[붙임 1] 공통 양식(참가 신청서)

# 「2023년 제11회 문화데이터 활용 경진대회」참가신청서

* 해당란에 ☑ 표시

| | | | |
|---|---|---|---|
| 공모 분야 | ☐ 제품·서비스 개발 | ☐ 아이디어 기획 | ☑ 데이터 분석 |
| 참가 구분 | ☐ 개인 | ☑ 팀(기업) | |
| **창업 구분** | ☐ 창업 (사업자등록번호 : , 법인등록일 : 년 월 일) | | |
| | ☐ 창업예정 | ☑ 해당없음 | |
| 팀 명 | **김家네** | | |
| 제품·서비스 명<br>(분석 주제명) | 서울특별시 내 시니어들의 특성에 따른 문화생활 참여 예측 및 소비 분석 | | |
| 제품·서비스<br>개요<br>(분석 개요) | 시니어들의 특성에 따른 문화 생활 참여도 예측 및 중요 요인에 따른 소비분석 | | |
| 활용데이터<br>분야<br>※복수체크가능 | ☑ 문화예술<br>☐ 문화유산<br>☑ 문화산업<br>☐ 관광<br>☐ 체육 | ☐ 문화홍보<br>☑ 정책지원<br>☐ 도서<br>☐ 미디어·콘텐츠<br>☐ 기타 ( ) | |

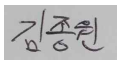
| 활용데이터 정보 ※복수기재가능 | 출처 | 제공기관명 | 데이터명 |
|---|---|---|---|
| | MDIS | 문화체육관광부 | 국민문화예술활동조사 |
| | 서울열린데이터광장 | 서울특별시 | 서울서베이 도시정책지표조사 정보 |
| | KOSIS | 통계청 | 가구주 연령별 가구당 월평균 가계수지 (전국,1인이상) |
| | 문화 빅데이터 플랫폼 | 컨슈머인사이트 | 여가문화 유형별 관심도 |

| 참가자 정보 | 성 명 | 소 속 | 연락처 | 이메일 |
|---|---|---|---|---|
| | 김종원 | X | 010-5548-1580 | rlawhddnjs6629@gmail.com |

| | | | | |
|---|---|---|---|---|
| | 김지우 | X | 010-5237-5315 | harutency@naver.com |
| | 김연진 | X | 010-5052-2968 | kyeonjin100@gmail.com |
| | 김예리 | X | 010-9457-2872 | ccuiri@naver.com |
| 이전 수혜 이력 및 입상 실적 | 년도 | 내용 | | |
| | | | | |
| | | | | |

본인(팀)은 '2023년 제11회 문화데이터 활용 경진대회' 참가와 관련하여 제출한 사항에 허위가 없으며, 유의사항을 숙지하고 진행에 필요한 사항에 성실히 응할 것을 동의합니다.

2023년 7 월 12일

신청인(대표자) 김종원 

[붙임 2-2] 데이터 분석 분야 양식

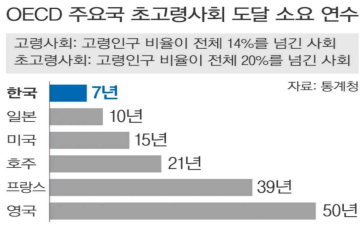
# 「2023년 제11회 문화데이터 활용 경진대회」분석보고서

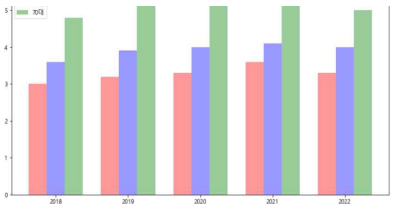
| 분석 프로그램 | ☑ Python | ☑ R | ☐ Tableau | ☐ 기타 |
|---|---|---|---|---|

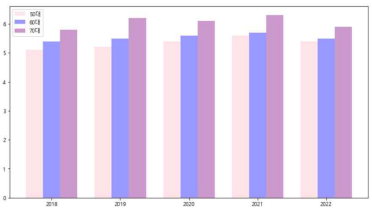
## 1) 분석 주제

서울특별시 내 시니어들의 특성에 따른 문화생활 참여 예측 및 소비 분석

## 2) 분석의 배경 및 목적





조사 결과에 따르면 한국은 타 선진국과 비교했을 때 가장 빠르게 고령화가 진행되어 2025년에는 초고령사회에 진입할 것으로 예상됩니다. 또한, **고령 인구 구매력의 증가율도 높아 고령 사회로 진입하는 과정에서 소비 패턴의 변화가 예상됩니다.** 이에 따라 저희는 서울특별시 내 시니어들의 특성에 따라 문화생활 참여도를 예측하고, 이를 응용하여 소비분석을 하고자 합니다.

<사전 분석 출처>
출처[1]: 2022 고령자 통계 [통계청]
출처[2]: 가구주 연령별 가구당 월평균 가계수지
출처[3]: 여가문화 유형별 관심도 [컨슈머 인사이트]

## 3) 활용 데이터 선정

### 데이터 목록

1. 문화체육관광부 - 국민문화예술활동조사
2. 서울특별시 - 서울서베이 도시정책지표조사 정보

### 선정이유

국민문화예술활동조사 데이터에서 중점적으로 사용하는 내용은 개인의 기본적인 정보와 개인의 문화활동 등에 대한 실태 파악 설문 결과이고, 도시정책지표조사 데이터에서는 개인의 기본적인 정보와 문화예술분야별 소비 금액에 대한 정보를 사용하고자 선정하게 되었습니다.

## 4) 분석 내용 및 결과

**데이터 전처리 방법**

- 데이터 타입 확인
- 분석에 필요한 컬럼 추출, 순서 설정, 컬럼명 변경
- 각 연도별 범주형 변수에 대해 설문 기준 통일
- 결측치 제거 및 대치
    - 서울시 문화활동 연간 평균 비용 통계 데이터에서 모든 문화활동 소비 금액이 빈 값일 경우, 문화활동에 참여하지 않았다고 간주하고 데이터 삭제
- 파생변수 추가
    - target 데이터 도출을 위해 dv_cnt와 mv_cnt를 더하여 view_cnt를 생성
- 2018, 2020년 데이터 병합 (2022년 데이터는 따로 진행)
- 머신러닝 예측 모델 생성 전 데이터 추가 전처리 진행
    - 필터링 및 범주 축소
    - 원핫인코딩, 스케일링, 로그변환

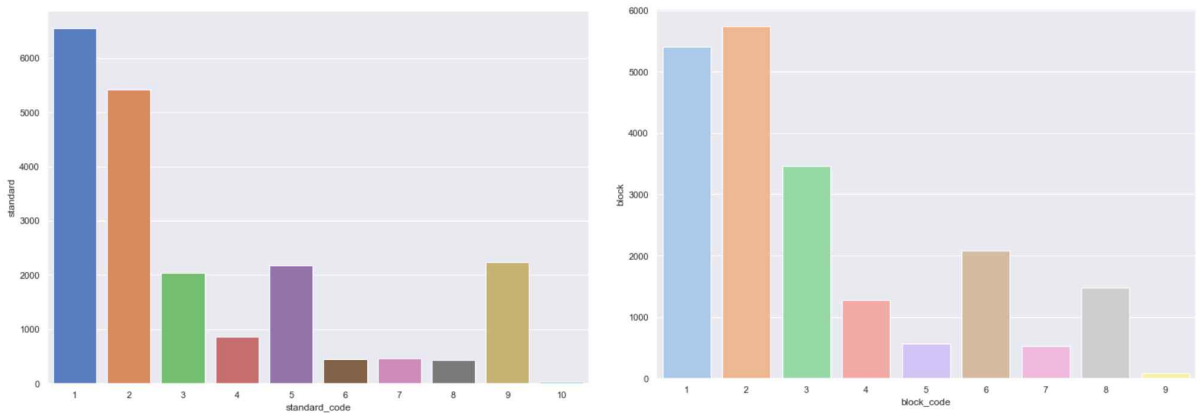
**사용된 변수**

| | |
|---|---|
| 성별 **sex** | 문화예술관련 많이 지출하는 항목 1순위 **expense** |
| 학력 **education** | 문화예술 관련 향후 지출을 늘리고 싶은 항목 1순위 **expense_f** |
| 혼인상태 **married** | 문화행사참여 시 가장 큰 어려움 **difficulty** |
| 종사상지위 **work** | 1년 이내 문화공간에서 개최하는 문화행사 참여 의향 **intention** |
| 가구소득 **income** | 역사문화 유적지 방문경험여부 **history** |
| 장애등록여부 **disabled** | 1년 이내 역사 문화유적지 방문 의향 **history_f** |
| 문화예술행사 선택기준 **standard** | 축제(거리 축제 포함) 방문경험 여부 **festival** |
| 문화예술행사 관람걸림돌 **block** | 1년 이내 축제 참여 의향 **festival_f** |
| 문화예술활동 공간 이용 횟수 총합 **where_cnt** | 문화예술활동 문화 행사 참석 횟수 총합 **attend_cnt** |
| 문화 관련 자원 봉사활동 참여 횟수 **volunteer_cnt** | 역사 문화 유적지 방문 경험 횟수 **history_cnt** |
| 축제(거리 축제 포함) 방문 경험 횟수 **festival_cnt** | 문화예술행사 참여 횟수 총합 **view_cnt** |
| 문화예술행사 직접관람 횟수 총합 **dv_cnt** | 매체를 이용한 문화예술행사 관람 횟수 총합 **mv_cnt** |

## 활용 알고리즘 설명

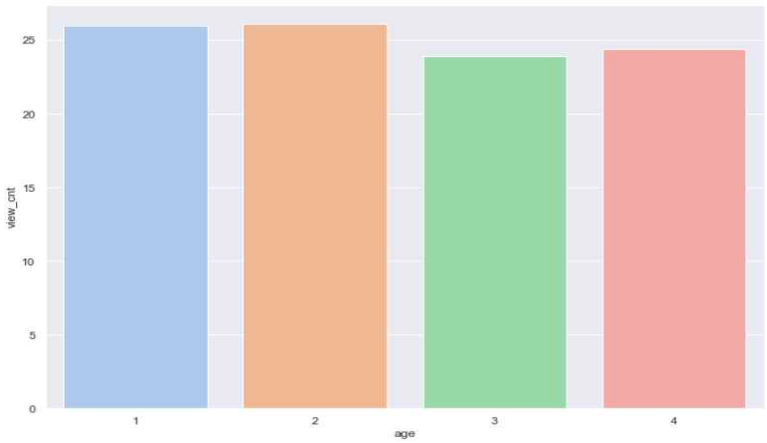
### 1. 기초 통계 분석

#### a. 문화예술행사 선택 기준 순위 / 문화예술행사 관람 걸림돌 순위 분석




- 위 막대그래프는 각각 문화예술행사 선택 기준 순위, 문화예술행사 관람 걸림돌 순위 그래프입니다.
- 왼쪽 그래프를 분석한 결과, **문화예술행사를 선택하는 기준은 '컨텐츠 및 퀄리티', '비용' 순으로 높았습니다**.
- 오른쪽 그래프의 경우, **문화예술행사 관람에 걸림돌이 되는 것은 '시간', '비용', '관심도 부족'의 순으로 높았습니다**.

#### b. 분산분석(ANOVA Test)



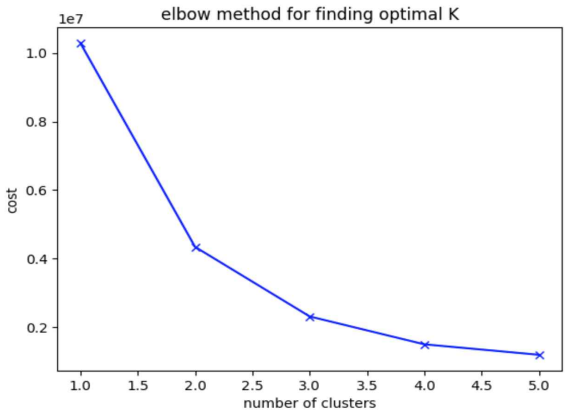

- 3개 이상 다수의 범주를 갖는 각 집단의 평균 값을 비교할 때 사용하는 가설검정 방법입니다.
- **연령층에 따라 문화예술행사 관람 횟수에 통계적으로 유의미한 변화가 있는지 분석했습니다.**
- 위 그래프는 연령대별 문화예술행사 참여 횟수를 나타낸 그래프입니다. 그래프로 봤을 때는 연령대 별로 큰 차이가 없어 보입니다. 이를 통계적 관점에서 차이의 유무를 알기 위해 ANOVA

Test를 진행했습니다.

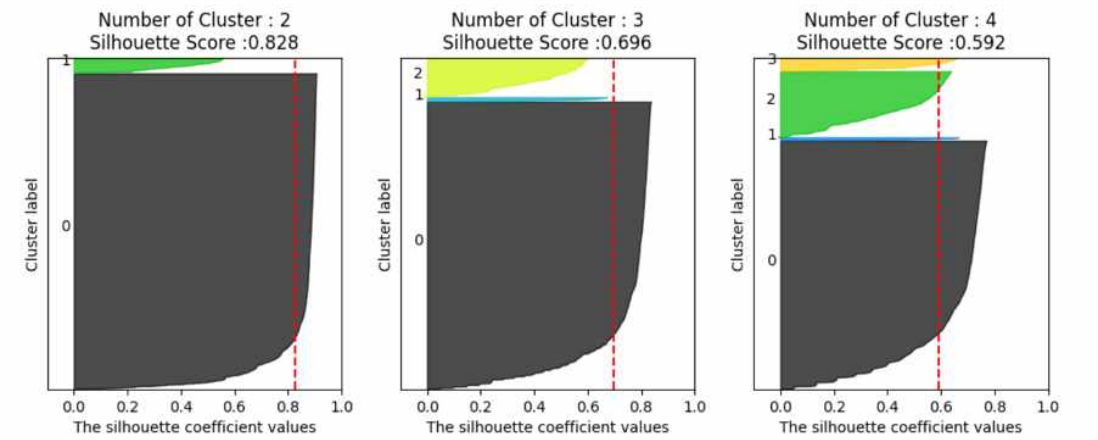
- p-value ≒ 0.0025 로 유의수준 0.05에서 유의합니다. 따라서 **연령층에 따라 문화예술행사 관람 횟수 평균의 차이가 있다고 볼 수 있습니다.**
- 특히, 5-60대의 사람들은 1~20대 및 3~40대인 사람들과 문화예술행사 관람 횟수에 유의미한 차이가 존재합니다. **즉, 5~60대인 사람들이 문화예술행사 관람 횟수가 다른 연령대에 비해 낮습니다.**

## 2. 머신러닝

### a. 비지도학습 - 군집화 (Clustering)




- 주어진 데이터 집합을 유사한 데이터들의 그룹으로 나누는 기법입니다.
- K-Modes, K-Means, K-Prototypes 세 가지 알고리즘 모두 적용해본 결과, **가장 결과가 좋은 K-Prototypes 모델을 채택**했습니다.
- 위는 최적의 군집 개수를 찾기 위해 엘보우 커브를 시각화한 그래프입니다. 그 결과, 군집이 2 ~ 4개인 경우 군집화가 효과적으로 이루어질 것이라는 결론을 도출했습니다.



```
# 0번 군집의 문화예술행사 관람횟수
CS5[CS5['cluster_3']==0][['view_cnt']].min()
CS5[CS5['cluster_3']==0][['view_cnt']].max()
CS5[CS5['cluster_3']==0][['view_cnt']].mean()
```

```
view_cnt    178
dtype: int64

view_cnt    672
dtype: int64

view_cnt    275.39759
dtype: float64
```

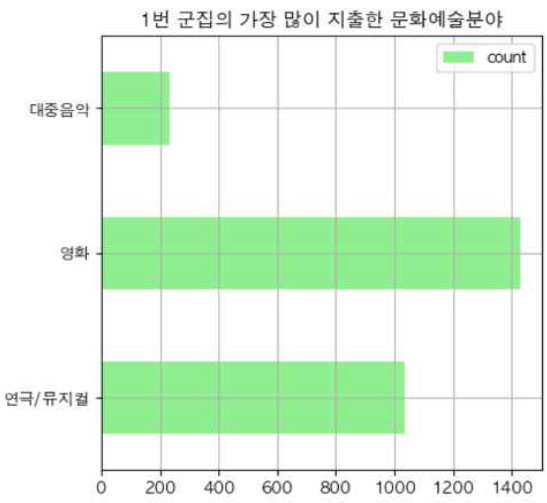
```
# 2번 군집의 문화예술행사 관람횟수
CS5[CS5['cluster_3']==2][['view_cnt']].min()
CS5[CS5['cluster_3']==2][['view_cnt']].max()
CS5[CS5['cluster_3']==2][['view_cnt']].mean()
```

```
view_cnt    45
dtype: int64

view_cnt    173
dtype: int64

view_cnt    75.414449
dtype: float64
```

```
# 1번 군집의 문화예술행사 관람횟수
CS5[CS5['cluster_3']==1][['view_cnt']].min()
CS5[CS5['cluster_3']==1][['view_cnt']].max()
CS5[CS5['cluster_3']==1][['view_cnt']].mean()
```

```
view_cnt    0
dtype: int64

view_cnt    44
dtype: int64

view_cnt    13.43228
dtype: float64
```

- 더 자세히 알아보기 위해 실루엣 분석 결과를 시각화하고, 군집화된 결과를 확인했습니다.
- 실루엣 스코어가 1에 가까우면서, 개별 군집의 실루엣 계수 평균값과 전체 실루엣 계수의 평균값의 편차가 크지 않아야 좋은 군집화라고 할 수 있습니다.
- 따라서 군집 개수가 3개인 경우 제일 적절한 군집화가 이루어질 것이라고 분석하였습니다.
- 군집의 개수가 3개일 때의 각 군집 별 문화예술행사 관람횟수의 기초통계량을 확인해본 결과, **문화예술행사 관람 횟수가 적은 그룹, 보통인 그룹, 많은 그룹으로 군집화가 잘 되었다는 결론을 내렸습니다**.



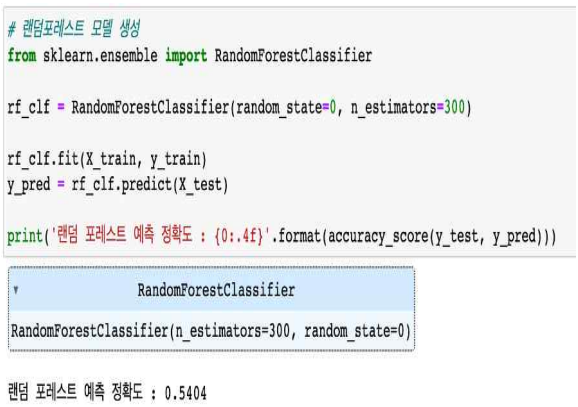


- 각 군집 별 특성을 파악한 결과, 문화예술활동 참여 횟수가 적은 그룹은 영화에 많이 지출하고, 많은 그룹은 연극/뮤지컬에 많이 지출하는 것을 확인할 수 있습니다.
- 영화에 비해 연극 및 뮤지컬이 상대적으로 티켓 가격도 비싸고, 접근성 부분에서도 차이가 있기 때문일 것이라고 보았습니다. 따라서 **참여횟수가 많은 그룹은 적은 그룹에 비해 문화예술에 지출하는 비용 또한 더 높을 것이라는 가설을 세울 수 있었습니다**. 후에 소비분석 파트에서 해당 가설이 맞는지, 자세히 다뤄보도록 하겠습니다.

### b. 지도학습 - 분류(Random Forest, LightGBM, Logistic Regression)

- 랜덤 포레스트, LightGBM, 로지스틱 회귀분석을 이용해 **문화생활 참여 정도를 3가지 범주로 나누어 분류하고 예측하였습니다**.
- 머신러닝 모델링 전, 데이터 전처리를 진행했습니다. 필터링을 통해 5-60대의 시니어, 수도권에 해당하는 지역의 사람들로만 데이터를 정제해주었습니다.
- 로지스틱 회귀분석의 경우, 범주형 feature 데이터에 대해 원-핫 인코딩을, 연속형 feature에 대한 스케일링 작업과, target 데이터의 분포가 한쪽으로 쏠려 있는 것을 감안하여 log 변환을 적용하여주었습니다.

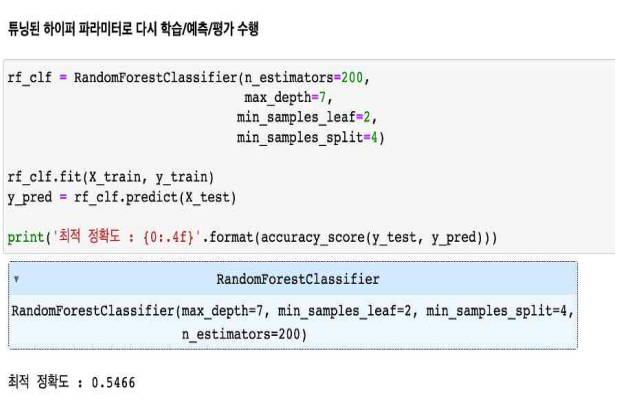
- 문화예술활동 관람 횟수 데이터인 target 데이터는 본래 연속형 데이터였지만, 카테고리화 하여 범주형 데이터로 바꿔주었습니다.
- **[상, 중, 하] 총 3개의 카테고리로 관람횟수(참여도) 범주 기준을 분리하였습니다**. 해당 기준은 아래와 같습니다. 또한 모든 모델에 동일한 기준을 적용하였습니다.
    - 하위 관람 그룹(target = 1) : 최솟값 ~ 제 1 사분위수
    - 평균 관람 그룹(target = 2) : 제 1 사분위수 ~ 평균
    - 상위 관람 그룹(target = 3) : 평균 ~ 최댓값



```
# 랜덤포레스트 모델 생성
from sklearn.ensemble import RandomForestClassifier

rf_clf = RandomForestClassifier(random_state=0, n_estimators=300)

rf_clf.fit(X_train, y_train)
y_pred = rf_clf.predict(X_test)

print('랜덤 포레스트 예측 정확도 : {0:.4f}'.format(accuracy_score(y_test, y_pred)))
```

```
RandomForestClassifier(n_estimators=300, random_state=0)
```

```
랜덤 포레스트 예측 정확도 : 0.5404
```



튜닝된 하이퍼 파라미터로 다시 학습/예측/평가 수행

```
rf_clf = RandomForestClassifier(n_estimators=200,
                                max_depth=7,
                                min_samples_leaf=2,
                                min_samples_split=4)

rf_clf.fit(X_train, y_train)
y_pred = rf_clf.predict(X_test)

print('최적 정확도 : {0:.4f}'.format(accuracy_score(y_test, y_pred)))
```

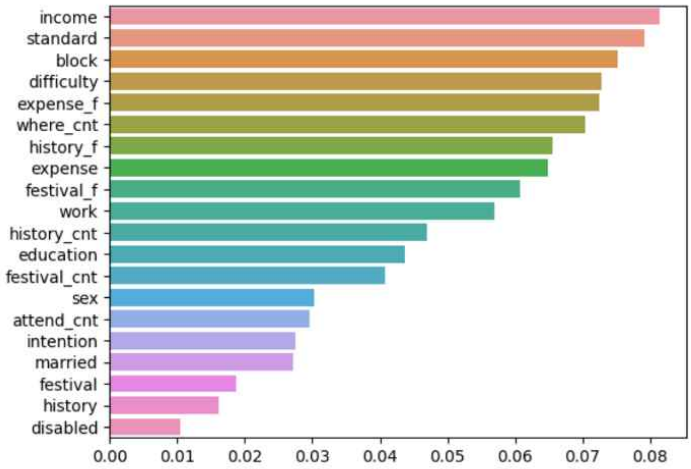
```
RandomForestClassifier(max_depth=7, min_samples_leaf=2, min_samples_split=4,
                       n_estimators=200)
```

```
최적 정확도 : 0.5466
```

- 각 AI 알고리즘에 대해 모델을 생성하고, 하이퍼 파라미터 튜닝을 통해 모델의 성능을 높여주었습니다.

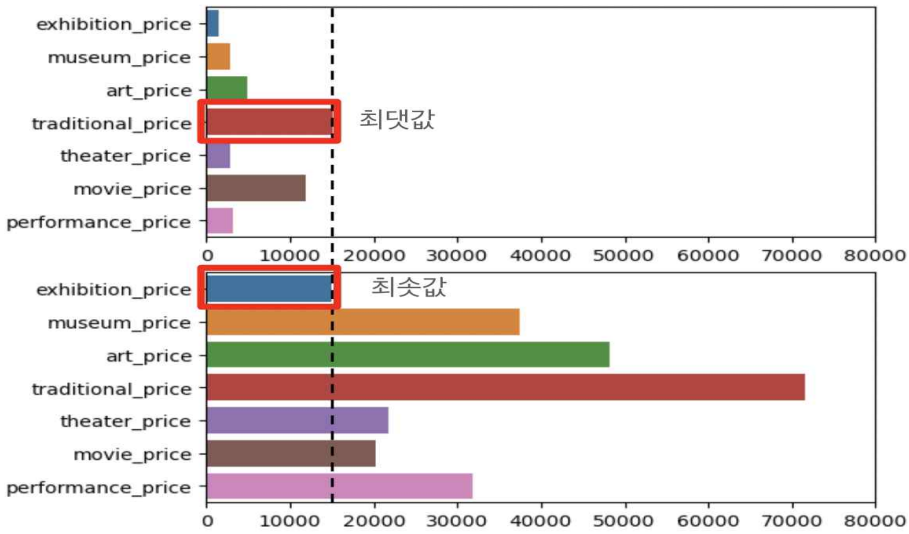
| 모델 | Logistic Regression | LightGBM Classifier | **Random Forest** |
|---|---|---|---|
| 정확도 | 0.4779 | 0.5291 | **0.5466** |

- 위 표는 각 모델의 정확도입니다.
- 하이퍼 파라미터 튜닝을 통해 각 모델의 성능을 향상시켜주었습니다.
- 그 결과, **분류 모델의 정확도가 Random Forest의 경우 제일 높았습니다**. 따라서 랜덤포레스트 분류 모델을 최종 모델으로 채택하고, 해당 모델의 예측값을 사용하여 후에 소비분석과 연계하였습니다.

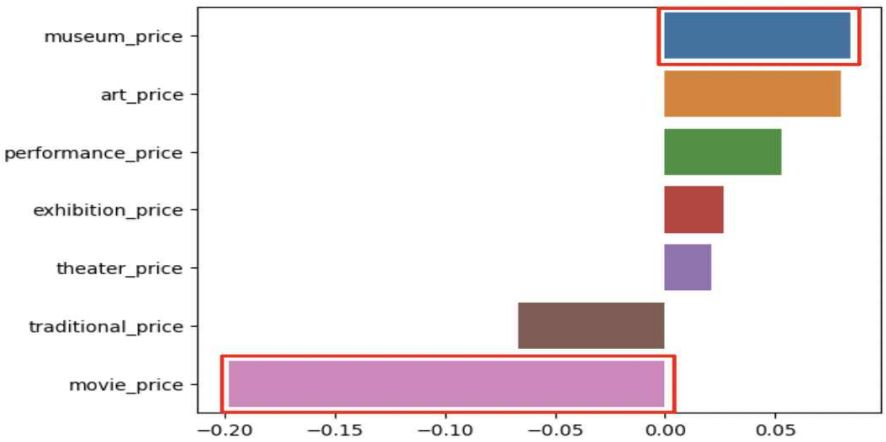



- 위 그래프는 최종 채택된 랜덤포레스트 분류 모델의 feature 중요도를 시각화한 그래프입니다. 소득(income)이 가장 높은 feature 중요도를 보이고, 문화예술행사 선택기준(standard)과 문화예술행사 관람 걸림돌(block)이 그 다음 순위를 보이는 것을 볼 수 있습니다.
- 따라서 **소득이 문화예술행사 참여에 제일 중요한 요인으로 생각하고 소비분석을 진행** 했습니다.

### 3. 소비분석




- 소비분석 결과 저소득층과 고소득층의 지출액은 현저한 차이를 보입니다. 위 그래프의 상단은 저소득층의 연간 문화활동 분야별 소비 비용 그래프이고, 하단은 고소득층의 연간 문화활동 분야별 소비 비용 그래프입니다.
- 그래프에서 볼 수 있듯이 **저소득층의 분야별 문화활동 지출 최댓값은 고소득층의 분야별 문화활동 지출 최솟값과 거의 동일한 수준임을 확인할 수 있습니다.**

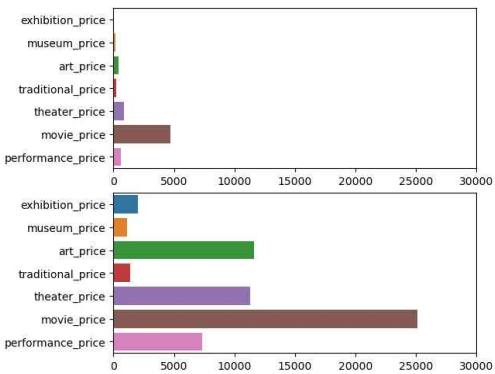


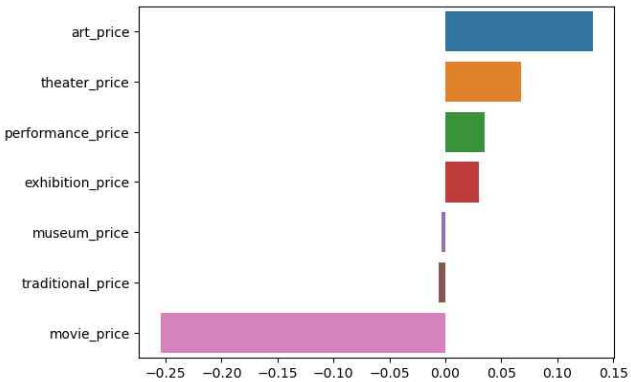

- **그렇다면 소득에 따라 향유하는 문화활동분야의 차이를 알아보기 위해서 분야별 소비 금액 평균을 비율로 계산하여 차이를 나타내보았습니다.** 고소득층의 경우, 박물관, 음악 및 무용 발표회 및 전시회에 소비하는 금액의 비율이 저소득층 보다 높았습니다. 반면 저소득층의 경우, 영화 관람에 소비하는 금액의 비율이 고소득층 보다 높았습니다. 이는 군집화 결과와 비슷한 양상을 띕니다.

- 소비 분석을 통해 문화생활에 소비하는 비용과 분야 측면에서 특정 문화예술분야에 저소득층의 문화 소외 현상이 발생한다는 것을 확인할 수 있었습니다. **문화생활에 소외되는 그룹의 원인 파악을 위해 문화예술행사에 적게 참여한 그룹의 관람 걸림돌 설문 데이터를 확인하였습니다.** 가장 큰 이유로 뽑히는 세 가지 관람 걸림돌은 다음과 같은 것을 확인할 수 있었습니다.


1. 비용이 많이 든다.
2. 시간이 좀처럼 나지 않는다.
3. 관심 있는 프로그램이 없다.

## 4. 최신 데이터(2022년)와 비교




- 위 그래프들은 2022년 데이터를 통해 소비분석을 진행한 결과를 나타낸 그래프입니다.
- 왼쪽은 문화예술활동 분야별 소비 금액 그래프입니다. **소득에 관계없이 영화 분야에 지출하는 금액이 현저히 큰 것을 확인할 수 있습니다.** 하지만 소득에 따라 지출하는 비용의 차이는 여전히 존재했습니다.
- 오른쪽의 소비 금액의 평균 차 그래프를 보면, 여전히 가구소득이 낮은 사람들은 영화 관람에 많이 소비하고, 소득이 높은 사람들은 다양한 활동에 소비하는 것을 확인할 수 있습니다.
- 소득에 관계 없이 영화 관람 소비가 너무 높은 것으로 보아, 코로나가 성행 중이던 2022년의 문화예술 활동 설문 데이터는 COVID-19로 인해 크게 왜곡 되었을 것으로 판단됩니다. 팬데믹 이전과 완전히 동일할 수는 없겠지만, 엔데믹이 선언된 현재의 상황에서는 COVID-19 이전인 2020년까지의 데이터를 사용하여 문화 예술 활동에 대한 분석을 하는 것이 올바를 것으로 판단됩니다. 따라서 저희는 2022년 데이터를 배제하여 분석하기로 결정하였습니다.

## 5) 시사점 및 기대효과

지금까지 분석을 통해 시니어의 문화생활 참여도에 가장 큰 영향을 주는 요인이 소득임을 확인하였습니다. 또한 저소득층과 고소득층이 문화생활에 소비하는 비용에 차이가 존재하고, 특정 문화분야별로 지출하는 금액에도 차이가 존재하는 것을 알아볼 수 있었습니다.

또한, 문화예술행사 참여도가 낮은 그룹의 관람에 걸림돌이 되는 부분은 '비용', '시간, '관심도', '관련 정보 부족' 순으로 나타났습니다.

이를 통해 시니어 저소득층의 문화 소외 현상을 해소하고 참여도를 고취시킬 방안을 제시하려합니다.

먼저, 저소득층이 고소득층에 비해 상대적으로 소외된 박물관과 음악 및 무용 발표회 등에 참여도를 높이기 위해 해당 문화예술분야에 더욱 적극적으로 복지 제도를 마련해야 합니다.

또한 각 관람에 걸림돌이 되는 상위 요인을 고려하여 국가가 지원할 수 있는 해결책으로 아래의 내용을 제안합니다.

가장 큰 걸림돌로 꼽힌 비용 문제는 국가의 금전적 지원을 통해 일정 부분 해소할 수 있습니다. 하지만 이는 국가의 지원에 너무 의존한다는 점, 악용 가능성과 지원 금액의 한계 등의 이유로 지속 가능한 해결책이라고 보기 힘들고, 문화누리카드 등의 문화격차 완화를 위해 소외계층을 대상으로 기존에 문화예술활동을 지원하는 제도가 이미 존재한다는 점 등을 미루어보아, 기존에 운영되던 금전적 비용 지원 제도의 타겟층을 시니어로 포함하여 범위를 넓히는 방안을 고려하는 것이 바람직합니다.

또 다른 관람 걸림돌로 꼽힌 관심도와 관련 정보 제공 미비 측면에서는, 저소득층의 참여가 저조한 문화예술활동 분야에 관련한 정보를 더 많이 제공할 수 있는 홍보 플랫폼을 마련하는 등의 방안을 고려해야합니다. 또한 저소득층 시니어를 대상으로 하는 문화예술행사 참여 사업 및 프로그램 개발 등을 고려할 수 있겠습니다.

현재 대한민국은 빠르게 초고령화사회로 진입하고 있고, 이에 따라 액티브 시니어의 등장이 주목받고 있습니다. 정년 퇴직 후 시간적, 경제적 여유를 기반으로 문화여가활동에 적극적으로 소비하는 액티브 시니어가 늘고있고, 이는 사회경제적으로도 바람직한 방향입니다.

하지만 경제적 여유가 있는 고소득층의 시니어만이 액티브 시니어가 될 수 있다면 그것은 양극화 현상을 야기할 것입니다. 따라서 국가적 차원에서 선제적으로 지원해주는 복지 정책을 통해 저소득층의 시니어 또한 액티브 시니어와 비교하여 상대적으로 소외되는 현상을 완화시킬 수 있을 것입니다.

[붙임 3] 공통 양식(참가 서약서)

# 참가 서약서

본인(팀)은 "2023년 제11회 문화데이터 활용 경진대회"에 출품하며 아래 사항을 숙지하고, 허위사실 기재 및 타인의 권리를 침해하는 등의 행위로 인하여 손해를 발생시키는 경우, 본인의 귀책으로 인하여 발생되는 손해에 관한 손해배상책임이 본인에게 있음을 확인합니다.

1. 이미 채택된 제안과 동일한 것, 표절 및 복제 등의 지적재산권 침해 작품, 타 공모전 입상작품 등은 심사에서 제외되며, 이에 따른 모든 책임은 참가자에게 있음
2. 제출한 작품이 제3자의 권리(소유권, 저작권, 이용권)를 침해하였거나 이와 관련한 분쟁이 발생한 사실이 없으며, 이로 인하여 발생하는 법적인 책임은 출품자에게 있음
3. 수상 이후 위반 사실이 밝혀질 경우 수상 취소 및 상금 환수(자진반납)에 이의를 제기하지 않음

본인은 유의사항을 충분히 숙지하였으며 대회진행에 필요한 주관기관의 요구사항에 성실히 응할 것에 동의합니다.

2023년 7월 12일

서약자 팀(기업)명 성명 김종원 김종원
성명 김지우 김지우
성명 김연진 김연진
성명 김예리 김예리

한국문화정보원장 귀중