신용카드 사용자 연체 예측 AI 경진대회

5

재벌 집 막내 조

LAST WEEK (3w)

5

변수의 유의성 검정 마무리

A

VIF 수치 확인

B

결측값 처리

C

변수의 유의성 검정

5

개별 값을 기준으로 로지스틱 검정

logistic income_total

Generalized Linear Model Regression Results

Dep. Variable:	output	No. Observations:	26451
Model:	GLM	Df Residuals:	26449
Model Family:	Binomial	Df Model:	1
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-inf
Date:	Tue, 17 Jan 2023	Deviance:	1.5665e+06
Time:	05:09:48	Pearson chi2:	9.09e+19
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.29e+15	8.64e+05	2.65e+09	0.000	2.29e+15	2.29e+15
feature	2.658e+08	4.050	6.56e+07	0.000	2.66e+08	2.66e+08

income_total

logistic DAYS_BIRTH

Generalized Linear Model Regression Results

Dep. Variable:	output	No. Observations:	26451
Model:	GLM	Df Residuals:	26449
Model Family:	Binomial	Df Model:	1
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-inf
Date:	Tue, 17 Jan 2023	Deviance:	1.5665e+06
Time:	05:09:48	Pearson chi2:	9.09e+19
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.037e+15	1.62e+06	1.26e+09	0.000	2.04e+15	2.04e+15
feature	1.898e+10	98.203	1.93e+08	0.000	1.9e+10	1.9e+10

DAYS_BIRTH

p-value 가 모두 0에 가깝게 나오고 통계량이 크게 나옴
Odds 비가 0 or inf 로 나옴 -> 변수 선택의 의미가 없다

변수의 유의성 검정

5

수치형 변수 전체를 이용해 로지스틱 검정

Generalized Linear Model Regression Results

Dep. Variable:	output	No. Observations:	26451
Model:	GLM	Df Residuals:	26442
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-inf
Date:	Tue, 17 Jan 2023	Deviance:	1.5665e+06
Time:	05:13:22	Pearson chi2:	9.09e+19
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.321e+15	2.8e+06	1.54e+09	0.000	4.32e+15	4.32e+15
income_total	-2.499e+07	4.090	-6.11e+06	0.000	-2.5e+07	-2.5e+07
DAYS_BIRTH	8.982e+09	3867.453	2.32e+06	0.000	8.98e+09	8.98e+09
DAYS_EMPLOYED	1.943e+12	4.79e+04	4.06e+07	0.000	1.94e+12	1.94e+12
family_size	1.683e+13	4.84e+05	3.48e+07	0.000	1.68e+13	1.68e+13
begin_month	2.144e+13	2.51e+04	8.55e+08	0.000	2.14e+13	2.14e+13
new_age	8.336e+11	1.41e+06	5.9e+05	0.000	8.34e+11	8.34e+11
근속연수	1.375e+14	1.45e+06	9.46e+07	0.000	1.37e+14	1.37e+14
근속월수	-6.934e+13	1.45e+06	-4.78e+07	0.000	-6.93e+13	-6.93e+13

결과

```
""", Intercept          inf
income_total      0.0
DAYS_BIRTH        inf
DAYS_EMPLOYED     inf
family_size       inf
begin_month       inf
new_age           inf
근속연수                inf
근속월수                0.0
dtype: float64)
```

모두 p-value 가 낮고
Odds 비가 0 or inf 나옴

의미 없음

변수의 유의성 검정

5

타겟변수 = credit (신용등급)
0,1,2 로 구분
0으로 갈수록 높은 신용등급 의미

로지스틱 회귀는
0 or 1 로만 하는 회귀

우리가 하는 데이터
의 타겟변수는 0,1,2

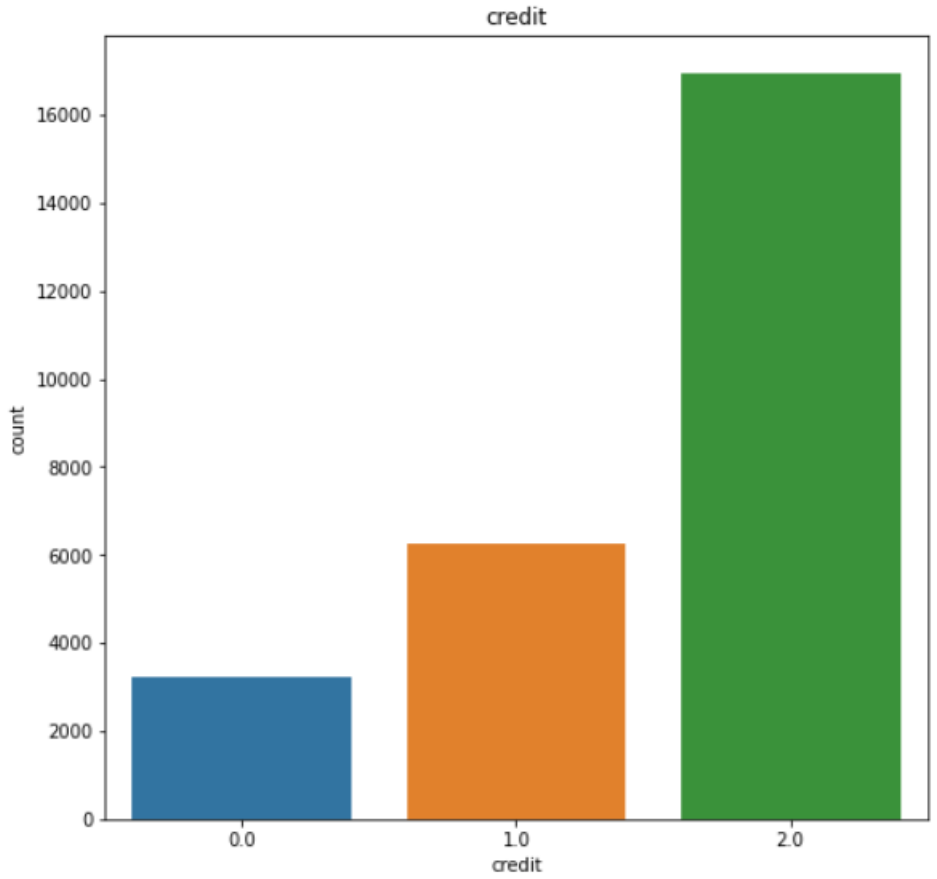
로지스틱 검정에
적합하지 않을 수 있음

Y≠2.0 인 (0.0, 1.0) 으로 로지스틱 회귀 실시

변수의 유의성 검정

5

Y≠2.0 인 (0.0, 1.0) 으로 로지스틱 회귀 실시 이유



0과 1의 분포는 비슷한데
0과 2, 1과 2 의 분포는 차이가 큼

이 분포의 차이 때문인지 올바른 결과가 나오지 않음

그래서 분포가 비슷한 0과 1로 로지스틱 회귀 실시

0과 2, 1과 2 는 SMOTE 오버 샘플링을 통해
분포를 비슷하게 해준 뒤 다시 실시할 예정!

변수의 유의성 검정

5

Y≠2.0 인 (0.0, 1.0) 으로 로지스틱 회귀 실시

Generalized Linear Model Regression Results

Dep. Variable:	output	No. Observations:	9489
Model:	GLM	Df Residuals:	9480
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-6035.1
Date:	Tue, 17 Jan 2023	Deviance:	12070.
Time:	05:10:07	Pearson chi2:	9.48e+03
No. Iterations:	4		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.2109	0.145	8.358	0.000	0.927	1.495
income_total	-9.933e-07	2.25e-07	-4.420	0.000	-1.43e-06	-5.53e-07
DAYS_BIRTH	-9.725e-06	0.000	-0.048	0.962	-0.000	0.000
DAYS_EMPLOYED	0.0057	0.003	2.249	0.025	0.001	0.011
family_size	-0.0221	0.025	-0.877	0.380	-0.072	0.027
begin_month	-0.0093	0.001	-7.399	0.000	-0.012	-0.007
new_age	0.0011	0.075	0.014	0.989	-0.145	0.147
근속연수	0.1177	0.077	1.526	0.127	-0.033	0.269
근속월수	-0.1813	0.077	-2.355	0.019	-0.332	-0.030

결과

```
""", Intercept        3.356457
income_total      0.999999
DAYS_BIRTH        0.999990
DAYS_EMPLOYED     1.005730
family_size       0.978098
begin_month       0.990721
new_age           1.001065
근속연수              1.124893
근속월수              0.834147
dtype: float64)
```

p-value < 0.05
Odds > 1 :
DAY_EMPLOYED

이후 VIF 결과와 비교해 변수 선택 필요

VIF 수치 확인

5

	VIF_Factor	Feature
0	inf	DAYS_BIRTH
1	inf	DAYS_EMPLOYED
2	inf	고용전 날 수
3	8.005880e+04	근속월수
4	1.552151e+03	new_age
5	5.269276e+02	근속연수
6	6.659122e+01	intercept
7	2.688998e+01	고용비율
8	2.898712e+00	인당 평균 부양비
9	2.806763e+00	income_total
10	1.543528e+00	child_num
11	1.537337e+00	연봉
12	1.014073e+00	begin_month

수치형 변수의 VIF값을 내림차순으로 산출

VIF 지수10 이상이면 다중 공산성 보유 가능성 높음

VIF 지수 10 이상의 변수를 삭제

오히려 모델의 정확도가 낮아지는 경우 발생

VIF 수치 확인

5

```
def vif(x):
    # vif 10 초과시 drop을 위한 임계값 설정
    thresh = 10
    # Filter method로 feature selection 진행 후 최종 도출 될 데이터 프레임 형성
    output = pd.DataFrame()
    # 데이터의 컬럼 개수 설정
    k = x.shape[1]
    # VIF 측정
    vif = [variance_inflation_factor(x.values, i) for i in range(x.shape[1])]
    for i in range(1,k):
        print(f'{i}번째 VIF 측정')
        # VIF 최대 값 선정
        a = np.argmax(vif) # np.argmax -> 가장 큰 값이 있는 인덱스 값을 반환하는 메서드
        print(f'Max VIF feature & value : {x.columns[a]}, {vif[a]}')
        # VIF 최대 값이 임계치를 넘지 않는 경우 break
        if (vif[a] <= thresh):
            print('₩n')
            for q in range(output.shape[1]):
                print(f'{output.columns[q]}의 vif는 {np.round(vif[q],2)}입니다.')
            break
        # VIF 최대 값이 임계치를 넘는 경우, + 1번째 시도인 경우 : if 문으로 해당 feature 제거 후 다시 vif 측정
        if (i == 1):
            output = x.drop(x.columns[a], axis = 1)
            vif = [variance_inflation_factor(output.values, j) for j in range(output.shape[1])]
        # VIF 최대 값이 임계치를 넘는 경우, + 1번째 이후 시도인 경우 : if 문으로 해당 feature 제거 후 다시 vif 측정
        elif (i > 1):
            output = output.drop(output.columns[a], axis = 1)
            vif = [variance_inflation_factor(output.values, j) for j in range(output.shape[1])]
    return(output)
```

필터 메서드

다중공선성이 높은 변수를 하나씩 제거할 때마다 다시 VIF값을 산출해서 변수를 하나씩 제거하는 방법을 통해 좀 더 정확하게 제거 할 변수를 판별

VIF 수치 확인

5

필터 메서드

다중공선성이 높은 변수를 하나씩 제거할 때마다 다시 VIF값을 산출해서 변수를 하나씩 제거하는 방법을 통해 좀 더 정확하게 제거 할 변수를 판별

```
1번째 VIF 측정
Max VIF feature & value : DAYS_BIRTH, inf
2번째 VIF 측정
Max VIF feature & value : 근속연수, 80058.7980739917
3번째 VIF 측정
Max VIF feature & value : 근속연수, 2084.3821386880277
4번째 VIF 측정
Max VIF feature & value : new_age, 518.2447877525036
5번째 VIF 측정
Max VIF feature & value : 고용전 날 수, 50.999325602834624
6번째 VIF 측정
Max VIF feature & value : new_age, 29.616854435513314
7번째 VIF 측정
Max VIF feature & value : income_total, 12.114627443730756
8번째 VIF 측정
Max VIF feature & value : DAYS_EMPLOYED, 4.987985884139117
```

6개의 변수

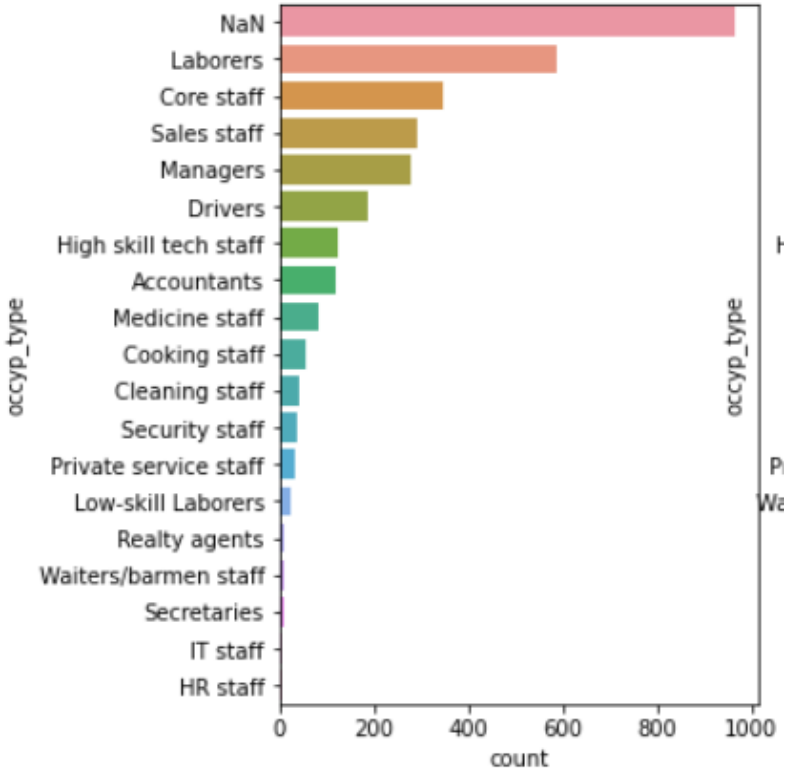
child_num의 vif는 1.42입니다.
DAYS_EMPLOYED의 vif는 1.92입니다.
begin_month의 vif는 3.33입니다.
new_age의 vif는 4.99입니다.
인당 평균 부양비의 vif는 3.47입니다.
연봉의 vif는 1.69입니다.

VIF 필터 메소드 결과
분산팽창요인이 10이하인
6개의 변수

결측값 처리

5

occyp_type 변수에 약 8000개의 결측값



Nan 값으로 결측값 채우기

```
Laborers는 전체의 25%입니다.
Core staff는 전체의 14%입니다.
Sales staff는 전체의 14%입니다.
Managers는 전체의 12%입니다.
Drivers는 전체의 9%입니다.
High skill tech staff는 전체의 6%입니다.
Accountants는 전체의 5%입니다.
Medicine staff는 전체의 5%입니다.
Cooking staff는 전체의 2%입니다.
Security staff는 전체의 2%입니다.
Cleaning staff는 전체의 2%입니다.
Private service staff는 전체의 1%입니다.
Low-skill Laborers는 전체의 1%입니다.
Waiters/barmen staff는 전체의 1%입니다.
Secretaries는 전체의 1%입니다.
Realty agents는 전체의 0%입니다.
HR staff는 전체의 0%입니다.
IT staff는 전체의 0%입니다.
```

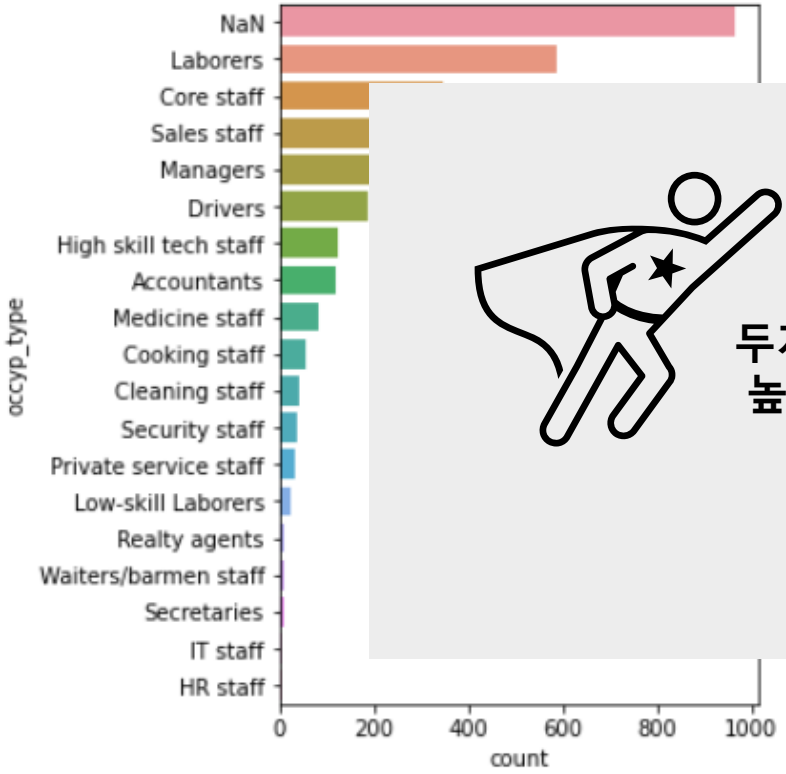
```
Laborers는 결측값의 2017개를 차지합니다.
Core staff는 결측값의 1183개를 차지합니다.
Sales staff는 결측값의 1135개를 차지합니다.
Managers는 결측값의 969개를 차지합니다.
Drivers는 결측값의 703개를 차지합니다.
High skill tech staff는 결측값의 465개를 차지합니다.
Accountants는 결측값의 403개를 차지합니다.
Medicine staff는 결측값의 386개를 차지합니다.
Cooking staff는 결측값의 204개를 차지합니다.
Security staff는 결측값의 190개를 차지합니다.
Cleaning staff는 결측값의 179개를 차지합니다.
Private service staff는 결측값의 109개를 차지합니다.
Low-skill Laborers는 결측값의 57개를 차지합니다.
Waiters/barmen staff는 결측값의 55개를 차지합니다.
Secretaries는 결측값의 43개를 차지합니다.
Realty agents는 결측값의 28개를 차지합니다.
HR staff는 결측값의 28개를 차지합니다.
IT staff는 결측값의 18개를 차지합니다.
```

직업 유형이 전체의 몇 퍼센트인지 확인하고
결측값 8000개에서 몇개를 채워줘야 하는지
비율과 개수 산출해서 결측값 채우기

결측값 처리

5

occyp_type 변수에 약 8000개의 결측값



Nan 값으로 결측값 채우기

두개의 방법 중 모델링 후 더 높은 성능인 방법으로 결정

```
Laborers는 결측값의 2017개를 차지합니다.
Core staff는 결측값의 1183개를 차지합니다.
Sales staff는 결측값의 1135개를 차지합니다.
Managers는 결측값의 969개를 차지합니다.
Drivers는 결측값의 703개를 차지합니다.
High skill tech staff는 결측값의 465개를 차지합니다.
Accountants는 결측값의 403개를 차지합니다.
Medicine staff는 결측값의 386개를 차지합니다.
Cooking staff는 결측값의 204개를 차지합니다.
Security staff는 결측값의 190개를 차지합니다.
Cleaning staff는 결측값의 179개를 차지합니다.
Private service staff는 결측값의 109개를 차지합니다.
Low-skill Laborers는 결측값의 57개를 차지합니다.
Waiters/barmen staff는 결측값의 55개를 차지합니다.
Secretaries는 결측값의 43개를 차지합니다.
Realty agents는 결측값의 28개를 차지합니다.
HR staff는 결측값의 28개를 차지합니다.
IT staff는 결측값의 18개를 차지합니다.
```

직업 유형이 전체의 몇 퍼센트인지 확인하고 결측값 8000개에서 몇개를 채워줘야 하는지 비율과 개수 산출해서 결측값 채우기

NEXT WEEK (4w)

5

SMOTE 오버 샘플링 후 변수 선택

A

인코딩

B

모델링

C

감사합니다