2023 DAB(Data Analytics for Business) 경진대회

예측모델을 통한 행사 참여 인원 예측 및 행사 주최사 컨설팅 서비스

- 올림픽공원을 중심으로

혼돌이팀

권소영 김수민 오지민 이세은 정성희

팀원 소개

권소영

통계학과 21학번

날씨 데이터 분석

행사인원 예측모델 제작

김수민

통계학과 21학번

날씨 데이터 분석

행사인원 예측모델 제작

오지민

수학과 19학번

행사 데이터 분석

행사인원 예측모델 제작

이세은

보건정책관리학부 21학번

교통량 데이터 분석

교통량 이상치 탐지모델 제작

정성희

경영학과 21학번

행사 데이터 분석

비즈니스 모델 고안

Table of Contents

1.1 제안배경

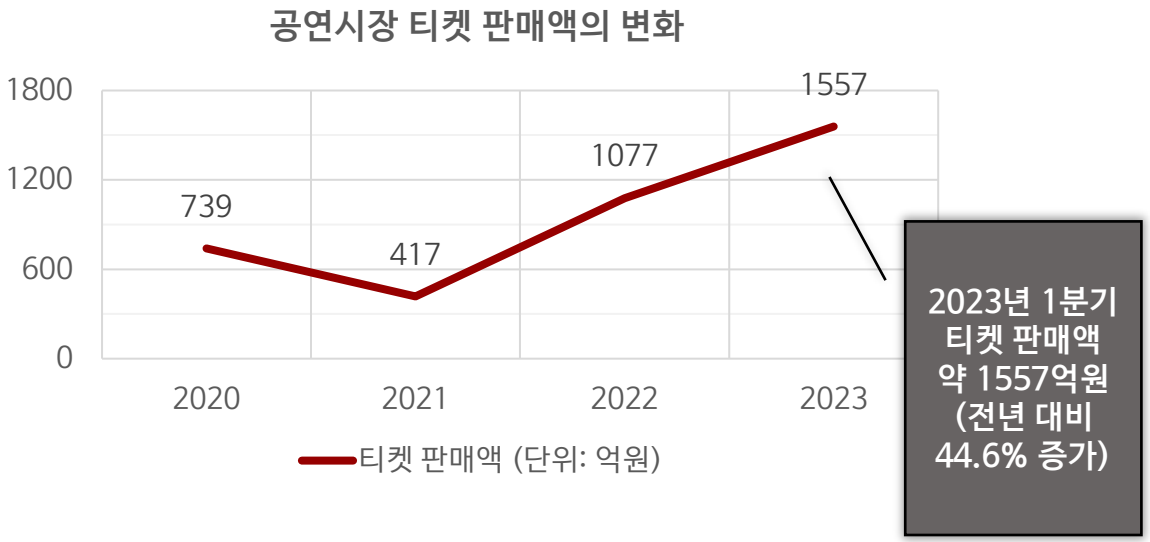
공연예술시장 회복 및 행사 안전 관리에 대한 관심 증가로 참여 인원 예측의 중요성 대두

공연 예술 및 행사의 활성화

- 코로나19로 침체기를 겪었던 공연예술시장의 회복세
- 협찬 및 광고 유치의 비용 절감과 티켓 판매의 수익 증대를 위해 참여 인원의 사전적인 예측 필요

안전 관리에 대한 관심 증가

- 사고예방을 위한 지자체의 다중밀집행사 안전 관리 점검 강화 및 철저한 안전관리계획 요구
- 교통 통제 방안 계획 수립 및 안전 장치 설치, 인력 배치를 위한 비용 추산 필요



'만원 지하철' 인파관리한다…안전관리지침 제정

정부, '다중밀집사고' 예방 대책 마련 나서

정부·지자체 안전관리 달라진다

1.2 사업 구조 및 목표

제안 사업의 기대 효과

수익성 극대화

- 수입 증대
  1. 기획 시, 행사 일자의 적절한 선정을 통해 흥행 가능성 증가
  2. 협력 업체 및 광고 유치 시, 정확한 계약 단가 책정 가능
- 비용 절감

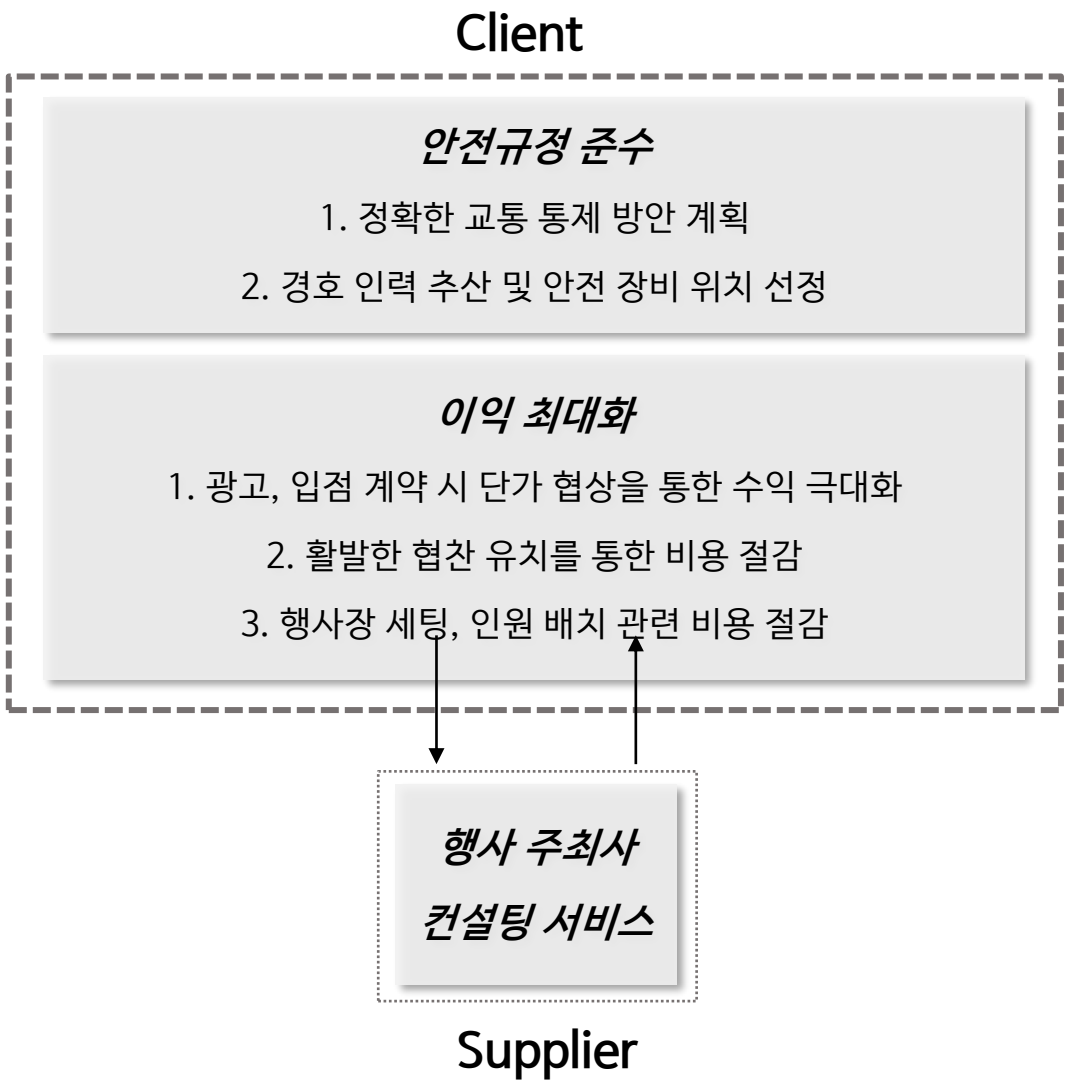
  인력배치, 물품배치 등의 비용 효율화

안전 관리의 혁신

- 외주를 통한 질서 유지, 경호 업무 중심의 기존 행사 안전관리
- 행사의 성격, 날짜, 날씨 등을 종합적으로 고려한 참여 인원 추산과 이를 바탕으로 본 서비스만의 맞춤형 안전관리 컨설팅 제공

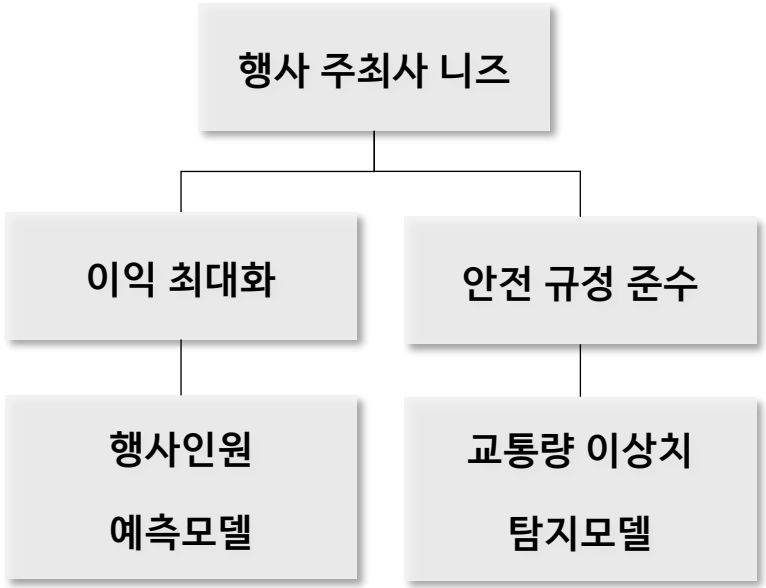
니치마켓 공략

1. 행사 진행 외주가 필요한 행사 주최자에게 인원 예측
2. 서비스 및 예측 인원에 기반한 행사 컨설팅 서비스 제공



1.3.1 분석 Overview

서비스 목표 실현을 위해 1) 행사인원 예측모델 2) 교통량 이상치 탐지 모델로 구분하여 진행



행사인원 예측 모델

Data.

- 올림픽공원 대관 정보, 유사공연 데이터, 날씨 정보

Method.

- 회귀분석 (SVM, 선형, Ridge, Lasso)
- 인공신경망

Expectation.

- 정확한 참여 인원 예측으로 수입 증대 및 비용 절감의 이익 최대화 실현

교통량 이상치 탐지 모델

Data.

- 서울시 역별 · 일별 · 시간대별 승하차 인원 정보

Method.

- SARIMA 모델
- 이상치 탐지 함수

Expectation.

- 날짜, 시간, 행사 이상치 탐지를 통해 질서 유지 계획 수립

1.3.2 올림픽공원 선정 이유

올림픽공원은 타 장소들의 특성을 포괄하고, 관련 데이터를 얻기에 용이한 등의 이유로 분석 대상으로 선정

올림픽 공원 주요 특성

행사 진행 테마의 다양성

- 콘서트, 팬미팅, 뮤지컬, 체육행사, 패션위크 등 다양한 유형의 행사가 진행되어 타 장소들의 특성 포괄

다양한 규모의 행사장

- 실내 행사장인 KSPO DOME, 올림픽홀 등과 실외 행사장인 88 잔디마당, 88 호수수변무대 등 존재
- 실내 행사장의 경우에도 연극, 콘서트 등 다양한 목적에 따른 다양한 규모의 행사장 존재

풍부한 데이터

- 공연 데이터 외에도 교통량 관련 데이터 다량 보유
- 특히, 올림픽 공원 내 진행된 행사의 참여 인원 데이터 존재

교통량 관련 해석의 용이

- 몽촌토성역, 올림픽공원역, 한성백제역 등 올림픽공원과 근접한 대중교통의 경우, 여타 유동 인구 밀집 구역의 영향을 적게 받아 행사로 인한 교통량 변화에 대한 해석 용이

프로젝트 방향

우선적으로 올림픽 공원 내 진행된 행사 데이터들을 대상으로 행사인원 예측 모델 및 교통량 이상치 탐지 방법론을 수립한다. 해당 방법론이 충분한 성능을 갖추면 타 생활체육, 문화 시설로 분석 범위를 넓힌다.

1.3.3 Doc2Vec를 활용한 선례 탐색

행사 정보가 주어지면, 이전에 개최된 행사들 중 가장 유사한 선례들을 알려주는 '선례 탐색 모델'을 제작하여 예측 모델 및 이상치 탐지 모형 제작에 활용하고자 함.

공연정보 데이터

올림픽공원 공연 외 타 공연장 정보 수집

- 기간: 2009년 – 2023년
- 정보: 공연 제목, 출연진, 공연 기간, 날짜 등등

Source:

- 인터파크 PLAY DB 109570 건의 공연 정보 데이터
  - 올림픽공원 공연 내 1463건의 공연 정보
  - 공공데이터 포털의 108107건 공연 정보

Doc2Vec 모델 생성(선례 탐색 모델)

모델 작동 예시

Input	2018 아이유 데뷔 10주년 아시아 투어 콘서트
Output	2018 자우림 정규앨범 발매 기념콘서트
	2018 샤이니 6번째 콘서트
	2015 슈퍼주니어 아시아 투어
	2018 WINNER PRIVATE STAGE
	2018 BLACKPINK SHOW

활용 방안

예측 모델에 있어서는 선례들의 행사 참여 인원을 가중평균하여 예측에 사용하고, (2.2 참조)

이상치 탐지 모델에는 교통량이 급증했던 선례들이 있을 경우 안전 계획 수립에 반영한다. (3.3 참조)

2.1 행사인원 예측모델 Overview

올림픽 공원 대관 정보와 유사 공연에 대한 인원 정보, 날씨 등의 추가 정보를 활용하여 행사인원 예측모델 제작

올림픽 공원 대관 정보

변수명

공연장소 공연유형 이름 시작일자
종료일자 신청인원 정산인원 내/외부
사용일수 주말/공휴일 여부
+) 유사 공연 인원, 날씨

기간

2012.01.01 ~ 2023.03.31

개수

1338개

예측변수 대관 정보 | 유사 공연 인원 정보 | 날씨 정보 등

유의 변수 선택

모델적합 선형 회귀 | ANN | Ridge | Lasso | SVM

예측 성공률 계산

예측 인원 도출 상위 성능 모델의 예측값 기준으로 가중평균 계산

최종 예측값 계산 및 모델 성능 평가

2.2 행사인원 예측모델 – 변수 선택

유사 공연 인원 정보를 담은 '선례 인원'과 대관 정보의 변수들 중 유의한 변수들을 선택하여 예측모델에 사용

선례 탐색 모델을 통한 가중 평균 계산

가중 평균 feature 생성

올림픽공원 행사 정보 데이터에 대하여 벡터화를 시킨다. 각 행사에 대하여 가장 유사한 행사 5개를 찾은 뒤, 유사도를 가중치로 한 가중 평균을 계산하여 **'선례 인원'**을 생성한다.

Ex)

10000	12000	13000	14000	9000
90%	91%	92%	85%	80%

$$\rightarrow \frac{10000*0.9+12000*0.91+13000*0.92+14000*0.85+9000*0.8}{0.9+0.91+0.92+0.85+0.8}$$

$$= 11639.27$$

선형 회귀를 사용한 변수 선택

인원 예측을 정교화하기 위해 '선례 인원'이 반영하지 못하는 정보 추가.
행사인원과 상관관계가 있는 변수를 선택 후, 최종 행사인원 예측에 사용.

1. 모든 변수를 포함한 모형을 수립
2. P-value가 큰 변수를 삭제한 모형 설정
3. 모형간 수정 결정계수 비교
4. 2, 3을 반복하여 유의한 변수 선정

후보 변수

- 대관 정보 변수
- 선례 인원
- 날씨
- 홍보 관련 지표

변수 선정 결과

- 공연장소
- 공연유형
- 신청인원
- 선례 인원

*변수 선정의 자세한 내용은 APPENDIX 참조

2.3 행사인원 예측모델 – 구동 원리

후보 예측모델들을 구현하여 모델 성능 비교. 성능 상위 3개의 모델을 베이스 모델로 선정하여 최종 예측치 계산

사용 변수 목록	모델 목록	Validation 모델 성능 (MSE)	*예측 성공률	최종 모델 제작
- 공연장소 - 공연유형 - 신청인원 - 선례 인원	SVM 회귀	3.05E+07	22.8%	모델 성능 상위 3개를 베이스 모델로 선정한 후, Bayesian Optimization으로 예측값에 가중치를 부여하여 최종 예측치를 구한다.
	선형 회귀 분석	7.94E+08	19.4%	
	Ridge 회귀	9.68E+08	17.2%	
	Lasso 회귀	9.68E+08	17.2%	
	인공 신경망	2.43E+08	13.1%	

***예측 성공률: 예측 인원의 실제 인원 ±10% 범위 포함 여부. 행사/공연 관계자와의 인터뷰에 기반하여 정의**

2.4 행사인원 예측모델 – 성능 검정

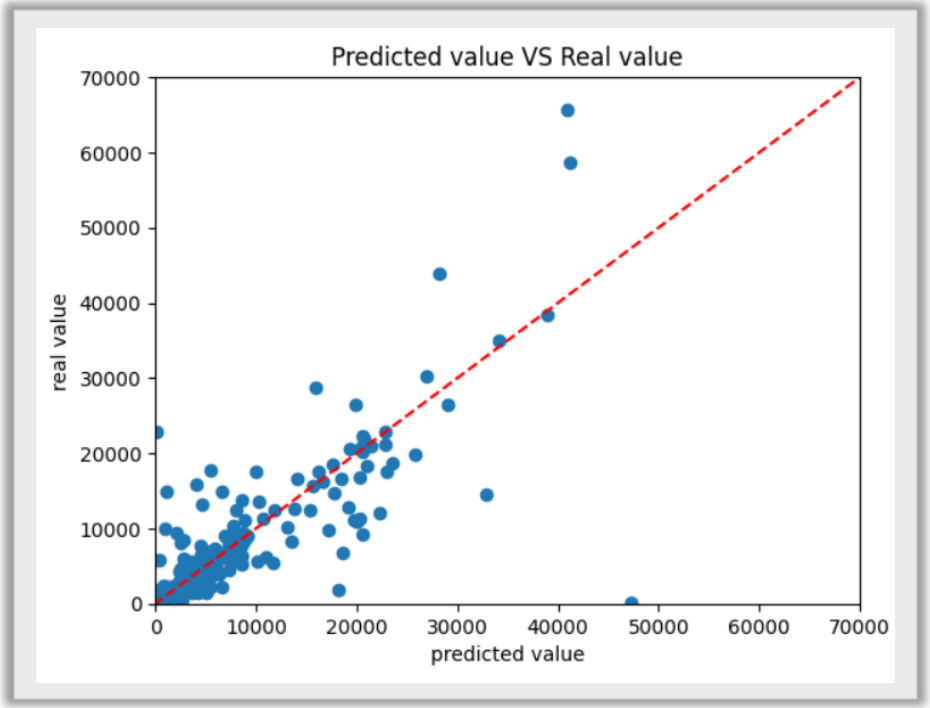
최종 인원 예측을 수행하였을 때 예측 성공률이 높지 않았으나, 관련 있는 다른 예측 변수를 찾아 적합한다면 더욱 좋은 성능을 보일 것으로 기대함

베이스 모델	가중치
선형 회귀 분석	0.034
Ridge 회귀	0.138
SVM 회귀	0.899

최종 예측 모델

MSE
3.71E+07

예측 성공률
29.5%



Implication

기존 목표로 설정한 성공률인 80%인 점을 감안하면, 최종 예측모델은 기대에 미치지 않았다.
현재 모델의 성능의 한계점은 행사의 인지도, 화제성 등을 반영하는 정보 등의 부재가 주요 원인으로 여겨지며, 이에 후속적으로 보다 다양한 정보들을 반영하는데 초점을 맞출 것이다.

3.1 교통량 이상치 탐지모델 Overview

서울교통공사에서 제공한 5년 간의 지하철 승하차 인원 정보를 기반으로 교통량이 급증하는 시기와, 해당 시기에 진행된 행사를 탐지하여 질서 유지 계획 수립에 반영한다.

역별 일별 시간대별 승하차 인원 정보 (서울교통공사)

변수	수송 일자	호선	역명	승하차 구분	6시 이전	6시-7시	…	23-24시	24시 이후
1	2022-01-01	8	몽촌토성(평화의문)	승차	21	74	…	30	
2	2022-01-01	8	몽촌토성(평화의문)	하차	17	90	…	42	

❖ **사용한 지하철역**

몽촌토성역(평화의문) 8호선, 올림픽공원역 5호선,
올림픽공원역 9호선, 한성백제역 9호선

❖ **기간: 2018.12.01 ~ 2023.03.31**

→ **Covid 19 전/중/후로 구분된 데이터셋 생성**

코로나 전: 2018.12.01 ~ 2020.01.19

코로나 중: 2020.01.20(국내 첫 확진자 발생) ~ 2022.05.01

코로나 후: 2022.05.02(야외 마스크 해제) ~ 2023.03.31

모델 적합

1. 각 역 별로 SARIMA 모델 적합
2. 표준화 잔차 4 이상인 값들을 이상치로 간주

선례 탐색

원하는 행사의 선례를 탐색한 후,
선례의 진행 시기가 이상치에 해당하는지 확인

이상치일 경우

선례 분석

선례가 실제로 교통량에 영향을 끼쳤는지 확인하고,
교통량 급증이 해당 행사에도 나타날 지 분석

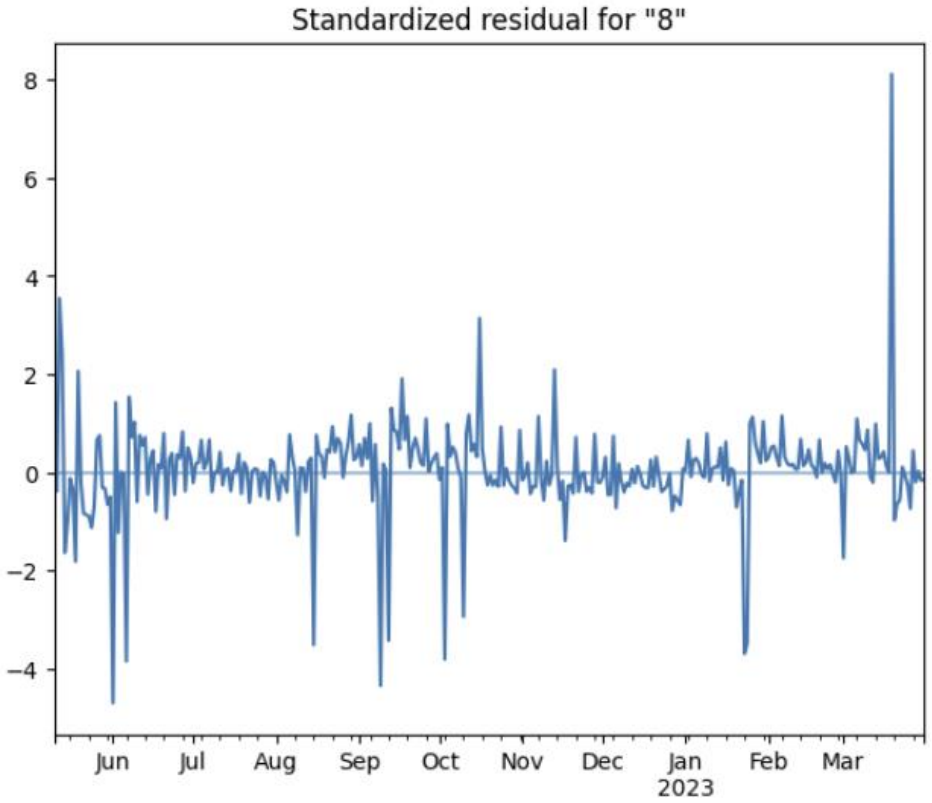
급증이 나타날 것이라 판단될 경우

분석 결과를 토대로 안전 계획 수립

3.2 SARIMA 모델을 활용한 이상치 여부 판단

계절성을 고려한 시계열 모델인 SARIMA로 지하철 승하차 인원의 추이를 모델링한 후, 행사로 인한 교통량 급증이 잔차에 반영된다는 점을 이용하여 이상치를 탐지함.

SARIMA 적합 후 잔차 이상치 파악



날짜, 시간, 행사 이상치 추출

날짜 / 시간	19	20	21	22	23
2018-12-07	-연극 〈그을린사랑〉 -2018 김동률 콘서트	-연극 〈그을린사랑〉 -2018 김동률 콘서트	·	·	-연극 〈그을린사랑〉 -2018 김동률 콘서트
2018-12-21	·	·	-윤주희 2집 앨범 〈Mother Nature〉 발매 기념 쇼케이스 -PSY CONCERT 올나잇스탠드 2018	-윤주희 2집 앨범 〈Mother Nature〉 발매 기념 쇼케이스 -PSY CONCERT 올나잇스탠드 2018	-윤주희 2집 앨범 〈Mother Nature〉 발매 기념 쇼케이스 -PSY CONCERT 올나잇스탠드 2018

3.3 선례 탐색을 활용한 교통량 분석

선례 탐색 모델을 통해 선례 행사들을 추출한 뒤 이상치 여부를 살펴보며 해당 공연이 교통량에 영향을 미칠지 분석함.

Ex)

선례 탐색 모델 사용

2018 국카스텐 연말 전국 투어

2018 아이유 데뷔 10주년 아시아 투어 콘서트

2018 자우림 정규앨범 발매 기념 콘서트

2018 은지원 PIRVATE STAGE

2018 김동률 콘서트 (교통량 이상치)

2018 성시경 콘서트

선례 분석을 통한 교통량 급증 여부 판단

만일 선례들 중 교통량 이상치가 존재한다면,
다음의 로직을 통해 교통 통제가 필요한지 결론을 낸다.

✓ '2018 김동률 콘서트'가 진행된 기간에 다른 행사가 있었는가?

- 연극 〈그을린 사랑〉이 존재하였으나, 참여 인원이 1400명인 반면, '2018 김동률 콘서트'의 경우 3만 명이 참가하였다.

✓ '2018 국카스텐 연말 전국 투어'는 '2018 김동률 콘서트'의 규모에 준하는가?

- 해당 공연에서 준비한 좌석이 8000여 명 가량으로, 이상치가 존재한 행사에 비해 교통량 급증의 우려는 덜하다.

✓ 만일 준할 경우 어떻게 대응하는가?

- '2018 김동률 콘서트'의 공연 시간이 오후 8시 ~ 11시였고, 붐비는 시간은 오후 7시, 8시, 11시였음을 감안하여 안전 계획 수립

4.1 사업화 전략

안전관리 컨설팅, 협찬사 및 광고 대행사 중개 진행 등의 차별화 전략

B2B 기반의 수익 모델

인원예측을 통한 이익 최대화

- 공연/행사장 물품 세팅 및 인력 배치 효율화
- 참여 인원 예측을 바탕으로 한 광고 단가 협상 및 활발한 협찬 유치 등 협찬사 및 광고 대행사 중개 진행

교통량 이상치 탐지를 통한 안전 관리 컨설팅

- 비상 출입구 확보 및 관람객 동선 설계
- 지하철 하차 인원 이상치 탐지를 통한 질서 유지 계획 수립
  - 유관 기관 협조 체계의 교통 통제
  - 인파 밀집 지역 질서 유지
  - CCTV를 이용한 모니터링

목표 고객층

행사 기획을 담당하는 기업

- 인원 예측 및 안전관리 방안에 관한 구체적인 가이드가 필요한 기업
- 티켓 판매 수익 외에도 광고 및 협찬 등 부가 수입 창출이 필요한 기업

경쟁사에 따른 전략

주 경쟁사

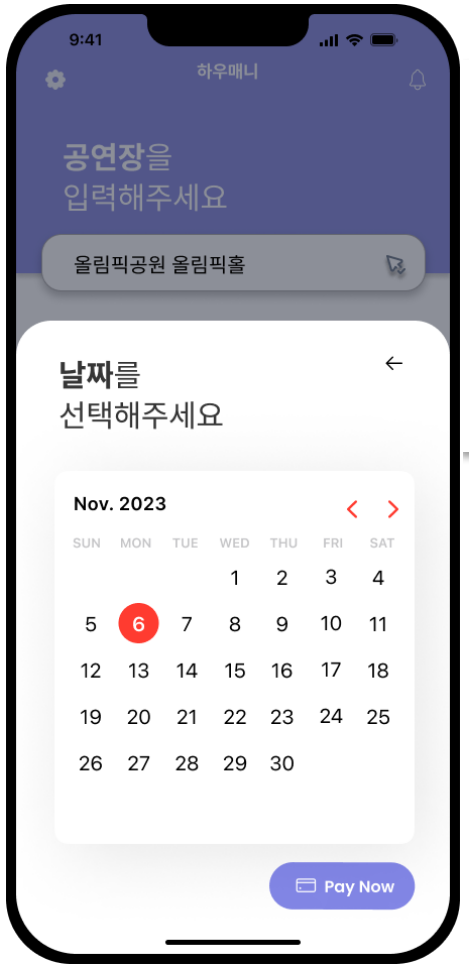
- 서비스행사 인원 통제 및 진행 보조 업체 혹은 행사 대행 업체

경쟁 상대에 대한 전략

- 고객 및 행사 데이터를 선점할 시 신규 경쟁자의 진입이 어려우므로, 시장에 빠르게 진입하는 것이 중요할 것으로 보임.

4.2 서비스 구체화

이용 단계의 서술을 통한 서비스 구체화 방안



의뢰

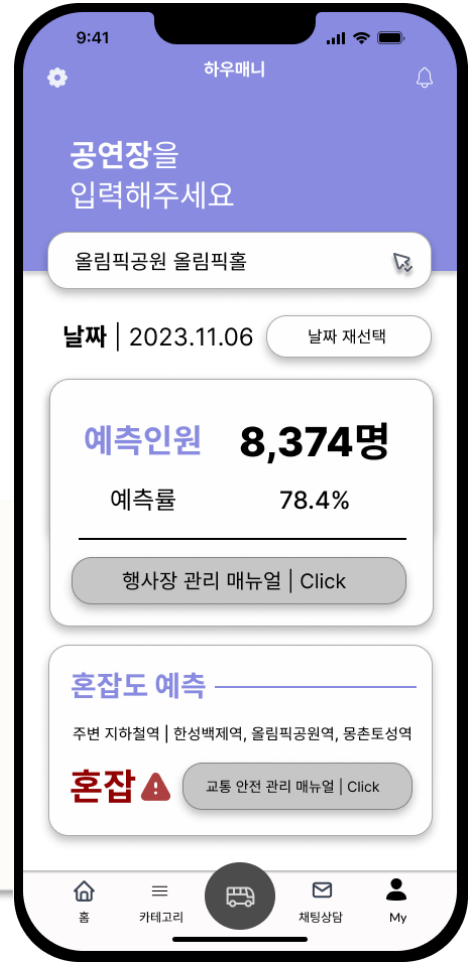
✓ **app을 통해 의뢰**

- 공연장 및 일시 선택
- 채팅상담 통해 의뢰비 조회
- App 내에서 결제 진행

결과 전달

✓ **app을 통해 결과 전달**

- 안전 관리 컨설팅, 광고대행사 및 협찬사 중개 서비스 의뢰 시 클라이언트와 미팅 후 진행



4.3 자체 피드백

성능 향상 및 사업 프로세스의 구체화

성능 향상 및 분석 대상 확대

- 마케팅, 홍보 관련 지수 만들어 반영
  - 다양한 티켓 판매사, 홍보매체를 반영하여 홍보정도를 측정
- 전국의 생활체육시설, 행사/공연장으로 서비스 제공 가능 장소 확대

서비스 구체화

- App 개발
- 안전 관리 컨설팅 위한 행정안전부 안전 관리 매뉴얼 숙지
- 광고 대행사와 파트너십 체결 및 협찬사들과 협력 환경 구축

기대효과

해당 프로젝트는 클라이언트의 행사 집행의 효율성 증진과 함께 군중 밀집 행사의 안전 확보를 통한 사회적인 기여를 하는 등의 효과를 기대할 수 있다.

End of Document

APPENDIX.

유의 변수 선택

OLS Regression Results

Dep. Variable:	EXCCLC_EVENT_NMPR_CO	**R-squared:**	0.896
Model:	OLS	**Adj. R-squared:**	0.894
Method:	Least Squares	**F-statistic:**	342.8
Date:	Wed, 27 Sep 2023	**Prob (F-statistic):**	0.00
Time:	13:59:18	**Log-Likelihood:**	-9615.4
No. Observations:	936	**AIC:**	1.928e+04
Df Residuals:	912	**BIC:**	1.939e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-310.1295	622.182	-0.498	0.618	-1531.205	910.946
C(FCLTY_NM)[T.2]	-574.6833	879.667	-0.653	0.514	-2301.089	1151.723
C(FCLTY_NM)[T.3]	1207.9186	872.303	1.385	0.166	-504.036	2919.873
C(FCLTY_NM)[T.4]	249.0058	1043.163	0.239	0.811	-1798.273	2296.284
C(FCLTY_NM)[T.5]	-1141.5221	957.715	-1.192	0.234	-3021.104	738.060
C(FCLTY_NM)[T.6]	-308.7243	1185.152	-0.260	0.795	-2634.666	2017.217
C(FCLTY_NM)[T.7]	71.4689	2127.678	0.034	0.973	-4104.246	4247.183
C(FCLTY_NM)[T.8]	-600.8682	1609.917	-0.373	0.709	-3760.442	2558.705
C(FCLTY_NM)[T.9]	-998.6422	1648.824	-0.606	0.545	-4234.572	2237.287
C(FCLTY_NM)[T.10]	-4008.4463	1615.498	-2.481	0.013	-7178.972	-837.920
C(FCLTY_NM)[T.11]	747.7013	2942.129	0.254	0.799	-5026.429	6521.832
C(FCLTY_NM)[T.12]	-331.5316	2718.748	-0.122	0.903	-5667.260	5004.197
C(FCLTY_NM)[T.13]	4813.5878	2182.948	2.205	0.028	529.402	9097.774
C(FCLTY_NM)[T.14]	-2874.2297	3466.814	-0.829	0.407	-9678.089	3929.630

회귀분석을 통한 R-squared 값으로 행사 인원

을 예측하기에 유의한 변수 선택

각 모델별 예측값 계산

	test	predict	pre_success
0	3080	4349.208131	False
1	5700	7408.204891	False
2	3130	2861.187453	True
3	3830	4241.838316	False
4	8643	6779.022805	False
5	12525	14214.587867	False
6	1404	2976.613164	False
7	870	357.162343	False
8	14675	17452.919816	False
9	1050	929.809078	False
10	1890	2360.351313	False
11	2230	5024.649365	False
12	350	201.712318	False
13	1500	3403.384487	False
14	992	895.564610	True

test: 실제 인원 | predict: 예측 인원

모델별 예측값을 통해 성능 평가

가중치 도출

iter	target	weights_1	weights_2	weights_3
1	0.05224	0.5206	0.3337	0.9334
2	0.05597	0.857	0.3194	0.725
3	0.05597	0.7681	0.07678	0.9861
4	0.1007	0.336	0.3546	0.1624
5	0.2388	0.4224	0.5784	0.0771
6	0.1828	0.4618	0.6824	0.03662
553	0.03731	0.8003	0.5042	0.783
554	0.08209	0.4654	1.0	0.0
555	0.03731	1.0	0.4329	0.6579
556	0.04104	0.5435	0.5095	0.9887
557	0.1381	0.0	0.8667	0.3731
558	0.1231	0.623	0.0	0.7533
559	0.1567	0.7611	0.2249	0.2589
560	0.2948	0.03398	0.1379	0.8992
561	0.04851	0.1837	0.8054	0.9983
562	0.2313	0.2044	0.3192	0.5286
563	0.2127	0.0	0.6995	0.3323
564	0.08955	0.328	0.2753	0.08829
565	0.09701	0.5922	0.7365	0.04927
566	0.02239	0.7618	0.9986	0.7071
567	0.1978	0.339	0.348	0.4697
568	0.07463	0.8066	0.3612	0.3074
569	0.1007	0.1454	1.0	0.2884

상위 성능 모델의 가중평균을 구하기 위한 가중

치를 베이지안 최적화를 통해 계산

APPENDIX.

SARIMA 적합 결과

Best SARIMA(1, 1, 1)x(0, 1, 1, 7) model >> AIC: 5012.041580211024

SARIMAX Results

Dep. Variable:	8	No. Observations:	334
Model:	SARIMAX(1, 1, 1)x(0, 1, 1, 7)	Log Likelihood	-2502.021
Date:	Mon, 02 Oct 2023	AIC	5012.042
Time:	11:20:13	BIC	5027.189
Sample:	05-02-2022	HQIC	5018.086
	- 03-31-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1551	0.046	3.353	0.001	0.064	0.246
ma.L1	-0.9418	0.025	-37.916	0.000	-0.991	-0.893
ma.S.L7	-0.9686	0.059	-16.451	0.000	-1.084	-0.853
sigma2	2.573e+05	9577.941	26.862	0.000	2.39e+05	2.76e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	4488.03
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	1.06	Skew:	0.37
Prob(H) (two-sided):	0.74	Kurtosis:	21.16