ML Competition

# 신용카드 사용자 연체 예측


**아머러파티**

김지은 노명진 심재민 이서연 임형빈

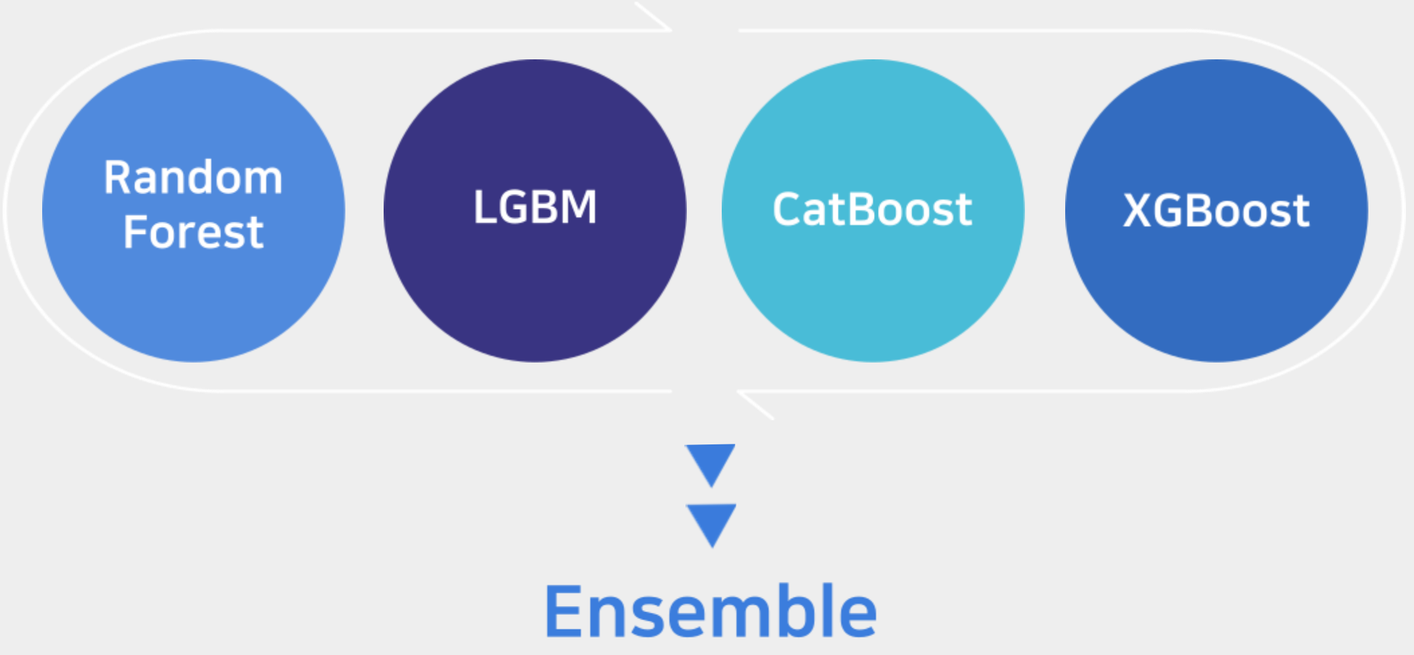

# CONTENTS

# 00 분석 주제 & 전략

신용카드 사용자들의
개인 신상정보 데이터

→

신용카드 사용자들의
대금 연체 정도 예측



# 01 Data Cleansing

· 범주형 변수

index
(인덱스 번호)

FLAG_MOBIL
(핸드폰 소유 여부)

▶▶ column 삭제

occyp_type
(직업 유형)

결측치 존재

DAYS_EMPLOYED > 0
-> 고용되지 않은 상태

▶▶ 'unemployed'

남은 결측치

▶▶ 'NaN'

# 01 Data Cleansing

· 수치형 변수

DAYS_BIRTH
(출생일)

begin_month
(신용카드 발급 월)

▶▶ 음수 -> 양수 변환

DAYS_BIRTH
(출생일)

▶▶ 365243 -> 0

음수 -> 양수 변환

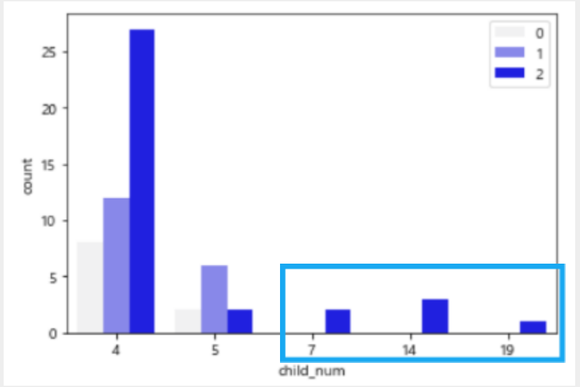
income_total
(연간 소득)

▶▶ 표준정규분포로 변환 후
-3 이하 및 3 이상인 값들 대체

# 01 Data Cleansing

## · 수치형 변수

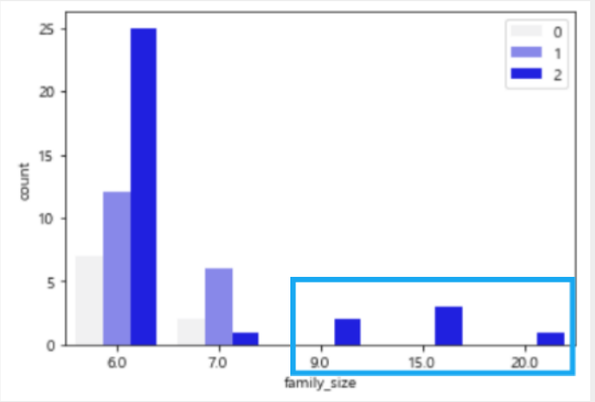
child_num
(자녀 수)




child_num이 7 이상은
target 값이 전부 2

▶▶ 7 이상은 7로 대체

family_size
(가족 규모)



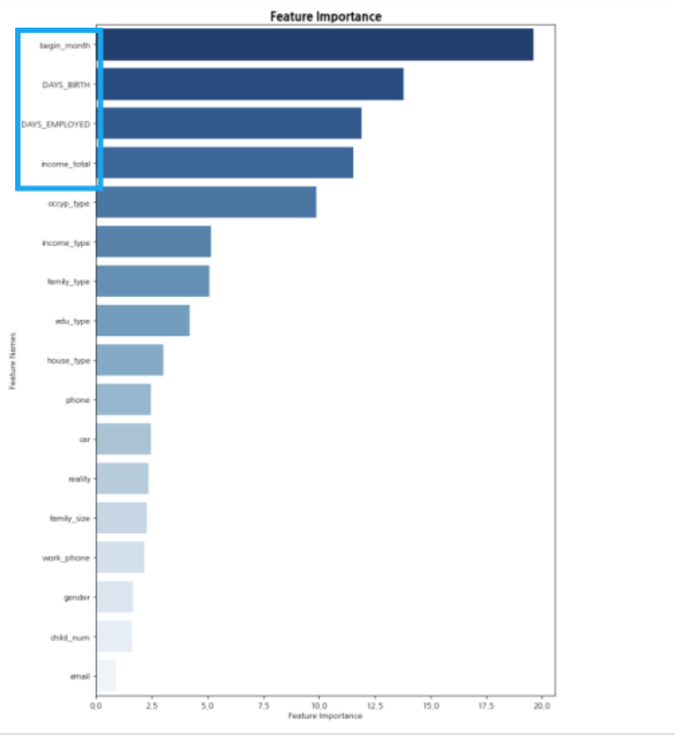

family_size가 9 이상은
target 값이 전부 2

▶▶ 9 이상은 9로 대체

# 02 Feature Engineering

## · feature 생성

- begin month를 제외한 모든 값이 중복되는 데이터 존재 ▶▶ 고객별 고유 ID




**income_total**

월 소득
총 소득
가구별 인당 평균 소득

**begin_month**

주 단위
년 단위
카드 발급 나이
고용된 후 몇 년 후에 발급 받았는지
10개월 기준 구간화

**DAYS_EMPLOYED**

주 단위
월 단위
년 단위
고용 여부
고용되기까지 걸린 날의 수
카드 사용 중 소득 변화 여부
소득 변화가 있는 사람의 월 소득

**DAYS_BIRTH**

주 단위
월 단위
년 단위
20대-60대 범주화

# 02 Feature Engineering

· feature 생성

수치형 변수 → · 사분위수로 범주화
· 수치형 변수끼리의 곱

범주형 변수 → · clustering
· groupby
-> 수치형 변수들의 평균, 합계, 표준편차

**총 102개의 feature 생성**

# 02 Feature Engineering

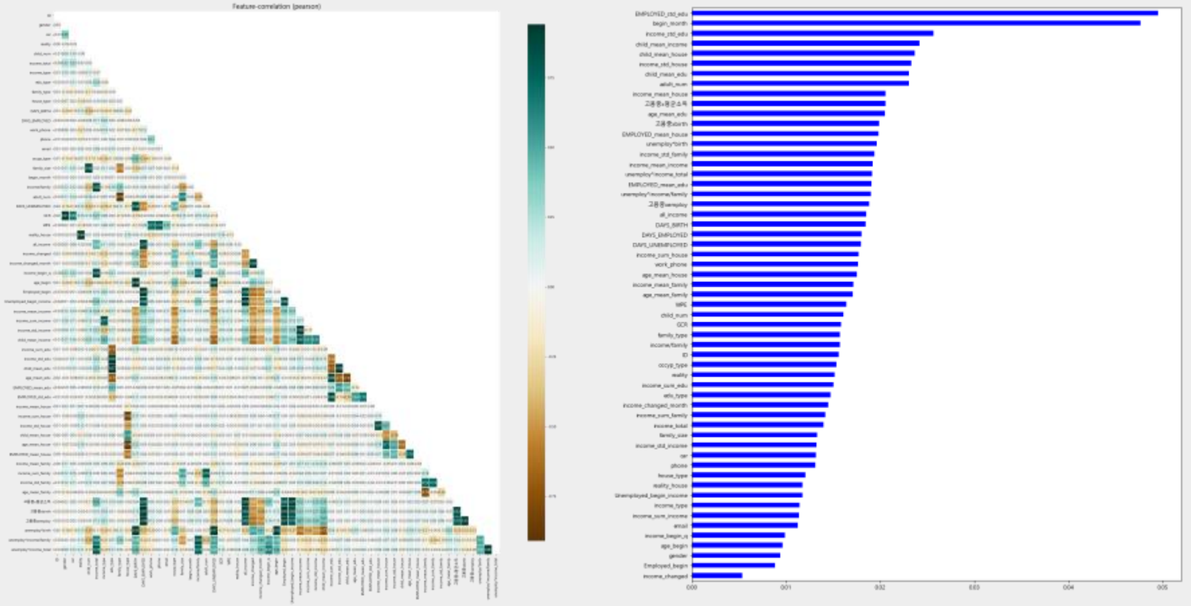
· feature selection

## 각 모델별로 적합한 feature 선택

- **Embedded Method를 활용한 Select From Model 사용**

```
{'ID': True,
 'gender': False,
 'car': False,
 'reality': False,
 'child_num': False,
 'income_total': True,
 'income_type': True,
 'edu_type': True,
 'family_type': True,
 'house_type': False,
 'DAYS_BIRTH': True,
 'DAYS_EMPLOYED': False,
 'work_phone': False,
 'phone': False,
 'email': False,
 'occyp_type': True,
 'family_size': False,
 'begin_month': True,
 'income/family': True,
 'adult_num': False,
 'age': False,
 'EMPLOYED': False,
 'MONTH_EMPLOYED': True,
 'YEAR_EMPLOYED': False,
 'YEAR_EMPLOYED_s': True
```

- **Feature간의 상관관계가 높은 것 중 feature importance가 더 작은 변수 삭제**



# 03 Scaling & Encoding

## · Scaling

| | child_num | income_total | DAYS_BIRTH | DAYS_EMPLOYED | family_size | begin_month | income/family | adult_num | age | MONTH_EMPLC |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 26457.000000 | 26457.000000 | 26457.000000 | 26457.000000 | 26457.000000 | 26457.000000 | 26457.000000 | 26457.000000 | 26457.000000 | 26457.00 |
| mean | 0.427411 | 184830.586410 | 15958.053899 | 2198.529538 | 2.195752 | 26.123294 | 62178.848922 | 1.768341 | 43.213478 | 72.88 |
| std | 0.728768 | 89387.516632 | 4201.589022 | 2370.140530 | 0.903769 | 16.559550 | 35043.384019 | 0.422616 | 11.513590 | 78.93 |
| min | 0.000000 | 27000.000000 | 7705.000000 | 0.000000 | 1.000000 | -0.000000 | 6300.000000 | -1.000000 | 21.000000 | 0.00 |
| 25% | 0.000000 | 121500.000000 | 12446.000000 | 407.000000 | 2.000000 | 12.000000 | 37500.000000 | 2.000000 | 34.000000 | 13.00 |
| 50% | 0.000000 | 157500.000000 | 15547.000000 | 1539.000000 | 2.000000 | 24.000000 | 52500.000000 | 2.000000 | 42.000000 | 51.00 |
| 75% | 1.000000 | 225000.000000 | 19431.000000 | 3153.000000 | 3.000000 | 39.000000 | 78750.000000 | 2.000000 | 53.000000 | 105.00 |
| max | 7.000000 | 492935.852351 | 25152.000000 | 15713.000000 | 9.000000 | 60.000000 | 246467.926175 | 2.000000 | 68.000000 | 523.00 |

수치형 변수들 간의 값 범위의 차이가 큼

▼

MinMax Scaler

일정한 범위로 조정

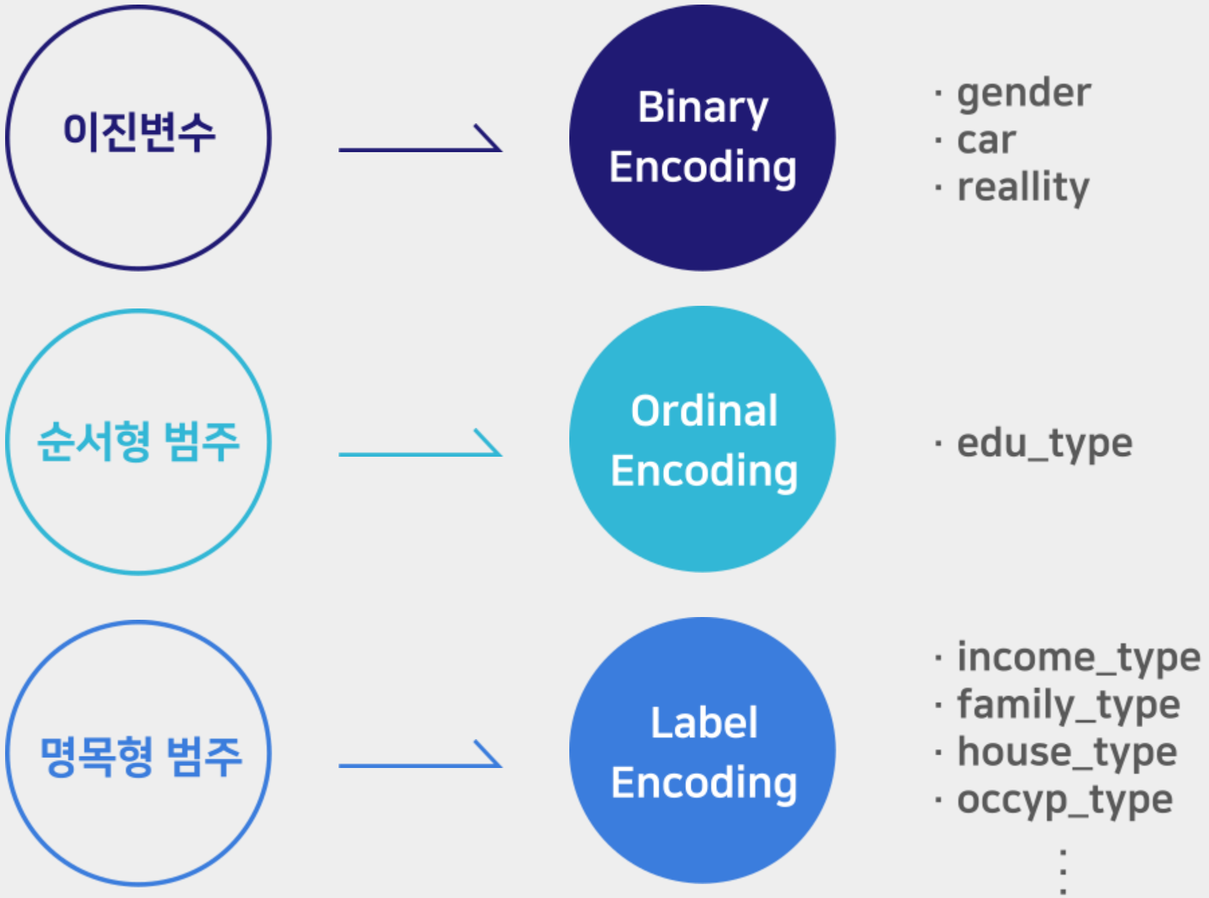
▼

로그 변환

정규분포에 가까운 형태

# 03 Scaling & Encoding

· Encoding



# 04 Modeling

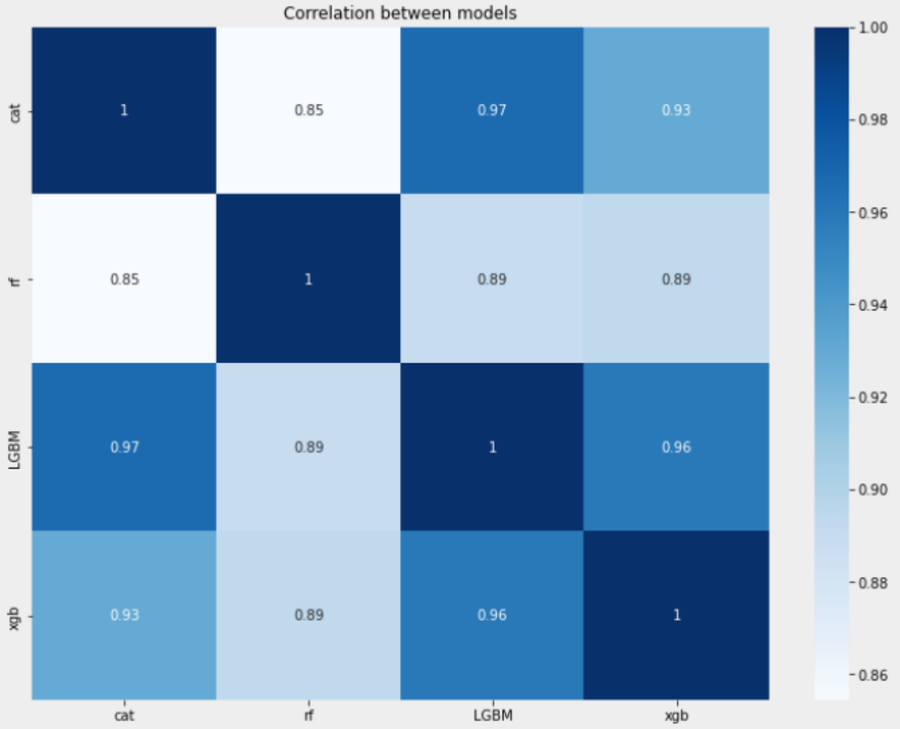
· Random Forest, LGBM, CatBoost, XGBoost

optuna → 하이퍼파라미터 최적화

| Random Forest | LGBM | CatBoost | XGBoost |
|---|---|---|---|
| public: 0.7439<br>private: 0.7267 | public: 0.7368<br>private: 0.7207 | public: 0.7293<br>private: 0.7135 | public: 0.7016<br>private: 0.6842 |

# 04 Modeling

· Ensemble




**상관계수가 낮은 모델들 선택**

**-> RandomForest, CatBoost, XGBoost**

# 04 Modeling

· Ensemble

가중평균

## Submission Ensemble


public: 0.7016
private: 0.6854

# THANK YOU

ML Competition

아머러파티

2022. 11. 29