3조 프로젝트 기획서

기획자 : 유호준

프로젝트 제목	A기업 퇴사자 현황 분석을 통한 인사정책 개선 방향 도출
제안 배경	○ A 신용카드사는 신용카드의 개인정보와 데이터를 활용해 신용 점수를 산정하여, 신청자의 향후 채무 불이행과 신용카드 대금 연체 가능성을 예측하려 한다.
프로젝트 목적	① 신용카드 사용자들의 개인 신상정보 데이터로 사용자 신용카드 대금 연체 정도를 예측할 수 있는 알고리즘 개발 ② 알고리즘을 통해 금융업계에 제안할 인사이트를 발굴

플로우 | ① EDA | 1

① 데이터셋 컬럼 파악

1) 인구통계 측면	성별
	자녀 수
	결혼 여부
	교육 수준
	생활 방식
	출생일
	가족 규모
2) 소득 측면	소득 분류
	연간 소득
	차량 소유 여부
	부동산 소유 여부
3) 직업 측면	직업 유형
	업무용 전화 소유 여부
	업무 시작일
4) 기타	핸드폰 소유 여부
	전화 소유 여부
	이메일 소유 여부
	신용카드 발급 월
5) 타겟 레이블	신용도

컬럼은 타겟 레이블을 제외하고 크게 **4측면으로 분류 가능**하다.
타겟 레이블(신용도)가 미리 주어지기에 **알고리즘은 지도학습의 회귀모델이 적합**하고, **신용도에 미치는 영향이 다양하기에 다중회귀 분석의 알고리즘을 채택**한다.

데이콘 데이터의 특성상 정제된 데이터를 제공 받았다. 따라서 전처리 과정에 들어가기 전에, **다중회귀 분석을 위한 피처들의 조합을 우선적으로 검토한다.**

2

② 피처 조합 선정

피처 조합은 **과대적합을 예방하기 위해 피처 수를 4개로 제한**한다.

시간이 다소 소요되어도, 정확도를 높이기 위해 모든 경우의 수를 고려한다.
4분류의 컬럼들을 리스트로 만들고, for 문을 이용하여 다중 회귀분석을 진행한다.

리스트 변수 명은 다음과 같이 지정한다.

```
demography_side = [ ]
income_side = []
occupation_side = []
etc_side = []
```

3

③ 전처리 진행

*** 결측값 확인**

```
In [2]:
1 # 결측값 확인
2 test_df.isnull().sum()
```

```
Out[2]:
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      3152
family_size        0
begin_month        0
dtype: int64
```

occyp_type 이외의 결측값은 없다.
occyp_type 결측값의 경우 무직이기 때문에, 전처리 없이 진행한다.

*** 중복 데이터 확인**

```
# 중복 데이터 확인
test_df.duplicated().sum()
```

```
0
```

중복 데이터는 존재하지 않는다.

예상대로 데이터셋이 정제되어 있다.

6

7

8

		13	
		14	

		15	