『공공 빅데이터 청년인턴 활동』 결과보고

체납정보와 회수등급을 이용하여 체·수납 대상자를 중심으로 한 빅데이터 분석 결과와 향후 계획을 보고하는 사항임

□ 분석 개요

○ 분석 기간 : 2020. 10. 5(월) ~ 2020. 12. 31(목)

○ 분석 대상 : 2020년 10월 체납액 고지서·안내문 대비 수납내역

- 10월 기준 체납액 고지서·안내문
  · 총 142,325명 / 306,048건 / 38,606백만원
- 10월 기준 수납결과
  · 총 8,390명 / 16,351건 / 625백만원

○ 추진 사항 : 지방세 체납징수 효율화를 위한 회수등급 고도화

□ 분석 결과

○ 탐색적 데이터 분석을 통한 체납 징수현황 도출

- 회수등급이 높을수록 징수율이 높으며, 납세자별 차이가 발생하였음.

○ 체납정보와 회수등급을 이용한 체·수납 대상자 예측

- 체·수납 대상자 구분에 유의한 결과를 확인하고, 향후 연구방안 제안함.

○ 체·수납 대상자 유형 분류를 통한 특성 파악

- 체납정보와 회수등급을 결합하여 최종 8개의 유형을 확인하였음

□ 향후 계획

○ 악성·생계형 체납대상자를 구분하기 위한 지방세 체납정보와 체납회수등급 갱신의 연구 필요성을 제안

[별첨] 인천광역시 지방세 체납 효율화 빅데이터 분석 결과보고서.pdf

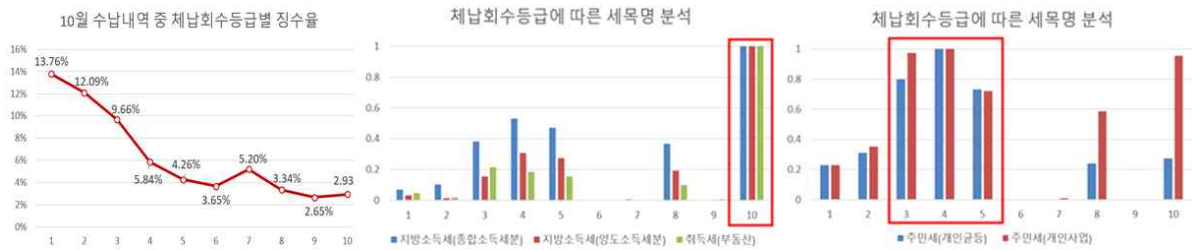
1. 결론

가. 분석결과 해석

1) 탐색적 데이터 분석(EDA)을 통한 징수현황 파악

[그림 1-1] 탐색적 데이터 분석(EDA) 결과

빅데이터 분석 기반인 '체납회수등급'을 10월 발송대비 수납내역을 활용하여 회수등급의 이해도와 활용도를 높이고자 탐색적 데이터 분석(EDA)을 진행하였다. 이는 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 더 잘 이해하고, 잠재적인 문제를 발견하고자 하였고, 그래프나 통계적인 방법으로 체납징수 현황에 직관적으로 바라보는 과정을 진행하였다.



[그림 1-2] 탐색적 데이터 분석(EDA) 결과

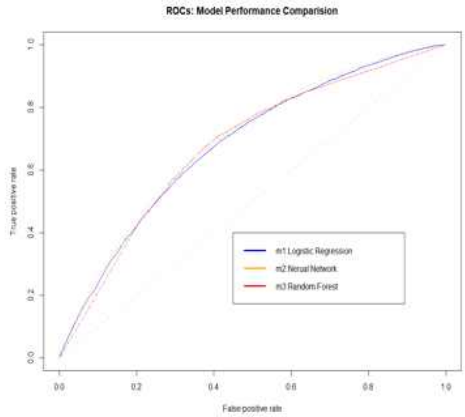
이를 통해 소액·고액 체납자에 대한 특성을 파악하였으며, 체납회수등급에 따라 징수율이 상이한 것을 확인하였다. 아울러 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 파악하지 못한 체납징수 현황에 대한 패턴을 발견하고, 이를 바탕으로 기존의 프로세스를 수정하거나 새로운 정책을 제안할 수 있을 것이다.

2) 앙상블(Ensemble) 예측 모델

모형구분	우·불량 예측(A)	우·불량 실제(B)	정확도(A/B)
Logistic Regression	45,499	71,592	63.55%
Nerual Network	46,097	71,592	64.39%
Random Forest	46,420	71,592	63.85%

[표 1-1] 사전 예측 결과와 실제 납부 형태 비교 결과

3개 모델을 결합한 앙상블 모형의 예측 정확도는 평균 63.93%로 확인하였다. 10월 회수등급 기준의 분석 데이터 셋은 체·수납 대상자 비율이 6:94로, 클래스 불균형 문제가 발생하였다. 비록 SMOTE를 통해 오버샘플링을 적용하였지만, 향후 징수 활동의 데이터를 추가하고 체·수납 대상자의 명확한 특성을 보이는 변수에 대해 조금 더 연구가 진행된다면, 실제 예측 정확도를 높일 수 있을 것이다.



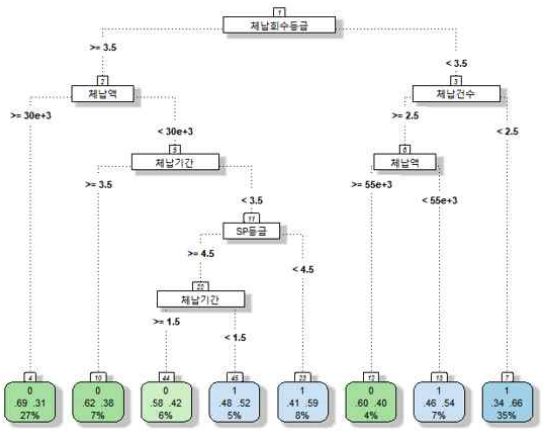
Model	AUC	KS	Gini
Logistic Regression	68.08	27.6	36.16
Nerual Network	69.19	29.07	38.38
Random Forest	67.59	29.81	35.18

Confusion Matrix	Logistic Regression	Nerual Network	Random Forest
Precision	0.6419	0.6549	0.6600
Accuracy	0.6355	0.6439	0.6485
Recall	0.6219	0.6164	0.6201
Specificity	0.6493	0.6717	0.6771
FP_rate	0.3126	0.3099	0.3117
F1 Score	0.6317	0.6351	0.6394
Kappa	0.2712	0.288	0.2971

[그림 1-3] 앙상블(Ensemble) 모델의 결과 비교

결론적으로 개발된 모델 보완의 필요성과 체납에 영향을 미치는 변수가 새로 도출이 필요할 것을 의미한다. 예측 분석에 사용할 기초 데이터 다양화와 분석 모델을 적용할 지자체의 확대를 통해 정확도를 더 높일 수 있도록 해야 할 필요성이 있는 것으로 분석되었다. 이는 데이터의 수집 종류 및 분석대상 데이터의 범위를 확대하여 기존 분석결과와 결합함으로써 분석 범위 확장이 필요하며, 초기 데이터와 추가 데이터의 구조적인 이질성, 데이터 항목 간의 의미와 데이터 값의 이질성 등을 식별하며 지속적으로 모델을 개선해 나갈 필요가 있다.

3) 의사결정 나무(Decision tree) 모형 기반 유형 분류 모델



유형	특성	비율
체납자	체납회수등급이 3.5등급 이상, 체납액이 30,190원 이상인 체납자	27%
	체납회수등급이 3.5등급 이상, 체납액이 30,190원 이하, 체납 기간이 3.5년 이상인 체납자	7%
	체납회수등급이 3.5등급 이상, 체납액이 30,190원 이하, 체납기간이 1.5년 이상 3.5년 이하, SP등급이 4.5등급 이상인 체납자	6%
	체납회수등급이 3.5등급 이하, 체납건수가 2.5건 이상, 체납액이 55,330원 이상인 체납자	4%
수납자	체납회수등급이 3.5등급 이상, 체납액이 30,190원 이하, 체납기간이 1.5년 이하, SP등급이 4.5등급 이상인 수납자	5%
	체납회수등급이 3.5등급 이상, 체납액이 30,190 이하 체납기간이 3.5년 이하이고 SP등급이 4.5등급 이상인 수납자	8%
	체납회수등급이 3.5등급 이하, 체납건수가 2.5건 이상, 체납액이 55,330원 이하인 수납자	7%
	체납회수등급이 3.5등급 이하, 체납건수가 2.5건 이하인 수납자	35%

[그림 1-4] 의사결정 나무(Decision tree) 모형 결과

체납·결손의 내부 데이터와 신용정보융합의 외부 데이터를 기반으로 의사결정(Decision tree) 모형을 활용하여 시각화하였고, 체·수납 대상자 유형 분류를 통해 납세자의 패턴을 확인하였다.

본 모형은 중요한 변수를 선별하여 시각적으로 표현하였기 때문에 모형에 대한 이해가 쉽고, 어떤 입력변수가 목표 변수를 설명하기에 좋은지 파악에 수월하며, 주요 변수의 선정이 용이하다. 따라서 체납 대상자의 유형을 분류하여 이들에게 어떠한 변수를 통해 특성과 패턴이 발생하였는지 확인할 수 있었다.

향후 이러한 모형을 통해 호화로운 생활을 하며 거액의 세금을 내지 않는 악성 체납자의 특성을 파악하는 것에 유용할 것으로 보인다. 이를 위해선 기존 악성 체납 사례를 분석 모형에 적용하여 특성을 비교하고 추출하는 등 관련된 연구의 필요성을 제기한다. 마지막으로 체·수납 대상자의 특성에 알맞은 차량 영치 활동, 압류 활동, 행정제재 등과 같이 분석한다면 디지털 기반 세무징수 우수사례로 발전할 것으로 보인다.

나. 활용방안 및 정책제언

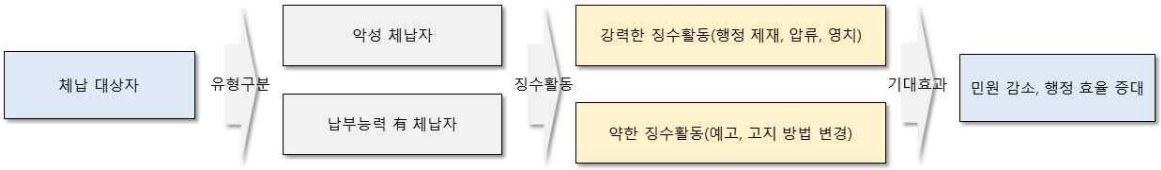
1) 활용 방안

□ 내·외부 데이터 기반 고지 방법 제안

[그림 1-5] 고지방법 차별화

회수등급뿐만 아니라, 연령, 성별, 지역, 체납액, 체납 기간 등 다양한 내·외부 데이터를 통해 차별화된 고지 방법으로 활용할 수 있을 것으로 보인다. 예를 들어, 체납자의 회수등급, 연령, 체납 기간을 동시 고려하여 기존 고지서를 통한 방법이 아닌, SMS를 통한 알림으로 고지 방법을 제안하는 등 체납자의 특성을 고려한 고지 방법으로 효율적인 징수 활동을 할 수 있을 것이다.

□ 체납자 유형 특성에 효율적인 징수 활동



[그림 1-6] 효율적인 징수활동

의사결정(Decision tree) 모형 기반 유형 분류 모델을 통해 체납자 유형 분류를 확인하였다. 체납자 특성의 유형 분류를 통해 악성 체납자는 출국금지, 명단공개 등과 같은 행정제재를 하거나, 신용조회를 통한 압류, 차량 영치 활동 등과 같은 강한 징수 활동을 시행하고, 납부능력이 있는 체납자는 압류예고 및 차량 영치 예고, 고지서 변경 방법으로 약한 징수 활동을 전개할 수 있을 것이다. 이를 통해 체납징수 활동에 민원을 감소시키고 행정의 효율을 높이며, 효율적인 징수 활동을 전개할 수 있을 것으로 기대한다.

□ 체납 모형 검증 작업을 통한 징수 실효성 확인

모형구분	우·불량 예측 (A)	우·불량 실제 (B)	정확도(A/B)
2020년 10월	45,499	71,592	63.93%
2020년 12월	53,132	72,783	73.00%
2021년 2월	56,420	67,739	83.29%

[표 1-2] 사전 예측 결과와 실제 납부 형태 비교 결과 예시

빅데이터 기반인 체납회수등급은 실질적인 징수 활동의 실효성을 확인하기 위해, 체납 대상자를 중심으로 고지서 발송 대비 수납 내역을 우·불량으로 예측할 필요성이 있다. 이는 다양한 내·외부 변수들을 활용한 회수등급 갱신이 실제 징수 활동에 실익 여부를 확인하는 작업이며, 주기적으로 활용하며 점검해야 할 것이다.

2) 정책제언

□ 국세청 등 타 세무기관 부서와의 정보 공유

디지털 기반 행정시대가 도래하면서, 국세 및 지방세에 대한 빅데이터 관련 사업을 동시적으로 진행하는 것을 확인하였다. 빅데이터는 정보가 많으면 많을수록 실익이 커지는 장점을 살려, '정보 교류'에 초점을 맞추고자 한다. 즉, 지역 맞춤형 징수 활동을 위해 '빅데이터를 활용한 체납징수 방법론'을 교류하고, 연구사례 공모전과 같은 선의의 경쟁을 통해 징수 업무의 고도화를 제안하고자 한다.

□ 효율적인 징수 활동을 위한 전문 구성원 필요

지방세 체납징수 업무에 대한 인력이 부족한 상황이면서 공공 빅데이터를 활용한 지방세 효율화 사업을 진행한다면 기존 인력에게 겸업을 요구하게 되고, 이에 따른 업무 과중과 전문성 부족이 발생해 결국 공공 빅데이터의 활성화 저해요인이 발생할 것이다. 따라서 세무 지식과 회수등급, 요인변수에 대한 이해도가 높고 통계 및 데이터에 유능한 전문인력이 필요할 것으로 보인다. 이를 통해 공공 빅데이터 개방 및 활용 활성화를 도모하는 이해당사자에게 실무적 시사점을 줄 것으로 기대한다.

다. 기대효과

1) 비용 절감 및 환수 금액 시뮬레이션 분석

	체납 건에 대한 징수 활동		
체납회수등급	SMS발송	예고	압류
1	①		
2			
3			
4		②	
5			
6			
7			
8			③
9			
10			

① 비용 감소 효과 가능 그룹 — 회수가능성이 높은 그룹이므로 예고 및 압류 대신 SMS 안내로 대체하여 비용 절감

② 조기 환수금액 증대 가능 그룹 — 회수가능성이 중간 그룹이므로 조기에 예고를 한다면 장기 체납 및 처분손실로 연결되지 않고 조기 환수가 가능

③ 적절한 조치가 이루어진 그룹 — 회수가능성이 낮은 등급에 실제로 압류가 이루어져 적절한 조치가 이루어짐

체납 건에 대하여 회수등급이 1~3 등급일 경우에 SMS발송, 4~7등급인 경우에 예고, 8~10등급인 경우 압류하는 정책을 가정함

[그림 1-7] 비용 및 환수 금액 시뮬레이션 정의

단위 : 백만원

		체납건수		체납금액	
체납회수등급		압류X	압류O	압류X	압류O
1					
2	고	105,306건	① 809건	1,768	165
3					
4					
5	중	162,799건	② 2,748건	2,968	507
6					
7					
8					
9	저	74,747건	③ 12,875건	27,876	26,598
10					

압류 건은 모두 환수된다는 것을 가정하였으며, 2020년 10월 기준 과세년월으로부터 60일 이상 경과하고, 체납금액이 10만원 이상인 경우 압류 대상으로 가정하였음.

	구분	기존방식	회수등급방식	비고
①	SMS 발송		40,450 원	
②	예고 안내		1,099,200 원	
③	압류 건수	31,012,000 원	25,750,000 원	건당 SMS 비용 : 50원
	합계	31,012,000 원	26,889,650 원	건당 예고 비용 : 400원
	비용차이	(4,122,350 원)		건당 압류 비용 : 2,000원
	압류건수		15,506 건	
	압류비용		26,889,650 원	
	(조기)환수금액		27,270,206,670 원	

[그림 1-8] 비용 및 환수 금액 시뮬레이션 결과

최종 산출된 회수등급을 토대로 실제 징수 업무에 적용 시, 절감 비용과 조기 환수 금액을 계산하였다. 다음 그림은 시뮬레이션 결과이며, 구분 징수 활동을 통해 약 400만원의 비용 절감 효과를 기대할 수 있는 것으로 나타났다. 아울러 2700만원을 환수비용으로 사용하여 약 270억 원의 체납액 조기 환수 효과를 기대할 수 있을 것으로 보인다.

이처럼 회수등급에 따라 체계적이고 차별화된 징수전략을 수행하여 불필요한 압류와 비용은 줄이고, 무분별한 징수 활동으로 인한 민원은 감소할 것으로 기대한다. 이를 통해 악의적 체납자에 대한 빠른 법적 조치로 회수율을 증대할 것이며, 생계형 체납자에겐 종합적인 사회복지서비스 제공하여 지자체의 고질적인 체납징수의 문제를 해결할 수 있을 것이다.

라. 분석의 한계점 제시

1) 지방세 정보 시스템과 세무 지식을 통한 분석의 필요성

지방세 효율화를 위한 빅데이터 분석에 있어 지방세 체납의 프로세스와 세무 관련된 지식의 깊이가 요구된다. 즉, 체납징수 활동 중 압류와 같은 체납처분, 공매, 행정제재에 대한 전반적인 업무 지식과 지방세 정보 시스템의 이해가 필요한 것을 한계로 파악하였다. 해당 모델을 구축하고 What-How의 중심으로 문제를 해결하고 효과를 파악하기 위해서, 세무 지식과 빅데이터에 대한 통계적인 지식을 겸허한 전문인력이 필요한 것으로 파악하였다. 이를 통해 고도화된 빅데이터 기반 지방세 업무처리 역량을 제고하여 조세정의를 구현하고, 재정 건전화를 도모할 수 있을 것이다.

2) 체납 우·불량 특성에 따른 연구의 필요성

체납에 대한 빅데이터 분석의 등장은 신용정보 회사의 우·불량 정보를 통한 등급과 스코어 모델에서 파생된 것으로 확인하였다. 이는 이용자의 소득과 같은 신용정보를 종합적으로 고려하여 대출, 사기탐지 등 리스크 점수를 중심으로 예측하였다. 체납회수등급은 체납 대상자의 소득을 중심으로 반영하였지만, 소득과 더불어 지역, 체납 대상자의 친인척, 신용정보조회 등 다양한 특성을 반영하고, 체납 특성의 맞춤형 변수를 발굴하여 예측 정확도를 높일 필요성을 발견하였다. 특히, 도메인의 이해를 갖춘 이해관계자가 세목 분석을 통해 특정 체·수납 대상자에게 발생하는 세목을 확인할 필요성이 있다. 이를 통해 구축한 모델의 변수로 추가한다면 조금 더 높은 예측 정확도를 나타낼 것으로 기대할 수 있을 것이다.

3) 수납 대상자의 특성을 파악하기 위한 데이터 필요

개발한 예측모형을 통해 약 63% 우·불량 예측 정확도를 보이며, 통계수치를 통해 모델 안정 지수가 상당히 낮은 것을 확인할 수 있다. 본 모델은 6%의 수납 대상자의 비율로, 클래스 불균형 문제를 오버샘플링으로 해결하여 분석을 진행하였지만, 실제로 체·수납 대상자를 예측에 한계가 있었다. 향후 체·수납 대상자의 데이터가 많아지고, 이들의 패턴을 확인한다면 체·수납 대상자를 예측하는데 수월하고, 예측 정확도의 한계를 개선할 것으로 보인다.

4) 빅데이터 분석 기간의 한계

실제 데이터를 탐색하기 전, 지방세 정보 시스템과 체납에 대한 용어와 프로세스를 이해하는데 상당한 시간이 소요되었다. 따라서 탐색적 데이터 분석(EDA)과정을 통해 문제를 정의하고 접근하는 과정의 기간이 짧아 세부적인 분석의 한계가 있었다. 아울러 정확한 결과를 산출하기 위해 앙상블(Ensemble) 모형과 의사결정(Decision tree) 모형에 대한 임계값과 파라미터를 수정하기 위한 시간이 다소 부족한 것을 한계로 확인하였다. 향후 충분한 분석 기간과 다양한 시도를 통해 유의미한 결과 산출을 할 것으로 기대한다.