1. 데이터셋 이름

대출 연체 여부 기록 데이터

1. 사용목적

딥러닝 및 기타 머신러닝 기법을 활용하여 대출 연체 및 미상환 예측 알고리즘 개발

1. 데이터 소스

한화생명에서 주최한 빅데이터 분석 공모전에서 제공된 데이터

1. 수집범위

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

1. 데이터에 대한 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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 예측하고자 하는 필드 |
| 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 |
| 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)~~ | ~~분석 필드에서 제외(결측값이 50% 이상)~~ |
| 22 | MATE\_OCCP\_NAME\_G | 배우자직업 | 산출일 기준 배우자의 대분류 직업 정보 (NULL, \*(비식별처리)) | Categorical(kNN imputation으로 결측값 처리) |
| 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 |
| 52 | AGE | 연령 | 한화생명 및 SKT고객이면서 대출정보가 있는 고객의 연령 (\*(비식별처리)) | Numeric (kNN imputation으로 결측값 처리) |
| 53 | SEX | 성별 | 한화생명 및 SKT고객이면서 대출정보가 있는 고객의 성별: 1(남자), 2(여자) | Categorical (kNN imputation으로 결측값 처리) |
| 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 : E, Q, S, W | Categorical |
| 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 : 90s, 00s, 10old, 10new) | Categorical (90년대 가입자 ~ 2010년대 후반 가입자로 범주화) |
| 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 |

필드 1번, 22번, 56번을 제외한 전체 66개의 필드를 분석에 사용  
각 변수간 상관성 등을 파악하여 추가적으로 변수 선택을 하고 있으며 변수선택에 따른 예측 모델의 성능을 튜닝중

서로 다른 산업의 3 가지 데이터가 결합되었으나 상관성을 가지는 필드들이 존재했으며  
이는 다중공선성의 문제보다는 타겟의 예측력을 높일 수 있는 설명력 있는 변수로 보임

1. 데이터 샘플 ( 12개의 레코드에 대한 전체 필드)







1. 보완 계획

결측값이 존재하는 필드의 레코드는 제거하기보다는 추후에 입력되는 레코드가 존재한다는 가정 하에 해당 레코드들도 대치될 수 있도록 분류 모델(classification) 또는 회귀 모델(regression)을 사용하여 결측값을 대치하여 예측 모델 학습에 사용

약 10만 명의 데이터 중에서 예측해야 하는 대출 미상환자는 약 4,000명으로 4% 정도의 심한 imbalanced 데이터 이므로 학습 시에 balancing 작업이 가장 중요할 것이며 underscaling으로 할지 overscaling으로 할 것인지 학습 결과를 비교하며 결정할 예정  
또는 imbalance 문제에 영향을 받지 않는 고급 딥러닝 모델을 사용할 예정. 기존 분석에서 활용하는 딥러닝은(Feed Forward Network) 방식이며 이 모델 이외에 추가적인 고급 딥러닝 모델 활용 계획.

데이터 분류 기준에 따라 추천할 만한 대출 상품을 제안하는 것으로 프로젝트를 마무리하는 것이 최종 목표임