


구매패턴 분석을 통한 결혼여부 예측


파이널 프로젝트

조 칠전팔기

조원 김이지, 이희원, 길선종



CONTENTS



01 배경 설명

02 EDA 및 시각화

03 특성공학

04 모델링 및 앙상블

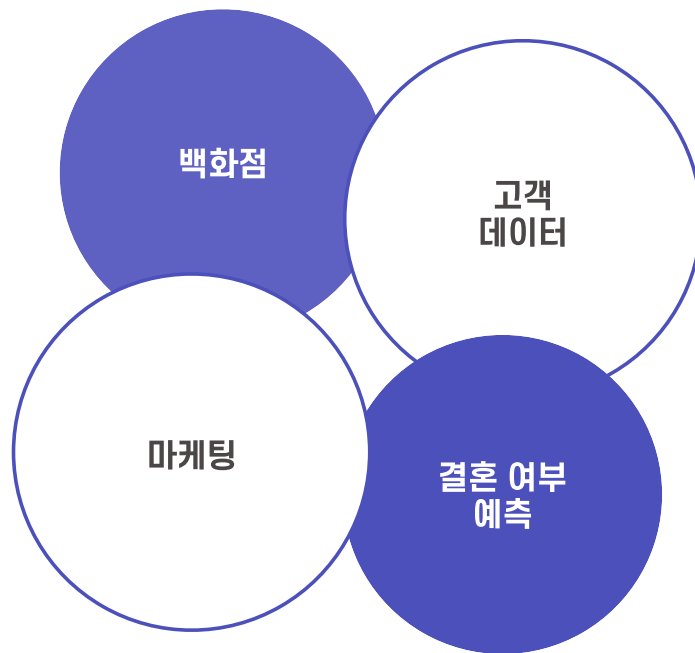
05 결론

06 도전 및 해결

01

배경 설명

프로젝트 진행 배경



- **고객 데이터**를 활용한 **맞춤형 마케팅 전략** 수립을 경쟁력을 강화하는 핵심 요소
- **백화점**을 대상으로 했을 때, **결혼 여부**는 고객이 소비패턴에 큰 영향을 줌.
- 하지만 결혼 여부는 개인정보 보호 문제 등에 의해 수집이 어려우므로 **간접적 예측이 필요**
- **데이터 기반의 접근법**을 통해, 결혼여부를 예측하여 **고객 경험을 향상**시키고 **백화점의 장기적 경쟁력을 강화**하는데 기여하고자 함.

PROFILE

백화점 고객

ID

구매일시

지점코드

대분류

중분류

브랜드코드

구매가격

— 02

EDA 및 시각화

feature importance

LGBM

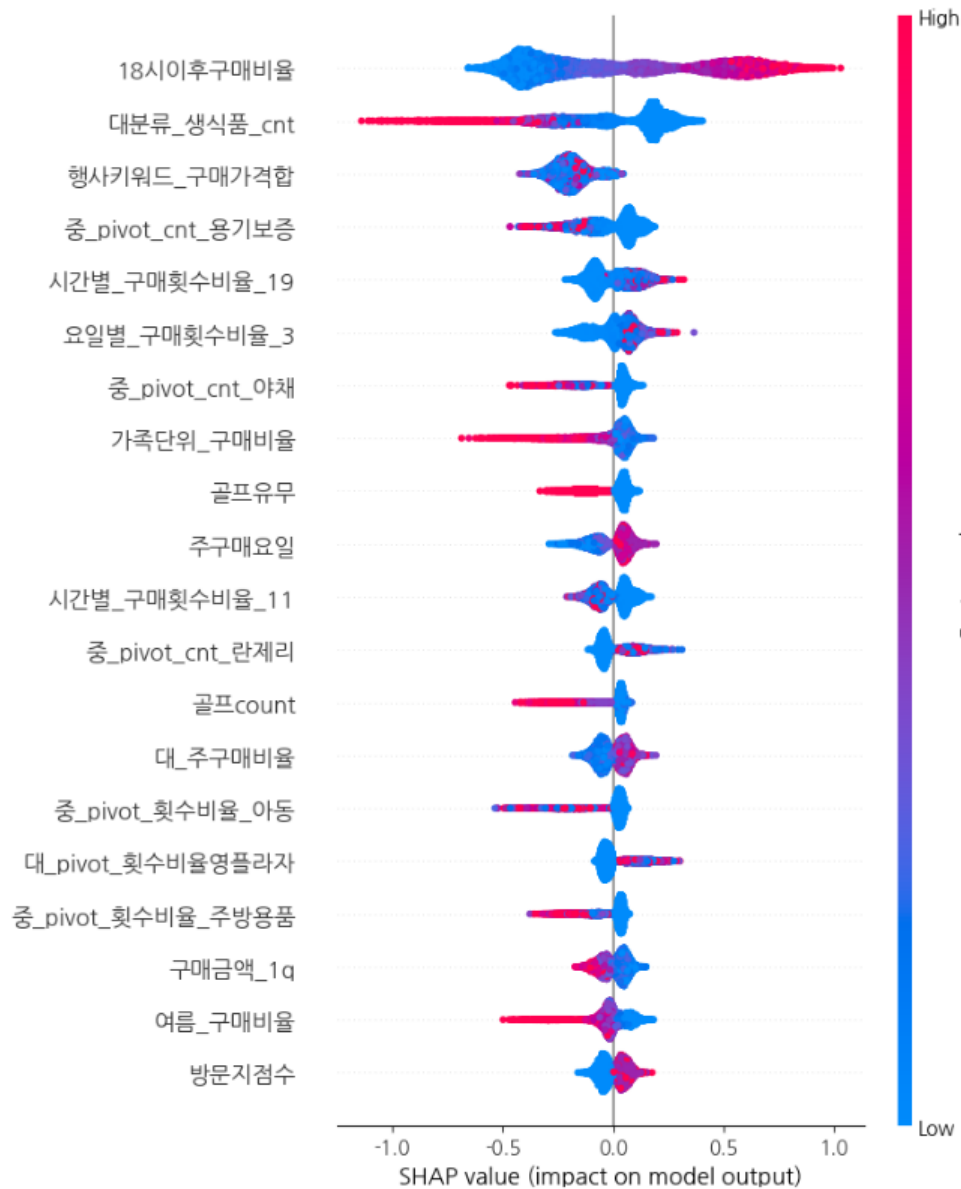
EDA 및 시각화 07

	Feature	Importance
430	가족단위구매비율	64
4	여름_구매비율	54
14	방문평균거래평균횟수	52
10	18시이후구매비율	49
411	요일별_구매횟수비율_3	46
32	대분류_생식품_cnt	40
2	구매주기	39
187	중_pivot_횟수비율_수입종합화장품	38
37	최소구매액	37
233	중_pivot_횟수비율_용기보증	36
9	12시이후_18시이전구매비율	35
209	중_pivot_횟수비율_아동	35
43	구매금액_1q	35
410	요일별_구매횟수비율_2	33
426	시간별_구매횟수비율_19	32
28	대_주구매비율	32
18	하루 구매 시간 간격	31
40	구매금액표준편차	31
36	최대구매액	29
125	중_pivot_횟수비율_란제리	28
157	중_pivot_횟수비율_상품군미지정	27
16	금오후토일방문비율	26



- 초반에 특성을 어느 정도 만든 이후 진행
- feature importance를 통해 어느 특성이 **영향**을 많이 끼쳤는지 확인
- 선택하여 사용할 특성을 **판단**하는 데에 사용
- **예시** : 가족단위구매비율, 여름_구매비율, 방문평균거래평균횟수, 18시 이후 구매비율 등

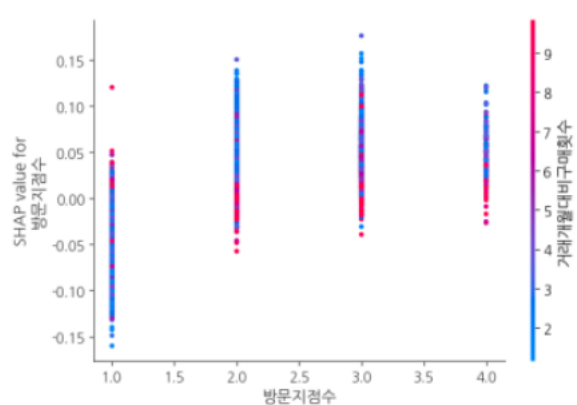
feature importance shap



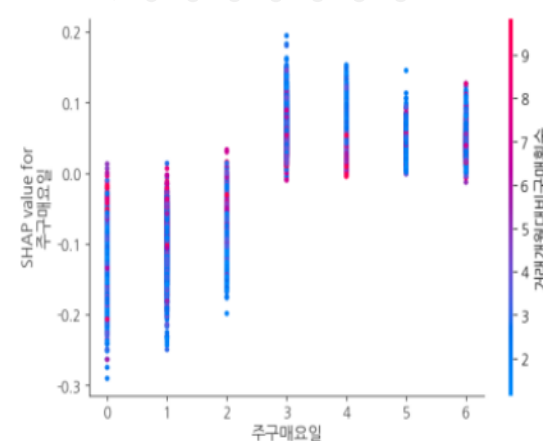
- XAI를 진행하기 전 1차로 판단
- 기혼과 미혼을 **비교**하여 특성 확인
- **예시** : 아동은 기혼에 더 영향을 많이 끼쳤지만 유아는 미혼에 더 영향을 많이 끼쳤다.
- 이후 각 특성에 대한 XAI로 2차 확인

XAI 활용

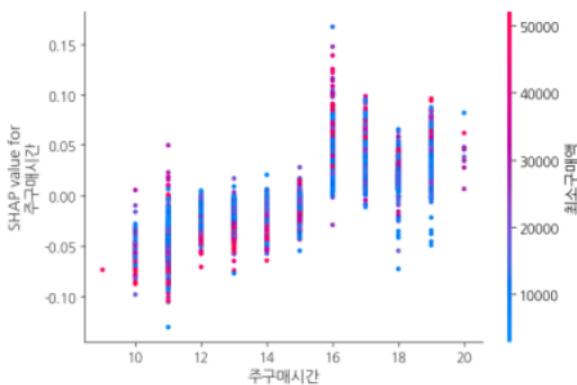
방문지점수



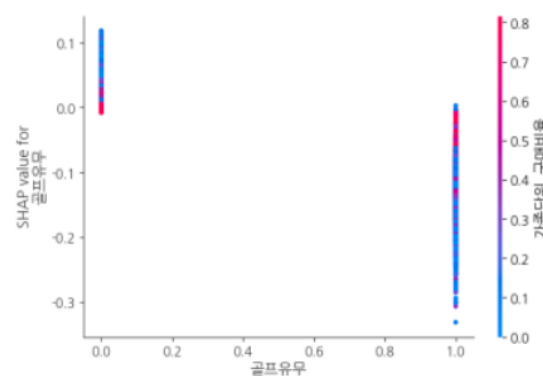
사건참여액



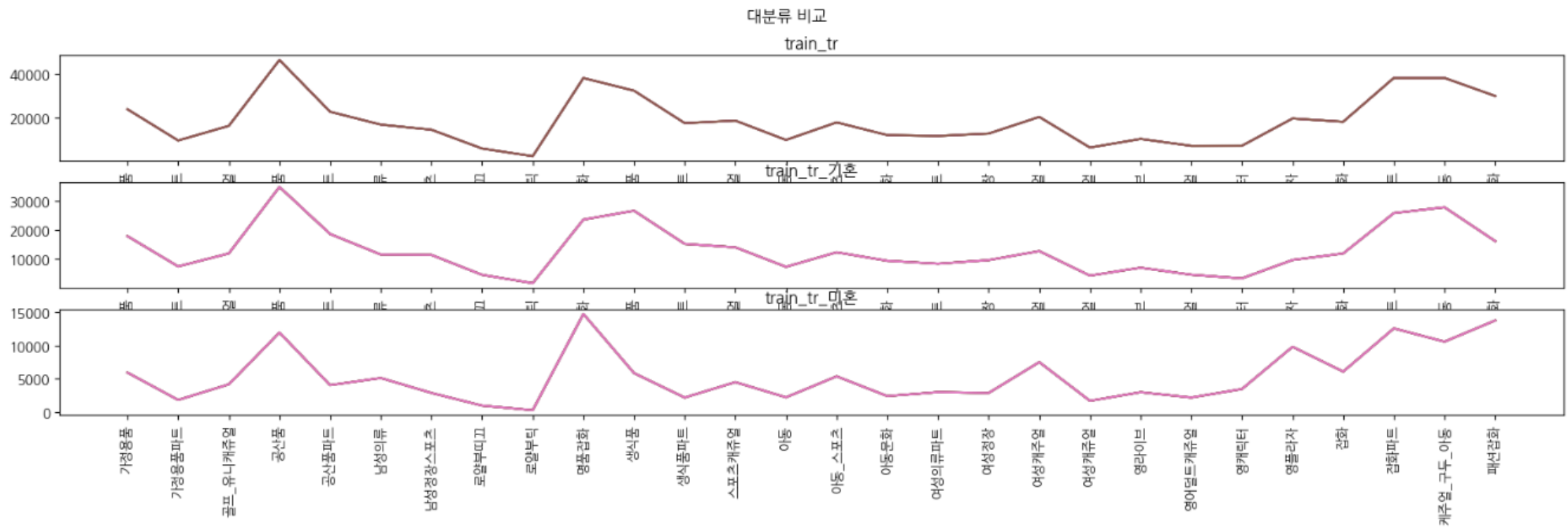
주구매시간



데뷔비야마

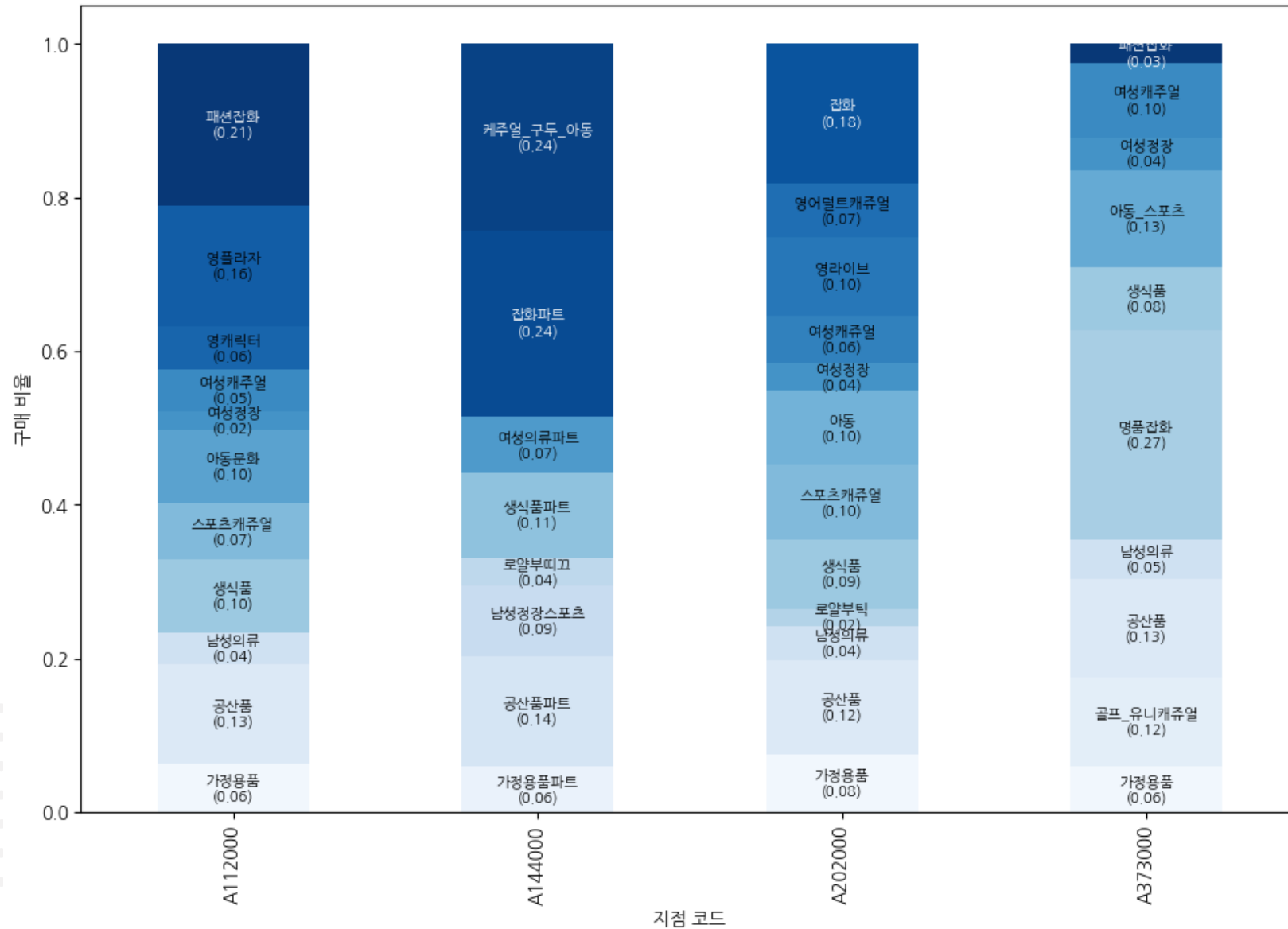


대분류별 기온미혼



지점별 대분류

지점별 대분류 구매 비율 (정렬된 비율)





03

특성 공학

train_tr 특성 추가

- train_tr['구매일시'] 형식 변환

object 형식으로 작성된 구매일시 컬럼을 pandas를 이용해 날짜 형식으로 변환

- 구매월, 구매일, 구매요일 추출

(예시) '2004-05-03 01:15:00'

날짜로 변경된 구매일시 컬럼에서 월, 일, 요일 추출

- 구매시간 추출

(예시) '2004-05-03 01:15:00'

날짜로 변경된 구매일시 컬럼에서 시간 추출

train_ft 특성 추가 (1)



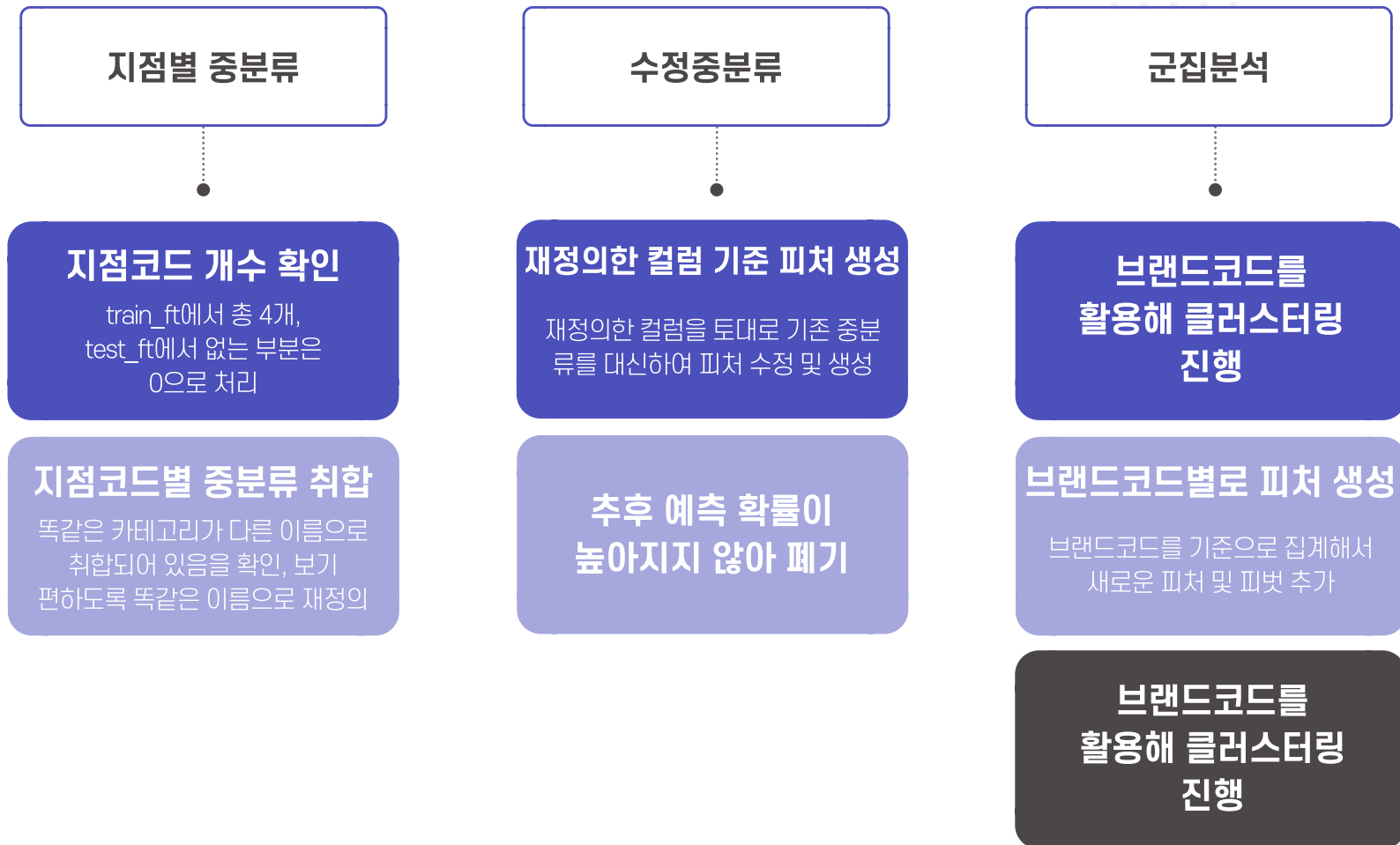
train_ft 특성 추가 (2)



train_ft 특성 추가 (3)

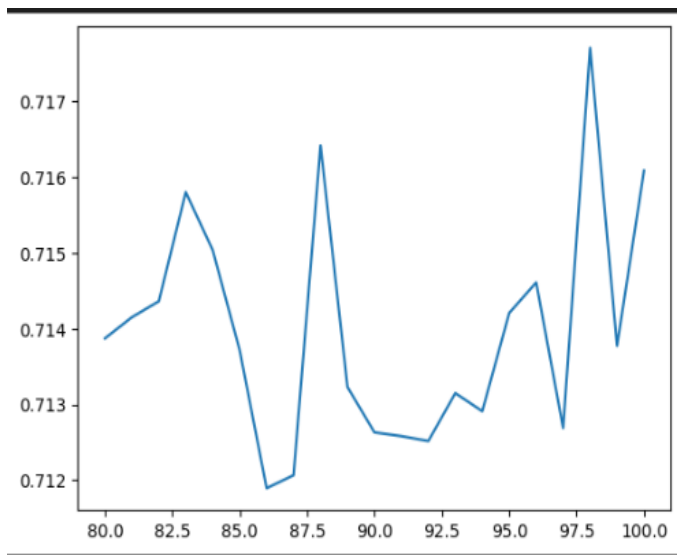


train_ft 특성 추가 (4)



특성선택

feature selection



sklearn.feature_selection의 SelectPercentile을 사용해서
점수가 가장 높은 피쳐들을 선택해서 사용

feature importance 0 제거

```
feature_importance_df[feature_importance_df["importance"]==0]
```

	Feature	Importance
488	최대구매액_대분류_20	0
431	체류시간	0
448	최소구매액_대분류_8	0
487	최대구매액_대분류_19	0
446	최소구매액_대분류_6	0
...
274	중_pivot_횡수비율_종합_수입	0
278	중_pivot_횡수비율_즉석조리	0
279	중_pivot_횡수비율_지갑_벨트	0
280	중_pivot_횡수비율_직수입침구	0
496	최대구매액_대분류_28	0

176 rows x 2 columns

피쳐 중요도가 0인 176개의 피쳐 제거 후
모델 점수가 많이 차이 나지 않으면 피쳐 제거

— 04

모델링 및 앙상블

교차검증

01

kfold 이용

cv 변수를 KFold(n_splits=5,
shuffle=True,
random_state=SEED)로 지정

02

cv.split 이용

for문을 활용해서 교차 검증과 동시에
5개로 분리해서 모델 학습

04

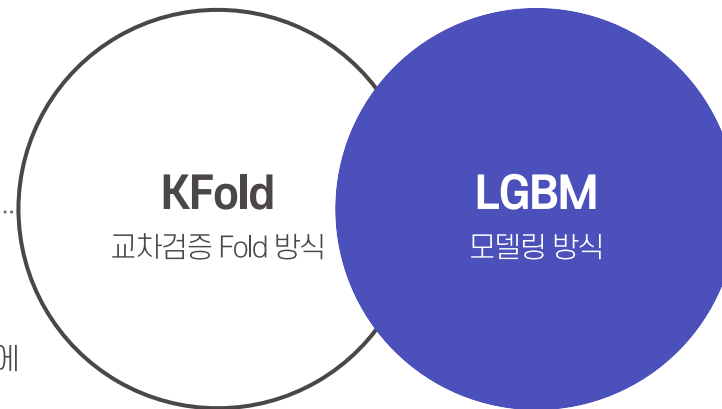
예측 진행

for문을 활용해서 리스트에 담은
모델의 test_ft 예측 확률을 계산해서
새로운 리스트에 저장
리스트에 저장된 확률의 **평균**을 계산한
후 1에 대한 예측 확률을 0과 1로 변경

03

5개 모델 리스트화

학습시킨 모델을 새로운 리스트에
저장



AutoML

catboost

boosting 앙상블 기법을 사용하는 모델
오버피팅을 피하기 위해 내부적으로 여러 방
법을 갖추고 있어 속도 뿐만 아니라
예측력도 굉장히 높다.

histgb

그래디언트 부스팅 트리를 기반으로함
효율적인 계산
상대적으로 낮은 과적합 경향
범주형 변수 처리

loss 점수 기준 lgbm, xgboost, catboost, histgb

기본세팅

```
metric : macro_f1
estimator_list = ['lgbm', 'rf', 'xgboost', 'extra_tree',
                  'xgb_limitdepth', 'lrl1', 'catboost', 'histgb']
task : classification
time_budget : 60*60
seed : 42
early_stop : True
```

XGBoost

```
class XG_Objective:
    def __init__(self, x, y, seed):
        self.x = x
        self.y = y
        self.seed = seed
        self.cv = KFold(n_splits=5, shuffle=True, random_state=self.seed)

    def __call__(self, trial):
        hp = {
            'n_estimators': trial.suggest_int('n_estimators', 50, 201),
            'learning_rate': trial.suggest_float('learning_rate', 0.01, 0.31,
step=0.01),
            'max_depth': trial.suggest_int('max_depth', 3, 12),
            'gamma': trial.suggest_float('gamma', 0.1, 5.1, step=0.1),
            'subsample': trial.suggest_float('subsample', 0.5, 1.0, step=0.1),
            'colsample_bytree': trial.suggest_float('colsample_bytree', 0.5,
1.0, step=0.1)
        }
        model = XGBClassifier(random_state=self.seed, **hp)
        scores = cross_val_score(model, self.x, self.y, cv=self.cv, scoring='f
1_macro', n_jobs=-1)
        return scores.mean()
```

```
xg_study = optuna.create_study(
    direction='maximize',
    sampler= sampler
)
xg_objective_func = XG_Objective(train_ft, target, SEED)
xg_study.optimize(xg_objective_func, n_trials=150)
```

LGBM

```
class LGB_Objective:
    def __init__(self, x, y, seed):
        self.x = x
        self.y = y
        self.seed = seed
        self.cv = KFold(n_splits=5, shuffle=True, random_state=self.seed)

    def __call__(self, trial):
        hp = {
            'n_estimators': trial.suggest_int('n_estimators', 50, 501),
            'learning_rate': trial.suggest_float('learning_rate', 0.01, 0.31, s
tep=0.01),
            'max_depth': trial.suggest_int('max_depth', 3, 16),
            'num_leaves': trial.suggest_int('num_leaves', 20, 101),
            'min_child_samples': trial.suggest_int('min_child_samples', 5, 31),
        }
        model = LGBMClassifier(random_state=self.seed, **hp)
        scores = cross_val_score(model, self.x, self.y, cv=self.cv, scoring='f
1_macro', n_jobs=-1)
        return scores.mean()
```

```
lgb_study = optuna.create_study(
    direction='maximize',
    sampler= sampler
)
lgb_objective_func = LGB_Objective(train_ft, target, SEED)
lgb_study.optimize(lgb_objective_func, n_trials=150)
```

앙상블

- predict threshold 조절

미혼 비율이 **train 데이터**에서 **0.393** 이나왔다.

threshold 값을 0.39 근처로 조정해서 0 과 1의 비율을 조정해서 recall 값을 올리고 precision을 낮췄다.

recall값이 얻는 이득이 precision이 손해보는 것보다 커서 f1-macro값이 커진다.

- Soft Voting 사용

각자 다른 모델링을 2~3개씩 뽑아서 총 7~9개의 모델을 **SoftVoting방식**을 사용해서 오버피팅감소효과로 인한 예측성능을 향상시키고 **일반화**했다.

개별성능이 어느정도 나오는 모델로 선정하기위해 **기준을 0.71이상**으로 정했다.

각 클래스의 예측확률을 평균내서 앙상블했다.



05

결론





결혼 여부

인사이트

- 소비목적과 패턴 차이
- 구매 주기 및 시간대
- 브랜드 및 카테고리 선호도

마케팅 전략

- 맞춤형 프로모션 및 혜택 제공
- 고객 분류에 따른 이메일/푸시 마케팅
- 지점별 상품 배치 최적화
- 충성도 프로그램 강화
- 라이프 이벤트 타겟팅

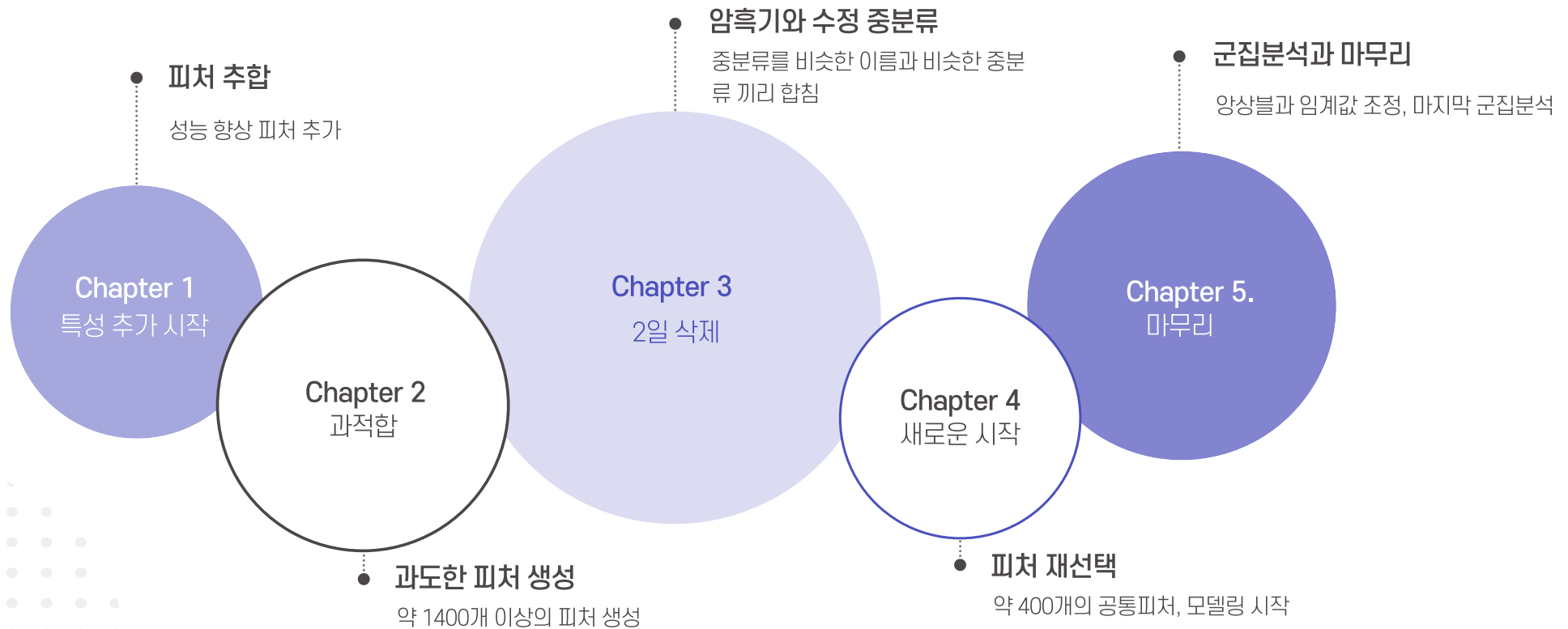


백화점

— 06

도전 및 해결

도전 타임라인



앞으로의 도전이 있다면..?

개선점

01. 빠른 판단

오래 걸리는 작업이 끝까지 해서 결과가 좋지 않게 나왔을 때 **환기**를 빨리하고 **다시 새롭게** 시작해야 된다는 것을 알았다.

03. 좋은 피처

피처의 개수가 무작정 많다고 좋은 모델이 아니라는 것을 배웠다. 피처선택션이 모든 것을 해결지 않고 pivot테이블과 시간데이터를 이용해서 무작정 피처를 늘렸는데 결과적으로는 의미 없는 피처가 많아지면서 피처 선택션도 의미가 많이 없어지고 시간만 많이 걸렸다. 최대한 **의미가 있는 피처만 생성**하고 무작정 피처를 늘리는 것은 좋지 않다는 것을 배웠다.

02. 기록

파일 정리를 제대로 해야 되겠다는 생각을 했다. 한 파일을 가지고 계속 쓰다보니까 어떤 식으로 모델링 했는지도 기억에 잘 남지 않고 결과가 어떤 것인지 구분이 잘 안됐다.

04. 다양한 접근

한 특성이 다른 특성과 관계가 있을 때는 함부로 정리를 하면 안된다는 것을 배웠다. 다른 요인이 관여될 경우 새롭게 정리한다고 해도 오히려 모델의 정확성을 떨어트린다는 것을 배웠다. **다양한 접근**을 통해 많은 도전을 해보고 **실험 정신**으로 결과를 확인해 봐야 한다는 것을 배웠다.



구매 패턴 분석을 통한 결혼 여부 예측

THANK
YOU

칠전팔기

김이지, 이희원, 길선종