머신러닝을 활용한

대출 연체 예측

산업정보시스템공학과 | 20212843 권용후 | 20211290 심승현 | 20211322 함유민

목 차

분석 배경

분석 배경

연구 소개

- 최근 데이터마이닝을 활용하여 고객의 금융 행태를 분석하고, 대출 위험도를 사전에 예측하려는 금융사들이 증가
- 고객 맞춤형 신용 평가와 대출 심사 시스템을 통해 금융기관은 연체 가능성이 높은 고객을 조기에 식별하고, 이에 따라 적절한 리스크 관리 전략을 수립
- 본 연구는 금융 고객의 다양한 속성(직업, 소득, 재무상태 등)을 기반으로 대출 연체 여부를 예측하는 모델을 데이터마이닝의 다양한 기법으로 구축하고자 함.

분석 배경

연구 필요성

- 금융 산업에서는 고객의 신용 리스크를 사전 예측하고 관리하는 것이 금융기관의 안정성과 직결됨.
- 대출 연체는 직접적 손실뿐 아니라 고객 신뢰 하락과 시스템 위험으로 이어질 수 있음
- 기존 신용평가 모델은 복잡한 데이터와 비선형적 패턴을 충분히 반영하지 못해 한계가 있음

머신러닝 기반의 정교한 분석 기법을 통해
연체 위험을 조기에 탐지해야 함

연구 목표

- 고객 특성 데이터를 기반으로 대출 연체 여부를 예측하는 최적의 머신러닝 모델을 구축
- 데이터 전처리, 이상치 및 불균형 처리, 피처 선택을 거쳐 RandomForest, SVM, XGBoost, LightGBM 다양한 모델을 비교·평가
- 다양한 평가지표를 통해 최적의 임계값을 설정

데이터 분석 및 전처리

데이터 탐색

데이터 소개

LoanID	Age	Income	LoanAmou	CreditScor	MonthsEm	NumCredit	InterestRat	LoanTerm	DTIRatio	Education	Employme	MaritalStat	HasMortga	HasDepen	LoanPurpo	HasCoSign	Default
I38PQUQS	56	85994	50587	520	80	4	15.23	36	0.44	Bachelor's	Full-time	Divorced	Yes	Yes	Other	Yes	0
HPSK72WA	69	50432	124440	458	15	1	4.81	60	0.68	Master's	Full-time	Married	No	No	Other	Yes	0
C1OZ6DPJ	46	84208	129188	451	26	3	21.17	24	0.31	Master's	Unemploy	Divorced	Yes	Yes	Auto	No	1
V2KKSFM3	32	31713	44799	743	0	3	7.07	24	0.23	High Scho	Full-time	Married	No	No	Business	No	0
EY08JDHTZ	60	20437	9139	633	8	4	6.51	48	0.73	Bachelor's	Unemploy	Divorced	No	Yes	Auto	No	0
A9S62RQ7	25	90298	90448	720	18	2	22.72	24	0.1	High Scho	Unemploy	Single	Yes	No	Business	Yes	1
H8GXPAOS	38	111188	177025	429	80	1	19.11	12	0.16	Bachelor's	Unemploy	Single	Yes	No	Home	Yes	0
0HGZQKJ3	56	126802	155511	531	67	4	8.15	60	0.43	PhD	Full-time	Married	No	No	Home	Yes	0
1R0N3LGN	36	42053	92357	827	83	1	23.94	48	0.2	Bachelor's	Self-emplc	Divorced	Yes	No	Education	No	1
CM9L1GTT	40	132784	228510	480	114	4	9.09	48	0.33	High Scho	Self-emplc	Married	Yes	No	Other	Yes	0
IA35XVH6Z	28	140466	163781	652	94	2	9.08	48	0.23	High Scho	Unemploy	Married	No	No	Education	No	0
Y8UETC3LS	28	149227	139759	375	56	3	5.84	36	0.8	PhD	Full-time	Divorced	No	No	Education	Yes	1
RM6QSRH	41	23265	63527	829	87	4	9.73	60	0.45	Master's	Full-time	Divorced	Yes	No	Auto	Yes	0
GX5YQOGI	53	117550	95744	395	112	4	3.58	24	0.73	High Scho	Unemploy	Single	No	No	Auto	Yes	0
X0BVPZLD	57	139699	88143	635	112	4	5.63	48	0.2	Master's	Part-time	Divorced	No	No	Home	No	0
O5DM5MF	41	74064	230883	432	31	2	5	60	0.89	Master's	Unemploy	Married	Yes	No	Auto	No	0
ZDDRGVTE	20	119704	25697	313	49	1	9.63	24	0.28	PhD	Unemploy	Single	Yes	No	Home	No	0
9V0FJW7Q	39	33015	10889	811	106	2	13.56	60	0.66	Master's	Self-emplc	Single	Yes	No	Other	No	0
O1IKKLC6S	19	40718	78515	319	119	2	14	24	0.17	Bachelor's	Self-emplc	Divorced	Yes	No	Education	No	1
F7487UU2I	41	123419	161146	376	65	4	16.96	60	0.39	High Scho	Self-emplc	Single	Yes	No	Other	Yes	0
7ASF0IHRI	61	30142	133714	429	96	1	15.58	12	0.65	PhD	Part-time	Divorced	No	Yes	Business	No	0

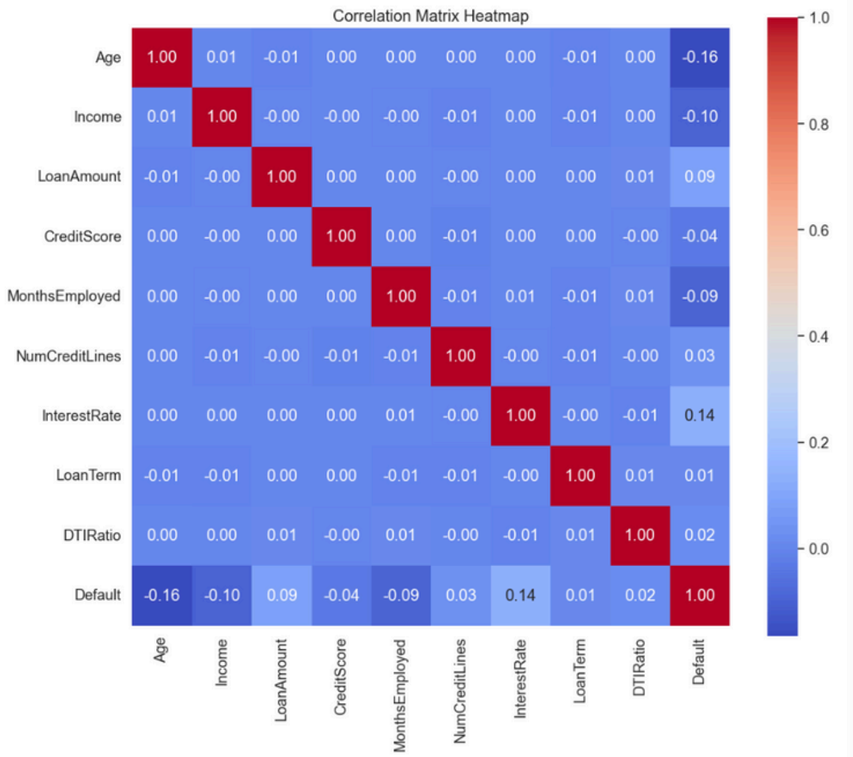
25,000 rows × 18 columns

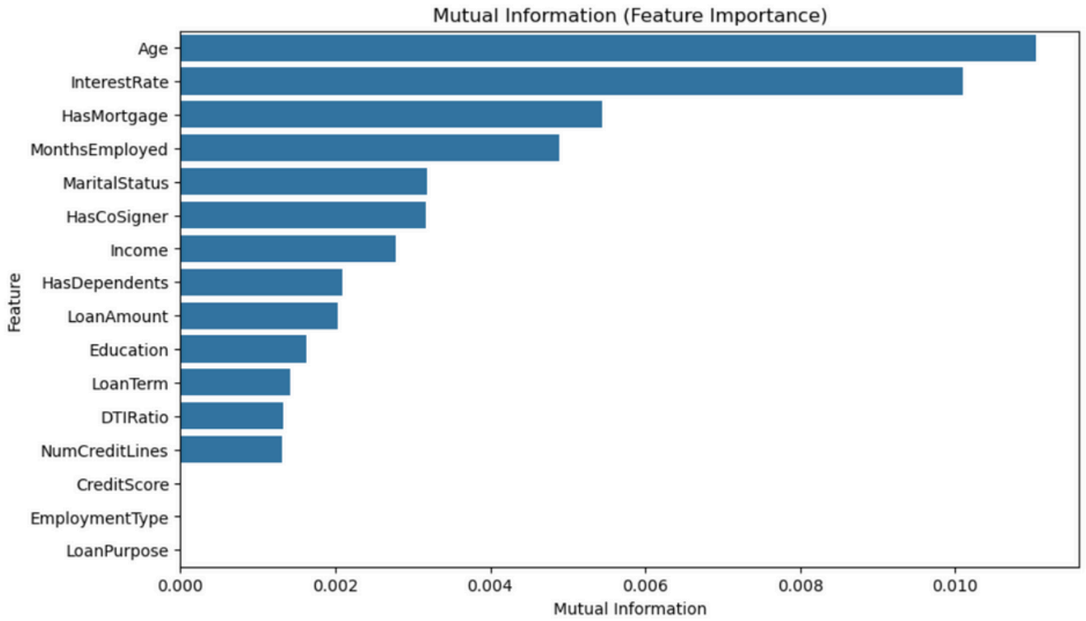
은행 고객들의 직업, 계좌 및 대출 개수, 연체 일수 등의 개인정보가 담긴 데이터

데이터 탐색

변수명	변수 이름	변수명	변수 이름
LoanID	대출 ID	Education	최종학력
Age	나이	EmploymentType	직업 종류
Income	수입	MartialStatus	결혼 여부
LoanAmount	대출 금액	HasMortgage	주택담보대출 여부
CreditScore	신용 점수	HasDependents	부양가족 여부
NumCreditLines	한도 대출 횟수	LoanPurpose	대출 목적
LoanTerm	대출 기간	HasCoSigner	공동서명자 여부
DTIRatio	충부채 상환 비율	Default	채무불이행 여부

상관 관계, 상호정보량 (MI값)



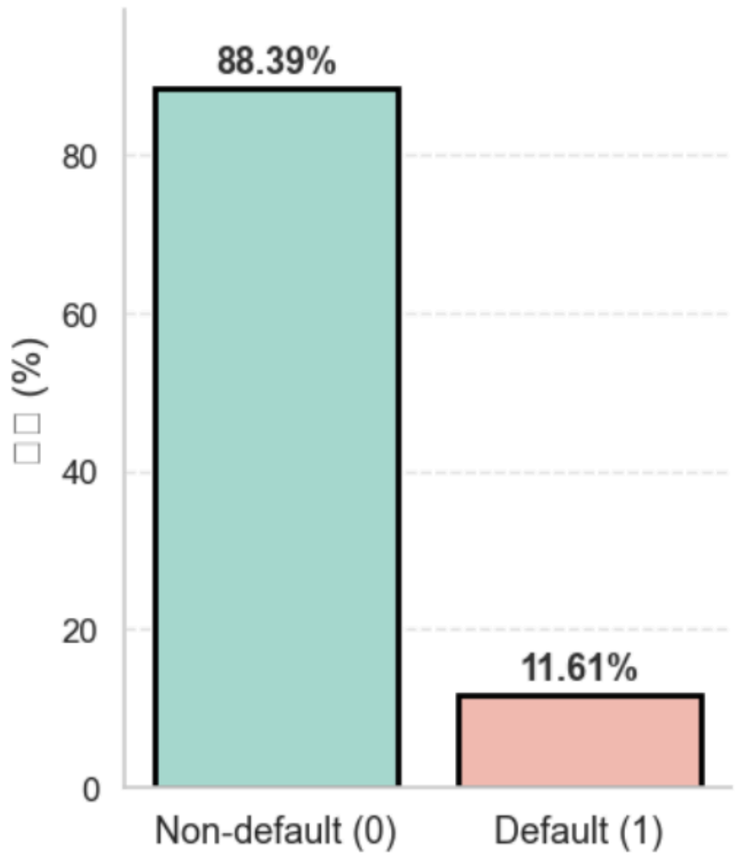


상관 관계 와 MI값을 기준으로 가장 관계가 약한 변수 제거 (4개)

Loan ID, Loan term, HasMortage, DTI Ratio

데이터 분석 및 전처리

전체 데이터의 **Default** 클래스



클래스의 불균형이 심한 **Imbalanced Dataset**

정확도가 높게 나올수 있지만 정밀도와 재현율이 낮을 가능성

대출 연체 예측을에서 소수 클래스를 예측하는 것이 더 중요

과적합 방지를 위해 샘플링 필요

데이터 전처리 : 결측치

```
print(df.isnull().sum())
```

```
Age                143
Income               0
LoanAmount           0
CreditScore          0
MonthsEmployed       0
NumCreditLines       0
```

Age컬럼의 결측치 제거 (143개)

24857 rows

데이터 전처리 : 이상치

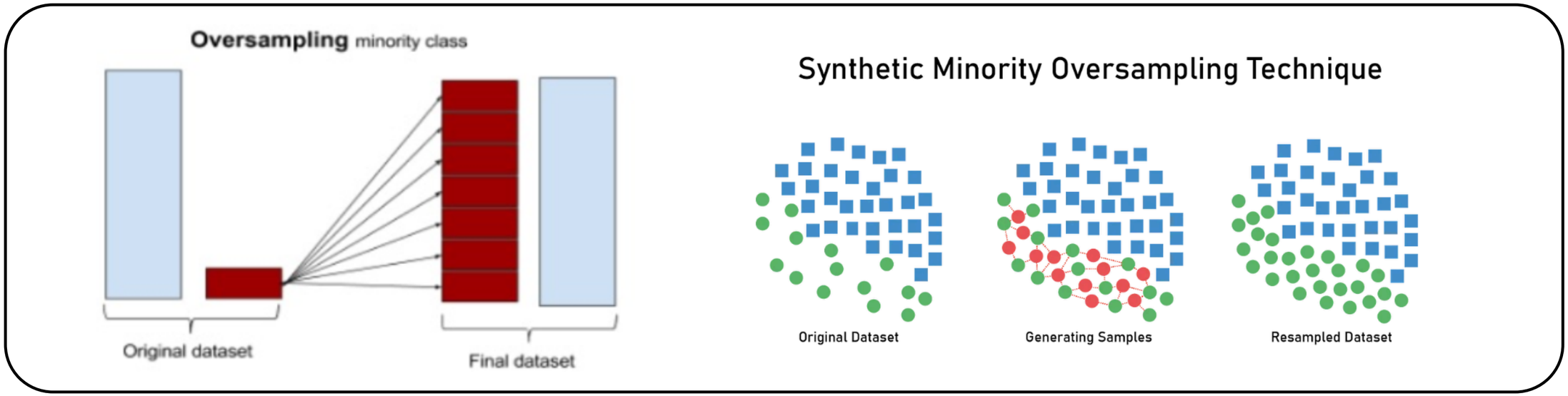
	Q1	Q3	IQR
Age	31.00	57.00	26.0
Income	48900.00	116174.00	67274.0
LoanAmount	66531.00	189292.00	122761.0
CreditScore	436.00	712.00	276.0
MonthsEmployed	30.00	90.00	60.0
NumCreditLines	2.00	4.00	2.0
InterestRate	7.76	19.26	11.5
LoanTerm	24.00	48.00	24.0
DTIRatio	0.30	0.70	0.4
Default	0.00	0.00	0.0

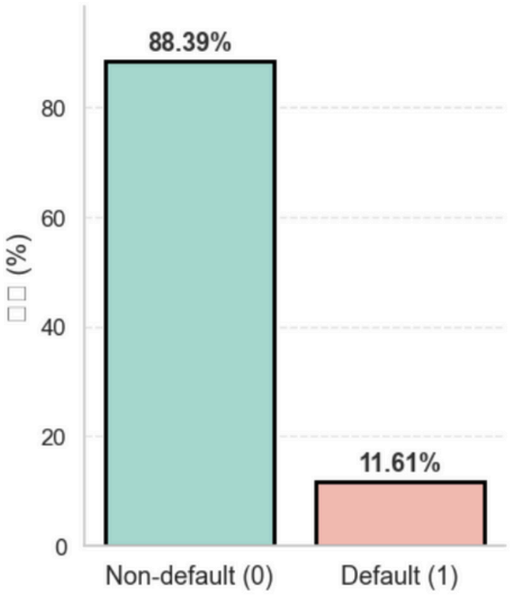
IQR 기반 이상치 탐지 결과, 본 데이터에서는 이상치가 존재하지 않는 것으로 판단

인코딩 방식

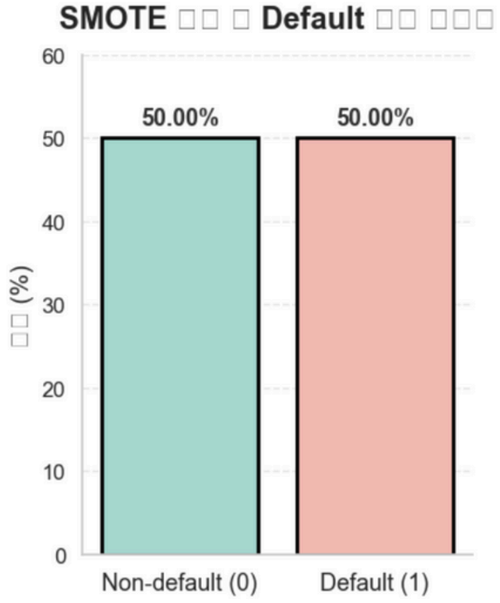
Binary Encoding	Yes는 1, No는 0으로 변환 , map() 함수 사용 대상: HasDependents, HasCoSigner
Ordinal Encoding	학력 수준 별로 순서 부여, 정수값 변환 (High School < Bachelor's < Master's < PhD) 대상: : Education
One-Hot Encoding	각 범주를 더미 변수(0/1)로 변환, 첫 번째 범주는 삭제 대상: EmploymentType, MaritalStatus, LoanPurpose

데이터 분석 및 전처리





SMOTE



모델 성능 비교

모델 성능 비교

	Random Forest	SVM	XGBoost	Catboost	LightGBM	KNN
Accurancy	0.84	0.84	0.83	0.87	**0.86**	0.75
Precision	0.25	0.26	0.29	0.38	**0.33**	0.20
Recall	0.18	0.21	0.29	0.15	**0.17**	0.35
F1 Score	0.21	0.23	0.29	0.21	**0.23**	0.25
ROC_AUC	0.68	0.65	0.68	0.70	**0.71**	0.62

ROC-AUC : 분류 모델의 전체적인 예측 성능

F1 score : Precision과 Recall의 조화 평균

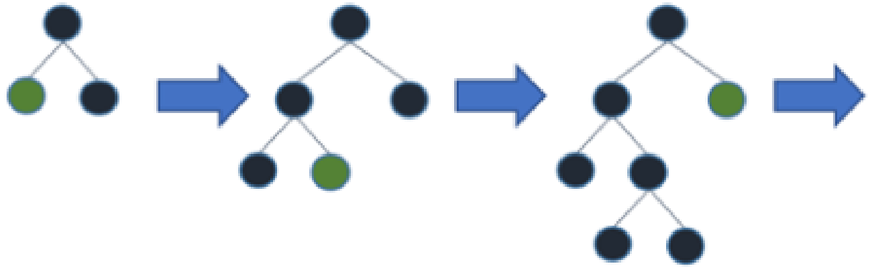
LightGBM 최적 모델 선정

LightGBM

<알고리즘 설명>

Gradient Boosting 기반의 트리 앙상블 알고리즘

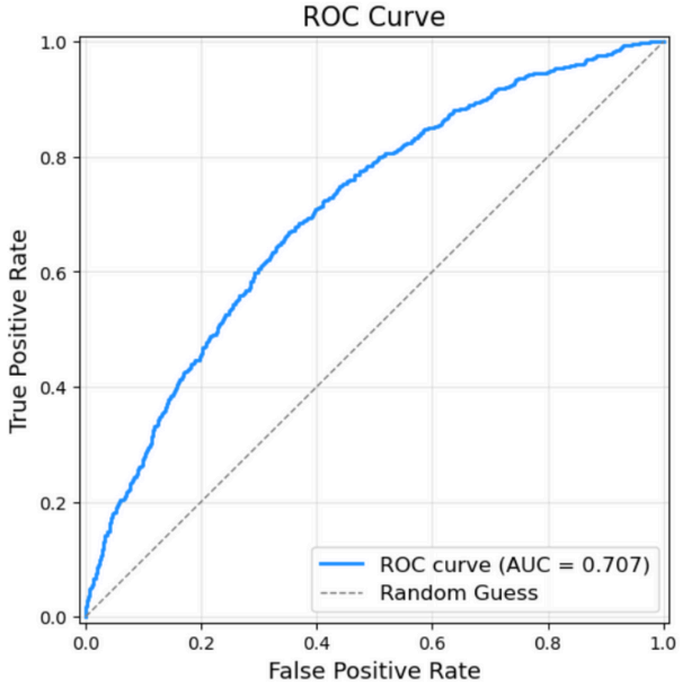
- 중요한 데이터에 더 집중해 학습 속도와 효율을 높임
- 비슷한 변수는 하나로 묶어 변수 수를 크게 줄임(EFB)
- 트리 구조 수직 확장(leaf-wise): 더 깊게 트리를 생성
- 기존 GBDT 대비 10~20배 빠른 속도 높은 예측 성능과 확장성



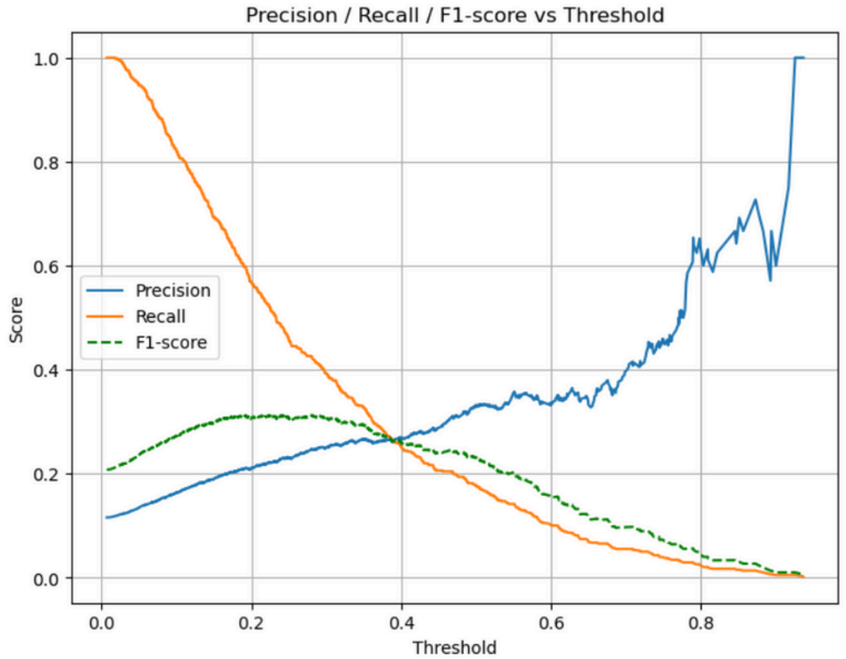
Leaf-wise tree growth

성능 평가

ROC_Curve



임계값 - (Recall/ Precision/F1)

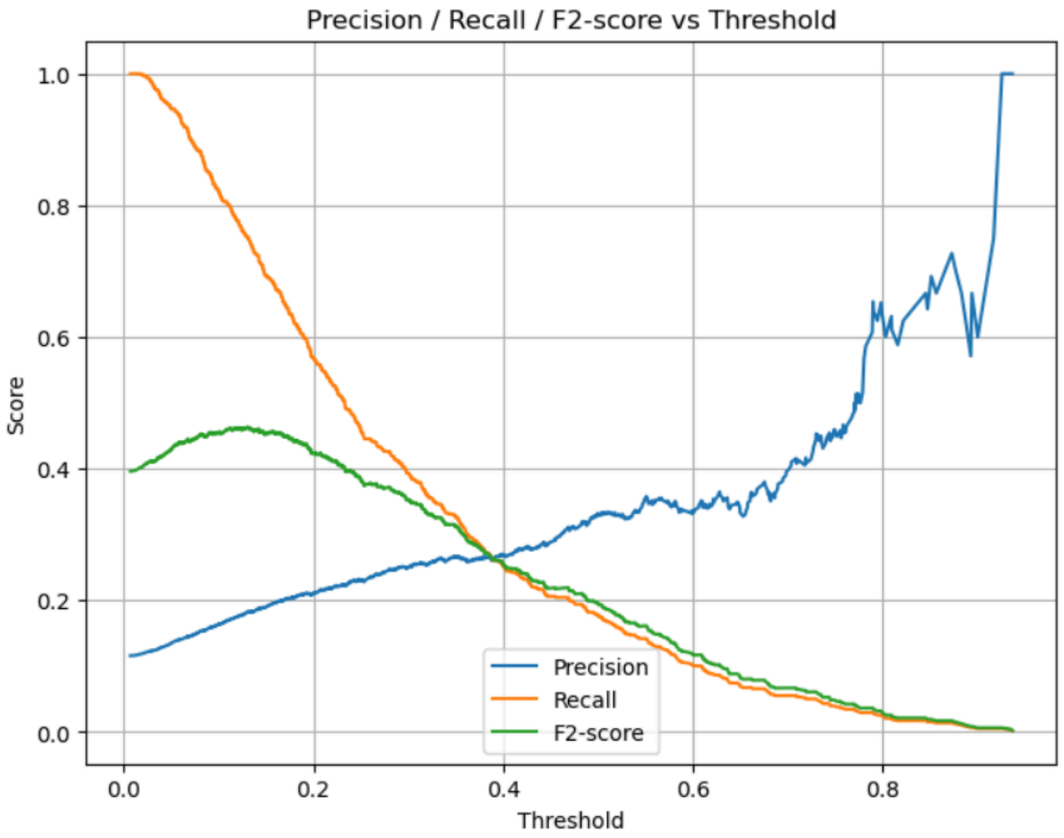


임계값 0.2 와 0.4의 구간에서 F1 score는 비교적 비슷하게 유지

하지만 Recall 값은 급격하게 하락

성능 평가

임계값 - (Recall /Precision / F2-score)



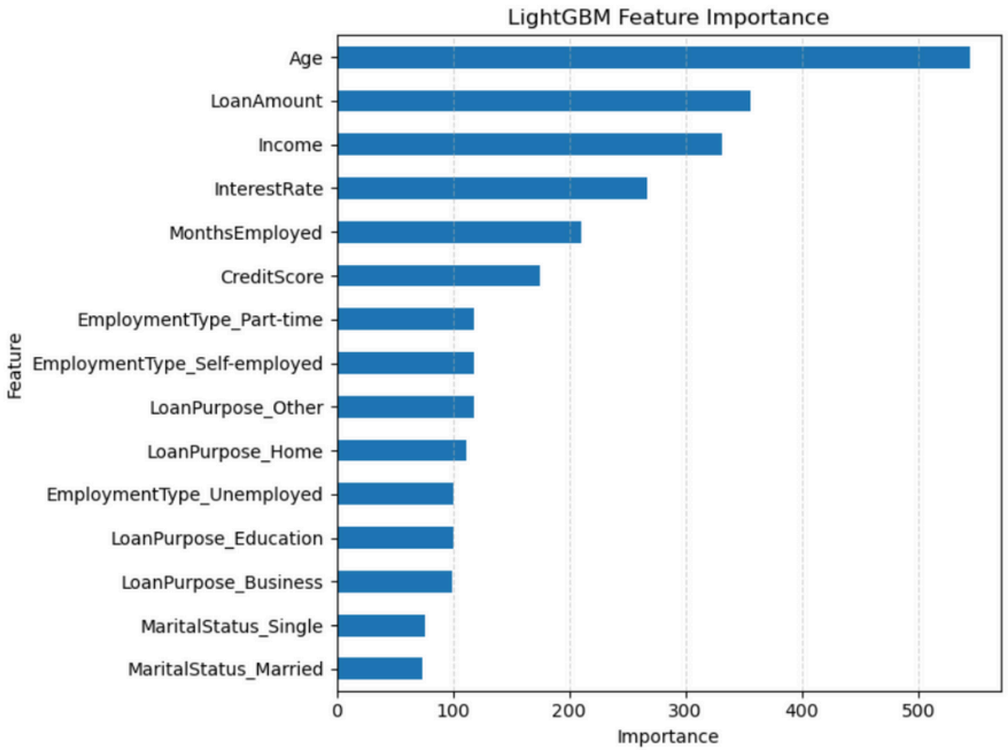
F2-score

- **Precision과 Recall 중 Recall에 더 큰 가중치를 두는 평가 지표**
- 실제 양성을 최대한 놓치지 않는 것이 목적
- F2 Score가 최대가 되는 0.18을 최적의 임계값으로 설정

Accuracy: 0.58
Precision: 0.18
Recall: 0.75
F1-score: 0.29
ROC-AUC: 0.71

모델 해석

Feature Importance



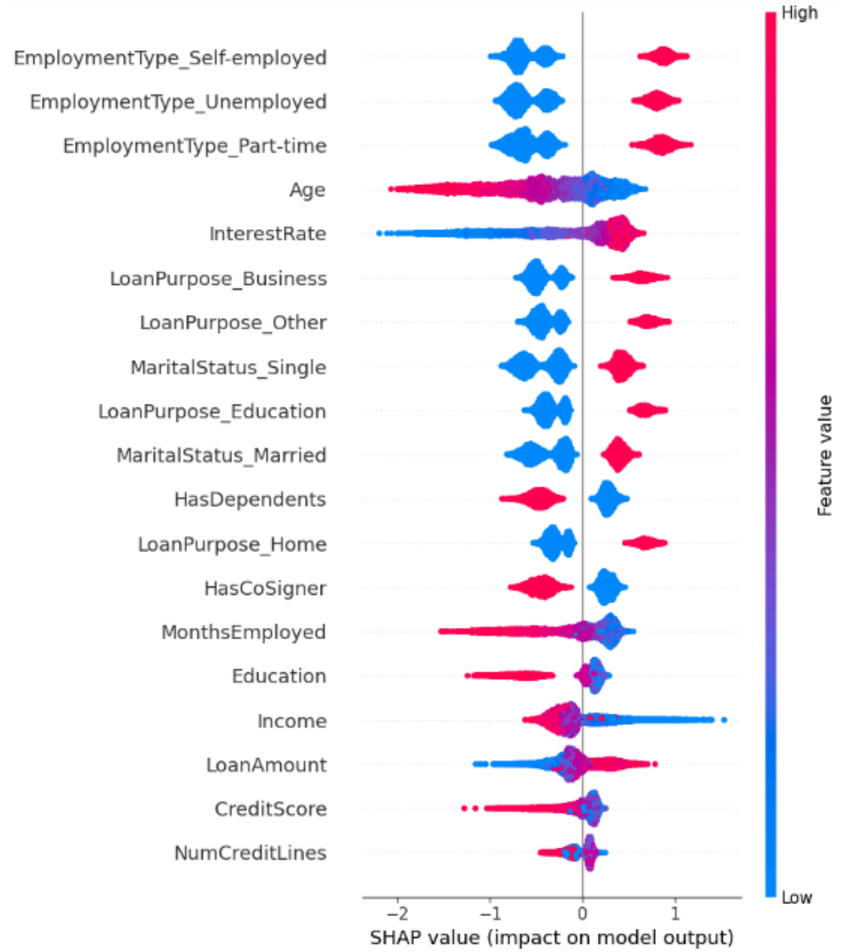
모델이 트리를 만들 때 '몇 번' 또는 '얼마나 많이' 각 변수를 분할에 사용했는지, 즉 변수의 사용 빈도와 기여도를 단순히 합산해서 중요도를 계산

변수 간 상호작용, 실제 예측에 끼치는 '방향성' 은 고려 안함

'Age', 'LoanAmount', 'Income', 'InterestRate' MonthEmployed 지표가 높게 나타남

모델 해석

SHAP values



각 변수의 SHAP value는 모델의 각 예측에서 해당 변수가 실제로 얼마나, 어느 방향(+), 얼마만큼'예측에 영향을 미쳤는지 정량적으로 계산

변수의 영향력뿐만 아니라, 예측값 변화에 기여한 실질적인 효과를 정밀하게 측정

EmploymentType(고용형태) 관련 변수들이 예측에 가장 큰 영향을 미쳤으며, 그 뒤로 Age(나이), InterestRate(이자율), LoanPurpose(대출 목적) 등이 중요한 특성으로 확인

한계점 및 개선방향

한계점 및 개선방향

한계점

임계값을 조정해 Recall값을 높였지만 정확도(Accuracy), 정밀도(Precision), F1-score등 다른 주요 평가지표에서는 낮은 수치를 기록하여 오탐율이 높아 예측의 신뢰도가 떨어지는 한계점이 존재하였다.

개선 방향

변수 간 상호작용과 비선형성을 반영할 수 있도록 다양한 피처 엔지니어링, 파생변수와 같은 변환 기법을 적용하고 스태킹이나 블렌딩과 같은 복합적 모델링 기법을 도입해 모델의 성능을 높인다. 대출 연체 예측은 모의 해석력도 중요하므로 설명 가능한 AI기법(SHAP, LIME 등)을 활용하여 예측의 신뢰성을 높이는 방식이 개선 방향으로 제시된다.

감사합니다