데이터분석, machine learning, deep-learning

Portfolio

지원자: 최은비
연락처 : 010-6341-1204
e-mail : p5653ceb@gmail.com
github : github.com/p5653ceb

Contents

1. Brunch 사용자를 위한 추천시스템

[machine learning] Brunch 사용자를 위한 추천시스템 개발

프로젝트 상세

프로젝트 내용

- Brunch 사용자의 구독 정보를 기반으로 새로운 글을 추천하는 추천 시스템 개발

활용 솔루션

- 3000명의 추천 대상 그룹에게 동일한 방법으로 추천을 해주는 것은 의미가 없을 것으로 판단
- 대상그룹을 세분화 하여 각자 다른 방식으로 추천

1) following based 추천
- 전체 독자의 98%는 follow하는 독자가 있음

2) magazine based 추천
- magazine의 글을 읽는 경향이 더 큼

3) popularity based 추천
- 등록된지 2주가 지난 후에는 글 소비 수가 거의 0에 가까워짐
- brunch platform은 지식 전달의 목적보다는 일상이나 자신의 생각을 전달하는 글이 대부분으로, steady seller와 같은 글이 거의 없음

4) interest based 추천 (collaborative filtering)
- keyword list를 활용하여 독자가 읽은 글의 키워드와, 작가가 쓴 글의 키워드로 취향이 유사한 작가와 독자를 선별하여 읽은 글 추천

주요 역할

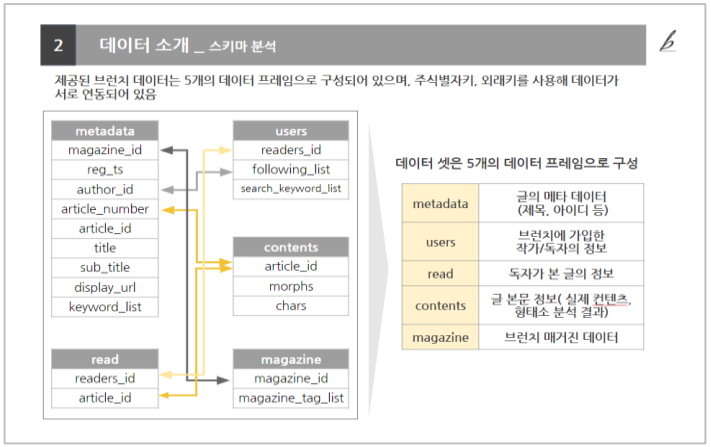
1) 데이터 전처리 및 EDA
- 최근 2주 이내 소비된 글을 보는 경향 : 글 추천 기간 한정
- 대상 그룹 구분 : 단 한 편의 글만 읽은 사람과 여러편의 글을 읽은 사람에게 추천은 다른 방식으로 진행되어야 함

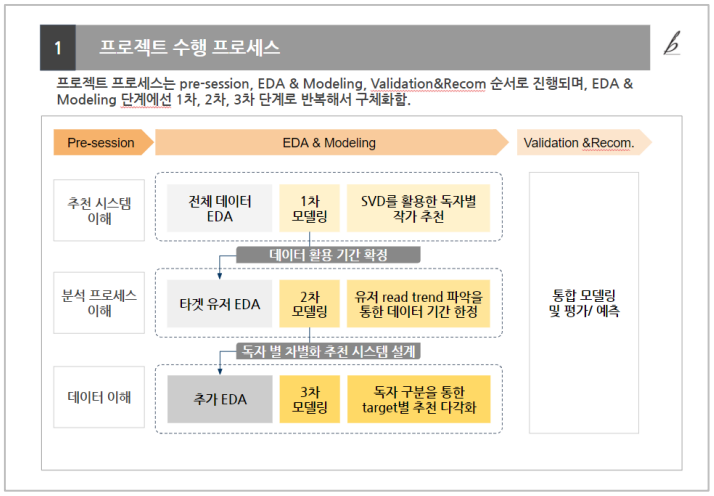
2) 협업필터링 모델 개발
- 독자-독자, 독자-작가 간 취향 유사도 확인
- tf-idf를 활용한 vector화 및 cosine 유사도 확인
- Doc2vec을 활용한 유사도 확인
- 취향 유사독자, 유사작가 10명씩을 뽑아 유사 독자가 읽었거나, 유사 작가가 쓴 최신/인기글 100개 추천

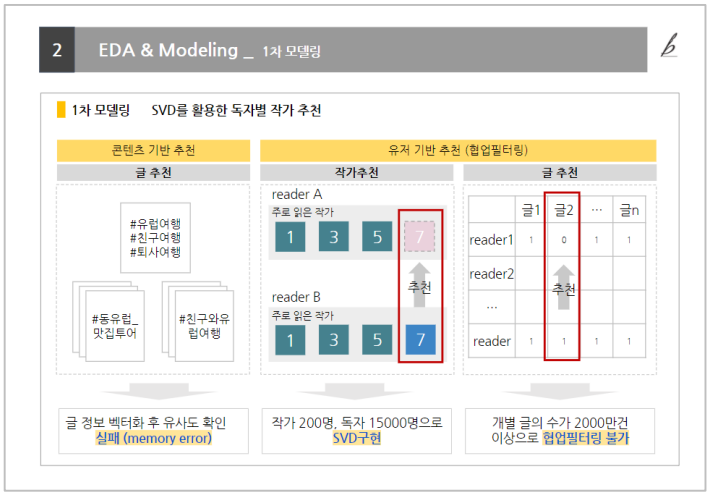
※ 깃헙 주소: https://github.com/p5653ceb/Regression_3_project

[machine learning] Brunch 사용자를 위한 추천시스템 개발

데이터 탐색

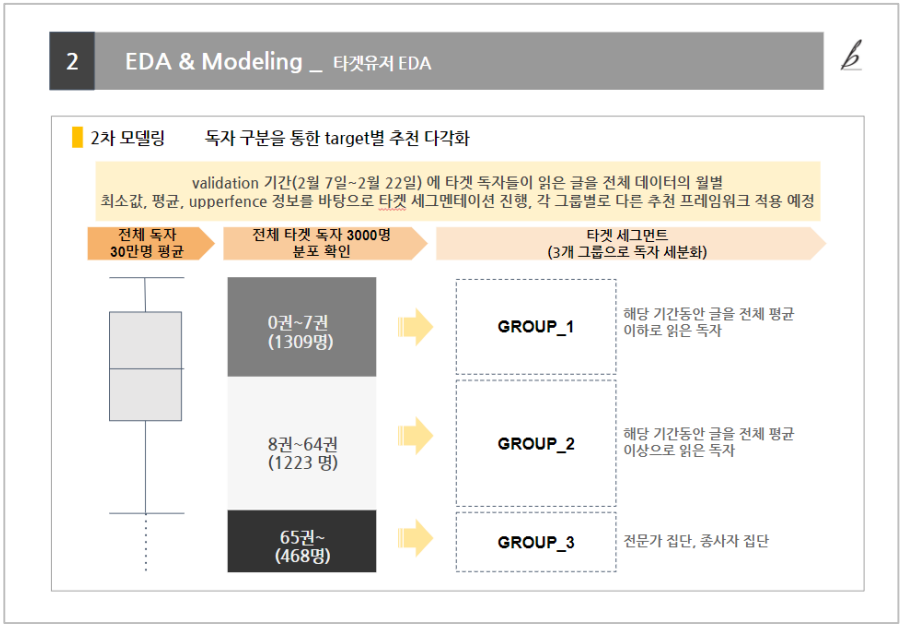


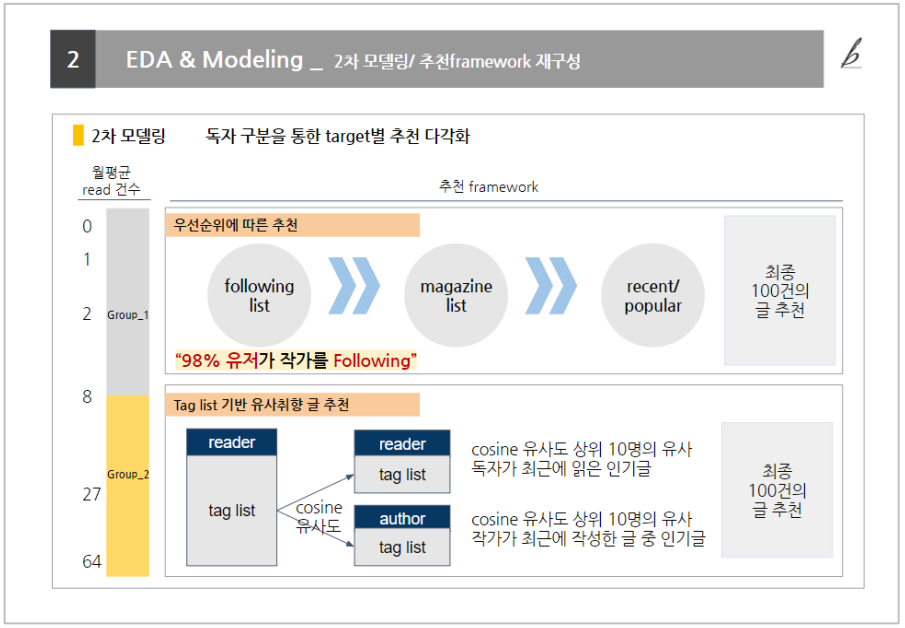




[machine learning] Brunch 사용자를 위한 추천시스템 개발

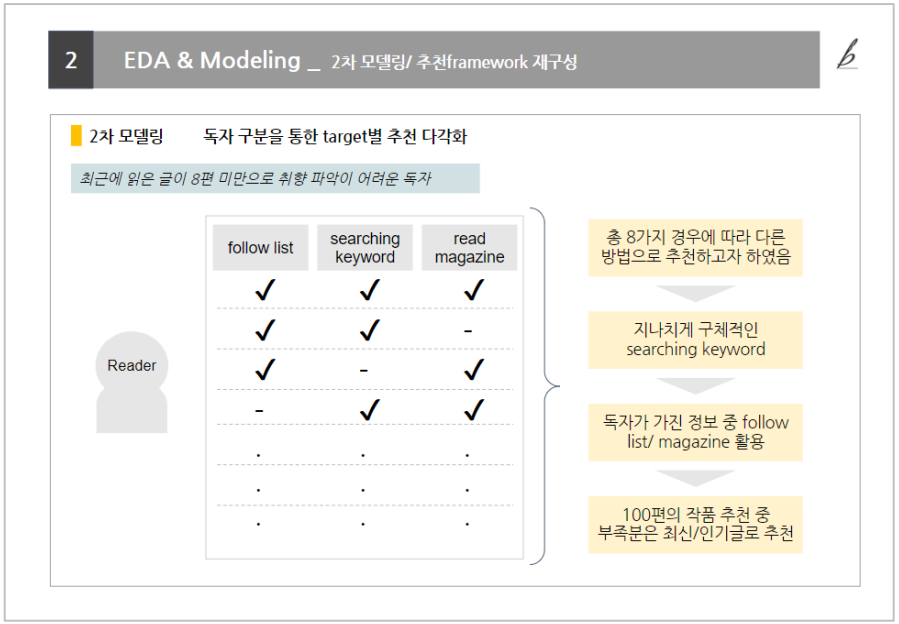
타켓유저 구분

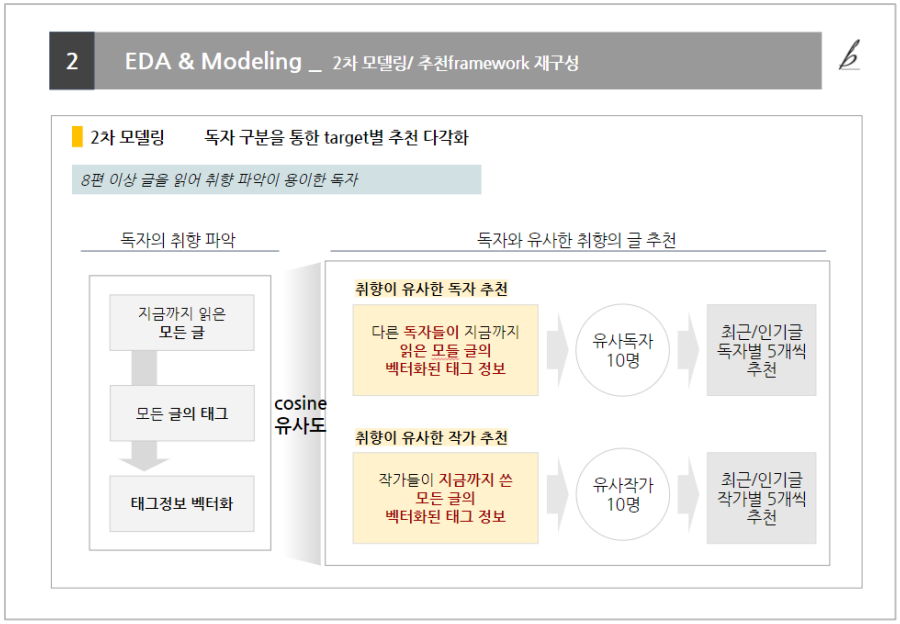




[machine learning] Brunch 사용자를 위한 추천시스템 개발

타겟별 추천 다각화





[machine learning] Brunch 사용자를 위한 추천시스템 개발

vectorizer활용

[Deep-Learning]

2. YOLO v5를 이용한 CCTV 인물 detection

[deep learning] YOLO v5를 이용한 CCTV 인물 detection

프로젝트 상세

프로젝트 내용

- 입구에서 촬영된 CCTV영상으로 학습 후 실내 CCTV에서 동일인물을 detection 하도록 하는 프로젝트

활용 솔루션

- YOLO v5
  - YOLO의 가장 최신 버전인 v5활용
  - 더 적은 양의 사진으로 더 정확하고 빠르게 학습이 필요하여 v5를 선택
  - 학습은 YOLO v5s를 사용하여 진행
- data augmentation
  - 선명하게 나온 사진을 많이 활용하기 위해 영상을 다시 자르지 않고 선명한 사진을 증강함
  - 1장의 사진을 12장의 사진으로 증강하여 학습에 사용 총 9장의 사진을 108장으로 증강하여 학습에 사용
- train
  - 1000 epoch , batch-size 64 설정
  - weights는 설정 없이 학습 진행 (학습 중 기존에 사용했던 best weight를 사용해 보았으나 성능차이가 거의 없었음)

주요 역할

1) 데이터 좌표 변경 및 라벨링
   - 기존의 json형식의 좌표를 YOLO v5에서 요구하는 좌표의 순서에 맞게 txt파일로 변경 (class, x, y, w, h)
   - BlackPink 데이터에 대해서는 YOLO mark를 사용하여 라벨링을 진행

2) 학습 및 augmentation
   - YOLO v5를 활용하여 학습
   - data augmentation 진행 : 기존에 가진 좌표도 함께 augmentation하기 위해 imgaug모듈 사용

※ 깃헙 주소: https://github.com/p5653ceb/deeplearning-repo-2

[deep learning] YOLO v5를 이용한 CCTV 인물 detection

CCTV dataset

2. Custom Dataset with YOLO v5

prepare dataset

json

```
"mark": [
  {
    "coordinates": [
      [
        830.4345999999999,
        298.59640000000013
      ],
      [
        920.9345999999999,
        298.59640000000013
      ],
      [
        920.9345999999999,
        567.5964000000001
      ],
      [
        830.4345999999999,
        567.5964000000001
      ]
    ],
    "label": {
      "Category": "pedestrian",
      "Clear face": "no",
      "Truncated": "non truncated",
      "Orientation": "left",
      "Occluded": "largely occluded"
    }
  }
],
"title": "10541_market/frames/01/00003.jpg"
```

```
# text file로 만들기
for filename in json_name_ls:
    with open(filename, 'r') as file:
        json_data = json.load(file)
    with open(filename.split('.')[0] +'.txt', 'w') as f:
        x = round(((json_data['mark'][0]['coordinates'][0][0]/1920) + (json_data['mark'][0]['coordinates'][1][0]/1920))/2, 3)
        y = round(((json_data['mark'][0]['coordinates'][0][1]/1080) + (json_data['mark'][0]['coordinates'][2][1]/1080))/2, 3)
        w = round((np.abs(json_data['mark'][0]['coordinates'][1][0]-json_data['mark'][0]['coordinates'][0][0]))/1920, 3)
        h = round((np.abs(json_data['mark'][0]['coordinates'][2][1]-json_data['mark'][0]['coordinates'][0][1]))/1080, 3)
        cl = filename.split('/')[-1].split('-')[0] # market명에 번호를 class번호 적어줌
        a = [cl, x, y, w, h]
        f.write(" ".join(map(str, a)))
```

01-00004.txt

```
0 0.4095303125 0.43210259259259254 0.060416666666666674 0.2537037037037037
```

txt

train.txt

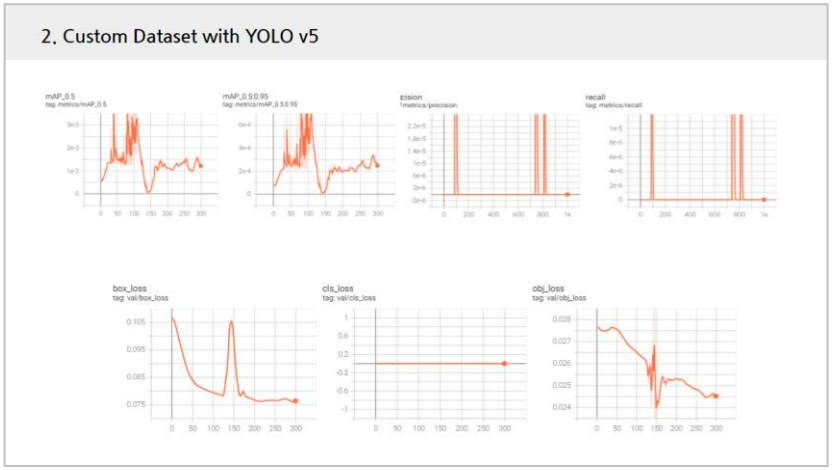
```
/content/drive/MyDrive/Data/blackpink/origin_data/frame208.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame199.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame202.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame204.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame201.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame206.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame203.jpg
/content/drive/MyDrive/Data/blackpink/origin_data/frame210.jpg
```

data.yaml

```
names: ['person']
nc: 1
train: /content/drive/MyDrive/Data/train.txt
val: /content/drive/MyDrive/Data/val.txt
```

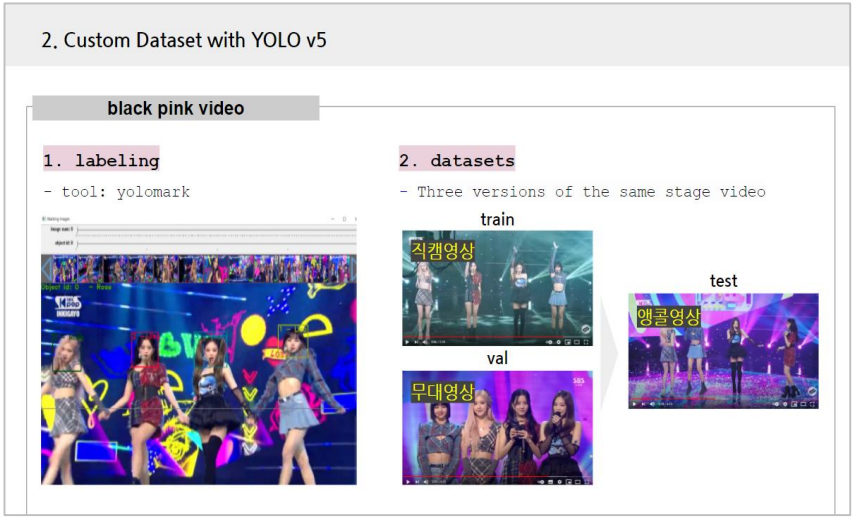
[deep learning] YOLO v5를 이용한 CCTV 인물 detection

augmentation



[deep learning] YOLO v5를 이용한 CCTV 인물 detection

BlackPink dataset




```
# 전체데이터를 train 시키기
!python train.py --data '/content/drive/MyDrive/Data/blackpink/data.yaml' --epochs 1000 --cfg '/content/drive/MyDrive/yolov5/models/yolov5s.yaml' --weights '' --batch-size 64 --name blackpink_detection
```

[deep learning] YOLO v5를 이용한 CCTV 인물 detection

결과

[deep learning] YOLO v5를 이용한 CCTV 인물 detection

결과

[Machine-Learning]

3. 여성 패널 데이터 기반 여성인력 임금 예측

[machine learning] 여성 인력 임금 예측

프로젝트 상세

프로젝트 내용

- 여성가족부 패널데이터를 활용, 여성인력의 임금을 예측하는 프로젝트

활용 솔루션

- RandomForest Regressor 사용 (max depth 3)

1) test 결과값을 확인할 수 있는 machine learning model을 활용

- R2 수치는 OLS에서 높게 나타났지만 train 결과에만 한정된 수치임을 확인하고 다른 모델 활용

2) 다양한 모델, 파라미터를 grid search로 찾음

- Linear, Decision Tree, Random Forest, GradientBoosting, XGBoost regressor를 모두 활용하고 grid search로 최적의 모델과 파라미터를 찾음
- R2는 0.25, RMSE 52 : 월급의 단위가 만원, 예측의 오차는 52만원 정도

주요 역할

1) 데이터 전처리 및 EDA
   - 데이터 결측치 제거, 데이터 EDA 및 시각화
   - 대상 한정 (조사회차 및 응답자 첫 직장 입직년도 기준으로 한정)
2) 예측 모델 개발
   - 변수 생성 (경력변수, 대졸여부 변수)
   - 모델 파이프라인 생성 및 하이퍼파라미터 튜닝

어려움

1) 패널데이터이다 보니 오래된 데이터가 많았음
2) 경력을 별도로 작성하지 않아 총 경력과 경력 단절에 대해 확인하기 어려웠음

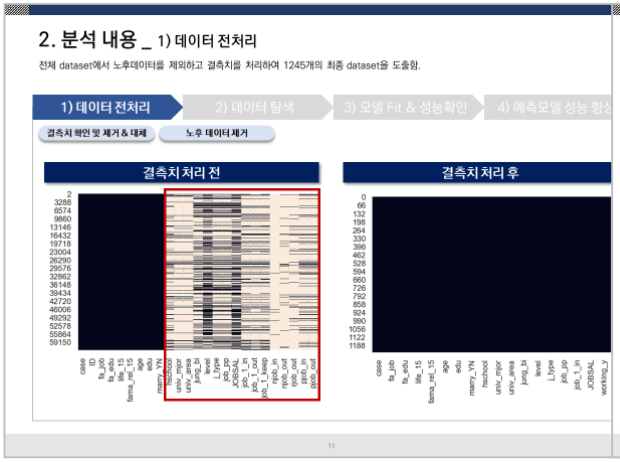
해결방안

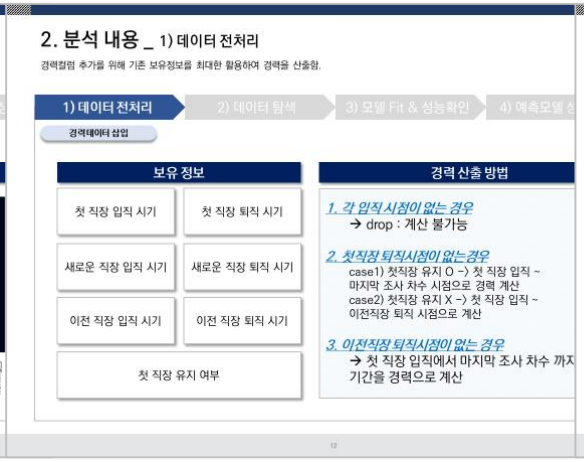
1) 2012년 이전 데이터는 분석에서 제외함
2) 경력데이터를 계산해서 삽입함 (첫 직장의 입직, 퇴직, 다음직장의 입직 퇴직 시기 등 고려)

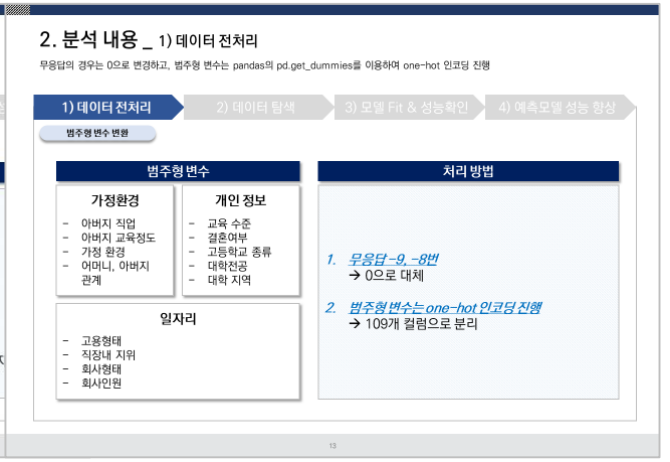
※ 깃헙 주소: https://github.com/p5653ceb/Regression_3_project

[machine learning] 여성 인력 임금 예측 _ 데이터 전처리

데이터 전처리

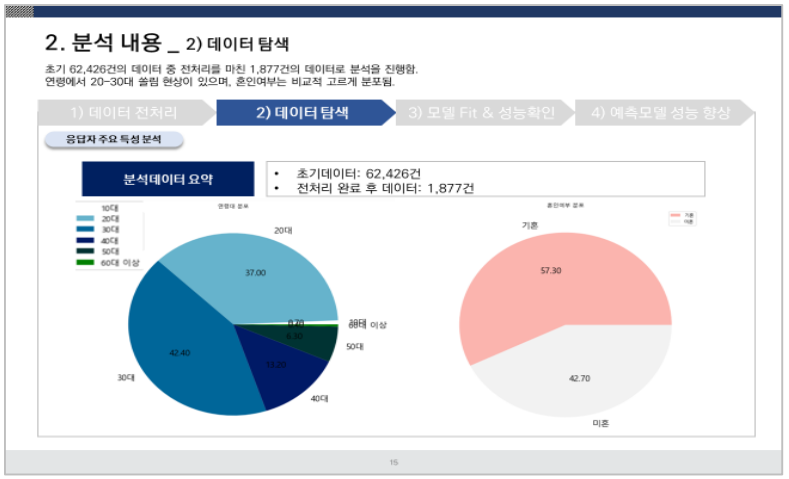


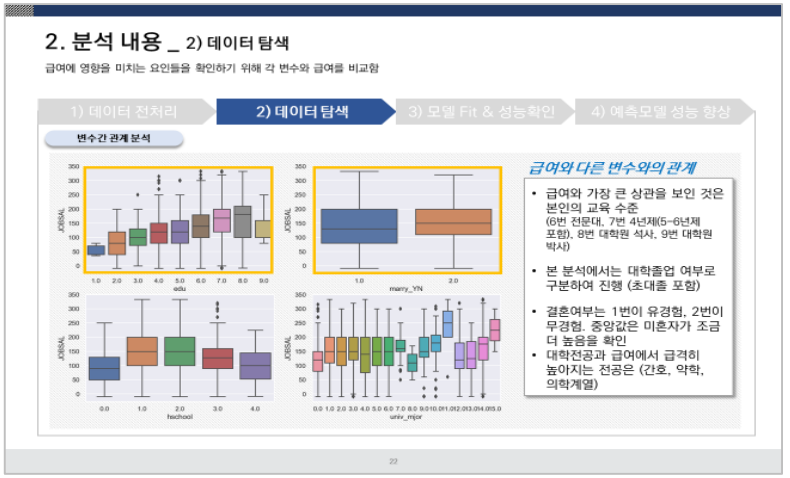




[machine learning] 여성 인력 임금 예측 _ 데이터 탐색

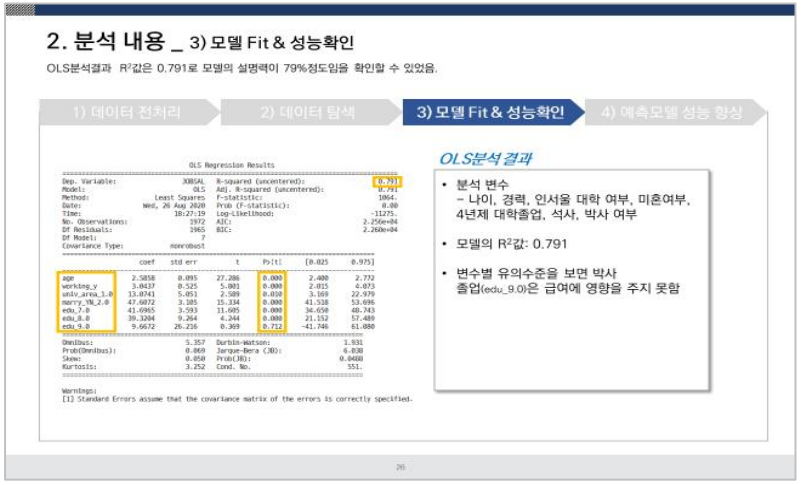
데이터 탐색

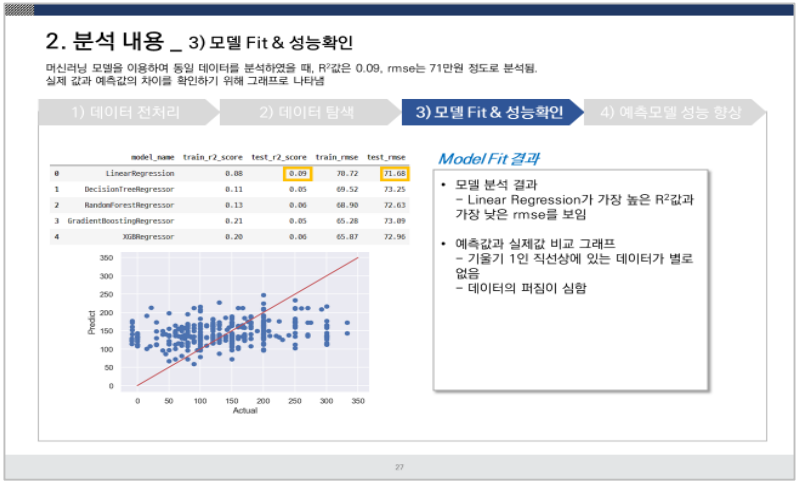




[machine learning] 여성 인력 임금 예측 _ 모델 fit & 성능확인

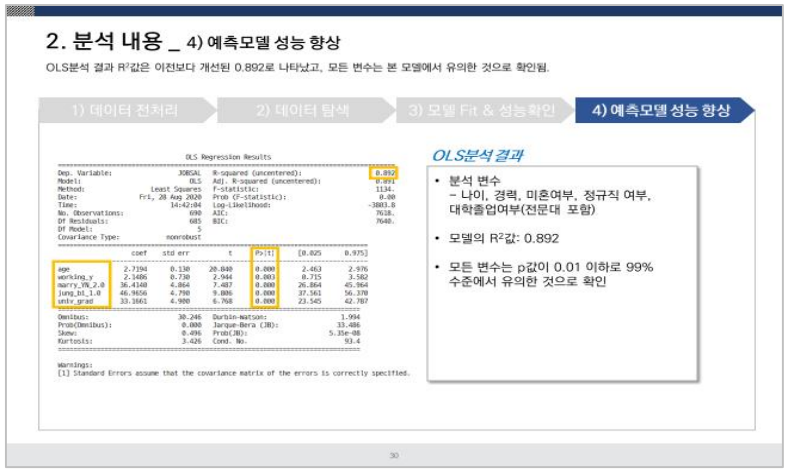
모델 fit & 성능확인

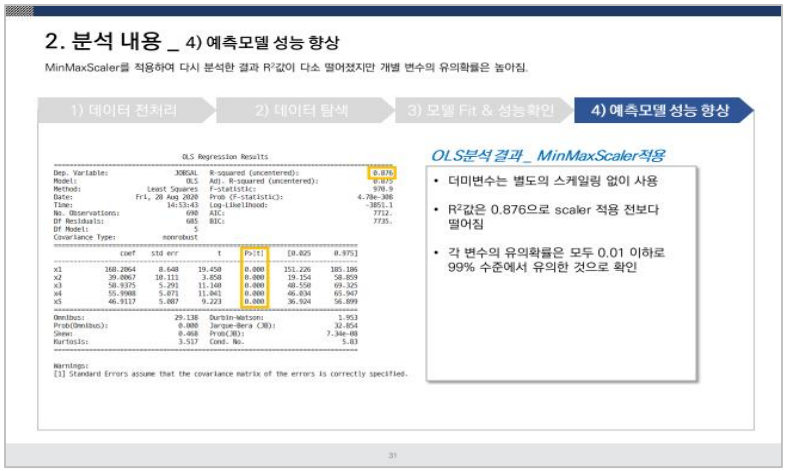


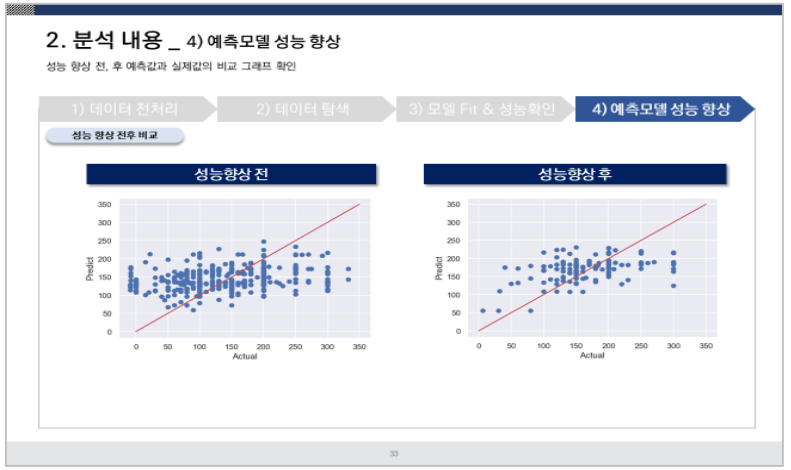


[machine learning] 여성 인력 임금 예측 _ 예측모델 성능 향상

예측모델 성능 향상







[Machine-Learning]

4. 홈쇼핑 매출 예측 & 최적 편성 제안

[machine learning] 홈쇼핑 매출 예측 & 최적 편성 제안

프로젝트 상세

프로젝트 내용

- 2019년 1년 상품 매출 데이터를 기반으로 2020년 6월의 매출을 예측하고 최적의 편성 제안

활용 솔루션

- 예측: LGBM과 catboost 모델을 활용하여 최종 예측 진행
  - 모델 선정 기준: 높은 정확도와 짧은 학습시간/ 데이터의 비선형성을 효과적으로 반영할 필요
  - 앙상블 계열 알고리즘 중 많은 양의 데이터에 적합하고 속도가 빠른 LGBM과 범주형 변수 처리 속도가 빠른 catboost 사용
- 편성 제안: 통계적 예측기를 활용한 예상 매출 예측 및 헝가리안 알고리즘으로 최적 배치
  - 프로젝트의 예측 데이터를 기반으로 최대값을 산출하는데 활용
  - 1주일을 168시간, 방송이 없는 시간을 제외하여 147시간으로 보고 1주일 편성 제안
  - 카테고리별 평균 취급액과 시간대별 평균 취급액의 가중치를 조정하여 실제값과 예측값의 RMSE를 최소로 만드는 가중치의 조합 도출
  - 147개 시간대 별로 최적 가중치를 적용하여 매출을 예측
  - 매출이 최대가 되는 최적 편성을 위해 헝가리안 알고리즘을 활용

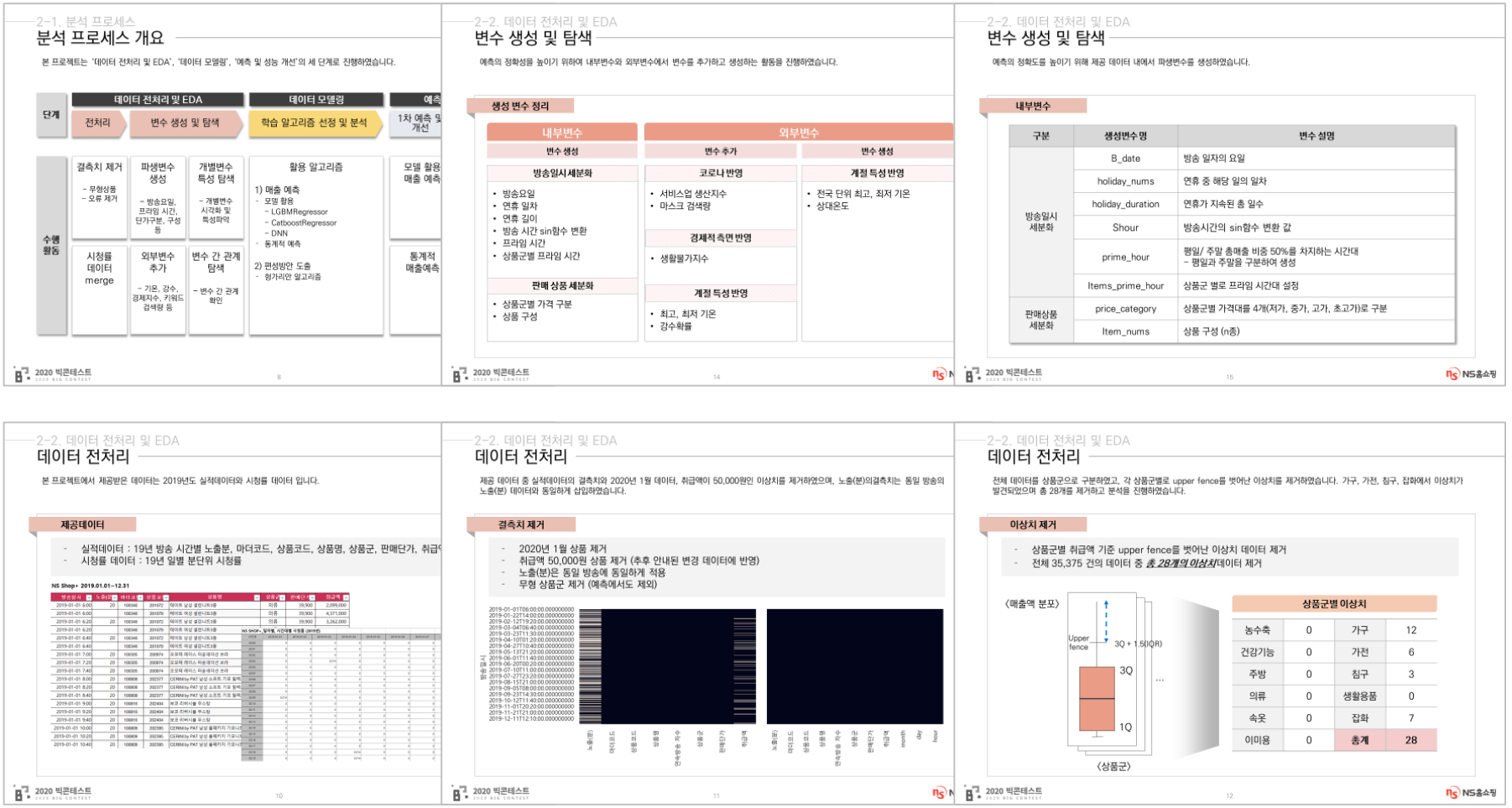
주요 역할

1) 데이터 전처리 및 EDA/ 예측모델 개발
   - 결측치 제거/ 경향성 EDA
   - 변수 생성 : 카테고리 관련 파생변수 생성
   - 외부변수 탐색 : 마스크 검색량/ 기온/ 상대기온 추가/ 타방송 시청률 크롤링 및 추가

2) 편성 모델 개발
   - 카테고리별 시간대별 평균 매출액을 활용하여 예상매출 matrix 생성 고안
   - 시간대/ 카테고리별 차별화된 가중치 적용
   - 헝가리안 알고리즘을 활용한 최적 편성 제안

[machine learning] 홈쇼핑 매출 예측 & 최적 편성 제안

프로젝트 상세



2-1. 분석 프로세스

분석 프로세스 개요

본 프로젝트는 '데이터 전처리 및 EDA', '데이터 모델링', '예측 및 성능 개선'의 세 단계로 진행하였습니다.

단계	데이터 전처리 및 EDA		데이터 모델링	예측
	전처리	변수 생성 및 탐색	학습 알고리즘 선정 및 분석	1차 예측 및 개선

수행 활동

- 결측치 제거 – 무형상품 – 오류 제거
- 파생변수 생성 – 방송요일, 프라임 시간, 단가구분, 구성 등
- 개별변수 특성 탐색 – 개별변수 시각화 및 특성파악
- 활용 알고리즘
  1) 매출 예측
  - 모델 활용
    - LGBMRegressor
    - CatboostRegressor
    - DNN
  - 통계적 예측
  2) 편성방안 도출
  - 정가리안 알고리즘
- 모델 활용 매출 예측
- 시청률 데이터 merge
- 외부변수 추가 – 기온, 강수, 경제지수, 키워드 검색량 등
- 변수 간 관계 탐색 – 변수 간 관계 확인
- 통계적 매출예측

2020 빅콘테스트

2-2. 데이터 전처리 및 EDA

변수 생성 및 탐색

예측의 정확성을 높이기 위하여 내부변수와 외부변수에서 변수를 추가하고 생성하는 활동을 진행하였습니다.

생성 변수 정리

내부변수	외부변수	
변수 생성	변수 추가	변수 생성
방송임시세분화	**코로나 반영**	**계절 특성 반영**
• 방송요일 • 연휴 일차 • 연휴 길이 • 방송 시간 sin함수 변환 • 프라임 시간 • 상품군별 프라임 시간	• 서비스업 생산지수 • 마스크 검색량	• 전국 단위 최고, 최저 기온 • 상대온도
판매 상품 세분화	**경제적 측면 반영**	
• 상품군별 가격 구분 • 상품 구성	• 생활물가지수	
	계절 특성 반영	
	• 최고, 최저 기온 • 강수확률	

2020 빅콘테스트

2-2. 데이터 전처리 및 EDA

변수 생성 및 탐색

예측의 정확도를 높이기 위해 제공 데이터 내에서 파생변수를 생성하였습니다.

내부변수

구분	생성변수 명	변수 설명
방송임시 세분화	B_date	방송 일자의 요일
	holiday_nums	연휴 중 해당 일의 일차
	holiday_duration	연휴가 지속된 총 일수
	Shour	방송시간의 sin함수 변환 값
	prime_hour	평일/ 주말 총매출 비중 50%를 차지하는 시간대 – 평일과 주말을 구분하여 생성
	Items_prime_hour	상품군 별로 프라임 시간대 설정
판매상품 세분화	price_category	상품군별 가격대를 4개(저가, 중가, 고가, 초고가)로 구분
	Item_nums	상품 구성 (n종)

2020 빅콘테스트 NS홈쇼핑

2-2. 데이터 전처리 및 EDA

데이터 전처리

본 프로젝트에서 제공받은 데이터는 2019년도 실적데이터와 시청률 데이터 입니다.

제공데이터

- 실적데이터 : 19년 방송 시간별 노출분, 마더코드, 상품코드, 상품명, 상품군, 판매단가, 취급액
- 시청률 데이터 : 19년 일별 분단위 시청률

NS Shop+ 2019.01.01~12.31

2020 빅콘테스트

2-2. 데이터 전처리 및 EDA

데이터 전처리

제공 데이터 중 실적데이터의 결측치와 2020년 1월 데이터, 취급액이 50,000원인 이상치를 제거하였으며, 노출(분)의결측치는 동일 방송의 노출(분) 데이터와 동일하게 삽입하였습니다.

결측치 제거

- 2020년 1월 상품 제거
- 취급액 50,000원 상품 제거 (추후 안내된 변경 데이터에 반영)
- 노출(분)은 동일 방송에 동일하게 적용
- 무형 상품군 제거 (예측에서도 제외)

2020 빅콘테스트

2-2. 데이터 전처리 및 EDA

데이터 전처리

전체 데이터를 상품군으로 구분하였고, 각 상품군별로 upper fence를 벗어난 이상치를 제거하였습니다. 가구, 가전, 침구, 잡화에서 이상치가 발견되었으며 총 28개를 제거하고 분석을 진행하였습니다.

이상치 제거

- 상품군별 취급액 기준 upper fence를 벗어난 이상치 데이터 제거
- 전체 35,375 건의 데이터 중 ***총 28개의 이상치***데이터 제거

〈매출액 분포〉

Upper fence

3Q + 1.5(IQR)

3Q

1Q

〈상품군〉

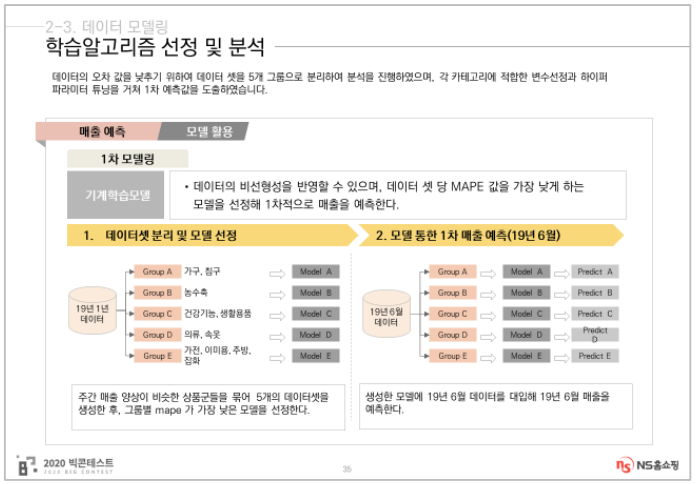
상품군별 이상치

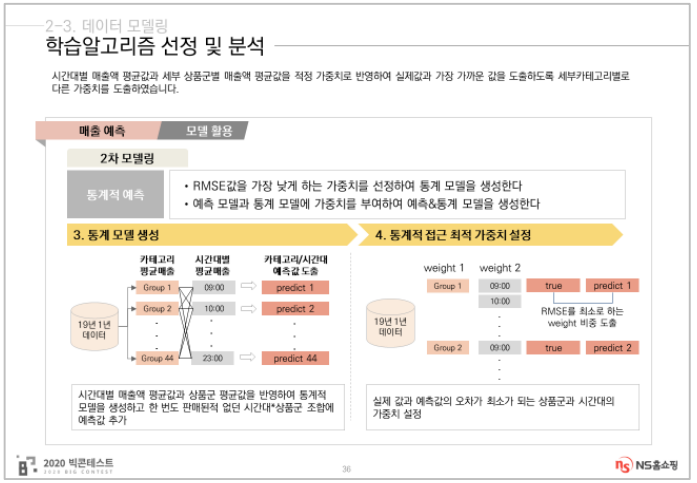
농수축	0	가구	12
건강기능	0	가전	6
주방	0	침구	3
의류	0	생활용품	0
속옷	0	잡화	7
이미용	0	총계	28

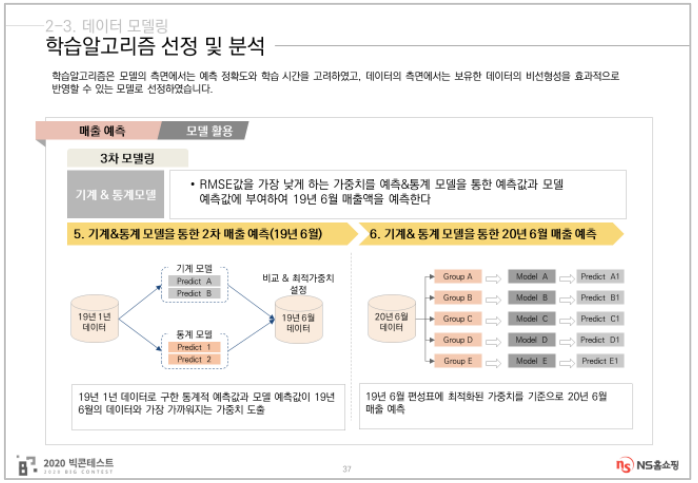
2020 빅콘테스트 NS홈쇼핑

[machine learning] 홈쇼핑 매출 예측 & 최적 편성 제안

매출예측 모델링







[machine learning] 홈쇼핑 매출 예측 & 최적 편성 제안

편성 제안

2-3. 데이터 모델링

학습알고리즘 선정 및 분석

통계적 예측의 정확도를 높이기 위해 카테고리별 평균 취급액과 시간대별 평균취급액의 가중치를 카테고리별로 다르게 적용하였습니다.

매출 예측 | 통계적 예측

실제 값과 예측값의 RMSE를 최저로 만드는 가중치 조합 도출

$$Sales = w_1 * SalesPerCaterogy + w_2 * SalesPerHour$$

* SalesPerCategory : 카테고리별 평균 취급액 , * SalesPerHour: 시간대별 평균 취급액 (147개 시간, 21h*7일)

구분	category1	...	44개 카테고리	sales/ hour
06:00				
07:00			w2	
...				
147개 시간대 (21h * 7일)	w1			
sales/category				

판매기록이 있는 Category와 시간대 조합의 **실제값**

오차 최소화

0부터 1까지 0.1 단위로 w1, w2를 조정해서 얻은 **예측값**

2020 빅콘테스트 | 38 | NS홈쇼핑

2-3. 데이터 모델링

학습알고리즘 선정 및 분석

예측한 매출을 바탕으로 1일 최대 매출 달성을 위한 방송 시간대별 최적의 카테고리를 헝가리안 알고리즘을 활용하여 할당하였습니다.

편성 방안 도출 | 편성 알고리즘

활용 알고리즘	헝가리안 알고리즘
알고리즘 특징	할당문제 최적화에 활용하는 알고리즘 - 문제 해결을 위한 최소 비용을 구하는 알고리즘 - 본 프로젝트에서는 예측 데이터를 기반으로 최대값을 산출
편성단위	1주일 주간 편성 (147개 시간대별 추천 상품군 제시)
도출 방법	• 주간 편성과 일간 편성 동시에 진행 • 주간, 일간 편성에서 시간대별 중복 상품 우선 배치 • 중복되지 않는 시간대의 경우, 높은 예상매출액의 상품 우선 배치

2020 빅콘테스트 | 39 | NS홈쇼핑

02. 프로젝트 세부내용

최종 예측 결과

모델을 활용한 예측에서는 앙상블계열 모델과 딥러닝 모델을 최종비교 하였고, 앙상블계열의 CatBoost Regressor가 가장 좋은 성능을 보였습니다. 2020년 6월의 예측 정확도와 오차는 구할 수 없어 2019년 6월의 예측치로 정확도와 오차를 확인하였습니다.

최종 결과 | 그룹별 모델 예측

모델 예측 성능 비교

- 앙상블 모델과 딥러닝 모델을 비교함
- 다섯 가지 모델 중 앙상블 모델인 catboost와 lgbm 성능이 가장 좋았음

기계 모델 예측 성능

model_name	R-sqaured	MAPE
XGBRegressor	0.61	68.80
LGBMRegressor	0.68	57.88
CatBoostRegressor	0.69	53.52
DNN	0.49	91

모델 자체의 성능으로 확인하였을 때, CatBoostRegressor의 R^2값이 0.69로 가장 높았으며, MAPE값 또한 53.52로 가장 낮게 분석됨

기계 모델 19년 6월 예측

model_name	R-sqaured	MAPE
XGBRegressor	-	-
LGBMRegressor	0.73	80.22
CatBoostRegressor		
DNN	-	-

2019년 6월 데이터의 예측치와 실제 값을 비교하였을 때 기계모델을 혼합하여 사용한 모델에서는 R^2가 0.73, MAPE가 80.22로 분석됨

2020 빅콘테스트 | 41 | NS홈쇼핑

02. 프로젝트 세부내용

최종 예측 결과

기계학습 모델만으로는 정확한 예측이 어렵다고 판단하여 통계적 예측기를 함께 사용하여 최종 예측을 진행하였습니다.

최종 결과 | 모델 예측 합계

모델 예측 성능 비교

- 통계적 접근으로 2019년 6월의 매출을 예측함
- 기계학습과 통계적 예측기를 혼합하여 2019년 6월의 매출을 예측한 결과

통계적 예측 결과

model_name	R-sqaured	MAPE
통계적 예측기	0.51	85.47

통계적 예측기로 2019년 6월의 예측값을 도출하였고, 실제값과 비교하였을 때, R^2값은 0.51, MAPE값은 85.47로 기계학습 모델보다는 다소 성능이 떨어짐

기계&통계 결과

model_name	R-sqaured	MAPE
기계 & 통계 예측기	0.25	52.89

기계학습과 통계적 예측기를 함께 사용한 모델의 결과는 R^2값이 0.25, MAPE값은 52.89로 도출되었음

2020 빅콘테스트 | 43 | NS홈쇼핑

End of the document