리스크 확산 방지를 위한
가계부채 누증과 채무 불이행 관계 분석

IBAS 알파테스터
중소기업 막내아들

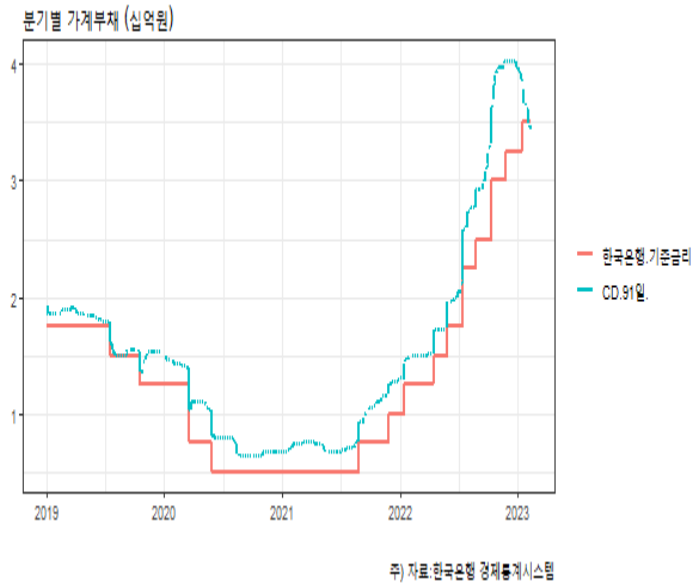
팀장
12183010 김호균

팀원
12181901 고영호
12192228 박지민

CONTENTS

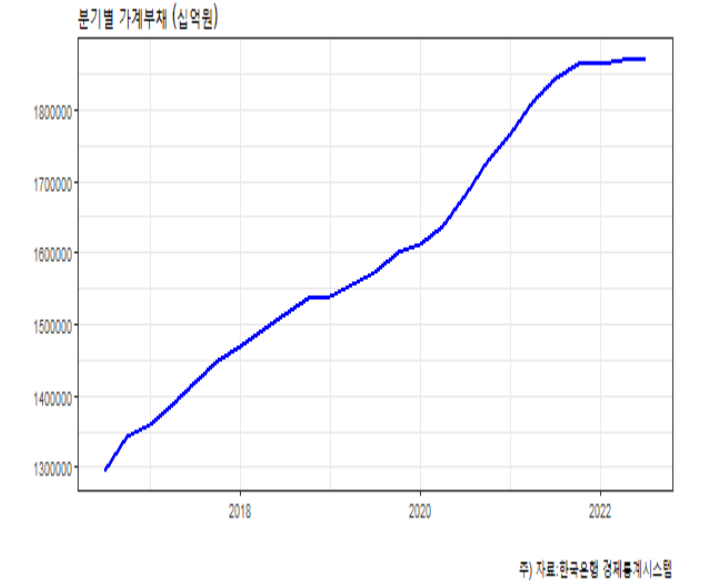
연구 동기

고물가 저성장 시대에서 경기 회복을 제한하는 가계 부채



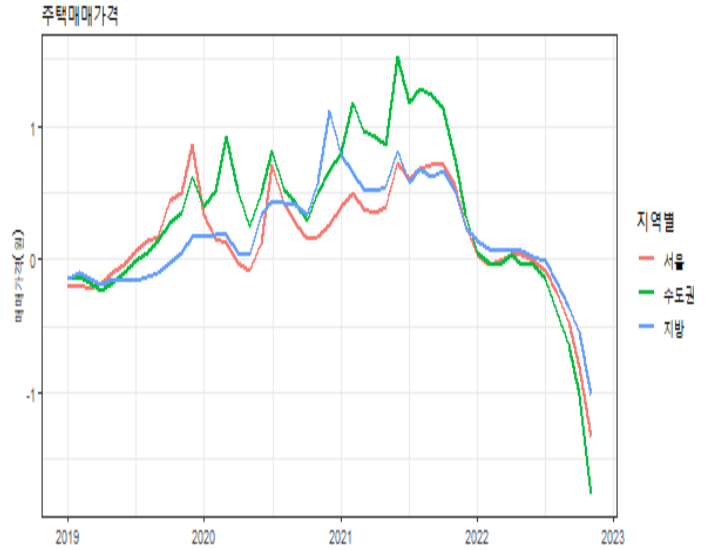
출처: 한국은행 경제통계시스템

한국은행의 금리인상 기조가 이어졌다. 지속된 높은 금리로 부채 증가 속도는 줄었으나, 여전히 하락세는 아니며, 코로나19로 부채는 크게 증가했다.



출처: 한국은행 경제통계시스템

지속적인 가계부채 누증으로, 한국의 CDP 대비 가계부채는 105.6%로 43개국 중 43위이다. (BIS) 가계부채는 금융시스템의 리스크를 늘리고, 경기의 하방 압력으로 작용해 통화정책을 제한하여 경기회복을 제한하고 물가 안정을 늦출 수 있는 요소이다.



출처: 한국은행 경제통계시스템

금리 인상 기조에 따라 주택 가격이 하락하고 이자부담이 늘면서, 가계 대출 상환 부담 증가가 크게 증가하고 있다. 이는 소비 감소로 이어지고 있다.

분석 목표

가계부채

경제 성장 하방 압력
물가 안정 방해 요소
금융시스템 위험 증가 요인

경기 안정을 위해 주목해야 할 요소

채무 불이행 문제

채무 이해당사자들에게
유동성 위험을 전파하기 때문에
금융 리스크가 급증

채무 불이행 위험 예측 모형

대출 신청자 정보를 통해 모형 수립
상존하는 위험을 현실화시키지 않기 위해
현재 가계 대출 상황 분석

데이터 수집

Home Credit Default Risk 데이터 (미국 자료)

ID	Target	Gender	Flag_ own_car	...	AMT_ CREDIT	DAYS_ BIRTH
100002	1	M	N		406597.5	-9461
100003	0	F	N		1293502.5	-16765

이 데이터는 이후에 credit risk 데이터로 부를 예정

307511 x 120

Finda 데이터 (한국 자료)

Application – Id	user_id	Gender	is_applied	...	yearly_ income	birth_year
1019382	186886	0	1		9.5e+07	1979
1117343	594274	1	0		3.5e+07	1993

377506 x 17
신청 승인된 정보만을 인덱싱

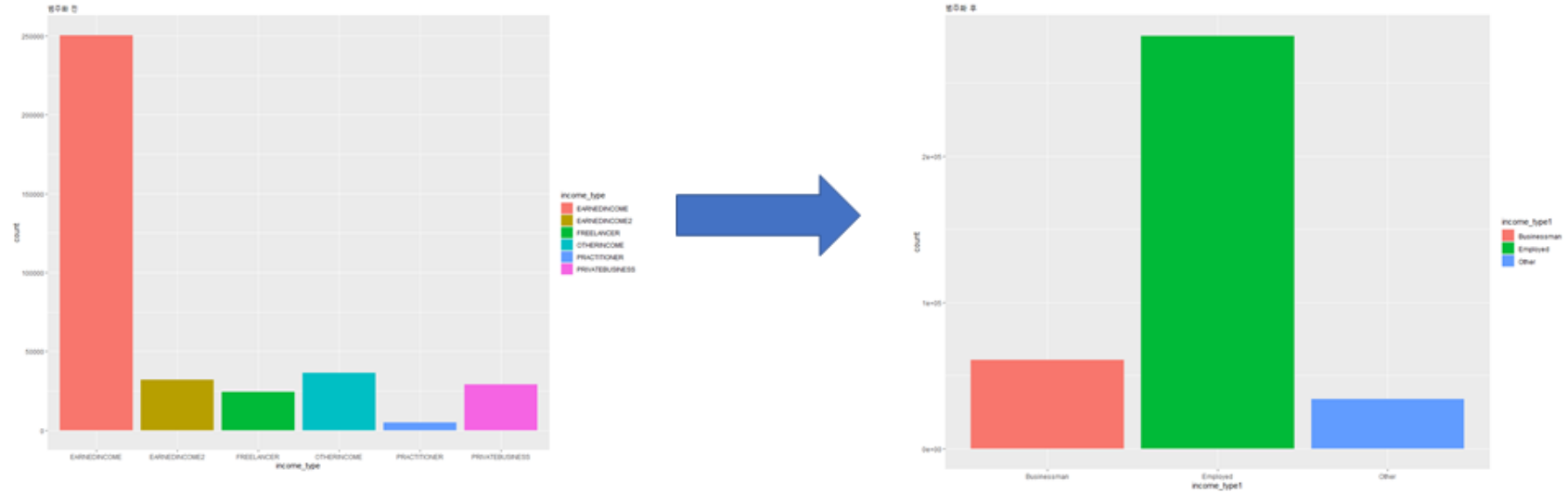
공통된 변수 추출 및 데이터 filtering

두 데이터의 변수들이 채무 불이행에 동일한 영향을 끼친다는 가정

변수명	변수 설명	변수타입
AMT_INCOME_TOTAL	연소득	연속형
AMT_CREDIT	대출희망금액	연속형
PREV_AMT_CREDIT	기대출금액	연속형
AGE	유저 나이	정수형
WORK_MONTH	입사개월수	연속형
WORK_AGE	입사년수	정수형
GENDER	유저 성별	범주형
INCOME_TYPE	근로 형태	범주형
HOUSING_TYPE	주거소유형태	범주형
PREV_APP_CNT	기대출횟수	이산형

데이터 전처리

범주형 데이터 전처리 과정



범주를 압축한 plot

소득 유형

✓ 고용인 / 피고용인 / 기타

주택 유형

✓ 자가주택 / 전월세 / 기타

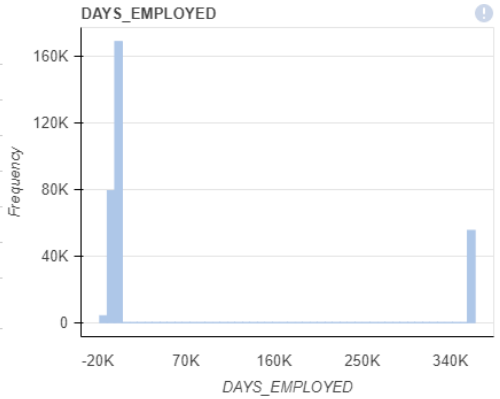
데이터 전처리

결측치 처리 과정: Credit risk 데이터

CODE_GENDER
XNA

결측치로 간주하여
최빈값인 F로(female을 의미) 대체

Mean	63815.0459
Minimum	-17912
Maximum	365243
Zeros	2
Zeros (%)	0.0%
Negatives	252135
Negatives (%)	82.0%



결측치의 비율은 없지만 수치적으로 이상치 존재
이상한 값으로 판단 후 결측치 처리

MAR 가정

- ✓ Missing value의 유형은 MCAR, MAR, MNAR로 세 가지
- ✓ 결측치가 랜덤으로 발생하지만 다른 변수와 관계 있음

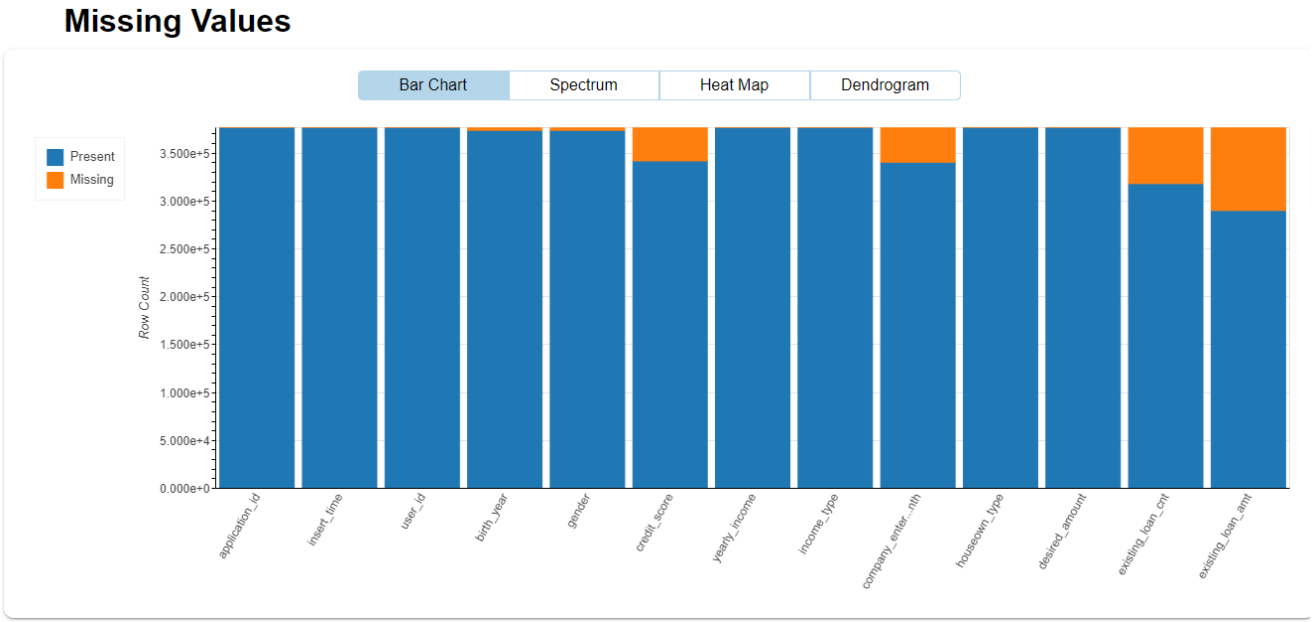
MICE 적용

- ✓ 다량의 결측치는 동 변수 하나만 존재
- ✓ MICE 알고리즘을 사용하여
  stochastic regression imputation을 진행

데이터 전처리

결측치 처리 과정: Finda 데이터

✓ Finda 데이터



기대출수와 기댓출횟수에 관련된 변수에 이상치 다량 존재

✓ 학습에 사용되지 않는 데이터

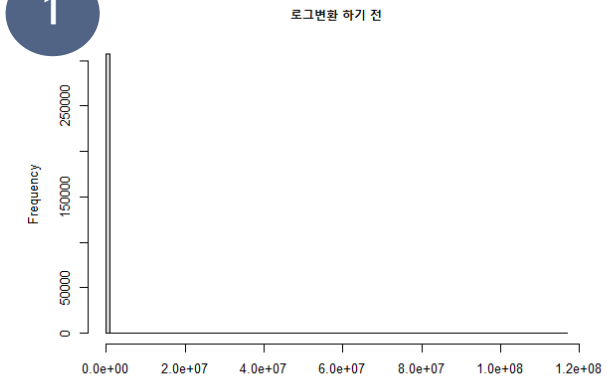
✓ 다량의 결측치이기에 mice로 보간했을 때 노이즈가 크게 생길 것

→ 해당 행을 삭제

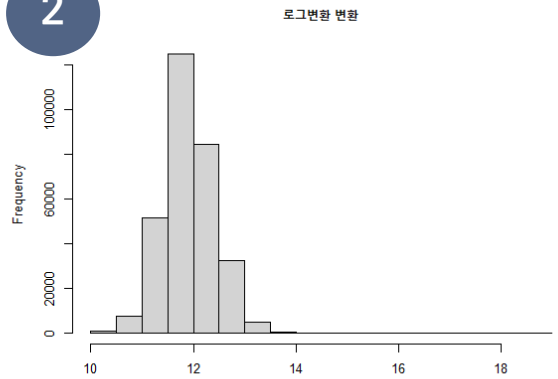
데이터 전처리

이상치 제거

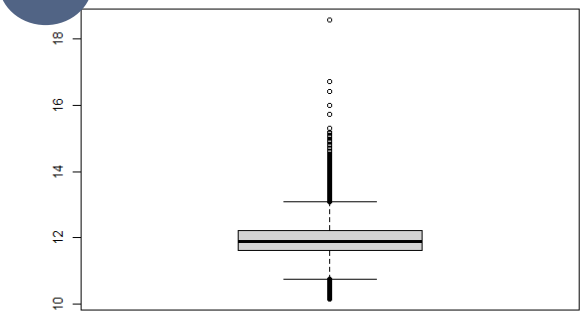
1



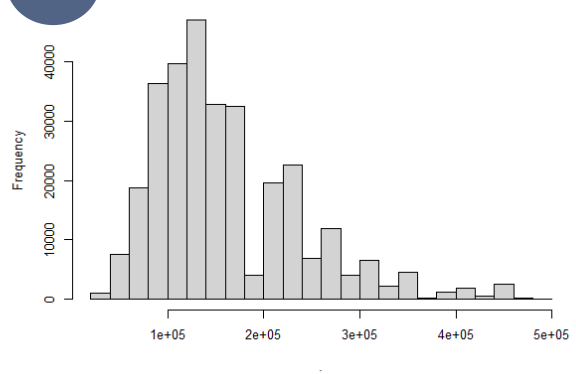
2



3



4



첨도와 왜도가 높은 데이터

- ✓ Skewed 된 변수와 아웃라이어가 있다고 판단되는 변수에 한해 **로그 변환** 진행
- ✓ Q3 + 1.5IQR(IQR = Q3 − Q1)을 벗어나는 데이터는 이상치로 간주 후 제거
- ✓ 이상치 제거하는 이유는 두 데이터의 단위가 맞지 않는 변수에 한해 scaler의 변동성을 제거하기 위함임

데이터 전처리

파생변수 생성

두 데이터를 통합하는 과정에서 사용하지 못한 변수 다량 존재 변수 부족 문제 해결 하기 위해 생성
Credit risk 데이터 120개 열 → 12개 열

변수명	생성식
INCOME_CREDIT_RATE	AMT_INCOME_TOTAL / AMT_CREDT
PRE_CURR_CREDIT_DIFF	AMT_CREDIT - PREV_AMT_CREDIT
WORK_AGE	AGE - WORK_MONTH / 12

INCOME_CREDIT_RATE

- ✓ 실제 부채가 과도한 수준인지 보기 위해 소득 수준과의 비율이 필요하다고 판단하여 생성

PRE_CURR_CREDIT_DIFF

- ✓ 대출자의 과거 기대출과 현재 신청한 대출 금액을 통해 현재 부채가 무리한 수준인지 파악하기 위해 생성

WORK_AGE

- ✓ 현재 나이에서 근속 기간을 빼 처음 일하기 시작한 나이를 산출
- ✓ 학력을 간접적으로 대변하는 변수

데이터 전처리

Robust Scaler 적용

$$Robust\ Scaler = \frac{x - Q1(x)}{Q3(x) - Q1(x)}$$

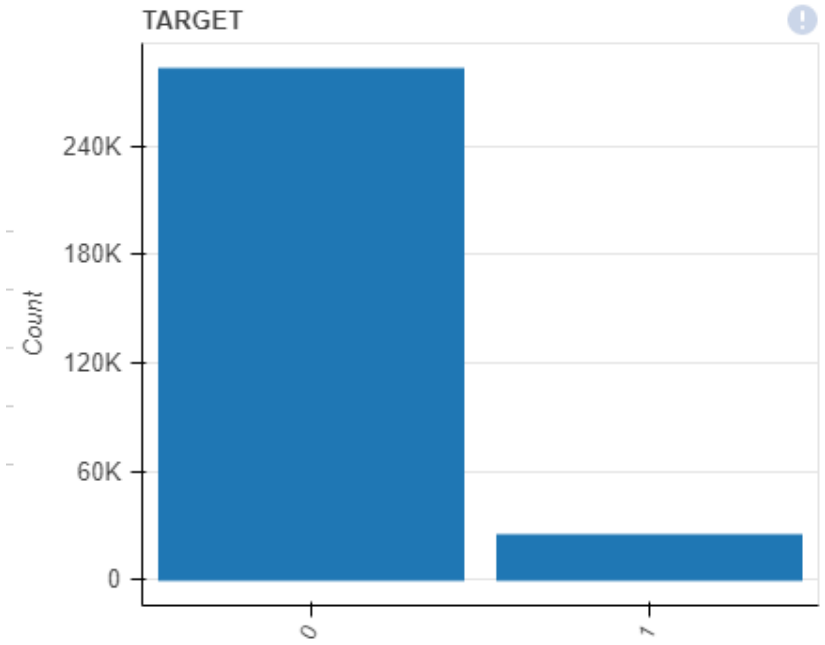
- Robust Scaler는 이상치에 덜 민감하다는 장점이 있음

사용 이유

- ✓ 통화 단위 일원화 위해
- ✓ Right-skewed 경향을 띄는 변수들에서 이상치로 보고 제거하기 어려움
  → Min_Max Scaler를 적용하기에는 여전히 이상치가 제거되지 않을 위험성 존재
- ✓ 분석 과정에서 MLP 등 딥러닝 관련 시도할 때 필요
  - 트리 기반 모델에는 알고리즘 특성상 scaling으로 성능 향상을 기대하기는 어려움

모델링

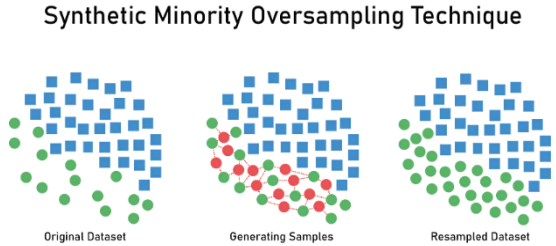
고려 사항
데이터 불균형 문제를 보완하기 위한 방법



Credit의 Target(0, 1) barplot 시각화

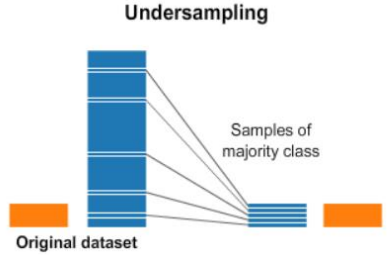
1 SMOTENC

- ✓ Over-sampling 기법으로써 범주형 변수도 작동할 수 있는 기법
- ✓ 범주형 변수가 있다는 것을 고려한 방법



2 UNDER SAMPLING

- ✓ 데이터의 양이 충분히 많기 때문에 랜덤 언더 샘플링 기법 고려



3 파라미터 튜닝(is_balanced = True)

- ✓ 불균형을 해소하기 위해 가중치를 주는 방식

모델링

고려 사항
데이터 불균형 문제를 보완하기 위한 방법

베이스라인 코드는 트리기반의 학습 알고리즘인 LGBM(Light Gradient Boosting Model) 사용

- 빠른 학습 속도와 기본 파라미터 튜닝으로 어느 정도의 성능을 보장하기 때문
- Sklearn api 모델과 달리 custom loss function 사용 가능하기에 imbalanced data에 다양한 시도를 할 수 있었기 때문
- Random Forest를 시도했으나 AUC_ROC score가 0.633으로 Lightgbm을 적용한 베이스라인이 성능이 더 좋기 때문에 LGBM 사용

모델 성능 판단 기준은 AUC-ROC score가 가장 높은 값을 임계점으로 설정

- 성능적 측면과 분석 목표에 근접한 모델을 만들기 위함임
- 이 임계점을 기반으로 confusion matrix를 생성 예정

방법	AUC_ROC score
베이스라인(lgbm)	**0.646**
SMOTENC	0.567
Under sampling	0.630
파라미터 튜닝	0.603

베이스라인의 score가 가장 높음

✓ 불균형을 보완하기 위한 방법들은
노이즈를 크게 발생시키거나 정보 손실을 야기한다고 판단

✓ 따라서 loss function을 통해 문제 해결 시도

모델링

고려 사항
Loss function을 통한 방법

	Predcited_neg	Predcited_pos
True_neg	13326	14943
True_pos	1025	1458

	Predcited_neg	Predcited_pos
True_neg	16660	11609
True_pos	993	1490

	Predcited_neg	Predcited_pos
True_neg	16697	11572
True_pos	954	1529

언급한 임계점을 기반으로 생성한 confusion matrix 및 Loss function 실험 결과
위에서부터 roc_auc, Focal cross entropy, Binary log-loss (lgbm 기본 제공)

Recall 값이 기준인 이유

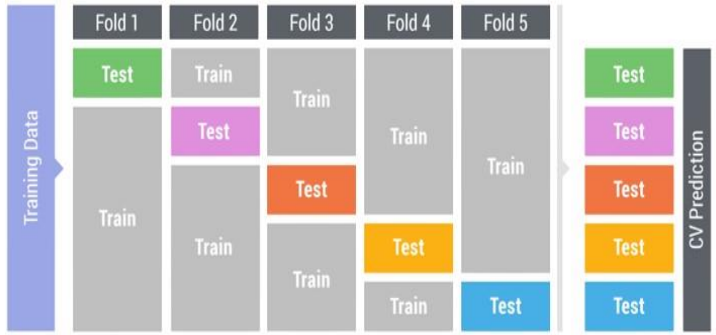
- ✓ 리스크 확산 방지가 분석 목적 중 하나(연쇄부도 방지)
- ✓ 채무 불이행을 민감하게 찾아내기 위해 민감도를 우선으로 하여 confusion matrix의 recall이 가장 높은 loss function을 선택

- ✓ Recall 값이 제일 높은 **binary log-loss function**을 loss function으로 사용
  Recall = True Positive / (True Positive + False Positive)

모델링

Stratified K-Fold Cross Validation Stacking Ensemble

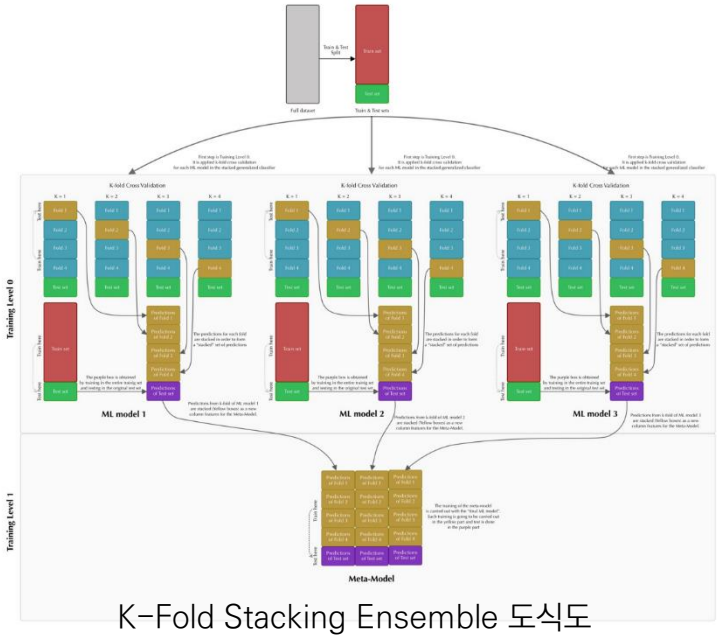
- train set, validation set, test set은 전체 데이터 셋에서 63%, 27%, 10%의 비율로 분석에 사용



K-Fold Cross Validation Process 그림

Stratified K-Fold Cross Validation

- ✓ K-Fold CV에서 각각의 Fold에 Target의 분포가 동일하게 Fold를 나누는 방식
- ✓ Credit 데이터의 Target 분포 시각화와 같이 범주의 분포가 균일하지 않기에 사용하기 적절하다고 판단



K-Fold Stacking Ensemble 도식도

Training Level 0

- ✓ Train, validation Set에 대해 처음 n개의 모델이 들어가 학습 진행

Training Level 1

- ✓ 각 모델의 predict를 이용해 Meta-Data 형성 후 Meta-Data를 통해 Meta-Model을 학습

모델링

Stratified K-Fold Cross Validation Stacking Ensemble

- train set, validation set, test set은 전체 데이터 셋에서 63%, 27%, 10%의 비율로 분석에 사용

Stacking Ensemble

Training Level 0

사용한 모델은
Lightgbm, Adaboost, Random Forest, ExtraTreesClassifier,
MLPClassifier, Logistic Regression (총 6개)

↓

Training Level 1

Lightgbm으로 학습시켜 모델 완성
모델 파라미터는 0 level 모델 6개, Meta-model 모두 optuna 패키지를 사용하여 최적화

Stacking Ensemble 사용 이유

- ✓ 기존 모델의 성능이 낮았기 때문에 최소한의 성능을 보장하기 위해 다양한 모델의 결과를 메타 학습하여 성능을 끌어올릴 목적
- ✓ 다양한 모델을 사용함으로써 각 모델의 feature importance 및 sharp value를 통해 모델별 변수의 영향/기여도 양상을 확인할 목적

모델링

Stratified K-Fold Cross Validation Stacking Ensemble

- train set, validation set, test set은 전체 데이터 셋에서 63%, 27%, 10%의 비율로 분석에 사용

K-Fold에서 k=5를 적용 및
0 step train에 사용된 모델들과 메타모델 모두 optuna 패키지를 사용하여 최적화

```
params_lgb = {
    "random_state": 4885,
    "verbosity": -1,
    "learning_rate": 0.05,
    "n_estimators": 10000,
    "objective": "binary",
    "reg_alpha": trial.suggest_float("reg_alpha", 1e-8, 3e-5),
    "reg_lambda": trial.suggest_float("reg_lambda", 1e-8, 9e-2),
    "max_depth": trial.suggest_int("max_depth", 1, 20),
    "num_leaves": trial.suggest_int("num_leaves", 2, 256),
    "colsample_bytree": trial.suggest_float("colsample_bytree", 0.4, 1.0)
}
```

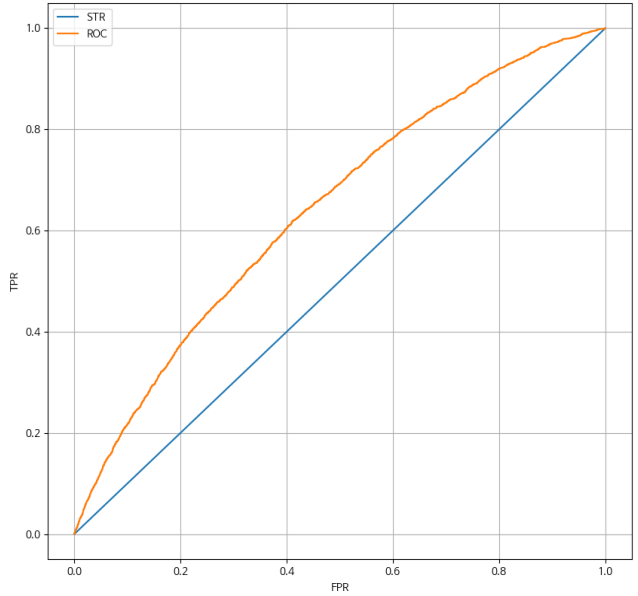
```
[I 2023-02-13 00:25:12,967] Trial 0 finished with value: 0.6555694901626302 and parameters: {'reg_alpha': 6.413115564166195e-06, 'reg_lambda': 0.05085443815925977, '
max_depth': 16, 'num_leaves': 16, 'colsample_bytree': 0.8052158771534885, 'subsample': 0.5529713643286454, 'subsample_freq': 1, 'min_child_samples': 9, 'max_bin': 28
0}. Best is trial 0 with value: 0.6555694901626302.
[I 2023-02-13 00:25:27,443] Trial 1 finished with value: 0.6477100146237463 and parameters: {'reg_alpha': 2.4130315144639867e-05, 'reg_lambda': 0.03513171636695353,
'max_depth': 3, 'num_leaves': 255, 'colsample_bytree': 0.8568306397309378, 'subsample': 0.4082364723197392, 'subsample_freq': 1, 'min_child_samples': 79, 'max_bin':
233}. Best is trial 0 with value: 0.6555694901626302.
[I 2023-02-13 00:25:32,905] Trial 2 finished with value: 0.6448370978280702 and parameters: {'reg_alpha': 1.4026688272191281e-05, 'reg_lambda': 0.00939566326902263,
'max_depth': 16, 'num_leaves': 69, 'colsample_bytree': 0.5601206431889586, 'subsample': 0.9192342357993655, 'subsample_freq': 4, 'min_child_samples': 69, 'max_bin':
248}. Best is trial 0 with value: 0.6555694901626302.
[I 2023-02-13 00:26:23,030] Trial 3 finished with value: 0.6442969759432275 and parameters: {'reg_alpha': 2.290907505666122e-05, 'reg_lambda': 0.06427553150407958, '
max_depth': 1, 'num_leaves': 190, 'colsample_bytree': 0.5160104910657785, 'subsample': 0.8247687059021509, 'subsample_freq': 9, 'min_child_samples': 86, 'max_bin': 2
01}. Best is trial 0 with value: 0.6555694901626302.
```

모델링 결과

Stratified K-Fold Cross Validation Stacking Ensemble

- train set, validation set, test set은 전체 데이터 셋에서 63%, 27%, 10%의 비율로 분석에 사용

AUC_ROC score = 0.65



0.65로 소폭 상승

Confusion Matrix

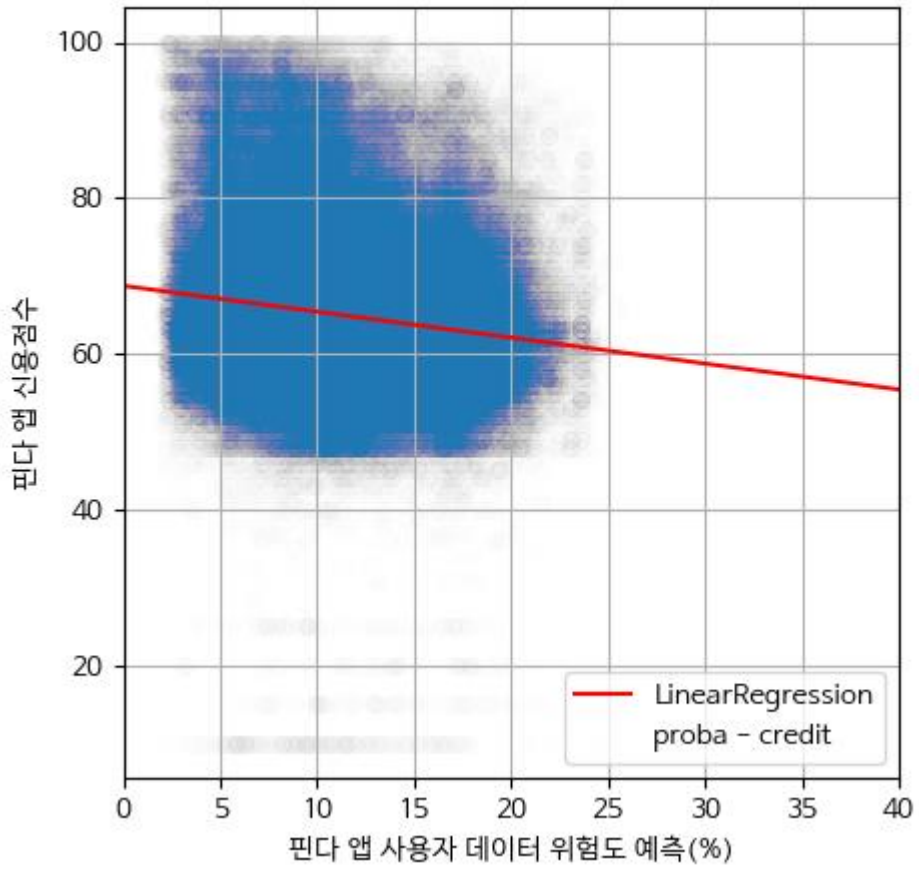
	Predcited_neg	Predcited_pos
True_neg	16697	11572
True_pos	954	1529

	Predcited_neg	Predcited_pos
True_neg	15885	12384
True_pos	882	1601

모델링 결과

Stratified K-Fold Cross Validation Stacking Ensemble

- train set, validation set, test set은 전체 데이터 셋에서 63%, 27%, 10%의 비율로 분석에 사용

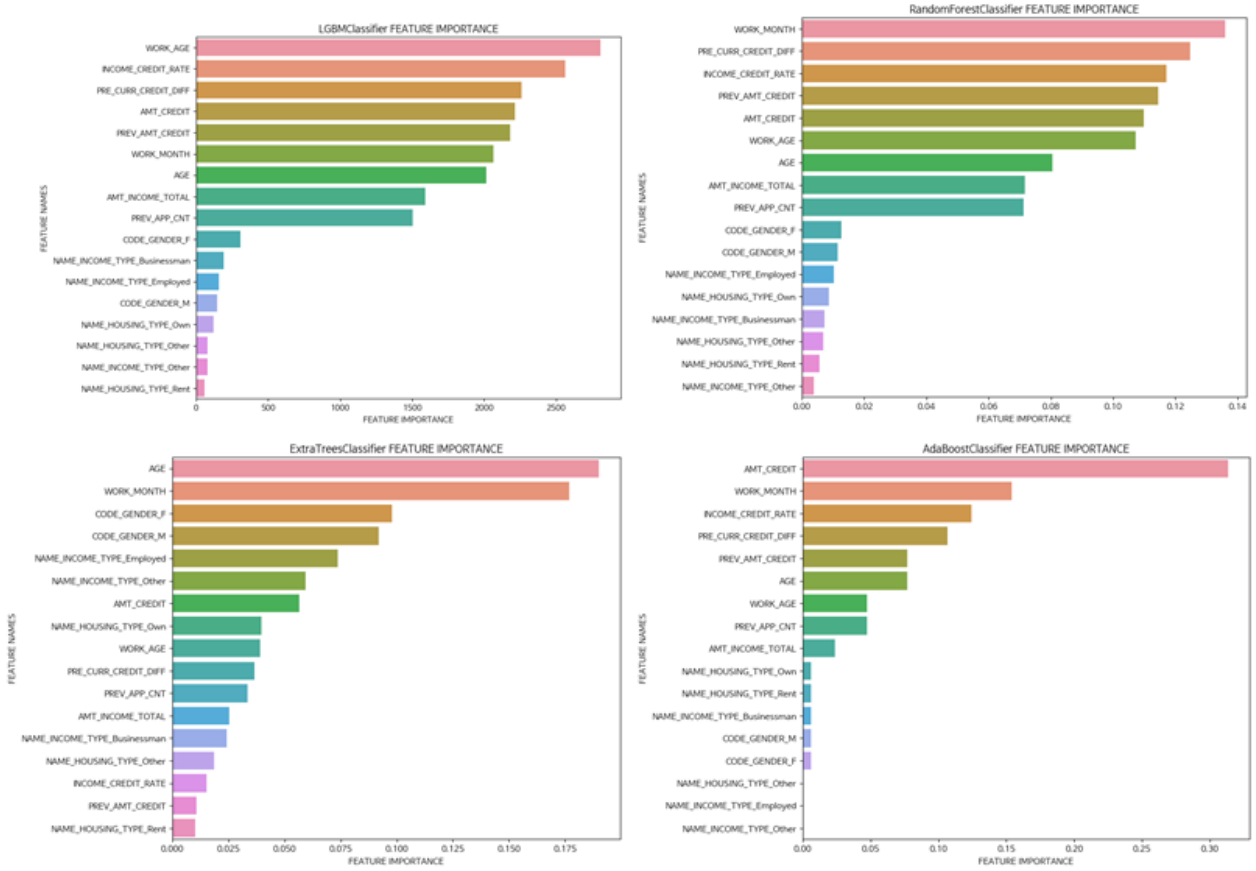



```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.029
Model:                            OLS   Adj. R-squared:                  0.029
Method:                 Least Squares   F-statistic:                     7455.
Date:                Mon, 13 Feb 2023   Prob (F-statistic):               0.00
Time:                        02:05:29   Log-Likelihood:            -9.4185e+05
No. Observations:              253971   AIC:                         1.884e+06
Df Residuals:                  253969   BIC:                         1.884e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         69.0461      0.051   1347.901      0.000      68.946      69.146
x1            -0.3430      0.004    -86.342      0.000      -0.351      -0.335
==============================================================================
Omnibus:                    31324.591   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            62536.416
Skew:                           0.784   Prob(JB):                         0.00
Kurtosis:                       4.857   Cond. No.                         33.9
==============================================================================
```

Finda 앱 데이터를 통해 예측한 위험도와 신용 점수를 통해
기존의 신용점수와 모델을 통해 산출한 위험도의 상관관계 파악
회귀계수가 약 -0.34로 **약한 음의 관계**를 볼 수 있었음

모델링 결과

Feature Importance



오른쪽 상단부터 시계방향
Random forest, Adaboast, ExtraTreeClassifier, LGBM

Feature Importance

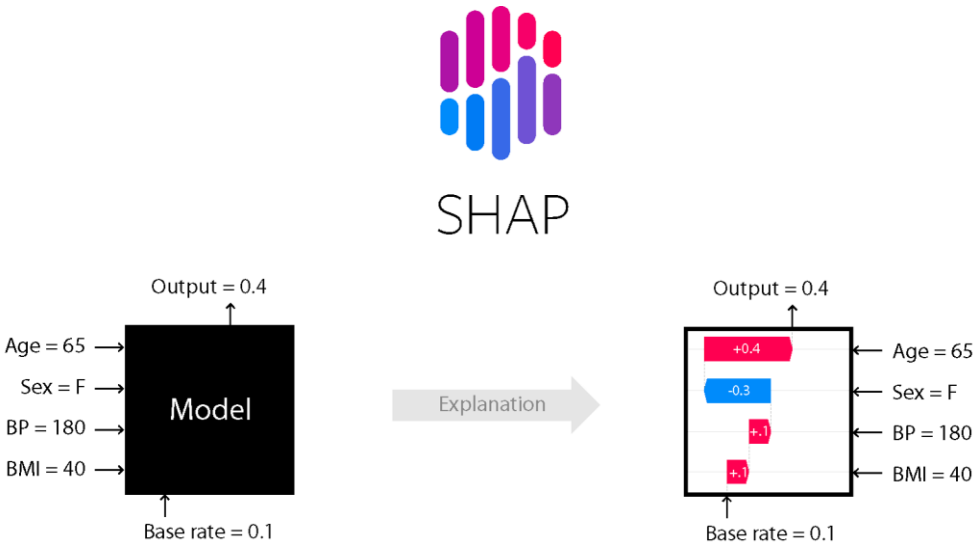
어떠한 변수가 모델의 예측 결과에 미치는 영향을 알 수 있음

한계점

- ✓ 어떤 방향으로 영향을 주는지는 알 수 없음
- ✓ 학습 데이터의 노이즈와 트리 모델의 샘플링 변동성에 영향을 크기 받음

→ **한계점**들로 인해 자세하고 신뢰받는 설명 지표로 **활용 불가**

SHAP(Shapley Addictive exPlanations)



Shap 그림

SHAP란

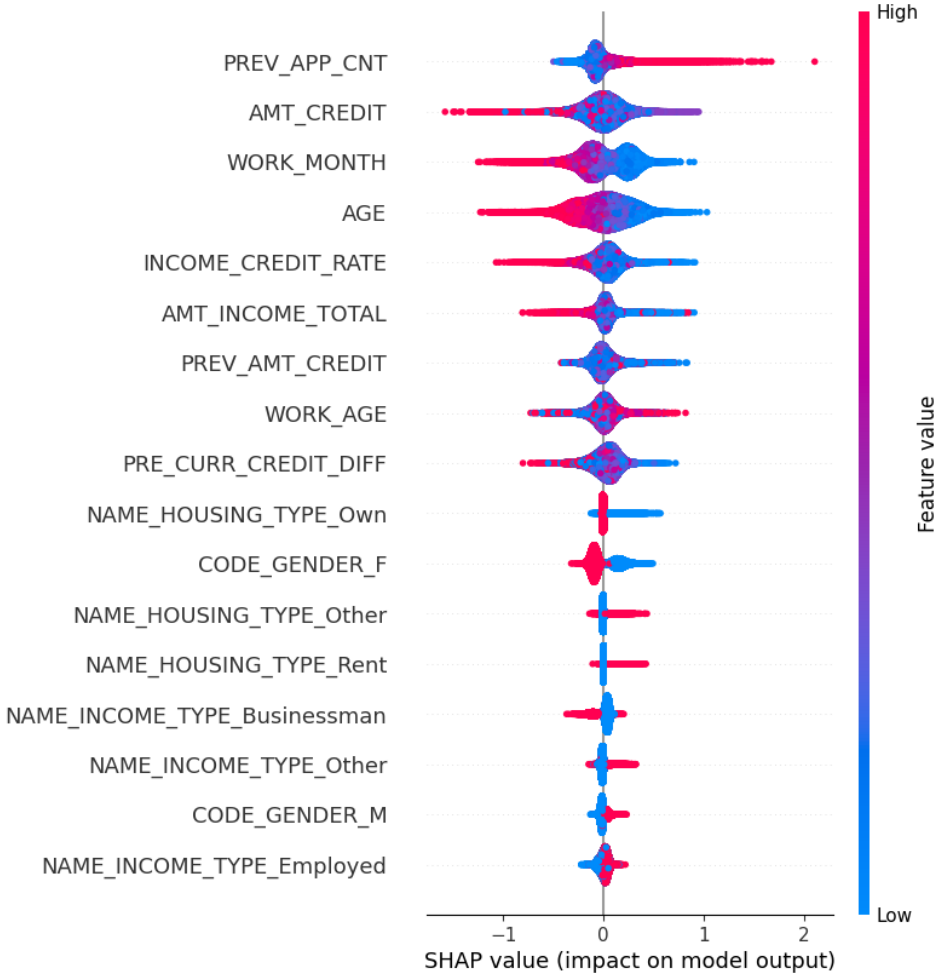
각각의 입력 변수에 대한 shapley value를 계산함으로써 입력 변수와 모델의 결과 값 사이의 관계를 분석하는 설명 가능한 인공지능 기법

- ✓ 게임 이론에서 각 참여자의 기여도를 계산하는 방법론을 기계학습 모델에 적용
- ✓ (m, n)의 학습 데이터에서 m은 게임 횟수, n은 참여자 수로 모델의 output은 게임의 결과로 생각하여 각 변수의 기여도를 계산
- ✓ 모든 행에서 **변수의 기여도를 계산**할 수 있기 때문에 모델을 설명하는 데에 있어 가장 좋은 방법이라 생각하여 채택

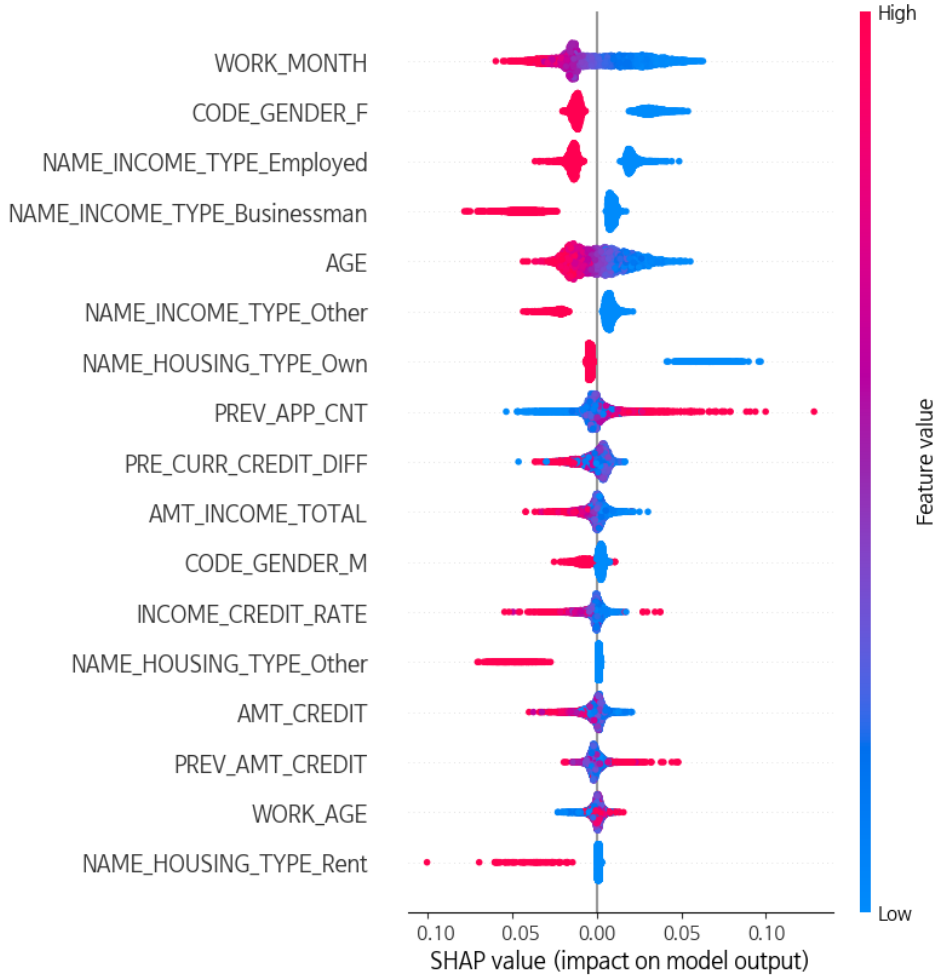
SHAP Plot

입력 변수들의 SHAP value 값이 표시된 분포도

- shapley value의 절댓값이 높은 순으로 나열되어 있음
- 점의 색은 각 입력 변수의 값이 높고 낮은 정도를 나타냄
- 중상선을 기점으로 왼쪽의 점은 예측 값이 낮아지는 데 영향을 줌
- 반대로 오른쪽은 예측 값이 높아지는 데 영향을 줌



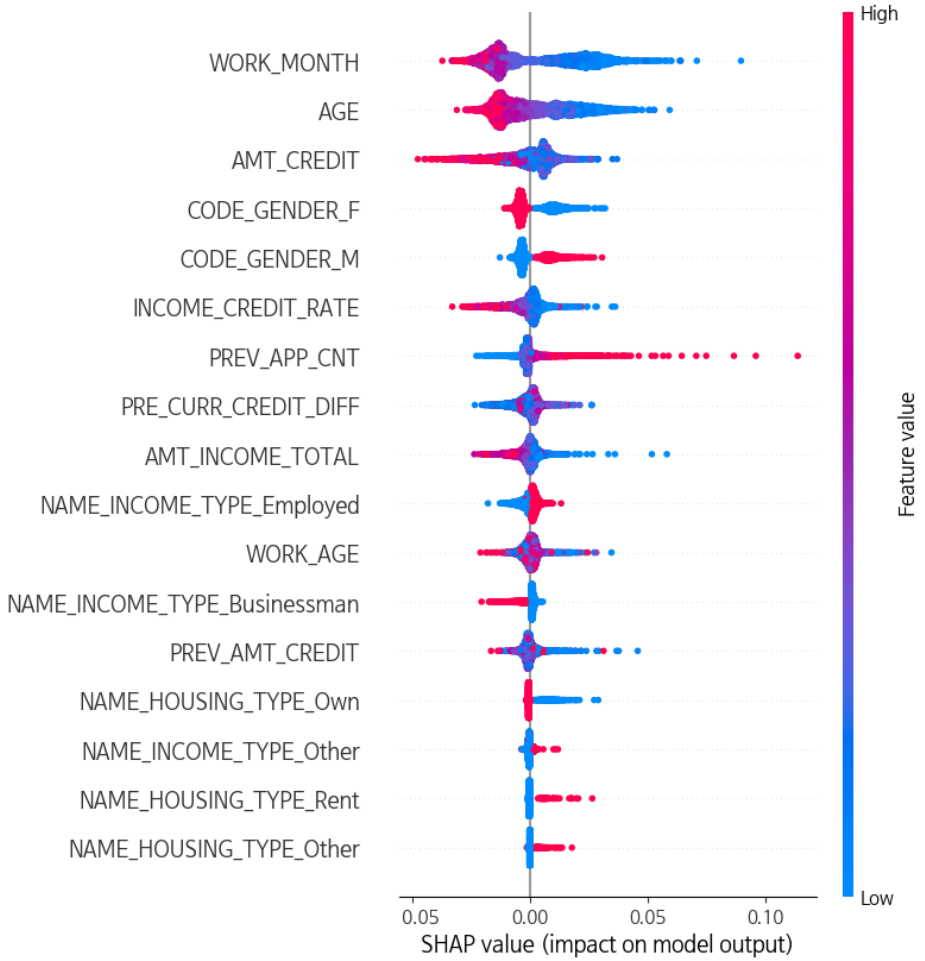
Lightgbm



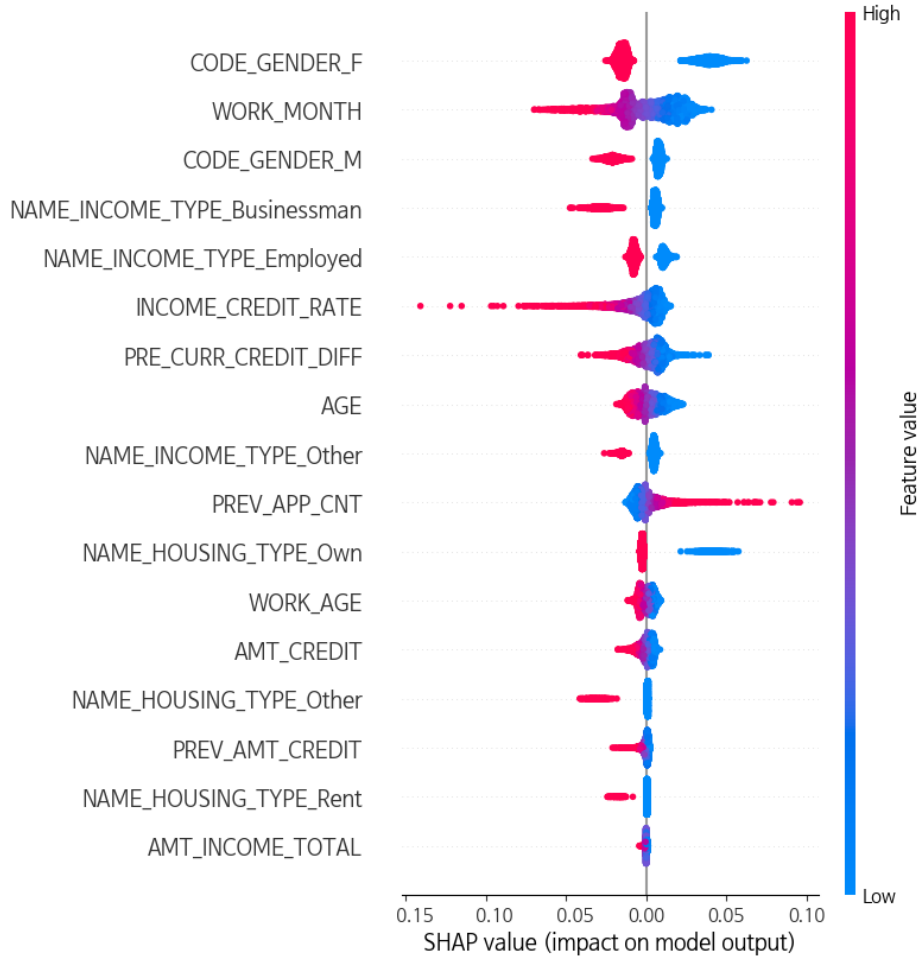
MLP

SHAP Plot

- shapley value의 절댓값이 높은 순으로 나열되어 있음
- 점의 색은 각 입력 변수의 값이 높고 낮은 정도를 나타냄
- 중상선을 기점으로 왼쪽의 점은 예측 값이 낮아지는 데 영향을 줌
- 반대로 오른쪽은 예측 값이 높아지는 데 영향을 줌



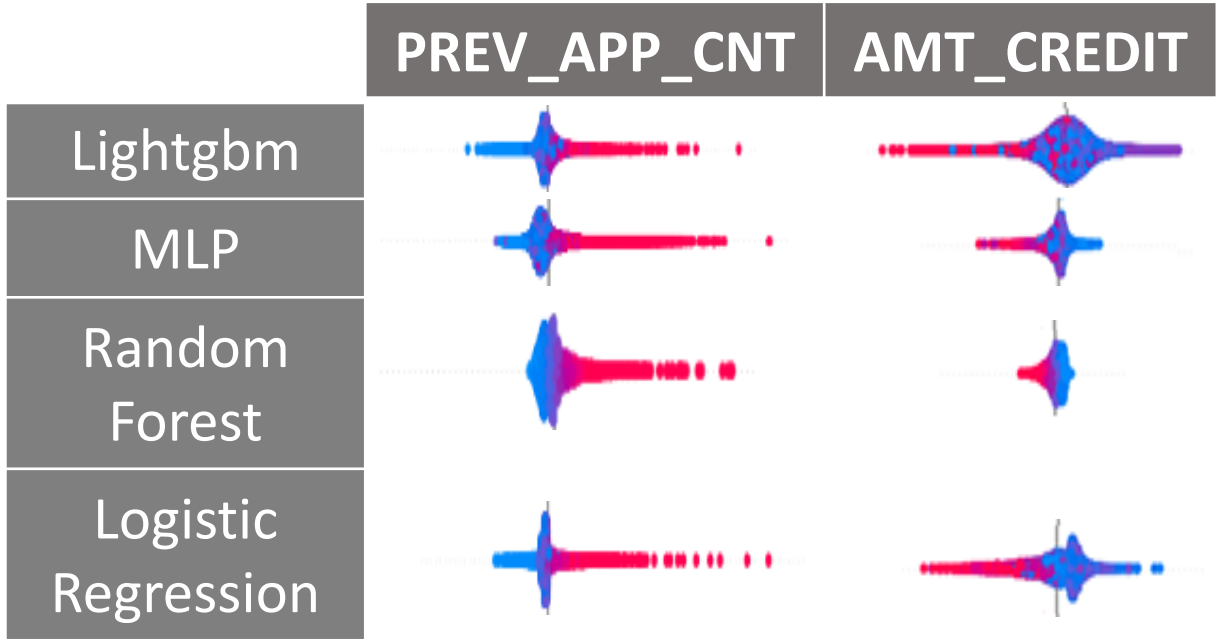
Random Forest



Logistic Regression

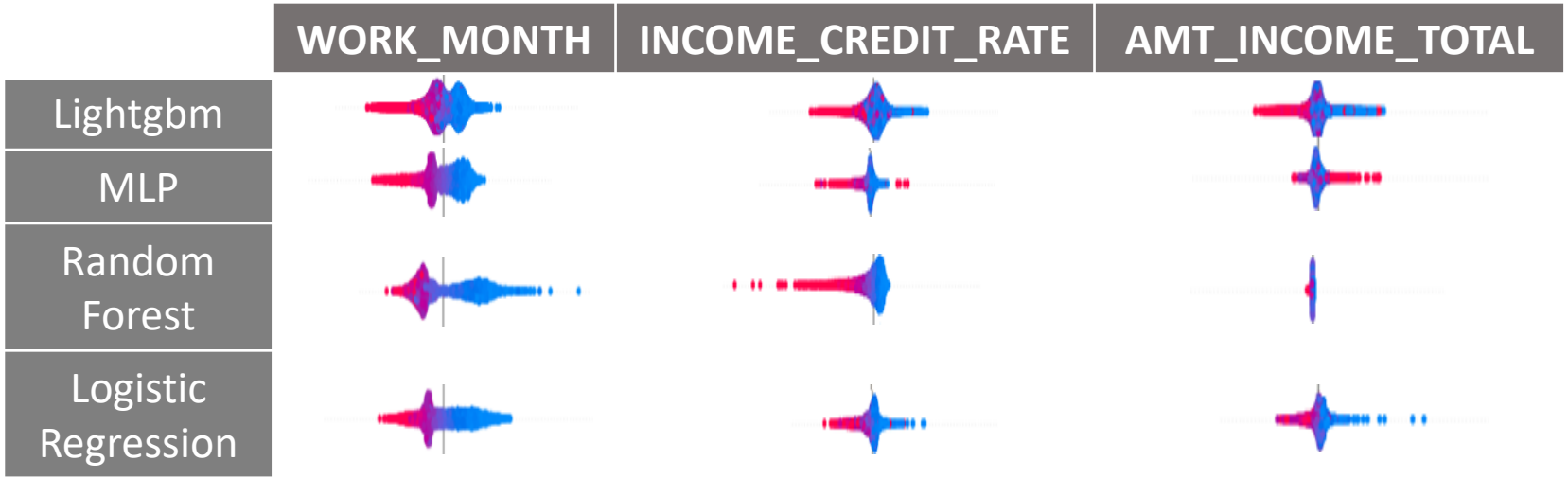
SHAP Plot

유의미한 변수 5개의 SHAP plot



SHAP Plot

유의미한 변수 5개의 SHAP plot



모델 해석

PREV_APP_CNT

기대출 횟수가(PREV_APP_CNT) 높으면
채무 불이행 위험을 높임

INCOME_CREDIT_RATE

연소득/대출희망금액(INCOME_CREDIT_RATE)이 낮으면
채무 불이행을 높이는 방향으로 학습이 이루어짐

WORK_MONTH

경력의 경우 근속월수가(WORK_MONTH)
낮으면 채무 불이행을 높이는 방향으로 학습이
이루어짐

AMT_INCOME_TOTAL

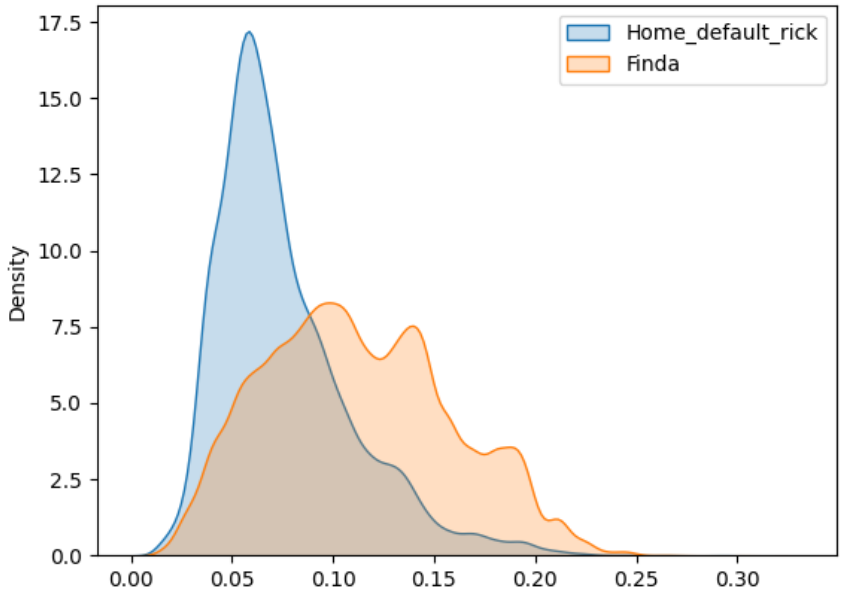
연소득에(AMT_INCOME_TOTAL) 대해서는 대체로 낮으면
채무 불이행 위험을 증가시킴

AMT_CREDIT

- ✓ 대출희망금액이(AMT_CREDIT) 낮으면 대출 불이행 위험을 높이는 것을 확인할 수 있음
- ✓ 이때의 해석은 고액 대출일 경우 은행에서 대출자의 신용을 면밀히 보거나 담보를 설정하고 대출하기에
  대출희망금액이 낮으면 대출 불이행 위험을 증가시켰다고 할 수 있음
- ✓ 따라서 소액 대출을 중심으로 채무 불이행이 일어날 것이라는 인사이트를 얻을 수 있었음

모델 해석

Credit risk 데이터에 Finda 데이터 적용



채무 불이행에 대한 분포

→ 한국의 부채 위험이 더 크다고 판단

한계점

데이터 측면의 한계점

- 적은 column 수로 인한 정보 손실 → Credit risk, Finda 데이터의 공통 column만을 추출이 원인
  - ✓ 사후 평가로써, score 높이기 위해 baseline 코드를 돌렸을 때 분석 과정에서 쓰지 못한 변수들이 유의미한 결과를 도출
- 정확한 거시경제 지표 수집이 어려움 → Credit Risk 데이터에서 날짜 데이터가 마스킹 되어 있음
- Credit Risk 데이터는 한 은행사, Finda는 여러 은행사에서 수집된 자료
  - ✓ 은행사마다 상품도 다르기 때문에, 은행사에 대한 동질성 검정 실시하지 못한 한계를 지님

가정 측면의 한계점

- 미국(Credit risk 데이터) – 한국(Finda 데이터)는 국가 차이로 인해 동질성 확보가 어려움
- Subsampling을 시도하지 못한 한계점을 지님
  - ✓ Subsampling은 데이터가 동질성이 맞지 않아 가정을 사후적으로 충족시키기 위해서 필터링 목적으로 시도할 수 있음
  - ✓ 예를 들면 재정이 건전한 사람들만을 추출하여, 재정이 건전한 상황하에서 채무 불이행율이 경제 상황으로 인해 바뀌는지 확인 가능성 존재
- MAR 검증 과정이 미흡하였음

기대 효과 및 시사점

기대 효과

연쇄 부도 막을 수 있는 위험 예측 지표로 활용

- 신용경계감이 높아진 지금 상황에서 연쇄부도를 막을 수 있는 위험 예측 지표로 활용
- 대출자의 위험도를 정확히 파악 및 예금 금리의 리스크 프리미엄을 감소시켜 예금 금리를 낮출 수 있음
- 채무자의 자금 조달 차질 완화 및 유동성 부족 문제 해결에 기여

시사점

한국은행 선행 연구를 현실 데이터으로 증명했다는 점에서 의의를 가짐

- 소액 대출은 채무 불이행이 높다는 모델 해석을 아래의 선행 연구에서 확인할 수 있음
  - ✓ 선행 연구 〈가계대출 민감도 분석 및 시사점, 조사통계월보 제76권 제9호 (2022.9)〉에 따르면,
    금리 인상은 가계부채 증가세의 억제에 기여 및 금융불균형을 완화
  - ✓ 취약계층은 생계비 목적으로 대출하고, 이들은 금리가 오르더라도 대출 억제 효과가 크지 않으며 채무상환부담이 더 많이 늘어나는 위험성을 야기
  - ✓ 금리상승압력이 큰 시기에는 취약계층의 부실 위험이 높아질 수 있으며,
    이들에 대한 대출 비중이 높은 비은행금융기관의자산 건전성이 저하될 가능성 존재
  - ✓ 본 연구는 리스크 현실화 가능성이 가장 높고, 자산 건전성이 악화될 수 있는 취약 계층에게 정책적 도움이 필요함을 시사함
- imbalanced 데이터 → 연체율이 낮다는 사실로부터 발생
  - ✓ 주제 특성상 생길 수밖에 없는 데이터의 특성

Q&A

감사합니다