[붙임 2-3] 데이터 분석 분야 기획서

「2025 문화 디지털혁신 및 데이터 활용 공모전」 기획서
- 데이터 분석 분야 -

공모 분야	데이터 분석
공 모 명	미술관 관람객 감소 요인 분석을 통한 정책 제안
분석 툴(Tool)	■ Python □ R □ Tableau □ 기타( )

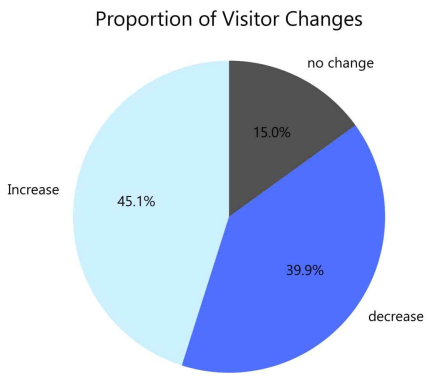
1) 분석 개요

1. 전국 미술관의 40%가 관람객 감소 현상 나타남
2. 2024년 관람객 수가 전년도보다 감소했는지를 기준으로 분류(Classification) 모델을 적용하여, 관람객 감소에 영향을 준 주요 요인 분석
3. 인사이트 발견 및 정책 제안

2) 분석 배경 및 목적

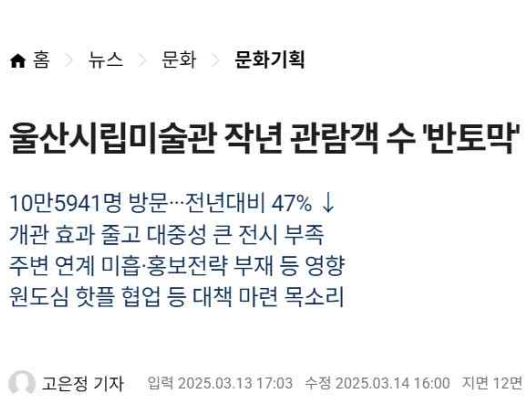
1. 분석 배경

문화예술에 대한 사회적 관심이 증가하고 있음에도 불구하고, 한국문화정보원의 2023년과 2024년 미술관 시설 및 현황통계 데이터에 따르면 전국 미술관 가운데 약 40%는 관람객 수가 감소하는 현상이 나타나고 있다[그림 1].



[그림 1]

국립현대미술관은 대형 기획전과 미디어 연계 홍보를 통해 지난해 약 38만 2천명의 관람객이 증가했지만, 이와 달리 중소규모 미술관, 지역 기반 공공 미술관은 관람객 확보에 어려움을 겪고 있다. 울산매일 지역신문에 따르면 울산시립 미술관은 2022년의 관람객이 총 19만 4,235명인 것에 비해, 2024년에는 10만5,941명으로 관람객 수가 약 47% 정도가 감소했다[그림 2]. 또한, OBS 뉴스는 경기도문화재단 소속 미술관의 관람객 수가 2023년 대비 2024년에 6만 명 감소했다고 보도했다[그림3]. 이러한 불균형은 미술관 간 콘텐츠 경쟁력, 주변 상업지역의 활성 유무, 관람 경험 설계 등 운영 전략의 편차로부터 기인하는 구조적 문제로 해석할 수 있다.



[그림 2]

[그림 3]

2. 필요성

관람객의 감소는 운영 수익 악화로 연결되며, 장기적으로는 휴관 또는 폐관과 같이 시설의 존속에도 영향을 미칠 수 있다. 따라서 관람객이 줄어들고 있는 미술관들의 감소 원인을 명확히 찾고, 실질적인 전략을 수립하는 일이 필수적이다. 이는 국민의 문화 접근성과 예술 향유 기회를 확대하는 결과로 이어질 것이다.

3. 목적

본 분석의 목적은 미술관 관람객 감소 요인을 정량적으로 분석하여 전략적인 대응 방안을 찾는 데에 있다. 관람객 감소는 단일한 요인에서 비롯되지 않는다. 전시 콘텐츠, 체험 프로그램 운영, 관람료, 미술관 공간, 주변 인프라 등 관람객 감소에 영향을 미칠 가능성이 큰 다양한 요인들 중 관련성이 가장 큰 요인을 찾고 그에 대한 해결 방안을 모색하여 정책을 제안한다. 무작정 방문객 수를 늘리기 위한 마케팅이나 홍보를 하는 것을 넘어, 어떤 요인이 관람객 유입에 기여하고 있는지에 대한 실질적인 원인을 파악하여 왜 관람객 이탈이 발생하는지를 이해하는 것이 핵심이다.

3) 분석 내용 및 결과 (데이터 분석 목적 : 전년도 대비 미술관 관람객 감소 요인 찾기)

[Step1. Data info check : 수집된 데이터의 기본 정보 확인]

1-1. 2024년 국내 문화체육관광 분야 미술관 시설 및 현황 통계 데이터 불러오기

	ID	LCLAS_NM	MLSFC_NM	FCLTY_NM	CTPRVN_NM	SIGNGU_NM	LEGALDONG_CD	LEGALDONG_NM	ADSTRD_CD	ADSTRD_NM	...
0	KCDMART23N000000001	문화시설	미술관	국립현대미술관(과천관)	경기	과천시	4129010400	막계동	4129056000	문원동	...
1	KCDMART23N000000002	문화시설	미술관	국립현대미술관(서울관)	서울	종로구	1111014200	소격동	1111054000	삼청동	...
2	KCDMART23N000000003	문화시설	미술관	국립현대미술관(덕수궁관)	서울	중구	1114016700	정동	1114052000	소공동	...
3	KCDMART23N000000004	문화시설	미술관	국립현대미술관 미술품수장센터(청주관)	충북	청주시	4311410200	내덕동	4311453000	내덕2동	...
4	KCDMART23N000000005	문화시설	미술관	DDP디자인뮤지엄	서울	중구	1114014900	을지로7가	1114059000	광희동	...
...	...	...	...	...	...	...	...	...	...	...	...

(1) 데이터 출처: 한국문화정보원(문화 빅데이터 플랫폼에서 다운)

(2) 데이터 형태: 286row x 97columns

(3) 분석에 필요한 칼럼 선택:

FCLTY_NM(시설명), CTPRVN_NM(시도명), SIGNGU_NM(시군구명), ADSTRD_NM(행정동), DATA_CO(보유 자료 수), TOT_PROGRM_CO(총 프로그램 수), VIEWNG_PRICE(일반 관람료), ARTGR_EMP_CO(미술관 직원 수), BULD_AR_VALUE(건물면적), DSPY_AR_CN(전시면적), FLAG_NM(운영 주체). VIEWNG_NMPR_CO(관람 인원 수)

(4) 데이터 정보 확인: DATA_CO, TOT_PROGRM_CO, VIEWNG_PRICE, ARTGR_EMP_CO, BULD_AR_VALUE, VIEWNG_NMPR_CO 칼럼에 Null값 존재.

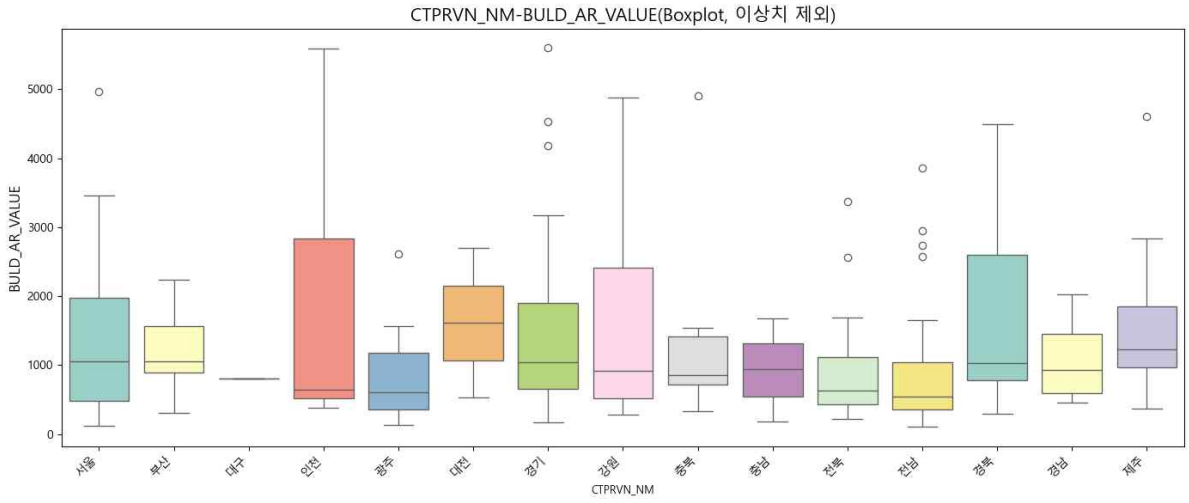
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 286 entries, 0 to 285
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   FCLTY_NM        286 non-null    object
 1   CTPRVN_NM       286 non-null    object
 2   SIGNGU_NM       286 non-null    object
 3   ADSTRD_NM       286 non-null    object
 4   DATA_CO         149 non-null    float64
 5   TOT_PROGRM_CO   214 non-null    float64
 6   VIEWNG_PRICE    96 non-null     object
 7   ARTGR_EMP_CO    173 non-null    float64
 8   BULD_AR_VALUE   283 non-null    float64
 9   DSPY_AR_CN      286 non-null    float64
 10  FLAG_NM         286 non-null    object
 11  VIEWNG_NMPR_CO  278 non-null    object
dtypes: float64(5), object(7)
memory usage: 26.9+ KB
```

(5) 불필요한 데이터 제거: VIEWNG_NMPR_CO 칼럼에 Null인 미술관은 모두 '휴관'이다. 휴관인 미술관은 대부분의 항목에서 데이터가 집계되지 않아 분석에서 제외하였다.
해당 미술관은 다음과 같다: 한광미술관, 남철미술관, 여진불교미술관, 일현미술관, 아산조방원미술관, 장전미술관, 제주유리의성, 뮤지엄한미

(6) NULL값 처리

• DATA_CO(보유 자료 수) 칼럼 NULL값 처리:
DATA_CO와 BULD_AR_VALUE의 상관계수는 0.5694로 중간 정도의 양의 상관관계를 보인다.
이에 따라 BULD_AR_VALUE를 기준으로 미술관들을 4분위로 나눈 뒤, 각 그룹 내에서의
평균 DATA_CO 값을 활용하여 NULL값을 처리했다. 평균을 집계할 때, BULD_AR_VALUE가 NULL인 경우는 집계에서 제외했다.

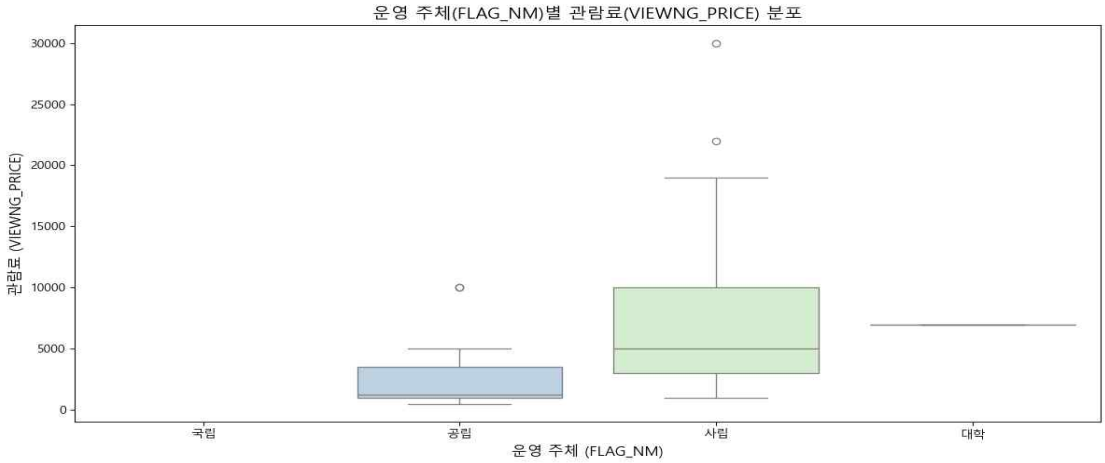
• BULD_AR_VALUE(건물면적) 칼럼 NULL값 처리:
CTPRVN_NM(시도명)을 기준으로 BULD_AR_VALUE의 분포를 Boxplot으로 시각화하였다.



Boxplot을 확인한 결과 시도별로 미술관 건물면적 분포에 차이가 있음을 확인할 수 있다.
이를 바탕으로 CTPRVN_NM(시도명)을 기준으로 그룹화하여 BULD_AR_VALUE(건물면적)의
이상치를 제외한 후 평균을 계산하였고, 해당 평균값을 활용해 BULD_AR_VALUE칼럼의 NULL을 처리했다.

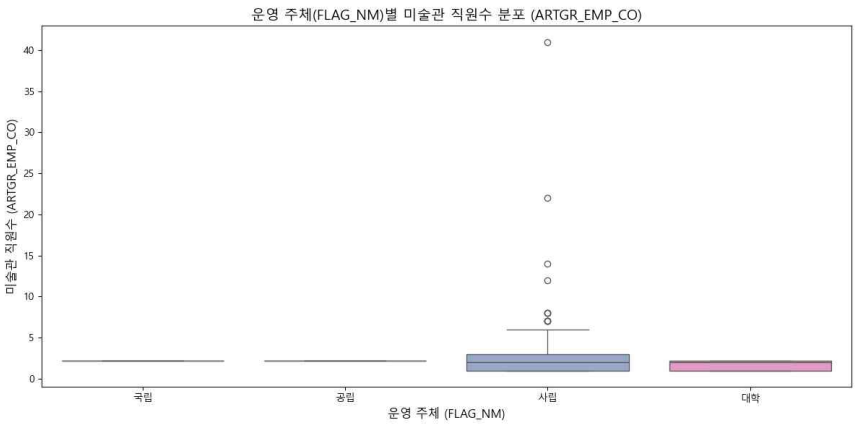
• TOT_PROGRM_CO(총 프로그램 수) 칼럼 NULL값 처리: TOT_PROGRM_CO가 집계되지 않은 미술관은 검색 결과 전시 중심 공간인 경우가 대부분이라 NULL값을 0으로 처리했다.

• VIEWNG_PRICE(일반 관람료) 칼럼 NULL값 처리: VIEWNG_PRICE 칼럼의 데이터 타입은 초기에는 object로 확인되었으며, 이는 문자열이 포함된 값들 때문이었다. 예를 들어, "전시별 상이", "13,000 (커피제공)", "야간:22000, 주간:10000"과 같이 숫자 이외의 문자가 포함된 값들이 존재하였다. Null값을 처리하기 위해 우선 해당 칼럼 내의 모든 값을 정제하여 숫자형으로 변환한 뒤, 데이터 타입을 숫자형을 변경하였다. 이후 FLAG_NM(운영주체)를 기준으로 일반 관람료의 분포를 Boxplot으로 시각화하였다.



Boxplot을 통해 운영 주체별로 일반 관람료 VIEWNG_PRICE의 분포에 차이가 있음을 확인할 수 있었다. 이에 따라 Null값 처리 방식을 운영 주체별로 다음과 같이 적용하였다.
FLAG_NM = '대학' : 관람료가 모두 무료인 것으로 확인되어, Null값을 0으로 처리하였다.
FLAG_NM = '국립' : 국립현대미술관(과천관, 서울관, 덕수궁관, 청주관) 등 4개 기관의 관람료가 모두 Null이 었으며, 공식 홈페이지 정보를 참조하여 수기로 Null값을 처리하였다.
FLAG_NM = '공립' 또는 '사립' : 두 운영 주체 각각을 그룹화하여 VIEWNG_PRICE의 평균을 산출한 후, 해당 그룹 평균값으로 Null값을 처리하였다.

• ARTGR_EMP_CO(미술관 직원 수) 칼럼 NULL값 처리:
FLAG_NM(운영주체)을 기준으로 ARTGR_EMP_CO의 분포를 Boxplot으로 시각화하였다.
Boxplot을 확인한 결과 운영주체별로 미술관 직원 수의 분포에 차이가 있음을 확인할 수 있다.



이를 바탕으로 FLAG_NM(운영주체)을 기준으로 그룹화하여 ARTGR_EMP_CO(미술관 직원 수)의 이상치를 제외한 후 평균을 계산하였고, 해당 평균값을 활용해 ARTGR_EMP_CO 칼럼의 NULL을 처리했다.

• VIEWNG_NMPR_CO(관람 인원 수) 칼럼 NULL값 처리: VIEWNG_NMPR_CO 칼럼에는 Null값으로 집계되지 않았지만 '미제출'값이 존재한다. 이를 Null값으로 바꾼 후 관람객 수 데이터 타입을 숫자형으로 변환하였다. Null값

처리 방식으로는 이전 행과 다음 행의 값을 평균 내어 소수점은 버린 뒤, 해당 위치의 Null값을 대체하였다.

1-2. 2023년 국내 문화체육관광 분야 미술관 시설 및 현황 통계 데이터 불러오기

	ID	LCLAS_NM	MLSFC_NM	FCLTY_NM	CTPRVN_NM	SIGNGU_NM	LEGALDONG_CD	LEGALDONG_NM	ADSTRD_CD	ADSTRD_NM	...
0	KCDMART23N000000001	문화시설	미술관	국립현대미술관(과천관)	경기	과천시	4129010400	막계동	4129056000	문원동	...
1	KCDMART23N000000002	문화시설	미술관	국립현대미술관(서울관)	서울	종로구	1111014200	소격동	1111054000	삼청동	...
2	KCDMART23N000000003	문화시설	미술관	국립현대미술관(덕수궁관)	서울	중구	1114016700	정동	1114052000	소공동	...
3	KCDMART23N000000004	문화시설	미술관	국립현대미술관 미술품수장센터(청주관)	충북	청주시	4311410200	내덕동	4311453000	내덕2동	...
4	KCDMART23N000000005	문화시설	미술관	DDP디자인뮤지엄	서울	중구	1114014900	을지로7가	1114059000	광희동	...
...	...	...	...	...	...	...	...	...	...	...	...

(1) 데이터 출처: 한국문화정보원(직접 요청)

(2) 데이터 형태: 293row x 97columns

(3)~(6) 프로세스는 1-1과정과 동일하므로 생략한다.

1-3. 2024년 소상공인시장진흥공단_상가(상권)정보 데이터 불러오기

	상가업소번호	상호명	지점명	상권업종대분류코드	상권업종대분류명	상권업종중분류코드	상권업종중분류명	상권업종소분류코드	상권업종소분류명	표준산업분류코드	...
0	MA0101202210A0084547	금강산노래광장	NaN	I2	음식	I211	주점	I21101	일반유흥주점	I56211	...
1	MA0101202208054309O3	엔젤	NaN	I2	음식	I201	한식	I20101	백반/한정식	I55109	...
2	MA0101202208054309 41	누베헤어	NaN	S2	수리·개인	S207	이용·미용	S20701	미용실	S96112	...
3	MA0101202208054309 46	공차	NaN	I2	음식	I212	비알코올	I21201	카페	I56229	...
4	MA0101202208054313 69	행운쉼터	NaN	I1	숙박	I101	일반숙박	I10103	펜션	I55104	...
...	...	...	...	...	...	...	...	...	...	...	...

(1) 데이터 출처: 소상공인진흥공단(공공데이터포털에서 다운)

(2) 데이터 형태: 2338043row x 39columns

(3) 분석에 필요한 칼럼 선택 및 필터링: 시도명, 시군구명, 행정동명, 상권업종대분류명 칼럼을 선택하고 '상권업종대분류명' = '음식'인 데이터로 필터링

(4) 파생 칼럼 생성: 행정동별로 음식점 상권의 분포를 파악하기 위해 '시군구명', '행정동'을 기준으로 그룹화한 뒤, 각 행정동별 음식점 상권 수를 집계하여 새로운 파생변수 '개수'를 생성하였다. 결과는 다음과 같다.

	시군구명	행정동명	개수
0	강릉시	강남동	177
1	강릉시	강동면	134
2	강릉시	경포동	237
3	강릉시	교1동	607
4	강릉시	교2동	166
...	...	...	...
3539	충주시	주덕읍	131
3540	충주시	중앙탑면	166
3541	충주시	지현동	94
3542	충주시	칠금.금릉동	399
3543	충주시	호암.직동	234

1-4. 수집된 데이터 JOIN하기

(1) 2024년 미술관 시설 데이터와 2024년 음식점 상권 데이터를 2024년 미술관 시설의 'SIGNGU_NM(시군구)', 'ADSTRD_NM(행정동)' 기준으로 'LEFT JOIN'을 실행하였다. 이때 미술관 시설 데이터를 기준으로 상권 데이터가 결합되었기 때문에, 일부 미술관 위치에는 매칭되는 상권 수 정보가 없어 NULL값이 발생하였다. NULL값 처리를 위해 먼저 SIGNGU_NM별로 상권 수의 평균값을 계산하였다. 이후 칼럼의 NULL값에 대해 동일 SIGNGU_NM의 평균값으로 대체하였다.

(2) 다음으로 2024년 미술관 시설 데이터와 2023년 미술관 시설데이터를 2024년 미술관 시설의 시설명 FCLTY_NM(시설명)을 기준으로 'INNER JOIN'을 진행하였다. 이 두 데이터는
연도만 다르고 동일한 칼럼명을 가지고 있어, 병합 후 칼럼들을 구분하기 위해 각 칼럼명 앞에 해당 연도를 접두어로 명시하였다.

[Step2. Data Readiness Check : 데이터 분석을 위한 사전 점검]

2-1. Target Label생성 : 분류 모델 학습을 위해 2024년 미술관 관람객 수가 전년 대비 감소했는지 나타내는 Target Label인 viewer_decline을 생성하였다.
(관람객 수가 감소하면 1, 감소하지 않으면 0)

```
df_art_Joined['viewer_decline'] = df_art_Joined.apply(
    lambda row: 1 if row['2024_VIEWNG_NMPR_CO'] < row['2023_VIEWNG_NMPR_CO'] else 0,
    axis=1
)
```

2-2. Target Ratio확인

```
df_art_Joined['viewer_decline'].value_counts()
```

결과 : 값이 0인 경우는 188개, 값이 1인 경우는 74개이다.

[Step3. Data Mart : 데이터 마트 기획 및 설계]

3-1. Data Mart 정의서

가설	범주	변수	설명	산식
전시 및 프로그램 수가 관람객 감소에 영향을 미칠 것이다	전시 및 프로그램 통계	2023_DATA_CO	2023년 보유 자료 수	
		2024_DATA_CO	2024년 보유 자료 수	
		DATA_CO_GROWTH_RATE	전시자료수 증가율	(2024년 - 2023년) / 2023년
		2023_TOT_PROGRM_CO	2023년 총 프로그램 수	
		2024_TOT_PROGRM_CO	2024년 총 프로그램 수	
		TOT_PROGRM_CO_GROWTH_RATE	프로그램 수 증가율	(2024년 - 2023년) / 2023년
관람료가 관람객 감소에 영향을 미칠 것이다.	관람료	2023_VIEWNG_PRICE	2023년 관람료	
		2024_VIEWNG_PRICE	2024년 관람료	
		VIEWNG_PRICE_GROWTH_RATE	관람료 증가율	(2024년 - 2023년) / 2023년
미술관의 인력수가 관람객 감소에 영향을 미칠 것이다.	인력	2023_ARTGR_EMP_CO	2023년 미술관 직원 수	
		2024_ARTGR_EMP_CO	2024년 미술관 직원 수	
		ARTGR_EMP_CO_GROWTH_RATE	미술관 직원 수 증가율	(2024년 - 2023년) / 2023년
미술관의 공간 및 면적이 관람객 감소에 영향을 미칠 것이다.	공간 및 면적	2023_DSPY_AR_CN	2023년 전시면적	
		2024_DSPY_AR_CN	2024년 전시면적	
		DSPY_AR_CN_GROWTH_RATE	전시면적 증가율	(2024년 - 2023년) / 2023년
		2023_BULD_AR_VALUE	2023년 건물면적	
		2024_BULD_AR_VALUE	2024년 건물면적	
미술관의 운영 주체가 관람객 감소에 영향을 미칠 것이다.	정보	FLAG_NM	운영주체	
미술관 주변 인프라가 관람객 감소에 영향을 미칠 것이다.	인프라	2024_food_count_by_dong	2024년 미술관 주변 음식점 수	

3-2. Feature Engineering

전체 데이터 개수가 262개로 비교적 적은 편이기 때문에, 모든 변수를 그대로 모델에 투입할 경우 과적합(Overfitting)이 발생할 우려가 있다. 이에 따라 모델의 성능을 향상시키고 해석력을 높이기 위해 Feature Engineering을 수행하였으며, 이 과정을 통해 의미 있는 일부 변수만을 선별하여 모델에 활용하고자 한다. 특히 본 분석은 이진 분류(Classification) 문제이므로, 변수 선택 과정에서 IV(Information Value) 기반의 Feature

Engineering 기법을 적용하였다. IV는 각 Feature가 타겟 변수와의 구분력을 얼마나 잘 가지는지를 수치화한 지표로, 모델에 유의미한 변수를 효과적으로 선별하는 데 활용된다.

• IV (Information Value) : IV는 하나의 Feature가 Good(Target=0)과 Bad(Target=1)을 구분해주는 정보량을 수치화한 값이다. 일반적으로 IV 값이 클수록 Target과 Non-target을 잘 구분하는 변수, 반대로 IV 값이 작을수록 정보량이 적은 변수로 간주된다.

• IV 수치를 구하기 위해 OptimalBinning패키지 활용

```
# 학습에 필요없는 Column 제거
df_art_IV = df_art_MT.drop(['CTPRVN_NM', 'SIGNGU_NM', 'ADSTRD_NM', '2024_FCLTY_NM', '2024_VIEWNG_NMPR_CO', '2023_VIEWNG_NMPR_CO'], axis=1)
```

numerical 변수에 대한 IV값을 구하기 위한 코드

```
from optbinning import OptimalBinning

# numerical
columns_except_target_and_flag = [col for col in df_art_IV.columns if col not in ['viewer_decline', 'FLAG_NM']]
integer_list = columns_except_target_and_flag
iv_df = []

for i in integer_list :
  variable = i
  x = df_art_IV[variable].values
  y = df_art_IV.viewer_decline
  # max_n_prebins
  optb = OptimalBinning(name=variable, dtype="numerical", solver="cp", max_n_prebins=3)
  optb.fit(x, y)
  # print("split points : ", optb.splits)

  binning_table = optb.binning_table
  v1 = binning_table.build()

  df = pd.DataFrame({'val' : variable,
                     'IV' : [v1.loc['Totals','IV']]})
  iv_df.append(df)

iv_df = pd.concat(iv_df).reset_index(drop=True)
iv_df.sort_values(by=['IV'], ascending = False)
```

Categorical 변수('FLAG_NM')에 대한 IV값을 구하는 코드

```
# Categorical
variable_cat = "FLAG_NM"
x_cat = df_art_IV[variable_cat].values
y_cat = df_art_IV.viewer_decline

# OptimalBinning 객체 생성 (categorical 타입)
optb = OptimalBinning(name=variable_cat, dtype="categorical", solver="mip", cat_cutoff=0.1)

# 학습
optb.fit(x_cat, y_cat)

# binning 테이블 생성
binning_table = optb.binning_table
binning_result = binning_table.build()

# IV 값만 출력
iv_value = binning_result.loc['Totals', 'IV']
print(f"Variable: {variable_cat}, IV: {iv_value:.6f}")
```

IV값이 0.1 이상이 나와야 모델에 유의미하게 기여 가능하다.
IV값이 0.1 이상인 Feature는 다음과 같다.

Value	IV
2023_BULD_AR_VALUE	0.243421
2023_DATA_CO	0.154039
2023_DSPY_AR_CN	0.143817
2023_VIEWNG_PRICE	0.134772
2024_VIEWNG_PRICE	0.126709
2023_TOT_PROGRM_CO	0.114284
ARTGR_EMP_CO_GROWTH_RATE	0.112792

2023_VIEWNG_PRICE와 2024_VIEWNG_PRICE는 서로 상관관계가 높아 다중공선성 문제가 발생할 수 있으므로, IV 값이 더 높은 2023_VIEWNG_PRICE를 선택하였다.

따라서 모델 학습에 사용할 변수는 다음과 같다.
2023_BULD_AR_VALUE, 2023_DATA_CO, 2023_DSPY_AR_CN, 2023_VIEWNG_PRICE,
2023_TOT_PROGRM_CO, ARTGR_EMP_CO_GROWTH_RATE

[Step4. Data Modeling]

4-1. 데이터 모델링을 위한 사전 준비

target ratio를 확인한 결과, 관람객 감소(1)에 해당하는 사례가 전체의 약 28.2%로, 클래스 간 비율이 크게 불균형한(Class Imbalanced) 상태이다. 이로 인해 모델은 감소하지 않은 시설(0)에 편향되어 학습될 가능성이 높으며, 결과적으로 감소 원인을 탐지하는 모델의 실효성이 저하될 수 있다. 이러한 문제를 완화하기 위해 소수 클래스 데이터를 기반으로 새로운 데이터를 생성하는 SMOTE(Synthetic Minority Over-sampling Technique) 기법을 적용하여 데이터의 균형을 맞추었다. SMOHE기법은 소수 클래스의 주변 이웃(K-Nearest Neighbors)을 기준으로 새로운 샘플을 선형 보간(Interpolation) 방식으로 생성하는 기법이다.

```
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

# ▶ X, Y 분리
X = df_art_MT.drop(['viewer_decline'], axis=1)
Y = df_art_MT['viewer_decline']

# ▶ SMOTE 적용하여 1:1 클래스 비율 맞추기
smote = SMOTE(random_state=1234)
X_balanced, Y_balanced = smote.fit_resample(X, Y)

# ▶ train-test split (이제 stratify 생략 가능: 이미 클래스 균형됨)
x_train, x_test, y_train, y_test = train_test_split(
    X_balanced, Y_balanced, test_size=0.3, random_state=1234
)
```

target ratio 감소하지 않은 경우(0)가 188건, 감소한 경우(1)가 74건인 상태에서 SMOTH적용 후 적용 후에는 두 클래스 모두 188건으로 균형을 이루게 되었다.

4-2. Model Selection

Classification 문제를 해결하기 위해서는 여러 대표적인 알고리즘을 테스트해보는 것이 유리하다. 본 분석에서는 다음과 같은 모델들을 사용하여 성능을 비교하였다:
Logistic Regression, Random Forest (Tree 기반 Bagging 앙상블), LightGBM (Boosting 기반 앙상블) 각 모델(Logistic Regression, Random Forest, LightGBM)에 대해 동일한 학습 및 평가 프로세스를 적용하였다. 구체적으로는, 훈련 데이터로 모델을 학습한 후 테스트 데이터에 대해 예측을 수행하였으며, 모델 성능 평가는 F1 Score와 AUC(Area Under the Curve)를 기준으로 진행하였고 결과는 다음과 같다.

model	f1_train	f1_test	AUC_train	AUC_test
LR_Standard	0.614841	0.568966	0.6369	0.593417
LR_Norm	0.646259	0.566667	0.634876	0.592163
RFC	1	0.728972	1	0.820219
LGBM	0.969466	0.716981	0.998178	0.809091
RFC(tuned)	0.806084	0.653846	0.898149	0.761755
LGBM(tuned)	0.976923	0.684685	0.999161	0.79232

앞선 모델 성능 비교 결과, Random Forest(RFC) 및 LightGBM(LGBM) 모델 모두 학습 데이터에 비해 테스트 데이터에서의 성능이 상대적으로 낮아, 과적합(Overfitting) 가능성이 제기되었다. 이에 따라, 보다 일반화 성능이 높은 모델을 도출하기 위해 BayesianOptimization 패키지를 활용하여 하이퍼파라미터 튜닝 작업을 수행하였다.
Random Forest (RFC) : n_estimator=160, max_depth=3, max_feature=0.7, oob_score=True
LightGBM (LGBM) : n_estimator=198, max_depth=4
모델 선택에 있어 가장 중요하게 고려한 기준은 과적합(overfitting)의 최소화였으며,
그 다음으로는 일관되고 안정적인 예측 성능 확보를 목표로 하였다. F1 Score와 AUC 지표를 기준으로 비교한 결과, Random Forest(RFC), LightGBM(LGBM) 및 하이퍼파라미터 튜닝을 적용한 모델(RFC(tuned), LGBM(tuned))은 학습 데이터에서 매우 높은 성능을 보였으나, 테스트 데이터에서는 상대적으로 큰 성능 하락폭을 보이며 과적합 가능성이 높게 나타났다. 반면, Logistic Regression (LR_Standard) 모델은 학습 데이터와 테스트 데이터 간 성능 차이가 가장 작고, 전반적으로 일관된 예측 성능을 유지하였다:

- F1 Score: Train 0.614 → Test 0.569 (차이 약 0.046)
- AUC: Train 0.637 → Test 0.593 (차이 약 0.044)

이는 모델이 복잡하지 않음에도 불구하고, 과적합 없이 비교적 안정적으로 일반화 성능을 확보했음을 의미한다. 따라서 본 분석에서는 Logistic Regression(Standard Scaling 적용)을 최종 분류 모델로 선정하였다. 본 분석의 핵심 목적은 Target 변수에 영향을 미치는 주요 요인을 도출하는 것에 있다.
따라서 단순한 예측 성능보다는 모델 해석의 신뢰도 확보가 우선되어야 하며, 이를 위해 모델 선택 시 과적합(overfitting) 여부를 가장 중요한 판단 기준으로 고려하였다.

Logistic Regression(Standard Scaling 적용) 모델은 다음과 같은 절차로 생성 및 학습하였다.
우선, 입력 변수에 대해 StandardScaler를 이용해 평균 0, 표준편차 1로 스케일링을 수행하였으며,
이후 LogisticRegression 클래스를 사용하여 학습 데이터를 기반으로 모델을 학습시켰다.

```
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression

# 표준화
scaler = StandardScaler()
x_train_standard = scaler.fit_transform(x_train)
x_test_standard = scaler.transform(x_test)

# 모델 학습
LR_standard = LogisticRegression()
LR_standard.fit(x_train_standard, y_train)
```

4-3. Model Explanation

모델의 설명력을 확보하기 위해 shap 패키지 활용

Shapley Value : 특정 Feature가 예측값에 얼만가 기여하는지 파악하기 위해 특정 변수와 관련된 모든 변수 조합들을 입력시켰을 때 나온 결과값과 비교를 하면서 변수의 기여도를 계산하는 방식, 즉 특정 Feature(on/off)가 예측값에 얼마나 영향을 끼쳤는지 탐색

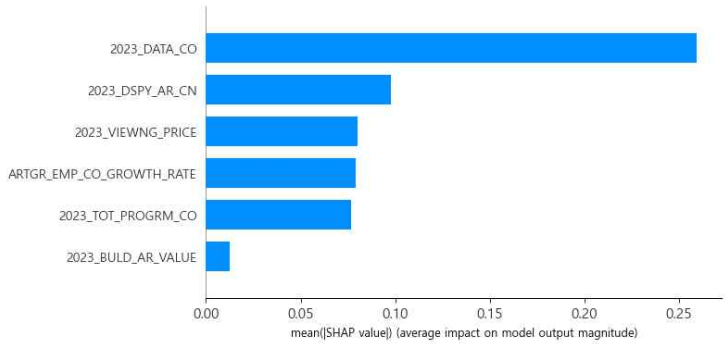
```
import shap
import matplotlib.pyplot as plt

shap.initjs()

# LR_Standard SHAP 계산
explainer = shap.LinearExplainer(LR_Standard.fit(x_train_Standard, y_train), x_train_Standard)
shap_values = explainer.shap_values(x_test_Standard)

# SHAP plot 표시 (show=False로 설정해서 폰트 제어 가능)
shap.summary_plot(shap_values, x_test_Standard,
                  feature_names=x_test.columns,
                  plot_type="bar",
                  class_names=y_test,
                  max_display=20,
                  show=False)
```

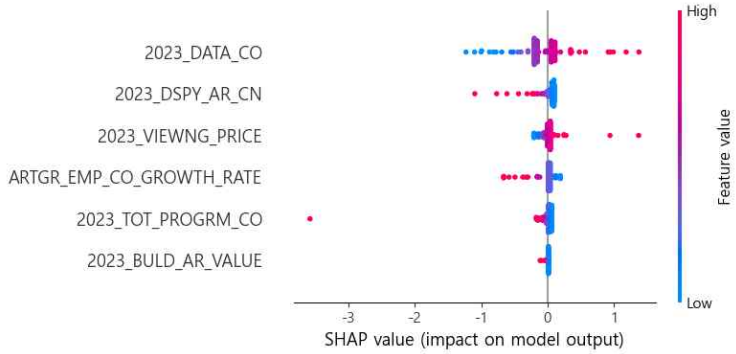
출력결과는 다음과 같다.



Feature들의 값의 크고 작음이, 관람객 감소 여부를 예측하는 데 있어서 모델(LR_Standard)이 가장 크게 참고한 변수를 내림차순으로 출력한 것이다.

```
# LR_standard shap
shap.summary_plot(shap_values, x_test_Standard, feature_names=x_test.columns)
```

의 출력결과는 다음과 같다.



- SHAP 값이 양수일수록 → 모델은 viewer_decline = 1 (감소)를 예측하는 쪽으로 기여
- SHAP 값이 음수일수록 → 모델은 viewer_decline = 0 (감소하지 않음)을 예측하는 쪽으로 기여

[Step5. Insight]

5-1. Data Mart 정의서(3-1참고)에서 IV값을 참고하면 2023년 건물면적, 2023년 보유 자료 수 2023년 전시면적, 2023년 관람료, 2023총 프로그램 수, 미술관 직원 수 증가율을 제외한 칼럼들은 미술관 관람객 감소 여부를 구분하는데 영향이 미미하다.

5-2. 모델이 관람객이 줄었는지를 판단할 때, 2023년에 보유한 자료 수가 많았는지 적었는지를 가장 중요하게 참고했습니다. 다시 말해, 보유 자료 수는 관람객 감소 여부를 예측하는 데 가장 큰 영향을 준 요소이다.

5-3. SHAP value해석

(1). 2023년에 보유 자료 수가 많았던 미술관일수록, 그 다음 해(2024년)에 관람객이 줄었을 가능성이 더 크다.

(2). 2023년에 전시면적이 넓은 미술관일수록, 2024년에 관람객 수가 유지되거나 오히려 줄지 않았을 가능성이 크다.

(3). 2023년에 관람료가 높았던 미술관일수록, 2024년에 관람객이 줄었을 가능성이 더 크다.

(4). 미술관 직원 수가 전년도 대비 많이 늘어난 곳일수록, 2024년 관람객이 줄지 않았을 가능성이 크다.

4) 시사점 및 기대효과

시사점

위의 5-3. SHAP value해석의 *(1). 2023년에 보유 자료 수가 많았던 미술관일수록, 그 다음 해(2024년)에 관람객이 줄었을 가능성이 더 크다*는 것은 보유 자료의 양과 관리 구조가 오히려 전시를 관람하는 데 부정적인 영향을 주고 있다고 해석될 수 있다. 즉, 보유 자료 수가 많아지면, 보존 및 관리 부담이 커져 실제 전시나 프로그램 운영에 집중하기보다는 자료 유지에 집중하게 되거나, 많은 자료 수로 인해 관람객이 오히려 피로와 인지적 부담을 느껴 관람 경험을 저하시킨다. 이에 따라 새로운 미술관 관리 정책을 통해 과잉 소장된 자료를 정리하고, 새로운 작품들을 보다 효율적으로 수용할 수 있도록 한다. 또한 *(3). 2023년에 관람료가 높았던 미술관일수록, 2024년에 관람객이 줄었을 가능성이 더 크다*는 결과는 관람료에 대한 관람객의 부담이 있으며 이에 대한 정책적인 방안이 필요하다는 것을 시사한다.

정책 제안

▶ 정책 개요

- 보유 자료의 이관, 기증 등의 정리를 통해 전시의 회전율을 높임으로써 방문 동기를 유발하여 미술관 방문을 유도한다.
- 기존의 문화누리카드 앱을 활용하여 문화포인트·스탬프 제도를 통해 관람료에 대한 부담을 줄인다.

▶ 정책 내용

1. 일정 시간 이상 지난 보유 자료는 이관, 기증 등을 통해 관리에 대한 부담 경감
2. 보유 자료의 순환구조 구축을 통해 신선한 콘텐츠를 지속적으로 공급
- 같은 공간, 다른 전시 → 분기 또는 시즌마다 전시 교체를 유도
- 소장품 감축 미술관에 전시 기획 예산 지원

3 문화누리카드 앱과 연계하여 "아트스탬프 – 미술관 도장 찍고 할인받자" 하위 정책 시행
- 문화누리카드 앱에 디지털 스탬프 기능 추가
- 예를 들어 3개 전시 방문 시 다음 전시 50% 할인 / 5개 방문 시 1회 무료
- 미술관별 참여 확대 유도

4. 문화포인트 환급형 할인제
- 관람료 일정 이상 지출 시, 문화포인트로 환급 (예: 관람료 5천 원당 1천 포인트)
- 포인트는 다음 관람 때 현금처럼 사용 가능

▶ 기대효과

1. 미술관 방문객 증가는 입장료, 기념품 등 미술관 수익이 증가로 이어지며 이는 미술관의 재정 안정화에 이바지할 수 있다.
2. 관람객 수는 해외 미술관과의 공동 전시 협약 및 작품 대여 등에 유리하게 작용하며 이는 보다 좋은 미술 전시를 유치게끔 할 수 있다.
3. 미술 전시를 통해 시민들의 감성·창의력·소통능력을 키우며 이는 미래 인재 육성의 기반이 될 수 있다.

5) 문화데이터 활용성

데이터명	제공 기관	URL	활용 내용
2024년 국내 문화체육관광 분야 미술관 시설 및 현황통계 데이터 (방문객 등 통계 포함)	한국문화정보원	https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=6c6b5bee-4264-441f-b232-8d26cddb5771	관람객 감소 확인 및 분류모델에 사용되는 관람객 감소 요인 (프로그램 수, 전시면적, 보관 자료 수, 관람료 등) 데이터로 활용
2023년 국내 문화체육관광 분야 미술관 시설 및 현황통계 데이터 (방문객 등 통계 포함)	한국문화정보원	개별적으로 데이터 파일 요청	관람객 감소 확인 및 분류모델에 사용되는 관람객 감소 요인 (프로그램 수, 전시면적, 보관 자료 수, 관람료 등) 데이터로 활용
소상공인시장진흥공단_ 상가(상권)정보	소상공인시장 진흥공단	https://www.data.go.kr/data/15083033/fileData.do	분류모델에 사용되는 관람객 감소 요인 중 미술관 주변 인프라 데이터로 활용