

발 간 등 록 번 호

11-1241140-100001-10



2025년 연구보고서

내검 분야의 AI 활용 기초연구: 해외 사례 중심으로

2026. 1.



<https://mods.go.kr/dsri>



국가데이터처
국가데이터연구원

연구보고서 2025-04

내검 분야의 AI 활용 기초연구: 해외사례 중심으로

조미현



Statistics Korea
Statistics Research
Institute

발간사

“데이터의 가치는 분석과 활용을 통해 의사결정을 지원하고, 혁신과 효율성 향상 등 구체적인 성과를 창출하는 데서 비롯됩니다.”

급변하는 불확실성의 시대에 데이터는 더 이상 단순한 숫자의 기록이 아니라, 미래를 예측하고 사회 문제를 해결하는 핵심 나침반으로 자리매김하고 있습니다. 국가데이터연구원은 이러한 시대적 요구에 부응하여 국민의 삶을 실질적으로 개선하고 AI 기반의 공공 AX 대전환을 뒷받침하기 위한 데이터 기반 연구에 지속적으로 매진해 왔습니다.

2025년 연구보고서에는 우리 사회가 직면한 환경 변화에 능동적으로 대응하고자 첨단 기술을 국가통계에 접목하기 위해 치열하게 고민한 연구 성과를 담았습니다.

첫째, 인공지능(AI) 기반 국가통계 기술혁신을 선도하고자 노력하였습니다.

생성형 AI 기술을 현장조사에 적용하기 위한 기초연구를 통해 조사자료의 내용검토 및 자동분류, 질의응답에 활용 가능성을 모색하였으며, 이는 통계 생산의 신속성과 정확성을 획기적으로 제고하는 토대가 될 것입니다. 아울러 생성형 AI를 활용한 나우캐스트 지표 서비스 제공 방안 연구는 통계서비스의 새로운 가능성을 여는 의미 있는 첫걸음이라 할 수 있습니다.

둘째, 점차 열악해지고 있는 조사환경에 대응하기 위해 새로운 통계방법론 연구와 국가통계 품질제고를 위한 연구를 강화하였습니다.

확률표본과 자원자표본을 통합한 추정 방안 연구는 응답자 조사 부담을 완화하고 비확률표본의 병행 활용 가능성을 제시하였으며, 데이터 과학기술을 활용한 자료수집 개선 연구와 데이터 통합방법 연구는 다양한 데이터의 연계·통합 방법을 보다 체계화하였습니다.

셋째, 사회적 사각지대를 조명하고 지속가능한 미래를 지원하기 위한 데이터 기반 정책 연구에 집중하였습니다.

최근 심각한 사회 문제로 대두된 ‘고립·은둔 청년’의 실태 파악을 위한 조사 문항 개발 연구를 비롯하여, 돌봄 분야 국가통계 활용 방안과 국내 최초의 기후변화 통계·지표 분석 연구는 데이터가 사회안전망 강화에 기여할 수 있음을 보여줍니다. 또한 소득이동통계 심층 분석 연구와 생애과정 이행에 대한 중·고령기 비교 연구는 관련 정책의 실효성과 활용도를 한층 높일 것으로 기대됩니다.

아울러 가계동향조사의 소비지표 작성 연구와 퇴직연금 적립금 배분 방법 연구는 국민의 체감 경기를 보다 정확히 진단하고 합리적인 경제정책 수립을 지원하는 든든한 기반이 될 것입니다.

2025년 10월부터 새롭게 출발한 국가데이터처 국가데이터연구원은 앞으로도 최신 기술과 사람을 잇는 데이터 연구를 통해 국가통계의 지평을 지속적으로 확장해 나가겠습니다.

본 연구보고서가 통계 생산자와 이용자 모두에게 실질적인 도움이 되고, 각계각층의 의사결정자에게 깊이 있는 통찰을 제공하기를 기대합니다.

많은 관심과 성원을 부탁드립니다.

2026년 1월

국가데이터연구원장

가진

목 차

제1장 서론	1
제1절 연구 배경 및 개요	1
제2절 개념 정의	3
제2장 해외 AI 기반 내검 연구사례	4
제1절 편집: 영국 통계청	4
제2절 대체: 벨기에 VITO	19
제3장 요약 및 시사점	29
제1절 요약	29
제2절 시사점	33
참고문헌	34
Abstract	35

요 약

최근 통계 분야에서 머신러닝 활용에 대한 관심이 빠르게 증가하고 있다. 많은 국가 통계기관 및 국제기구에서 공식통계의 품질을 향상시키기 위한 머신러닝 활용 방안을 탐색하고 있는데, 그 일환으로 UNECE 산하 HLG-MOS에서 ‘머신러닝 프로젝트(2019)’를 출범하였고, 다양한 분야에서 머신러닝을 적용한 연구 결과를 정리하여 「공식 통계를 위한 머신러닝(Machine Learning for Official Statistics, 2021)」 보고서를 발간하였다.

본 연구는 위 보고서를 참고하여 HLG-MOS에서 수행한 ‘편집 및 대체’ 분야에 머신러닝을 활용한 연구 과정 및 결과를 정리하고, 추후 국가데이터처 작성 통계 내검에 AI 적용 여부를 검토하기 위한 기초연구를 수행하는 데 목적을 두고자 한다.

먼저, UNECE 보고서상 정의에 따라 ‘내검’의 유형을 ‘편집’과 ‘대체’로 나누었고, 각각 1가지씩 총 2가지 연구 사례를 정리하였다. 편집 부문에서는 영국 통계청 ‘생계비 및 식료품 조사’ 연구 결과를 다루었고, 대체 부문에는 벨기에 VITO ‘에너지 균형통계’ 연구 결과를 수록하였다. 연구 결과, 머신러닝의 편집 및 대체 결과를 원본 데이터 및 기존 처리 방법 결과와 비교해 봤을 때, 머신러닝 분류 및 예측 결과가 꽤나 우수한 것으로 나타났음을 알 수 있었다.

이후 연구별 주요 내용을 표로 정리하고 요약하였고, 시사점으로 머신러닝을 내검에 적용했을 때의 장점과 단점, 향후 과제에 대해 언급하며 마무리하였다. 머신러닝을 실무에 적극적으로 활용하기 위해서는 머신러닝을 활용한 내검 방식이 기존 방법보다 효율성, 신뢰성 측면에서 더 뛰어남을 증명해야 하고, 이를 평가하기 위한 기준 지표에 대해서도 고민이 필요하다.

주요 용어: AI, 내검, 편집, 대체, 에디팅(Editing), 임putation(Imputation)

제 1 장

서 론

제1절 연구 배경 및 개요

현재 사회의 디지털화로 인해 다양한 출처에서 수많은 데이터가 발생하고 있다. 이에 따라 인간이 분석하거나 이해하기 어려울 정도로 거대한 양의 데이터를 보유한 분야에서는 데이터 분석에 인공지능 활용을 확대하고 있다.

통계 분야에서도 마찬가지로 머신러닝에 대한 관심이 빠르게 증가했다. 머신러닝은 직접 프로그래밍하지 않고 컴퓨터에게 학습할 능력을 부여하는 학문으로, 통계·컴퓨터공학·인공지능과 밀접한 관련이 있다. 머신러닝은 특정 프로세스를 자동화하거나 프로세스 수행 능력을 향상하는 데 도움을 주어 통계 생산성을 효율적으로 증가시킬 수 있다. 또한, SNS 데이터, 이미지 등 새로운 유형의 데이터를 사용할 수 있게 한다.

현재 많은 국가 통계기관 및 국제기구는 신뢰할 수 있는 정보와 빠르게 접근 가능한 기술에 대한 수요가 증가하는 환경에서 공식 통계의 품질을 향상시키기 위한 머신러닝 활용 방안을 탐색하고 있다. 그동안 통계기관들은 공식 커뮤니티를 통해 지식 및 경험을 공유함으로써, 공통 머신러닝 솔루션 개발을 위해 협력해 왔다.

2019년 3월, UNECE 산하 ‘공식통계 현대화 고위급 그룹(High-Level Group on Modernisation of Official Statistics, 이하 HLG-MOS)’은 공식통계에 머신러닝 활용을 확대 또는 발전시키고, 각 통계기관에서 머신러닝 활용을 통해 얻은 교훈을 공유하기 위해 ‘머신러닝 프로젝트’를 출범하였다. 이 프로젝트는 머신러닝의 가치를 입증하는 것을 목표로, 통계 생산 과정에 머신러닝을 적용할 때 발생하는 공통적인 과제들을 해결하고자 했다.

UNECE HLG-MOS는 머신러닝 활용이 사람이 수행하던 프로세스를 자동화하고 업무 효율성을 향상시킴으로써, 통계 발전에 기여할 수 있다고 간주하였다. 또한, 통계 생산 과정에서 머신러닝을 활용함으로써 그동안 사용하지 않았던 새로운 데이터 소스를 활용하여 통계의 연관성과 적시성을 향상할 수 있다고 여겼다. 또한, 머신러닝을 통해 데이터 세트 안에서 처리되는 레코드 수를 늘릴 수 있으며, 자동화 방식을 통해 정확하고 빠른 결과를 제공함으로써 통계 생산 프로세스를 개선할 수 있다

고 생각했다.

HLG-MOS 머신러닝 프로젝트에 따르면, ‘①텍스트 데이터 분류 및 코딩, ②편집 및 대체, ③이미지 분석’ 3가지 분야에서 예비 연구가 수행되었고, 2021년 해당 연구 결과를 「공식 통계를 위한 머신러닝(Machine Learning for Official Statistics)」 보고서에 수록하여 발간하였다. 본 연구보고서에서는 이 중에서 ‘②편집 및 대체’ 연구 결과를 자세히 다루고자 한다.

통계기관에서는 설문조사, 행정자료, 웹 스크래핑(Web Scraping) 등을 통해 수집된 데이터 안에 문제가 있거나 누락된 값을 식별하고 처리해야 한다. 편집과 대체는 통계 생산 과정에서 데이터의 품질을 보장하는 중요한 단계이다. 그동안 통계기관에서는 해당 작업을 수행하기 위해 다양한 방법을 활용해 왔다. 예를 들어, 규칙에 따라 데이터 허용 범위의 충족 여부를 확인하는 규칙 기반 방법, 분산 비교 등을 통해 의심되는 값을 감지할 수 있다. 이때 타당하지 않다고 식별된 데이터는 제거 또는 대체 처리된다.

HLG-MOS 머신러닝 프로젝트에서는 편집 및 대체 과정에 머신러닝 방법의 단독 적용을 연구하는 것이 아니라, 기존 프로세스(통계적 방법, 수작업 등) 적용 결과에 머신러닝 적용 결과를 비교하는 방식으로 연구 과정을 서술하고자 하였다. 본 연구는 HLG-MOS에서 수행한 편집 및 대체에 머신러닝을 적용한 연구 과정 및 결과를 정리하고, 추후 국가데이터처 작성 통계 내검에 AI 적용 여부를 검토하기 위한 기초 연구 수행에 목적이 있다.

제1장 제1절 연구 배경 및 개요 부분에서 통계 분야의 머신러닝 활용 확대, UNECE HLG-MOS 머신러닝 프로젝트 소개, 머신러닝을 적용한 편집 및 대체 연구 개요, 본 연구의 목적에 대해서 언급하였다.

그리고 제1장 제2절에서는 본 연구에서 다룬 내검의 유형을 ‘편집’과 ‘대체’로 구분하고, 개념 정의를 명시하였다.

제2장에서는 본격적으로 해외 AI 기반 내검 연구 사례를 다뤘다. 제1절 편집 부분에서는 영국 통계청의 연구 결과를 다루었고, 제2절 대체 부분에서는 벨기에 VITO의 연구 결과를 수록하였다.

마지막으로 제3장 제1절에서는 연구별 주요 내용을 표로 정리하고 요약하였다. 그리고 제2절 시사점 부분에서 머신러닝을 내검에 적용했을 때의 장점과 단점, 앞으로의 과제에 대해 언급하며 마무리하였다.

제2절 개념 정의

다음 장에서는 UNECE HLG-MOS에서 수행한 해외 AI 기반 내검 연구 사례에 대해 자세히 다루고자 한다. 연구 내용을 본격적으로 서술하기 전, 관련 개념을 명확하게 정의하고자 한다.

통계 작성 과정에서 본격적인 분석 작업을 실행하기 전, 통계 담당자는 데이터 세트에 잘못된 값이나 누락된 값이 있는지 반드시 검토해야 하며, 필요하다면 새로운 값을 추정하여 삽입해야 한다. 현재 실무적으로 진행되는 이 일련의 과정을 ‘내검’이라고 부르고 있다.

본 연구에서는 UNECE 「공식 통계를 위한 머신러닝(Machine Learning for Official Statistics)」 보고서상 정의를 반영하여, 내검을 두 가지로 분류하기로 한다.

먼저, 데이터 세트 안에서 누락되거나 문제가 있는 값을 식별하는 작업을 ‘편집(Editing)’이라고 정의한다. 또한, 문제가 있다고 식별된 값을 더 나은 값 또는 올바른 값으로 대체하여 삽입하는 작업을 ‘대체(Imputation)’라고 정의한다. 그리고 편집과 대체 두 가지 개념을 포괄하여 ‘내검’이라고 정의한다.

<표 1-1> UNECE ‘편집 및 대체’의 개념 정의

분류	내검	
	편집(Editing)	대체(Imputation)
개념	· 데이터 세트의 누락 또는 문제가 있는 데이터를 식별하는 작업	· 잘못된 값을 수정하거나 누락된 값을 삽입하는 작업
대상	· 이상값(Outlier), 레코드 내 모순	· 편집 과정의 식별된 값, 무응답 등

제 2 장

해외 시 기반 내검 연구사례

제1절 편집: 영국 통계청

머신러닝 모델은 이전에 통계 작성자가 분석한 데이터를 스스로 학습하여 일정한 규칙을 발견할 수 있다. 이러한 머신러닝의 특성을 편집 과정에 활용할 수 있는데, 머신러닝 모델은 이전에 수행된 편집 결과를 스스로 학습하여, 데이터 세트 안에서 문제가 되었던 레코드 규칙을 습득할 수 있다. 즉, 머신러닝 모델이 데이터 세트의 모든 레코드를 ‘타당함’ 또는 ‘타당하지 않음’으로 자동 분류하고, 분류 방법의 기준이 되는 규칙을 추출할 수 있다.

또한, 시간이 지나 통계 작성자가 바뀌어도 관련 지식을 보존하고 해당 방법을 공식화할 수 있으며, 성능을 지속적으로 개선할 수 있다. 따라서 통계 작성자가 내검 과정에 시간과 노력을 아낄 수 있어, 보다 중요하거나 영향력이 큰 레코드에 집중할 수 있다는 장점이 있다.

머신러닝은 통계적 규칙이 없는 편집 과정에도 사용될 수 있다. 비지도형 머신러닝¹⁾ 모델에서 데이터에 대한 선형적 모델(Priori Model)이 없더라도, 데이터의 은닉 구조(Hidden Structure)를 분석하는 데 머신러닝을 사용할 수 있다. 즉, 규칙이 없는 편집 과정에서 머신러닝을 활용함으로써, 직감 또는 규칙을 통해서 찾기 어려운 ‘문제 있는 셀’을 감지할 수 있다.

UNECE HLG-MOS에서 발간한 「공식 통계를 위한 머신러닝(Machine Learning for Official Statistics)」 보고서에서는 머신러닝을 편집에 적용한 연구 2가지(이탈리아 Istat, 영국 통계청)를 수록하였는데, 둘 중 영국 통계청의 연구 과정 및 결과를 자세히 설명하고자 한다.

1) 입력값만 있는 훈련 데이터를 이용하여 입력들의 규칙성을 찾는 머신러닝 방법. 즉, 훈련 데이터(Training Data)를 사용하여 출력 없이 입력만 제공되는 상황을 문제(입력)의 답(출력)을 가르쳐 주지 않는 것에 비유해 ‘비지도형 머신러닝’이라고 한다(출처: 한국정보통신기술협회).

1. 연구 개요

2020년 영국 통계청에서는 ‘생계비 및 식료품 조사 소득 데이터(Living Cost and Food Survey Income Data)’에 머신러닝을 활용한 편집 관련 예비 연구를 수행하였다. 기존 영국 통계청에서는 데이터의 허위 레코드를 통계 작성자가 직접 감지하여 일일이 수작업으로 처리해 왔다. 이 연구의 목적은 ‘생계비 및 식료품 조사(Living Cost and Food Survey, 이하 LCF)’에서 편집 처리가 필요한 레코드를 식별하는 작업에 머신러닝을 적용할 수 있는지 검토하는 것이다.

LCF 조사는 ‘생활 환경 조사(Survey of Living Conditions, 이하 SLC)’ 및 ‘부 및 자산 조사(Wealth and Asset Survey, 이하 WAS)’와 결합되어 최종적으로 ‘가계금융조사(Household Financial Survey, 이하 HFS)’를 형성한다. 따라서 이번 LCF 머신러닝 편집 연구를 통해, 궁극적으로 HFS 조사를 위한 머신러닝 솔루션을 구축하는 것이 영국 통계청의 최종 목표라고 할 수 있다.

<표 2-1> 영국 통계청 LCF, SLC, WAS 조사 비교

	LCF (Living Cost and Food Survey)	SLC (Survey of Living Conditions)	WAS (Wealth and Asset Survey)
조사 주제	지출, 식료품 및 영양	생활환경	부 및 자산
가구 수	5,000	12,000	1,000
기존 편집 과정	전체 가구 수기 편집	스크립트로 이상값(Outliers) 탐지 후 수정	

<표 2-1>에 따르면, 현재 각 조사별 가구 수는 LCF 5,000가구, SLC 12,000가구, WAS 1,000가구로 SLC에서 가장 많은 가구를 조사하고 있다. LCF는 지출, 식료품, 영양 관련 조사를 하고 있으며, SLC는 생활 환경, WAS는 부 및 자산을 주제로 한 통계조사로, 이 3가지 조사는 HFS 조사의 하위 표본으로 되어 있다.

각 조사의 기존 편집 과정을 살펴보면, LCF는 영국 통계청(Official National Statistics UK, 이하 ONS) 통계 작성자가 수작업으로 전체 가구 레코드에 대해 편집과 대체를 수행한다. 반면, SLC와 WAS는 소득 데이터의 이상값(Outliers)을 스크립트로 탐지한 후, 탐지된 레코드에 대해 통계 검증 담당자가 편집과 대체를 수행한다. LCF는 전수 편집 과정을 거치기 때문에 편집 작업량이 과도하여, 시간이 많이 소요된다는 단점이 있다.

LCF 소득 데이터에 SLC 스크립트 기반의 이상값 탐지 과정을 거친 결과, LCF 수기 편집 담당자가 수행한 총 변경 내역의 10%만 변경되었다. 이에 따라 SLC와 WAS

에 수행된 스크립트 기반 이상값 탐지 시스템의 정확성에 의문이 제기될 수 있으며, SLC와 WAS 데이터가 충분히 편집 처리되지 않았을 가능성을 시사한다.

LCF의 수기 편집된 데이터는 전체 데이터를 대상으로 하기에 우리가 가지고 있는 ‘골든 데이터 세트(Golden Data Set)’ 또는 ‘실제 값(Ground Truth)’과 가장 가깝다고 할 수 있다. 따라서 본 연구에서는 LCF 데이터로 해당 연구를 수행하는 것이 가장 적합하다고 간주하였다.

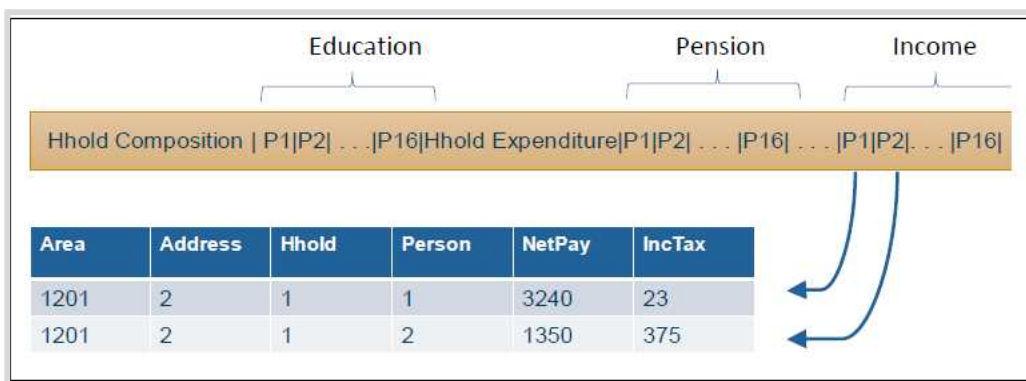
즉, LCF 조사에서 원본 조사 데이터와 수기 편집 및 대체된 데이터 모두 이용 가능하기 때문에, 영국 통계청에서 연구 범위를 LCF 소득 데이터에 한정하여 연구를 수행했다고 볼 수 있다.

2. 연구 과정

가. 입력 데이터

영국 통계청의 궁극적인 목표는 HFS 조사에서 편집 과정에 머신러닝을 적용하는 것이지만, 이번 연구에서는 편집 및 대체된 데이터와 원본 조사 데이터가 모두 제공되는 LCF 조사 데이터를 사용하였다.

영국 통계청은 LCF 데이터에 수기로 편집 및 대체 처리한 데이터 변경 사항을 라벨링(Labeling)하고, 이를 머신러닝 학습에 사용하였다. LCF 통계조사 담당자는 수기 편집 및 대체 과정에서 이루어진 변경 사항 중 약 80%가 정확하다고 추정하는데, 머신러닝을 활용하더라도 최소 해당 수준의 정확도가 유지되어야 한다고 주장한다.



<그림 2-1> LCF 조사 가구 데이터 레코드 구조

LCF 레코드의 구조는 위 <그림 2-1>에 나타나 있다. 위 구조는 주제별 설문 문항과 데이터 블록으로 구성되어 있다. 블록은 최대 16명의 가구 구성원 배열로 되어 있으며, 소득(Income), 연금(Pension) 및 교육(Education) 블록 등이 있다.

연구에 사용된 데이터 세트는 원본 데이터에서 추출되었다. 원본 데이터는 조사 대상자가 통계청에 제출한 데이터로 수기 편집 및 대체된 2018년 2분기(이하 8Q2)와 3분기(이하 8Q3)의 데이터로 구성된다. 이로써 4개(원본 8Q2와 8Q3, 편집 또는 대체된 8Q2와 8Q3)의 개인 소득 데이터 세트가 CSV(Comma Separated Values) 파일 형식으로 생성되었다.

본 연구에서는 2,912개 레코드를 포함한 8Q2 LCF 데이터를 테스트 데이터(Test Data)로, 3,059개의 레코드를 포함한 8Q3 데이터를 훈련 데이터(Training Data)로 사용하였다.

나. 데이터 준비

영국 통계청에서는 머신러닝 모델을 훈련하고 성능을 향상시키기 위해 다양한 특성 변수들을 사전에 검토했는데, 이를 ‘데이터 준비’ 과정으로 볼 수 있다. 즉, 본격적인 머신러닝 모델 훈련 전, 영향이 큰 특성 변수를 골라내기 위한 사전 검토 과정을 거쳤다고 할 수 있다.

변수 검토 과정에서 모든 조사 변수(Survey Variables)가 아닌 일부 변수가 모델에 영향을 미치는 것으로 보고, 이를 탐색해 보고자 했다. 그 과정에서 총 2,000개의 특성(조사 변수) 중 다음 영역의 91개 숫자형 또는 범주형 변수가 선택되었다.

- a. 소득 및 세금
- b. 교육
- c. 가족 상황
- d. 직업 및 부업의 소득과 세금
- e. 행복과 웰빙
- f. 취미용품, 의류, 신발 구매 가능성

조사 변수에는 숫자형 또는 범주형 특성 변수만 사용되지만, ‘모른다’, ‘거부’, ‘적용되지 않음’과 같은 값이 존재하는데, 해당 값들을 ‘-1’로 대체하여 처리하였다.

범주형 특성의 One-Hot-Encoding(이하 OHE) 처리를 하는 과정에서, 범주형 특성의 일부 옵션이 8Q2 레코드에서만 사용되고 8Q3 레코드에서는 사용되지 않거나, 혹은

그 반대인 경우가 있다. 원활한 분석을 위해 범주형 특성의 모든 값은 두 데이터 세트 모두에서 OHE 특성으로 표현되어야 한다. 그렇지 않으면 훈련 데이터와 테스트 데이터 세트 간 특성 변수의 갯수와 이름이 일치하지 않을 수 있기 때문이다.

예를 들어, NetPd(급여 지급기간) 특성 변수는 응답자가 마지막으로 받은 급여가 포함된 기간을 나타내기 위해 선택할 수 있는 15가지 옵션을 뜻하는데, 옵션 8은 급여를 연간 8회 받음을 나타낸다. OHE 처리 과정에서 해당 옵션이 선택된 레코드가 하나라도 있으면 NetPd_8이라는 특성이 생성된다. 8Q2 데이터 세트에서 이 옵션을 선택한 응답자가 없으면 NetPd_8은 생성되지 않는다. 그러나 8Q3 데이터에서 최소한 한 명이라도 이 옵션을 선택한 경우, NetPd_8이 생성되고 훈련 데이터와 테스트 데이터 간 특성 불일치가 발생할 수 있다.

순수급여(Net Pay) 및 총급여(Gross Pay)를 연간 금액으로 변환하는 과정에서 해당 특성들은 급여기간(Net Period)과 쌍을 이룬다. 즉, NetPay(순수급여) 값은 해당 급여가 발생한 기간을 설명하는 특성 NetPd(급여기간)와 함께 제공된다. 예를 들어, 응답자가 순수급여가 2달 동안 £1,800(파운드)라고 말했다면, 이를 연간화한 값은 $6 \times £1,800 = £10,800$ 로 계산할 수 있다. 또한, 기존 변수 중 삶의 질 관련 4가지(만족도, 가치, 행복, 불안)를 ‘웰빙(Wellbeing)’이라는 새로운 특성 변수 1개로 집계하였다.

머신러닝 모델을 활용하여 본격적으로 편집 및 대체 작업을 수행하고 변경 전후를 비교하기 위해, 해당 연구에서는 ‘변화 벡터(Change Vector)’를 활용하고자 하였다. ‘변화 벡터(Change Vector)’란, 편집 및 대체 과정을 통해 데이터에 변화가 있었는지(변경 또는 변경 없음) 여부를 라벨링(Labelling)한 것이다.

앞에서 선택한 91개 변수 중 결과에 유효한 영향을 미치는 변수만을 사용하기 위해서, 다시 25개의 특성 변수를 선택하여 연구를 시행하였다. 이 25개 특성 변수는 수기 편집 작업 팀이 사용하는 ‘소득 편집 및 대체 지침’을 기반으로 한다. ‘이진 변화 벡터(Binary Change Vector)’에 의해 25개 특성 변수 중 해당 변수에 하나라도 변경된 데이터가 있으면 ‘1’로 표시하고, 변경이 없으면 ‘0’으로 표시했다.

실험 결과, 변경 빈도가 낮은 20개의 특성 변수를 제거하면 머신러닝 알고리즘의 예측 성능이 향상되는 것을 발견하였다. 즉, 20개의 변수를 제거하니 알고리즘의 노이즈가 줄고 데이터 특성을 더 잘 탐지할 수 있었다. 예를 들어, SeTaxAmt(자영업 소득세 총액) 특성 변수는 테스트 데이터 2,912개 레코드 중 6개만 변경되었는데, 이 특성을 포함한 20개 특성에 대한 오류는 알고리즘에 의해 거의 예측되지 않았으므로 해당 변수를 제거하고자 했다.

반면, 변경 빈도가 높은 나머지 5개 특성 변수(NetNorm, IncTax, Nlns, GrossNorm, DedPenAm)는 남겨졌다. 해당 특성 변수들은 데이터에 사소한 변경이 있더라도 ‘Change(=1)’로 레이블이 지정되었다. 예를 들어, IncTax(소득세) 값은 £0.01만 변경되

있더라도 ‘Change(=1)’로 표시되었다.

그 이후에는 남겨진 5개 변수를 대상으로 ‘상대 변화 벡터(Relative Change Vector)’를 사용하여 변화율을 계산하였다. 그 결과, 10%의 상대 변화 임계값이 최적의 예측 결과를 낸다는 것이 밝혀졌다. ‘10%의 상대 변화 임계값’이란, 10% 이상의 변화를 ‘Change’로 간주하여 변화 벡터에 ‘1’로 기록하고, 10% 미만의 변화는 ‘No-Change’로 간주하여 ‘0’으로 기록한다는 뜻이다.

<표 2-2> 테스트 데이터(8Q2)의 특성 변수별 변화 빈도 비교

특성 변수	설명	변화 빈도	변화 빈도 (10% 임계값 적용 시)
NetNorm	연간 순소득(공제 후)	89	80
IncTax	소득세	270	268
NIns	국민보험료(National Insurance) 납부액	285	283
GrossNorm	연간 총소득(공제 전)	344	231
DedPenAm	연금 공제액	113	112

<표 2-3> 훈련 데이터(8Q3)의 특성 변수별 변화 빈도 비교

특성 변수	설명	변화 빈도	변화 빈도 (10% 임계값 적용 시)
NetNorm	연간 순소득(공제 후)	52	49
IncTax	소득세	247	246
NIns	국민보험료(National Insurance) 납부액	275	273
GrossNorm	연간 총소득(공제 전)	319	207
DedPenAm	연금 공제액	93	93

테스트 데이터(8Q2)의 변화 빈도는 <표 2-2>에, 훈련 데이터(8Q3)의 변화 빈도는 <표 2-3>에 나와 있다.

본 연구에서는 5개 변수 중 하나라도 변경된 경우 이를 변경 레코드로 간주하는 ‘단일 열 이진 변화 벡터(single-column binary change vector)’를 사용하였는데, 이 방법은 예측 결과를 개선한다는 장점이 있으나, 큰 영향을 미치는 개별적 특성 변수와 레코드를 구체적으로 식별하지 못한다는 단점이 있다.

‘Change’로 라벨링된 레코드 수를 살펴보면 다음과 같다.

- 테스트 데이터(8Q2): **2,912개 중 474개**가 ‘Change’로 라벨링
⇒ 10% 변화 임계값 적용 시 **387개**로 감소
- 훈련 데이터(8Q3): **3,059개 중 464개**가 ‘Change’로 라벨링
⇒ 10% 변화 임계값 적용 시 **384개**로 감소

다. 머신러닝 솔루션

영국 통계청은 sklearn Python 라이브러리에서 제공하는 지도 학습용 머신러닝 알고리즘을 사용하였다. 연구에 사용한 알고리즘 모델별 결과는 다음과 같다.

- ① 의사결정 나무(Decision Tree Classifier): 초기 실험에서 ‘결정 나무(Decision Tree)’를 시각화하고, 데이터 패턴에서 트리가 생성한 규칙을 추출하기 위해 사용되었으나, 예측 결과가 좋지 않았고, 추출된 규칙도 유용하지 않았다.
- ② 신경망(Supervised Neural Network): 성능이 좋지 않아 더 이상 사용되지 않았다.
- ③ 랜덤 포레스트(Random Forest): 매우 성공적인 결과를 보였다.

따라서 sklearn Python 라이브러리의 ‘랜덤 포레스트(Random Forest)’를 최종적으로 선택하여 머신러닝 모델을 훈련시켰다.

<표 2-4> 하드웨어 정보

구분	정보
모델명	ThinkPad T490
CPU	Intel Core i5-8365U 1.60GHz
RAM	8GB RAM
스토리지	256GB SSD

위 <표 2-4>에 따르면, 연구에 사용된 하드웨어 정보가 설명되어 있다. 전체가 아닌 일부 데이터를 선택 사용하였기 때문에, 고사양이 아닌 평범한 수준의 하드웨어

로 연구를 진행하였는데도 부족함 없이 연구를 수행할 수 있었음을 알 수 있다.

훈련 데이터로 총 3,056개의 사례(레코드)와 22개의 특성 변수를 사용하였으며, 모델을 훈련하는 데 총 2.71초가 소요되었다.

라. 특성 선택

연구에 사용할 수 있는 훈련 데이터(8Q3)는 3,059건으로 비교적 적다. 한편, 원 가구 레코드에는 2,000개 이상의 특성 변수를 가진 수많은 양의 데이터가 있기 때문에, 본 연구를 수행하는 데 ‘특성 선택(=변수 선택, Feature Selection)’이 매우 중요하다.

알고리즘 모델을 훈련하는 데 모든 변수를 사용하는 경우 과도한 노이즈(noise)를 생성할 수 있기 때문에, 일부 특성 변수를 선택하는 것이 좋다. 따라서, 사전에 ‘나. 데이터 준비’에서 영향이 큰 특성 변수를 살펴본 것을 참고하여, 모델 훈련을 위한 일부 특성 변수를 선택하고자 한다.

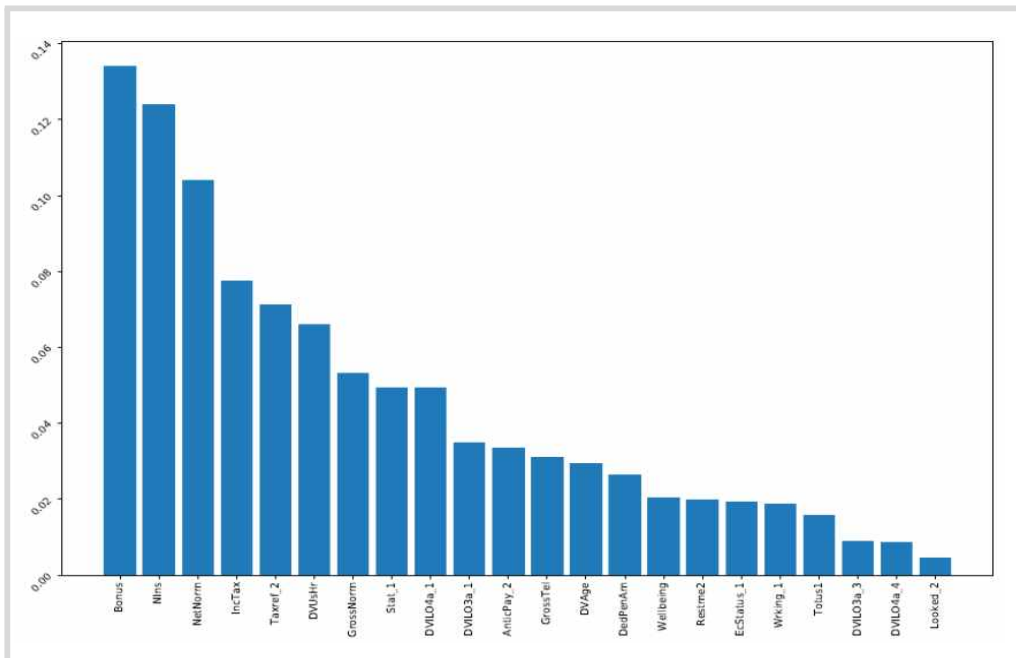
앞서 ‘나. 데이터 준비’ 과정에서 약 2,000개 중 91개의 주요 특성 변수가 선택되었는데, 여기에는 55개의 범주형 특성(Categorical Features)이 포함되어 있다. 이 범주형 변수에 OHE 작업을 수행한 결과, 특성 변수가 246개로 증가했다. 예를 들어 특성 변수 Stat_1은 OHE로 변환된 Stat 특성의 옵션 1을 나타낸다.

246개 변수를 사용하여 ‘특성 중요도(Feature Importance)’를 실험한 결과, 중요도가 0.01 이상인 특성만 사용하는 것이 모델 성능을 최적화하는 것으로 나타났다. 따라서, 머신러닝 모델을 훈련하는 데, 중요도가 0.01 이상인 22개의 특성 변수만을 선택하여 사용하기로 했다.

<표 2-5> 특성 변수별 중요도 및 설명(22개)

특성 변수	중요도	설명
Bonus	0.134	연간 보너스
NIns	0.123	국민 보험료(National Insurance) 납부액
NetNorm	0.104	연간 순수입 금액 (공제 후)
IncTax	0.077	주어진 기간 동안 지급된 소득세
Taxref_2	0.071	세금 환급이 안 된 마지막 급여
DVUsHr	0.066	주간 총 근무 시간
GrossNorm	0.053	연간 총 급여 (공제 전)
Stat_1	0.049	직원
DVIL04a_1	0.049	고용 중

특성 변수	중요도	설명
DVIL03a_1	0.034	고용 중
AnticPay_2	0.033	선 지급금
GrossTel	0.031	공제 전 총 소득
DVAge	0.029	나이
DedPenAm	0.026	주어진 기간 동안 지급된 연금 금액
Wellbeing	0.020	삶의 질 지표 4개 합계
Restme2	0.019	사람이 주소지에 거주한 연수
EcStatus_1	0.019	정규직으로 근무
Wrking_1	0.018	지난 7일 동안 수행한 유급 업무
Totus1	0.015	주당 보통 근무 시간
DVIL03a_3	0.008	경제활동 안 함
DVIL04a_4	0.008	경제활동 안 함
Looked_2	0.004	지난 4주 동안 어떠한 유급 일자리도 찾지 않음



<그림 2-2> 특성 변수별 중요도 그래프(X축: 특성 변수, Y축: 중요도)

<표 2-5>와 <그림 2-2>에 22개 변수로 수행한 연구의 특성 변수별 중요도가 나와 있다. 22개 특성 변수만을 갖고 수행한 후속 연구 결과를 보면, 중요도가 0.01 이하

인 변수 3개(DVILO3a_3, DVILO4a_4, Looked_2)가 추가로 발생했으나 최종적으로 제거하지 않은 것으로 보인다.

특성 변수 중요도 실험 과정을 정리하면 다음과 같다.

- ① 246개 모든 특성을 사용하여 ‘랜덤 포레스트’로 모델 훈련
- ↓
- ② 중요도가 0.01 이상인 22개의 특성 목록을 작성
- ↓
- ③ 앞 단계에서 선택된 22개의 특성만을 사용하여 랜덤 포레스트를 다시 훈련하고, 각 특성의 중요도를 다시 계산
(새롭게 계산된 22개의 특성 중요도는 <표 2-5>와 <그림 2-2>에 표시)
- ↓
- ④ <표 2-5>의 마지막 3개 특성은 중요도가 0.01 미만으로 나타남

마. 데이터 출력

영국 통계청은 위 과정을 거쳐 ‘랜덤 포레스트’로 머신러닝 모델을 학습하였고, 마침내 편집이 필요한 레코드를 식별하였다. 랜덤 포레스트 예측 결과는 각 레코드에 대해 2개의 숫자 쌍을 제공한다.

예를 들어 랜덤 포레스트 예측 결과, 한 레코드의 특정 변수 숫자 쌍이 [0.26, 0.64]로 나왔을 때, 이는 ‘가짜 확률(Fake Probabilities)’이 아닌 해당 레코드에 사용된 트리의 ‘투표 점수(Voting Scores)’라고 말할 수 있다. 위 숫자 쌍을 예시로 설명해 보자면, 1,000개 중 260개의 트리가 ‘변경 없음(No-Change)’에 투표했고, 640개의 트리가 ‘변경(Change)’에 투표했음을 뜻한다.

만약 예측 임계값이 35%인 경우, 이 레코드는 변경 투표 비중이 $640/1,000(=64\%)$ 이기 때문에 ‘Change’로 간주된다. 정리하자면, 머신러닝 모델은 변경 투표 비중이 예측 임계값을 넘는 경우, 해당 레코드를 ‘Change’로 간주하여 변경이 필요한 레코드로 식별하고(=편집), 적절한 예측값을 삽입하는 과정(=대체)을 거쳐 결과를 출력한다.

3. 연구 결과

레코드를 살펴보면, ‘Change’와 ‘No-Change’가 불균형을 이루고 있다. 그 이유는 연구 목적이 수기 편집자가 데이터를 변경한 경우를 ‘탐지’하는 것에 그치기 때문이다. 불균형 데이터 세트를 사용할 경우, 예측 성능이 낮은 모델이라도 모든 사례를 음성 범주(Negative Class, 본 연구에서는 ‘No-Change’)로 예측함으로써 높은 ‘정확도(Accuracy)’를 달성할 수 있다. 그 이유는 더 많은 데이터가 음성 범주로 구성되어 있기 때문이다.

정확도 관련 개념을 설명함에 앞서, 범주형 변수가 이진 분류(Binary Classification)에 속하는 경우(양성 또는 음성) 머신러닝 모델의 예측을 구분하는 개념은 다음과 같다.

- 진양성(True Positive, TP): 모델이 양성 범주를 올바르게 예측한 경우
- 위양성(False Positive, FP): 모델이 음성 범주를 양성 범주로 잘못 분류한 경우
- 위음성(False Negative, FN): 모델이 양성 범주를 음성 범주로 잘못 분류한 경우
- 진음성(True Negative, TN): 모델이 음성 범주를 올바르게 예측한 경우

위 개념을 활용한 정확도의 정의는 다음과 같다.

- 정확도(Accuracy): 전체 예측 중 올바른 예측의 비율을 측정한 것

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

모든 레코드가 ‘No-Change’ 클래스에 속한다고 예측해도 꽤나 높은 정확도를 얻을 수 있다. 테스트 데이터(8Q2) 총 2,912개의 사례 중 387개가 ‘Change’로 지정되었음을 감안할 때, 기본 정확도(Baseline Accuracy, A)는 다음과 같다.

$$A = \frac{TN}{TP + TN + FP + FN} = \frac{2,912 - 387}{2,912} = 86.7\%$$

위 산식에 따르면, 모든 레코드가 ‘No-Change’ 클래스에 속한다고 예측해도 86.7%라는 높은 정확도를 얻을 수 있다. 따라서, 이러한 상황에서는 정확도보다는 ‘재현율(Recall)’과 ‘정밀도(Precision)’가 모델 성능을 평가하는 데 더 유용한 방법이라고 할 수 있다.

재현율과 정밀도의 정의는 다음과 같다.

- 재현율(Recall): 실제 양성인 범주($TP+FN$) 중 모델이 올바르게 양성으로 분류(TP)하는 능력을 측정한 것

$$Recall = \frac{TP}{TP + FN}$$

- 정밀도(Precision): 모델이 관련 사례를 정확히 예측하는 능력으로, 모델의 양성 예측($TP+FP$) 중 실제로 양성인 비율(TP)을 나타냄

$$Precision = \frac{TP}{TP + FP}$$

각 지표는 알고리즘 성능에 대해 서로 다른 관점을 제공한다. 재현율은 모델이 실제 양성 범주를 놓치지 않았는지를 평가하는 지표로 ‘위음성(False Negative, FN)’을 최소화하는 데 중점을 두고, 정밀도는 모델의 예측이 얼마나 정확한지를 평가하는 지표로 ‘위양성(False Positive, FP)’을 최소화하는 데 중점을 둔다.

따라서 재현율과 정밀도의 목표가 다르기 때문에, 두 지표 간 ‘긴장 관계(Tension)’가 존재한다. 즉, 정밀도를 높이면 재현율이 낮아질 수 있으며, 반대로 재현율을 높이면 정밀도가 낮아질 수 있다. 예를 들어, 모델이 양성 범주를 더 많이 탐지하려고 하면, 더 많은 위양성(False Positive, FP)이 발생할 수 있다. 따라서, 무엇을 더 우선하는지에 따라 정밀도와 재현율을 조정하는 ‘절충(Trade-off)’이 필요하다.

머신러닝 모델은 각 사례가 ‘Change’ 범주에 속할 가능성을 백분율로 예측하는데, ‘예측 임계값(Prediction Threshold)’을 초과하는 사례를 ‘Change’ 범주에 속한 것으로 간주한다. 아래 <표 2-6>은 예측 임계값에 따른 평가지표 결과를 보여준다.

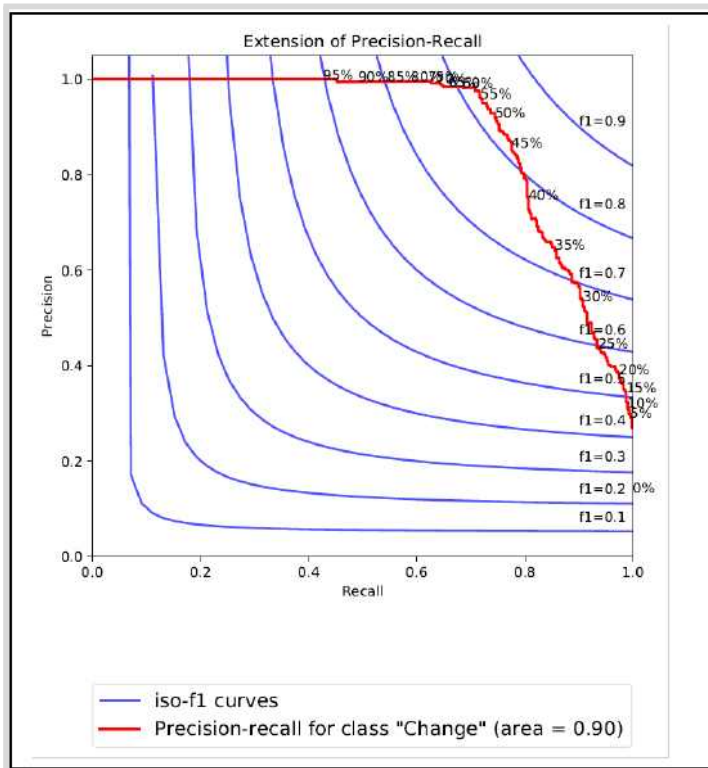
<표 2-6> 예측 임계값에 따른 평가지표 결과

예측 임계값	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%
Recall	97.4%	93.5%	90.7%	85.5%	80.6%	77.5%	74.4%	71.8%	68.5%	65.9%	63.6%
Precision	38.1%	43.6%	53.6%	64.3%	74.6%	85.7%	92.0%	95.9%	98.1%	98.5%	99.2%
F1-Score	54.8	59.5	67.4	73.4	77.5	81.4	82.3	82.1	80.7	78.9	77.5
TP	377	362	351	331	312	300	288	278	265	255	246
FP	612	468	304	184	106	50	25	12	5	4	2

* F1-Score: 정밀도(Precision)와 재현율(Recall)의 조화 평균

각 임계값에서 머신러닝에 의해 예측된 사례의 총합은 ‘진양성(True Positive, TP)’과 ‘위양성(False Positive, FP)’ 사례의 합(TP + FP)이다. 이 합은 임계값이 증가할수록 감소한다. 머신러닝 알고리즘은 임계값이 높을수록 해당 사례가 실제로 Change 클래스에 속할 가능성이 높다고 판단한다. <표 2-6>의 결과에 따르면 임계값이 증가함에 따라 TP와 FP의 수가 모두 감소하지만, FP 수가 더 빠르게 감소하기 때문이다.

<표 2-6>에서 ‘F1-점수(F1-score)’는 ‘정밀도(Precision)와 재현율(Recall)의 조화 평균’을 낸 것으로 적절한 품질 측정 지표로 간주되며, 약 50% 임계값에서 최고치를 기록한다. 그러나 아래 <그림 2-3>에서도 확인할 수 있듯이, 이 임계값에서 재현율(Recall, X축)은 74.4%로 충분히 높지 않을 수 있다.

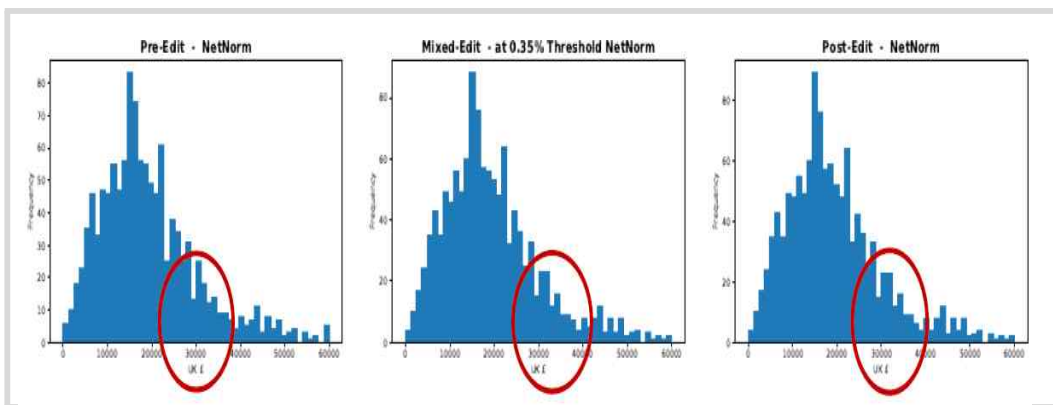


<그림 2-3> 정확도-재현율 곡선(Precision-Recall Curve)

<그림 2-3> ‘정밀도-재현율 곡선(Precision-Recall Curve)’에 따르면, 예측 임계값이 약 55%일 때를 기점으로 X축인 재현율(Recall)이 증가함에 따라 Y축 정밀도(Precision)가 급격히 감소함을 알 수 있다. 따라서 임계값 설정 시, 모델 개발자와 통계 담당자 간 논의를 통해 우선순위에 맞게 재현율과 정밀도 값을 조율할 필요가 있다.

연구에서 임계값을 설정하는 기준은 명확히 존재하지 않는다. 그러나, 임계값 설정값에 따라 재현율, 정밀도 등의 결과 평가지표 해석이 달라질 수 있으므로, 기준선(Baseline)이 되는 임계값을 설정하는 것은 매우 중요하다고 할 수 있다.

본 연구에서는 먼저 예측 임계값을 설정한 후, 유효한 영향을 미친다고 판단된 특성 변수를 선택하여, 최소 1개 이상 변수에 임계값 이상의 상대적 변화가 있는 레코드에 ‘Change’ 표시를 한 후 머신러닝 알고리즘을 훈련했다. 다음 부분에서 원본 데이터, 머신러닝 예측 결과, 수기 처리 결과, 3가지 방법을 그래프로 비교하여, 머신러닝이 편집이 필요한 데이터를 잘 분류했는지 여부를 분석해 보겠다.



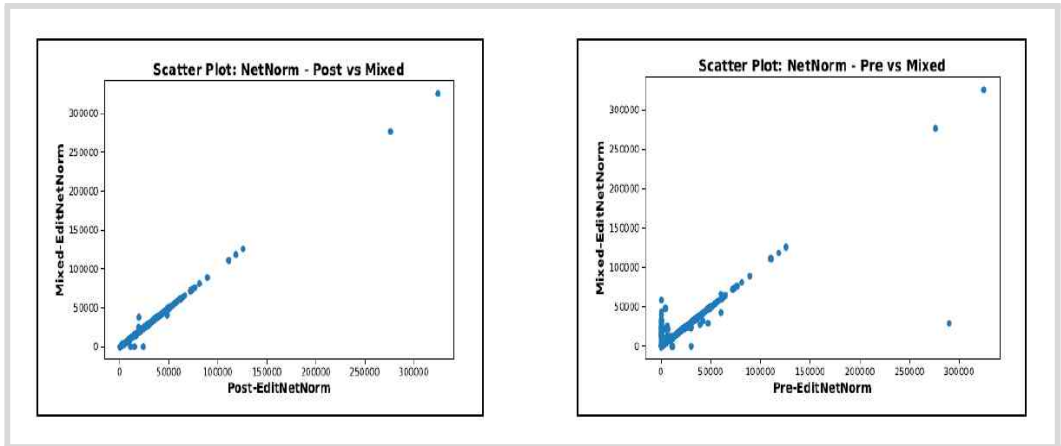
* X축: 예측값, Y축: 레코드 빈도

<그림 2-4> 35% 임계값에서의 NetNorm 데이터 비교

<그림 2-4>에 있는 3개의 히스토그램은 NetNorm(연간 순소득) 변수의 데이터 분포를 나타낸다. 왼쪽에서부터 ‘원본 데이터(Pre-Edit)’, ‘머신러닝 예측 결과(Mixed-Edit)’, ‘수기 처리 결과(Post-Edit)’에 해당한다. 즉, 중앙의 Mixed-Edit 그래프는 원본 데이터를 머신러닝으로 편집 및 대체한 것이고, 오른쪽의 Post-Edit 그래프는 원본 데이터를 수작업으로 편집 및 대체한 기존 방법을 뜻한다.

히스토그램의 빨간색 원을 보면, 중앙의 Mixed-Edit 그래프 모양이 우측 Post-Edit 그래프 모양과 매우 유사하고, 좌측 Pre-Edit 그래프 모양과는 차이가 있음을 보여준다. 즉, 원본 데이터 그래프와 비교했을 때, 수기 처리 결과와 머신러닝 예측 처리 결과가 비슷한 것으로 보인다.

‘2. 연구 과정’의 ‘가. 입력 데이터’ 부분에서 언급했듯이, LCF 통계조사 담당자는 수기 편집 및 대체 결과의 정확도를 80%로 추정하는데, 위 그래프에 따르면 머신러닝에 의한 편집이 수기 처리에 의한 편집만큼 예측력이 높은 것으로 보인다.



* X축: 수기 예측값(좌) 또는 원본(우), Y축: 머신러닝 예측값

<그림 2-5> NetNorm 변수의 산점도

<그림 2-5>의 산점도 또한 마찬가지로 NetNorm(연간 순소득) 변수 데이터를 사용하였다. 먼저, 좌측 산점도의 Y축은 머신러닝 예측 결과(Mixed), X축은 수기 처리 결과(Post-Edit)에 해당한다. 우측 산점도의 Y축은 좌측과 동일하게 머신러닝 예측 결과(Mixed), X축은 원본 데이터(Pre-Edit)에 해당한다.

좌측 ‘Post vs Mixed(수기 vs 머신러닝)’ 산점도에서 결과값들이 대부분 직선에 위치해 있는 것으로 볼 때, 수기 처리 결과와 머신러닝 처리 결과가 대체적으로 매우 비슷하다고 해석할 수 있다. 즉, 머신러닝이 편집이 필요한 데이터를 대부분 잘 예측하였으므로, 추가로 편집이 필요한 레코드가 매우 적음을 알 수 있다.

반면, 우측 ‘Pre vs Mixed(원본 vs 머신러닝)’ 산점도를 보면, 원본 데이터와 머신러닝 처리 결과가 일직선에 위치 하지 않는 경우가 많다. 즉, 머신러닝 예측 결과와 원본 데이터와 차이가 있는 걸로 봐서, 머신러닝이 꽤 많은 원본 데이터에 편집 및 대체 작업을 수행했음을 알 수 있다.

추후 추가적인 연구를 통해, 데이터 품질 향상에 가장 큰 기여를 한 변수와 레코드를 찾아내는 방법을 검토할 필요가 있다. 이 방법을 연구하는 것은 복잡할 것으로 예상되나, 머신러닝 초기 연구를 실제 활용 단계로 전환하기 위해 매우 중요한 것으로 보인다.

또한, 영국 통계청은 추후 세 가지 조사(LCF, SLC, WAS)를 통합하여 가계금융조사(HFS)를 위한 머신러닝 솔루션을 구축하려는 목표를 갖고 있다. 이때 LCF, SLC, WAS 조사의 기준선(Baseline), 즉, 임계값이 서로 다를 가능성이 높으므로, ‘조화된 임계값(Harmonised Thresholds)’을 도출하는 방법에 대해서도 논의가 필요하다.

제2절 대체: 벨기에 VITO

통계 조사에는 단위 무응답(참가자가 조사를 완료하지 않음) 또는 항목 무응답(특정 질문에 누락된 값이 존재)으로 인한 어려움이 수반된다. 이때 누락된 데이터를 적절하게 처리하지 않으면 통계 분석에 편향이 생길 수 있다. 그러나 현실적으로 누락된 데이터 없이 완전한 분석을 하는 것은 불가능하기 때문에, 통계 작성자는 누락된 값을 찾아 대체 처리하고, 입력한 값이 실제 값인 것처럼 분석을 진행한다.

최근 머신러닝의 발전으로 고차원 데이터와 복잡한 모델을 분석하기가 용이해졌다. 이러한 발전은 머신러닝 모델을 기반으로 한 누락 데이터 대체 방법이 기존 통계적 방법의 이론적 또는 계산적 한계를 해결할 수 있을 것이라는 기대를 갖게 한다.

머신러닝 기반 대체 방법은 이론적으로 기존 통계 모델에 비해 여러 가지 장점이 있다. 먼저, 분포 가정을 하지 않아도 되고, 혼합된 데이터 유형을 쉽게 처리할 수 있다. 또한, 변수 간의 비선형 관계를 모델링할 수 있으며, 고차원의 복잡한 설정에서도 기존 방법에 비해 손쉽게 좋은 성능을 보일 것으로 예상된다.

UNECE HLG-MOS에서 발간한 「공식 통계를 위한 머신러닝(Machine Learning for Official Statistics, 2021)」 보고서에서는 머신러닝을 대체(Imputation)에 적용한 연구 4가지(벨기에 VITO, 독일 연방통계청, 이탈리아 Istat, 폴란드 통계청)를 수록하였다. 본 연구보고서에서는 이 중 벨기에 VITO²⁾에 의해 수행된 대체 연구 과정 및 결과를 자세히 다루고자 한다.

벨기에 VITO는 통계에 머신러닝을 활용하여 결측값을 대체하고, 불완전한 데이터를 활용하여 분석 가능한 결과를 도출하는 것을 목표로 한다. 이 연구는 에너지 생산과 소비에 관한 ‘에너지 균형 통계’를 기반으로 하며, 통계 분석 결과는 매년 9월 또는 10월에 전년도 결과를 공표하는 연간 통계로 제공된다. 벨기에 VITO는 누락된 데이터 셀을 머신러닝 예측값으로 채움으로써, 최종적으로 통계 제공 시기를 앞으로 당길 수 있는지를 검토하고자 한다.

2) 벨기에 플랑드르 지역에 위치한 독립 연구기관으로, 약어로 VITO(Vlaamse Instelling voor Technologisch Onderzoek)라고 불린다. 에너지, 화학, 재료, 건강 및 토지 이용 분야에서 지속 가능한 사회로의 전환을 위한 과학적 조언과 기술 혁신을 제공한다.

1. 연구 개요

현재 전 세계 국가들은 매년 에너지 소비 목표치를 충족하기 위해 노력하고 있다. 벨기에는 에너지 소비 현황을 이해하고 미래를 위한 정책을 만드는 데 활용하고자, 매년 ‘에너지 균형 통계’를 작성해 왔다.

벨기에는 플랑드르, 왈론, 브뤼셀 3개 지역으로 나뉘며, 지역별로 자체 에너지 소비 계획과 목표를 가지고 있다. 벨기에 VITO는 플랑드르 지역에 있는 청정기술 및 지속 가능한 발전 연구기관으로, 매년 ‘에너지 균형 보고서’를 작성한다. VITO의 에너지 균형 보고서는 경제 부문별 및 에너지원별로 구성되며, 플랑드르 지역의 에너지 사용, 화석 연료 및 재생 에너지 생산량에 대한 전반적인 개요를 제공한다.

에너지 균형 보고서는 재생 에너지 비중을 모니터링하고, 벨기에의 에너지 의존도 파악 및 에너지 공급의 안정적 확보에 중요한 역할을 하고 있다. 또한, 에너지 관련 환경 정책을 평가하고, 정책의 영향을 정량화하는 데에도 사용될 수 있다. 더불어 VITO 내부 또는 외부에서 수행되는 타 연구의 기반으로도 활용된다.

에너지 균형 보고서를 작성하는 과정은 시간이 많이 소요되는데, 이는 대용량 엑셀 시트를 활용하는 등 구식 작업 방식을 사용하기 때문이다. 복잡한 엑셀 파일의 사용은 한 셀의 변경이 다른 프로세스에 영향을 미치는 등 데이터 오류로 이어질 수 있다. 또한, 데이터 수집 및 사용 가능 시기를 예측하기 어렵고, 데이터의 완전성을 명확히 보장하기 어려운 상황이다. 이런 문제들로 인해 VITO에서는 통계 작성 일정을 확실히 정하기 어려웠고, 보고서 완성 일정이 항상 지연되어 왔다.

현재 많은 통계기관이 적시에 효율적으로 통계를 생산하고자 노력하고 있고, 이를 위해 공식통계에 머신러닝을 활용하는 방안에 관심을 갖고 있다. VITO는 이에 뒤처지지 않기 위해 꾸준히 관련 연구를 진행하고 있으며, 이 연구도 그러한 노력의 일환으로 볼 수 있다. VITO에서는 대체 작업에 머신러닝을 활용함으로써, 보다 효율적으로 통계를 작성하고자 했다.

VITO는 48개의 경제 지표와 2개의 기상 지표를 바탕으로 에너지 균형 보고서의 에너지 소비량을 예측하고자 했다. 통계 제공 시기를 앞당기려면 누락된 데이터를 예측으로 보완해야 한다. 경제 지표와 기상 지표는 연초 정해진 시기에 사용할 수 있으므로, 이러한 비에너지 관련 지표를 사용하면 데이터의 지연 문제를 극복하고 더 빠른 예측을 생성할 수 있다.

2. 연구 과정

가. 입력 데이터

초기에는 1995년부터 2012년까지 플랑드르 지역의 연간 데이터를 사용하는 것으로 시작했으나, 더 많은 데이터를 사용하기 위해 벨기에의 분기별 데이터를 사용하는 것으로 변경하였다. 즉, 2000년 1분기부터 2019년 1분기까지의 벨기에 분기별 데이터를 사용하여 연구를 수행하였다. 변수는 다음과 같이 세 가지 유형으로 구분된다.

- 목표 변수(Target Variables, 8개): 전기 공급량
- 경제 변수(Economic Variables, 48개): GDI, GVA, GDP, 인구 등
- 기상 변수(Weather Variables, 2개): 태양 흑점, 난방도일

경제 변수와 기상 변수는 예측 변수 또는 보조 변수(X)로 사용된다. 우리가 예측하고자 하는 목표 변수(Y)인 에너지 변수의 범위는 벨기에 내 전기 공급량으로, 전기 공급량은 벨기에에 공급되는 전체 에너지 공급량의 일부이다. 자세한 변수 설명은 아래 <표 2-7>~<표 2-9>에 나열되어 있다.

<표 2-7> 목표 변수(Target Variables, 8개)

변수명	설명
+ EnrgCombustibleFuels	· 가연성 연료
+ EnrgNuclearNuclear	· 핵융합
+ EnrgHydroHydro	· 수력
+ EnrgGeothermalOther	· 지열/기타
= EnrgIndigenousProd	· 국내 생산량
+ EnrgImportsImports	· 수입
- EnrgExportsExports	· 수출
= EnrgElectricitySupplied	· 전체 전기 공급량 (= 국내 생산량 + 수입 - 수출)

<표 2-8> 경제 변수(Economic Variables, 48개)

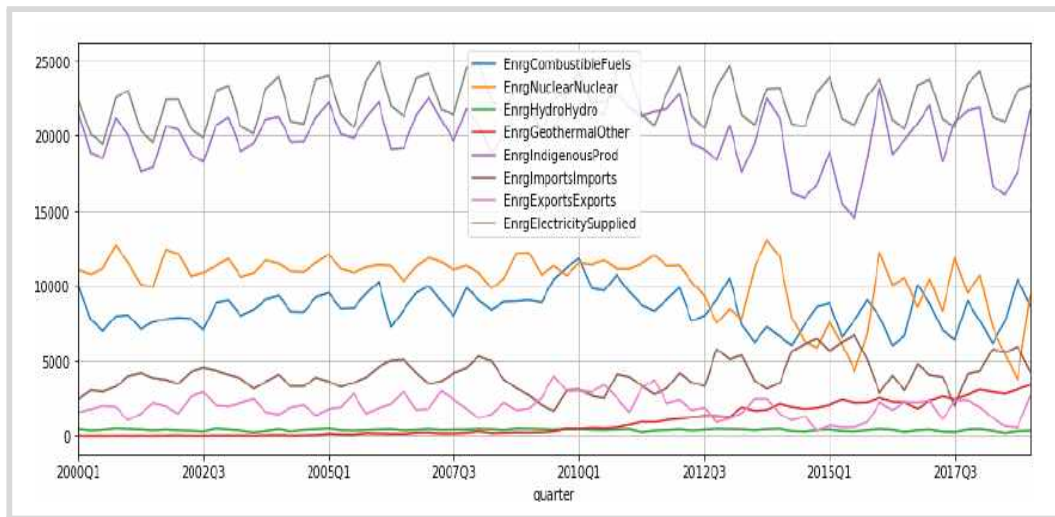
변수명	설명
Gross Domestic Income(GDI)	· 국내총소득(GDI): 국가 내 모든 경제 부문에서 받는 총소득
Gross Value Added(GVA)	· 총 부가가치(GVA): 특정 지역, 산업 또는 경제 부문에서 생산된 상품과 서비스의 가치를 측정하는 지표
Gross Domestic Product(GDP)	· 국내총생산(GDP): 특정 기간 동안 생산된 모든 최종 재화와 서비스의 시장 가치를 화폐 단위로 측정한 것
Population (in thousand persons)	· 인구(천명 단위)
그 외 44개 변수	· 고정 자본 소비, 순 국민소득, 최종 소비 지출, 고정 자본 투자, 경제 부문별 GVA(농업, 임업 및 어업, 건설업, 건설업 제외한 산업, 도소매업 등), 가계 최종 소비 지출, 일반 정부의 최종 소비 지출, 기업의 고정 자본 형성, 가구의 고정 자본 형성, 공공 행정의 고정 자본 형성, 재고 변화, 귀중품 취득 및 처분, 상품 및 서비스의 외부 균형 등으로 구성

<표 2-9> 기상 변수(Weather Variables, 2개)

변수명	설명
Sun spots	· 태양 흑점: 태양의 광구에서 일시적으로 나타나는 현상으로, 주변보다 어둡고 차가운 점처럼 보이는 현상
Degree days	· 난방도일: 건물의 난방에 필요한 에너지 수요를 정량화하기 위해 설계된 측정 방법

나. 데이터 준비

본 연구에서는 2000년 1분기부터 2019년 1분기까지의 분기별 데이터를 사용하였다. 변수는 목표 변수(Y)와 보조 변수(X)로 구분되며, 목표 변수는 <표 2-7>에 나와 있는 에너지 변수들로, 아래 <그림 2-6>은 목표 변수 8개의 시계열 그래프이다.



<그림 2-6> 에너지 변수(Y, 8개)의 시계열

예측 변수이자 보조 변수(X)로 경제 변수와 기상 변수를 사용하며, 각각 <표 2-8>과 <표 2-9>에 서술하였다. 추가로 본 연구에서는 에너지 변수의 ‘과거 관측치’를 보조 데이터로 고려하는데, 구체적으로 각 에너지 변수에 대해 이전 관측치 중 4분기 전의 값인 1년 전 동분기의 수치를 추가로 고려한다.

본 연구는 데이터를 두 개의 세트로 나누어, 각각 머신러닝 모델의 훈련과 테스트에 사용하였다. 총 77개 분기 중 훈련 데이터(Training Data)로 약 80%인 61개 분기, (2000년 1분기부터 2015년 1분기 시계열), 나머지 20%인 16개 분기(2015년 2분기부터 2019년 1분기 시계열)를 테스트 데이터(Test Data)로 사용하였다.

훈련 단계에서 먼저 훈련 데이터 세트의 보조 변수(X)를 머신러닝 모델에 입력한 후, 목표 변수(Y)를 함께 사용하여 모델을 최적화했다. 그러나 테스트 단계에서는 머신러닝 알고리즘에 테스트 세트의 목표 변수(Y)를 사용하지 않았다. 이 경우, 테스트 세트의 보조 데이터(X)는 단순히 훈련된 알고리즘에 입력되고, 목표 변수(Y)는 모델의 예측 성능을 평가하는 용도로만 사용되었다.

일부 알고리즘은 변수값이 넓게 분포되어 있을 때 수치적으로 불안정해질 수 있으므로, 알고리즘을 적용하기 전에 데이터를 표준화했다. 표준 점수 z 는 훈련 데이터 세트의 샘플 x 의 평균 u 와 표준 편차 s 를 계산하여 $z = (x - u) / s$ 로 계산한다.

다. 머신러닝 알고리즘 및 데이터 출력

머신러닝 알고리즘의 장점은 고차원의 데이터를 쉽게 처리할 수 있다는 것이다. 따라서 머신러닝을 사용할 경우, 특정 변수를 통계 작성자가 명시적으로 선택할 필요가 없으며, 알고리즘이 암묵적으로 중요한 변수를 결정하여 선택한다.

머신러닝 모델의 출력 결과는 입력 데이터(X)를 기반으로 한 에너지 변수(Y)에 대한 예측값이라고 볼 수 있다. 본 연구에서는 훈련된 모델을 사용하여 2015년 2분기부터 2019년 1분기까지의 에너지 변수에 대한 예측을 수행하고, 실제 값과 비교하였다. 데이터를 정규화했기 때문에, 예측 정확도와 품질을 평가하기 전 모델의 출력 결과를 본래 스케일(Scale)로 역변환 후 비교를 수행하였다.

또한, 머신러닝 수행 결과와 비교하기 위한 ‘기준 추정치’를 계산하였다. 기준 추정치는 계산하기 매우 쉽고 간단하며, t 시점을 기준으로 시계열 y의 값을 이전 연도의 값, 즉, 4분기 전 값인 $Y_t = Y_{t-4}$ 로 예측한다.

연구에서는 예측에 다음 5가지 머신러닝 알고리즘을 사용하였다.

- ① 선형 회귀 모델: 재귀적 변수 제거(Recursive Feature Elimination, 이하 RFE)를 사용하여 전체 모델에서 변수를 하나씩 제거한 후 남길 예측 변수를 결정하였다. 해당 연구에서는 3개가 남을 때까지 이를 반복하였다.
- ② 릿지(Ridge) 및 라쏘(Lasso) 회귀: 회귀 계수에 작은 페널티 항을 포함하였다. 계수가 작을수록 좋으며 이를 통해 회귀를 안정화하고 과적합을 방지할 수 있다.
- ③ 랜덤 포레스트: 회귀 나무의 앙상블 모델로, 하나의 나무는 데이터 세트를 점점 더 작은 그룹으로 분할하면서 내부 분산을 최소화하고, 외부 분산을 최대화한다. 여러 개의 나무를 사용하므로 단일 나무보다 더 견고하고 변수 선택이 필요하지 않으며, 알고리즘이 자동으로 선택하도록 설계되어 있다.
- ④ 신경망: 뇌가 작동하는 방식과 비슷하며, 함수로 단일 출력 결과를 계산하는 노드(node)들로 구성된 알고리즘이다. 한 노드의 출력은 다른 노드의 입력으로 사용될 수 있고, 값이 전달될 때마다 가중치에 곱해진다. 주로 입력과 출력의 관계가 비선형인 경우에 사용한다.
- ⑤ 앙상블 예측: 단일 모델이 모든 변수에 대해 항상 최고의 성능을 발휘하지 못할 때 사용하는 방법으로, 본 연구에서는 ‘선형 회귀, 릿지 회귀, 라쏘 회귀, 랜덤 포레스트, 신경망’ 5가지 예측 결과들의 평균을 계산하여 앙상블 예측을 생성한다. 이때, 가중치는 동일하게 부여된다.

3. 연구 결과

벨기에 VITO는 48개 경제 변수 및 2개 기상 변수(X)를 사용하여 8개 에너지 관련 변수(Y)를 예측하고자 하였다. 훈련 데이터와 테스트 데이터에 머신러닝 알고리즘 모델을 적용하였고, 결과 평가 품질 지표로 RMSE, ME, MAE, MAPE 4가지를 사용하였다. 다음 부분에서 각 품질 지표의 개념을 살펴보고, 머신러닝 알고리즘별 품질 평가 결과를 비교해 보겠다. 예측값 \hat{y}_i 은 참값 y_i 와 비교되며, 예측 정확도(Predictive Accuracy)를 판단하는 데 사용된다.

- RMSE(Root Mean Squared Error): 예측오차의 제곱합에 root를 씌운 값이다. 지표 자체가 직관적이며 예측변수와 단위가 같다. 스케일에 의존적이며 제곱 후 루트를 씌우기 때문에 실제 값의 과소·과대 추정 여부를 파악하기 힘들다.

$$RMSE = \sqrt{\frac{1}{n} \sum (\hat{y}_i - y_i)^2}$$

- ME(Mean of Error): 예측오차의 산술평균을 의미하는 ‘평균 제곱 오차’를 뜻한다.

$$ME = \frac{1}{n} \sum (\hat{y}_i - y_i)$$

- MAE(Mean Absolute Error): 예측값과 실제값 차이의 합을 의미하며, 직관적으로 차이를 알 수 있다는 장점이 있다. 이상치에 덜 민감하여 오차값이 Outlier의 영향을 상대적으로 적게 받고, 절댓값을 취하므로 과소·과대 추정 여부를 파악하기 어려우며 스케일에 의존적이다.

$$MAE = \frac{1}{n} \sum |\hat{y}_i - y_i|$$

- MAPE(Mean Absolute Percentage Error): MPE(Mean Percentage Error)는 예측오차를 단순 합계했기 때문에 0이 되는 경우가 있을 수 있다. MAPE는 이 문제점을 개선하여 예측오차의 절대값을 계산하지만 과소·과대 여부를 파악하기 어렵다.

$$MAPE = \frac{1}{n} \sum \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

<표 2-10> 연구 결과

Variable (8개)	methods (7개)	RMSE	MAE	ME	MAPE
EnrgCombustibleFuels (가연성 연료)	yminus4	913	853	-123	10.8%
	linmod	1251	1064	88	13.5%
	ridge	1492	1200	-690	14.6%
	lasso	1570	1408	-402	18.2%
	randomforest	1188	1039	544	14.1%
	neural net	2022	1542	-985	18.5%
	ensemble	1321	1114	-289	13.8%
EnrgElectricitySupplied (전체 전기 공급량)	yminus4	382	287	-25	1.3%
	linmod	750	655	-416	3.0%
	ridge	1129	941	889	4.2%
	lasso	1185	1048	941	4.7%
	randomforest	425	340	253	1.5%
	neural net	438	367	-347	1.6%
	ensemble	466	378	264	1.7%
EnrgExportsExports (수출)	yminus4	1032	834	-101	81.0%
	linmod	921	761	335	91.1%
	ridge	1299	1065	-784	69.7%
	lasso	1318	1061	-743	74.1%
	randomforest	888	794	-163	70.5%
	neural net	1162	914	-626	58.1%
	ensemble	998	842	-396	62.9%
EnrgGeothermalOther (지열/기타)	yminus4	425	373	-279	14.5%
	linmod	2105	2075	-2075	80.6%
	ridge	453	391	-310	15.0%
	lasso	495	417	-348	15.6%
	randomforest	812	712	-706	26.0%
	neural net	1182	1116	-1116	42.1%
	ensemble	975	911	-911	34.2%
EnrgHydroHydro (수력)	yminus4	61	47	14	14.9%
	linmod	80	64	49	23.2%
	ridge	66	51	23	18.3%
	lasso	77	60	53	22.5%
	randomforest	65	49	37	18.2%
	neural net	122	104	104	36.9%
	ensemble	79	62	53	22.8%
EnrgImportsImports (수입)	yminus4	1797	1382	145	35.0%
	linmod	1420	1120	-638	25.3%
	ridge	1882	1606	1239	44.6%
	lasso	2098	1804	1488	50.3%
	randomforest	1626	1377	953	39.7%
	neural net	1376	1146	446	32.3%
	ensemble	1479	1223	698	34.9%
EnrgIndigenousProd (국내 생산량)	yminus4	2658	2084	-271	11.1%
	linmod	3183	2596	1674	14.9%
	ridge	2595	2138	-1420	10.7%
	lasso	2773	2222	-1625	11.1%
	randomforest	2254	1944	-654	10.2%
	neural net	2509	2208	-1403	11.1%
	ensemble	2222	1947	-686	10.1%
EnrgNuclearNuclear (핵융합)	yminus4	3284	2614	117	37.8%
	linmod	3509	2761	2495	47.5%
	ridge	2902	2370	-581	33.3%
	lasso	2756	2300	-435	32.2%
	randomforest	2681	2369	-902	31.6%
	neural net	2847	2389	-11	34.1%
	ensemble	2580	2145	113	32.4%

* 변수별로 품질 지표마다 오차가 가장 작은 셀에 강조 표시함

위 표는 목표 변수(Y)의 모델별 품질 지표 결과를 보여준다. 위 표에 따르면 8개의 변수 중 4개에서 yminus4($Y_t = Y_{t-4}$, 기준선 방법)가 다른 머신러닝 모델에 비해 가장 좋은 성능을 보인다.

특히, EnrgElectricitySupplied(전체 전기 공급량) 변수의 경우, yminus4 기준선 예측이 이미 매우 우수하여 MAPE가 1.3%에 불과하며, 이는 매우 낮은 수치다.

EnrgCombustibleFuels(가연성 연료), EnrgGeothermalOther(지열/기타), EnrgHydroHydro(수력) 변수도 yminus4 방법이 가장 좋은 점수를 받았지만, 상대적으로 MAPE는 더 크다.

위 결과를 봤을 때, 머신러닝 방법은 기준선 방법보다 좋은 결과를 낼 수 없는 것으로 보인다. 그러나, 본 연구에서 사용된 예측 변수(Predictors, X)가 충분한 예측력을 가지고 있지 않은 경우, 연구 결과를 다르게 해석할 수도 있다. 또한, 평가지표에 따라 변수별로 가장 좋은 머신러닝 알고리즘이 다를 수 있다.

EnrgIndigenousProd(국내 생산량)의 경우, ME를 기준으로 봤을 때 yminus4의 값이 가장 작지만, 다른 평가 지표를 고려했을 때 머신러닝 방법의 결과값이 더 좋은 경우도 있으므로, 머신러닝 예측력이 꽤나 좋은 것으로 보인다. ME 외 다른 품질 지표 기준으로 좋은 알고리즘은 랜덤 포레스트와 앙상블 방법을 꼽을 수 있다.

위 결과에 따르면, 항상 동일한 머신러닝 방법이 모든 변수에서 최고 점수를 나타내지는 않는다. EnrgExportsExports(수출)의 경우, 랜덤 포레스트는 RMSE에서 최고점을, 선형 모델은 MAE에서 최고점을, 기준선 방법은 ME에서 최고점을, 신경망은 MAPE에서 최고점을 기록했다.

반면, EnrgNuclearNuclear(핵융합)의 경우, 앙상블 방법이 RMSE와 MAE에서 최고 점수를 얻었으나, 랜덤 포레스트는 MAPE에서 최고점을, 신경망은 ME에서 최고점을 기록했다.

EnrgImportsImports(수입)에서는 선형 모델의 예측력이 높았고, 릿지와 라쏘 회귀는 최고점을 기록하지 못했다.

<표 2-11> 품질 지표(4개) 및 예측 모델(7개)별 평균 순위

품질 지표	yminus4 평균 순위	linmod 평균 순위	ridge 평균 순위	lasso 평균 순위	randomforest 평균 순위	neural net 평균 순위	ensemble 평균 순위
RMSE	5.0	3.1	3.4	2.6	5.6	3.5	4.8
MAE	5.4	3.4	3.4	2.9	5.5	3.0	4.5
ABS(ME)	6.6	3.4	3.1	2.6	4.4	3.5	4.4
MAPE	4.8	2.8	4.0	3.0	5.0	3.4	4.8
종합 순위 (평균)	5.4	3.2	3.5	2.8	5.1	3.3	4.6

* 순위는 낮을수록 좋으며(오차가 작다는 뜻), 각 품질 지표별로 최고 점수 셀 강조 표시

본 연구에서는 모든 품질 지표가 그 자체의 장점을 갖고 있고 모델 평가에 골고루 사용되기 때문에, 결과를 분석할 때 다양한 품질 지표를 사용하여 종합적으로 설명하고자 하였다.

<표 2-10> 결과를 기준으로, 8가지 목표변수에 대해 4가지 품질 지표 및 7가지 예측 모델별로 오차 값이 높은 순으로 1위부터 7위까지 순위를 매겼고, 순위의 평균을 내어 위 <표 2-11>에 정리하였다.

위 표에 따르면, 순위가 높을수록(1위) 품질 지표 값이 높음(오차 값이 높으므로 나쁜 예측)을 의미하며, 순위가 낮을수록(7위) 지표 값이 낮아져(오차 값이 낮으므로 좋은 예측) 좋음을 의미한다. 측정 오차(ME) 지표의 경우, 0에 가까운 값이 선호되기 때문에 ABS(ME) 절대값을 기준으로 순위를 매겼다.

연구 결과에 따르면 ‘랜덤 포레스트’가 3가지 품질 지표(RMSE, MAE, MAPE)에서 가장 최상위를 차지하는 것을 볼 수 있다. ‘yminus4(기준선 방법)’는 ABS(ME)에서 최고 점수를 얻었으며, 종합 순위에서도 최고점을 기록했다.

즉, 종합적으로 ‘기준선 방법’이 가장 우수한 예측 결과를 보였다고 할 수 있으며, 머신러닝 알고리즘 중에서는 ‘랜덤 포레스트’ 결과가 가장 좋았다고 정리할 수 있다.

제 3 장

요약 및 시사점

제1절 요약

앞서 제2장에서 해외 AI 기반 내검 연구사례 2가지를 자세히 살펴보았다. 본 연구 보고서에서는 UNECE 「공식 통계를 위한 머신러닝(Machine Learning for Official Statistics, 2021)」 보고서상 정의를 반영하였는데, ‘내검’을 ‘편집’과 ‘대체’로 나누어 편집 부문에서는 영국 통계청의 ‘생계비 및 식료품 조사’, 대체 부문에서는 벨기에 VITO의 ‘에너지 균형통계’ 연구 결과를 정리하였다.

가. 편집: 영국 통계청

영국 통계청에서 수행한 머신러닝 편집 활용 연구의 주요 내용을 표로 정리하면 다음과 같다.

<표 3-1> 영국 통계청 연구결과 주요 내용 정리

	편집: 영국 통계청 ‘생계비 및 식료품 조사’
데이터	2018년 2분기, 3분기 조사 데이터
현 편집 방법	통계 작성자가 수작업으로 허위 레코드 감지(수기 편집)
ML 훈련 사용변수	22개(Bonus, NIns, NetNorm, IncTax, Taxref_2 등)
ML 알고리즘	Python 사용: 의사결정나무, 신경망, 랜덤 포레스트
ML 적용결과	랜덤 포레스트가 가장 성공적인 예측력을 보임
결과 평가지표	재현율(Recall), 정밀도(Precision), F1-점수(F1-score), TP(진양성), FP(위양성)
결과 시사점	원 데이터와 머신러닝 예측 결과를 비교했을 때, 몇 % 이상 변화의 경우 ‘Change’로 간주하여 편집과 대체를 수행할 것인지 ‘임계값(%)’ 설정 필요

영국 통계청에서는 2018년 2분기와 3분기 ‘생계비 및 식료품 조사’ 데이터를 사용하여 연구를 수행하였다. 훈련 데이터로는 3,059개의 2018년 3분기 데이터를, 테스트 데이터로는 2,912개의 2018년 2분기 데이터를 사용하였다.

기존에 편집을 수행할 때, 영국 통계청에서는 통계 작성자가 모든 데이터를 직접 검토하여 수기로 편집 처리를 하고 있다. 본 연구에서는 머신러닝 알고리즘을 활용하여 편집 과정을 수행한 결과와 수기로 편집을 수행한 결과를 비교하고자 하였다.

원본 데이터에는 2,000개 이상의 수많은 특성 변수들이 있는데, 본 연구에서는 통계적으로 의미 있는 일부 변수를 선택하여 모델 훈련에 사용하고자 하였다. 처음에 91개의 숫자형 또는 범주형 특성 변수를 선택하였고, 범주형 특성에 One-Hot-Encoding 작업을 수행한 결과 총 246개 특성 변수가 나왔다.

246개 변수를 사용하여 특성 중요도를 실험한 결과, 중요도가 0.01 이상인 22개 변수만 사용하는 것이 모델 성능을 최적화한다는 결론을 내렸다. 따라서, 본 연구에서는 최종적으로 결과에 유효한 영향을 미치는 변수인 22개의 특성 변수를 사용하여 머신러닝 모델을 훈련하였다.

분석 프로그램으로는 파이썬을 사용하였고, ‘의사결정나무, 신경망 모델, 랜덤 포레스트’ 3가지를 사용하여 머신러닝 적용을 시행하였다. 적용 결과를 봤을 때, 다른 방법들보다 ‘랜덤 포레스트’의 예측력이 가장 좋았기 때문에 최종적으로 랜덤 포레스트를 사용하여 모델을 훈련하였다.

머신러닝 모델이 제대로 잘 예측했는지를 판단하기 위한 지표로는 여러 가지가 있을 수 있는데, 본 연구에서는 ‘재현율(Recall), 정밀도(Precision), F1-점수(F1-score), TP(진양성), FP(위양성)’을 사용하였다. 이 평가 지표들은 머신러닝의 분류 효과를 판단하는 보편적인 평가 기준이라고 할 수 있다.

각 레코드가 변화된 정도를 뜻하는 예측 임계값(20~70%)에 따른 결과 지표는 <표 2-6>에 나와 있는데, 임계값에 따라 재현율, 정밀도 값이 달라지고, 두 가지 지표가 상충되기 때문에, 연구자가 알맞은 기준 임계값을 설정할 필요가 있다.

본 연구는 35% 예측 임계값을 기준으로 수기 편집 결과와 머신러닝 예측 결과를 비교하였는데, <그림 2-4> 히스토그램과 <그림 2-5> 산점도에 의하면 수기 편집 결과와 머신러닝 예측 결과가 꽤나 비슷함을 알 수 있다. 즉, 머신러닝 알고리즘이 수기 편집만큼 변경이 필요한 레코드를 잘 예측했다는 결론을 내릴 수 있다. 그러나, 아직 임계값 설정에 대한 명확한 기준이 미비하기 때문에 이 부분에 대한 보완 연구가 필요한 상황이다.

나. 대체: 벨기에 VITO

벨기에 VITO에서 수행한 머신러닝 대체 활용 연구의 주요 내용을 표로 정리하면 다음과 같다.

<표 3-2> 벨기에 VITO 연구결과 주요 내용 정리

대체: 벨기에 VITO '에너지 균형통계'	
데이터 및 변수 설명	분기별 데이터(2000년 1분기 ~ 2019년 1분기) - X: 경제 변수(48개, GDI·GDP·GVA 등), 기상 변수(2개, 태양 흑점·난방도일) - Y: 전기 공급량(8개)
현 대체 방법	규모가 크고 복잡한 엑셀 시트 수작업 대체
ML 알고리즘	Python 사용: 선형 회귀, 릿지·라쏘 회귀, 랜덤 포레스트, 신경망, 앙상블 예측
ML 적용결과	기준선 방법($Y_t = Y_{t-4}$) 및 랜덤 포레스트가 가장 성공적인 예측력을 보임
결과 평가지표	RMSE, MAE, ME, MAPE
결과 시사점	- 목표 변수(Y)별 가장 좋은 예측력을 보이는 머신러닝 알고리즘이 다름 - 기준선 방법($Y_t = Y_{t-4}$)이 머신러닝보다 더 좋은 결과를 내는 것으로 보임

벨기에 VITO에서는 2000년 1분기부터 2019년 1분기까지의 '에너지 균형통계' 데이터를 사용하여 연구를 수행하였다. 해당 통계는 경제 변수 48개와 기상 변수 2개를 사용하여 전기 공급량 8개를 예측하는데, 본 연구에서는 훈련 데이터로 2000년 1분기부터 2015년 1분기 데이터(61개)를 사용하였고, 테스트 데이터로 2015년 2분기부터 2019년 1분기 데이터(16개)를 사용하였다.

현재 대체를 수행할 때, 벨기에 VITO에서는 대용량 엑셀 시트에 수작업으로 처리하고 있는데, 이는 시간이 많이 소요될 뿐 아니라, 한 셀의 변경이 다른 프로세스에 영향을 미치는 등 데이터 오류가 빈번하게 발생하고 있다. VITO는 이 문제를 해결하고자 하였고, 그 일환으로 대체 작업에 머신러닝을 활용하는 연구를 수행하였다.

분석 프로그램으로는 파이썬을 사용하였고, 머신러닝 알고리즘으로 '선형 회귀 모델, 릿지·라쏘 회귀, 랜덤 포레스트, 신경망, 앙상블 예측'을 적용하여 분석을 시행한 후, 결과를 비교하였다.

본 연구에서는 머신러닝 결과 값을 연속적인 값을 예측하는 ‘회귀’ 예측으로 간주하여, 주요 성능지표로 예측값과 실제값의 차이, 즉, 오차를 활용한 RMSE, MAE, ME, MAPE를 사용하였다.

<표 2-10>에 의하면, 목표 변수별로 가장 좋은 예측을 보이는 머신러닝 모델이 각각 다르게 나왔음을 알 수 있다. 여기서 특히 주목할 점은 머신러닝 알고리즘이 아닌 1년 전 값으로 예측하는 ‘기준선 방법($Y_t = Y_{t-4}$)’이 가장 좋은 예측력을 보여줬다는 것이다. 머신러닝 알고리즘 중에서는 대체적으로 ‘랜덤 포레스트’가 좋은 예측력을 보여줬다.

벨기에 VITO에서 시행한 머신러닝 대체 활용 연구에서, 머신러닝 모델이 기준선 방법에 비해 낮은 예측력을 보이는 등 좋은 성능을 보이지 못하는 원인으로 크게 2가지를 들 수 있다.

먼저, 훈련 데이터 수가 매우 부족했기 때문이다. 머신러닝 모델을 훈련하는 데는 기존 통계 모델에 비해 훨씬 더 많은 양의 데이터가 필요하다. 그러나 연구에 제공된 데이터는 77개 분기(2000년 1분기부터 2019년 1분기 시계열) 데이터뿐이었으며, 그중 훈련 데이터(Training Data)로 61개 분기(2000년 1분기부터 2015년 1분기 시계열)를 사용하였다. 61개는 머신러닝 모델을 훈련하기에 턱없이 부족한 수로, 랜덤 포레스트의 경우 일반적으로 1,000개 이상의 훈련 데이터를 요구한다.

또한, 연구에 적용된 머신러닝 방법론이 적절하지 않았을 수 있다. 이 연구는 2020년에 시행된 연구로 그 이후로 머신러닝 방법론에 많은 발전이 있었다. 이 연구에 활용할 수 있는 최신 머신러닝 방법론으로 딥러닝 신경망 모델의 일종인 시계열 예측에 많이 쓰이는 ‘RNN(Recurrent Neural Network)³⁾’과 결측값 대체에 활용 가능한 ‘GAN(Generative Adversarial Network)⁴⁾’을 제안할 수 있다. 전통적 머신러닝 방법론 이외에, 많은 양의 데이터를 확보하여 딥러닝 신경망 모델과 같은 최신 머신러닝 방법론을 적용하여 모델을 훈련시키고, 그 결과를 분석해 볼 필요가 있다.

3) RNN(Recurrent Neural Network): 시퀀스 데이터를 예측하는 딥러닝 신경망 구조로, 시간이나 순서에 따라 변하는 데이터를 처리하는 데 특화되어 있다. 이전 시간 단계의 출력을 현재 입력으로 사용하는 구조를 갖고 있는데 이는 ‘은닉 상태’로 저장되며, 은닉 상태는 이전 정보를 기억하고 있어 시퀀스 내에서의 맥락을 유지할 수 있다.

4) GAN(Generative Adversarial Network): 입력값인 실제 데이터와 비슷한 특징을 갖는 데이터를 생성하는 딥러닝 신경망 구조로, 실제처럼 보이는 가짜 데이터를 생성하여 판별자를 속이도록 생성자를 훈련시키고, 실제 데이터와 생성된 가짜 데이터를 구분하도록 판별자를 훈련시킨다. 생성자와 판별자가 상호 경쟁하며 모델을 학습함으로써 생성자는 매우 현실적인 데이터를 생성할 수 있게 되고, 판별자는 데이터를 완벽하게 구분할 수 있게 된다.

제2절 시사점

연구 결과에 따르면 머신러닝을 활용하는 경우, 수기 내검 방식에 비해 많은 양의 데이터를 더욱 빠르고 일관성 있게 내검할 수 있다. 이는 더 높은 데이터 품질로 이어질 수 있으며, 내검 과정 자동화를 통해 시간을 단축함으로써 최종 분석 결과를 더 빠르게 공개할 수 있다는 장점이 있다. 또한, 시간이 지나 통계 작성자가 바뀌어도 관련 지식을 보존할 수 있으며, 통계 작성자가 내검 작업에 시간과 노력을 아끼 수 있어, 보다 중요한 레코드에 집중할 수 있다는 장점이 있다.

그러나 이러한 머신러닝 방법이 기존에 수행하던 방법보다 무조건 더 좋다는 결론을 내긴 어렵다. 머신러닝이 데이터를 훈련하고 모델 및 결과를 유지 관리하는 데에는 지속적인 노력이 필요하고, 관련 전문 지식이 계속해서 요구된다. 또한, 머신러닝 처리 과정은 ‘블랙박스’에 비유될 수 있다. 수기 내검은 일정한 규칙에 따라 데이터를 처리함으로써 프로세스를 논리적으로 설명할 수 있으나, 머신러닝은 자체적으로 다소 복잡한 처리 과정을 통해 편집 및 대체 결과를 표출하기 때문에 설명이 어려울 수 있다. 따라서 신뢰성 측면에서도 기존 방법이 더 우수하다고 보는 관점이 있다.

머신러닝은 현재 사용되는 방법, 즉, 전통적인 내검 방법보다 더 나은 결과를 가져온다고 판단되는 경우에 실무에 활용될 수 있을 것으로 기대된다. 현업에서 머신러닝을 추후 활용할 수 있을지를 판단하기 위해서는 머신러닝을 활용한 내검 방식이 기존의 방법보다 효율성, 신뢰성 측면에서 더 뛰어난 것을 증명해야 한다. 또한, 이를 평가하기 위한 기준으로 어떠한 지표를 사용할지에 대해서도 고민이 필요하다.

추가로 머신러닝을 실무에 적용하려면 더 많은 연구가 필요하며, 지속적인 노력과 헌신이 필요하다. 그 예로 통계 작성자가 전통적 통계 지식 외에 데이터 사이언스 기술(프로그래밍, 코딩 능력 등)을 학습해야 한다. 또한, 이러한 기술을 가진 프로그래머, 통계학자, 분야별 전문가가 적극적으로 협력하여야 한다. 그리고 활발한 연구를 위해 머신러닝 모델 훈련에 사용하기 위한 방대하고 충분한 양의 데이터를 우선적으로 확보하고, 다양한 최신 머신러닝 방법론을 적용하여 그 결과를 분석해 볼 수 있어야 한다.

참고문헌

한만규 외. (2016). *패널형 사업체 자료에 선택적 에디팅 적용 연구*. 한국자료분석학회.

Anneleen, G. et al. (2020). *Early estimates of energy balance statistics using machine learning*, VITO.

Claus, S. (2020). *Editing of LCF Survey Income data with Machine Learning*. ONS UK.

Fabiana, R. et al. (2021). *Adopting Data Science in Official Statistics to Meet Society's Emerging Needs*. Statistics Canada.

UNECE. (2021). *Machine Learning for Official Statistics*.

Zhenhua, W. et al. (2022). *Are deep learning models superior for missing data imputation in surveys? Evidence from an empirical comparison*. Statistics Canada.

Abstract

A Basic Research on AI Utilization in the Editing Area: Focused on Overseas Case Studies

Mihyeon Cho

Recently, interest in utilizing machine learning techniques in the field of statistics is rapidly growing. A lot of national statistical offices and international organizations are exploring methods to use machine learning so as to improve the quality of official statistics. As part of these efforts, under the auspices of the United Nations Economic and Social Council (UNECE), the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) launched the 'Machine Learning Project' in 2019. After summarizing the research results using machine learning in various fields, a report of Machine Learning for Official Statistics (2021) was published.

With reference to the above-mentioned report, this research summarized research processes and results using machine learning in the 'editing and imputation' area that was conducted by the HLG-MOS, and aimed to carry out basic research to apply AI to the editing of statistics produced by the Ministry of Data and Statistics.

First of all, according to the definitions in the UNECE report, the types of 'edit' were classified into 'editing' and 'imputation', and two case studies -one for each category- were summarized in this research. This thesis examined the research results of 'Cost of Living and Food Survey' by the UK Office for National Statistics (ONS) for the editing area, and the findings of 'Energy Balance Statistics' by VITO of Belgium for the imputation area. It was found that the classification and estimation results using machine learning were quite superior, when comparing the editing and imputation results using machine learning techniques with original data and existing data processing methods.

Finally, this thesis summarized the major findings of each case study, and presented the implications after looking into advantages and disadvantages of applying machine learning to editing, as well as future challenges. To actively utilize machine learning in practice, it is necessary to demonstrate that the editing methods using machine learning are superior to existing methods, in terms of efficiency and reliability, and criteria for the assessment of these new methods should be established.

Key words: AI, editing, substitution, imputation

연구진

○ 조미현(국가데이터처 국가데이터연구원 통계방법연구실 주무관)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2025-04

내검 분야의 AI 활용 기초연구: 해외사례 중심으로

인 쇄 2026년 1월

발 행 2026년 1월

발 행 인 김 진

발 행 처 국가데이터처 국가데이터연구원
35220 대전광역시 서구 한밭대로 713
TEL.(042)366-7100 Fax.(042)366-7123

홈페이지 <https://mods.go.kr/dsri>

ISSN(Online) 2733-4120





국가데이터처
국가데이터연구원

