신용카드 사용자 연체 예측 AI 경진대회

5

재벌 집 막내 조

LAST WEEK (1w)

5

변수의 유의성 검정

A

결측값 처리

B

파생변수 생성 마무리

C

파생변수 생성 마무리

5

```
df['근속연수'] = df['DAYS_EMPLOYED'] // 365 # 근속연수
df['근속월수'] = df['DAYS_EMPLOYED'] // 30 # 근속월수
df['임용 월'] = np.floor(df['DAYS_EMPLOYED'] / 30) - ((np.floor(df['DAYS_EMPLOYED'] / 30) / 12).astype(int) * 12) # 고용된 달
df['임용 주'] = np.floor(df['DAYS_EMPLOYED'] / 7) - ((np.floor(df['DAYS_EMPLOYED'] / 7) / 4).astype(int) * 4) # 고용된 주
df["고용전 날 수"] = df["DAYS_BIRTH"] - df["DAYS_EMPLOYED"]
```

```
df['고용비율'] = df['DAYS_EMPLOYED'] / df['DAYS_BIRTH'] # 인생 살면서 일한 비율
df['인당 평균 부양비'] = df['income_total'] / df['family_size']
```

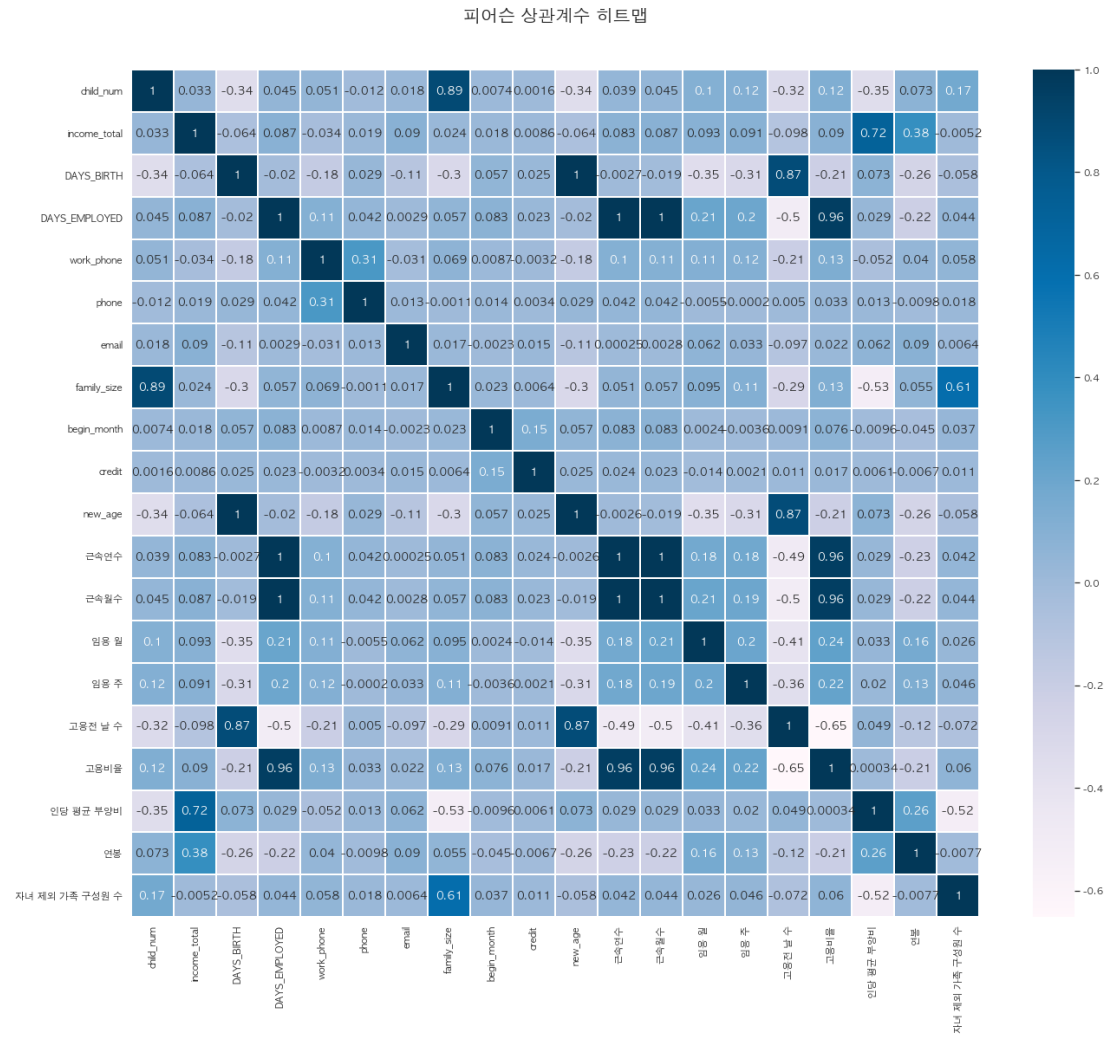
```
df['연봉'] = df['income_total'] / (df['근속연수'])
```

gender	car	reality	child_num	income_total	income_type	edu_type	family_type	...	new_age	근속연수	근속월수	임용월	임용주	고용전 날 수	고용비율	인당 평균 부양비	연봉	자녀 제외 가족 구성원 수
F	N	N	0	202500.0	Commercial associate	Higher education	Married	...	38	12	156	0.0	0.0	9190	0.338801	101250.0	16875.0	2.0
F	N	Y	1	247500.0	Commercial associate	Secondary / secondary special	Civil marriage	...	31	4	51	3.0	0.0	9840	0.135325	82500.0	61875.0	2.0
M	Y	Y	0	450000.0	Working	Higher education	Married	...	52	12	147	3.0	1.0	14653	0.232305	225000.0	37500.0	2.0

신용카드 사용기간 변수 추가 생성 (예정)

파생변수 생성 마무리

5



상관계수 파악 (파생변수 포함)

독립 변수간의 다중공선성 확인
→ VIF 지수 확인 필요

범주의 유의성 검정

5

	독립변수	종속변수
t검정	범주형	수치형
분산분석 (일원 분산분석)	범주형	수치형
카이제곱검정	범주형	범주형
상관분석 (피어슨)	수치형	수치형
회귀분석 (단순 선형)	수치형	수치형
로지스틱 회귀분석	수치형 (or 범주형)	범주형

카이제곱 독립성 검정

두 가지 범주형 또는 명목형 변수가 관련될 가능성 여부를 확인하는데 사용되는 통계적 가설 검정

H0 : 독립변수와 종속변수는 독립이다
H1 : 독립변수와 종속변수는 독립이 아니다

H0 : 독립변수와 종속변수는 관련성이 없다
H1 : 독립변수와 종속변수는 관련성이 있다

범주의 유의성 검정

5

	chi_2	p-value	df
gender	0.742683	6.898085e-01	2
car	9.366187	9.250354e-03	2
reality	11.230277	3.642304e-03	2
child_num	19.978358	2.945805e-02	10
income_type	23.800389	2.475172e-03	8
edu_type	8.886748	3.519398e-01	8
family_type	46.383397	2.009568e-07	8
house_type	37.725432	4.236293e-05	10
work_phone	0.385865	8.245374e-01	2
phone	7.995643	1.835558e-02	2
email	6.107595	4.717942e-02	2
occyp_type	88.791126	8.757964e-07	34
임용 월	66.017750	2.756842e-06	22
임용 주	7.283921	2.953866e-01	6
자녀 제외 가족 구성원 수	25.970734	2.254547e-04	6
family_size	39.212899	9.706224e-05	12

P-value < 0.05

통계량 > 10

	chi_2	p-value	df
reality	11.230277	3.642304e-03	2
child_num	19.978358	2.945805e-02	10
income_type	23.800389	2.475172e-03	8
family_type	46.383397	2.009568e-07	8
house_type	37.725432	4.236293e-05	10
occyp_type	88.791126	8.757964e-07	34
임용 월	66.017750	2.756842e-06	22
자녀 제외 가족 구성원 수	25.970734	2.254547e-04	6
family_size	39.212899	9.706224e-05	12

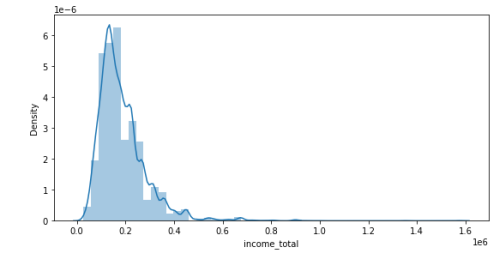
전체 범주형 변수 : 16개

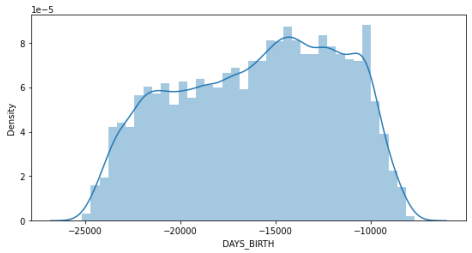
변수 선택 후 범주형 변수 : 9개

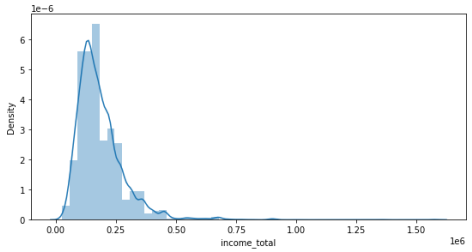
Logistic 수치형 변수 검정 & VIF 지수 확인 필요

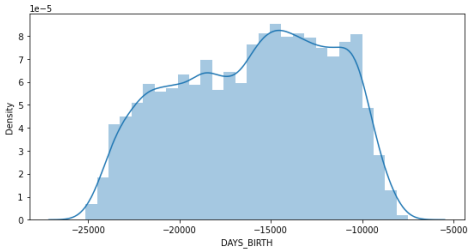
결측값 처리

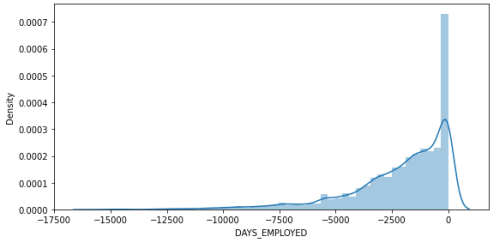
5

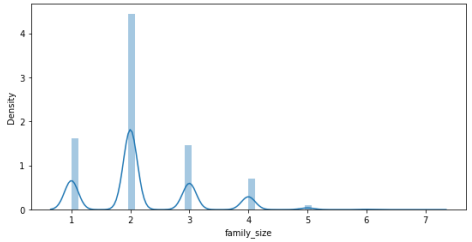


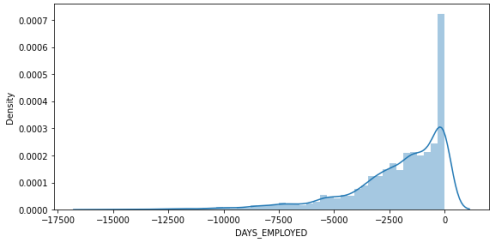


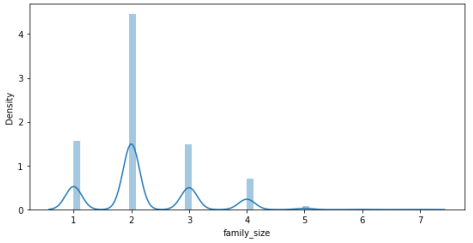


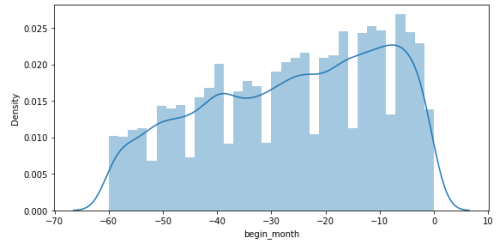




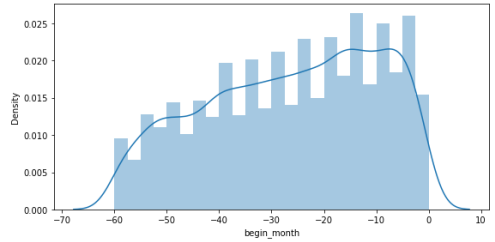








TRAIN



TEST

Train test 의 큰 차이 존재 X

결측치 처리의 어려움 → 고민 필요

NEXT WEEK (3w)

5

변수의 유의성 검정 마무리

A

결측값 처리 마무리

B

VIF 수치 확인

C

감사합니다