

빅데이터 분석 기반 체납징수 현황을 통한

인천광역시 지방세 체납 효율화 분석 결과보고서

내·외부 데이터를 결합한
체·수납 대상자 예측과 유형 분류를 중심으로



김성범

2020. 12. 31.

목 차

1. 분석 개요	4
가. 분석 배경	4
나. 분석 목적	5
다. 분석 주요내용	6
2. 데이터 수집 및 정제	9
가. 활용 데이터 목록	9
나. 데이터 상세 설명	10
1) 체납자 대상 NICE 신용등급	
2) 체납자 대상 발송 내역	
3) 체납자 대상 수납 내역	
4) 인천광역시 시·군구별 체납정리 세무 공무원 현황	
다. 데이터 정제 방안	12
1) 인천광역시모형 NICE 체납회수등급	
2) 10월 발송 내역	
3) 10월 수납 내역	
3. 탐색적 데이터 분석(EDA)	19
가. 인천광역시 연도별 체납현황	19
나. 인천광역시 등급별 현황 및 상관분석	20
다. 10월 발송내역 중 등급별 현황 및 상관분석	23
라. 10월 수납내역 중 등급별 현황 및 상관분석	26
마. 10월 발송 대비 수납 등급별 징수율	29
바. 체납 세목명 분석	30
사. 시·군구별 체납징수 현황	33


4. 모델링 개발 과정	34
가. 분석 데이터 셋 생성	34
1) 수납 및 체납 데이터 병합	
2) 클래스 불균형 해결(SMOTE)	
3) 체납액 이상치 확인	
나. 변수 선택 과정 및 데이터 분할	37
1) Information Value 값	
2) WOE 값 생성 및 확인	
3) Logistic Regression Stepwise	
4) 데이터 정규화 및 분할	
다. 예측 모델 형성	39
1) 로지스틱 회귀(Logistic Regression) 모형	
2) 신경망(Neural Network) 모형	
3) 랜덤포레스트(Random Forest) 모형	
4) 각 모형의 성능 비교	
라. 유형 분류 모델 형성	49
1) 의사결정 나무(Decision tree) 모형	
5. 결론	53
가. 분석결과 해석	53
나. 활용방안 및 정책제언	56
다. 기대효과	58
라. 한계점	59
6. 부록	61
가. 참고문헌	61

1. 분석 개요

가. 분석 배경

1) 인천광역시 지방세 체납징수 활동 현황

□ 체납액 증가에 따른 데이터 기반 정책 마련의 필요성

 인천투데이 | 2019.02.27.

인천시, 지방세 체납액 1700억원 ... 전국 두번째로 많아

인천시의 지방세 이월체납액이 무려 1700억원에 달하는 것으로 나타났다. 이는 서울을 제외한 전국 6개 광역시중 가장 많은 수치다. 27일 인천시에 따르면 2018년...

시도별	지방세	지방세	지방세	지방세	지방세	지방세
서울	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
부산	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
대구	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
대전	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
광주	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
울산	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
경북	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
충북	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
충남	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
전북	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
전남	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
경남	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247
제주	1,017,212	9,715,247	10,111,247	10,111,247	10,111,247	10,111,247

인천시, 지방세 체납액 규모 축소 총력전 나선다 매일일보 | 2019.02.27.

인천시, '지방세 체납징수를 제고 3개년' 추진...202... 아시아투데이 | 2019.02.27.

[그림 1-1] 인천시 지방세 체납액 관련 기사

2018년 기준, 인천시 지방세 체납건수는 923,015건, 체납액은 1,094억 원으로 인구를 고려할 때 전체 시도에서 체납건수는 13위, 체납액은 제주, 경기 다음인 3위로 나타났다. 지방세 체납은 지방세 징수권의 소멸시효 완성 시 세수 손실과 원활한 공공서비스 제공 차질 등 지방 정부 재원확보에 어려움을 발생한다.

이에 따라 인천시 체납특성에 따른 '맞춤형 회수등급'을 통해 효율적 징수 활동 지원 및 불필요한 징수 활동으로 야기되는 민원을 감소하기 위한 데이터 기반의 정책방안 마련이 필요할 것으로 보인다.

□ 지방세 체납·결손 징수를 증대를 위한 빅데이터 분석사업

인천시, 체납정리 빅데이터로 '핀셋 정리'
- 꼭꼭 숨은 악성체납자는 철퇴, 생계형 체납자는 사회복지 연계 -

인천시, 체납자 빅데이터로 추적 관리한다
- 2개 부서 협업으로 내/외부 정보 활용 체납회수 -

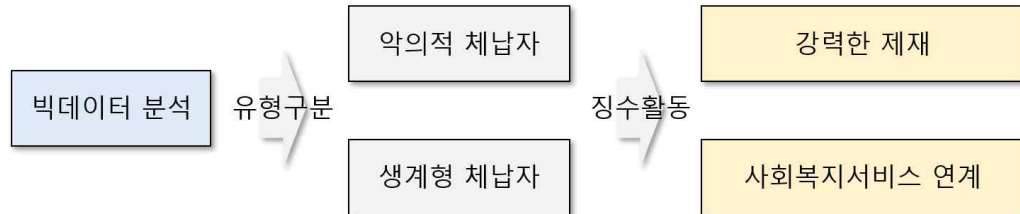
[그림 1-2] 빅데이터를 통한 체납징수 활동 전개 보도자료

인천광역시는 데이터 기반 디지털 행정 실현을 위해, 지방세 체납정보를 빅데이터 분석으로 체납징수 증대 및 정책수립 기반 마련을 목표로 하고 있다. 따라서 체납자 정보(27종)와 신용정보(16종)의 데이터 융합을 통해 체납자의 특성을 파악하는 체납회수 모형을 만들고 예측하는 사업을 진행하는 등, 인천만의 고유의 회수전략에 대한 '빅데이터 분석'으로 신규 업무 발굴에 관심을 쏟고 있다.

나. 분석 목적

1) 데이터 기반 세무행정을 위한 빅데이터 분석

□ 체납자 특성에 따른 차별적인 징수 활동 재개

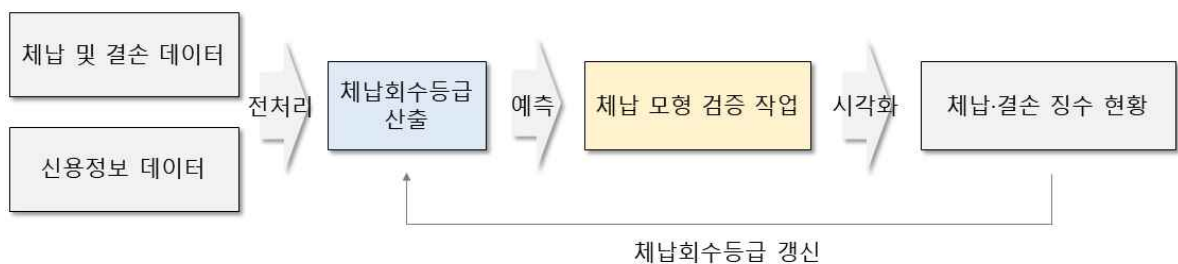


[그림 1-3] 빅데이터 분석을 통한 차별적인 체납징수 활동

그동안 체납징수 담당 공무원은 담세력에 따라 체납자별 맞춤형 징수 활동이 어려웠지만, 빅데이터 분석을 이용하여 정밀한 체납분석이 가능하게 되면 악의적 체납자에게는 보다 신속하고 강력한 제재를, 그리고 생계형 체납자에게는 종합적인 사회복지서비스 제공하는 등 차별적인 징수 활동을 전개하고자 한다.

이러한 빅데이터 기반의 세무행정 구축으로 체납액 징수 가능성 예측과 같은 산출물의 결과로, 체납자 개개인별 효율적인 체납처분이 이루어져 인천시 조세정의 구현에 기여하고자 한다.

□ 체납 징수현황 분석을 통한 체납회수등급 갱신



[그림 1-4] 체납 모형 검증 작업을 통한 체납회수등급 갱신

‘지방세 체납·결손 징수율 증대를 위한 빅데이터 분석사업’은 체납정보 및 결손정보, 신용정보를 활용하여, 기초현황분석, 데이터전처리, 신용정보융합분석 등을 실시하여 체납회수등급을 도출하였다. 이를 이용하여 시·군구별 체납·결손 징수현황을 파악하고, 시각화된 이미지를 제공하여 이해와 활용도를 높이하고자 한다. 아울러 회수등급의 실효성을 확인하기 위해 내부 체납·결손 데이터와 외부 신용정보 데이터로 파생변수를 추가하여 체납 모형 검증 작업을 진행하고, 이러한 과정을 체납회수등급 갱신을 통해 예측의 정확도를 향상하고자 한다.

다. 분석 주요내용

1) 분석 프레임 워크

□ 데이터 분석 프로세스



[그림 1-5] 지방세 효율화 분석 프로세스

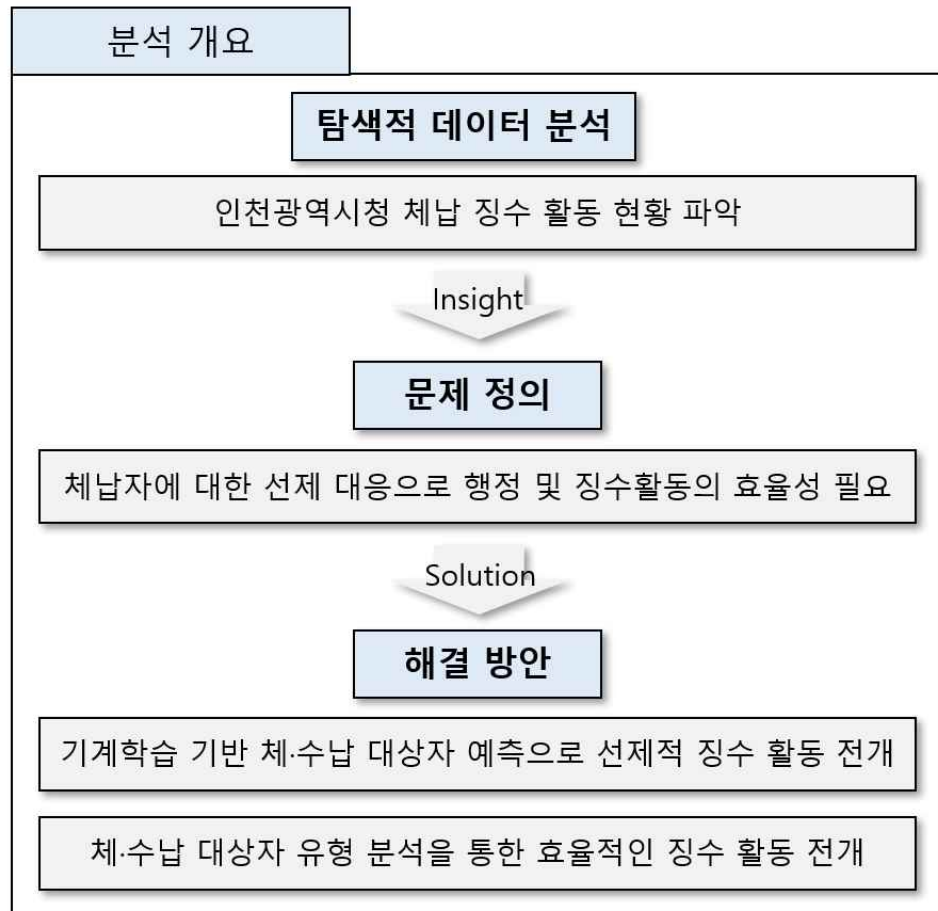
체납 모형 검증 작업을 위한 내부 데이터와 외부 데이터를 수집하여 각각 데이터에 대한 정제과정을 진행하였다. 이를 통해 체납회수등급 기반 인천광역시 체납현황, 10월 발송·수납 대비 징수율, 시·군구별 체납현황을 탐색적 데이터 분석(EDA)의 시각화 과정을 통해 이해관계자들의 이해와 활용도를 높이고, 인사이트(Insight)를 도출하였다.

분석 목적에 따라 데이터를 병합하고, 클래스 불균형 문제를 해결하여 학습과 테스트 데이터로 분할하는 데이터 전처리 단계를 통해 최종적으로 분석 데이터 셋을 형성하였다.

사전에 정의한 체·수납 대상자 예측과 유형 분류의 문제를 해결하기 위해 앙상블(Ensemble) 예측모형을 구축하여 각 통계적 수치를 비교하였고, 의사결정(Decision tree) 나무의 유형 분류 기반 모형을 진행하여 시각화하였다.

결론적으로 체납회수등급을 기반으로 탐색적 데이터 분석(EDA)와 체·수납 대상자 예측과 유형 분류를 통해 어떠한 변수가 체납에 영향을 미쳤는지에 대해 확인하는 실효성을 확인하였다. 아울러 향후 체납회수등급 갱신에 영향을 줄 새로운 변수를 모색하였으며, 모형개발을 통한 해석과 향후 연구 방향을 제시하고, 활용방안과 정책제언을 제시하였으며, 이에 대한 기대효과를 정량적으로 제시하였다.

□ 분석 개요

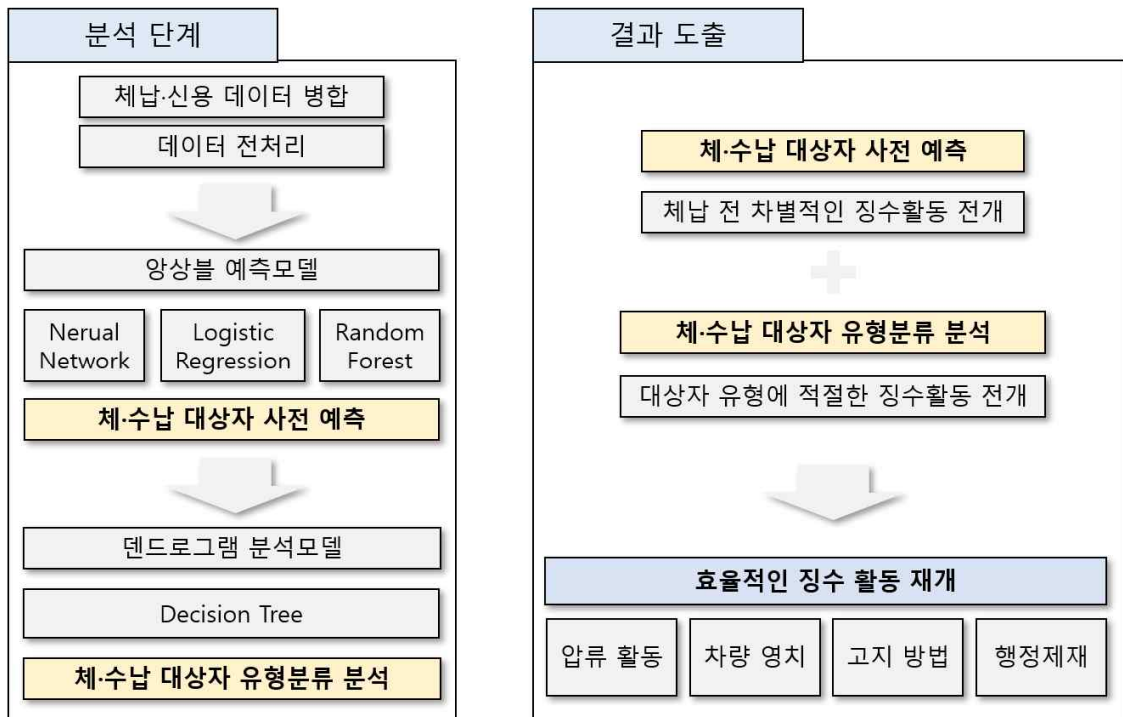


[그림 1-6] 체·수납 대상자 예측을 통한 선제 대응 프레임워크

인천광역시 지방세 과세정보 27종과 민간 신용정보 16종을 결합하여 완성한 맞춤형 체납회수등급을 이용하여 탐색적 데이터 분석(EDA)을 진행하였다. 이를 통해 인천광역시의 체납징수 활동 현황을 파악하고, 인사이트(Insight)를 도출하여 문제를 정의하고자 하였다.

지방세 체납문제는 국가 재정에 심각한 문제를 초래하기도 하며, 민원 문제로 인한 불필요한 행정 인력배치로 징수 활동에 어려움을 겪고 있다. 따라서 지방세 체납문제에 대한 개선의 여지가 필요하여 체납자에 대한 선제 대응으로 행정 및 징수 활동의 효율성이 필요하다는 문제를 정의하였고, 이를 빅데이터 분석 기법 중 ‘기계학습을 기반으로 체·수납 대상자 예측으로 선제적 징수 활동’을 전개하고, ‘체·수납 대상자 유형 분석을 통한 효율적인 징수 활동’으로 해결하고자 한다.

□ 분석 단계 및 결과 도출



[그림 1-7] 체·수납 대상자 예측을 통한 선제 대응 프레임워크

인천시의 내부 데이터인 부과·체납·결손 테이블과 외부 데이터인 민간 신용 데이터를 병합하고, 전처리를 통해 분석 데이터 셋을 형성하였다. 이를 3개의 기계학습 모델을 결합한 앙상블(Ensemble) 모델을 구축하여 체·수납 대상자를 사전에 예측하고자 한다. 아울러 덴드로그램 분석을 통해 체·수납 대상자의 유형을 파악하고, 납세자별 어떠한 특성을 보이는지 확인하고자 하였다.

체·수납 대상자를 사전 예측을 통해 납세자가 체납하기 전 선제적으로 징수 활동을 전개하여 체납을 방지하고, 체·수납 대상자의 유형 분류 분석을 통해 어떠한 특성을 보이는지 확인하였다.

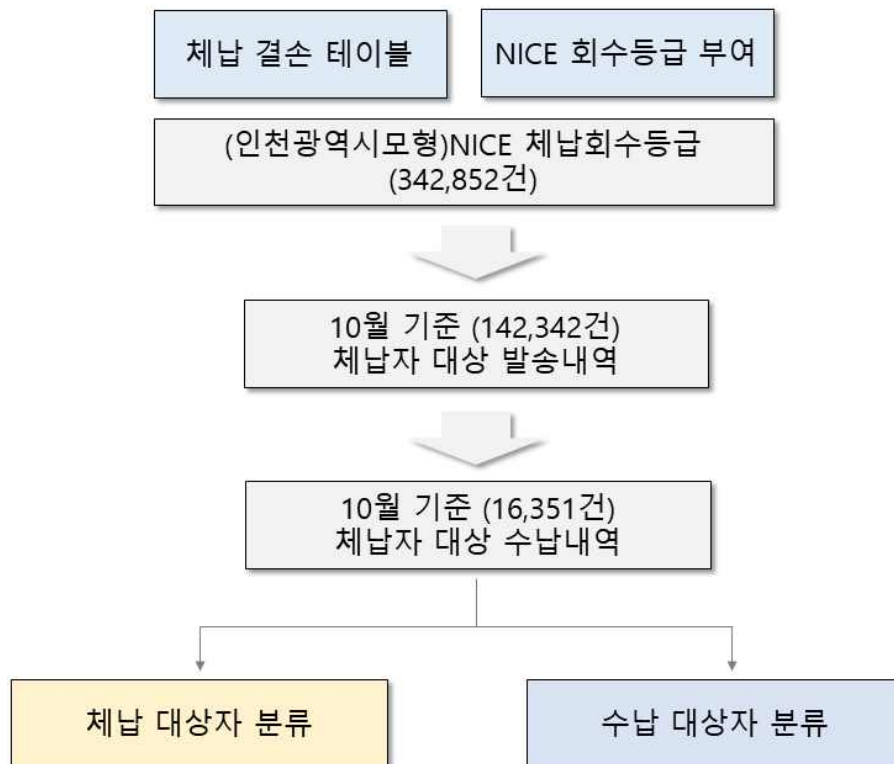
이를 통해 체납자의 특성을 고려한 급여, 매출채권 등의 압류 활동, 차량 영치 및 예고 여부, 납세자별 고지 방법의 차별화, 행정제재를 통해 효율적인 징수 활동을 제안하고자 한다.

2. 데이터 수집 및 정제

가. 활용 데이터 목록

활용 데이터		구분	중요도	생성주기	특성
체납자 대상 NICE 신용등급	(인천광역시모형)NICE체납회수등급	정형/외부	필수	년단위	
체납자 대상 발송내역	10월 고지서-최종	정형/내부	필수	월단위	체납자
	10월 안내문-최종	정형/내부	필수	월단위	
체납자 대상 수납내역	10월 수납현황	정형/내부	필수	월단위	수납자
인천광역시 시·군구별 체납정리 세무 공무원 현황		정형/내부	중간	년단위	조직도

[표 2-1] 분석데이터 목록



[그림 2-1] 활용 데이터를 통한 분석 프로세스

(인천광역시모형)NICE 체납회수등급을 통해 체납자 대상 발송·수납내역에 신용등급을 부여하였다. 이를 통해 10월 기준 체납자 대상 발송내역에서 수납한 대상자의 신용정보를 확인하고, ‘체납 대상자’와 ‘수납 대상자’를 분류하였다.

나. 데이터 상세 설명

1) 체납자 대상 NICE 신용등급

□ (인천광역시모형) NICE 체납회수등급(342,852건)

■ 체납자정보(납세자번호, 시군구코드, 과세년월, 세목명, 부과금액, 가산금, 체납액)

대장	시군구코드	과세년월	과세구분	과세번호	세목코드	세목명	납세자번호_주민등록번호	납세자명	부과금액	가산금	총체납액_또는_결손금액
체납	28110	201501	고지	100	101001	귀취특세(부동산)	123456-1234567	홍길동	12000	1000	130000
체납	28110	201501	고지	200	101501	취특세(부동산)	123456-1234567	홍길동	12000	1000	130000
체납	28110	201501	고지	300	101501	취특세(부동산)	123456-1234567	홍길동	12000	1000	130000
체납	28110	201501	고지	400	101501	취특세(부동산)	123456-1234567	홍길동	12000	1000	130000
체납	28110	201501	고지	500	101501	취특세(부동산)	123456-1234567	홍길동	12000	1000	130000

[그림 2-2] (인천광역시모형) NICE 체납회수등급 - 체납자 정보

■ NICE등급(CB점수, 소득추정등급, 소득추정금액, 체납회수등급, SP등급)

CB점수	소득추정등급	소득추정금액_단위_만원	체납회수등급	회수율_우량률	SP등급
350	8	3000	10	0%	2
243	9	2000	10	0%	3
526	8	2500	1	40%	4
350	11	2700	2	26%	7
753	14	1200	4	0%	4

[그림 2-3] (인천광역시모형) NICE 체납회수등급 - NICE 신용 정보

변수명	설명	구성
CB점수	전체 경제활동인구에 대한 종사종인 신용도를 측정하여 신용점수로 계량화 한 지표	0점(불량) ~ 1,000점(우량)
소득추정지수	실소득, 국민연금·건강보험 납부내역, 신용정보, 직장정보, 부동산정보를 통합 반영하여 경상소득(근로소득/사업 및 부업소득)을 통계적으로 추정	1등급(8,000만원 이상) ~ 15등급(1,000만원 이상)
체납회수등급	체납 발생일 이후 일정기간 내 회수 가능성을 다양한 변수들을 고려하여 하나의 등급으로 예측한 지표	1등급 ~ 10등급
SP등급	SP등급은 카드/대부업대출 등을 이용하는 고객을 대상으로 신용등급을 세분화하여 Sub prime등급을 부여	1등급 ~ 15등급

■ 카드개설정보(기관명1, 기관명2, 기관명3)

기관명1	카드종류1	개설일자1	기관명2	카드종류2	개설일자2	기관명3	카드종류3	개설일자3
기업은행카드	카드(체크카드)	2018.11.02	새마을금고	카드(체크카드)	2014.04.25	우정사업본부	카드(체크카드)	2013.04.01
농협카드	카드(체크카드)	2018.08.01	신협중앙회본부	카드(체크카드)	2016.10.11	기업은행카드	카드(체크카드)	2013.05.22
신협중앙회본부	카드(체크카드)	2015.02.02	농협카드	카드(체크카드)	2013.06.21	롯데카드	카드(일반)	2012.11.26
우정사업본부	카드(체크카드)	2019.04.02						
농협카드	카드(체크카드)	2012.02.29						

[그림 2-4] (인천광역시모형) NICE 체납회수등급 - 카드개설정보

2) 체납자 대상 발송 내역

□ 10월 고지서-최종

■ 고지서발송대장(납세자번호, 세목명, 금액)

■ 20년 10월 기준 : 56,554건

세목명	과세대상	발송지	우편번호	금액	납세자구분	납세자번호	납세자명
주민세(개인균등)	2019년 개인균등분 정기분 (홍길동)	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	12345	12870	납세자	1234561234567	홍길동
주민세(개인균등)	2017년 개인균등분 정기분 (홍길동)	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	23456	12870	납세자	1234561234567	홍길동
주민세(개인균등)	2017년 개인균등분 정기분 (홍길동)	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	34567	12870	납세자	1234561234567	홍길동
주민세(개인균등)	2019년 개인균등분 정기분 (홍길동)	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	45657	12870	납세자	1234561234567	홍길동
주민세(개인균등)	2015년 개인균등분 정기분 (홍길동)	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	56789	12870	납세자	1234561234567	홍길동

[그림 2-5] 체납자 대상 발송내역 - 10월 고지서 샘플 데이터

□ 10월 안내문-최종

■ 안내문발송대장(납세자번호, 세목명, 체납합계)

■ 20년 10월 기준 : 85,840건

납세자번호	성명	우편번호	현주소	과세년월	세목명	체납액	체납합계
1234560000000	홍길동	123-12	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	2006년10월	등록세(차량)	1,210,890	1,771,330
1234560000000	홍길동	123-12	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	2019년12월	주민세(개인균등)	12,870	25,740
1234560000000	홍길동	123-12	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	2015년08월	주민세(법인균등)	96,560	193,120
1234560000000	홍길동	123-12	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	2014년01월	취득세(부동산)	620,450	824,940
1234560000000	홍길동	123-12	인천광역시 부평구 평천로123번길 12 인천아파트 101동 101호	2017년08월	주민세(법인균등)	96,560	115,490

[그림 2-6] 체납자 대상 발송내역 - 10월 안내문 샘플 데이터

3) 체납자 대상 수납 내역

□ 10월 수납현황

■ 수납대장(주민번호, 수납금액)

■ 20년 10월 기준 : 16,351건

자치단체	세목	은행수납일	회계일	주민번호	N10월 발송	이름	수납금액
28237	104001	20201102	20201102	1234560000000	고지서	홍길동	12870
28245	140002	20201020	20201020	1234560000000	고지서	홍길동	83820
28185	104003	20201027	20201027	1234560000000	고지서	홍길동	96020
28245	104003	20201031	20201102	1234560000000	안내문	홍길동	64370
28245	140003	20201031	20201102	1234560000000	안내문	홍길동	19860

[그림 2-7] 체납자 대상 수납내역 - 10월 수납현황 샘플 데이터

4) 인천광역시 시·군구별 체납정리 세무 공무원 현황

□ 12월 기준 세무공무원 조직도 현황

■ 세무공무원 조직도(시군구명, 인원)

■ 20년 12월 기준 : 123명

시군구명	본청	중구	동구	미추홀구	연수구	남동구	부평구	계양구	서구	강화군	옹진군	합계	비고
인원	42	5	5	16	8	14	9	9	10	3	2	123	(실무직원기준, 임기제 포함)

[그림 2-8] 시·군구별 체납정리 세무 공무원 현황

다. 데이터 정제 방안

1) 인천광역시모형 NICE 체납회수등급

□ 1. 데이터 사용 변수 지정

시군구코드	과세년월	세목명	납세자번호	부과금액	가산금	체납액	CB점수	소득추정등급	소득추정금액	체납회수등급	회수율	SP등급	카드1	카드2	카드3
1	28110	201504	구)취득세(부동산)	282160	743610	1025770	350	9	3200	10	0%	4	기업은행카드	새마을금고	우정사업본부
2	28110	201501	취득세(부동산)	292600	517860	810460	243	10	2700	10	0%	5	농협카드	신협중앙회본부	기업은행카드
3	28110	201504	취득세(부동산)	95380	2840	98220	526	9	3000	4	40%	5	신협중앙회본부	농협카드	롯데카드
4	28110	201507	취득세(부동산)	37810	1120	38930	350	12	2300	5	26%	5	우정사업본부	NA	NA
5	28110	201507	취득세(부동산)	239680	7180	246860	753	14	1500	10	0%	2	농협카드	NA	NA

[그림 2-9] (인천광역시모형)NICE체납회수등급

총 60개의 변수 중 최종적으로 16개(시군구코드, 과세년월, 세목명, 납세자번호, 부과금액, 가산금, 체납액, CB점수, 소득추정등급, 소득추정금액, 체납회수등급, 회수율, SP등급, 카드1, 카드2, 카드3)의 변수 사용하였다.

납세자번호	세목명	체납액
1	주민세(개인균등)	12870
2	주민세(개인균등)	12870
3	주민세(개인균등)	12870
4	주민세(개인균등)	12870
5	주민세(개인균등)	12870

시군구코드	세목명	납세자번호	수납금액
1	28237	104001	12870
2	28245	140002	83820
3	28185	104003	96020
4	28245	104003	64370
5	28245	140003	19860

[그림 2-10] 10월 고지서 & 안내문 -최종

[그림 2-11] 10월 수납현황

고지서 & 안내문 25개 변수 중 3개(납세자번호, 세목명, 체납액) 변수 사용하였으며, 수납내역 8개 변수 중 4개(시군구코드, 세목명, 납세자번호, 수납금액) 변수 사용하였다.

□ 2. 세목명 정제 인천광역시청 NICE 신용등급 정제

[1] "구)취득세(부동산)"	"취득세(부동산)"	"취득세(차량)"	"취득세(이륜차량)"
[5] "취득세(기계장비)"	"취득세(선박)"	"취득세(기타)"	"주민세(개인균등)"
[9] "주민세(개인사업)"	"주민세(법인균등)"	"지역자원시설세(특자)"	"지방소득세(종합소득세분)"
[13] "지방소득세(양도소득세분)"	"지방소득세(법인세분)"	"지방소득세(특별징수분)"	"등록면허세(면허)"
[17] "등록면허세(등록)"			
[1] "취득세(부동산)"	"취득세(차량)"	"취득세(이륜차량)"	"취득세(기계장비)"
[5] "취득세(선박)"	"취득세(기타)"	"주민세(개인균등)"	"주민세(개인사업)"
[9] "주민세(법인균등)"	"지역자원시설세(특자)"	"지방소득세(종합소득세분)"	"지방소득세(양도소득세분)"
[13] "지방소득세(법인세분)"	"지방소득세(특별징수분)"	"등록면허세(면허)"	"등록면허세(등록)"

[그림 2-12] 세목명 통일을 위한 변경

추후 데이터 병합과 세목명 분석을 위해 '구)취득세(부동산)'의 세목명을 '취득세(부동산)'으로 변경하였다.

□ 3. 카드개수 파생변수 생성

카드1	카드2	카드3	카드 개수
기업은행카드	새마을금고	우정사업본부	3
농협카드	신협중앙회본부	기업은행카드	3
신협중앙회본부	농협카드	롯데카드	3
우정사업본부	NA	NA	1
농협카드	NA	NA	1

[그림 2-13] 카드개수 파생변수 생성

사용하는 카드 개수에 따라 체납에 영향을 미칠 것을 가설로 세워, 카드 유무를 통해 카드개수의 파생변수를 생성하였다.

□ 4. 납세자번호별 인당 체납액 정제

	납세자번호	부과금액	가산금	체납 액
1		12500	370	12870
2		12500	370	12870
3		16020	480	16500
4		12500	370	12870
5		93750	2810	96560

[그림 2-14] 납세자번호별 부과금액/가산금/체납액

체납 데이터는 ‘체납자 명수’가 아닌 ‘체납자 건수’이기 때문에 중복된 납세자번호가 존재한다. 따라서 342,852건의 체납 데이터 중 납세자 번호별 기준으로 부과금액, 가산금, 체납액을 합산하여 138,020명으로 정제하였다.

□ 5. 세목 건수 파생변수 생성

	납세자번호	세목명	세목 명건수
1		취득세(부동산), 주민세(개인균등)	2
2		취득세(부동산)	1
3		취득세(부동산), 지방소득세(양도소득세분)	2
4		취득세(부동산)	1
5		취득세(부동산)	1

[그림 2-15] 세목명 건수 파생변수

체납한 세목명의 개수에 따라 체납에 영향을 미칠 것이라는 가설로 세우고, 세목건수의 파생변수를 생성하였다.

□ 6. 인당 체납건수 파생변수 생성

	납세자번호	세목명	체납건수
1		취득세(부동산), 주민세(개인균등), 주민세(개인균등), 주민세...	4
2		취득세(부동산)	1
3		취득세(부동산), 지방소득세(양도소득세분)	2
4		취득세(부동산)	1
5		취득세(부동산)	1

[그림 2-16] 납세자번호 당 체납 건수

납세자의 체납 건수가 체납에 영향을 미칠 것이라는 가설을 세우고, 납세자번호 당 체납 건수의 파생변수를 생성하였다.

□ 7. 생성한 파생변수와 NICE 신용정보 병합

납세자번호	부과금액	가산금	체납액	세목명	세목명건수	체납건수	시군구코드	과세년월	CB점수	소득추정등급	소득추정금액	체납회수등급	회수율	SP등급	카드개수
1	12500	370	12870	주민세(개인균등)	1	1	28177	201508	0	0	0	5	26%	0	0
2	12500	370	12870	주민세(개인균등)	1	1	28237	201508	0	0	0	5	26%	0	1
3	16020	480	16500	지방소득세(양도소득세분)	1	1	28177	201602	0	0	0	5	26%	0	0
4	12500	370	12870	주민세(개인균등)	1	1	28200	201912	0	0	0	5	26%	0	0
5	93750	2810	96560	주민세(법인균등)	1	1	28200	201808	0	0	0	5	26%	0	0

[그림 2-17] 체납 테이블과 NICE 신용정보와 결합

체납 테이블을 기반으로 파생변수를 생성한 데이터와 NICE 신용정보 데이터를 납세자번호를 기준으로 병합하여 총 16개의 변수를 생성하였다.

□ 8. NICE 정보의 결측값 처리

납세자번호	부과금액	가산금	체납액	세목명	세목명건수	체납건수	시군구코드	과세년월	CB점수	소득추정등급	소득추정금액	체납회수등급	회수율	SP등급	카드개수
1	12500	370	12870	주민세(개인균등)	1	1	28177	201508	0	0	0	5	26%	0	0
2	12500	370	12870	주민세(개인균등)	1	1	28237	201508	0	0	0	5	26%	0	1
3	16020	480	16500	지방소득세(양도소득세분)	1	1	28177	201602	0	0	0	5	26%	0	0
4	12500	370	12870	주민세(개인균등)	1	1	28200	201912	0	0	0	5	26%	0	0
5	93750	2810	96560	주민세(법인균등)	1	1	28200	201808	0	0	0	5	26%	0	0

[그림 2-18] 결측 변수 확인

[1] "체납자 중 부과금액의 결측값은 0.12 % 입니다."
 [1] "체납자 중 가산금의 결측값은 0.01 % 입니다."
 [1] "체납자 중 CB점수의 결측값은 1.36 % 입니다."
 [1] "체납자 중 소득추정등급의 결측값은 5.29 % 입니다."
 [1] "체납자 중 소득추정금액의 결측값은 5.29 % 입니다."
 [1] "체납자 중 SP등급의 결측값은 5.33 % 입니다."
 [1] "체납자 중 카드개수의 결측값은 3.76 % 입니다."
 [1] "CB점수발송 내역 전처리 완료"
 [1] "소득추정등급발송 내역 전처리 완료"
 [1] "소득추정금액발송 내역 전처리 완료"
 [1] "SP등급발송 내역 전처리 완료"
 [1] "체납자 중 부과금액의 결측값은 0.12 % 입니다."
 [1] "체납자 중 가산금의 결측값은 0.01 % 입니다."
 [1] "체납자 중 카드개수의 결측값은 3.76 % 입니다."

[그림 2-19] 결측 변수 당 비율 확인

납세자번호	부과금액	가산금	체납액	세목명	세목명건수	체납건수	시군구코드	과세연월	CB점수	소득추정등급	소득추정금액	체납회수등급	회수율	SP등급	카드개수
1	12500	370	12870	주민세(개인균등)	1	1	28177	201508	521	10	3068	5	26%	5	0
2	12500	370	12870	주민세(개인균등)	1	1	28237	201508	521	10	3068	5	26%	5	1
3	16020	480	16500	지방소득세(양도소득세분)	1	1	28177	201602	521	10	3068	5	26%	5	0
4	12500	370	12870	주민세(개인균등)	1	1	28200	201912	521	10	3068	5	26%	5	0
5	93750	2610	96560	주민세(법인균등)	1	1	28200	201808	521	10	3068	5	26%	5	0

[그림 2-20] 결측값 대체

CB점수, 소득추정등급, 소득추정금액, SP등급에서 결측(0)인 값들은 체납회수등급을 기준으로 해당 변수의 평균값으로 대체하였다. 예를 들어 CB점수 결측(0점) 값의 체납회수등급은 5등급이고, 평균 체납회수 5등급의 CB점수 평균값으로 변경하였다. 부과금액, 가산금, 카드개수는 0에 대한 고유한 의미를 지니므로, 결측 처리하지 않았다.

□ 9. 금액 이상치 처리

부과금액	가산금	체납액
Min. : -7533050	Min. : -790890	Min. : -8323940
1st Qu.: 12500	1st Qu.: 370	1st Qu.: 12870
Median : 25000	Median : 740	Median : 25740
Mean : 202971	Mean : 33314	Mean : 236285
3rd Qu.: 62500	3rd Qu.: 1850	3rd Qu.: 64350
Max. : 439948270	Max. : 91637760	Max. : 479543590



부과금액	가산금	체납액
Min. : 0	Min. : 0	Min. : 10000
1st Qu.: 12500	1st Qu.: 370	1st Qu.: 12870
Median : 25000	Median : 740	Median : 25740
Mean : 206364	Mean : 33870	Mean : 240234
3rd Qu.: 62500	3rd Qu.: 1850	3rd Qu.: 64350
Max. : 439948270	Max. : 91637760	Max. : 479543590

[그림 2-21] 부과금액, 가산금, 체납액 이상치 확인제거

체납 데이터 중 부과금액, 가산금, 체납액에 음수 값을 제거하고, 체납액 기준 10,000원 이상만 추출하였다.

□ 10. 과세년월 변경 및 체납기간 부여

납세자번호	부과 금액	가산 금	체납 액	세목명	세목명 건수	체납 건수	시군구 코드	과세 년월	CB 점수	소득 추정 등급	소득 추정 금액	체납 회수 등급	회수 율	SP 등급	카드 건수
1	12500	370	12870	주민세(개인균등)	1	1	28177	201508	521	10	3068	5	26%	5	0
2	12500	370	12870	주민세(개인균등)	1	1	28237	201508	521	10	3068	5	26%	5	1
3	16020	480	16500	지방소득세(양도소득세분)	1	1	28177	201602	521	10	3068	5	26%	5	0
4	12500	370	12870	주민세(개인균등)	1	1	28200	201912	521	10	3068	5	26%	5	0
5	93750	2810	96560	주민세(법인균등)	1	1	28200	201808	521	10	3068	5	26%	5	0

[그림 2-22] 과세년월 변수 확인

납세자번호	부과 금액	가산 금	체납 액	세목명	세목명 건수	체납 건수	시군구 코드	발생 년도	CB 점수	소득 추정 등급	소득 추정 금액	체납 회수 등급	회수 율	SP 등급	카드 건수	체납 기간
1	12500	370	12870	주민세(개인균등)	1	1	28177	2015	521	10	3068	5	26%	5	0	5
2	12500	370	12870	주민세(개인균등)	1	1	28237	2015	521	10	3068	5	26%	5	1	5
3	16020	480	16500	지방소득세(양도소득세...	1	1	28177	2016	521	10	3068	5	26%	5	0	4
4	12500	370	12870	주민세(개인균등)	1	1	28200	2019	521	10	3068	5	26%	5	0	1
5	93750	2810	96560	주민세(법인균등)	1	1	28200	2018	521	10	3068	5	26%	5	0	2

[그림 2-23] 컬럼 변경 후 체납기간 파생변수 생성

‘과세년월’ 변수의 앞 4글자를 추출하여 ‘발생년도’의 변수로 변경하였다. 또한, 체납 기간이 오래될수록 체납에 영향을 미칠 것이라는 가설을 통해 현 시점 (2020년) 기준으로 발생년도를 차감한 ‘체납기간’의 파생변수를 생성하였다.

□ 11. 체납액 중 가산금 비율 파생변수 생성

납세자번호	부과 금액	가산 금	체납 액	세목명	세목명 건수	체납 건수	시군구 코드	발생 년도	CB 점수	소득 추정 등급	소득 추정 금액	체납 회수 등급	회수 율	SP 등급	카드 건수	체납 기간
1	12500	370	12870	주민세(개인균등)	1	1	28177	2015	521	10	3068	5	26%	5	0	5
2	12500	370	12870	주민세(개인균등)	1	1	28237	2015	521	10	3068	5	26%	5	1	5
3	16020	480	16500	지방소득세(양도소득세분)	1	1	28177	2016	521	10	3068	5	26%	5	0	4
4	12500	370	12870	주민세(개인균등)	1	1	28200	2019	521	10	3068	5	26%	5	0	1
5	93750	2810	96560	주민세(법인균등)	1	1	28200	2018	521	10	3068	5	26%	5	0	2

[그림 2-24] 가산금과 체납액 변수

납세자번호	부과 금액	가산 금	체납 액	세목명	세목명 건수	체납 건수	시군구 코드	발생 년도	CB 점수	소득 추정 등급	소득 추정 금액	체납 회수 등급	회수 율	SP 등급	카드 건수	체납 기간	가산 비율
1	12500	370	12870	주민세(개인균등)	1	1	28177	2015	521	10	3068	5	26%	5	0	5	2.875
2	12500	370	12870	주민세(개인균등)	1	1	28237	2015	521	10	3068	5	26%	5	1	5	2.875
3	16020	480	16500	지방소득세(양도소득세분)	1	1	28177	2016	521	10	3068	5	26%	5	0	4	2.909
4	12500	370	12870	주민세(개인균등)	1	1	28200	2019	521	10	3068	5	26%	5	0	1	2.875
5	93750	2810	96560	주민세(법인균등)	1	1	28200	2018	521	10	3068	5	26%	5	0	2	2.910

[그림 2-25] 가산비율 파생변수 생성

체납액 중 가산금이 많을수록 체납에 영향에 높을 것을 가설로 세워, 체납액 대비 가산금의 비율을 ‘가산비율’이라는 파생변수를 생성하였다.

□ 12. 세목명 별 가산점 파생변수 생성

	체납회수 등급	세목명	체납액	등급별가산점	전체가산점
1	1	주민세(개인균등)	182613450	7.523570e-01	5.597254e-03
2	1	주민세(개인사업)	29144190	1.200724e-01	8.932936e-04
3	1	취득세(부동산)	11011490	4.536671e-02	3.375113e-04
97	9	주민세(법인균등)	96560	1.247518e-03	2.959644e-06
98	9	등록면허세(면허)	86520	1.117805e-03	2.651910e-06
99	10	지방소득세(종합소득세분)	16896317940	6.250620e-01	5.178862e-01
100	10	지방소득세(양도소득세분)	6985921140	2.584370e-01	2.141243e-01
101	10	취득세(부동산)	2015345440	7.455565e-02	6.177201e-02

[그림 2-26] 세목별 가산점수

특정 세목명에 따라 체납에 영향을 미치는지 알아보고자, 체납회수등급 별 체납액 기준에 따라 세목명을 나열하였다. ‘등급별가산점’은 해당 체납회수등급 체납액 합계 대비 세목 체납액을 나눈 비율이며, ‘전체가산점’은 전체 체납액 합계 대비 해당 세목 체납액을 나눈 비율을 의미한다.

납세자번호	부과 금액	가 산 금	체 납 액	세목명	세목명건수	체 납 건 수	시 군 구 코 드	발 생 년 도	CB 점 수	소 독 주 정 등 급	소 독 주 정 금 액	체 납 회 수 등 급	회 수 율	SP 등 급	카 드 개 수	체 납 기 간	가 산 비 율	등급별 가산	전체가산	
1		12500	370	12870	주민세(개인균등)	1	1	26177	2015	521	10	3068	5	26%	5	0	5	2.875	0.640352930	2.459773e-02
2		12500	370	12870	주민세(개인균등)	1	1	26237	2015	521	10	3068	5	26%	5	1	5	2.875	0.640352930	2.459773e-02
3		16020	480	16500	지방소득세(양도소득세분)	1	1	26177	2016	521	10	3068	5	26%	5	0	4	2.909	0.016550225	6.357399e-04
4		12500	370	12870	주민세(개인균등)	1	1	26200	2019	521	10	3068	5	26%	5	0	1	2.875	0.640352930	2.459773e-02
5		93750	2810	96560	주민세(법인균등)	1	1	26200	2018	521	10	3068	5	26%	5	0	2	2.910	0.000856023	3.268221e-05

[그림 2-27] 등급별 가산 및 전체 가산점 파생변수 생성

체납회수등급 별 체납액에 따른 세목명의 가산 점수를 납세자 번호를 기준으로 ‘등급별가산점’과 ‘전체가산점’의 파생변수를 병합하였다.

2) 10월 발송 내역 정제

□ 10월 발송 체납 대상자 NICE 등급 부여

납세자번호	체납액	세목명	세목명건수	체납건수	시군구코드	발생년도	CB점수	소독추정등급	소독추정금액	체납회수등급	회수율	SP등급	카드개수	체납기간	가산비율	등급별가산	전체가산	전체가산
		12870 주민세(개인균등)	1	1	28177	2015	521	10	3068	5	26%	5	0	5	2.875	0.640352930	2.459773e-02	2.459773e-02
		12870 주민세(개인균등)	1	1	28237	2015	521	10	3068	5	26%	5	1	5	2.875	0.640352930	2.459773e-02	2.459773e-02
		16500 지방소득세(양도소득세..	1	1	28177	2016	521	10	3068	5	26%	5	0	4	2.909	0.016550225	6.357399e-04	6.357399e-04
		12870 주민세(개인균등)	1	1	28200	2019	521	10	3068	5	26%	5	0	1	2.875	0.640352930	2.459773e-02	2.459773e-02
		96560 주민세(법인균등)	1	1	28237	2019	521	10	3068	5	26%	5	0	1	2.910	0.000856023	3.288221e-05	3.288221e-05

[그림 2-28] 10월 발송 내역 : 119,618건 18개 변수 사용

10월 고지서 & 안내문 발송 대상자 중 납세자번호 별 체납액을 합산하고, 사전에 가설에 따른 파생변수와 NICE 신용등급을 결합하였다.

3) 10월 수납 내역 정제

□ 10월 수납 대상자 NICE 등급 부여

납세자번호	수납액	세목명	세목명건수	체납건수	시군구코드	발생년도	CB점수	소독추정등급	소독추정금액	체납회수등급	회수율	SP등급	카드개수	체납기간	가산비율	등급별가산	전체가산	전체가산
		96560 주민세(법인균등)	1	1	28200	2018	521	10	3068	5	26%	5	0	2	2.910	0.000856023	3.288221e-05	3.288221e-05
		12870 주민세(개인균등)	1	1	28260	2015	521	10	3068	5	26%	5	0	5	2.875	0.640352930	2.459773e-02	2.459773e-02
		12870 주민세(개인균등)	1	1	28237	2015	521	10	3068	5	26%	5	0	5	2.875	0.640352930	2.459773e-02	2.459773e-02
		12870 주민세(개인균등)	1	1	28200	2016	521	10	3068	5	26%	5	0	4	2.875	0.640352930	2.459773e-02	2.459773e-02
		12870 주민세(개인균등)	1	1	28200	2015	753	10	3068	4	40%	5	1	5	2.875	0.679730000	3.279520e-02	3.288221e-05

[그림 2-29] 10월 수납 내역 : 8,124건 18개 변수 사용

10월 수납 대상자 중 납세자번호 별 수납액을 합산하고, 사전에 가설에 따른 파생변수와 NICE 신용등급을 결합하였다.

3. 탐색적 데이터 분석(EDA)

가. 인천광역시 연도별 체납현황

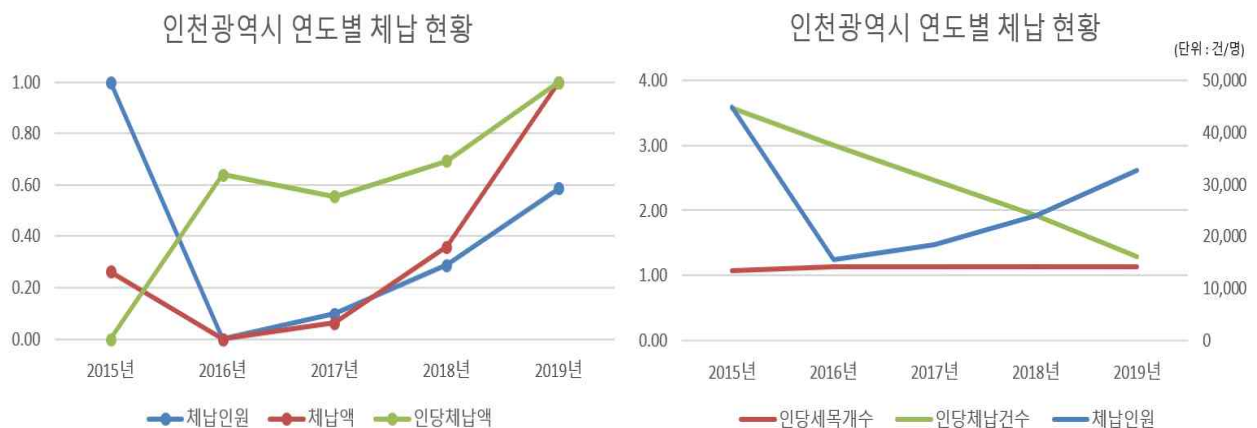
1) 인천광역시 연도별 체납액 현황 추이

발생년도	체납인원	체납액	인당세목개수	인당체납건수	인당체납액
2015년	44,771	₩5,482,764,740	1.08	3.57	₩122,462
2016년	15,529	₩3,797,218,470	1.13	3	₩244,524
2017년	18,421	₩4,207,936,030	1.14	2.46	₩228,431
2018년	23,980	₩6,106,924,270	1.14	1.92	₩254,667
2019년	32,680	₩10,233,467,800	1.13	1.3	₩313,142

[그림 3-1] 인천시 연도별 체납액 현황

2015년부터 2019년까지 체납액과 인당 체납액도 같이 증가하며, 체납 인원도 2016년 이후로 꾸준히 증가하는 추세를 확인할 수 있다.

2) 인천광역시 연도별 체납액 현황 그래프



좌측 그래프를 통해 체납액이 체납 인원보다 많아지면서, 인당 체납액이 313,142원으로 증가하는 흐름을 파악할 수 있다. 우측 그래프를 통해 인당체납 건수는 19년 기준 1.3건으로 감소하였지만, 인당 체납액이 높은 것으로 보아 체납징수 현황에서 '1인 체납자'에 대한 영향이 커짐을 확인할 수 있다.

나. 인천광역시 등급별 현황 및 상관분석

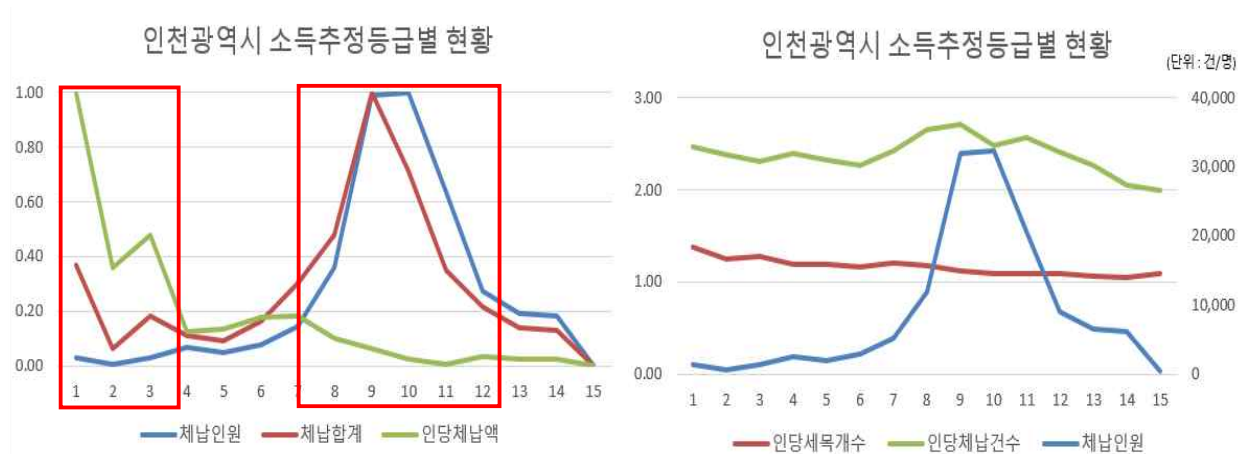
1) 인천광역시 소득추정등급별 체납액 현황 추이

소득추정등급	체납인원	체납합계	인당세목개수	인당체납건수	인당체납액
1	1,446	₩2,760,351,610	1.38	2.47	₩1,908,957
2	698	₩533,411,910	1.26	2.38	₩764,200
3	1,438	₩1,405,790,350	1.27	2.31	₩977,601
4	2,562	₩879,630,230	1.19	2.4	₩343,337
5	1,989	₩714,729,950	1.19	2.32	₩359,341
6	2,913	₩1,263,268,010	1.16	2.27	₩433,666
7	5,117	₩2,289,659,130	1.2	2.42	₩447,461
8	11,912	₩3,577,743,610	1.17	2.65	₩300,348
9	31,999	₩7,410,734,750	1.12	2.71	₩231,593
10	32,345	₩5,278,602,800	1.09	2.48	₩163,197
11	20,669	₩2,631,742,810	1.09	2.56	₩127,328
12	9,093	₩1,650,417,190	1.08	2.41	₩181,504
13	6,620	₩1,084,424,010	1.07	2.26	₩163,810
14	6,205	₩1,006,648,850	1.05	2.05	₩162,232
15	448	₩53,519,160	1.09	1.99	₩119,462

[그림 3-2] 인천시 소득추정등급별 체납액 현황

소득추정등급이 높을수록 인당 체납액이 많고, 중간(7~12)등급일수록 체납 인원
과 체납합계가 많이 분포된 것을 확인할 수 있다.

2) 인천광역시 소득추정등급별 체납액 현황 그래프



좌측 그래프를 통해 소득추정등급이 높을수록 인당 체납액이 많은 것을 확인할 수 있고, 체납합계와 체납 인원은 소득추정등급에 따라 유사한 흐름을 보였다. 우측 그래프를 통해 소득추정등급에 따라 인당세목개수와 인당체납건수는 큰 차이는 없음을 확인할 수 있다.

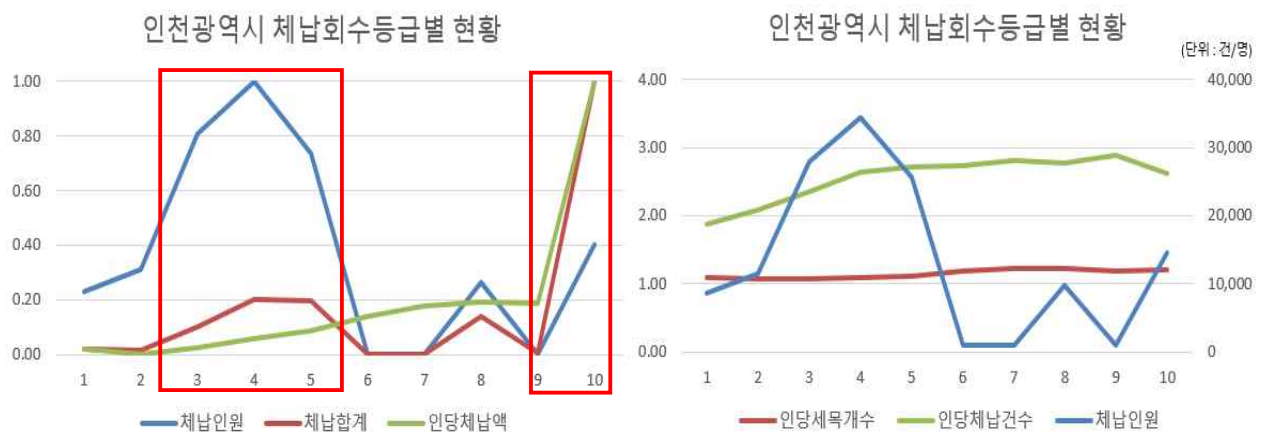
3) 인천광역시 체납회수등급별 체납액 현황 추이

체납회수등급	체납인원	체납합계	인당세목개수	인당체납건수	인당체납액
1	8,657	₩605,252,600	1.08	1.87	₩69,915
2	11,462	₩491,272,920	1.07	2.09	₩42,861
3	27,996	₩2,009,171,390	1.08	2.35	₩71,766
4	34,448	₩3,858,355,860	1.1	2.64	₩112,005
5	25,552	₩3,788,415,040	1.11	2.71	₩148,263
6	1,025	₩217,980,650	1.2	2.74	₩212,664
7	1,035	₩269,734,970	1.22	2.8	₩260,613
8	9,770	₩2,705,780,330	1.22	2.78	₩276,948
9	1,027	₩281,703,830	1.19	2.89	₩274,298
10	14,482	₩18,313,006,780	1.21	2.63	₩1,264,536

[그림 3-3] 인천시 체납회수등급별 체납액 현황

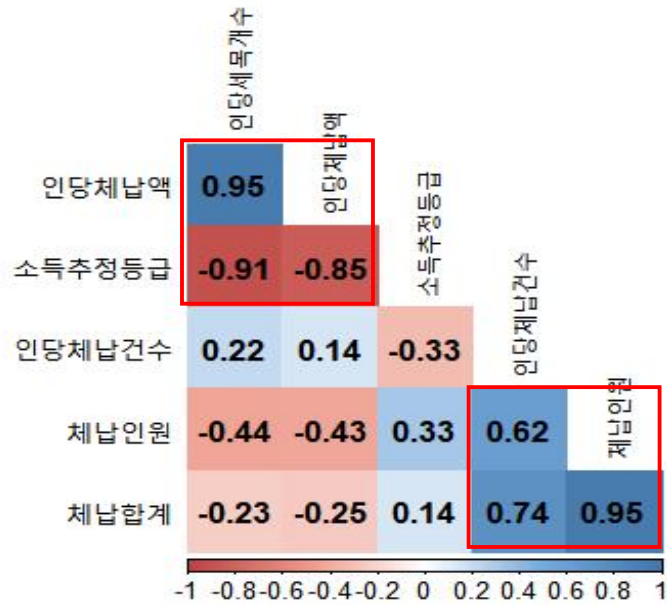
체납회수등급이 중간(3~5)등급일수록 체납 인원이 가장 많은 것을 확인할 수 있으며, 체납회수등급이 10등급인 경우 인당체납액이 1,264,535원으로 가장 많은 것으로 확인하였다.

4) 인천광역시 체납회수등급별 체납액 현황 그래프



좌측 그래프를 통해 체납회수등급이 3~5등급일수록 체납 인원이 가장 많았으며, 인당 체납액은 10등급이 가장 많음을 확인할 수 있다. 우측 그래프를 통해 체납 회수등급에 따른 인당세목개수와 인당체납건수의 큰 차이는 없는 것으로 확인할 수 있다.

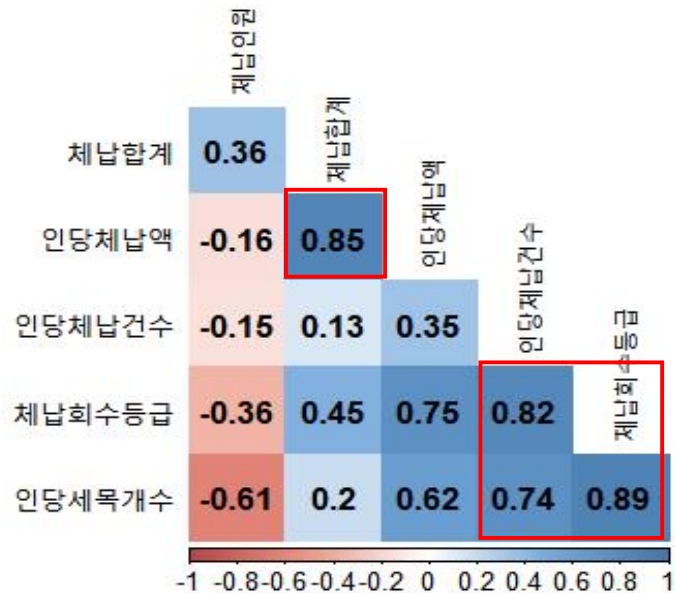
5) 인천광역시 소득추정등급별 상관분석



[그림 3-4] 소득추정등급별 상관분석

소득추정등급별 상관분석을 통해, 등급이 낮을수록 인당세목개수와 인당체납액이 낮아지는 음의 상관관계를 확인하였고, 인당 체납액이 많을수록 인당세목개수도 많아지며, 체납합계와 체납 인원도 많아지는 양의 상관관계를 확인하였다.

6) 인천광역시 체납회수등급별 상관분석



[그림 3-5] 체납회수등급별 상관분석

체납회수등급별 상관분석을 통해, 등급이 낮을수록 인당체납건수, 인당세목개수, 인당 체납액이 많아지고, 인당 체납액과 체납합계가 동시에 증가하는 양의 상관관계를 확인하였다.

다. 10월 발송내역 중 등급별 현황 및 상관분석

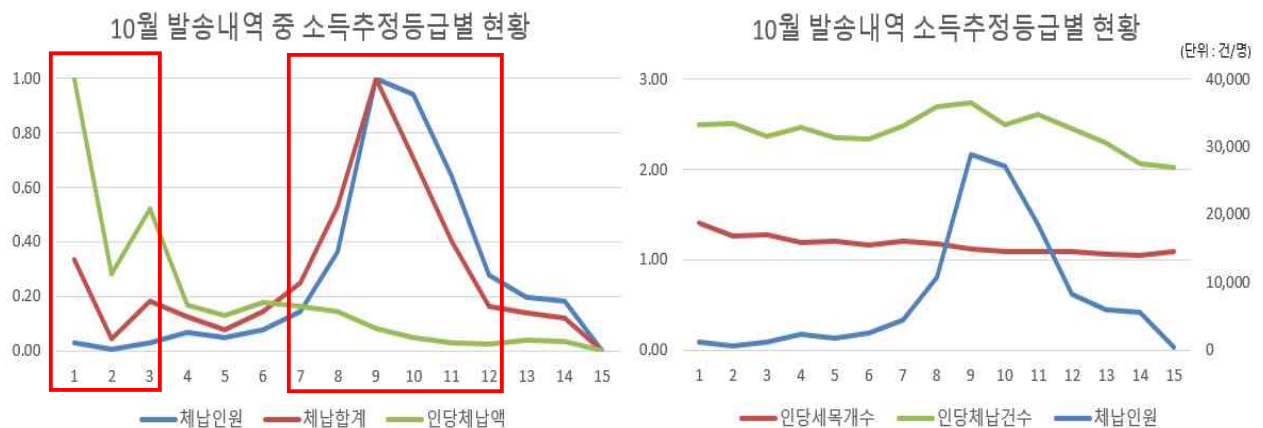
1) 10월 발송내역 중 소득추정등급별 현황 추이

소득추정등급	체납인원	체납합계	인당세목개수	인당체납건수	인당체납액
1	1,268	₩2,227,496,750	1.41	2.5	₩1,756,701
2	595	₩333,238,230	1.26	2.51	₩560,064
3	1,266	₩1,212,341,430	1.28	2.37	₩957,616
4	2,267	₩843,259,070	1.19	2.46	₩371,971
5	1,763	₩536,019,480	1.2	2.35	₩304,038
6	2,547	₩987,879,110	1.16	2.33	₩387,860
7	4,517	₩1,653,000,250	1.21	2.48	₩365,951
8	10,671	₩3,497,703,980	1.18	2.69	₩327,777
9	28,781	₩6,553,394,220	1.13	2.74	₩227,699
10	27,156	₩4,647,338,340	1.09	2.49	₩171,135
11	18,643	₩2,642,102,510	1.09	2.6	₩141,721
12	8,226	₩1,114,977,340	1.09	2.45	₩135,543
13	5,976	₩940,371,080	1.07	2.3	₩157,358
14	5,537	₩834,455,240	1.05	2.06	₩150,705
15	405	₩37,010,310	1.09	2.03	₩91,383

[그림 3-6] 10월 발송내역 소득추정등급 현황

10월 발송내역은 인천광역시 체납자 대상으로 고지서 및 안내문을 발생한 대상자로, 소득추정등급이 높을수록 인당 체납액이 많고, 중간(7~12)등급일수록 체납 인원과 체납합계가 많이 분포된 것을 확인할 수 있다.

2) 10월 발송내역 중 소득추정등급별 현황 그래프



좌측 그래프를 통해 소득추정등급이 높을수록 인당 체납액이 많은 것을 확인할 수 있고, 체납합계와 체납 인원은 소득추정등급에 따라 유사한 흐름을 보였다. 우측 그래프를 통해 소득추정등급에 따라 인당 세목 개수와 인당 체납 건수는 큰 차이는 없음을 확인할 수 있다.

3) 10월 발송내역 중 체납회수등급별 현황 추이

체납회수등급	체납인원	체납합계	인당세목개수	인당체납건수	인당체납액
1	7,305	₩566,416,260	1.09	1.93	₩77,538
2	9,964	₩452,075,560	1.07	2.14	₩45,371
3	24,877	₩1,959,939,290	1.08	2.4	₩78,785
4	30,445	₩3,721,562,110	1.1	2.67	₩122,239
5	22,041	₩3,816,234,700	1.12	2.73	₩173,143
6	931	₩218,762,570	1.2	2.76	₩234,976
7	943	₩217,042,450	1.22	2.81	₩230,162
8	8,947	₩2,538,871,990	1.22	2.79	₩283,768
9	943	₩319,501,850	1.19	2.89	₩338,814
10	13,222	₩14,250,180,560	1.21	2.65	₩1,077,763

[그림 3-7] 10월 발송내역 체납회수등급 현황

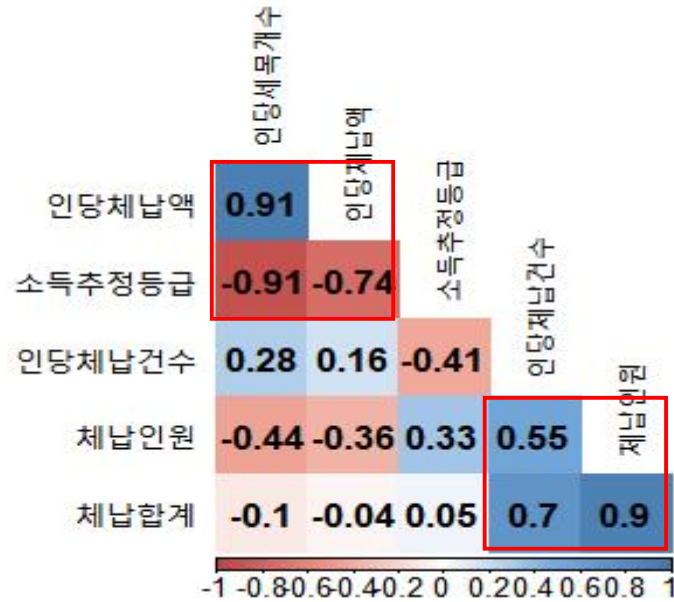
10월 발송 내역 중 체납회수등급이 중간(3~5)등급일수록 체납 인원이 가장 많은 것을 확인할 수 있으며, 체납회수등급이 10등급인 경우 인당 체납액이 1,077,763원으로 가장 많은 것으로 확인하였다.

4) 10월 발송내역 중 체납회수등급별 현황 그래프



좌측 그래프를 통해 체납회수등급이 3~5등급일수록 체납 인원이 가장 많았으며, 인당 체납액은 10등급이 가장 많음을 확인할 수 있다. 우측 그래프를 통해 체납 회수등급에 따른 인당세목개수와 인당체납건수의 큰 차이는 없는 것으로 확인할 수 있다.

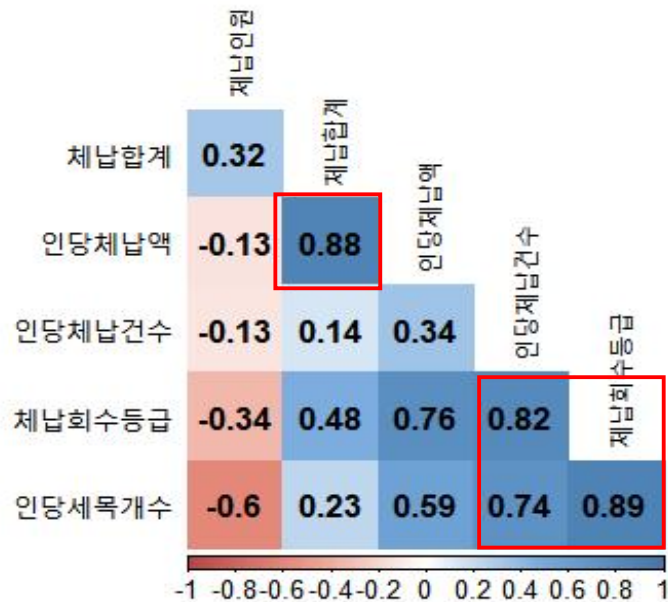
5) 10월 발송내역 중 소득추정등급별 상관분석



[그림 3-8] 소득추정등급별 상관분석

소득추정등급별 상관분석을 통해, 등급이 낮을수록 인당세목개수와 인당체납액이 낮아지는 음의 상관관계를 확인하였고, 인당 체납액이 많을수록 인당세목개수도 많아지며, 체납합계와 체납 인원도 많아지는 양의 상관관계를 확인하였다.

6) 10월 발송내역 중 체납회수등급별 상관분석



[그림 3-9] 체납회수등급별 상관분석

체납회수등급별 상관분석을 통해, 등급이 낮을수록 인당체납건수, 인당세목개수, 인당 체납액이 많아지고, 인당 체납액과 체납합계가 동시에 증가하는 양의 상관관계를 확인하였다.

라. 10월 수납내역 중 등급별 현황 및 상관분석

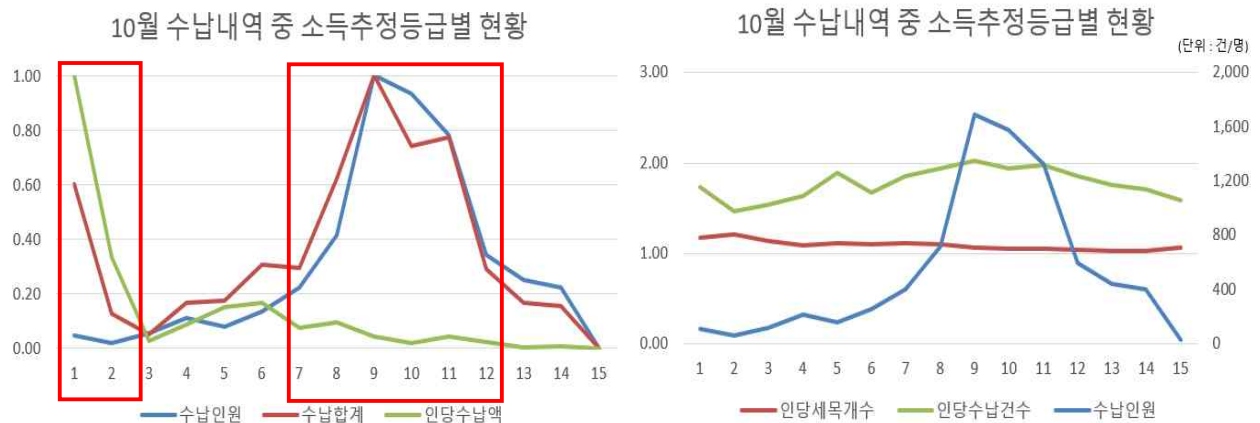
1) 10월 수납내역 중 소득추정등급별 현황 추이

소득추정등급	수납인원	수납합계	인당세목개수	인당수납건수	인당수납액	징수율
1	110	₩62,139,860	1.17	1.74	₩564,908	8.68%
2	65	₩13,928,630	1.22	1.46	₩214,287	10.92%
3	121	₩6,292,560	1.13	1.54	₩52,005	9.56%
4	219	₩18,346,780	1.09	1.64	₩83,775	9.66%
5	161	₩18,956,120	1.12	1.89	₩117,740	9.13%
6	254	₩32,218,810	1.1	1.67	₩126,846	9.97%
7	399	₩31,208,400	1.12	1.85	₩78,217	8.83%
8	720	₩64,426,700	1.1	1.94	₩89,482	6.75%
9	1,689	₩102,453,320	1.07	2.02	₩60,659	5.87%
10	1,580	₩76,304,620	1.05	1.94	₩48,294	5.82%
11	1,325	₩79,659,460	1.05	1.98	₩60,120	7.11%
12	600	₩30,806,920	1.04	1.86	₩51,345	7.29%
13	446	₩18,261,160	1.03	1.75	₩40,944	7.46%
14	403	₩16,920,540	1.02	1.71	₩41,986	7.28%
15	32	₩1,222,080	1.06	1.59	₩38,190	7.90%

[그림 3-10] 10월 수납내역 소득추정등급 현황

10월 수납내역은 인천광역시 체납자 대상으로 고지서 및 안내문을 발생 중 수납한 대상으로, 소득추정등급이 높을수록 인당 납부액이 많고, 중간(7~12)등급일수록 수납 인원과 수납합계가 많이 분포된 것을 확인할 수 있다.

2) 10월 수납내역 중 소득추정등급별 현황 그래프



좌측 그래프를 통해 소득추정등급이 높을수록 인당 납부액이 많은 것을 확인할 수 있고, 수납합계와 수납 인원은 소득추정등급에 따라 유사한 흐름을 보였다. 우측 그래프를 통해 소득추정등급에 따라 인당세목개수와 인당수납건수는 큰 차이는 없음을 확인할 수 있다.

3) 10월 수납내역 중 체납회수등급별 현황 추이

체납회수등급	수납인원	수납합계	인당세목개수	인당수납건수	인당수납액	징수율
1	1,005	₩27,200,100	1.05	1.53	₩27,065	13.76%
2	1,205	₩37,258,000	1.05	1.73	₩30,920	12.09%
3	2,403	₩140,524,930	1.06	1.87	₩58,479	9.66%
4	1,777	₩92,685,650	1.06	2.07	₩52,158	5.84%
5	939	₩57,052,590	1.08	2.13	₩60,759	4.26%
6	34	₩4,712,340	1.03	1.82	₩138,598	3.65%
7	49	₩14,937,350	1.14	2.76	₩304,844	5.20%
8	299	₩26,261,300	1.16	2.1	₩87,830	3.34%
9	25	₩1,271,740	1.2	2.48	₩50,870	2.65%
10	388	₩171,241,960	1.09	1.93	₩441,345	2.93%

[그림 3-11] 10월 수납내역 소득추정등급 현황

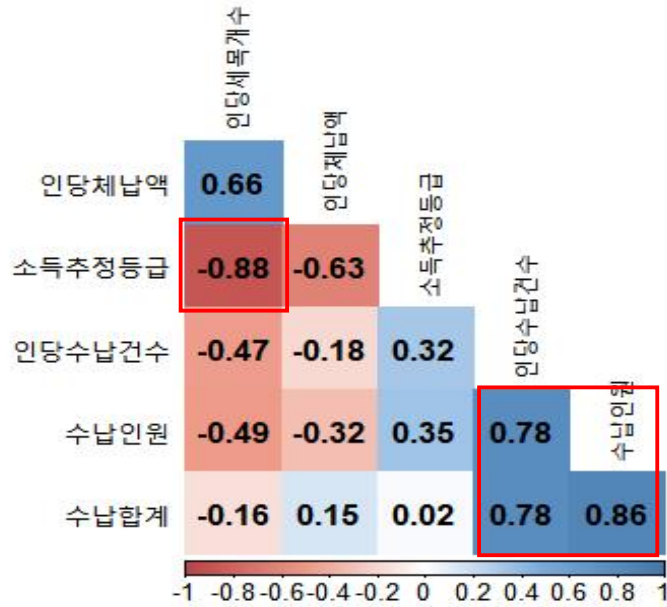
10월 수납내역 중 체납회수등급이 높은(1~4)등급일수록 수납 인원과 수납합계가 가장 많은 것을 확인할 수 있으며, 체납회수등급이 높을수록 발송 내역 대비 수납 비율인 '징수율'이 높아지는 것을 확인하였다.

4) 10월 수납내역 중 체납회수등급별 현황 그래프



좌측 그래프를 통해 체납회수등급이 1~4등급일수록 수납 인원이 가장 많았으며, 인당 수납액은 회수등급 10등급에 가장 많음을 확인하였다. 우측 그래프를 통해 체납회수등급이 높을수록 수납 인원은 높은 반면, 회수등급이 낮을수록 인당수납 건수는 높은 것을 보인다. 체납회수등급에 따른 인당세목개수 큰 차이는 없는 것으로 확인하였다.

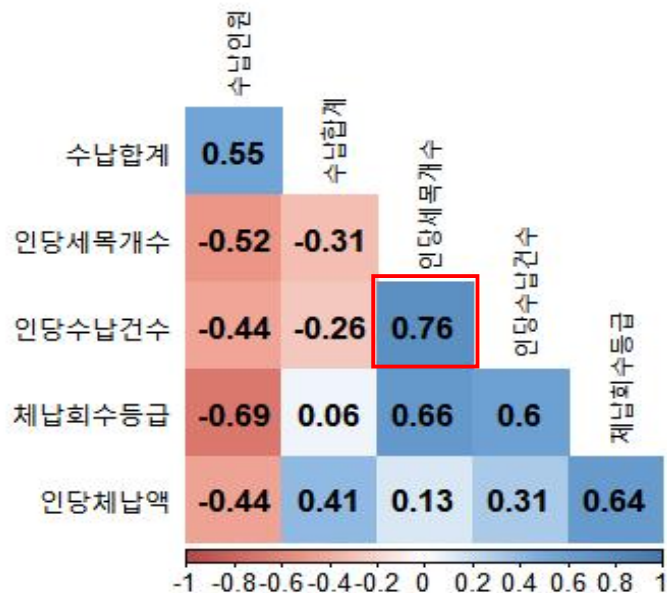
5) 10월 수납내역 중 소득추정등급별 상관분석



[그림 3-12] 소득추정등급별 상관분석

소득추정등급별 상관분석을 통해, 등급이 낮을수록 인당 세목 개수가 낮아지는 음의 상관관계를 확인하였고, 수납합계, 인당수납건수, 수납 인원 각각 양의 상관관계가 있는 것을 확인하였다.

6) 10월 수납내역 중 체납회수등급별 상관분석

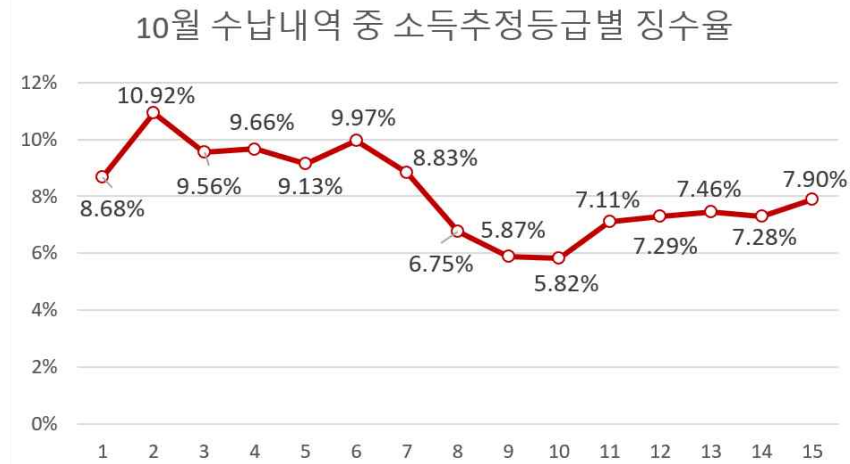


[그림 3-13] 체납회수등급별 상관분석

체납회수등급별 상관분석을 통해, 등급에 따른 상관성은 없는 것으로 확인하였고, 인당수납건수와 인당세목개수 간 양의 상관관계가 있는 것으로 확인하였다.

마. 10월 발송 대비 수납 등급별 징수율

1) 10월 발송 대비 수납 소득추정등급별 징수율



[그림 3-14] 소득추정등급에 따른 징수율

10월 발송 대비 수납한 사람들에 대한 비율인 ‘징수율’에 대한 그래프로, 소득추정등급에 따른 징수율의 변화는 최대 5.07%p이며, 평균 2~3%p의 차이로 큰 변화가 없는 것으로 보인다.

2) 10월 발송 대비 수납 체납회수등급별 징수율



[그림 3-15] 체납회수등급에 따른 징수율

반면 체납회수등급에 따른 징수율의 변화는 최대 9.52%p로 가장 큰 차이가 있으며, 회수등급이 낮을수록 동시에 징수율이 낮아지는 현상을 확인하였다. 이를 통해 체납 대상자를 고지서 발송으로 한 ‘체납 모형 검증 작업’을 통해 체납회수등급의 의미가 있는 것으로 확인하였다. 따라서 회수등급이 낮은 대상자에 집중하여 징수하는 등 회수등급에 따른 차별화된 체납징수 활동을 해야 할 필요성을 확인하였다.

바. 체납 세목명 분석

1) 체납액 상위 5개 세목명 분석

세목명	체납합계
지방소득세(종합소득세분)	₩17,825,118,820
지방소득세(양도소득세분)	₩7,080,223,810
주민세(개인균등)	₩3,807,416,610
취득세(부동산)	₩2,065,512,940
주민세(개인사업)	₩1,000,138,200

[그림 3-16] 체납합계 기준 세목명

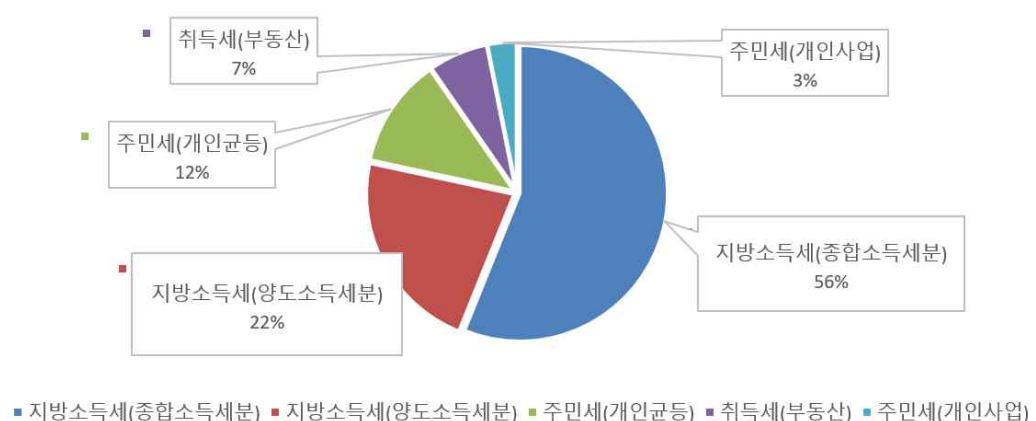
세목명	체납건수
주민세(개인균등)	124,120
지방소득세(종합소득세분)	14,198
주민세(개인사업)	7,347
지방소득세(양도소득세분)	1,832
취득세(부동산)	373

[그림 3-17] 체납건수 기준 세목명

다음 그림은 2020년 8월 기준 체납된 세목에 대한 분석이며 전체 체납액의 97%에 해당하는 세목명이다. 지방소득세(종합소득세분)는 178억 원으로 가장 많은 체납액이며 주민세(개인사업)는 10억 원으로 가장 낮은 체납액으로 확인하였다.

주민세(개인균등)와 주민세(개인사업)은 체납액 대비 체납 건수가 가장 많은 것을 확인하였고, 이는 인당 체납액이 낮은 소액체납자가 많다는 것을 의미한다. 반면 지방소득세(종합소득세분)과 지방소득세(양도소득세분)은 체납액 대비 체납 건수가 작은 것을 통해 인당 체납액이 많은 고액체납자가 많다는 것으로 분석되었다.

체납액 상위 5위 세목명 그래프



[그림 3-18] 체납액 기준 상위 5개 항목 세목명

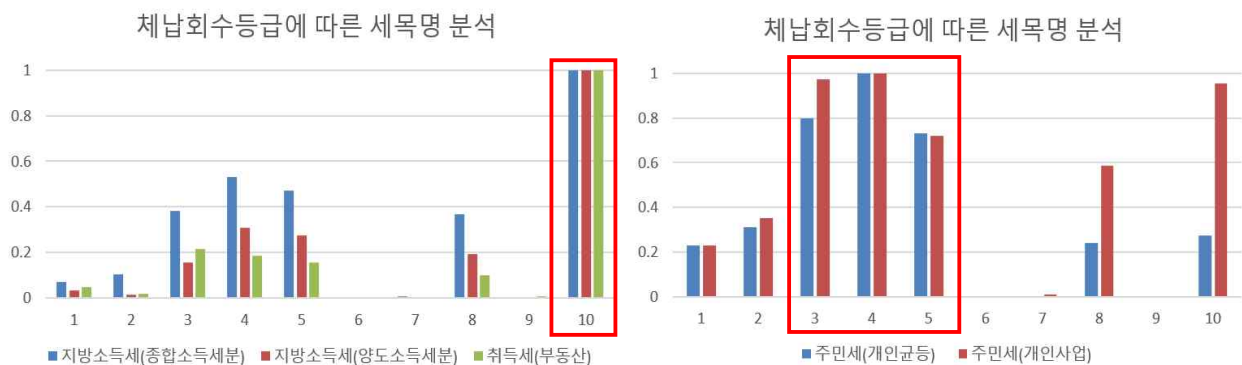
체납액 기준 상위 5개의 그래프로, 지방소득세(종합소득세분) 56%, 지방소득세(양도소득세분) 22%, 주민세(개인균등) 12%, 취득세(부동산) 7%, 주민세(개인사업) 3%으로 나타났다.

2) 체납회수등급 별 상위 5위 세목명 분석

체납회수등급	지방소득세_종합소득세분	지방소득세_양도소득세분	주민세_개인균등	취득세_부동산	주민세_개인사업
1	452	48	8,340	11	393
2	592	33	10,952	5	556
3	1,805	148	26,413	46	1,399
4	2,451	275	32,912	40	1,438
5	2,190	245	24,357	34	1,056
6	159	20	934	1	79
7	162	23	932	1	95
8	1,748	178	8,678	22	876
9	136	20	933	3	81
10	4,503	842	9,669	210	1,374

[그림 3-19] 체납회수등급별 상위 5위 세목명 체납건수

체납액 기준 상위 5위 항목을 대상으로 체납회수등급에 따른 체납 건수를 확인하였다. 지방소득세(종합소득세분), 지방소득세(양도소득세분), 취득세(부동산)는 체납회수등급이 낮을수록 체납 건수가 많아지는 것을 확인할 수 있으며, 주민세(개인균등)와 주민세(개인사업)는 체납회수등급이 높을수록 체납 건수가 많은 것을 확인하였다.



좌측그래프를 통해 지방소득세(종합소득세분), 지방소득세(양도소득세분), 취득세(부동산)는 체납회수등급의 10등급에서 많이 발생하는 것을 확인하였고, 우측 그래프를 통해 주민세(개인균등)와 주민세(개인사업)는 체납회수등급이 3~5등급에서 많이 발생하는 것을 확인하였다.

3) 체납액 기준 다중 세목명 발생 유형

세목명	체납합계
주민세(개인균등),지방소득세(종합소득세분)	₩4,923,047,000
주민세(개인균등),지방소득세(양도소득세분)	₩1,474,533,320
주민세(개인균등),지방소득세(종합소득세분),지방소득세(양도소득세분)	₩973,447,190
주민세(개인균등),주민세(개인사업),지방소득세(종합소득세분)	₩895,560,400
지방소득세(종합소득세분),지방소득세(양도소득세분)	₩713,077,610

[그림 3-20] 체납액 기준 다중 발생 세목명

납세자번호별 동시 발생 체납 세목명을 분석하여 주민세와 지방소득세가 같이 발생하는 것을 확인하고, 이에 따른 체납액이 가장 많은 것으로 확인하였다.

체납회수등급	A	B	C	D	E
1	252	32	4	51	0
2	295	20	0	52	2
3	770	80	10	154	5
4	1,150	173	18	234	5
5	1,070	155	19	220	2
6	75	11	2	13	0
7	67	12	2	18	0
8	767	92	18	145	8
9	53	9	3	13	1
10	775	100	14	163	43

[그림 3-21] 체납회수등급별 다중 발생 세목명 체납건수

A는 ‘주민세(개인균등)와 지방소득세(종합소득세분)’를 의미하며, E는 ‘지방소득세(종합소득세분)와 지방소득세(양도소득세분)’를 의미한다.

그림과 같이 체납회수등급이 3~5등급일 때 가장 많은 건수를 보이고, 다음으로 10등급에서 많은 것을 확인할 수 있다. 이를 통해 다중 체납 세목 당 회수등급에 따른 발생이 차이가 있는 것으로 확인하였다.

사. 시·군구별 체납징수 현황

1) 시·군구별 체납 담당자 공무원을 통한 체납 징수현황 파악

구분	체납건수	체납명수	인당체납건수	체납액	인원
부평구	59,877	24,216	2.47	₩2,220,866,600	9
서구	63,137	24,138	2.62	₩3,026,341,260	10
남동구	61,758	24,332	2.54	₩1,660,767,240	14
미추홀구	66,232	27,323	2.42	₩2,059,773,750	16
중구	20,498	8,982	2.28	₩798,792,630	5
계양구	33,523	13,443	2.49	₩1,203,962,480	9
연수구	28,427	11,591	2.45	₩2,109,204,190	8
동구	8,199	3,436	2.39	₩248,793,730	5
옹진군	702	330	2.13	₩16,250,040	2
강화군	499	229	2.18	₩139,449,920	3

[표 3-1] 체납 담당자 공무원 조직도를 통한 징수 인원 현황

시·군구별 체납 건수와 체납 명수는 66,232건과 27,323명으로 미추홀구가 가장 많은 반면, 인당체납건수와 체납액은 2.54건과 30억 원으로 가장 많았다. 이에 따른 체납 담당자 공무원 현황을 통해 체납징수 활동대비 인력배치가 적절히 이루어졌는지 확인하고자 하였다.

구분	체납 담당자당 미정리 체납건수	체납 담당자당 미정리 체납명수	미정리 체납건당 투입가능 시간(분)
부평구	6653.0	2690.7	18.04
서구	6313.7	2413.8	19.01
남동구	4411.3	1738.0	27.20
미추홀구	4139.5	1707.7	28.99
중구	4099.6	1796.4	29.27
계양구	3724.8	1493.7	32.22
연수구	3553.4	1448.9	33.77
동구	1639.8	687.2	73.18
옹진군	351.0	165.0	341.88
강화군	166.3	76.3	721.44

[표 3-2] 시·군구별 미정리 체납건당 투입가능 시간

체납 담당자당 미정리 체납 건수와 체납 명수는 배치 인원 대비 체납 건수와 체납 명수를 나눈 값으로, 부평구와 서구가 많음을 확인하였다. 체납징수업무 담당 인력 규모를 동시에 고려하여 미정리 체납 건당 투입 가능 시간은 세무공무원의 연간 근무시간을 2,000시간(주 40시간 * 50주)기준으로 산정하였으며, 부평구와 서구는 18,04분, 19,01분으로, 타 시·군구보다 체납징수업무를 담당하는 인력이 절대적으로 부족함을 알 수 있다.

4. 모델링 개발 과정

가. 분석 데이터 셋 생성

1) 수납 및 체납 데이터 병합

□ 수납 대상자 데이터 우량 정의

	체납액	세목명건수	체납건수	CB점수	소득추정등급	소득추정금액	체납회수등급	SP등급	카드개수	체납기간	가산비율	등급별가산	전체가산	여부
1	96560	1	1	521	10	3068	5	5	0	2	2.910	0.000856023	3.288221e-05	1
2	12870	1	1	521	10	3068	5	5	0	5	2.875	0.640352930	2.459773e-02	1
3	12870	1	1	521	10	3068	5	5	0	5	2.875	0.640352930	2.459773e-02	1
4	12870	1	1	521	10	3068	5	5	0	4	2.875	0.640352930	2.459773e-02	1

[그림 4-1] 10월 수납 대상자 수납내역 데이터

10월 체납자 대상 발송 중 수납 대상자인 8,124명에 대해 우량(1)을 부여하였다.

□ 체납 대상자 데이터 불량 정의

	체납액	세목명건수	체납건수	CB점수	소득추정등급	소득추정금액	체납회수등급	SP등급	카드개수	체납기간	가산비율	등급별가산	전체가산	여부
1	12870	1	1	521	10	3068	5	5	0	5	2.875	0.640352930	2.459773e-02	0
2	12870	1	1	521	10	3068	5	5	1	5	2.875	0.640352930	2.459773e-02	0
3	16500	1	1	521	10	3068	5	5	0	4	2.909	0.016550225	6.357399e-04	0
4	12870	1	1	521	10	3068	5	5	0	1	2.875	0.640352930	2.459773e-02	0

[그림 4-2] 10월 체납 대상자 발송 중 수납 대상자 제외 데이터

10월 체납자 대상 발송 중 수납 대상자를 제외한 119,618명에 대해 불량(0)을 부여하였다.

□ 데이터 최종 병합

	체납액	세목명건수	체납건수	CB점수	소득추정등급	소득추정금액	체납회수등급	SP등급	카드개수	체납기간	가산비율	등급별가산	전체가산	여부
1	12870	1	1	521	10	3068	5	5	0	5	2.875	0.640352930	2.459773e-02	0
2	12870	1	1	521	10	3068	5	5	1	5	2.875	0.640352930	2.459773e-02	0
3	16500	1	1	521	10	3068	5	5	0	4	2.909	0.016550225	6.357399e-04	0
4	12870	1	1	521	10	3068	5	5	0	1	2.875	0.640352930	2.459773e-02	0

[그림 4-3] 최종 분석 데이터 병합

10월 기준 체납 대상자와 수납 대상자를 병합하여 총 127,742 건의 데이터 셋을 구성하였다.

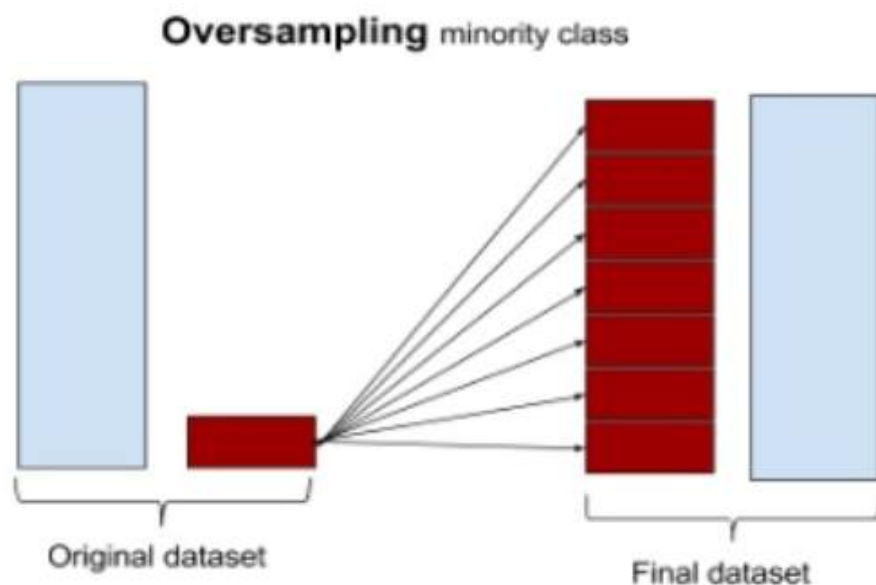
2) 클래스 불균형 해결(SMOTE)

□ SMOTE(synthetic minority oversampling technique)

구분	우량 건수	불량 건수	우량 비율	불량 비율
클래스 불균형	8,124	119,618	6.4%	93.6%

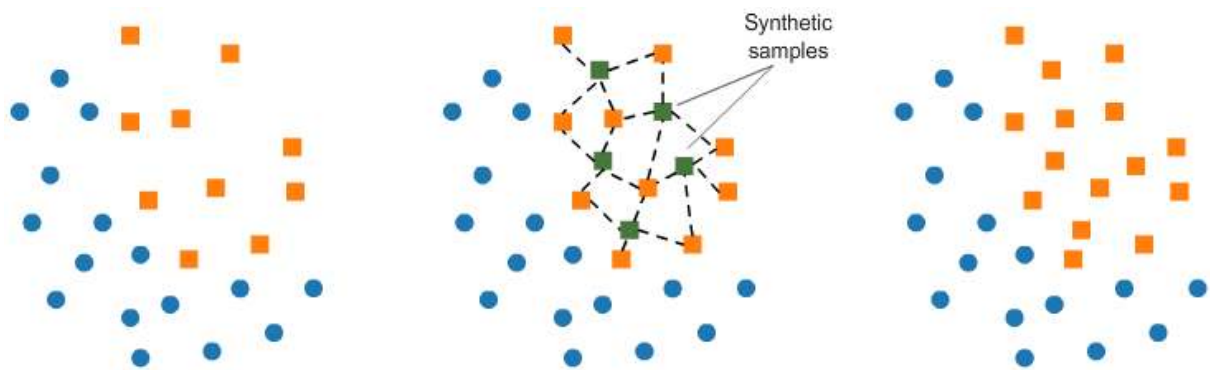
[표 4-1] 분석 데이터 셋의 클래스 불균형

총 127,742건의 데이터 셋은 수납자인 우량(8,124건)과 체납자인 불량(119,336건)은 6 : 94의 비율로 데이터 클래스의 불균형 문제를 초래함을 확인할 수 있다. 데이터에서 각 클래스의 개수가 현저하게 차이가 난 상태로 모델을 학습하면, 다수의 범주로 패턴 분류하는 문제가 생기고 이는 곧 모델의 성능에 영향을 끼치게 된다.



[그림 4-4] 오버 샘플링 적용

따라서 불균형 데이터(imbalanced data)의 문제를 해결할 수 있는 SMOTE(Synthetic Minority Oversampling Technique)을 통해 데이터의 개수가 적은 클래스의 표본을 가져온 뒤 임의의 값을 추가하여 새로운 샘플을 만들어 데이터에 추가하는 오버샘플링 방식으로 해결하고자 한다.



[그림 4-5] Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE의 오버샘플링은 Minor class의 instance를 단순히 복제함으로써 생기는 과적합(overfitting)을 최소화하기 위해 사용한다. 소수 클래스로부터 데이터의 subset을 뽑아낸 다음에 새롭게 비슷한 instance를 생성하며, 새롭게 만든 instance를 원래 데이터에 더해지고 최종 학습 데이터 셋으로 사용하였다. 따라서 synthetic example은 랜덤하게 오버샘플링하는 것보다 과적합(overfitting)이 일어날 가능성이 줄어들며 정보가 손실될 우려가 없다.

구분	우량 개수	불량 개수	우량 비율	불량 비율
클래스 불균형	8,124	119,618	6.4%	93.6%
클래스 균형 (SMOTE)	119,197	111,618	49.9%	50.1%

[표 4-2] 분석 데이터 셋의 클래스 불균형 해소 비교

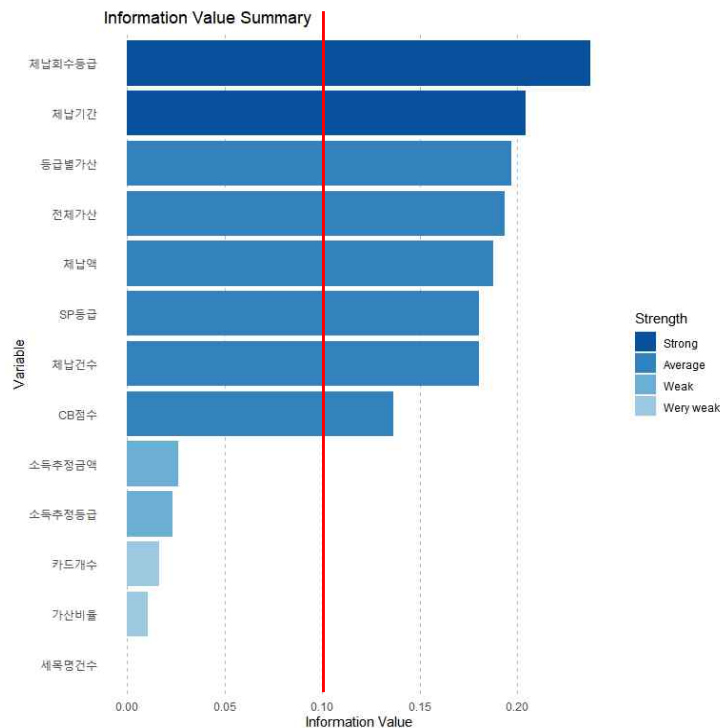
3) 체납액 이상치 확인

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10,000	12,870	25,740	152,947	51,480	376,882,410
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10,000	12,870	25,740	120,218	51,480	10,000,000

데이터의 범위가 큰 이상치는 학습 모델 입장에 영향을 미쳐, 이를 적합하게 처리해야 할 필요성이 있다. 따라서 체납 담당자와 인터뷰를 통해 체납액 ‘1,000만원 이상’과 ‘1만원 이하’를 이상치로 선정하여 분석 데이터 셋으로 구성하였다.

나. 변수 선택 과정 및 데이터 분할

1) Information Value 값



[그림 4-6] Information Value 그래프

Information Value는 정의된 우·불량('1' 또는 '0')값의 통계학적 계산으로 변수 채택 과정에서 가장 중요한 판단 기준이 된다. 보통의 경우 IV값이 '0.10' 또는 '0.15' 이상일 때 변수로서 의미가 있다고 판단한다. 따라서 채납회수등급, 채납기간, 등급별 가산, 전체가산, 채납액, SP등급, 채납건수, CB점수를 모형개발에 사용 변수로 채택하였다.

2) WOE 값 생성 및 확인

여부	채납액_woe	세목명건수_woe	채납건수_woe	CB점수_woe	소득추정등급_woe	소득추정금액_woe	채납회수등급_woe	SP등급_woe	카드개수_woe	채납기간_woe	가산비율_woe	등급별가산_woe	전체가산_woe
0	-0.43032825	0	-0.35621670	0.39657467	0.1261308	0.136852	0.4867017	0.4896812	0.33443476	0.6625554	-0.04205341	0.2826061	0.02309056
0	-0.43032825	0	-0.35621670	0.39657467	0.1261308	0.136852	0.4867017	0.4896812	0.33443476	0.6625554	-0.04205341	0.2826061	0.02309056
0	-0.43032825	0	-0.35621670	0.39657467	0.1261308	0.136852	0.4867017	0.4896812	0.33443476	0.1606114	0.24962544	0.2826061	-0.34576939
0	-0.43032825	0	-0.35621670	0.39657467	0.1261308	0.136852	0.4867017	0.4896812	0.33443476	-0.5293126	-0.04205341	0.2826061	0.02309056

[그림 4-7] WOE 값 확인 결과

실행결과에서 볼 수 있듯이 각 변수와 변수 내 특성이 반영된 구간에 해당하는 새로운 WOE 값 변수가, '변수명_woe'로 생성된 것을 확인할 수 있다.

3) Logistic Regression stepwise

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.004198   0.004318  -0.972   0.3310
체납회수등급_woe -0.607236   0.013861 -43.808 <2e-16 ***
체납기간_woe    -0.535281   0.012451 -42.992 <2e-16 ***
SP등급_woe     -0.424876   0.013934 -30.491 <2e-16 ***
체납건수_woe   -0.355609   0.019706 -18.046 <2e-16 ***
체납액_woe     -0.409431   0.016892 -24.237 <2e-16 ***
전체가산_woe    0.029179   0.014972   1.949  0.0513 .
등급별가산_woe  0.191674   0.015747  12.172 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 330820  on 238636  degrees of freedom
Residual deviance: 306262  on 238629  degrees of freedom
AIC: 306278

Number of Fisher Scoring iterations: 4

```

[그림 4-8] Stepwise 로지스틱 회귀모형 실행결과

앞 단계에서 IV 값을 기준으로 채택된 8개의 변수를 기준으로, 로지스틱 회귀모형을 산출하고 Stepwise 기법을 통해 다시 한번 변수의 유의미함을 파악하고자 진행하였다.

로지스틱 회귀의 Stepwise를 실행한 결과 'CB점수'의 변수를 제거하고, 남은 7개의 변수 중 P-value (0.05) 값을 기준으로 유의하지 않는 '전체가산' 변수를 제거하였다.

4) 데이터 정규화 및 분할



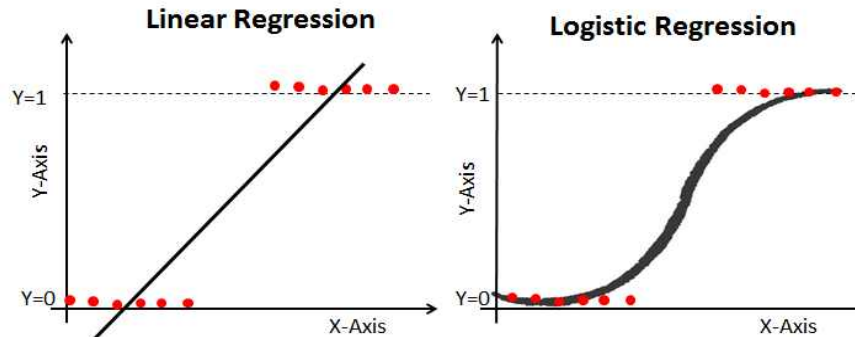
[그림 4-9] minmax정규화와 데이터 분할 구조도

데이터의 원활한 비교를 위해 minmax 정규화를 통해 모든 값들을 0부터 1까지 변경하였고, 이를 하나의 Data 셋에서 7:3의 비율로 Train과 test셋으로 분할하였다.

다. 예측 모델 형성

1) 로지스틱 회귀(Logistic Regression) 모형

□ 모델 설명



[그림 4-10] 로지스틱 회귀(Logistic Regression) 모형도

로지스틱 회귀모형은 종속변수와 독립변수의 관계를 설명한다는 점에서 회귀분석과 동일하지만, 종속변수가 범주형 변수여야 한다는 차이를 가지고 있다. 즉, 연속변수가 아닌 우·불량과 같은 이진 분류를 하는 기계학습 모형의 한 종류로, 납세자의 납세 및 체납여부인 우·불량을 분류하고자 한다.

□ 모델 구축

Call:

```
glm(formula = 여부 ~ 체납액_woe + 체납회수등급_woe + SP등급_woe + 체납건수_woe,
+ 체납기간_woe + 등급별가산_woe, family = binomial(), data = train_set)
```

Deviance Residuals:

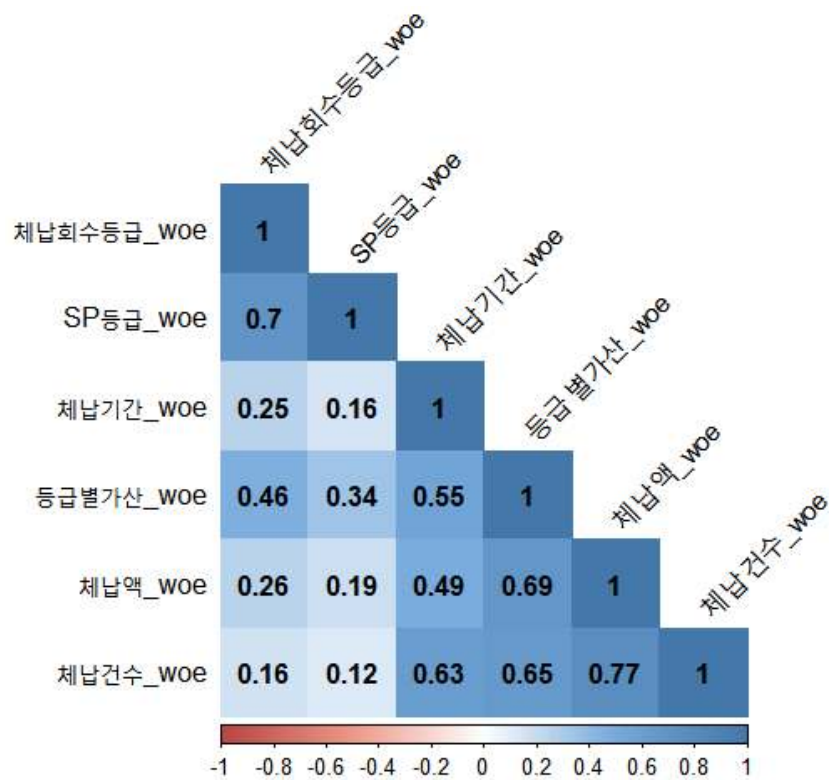
Min	1Q	Median	3Q	Max
-1.7458	-1.0727	-0.6018	1.0676	1.8760

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.07364	0.01300	82.570	<2e-16 ***
체납액_woe	-0.42668	0.02157	-19.785	<2e-16 ***
체납회수등급_woe	-0.84968	0.02220	-38.266	<2e-16 ***
SP등급_woe	-0.48401	0.01901	-25.460	<2e-16 ***
체납건수_woe	-0.45746	0.02849	-16.059	<2e-16 ***
체납기간_woe	-0.63132	0.01766	-35.740	<2e-16 ***
등급별가산_woe	0.23965	0.02535	9.453	<2e-16 ***

glm()를 사용한 로지스틱 회귀모형의 독립변수는 체납액, 체납회수등급, SP등급, 체납건수, 체납기간, 등급별가산 모두 P-value (0.05)를 통해 모든 변수가 유의하다는 것을 확인하고, 사전에 구축한 train data set으로 학습을 진행하였다.

□ 상관계수 및 그래프 도출



[그림 4-11] 채택변수의 상관계수 그래프 도출

최종 모형에서 채택된 6개의 변수를 대상으로 상관관계를 나타내는 상관계수를 확인하고 변수채택 여부를 최종 판단하였다.

채택변수의 상관계수 확인을 위해, 변수 간 상관계수가 0.8 이상의 변수가 없는 것으로 확인하였다. (일반적으로 상관계수가 0.8 이상이면 변수 간 강한 상관관계가 있음을 의미하여, 상관관계를 갖는 변수 중 하나를 제거한다.)

□ 다중공선성 확인

체납액	체납회수등급	SP등급	체납건수	체납기간	등급별가산
2.787	2.195	1.876	3.236	1.685	2.5940

[표 4-3] 로지스틱 회귀모형의 다중공선성 확인

다중공선성은 회귀분석에서 사용된 모형의 일부 예측 변수가 다른 예측 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상을 말한다. 따라서 `vif()`함수를 이용하여 문제가 되는 지수 10 이상이 없는 것으로 확인하였다.

□ Confusion Matrix and Statistics

		Reference	
		0	1
Prediction	0	22,380	12,485
	1	13,608	23,119

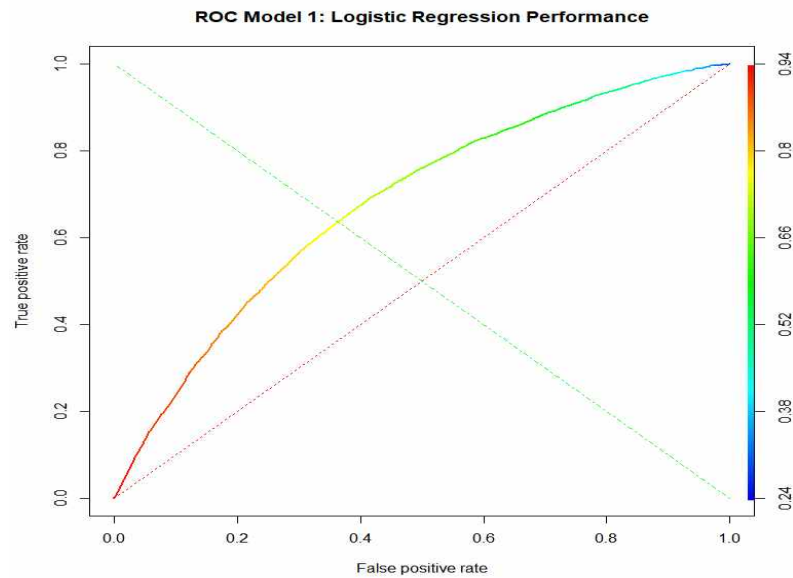
Confusion Matrix	값
Precision	0.6419
Accuracy	0.6355
Recall	0.6219
Specificity	0.6493
FP_rate	0.3126
F1 Score	0.6317
Kappa	0.2712

로지스틱 회귀모형의 오차 행렬(Confusion Matrix)은 납세자의 우·불량에 대한 분류기(classifier)의 성능을 평가하는 지표로, 예측과 실제로부터 평가를 다음과 같이 확인하였다.

오차 행렬(Confusion Matrix)

- ◆ Precision : Y로 예측된 것 중 실제로도 Y인 비율
- ◆ Accuracy : 전체 예측에서(예측이 Y든 N이든 무관하게) 옳은 예측의 비율
- ◆ Recall : 실제로 Y인 것들 중 예측이 Y로 된 경우의 비율
- ◆ Specificity : 실제로 N인 것들 중 예측이 N으로 된 경우의 비율
- ◆ FP Rate : Y가 아닌데 Y로 예측된 비율. 1-Specificity와 같은 값
- ◆ F1 Score : Precision과 Recall의 조화 평균, 시스템의 성능을 하나의 수치로 표현하기 위해 사용하는 점수로, 0~1사이의 값을 갖는다. Precision과 Recall 중 한쪽만 클 때보다 두 값이 골고루 클 때 큰 값을 가진다.
- ◆ Kappa : 코헨의 카파는 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 0~1 사이의 값을 가진다. 이는 우연히 일치할 확률을 제외한 뒤의 점수다.

□ 모형검정



[그림 4-12] ROC Model1 : Logistic Regression Performance

로지스틱 회귀모형의 ROC curve를 통해 적중확률(True Positive Rate)과 오경보확률(False Positive Rate)을 파악하고, 확인 가능한 주요 통계수치를 통해 실제 모형의 성능을 확인하였다.

Model	AUC	KS	Gini
Logistic Regression	68.08	27.6	36.16

AUC, KS, Gini 계수

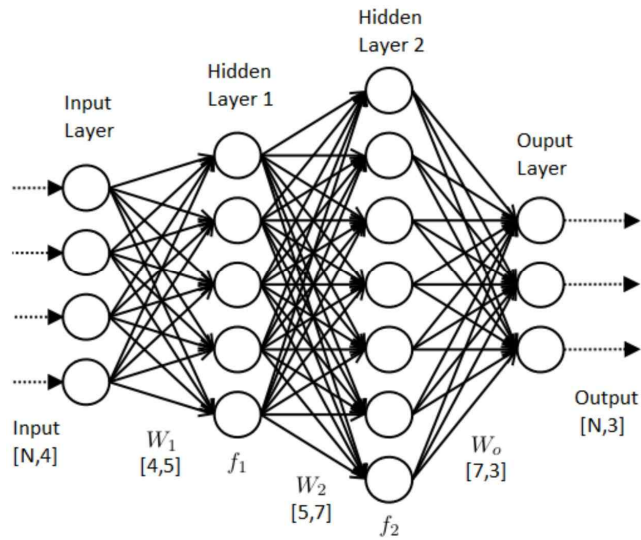
◆ AUC : ROC curve의 면적으로서 판별된 적중확률과 오경보확률의 계산을 통해 값을 산출한다. 1에 가까울수록 우수한 모형으로 판단한다.

◆ KS 통계량 : 우량집단과 불량집단의 누적분포의 차이를 나타내는 지표로 변별력 평가 시 주요 판별 통계량으로 활용한다. 50 이상이면 우수한 모형으로 판단한다.

◆ Gini 계수 : 불량누적분포를 이용하여 모형의 변별력을 판별하는 지표로 활용한다. 60 이상이면 우수한 모형으로 판단한다.

2) 신경망(Neural Network) 모형

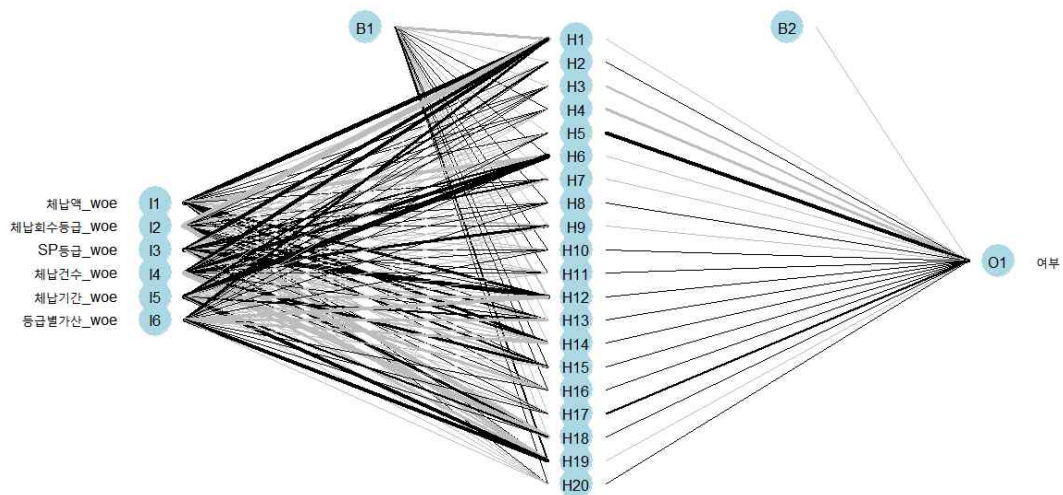
□ 모델 설명



[그림 4-13] 신경망(Deep Neural Network) 모형도

신경망 모형은 말 그대로 인체의 신경망(Neural Network)의 복잡계(Complex System)를 활용하여 학습모형을 만드는 것이다. 즉, 은닉층(Hidden Layer)과 각 연결 지점(Node)을 통해 신경망 모형을 산출하고 향후 관측값을 예측하는 기계학습 모형으로 납세자의 우·불량을 분류하고자 한다.

□ 모델 구축



[그림 4-14] 단층 은닉층으로 형성한 Neural Network

앞서 확인한 로지스틱 회귀모형과 마찬가지로 총 6개의 변수를 통해, 20개의 연결 지점(Node)을 갖는 1개의 은닉층(Hidden Layer)을 생성함으로써 신경망 모형이 완성됨을 확인할 수 있다.

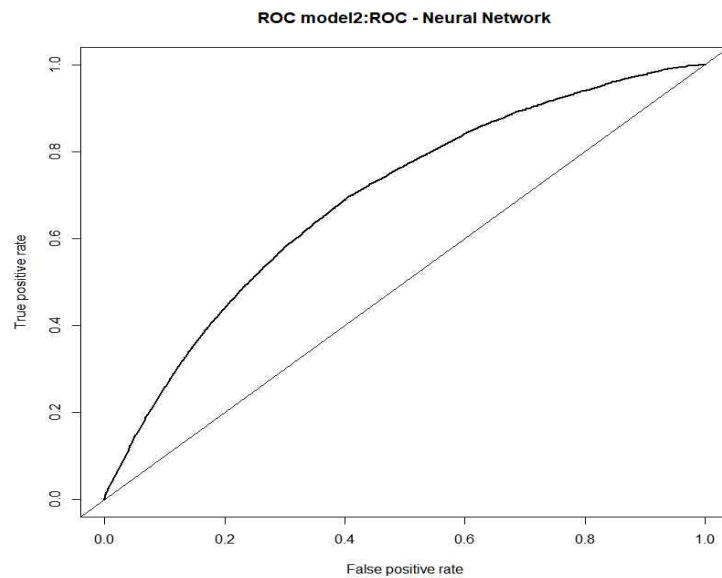
□ Confusion Matrix and Statistics

		Reference	
		0	1
Prediction	0	22,183	11,690
	1	13,805	23,914

Confusion Matrix	값
Precision	0.6549
Accuracy	0.6439
Recall	0.6164
Specificity	0.6717
FP_rate	0.3099
F1 Score	0.6351
Kappa	0.288

신경망 모형의 오차 행렬(Confusion Matrix)은 납세자의 우·불량에 대한 분류기(classifier)의 성능을 평가하는 지표로, 예측과 실제로부터 평가를 다음과 같이 확인하였다.

□ 모형검정



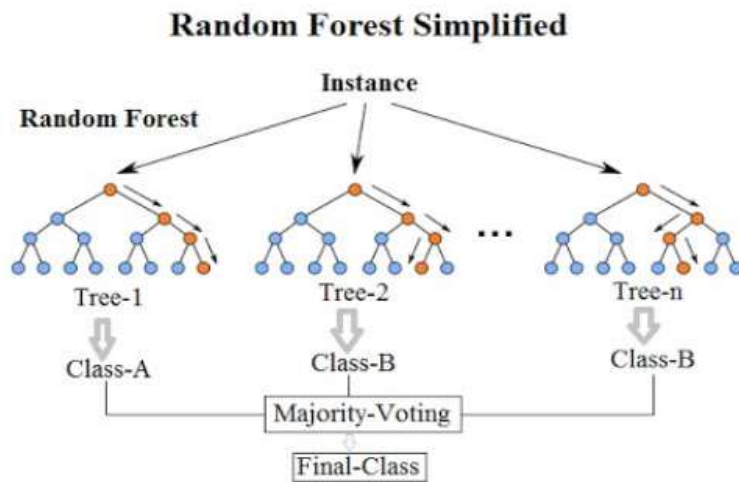
[그림 4-15] ROC Model2 : Neural Network Performance

Model	AUC	KS	Gini
Nerual Network	69.19	29.07	38.38

신경망 모형의 ROC curve를 통해 적중확률(True Positive Rate)과 오경보확률(False Positive Rate)을 파악하고, 확인 가능한 주요 통계수치를 통해 실제 모형의 성능을 확인하였다.

3) 랜덤 포레스트(Random Forest) 모형

□ 모델 설명



[그림 4-16] 랜덤포레스트(Random Forest) 모형도

랜덤포레스트 모형은 여러 개의 결정 트리들을 임의적으로 학습하는 다중 트리 모형으로서 앙상블 기법 중 한 가지이다. 데이터 상에서 확인 가능한 분류 및 세분화 학습을 통해 향후 관측값을 예측 및 분류하는 기계학습 모형이다.

□ 모델 구축

```
# randomForest 모델 생성
model2 <- randomForest(여부 ~ 체납액_woe + 체납회수등급_woe + SP등급_woe + 체
납건수_woe, + 체납기간_woe + 등급별가산_woe, data = train_set, importance =
TRUE)

# 중요도
importance(model2)
```

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
체납액_woe	78.463491	20.803870	133.57309	1918.122
체납회수등급_woe	46.106261	17.674363	83.08375	2840.584
SP등급_woe	53.285031	2.732736	93.86941	1083.851
체납건수_woe	14.104797	29.495217	78.70361	1202.839
체납기간_woe	-37.945940	100.710310	119.95666	1739.873
등급별가산_woe	1.955342	35.136715	71.02065	1353.891

randomForest()를 사용한 랜덤포레스트 모형으로 기존 독립변수들을 이용한 train data set으로 학습을 진행하였다. importance()를 통해 Random Forest 에서 가지를 칠 때 얼마나 정확도가 올라가는지에 대한 중요한 변수들을 나타내고 있다.

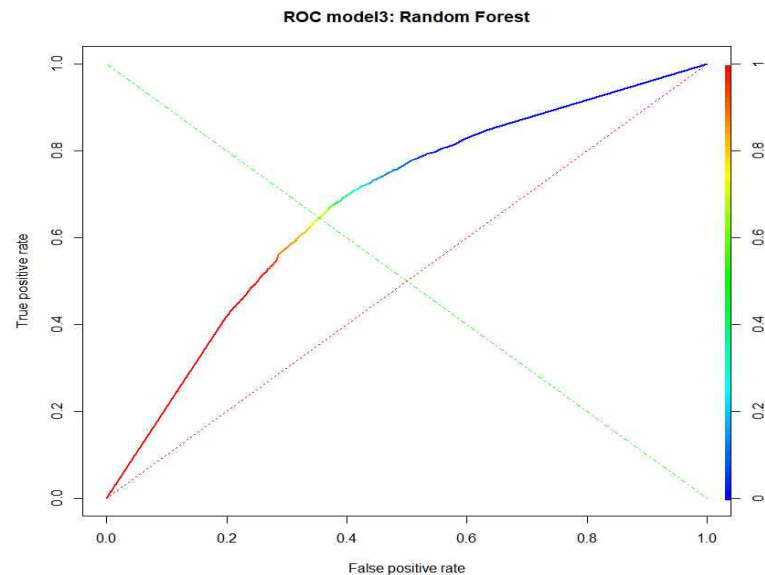
□ Confusion Matrix and Statistics

		Reference	
		0	1
Prediction	0	22,308	11,492
	1	13,680	24,112

Confusion Matrix	값
Precision	0.6600
Accuracy	0.6485
Recall	0.6201
Specificity	0.6771
FP_rate	0.3117
F1 Score	0.6394
Kappa	0.2971

랜덤 포레스트 모형의 오차 행렬(Confusion Matrix)은 납세자의 우·불량에 대한 분류기(classifier)의 성능을 평가하는 지표로, 예측과 실제로부터 평가를 다음과 같이 확인하였다.

□ 모형검정



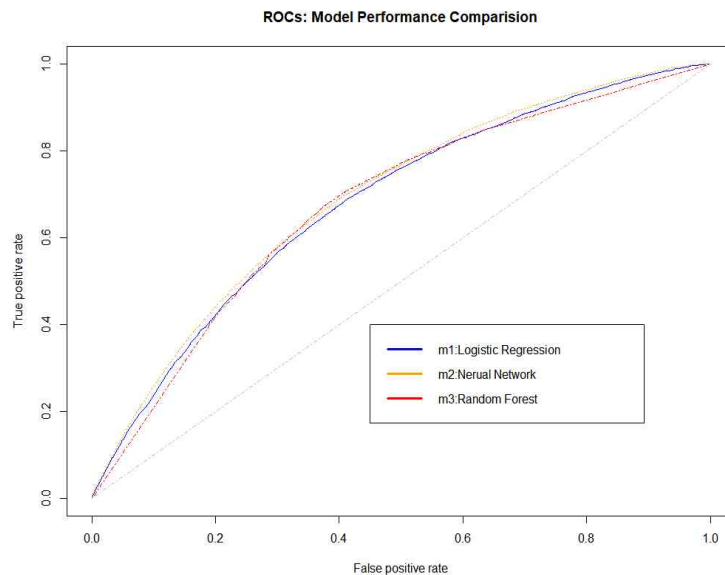
[그림 4-17] ROC Model3 : Random Forest

Model	AUC	KS	Gini
Random Forest	67.59	29.81	35.18

랜덤포레스트 모형의 ROC curve를 통해 적중확률(True Positive Rate)과 오경보확률(False Positive Rate)을 파악하고, 확인 가능한 주요 통계수치를 통해 실제 모형의 성능을 확인하였다.

4) 각 모형의 성능 비교

□ ROC curve



[그림 4-18] 각 모형의 ROC curve 비교

총 세 가지 모형을 개발하였고, 이러한 모형들의 주요 통계량을 비교하기 위해 세 가지의 ROC curve를 비교하였으나, 세 가지 모형 모두 성능에 있어서 눈에 띄는 차이를 보인다고 보기 어려움에 있다.

□ 주요 통계수치 확인

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

[그림 4-19] 각 모형의 통계수치 비교

ROC curve 비교를 통해 성능의 차이가 뚜렷하게 보이지 않았지만, AUROC, KS통계량, Gini 계수 비교를 통해 성능의 차이를 구분해 낼 수 있다. AUROC 기준으로 신경망 모형, 로지스틱 회귀 모형, 랜덤포레스트 모형 순으로 성능이 좋은 것을 확인할 수 있다.

그렇지만 성능의 차이가 뚜렷하게 나타나지 않고 회수 가능성 예측모형의 개발 이후 그 적용에 있어 모형 내 각 변수가 지닌 의미를 설명할 수 있어야 하므로 활용과 접근이 용이한 로지스틱 회귀모형으로 최종 모형으로 선택하였다.

□ Confusion Matrix 비교 확인

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

[표 4-4] 각 모형의 Confusion Matrix 비교

오차 행렬(Confusion Matrix)의 비교를 통해 랜덤포레스트 모형, 신경망 모형, 로지스틱 회귀모형의 순으로 성능이 좋은 것을 확인할 수 있다. 납세자의 우·불량에 대한 분류기(classifier)의 성능을 평가하는 지표로, 예측과 실제로부터 평가를 다음과 같이 비교할 수 있으며, 이를 해석하자면 다음과 같다.

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 4-5] 사전 예측 결과와 실제 납부 형태 비교 결과

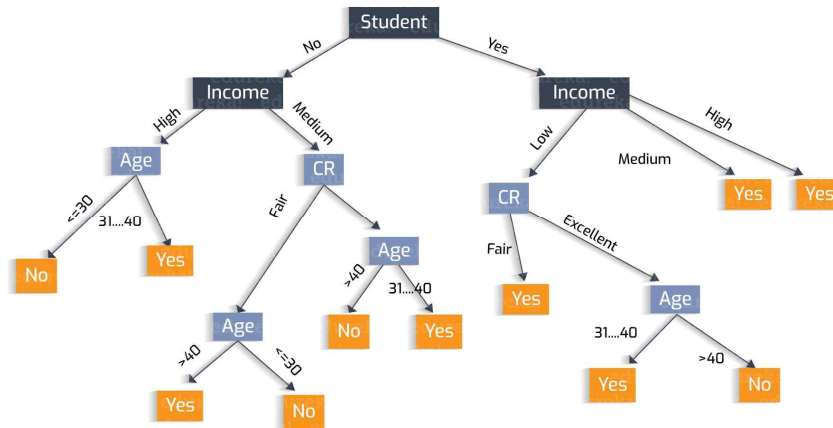
이번 분석에서 개발한 2020년 10월 고지서 발송대비 수납 기준으로 우·불량 대상자 예측과 실제 비교 모형 결과는 표와 같으며, 신경망(Nerual Network) 모형에서 64.39%으로 가장 높은 정확도를 보였다.

하지만 정확한 우·불량 예측을 하기에 한계가 있는 정확도이며, 체납회수등급과 같은 사용된 변수에 있어서 징수 활동을 하기 위해선 조금 더 세분화하고 정확도를 높일 수 있는 다양한 변수의 필요성이 있는 것으로 분석되었다.

라. 유형 분류 모델 형성

1) 의사결정 나무(Decision tree) 모형

□ 모델 설명



[그림 4-20] 의사결정 나무(Decision tree) 모형도

의사결정 나무(Decision Tree)는 각 데이터들이 가진 속성들로부터 패턴을 찾아내서 분류 과제를 수행할 수 있도록 하는 지도학습 머신러닝 모델이다. 해당 분석에서는 납세자를 대상으로 어떠한 변수들을 얼마나 이용하여 체납 또는 수납이 발생하였는지 대상자의 특성을 확인하고자 한다.

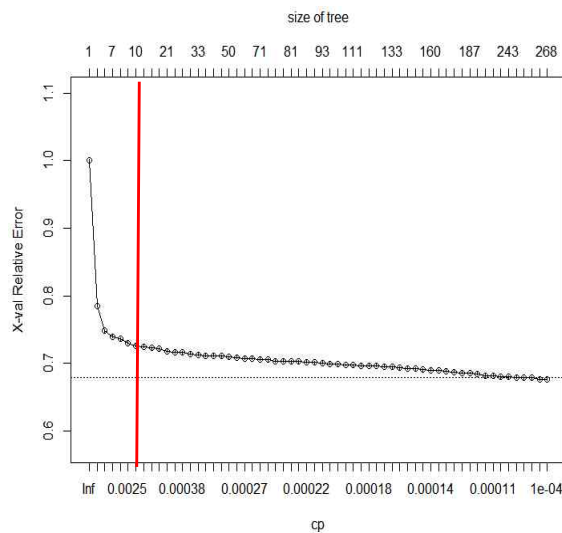
□ 사용 변수 선정

	체납회수등급	체납기간	SP등급	체납건수	체납액	여부
1	5	5	5	1	12870	0
2	5	5	5	1	12870	0
3	5	4	5	1	16500	0
4	5	1	5	1	12870	0
5	5	1	5	1	96560	0

[그림 4-21] 모형에 적합한 변수 선정

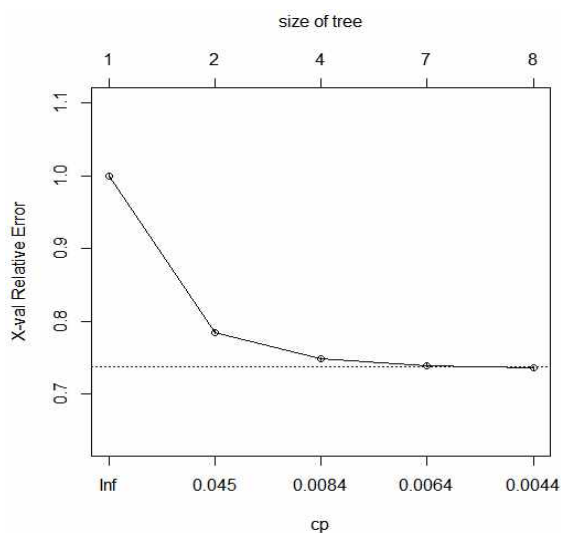
앞 단계에서 IV(Information Value) 값을 기준으로 채택된 8개의 변수에서 Stepwise 로지스틱 회귀와 이해관계자가 이해 가능한 수치 변수를 통해 최종 체납회수등급, 체납기간, SP등급, 체납건수, 체납액으로 총 5개의 독립변수를 선정하였다.

□ 최적 CP값 탐색



	CP	nsplit	rel error	xerror	xstd
1	0.2152831457	0	1.0000000	1.0000000	0.002052190
2	0.0092085364	1	0.7847169	0.7847169	0.002003184
3	0.0077381953	3	0.7662998	0.7484120	0.001985184
4	0.0052932280	6	0.7364729	0.7392539	0.001980177
5	0.0036464460	7	0.7311796	0.7359771	0.001978339
6	0.0017812132	8	0.7275332	0.7303142	0.001975105
7	0.0010922534	9	0.7257520	0.7259284	0.001972551
8	0.0007253683	10	0.7246597	0.7244245	0.001971664
9	0.0006889598	13	0.7224836	0.7226433	0.001970608
10	0.0006329468	14	0.7217947	0.7213410	0.001969831
11	0.0004200975	20	0.7176273	0.7179718	0.001967803
12	0.0004032936	21	0.7172072	0.7165182	0.001966919
13	0.0003596034	25	0.7151991	0.7161065	0.001966668

의사결정 나무(Decision tree)의 적절한 가지 수를 선택하기 위해 오차가 최소가 되는 지점의 복잡도를 확인하였다. 7번의 분기(nsplits)가 발생할 때 오류율(rel error)이 가장 적으므로 해당 복잡성(cp) 값을 선택하였다.



	CP	nsplit	rel error	xerror	xstd
1	0.215283146	0	1.0000000	1.0000000	0.002052190
2	0.009208536	1	0.7847169	0.7847169	0.002003184
3	0.007738195	3	0.7662998	0.7484120	0.001985184
4	0.005293228	6	0.7364729	0.7392539	0.001980177
5	0.003646446	7	0.7311796	0.7359771	0.001978339

그래프를 보면, 오른쪽으로 갈수록 오류율(rel error)은 적어지지만, 분기(nsplits)가 커지는 것을 확인할 수 있다. 따라서 오류율이 크게 감소한 기점까지의 포인트를 복잡성(cp) 값을 최종 선정하였다.

□ 모델 구축

```
# Decision tree 모델 생성
rpart_model <- rpart(여부 ~ 체납액 + 체납회수등급 + SP등급 + 체납건수 + 체납기간,
data = Decision_tree_data, control = rpart.control(cp=0.003646446))

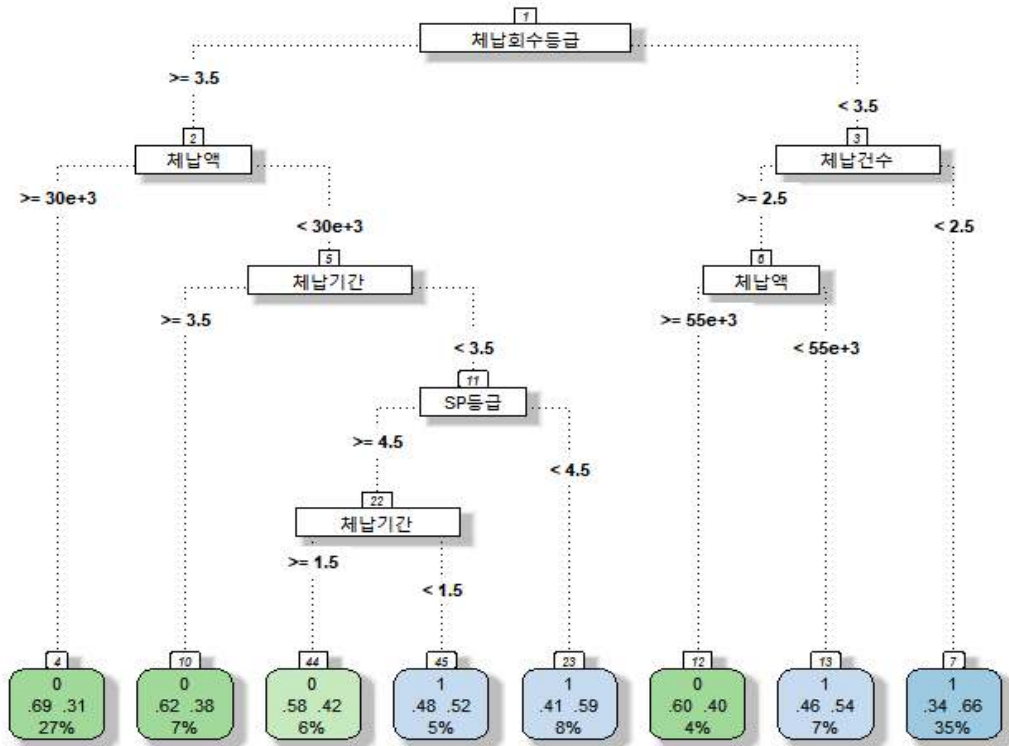
n= 238637

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 238637 119020 0 (0.5012509 0.4987491)
  2) 체납회수등급>=3.5 128722 51251 0 (0.6018474 0.3981526)
    4) 체납액>=30190 64073 20100 0 (0.6862953 0.3137047) *
    5) 체납액< 30190 64649 31151 0 (0.5181519 0.4818481)
      10) 체납기간>=3.5 16891 6351 0 (0.6240009 0.3759991) *
      11) 체납기간< 3.5 47758 22958 1 (0.4807153 0.5192847)
        22) SP등급>=4.5 27912 13102 0 (0.5305962 0.4694038)
          44) 체납기간>=1.5 15000 6331 0 (0.5779333 0.4220667) *
          45) 체납기간< 1.5 12912 6141 1 (0.4756041 0.5243959) *
        23) SP등급< 4.5 19846 8148 1 (0.4105613 0.5894387) *
  3) 체납회수등급< 3.5 109915 42146 1 (0.3834418 0.6165582)
    6) 체납건수>=2.5 27188 13086 0 (0.5186847 0.4813153)
      12) 체납액>=55330 10714 4261 0 (0.6022961 0.3977039) *
      13) 체납액< 55330 16474 7649 1 (0.4643074 0.5356926) *
    7) 체납건수< 2.5 82727 28044 1 (0.3389945 0.6610055) *
```

rpart()를 사용한 의사결정 나무(Decision tree) 모형으로 기존 독립변수들을 이용하여 유형 분석을 진행하였다. 최적 복잡성(cp) 값은 0.003646446으로 노드와 가지에 있어 체·수납 대상자의 유형 분류를 하는데 수월하였다. 이를 통해 2020년 10월 고지서 발송대비 수납 기준으로 의사결정 나무(Decision tree) 모형을 적용하여 4개의 체납 유형과 4개의 수납 유형을 확인하였다.

□ 결과 해석



[그림 4-22] 의사결정 나무(Decision tree) 모형을 통한 시각화

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%

[표 4-6] 체·수납 대상자 유형 구분

의사결정 나무(Decision tree) 모형을 이용하여 덴드로그램으로 시각화하고 이에 대한 해석을 진행하였으며, 체·수납 대상자를 8개의 유형으로 확인하였다.

5. 결론

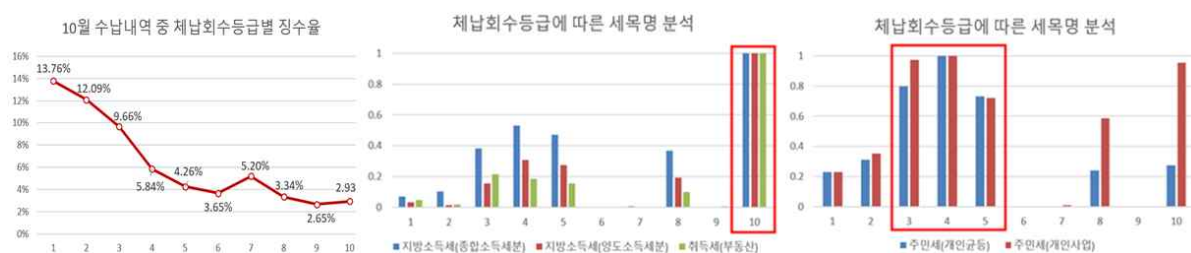
가. 분석결과 해석

1) 탐색적 데이터 분석(EDA)을 통한 징수현황 파악



[그림 5-1] 탐색적 데이터 분석(EDA) 결과

빅데이터 분석 기반인 ‘체납회수등급’을 10월 발송대비 수납내역을 활용하여 회수등급의 이해도와 활용도를 높이하고자 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 체납 징수 현황에 직관적으로 바라보는 과정을 진행하였다.



[그림 5-2] 탐색적 데이터 분석(EDA) 결과

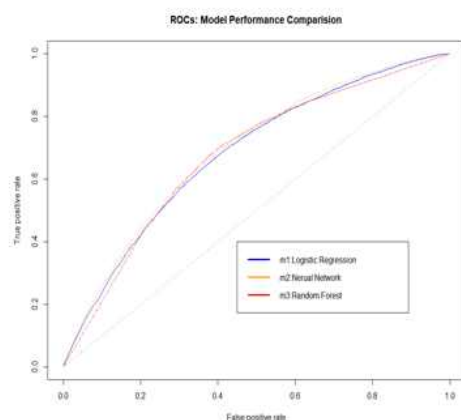
이를 통해 소액·고액 체납자에 대한 특성을 파악하였으며, 체납회수등급에 따라 징수율이 상이한 것을 확인하였다. 아울러 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 파악하지 못한 체납징수 현황에 대한 패턴을 발견하고, 이를 바탕으로 기존의 프로세스를 수정하거나 새로운 정책을 제안할 수 있을 것이다.

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 5-1] 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.



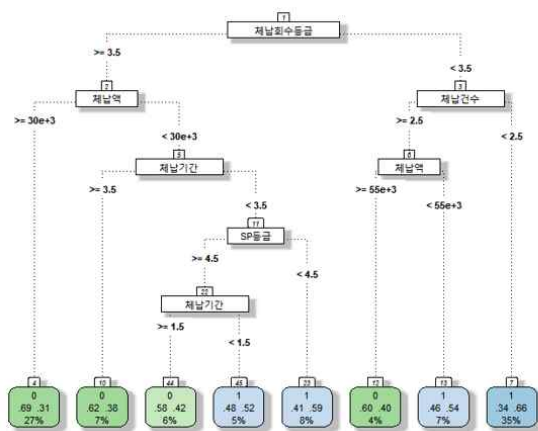
Model	AUC	KS	Gini
Logistic Regression	68.08	27.6	36.16
Nerual Network	69.19	29.07	38.38
Random Forest	67.59	29.81	35.18

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 5-3] 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, 체납기간이 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 5-4] 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 5-5] 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 5-6] 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 5-2] 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 대하여 회수등급이 1~3 등급일 경우에 SMS발송, 4~7등급인 경우에 예고, 8~10등급인 경우 압류하는 정책을 가정함

[그림 5-7] 비용 및 환수 금액 시뮬레이션 정의

단위 : 백만원								
체납회수등급	체납건수		체납금액		구분	기존방식	회수등급방식	비고
	압류X	압류O	압류X	압류O				
1	고	① 809건	1,768	165	①	SMS 발송	40,450 원	건당 SMS 비용 : 50원 건당 예고 비용 : 400원 건당 압류 비용 : 2,000원
2					②	예고 안내	1,099,200 원	
3					③	압류 건수	31,012,000 원	
4	중	② 2,748건	2,968	507	합계	31,012,000 원	26,889,650 원	
5					비용차이		(4,122,350 원)	
6					압류건수		15,506 건	
7	저	③ 12,875건	27,876	26,598	압류비용		26,889,650 원	
8					(조기)환수금액		27,270,206,670 원	
9								
10								

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

[그림 5-8] 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 체납 징수의 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 신용 정보를 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.

6. 부록

가. 참고문헌

- ◆ 행정안전부, 공공빅데이터 표준분석모델 매뉴얼(지방세 체납 징수), 2018
- ◆ 김은찬·유병준, 체납자 회수 가능성 예측모형 개발 및 개발 및 체납정보 시스템 개선 한국경영정보학회, 2018
- ◆ 박명호, 지방세 체납징수의 문제점과 개선방안 한국조세연구원, 2011
- ◆ 행정안전부, 지방세 통계연감(2019년 실적), 2020
- ◆ 정재안, 빅데이터 기반 체납 수용가 예측 모델 개발 배재대학교, 2019
- ◆ 전병진·김희웅, 공공 빅데이터 개방 및 활용 활성화 방안에 대한 연구 한국정보화진흥원, 2017