보험 Data와 통신 Data, 신용평가 Data를 활용한 대출 연체 여부 예측

2016. 12. 13

오성우, 김제현

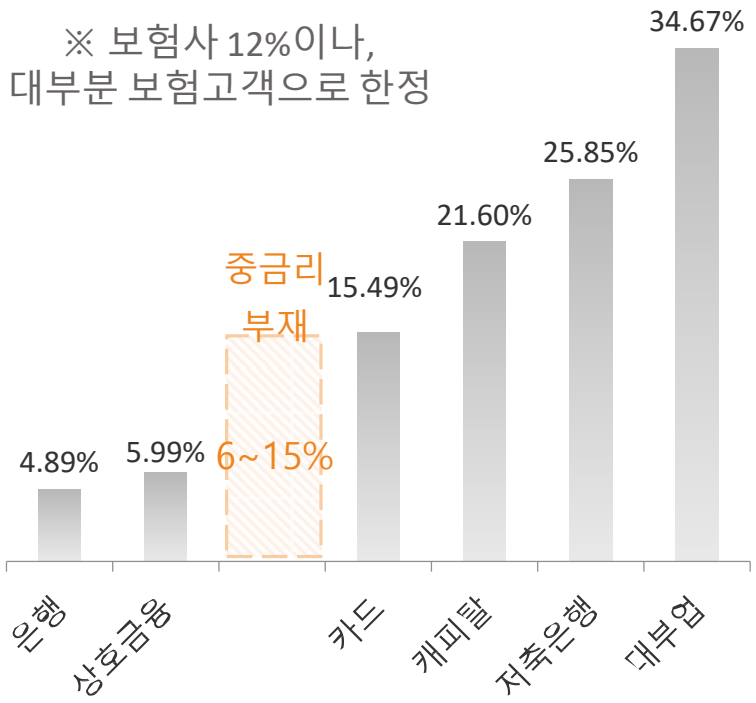
목 차

Background

새로운 신용평가방식 필요

국내 대출시장은 중금리수요 대비 공급이 부족하고, 은행과 비은행권 간의 금리 양극화가 존재해 왔음. 최근 들어 중신용자를 대상으로 한 시장확대 노력이 진행되고 있으나 기존의 신용평가 방식 활용만으로는 고객의 정확한 상환능력/의지를 파악하기에 한계가 있음

업계별 신용대출 평균금리



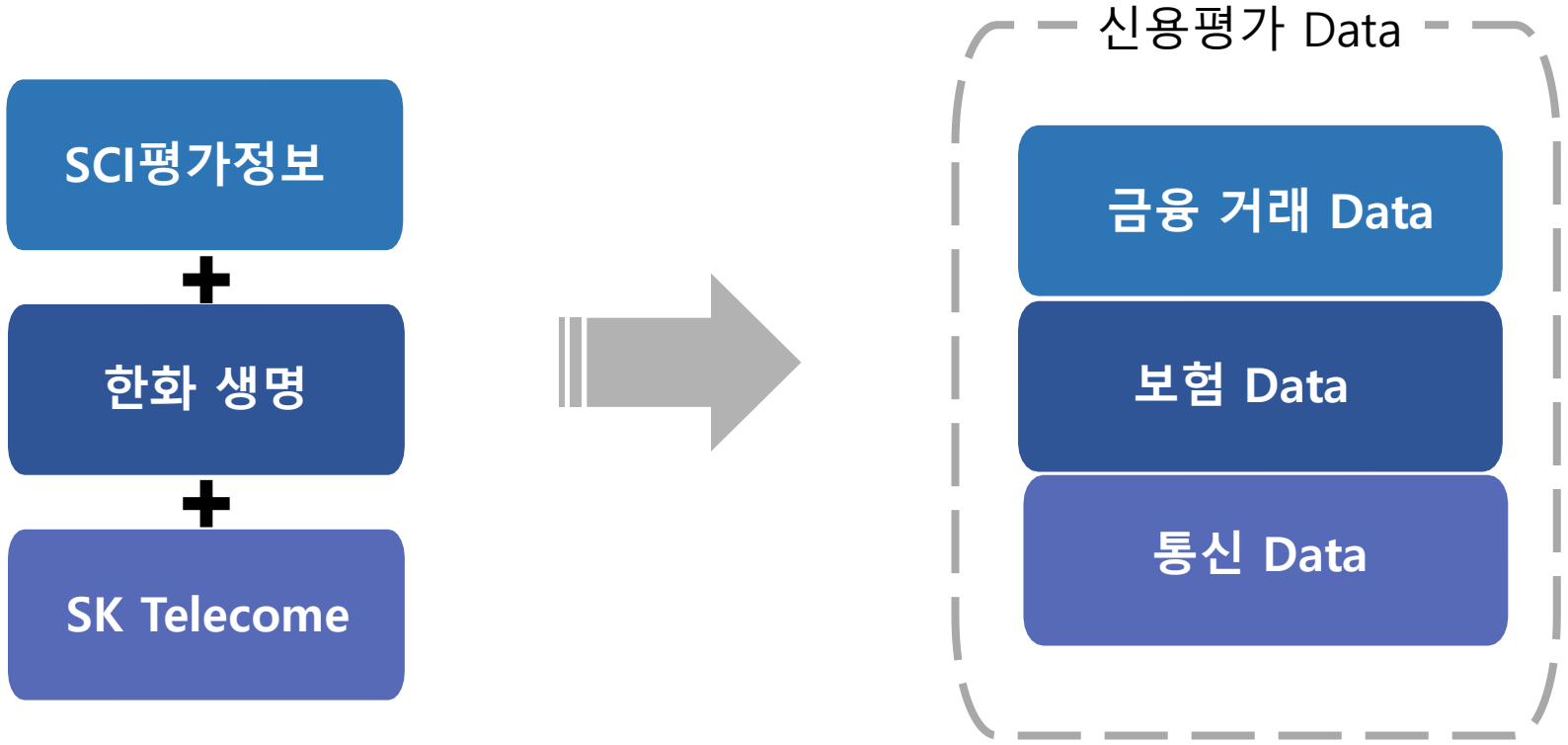
전통적인 신용평가 방식의 한계

- 중금리 대출 경험 Data 부족
- 신용정보사 Data 및 자사 Data 활용
- Rule 기반의 보수적인 신용평가

- 금융거래 실적 부족
- 금융 이외 기관과의 다양한 거래
- 정교화 된 가치평가 (금리 인하)

대출 연체여부 예측을 통한 신용평가모델

- 에서는 기존 활용하던 금융거래관련 Data(SCI평가정보) 이외에 보험(한화생명) 및 통신(SKT) 데이터 등 다양한 산업의 Data를 활용하여 **대출의 연체여부를 예측하는 알고리즘을 개발함으로써 Alternative 신용평가 모델 개발의 가능성을 검증**하고자 함



분석 절차

Data 탐색	Data 특성 파악
⬇	
Data 처리	결측치, 이상치 등 처리
⬇	
oversampling	불균형 자료를 균형 자료로 샘플링
⬇	
Data 분석	각 알고리즘에 따라 분류 모델 생성
⬇	
모델 평가	생성 된 모델에 test data를 넣어 모델의 성능을 평가
⬇	
모델 비교	각 모델간의 성능을 비교함.

Data 정의

Data 탐색

Data 출처

한화생명에서 주최한 빅데이터 분석 공모전에서 제공된 데이터

수집 범위

약 10만 명의 금융거래 및 보험가입 정보, 통신가입 정보 등을 비식별화하여 결합한 융합된 데이터, 데이터의 크기는 약 30MB이며 전체 3개의 테이블(금융거래정보-SCI, 보험가입정보-한화생명, 통신가입정보-SKT)로 구성되어 있으며 비식별화된 고객 기본기(PRIMARY KEY)로 조인이 가능. 3개의 테이블 조인 시, 전체 69개 필드와 102,252 레코드로 구성

Data 설명

실제 기업내부 데이터 기반의 데이터로 구성되어 있어 Null 값 등이 존재하며 파생변수 등의 다수 업종에 대한 이해가 필요함. 여러 데이터 결합 시 발생하는 개인정보 식별 가능성 때문에 비식별 처리가 되었으며 이 과정에서 데이터의 추가적인 가공으로 인해 삭제되고 마스킹, 범주화 등으로 데이터의 정보 손실이 발생함. 그리고 데이터의 특성 상 대출상환을 하는 이들의 비율이 적기 때문에 타겟 변수의 값이 imbalanced 되어 있음. 데이터 필드에 대해서는 다음 장에서 표로 설명

Data(금융거래)

No	변수영문명	변수명	변수 설명	비고
0	~~CUST_ID~~	~~고객_ID~~	~~임의로 부여한 고객번호~~	~~분석 필드가 아님~~
1	TARGET	대출연체여부	대출연체 발생 여부: 미발생(0), 발생(1)	Binary
2	BNK_LNIF_CNT	대출정보 현재 총 건수[은행]	산출일 기준 은행권에서 발생된 총 대출 건수	Numeric
3	CPT_LNIF_CNT	대출정보 현재 총 건수[카드사/할부사/캐피탈]	산출일 기준 카드사/할부사/캐피탈에서 발생된 총 대출 건수	Numeric
4	SPART_LNIF_CNT	대출정보 현재 총 건수[2산업분류]	산출일 기준 2산업분류에서 발생된 총 대출 건수	Numeric
5	ECT_LNIF_CNT	대출정보 현재 총 건수[기타]	산출일 기준 기타 금융권에서 발생된 총 대출 건수	Numeric
6	TOT_LNIF_AMT	대출정보 현재 총 금액	산출일 기준 총 대출 금액	Numeric
7	TOT_CLIF_AMT	대출정보 현재 총 금액[신용대출]	산출일 기준 총 신용대출 금액	Numeric
8	BNK_LNIF_AMT	대출정보 현재 총 금액[은행]	산출일 기준 은행권에서 발생한 총 대출 금액	Numeric
9	CPT_LNIF_AMT	대출정보 현재 총 금액[카드사/할부사/캐피탈]	산출일 기준 카드사/할부사/캐피탈에서 발생한 총 대출 금액	Numeric
10	CRDT_OCCR_MDIF	대출정보 최근 개설일로부터 현재까지 유지기간[신용대출]	신용대출 개좌 개설일부터 산출일까지 유지 개월 수	Numeric
11	SPTCT_OCCR_MDIF	대출정보 최근 개설일로부터 현재까지 유지기간[2산업분류-신용대출]	2산업분류에서 신용대출 개좌 개설일부터 산출일까지 유지 개월 수	Numeric
12	CRDT_CARD_CNT	개설정보 현재 신용개설 총 건수[신용카드]	산출일 기준 신용카드 발급 수	Numeric
13	CTCD_OCCR_MDIF	개설정보 최초 개설일로부터 현재까지 유지기간[신용카드]	신용카드개설일부터 산출일까지 유지 개월 수	Numeric
14	CB_GUIF_CNT	보증정보 현재 보증 총 건수	산출일 기준 총 보증 건수	Numeric
15	CB_GUIF_AMT	보증정보 현재 보증 총 금액	산출일 기준 총 보증 금액	Numeric

Data(보험사)

No	변수영문명	변수명	변수 설명	비고
16	OCCP_NAME_G	직업	산출일 기준 대분류 직업 정보 (NULL, *(비식별처리))	Categorical
17	CUST_JOB_INCM	추정소득	직업정보기반 추정 소득 금액	Numeric
18	HSHD_INFR_INCM	가구추정소득	가계 합산 추정 소득	Numeric
19	ACTL_FMLY_NUM	실가족원수	산출일 기준 입력된 가족원 수	Numeric
20	CUST_FMLY_NUM	보험가입가족원수	산출일 기준 보험가입이력이 있는 가족원 수	Numeric
21	LAST_CHLD_AGE	막내자녀나이	산출일 기준 입력된 막내 자녀의 나이 (0 = NULL)	Numeric
22	~~MATE_OCCP_NAME_G~~	~~배우자직업~~	~~산출일 기준 배우자의 대분류 직업 정보 (NULL, *(비식별처리))~~	~~분석 필드에서 제외~~
23	MATE_JOB_INCM	배우자추정소득	배우자 직업 또는 주소 기반 추정 소득 금액	Numeric
24	CRDT_LOAN_CNT	신용대출건수	산출일 기준 한화생명에서 실행된 총 신용대출 건수	Numeric
25	MIN_CNTT_DATE	최초대출날짜	한화생명에서 실행된 최초의 신용대출의 년월	Numeric
26	TOT_CRLN_AMT	한화생명신용대출금액	산출일 기준 한화생명에서 실행된 총 신용대출 금액	Numeric
27	TOT_REPY_AMT	한화생명신용상환금액	산출일 기준 한화생명에서 실행된 총 신용대출 금액 중 총 상환된 상환금액	Numeric
28	CRLN_OVDU_RATE	신용대출연체율	한화생명에서 실행된 신용대출이후 경과월수 중 연체경험월수의 비율	Numeric
29	CRLN_30OVDU_RATE	30일이내신용대출연체율	한화생명에서 실행된 30일이내 연체경험월수/ 30일이내 신용대출월수*100	Numeric
30	LT1Y_CLOD_RATE	최근1년신용대출연체율	한화생명에서 실행된 최근1년 연체경험월수/ 최근1년 신용대출월수*100	Numeric
31	STRT_CRDT_GRAD	최초신용등급	한화생명에서 실행된 가장 오래된 대출시점의 신용등급 (0(등급없음))	Categorical
32	LTST_CRDT_GRAD	최근신용등급	한화생명에서 실행된 가장 최근 대출시점의 신용등급 (0(등급없음))	Categorical
33	PREM_OVDU_RATE	보험료연체율	총납입보험료 횟수 중 연체한 보험료 횟수의 비율	Numeric
34	LT1Y_PEOD_RATE	최근1년보험료연체율	최근1년 연체납입횟수/총납입횟수*100	Numeric
35	AVG_STLN_RATE	평균약대율	월별 약관대출가능 금액 중 약관대출 받은 금액의 비율의 연중 평균	Numeric
36	STLN_REMN_AMT	약관대출가능잔액	약관대출 받은 금액	Numeric
37	LT1Y_STLN_AMT	최근1년약대금액	최근1년 약관대출 받은 금액	Numeric
38	LT1Y_SLOD_RATE	최근1년약대연체율	최근1년 약관대출연체경험월수/ 최근1년 약관대출월수*100	Numeric
39	GDINS_MON_PREM	非연금저축상품월납입보험료	유효한 계약 중 납입중인 보장성 상품의 월납환산보험료(일시납 제외)	Numeric
40	SVINS_MON_PREM	연금저축상품월납입보험료	유효한 계약 중 납입중인 저축성 상품의 월납환산보험료(일시납 제외)	Numeric
41	FMLY_GDINS_MNPREM	非가구연금저축상품월납입보험료	가계 합산 기준 유효한 계약 중 납입중인 보장성 상품의 월납환산보험료(일시납 제외)	Numeric
42	FMLY_SVINS_MNPREM	가구非연금저축상품월납입보험료	가계 합산 기준 유효한 계약 중 납입중인 저축성 상품의 월납환산보험료(일시납 제외)	Numeric
43	MAX_MON_PREM	최대월납입보험료	기준일 이전 납입한 월납입보험료 中 최대보험료	Numeric
44	TOT_PREM	기납입보험료	유효한 계약의 총납입보험료	Numeric
45	FMLY_TOT_PREM	가구기납입보험료	가계 합산 기준 유효한 계약의 총납입보험료	Numeric
46	CNTT_LAMT_CNT	실효해지건수	계약해지 또는실효난 계약건수	Numeric
47	LT1Y_CTLT_CNT	최근1년 실효해지건수	최근1년 계약해지 또는 실효난 계약건수	Numeric
48	AUTR_FAIL_MCNT	자동이체실패월수	산출일 기준 총 자동이체실패월수	Numeric
49	FYCM_PAID_AMT	가구총지급보험금액	가계 합산 보험금지급 총액	Numeric
50	FMLY_CLAM_CNT	가구총보험금청구건수	가계 합산 총 보험금청구 건수	Numeric
51	FMLY_PLPY_CNT	가구만기완납경험횟수	가구단위 만기까지 보험료를 완납한 증번의 갯수	Numeric

Data(통신사)

No	변수영문명	변수명	변수 설명	비고
52	AGE	연령	한화생명 및 SKT고객이면서 대출정보가 있는 고객의 연령 (*(비식별처리))	Numeric
53	SEX	성별	한화생명 및 SKT고객이면서 대출정보가 있는 고객의 성별: 1(남자), 2(여자)	Categorical
54	AVG_CALL_TIME	월통화시간_분	월평균 통화시간 분단위	Numeric
55	AVG_CALL_FREQ	월통화빈도	월평균 통화횟수	Numeric
56	~~TEL_MBSP_GRAD~~	~~멤버쉽등급~~	~~SKT멤버쉽 등급~~	~~분석필드에서 제외 결측값이 50% 이상이며 멤버쉽 등급은 납부요금에 결정되므로 다중공선성 문제 발생~~
57	ARPU	가입자매출_원	월기준 회선당 평균 수익금	Numeric
58	MON_TLFE_AMT	납부요금_원	월기준 서비스 납부요금	Numeric
59	CBPT_MBSP_YN	결합상품가입여부	인터넷, TV등 결합상품가입 여부: Y(가입) , N(미가입)	Categorical
60	MOBL_FATY_PRC	단말기가격_원	사용중인 핸드폰단말기 출고가액	Numeric
61	TEL_CNTT_QTR	가입년월_분기	SKT가입년월_분기단위: YYYYQ	Numeric
62	NUM_DAY_SUSP	정지일수	회선의 사용정지일수	Numeric
63	CRMM_OVDU_AMT	당월연체금액_원	해당월 납부요금의 연체금액	Numeric
64	TLFE_UNPD_CNT	납부일미준수횟수	핸드폰 납부요금의 납입일 미준수한 횟수	Numeric
65	LT1Y_MXOD_AMT	년간최대연체금액_원	산출일 기준 최근1년이내 납부요금 연체금액 中 최대 연체금액	Numeric
66	PAYM_METD	납부방법	납부요금의 납부 방법	Categorical
67	LINE_STUS	회선상태	산출일 기준 회선의 상태: S(정지), U(사용)	Categorical
68	MOBL_PRIN	남은할부금_원	산출일 기준 남아있는 핸드폰 단말기 할부원금	Numeric

1

오버샘플링에 전후에 따른 성능 비교

2

알고리즘에 따른 성능 비교

Data 탐색 및 전처리

Data 탐색 및 전처리

Data 탐색

• **Data 전체 100233 레코드 69개의 변수**

- 보험사, 통신사, 금융거래 데이터 통합을 위한 key Customer_id 제외 총 68개 변수

- 변수의 형태는 이분(Binary), 명목(Nomial), 연속등이 존재함.

Data 처리

• **잘못 지정된 Data 유형 및 값 처리**

- ex) Y/N -> 0/1 등(파이썬에서 값들 인식되서 바꾼 것들 수정)

• **결측치 처리**

- 결측값이 절반 이상을 차지하는 변수 LAST_CHLD_AGE, OCCP_NAME_G, MATE_OCCP_NAME_G 제거(대체하면 전체를 왜곡할 가능성이 있음)

- 결측값이 40% 가까이 존재하지만, 해당 변수에 따라 연체율의 차이가 존재하는 변수인 통신사 멤버십 변수는 사용.

(결측치를 포함하는 레코드 제거. 해당 레코드를 제거해도 모집단의 분포와 동일)

- 결측치 비율이 1% 이하인 변수들에 한하여 결측치를 대체함.

Data 변수 선택

범주형 변수

- 결측치가 많아 제거한 변수 3개를 제외한 각 변수에 따른 타겟 분포 확인

지불 방법

전체	G	K	O	R
4.0%	3.8%	1.7%	3.4%	14.9%

멤버십 등급

전체	E	Q	R	W
4.0%	**2.8%**	**4.5%**	**2.9%**	**5.7%**

결합상품 가입여부

전체	Y	N
4.0%	3.1%	5.4%

성별

전체	남자	여자
4.1%	**3.9%**	**4.3%**

회선 사용

전체	Use	Stop
4.1%	4.2%	17.2%

Data 변수 선택

연속형 변수

- 결측치가 많아 제거한 변수 3개, ID 1개를 제외한 각 변수에 따른 타겟 분포 확인
- 변수 중 두 집단간의 평균의 차이가 존재하지 않고, 분포가 동일한 변수는 제거

변수명	평균 비교	분포 비교
CUST_JOB_INCM	p-value = 0.3838	p-value = 0.1991
ACTL_FMLY_NUM	p-value = 0.3457	p-value = 0.3774
CUST_FMLY_NUM	p-value = 0.1131	p-value = 0.6106
…	…	…
SVINS_MON_PREM	p-value = 0.05304	p-value = 0.3443
FMLY_SVINS_MNPREM	p-value = 0.1906	p-value = 0.7068
FYCM_PAID_AMT	p-value = 0.7074	p-value = 0.7806
AVG_CALL_TIME	p-value = 0.1696	p-value = 0.05149
TEL_CNTT_QTR	p-value = 0.9614	p-value = 0.06836

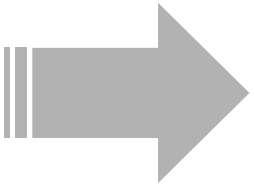
Data 불균형 자료 처리

oversampling

- 관측수가 큰 클래스를 모두 사용하고, 관측수가 작은 클래스의 관측수를 증대시키는 방법
- 기존 데이터를 중복하여 사용하는 방법과, 노이즈를 포함시키는 두가지 방법이 있음.
- 본 분석에서 두 방법 모두를 사용함.

원 데이터

전체	0(상환)	1(연체)
52843	50735	2108
100%	96%	4%

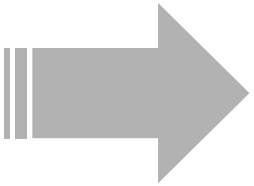


중복 oversampling

전체	0(상환)	1(연체)
101470	50735	50735
100%	50%	50%

원 데이터

전체	0(상환)	1(연체)
52843	50735	2108
100%	96%	4%



노이즈를 포함하는 oversampling

전체	0(상환)	1(연체)
46376	23188	23188
100%	50%	50%

분석 결과

분석 모형

- **Raw data**

전체	Positive (상환)	Negative (연체)
Positive (상환)	13651	312
Negative (연체)	1569	320

나이브 베이즈

- Accuracy : 0.8813
- Sensitivity : 0.8969
- Specificity : 0.5063
- Precision : 0.9777

전체	Positive (상환)	Negative (연체)
Positive (상환)	14731	489
Negative (연체)	285	347

랜덤포레스트

- Accuracy : 0.9512
- Sensitivity : 0.9810
- Specificity : 0.4151
- Precision : 0.9679

전체	Positive (상환)	Negative (연체)
Positive (상환)	14660	560
Negative (연체)	345	287

DeepLearning

- Accuracy : 0.9429
- Sensitivity : 0.9770
- Specificity : 0.6612
- Precision : 0.9632

분석 모형

- 나이브 베이지안 모델

전체	Positive(상환)	Negative(연체)
Positive(상환)	12913	226
Negative(연체)	2307	406

중복 upsampling

- Accuracy : 0.8402
- Sensitivity : 0.8484
- Specificity : 0.6424
- Precision : 0.9828

전체	Positive(상환)	Negative(연체)
Positive(상환)	12777	185
Negative(연체)	2443	447

노이즈 포함 upsampling

- Accuracy : 0.8342
- Sensitivity : 0.8395
- Specificity : 0.7073
- Precision : 0.9857

분석 모형

랜덤포레스트 모델

전체	Positive(상환)	Negative(연체)
Positive(상환)	15050	170
Negative(연체)	194	438

중복 upsampling

- Accuracy : 0.9770
- Sensitivity : 0.9873
- Specificity : 0.7204
- precision : 0.9888

전체	Positive(상환)	Negative(연체)
Positive(상환)	15035	185
Negative(연체)	181	451

노이즈 포함 upsampling

- Accuracy : 0.9769
- Sensitivity : 0.9881
- Specificity : 0.7091
- precision : 0.9878

분석 모형

딥러닝

전체	Positive(상환)	Negative(연체)
Positive(상환)	14854	366
Negative(연체)	165	467

중복 upsampling

- Accuracy : 0.9665
- Sensitivity : 0.9890
- Specificity : 0.5606
- precision : 0. 9760

전체	Positive(상환)	Negative(연체)
Positive(상환)	14810	410
Negative(연체)	173	459

노이즈 포함 upsampling

- Accuracy : 0.9632
- Sensitivity : 0.9884
- Specificity : 0.5281
- precision : 0.9731

Q&A