

멀티모달 스택킹 모델을 활용한 112 신고 건수 예측

김서연¹ · 이민정¹ · 이지윤¹ · 황하은¹ · 장광호² · 김희두² · 김종윤² · 김성범^{1*}

¹고려대학교 산업경영공학과 / ²경찰대학교 치안정책연구소

Early Prediction of 112 Emergency Calls Using Multimodal Stacking Model

Seo-Yeon Kim¹ · Minjung Lee¹ · Jiyeon Lee¹ · Ha Eun Hwang¹ · Gwang Ho Jang² · Hee Dou Kim² · Jong Yoon Kim² · Seoung Bum Kim¹

¹Department of Industrial and Management Engineering, Korea University

²Police Science Institute, Korea National Police University

Maintaining a safe society is a fundamental value of the nation, and is the primary purpose of policing. Recently, a new approach of using artificial intelligence for policing, known as smart policing, has been proposed. While various studies on smart policing have been discussed, most studies only focus on predicting the hotspot regions or the number of major crimes. The Republic of Korea has a policing system called the 112 system, and the need to incorporate smart policing to the 112 system is increasing gradually. However, the current 112 system is operating less efficiently by deploying the same number of police forces regardless of the number of expected 112 emergency calls. Moreover, studies related to the number of 112 emergency calls are not only insufficient, but also have limitations in that the number of emergency calls in the past has not been utilized. In this study, we propose a multimodal stacking model (MSM) that can predict 112 emergency calls by reflecting the temporal characteristics of the policing data and facilitate interpretation of the important variables. A deep learning-based recurrent neural network and a gradient boosting model are used together in MSM. The proposed MSM yielded excellent predictive performance in experiments using various kinds of regional characteristic data and 112 emergency call data.

Keywords: 112 Emergency Call, Multimodal Learning, Stacking Model, Machine Learning, Deep Learning

1. 서론

모든 인간은 위협이나 위협으로부터 자신을 보호하고 불안을 회피하려는 '안전'에 대한 욕구를 필수적으로 갖고 살아간다. 이러한 기본적인 안전 욕구를 충족하기 위하여 치안의 중요성이 날로 증대되고 있다. 우리나라 경찰은 1957년부터 국민의 안전과 치안에 집중하고자 범죄신고 즉응 체제인 112 시스템을 구축하여 운영하고 있다. 112 시스템은 경찰관이 현장에 출동할 필요가 있는 범죄 발생 시 전화번호 112를 통해 신고 접수하고, 이후 최단시간 내에 경찰력이 현장에 투입되도록

설계되어 있다. 범죄 발생 확률이 높은 시간대와 장소에 경찰력을 미리 투입해 놓는다면 범죄 예방에 매우 효과적인 것이다. 하지만, 현재 112 시스템은 주로 과거 경험에 의존해 경찰력 배치를 결정하고 있다. 신고 건수는 지역과 날씨, 환경 등에 따라 달라지기 때문에 경험에 의한 경찰력 배치는 객관성이 떨어지며 배치 업무 담당자들에 따라 배치 결과가 달라질 수 있는 한계점이 있다. 따라서, 보다 일관적이고 효율적인 경찰력 활용을 위해서는 112 신고 건수에 대한 정확한 예측과 이에 영향을 미치는 요인 파악이 필요한 상황이다.

현 112 시스템의 한계점을 보완하고자 치안문제 의사결정

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2018-0-00440, 위협 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발).

* 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로 145, 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888,

E-mail : sbkim1@korea.ac.kr

2021년 5월 10일 접수; 2021년 6월 28일 수정본 접수; 2021년 7월 28일 게재 확정.

에 인공지능을 활용하는 ‘스마트 치안’에 대한 관심이 높아지고 있으며, 다양한 연구들이 수행되고 있다. Kim and Jung(2021)은 지역을 세분화하여 범죄 발생 고위험 지역 여부를 격자 레벨로 예측하였다. 예측 방법론으로는 랜덤포레스트(random forest), 의사결정나무(decision tree), 서포트 벡터 머신(support vector machine), K 근접 이웃(k-nearest neighbor) 등을 사용하였다. 또한, 지역별 기상 및 인구 통계량 등의 지역 특성 데이터를 기반으로 해당 지역의 절도, 폭력, 강도, 강간, 살인과 같은 주요 범죄 발생 건수를 예측하는 연구들이 수행되었다(Heo *et al.*, 2018; Kang *et al.*, 2017). 이 외에도 Joo(2012)는 주요 범죄에 대한 발생 건수를 시계열 예측 모형인 autoregressive integrated moving average(ARIMA) 모형과 지수 평활법(exponential smoothing) 모형을 활용하여 예측하였다. 하지만, 위의 언급된 연구들은 주로 범죄가 발생한 지역의 위험도를 예측하거나 특정 범죄 발생 건수만을 예측하는 연구이기에 전체 지역과 포괄적인 범죄 항목을 대상으로 하는 치안 정책 수립에 직접적으로 활용되기 어렵다는 한계점을 갖고 있다.

112 신고 건수 예측은 특정 범죄 발생 건수 예측에 비해 충분한 연구가 진행되지 않았으며 소수의 연구들이 전통적인 예측모형을 사용하여 수행되었다. Choi(2018)은 음이항 회귀분석과 신경망 모델에 기상, 인구, 면적, 외국인 수, 주거 유형 등의 지역 특성 데이터를 활용하여 112 신고 건수를 예측하였다. Kim and Lee(2020)은 유동인구 데이터와 순찰 지점 및 경찰 인원의 수, 온도, 공휴일, 계절 등 다양한 지역 특성 데이터를 활용하였다. 랜덤포레스트 모델을 활용하여 예측된 대구광역시 내 법정동별 112 신고 건수로 지역별 신고사건 발생 위험도를 산출하고 신규 순찰 지점을 추천하였다. 하지만, 해당 연구는 시간 흐름에 따른 112 신고 건수의 변화는 반영하지 않고 각 법정동별로 발생할 신고 건수를 단순 예측하였다. 또한, 해

외 연구 사례 중 하나인 Chohlas-Wood *et al.*(2015)은 랜덤포레스트를 사용하여 뉴욕시 911 신고 건수를 예측하였다. 해당 연구들은 모두 기상, 유동인구 등의 지역 특성 데이터만을 활용하여 신고 건수 예측을 진행하였다는 한계점을 갖고 있다.

112 신고 건수 데이터는 기존 연구들에서 활용되던 주요 범죄 건수 데이터와는 다른 특성을 보인다. 주요 범죄의 경우 우발적으로 일어나는 경우가 많아 특정 패턴을 파악하여 분석하기에 어려움이 존재한다(Park *et al.*, 2017). 하지만 112 신고의 경우, 가까운 시점의 정보가 미래 예측에 더 많은 영향을 미치는 시계열적 패턴을 가지고 있다. <Figure 1>은 2014년 1월부터 2019년 5월까지 서울특별시 내 관할관서들의 일자 별 평균 112 신고 건수를 보여주고 있다(데이터 보안 정책에 따라 y축에 정확한 신고 건수는 기입하지 못했음). 신고 건수는 매년 여름에 증가하며, 겨울에는 감소하는 뚜렷한 계절적 주기성 추세를 보이고 있음을 알 수 있다. 따라서 신고 건수 예측에 신고 건수 데이터가 가진 고유한 시계열적 특성들을 포괄적으로 반영하는 모델을 활용한다면 보다 정확한 예측이 가능할 것이다.

본 논문은 순환 신경망(recurrent neural network) 모델과 그라디언트 부스팅(gradient boosting) 모델을 함께 활용한 멀티모달 스택킹(multimodal stacking) 구조의 multimodal stacking model(MSM)을 제안한다. 112 신고 건수 데이터의 시계열적 패턴을 반영하고자 순환 신경망 모델을 활용하였으며, 신고 발생 지역의 특성을 112 신고 건수 예측에 효과적으로 반영하기 위하여 그라디언트 부스팅 모델을 사용하였다. MSM은 딥러닝 기법을 112 신고 건수 예측에 적용한 첫 사례이며, 치안에 영향을 미치는 것으로 밝혀진 다양한 지역 특성 데이터를 함께 활용하고 이에 대한 해석까지 제공할 수 있다는 특징을 가지고 있다. 본 연구에서는 MSM을 통해 1일 후부터 7일 후, 14일 후, 30일 후까지의 예측 결과를 산출하였다.

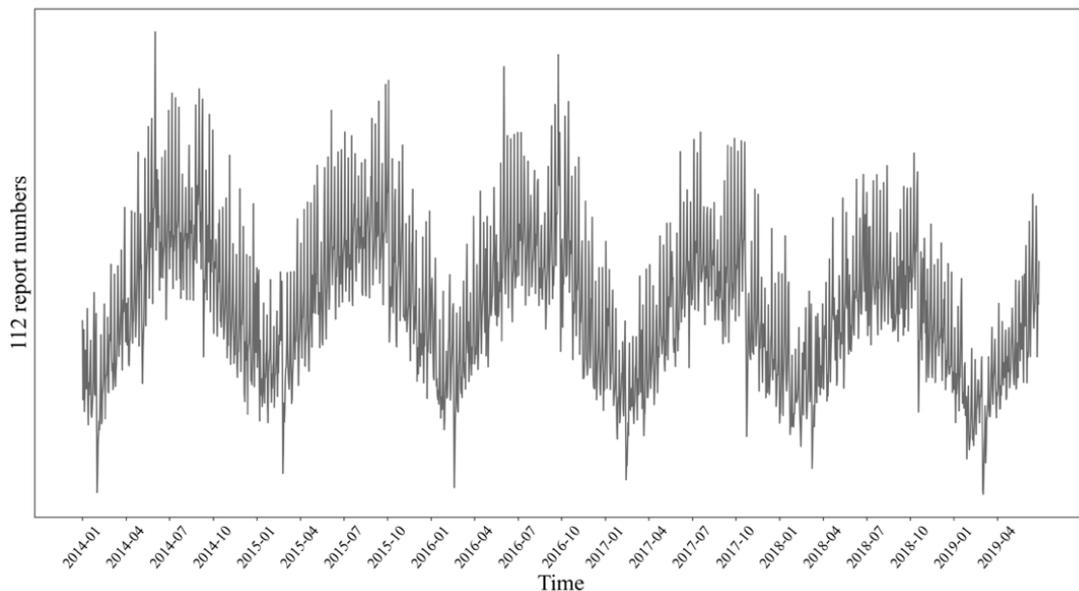


Figure 1. The Trend of 112 Emergency Calls in Seoul, Republic of Korea

본 논문의 구성은 다음과 같다. 제2장에서는 제안 방법론에서 활용되는 관련 모델을 소개하며, 제3장에서는 분석에 활용한 데이터와 제안 방법론에 대하여 자세히 설명하였다. 제4장에서는 제안 모델의 112 신고 건수 예측 실험 결과를 비교한다. 제5장에서는 본 연구의 결론 및 기대 효과에 대하여 다루도록 한다.

2. 관련 모델

2.1 게이트 순환 유닛 모델

게이트 순환 유닛(gated recurrent unit, GRU) 모델은 기존에 순환 신경망 모델의 장기 의존성 문제를 해결하기 위하여 제안된 장기-단기 기억 신경망(long short-term memory model, LSTM)의 구조를 단순화시킨 모델이다(Cho *et al.*, 2014). GRU 모델의 구체적인 구조는 <Figure 2>와 같다.

GRU 모델은 리셋 게이트(reset gate)와 업데이트 게이트(update gate)의 두 가지 게이트로 이루어져 있으며, 순차적인 연산 과정을 통하여 과거 정보를 활용한 연산을 진행한다. 리

셋 게이트에서는 과거 정보와 현재 시점의 정보를 기반으로 기억할 확률($r^{(t)}$)을 계산하며 이는 식 (1)로 표현된다. 업데이트 게이트에서 과거 정보와 현재 시점의 정보를 기반으로 다음 결과값 계산에 활용될 확률($z^{(t)}$)을 식 (2)와 같이 결정한다. 리셋 게이트의 확률 값($r^{(t)}$)과 과거 정보($h^{(t-1)}$) 및 현재 시점의 정보($x^{(t)}$)만으로 현재 시점의 은닉층 값의 후보($\tilde{h}^{(t)}$)를 미리 계산한 후, 이를 과거 정보와 결합하여 현재 시점의 최종 은닉층 값($h^{(t)}$)을 구하며, 관련 수식은 식 (3)과 식 (4)로 표현된다.

$$r^{(t)} = \sigma(W_r h^{(t-1)} + U_r x^{(t)}) \quad (1)$$

$$z^{(t)} = \sigma(W_z h^{(t-1)} + U_z x^{(t)}) \quad (2)$$

$$\tilde{h}^{(t)} = \tau(W_h^{(t-1)} * r^{(t)} + U_h x^{(t)}) \quad (3)$$

$$h^{(t)} = (1 - z^{(t)}) * h^{(t-1)} + z^{(t)} * \tilde{h}^{(t)} \quad (4)$$

2.2 그래디언트 부스팅 모델

그래디언트 부스팅은 정확한 예측을 수행하기 위하여 의사결정나무 모델의 순차적인 업데이트를 진행하는 앙상블(ensemble) 모델이다(Friedman, 2001). 그래디언트 부스팅의

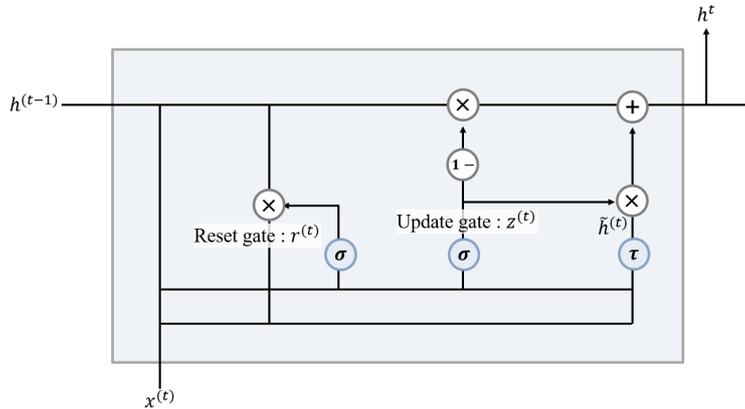


Figure 2. Overview of a Gated Recurrent Unit Cell

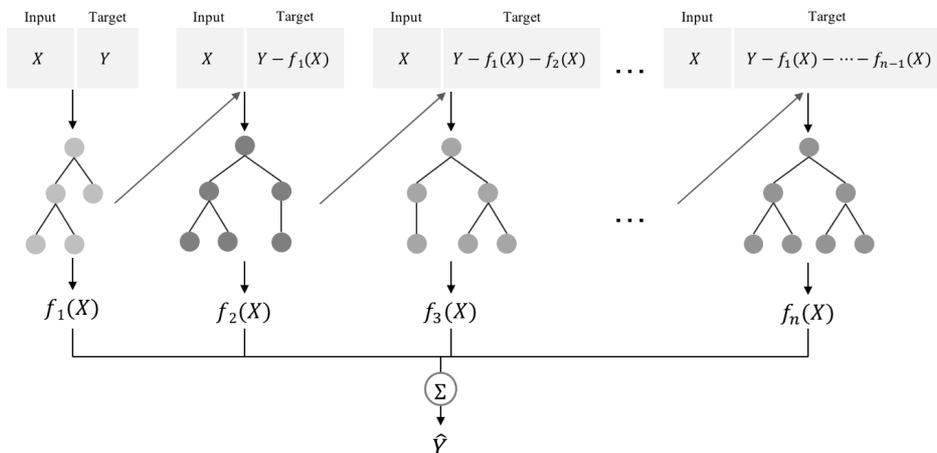


Figure 3. Overview of a Gradient Boosting Model

구체적인 모델 구조는 <Figure 3>과 같다. 처음 생성된 의사결정나무를 제외하고 두 번째 생성된 의사결정나무부터는 이전 모델의 예측 값과 실제 값 사이의 차이인 잔차를 예측하도록 학습한다. 이와 같은 순차적 학습 과정을 통하여 이전 의사결정나무 모델에서 부족하였던 예측 성능을 보완한다. 해당 과정을 반복하여 여러 개의 의사결정나무 모델을 구축한 후에, 최종 예측 값(\hat{Y})을 산출하기 위하여 각 의사결정나무의 결과 값의 가중 합을 계산한다. 최종 예측 값에 대한 수식은 식 (5)부터 식 (7)로 정의된다. 식 (6)은 가중 합 연산의 가중치(γ_i)를 구하는 수식으로, 이전까지 생성된 모델들의 가중 합 결과($F_m(X)$)인 식 (7)을 활용하여 잔차를 최소화하도록 계산한다.

$$\hat{Y} = \sum_{i=1}^n \gamma_i f_i(X) \quad (5)$$

$$\text{where } \gamma_i = \operatorname{argmin} L(Y, F_{m-1} + \gamma f_m(X)) \quad (6)$$

$$F_m(X) = F_{m-1}(X) + \gamma_m f_m(X) \quad (7)$$

그라디언트 부스팅은 다변량 데이터를 사용한 수치 예측 문제에서 우수한 성능을 보인다고 알려져 있을 뿐만 아니라 변수 중요도까지 산출된다는 장점이 있다. 각 의사결정나무 내에서 특정 변수가 사용된 경우와 사용되지 않은 경우를 비교하여 정보 획득 정도(information gain)를 계산하고, 이를 활용하여 특정 변수가 주어진 데이터를 얼마나 잘 예측하는지 수치화 할 수 있다. 그라디언트 부스팅 모델 내 모든 의사결정나무에서 특정 변수에 대한 정보 획득 정도를 산출한 이후, 평균을 취하여 변수 중요도를 결정한다. 이와 같은 방법으로 주요 변수를 파악함으로써 신고 건수 발생 예측에 기여한 주요 변수를 확인할 수 있고 이는 실제 치안 문제 예방에 유용하게 활용될 수 있다.

3. 실험 방법

3.1 데이터 수집 및 전처리

본 연구에서는 112 신고 건수, 기상, 공휴일, 집회 건수, 지역별 보육 시설 및 5대 범죄 발생 통계량 등의 자치구 통계 데이터를 사용하였다. 112 신고 건수의 경우 경찰대학 치안정책연구소에서 2014년 1월 1일부터 2019년 5월 31일까지의 서울특별시 데이터를 제공받았으며, 접수 일자, 관할관서, 당일 신고 건수의 변수를 활용하였다. 112 신고 건수를 제외한 나머지 데이터들은 지역의 특성을 반영하고 있기 때문에 신고 건수 예측 결과 설명에 유의한 영향을 줄 것으로 판단하였다. 아울러 기상 정보가 범죄 발생에 영향을 미친다는 기존 연구에 근거하여, 112 신고 건수 데이터 기간과 동일하게 기상 데이터를 수집하여 예측 모델을 구축하는데 활용하였다(Kim *et al.*, 2012; Ranson, 2014). 서울특별시 기상 데이터는 기상자료 개

방 포털(<https://data.kma.go.kr/cmmn/main.do>)에서 제공하는 기온, 강수량, 풍속, 풍향, 습도, 적설, 전운량 등의 자료를 크롤링하여 수집하였다.

기존 연구에 의하면, 해당 지역의 인구 밀집도, 상주 인구 및 유동 인구 비율 등의 변화는 범죄 발생과 밀접하게 관련되어 있으며(Lee and Cho, 2006), 법정 공휴일 또는 주말에 범죄 건수가 증가한다는 특징이 있다(Roh, 2017; Uittenbogaard and Ceccato, 2012). 이에 근거하여 본 연구에서는 신고 건수 예측에 유동 인구 정보를 함께 반영하기 위하여 공휴일과 집회 데이터를 활용하였다. 집회 데이터는 서울 지방 경찰청 사이트(<https://www.smpa.go.kr/user/nd54882.do>)에서 수집하였으며, 각 집회 날짜 별 예상 참가 인원과 집회 시간대, 관할관서 변수로 구성되어 있다. 이는 특히 집회가 빈번한 관서의 신고 건수 예측에 유용하게 활용될 수 있을 것으로 판단하였다.

아울러 범죄 발생은 지역의 특성에 따라 영향을 받는다고 알려져 있기 때문에(Kubrin and Weitzer, 2003), 본 연구에서는 서울특별시 각 자치구의 다양한 통계 데이터를 활용하였다. 서울 열린데이터광장(<https://data.seoul.go.kr/>)에서 제공하는 여러 데이터 중 공공체육시설, 공원, 노인 여가시설 및 복지시설, 5대 범죄 발생 통계량, 보육시설, 빈집 현황, 지역용도 등 22개의 데이터를 직접 크롤링하여 수집하였다. 이후 관할관서와 지역구를 매칭시키는 방식으로 112 신고 건수 데이터와 병합하여 최종 데이터를 구축하였다. 모델 학습에 사용된 최종 데이터의 변수 명과 설명을 <Table 1>로 요약하였다.

3.2 제안 방법론

본 연구는 여러 기간의 112 신고 건수를 예측하기 위하여 주 모듈인 순환 신경망 모델과 보조 모듈인 그라디언트 부스팅 모델을 결합하는 멀티모달 스택킹 방법론을 제안하였다. 112 신고 건수 데이터의 경우 시계열적 특성을 반영하도록 딥러닝 계열의 순환 신경망 모델을 사용하였으며, 지역 특성 데이터인 기상, 공휴일, 집회 건수, 자치구 통계 데이터를 추가적으로 활용하고자 다변량 데이터의 특성에 적합하면서 중요 변수에 대한 추출이 가능한 그라디언트 부스팅 모델을 선정하였다. 두 모델에서 얻은 112 신고 건수 예측 값은 신경망 모델을 통해 최종 예측 값으로 산출된다. 제안 모델인 MSM을 사용한 예측 과정을 <Figure 4>에 요약하였다.

본 제안 모델 MSM의 주 모듈인 순환 신경망은 딥러닝 모델로 112 신고 건수의 시계열적 패턴을 효과적으로 반영하여 미래 신고 값을 예측하고자 사용하였다. 순환 신경망 모델은 입력과 출력을 순차적으로 처리하면서 이전 시점의 요약 값을 다음 예측 값에 함께 활용하여 시계열 데이터 예측을 수행한다. 특히, 본 연구에서는 다양한 순환 신경망 종류 가운데 GRU 모델을 사용하였다. 반복 실험을 통하여 검증 단계에서 우수한 성능을 보인 값으로 선정된 하이퍼파라미터는 30개의 노드를 가진 은닉층 1개와 양방향성을 추가하는 것으로, 해당 값으

로 신고 건수 데이터를 더 정확하게 파악하도록 학습하였다. 최적화 방식(optimizer)으로는 학습률이 0.0001인 adaptive moment estimation(Kingma and Ba, 2014)을 사용하였다. 과거 60

일의 데이터를 학습한 후 특정 예측 기간의 신고 건수를 예측하도록 순환 신경망 모델의 구조를 생성하였고 예측 기간을 7일로 설정한 예시는 <Figure 5>와 같다.

Table 1. Variables Description

Data Name	Variable Name	Description
Time series police data	Registration Date	Date of registration for each number of emergency calls
	Police Station	Name of the patrol division or police box
	District	The district which the police station is included
	Number of Emergency Calls	Number of emergency calls for the registration date
	Year	Year of the registration date
	Month	Month of the registration date
Weather forecast data	Temperature	Average temperature
	Precipitation	Average amount of precipitation
	Wind Speed	Average wind speed
	Wind Direction	Average wind direction
	Humidity	Average humidity
	Amount of Snowfall	Average amount of snow
National holiday data	Total Cloud	Average amount of cloud
	National Holiday	Whether the registration date is a national holiday
Number of rallies	Weekend	Whether the registration date is a weekend
	Number of Rallies	The number of rallies that happened during the registration date
Statistics for Seoul districts	Area Usage	Percentage of each area usage
	Five Major Crimes	Number of five major crimes(Murder, Robbery, Forced harassment, Theft, Violence)
	Public Sport Facilities	Number of public sports facilities for each district
	Number of Apartment Houses	Number of apartment houses with different sizes
	Park	Number of different types of parks
	Leisure Welfare Facility for Seniors	Number of different leisure welfare facilities for seniors
	Residential Welfare Facility for Seniors	Number of different residential welfare facilities for seniors
	Number of Seniors Living Alone	Number of seniors living alone, classified according to income level
	Childcare Facility	Number of childcare facilities
	Empty Houses	Number of empty houses, classified according to the type of housing
	Death Statistics	Number of deaths, classified according to different causes
	Water Supply Status	Statistics of water supply for each district
	Market	Number of markets, classified according to the market type
	Overall Budget	Overall budget for each district
	Medical Institution	Number of medical institutions
	Public High Schools	Specific statistics related to public high schools
	Suicide Rate	The suicide rate per 100,000 people for each district
	People with Disabilities	Number of people with disabilities, classified according to different disability types
	Low-income Single Parent Family	Number of single parent families with low income
	Registered Population	Number of registered populations
Parking Lot Ratio	Ratio of parking lots	
Public Middle Schools	Specific numbers related to public middle schools	

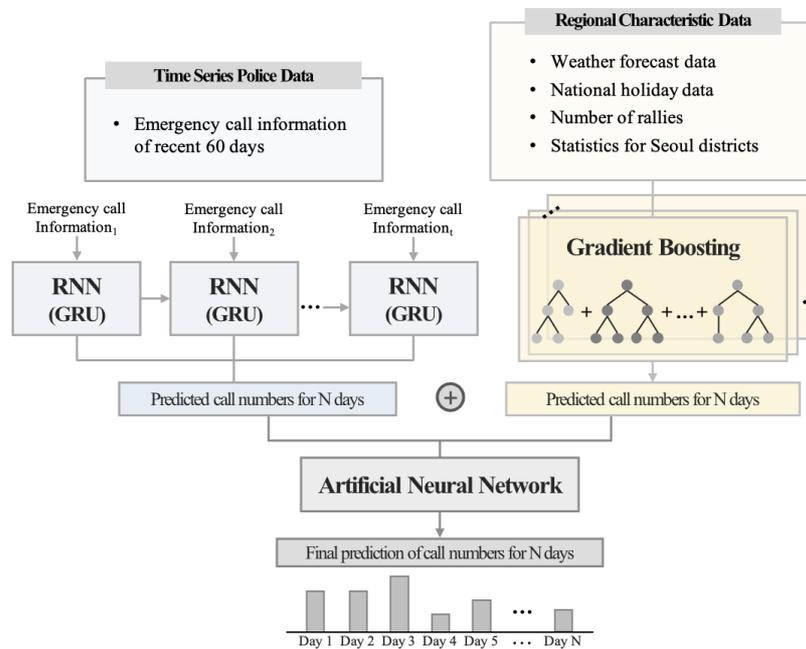


Figure 4. Overview of the Proposed Multimodal Stacking Model(MSM)

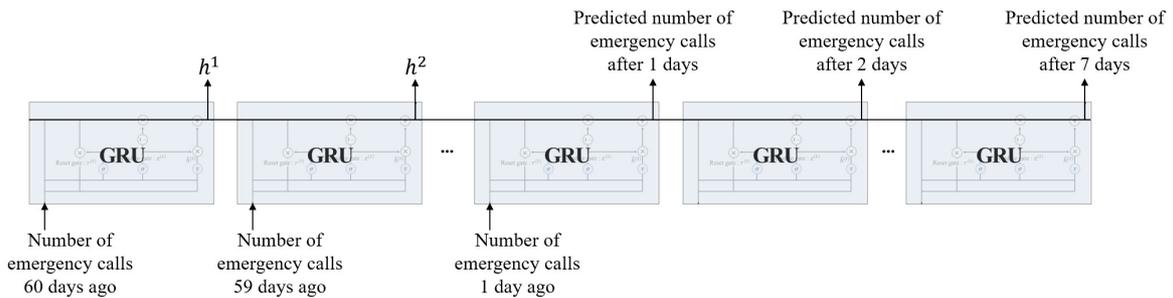


Figure 5. Overview of the Full GRU Model for the Proposed MSM

MSM의 보조 모듈로는 지역 특성 데이터의 추가적인 활용이 가능하며 변수 중요도를 추출할 수 있는 그래디언트 부스팅 모델을 선정하였다. 이를 통하여 각 지역별 특성을 효과적으로 반영하는 다변량 데이터 예측을 수행한다. 그래디언트 부스팅 모델 역시 과거 60일의 데이터를 학습한 후 특정 예측 기간의 미래 신고 건수를 예측하도록 설계하였다. 실험에서 활용한 하이퍼파라미터는 그리드 서치(grid search) 방식을 이용하여 의사결정나무 깊이 4, 학습률 0.1, 의사결정나무의 개수 75개로 결정하였다.

본 논문에서 제안하는 MSM은 주 모듈인 GRU 모델과 보조 모듈인 그래디언트 부스팅 예측 결과를 모두 활용할 수 있는 스택킹 구조를 활용하였다(<Figure 4>). 스택킹이란 머신러닝의 앙상블 기법 중 하나로, 이전 모델에서 나온 예측 값들을 훈련용 데이터로 삼아 추가적인 모델을 학습시켜 최종 예측 값을 얻는 과정이다. 본 연구에서는 두 모델의 예측 결과를 인공 신경망 모델로 학습하여 추후 미래의 최종 112 신고 건수 예측 값을 도출하였다. 이 때 예측 기간은 사용자의 필요에 따라 적

접 지정할 수 있다. 스택킹에 활용한 인공 신경망 모델은 6개의 노드로 구성된 1개의 은닉층을 가지고 있다. 이와 같이 스택킹 과정을 추가적으로 활용함으로써 두 모델의 결과값을 체계적으로 결합하여 더욱 정확하게 예측할 수 있다는 장점을 갖고 있다.

4. 실험 결과

실험은 2014년 1월 1일부터 2018년 6월 30일까지를 훈련용 데이터, 2018년 7월 1일부터 2019년 5월 31일까지를 평가용 데이터로 분할한 후 진행하였다. 112 신고 건수 데이터는 시간의 흐름이 중요하기 때문에 시간 순서대로 훈련용과 평가용 데이터로 분할하였다. 실험 결과는 평가용 데이터의 결정 계수(R-squared, R^2)와 평균 제곱근 편차(root mean squared error, RMSE)를 사용하여 정리하였다. 결정 계수는 데이터의 총 변동성 중 모델로 설명 가능한 변동성의 비율이며, 0과 1 사이의

값을 갖는다. 더 큰 값을 가질수록 모델이 데이터를 많이 설명함을 의미하기 때문에 더 좋은 성능을 보인다고 평가할 수 있다. 결정 계수의 구체적인 수식은 식 (8)로 표현된다. 평균 제곱근 편차는 실제 값과 예측 값 사이의 차이를 계산하는 평가 지표로, 작은 값을 가질수록 예측 성능이 높음을 의미한다. 이에 대한 수식은 식 (9)이며, 식 (8)과 식 (9)에서 등장하는 Y_i 는 실제 신고 건수 값, (\hat{Y}_i) 는 모델을 통하여 얻은 예측 신고 건수 값, \bar{Y} 는 실제 신고 건수 값의 전체 평균을 의미한다.

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (8)$$

$$RMSE(Y_i, \hat{Y}_i) = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (9)$$

실험을 통해 본 논문에서 제안한 예측 모델인 MSM의 성능을 다른 모델과 비교 분석하고 모델 구조의 타당성을 입증하였다. 앞서 설명한대로 MSM은 112 신고 건수 데이터를 요약하는 주 모듈에서는 시계열 예측 모델을, 지역 특성 데이터를 요약하기 위한 보조 모듈에서는 다변량 예측 모델을 사용하며, 이는 멀티모달 스택킹 구조로 구성된다. 구체적으로 주 모듈로는 GRU 모델을, 보조 모듈로는 그래디언트 부스팅 모델을 활용하였다. 따라서, 세부 모듈을 구성하는 모델들의 사용 타당성을 입증하고자 시계열 예측 모델(4.1절), 다변량 예측 모델(4.2절), 멀티모달 스택킹 모델(4.3절)의 세 가지 비교 실험에 대한 결과를 분석하였다. 또한, 예측 기간은 1일 후부터 7일 후, 14일 후, 30일 후 등 다양한 기간의 성능 비교를 진행하였다.

4.1 시계열 예측 모델 비교

MSM의 주 모듈인 신고 건수 데이터의 내재된 시계열 추세를 반영하는 모델을 구축하고 성능을 비교하기 위하여 시계열

예측 모델 실험을 진행하였다. 본 실험에서는 대표적인 시계열 예측 모델인 ARIMA, seasonal autoregressive integrated moving average(SARIMA), 지수 평활법, 딥러닝 기반 LSTM 및 GRU 모델을 비교하였다. ARIMA 모델은 과거 시계열 데이터를 기반으로 미래를 예측하는 모델로 과거 자기 자신의 정보를 활용하는 자기회귀(autoregressive) 모델과 과거 오류 정보를 활용하는 이동평균(moving average) 모델을 결합한 모델이다(Box et al., 2015). ARIMA 모델은 시계열 데이터가 비정상성을 가진 경우 차분의 과정을 통하여 데이터의 정상성을 확보하여 예측을 진행한다. SARIMA 모델은 ARIMA 모델에 계절형 자기회귀(seasonal autoregressive) 항과 계절형 이동평균(seasonal moving average) 항을 추가하여 계절성을 반영한 모델이다(Box et al., 2015). 지수 평활법은 시계열 데이터가 계절의 변동성을 포함하지만 변동 정도가 일정한 경우 활용되는 방법론이다(Holt, 2004). 과거 시점의 데이터로 갈수록 점점 작은 가중치를 부여하는 방식으로 전체 데이터를 이용하여 미래를 예측한다. LSTM 모델은 순환 신경망 계열의 모델로 기존 순환 신경망 모델에서 과거 정보를 오래 기억하지 못하고 소실되어버리는 장기 의존성 문제를 해결하기 위하여 제안되었다(Hochreiter and Schmidhuber, 1997). 과거 정보를 효과적으로 기억할 수 있도록 메모리 셀을 추가적으로 활용하여 이전 정보의 특성을 효과적으로 학습할 수 있다. GRU 모델은 2.1절과 3.2절에서 설명한 바와 같이 본 제안 모델의 주 모듈로 사용하였다.

ARIMA, SARIMA, 지수 평활법, LSTM, GRU의 다섯 가지 모델을 비교한 결과를 <Table 2>에 정리하였다. 비교에 활용한 모든 시계열 예측 모델 중 GRU 모델이 두 가지 평가 지표에서 가장 좋은 성능을 보였다. 또한, 예측 기간이 7일 후, 14일 후, 30일 후로 점차 확장됨에 따라 예측 성능이 저하됨을 확인할 수 있다. 전반적으로 순환 신경망 계열의 모델들이 예측 기간이 증가하더라도 꾸준히 좋은 성능을 보이며, 다른 시계

Table 2. Model Performance for ARIMA, SARIMA, Exponential Smoothing, LSTM, and GRU

Model Name	ARIMA		SARIMA		Exponential Smoothing		LSTM		GRU	
	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE
1 Days Later	0.6945	11.3822	0.6986	11.2990	0.7071	10.9683	0.7263	11.4968	0.7538	10.2289
2 Days Later	0.6634	12.0122	0.6818	11.6442	0.6950	11.2238	0.7145	11.6749	0.7376	10.5706
3 Days Later	0.6580	12.1213	0.6789	11.7063	0.6882	11.3682	0.7141	11.6821	0.7367	10.5824
4 Days Later	0.6615	12.0381	0.6753	11.7657	0.6730	11.6959	0.7155	11.6327	0.7403	10.4566
5 Days Later	0.6726	11.8158	0.6767	11.7294	0.6624	11.9326	0.7183	11.5714	0.7460	10.3494
6 Days Later	0.6859	11.5655	0.6843	11.5853	0.6508	12.2034	0.7226	11.5021	0.7534	10.2183
7 Days Later	0.6839	11.6161	0.6867	11.5354	0.6365	12.5204	0.7196	11.5355	0.7495	10.2808
14 Days Later	0.6455	12.6280	0.6543	12.1111	0.5168	14.4201	0.7113	11.6400	0.7355	10.4262
30 Days Later	0.5631	14.2259	0.6011	13.4901	0.3111	23.6860	0.6782	12.1497	0.6981	11.1128
Total Average	0.6587	12.1561	0.6709	11.8741	0.6157	13.3354	0.7134	11.6539	0.7390	10.4696

열 모델들은 먼 미래를 예측하는 데에 어려움을 보이고 있었다. 정확한 예측을 위해서는 약 14일 이내로 예측 기간으로 설정하는 것이 필요함을 보여주는 결과이다.

4.2 다변량 예측 모델 비교

MSM의 보조 모듈에 해당하는 다변량 예측 모델 실험에서는 기상, 집회 건수, 공휴일, 자치구 통계 데이터 등의 지역 특성 데이터를 활용하였다. 이는 지역 특성 데이터에 음이항 회귀분석, 랜덤포레스트 등의 전통적인 머신러닝 모델을 적용한 기존 연구들의 모델 학습 방식을 참고한 것이다(Choi, 2018; Kim and Lee, 2020). 다양한 머신러닝 모델 가운데 주요 변수 선택 기능이 있는 캣부스트(catboost), 랜덤포레스트, 그래디언트 부스팅의 세 가지 모델을 비교하였다. 캣부스트 모델은 학습 단계에서는 관측되지 않았던 범주가 테스트 데이터로 모델에 입력되었을 때 발생하는 문제를 다루기 위해 순서형 종속 변수 통계량(ordered target statistics)과 순서형 부스팅(ordered boosting) 방법을 사용한다(Prokhorenkova et al., 2017). 기존 그래디언트 부스팅 계열 모델들에서는 범주형 변수를 처리하기 위해 종속 변수 통계량을 범주 값에 치환하는 방식을 사용하였다. 수치형으로 치환된 변수는 범주 별로 모두 동일한 값을 갖게 된다. 하지만, 해당 방식은 종속 변수 통계량이 종속 변수 예측에 사용되어 모델 과적합을 야기한다고 알려져 있다. 이러한 문제를 방지하기 위하여, 캣부스트 모델은 각 데이터에 임의의 순서를 부여하고, 순서에 따라 범주 별 순서형 종속 변수 통계량을 계산하는 방식을 사용하였다. 임의의 순서대로 해당 데이터 이전까지의 종속 변수 정보만으로 종속 변수 통계량을 계산하기 때문에 동일한 범주일지라도 서로 다른 종속 변수 통계량 값을 갖게 된다. 더 나아가, 모델의 잔차를 예측하는 순차적인 업데이트 단계에서 임의의 순서에 따라 학습용 데이터를 추가하는 순서형 부스팅 방식을 활용하여, 모델을 통과한 학습용 데이터와 평가용 데이터의 분포가 서로 달라지

는 한계점을 해결하였다. 랜덤포레스트 모델은 학습 데이터로부터 복원 추출을 진행한 후 다양한 의사결정나무를 구성하여 이를 앙상블 하는 모델이다(Breiman, 2001). 매 단계에서 생성되는 의사결정나무는 학습용 데이터의 부분 집합과 변수들의 랜덤 부분 집합으로 구축되며, 매번 다른 부분 집합이 사용되는 만큼 다양한 특성을 담게 된다. 서로 다른 특징을 가진 의사결정나무 모델을 여러 개 구축한다는 점에서 다양성을 확보하게 되며, 이는 단일 의사결정나무 모델 대비 높은 성능 뿐만 아니라 강건한 결과를 보여준다. 그래디언트 부스팅은 본 제안 모델의 보조 모듈로 사용되었으며 2.2절에서 설명하였다.

총 세 가지 모델에 대하여 비교 실험을 진행한 결과를 <Table 3>에 정리하였다. 두 가지 평가 지표를 기준으로 모든 예측 시점에서 그래디언트 부스팅 모델의 성능이 가장 좋음을 확인할 수 있었다. 또한, 그래디언트 부스팅 모델의 경우 예측 기간이 14일 후, 30일 후로 증가하더라도 일정한 성능을 보이고 있다. 하지만, 결정 계수 평균 값을 기준으로 예측 성능은 0.4194에 그쳤으며 이는 그래디언트 부스팅 모델만으로는 112 신고 건수 예측을 정확히 할 수 없음을 의미한다.

4.3 멀티모달 스택킹 모델 비교

마지막으로, 멀티모달 스택킹 구조에 대한 성능을 확인하고자 주 모듈인 GRU 모델과 다변량 예측 모델(random forest, catboost, gradient boosting)을 결합한 멀티모달 스택킹 모델을 비교하였으며 실험 결과는 <Table 4>에 정리하였다.

GRU 모델과 그래디언트 부스팅 모델을 멀티모달 스택킹 구조로 함께 결합한 MSM이 결정 계수 평균 기준 0.7691의 성능을 보였으며, 비교 모델 대비 가장 우수하였다. 랜덤포레스트 모델을 결합한 조합은 0.7318로 멀티모달 스택킹 모델 중 가장 정확도가 낮았다. 캣부스트 모델을 결합한 경우에는 0.7652의 평균 성능으로 그래디언트 부스팅을 결합한 모델과 성능 차이는 거의 없었으나 다른 예측 모델들에 비해

Table 3. Model Performance for CatBoost, Random Forest, and Gradient Boosting

Model Name	CatBoost		Random Forest		Gradient Boosting	
	R ²	RMSE	R ²	RMSE	R ²	RMSE
1 Days Later	0.1177	19.1872	0.3604	17.3016	0.4120	16.3849
2 Days Later	0.1125	19.2425	0.3709	16.6654	0.4230	16.1882
3 Days Later	0.1192	19.1842	0.3413	17.5548	0.3948	16.2613
4 Days Later	0.1273	19.0421	0.3369	17.5754	0.4178	16.1108
5 Days Later	0.1170	19.1325	0.3331	17.6539	0.4317	16.0463
6 Days Later	0.1268	19.0353	0.3331	17.6701	0.4211	16.1554
7 Days Later	0.1134	19.1818	0.3319	17.7460	0.4188	16.1763
14 Days Later	0.1133	19.2251	0.3481	16.8887	0.4268	15.8310
30 Days Later	0.1024	20.4468	0.2969	17.7959	0.4288	15.5727
Total Average	0.1166	19.2975	0.3392	17.4280	0.4194	16.0808

Table 4. Model performance for GRU + CatBoost, GRU + Random Forest and GRU + Gradient Boosting

Model Name	GRU + Random Forest		GRU + CatBoost		GRU + Gradient Boosting (MSM, proposed)	
	R ²	RMSE	R ²	RMSE	R ²	RMSE
1 Days Later	0.7515	10.2489	0.7810	9.4670	0.7913	9.3165
2 Days Later	0.7363	10.5306	0.7649	9.8196	0.7742	9.7021
3 Days Later	0.7275	10.7036	0.7696	9.7208	0.7739	9.7060
4 Days Later	0.7241	10.8667	0.7722	9.6448	0.7787	9.5817
5 Days Later	0.7214	10.7706	0.7689	8.7012	0.7773	9.6001
6 Days Later	0.7332	10.5711	0.7680	9.7272	0.7755	9.6447
7 Days Later	0.7459	10.3246	0.7760	9.5596	0.7813	9.5213
14 Days Later	0.7436	10.0879	0.7590	9.7670	0.7491	9.9673
30 Days Later	0.7025	10.8561	0.7272	10.3947	0.7204	10.5194
Total Average	0.7318	10.5511	0.7652	9.6447	0.7691	9.7288

매우 느린 연산 속도를 보인다는 단점을 갖고 있다. 따라서, 성능 및 활용도 측면에서 GRU 모델과 그래디언트 부스팅 모델을 결합한 MSM 구조가 가장 타당한 것으로 결론 내릴 수 있다.

전반적인 실험 결과를 통하여 GRU 모델을 단독으로 사용하여도 112 신고 건수 예측 성능이 결정 계수 평균 기준 0.7390을 기록함을 확인하였으며, MSM의 성능 향상에 주된 역할을 담당함을 알 수 있었다. 보조 모듈로 사용한 그래디언트 부스팅 모델은 MSM의 실용성을 향상시키기 위한 접근으로, 112 신고 건수 발생에 영향을 미치는 지역 특성 데이터들을 함께 학습하고 변수 중요도 도출을 통한 구체적인 요인 분석이 가능하다는 장점을 갖고 있다. 더불어 GRU 모델과 그래디언트 부스팅 모델을 멀티모달 스택킹 구조로 결합한 실험에서 결정 계수 평균 기준 0.7691의 성능을 보이며, GRU 모델만을 단독으로 사용한 경우보다 약 3%의 성능 향상도 이루어짐을 확인하였다.

4.4 MSM 모델 결과 해석

<Figure 6>은 MSM 보조 모듈을 통하여 도출된 변수 중요도를 히트맵(heatmap)으로 시각화한 결과이다. 해당 그림은 1일 후 예측 모델부터 7일 후 예측 모델에 해당하는 총 7개 그래디언트 부스팅 모델들의 변수 중요도가 각 행으로 표현되었다. 또한, 히트맵 그림의 각 열은 앞서 Table 1에서 설명한 변수들로 이루어져 있다. 히트맵 내의 색깔은 짙은 파란색일수록 분석에서 더욱 중요하게 활용되었음을 의미한다. 변수 중요도 결과를 통하여 기존 정보가 신고 건수 예측에 가장 많은 영향을 미쳤다는 사실을 확인할 수 있다. 또한, 보육 시설, 5대 범죄 발생 통계량, 주거지역 및 상업지역 등과 같은 해당 지역 용도 정보 변수 역시 주요한 변수로 선정되었다. 이외에도 주말 변수가 중요하게 활용되고 있어 유동인구의 수에 따른 112 신고 건수의 변화가 존재함을 알 수 있다. 해석 결과를 통하여 관할관서에서는 관할 구역의 지역 용도 정보를 이용하여 그 특성에 맞는 순찰 활동을 수행하고 주말 여부나 기존

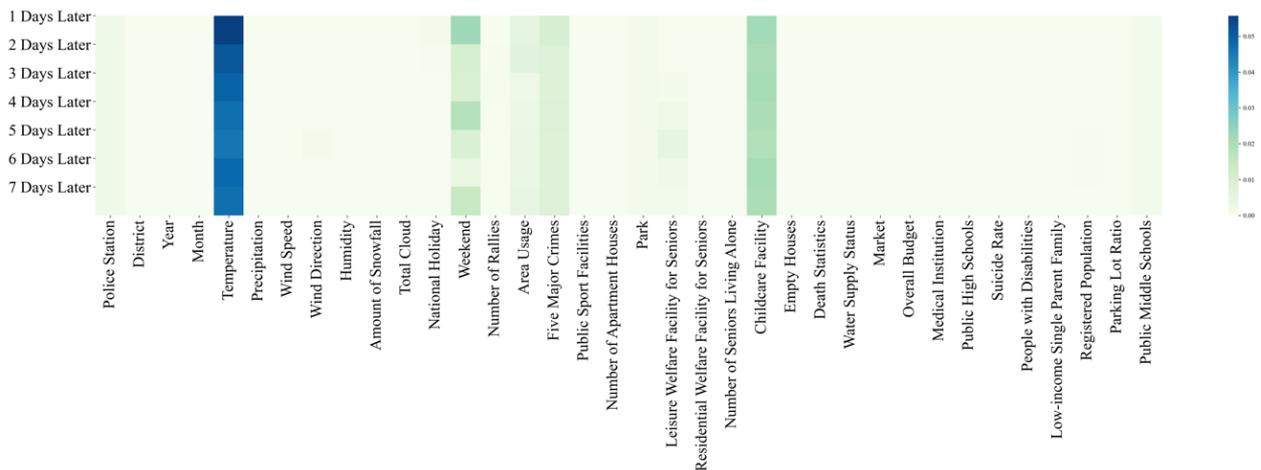


Figure 6. Feature Importance from the Gradient Boosting Models

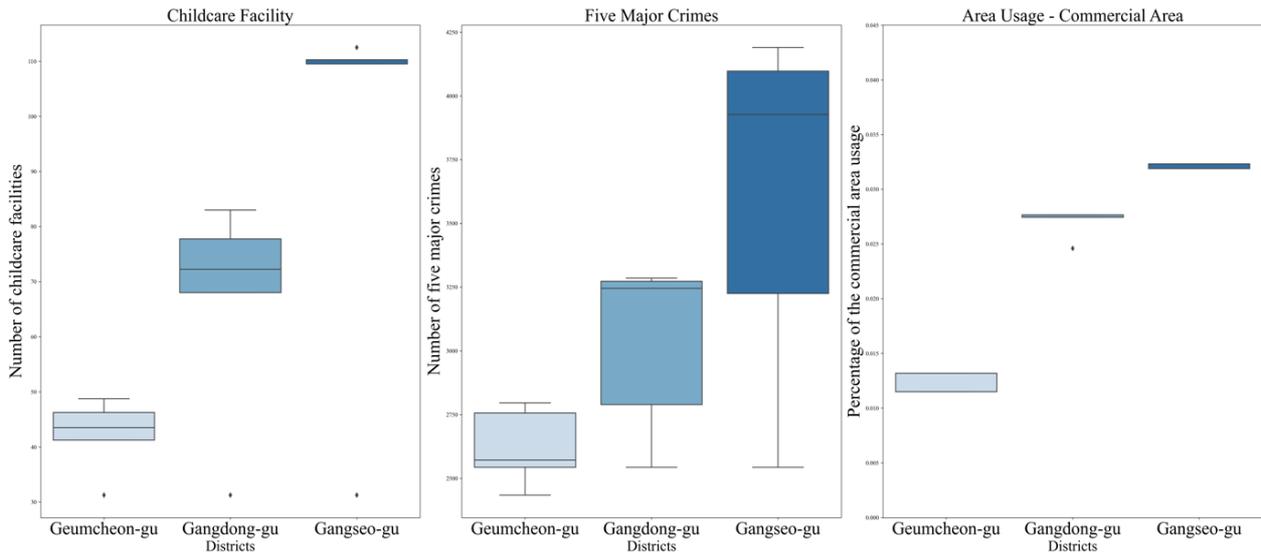


Figure 7. Box plot of Geumcheon-gu, Gangdong-gu, and Gangseo-gu for three different variables

정보에 따른 경찰력의 탄력적 배치가 필요하다는 결론을 도출할 수 있다.

주요 변수들 중 자치구 통계 데이터에 해당하는 보육 시설(childcare facility), 5대 범죄 발생 통계량(five major crimes), 지역 용도 정보 중 상업 지역의 비율 변수(area usage - commercial area)를 기준으로 자치구 별 차이를 살펴보았다. 비교를 진행한 세 자치구로는 예측 성능이 가장 높은 강서구와 가장 낮은 금천구, 그리고 평균적인 성능을 보인 강동구를 선정하였다. 예측 성능이 높아지는 금천구, 강동구, 강서구 순으로 변수별 상자 그림(box plot)을 그려 <Figure 7>와 같이 비교하였다. 상자 그림은 수치형 자료에 대한 최솟값, 사분위 값, 최댓값을 한번에 표현하여 데이터의 분포를 시각적으로 파악할 수 있는 그림이다. 시각화 결과에 따라 세 가지 변수 모두에 대하여 자치구별 차이를 확인할 수 있었다. 성능이 가장 우수했던 강서구와 가장 낮았던 금천구의 결과를 비교해보면 보육 시설의 수와 5대 범죄 발생 건수가 많을수록, 상업 지역의 비율이 클수록 신고 건수 예측 성능이 높아진다. 이와 같이 자치구별로 특성 차이를 보이며, 신고 발생 지역의 특성 차이를 나타내는 변수들이 실제 신고 건수 예측에 주요하게 활용되므로 해당 변수들을 반영하는 멀티모달 스택킹 모델이 효과적으로 구축되었다는 결론을 내릴 수 있다.

5. 결론

본 연구는 112 신고 건수를 예측하는 모델을 개발하여 대한민국 치안 시스템의 고도화 및 효율적인 경찰력 활용을 도모하기 위해 진행되었다. 112 신고 건수 데이터 외에 신고 발생에 영향을 미치는 기상, 공휴일, 집회 건수, 자치구 통계 데이터와 같은 지역 특성 데이터를 함께 수집하여 예측에 사용하였다.

신고 건수 데이터의 시계열 패턴을 반영하기 위한 GRU 모델에 지역 특성 데이터들의 효과적인 반응을 위한 그래디언트 부스팅 모델을 추가적으로 활용하였다. 각 모델의 예측 값에 신경망 모델을 이용한 스택킹 기법을 사용하여 최종 예측 값을 도출하는 MSM을 제안하였다.

MSM은 112 신고 건수 데이터에 딥러닝 기반 모델을 사용하여 정확도를 향상시키고 동시에 주요 변수 도출을 통한 결과 해석이 가능한 모델이라는 특징을 가지고 있다. 또한, 실험을 통하여 시계열 예측 모델과 다변량 예측 모델을 개별적으로 사용했을 때보다 더욱 향상된 예측 성능을 검증하였으며, 사용자의 필요에 따라 예측 기간을 유동적으로 설정할 수 있다는 효용성도 확인하였다. MSM로부터 도출된 112 예측 신고 건수와 이에 대한 해석 결과는 치안 정책 수립에 보다 과학적인 근거로 활용될 것으로 기대한다. 아울러 수사관들의 주관적인 경험에 근거한 경찰력 배치가 아닌 모델의 예측 결과 기반의 객관적이며 탄력적인 인력 배치에 도움을 줄 것으로 보인다. 본 연구에서는 서울특별시 내 관할관서 데이터만 활용하였지만, 향후 전국 단위의 112 신고 건수 데이터에 MSM을 적용하거나, 시간대 및 지역별 유동 인구 수와 같이 신고 건수에 영향을 미치는 추가적인 데이터를 모델 구축에 활용할 경우 더 높은 효용성을 가질 것으로 판단된다.

참고문헌

- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015), *Time series analysis: Forecasting and control*, John Wiley & Sons.
- Breiman, L. (2001), Random forests, *Machine Learning*, **45**(1), 5-32.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014), Learning Phrase Representations Using RNN Encoder-decoder for Statistical Machine

- Translation, ArXiv Preprint ArXiv:1406.1078.
- Chohlas-Wood, A., Merali, A., Reed, W., and Damoulas, T. (2015), Mining 911 calls in New York City: Temporal Patterns, Detection, and Forecasting, *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Choi, J. (2018), The Data-based Prediction of Police Calls Using Machine Learning, *The Korea Journal of BigData*, 3(2), 101-112.
- Friedman, J. H. (2001), Greedy Function Approximation: A Gradient Boosting Machine, *Annals of Statistics*, 1189-1232.
- Heo, S.-Y., Kim, J.-Y., and Moon, T.-H. (2018), Predicting Crime Risky Area Using Machine Learning, *Journal of the Korean Association of Geographic Information Studies*, 21(4), 64-80.
- Hochreiter, S. and Schmidhuber, J. (1997), Long Short-term Memory, *Neural Computation*, 9(8), 1735-1780.
- Holt, C. C. (2004), Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages, *International Journal of Forecasting*, 20(1), 5-10.
- Joo, I.-Y. (2012), A Case Study on Crime Prediction using Time Series Models, *Korean Security Journal*, 30, 139-169.
- Kang, S.-H., Yang, J., Han, B., and Lee, D.-H. (2017), Development of Crime Prediction Model: A Case Study in Anyang City, *Journal of the Korean Operations Research and Management Science Society*, 42(4), 135-146.
- Kim, D. and Jung, S. (2021), Comparison of Crime Forecasting Models based on Spatio-Temporal Data and Machine Learning, *Journal of the Architectural Institute of Korea*, 37(1), 135-143.
- Kim, J. M., Kim, M., Yoo, S., and Lee, D. (2012), Anlysis of Relationship between Reasons for Mereorological Change and Crime, *Convergence Security Journal*, 12(6), 107-113.
- Kim, J. and Lee, T. H. (2020), Big Data Analyses Models for Efficient Patrol and 112 Crime Call Response: Focusing on the Case of Daegu Metropolitan City, *Journal of Korean Public Police and Security Studies*, 17(3), 91-106.
- Kingma, D. P. and Ba, J. (2014), Adam: A Method for Stochastic Optimization, ArXiv Preprint ArXiv:1412.6980.
- Kubrin, C. E. and Weitzer, R. (2003), New Directions in Social Dis-Organization Theory, *Journal of Research in Crime and Delinquency*, 40(4), 374-402.
- Lee, S. W. and Cho, J. K. (2006), The Effects of Spatial and Environmental Factors on Crime Victimization, *Seoul Studies*, 7(2), 57-76.
- Park, J., Lim, H., and Lee, S. (2017), Determinants of the Five Major Crime Incidences in Seoul: Application of Panel Models, *Journal of The Korean Regional Development Association*, 29(2), 133-160.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2017), CatBoost: Unbiased Boosting with Categorical Features, ArXiv Preprint ArXiv:1706.09516.
- Ranson, M. (2014), Crime, Weather, and Climate Change, *Journal of Environmental Economics and Management*, 67(3), 274-302.
- Roh, S. (2017), Weather, Time, and Domestic Violence, *Korean Criminological Review*, 28(3), 69-95.
- Uittenbogaard, A. and Ceccato, V. (2012), Space-time Clusters of Crime in Stockholm, Sweden. *Rev. Eur. Stud.*, 4, 148.

저자소개

김서연 : 한양대학교 산업공학과에서 2020년 학사 학위를 취득하고 고려대학교 산업경영공학과에서 석사과정에 재학 중이다. 연구 분야는 인공지능, 머신러닝, Explainable Artificial Intelligence이다.

이민정 : 홍익대학교 산업공학과에서 2017년 학사 학위를 취득하고 고려대학교 산업경영공학과에서 박사과정에 재학 중이다. 연구 분야는 인공지능, 머신러닝, Artificial Intelligence for Multivariate Time-Series Data이다.

이지윤 : 명지대학교 산업공학과에서 2017년 학사 학위를 취득하고 고려대학교 산업경영공학과에서 박사과정에 재학 중이다. 연구 분야는 인공지능, 머신러닝, Explainable AI for Time-Series Data이다.

황하은 : 건국대학교 응용통계학과에서 2019년 학사 학위를 취득하고 고려대학교 산업경영공학과에서 석사과정에 재학 중이다. 연구 분야는 인공지능, 머신러닝이다.

장광호 : 경찰대학 행정학과에서 1997년 행정학사를 취득하고, 고려대학교에서 법학석사 학위, 명지대학교에서 2018년 행정학박사를 취득했다. 관심분야는 경찰데이터에 최신 기술을 적용하는 전략과 기술정책이다.

김희두 : 경찰대학 법학과에서 2014년 학사학위를 취득하고 고려대학교 빅데이터융합학과에서 석사과정에 재학 중이다. 경찰대학 치안정책연구소 스마트치안기능센터에서 재직하고 있다. 관심분야는 범죄분야 인공지능이다.

김중윤 : 경찰대학 치안대학원 범죄학과에서 2021년 석사학위를 취득하고 경찰청 위기관리센터에서 재직하고 있다. 관심분야는 범죄데이터 활용 머신러닝, 시계열분석 등이다.

김성범 : 한양대학교 산업공학과에서 1999년 학사를 취득하고 2001년과 2005년 미국 Georgia Institute of Technology에서 산업공학 석사학위, 박사학위를 취득하였다. 미국 텍사스주립대학교 교수를 역임하고 2009년부터 고려대학교 산업경영공학부 교수로 재직하고 있다. 연구 분야는 인공지능, 머신러닝, 최적화이다.