1. 결측값 처리
2. 결측값 판단

* 데이터가 불확실하므로 변수간의 관계를 통한 추론을 포기하고 간단한 전처리 시행
* 대부분의 데이터는 numeric 처리하였음
* Numeric의 경우 NULL은 0으로 처리
* Factor형의 경우 NULL을 하나의 범주로 처리

1. 비식별 처리

* 비식별은 특수한 경우의 값으로 가정하고 하나의 범주로 가정

1. 이상치 처리

* 고려하지 않음

1. 자료 분포

* 변수 분포 미확인

1. 가정 위반 여부

* Logistic과 LASSO의 별도의 가정 확인하지 않음

1. 범주 축소 / 변수 변환

* CUST\_ID 제거
* Factor : TARGET, OCCP\_NAME\_G, MATE\_OCCP\_NAME\_G, MIN\_CNTT\_DATE(년도별로 범주 축소), STRT\_CRDT\_GRAD, LTST\_CRDT\_GRAD, AGE, SEX, TEL\_MBSP\_GRAD, CBPT\_MBSP\_YN, PAYM\_METD, LINE\_STUS
* Numeric : 그 외 모두 numeric

1. 샘플링

* Rare Case와 동일한 비율로 샘플링하여 Target을 1:1로 맞춰주는 Oversampling 활용하여 Train set 생성 (3001 + 3001 = 6002)
* Training에 쓰인 데이터를 제외하고 모집단의 타겟 비율(약4%정도)에 맞춰 Validation set 생성 (28770 + 1286 = 30056)

1. 변수 선택

* LASSO를 통한 자동 선택

1. 모델링

* LASSO를 통해 beta를 추정한 Binomial logistic
* Grid 10의 -5승과 5승 사이의 랜덤한 lambda를 통해 최적의 lambda도출
* 최적 lambda 선정 기준으로 mae보다 auc가 오분류율이 높아 auc사용
* logistic 예측값 0.83 이상을 연체로 판단하는 기준으로 삼을 때 F-measure이 0.377로 가장 높았으나, 실제 연체자 중, 많은 연체자를 미연체자로 분류하므로 F-measure를 0.5~0.83사이로 잘 정하는 것이 필요.