JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 3, pp. 739-747, Mar. 2025

Check for updates

머신러닝을 이용한 뮤지컬 잔여 좌석 예측 프레임워크를 통한 인자 분석

김 민 선[1†] · 서 준 혁[2†] · 최 승 호[3*]
[1†] 한성대학교 AI응용학과 학사과정 [2†] 한성대학교 지능시스템학과 학사과정 [3]한성대학교 기초교양학부 조교수

Factor Analysis with a Framework for Predicting Musical Residual Seats Using Machine Learning

Min-Sun Kim[1‡] · Jun-Hyuk Seo[2‡] · Seoung-Ho Choi[3*]

[1‡] Bachelor's Course, Department of AI Application, Hansung University, Seoul 02876, Korea
[2‡] Bachelor's Course, Department of Intelligent System, Hansung University, Seoul 02876, Korea
[3]Assistant Professor, Collage of Liberal Arts Faculty of Basic Liberal Art, Hansung University, Seoul 02876, Korea

[요 약]

뮤지컬 배우의 인기도와 티켓 판매율은 밀접한 관련이 있다. 하지만 스타 캐스팅으로 인한 제작비 상승으로 티켓값 인상과 뮤지컬 접근성 하락이라는 악순환이 계속되고 있다. 이를 방지하기 위해, 티켓 판매에 영향을 미치는 주요 인자를 식별하고, 머신러닝 기반의 뮤지컬 잔여 좌석 예측 프레임워크를 통해 인자 분석 방법을 제안하였다. 제안한 프레임워크는 뮤지컬 잔여 좌석 예측에 큰 영향을 주는 인자가 무엇인지에 대해서 분석하였다. 이를 검증하기 위해 통계적 분석과 함께 9가지 머신러닝 모델을 이용하여 실험을 진행했다. 인자 분석을 위해 SHAP을 활용하여 모델의 예측 결과를 분석했다. Gradient Boosting Regressor 모델이 가장 좋은 예측 성능을 보였다. 할인율이 티켓 판매에 가장 중요한 요인임을 확인했다. 제안한 프레임워크를 통해 잔여 좌석을 효율적으로 예측하고, 잔여 좌석을 예측하는 데에 있어 가장 중요한 인자가 무엇인지를 확인했다.

[Abstract]

The popularity of musical actors and ticket sales are closely related. However, owing to the increase in production costs caused by star casting, a vicious cycle of ticket price increases and decreased accessibility to musicals continues. To prevent this, we identify the major factors affecting ticket sales and propose a factor analysis method via a machine learning-based musical remaining seat prediction framework. The proposed framework analyzes the factors that have a significant impact on the prediction of remaining musical seats. To verify this, we conducted an experiment using nine machine learning models along with statistical analytical techniques. We analyzed the prediction results of the models using Shapley Additive exPlanations for factor analysis. The Gradient Boosting Regressor model showed the best prediction performance. We confirmed that the discount rate was the most important factor in ticket sales. Through the proposed framework, we efficiently predicted remaining seats and confirmed the most important factors in predicting remaining seats.

색인어 : 뮤지컬 잔여 좌석 예측, 머신 러닝, 프레임워크, SHAP, 요인 분석

Keyword : Musical Residual Seat Prediction, Machine Learning, Framework, Shapley Additive exPlanations, Factor Analysis

http://dx.doi.org/10.9728/dcs.2025.26.3.739

Received 23 December 2024; **Revised** 06 February 2025
Accepted 26 February 2025

‡ **These authors contributed equally to this work**

***Corresponding Author; Seoung-Ho Choi**

Tel: ※ 개인정보 표시제한
E-mail: jcn99250@naver.com

Ⅰ. 서 론

뮤지컬은 사회적 변화와 밀접한 관련이 있다. 다양한 예술 형식 중의 하나로써 사회의 문화적, 정치적, 사회적 풍토를 반영하고 형상화한다. 따라서 뮤지컬은 사회적 변화와 함께 발전해왔으며, 사회적 이슈와의 상호작용을 통해 더욱 풍부하고 다채로운 예술 경험을 제공한다. 이러한 뮤지컬은 감성 중심의 문화적 창의성이 중요한 시대적 배경 속에서, 현대인들의 소득 수준 향상과 여가 시간이 증가함에 따라 삶의 질이 개선되고, 이에 따라 예술 향유에 대한 소비자의 욕구가 커지면서 공연예술에 대한 수요도 증가하게 된다[1]. 이에 기반해서 2022년 티켓 매출은 4,253억원을 기록한 데 이어, 2023년 상반기 매출은 2022년 해 대비 42% 늘어났다. 또한 2022년 잠정 집계된 뮤지컬 총매출액은 4,590억원으로 공연 시장 내 비율은 36.2%에 달했다[2].

앞선 예시를 통해서 볼 수 있듯이 뮤지컬 관람을 즐기는 사람들이 점점 증가하고 있다. 뮤지컬을 많이 보게 되는 이유 중에는 뮤지컬은 노래와 춤, 연기가 어우러져 연극보다 재미있으며, 극적 진행에 음악이 적극적으로 강조된다는 점이 있다. 또한, 원작에선 한계가 있더라도 작품을 재탄생시키는 과정에서 충분히 변화를 꾀할 수 있다. 또한 무대를 통해 타인의 삶을 살아볼 수 있는 경험 때문도 있다[3].

월 7회 이상 뮤지컬을 관람하는 관객의 경우 뮤지컬 관람의 주요인이 뮤지컬 배우였으며[4], 이는 캐스팅과 뮤지컬 배우의 인기도는 밀접한 관련이 있음을 시사한다. 실제로 티켓 예매 사이트에 접속해보았을 때, 배우의 인기도와 공연의 매진율이 비례함을 알 수 있다.

이러한 이유로 뮤지컬 업계에서 '스타 캐스팅'은 필수불가결적 요소가 되었다[5]. 하지만, 스타 캐스팅으로 인한 제작비 증가는 티켓값 상승. 이른바, 티켓플레이션(티켓+ 인플레이션)을 야기하게 되었다[6]. 공연예술통합전산망(KOPIS) 데이터를 분석한 결과, 티켓예매 수는 지난해에 비해 10만개 이상 감소했지만, 매출액은 20억 정도 상승했다[7]. 단기적으로 보았을 땐 뮤지컬 시장의 성장을 나타내는 지표로 볼 수 있겠지만, 장기적인 관점으로 보았을 땐 뮤지컬 산업의 지속가능성을 위협하는 요인으로 작용할 수 있다. 따라서. 티켓값 상승과 뮤지컬 접근성 하락의 악순환을 방지하기 위해 뮤지컬 잔여 좌석 예측에 중요한 인자를 찾는 것이 필요하다. 본 논문의 공헌도는 아래와 같다.

머신러닝을 이용한 뮤지컬 잔여 좌석 예측 프레임워크를 이용하여 인자 분석할 수 있는 방법을 처음으로 제안한다.

뮤지컬 잔여 좌석 예측을 수행하기 위해서 직접 뮤지컬 티켓팅 관련 정보를 수집했다.

수집한 데이터를 통계적으로 분석하기 위해서 상관관계 분석 그리고 종속변수와 독립변수 간의 연관관계 분석을 수행했다.

뮤지컬 잔여 좌석을 예측하는데 있어서 9가지 머신러닝을 이용하여 어떤 변수가 중요한지를 SHAP을 이용하여 분석했다.

머신러닝 예측 결과의 신뢰성을 확보하기 위해 Q-Q plot을 사용하여 모델의 성능을 확인했다.

2장에서는 관련 연구 관련해서 설명한다. 3장에서는 제안 방법 관련해서 이야기한다. 4장에서는 실험방법에 관해서 이야기한다. 5장에서는 실험 결과, 6장에서는 결론에 관해서 이야기한다.

Ⅱ. 관련 연구

예측 모델링은 과거 데이터를 기반으로 미래의 사건을 예측하는 기법이다. 다양한 분야에서 활용되며, 특히 경제학이나 마케팅 분야에서 적용된다. 진동현의 연구[8]에서는 한국은행 경제통계시스템의 데이터를 활용하여 국내 주식 포트폴리오의 수익률을 예측했다. 김찬수 등의 연구[9]에서는 한국 주식시장의 고빈도 데이터를 활용하여 장중 주가 움직임을 예측하기 위한 머신러닝 모형을 개발하고, 수익성을 분석했다. 서민석의 연구[10]에서는 LIME을 활용하여 한국어 텍스트 데이터를 양과 음의 의미를 가진 벡터로 변환했다. 이를 기반으로 주식시장 예측모형을 구축했다. 이성주의 연구[11]에서는 중소 전자상거래 판매상들의 프로모션 데이터를 이용하여 프로모션 시행 시 매출의 변화를 예측했다. 데이터에는 다섯 회사의 프로모션 정보가 포함되었으며, 품목별 특성과 프로모션의 복잡한 영향을 분석했다. 안세희와 정재윤의 연구[12]에서는 M5 Competition의 Walmart 데이터셋 중 CA_1 매장의 판매량 데이터를 이용하여 판매량을 예측했다. TFT (Temporal Fusion Transformer) 모형을 적용하여 실험을 진행했고, M5 Competition에서 우승한 DRFAM 기법과 비교하여 성능을 평가했다. 이 연구는 TFT 모델의 성능이 DRFAM 기법보다 평균적으로 더 우수하다는 점을 입증했다. 윤동준의 연구[13]에서는 도서 검색량, 작가 인지도, 날씨, 주말/공휴일, 출판 시즌 등의 변수를 고려하여 신간 도서의 판매량을 예측했다. 35만 종의 도서 정보와 3,572만 건의 판매 데이터를 활용해 모델을 개발하고 실험한 결과, 기존 예측 방법보다 높은 정확도를 보여주었다.

데이터 기반 접근법은 이미 다양하게 적용되고 있지만, 그 중 공연 예술 분야에서는 최근 들어 그 중요성이 강조되고 있다. 뮤지컬, 연극, 콘서트와 같은 공연 예술은 과거에는 주로 예술적, 감성적 측면에서 평가되었지만, 최근에는 데이터 분석을 통해 관객의 수요를 예측하고 관리하는 연구가 이루어지고 있다[14].

예를 들어, 프로모터 기업 마이 뮤직 테스트는 아티스트에 대한 데이터를 수집하여 티켓 판매량을 예측한다. 이를 바탕으로 공연의 규모를 정하고, 공연의 매출을 최대화한다[15]. M. G. S et al. 의 연구[16]에서는 다변량 회귀 분석과 인공신경망(ANN)을 사용하여 클래식 음악 유형, 콘서트 스폰서

십 여부 등 티켓 판매에 영향을 줄 수 있는 데이터 특징을 사용하여 클래식 공연 티켓 판매량을 예측했다. 하지만 이러한 프로그램은 회차마다 배우나 티켓 할인율, 이벤트 여부가 달라지는 뮤지컬과 양상이 다르며, 지금까지 잔여 좌석을 예측하는 연구는 아직 없다.

서범근의 연구[17]에서는 Random Forest Regressor, XGB Regressor, LGBM Regressor 등의 모델을 사용하여 유튜브 콘텐츠의 인기를 예측하는 프로그램은 제안했고, 김진웅의 연구[18]에서는 Summary plot을 사용하여 모델에 큰 영향을 주는 변수에 대해 분석했다.

기존 연구에서는 공연 예술 관련 데이터 분석이 주로 티켓 판매량 예측(예: 콘서트, 스포츠 경기 등)과 관련되어 있었다. 그러나 본 연구는 뮤지컬의 잔여 좌석 예측에 초점을 맞춘다는 점에서 차별점을 가진다.

이와 같은 기존 연구의 기법을 확장하여 뮤지컬 티켓팅 데이터를 기반으로 잔여 좌석 예측 모델을 개발한다. 데이터 샘플링 기법과 사용한 모델에 따라 결과를 비교하고, SHAP을 통해 변수의 중요도를 분석함으로써, 잔여 좌석 수에 영향을 미치는 주요 요인을 식별하고 뮤지컬 접근성을 개선하는 데 기여하고자 한다.

Ⅲ. 제안 방법

그림 1은 잔여 좌석 예측을 위해 제안하는 프레임워크이다. 첫 번째 단계에서는 데이터 취득과 취득한 데이터에 대한 전처리를 진행한다. 그리고 취득한 데이터에 대한 탐색적 데이터 분석(Exploratory data analysis)을 진행한다. 두 번째 단계에서는 데이터에 대한 샘플링을 진행하고, 세 번째 단계에서는 머신러닝 모델을 학습한다. 학습한 머신러닝 모델을 R Score, MSE, 그리고 MAE를 이용하여 평가한다. 네 번째 단계에서는 머신러닝 모델의 인자 분석을 위해서 SHAP의 Decision plot, Dependency plot, Summary plot 그리고 Q-Q plot을 수행한다.

우리는 잔여 좌석 예측을 하기 위해서 2024년 4월부터 5개월간 데이터를 직접 크롤링해서 수집한다. 각 변수는 seat, date, cast1, cast2, cast3, cast4, weekend, day, dc, evt이다. seat은 잔여 좌석 수, date는 진행된 공연의 날짜, cast1-cast4는 배역, weekend는 요일, day는 낮 공연 혹은 밤 공연, dc는 티켓 할인율, 그리고 evt는 커튼콜과 같은 이벤트의 진행 여부를 의미한다.

획득한 데이터 셋을 증강하여 머신러닝 모델에 입력하기 위해서 데이터 샘플링 기법 SMOTE, Random 그리고 Borderline SMOTE에 대해서 비교 분석한다.

이를 기반으로 머신러닝 모델의 데이터 구성을 Non-stratified와 Stratified로 나누어 실험한다. Non-stratified는 sklearn 라이브러리에서 제공하는 train test split 함수에 seat 변수만 포함하여 훈련 및 테스트 데이터를 분할한 경우이며, Stratified는 sklearn 라이브러리에서 제공하는 train test split 함수에 seat 변수 및 musical 뮤지컬 명까지 포함하여 훈련 및 테스트 데이터를 분할 한 경우이다.

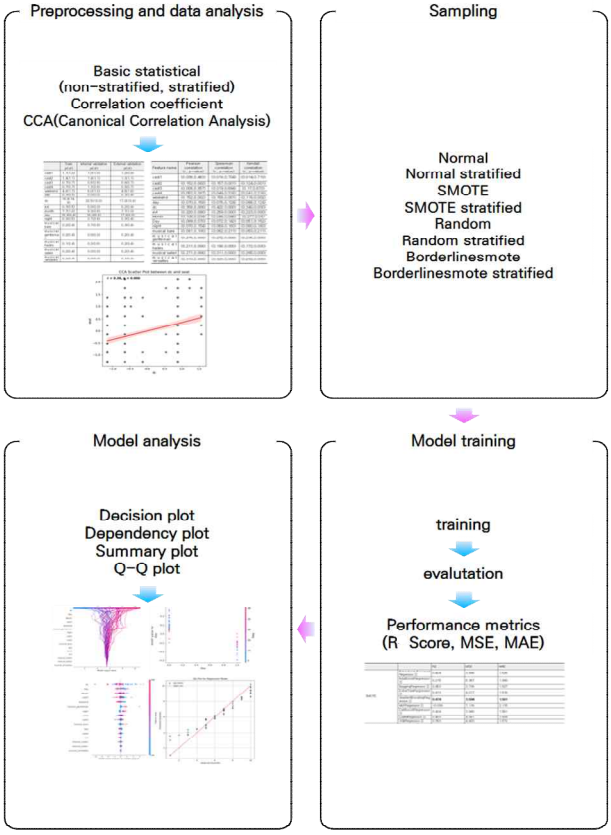

그림 1. 제안하는 프레임워크
Fig. 1. Proposed framework

머신러닝 모델의 분석을 위해 SHAP의 Decision plot, Dependency plot, 그리고 Summary plot을 수행한다.

Ⅳ. 실험 방법

수집한 데이터는 다음과 같이 전처리했다. 변수 day는 day와 night로 나누어 낮 공연인지, 밤 공연인지 표시했다. 하루에 공연이 두 번 진행되는 경우, 두 번째 공연은 밤 공연으로 처리했으며, 그 외 모든 경우는 낮 공연으로 처리했다. 변수 date는 Month와 Day로 나누어 월과 일을 표시했다. 크롤링 대상이 된 5가지 뮤지컬을 각각 musical '뮤지컬 제목'으로 표기하여 원-핫 인코딩 했다. 따라서 독립 변수로 사용한 변수는 cast1, cast2, cast3, cast4, weekend, day, dc, evt, Month, Day, night, musical bare, musical gentleman, musical hades, musical salieri, 그리고 musical versailles로 구성했다.

종속변수로는 seat을 활용했다. 이는 전체 좌석 수에 대한 잔여 좌석 비율을 10등급으로 나눈 변수로, 숫자가 커질수록 티켓 판매율이 낮음을 의미한다. 종속변수와 독립변수 간 상관관계를 파악하기 위하여 상관 분석(Correlation Analysis)를 진행했다. 가장 널리 사용되는 피어슨(Pearson) 상관계수를 통하여 독립변수와 종속변수 간 선형관계가 있는지 파악했다. 단, 피어슨 상관계수는 변수들이 정규 분포를 가정한다는 전제하에서 해석이 적합한 방식이다. 그러므로, 정규분포를 만족하지 않는 경우에서 사용하는 스피어맨(Spearman) 상관계수와 켄달(Kendall) 상관계수를 활용하여 단순한 선형적 특성에 국한되지 않고, 비 정규성 혹은 순위 차원에서 의미있는 패턴이 있는지 확인했다.

또한, 종속변수와 독립변수 간 상관관계를 파악하기 위하여 정준 상관 분석(CCA)를 진행했다. 시각적으로 상관관계가 어떻게 이루어져 있는지 확인했다. 상단의 Pearson, Spearman, 그리고 Kendall 상관계수 분석을 진행했다.

독립변수들 간에 강한 선형 상관관계가 존재하는 것을 '다중공선성 문제(Multicolinearity)가 있다.'라고 표현한다[19]. 이러한 다중공선성은 회귀 분석을 하는 경우 회귀계수의 불안정성을 높여 예측력을 저하시키고, 모델 해석의 어려움을 증가시킨다. 우리는 이러한 점을 감안하여 독립변수의 VIF (Variance Inflation Factor)를 측정하여 다중공선성의 정도를 확인했다.

머신러닝 모델에서 데이터의 양이 부족하다고 판단하여 데이터 증강기법을 적용한 비교분석을 수행했다. 먼저 증강 방법을 적용하지 않은 경우를 기준으로 SMOTE, RandomOverSampling, 그리고 BorderlineSMOTE 기법을 적용한 경우와의 성능을 비교했다. SMOTE는 소수 클래스 샘플과 k-최근접 이웃(k-NN) 사이에서 새로운 데이터를 선형 보간하여 합성 샘플을 생성하는 기법이다. Random OverSampling은 소수 클래스의 기존 데이터를 단순 복제하여 데이터 불균형을 완화하는 기법으로, 데이터 다양성이 증가하지 않아 과적합의 위험이 존재할 수 있다. Borderline-SMOTE는 소수 클래스 중 다수 클래스와의 경계에 위치한 샘플을 중심으로 새로운 데이터를 생성하여, 결정 경계를 보다 정교하게 학습할 수 있도록 설계된 기법이다.

머신러닝 모델은 9개 모델(AdaBoost Regressor, Bagging Regressor, CatBoost Regressor, ExtraTrees Regressor, Gradient Boosting Regressor, LGBM Regressor, MLP Regressor, RandomForest Regressor, 그리고 XGB Regressor)을 이용하여 실험을 진행했다. 사용한 모델은 각각 부스팅(AdaBoost, Gradient Boosting, CatBoost, LGBM, XGB), 배깅(Bagging, RandomForest, ExtraTrees), 신경망(MLP) 기반 접근 방식을 포함하여 다양한 학습 패러다임을 제공하며, 비선형 관계를 효과적으로 학습하고 예측 성능을 극대화할 수 있는 특징을 갖기 때문에 선정하였다.

모델의 일반화 성능을 측정하기 위해 Nested Cross-Validation (Nested CV)을 적용하였으며, 내부 교차 검증(inner CV) 과정에서 Grid Search를 활용한 하이퍼파라미터 튜닝을 수행했다. 실험에 사용된 모델의 파라미터는 표 1과 같다.

Nested CV는 5-Fold Stratified K-Fold를 기반으로 구성되었으며, 외부 교차 검증(outer CV)에서는 모델의 성능을 평가하고, 내부 교차 검증(inner CV)에서는 최적의 하이퍼파라미터를 탐색하는 방식으로 진행되었다. 구체적으로, 내부 교차 검증에서는 Grid Search를 활용하여 최적의 하이퍼파라미터를 찾고, 이를 기반으로 외부 교차 검증의 테스트 데이터셋에 대해 최종 성능을 평가하였다.

각 모델별로 다양한 하이퍼파라미터 조합을 탐색한 후, 외부 교차 검증의 각 폴드에서 최적의 하이퍼파라미터를 선택하였다. 그 후, 최적의 하이퍼파라미터를 모델에 적용하여 전체 학습 데이터(X_train, y_train)로 학습한 후, 독립적인 테스트 데이터(X_test, y_test)에 대해 최종 성능을 평가하였다.

표 1. 실험에서 사용한 모델의 파라미터
Table 1. Parameters of the model used in the experiment

RandomForestRegressor[20]		AdaBoostRegressor[21]	
n_estimators	[50, 100]	n_estimators	[50, 100]
max_depth	[5, 10, None]	learning_rate	[1.0, 0.1]
BaggingRegressor[22]		ExtraTreesRegressor[23]	
n_estimators	[50, 100]	n_estimators	[50, 100]
max_samples	[0.8, 1.0]	max_depth	[5, 10, None]
max_features	[0.8, 1.0]	min_samples_split	[2, 5]
GradientBoostingRegressor[24]		MLPRegressor[25]	
n_estimators	[50, 100]	hidden_layer_sizes	[(50,), (100,), (50, 50)]
learning_rate	[0.1, 0.01]	activation	['tanh', 'relu']
max_depth	[3, 5]	alpha	[0.0001, 0.001]
CatBoostRegressor[26]		LGBMRegressor[27]	
loss	['squared_error', 'huber', 'epsilon_insensitive']	C	[0.1, 1, 10]
penalty	['l2', 'l1', 'elasticnet']	kernel	['linear', 'rbf']
max_iter	[1000, 2000]	gamma	['scale', 'auto']
XGBRegressor[28]			
n_estimators	[50, 100]		
learning_rate	[0.01, 0.1]		
max_depth	[3, 5]		

머신러닝의 모델의 성능 평가 지표는 R score, MSE (Mean Square Error), 그리고 MAE (Mean Absolute Error)을 수행했다. R Score는 모델이 주어진 데이터를 얼마나 잘 설명하고 있는지 나타내는 지표로써, 실제값의 변동중 모델이 설명할 수 있는 비율을 의미한다. 간혹 음수가 나오는

경우가 있는데, 단순 일괄 평균으로 예측하는 경우보다 성능이 떨어지는 것을 의미한다. MSE (Mean Squared Error)는 예측값과 실제값의 차이 제곱을 모두 더하고, 데이터 수로 나눈 평균 제곱 오차이다. 이 값이 작을수록 모델의 예측이 실제 값에 가까운 것을 의미한다. MAE (Mean Absolute Error)는 예측값과 실제값의 차이 제곱을 모두 더한 뒤 데이터 수로 나눈 평균 절댓값 오차로써, 동일하게 값이 작을수록 모델 성능이 우수하며, 제곱이 아닌 절댓값이므로 이상치(Outlier)에 덜 민감한 특성을 가진다.

SHAP Value를 계산하여 모델의 결과 및 출력을 이해하고자 했다. 이는 해석할수 없는, '블랙박스'로 인식되는 머신러닝 모델이 어떻게 예측을 도출하는지, 각 독립변수가 어떤 방식으로 예측 결과에 기여하는지 구체적으로 확인할 수 있는 기법이다. SHAP value란 게임이론의 Shapely Value 개념을 바탕으로, 각 독립변수가 예측 결과에 미치는 기여도를 정량화한 지표이다. 단일 샘플에 대해 모든 독립변수를 제거하거나 변화시켜 모델의 출력이 어떻게 달라지는지를 고려하기 때문에, 어떤 독립변수가 예측값(확률)을 증가시키는지, 어떤 독립변수가 예측값(확률)을 감소시키는지 직관적으로 해석할 수 있다. 이를 통하여 모델의 신뢰성을 높임과 동시에 변수의 중요도에 대해서 분석했다.

Q-Q Plot은 예측값과 관측값의 분위 수가 얼마나 일치하는지 알아보기 위해 사용했다. 점들이 빨간색 점선(이상적인 선)에 가까울수록 모델의 예측 성능이 좋음을 의미한다.

V. 실험 결과

본 논문에서는 예측 모델의 입력으로 사용하기 전에 다중공산선의 영향성을 측정했다. variance inflation factor 결과는 seat 1.468, cast1 1.082, cast2 1.091, cast3 1.148, cast4 1.145 weekend 1.295, day 1.555, dc 8.779, evt 1.942, Month 1.693, Day 1.193, 그리고 night 1.521의 수치를 보였다. 값이 10보다 큰 경우, 그 독립변수는 다중공산성이 있다고 보는데, variance inflation factor 결과엔 10보다 큰 값이 없으므로 다중 공산성을 가지는 변수는 없음을 확인했다.

표 2에서는 변수 cast2, weekend, dc, evt, musical gentleman, musical hades, musical salieri, musical versailles가 Pearson, Spearman, Kendall 상관계수 모두에서 높은 상관관계를 보였다. 이 결과를 통해, 어떤 배우가 출연하는지, 어떤 요일인지, 할인율이 얼마나 높은지, 이벤트가 있는지, 어떤 뮤지컬인지와 같은 요소들이 잔여 좌석 수와 높은 관계가 있다고 볼 수 있다. 또한 독립변수 중 Month가 높은 상관관계에 있으므로 보아, 뮤지컬의 공연 시기가 몇 월인지가 하나의 중요한 요인으로 확인했다.

계층화는 데이터의 클래스나 특징 분포를 균형 있게 나누는 방식이다. 표 3는 non-stratified와 stratified에 따른 데이터셋의 기초 통계량을 분석한 결과이다. 예를 들어, cast3과 cast4는 Non-stratified보다 stratified에서 표준편차(σ)가 더 낮은 경향을 볼 수 있으며, 이는 stratified 방식이 데이터의 안정화에 기여했음을 알 수 있다.

표 2. 독립변수와 종속변수간의 상관계수 분석
Table 2. Analysis of correlation coefficients between independent and dependent variables

Feature name	Pearson correlation (r , p-value)	Spearman correlation (r , p-value)	Kendall correlation (r , p-value)
cast1	(0.036,0.463)	(0.018,0.704)	(0.014,0.710)
cast2	(0.152,0.002)	(0.157,0.001)	(0.124,0.001)
cast3	(0.008,0.857)	(0.019,0.694)	(0.17,0.670)
cast4	(0.052,0.287)	(0.049,0.316)	(0.041,0.316)
weekend	(0.152,0.002)	(0.158,0.001)	(0.116,0.002)
day	(0.070,0.158)	(0.076,0.124)	(0.066,0.124)
dc	(0.358,0.000)	(0.422,0.000)	(0.340,0.000)
evt	(0.220,0.000)	(0.259,0.000)	(0.223,0.000)
Month	(0.105,0.034)	(0.098,0.046)	(0.077,0.08)
Day	(0.089,0.070)	(0.072,0.142)	(0.051,0.152)
night	(0.070,0.154)	(0.069,0.160)	(0.060,0.160)
musical bare	(0.081,0.100)	(0.062,0.211)	(0.053,0.211)
musical gentleman	(0.276,0.000)	(0.272,0.000)	(0.235,0.000)
musical hades	(0.211,0.000)	(0.198,0.000)	(0.172,0.000)
musical salieri	(0.271,0.000)	(0.311,0.000)	(0.265,0.000)
musical versailles	(0.310,0.000)	(0.326,0.000)	(0.282,0.000)

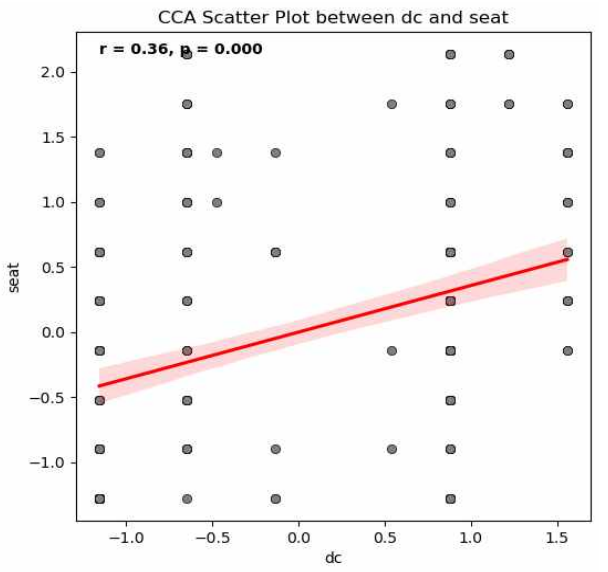

그림 2. 변수 seat과 dc에 대한 CCA 분석
Fig. 2. CCA analysis for variables seat and dc

그림 2는 종속 변수와 독립 변수 간의 상관관계를 분석한 결과 중, 가장 높은 R score을 기록한 변수 dc와 seat을 산

표 3. stratified 여부에 따른 데이터 셋의 기초 통계량 분석
Table 3. Basic statistics and analysis of datasets based on whether they're stratified or not

	Train $\mu(\sigma)$		Internal validation $\mu(\sigma)$		External validation $\mu(\sigma)$	
data	Non-stratified	Stratified	Non-stratified	Stratified	Non-stratified	Stratified
cast1	1.100 (1.000)	1.000 (0.900)	1.000 (1.000)	1.300 (0.900)	1.200 (0.900)	1.100 (1.000)
cast2	1.400 (1.100)	1.400 (1.100)	1.400 (1.100)	1.400 (1.100)	1.300 (1.100)	1.300 (1.000)
cast3	0.700 (0.700)	0.700 (0.700)	0.600 (0.600)	0.700 (0.700)	0.600 (0.700)	0.500 (0.600)
cast4	0.700 (0.700)	0.700 (0.700)	1.300 (0.600)	0.700 (0.700)	0.800 (0.700)	0.700 (0.800)
weekend	4.800 (1.700)	4.800 (1.700)	6.000 (1.000)	4.900 (1.800)	4.600 (1.600)	4.800 (1.600)
day	0.300 (0.500)	0.300 (0.500)	0.000 (0.000)	0.300 (0.400)	0.300 (0.400)	0.300 (0.400)
dc	16.800 (14.900)	16.800 (14.800)	22.500 (13.000)	16.900 (14.800)	17.000 (13.900)	18.100 (14.700)
evt	0.300 (0.500)	0.200 (0.400)	0.000 (0.000)	0.300 (0.500)	0.200 (0.400)	0.300 (0.500)
Month	7.700 (1.000)	7.700 (1.000)	8.300 (0.600)	7.700 (1.000)	7.700 (1.000)	7.700 (1.100)
Day	15.800 (8.400)	16.400 (8.200)	16.000 (6.600)	15.700 (9.100)	17.800 (8.800)	16.100 (9.200)
night	0.300 (0.500)	0.300 (0.500)	0.700 (0.600)	0.400 (0.500)	0.300 (0.400)	0.400 (0.500)
musical bare	0.200 (0.400)	0.200 (0.400)	0.700 (0.600)	0.200 (0.400)	0.300 (0.400)	0.300 (0.400)
muscial gentleman	0.200 (0.400)	0.200 (0.400)	0.000 (0.000)	0.200 (0.400)	0.200 (0.400)	0.200 (0.400)
musical hades	0.100 (0.400)	0.200 (0.400)	0.300 (0.600)	0.200 (0.400)	0.200 (0.400)	0.100 (0.300)
musical salieri	0.200 (0.400)	0.200 (0.400)	0.000 (0.000)	0.200 (0.400)	0.200 (0.400)	0.300 (0.400)
musical versailles	0.200 (0.400)	0.200 (0.400)	0.000 (0.000)	0.200 (0.400)	0.100 (0.300)	0.100 (0.300)

포도로 시각화한 것이다. 분석 결과, 두 변수의 R score값은 약 0.1296으로 낮은 설명력을 보였다. 또한, 산포도를 통해 넓게 퍼진 점들의 분포를 확인했다.

표 4는 샘플링 기법별로 성능이 가장 좋았던 모델에 대한 R Score, MSE, MAE 결과를 정리한 결과이다. 표에 나타난 바와 같이, SMOTE stratified 기법을 적용한 GradientBoosting Regressor 모델이 R Score 값이 0.608로 가장 높은 성능을 보였다. 이는 모델이 잔여 좌석 예측에 있어 가장 높은 정확도를 달성했음을 의미한다. 반면, Random stratified 기법을 적용한 GradientBoostingRegressor 모델 역시 R Score 값 0.608을 기록하였고, 다른 기법들과 비교하여 비교적 우수한 성능을 보였다. MSE와 MAE 값 또한 Random stratified GradientBoostingRegressor 모델이 상대적으로 낮은 값을 기록하며, 예측 오차가 가장 적다는 것을 나타낸다. 이 모델은 과적합을 방지하고, 예측 정확도를 높이는 데 효과적인 기법으로 평가된다.

반면, Normal stratified CatBoostRegressor는 R Score 값 0.580으로 상당히 좋은 성능을 보였으나, MSE와 MAE는 상대적으로 높은 값을 나타내어 예측 정확도 면에서 다른 기법보다는 다소 떨어지는 성과를 보였다.

SMOTE, BorderlineSMOTE, Random 기법에서 각각 다른 모델들이 성능을 나타냈지만, Random stratified Gradient BoostingRegressor와 SMOTE stratified LGBMRegressor가 전반적으로 가장 우수한 성능을 보였다.

표 4. 샘플링 기법에 따른 머신러닝 성능 비교
Table 4. Comparison of machine learning performance by sampling technique

	Model Name	R Score	MSE	MAE
Normal	ExtraTreeRegressor [23]	0.500	3.422	1.446
Normal stratified	CatBoostRegressor [26]	0.580	2.925	1.396
SMOTE	GradientBoosting Regressor [24]	0.474	3.594	1.501
SMOTE stratified	LGBMRegressor [27]	0.557	3.085	1.457
Random	ExtraTreeRegressor [23]	0.466	3.650	1.537
Random stratified	GradientBoosting Regressor [24]	**0.608**	**2.727**	**1.309**
Borderlin esmote	CatBoostRegressor [26]	0.492	3.474	1.446
Boderline smote stratified	XGBRegressor [28]	0.590	2.850	1.388

그림 3은 non-stratified와 stratified의 R Score를 샘플링 기법별로 비교하여 정리한 그림이다. Non-stratified의 경우 샘플링을 적용하지 않은 방법이, stratified의 경우 Random 샘플링의 성능이 가장 높았으며, 전반적으로 non-stratified보다 stratified의 성능이 더 높았다. 이를 통해 데이터의 불균형을 완화해주는 것이 모델 성능 향상에 기여함을 알 수 있다.

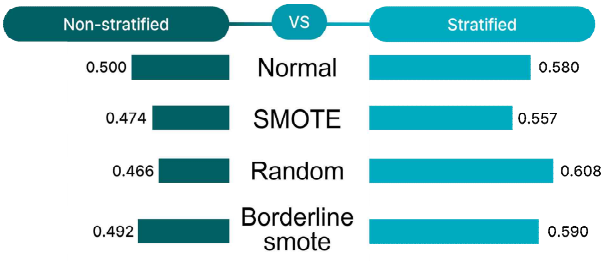

그림 3. Non-stratified와 Stratified의 샘플링 기법별 R Score 비교차트
Fig. 3. R score comparison chart for non-stratified vs. stratified sampling techniques

그림 4는 머신러닝 모델 중 성능이 가장 우수했던 Random stratified 데이터를 활용한 Gradient Boosting Regressor 모델의 SHAP 분석 결과를 보여준다. 그림 4(a)에서 보이듯이 seat을 예측하는 데에 있어서 가장 중요한 인자는 변수 dc가 중요하게 나왔음을 확인했다. 그림 4(b)에서 변수 dc가 결정을 내리는데 가장 많은 영향성을 끼치는 것을 확인했다. 그림 4(c)에서 변수 dc가 30 아래인 경우, 즉 할인율이 30% 미만인 경우 예측값을 1.50 이상 증가시키고, 30% 이상인 경우 1.50미만으로 감소시키는 것을 확인했다.

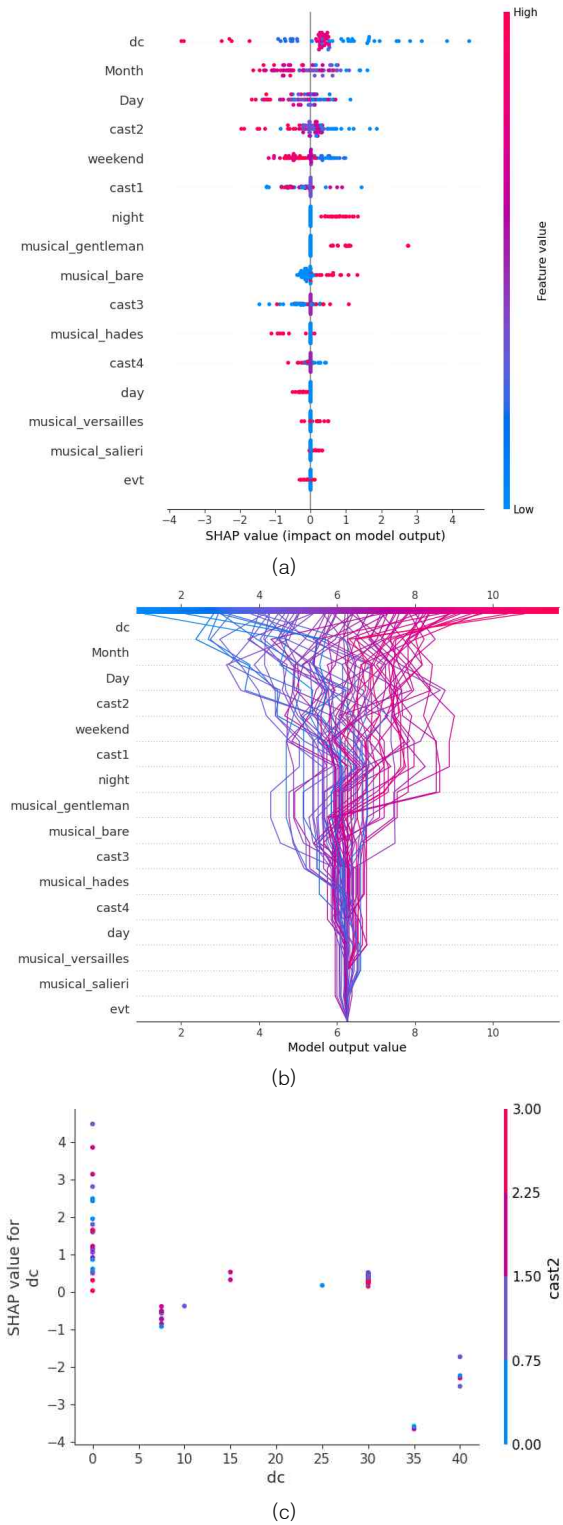

그림 4. Gradient Boosting Regressor에서 SHAP 분석, (a) Summary plot, (b) Decison plot, (c) Dependencey plot
Fig. 4. SHAP analysis in Gradient Boosting Regressor, (a) Summary plot, (b) Decision plot, (c) Dependencey plot

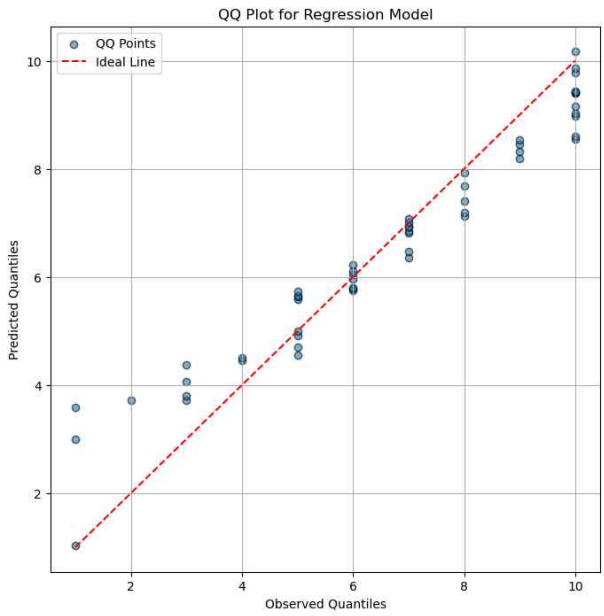

그림 5. Gradient Boosting Regressor에 대한 Q-Q Plot
Fig. 5. Q-Q Plot for Gradient Regressor

그림 5는 성능이 가장 좋았던 Gradient Boosting Regressor 모델의 결과이다. 점이 빨간색 점선을 따라 분포하는 것을 보아, 높은 성능을 보이는 것을 알 수 있다. 이는 모델이 데이터의 패턴을 효과적으로 학습하여 예측을 수행했다.

Ⅵ. 결 론

본 연구에서는 처음으로 뮤지컬 잔여 좌석 예측을 위한 프레임워크를 제안했다. 해당 프레임워크에서는 데이터 샘플링 기법을 활용하여 각 샘플링 방법에 따른 머신러닝 모델의 예측 성능을 비교했다. 또한, 9개의 머신러닝 모델을 분석한 결과, Gradient Boosting Regressor가 가장 우수한 성능을 보였다. SHAP 분석을 통해 변수 dc가 잔여 좌석 예측에 가장 중요한 영향을 미친다는 것을 확인했다.

향후 연구에서는 티켓팅 예측의 해석력을 강화하고 예측 성능을 더욱 향상시킬 수 있는 새로운 모델의 개발이 요구된다.

참고문헌

[1] J.-H. Lee and G.-E. Chung, “A Study on the Development Status and Change of Korean Musical Industry,” *Journal of Culture Industry*, Vol. 13, No. 4, pp. 43-53, December 2013.

[2] Chosun Ilbo. Performance Arts Ticket Insights [Internet]. Available: https://www.chosun.com/culture-life/performance-arts/2024/01/16/KWSSXKNOEVEHPILXI35TAHPWSA/.

[3] Brunch. Analysis of Musical Audience Preferences [Internet]. Available: https://brunch.co.kr/@akzkfldh/6.

[4] J. Kim, Big Data Analysis on Audience Preference and Public Relation Factors in Performing Arts, Master’s Thesis, Hongik University, Seoul, February 2018. https://doi.org/10.23174/hongik.000000022351.11064.0000294

[5] Dailian News. Musical Ticketing Trends and Analysis [Internet]. Available: https://www.dailian.co.kr/news/view/1390784.

[6] News N Yonhap. Ticket Inflation Impact on Musical Industry [Internet]. Available: https://www.newsyonhap.com/news/12466.

[7] The Korea Economic Daily. Musical Industry Revenue Growth Trends [Internet]. Available: https://www.hankyung.com/article/2024112287617.

[8] D. H. Jin, Economic Scenario Analysis for Stock Portfolios Based on Machine Learning: The Case of South Korea, Master’s Thesis, Kyung Hee University, Yongin, February 2023.

[9] C. S. Kim, S. W. Kim, and H. S. Choi, “Machine Learning-Based Intraday Stock Price Prediction: A Firm-Level Analysis Using High Frequency Data,” *Journal of Intelligence and Information Systems*, Vol. 30, No. 2, pp. 85-106, June 2024. https://doi.org/10.13088/jiis.2024.30.2.085

[10] M. S. Seo, A Study on Dynamic Text Feature Weighting for Stock Market Prediction with Deep Learning, Master’s Thesis, Yonsei University, Seoul, February 2022.

[11] S. J. Lee, Promotion Demand Forecasting Design for Small and Medium e-Commerce Sellers: Focusing on Integrated Data Utilization, Master’s Thesis, Incheon National University, Incheon, February 2024.

[12] S. An and J.-Y. Jung, “Multi-Step Time Series Forecasting for Hypermarket Sales Using Temporal Fusion Transformers,” *The Journal of Society for e-Business Studies*, Vol. 28, No. 3, pp. 43-53, August 2023. https://doi.org/10.7838/jsebs.2023.28.3.043

[13] D. J. Yoon, Predict New Book Sales by Online and Offline Sales Channel with Ensemble Machine Learning, Master’s Thesis, Korea University, Seoul, August 2023. https://doi.org/10.23186/korea.000000276385.11009.0000132

[14] H. N. Oh and Y. T. Kim, “A Study on Musical Purchase Factors among the Domestic and International Audience,” *Journal of Korea Service Management Society*, Vol. 19, No. 2, pp. 217-235, June 2018. https://doi.org/10.15706/jksms.2018.19.2.010

[15] MyMusicTaste Official Blog. Ticket Sales Estimation Modeling [Internet]. Available: https://medium.com/mymusictaste-official/ticket-sales-estimation-modeling-3-1af24baaf85b.

[16] M. G. Saerodji, Forecasting Method Analysis for Predicting Ticketing Sales of Classical Music Concert in Taipei, Master's Thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2016.

[17] B. Seo and H. Lee, "A Machine Learning-Based Popularity Prediction Model for YouTube Mukbang Content," *Journal of Internet Computing and Services*, Vol. 24, No. 6, pp. 49-55, December 2023. https://doi.org/10.7472/jksii.2023.24.6.49

[18] J.-W. Kim, J.-W. Seo, C.-H. Son, and S. H. Choi, "Analysis of Influential Factors for the Prediction of YouTube Views Using Tree-Based Machine Learning," *Journal of Digital Contents Society*, Vol. 26, No. 1, pp. 175-182, January 2025. http://dx.doi.org/10.9728/dcs.2025.26.1.175

[19] J. H. Kim, "Multicollinearity and Misleading Statistical Results," *Korean Journal of Anesthesiology*, Vol. 72, No. 6, pp. 558-569, December 2019. https://doi.org/10.4097/kja.19087

[20] Scikit-Learn Documentation. RandomForestRegressor [Internet]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[21] Scikit-Learn Documentation. AdaBoostRegressor [Internet]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.AdaBoostRegressor.html.

[22] Scikit-Learn Documentation. BaggingRegressor [Internet]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.BaggingRegressor.html.

[23] Scikit-Learn Documentation. ExtraTreesRegressor [Internet]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html.

[24] Scikit-Learn Documentation. GradientBoostingRegressor [Internet]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

[25] Scikit-Learn Documentation. MLPRegressor [Internet]. Available: https://scikit-learn.org/1.6/modules/generated/sklearn.neural_network.MLPRegressor.html.

[26] CatBoost Documentation. CatBoostRegressor [Internet]. Available: https://catboost.ai/docs/en/concepts/python-reference_catboostregressor.

[27] LightGBM Documentation. LGBMRegressor [Internet]. Available: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html.

[28] XGBoost Documentation. Parameter Documentation [Internet]. Available: https://xgboost.readthedocs.io/en/stable/parameter.html.

김민선(Min-Sun Kim)

2022년 3월～현 재: 한성대학교 AI응용학과 (학부과정)
※관심분야 : 딥러닝(Deep Learning), 컴퓨터 비전(Computer Vision), 의료 인공지능(Medical AI)

서준혁(Jun-Hyuk Seo)

2019년 3월～현 재: 한성대학교 IT융합공학부 (학부과정)
※관심분야 : 데이터 과학, 컴퓨터 비전, 자연어 처리

최승호(Seoung-Ho Choi)

2018년 : 한성대학교 전자정보공학과 (공학사)
2020년 : 한성대학교 전자정보공학과 (공학석사)
2023년～현 재: 한성대학교 기초교양학부 조교수
※관심분야 : 딥러닝