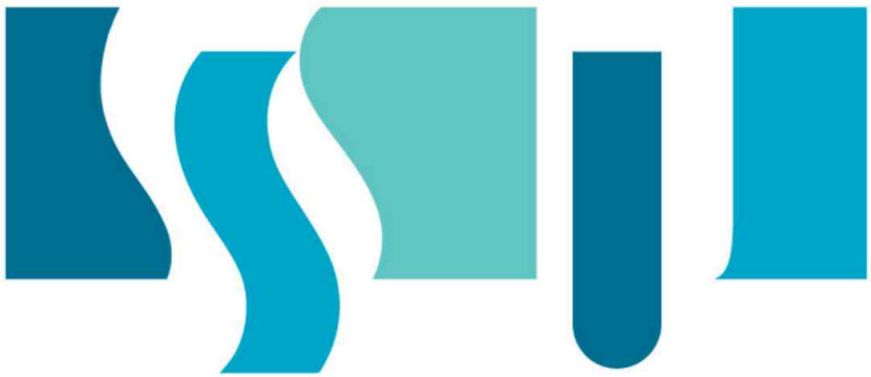
머신러닝을 활용한
대출 연체 예측



(가)반

산업정보시스템공학과	
20212843	권용후
20211290	심승현
20211322	함유민

목차

1. 서론

1.1. 연구 소개

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

1.2. 연구 동기

대출 연체는 금융기관의 수익성과 직결되는 주요 위험 요인으로, 이를 정확히 예측하는 것은 금융 비즈니스의 안정성에 있어서 매우 중요하다. 기존에는 재무제표나 신용평점 등 제한된 정형 데이터와 전문가의 정성적 판단에 의존해왔다. 그러나 최근에는 모바일, 온라인 거래 등 다양한 채널을 통해 수집되는 대규모·비정형 데이터를 활용하여 고객의 연체 가능성을 보다 정밀하게 예측할 수 있게 되었다. 이에 따라 금융기관은 머신러닝 기반의 정량적이고 자동화된 예측 시스템 도입을 통해 리스크 대응 역량을 강화할 필요가 있다.

1.3. 연구의 필요성

금융 산업에서 고객의 신용 리스크를 사전에 예측하고 효과적으로 관리하는 것이 금융기관의 안정적 운영과 건전성 유지에 있어 매우 중요한 과제가 되었다. 특히, 대출 연체가 발생할 경우 금융사의 직접적인 손실뿐만 아니라, 고객 신뢰 하락 및 금융 시스템 전반의 위험 확대로 이어질 수 있다. 기존의 전통적인 신용평가 모델은 복잡한 데이터 구조와 비선형적 패턴을 충분히 반영하지 못하는 한계가 있으며, 실제 연체자를 놓치는 경우가 빈번하게 발생하고 있다. 이에 따라, 최신 데이터마이닝 및 머신러닝 기법을 활용해 연체 위험군을 조기에 탐지하고, 리스크 관리의 효율성을 높이는 것이 필수적인 상황이다.

1.4. 연구 목표

프로젝트의 목표는 다양한 고객 특성 데이터를 바탕으로 대출 연체 여부를 효과적으로 예측할 수 있는 최적의 머신러닝 모델을 구축하는 데 있다. 이를 위해, 데이터 전처리, 이상치 및 불균형 데이터 처리, 피처 선택 등 체계적인 데이터마이닝 과정을 거쳐 RandomForest, SVM, XGBoost, LightGBM 등 여러 예측 모델을 적용하고, 성능을 다각적으로 비교·평가한다. 또한 Feature Importance 및 SHAP 등 설명 가능한 AI 기법을 통해 주요 변수의 영향력을 해석하고, 실무적 의사결정에 활용할 수 있는 실질적 인사이트를 제공하는 것을 최종적 목표로 한다.

2. 관련 연구

① 상호 저축은행 연체 예측을 위한 신용평점 도출에 관한 연구

이 논문에서는 상호저축은행의 대출 연체 여부를 더 정확히 예측하기 위해 기존 신용평가사의 평가와 데이터마이닝 기반 예측모델의 성능을 비교하는 것이 목적이다. 국내 모 상호저축은행의 실제 대출 고객 데이터 7,923건을 활용하여 C5.0, CHAID, C&RT, 신경망, 로지스틱 회귀 등 5가지 지도학습 알고리즘을 적용하였고, 여기서 C5.0 모델이 가장 정확도가 높았다. 예측된 신용 점수는 MDLP 알고리즘을 통해 등급화되었으며, 기존 신용평가사 평가보다 연체 고객과 우량 고객의 구분력이 뛰어났고, Divergency 지표를 통해 데이터마이닝 모델이 신용평가사보다 고객 구분력이 훨씬 높다는 점이 입증되었다. 특히 '직장인 대출 상품'에서 데이터마이닝 모델은 1등급 연체율 0%, 3등급 연체율 99.59%로 매우 선명한 구분 결과를 보였다.결론적으로 이 논문은 상호저축은행은 신용평가사 정보에만 의존하기보다는 자체 데이터와 데이터마이닝 모델을 병행 활용하여 신용평가 시스템을 구축해야 한다고 제안한다.

② 데이터마이닝 기법을 활용한 개인 신용평점모형 개발

이 논문은 기존 단변량 분석 방식의 한계인 변수 간 결합 효과가 반영되지 않는다는 문제를 지적하고, 이를 보충하기 위해 데이터마이닝을 기반으로 한 변수 선택 기법을 제안하였다. 변수 선택에는 종속변수와의 높은 연관성과 낮은 중복성을 고려한 mRMR(Minimum Redundancy Maximum Relevance) 알고리즘과, 변수의 예측 기여도를 측정하는 PLS-VIP(Partial Least Squares - Variable Importance in Projection) 기법을 활용하였으며, 선정된 변수들을 바탕으로 로지스틱 회귀 모델을 구축하였다. 모델 성능은 AUROC, K-S 통계량 등을 통해 평가되었고, 기존 방식에 비해 우수한 예측력을 보였다. 이를 통해 데이터마이닝 기반 변수 선택이 신용평가 모델의 성능 향상에 효과적임을 증명하였다.

③ 설명 가능한 인공지능을 활용한 은행 대출 연체 예측 모델 연구

이 논문에서는 기계학습 기법인 XGBoost와 설명 가능한 인공지능(XAI) 기법인 SHAP을 결합하여 은행 대출 연체 여부를 예측하는 모델을 구축하였으며, 국내 A 은행의 2020~2022년 가계 대출 계좌 75만 건 데이터를 활용하면서 클래스 불균형 문제를 해결하였다. XGBoost는 기존 로지스틱 회귀 모델 대비 높은 예측 성능을 보여주었고, 특히 재현율, F1 Score, Kappa 값에서 훌륭한 결과를 보였다. SHAP 분석을 통해 영향을 주는 주요 변수로는 신용평점, 예금잔액, 금리, 채무위험등급이 확인됐으며, 신용평점과 예금잔액이 낮고, 금리와 채무위험등급이 높을수록 연체 가능성이 증가하는 경향이 도출되었다. 또한 SHAP Force Plot을 활용해 개별 계좌에 대한 변수 영향력을 시각화함으로써 모델의 해석력을 확보하였다. 결과적으로, XAI를 결합한 XGBoost 모델은 높은 예측력과 설명력을 동시에 갖춘 대출 연체 예측 도구로, 실무 금융 현장에서 로지스틱 회귀의 효과적인 대안이 될 수 있음을 제시하였다.

3. 본론

3.1. 데이터 소개

이 데이터는 은행 고객들의 직업, 계좌 및 대출 개수, 연체 일수 등의 개인정보가 담긴 자료로, 총 25,000개의 행과 18개의 열로 이루어져 있다. 10개의 수치형 데이터, 8개의 범주형 데이터로 이루어져 있고, 각 컬럼에 대한 설명은 아래와 같다.

데이터 출처 (Kaggle - Loan Default Prediction Data set)

LoanID	대출 ID
Age	나이
Income	수입
LoanAmount	대출 금액
CreditScore	신용 등급
NumCreditLines	한도 대출 횟수
InterestRate	이자율
LoanTerm	대출 기간
DTIRatio	총부채 상환 비율
Education	최종학력
EmploymentType	직업 종류
MaritalStatus	결혼 여부
HasMortgage	주택 담보 대출 여부
HasDependents	부양가족 여부
LoanPurpose	대출 사유
HasCoSIgner	대출 공동 서명자 여부
Default	채무 불이행 여부

3.2 데이터 탐색

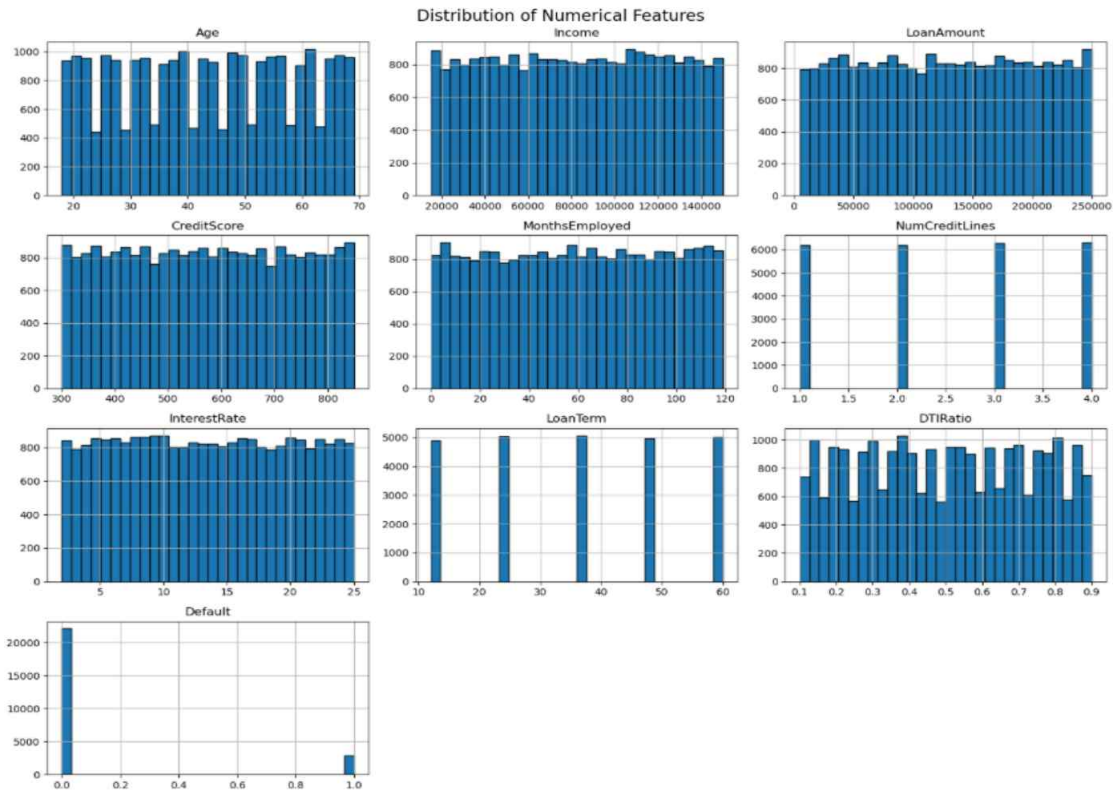
3.2.1 주요 변수별 기초통계량 요약

먼저 describe() 함수를 통해 데이터셋의 기초 통계량에 대해 확인하였다. describe() 함수의 결과는 각 변수의 데이터 분포 특성(중앙값, 평균, 최소/최대, 사분위수 등)을 파악할 수 있도록 해주며, 이를 통해 데이터의 이상치 여부, 분포의 비대칭성, 스케일 차이 등도 1차적으로 확인하였다.

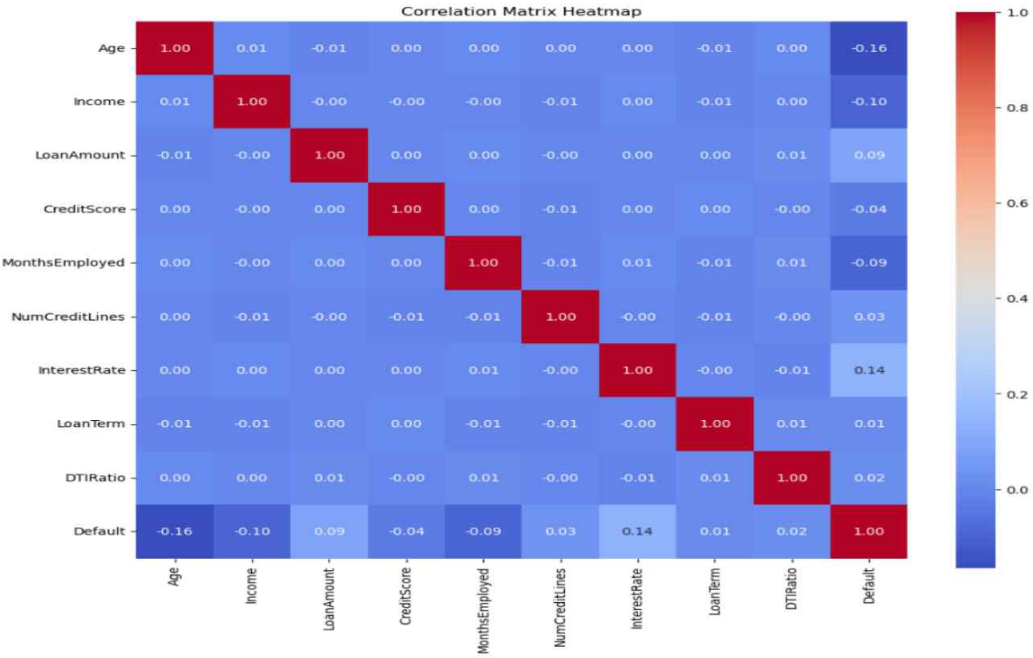
	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio	Default
count	24857.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000
mean	43.609124	82698.812560	128022.278320	574.446400	59.833200	2.509520	13.475750	36.077280	0.499893	0.116120
std	15.016923	38931.273391	70808.416376	159.127712	34.798041	1.118266	6.640217	16.906965	0.230697	0.320375
min	18.000000	15010.000000	5000.000000	300.000000	0.000000	1.000000	2.000000	12.000000	0.100000	0.000000
25%	31.000000	48886.250000	66439.000000	436.000000	30.000000	2.000000	7.747500	24.000000	0.300000	0.000000
50%	44.000000	82900.500000	128095.000000	574.000000	60.000000	3.000000	13.420000	36.000000	0.500000	0.000000
75%	57.000000	116176.750000	189281.250000	712.000000	90.000000	4.000000	19.260000	48.000000	0.700000	0.000000
max	69.000000	149994.000000	249976.000000	849.000000	119.000000	4.000000	25.000000	60.000000	0.900000	1.000000

3.2.2 데이터 시각화를 통한 EDA

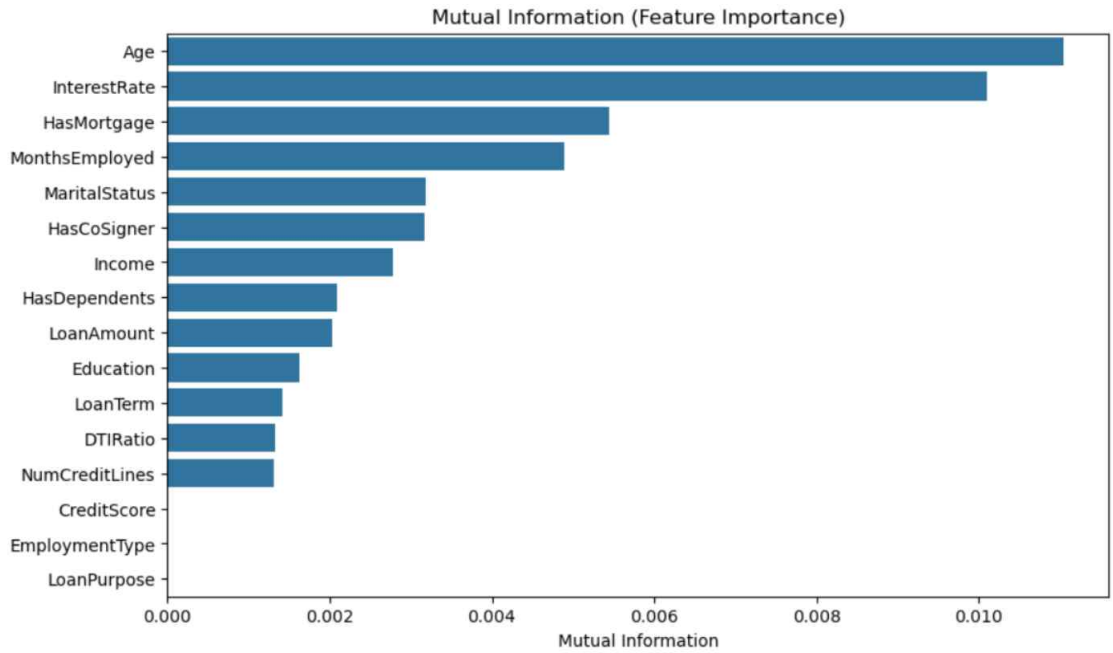
개별 변수의 데이터 분포



변수별 상관관계 시각화



변수별 MI (Mutual Information)값 시각화



개별 변수의 데이터 분포를 확인하였을때 대부분의 수치형 변수는 전반적으로 고르게 분포되어있는 상태(로그변환 필요 없음). Default(연체)와 가장 상관이 높은 변수는 Age, Income, LoanAmount, Interate이나 이 변수들 또한 절대값 0.2 이하로 전체적으로 약한 상관관계를 가진다. 상호정보량 기준으로도 Age(나이)와 InterestRate(이자율), HasMortgage(주택담보대출 보유 여부)등이 연체 예측에 중요한 피처로 나타나지만 MI값 역시 절대값 자체가 높지는

않다는 것을 확인 할 수 있었다. 단일 변수로는 연체 여부를 명확히 구분하기 어렵고, 여러 변수를 조합한 복합적인 분석 및 모델링 필요하고 Feature 선택 진행시 Filter Method보다 Wrapper Method가 더 적합하다고 판단하였다.

3.3 데이터 전처리

3.3.1 데이터 결측치 처리

데이터 셋의 결측치를 확인 해본 결과 Age컬럼에 143개의 결측치가 존재하였다. 위의 상관관계 분석에서 Default(타깃값)과 Age컬럼의 상관관계가 0.16으로 가장 높은 수준이기에 결측치 대체가 오히려 잘못된 정보를 줄 수 있다고 판단해 143개의 결측치를 삭제하였다.

3.3.2 데이터 이상치 처리(IQR)

	Q1	Q3	IQR	Lower Bound	Upper Bound
Age	31.00	57.00	26.0	-8.00	96.00
Income	48900.00	116174.00	67274.0	-52011.00	217085.00
LoanAmount	66531.00	189292.00	122761.0	-117610.50	373433.50
CreditScore	436.00	712.00	276.0	22.00	1126.00
MonthsEmployed	30.00	90.00	60.0	-60.00	180.00
NumCreditLines	2.00	4.00	2.0	-1.00	7.00
InterestRate	7.76	19.26	11.5	-9.49	36.51
LoanTerm	24.00	48.00	24.0	-12.00	84.00
DTIRatio	0.30	0.70	0.4	-0.30	1.30
Default	0.00	0.00	0.0	0.00	0.00

데이터의 1사분위(Q1)와 3사분위(Q3) 사이의 범위를 이용해 이상치를 탐지하는 IQR방식을 활용하여 이상치를 탐지하였지만 분석에 방해가 되는 이상치는 없다고 판단하였다.

3.3.3 범주형 변수의 인코딩 전략

이진 인코딩 (Binary Encoding)

두 가지 값(Yes/No, 1/0)만 가지는 이진 변수에 대해 각 값을 0과 1로 변환한는 인코딩 방식.

적용 변수 : HasMortgage, HasDependents, HasCoSigne

위의 변수들은 단순 이진 변수이므로, 복잡한 변환 없이 0/1로만 처리하는 것이 효율적이다. 원-핫 인코딩과 같이 변수 수가 불필요하게 늘어나지 않기에 적용하였다.

순서형 인코딩 (Ordinal Encoding)

순서가 존재하는 범주형 변수(학력)는 각 범주를 순서에 맞는 정수로 변환하는 인코딩 방식.

적용 변수 : Education (High School < Bachelor's < Master's < PhD)

단순 원-핫 인코딩은 순서 정보를 표현하지 못하기 때문에 교육 수준처럼 순서가 중요한 변수는 순서 정보를 수치적으로 반영하는 순서형 인코딩 방식을 사용하였다.

원-핫 인코딩 (One-hot Encoding)

값의 우열이나 순서가 없는 명목형 변수를 각 카테고리 별로 새로운 컬럼을 만들어 1 또는 0으로 표현하는 인코딩 방식.

적용 변수 : EmploymentType, MaritalStatus, LoanPurpose

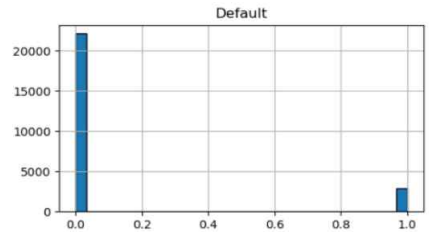
위의 변수들은 순서가 없으므로 원-핫 인코딩을 통해 모델에 불필요한 순서 정보를 반영하지 않고 인코딩 실행. drop_first=True를 통해 첫번째 카테고리를 삭제하여 변수들간의 상호작용을 강화시킬 수 있는 다중공선성을 방지하였다.

3.3.4 모델링 준비

Train Set/Test set 분할

전체 데이터를 학습 데이터(Train Set)와 테스트 데이터(Test Set)로 분할하였다.
train_test_split함수를 사용해 전체 데이터의 80%는 학습용, 20%는 테스트용으로 나누었으며, 타깃 변수의 클래스 비율이 학습/테스트셋에 동일하게 유지되도록 stratify옵션을 적용하였다.

클래스 불균형 처리 (SMOTE, Over-sampling, Under-sampling, ADASYN)




```
Default 컬럼 클래스별 비율(%):
 Default
0    88.39
1    11.61
Name: proportion, dtype: float64
```

현재 학습 데이터는 Default 클래스가 모델이 소수 클래스 예측을 잘 못할 수 있는 위험이 존재한다. 이를 보완하기 위해 SMOTE(Synthetic Minority Over-sampling Technique)를 사용해 학습(Train) 데이터 내 소수 클래스(Defult)의 표본을 인위적으로 생성하여 클래스의 균형

을 유지하였다. 오버샘플링 외에 언더샘플링(다수 클래스 데이터축소) 기법이나 ADASYN 기법도 고려할 수 있으나, 데이터 손실이 없는 SMOTE, ADASYN을 우선 선택하였고 베이스라인 성능평가에서 SMOTE 기법이 더 좋은 평가를 받아 최종적으로 SMOTE기법을 적용하였다.

```
SMOTE 적용 후 Default 클래스별 개수:
 Default
0    17576
1    17576
Name: count, dtype: int64
SMOTE 적용 후 Default 클래스별 비율(%):
 Default
0    50.0
1    50.0
Name: proportion, dtype: float64
```

수치형 변수 표준화 (Standard Scaling)

모델링 과정에서는 각 변수의 스케일(단위, 범위)이 서로 크게 다를 경우 성능이 저하될 위험이 존재한다. 이를 방지하기 위해 수치형 변수에 대해 평균 0, 표준편차 1로 변환하는 표준화를 적용하였다. ColumnTransformer로 사용하여 지정된 수치형 변수에만 변환을 적용하고 나머지 변수는 원본 그대로 유지하였다.

3.4 모델링

3.4.1 활용 알고리즘(RandomForest, LightGBM, SVM, XGBoost, CatBoost, KNN)

Random Forest

Random Forest는 여러 개의 결정 트리(Decision Tree)를 조합해 예측을 수행하는 대표적인 앙상블 학습 방법이다. 각 트리는 원본 데이터에서 일부 샘플(bootstrapping)과 일부 피처를 무작위로 선택하여 독립적으로 학습한다. 이렇게 학습된 트리들의 예측을 다수결(분류) 또는 평균(회귀) 방식으로 통합하여 최종 결과를 도출한다. Random Forest는 개별 트리보다 예측 성능이 높고, 과적합(overfitting)을 효과적으로 방지한다는 장점이 있다. 또한, 피처 중요도(feature importance) 평가에도 유용하여, 어떤 변수가 예측에 영향을 많이 미치는지 파악할 수 있다.

LightGBM

LightGBM(Light Gradient Boosting Machine)은 부스팅 기반의 결정 트리 알고리즘으로, 대규모 데이터셋에도 빠른 학습 속도와 뛰어난 성능을 제공한다. 기존의 gradient boosting 방법과 달리, 히스토그램 기반 학습과 리프 중심(leaf-wise) 성장 방식을 선택해 연산 효율성을

크게 개선한다. 이로 인해 메모리 사용량을 절감하면서도 복잡한 데이터에서도 강력한 예측력을 보인다. LightGBM은 범주형 변수 자동 지원, 결측치 처리 등 활용도가 높다.

SVM

SVM(Support Vector Machine)은 데이터를 분리하는 최적의 초평면(hyperplane)을 찾는 지도학습(supervised learning) 알고리즘이다. SVM은 마진이 최대가 되는 경계를 찾아 분류의 신뢰도를 높이는 것이 특징이다. 데이터가 선형적으로 구분되지 않는 경우에도 커널 트릭(kernel trick)을 이용해 데이터를 고차원 공간으로 매핑함으로써 복잡한 패턴을 효과적으로 학습할 수 있다.

XGBoost

XGBoost(eXtreme Gradient Boosting)는 부스팅(Boosting) 방식의 결정 트리 알고리즘 중 하나로, 속도와 성능 모두에서 우수하다는 평가를 받는다. L1, L2 정규화 적용, 결측값 자동 처리, 병렬 학습, early stopping 등 다양한 기능을 제공하여, 데이터가 복잡하고 대규모인 상황에서도 높은 예측력을 발휘한다. 또한, 하이퍼파라미터 튜닝을 통해 모델의 복잡도와 일반화 성능을 세밀하게 조절할 수 있다.

CatBoost

CatBoost(Categorical Boosting)는 gradient boosting 기반의 알고리즘으로, 범주형 변수 처리에 강점을 가진다. 데이터 전처리 없이 범주형 피처를 자동으로 인코딩할 수 있으며, 과적합 방지 및 결측값 처리 기능도 내장되어 있다. CatBoost는 트리 기반 부스팅 모델의 장점은 그대로 유지하면서 자주 발생하는 범주형 데이터의 효율적인 처리로 실질적인 성능 향상이 가능하다.

KNN

KNN(K-Nearest Neighbors)은 예측하고자 하는 데이터 포인트와 가장 가까운 K개의 이웃을 찾아, 이웃들의 결과(분류: 최빈값, 회귀: 평균)를 바탕으로 예측하는 비모수 알고리즘이다. 데이터의 분포나 패턴을 사전에 가정하지 않기 때문에 단순하고 직관적이며, 설명력이 높은 편이다. 하지만 데이터가 많아질수록 거리 계산 등으로 인해 계산량이 크게 증가하는 단점이 있다. 또한, 이상치(outlier)나 스케일(거리 기준)에 민감하기 때문에 사전 데이터 전처리(정규화)가 필요하다.

위의 6가지 알고리즘**(**RandomForest, LightGBM, SVM, XGBoost, CatBoost, KNN**)**을 활용하여 모델링을 진행하였다.

4. 평가 및 분석

4.1 평가 지표 (Accuracy, Precision, Recall, F1, F2 Score ,ROC_AUC)

Accuracy(정확도)
전체 샘플 중에서 모델이 올바르게 예측한 비율. 클래스 불균형이 심할 경우 한계 존재.

Precision (정밀도)

Precision_0: 0(비연체)로 예측한 것 중 실제로 비연체인 비율

Precision_1: 1(연체)로 예측한 것 중 실제로 연체인 비율

정밀도가 높을수록 잘못된 양성 예측(False Positive)이 적음을 의미한다.

Recall (재현율)

Recall_0: 실제 비연체 중 비연체로 맞춘 비율

Recall_1: 실제 연체 중 연체로 맞춘 비율

재현율이 높을수록 놓치는 양성(False Negative)이 적음을 의미한다.

F1_Score (F1 점수)

F1_0: 비연체에 대한 F1 점수
F1_1: 연체에 대한 F1 점수

정밀도와 재현율의 균형을 반영한 종합 지표로, 불균형 데이터에 유용하다.

F2_Score (F2 점수)

F2_0: 비연체(0)에 대한 F2 점수
F2_1: 연체(1)에 대한 F2 점수
F2 점수는 정밀도(Precision)와 재현율(Recall)의 조화평균 중, 재현율에 더 큰 가중치(β=2)를 둔 종합 지표. 특히 놓치는 양성에 더 민감하게 반응해야 할 때 유용하며, 재현율의 중요성이 정밀도보다 2배 더 클 때 사용된다.

ROC_AUC
ROC 곡선 아래 면적으로, 1에 가까울수록 연체/비연체 구분을 잘함을 의미한다. 클래스 불균형에서도 신뢰성이 높다.

4.2 모델 성능 평가표

	Model	Accuracy	Precision_0	Precision_1	Recall_0	Recall_1
0	RandomForest	0.858407	0.894085	0.280277	0.952673	0.140381
1	XGBoost	0.875704	0.895663	0.396985	0.972696	0.136915
2	LightGBM	0.879525	0.894595	0.432099	0.979067	0.121317
3	CatBoost	0.877514	0.895694	0.414894	0.974972	0.135182
4	KNN	0.735519	0.906762	0.188870	0.781115	0.388215
5	SVM	0.842518	0.897973	0.262673	0.927190	0.197574

	F1_0	F1_1	ROC_AUC
0	0.922450	0.187067	0.685374
1	0.932592	0.203608	0.700426
2	0.934927	0.189445	0.719145
3	0.933653	0.203922	0.707143
4	0.839262	0.254112	0.630765
5	0.912347	0.225519	0.651356

여러 분류 모델의 성능을 비교한 결과 표에서 알 수 있듯이, 전체적으로 모든 모델의 Recall_1(실제 연체자를 예측하는 비율) 값이 낮게 나타났으며, 이는 클래스 비율의 불균형과 임계값(threshold) 설정 문제에 관련된 것으로 판단된다. 이러한 상황에서 Accuracy, Precision, Recall, F1-score 등의 지표보다, 임계값 변화에 영향을 받지 않으며 불균형 데이터에서도 신뢰할 수 있는 ROC_AUC를 평가 기준으로 삼는 것이 더 적합하다고 판단하였다. 따라서 ROC_AUC가 가장 높은 LightGBM 모델을 최적의 예측 모델로 최종 선정하였다.

4.3 모델 성능 향상

4.3.1 클래스 가중치 부여

클래스 가중치 부여 후 평가지표 변화

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.76	0.84	4395
1	0.24	0.56	0.34	577
accuracy			0.74	4972
macro avg	0.58	0.66	0.59	4972
weighted avg	0.85	0.74	0.78	4972

클래스 불균형 문제를 해결하기 위해 클래스 가중치(class_weight='balanced') 기법을 적용하여 모델의 성능 변화를 비교하였다. 클래스 가중치 적용 전에는 전체 정확도(Accuracy)가

0.88로 높게 나타났지만 소수 클래스(1)에 대한 Recall이 0.12, F1-score가 0.19에 불과해 소수 클래스에 대한 분류 성능이 매우 낮은 것으로 나타났다. 모델이 주로 다수 클래스(0)로 예측하는 경향이 있어 불균형 데이터의 문제점을 나타냈다. 반면, 클래스 가중치 기법을 적용한 후에는 전체 정확도가 0.74로 소폭 하락했지만 소수 클래스(1)에 대한 Recall이 0.56, F1-score가 0.34로 크게 상승하였다. Precision은 0.24로 다소 낮아졌으나, 실제로 분류하고자 하는 소수 클래스의 탐지율이 크게 개선되었다.

4.3.2. 하이퍼파라미터 튜닝 (GridsearchCV)

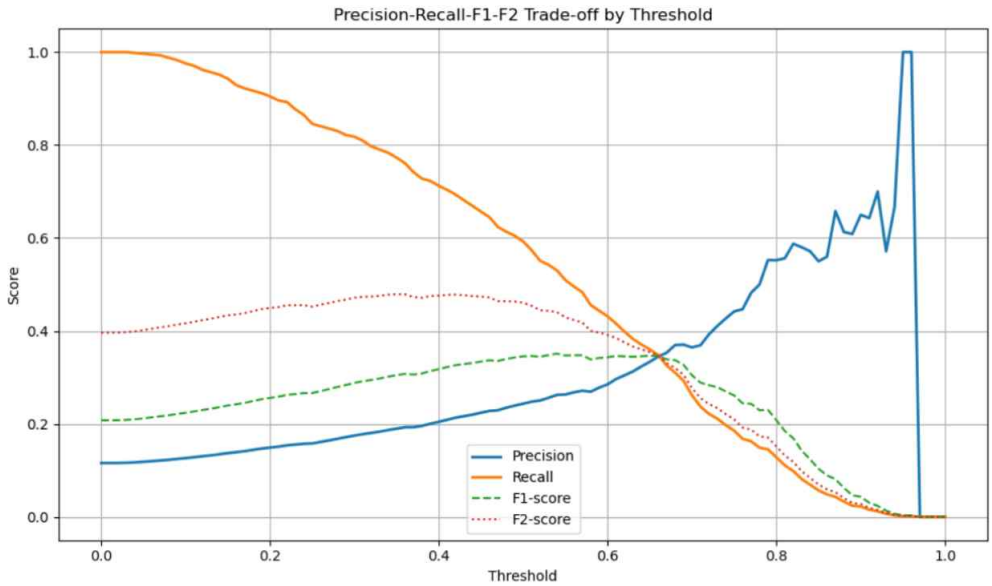
```
Best Parameters: {'learning_rate': 0.05, 'max_depth': 7, 'n_estimators': 200, 'num_leaves': 31, 'subsample': 0.8}
Best F1 Score: 0.3325824159613613
Test Accuracy: 0.7393403057119872
Test Precision: 0.24376336421952957
Test Recall: 0.5927209705372617
Test F1-score: 0.34545454545454546
Test ROC-AUC: 0.7372703738098477

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.76      0.84      4395
           1       0.24      0.59      0.35       577

    accuracy                           0.74      4972
   macro avg       0.59      0.68      0.59      4972
weighted avg       0.85      0.74      0.78      4972
```

LightGBM 모델의 성능을 극대화하기 위해 GridSearchCV를 활용하여 하이퍼파라미터 최적화를 실시하였다. 그 결과, learning_rate 0.05, max_depth 7, n_estimators 200, num_leaves 31, subsample 0.8의 조합이 최적임을 확인할 수 있었다. 이러한 하이퍼파라미터를 적용한 모델은 테스트 데이터에서 소수 클래스(1)에 대한 Recall이 0.59, F1-score가 0.35, 그리고 ROC-AUC가 0.74로 나타나 기존 대비 소수 클래스 탐지력이 더욱 향상된 성능을 보였다. 여기서 사용한 GridSearchCV는 사용자가 지정한 하이퍼파라미터 후보 값들의 모든 조합에 대해 체계적으로 모델을 학습 및 평가하여, 최적의 조합을 찾는 탐색 방법이다. 각 조합에 대해 교차검증을 수행하여 가장 성능이 우수한 하이퍼파라미터를 선택하므로, 모델의 성능을 최대화하는 데 효과적이다. 하지만 가능한 조합의 수가 많아질수록 연산 시간이 길어지는 단점도 존재한다.

.

4.3.3. 최적의 임계값 선정

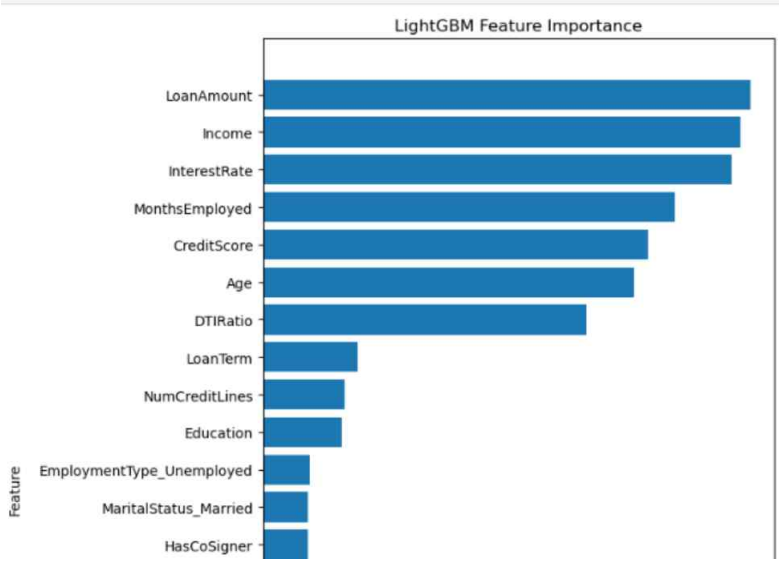
정확도, 재현율, F1, F2의 임계값에 따른 trade-off 그래프



임계값(threshold)에 따른 Precision, Recall, F1-score, F2-score의 변화를 살펴본 결과, 임계값이 낮을수록 모델의 Recall 값이 크게 높아지는 반면 Precision은 낮아지는 전형적인 트레이드오프가 뚜렷하게 나타났다. 임계값 0.2와 0.6사이의 구간에서 precision은 비교적 완만하게 상승하는 반면 Recall 값은 급격하게 하락하는 모습을 확인할 수 있었다. 프로젝트의 주 목적은 연체자(소수 클래스)를 최대한 놓치지 않고 예측하는 데 있으므로 Precision보다 Recall의 중요도가 더 높다고 할 수 있다. 따라서 Precision과 Recall의 조화 평균인 F1-score뿐 아니라 Recall에 더 높은 가중치를 두는 F2-score를 주요 성능 지표로 활용하였다. 그래프를 분석한 결과 F2-score가 가장 높게 나타나는 임계값은 0.35였으며 이때 Precision은 0.1899, Recall은 0.7730으로 연체자를 예측하는 데 있어 실제 연체자의 약 77%를 탐지할 수 있는 수준임을 확인할 수 있었다. Precision(정밀도)이 0.19로 다소 낮지만, 실제 연체자를 놓치지 않는 것이 중요한 프로젝트 목적상 이와 같은 threshold 적용이 합리적이라고 판단하여 F2-score가 최대가 되는 지점인 0.35를 클래스를 분류하는 최적의 임계값으로 설정하였다.

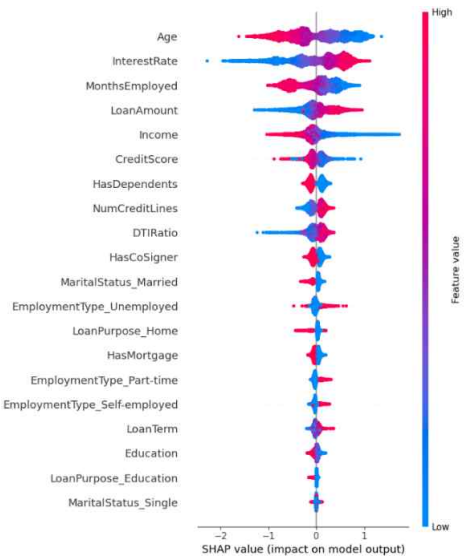
4.4 모델 해석

4.4.1 LightGBM Feature Importance



LightGBM 모델의 feature importance 분석 결과, 대출 연체 예측에 있어 LoanAmount(대출 금액), Income(소득), InterestRate(이자율), MonthsEmployed(재직 월수), CreditScore(신용점수), Age(나이)등이 가장 중요한 변수로 도출되었다. 이는 모델이 연체 가능성을 예측할 때 대출자의 경제적 여력과 상환 능력(예: 소득, 대출금액, 이자율), 신용 상태(신용점수, 재직개월수, 나이)와 직접적으로 관련된 정보를 주로 활용함을 의미한다. 반면, HasCoSigner, MaritalStatus_Married 등은 상대적으로 영향력이 낮은 변수로 평가되었다.

4.4.2 shap values



LightGBM 모델의 예측 결과에 대한 해석력을 높이기 위해 SHAP value(Shapley Additive Explanations)를 활용하였다. 아래의 summary plot은 각 변수별로 SHAP value가 어떻게 분포하는지를 시각적으로 보여준다. x축은 SHAP value(변수가 예측 결과에 미치는 영향의 크기와 방향)를, y축은 변수명을 의미하며, 각 점은 개별 샘플을 나타낸다. 색상은 해당 변수의 값이 높을수록 분홍색, 낮을수록 파란색으로 표시된다. 위의 해석 기법을 통해 InterestRate(이자율), Age(나이), MonthsEmployed(재직개월수), LoanAmount(대출금액), Income(소득), CreditScore(신용점수) 등이 대출 연체 예측에 중요한 역할을 하고 있음을 알 수 있다. InterestRate가 높거나 Age가 낮은 경우, 연체 예측값이 증가하는 경향이 나타난다. 반대로 MonthsEmployed, Income, CreditScore 값이 높을수록 연체 위험이 낮아지는 효과가 확인되었다.

5. 결론

5.1 분석의 문제점 및 한계

프로젝트에서 소수 클래스 가중치 부여와 임계값 조정과 을 통해 연체자(소수 클래스) 탐지율(Recall)을 크게 향상시킬 수 있었으나, 이로 인해 정밀도(Precision)와 F1-score, 그리고 전체 정확도(Accuracy)는 낮은 값을 기록하는 결과를 보였다. 임계값을 0.35로 조정하여 Recall이 77%까지 상승했으나, 이때 Precision은 약 0.19에 불과해, 연체자로 예측한 사례 중 실제로 연체자인 비율이 매우 낮았다. 소수클래스의 한계를 극복하지 못해 실제 연체자가 아닌 대상을 연체자로 잘못 분류하는 오탐률이 높아지면서 예측 결과의 신뢰성이 전반적으로 저하되는 한계가 존재했다.

5.2 향후 발전 방향

향후 연구에서는 변수 간 낮은 상관관계 및 비선형성을 효과적으로 반영하기 위해 다양한 피처 엔지니어링과 파생변수 생성 기법을 적극적으로 도입할 필요가 있다. 다양한 전처리 방법과 함께, 스태킹(Stacking)이나 블렌딩(Blending)과 같은 앙상블 기반의 복합적 모델링 기법을 활용함으로써 전체 모델의 예측 성능을 한층 더 향상시킬 수 있을 것으로 기대된다. 아울러, 대출 연체 예측 모델은 실제 금융 의사결정에 적용되는 만큼, 모델의 해석력 또한 매우 중요하다. 이에 따라 SHAP, LIME 등 설명 가능한 인공지능(XAI) 기법을 활용하여, 모델의 예측 결과에 대한 신뢰성과 투명성을 높이는 방안을 발전 방향으로 제시 할 수있다.

6. 참고 자료

김영재. (2014). 데이터마이닝 기법을 활용한 개인 신용평점모형 개발 [석사학위논문, 연세대학교 공학대학원].

최봉진, 이성우, & 김연국. (2024). 설명 가능한 인공지능을 활용한 은행대출연체 예측 모델 연구. 전산회계연구, 22(2), 43–61.

허준, & 고승곤. (2008). 상호 저축은행 연체 예측을 위한 신용 평점도출에 관한 연구. Journal of the Korean Data Analysis Society, 10(5), 2795–2809.

데이터 증강 기법의 앙상블을 통한 레이블 불균형 해소: 설명 가능한 신용평가 모델을 중심으로. (정지영 1, 이소연 2, 용예린 3, 김민준 41(주)엠로2 숙명여자대학교 경영학과 석사과정 3 서강대학교 경제학과 학부생4 한양대학교 경제금융학과 학부생)

Research on the Prediction Method for PersonalLoan Default Based on Two-Layer StackingEnsemble Learning Model (Zhirui Ma1 and Qinglie Wu)