

클래식 공연 활성화를 위한 가격 모델 개발

팀 명 : 한탕

조원 이름 : 이충은(rladjsdir123@naver.com)

강현규(kgh2000v@korea.ac.kr)

곽동호(kwakdh1@korea.ac.kr)

목차

INDEX

1 단계
개요

2 단계
데이터 전처리

3 단계
EDA

4 단계
가설검정

5 단계
SPI 지수

6 단계
모델링

7단계
결론

분석 배경
분석 목적 및 필요성

결측치 제거
데이터 요약

시각화
변수추출

가설 설정
가설 검정

사용 변수 설정
PCA 및 계산

변수 인코딩
성능 평가 및 선정

결과 요약
시사점

개요

분석 배경

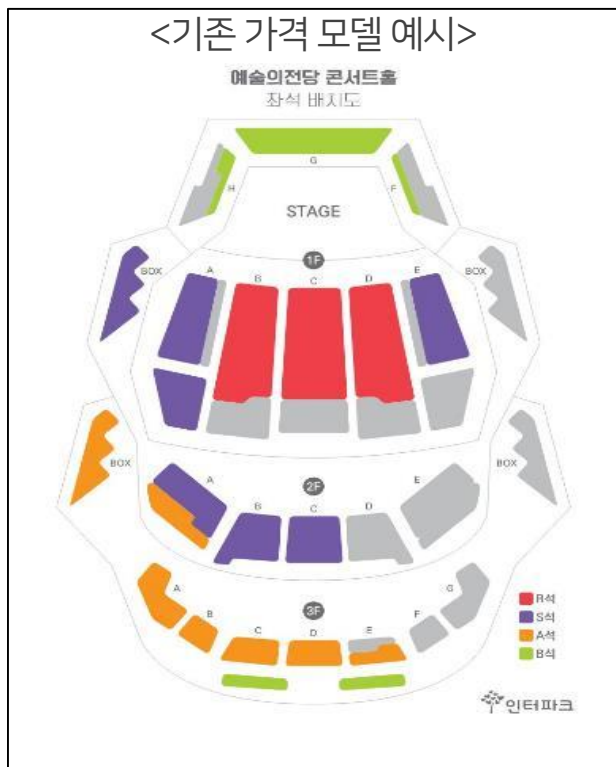
- 코로나 19 팬데믹 이후 클래식 공연 산업은 계속 된 성장세를 보임.



개요

분석 배경

- 그러나, 기존 가격 모델은 명확한 한계점을 지님



- 임의로 좌석 등급을 설정하고, 같은 등급의 좌석에 대하여 같은 가격을 가짐
- 같은 등급의 좌석에 대한 고객 만족도가 상이함



고객들의 가격 대비 공연 만족도 유지 및 수익성 확보를 위한
새로운 가격모델 필요

개요

분석 목적

기존에서 크게 벗어나지 않으면서 선호도를 잘 반영하는
좌석 등급 예측 및 가격 책정 모델

데이터 분석을 통한
좌석 별 선호도 측정



좌석 등급 재분류 및 다양화



등급 예측 및 적정 가격 제시



관객 만족도 증가 및 클래식 공연 활성화

분석절차

데이터 전처리

주어진 데이터셋

#	Variable Name	변수 명	자료형
1	age	나이	bigint
2	gender	성별	text
3	membership_type_1	멤버십종류1	text
4	membership_type_2	멤버십종류2	text
5	membership_type_3	멤버십종류3	text
6	membership_type_4	멤버십종류4	text
7	membership_type_5	멤버십종류5	text
8	membership_type_6	멤버십종류6	text
9	tran_date	예매 거래일자	text
10	tran_time	예매 거래시간	text
11	play_date	공연날짜	text
12	play_st_time	공연시작시간	text
13	seat	좌석번호	text
14	price	최종예매가격	bigint
15	ticket_cancel	예매취소여부	text
16	discount_type	할인내역	text
17	performance_code	공연명 코드	text
18	pre_open_date	선예매시작일	text
19	open_date	예매시작일	text
20	genre	장르	text
21	place	공연장소	text
22	running_time	러닝타임	bigint
23	intermission	휴게시간	bigint
24	member_yn	회원여부	text



전처리된 데이터셋 (ConHall_m)

#	Variable Name	변수 명
1	Genre	장르
2	performance_code	공연명 코드
3	Running_time	러닝타임
4	Intermisssion	휴게시간
5	Days_difference	예약까지 걸린 날짜
6	Undiscounted_price_rounded	할인 전의 가격(반올림)
7	Floor	층
8	Block	블록
9	Row	행
10	Seat_num	좌석 번호
11	Seat_group	좌석 그룹
12	Weekend	주중/주말
13	Pm	오전/오후
14	Member_yn	멤버십 유무
15	Highest_paid_membership	가장 높은 등급의 멤버십
16	Free_membership	무료 멤버십 종류
17	Membership_score	멤버십 점수

분석절차

데이터 전처리

멤버십

Variable Name
membership_type_1
membership_type_2
membership_type_3
membership_type_4
membership_type_5
membership_type_6

→ Highest paid_membership
(6개의 멤버십 중 가장 높은 등급의 유료 멤버십)

공연 관련 정보

pre_open_date의 결측치
(open_date의 값으로 대체) - tran_date → days_difference
(예매 시작일과 실제 예매일의 차이)

Play_date → Weekend(주중/주말)

Play_st_time → Pm(오전/오후)

취소 티켓 및 초대권 처리, 데이터셋 분류

가격이 0인 예매정보(초대권) 제외

유료 티켓 중 취소표 제외

좌석 그룹화

Seat
(1층 C블록 3열 4)

Floor(1)

Block(C)

Row(3)

} Seat_group(1층C3)

분석절차

데이터 전처리

장르

Genre 결측치 제거

샘플 크기 2000개 이하인 장르 제거
(재즈, 기타, 크로스오버, 가족극)

티켓 가격

Discount_type의 할인율 추출

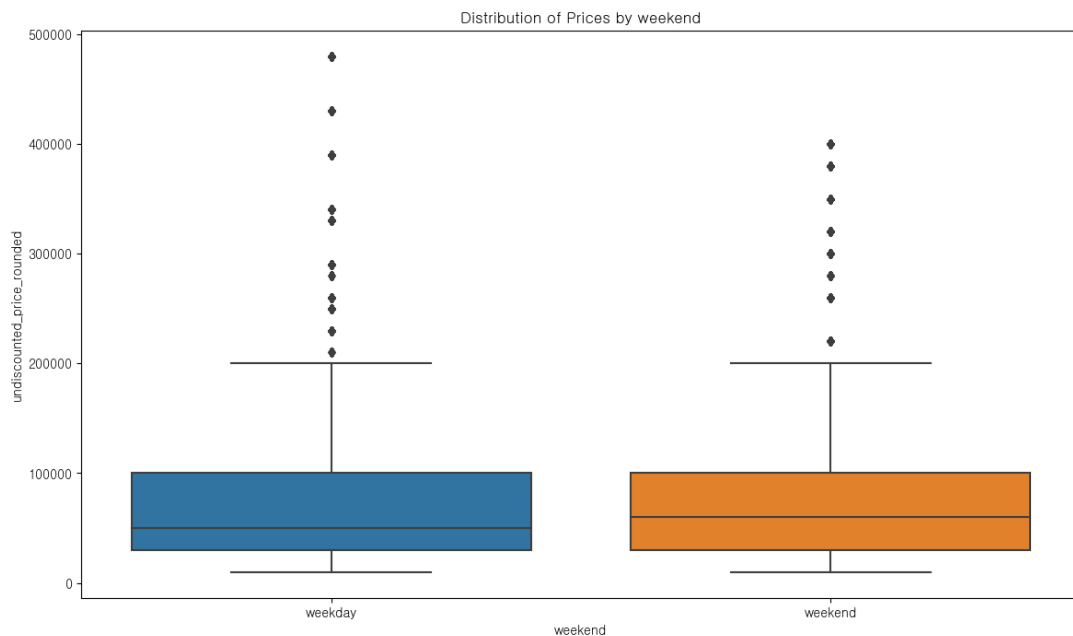
→ $\text{Price} \times 1/\text{할인율}$ & 만원 단위에서 반올림

→ Undiscounted_price_rounded (티켓 원가)

분석절차

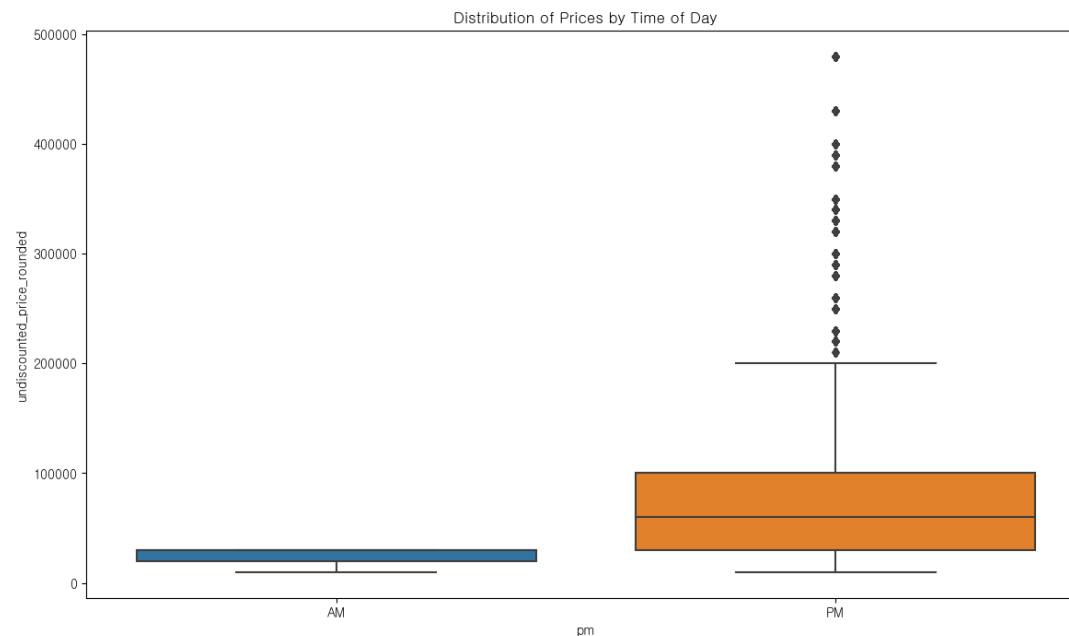
EDA

주중/주말 티켓 가격의 분포



주중 공연의 티켓 가격이 전반적으로 주말 공연보다 낮지만
전체적인 분포에는 큰 차이가 없는 것으로 보임
: 가설검정을 통해 확인해 볼 필요성 있음

오전/오후 티켓 가격의 분포

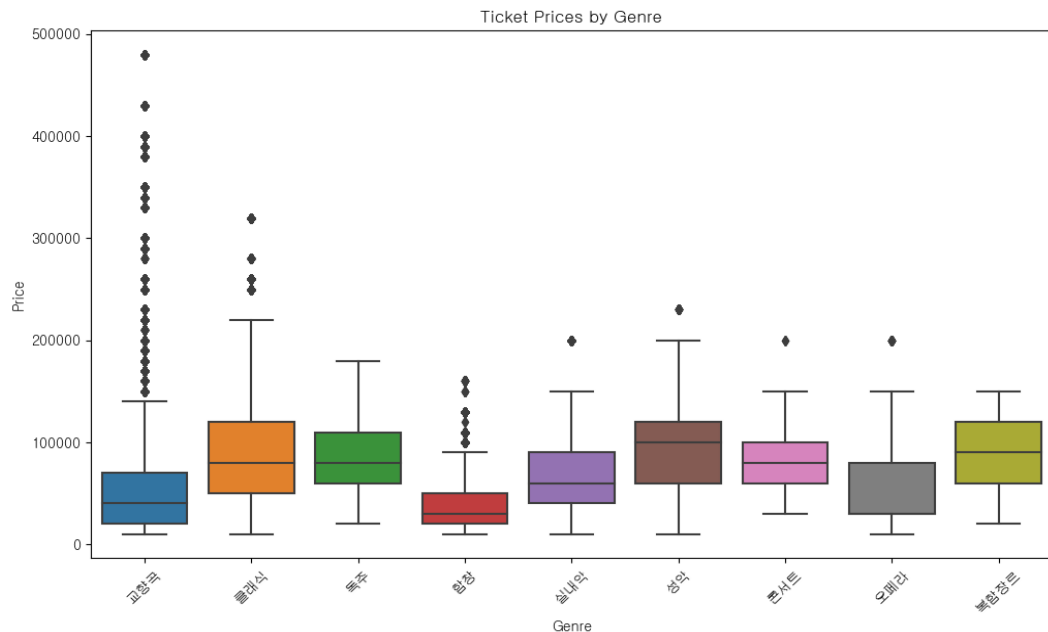


오전 공연이 오후 공연에 비해 횟수가 적으며
티켓 가격에 차이가 있음

분석절차

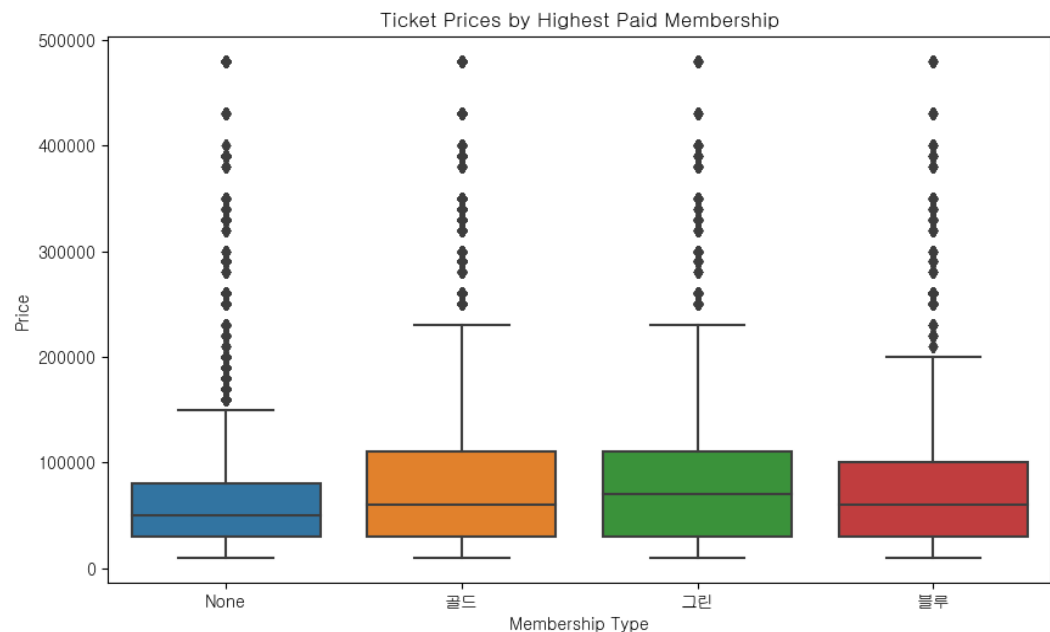
EDA

장르 별 티켓 가격의 분포



장르 별로 평균적인 티켓가격이 다른 것으로 보임

멤버십 등급 별 구매 티켓 가격의 분포



원가가 높은 티켓들의 경우 유료 멤버십을 보유한 사람이 더 많이 구매하는 것으로 보임

분석절차

가설검정

: EDA 결과에 대해 가설검정을 통한 통계적 유의성 확인

EDA를 통한 인사이트

- ① 장르별 가격은 차이가 있다
- ② 주중과 주말의 가격에는 차이가 있다
- ③ 오전과 오후의 가격에는 차이가 있다
- ④ 멤버십에 따른 가격에는 차이가 있다

H1 : 대립가설로 설정



차이가 없을 것이다

H0 : 귀무가설로 설정

분석절차

가설검정

: 정규성 및 등분산성 만족 시 ANOVA 이용

- 정규성 검정 : 샤피로-윌크 검정(Shapiro-Wilk test)

- 등분산성 검정 : Levene's test

① 장르별 가격은 차이가 있다

Genre	W	P-value
독주	0.9534	0.0
클래식	0.9116	0.0
오페라	0.8062	0.0
합창	0.7828	0.0
실내악	0.9204	0.0
교향곡	0.6628	0.0
콘서트	0.9311	2.0435
복합장르	0.9361	2.0326
성악	0.9459	9.8091
#	Stat	P-value
1	576.4419	0.0

② 주중과 주말의 가격에는 차이가 있다

#	W	P-value
0	0.7509	0.0
1	0.8023	0.0
#	Stat	P-value
1	4034166	1.1455

③ 오전과 오후의 가격에는 차이가 있다

pm	W	P-value
1	0.7734	0.0
0	0.6236	0.0
#	Stat	P-value
1	1545.1704	0.0

④ 멤버십에 따른 가격에는 차이가 있다

Membership	W	P-value
None	0.7509	0.0
블루	0.7933	0.0
그린	0.8094	0.0
골드	0.8023	0.0
#	Stat	P-value
1	4230.9462	0.0

➡ 4가지 가정 모두 정규성과 등분산성을 만족하지 않으므로 비모수적 검정을 이용해야함

분석절차

가설검정

: 비모수적 방법론을 이용한 검정 = 크루스칼-왈리스 검정(Kruskal-Wallis H test)

	Stat	P-value
Genre	45947.7261	0.0
Weekend	1137.0372	0.0
Pm	7384.6040	0.0
membership	8923.9338	0.0

비모수적 검정 결과 p-value가 모두 0에 가까우므로

귀무가설을 기각하고 대립가설을 채택한다.

다만 장르와 멤버십의 경우 집단 간 차이를 확인해야 하므로

사후검정을 진행하였으며 결과는 아래와 같다.

- ① 장르별 가격은 차이가 있다 (SPI 산출 시 활용)
- ② 주중과 주말의 가격에는 차이가 있다 (모델링 시 활용)
- ③ 오전과 오후의 가격에는 차이가 있다 (모델링 시 활용)
- ④ 멤버십에 따른 가격에는 차이가 있다 (SPI산출 및 모델링 시 활용)
- ⑤ 골드와 그린 멤버십의 다른 가격에는 유의미한 차이가 없다 (SPI산출 시 활용)

분석절차

가설검정

: 사후검정을 통한 결론도출

EDA를 통한 인사이트

- ① 장르별 가격은 차이가 있다
- ② 주중과 주말의 가격에는 차이가 있다
- ③ 오전과 오후의 가격에는 차이가 있다
- ④ 멤버십에 따른 가격에는 차이가 있다

H1 : 대립가설로 설정



차이가 없을 것이다

H0 : 귀무가설로 설정

분석절차

좌석 선호도 지수(SPI)

분석 목표 : 좌석 별 선호도 수치화 → 좌석 등급 분류 및 가격 책정

선호도를 나타낼 수 있는 변수 설정

- ① days_difference: 해당 좌석이 예매 시작일로부터 실제 예매되기까지 얼마나 걸렸는가?
- ② booking_frequency: 해당 좌석이 얼마나 자주 예약되었는가? (선호되는 좌석은 초대권일 가능성이 높으므로 삭제했던 초대권 데이터 고려)
- ③ membership score: 해당 좌석을 예약한 고객의 멤버십 등급이 어떻게 되는가? (골드/그린 = 2, 블루 = 1, 없음 = 0)

분석절차

좌석 선호도 지수(SPI)

각각의 좌석 별 선호도를 고려하는 경우

: 공연 별 특성, 이상치에 크게 영향을 받을 수 있음



각 블록의 같은 열을 하나의 좌석으로 간주 (Robust)

ex) 1층 C블록 1열 3, 5, ... = 1층 C블록 1열

각 열에 포함 되는 각 좌석들의

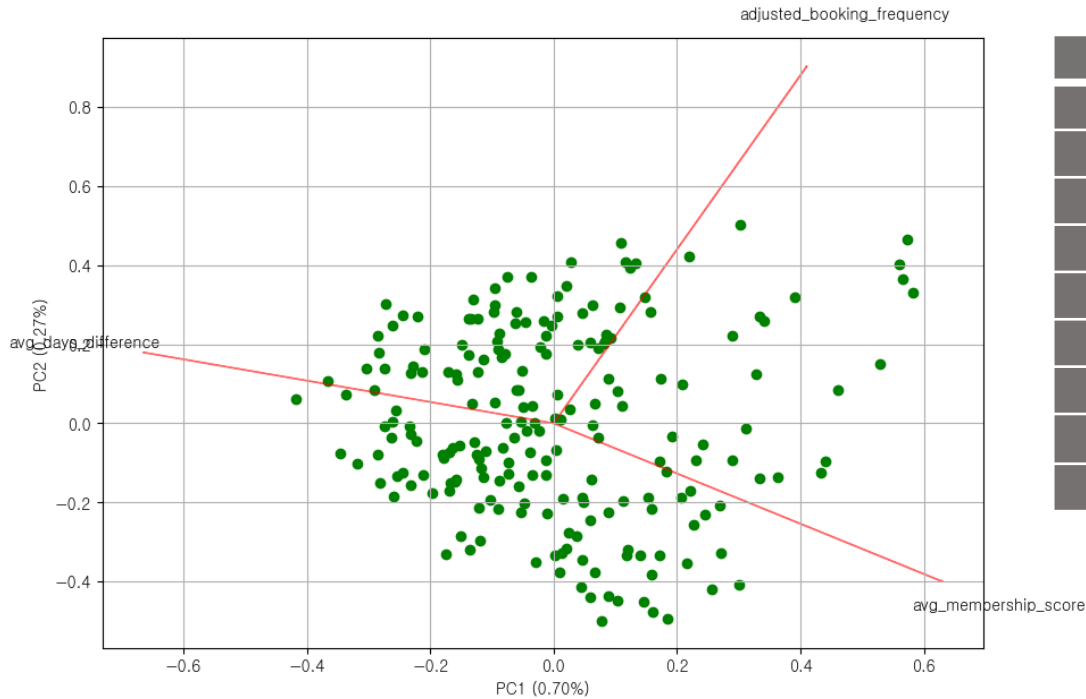
days_difference, booking_frequency, membership_score의 평균 계산

	floor	block	row	booking_frequency	avg_days_difference	avg_membership_score	seat_count	adjusted_booking_frequency
0	1층	A	1	1263	20.404711	0.982869	9	140.333333
1	1층	A	10	1947	23.408582	1.104478	11	177.000000
2	1층	A	11	1877	26.499102	1.016158	11	170.636364
3	1층	A	12	1998	29.076636	1.082243	12	166.500000

분석절차

좌석 선호도 지수(SPI)

3개의 변수를 이용해 주성분분석(PCA) 진행



첫번째 주성분(PC1)이 전체 분산의 70%를 설명하고 있음

장르별로 첫번째 주성분의 계수 확인

Genre	Adjusted_booking_frequency	Avg_days_difference	Avg_membership_score
독주	0.5762	-0.5730	0.5829
클래식	0.1906	-0.6767	0.7111
오페라	-0.2356	-0.6926	0.6817
합창	0.4933	-0.6203	0.6098
실내악	0.4237	-0.6429	0.6381
교향곡	0.3734	-0.6609	0.6510
콘서트	0.3497	-0.6696	0.6552
복합장르	0.5417	-0.5034	0.6731
성악	0.5108	-0.6017	0.6140

일반적으로 booking_freq와 membership_score에 대해서는 양의 계수를, days_difference에 대해서는 음의 계수를 가짐

일반적으로 예매빈도와 멤버십 소유 고객들의 좌석 선택 비율이 높을수록, 예매 시작일로부터 예매까지 걸린 시간이 짧을수록 선호되는 좌석



PC1 = 선호도를 나타내는 주성분: 이를 이용해 SPI 산출

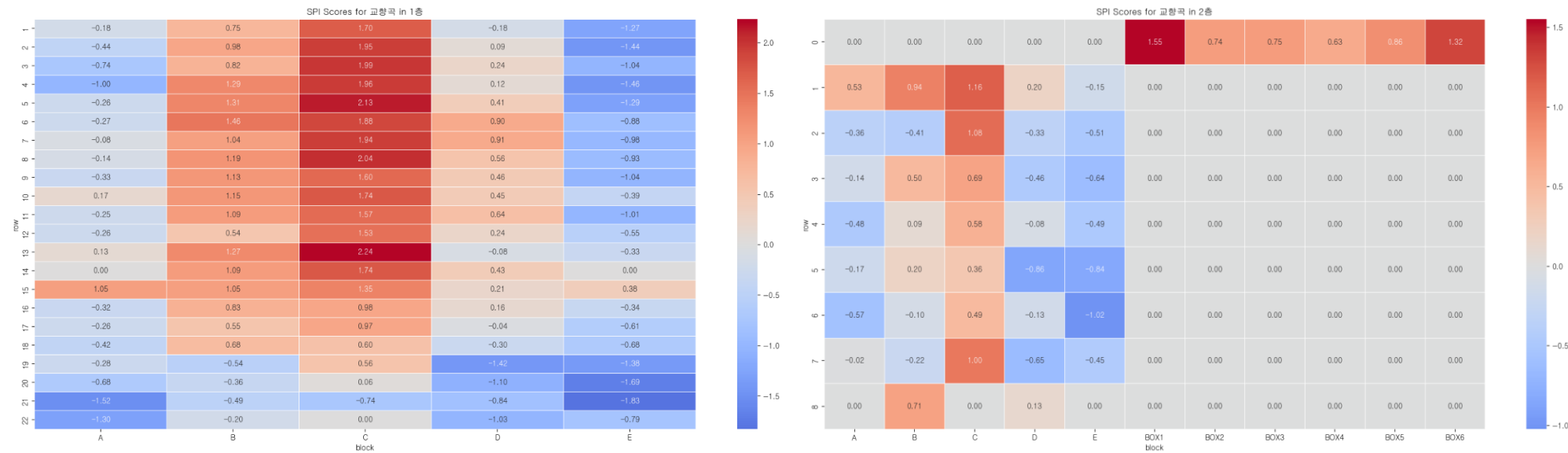
분석절차

좌석 선호도 지수(SPI)

좌석 선호 지수(Seat Preference Index, SPI) 산출식

$$SPI = [PC1] \times [booking_freq, avg_day_diff, avg_membership_score]$$

교향곡 장르 1층, 2층에 대한 SPI 산출 결과



현재 좌석 등급과 전반적으로 유사함: SPI가 좌석의 선호도를 잘 반영하고 있음

같은 등급 내 좌석들이 서로 다른 SPI를 가짐: 현재 가격 정책은 좌석의 선호도를 완전히 반영하지 못하고 있음

분석절차

좌석 선호도 지수(SPI)

결론 : SPI를 활용하면 기존 좌석 등급에서 크게 벗어나지 않으면서도, 기존 가격 보다 적절하게 선호도를 반영한 가격 책정 가능

장르별로 10000원 단위로 반올림한 티켓 그룹화

그룹 별 SPI의 평균, 표준편차 계산

SPI지수를 바탕으로 가격 상향 및 하향 조정

- 장르별로 티켓 가격이 다르다는 가설 검정 결과 반영
- 공연별로 티켓 가격이 상이하다는 일반적인 사실 반영

: 장르와 공연의 영향을 제외하고
좌석 자체의 선호도만 반영 가능

$$adjusted\ price = (original\ price) * (1 + 0.1 * \frac{d_{ij}}{std_{ij}})$$

SPI가 평균에서 1표준편차만큼 벗어날 때마다 10%씩 조정
: 기존 좌석 등급을 유지하면서도 선호도 반영 가능

기존 가격	SPI
100,000	2.1
100,000	1.0
100,000	-0.8
50,000	1.4
50,000	-1.4
50,000	0.2



그룹 별 평균 및 표준편차 계산

그룹 별 평균 및 표준편차 계산

조정 가격
110,000
100,000
90,000
55,000
50,000
45,000

분석절차

가격 등급 예측 모델링

새로운 공연의 입력

- ① 장르
- ② 공연 정보
- ③ 주말/오전 오후 여부
- ④ 좌석 정보



출력

각 좌석에 대한 예상 등급

1. 새로운 퍼포먼스에 대해 좌석등급을 예측
2. 새로운 퍼포먼스에 대해 가격 기준을 제시

분석절차

가격 등급 예측 모델링

: 파생변수 추가 및 등급 분류

인기 지수 제작

공연의 모든 좌석의 days_difference 평균

Popularity 변수로 추가

Very unpopular ~ Very popular 까지 5개의 척도로 구분, 범주화

도메인 지식에 따라 1~5까지 새로운 공연의 인기도 입력 가능

등급 분류

좌석 등급을 특정 비율과 그 범위를 설정하여 설정

선호도를 세부적으로 반영할 수 있도록
임의의 등급 책정

저평가 및 고평가된 좌석 등급 자동 조정

등급	비율
RR	0.05
R	0.25
RS	0.2
S	0.2
A	0.15
B	0.15

등급의 개수와 비율을 자유롭게 조정하여 입력 가능

분석절차

가격 등급 예측 모델링

: 범주형 변수 인코딩 및 Train-Test set 분리

순서 범주형 변수

popularity



Labelencoder 사용
1,2,3,4,5 수치형으로 변환

명목 범주형 변수

genre
seat group



pd.get_dummies 사용
dummy variable 변환

Train-Test set 분리



Cross-Validation

분석절차

가격 등급 예측 모델링

: 데이터 구조 구조

입력: 장르, 좌석, 인기도 등

row	genre_독주	genre_클래식	...	seat_group_1층 A 2	seat_group_1층 A 3	...	popularity
1	1	0		1	0	...	5
2	0	1		0	1		3



출력: 좌석 등급

R A

RR B

분석절차

가격 등급 예측 모델링

: 사용 모델 및 accuracy

사용 모델

RandomForest

다수의 결정트리 학습,
예측을 평균 내어 결과값 도출

CV



Accuracy

0.9330

Xgboost

트리 구축
이전 트리의 오류를 줄이는 방향 학습

CV



0.5394

LightGBM

리프 중심 트리 분할 사용

CV



0.7092

: RandomForest 모델 선택, 선호도가 반영되어 조정된 각 좌석의 등급을 93.3%로 예측

등급 분류 개수, 비율의 변경에도 일관성 있는 정확도 유지!

분석절차

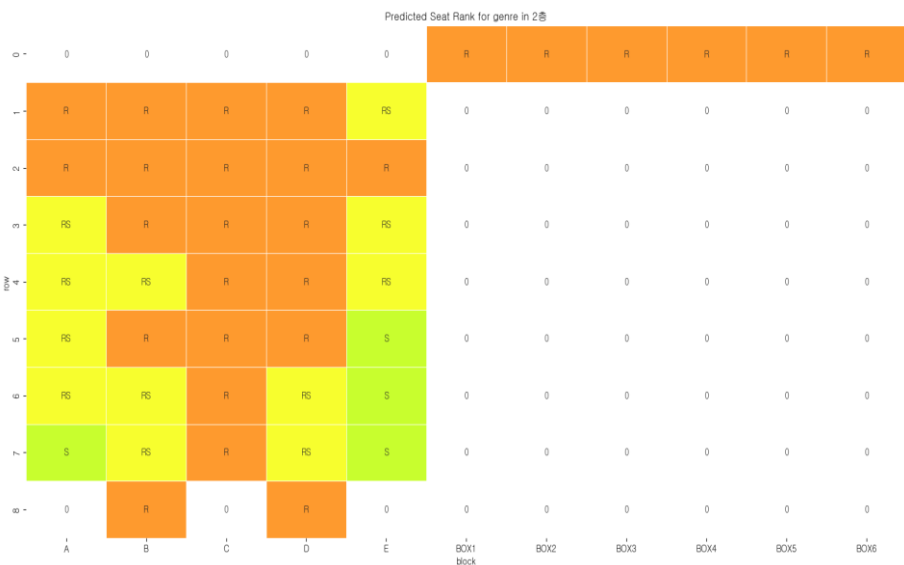
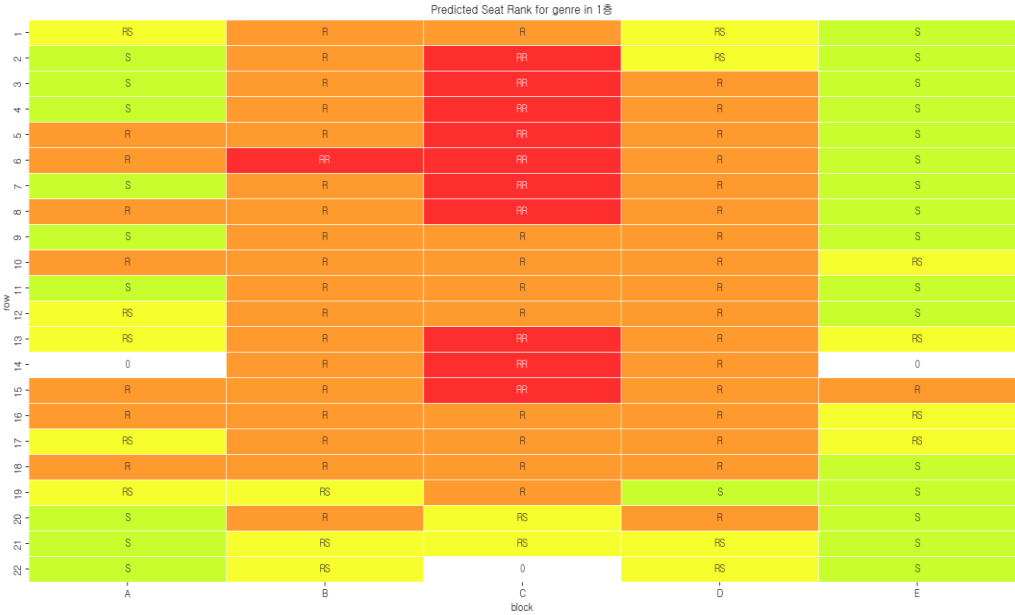
가격 등급 예측 모델링

새로운 공연

교향곡 120분 공연 15분 휴식

주말 오후 공연 인기도 높음 예상

공연에 대한 1, 2층 좌석 예측 결과



분석절차

가격 등급 예측 모델링 : 가격 기준표 제시

SPI로 조정한 가격 장르와 등급 구분에 따라 평균을 내어 표로 정리

예시 가격표

Genre	B	A	S	SS	R	RR
독주	20000	30000	50000	70000	110000	130000
클래식	40000	50000	70000	100000	120000	140000
오페라	50000	60000	80000	100000	120000	140000
합창	40000	50000	80000	110000	130000	150000
실내악	30000	40000	60000	80000	100000	120000
교향곡	30000	30000	40000	60000	80000	100000
콘서트	40000	60000	80000	90000	110000	130000
복합장르	40000	50000	70000	100000	130000	150000
성악	20000	20000	30000	40000	60000	70000

각 공연에 대한 정보를 조합해 최종 등급별 가격 계산에 활용 가능!

결론

주요 결과

① 코로나 등 각종 변인에 대해 robust한 분석 추구



분석 단위를 개별 좌석이 아닌
하나의 열을 묶은 좌석 그룹으로 설정

② 세운 가설을 기반으로, 장르, 주말/주중,
오전/오후, 멤버십을 모델링에 사용할 변수로 지정



멤버십 간 구매 가격 분포에 통계적으로 유의한
차이가 없음을 확인하고 같은 수치로 인코딩

③ PCA를 이용한 가중치를 통해 각 장르별로
모든 좌석 그룹에 대한 SPI를 산출



우리의 직관적인 좌석 선호도와 같은 방향성을 띠



SPI가 선호도를 합리적으로 반영

결론

주요 결과

④ 산출된 SPI를 기반으로 기존 가격을 조정



선호도가 가격에 직접적으로 반영

⑤ 반영된 가격을 기반으로 좌석 등급을 나누어
랜덤 포레스트 모델을 통해 새로운 공연에 대해
모든 좌석의 예측 좌석 그룹이 출력되게 설정



93.3%의 정확도로 새로운 공연에 대해 선호도를 반영한
일관성 있는 등급 구성과 분류를 제시

⑥ 등급 분류를 새로 하더라도 각 장르별로 새로운 좌석 분
류의 기준이 되는 **가격 기준표**를 제작 가능



등급 분류 후 실제 가격을 선정할 때에 기준이 됨

↓
과거 가격을 선호도를 반영한
가격으로 조정했기 때문

결론

결과 활용 및 시사점

▪ SPI 활용 가격 모델의 효과

- ① 합리적이고 다양한 좌석 선택지로 인한 **고객 만족도** 상승
- ② 관객들의 구매의사 증대로 인한 **매진율 및 수익** 증대
- ③ **지속적인** 모델 업데이트를 통한 **유연한 전략 수립** 가능성
- ④ 클래식 공연의 **활성화** 및 새로운 고객 유입

감사합니다

Team. 한탕

참고문헌

윤정연. (2020). 코로나 이후의 클래식 공연. 한국예술연구, (29), 73-90.

조권중(2004), 서울시민의 문화에 대한 인식과 문화예술 향수 수준, 서울 연구 포커스, 23호

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001).

Dinno, A. (2015). Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal*, 15(1), 292–300.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794.

Satoshi Kawase, Factors influencing audience seat selection in a concert hall: A comparison between music majors and nonmusic majors, *Journal of Environmental Psychology*, Volume 36, 2013, Pages 305-315, 0272-4944,