기획논문 「정부학연구」제26권 제2호(2020): 29~61
DOI https://dx.doi.org/10.19067/jgs.2020.26.2.2

원고투고일: 2020.08.21.
원고수정일: 2020.08.30.
게재확정일: 2020.08.30.

데이터 사이언스를 활용한 사회안전망 강화: 의료보장제도 가입자의 위험 예측 모형 구축

정우진*

본 연구의 목표는 데이터 사이언스를 활용하여 공적의료보장제도 가입자의 위험률을 예측하는 모델을 구축하는 것이다. 이를 위해 먼저 본고는 데이터 사이언스적 접근의 개념과 최근 연구 추세를 살펴보았다. 다음으로는 이러한 논의를 적용하여 미국 공적의료보장제도 보험가입자의 입원율을 예측하는 모형을 구축하였다. 전통적 회귀모형, 일반 기계 학습모형, 딥러닝 모형을 포함한 6개의 모형을 비교했을 때, 딥러닝 모형들이 가장 재현율이 높았다. 특히 전방향 다중 신경망 모델은 사회 인구학적 정보를 통해 80%의 정확도로 환자의 재원 여부를 예측할 수 있었다. 결론에서는 모형의 예측력을 높일 수 있는 방안과 함께, 고위험 대상자에 대한 예방적 개입을 실시함에 있어 시사점을 논의하였다. 이러한 논의는 인공지능 기술을 이용하여 사회안전망을 강화하고 사회복지 재원이 위기 가정 및 개인에게 적절히 전달될 수 있도록 하기 위한 학술 및 정책 연구의 일환으로 기여할 수 있을 것이다.

주제어: 데이터 사이언스, 인공지능, 빅데이터, 사회안전망

I. 서론

최근 2020년 7월 정부는「한국판 뉴딜」종합계획을 발표하면서 지능형 정부로의 혁신을 천명하였다 (관계부처합동, 2020). 뉴딜은 공공데이터 14.2만 개 및 제조, 의료, 바이오 등 분야별 데이터를 수집, 가공, 연계 하면서 개인맞춤형 공공서비스를 신

* University of California, Berkeley 대학교에서 사회복지학/개발공학(Development Engineering)을 전공으로 박사학위를 취득하고, 현재 Rutgers, 뉴저지주립대학교 사회복지학과 교수로 있다. 주요 관심분야는 국제빈곤, 인공지능 및 데이터 사이언스를 통한 개발협력정책, 취약국 지역개발 등이다(wj153@ssw.rutgers.edu).

속 처리하는 방안을 담고 있다. 특히 사회복지 정책 관련해서는 국가보조금, 연금 맞춤형 안내, 취약계층 대상 디지털 안전진료 등 비대면 맞춤행정에 투자하고 관련제도를 개선하는 방안이 눈에 띈다. 이러한 정부의 계획은 최근 증거기반 정책을 강화하고 반복적 행정 업무를 처리하기 위한 도구로 빅데이터와 기계 학습이 활용되는 추세(Coglianese & Lehr, 2016)와 맞닿아 있다.

국내 학계에서도 인공지능(Artificial Intelligence: AI) 및 기계학습(Machine learning: ML) 기법을 정책, 행정, 사회복지, 보건학 분야에 적용하는 연구가 이루어지고 있다. 현재까지의 주요 연구들은 인공지능과 빅데이터의 발전을 위한 정부의 역할이나 이러한 방법론을 도입할 때 나타나는 변화에 대한 거시적 시론적 연구가 주가 되고 (김병조 · 은종환, 2020), 실증적 데이터 분석을 통해 공공정책과 서비스를 개선시킬 수 있는 방안을 탐색하는 연구는 아직 초기단계이다 (이제복 · 최상옥, 2018). 선행 연구들은 크게 세 가지로 나눠 볼 수 있다. 먼저 정부의 역할에 관한 연구로 주로 정부대책을 다룬다. 여기에는 데이터의 생산, 연계 및 활용을 (윤상오 · 김기환, 2016; 지광석, 2014; 임상규, 2014) 비롯한 데이터 거버넌스 시스템 구축 (이동규, 2016), 공공부문 빅데이터 관련 정책 및 연구 과제 제안 (성욱준, 2016; 오철호, 2017; 권설아 · 김지은 · 이재은, 2016) 등이 포함된다. 다음으로는 개념 및 사례분석을 통한 새로운 방법론에 대한 논의가 있다. 빅데이터를 활용한 기계학습과 전통적 연구 방법론과의 차이 (김선영, 2020; 김기환, 2013), 머신러닝 기반 의사결정 사례 분석 (김병조 · 은종환, 2020), 공공부문 서비스에 인공지능을 도입한 사례분석 (이제복 · 최상옥, 2018), 보이스피싱과 같은 위험을 예측하는 방법론 제안 (이승용 · 이주락, 2020) 등이 이에 해당한다. 마지막으로는 실질적인 데이터 분석을 통해 정책수요 분석이나 정책방향 설정 및 평가에 사용하고자 하는 연구가 있다. 인터넷 사용자의 검색활동을 분석하여 기후변화 정책수립이나 재난 위기 대응과 같은 정부 의사 결정에 반영 가능성을 본 논문들 (정예림 · 강정은, 2019; 이은미, 2015)이 그 예이다.

정책, 행적 분야에서 실증적 사례 연구는 AI/ML 기법을 통한 빅데이터의 활용 가능성을 보여주고 있지만 몇 가지 한계도 있다. 먼저, 정예림 · 강정은 (2019), 송태민 · 송주영 (2016), 이은미 (2015), 송태민 외 (2014) 연구의 방법론이 모두 일반대중이 기여하는 크라우드소싱(crowdsourcing) 방법에 편중되어 있다는 것이다. 소셜 미디어나 검색엔진을 활용하여 얻어진 문자 데이터 분석은 방대한 자료를 비교적 손쉽게 얻을 수 있는 장점이 있으나 데이터의 진위나 질에 대한 검증이 필요하다. 한편 보건 복지 분야에서는 위험 예측모델 (predictive risk modeling)이 활용되는데, 이는 정기적으로 수집되는 행정 데이터를 활용하여 미래의 부정적인 결과를 미리 예측하고 재원을

분배하는 것이다. 예를 들어 복지사각지대발굴을 위한 복지수급 예측모형 개발 연구(오미애 외, 2017)에서는 딥러닝(deep learning)의 한 종류인 합성곱신경망이 사용되었으나, 사용된 데이터베이스가 모의자료이고 횡단면 데이터이다. 중증도를 보정하여 환자의 재원일수를 예측하는 연구들은 (최병관 · 함승우 · 김촉환 · 서정숙 · 박명화 · 강성홍, 2018; 박종호 · 강성홍, 2019) 시계열 자료 분석을 위해 딥러닝의 한종류인 다층퍼셉트론을 활용하였다. 이러한 연구들은 기존 이론과 선행 연구를 통해 분석에 사용될 변수를 정리, 정제 하였다는 면에서 머신러닝의 자동화된 변수선택 기능을 이용하기보다는 전통적인 분석 방법을 혼합 하였다고 볼 수 있다. 또한 독립변수가 인구사회학적 변수라기 보다 주로 의학 정보이고, 시퀀스 데이터의 특성을 활용할 수 있는 순환적 신경망은 분석되지 않았다.

본 연구는 이러한 선행연구를 바탕으로 보건복지 정책 및 서비스와 관련된 실증적 분석 사례를 통해 예측모델의 가능성을 살펴보면서도 연구의 의미를 데이터 사이언스라는 관점에서 짚어 보고자 한다. 최근 연구들은 빅데이터, 기계학습, 인공지능과 같은 새로운 개념들을 사회복지, 보건, 정책, 행정 분야에 도입하고 있는데, 아직 이러한 개념들을 통합하여 데이터사이언스의 관점을 적용한 연구는 찾아보기 어렵다. 따라서 본 연구의 목적은 사회과학에서 데이터 사이언스적인 접근을 고찰해 보고, 이러한 접근법을 활용하여 공적의료보장제도 가입자의 위험률을 예측하는 모델을 구축하며, 정책적 함의에 대해 논하는 것이다. 본 연구는 먼저 데이터사이언스적 접근의 개념과 최근 연구추세 및 정책 활용성, 그리고 사회과학적 관점에서 짚어봐야 할 시사점을 논의한다. 다음으로는 실증적 분석 예시를 통해 이러한 데이터 사이언스적 접근이 어떻게 예측분석 모형개발과 활용에 적용될 수 있는지 논하고자 한다. 신경망 분석 모델 분석 과정에 대한 기술이 부족한 기존 연구와 달리, 본 연구는 모형의 선택이유와 모델링 과정에 대해 보다 구체적으로 논함으로서 후속 연구 시 유사 모형을 적용하고 비교, 발전시킬 수 있는 가능성을 높이고자 했다.

II. 데이터 사이언스의 활용

1. 개념

데이터 사이언스는 분야별 전문성과 컴퓨터 과학 및 프로그래밍 기술, 그리고 통계 · 수학적 지식이 결합되어 다양한 사회적 현상을 분석하고 이를 문제해결에 활용하는 융

합 학문이라 볼 수 있다[1]. 특히 응용사회과학에서 데이터 사이언스적 접근은 인류가 당면한 중요한 사회적 문제를 연구하고 해결하기 위해 데이터를 발견하고 구조화 하며 가설 형성과 검증을 통해 데이터로부터 실천 가능한 지식을 추출하는 하는 것이라 할 수 있다[2]. 데이터 사이언스의 요소를 보다 자세히 살펴보면 i) 분야 전문성에 기반을 두어 연구 목적에 맞는 데이터를 수집하고, ii) 프로그래밍 기술을 통해 원 자료를 분석가능 데이터로 표현하고 구조화하며, iii) 여기서 통계적, 수학적으로 데이터를 모델링하여 통찰을 얻고, iv) 데이터를 시각화하여 묘사하며, v) 기술기록보관 및 색인과 데이터 관리정책을 다루는 일련의 과정으로 이루어져 있다 (Brady, 2019).

데이터 사이언스보다 더 많이 개념화 되어 있고 많이 쓰이는 용어 중에 빅데이터와 인공지능, 그리고 기계학습이 있다. 이러한 개념들은 서론에서 논의한 방법론 분야 연구들이 문헌정리를 통한 정의를 제시하고 있다. 따라서 본 연구에서는 이러한 유사 개념들이 데이터사이언스와 어떻게 연관되는지에 초점을 두고 관련 연구 동향을 살펴보고자 한다.

2. 연구 추세

데이터 사이언스적 접근을 활용한 현재까지의 연구추세는 크게 세 가지로 나누어 볼 수 있다. 이는 i) 새로운 변수의 개발, ii) 새로운 변수를 활용한 전통적인 측정, iii) 새로운 알고리즘을 통한 관계분석이다. 이 세 연구 분야는 각각 빅데이터, 통계, 기계학습과 밀접하게 관련되어 있다. 실제 연구에서는 모든 요소가 다 사용되지만 각 분야별로 특히 중요한 요소에 대해서 생각해 볼 수 있다. 먼저 새로운 변수의 개발을 위해서 빅데이터가 주로 활용되며, 다음으로 새로운 변수를 활용한 전통적인 측정에는 통계와 수학적 모델링이 강조되고, 마지막으로 새로운 관계분석에는 기계학습을 통한 데이터의 구조와 양식(pattern) 파악, 그리고 인간의 지능을 모방한 자동화된 의사결정, 즉 인공지능이 유용하다. 세 가지 추세를 보다 자세히 살펴보면 아래와 같다.

첫 번째 연구의 흐름은 새로운 데이터 원천을 이용해 새로운 변수를 개발하려는 것이다. 이는 빅테이터, 즉 빠른 처리속도를 요구하는 대규모의 데이터, 출처가 다원화되

1) Drew Conway의 벤다이어그램은 분야 전문성, 프로그래밍 기술, 통계적 지식이 골고루 접목되어야 함을 보여준다. 만약 분야 전문성과 통계 지식에만 의존한다면 전통적 연구의 영역에 가깝고, 프로그래밍 기술과 통계적 지식에만 의존한다면 기계학습에 가까울 것이다. 한편 분야 전문성과 프로그래밍 기술에만 의존한다면 분석된 결과에 대한 해석과 검증이 어려울 것이다.

2) (NIST, 2015)를 참고로 사회과학 분야에서의 데이터 사이언스를 개념화 함.

고, 미검증된 정보를 포함하는 데이터를 이해하고 활용하는 것과 관련된다[3]. 즉 비 정형적이고 비 구조화된 대규모의 데이터를 다양한 경로를 통해 수집, 분석하여 기존의 전통적인 변수를 보완하는 새로운 변수를 개발할 수 있다. 빅데이터에는 보통 시·공간적 정보가 포함되어 있으며 더 나아가 실시간으로 생성 확장이 가능한 경우도 있다. 관련 연구로는 핸드폰에서 얻어진 전화통화기록 메타 데이터(Call Detailed Records, CDR)를 빈곤 대용지표로 활용할 수 있음을 검증한 경우나 (Blumenstock, Cadamuro & On, 2015), 구글 거리 보기에 포착된 자동차의 종류를 통해 지역별 인구사회학적 특성의 대용지표로 사용한 연구 (Gebru et al., 2017), 착용가능한 센서(sociometer)에서 나오는 신호를 통해 사람간 거리를 측정하고 이를 사회적 상호작용이라는 대용지표로 만들어 낸 연구 (Onnela, Waber, Pentland, Schnorf, & Lazer, 2014) 예시가 있다. 이러한 데이터는 보통 고해상도이면서 높은 밀도로 수집되고 수집방법도 자동화 되어 있어서 기억력과 자가보고에 의존하는 조사방법보다 정보의 편향(예: 사회적 바람직성 편향)을 줄일 수 있고, 시공간적으로 촘촘한 데이터가 필요할 때 효과적이며, 비용 효율적이다. 그러나 데이터의 접근성, 복잡한 데이타 처리 및 분석과정이 문제일 수 있다.

둘째, 빅데이터에서 얻어진 새로운 변수를 포함하되, 변수들간의 관계는 전통적인 계량경제학이나 통계학적 방법으로 분석 할 수 있다. 여기서는 전통적 통계분석이 강조된다. 지역 간 이동이라는 변수를 모바일폰이라는 빅데이터로 측정하되, 말라리아 발병율이 높은 지역에서 온 여행자가 특정지역의 발병율 확산에 미치는 영향을 모수적인 방법으로 측정한 연구 (Milusheva, 2016)가 그 예이다. 인종비하 단어의 검색빈도를 인종편견을 측정하는데 활용한 후 이를 흑인대통령 후보에 대한 투표성향과 연계한 경우도 (Stephens-Davidowitz, 2014) 이에 해당된다.

셋째, 인공지능, 기계학습 분야에서 개발된 알고리즘을 복잡한 변수들 간의 관계 규명에 사용하는 접근법이 있다. 여기서는 인공지능이나 기계 학습적 개발과 활용에 초점을 맞춘다. 사회과학에서 전통적 분석은 특정 이론과 가정 하에 연구자가 수학적 통계적 모형을 통해 설명변수와 반응변수($\hat{y}$)의 관계를 설명하고, 주요 요인 변수들을 규명하는 것이다. 그러나 기계학습은 비선형적, 고차원적 데이터세트를 비모수적(non-parametric)으로 유연하게 탐험하므로 특별한 이론적 분석모형 설정 없이도 $\hat{y}$의 예측을 높일 수 있다.

---

3) 빅데이터의 특징은 4V라고 불리는데, 양(volume)이 많고, 빠른 처리속도(velocity)를 요구하며, 정보가 다원화(variety)되어 있고, 정보에 대한 정확성과 검증(veracity)이 요구되기 때문이다.

세 번째 분야는 크게 두 가지로 나뉘는데, 먼저 컴퓨터 과학적 입장에서 알고리즘을 향상시키는데 초점을 맞추는 연구가 있다. 자연어처리 분야(natural language processing, NLP)에서 설득력 있는 단어의 속성을 포착할 수 있는 알고리즘을 개발하는 연구 (Atkin, Srinivasan, Tan, 2019)가 그 예이다. 이와 함께 원격 탐사 데이터의 기저 특징을 자율적으로 학습(unsupervised learning)할 수 있는 알고리즘을 개발하는 노력 (Jean, Samar, Zaari, Lobell, & Ermon, 2019)이 또 다른 예이다. 이는 기계학습 분야에서 라벨(label)이라고 불리는 종속/타깃 변수가 부족하고, 라벨링 과정과 데이터 훈련의 품질을 담보할 수 없는 단점 (Geiger, Yu, Yang, Dai, Qui, Tang, Huang, 2019)[4]을 개선하고자 하는 노력의 일환이다.

사회과학적으로는 알고리즘 개발 자체보다는 이미 개발된 알고리즘을 응용하여 예측분석(predictive analytics)을 하고 이를 전략적 자원 분배에 활용하는 접근법이 관심을 받고 있다. 존재하는 방대한 데이터를 활용하고 일관적으로 규칙을 적용하는 기계학습의 장점은 많은 연구에서 보여지고 있다. 범죄행정학 분야에서 인간판사가 내리는 보석 결정보다 알고리즘이 내리는 결정이 재범죄율을 낮추는데 효과적이라는 점은 판사들이 재범죄라는 결과를 예측함에 있어 유용한 신호보다는 '잡음'에 반응한다는 것을 보여 준다 (Kleinberg, Lakkaraju, Leskovec, Ludwig & Mullainathan, 2018). 재무, 금융 분야에서도 ML 알고리즘이 투자자 집단보다 채무불이행 예측에 더 우수하며, 이를 활용할 경우 투자자와 대출 신청자 모두에게 이익이 될 수 있다는 연구가 있다 (Fu, Huang & Singh, 2018). 사회복지 분야에서는 아동학대 신고 전화 접수에 대해서 추가적인 개입의 필요성을 결정하는데 도움을 주는 알고리즘이 연구·개발되고 있다 (Chouldechova et al., 2018). 고위험군 임산부를 위한 산전관리 프로그램의 수급 결정에서도 기계학습이 (assessment)이 서류기반의 사정보다 더 효과가 있음이 연구되었다 (Pan et al., 2017).

3. 예측모형구축

데이터 사이언스는 자동화된 의사결정에 의존하는 것처럼 보이지만 각 단계별로 사회과학적 통찰력 및 정책·실무적 전문성이 요구되는 판단을 필요로 한다. 특히 본 연구가 초점을 두는 예측 모형의 경우 분석과정에서 다음과 같은 판단이 필요할 것이다.

먼저 질문의 형성에서는 정확성과 계산 비용이라는 두 가치가 상충하는 상황에서

4) 표준화된 절차를 지켜 라벨링 작업을 하지 않는다면 라벨의 질이 떨어질 것이고, 따라서 이 타깃변수에 맞추어 훈련된 데이터의 예측력도 떨어질 수밖에 없다.

데이터 셋(data set)의 크기를 결정해야 할 것이다. 그리고 타깃으로 정한 반응 변수는 보통 관찰 가능한 변수인데 이것이 실제 관심 있는 결과나 잠재변수(latent variable)를 대변하는지 여부를 살펴야 한다. 예를 들어 펜실베니아 주 알레기니 카운티(Allegheny County) 아동국에서는 아동학대가 신고된 아동에 대한 추가의뢰(referral)가 이루어진 경우 해당 아동이 '2년 내 가정 외 보호라는 분리조치를 받는지 여부'를 가지고 훈련되었는데 이는 '반복적이고 심각한 아동학대'라는 잠재변수를 나타내는 여러 관측변수 중 하나이다. 이 관측 변수가 잠재변수를 적절히 대변하는지, 기간을 2년으로 할 것인지 1년으로 할 것이지 등에는 사회복지 전문가나 아동보호전문기관의 판단이 필요할 것이다. 다음으로 얼마나 많은 설명변수를 사용할 것이며, 특히 민감한 변수를 사용할 것인지에 대한 결정을 하여야 한다. 예측하고자 하는 사건의 위험도가 높고 (예: 반복되는 아동학대) 낮음에 따라 민감한 개인정보 (예: 전과기록)나 통계적 차별성, 편향성(bias)을 가진 정보 (예: 인종, 우편번호 등)의 사용여부가 달리 결정될 수 있을 것이다(Shroff, 2017)[5]. 또한 선별적 반응변수(selective label), 즉 훈련에 사용할 표본이 편향되어 있음에서 오는 문제를 해결해야 한다. 예를 들어 판사의 보석결정 연구에서는 보석결정을 받은 사람들의 재범율만 알 수 있고, 아동학대 핫라인의 경우도 추가조사를 실시하기로 한 아동만 2년 내 학대로 인한 분리조치가 있었는지 알 수 있으므로, 이에 속하지 않은 집단의 결과는 예측 할 수 밖에 없다.

그 밖에도 전처리(pre-processing)시 분석의 단위, 결측치, 행정 데이터에서 흔한 비일관적인 기록에 대해서 다루어야 한다. 알고리즘 결정에서는 예측력과 해석력에 따라서 모형과 모형평가 방법을 선택할 수 있다. 결과 및 해석 단계에서는 인간이 결정을 내리는데 보조 자료로 활용되기 위한 위험점수를 보여줄 수도 있고, 가부 여부를 보여주는 분류 과제로 만들 수도 있다. 위험점수의 경우는 점수의 범위와 위험 기준점을 결정해야 한다. 그리고 결과 이후에 적절한 후속조치와 보상(payoffs)이 무엇인지 결정해야 한다.

4. 정책 활용

데이터 사이언스는 정책 타깃팅, 모니터링, 그리고 평가에 활용될 수 있다. 위에서 논한 위험예측모형은 정책에서 주로 수요 분석 등을 통한 정책 표적화, 즉 타깃팅

5) 데이터 자체가 가지는 편향성이 있다. 예를 들어 사회복지 수급자로서 정부 행정 데이터의 많은 비중을 차지하는 빈곤한 가정, 유색인종 가정들은 그 그룹에 대한 데이터 자체가 많이 쌓여 있기 때문에 어떤 위험률 예측 모형에서도 위험도가 높은 가정으로 나타날 가능성이 높다.

(targeting)과 연결이 된다. 위성사진과 딥러닝으로 마을별 빈곤을 측정하여 지역개발 원조가 빈곤한 지역으로 가고 있는지 측정한 연구 (Jung, 2019a)도 첫 번째 사례에 해당한다. 다음으로 데이터 사이언스를 현재 시행중인 정책이나 프로그램을 모니터링 하고 이를 개선시키는데 활용할 수 있다. 환경오염을 줄일 수 있는 친환경적 화덕을 지원한 프로젝트에서 화덕에 센서를 부착하여 화덕 사용 데이터를 실시간으로 관찰하였던 연구가 그 예이다 (Wilson et al., 2016). 어떤 개입의 영향을 평가하기 위해서는 먼저 참가자들이 그 프로그램을 정해진 데로 따르는 것이 선행돼야 하므로 사용데이터에 대한 분석은 정책대응과 사회경제적 결과 간의 잃어버린 고리를 찾는데 도움이 된다 (Jung, 2019b). 인과관계 검증을 해야 하는 평가의 영역은 Athey (2017)와 Clark & Golder (2014) 등이 지적했듯 반사실적 가정(counterfactual)과 식별전략(identification strategy)을 갖추는 연구 설계가 핵심이므로 빅데이터와 ML을 통한 접근이 한계가 있다. 다만 전통적인 영향평가의 보조적인 방법으로 쓰일 수 있는데, 유치원 교육에 관한 정책실험에 참가한 유치원생들의 추후 소득공제에 관한 빅데이터를 활용한 연구 (Chetty et al., 2011)가 한 예이다. 또한 무작위로 올린 인터넷 포스트(post) 중 어떤 내용이 검열을 받는지 실험을 통해 중국정부의 검열을 평가한 방법도 있다 (King, Pan, & Roberts, 2014). 이러한 세 부분 중 지금까지 가장 활발하게 이루어지는 것은 타깃팅이며 본 연구도 그에 초점을 맞추고 있다.

다음으로, 더 구체적인 논의를 위해 위에서 살펴본 데이터 사이언스적 접근을 실제의 예측모형 구축에 적용해 보겠다.

III. 분석 예시[6)]

1. 배경

보건 의료산업은 개인에 대한 대규모의 정보를 수집한다. 그러나 보건의료분야 빅데이터는 개인정보 보호 등 여러 가지 이유로 충분히 활용이 되기 어렵고, 활용이 된다 하여도 데이터의 소유 주체마다 단편적으로 운용이 될 수 있어 데이터로서 규모의 경제를 살리기 어렵다. 개인정보보호법 등 관련 법률을 자의적으로 해석하여 수집된 개

6) 본 분석은 캘리포니아대학교 버클리캠퍼스에서 2018년에 실시한 Data Mining and Analytics 의 과제를 발전시킨 것으로 분석에는 Woojin Jung, David Proudman, Kyungna Kim, Zhiling Pan, Dunyang Wang, Michael Fermanian이 참여하였다.

인 데이터를 공공의 목적으로 활용하기 위해 제삼자에게 제공하는데 소극적인 기관도 적지 않다 (송태민 외, 2014). 이와 반면, 정부나 보험회사의 경우는 보험가입자들과 보건의료 서비스 제공자들에 대한 통합적인 데이터를 가지고 있어 이를 활용하면 보건의료 정책과 실무에 활용하고 비용을 절감할 수 있을 것이다.

따라서 본 연구에서는 미국의 공적의료보장제도인 메디케어(Medicare, 미국의 공적의료보험)와 메디케이드(Medicaid, 미국의 공적의료부조로 한국의 의료급여제도에 해당) 보험 청구 데이터베이스를 활용하여 환자의 재입원율의 예측을 다양한 분석 모델을 비교해 보고자 한다. 전통적으로 연구자들은 관측 데이터를 분석할 때 회귀분석 모형을 사용해 왔다. 특히 로지스틱 회귀분석은 임상적 시계열 데이터를 통한 예측 모델 개발에 빈번하게 사용되고 있다. 그러나 이러한 모수적인 방법은 데이터 구조에 대한 특정한 가정을 기반으로 하는데, 이 가정이 위배될 수도 있고, 이론적으로 잘 정립되지 않은 분야나 잘 알려지지 않은 데이터 구조를 탐험하기에는 한계가 있을 수 있다. 또한 표본 수에 비해 모수의 수가 지나치게 많고, 모수가 무한차원의(hyper-dimensional) 함수형태 모형인 경우, 또 변수들끼리 상관관계가 커 예측 값과 실측값의 차이인 편향이 클 경우 분석의 어려움이 있다.

본 연구에서는 먼저 보험청구 데이터를 통해 병원 입원여부를 예측하는 모델을 만들고, 이와 함께 예측에 영향을 미치는 변수들을 분석하고자 한다. 병원 입원 여부는 환자의 진단명과 관련 없이 '심각한 건강 위험'이라는 잠재변수를 나타내는 대표적 지표이므로 반응변수로서 적절할 것이다. 입원여부 대신 입원율이라는 확률계산이나 입원위험 점수를 계산할 수도 있으나, 입원과 비입원은 연속적이라기보다 서로 분리된 이산형(discrete) 사건이므로 분류과제(classification task)로 보았다. 대안적인 지표로는 소수의 고위험 환자에 대한 신호를 파악할 수 있는 재입원 여부를 사용할 수 있을 것이다. 하지만 병원 입원에 드는 막대한 개인적, 국가적 비용을 고려한다면 입원이라는 위험 자체가 횟수와 관련 없이 단 한번이라도 일어나는지 파악하는 것이 중요할 것이다. 입원 가능성에 대한 정확한 예측은 위험 환자에 대한 예방적 개입을 함으로서 사후적 치료에 드는 의료비와 사회적 비용을 절감할 수 있게 한다. 이를 통해 넓게는 사회 전체적으로 요구되는 의료자원에 대한 수요와 공급을 가늠하여 희소한 자원을 보다 효율적으로 분배할 수 있게 할 것이다.

2. 데이터

본 연구의 데이터 셋은 비식별화 되었으며, 미국의 공적의료보장인 메디케어와 메디

케이드 수급자 45,000명의 데이터로 이루어져 있다. 메디케어 가입자들이 메디케어 수급 자격조건을 갖게 된 주 이유는 65세 이상의 고령이 되었거나 지난 2년간 장애를 가졌기 때문이고, 메디케이드는 저소득 자격을 만족하였기 때문이다. 본 데이터 셋에는 보험 가입자의 건강 상태 및 의학 관련 정보, 인구 사회적 특성(feature) 관련 939 개의 지표가 포함되어 있으며 2014년에서 2015년까지 8개의 분기별 정보가 들어있다.

반응 변수는 본 분석에서 예측하고자 하는 목표변수인 2016년 병원 입원 여부이다. 이는 환자별 2016년의 병원입원 횟수를 변환한 지표이다. 한편 주요 설명 변수로는 연령과 성별 특정 쿼터에 특정 의료 기관 및 서비스 이용 횟수 (예: 1분기인 2014년 1월-3월에 외래 방문), 특정 쿼터에 특정 질병과 연관된 의료 기관 이용 횟수 (예: 2분기인 2014년 4월-5월에 관상동맥질환으로 내원)가 있다. 그 밖에 보험가입기간, 처방 약 그룹별(예: 백신, 항생제), 미국 보건복지부(Department of Health and Human Service) 산하의 보험청(Center for Medicare and Medicaid Services: CMS)에 의해 계산된 의학적 위험 지수, 메디케어 수급이유, 메디케어 메디케이드 이중 수급자격, 저소득층 보조금 수급 자격, 보험보장 종류, 소득, 빈곤율, 가구원 수, 집 가치, 주거형태, 거주기간, 거주지역이 속하는 센서스 블록그룹(block group) 등을 포함하고 있다. 대부분의 변수들은 범주형 변수, 특히 이진변수이며 연속형 변수는 소득과 관련된 지표들이 있다. 10%의 표본은 일부 특성에서 결측값이 있었다. 보다 정교하게 결측값과 관련된 분석을 하고 결측값 대체를 할지 여부와 어떤 모형으로 결측값 대체를 할 것인지 설정해야 하나, 본 분석에서는 간편하게 평균값이나 중위 값을 사용하였다.

〈표 1〉은 분석대상자의 일반적 특성을 파악하기 위한 기술통계 표이다. 표에서 보이듯 본 데이터에서 예측하고자 하는 목표 지표인 "2016년 입원"이라는 결과 값은 전체 환자의 약 20%에서만 보이므로 비입원의 결과값 80%에 비해 훨씬 적은 불균형을 보였다. 입원 경험이 있는 환자의 경우 평균 입원 횟수는 2회로, 80%가 2회 이하로 입원하였으나, 20%는 3회 이상에서 최대 19회까지 입원을 하였다. 전체 환자의 4.4%가 재입원을 하였다.

본 데이터 셋은 40세 이상의 성인을 대상으로 하였으며 메디케이드 수급자격이 연령이므로 평균연령대가 70대인 노년층이 많았다. 장애인이 전체의 35%이고, 성별은 균형을 이루었다. 평균 빈곤선 이하 80%에 살고 있고, 연간 추정 소득은 $53,851 정도이며, 평균적으로 2인 가구였다. 교육은 고등학교 졸업에서 약간의 전문대학 교육을 받은 정도이고, 주요 주거형태는 단독 주택이었다. 보건복지부 보험청에서 계산한 의학적 위험지수는 대부분 1.3으로 낮은 편이었으며 대부분의 표본이 0-6사이에 있어서

오른쪽으로 꼬리가 긴 분포를 보여주었다. 보험청에서 계산한 처방 위험성 지수 역시 평균 1.01으로 양의 왜도(positive skewness)를 보였으나 표준편차와 최대값이 의학적 위험지수보다 낮았다.

데이터 전처리 과정에서는 다음과 같은 점을 고려하였다. 먼저 본 데이터 셋은 보험 가입자 전체에 대한 입원 및 비입원 결과변수가 모두 있기 때문에 선택적 라벨로 인한 표본편향 문제는 없었다. 또한 기계학습에서는 가용 가능한 모든 특성을 활용하는 것이 일반적인 철학이므로 (Shroff, 2017), 이론적 지식에 따라 사전에 특성을 선택하기 보다는 가능한 모든 특성을 분석에 활용하였다. 다음으로 변수를 분석 가능한 형태로 만들기 위해 범주형 지표를 원핫코딩(one-hot-encoding)의 이진 변수로 변환하였다. 그밖에 정규화(normalization)와 평균중심화(centering), 스케일링(scaling)을 하였으며, 결측값이 많은 변수와 2016년 변수 중 입원 전후의 시간순서가 분명치 않은 변수를 삭제하였다. 정규화는 단위가 전혀 다른 두개의 데이터가 존재할 때[7] 단위가 큰 특성이 과대평가 되고 예측력에 영향을 미칠 수 있는 경우 필요하다.

〈표 1〉 메디케어 메디케이드 보험청구 환자의 특성

변수	평균	표준편차	최소	최대
입원율 (2016년)	19.33%			
입원횟수	0.37	1.01	0	19
재입원율 (2016년)	4.33%			
재입원횟수	0.06	0.4	0	12
나이 (2014년 12월)	70.54	9.58	40.00	95.00
추정소득 (달러/연)	53,851	44,718	7,500	500,000
빈곤선 이상 (%)	86.49	10.47	0	99.00
가구수	2.46	1.79	1.00	16.00
의학적 위험지수	1.30	1.05	0	10.60
처방 위험지수	1.10	0.61	0	6.60
성별	여성 50.56%, 남성 49.44%			
수급 이유	나이:66.55%, 장애: 33.37%, 말기신질환: 0.06%, 장애+말기신질환: 0.02%			
주거 종류	단독주택 80.27%, 아파트 10.64%			
교육수준	고등학교 48.23%, 약간의 전문대학 교육 40.5%			

7) 예를 들어 나이의 범위는 40-95임에 반해 소득의 범위는 7,500-500,000으로 소득이 범위가 훨씬 크다.

3. 분석방법

본고에서는 인공지능 및 기계학습을 통한 분석 사례의 적용성을 보기 위해, 지도학습(supervised learning)을 통한 예측모형 개발을 주 분석방법으로 삼았다. 최적인 모형을 찾기 위해서는 하나의 모델 보다는 여러 통계모형들을 비교하여 분석하고 이 중 자료를 최적으로 설명하는 모형을 찾아낼 필요가 있다. 따라서 본고에서는 기본적인 로지스틱 회귀모형을 일반적 기계학습 및 딥러닝과 비교해 볼 예정이다. 본 연구는 오픈소스 프로그래밍 언어인 파이썬(Python)을 통해 예측모형 개발 및 평가를 실시하였다.

좋은 모형의 조건은 모델이 현실을 잘 대표할 수 있도록 타당해야 하고, 이를 가급적 간단하고 이해하기 쉬운 모형으로 구현하는 효율성을 가지고 있어야 한다. 또 작은 가정 및 모수의 변화에 안정적이어야 하고, 잡음에 대응할 수 있어야 하며, 큰 규모의 데이터를 다룰 수 있는 확장성이 있어야 한다. 이러한 여러 고려 요소 중에서도 예측을 중심으로 하는 모델에서는 가장 중요한 요소는 예측력이 높은지 하는 것이다 (오미애 외, 2017). 따라서 모형의 평가는 예측을 위해 만든 모형이 임의의 모형보다 예측력에 있어 우수한지 비교해 볼 수 있다.

1) 평가방법

예측력을 평가하는 데는 다양한 방법이 있으나 본 분석은 의료보장제도 가입자의 사회인구학적 지표와 의료 기록을 통해 특정 시점에 병원 입원이라는 사건이 일어났는지를 예측하는 것을 목적으로 하기 때문에 오분류율을 활용하였다. 오분류율을 산정하는 것은 목표변수의 실제 범주, 즉 환자 i 가 2016년에 실제 입원을 했는가와 예측된 결과의 범주, 즉 모형이 환자 i 가 2016년에 입원을 했는지 예측한 결과가 얼마나 비슷한지 분류 성능을 평가하는 것이다. 이를 위해 다양한 오류 행렬(Error metrics)을 참고할 수 있는데, 아래는 이러한 분류결과를 분석할 때 쓰이는 혼동행렬(confusion matrix), 즉 오분류표를 보여준다.

〈표 2〉 오분류표

	참 (True, T)		거짓 (False, F)	
양성 (Positive, P)	$\frac{TP}{P=TP+FP}$ 정밀도 (Precision)	$\frac{TP}{TP+FN}$ 재현율 (Recall) 민감도 (Sensitivity) 참양성 비율 (True Positive Rate: TPR)	위양성 비율(False Positive Rate: FPR, Fall-out) $\frac{FP}{TN+FP}$ =1-특이도 Type I 오류	
음성 (Negative, N)	특이도 (Specificity) $\frac{TN}{TN+FP}$ Type II 오류			
	정확도 (Accuracy) $\frac{True=TP+TN}{All=P+N}$		오류항(Error term) $\frac{False=FP+FN}{All=P+N}$	총합 (ALL)

- F1 스코어: 정밀도와 재현율의 조화 평균

$$\frac{2}{\frac{1}{\text{정밀도}}+\frac{1}{\text{재현율}}}$$

- 곡선아래면적 (Area Under Curve, AUC): TPR (y축, 실제 양성을 양성으로 예측)와 FPR (x축, 실제 음성을 양성으로 예측) 사이의 Receiver Operating Characteristic (ROC) 곡선 아래의 면적

여러 기준 중 본 분석은 Chouldechova et al. (2018)의 연구와 같이 정밀도와 재현율을 사용하였다. 데이터의 분류가 한쪽으로 편중될 경우, 정확도만 사용한다면 지나치게 높은 예측결과를 보여줄 수 있다. 입원율이 비입원율 보다 훨씬 낮을 경우, 비입원율을 정확히 맞추는 것은 어렵지 않기 때문이다. 이에 반해 정밀도는 모델이 양성으로 분류한 모든 사례 중 실제로 양성이었던 경우를 측정한다. 한편 재현율은 실제 양성인 사례 중 모델이 올바르게 양성으로 예측한 경우의 비율로 본 분석에서 가장 의미있는 지표이다. 다만, 정밀도의 경우 보수적으로 확실한 양성 사례만 양성으로 분류하고 확실하지 않은 경우 모두 음성이라고 한다면 예측력을 극대화 할 수 할 수 있고, 반면 재현율은 확실한 음성의 경우만 음성으로 분류하고 대부분을 양성으로 분류한다면 예측력을 극대화 할 수 있다. 따라서 상충 관계에 있는 정밀도와 재현율 두 지표를 종합해서 보여주는 F1 점수도 참고할 수 있다.

2) 분석 모형

본 연구는 안정적이며 직관적 해석이 가능한 로지스틱 회귀 모형을 기본 모델로 사용하였다. 이는 이항(binary) 종속 변수와 독립변수가 연결 함수를 통해 선형적 관계로 연결되어 있다는 가정에서 출발한 것이다. 하지만 이러한 가정이 위배될 여지가 있다고 보고, 기계학습을 활용하였다. 먼저 일반 기계학습 모형으로 의사결정 나무를 사용하였다. 의사결정 나무는 지도학습(supervised learning) 중 특정 임계치를 넘었을 때 사건이 일어난다고 보는 분류 과제를 수행할 수 있는 모형이기 때문이다. 그리고 다수의 예측이 하나의 예측보다는 더 정확하다는 가정 하에 다수의 의사결정나무를 종합해서 모형을 도출하는 앙상블(ensemble) 방법을 사용하였다. 또한 표본 수는 한정적인데 1,000여개 가까운 설명변수를 효과적으로 축소시키기 위한 방법으로 서포트 벡터 기계(Support Vector Machine: SVM)를 차용하였다. 마지막으로 수많은 비선형적 종속변수와 설명변수가 어떻게 연결되는지에 대한 이론이 부족하다는 면에서 복잡한 대규모 데이터의 구조를 발견할 수 있는 효과적인 방법으로 제시된 딥러닝(LeCun, Bengio & H inton, 2015)을 사용하였다. 각각의 모형에 대해서는 아래에 보다 자세히 다루도록 하겠다.

① 일반적 기계학습

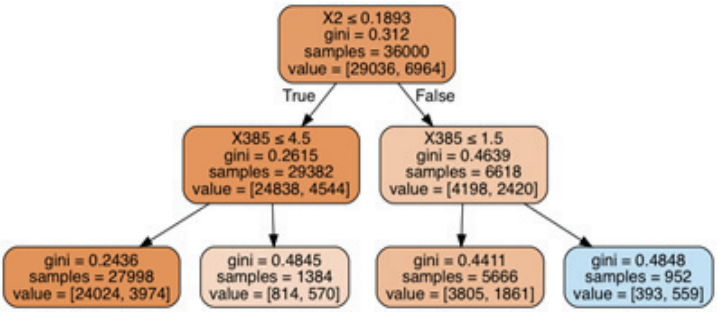
a. 의사결정 나무(decision tree)

의사결정 나무는 만약 A 라면 B라는 “if and then” 의 설명 변수간 규칙을 배워 타깃인 Y에 적용하며 모형에 대한 가정이 필요 없는 비모수적 방법이다. 특정 임계치를 넘으면 입원과 비입원을 구분하는 분할점이 생기는데 가장 설명력이 있는 변수에 대해 최초로 분리가 일어난다. 직관적이고 이해하기 쉬운 규칙을 생성하므로 지도학습 문제에서 최종모형의 해석력이 중요할 경우 사용할 수 있으며 이상치(outlier)에 덜 민감하다는 장점이 있다. 그러나 자료에 약간의 변화가 있는 경우에 전혀 다른 변화를 줄 수도 있어 안정성이 떨어질 수 있다 (Kotu & Deshpande, 2014).

비입원율에 대한 입원율의 비율이 1:4로 큰 차이가 있기 때문에 전체적인 오류를 줄이기 위해 모든 모델에서 입원율이 과소 측정되는 경향이 있었다. 이를 보정하기 위한 다양한 방법이 있지만 본 연구에서는 비입원 환자를 적게 표본 추출하여 입원, 비입원간 분류, 즉 클래스(class)의 균형을 맞추었고, 이를 통해 재현율을 높일 수 있었다. 최대나무 깊이와 같은 정지규칙은 재현율을 최대화 하면서도 과적화(overfitting)

를 피할 수 있는 50으로 결정하였다. 아래 의사결정 나무 〈그림 1〉은 미국 보험청에서 계산한 위험점수 (x=2)와 2015년 10월에서 12월 요양시설에 방문한 횟수 (x=385)가 상대적으로 가장 설명력 있는 두 개의 변수임을 보여준다.

〈그림 1〉 의사결정 나무



b. 앙상블: 무작위 숲(Random Forest)

무작위 숲, 즉 랜덤포레스트는 개별 결정 나무들의 예측을 다수결 방식으로 조합하는 앙상블 방법이며, 숲을 구성하는 모든 나무들은 동일한 분포로부터 독립적으로 표본 추출된 무작위 벡터로 만들어 진다 (Brieman, 2001). 무작위 숲은 무작위중복표본추출(bootstrapped)된 샘플이 각기 다른 특성변수의 조합으로 구성되어 있어 과적화 및 다차원성을 줄이는데 효과적이며, 대체로 좋은 예측력과 정확성, 안정성, 사용의 편의성 때문에 널리 쓰이는 방법이다. 스스로 중요한 특징들을 자동으로 선택하며 특정 설명변수/특성을 사용하는 노드가 얼마나 분류의 불순성(impurity)을 줄일 수 있는지를 통해, 상대적 중요성 지수를 제공한다. 그러나 모델 간 독립성을 달성하기 어렵고 모델 안에서 어떤 작용이 일어나는지 이론적 설명이 부족하며 최종 결과에 대한 해석이 어렵다.

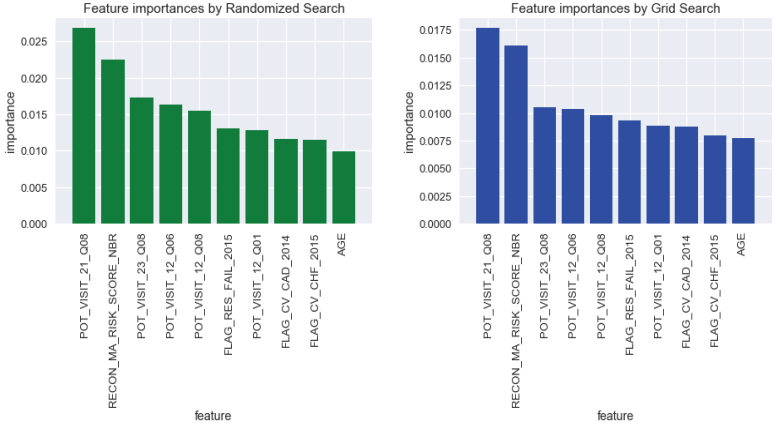
본 분석에서는 초매개변수(hyper-parameter)의 조율(tuning)을 위한 무작위 탐색(randomized search) 및 격자 탐색(grid search) 알고리즘을 통해서 자동으로 중요 특성을 선택하고 상대적 중요성 지수를 구하였다. 격자탐색의 경우 무작위표본추출, 불확실성 (entropy) 기준 사용, 최대특성 10, 최소 나뭇잎 3, 최소 분열(min sample split) 3으로 설정하였다[8].

8) 훈련, 시험 데이터를 나누기 전에 변수선택 과정을 거치면 실제보다 더 예측력이 높게 판명

〈그림 2〉의 상대적 중요 변수를 보면 무작위 숲에서는 격자탐색방법으로 초매개변수를 조율할 때 의사결정 나무와 마찬가지로 위험 점수와 나이가 상대적으로 가장 중요한 변수로 나타났다. 무작위 탐색 방법을 사용할 때는 8분기의 외래방문과 위험점수가 상대적으로 중요한 변수로 나타났다. 이와 달리 로지스틱 회귀분석에서는 6분기에서의 약국방문과 1분기에서의 이동진료소 방문이 가장 큰 결정계수로 나타났다.

〈그림 2〉 상대적으로 중요한 변수 (로지스틱 회귀분석과 무작위 숲 비교)

Logistic regression (M0)		Randomized search (M1)		Grid Search (M2)	
Feaetures	Coefficients	Feataures	Importances	Feataures	Importances
'POT_VISIT_01_Q06',	2.24	'POT_VISIT_21_Q08',	0.03	'RECON_MA_RISK_SCORE_NBR'	0.02
'POT_VISIT_15_Q01',	1.78	'RECON_MA_RISK_SCORE	0.02	'AGE',	0.01
'RECON_MA_RISK_SCORE_NBR',	1.46	'POT_VISIT_23_Q08',	0.02	'RECON_RX_RISK_SCORE_NBR',	0.01
'HOSPICE_IND=N',	1.46	'POT_VISIT_12_Q06',	0.02	'Length_residence',	0.01
'POT_VISIT_60_Q07',	1.07	'POT_VISIT_12_Q08',	0.02	'Pct_above_poverty_line',	0.01
'MCO_PROD_TYPE_CD_LPPO',	-0.67	'FLAG_RES_FAIL_2015',	0.01	'Home_value',	0.01
'POT_VISIT_62_Q03',	0.66	'POT_VISIT_12_Q01',	0.01	'Population_density_centile_ST'	0.01
'MCO_PROD_TYPE_CD_HMO',	-0.65	'FLAG_CV_CAD_2014',	0.01	'Population_density_centile_US	0.01
'POT_VISIT_71_Q05',	0.64	'FLAG_CV_CHF_2015',	0.01	'POT_VISIT_12_Q06',	0.01
'POT_VISIT_15_Q05'	-0.56	'AGE'	0.01	'POT_VISIT_11_Q08'	0.01



c. 서포트 벡터 기계(Support Vector Machine, SMV)

서포트 벡터 기계는 커넬(kernel, 벡터부등식)을 통해 데이터를 고차원의 벡터 공간에 표현한 후, 데이타 간 거리(margin)를 최대화하여 각 데이터를 분류하기 위한 최적의 선형 결정경계 (decision boundary, decision surface), 즉 초평면 (hyper-plane)을 찾는 기법이다. 로지스틱 회귀분석은 설명변수들이 주어졌을 때 반응변수의 조건부

되는 결과가 나올 수 있으므로, 시험 데이터 안에서 변수선택 과정을 거칠 수 있다.

확률을 구하나, 서포트벡터기계는 확률 계산 없이 예측만을 하고, 데이터 포인트 중 서포트 벡터, 즉 입원과 비입원의 경계선상에 있는 환자의 데이터를 주로 활용하기 때문에 효율적이고 예측력이 좋다. 특히 표본의 개수에 비해 차원이 큰 경우에 장점이 있다. 과적화를 안정적으로 해결하며, 입력 데이터의 작은 차이에 민감하지 않아 비선형적 관계의 모델링하는데 좋다. 그러나 서포트 벡터와 같은 초매개 변수를 설정하여 결정 표면을 정의하는 최적화의 문제가 간단치 않고, 모형 해석도 쉽지 않다.

여기서는 격자탐색법을 이용하여 초매개변수를 세팅하였으며, 데이타를 고차원적인 공간에 투시할 커넬 함수로 비선형 함수인 방사기저함수 (Radial Bias Function, RBF), 다항식, 지그모이트(Sigmoid) 커넬을 시도하였다. Hsu & Chang (2003)에 따라 c(오류 허용값)과 감마(gamma, 결정 경계의 유연성)을 변화시켰으며, C 10과 감마 1, 그리고 5차 다항식 커넬에서 가장 재현율이 높았다.

② 딥러닝: 다중 신경망학습(Deep Neural Network)

신경망 학습은 생물학의 신경망에서 영감을 얻은 방법으로, 행렬곱셈의 시리즈를 설명변수에 적용하여 종속변수를 분류한다. 신경망 모델은 입력층, 은닉층, 출력층으로 구성되어 있고, 은닉층이 다층으로 된 인공신경망을 학습하는 알고리즘을 딥러닝(deep learning) 이라고 한다. 입력값은 데이터의 특성을 나타내는 값으로 이루어져 있고, 각 층에서 순입력 함수값 (특성값에 가중치를 더한 합)을 계산 후 활성함수(activation function)에 의한 결과값과 실제값이 허용오차 이내가 되도록 경사하강법(gradient descent) 알고리즘을 사용, 각층에서 가중치를 업데이트 한다. 과적화 방지를 위한 정규화 기법으로, 학습시 무적위로 일정한 비율의 노드들을 비활성화 하는 하차(drop-out) 기법 등을 사용한다.

딥러닝은 언어, 이미지, 오디오 등 비정형적인 데이터를 다룰 수 있고, 데이터의 잡음에 크게 구애받지 않으며 차원축소(dimension reduction)에 획기적이다. 그러나 많은 연산량으로 데이터 훈련 시간이 길고 초매개변수가 많으며 결과에 대한 직관적 설명이 어렵다.

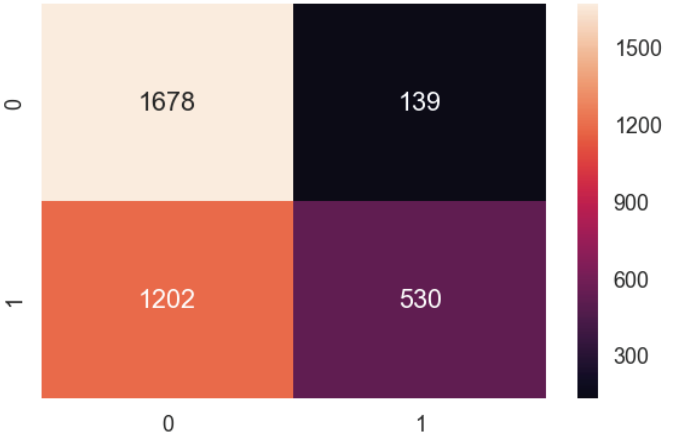
d. 순전파 다층신경망(Feedforward Multi-Layer Perceptron, MLP)

인공신경망의 단층 퍼셉트론은 비선형으로 분리되는 데이터에 대해서는 제대로 된 학습이 어려우므로 대신 순전파 다층퍼셉트론을 활용할 수 있다. 입력 층에서 은닉층으로 이동시 각 입력에 가중치가 곱해지고 가중치의 합이 각층의 출력 값이 되며, 이

는 다음 층의 입력 값이 된다. 전형적인 순전파 MLP는 이전 층의 노드(node)와 다음 층의 노드가 완전히 연결되어 있다. 지도학습에서는 출력 층에서 발생하는 오차 값을 이용하여 은닉층으로 역전파(backpropagation) 시켜 은닉층의 가중치를 업데이트 하게 된다. 완만하게 증가하는 지그모이드(Sigmoid) 함수를 사용하면 은닉층이 깊어질수록 가중치가 0으로 수렴되는데 이를 방지하기 위해 최근 Relu(Rectified linear unit, ReLu, f(x)=max(0,x)) 알고리즘을 사용한다.

한번 훈련을 반복할 때의 표본 사이즈(batch size)는 256, 순전파와 역전파를 1회 완료 하는 순환 학습(epochs) 횟수는 10세대로 설정하였으며, 배치정규화(batch normalization)를 도입하여 학습속도를 개선시켰다. 이는 또한 초기 값의 의존도를 줄이며, 기울기가 줄어드는 것을(gradient vanishing)을 방지하고자 했다. 또한 하차 기법으로 과적합을 방지하여 모델의 일반화 가능성을 높였다. 클래스 불균형을 맞추기 위해 전체 39,549개의 샘플과 833개의 특성을 이용했고, 훈련 세트 73%, 검증세트 18%, 시험세트 9%로 나누었다. 전체 348,674개의 모수들 중 훈련이 가능한 347,138개의 모수를 훈련시켜 〈그림 3〉의 혼동행렬(confusion matrix)을 얻었다. 과적화와 과소적화 없이 적절히 훈련되었음을 그림 4의 정확도 곡선 및 정보손실 곡선이 보여준다.

〈그림 3〉 다층신경망 분석으로 얻은 혼동행렬

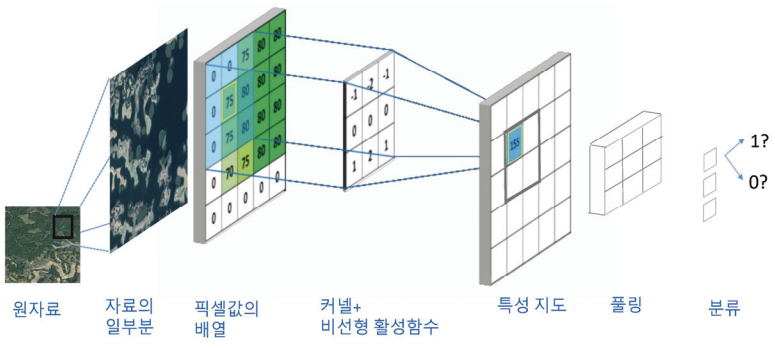


〈그림 4〉 순환학습 횟수별 정확도 (왼쪽) 및 순환학습 횟수별 손실 곡선 (오른쪽)

e. 합성곱신경망(Convolution Neural Network, CNN)

합성곱신경망은 가공되지 않은 데이터로부터 자동으로 특성을 추출하고 학습하는 대표적인 알고리즘이며 이미지 인식에 우수한 성능을 보인다. CNN은 완전히 연결된 층들 이전에 합성곱층(convolution layer)과 풀링층(pooling layer)이라고 하는 새로운 층을 추가함으로써 여과처리(flitering)된 이미지에 대해 효과적으로 분류연산이 수행되도록 구성된다. 커널은 데이터에서 중요한 상위차원의 특성들을 (예: 이미지의 모서리, 모양, 위치 등) 파악하는 역할을 하며 이러한 합성곱 채널의 출력을 특성지도라고 부른다 〈그림 5〉. 풀링층은 하위 표본을 추출하는 방법으로 여러 개의 픽셀들을 하나의 픽셀로 지도화하여 차원을 감소시킬 수 있다. 이를 반복적으로 거치면, 추출된 주요 특징은 평평한 층(flatten layer)을 통해 벡터의 1차원 자료로 변환되어 완전결합된 순전파 신경망에 전달되고 다음으로는 다층신경망과 같이 순전파와 역전파 작업이 되풀이 된 후 분류작업이 일어난다.

〈그림 5〉 미얀마 마을의 위성사진을 CNN으로 분석하는 과정



합성곱망의 장점은 이미지와 같은 원 자료를 처리가 손쉬운 데이터로 축소할 때, 예측에 필요한 주요 특성을 보존하는 행렬 형태의 데이터를 입력 받기 때문에 인접 픽셀간의 공간적 시간적 자기상관성(autocorrelation)을 이용할 수 있다. 또한 커넬과 활성함수가 이미지 전체에 적용되는 것이 아니라 일부에 적용되므로, 완전히 연결된 신경망처럼 기하급수적으로 많은 모수들이 필요치 않아 빠른 연산이 가능하다. 하나의 필터가 입력 이미지를 순회하면서 적용된 결과 값을 모아 출력 이미지가 생성되므로 적용 가중치가 동일하고(parameter sharing, 모수 공유) 이는 학습해야 할 가중치 수를 현저히 줄인다.

본 분석에서는 단차원적 합성곱층(1-D convolution layer)을 넣어서 분석을 하였다. 최적화 알고리즘은 아담(Adam), 손실함수는 범주형 교차 불확실성(categorical cross entropy), 보폭(stride) 은 1, 패딩(padding) 크기는 같게 하여 모델을 평가하였다. 풀링시에는 최대값을 반환하는 형태의 맥스풀링(max-pooling) 기법으로 지역적 사소한 변화에 영향을 받기보다 우세한 특성을 추출하고자 했다.

f. 순환신경망(Recurrent Neural Network, RNN)

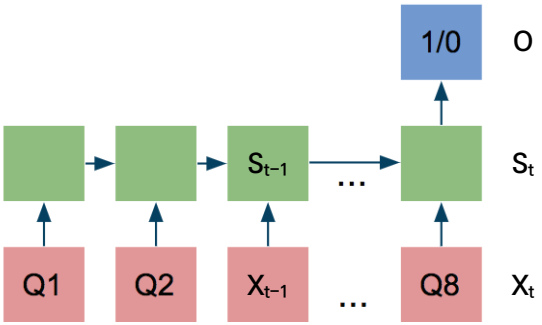
순환신경망은 시퀀스 데이터를 모델링하기 위한 방법으로 내부의 루프(loop)가 정보가 지속되는 것을 도와준다. RNN은 다음 정보가 신경망에 통과될 때 이전 정보 값이 다시 고려되어 연속적 성질을 가진 데이터를 학습할 때 효과적이다. 기존의 네트워크와 다른 점은 은닉상태(hidden state)라는 기존 입력데이터를 요약한 정보를 갖고 있다는 점이다.

〈그림 6〉에서 X_t는 t시간 스텝에서의 입력벡터, S_t는 t시간 스텝에서 기억을 담당하는 은닉상태, O는 1과 0의 값으로 이루어진 출력벡터라고 한다면, St는 입력 X_t와 과거의 기억 S_{t-1}을 조합하여 만들어진다. 조합의 방식은 모수/가중치에 의해 결정되며, 활성함수는 비선형 함수인 Tanh나 ReLu가 주로 사용된다. 출력 벡터를 확률값으로 변환할 때는 소프트맥스(Softmax) 함수를 적용한다.

RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파의 기울기, 즉 그라디언트(gradient)가 점차 줄어드는 문제가 있다. 이를 극복하기 위해 장기의존성을 학습할 수 있도록 장단기 기억(Long Short Term Memory, LSTM) 모델을 활용한다. LSTM은 은닉상태에 컨베이터 벨트와 같은 셀 상태(cell-state)의 개념과 어떤 정보를 버릴지 저장할지 출력할지 결정하는 문 개념을[9] 도입하여 그라디언트가 잘 전파되게 한다.

9) 망각문(forget gate), 입력문 (input gate), 출력문(output gate).

〈그림 6〉 순환신경망을 활용한 분석 도식



본 연구는 〈그림 6〉과 같이 8개의 쿼터별로 각 환자의 질환 및 이용한 의료기관과 서비스에 대한 시퀀스 데이터(sequence data)를 활용하므로 RNN 도입이 가능하다. 이론적으로도, 의료서비스 이용이 시간에 따라 증가했다면 해당 환자의 입원 가능성이 증가할 것으로 가정할 수 있다.

이러한 모델을 훈련시키기 위해서, 데이터를 각 쿼터별로 재정비 하였다. 이는 종단 자료를 긴 형태(long form)에서 넓은 형태(wide form)로 바꾸는 과정과 유사하다고 볼 수 있다. 먼저 쿼터별로 의료기관 별 방문 변수와 환자의 질병 코드별 방문 변수를 합하였다[10]. 예를 들어 첫 번째 분기(Quarter 1, Q1)인 2014년 1월과 3월 사이에 일어나는 방문 및 질병관련 변수를 모두 합한 후 0이 아닌 변수를 제외하면 68개였다. 여기에 소득, 성별, 연령과 같은 시간에 따른 큰 변화가 없다고 가정한 데이터를 각 쿼터별로 포함하였다[11]. 이러한 재배열로 인해 변수의 수는 8,292개로 늘어났다. 모델의 구조는 여러 시퀀스의 입력이 1개의 출력을 예측하는 다 대 일 대응이다. LSTM이 이용되었고 64개의 셀이 모델의 기본이었다. 이를 하차층에 30% 제거 비율로 연결하였는데, 이러한 초매개변수는 교차검증에 의해 결정되었다. 마지막으로는 완전 연결된 출력층에 연결하여 이분적 분류에 강한 지그모이드 활성함수에 통과시켰다. 최적화알고리즘(optimizer)은 이분적 불확실성 손실함수(binary cross entropy loss function)에서 훈련된 아담(Adam)을 활용하여 학습의 속도(learning rate)와 방향(gradient)를 개선하고자 했다. 4번째의 순환 학습을 마쳤을 때 가장 높은 정확도를 나타내었다.

---

10) 총의료기관 및 서비스의 종류는 64개이며, 총 환자의 질병코드와 관련된 변수는 34개이다.
11) 연령은 2015년과 2016년에 1년씩 증가하나 본 분석에서는 고정된 것으로 보았다.

3) 모형별 평가

모형별 예측력 평가표 〈표 3〉을 보면 가장 성능이 높은 모형은 80% 가까운 비교적 높은 예측율을 보임을 알 수 있다. 재현율과 정밀도 모두 딥러닝 모델들에서 높게 나왔다. 재현율은 전방향 다중신경망 MLP가 0.78로 가장 높고, 정밀도는 CNN이 0.68로 MLP의 0.60보다 약간 더 높게 나왔다. 재현율과 정밀도를 종합한 F1점수 역시 MLP가 가장 높았다. 모형 분류별로 보면 전통적 로지스틱 회귀분석보다 일반 기계학습인 의사결정 나무가 더 높은 예측력을 보였고, 일반적 기계학습보다 딥러닝인 MLP의 예측력이 더 높았다.

〈표 3〉 모형별 예측력 평가

모형	정밀도	재현율	F1 점수
로지스틱 회귀분석	0.58	0.17	0.26
의사결정 나무	0.58	0.55	0.56
전방향 다중신경망 (MLP)	0.60	**0.78**	**0.68**
합성곱신경망 (CNN)	**0.68**	0.61	0.64
순환신경망 (RNN)	0.57	0.17	0.26
서포트 벡터 기계 (SVM)	0.40	0.41	0.40
무작위 숲	0.55	0.57	0.57

본 분석은 기계학습의 장점으로 알려진 종속변수에 대한 예측력이 전통적 방법보다 높음을 보여준다. 기존연구와 비교했을 때 딥러닝이 예측력, 특히 재현율이 높은 것이 유사했다. 보건의료 분야에서 중증도 보정 재원일수 예측모형을 개발한 박종호 · 강성홍(2019) 및 최병관 외(2018)의 연구에서도 딥러닝이 모형설명력(R^2)과 예측력이 높은 것으로 나왔다. 위 연구에서도 회귀분석, 의사결정나무, 랜덤포레스트, 서포트 벡터 회귀분석, 신경망이 비교되었는데 어떤 신경망 모델인지는 정확히 논의되지 않았다. 사회복지 분야에서는 오미애 외(2017)가 인구사회학적 변수를 통해 생계급여 수급여부를 예측한 모의실험에서 CNN을 사용하였는데 본 연구처럼 민감도, 즉 재현율이 일반 MI보다 높았다. 그러나 전반적으로는 앙상블 방법의 하나인 부스팅(boosting)이 우수한 성능을 보였다. 부스팅은 오분류에 대한 훈련을 강화함으로서 편향과 분산을 줄일 수 있는 방법이다. 딥러닝을 활용하지 않은 연구들의 경우 앙상블 방법, 특히 부스팅이 효과적으로 보인다. 실제 인구 사회학적 데이터를 활용한 복지 사각지대 연구인 최현수 외(2018)와 Chouldechova (2018)에서는 특히 XGBoost가 높은 예측력을 보였다.

범죄행정학 관련 데이터를 사용한 Kleinberg et al. (2018)도 그라디언트 부스팅 (gradient boosting) 알고리즘을 사용하였다.

이와 같이 본 연구에서 도출한 가장 적합한 모델이 다른 유사한 데이터를 사용한 연구로 일반화 될 수 있느냐 하는 부분은 이 연구의 한계일 것이다. 위험 예측 연구별로 데이터 셋의 차이, 데이터 전처리 및 분석방법의 차이가 존재하며, 모델의 선택과 구축에 있어 가치적, 자의적 판단이 들어가기 때문이다. 대개는 정확한 예측, 효율적 분석, 투명성, 공정성, 책무성이라는 윤리적 판단의 상충관계 속에 각 분석의 목적에 가장 부합하는 선택을 해야 하고, 이는 상황이나 맥락에 따라 달라질 수 있다.

다만 분야별로 이러한 연구가 쌓이면서 일반적인 교훈을 도출해 가는데 도움을 줄 수 있을 것이다. 본 연구와 기존 연구를 종합해 보면, 예측력이 비교적 우수한 두 모델은 딥러닝과 앙상블 모델이라고 보인다. 고위험 대상을 찾아내어 예방적 대응을 하는 자체가 중요한 과제는 딥러닝과 같이 예측력이 높은 모형이 유용할 것으로 보이고, 주요 변수의 영향도를 파악하여 대응 방안에 대한 시사점을 얻고 이를 정책 결정자들과 의사소통 할 때는 앙상블과 같이 변수를 보여주는 모형이 유용해 보인다. 기술적으로는 어떤 종류의 딥러닝 모델을 어떻게 활용했는지에 관한 차이 역시 존재할 것이지만 선행연구에는 딥러닝의 종류나 분석과정이 자세히 나와 있지 않아 자세한 비교가 어렵다.

본 분석에서 딥러닝 중 하나인 MLP의 높은 성과는 매우 복잡하게 연결되어 있는 비선형적인 관계를 포착하는데 우수하기 때문일 것으로 보인다. 그러나 예상과 달리, 보다 진화된 형태의 딥러닝인 RNN과 CNN의 예측력이 기본적 딥러닝인 MLP보다 낮았다. 여러 가지 요인이 있겠으나 RNN의 경우는 데이터 셋의 크기에 비해서 모델이 지나치게 복잡한 면이 있을 것이다. 데이터를 8개의 쿼터로 쪼갬으로서 특성공간의 수를 급격히 늘렸지만, 시계열성에서 얻는 기억 학습의 이득이 크게 없을 수 있다. 예를 들어 환자의 의료기관 방문 수나 패턴이 중요하기 보다는 환자의 상태의 심각성이 중요해서, 그 심각성이 특정 임계점을 넘으면 입원이라는 사건이 발생한다고 볼 수 있다. 몇 번의 학습 이후 과적화가 일어나는 것을 보면 지나치게 복잡한 모델을 적용하기에는 데이터가 작다고 볼 수도 있다. CNN의 경우는 이미지와 같이 서로 연관된 부분의 공간 상관성이 크거나 경우나 지엽적인 특징을 포착하는 데 높은 성능을 보인다. 그런데 데이터에서 지엽적 특징이나 인접 데이터 간 시공간적 의존성이 뚜렷치 않다면 CNN의 알고리즘이 예측력 향상에 큰 영향을 미치지 않은 것으로 추측해 볼 수 있다.

4) 정책적 함의

본 분석은 환자의 사회 인구학적 및 보건 의학적 데이터를 통해 입원율을 80% 정

도 예측할 수 있는 것으로 나타났다. 이러한 시험적 예측모형은 예방적 서비스의 수요를 파악하고 이를 적절히 분배하는 데 활용될 수 있다. 즉 입원 위험도가 높은 환자를 대상으로 하는 비용효과적이고 예방적 개입을 할 수 있을 것이다. 그 예시로는 처방약을 제때 타는지, 처방법과 용량을 준수하는지, 정기적으로 의사와 상담하는지 등을 체크해 보고 주위를 환기시키거나(nudge) 관련된 위험 질환에 대한 조기 검사(screening)를 하는 것이다[12].

이 모델을 비롯하여 앞으로 개발될 모델의 적용 가능성은 정밀도, 재현율 등의 예측력에 달려 있을 것이다. 예를 들어 아래와 같은 부등식을 만들어 볼 수 있다.

정밀도× (병원 입원비용 − 예방적 개입의 비용) − (1 − 재현율) ×입원율 비용

이는 예측모델이 선정한 환자에 대한 예방적 케어를 함으로서 절약되는 비용이 예측모델이 미처 찾아내지 못한 환자에 대해 치료하는 비용보다 클 때 유용하다는 것을 보여준다. 여기서 비용효율성을 높게 하는 방법은 정밀도와 재현율을 높이는 것이다.

특정 기계학습 모형은 입원율을 낮추기 위한 방안과 관련된 설명변수에 대한 정보를 보여준다. 입원율과 같은 위험과 연관이 되어 있는 직·간접적인 변수들이 의사결정나무나 랜덤포레스트에서는 중요 특성이라는 지표로 나타나기 때문이다. 본 연구에서는 전통적 회귀분석과 기계학습 모두, 질병의 종류와 관련된 컨디션코드(condition code) 보다는 의료 서비스를 받은 장소가 입원 예측에 중요하다고 나타났다. 차이점으로는 회귀분석은 2015년 4-6월 사이 외래 방문을 높은 위험으로 보는 반면, 의사결정나무와 랜덤포레스트 같은 기계학습은 모두 미국 보건복지부 보험국이 계산한 의학적 위험 지수가 입원율과 관련성이 높다고 나타냈다. 이는 의학적 위험지수라는 전통적 지표가 입원율과 관련이 있다는 것을 보여주며, 기계학습은 이러한 지표를 검증하고 향상시키는 데도 사용될 수 있을 것이다.

회귀 분석은 대개 이론과 선행연구를 바탕으로 수작업을 통해 연구자가 이미 선택되어 있는 변수를 가지고 분석하기 때문에 이미 알려져 있지 않은 의외의 요인들을 찾아내기 쉽지 않을 수 있으며 특히 표본이 적은 경우는 다룰 수 있는 변수 수의 제약이 클 것이다. 또한 타 변수들에 의해 효과가 상쇄되기 쉬운 변수, 직접관찰에 의한 파악이 어려운 잠재변수, 다중공선성을 가진 변수들이나 교란변수가 존재 경우, 교호작용이 복잡할 경우 전통적 설명요인을 보완할 수 있는 새로운 통찰을 줄 가능성이 있을 것이다.

---

12) 이러한 일상의 자기관리 부족은 심각한 문제를 초래할 수 있는데, 이런 문제를 다루기 위해 Clover Health 와 같은 보험회사가 생겼을 정도이다.

Ⅳ. 결론

본 연구는 데이터사이언스적인 접근을 적용하여 공적의료보장 제도 가입자들의 위험률을 예측하는 모형을 구축하였다. 이러한 시도는 방대하게 축적돼 온 공공데이터를 활용하여 비교적 신속하고 일관성 있게 서비스 대상자를 파악하고 개입해야 할 때 유용할 것이다. 특히 더 가속화되고 있는 비대면 시대에 AI 기술이 접목되어서 가장 필요한 대상에게 우선적으로 서비스가 배분된다면 희소한 사회복지 자원의 분배에 도움이 될 것이다. 이러한 연구는 정부의 「한국판 뉴딜」 종합계획을 실행하는 데도 참고가 될 수 있을 것이다.

본 분석은 해외 데이터를 사용하였지만, 우리나라 공공데이터의 분석과 활용에도 시사점이 있을 것이다. 기존의 사회보장 연구는 주로 보건사회연구원, 사회보장원 등과 같은 공공기관이나 국책기관에서 혹은 대학과 연계해서 이루어진 경향이 있었다. 앞으로는 개별연구자 단위에서도 보건의료 빅데이터 개방 시스템이나 보건복지부 R & D 사업, 그리고 기관 연계 맞춤형 데이터베이스[13] 등의 플랫폼을 활용하면 다양한 연구가 가능할 것으로 본다. 후속 연구로는 작은 규모의 복잡한 행정 데이터, 시퀀스의 길이가 다른 데이터, 불규칙적 샘플링, 많은 결측치가 있는 데이터와 같이 보다 더 도전적인 사례를 가지고 분석해 볼 수 있을 것이다.

본 연구의 함의는 데이터 사이언스를 활용한 사회과학 연구, 그리고 이를 정책에 적용하는 것이라 볼 수 있다. 본 고에서는 데이터를 설계 수집하고, 예측 모형을 개발, 실행, 평가, 환류하는 것은 분야 전문성, 기계학습, 통계가 유기적으로 연계되는 영역임을 논했다.

먼저 연구에 필요한 데이터를 설계하여 수집하거나 이미 존재하는 데이터를 발견, 연계, 병합해서 활용할 때는 분야 전문성이 중요할 것이다. 이 전문성에는 공공의 가치창출이라는 목적과 더불어 연구 설계에 필요한 학술적 이론적 지식, 데이터 접근 가능성, 데이터 획득 과정의 윤리성이 조화되어야 한다. 케냐 통신위원회는 말라리아 유병률 추적의 명분하에 15만 케냐 국민의 통신 데이터를 가입자의 동의 없이 하버드 공중보건대학원에 제공하였는데 (Nation, 2013), 이는 데이터 획득의 합법성과 투명성이 부족한 예이다. 한편 연구목적 및 필요성과 별도로 일반 연구자들에게는 데이터 자체에 대한 접근 장벽이 클 수 있다. 예를 들어 코로나 바이러스 감염경로 추적 연구 시 통화기록, 신용카드 사용기록, CCTV 영상과 같은 민감한 데이터는 접근이 어려울

13) 건강보험공단, 건강보험 심사평가원, 질병관리 본부 등

수 있으므로, 출처와 경로를 다각화 한 데이터 발굴이 필요하다. 일단 데이터에 접근을 하였더라도 데이터의 특성에 대한 지식이 필요하다. 데이터가 포함하지 않은 정보, 따라서 '데이터 사각지대'에 있는 집단의 특성을 파악하는 것이 그 예이다. 자원분배를 위해서 예측이 필요한 경우는 데이터에 빠진 집단이 더 취약한 계층일 수 있다. 예를 들어 사회복지 사각지대 예측모델의 경우 주거정보가 중요한 축인데 노숙자와 같은 취약계층은 모델 개발에 반영이 되기 어렵기 때문이다. 반대로 제재를 위주로 하는 범죄위험율 예측의 경우 데이터에 취약계층의 특성이 과도하게 반영되는 편향에 대해 고려해야 할 것이다.

다음으로, 모형의 개발과 평가에 있어서는 기계학습과 통계적 지식이 주로 활용되지만, 분야 전문성 역시 뒷받침 되어야 할 것이다. 대표적으로 데이터 전처리 과정에서 설명변수와 반응 변수를 선택하고 이들을 분석에 최적화된 형태로 변환 하는 과정을 생각해 볼 수 있다. 예를 들어 변수 선택에 있어서는 학술적 이론, 선행연구, 회귀분석의 상관계수, 전문가 의견 수렴, 이해관계자 표적 집단면적 등을 통해 미리 선택, 정제한 변수를 사용한 경우와 모든 변수를 활용한 경우, 알고리즘이 자동으로 선택해 준 변수를 넣은 경우를 비교하여 최적의 모형을 개발할 수 있을 것이다. 또한 연구나 정책 필요에 따라 데이터를 특성별로 층화하여 성별, 연령별, 가구 형태별, 지역별, 위험정도별 등으로 특화된 모델을 만들 수도 있을 것이다. 이를 위해서는 다학문적 연구팀의 구성이 필요할 것이다.

모형의 평가에서는 기술적인 면과 이론적 판단의 균형이 필요할 것이다. 예측하고자 과제, 위험의 종류, 예측을 필요로 하는 위험 대상인구의 규모 등에 따라 오류행렬의 종류와 가중치, 예측목표 등을 달리해야 하기 때문이다. 어떤 평가 지표를 사용할 것인지는 먼저 타깃변수 유형이 범주형인지 연속형인지, 예측의 질을 어떻게 판단할 것인지에 따라 다르다. 본 분석과 같은 입원율을 파악하는 과제와 신용점수를 예측하는 과제, 3차원에서 어떤 물체의 위치를 파악하는 과제는 서로 다른 평가지표를 필요로 할 것이다. 또한 위험의 종류에 따라 서로 상충관계에 있는 지표의 가중치를 조절해야 한다. 다수의 재입원 환자나 반복되는 아동학대의 경우, 안전과 직결된 위기 상황에서는 예산이 많이 발생하더라도 민감도를 높이는 것이 특이도를 높이는 것보다 중요할 것이다. 대상 인구 규모에 따라서도 얼마나 정확한 예측을 필요로 하는지 기준이 달라질 것이다. 본 분석에서 8,696 명의 입원환자 중 오분류 가능성이 있는 20%는 174명이지만 특정 지자체의 주민 혹은 국민전체를 대상으로 할 때는 1%도 절대적으로 큰 수치가 될 수 있다.

이와 함께 모델 예측 다음에는 어떤 대응을 할 것인가 하는 정책적 영역도 있다. 예를 들어 건강 위험신호가 높은 사람들에게 자동으로 문자 메세지를 보내 관련된 서비스를 안내하는 것은 비교적 비용 효과적이고 많은 이들에게 제공할 수 있는 활동이다. 이와 달리 위험군으로 분류된 가입자들에게 의료관계자가 전화를 해서 개별 상담을 하거나, 더 나아가 선별적 공공 의료 서비스를 제공하는 것은 다른 차원의 개입이 될 것이다. 의학정보나 바람직한 행동을 환기시키는 자동안내는 실질적 효과가 작을 수도 있지만 환자 분류를 잘못했을 때의 영향 또한 크지 않을 것이다. 반면 중요한 예방적 의료서비스가 예측환자에게만 제공된다면, 효과 못지않게, 잘못된 분류의 파급력이 클 수 있다.

예측 결과를 해석하고 이를 실제에 활용하기 위해서는 일선 기관의 리더십과 함께 제도적 뒷받침이 필요할 것이다. 먼저 일선기관에서 위험도 지수를 보여주는 모델을 활용한다면, 위험도 지수의 점수나 등급, 개입이 필요한 기준점을 실무적으로 결정해야 할 것이다. 또한 담당자들이 기계의 신호를 얼마나 신속히, 어떤 가중치를 두고 반영해야 하는지, 자율적으로 따를 것인지, 강제력이 있는 것인지에 대한 판단도 기관 및 지자체 자원, 넓게는 국가 차원에서 내려야 할 것이다. 예를 들어 우리나라에서는 2018년 3월부터 e아동행복지원시스템을 통해 위험 징후 정보를 연계하여 위기 아동을 발굴, 개입해왔다. 그러나 위기 아동으로 발굴된 후에도 지자체별로 대응이 다르며 가정방문조사가 이루어지지 않는 경우, 이를 보완할 어떤 장치가 필요한지는 사회적으로 합의되어야 할 부분이다.

마지막으로 평가의 환류를 통한 모형의 지속적인 활용을 위해서는 부문 간 협력이 중요하다. 예측과제가 복잡해질수록 고도화된 머신러닝 기술을 활용하기 위해서 민간부문의 역할이 커질 것이다. 다만 이를 위해서는 긴밀한 민관협력 및 산학협력이 전제되어야 한다. 미국 일리노이주 아동국에서는 위기가정이 아니라고 예측된 가정의 아동이 사망하는 사고가 있었으나, 모형 개발 업체가 알고리즘의 공식을 밝히기를 거부하여 결국 알고리즘의 사용이 중지되었다 (Hurley, 2018). 이 사례는 모형 개발을 순수한 기술의 영역에만 맡길 경우 유기적이고 투명한 평가 및 환류의 과정이 어려울 수 있음을 보여준다. 결론적으로 데이터 사이언스의 다 학문적, 다부문적 접근은 모형 활용의 전 과정에서 사회과학 연구자들과 정책 입안자들이 공적 가치 창출이라는 목적을 달성할 수 있도록 충분한 관여를 해야 함을 의미한다.

▣ 참고문헌

관계부처합동. 2020(7월 14일). ≪한국판 뉴딜 종합계획: 선도국가로 도약하는 대한민국의 대전환≫.

권설아 · 김지은 · 이재은. 2016. “Future Directions of Research on Crisis Management Using Big Data.” ≪Crisisonomy≫, 12(10): 133-148.

김기환. 2013. “공공부문 빅데이터의 활용성과 위험성.” ≪정책분석평가학회보≫, 23(2): 1-27.

김병조 · 은종환. 2020. “행정-정책 의사결정에서 머신러닝 방법론 도입의 정책적 함의: 기계의 한계와 증거기반 의사결정.” ≪한국행정학보≫, 54(1): 261-285.

김선영. 2020. “증거기반정책에서의 빅데이터에 관한 연구.” ≪한국정책학회보≫, 29(1): 69-90.

박종호 · 강성홍. 2019. “머신러닝을 이용한 신경계통 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발.” ≪보건사회연구≫, 39(1): 390-427.

송태민 · 진달래 · 박대순 · 박현애 · 안지영 · 김정선. 2014. “보건복지 빅데이터 효율적 관리 방안 연구.” 연구보고서 2014-05. ≪보건사회연구원≫.

성욱준. 2016. “공공부문 빅데이터 정책 활성화 연구.” ≪한국정책학회보≫, 25(2): 125-150.

송태민 · 송주영. 2016. “소셜빅데이터 기반 보건복지 정책 미래신호 예측.” ≪Journal of Health Informatics and Statistics≫, 41(4): 417-427.

오미애 · 최현수 · 김수현 · 장준혁 · 진재혁 · 천미경. 2017. “기계학습 기반 사회보장 빅데이터 분석 및 예측 모형.” ≪보건사회 연구원≫.

오철호. 2017. “문제제기: 데이터 기반 정책분석평가의 연구와 적용.” ≪정책분석평가학회보≫, 27(2): 155-167.

윤상오 · 김기환. 2016. “빅데이터 시대의 한국과 영국간 개인데이터 활용정책 비교 연구.” ≪한국정책과학학회보≫, 20(1): 29-56.

이동규. 2016. “재난관리 예측적 거버넌스 시스템 구축을 위한 시론적 검토: 미래예측적 이상신호 감지를 위한 협력적 재난관리 의사결정 시스템 제언을 중심으로.” ≪Crisisonomy≫, 12(2): 35-52.

이승용 · 이주락. 2020. “빅데이터와 FDS를 활용한 보이스피싱 피해 예측 방법 연구.” ≪시큐리티연구≫, 62: 185-204.

이은미. 2015. “빅데이터의 정부 의사결정 반영에 관한 탐색적 연구: 사회적 관심의 재난위기단계 적용을 중심으로.” ≪한국정책학회보≫, 24(4): 491-511.

이제복 · 최상옥. 2018. “공공서비스 인공지능 ML 적용과 공공가치.” ≪정부학연구≫, 24(1): 3-27.

임상규. 2014. “빅 데이터를 활용한 스마트 재난관리전략.” ≪Crisisonomy≫, 10(2): 23-43.

지광석. 2014. “공공데이터의 생산과 제공에 대한 정책적 함의: 소비자 공공데이터 사례분석을 중심으로.” ≪국가정책연구≫, 28(3): 323-348.

정예림 · 강정은. 2019. “기후변화 정책 수립 지원을 위한 소셜 빅데이터 분석.” ≪환경정책≫, 27(4): 211-239.

최병관 · 함승우 · 김촉환 · 서정숙 · 박명화 · 강성홍. 2018. “인공지능을 이용한 급성 뇌졸중 환자의 재원일수 예측모형 개발.” ≪디지털융복합연구≫, 16(1): 231-242.

최현수 · 오미애 · 천미경 · 김은하 · 추병주 · 박선미 · 이인수. 2018. “사회보장정보시스템을 활용한 복지사각지대 발굴 확대방안.” 정책보고서 2018-33. 연구보고서 2017-46. ≪보건사회연구원≫.

Agrawal, Ajay, Joshua Gans, & Avi Goldfarb. 2019. “Economic policy for artificial intelligence.” I*nnovation Policy and the Economy*, 19(1): 139-159.

Athey, Susan. 2017. “Beyond prediction: Using big data for policy problems.” *Science,* 355(6324): 483-485.

Atkinson, David, Kumar Bhargav Srinivasan, & Chenhao Tan. 2019. “What Gets Echoed? Understanding the “Pointers” in Explanations of Persuasive Arguments.” *arXiv preprint arXiv:*1911.00523.

Blumenstock, Joshua, Gabriel Cadamuro, & Robert On. 2015. “Predicting poverty and wealth from mobile phone metadata.” *Science*, 350(6264): 1073-1076.

Brady, Henry E. 2019. “The challenge of big data and data science.” *Annual Review of Political Science*, 22: 297-323.

Breiman, Leo. 2001. “Random forests.” *Machine learning*, 45(1): 5-32.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane

Whitmore Schanzenbach, & Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics,* 126(4): 1593-1660.

Chouldechova, Alexandra, Emily Putnam-Hornstein, Dianan Benavides-Prado, Oleksandr Fialko, & Rhema Vaithianathan. 2018. January. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134-148).

Clark, William Roberts, & Matt Golder. 2015. "Big data, causal inference, and formal theory: Contradictory trends in political science?: Introduction." *PS: Political Science & Politics.* 48(1): 65-70.

Coglianese, Cary, & David Lehr. 2016. "Regulating by robot: Administrative decision making in the machine-learning era." Geo. LJ, 105, 1147.

Ellen, Ingrid Gould, Keren Mertens Horn, & Amy Ellen Schwartz. 2016. "Why don't housing choice voucher recipients live near better schools? Insights from Big Data." *Journal of Policy Analysis and Management* 35(4): 884-905.

Fu, Runshan, Yan Huang, & Param Vir Singh. 2018. "Crowd, Lending, Machine, and Bias." available at https://doi.org/10.2139/ssrn.3206027

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, & Li Fei-Fei L. 2017. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." *Proceedings of the National Academy of Sciences*, 114(50): 13108-13113.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, & Jenny Huang. 2020(January). "Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?." In Proceedings of the 2020 *Conference on Fairness, Accountability, and Transparency* (pp. 325-336).

Grimmer, Justin. 2015. "We are all social scientists now: How big data, machine learning, and causal inference work together." *PS, Political Science & Politics*, 48(1): 80.

Hsu, Chih-Wei, Chih-Chung Chang, & Chih-Jen Lin. 2003. *"A practical guide to support vector classification."*

Hurley, Dan. 2018(January, 2). Can an Algorithm Tell When Kids Are in Danger? *The New York Times*.

Jean, Neal, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, & Stefano Ermon. 2019(July). "Tile2Vec: Unsupervised representation learning for spatially distributed data." In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 3967-3974).

Jung, Woojin. (2019a). *Combating Poverty through Aid: Critical Analysis of Alternative Models*, [Doctoral dissertation, University of California, Berkeley]. ProQuest Dissertations Publishing.

_____. (2019b, November). Mapping community development projects: Spatial analysis in Myanmar. Paper presented at the conference of *Association for Public Policy Analysis & Management*, Denver, CO

King, Gary, Jennifer Pan, & Margaret E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science*, 345(6199): 1251722.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, & Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *The quarterly journal of economics*, 133(1): 237-293.

Kotu, Vijay, & Bala Deshpande. 2014. Predictive Analytics and Data Mining: Concepts and Practice with Rapidminer. Waltham, MA: Morgan Kaufmann.

Kum, Hye-Chung, Joy C. Stewart, Roderick A. Rose, & Dean F. Duncan. 2015. "Using big data for evidence based governance in child welfare." *Children and Youth Services Review*, 58: 127-136.

LeCun, Yann, Yoshua Bengio, & Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553): 436-444.

Milusheva, Sveta. 2020. "*Using Mobile Phone Data to Reduce Spread of Disease*." World Bank.

NIST Big Data Public Working Group. 2015. "Big data interoperability framework: Volume 1, definitions." *Gaithersburg, MD: National Institute of Standards and Technology.* 1500-1. https://bigdatawg.nist.gov/

_uploadfiles/NIST.SP.1500-1.pdf

Onnela, Jukka-Pekka, Benjamin N. Waber, Alex Pentland, Sebastian Schnorf & David Lazer. 2014. "Using sociometers to quantify social interaction patterns." *Scientific reports*, 4: 5604.

Pan, Ian, Laura B. Nolan, Rashida R. Brown, Romana Khan, Paul van der Boor, Daniel G. Harris, & Rayid Ghani. 2017. "Machine learning for social services: a study of prenatal case management in Illinois." *American journal of public health*, 107(6): 938-944.

Pan, Yue, Hongmei Liu, Lisa R. Metsch, & Daniel J. Feaster. 2017. "Factors associated with HIV testing among participants from substance use disorder treatment programs in the US: A machine learning approach." *AIDS and Behavior*, 21(2): 534-546.

Shroff, Ravi. 2017. "Predictive Analytics for City Agencies: Lessons from Children's Services." *Big data*, 5(3): 189-196.

Stephens-Davidowitz, Seth. 2014. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics*, 118: 26-40.

Wilson, Daniel L., Jeremy Coyle, Angeli Kirk, Javier Rosa, Omnia Abbas, Mohammed Idris Adam, & Ashok J. Gadgil. 2016. "Measuring and increasing adoption rates of cookstoves in a humanitarian crisis." *Environmental science & technology*, 50(15): 8393-8399.

Nation 2013(August 10, Saturday). US scientists 'spied' on phone users.

Using Data Science to Strengthen the Social Safety Net: Predicting Risk for Medicare and Medicaid Insurers*

Woojin Jung

This paper explores how data science can be used to strengthen the social safety net. As one such approach, the paper develops a model to predict the likelihood of hospital admission for Medicare and Medicaid insurers in the U.S. The analysis draws from the health insurance claims data of 45,000 patients with 939 features, spanning eight quarters from 2014 and 2015. Six models are adopted to predict patients' hospital admissions in 2016, based on their sociodemographic and health-related characteristics. The paper presents the rationale, processes, and results of analysis from logistic regression, Decision Tree, Random Forest, Support Vector Machine, feed-forward Multi-layer Perceptron (MLP), Convolution Neural Network, and Recurrent Neural Network. The best performing model, evaluated against recall and precision scores, is the MPL. This simple deep learning model was correct about 80% of the time for patient admissions to hospital. Additionally, tree-based algorithms provide important features related to hospital admission, such as medical risk scores. As a policy implication, the paper discusses predictive risk modeling to provide preventive care for at-risk populations. The paper concludes by suggesting strategies for using the data science approach in the allocation of social welfare programs and services.

※ Key Words: Data science, Artificial intelligence, Big data, Social safety net

---

* The analysis of this paper was conducted as a part of Data Minig and Analytics Project at University of California Berkeley In 2018. Woojin Jung, David Proudman, Kyungna Kim, Zhiling Pan, Dunyang Wang, and Michael Fermanian participated in the project.