**1. 서론**

**1-1. 주제 선정 배경/목표**

최근 서울시를 포함한 대도시에서는 공휴일이나 축제 기간 등 특정 시점에 특정 지역으로 인구가 급격히 집중되는 현상이 반복적으로 나타나고 있다. 이와 같은 인구 집중은 교통 혼잡, 환경 문제, 안전 사고, 지역 상권의 과부하 등 다양한 사회적 문제를 유발한다.  
혼잡도 예측을 위한 시스템과 알고리즘이 시중에 다수 존재하지만, 실제 현장에서는 여전히 사후적 대응에 그치는 경우가 많다. 특히 소상공인이나 자율적으로 운영되는 상권에서는 예기치 못한 고객 유입에 대해 인력 및 자원 확보 측면에서 즉각적인 대응이 어려운 경우가 빈번하다.  
이에 따라 본 프로젝트는 서울시 공공데이터를 활용하여, 특정 지역의 생활인구(상주 인구 + 유동 인구)를 외부 요인에 기반해 예측하고, 이를 통해 지역 혼잡 가능성을 사전에 파악할 수 있는 모델을 개발하는 것을 목표로 잡았다.

**1-2. 연구 대상 및 분석 조건 설정**  
이번 프로젝트는 서울시 주요 지역의 생활인구 예측을 통해 지역 혼잡도를 사전에 파악하고, 이에 기반한 공공 및 민간의 대응 방안을 제시하고자 한다. 이를 위해 필요한 분석 조건을 설정하였다:

**분석 기간:** 2024년 5월 1일 ~ 2025년 4월 30일  
→ 계절성, 명절·공휴일·연휴 등 주요 요인을 포함할 수 있으며, 대부분의 축제가 연례적으로 반복된다는 점을 고려하였다.  
→ 정책, 문화행사 등 공공 및 민간 사업이 연 단위로 기획·집행되는 점에서 분석 해석의 실효성 확보.  
→ 시간 단위 데이터도 고려하였으나, 데이터량 증가(약 24배) 대비 feature의 시간대별 변동성이 크지 않아 최종적으로 일 단위로 제한.

**분석 지역:**  
서울시 내 상업, 업무, 관광, 교통, 문화 기능이 집중된 16개 주요 행정동을 선정하였다. 해당 지역들은 일반 주거지보다 외부 요인에 따른 생활인구 변동성이 크고, 혼잡도 예측의 실효성이 높다고 판단되었다.

**1-3. 주요 병합 데이터셋 소개**  
생활인구 예측을 위해 서울시 공공데이터 및 외부 요인을 반영한 데이터를 병합하여 분석용 통합 데이터셋을 세팅했다. 각 데이터는 기준일(STDR\_DE\_ID)과 행정동 코드(ADSTRD\_CODE\_SE)를 기준으로 병합되었으며, 구성은 아래와 같다:

* **생활인구 데이터** 
  + 출처: 서울 열린데이터광장
  + <https://data.seoul.go.kr/dataList/OA-14991/S/1/datasetView.do>
  + 설명: 서울시 행정동 단위의 내국인 생활인구 데이터
  + 활용: 예측 대상 변수(TOT\_LVPOP\_CO)로 사용. 시간대별 수치를 일별 총합으로 변환
* **기상 데이터**
  + 출처: 기상청 기상자료개방포털
  + <https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36>
  + 변수: 평균 기온, 최고/최저 기온, 강수량, 풍속, 습도, 지면 온도 등
  + 설명: 서울시 평균값을 모든 행정동에 동일 적용
* **공휴일 정보**
  + 출처: 공공데이터포털 OpenAPI
  + <https://www.data.go.kr/data/15012690/openapi.do>
  + 변수: is\_holiday (공휴일 여부 이진 처리)
  + 설명: API 매핑 기반 수집 후 누락값은 수작업으로 보완
* **문화행사 정보**
  + 출처: 서울 열린데이터광장
  + <https://data.seoul.go.kr/dataList/OA-15486/S/1/datasetView.do>
  + 변수: culture\_event\_count, event\_type(유무료 구분), has\_culture\_event
  + 설명: 위도·경도 좌표 기반 데이터를 카카오 API로 행정동 코드로 변환(reverse geocoding) 후 병합
* **교통 데이터 (지하철 승하차)**
  + 출처: 서울 열린데이터광장
  + [https://data.seoul.go.kr/dataList/OA-12914/S/1/datasetView.do#](https://data.seoul.go.kr/dataList/OA-12914/S/1/datasetView.do)
  + 변수: subway\_inflow, subway\_outflow
  + 설명: 지하철역과 행정동 간 수작업 매핑. 다만 결측치 및 노이즈 문제로 주요 모델에서는 제외됨

**1-3. EDA, 시각화**

**2. 본론: 예측 및 모델링**

**2-1. 예측 목표 및 접근 전략**생활인구는 날씨, 공휴일, 문화행사 등 다양한 외부 요인의 영향을 받기 때문에, 이러한 변수들을 반영한 예측 모델이 필요하다고 판단하였다.  
이에 따라 모델 구조는 총 세 가지 방식으로 구분하여 구성하였다:

* 개별 모델: 각 행정동별로 독립적인 모델을 학습하는 방식
* 통합 모델 (전체 기준 정규화): 전체 데이터를 하나로 통합한 후, target을 전체 기준으로 정규화하여 학습
* 통합 모델 (행정동별 정규화): 통합된 데이터 구조를 유지하면서도 target은 행정동별로 정규화하여 학습

세 가지 모델은 각각 해석력, 성능, 확장성 측면에서 차이가 있기 때문에 이를 비교해 가장 효과적인 구조를 도출하고자 하였다.

**2-2. 데이터 전처리 개요**  
모델 학습에 앞서 데이터는 다음과 같은 방식으로 전처리되었다:

* 기준일에서 월 정보를 추출하여 계절성 변수(month)를 생성
* 주말 및 공휴일 정보를 활용해 이진 변수 is\_holiday를 구성
* 문화행사 관련 정보 중 설명력이 낮은 분류 변수는 제거하고, 유무/유료 여부를 반영해 event\_type 변수 생성 (0: 없음, 1: 유료, 2: 무료)
* 생활인구(target)는 지역 간 규모 차이를 보정하기 위해 행정동별로 개별 MinMax Scaling 수행
* 지하철 승하차 변수는 예측 이후에만 확인 가능한 정보이므로 제외함

**2-3. 모델 구조 및 설계 방식**  
(1) 행정동별 개별 모델  
각 행정동에 대해 독립적인 모델을 학습하였다. 총 16개 동을 대상으로 XGBoost와 Random Forest (Grid SearchCV 적용) 모델을 구성하고, target은 각 동별로 정규화하였다. 지역별 특성을 민감하게 반영할 수 있다는 장점이 있지만, 데이터 분산과 모델 관리의 부담이 존재하였다.

(2) 통합 모델 (전체 기준 scaling)  
전체 데이터를 통합하고 행정동 코드를 범주형 변수로 포함한 단일 모델을 학습하였다. 이때 target은 전체 데이터 기준으로 정규화하였다. 수치상으로는 RMSE 0.03, R² 0.97로 높은 성능을 보였지만, 모델이 행정동 정보를 통해 평균값만 예측하는 구조적 한계가 드러났다.

(3) 통합 모델 (행정동별 target scaling)  
통합 모델 구조를 유지하면서도 target은 행정동별로 정규화하였다. 이 방식은 전체 데이터를 활용하면서도 각 동의 상대적 생활인구 변동을 반영할 수 있었고, 해석력과 예측력 모두에서 균형 있는 성과를 보였다.

**2-4. 성능 평가 및 결과 해석**  
모델 성능은 RMSE와 결정계수 R²를 기준으로 평가하였다.

(1) 개별 모델 결과  
대다수 지역에서 중간 이상의 성능을 보였으며, Random Forest가 XGBoost보다 소폭 우수한 결과를 나타냈다. 성수동, 을지로동, 역삼동은 R² 0.85 이상을 기록했으며, 반면 송파1동, 서교동 등은 낮은 R²를 보여 지역 간 편차가 확인되었다.

(2) 통합 모델 (전체 기준 scaling)  
R² = 0.97, RMSE = 0.03으로 수치상 우수했으나, 모델이 각 지역의 평균값만을 예측하는 경향을 보여 실제 활용에 적합하지 않았다.

(3) 통합 모델 (행정동별 scaling)  
R²는 0.88 수준으로, 평균값 대비 상대적 변동을 잘 학습했다. 외부 요인(공휴일, 행사, 날씨 등)이 지역별 생활인구에 어떤 영향을 주는지 모델이 파악할 수 있도록 설계되었다.

(4) 종합 비교

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 모델 방식 | 데이터 구성 | Target scaling | 평균 R² | 해석력 | 확장성 |
| 개별 모델 (XGB/RF) | 행정동별 분할 | 행정동별 | 0.59~0.65 | 높음 | 낮음 |
| 통합 모델(전체scaling) | 전체 통합 | 전체 기준 | 0.97 | 낮음(형식적 예측) | 높음 |
| 통합 모델(행정동별scaling) | 전체 통합 | 행정동별 | 0.88 | 높음 | 높음 |

결과적으로, 개별 모델은 지역 특화 분석에는 유리했지만 비효율적이었고, 전체 기준 scaling 모델은 수치상 우수하나 실효성이 낮았다. 행정동별 scaling을 적용한 통합 모델은 예측력, 해석력, 확장성 측면에서 가장 균형 잡힌 방식으로 평가된다.

**3. 결론**

**3-1. 종합 요약**  
예측 모델은 구성 방식과 target 처리 방식에 따라 세 가지 구조로 구분하여 비교하였다.  
① 각 행정동별 개별 모델, ② 통합 모델 (전체 target scaling), ③ 통합 모델 (행정동별 target scaling) 방식이다.  
실험 결과, 전체 기준으로 정규화한 통합 모델은 수치상 가장 높은 성능(R² = 0.97)을 보였지만, 행정동 feature를 통해 평균값을 그대로 반영하는 구조적 한계가 있었다.  
반면, 행정동별로 target을 정규화한 통합 모델은 안정적인 성능(R² = 0.88)과 해석력을 동시에 확보하였다.  
개별 모델은 지역 민감도 분석에는 효과적이었지만, 데이터의 분산과 모델 관리 측면에서 비효율적이었다.

**3-2. 기대 효과 및 활용 가능성**  
생활인구 예측은 지역의 혼잡도를 사전에 파악하고 대응할 수 있는 기반이 될 수 있다.  
특정 시점에 인구가 몰릴 가능성을 미리 예측할 수 있다면, 교통, 환경, 안전 등의 행정 운영뿐 아니라 지역 상권에서도 인력 및 자원 배치를 유연하게 조정할 수 있다.  
이번 분석을 통해 공휴일, 날씨, 문화행사 등의 외부 요인이 생활인구 변화에 미치는 영향을 정량적으로 파악할 수 있었으며, 이를 통해 생활인구는 일정한 흐름을 따르기보다는 외부 요인에 민감하게 반응하는 변수임을 확인할 수 있었다.  
또한, 행정동별 모델 성능 차이를 통해 지역에 따라 생활인구의 반응 패턴이 상이함을 발견하였으며, 이는 향후 혼잡 대응 및 마케팅 전략 수립 시 지역 맞춤형 대응 체계의 필요성을 시사한다.  
특히, 문화행사와 같은 이벤트성 변수는 전체 데이터 내 비중은 작지만, 존재하는 경우 생활인구 변화에 일정 수준의 영향을 줄 수 있음을 확인하였다.  
또한, 통합 모델 구조는 데이터 효율성과 정책적 활용 가능성 측면에서 장점이 크며,  
실제 행정 혹은 플랫폼 설계 과정에서 하나의 통합 예측 모델을 운영 기반으로 적용할 수 있는 가능성을 보여준다.

**3-3. 한계점 및 추후 개선 방안**  
1. 문화행사 데이터의 경우 위치 정보가 부정확하거나 기간이 누락된 경우가 다수 존재해, 전처리 과정에서 유실된 레코드가 발생하였다. 이로 인해 일부 지역의 이벤트 반영이 제한되어 정밀도 저하의 원인이 될 수 있다.  
2. 생활인구 데이터가 일 단위로 집계되는 특성상, 시간대별 혼잡 양상이나 특정 시간의 인구 흐름을 반영하기 어려웠다. 특히 상권 운영이나 실시간 정책 대응과 연결하기에는 정보가 부족하였다..  
3. 교통 관련 변수는 지하철 승하차 정보를 일부 매핑하여 시도하였으나, 해당 데이터가 관측 이후에만 확인 가능한 구조여서 실시간 예측과 충돌하였었다. 또한 버스 이용량이나 도로 혼잡도와 같이 보다 세부적인 교통 흐름 정보는 데이터 수집·병합 측면의 기술적 제약으로 활용하지 못하였다.  
4. 민간 행사나 최근 중요하게 여겨지는 SNS 반응과 같은 비정형 외부 요인은 수집·정량화 과정의 한계로 분석에 포함하지 못하였으며, 이로 인해 생활인구 변화의 특정 패턴을 충분히 설명하지 못하였다.

5. 행정동 코드 기준으로 분석을 진행하여 행정동 코드가 동일한 종로1,2,3,4동은 각각의 개별적인 특성은 해석할 수 없었다.

향후에는 시간 단위 또는 실시간 유동인구 데이터를 포함하여 보다 세분화된 인구 흐름 분석이 가능한 예측 모델로 확장하는 것이 좋아 보인다.  
또한, 교통량, 축제 규모, 온라인 반응 등 외부 요인의 정량화를 위한 기술적 기반이 확보된다면, 예측의 정밀도뿐만 아니라 정책 수립 및 지역 대응 전략 수립 측면에서도 훨씬 높은 활용 가능성을 기대할 수 있을 것 같다.