

# 머신러닝과 XAI 기반 신용카드 연체 예측

김선오\*<sup>1</sup>, 손지우\*<sup>2</sup>, 유지현\*<sup>3</sup>, 정성균\*<sup>4</sup>, 이한준\*\*

## Machine Learning and XAI-based Credit Card Delinquency Prediction

Seono Kim\*<sup>1</sup>, Jiwoo Son\*<sup>2</sup>, Jihyeon Yu\*<sup>3</sup>, Sungkyun Jung\*<sup>4</sup>, and Hanjun Lee\*\*

### 요약

신용카드 연체 예측은 신용 위험 관리의 핵심 과제이다. 기존 모델은 주로 전통적 금융 변수와 로지스틱 회귀 모형으로 접근하였으나, 최근 연구는 비재무적 특성과 머신러닝 결합이 예측 정확도를 향상시킬 수 있음을 보여준다. 본 연구는 Kaggle의 26,457명 신용카드 사용자 데이터를 활용하여 LightGBM, XGBoost, Random Forest 알고리즘을 적용하고, Optuna로 하이퍼파라미터를 최적화하였다. K-폴드 교차검증으로 성능을 검증하고 최적 폴드 수를 분석했으며, SHAP 분석을 통해 각 특성이 예측에 미치는 영향을 해석하였다. LightGBM이 가장 높은 성능을 보였으며 신용카드 발급 경과 월 수, 출생·취업 경과 일 수, 총소득 등 신용연체에 영향을 미치는 핵심 변수들을 확인할 수 있었다.

### Abstract

Credit card delinquency prediction is vital in credit risk management. Traditional models have mainly relied on financial variables and logistic regression, while recent studies show that machine learning can improve predictive accuracy. This study employs a Kaggle dataset of 26,457 credit card users to evaluate three algorithms—LightGBM, XGBoost, and Random Forest—tuning hyperparameters with Optuna and validating performance via k-fold cross-validation. We also assess the optimal number of folds for model stability. SHAP was applied to interpret feature contributions. Results show that LightGBM achieved the best performance, and key variables such as month\_card\_issued, days\_birth, days\_employed, and income\_total were identified as critical factors influencing delinquency.

### Keywords

machine learning, explainable AI, XAI, SHAP, credit card delinquency, Kaggle

\* 명지대학교 경영정보학과 학부과정

- ORCID<sup>1</sup>: <https://orcid.org/0009-0000-2354-4471>

- ORCID<sup>2</sup>: <https://orcid.org/0009-0007-2351-5700>

- ORCID<sup>3</sup>: <https://orcid.org/0009-0007-7467-8171>

- ORCID<sup>4</sup>: <https://orcid.org/0009-0002-4287-0437>

\*\*명지대학교 경영정보학과 부교수(교신저자)

- ORCID: <https://orcid.org/0000-0002-9005-3661>

· Received: Aug. 18, 2025, Revised: Sep. 30, 2025, Accepted: Oct. 03, 2025

· Corresponding Author: HanJun Lee

Dept. of Management Information Systems, Myongji University, Korea

Tel.: +82-2-300-0772, Email: [hjlee1609@gmail.com](mailto:hjlee1609@gmail.com)

## I. 서론

우리가 일상에서 사용하는 신용카드는 단순한 결제 수단을 넘어 현대 경제 활동의 핵심 도구로 자리 잡았다. 신용카드 사용 이력은 개인의 금융 신뢰도를 평가하는 중요한 지표로 활용되며, 연체 발생 시 신용도가 하락해 대출 가능성, 금리, 나아가 주거·고용 등 사회 전반에 영향을 미친다. 전 세계적으로 가계신용이 꾸준히 증가하는 상황에서, 신용평가의 정확성과 신뢰성 확보는 금융기관뿐 아니라 국가 경제 건전성을 위해 필수 과제이다. 가계신용은 가계대출과 신용카드 이용 및 할부거래 등을 포함한 판매신용을 합한 총액을 의미한다[1].

가계신용이 적정 수준에서는 소비와 투자를 촉진하지만, 과도하면 금융기관 건전성을 약화시키고 금융위기로 이어질 수 있다. 1997년 아시아 금융위기와 1930년대 미국 대공황은 부정확한 신용평가와 대규모 채무불이행이 결합해 경제에 심각한 타격을 준 사례다. 정확한 신용평가 체계는 금융 시스템의 안전망 역할을 하며, 신용도 과대평가는 상환 능력 부족 차주에 대한 과도 대출을, 과소평가는 상환 능력 있는 차주의 금융 접근 제한을 초래한다.

현재 널리 사용되는 전통적 신용평가 방식은 공적 기록, 채권 회수 정보, 상환 이력 등 금융 정보에 기반한다[2]. 그러나 이러한 방식은 적시성이 떨어지고, 급변하는 경제 환경이나 개인 소비 패턴 변화를 충분히 반영하지 못하는 한계가 있다. 이에 따라 금융기관과 학계에서는 기존 지표 외 다양한 요인을 반영한 신용평가 가능성을 탐색해 왔다. 최근에는 머신러닝 기반 신용평가 모델이 빠르게 확산되고 있으며, 신용카드 기본 데이터를 활용한 머신러닝·딥러닝 기법 비교 연구[3], 상업은행 데이터를 이용한 연체 상환행태 예측 연구[4] 등 다양한 시도가 이루어지고 있다.

전통적 로지스틱 회귀 모델과 달리, 머신러닝은 방대한 데이터와 다양한 변수를 유연하게 처리하며 높은 예측력을 보인다[5]. 금융뿐 아니라 이직률 예측 등 산업 전반에도 활용된다[6]. 특히 XAI(Explainable AI, 설명 가능한 인공지능) 기법과 결합해 ‘블랙박스’ 문제를 완화하려는 시도가 늘고 있다. SHAP(Shapley Additive Explanations)기반 분석처

럼 변수별 영향도를 정량화하면, 정확한 예측뿐 아니라 결과 해석이 가능해 금융기관의 정책·심사 과정에 직접 반영할 수 있다[7][8].

본 연구는 이러한 흐름 속에서, 인구통계학적 특성과 재무 정보를 포함한 데이터를 활용하여 머신러닝 기반 신용카드 연체 예측 모델을 구축하고자 한다. 또한 XAI 기법을 적용하여 주요 변수의 기여도를 분석함으로써 결과의 투명성과 해석력을 확보한다. 이를 통해 신용등급 평가 기준 개선, 대출 기회 확대, 금융 리스크 최소화에 기여할 수 있을 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 검토하고, 3장에서는 데이터 수집과 모델링을 설명한다. 4장에서는 실험 결과 및 성능 비교 결과를 제시하고 5장에서는 시사점과 향후 연구 방향을 논의한다.

## II. 선행연구

신용위험 평가는 1960년대 로지스틱 회귀와 선형 판별분석 등 통계적 방법에서 출발하였다[1][4][7].

2000년대 중반 이후 서포트 벡터 머신(SVM, Support Vector Machine)과 결정트리 등 다양한 머신러닝 기법이 적용되기 시작하였고[1], 최근에는 딥러닝과 앙상블 알고리즘이 본격 도입되면서 전통적 통계 모델의 한계를 보완·대체하려는 시도가 활발하다[7].

국내 저축은행 가계대출 연체 예측 연구에서는 2016 - 2019년 대출 45,851건을 대상으로 네 가지 전처리 조건을 비교하고, 딥러닝을 포함한 여러 모델의 성능을 대조하였다. CatBoost와 합성곱신경망(CNN, Convolution Neural Network)이 F1-score와 ROC-AUC(Receiver Operating Characteristic-Area Under the Curve)에서 우수한 성과를 보였으며, 특히 고위험 차주 식별에서 상대적 강점을 나타냈다. 핵심 변수로는 대출기간, 연체이율, 연령이 제시되었다[8].

가계금융복지조사 데이터를 기반으로 수행된 선행연구에서는 로지스틱 회귀와 의사결정나무를 비교하고, 클래스 불균형을 완화하기 위해 오버샘플링을 적용하여 민감도를 개선한 바 있다. ROC 분석

에서는 로지스틱 회귀가 연체 예측에 보다 효과적임을 확인했다[4].

또한 한국신용정보원 3년간 74만 명의 신용거래 기록을 활용한 연구에서는 RNN(Recurrent Neural Network)을 통해 성별·연령·계절 등 시간·인구통계학적 요인이 반영된 연체 패턴을 제시하였다. 예컨대 40대 남성이나 여름철과 같은 특정 조합에서 연체 위험이 상대적으로 높게 나타났다[9].

한편 머신러닝 모델은 높은 예측력에도 해석 가능성이 낮다는 한계를 지닌다. 규제 준수와 내부감사, 고객 설명 의무를 고려할 때 XAI 기법의 중요성이 커지고 있으며, XGBoost 기반 연체 예측에 SHAP을 적용해 변수 기여도를 시각화하여 설명하는 연구가 수행되었다[7][8].

본 연구는 이러한 선행연구를 토대로 금융(연체 이력 등)과 비금융(고용 유형, 인구통계학) 변수를 통합한 신용카드 연체 예측 모델을 제안한다. 다양한 머신러닝 알고리즘을 비교하고, SHAP을 통해 변수 영향도를 정량화함으로써 모델의 투명성과 실무 활용성을 높이고자 한다.

본 연구에서는 신용카드 연체 여부를 예측하기 위한 설명 가능한 머신러닝 기반 모델을 제안한다. 연구 절차는 그림 1과 같이 데이터 수집 및 전처리, 모델 구축, 성능 평가 및 해석 단계로 구성된다.

### 3.1 데이터 수집 및 전처리

본 연구에서는 캐글 플랫폼으로부터 26,457개의 행으로 구성된 신용카드 사용자 데이터를 수집하였다. 변수들은 성별, 차량 소유 여부, 부동산 소유 여부, 출생 후 경과 일 수(days\_birth, 분석 과정에서는 원본 단위를 그대로 사용), 교육 수준, 피고용 경과 일 수(days\_employed, 고용 이력이 없으면 0으로 치환), 가족 구성원 수, 자녀 수, 가족 유형 등 인구학적 변수, 그리고 총소득(income\_total, 단위가 명시되지 않은 원본 데이터를 원단위 연소득으로 해석), 소득 유형, 직업 유형, 주거 형태 등 경제 관련 변수, 그 외 개인용 및 업무용 휴대전화 소유 여부, 이메일 소유 여부, 신용카드 발급 경과 월 수, 그리고 신용카드 연체 경과 월 수 변수로 구성되어 있다. 이 중 수치형 변수들의 분포는 표 1에서 보는 바와 같다.

### III. 연구 방법

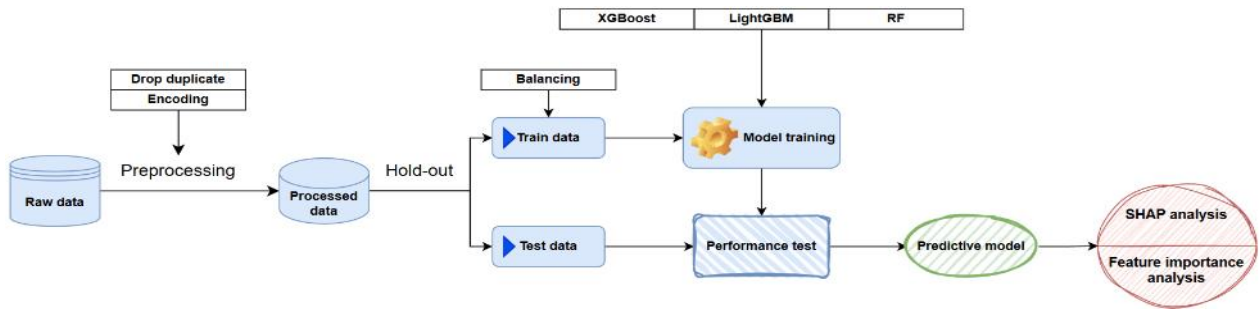


그림 1. 연구 흐름도

Fig. 1. Overall research process

표 1. 수치형 변수의 기술통계량

Table 1. Descriptive statistics of numeric variables

	Income_total	Days_birth	Days_employed	Family_size	Month_card_issued
Count	26457	26457	26457	26457	26457
Mean	187306.52	15958.05	2198.65	2.19	26.12
Std	101878.37	4201.59	2362.04	0.92	16.56
Min	27000.0	7705.0	0.0	1.0	0.0
25%	121500.0	12446.0	407.0	2.0	12.0
50%	157500.0	15547.0	1539.0	2.0	24.0
75%	225000.0	19431.0	3153.0	3.0	39.0
Max	1575000.0	25152.0	15713.0	20.0	60.0

범주형 변수에는 교육 수준, 결혼 여부, 주거 형태, 소득 유형, 직업 유형 등이 있었으며 이들은 원 핫 인코딩을 적용하여 변환하였다.

### 3.2 밸런싱

아울러, 본 연구의 목표는 신용카드 연체를 예측하는 것이므로 delinquency 변수(0: 안정, 1: 주의, 2: 위험)를 활용하여 목표 변수를 설정하였다. 구체적으로 안정(0)은 0으로, 주의(1)와 위험(2)는 1로 통합하여 이진 분류 문제로 재정의하였다. 이는 본 연구의 목적이 연체 여부를 판별하는 데 있으며, 연체 심각도를 세분화하는 것은 연구 범위에 포함되지 않기 때문이다. 한편, 목표 변수의 클래스 분포는 위험(1) 16,264건, 안정(0) 2,225건으로 심각한 불균형을 보였다. 이러한 불균형은 소수 클래스(안정)의 예측 정확도를 저하시킬 뿐 아니라 모델이 다수 클래스(위험)에 편향되도록 만든다는 점이 지적되어 왔다[10]. 실제로 본 연구에서도 불균형을 해소하지 않은 경우 특정 변수에 중요도가 과도하게 집중되는 편향이 확인되었다. 이를 보완하기 위해, 단순 복제의 중복·과적합 위험을 피하면서 다수 클래스 분포를 유지한 채 소수 클래스의 k-최근접 이웃(k-nearest neighbors)을 이용해 합성 표본을 생성하는 SMOTE를 학습 세트에 한해 적용하였다[11]. 학습용 데이터에서 소수 클래스 샘플을 합성하였으며, 최종적으로 두 클래스가 모두 16,264건으로 균형을 이루도록 하였다.

전체 데이터셋은 학습용과 테스트용으로 7:3 비율로 분할하였으며, 클래스 비율을 유지하기 위해 층화추출(Stratified sampling)을 적용하였다.

### 3.3 모델링

본 연구에서는 Random Forest, XGBoost, LightGBM 세 가지 머신러닝 모델을 기반으로 신용카드 연체 여부를 예측하고, 이들 간의 성능을 비교하였다. 세 모델은 모두 분류 문제에서 효과적으로 적용될 수 있으며, 각각의 알고리즘이 갖는 구조적 차이와 학습 방식을 통해 다양한 관점에서 데이터를 분석할 수 있다는 장점이 있다. 세 모델 모두 다수의 트리를 생성하여 예측을 수행하는 앙상블 기반 알고리즘이지만, 학습 방식에서는 차이가 있다. Random Forest는 배깅(Bagging) 기반의 방식으로 각 트리를 병렬적으로 학습하는 반면, XGBoost와 LightGBM은 부스팅(Boosting) 방식으로 이전 트리의 오류를 보완하는 순차적 학습을 수행한다.

모델의 성능을 극대화하기 위해 하이퍼파라미터 최적화 도구인 Optuna를 활용하였으며, 각 모델에 맞는 고유의 매개변수 공간을 정의하여, 반복적인 탐색 과정을 통해 최적의 파라미터 조합을 도출하였다. 이 과정에서 모델의 일반화 성능을 검증하기 위해 K=5의 Stratified K-Fold 교차검증을 적용하여, 데이터의 불균형을 최소화 하면서 각 폴드별 성능을 평가하였다.

표 2. 모델별 하이퍼파라미터 조합

Table 2. Hyperparameter settings for the models

Model	N_estimators	Learning_rate	Max_depth	Subsample	Colsample_bytree	Gamma	Num_leaves
	max_delta_step	reg_alpha	reg_lambda	min_samples_split	min_samples_leaf	min_child_samples	min_child_weight
RF	279	-	15	-	-	-	-
	-	-	-	5	4	-	-
LGBM	539	0.05	10	0.63	0.65	-	123
	-	0.91	0.34	-	-	17	-
XGB	549	0.07	10	0.98	0.65	0.25	-
	2	0.65	0.06	-	-	-	1

각 모델의 튜닝에서는 알고리즘의 특성에 맞춰 매개변수를 설정하였다. 예를들어 Random Forest에서는 n\_estimators, max\_depth, min\_samples\_split 등의 구조적 매개변수를 중심으로 설정하였고, XGBoost와 LightGBM에서는 learning\_rate, colsample\_bytree, reg\_alpha 등 고유의 규제 관련 항목도 함께 고려하였다. 각 모델의 최적 하이퍼파라미터 조합과 상세한 설정값은 표 2에서 확인할 수 있다.

#### IV. 연구 결과

##### 4.1 성능평가

모델의 성능평가는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수를 기준으로 진행하였다. 표 3은 Random Forest, LightGBM, XGBoost 세 모델의 성능평가 결과를 보여준다. 세 모델 간 큰 차이는 없었으나 LightGBM이 재현율과 F1 점수를 기준으로 가장 높은 성능을 보여 최종 모델로 선정하였다.

표 3. 모델별 성능평가  
Table 3. Performance evaluation of the Models

Model	Accuracy	Precision	Recall	F1
RF	0.85	0.90	0.94	0.92
LGBM	0.88	0.89	0.98	0.94
XGB	0.88	0.90	0.96	0.93

##### 4.2 변수중요도 분석

LightGBM 모델에 대한 변수중요도 분석을 진행하였다. 무작위성으로 인한 편차를 줄이기 위해 random seed를 고정하고 10회 반복 실험하여 산출한 평균값을 기반으로 하였다. 분석 결과는 그림 2에서 보는 바와 같다.

변수 중 신용카드 발급 경과 월 수(month\_card\_issued)가 가장 높은 중요도를 보였으며, 가족 구성원 수(family\_size), 출생 후 경과 일 수(days\_birth), 취업 후 경과 일 수(days\_employed), 총 소득(income\_total) 순으로 나타났다. 이는 신용카드 사용 기간(month\_card\_issued)이 연체 가능성에 직접적인 영향을 미친다는 점을 보여준다. 또한 가족 구성원 수(family\_size)와 출생 및 취업 경과 일수(days\_birth, days\_employed)와 같은 개인의 생애와 경제적 맥락이 연체 위험을 예측하는 데 중요한 요인임을 의미한다. 총 소득(income\_total) 역시 주요 변수로 확인되어, 사용자의 금융 능력이 연체 여부 판단에 핵심적 역할을 하고 있음을 시사한다.

이외에도 교육 수준(edu\_type), 결혼 여부(family\_type\_married), 직업 유형(occup\_type) 관련 변수가 중위권 이상의 중요도를 보여, 사회경제적 배경이 신용카드 연체 가능성에 유의미한 영향을 미친다는 사실을 확인할 수 있었다.

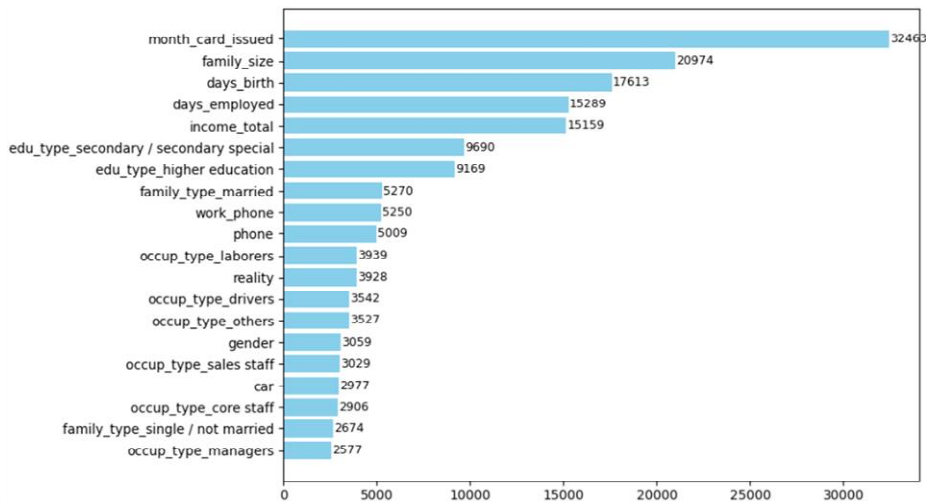


그림 2. 변수중요도

Fig. 2. Feature importance of variables

### 4.3 SHAP 분석

변수 중요도 분석에 이어 본 연구에서는 SHAP 분석을 진행하였다. SHAP 분석 결과는 그림 3에서 보는 바와 같다. 변수들은 목표 변수에 대한 영향이 큰 순서대로 나열되어 있다. SHAP 분석은 각 변수의 값이 예측값에 미치는 영향을 정량적으로 측정하며, 변수별 영향력을 보다 정밀하게 해석할 수 있어 모델의 투명성과 설명 가능성을 높일 수 있다 [4][5]. SHAP 값의 부호는 목표 변수에 대한 정(+) 또는 부(-)의 영향을 나타낸다. 붉은 색은 해당 변수의 값이 클 때, 푸른 색은 값이 작을 때를 의미한다.

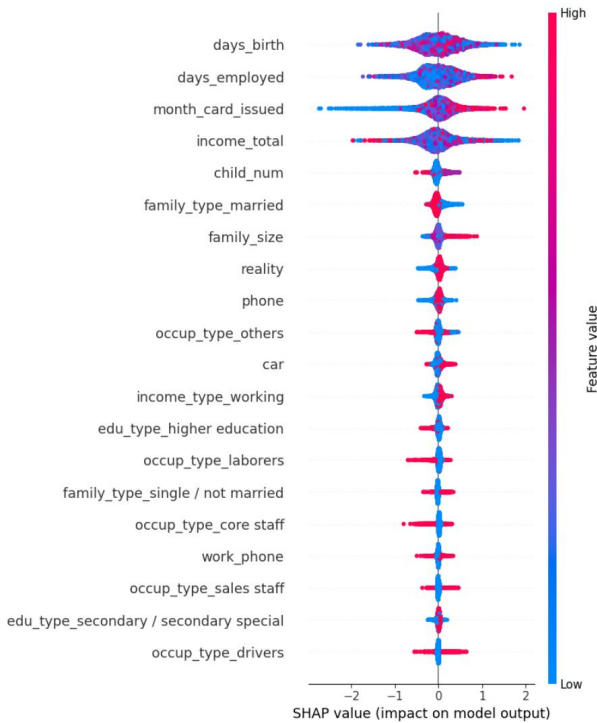


그림 3. SHAP 분석 결과  
Fig. 3. Result of SHAP analysis

분석 결과, 중요도는 출생 후 경과 일 수 (days\_birth)와 취업 후 경과 일 수(days\_employed), 신용카드 사용 기간(month\_card\_issued), 총소득 (income\_total), 순이었으며 해당 네 변수의 영향력이 상대적으로 타변수 대비 큰 것으로 나타났다.

먼저, 출생 후 경과 일 수(days\_birth)는 SHAP 분석에서 가장 높은 중요도를 기록하였으며, 전반적으로 양(+),음(-) 기여가 혼재되어있으며, 특히 일부 젊은 집단 에서는 양(+) 기여가 나타나 연체 위험

증가와 연관될 가능성이 있다. 또한 특정 연령 구간마다 기여 방향이 달라지는 패턴을 확인되어, 단순히 ‘젊다’ 또는 ‘나이가 많다’로 설명될 수 없고, 다른 변수들과의 결합에 따라 방향성이 달라질 가능성이 있음을 시사한다.

두 번째로 높은 중요도를 보인 변수는 취업 후 경과 일 수(days\_employed)이다. 대부분의 근로자가 연체 ‘안정’ 방향으로 예측되도록 기여한 변수였다. SHAP 분석 결과, 근속 기간이 오래된 사람들은 전반적으로 SHAP 값이 양(+) 방향으로 나타나 연체 위험을 높이는 데 영향을 미쳤다. 반대로, 근로 이력이 없거나 불안정한 경우는 다소 음(-) 값을 보여 연체 위험을 낮추는 요인으로 작용했다.

셋째, 신용카드 발급 경과 월 수 (month\_card\_issued)가 낮은 경우에는 연체 가능성이 낮을 수 밖에 없지만, 높을수록 연체 위험이 증대되는 것으로 나타났다. 즉, 연체에 영향을 크게 끼치는 신용카드 사용자들은 장기 고객인 경우가 많은 것으로 나타났다.

한편, 총소득(income\_total)은 값이 클수록 전반적으로 음(-) 기여가 우세해 소득이 높을수록 연체 위험이 낮아지는 경향을 보인다. 그러나, 중·저소득 구간에서는 양(+)과 음(-)의 기여가 혼재하여, 동일 소득이라도 소비 구조, 부채 비율 등 다른 요인에 따라 위험도가 상이할 수 있음을 시사한다.

가족 관련 변수들 가운데 자녀 수(child\_num)는 상대적으로 높은 중요도를 보였으나, SHAP 분석 결과에서는 뚜렷한 방향성을 확인하기 어려웠다. 자녀 수가 많거나 적더라도 예측 기여는 일정하지 않았으며, 이는 변수 자체의 영향보다는 다른 요인과의 상호작용 속에서 중요도가 산출된 것으로 해석된다. 반면, 가족 규모(family\_size)는 구성원이 많을수록 연체 위험이 높아지는 경향이 나타났는데, 이는 자녀 수와 달리 친척이나 부양가족 등까지 포함하는 특성이 신용 위험을 가중시키는 요인으로 작용했음을 시사한다. 나아가, 가족 규모 변수와 더불어 가족의 구조적 특성을 반영하는 결혼 여부 (family\_type\_married) 또한 중요한 요인으로 확인되었다. 분석 결과, 기혼자의 경우 연체 위험이 낮아지는 일관된 안정 효과를 보였으며, 이는 가족 규모가 커질수록 위험이 증가하는 양상과는 대조적이다.

즉, 가족 관련 변수들은 규모와 구조라는 서로 다른 측면에서 신용 안정성에 상반된 영향을 미치고 있음을 보여준다.

종합하면, 본 분석은 시간 경과 변수·구조적 변수·재무 변수라는 유형별 차원에서 신용 위험의 작용이 다르게 나타난다는 점을 보여주며, 단순 중요도 순위만으로는 위험 패턴을 설명하기 어렵다는 한계를 드러낸다. 따라서 신용평가에서는 개별 변수의 크기만을 해석하기보다, 변수 유형별 특성과 상호작용을 종합적으로 고려할 필요가 있다. 특히, 겉보기에 유사해 보이는 변수들조차 서로 상반된 영향을 줄 수 있음을 시사한다.

## V. 결론 및 시사점

본 연구는 신용카드 사용이력 데이터를 활용하여, 머신러닝 기법과 설명 가능한 인공지능(XAI)을 결합한 신용카드 연체 예측 모델을 제안하였다. 다양한 분류 알고리즘 중 LightGBM이 가장 높은 예측 성능을 보였으며, SHAP 분석을 통해 주요 변수의 영향력과 방향성을 시각화함으로써 모델 해석 가능성을 확보하였다. 이를 통해 단순한 예측 성능 향상뿐만 아니라, 연체 발생 메커니즘에 대한 실증적 이해를 도모하였다. 이는 기존 연구가 제시하지 못한 정확성과 해석가능성의 동시 확보라는 이중 목표를 달성한 데 의의가 있다.

실무적으로는 재무정보뿐 아니라 인구통계 변수를 함께 고려함으로써 금융기관이 다양한 고객 특성을 반영할 수 있는 가능성을 제시하였다. 이는 금융 소외 계층을 포함한 ‘대안적 신용평가 모델’의 기초가 될 수 있으며, 개인화된 리스크 관리와 대안 신용평가 체계의 확장에 기여할 수 있을 것으로 기대된다.

다만, 본 연구는 다음과 같은 한계를 가진다. 첫째, Kaggle 데이터셋에 의존함에 따라 변수 구성이 제한적이었다. 특히 소비 행태, 거시경제 지표, 대체 데이터 등 다양한 비재무 요인이 포함되지 못했다. 둘째, 설명 가능한 인공지능 기법 중 SHAP 분석에만 의존하여 모델 해석의 다양성 측면에서 다소 제약이 있었다.

향후 연구에서는 다국적·장기 패널 데이터를 활

용한 일반화 가능한 모델 구축, 소비 행태 및 거시경제 지표 등 외부 요인을 포함한 변수 확장, TabNet이나 CatBoost와 같은 최신 모델의 도입, 그리고 다양한 XAI 기법 간의 비교 분석을 통해 모델의 견고성과 설명력을 강화할 필요가 있다. 이러한 후속 연구는 보다 정교하고 신뢰성 높은 신용 위험 예측 시스템 개발에 기여할 수 있을 것으로 기대된다.

## References

- [1] Y. Yoo, S. Y. Kim, and J. Yeo, "A Study on Debt Delinquency Prediction for Korean Households across Income Level Using a Decision Tree Analysis", *Journal of Consumer Studies*, Vol. 32, No. 1, pp. 1-15, Feb. 2021. <https://doi.org/10.35736/JCS.32.5.1>.
- [2] S. W. Lim, "A Study on Modeling of Predicting Overdue with Household Loan Borrowers of the Savings Bank", Master's thesis, Sungkyunkwan University, Jun. 2020.
- [3] F. Wahab, I. Khan, and S. Sabada, "Credit Card Default Prediction using ML and DL Techniques", *Internet of Things and Cyber-Physical Systems*, Vol. 4, pp. 293-306, Sep. 2024. <https://doi.org/10.1016/j.iotcps.2024.09.001>.
- [4] J. Li, J. Kang, J. Wu, H. Wang, and X. Yang, "Research on credit card default repayment prediction model", *The Journal of Finance and Data Science*, Vol. 10, pp. 100136, Jul. 2024. <https://doi.org/10.1016/j.jfds.2024.100136>.
- [5] B. J. Choi, S. W. Lee, and Y. K. Kim, "Developing Consumer Bank Loan Delinquency Prediction Model Using XAI", *Journal of Computational Accounting*, Vol. 22, No. 1, pp. 43-61, Aug. 2024. <https://doi.org/10.32956/kaoca.2024.22.2.43>.
- [6] X. Wang and J. Zhi, "A Machine Learning-based Analytical Framework for Employee Turnover Prediction", *Journal of Management Analytics*, Vol. 8, No. 3, pp. 351-370, Aug. 2021. <https://doi.org/>

10.1080/23270012.2021.1961318.

- [7] M. J. Kim, S. W. Kim, and J. H. Moon, "Explainable Credit Default Prediction using SHAP", Proc. of the 2024 Winter Conference of The Korea Society of Computer and Information (KSCI), Asan Corporation, Korea, Vol. 32, No. 1, pp. 39-40, Jan. 2024.
- [8] S. Park and D. Choi, "The Analysis of Loan Status and Comparison of Default Prediction Performances based on Personal Credit Information Sample Database", Journal of Korean Institute of Information Scientists and Engineers, Vol. 46, No. 7, pp. 627-635, Jul. 2019. <https://doi.org/10.5626/JOK.2019.46.7.627>.
- [9] J. Lee, "Machine Learning Applications to Households' Insolvency with Imbalanced Data", Journal of Consumer Studies, Vol. 30, No. 6, pp. 97-118, Dec. 2019. <https://doi.org/10.35736/JCS.30.6.5>.
- [10] Y.-J. Han and I.-W. Joe, "Imbalanced Data Improvement Techniques Based on SMOTE and LightGBM", KIPS Transactions on Computer and Communication Systems, Vol. 11, No. 12, pp. 445-452, Dec. 2022. <https://doi.org/10.3745/KTCCS.2022.11.12.445>.
- [11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, Jun. 2002. <https://doi.org/10.1613/jair.953>.

저자소개

김 선 오 (Seono Kim)



2019년 3월 ~ 현재 : 명지대학교  
경영정보학과 학부과정  
관심분야 : 머신러닝, 딥러닝,  
자연어처리

손 지 우 (Jiwoo Son)



2021년 3월 ~ 현재 : 명지대학교  
경영정보학과 학부과정  
관심분야 : 데이터 분석, AI, ERP

유 지 현 (Jihyeon Ryu)



2022년 3월 ~ 현재 : 명지대학교  
경영정보학과 학부과정  
관심분야 : 머신러닝, 자연어 처리,  
데이터 분석, AI

정 성 균 (Sungkyun Jung)



2021년 3월 ~ 현재 : 명지대학교  
경영정보학과 학부과정  
관심분야 : 데이터 분석, AI, ERP

이 한 준 (Hanjun Lee)



2001년 2월 : 서울대학교  
컴퓨터공학과(공학사)  
2004년 2월 : 서울대학교  
컴퓨터공학과(공학석사)  
2016년 8월 : 고려대학교 경영학과  
MIS 전공(경영학박사)  
2020년 3월 ~ 현재 : 명지대학교

경영정보학과 부교수

관심분야 : 머신러닝, 자연어 처리, 정보시스템, 정보화  
정책