

XGBoost를 활용한 국세 체납자의 납부가능성 추정 연구*

정 영 우**

요약

본 연구는 머신러닝의 일종인 XGBoost를 활용하여 국세 체납자의 납부 가능성과 주요 결정을 추정하는데 그 목적이 있다. 분석을 위해 2013년부터 2022년까지 국세청이 캠프에 위탁한 국세 체납자 정보 총 434,294건을 활용했다. 분석 결과, Shapley Value 기준으로 총체납액, 여성여부, 체납자 연령, 전회수신건수, 체납경과기간 등이 국세 체납자의 납부 가능성에 영향을 미칠 것으로 추정되는 주요 변수임을 확인했다. 다음으로 모형의 분류정확도를 확인한 결과, XGBoost의 분류정확도는 로지스틱 회귀모형의 정확도와 비교해 약 10%p 가량 높은 수치를 기록했고, F1 스코어 또한 비교적 우수한 결과를 도출함으로써 모형의 분류성능이 전통적 방식 모형에 비해 우수함을 확인할 수 있었다. 본 연구는 국세 체납자의 실제 데이터를 활용하여 머신러닝의 방법을 통해 납부가능성을 추정했다는 점에서 의의가 있다.

주제분류 : B030500

핵심 주제어 : 국세체납, 기계학습, XGBoost, 위탁징수

I. 서론

대한민국 국민은 헌법 제38조에 따라 납세의 의무를 가진다. 납세는 반대급여 또는 보상없이 중앙정부 및 지방자치단체가 국민에게 일방적으로 부

* 본 연구는 한국자산관리공사 캠프연구소가 수행한 '국세체납 위탁징수채권 회수 가능성 추정 분석' 결과에 기반하여 학술적인 시사점을 도출하기 위해 재가공한 것이다. 본 고의 질적 향상을 위해 유익한 논평을 해주신 익명의 심사자분들께 감사드린다.

** 한국자산관리공사 캠프연구소 차장대우, e-mail: jyoungwoo@kamco.or.kr

과하는 조세의 납부를 의미하는데, 조세는 국가 재정의 주된 재원에 해당하므로 국가운영에 있어서는 필수불가결한 요소이다(정덕주·김광윤, 2004).

그러나 최근 고물가·고금리·고환율의 지속에 따른 경기하방압력 확대로 유동성 확보에 문제가 발생한 소상공인과 자영업자를 중심으로 조세를 체납하는 사례가 증가하면서 국가 재정에 부담으로 작용하고 있다. 조세 체납은 국가세수 확보에 부정적인 영향을 미쳐서 국가재정사업을 위축시킬 뿐만 아니라 성실 납세자와의 형평성 문제를 야기할 수 있으므로(한상국 외, 2007), 국세청은 체납자에 사회적, 경제적인 불이익을 가하는 등 세금징수를 효과적으로 도모하기 위한 다양한 노력을 경주하고 있음에도 국세 체납액은 매년 꾸준히 증가하는 추세를 보인다. 2022년말 기준 국세청 누적 체납액이 100조원을 상회하는 등 체납관리 방안을 효과적으로 개선할 필요가 있다는 목소리가 점차 커지고 있다.

납세의 의무는 국가 운영을 위한 경비 조달을 위해 필요한 모든 국민이 마땅히 그 의무를 준수해야하는 국민의 5대 의무에 해당하지만, 직접적인 대가가 불분명한 상태로 개개인에게 부과된 조세의 특성상 모든 납세자에게 자발적인 납부를 기대할 수 없으므로 공권력에 의한 강제징수권에 해당하는 체납정리제도가 존재한다(서희열 외, 2012). 체납정리는 납세의 고지를 받은 납세자가 납부기한까지 납세를 이행하지 않은 경우 일련의 독촉 절차를 거친 후 재정공권력에 따라 납세자의 재산을 압류하고 이를 환가하는 절차를 말하는데, 체납정리에는 독촉, 재산조사 및 압류, 매각처분 등의 다양한 절차가 필요하므로 국세청의 한정된 인원으로서는 모든 체납정리 절차를 효과적으로 처리하는 것이 사실상 불가능하다.

국세청은 국세체납 징수의 효율성을 도모하기 위해 2013년 국세징수법을 개정하여 체납자별 체납액이 1억원을 상회하거나 체납자 명의의 재산이 없는 등의 사유로 체납액 징수가 어렵다고 판단할 경우 한국자산관리공사(이하 캠코)에 체납자의 주소 또는 거소 확인, 체납액 납부를 촉구하는 안내장 발송 등의 체납액 징수 관련 사실행위를 위탁할 수 있는 근거를 마련했다. 국세청이 제한적인 범위에 한해서만 위탁징수업무를 캠코에 위탁한 것은 국민의 권리 또는 의무와 직접적으로 관련이 없는 단순 사실행위인 행정작용, 공익성보다 능률성이 현저히 요구되는 단순 행정사무 등에 해당하는 경우에만 민간위탁이 가능하다는 규정 때문이다(변혜정, 2016). 이에

캠코는 2013년부터 국세청으로부터 체납처분의 실익이 없는 장기국세 체납에 대한 징수업무를 위탁받아 수행하고 있으며, 2015년 8월부터는 현장밀착 관리체제로 전환하여 징수활동을 강화하고 있다. 그럼에도 캠코로 위탁되는 국세체납 건은 행정청의 징수활동에도 불구하고 체납자 재산이 없거나 폐업 등의 사유로 징수가 곤란하다고 판단하여 징수를 보류한 악성채권에 해당하므로, 2022년 상반기 기준 징수율¹⁾이 1.97%에 머무는 등 위탁징수 관리의 효율성을 극적으로 향상시키는 것이 쉽지 않은 상황이다.

본 연구는 캠코가 국세청으로부터 징수관리를 위탁받은 국세 체납자 정보를 활용하여 체납자별 납부가능성을 추정하고, 그에 미치는 주요 결정요인이 무엇인지 확인하는 실증분석을 진행하고자 한다. 본 분석을 통해 납부가능성이 높은 체납자를 사전에 선별할 수 있다면 선택과 집중을 통한 납부촉구활동의 실질적인 실효성 제고를 기대할 수 있을 것이다. 본 연구에서는 국세 체납자의 납부가능성을 추정하기 위해 머신러닝의 일종인 XGBoost를 분석모형으로 활용하고자 한다. XGBoost는 의사결정나무에 기초하여 추정된 결과를 모형에 순차적으로 학습시키는 부스팅 기법을 사용하므로 타 추정모형과 비교해 분석 데이터에 대한 예측 정확도가 높다는 장점이 있다 (오재영 외, 2019). 본 연구에서도 XGBoost를 활용했을 경우 전통적 방식인 로지스틱회귀모형과 비교해 예측결과가 우수한 것으로 나타났는데 자세한 내용은 후속 지면을 통해 서술하도록 하겠다.

II. 선행연구 검토

조세 체납과 관련한 연구는 주제가 가지는 중요성에도 불구하고 외부에서 자료를 접근하기 어렵다는 문제가 있어 체납자의 성향이나 특성 등을 파악하기 위해 실증적인 분석을 수행한 선행연구를 찾는 것은 쉽지 않다.

먼저 납세자의 의식 정도가 성실납세 등에 미치는 효과 등을 확인할 목적으로 설문조사 등을 활용했던 연구로 전태영·변용환(2013)은 납세의식에 영향을 미치는 주요 요인을 추정하기 위해 각 국가별로 인터뷰를 진행하여

1) 당해 위탁액 대비 당해 징수액으로 계산한 징수율을 의미한다.

수집한 자료에 해당하는 세계가치관조사(World Values Survey) 자료 중에서 1,200명의 한국인 데이터를 활용하여 순서회귀분석을 통해 분석한 결과, 민주주의를 중요시하고 준법성이 강할수록 납세의식이 강화되는 결과도 출되었다고 언급하면서 사회지도층이 솔선수범하는 모습을 보일 경우 납세의식 전반에 긍정적인 영향을 미칠 수 있을 것이라 주장했다. 강민조(2021)는 한국조세재정연구원의 10차 재정패널데이터를 활용하여 Lasso 회귀모형을 통해 성실납세의향에 영향을 미치는 요인을 추정한 결과, 납세자의 주관적 건강상태 등이 탈세를 결정하는 중요한 영향을 미칠 수 있음을 밝히며 납세자의 성실납세를 위해서는 주관적 건강상태와 같은 재무적 위험 회피 성향에 관심을 가질 필요가 있음을 주장한 바 있다. 박명호(2016)는 만 24~64세에 해당하는 성인 남녀 2,299명을 대상으로 수집한 설문조사 결과를 토대로 납세순응행위에 영향을 미치는 심리적 요인을 확인하는 연구를 수행하여 납세자의 성별, 거주지역 등의 사회경제적 요인과 탈세 적발 가능성 등의 과세행정 관련 요인을 통제하더라도 성실납세의향과 같은 심리적 요인은 납세순응과 양(+)의 관계가 있음을 언급하며, 국세청이 국민들이 신뢰할 수 있는 납세환경을 만드는 것이 징수에 도움이 될 것이라 주장한 바 있다. 정재환·이영환(2020)은 국내 시중은행 PB센터의 고객 데이터베이스에 등재된 291명의 개인기업 대표자들을 대상으로 설문조사를 진행하여 규범과 공정성에 대한 인식이 납세순응에 미치는 영향에 대해 분석한 결과, 사회적 규범이나 개인적 규범과 같은 사회적 요인이 납세순응과 정(+)의 관계가 있음을 언급하며 납세자의 순응도를 높이기 위해 정부는 조세제도에 대한 공정성 인지도를 상승시킬 필요가 있음을 덧붙였다.

다음으로 체납 전반에 관하여 분석한 연구로 임주영·현진권(2002)은 국세체납이 발생하는 이유는 신고납부제도로 운영되는 종합소득세와 법인세 등의 세목에서 무신고 및 불성실납부 등이 발생하는 비중이 높고, 조세불복에 따른 체납도 상당했음을 언급하며 성실납부 유도를 위해서는 불성실 신고 및 납부자에 대한 벌칙 강화와 납세를 위한 홍보 및 교육기능을 활성화할 필요가 있음을 주장했다. 김태호·정송이(2015)는 압류·공매 등의 직접적인 강제징수와 관허사업제한, 출국금지 등의 간접방식 등으로 체납세금을 징수하고 있으나 체납담당 인력의 태부족과 징수포상금 부재 등의 문제로 징수율이 높지 않음을 지적하며, 징수율 제고를 위해 체납징수조직의 경

제화와 별도의 포상제도를 마련하는 등 제도적 보완이 필요함을 언급했다. 홍범교(2017)는 현행 체납국세 위탁제도에 대한 개선방향을 검토하는 연구를 수행한 결과, 현행 위탁제도의 문제점으로 국세청이 정해진 예산 범위 내에서 임의로 위탁 대상을 정하고 있으므로 민간위탁기관은 위탁되는 체납건의 유형에 대해 사전에 예측하기 어렵고, 징수금액이 없을 경우에는 수수료를 받지 못하는 구조로 인해 위탁징수효율이 떨어지고 있음을 언급하면서 체납징수 대상의 이원화와 수수료 체계의 개편을 위탁징수제도의 효율성 제고를 위해 검토할 필요가 있음을 주장했다. 한편, 류지민(2017)은 국세 체납액 징수업무를 위탁할 때 캠프가 수행할 수 있는 업무가 극히 제한적인 범위로 한정되고 있어 가시적인 징수실적 개선이 나타나지 않음을 주장하면서 징수효율을 높이기 위해서는 민간위탁 범위를 확대하기 위한 사회적 합의가 도출될 필요가 있으며, 징수조직의 개편 및 확대를 추진하여 자체적인 징수역량 강화 노력을 수행할 필요가 있다고 덧붙였다.

한편, 본 연구에서 분석방법으로 활용하는 XGBoost와 주요 결정요인의 영향력을 간접적으로 확인하는 목적으로 활용하는 SHAP 기법을 활용한 연구가 사회과학 분야에서도 서서히 등장하고 있다. 먼저, Park et al. (2023)은 이익발표 후 잔류현상(PEAD, Post-Earnings -Announcement Drift)을 분석하기 위해 PEAD에 영향을 미칠 것으로 예상되는 89개 변수를 활용하여 XGBoost와 LightGBM 등 2개의 머신러닝 모델을 통해 분석한 결과, LightGBM의 추정결과에 기반하여 구성된 롱-숏 포트폴리오의 수익률이 일반적인 포트폴리오보다 2.1배 높았음을 주장한 바 있다. 이재준 외(2021)은 2019 대졸자직업이동경로조사(GOMS) 자료를 활용하여 XGBoost와 SHAP 통해 근로자 이직 가능성을 추정한 결과, 전통적 방식인 로지스틱회귀모형이나 의사결정나무를 추정모형으로 활용할 때보다 XGBoost 모형에서 높은 예측 정확도를 산출했음을 언급하며, 블랙박스과 같은 머신러닝의 추정결과를 보다 직관적으로 해석할 수 있도록 도움을 주는 SHAP 등의 방법을 병행할 경우 머신러닝 모형의 활용도를 제고할 수 있음을 덧붙였다.

본 연구가 기존의 선행연구와 구별되는 차별점은 실제 체납자 정보를 활용하여 체납자의 특성에 따른 납부가능성을 추정하기 위해 실증분석을 수행했다는 점을 들 수 있다. 만약 본 분석을 통해 체납자의 특성에 따라 체납

액 상환행태가 유의미한 차이가 있음을 확인할 수 있다면 효율적인 징수관리방안을 수립하기 위한 기본토대를 마련할 수 있을 것으로 기대한다. 또한, 머신러닝의 일종인 XGBoost를 활용하여 보다 실체에 가까운 체납액 납부가능성에 미치는 체납자의 특성별 영향력을 추정했다는 것도 본 연구결과의 신뢰성을 높이는 요인으로 작용할 것이라 판단한다.

Ⅲ. 캠프의 국세체납 위탁징수관리 현황

캠프는 2013년부터 국세징수법 제11조에 따라 국세청으로부터 체납액 징수업무를 위탁받아 업무를 수행 중에 있다. 관할 세무서장은 국세징수법 시행령 제4조에 따라 체납자별 체납액이 1억원 이상 또는 체납자 명의의 재산이 없거나 폐업 등으로 체납액 징수가 어렵다고 판단한 경우 캠프에 체납액 징수 관련 사실행위를 위탁할 수 있다. 행정권한의 위임 및 위탁에 관한 규정 제11조에서는 행정기관이 민간에 위탁할 수 있는 범위를 단순 사실행위인 행정작용 등 국민의 권리·의무와 직접 관계되지 않는 것이라 명시하고 있으므로 캠프의 관련업무 범위는 체납자의 주소 또는 거소 확인, 재산조사, 체납액의 납부를 촉구하는 안내문 발송과 전화 또는 방문 상담 등 단순사실행위로 한정된다.

〈표 1〉 국세체납자의 유형별 분포(The status of consignment collection for national tax delinquents)

(단위: 건, 억원, %)

구분	위탁건수(당해)	위탁액(당해, A)	징수액(당해, B)	징수율(B/A)
2017년	212,301	22,569	378	1.67
2018년	201,195	23,852	390	1.64
2019년	166,928	25,390	429	1.69
2020년	185,780	25,057	397	1.58
2021년	256,129	24,688	435	1.76
2022년	199,865	24,184	477	1.97

출처: 한국자산관리공사 홈페이지.
source: KAMCO website.

2022년 12월말 현재 캠프는 총 21차에 걸쳐서 총 20조 3,247억원(체

납자 기준 377,562명²⁾) 규모의 국세 체납액을 수탁했으며, 총 3,087.5억 원을 징수한 바 있다. 캠프에 위탁된 체납국세는 징수가 곤란한 악성채권에 해당하여 징수율이 높은 편은 아니나, 최근 들어 중점관리 대상자 선정을 통한 납부촉구 활동 확대 등 적극적인 징수활동의 전개한 결과 징수액과 징수율 모두 동반 상승하는 추세를 보이고 있다.

IV. 연구 방법

1. XGBoost(eXtreme Gradient Boosting)

본 연구에서 활용하는 XGBoost는 처음부터 완벽한 모형을 구축하는데 시간을 소비하는 것이 아닌 성능이 약한 분류기(weak learner)를 순차적으로 개선하면서 최종적으로 강력한 모형으로 변모시켜 나가는 Gradient Boosting 알고리즘에 기반한다(이용준·선종완, 2020). Gradient Boosting은 이전의 분류결과를 다음 학습에 활용하여 최적화를 진행한다는 구조적 특징으로 인해 모형의 정확성이나 효율성 측면에서 매우 우수한 것으로 평가받고 있으나(Friedman, 2001), 최종 모형 구축을 위해서는 반복적인 계산이 필수적이라는 점에서 데이터 규모가 커질수록 많은 분석시간이 필요하다는 단점이 존재한다(장병탁, 2007).

이러한 Gradient Boosting의 단점을 해결하기 위해 XGBoost는 데이터 연산속도를 향상시킬 목적으로 표본을 특정 변수 기준으로 복수의 하위 집단으로 나누고, 연산제어장치별로 하위집단을 할당하여 동시에 연산을 수행하도록 하는 병렬처리를 채택함으로써 연산시간을 크게 줄였다. 또한, 모형 훈련에 활용했던 데이터의 분류에만 특화되는 과적합(over-fitting) 문제를 방지하기 위해 정규화항(regulation parameter)을 활용함으로써 모형 결과의 신뢰도를 향상시켰다는 것이 특징이다(Chen and Guestrin, 2016). XGBoost는 약한 분류기로 활용할 모형으로 데이터를 나무의 형태로 분할하는 의사결정나무(Decision Tree)를 활용한다.

2) 동일 체납자에 대한 차수별 중복위탁을 제외할 경우 관리 체납자는 총 372,285명이다.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in F \quad (1)$$

식 (1)은 복수 개의 의사결정나무에서 추정된 분류결과를 결합하는 앙상블 학습(Ensemble Learning)의 구조를 수식으로 표현한 것으로, \hat{y}_i 는 i 번째 표본의 예측결과, K 는 의사결정나무의 개수, F 는 활용할 수 있는 모든 의사결정나무의 집합, f_k 는 k 번째 의사결정나무를 의미한다. 위 식을 활용하여 XGBoost의 목적함수를 다음과 같이 표현할 수 있다.

$$L(\Phi) = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ (2)

식 (2)와 같이 XGBoost의 목적함수는 예측값과 실제값의 차이를 나타내는 좌측항의 손실함수(loss function)와 모형의 과적합 문제를 제어하기 위한 우측의 정규화항으로 구성되어 있다. 우측항을 구성하는 요소 중 T 는 의사결정나무의 잎의 개수, w 는 잎의 가중치를 의미하며, γ 와 λ 는 각 요소에 부과되는 규제항(penalty parameter)을 의미한다. XGBoost는 의사결정나무의 잎의 개수를 효과적으로 제어함으로써 모형의 필요 이상으로 복잡해지지 않도록 한다.

한편, 기계학습을 위한 목적함수가 설정된 경우 모형의 분류성능을 극대화하기 위해서는 분석 데이터에 대한 최적의 하이퍼파라미터(hyperparameter) 값을 설정해야만 한다(Hutter et al., 2011; Duarte and Wainer, 2017). 하이퍼파라미터는 모형의 학습환경을 관리하는데 사용하는 외부 구성변수를 의미하는데, 이 변수들은 이론적 근거에 따라 사전적으로 정해지는 것이 아니라 학습하려는 데이터의 성격을 고려하여 경험적으로 설정해야 한다(오재영 외, 2019). 그런데 하이퍼파라미터의 크기변화에 따라 예측성능이 어떻게 달라지는지를 측정하기가 매우 복잡하여 시행착오법에 따라 사후적으로만 확인해야 하며, 데이터 특성에 따라서도 영향력이 크게 변동하게 되므로 연구자가 모든 조합을 고려하여 최적의 파라미터값을

찾아내는 것은 사실상 불가능하다(Bergstra and Bengio, 2012). 따라서 하이퍼파라미터의 최적화를 위해서는 격자 탐색(grid search)과 랜덤 탐색(random search), 베이지안 최적화(bayesian optimization)를 주로 활용하는데, 본 연구에서는 통계적인 방법을 활용하여 최적의 하이퍼파라미터를 탐색하는 베이지안 최적화를 활용하도록 한다. 베이지안 최적화는 임의로 복수 개의 지점을 설정하여 예측 성능을 확인하고, 가우시안 분포(Gaussian distribution)에 따라 아직 탐색되지 않은 부분의 성능 분포를 확률적으로 추정함으로써 최적의 파라미터값을 탐색하려는 것으로, 복잡한 기계학습 알고리즘을 위한 하이퍼파라미터를 찾는 데 적합한 방법으로 알려져 있다(Mockus, 2012). 본 연구에서는 하이퍼파라미터의 최적값 도출을 위해 파라미터별 탐색구간을 고병건 외(2021)이 제안한 구간을 기본으로 하되, 일부 변수의 구간은 저자의 판단에 따라 기존의 범위보다 확대하여 적용했다.³⁾

2. SHAP(SHapley Additive exPlanations)

XGBoost를 포함한 머신러닝 모형은 분석을 위해 선형성 가정 등이 필요하지 않아 고차원 입력변수로 구성된 비선형성 관계를 효과적으로 추정할 수 있다는 장점이 있으나(고병건 외, 2021), 블랙박스과 같이 모형의 내부를 확인할 수 없어 변수 간 인과관계를 구체적으로 확인할 수 없다는 단점이 존재한다(구소정·최필선, 2019).

이에 Lundberg와 Lee(2017)는 결과 값에 영향을 미치는 변수들의 관계를 간접적으로나마 해석하기 위한 수단으로 Shapley Value를 활용하는 방법을 제안하였다. Shapley Value는 게임이론을 바탕으로 특정 변수가 결과 값에 얼마나 기여했는지를 확인하기 위해 여러 변수들을 활용하여 복수의 조합을 만들고, 종속변수에 미치는 기여도를 확인하고자 하는 변수의

3) 본 연구에서는 learning rate는 0.01~0.8, n_estimator는 100~1,000, max depth는 3~10, min child weight는 1~10, gamma는 0.1~3, min subsample는 0.5~1, colsample bytree는 0.5~1을 활용했다. 또한, 베이지안 최적화를 위해 학습표본의 10겹 교차검증을 통한 평균 정확도를 활용했으며, 하이퍼파라미터의 탐색은 총 1,000회 수행했다. 각 하이퍼파라미터의 의미는 지면 관계상 생략하도록 한다.

유무에 따라 달라지는 분류성과의 평균적인 차이를 수치화한 것이다. 다시 말해서, 특정 변수를 제외했을 때의 평균적인 성과와 해당 변수가 포함되었을 때의 평균적인 성과의 차이를 수치적으로 계산하여 변수의 기여도를 측정한다.

$$\Phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(n-|z'|-1)!}{n!} [f_x(z') - f_x(z'_{w/o-i})] \quad (3)$$

식 (3)은 SHAP기법에서 종속변수에 주요한 영향을 미친 변수를 확인하는데 활용하는 Shapley Value를 추정하는 식이다. 여기서 Φ_i 는 i 번째 변수의 Shapley Value를 의미하고, n 은 전체 변수의 수를 의미한다. 또한, $f_x(z')$ 는 2개 이상의 특성값을 가지는 범주형 변수(x)를 해석하기 용이하도록 0과 1의 값을 가지는 이진변수(z)로 변환하여 예측성과를 측정하는 함수를 의미하고, $f_x(z'_{w/o-i})$ 는 i 번째 변수를 제외하고 측정한 예측성과를 의미한다.

SHAP기법은 예측에 활용했던 변수별로 기여도를 계산함으로써 중요도 순으로 변수 간 우선순위를 정할 수 있는데, 경우에 따라 어떠한 특정변수의 Shapley value는 음(-)의 값을 가질 수도 있다. 이러한 경우 해당 변수는 종속변수에 음(-)의 영향을 미쳤다고 해석할 수 있다(안재현, 2020). 이와 같이 SHAP기법은 특정 변수의 영향력을 긍정 또는 부정의 개념을 더하여 수치적으로 표시함으로써 직관적으로 쉽게 이해할 수 있다는 장점이 있으나, 변수별 기여도 측정을 위해 표본마다 개별 연산을 수행하여 평균 수치를 도출해야 하므로 데이터 규모가 매우 클 경우 수치를 산출하기까지 많은 시간이 소요되고, 표본에 따라 특성을 구분하는 변수의 수가 가변적일 경우에는 적용하기 어렵다는 단점이 있다(천예은 외, 2021).

3. 분석표본 및 연구설계

본 연구는 XGBoost를 활용하여 국세체납자의 납부가능성을 사전적으로 추정하는데 목적이 있다. 이러한 목적을 달성하기 위해 2013년부터 2022년까지 캄코가 국세청으로부터 위탁받은 국세체납 정보 총 434,294건을

분석에 활용했다. 분석대상을 2022년까지 위탁된 국세체납 정보로 한정하는 것은 위탁징수관리 이후 체납액 징수를 위한 기간을 최소 6개월 이상 부여하기 위함이다. 본 연구에서 분석표본으로 활용한 국세체납자의 특성별 분포는 다음과 같다.

〈표 2〉 국세체납자의 유형별 분포(Distribution of national tax delinquents by characteristics)

(패널1) 범주형 자료					
변수명	분포				
납부여부 ⁴⁾	납부(10.5%)		미납부(89.5%)		
위탁청	인천지방(16.1%)	충북지방(21.1%)	대전지방(8.6%)	부산지방(14.8%)	
	광주지방(7.2%)	대구지방(7.6%)	서울지방(24.6%)		
체납자구분	개인(74.4%)		법인(25.6%)		
성별	남자(41.7%)	여자(25.3%)	정보없음 ⁵⁾ (33.0%)		
소재지역 ⁶⁾	강원권(2.4%)	충청권(9.4%)	전라권(8.8%)	수도권(57.6%)	경상권(21.8%)
대표세목 ⁷⁾	양도소득세(11.9%)	종합소득세(26.0%)	부가가치세(55.5%)	법인세(4.2%)	기타(2.4%)
기초생활수급 ⁸⁾	No(98.9%)		Yes(1.1%)		
차상위계층 ⁹⁾	No(99.9%)		Yes(0.1%)		

(패널2) 연속형자료 ¹⁰⁾							
변수명	최소값	10%	25%	50%(중위수)	75%	90%	최대값
채무자연령 ¹¹⁾ (세)	0(법인등)	0	0	48	57	65	110
체납건수(건)	1	1	1	2	5	8	30
체납액합계(백만원)	0.00	3.42	8.93	22.04	56.25	120.96	400.00
체납경과개월 ¹²⁾ (개월)	0	21	31	47	102	161	270
전화수신 ¹³⁾ (건)	0	0	0	0	0	0	5
전화송신(건)	0	0	0	0	1	1	10

- 4) 체납자의 납부여부 판정은 국세체납 위탁징수관리를 수행하는 한국자산관리공사 조세채권관리처의 실무담당자와 협의하여 위탁 관리를 시작한 이후에 만원 이상 납부 이력이 존재하는 경우 체납액이 납부된 것으로 간주하였다.
- 5) 법인이나 주민등록번호 오류 등으로 성별을 추정할 수 없는 경우 정보없음으로 구분하였다.
- 6) 체납자의 소재지역 구분 시 제주지역을 전라권에 포함하였다.
- 7) 대표세목은 체납자별로 합산체납액이 가장 큰 세목을 의미하며, 세목별 비중 기준으로 상위 4개 외의 세목은 기타로 합산하여 표시하였다.
- 8) 체납자가 위탁시점 이전에 기초생계, 기초의료, 기초주거, 기초교육 자격을 취득한 경우 기초생활수급자에 해당하는 것으로 간주했다.
- 9) 체납자가 위탁시점 이전에 장애인연금, 차상위장애인, 차상위본인부담경감, 차상위자활, 차상위계층확인, 한부모가정 자격을 취득한 경우 차상위계층에 해당하는 것으로 간주했다.
- 10) 연속형자료의 경우 이상치(outlier)의 영향력이 비정상적으로 커지는 것을 방지하

〈표 2〉에서 확인할 수 있는 바와 같이 본 연구에서 활용한 국세체납 데이터는 체납액을 납부한 체납자와 미납부자의 비중이 상당한 차이가 나는 불균형 데이터에 해당한다. 납부자의 비중이 높지 않은 것을 보면 캠프코의 위탁징수 관리가 효율적으로 운영되지 못한 것으로 보일 수 있으나, 국세청에서 자체적으로는 징수가 어려울 것으로 판단되는 체납 건만 캠프코로 위탁징수관리를 맡긴 것을 고려한다면 단순히 징수실적만으로 해당 제도의 효용성을 논하기는 어렵다(홍범교, 2017).

대표세목 기준으로는 부가가치세가 대표세목인 체납자가 전체 체납자 중 과반 이상을 차지했는데, 부가가치세는 세금을 부담하는 담세자와 납세의무자가 상이한 간접세임에도 불구하고 체납자의 비중이 다른 세목에 비해 높게 나타난 것은 다소 흥미롭다. 거래단계에서 세금이 원천징수되는 부가가치세에서 체납이 발생하는 것은 비교적 재무 건전성이 떨어지는 소상공인, 자영업자 등이 경기침체로 인해 소비자로부터 미리 수취한 부가가치세도 납부하지 못할 만큼 자금사정이 악화되면서 체납이 발생했을 가능성이 높다(김재진, 2013).

본 연구에서는 국세체납자의 납부가능성을 추정하기 위해 결과변수는 국세체납자의 위탁징수관리 이후 체납액 납부여부를, 설명변수로는 〈표 2〉과 같이 국세체납액 납부에 영향을 미칠 것으로 예상되는 총 13개 특성변수를 활용하여 XGBoost 분석을 수행한다. 본 연구는 사실상 국세체납자의 납부가능성을 사전적으로 예측하는데 목적이 있으므로 분석의 정확성을 높이기 위해 설명변수들은 위탁 당시에 입수할 수 있는 변수로만 한정하였다.

본 연구에서는 원활한 XGBoost 분석을 위해 범주형 변수들을 원-핫 인코딩(One-Hot Encoding)하였다. 원-핫 인코딩은 범주형 변수의 속성 개수만큼 더미변수를 생성하고, 분석 표본이 특정 속성에 해당할 경우 1, 그

기 위해 상위 0.5%를 초과하는 수치는 상위 0.5%의 값으로 치환하는 윈저화(winsorization)를 진행했다(연령: 110세, 체납건수: 30건, 체납액합계: 4억원, 체납경과개월: 270개월, 전화수신: 5건, 전화송신: 10건).

- 11) 연령은 위탁일 당시의 체납자의 만 연령을 의미하며, 주민등록번호 오류 등으로 체납자 연령을 계산할 수 없는 경우 법인과 동일하게 0세로 입력하였다(연령효과를 미고려하기 위함).
- 12) 체납경과개월은 최초납부기한부터 위탁일까지의 경과개월수를 의미한다.
- 13) 전화수신은 체납자로부터 연락이 온 경우를 의미하며, 전화송신은 체납자에게 연락을 한 경우를 의미한다.

량지 않을 경우 0의 값을 가지도록 처리한다. 원-핫 인코딩은 표본의 속성을 구별하는데 효과적인 매우 직관적인 방법에 해당하나, 속성의 종류에 비례하여 더미변수의 개수가 증가함에 따라 데이터 처리 난이도가 상승하고 인코딩 이후 단어의 문맥정보가 소멸되는 단점이 있다(유승의, 2021). 그러나 본 연구에서는 표본 특성을 고려하는 용도로만 원-핫 인코딩을 활용하므로 비정형 분석과 같이 단어의 유사도를 고려할 필요가 없으며, 설명변수로 고려하려는 변수의 속성 수만큼 더미변수를 확대하더라도 기계학습을 수행하는 데는 무리가 없는 것으로 판단된다.

본 연구에서는 XGBoost 분석을 위해 R 프로그램의 'xgboost' 라이브러리를 활용했다. 분석결과 최적화를 위한 조기종료 최저회수는 150회로 지정했고, 목표함수의 최적조건은 AUC(Area Under Curve)의 최대화로 지정했다. 또한, 추후 모형의 성능평가를 위해 표본을 8대 2로 무작위 분류하여 표본의 80%는 모형의 학습목적(train)으로, 나머지 20%는 사후 테스트 목적(test)으로 활용하도록 한다.¹⁴⁾

V. 분석 결과

1. 체납자 유형별 납부율 비교

본 절에서는 국세체납자의 납부가능성을 추정하기에 앞서 체납자의 특성에 따라 납부율에 차이가 있는지를 확인하기 위해 체납자를 특성별로 분류한 후 집단별 평균 납부율¹⁵⁾을 비교하도록 한다.

먼저, 위탁청, 성별, 소재지역 등 체납자의 범주형 변수에 따라 분류한 집단별 평균 납부율은 다음과 같다.

먼저, 위탁징수관리를 맡긴 지방국세청에 따라 국세체납자의 평균 납부율은 다소간 차이가 존재했다. 이러한 현상이 발생한 이유는 국세청마다 캠프

14) 훈련 목적 데이터(347,435건, 전체 대비 80.0%), 테스트 목적 데이터(86,859건, 전체 대비 20.0%).

15) 납부율은 전체 체납자 대비 위탁징수관리 이후 만원 이상 납부한 이력이 있는 체납자의 비중을 의미한다.

에 위탁징수관리를 맡긴 체납자의 세금 납부능력이 상이했기 때문에 추정되는데, 전술한 바와 같이 국세징수법 시행령 제4조에 따르면 체납자의 징수가능성을 판단하는 것은 지방국세청 산하의 관할 세무서장에 권한이 있는 것으로 명시되어 있으므로 지방국세청마다 위탁할 체납건을 선별하는 판단 기준이 달라 위탁된 체납자의 조세납부능력 등이 상이했을 가능성이 있다.

〈표 3〉 국세체납자의 범주형 변수별 납부율(The ratios of the possibilities to pay of the national tax delinquents according to their categorical variables)

변수명	납부율				
위탁청	인천지방(11.2%)	충북지방(11.1%)	대전지방(11.1%)	부산지방(10.4%)	
	광주지방(10.3%)	대구지방(10.2%)	서울지방(9.6%)		
체납자구분	개인(13.2%)			법인(2.9%)	
성별	남자(9.8%)		여자(13.7%)	정보없음(2.9%)	
소재지역	강원권 (11.7%)	충청권 (11.3%)	전라권 (10.5%)	수도권 (10.5%)	경상권 (10.2%)
대표세목	양도소득세 (12.2%)	종합소득세 (11.8%)	부가가치세 (10.3%)	법인세(2.3%)	기타(8.2%)
기초생활수급	No(10.6%)			Yes(4.4%)	
차상위계층	No(10.5%)			Yes(7.2%)	

다음으로 법인체납자의 평균 납부율은 개인체납자 대비 현저하게 낮은 2.9%를 기록했다. 법인체납자의 납부율이 상대적으로 낮았던 것은 법인체납자는 경우에 따라 납부를 촉구할 대상이 뚜렷하지 않을 수 있다는 점을 이유로 들 수 있다. 국세기본법 제39조에 따르면 법인이 세금을 납부하지 못할 경우 납세의무 성립일 현재 과점주주가 제2차 납세의무를 지도록 명시되어 있으나, 과점주주로 간주할 수 있는 조건이 주주 또는 그 특수관계자의 소유 주식 합계가 해당 법인의 발행주식 총수의 100분의 50을 초과하고, 그 법인의 경영에 대하여 실질적으로 권리를 행사한 자로 명시되어 있어서 과점주주로 인정받는 조건이 매우 까다롭다. 이와 관련하여 홍범교(2017)는 법인이 폐업할 경우 대부분 납부촉구 대상도 같이 소멸하게 되어 징수율이 낮아질 수 있음을 언급한 바 있다.

다음으로 개인체납자의 성별에 따라 분류한 결과에서는 남성체납자가 여성체납자 대비 납부율이 낮은 것으로 나타났다. 이는 일반적으로 남성의 성실납세의향이 여성에 비해 떨어지기 때문으로 해석할 수 있는데, 박명호

(2011)의 주장에 따르면 일반적으로 여성에 비해 소득수준이 높은 남성의 경우 세부담이 상대적으로 크다고 인식하는 경우가 많아 납세순응행위나 조세제도불평등 인식지수 등이 모두 낮게 나타나는 등 성별에 따른 조세저항이 일정부분 차이가 있을 수 있음을 언급한다.

체납자를 대표세목 기준으로 분류했을 경우에도 세목에 따라 납부율에서 상당한 차이가 존재한다. 개인에게 직접 과세되는 양도소득세와 종합소득세가 대체로 납부율이 높게 나타났으며, 간접세에 해당하는 부가가치세가 다음 위치를 차지했다. 전술한 바와 같이 거래단계에서 원천징수하는 부가가치세를 대납하지 못한 체납자는 미리 원천징수한 자금을 다른 용도로 전용할 만큼 자금사정이 악화되었을 가능성이 높아 앞선 두 가지 세목에 비해 납부율이 낮았을 가능성이 있다. 한편, 법인에 부과되는 법인세가 비교적 낮은 납부율을 기록한 것은 법인체납자의 납부율이 개인체납자에 비해 낮았던 것과 궤를 같이하는 결과로 판단된다.

다음으로 체납자의 연속형 변수에 따라 체납액 납부율에 차이가 있는지를 확인하기 위해 표본을 체납액 납부자와 미납부자 집단으로 구분하여 T-test를 진행하였다. 분석 결과, 연속형 변수로 고려한 6개 변수 모두 체납액 납부자 집단과 미납부자 집단 간 통계적으로 유의미한 차이가 확인될 수 있었다.

〈표 4〉 국세체납자의 연속형 변수별 T-test 결과(The t-test results of the national tax delinquents' continuous variables)

변수명	납부자 평균(A)	미납부자 평균(B)	차이(A-B)	T-value
채무자연령 ¹⁶⁾ (세)	34.757	40.300	-5.544	-40.498
체납건수(건)	3.329	3.858	-0.528	-33.359
체납액합계(백만원)	25.360	48.310	-22.950	-113.540
체납경과개월 ¹⁷⁾ (개월)	62.178	72.446	-10.267	-39.446
전화수신 ¹⁸⁾ (건)	0.211	0.036	0.174	64.532
전화송신(건)	0.516	0.468	0.048	11.003

16) 체납액 납부자 집단과 미납부자 집단 간 평균 연령 차이가 0세로 입력한 법인 체납자의 존재로 왜곡되는 현상이 발생하지 않도록 개인 체납자로만 표본을 재구성하여 연령 차이를 확인했다.

17) 체납경과개월은 최초납부기일부터 위탁일까지의 경과개월수를 의미한다.

18) 전화수신은 체납자로부터 연락이 온 경우를 의미하며, 전화송신은 체납자에게 연락을 한 경우를 의미한다.

먼저, 체납자의 연령을 확인할 경우 체납액 납부집단의 평균 연령은 34.757세로 미납자 집단의 평균연령 대비 5.544세 어린 것으로 나타났다. 일반 금융기관의 부실채권에서도 연령이 어릴수록 채무자의 상환가능성이 높아지는 경향을 보이는데(김형준 외, 2019; 허원·정영우, 2022a), 이는 연령이 낮을수록 신용회복을 통해 정상적인 경제생활을 영위하고자 하는 의지가 컸기 때문으로 사료된다.

체납건수의 경우 납부자 평균 건수는 미납부자의 평균 대비 0.528건 적은 3.329건을 기록했다. 세목에 따라 과세근거나 과세율, 과세목적 등이 상이함에도 불구하고 다중체납이 발생했다는 것은 간접적으로 그만큼 체납자가 자금유동이 어려운 상황임을 의미한다고 볼 수 있다. 허원·정영우(2022b)는 연체가 발생한 계좌수가 많을수록 채무조정을 위한 약정 체결 비중이 낮아지는 등 채무자의 상환의지가 약화됨을 주장한 바 있는데, 다중 체납자의 경우에서도 체납 건수의 증가가 체납자의 납부의지를 꺾는 요인으로 작용했을 가능성이 높다.

체납액 합계액을 기준으로 살펴보면 체납액 납부자 집단의 평균 총체납액 규모는 미납부자 집단과 비교해 작은 것으로 나타났다. 이러한 현상의 원인은 국회예산정책처(2023) 자료에서 찾아볼 수 있는데, 해당 자료에서는 2억원 이상 체납한 고액체납자의 2022년 현금정리 비율은 6.9%에 불과하다고 지적하며 고액체납자의 징수율을 높이기 위한 징수방안을 강구할 필요가 있음을 주장한 바 있다.

다음으로 최초 납부기한일로부터 위탁시점까지의 경과개월수를 의미하는 체납경과개월의 경우 체납액 납부자들의 평균 경과개월은 62.178개월로 미납부자의 평균 경과개월 대비 짧은 것으로 나타났다. 국세를 체납할 경우 최대 60개월 동안 가산세가 매월 추가되며, 체납에 따른 사회적·경제적 불이익이 존재하므로 체납자들은 정상적인 경제활동을 위해 가급적이면 체납액을 납부할 유인이 있다. 그럼에도 체납액을 정리할 수 없는 상태가 지속됨에 따라 체납경과개월이 증가한다는 것은 그만큼 체납자의 자금여건이 녹록치 상황임을 짐작할 수 있다.

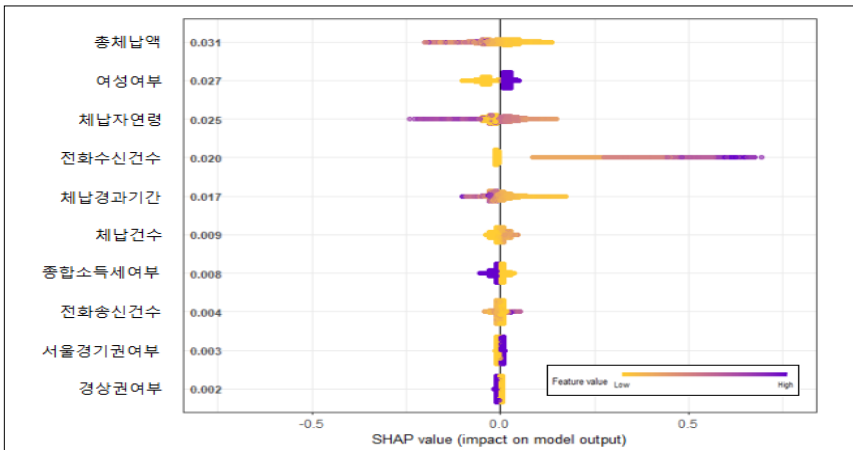
한편, 체납액 납부자 집단의 경우 평균적으로 연락횟수가 미납부자 집단에 비해 많은 것도 흥미롭다. 이는 홍범교(2017)의 연구에서 그 원인을 찾을 수 있는데, 체납자와 연락이 될 경우 체납세액의 존재를 알리는 동시에

체납자 상황에 가장 알맞은 체납세액 납부방법 등을 고지할 수 있으므로 국세체납을 인지하지 못해 납부하지 못하는 단순 체납의 가능성이 크게 줄일 수 있다.

2. XGBoost를 활용한 납부가능성 및 주요 결정요인 추정

본 절에서는 XGBoost를 활용하여 국세체납자별 납부가능성을 추정하고, 체납자의 세금 납부에 영향을 미치는 주요 결정요인을 확인하고자 한다. 본 분석을 위해 사전에 난수생성을 통해 추출한 전체 표본의 80%에 해당하는 훈련목적 데이터를 활용했으며, 종속변수는 위탁징수관리 이후 체납자의 납부여부를, 설명변수는 앞서 분류했던 체납자별 특성변수를 활용했으며, 범주형 변수는 원-핫 인코딩을 통해 속성 값의 개수만큼 생성된 별도의 더미 변수를 활용했다. 한편, 머신러닝 모형의 경우 전술한 바와 같이 일종의 블랙박스의 특성을 보이면서 모형이 변수별 영향력을 어떻게 추정했는지 확인하기 힘든 측면이 있으므로, SHAP (SHapley Additive exPlanations) 기법을 활용하여 국세 체납자의 납부가능성에 상당한 영향을 미치는 것으로 예상되는 주요한 결정요인을 확인하도록 한다. SHAP 기법에 따라 체납자의 납부가능성에 영향을 미치는 변수 중 중요도 순으로 상위 10개 요인의 SHAP plot은 다음과 같다.

〈그림 1〉 XGBoost 추정 기준 중요도 상위 10개 변수의 SHAP plot(The SHAP plot of top10 Important Variables by XGBoost)



SHAP plot은 체납자의 특성변수별로 측정된 Shapley value의 분포를 시각화한 것으로, y축은 체납자의 각 특성을, x축은 각 표본들의 Shapley value를 의미한다. 해당 특성변수의 값이 클 경우 표본별로 추정된 Shapley value는 짙은 색을 띠게 되며, 특성 값이 작아질수록 색이 점차 옅어진다. 또한, 평균 Shapley value가 가장 큰 순서대로 y축의 특성변수가 정렬되어 표시된다.

먼저 예측결과에 가장 큰 영향을 미칠 것으로 추정된 특성변수는 총체납액이다. 소액체납의 경우 세금체납에 따른 불이익을 감수하기 보다는 자금을 조달해서라도 세금을 납부하려는 유인이 크고, 단순 부주의나 일시적인 사유 등으로 납부기한을 놓치는 경우도 종종 발생할 수 있으므로 징수율이 비교적 높게 나타난다(홍범교, 2017). 반면, 체납액이 고액일 경우에는 세금 납부를 위한 자금 확보에 실패하여 체납에 따른 불이익을 감수해야만 하는 경우도 상당할 것으로 사료된다.

두 번째로 체납액 납부여부에 영향을 크게 미치는 것으로 예상되는 특성변수는 체납자의 성별이 여성인지를 구분하는 더미변수였다. <표 3>에서도 확인할 수 있듯이 여성 체납자의 위탁징수관리 이후 납부율은 13.7%로, 남성 체납자의 납부율 대비 3.9%p 높게 나타나면서 성별에 따라 납부율에 유의미한 차이가 있음을 확인할 수 있다. Young(1994)은 여성에 비해 남성이 대체로 높은 조세회피 성향을 보인다고 주장한 바 있으나, 관련 후속 연구에서는 여성의 사회적 진출이 확대되면서 성별과 납세순응도의 관계는 거의 소멸했다고 밝힌 바 있으므로(홍성훈 외, 2017), 성별에 따른 체납액 납부율 차이에 대한 명확한 원인을 파악하기 위해서는 좀 더 면밀한 분석이 필요할 것으로 보인다.

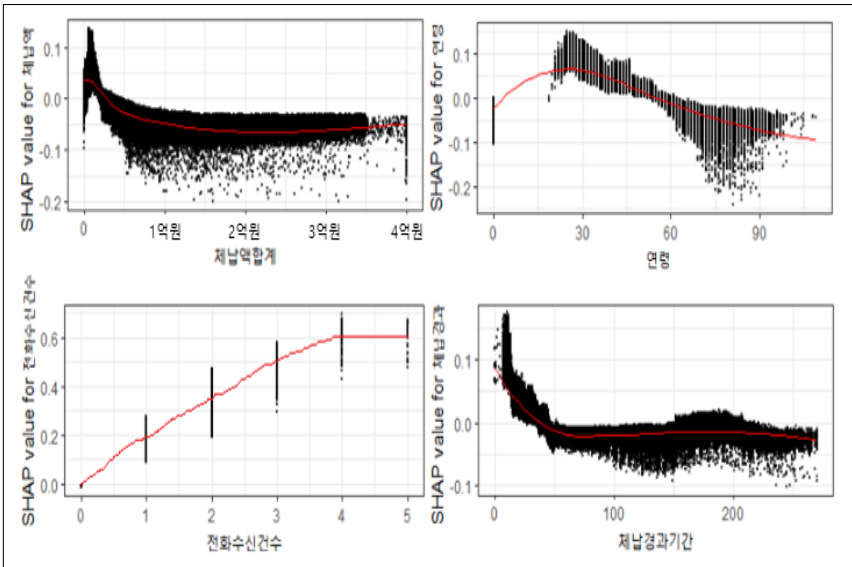
다음으로 체납자의 연령 또한 체납액 납부여부를 예측하는데 상당한 영향을 미칠 것으로 추정된 주요변수였는데, 이는 <표 3>에서 확인한 바와 같이 체납자의 연령이 낮을수록 체납액을 납부하려는 경향을 보였기 때문으로 해석할 수 있다. 이러한 결과는 연령이 낮을수록 정상적 경제주체로 복귀하기 위해 신용회복을 하려는 의지가 커졌던 것이 주된 원인으로 판단된다(김형준 외, 2019). 반면, 이 결과는 연령대가 높아질수록 사회적 제재에 대해 민감하므로 납세의식이 강해진다고 주장한 기존 선행연구(전태영·변용환, 2013; 박명호·심태섭, 2008)들의 주장과는 상반된다는 점에서 흥미로운

측면이 있다.

이 외에도 전화수신건수와 체납경과기간, 체납건수 등도 국세체납자의 납부가능성에 상당한 영향을 미칠 것으로 보이는 중요 결정요인으로 선정되었는데, 각 변수별 영향력에 관한 해석은 앞선 1절에서 서술한 내용으로 대신 하도록 한다.

아래 <그림 2>는 표본별로 특정 연속형 변수와 Shapley value를 매칭하여 점으로 표시한 SHAP dependence plot로, 연속형 변수의 크기 변화에 따라 해당 변수의 영향력이 어떻게 달라지는지를 대략적으로 확인할 수 있다. 체납액 납부가능성에 상당한 영향을 미칠 것으로 추정됐던 총체납액, 체납자연령, 전화수신건수, 체납경과기간 모두 해당 변수의 크기 변화에 따라 영향력이 비선형적으로 변화하는 것을 확인할 수 있다. 이러한 결과는 변수 간 선형적 관계가 존재함을 가정하고 분석하는 로지스틱회귀모형 등의 전통적 방식을 활용해서는 해당 변수들의 영향력을 정확히 추정하지 못할 수 있음을 의미한다.

<그림 2> 중요 연속형 변수의 SHAP dependence plot(The SHAP dependence plot of Important Continuous Variables)



3. 추정모형 예측 결과

본 절에서는 추정모형의 분류성능을 확인하기 위해 앞서 테스트 목적으로 분류한 전체 표본의 20%에 해당하는 총 83,456건의 실제 체납자의 위탁 관리 이후 납부결과와 모형의 예측결과를 아래의 혼동행렬(Confusion Matrix)¹⁹⁾을 통해 확인하였다. 또한, 추정모형의 분류 성능의 객관적 비교를 위한 전통적 방식인 로지스틱 회귀모형을 활용한 분류결과도 함께 표시하였다. 각 모형의 추정 납부가능성을 납부여부로 구분하는 분류기준점(threshold)는 모형별 분류결과의 F1-score가 최대가 되는 지점²⁰⁾으로 정했다.

〈표 5〉 추정모형별 분류 결과(The classification result by estimation model)

XGBoost				로지스틱회귀모형			
		예측결과				예측결과	
		납부	미납부			납부	미납부
실제 결과	납부	3,877(A)	5,232(C)	실제 결과	납부	4,375(A)	4,734(C)
	미납부	10,625(B)	67,125(D)		미납부	19,738(B)	58,012(D)
정확도: 81.7%(=(A+D)/(A+B+C+D)) 민감도: 42.6%(=(A)/(A+C)) 특이도: 86.3%(=(B)/(B+D)) 정밀도: 26.7%(=(A)/(A+B)) AUC: 0.7451 F1-score: 0.3284				정확도: 71.8%(=(A+D)/(A+B+C+D)) 민감도: 48.0%(=(A)/(A+C)) 특이도: 74.6%(=(B)/(B+D)) 정밀도: 18.1%(=(A)/(A+B)) AUC: 0.7042 F1-score: 0.2634			

두 모형의 분류결과를 대조한 결과, XGBoost 모형의 분류 정확도는 전통적 방식인 로지스틱 회귀모형의 정확도보다 9.9%p 높은 81.7%를 기록하여 비교적 성능이 우수한 것을 확인할 수 있었다. 한편, 본 연구에서 활용한 국세체납 데이터는 실제로 체납액을 납부한 체납자와 미납부자의 비중이 상당한 차이가 나는 불균형 데이터에 해당하는데, 데이터의 불균형 정도가 심할수록 단순히 전체 표본의 분류결과를 기반으로 측정한 정확도 지표로는

19) 혼동행렬은 추정모형을 통한 테스트 표본의 분류결과를 행렬의 형태로 시각화한 표를 의미하는데, 두가지 이상의 유형으로 표본을 분류하는 모형의 성능을 평가하는 목적으로 활용한다(Sammut and Webb, 2011).
 20) 본 연구에서 분석을 위해 활용한 R프로그램의 scorecard 라이브러리에 속한 perf_eva 함수를 통해 추정된 결과, XGBoost의 최적 분류기준점은 0.1501, 로지스틱회귀모형은 0.1560으로 나타났다.

모형의 분류성능을 왜곡할 수 있는 문제가 있다(Longadge et al., 2013). 따라서, 이러한 문제를 해결하기 위해 상충관계에 있는 민감도(실제결과가 참인 표본의 결과를 정확하게 예측한 비율)와 정밀도(예측결과가 참인 표본의 결과를 정확하게 예측한 비율)의 조화평균인 F1-score를 모형의 성능을 확인하는 지표로 활용할 수 있다. F1-score에서도 XGBoost모형이 로지스틱회귀모형보다 우수한 성능이 보인 것으로 나타났는데, 이는 XGBoost가 표본의 다수를 차지하는 미납부 체납자의 납부결과를 비교적 잘 분류했기 때문으로 사료된다.

VI. 요약 및 시사점

본 연구는 국세 체납자의 납부가능성과 주요 결정요인을 추정하기 위해 머신러닝의 일종인 XGBoost를 활용하여 2013년부터 2022년까지 국세청이 캠프에 위탁징수업무를 맡긴 구세 체납자 정보 총 434,294건을 분석했다. 또한 임의로 분류한 테스트 목적 데이터를 활용하여 전통적 방식인 로지스틱 회귀모형과 XGBoost의 분류 정확도를 검증하였다. XGBoost는 분석 회차마다 산출되는 결과를 추정모형에 순차적으로 학습시킴으로써 예측 정확도를 향상시키는 부스팅 기법에 기반하여 구축된 머신러닝의 일종이므로 본 연구의 목적인 국세 체납자의 납부가능성을 추정을 위해 활용할 경우 분석결과와 신뢰도를 높여줄 것이라 기대한다.

분석 결과, Shapley value 기준으로 총채납액, 여성여부, 체납자 연령, 전화수신건수, 체납경과기간 등이 국세 체납자의 납부가능성에 영향을 미칠 것으로 추정되는 주요 변수임을 확인했다.

먼저 국세 체납자의 납부가능성에 가장 큰 영향을 미칠 것으로 추정된 총채납액의 경우 대체로 체납액이 증가할수록 납부가능성이 낮아지는 추세를 보였는데, 이는 소액 체납의 경우 체납상태를 지속하면서 불이익을 감수하기 보다는 체납액을 납부하려는 유인이 크게 작용했던 반면, 고액 체납자는 체납액 납부를 위한 자금 확보가 녹록치 않아 어쩔 수 없이 체납상태를 지속하는 경우가 많은 것으로 해석할 수 있다.

다음으로 여성 체납자의 경우 상대적으로 국세 체납액을 납부할 가능성이

높은 것으로 추정됐는데, 성별에 따라 조세회피성향이 상이하다고 주장한 연구와 성별과 조세회피는 무관함을 주장한 연구들이 혼재되어 있는 상황이므로 성별에 따라 체납자별 납부율이 차이가 나는 원인을 제대로 파악하기 위해서는 보다 심층적인 분석을 수행할 필요가 있다.

체납자의 연령도 국세 체납자의 납부가능성에 상당한 영향을 미치는 주요 변수로 판명되었는데, SHAP plot로 확인된 추세를 보면 체납자의 연령이 높아질수록 납부가능성은 점차 감소한다고 유추할 수 있다. 이러한 결과는 연령이 낮을수록 신용회복을 통해 정상적인 경제주체로 회귀하려는 의지가 컸기 때문이라 해석할 수 있다.

전화수신건수 또한 국세 체납자의 납부가능성에 유의미한 영향력을 행사하는 변수임을 확인했는데, 이는 소액 또는 단순체납에 해당할 경우 고지와 독촉 등으로 체납사실을 체납자에게 상기시켜 주는 것만으로도 징수효과가 발생할 수 있음을 의미하므로 정책적으로 시사하는 바가 있다.

그리고 체납경과기간이 길어질수록 국세체납액의 납부가능성이 점차 감소하는 것으로 나타난 결과는 체납 지속에 따른 사회적·경제적 불이익이 존재함에도 장기 체납자들은 세금 납부를 위한 자금확보에 어려움이 있어 체납액을 정리하지 못함에 따라 체납경과개월이 점차 길어지는 결과를 초래한 것으로 해석할 수 있다.

다음으로 모형의 성능을 측정하기 위해 전통적 모형인 로지스틱 회귀모형과 XGBoost의 분류성능을 확인한 결과, XGBoost의 분류정확도는 로지스틱 회귀모형의 정확도와 비교해 약 10%p 가량 높은 수치인 81.7%를 기록하며 국세 체납자의 납부여부를 사전에 예측하는 능력이 비교적 우수했음을 확인할 수 있었다. 또한 전반적인 모형의 성능을 가늠할 수 있는 지표인 F1 스코어 또한 0.3284로, 로지스틱 회귀모형 대비 우수한 결과를 도출했다.

본 연구는 국세 체납자의 특성변수를 활용하여 납부가능성이 높은 체납자를 사전에 분류할 수 있음을 확인했다는 점에서 의의가 있다. 국세 체납징수의 경우 한정된 징수인력으로 다수의 체납자를 관리하는 등 만성적인 인력부족 상태에 놓여 있다. 그러므로 본 연구 결과를 토대로 체납자의 납부가능성에 기반한 관리방안을 수립한다면 보다 효율적인 국세체납 징수관리가 가능할 것이다.

반면, 본 연구에서 국세 체납자의 납부가능성으로 추정하기 위해 활용한

체납자의 특성변수가 연령, 소재지역 등의 일부 변수로 한정된 점은 다소 아쉬운 측면이 있다. 향후 후속연구에서는 체납자의 납부능력을 파악하기 위한 직업정보나 소득정보 등을 함께 고려하여 분석한다면 보다 실체에 가까운 체납자의 납부행태 분석을 수행할 수 있을 것이라 기대한다.

투고 일자: 2023. 11. 8. 심사 및 수정 일자: 2023. 12. 7. 게재 확정 일자: 2023. 12. 14.

◆ 참고문헌 ◆

- 강민조 (2021), “Lasso 회귀모형을 이용한 성실납세의향의 결정요인 연구,” 『조세연구』, 21(4), 7-30.
- Kang, Minjo (2021), “A Study on the Determinants of Voluntary Tax Compliance Tendency using Lasso Regression,” *Journal of Tax Studies*, 21(4), 7-30.
- 고병건 · 신지훈 · 차윤경 (2021), “베이지안 최적화를 통한 저서성 대형무척추동물 종분포모델 개발,” 『상하수도학회지』, 35(4), 259-274.
- Go, Byeonggeon, Jihoon Shin, and Yoonkyung Cha (2021), “Development of Benthic Macroinvertebrate Species Distribution Models using the Bayesian Optimization,” *Journal of Korean Society Water and Wastewater*, 35(5), 259-274.
- 구소정 · 최필선 (2019), “미국 모기지 대출의 인종차별성에 대한 랜덤 포리스트 분석,” 『한국데이터정보과학회지』, 30(2), 261-270.
- Qui, Xiaoting, and Pilsun Choi (2019), “A Study on Discrimination in Mortgage Lending in the United States: A Revisit by Random Forest Method,” *Journal of the Korean Data & Information Society*, 30(2), 261-270.
- 국회예산정책처 (2023), 『한눈에 보는 재정 · 경제 주요 이슈』, 기획보고서.
- National Assembly Budget Office (2023), *Fiscal & Economic Issues at a Glance*, Special Report.
- 김원대 · 정영우 (2020), “채무조정 차주의 기한이익상실에 미치는 결정요인 연구,” 『경영경제연구』, 42(4), 45-74.
- Kim, Won-dae, and Youngwoo Jeong (2020), “The Study on Re-default of th Written-off Debtors and its Determinants,” *Journal of*

- Managment and Economics*, 42(4), 45-74.
- 김재진 (2013), “부가가치세 매입자납부제도의 도입 방안,” 『조세재정Brief』, 2013-7, 1-28.
- Kim, Jae-Jin (2013), “An Introduction to Value Added Tax Purchaser Payment System,” *KIPF Issue Paper*, 2013-7, 1-28.
- 김태호·정송이 (2015), 『지방세 체납관리 효율화 방안』, 한국지방세연구원, 1-214.
- Kim, Taeho, and Jeong, Song-I, *The Efficiency Plan for Local Tax Delinquency Management*, Korea Institute of Local Finance, 1-214.
- 김형준·류두진·조 훈 (2019), “개인부실채권 상환약정체결 및 완제행태 분석,” 『재무연구』, 32(2), 187-219.
- Kim, Hyeongjun, Doojin Ryu, and Hoon Cho (2019), “Empirical Analysis of Consumer Credit Recovery,” *Asian Review of Financial Research*, 32(2), 187-219.
- 류지민 (2017), 『지방자치단체의 체납국세 수탁징수 방안』, 한국지방세연구원, 1-107.
- Ryu, Jimin (2017), *Local Governments' Consignment Collection Plan for Delinquent National Taxes*, Korea Institute of Local Finance, 1-107.
- 박명호 (2016), “납세자의 심리적 요인이 납세순응행위에 미치는 영향,” 『조세재정Brief』, 2016-5, 1-24.
- Park, Myung-Ho, “The Impact on Taxpayers' Psychological Factors on Tax Compliance Attitudes,” *KIPF Issue Paper*, 2016-5, 1-24.
- 박명호·심태섭 (2008), “우리나라 납세자의 납세순응태도에 영향을 미치는 요인,” 『세무와회계저널』, 9(4), 55-84.
- Park, Myung-Ho, and Shim, Tae-Sup, “Determinants of Tax Compliance Attitudes in Korea,” *Journal of Taxation and Accounting*, 9(4), 55-84.
- 변혜정 (2016), 『체납지방세 관리방안에 관한 연구』, 한국지방세연구원, 1-91.
- Byun, Hyejung (2016), *A Study on the Management of Delinquent Local Taxes*, Korea Institute of Local Finance, 1-91.
- 서희열·이준규·서정화 (2012), “공정사회 구현과 국세 체납세액 징수방안,” 『세무학연구』, 29(4), 9-42.
- Suh, Hi-Youl, Juneq Lee, and Junghwa Suh (2012), “Realization of

- Fair Society and Collection of Delinquent Taxes,” *Korean Journal of Taxation Research*, 29(1), 9-42.
- 안재현 (2022), 『XAI 설명 가능한 인공지능, 인공지능을 해부하다』, 위키북스.
- Ahn, Jae-Hyun (2022), *XAI Explainable Artificial Intelligence, Dissecting Artificial Intelligence*, Wikibooks.
- 오재영 · 함도현 · 이용건 · 김기백 (2019), “XGBoost 기법을 이용한 단기 전력 수요 예측 및 하이퍼파라미터 변화에 따른 영향 분석,” 『전기학회논문지』, 68(9), 1073-1078.
- Oh, Jae-Young, Do-Hyeon Ham, Yong-Geon Lee, and Gibak Kim (2019), “Short-term Load Forecasting Using XGBoost and the Analysis of Hyperparameters,” *The Transactions of the Korean Institute of Electrical Engineers*, 68(9), 1073-1078.
- 유승의 (2021), “인공지능과 자연어 처리 기술 동향,” 『주간기술동향』, 제1984호, 2-12.
- Ryu, Seung-Eui (2021), “Artificial Intelligence and Natural Language Processing Technologies Trends,” *Weekly ICT Trends*, 1984, 2-12.
- 이용준 · 신종완 (2020), “XGBoost를 활용한 고속도로 콘크리트 포장 파손 예측,” 『한국건설학회논문집』, 21(6), 46-55.
- Lee, Yong-Jun, and Jong-Wan Sun (2020), “Predicting Highway Concrete Pavement Damage using XGBoost,” *Korean Journal of Construction Engineering and Management*, 21(6), 46-55.
- 이재준 · 이윤희 · 임도현 · 안현철 (2021), “XGBoost와 SHAP 기법을 활용한 근로자 이직 예측에 관한 연구,” 『정보시스템연구』, 30(4), 21-42.
- Lee, Jae Jun, Yu Rin Lee, Do Hyun Lim, and Hyun Chul Ahn (2021), “A Study the Employee Turnover Prediction using XGBoost and SHAP,” *The Journal of Information Systems*, 30(4), 21-42.
- 임주영 · 현진권 (2002), “국세 체납방지를 위한 정책과제,” 『세무학연구』, 19(1), 223-250.
- Lim, Ju-Young, and Jin-Kwon Hyun (2002), “Policy Issues in Tax Arrears,” *Korean Journal of Taxation Research*, 19(1), 223-250.
- 장병탁 (2007), “차세대 기계학습 기술,” 『통신정보 합동학술대회』, 17(11), 96-107.
- Jang, Byoung-Tak (2007), “Next-Generation Machine Learning Technologies,” *Joint Conference on Communications and*

Information, 17(11), 96-107.

전태영·변용환 (2013), “우리나라 납세자의 납세의식에 영향을 미치는 요인,” 『세무학연구』, 30(4), 185-208.

Jun, Tae-Young, and Young-Hwan Byun (2013), “Factors Affecting Tax Morale in Korea,” *Korean Journal of Taxation Research*, 30(4), 185-208.

정덕주·김광윤 (2004), “조세 징수관리의 현황과 개선방안,” 『세무학연구』, 21(4), 39-61.

Jeoung, Duck-Joo, and Kwang-Yoon Kim (2012), “Status Quo of Tax Collection Management and Its Enhancement,” *Korean Journal of Taxation Research*, 21(4), 39-61.

정재환·이영환 (2020), “규범과 공정성에 대한 인식이 납세순응에 미치는 영향에 관한 연구,” 『조세연구』, 20(2), 99-128.

Jung, Jae-Hwan, and Young-Hwan Lee (2020), “A Study on the Effects on Awareness of Norms and Fairness on Taxpayer Compliance,” *Journal of Tax Studies*, 20(2), 99-128.

천예은·김세빈·이자윤·우지환 (2021), “설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구,” 『한국데이터정보과학회지』, 32(2), 283-295.

Chun, Ye-Eun, Se-Bin Kim, Ja-Yun Lee, and Ji-Hwan Woo (2021), “Study on Credit Rating Model using Explainable AI,” *Journal of the Korean Data & Information Society*, 32(2), 283-295.

한상국·안경봉·마정화 (2007), “경매·공매제도 비교를 통한 채납조세 징수업무의 합리화 방안,” 『세법연구』, 7(1), 1-45.

Han, Sang-Kook, Kyung-Bong Ahn, and Jeong-Hwa Ma (2007), “Rationalization Plan for Delinquency Investigation and Collection Work through Comparison of Auction and Public Auction Systems,” *Tax Law and Administration*, 7(1), 1-45.

허 원·정영우 (2022a), “취약 채무자의 채무조정 약정에 미치는 결정요인 추정,” 『국제경제연구』, 28(1), 1-27.

Heo, Won, and Youngwoo Jeong (2022a), “A Study on the Commitment of the Troubled Debt Restructuring and its Determinants,” *Kukje Kyungje Yongu*, 28(1), 1-27.

_____ (2022b), “취약 채무자의 성실상환지속에 관한 실증분석,” 『금융정보연구』, 11(3), 27-60.

Heo, Won, and Youngwoo Jeong (2022b), “The Analysis of the Faithful

- Repayment Period of the Vulnerable Debtors,” *Review of Financial Information Studies*, 11(3), 27-60.
- 홍범교 (2017), “체납국세 위탁징수제도의 개선에 관한 소고,” 『재정포럼』, 258, 6-34.
- Hong, Beom-Kyo (2017), “A Review on the Improvement of the Delinquent National Tax Consignment Collection System,” *Monthly Public Finance Forum*, 258, 6-34.
- 홍성훈 · 노희찬 · 정희선 · 김용수 (2017), 『납세의식 제고를 위한 조세행정 개선 방향 연구』, 조세재정연구원.
- Hong, Sunghoon, Heechun Roh, Heesun Chung, and Yongsoo Kim, *A Study on the Tax Administration Improvement inducing Tax Compliance*, Korea Institute of Public Finance.
- Bergstra, J., and Bengio, Y. (2012), “Random Search for Hyperparameter Optimization,” *Journal of Machine Learning Research*, 13, 281-305.
- Chen, T., and C. Guestrin (2016), “XGBoost: A Scalable Tree Boosting System,” *Knowledge Discovery and Data Mining*, 785-794.
- Duarte, E., and J. Wainer (2017), “Empirical Comparison of Cross-validation and Internal Metrics for Tuning SVM Hyperparameters,” *Pattern Recognition Letters*, 88, 6-11.
- Freidman, J. H. (2001), “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, 29(5), 1189-1232.
- Hutter, F., H. H. Hoos, and K. Leyton-Brown (2011), “Sequential Model-based Optimization for General Algorithm Configuration,” *International Conference on Learning and Intelligent Optimization*, Italy.
- Mockus, J. (2012), “Bayesian Approach to Global Optimization: Theory and Applications,” *Springer Science and Business Media*.
- Longadge, R., S. S. Dongre, and L. Malik (2013), “Class Imbalance Problem in Data Mining: Review,” *International Journal of Computer Science and Network*, 2(1).
- Lundberg, S. M., and S. Lee (2017), “A Unified Approach to Interpreting Model Predictions,” *Conference on Neural Information Processing Systems*, USA.
- Park, Dojoon, Jihoon Jung, and Zoonky Lee (2023), “A Study of

Machine Learning Approaches for Analyzing Post-Earnings-Announcement Drift in Korea,” *Asian Review of Financial Research*, 36(1), 1-30.

Sammut, C., and G. I. Webb (2011), “Encyclopedia of Machine Learning,” *Springer*.

Young, I. M. (1994), “Gender as Seriality: Thinking about Women as a Social Collective,” *Signs*, 19(3), 713-738.

〈부 록〉

〈부표 1〉 로지스틱회귀분석 추정결과(The logistic regression estimation result)

변수명	계수	표준오차	Z-value
α	-2.1431 ***	0.0060	-354.775
성별_woe	0.3660 ***	0.0206	17.729
연령_woe	0.6539 ***	0.0171	38.228
체납건수_woe	-0.9890 ***	0.0442	-22.197
체납액_woe	0.8375 ***	0.0138	60.740
대표세목_woe	-0.0846 **	0.0311	-2.722
체납경과기간_woe	0.7802 ***	0.0239	32.668

Note: 1) ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

2) These variables used in the logistic regression model were converted according to Weight of Evidence(WOE)²¹⁾ method. And also the final variables were selected by a formula-based stepwise model for maximizing AIC.

21) 증거비중(Weight of Evidence)은 범주화된 집단별로 종속변수가 Y인 개체와 N의 값을 가지는 개체 간 상대적 비중을 의미하는 것으로, WOE를 로지스틱모형의 설명 변수로 활용할 경우 로그값을 표시되는 종속변수와 강한 선형성을 띄게 되어 설명력이 높아지는 장점이 있다(김원대·정영우, 2020).

A Study on the Estimation of the Probability of Payment of the National Tax Delinquents using XGBoost

Youngwoo Jeong*

Abstract

The aim of this study is to estimate the likelihood of payment of the national tax delinquents and its determinants using XGBoost, a kind of machine learning method. For doing this, a total of 434,494 cases of national tax delinquents information, which were entrusted to KAMCO for consignment collecting between 2013 to 2022, were used for empirical analysis. As a result of this analysis, it was confirmed that several variables such as total amount of arrears, female or not, delinquent's age were major factors to affect the possibility of payment of national tax delinquents, based on the Shapley Value. And also, it was found that the classification accuracy of XGBoost was about 10%p higher than that of the logistic regression model, and the F1 score also produced relatively excellent result. Therefore, it is certain that XGBoost model is superior to a traditional model for predicting the possibility to pay of the national tax delinquents. The contribution of this paper is to confirm that the possibility of payment can be estimated through machine learning method using actual data of national tax delinquents.

KRF Classification : B030500

**Key Words : national delinquent taxes, machine learning,
XGBoost, consignment collection**

* Deputy General Manager, KAMCO research Institute, Korea Asset Management Corporation, e-mail: jyoungwoo@kamco.or.kr