J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

머신러닝을 이용한 정부통계지표가 소매업 매출액에 미치는 예측 변인 탐색: 약국을 중심으로

Exploring the Predictive Variables of Government Statistical Indicators on Retail sales Using Machine Learning: Focusing on Pharmacy

이 광 수[1*]
Gwang-Su Lee

요 약

본 연구는 데이터, 네트워크, 인공지능을 기반으로 산업 생태계 조성을 위해 구축된 정부통계지표가 약국 매출액에 영향을 미치는지 머신러닝을 이용하여 변인을 탐색하고 약국 매출액 예측에 적합한 분석 기법을 제공하고자 한다. 이에, 본 연구는 28개 정부통계지표와 소매업종인 약국을 대상으로 2016년 1월부터 2021년 12월까지의 분석 데이터를 활용하여 머신러닝 기법인 랜덤 포레스트, XGBoost, LightGBM, CatBoost을 통해 예측 변인 및 성능을 탐색하였다. 분석결과 경기관련 지표인 경제심리지수, 경기동행지수순환변동치, 소비자심리지수는 약국 매출액에 영향을 미치는 중요한 변인으로 나타났고, 회귀성능은 지표 MAE, MSE, RMSE를 살펴본 결과 랜덤 포레스트가 XGBoost, LightGBM, CatBoost 보다 성능이 가장 우수하게 나타났다. 이에, 본 연구는 머신러닝 결과를 토대로 약국 매출액에 영향을 미치는 변인과 최적의 머신러닝 기법을 제시하였으며, 여러 시사점과 후속연구를 제안하였다.

☞ 주제어 : 머신러닝, 랜덤 포레스트, XGBoost, LightGBM, CatBoost, 정부통계지표

ABSTRACT

This study aims to explore variables using machine learning and provide analysis techniques suitable for predicting pharmacy sales whether government statistical indicators built to create an industrial ecosystem based on data, network, and artificial intelligence affect pharmacy sales. Therefore, this study explored predictive variables and performance through machine learning techniques such as Random Forest, XGBoost, LightGBM, and CatBoost using analysis data from January 2016 to December 2021 for 28 government statistical indicators and pharmacies in the retail sector. As a result of the analysis, economic sentiment index, economic accompanying index circulation change, and consumer sentiment index, which are economic indicators, were found to be important variables affecting pharmacy sales. As a result of examining the indicators MAE, MSE, and RMSE for regression performance, random forests showed the best performance than XGBoost, LightGBM, and CatBoost. Therefore, this study presented variables and optimal machine learning techniques that affect pharmacy sales based on machine learning results, and proposed several implications and follow-up studies.

☞ keyword : Machine Learning, Random Forest, XGBoost, LightGBM, CatBoost, Government Statistical Indicators

1. 서 론

2016년 세계경제포럼에서 4차 산업 혁명이 언급되고 인공지능, 사물인터넷, 빅데이터 등이 정보통신기술과 융합되면서 사회 전반에 혁신적인 변화를 불러오고 있다. 이에 정부에서는 데이터(Data), 네트워크(Network), 인공지능(AI) 일명 D.N.A를 핵심 인프라로 강조하고 있으며, 데이터 3법 개정, 가명정보·마이데이터 등 신규 제도 도입, 데이터산업법 제정, AI윤리 기준 마련 및 AI 학습용 데이터 구축 등 D.N.A와 관련된 법·제도 등을 마련하여 산업 생태계를 만들어가고 있는 실정이다.

이러한 산업 생태계 조성의 일환으로 국가정책 수립, 점검 및 정책성과 측정 등을 목적으로 38개 중앙행정기관에서 관리하고 있는 지표와 민간의 통계자료 및 행정자료를 토대로 700여개의 지표 정보를 제공하여 사회·경제상황을 한곳에서 파악할 수 있는 웹사이트를 구축·운영 중에 있다. 웹사이트를 통해 관리되고 있는 사회·경제 관련 지표들은 물가, 경기, 고용, 금리 등이 있으며, 이런 지표들은 중소기업 및 가계 경제에도 영향을 미치고 있다.

1 Department of Computer Education, Sungkyunkwan University, Seoul, 110-745, Korea
* Corresponding author (73gslee@gmail.com)
[Received 4 May 2022, Reviewed 17 May 2022, Accepted 6 June 2022]

대표적으로 물가란 시장에서 거래되는 개별 상품의 가격 및 서비스의 요금을 경제생활에서 차지하는 중요도를 고려하여 평균한 종합적인 가격수준을 말하며, 물가변동이란 시장에서 제품에 대한 수요와 공급에 의해 결정된다.

물가변동은 단순히 수요와 공급에 의해서만 결정되는 것이 아니라 원자재의 대부분을 수입에 의존하고 있는 현실에서 국제원자재 가격의 변동이나 환율에 의한 수입품의 가격변동, 유통구조나 경쟁과 같은 시장 구조적 요인, 통화량 등도 물가변동에 미치는 다양한 요인으로 작용하고 있다.

이렇듯 우리는 일상생활에서 물가가 오르고 내리는 것을 느끼며 살고 있으나, 실제 피부로 느끼는 물가와 정부에서 발표하는 물가와는 차이가 있다는 느낌을 받곤 한다. 이는 개인마다 소비하는 상품과 서비스의 조합이 다르기 때문에 개인마다 체감물가와 차이를 보일 수 있고, 개인이 가격을 비교하는 시점과 지표물가의 비교시점이 서로 다른 것도 요인이 될 수 있다.

또한 2019년 하반기에 코로나19가 발생한 후 현재까지 사회·경제 분야에서 직·간접적으로 코로나19의 영향을 받고 있으나, 사회·경제 상황을 반영한 지표들은 단순히 전월, 전분기 또는 전년대비 몇 % 상승 및 하락하고 있다는 수치만 제공하고, 어떤 지표들이 중소기업, 소상공인 또는 어떤 업종에 밀접하게 영향을 주고 있는지 알 수가 없는 실정이다.

이에, 본 연구는 정부에서 발표 및 관리하고 있는 사회·경제관련 지표들이 소매업에 영향을 미치는지를 탐색하고자 한다. 이를 위해 전통적인 통계 기법이 아닌 최근 화두가 되고 있는 머신러닝을 통해 분석을 실시하였으며, 분석 데이터로는 사회·경제관련 정부통계지표를 활용하기 위해 통계청 e-나라지표와 한국은행 경제통계시스템에서 제공하는 통계 지표 28개를 독립변수로 사용하였고, 종속변수로는 소매업종인 약국의 매출액을 사용하였다.

머신러닝 분석을 통해 사회·경제관련 정부통계지표 중 어떤 지표가 약국 매출액에 가장 영향을 미치는 변수인지 탐색하고, 각 머신러닝 기법 간의 성능을 비교함으로써 약국 매출액 예측에 적합한 머신러닝 기법을 제안하고자 한다.

2. 이론적 배경

최근 빅 데이터 관련 연구에서는 전통적인 통계분석 시 한정된 변수 사용으로 중요한 변수를 간과하는 점과 변수가 많을수록 다중공선성의 문제로 성능이 떨어지는 문제점을 개선하고자 머신러닝을 이용한 새로운 통계적 분석 방향을 제시하는 연구들이 활발히 진행되고 있다.

머신러닝을 이용한 연구들은 하나의 머신러닝 기법만을 사용하여 예측변수를 탐색하거나, 다수의 머신러닝 기법을 사용하여 분석 데이터, 예측변수 수, 종속변수 유형 등에 맞는 최적의 머신러닝 기법 제시 및 예측변수를 탐색하고자 하는 연구로 구분되어 진다.

하나의 머신러닝 기법을 사용한 연구로 Go et al.(2018)은 전통적인 다중회귀방법에서 변수가 많을수록 다중공선성의 문제로 성능이 떨어지는 점을 개선하기 위해 교육부와 한국교육학술정보원에서 제공하는 EDSS 에듀데이터를 기반으로 랜덤 포레스트와 다중회귀분석의 설명력을 비교하고, 랜덤 포레스트 기법을 활용하여 학업성취도에 영향을 주는 변수를 분석하였다. 사용된 변수는 학교특성 변수, 교육활동 변수, 동아리 활동 변수 등 78개 변수를 사용하였고, 분석결과 랜덤 포레스트의 설명력은 0.70으로 전통적 통계기법인 다중회귀분석의 설명력 0.58보다 높게 나타났다. 학업성취도에 영향을 주는 변수로 전문대진학생비율, 상반기교과강좌1개당 최대학생수, 전체대학진학생수, 기타진로학생비율 등을 제시하였고, 분석 결과를 기반으로 새로운 통계적 접근 방식을 제시하였다[1].

Lee and Kim(2019)는 장애인의 취업을 예측하는데 영향을 미치는 변수를 탐색하기 위해 머신러닝 기법인 랜덤 포레스트를 사용하였다. 사용된 변수는 한국장애인고용공단 고용개발원의 장애인고용패널조사 데이터를 기반으로 취업을 위한 노력과 지원, 직업적 능력, 취업 태도와 환경 등으로 구성된 179개 변수를 사용하였고, 분석결과 변수 중요도는 1차 연도 조사 당시 기존일자리의 종사상 지위, 업무경력, 전반적인 건강상태, 2016년 가구소득 총계 순으로 나타났다. 분석 결과를 기반으로 장애인의 취업지원 대책을 수립함에 있어 정책 대상 및 내용에 차별화가 이루어질 필요가 있음을 제시하였다[2].

Kim et al.(2019)는 전통적인 회귀분석 시 한정된 변수만을 사용하는 문제점을 개선하고 대학 수시 입학에 영향을 미치는 요인을 탐색하기 위해 머신러닝 기법인 랜덤 포레스트를 사용하였다. 사용된 변수는 한국직업능력개발원 한국교육고용패널 데이터를 기반으로 학교관련, 가정생활, 여가생활, 재학 중 근로경험 등으로 구성된 424개 변수를 사용하였고, 분석결과 변수 중요도는 학교의 서울4년제 대학 진학자수, 수리등급, 언어등급, 외국

어 등급, 지방전문대학진학자수 순으로 나타났다. 분석 결과를 기반으로 부모의 사회경제적 배경, 학생의 가정 배경이 대학 수시 입학에 영향을 미치는 요인이 될 수 있지만 우선 순위에 있어 학교유형, 학업 분위기, 학교규모 등 학교 특성의 영향력이 훨씬 큰 것으로 나타나 향후 수시 전형의 개선을 위한 정책적 우선순위의 방향성을 제시하였다[3].

Cho and Lee(2020)은 수도권 서·남부 지역의 청소년을 대상으로 청소년의 디지털 성범죄 중 유포 및 소비형 가해행위와 관련된 예측 요인을 분석하기 위해 머신러닝 기법인 랜덤 포레스트를 사용하였다. 사용된 변수는 콘텐츠 판매, 구입, 업로드 별로 가출, 온라인도박, 음주, 흡연, 성매매, 성인방송 등 20개 변수를 사용하였으며, 분석결과 판매와 구입 행위에서는 가출, 온라인 도박, 음주, 흡연이 높은 변수 중요도를 보였고, 업로드 행위에서는 성매매, 성인방송, 온라인 도박, 가출이 높은 변수 중요도를 보였다. 분석 결과를 기반으로 세부 디지털 성범죄 가해행위마다 개별화된 접근 방식의 필요성을 제시하였다[4].

다수의 머신러닝 기법을 사용한 연구로 Jho(2018)는 대학에서 이러닝 강의의 학습 데이터를 토대로 학업성취 수준을 예측하기 위해 머신러닝 기법인 KNN, 서포트 벡터 머신, 랜덤 포레스트, 의사결정나무, 그레디언트 부스팅, 인공신경망을 사용하여 예측 성과 및 주요 변수를 비교·분석하였다. 사용된 변수는 출결, 과제, 중간고사, 기말고사를 사용하였고, 분석결과 정확도 예측성과가 서포트 벡터머신이 가장 좋게 나타났으며, 변수 중요도는 기말고사, 중간고사, 과제, 출결 순으로 나타났다. 분석 결과를 기반으로 이러닝 학습 성과를 예측함에 있어, 출결의 효과가 미미함으로 과제나 토론 등을 함께 분석하여 다른 학습 방법 마련을 제시하였다[5].

Kim and Kim(2019)는 전통적인 사회과학 연구에서 분석모형이나 변수들이 제한적이어서 중요한 변수를 간과하는 문제점을 개선하기 위해 머신러닝 기법인 랜덤 포레스트, 나이브 베이즈 분류 모형, 서포트 벡터머신, 인공신경망 모형을 사용하여 고등학생의 사교육 참여를 예측하는 주요 변수들을 제시하고, 모형간의 예측성과를 비교·분석하였다. 사용된 변수는 학생, 가구 대상으로 학교생활, 가정생활, 여가생활, 교육환경, 문화환경 등의 범주로 구성된 250개 변수를 사용하였고, 분석결과 정확도 예측성과에서 랜덤 포레스트가 다른 머신러닝 기법보다 가장 좋게 나타났으며, 사교육 참여를 예측하는 주요 변수는 각 머신러닝별로 다르게 나타났다. 이는 머신러닝 기법에 따라 분석 결과가 달라질 수 있고, 같은 모형이더라도 분석 옵션에 따라 분석 결과가 다르게 나타날 수 있다는 것을 제시하는 것으로 전통적인 사회과학 연구에서 중요한 변수를 간과하는 문제점을 개선할 수 있는 새로운 연구 방법을 제시하였다[6].

Lee et al.(2020)는 학습과정 중 데이터를 중심으로 학업성취에 영향을 주는 요인을 구명하고자 동영상의 학습자 로그 데이터를 기반으로 머신러닝 기법인 k-근접이웃, 서포트 벡터 머신, 인공신경망, 랜덤 포레스트를 사용하여 학업성취에 영향을 주는 주요 변수들을 제시하고, 모형간의 예측성과를 비교·분석하였다. 사용된 변수는 동영상 학습 플레이어 구동과 관련된 행동빈도, 행동시간의 21개 변수를 사용하였고, 분석결과 정확도 예측성과가 서포트 벡터머신이 다른 머신러닝 기법보다 가장 좋게 나타났으며, 변수 중요도는 각 머신러닝 기법 모두에서 동료 학습자가 작성한 코멘트를 확인하는 것으로 나타났다. 분석 결과를 기반으로 양 방향적 동영상 학습이 가능한 학습 환경에서 학습자 상호작용 행동이 학업성취에 영향을 미친다는 결과를 제시하였다[7].

Park and Chung(2020)는 중학생들의 진로 결정을 예측하는데 주요한 변수를 탐색하기 위해 머신러닝 기법인 의사결정나무, 랜덤 포레스트, Adaptive LASSO, 서포트 벡터머신, 그레디언트 부스팅, 인공신경망을 사용하여 진로 결정에 영향을 주는 주요 변수들을 제시하고, 예측성과를 비교·분석하였다. 사용된 변수는 한국청소년정책연구원의 한국아동·청소년패널조사 데이터를 기반으로 학생변수, 보호자변수로 구성된 320개 변수를 사용하였고, 분석결과 정확도 예측성과가 그레디언트 부스팅이 가장 좋게 나타났으며, 변수 중요도는 학생변수 중 우울과 관련된 문항인 장래가 희망적이지 않은 것 같다가 중요도 지수가 가장 높았으며, 여가시간, SNS이용, 창의적 성격의 변수가 새롭게 도출되었다. 분석 결과를 기반으로 가정과 학교에서 중학생의 진로 결정과 관련하여 진로 탐색의 접근 방식을 제시하였다[8].

Jeong et al.(2021)는 기존 재범위험요인을 평가하는 방법인 전통적 통계기법에서 나아가 새로운 통계적 분석 방향을 제시하기 위해 머신러닝 기법인 의사결정나무, 랜덤 포레스트를 사용하여 재범위험요인 변수 중 재범에 영향을 주는 주요 변수들을 제시하고, 예측성과를 비교·분석하였다. 사용된 변수는 성범죄 전자감독 248명을 대상으로 범죄와 관련된 55개 변수를 사용하였고, 분석결과 정확도 예측성과에서 랜덤 포레스트가 가장 좋게 나타났으며, 변수 중요도는 각 머신러닝 기법 모두에서 감독 기간 내 문제행동 여부 변수가 가장 영향력이 있는 것

으로 나타났다. 분석 결과를 기반으로 진화하는 범죄만큼이나 재범위험요인 평가방법의 새로운 통계적 분석 방향을 제시하였다[9].

Lee et al.(2021)는 부동산 시장의 지표가 되는 부동산 지수를 대상으로 머신러닝 예측의 적합성과 활용성을 확인하기 위해 머신러닝 기법인 랜덤 포레스트, XGBoost, LSTM를 사용하여 각 머신러닝 모델 간의 지수별 예측력을 비교·분석하였다. 사용된 변수는 한국감정원에서 제공하는 아파트 매매지수, 전세가격지수, 지가지수, 부동산 심리지수를 사용하였고, 분석결과 부동산 모든 지수에서 정확도 예측성과가 LSTM 모델이 가장 좋게 나타났으며, 부동산 지수의 예측은 지수의 주기특성과 데이터 특성에 따라 머신러닝 모델별 예측 정확도 차이가 있다는 것을 제시하였다[10].

이렇듯 머신러닝 기법을 사용한 연구에서는 전통적 통계기법의 한계점을 개선하고 분석 데이터 특성에 맞는 최적의 머신러닝 기법 및 새로운 통계적 접근 방식을 제시하고 있으나, 사회·경제 상황을 반영한 지표들을 기반으로 한 연구와 최근 주목 받고 있는 머신러닝의 앙상블 기법을 이용한 연구는 미흡한 실정이다.

이에, 본 연구는 통계청 e-나라지표와 한국은행 경제통계시스템에서 제공하는 통계 지표를 활용하여 정부통계지표가 소매업종인 약국 매출액에 영향을 미치는 예측변인을 탐색하고자 머신러닝의 앙상블 기법인 랜덤 포레스트, XGBoost, LightGBM, CatBoost을 적용하여 분석 데이터 특성에 맞는 최적의 머신러닝 기법을 제시하고, 약국 매출액에 영향을 미치는 주요 예측변수가 무엇인지를 탐색 및 제안하고자 한다.

3. 연구방법

본 연구는 통계청 e-나라지표와 한국은행 경제통계시스템에서 제공하는 통계 지표를 활용하여 정부통계지표가 소매업종 중 약국 매출액에 영향을 미치는 예측 변인을 탐색하기 위해 월별 통계 지표인 물가관련 지표 4개, 경기관련 지표 6개, 고용관련 지표 4개, 산업활동 관련 지표 5개, 금리관련 지표 3개, 증권관련 지표 2개, 통화관련 지표 2개, 기타 지표 2개로 총 28개 지표를 독립변수로 선정하였다. 표 1은 분석대상 주요 통계지표를 나타낸 것이다. 종속변수로는 2016년 1월부터 2021년 12월까지 약국의 매출 데이터 86,203건을 사용하였다.

(표 1) 주요 통계지표
(Table 1) Major statistical indicators

구분	변수명	지수명	기간
물가	con_index	소비자물가지수	65.1~21.12
	col_index	생활물가지수	95.1~21.12
	lea_index	주택전세가격지수	03.11~21.12
	sal_index	주택매매가격지수	03.11~21.12
경기	gci_pchange	경기선행지수순환변동치	70.1~21.12
	gci_achange	경기동행지수순환변동치	70.1~21.12
	cos_index	소비자심리지수	08.7~21.12
	esi_index	경제심리지수	03.1~21.12
	sbo_bsi	소상공인체감경기지수	02.1~21.12
	tra_bsi	전통시장체감경기지수	02.1~21.12
고용	ger_index	일반 고용률	00.1~21.12
	gur_index	일반 실업률	00.1~21.12
	yer_index	청년 고용률	00.1~21.12
	yur_index	청년 실업률	00.1~21.12
산업활동	csi_index	소매판매액지수	95.1~21.12
	wri_index	도소매업지수	00.1~21.12
	mur_index	제조업 가동률지수	71.1~21.12
	mei_index	제조업 출하지수	75.1~21.12
	cre_index	개인신용카드사용액	03.1~21.12
금리	mar_index	시장금리(국고채 3년)	97.1~21.12
	cd_index	CD 금리	97.1~21.12
	cal_index	콜금리	97.1~21.12
증권	kos_index	코스피지수	70.1~21.12
	kod_index	코스닥지수	70.1~21.12
통화	cur_index	본원통화	97.1~21.12
	nec_index	M1(협의통화)	97.1~21.12
환율	exc_index	원/달러 환율	00.1~21.12
유가	oil_index	국제유가	86.1~21.12

분석 절차는 데이터 전처리, 데이터 매칭, 데이터 분석, 데이터 평가 단계로 진행하였다. 데이터 전처리 단계에서는 독립변수와 종속변수의 데이터 특성을 분석한 결과 데이터 제공기준 및 기간이 상이하였다. 데이터 제공기준은 독립변수 28개 정부통계지표에서 월별 데이터로 제공하고 있고, 종속변수 약국 매출액은 일별 데이터로 제공하고 있어, 데이터 제공기준을 통일하기 위해 종속변수인 일별 약국 매출액을 월별로 집계 처리하여 월별 데이터로 변환하였다. 더불어 데이터 제공기간은 독립변수와 종속변수 데이터가 공통적으로 존재하는 기간인 2016년 1월부터 2021년 12월까지를 데이터 제공기간으로 통일하였다.

데이터 매칭 단계에서는 28개 독립변수와 종속변수를 결합하기 위해 데이터 특성을 분석한 결과 고유 식별번호는 없고, 공통변수만 존재하기 때문에 데이터 매칭 방법 중 통계적 매칭 방법인 비 모수 매칭을 활용하여 데이

터를 결합하였다[11-13].

데이터 분석 단계에서는 머신러닝을 위해 독립변수와 종속변수의 값 범위를 일정한 수준으로 맞추는 작업인 피처 스케일링을 위해 모든 변수를 정규분포 형태로 만들기 위해서 로그 값으로 변환하였으며, 머신러닝 분야의 앙상블 기법인 랜덤 포레스트, XGBoost, LightGBM, CatBoost을 적용하여 약국 매출액에 영향을 미치는 예측 변인 추정 및 머신러닝 기법 간의 성능을 비교하기 위해 회귀 분석을 실시하였다.

데이터 평가 단계에서는 회귀 성능을 평가하는 지표인 MSE(Mean Squared Error)와 RMSE(Root Mean Squared Error), MAE(Mean Absolute Error)를 사용하여 실제 값과 예측 값의 차이를 회귀 성능 지표별로 도출하여 머신러닝별로 비교·분석하였다.

본 연구는 머신러닝 분석을 위하여 파이션(version 3.9.7)을 이용하였으며, 회귀분석을 위해 머신러닝 앙상블 기법인 랜덤 포레스트, XGBoost, LightGBM, CatBoost 별로 표 2와 같이 분석 모형에 따른 패키지를 사용하였다.

(표 2) 분석 모형 패키지
(Table 2) Analysis model package

구분	패키지
RandomForest	RandomForestRegressor
XGBoost	XGBRegressor
LightGBM	LGBMRegressor
CatBoost	CatBoostRegressor

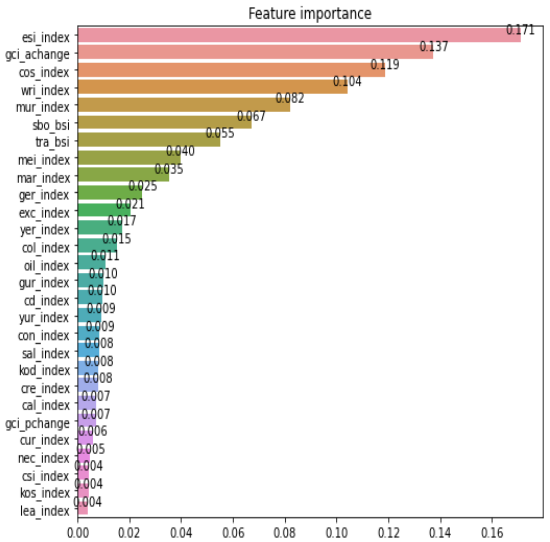
분석 자료는 모형설정 및 예측을 위해 랜덤 함수를 통해 80%는 학습 데이터로, 20%는 시험 데이터로 설정하였다. 보통 시계열 데이터는 과거 데이터를 학습 데이터로 훈련하고 최근 데이터를 시험 데이터로 성능평가를 하지만 고정된 시험 데이터로 회귀 성능을 검증 및 수정 과정을 반복하면 시험 데이터에 과적합되는 문제점이 있다. 이런 문제점을 방지하기 위해 교차검증을 실시한 후, 각 머신러닝별로 회귀분석을 실시하였다. 교차검증은 학습 데이터를 별도의 여러 셋으로 구성된 학습 데이터와 검증 데이터로 셋으로 구성하고 학습과 평가를 수행하여 데이터의 편중을 방지한다.

회귀분석을 실시 한 후 머신러닝별로 독립변수들이 종속변수에 얼마만큼 영향을 주는가를 수치로 나타낸 변수 중요도와 실제값과 예측값의 차이를 절대값으로 변환해 평균한 MAE, 실제값과 예측값의 차이를 제곱해 평균한 MSE, MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 RMSE를 도출하였다[14]. 즉, MAE, MSE, RMSE는 수치가 0으로 가까워지고, 작을수록 높은 예측 성능을 나타낸다. 변수 중요도는 머신러닝 앙상블 기법별 트리성장 방식에 따라 계산하는 방식이 서로 다르다. 랜덤 포레스트는 모형 트리에서 분할이 있어날 때 불순도 감소와 관련된 지수를 계산한 것이고, XGBoost, LightGBM, CatBoost는 모형트리에서 트리성장방식에 따라 데이터 분할 횟수를 계산한 것이다. 즉 랜덤 포레스트는 수치가 0에서 1로 가까워지고, XGBoost, LightGBM, CatBoost는 수치가 클수록 변수 중요도가 높은 것으로 판단한다[15].

4. 연구결과

4.1 랜덤 포레스트 분석결과

랜덤 포레스트를 활용한 모형에서 변수 중요도가 높은 순으로 변수들을 추출한 결과 그림 1과 같이 나타났다.



(그림 1) 랜덤 포레스트 변수 중요도
(Figure 1) RandomForest feature importance

추출한 변수 중요도 상위 5개 지표를 확인하면 경제심리지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수, 제조업 가동률지수 순으로 나타났으며, 그중 경제심리지수 변수가 약국 매출액에 대한 예측력이 가장 큰 것으로 나타났다. 표 3은 랜덤 포레스트 상위 5개 변

수를 나타낸 것이다.

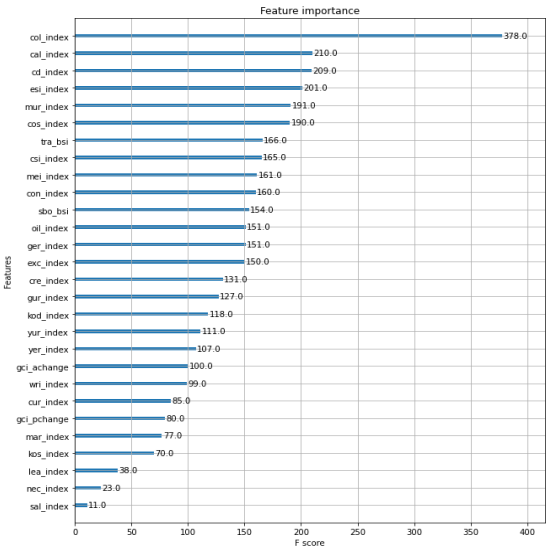
회귀성능 평가지표인 MAE, MSE, RMSE를 살펴보면 MAE 0.1466, MSE 0.0286, RMSE 0.1692로 나타났다.

(표 3) 랜덤 포레스트 상위 5개 변수
(Table 3) RandomForest Top 5 variables

순번	변수명	지수
1	esi_index	경제심리지수
2	gci_achange	경기동행지수순환변동치
3	cos_index	소비자심리지수
4	wri_index	도소매업지수
5	mur_index	제조업 가동률지수

4.2 XGBoost 분석결과

XGBoost를 활용한 모형에서 변수 중요도가 높은 순으로 변수들을 추출한 결과 그림 2와 같이 나타났다.



(그림 2) XGBoost 변수 중요도
(Figure 2) XGBoost feature importance

추출한 변수 중요도 상위 5개 지표를 확인하면 생활물가지수, 콜금리, CD금리, 경제심리지수, 제조업 가동률지수 순으로 나타났으며, 그중 생활물가지수 변수가 약국 매출액에 대한 예측력이 가장 큰 것으로 나타났다. 표 4는 XGBoost 상위 5개 변수를 나타낸 것이다.

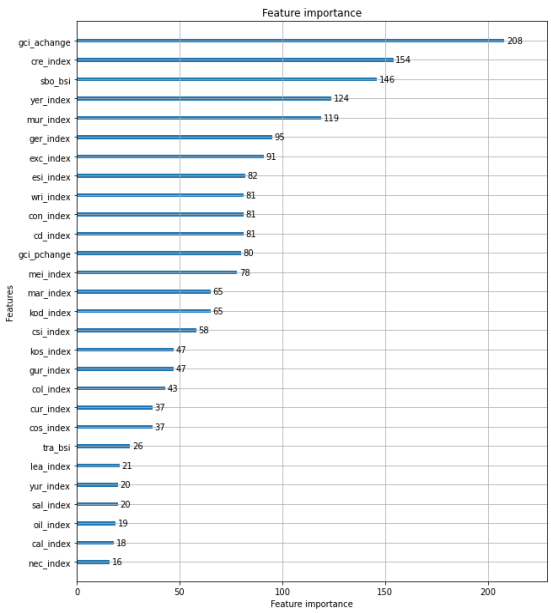
회귀성능 평가지표인 MAE, MSE, RMSE를 살펴보면 MAE 0.1583, MSE 0.0477, RMSE 0.2184로 나타났다.

(표 4) XGBoost 상위 5개 변수
(Table 4) XGBoost Top 5 variables

순번	변수명	지수
1	col_index	생활물가지수
2	cal_index	콜 금리
3	cd_index	CD 금리
4	esi_index	경제심리지수
5	mur_index	제조업 가동률지수

4.3 LightGBM 분석결과

LightGBM를 활용한 모형에서 변수 중요도가 높은 순으로 변수들을 추출한 결과 그림 3과 같이 나타났다.



(그림 3) LightGBM 변수 중요도
(Figure 3) LightGBM feature importance

추출한 변수 중요도 상위 5개 지표를 확인하면 경기동행지수순환변동치, 개인신용카드사용액, 소상공인 체감경기지수, 청년고용률, 제조업 가동률지수 순으로 나타났으며, 그중 경기동행지수순환변동치 변수가 약국 매출액에 대한 예측력이 가장 큰 것으로 나타났다. 표 5는 LightGBM 상위 5개 변수를 나타낸 것이다.

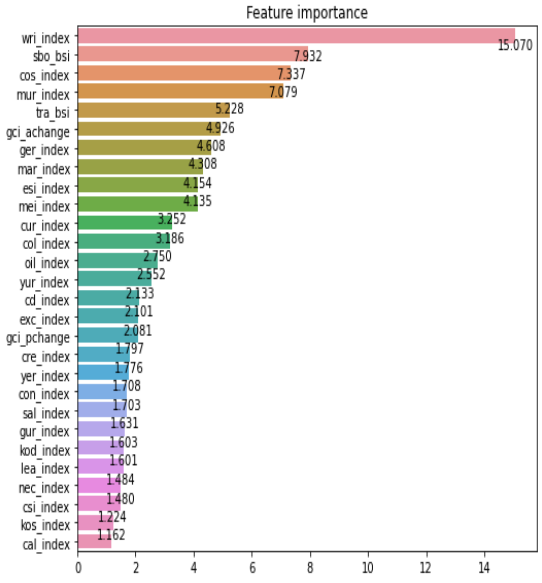
회귀성능 평가지표인 MAE, MSE, RMSE를 살펴보면 MAE 0.2199, MSE 0.0797, RMSE 0.2824로 나타났다.

(표 5) LightGBM 상위 5개 변수
(Table 5) LightGBM Top 5 variables

순번	변수명	지수
1	gci_achange	경기동행지수순환변동치
2	cre_index	개인신용카드사용액
3	sbo_bsi	소상공인 체감경기지수
4	yer_index	청년 고용률
5	mur_index	제조업 가동률지수

4.4 CatBoost 분석결과

CatBoost를 활용한 모형에서 변수 중요도가 높은 순으로 변수들을 추출한 결과 그림 4와 같이 나타났다.



(그림 4) CatBoost 변수 중요도
(Figure 4) CatBoost feature importance

추출한 변수 중요도 상위 5개 지표를 확인하면 도소매업지수, 소상공인 체감경기지수, 소비자심리지수, 제조업 가동률지수, 전통시장 체감경기지수 순으로 나타났으며, 그중 도소매업지수 변수가 약국 매출액에 대한 예측력이 가장 큰 것으로 나타났다. 표 6는 CatBoost 상위 5개 변수를 나타낸 것이다.

회귀성능 평가지표인 MAE, MSE, RMSE를 살펴보면 MAE 0.1772, MSE 0.0666, RMSE 0.258로 나타났다.

(표 6) CatBoost 상위 5개 변수
(Table 6) CatBoost Top 5 variables

순번	변수명	지수
1	wri_index	도소매업지수
2	sbo_bsi	소상공인 체감경기지수
3	cos_index	소비자심리지수
4	mur_index	제조업 가동률지수
5	tra_bsi	전통시장 체감경기지수

4.5 머신러닝 기법 간 비교결과

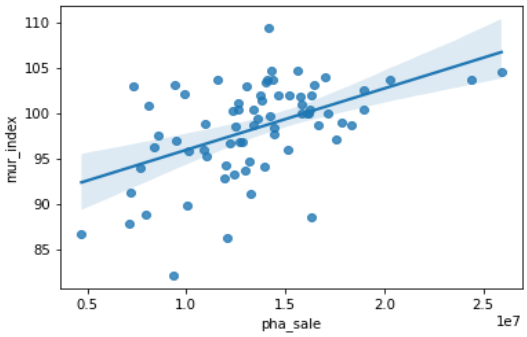
랜덤 포레스트, XGBoost, LightGBM, CatBoost 분석 결과에서와 같이 동일한 자료를 활용하였음에도 머신러닝 기법에 따라 약국 매출액에 영향을 미치는 변수들이 서로 다르게 나타났으며, 회귀분석 결과도 차이가 나타났다. 이는 분석 기법에 따라 분석 데이터, 종속변수 유형, 예측변수 수 등의 차이로 인해 머신러닝의 예측성과가 달라질 수 있기 때문에 당연한 결과일 것이다.

랜덤 포레스트, XGBoost, LightGBM, CatBoost별 서로 다르게 나타난 상위 5개 변수 중 공통적으로 도출된 변수를 확인해 본 결과, 4가지 기법 모두에서 나타난 변수는 제조업가동률지수이며, 3가지 기법에서 나타난 변수는 존재하지 않았으며, 2가지 기법에서 나타난 변수는 경제심리지수, 소상공인체감경기지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수로 나타났다. 표 7는 공통변수 빈도 결과와 그림 5는 제조업가동률지수와 매출액과의 관계를 시각화한 것이다.

(표 7) 공통변수 빈도 결과
(Table 7) Common Variable Frequency Results

빈도	지수
4회	제조업가동률지수
3회	-
2회	경제심리지수, 소상공인체감경기지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수

그림 5에서 제조업가동률지수와 매출액과의 관계를 살펴보면 제조업가동률지수가 낮을 때 매출액이 감소하고 지수가 높을 때 매출액이 증가하는 관계를 보여주고 있다. 즉 제조업가동률지수와 매출액과의 관계는 양의 상관관계를 보여주고 있다.



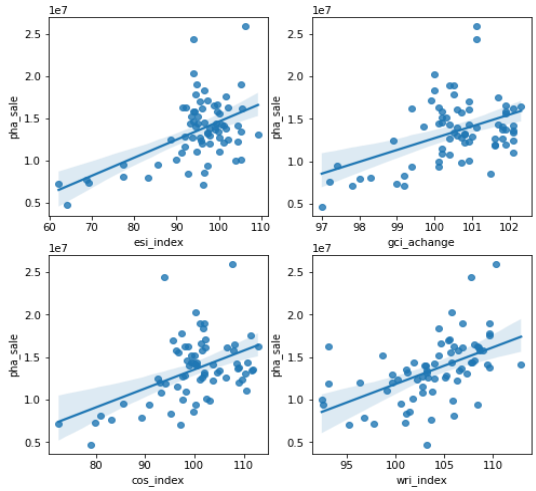
(그림 5) 제조업가동률지수와 매출액 관계
(Figure 5) Relationship between manufacturing operation rate index and sales

각 머신러닝별 회귀 성능평가 결과를 살펴보면 MAE는 랜덤 포레스트 0.1466, XGBoost 0.1583, LightGBM 0.2199, CatBoost 0.1772으로 나타났고, MSE는 랜덤 포레스트 0.0286, XGBoost 0.0477, LightGBM 0.0797, CatBoost 0.0666으로 나타났다. RMSE는 랜덤 포레스트 0.1692, XGBoost 0.2184, LightGBM 0.2824, CatBoost 0.258로 나타나, 랜덤 포레스트가 XGBoost, LightGBM, CatBoost 보다 회귀 성능이 가장 우수하다고 나타났다. 즉 정부통계지표가 약국 매출액에 영향을 미치는지 예측하는데 있어 랜덤 포레스트가 XGBoost, LightGBM, CatBoost보다 최적의 머신러닝 기법이라고 할 수 있다. 표 8은 머신러닝별 회귀분석 결과를 나타낸 것이다.

(표 8) 머신러닝별 회귀분석 결과
(Table 8) Regression analysis results by machine learning

구분	MAE	MSE	RMSE
RandomForest	0.1466	0.0286	0.1692
XGBoost	0.1583	0.0477	0.2184
LightGBM	0.2199	0.0797	0.2824
CatBoost	0.1772	0.0666	0.2580

회귀 성능이 가장 우수한 랜덤 포레스트의 변수 중 그림 1에서 상위 3개 변수인 경제심리지수, 경기동행지수순환변동치, 소비자심리지수는 정부 주요 통계지표인 표 1에서 확인해 본 결과 경기관련 지수에 포함되어 있다는 것을 알 수 있다. 즉 약국 매출액은 정부통계지표 중 물가관련 지표, 고용관련 지표, 금리관련 지표, 증권관련 지표, 통화관련 지표 보다 경기관련 지표에 가장 영향을 많이 받는 것으로 확인되었다. 그림 6은 랜덤 포레스트 상위 4개 변수인 경제심리지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수와 약국 매출액과의 관계를 시각화한 것이다.



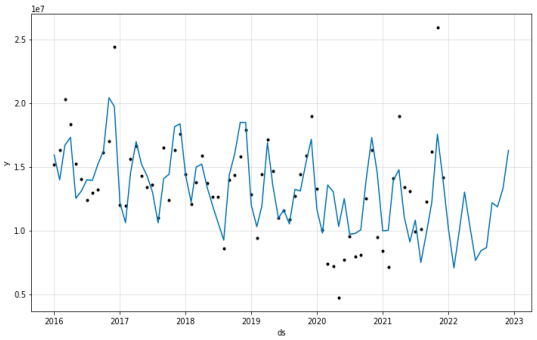
(그림 6) 랜덤 포레스트 상위 4개 변수와 매출액 관계
(Figure 6) Relationship between the Top 4 variables and sales of random forest

그림 6에서 나타난 경제심리지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수와 약국 매출액과의 관계를 살펴보면 경제심리지수는 지수가 낮을 때 매출액이 감소하고 지수가 높을 때 매출액이 증가하는 관계를 보여주고 있다. 즉 경제심리지수와 매출액과의 관계는 양의 상관관계를 보여주고, 경기동행지수순환변동치는 지수가 낮을 때 매출액이 감소하고 지수가 높을 때 매출액이 증가하는 관계를 보여주고 있다. 즉 경기동행지수순환변동치와 매출액과의 관계는 양의 상관관계를 보여준다. 소비자심리지수는 지수가 낮을 때 매출액이 감소하고 지수가 높을 때 매출액이 증가하는 관계를 보여주고 있다. 즉 소비자심리지수와 매출액과의 관계는 양의 상관관계를 보여주고, 도소매업지수는 지수가 낮을 때 매출액이 감소하고 지수가 높을 때 매출액이 증가하는 관계를 보여주고 있다. 즉 도소매업지수와 매출액과의 관계는 양의 상관관계를 보여준다.

결과적으로 랜덤 포레스트의 변수 중요도 상위 5개 변수인 경제심리지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수, 제조업가동률지수는 그림 5와 그림 6에서 나타났듯이 약국 매출액에 양의 상관관계를 보여주고 있다.

더불어 2016년 1월부터 2021년 12월까지 약국 매출 데

이터를 기반으로 2022년 매출액 추이를 분석하였다. 그림 7는 2016부터 2022년까지 약국 매출액 추이를 시각화한 것이다.



(그림 7) 매출액 추이분석
(Figure 7) Sales trend analysis

분석결과 22년도 상반기에는 매출액이 상승 후 하락하겠으나, 하반기에는 매출액이 계속 상승할 것으로 예측되었다. 즉 경기관련 지표는 약국 매출액에 양의 상관관계를 보여주고 있어 이를 기반으로 22년도 하반기부터 경기가 상승할 것으로 판단된다.

5. 결론 및 제언

본 연구는 4차 산업혁명으로 인해 인공지능, 사물인터넷, 빅데이터 등이 정보통신기술과 융합되고 데이터, 네트워크, 인공지능을 기반으로 산업 생태계 조성을 하고 있는 상황에서 각 중앙행정기관에서 관리하고 있는 지표와 민간의 통계자료 등을 토대로 만들어진 정부통계지표가 소매업에 영향을 미치는 예측 변인을 탐색하고 모형간의 예측 결과를 비교·분석하여 최적의 머신러닝 기법을 제시하는 것이다.

이러한 연구목적을 위해 분석 데이터로 통계청 e-나라지표와 한국은행 경제통계시스템에서 제공하고 있는 28개 정부통계지표와 소매업종인 약국 매출액을 사용하였으며, 분석에 앞서 통계적 매칭 방법인 비 모수매칭을 적용하여 28개 정부통계지표와 약국 매출액 데이터를 결합하였다.

데이터 분석은 머신러닝의 앙상블 기법인 랜덤 포레스트, XGBoost, LightGBM, CatBoost를 사용하여 약국 매출액에 영향을 미치는 예측 변인과 회귀성능 평가지표인 MAE, MSE, RMSE를 도출하여 머신러닝별로 비교·분석하였다.

분석결과 랜덤 포레스트, XGBoost, LightGBM, CatBoost 별로 약국 매출액에 영향을 미치는 상위 5개 변수들이 서로 다르게 나타났으며, 4가지 기법 모두에서 나타난 변수는 제조업가동률지수이며, 3가지 기법에서 나타난 변수는 존재하지 않았으며, 2가지 기법에서 나타난 변수는 경제심리지수, 소상공인체감경기지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수로 나타났다. 회귀성능 평가결과를 살펴보면 표 8에서 랜덤 포레스트가 XGBoost, LightGBM, CatBoost보다 MAE, MSE, RMSE에서 수치가 가장 적게 나타나, 정부통계지표가 약국 매출액에 영향을 미치는지 예측하는데 있어 최적의 머신러닝 기법이라고 할 수 있다.

또한 회귀 성능이 가장 우수한 랜덤 포레스트의 변수 중 그림 1에서 상위 3개 변수인 경제심리지수, 경기동행지수순환변동치, 소비자심리지수는 정부 주요 통계지표인 표 1에서 경기관련 지수로 나타나 약국 매출액은 물가관련 지표, 고용관련 지표, 금리관련 지표, 증권관련 지표, 통화관련 지표 보다 경기관련 지표에 가장 영향을 많이 받는 것으로 확인되었다.

결과적으로 랜덤 포레스트의 변수 중요도 상위 5개 변수인 경제심리지수, 경기동행지수순환변동치, 소비자심리지수, 도소매업지수, 제조업가동률지수는 약국 매출액과 양의 상관관계를 보여주고 있고, 약국 매출 데이터를 기반으로 추이를 분석한 결과 22년도 하반기부터 경기가 상승할 것으로 예측되었다.

본 연구는 데이터, 네트워크, 인공지능을 기반으로 산업 생태계 조성을 위해 구축된 물가, 경기, 고용, 금리관련 통계지표들 중 어떤 지표가 약국 매출액에 영향을 미치는지 머신러닝을 활용한 연구로서, 다음과 같은 시사점이 있다.

첫째, 연구를 위한 분석 데이터의 범위를 확대·적용하였다. 전통적인 통계기법은 단일 주제의 데이터를 기반으로 실증분석 및 인과관계를 제시하였으나, 본 연구는 여러 주제의 데이터 세트 즉 물가, 경기, 고용, 금리관련 데이터 세트를 통합하여 머신러닝 분석을 통해 연구 목적에 맞는 새로운 변수를 찾아내는 탐색적 연구를 수행하였다.

둘째, 새로운 통계적 분석 방법을 확대하여 머신러닝 앙상블 기법인 XGBoost, LightGBM, CatBoost를 적용하였다. 머신러닝을 이용한 기존 연구들은 대부분 랜덤 포레스트, KNN, 서포트 벡터 머신, 인공신경망 등을 이용하여 분석하였으나, 본 연구는 기존 앙상블 기법인 랜덤 포

레스트 외에 최근 주목받고 있는 앙상블 기법인 XGBoost, LightGBM, CatBoost를 적용하여 데이터를 비교·분석하고 데이터 특성에 맞는 최적의 머신러닝 기법을 제시하였다.

셋째, 약국 매출액에 가장 영향력이 있는 정부통계지표를 제시하였다. 정부통계지표들은 각 영역별로 관리 및 작성 기준에 의해 산출되고 있으나, 산출된 지표들이 사회·경제적으로 다른 영역에 영향을 미치는지 실증 분석한 연구들은 미흡한 실정이다. 이에 본 연구는 물가, 경기, 고용, 금리관련 28개 정부통계지표와 소매업종인 약국 매출액과의 관계를 머신러닝으로 분석하여, 가장 영향력 있는 정부통계지표를 도출하였다.

본 연구는 정부통계지표가 약국 매출액에 미치는 변수 및 데이터 특성에 맞는 최적의 머신러닝 기법을 제시하였으나, 다음과 같은 한계점이 있다. 첫째, 본 연구는 소매업 중 약국에 한정되어 분석한 것으로 모든 소매업으로 연구결과를 일반화하는데 한계가 있다. 둘째, 약국 매출액에 가장 영향력이 있는 정부통계지표를 제시하였으나, 이는 통계청 e-나라지표와 한국은행 경제통계시스템의 정부통계지표 중 일부분인 28개 지표에 한정되어 분석한 것으로 연구결과를 일반화하는데 한계가 있다. 셋째, 1인 1스마트폰 시대에 정부통계지표 중 모바일을 이용한 지역별 인구 이동량 등에 대한 통계 데이터 부재로 약국 매출액과 인구 이동량에 대한 관계를 분석할 수 없는 한계가 있다.

따라서 이런 시사점 및 한계점들을 보완하여 새로운 통계적 분석 방법으로 머신러닝을 이용한 다각도 연구가 필요할 것으로 사료된다.

참고문헌(Reference)

[1] D. W. Go, M. S. Choi, Y. B. Yoo, & H. W. Jang, “Exploratory correlation analysis between school variables and high school academic performance using random forest techniques,” The Korean Society For The Study Of Education, pp. 149-160, 2018.

[2] K. J. Lee, & Y. S. Kim, “Exploring a Predictive Model for the Employment of Persons with Disabilities Utilizing a Random Forest method: Focusing on the Status and Quality of Employment,” Disability & Employment, 29(3), 145-165, 2019. http://dx.doi.org/10.15707/disem.2019.29.3.006

[3] Y. S. Kim, E. J. Lee, & H. H. Joo, “Exploring a predictive variables for the university entrance through Korean-Type Early Decision Programs,” Journal of educational studies, 50(4), 233-255, 2019. http://dx.doi.org/10.15854/jes.2019.12.50.4.233

[4] C. B. Cho, & H. Lee, “A Predictive Model for Digital Sexual Crime of Adolescents Using Random Forests,” The Journal of Humanities and Social science, 11(6), 3127-3141, 2020. http://dx.doi.org/10.22143/HSS21.11.6.219

[5] H. K. Jho, “Exploration of Predictive Model for Learning Outcomes of Students in the E-learning Environment by Using Machine Learning,” Journal of Learner-Centered Curriculum and Instruction, 18(21), 553-572, 2018. http://dx.doi.org/10.22251/jlcci.2018.18.21.553

[6] Y. S. Kim, & H. H. Kim, “An inquiry for the predictive variables on the demand for the private tutoring utilizing machine learning approaches,” The Journal of economics and finance of education, 28(3), 29-52, 2019. http://dx.doi.org/10.46967/jefe.2019.28.3.29

[7] J. E. Lee, D. S. Kim, & H. I. Jo, “Exploration of Predictive Model for Learning Achievement of Behavior Log Using Machine Learning in Video-based Learning Environment,” Journal of Korean Association of Computer Education, 23(2), 53-64, 2020. https://doi.org/10.32431/kace.2020.23.2.006

[8] S. Y. Park, & H. W. Chung, “Exploring Variables Affecting Career Decision of Middle School Students : An Application of Machine Learning Approaches,” Asian Journal of Education, 21(3), 727-753, 2020. http://dx.doi.org/10.15753/aje.2020.09.21.3.727

[9] Y. C. Jeong, H. Y. Ryu, S. J. Lee, D. J. Seo, & C. G. Park, “Identification recidivism risk factors study based on machine learning : Using decision tree analysis and random forest algorithm,” Korean Police Studies Review, 20(1), 323-350, 2021. https://doi.org/10.38084/2021.20.1.14

[10] J. M. Lee, S. H. Park, S. H. Cho, & J. H. Kim, “Comparison of Models to Forecast Real Estates Index Introducing Machine Learning,” Journal of the Architectural Institute of Korea, 37(1), 191-199, 2021.

https://doi.org/10.5659/JAIK.2021.37.1.191

[11] M. A. Oh, “On the Need for Data Linkage in the Health and Welfare Sectors,” Health and welfare policy forum, 9, pp.17-28, 2015.

[12] Y. C. Jeong, W. T. Lee, H. Jeong, Y. H. Kim, S. S. Yoo, B. Y. Jeong, Y. S. Oh, M. G. Park, H. Y. Kwon, & H. N. Oh, “A Study on the Innovation Plan of the ICT Statistical Production System in Response to Changes in the Survey Environment(Ⅱ) General Report,” Korea Information Society Development Institute, pp.1-237, 2017.

[13] K. M. An, “Developing a Prediction Model for Firm Innovation and Performance Using Statistical Matching and Machine Learning Ensemble Techniques,” Doctoral dissertation, Dongguk University, Seoul, 2021.

[14] C. M. Kwon, “Python Machine Learning Complete Guide,” Wikibook, 2020.

[15] G. S. Lee, “Exploring the predictive variables of major economic-related statistical indicators on SME sales using machine learning: Focusing on small and medium-sized businesses in print-related shopping,” Journal of Korean Association of Computer Education, 25(3), 79-89, 2022.
http://dx.doi.org/10.32431/kace.2022.25.3.007

◐ 저 자 소 개 ◑

이 광 수(Gwang-Su Lee)
2005년 한국외국어대학교 컴퓨터교육과 졸업(석사)
2011년 성균관대학교 컴퓨터교육과 졸업(박사)
2014년~현재 성균관대학교 컴퓨터교육과 겸임교수
관심분야 : 빅데이터, 인공지능, 정보기술아키텍처, 정보교육, etc.
E-mail : 73gslee@gmail.com