안녕하세요 이어서 발표를 맡은 1조 김지우입니다.

이어서 군집화부터 발표를 재개하도록 하겠습니다.

저희는 또 다른 분석으로 군집화를 진행해보았습니다.

(슬라이드 넘김)

군집화를 진행한 목적은 머신러닝에 들어가기에 앞서 시니어로 필터링한 데이터를 군집화하여 군집 별로 특성을 파악하기 위해 군집화를 해보았습니다.

저희는 K-Prototype 이라는 군집화 알고리즘을 선택했는데요, 아마 수업시간에 배우지 않아서 해당 알고리즘에 대해 모르시는 분들이 있으실 거라고 생각합니다.

짧게 설명하자면, 기존에 수업시간에 배웠던 K-Means 알고리즘은 수치형 데이터일 경우에만 사용가능하고, 또 K-Modes라는 알고리즘은 범주형 데이터일 경우에만 사용 가능합니다. 저희가 메인으로 쓴 K-Prototypes 라는 알고리즘은 수치형 데이터와 범주형 데이터가 섞여있을 때 사용 가능한 알고리즘으로, 저희는 위의 3가지 방법을 다 해 본 결과, 가장 군집화 결과가 좋았던 K-Prototypes 군집화 모델에 대해서 말씀드리려고 합니다.

(슬라이드 넘김)

먼저 K-Prototypes 모델에 적용할 최적 군집의 개수를 찾기 위해 엘보우 커브를 시각화 해 보았습니다. 그 결과, K가 2,3,4 일 때 군집화가 효과적일 것이라는 결과를 도출할 수 있었습니다.

(슬라이드 넘김)

다음은 K가 2,3,4일 때 실루엣 분석을 한 결과를 시각화 해보았습니다. 실루엣 분석 결과, 군집의 개수가 2개일 때 실루엣 스코어가 0.828으로 가장 높고, 군집의 개수가 4개일 때 실루엣 스코어가 0.592로 가장 낮지만, 개별 군집의 실루엣 계수 평균값과 전체 실루엣 계수의 평균값의 편차가 크지 않고, 즉 빨간 점선을 관통할 경우에 군집화가 잘 되어있다고 판단할 수 있으므로, 두 가지 기준을 미루어봤을 때 그 중간값인 군집의 개수가 3인 경우를 채택하여 군집화를 진행하였습니다.

(슬라이드 넘김)

다음은 군집의 개수가 3개일 때의 군집화 결과를 말씀드리겠습니다.

각 군집별 관람횟수를 기준으로 최소, 최대, 평균값을 도출해봤습니다.

그 결과 1번 군집은 관람횟수의 최솟값이 0, 최댓값이 44,

2번 군집은 관람횟수의 최솟값이 45, 최댓값이 173,

0번 군집은 관람횟수의 최솟값이 178, 최댓값이 672로,

관람횟수가 적은 그룹, 보통인 그룹, 많은 그룹으로 군집화가 잘 되었다는 결론을 도출하였습니다.

(슬라이드 넘김)

이렇게 군집화가 완료되고, 이제 군집 별 특성을 파악해보았는데요, 그 중 가장 유의미한 차이가 있었던 결과를 시각화해보았습니다. 참여횟수가 가장 적은 그룹이었던 1번 군집의 경우, 가장 많이 지출한 문화예술분야는 ‘영화’가 가장 높았고, 반면 참여횟수가 가장 많은 그룹이었던 0번 군집의 경우 가장 많이 지출한 문화예술분야가 ‘영화’에 비해 ‘연극 및 뮤지컬’이 월등히 높은 것을 파악할 수 있었습니다.

영화에 비해 상대적으로 연극 및 뮤지컬이 티켓 가격도 비싸고, 접근성 부분에서도 차이가 있기 때문일 것이라고 보았고요, 그렇기 때문에 참여횟수가 많은 그룹은 적은 그룹에 비해 문화예술에 지출하는 비용이 더 높을 것이라는 가설을 세울 수 있었습니다. 후에 소비분석 파트에서 해당 가설이 맞는지, 자세히 다뤄보도록 하겠습니다.

(슬라이드 넘김)

다음으로는 머신러닝에 대해 다뤄본 결과를 말씀드리도록 하겠습니다.

(슬라이드 넘김)

저희 조는 4명이 각자 하나의 머신러닝 기법을 맡아서 전처리부터 모델 생성, 하이퍼 파라미터 튜닝 및 결과 도출까지 해보았습니다.

4개의 알고리즘은 다중 회귀, 로지스틱 회귀, lightgbm, 랜덤포레스트 이렇게 총 회귀 알고리즘 1가지, 분류 알고리즘 3가지를 해보았습니다.

(슬라이드 넘김)

저희가 머신러닝에 사용한 Feature와 Target 열은 이러합니다.

타겟 데이터로는 문화예술행사 관람횟수의 총합을 사용하였고,

피쳐데이터로는 (성별, 학력, 혼인상태, 종사상지위, 가구소득, 장애등록여부, 문화예술행사 선택기준, 그리고 관람걸림돌) 이렇게 사용하였습니다.

(슬라이드 넘김)

처음으로는 다중회귀 결과에 대해서 말씀드리겠습니다. 저희는 다중회귀중 릿지와 라쏘를 사용하여 머신러닝을 진행해보았습니다.

(슬라이드 넘김)

모델을 생성하기 전, 먼저 시니어와 수도권에 해당하는 데이터를 필터링 해준 이후 타겟 데이터를 로그 변환 해주었습니다. 오른쪽에 보시는 데이터프레임은 로그 변환 이후의 데이터입니다.

(슬라이드 넘김)

그 다음, 범주형 feature 데이터에 대해서 원-핫 인코딩을 진행해주었습니다.

(슬라이드 넘김)

그 다음 다중회귀 모델을 만들어 학습 및 예측을 진행해본 결과, RMSE는 0.969, R제곱 값으로는 0.222의 결과가 나왔습니다. RMSE는 숫자가 작을수록 성능이 좋고, 반대로 R제곱은 높을수록 좋은 지표인데요. 저희 만든 회귀 모델은 RMSE는 높게 나오고 R제곱은 낮게 나와서 이에 한계점을 느끼고, 분류 분석을 진행해보기로 했습니다.

(슬라이드 넘김)

그 다음으로는 분류 분석을 진행해보았습니다.

(슬라이드 넘김)

저희가 분류 분석에 사용한 알고리즘은 총 3가지인데요, 이를 위해 타겟 데이터를 pd.cut 함수를 이용하여 카테고리화 해주었습니다. 범주를 나눈 기준에 대해 설명해드리겠습니다.

저희는 관람횟수가 적은 그룹, 평균 정도인 그룹, 많은 그룹 이렇게 3가지 범주로 나누었는데요, 관람횟수의 4분위수와 평균값을 구해서, 하위 관람 그룹은 최솟값부터 제 1사분위수까지, 평균 관람그룹은 제 1사분위수에서 평균까지, 상위 관람 그룹은 평균값부터 최댓값까지 이렇게 나누었습니다.

(슬라이드 넘김)

첫번째로 로지스틱 회귀 결과에 대해 말씀드리겠습니다.

(슬라이드 넘김)

모델을 생성하기 전에 연속형 feature에 대해 스케일링과 로그 변환을 해주었습니다.

그리고 엘라스틱넷으로 로지스틱 회귀 모델을 만들고 하이퍼 파라미터 튜닝까지 완료한 모델의 정확도가 0.4662로 나왔습니다.

(슬라이드 넘김)

그리고 타겟 값을 예측하는 확률 별 회귀계수를 구해봤는데요,

where\_cnt, 문화예술행사를 어디서 많이보는지에 대한 변수가 모든 타겟에 걸쳐 예측 확률이 가장 높게 나온 것을 확인 할 수 있었습니다.

(슬라이드 넘김)

그 다음으로는 LightGBM 을 해보았습니다.

(슬라이드 넘김)

저희가 LightGBM을 시도해본 이유가 이 알고리즘은 범주형 변수가 많을 때 강한 성능을 보여주는 알고리즘이어서 시도를 해봤습니다. 그 결과 정확도가 0.5285가 나왔습니다. 하지만 해당 알고리즘의 경우 데이터가 10000개 이하일 경우 과적합이 발생할 확률이 높은 알고리즘이라는 한계점이 있을 수 있겠습니다.

(슬라이드 넘김)

그 다음은 해당 LightGBM 모델을 하이퍼 파라미터 튜닝을 진행한 과정을 담아보았습니다.

(슬라이드 넘김)

그리고 LightGBM 모델의 피쳐 중요도를 시각화 해보았습니다. 피쳐 중요도 1위는 로지스틱 회귀와 동일하게 where\_cnt로 나왔고, 2순위로는 문화예술행사에 향후 지출하고 싶은 분야, 3순위로는 문화예술행사 참여시 어려움이었습니다.

(슬라이드 넘김)

이후 저희는 랜덤포레스트 분류 분석을 진행하였습니다.

(슬라이드 넘김)

랜덤포레스트 모델 생성 전, 비용에 관련된 피쳐데이터 두 가지, (문화예술에 가장 많이 지출한 분야, 그리고 향후 지출할 의향이 있는 분야) 피쳐에 대해 범주를 축소해주었습니다.

그 후 모델 생성을 해보니, 정확도가 0.5404로 지금까지 진행한 분류 분석 알고리즘 중 가장 높은 정확도를 보였습니다.

(슬라이드 넘김)

그리고 성능 향상을 위해 하이퍼 파라미터 튜닝을 진행해주었고, 이 후 최적의 estimator를 통해 다시 학습 및 예측을 진행한 결과 모델의 정확도가 0.5466으로 소폭 향상되었습니다.

(슬라이드 넘김)

이 다음 도출한 모델의 피쳐 중요도를 시각화해보았는데요, 가장 중요도를 보이는 피쳐는 소득이었고, 그 다음은 선택 기준, 관람걸림돌 순이었습니다.

(슬라이드 넘김)

이 다음으로는 저희가 진행한 머신러닝의 결과를 표로 정리해보았습니다.

첫번째로 진행했던 다중회귀분석의 경우 저희가 예측하고 싶었던 결과가 관람횟수의 연속형 수치값이었기 때문에 회귀분석을 진행하였지만, 이후 관람횟수를 범주화하여 범주화된 결과를 얻고 싶었기 때문에 분류분석으로 방향을 바꿔서 진행해주었습니다. 그 결과 왼쪽 세 가지 분류분석의 정확도를 비교하면 랜덤포레스트가 가장 높았기 때문에 이후 소비 분석으로 연계하기 전 최종 모델을 랜덤포레스트로 채택하였습니다.

(슬라이드 넘김)

저희는 지금까지 시니어들의 특성에 따라 문화예술행사의 참여도를 예측하였는데요, 이 예측 결과를 가지고 다른 데이터셋과 연계하여 서울특별시 내 시니어의 소비 패턴 분석을 진행해주었습니다.

(슬라이드 넘김)

소비 분석 데이터에서도, 기존에 참여도를 나눴던 범주 기준과 동일한 기준을 적용하여 소비 그룹을 카테고리화해주었습니다.

그 결과 하위 소비 그룹은 연간 평균 12000원을 소비한다는 결과를 얻을 수 있었고,

평균 소비 그룹은 연간 평균 45000원을, 상위 소비 그룹은 연간 평균 23만원을 소비한다는 결과를 얻을 수 있었습니다.

(슬라이드 넘김)

다음으로는 그룹별 총 소비금액을 시각화해보았습니다.

상위 그룹은 총 소비금액의 71%를 소비하는 것으로

하위 그룹은 총 소비금액의 5%를 소비하는 것으로

두 그룹간의 소비 금액의 편차가 심한 것을 알 수 있었습니다.

(슬라이드 넘김)

저희는 기존 머신러닝 모델의 분류 기준으로 가장 영향을 많이 끼쳤던 피쳐인 소득을 기준으로 기존 문화예술행사 데이터셋에서 참여도를 기준으로 소득 분포를 시각화해보았습니다.

그래프의 X축 범주 코드 기준은

1번은 100만원 미만, 2번은 100-200만원, 그리고 7번은 600만원 이상입니다.

그 결과 참여도가 적은 그룹은 하위 소득 구간에 분포가 더 많고, 상위 소득구간으로 갈수록 빈도수가 적어지는 반면,

참여도가 높은 그룹은 하위 소득 구간에 분포가 적고, 상위 소득구간으로 갈수록 빈도가 높아지는 경향을 보이는 것을 확인했습니다.

그렇다면 참여도 별 소득구간에는 차이가 있다는 것을 확인하였고,

그럼 소득구간에 따른 문화 소비 비용에도 차이가 있는지 다음 슬라이드에서 연계하여 알아보도록 하겠습니다.

(슬라이드 넘김)

저소득층의 연간 문화활동 소비 비용 평균 금액은 42000원,

고소득층의 연간 문화활동 소비 비용 평균 금액은 245000원인 것을 확인하였습니다.

그리고 저소득층이 가장 많이 소비하는 분야인 전통예술공연 분야의 소비 금액과

고소득층이 가장 적게 소비하는 분야인 전시회 관람 분야의 소비 금액이 거의 비슷한 것을 볼 수 있습니다.

이뿐만 아니라 저소득층의 연간 문화활동 소비 비용은 전체적으로 고소득층의 비용보다 훨씬 낮은 것을 확인할 수 있습니다.

그만큼 소득간 문화예술활동의 격차가 크다는 것을 의미하는데요,

그렇다면 세부적으로 어떤 문화예술분야에서 차이가 큰 건지를 다음 슬라이드에서 알아보았습니다.

(슬라이드 넘김)

비교를 하기 위해서 비율로 값을 구한 그래프가 왼쪽에 나와 있는 그래프이고요,

그 둘의 차를 구한 그래프가 오른쪽에 나와있는 그래프입니다.

예를 들면, 왼쪽의 그래프의 파란색 상자를 보시면 저소득층에서는 문화예술분야 중 영화에 소비하는 경향이 큰 반면,

상대적으로 고소득층에서는 영화에 소비하는 경향이 다른 분야에 비해 적은 것을 확인할 수 있습니다.

이 오른쪽 그래프, 즉 소득에 따라 향유하는 문화예술활동 분야의 차이가 나와있는 그래프에 대해 조금 더 자세히 봐보도록 하겠습니다.

(슬라이드 넘김)

그래프를 보면, 고소득층이 저소득층에 비해 가장 많이 향유하는 문화예술활동의 차이가 큰 분야는 전시회 관람과 음악 및 무용 발표회 관람임을 확인할 수 있고,

반대로 저소득층이 고소득층에 비해 가장 많이 향유하는 문화예술활동은 영화임을 확인할 수 있었습니다.

(슬라이드 넘김)

저희는 지금까지 분석을 통해 내린 결론들로 저소득층의 문화 소외 현상을 해소하고 참여도를 고취시킬 방안을 제시하려합니다.

(슬라이드 넘김)

먼저 저소득층이 고소득층에 비해 상대적으로 소외된, 전시회와 음악 및 무용 발표회 등에 참여도를 높이기 위해 해당 문화예술분야에 더욱 적극적으로 복지 제도를 마련하고, 이를 해소하는 방안으로는 기존 국민문화예술활동 조사 데이터의 관람 걸림돌 설문조사 부분을 분석한 결과,

관람걸림돌이 비용, 시간, 관심도의 순으로 집계된 것을 확인했습니다.

따라서 첫번째로는 해당 분야에 대해 비용적인 부분을 고려하되, 3번으로 집계된 관심도가 적다는 부분도 고려하여 해당 분야에 관련한 정보를 소외계층에 더 많이 제공할 수 있는 플랫폼을 마련하는 등의 정보 제공을 통한 관심도 고취에 관한 부분도 정책적으로 고려를 해야한다고 분석하였습니다.