 절차에서 모델링이 어떠한 방식으로 이루어졌는지에 대해 자세히 기술되어 있다.

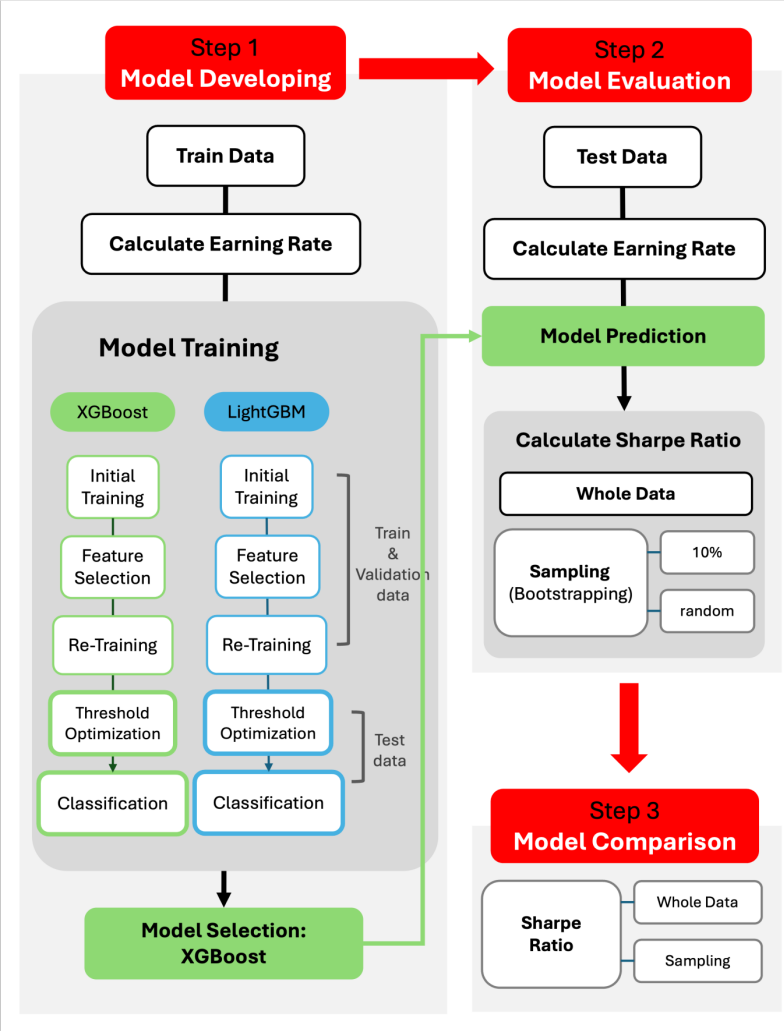


Figure 2. Modeling flow chart

(3)-1. Model Developing

(3)-1-①. Calculate Earning Rate

LS model은 sharpe ratio maximization을 위해 threshold objective function과 loan approval policy를 지정하였다. 이 두 objective들과 sharpe ratio를 계산하기 위해서는 earning rate가 필요하기에, 본격적인 model training에 들어가기 앞서 earning rate를 계산했다 (Equation 7).

$$\text{LS 초과수익률} = \sum_{i=1}^{N} (\text{TNer}_i - \text{FNer}_i)$$

Equation 7. Earning Rate

(3)-1-②. Model Training

가. Initial training

Processed된 train data를 입력하여 XGBoost와 LightGBM의 모델을 구축하였다. 해당 모델은 아래와 같은 하이퍼파라미터를 가지고 구동되었다 (Table 2). Train data, validation data, test data의 비율은 3:1:1이며, 이는 향후 train-test 모델링에도 모두 동일하게 적용되었다.

XGBoost		LightGBM	
Hyperparameter	Rate	Hyperparameter	Rate
objective	binary:logistic	objective	binary
eval_metric	logloss	metric	binary_logloss
random_state	40	seed	40
num_boost_round	500	num_boost_round	500
early_stopping_rounds	10	early_stopping_rounds	10

Table 2. Hyperparameters

나. Feature Selection

Initial model에 사용된 값을 feature importance 순서대로 나열하였다. 그 후 상위 30개의 feature를 선정하여 이를 re-training model에 입력하였다. 단, 위에서 언급했듯 데이터 사이언스를 다룰 때 도메인 지식 없이 모델링을 구현하는 것은 risky한 일이므로, Table 1에서 지정된 부도 여부에 큰 영향을 끼치는 것으로 알려진 feature(dti, revol_util, fico_avg)들은 반드시 포함하도록 코딩하였다. 즉, 총 3개의 domain feature는 모델 training에 고정적으로 사용되었으며, 나머지 27개의 feature는 도메인 feature가 아님과 동시에 initial model의 feature importance top 30에 속하는 것들이다.

다. Re-Trainig

Feature selection을 통해 선택된 30개의 feature를 입력 변수로 지정하여 XGBoost와 LightGBM 모델을 re-training하였다. Selected된 feature만을 사용해야 하므로 초기에 지정

된 train, validation, test data를 재조정하는 과정이 필요하기에, 해당 절차가 모델링에 포함되도록 코딩하였다. 모델에는 Table 2의 하이퍼 파라미터들이 동일하게 적용되었다. 모델 구축 후 feature importance mapping을 plot하였고, 최종 모델에 사용된 feature와 각각의 feature importance를 시각화하였다.

전통적인 classification의 경우 re-training을 진행하지 않지만, LS model의 경우 2가지를 이유로 인해 re-training의 과정을 진행하였다. 첫째, domain 지식과 결합된 부도 예측 모델을 제작하기 위해서다. Re-trained LS model은 domain feature를 반드시 포함하도록 디자인되었다. 이는 domain feature가 feature importance 내림차순 행렬의 상위 30등에 포함하지 않더라도 모델의 input feature로 사용되게 하기 위함이다. 데이터 모델링의 관점에서 바라보았을 때 해당 과정은 비효율적이라 판단될 수 있으나, domain knowledge와 model의 feature selection을 동시에 고려해야 한다는 측면에서 바라보았을 때에는 매우 합리적임과 동시에 안전한 선택으로 여겨질 수 있다. 둘째, 유의미한 feature들을 토대로 모델의 분류 능력을 높이기 위함이다. Initial training과정에서 모델에 input되는 feature들은 loan_status와 연관된 feature들의 조합이 아니라, 부도 예측이 진행되기 전에 한 대출희망자로부터 얻을 수 있는 모든 feature들의 조합이다. 어떠한 방식으로든 feature selection이 일어나지 않은 모델의 경우, feature selection을 한 모델보다 분류 성능이 높기는 힘들다. 그러므로 initial training에서 얻은 feature importance를 토대로 feature를 고르는, 즉 수동적인 매커니즘을 통해서라도 feature를 select함으로써 모델의 분류 능력을 높이고자 하였다.

라. Threshold Optimization

Re-train된 모델에 adjusted test dataset을 입력하여 각 투자상품 별 부도확률을 예측하였다. 해당 모델로부터 도출된 예상 부도 확률을 토대로 초과수익률을 maximize할 수 있는 값을 threshold로 설정하였다 (Equation 5 참고).

마. Classification

Threshold optimization을 통해 선정된 최적의 threshold에 따라, 각 투자상품에 대한 부도 확률을 기반으로 부도여부를 판단하였다.

(3)-1-③. Model Selection

Model selection이란 최종적으로 사용할 모델을 선정하는 단계로, 본 연구의 최종적인 목적인 sharpe ratio maximization과 직접적으로 연관된 과정이다. 앞서 언급했듯 sharpe ratio의 분모는 분자의 값이 온전히 도출되어지만 산정될 수 있으므로, 분자의 값이 모두 추출되기 이전의 상태인 model selection의 단계에서 고려될 수 있는 것은 오직 분자(초과수익률) maximization이다. 그러므로 sharpe ratio maximization을 위해, 두 모델 중 더 높은 초과수익률을 가지는 모델을 LS model로 선정하였고, 그 결과 XGBoost가 선정되었다.

(3)-2. Model Evaluation

(3)-2-①. Caculate Earning Rate

Model developing 단계에서 진행한 것과 동일한 목적과 방식으로 earning rate를 계산하였다.

(3)-2-②. Model Prediction

Model developing 과정에서 선택된 XGBoost 모델을 사용하여 입력된 test dataset의 부도확률을 계산하였고, XGBoost모델에서 도출된 optimized threshold를 사용하여 부도 여부를 predict하였다. Trained model에서 사용된 feature의 조합으로 test data가 input되어야만 trained model의 파라미터를 활용한 prediction이 이루어질 수 있다. 그러므로 코드 상에서 이를 맞추어 주는 과정이 필요하다. 허나 기존의 XGBoost의 모델과 달리 LS 모델은 re-training을 시킨 모델이므로, 다른 random seed를 입력할 때마다 다른 feature의 조합이 select되어 re-training model에 input될 것이다. 그러므로 dynamic하게 test data의 input feature를 trained data의 input feature와 동일하게 맞추어 주는 과정이 필요하다. 그러기에, LS model에서는 trained model의 selected feature를 토대로 사용되지 않는 feature들을 drop 시키는 함수를 설계하였고 이를 prediction 과정 직전에 실행되도록 모델링을 하였다.

Prediction이 이루어진 이후 loan approval policy에 의거하여 fully paid로 predict된 투자 상품에 한해서 대출을 승인하였다 (Equation 6 참고).

(3)-2-③. Calculate Sharpe Ratio

승인된 대출 상품들을 토대로 sharpe ratio를 계산하였다. Sharpe ratio는 대출이 승인된 투자상품들의 earning rate histogram의 mean과 std을 구하고 이를 나누어줌으로써 계산된다.

(3)-3. Model Comparison

LS 모델은 LC 모델보다 더 높은 sharpe ratio를 가지는 것을 목표로 하고 있다. 그래서 크게 2종류로 분류된 총 3가지 방법을 통해 LC 모델 대비 LS 모델의 성능을 평가하였다. 첫 번째는 모평균과 모분산을 보는 것이다. 이는 모든 earning rate의 기댓값과 분산을 통해 구해지며, histogram을 통해 시각화될 수 있다. 두 번째는 boostrapping이다. Bootstrapping은 복원추출의 기법 중 하나로, sample에 대한 통계값 계산이 가능하도록 한다. Bootstrapping을 적용한 이유는 실제 P2P 대출의 진행이 특정 distribution하에서 bootstrapping된 상품들에 대한 평가를 내리는 것으로 해석이 될 수 있기 때문이다. 그러기에, 특정 distribution에서 random하게 복원추출하는 과정을 통해 만들어진 sample의 sharpe ratio를 계산하였고, 이를 histogram의 형식으로 시각화하여 LS와 LC 모델을 평가하였다.

Bootstrapping은 총 2가지 방식으로 진행되었다[6]. 첫 번째는 fixed rate bootstrapping이다. 위 방식을 통해 각 월마다 승인되는 P2P 대출의 비슷하게 유지되는 시기에 대한 모델의 성능을 평가할 수 있다. 이를 위해 fixed rate를 임의로 10%로 선정하였다. 하지만, 경제적인 상황에 따라 개인의 대출 선호도는 변화하기에, random한 sampling이 진행될 때의 모델 성능도 측정할 필요가 있다. 그러므로 본 연구팀은 random rate bootstrapping을 추가로 진행하였다.

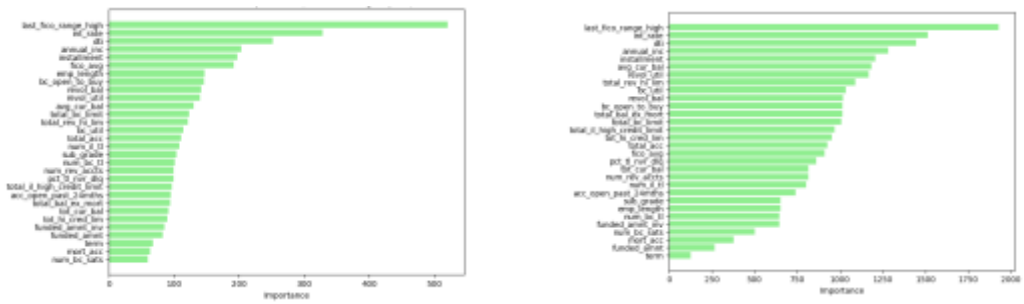
4. Result and Discussion

4.1. Base Model

(1) XGBoost

XGBoost 모델의 feature importance는 Figure 3과 같다. 전반적으로 많은 feature들이 고루 높은 importance rate를 가지는 방식으로 모델이 구성된 것을 확인할 수 있다. 신기한 점은 re-trained 모델에서 전반적으로 각 feature들의 importance가 상승하였다는 것이다. 이는 re-training의 과정에서 상대적으로 낮은 feature importance를 가진 feature들을 training에서 제외시키는 방향의 학습이 유의미했음을 보여준다. Domain feature로 선정되었던 dti, revol_util, fico_avg가 각각 re-trained model의 feature importance mapping에서 3등, 7등, 17등을 했다. XGBoost가 LS 모델로 선정되었음을 감안하고 바라보았을 때, domain feature 중 dti와 revol_util은 부도 예측에 유의미한 영향을 주는 변수들이라 판단할 수 있다. 다만, fico_avg의 경우 타 domain feature들보다 상대적으로 feature importance가 낮게 도출되었다는 사실과 last_fico_range_high가 분류모델에서 가장 높은 feature importance를 가진다는 점을 고려해보았을 때, fico 점수의 극단값이 평균값보다 부도 예측에서 유의미한 영향을 끼친다고 판단할 수 있다. 선택된 top 30 feature는 Table 3에 기록되어 있다.

6) 각각의 P2P 대출 투자상품의 경우 자신만의 시계열 구조를 가지고 있기에, 각 투자상품의 earning rate histogram을 하나의 P2P 대출 distribution으로 보는 것에 대해 반대하는 견해가 존재할 수 있다. 허나, 인류의 역사는 반복된다는 역사적 유추와 귀납적 추론의 관점을 가지고 바라본다면, 시계열을 가지는 상품일지라도 미래에 동일하거나 유사한 상품이 발생할 가능성이 존재하기에, 과거에 특정 timeflow를 거치며 형성된 데이터의 경우에도 동일시기의 P2P 대출의 distribution으로 간주할 수 있다고 생각한다.



(a) Initial Training (b) Re-Training

Figure 3. Feature Importance Mapping of XGBoost

	Feature	Description
1	last_fico_range_high	최종 FICO 점수 상한 구간
2	int_rate	대출 이자율
3	dti	총 부채 상환액 대비 소득 비율
4	annual_inc	신청 시 보고한 연간 소득
5	installment	월 상환 금액
6	avg_cur_bal	모든 계좌의 평균 현재 잔액
7	revol_util	리볼빙 한도 대비 사용률
8	total_rev_hi_lim	리볼빙 총 한도
9	bc_util	은행카드 한도 대비 잔액 비율
10	revol_bal	리볼빙 신용 잔액 총액
11	bc_open_to_buy	리볼빙 은행카드 사용 가능 한도
12	total_bal_ex_mort	주택담보 제외 총 잔액
13	total_bc_limit	은행카드 총 한도
14	total_il_high_credit_limit	할부 계좌 총 한도
15	tot_hi_cred_lim	총 최고 신용 한도
16	total_acc	총 신용 계좌 수
17	fico_avg	신청 시 FICO 신용점수 상/하한 구간 평균
18	pct_tl_nvr_dlq	연체 경험 없는 계좌 비율
19	tot_cur_bal	모든 계좌 현재 잔액 합계
20	num_rev_accts	리볼빙 계좌 수
21	num_il_tl	할부 계좌 수
22	acc_open_past_24mths	최근 24개월 동안 개설된 신용거래 건수
23	sub_grade	LC 부여 대출 하위 등급
24	emp_length	근무 기간
25	num_bc_tl	은행카드 계좌 총 수
26	funded_amnt_inv	투자자 승인 총 금액
27	num_bc_sats	양호한 은행카드 계좌 수
28	mort_acc	주택담보대출 계좌 수
29	funded_amnt	현재까지 승인된 총 대출 금액
30	term	상환 기간 (36/60개월)

Table 3. Selected Features for Re-Training in XGBoost

다음은 XGBoost 모델의 classification 결과다. Threshold는 0.51이며, earning rate는 1824159.799999922이다 (Table 4). Table 5은 XGBoost 모델의 confusion matrix이며, TN값은 155611이고, FN 값은 10604이다. Recall은 0.7307이고, precision은 0.7562이다,

Metric	Rate
threshold	0.51
1earning rate	1824159.799999922
accuracy	0.9027
precision	0.7562
recall	0.7307
f1 score	0.7432

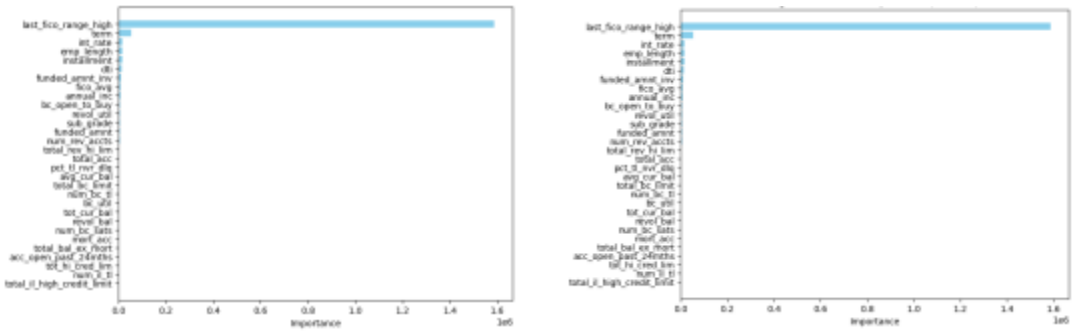
Table 4. Overall Metrics of XGBoost

		Prediction	
		Fully Paid	Default
Actual	Fully Paid	155611 (TN)	9280 (FP)
	Default	10604 (FN)	28779 (TP)

Table 5. Confusion Matrix of XGBoost

(2) LightGBM

LightGBM의 feature importance는 Figure 4와 같다. 전반적으로 last_fico_range_high의 feature importance 값이 높게 책정된 것이 확인된다. 두 번째로 높은 feature importance 값을 가진 feature인 term이 1등인 last_fico_range_high와 비교하였을 때 매우 작은 것을 통해, 하나의 feature만을 지배적으로 활용하여 분류모델을 수행했다고 판단할 수 있다. 해당 모델이 LS 모델로 선정되지 않았다는 사실을 통해, 한가지 feature만을 중점적으로 활용하여 분류하는 모델의 경우 성능이 좋지 않음을 유추할 수 있다. LightGBM의 경우에도 domain feature인 dti, revol_util, fico_avg가 각각 feature importance에서 6등, 11등, 8등을 했음이 확인되지만, 3등이부터는 feature importance rate 간의 차이가 크지 않았다. 이는 domain feature가 classification에 거의 영향을 끼치지 않았다고 결론지을 수 있다. Re-trained LightGBM의 input feature는 Table 6에 기록되어 있다.



(a) Initial Training (b) Re-Training

Figure 4. Feature Importance Mapping of LightGBM

	Feature	Description
1	last_fico_range_high	최종 FICO 점수 상한 구간
2	term	상환 기간 (36/60개월)
3	int_rate	대출 이자율
4	emp_length	근무 기간
5	installment	월 상환 금액
6	dti	총 부채 상환액 대비 소득 비율
7	funded_amnt_inv	투자자 승인 총 금액
8	fico_avg	신청 시 FICO 신용점수 상/하한 구간 평균
9	annual_inc	신청 시 보고한 연간 소득
10	bc_open_to_buy	리볼빙 은행카드 사용 가능 한도
11	revol_util	리볼빙 한도 대비 사용률
12	sub_grade	LC 부여 대출 하위 등급
13	funded_amnt	현재까지 승인된 총 대출 금액
14	num_rev_accts	리볼빙 계좌 수
15	total_rev_hi_lim	리볼빙 총 한도
16	total_acc	총 신용 계좌 수
17	pct_tl_nvr_dlq	연체 경험 없는 계좌 비율
18	avg_cur_bal	모든 계좌의 평균 현재 잔액
19	total_bc_limit	은행카드 총 한도
20	num_bc_tl	은행카드 계좌 총 수
21	bc_util	은행카드 한도 대비 잔액 비율
22	tot_cur_bal	모든 계좌 현재 잔액 합계
23	revol_bal	리볼빙 신용 잔액 총액
24	num_bc_sats	양호한 은행카드 계좌 수
25	mort_acc	주택담보대출 계좌 수
26	total_bal_ex_mort	주택담보 제외 총 잔액
27	acc_open_past_24mths	최근 24개월 동안 개설된 신용거래 건수
28	tot_hi_cred_lim	총 최고 신용 한도
29	num_il_tl	할부 계좌 수
30	total_il_high_credit_limit	할부 계좌 총 한도

Table 6. Selected Features for Re-Training in LightGBM

Re-trained LightGBM은 threshold가 0.6이고, earning rate는 1740265.869999929이다. XGBoost의 결과값과 비교해보았을 때 threshold는 0.1이 더 높고 earning rate 89394.56999999정도 더 작다. 부도에 대한 정확한 분류도를 보여주는 metric인 recall 값을 보았을 때에도 0.7307인 XGBoost 모델보다 0.0718 작은 0.6589임을 알 수 있다 (Table 7). 즉, LightGBM 모델은 XGBoost 모델과 비교했을 때 earning rate와 recall 값이 모두 작다. 이는 Table 5와 8의 confusion matrix 비교를 통해서도 쉽게 알 수 있다.

Metric	Rate
threshold	0.6
earning rate	1740265.869999929
accuracy	0.8698
precision	0.6634
recall	0.6589
f1 score	0.6612

Table 7. Overall Metrics of LightGBM

		Prediction	
		Fully Paid	Default
Actual	Fully Paid	151727 (TN)	13164 (FP)
	Default	13434 (FN)	25949 (TP)

Table 8. Confusion Matrix of LightGBM

4.2. Selected Model : XGBoost

(1) Confusion Matrix

Table 9는 LS 모델로 선정된 XGBoost 모델의 metrics table이다. Test data 상의 recall 값인 0.7317이 train data상의 recall 값보다 0.0010 더 크다는 사실을 통해, LS model이 overfitting되지 않았다고 판단할 수 있다. 모델의 분류성능은 confusion matrix를 통해서 확인 가능하다 (Table 10).

Metric	Rate
accuracy	0.9018
precision	0.7536
recall	0.7317
f1 score	0.7425

Table 9. Overall Metrics of LS (threshold = 0.5)

		Prediction	
		Fully Paid	Default
Actual	Fully Paid	517647 (TN)	31518 (FP)
	Default	35356 (FN)	96408 (TP)

Table 10. Confusion Matrix of LS

(2) Model Comparison : Sharpe Ratio

Figure 5-6는 LS 모델의 histogram이고 Figure 7-8는 LC 모델의 histogram이다. 모든 histogram의 경우 모두 평균에서 가장 높은 frequency가 관측되고 tail로 갈수록 frequency가 줄어드는 normal distribution과 유사한 분포를 가지고 있다. 두 모델 모두 1000회의 bootstrapping을 실행하였는데, 해당 distribution의 평균값이 모평균과 거의 유사함을 확인하였다. 이를 통해, 일정한 대출선호를 가지는 상황과 급변하는 대출 선호도를 가진 상황 모두에서 두 모델은 평균적으로 전체 distribution과 유사한 성능을 보인다. 또한, 두 모델 모두에서 fixed rate bootstrapping보다 random rate bootstrapping에서 더 작은 분산이 관측되었다.[7)]

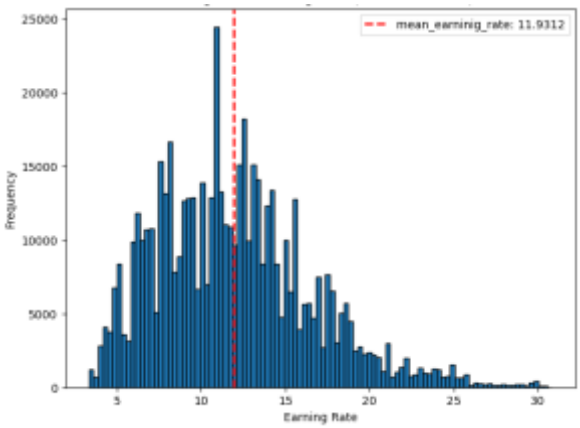


Figure 5. Whole dataset : earning rate histogram (LS)

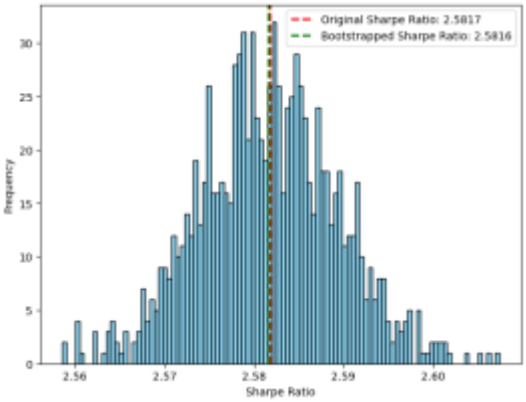


Figure 6-(a). Fixed Rate Bootstrapping

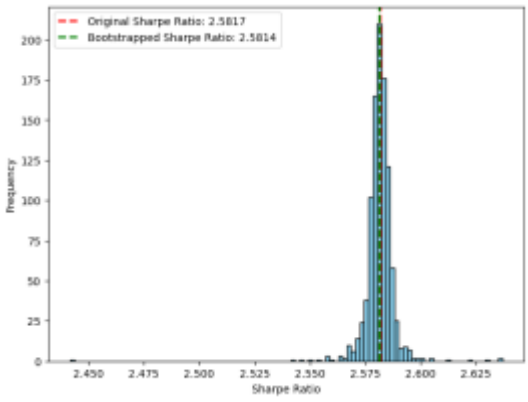


Figure 6-(b).
Random Rate Bootstrapping

Figure 6. Histogram of sharpe ratio bootstrapping (LS)

7) 이러한 결과의 원인에 대해 고민해보았으나, 아직 명료한 답안을 얻지 못하였다. 향후 더 많은 통계학적 추론 및 분석을 통해 해당 문제에 대한 답을 고민해볼 계획이다.

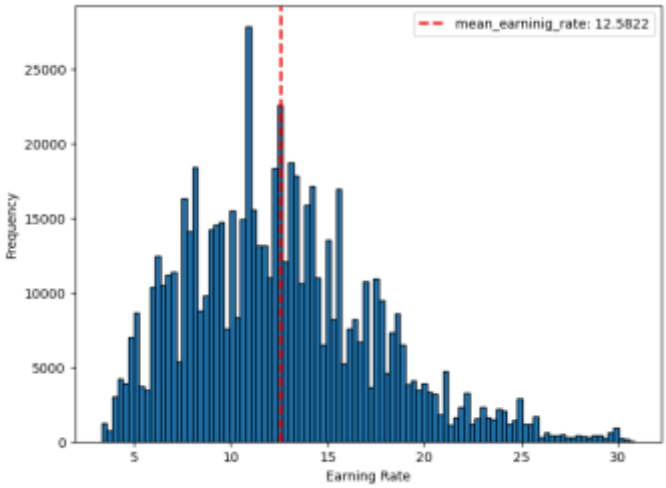


Figure 7. Whole dataset : earning rate histogram (LC)

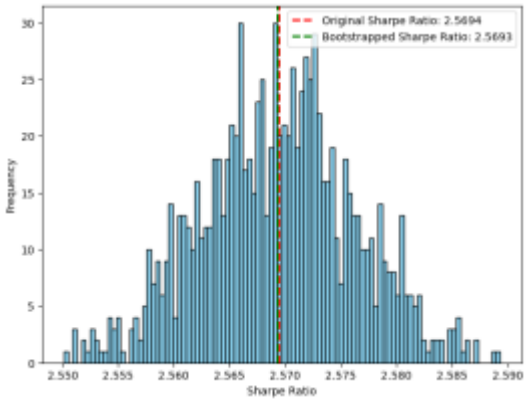


Figure 8-(a). Fixed Rate Bootstrapping

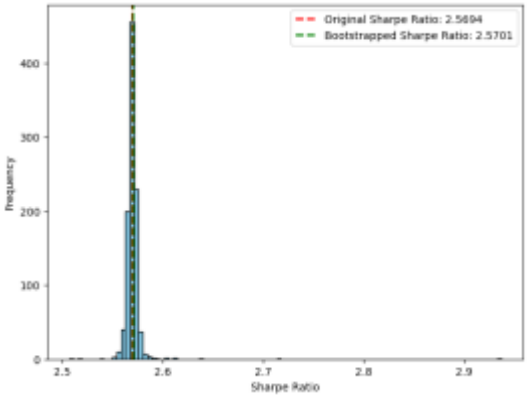


Figure 8-(b).
Random Rate Bootstrapping

Figure 8. Histogram of sharpe ratio bootstrapping (LC)

Table 11의 결과를 통해 LS 모델이 LC 모델보다 더 좋은 성능을 보임이 확인된다. 우선 전체 데이터셋을 기준으로 평가를 한 whole을 보자. LC 모델에서 mean이 더 높으나 std가 더 크기 때문에, 궁극적으로 sharpe ratio의 측면에서는 LS 모델보다 더 낮은 값을 가지게 된다. 즉, LC가 운용하던 모델은 초과수익률이 더 높은 모델이나 위험이 높은 모델로 해석될 수 있다. LS 모델은 더 적은 평균 초과수익률을 가지나 정확한 부도 예측으로 인해 P2P 대출에 내재되어 있는 부도 리스크를 LC 모델보다 더 minimize시켰기 때문에, LC 모델보다 더 안정적인 수익률을 보장하는 모델이 되었다. 더 나아가 이는 본 연구팀이 P2P 대출 상품을 채권과 유사한 금융 상품으로 바라보고 risk minimization을 우선적으로 고려하는 방향으로 모델을 설계한 것이 올바른 방향이었음을 증명해주는 결과이기도 하다.

동일한 결과는 bootstrapping에서도 관측된다. LS 모델이 fixed rate과 random rate bootstrapping 모두에서 LC 모델보다 더 높은 mean을 가지고 있다. 이를 통해, LS 모델이 평균적으로 더 높은 sharpe ratio를 가지고 있음을 알 수 있다. 2^{nd} moment인 std를 보면, LS 모델이 fixed rate에서는 높은 값을 보이나, random rate에서는 더 낮은 값을 보인다. 하지만, 95% 신뢰구간으로 보았을 때, LS 모델이 fixed rate과 random rate bootstrapping 모두에서 더 높은 값을 가지고 있으므로, LS 모델이 LC 모델보다 더 높은 sharpe ratio를 안정적으로 보장하는 모델이라고 판단할 수 있다.

Range		Statistics	LC	LS
whole		mean	12.5822	11.9312
		std	4.8969	4.6214
		sharpe ratio	2.5694	2.5817
Sampling (bootstrapping)	fixed rate	mean	2.5693	2.5816
		std	0.0069	0.0078
		2.5^{th} percentile	2.5554	2.5668
		97.5^{th} percentile	2.5823	2.5971
	random rate	mean	2.5701	2.5814
		std	0.0139	0.0077
		2.5^{th} percentile	2.5597	2.5689
		97.5^{th} percentile	2.5790	2.5933

Table 11. Selected Features for Re-Training in XGBoost

5. Conclusion

5.1. Summary

본 연구팀은 LC로 대표되는 P2P 대출 플랫폼의 근본적인 문제점, 즉 운영 주체의 moral hazard로 인해 발생하는 agency problem과 투자자 유치 어려움을 부도 예측 모델인 LS 모델의 개발로 해결하고자 했다. P2P 대출이 승인된 건수에서 채무 불이행이 발생할 시, 투자자는 손실로 인한 투자 중단을 할 가능성이 크며, 그로인해 운영주체도 투자건수 감소로 재정난을 경험할 가능성이 크다. 즉, P2P 대출업에서 부도는 중점적으로 관리해야하는 리스크 중 하나다. 하지만, 타 투자상품에 비해 투자자 유치에 어려움을 겪고 있다는 점을 통해 P2P업의 리스크가 투자자들에게 명확히 인지되지 못하고 있음과 운영주체들이 빅데이터를 활용하여도 충분히 좋은 부도 탐지 모델을 구축하지 못하고 있다는 점을 유추해볼 수 있다. 그래서 본 연구팀은 투자자들에게 명확한 위험을 고지하기 위해, LS model의 sharpe ratio를 통지하는 방안을 생각하였다. 더 나아가, 운영주체의 리스크를 최소화하기 위해 sharpe ratio를 maximization을 목표로 하는 부도 예측 모델의 개발하는 것을 목표로 삼았다. 그래서 본 연구팀은 "Sharpe Ratio를 최대화하는 부도 예측 모델"인 LS model을 구축하였다. 해당 모델은 '위험을 최소화하는 것이 최우선적으로 고려되어야 한다'는 점과 'P2P 대출 운영회사의 수익은 거래 건수에 비례하는 수수료 수익 증가로 결정된다'는 점이 동시에 고려될 수 있는 방

향으로 설계되었다.

먼저, 경제적으로 유의미한 모델을 만들기 위해 수익률을 고려한 부도 예측 모델의 평가 기준 및 threshold를 설정했다. Domain knowledge를 활용하여 부도와 관련된 주요 feature(dti, revol_util, fico_avg)를 선정하고, bayesian approach에 기반하여 빅데이터를 활용한 feature selection을 진행했다. 그리고 중간 발표 피드백을 반영하여 sharpe ratio를 재정의하고, 부도 예측 결과를 바탕으로 비부도 예측 대상에게만 대출을 승인하는 방식으로 위험을 최소화했습니다. 모델링에는 데이터 불균형 처리에 용이하고 학습 속도가 빠르고 feature selection에 대해 필요로 하는 개입의 정도가 작은 XGBoost와 LightGBM을 사용했다. 모델 학습 과정은 earning rate 계산, model training (initial training, feature selection, re-training), threshold optimization, classification 단계로 진행되었고, 최종적으로 XGBoost 모델이 선정되었다.

선정된 XGBoost 기반 LS 모델은 LC 모델과 비교했을 때, 더 높은 sharpe ratio를 보였다. 이는 LS 모델이 더 적은 평균 초과 수익률을 가짐에도 불구하고 정확한 부도 예측을 통해 P2P 대출에 내재된 부도 리스크를 최소화했기 때문에 가능한 결과였다. Bootstrapping에서도 동일한 결과가 확인되었다. Fixed rate와 random rate bootstrapping에서 95% 신뢰구간에 속하는 sharpe ratio의 under and upper boundary 둘다 LS model에서 더 높았다. 이를 통해, 본 연구팀이 개발한 LS 모델이 기존 LC 모델보다 안정적으로 더 높은 수익률을 보장하는 모델, 투자자와 운영 주체 모두에게 위험 대비 최대의 초과수익률을 보장하는 모델임을 입증하였다.

5.2. Limitations and Future Research

(1) Limitations of LS Model

LS 모델은 LC 모델보다 성능이 좋으나, 더 좋은 성능을 가지기 위해서는 보완해야 할 지점들이 관찰된다.

첫째, initial training을 진행할 때, 변수의 분류없이 모든 feature가 사용된었다. 대출 시기 직전의 모든 데이터를 분류 모델에 사용하는 것은 해당 시기까지의 모든 데이터가 부도 여부와 관련이 있다는 가정 하에서만 유의미한 선택이다. 하지만 LS 모델의 경우 그러한 가정 하에 input data를 입력한 것이 아니기 때문에, 해당 측면에서의 개선이 필요하다. 이는 train model의 input feature를 선택할 때 부도와의 상관관계를 검증하는 등, 각 변수의 기여도를 고려하여 initial model의 input feature의 조합을 구성해봄으로써 구현될 수 있다. 비록 본 연구에서는 re-training 방식으로 feature selection을 부분적으로 구현하였지만, initial training부터 부도 예측에 유의미한 feature들을 input feature로 선택한다면 더 좋은 성능을 가진 모델이 개발될 것이다.

둘째, re-training model의 input feature 갯수를 30개로 제한한 것이다. 현재 re-training model을 구동할 때에는 initial training의 classification에서 feature importance score가 높은 상위 27개의 feature + 3개의 domain feature를 선택하여 re-training model의 input feature로 사용하고 있다. 하지만 feature 개수에 대한 최적화는

진행하지 않았기에, 30개가 가장 최적의 input feature의 수가 아닐 가능성이 농후하다. 그러기에, 30이 아닌 새로운 N의 값을 랜덤으로 산정하여, 그 중 가장 좋은 성능을 보이는 N을 feature max값으로 지정하여 re-training model의 input feature로 구현한다면, re-trained model의 feature 개수 문제는 해결될 수 있을 것이다.

셋째, threshold optimization이 진행되지 않았다. 본 연구에서는 threshold optimization function에 의해 threshold를 즉각적으로 도출하였다. 하지만 위와 같은 방식으로 도출된 threshold는 train data에 overfitting될 위험이 존재한다. 그러기에 threshold 파라미터 선정할 때 k-fold의 방식을 활용한 update로 진행한다면, overfitting의 가능성을 낮출 수 있을 것이다.

(2) Develop MAX_Sharpe_Ratio Model Using Reinforcement Learning

Sharpe ratio는 본래 포트폴리오 구성에서 주로 사용되는 개념이다. 본 연구팀은 포트폴리오를 구성하는 방식으로 문제를 해결하려고 하였으나, computing power 및 수학적 지식의 한계로 인해 정해진 기간 내에 코드의 형태로 구현해내지 못하였다. 그래서 이번 장에서는 본 연구팀이 논의했던 모든 대출 거래 데이터를 고려한 포트폴리오 구축과 sharpe ratio maximization에 대해서 이야기해보고자 한다.

Maxmized된 sharpe ratio를 가진 포트폴리오는 아래와 같은 방식으로 산출된다. 먼저, 기대수익률, 표준편차, 상관관계를 고려한 efficient frontier를 그래프로 그린 후, 무위험 수익률을 포함한 직선의 자본배분선을 efficient frontier에 접하도록 그린다. 즉, 두 그래프가 접하는 접점이 발생할 것이며, 그 접점은 시장 포트폴리오이다. 시장 포트폴리오를 지나는 자본배분선은 sharpe ratio가 가장 높은 자본시장선이 되므로, 본 연구팀이 목표하는 sharpe ratio의 극대화가 실현되는 지점이다.

이제 포트폴리오 이론을 코드로 구현할 때, 어떠한 지점들을 고려하면서 구현되어야 하는지에 대해 설명하겠다.

Step 1. Efficient Frontier 그리기

: 포트폴리오는 자산의 기대수익률, 표준편차, 상관관계를 계산하여 수익률과 분산을 그래프로 표현하기만 하면 되는 간단한 작업이지만, LC의 경우 행의 수가 과도하게 많기에 하나의 대출 거래 데이터를 개별 자산으로 치부하여 efficient frontier를 도출하는 전통적인 접근법으로는 efficient frontier를 조성하는 것이 불가능에 가깝다. 이를 극복하기 위해서는 추가적인 전처리 작업이 요구된다. 이는 자산 수익률을 공통 요인으로 묶어 변수의 개수를 감소시키는 것이다. LC 대출 상품의 경우 각 채무자 별 신용등급이 산정되어 있기 때문에, 이를 공통 요인으로 활용할 수 있다. 즉, 더 적은 수의 데이터가 활용되는 방식이기에, 신용등급 별 대표값을 선정하여 자산군을 구성한다면 efficient frontier를 구할 수 있을 것이다.

Step 2. 자본배분선 그리기

: LC의 총 대출 거래 데이터의 기간에 해당하는 3개월 T-bill의 평균 이자율을 도출하여, 자본배분선을 그리기 위한 y절편(무위험 수익률)으로 지정한다. 해당 y절편을 기준으로 기울기 값(sharpe ratio)을 조절하여 efficient frontier에 접하는 접점을 찾는다.

Step 3. 자본시장선 찾기

: 자본배분선의 기울기 값(sharpe ratio)을 조절하여 efficient frontier에 접하는 접점은 시장 포트폴리오이다. 해당 지점에서는 sharpe ratio가 가장 극대화 되는 포트폴리오를 구성할 수 있다.

1~3의 과정에서 강화학습을 활용한다면 모든 경우의 수에 대한 연산을 수행하지 않더라도 sharpe ratio가 maxmized된 포트폴리오를 구할 수 있을 것이다. '시뮬레이션-행동-보상'의 사이클을 통해 고차원 최적화 문제를 인간의 직관을 넘어서 정밀하게 도출 해낼 수 있다고 기대된다.

5.3. Advanced LS Model to Achieve Operational Efficiency

Sharpe ratio가 본래 펀드매니저의 투자 포트폴리오 성과의 평가에 활용되는 지표이기 때문에, 해당 개념을 각 개별 P2P 대출 데이터의 초과 수익률 극대화에 적용하는 데에는 다소 억지스러운 부분이 존재한다. 그러나, sharpe ratio의 극대화가 과제의 목적이었기에, sharpe ratio의 개념적 정의인 '위험 대비 초과수익률'을 활용하여 sharpe ratio가 도출될 수 있는 방향으로 새롭게 문제 상황을 정의하였고, 이를 토대로 LS 모델을 개발하였다. 하지만 LS 모델의 경우 분배의 효율성에 집중하여 개발된 모델이기에, 운영의 효율성까지 고려하지 못했다는 한계점을 가진다. 그래서 본 연구팀은 운영의 효율성까지 고려하는 사업모델을 개발하기 위해 노력하였다.

본 연구팀이 생각해낸 방안은 대출신청자들의 신용등급을 기반으로 tranche를 구성하여 개별 대출 건수에 대한 위험을 pooling하는 것이다. CDO의 개념을 차용하여 신용도에 따라 차등적인 투자 수익률을 제공하는 대출 포트폴리오가 구성된다면, 포트폴리오가 발생함에 따라 sharpe ratio를 포트폴리오 별로 계산할 수 있게 된다. CDO는 'tranche 별 포트폴리오 內 자산들 간에는 낮은 상관관계가 있어야 한다'는 조건을 충족해야 한다. 하지만 경기 상황에 따라 대출금 상환에 직접적인 영향을 받는 개인의 경우, 대출 상품 상에서 상호 연관성이 비교적 높게 관찰될 확률이 크기에, P2P 대출에 대해 CDO의 방식을 착안한 포트폴리오의 구성은 개념적으로 불가능해 보인다.

하지만, CDO 방식을 도입하여 P2P 대출을 운용하게 된다면, 회사는 운영의 효율성 극대화를 이루어낼 수 있을 것이다. P2P 대출 상품을 토대로 상관관계가 0에 수렴하는 포트폴리오를 구성하는 데에 성공하여 CDO개념의 도입이 가능해졌다고 가정해보자. 그 경우, 하나의 P2P 대출 상품에서 default가 발생했을 때에도 투자금이 pooling되어 있기 때문에, 투자자들

은 tranche에 따라 원금을 회수할 수 있을 것이다. 또한 CDO의 경우 equity 등급이 존재하기에, 부도 채무자가 다량으로 발생했을 경우에도 equity 등급에 투자한 투자자들에게 책임을 전가하는 방식으로 운영될 수 있다. 즉, LC의 대출금 상환 의무는 equity 등급에 투자한 투자자들에게 전가되며, 이로 인해 책임 소재의 측면에서 LC는 대출금 상환의 의무로부터 자유로워질 수 있다. 그러므로 LC의 P2P lending 사업에 CDO의 개념을 도입된다면, 채무자의 default 여부와 회사의 자금운용이 CDO를 도입하기 전보다 독립적으로 운용될 수 있을 것이다. 더 나아가 CDS 계약의 체결까지 이루어질 수 있으므로, 기업은 최악의 상황이 발생할 때에도 자금의 유동성을 확보할 수 있다는 점에서, 보장성이 갖춰진 자금 운용 정책을 수행할 수 있게 된다. 이러한 운용의 효율성 이점을 취하기 위해서, 'P2P 대출 포트폴리오 간 상관관계 최소화가 이루어질 수 있는 방향'의 모색이 필요하며, 이에 대해서는 추가적인 고민과 연구가 필요할 것으로 사료된다.

Reference

[1] LendingClub, https://www.lendingclub.com

[2] Rebecca Potters, “LendingClub Personal Loan Review: Everything You Need to Know”, BUSINESS INSIDER, 2024-09-04, https://www.businessinsider.com/personal-finance/personal-loans/lendingclub-personal-loans-review (2025-03-11)

[3] Lindsay Frankel, “LendingClub Review [2025]: Easy Online Personal Loans and More”, FINANCE BUZZ, 2024-11-11, https://financebuzz.com/lendingclub-personal-loan-review (2025-03-13)

[4] Maureen Milliken, “LendingClub”, Debt.org, 2024-12-24, https://www.debt.org/credit/loans/personal/lending-club-review/ (2025-03-12)

[5] Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L., 2015, “Determinants of default in P2P lending”, PloS one, 10(10), e0139427.

[6] Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics, 47(1), 54-70.

Supplementary Material

Ⅰ. Data Selection

본문에 언급한 것처럼 데이터는 시점에 따라 나누었고 본 연구에서는 [02] 시점까지의 데이터까지 사용을 하여 분석을 진행하였다. 각 시기의 데이터를 알고 싶다면 아래의 표를 참고하면 된다.

[01] Lending Club 방문 전 발생한 데이터 (59개)

term	emp_length	annual_inc	dti
delinq_2yrs	fico_avg	inq_last_6mths	open_acc
pub_rec	revol_bal	revol_util	total_acc
acc_now_delinq	tot_cur_bal	mths_since_rcnt_il	total_bal_il
il_util	max_bal_bc	all_util	total_rev_hi_lim
total_cu_tl	avg_cur_bal	bc_open_to_buy	bc_util
chargeoff_within_12_mths	mjort_acc	num_accts_ever_120_pd	num_actv_rev_tl
num_bc_sats	num_bc_tl	num_op_rev_tl	num_rev_accts
num_rev_tl_bal_gt_0	num_sats	pct_tl_nvr_dlq	percent_bc_gt_75
pub_rec_bankruptcies	tax_liens	tot_hi_cred_lim	total_bal_ex_mort
total_bc_limit	total_il_high_credit_limit	hone_ownrship	verification_status
open_acc_6m	open_act_il	open_il_12m	open_il_24m
open_rv_12m	open_rv_24m	inq_fi	inq_last_12m
acc_open_past_24mths	num_actv_bc_tl	num_il_tl	num_tl_120dpd_2m
num_tl_30dpd	num_tl_90g_dpd_24m	num_tl_op_past_12m	

[02] Lending Club 방문 시 발생한 데이터([01]까지 발생한 데이터 + 6개)

loan_amnt	int_rate	installment	grade
sub_grade	issue_d	last_fico_range_high	last_fico_range_low

[03] 대출이 실행이후 발생하는 데이터([2]까지 발생한 데이터 + 8개)

funded_amnt	funded_amnt_inv	loan_status	out_prncp
out_prncp_inv	total_rec_late_fee		

Ⅱ. Data Processing

본문에는 개괄적으로 data preprocessing이 어떻게 진행되었는지를 서술해두었다. 구체적으로 어떠한 방식으로 data processing이 진행되었는지에 대해에서는 아래에 서술되어 있다.

[Step 01] 모델링에 불필요하다고 판단된 열 제거

id	title	application_type	next_pymnt_d
policy_code	delinq_amnt	hardship_flag	zip_code
emp_title	purpose	addr_state	earliest_cr_line
initial_list_status	last_credit_pull_d	debt_settlement_flag	mo_sin_rcnt_rev_tl_op
mo_sin_rcnt_tl	last_pymnt_amnt	pymnt_plan	total_rec_prncp
total_rec_int	total_pymnt	total_pymnt_inv	issue_d를 제외한 날짜 관련 칼럼

[Step 02] 채무불이행 이후 발생하는 열 제거

recoveries	collection_recovery_fee	collections_12_mnths _ex_med	tot_coll_amt

[Step 03] 타겟 변수인 loan_status의 칼럼값이 결측치인 행은 전부 제거

[Step 04] 범주형 변수 (순서가 있는 변수)

칼럼명	처리방법	이유
emp_length	무직인 경우만 특별한 값(-1)로 매핑한 후 나머지는 근속연수에 맞춰 동일한 숫자로 매핑	직업이 아예 없는(없었던) 경우를 제외하고는 근속연수가 길수록
term	'36 months'와 '60 months'을 36, 60으로 매핑	현재 칼럼 값이 36과 60밖에 없어서 바이너리로 매핑해도 되지만, 엄연히 기간은 순서가 있는 변수이고 추후 확장성을 고려해 실제 기간 값을 매핑
grade	알파벳 순으로 0~n까지 동일 간격 정수를 매핑	신용 등급은 순서가 있는 변수이므로 동일 간격의 수치로 매핑해주어야 순서가 유지됨
sub_grade	알파벳과 숫자 오름차순에 따라 0~n까지 동일간격 정수를 매핑	grade와 동일

[Step 05] 범주형 변수 (순서가 없는 변수)

칼럼명	처리방법	이유
loan_status	Fully paid = 0, Charged off(default) = 1 나머지 칼럼값을 가지는 행은 삭제	상환이 완료된 시점에서 정상적으로 상환이 완료된 케이스와 부도가 난 케이스만을 남기고 나머지 클래스는 제외함
home_ownership, verification_status	One-HotEncoding (Dummy Variables)	칼럼값 확인 결과 차원이 크게 늘어나지 않는다고 판단하여 One-Hot Encoding

[Step 06] 수치형 변수

칼럼명	처리방법	이유
revol_util, int_rate	"%"기호 제거 후 실수형 변환, 100으로 나눠 소수점 실수값으로 표기. 결측치는 평균값으로 대체	현재 %기호를 포함한 문자열로 되어 있음. 퍼센티지 데이터 값은 소수점실수값으로 표기
tot_coll_amt, tot_cur_bal,	결측치를 각각 0, 중앙값 등 적절한 값으로 대체	

total_rev_gi_lim, acc_open_past_24mths, avg_cur_bal, bc_open_to_buy		
bc_util	값을 100으로 나눠 소수점 형태로 변환한 후, 결측치는 평균값으로 대체	퍼센티지 데이터 값은 소수점 실수값으로 표기
dti	100으로 나눠 소수점 실수값으로 변환	퍼센티지 데이터 값은 소수점 실수값으로 표기
mort_acc, num_accts_ever_120_pd	결측치는 중앙값으로 대체	+ 또는 - 왜도 형태를 띄는 칼럼이므로 중앙값으로 대체
여러 계좌 관련 칼럼 및 신용 한도, 잔액 관련 칼럼	결측치는 중앙값 또는 평균값으로 대체	

[Step 07] 파생변수

칼럼명	처리방법	이유
fico_avg	(fico_range_high + fico_range_low) / 2	두 변수의 차이(fico 점수 범위)가 매우 작음. 또한 두 변수가 비슷한 정보를 제공함으로 하나의 변수로 줄여 차원 축소

[Step 08] 결측값 처리

Train set, Test set의 결측치가 있는 칼럼을 결측치 개수, 결측치 비율을 계산해 각각 출력한 후 결측치 비율을 내림차순해서 보면, train dataset과 test dataset 모두 약 45%의 결측치를 보이는 칼럼을 마지막으로 이후 칼럼의 결측치 비율은 10% 미만으로 줄어든다. 두 데이터 셋의 결측치가 40% 이상인 공통 칼럼의 절대적인 결측치 개수가 30만개 이상이기 때문에, 결측값을 채워 넣어 학습에 사용하는 것은 부적절하다고 판단하였고 그리하여 제거하였다. 또한 남은 칼럼들의 공통 결측치 개수가 10만개 미만이므로 결측치가 있는 행을 제거하여 결측치 문제를 해결하였다.