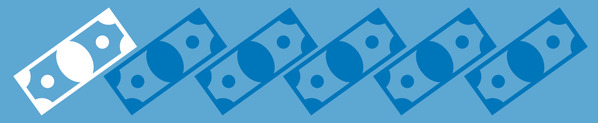
BIG CONTEST 2017

Loan Repayment Forecast

Sohyun Jeon

https://github.com/SohyunJeon
http://sherry-data.tistory.com/



대출의 부실..

금융권에 있어 가장 중요한 문제

부실은 과도한 충당금의 유발로

건전성에 영향

비대면 채널의 확대 및

금융 상품의 다양성

심사에 있어 비금융 정보의 중요성 부각

Target 비율의 비대칭성과 다양한 결손값으로 예측이 어려움

두 가지 이슈가 주요 해결 과제

Target 비율

	TARGET	per
0	95946	0.95723
1	4287	0.04277

매우 불균형한 데이터임이 확인된다

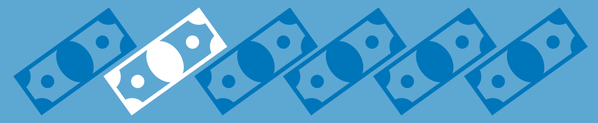
결손값 개수

OCCP_NAME_G	464
LAST_CHLD_AGE	1027
MATE_OCCP_NAME_G	45709
TEL_MBSP_GRAD	46015
PAYM_METD	2833

반 이상이 결손값인 변수도 존재한다

'*'의 존재 (비식별)

OCCP_NAME_G	0	per	
0	*	1189	1.186236
1	1차산업 종사자	1178	1.175262
2	2차산업 종사자	9601	9.578682
3	3차산업 종사자	8275	8.255764
4	고소득 전문직	1223	1.220157
5	공무원	5091	5.079166
6	기업/단체 임원	1041	1.038580
7	기타	1672	1.668113
8	단순 노무직	821	0.819092
9	단순 사무직	4107	4.097453
10	사무직	16581	16.542456
11	예체능계 종사자	936	0.933824
12	운전직	2126	2.121058
13	자영업	9485	9.462951
14	전문직	5043	5.031277
15	주부	27565	27.500923
16	학생	3835	3.826085



OCCP_NAME_G, MATE_OCCP_NAME_G

- 직업란에 '기타'가 따로 있는데 결손값이 생긴 이유는 무엇일까
- 결손값의 연체율은 전체의 평균에 비해 낮고, 연령도 중년인 것을 보아 불로소득을 추정 해 본다
- '기타'는 직업에 기반한 추정소득이 0이고 신용대출 연체율이 전체 집단보다 많이 높은 것을 보아 '무직'이라고 추정한다

변수처리) '*'와 null 을 한 집단으로 묶는다 -> '비식별'이라는 값 별도 생성

```
1 check_data.loc[check_data['OCCP_NAME_G']=='기타'].describe()
```

	CUST_JOB_INCM	CRLN_OVDU_RATE
count	1672.0	1672.000000
mean	0.0	4.131579
std	0.0	15.232471
min	0.0	0.000000
25%	0.0	0.000000
50%	0.0	0.000000
75%	0.0	0.000000
max	0.0	99.000000

```
1 check_data.loc[check_data['OCCP_NAME_G']=='*'].describe()
```

	CUST_JOB_INCM	CRLN_OVDU_RATE
count	1189.000000	1189.000000
mean	3236.417157	1.983179
std	2875.961724	11.530141
min	0.000000	0.000000
25%	0.000000	0.000000
50%	3900.000000	0.000000
75%	5100.000000	0.000000
max	10000.000000	92.000000

LAST_CHLD_AGE

- 중요한 변수로 파악되지 않는다

변수처리) 0으로 값을 채운다

```
nan 의 개수는 1027
0.0     50125
24.0    10774
19.0     9651
29.0     7905
14.0     5801
34.0     5600
39.0     4297
9.0      2862
44.0     1221
4.0       544
49.0      330
54.0       46
60.0       33
59.0       17
Name: LAST_CHLD_AGE, dtype: int64
```

PAYM_METD

- 카테고리 변수는 추정할 수 없다

변수처리) N으로 치환한다

TEL_MBSP_GRAD

- SKT멤버쉽은 4단계로 나누어져 있는데 반 정도의 결손값은 일반 등급이거나,
- 현재 아직 등급이 생성되지 않았다고 판단된다

변수처리) 새로운 값 N으로 값을 채운다

AGE

- 결손값은 평균으로 채울 수 있다

변수처리) *을 제외한 나이의 평균인 46으로 채운다

SEX

- 추정하기 위해 직업, 배우자직업, 연봉을 함께 확인한다
- 전체 소득의 중앙값보다 적으면 여자, 많으면 남자로 구분한다
- 대신 추정소득 3,600만원 이하인 사람 중 배우자의 직업이 주부인 사람은 남자로 구분한다

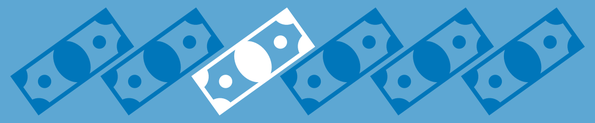
변수처리) 상기 조건으로 값을 치환한다

```
추정 소득의 기술통계량
:count     100233.000000
mean         2788.233416
std          2472.287102
min             0.000000
25%             0.000000
50%          3600.000000
75%          4700.000000
max         10000.000000
Name: CUST_JOB_INCM, dtype: float64
```

MIN_CNTT_DATE, TEL_CNTT_QTR

- 일반적으로 신용기간및 거래 기간은 등급 산정에 영향을 미친다
- 최초 신용대출 일자 및 통신 가입일자와 현재의 일자를 월단위로 변환한다

변수처리) 두 값의 차이로 월단위 기간으로 대체한다



단위는 '원'으로 통일 후 로그변환 취한 변수를 추가한다

- 오른쪽으로 쏠린 데이터가 많아 로그 변환을 취할 시 정규 분포에 가까워 진다
- 천원: 'TOT_LNIF_AMT', 'TOT_CLIF_AMT', 'BNK_LNIF_AMT', 'CPT_LNIF_AMT', 'CB_GUIF_AMT'
- 만원 : 'CUST_JOB_INCM', 'HSHD_INFR_INCM','MATE_JOB_INCM'

변수처리) '원' 변환 변수 + 로그변환 변수

OCCP_NAME_G

MATE_OCCP_NAME_G

SEX

TEL_MBSP_GRAD

CBPT_MBSP_YN

PAYM_METD

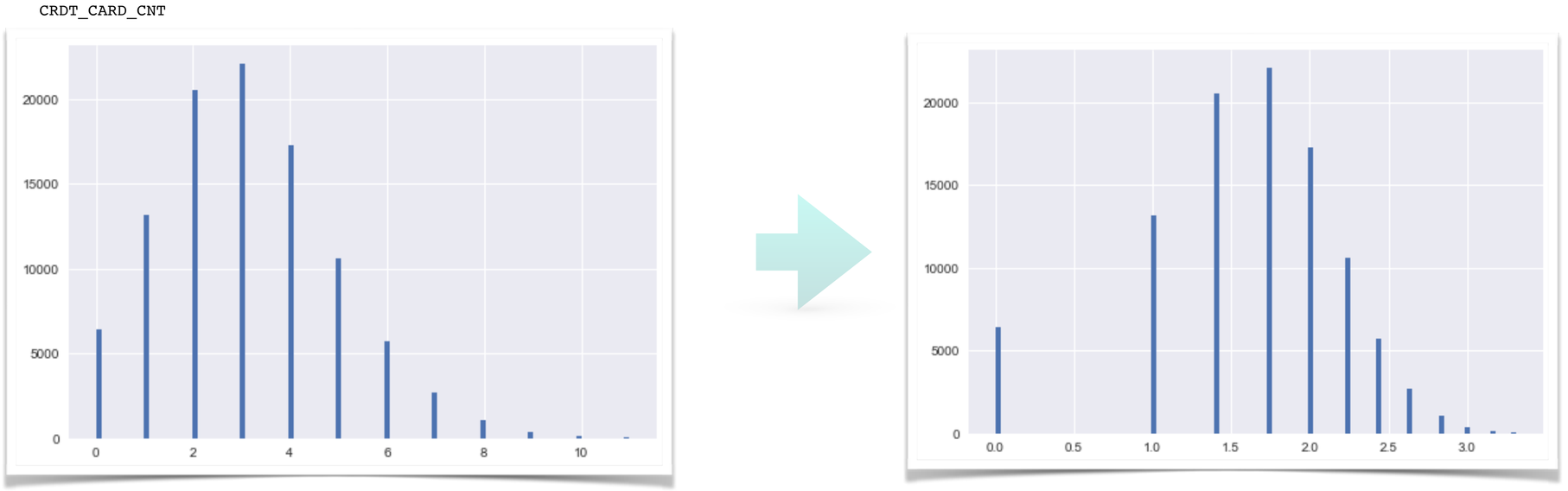
LINE_STUS

변수처리) 모든 질적 변수는 dummy처리한다

CRDT_CARD_CNT

- 건수 관련 변수 중 CRDT_CARD_CNT 는 로그 변환 시 정규 분포가 된다

변수처리) 로그변환 값을 변수로 추가한다



‘housing_prop’

- 담보대출의 비율

변수처리) CB_GUIF_AMT / TOT_LNIF_AMT

‘bank_grade’

- 은행권 대출금액으로 추정하는 은행권 내의 본인 등급

변수처리) BNK_LNIF_AMT / CUST_JOB_INCM

‘loan_ver_earn’

- 소득대비 대출금액

변수처리) TOT_LNIF_AMT / HSHD_INFR_INCM

‘complete_ins’

- 만기완납 한 보험 대비해서, 실효된 보험의 비율

변수처리) CNTT_LAMT_CNT / FMLY_PLPY_CNT

‘bank_loan_prop’

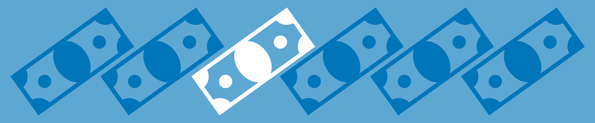
- 은행권 대출의 비율

변수처리) BNK_LNIF_AMT /TOT_LNIF_AMT

‘danger_loan’

- 위험대출비중

변수처리) CPT_LNIF_AMT /TOT_LNIF_AMT



‘insur_fam_prop’

- 가구 내 보험가입 비중

변수처리) CUST_FMLY_NUM / ACTL_FMLY_NUM

‘cost_prop’

- 월소득 대비 통신비, 보험료

변수처리) (FMLY_GDINS_MNPREM + FMLY_SVINS_MNPREM+ MON_TLFE_AMT) / HSHD_INFR_INCM

‘TEL_CNTT_QTR_duration_day’

- 통신요금 정지비중

변수처리) NUM_DAY_SUSP / TEL_CNTT_QTR_duration

‘have_child’

- 자녀의 유무

변수처리) 막내자녀 나이의 변수로 자녀의 유무만 따진다

‘bad_loan_prop’

- 은행권 대출 개수 대비 기타대출 비중

변수처리) (CPT_LNIF_CNT + SPART_LNIF_CNT + ECT_LNIF_CNT)/ BNK_LNIF_CNT

‘credit_prop’

- 소득대비 신용대출의 비중

변수처리) TOT_CLIF_AMT / HSHD_INFR_INCM

‘credit_long_overdue’

- 한화생명 대출 연체 중, 최근 1년과 30일 내의 연체를 비교
- 최근 1년에는 있지만 30일 내의 기록이 없을 시, 상환의 의지가 있다고 판단

변수처리) LT1Y_CLOD_RATE & CRLN_30OVDU_RATE 둘 다 값이 있을 시1,
나머지는 0

67개 변수 → 145개 변수

‘PREM_OVDU_RATE_section’

‘AVG_STLN_RATE_section’

변수처리) 10단위로 구간을 나눔

불균형 데이터

이벤트 비율이 5%이내만
머신러닝은 소수 클래스를 잡음으로 간주해 종종 무시한다

(Random)under-sampling

무작위로 다수 클래스를 제거, 소수 클래스와 비율을 맞춤
정보 손실의 단점

over-sampling

소수 클래스의 인스턴스를 늘려 비율을 맞춘다
다양한 방법이 있다

- Random over-sampling
- cluster-based over-sampling
- ✓ SMOTE(Synthetic Minority over-sampling Technique)

데이터의 부분집합은 소수 클래스로 부터 추출되고,
새로운 인스턴스가 생성된다

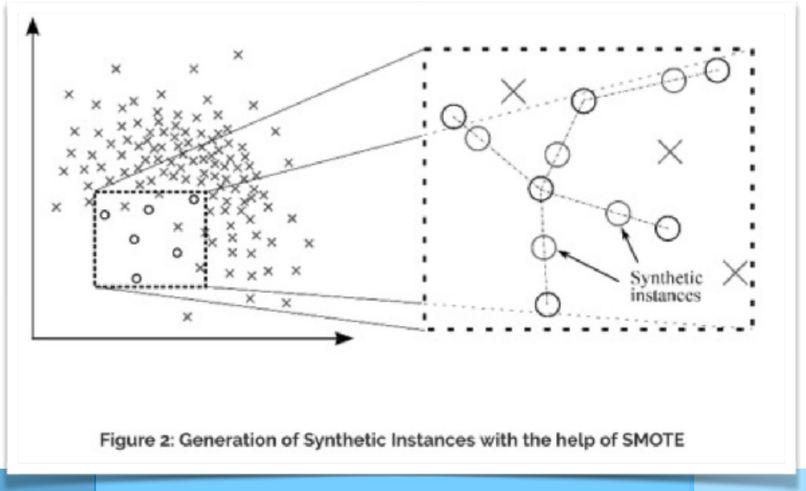
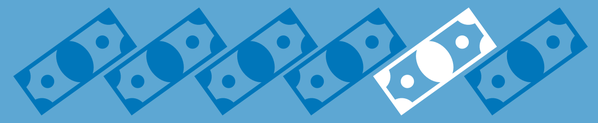


Figure 2: Generation of Synthetic Instances with the help of SMOTE



원 데이터

1) 전체 데이터 개수 : 100233
2) 상환 개수 : 95946
3) 부실 개수: 4287
4) 부실데이터 비율: 0.04

Scaling

SMOTE는 K-means 기반으로
새로운 인스턴스를 생성

Random Forest

F1 score:0.254

Border SMOTE

1) 전체 데이터 개수 : 191892
2) 상환 개수 : 95946
3) 부실 개수: 95946
4) 부실데이터 비율: 0.50

Random Forest

F1 score:0.982

XGBoost

F1 score:0.974

LightGBM

F1 score:0.980

보험 이탈

카드 사기

마케팅 성공률

비대칭 문제 해결

대출 부실

직원의 퇴사

사회, 산업 전반적인 문제 해결의 실마리