

¹다중 선형 회귀 모델을 이용한 서울특별시 성범죄 발생 건수 예측

한다송, 양희정, 이민희, 이영원, 김현희
동덕여자대학교 정보통계학과

20210868@dongduk.ac.kr, 20210847@dongduk.ac.kr, 20210853@dongduk.ac.kr, 20210855@dongduk.ac.kr,
heekim@dongduk.ac.kr

Prediction of the Number of Sexual Crimes in Seoul Using a Multiple Linear Regression Model

Da-Song Han, Hee-Jung Yang, Min-Hee Lee, Young-Won Lee, Hyon Hee Kim
Dept. of Statistics and Information Science,
Dongduk Women's University

최근 다양한 사회적 요인으로 인해 증간하는 성범죄는 심각한 사회 문제로 대두되고 있다. 이에 본 연구는 서울시 성범죄 발생 건수를 다중회귀분석을 통해 예측하고 예측값과 실제 발생 건수 간의 차이를 분석하여 서울시 행정구역별 맞춤형 치안 강화 방안을 제안한다. 이는 성범죄 발생 이전에 통제가능한 환경 요인을 제어하는 데 기여할 수 있는 기초 자료를 제공하는 데 의의가 있다.

1. 서론

최근 대한민국 대도시에서의 성범죄 발생 추세가 가파르게 상승하고 있다. 서울시의 경우 다양한 사회적 요인으로 인해 성범죄가 지속해서 증가하고 있으며, 이는 심각한 사회 문제로 대두되고 있다. 이에 따라 성범죄 예방을 위한 “여성안심지킴이집서비스”를 비롯한 여러 제도가 시행되었지만, 실질적인 효과는 명확히 평가되지 않았으며, 일부 지역에서는 범죄 예방에 미미한 성과를 보였다.

Jung 과 Lee(2015)에 따르면 대도시 성범죄에 영향을 미치는 요인으로 인구사회학적 특성과 물리적 특성을 나타내는 지표가 유의미한 예측변수로 작용한다. 또한, 사회경제적 요인이 범죄율에 미치는 영향을 분석한 연구[1]에 따르면, “소득 불평등”과 “실업률”이 범죄율 증가에 밀접한 관련이 있다는 결과가 도출되었다. 이에 본 연구는 ‘인구사회학적 특성’, ‘물리적 특성’, ‘경제적 특성에 해당하는 지표들을 활용한 다중회귀분석을 통해 서울특별시의 구별 성범죄 발생 주요 요인을 도출하고 구별 성범죄 발생의 취약점을 분석하여 집중적인 치안 강화 방안을 제

안하고자 한다.

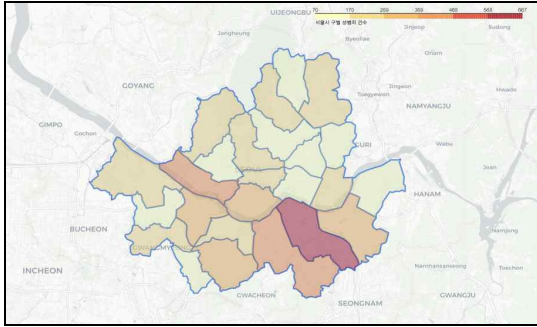
본 연구에서는 2023 년도 기준 서울시 공공데이터를 활용해 ‘인구사회학적 특성’, ‘물리적 특성’, ‘경제적 특성’에 해당하는 데이터를 수집 후 변수들 간의 상관관계 분석을 유의미한 변수들을 선택하였다. 후에, 성범죄 건수를 예측하는 다중회귀모델을 구축하고 모델의 유의적 검정을 통해 얻은 결과를 토대로 모형을 재구성하여 서울시 성범죄에 주요한 영향을 미치는 요인을 분석하였다.

2. 데이터 수집 및 전처리

서울시 성범죄 발생 건수 예측을 위한 데이터는 2023 년을 기준으로 서울시의 25 개 행정구역에 대해 통계청, 성범죄자알림 e, 서울시 열린 데이터광장을 활용하여 행정구역의 성범죄 발생건수와 총 18 개의 예측변수를 담은 데이터셋을 구성하였다. 이후, Feature-Selection 을 통해 선택된 변수들에 대해 Min-Max 스케일러를 적용하여 데이터 스케일링 작업을 수행하였다. 최종적으로 선택된 11 개의 변수는 다음과 같다: 'CCTV 개수', '청년 인구 비율(청년 인

¹ * 이 성과는 중소벤처기업부의 2024 년 SW/콘텐츠 분야 대학 지원 사업인 벤처스타트업 아카데미의 지원을 받아 수행된 연구임.

구/총인구)', '여성 인구 비율(여성 인구/총인구)', '일반 호텔 수', '지역 내 총생산(GRDP)', '면적', '인구 밀도(명/면적)', '성범죄자 거주 인구', '실업률', '경제활동 참가율', '유흥업소 밀도(단란주점 및 유흥주점 수/면적)'



(그림 1) 구별 성범죄 발생 건수

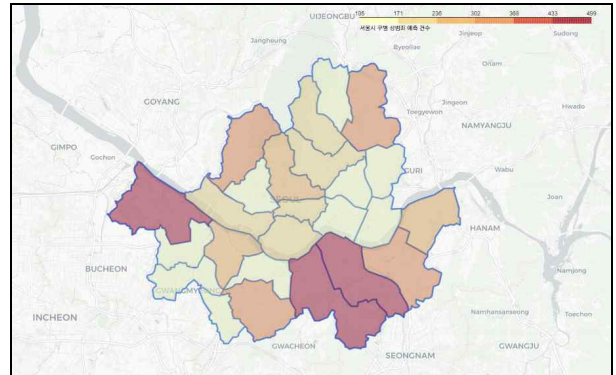
다중회귀모형을 구성하기 전에 예측모형의 결과값인 성범죄 발생 건수에 대한 데이터 탐색을 진행한 결과, 그림 1 과 같이 강남구가 유의미하게 높은 것을 확인하였다.

3. 서울시 성범죄 발생 예측 다중회귀모델

전처리 결과 도출된 인구사회학적 특성, 경제적 특성, 물리적 특성에 포함되는 총 11 개의 예측변수값들을 이용하여 다중회귀모형을 생성하였다. 전체 데이터의 80%는 훈련 데이터로, 나머지 20%는 테스트 데이터로 할당 후 Grid-Search 기법을 사용하여 최적의 하이퍼 파라미터를 도출하고 다중회귀모형에 적용하였다.

4. 성능 평가 및 예측 결과

다중회귀모델의 성능을 나타내는 결정계수(R-squared)는 0.94, F-통계량(F-statistic)은 11.38, p-값이 0.000992 로 매우 유의미한 수준임을 나타낸다. 개별 변수들의 회귀계수를 살펴보면, 'x6(면적)' 과 'x11(유흥업소 밀도)' 이 p-값이 각각 0.021, 0.029 를 보여 95% 신뢰 수준에서 유의하였다. 이에 따라 x6 과 x11 만을 포함한 새로운 단순화된 회귀모델을 생성하였으며, 해당 모델의 최종 결정계수(R-squared)는 0.308 이었다. 단순화된 회귀모형으로 성범죄 건수를 예측하였을 때의 결과는 그림 2 와 같다.



(그림 2) 구별 성범죄 발생 건수 예측값

이후에 실제 성범죄 건수와 단순화된 다중회귀모델로 예측된 성범죄 건수와의 차이를 사후분석한 결과 단순화된 다중회귀모델의 경우, 강남구(-205.238)와 마포구(-227.753)에서 성범죄 발생 건수를 크게 과소평가하고 있다. 반면, 강서구(+206.316)와 노원구(+158.543)에서는 과대평가하고 있다. 중랑구(+3.061), 양천구(+4.206), 동대문구(-4.956)에서는 차이가 크지 않아, 모델이 비교적 신뢰할 수 있는 예측을 하고 있었다.

5. 결론 및 향후 연구

본 연구는 다중 선형 회귀 분석을 통해 서울특별시 25 개 구의 성범죄 발생 건수를 예측하고, 예측값과 실제 발생 건수 간의 차이를 분석하였다.

강서구와 노원구에서는 성범죄 발생 예측값이 실제 값보다 높게 나타났으며, 이는 '청년 인구', '여성 인구', '유흥시설' 등의 변수가 실제 범죄 발생과 과대 상관되었을 가능성을 보여준다. 이러한 지역에서는 현재 설치된 CCTV 와 같은 방범장치의 효과를 점검하고, 추가적인 범죄 예방 조치, 예를 들어 순찰 강화와 같은 대책이 필요하다.

반면, 마포구와 강남구에서는 예측값이 실제 발생 건수보다 낮게 나타났으며, 이들 지역은 외부 방문객이 많고 유흥업소가 밀집된 특성이 있어 범죄 발생 가능성이 높다. 따라서 해당 지역에서는 보다 적극적인 범죄 예방 조치가 필요하다. 특히, 야간 순찰 강화, 유흥업소 밀집 지역에 대한 집중 감시, (여성 보호를 위한 안전지킴이집 확충) 등이 고려될 수 있다.

본 연구는 2023 년도 데이터를 기반으로 분석되었기 때문에 장기적인 성과를 위해서는 최근 10 년간의 서울시 성범죄 데이터와 관련 요인을 추가로 분석할 필요가 있다. 이를 통해 더욱 신뢰할 수 있는 예측 모델을 구축하고, 보다 정밀한 성범죄 예방 방안을

제시할 수 있을 것으로 기대된다.

참고문헌

- [1] 정승우, 이경훈, “대도시성범죄발생에 영향을 미치는 환경요인 분석”, 대한건축학회논문집 계획계, Vol.31 No.11, pp.179-186, 2015.
- [2] 변정희, “성범죄 피해에 대한 공간적 영향요인 연구 : 부산의 성매매와 유흥업소 밀집 지역을 중심으로”, 석사학위논문, 부산대학교, 2023.
- [3] 염윤호, 정진욱, “신상정보공개 성범죄자 거주지의 지리적 분포와 사회구조적 요인들 간의 관계에 대한 공간회귀분석”, 형사정책연구, 제 30 권 제 1 호, pp43-67, 2019.
- [4] Anser, M.K., Yousaf, Z., Nassani, A.A. et al. Dynamic linkages between poverty, inequality, crime, and social expenditures in a panel of 16 countries: two-step GMM estimates. Economic Structures 9 , 43 (2020).