스크립트

1~3p

안녕하세요. 1조는 1밖에 몰라 팀입니다.

저희 발표 주제는 **서울형 키즈카페 입지선정 추천모델**입니다.

프로젝트 소개 후 데이터 수집과 전처리 과정, 기초 분석과 모델링을 어떻게 했는지 그리고 기대효과와 2차 프로젝트 계획까지 말씀드리겠습니다.

함께한 팀원들입니다.

프로젝트 소개 및 기획 배경

5p

서울형 키즈카페는 서울시가 운영하는 공공 놀이시설로, **저렴한 입장료**가 큰 특징입니다. 서울시는 저출산 대응 정책의 하나로 서울형 키즈카페 **확대**를 추진 중인데요,

6p

하지만 현재 운영 중인 서울형 키즈카페는 기대보다 **낮은 이용률**을 보이며, 위치 또한 접근성이나 수요보다는 **행정적 편의 기반**이라는 문제점을 발견하였습니다.

7p

이에 저희는 **데이터 기반**의 입지 추천 모델을 구축해 **정책 효과**를 높이고, **돌봄시설수요를 충족**하는데 기여하며 **소외 계층**을 포함한 다양한 시민들에게 이용 기회를 확대하고자 합니다.

8p

이러한 기획의도와 목표를 위해 다음과 같은 흐름으로 프로젝트를 진행했습니다.

1단계로 격자 기반의 공간 단위를 설정하고 2단계로 수집한 데이터를 매핑시켰으며, 3단계로 기초분석을 통해 가중치를 뽑아냈고 점수화를 거쳤습니다. 그렇게 최적의 입지를 도출했습니다.

우측 상단 이미지는 서울을 1km 단위 격자로 나눈 공간 데이터입니다. 이후 자주 등장할 '격자 기반 공간 데이터'를 의미합니다.

데이터 수집 및 전처리

9p

데이터 수집 및 전처리 과정에 대해 설명드리겠습니다.

10p

모델 구축을 위해 서울열린데이터광장을 비롯해 API 크롤링, GitHub 등에서 다양한 데이터를 수집했습니다. 키즈카페 관련 데이터는 일반키즈카페 약 400개, 서울형 키즈카페 65개, 민간인증제 키즈카페 53건를 활용했습니다. 그리고 입지에 영향을 주는 위치 기반 정보 2만 개 정도, 3개년의 인구와 소득소비 데이터, 지리정보 데이터가 포함됩니다. 총 10만 건을 데이터를 활용했습니다.

참고로 서울형 민간 인증제 키즈카페는 서울시가 인증한 민간 시설입니다. 일반인이 운영하는 키즈카페이지만 할인된 가격에 이용이 가능합니다.

11p

저희는 일반데이터와 공간데이터 **투 트랙**으로 모델링을 진행하기 위해 데이터 전처리 또한 나눠서 시행했습니다. (클릭 2번-애니메이션) 이렇게 진행한 이유는 일반 데이터는 **기초분석과 가중치 수치화**를 위해, 공간 데이터는 **공간적 특성**을 반영한 입지 추천과 가중치 산출을 위해 필요했기 때문입니다.

그리고 모든 데이터들의 분석 단위는 행정동 코드입니다. 이후 이 코드를 기준으로 병합을 진행했습니다.

12p

먼저 **일반 데이터 전처리 과정**을 보여드리겠습니다. 일반 데이터는 기초 전처리 시행 후 api로 위경도 좌표와 행정동 이름을 수집했고, 각 데이터의 행정동에 매칭되는 행정동 코드를 매핑시켰습니다. 이렇게 코드가 심어진 모든 데이터들을 코드 기준로 최종 병합을 진행했습니다.

13p

이미지와 함께 보여드리자면, api를 기반으로 위경도 좌표를 수집한 다음에 행정동명 또한 수집을 했고

14p

이런 식으로 수집한 모든 데이터들에 행정동 코드를 삽입했습니다.

15p

저희가 수집한 데이터는 크게 시간 단위로 된 데이터 그리고 위치 기반으로 된 데이터 이렇게 두 개로 구분할 수 있었습니다. 따라서 1차적으로 시간 단위 데이터들끼리 하나로 병합을 시키고, 위치 기반 데이터들끼리 하나로 병합한 후에 이 두 개의 데이터를 **행정동 코드 기반**으로 최종병합시켰습니다.

16p

완성된 일반 데이터의 전처리 결과입니다. 보시는 것처럼 행정동 코드 단위로 정리되어 있으며, 위치 데이터는 각 행정동에서 몇개씩 있는지 알 수 있습니다.

17p

다음으로는 **격자 기반 공간 데이터 전처리**에 대하여 설명드리겠습니다.
먼저 1km의 격자를 생성한 후 행정동 경계를 기반으로 격자를 다시 나누는
오버레이 작업을 진행했습니다.

그 이후 위치 기반 데이터, 시간 단위 데이터와 인구 데이터를 격자에 매핑하였습니다. 각 격자별로 모든 정보를 종합할 수 있게 만든 것이라고 이해하시면 됩니다.

18p

이미지와 함께 공간 데이터 전처리 과정을 보여드리자면,

먼저 정사각형 기본 격자를 생성한 후에 행정경계에 맞춰 오버레이를 진행한 모습입니다.

이 과정을 통해 2427개의 격자를 확보하였고, 이후 모든 분석은 이 격자를 기준으로 집계 및 모델링이 이뤄졌습니다. 공간 단위를 명확히 설정함으로써 **정확도와 해석력**을 높일 수 있었습니다.

19p

그리고 만들어진 공간 격자 데이터에 다음 데이터들을 매핑시켰습니다.

위경도 좌표를 이용해 위치 기반 데이터들을 할당했고,

행정동 코드를 기준으로 소득소비 데이터를 할당했습니다.

정확한 분석을 위해 지역 특성을 나타내는 **지적편집도**도 활용했는데요. 이 지적편집도를 기반으로 각 구역을 행정동 단위로 매핑해서 **다각형 형태**의 공간 데이터를 만들었습니다.

마지막으로 지적편집도를 기반으로 (거주 비율과 주요 도심지 거리를 고려해서) **가중치 설정**해 인구데이터 분배했습니다. 보다 현실성 있는 인구 분포 추정을 하기 위함이었습니다.

20p

왼쪽은 위치 기반 데이터들을 매핑한 결과입니다. 일부 부분을 확대해보면, (슬라이드 다음 장)

21p

이런 식으로 대중교통은 분홍색, 일반키즈카페는 하늘색 점으로 나타나 위치 분포가 확인됩니다.

오른쪽은 소득소비 데이터를 매핑한 지도이며 진할 수록 소득이 높음을 의미합니다.

22p

그리고 가중치를 이용해 인구데이터를 매핑한 모습입니다. 두 그림 모두 붉어질 수록 인구가 많은 곳입니다. 왼쪽은 주거 지역을 중심으로 아동인구 데이터를 넣은 결과이고, 오른쪽 지도에서는 출퇴근 시간에 따른 유동인구의 변화 상태 등도 확인할 수 있습니다.

세번째 목차는 다음 발표자가 이어 발표하겠습니다.

23p

이어서 **기초분석 및 모델링 파트**를 설명드리겠습니다.

24p

저희는 총 9가지 기초분석을 수행했으며, 이 중 **정책적 활용이 가능한 4가지 분석 결과**를 선별하여 사용했습니다. 다양한 분석 기법을 통해 **일관되게 중요한 변수**를 도출하고, **정책 점수 산정의 신뢰성**을 높이고자 했습니다.

25p

첫 번째, **LASSO 회귀분석**입니다. 많은 변수 중에서 **정말 영향력이 큰 변수만 선택하는 기법**인데요, 그 결과 '소득'과 '유흥시설 수'가 중요 변수로 도출되었고, **두 변수 모두 음의 방향**을 가지는 것으로 나타났습니다.

26p

두 번째는 랜덤포레스트와 SHAP 분석입니다. 랜덤포레스트는 변수의 중요도를 알려주지만, **방향성은 파악하기 어렵기 때문에**, 변수별 기여 방향을 설명할 수 있는 **SHAP 분석을 함께 적용**했습니다.

그 결과, '총생활인구', '소득', '일반 키즈카페 수', '유아시설 수', '아동 인구' 정도가 **키즈카페 입지에 유의미한 변수**로 도출되었고, 특히 **'소득'은 음의 방향**, 즉 소득이 높을수록 키즈카페의 입지가 적합하지 않은 지역으로 나타났습니다.

27p

세 번째는 **포인트-바이시리얼 분석**입니다. 0과 1로 나타낸 키즈카페 유무와 소득, 초등학교 수 와 같은 연속형 변수 간의 상관관계를 파악한 분석인데요, 결론적으로 '초등학교 수'는 **양의 방향**, '유흥시설 수'는 **음의 방향**으로 유의미한 결과가 도출되었습니다.

28p

마지막으로, **정준상관분석(CCA)**을 통해 지역 특성과 키즈카페 설치 수 간의 구조적 상관관계를 분석한 결과, 다시 한 번 '소득'과 '유흥시설 수'가 **음의 방향**으로 도출되었습니다.

29p

기초분석 결과를 정리하면, **일반 키즈카페 수, 소득, 총생활인구**는 반복적으로 높은 중요도가 확인되어 **정책 점수 산정 시 핵심 변수**로 반영했습니다. 반면, '소득'과 '유흥시설 수'는 **회피 요인**으로, '초등학교 수', '키움센터 수', '아동 인구'는 **보조적인 판단 변수**로 활용했습니다.

30p

저희는 앞서 도출한 기초 분석 결과를 바탕으로 랜덤포레스트 중요도와 SHAP 방향성을 결합한 예측 모델을 만들어보았습니다.

구체적으로는, **랜덤포레스트로 도출한 변수의 중요도에, SHAP 분석을 통해 확인된 영향 방향(+ 또는 –)**을 곱해 **최종 가중치**를 산출했습니다.

예를 들어, '총생활인구' 변수는 중요도 0.1833, SHAP 방향성 +1이므로 **최종 가중치는 +0.1833**으로 계산됩니다.

모든 변수는 StandardScaler를 통해 정규화해 주었고, 각 행정동별로 '변수 값 × 가중치'를 모두 더해 예측 점수를 계산했습니다.

(이 점수는 **서울형 키즈카페 수요를 정량적으로 수치화한 결과**로, 향후 **정책 대상지 우선순위를 설정하는 데 활용 가능한 기준**이 될 수 있습니다.)

31p

하지만, 예측 모델만 사용할 경우 **이미 키즈카페가 많은 지역이 또 높은 확률로 나올 수 있는 한계**가 있습니다. 그래서 저희는 **정책적 우선순위를 수치화한 '정책 점수 모델'**을 함께 도입했습니다.

앞선 기초분석 결과를 참고해 '총생활인구(0.30)', '보육시설 수(0.25)' 등 주요 변수에 높은 가중치를 부여했고, **소득이 지나치게 높거나, 유흥시설 밀집 지역**, **기존에 키즈카페가 설치된 지역**은 감점 처리했습니다.

이 점수는 예측 모델과 함께 종합적으로 고려되어 뒤에서 설명할 **최종 입지 모델**의 기반이 되었습니다.

기초 분석 및 모델링

32p

저희는 본 프로젝트에서 두 가지 주요 모델링 접근법을 병행해 활용했습니다.

먼저, 좌측 모델은 KMeans와 DBSCAN을 활용하여 가중치를 산출하고 **다기준의사결정(MCDA)**을 수행하고, 이를 바탕으로 **최대 커버리지 모델(MCLP)**을 적용해 입지 후보를 선정했습니다. 여기서 KMeans와 DBScan, 즉 클러스터링 분석을 활용한 이유와 그 과정은 뒤에서 설명드리겠습니다.

반면, 우측 모델은 Random Forest와 XGBoost에 SHAP 분석을 결합해 공간 데이터의 가중치를 계산하고, 여기에 일반 데이터 가중치 평균값을 더해 최종 가중치를 도출한 다음, **앙상블 기반 확률형 모델**을 구축했습니다.

이렇게 도출된 MCDA 기반 점수와 앙상블 예측확률값을 결합하여 보다 정밀한 **최종 입지 추천 모델**을 완성하였습니다.

33p

먼저, MCDA(다기준 의사결정 분석)는 소득, 인구, 접근성 등 여러 요소에 **가중치를 부여**하여 **각 격자의 입지 적합도를 점수화**하는 방식입니다. 이어서, MCLP(최대 커버리지 모델)을 통해 제한된 수의 시설(예: 270개)**로, **가장 많은 수요를 커버할 수 있는 격자**들을 최적 위치로 도출했습니다.

(클릭 시 지도 등장) 보시는 화면처럼, 붉은색일수록 추천도가 높은 지역입니다.

동시에, 저희는 머신러닝 확률 기반의 모델링도 병행했습니다. 로지스틱, 랜덤포레스트, XGBoost 세 가지 모델을 결합한 **Voting 앙상블 모델**을 구성했고, 각 모델의 예측 확률을 평균 내는 **Soft Voting 방식**을 적용했습니다. (특히 XGBoost는 학습 과정에서 예측 오차를 반복적으로 보완하는 방식으로, 더 정교한 패턴 학습이 가능했습니다.)

이 모델은 각 격자에 대해 서울형 키즈카페가 존재할 가능성, 즉 **입지 가능성 자체를 수치화하여** 상위 N개의 격자를 보여줍니다.

다만, 기존 키즈카페의 입지가 최적이 아니라는 현실을 반영하여, 저희는 이를 **정책 점수 모델의 보조적인 지표로 활용**하고자 했습니다.

(여기부터는 다음 발표자가 이어서 설명드리겠습니다.)

34p

다음으로 MCDA와 MCLP 모델에서 활용한 클러스터링 결과물들에 대해 설명드리겠습니다. 화면에서 보시는 KMeans와 DBScan 부분입니다.

35p-36p

엘보우 차트와(피피티 넘기는 것 잊지 말기) 실루엣 분석은 클러스터의 개수를 결정하는 방법으로, 두 분석에서 적절한 클러스터 수는 모두 4개로 도출되었습니다.

37p

이제 이 결과를 바탕으로 군집분석을 수행합니다. KMeans는 중심 기반 클러스터링 방법입니다.

군집별 해석은 다음과 같은습니다. 녹색 군집은 시설이 이미 충분합니다. 주황색과 파란색 군집은 **신규 입지 추천에 유리**하며, 분홍색은 소득이 높지만 **공공성이 낮은 지역**으로 분류됩니다.

38p

다음으로는 DBSCAN입니다. 해당 방법론은 유사한 데이터는 서로 근접하게 분포시키며, 최적의 k값을 자동으로 결정합니다.
초록색은 일반 밀집지역, 주황색은 특이성을 지닌 개별지역으로 나타나고 있으며, 파란색은 **고밀도 특화지역으로 나타남과 동시에 공공시설 도입 효과가 클 것으로** 예상됩니다.

39p

군집분석 2가지를 통해 얻은 결과로 추출한 추천 행정동은 다음과 같습니다.

40p

이어 K-means와 DBSCAN을 통해 얻은 결과를 바탕으로, 정책기반 모델링이자 다기준 의사결정 모델인 MCDA점수에 근거하여 나온 추천 행정동 리스트는 다음과 같습니다. 이는 MCDA 단독 모델의 결과입니다.

42p + 43p

이번에는 앙상블 기반 확률형 모델 기초분석입니다. 공간 데이터 특성 상 비선형 모델이 적합하며, 랜덤포레스트와 XGBoost를 사용했습니다. 두 모델 모두 SHAP 분석과 병행하였고, 공통적으로 도출된 주요 변수 상위 5개는 총 생활인구, 소득, 아동인구, 대중교통, 유아시설이었습니다.

44p

비선형 2개 모델의 기초 분석 결과를 가지고 모델링을 위한 가중치를 도출합니다. 이 계산에서 중요도와 SHAP의 영향력은 7:3입니다. 이는 모델의 판단력과 해석력을 함께 고려한 것으로, 키즈카페 수의 부족으로 SHAP 값이 과소평가되어 가중치가 작아지는 문제를 보완하기 위한 조치입니다. 그리고, 랜덤포레스트와 XGBoost의 평균치를 공간 데이터의 가중치로 채택하였습니다.

45p

구체적인 수치는 다음과 같습니다. 음의 방향성을 보인 소득과 절대적 기피시설인 유흥시설에는 음의 값을 부여했습니다.

47p

이제 일반 데이터 가중치와 공간 데이터 가중치의 평균값을 산출합니다. 공간 데이터는 편향 여부 확인이 어려워, **일반 데이터와의 평균값으로 가중치를 보정**했습니다. 또한, 서울형 키즈카페 입지가 '편의성 중심'이라, 최적 입지의 불완전성이 있어, 이를 보완하기 위한 **정책 점수의 반영**이 필요한 점 또한 일반 데이터와의 평균값이 필요했던 이유입니다. 최종 가중치는 **정규화하여 MCDA 등 다른 모델에도 활용**할 수 있도록 정비했습니다.

49-50p

이제 최종 가중치를 적용해 앙상블 모델을 단독 학습한 결과입니다. 전체 예측 비율인 정확도와 전체 성능인 AUC는 높았지만, 실제 키즈카페 설치 수의 문제로 '없음'만 예측해도 성능이 높게 나오는 구조라, 설치 예측 중 실제 설치된 비율인 **정밀도**는 8%에 불과합니다. 하지만 실제 설치 중 설치 예측한 비율인 재현율은 60%로 일부 지역은 포착되고 있습니다. 이는

기존 키즈카페 수가 적고 입지가 불완전하다는 구조적 한계 때문이며, 추가적 보완이 필요함을 시사합니다.

다음은 이 모델을 바탕으로 예측한 행정동 순위와 시각화 지도입니다. 지도에서는 입지 점수가 낮을수록 파란색, 높을수록 붉은색으로 표시됩니다.

52p

저희의 최종 모델은 앙상블 기반 확률형 모델에 MCDA를 결합한 모델입니다. 따라서, 정책적 기준에 따른 점수와 데이터를 바탕으로 한 예측확률을 결합하는 것이 됩니다. 이를 결합할 때 수치를 정규화 하여 곱하는 방식과 평균을 내는 방식이 있는데, 교차 검증 결과, 정규화 곱의 성능이 더 좋은 것으로 나타났으므로, 해당 값을 추천 모델의 최종 값으로 채택하였습니다.

53p

최종 모델이 추천한 행정동명과 세부 격자의 ID는 다음과 같습니다. 동일 행정동명이 중복 추천된 것은 동일 행정동에 여러 격자가 있고, 각 격자의 점수가 동시에 높기 때문입니다.

54p

이를 시각화 하면 다음과 같고, 색상이 남색으로 진해질수록 입지에 적합한 지역입니다. 그리고 상위 5개의 지역을 마커로 표시해 잘 보이도록 해 보았습니다.

55p

저희는 **서울형 키즈카페 예약현황**을 활용해, 예측된 입지 점수가 실제 높은 예약률과 연결되는지를 확인했으며, 이는 현실적으로 접근 가능한 검토 방법이었습니다. 가능한 기간 동안 크롤링으로 데이터를 수집하고, 모델의 **정규화 점수**와 **평균 예약률** 간의 **스피어만 상관계수**를 계산한 결과, **0.3 수준의 양의 상관관계**가 나타났고, p**값은 0.0151**로 통계적으로 유의했습니다. 산점도에서도 약한 양의 비례관계가 시각적으로 확인됩니다. 물론 예약률은 입지 외에도 홍보, 운영 방식 등 다양한 요인의 영향을 받으며, 현장 예약은 수집이 어렵다는 한계도 있지만, 그럼에도 불구하고 **입지 예측 점수가 실제 수요를 어느 정도 설명해낸다는 점에서 유의미한 결과**라고 판단됩니다.

56p

다음은 두번째 최종모델인 MCLP입니다. MCLP는 정책의 목표치를 입력해 타겟팅 할 수 있어 저희의 모델과 가장 적합성이 있으나, 아쉽게도 세부 격자 ID와의 매핑에는 실패하여 행정동 단위의 추천만 할 수 있었습니다. 따라서, 마지막에 해당 결과를 첨부했으며, 최종모델에 나타난 행정동과 상당히 유사한 것을 보실 수 있습니다.

기대효과 및 추후계획

58~59p

이제 본 프로젝트의 기대효과와 추후 계획을 말씀드리겠습니다. 기대효과는 정책결정자, 사업자, 이용자 관점으로 나뉩니다. 먼저 **정책결정자**는 데이터 기반의 효율적인 입지 선정 가능성과 자원 배분의 효율성을 얻을 수 있으며 다른 공공시설과의 연계 가능성도 높아집니다. **사업자**는 예측 정보를 바탕으로 리스크를 줄이고, 성공적인 입지 전략 수립이 가능해지며, 민간 인증제를 통한 민간 협력 기반도 마련됩니다. **이용자**는 생활권 내 접근성이 높아져 육아 부담이 줄어들 수 있고, 지역 간 보육 격차 해소 효과도 기대할 수 있습니다.

60p

다음은 이러한 기대효과를 실현하기 위한 추후 계획입니다.

첫째, **인터랙티브 지도**를 통해 행정동별 점수, 인구, 주변 시설 정보를 클릭 한 번에 확인할 수 있도록 시각화할 예정입니다. 이를 통해 입지의 장단점을 직관적으로 비교할 수 있습니다. 둘째, **창업자 맞춤형 입지 체크리스트**와 **사업성 시뮬레이션**을 제공해, 수요 예측과 경쟁 분석 기반으로 입지 타당성을 판단할 수 있도록 합니다. 셋째, **운영 피드백 게시판**을 마련해 정책 개선을 위한 의견을 수렴하고, 이용자와 운영자의 목소리를 반영해 입지 모델을 지속적으로 고도화하여 지속가능한 정책 발전을 이룰 수 있도록 해보려 합니다.