Journal of Korean Home Management Association
Vol.38 No.3 Serial No.159 2020.9. 1~16
ISSN(print) 1229-1331
https://doi.org/10.7466/JKHMA.2020.38.3.1

데이터마이닝 기법을 이용한 중년층의 가계부채 연체 가능성 분류 연구

Decision Tree Analysis for Payment Delinquency Among Middle-Aged Borrowers

이종희(Jonghee Lee)* https://orcid.org/0000-0002-0235-2689

Department of Consumer Science, Inha University, Associate Professor

<Abstract>

This study aims to detect households at high risk of insolvency among middle-aged debtors using The Survey of Household Finances and Living Conditions in 2019. This study assesses the performance of classifier such as the Decision Tree model in machine learning. It is well known that when the proportion of one class in a dataset is dominant, the prediction performance of classifiers becomes problematic. In order to address the degree of imbalance of two classes in data sets, the ROSE (random oversampling examples) technique was considered. It was found that the ROSE improved the sensitivity and AUC, helping to improve classification prediction accuracy while avoiding overfitting problems. In addition, 1) those with debt-to-asset ratio greater than 0.8, 2) those with debt-to-asset ratio greater than 0.4 and less than 0.8 and got a loan through a savings bank, 3) those with debt-to-asset ratio less than 0.4, non-regular workers, those who did not own a house and those who got a loan for a business were more likely to be delinquent on their debt payment. This study found that the level of Debt to Asset ratio, Debt to Financial Asset ratio, the kind of financial institution, and reason for borrowing money were significant factors of the payment delinquency.

▲주제어(Keywords) : 의사결정나무분석(decision tree analysis), 중년층 가계(middle-aged household), 가계부채(household debt), 부채상환연체(debt payment delinquency)

I. 서 론

글로벌 금융위기 이후 세계 각국에서 금융위기의 원인 규명 및 대책 수립을 위한 다각적인 노력을 경주하고 있다. 한국의 경우 다음과 같은 점에서 가계부채 건전성에 대한 우려가 제기된다. 첫째, 가계부채의 규모가 지속적으로 증가하였다. 미국을 포함한 주요 국가들이 서브프라임 사태이후 디레버리징을 경험한 반면 한국의 가계 부채는 지속적으로 증가하고 있다(한국은행, 2019). 2009년 776조원이었던 가계신용은 2019년 현재 1,540조원으로 증가하였다(나라지표, 2019). 이는 한국은행이 부채 관련 통계를 집계하기 시작한 2002년 이후 가장 높은 수준이다. 2018년도 가계의 처분가능소득 대비 가계부채 비율은 162.7%이며, 이 비율은 2014년 이후 지속적으로 늘고 있는 추세다. 둘째, 가계부채의 연령별 구성에서 은퇴시점 이후의 소득이 급격히 감소하는 중년기의 가구부채비중이 매우 높으며, 이들 집단의 부채의 규모도 점차 증가하고 있다. 가구주 연령별로 부채 규모를 살펴보면 40대가 9,896만원, 50대가 8,602만원 순으로 컸으며, 가계부채 증가폭은 40대가 전년대비 14.6%로 가장 빠른 증가율을 보

• 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A5A8027769).
* Corresponding Author: Jonghee Lee, Department of Consumer Science, Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, Rep. of Korea. Tel: +82-32-860-8112, E-mail: jongheelee@inha.ac.kr

였다(대한민국정책브리핑, 2018). 이러한 상황 속에서 한국의 가계부채에 대한 지속적인 관심이 요구되며, 중년기의 가계부채 부실화 가능성에 대한 면밀한 분석이 필요하다.

가계 부채를 이해하기 위해서는 소비와 저축에 대한 전통적 이론인 생애주기가설 (Life Cycle Hypothesis) (Ando & Modigliani, 1963)을 이용할 수 있다. 이 이론에 의하면 소비 주체인 가계는 일반적으로 전생애에 걸쳐 벌어들인 소득에 맞추어 현재의 소비를 결정하며, 효용을 극대화한다. 일반적으로 가계는 대부분 일생 동안 소득과 소비의 불균형의 상태를 경험하게 되며, 저축 또는 차용이라는 방식을 통해 그 격차를 조절하며, 가계의 후생을 극대화시켜 나간다. 이러한 점에서 가계의 대출은 가계의 유동성 제약을 완화시켜주어 가계의 경제적 안정에 기여할 수 있다. 그러나 가계의 입장에서 부채상환을 위하여 소득의 일정부분을 지출하여야 하므로, 가계의 실질적인 소득 감소와 같은 효과가 발생할 수 있다. 또한 지나친 차입은 부채상환 부담을 가중시킬 뿐만 아니라, 지불연체, 가계파산과 같이 가계 경제의 문제를 야기시킬 수 있다(신용회복위원회, 2019). 거시적으로는 가계의 가계부채 부실위험의 증가가 금융기관의 부실로 이어질 수 있으며, 금융시스템 전반의 건전성에 영향을 미치는 잠재적 위험요인이 될 수 있다. 금리변동 위험이 확대될 경우 소비와 경제성장을 저해할 수 있다(강종구, 2017). 이러한 점에서 한국가계의 가계부채 부실화를 이해하기 위해서는 전통적인 이론을 보완하는 다양한 접근이 필요하다.

최근 한국은 가계부채와 관련한 통계정보의 생산과 생산된 정보의 분석이 증가해왔다(한국조사연구학회, 2016). 가계부채와 관련된 기존의 정보는 주로 금융기관이 중심이 되는 정보로 금융기관의 안정성을 파악하고, 금융기관을 대상으로 한 정책을 제안할 경우에는 이러한 자료의 사용이 적절하다. 반면 가계부채를 과소추정하게 될 가능성이 있다(한국조사연구학회, 2016). 가계의 안전성을 파악하고, 복지정책, 주거정책 등과 같은 다양한 분야와 다양한 계층의 가계부채 관련 정보 파악하고자 할 경우 미시적인 관점에서 가계의 부채행동에 대한 분석을 수행하는 것이 필요하다. 2000년대 이루어진 가계부채와 관련된 다수의 실증연구들은(e.g. 김영일, 변동준, 2012; 김영일, 유주희, 2013; 김영일, 2019; 김현정, 2010; 박인수, 박창수, 2018; 박윤태, 노정현, 2017; 박연우, 허석균, 2018; 허석균, 박연우, 변동준, 심혜인, 2016) 차입자의 재무적 변수(e.g. 소득. 자산, 주택보유) 뿐 아니라 차입자의 인구사회학적 변수(e.g. 성별, 연령, 교육수준)가 차입자의 의사결정에 영향을 미친다는 점을 밝혔다. 인구사회학적 변수 중 차입자의 연령은 가계부채와 관련된 주요한 설명변수인 것으로 나타났다. 대부분의 선행연구들은 차입자의 연령과 같은 인구사회학적 변수들을 통제변수로서 연구모형에 포함하였다. 가령 김지섭(2014)은 한국의 가계부채는 가구주 연령이 40-50대인 가구에서 보유하고 있는 비중이 다른 연령대와 비교하여 높은 편이라는 것을 밝힌 바 있다.

다른 연령층보다 중년기 가계의 부채비중이 높음에도 불구하고, 중년층 차입자들을 대상으로 한 가계부채 관련 연구는 부족한 상황이라고 할 수 있다. 한국 중년기의 가계부채 상환 및 부실화 정도는 몇몇 보고서(e.g. 김지섭, 2014; 성현구, 박범기, 2018)에서 부분적으로 다루어졌다. 이들 보고서 간 중년기 차입자들의 가계부채 부실화 정도에 대한 평가는 일치하지 않는 편이었다. 중년기의 가계부채를 긍정적으로 바라보는 측면에서는 중년기이 다른 연령대와 비교하여 상대적으로 소득 및 자산 수준이 높기 때문에(성현구, 박범기, 2018) 다른 연령대와 비교하면 이들의 가계부채의 상환 여력이 양호한 수준이라고 진단하였다. 반면 중장년층이 다른 연령대에 비하여 상대적으로 많은 자산을 보유하고 있지만, 중장년층에 가계부채가 집중되어 있는 상황에서 이들이 소득이 급감하는 은퇴 이후에도 부채를 상환하지 않고 상당 부분 그대로 부채를 보유하게 될 가능성이 크다(김지섭, 2014)고 진단하기도 하였다. 또한 동세대 효과(cohort effect)를 감안해 본다면, 이들이 은퇴하게 되는 10-20년 후 우리나라의 가계부채 문제가 현재보다 심각해질 가능성을 배제하기 어렵다(김지섭, 2014). 따라서 이들 집단을 중심으로 한 가계부채의 증가 및 부실위험과 관련된 연구가 필요하다고 하겠다.

가계부채 관련 선행연구들의 대부분은 통계적 가정을 기본으로 하는 모수적 (parametric) 방법을 이용하여 가계부채 규모 및 가계부채 비율의 결정요인을 분석하는 경향이 있었다. 그러나 가계대출을 예측할 수 있는 잠재적인 설명변수의 수가 많으므로, (이창훈, 강규호, 목정환, 2018) 특정 함수형태를 고려한 추정방법은 한계를 지닐 수 있다. 이런 점에서 전통적인 선형모형보다 유연한 추정모형을 구현할 필요성이 제기된다. 또한 최근 한국의 가계부채에 대한 관심의 증대로 가계부채와 관련한 통계정보를 생산하고, 생산된 정보를 분석하거나 신규통계의 개발에 관심이 증가하고 있다 (한국조사연구학회, 2016). 가계부채와 관련된 기존의 정보는 주로 금융기관이 중심이 되는 정보로 금융기관의 안정성을 파악하고, 금융기관을 대상으로 한 정책을 제안할 경우에는 이러한 자료의 사용이 적절하다. 반면 가계부채를 과소추정하게 될 가능성이 있다(한국조사연구학회, 2016). 가계의 안전성을 파

악하고, 복지정책, 주거정책 등과 같은 다양한 분야와 다양한 계층의 가계부채 관련 정보 파악하고자 할 경우 미시적인 관점에서 가계의 부채행동에 대한 분석을 수행하는 것이 필요하다. 따라서 보다 유연한 추정방법을 활용하며, 미시적인 관점에서 중년기 가계의 가계부채 부실위험 수준은 어떠한지, 혹은 가계부채 부실위험이 높은 중년기 가계를 분류하고 예측할 수 있는 연구들이 필요하다고 사료된다.

본 연구의 목적은 한국 중년기의 가계부채 부실위험 수준을 파악하고, 중년기 가계가 가계부채 부실위험에 이르는 경로를 도출하여, 부실위험이 높은 중년기 가계를 분류 및 예측하는 것이다. 연구의 목적을 달성하기 위하여 구체적인 연구의 목표를 다음과 같이 설정하였다. 중년기 가계의 가계부채 부실위험 관련 요인들을 포괄적으로 분석하기 위하여 머신러닝 기반의 의사결정나무 분석법(Decision Tree Analysis)을 적용하였다. 이러한 연구방법을 바탕으로 부실위험이 높은 중년기 가계를 분류 및 예측할 수 있는 모형을 구축하고자 한다. 이러한 시도를 통하여 그동안 연구가 많이 진행되지 못했던 중년층의 가계부채 문제를 사회적으로 환기시키는 계기가 될 수 있을 것이다. 본 연구를 통하여 도출된 결론을 바탕으로 한국의 중년기 차입자들이 직면한 가계부채 부실화 관련 요인들을 기술하고, 정책 방안을 제시하고자 한다. 또한 본 연구에서 제안한 분석틀과 연구방법들이 향후 중년기의 부채에 관심을 두고 있는 후속 연구자들에게 기초적 연구로서 제공될 수 있을 것으로 기대한다.

II. 이론적 배경

1. 중년기의 개념

본 연구의 주요한 연구 대상은 생애주기(life cycle)에서 중년기에 속하는 중년기 차입자들이다. 중년기는 일반적으로 초기 성인기 과업이 달성되는 시기부터 노년기에 이르는 시기까지의 기간이다(Berk, 2007). 인생의 전반에서 후반으로 바뀌어 가는 전환점으로서 성숙기 혹은 갱년기로 묘사된다(전산초, 최영희, 1985). 이 시기는 인간의 생애주기적 특성상 사회경제적으로 중추적인 역할을 담당하며 자녀의 양육 및 교육에 대한 책임을 완수하면서 인생후반부를 대비하는 중요한 시기이다 강신기, 조성숙, 2013).

중년기의 범위는 학자들에 따라, 연구관점에 따라 구분이 다르나(박재순, 최의순, 1995), 고기숙(2003)은 중년기를 세 가지 기준으로 구분할 수 있다고 설명하였다. 첫째는 연령을 기준으로 하는 것이며, 둘째는 가족생활주기와 연령을 동시에 고려하는 것이며, 셋째는 가족생활 주기를 기준으로 하는 것이다. 첫 번째 방법은 연령만을 고려하는 중년기를 구분하는 방법이다. 이 방법은 연구 대상을 명확히 구분하는데 용이하므로 일반적이며 보편적으로 사용된다 (정성훈, 2013). 김명자(1998)와 김혜영과 고효정(1997)은 중년기를 40-59세로 정의하였으며, Levinson, Darrow, Klein, Levinson과 McKee(1978)은 중년기를 40-60세 시기로 신체적 능력은 다소 감소하지만 사회적 책임이 더 커지는 시기로 정의하였다. 둘째는 연령과 가족주기를 함께 고려하여 중년기를 구분하는 방법이다. 이 구분에 의하면 연령이 40-59세에 속하고 막내 자녀가 중학교 이상 재학 중인 경우를 중년기로 본다(신기영, 옥선화, 1997) 이 구분은 막내 자녀가 사춘기에 접어들면서 부모로부터의 심리적 독립한다는 가정에서 설정된 것으로 한국 가계에 그대로 적용하는 데에 무리가 있다(강덕진, 2010). 셋째, 가족생활주기를 기준으로 하여 생활사건과 가족관계의 변화에 따라 중년기를 구분할 수 있다. 막내 자녀의 독립부터 시작해서 자신의 은퇴까지의 시기로서 탈부모기, 진수기, 빈둥지기 등으로 나누어 중년기를 설명한다(Borland, 1978). 그러나 막내자녀의 독립시기를 고등학교 졸업 이후로 설정하는 것은 서구의 실정에 맞으며, 한국은 결혼 전까지 혹은 결혼 후에도 부모에게 의존하고 있는 경우가 많으므로 한국 가계에 적용하기에는 문제가 있다(강덕진, 2010). 따라서 본 연구에서는 첫 번째 방법인 응답자의 연령을 기준으로 중년기를 40세부터 59세까지로 정의하였다.

2. 중년기의 가계부채

생애주기에서 중년기에 속하는 중년기 차입자만을 대상으로 수행한 가계부채 상환 연구는 부족한 편이다. 그러나 여러 연령대 중에서 부분적으로 중년기 차입자의 가계부채의 보유 현황 및 규모를 분석한 연구들이 소수 존재한다. 가령 김지섭(2015a)은 한국과 미국의 가계부채를 비교 분석한 결과, 한국의 고령층은 가계소득이 상대적으로 낮고 보유하고 있는 부채는 상대적으로 많은 편이었다. 한국의 경우 50대가 보유하고 있는 부채의 비중은 전체 가계부채의 약 33%로 매우 높은 상황이었으며, 평균부채의 규모 또한 전체 평균보다 30% 높은 것으로 나타났다. 김우영과 김현정(2010)은 한국노동패널자료(2000-2007년)를 이용하여 한국가계의 부채보유여부, 부채규모, 소득대비 부채비율 등에 영향을 미치는 변수들을 분석하였다. 분석 결과, 한국가계의 부채보유 확률은 45세까지

증가하다가 그 이후부터 감소하고, 부채규모는 55세를 기점으로 감소하는 것으로 나타났다. 또한 가구주의 교육수준이 높거나, 자영업자인 경우 교육비의 부담이 큰 가계인 경우, 부채보유확률 및 부채규모가 증가하였다. 또한 거시환경 변수로 부동산 가격이 상승한 경우 부채보유 확률 및 부채규모에 영향을 미치는 것으로 나타났다.

가계부채의 상환여부에 초점을 둔 연구도 존재한다. 중년기 차입자들의 가계부채 부실화 가능성을 평가하는 것은 단순하지 않다. 중년기는 다른 연령대의 차입자들과 비교하여 상대적으로 소득 및 자산 수준이 높기 때문에 (성현구, 박범기, 2018) 이들의 가계부채의 상환 여력이 비교적 양호할 것이라고 판단할 수 있다. 그러나 한국의 가계대출은 단기거치식 상환방식이 차지하는 비중이 높으며, 대출을 받은 후 약 3-5년 후에 차환대출을 받는 식으로 부채를 상환하므로 부채의 원금이 시간이 경과함에 따라 크게 감소 되지 않는 구조이다(김지섭, 2015a). 이러한 경우 중년기 집단은 시간이 경과하더라도 부채를 상환하지 않고 상당 부분 그대로 유지하게 될 가능성이 높다 (김지섭, 2015a). 따라서 중년기가 시간에 흐름에 따라 노년층으로 진입하게 되면 소득의 하락으로 인하여 유동성 측면에서 문제가 발생할 수 있다.

3. 가계부채 상환 관련 연구

최근 한국의 지속적인 가계부채 증가와 그로 인한 관심의 증대로 가계부채 관련 연구들이 지속적으로 이루어져 왔다. 초기의 연구들은 가계부채의 상환을 연구하기 위하여 직접적인 연체정보를 활용하는 대신 자산 대비 부채 비율 혹은 소득 대비 부채상환액 비율 등의 재무비율을 활용하여 가구수준의 부채 상환위험을 간접적으로 평가하였다. 이러한 경향은 과거에는 실증분석에 활용할 만큼 충분한 자료가 축적되지 않아 직접적으로 연체정보를 활용하는 것이 어려웠기 때문이다(김영일, 2019). 가령 김영일과 변동준(2012)은 차입자 단위의 연구자료를 이용하여 지역별, 연령별, 대출업권별, 다중채무여부 등을 기준으로 가계부채의 상환취약성을 분석하였다. 분석 결과 과대채무자, 다중채무자, 자영업 채무자, 저소득 및 저신용 채무자가 가계부채의 상환취약성이 높은 것으로 나타났다. 특히 소득 하락과 금리인상 스트레스 상황에 노출될 경우, 비은행권 차입자, 자영업자, 다중채무자의 부실 위험이 상승할 것으로 예상하였다. 박인수와 박창수(2018)는 2016년 가계금융조사를 이용하여 가구특성 변수와 연체위험 간 관련성을 분석하였다. 이를 위하여 DSR 40% 또는 DTA 70%를 기준으로 일반가계부채 가구와 한계가계부채 가구로 나누어 두 집단의 연체 가능성을 낮추는 요인들을 이항 로짓 모형을 이용해 추정하였다. 추정결과 경상소득 기준으로는 4분위와 5분위, 그리고 자산총액 기준으로는 3~5분위의 한계가계부채 가구들과 일반가계부채 가구들이 연체 가능성이 낮은 것으로 나타났다. 이종희(2018)는 가계금융복지조사를 활용하여 자산부채 부채비율(DTA)과 원리금상환비율(DSR)이 임계치를 초과하는 취약가계를 분석하였는데, 연구결과 자가 소유, 금융자산 및 실물자산 규모는 가계의 부실화가능성을 낮추는 변수로 나타났다.

비교적 최근의 연구들은 가계금융조사 등의 일부 가구조사 자료에서도 채무불이행 관련 정보를 본격적으로 파악하기 시작함으로써 가구 단위의 상환연체를 파악할 수 있는 자료 여건이 조성되기 시작하였다(김영일, 2019). 박윤태와 노정현(2017)은 2016년 가계금융복지조사를 활용하여 가구를 연령별로 구분하여, 가계부채 상환위험과 관련된 변수들을 분석하였다. 연구결과 20-40대 비 근로자 집단의 상환위험이 높았는데, 이는 이들 집단의 소득 및 누적 자산이 적기 때문인 것으로 해석하였다. 50대의 경우 부동산 담보대출이 높을수록 상환위험이 높았는데, 이는 부동산을 이미 소유한 상태에서 다른 부동산을 구입하기 위하여 담보대출을 활용하기 때문이라고 설명하였다. 60대의 경우 은퇴 이후 근로 소득의 감소가 상환위험에 영향을 미치는 것으로 나타났다. 중년기는 가계부채 상환에 있어서 경제활동의 은퇴를 앞두고 있기 때문에 매달 납입하고 있는 월세와 투자 및 사업자금 마련은 부동산담보대출을 활용하여 조달하는 것으로 나타났다. 또한 타 연령층과 달리 중년기는 주거 마련을 통해서 상환위험을 낮추고 있기 때문에 주택시장의 변화에 따라 가장 민감한 그룹으로 분석되었다.

4. 머신러닝을 활용한 가계부채 연구

머신러닝이 비약적인 발전을 통해 과거보다 뛰어난 성능의 모형을 구축하는 것이 가능해지면서, 최근 사회과학 실증분석에서도 머신러닝 기법을 적극적으로 활용하기 시작하였다. 특히 경제분석과 관련하여 머신러닝을 활용하는 방안에 대해 논의가 이루어지기 시작하였으며, 금융서비스, 의료서비스, 정부부처와 공공 서비스를 제공하는 기관에서 머신러닝의 활용이 적극적으로 이루어지고 있다(이상길, 2018).

가계부채 관련 선행연구들의 연구방법론을 살펴보자면 통계적 가정을 기본으로 하는 모수적 (parametric) 방법을 이용하여 가계부채 규모 및 가계부채 비율의 결정요인을

분석하는 연구가 주류를 이루고 있다. 선행연구들을 살펴보면 김주영과 장희순(2016)은 가계의 부채보유를 분석하기 위하여 2006년 2014년 한국복지패널자료에 로지스틱 회귀분석을 실시하였다. 연구결과 가계소득, 연령, 성별, 가구원수, 교육, 결혼상태, 주택유형, 거주지역, 자산, 주택보유가 가계의 부채보유에 영향을 미치는 것으로 나타났다. 박대근과 최우주(2015)는 가계부채비율을 분석하기 위하여 2003-2012년 12개 국가의 패널자료를 활용하였으며, 다중 회귀분석을 통하여 분석을 시도하였다. 연구결과 주택가격상승률, 대출심사기준과 같인 시장요인이 가계부채 비율에 영향을 미치는 것으로 나타났다. 전승훈과 임병인(2008)은 부채규모를 분석하기 위하여 2006년 가계자산조사 및 2000년 가구소비실태조사에 이항 로지스틱회귀분석을 적용하였다. 연구결과 연령, 성별, 교육, 가구원수, 결혼상태, 직업, 소득, 자산, 부동산, 월세 등이 유의한 변수인 것으로 나타났다. 그러나 가계대출과 관련된 요인이 유형 별로 상이할 뿐만 아니라 잠재적인 설명변수의 수가 대단히 많기 때문에(이창훈, 강규호, 목정환, 2018) 특정 함수형태를 고려한 모수적 추정방법은 함수형태에 따라 문제점이 발생할 수 있다. 이런 점에서 전통적인 선형모형보다 유연한 추정모형을 구현할 필요성이 제기된다. 전술한 바와 같이 가계부채 선행연구들은 대부분 모수적 방법을 이용하였다. 그러나 전통적인 선형모형보다 유연한 추정모형을 구현할 필요성이 있다고 보고, 본 연구에서는 머신러닝 기법을 활용한 예측모형을 구축하고자 한다. 특히 본 연구는 머신러닝(Machine Learning) 분류 모형 중 의사결정나무 분석법을 실시한다.

III. 연구방법

1. 연구자료

본 연구는 중년기 차입자의 가계부채 상환연체 가능성을 분류하고 예측하고자 한다. 이를 위하여 통계청, 금융감독원, 한국은행이 공동으로 실시한 최신의 2019년「가계금융・복지조사」에 머신러닝의 기법을 적용한다. 이 자료는 통계청, 금융감독원, 한국은행이 전국의 2만 표본가구를 대상으로 하며, 가계의 자산, 부채, 소득, 지출 등을 통해 재무건전성을 파악하고, 경제적 삶의 수준 및 변화 등을 미시적으로 파악하기 위한 자료이다 (통계청, 2019). 중년기의 가계부채 부실위험에 영향을 미치는 요인을 도출하고 규칙을 발견하기 위한 최적의 연구자료라고 사료된다.

2. 주요변수

본 연구의 종속변수의 측정을 위하여 중년기 가계의 상환연체 여부를 문항을 활용하였다. 2019년 가계금융복지조사가 이루어지기 이전 1년간 응답자 가계에서 원금을 상환하거나 이자를 낼 날짜를 지나친 적이 있는지 여부로 측정한다. 답변은 '있다' 혹은 '없다'로 나뉘며, '있다'고 응답한 가계의 경우 원금 혹은 이자 상환을 연체한 가계로 간주한다.

설명변수로는 인구사회학적 변수, 재무적 변수, 부채관련 변수 등을 포함한다. 인구사회학적 변수로는 성별, 혼인상태, 교육수준, 거주지, 가구원수를 포함하며, 재무적 변수로는 입주형태 및 종사상 지위를 포함한다. 부채관련 변수로는 주요변수로 단기부채부담지표인 월평균가처분소득 대비 월평균부채상환액 비율, 중기부채부담지표인 금융자산 대비 총부채 비율, 장기부채부담지표인 총자산 대비 총부채를 포함한다.

3. 불균형데이터 문제 및 해소

데이터 불균형 문제는 목표변수가 이항형이며, 두 집단에 속하는 데이터 수의 비율이 크게 차이가 나는 경우를 의미한다. 특히 연구자의 관심의 대상이 되는 집단에 속하는 데이터의 수가 현저하게 낮은 비율로 발생하는 경우이다(오장민, 장병탁, 2001). 이렇듯 분류 문제에서 연구자가 관심을 갖는 집단에 속하는 데이터의 수가 현저하게 작은 경우, 머신러닝 알고리즘의 성능이 저하될 수 있다(강필성, 이형주, 조성준, 2004). 그 이유는 한 쪽의 집단에 속한 데이터의 수가 불균형적으로 큰 경우 머신러닝 알고리즘은 전체적인 오분류를 작게 하기 위해서 데이터의 수가 불균형적으로 큰 집단으로 패턴 분류를 하게 되고, 이 경우 소수의 집단에 속한 데이터를 다수의 집단에 속한 데이터로 취급하여 올바른 분석을 진행할 수 없기 때문이다(Weiss & Provost, 2001; 허준, 김종우 2007 재인용). 이러한 경우 분류모형의 정확도(accuracy)는 매우 높으나, 실제 부실을 부실이라고 예측하는 확률인 재현율(recall)이 매우 작아지는 현상이 발생하게 된다. 정확도만을 고려한다면 매우 성능이 좋은 분류 모형인 것처럼 보이나, 실제로는 연구자의 관심의 대상이 되는 가계부채 상환 연체가계에 대한 예측 능력이 매우 낮으므로, 적절한 모형이라고 할 수 없다.

머신러닝 알고리즘의 성능을 저하시킬 수 있는 데이터 불균형과 같은 문제가 발생하면 이러한 문제를 해소하는 방법으로 샘플링(sampling) 기법을 활용할 수 있다. 불균

형한 이항 자료를 분석할 때 가장 널리 사용되는 두 가지 샘플링 방법은 다운샘플링(Down‐sampling) 기법과 오버샘플링(Over-sampling) 기법이다. 다운샘플링 기법은 다수 범주 집단에서 임의적 샘플링을 하여, 소수 범주와 균형을 이루도록 하는 기법이다. 이 경우 연구자의 관심 대상이 아닌 집단의 데이터를 랜덤하게 제거함으로써 데이터의 불균형 문제를 해소한다 (김한용, 이우주, 2017). 오버샘플링 기법은 관심의 대상이 되는 소수의 집단 랜덤하게 복제함으로써 다수 범주 집단과 균형을 맞추어 불균형 문제를 해소하는 식이다(김한용, 이우주, 2017).

전술한 샘플링 방법은 다음과 같은 한계점을 갖는다. 가령 다운샘플링의 경우 다수 범주에서 랜덤하게 데이터를 제거하는 하면서 정보손실의 문제가 발생할 수 있다. 또한 오버샘플링의 경우 소수의 집단 랜덤하게 복제함으로써 다수 범주 집단과 균형을 맞추어 불균형 문제를 해소하면서 과적합의 문제가 발생할 수 있다. 이러한 점에서 안철휘와 안현철(2018)은 기존의 샘플링 기법들이 가질 수 있는 문제점들을 해소할 수 있는 방법을 제안하였다. 그들은 효과적인 기업부도 예측 모형 학습을 위하여 2014년에 Menardi와 Torelli가 제안한 ROSE(random over sampling examples)(Menardi & Torelli, 2014)기법을 활용하였다. 이 방법을 통하여 과적합화 문제를 피하면서도 분류 예측 정확도 개선에 도움을 줄 수 있음을 확인하였다. 이에 본 연구에서도 학습에 사용될 사례를 반복적으로 새롭게 합성하여 생성(synthetic generation)하는 기법인 ROSE 기법을 중년기의 가계부채 상환 예측 모형 구축에 활용하였다.

4. 통계적 방법

1) 기술통계

본 연구는 중년기의 가계부채 상환연체 가능성에 영향을 주는 분류기준을 파악하고 연체가능성을 예측하고자 한다. 중년기 차입자의 가계부채 원리금 상환여부를 목표변수로 (target variable)로 두고, 이에 영향을 주는 투입변수(input variable)을 선별하고자 한다. 우선 기술적 통계 방법을 활용하여 중년기와 조사대상자 전체의 인구사회학적 특성, 재무적 특성, 부채관련 특성을 살펴본다. 또한 연령대별 가계부채와 자산 분포를 보다 직관적으로 살펴보기 위하여 히트맵(heatmap)를 이용하여 분석을 수행하였다. 히트맵은 데이터의 값을 색으로 변환시켜 시각적인 분석을 가능하게 하는 데이터 시각화 기법 중 하나이다. 히트맵은 가계의 부채와 자산을 이들 조합의 비중을 색의 농도로 표현한 것이다.

2) 머신러닝

선행연구(e.g. 김주영, 장희순, 2016; 박대근, 최우주, 2015; 전승훈, 임병인, 2008)을 고찰해 본 바와 같이 가계부채 관련 선행연구들은 통계적 가정을 기본으로 하는 모수적 (parametric) 방법을 주로 이용하였다. 이와 같이 선형회귀를 중심으로 한 기존의 계량경제의 많은 분석방법들이 특성변수와 반응변수의 관계를 사전적으로 모형화하고 모수추정치를 추정하는 것에 집중하였다(정재현, 2019). 그러나 가계부채를 결정하는 데에는 차입자의 의사결정과정 속에 포함된 비선형성과 불확실성이 존재할 수 있고, 가계대출과 관련된 잠재적인 설명변수의 수가 대단히 많기 때문에(이창훈, 강규호, 목정환, 2018), 특정 함수형태를 고려한 추정방식은 한계를 가질 수 밖에 없다. 따라서 모델에 대한 특정한 가정 없이 반응변수의 예측치의 정확도를 높일 수 있는 알고리즘을 제공하는 머신러닝 기법을 활용하는 것이 중요하다고 사료된다. 따라서 본 연구에서는 부실위험이 높은 중년기 가계를 분류하고 예측할 수 있는 모형 구축을 위하여 의사결정나무 분석을 실시하였다.

의사결정나무 분석은 인공지능, 기계학습, 통계분석에서 많이 활용되고 있는 알고리즘으로 데이터마이닝 분석의 대표적인 분석 방법이다. 의사결정나무 분석모형은 표본집단을 특정 기준값에 의해 집단을 분류하고, 분류된 하위집단을 다시 특정 기준을 찾아 분류하는 과정을 반복함으로써 자료 내에 존재하는 관계, 규칙 등을 탐색하는데 유용하다(박명화, 최소라, 신아미, 구철회, 2013). 한편 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하고자 하는 경우에 사용된다(최종후, 한상태, 강현철, 김은석, 1998). 이 분석 방법은 목표변수에 영향을 미치는 투입변수를 찾아내고 각 투입변수의 중요도를 결정할 수 있다는 점에서 유용하다.

3) 불균형데이터 문제

본 연구에서 사용하는 목표변수를 측정하기 위하여 응답자가 가계부채 원리금을 상환하였는지 여부를 묻는 문항을 활용하였다(참조 3.주요변수). 응답자 중 가계부채 원리금을 연체하였다고 답한 응답자는 10.2%이며, 원리금을 연체하지 않았다고 응답한 응답자가 89.8%다. 이 경우는 목표변수가 두 가지의 결과만을 갖는 이항형 문항으로 관심 범주의 사례 수가 상당히 낮은 비율로 발생하는 불균형데이터 환경이라고 할 수 있다. 전술한 바와 같이 불균형 데이터 상황에서는 머신러닝 알고리즘의 성능저

하의 문제가 발생할 수 있다. 따라서 본 연구에서는 이러한 불균형 데이터 문제를 해소하고자 Menardi와 Torelli (2014)가 제안한 ROSE 기법을 활용한다.

4) 연구모형의 성능평가

머신러닝에서 성능 측정은 필수적인 작업이다. 전술한 연구방법을 적용한 후 도출된 분류 또는 예측 모형의 유용성을 판단할 필요가 있다. 즉 모형의 평가단계를 거쳐야 하는데, 분류 모형의 성과를 평가하는 방법으로는 분류 정확도를 측정하는 방법과 AUC-ROC 곡선을 활용하는 방법이 있다. 분류정확도는 정오분류표(confusion matrix)로 불리우는 분류행렬표(classification matrix)로 측정된다. 머신러닝 모형의 정확도를 평가하기 위한 사용되는 평가 측도이다(박수호 등, 2018). 이 분류행렬표는 분류모형이 특정 데이터 집합에 대해 수행한 정분류와 오분류의 요약정보를 보여주며, 정오분류표의 행과 열은 각각 실제집단(Actual Value)과 예측집단(Predicted Value)과 대응된다. 한편 분류모형은 모든 측면에 있어서 우세한 집단에 속하도록 분류하는 단순규칙(naïve rule)보다는 최소한 나은 분류를 해야 한다. 특히 본 연구에서 불균형적인 이항자료를 분류할 때의 예측성능을 평가하기 위하여 여러 지표 중 주로 재현율(Recall rate)의 상승에 주목한다. 재현율은 범주 1을 범주 1로 구분하는 확률을 뜻하고, 특이도는 범주 0을 범주 0으로 구분하는 확률을 나타낸다. 본 연구에 적용하자면 재현율은 중년기 차입자 중 가계부채 부실위험이 존재하는 가계를 중년기 부실가계로 진단하는 것을 의미한다. 그 외 오분류율을 이용하기도 하는데, 이는 전체 데이터에서 잘못 분류한 자료의 비율을 의미한다. 그러나 오분류율은 기본적으로 대칭 손실을 가정하기 때문에 대부분의 불균형 자료에 적용하기에는 한계가 있다(김동아, 강수연, 송종우, 2015).

		Actual Value	
		0	1
Predicted Value	0	3,415	221
	1	1,422	327

그림 1. Confusion Matrix

분류 모델의 성능을 확인하는데 또 다른 방법은 ROC (Receiver Operating Characteristics) AUC(Area Under The Curve)를 활용하는 것이다. Roc AUC는 분류 문제에 가장 일반적으로 사용되는 지표 중 하나이다. ROC 곡선은 FP(False Positive)의 비율에 대한 TP(True Positive)의 비율을 표시함으로써 분류자의 재현율을 보여준다. AUC가 ROC 곡선 아래의 영역으로 AUC가 높을수록 모델은 0을 0으로, 1을 1로 잘 예측한다는 것을 의미한다. 기본적으로 AUC를 극대화하는 것이 선호된다. 가령 AUC가 0.7이면 모형이 양성 클래스와 음성 클래스를 구별 할 수있는 확률이 70 %라는 것을 의미한다. 일반적으로 AUC가 0.6이상이면 만족스러운 수준이며, 0.9이상이면 대단히 우수한 모형으로 판단한다.

표 1. AUC Value 기준

AUC Value	Test Quality
0.9-1.0	Excellent
0.8-0.9	Very Good
0.7-0.8	Good
0.6-0.7	Satisfactory
0.5-0.6	Unsatisfactory

최종적으로 본 연구에서는 모형의 타당성을 확인하기 위해 교차타당성 검증(cross validation)을 실시한다. 교차타당성 검증은 훈련용 데이터와 평가용 데이터를 서로 변경하면서 모형의 타당성을 검증하는 방법으로, 샘플 데이터에 대한 과적합을 방지하는데 기여하는 것으로 알려져 있다(오준병, 허원창, 이혜민, 2019). 한편 이러한 기본적인 개념을 바탕으로 엄밀성과 정확성을 높인 방법인 K겹 교차타당성 검증(K-fold-cross validation)을 수행한다. 이 방법은 하나의 데이터셋을 k개로 분할하고 그 중 하나는 테스트 용으로 사용하고 나머지 k-1개는 훈련용으로 사용하는 것이다(Tan, Steinbach, & Kumar, 2006). 이러한 방법을 통하여 부실화 위험이 높은 중년기 차입가구의 분류 및 예측을 위한 추천모형을 선정하고자 한다.

IV. 연구결과

본 연구의 목적은 한국 중년기의 가계부채 부실위험 수준을 파악하고, 중년기 가계가 가계부채 부실위험에 이르는 경로를 도출하여, 부실위험이 높은 중년기 가계를 분류 및 예측하는 것이다. 이러한 연구의 목적을 달성하기 위하여 중년기 가계를 포함한 전 연령대의 소득, 자산, 부채의 비중 및 규모를 조망한 후 머신러닝을 수행하고자 한다. 전 연령대의 소득, 자산, 부채를 조망하기 위한 기술통계에서는 18,637가계를 분석하였으며, 부채 상환에 응답자를 대상으로 하는 기술통계의 경우 9,851 가계를 대상으로 분석을 수행한다. 한편 머신러닝의 수행을 위해서

는 부채상환에 응답한 중년기 가계인 5,385가계를 최종 분석의 대상으로 한다.

본 연구에서는 학습데이터의 불균형 문제는 머신러닝 알고리즘 성능을 저하시키는 주요한 원인이 된다는 점에서 불균형 데이터 문제를 해소하고자 한다. 그 방법으로 ROSE(random over-sampling examples) 기법을 활용하고, 샘플링을 하지 않은 모형과 비교한다. 머신러닝기법을 활용하기 위하여 훈련데이터와 평가데이터를 나누는 비율은 연구자에 의해 결정될 수 있는데, 주로 70/30 rule을 활용한다(Genkin, Dehne, Navarro, & Zhou, 2019). 이에 따라 본 연구에서는 훈련데이터로 3,770가계를 이용하고, 평가데이터로 1,615가계를 이용한다.

1. 조사대상자의 소득, 자산, 및 부채

중년기 가계의 부채부실화 가능성을 조망하기 위하여, 응답자들을 20-30대, 40-50대, 60대이상의 연령대로 구분하고, 소득, 자산, 부채의 비중 및 규모를 조망한다. <표 2>는 연령대별 전체 소득, 자산, 부채의 비중을 나타낸 결과이다. 2019년 가계금융복지조사에서 20세-39세는 2,858가계 (15.34%), 40-50대 (40-59세)는 8,195가계(43.97%), 60대 이상은 7,584가계(40.69%)이었다. 20-30대는 총소득에서 16.38%, 40-50대는 62.41%, 60세 이상은 21.21%에 해당한다. 총자산의 경우 20-30대, 40-50대, 60대 이상이 각각 11.75%, 57.50%, 30.75%에 해당한다. 총부채의 경우 총부채의 절반이상(59.10%)을 40-50대가 보유하였다. 총부채는 금융부채와 임대보증금의 합이며, 금융부채는 담보대출, 신용대출, 기타대출로 구성된다. 담보대출과 신용대출의 경우에도 40-50대가 절반이상을 보유하여 다른 연령대보다 상대적으로 높은 것으로 나타났다.

40-50대 응답자들은 전체 응답자의 절반이하 (43.97%)이었으나, 이들이 보유하는 총 가계부채는 50%이상이며, 담보대출 및 신용대출의 경우에도 이들 집단의 보유규모가 다른 연령대보다 컸다. 그러나 중년기는 다른 연령대보다 상대적으로 소득 및 자산 수준이 높은 것으로 나타났다. 이러한 점에서 이들 집단이 많은 양의 부채를 감내할 수 있으며, 가계부채의 상환 여력이 비교적 양호하다고 판단할 있다. 그러나 10여년 이후 이 집단이 가구소득이 손실 혹은 급감하는 은퇴를 맞이할 경우 고령층의 가계부채 문제로 이어질 가능성이 존재한다.

표 2. 연령대별 소득, 자산, 부채의 비중

단위 : %

	총소득 비중	총자산 비중	총부채 비중	담보대출 비중	신용대출 비중
20세-39세 (n=2,858)	16.38%	11.75%	15.54%	16.17%	21.09%
40세-59세 (n=8,195)	62.41%	57.50%	59.10%	59.64%	61.04%
60세 이상 (n=7,584)	21.21%	30.75%	25.36%	24.19%	17.87%

연령대별 가계부채와 자산 분포를 보다 직관적으로 살펴보기 위하여 히트맵(heatmap)을 작성한 결과를 <그림 2>에 제시한다. 히트맵은 가계의 부채와 자산을 이들 조합의 비중을 색의 농도로 표현한 것이다. 농도가 짙을수록 조합의 비중이 높은 것을 의미한다. 열 지도는 가계의 부채・자산을 각각 10분위로 나눈 후 시각화를 시도하였는데, 자산의 경우 10단위로 분할하고, 부채의 경우 8분위로 나누어 분석한 것이 시각적으로 해석이 용이하여 최종적으로 이를 활용한다. X축은 부채이며, 값이 커질수록 오른쪽으로 이동한다. Y축은 자산이며, 값이 커질수록 하단으로 이동한다. 분석 결과 20-30대와 40-50대의 경우 45

20-39 세 40-59 세 60세 이상

그림 2. 연령대별 부채 및 자산 분포

도선을 중심으로 분포가 집중되어 있는데 이는 부채가 많을수록 자산도 많은 가구의 집중도가 높음을 의미한다. 40-50대의 경우 특히 우하향 지점으로 갈수록 농도가 짙은데, 이것은 중년기의 가계부채와 자산의 연관성이 높음을 의미한다. 반면 60대이상의 경우 45도선 부근의 집중도가 낮아 부채와 금융자산간 연계성이 상대적으로 낮은 것으로 나타났다.

<표 3>은 가계부채 원리금 상환 여부에 응답한 차입자들을 대상으로 인구사회학적 특성 및 재무적 특성을 제시한 표이다. 본 연구의 관심대상인 40-50대 조사대상자(중년기)와 청년기(20-39세), 노년기(60세 이상)의 응답자의 결과를 함께 제시하였으며, 부채상환을 상환한 가계(연체○)와 상환을 연체하지 않은 가계로 나누어 결과를 제시하였다(연체×). 또한 변수의 유목에 따라서 통계적으로 유의미한 차이가 있는지 평가하기 위해서 카이제곱분석을 사용하여 자료를 통계 처리하였다. 분석결과 가구주의 성별에서는 중년기, 노년기의 경우 남성 가구주의 비율이 높았고, 이들 집단의 경우 기혼의 비율 및 자가 보유 비율이 높았다. 교육수준에서는 청년기와 중년기의 경우 대졸 이상의 비율이 높았으며, 이들 집단의 경우 상용근로자로 근무하는 경우가 많았다.

연구모형에 포함된 부채 관련 변수는 단기부채부담지

표 3. 조사대상자의 인구사회학적/재무적 특징

단위: 가계(%)

변수		연령구분								
		20-39세 (n=1,753)			40-59세 (n=5,385)			60세 이상 (n=2,713)		
		연체×(%)	연체○(%)	p-value	연체× (%)	연체○(%)	p-value	연체× (%)	연체○(%)	p-value
성별	여성	261 (16.2)	20 (14.5)	.6952	803 (16.6)	124 (22.6)	.0005	558 (22.6)	83 (33.5)	.0020
	남성	1,354 (83.8)	118 (85.5)		4,034 (83.4)	424 (77.4)		1,907 (77.4)	165 (66.5)	
혼인상태	기혼 외	417 (25.8)	40 (29.0)	.4765	1,072 (22.2)	182 (33.2)	<.0001	649 (26.3)	117 (47.2)	<.0001
	기혼	1,198 (74.2)	98 (71.0)		3,765 (77.8)	366 (66.8)		1,816 (73.7)	131 (52.8)	
교육정도	그 외	341 (21.1)	55 (39.9)	<.0001	,2283 (47.2)	347 (63.3)	<.0001	1,930 (78.3)	208 (83.9)	.0493
	대졸이상	1,274 (78.9)	83 (60.1)		,2554 (52.8)	201 (36.7)		535 (21.7)	40 (16.1)	
종사상지위	그 외	409 (25.3)	58 (42.0)	<.0001	2,117 (43.8)	337 (61.5)	<.0001	2,003 (81.3)	213 (85.9)	.0872
	상용근로자	1,206 (74.7)	80 (58.0)		2,720 (56.2)	211 (38.5)		462 (18.7)	35 (14.1)	
입주형태	그 외	863 (53.4)	81 (58.7)	.2711	1,536 (31.8)	276 (50.4)	<.0001	596 (24.2)	127 (51.2)	<.0001
	자가	752 (46.6)	57 (41.3)		3,301 (68.2)	272 (49.6)		1,869 (75.8)	121 (48.8)	
수도권	그 외	697 (43.2)	55 (39.9)	.5074	1,704 (35.2)	189 (34.5)	.7669	847 (34.4)	55 (22.2)	.0010
	수도권	918 (56.8)	83 (60.1)		3,133 (64.8)	359 (65.5)		1,618 (65.6)	193 (77.8)	

표 4. 조사대상자의 가계부채 비율 특징

변수		연령구분								
		20-39세 (n=1,753)			40-59세 (n=5,385)			60세 이상 (n=2,713)		
		연체×(%)	연체○(%)	p-value	연체×(%)	연체○(%)	p-value	연체×(%)	연체○(%)	p-value
DTA[1]	DTA≤0.4	995 (61.6)	52 (37.7)	<.0001	3599 (74.4)	246 (44.9)	<.0001	1982 (80.4)	109 (44.0)	<.0001
	0.4<DTA≤0.8	504 (31.2)	54 (39.1)		976 (20.2)	156 (28.5)		354 (14.4)	53 (21.4)	
	0.8<DTA	116 (7.2)	32 (23.2)		262 (5.4)	146 (26.6)		129 (5.2)	86 (34.7)	
DSR[2]	DSR≤0.2	881 (54.6)	58 (42.0)	.0109	2816 (58.2)	279 (50.9)	.0004	1440 (58.4)	142 (57.3)	.5481
	0.2<DSR≤0.4	346 (21.4)	42 (30.4)		996 (20.6)	114 (20.8)		479 (19.4)	44 (17.7)	
	0.4<DSR	388 (24.0)	38 (27.5)		1025 (21.2)	155 (28.3)		546 (22.2)	62 (25.0)	
DTFA[3]	DTFA≤1	884 (54.7)	56 (40.6)	.0025	2607 (53.9)	174 (31.8)	<.0001	1201 (48.7)	80 (32.3)	<.0001
	1<DTFA≤5	484 (30.0)	49 (35.5)		1551 (32.1)	220 (40.1)		660 (26.8)	66 (26.6)	
	5<DTFA	247 (15.3)	33 (23.9)		679 (14.0)	154 (28.1)		604 (24.5)	102 (41.1)	

[1]DTA=총부채/총자산
[2]DSR=원리금상환액/소득
[3]DTFA=총부채/금융자산

표인 가처분소득 대비 부채상환액 비율, 중기부채부담지표인 금융자산 대비 부채 비율, 장기부채부담지표인 자산대비 부채 비율이다. 이들 비율을 중년기, 청년기, 노년기로 나누어 분석한 결과를 <표 4>에 제시하였다. 자산대비 부채가 0.8보다 큰 가계, 가처분소득 대비 부채상환액 비율이 0.4보다 큰 가계, 금융자산 대비 부채 비율이 5보다 큰 가계의 경우 가계부채 부담수준이 높은 가계로 볼 수 있다. 이러한 가계부채 비율에 해당하는 가계의 중심으로 살펴보면, 자산대비부채비율이 0.8을 초과하는 가계의 비율은 노년기의 가계가 가장 높았고, 가처분소득 대비 부채상환액 비율이 0.4를 초과하는 가계의 비율도 노년기의 가계가 가장 높았다. 한편 금융자산 대비 부채비율이 5를 초과하는 가계의 비율은 중년기의 가계가 가장 높았다.

조사대상자의 가계부채 대출용도 및 대출기관의 차이를 중년기, 청년기, 노년기로 나누어 분석한 결과를 <표 5>에 제시하였다. 가계부채 상환 연체를 한 가계를 중심으로 특징적인 몇 가지를 살펴보면 다음과 같다. 중년기 가계의 경우 거주주택 마련을 위하여 담보대출을 이용한 가계의 비율이 높았고, 사업자금 마련 및 생활비 마련을 위하여 신용대출을 이용한 비율이 다소 높았으며, 은행을 통한 담보대출의 비율도 높았다. 청년기 가계의 경우 생활비 마련을 위하여 신용대출을 받은 비율이 높았다. 노년기 가계의 경우 거주주택 마련을 위해 담보대출 혹은 신용대출을 받은 비율이 낮은 편이었다.

표 5. 조사대상자의 가계부채 용도 및 대출기관 특징

단위: 가계 (%)

변수			연령대								
			20-39세 (n=1,753)			40-59세 (n=5,385)			60세 이상 (n=2,713)		
			연체×(%)	연체○(%)	p-value	연체×(%)	연체○(%)	p-value	연체×(%)	연체○(%)	p-value
담보대출용도	거주주택마련	0	1002 (62.0)	97 (70.3)	.0671	3023 (62.5)	382 (69.7)	.0011	1903 (77.2)	212 (85.5)	.0035
		1	613 (38.0)	41 (29.7)		1814 (37.5)	166 (30.3)		562 (22.8)	36 (14.5)	
	부채상환	0	1606 (99.4)	136 (98.6)	.2125	4748 (98.2)	537 (98.0)	.9140	2406 (97.6)	239 (96.4)	.3304
		1	9 (0.6)	2 (1.4)		89 (1.8)	11 (2.0)		59 (2.4)	9 (3.6)	
	사업자금	0	1563 (96.8)	130 (94.2)	.1360	4350 (89.9)	473 (86.3)	.0107	2181 (88.5)	215 (86.7)	.4651
		1	52 (3.2)	8 (5.8)		487 (10.1)	75 (13.7)		284 (11.5)	33 (13.3)	
	생활비	0	1567 (97.0)	129 (93.5)	.0395	4548 (94.0)	501 (91.4)	.0218	2252 (91.4)	220 (88.7)	.2003
		1	48 (3.0)	9 (6.5)		289 (6.0)	47 (8.6)		213 (8.6)	28 (11.3)	
담보대출기관	은행	0	681 (42.2)	81 (58.7)	.0002	2294 (47.4)	317 (57.8)	<.0001	1454 (59.0)	176 (71.0)	.0003
		1	934 (57.8)	57 (41.3)		2543 (52.6)	231 (42.2)		1011 (41.0)	72 (29.0)	
	저축은행	0	1594 (98.7)	135 (97.8)	.4295	4776 (98.7)	535 (97.6)	.0544	2435 (98.8)	244 (98.4)	.5460
		1	21 (1.3)	3 (2.2)		61 (1.3)	13 (2.4)		30 (1.2)	4 (1.6)	
	비은행	0	1534 (95.0)	128 (92.8)	.3503	4385 (90.7)	488 (89.1)	.2557	2050 (83.2)	195 (78.6)	.0866
		1	81 (5.0)	10 (7.2)		452 (9.3)	60 (10.9)		415 (16.8)	53 (21.4)	
신용대출용도	거주주택마련	0	1545 (95.7)	134 (97.1)	.5588	4732 (97.8)	537 (98.0)	.9246	2443 (99.1)	246 (99.2)	<.0001
		1	70 (4.3)	4 (2.9)		105 (2.2)	11 (2.0)		22 (0.9)	2 (0.8)	
	부채상환	0	1590 (98.5)	127 (92.0)	.0001	4750 (98.2)	528 (96.4)	.0054	2442 (99.1)	235 (94.8)	<.0001
		1	25 (1.5)	11 (8.0)		87 (1.8)	20 (3.6)		23 (0.9)	13 (5.2)	
	사업자금	0	1529 (94.7)	126 (91.3)	.1439	4390 (90.8)	465 (84.9)	<.0001	2225 (90.3)	214 (86.3)	.0616
		1	86 (5.3)	12 (8.7)		447 (9.2)	83 (15.1)		240 (9.7)	34 (13.7)	
	생활비	0	1361 (84.3)	100 (72.5)	.0006	4023 (83.2)	418 (76.3)	.0001	2143 (86.9)	203 (81.9)	.0329
		1	254 (15.7)	38 (27.5)		814 (16.8)	130 (23.7)		322 (13.1)	45 (18.1)	
신용대출기관	은행	0	1111(68.8)	96 (69.6)	.9264	3490 (72.2)	391 (71.4)	.7291	2028 (82.3)	187 (75.4)	.0100
		1	504 (31.2)	42 (30.4)		1347 (27.8)	157 (28.6)		437 (17.7)	61 (24.6)	
	저축은행	0	1565(96.9)	125 (90.6)	.0009	4761 (98.4)	527 (96.2)	.0003	2436 (98.8)	244 (98.4)	.5373
		1	50 (3.1)	13 (9.4)		76 (1.6)	21 (3.8)		29 (1.2)	4 (1.6)	
	비은행	0	1579(97.8)	130 (94.2)	.0188	4588 (94.9)	506 (92.3)	.0178	2236 (90.7)	218 (87.9)	.1867
		1	36 (2.2)	8 (5.8)		249 (5.1)	42 (7.7)		229 (9.3)	30 (12.1)	

2. 의사결정나무모형 결과

가계부채를 상환 하지 않는 중년기 차입자 가계의 비율은 응답자 중 약 10.18%이다. 이 경우는 목표변수가 이항변수이며 관심 범주의 사례 수가 상당히 낮은 비율로 발생하는 불균형데이터 환경이라고 할 수 있다. 데이터 불균형 문제는 머신러닝 알고리즘의 성능을 저하시킬 수 있으므로, 이를 해소하기 위한 방법으로 ROSE 기법을 활용하였다.

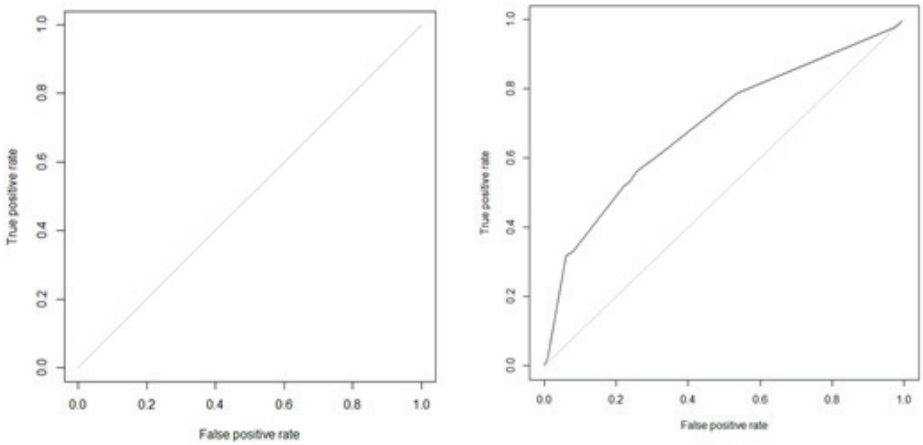
최종적으로 생성된 모형의 과적합(Overfitting)과 일반화의 오류를 평가하기 위해 분류의 검증은 K겹 교차검증을 활용한다. 분류 및 예측 모형의 성능을 평가하기 위한 분류행렬표는 <표 6>에 제시한다. 연구결과 샘플링을 하지 않은 분석의 경우 정확도는 89.86%이며, 재현율과 특이도는 각각 2.74%와 99.73%이다. ROSE 샘플링을 통한 연구모형의 정확도는 69.49%이며, 재현율과 특이도는 각각 59.67%, 70.60%이다. 본 연구에서 모형의 재현율이 많이 향상되었는데, 재현율이 2.74%에서 59.67%로 약 22배의 상승을 보였다.

샘플링 방법에 따른 모형 적합도를 알아보기 위한 ROC curve 검증결과를 <그림 3>에 제시하였다. 전술한 바와 같이 ROC 곡선은 실제로는 양성(+)이 아닌데 양성(+)으로 분류하는 FP(False Positive)의 비율에 대한 실제로는 양성(+)인데 양성(+)으로 분류하는 TP(True Positive)의 비율을 표시한다. ROC 곡선 아래의 영역이 높을수록 모델은 0을 0으로, 1을 1로 잘 예측한다는 것을 의미한다. 샘플링을 하지 않은 의사결정나무분석의 AUC는 0.50이며, ROSE 기법을 활용한 의사결정나무분석의 AUC 는 0.69이었다. 따라서 이 기법을 활용한 분류모형이 만족스럽지 않은 상태(unsatisfactory)에서 만족스러운(satisfactory)수준으로 향상된 것으로 판단하였다. 불균형 데이터를 이용한 모델링의 경우 정확도, 재현율, 특이도, ROC curve를 모두 고려하여 모형의 성능을 평가하고, 모든 지표에서 우수한 모형을 선택하는 것이 이상적이다. 그러나 연구결과 모든 지표에서 절대적으로 우위에 있는 모형이 존재하지 않으므로, 이론적 배경에서 고찰한 바와 같이 높은 재현율과 AUC 수준을 보여주는 연구모형의 경우 적합도가 향상된 것으로 판단하였다.

의사결정나무 알고리즘을 이용한 다중나무구조(Multi-tree structure)의 분류결과를 <그림 4>에 제시한다. 의사결정나무를 구축한 결과 모두 19개의 노드와 10개의 터미널 노드, 6개의 계층이 형성되었다. 연체가능성이 높은 차입가계들은 주로 다음과 같다. 상환연체자의 비율이 높은 집단은 노드 13, 7, 23에 해당한다. 7번노드의 경우 DTA가 0.8초과인 그룹으로 이들 중 81%가 상환연체 차입자에 해당한다. 13번 노드의 경우 DTA가 0.4초과 0.8이하이며, 저축은행을 통해 대출을 받은 중년기 차입자를 의미하는데, 이들중 84%가 상환연체를 한 것으로 나타났다. 23번 노드의 경우 DTA가 0.4이하이며, 상용근로자가 아니며, 자가를 소유하지 않고, 사업자금마련을 위하여 대출을 받은 경우로 이들 중 64%가 상환연체자로 나타났다. 상환연

표 6. 분류행렬표

	No Sampling	ROSE
정확도	89.86	69.49
재현율	2.74	59.67
특이도	99.73	70.60
오분류율	10.14	30.51



AUC = .50
No Sampling

AUC=.69
ROSE

그림 3. ROC curve 검증결과

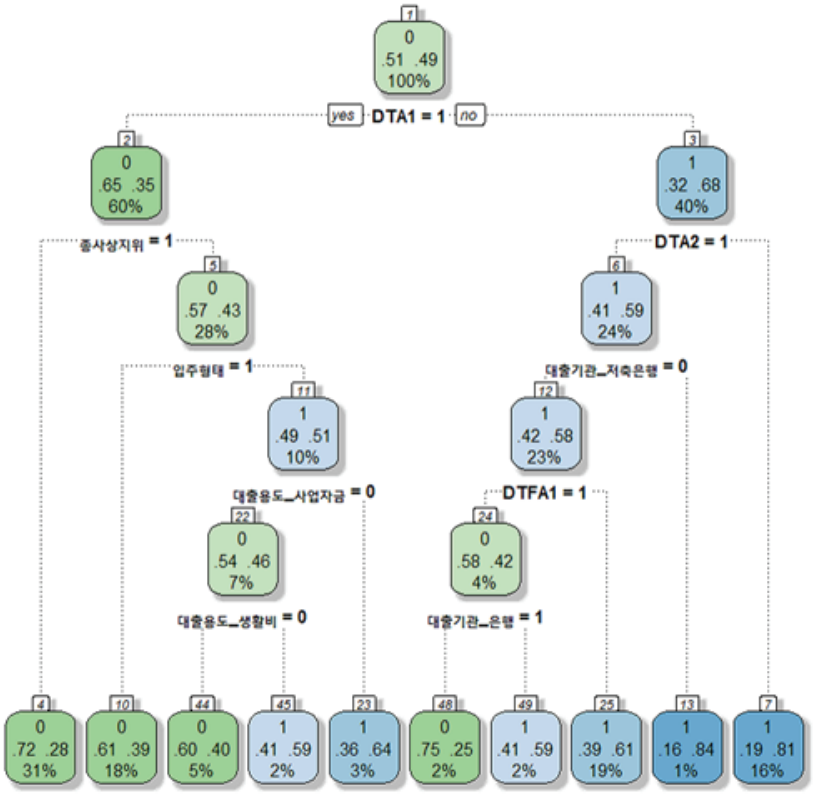


그림 4. ROSE를 활용한 의사결정나무모형

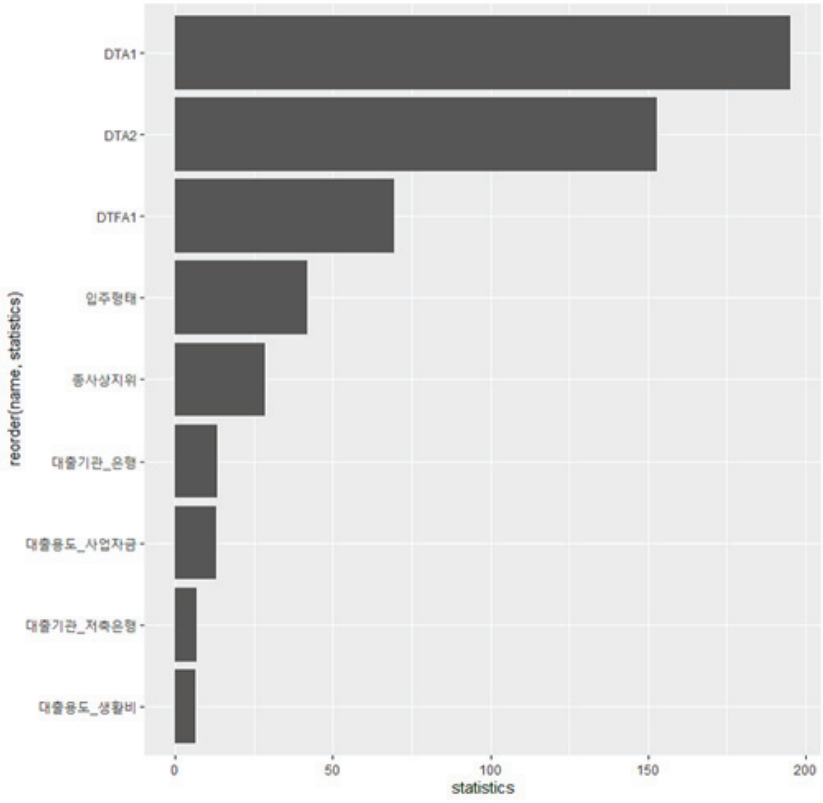


그림 5. ROSE를 활용한 의사결정나무모형의 변수 중요도

체자의 비율이 낮은 집단은 노드 48, 4, 10에 해당한다. 노드 48은 자산대비 부채비율이 0.4초과 0.8이하이며, 금융자산대비 부채비율이 1이하이며, 은행을 통하여 대출을 받은 집단으로 이들 중 25%가 상환연체를 하는 것으로 나타났다. 노드 4는 자산대비부채비율이 0.4이하이며, 상용근로자의 경우 연체자의 비율이 28%이었다. 자산대비 부채비율이 0.4이하이면서, 상용근로자가 아닌 경우, 자가를 소유한 경우 연체자의 비율이 39%이었다.

<그림 5>의 의사결정나무모형의 변수 중요도에 나타난 바와 같이 자산 대비 부채비율(DTA), 금융자산 대비 부채비율(DTFA), 입주형태, 종사상 지위, 은행을 통한 대출인지 여부, 사업자금 마련을 위한 대출인지 여부, 저축은행을 통한 대출인지 여부, 생활비 마련을 위한 대출인지여부 순으로 중년기의 가계부채 상환 연체 여부를 분류하는 주요 요인인 것으로 나타났다. 장기 부채부담지표과 중기 부채부담지표가 상환연체에 영향을 주는 변수로 나타난 반면 단기 부채부담지표인 소득 대비 원리금상환액(DSR)은 유의미한 설명변수로 나타나지 않았다. 이러한 결과는 가계의 단기적 부채부담정도 보다는 지속적, 중장기적으로 가계경제에 영향을 미치는 부채부담이 부채상환에 직접적인 영향을 미친다고 할 수 있겠다.

V. 결론 및 제언

본 연구의 목적은 통계청, 한국은행, 금융감독원이 발표하는 대용량 자료인 가계금융복지조사를 바탕으로 머신러닝 기반인 의사결정나무 분석모형을 활용하여 중년기 차입자의 가계부채 상환연체 가능성을 파악할 수 있는 예측 모델을 개발하고 관련 규칙을 도출하는 것이다.

연구결과를 정리하면 다음과 같다. 첫째, 기술통계 결과 본 연구의 주요한 연구대상인 중년기는 다른 연령대와 비교하여 상대적으로 소득 및 자산 수준이 높았다. 기술통계분석과 히트맵을 통하여 분석한 결과 중년기 차입자들의 경우 부채가 많을수록 자산도 많은 가구의 집중도가 높았다. 이러한 점에서 가계부채의 상환 여력이 다른 연령대보다 비교적 양호할 것으로 판단할 수 있다. 그러나 한국의 경우 단기 일시상환 방식의 계약구조가 대부분을 차지하며, 차환대출을 하는 경우도 많아 시간이 경과하더라도 부채의 원금이 크게 감소하지 않을 가능성이 크다. 이러한 경우 중년기 가계부채 문제가 노년기의 문제로 이어질 가능성이 존재한다. <표 3>, <표 4>, <표 5>의 내용을 간단하게 정리하면 중년층 응답자들은 가구주성별, 기혼 여부, 대졸 학력, 상용근로자, 수도권 거주의 비율이 20-30대 응답자와 60대 이상의 응답자의 중간정도에 해당하는 것으로 나타났다. 또한 부채관련 비율의 경우 DTA가 0.8을 초과하거나, DTFA가 5를 초과하는 응답자의 비율은 노년층에서 가장 높게 나타났고, 중년층의 경우에는 단기부담지표인 DSR이 0.4를 초과하는 비율이 높았다. 한편 중년층의 경우 거주주택 구입을 위하여 담보대출을 이용하거나, 사업 비용을 충당하기 위하여 담보대출을 이용하는 비율이 가장 높았다. 의사결정나무모형에 의하면 자산 대비 부채비율, 금융자산 대비 부채비율, 입주형태, 종사상 지위, 대출기관, 대출목적 순으로 중년기의 가계부채 상환 연체 여부를 분류하는 기준이 되는 것로 나타났다. 특히 단기 부채부담지표인 소득 대비 원리금상환액(DSR)은 유의미한 설명변수로 나타나지 않은 반면 장기부담지표과 중기 부채부담지표가 상환연체에 영향을 주는 변수로 나타났다. 이것은 가계의 단기적 부채부담 정도보다는 지속적, 중장기적인 부채부담이 부채상환에 영향을 미친다고 볼 수 있다. 이러한 점에서 중년기 차입자의 가계부채 연구에 대한 관심이 필요하다고 하겠다.

둘째. 불균형 데이터셋을 이용한 머신러닝의 경우 분류모형의 정확도는 매우 높으나, 실제 부실을 부실이라고 예측하는 확률인 재현율이 매우 작아지는 현상이 발생하게 되는데, 본 연구에서도 그러한 현상이 발견되었다. 즉 샘플링을 활용하지 않을 경우 현저하게 낮은 재현율이 나타나는 현상을 발견하였다. 따라서 Menardi와 Torelli (2014)가 제안한 ROSE기법을 활용하였다. 그 결과 연구모형의 재현율이 상당히 향상되었고, 분류모형이 만족스러운 수준으로 향상되었다고 판단하였다.

셋째, 본 연구에서 사용된 의사결정나무 분석모형은 표본집단을 특정 기준값에 의해 유사한 집단으로 분류하고, 분류된 하위집단을 다시 특정 기준을 찾아 분류하는 과정을 반복함으로써 목표변수와 입력변수들 간의 패턴이나 관계를 찾아내는데 유용하다(박명화 외, 2013). 또한 변수들 간의 상호작용효과(interaction effects)를 파악할 수 있다는 장점이 있다. 분석 결과 자산대비부채비율, 금융자산대비부채비율, 입주형태, 종사상지위, 대출금융기관, 대출목적이 중년기의 가계부채 상환 연체 여부를 분류하는 주요한 변수인 것으로 나타났다.

연구결과를 바탕으로 함의를 모색해보면 다음과 같다. 전체 가계부채의 상당부분을 중년기 가계가 보유하는 가운데 우리나라의 가계부채가 빠른 속도로 증가하고 있다. 중년기가 다른 연령대와 비교하여 상대적으로 소득 및 자산 수준이 높으므로, 중년기의 가계부채의 상환 여력이 양호한 수준일 것이라고 판단할 수 있다. 그러나 부채를 한 번 가지면 그 보유 연한이 상당히 긴 기간에 걸쳐 나타

나고 있으며, 인구구조의 고령화를 감안할 때 장기적으로 중년기의 가계부채 문제는 고령층을 중심으로 현재보다 심각해질 가능성이 있다(김지섭, 2015b). 따라서 부실화 가능성이 높은 중년기 차입자들의 상환을 할 수 있는 구조를 만들어야 한다. **중**년기 차입자들의 경우 조기퇴직 등으로 경제활동 상태가 완전하지 않은 상태에서 자녀 학비, 주거비 마련 등을 목적으로 가계부채가 늘어나고 있다(김태완, 김문길, 이주미, 김기태, 김명중, 홍성우, 2016)는 선행연구와 유사하게, 본 연구의 기술적 통계 분석결과 중년기에 해당되는 상환연체 차입자들은 주택 마련 및 사업자금을 마련하기 위하여 차입하는 비율이 다른 연령대보다 높았다. 40대와 50대의 경우 주택구입에 따른 부채의 증가로 유동성의 제약을 겪을 가능성이 높음에도 불구하고, 20~30대와 60대 이상을 대상으로 가계부채 관련 정책이 상대적으로 많은 반면 장년층을 대상으로 한 정책은 상대적으로 적다 (김민철 등, 2016). 따라서 중년기 차입자들에 대해서는 자금 유동화를 지원하는 등의 생애주기 맞춤형 지원 정책 등을 강구할 필요가 있다. 한편 평생교육을 통한 신용교육과 재무상담 등이 활성화하여 가계부채 문제 해소할 수 있도록 제도적 개선이 필요하다.

금융소비자를 위한 제언은 다음과 같다. 본 연구의 결과 부채비율, 입주형태, 종사상지위, 대출금융기관, 대출목적이 중년기의 가계부채 상환 연체 가능성에 영향을 미치는 주요 변수인 것으로 나타났다. 따라서 이러한 변수들이 자신의 가계 특성에 해당한다면 부채 차입 시 자신의 상환능력을 감안하여 과도한 차입이 발생하지 않도록 하여야 할 것이며, 신규 차입을 늘리는 대신 자발적인 디레버리징이 필요할 것이다. 가령 DTA가 0.8초과인 가계는 중년기 차입자의 약 7.58%이었으며, 이들 가계 중 약 81%가 상환연체가 해당하였다. 금융부채 대비 자산 비율이 1을 초과하는 가계가 48.41%에 달하였다. 본 연구는 DTA가 0.8초과인 가계에 대한 추가적인 분석을 수행하였다. 추가적인 분석을 통하여 DTA가 0.8초과인 가계인 중년기 가계는 경제적으로 보다 열악한 환경에 처해있다는 것을 발견할 수 있었다. 이들 가계의 경우 가구주의 종사상 지위가 상용근로자인 비율이 40.8%이었는데, 이는 40대의 부채상환연체 차입가계의 평균적인 상용근로자비율인 61.5%보다 유의미하게 낮았다. 또한 소득의 경우 DTA가 0.8초과인 가계의 평균은 35,890,000원이었으나, 이는 40대의 부채상환연체 차입가계의 평균소득인 42,480,000원보다 유의미하게 낮았다. 자산의 경우 DTA가 0.8초과인 가계의 평균은 121,090,000원이었으나, 이는 40대의 부채상환연체 차입가계의 평균인 311,436,000원보다 낮았다. 한편 부채상환 혹은 생활비 충당을 위하여 담보대출 및 신용대출을 이용한 비율은 DTA가 0.8초과인 가계가 40대의 부채상환연체 차입가계의 평균적 비율보다 높았다.

후속연구자를 위한 제언은 다음과 같다. 중년기의 가계부채를 주제로 머신러닝을 활용하여 분석을 하고자 하는 경우 불균형데이터 문제를 해결하여야 할 필요가 있다. 전술한 바와 같이 불균형상태는 머신러닝 알고리즘의 성능을 저하시킬 수 있으므로, 다양한 샘플링 기법을 적용해보고 개선의 여지가 있는 방법을 적용하여야 할 것이다. 정보손실 및 과적합문제 등 기존의 샘플링 기법들이 가질 수 있는 문제점들을 해소하는 방법으로 ROSE를 활용하였다. 그러나 불균형한 이항자료를 다룰 때 어떤 특정한 분석기법이나 샘플링 기법이 절대적으로 유리하다고 할 수 없으므로 가계부채 상환연체와 같은 데이터불균형 상황에서 분석을 시도하는 연구자라면 다양한 샘플링 기법을 적용해 보는 것이 필요할 것이다.

본 연구의 의의에도 불구하고 연구의 한계점이 존재한다. 이는 본 연구에서 사용한 데이터와 관련이 있다. 본 연구에서 종속변수로 사용한 문항의 경우 2019년 가계금융복지조사가 이루어지기 이전 1년간 응답자 가계에서 원금을 상환하거나 이자를 낼 날짜를 지나친 적이 있는지를 묻는 것으로, 응답은 '예', '아니오'로 나뉜다. 본 연구에서는 '예'라고 응답한 경우를 상환연체를 한 경우라고 보았다. 그러나 '예'라고 응답한 경우 일시적 상환곤란, 실수에 의한 것인지 만성적인 상환 곤란인지를 구분하기 어렵다. 향후에 이러한 부분을 해소할 수 있는 문항을 연구에 포함할 수 있다면 중년기의 가계부채 상환연체 문제를 보다 심도 있게 해석할 수 있는 연구결과가 도출될 것이다.

참고문헌

강덕진(2010). **중년기 남성 위기와 심리적 요인과의 관계성 연구**. 총신대학교 목회신학전문대학원 박사학위논문.

강신기, 조성숙(2013). 중년기의 재무교육 및 퇴직태도가 노후준비인식에 미치는 영향 요인. *디지털융복합연구, 11*(11), 117-132. https://doi.org/10.14400/JDPM.2013.11.11.117

강종구(2017). 가계부채가 소비와 경제성장에 미치는 영향: 유량효과와 저량효과 분석. *경제분석, 23*(2), 28-57. https://dx.doi.org/10.2139/ssrn.2901797

강필성, 이형주, 조성준(2004). 데이터 불균형 문제에서의 SVM 앙상블 기법의 적용. *한국정보과학회 가을 학술발표논문집, 31*(2), 706-708.

김동아, 강수연, 송종우(2015). 불균형 자료에 대한 분류분석.

응용통계연구, *28*(3), 495-509. https://doi.org/10.5351/KJAS.2015.28.3.495

김민철, 김근용, 천현숙, 강미나, 이윤상, 배순석, 송준혁, 김덕례(2016). *생애주기별 주거소비 특성을 반영한 정책 방안 연구*. 안양: 국토연구원.

김명자(1998). **중년기 발달**. 서울: 교문사.

김태완, 김문길, 이주미, 김기태, 김명중, 홍성우(2016). *저소득층 빈곤환경 실태와 자활지원 연계 방안*. 세종: 한국보건사회연구원.

김우영, 김현정(2010). 가계부채의 결정요인 분석. *금융경제연구*. *16*(1), 39-78. https://doi.org/10.17298/kky.2010.16.1.003

김영일, 변동준(2012). *우리나라 가계부채의 주요 현황과 위험도 평가: 차주단위 자료를 중심으로*. 서울: 한국개발연구원.

김영일, 유주희(2013). 가계부채 부실위험에 대한 스트레스 테스트: 가구자료를 중심으로. *경제분석*, *19*(2), 59-95.

김영일(2019). 가계의 종합적 특성을 고려한 연체위험분석. *통계연구*, *24*(4), 48-74. https://doi.org/10.22886/jkos.2019.24.4.48

김주영, 장희순(2016). 가계부채의 결정요인과 변화특성 분석: 한국복지패널자료를 이용한 미시분석을 중심으로. *주거환경*, *14*(1), 221-230.

김지섭(2014). *가계부채 증가원인과 감축방안: 낙과전 기대가 미치는 영향을 중심으로*. 세종: 한국개발연구원.

김지섭(2015a). 연령별 가계부채 분포의 구조적 변화: 우리나라와 미국과의 비교를 중심으로. *부동산포커스*, *84*, 84-91.

김지섭(2015b). *고령층 가계부채의 구조적 취약성*. 세종: 한국개발연구원.

김한용, 이우주(2017). 불균형적인 이항 자료 분석을 위한 샘플링 알고리즘들: 성능비교 및 주의점. *응용통계연구*, *30*(5): 681-690. https://doi.org/10.5351/KJAS.2017.30.5.681

김현정(2010). 우리나라 가계부채의 특징과 민감도 분석. *한국경제포럼*, *3*(3), 77-94.

김혜영, 고효정(1997). 중년기 여성의 우울과 자아정체감에 관한 연구. *여성건강간호학회지*, *3*(2), 129-156.

고기숙(2003). **중년기 남성의 심리적 위기에 관한 연구**. 성균관대학교 일반대학원 박사학위논문.

나라지표(2019). 가계신용 동향. http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1076에서 인출.

대한민국정책브리핑(2018). 2018년 가계금융・복지조사 결과. http://www.korea.kr/news/policyBriefingView.do;JSESSIONID_KOREA=8Pszc6LGsRrh9WvkGvFXyDSzQkpRGm1ykblcx7tt3vpr3Gfkvvvx!93258451!1681798175?newsId=156310018에서 인출.

박대근, 최우주(2015). 가계부채의 결정요인에 대한 패널자료 분석: 주택가격과 대출심사기준을 중심으로. *경제연구*, *33*(1), 75-98.

박명화, 최소라, 신아미, 구철회(2013). 의사결정나무 분석법을 활용한 우울 노인의 특성분석. *대한간호학회지*, *43*(1), 1-10. https://doi.org/10.4040/jkan.2013.43.1.1

박수호, 김흥민, 김범규, 황도현, 엥흐자리갈 운자야, 윤홍주(2018). 불균형 데이터 환경에서 로지스틱 회귀모형을 이용한 Cochlodinium polykrikoides 적조 탐지 기법 연구. *한국전자통신학회 논문지*, *13*(6), 1353-1363. https://doi.org/10.13067/JKIECS.2018.13.6.1353

박인수, 박창수(2018). 일반가계부채 가구와 한계가계부채 가구의 부채 연체의 결정요인 비교 분석. *산업경제연구*, *31*(3), 1113-1133. https://doi.org/10.22558/jieb.2018.06.31.3.1113

박윤태, 노정현(2017). 가구의 소득분위별 가계부채 주관적 상환위험요인에 관한 연구. *한국콘텐츠학회논문지*, *17*(9), 145-158. https://doi.org/10.5392/JKCA.2017.17.09.145

박연우, 허석균(2018). 가계부채 결정요인과 부채부담취약계층의 재무곤경 및 신용위험 분석. *금융정보연구*, *7*(1), 1-31. https://doi.org/10.35214/rfis.7.1.201802.001

박재순, 최의순(1995). 중년여성의 월경상태에 따른 건강증진 생활양식. *여성건강간호학회지*, *1*(2), 234-242.

신기영, 옥선화(1997). 중년기 주부의 가족 역할 수행과 심리적 복지에 관한 연구. *대한가정회지*, *35*(1), 111-128.

신용회복위원회(2019). *신용상담을 위한 재무관리*. 서울: 신용회복위원회.

성현구, 박범기(2018). *세대별 가계부채의 특징 및 시사점*. 서울: 한국은행.

안철휘, 안현철(2018). 효과적인 기업부도 예측모형을 위한 ROSE 표본추출기법의 적용. *한국콘텐츠학회논문지*, *18*(8), 525-535. https://doi.org/10.5392/JKCA.2018.18.08.525

이상길(2018). *국내외 AI 활용 현황과 공공 적용*. 대전: 정보통신기술진흥센터.

이종희(2018). 소득계층별 한국 차입 가계의 부실화 가능성 연구. *한국가족자원경영학회지*, *22*(1), 63-78. https://doi. org/10.22626/jkfrma.2018.22.1.004

이창훈, 강규호, 목정환(2018). 은행권 및 비은행권 가계대출 결정요인 분석과 장단기 예측. *계량경제학보*, *29*(3), 23-57. https://doi.org/10.22812/jetem.2018.29.3.002

오장민, 장병탁(2001). 불균형 데이터의 효과적 학습을 위한 커널 퍼셉트론 부스팅 기법. *한국정보과학회 봄 학술발표논문집*, *28*(1), 304-306.

오준병, 허원창, 이혜민(2019). *기계학습(Machine Learning)을 이용한 인천지역 노동공급이탈 예측모형*. 인천: 한국은행 인천본부.

전산초, 최영희(1985). **노인간호학**. 서울: 수문사.

전승훈, 임병인(2008). 2000년 이후 가계의 자산 및 부채 보유 실태의 변화분석. *재정학연구*, *1*(2), 133-162.

정성훈(2013). **중년 남성의 심리적 위기감과 자기성찰이 심리적 안녕감에 미치는 영향**. 경성대학교 교육대학원 석사학위논문.

정재현(2019). 머신러닝을 활용한 정책설계: 출산 결정요인을 중심으로. *재정포럼, 279*, 12-35.

최종후, 한상태, 강현철, 김은석(1998). **AnswerTree를 이용한 데이터마이닝 의삭결정나무분석**. 서울: SPSS아카데미.

통계청(2019). 2019년 가계금융복지조사 결과. http://kostat.go.kr/portal/korea/kor_nw/1/1/index.board?bmode=read&aSeq=379367에서 인출.

한국은행(2019). *금융안정보고서*. 서울: 한국은행.

한국조사연구학회(2016). *빅데이터 활용 통계생산방법론 연구용역 결과보고서*. 대전: 통계청.

허석균, 박연우, 변동준, 심혜인(2016). *가계부채 현황분석 및 정책적 대응방안*. 서울: 국민경제자문회의지원단.

허준, 김종우(2007). 불균형 데이터 집합에서의 의사결정나무 추론: 종합병원의 건강 보험료 청구 심사 사례. *Information Systems Review*, *9*(1), 45-65.

Ando, A., & Modigliani, F. (1963). The “life cycle” hypothesis of savings: Aggregate implications and tests. *The American Economic Review*, *53*(1), 55-84.

Berk, L. E. (2007). **생애발달**(이옥경, 박영신, 이현진, 김혜리, 정윤경, 김민희 역). (*Development through the lifespan*. 4th ed.). 서울: 시그마프레스.

Borland, D. C. (1978). Research on middle age: An assessment. *The Gerontologist, 18*(4), 379-386. https://doi.org/10.1093/geront/18.4.379

Genkin, M., Dehne F., Navarro, P., & Zhou S. (2019). Machine-learning based spark and hadoop workload classification using container performance patterns. In C. Zheng & J. Zhan (eds.). *Benchmarking, Measuring, and Optimizing* (pp. 118-130). Seattle, WA: Springer. https://doi.org/10.1007/978-3-030-32813-9_11

Levinson, D. J., Darrow, C. N., Klein, E. B., Levinson, M. H., & McKee, B. (1978). **남자가 겪는 인생의 사계절**(김애순 역). (*The seasons of man’s life*). 서울: 이화여자대학교 출판부.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*, 92-122. https://doi.org/10.1007/s10618-012-0295-5

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Addison Wesley.

Weiss, G. M., & Provost, F. (2001). *The effect of class distribution on classifier learning: An empirical study*. New Brunswick, NJ: Rutgers University. https://doi.org/doi:10.7282/t3-vpfw-sf95

☐ Received: April 1. 2020
☐ Revised: June 30. 2020
☐ Accepted: September 6. 2020