| **제11회「2023 빅콘테스트」결과보고서** |
| --- |

|  | \* 해당란에 ☑ 표시 | | |
| --- | --- | --- | --- |
| **참가분야** | □ 생성형AI 분야 □ 데이터신기술 분야  □ 정형데이터 분석 분야 □ 비정형데이터 분석 분야  □ 빅데이터플랫폼 활용 분야 | | |
| **세부리그**  \*해당시 체크 | □ 어드밴스드 리그 □ 스타터 리그  \*정형데이터 분석분야에 한함(선택) | | |
| □ 지정주제 리그 □ 자유주제 리그  \*빅데이터플랫폼 활용분야에 한함(선택) | | |
| **개인/팀여부** | □ 개인 □ 팀(총 3 명) | **개인/팀명** | SolMA |
| **지도교사명** | \*스타터 리그에 한함(선택) | | |
| **대표ID** | ppq390@gmail.com | | |



| **결과보고서 작성 안내 사항** | |
| --- | --- |
| **목차**  **(예시)** | **Ⅰ. 개요**  1. 클래식 공연 활성화를 위한 예술의전당 콘서트홀의 효과적 가격 모델 수립  2. 목적 및 필요성  3. 분석 수행 범위  **Ⅱ. 문제 수행 내용**  1. 분석(수행) 절차  2. 분석(수행) 내용 및 결과  **Ⅲ. 주요 결과 및 시사점**  1. 주요 결과 요약  2. 결과 활용 및 시사점 |
| **작성방향** | - 결과보고서는 **30장 내외**로 목차를 준수하여 작성하여야 하며, 필요시 목차  구성에 항목을 추가하여 자유롭게 작성  - 그림 및 도표 등 활용 가능  - 출처 명시(참고 문헌/논문, 이미지, 저자, 사이트 URL 등)  - **생성형AI분야의 경우**, 활용하는 생성형AI 툴 종류 명시 및 생성형AI를 활용한 히스토리 필수 제출(생성형AI에 input/output한 내용이 드러나는 소스) |
| **글꼴 및**  **글자크기** | - 본문 글꼴 : 맑은 고딕  - 대분류[1, 2, 3] 항목 : 13포인트(굸게)  - 중분류[가, 나, 다] 항목 : 12포인트(굵게)  - 소분류[ 1), 2), 3) ] 항목 : 12포인트, 본문내용 : 10포인트 |

| **Ⅰ.** |  | **개요** |
| --- | --- | --- |

**1. 배경**

**클래식 공연 활성화를 위한 예술의전당 콘서트홀의 효과적 가격 모델 수립**

1. 예술의 전당 데이터를 보고 가설 추측

공연예술의 동적가격의 적용가능성

* 동적 가격 제도(dynamic pricing): 특정 기준에 따라 가격을 달리 함으로써 매출을 늘리기 위한 전략.
* 서비스 산업에서도 동적가격이 보편적인 전략으로 변화되고 있음을 감안한다면 공연예술산업에서도 동적가격전략의 적극적인 도입이 필요한 시점으로 보임
* \*동적 가격은 수요와 공급, 그리고 고객의 정보 등을 잘 활용할 수 있는 정보기술을 활용함. 적용가능성:가격공정성 지각을 중심으로Applicability of Dynamic Pricing in Performing Arts:Price Fairness Perspective)

예술경제학으로 주 고객층 양성가능성

* 일반적으로 공연예술의 분야는 공공재의 인식이 있기 때문에 다른 산업군에 비해서 비교적으로 상품의 범용화 단계에 있다고 보기도 어렵고, 탄력성에 관한 상황도 동적가격 도입을 위한 조건을 충족하지 않는 경우가 있음. (가격이 비탄력적임)
* 그러나, 예술업계에서 기부금과 멤버십으로 사용하는 시스템으로 이러한 기부금으로 비영리 공연예술조직을 설립하여 신규 단골 고객을 포섭하는 방식을 사용하고 있음.
* 실제로, 한 연구결과에 따르면, 공연예술 구매자들의 가격 공정성 지각이 동적 가격제의 도입에 따라 유의미한 결과가 있는지를 연구하자, 부정적인 의견이 적을 것이라는 결과가 있음.  
  출저(예술경제학의 이해)

이와 같은 증명을 위해서 할인율과 할인 종류를 기준으로 하여 공연의 예매율을 분석한 후, 동적 가격적용전과 후를 비교 목표

1. 데이터 전처리

* Membershiptype = > 타입1~6에 입력되는 기준을 등급순으로 순차적 배정
* SEAT = > 공연마다 좌석 기준이 다르고 좌석도 여러개여서 하나씩 확인하여 장소별로 분류하고 기준을 정함
* tokenized\_discount: 할인 유형에 따라 8가지의 할인 유형을 만들어 토큰화 변환
* discount\_type\_count: 공연당 사용된 할인 유형의 갯수
* Time\_category: 기준에 따라 오전/오후/저녁으로 분류
* date\_difference: 공연날짜 - 공연예매일
* day\_of\_week: 요일 토큰화
* is\_holiday: 공휴일 유무
* reservation\_rate: 예매율
* corona\_reg: 공연당일 단계별 코로나 규제 단계
* corona\_infected: 서울시 당일 코로나 추가 확진자

3)**상당한 결측치들의 존재를 보고 결측치를 채우는 걸 목표로 정함**

    - age/gender의 대한 결측치를 해결하기 위해서 확정으로 채울수 있게 할인권들로 확정적 유추가 되는 성별/나이의 값을 채워 넣음

   - 멤버쉽과 좌석,공연장을 머신러닝에 학습시키기위해서 각각 토큰화 작업 실행

       타입1~6에 입력되는 기준을 등급순으로 순차적 배정

       - 공연마다 좌석 기준이 다르고 좌석도 여러개여서 하나씩 확인하여 장소별 분류하고 기준을 정함

       - 공연장을 3가지 토큰화로 변환

       - 할인율을에 따라 8가지의 할인 유형을 만들어 토큰화 변환

- genre는 performance\_code를 보고 관련있는 genre로 채워놓고 토큰화 작업

여러가지의 AI 모델을 사용해 결측치 채우기와 정확도 확인  
   사용한 모델:KNN,결정트리,랜덤 포레스트,Gredient Boosting,Imputation Using Multivariate

* 사용모델들중 제일 정확도가 높은 모델로 중심으로 교차검증 후 정확도가 높았던 머신러닝으로 결측치 성별/나이 채우기 실행
* 정확도가 높았던 (랜덤 **포레스트 -> 성별, KNN -> 나이)**  결측치 채우기
* 데이터의 경량화를 위해 스케일링으로 전처리를 실행 후 성별/나이 결측치 채우기 실행
* 이후 완성된 데이터로 가계예측을 위한 여러가지 머신러닝을 사용해서 가격 예측화
* 성별 예측 모델 정확도 -> 73%

나이 예측 모델 정확도 -> 52%

**3.목표 변경**

* 머신러닝들의 정확도가 현저히 떨어지는 문제발생
* 데이터의 결측치가 너무 많아 데이터 학습이 원치않는 방안으로 진행되어 새로운 방법을 찾기로 계획이 변경
* 성별/나이에 대한 데이터가 문제가 되는것으로 판단되어 제외하는 걸로 다시 계획을 변경

| **Ⅱ.** |  | **문제 수행 내용** |
| --- | --- | --- |

**1.분석(수행) 절차 및 문제 분석**

1.age/gender의 대한 결측치 채우기에는 너무나도 큰 결측치

2.정확도가 급격히 떨어지는 데이터가 존재할 것 같은 생각으로 age/gender를 폐기하는 것으로 정함

3.공연별로 예매수를 확인하여 공연장 좌석수를 비교하여 예매율 값을 구함

4.랜덤 포레스트 방식의 머신러닝이 80%의 정확도가 나와 예측모델로 사용하는걸 선택

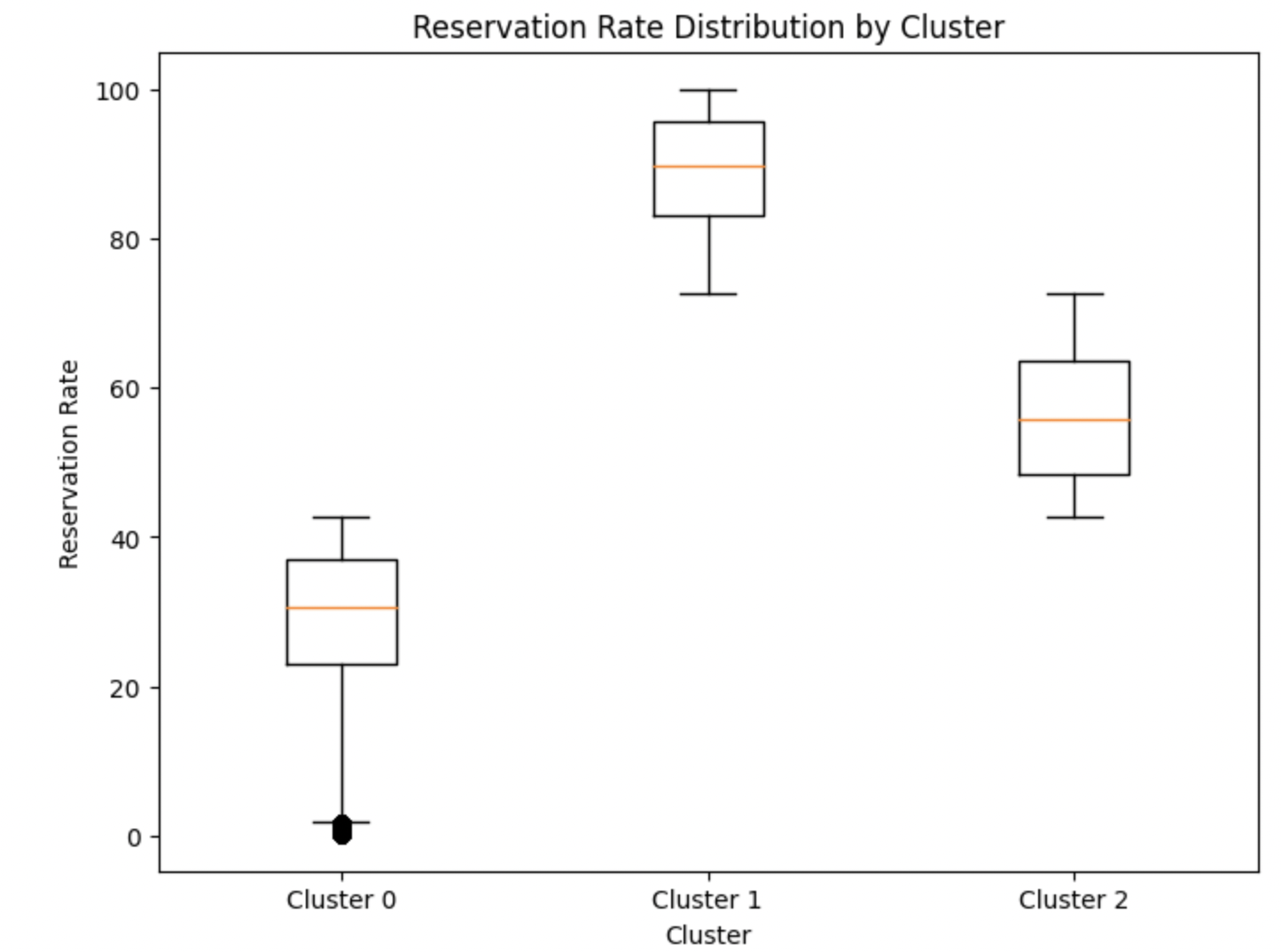
5.예매율과 예매 취소율의 관계성이 있는지 확인

**2.예매율과 모델 선정**

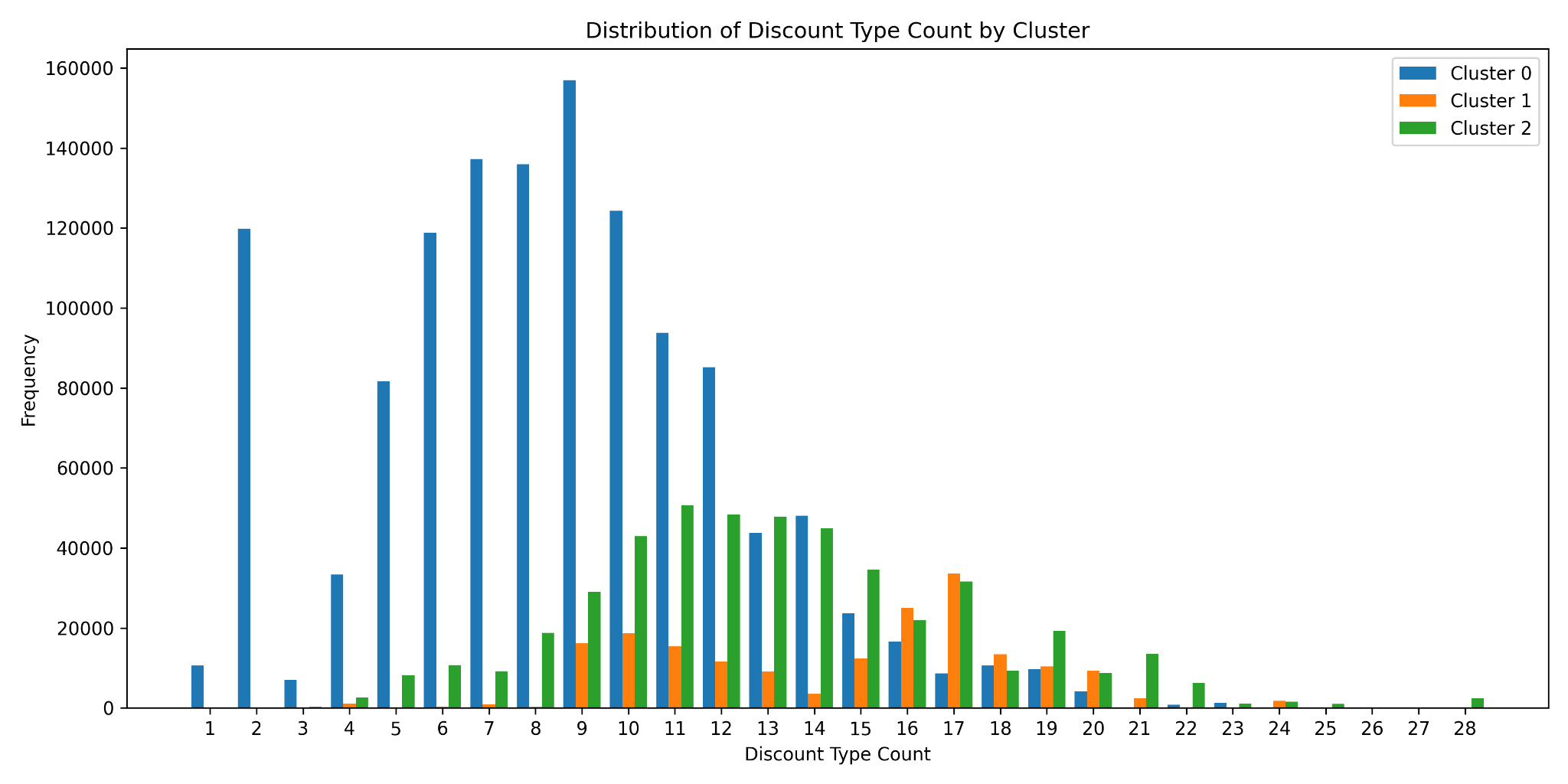
* 공연별 중복값으로 카운트하여 예매된 숫자를 중복값으로 추출하여 예매율 계산  
  (**place,performance\_code,play\_date,ticket\_cancel,play\_st\_time**)
* 공연장마다 좌석수가 정해져있음  
  IBK챔버홀:600석  
  리사이틀홀:354석  
  콘서트홀:2505석  
  정해진 좌석들과 예매가 된 중복값의 수로 예매율의 결과치를 얻음
* 할인율과 예매취소율의 상관관계성을 머신러닝(랜덤 포레스트)으로 확인

**3**.**군집화**

비인기/인기 공연들을 구별하기위해서 예매율로 3가지 형태로 군집화 진행

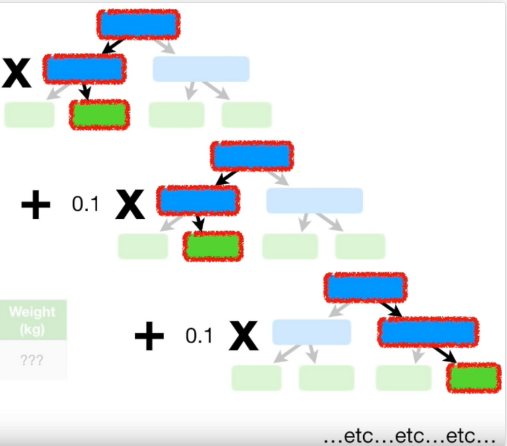


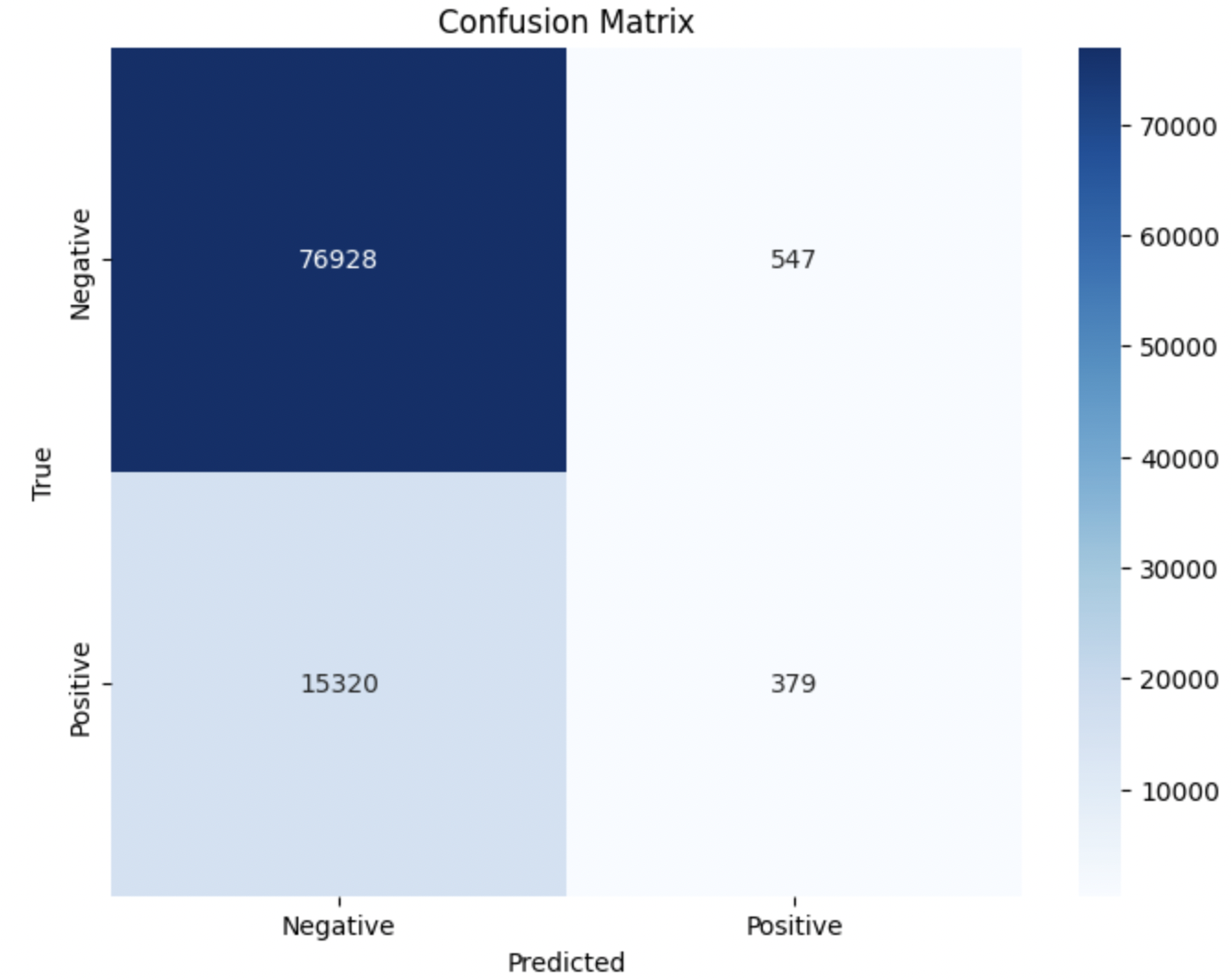
* 예매율을 기준으로 3개의 집단으로 나눔.
* 그중, 동적 가격제를 적용했을 때, 가장 변화가 두드러지는 0군을 택함.
* 0군에서의 예매 취소 샘플 수가 부족하여, 2군에서의 취소된 티켓 샘플 10만 데이터를 추가로 보충함.

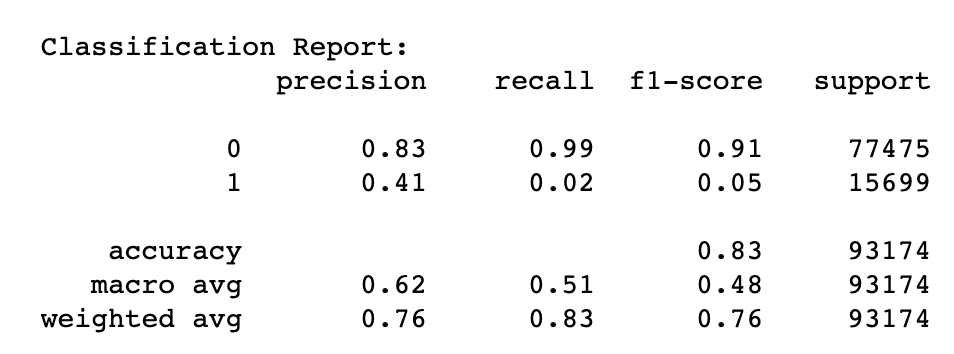


* 예매율이 낮은 0군은 타군에 비해 할인 유형의 갯수가 상대적으로 작음.
* 예매율이 높은 1군은 타군에 비해 할인 유형의 갯수가 상대적으로 높음.

4. 예측 모델 선정



* ticket\_cancel을 예측하는 모델을 선정하기 위해, 결정트리모형, 랜덤 포레스트, 그리디언트 부스팅 등을 훈련시킨 결과, 가장 적합했던 모델은 그리디언트 부스팅 모델이였다.
* 



=> 재현율 / f1-score / support 수치

| **Ⅲ.** |  | **주요 결과 및 시사점** |
| --- | --- | --- |

**1. 분석 결과**

머신러닝 **GradientBoostingClassifier**을 돌려본 결과로 군집화 0번 그룹(비인기 그룹) 그 중 가장 성능이 좋았던 **2구간**의 **5%**를 적용했을 때의 매출 상승 **24. 96% (약 1억원 )**

**2. 결과 활용 및 시사점**

분석된 결과 구매층이 할인율이 높은 것을 선호

중간과정인 기획사,할인업체,초대권 등으로 신규고객이 될 단골층으로 신규포섭 방법 제시 (할인유형의 갯수를 늘림으로써 할인 접근성 증가)  
=> 홍보 효과 증진 및 장기적 고객 유치 가능

예매율이 높은 공연은 예매취소율이 낮으니 인지도가 높은 공연들을 새롭게 만들어내서 소비자들이 눈높이와 가격대를 이익교차점을 찾기

적절한 할인 정책 찾기   
 => 너무 높은 할인율은 오히려 장기적으로는 매출 감소로 이어질 수 있는데 우리 팀이 추가적으로

전처리 작업을 한 공연 장르, is\_holiday(공휴일 유무), day\_of\_week(요일 토큰화)등 여러

요소를 고려하여 유동적으로 관리할 수 있다

모델 업데이트 및 개선 계속하기

=> 시간이 지남에 따라 데이터(티켓 내역)가 업데이트 되므로 그 값에 따라 주기적으로 모델을

업데이트하며 모델을 항상 최신 상태 유지

=> 특정 집단군 고객의 요구와 행동 패턴 변화에 적응할 수 있는 강력한 경쟁력 제공