

KHUDA 컨퍼런스

신용카드 사용자 연체 예측 및 마케팅 제안



2023.01.24

강현빈, 김효준, 신영서, 차민재

목차

- 01 / 프로젝트 목적
- 02 / 전처리 과정
- 03 / 예측 모델 수립
- 04 / 마케팅 방안
- 05 / 결론 및 제언

1.프로젝트 목적

개인
신상정보
데이터



- ☑ 신용카드 신청자의 향후 채무 불이행 예측
- ☑ 신용카드 신청자의 대금 연체 가능성 예측
- ☑ 예측한 신용도 기반으로 마케팅 고도화
- ☑ 개인 신상정보 기반으로 마케팅 고도화



신용카드 사용자들의 개인 신상정보 데이터를 기반으로
사용자의 신용카드 대금 연체 정도(신용도)를 예측 &
신용도와 신상정보 데이터에 기반한 마케팅 방안 제안

2. 전처리 과정

/ 데이터 전처리

1. 직업 유형 결측치 처리

- ☑ 일을 하지 않는 사람의 경우, 무직자 처리

[1] 직업 유형 == NaN
[2] 연금을 수령 받는 사람들

- ☑ 그 외의 결측치는 최빈값인 “Laborers”로 대체

2. 이상치 처리

- ☑ 자녀 수 7명 이상인 데이터 제거
- ☑ IQR 이용하여 이상치 처리

[1] 연간 소득
[2] 근무하고 있는 사용자 데이터 한정

3. 중복 값 & 불필요한 컬럼 제거

- ☑ 중복 값 제거
: index 제외 모든 특성이 동일한 경우,
같은 고객으로 처리
- ☑ 불필요한 컬럼 제거

[1] FLAG_MOBIL : 모두 1로 동일
[2] child_num : family_size와 상관관계 강함

4. 파생변수 생성

- ☑ numeric 변수
: 최대한 많은 변수를 조합 및 생성하고 이와
다중공선성을 보이는 컬럼 제거
- ☑ categorical 변수 : begin_month(신용카드 발급일) 컬럼과 child_num(자녀 수)를
제외한 모든 변수를 합친 ID 생성

/ 데이터 샘플링

credit (신용도)별로 데이터 불균형이 존재하여
imblearn 모듈을 이용하여 샘플링 진행

```
1 train['credit'].value_counts()
2      14383
1       5465
0       2771
```

- Over Sampling

- RandomOverSampler
- SMOTE

- Under Sampling

- RandomUnderSampler
- NearMiss

- Raw Data

3. 예측 모델 수립

/ 예측 모델 수립

- 데이터 결정

raw 데이터가 오버, 언더 샘플링보다 정확도가 높아 raw 데이터로 최종 결정

- 모델 결정

모델	Accuracy	하이퍼파라미터 튜닝 후 Accuracy
XGBoost	0.68	0.70
LGBM	0.69	0.70
CatBoost	0.73	0.72 (log loss 0.66)
Random Forest	0.69	0.70

/ 예측 모델 수립

● 모델 고도화

전처리 과정 변경

- 다양한 파생변수 생성,
실험적으로 적용
- Scaling 방법

샘플링 기법 변경

- RandomOverSampling
- SMOTE

하이퍼파라미터 튜닝

- Grid Search
- optuna

```
# logloss 가장 낮게 나온 모델
from sklearn.model_selection import StratifiedKFold
from sklearn.cluster import KMeans
from catboost import CatBoostClassifier, Pool

skfold = StratifiedKFold(n_splits=n_fold, shuffle=True, random_state=seed)
folds=[]
for train_idx, valid_idx in skfold.split(X, y):
    folds.append((train_idx, valid_idx))

cat_pred = np.zeros((X.shape[0], n_class))
cat_pred_test = np.zeros((X_test.shape[0], n_class))
cat_cols = ['income_type', 'edu_type', 'family_type', 'house_type', 'occyp_type', 'ID']

for fold in range(n_fold):
    print(f'##### Fold {fold} #####')
    train_idx, valid_idx = folds[fold]
    X_train, X_valid, y_train, y_valid = X.iloc[train_idx], X.iloc[valid_idx], y[train_idx], y[valid_idx]
    train_data = Pool(data=X_train, label=y_train, cat_features=cat_cols)
    valid_data = Pool(data=X_valid, label=y_valid, cat_features=cat_cols)

    model_cat = CatBoostClassifier()
    model_cat.fit(train_data, eval_set=valid_data, use_best_model=True, early_stopping_rounds=100, verbose=100)

    cat_pred[valid_idx] = model_cat.predict_proba(X_valid)
    cat_pred_test += model_cat.predict_proba(X_test) / n_fold
    print(f'CV Log Loss Score: {log_loss(y_valid, cat_pred[valid_idx]):.6f}')

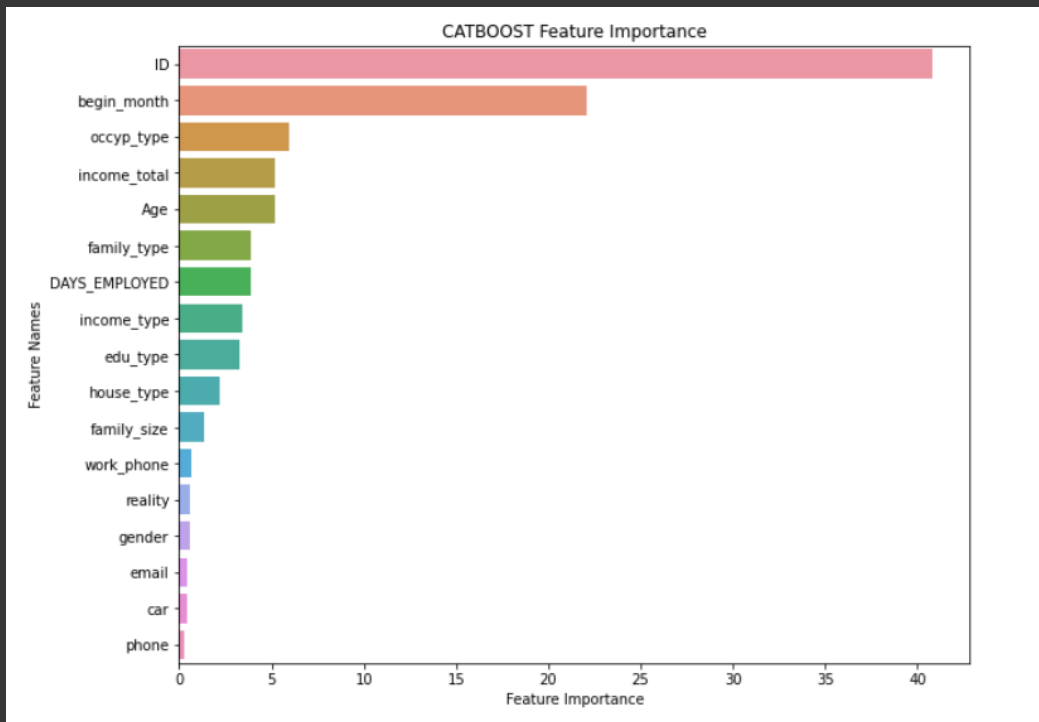
print(f'##tLog Loss: {log_loss(y, cat_pred):.6f}')
```

- **결과**

ID 생성이 중요하게 작용
최종 log loss 결과는 0.66

```
bestTest = 0.6697810599  
bestIteration = 352
```

```
Shrink model to first 353 iterations.  
CV Log Loss Score: 0.669781  
Log Loss: 0.665006
```



● 제출 결과

전체 랭킹 >					
● WINNER ● 1% ● 4% ● 10%					
#	팀	팀 멤버	점수	제출수	등록일
18	KHUDA	KH	0.66684	1	2시간 전
1	다냐나라	다냐나라	0.66408	37	일 년 전
2	chopin	ch	0.66414	8	일 년 전
3	새싹치지마	어빙	0.66448	12	일 년 전
4	비회원		0.66509	23	일 년 전
5	인생사는갓김치	nl	0.66513	13	8시간 전
6	닉네임11	HA	0.66598	22	일 년 전

4. 마케팅 방안

/ 신용 점수를 활용한 마케팅

- 프로젝트 목적

신용카드 사용자 데이터를 이용하여 신용 점수 예측하는 것

카드사는 이 신용 점수를 활용해 신청자의 채무 불이행, 신용카드 대금 연체 가능성 예측 가능

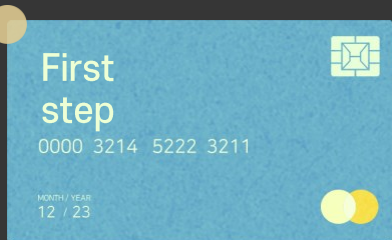


고신용자들을 대상으로
새로운 마케팅 전략 수립

- 마케팅 방안

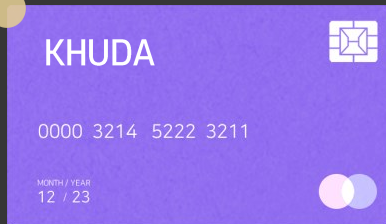
신용 점수가 0, 1인 신청자 대상

성실하게 신용거래를 하도록 유도하여 신용 등급을 유지 또는 높일 수 있도록 사용자 맞춤형 카드 개발



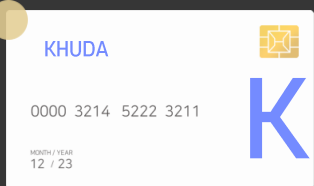
사회초년생 카드

- 신용등급이 높지 않은 신파일러
- 주 사용층 : 34세 이하의 근무기간 3개월 이상 직장인
- 특징 : 편의점, 대중교통, 점심식사 할인



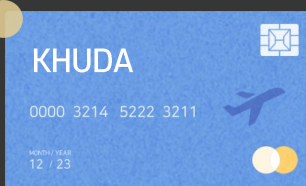
1인 가구 카드

- 주 사용층 : 미혼 1인 가구
- 특징 : 서적, 플랫폼 구독, 배달어플 캐시백, 가정 내 홈 CCTV



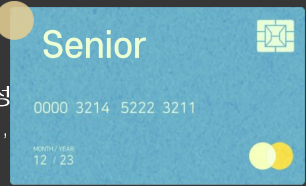
교육 카드

- 주 사용층 : 기혼의 자녀가 있는 30, 40대 여성
- 특징 : 서적, 학원비, 인터넷 강의 할인



세계로 카드

- 주 사용층 : 30~50대 남성
- 특징 : 항공 마일리지 적립, 해외 숙박 결제 할인



시니어 카드

- 주 사용층 : 60세 이상
- 특징 : 병원비, 약국 할인

고객의 성별, 연령대, 직업유형 등을 고려하여
특징을 나눠 카드상품 기획

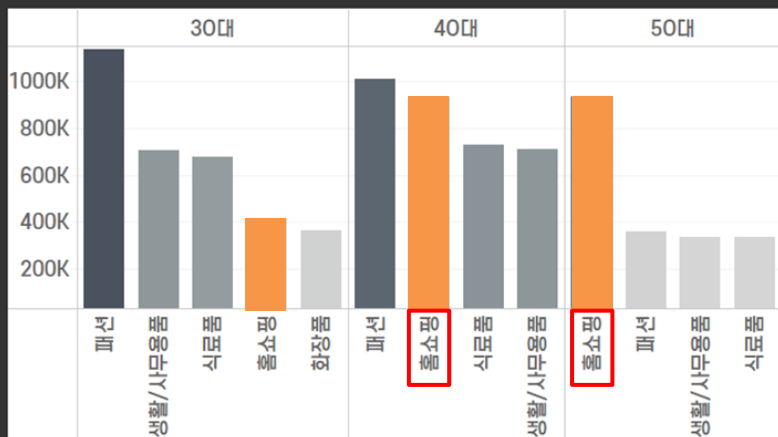
참고: 신한카드의 코드나인 시리즈



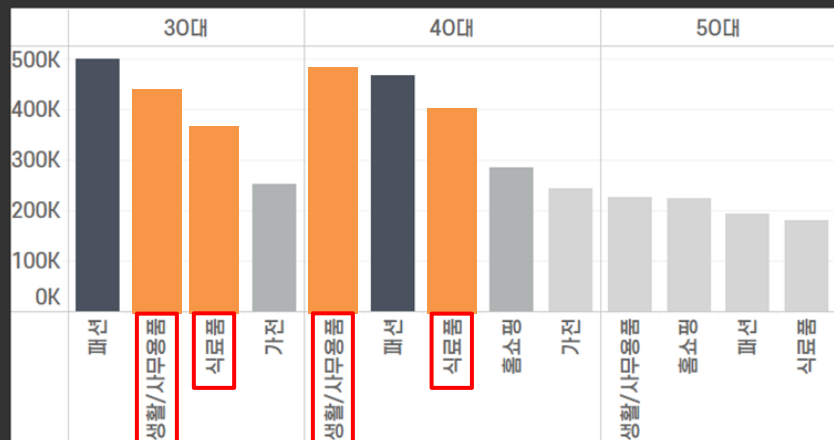
고객 개인의 필요를 고려한 마케팅

/ 고객 세분화를 통한 카드 상품 추천

40-50대 여성을 타겟팅 했을 때
홈쇼핑 수요가 증가



30-40대 남성이 관심 많은
생활/사무용품, 식료품



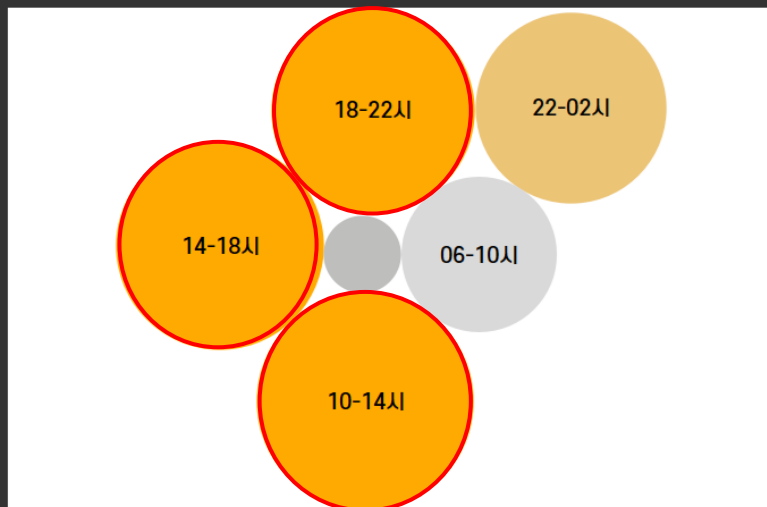
데이터 출처: <금융데이터 거래소> 삼성카드 '온라인쇼핑 요일/시간대별 이용 특징'

/ 고객 세분화를 통한 카드 상품 추천

휴일보단 평일에!

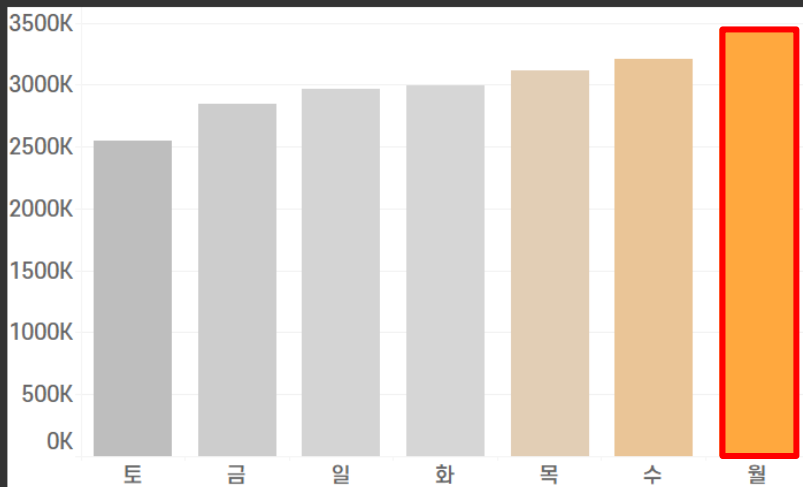
평일 패션	평일 홈쇼핑	휴일 패션	휴일 홈쇼핑
평일 생활/사무용품	평일 식료품	휴일 생활/사무용품	휴일 식료품

점심시간과 취침 전 증가하는 구매율

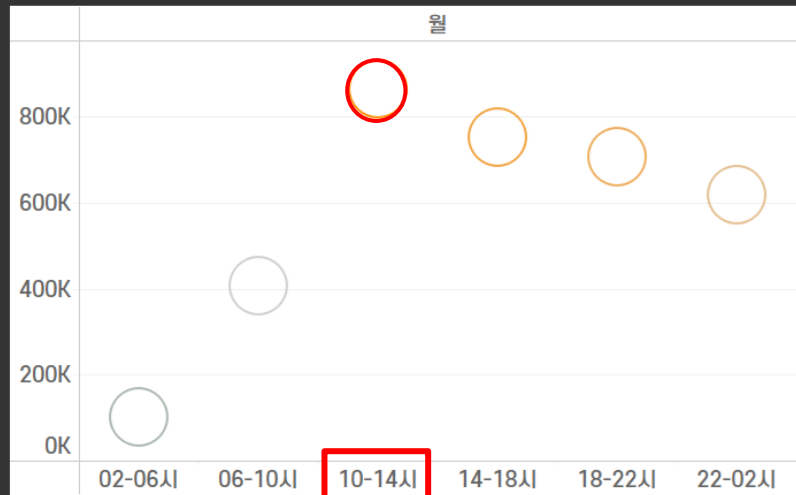


/ 고객 세분화를 통한 카드 상품 추천








쇼핑을 가장 많이하는 요일은 '월요일'



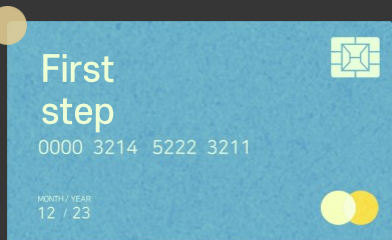
월요일 10-14시에 구매율 up!



5. 결론 및 제언

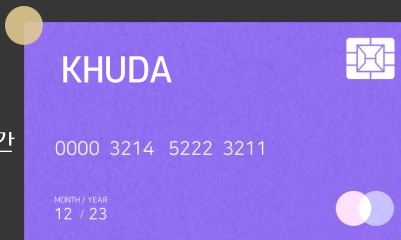
<div> ● WINNER ● 1% ● 4% ● 10% </div> <div>전체 랭킹 ></div>					
#	팀	팀 멤버	점수	제출수	등록일
18	KHUDA		0.66684	1	2시간 전
1	다나나라		0.66408	37	일 년 전
2	chopin		0.66414	8	일 년 전
3	새싹치지마		0.66448	12	일 년 전
4	비회원		0.66509	23	일 년 전
5	인생사는갓김치		0.66513	13	8시간 전
6	닉네임11		0.66598	22	일 년 전

파생변수 생성, 데이터 불균형 문제 고려 등의 전처리 고도화를 통해
XGB, RF, LGBM, CatBoost와 같은 예측 모델 성능 개선



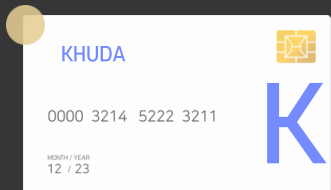
사회초년생 카드

- 신용등급이 높지 않은 신파일러
- 주 사용층 : 34세 이하의 근무기간 3개월 이상 직장인
- 특징 : 편의점, 대중교통, 점심식사 할인



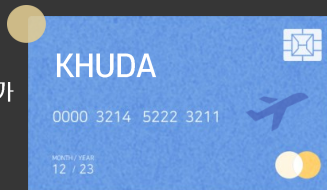
1인 가구 카드

- 주 사용층 : 미혼 1인 가구
- 특징 : 서적, 플랫폼 구독, 배달어플 캐시백, 가정 내 홈 CCTV



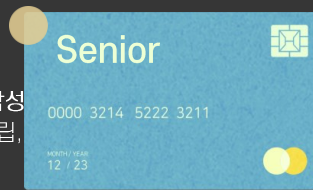
교육 카드

- 주 사용층 : 기혼의 자녀가 있는 30, 40대 여성
- 특징 : 서적, 학원비, 인터넷 강의 할인



세계로 카드

- 주 사용층 : 30~50대 남성
- 특징 : 항공 마일리지 적립, 해외 숙박 결제 할인



시니어 카드

- 주 사용층 : 60세 이상
- 특징 : 병원비, 약국 할인

사용자 신상정보와 신용도 예측 데이터, 신한카드 온라인 소비 분석 데이터를 종합하여 타겟형 신용카드 마케팅 방안 제시

기존 사용자 정보에
기반하여 신용, 대출을
평가하고 등급 및 자산별
맞춤형 관리 서비스 제공

자체적으로 신용도를
예측하고 평가할 수 있는
프로세스를 구축하여
대환대출 평가에 사용

KHUDA 컨퍼런스

감사합니다.

강현빈, 김효준, 신영서, 차민재