Journal of KIIT. Vol. 23, No. 9, pp. 23-31, Sep. 30, 2025. pISSN 1598-8619, eISSN 2093-7571 23
http://dx.doi.org/10.14801/jkiit.2025.23.9.23

청년층 신용위험 예측을 위한 머신러닝 모형과 SHAP 기반 해석

김태섭*[1], 김은지*[2], 김이현*[3], 이한준**

Predicting Credit Risk among Young Adults : A Machine-Learning Approach with SHAP Interpretation

Taeseop Kim*[1], Eunji Kim*[2], Yeehyun Kim*[3], and Hanjun Lee**

요 약

본 연구는 20대 신용카드 사용자의 연체 가능성을 예측하는 머신러닝 모델을 구축하고, SHAP(Shapley Additive Explanations) 기법을 활용하여 주요 영향을 미치는 변수를 분석하였다. 연구에서는 AI-Hub의 금융합성데이터를 활용하여 신용카드 사용자들의 소비 패턴과 신용 정보를 포함한 232,716건의 데이터셋을 구성하였다. Random Forest와 XGBoost 모델을 비교하여 신용 예측의 성능을 평가하였으며, SHAP 분석을 통해 카드 이용 한도 대비 사용 비율, 비필수 소비 지출 비중이 연체 위험을 증가시키는 중요한 요인으로 확인되었다. 본 연구는 금융기관이 20대 소비자의 금융 습관을 고려한 보다 정밀한 신용평가 기준을 마련하고, 맞춤형 금융 전략을 개발하는 데 기여할 수 있는 실증적 근거를 제공한다.

Abstract

This study developed a machine learning model to predict the probability of delinquency among credit card users in their 20s and analyzed key influencing variables using the Shapley Additive Explanations (SHAP) method. The study utilized financial synthetic data from AI-Hub, constructing a dataset of 232,716 cases that included credit card users' spending patterns and credit information. The predictive performance of the models was evaluated by comparing Random Forest and XGBoost, and SHAP analysis identified the credit card utilization rate and the proportion of discretionary spending as key factors increasing the risk of delinquency. This study provides empirical evidence that can help financial institutions establish more precise credit evaluation criteria and develop customized financial strategies that consider the financial habits of consumers in their 20s.

Keywords
credit prediction, machine learning, SHAP, random forest, AI-hub

* 명지대학교 경영정보학과 학부과정
- ORCID[1]: https://orcid.org/0009-0004-0674-151X
- ORCID[2]: https://orcid.org/0009-0004-3689-8171
- ORCID[3]: https://orcid.org/0009-0005-4812-7216
** 명지대학교 경영정보학과 부교수(교신저자)
- ORCID: https://orcid.org/0000-0002-9005-3661

· Received: Jun. 19, 2025, Revised: Jul. 17, 2025, Accepted: Jul. 20, 2025
· Corresponding Author: Corresponding Author: Hanjun Lee
Dept. of Management Information Systems, Myongji University, Korea
Tel.: +82-2-300-0772, Email: hjlee1609@gmail.com

Ⅰ. 서 론

최근 거시경제 불확실성 심화와 고금리·고물가 국면의 장기화로 청년층의 신용 취약성이 급격히 확대되고 있다. 특히 경제적 자립이 완전하지 않은 20대는 학자금 대출·생활비 대출·소비성 대출이 중첩되면서 조기 채무불이행 위험에 노출되는 사례가 빈번하다. 서울회생법원 「2023년 개인회생·파산사건 통계」에 따르면 2023년 한 해 20대의 개인회생 신청 건수는 약 3,200건으로 전년 대비 45% 증가하였다[1]. 이러한 지표는 20대 신용위험이 이미 구조적 문제로 진화하고 있음을 시사한다.

20대 신용불량자 증가는 금융기관의 손실 증가, 신용평가 모델의 불확실성 확대, 청년층 경제활동 위축, 나아가 국가 전체의 소비·투자 기반 약화로 이어질 수 있다는 점에서 정책적·산업적 대응이 시급하다. 그러나 현재 국내 금융기관은 주로 전통적 신용등급 체계를 활용해 대출 심사와 사후 관리를 수행하고 있어, 연령별 특성을 반영한 세분화된 위험 관리는 사실상 어려운 실정이다.

선행연구 역시 기업 연체 예측[2][3] 또는 대출 상품별 리스크 관리[4]에 집중되어 있으며, 개인 차원 연구라고 하더라도 신용정보시스템 내 정형 변수, 예를 들어 기존 대출 이력이나 소득 수준 등에 치우쳐 있다[5]. 하지만 청년층 소비 행동·디지털 금융 이용 패턴·단기 대출 의존도 등은 전통적 변수만으로 설명 · 예측하기 어렵다. 20대를 대상으로 한 연구가 아직 부족한 이유도 이러한 비정형·행동 데이터를 함께 활용할 수 있는 분석 틀의 부재와 해석 가능성 결여에 기인한다.

이에 본 연구는 20대 신용카드 사용자의 소비·금융 행태 데이터를 다차원적으로 통합하고, 머신러닝 알고리즘을 적용해 연체 여부를 분류하며, SHAP(Shapley Additive Explanations) 기법으로 모델 예측에 기여하는 주요 요인을 해석함으로써 다음과 같은 학술·실무적 기여를 제공하고자 한다. 첫째, 20대 특유의 소비 패턴과 금융 행태를 반영한 연령 특화 신용위험 예측 틀을 제시함으로써 기존 신용등급 대비 예측 성능을 개선한다. 둘째, SHAP 분석을 통해 변수별 기여도와 방향성을 시각화·정량화하여 정책 설계 및 금융기관 실무 적용 시 해석 가능성을 확보한다. 셋째, 주요 위험 요인을 파악해 청년층 대상 채무 조기경보 시스템, 소비 습관 개선 프로그램, 연체 예방 금융교육 등을 위한 근거 자료를 제시한다.

본 논문의 구성은 다음과 같다. 제2장에서는 청년층 신용위험 및 머신러닝 기반 신용 예측에 관한 선행연구를 고찰하고, 제3장에서는 데이터 수집·전처리 및 모델링 절차를 설명한다. 제4장에서는 성능 평가와 SHAP 분석 결과를 해석하고, 제5장에서 결론 및 시사점을 논의한다.

Ⅱ. 선행 연구

2.1 신용 예측 연구

신용위험 예측은 그동안 기업 부도 가능성과 대출 상품 건전성을 중심으로 발전해 왔다. 기업을 대상으로 선행 연구에서는 재무제표 같은 정형 데이터와 뉴스 기사 같은 비정형 데이터를 통합해 Random Forest 기반 부도 예측 모형을 구축한 사례가 있으며, 전통적 로지스틱 회귀보다 우수한 성능을 확인한 바 있다[2]. 또한 KOSPI·KOSDAQ 상장 제조업체를 분석해 Random Forest 모델이 기존 신용평가 모델보다 높은 정확도를 보였다는 결과도 보고되었다[3].

대출 상품 분야에서는 학자금 대출자를 대상으로 의사결정나무와 로지스틱 회귀를 활용해 부실 고위험군을 식별한 연구가 대표적이라 할 수 있다[4]. 개인 차원 연구에서는 금융거래 이력이 부족한 사례를 보완하기 위해 통신사의 Call Data Records(가입·단말·통화·빌링·결제·상품 정보) 같은 비금융 데이터를 도입하여 로지스틱 회귀와 Segmentation 기법으로 연체 위험 고객을 분류하였다[5]. 또 다른 선행연구[6]에서는 한국신용정보원 DB 200만 건을 Random Forest 알고리즘으로 분석해 대출·연체 금액을 예측하면서 소비자물가지수 등 거시 변수와의 상관관계를 제시한 바 있다.

이처럼 선행연구에서는 기업·상품 단위 또는 금융 정보 중심 개인 데이터에 편중되어 있으며, 특정

연령대에 특화되거나 소비 행태 변수를 반영한 개인 신용위험 연구는 드물다.

특히 기존 연구들은 전통적이고 정형화된 금융 변수에 과도하게 의존하여 청년층의 미묘한 금융 행동을 포착하기에 불충분했다. 또한, Random Forest와 같은 복잡한 머신러닝 모델의 제한된 해석 가능성은 실질적인 정책 수립이나 금융기관의 실무 적용에 한계를 보였다. 이에 본 연구는 20대 신용카드 사용자의 연체 가능성을 예측하는 데 집중하며, 다차원적 소비 및 금융 행태 데이터를 기반으로 예측 모델을 구축한다. 그리고, 설명가능한 인공지능(eXplainable AI) 기법을 통해 20대 특유의 연체 요인을 식별하고자 한다.

2.2 Random Forest와 SHAP 분석

Random Forest는 다수의 결정 트리를 배깅(Bagging)으로 결합해 과적합을 완화하고 이상치·노이즈에 강인하며, 불균형 데이터 환경에서도 견고한 성능을 제공한다[7]. 앞서 언급한 기업·대출 관련 연구들은 Random Forest가 전통적 로지스틱 회귀 대비 높은 예측력을 보임을 입증하였다[2][3]. 그러나 Random Forest는 블랙박스로 작동해 예측 결과를 해석하기 어렵다는 한계를 갖는다. 이를 보완하기 위해 SHAP 기법이 도입되었으며, SHAP은 게임 이론 기반의 설명 가능한 AI 기법으로, 각 예측에 대해 개별 특성(변수)들이 미치는 영향(기여도)을 공정하게 분배하여 정량화하고 그 영향의 방향성(긍정적/부정적 영향)을 시각화할 수 있다[8]. 이러한 특성은 금융 규제기관의 모델 투명성 요구를 충족시킨다[9].

본 연구는 Random Forest 기반 연체 예측 모델에 SHAP을 결합함으로써 예측 성능과 해석 가능성을 동시에 확보하고, 20대 소비 행태별 주요 위험 요인을 미시적으로 식별하여 금융기관의 맞춤형 연체 관리 전략 수립에 실무적 근거를 제공한다.

III. 연구 방법

본 연구에서는 20대 신용카드 사용자의 연체 가능성을 효과적으로 예측하기 위해 머신러닝 기법을 활용하였다. 연구 수행은 데이터 수집, 전처리 및 변수 추출, 모델 구축, 평가 및 사후분석 순으로 진행하였다.

3.1 데이터 수집

본 연구에서는 AI-Hub(https://www.aihub.or.kr/)에서 제공하는 데이터를 분석한다. AI-Hub는 한국지능정보사회진흥원이 운영하는 AI 통합 플랫폼으로서 AI 서비스 개발을 지원하고자 금융 분야를 포함한 15개 분야의 학습용 데이터를 제공하고 있다. 본 연구에서는 그중 금융합성데이터라는 제목의 데이터셋을 활용한다. 해당 데이터셋은 카드사, 신용평가회사(CB), 은행 등으로부터 수집된 실제 금융 데이터를 바탕으로 합성 알고리즘을 통해 생성된 가상 데이터로, 실제 데이터와 유사한 분포를 지니면서도 개인 정보 유출 위험이 없도록 설계되어 있다. 데이터는 총 12종으로 구성되어 있으며 이 중 10종은 개인 관련 데이터, 1종은 기업, 1종은 상품 관련 데이터이다. 개인 신용카드에 관한 데이터는 회원 정보, 신용 정보, 승인 매출 정보, 잔액 정보 등 8개 범주로 이루어져 있으며 약 300만 명의 개인 기록을 포함한다. 이 8종의 개인 데이터는 동일한 모집단에서 추출되었기에 병합 가능하며, 2018년 7월부터 12월까지 월별로 나누어진 시계열 형식으로 제공된다. 본 연구에서는 이 중 2018년 7월부터 12월까지의 회원 정보, 신용 정보, 승인 매출 정보, 잔액 정보를 결합하여 20대 신용카드 사용자에 대한 연구용 데이터셋을 구성하였다.

데이터의 각 세부 내용은 다음과 같다. 첫째, 회원 정보 데이터는 사용자의 인적 사항(예: 성별, 연령, 거주지, 직장 정보 등)과 기초 신용카드 정보(예: 보유 신용카드 수, 카드 가입 경과 기간 등)로 구성되어 있으며 총 79개의 변수를 포함한다. 둘째, 신용 정보 데이터는 카드 한도 관련 정보(예: 카드 이용 한도 금액, 신용한도 증감 횟수 등)와 리볼빙 관련 정보(예: 리볼빙 이자율, 리볼빙 신청일자 등)로 구성되어 있으며 총 42개의 변수를 포함한다. 셋째, 승인 매출 정보는 사용자의 카드 사용 내역(예:

이용 건수, 이용 금액, 업종별 이용 금액 등)을 포함하며 총 430개의 변수가 있다. 마지막으로, 잔액 정보는 카드 이용 후 미결제된 금액과 관련된 정보(예: 연체 잔액, 연체 발생일자, 연체 횟수 등)로 구성되어 있으며 총 83개의 변수를 담고 있다.

3.2 데이터 전처리 및 변수 추출

본 연구의 예측 목표는 20대 신용카드 사용자를 연체 위험군과 비연체군으로 분류하는 것이다. 원본 데이터셋에는 이미 사용자의 연체 경험 여부가 포함되어 있으나, 일회성 연체자와 다회성 연체자를 동일하게 취급할 경우 모델의 변별력이 저하될 수 있다고 판단하였다. 따라서 기존 선행연구의 기준을 참고하여 우량고객(비연체)과 불량고객(연체)을 구분하는 목표 변수 기준을 새롭게 정의하였다. 우선, 잔액 정보의 연체 발생일자 변수를 활용하여 기준시점(2018년 12월)으로부터 6개월 이내에 최장 연체일수가 30일 이상인 사용자를 불량(연체) 고객으로 정의하였다. 그리고 6개월 이내 최장 연체일수가 5일 미만인 사용자를 우량(비연체) 고객으로 정의하였다. 위 두 기준에 모두 해당하지 않는 중간 경우는 '판단미정' 그룹으로 분류하였다.

상기의 우량/불량 기준을 20대 사용자 데이터에 적용한 결과, 전체 240,101명 중 우량 고객은 236,314명, 불량 고객은 3,593명으로 분류되었으며 나머지 194명은 판단미정으로 나타났다. 판단미정에 해당하는 194명을 제외하여 총 239,907명의 데이터를 유효 표본으로 확보하였다. 이 중에서도 기준시점 당시 카드 가입 기간이 6개월 미만이거나 신용카드를 보유하지 않은 사용자 등 분석에 부적합한 사례를 제외한 결과, 최종적으로 232,716명의 데이터를 분석 대상으로 선정하였다.

이후 주요 변수를 해석하기 쉽도록 일부 수치형 변수들을 비율 변수로 변환하였다. 예를 들어, 승인매출 정보의 가맹점별 이용 금액은 쇼핑, 요식, 교통, 의료 등 여러 업종 분류로 세분화되어 있으므로, 각 사용자의 업종별 이용금액을 전체 이용금액 대비 비율로 계산하여 새로운 변수를 생성하였다. 이와 마찬가지로 카드 결제 유형별 (일시불, 유이자할부, 무이자할부, 현금서비스 등) 이용 금액 및 이용 횟수도 전체 대비 비율로 환산하여 변수화하였다. 이러한 처리 방식은 변수가 갖는 의미를 유지하면서도, 각 사용자의 소비 특성을 보다 직관적으로 파악할 수 있도록 돕는다.

한편, 연체 비율의 불균형 문제를 해결하기 위해 언더샘플링 기법을 적용하였다. 원 데이터에서 비연체(우량) 사례의 수가 연체(불량) 사례에 비해 월등히 많았으므로, 비연체 그룹의 일부 데이터를 무작위로 제거하여 두 그룹의 규모를 균형 있게 맞추었다. 그 결과 최종 전처리된 데이터셋은 총 7,132건의 사례와 43개의 변수로 구성되었다. 해당 균형 데이터셋을 이후 머신러닝 모델의 학습과 평가에 활용하였다.

3.3 모델 구축

모델 구축 단계에서는 먼저 훈련 데이터와 테스트 데이터를 7:3의 비율로 분할하였다. 이후 청년층 신용위험 예측을 위한 최적의 모델을 선정하고자 높은 예측 성능으로 널리 알려진 Random Forest 모델과 XGBoost 모델을 후보군으로 설정하고, 두 모델의 성능을 비교 분석하였다. 두 모델은 각각 배깅과 부스팅의 서로 다른 앙상블 방식을 가지며 모델 간의 비교를 통해 연구 방법론의 견고성을 확보하고, 주어진 데이터셋과 예측 문제에 가장 적합한 알고리즘을 도출하는 것이 중요하다고 판단하였다.

구체적으로 본 연구에서는 RandomizedSearchCV와 GridSearchCV 기법을 조합하여 교차검증을 3회 수행하고, 분류 정확도를 기준으로 가장 높은 성능을 보이는 하이퍼파라미터 값들을 선정하였다. 두 모델 모두에 대해 하이퍼파라미터 튜닝 후 성능이 향상되었으며, 튜닝 결과에 따른 모델별 성능 지표를 표 1에 요약하였다. 튜닝 완료 후, 최적 파라미터를 적용한 RF 모델과 XGBoost 모델을 최종 후보 모델로 확정하고, 다음 단계에서 두 모델의 예측 성능을 비교·평가하였다. 참고로 본 연구에서는 모델 구축에 Python 언어와 scikit-learn, XGBoost 등 관련 라이브러리를 활용하였다.

표 1. 모델별 하이퍼파라미터 튜닝
Table 1. Model-specific hyperparameter tuning

Random forest			
Parameter	Value	Parameter	Value
max_depth	None	n_estimators	100
min_samples_split	2	min_samples_leaf	1
Accuracy	0.828504		
XGBoost			
Parameter	Value	Parameter	Value
sub_samples	1.0	reg_lambda	2
reg_alpha	1	n_estimators	190
max_depth	9	learning_rate	0.10666
gamma	0.2	colsample_bytree	0.6
Accuracy	0.825701		

IV. 연구 결과

4.1 모델 성능 비교

앞서 구축한 Random Forest 모델과 XGBoost 모델의 예측 성능을 비교하였다. 성능 평가는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 및 F1 값의 네 가지 지표를 사용하였다. 표 1의 튜닝 후 성능 비교 결과에서 알 수 있듯이, Random Forest 모델의 성능이 XGBoost 모델보다 소폭 우수하게 나타났다. 예를 들어, 정확도의 경우 RF 모델이 약 0.834로 XGBoost 모델의 0.826보다 다소 높게 나타났으며, 다른 지표들에서도 RF 모델이 근소하게 앞서는 결과를 보였다. 이에 따라 RF 모델을 최종 모델로 선정하였다.

다음 단계로서 해당 모델의 변수 중요도를 분석하여 예측에 기여하는 주요 변수를 확인하고자 RFECV(Recursive Feature Elimination with Cross-Validation) 기법을 적용하여 중요 변수를 선별하였다. RFECV를 통해 초기 입력 변수 43개 중 중요도가 낮은 변수를 순차적으로 제거한 결과, 최적의 성능을 내는 변수 조합은 38개 변수로 파악되었다. 선택된 38개 변수만을 사용하여 모델을 재훈련한 결과 모델 성능이 다소 향상되었다. RFECV 적용 전후의 정확도가 0.826에서 0.834로 상승하는 등 전반적인 지표가 개선되었는데, 이러한 성능 향상 효과는 표 2에 요약되어 있다.

표 2. RFECV후 Random Forest 모델의 성능개선
Table 2. Performance improvement of the random forest model after RFECV

	Before	After
Number of features	43	38
Accuracy	0.826	0.834
Precision	0.827	0.837
Recall	0.826	0.834
F1 score	0.825	0.833

4.2 변수중요도 분석

최종 선정된 38개 변수에 대해 변수 중요도를 분석한 결과, 그 영향력이 큰 상위 10개 변수를 선별할 수 있었다. 그림 1은 이 상위 10개 변수의 중요도 값을 시각화한 것이다.

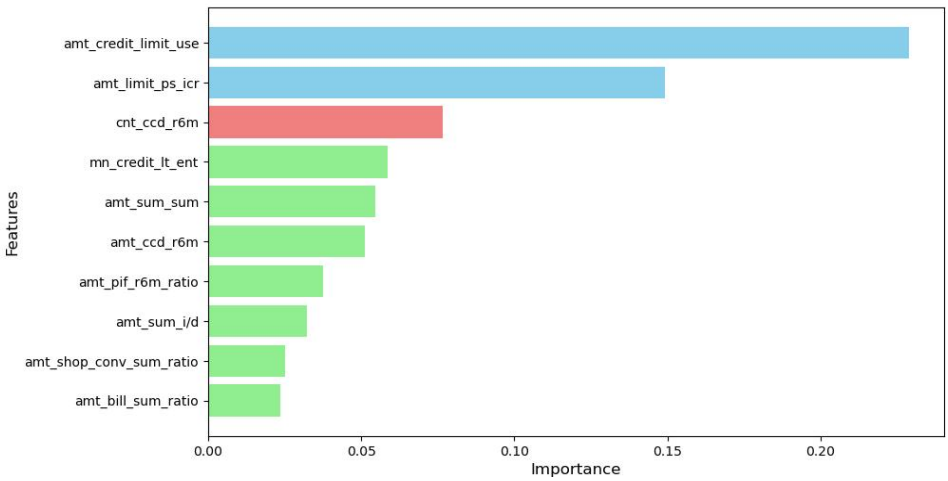

그림 1. 변수중요도 분석 결과
Fig. 1. Result of feature importance analysis

그림 1에서 볼 수 있듯이, "최근 6개월간 카드 이용 한도 금액" 변수가 다른 변수들에 비해 예측에 미치는 영향력이 가장 큰 것으로 나타났다. 이외에도 한도 증액 가능 금액, 최근 6개월 총 이용 금액, 이용 건수 등 여러 변수들이 중요한 요인으로 식별되었다. 상위 10개 변수의 명칭과 간략한 설명은 표 3에 제시되어 있다. 이러한 결과를 통해, 20대 신용카드 사용자의 신용카드 한도 관리 및 사용량이 연체 여부 예측에 핵심적인 역할을 함을 알 수 있다.

표 3. 상위 10개 중요 변수 목록
Table 3. List of top 10 important features

Variable	Type	Description
amt_credit_limit_use	C	Card credit limit usage amount
amt_limit_ps_icr	C	Available credit limit increase amount
cnt_ccd_r6m	M	Number of transactions (lump-sum, installment, cash advance, card loan) in the last 6 months
mn_credit_lt_ent	A	Number of months since credit card issuance
amt_sum_sum	A	Total amount spent over 6 months
amt_ccd_r6m	A	Total spending amount (lump-sum, installment, cash advance, card loan) in the last 6 months
amt_pif_r6m_ratio	A	Lump-sum spending amount in the last 6 months
amt_sum_i/d	A	Change in spending amount compared to the previous month
amt_shop_conv_ratio	A	Shopping and convenience store spending in the last 6 months
amt_bill_ratio	A	Payment amount for the last 6 months

M: Member, C: Credit, A: Approval sales

4.3 SHAP 기반 해석

변수 중요도 분석만으로는 각 변수가 연체 가능성에 어떻게 영향을 미치는지 구체적으로 파악하기 제한된다. 이에 최종 모델에 대해 SHAP(Shapley Additive Explanations) 분석을 수행하여 개별 변수들의 영향 방향과 규모를 심층적으로 해석하였다. 그림 2는 RF 최종 모델에 대한 SHAP 분석 결과를 보여준다.

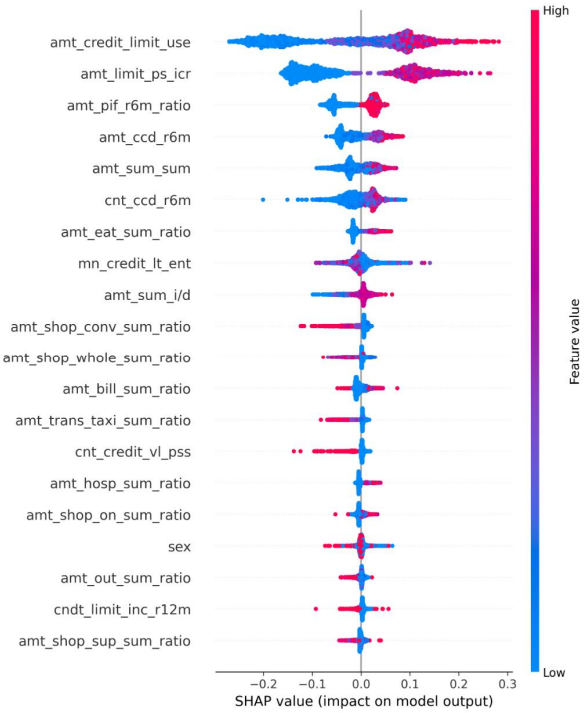

그림 2. SHAP 분석 결과
Fig. 2. Result of SHAP analysis

신용 관련 변수 중에서는 카드 이용 한도금액(amt_credit_limit_use)이 낮을수록 해당 사용자의 연체 가능성은 높아지는 경향을 보였다. 또한 상향 가능한 카드 한도액(amt_limit_ps_icr)이 클수록 연체 위험이 증가하는 것으로 나타났다. 이러한 결과는 젊은층 카드 사용자의 가용 신용한도 수준이 연체 위험과 밀접한 관련이 있음을 보여준다. 다시 말해, 카드 이용 한도가 낮거나 추가 한도 여력이 큰 소비자일수록 신용위험이 높아질 가능성이 있다는 의미이다.

소비 행태 변수에 관한 분석 결과도 주목할 만하다. 최근 6개월간의 총 신용카드 사용 금액(amt_ccd_r6m)이 많을수록 연체 가능성이 증가하는 것으로 나타났으며, 일시불 사용 금액 비율(amt_pif_r6m_ratio)이 높을수록, 그리고 최근 6개월간의 총 승인 금액(amt_sum_sum)이 클수록 신용 연체 위험이 커지는 경향을 보였다.

아울러 동일 기간 동안의 신용카드 이용 건수(cnt_ccd_r6m)가 많을수록 연체 위험이 증가하는 것으로 나타났다. 특히 전월 대비 카드 사용 금액 증감률(amt_sum_i/d) 역시 연체 가능성과 상관성이 있었는데, 이 값이 크다는 것은 이전 달보다 소비 금액이 크게 증가했음을 의미하며, 그러한 급격한 소비 패턴의 변화가 연체 위험을 높이는 요인임을 시사한다.

소비 패턴 구성비 변수들에 대한 SHAP 값도 흥미로운 시사점을 제공한다. 요식업종 소비 비율(amt_eat_sum_ratio)이 높을수록 연체 가능성이 증가하는 것으로 나타났고, 쇼핑/편의점 소비 비율(amt_shop_conv_sum_ratio)과 쇼핑/도소매 소비 비율(amt_shop_whole_sum_ratio)이 높을수록 연체 위험이 증가하는 것으로 분석되었다. 반면 택시 교통비 소비 비율(amt_trans_taxi_sum_ratio)이 낮은 경우일수록 오히려 연체 가능성이 높아지는 특징을 보였다. 즉, 교통비 지출을 최소화하는 소비자일수록 재정적으로 어려움을 겪고 있을 가능성이 상대적으로 높으며, 이는 곧 해당 집단의 신용카드 연체 위험이 증가할 수 있음을 의미한다.

Ⅴ. 결론 및 시사점

본 연구는 20대 청년층 신용카드 사용자들의 연체 가능성을 예측하는 머신러닝 모델을 개발하고, SHAP 기법을 활용하여 예측에 영향을 미치는 핵심 요인들을 분석하였다. 주요 연구 결과, 신용카드 연체 여부는 단순히 개인의 신용등급에만 좌우되지 않으며 소비 패턴, 신용카드 한도 관리 수준, 대출 및 결제 방식 등 다양한 요인들의 영향을 받는 것으로 나타났다. 예를 들어, 카드 이용 한도 사용량이 낮거나 추가 한도 여력이 큰 경우, 그리고 최근 몇 달간 카드 사용액 및 이용 빈도가 많은 경우 연체 위험이 증가하는 경향을 보였다. 특히 '상향 가능한 카드 한도액(amt_limit_ps_icr)이 클수록 연체 위험이 증가한다'는 결과는 전통적인 신용평가 관점(일반적으로 높은 한도 증액 여력은 높은 신용도를 의미함)과는 다소 상이할 수 있으나, 이는 본 연구가 밝혀내고자 했던 20대 특유의 신용 행태를 반영하는 중요한 발견이다. 즉, 경제적 자립이 완전하지 않은 20대의 경우 잠재적으로 높은 신용 한도를 부여받을 수 있는 능력(예: 초기 신용 이력 또는 금융기관의 공격적인 한도 부여)이 있더라도, 오히려 이를 활용하여 과도한 소비를 하거나, 재정적 어려움을 해결하기 위해 한도를 적극적으로 늘리려는 시도(이른바 '돌려막기'와 같은 행태)와 연관될 가능성을 시사한다. 또한 일상 소비에서 요식업 및 쇼핑 관련 지출 비중이 높은 경우 연체 가능성이 높아지는 반면, 택시 교통비 등 필수적인 교통비 지출 비중이 지나치게 낮은 경우에도 연체 위험이 상대적으로 높게 나타났다. 이러한 결과는 20대 소비자의 과도한 지출 습관이나 불균형한 소비 패턴이 신용위험과 유의미하게 연관되어 있음을 실증적으로 시사한다.

본 연구의 가장 큰 학술적·실무적 의의는 다음과 같다. 첫째, 20대 특유의 소비 패턴과 금융 행태를 반영한 연령 특화 신용위험 예측 틀을 제시함으로써 기존 신용 등급 중심의 평가 방식 대비 예측 성능을 개선하고, 20대 신용위험 예측 연구의 학술적 공백을 메웠다는 점이다. 둘째, Random Forest와 같은 '블랙박스' 모델의 한계를 극복하기 위해 SHAP 분석을 도입하여, 각 변수가 연체 가능성에 미치는 기여도와 방향성을 정량화하고 시각화함으로써 금융기관의 정책 설계 및 실무 적용 시 해석 가능성과 모델 투명성을 확보했다. 이는 단순히 예측 모델을 제시하는 것을 넘어, 도출된 결과를 실제 금융 전략 수립에 활용할 수 있는 구체적인 근거를 제공한다. 셋째, 본 연구에서 밝혀진 주요 위험 요인들(카드 한도 대비 높은 사용 비율, 비필수 소비 지출 증가, 특정 업종 소비 편중 등)은 금융기관이 20대 연령층에 적합한 정교한 신용평가 기준을 마련하고, 연체 가능성이 높은 소비 패턴을 조기에 탐지하여 보다 효과적인 연체 관리 전략을 수립하는 데 활용될 수 있는 실증적 근거를 제공한다. 예를 들어, 금융기관은 본 연구의 결과를 바탕으로 높은 위험군 20대 고객을 선별하여 사전에 금융 교육을 실시하거나 맞춤형 부채 관리 프로그램을 제공함으로써 연체 발생률을 낮추는 데 기여할 수 있다. 이는 청년층의 건전한 신용 구축뿐만 아니라 금융기관의 신용위험 관리 효율성 제고에도 이바지할 수 있을 것으로 기대된다.

물론 본 연구에는 몇 가지 한계점이 존재한다. 연구에 활용된 데이터셋이 2018년 하반기(7월~12월)라는 특정 시점의 소비 행태만을 반영하고 있어, 장기적인 추세 변화나 시계열적 요인을 충분히 고려하지 못했다. 향후 연구에서는 보다 장기간의 데이터를 수집하여 종단적(Longitudinal) 연구를 수행함으로써 연체 발생 과정과 그 영향 요인을 더욱 면밀히 분석할 필요가 있다. 또한 본 연구는 분석 대상을 20대로 한정하였으므로, 다른 연령층에 본 모델을 적용할 경우에도 유사한 예측력이 확보되는지 추가 검증이 필요하다. 향후에는 연령대를 확대하거나 세분화하여 모델을 적용함으로써 범용적인 신용위험 예측 모델로 발전시킬 수 있을 것이다. 결론적으로, 본 연구는 20대 신용카드 이용자의 연체 가능성을 예측하는 새로운 접근법을 제시하고, 연체 위험을 높이는 주요 요인들을 규명함으로써 금융권의 신용 위험 관리에 유용한 실증적 근거를 제공하였다. 이러한 연구 결과는 금융기관이 청년층 맞춤형 신용평가 및 연체 예방 전략을 수립하는 데 활용될 수 있으며, 더 나아가 향후 연구를 통해 모델의 범위를 확장하고 한계를 보완함으로써 금융권의 신용 관리 시스템을 한층 고도화하는 데 기여할 수 있을 것으로 기대된다.

References

[1] Seoul Rehabilitation Court, "2023 Report on the Statistics of Individual Rehabilitation and Bankruptcy Cases", Seoul Rehabilitation Court, 2023.

[2] Y. H. Kim, "A study on machine learning-based corporate credit rating model using unstructured data", Doctoral dissertation, Soongsil University, Dec. 2023.

[3] S. J. Kim, "A corporate credit rating model with Random Forest", Master's thesis, Kookmin University, Dec. 2015.

[4] J. S. Choi, J. T. Han, M. J. Kim, and J. A. Jeong, "Developing the high risk group predictive model for student direct loan default using data mining", Journal of the Korean Data & Information Science Society, Vol. 26. No. 6, pp. 1417-1426, Nov. 2015. https://doi.org/10.7465/jkdi.2015.26.6.1417.

[5] J. Y. Kim, "Development of personal credit evaluation model (telecom score) using telecommunication big data", Doctoral dissertation, Soongsil University, Jun. 2019.

[6] B. W. Seo, "Loan amount and delinquency amount prediction method of household loan using Random Forest", Master's thesis, Soongsil University, Jun. 2022.

[7] I. H. Sarker, "Machine Learning: Algorithms, Real-world Applications and Research Directions", SN Computer Science, Vol. 2, No. 3, pp. 1-21, Mar. 2021. https://doi.org/10.1007/s42979-021-00592-x.

[8] S. S. Fatima, M. Wooldridge and N. R. Jennings, "A Linear Approximation Method for the Shapley Value", Artificial Intelligence, Vol. 172, No. 14, pp. 1673-1699, Sep. 2008. https://doi.org/10.1016/j.artint.2008.05.003.

[9] M. Lee, J. Lee, and H. Lee, "Cognitive dysfunction prediction model with lifelog dataset based on Random Forest and SHAP", Journal of Korean Institute of Information Technology, Vol. 22, No. 1, pp. 1-8, Jan. 2024. https://doi.org/10.14801/jkiit.2024.22.1.1.

저자소개

김 태 섭 (Taeseop Kim)

2022년 3월 ~ 현재 : 명지대학교 경영정보학과 학부과정
관심분야 : 데이터 분석, AI, ERP

김 은 지 (Eunji Kim)

2022년 3월 ~ 현재 : 명지대학교 경영정보학과 학부과정
관심분야 : 데이터 분석, AI, ERP

김 이 현 (Yeehyun Kim)

2022년 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