21.08.04(수) - 경기도 고양시 공공자전거 정리

INDEX

1. EDA
    1. 자전거 스테이션 입지 및 이용 현황
    2. 불만 상황 파악
    3. 분석 방향 및 기대효과
2. MODELING
    1. 모델링 개요
    2. 데이터셋 설명 및 전처리
    3. 학습용 데이터와 예측용 데이터 구축
    4. 모델 학습
    5. 스테이션 배치
3. EVALUATION
    1. 배치 결과와 수요 충족도
    2. 한계점 및 의의

현재는 인구 분포도 에 따라 하고 있음

반납이 안되는데 있어서는 기존의 인구 분포도에 따른 영향이 있을 수 도 있다.

사용비율 같은걸 볼 수도 있다.

수요가 많은쪽이 생기는 곳에 더 많으면 좋을 듯

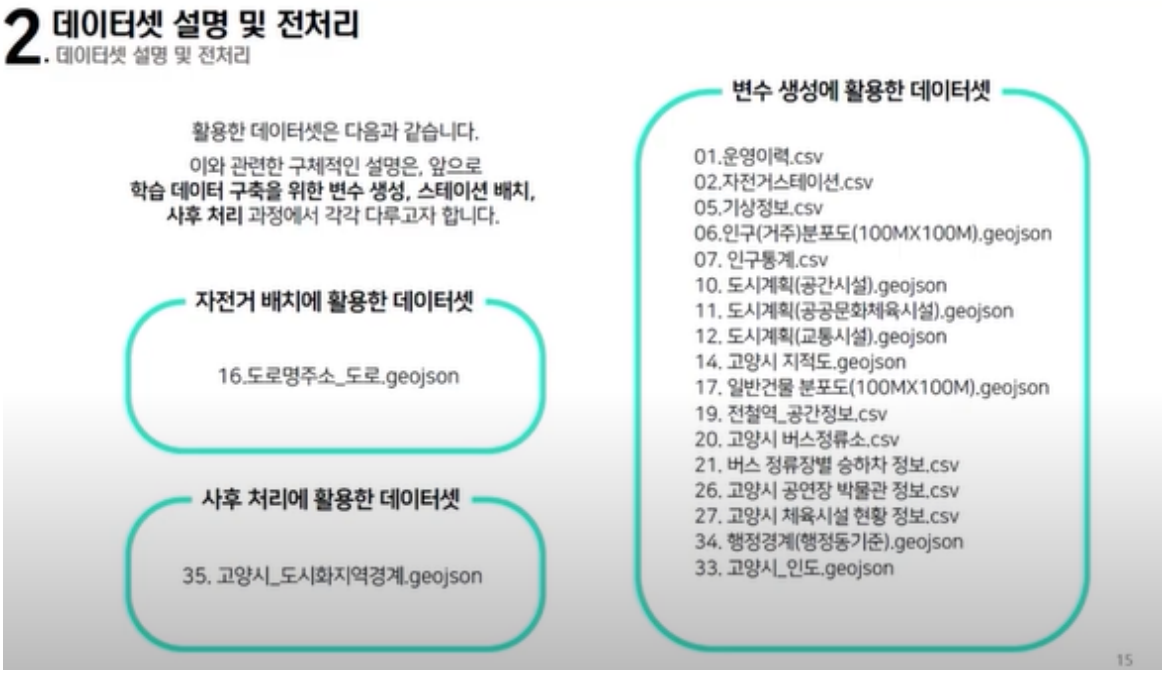
새로운 / 잠재적인 수요를 반영하는 변수들을 포함하여
각 지역수요를 예측하자.

수요가 있는곳에 확실히 네트워크를 구축하자.

모델링 개요

2020년의 하루 대여건수를 예측

1. 고양시의 50m*50m에 해당하는 인구분포도를 이용하여 후보군 생성
2. 주어진 데이터를 이용하여 후보군 주변 정보를 변수화
3. 기존에 설치된 스테이션 정보를 통해 모델 구축 및 후보군의 수요 예측
4. 고양시 도시화 계획 구간의 예측값 및 인구분포도를 기준으로 사후처리 진행
5. 후보군들 대상으로 클러스터링 진행 후, 각 클러스터에 해당하는 자전거 스테이션 개수 결정
6. 최종적으로 거치대 수량 및 스테이션의 좌표 생성



시간에 따른 변수는 인구추이 같은걸 볼 수 있고 인구 분포나 건물 정보를 볼 수 있다.

예측용 데이터 구축

50M
각각의 점들에 대한 정보를 변수로 만들어 데이터 프레임으로 만들었다.

변수생성
인구 추이나 기타등등

기존의 164개의 데이터를 가지고 학습하면 오버피팅이 되서 데이터를 늘리는 작업을 하였다. - 자전거 빌리는 곳 위치

시간 프레임으로 만들었다.

이것을 매일 빌린 대여량을 통해 하나의 후보군으로 만들어줬다. 그래서 1년치이면 365개의 후보군이 생기고 3년치면 365*3 의 후보군이 생긴다.

최종적으로 활용한 예측용 데이터는

각 후보군 별로 월(1~12) * 요일(월~일) 로 84개의 행을 구했다.

이 84개 예측값의 평균값을 해당 후보군의 수요로 예측을 하였다.

모델학습

미래 지향적인것이기 때문에 (미래에 수요가 있어야 되는 것이기 때문에)

Time Series Cross Validation 을 사용하였다.

그리고 성능을 올리기 위해 permutation Test를 통해 변수를 추리는 과정을 할 수 있다.

예측값과 한계점

애초에 빌린 곳에서 밖에 데이터를 가지고올수 밖에 없기때문에

이건 모든 지역에 수요가 있다고 나올 수 밖에 없다.

그래서 사후처리를 했다.

스테이션 배치

1. 실질적 수요 존재 지역 선정
    1. 도시화 지역 경계 인구분포도와 건물 분포도를 기준으로 실질적 수요가 적은 후보군 필터링 - o
    2. 밀도기반 클러스터링을 통해 noise제거 - x
    3. 도시화 지역 경계에 해당하는 클러스들의 예측값을 기준으로 실질적 수요가 적은 후보군 필터링 - ?
    4. 후보군을 도로에 배치 - ?
    5. 클러스터를 구성하는 후보군의 개수를 기준으로 면적이 적은 클러스터 필터링 - ?
2. 스테이션 배치
    1. 클러스터의 평균 예측값과 클러스터의 넓이를 고려하여, 각 클러스터에 적절한 스테이션 개수 할당 => 도서관 크기(다각형을 구하는 걸 통해 구함)
    2. 각 클러스터에 할당된 스테이션 개수만큼 스테이션 설치 - k means로 배치
3. 거치대 수량 결정
    1. 예측값의 상위 25% 값을 이용하여 각 스테이션의 거치대 수량 결정
    2. 순환율이 저조한 것으로 보이는 스테이션 근처에 추가적인 배치

최종 결과

새롭게 인구분포가 높아지는 지역 / 인구분포도가 원래 높았던 지역

기존 도서관 배치랑 비교