흙속의 진주,
저신용자 중 우량고객을 찾아라

머신러닝 기반 대출 상환 예측 모델 개발

멘토 | SK C&C 박병선
알파 | 김민지 김수인 안이찬 이지훈 피하영

SK planet SK 주식회사 C&C

CONTENTS

1장 프로젝트 개요 |

01 주제 및 배경

제 1금융권 대출의 높은 문턱으로 중 · 저 신용자들은 금융 혜택의 사각지대에 놓여 ..

226만명 신용회복해도 줄줄이 '대출 거절'...

해럴드 경제 ooo 기자 2024년 5월 5일

저신용자였다가 신용사면으로 중신용자가 된 소비자도 카드 발급이 어렵거나, 금융소외 계층에 속한 사람들이 줄줄이 1금융권 대출이 거절되면서 소비자들은 다시 제도권 밖으로 밀려나는 모양새다.

.
.
.

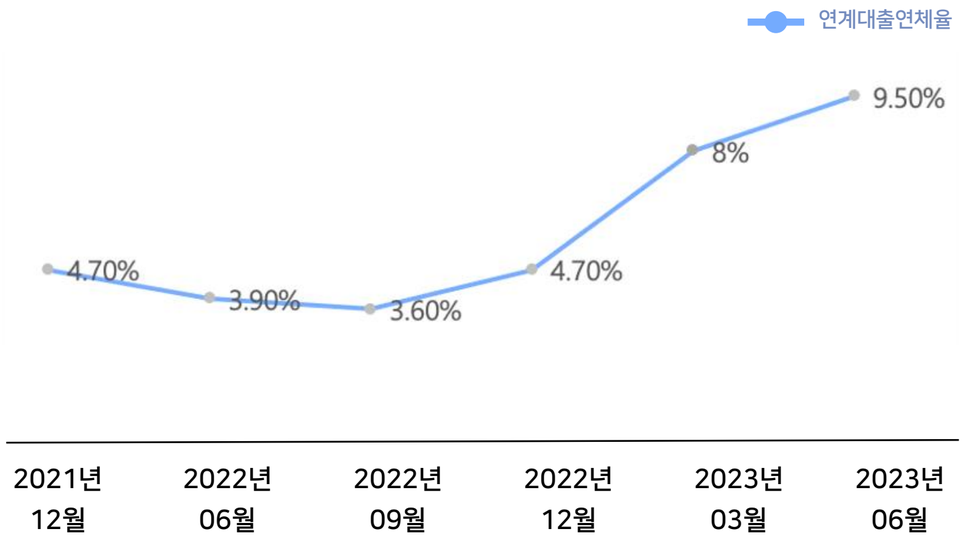
1금융권 대출 문턱이 높아지면서 신용사면 이후에도 2금융권을 다시 찾는 소비자들이 늘고 있다.

출처: https://biz.heraldcorp.com/view.php?ud=20240504050116

*** 금융소외계층 이란?**
적은 금융거래 실적으로 인해 불리한 신용평가를 받는 '신 파일러'를 포함하여 저소득층, 중·저 신용등급자등의 계층으로 대출과 같은 신용거래에서 불리한 조건에 놓이는 사람들.

01 주제 및 배경

제 2금융권, P2P 플랫폼 등에서는 연체율 관리를 위한 정확한 고객 선별이 보다 중요



출처: https://biz.heraldcorp.com/view.php?ud=20240504050116

*** P2P(Peer-to-Peer)란?**
인터넷상에서 온라인 플랫폼을 통해 개인 투자자가 대출신청자에게 전통적인 금융기관을 거치지 않고 직접 대출을 할 수 있게 연결해주는 서비스

01 주제 및 배경

“

우·불량 고객 분류하는 예측력

개별 고객에 대응할 수 있는 해석력

대출 상환 예측 시스템 개발

”

02 프로세스

모델 개발 프로세스

프로젝트 기획	데이터 탐색	데이터 전처리	모델 학습 및 평가	최종 모델 선정
비즈니스 이해	기술 통계량 확인	이상치 처리	후보 모델 선택	best 모델 선택
타겟 정의	EDA 및 시각화	결측치 처리	후보 모델 학습	변수 중요도 확인
평가 지표 정의	가설 검정	피처 엔지니어링	모델 최적화	XAI 모델 해석
분석 대상 정의	변수 간 관계 파악	파생변수 생성	교차 검증	

단계별 세부 사항

• 여신 비즈니스 이해 • target 이진화 작업 • 선행연구 조사 • 변수 의미 파악	• histogram, boxplot 등 시각화를 통한 데이터 탐색 • pearson, spearman 상관관계를 통해 변수 간 관계 파악	• One hot Encoding • Min Max Scaling • Target Encoding • box-cox 변환 • 범주형 변수 그룹화	• RF, CAT, XGB, LGBM 후보 모델로 선정 • Optuna, GridSearch, RandomSearch를 통한 성능 최적화 • k-fold 교차검증 • 타겟 불균형 처리	• SHAP를 통한 모델 해석 • Best 모델의 feature importance 추출 • f1 score, AUC 등 확인

03 프로젝트 일정

주간단위 스터디(이론)와 과제(수행)를 동시에 진행하여 시너지 제고

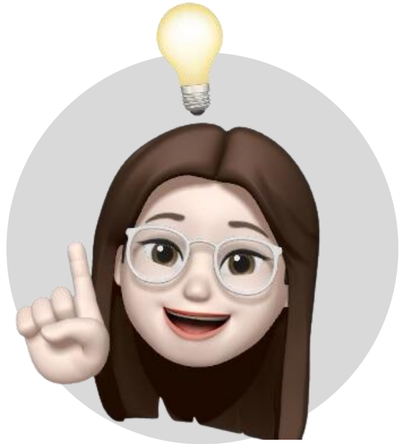
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
데이터 이해	140+개 변수 의미와 관계성 파악									
변수 선택			필터링			최종 변수 선택				
피처 엔지니어링			이상치 제거, 인코딩/스케일링, 파생변수 생성							
모델 개발						RF, XGB, LGBM, CAT, LR 학습				
XAI를 통한 모델 해석								SHAP		
대시보드 기획 및 구현								Tableau		
최종 발표 준비									PPT 제작 및 시연 준비	
세션 및 스터디		금융 도메인, 회귀분석 통계 이론, 머신러닝 모델, XAI, 피처 엔지니어링								

04 팀원 소개

팀장

김민지

EDA
전처리
모델링
XAI
PPT



팀원

김수인

EDA
전처리
모델링
XAI
PPT

팀원

안이찬

EDA
전처리
모델링
XAI
하이퍼파라미터 튜닝

팀원

이지훈

EDA
전처리
모델링
XAI
하이퍼파라미터 튜닝

팀원

피하영

EDA
전처리
모델링
XAI
대시보드 제작

2장 데이터 전처리 |

01) 데이터 설명

02) EDA

03) 변수 선택

04) 피처 엔지니어링

05) 최종 변수

01 데이터 설명

Lending club은 미국의 P2P 대출 플랫폼으로, 개인들이 은행을 통하지 않고 직접 대출과 투자가 가능

데이터셋은 2007년~2020년 **개인 별 정보**로 구성

타겟 변수를 포함하여 **총 141개의 변수**과 **약 300만 개의 행**으로 구성

개인 정보, 신용 기록, 대출 정보, 정책 자금 등으로 이루어져 있음

데이터셋 예시)	직업	근속년수	소득($)	거주 주	대출 목적	대출신청액($)		대출상태
	의사	10y+	255K	NY	주택	150K		상환
	경찰관	5y	65K	TX	부채 통합	20K	...	상각
⋮								
	요리사	~1y	30K	LA	자동차 구매	15K		상환

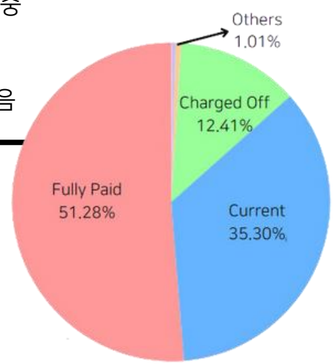
01 데이터 설명

Target 변환 전

Fully Paid	• 대출금 완납 상태
Charged Off	• 감가상각된 상태
Default	• 부도난 상태
Late	• 연체된 상태
In grace Period	• 유예기간에 있는 상태
Current	• 정상 상환 진행 중
Issued	• 대출이 발행 됨
Dose not meet the credit policy	• 정책에 맞지 않음

데이터 수

2,925,492



대출의 상태

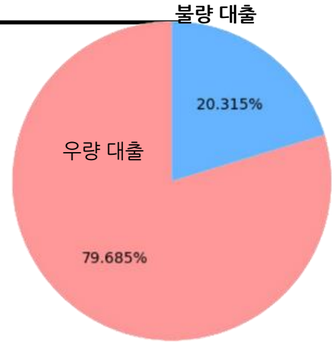
Target 변환 후

상환과 연체 사이의 상태는 삭제 후 이진화 작업 수행

우량 대출	• Fully Paid(완납)
불량 대출	• Charged off(상각) • Default(부도) • Late(연체)

데이터 수

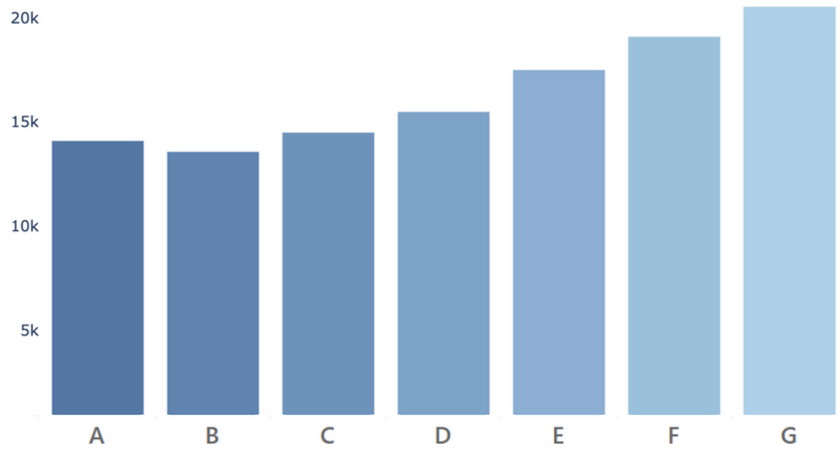
1,879,637



새로 정의한 대출의 상태

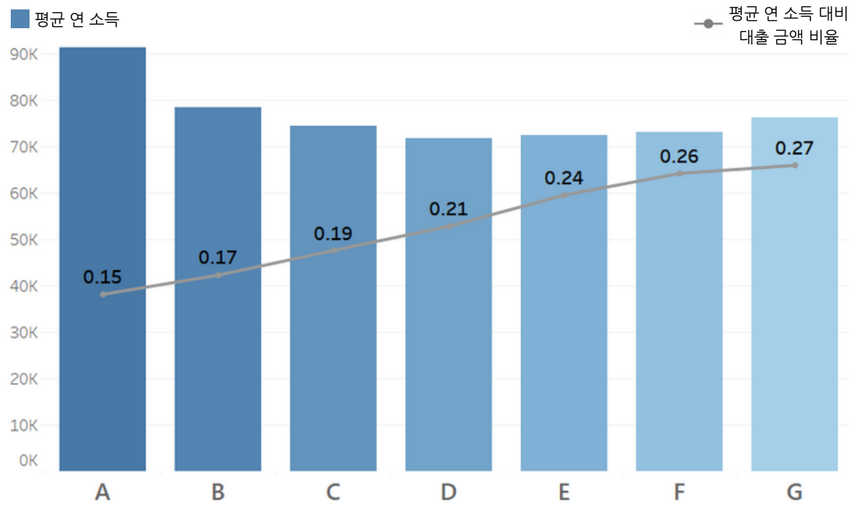
02 EDA 등급별 대출 금액 및 소득 대비 대출 비율 분석

등급별 평균 대출 금액



- 등급이 낮아질수록 대출 금액이 증가함

등급별 평균 연 소득 대비 대출 금액 비율

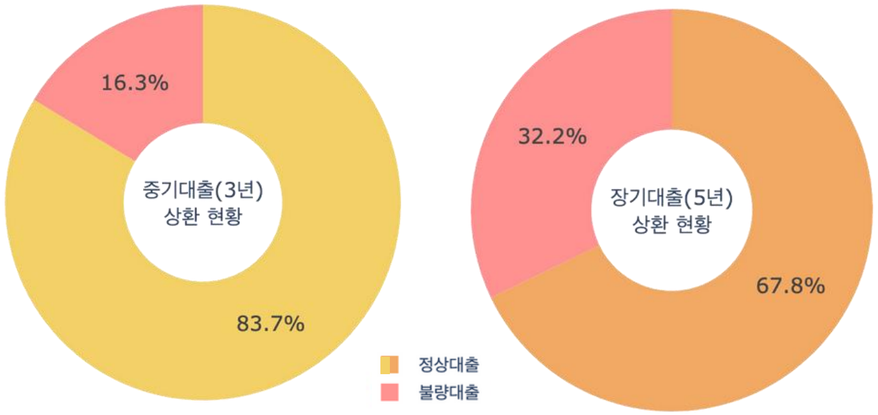


- 등급이 낮아질수록 연간 소득 대비 대출 금액이 높아짐

→ 저신용자가 오히려 소득 대비 과도한 대출 경향이 뚜렷함

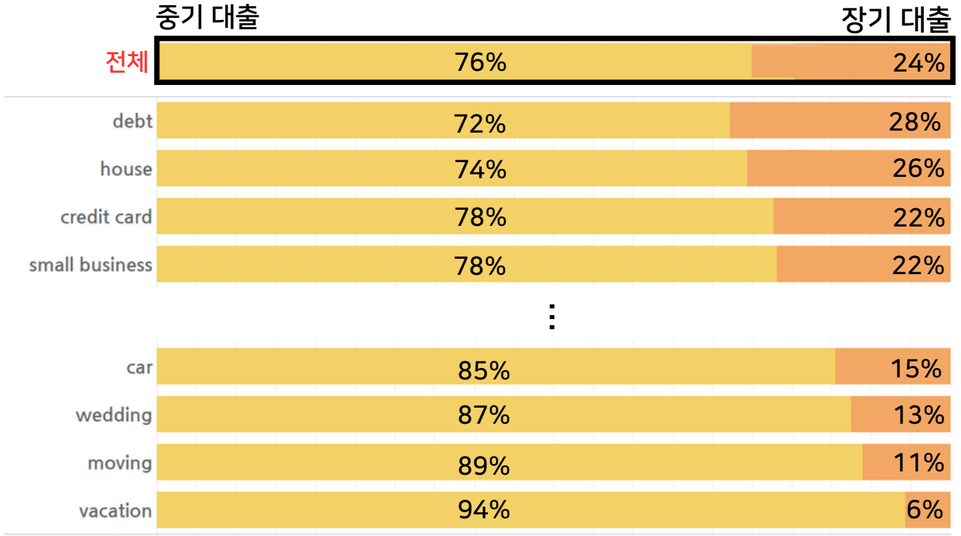
02 EDA 대출 기간 및 대출 목적 분석

대출 기간에 따른 대출 상태



- 장기 대출자가 불량 대출률 2배 높음

대출 목적에 따른 대출 기간 비율

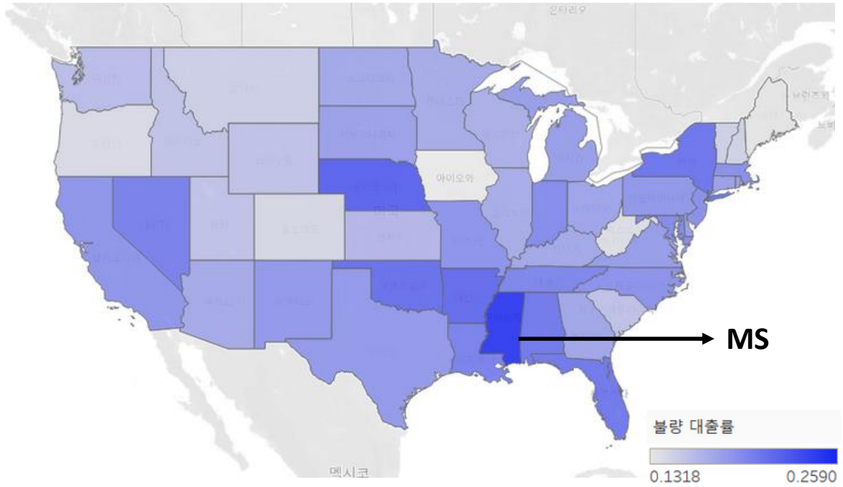


- 장기 대출이 선호하는 목적은 부채 대환, 주택 자금
- 중기 대출이 선호하는 목적은 차량 구매, 결혼

→ 목적마다 선호하는 대출 기간이 존재

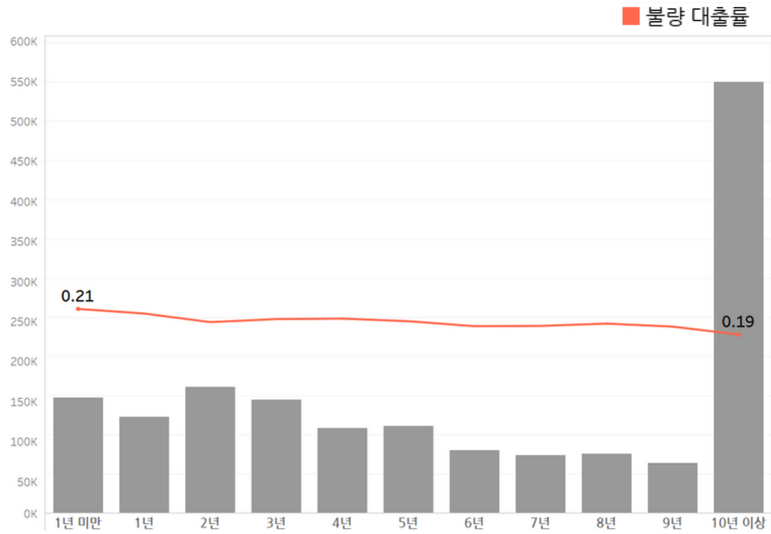
02 EDA 지역별 및 재직 기간별 불량 대출률 분석

지역별 불량 대출률 분포



- MS(미시시피) 불량 대출률이 가장 높음

재직 기간별 대출 건수 및 불량 대출률



- 재직 기간별 불량 대출률에 큰 차이가 없음
- 사회초년생 중에서도 우량 고객 발굴 가능성 有

→ 불량률에 있어, 지역은 유의미한 요인,
상식과 달리 사회 초년생과 장기 근속자의 유의미한 차이 없음

03 변수 선택

타겟을 제외한 140개 변수에 대하여 다음과 같은 4개의 기준으로 변수 필터링을 수행

1 결측값이 많은 변수	• 공동대출신청인 정보 및 정책자금 정보
2 카디널리티가 높은 변수	• emp title : 대출 신청인이 작성한 직업 정보 • 고객 ID
3 다른 변수에 종속인 변수	• grade 및 sub grade
4 Data Leakage 변수	• total pymnt : 현재까지 회수한 금액으로 대출발행 후 발생 • last fico : 최근 fico 신용점수로 대출발행 후 발생

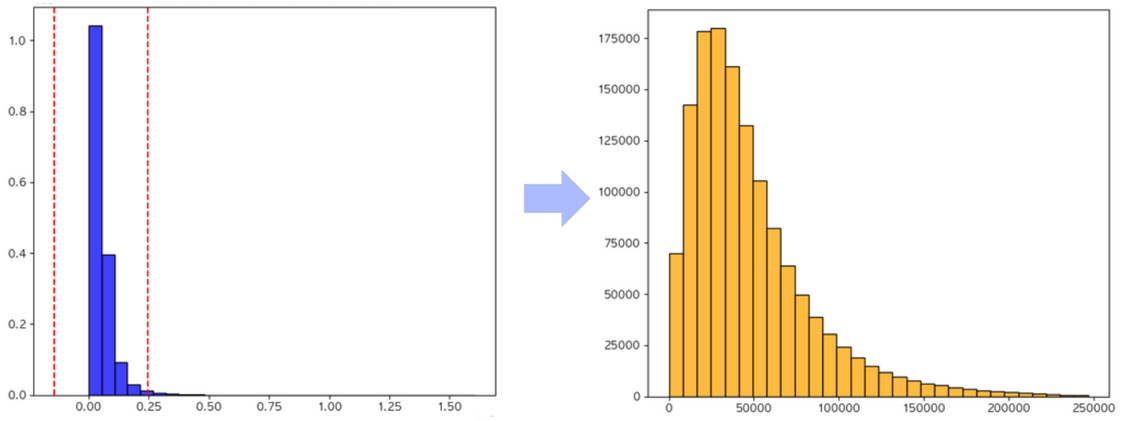
필터링 후 남은 변수

44개

04 피처 엔지니어링 수치형 데이터

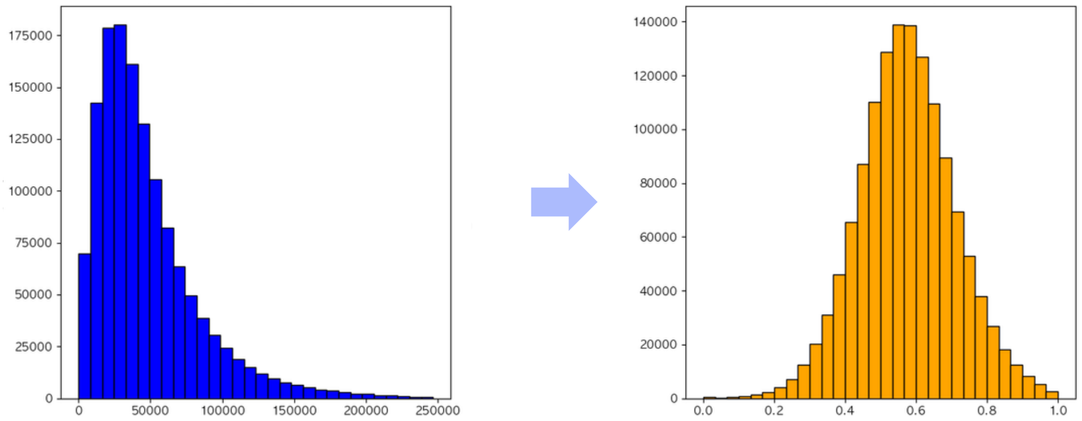
이상치 처리

z-score 기준으로 극단값을 제거하여
분포에 영향을 미치는 이상치를 처리



스케일링

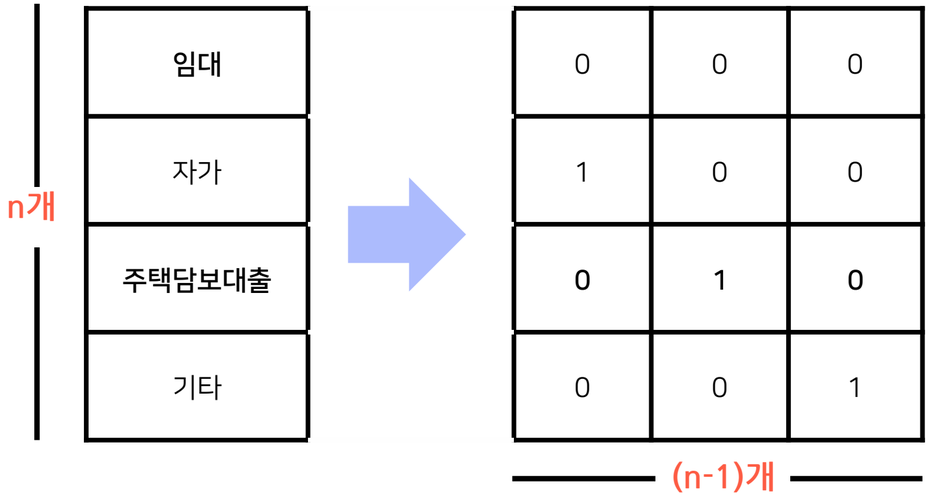
왜도가 심한 변수를 Box-Cox 변환으로 정규분포에 가깝게 변환한 후,
Min-Max 스케일링으로 동일한 범위로 조정



04 피처 엔지니어링 범주형 데이터

원 핫 인코딩

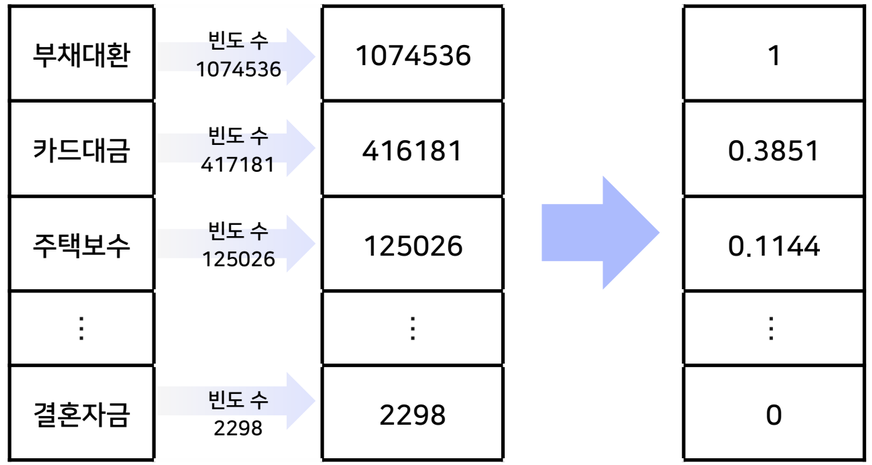
ex) 주거 형태



주거 형태, 소득 확인 유무, 대출 초기 상태 변수
원 핫 인코딩 수행

카운터 인코딩

ex) 대출 목적



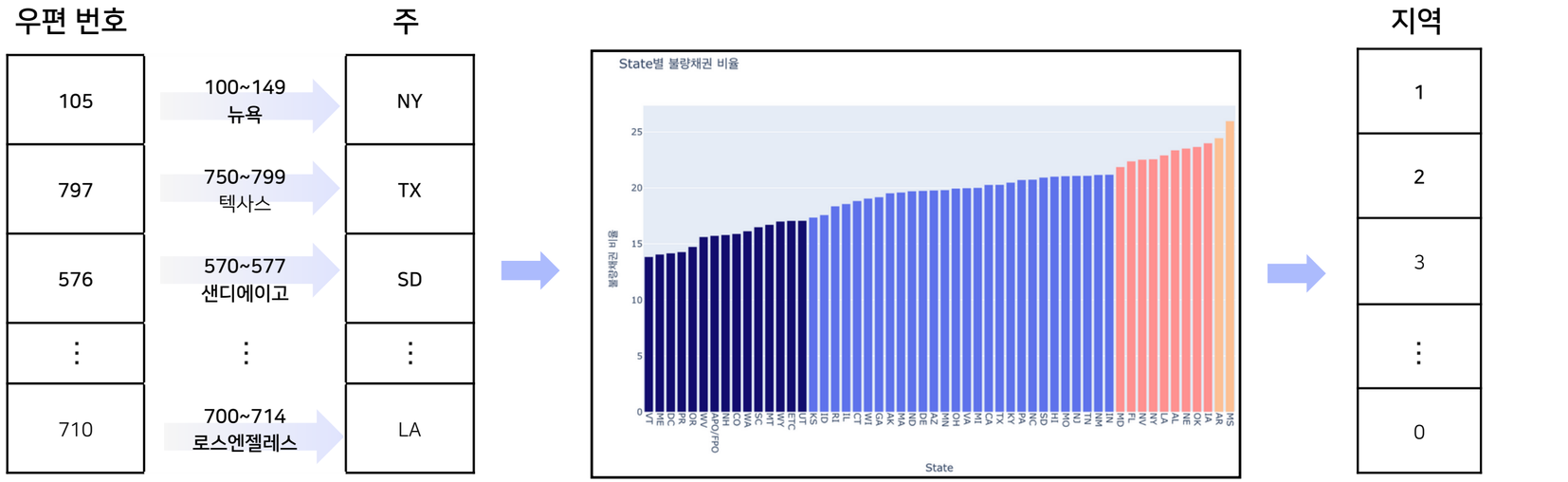
대출 목적 변수 카운터 인코딩 수행
Min-Max 변환

04 피처 엔지니어링 파생 변수 생성

고유값이 많지만 변별력이 있는 변수를 그룹핑하여 파생변수를 생성

ex) 지역

우편 번호와 거주 지역이 잘못 매칭 된 행들이 존재하여 새로운 지역 변수 생성



각 우편번호에 맞는 주를 생성

불량 대출률 순으로 정렬하여
변화가 뚜렷한 구간을 기준으로 그룹핑

총 4개
ordinal 인코딩 수행

05 최종 변수

변수 필터링 후 남은 44개 변수에 대해 다음과 같은 기준으로 최종 변수를 선택

정량적 평가

1 통계 기반 선택

- 변수들의 상관관계와 다중공선성을 분석
- F-검정 을 통해 통계적 유의성 파악
- 우량, 불량 판별을 위한 변수 예측력 평가방법 iv 활용
- 중요도가 낮은 변수를 제거하며 변수를 선택하는 RFE 방식 활용

2 모델 기반 선택

- LR, LDA를 통해 회귀계수 확인
- RF, XGB, LGB, Cat 모델의 feature importance 확인

정성적 평가

3 비즈니스 지식 기반 선택

- 비즈니스 목표와 직접 연관된 이자율, 대출 금액과 같은 대출 관련 정보 선택

최종 선택된 변수

22개

05 최종 변수

대출 관련 정보 (4)

변수명	설명
int_rate	이자율
funded_amnt	대출 승인액
term	상환기간
purpose	대출 목적

대출신청인의 개인정보 (5)

변수명	설명
emp_length	고용 기간
home_ownership	주택 소유 상태
verification_status	소득 확인 상태
state	지역
annual_inc	연간 소득

대출신청인의 신용정보 (13)

변수명	설명
dti	소득 대비 부채 비율
mo_sin_old_il_acct	최초 할부 계좌 개설 경과 기간
mo_sin_old_rev_tl_op	최초 리볼빙 계좌 개설 경과 기간
num_tl_op_past_12m	지난 12개월간 개설된 계좌 수
num_op_rev_tl	리볼빙 계좌 수
num_rev_tl_bal_gt_0	잔액이 있는 리볼빙 계좌 수
revol_bal	리볼빙 계좌 총 잔액
tot_hi_cred_lim	리볼빙 계좌 최대 신용 한도
mort_acc	모기지 계좌 수
fico_range_high	Fico 점수
revol_util	리볼빙 신용한도 사용률
percent_bc_gt_75	고한도 사용률
months_difference	신용계좌 개설 후 개월 수

3장 모델링

01) 데이터 불균형 처리

02) 모델 및 평가지표 선정

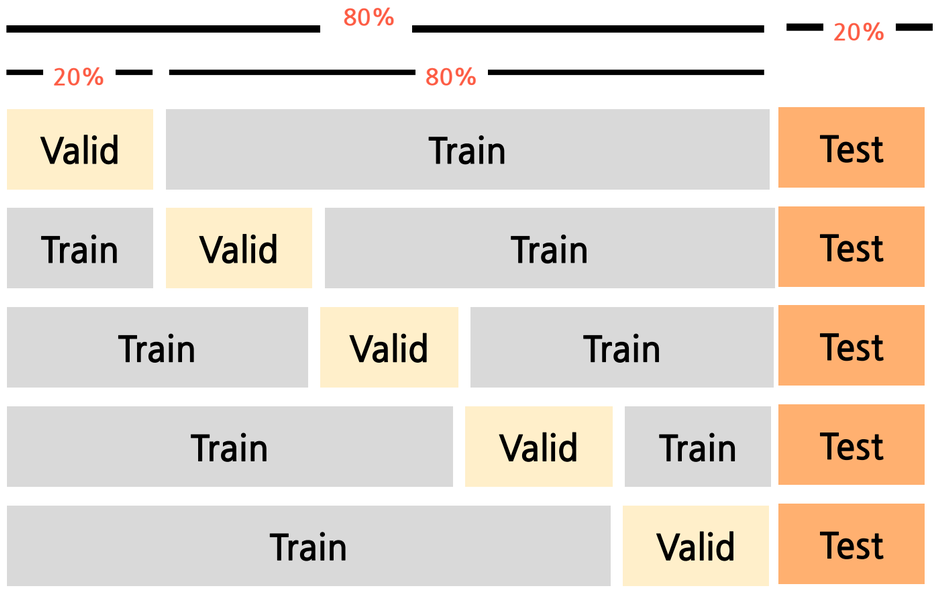
03) 최종 모델 선택

04) 모델 해석

01 데이터 불균형 처리

데이터 과적합 방지를 위한

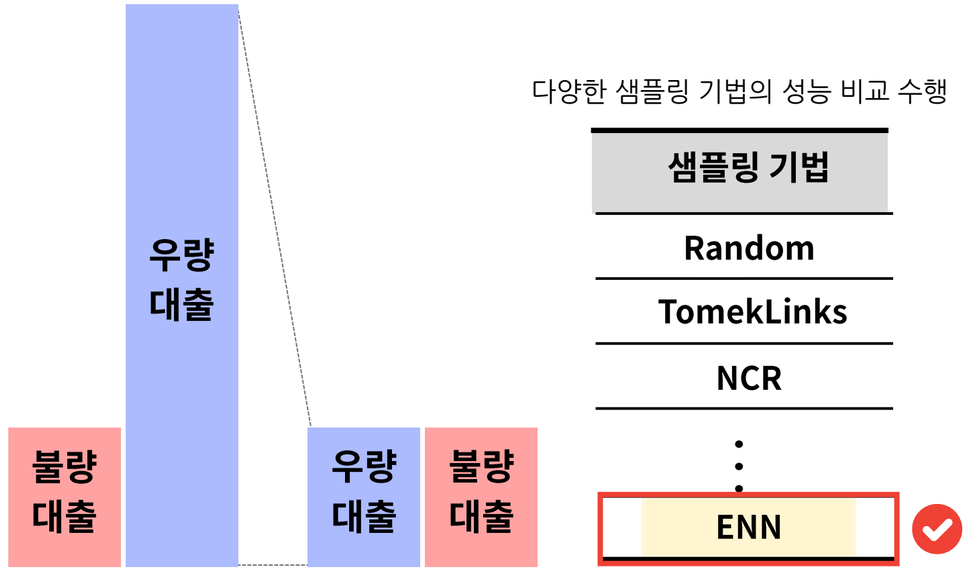
데이터 분할



K-Fold 교차 검증으로
과적합 방지 및 모델의 일반화 성능 향상

데이터 불균형 해소를 위한

데이터 샘플링



다양한 샘플링 기법의 성능 비교 수행

샘플링 기법
Random
TomekLinks
NCR
⋮
ENN

데이터 정보를 잃지 않기 위해
ENN 언더 샘플링 기법 적용

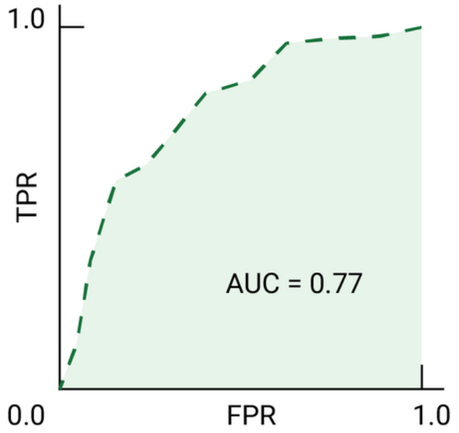
02 모델 및 평가지표 선정

고차원 변수 셋에 적합한
후보 모델

RANDOM FOREST
XGBoost
CatBoost
LightGBM

안정된 예측력의 트리기반 앙상블 알고리즘이 주 후보군
실험 결과, 예측력이 떨어지는 선형 모델은 제외

균형잡힌 시각을 제공하는
평가지표



- **TPR(True Positive Rate) = Recall**
  실제 불량 대출 중 모델이 맞게 예측한 비율
- **FPR(False Positive Rate)**
  실제 우량 대출 중 모델이 잘못 예측한 비율

AUROC가 1에 가까워질수록 모델이 우수함을 나타냄
특정 임계값에 의존하지 않고 모델의 전반적인 성능을 평가

03 최종 모델 선택

모델 성과

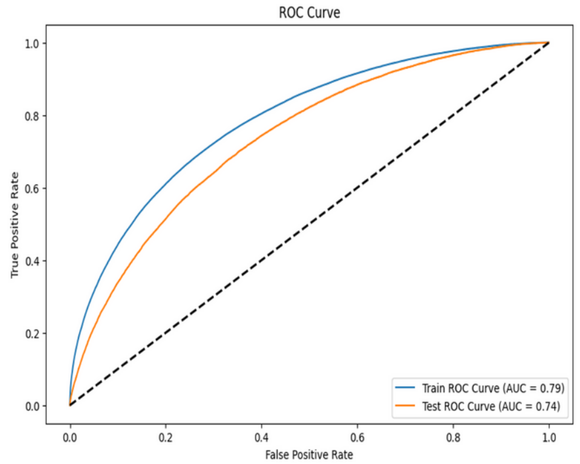
- 베이스 모델은 언더샘플링을 진행하지 않고, 전처리된 데이터만 활용하여 학습
- 오른쪽 결과표는 언더샘플링과 하이퍼파라미터 튜닝 후의 결과
- 하이퍼파라미터 튜닝방식은 GridSearch/RandomSearch와 Optuna 활용

	AUC	F1-score
Random Forest	0.709	0.421
XGBoost	0.715	0.434
LightGBM	0.711	0.425
CatBoost	0.719	0.448

베이스 모델 성능

언더샘플링
하이퍼파라미터 튜닝

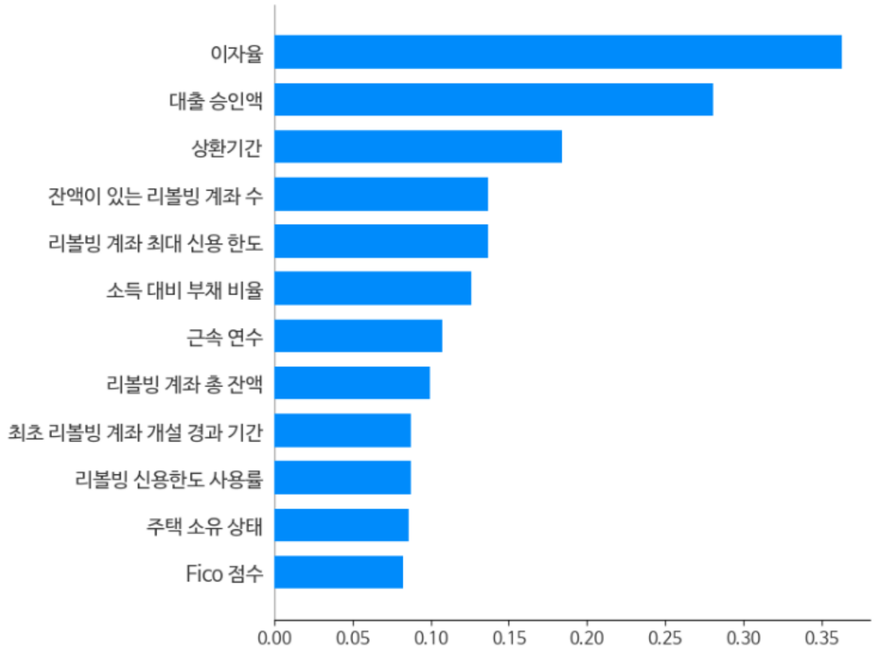
	AUC	F1-score
Random Forest	0.728	0.478
XGBoost	0.731	0.486
LightGBM	0.731	0.485
CatBoost	0.744	0.487



최종 모델 성능

03 최종 모델 선택

상대적인 기여도로 살펴본 변수 중요도



순위	변수	변수 중요도
1	이자율	0.37
2	대출 승인액	0.28
3	상환기간	0.18
4	잔액이 있는 리볼빙 계좌 수	0.14
5	리봉빙 계좌 최대 신용 한도	0.14
6	소득 대비 부채 비율	0.12
7	근속 연수	0.11
8	리볼빙 계좌 총 잔액	0.10
9	최초 리볼빙 계좌 개설 경과기간	0.09
10	리볼빙 신용한도 사용률	0.09
11	주택 소유 상태	0.09
12	Fico 점수	0.08

신용점수가 우량대출고객과 불량대출고객을 구분하는 절대적인 기준은 아니었음

03 최종 모델 선택

참고) Data Leakage 변수를 사용한 잘못된 연구 사례

Resulting features are: "recoveries", "total_rec_prncp", "collection_recovery_fee", "last_fico_range_high", "last_fico_range_low", "last_pymnt_amnt", "total_pymnt_inv", "total_pymnt", "funded_amnt", "installment", "loan_amnt", "funded_amnt_inv", "debt_settlement_flag_Y", "debt_settlement_flag_N", "total_rec_int", "term", "total_rec_late_fee", "int_rate", "issue_d", "grade_A".

We eliminate recovery-related variables because they are a trivial explanatory for defaulted or delinquent loans. LendingClub charges fees for recovery of principal and interest rate for late or no payments. Also, we delete redundant features with a correlation coefficient over 0.80. We maintain the following variables for further analysis:

Table 1. Selected features description.

Feature name	Description	Type
last_fico_range_high	The upper boundary ranges the borrower's last FICO pulled belongs to.	Numeric
last_pymnt_amnt	Last total payment amount received	Numeric
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	Numeric
debt_settlement_flag_Y	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.	Categoric: 0 for 'YES,' 1 for 'NO'
total_rec_int	Interest received to date	Numeric
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	Numeric
int_rate	Interest Rate on the loan	Numeric
issue_d	The month which the loan was funded	Numeric
grade_A	LC assigned loan grade	Categoric: 1 for grade A, 0 for other grades

Source: From the LendingClub data dictionary

출처: Loan Default Prediction: A Complete Revision of LendingClub

동일한 데이터를 활용한 많은 선행 연구 논문에서 Data Leakage 변수 사용

- last_fico_range_high : 대출 시점 이후, 최근 신용 점수
- last_pymnt_amnt : 대출 시점 이후, 최근 상환 금액

이는 **예측 시점에서 얻을 수 없는 정보이기** 때문에 잘못된 모델링 작업

	최종 모델 + Data Leakage 변수	논문 성능 (Decision Tree)
AUROC	0.97	0.971

Data Leakage 변수로 인해 과도한 예측력을 보임

03 최종 모델 선택

참고) Data Leakage 변수를 사용하지 않은 연구 사례

〈표 2〉 변수 설명

변수 유형	변수 명	설명	속성
종속변수	Loan status	현재 대출 상태	범주형
독립변수	Address state	대출 신청서 상 차주의 거주 지역	범주형
	Annual income	차주가 자체 보고한 연간 소득	수치형
	Application type	개인 대출/공동 대출 여부	범주형
	Dept to income	주택담보대출, LC대출을 제외한 총 채무에 대하여 차주의 월 부채 상환액을 자진신고한 월 소득으로 나누어 계산한 비율	수치형
	Earliest credit line	차주의 신용한도가 개설된 최초 월	범주형
	Employment length	고용 기간 (0~10 사이의 수)	수치형
	FICO range high	대출 개시 시 차주의 FICO 상위 경계범위	수치형
	FICO range low	대출 개시 시 차주의 FICO 하위 경계범위	수치형
	Home ownership	주택 소유 상태 (임대, 소유, 저당, 기타)	범주형
	Initial list status	대출의 초기 상장 상태	범주형
	Installment	대출 발생 시 차주가 지불해야 하는 월별 금액	수치형
	Interest rate	대출 이자율	수치형
	Issue date	대출 시행 월	범주형
	Loan amount	차주에게 대출된 총 금액	수치형
	Mortage account	모기지 계좌 수	수치형
	Open account	차주의 신용보고서에 있는 개방된 신용계좌 수	수치형
	Public records	부정적인 공공 기록의 수	수치형
	Public record bankruptcies	공공 기록의 파산 횟수	수치형
	Purpose	차주의 대출 신청에 따른 대출 구분	범주형
	Revolving Balance	총 리볼빙에 대한 미지급 잔액	수치형
	Revolving Utilization	총 리볼빙에 대한 사용자 이용률	수치형

출처 : https://koreascience.or.kr/article/JAKO202328958570439

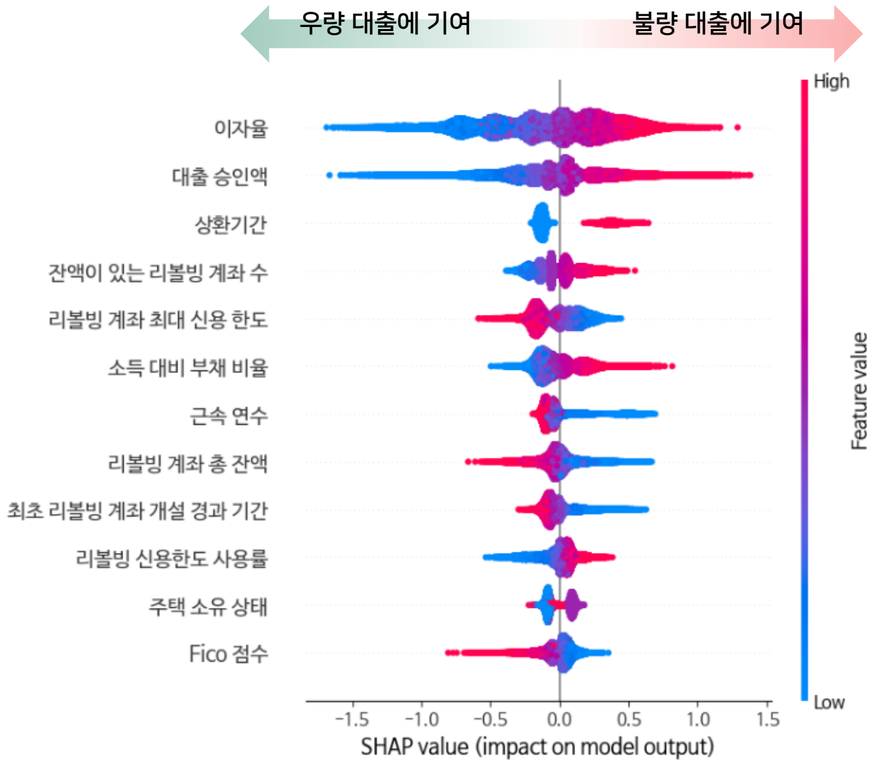
	최종모델	논문 성능 (머신러닝)	논문 성능 (딥러닝)
AUROC	0.744	0.721	0.770

동일한 데이터를 활용한 다른 논문의 머신러닝 모델 성능보다 우수

딥러닝 모델 성능은 약간 높지만, 해석성이 없음

04 모델 해석

XAI를 통해 변수 변화에 따른 해석성 (방향+강도)을 제공



SHAP란?

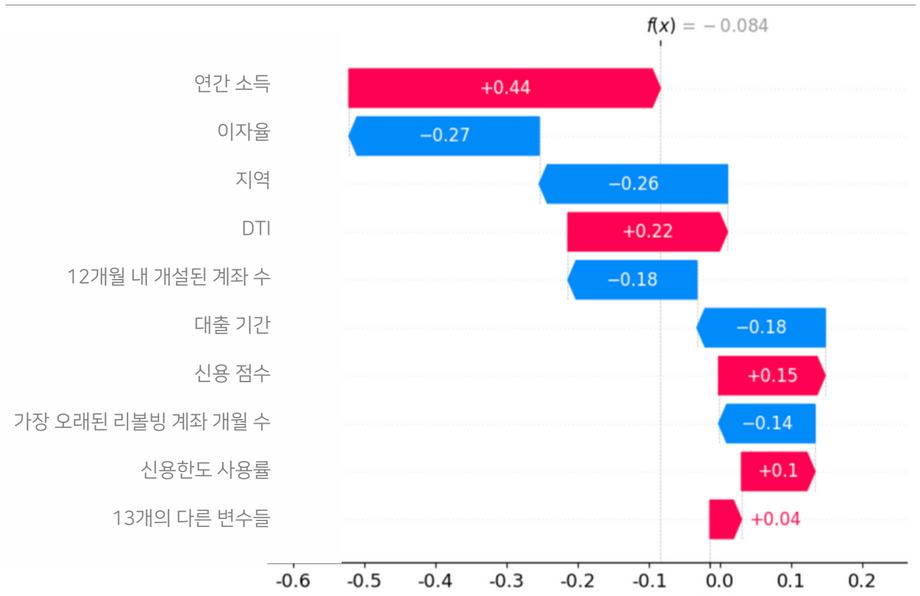
변수 간의 상호작용을 고려하여, 모델의 예측에 어떻게 기여하는지를 설명하는 XAI 방법론

- 모델 전체에서 어떤 변수가 중요한지를 파악
- 색상은 값의 크기를 나타냄
  - 파란색은 feature 값이 낮음을 의미
  - 빨간색은 feature 값이 높음을 의미
  - 이자율이 높을수록 불량 대출에, 낮을수록 우량 대출에 기여
  - 신용한도가 낮을수록 불량 대출에, 높을수록 우량 대출에 기여
- 변수가 상위에 존재할 경우 타겟에 대한 영향력이 큼

XAI 를 통해 불량 대출에 기여한 변수를 쉽게 파악 가능

04 모델 해석

XAI를 통해, 개별 고객에 대한 해석성 제공



- 연간 소득 : 하위 8%
- 신용 점수 : 하위 20%
- DTI : 하위 28%
- 신용한도 사용률 : 84%

- 특정 데이터 포인트에 대한 모델 예측이 어떻게 이루어졌는지 파악
- 색상은 feature 가 타겟에 미치는 영향
  - 파란색은 우량 대출를 예측하는데 영향을 끼침
  - 빨간색은 불량 대출를 예측하는데 영향을 끼침
- f(x)는 feature SHAP값의 합
  - f(x) ≤ 0 이면 우량 대출로 예측
  - f(x) > 0 이면 불량 대출로 예측

신용 상태가 좋지 않아 대출 거절이 예상되지만,
실제로는 우량 대출 고객인 사례

4장 결론 및 기대효과 |

01) 결론

02) 대시보드

03) 기대효과

04) 한계점 및 향후 과제

05) 회고록

01 결론

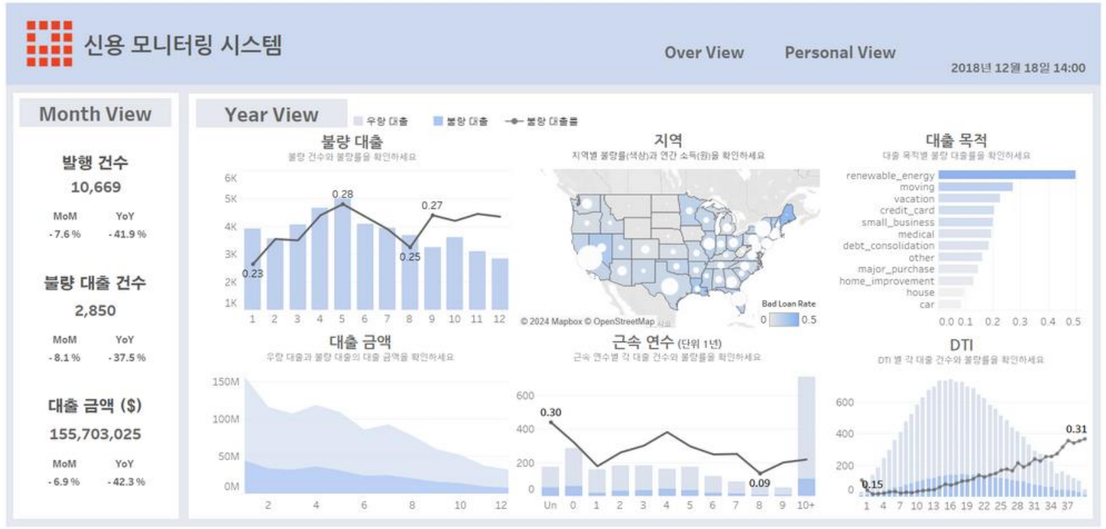
- 신용도가 낮은 P2P 고객군의 특성상 건전성이 떨어지는 문제점 발견
  예측력과 해석력을 갖춘, 대출 상환 예측 시스템 필요
- Lending club 데이터에서 22개의 변수를 활용, 최종 성능 AUROC 0.74의 모델 개발
- 신용 점수가 대출 상환 여부를 판단하는 절대적인 기준은 아님
  개별 고객 맞춤형 분석으로 해석성 제공

02 대시보드

목적 : 신용 리스크를 체계적으로 관리하고, 데이터를 기반으로 한 최적의 의사결정을 기여

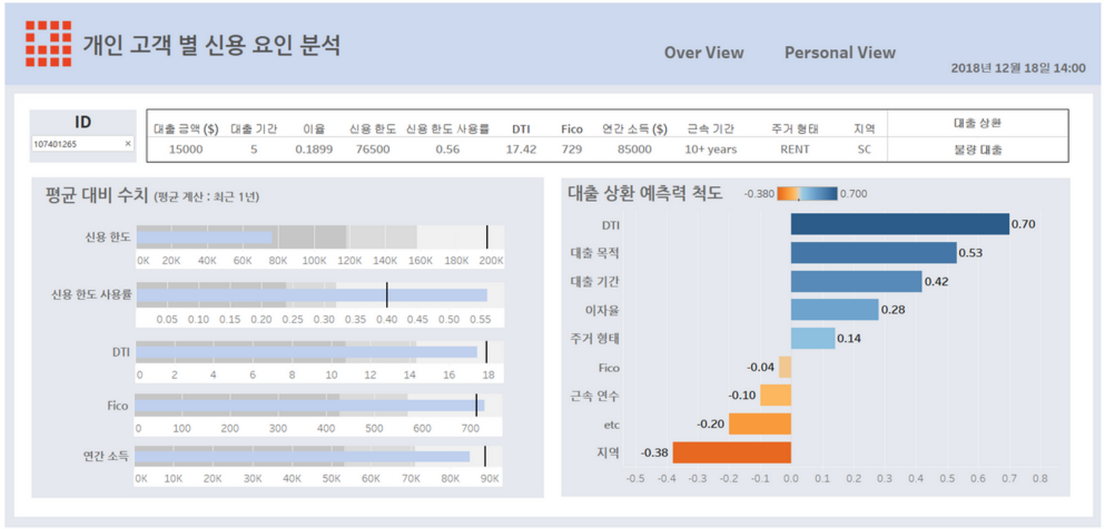
대상 : 대출 신용 회사

1. 신용 모니터링 시스템



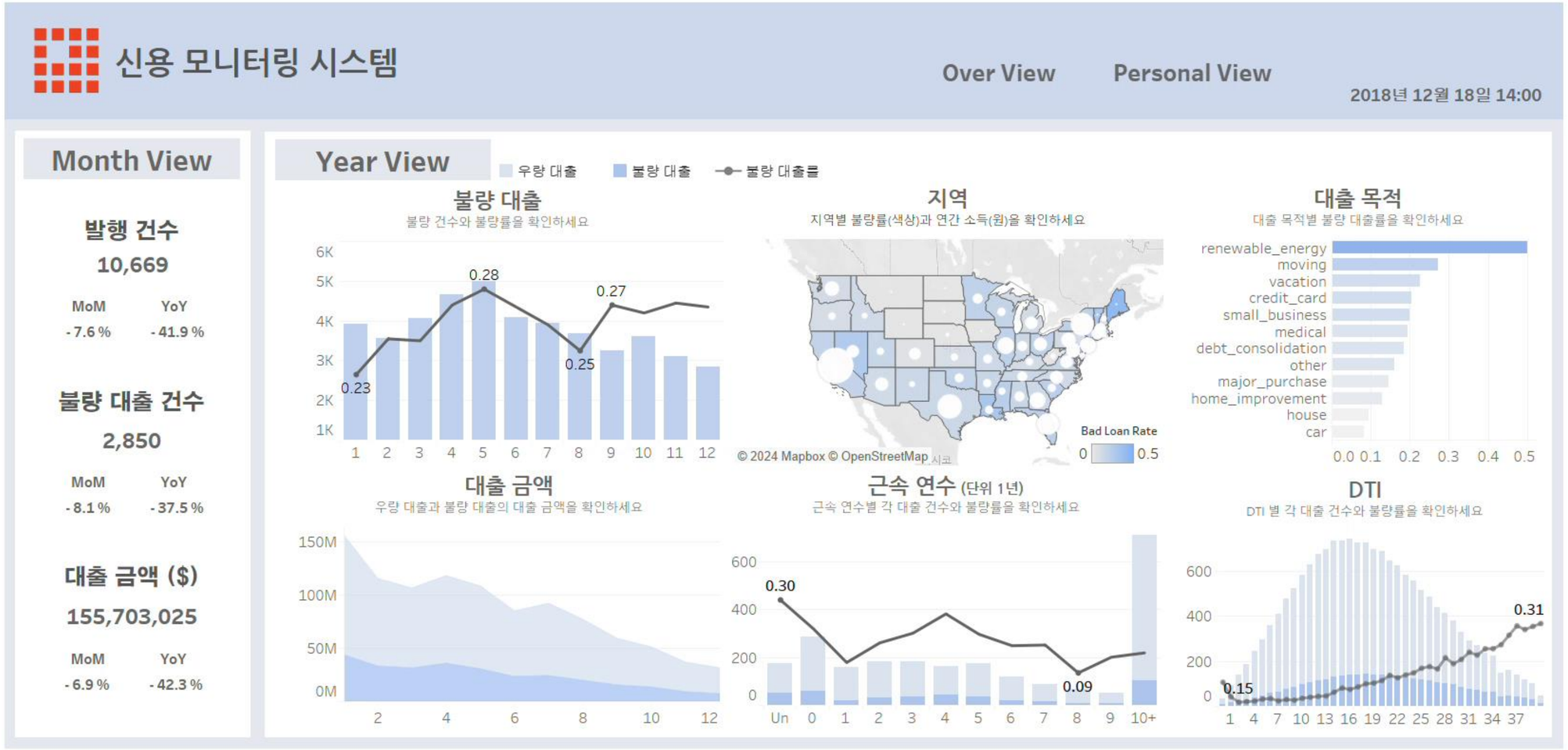
대출 포트폴리오 상태를 종합적으로 모니터링

2. 개인 신용 요인 분석



고객의 신용 프로파일을 분석해 효과적으로 관리

02 대시보드


신용 모니터링 시스템
Over View
Personal View
2018년 12월 18일 14:00
Month View
발행 건수
10,669
MoM
-7.6 %
YoY
-41.9 %
불량 대출 건수
2,850
MoM
-8.1 %
YoY
-37.5 %
대출 금액 ($)
155,703,025
MoM
-6.9 %
YoY
-42.3 %
Year View
우량 대출
불량 대출
불량 대출률
불량 대출
불량 건수와 불량률을 확인하세요
0.23
0.28
0.25
0.27
지역
지역별 불량률(색상)과 연간 소득(원)을 확인하세요
Bad Loan Rate
0
0.5
© 2024 Mapbox © OpenStreetMap
대출 목적
대출 목적별 불량 대출률을 확인하세요
renewable_energy
moving
vacation
credit_card
small_business
medical
debt_consolidation
other
major_purchase
home_improvement
house
car
대출 금액
우량 대출과 불량 대출의 대출 금액을 확인하세요
근속 연수 (단위 1년)
근속 연수별 각 대출 건수와 불량률을 확인하세요
0.30
0.09
DTI
DTI 별 각 대출 건수와 불량률을 확인하세요
0.15
0.31

03 기대 효과

금융 포용성 확대

신용 이력이 부족한
금융소외계층 인구에게도
금융 서비스 제공 가능

XAI를 통한 해석성 제공

해석성을 제공함으로써, 고객에게
대출 승인 거절 사유를 명확하고
투명하게 제시

리스크 관리

대출 상환 불이행의 가능성을
줄이고, 안정적인 리스크 관리를
통한 투자자 확보

04 한계점 및 향후 과제

한계점

1. 변수 추출 시점에 대한 명확한 정의 부족으로 변수 정의의 정확성이 떨어짐
2. 일부 변수에서 -1과 999 값의 의미 해석이 어려움
3. 시의적절한 경제 상황 반영의 부족

향후 과제

1. 특수한 상황에만 작용하는 변수 (정책 자금, 공동 대출인 정보) 를 새로운 모델로 구축 성능 향상 기대
2. P2P 사업의 주 고객층인 저·중 신용자를 위한 대안 변수 데이터를 추가하여 대안 신용 평가 모형 구축, 대안 요소 연구 및 적용 방안을 탐색

05 회고록

김민지

데이터 EDA 과정에서 변수 간의 관계성을 면밀히 살펴보며 데이터 전처리 및 변수 선택 과정에서 오류를 방지할 수 있는 방법을 체득했다. 데이터의 불균형을 해소하기 위한 샘플링 기법을 적용하고 불균형 데이터로 학습시킨 모델을 올바르게 평가하는 방법을 습득할 수 있었으며, 전반적인 데이터 분석 스킬을 성장시킬 수 있는 시간이었다.

김수인

모델링 과정에서 성능 향상이 쉽지 않아 많은 어려움을 겪었지만, 그 과정을 통해 선행 연구들과 차별성을 갖춘 신뢰성 있는 모델을 완성했다는 점에서 큰 보람을 느꼈다. 특히, 스터디와 세션을 통해 그동안 미뤄왔던 통계 이론 공부를 다시 시작하게 되었고, 스스로 성장하고 있음을 느낄 수 있는 값진 시간이었다.

안이찬

lending club 데이터를 다룬 다른 프로젝트는 많지만 우리 만큼 상세히 EDA를 하고 의미들을 탐색하며 모델링한 프로젝트는 없을 것이라 자부한다. 데이터의 한계로 인해 많은 노력을 기울였음에도 모델의 성능을 높이지 못한 것은 아쉽지만 그 과정에서 많은 스킬들을 알아보고 활용하는데 도움이 되었다.

이지훈

낮은 성능을 개선하기 위해 여러 모델튜닝을 시도하며 알고리즘, 샘플링 기법, 평가지표, XAI 등을 깊게 공부하게 되어 유익했다.
하지만 드라마틱한 성능 향상을 위해서는 모델튜닝에 의존하기 보단 첫단계로 돌아가 적절한 데이터셋을 다시 구축하는 것이 더 중요하다는 것을 다시 한번 깨달았다.

피하영

EDA 과정에서 도메인 지식은 데이터의 맥락을 이해하고, 의미 있는 인사이트를 도출하는 데 필수적임을 깨달았다.
특히 금융권 데이터를 다루면서 시행착오를 통해 이 지식의 중요성을 더욱 깊이 인식하게 되었다.
앞으로 이 분야에 대한 지식을 쌓아 도메인과 관련된 소프트 스킬도 함께 길러야겠다.

THANK YOU