

공개 데이터 기반 머신러닝 기법을 활용한 상권 매출 예측 모델 연구

신승민¹, 신창영¹, 임성훈¹, 임지성¹, 이택², 김정동², 이정빈²

¹선문대학교 컴퓨터공학부 학부생

²선문대학교 컴퓨터공학부 교수

{dolkuji1, toyoaki900, paulodybi, ljsoung0923, comtaek, kjd4u, jungbini}@sunmoon.ac.kr

A Machine Learning Approach to Revenue Prediction in Commercial Districts Using Open Data

Seung-Min Shin¹, Chang-Young Shin¹, Sung-Hoon Lim¹, Ji-Sung Lim¹,
Taek Lee², Jeong-Dong Kim², Jung-Been Lee²

^{1,2}Div. of Computer Science and Engineering, Sun Moon University

요약

최근 자영업자 수는 매출 부진과 원자재 가격 상승 등의 이유로 급격히 감소하고 있으며, 많은 자영업자가 폐업을 고려하고 있다. 본 연구는 이러한 문제를 해결하기 위해 서울시 공개 데이터를 활용하여 K-평균 알고리즘을 사용해 서울시 지역을 259개의 소지역으로 군집화한 뒤, 군집별로 유동 인구, 매출 등 다양한 정보를 수집하고, 이를 기반으로 회귀 모델을 구축하여 군집별 매출을 예측하였다. 그 결과, 변수별 상관관계 및 중요도 분석을 통해 매출에 영향을 주는 주요 요인을 도출할 수 있었다.

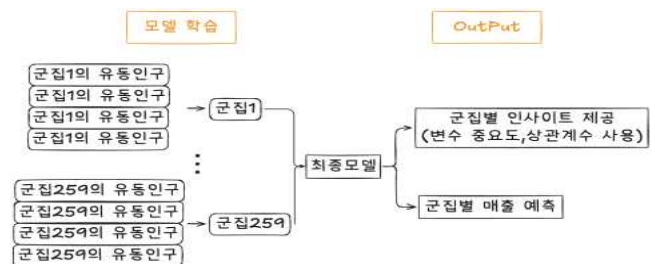
1. 서론

최근 자영업자의 수는 매출 부진과 원자재 가격 상승 등의 복합적인 요인으로 인해 급격히 감소하고 있다[1]. 서울시는 공공데이터 형태로 유동 인구, 매출 관련 정보를 제공하고 있으나, 해당 데이터를 직접 수집·분석하여 활용하는 것은 자영업자 개인에게 현실적으로 어려운 일이다.

따라서, 본 연구는 서울시의 공공데이터를 활용하여 상권별 매출을 예측하고, 이를 바탕으로 재고 관리 및 인력 운영 등 경영 판단에 도움을 줄 수 있는 기계학습 모델 및 애플리케이션을 개발하였다. 특히, K-평균 알고리즘을 적용하여 서울시를 여러 개의 군집으로 분류함으로써 군집별 특성을 반영한 예측이 가능하도록 하였다. 나아가 매출에 영향을 미치는 변수들의 상관관계와 중요도를 군집별로 분석하여, 자영업자가 타겟 고객층을 이해하고 최적의 운영 전략 수립이 가능한 애플리케이션을 개발하였다.

2. 상권 매출 예측 모델 개발 및 분석 결과

기계학습 기반 상권 매출 예측 모델 개발 과정은 그림 1과 같다. 엘보우 기법, 실루엣 계수 및 데이터 셋 결합 과정의 손실률 분석을 통해 K-평균 알고리



(그림 1) 상권 매출 예측 모델 개발

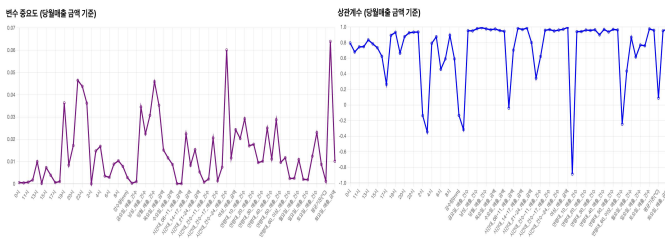
즘에 사용되는 최적의 K값을 259로 설정하였다. 유동인구, 기상, 공휴일 정보 등 다양한 변수를 사용해 매출을 예측하는 회귀분석 모델을 개발하였고 선형 회귀, 랜덤 포레스트, XGBoost를 포함한 모델들을 단계적으로 결합하는 앙상블 기법을 사용하여 예측 정확도를 향상했으며, 이는 기존 연구에서 제시된 다중 모델 기반 스택킹과 유사한 구조를 일부 공유한다는 점에서 타당성을 갖는다[2,3].

본 연구에서는 광화문역을 예시로 하여 해당 지역이 포함된 군집의 매출을 분석하였다. 그림 2는 광화문역이 속한 군집을 나타내고 있고, 그림 3은 상관 변수 및 중요도 변수를 시각화한 그래프이다. 광화문 군집을 분석한 결과, ‘여성_매출_금액’이 매출과 가장 높은 상관관계를 보였고, ‘화요일_매출_건수’가 매출에 가장 큰 영향을 미치는 변수로 확인되었다. 이를 통해 자영업자는 여성 고객층을 주요 타

것으로 설정하고, 화요일을 중심으로 한 맞춤형 이벤트나 프로모션을 기획함으로써 매출 증대를 기대해 볼 수 있다.



(그림 2) 광화문 군집 시각화 결과

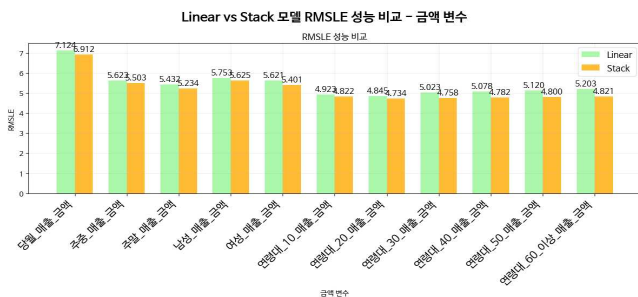


(그림 3) 광화문 군집 상관 변수 및 중요도 분석 결과

3. 모델 성능 평가

본 연구에서는 서울시 상권 데이터를 바탕으로 군집별 매출을 예측하기 위해 다양한 머신러닝 회귀 모델을 활용하였으며, 그 성능을 비교 평가하였다. 먼저 선형 회귀 모델을 베이스라인으로 설정하고, 이후 RandomForest, LightGBM, XGBoost 및 선형 회귀를 단계적으로 결합한 앙상블 모델을 구성하였다. 각 모델은 동일한 입력 변수와 교차 검증 조건 하에 학습되었으며, 성능 평가 지표로는 RMSLE를 사용하였다.

그림3에서 보는 바와 같이 스택킹 모델은 개별 모델과 비교했을 때 RMSLE 평가 지표 기준으로 모든 종속변수에서 개선된 성능을 보였다. 특히, 당월 매출 금액 변수의 경우 RMSLE 값이 선형 회귀 모델의 7.124에서 스택킹 모델의 6.912로 감소하여, 예측 정확도의 향상을 확인할 수 있었다.



(그림 4) RMSLE 평가에 대한 비교 지표

4. 결론

본 연구에서는 서울시 공공데이터를 활용하여 자영업자의 상권 매출 예측을 위한 기계학습 모델 및 웹 기반 애플리케이션을 개발하였다. 제안된 시스템은 군집별 유동 인구, 기상, 공휴일, 매출 데이터를 활용하여 머신러닝 기반 분석 모델을 구성하였다. 특히, 선형 회귀를 포함한 여러 예측 모델의 결과를 단계적으로 결합하는 기법을 적용함으로써 예측 성능을 향상시켰으며, 변수 중요도와 상관관계를 분석해 인사이트를 도출하는 기능을 구현하였다.

사용자는 웹 애플리케이션을 통해 자신의 위치 정보만 입력하면 해당 상권의 매출에 영향을 미친 요인들을 시각적으로 확인할 수 있으며, 이를 바탕으로 재고 및 인력 관리 전략을 수립할 수 있다. 결과적으로 자영업자의 경영 판단을 데이터 기반으로 지원함으로써 매출 극대화 및 경영 효율화에 기여할 수 있을 것으로 기대된다.

향후에는 예측 정확도를 높이기 위한 딥러닝 모델의 도입, 다른 지역 데이터 확장, 사용자 맞춤형 경고 시스템이나 자동 알림 기능 등의 고도화를 통해 상권 분석 플랫폼으로의 발전 가능성을 모색할 예정이다.

감사의 글

이 논문은 2025년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업(2024-0-00023) 및 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2023-00243114).

참고문헌

- [1]한국경제인협회, 자영업자 2024년 실적 및 2025년 전망 설문조사, 2025.02.24, https://www.fki.or.kr/kor/news/statement_detail.do?bbs_id=00036033&category=ST
- [2] Sitorus, Z., Syahputra, I., Angkat, C.I., & Sartika, D. Implementation of K-Means Clustering for Inventory Projection., International Journal of Science, Technology & Management, Vol.2, 673-678. 2023
- [3] Sun, C., Forecasting Retail Sales Via the Use of Stacking Model. In Proceedings of the 2022 International Conference on Education, Development and Business Communication, University of North Texas in Dallas, TX, 2022, pp. 405-411.