데이콘 Basic,
“여행 상품 신청 여부 예측 경진대회”

여행 상품 신청 여부 예측

* 대회 설명

주어진 고객 데이터셋을 이용하여 여행 패키지 신청 여부를 예측
https://www.dacon.io/competitions/official/235959/overview/description

* 변수 설명

- id : 샘플 아이디
- Age : 나이
- TypeofContact : 고객의 제품 인지 방법 (회사의 홍보 or 스스로 검색)
- CityTier : 주거 중인 도시의 등급. (인구, 시설, 생활 수준 기준)
  (1등급 > 2등급 > 3등급)
- DurationOfPitch : 영업 사원이 고객에게 제공하는 프레젠테이션 기간
- Occupation : 직업
- Gender : 성별
- NumberOfPersonVisiting : 고객과 함께 여행을 계획 중인 총 인원
- NumberOfFollowups : 영업 사원의 프레젠테이션 후 이루어진 후속 조치 수
- ProductPitched : 영업 사원이 제시한 상품
- PreferredPropertyStar : 선호 호텔 숙박업소 등급
- MaritalStatus : 결혼여부
- NumberOfTrips : 평균 연간 여행 횟수
- Passport : 여권 보유 여부 (0: 없음, 1: 있음)
- PitchSatisfactionScore : 영업 사원의 프레젠테이션 만족도
- OwnCar : 자동차 보유 여부 (0: 없음, 1: 있음)
- NumberOfChildrenVisiting : 함께 여행을 계획 중인 5세 미만의 어린이 수
- Designation : (직업의) 직급
- MonthlyIncome : 월 급여
- ProdTaken : 여행 패키지 신청 여부 (0: 신청 안 함, 1: 신청함)

* 가설

TypeofContact, CityTier, NumberOfTrips, MonthlyIncome 이 변수가 패키지 신청 여부와 관계가 클 것이다

여행 상품 신청 여부 예측

EDA & 전처리

1

```
Fe Male      56
```

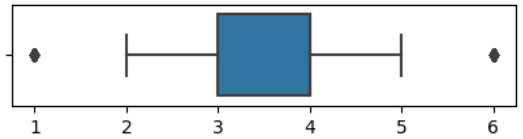
Fe Male을 Female로 바꿈

2

```
Divorced     375
Single       349
```

Divorced와 Single을 Unmarried로 바꿈

3



이상값 제거

4 # 이번 대회의 고득점 방법 ! (고득점자 코드에서 따온 것)

```
imputer = KNNImputer(n_neighbors=3)
imputed_train = imputer.fit_transform(train)
```

KNNImputer를 이용해서 결측값 보정

5

```
train.drop(['id', 'NumberOfChildrenVisiting', 'MonthlyIncome','NumberOfPersonVisiting', 'ProdTaken'], axis = 1)
```

예측에 도움 안되는 변수 제거

6 # 이번 대회의 고득점 방법 ! (고득점자 코드에서 따온 것)

```
classes = np.unique(y_train)
weights = compute_class_weight(class_weight = 'balanced', classes = classes, y = y_train)
class_weights = dict(zip(classes, weights))
```

타겟값 비율의 불균형을 compute_class_weight로 보정
(클래스 별로 가중치를 다르게 부여하는 방법)

여행 상품 신청 여부 예측

* 모델링

catboost 단일 모델
optuna로 하이퍼파라미터 조정 → Best parameters를 이용한 최종 학습의 accuracy_score

0.9018

* 인사이트

! Passport, Age, ProductPitched (여권 보유 여부, 나이, 제안 받은 패키지 상품)이 패키지 신청 여부에 가장 큰 영향을 준다는 것을 알 수 있음

? MonthlyIncome을 뺐을 때가 성능이 더 높게 나오는데 비슷한 변수라고 생각했던 CityTier, Occupation, Designation은 빼지 않는 게 더 나은 성능이 나오는 게 신기했음

? 가설에서 생각했던 TypeofContact, CityTier, NumberOfTrips, MonthlyIncome이 패키지 신청 여부와 크게 관계가 없다는 것도 신기했음

* 배운 점

- KNNImputer를 이용해서 결측값을 보정하는 것
- compute_class_weight로 타겟 값의 비율 불균형을 보정하는 것
- permutation_importance를 통한 피처 중요도 확인하는 것

여행 상품 신청 여부 예측

피쳐 중요도에 따른 추가 EDA

Passport

	O	X
여권 O	10%	18%
여권 X	8%	62%

여권이 없는 사람일 수록 여행 상품
신청을 하지 않는 경향이 있음

Age

	O	X
10대	0.5%, 0.5	0.5%, 0.5
20대	5%, 0.29	12%, 0.7
30대	7%, 0.17	33%, 0.82
40대	3%, 0.13	20%, 0.86
50대	2%, 0.18	9%, 0.81
60대	0%, 0	0.7%, 100

· 30-40-20-50-10-60대 순으로 상담을 받았음

· 10-20-50-30-40-60대 순으로 여행 상품 신청 전환율이 높음

· 20대 / 30-40-50대 / 10대와 60대 순으로 전환율을 나눌 수 있음

ProductPitched

	O	X
Basic	11%, 0.26	30%, 0.73
Standard	4%, 0.13	26%, 0.86
Deluxe	2%, 0.14	12%, 0.85
Super Deluxe	0.5%, 0.07	6%, 0.92
King	0.4%, 0.09	4%, 0.9

· 베이직 상품을 많이 제시했고
킹 상품을 가장 적게 제시했음

· 베이직 상품이 여행 상품 신청 전환율이 높고
스탠다드와 디럭스의 전환율이 비슷하고
슈퍼디럭스와 킹의 전환율이 비슷함