

재벌 집 막내 조

월간 데이콘 신용카드 사용자 연체 예측 AI 경진대회

알고리즘 | 정형 | 분류 | 금융 | LogLoss

목적

신용카드 사용자 데이터를 보고 사용자의 대금 연체 정도를 예측하는 알고리즘 개발

주최/주관

주최: 데이콘

주관: 데이콘

팀원

최지혁, 신주연, 박은비, 김진호



재벌 집 막내 조

Train.csv

	index	gender	car	reality	child_num	income_total	income_type	edu_type	family_type	house_type	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	work_phone	phone	email	occyp_type	family_size
0	0	F	N	N	0	202500.0	Commercial associate	Higher education	Married	Municipal apartment	-13899	-4709	1	0	0	0	NaN	2.0
1	1	F	N	Y	1	247500.0	Commercial associate	Secondary / secondary special	Civil marriage	House / apartment	-11380	-1540	1	0	0	1	Laborers	3.0
2	2	M	Y	Y	0	450000.0	Working	Higher education	Married	House / apartment	-19087	-4434	1	0	1	0	Managers	2.0
3	3	F	N	Y	0	202500.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-15088	-2092	1	0	1	0	Sales staff	2.0

거의 모든 데이터가 범주형 데이터와 소득과 같은 연속형 데이터들로 이루어짐

따라서 **Feature Engineering**

즉, 파생변수를 적절히 생성하는 것이 굉장히 중요하다고 판단

모델보다는 전처리와 파생 변수 생성에 힘쓸 예정이다.

재벌 집 막내 조

Train.csv

```
[5]: df.isnull().sum()
```

```
[5]: index          0
gender            0
car              0
reality          0
child_num        0
income_total     0
income_type      0
edu_type         0
family_type      0
house_type       0
DAYS_BIRTH       0
DAYS_EMPLOYED    0
FLAG_MOBIL       0
work_phone       0
phone            0
email            0
occyp_type      8171
family_size      0
begin_month      0
credit           0
dtype: int64
```

	index	child_num	income_total	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	work_phone	
count	26457.00	26457.00	26457.00	26457.00	26457.00	26457.00	26457.00	2
mean	13228.00	0.43	187306.52	-15958.05	59068.75	1.00	0.22	
std	7637.62	0.75	101878.37	4201.59	137475.43	0.00	0.42	
min	0.00	0.00	27000.00	-25152.00	-15713.00	1.00	0.00	
25%	6614.00	0.00	121500.00	-19431.00	-3153.00	1.00	0.00	
50%	13228.00	0.00	157500.00	-15547.00	-1539.00	1.00	0.00	
75%	19842.00	1.00	225000.00	-12446.00	-407.00	1.00	0.00	
max	26456.00	19.00	1575000.00	-7705.00	365243.00	1.00	1.00	

재벌 집 막내 조

Train.csv

```
[7]: df['edu_type'].value_counts()
```

```
[7]: Secondary / secondary special    17995  
     Higher education                 7162  
     Incomplete higher                1020  
     Lower secondary                  257  
     Academic degree                  23  
     Name: edu_type, dtype: int64
```

```
[8]: df['income_type'].value_counts()
```

```
[8]: Working                13645  
     Commercial associate    6202  
     Pensioner               4449  
     State servant           2154  
     Student                  7  
     Name: income_type, dtype: int64
```

```
[9]: df['family_type'].value_counts()
```

```
[9]: Married                18196  
     Single / not married   3496  
     Civil marriage         2123  
     Separated              1539  
     Widow                 1103  
     Name: family_type, dtype: int64
```

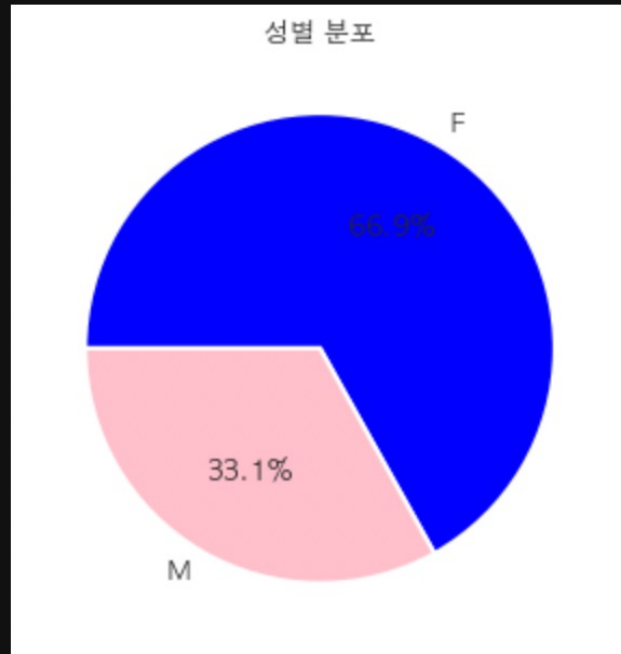
```
[10]: df['house_type'].value_counts()
```

```
[10]: House / apartment       23653  
     With parents          1257  
     Municipal apartment    818  
     Rented apartment       429  
     Office apartment       190  
     Co-op apartment        110  
     Name: house_type, dtype: int64
```

재벌 집 막내 조

Train.csv

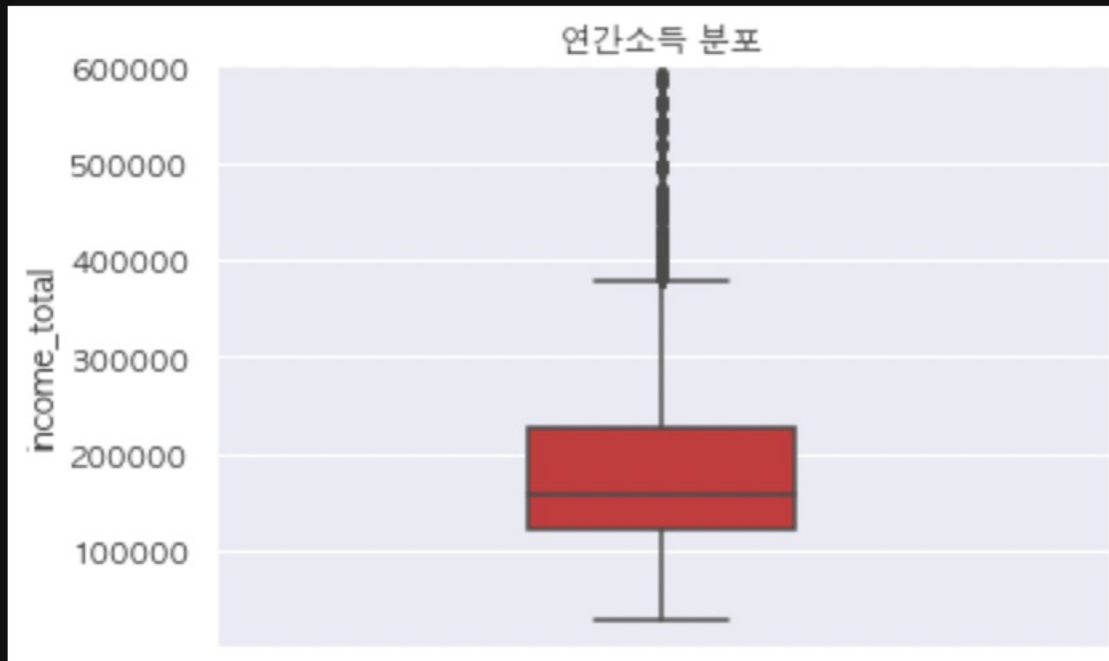
```
[17]: df1= df.groupby('gender').count()['index']  
fig = plt.figure(figsize=(5,5)) ## 캔버스 생성  
fig.set_facecolor('white')  
plt.pie(df1, labels=df1.index,colors=['blue','pink'], startangle=180,autopct='%1.1f%%',co  
plt.title('성별 분포')  
plt.show()
```



재벌 집 막내 조

Train.csv

```
[24]: sns.boxplot(data=df, y="income_total", width=0.3, color='tab:red')  
plt.title('연간소득 분포')  
plt.ylim(1000, 600000)  
plt.show()
```



재벌 집 막내 조

앞으로의 계획

변수 간의 관계확인
및 EDA



결측, 이상치 등
전처리 진행



Feature
Engineering



제출



적절모델 설정 후
모델링 진행



다중공선성 처리 및
인코딩 등 전처리
마무리