포스코 청년 AI·Bigdata 아카데미

영화 선별 알고리즘 향상을 통한 OTT 비즈니스 수익성 개선안

A분반 3조

김진명 김채은 김한빈 이경원 이다연 이상엽

CONTENTS

01
추진 배경

02
현황 및 개선기회

03
분석 계획 및 결과

04
개선안 적용 방안

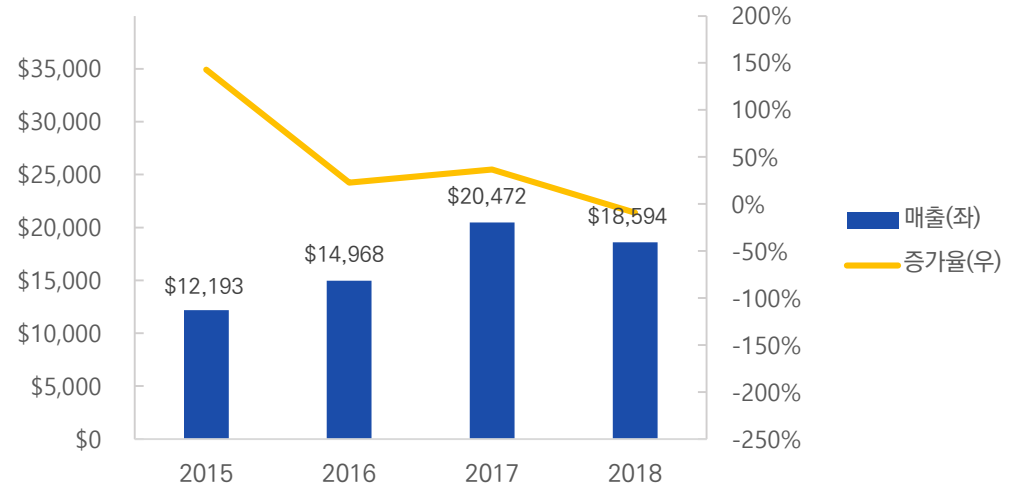
실적 부진과 어두운 전망으로 경영 위기에 처한 ㈜좋은 영화

경쟁 심화로 인한 실적 부진과 다운로드형 수익모델의 부정적 전망

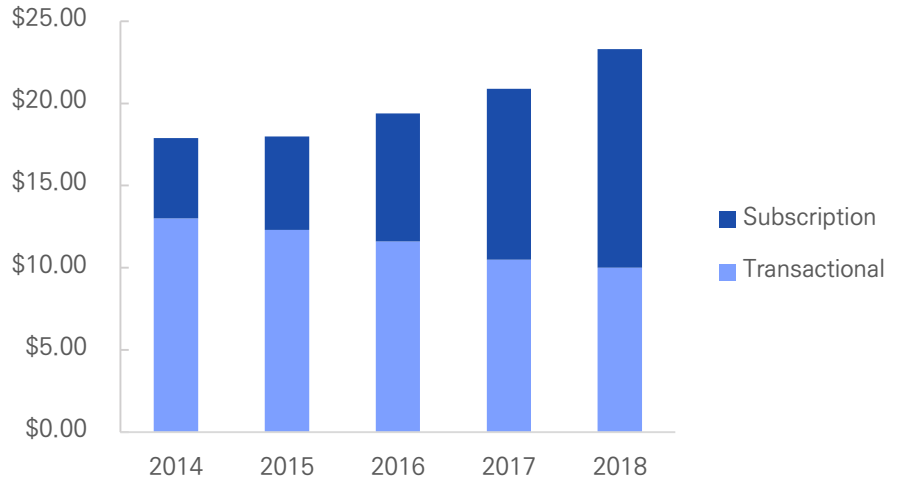
- OTT시장 경쟁 심화
  - 미국 내 OTT서비스 2014년 120개 → 2018년 230개
- ㈜좋은 영화 실적 부진
  - 매출액 2017년 $20,472 → 2018년 $18,594
  - 창사이래 첫 매출액 감소

- 다운로드형 모델 전망 악화
  - 다운로드형 모델을 한 번도 이용하지 않은 인구 56%
    (구독형 모델의 경우 26%에 불과)
- 미국 OTT 시장 내 다운로드 서비스 매출 감소
  - 소비자의 거래형 모델(VOD, EST 등) 지출 규모 감소
  - 반면 구독형 모델은 지속적으로 증가

㈜좋은 영화 연간 매출액 및 매출 증가율 추이



미국 디지털 엔터테인먼트 소비자 지출액



II. 현황 및 개선 기회

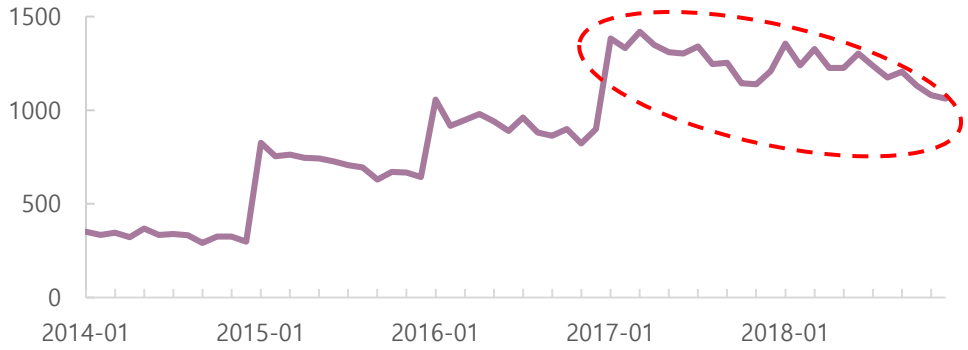
영화 추천 알고리즘 개선을 통해 매출 부진 해결 가능

㈜좋은 영화 매출 감소 원인 분석

원인 1 : 활동 유저 수 감소

- 2017년 1월 1,382명 → 2018년 12월 1,061명

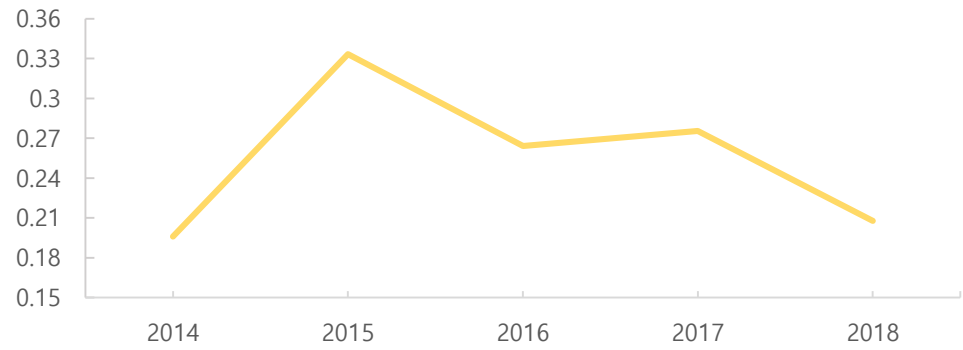
최근 5년간 월간활성사용자(MAU*) 추이



원인 2 : 계약 영화 수익성 감소

- 2015년 33.3% → 2018년 20.7%
- 투자금 회수에 걸리는 시간 3년에서 5년으로 증가

최근 5년간 계약 영화 ROI 추이



Solution : 영화 추천 알고리즘 개선

예상 효과 1 : 활성사용자 증가

추천 정확도 향상
↓
재방문율 증가
↓
사용자 및 사용시간 증가

예상효과 2 : 영화 수익성 개선

고객 반응이 높은 영화 선정
↓
ROI 상승
↓
투자금 회수 기한 단축 및 재투자 효율 증가

목표

1) 매출의 증가세 전환

2) 활성사용자 수 유지

*MAU(Monthly Active User): 1개월 간 한 번이라도 서비스를 이용한 유저 수

DB의 양적 질적 확장을 통한 분석 및 추천 알고리즘 구현

데이터 분석 계획

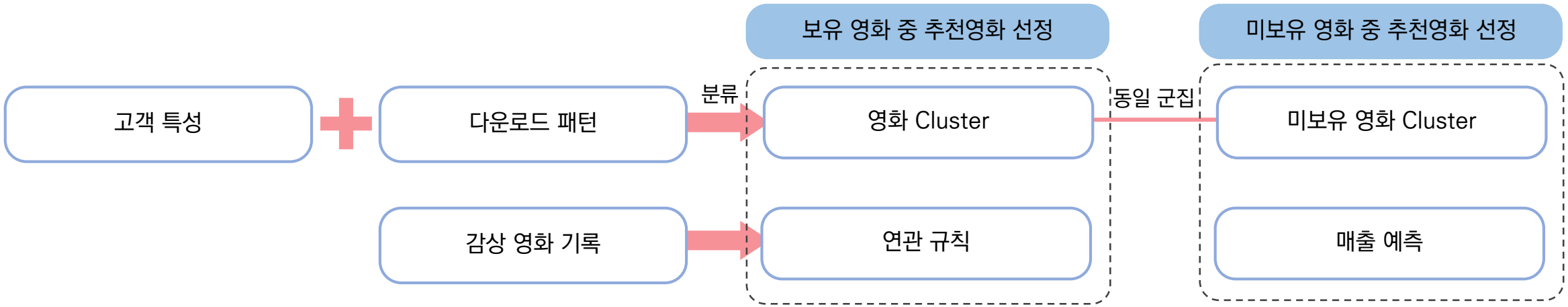
DB확장	EDA	분석	
Web Scraping	영화 DATA	군집분석	전체 DB 영화별 군집 형성
	고객 DATA	연관규칙	보유 영화의 다운로드 DATA 활용
영화 줄거리 감성분석	다운로드 DATA	예측	고객 특성을 이용한 고객별 매출 예측
	가격 DATA		영화 특성을 이용한 영화별 매출 예측

추천 알고리즘

군집분석과 연관규칙을 활용한 영화 추천 및 매출 예측

데이터 분석 계획

추천 알고리즘



스크래핑과 텍스트 분석을 통한 데이터의 양과 질 증가

분석 결과

Web Scraping

- IMDb, The Numbers 등 영화 DB 웹사이트를 스크래핑하여 변수(feature) 및 관측치 수정·보완
  - # of Obs : 2,184 → 4,560
  - # of Features : 31 → 57 (metascore, 언어, 개봉 첫 주 스크린 수, 시리즈, 원작, 제작방식, 인플레이션 조정 매출, 수상기록 등)

영화 줄거리 분석을 통한 감성분석

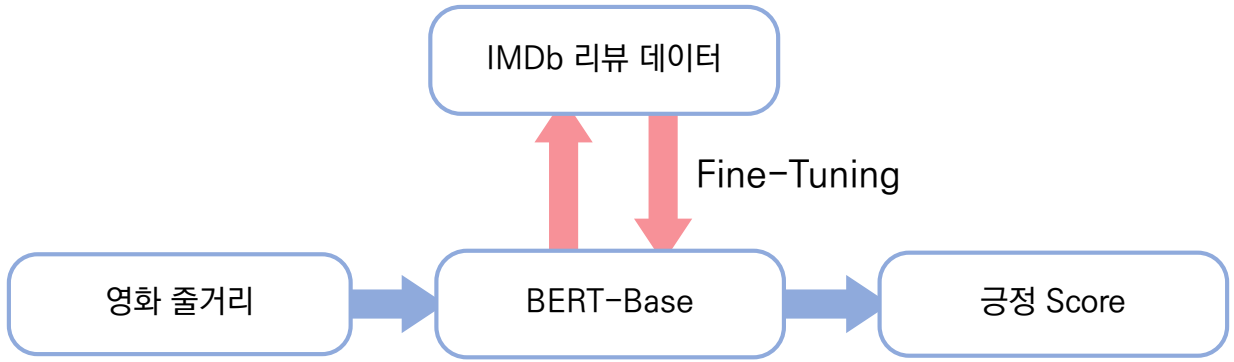
- 영화의 줄거리를 긍정도에 따라 점수화
- 임베딩 모델: BERT vs. Word2Vec
  - 문장 단위 분석에 사용되는 BERT를 활용하여 분류
  - 단어 기반의 Word2Vec은 줄거리 분석에서 분류 성능이 떨어짐

ex)

Movie	Word2Vec	BERT
Avengers: Infinity War*	0.5	0.155

*해당 영화는 인류 절반이 죽는 비극적 줄거리이다.

Sentiment Analysis(감성 분석)



(주)좋은 영화의 보유 영화는 전체 영화와 다른 특징을 가진 데이터로 분석 및 예측에 유의

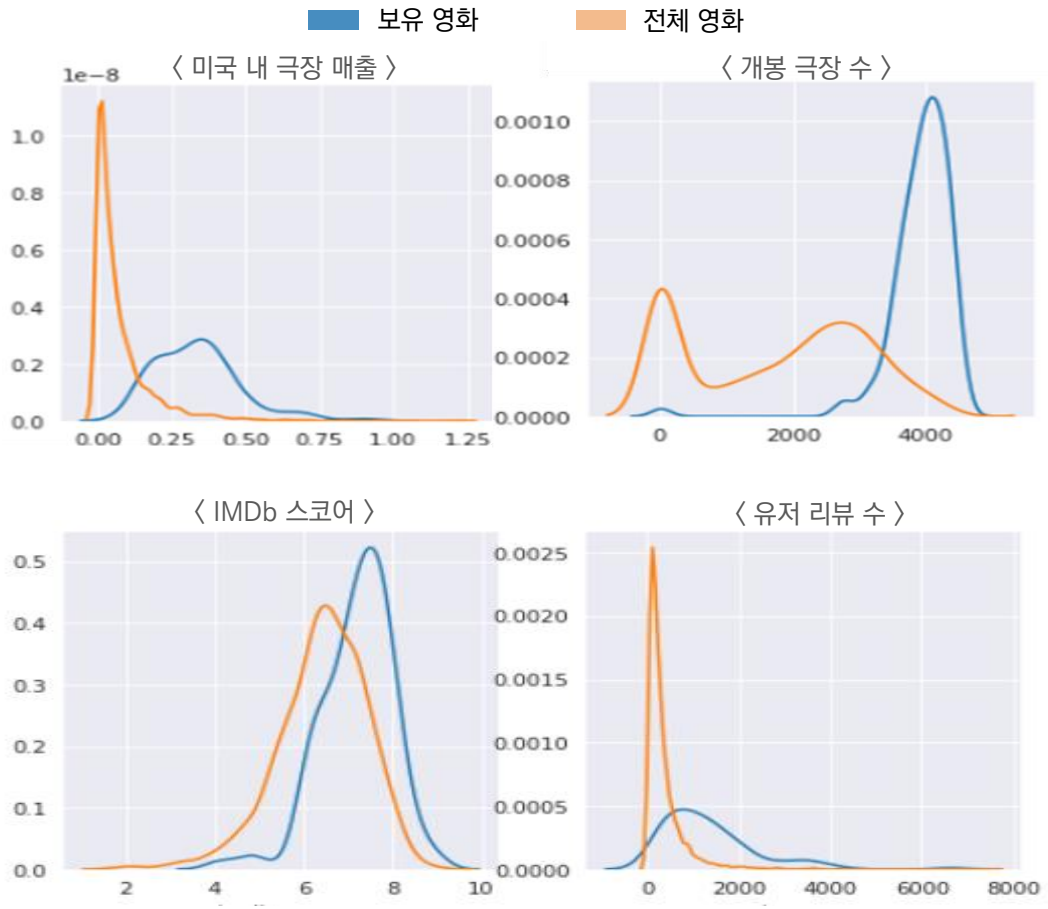
EDA

㈜좋은 영화가 보유하고 있는 영화는 극장에서 흥행 성공한 영화

- 보유 영화와 전체영화의 성질이 다르므로 분석에 유의

	보유 영화	전체 영화
개봉연도 평균	2013	2003
주요 장르	액션, 어드벤쳐, 판타지, SF	드라마, 코미디
MPA 등급 최빈값	PG-13 (전체관람가)	R (17세미만 단독관람불가)

보유 영화와 전체 영화 DATA 분포 비교



군집분석을 통해 다중 분류 모델의 목표변수 생성

군집분석

PCA

- 변수 개수가 너무 많아(57개) 차원의 저주 발생 가능성

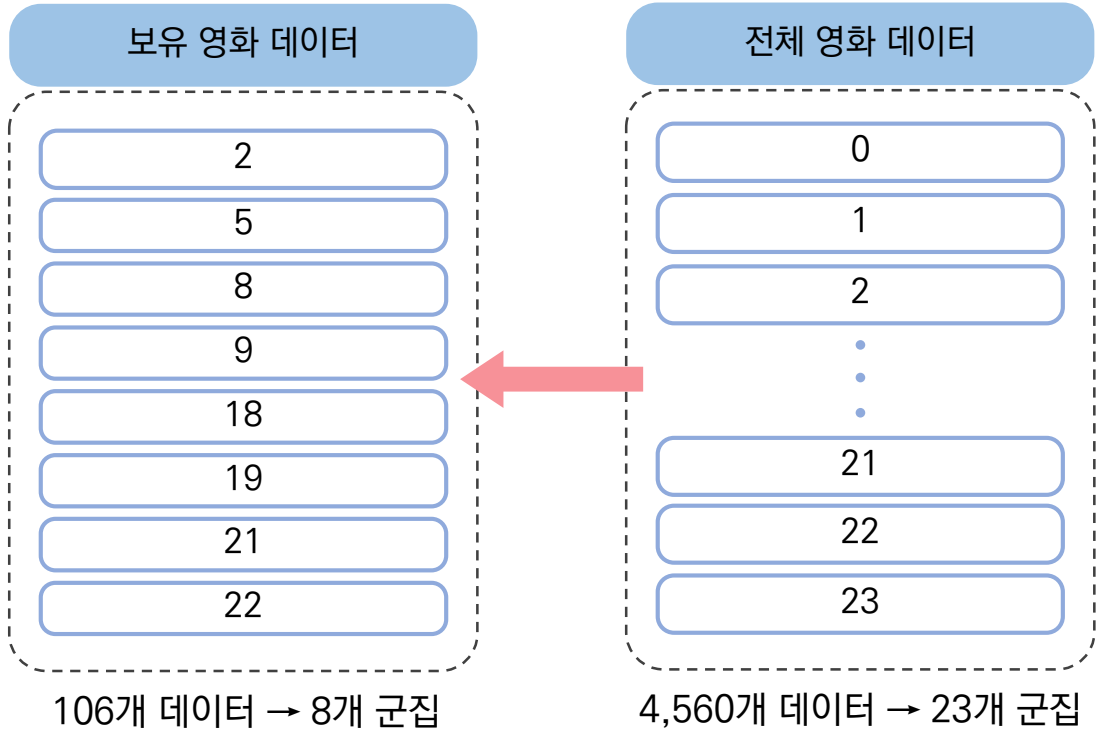
  → PCA를 통해 Feature 축소(차원 축소)

- 주성분의 Eigen value가 1이상인 주성분만 선택(6개)

K-means

- PCA를 통해 산출된 주성분 6개로 군집분석 진행
- Silhouette 계수의 값이 가장 큰 23개로 군집 결정

변수 개수 57개 → 6개로 축소



연관규칙의 전반적인 지지도가 낮으므로 추천 알고리즘의 보조적 수단으로 사용

연관 규칙

- 연관규칙 산출 조건
  - 지지도 : 0.1, 신뢰도 0.5 임계치 설정
  - 분석 대상 고객 수 : 3,277, 분석 대상 Item 수 : 106

지지도 기준 상위 규칙 5개

선행 사건	후행 사건	선행 지지도	후행 지지도	지지도	신뢰도	향상도
American Sniper	The Hobbit: The Battle of the Five Armies	0.445	0.444	0.276	0.6198	1.3946
The Hobbit: The Battle of the Five Armies	American Sniper	0.444	0.445	0.276	0.6203	1.3946
American Sniper	Guardians of the Galaxy	0.445	0.448	0.265	0.5960	1.3311
Guardians of the Galaxy	American Sniper	0.448	0.445	0.265	0.5922	1.3311
The Hunger Games: Mockingjay – Part 1	The Hobbit: The Battle of the Five Armies	0.444	0.444	0.264	0.5959	1.3406

신뢰도 기반 추천

American Sniper를 다운로드한 고객 중, 62%의 고객이 The Hobbit을 다운로드	1순위 추천
American Sniper를 다운로드한 고객 중, 59%의 고객이 Guardians of the Galaxy를 다운로드	2순위 추천

영화 특성을 통한 영화별 매출 예측

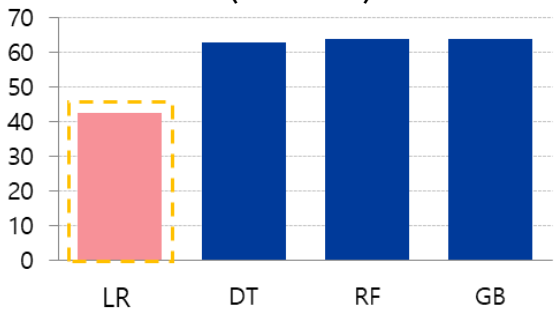
영화별 매출 예측

낮은 적합도로 예측의 정확도는 낮으나, Vital Few 확인 가능

다중회귀모형

- 모든 평가지표에서 다중회귀 분석이 ML 분석 결과에 앞서므로 다중회귀모형 채택

〈 RMSE 〉



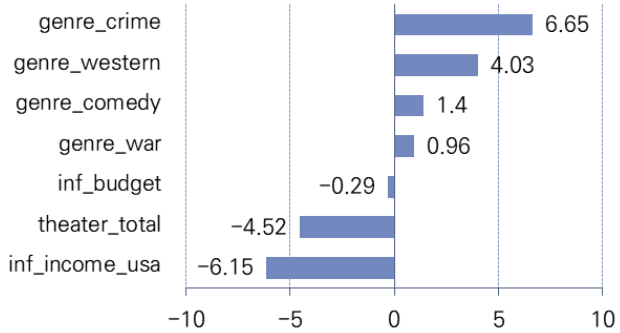
낮은 예측력

- 설명력이 가장 높은 모델임에도 불구하고 *adj.* R^2는 0.440으로 낮은 편
- 이는 106개라는 절대적으로 적은 관측치 개수에 기인한 것으로 보임

R^2	$adj.R^2$
0.594	0.440

Vital Few

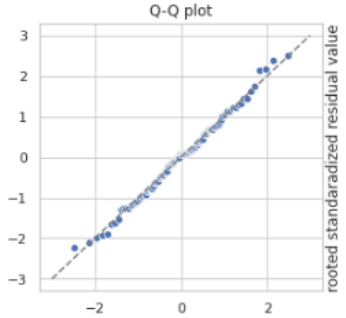
- 통념과 다른 결과, 적절한 해석 필요

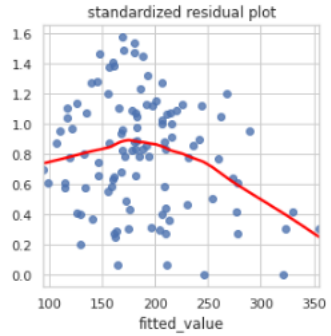


- 대중적이지 않은 장르에서 더 높은 매출 / 극장 성적이 더 낮은 영화에서 더 높은 매출

잔차는 정규분포를 따르고 독립적

- 낮은 예측력과 무관하게 모형은 올바름





고객별 매출 예측 및 Activation_day로 목표변수 대체

고객별 매출 예측

고객별 매출 예측

- 고객별 특징에 따라 고객별 예상 매출 예측 및 Vital Few 확인
- 회귀 모델의 적합도 값(0.04)이 너무 낮아 설명력 부족
  - 이는 변수들 간의 상관성이 낮아서 생기는 문제로 파악
- 고객별 매출을 대체하기 위한 변수로 Activation_day를 예측하는 방안으로 진행

Activation_day* 예측

- 고객 매출 예측의 대안으로써 고객 활동 일수 예측모델 생성
- 고객별 특성을 대상으로 활동일수 예측 모델링 및 Vital Few 확인
- 최종 설명력은 0.544이나 LASSO 정규화 이후 0.577까지 상승

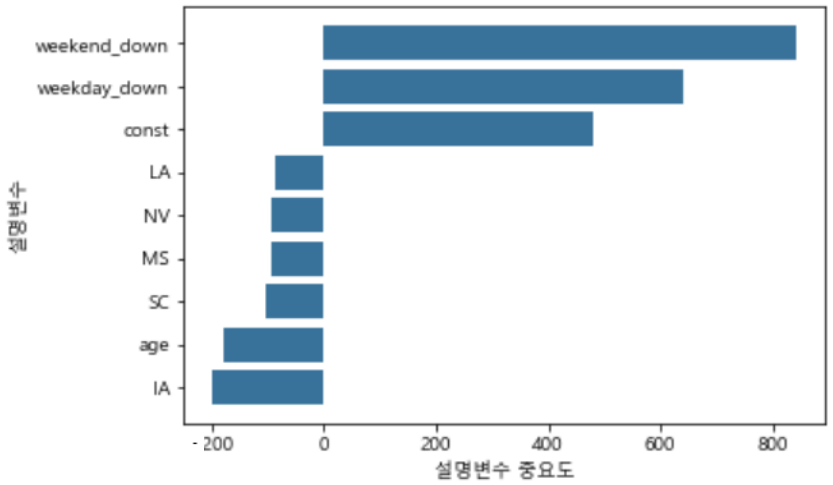
*Activation_day : 첫 다운로드 날짜부터 마지막 다운로드 날짜까지의 기간

Activation_day 예측

예측 모델

R^2	$adj.R^2$
0.545	0.544

Vital Few



분류 모델을 통한 고객 특성별 영화 군집 번호 부여

고객 특성에 따른 영화 군집 분류

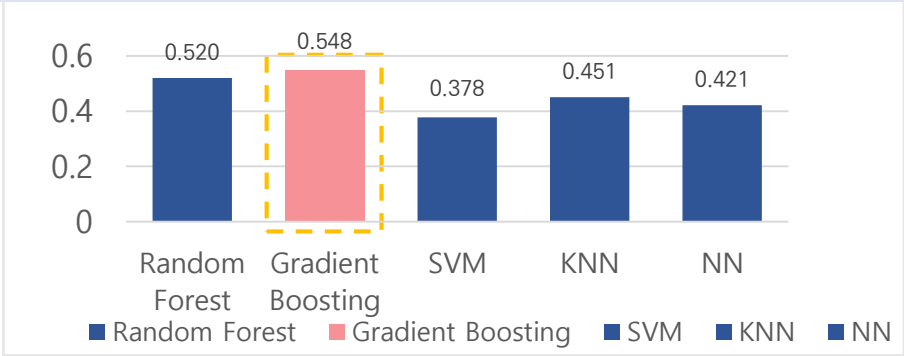
분류 모델 학습

- 학습 데이터 : 고객별 특성 정보와 다운로드 내역을 포함한 데이터
  - 고객 데이터와 다운로드 데이터를 병합
  - 변수 54개 X 관측치 73085개
  - train set : test set = 0.7 : 0.3 으로 나누어 학습

	down_year	weekday	down_price	gender	age	married	kids_under12	register_year	drop_flag	customer_sales	...
0	2014	6	0.946685	0	51	1	0	2014	1	10.997323739842438	...
1	2014	6	0.946685	0	51	1	0	2014	1	10.997323739842438	...
...	...	...	...	...	...	...	...	...	...	...	...
73083	2018	7	1.000000	1	64	0	0	2018	0	3.8	...
73084	2018	5	1.000000	1	64	0	0	2018	0	3.8	...

다운로드 내역 정보 | 고객 특성 정보

모델 선정 : Gradient Boosting



학습 결과 분류모델 중 test score가 가장 높았던
Gradient Boosting을 최종 모델로 선정

모델 적용 및 활용

- 고객 특성 및 다운로드 이력에 알맞은 영화 군집 번호 부여
  → 고객의 취향을 반영한 동일 군집 내의 다른 영화 추천 가능
  : 영화 추천 알고리즘 성능 향상에 기여

고객별 영화 군집 분류 및 연관규칙을 통한 영화추천 알고리즘 개선

영화 추천 알고리즘 개선

1단계
고객 특성별 영화 군집 번호 부여

고객 특성 데이터 → 분류모델 → 영화 데이터 군집 번호

- 입력 : 다운로드 내역 데이터를 포함한 고객 특성 입력
  - 고객의 인구사회학적 특성, 다운로드 내역 등의 정보로 이루어진 데이터
- 분류모델 : 다운로드 + 고객 데이터를 병합한 7만 개 데이터를 통해 학습
- 출력 : 고객 특성을 보유영화 데이터 내 군집 번호로 분류

2단계
보유 영화 중 동일 군집 내 다른 영화 추천

- 연관규칙을 통한 우선 추천
  - 동일 군집 내에 선행 사건에 해당하는 영화 존재 시 신뢰도가 높은 후행 사건 우선 추천
- 해당 고객이 이미 본 영화는 추천 대상에서 제외
  - 1단계 분류모델 입력 정보에 포함되어 있는 다운로드 내역 사용

3단계
나머지 영화들에 대해 예상매출 기준 추천

- 동일 군집 내 영화들 중 이전 단계에서 추천 대상이 아닌 영화들은 **예상매출 순**으로 추천
  - 연관분석을 통해 추천되지 않은 영화들, 보유하지 않은 영화들 등

고객별 영화 군집 번호 부여 및 연관규칙을 통한 영화추천 알고리즘 개선

추천 알고리즘 예시

1단계

고객 A 정보 입력 → 분류모델 → 번호 3 출력

고객 A의 특성에 맞는 영화 군집 = 군집번호 #3

2단계

고객 A가 이미 본 영화 제거

영화 ID	보유 여부
tt01	1
tt02	1
~~tt03~~	~~1~~
tt04	0
~~tt05~~	~~0~~
tt06	0
tt07	0

군집 번호 #3 영화 목록

선행	후행	지지도	신뢰도	향상도
tt01	…	…	…	…
tt02	…	…	…	…
tt03	tt02	0.5	0.7	1.1
tt03	tt01	0.5	0.5	1.1
tt04	…	…	…	…
…	…	…	…	…

연관규칙 결과

신뢰도 0.7 〉 0.5 이므로
tt01보다 tt02 우선 추천

3단계

[고객 A에 대한 영화 추천 최종 순서]

2단계 과정에 따른 우선 추천 tt02 – tt01 – tt07 – tt06 – tt04

보유하지 않은 영화를 비롯한 그 밖의 경우
예상 매출 순으로 추천

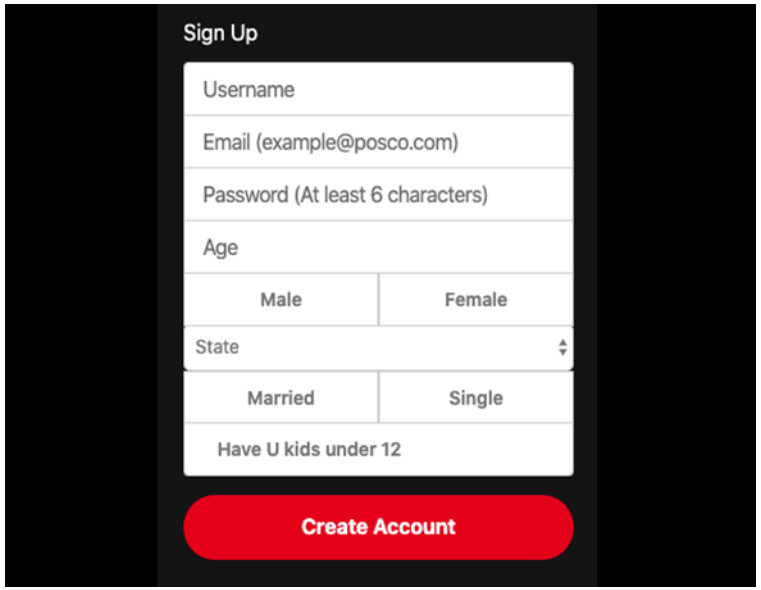
개선된 영화추천 알고리즘을 이용한 웹 서비스 구현

영화 추천 서비스 예시

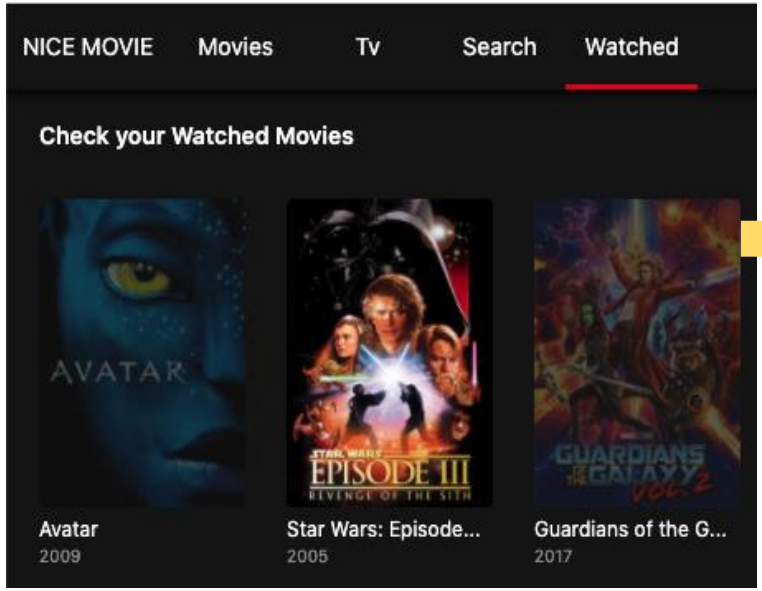
- 회원 가입 시 입력되는 고객 특징과 고객의 관람 정보를 조합하여 추천 영화 제시

Input

고객 정보 입력

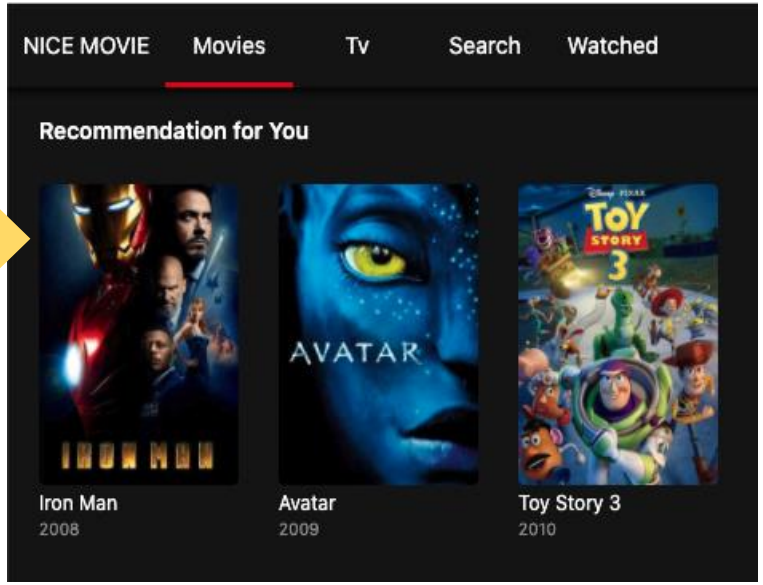


고객 관람 정보



Output

영화 추천



기대효과

1) 서비스 체류 시간 증가

2) 활성 사용자 이탈율 감소

ROI 기준 영화 선별 시 가능한 매출 상승 시나리오 구성

영화 선별을 통한 ROI 증가

전체 영화 정보 필터링

- 보유 영화와 같은 특징(흥행 영화)을 가진 영화만 예측
- 기준: 미국 매출, 제작 예산

영화별 연간 매출 예측

- 다중회귀모델로 추산한 회귀식 이용

ROI 추산

- $\text{ROI} = \frac{\text{연간 예상 매출}}{\text{추정 판권 가격}}$

투자

- ROI 순으로 정렬하여 높은 것부터 순서대로 판권 구입

향후 3년간 매출 추이 시뮬레이션

고객 수 가정

- 추천 알고리즘 개선으로 고객 수 유지
- 추천 알고리즘 실패로 고객 수 감소(18년도 이탈율(9%)로 감소)

매출 예측 가정

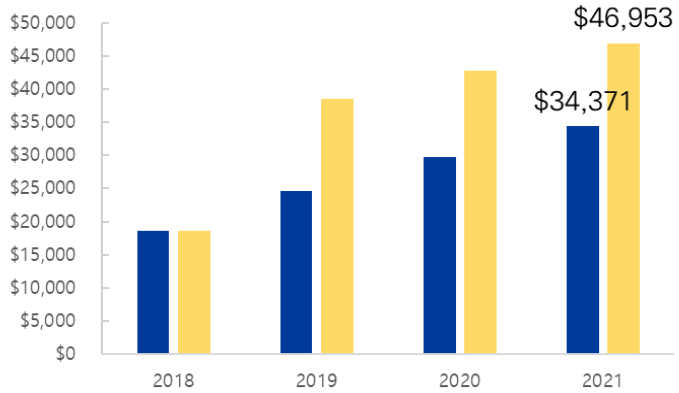
- 영화 선별 성공 : 예상된 ROI 대로 매출 증가
- 영화 선별 실패 : 18년도 계약 영화 ROI로 매출 증가

	고객 수 유지	고객 수 감소
영화 선별 성공	시나리오1	시나리오2
영화 선별 실패	시나리오3	시나리오4

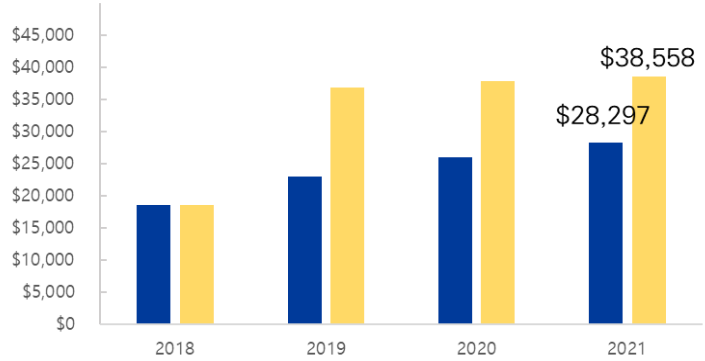
목표 달성 : 시나리오4의 적극적 투자 가정을 제외하고 매출 증가 달성

향후 3년간 매출 추이 시뮬레이션

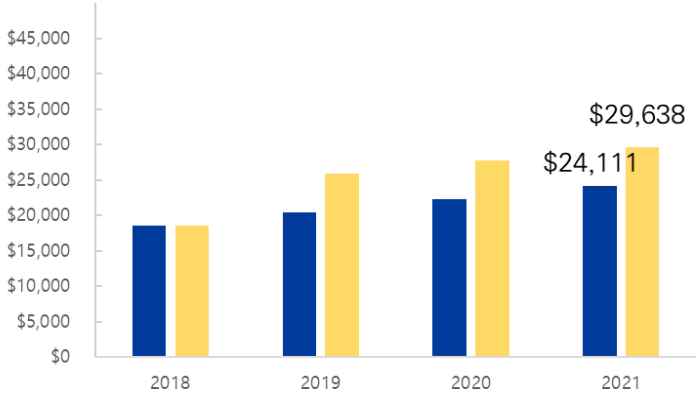
시나리오1 : 고객 수 유지 성공 & 영화 선별 성공



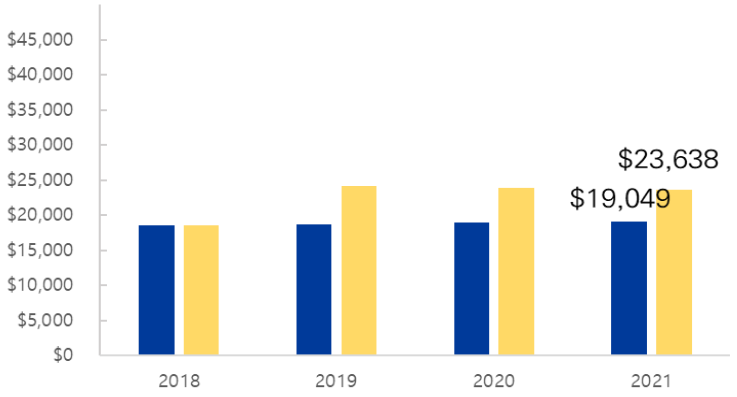
시나리오2 : 고객 수 유지 실패 & 영화 선별 성공



시나리오3 : 고객 수 유지 성공 & 영화 선별 실패



시나리오4 : 고객 수 유지 실패 & 영화 선별 실패



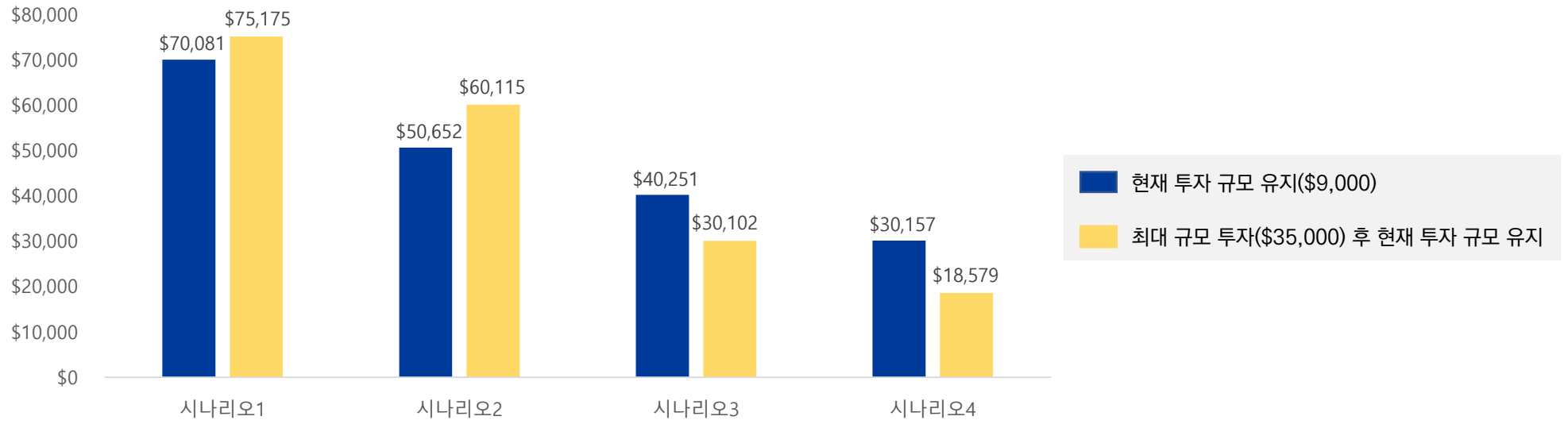
현재 투자 규모 유지($9,000)

최대 규모 투자($35,000) 후 현재 투자 규모 유지

CEO의 판단에 따라 적절한 투자 규모 선택 필요, 단 선택의 리스크는 크지 않은 것으로 판단

향후 3년간 매출 추이 시뮬레이션

투자금 제외 3년간 순수익 비교



- 시나리오1과 2는 적극적 투자 시 순수익이 더 크고, 시나리오 3과 4의 경우 소극적 투자 시 순수익이 더 큼
- CEO는 각 시나리오 타당성을 검토하고 가장 가능성이 높다고 판단되는 시나리오 선택
- TF팀의 입장 : 어떤 시나리오에도 ㈜좋은 영화의 순수익이 감소하는 경우는 없으므로 적극적 투자 권장.
  단, 다운로드 수익모델은 장기적으로 한계가 있으므로 BM 전환을 위해 자금을 전용하는 경우 소극적 투자도 합리적 판단으로 보임.

Lesson Learned

공통 교훈

데이터 수집 및 전처리	우리 팀 프로젝트 중 가장 많은 시간과 노력을 할애한 부분, 이상치를 제거하고 결측치를 대체함에 있어서 뚜렷한 기준보다는 경험을 바탕으로 한 직관이 필요
평가지표 활용	어떤 상황에서도 완벽한 기준이 되는 평가지표는 없음. 따라서 진행하는 모델링의 유형과 목적을 명확히 파악하고 가장 적절한 평가지표를 활용해야 함
목표 설정	영화 추천 개선도처럼 정량적 평가가 어려운 경우, 목표 설정이 난해함. 본 과제에서는 고객 수 유지를 목표로 설정하였으나 실제 관측이 불가능한 지표이므로 정량적 측정이 불가능한 것은 여전히 아쉬운 부분으로 남음
도메인 지식	분석 기법의 도구로서의 중요성을 간과할 수는 없지만, 직관이 수반된 도메인 지식이 없다면 의미 없는 분석만 하게 될 것임
분석 기준	협업을 함에 있어서 구성원이 합의한 뚜렷한 기준으로 분석을 진행해야 추후 작업물을 병합하는 과정에서 이상이 없음

개별 소감

- 김진명 : 데이터 분석은 단순한 기법이 아니라 의사소통을 통해 문제해결에 도움을 주고 협업하는 과정이라는 생각이 들었습니다.
- 김채은 : 통계와 데이터분석을 처음 배우면서 머릿속에서 산재된 지식들이 정리되어가는 과정이었습니다. 데이터 전처리 및 정제에 많은 에너지를 쓰면서 양질의 데이터가 분석에 매우 중요한 요소임을 절감했고, 다양한 방법과 접근의 분석을 시도했지만 최종적으로는 극히 일부만이 선정되는 것을 보며 분석가에게는 높은 질과 양의 경험이 중요한 자산임을 느꼈습니다.
- 김한빈 : 한때 EDA의 필요성을 무시한 적이 있으나 이번 프로젝트를 통해 그 중요성을 절감했습니다. 또한 프로젝트를 진행하며 제 자신의 부족함을 더 느끼고 각성하는 계기가 되었습니다.
- 이경원 : 실제 데이터를 활용한 분석을 해보니 정제된 데이터로 분석하는 것보다 많은 어려움이 있었습니다. 또한 데이터 분석을 위해서는 전체적인 데이터에 대한 이해가 필수적이라는 사실을 경험할 수 있었습니다.
- 이다연 : 실제 데이터로 분석하면서 시각화, 모델링도 중요하지만 데이터 정제 및 전처리과정이 중요하다는 것을 느꼈습니다. 또한 왜 이 분석이 요구되는 지 그 필요성을 사전에 인지하고 분석하는 과정이 필수적이라고 몸소 느끼는 시간이었습니다.
- 이상엽 : 빅데이터 기반의 분류, 예측을 통해 데이터 분석이라는 새로운 분야를 접해볼 수 있어 흥미로웠고, 해당 분야에 대한 지식이 미흡 했지만 조원들과 의견을 나누며 많은 걸 배울 수 있는 시간이었습니다.

감사
합니다

PRESENTOR KIM JIN MYEONG (TEAM 3)

PROJECT POSCO AI Big Data Project

DATE 2020.05.11