2팀 정리

2팀 정리 보고서

신 용 카 드 사 용 자 연 체 예 측 A I 경 진 대 회

PROEJCT

INDEX

PROEJCT

RED PAGE

데이터 살펴보기

PROEJCT

RED PAGE

데이터 살펴보기

colum별 의미 파악.

datainfo

	name	dtype	missing	nunique	values :5
0	index	int64	0	26457	[0, 1, 2, 3, 4]
1	gender	object	0	2	[F, M]
2	car	object	0	2	[N, Y]
3	reality	object	0	2	[N, Y]
4	child_num	int64	0	9	[0, 1, 2, 3, 4]
5	income_total	float64	0	249	[202500.0, 247500.0, 450000.0, 157500.0, 27000...
6	income_type	object	0	5	[Commercial associate, Working, State servant,...
7	edu_type	object	0	5	[Higher education, Secondary / secondary speci...
8	family_type	object	0	5	[Married, Civil marriage, Separated, Single / ...
9	house_type	object	0	6	[Municipal apartment, House / apartment, With ...
10	DAYS_BIRTH	int64	0	6621	[-13899, -11380, -19087, -15088, -15037]
11	DAYS_EMPLOYED	int64	0	3470	[-4709, -1540, -4434, -2092, -2105]
12	FLAG_MOBIL	int64	0	1	[1]
13	work_phone	int64	0	2	[0, 1]
14	phone	int64	0	2	[0, 1]
15	email	int64	0	2	[0, 1]
16	occyp_type	object	8171	18	[nan, Laborers, Managers, Sales staff, High sk...
17	family_size	float64	0	10	[2.0, 3.0, 4.0, 1.0, 5.0]
18	begin_month	float64	0	61	[-6.0, -5.0, -22.0, -37.0, -26.0]
19	credit	float64	0	3	[1.0, 2.0, 0.0]

데이터 생각.

몇몇개는 숫자. 몇몇개는 범주로 되어 있다. income_total , DAYS_BIRTH , DAYS_EMPOLYED 만 역속형 변수다.

범주형 데이터를 다루는것이 중요하다고 판단했다.

또, DAYS_BIRTH와 DAYS_EMPOLYED에서 마이너스가 보인다. 둘의 의미는 변수설명란에 태어난지(일한지) 몇일 의 의미다. 즉, 뒤로 갈수록 오래일했고, 태어난지 됐다는 뜻이다.

begin_month는 발급월로 -1은 월의 의미다.

occyp에서 nan 값이 보인다. 결측값 처리도 생각해야한다.

FLAG_MOBIL은 값이 하나다. test도 보니 1개 였다.

주의. test는 아예 모른다고 가정한다. 하지만, test도 보니까 1만있었다.

YELLOW PAGE

데이터 분석

PROEJCT

YELLOW PAGE

시각화를 통한 분포 확인

기술 통계와 함께.

종속변수의 분포 확인

독립변수의 분포 확인

변수들의
기술통계 확인

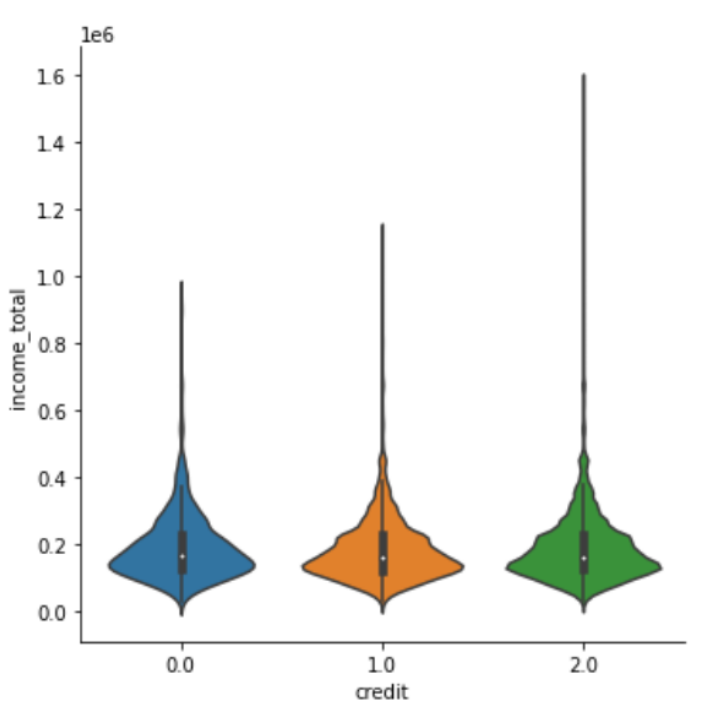
전체적으로
훑어본다.

시각화를 통한 분포

1

credit

credit(종속변수)와 income(수입의 관계)



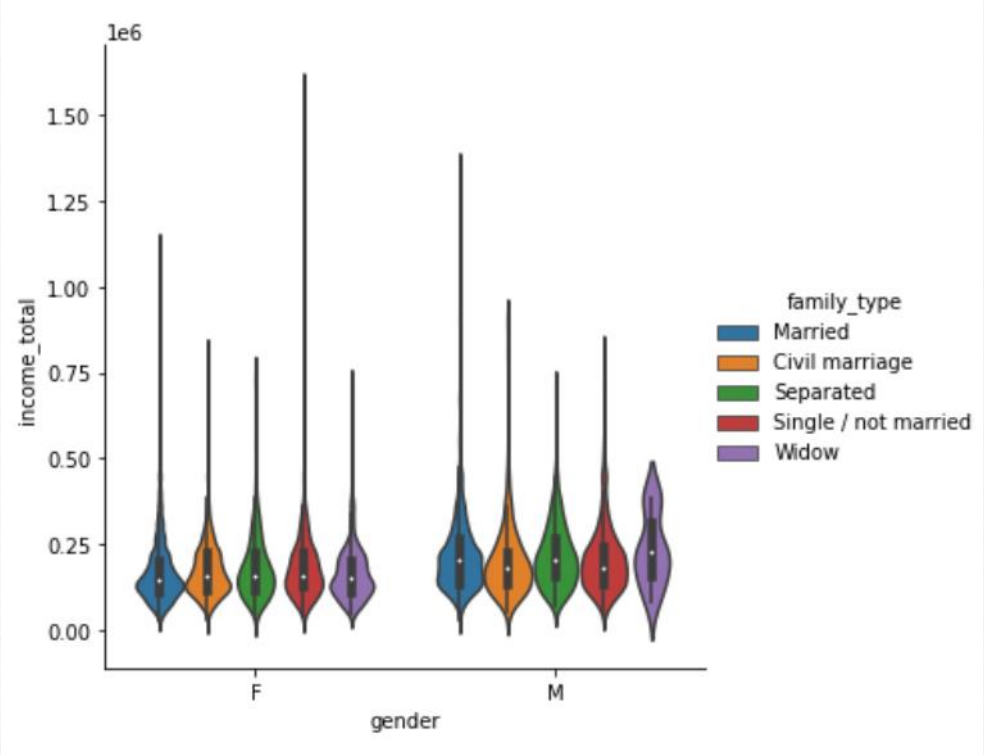
그 외에도 변수에 따라 0,1,2 의 분포가 달라지진 않았다.
주목할만한 점은 2가 넓게 퍼져있다는 것.

시각화를 통한 분포

다른 변수들에서
그렇게 눈에 띄는것이 없다.

2
성별과 수입

credit(종속변수)와 income(수입의 관계)



성별에 따른 수입 분포.
여자가 보통 수입이 많지만, 남자가 가족별 나눴을 때
돈 많이 버는 사람이 여자보다 많다.

시각화를 통한 분포

3
기술통계

이상치

child_num	DAYS_EMPLOYED	family_size
26457.000000	26457.000000	26457.000000
0.428658	59068.750728	2.196848
0.747326	137475.427503	0.916717
0.000000	-15713.000000	1.000000
0.000000	-3153.000000	2.000000
0.000000	-1539.000000	2.000000
1.000000	-407.000000	3.000000
19.000000	365243.000000	20.000000

자식이 19명, 본인포함 20명. 이상치라고 가정한다.
실제로 나이가 30대인데 자식이 14명인 경우도 있다.
위 두 경우를 제거했다.
일한날짜가 양수인경우는 일을 안하는 상태라고 했다.

시각화를 통한 분포

3

기술통계

범주형 변수의 기술통계

분포

	house_type
count	26457
unique	6
top	House / apartment
freq	23653

전체 데이터셋의 8~90%가 house다. 비대칭 하다.

시각화를 통한 분포

4

중복

중복의 이유

birth_date와 credi을 제외한 columns 값들이 같은 경우가 있다. 그 경우 제거 해줬더니 약 8700개의 데이터가 남는다.

첫번째 가정. 새로 발급 받은 같은 사람의 경우.
두번째 가정. 같은 사람이 아닐 경우.

데이콘측은 최신 발급 받은 경우가 있다고 했다.

첫번째 가정으로 일단 진행한다.

시각화를 통한 분포

음수와 결측값의 처리

음수를 갖는 column은 연속형의 의미를 띈다. 즉, 엄청 작게 하든, 크게 하든 상관없을 것이라고 생각했다.

occpy_type의 결측값은 unknown으로 채운다.

BLUE PAGE

데이터 전처리와 모델링

PROEJCT

BLUE PAGE

데이터 전처리

benchmark를 정한다.

benchmark 전처리

```
## 제거
train=train.drop(['FLAG_MOBIL','index'],axis=1)
test=test.drop(['FLAG_MOBIL','index'],axis=1)

## 중복 확인 및 제거
cols=['gender', 'car', 'reality', 'child_num', 'income_total',
      'income_type', 'edu_type', 'family_type', 'house_type', 'DAYS_BIRTH',
      'DAYS_EMPLOYED','work_phone', 'phone', 'email',
      'occyp_type', 'family_size']

train=train.drop_duplicates(subset=cols,keep='last')

## 가족수 이상치
train=train[(train['family_size']!=20)|(train['family_size']!=15)]
# 결측값
train['occyp_type']=train['occyp_type'].fillna('unknown')
test['occyp_type']=test['occyp_type'].fillna('unknown')
```

benchmark를 정해두기 위한 간단한 전처리 .

범주형 변수들은 더미화 시켰다.

후에 실험을 통해 확인 한다.

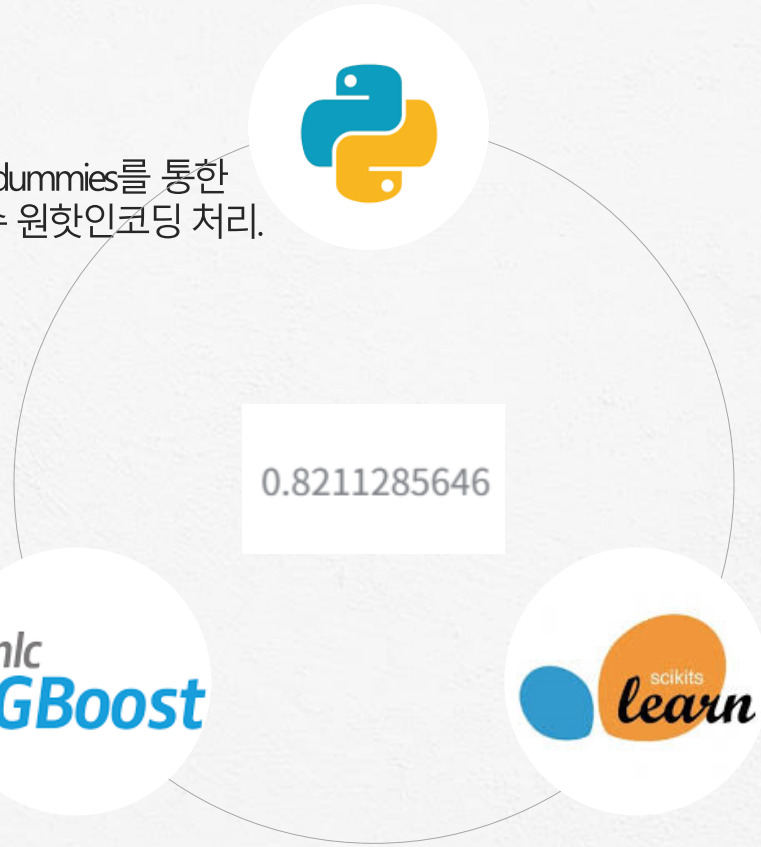
BLUE PAGE

모델링

benchmark

더미화

python get_dummies를 통한
범주형 변수 원핫인코딩 처리.



예측 알고리즘

여러 개의 클래스를 예측하는
xgboost의 multi:softprob

홀드아웃 기법

dataset을 85%,15% 나눠 검증

PROEJCT

BLACK PAGE

모델 개선

PROEJCT

- 20210501 중복제거 안했을 때 0.74 성능 좋아졌다.
- 20210502 모든 column에 한해서 normalize 했을 때 0.73으로 좋아졌다.
- 20210503 #을 통한 범주형의 결합을 해서 label encoding으로 했을 때 0.74로 비슷했다.
- 20210504 음수를 양수로 바꿔줘도 성능향상이 없었다. normalize와 함께 했을 때 0.75나왔다.
- 20210506 family_num과 child_num의 높은 상관관계를 확인했다. 하나 제거했는데 성능 안좋아졌다.
- 20210507 family_num과 child_num의 곱과 덧셈,뺄셈을 통한 파생변수를 만들었지만 성능향상은 없었다.
- 20210508 숫자로 표현된 범주형 변수들의 구간화를 진행했다.
- 20210509 구간화 된 범주와 income의 그룹평균을 통한 통계값을 추가해서 진행했지만 0.78로 떨어졌다.
- feature engineering으로 인한 점수 상승은 중복제거와 정규화말고 없었다.
- 20210510 pytorch 진행. 성능하락.
- 20210510 random_forest와 xgboost의 결과값의 합을 평균으로 했더니 성능향상이 됐다. 0.73
- lightgbm와 svc까지 멱평균으로 진행. 0.71 성능향상. svc는 이상치를 커버해주는 장점 확인.
- 20210511 중복값의 제일 오래된 행만 제거 후 진행. 성능하락.
- 결국 전처리가 주는 효과보다 모델이 주는 효과가 크다는 것을 확인.
- 20210512 중복값이 주는 힌트를 중복된 수라고 가정하고 중복된 행을 그룹화해서 카운트. 성능하락.

- 20210514 중복값이 있다면 1 없으면 0으로 진행. 성능하락.
- 20210518 모델 앙상블 with fold. 성능 0.704 제일 좋다. 모델 쌓는것이 더 좋다.
- 20210520 모델 앙상블 extra tree 와 catboost 더하기 with fold. 성능하락.
- 20210521 모델 앙상블 가중치 with fold. 가중치는 소용이 없다.
- 20210523 끝. 최종모델 lgb,xgb,randomforest. 가중치 1:1:2

THANK YOU

끝 까 지 봐 주 셔 서 감 사 합 니 다

PROEJCT

THANK YOU