멋쟁이 사자처럼
데이터 분석 3기

Final Project

고객 결혼 여부 예측 모델링

5아시스

배승준
정지훈
신부철
최　진

Table of Contents

발표 순서

프로젝트 배경과 목표

배경

- 소매업체는 고객 데이터를 기반으로 맞춤형 마케팅을 제공하는 것이 경쟁력 강화의 중요한 요소
- 고객의 결혼 여부는 소비 패턴에 큰 영향을 미치는 요소로, 백화점은 고객의 결혼 여부를 데이터로 직접 수집하기 어려움
- 고객의 구매 이력을 분석하여 간접적으로 파악하고, 이를 기반으로 개인화된 마케팅 전략을 수립할 수 있음

목표

- 고객의 결혼 여부를 정확하게 예측하는 모델을 개발하여, 마케팅 부서 맞춤형 캠페인 수립에 기여
- 고객의 구매 데이터를 바탕으로 결혼 여부를 예측하여, 더 효과적인 타겟팅 가능

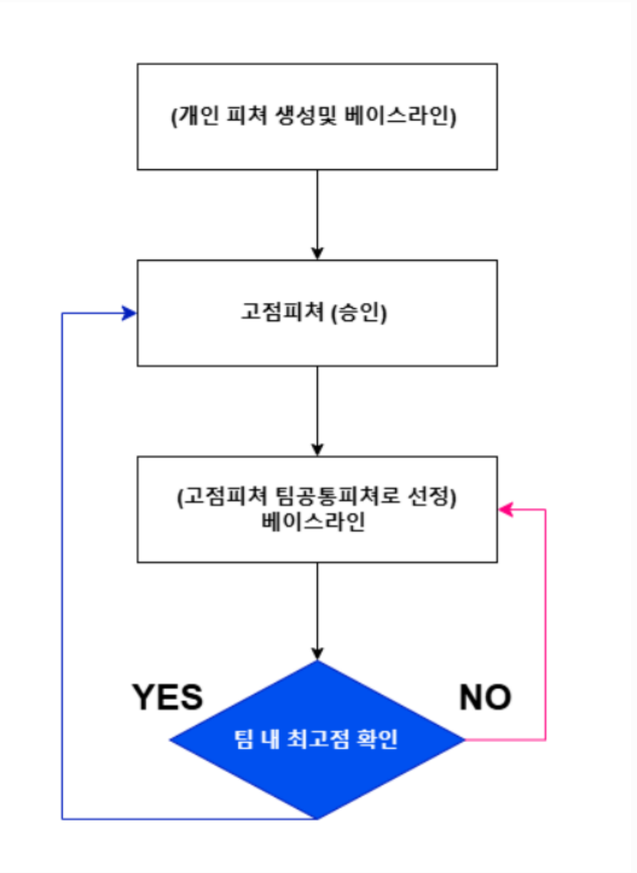
진행과정

- 매일 오전 10시, 오후 5시에 회의 진행

1. 공통피쳐(베이스라인 + 개인 피쳐) 제작
   a. 고점피쳐가 갱신되면, 팀 공통피쳐로 선정
   b. 갱신된 베이스라인 코드에 개인피쳐 추가
2. 공통피쳐 선정
3. 개인 모델링
4. 앙상블 테스트

Time Line

진행 logic



멋쟁이 사자처럼
데이터 분석 3기
Final Project

EDA | 공동피쳐 제작

EDA

store_train_transactions

- 구매시간이 같아도 여러 행으로 결과가 나옴
  - 데이터 하나는 한 번의 결제가 아닌, 하나의 물건 구매 이력

	ID	구매일시	지점코드	대분류	중분류	브랜드코드	구매가격
0	train_0	2004-07-29 17:13:00	A112000	패션잡화	싸롱화	5379	165000
1	train_0	2004-12-09 12:30:00	A144000	남성정장스포츠	골프웨어	5409	-205000
2	train_0	2004-09-01 17:53:00	A112000	가정용품	크리스탈	6164	234000
3	train_0	2004-12-09 12:30:00	A144000	남성정장스포츠	골프웨어	5452	968400
4	train_0	2004-12-09 12:30:00	A144000	남성정장스포츠	골프웨어	5452	-1076000

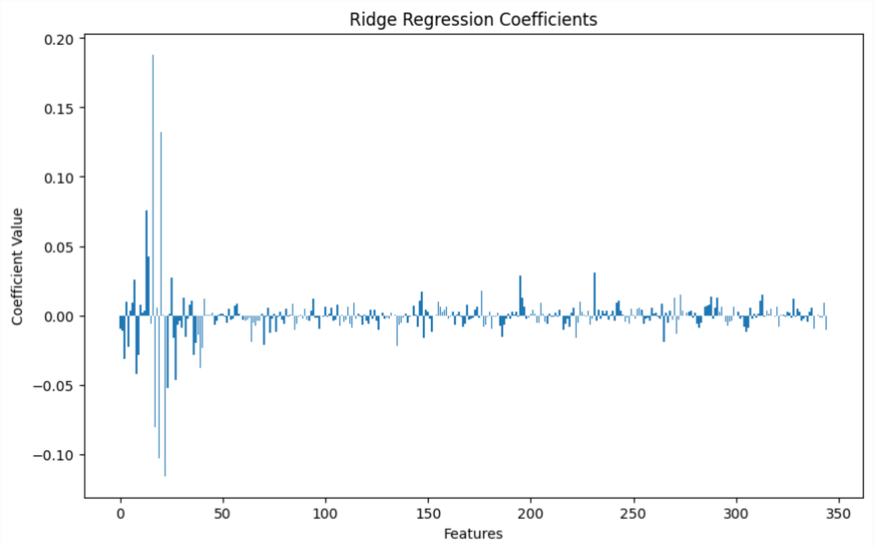
train_target

- 0.0 - 기혼자 9066명
- 1.0 - 미혼자 5874명

  -> 기혼자에 치우쳐진 target

target	count
0.0	9066
1.0	5874

- 독립변수와 종속변수 간의 상관 릿지 분석
  - 특성추출_베이스라인 코드 결과를 바탕으로 진행



공통피쳐 제작

- 고객 ID별
  - 지점코드, 대분류 카운트
  - 왜도, 첨도, 표준편차
  - 가장 많이 구매한 브랜드
- 구매일시
  - 일별 평균 구매 건수
  - 요일별 방문 비율
  - 주당 내점 횟수
  - 1,2,3,4분기 구매비율
  - 주구매요일
  - 구매 시간대에 따른 pivot
- 구매금액
  - 연도별 금액 합을 사분위수로 나눠 등급으로 분류
    → 구매 총합과 등급을 pivot
- 대분류
  - 주구매_대분류
  - 대분류, 지점코드에 따른 pivot
  - 생식품 평균 구입 가격
- 중분류
  - 주구매_중분류
- 지점
  - 지점별 구매 가격
- 브랜드코드
  - 최다구매_브랜드
  - 최대구매_브랜드코드의 중분류

개인 피쳐 생성

피쳐 생성_배승준

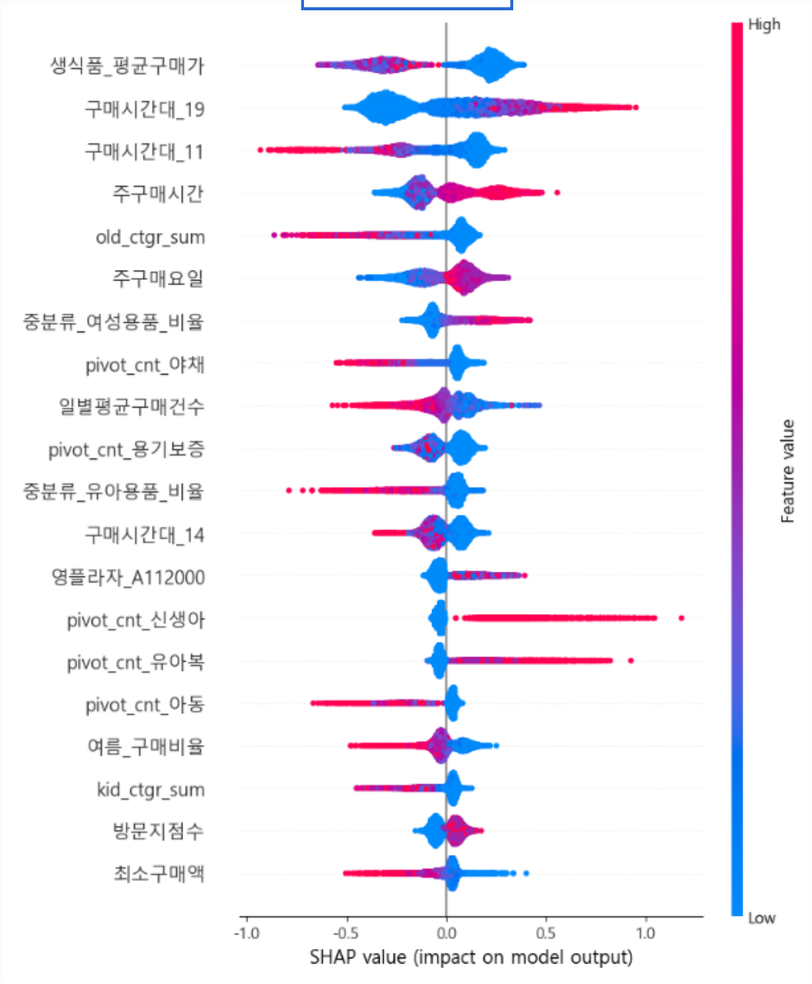
- 중분류에서 solo, fam, old, young 별로 카테고리 묶음 생성

  각 계층에 적합할 것으로 예상되는 요소를 추가하여 구매가격 sum

  XAI 결과를 바탕으로, 신생아와 유아복 / 야채 / 19시, 11시 칼럼 추가
  - 신생아와 유아복의 경우 소비 금액이 높은 경우 미혼
  - 야채의 소비 금액이 높으면 기혼일 것으로 예측
  - 19시, 11시 칼럼의 경우 피쳐 중요도가 높게 나왔지만 결과적으로 정확도는 하락

  브랜드 코드 활용하기
  MinMaxScaler 에 변환되는 브랜드코드를 활용하기 위해 여러 시도
  - 바이너리 인코딩
  - 브랜드 평균가를 타겟 인코딩
  - 원핫인코딩 이후 PCA 차원축소

XAI



피쳐 생성_신부철

- 결혼 예측 칼럼 생성

  릿지 회귀 분석 결과에서, 종속변수에 영향을 끼치는 피쳐를 대상으로 가중치 부여

  높은 가중치(0.8~1.0)
  - 구매주기, 총구매액, 구매건수, 최대구매액, 최소구매액
    → 결혼 후에는 생활비 증가, 빈도도 높을 것으로 예상
  - 방문지점수, 주구매지점
  - 주말방문비율, 주중방문비율

  중간 가중치(0.5~0.7)
  - 구매시간대 - 특정 시간대에 구매활동이 집중되면 생활패턴 관측 가능, 결혼 여부에 따라 구매시간대에 차이가 생길 수 있음
  - 브랜드코드•중분류•대분류_nunique
    → 기혼 가정의 경우 구매 브랜드가 다양할 것
  - 거래개월수, 계절별 구매비율 : 구매 기간이 지속적인지 확인

가중치 함수(infer_marriage_status(row)피쳐는 0.006~0.010 가량의 변동치
시간과, 횟수 제한으로 인해 적용하는 데 한계점이 있었음

피쳐 생성_신부철

- 최적의 피처 찾기
LightGBM 모델을 사용해 데이터를 학습하고, **특징 중요도 (Feature Importance)**를 기반으로 최적의 특징(feature) 개수를 선택하는 과정을 포함하고 있음.

최적 피쳐 :
- best_nu_features : 최적의 피처 개수.
- top_teatures: 최적의 피처 목록

최종 데이터셋 :
- 선택된 피처를 기중으로 최종 학습 및 데이터 생성

결과 :
- best_score : 최적의 피처로 얻은 최고의 F1 Macro Score.
- best_num_feature: 최적의 피처 개수.
- train_ft_selected, test_ft selected.shape: 최종 선택된 학습 및 테스트 데이터의 크기
- 결과 900 피처 ⟹ 227개 피쳐
- CV 점수하락

- top_features = feature_importances.head(best_num_features).index
- train_ft_selected = train_ft[top_features]
- test_ft_selected = test_ft[top_features]
- train_ft_selected.shape, test_ft_selected.shape

개선 가능 사항

1. 고급 탐색 기법 사용:
   - 순차적 피처 선택(Sequential Feature Selection) 알고리즘 활용
2. 모델 변경 적용:
   - Light 외에 모델 적용하여 중요도 피반 피쳐 셀렉션 시도

피쳐 생성_최진

1. XAI로 피쳐 특성을 살펴 봤을 때, 영향이 컸던 피쳐들끼리 묶기 위해 [구매 일시]관련 피쳐에서는 두 달씩 묶어 분석을 진행
   → 월별로 나눈 것보다 금액의 차이가 극명하게 보임

```
('12_1월_구매비율', lambda x: np.mean(x.dt.month.isin([1,12]))),
('2_3월_구매비율', lambda x: np.mean(x.dt.month.isin([2,3]))),
('4_5월_구매비율', lambda x: np.mean(x.dt.month.isin([4,5]))),
('6_7월_구매비율', lambda x: np.mean(x.dt.month.isin([6,7]))),
('8_9월_구매비율', lambda x: np.mean(x.dt.month.isin([8,9]))),
('10_11월_구매비율', lambda x: np.mean(x.dt.month.isin([10,11]))),
```

2. 주당 방문한 횟수를 확인하여 기혼 가정의 경우 1인가구보다 방문비율이 높을 것으로 예상

```
('주당내점횟수', lambda x: x.count() / x.dt.to_period('W').nunique()),
```

3. 구매시간대에 19시는 미혼이 많고, 11시와 14시에 기혼이 많은 점을 참고하여 월화수 / 목금토일로 나눠 낮과 밤의 방문비율을 체크

```
('월화수_밤_방문비율', lambda x: np.mean( (x.dt.weekday<3) & (x.dt.hour>=18) )),
('목금토일_밤_방문비율', lambda x: np.mean( (x.dt.weekday>=3) & (x.dt.hour>=18) )
('월화수_낮_방문비율', lambda x: np.mean( (x.dt.weekday<3) & (x.dt.hour<18) )),
('목금토일_낮_방문비율', lambda x: np.mean( (x.dt.weekday>=3) & (x.dt.hour<18) ))
```

4. 고객별로 최다 구매한 브랜드를 뽑고, 해당 브랜드의 중분류를 추가

```
brandMaxTrain = train_copy.groupby('ID')['브랜드코드'].agg(lambda x: x.mode()[0]).reset_index()
brand_categories = train_copy[['브랜드코드', '중분류']].drop_duplicates()
```

- 아쉬운 점
  - 명절(2004년 추석, 2005년 설) 전 2주 간 / 전후 10일의 구매금액, 구매빈도를 뽑아 보았으나 cv점수는 오히려 떨어짐
  - 상관계수가 0.01 미만의 피쳐들을 삭제하여 모델링을 돌려 f1 score는 상승하였으나, public 점수는 하락
    → 0.05로 기준치를 낮췄다면 점수가 올라갔을지 궁금

멋쟁이 사자처럼
데이터 분석 3기
Final Project

모델링

스케일링

RobustScaler

이상치를 제거하는 스케일러
평균과 분산 대신에
중간값과 사분위값 사용

-> 극단값에 영향을 받지 않음

MinMaxScaler

데이터 값을 0과 1사이로
변환하는 스케일러

이상치 민감도가 높음

MaxAbsScaler

데이터를 -1과 1 사이로
변환하는 스케일러

희소행렬이나 음수 값을 포함한
데이터에서 유용하게 사용

이상치로 인한 성능저하를
극복하기 위해
1단계로 사용

RobustScaler를
먼저 적용하여
이상치 영향을 낮춘 후
스케일링 적용

팀원들 중,
MinMaxScaler보다 성능
향상이 되는 경우
선택하여 사용

모델링

- 공통적인 모델링

XGBoost
- 모든 팀원들에게 가장 좋은 지표를 보여준 모델
- 성능이 가장 잘 나와서 데이터셋에 적합하다고 판단

CatBoost
- 일부 팀원들에게 좋은 지표가 나온 모델
- 좋은 결과가 나온 경우 앙상블에 사용함

LGBM
- 가장 가볍고 빠른 모델
- CV 점수는 가장 좋게 나오지만 과적합 의심
  → 앙상블의 일부 모델로만 사용

- 일부 팀원들이 사용한 모델링

Logistic Regression
- Stacking 앙상블에서 final_estimator 로 사용했을 때 가장 좋은 성능 발휘

KNeighbors, DecisionTree Classifier
- Voting 앙상블에서 사용한 모델

하이퍼 파라미터 튜닝

Grid Search CV

- 설정한 모든 경우의 수를 모두 학습
- 한 팀원의 경우 CatBoost에 적용하였지만, 학습시간이 20시간 넘게 걸려 사용하지 않음

Randomized Search

- 학습시간을 유동적으로 조절할 수 있다는 장점
- 더 광범위한 범위를 비교하기 위해 채택

AutoML

- 복잡한 머신러닝 파이프라인을 자동화
- 모델을 일일이 구축하고 평가하는 시간 절약
- 여러 모델들을 사용할 수 있다는 장점

Optuna

- Optuna의 최적화 알고리즘을 적용하여 시도
- Catboost 모델을 활용하였지만 성능이 낮아서 최종결과 코드에는 사용하지 않음

앙상블

앙상블

Stacking
- 대부분의 팀원들이 사용한 기법
- 가장 성능이 좋게 나와서 사용

Voting
- 일부에게는 성능이 좋지 않아서 사용하지 않음.

앙상블(np.mean)
- AutoML과 여러 모델들의 예측결과를 직접적으로 앙상블함

그외
- Stacking 기법의 결과물과 특정 모델의 예측결과를 직접 앙상블한 경우도 있음

앙상블

최종 앙상블의 target을 0.4로 매핑한 이유

```
# 교차 검증을 통해 예측 확률을 얻기 (predict_proba 사용)
probs = cross_val_predict(stack_model, train_ft, target,
cv=StratifiedKFold(5, shuffle=True, random_state=42),
method='predict_proba', n_jobs=-1)

# 임계값 0.4로 예측을 0과 1로 변환
threshold = 0.5
y_pred = (probs[:, 1] >= threshold).astype(int)  # 1번 클래스의 확률을
기준으로 예측

# 정확도 (accuracy)와 재현율 (recall) 계산
accuracy = accuracy_score(target, y_pred)
recall = recall_score(target, y_pred)
```

0.5
Accuracy: 0.7411
Recall: **0.6021**

0.4
Accuracy: 0.7385
Recall: 0.7034

Recall 값이 유의미하게 상승
하지만 Private score는 0.45가 최고점

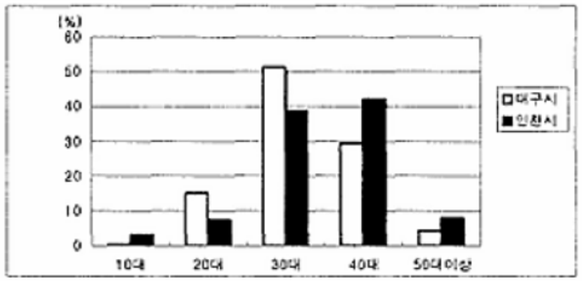
아쉬운 점

- Catboost에 optuna로 최적화 알고리즘을 적용, K-Fold로 5번 나눠 학습시킨 예측 모델
  - public 점수가 0.714로 낮아 앙상블에서 배제, private는 0.731점으로 꽤 높은 점수라 최종 앙상블에 반영하지 않은 점

- 2003년 4월 발간된 서울대학교 국토문제연구소 논문집 [대형할인점의 입지유형과 소비자 구매 행태]에 따르면, 고객 연령층은 3,40대가 가장 많았으며, 고객 직업으로는 전업주부가 가장 많음
  - 해당 결과를 반영하는 피쳐를 만들지 못한 아쉬움

- 피쳐 생성에 있어 target값의 0과 1을 혼동하여 미혼(1)을 나타내는 피쳐를 더 추가하지 못하고 기혼 중심의 피쳐로 치우져진 점

- optuna와 RandomSearchCV 등 하이퍼파라미터 최적화 모델을 사용하는 것도 좋은 공부가 되었지만, 직접 각각의 파라미터 값을 조정하며 결과값을 비교해보지 못한 점



<그림 VI-1> 대형할인점 고객의 연령분포

<표 VI-3> 자본계열별 대형할인점 고객의 직업

단위 : 명, (%)

직업	중앙자본	지방자본	외국자본	계
전업주부	77 (35.5)	61 (49.5)	113 (35.2)	241 (37.6)
사무직	55 (25.4)	24 (23.3)	84 (26.2)	163 (25.4)
판매,서비스직	17 (7.8)	4 (3.9)	42 (13.1)	63 (9.8)
전문직	17 (7.8)	4 (3.9)	35 (10.9)	56 (8.7)
기타주)	51 (23.5)	20 (19.5)	47 (14.6)	118 (25.7)
계	217 (100.0)	103 (100.0)	321 (100.0)	641 (100.0)

주 : $\chi^2 = 26.8$, (유의수준 : 0.001)
주) 운수업, 자영업, 학생 등이 이에 포함됨.
자료: 설문조사에 의함.

결론

2004, 2005년 고객 결혼 여부 예측 데이터는 20년전의 데이터이다 보니 지금과는 다른 배경 지식들이 많았고 이로 인해 저희가 세운 가설들이 예상과는 많이 달라 배경지식이 중요함을 다시 한번 느끼는 시간이 되었습니다.

기혼자들은 아이가 있다는 가정으로 신생아와 유아복 구매가격이 많으면 오히려 미혼인 경향이 나오는 상황과 20년전에는 지금과 달리 마트가 많이 없었던 것도 알게 되었습니다.

특정 모델에서 더 높은 정확도가 나오는 데이터들을 알게 되었고 cv점수가 낮아도 앙상블시 더 좋은 결과가 나오기도 하고 많은 것들을 알아갈 수 있어서 좋았던 프로젝트입니다.

저희가 가공한 데이터로는 XGBoost가 시간을 투자할 수록 가장 잘 나오는 모델이였고 LGBM같은 경우는 점수는 잘 나왔지만 실제 결과에서는 잘 나오지 못하는 결과를 확인하였습니다.

멋쟁이 사자처럼
데이터 분석 3기
Final Project

들어주셔서 감사합니다.

THANK YOU

5아시스