‘Give Me Some Credit’ 분석 보고서

201511508 통계학과 김민철

201611539 통계학과 하성진

부산대학교 통계학과
PUSAN NATIONAL UNIVERSITY

I. 분석목적

2007년 세계 금융 위기가 발발하고 미국의 수많은 금융기관과 가구들이 파산했다. 특히 경기의 악화로 인해 투자 심리는 얼어붙고 금융기관들은 정부의 구제금융 정책에도 불구하고 트라우마에 빠져 대출의 진입장벽을 높여 시중 유동성을 악화시켰다. 2011년 당시 신용경색이 발생한 미국 경제의 상황이 나아지기 위해서는 시민들을 통한 유동성 확보가 절실했고 이 역할을 수행해야 할 은행은 파산할 고객을 분류해 대출을 실행해야 한다. 이를 위해 머신러닝을 통한 예측을 수행하고자 하였다.

II. 분석 데이터

1. 데이터 개요

각 행당 개인 고객 관련 정보를 담고있는 데이터로써, 고객의 파산여부('SeriousDlqin2yrs') 관련 정보가 있는 Train data와 관련 정보가 없는 Test data로 구성되어 있다. Train data는 12개의 변수로 이뤄진 15만개 샘플로 구성됐고, Test data의 경우 12개의 변수로 구성된 101503개의 샘플로 이뤄졌다. 12개의 변수 중에서 7개의 변수는 해석에 유의할 부분이 있으므로 추가 설명이 필요했다.

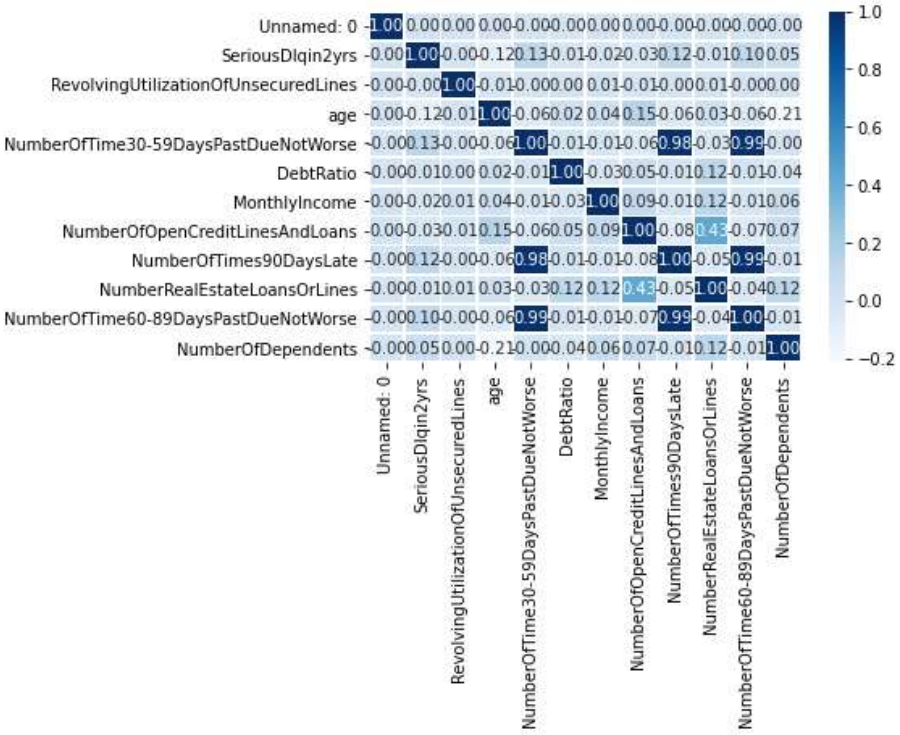
2. 기초 통계량 확인

✓ 분포 및 이상치 확인

	Unnamed: 0	SeriousDlqin2yrs	RevolvingUtilizationOfUnsecuredLines	age	NumberOfTime30-59DaysPastDueNotWorse
count	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000
mean	75000.500000	0.066840	6.048438	52.295207	0.421033
std	43301.414527	0.249746	249.755371	14.771866	4.192781
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	37500.750000	0.000000	0.029867	41.000000	0.000000
50%	75000.500000	0.000000	0.154181	52.000000	0.000000
75%	112500.250000	0.000000	0.559046	63.000000	0.000000
max	150000.000000	1.000000	50708.000000	109.000000	98.000000

기초 통계량을 통해 데이터의 분포를 가늠하고 이상치의 유무를 확인.

✓ 상관관계

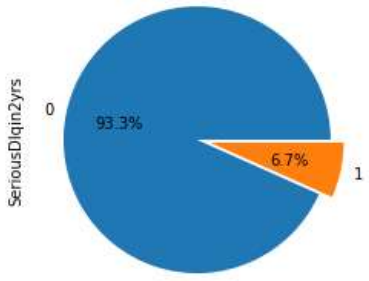


변수간 상관관계가 높은 세 변수가 존재하나 나머지는 낮은 상관관계를 보이기에 다중공선성 문제가 심하지 않을 것이라 판단했다.

3. 데이터 변수 탐색

✓ SeriousDlqin2yrs

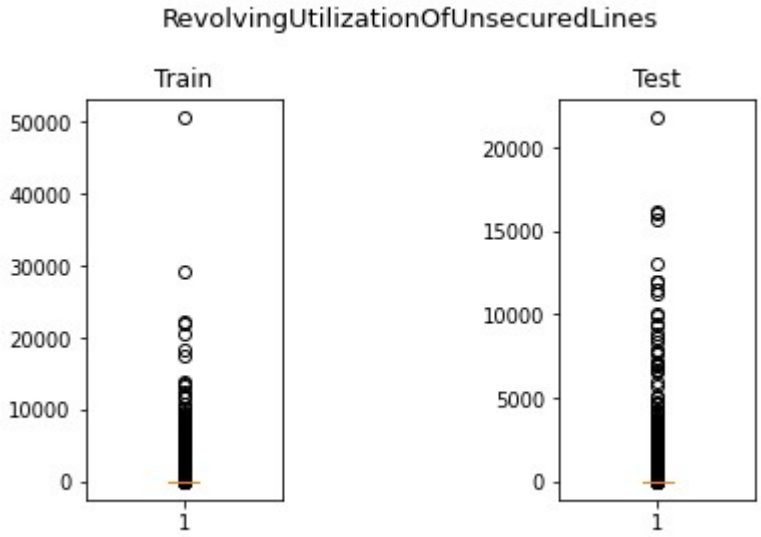
target 변수, 2년 내에 초과 연체 여부



전체 Train 데이터중 약 6.7%만이 2년 내에 연체하는 것으로 확인된다. 주의할 점은 모델을 학습시킬 때 연체를 안 하는 경우에 대해서 많이 학습하게 됨으로 resampling을 고려할 필요가 있다.

✓ RevolvingUtilizationOfUnsecuredLines

신용한도에서 빌릴 수 있는 돈의 한도에서 실제로 빌린 돈의 비율.



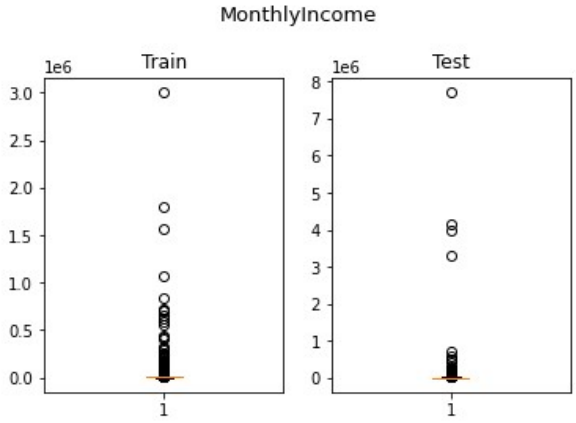
분포에서 떨어진 한, 두 개의 이상치들이 보인다.

✓ DebtRatio

총 월소득에서 갚아야 하는 빚의 비율.

✓ MonthlyIncome

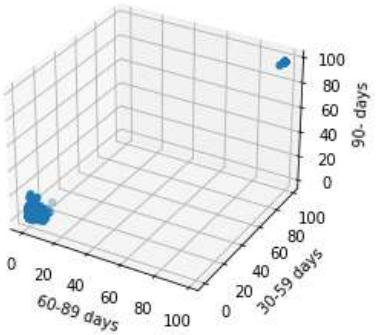
Debt을 제외한 실제 소득을 의미. 위의 RevolvingUtilizationOfUnsecuredLines(앞으로 RU라고 하겠다.) 변수와 DebtRatio(이 변수도 DR로 축약하겠다.) 변수하고 같은 맥락의 의미를 지니지 않을까 하는 생각이 든다. 개인이 좋은 직장을 가졌거나 자산을 가지고 있다면 변수 RU와 DR은 수치가 상대적으로 낮게 나오고 MonthlyIncome 수치가 높게 나올 것이다. 하지만 세계 금융위기로 인한 실물 경제 악화(고용률 감소)와 더불어 부의 양극화가 심화된 상황에서 MonthlyIncome 수치가 낮고 학력이 낮은 사람들은 앞으로도 이전보다 소득이 높은 직장을 가지기 힘들 것이며 파산의 위험성 또한 높아질 것이다.



✓ 변수 RU와 마찬가지로 한, 두 개의 이상치들을 확인할 수 있으며 나머지 변수들에서도 이와 비슷한 결과가 나왔다.

✓ NumberOfTime **N** DaysPastDueNotWorse

N일 이상 연체한 횟수를 의미하는 변수. 현재 데이터에서 제시된 N의 종류는 30~59, 60~89, 90~로 총 3가지의 변수가 들어있다. 데이터의 특징이 있다면 각 변수에서의 최대값과 그 다음 최대값의 차이가 큼에도 불구하고 그 사이에 아무 값이 없는 것이다. 신용경색의 심화로 인해 유동성 압박에 직면한 일반 사람들은 이 수치가 더욱 높게 나올 것으로 예상된다.



NumberOfTime(N)DaysPastDueNotWorse 변수에서 이상치를 가진 고객들은 모두 일수(30-59 일, 60-89 일, 90 일 이상)에 관계없이 모든 종류의 일수에서 100 근처의 이상치를 가지고 있음을 확인할 수 있다. 이는 Test 와 Train 데이터 모두에서 나타나고 있다.

✓ NumberofDependents

배우자나 아이 등 부양 가족의 수를 의미. 최근의 전세계적으로 1인 가구의 수가 증가되고 있다. 데이터 내에서도 가장 많은 부양 가족 수는 0이었다. 이 변수에 대해서 해석을 주의할 점이 있다. 파산한 데이터에서 가장 많은 부양 가족 숫자는 0이었다. 그리고 부양 가족이 0인 경우의 수는 60대의 비율이 가장 높았다. 하지만 실제로 파산한 데이터의 연령 분포를 보면 20대부터 40대 사이가 주를 이루고 있다. 이를 통해 부양 가족이 없는 사람이 파산할 가능성이 높다는 해석보다는 가족을 부양할 수 있을 정도의 경제력이 없

는 20대부터 40대의 세대들이 파산할 확률이 높을 것이라는 생각을 할 수 있다.

✓ NumberRealEstateLoansOrLines

home equity lines를 포함한 부동산 담보 대출이나 모기지. 당시 금융위기 이후, 금융 트라우마에 빠진 금융기관들은 대출에 있어 신중한 자세를 취하고 있었다. 그럼에도 불구하고, 대출 횟수가 높은 샘플들이 있었는데, 이는 시대적 배경과 무관하게 금융기관으로부터 여러 번의 대출을 받을 수 있을 정도의 고액 자산들을 보유한 사례로 예상된다. 따라서, 이들은 파산위험성이 다른 사람에 비해 낮을 것으로 예측된다.

✓ NumberOfOpenCreditLinesAndLoans

개인이 이용할 수 있는 자금 동원 횟수. 여기에는 신용카드, 마이너스 통장, 단기 담보 대출 등, 이미 금융사로부터 보증을 받아 신속하게 자금을 동원할 수 있는 방법을 의미할 것으로 예측된다.

4. 데이터 전처리

(1) 결측지 보정

✓

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 146250 entries, 0 to 149999
Data columns (total 12 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   Unnamed: 0                            146250 non-null  int64
 1   SeriousDlqin2yrs                      146250 non-null  int64
 2   RevolvingUtilizationOfUnsecuredLines  146250 non-null  float64
 3   age                                   146250 non-null  int64
 4   NumberOfTime30-59DaysPastDueNotWorse  146250 non-null  int64
 5   DebtRatio                             146250 non-null  float64
 6   MonthlyIncome                         146250 non-null  float64
 7   NumberOfOpenCreditLinesAndLoans       146250 non-null  int64
 8   NumberOfTimes90DaysLate               146250 non-null  int64
 9   NumberRealEstateLoansOrLines          146250 non-null  int64
 10  NumberOfTime60-89DaysPastDueNotWorse  146250 non-null  int64
 11  NumberOfDependents                    142556 non-null  float64
dtypes: float64(4), int64(8)
memory usage: 14.5 MB
```

그림 1. Train data 결측치 수 확인 결과

결측치 확인 결과, 'MonthlyIncome' 과 'NumberOfDependents' 총 2개의 변수에서 결측치가 존재함을 확인할 수 있었다. MonthlyIncome 변수의 경우, 이상치가 가중되어 영향이 끼칠 수 있으므로, 이와 무관한 중의수로 결측치를 대체하였다. 그리고, NumberOfDependents 변수의 경우, 데이터 분포를 확인해본 결과, 절대 다수의 데이터가 0으로 구성되어 있어, 결측치를 최반값인 0으로 대체하였다.

(2) 이상치 보정

	NumberOfTime30-59DaysPastDueNotWorse
0	126018
1	16033
2	4598
3	1754
4	747
5	342
6	140
7	54
8	25
9	12
10	4
11	1
12	2
13	1
96	5
98	264

	NumberOfTime60-89DaysPastDueNotWorse
0	142396
1	5731
2	1118
3	318
4	105
5	34
6	16
7	9
8	2
9	1
11	1
96	5
98	264

	NumberOfTimes90DaysLate
0	141662
1	5243
2	1555
3	667
4	291
5	131
6	80
7	38
8	21
9	19
10	8
11	5
12	2
13	4
14	2
15	2
17	1
96	5
98	264

그림 2. 각 변수의 수치의 개수

NumberOfTime**N**DaysPastDueNotWorse 변수의 경우, 분포에서 벗어난 이상치 그룹들을 각 변수들의 이상치 그룹을 제외한 데이터에서의 최댓값인 11, 13, 17로 대체하였다.

(3) 변수 추가 생성

예측에 도움이 될 것 같은 다양한 변수들을 추가했다.

✓ MonthlyIncomePerPerson

각 고객의 월별 임금을 본인을 포함한 부양가족수로 나누어, 고객의 부양 가족 한명당 실질적으로 생활하는데 사용되는 임금을 계산하고자 하였다.

✓ MonthlyDebt

각 고객의 월별 임금에 임금대비 부채 비율을 곱하여, 한 달마다 상환해야하는 부채를 계산하고자 하였다.

✓ isRetired

각 고객의 연령을 참고하여, 65세 이상은 직장에서 은퇴한 고객으로 분류하여 고객의 은퇴여부를 나타내고자 하였다.

✓ hasMultipleRealEstates

각 고객의 주택 및 토지 담보 대출 개수를 바탕으로, 주택이나 토지를 2개 이상 가진 고객을 구분해보고자 하였다.

III. 분석 절차 (예측모델 개발)

1. 분석 원리

✓ 회귀분석

회귀분석은 여러 설명변수를 사용하여 수량형 반응변수를 예측하기 위한 가장 간단하고 유용한 모형이다. 단, 반응변수가 독립이고 동이란 분포를 따른다는 가정이 필요하다.

✓ 의사결정나무

의사결정나무는 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 이때 몇 개의 소집단(노드)로 분류될 때, 이 때 소집단내에선 동질성이 커지고, 소집단간에는 이질성이 가장 커지도록 선택된다. 이때, 카이제곱통계량[1] P값과 지니지수[2]가 모두 작아지는 방향으로 분류를 진행한다. 이 원리는 분석기법으로 활용된 랜덤포레스트, XGboost, lightGBM에 활용되었다.

✓ 배깅

동일한 알고리즘(ex.의사결정나무)로 여러 개의 분류기를 만들어서 투표를 통해 최종 결정하는 알고리즘이다. 분류기들에 가중치를 주어 선형결합을 통해 최종 결과를 예측하는 방법에 해당한다. 쉽게 말해 기본적인 의사결정나무와는 달리 부트스트랩[3] 방식을 통해 전체데이터에서 중복을 허용하여 n개의 데이터를 추출하고, 또한 피처값 중에서 중복 허용 없이 k개를 추출하여 만든 데이터를 바탕으로 분류기에 넣어 데이터를 예측한다. 이러한 과정을 여러 번 반복하여 여러 개의 분류기에서 나온 결과값을 바탕으로 예측을 진행한다.

✓ 부스팅

부스팅 알고리즘은 분포에 대해 약한 학습자를 반복적으로 학습시켜 최종적으로 강한 학습자를 만드는 것을 목표로한다. 강한 학습자를 만들기 위해 약한 학습자들의 잔차가 큰 잘못 예측한 데이터에 대해 가중치를 부여하여 예측성을 향상시키고자 한다. 의사결정나무를 기반으로 분석하는 부스팅 기법에는 LightGBM과 XGboost가 있으며, 타 부스팅 기법 대시 분석 소요 시간을 획기적으로 줄이면서도,

---

[1] 카이제곱통계량 P값이 작을수록 의사결정나무에서 자식노드 간의 이질성이 커짐

[2] 지니지수 값이 클수록 자식노드 내의 이질성이 커짐

[3] 통계 메트릭을 계산하기 전에 복원추출법을 적용하는 방법을 일컫음. 확률 변수의 정확한 확률분포를 모르는 경우나 측정된 샘플이 부족한 경우에 사용

예측력은 그대로 유지하고 있거나 더 뛰어나다.

2. 분석 방법

(1) 분석 방법 선정

의사결정나무 기반이자, 배깅 방식으로 분류기를 생성하여 진행하는 방식인 랜덤포레스트 방식과, 동일한 의사결정나무 기반이지만, 부스팅 기법을 활용해 높은 예측성 및 정확성을 가진 LightGBM과 XGboost를 활용하고자 하였다. 이외의 추가적으로, 분류의 대표적인 모델인, 로지스틱 회귀도 적용하였다.

(2) 하이퍼파라미터 선정

전처리한 데이터를 바탕으로, 분석 기법에 대한 하이퍼 파라미터를 조정하여 데이터에 대한 예측력을 향상시키고자 하였다. 고려한 파라메터 중 핵심 파라메터는 아래와 같다.

✓ Regularization parameter

Logistic regression 모델이 과다적합이 되는 것을 방지하기 위한 파라미터. Lasso, Ridge, C-value가 있으며 변수선택과 모델의 안전성을 높이기 위해 보통 Lasso를 사용하고 규제를 크게 하고 싶은 경우 C값을 낮춘다.

✓ learning_rate

LightGBM와 XGboost에서 사용되는 하이퍼 파라메터로써, 학습률또는 훈련량을 의미한다. 학습률은 초매개변수[4]를 조정하는 비율을 조정하는 값에 해당한다. 학습률이 낮을수록 경사하강법 내에서 최소값을 지나쳐버리는 경우 없이 정밀하게 최적값을 찾아가지만, 그만큼 실행속도가 느려서 많은 시간이 소요된다. 반대로, 학습률이 높을수록 실행속도는 빨라지나, 최소값을 지나쳐 최적 값을 놓칠 가능성이 있다.

✓ max_depth

랜덤포레스트, LightGBM, XGboost 모두 사용되는 하이퍼 파라메터로써, 의사결정나무의 깊이를 의미하며, -1 설정 시 제한 없이 분기하며, 자연수 n으로 설정하면 최대 n회 분기한다. 복잡한 데이터 사용시 n을 높일 필요가 있으며, 지나치게 높일 시, 과적합을 일으킬 가능성이 있다.

---

[4] 머신러닝 알고리즘에서 조정하는 값

✓ num_leaves

LightGBM와 XGboost에서 사용되는 하이퍼 파라메터로써, 의사결정나무에서 분기하여 최종적으로 가지는 최대 잎사귀 수를 의미한다. 의사결정나무는 여러 차례 분기를 통해 잎사귀를 늘려갈 수 있으며, 이를 작은 값으로 설정할 시 과적합을 방지할 수 있는 규제로 작용할 수 있다.

✓ num_iterations/nrounds

LightGBM와 XGboost에서 사용되는 하이퍼 파라메터로써, 의사결정나무 과정에서 몇 번의 부스팅 과정을 진행할지를 결정한다. 부스팅 과정을 많이 진행할수록 데이터에 대한 예측력은 높아지나, 과적합될 위험성을 배제할 수 없다.

✓ max_features

의사결정나무에서 사용되는 하이퍼 파라메터로써, 랜덤포레스트에서 배깅의 과정에서 선정할 최대 피처 개수를 의미한다. 최대 피처갯수가 전체 피처갯수에 가까워질수록 기존의 의사결정나무와 큰 차이를 보이지 않을 수 있으며, 최소한으로 제한할 경우 과소적합하는 문제가 발생할 수 있다.

(3) 분석 진행

시간적 효율성을 고려하여, 사이킷런 라이브러리에서 제공하는 GridSearchCV를 이용하여 각 기법 별 최적 하이퍼 파라미터를 선정하였다. 선정된 하이퍼 파라미터를 바탕으로 캐글에서 제공하는 Public Score[5]와 Private Score[6]를 함께 고려하여, 최적 기법을 선정하고자 하였다.

분석기법	하이퍼파라미터	값
RandomForest	max_depth	1,3,5,9,10,11,12,13
	max_features	1,3,5,10
	min_samples_split	1,2,3,4,5,10
	max_leaf_nodes	1,3,5,9,10,11,13
XGboost	learning_rate	0.1,0.5,1,3
	max_depth	-1,1,3,5,10
	n_estimators	1,3,5,10
	num_iterations	100,500,1000

[5] Test Data의 70%를 램덤추출하여 점수 산정

[6] Train Data의 30%를 랜덤추출하여 점수 산정

	random_state	1,3,5,10
LightGBM	learning_rate	0.1,0.3,0.5,0.8,1,3
	max_depth	-1,1,3,5,10
	n_estimators	1,3,5,10
	num_iterations	100,500,1000, 1500
	random_state	1,3,5,10

(표에서는 생략됐지만 Logistic regression모델도 학습시켰다. Grid-search의 결과 C-value는 0.9가 나왔는데 이는 규제를 덜하는 것이 모델의 최적학습에 이롭다는 의미다. 참고로 Logistic Regression에 규제를 많이 걸어야 할 때는 SVM 모델을 사용한다.)

IV. 분석 결과

그림 3. 제출한 케글 점수 데이터

(왼쪽부터, Private Score와 Public Score)

ROC_AUC를 기준으로 랜덤포레스트 모델에서 가장 높은 Private 점수 0.867이 나왔다. ROC_AUC 점수는 ROC 곡선에 대한 면적을 계산한 값으로, ROC 곡선은 임곗값의 변화에 따라 예측한 것 중에서 맞을 확률과 정답 중에서 맞춘 확률을 그래프에 나타낸 곡선이다. 값이 높다는 의미는 임곗값을 바꾸어도 모델이 예측 성능이 우수함을 의미한다.

여기서 ROC_AUC 점수는 1에 가까울수록, 파산할 가능성이 높은 사람을 잘 파악해내면서도, 파산할 것이라고 예측한 사람이 실제로 파산한 확률이 높은 것을 의미한다. 그러므로, 예측결과 ROC_AUC값이 0.867이 나온 랜덤포레스트 모델이 다른 모델들에 비해 상대적으로 더 높은 예측 성능을 가진 것을 의미한다.