

대출 상환 예측 머신러닝 모델 및 데이터 분석

경기도 '기본대출' 이용자의 대출 상환 가능 여부를 판단하는 머신러닝 모델과 대출
데이터 분석

2021.04.12 김지연 & 오세광

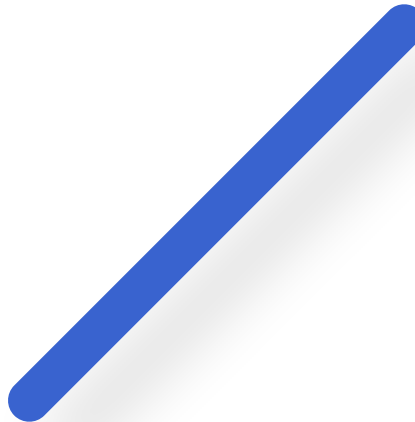
프로젝트2

TABLE OF CONTENTS

- 1 프로젝트 취지와 목적
- 2 데이터 소개
- 3 성능평가지표
- 4 가설 및 예상결과
- 5 EDA 진행 & 기본대출 분석
- 6 모델 소개
- 7 결과

1. 프로젝트 취지와 목적

- Home Credit 소개
- 기본대출 소개



1. 프로젝트 취지와 목적

Home Credit 데이터로 미리 살펴본 경기도 '기본대출' 시스템

Home Credit

비은행 금융 기관으로
기존의 은행 시스템에서
대출을 못받던
사람들에게
안전한 대출을 해준다.

경기도 기본대출

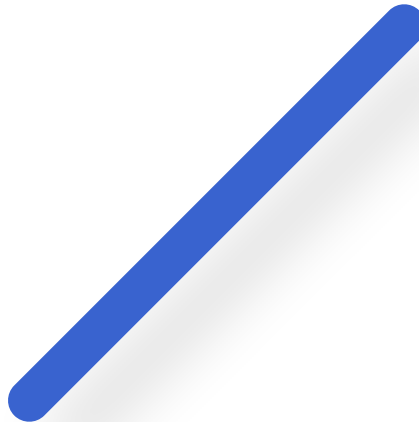
최대 1000만원을 연
이자 3%로 은행권이
제공하고자 하는 정책

데이터 분석 및 머신러닝 모델

Home Credit 데이터를
통해 기본대출 정책을
살펴보고 머신러닝 모델
구현하여 기대효과와
우려점 파악

2. 데이터 소개

- Home Credit 데이터 요약
- 타겟 설명



2. 데이터 소개

Home Credit 데이터 요약 : 120 여개의 컬럼 존재

금융관련 정보

대출 총액, 대출 종류,
수입액, 소득 종류 등

주거관련 정보

집 보유여부, 주거 현황,
거주지역 인구수 등

개인 정보

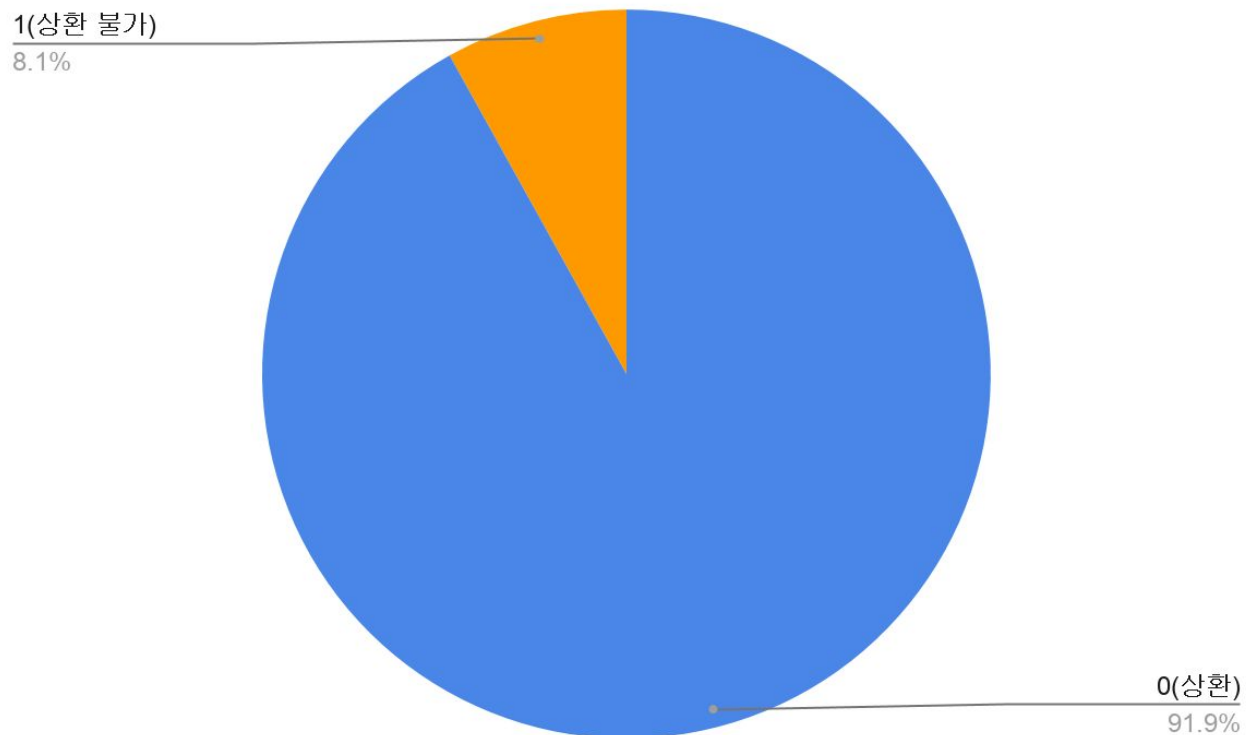
성별, 나이, 학력,
근로연수 등

외부 정보

외부 데이터 소스(신용)
점수,
고객 주변인의 연체 정보,
고객에 대한
신용평가사로의 문의 등

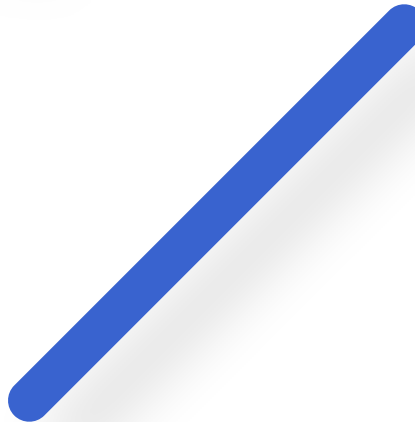
2. 데이터 소개

예측할 타겟(label) 설명 : 상환가능(0), 상환불가(1)



3. 성능평가지표

- Precision
- Recall
- F1 score



3. 성능평가지표

정밀도(Precision), 재현율(Recall), F1-score

CONFUSION MATRIX	ACTUAL	
	True Positive (TP)	<u>False Positive (FP)</u>
PREDICTED	<u>False Negative (FN)</u>	True Negative (TN)

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

3. 성능평가지표

정밀도(Precision), 재현율(Recall), F1-score

Precision

$$Precision = \frac{TP}{TP+FP}$$

모델이 True로 예측한 것 중
실제로 True인 것

예시:

상환가능(0)으로 예측한 것 중
실제 상환한 사람들의 비율

or

상환 불가능(1)으로 예측한 것
중 실제 상환하지 않은 사람들의
비율

Recall

$$Recall = \frac{TP}{TP+FN}$$

실제 True 중 모델이 True로
예측한 것

예시:

실제 상환한 사람들 중 상환가능
(0)할 것으로 예측한 비율

or

실제 상환하지 않은 사람들 중
상환 불가능(1)으로 예측한 비율

F1-score

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision과 Recall의 조화평균

4. 가설 및 예상결과

- 가설 및 예상결과



4. 가설 및 예상결과

중요한 특징 확인

EDA를 통해 대출상환여부와 관련된 특징을 살펴본 후, 대출상환여부와 높은 연관관계가 있는 특징을 확인한다.

타겟 불균형

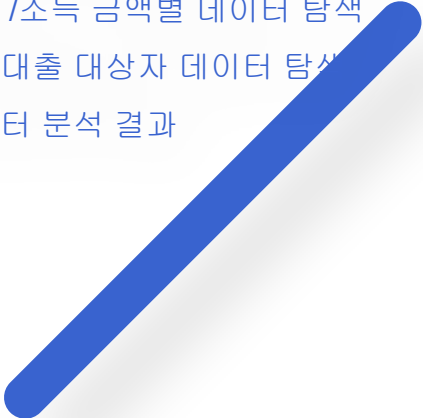
데이터의 타겟이 불균형하기 때문에, 이 문제를 해소할 수 있는 방법을 찾아서 성능을 개선한다.

모델의 성능을 개선

모델 성능을 개선하는 과정에서 모델지표 중 **recall** 값을 개선하여, 이용자들의 대출비율을 높이고, 그에 따른 손실액 예측을 높인다.

예상결과 : 타겟 불균형을 해소하고 주요 특징을 이용하여 베이스라인 모델의 **F1 score** 이상의 성능을 내는 모델을 구현할 수 있다.

5. EDA 진행 및 기본대출 분석

- 소득분위별 데이터 탐색
 - 연령대별 데이터 탐색
 - 소득분위/연령대별 데이터 탐색
 - 대출 /소득 금액별 데이터 탐색
 - 기본대출 대상자 데이터 탐색
 - 데이터 분석 결과
- 

5. EDA 진행 및 기본대출 분석

소득분위별 데이터 분석

- 소득분위별 상환여부 확인
- 전체비율 91.9%
- 1,2,3,4분위의 상환비율 비슷
- 소득이 가장 높은 분위의 경우만 2%P가량 차이남
- 소득분위별 상환여부를 판단하는 데 크게 의미 없었음.

```
income_rank  TARGET
1            0      0.917938
              1      0.082062
2            0      0.914117
              1      0.085883
3            0      0.913153
              1      0.086847
4            0      0.919431
              1      0.080569
5            0      0.934802
              1      0.065198
Name: TARGET, dtype: float64
```

5. EDA 진행 및 기본대출 분석

연령별 데이터 분석

- 평균상환비율(91.9%)에 비해
- 20대의 경우, 3.4%P가량 (88.5%) 떨어짐.
- 30대의 경우, 1.5%P가량 (90.4%)떨어짐.
- 20~30대의 상환율이 평균보다 떨어짐

age_group	TARGET	
20	0	0.885431
	1	0.114569
30	0	0.904165
	1	0.095835
40	0	0.923492
	1	0.076508
50	0	0.938703
	1	0.061297
60	0	0.950786
	1	0.049214

Name: TARGET, dtype: float64

5. EDA 진행 및 기본대출 분석

소득분위별, 연령별 데이터 분석

- 20~30대의 경우
소득분위가 낮을 수록
상환율이 떨어짐을 알
수 있음.
- 20대의 경우,
- 1분위 4.7%P(87.2%),
2분위 4.3%P(87.6%),
3분위 3.9%P(88.0%)
- 30대의 경우,
- 1분위 2.8%P(89.1%),
2분위 2.1%P(89.8%),
3분위 2.6%P(89.3%)

income_rank	age_group	TARGET	
1	20	0	0.872648
		1	0.127352
	30	0	0.891724
		1	0.108276
	40	0	0.917444
		1	0.082556
	50	0	0.942062
2		1	0.057938
	60	0	0.950395
		1	0.049605
	20	0	0.876351
		1	0.123649
	30	0	0.898935
		1	0.101065
3	40	0	0.919001
		1	0.080999
	50	0	0.936548
		1	0.063452
	60	0	0.950638
		1	0.049362
	20	0	0.880092
		1	0.119908
	30	0	0.893202
		1	0.106798
	40	0	0.918872
		1	0.081128
	50	0	0.940186
		1	0.059814
	60	0	0.946544
		1	0.053456

5. EDA 진행 및 기본대출 분석

대출금액에 따른 데이터 분석

- 대출액이 50,000이하일 경우
- 상환비율 4%P(95.9%) 높음
- 대출액이 100,000이하일 경우
- 상환비율 2.5%P(94.4%) 높음
- 대출금액이 낮을 때 상환비율이 높아짐

50,000이하 대출자들 상환비율

0 0.959002

1 0.040998

Name: TARGET, dtype: float64

100,000이하 대출자들 상환비율

0 0.94487

1 0.05513

Name: TARGET, dtype: float64

전체 대출자들 상환비율

0 0.919271

1 0.080729

Name: TARGET, dtype: float64

5. EDA 진행 및 기본대출 분석

소득(수입)금액에 따른 데이터 분석

- 수입액이 30,000이하일 경우
- 상환비율 94.3%
- 수입액이 40,000이하일 경우
- 상환비율 91.6%
- 수입액이 50,000이하일 경우
- 상환비율 92.4%
- 수입액이 100,000이하일 경우
- 상환비율 91.7%
- 수입액에 따른 상환비율이 증가 혹은 감소가 일정하지 않음.

```
수입 $30,000이하 대출자들 상환비율
0      0.943262
1      0.056738
Name: TARGET, dtype: float64
수입 $40,000이하 대출자들 상환비율
0      0.916926
1      0.083074
Name: TARGET, dtype: float64
수입 $50,000이하 대출자들 상환비율
0      0.924065
1      0.075935
Name: TARGET, dtype: float64
수입 $100,000이하 대출자들 상환비율
0      0.917972
1      0.082028
Name: TARGET, dtype: float64
전체 대출자들 상환비율
0      0.919271
1      0.080729
Name: TARGET, dtype: float64
```

5. EDA 진행 및 기본대출 분석

기본대출(금융) 정책 대상자 따른 데이터 분석

- 1000만원 이하 한도 내
대출을 금리 3%이하로 가능
- 만 25~26 or
결혼적령기대상 (남성 만
33~34세, 여성 만29~30세)
시범운영 계획 중
- 만 25~26세 경우
- 상환비율 2.2%P (89.7%)
떨어짐
- 만33~34세 남성 경우,
- 상환비율 3.8%P(88.1%)
떨어짐
- 만29~30세 여성 경우,
- 상환비율 2.2%P(89.7%)
떨어짐
- 전반적으로 시범운영
대상자의 상환비율이
떨어져 손실우려 있음

만25~26세

0 0.897354

1 0.102646

Name: TARGET, dtype: float64

결혼적령기 남성

0 0.881746

1 0.118254

Name: TARGET, dtype: float64

결혼적령기 여성

0 0.897167

1 0.102833

Name: TARGET, dtype: float64

5. EDA 진행 및 기본대출 분석

데이터 분석 결과 정리

- 소득분위별 상환비율은 크게 차이가 없었음.
- 20~30대의 상환비율이 모든연령대를 포함한 평균상환율보다 떨어짐(20대 3.4%P, 30대 1.5%P)
- 소득분위가 낮을 수록 20~30대의 상환비율이 떨어짐.
- 대출금액이 낮을수록 상환비율이 높아짐
- 소득액에 따른 상환비율은 비례하지 않음.
- 기본대출 예상대상자의 만 25~26세, 만33~34세 남성, 만29~30세 여성
- 각각 상환비율 (2.2%P, 3.8%P, 2.2%P) 떨어짐

6. 모델 소개

- XGBoost
- Hyper Parameter
- F1 score
- 모델 성능 비교
- 기본대출 대상자



6. 모델 소개

XGBoost 모델 및 성능

Model

XGBoost Classifier
: Gradient Descent
알고리즘을 이용한 모델,
Hyper Parameter의 튜닝이
중요하다.

Hyper Parameter

타겟 불균형의 해소를
위한 Hyper Parameter
사용
(scale_pos_weight)

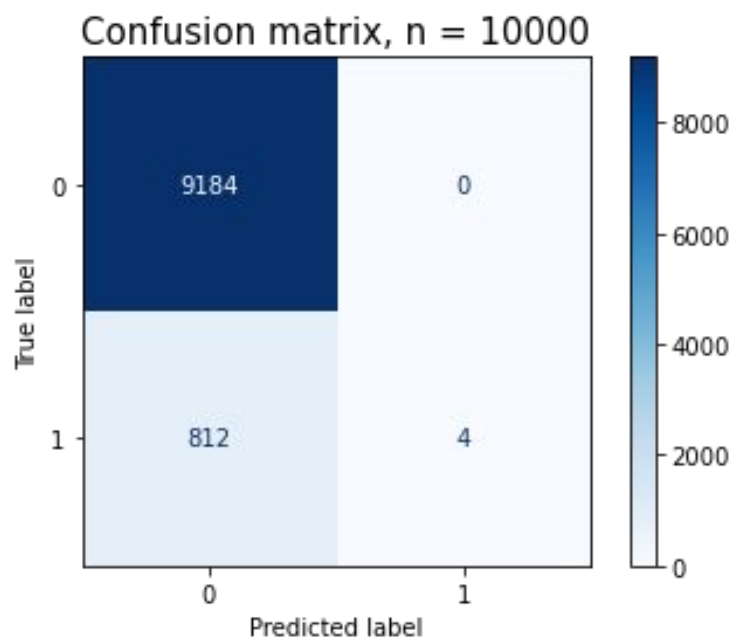
Score

	F1: 0	F1: 1
Baseline	0.96	0.01
XGBoost	0.83	0.27

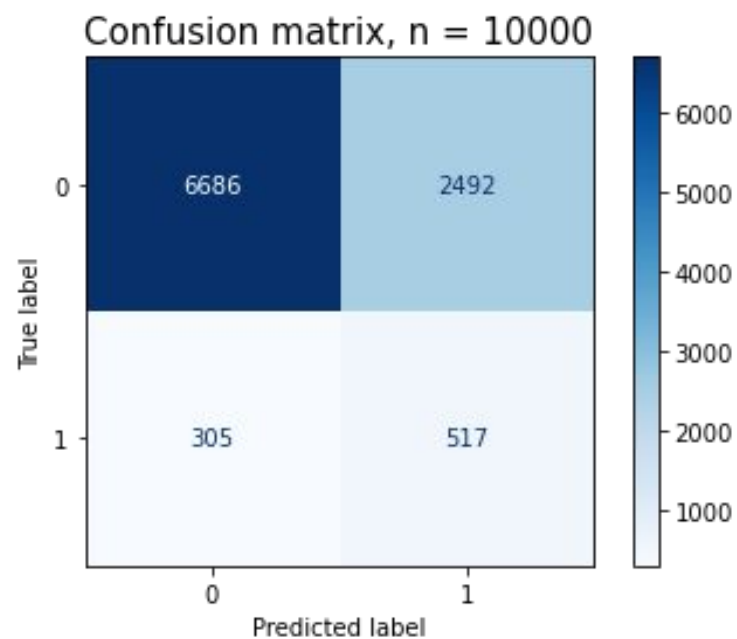
6. 모델 소개

모델 성능 비교

Baseline Model



XGBoost Model



6. 모델 소개

기본대출 대상으로 상환가능여부 예측하였을 때

Baseline Model

상환가능 인원예측(TP):
 $9184 \text{명} * 1000 \text{만원} * 3\%$
 이자율
 = 27억 5천만원(수익)
 상환 불가능 인원예측(FN):
 $812 \text{명} * 1000 \text{만원}$
 = 81억 2천만원(손실)

XGBoost Model

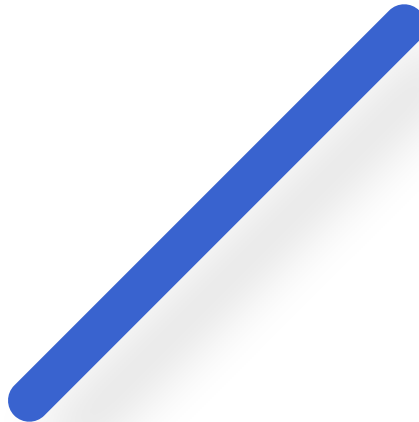
상환가능인원예측(TP):
 $6686 \text{명} * 1000 \text{만원} * 3\%$
 이자율
 = 20억(수익)
 상환 불가능 인원(FN) 예측 :
 $305 \text{명} * 1000 \text{만원}$
 = 30억 5천만원(손실)

Compare

	수익	손실	손익액
Baseline	27.5억	81.2억	-53.7억
XGBoost	20억	30.5억	-10.5억

7. 결과

- 가설 및 예상결과와 비교



7. 결과

첫 번째

F1 score 개선 없음

두 번째

0에 대한 F1 score는 감소
1에 대한 F1 score는 증가

세 번째

상환 가능(0)에 대한
recall값이 감소
상환 불가능(1)에 대한
recall 값은 증가

예상결과 : 타겟 불균형을 해소하고 주요 특징을 이용하여
베이스라인 모델의 **F1 score** 이상의 성능을 내는 모델을 구현할 수 있다.

결과 : 상환 가능(0)에 대한 예측 성능은 감소, 상환 불가능(1)에 대한 예측 성능은 증가,
상환이 불가능한 경우를 잘 예측하게 됨에 따라 대출 부실율이 감소할 것이다.

7. 결과

표본이 1만명일 때,
상환 불가능(1)에 대한 **recall**값을
개선함으로써, **43.2억**의 손실을 줄일 수
있다.

감사합니
다!