

# 신용카드 사용자 연체 예측 AI 경진대회

TEAM3 (2조) 팀프로젝트 중간 발표

안동현, 김태용, 안준용, 이문형, 이종섭

# 목차

1. 팀원 소개
2. 프로젝트 개요
3. 프로젝트 필요성
4. 데이터 설명 및 탐색
5. 프로젝트 진행사항
6. 프로젝트 계획

# 1. 팀원 소개

팀명 : TEAM3

팀장 : 안동현

팀원 : 김태용, 안준용, 이문형, 이종섭



안동현



김태용



안준용



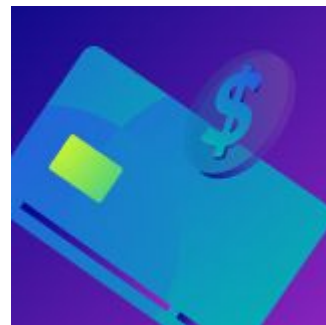
이문형



이종섭

## 2. 프로젝트 개요

데이콘 - 신용카드 사용자 연체 예측 AI 경진대회



### □ 분석개요

고객의 소득, 직업, 가족 구성, 부동산 등의 정보를 활용하여 신용도를 예측함  
신용도를 구하는 공식은 주어지지 않음 (연체 일수, 연체 횟수 등에 의해 부여)

※ 경진대회 특성 상, 주어진 변수 및 데이터를 기반으로 예측 모델을 만들기 때문에  
실제 신용도 예측과는 다소 차이가 있을 수 있음

## 2. 프로젝트 개요

데이콘 - 신용카드 사용자 연체 예측 AI 경진대회

### □ 분석 내용

사용자별 조건에 따른 신용도를 예측함

- 중복되는 사용자가 존재함 (같은 사람이 같은 날에 여러 카드 발급,  
같은 사람이 다른 날에 여러 카드 발급 등)

신용도 : 사용자의 신용카드 대금 연체를 기준으로 한 신용도(0, 1, 2)를 예측함  
(낮을수록 높은 신용의 신용카드 사용자)

### 3. 프로젝트 필요성

#### ❑ 기존 신용 평가 방식의 한계점

- 신용등급 정보 적시성 저하
- 정성적 평가에 있어 객관적 근거 제시 미흡
- 신용도와 비선형적인 관계를 가지는 요소에 대한 부정확한 판단 가능성

출처 : 빅데이터와 AI를 활용한 신용평가의 변화 시도 - 서울신용평가

#### ❑ 기업의 입장에서서는 신용도가 불량한 사람을 미리 예측해야 함

- 이자수익 뿐만 아니라 대출금 자체의 손실 위험이 있음
- Interpretability 측면에서도 중요함 (억울한 사람 및 남용 방지)

## 4. 데이터 설명 및 탐색

### □ 정량적 분석

- 훈련 데이터 셋 : 26,457 rows X ( 18 features + 1 target )

```
[124] train.head(3)
```

|       | 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 | begin_month | credit |
|-------|--------|-----|---------|-----------|--------------|----------------------|-------------------------------|----------------|---------------------|------------|---------------|------------|------------|-------|-------|------------|-------------|-------------|--------|
| index |        |     |         |           |              |                      |                               |                |                     |            |               |            |            |       |       |            |             |             |        |
| 0     | F      | N   | N       | 0         | 202500.0     | Commercial associate | Higher education              | Married        | Municipal apartment | -13899     | -4709         | 1          | 0          | 0     | 0     | NaN        | 2.0         | 6.0         | 1.0    |
| 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         | 5.0         | 1.0    |
| 2     | M      | Y   | Y       | 0         | 450000.0     | Working              | Higher education              | Married        | House / apartment   | -19087     | -4434         | 1          | 0          | 1     | 0     | Managers   | 2.0         | 22.0        | 2.0    |

- 평가 데이터 셋 : 10,000 rows X ( 18 features )

```
[125] test.head(3)
```

|   | 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 | begin_month |
|---|-------|--------|-----|---------|-----------|--------------|---------------|-------------------------------|----------------|-------------------|------------|---------------|------------|------------|-------|-------|------------|-------------|-------------|
| 0 | 26457 | M      | Y   | N       | 0         | 112500.0     | Pensioner     | Secondary / secondary special | Civil marriage | House / apartment | -21990     | 365243        | 1          | 0          | 1     | 0     | NaN        | 2.0         | -60.0       |
| 1 | 26458 | F      | N   | Y       | 0         | 135000.0     | State servant | Higher education              | Married        | House / apartment | -18964     | -8671         | 1          | 0          | 1     | 0     | Core staff | 2.0         | -36.0       |
| 2 | 26459 | F      | N   | Y       | 0         | 69372.0      | Working       | Secondary / secondary special | Married        | House / apartment | -15887     | -217          | 1          | 1          | 1     | 0     | Laborers   | 2.0         | -40.0       |

## 4. 데이터 설명 및 탐색

### □ 데이터 특성

| 특성 이름        | 설명        | 데이터 유형   | 특성 이름         | 설명           | 데이터 유형   |
|--------------|-----------|----------|---------------|--------------|----------|
| gender       | 성별        | binary   | DAYS_BIRTH    | 나이 (일수로 계산)  | numeric  |
| car          | 차량 소유 여부  | binary   | DAYS_EMPLOYED | 근속일 수        | numeric  |
| reality      | 부동산 소유 여부 | binary   | FLAG_MOBIL    | 핸드폰 소유 여부    | binary   |
| child_num    | 자녀 수      | numeric  | work_phone    | 업무용 전화 소유 여부 | binary   |
| income_total | 연간 소득     | numeric  | phone         | 전화 소유 여부     | binary   |
| income_type  | 소득 분류     | category | email         | 이메일 소유 여부    | binary   |
| edu_type     | 교육 수준     | category | occyp_type    | 직종           | category |
| family_type  | 결혼 여부     | category | family_size   | 가족 규모        | numeric  |
| house_type   | 거주 형태     | category | begin_month   | 신용카드 이용 기간   | numeric  |



## 4. 데이터 설명 및 탐색

### □ 데이터 분류

#### 1. 범주형 데이터 (12개 )

- 성별, 자동차 유무, 부동산 소유여부, 소득 분류, 교육 수준, 결혼 여부, 거주 형태, 핸드폰 소유 여부, 전화 소유 여부, 업무용 전화 소유 여부, 이메일 소유 여부, 직종

#### 2. 수치형 데이터 ( 6개 )

- 자녀 수, 연간 소득, 나이, 근속일 수, 가족 규모, 신용카드 이용 기간

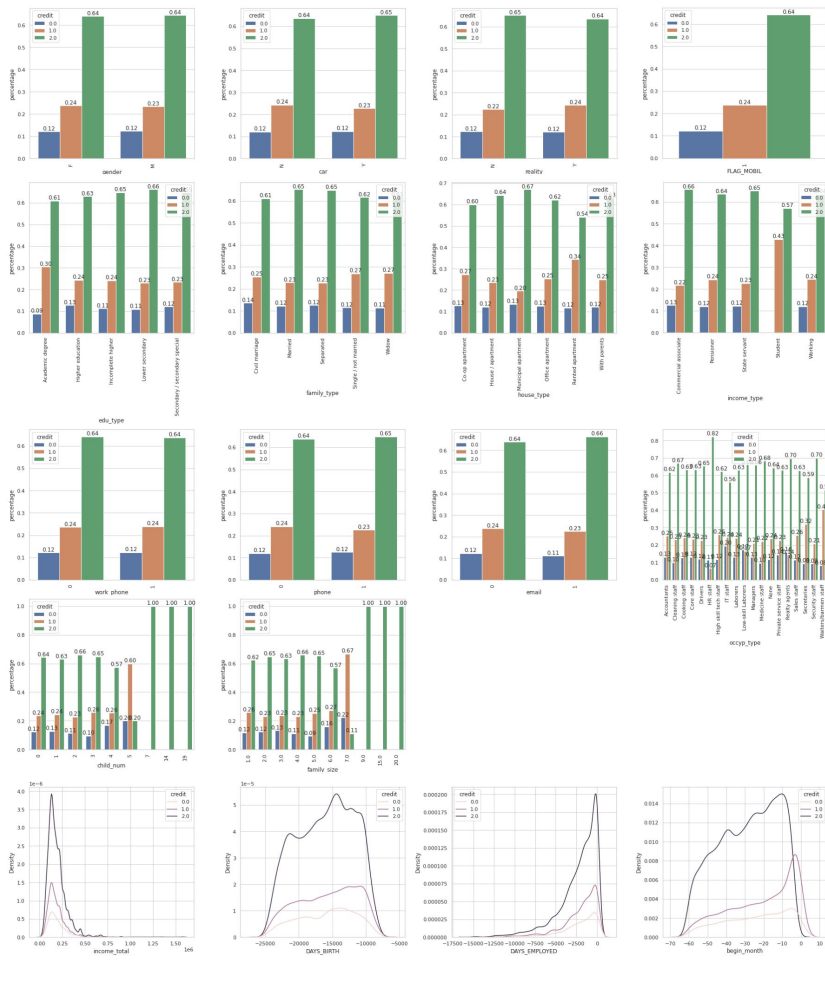
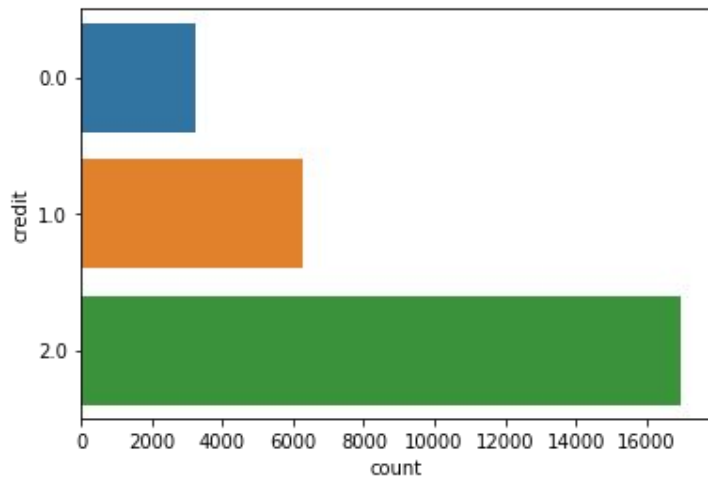
# 4. 데이터 설명 및 탐색

## □ 훈련 데이터 분포 시각화

- 레이블 별 데이터 불균형
- 각 특성 별 데이터 분포

```
# 결측치 처리
train['occyp_type'].fillna('None', inplace=True)

# 무직일 경우 근무 일수 0
train['DAYS_EMPLOYED'].replace(365243, 0, inplace=True)
```



## 4. 데이터 설명 및 탐색

### □ 훈련 데이터 결측치 확인

- 누락 데이터 : occyp\_type 특성 / 8,171개

```
raw_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 26457 entries, 0 to 26456  
Data columns (total 19 columns):  
#   Column              Non-Null Count  Dtype    
---  ---                
0   gender              26457 non-null  object   
1   car                  26457 non-null  object   
2   reality              26457 non-null  object   
3   child_num           26457 non-null  int64    
4   income_total        26457 non-null  float64  
5   income_type         26457 non-null  object   
6   edu_type            26457 non-null  object   
7   family_type         26457 non-null  object   
8   house_type          26457 non-null  object   
9   DAYS_BIRTH          26457 non-null  int64    
10  DAYS_EMPLOYED        26457 non-null  int64    
11  FLAG_MOBIL           26457 non-null  int64    
12  work_phone           26457 non-null  int64    
13  phone                26457 non-null  int64    
14  email                26457 non-null  int64    
15  occyp_type           18286 non-null  object   
16  family_size          26457 non-null  float64  
17  begin_month          26457 non-null  float64  
18  credit               26457 non-null  float64  
dtypes: float64(4), int64(7), object(8)  
memory usage: 4.0+ MB
```

```
raw_train.isnull().sum()
```

```
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
```

## 4. 데이터 설명 및 탐색

### □ 평가 데이터 결측치 확인

- 누락 데이터 : occyp\_type 특성 / 3,152개

```
[122] test.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10000 entries, 0 to 9999  
Data columns (total 19 columns):  
#   Column             Non-Null Count  Dtype  
---  ---  
0   index              10000 non-null  int64  
1   gender             10000 non-null  object  
2   car                10000 non-null  object  
3   reality            10000 non-null  object  
4   child_num          10000 non-null  int64  
5   income_total       10000 non-null  float64  
6   income_type        10000 non-null  object  
7   edu_type           10000 non-null  object  
8   family_type        10000 non-null  object  
9   house_type         10000 non-null  object  
10  DAYS_BIRTH         10000 non-null  int64  
11  DAYS_EMPLOYED      10000 non-null  int64  
12  FLAG_MOBIL         10000 non-null  int64  
13  work_phone         10000 non-null  int64  
14  phone              10000 non-null  int64  
15  email              10000 non-null  int64  
16  occyp_type          6848 non-null   object  
17  family_size         10000 non-null  float64  
18  begin_month         10000 non-null  float64  
dtypes: float64(3), int64(8), object(8)  
memory usage: 1.4+ MB
```

```
[121] test.isnull().sum()
```

```
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
```

## 4. 데이터 설명 및 탐색

### ❑ 중복 데이터 이슈

- Target을 포함한 모든 열의 데이터가 일치하는 경우
- 3,155개의 완전 중복 데이터
- 해석 : 사용자가 복수 개의 카드를 신청하여 같은 신용도를 얻음

```
train [ train.duplicated(train.columns, keep=False) ]
```

| 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 | begin_month | credit |
|-------|--------|-----|---------|-----------|--------------|----------------------|-------------------------------|----------------|---------------------|------------|---------------|------------|------------|-------|-------|-------------|-------------|-------------|--------|
| 19    | F      | N   | Y       | 0         | 180000.0     | Working              | Secondary / secondary special | Married        | House / apartment   | -13727     | -6031         | 1          | 0          | 0     | 0     | None        | 2.0         | -7.0        | 2.0    |
| 21    | F      | N   | N       | 0         | 157500.0     | Pensioner            | Secondary / secondary special | Married        | House / apartment   | -21253     | 0             | 1          | 0          | 1     | 0     | None        | 2.0         | -7.0        | 2.0    |
| 29    | F      | N   | Y       | 0         | 121500.0     | Commercial associate | Secondary / secondary special | Married        | Co-op apartment     | -12017     | -1711         | 1          | 0          | 1     | 0     | Sales staff | 2.0         | -22.0       | 0.0    |
| 48    | F      | N   | Y       | 0         | 99000.0      | Pensioner            | Secondary / secondary special | Married        | House / apartment   | -23585     | 0             | 1          | 0          | 0     | 0     | None        | 2.0         | -4.0        | 2.0    |
| 56    | F      | Y   | Y       | 0         | 130500.0     | Working              | Secondary / secondary special | Married        | House / apartment   | -16137     | -9391         | 1          | 0          | 1     | 0     | Laborers    | 2.0         | -29.0       | 2.0    |
| ...   | ...    | ... | ...     | ...       | ...          | ...                  | ...                           | ...            | ...                 | ...        | ...           | ...        | ...        | ...   | ...   | ...         | ...         | ...         | ...    |
| 26374 | F      | Y   | N       | 0         | 112500.0     | Working              | Secondary / secondary special | Married        | Municipal apartment | -17372     | -978          | 1          | 0          | 1     | 0     | Sales staff | 2.0         | -15.0       | 1.0    |
| 26393 | M      | Y   | Y       | 0         | 247500.0     | Working              | Secondary / secondary special | Married        | House / apartment   | -14122     | -3383         | 1          | 0          | 0     | 0     | Managers    | 2.0         | -31.0       | 2.0    |
| 26428 | F      | Y   | Y       | 2         | 270000.0     | Working              | Secondary / secondary special | Civil marriage | House / apartment   | -12745     | -525          | 1          | 0          | 0     | 1     | Core staff  | 4.0         | -23.0       | 1.0    |
| 26446 | F      | N   | Y       | 0         | 135000.0     | Working              | Secondary / secondary special | Civil marriage | House / apartment   | -16300     | -9698         | 1          | 0          | 0     | 1     | Managers    | 2.0         | -41.0       | 2.0    |
| 26451 | F      | N   | Y       | 0         | 202500.0     | Working              | Higher education              | Married        | House / apartment   | -12831     | -803          | 1          | 1          | 1     | 0     | Accountants | 2.0         | -44.0       | 1.0    |

3155 rows x 19 columns

## 4. 데이터 설명 및 탐색

### ❑ 중복 데이터 이슈

- Target을 제외한 모든 열의 데이터가 일치하는 경우
- 4,497개의 중복 데이터
- 해석 : 1342(=4497-3155)개의 경우는 사용자가 복수 개의 카드를 신청하여 다른 신용도를 얻은 결과임

```
[22] train.iloc[:, 1:-1]
train [ train.iloc[:, 1:-1].duplicated(train.iloc[:, 1:-1].columns, keep=False) ]
```

| 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 | begin_month | credit |
|-------|--------|-----|---------|-----------|--------------|----------------------|-------------------------------|----------------------|-------------------|------------|---------------|------------|------------|-------|-------|-------------|-------------|-------------|--------|
| 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         | -5.0        | 1.0    |
| 2     | M      | Y   | Y       | 0         | 450000.0     | Working              | Higher education              | Married              | House / apartment | -19087     | -4434         | 1          | 0          | 1     | 0     | Managers    | 2.0         | -22.0       | 2.0    |
| 19    | F      | N   | Y       | 0         | 180000.0     | Working              | Secondary / secondary special | Married              | House / apartment | -13727     | -6031         | 1          | 0          | 0     | 0     | None        | 2.0         | -7.0        | 2.0    |
| 21    | F      | N   | N       | 0         | 157500.0     | Pensioner            | Secondary / secondary special | Married              | House / apartment | -21253     | 0             | 1          | 0          | 1     | 0     | None        | 2.0         | -7.0        | 2.0    |
| 24    | F      | N   | N       | 0         | 202500.0     | Pensioner            | Secondary / secondary special | Single / not married | House / apartment | -22361     | 0             | 1          | 0          | 1     | 0     | None        | 1.0         | -5.0        | 2.0    |
| ...   | ...    | ... | ...     | ...       | ...          | ...                  | ...                           | ...                  | ...               | ...        | ...           | ...        | ...        | ...   | ...   | ...         | ...         | ...         | ...    |
| 26430 | F      | N   | Y       | 0         | 112500.0     | Working              | Incomplete higher             | Civil marriage       | House / apartment | -9301      | -1751         | 1          | 0          | 0     | 0     | None        | 2.0         | -19.0       | 2.0    |
| 26431 | F      | N   | Y       | 0         | 225000.0     | Pensioner            | Secondary / secondary special | Widow                | House / apartment | -21151     | 0             | 1          | 0          | 0     | 1     | None        | 1.0         | -60.0       | 1.0    |
| 26432 | F      | Y   | Y       | 0         | 72000.0      | Pensioner            | Secondary / secondary special | Married              | House / apartment | -22314     | 0             | 1          | 0          | 1     | 0     | None        | 2.0         | -17.0       | 1.0    |
| 26446 | F      | N   | Y       | 0         | 135000.0     | Working              | Secondary / secondary special | Civil marriage       | House / apartment | -16300     | -9698         | 1          | 0          | 0     | 1     | Managers    | 2.0         | -41.0       | 2.0    |
| 26451 | F      | N   | Y       | 0         | 202500.0     | Working              | Higher education              | Married              | House / apartment | -12831     | -803          | 1          | 1          | 1     | 0     | Accountants | 2.0         | -44.0       | 1.0    |

4497 rows x 19 columns

## 4. 데이터 설명 및 탐색

### ❑ 중복 데이터 이슈

- Target과 begin\_month 특성을 제외한 데이터가 일치하는 경우
- 23,208 개의 중복 데이터
- 해석 : 카드를 두 개 이상 발급한 이용자의 모든 기록 (신용카드 발급 기간에 따라 신용도가 측정된 결과가 포함된 것임)

```
[23] train.iloc[:, 1:-2]  
train [ train.iloc[:, 1:-2].duplicated(train.iloc[:, 1:-2].columns, keep=False) ]
```

|       | 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 | begin_month | credit |
|-------|--------|-----|---------|-----------|--------------|----------------------|-------------------------------|----------------|---------------------|------------|---------------|------------|------------|-------|-------|-------------|-------------|-------------|--------|
| index |        |     |         |           |              |                      |                               |                |                     |            |               |            |            |       |       |             |             |             |        |
| 0     | F      | N   | N       | 0         | 202500.0     | Commercial associate | Higher education              | Married        | Municipal apartment | -13899     | -4709         | 1          | 0          | 0     | 0     | None        | 2.0         | -6.0        | 1.0    |
| 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         | -5.0        | 1.0    |
| 2     | M      | Y   | Y       | 0         | 450000.0     | Working              | Higher education              | Married        | House / apartment   | -19087     | -4434         | 1          | 0          | 1     | 0     | Managers    | 2.0         | -22.0       | 2.0    |
| 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         | -37.0       | 0.0    |
| 6     | F      | N   | N       | 0         | 315000.0     | Working              | Secondary / secondary special | Separated      | House / apartment   | -17570     | -1978         | 1          | 0          | 0     | 1     | Core staff  | 1.0         | -41.0       | 2.0    |
| ...   | ...    | ... | ...     | ...       | ...          | ...                  | ...                           | ...            | ...                 | ...        | ...           | ...        | ...        | ...   | ...   | ...         | ...         | ...         | ...    |
| 26447 | M      | N   | Y       | 2         | 99000.0      | Working              | Secondary / secondary special | Married        | House / apartment   | -14226     | -1026         | 1          | 1          | 1     | 0     | Laborers    | 4.0         | -43.0       | 2.0    |
| 26448 | M      | N   | Y       | 0         | 292500.0     | Commercial associate | Higher education              | Married        | House / apartment   | -16280     | -887          | 1          | 0          | 0     | 0     | Laborers    | 2.0         | -23.0       | 0.0    |
| 26449 | F      | N   | N       | 0         | 90000.0      | Working              | Secondary / secondary special | Married        | House / apartment   | -10498     | -2418         | 1          | 1          | 1     | 0     | None        | 2.0         | -2.0        | 1.0    |
| 26451 | F      | N   | Y       | 0         | 202500.0     | Working              | Higher education              | Married        | House / apartment   | -12831     | -803          | 1          | 1          | 1     | 0     | Accountants | 2.0         | -44.0       | 1.0    |
| 26452 | F      | N   | N       | 2         | 225000.0     | State servant        | Secondary / secondary special | Married        | House / apartment   | -12079     | -1984         | 1          | 0          | 0     | 0     | Core staff  | 4.0         | -2.0        | 1.0    |

23208 rows x 19 columns

## 4. 데이터 설명 및 탐색

❑ 중복 데이터 이슈 요약 : **Target**을 포함한 모든 열의 데이터가 일치하는 경우 발견함

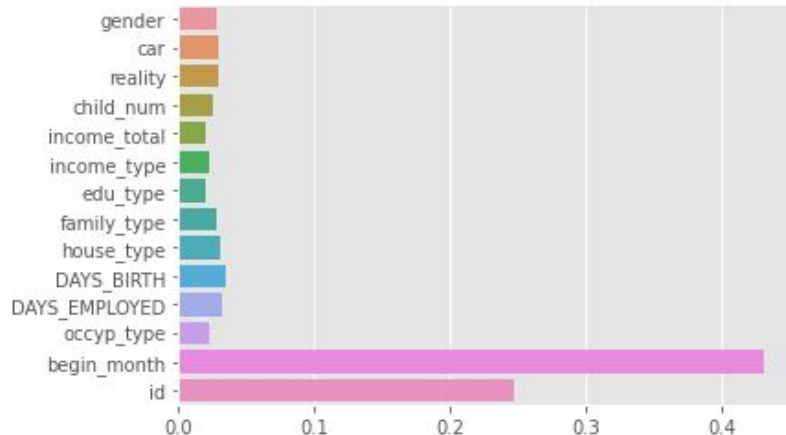
모든 값이 일치하거나, 2) 카드 발급 월을 제외한 경우, 또는

3) 카드 발급 월과 신용도를 제외한 경우에 값이 일치하는 **row**가 상당히 많음

❑ 이로 인해 모델이 학습할 때 카드 발급 월, 중복 여부에 큰 영향을 받음

```
[ ] train_dup = train[train.duplicated(['gender', 'income_total', 'income_type', 'edu_type', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'occyp_type', 'child_num',  
                                     'family_size', 'car', 'reality', 'family_type', 'house_type', 'work_phone', 'phone', 'email', 'FLAG_MOBIL'], keep=False))  
  
train_not_dup = train.drop_duplicates(['gender', 'income_total', 'income_type', 'edu_type', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'occyp_type', 'child_num',  
                                     'family_size', 'car', 'reality', 'family_type', 'house_type', 'work_phone', 'phone', 'email', 'FLAG_MOBIL'], keep=False)  
  
train_dup_group = train_dup.groupby('DAYS_BIRTH').groups  
  
print("전체 행 수:", len(train))  
print("중복 행 수:", len(train_dup))  
print("중복 아닌 사람 수:", len(train_not_dup))  
print("중복인 사람 수:", len(train_dup_group))  
print("전체 사람 수", len(train)-len(train_dup)+len(train_dup_group))  
  
train_dup['id'] = '-1'  
train_not_dup['id'] = '-1'
```

전체 행 수: 26457  
중복 행 수: 23208  
중복 아닌 사람 수: 3249  
중복인 사람 수: 4617  
전체 사람 수 7866

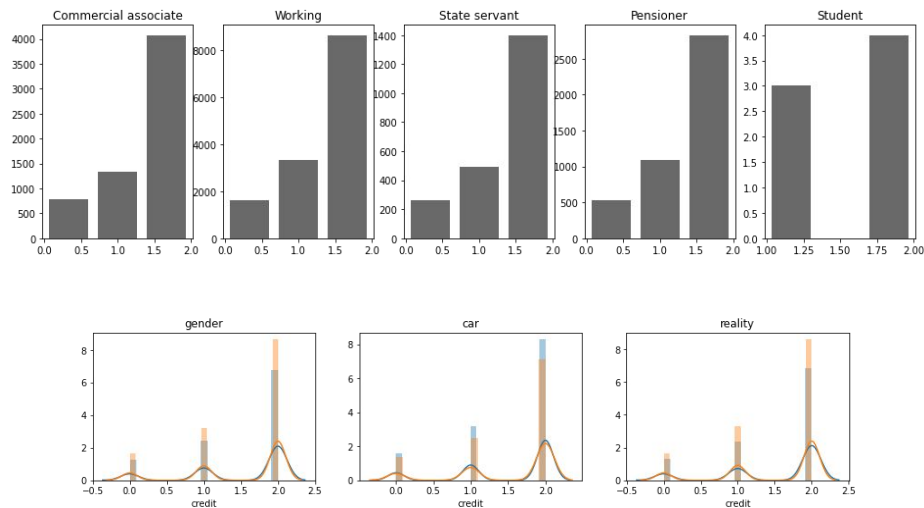
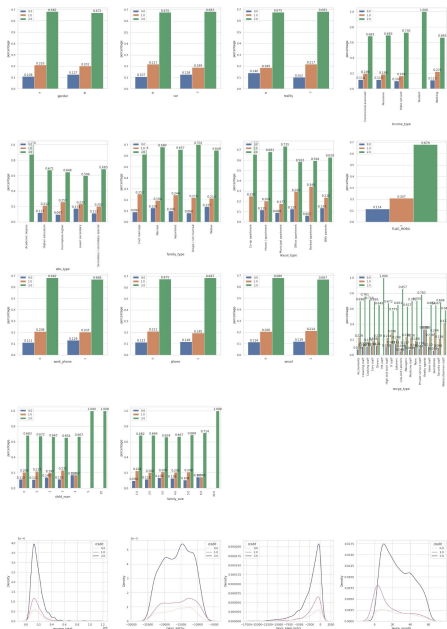




## 4. 데이터 설명 및 탐색

### □ 불균형 데이터 문제 및 전체적으로 고른 신용도 분포

대부분의 경우에, 변수와 신용도 간의 분포를 살펴보면 큰 차이를 보이지 않음

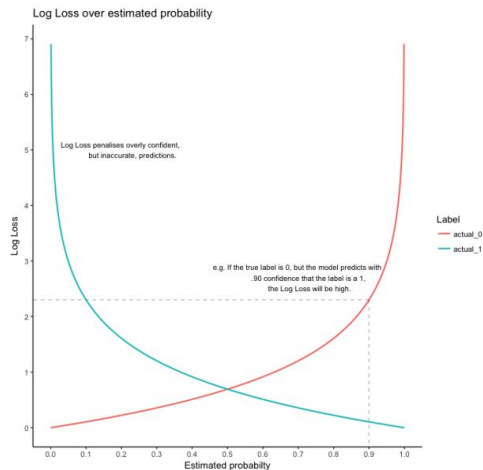


→ 연속형 변수의 경우 왜도를 줄이기 위한 분포 변환 및 비닝(범주화)를 진행함

## 4. 데이터 설명 및 탐색

### □ 평가 지표 : logloss

오분류 확률에 높은 패널티를 부여함



mlogloss for multi-classification

$$- L = - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{i,j} \log(p_{i,j})$$

Where,

$N$  No of Rows in Test set

$M$  No of Fault Delivery Classes

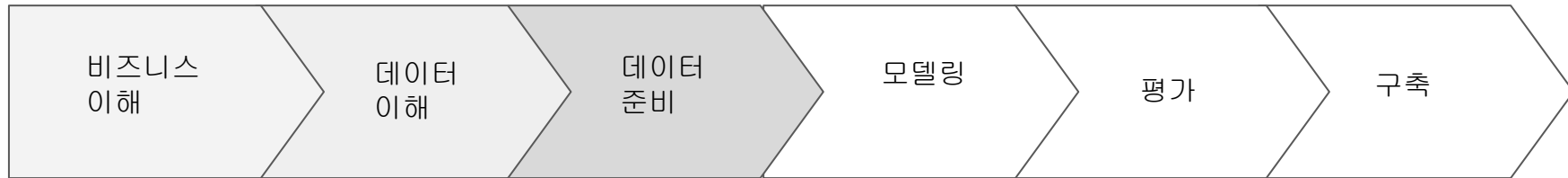
$Y_{ij}$  1 if observation belongs to Class  $j$ ; else 0

$p_{ij}$  Predicted Probability that observation belong to Class  $j$

※ 불균형 데이터 문제 및 신용평가 모델 특성을 반영하여, 모델 앙상블 시 F1-score 등을 추가 고려함

## 5. 프로젝트 진행사항

대회 기간 : 2021.04.05 ~ 2021.05.24



도메인 지식 습득  
목표 설정  
프로젝트 계획 설정

데이터 수집  
데이터 탐색

데이터 전처리  
**feature engineering**  
데이터 파티셔닝

**feature selection**  
알고리즘 구현  
하이퍼 파라미터  
튜닝  
양상블

제출 결과 평가  
검토 및 개선

최종 결과 제출  
최종 발표자료 작성









- ❑ 경진대회 참가가 가능한 인원수가 최대 3명이라 내부적으로 팀을 나눔
- ❑ 개별적으로 전처리 및 피쳐 엔지니어링을 진행하면서 결과에 대한 토론을 진행 중임

### ❑ 발견한 인사이트

- 중복 데이터 문제 → 카드 발급 일수를 제외한 같은 데이터로 인해 카드 발급 일수에 큰 영향
- 불균형 데이터 문제
- 범주가 너무 세분화 되어 있는 경우 또는 희소 범주는 병합하는 것이 좋음
- 출생일, 업무 시작일 등의 수치형 변수는 비닝 처리하는 것이 좋음
- 전화 관련 변수는 분석에 무의미한 피쳐
- 고객id, 보유 카드 수 등의 피쳐 추가

# 5. 프로젝트 진행사항

- 경진대회 참가가 가능한 인원수가 최대 3명이라 내부적으로 팀을 나눔
- 개별적으로 전처리 및 피쳐 엔지니어링을 진행하면서 결과에 대한 토론을 진행 중임

| Public Ranking Chart 순위기준 |          |   |         |     |        |
|---------------------------|----------|---|---------|-----|--------|
| ● WINNER ● 1% ● 4% ● 10%  |          |   |         |     |        |
| #                         | 팀        | 팀 멤버  | 점수      | 제출수 | 등록일    |
| 319                       | 남탕       |    | 0.80408 | 11  | 6시간 전  |
| 1                         | 렛서렌터     |    | 0.66934 | 25  | 13일 전  |
| 2                         | TYKIM    |    | 0.67115 | 18  | 16일 전  |
| 3                         | 초보산님     |    | 0.6758  | 15  | 3시간 전  |
| 4                         | 안녕하세요선생님 |    | 0.67604 | 29  | 7일 전   |
| 5                         | 진니       |    | 0.67943 | 19  | 15시간 전 |

| 디스커션                           |               |  |  |  |  |
|--------------------------------|---------------|--|--|--|--|
| 파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 |               |  |  |  |  |
| 이동성님이 24분 전에 마지막으로 수정했습니다.     |               |  |  |  |  |
| 1118                           |               |  |  |  |  |
| A                              | B             | C  | D  | E  | F  |
| 58                             |               |  |  |  |  |
| 59                             | 설명            | 유형   | 가능한 처리 방법  |  |  |
| 60                             | index         | 같은 사용자도 index는 다를 수 있음 (같은 사람이 복수 개의 카드를 빌균) |  |  |  |
| 61                             | gender        | 성별   | gender, car, reality 조합 파생변수 추가 or 삭제                                      |  |  |
| 62                             | car           | 차량 소유 여부                                     |  |  |  |
| 63                             | reality       | 부동산 소유 여부                                    |  |  |  |
| 64                             | child_num     | 자녀 수   | 이상치 이슈(재거, 병합 기준 - 3이상을 같은 범주), family_size와 높은상관관계(물 중 하나 재거 고려)->성능은 좋아짐 | 삭제 19,15   | 이상치family_size < cl [5825, 14800, 16110,   |
| 65                             | income_total  | 연간 소득  | num  | 4분위  |  |
| 66                             | income_type   | 소득 분류  | category   | working, commercial 이외 회소, student 극회소                                   |  |
| 67                             | edu_type      | 교육 수준  | category   | 서열적도달음, secondary, higher 이외 회소  | ordinal 레이블로 변환  |
| 68                             | family_type   | 결혼 여부  | category   | married 이외 범주 회소 - 재거, 병합 등  | 결혼 여부에 따라 cred   |
| 69                             | house_type    | 생활 방식  | category   |  |  |
| 70                             | DAYS_BIRTH    | 출생일  | num  | 비닝   |  |
| 71                             | DAYS_EMPLOYED | 업무 시작일                                       | num  | 양수들 0으로 처리, 비닝, 분포 변환(sqrt)  |  |
| 72                             | FLAG_MOBIL    | 핸드폰 소유 여부                                    | binary   | 삭제   |  |
| 73                             |               |  |  | oocyp과 비교 후 oocyp_type의 결측값 처리 시 참고, work_phone, phone, email은 삭제하는 것 고려 |  |
| 74                             | work_phone    | 업무용 전화 소유 여부                                 | binary   |  | 삭제   |
| 75                             | phone         | 전화 소유 여부                                     | binary   |  | 삭제   |
| 76                             | email         | 이메일 소유 여부                                    | binary   |  | none 값 대체  |
|                                |               |  |  |  | oocyp_type이 기입된 데이터 -> train oocyp_type 이 없고, DAYS_EMPLOYED가 0이 아닌 데이터. 학습에 이용한 특성 : 'gender', 'reality', 'income_type', 'edu_ty |

## 5. 프로젝트 계획

대회 기간 : 2021.04.05 ~ 2021.05.24

|            | 비즈니스 이해 | 데이터 이해 | 데이터 준비 | 모델링 | 평가 | 구축 |
|------------|---------|--------|--------|-----|----|----|
| 5월 첫째 주    |         |        |        |     |    |    |
| 5월 둘째 주    |         |        |        |     |    |    |
| 5월 셋째 주    |         |        |        |     |    |    |
| 5월 넷/다섯째 주 |         |        |        |     |    |    |

### □ 추가 반영 계획

- 중복 데이터 문제 → 중복 및 비중복 데이터를 위한 2개의 별도 모델 구현 및 통합 모델 구현
- **feature engineering** (범주형 변수 처리 및 인코딩 등)
- 모델링 개선

감사합니다.

TEAM3