2017

BIG CONTEST

# Boosting에 기반한 대출연체 예측 앙상블 모델

연체동물 해부학

김동현 이기웅 김선빈 류지승 정다인

# 목차


## 01 서론

1.1 분석 요약 (1p)

1.2 보고서의 구성 (1p)

## 02 본론

2.1 비즈니스 배경 (2p)

2.2 유의변수 지정 및 탐색 (2-4p)

2.3 데이터 전처리 (4p)

2.4 이상치 및 결측치 처리 (5-6p)

2.5 파생변수 생성 및 개발 (6-11p)

2.6 모델링 (11-15p)

## 03 결론

3.1 요약 (16p)

3.2 한계 및 향후 과제 (16-17p)

# 1 서론

## 1. 분석 요약

본 보고서에서는 고객별 대출상환 예측을 위해 기본변수 및 파생변수를 이용하여 GBDT알고리즘 중 GBM과 XGBoost의 모형별 예측값에 대해 평균을 적용한 앙상블 알고리즘을 설명한다. 제공된 데이터에서는 다음과 같은 특성이 있다.

| |
|---|
| 첫째, 목표변수가 4%에 불과한 불균형 데이터이다. |
| 둘째, 3사(SCI평가정보, 한화생명, SKT)의 요약데이터이다. |
| 셋째, 핵심변수로 예상되는 변수들의 값이 약 90%가 결측치였다. |
| 넷째, 상당 수의 변수들이 0으로 편중된 분포를 가졌다. |

**[표 1] 제공된 데이터의 특성**

**[표 1]**에서 제시한 데이터 특성에 대한 접근방법으로서 최적의 기계학습 알고리즘 탐색, 파생변수 개발, 핵심변수의 결측치 예측, 0에 편중된 수치형 변수의 분포를 변환하였다. 변수 선택방법으로 변수별 연체자 분포와 변수 중요도를 종합적으로 고려한 후진제거법을 적용하였다. 기계학습 알고리즘 선택방법으로 기본 67개의 변수를 사용한 알고리즘 벤치마크 테스트를 하여 GBDT기반. GBM, XGboost 2가지를 선정하였다. 모델의 성능향상을 위해 둘의 예측값에 대해 평균값을 적용하여 앙상블하였다. 본 팀의 대출연체 알고리즘 최종 예측성능은 0.46~0.48의 f1값으로 예상된다.

## 2. 보고서의 구성

**제 1장 (서론)**에서는 분석과정과 결과를 요약하고 및 분석 절차에 대해서 언급하였다.

**제 2장 (본론)**에서는 데이터 탐색과정과 전처리과정을 언급하고, 생성된 파생변수에 관해 설명하고 모델 구축 및 변수 선택과정, 모델 개선방법의 내용을 기술하였다.

**제 3장 (결론)**에서는 모델을 평가하여 모델을 확정하고, 모델의 성능을 높이기 위한 방법을 제시한다.

*※ 개발절차 : 데이터 탐색 -> 데이터 전처리 -> 모델 구축 및 평가 -> 모델 확정*

# 2 본론

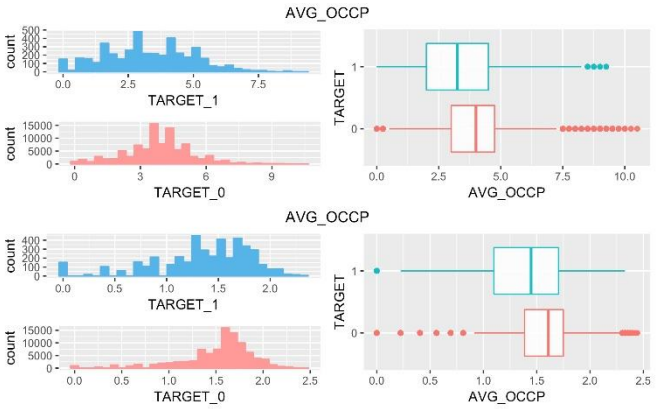
본 팀은 대출연체에 관한 도메인 지식을 어떻게 모델에 적용시킬 수 있을 지 고민하였다. 그 결과, 대출연체 및 상환의 여부는 신용과 밀접하다 판단하였다. 우리는 신용평가에 직접 또는 간접적으로 영향을 미치는 원인들에 대해서 조사하였다. 신용등급이 낮아지는 몇 가지 방법에 대해서는 통신비, 카드대금의 대출연체기록, 대출한 금융기관, 높은 빈도의 현금서비스사용 등에 따라 신용등급이 하락할 수 있다. 반대로 신용카드 유지기간 길수록, 연체정보가 없을수록, 사금융을 이용하지 않을수록, 주거래은행과 꾸준한 거래를 유지할수록 신용등급을 상승시킬 수 있는 것으로 파악되었다. 이러한 도메인 지식을 바탕으로, 다음과 같은 데이터 분석을 시행하였다.

## 2. 유의변수 지정 및 탐색

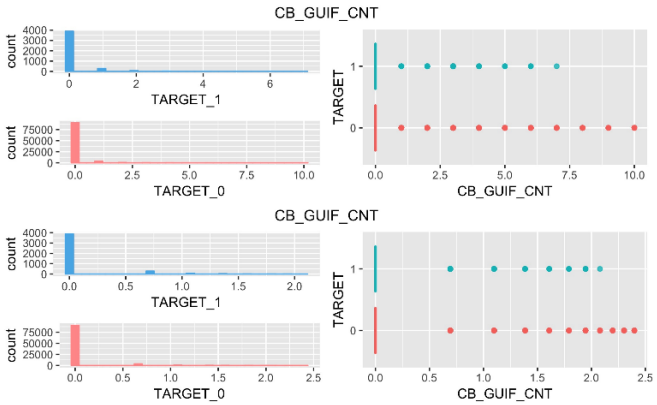
대출연체 및 상환예측을 위해서 신용등급을 하락시키거나, 상승시킬 수 있는 변수들을 살펴보았다. 아래에서는 SCI평가정보, 한화생명, SKT로 주제를 나누어 설명한다.

### 2.1) SCI신용정보

SCI평가정보의 데이터는 신용평가를 시행하고 있는 회사로서, 주어진 데이터가 모두 유의하다고 판단했다. 특히 은행, 2산업분류에서 대출한 금액과 건수 등 고객의 금융정보와 대출패턴을 알 수 있는 정보가 포함되어있다. 현재 신용평가에서 밀접한 연관이 있는 신용카드 유지기간의 정보와 사금융 이용여부 또한 파악할 수 있다. 아래는 이에 대한 예시이다.




[그림 1] 직업별 연체자 분포



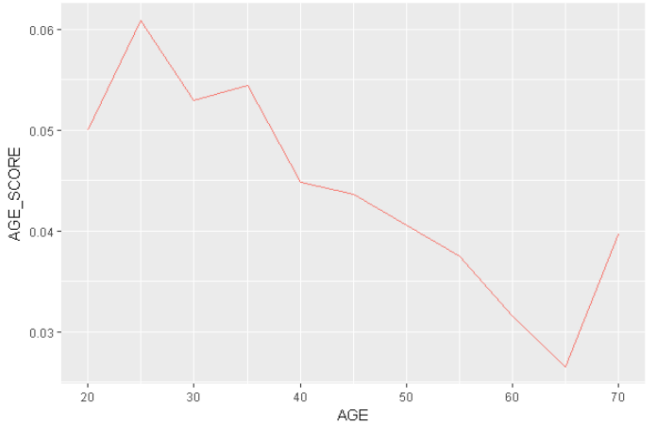

[그림 2] 보증금액에 따른 연체자 분포

### 2.2) 한화생명

한화생명 데이터는 0으로 편중된 변수가 대체적으로 많았다. 제공된 정보는 크게 고객 개인정보, 한화생명 대출정보, 보험관련 정보 3가지이다.

### 가. 고객 개인정보

한화생명에서는 개인, 배우자, 가족의 데이터가 주어진다. 본 팀은 나이와 직업이 대출연체에 유의미할 것이라 판단하여, 나이, 개인추정소득, 가구추정소득 등을 유의변수로 두었다. 가족의 데이터에서는 '부양가족이 대출연체에 영향을 준다'는 가설을 세워 접근하였다. 본 팀은 본인직업과 배우자직업의 패턴에 따라 연체율이 다를 것이라는 가설과 나이, 즉 생애주기에 따라 연체패턴이 다를 것이라는 가설을 세웠다.




**[그림 3] 나이에 따른 연체율 점수화**

### 나. 한화생명 대출정보

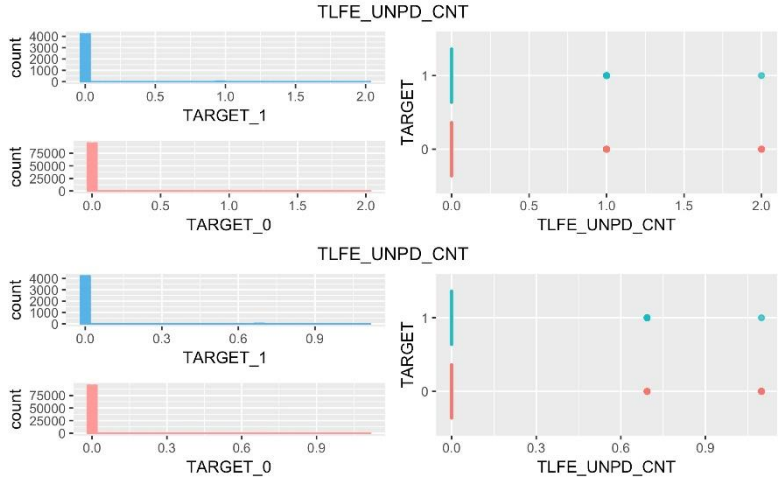
한화생명의 대출정보에는 신용대출 건수, 대출금액, 상환금액, 연체율, 신용등급 등 연체율과 밀접한 변수들이 많았다. 하지만 값이 있는 고객이 전체 10%에 불과하여, 한화생명에서 대출경험이 있는 고객 세그먼트를 위한 모델링을 시도하였다. 하지만 좋은 성능을 보이지 않아 세그먼트 모델로 사용하지 않았다.

### 다. 보험 정보

한화생명의 보험정보는 보험료 연체율, 만기완납경험횟수, 자동이체실패월수, 저축성 또는 비저축성 보험 관련 정보를 담고 있다. 보험의 가장 중요한 기능은 리스크 회피이다. 높은 보험료나 보험가입율은 리스크 회피에 대한 비용을 높게 지불하는 것을 감수함을 의미한다. 본 팀은 리스크 회피 관심도와 연체에 대한 유의미성에 초점을 맞추어 보험정보들을 살펴보았다. 더불어, 보험의 특성별로 연체자 분포에 차이가 있을 것이란 가설을 세웠고, 저축성 보험과 비저축성보험에 대해 모델을 나누어 살펴보았다.

## 2.3) SKT

SKT의 데이터에는 고객의 멤버쉽등급, 납부일미준수횟수, 년간최대연체금액, 납부일미준수횟수 등 대출연체와 밀접한 것으로 추정되는 변수들이 존재했다. 하지만 멤버쉽등급은 변수가 53,982행으로 해당 변수의 약 54%가 누락되어 있었다. 납부일미준수횟수는 0으로 편중된 분포와 함께 3개의 Level을 가져 설명력과 영향력을 가지기 힘들 것으로 판단되었다.




**[그림 4] 0으로 편중된 납부일미준수횟수**

## 3. 데이터 전처리

파생변수를 포함한 전체 변수는 112개이며, 이는 104개의 수치형 변수와 8개의 범주형 변수로 구성되어 있다. 다음은 전처리 대상변수를 특징 별로 정리한 것이다.

### 3.1) 금액단위 통일

| 조정값 | 대상변수 |
|---|---|
| * 0.1 | TOT_LNIF_AMT, TOT_CLIF_AMT, BNK_LNIF_AMT, CPT_LNIF_AMT, CB_GUIF_AMT |
| * 0.0001 | TOT_CRLN_AMT, TOT_REPY_AMT, STLN_REMN_AMT, LT1Y_STLN_AMT, GDINS_MON_PREM,SVINS_MON_PREM,FMLY_GDINS_MNPREM, FMLY_SVINS_MNPREM, MAX_MON_PREM, TOT_PREM, FMLY_TOT_PREM, FYCM_PAID_AMT, ARPU, MON_TLFE_AMT, MOBL_FATY_PRC, CRMM_OVDU_AMT, LT1Y_MXOD_AMT, MOBL_PRIN |

### 3.2) 도수 간 간격조정

일부 데이터는 금액 변수 단위 통일을 거친 후에 도수 간격이 0.1에서 300.1로 불균일하였다. 이 간격을 균일화 하기 위해 149.9를 0이외의 모든 도수에 더해주었다.

대상변수 : TOT_LNIF_AMT, TOT_CLIF_AMT, BNK_LNIF_AMT, CPT_LNIF_AMT

### 3.3) 명목형 변수의 수치화

순서를 갖는 명목형의 변수를 수치형으로 변환해주었다.

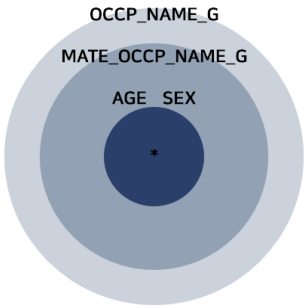
대상변수 : CRDT_OCCR_MDIF, SPTCT_OCCR_MDIF, CTCD_OCCR_MDIF, LT1Y_PEOD_RATE, CBPT_MBSP_YN, LINE_STUS

## 4. 이상치 및 결측치 처리

### 4.1) 나이 및 직업변수

| 변수 | Data_set | Test_set |
|---|---|---|
| 성별 & 나이 | 430 | 4 |
| 본인직업 | 1027 | 14 |
| 배우자직업 | 1189 | 15 |

**[표 2] 이상치 *에 대한 분포**




**[그림 5] * 이 있는 행들의 포함 관계**

본 팀은 예측 성능을 저하시킬 수 있는 극소량의 이상치 제거를 위해, Data_set에서 AGE와 SEX 변수에 * 이 있는 행들은 이상치로 판단하고 제거하였다. 따라서 Test_set의 나이가 * 인 4명의 고객은 알고리즘의 예측 대상에서 제외되었다. 그리고, 직업 변수(OCCP_NAME_G, MATE_OCCP_NAME_G)에 있는 * 은 제거하지 않았다. 이는 기타 이외에 또 다른 직업군으로서의 의미가 있다고 판단했기 때문이다.

대상 변수 : SEX, AGE, OCCP_NAME_G, MATE_OCCP_NAME_G

### 4.2) 신용등급

핵심변수로 예상되는 최초신용등급과 최근신용등급 변수는 한화생명에서 대출경험이 전제되는 변수이기 때문에 결측치가 많았다. 이에, 결측치를 예측하는 신용등급 예측모형 개발을 시도했다. 하지만 신용등급 예측결과가 낮아 이 모델은 활용하지 않았다.

모델 : Random Forest

변수 : 한화 데이터 (36개 변수)

결과 : Accuracy 0.34

| | | Reference | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Prediction | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 3 | 1 | 21 | 9 | 13 | 3 | 0 | 1 | 1 | 0 | 0 |
| | 4 | 11 | 30 | 29 | 95 | 43 | 25 | 7 | 1 | 0 | 0 |
| | 5 | 4 | 28 | 29 | 89 | 124 | 90 | 23 | 12 | 1 | 1 |
| | 6 | 1 | 11 | 9 | 41 | 107 | 194 | 97 | 46 | 10 | 10 |
| | 7 | 0 | 1 | 1 | 12 | 24 | 76 | 58 | 47 | 15 | 23 |
| | 8 | 0 | 0 | 0 | 0 | 4 | 2 | 3 | 11 | 6 | 13 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 12 |
| | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

### 4.3) 멤버십 등급

핵심변수로 예상되는 멤버십등급에서도 결측치가 존재하여, 값이 누락된 고객에 대해 멤버십을 예측하는 모형개발을 시도했다. 하지만 멤버십 예측결과가 기대보다 낮아 이 모형도 활용하지 않았다.

모델 : Random Forest

변수: SKT 데이터 (16개 변수) + TEL_CUS_MDIF (1개 파생변수)

결과: Accuracy 0.69

| | | reference | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Prediction | 1 | 2958 | 314 | 90 | 27 |
| | 2 | 1005 | 3522 | 1299 | 95 |
| | 3 | 160 | 1050 | 4067 | 546 |
| | 4 | 18 | 22 | 359 | 527 |

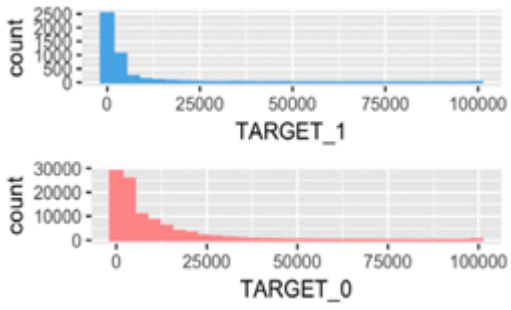
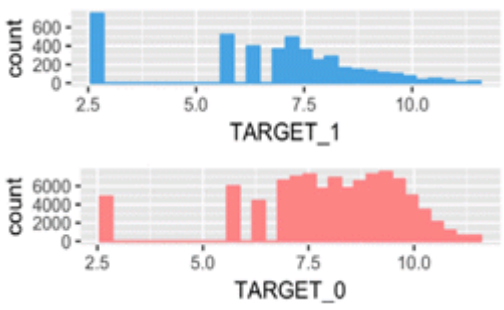
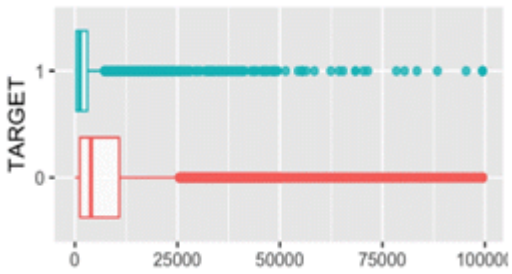
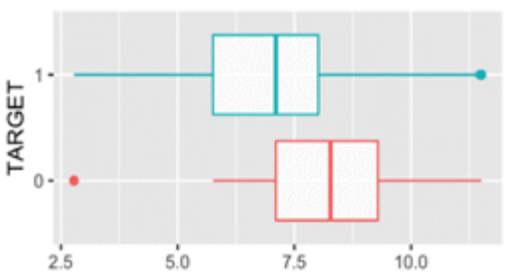
## 5. 파생변수 생성 및 개발

연체정보에 대한 설명력있는 변수를 만들기 위해 사용된 개발방법들을 소개한다.

### 5.1) 파생변수 개발방법

#### 가. 로그변환

왼쪽(0)으로 치우친 특정 변수들에 편향된 분포와 변수 간의 관계가 잘 드러나지 않는 문제를 해결하기 위하여 Log 변환법을 적용하였다. 그 결과, 모델 내에서 변수의 활용도를 향상시키는 결과를 얻을 수 있었다. 유의한 변수를 파악하기 위해서 변수별 연체/비연체자 분포를 비교분석하였다. 두 집단의 분포차이를 쉽게 파악하기 분포를 각각의 히스토그램과 박스 plot으로 시각화하여 분석하였다. 전체 변수에 대한 시각화자료는 **'[붙임1]** 변수별 연체자 분포 시각화 자료' 에 첨부하였다.

| TOT_LNIF_AMT의 기존 분포 | 로그변환 후 분포 |
|---|---|
|  |  |
|  |  |

### 나. 단순 연산을 통한 파생변수

#### 나.1) 비율을 사용한 파생변수

본 팀은 대출을 어디서 얼마나 받았는가에 대한 정보를 표현하기 위해, 전체대출 건수 중 은행대출 건수 비율과 같이 비율을 사용한 다양한 파생변수들을 만들었다.

ex) BNK_LNIF_CNT_PREM, INCM_LNIF_RATE_IND_LOG, TOT_CLIF_AMT_PREM,…

#### 나.2) 평균을 사용한 파생변수

본 팀은 은행대출 1회시 금액의 평균, 신용카드 유지기간 평균 등 대출연체에 있어서 긍정적 작용과 부정적 작용이 될 요인들의 평균값을 구해 다양한 파생변수들을 만들었다.

ex) AVG_OCCR, BNK_LNIF_AMT_AVG, CNT_SCORE_AVG,…

#### 나.3) 연체자 비율을 점수화한 파생변수

본 팀은 직업, 대출금액, 대출건수별 등의 변수를 범주형으로 나눠, 각 범주에 연체자 비율을 적용해 파생변수들을 만들었다.

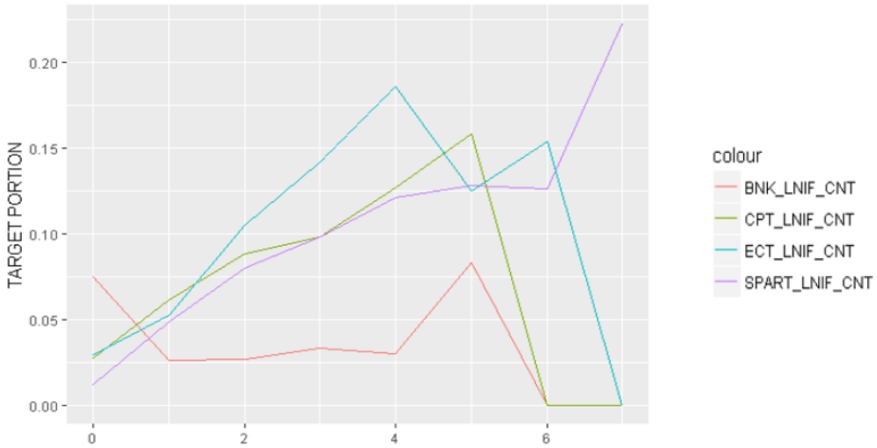
ex) JOB_RATE, BNK_CNT_SCORE, INCM_GRUP …

### 다. 도메인 지식을 활용한 파생변수

데이터 탐색과정에서 의미를 파악하고 그에 따른 접근으로 만든 파생변수들이다.

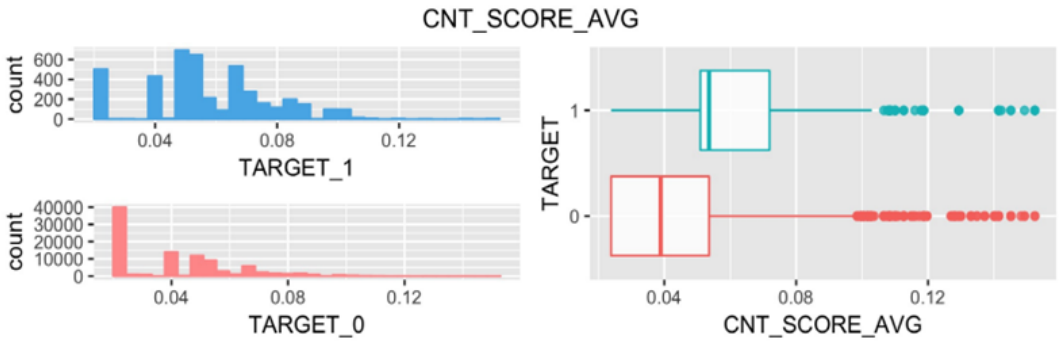
#### 다.1) CNT_SCRORE_AVG

CNT_SCORE_AVG는 SCI 제공 데이터 중 대출정보 건수별 연체비율 합의 평균을 의미한다. 아래의 그림은 대출기간별 건수에 대한 연제자의 비율을 시각화한 자료이다.




**[그림 6] CNT_SCORE_AVG에 대한 연체자 비율**

위에서 계산된 연체자비율로 각 기관의 대출횟수별 연체확률 테이블을 만들었다. 고객별로 기관별 대출횟수에 해당하는 연체확률 값을 더하여 평균한다. CNT_SCORE_AVG는 최소 0부터 최대 7의 값을 가지며 다음과 같은 분포를 갖는다.




**[그림 7] CNT_SCORE_AVG의 연체자 분포**

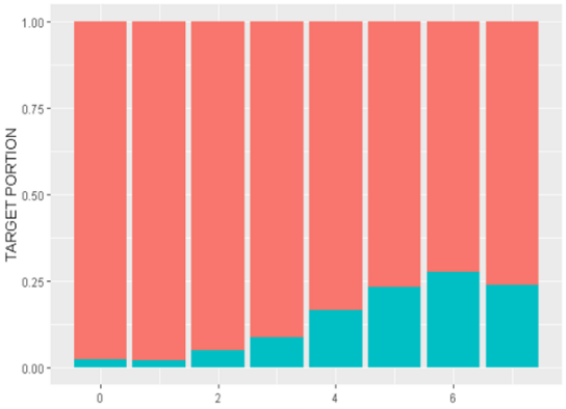
분석 결과 : 대부분의 연체자는 0.04 이후의 CNT_SCORE_AVG값을 가지며, 비연체자는 0에 집중 분포되어있다. 이를 통해 기관별 대출건수를 종합하여 살펴봤을 때, 두 집단 간 분명한 차이가 있음을 확인했다.

다.2) OVDU_CNT

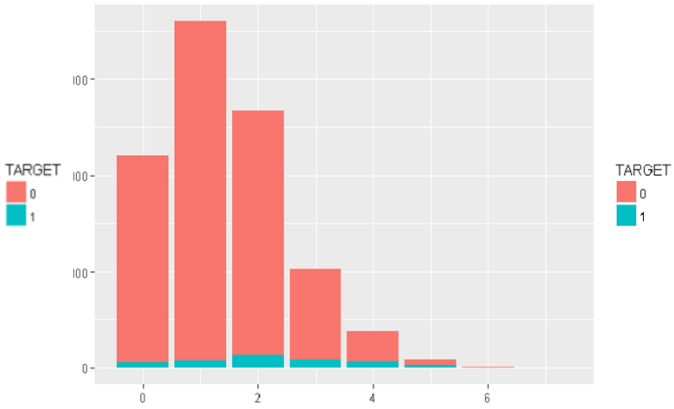
OVDU_CNT는 3사 제공변수 전체에서 0에 편중된 연체관련 변수들에 한해 0과 1로 이진화하여 합계한 변수이다. 활용한 연체 관련 변수들은 다음과 같다. 이 중 LT1Y_CLOD_RATE(최근 1년 신용대출 연체율)는 0의 비율이 99%이상으로 활용하지 않았다.

| 변수명 | 변수 설명 |
|---|---|
| CRLN_OVDU_RATE | 신용대출연체율 |
| CRLN_30OVDU_RATE | 30일이내 신용대출 연체율 |
| PREM_OVDU_RATE | 보험료 연체율 |
| LT1Y_PEOD_RATE | 최근 1년 보험료 연체율 |
| LT1Y_SLOD_RATE | 최근 1년 약대 연체율 |
| CRMM_OVDU_AMT | 당월 연체금액_원 |
| LT1Y_MXOD_AMT | 년간 최대 연체금액_원 |

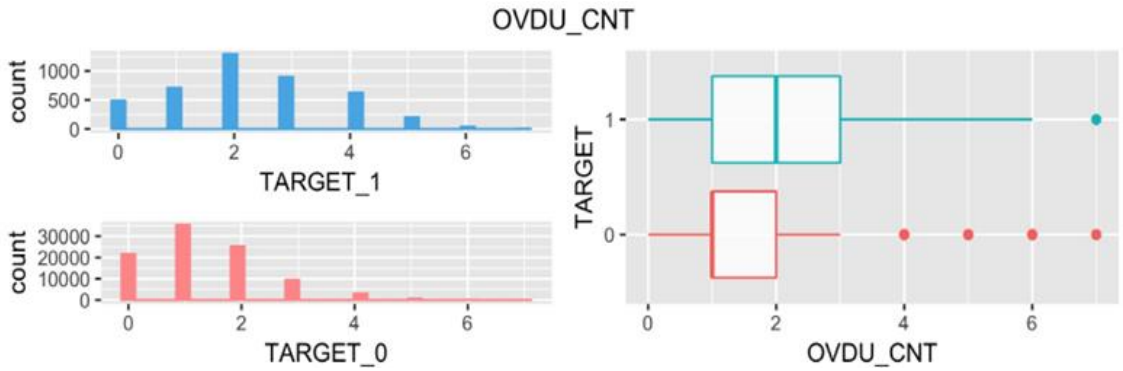
결과적으로 OVDU_CNT는 최소 0부터 최대 7의 값을 가진다. OVDU_CNT의 도수별 빈도의 분포는 아래 왼쪽으로 약간 치우친 모습을 확인할 수 있다. [그림 9]. 또한 이 변수의 도수값이 높아질수록 연체자의 비율이 상승하였다. [그림 10]




**[그림 8] OVDU_CNT별 연체자 비율**




**[그림 9] OVDU_CNT별 연체자 비율2**




**[그림 10] OVDU_CNT의 연체자 분포**

## 5.2) SCI신용평가

SCI신용평가의 데이터는 은행, 카드사, 기타금융권에서 대출받은 정보를 중심으로 연체율과 관련이 깊을 것으로 생각되는 것들이 많았다. 특히, 대출이 발생한 금융권의 건수별, 금액별 비율과 연체율을 점수화하여 총 28개의 파생변수를 개발하였다.

| 연번 | 변수명 | 변수설명 | 비고 |
|---|---|---|---|
| 1 | TOT_LNIF_CNT | 총 대출건수 | |
| 2 | BNK_LNIF_CNT_PREM | 전체대출 중 은행대출 비율 | |
| 3 | SPART_LNIF_CNT_PREM | 전체대출 중 2산업분류 대출 비율 | |
| 4 | CPT_TOT_CNT_PREM_LOG | 은행외의 대출 중 카드사비율 | 로그변환 |
| 5 | ECT_TOT_CNT_PREM | 은행외의 대출 중 기타비율 | |
| 6 | TOT_MTIF_AMT | 전제대출 중 신용대출이 아닌 금액 | |
| 7 | CPT_LNIF_CNT_PREM_LOG | 전체대출건수 중 카드사 비율 | 로그변환 |
| 8 | ECT_LNIF_CNT_PREM_LOG | 전체대출건수 중 기타 비율 | 로그변환 |
| 9 | BNK_CNT_SCORE | BNK건수별 연체비율 점수 | |
| 10 | CPT_CNT_SCORE | CPT건수별 연체비율 점수 | |
| 11 | SPART_CNT_SCORE | SPART건수별 연체비율 점수 | |
| 12 | ECT_CNT_SCORE | ECT건수별 연체비율 점수 | |
| 13 | CNT_SCORE_AVG_LOG | 건수별 연체비율 점수 평균 | 로그변환 |
| 14 | BNK_GUIF_AMT_LOG | 신용대출 외 대출금액 | 로그변환 |
| 15 | BNK_LNIF_AMT_AVG_LOG | 산1회 은행대출 평균금액 | 로그변환 |
| 16 | CPT_LNIF_AMT_AVG_LOG | 1회 카드사대출 평균금액 | 로그변환 |
| 17 | SPART_ECP_CPT_CNT | 1회 기타대출 평균금액 | |
| 18 | INCM_LNIF_RATE_IND_LOG | 개인소득대비 부채비율 | 로그변환 |
| 19 | INCM_LNIF_RATE_COU_LOG | 개인 + 배우자 소득대비 부채비율 | 로그변환 |

| | | | |
|---|---|---|---|
| 20 | INCM_LNIF_RATE_FAM_LOG | 가구소득대비 부채비율 | 로그변환 |
| 21 | TOT_CLIF_AMT_PREM_LOG | 전체 대출금액 중 신용대출 금액 | 로그변환 |
| 22 | TOT_MTIF_AMT_PREM_LOG | 전체 대출금액 중 신용대출 외 금액 | 로그변환 |
| 23 | BNK_LNIF_AMT_PREM_LOG | 전체 대출금액 중 은행대출 | 로그변환 |
| 24 | CPT_LNIF_AMT_PREM_LOG | 전체 대출금액 중 카드사 대출 | 로그변환 |
| 25 | ECT_LNIF_AMT_LOG | 은행, 카드사 외 대출 금액 | 로그변환 |
| 26 | ECT_TOT_AMT_PREM | 전체 대출금액 중 기타 대출 금액 | |
| 27 | AVG_OCCR | 신용대출, 신용카드 유지개설 기간 평균 | |
| 28 | AVG_OCCR2 | 신용대출 유지개설 기간 평균 | |

### 5.3) 한화생명

한화생명의 데이터는 고객의 개인특성이 잘 나타나 있다. 직업과 수입, 가족과 보험, 한화생명 대출정보, 신용등급, 자동이체 실패경험 등 연체율과 관련이 높을 것으로 보이는 데이터들을 활용하여 총 11개의 파생변수를 개발했다.

| 연번 | 변수명 | 변수설명 | 비고 |
|---|---|---|---|
| 1 | JOB_RATE | 본인과 직업패턴기준 연체자비율 | |
| 2 | OCCP_NAME_G2 | 본인직업 연체자 비율점수화 | |
| 3 | OTHR_INCM | 가구추정소득에서 본인, 배우자 소득 차 | |
| 4 | CUST_ACTL_INS_PREM | 가족 중 보험가입 인원의 비중 | |
| 5 | INCM_GRUP | 가구추정소득 기준 소득분위 | |
| 6 | DIFF_OVDU_RATE | 보험료연체율과 최근1년보험료연체율 차이 | |
| 7 | PREM_OVDU_AMT_LOG | 보험료 연체 금액 | 로그변환 |
| 8 | PER_TOT_PREM_LOG | 기납입보험료비율 | 로그변환 |
| 9 | PER_GDINS_PREM | 비연금저축상품 월납입보험료 비율 | |
| 10 | AUTR_FAIL_MCNT_FREQ | 자동이체 빈도수 기준 범주화 | |
| 11 | MON_TLFE_AMT_PREM | 최초대출 날짜부터 오늘까지 기간 (월) | |

### 5.4) SKT

SKT의 데이터에서는 핸드폰 납부요금 연체경험의 정보를 중심으로 총 4개의 파생변수를 개발하였다.

| 연번 | 변수명 | 변수설명 | 비고 |
|---|---|---|---|
| 1 | TEL_CUS_MDIF_LOG | 통신사를 SKT로 유지한 기간(월) | 로그변환 |
| 2 | MXOD_YN_LOG | 납부요금 연체경험 | 로그변환 |
| 3 | PAYM_METD_NUM | 납부방법 수치화 | |
| 4 | LT1Y_MXOD_Y_PROP | 최근 1년 납부요금 연체유무 비율 | |

### 5.5) 한화생명, SKT의 데이터

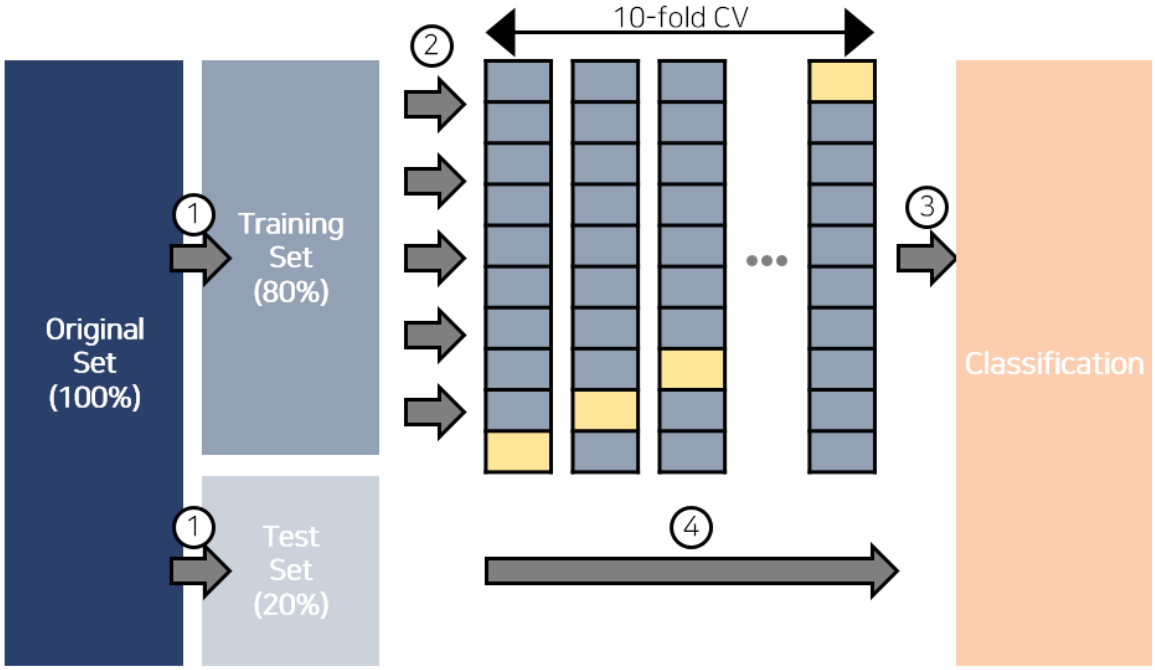
한화생명과 SKT의 데이터에서 연체율과 관련된 변수들에 집중했다. 연체 관련 변수들의 횟수를 더하거나, 범주별 연체자 비율을 점수화를 시켜 총 2개의 파생변수를 개발했다.

| 연번 | 변수명 | 변수설명 | 비고 |
|---|---|---|---|
| 1 | OVDU_CNT | 연체관련 변수의 합 | |
| 2 | OVDU_CNT_SCORE | 연체관련 변수 연체율 점수화 | |

## 6. 모델링

### 6.1) 기계학습 알고리즘 별 성능비교

모델의 과적합(Over-fitting)을 피하기 위해서 Validation Set을 활용하는 K-fold Cross Validation(CV)을 적용했다. 원본 데이터를 Training Data Set(80%)과 Test Data Set(20%)으로 분리한 후, 10-fold CV을 진행했다.




**[그림 11] 모형 적용 방법**

각 알고리즘별 성능 비교를 위해 다음의 조건으로 실험을 진행하였다.

첫째, 예측변수로 기본변수 67개를 사용했다.
둘째, H2O패키지의 GBM, Random Forest, XGBoost를 실험 대상으로 설정했다.
셋째, 각 알고리즘별 Hyper Parameter인 ntree값을 조정했다
넷째, Threshold는 10-Cross Validation에 대한 max f1 값이 나타나도록 하는 평균값을 선택했다.

| Classifier | ntrees | Accuary | Recall | Precision | F-score |
|---|---|---|---|---|---|
| GBM | 50 | 0.949 | 0.525 | 0.431 | 0.472 |
| | 100 | 0.952 | 0.511 | 0.459 | 0.481 |
| | 200 | 0.951 | 0.528 | 0.450 | 0.484 |
| | 500 | 0.952 | 0.512 | 0.461 | 0.479 |
| | 1000 | 0.952 | 0.512 | 0.461 | 0.479 |
| Random Forest | 50 | 0.947 | 0.501 | 0.413 | 0.451 |
| | 100 | 0.950 | 0.489 | 0.437 | 0.460 |
| | 200 | 0.950 | 0.501 | 0.433 | 0.462 |
| | 500 | 0.948 | 0.514 | 0.423 | 0.462 |
| | 1000 | 0.949 | 0.512 | 0.429 | 0.464 |
| XGBoost | 50 | 0.948 | 0.532 | 0.432 | 0.473 |
| | 100 | 0.950 | 0.501 | 0.436 | 0.463 |
| | 200 | 0.949 | 0.492 | 0.424 | 0.453 |
| | 500 | 0.950 | 0.487 | 0.439 | 0.458 |
| | 1000 | 0.950 | 0.496 | 0.432 | 0.460 |

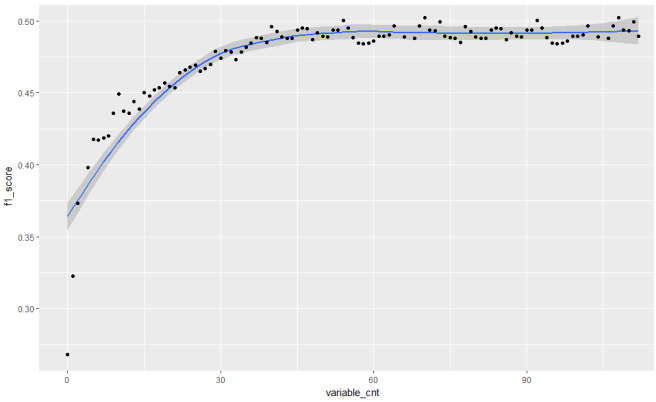
**[표 3] 모델 별 성능비교**

본 팀은 앞선 과정에서 변수의 선택 과정으로서 여러가지 도메인 지식을 활용하여 변수선택을 했기 때문에 모델 선택에 있어서는 성능을 우선하여 알고리즘을 선정하였다. 로지스틱 회귀모형은 설명력은 좋지만 변별력은 떨어진다. 이에 반해, Bagging 및 Boosting 모형 및 신경망 모형은 설명력이 낮지만 변별력이 높다. 따라서 알고리즘 선정 후보로 이상치의 영향을 적게 받고, 높은 변별력을 가진 GBM, RF, XGBoost 를 실험 대상으로 선정하였다. 그 결과, 세 모델은 비슷한 성능을 보이며 대체로 GBM 이 근소한 차이로 좋은 성능을 보였다. GBM 모델의 경우, ntree 를 증가시켰을 때 점진적인 성능 향상이 일어나며, ntree 가 200 이상일 때에는 그 차이가 미미했다. RF 모델의 경우는 1000, XGBoost 는 50 일 때가 가장 성능이 좋았다.

## 6.2) 변수 선택방법

### 가. 후진제거법

모델링의 결과 변수 중요도를 내림차순으로 정렬하고 가장 최하위 중요도를 가지는 변수 한 개를 제거한다. 그리고 다시 모델링하는 과정을 반복 시행하는 후진제거법을 시행하였다. 초기 투입변수(234개)를 모두 투입하여, f1의 하락시점까지 최하위 중요도 변수를 제거하였다. 그 결과 64개의 변수로 확정시켰다. 우측 그림은 ntree 200의 GBM모델에서 전체 변수(234개)와 f1의 plot이다.




**[그림 12] 변수 개수와 f1 산점도**

| 연번 | 변수명 | 변수설명 | 비고 |
|---|---|---|---|
| 1 | AGE | 고객의 연령 | |
| 2 | ARPU | 월기준 회선당 평균 수익금 | |
| 3 | AUTR_FAIL_MCNT_LOG | 총 자동이체실패월수 | 로그변환 |
| 4 | AVG_CALL_FREQ_LOG | 월평균 통화횟수 | 로그변환 |
| 5 | AVG_CALL_TIME_LOG | 월평균 통화시간 | 로그변환 |
| 6 | AVG_OCCR | 신용대출, 신용카드 유지개설 기간 평균 | |
| 7 | AVG_OCCR2 | 신용대출 유지개설 기간 평균 | |
| 8 | BNK_GUIF_AMT_LOG | 신용대출 외 대출금액 | 로그변환 |
| 9 | BNK_LNIF_AMT_AVG_LOG | 1회 은행대출 평균금액 | 로그변환 |
| 10 | BNK_LNIF_AMT_PREM | 전체 대출금액 중 은행대출비율 | |
| 11 | CB_GUIF_AMT_LOG | 총 보증 금액 | 로그변환 |
| 12 | CNT_SCORE_AVG | 총 건수별 연체비율 합의 평균 | |
| 13 | CPT_LNIF_AMT | 카드사에서 발생한 총 대출 금액 | |
| 14 | CPT_LNIF_AMT_AVG_LOG | 카드사에서 1회 대출 평균 | 로그변환 |
| 15 | CPT_LNIF_AMT_PREM | 전체 대출금액 중 카드사 대출 | |
| 16 | CPT_LNIF_CNT | 카드사에서 발생된 총 대출 건수 | |
| 17 | CPT_LNIF_CNT_PREM | 전체대출건수 중 카드사 비율 | |
| 18 | CPT_TOT_CNT_PREM | 은행외의 대출 중 카드사비율 | |
| 19 | CRDT_CARD_CNT | 신용카드 발급 수 | |
| 20 | CRDT_OCCR_MDIF | 신용대출 계좌 유지 개월 수 | |
| 21 | CRMM_OVDU_AMT | 해당월 납부요금의 연체금액 | |
| 22 | CTCD_OCCR_MDIF | 신용카드 유지 개월 수 | |
| 23 | CUST_GUIF_AMT_PREM_LOG | 보증금액/본인수입 | 로그변환 |
| 24 | CUST_JOB_INCM | 직업정보기반 추정 소득 금액 | |
| 25 | DIFF_OVDU_RATE_LOG | 보험료연체율 - 최근1년보험료연체율 | 로그변환 |
| 26 | ECT_LNIF_AMT_LOG | 은행, 카드사이 외 대출금액 | 로그변환 |
| 27 | ECT_TOT_AMT_PREM_LOG | 전체 대출금액 중 기타 대출 | 로그변환 |
| 28 | FMLY_CLAM_CNT_LOG | 가계 합산 총 보험금청구 건수 | 로그변환 |
| 29 | FMLY_PLPY_CNT_LOG | 가구단위 만기까지 보험료 완납 갯수 | 로그변환 |
| 30 | FMLY_TOT_PREM_LOG | 가계 합산 유효계약의 총납입보험료 | 로그변환 |
| 31 | FYCM_PAID_AMT_LOG | 가계 합산 보험금지급 총액 | 로그변환 |
| 32 | HSHD_INFR_INCM | 가계 합산 추정 소득 | |
| 33 | INCM_LNIF_RATE_COU_LOG | 개인 + 배우자 소득대비 부채비율 | 로그변환 |

| | | | |
|---|---|---|---|
| 34 | INCM_LNIF_RATE_FAM_LOG | 가구소득대비 부채비율 | 로그변환 |
| 35 | INCM_LNIF_RATE_IND_LOG | 개인소득대비 부채 비율 | 로그변환 |
| 36 | JOB_RATE | 본인과 직업패턴기준 연체자비율 | |
| 37 | LINE_STUS | 회선상태 | |
| 38 | LT1Y_CTLT_CNT_per | 최근1년 실효해지건수 별 연체자비율 | |
| 39 | LT1Y_MXOD_AMT | 최근1년 연체금액 中 최대 연체금액 | |
| 40 | LT1Y_PEOD_RATE | 최근1년 연체납입횟수/총납입횟수*100 | |
| 41 | MATE_OCCP_NAME_G | 배우자 직업정보 | |
| 42 | MAX_MON_PREM_LOG | 납입한 월납입보험료 中 최대보험료 | 로그변환 |
| 43 | MOBL_FATY_PRC | 사용중인 핸드폰단말기 출고가액 | |
| 44 | MOBL_PRIN_LOG | 남아있는 핸드폰 단말기 할부원금 | 로그변환 |
| 45 | MON_TLFE_AMT_LOG | 월기준 서비스 납부요금 | 로그변환 |
| 46 | OCCP_NAME_G | 고객 직업정보 | |
| 47 | OTHR_INCM | 직업수입 외 추정소득 | |
| 48 | OVDU_CNT | 연체관련변수 Y/N변환한 값들의 합 | |
| 49 | PAYM_METD | 납부요금의 납부 방법 | |
| 50 | PREM_OVDU_RATE_LOG | 납입횟수 중 연체횟수의 비율 | 로그변환 |
| 51 | SEX | 고객의 성별 | |
| 52 | SPART_ECP_CPT_CNT | 카드사를 제외한 제2산업분류 대출건수 | |
| 53 | SPART_LNIF_CNT | 2산업분류에서 발생된 총 대출 건수 | |
| 54 | SPTCT_OCCR_MDIF | 2산업분류에서 신용대출 계좌유지 | |
| 55 | ST_LT_CRDT_GRAD | 최초신용등급, 최근신용등급 묶음 | |
| 56 | TEL_CNTT_QTR_LOG | 가입년월_분기 | 로그변환 |
| 57 | TEL_CUS_MDIF_LOG | 현재부터 가입년월 차이 | 로그변환 |
| 58 | TEL_MBSP_GRAD | 멤버십등급 | |
| 59 | TOT_CLIF_AMT_LOG | 대출정보 현재 총 금액[신용대출] | 로그변환 |
| 60 | TOT_CLIF_AMT_PREM | 전체 대출금액 중 신용대출 금액 | |
| 61 | TOT_LNIF_AMT_LOG | 산출일 기준 총 대출 금액 | 로그변환 |
| 62 | TOT_LNIF_CNT | 총 대출건수 | |
| 63 | TOT_MTIF_AMT_PREM_LOG | 전체 대출금액 중 신용대출 외 금액 | 로그변환 |
| 64 | TOT_PREM_LOG | 유효한 계약의 총납입보험료 | 로그변환 |

**[표 4] 최종 선택된 변수 목록**

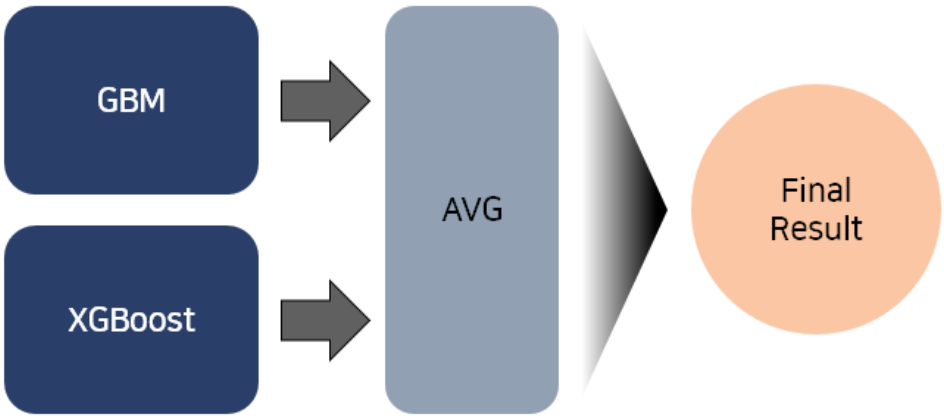
모델별 최적의 변수셋을 찾기 위해 기본 변수셋(261개), 112개 변수셋, 64개 변수셋으로 성능 비교를 하였다. Threshold 설정방법은 10-CV의 시도별로 최대의 f1을 갖는 Threshold의 평균값이다. 판단 지표의 우선순위는 f1, Recall, Accuracy순이다. Accuracy보다 Recall과 f1이 높아야 하는 이유는 연체자를 비연체자라고 했을 때 발생하는 손실을 최소화하기 위함이다. 분석결과, Recall과 f1이 가장 좋은 것은 각 모델에서 112개 변수셋이다. Accuracy는 GBM에서 기본 변수셋이, XGBoost에서 64개 변수셋이 가장 높았다. 본 팀은 위의 판단 기준에 따라 112개 변수셋을 최종 선정하였다.

| 기계학습 알고리즘 | 변수 개수 | Accuracy | Recall | Precision | f1 |
|---|---|---|---|---|---|
| GBM | 기본 | 0.951739 | 0.529246 | 0.449788 | 0.484349 |
| | 112 | 0.949942 | 0.558774 | 0.439441 | 0.488991 |
| | 64 | 0.950866 | 0.534727 | 0.444216 | 0.483164 |
| XGBoost | 기본 | 0.948971 | 0.532775 | 0.432782 | 0.473116 |
| | 112 | 0.949708 | 0.545643 | 0.434693 | 0.481118 |
| | 64 | 0.950277 | 0.530353 | 0.437781 | 0.477345 |

**[표 5] 변수 개수 별 성능비교**

### 6.3) 최종 선정모델 및 성능향상

시드에 따른 알고리즘별 예측력의 편차가 심했다. 이를 보완하는 방법으로써, 세 가지 모형 중 가장 우수한 성능을 보이는 GBM과 XGBoost의 앙상블 모형을 최종 선정하였다. 각 모형별로 앞선 실험 결과에 따른 최적의 트리 개수로서 GBM에서 200개, XGBoost에서 50개를 적용하였다. 그리고 앞선 변수 개수 별 성능비교에서 발견된 최적값인 112개의 변수셋을 적용하였다. 결과적으로, 예측의 변동성을 줄이고, 변별력을 안정적으로 향상시켰다. f1값을 약 0.1~0.2 향상시켰다.




**[그림 13] 최종 선정모델 흐름도**

| | XGBoost | GBM | 앙상블 (p>0.45) |
|---|---|---|---|
| F1 | 46.9874 | 47.8625 | 49.0433 |

**[표 6] 앙상블을 통한 성능개선 표**

# 3 결론

## 1. 요약

본 알고리즘은 대출연체를 예측하기 위해, 도메인 지식에 바탕을 둔 65개의 파생변수와 모델성능 향상을 위한 다양한 방법을 시도하여 개발되었다. 초기 성능인 0.3 수준의 f1 알고리즘을 0.4 중후반 수준까지 올리는데 성공하였다. 최적의 모델 선정을 위해 여러 알고리즘의 성능을 비교해보았다. 모델 성능 향상을 위해 변수 개발 및 선택과 Hyper Parameter 최적화를 위한 다양한 실험을 하였다.

기본 제공 데이터는 0이나 결측치가 매우 많아 분석으로의 활용에 어려움이 있었다. 이에 본 팀은 이상치 제거, 결측치 대체, 로그 변환 등 데이터 전처리에 많은 노력을 기울였다. 총 132개의 변수 중에서 선별력 있는 변수들만을 선정하고자 하였다. 이 과정에서 후진제거법, 모델 내 변수 분포 및 성능에 따라 최종적으로 112개의 변수를 선정하였다. 선정된 112개의 변수를 활용하여 모델 선정의 과정을 밟았다. 모델선정의 후보로는 GBM, XGBoost, Random Forest 세 가지 학습 모형이 있었다. 다양한 모델 적합시도 결과, GBM과 XGBoost의 앙상블이 가장 변동성이 낮으면서도 뛰어난 성능을 보였다.

## 2. 한계 및 향후 과제

### 2.1) 한계점

대출연체 및 상환예측 알고리즘을 개발하기 위해 본 팀은 제공된 67개의 데이터로부터 65개의 파생변수를 개발하였다. 기계학습 알고리즘은 Random Forest , GBM, XGBoost 에 대해 다양한 파라미터 값 조정실험을 진행했다. 최종적으로 112개의 변수와 GBM과 XGBoost의 평균값을 활용한 모델을 사용했다. 해당 모델의 성능을 개선하기 위해서 연체예측의 대상을 분리해서 모델을 만든다면 더 좋은 성능이 기대된다는 가설을 세우고 진행했다.

| |
|---|
| 첫째, 생애주기에 따라 대출 연체에 영향을 미치는 변수들이 다를 것이다.<br>둘째, 저축성 보험을 든 사람과 들지 않은 사람이 대출연체에 영향을 미치는 요인이 다를 것이다.<br>셋째, 발급받은 카드 수에 따라 연체를 결정하는 요인이 다를 것이다.<br>넷째, 소득분위에 따라 연체를 결정하는 요인이 다를 것이다. |

**[표 7] 대출 연체자에 대한 가설**

가설의 결과는 다음과 같다. 카드 발급 수 2개 이하의 대상으로 만든 모델은 f1이 0.52~0.53이 나왔고, 핸드폰 연체경험이 있는 고객 대상으로 만든 모델도 f1이 0.52~0.53이 나왔다. 반면 나이가 60대 이상의 고객들을 대상으로 만든 모델은 f1이 0.22~0.25로, 성능이 떨어졌다. 이외의 가설도 분리된 데이터 셋을 만들고 기존모델과 비교하여 검증했지만, 성능개선을 기대했던 것과는 상반되게 보다 낮은 성능을 보였다. 본 팀은 이 결과가 해당 데이터셋에서 연체자 비율이 많았기 때문에, 제공된 데이터 셋에서만 나타나는 결과로 결론지었다.

대출연체 및 상환예측 알고리즘의 성능이 낮게 나온 주요 원인 중 하나는 4,273명의 연체자 데이터가 전체 대출 연체자를 예측하기에 한계가 있기 때문이다. 연체자 데이터가 보다 충분하였다면 알고리즘의 성능이 더 높을 것으로 기대된다.

## 2.2) 발전 방향

발전 방향으로서 대출연체 및 상환예측 알고리즘 개발에 있어서 추가되어야 데이터를 제안하고, 해당 데이터의 접근방법을 제시하겠다.

| |
|---|
| 첫째, 대출의 목적이다.<br>학비를 위해 학자금 대출을 받는 학생과 돈을 벌기 위해 사업을 시작하는 학생의 대출은 다르다. 대출의 목적은 대출연체에도 큰 영향을 줄 것이라 보여져 수집의 필요성이 있다.<br>둘째, 대출의 시점이다.<br>최근에 대출을 여러 번 받았다면, 대출연체의 가능성이 높아진다는 가설을 세워 접근할 수 있다. 대출의 시점 또한 대출연체에 영향을 줄 것이라 보여진다.<br>셋째, 최근/최초 신용등급의 정보이다.<br>현재 신용에 따라 연체발생의 요인이 다를 것이라는 가설을 세워 접근할 수 있다. 주어진 데이터는 최근 신용등급이 90%가 누락되어 있다. 대출상품의 수요가 있기 때문에, 신용에 따른 대출연체 알고리즘이 개발되기 위해서는 질 높은 수준의 최근 신용등급의 데이터가 있어야한다. |

**[표 8] 발전 및 개선 가능방안 제시**

본 팀은 제공된 데이터를 기반한 대출연체 및 상환 예측 알고리즘 개발결과로서 0.46~0.50의 f1스코어를 기대한다. 위에서 제안한 데이터가 추가된다면 더 나은 성능을 보이는 대출연체 및 상환예측 알고리즘을 개발 할 수 있을 것이다.