2024 MEDICI FINAL PROJECT.

대출 미상환 예측: 위험한 대출자를 찾아라

P2P 대출 상환 예측을 위한 AI 모델 개발

Team. GeekGoing
하요한 . 곽용현. 박규리. 정혜영

팀 소개

GEEK GOING

팀장 하요한	**팀원** 곽용현	**팀원** 박규리	**팀원** 정혜영
EDA	EDA	EDA	EDA
데이터 전처리	데이터 전처리	데이터 전처리	데이터 전처리
모델링	모델링	모델링	모델링
프로젝트 총괄	대시보드 구현	발표자료 제작	발표자료 제작

인사이트 도출, 비즈니스 전략 수립

분석 환경

python

시각화 및
데이터 분석

NumPy

pandas

matplotlib

seaborn

Shap

Folium

머신러닝

scikit learn

Imbalanced learn

PYCARET

대시보드

Streamlit

프로젝트 진행 일정

기간	수행 목표	일수
2024.01.20~01.24	**금융 데이터 이해 및 컬럼 리뷰**	5일
2024.01.25~02.10	**EDA**	17일
2024.02.10~02.23	**데이터 전처리 (특성 엔지니어링)**	14일
2024.02.23~02.29	**모델링 및 파라미터 튜닝**	7일
2024.03.04~03.12	**발표 자료 준비**	8일
2024.03.04~03.15	**대시보드 설계**	11일
2024.03.12~03.19	**피드백 및 최종 수정**	7일
2024.03.20	최종 발표회	**63일**

분석로드맵

01. 프로젝트 배경

- 문제 제기
- 문제 배경

02. 프로젝트 개요

- 프로젝트 목표
- 데이터 소개 및 필터링

03. EDA 및 전처리

- 특성 선택
- 이상치 결측치 처리
- 전처리

04. 모델링

- 성능 평가지표
- 모델 선정 및 평가
- 모델 벤치마크
- 최종 모델 선정

06. 아이디어 제안

- 비즈니스 아이디어 도출
- 추후 추가 분석 제안

05. 대시보드

- Streamlit 대시보드 구성

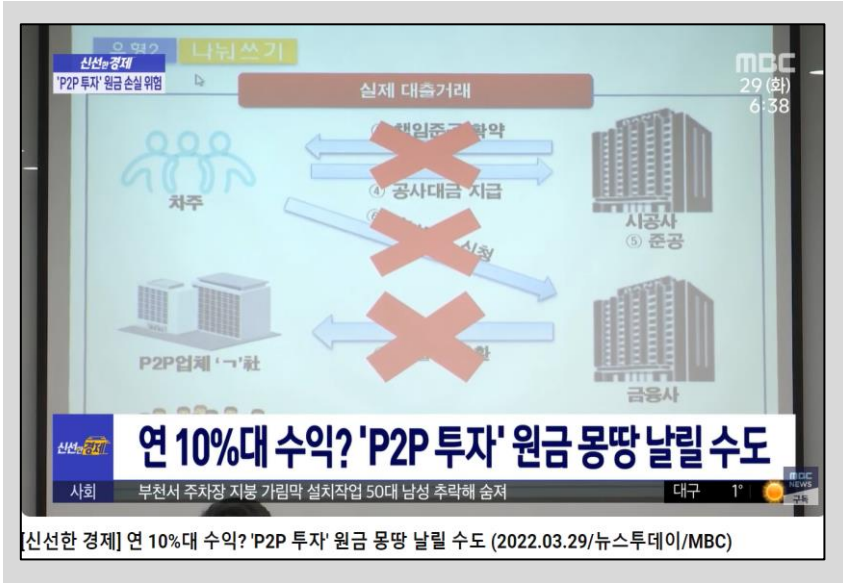
STEP 1 문제정의 → STEP 2 데이터 분석 및 예측 모형 설계 → STEP 3 비즈니스 인사이트 도출

1장. 프로젝트 배경

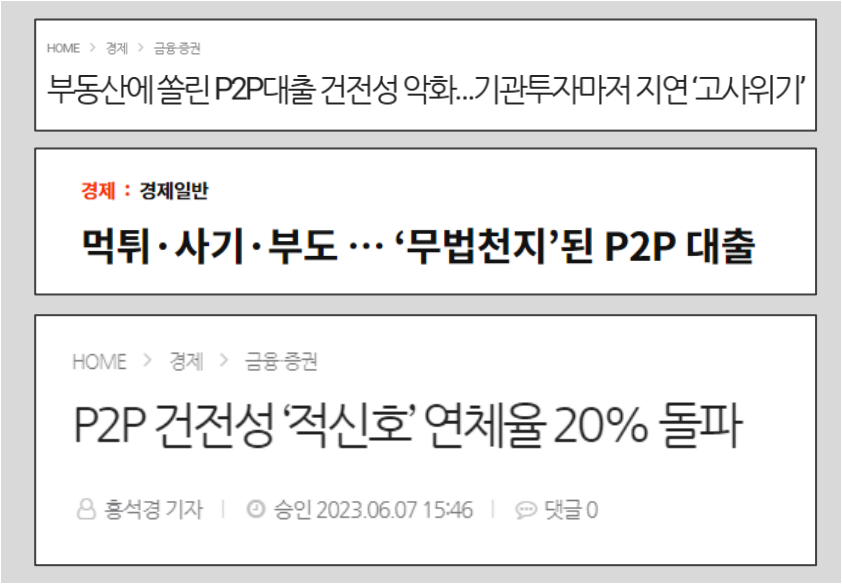
문제 제기

#건전성 이슈

P2P 대출에서는 건전성 문제가 일반 금융기관보다 심각한 것으로 보여집니다.



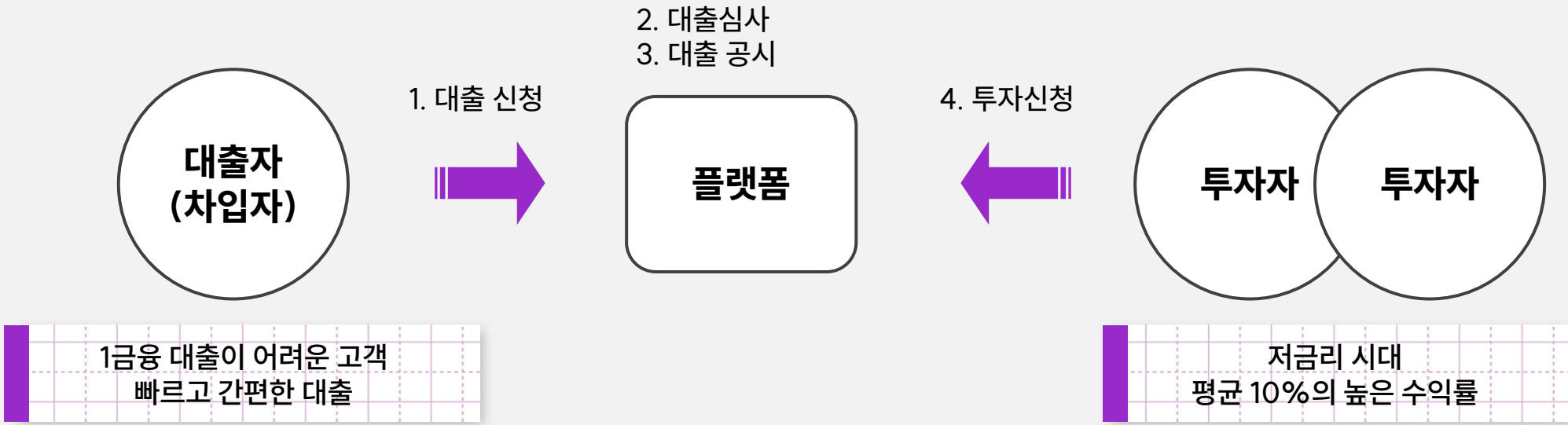
[신선한 경제] 연 10%대 수익? 'P2P 투자' 원금 몽땅 날릴 수도 (2022.03.29/뉴스투데이/MBC)



P2P 대출의 높은 연체율과 대출 부도로 인한 위험이 비즈니스 부작용으로 나타나고 있습니다.

P2P 대출

P2P(Peer-to-peer) 대출은 금융기관의 개입 없이 온라인에서 개인간 대출을 중개해주는 핀테크 기반의 금융 방식입니다.



P2P 대출의 장점과 성장 동향

P2P 대출의 장점을 통해 P2P 대출 시장 수요는 점차 증가하고 있습니다.

편리한 접근성

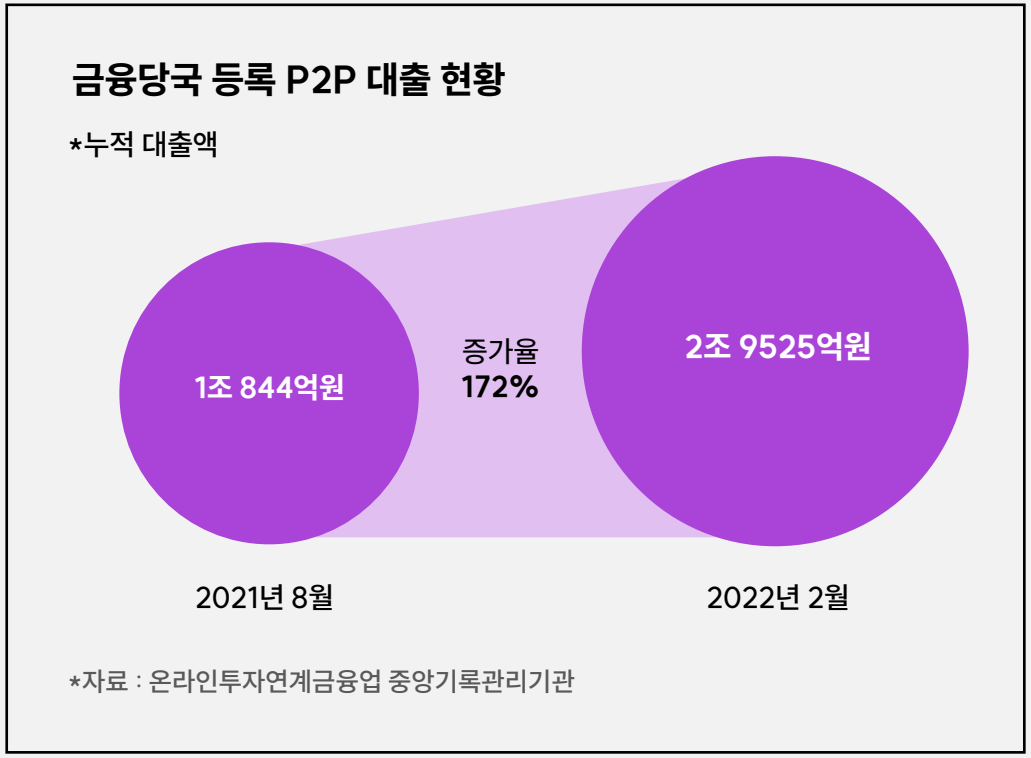
신속하고 간편하게
온라인 플랫폼을 이용한 접근이 가능합니다.

신용 기준 완화

전통적인 은행에 비해
신용기준이 완화되었습니다.

투자 기회 제공

자금 운용을 할 수 있는
다양한 옵션을 제공합니다.



P2P 대출의 문제점

P2P 대출은 신용위험, 부채 문제, 투자 위험 등 다양한 위험을 야기하고 있습니다.

신용 위험

개인이 대출자의 신용을
평가하기에는 한계가 있습니다.

부채 문제

비교적 쉬운 대출로 인한
대출자의 부채 누적 문제가 있습니다.

투자 위험

대출 채무 불이행으로 인한
투자자의 투자 손실 위험이 있습니다.

2장. 프로젝트 개요

프로젝트 목표

1

플랫폼의 **효과적인 리스크 관리**를 위한
P2P 대출 상환 예측 모델을 개발

2

모델을 활용하고 **비즈니스 인사이트 도출**을 위한
Streamlit을 이용한 분석 대시보드 설계

프로젝트 주요 고려사항

금융 비즈니스 상황에 맞게 투명성과, 빠른 속도, 타겟 불균형 문제 해결을 중심으로 프로젝트를 진행했습니다.

설명 가능성

"모델을 통한 심사 결과를
어떻게 고객에게 설명할 수 있을까?"

신속성

"모델을 통해 대출심사 결정을
빠르게 처리할 수 있을까?"

타겟 불균형 조정

"미상환은 상환보다 드물게 발생한다.
데이터의 타겟 불균형을 어떻게 해결할까?"

데이터 셋 소개

2007년 ~2018년까지의 Kaggle 에 공개된 Lending Club P2P 대출 플랫폼의 승인된 대출 데이터입니다.

- 물리적 파일명 : accepted_2007_to_2018Q4.csv
- 총 1,319,510 건의 대출 데이터

미상환 **Charged Off**	상환 **Fully Paid**
262,215건	1,057,295 건

kaggle LendingClub

 # 개인 대출에 관한 135개의 컬럼 (타겟 변수 포함)

의미를 중심으로 다음과 같이 특성을 그룹화하여 프로젝트 진행하였습니다.

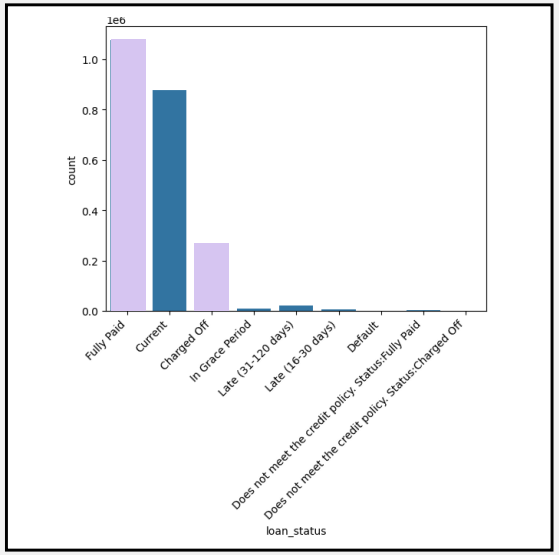
개인정보 | 신용기록 | 신용조회 | 신용거래 사용 | 대출 정보 | 할부 계정 | 리볼빙 계정

신용카드 | 연체, 미납 | hardship | settlement

타겟 설정 및 데이터 필터링

프로젝트 목표에 따라, 타겟의 이진화, 개인 대출 예측을 중점적으로 다루면서 시계열 특성을 배제하여 문제를 단순화하고 분석했습니다.

각 타겟의 개수 (count plot)

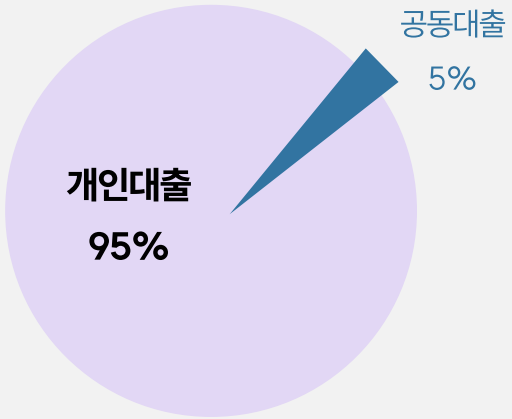


타겟 변수인 'loan_status'에서

Charged off, Fully paid를

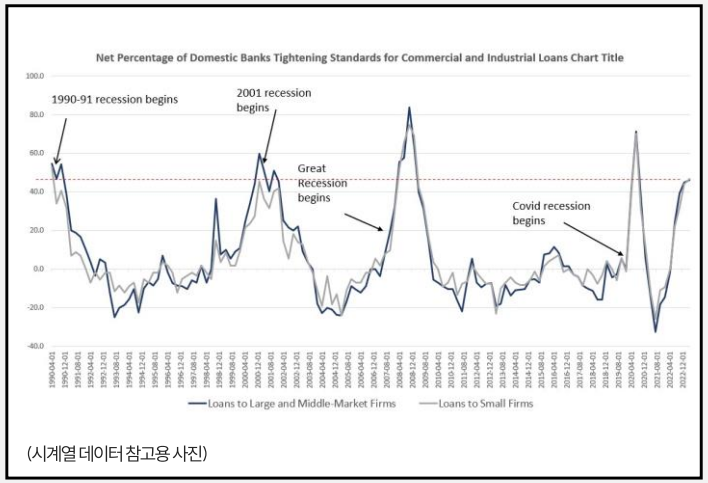
예측 타겟으로 설정하였다.

'application_type' 범주 값 비율



개인대출에 대해서만 진행하였다.

경제 침체에 따른 대출 긴축현상



(시계열 데이터 참고용 사진)

대출 데이터는 **거시 경제의 영향으로**

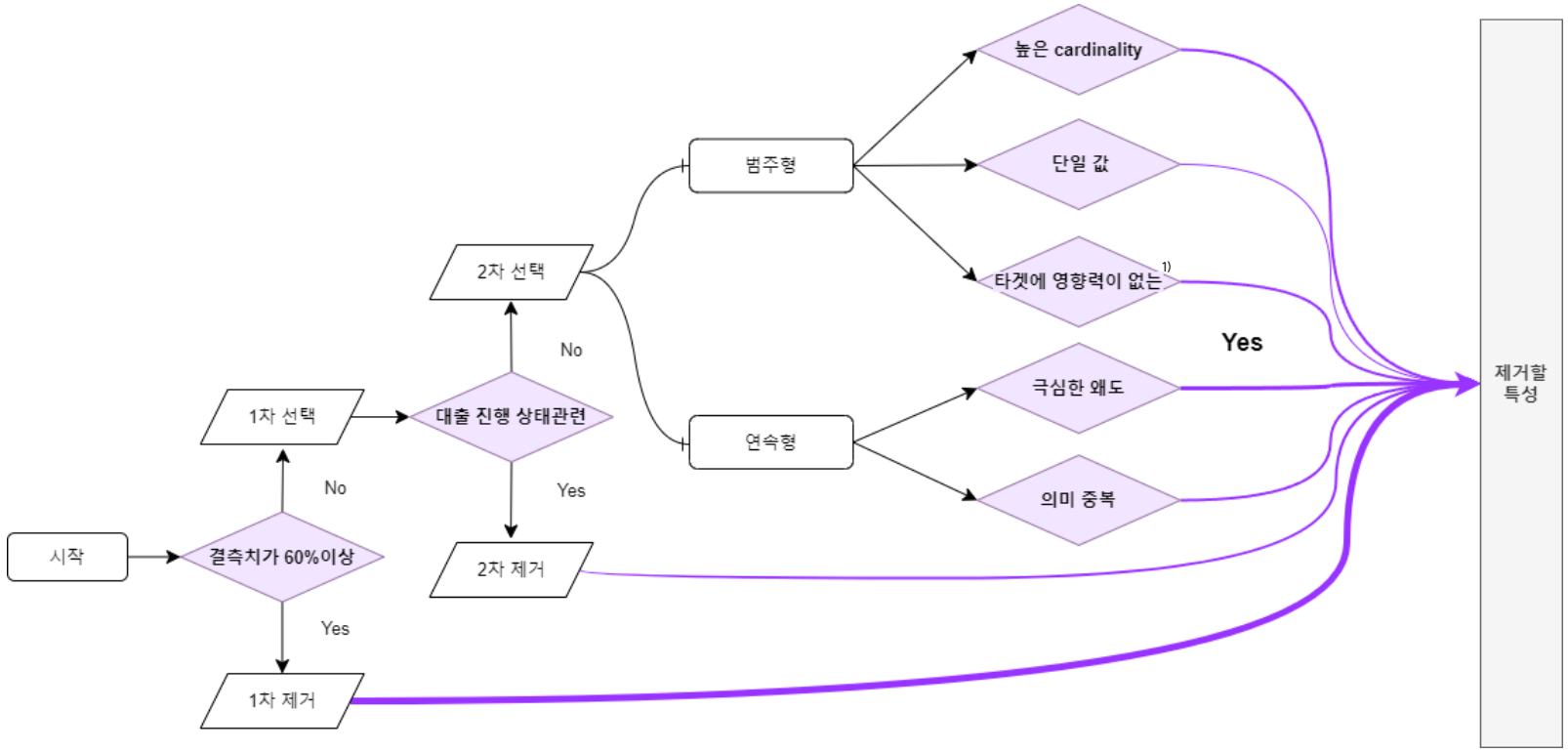
시기적 특성을 띄기 때문에

시계열 데이터는 제외하였다.

3장. EDA 및 전처리

특성 선택 과정 요약

총 134개의 특성 중, 특성 선택 과정을 통해 최종적으로 16개의 특성을 선택했습니다.



1) 타겟 영향도가 없다는 것은 범주형 타겟의 비율의 차이가 없음을 의미합니다.

특성 선택 과정– 아이디어 검증 1 (grade, sub_grade)

Q1. 각 등급의 상위는 미상환율이 더 낮을까? ex. 대출등급 C1는 B4인 사람보다 상환을 더 잘 할 것이다.

그림 1. 대출세부등급(sub_grade)별 미상환율

낮은 등급일수록 미상환율이 우상향하는 것을 볼 수 있습니다.

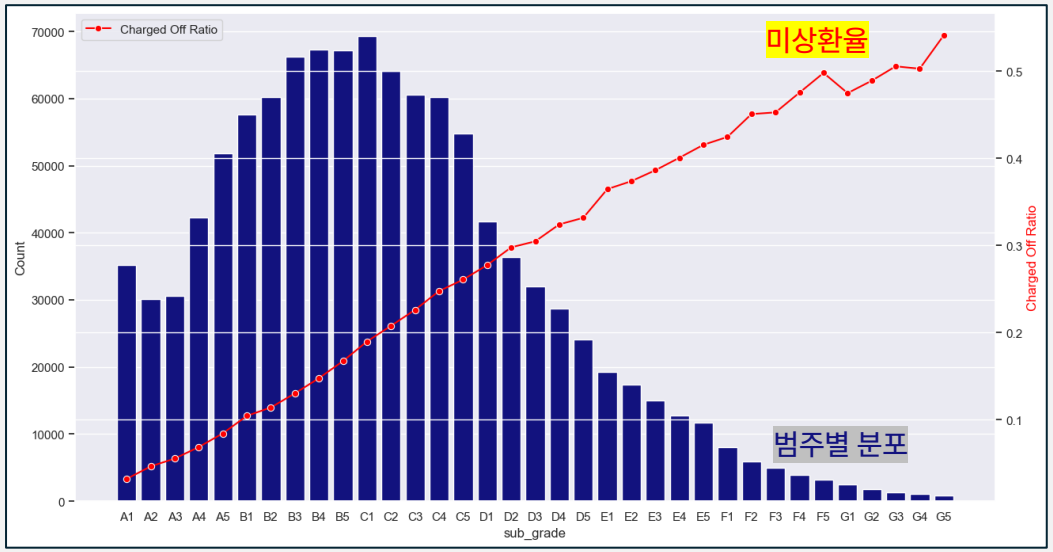
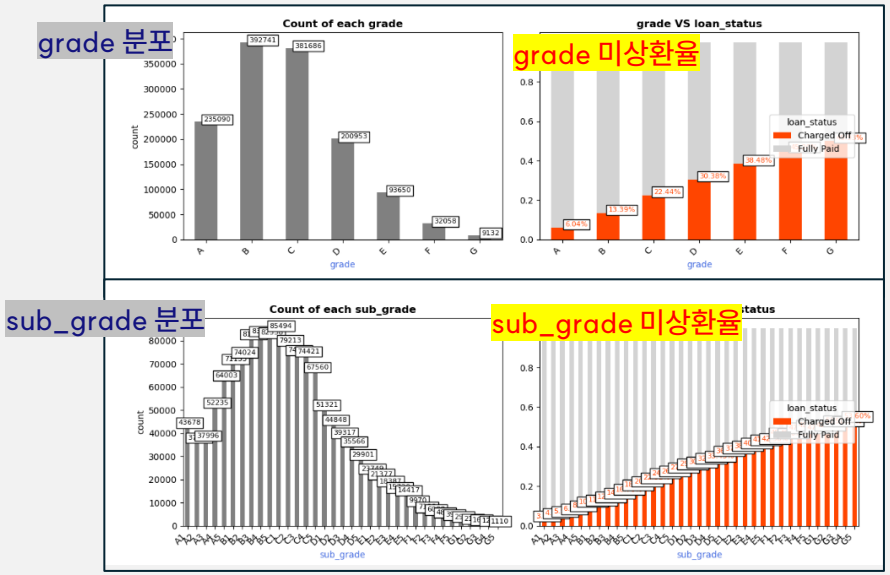


그림 2. 대출등급(grade) / 대출세부등급(sub_grade) 비교

두 특성 값의 분포와 미상환율의 차이가 보여지지 않습니다.



더 자세한 정보를 가진 **대출세부등급(sub_grade) 을 선택하기로 결정했습니다.**

특성 선택 과정– 아이디어 검증 2 (emp_length)

Q2. 재직기간(emp_length) 이 짧은 사람은 위험도가 더 높을까? ex. 재직기간이 긴 사람일수록 대출 상환 안정성이 좋을 것이다.

그림 1. 재직기간(emp_length) 별 평균 annual_inc(연간 소득)
평균 연간소득은 재직기간이 늘수록 높아지는 경향을 보입니다.

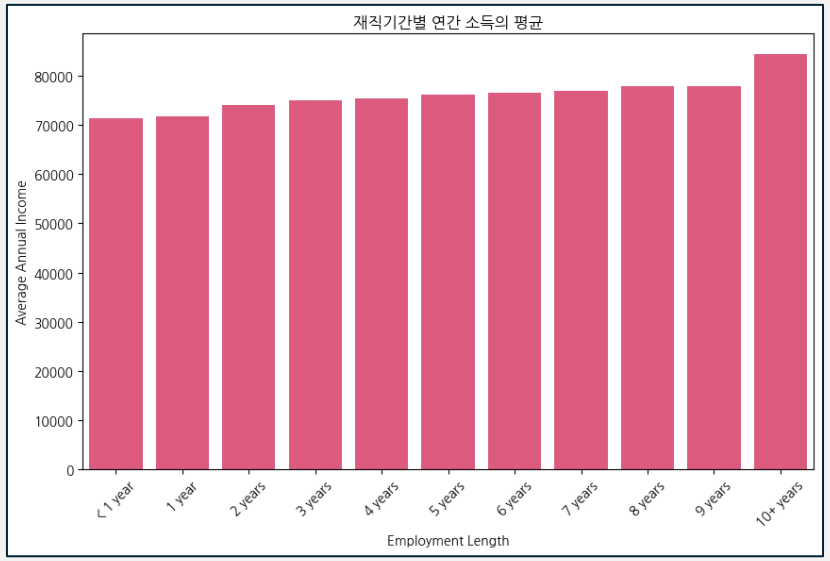
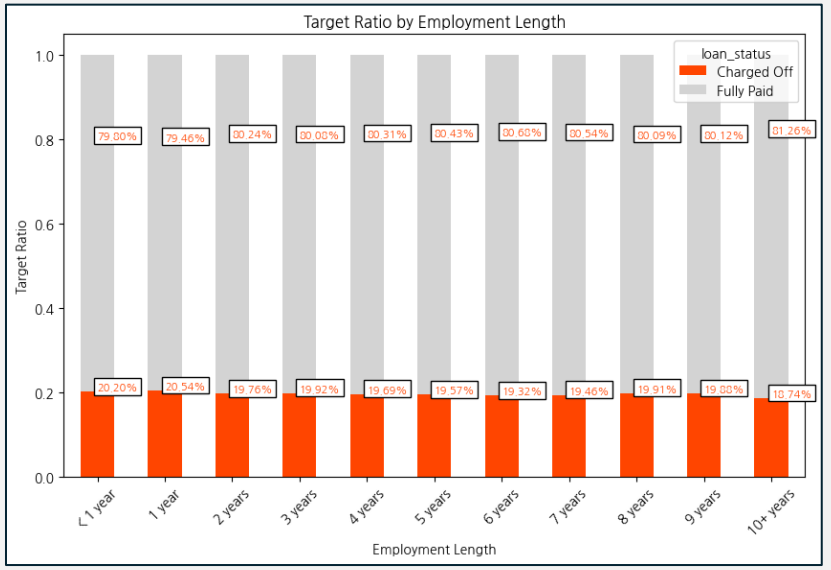


그림 2. 재직기간(emp_length)별 미상환율
재직 기간에 따른 대출 미상환율의 차이가 거의 유사한 비율입니다.



emp_length 범주별 유의미한 차이를 발견하지 못해 **제외**하기로 결정했습니다.

최종 모델링 특성 선택

최종적으로 선택된 16개의 특성은 다음과 같습니다.

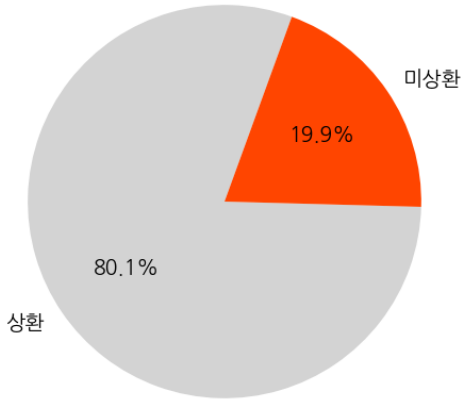
개인정보 (5)	addr_state ,last_fico_range_high, fico_range_high, dti, verification_status
대출 정보 (4)	loan_amnt, term, int_rate, sub_grade
리볼빙 계정 (3)	revol_util, mo_sin_old_rev_tl_op, mo_sin_rcnt_rev_tl_op
신용거래 사용 (4)	open_acc, total_acc, avg_cur_bal, pct_tl_nvr_dlq

결측치 처리

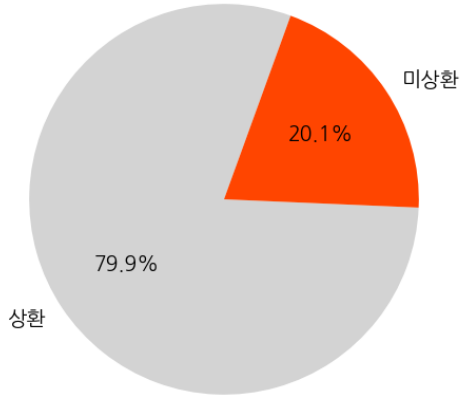
결측행을 제거한 데이터 셋을 구성했습니다.

선택된 16개의 특성 중 6개의 특성이 결측을 가지고 있고, 평균 결측치 비율이 전체 데이터의 5%로 제거를 결정했습니다.

제거 전 : 1,319,510 행



제거 후 : 1,251,098행



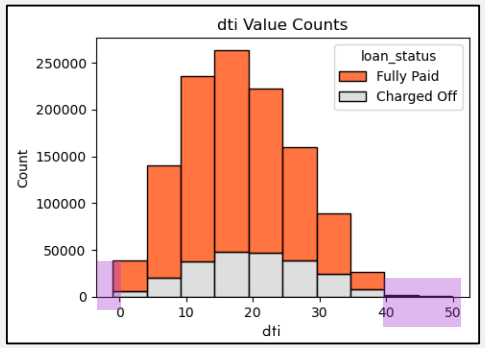
제거 후 1,251,098 건의 데이터가 확보되기 때문에 프로젝트에 진행에 충분한 크기라고 판단됩니다.

이상치 처리

각 특성 별 정상범위를 벗어나는 데이터들을 처리했습니다.

*이상치 판단 영역

1. 총부채상환율(dti) : 전체 0.15%



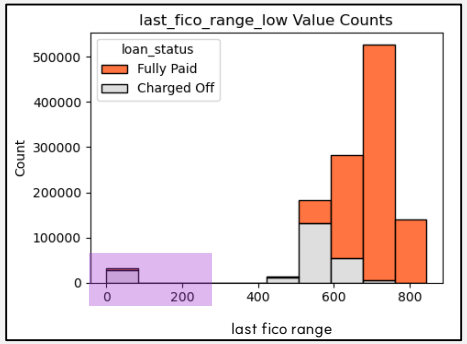
랜딩 클럽에서 대출승인자의 dti는

일반적으로 0~ 최대 40%까지 적용됩니다.

▽

0%이하 값과 40% 초과 값은 제거했습니다.

2. 최근신용점수하한(last_fico_range_low): 전체 2.72%



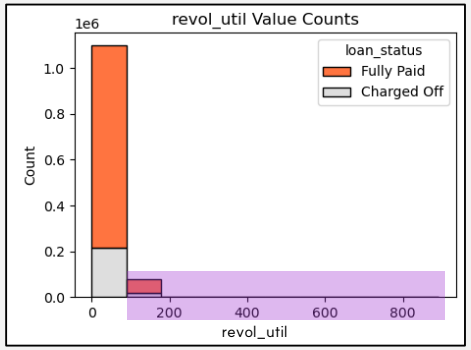
신용 점수의 경우, 랜딩 클럽에서

300~850의 값을 가진다고 명시되어 있습니다.

▽

300 이하의 값 제거했습니다.

3. 리볼빙사용률(revol_util) 전체 0.37%



비율이 100을 넘는 값을 제거했습니다.

이상치 처리 결과 **총 40,218행(전체의 3.21%)**이 제거되었습니다.

특성 인코딩

순서형 특성, 명목형 특성 등 특징을 나누어 다음과 같은 전처리를 진행했습니다.

1. 순서형 특성	2. 고차원 범주형 특성	3. 명목형 특성	4. 대출기간(term) 정수화
세부 신용등급(sub_grade)	거주 지역 : 미국 주 약자 (addr_state)	소득 검증 상태 (verification_status)	대출기간 (term) 의 값 : '36months', '60 months'
↓	↓	↓	↓
OrdinalEncoder	**CountEncoder**	**OneHotEncoder**	**정수화**
"F5": 0, "F4": 1, "F3": 2, "F2": 3, "F1": 4, ... "A4": 26, "A3": 27, "A2": 28, "A1": 29	특성의 빈도수로 변환		36,60으로 변환 1. 직관적으로 파악이 가능 2. 연산 속도 증가, 데이터 메모리 감소

특성 생성

일부 연속형 특성들을 조합하여 새로운 특성을 생성했습니다.

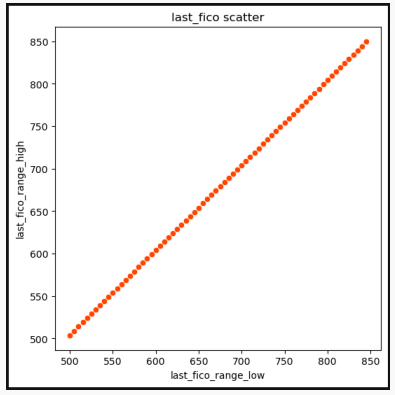
1. 월 상환금(installment)

$$installment = loan\_amnt \times \frac{int\_rate \times 0.01}{12} * \frac{(1+\frac{int\_rate \times 0.01}{12})^{term}}{(1+\frac{int\_rate \times 0.01}{12})^{term}-1}$$

기존 월상환금(installment) 특성이
대출금(loan_amnt), 이자율(int_rate), 대출 기간(term)과
대조하여 계산이 맞지 않는 값들을 발견해,
스냅샷에 의한 이상치로 판단하여 기존 월상환금(installment)
을 삭제하고 직접 월상환금(installment) 특성을 생성했습니다.

2. 최근신용점수(last_fico_score)

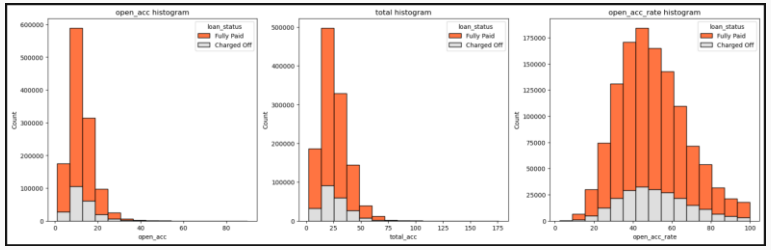
$$last\_fico\_score = \frac{last\_fico\_range\_low + last\_fico\_range\_high}{2}$$



최근신용점수 상/하한(last_fico_range_low / high)
특성의 상관계수가 1이고 중복된 의미를 가지므로,
두 특성의 평균값으로
최근신용점수(last_fico_score)를 생성했습니다.

3. 활성계좌비율(open_acc_rate)

$$open\_acc\_rate = \frac{open\_acc}{total\_acc} \times 100$$



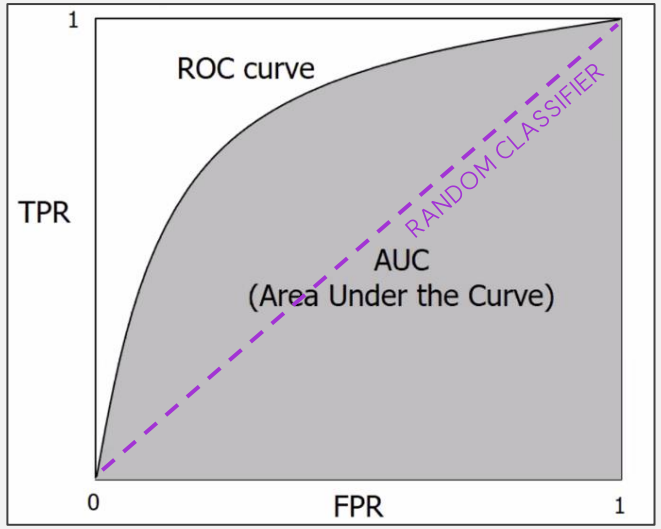
큰 왜도와 높은 상관계수를 가지고,
중복된 의미를 가진 활성계좌수(open_acc)와
총계좌수(total_acc)를 조합하여 활성계좌비율
(open_acc_rate)을 생성하였습니다.

4장. 모델링 : 분류기 성능 최적화

성능 평가 지표: ROC AUC & Recall

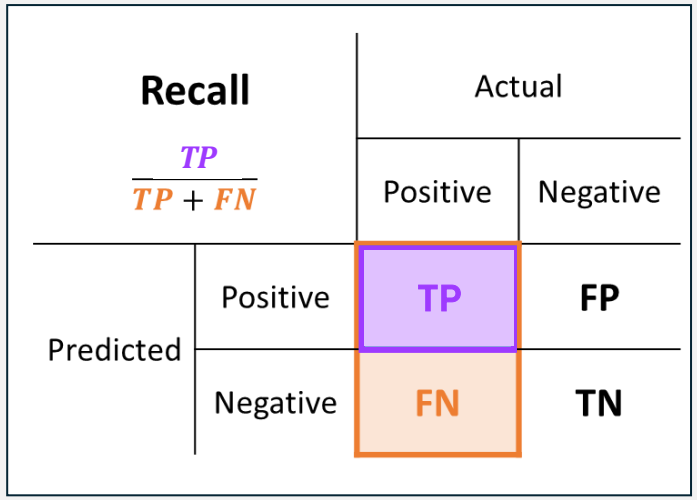
플랫폼의 신뢰성 상승과 대출 거래 활성화를 위한 비즈니스 목표를 고려하여 ROC-AUC와 RECALL을 평가지표로 선택했습니다.

1. ROC AUC



민감도와 특이도의 trade-off를 고려

2. Recall



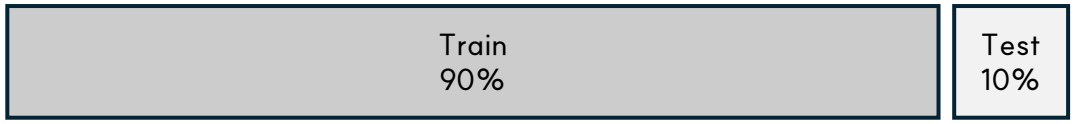
양성(Charged Off) 예측에 대한 신뢰도를 고려

* TPR(True Positive Rate) : 미상환할 사람을 미상환할 것으로 정확하게 예측한 비율
* FPR(False Positive Rate) : 미상환할 사람을 상환할 것으로 잘못 예측한 비율

타겟 불균형 조정

1. 훈련, 테스트 세트 분리

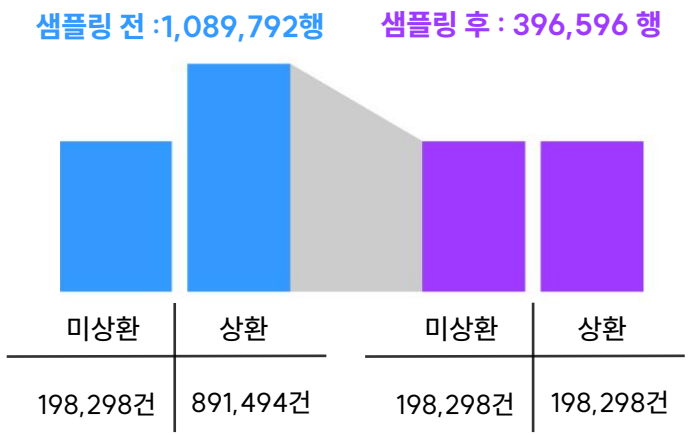
Scikit-learn의 train_test_split(stratify=target)사용하여 90:10으로 훈련, 테스트 세트를 분리했습니다.



2. 훈련세트에 대해 imbalanced-learn의 RandomUnderSampling 적용

타겟의 비율이 약 8:2로, 타겟의 불균형은 모델의 편향이 발생하고 성능측정이 왜곡될 수 있습니다.

상환(Fully Paid)데이터에 대해서 리샘플링 이후에도, 총 데이터의 크기가 프로젝트 진행에 충분하다 판단하여, 데이터를 언더샘플링 하기로 결정했습니다.



*** RandomUnderSampling**

단순하게 무작위로 다수 타겟의 샘플을 선택하여 제거하는 샘플링 기법입니다.

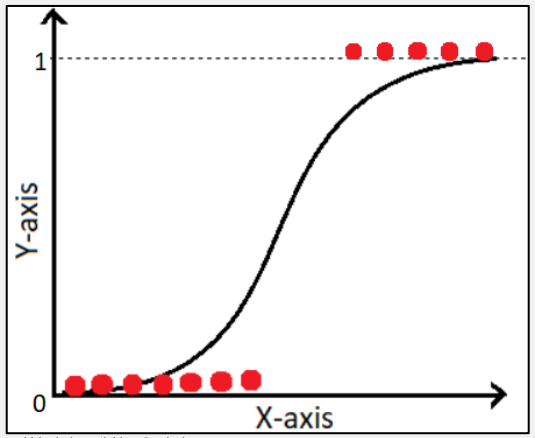
즉, 소수 타겟의 샘플 수와 동일한 수의 다수 타겟의 샘플을 무작위로 선택하여 제거합니다.

실행시간이 짧으며 데이터의 용량을 줄일 수 있습니다.

모델 선정

설명 가능한 모델들을 찾아 각 모델의 특징과 장·단점을 살펴 보았습니다.

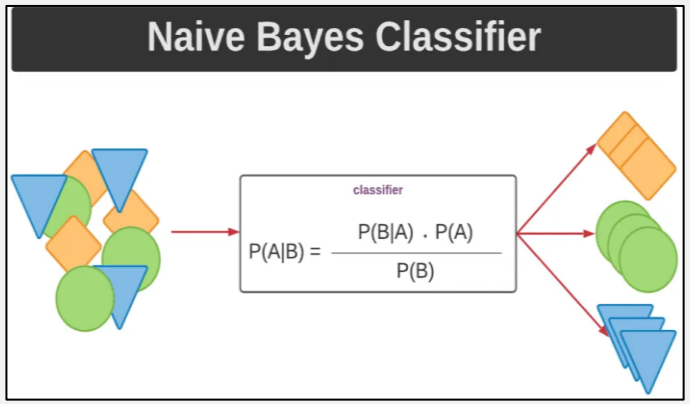
선형 기반 모델 : Logistic Regression



(선형 기반 모델 참고용 사진)

로지스틱 함수를 사용해 확률을 예측하며,
이진분류에 적용되는 모델입니다.
해석이 용이하고, 계신 효율성이 뛰어납니다.
비선형 문제에 적합하지 않고, 과적합의 가능성이 있습니다.

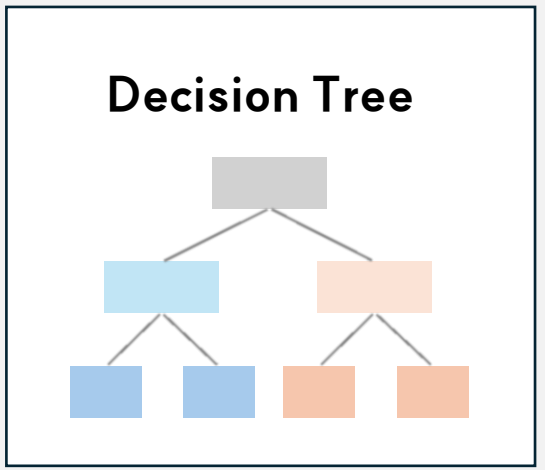
선형 기반 모델 : Naïve Bayes



(선형 기반 모델 참고용 사진)

확률 기반의 분류 알고리즘입니다.
특성들의 독립성을 가정합니다.
효율적인 계산과 소량의 데이터에서도 성능이 우수합니다.
자연어 처리에서 효과적으로 사용됩니다.

트리 기반 모델 : Decision Tree



(트리 기반 모델 참고용 사진)

데이터 특성에 기반한 질문 반복으로 결정을 내리는 모델입니다.
데이터는 트리의 각 노드에서 특정 기준으로 분할합니다.
해석이 용이하고, 비선형 관계를 모델링할 수 있습니다.
과적합에 취약할 수 있습니다.

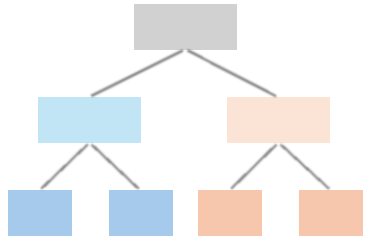
모델 선정

모델의 특징과 비즈니스 관점을 고려하여 Decision Tree를 사용하기로 결정했습니다.

1. 해석이 가능한 모델

2. 비선형 관계 탐지

Decision Tree



3. 특성간 상호작용 파악

4. 간단한 데이터의 전처리 과정

교차 검증

모델의 교차 검증을 통해 과대적합의 가능성이 있음을 확인했습니다.

교차 검증을 통한 평가지표 확인

베이스라인 모델 Cross Validate

		Fold_1	Fold_2	Fold_3	Fold_4	Fold_5	Mean
ROC_AUC	훈련	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	검증	0.8283	0.8283	0.8281	0.8275	0.8263	0.8277
RECALL	훈련	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	검증	0.8254	0.8222	0.8259	0.8243	0.8252	0.8246
시간 (Second)	훈련	11.7056	11.7326	11.7516	11.6876	11.9607	11.7676
	검증	0.1200	0.1250	0.1300	0.1260	0.1410	0.1284

교차 검증 결과(평균)

ROC-AUC	Recall
훈련 : 1.0 검증 : 0.8277	훈련 : 1.0 검증 : 0.8246

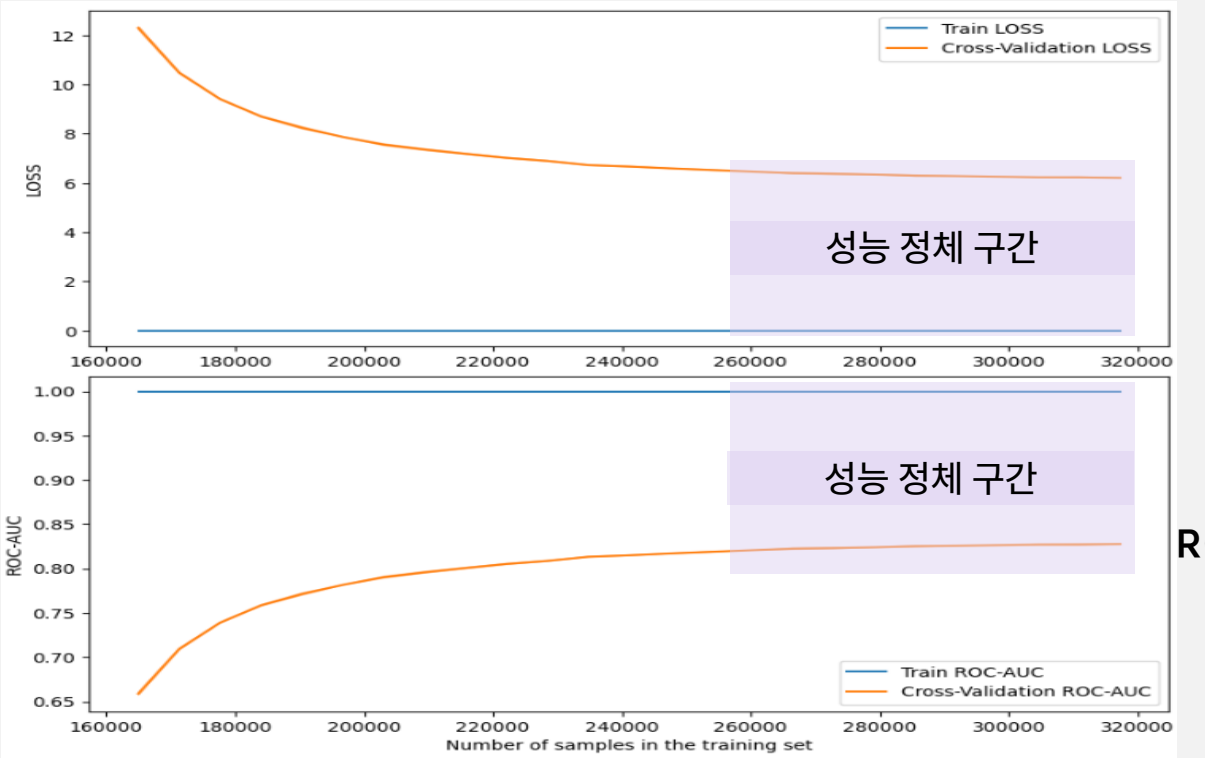
ROC-AUC를 비교했을 때, 훈련 세트의 성능에 비해
검증 세트의 성능이 비교적 낮은 것을 확인할 수 있습니다.

모델이 훈련데이터에 과대적합일 가능성이 있습니다.

과대적합 검증

학습곡선(learning curve) 확인 결과, 모델이 훈련세트에 과대적합(Over-fitting)되었다 판단했습니다.

학습곡선(learning curve)을 통한 과대적합 확인



훈련세트의 데이터가 증가할 수록
LOSS가 감소하는 모습을 볼 수 있습니다.

훈련세트의 데이터가 증가할 수록
ROC-AUC가 증가하는 모습을 볼 수 있습니다.

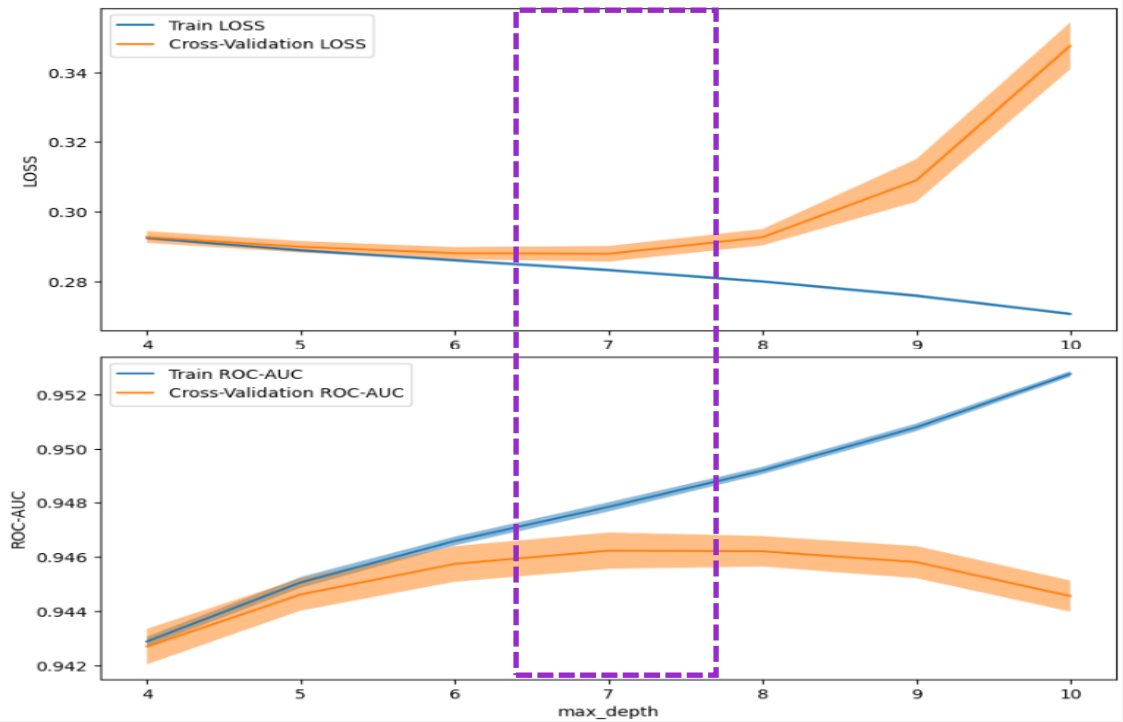
일정 데이터 수 이상부터는
훈련&검증 세트 간의
LOSS와 ROC-AUC의 간격이
더 이상 좁혀지지 않고
일정한 오차를 보이는 것을
확인할 수 있습니다.

➡ **과대적합 (Over-fitting)**

하이퍼파라미터 튜닝

검증곡선(validation curve) 확인결과, max_depth의 값이 7일 때 모델이 가장 좋은 성능을 보입니다.

검증곡선(Validation curve)을 통한 최적의 하이퍼파라미터 탐색



max_depth가 증가할 수록
LOSS의 오차가 증가합니다.

max_depth가 증가할 수록
ROC-AUC의 오차가 증가합니다.

max_depth <= **6**

분산에 따라 과소적합하는 경향이 보입니다.

따라서, max_depth = **7**

모델이 **최대 성능**을 가집니다.

하이퍼파라미터 튜닝

최적의 하이퍼파라미터를 튜닝 후 성능을 확인 해보니 과대적합이 해결되어 적합한 성능을 보입니다.

▶ PyCaret의 Tune Model을 이용한 최적의 하이퍼파라미터

하이퍼파라미터	값
criterion	"entropy"
max_depth	7
min_samples_split	26
min_samples_leaf	18
max_features	None

하이퍼파라미터 튜닝 후 성능 확인

튜닝 모델 Cross Validate

		Fold_1	Fold_2	Fold_3	Fold_4	Fold_5	Mean
ROC_AUC	훈련	0.9478	0.9481	0.9476	0.9480	0.9478	0.9478
	검증	0.9460	0.9453	0.9473	0.9460	0.9465	0.9462
RECALL	훈련	0.9151	0.9116	0.9067	0.9109	0.9152	0.9119
	검증	0.9132	0.9093	0.9078	0.9095	0.9150	0.9110
시간 (Second)	훈련	6.5024	6.4299	6.4770	6.5385	6.4204	6.4736
	검증	0.1530	0.1360	0.1490	0.1450	0.1430	0.1452

교차 검증 결과(평균)

ROC-AUC	Recall
훈련 : 0.9478 검증 : 0.9462	훈련 : 0.9119 검증 : 0.9110

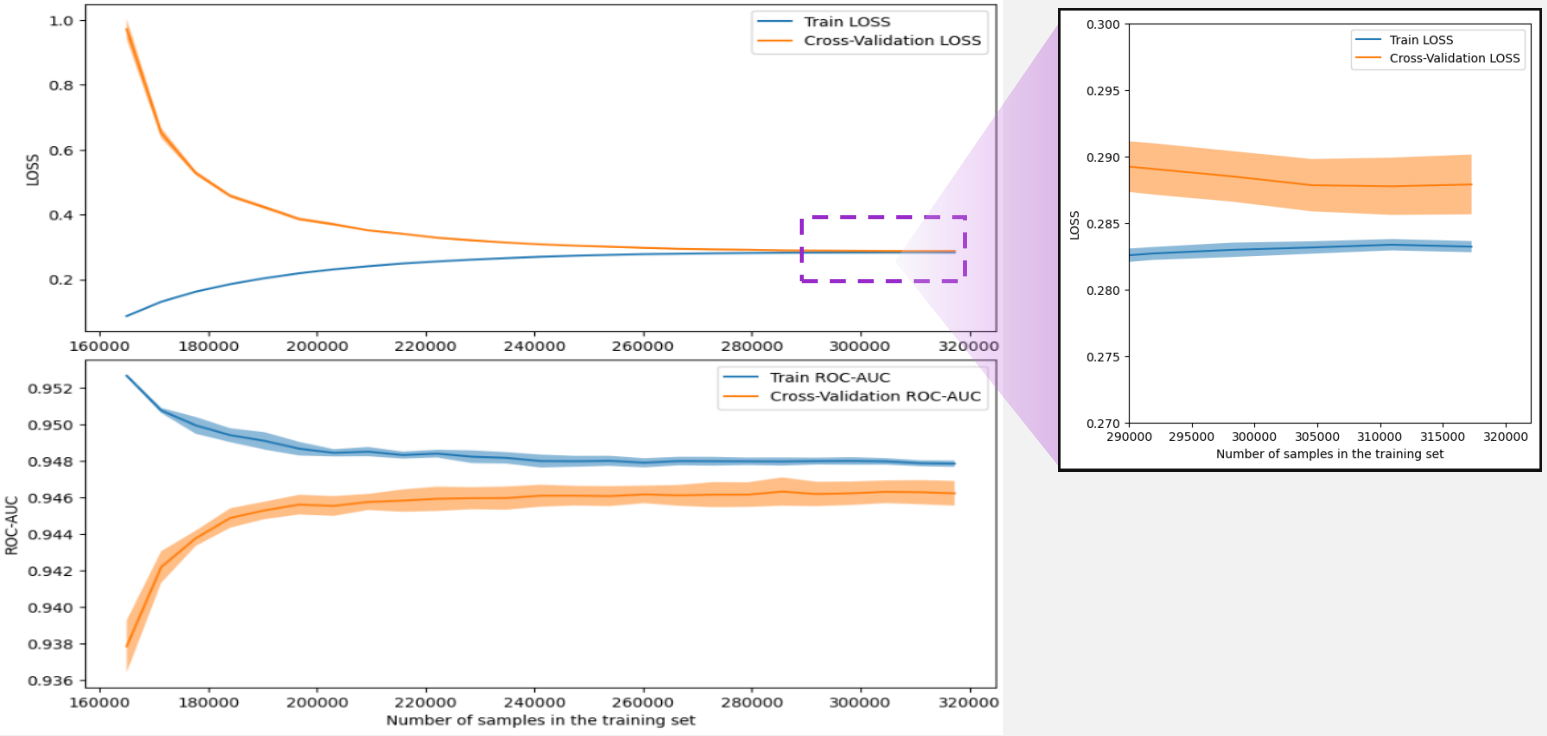
ROC-AUC 성능 점수를 비교했을 때,
훈련 세트와 검증 세트의 성능 점수 격차가 감소해
과대적합이 해소된 것을 확인 할 수 있습니다.

모델이 데이터에 적합한 성능을 보입니다.

과대적합 해소

훈련 데이터 샘플 증가에 따라 모델 점수 차이와 분산이 감소해 일정 수치에 수렴하며, 이는 언더샘플링된 데이터로도 수렴합니다.

학습곡선(learning curve)을 통한 성능 확인



훈련 세트 데이터의 증가에 따라 LOSS와 ROC-AUC의 훈련 및 검증 점수 격차가 감소하고, 분산도 작아지며 값이 일정 수치로 수렴하는 것을 확인할 수 있습니다.

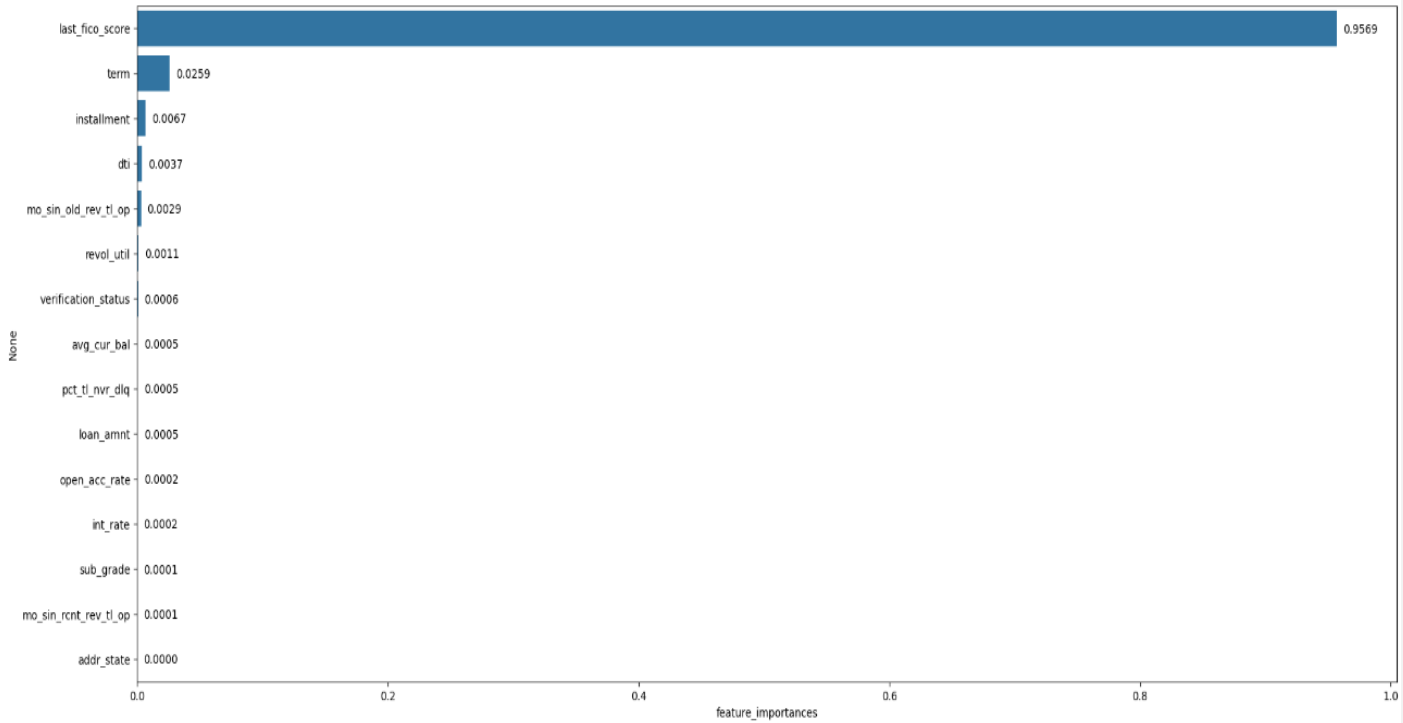
또한, 언더 샘플링된 데이터 샘플 수로 충분히 일정 수치에 수렴합니다.

이는 해당 모델이 데이터에 잘 적합하고 일반화 되었다는 것을 알 수 있습니다.

모델 해석

특성중요도를 통해 예측에 가장 중요한 특성은 최근신용점수(last_fico_score)임을 확인할 수 있습니다.

특성중요도(Feature Importance)



***최근신용점수(last_fico_score)**

최근신용점수 상,하한 (last_fico_range_high / low)의 평균값

대출자의 최근신용점수

*** 신용점수** = 개인의 신용 이력을 종합적으로 평가해 산출된 점수.

예측에 가장 중요한 특성 :

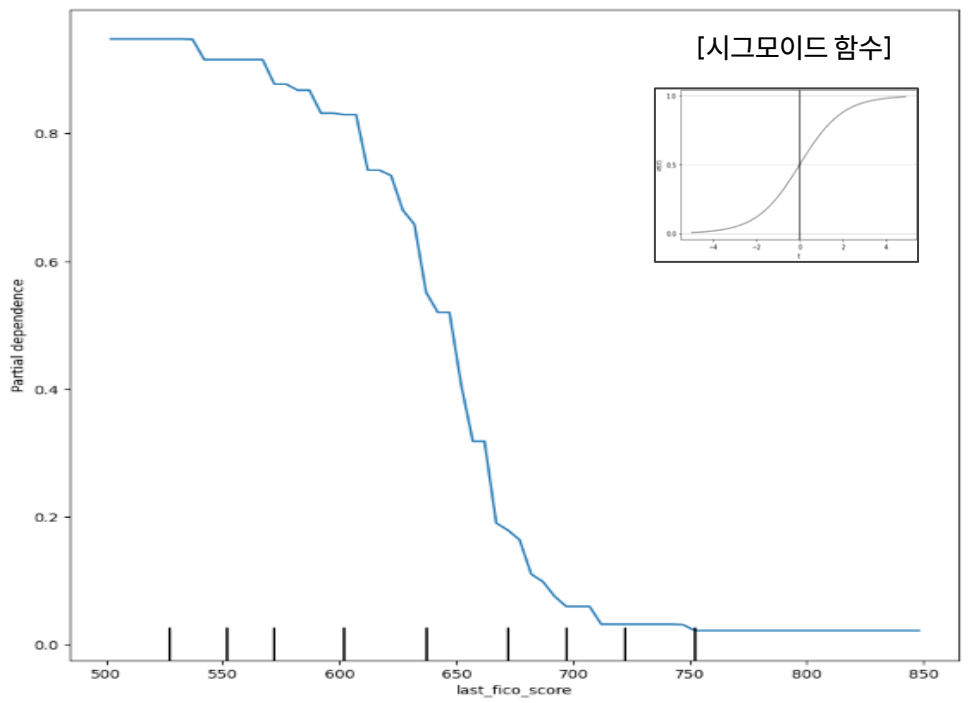
최근신용점수(last_fico_score)

데이터 내에서 확인 가능한 여러가지 신용을 나타낸 특성들을
포괄적 계산하여 하나의 수치로 나타낸 값으로,
모델 예측에 가장 중요한 특성으로 사용되었습니다.

모델 해석

가장 중요한 특성으로 판단한 최근신용점수(last_fico_score)를 PDP으로 검증, 시그모이드 함수의 대칭 함수 모양을 나타냅니다.

부분 의존성 플롯(Partial Dependence Plot)



최근신용점수(last_fico_score)에 대한 부분 의존성 플롯이
시그모이드 함수와 대칭 함수의 모양으로 그려지는 모습을 보입니다.

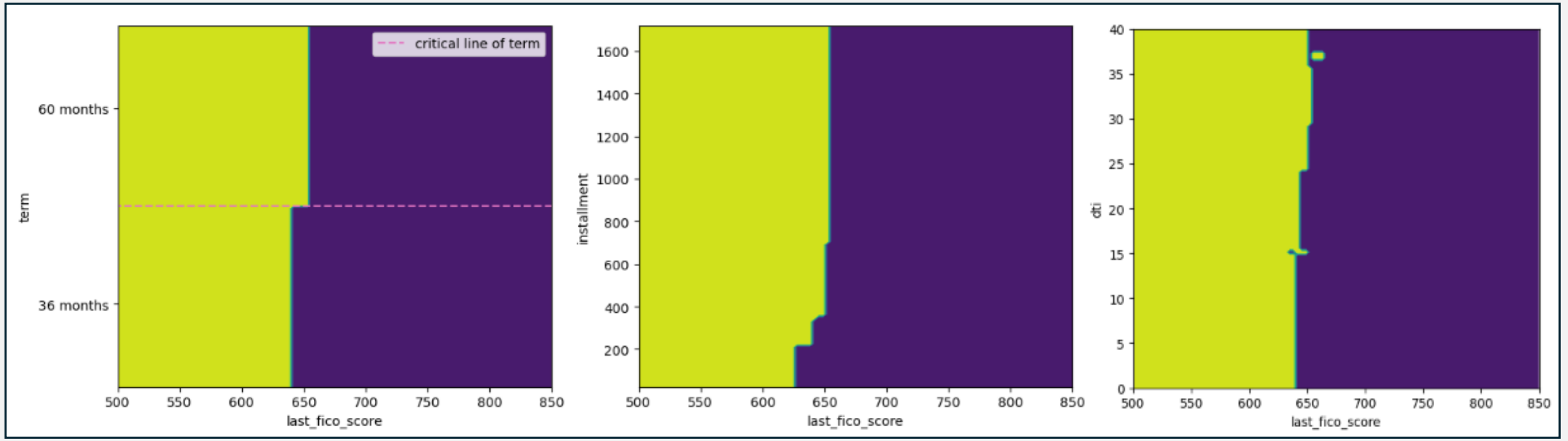
예측 확률에 가장 큰 영향력을 주는 특성임을 알 수 있습니다.

최근신용점수(last_fico_score)가 낮을 수록
미상환(Charged off) 가능성이 높아집니다.

모델 해석

최근신용점수(last_fico_score)와 다른 특성 간의 결정경계를 시각화하여 특성이 타겟에 미치는 경향을 확인할 수 있습니다.

결정경계 시각화



신청자의 대출기간(term)이 길수록
미상환(Charged off) 인 경향을 보입니다.

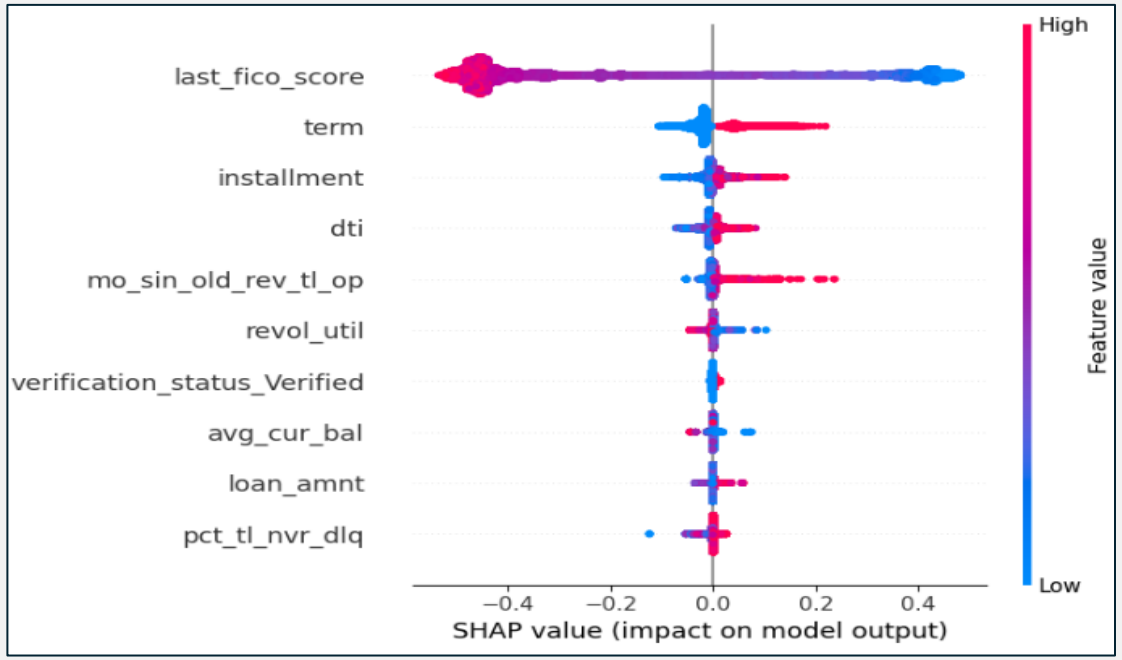
신청자의 월 원리금(installment)이 높을수록
미상환(Charged off) 인 경향을 보입니다.

신청자의 총부채상환비율(dti)이 높을수록
미상환(Charged off) 인 경향을 보입니다.

모델 해석

SHAP 그래프를 활용해 특성 별 영향도와 경향성을 확인 해 본 결과, last_fico_score이 가장 큰 영향력을 가진 것으로 판단되었습니다.

summary_plot을 이용한 특성 별 shap value, 영향도



- **최근신용점수(last_fico_score)** 예측값과 음의 상관성

낮아질수록 미상환(Charged off)인 경향성이 보입니다.
가장 영향력이 큰 특성으로 확인됩니다.

- **대출기간(term)** 예측값과 양의 상관성

길수록 미상환(Charged off),
짧을수록 상환(Fully paid)인 경향성을 보입니다.

- **월 상환금(installment), 총 부채 상환율(dti)** 예측값과 양의 상관성

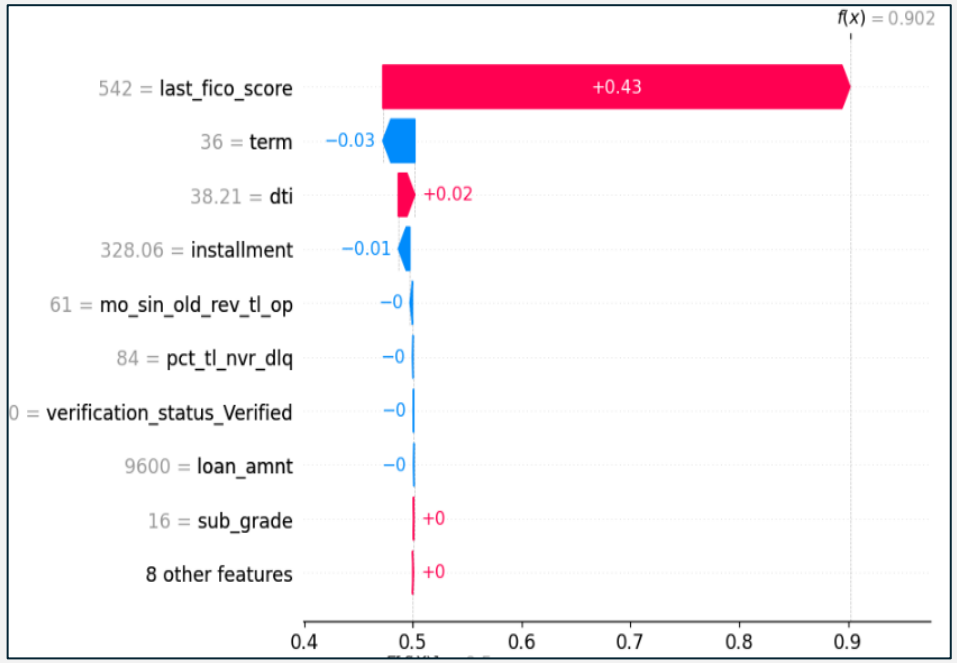
작을수록 상환 (Fully paid),
높을수록 미상환(Charged off)인 경향성을 보입니다.

모델 예측 해석

최근신용점수(last_fico_score)가 예측에 큰 영향을 끼쳤다고 해석할 수 있습니다.

각 예측 별 특성의 영향력

▶ 회원 ID : 42375130



긍정적 영향

last_fico_score : 평균보다 낮은 최근 신용점수
dti : 높은 총 부채 상환율

부정적 영향

term : 짧은 대출기간
Installment : 낮은 월 상환금

예측 결과

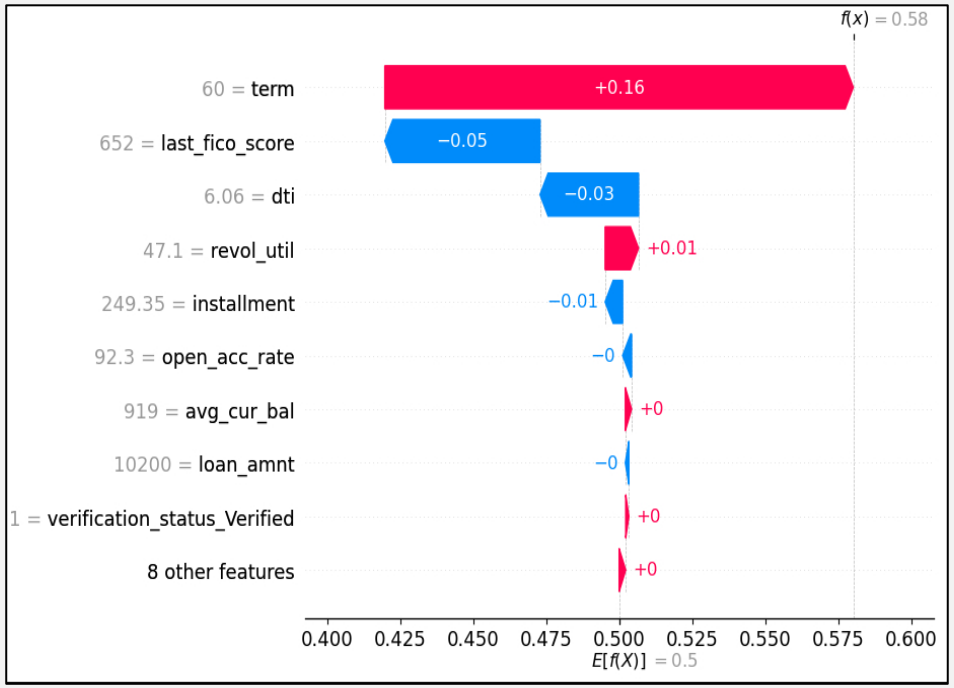
Charged off

모델 예측 해석

평균적인 최근신용점수(last_fico_score)로 인해 여러 특성들의 영향도가 높아짐을 확인할 수 있습니다.

각 예측 별 특성의 영향력

▶ 회원 ID : 10076060



긍정적 영향

last_fico_score : 평균적인 최근 신용점수
dti : 낮은 총 부채 상환율

부정적 영향

term : 긴 대출기간
revol_util : 높은 리볼빙 사용률

예측 결과

Fully paid

모델 최적화

특성 중요도가 0인 특성 제거 후 평가지표를 확인한 결과, 평가지표에 영향이 없음을 확인했습니다.

특성 중요도 = 0인 특성 제거

> 거주지역(addr_state) 제거

- **거주지역(addr_state) 지도시각화 분석**

그림 1. 주별 미상환 비율

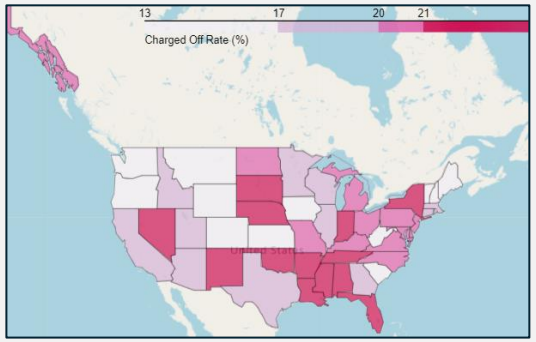
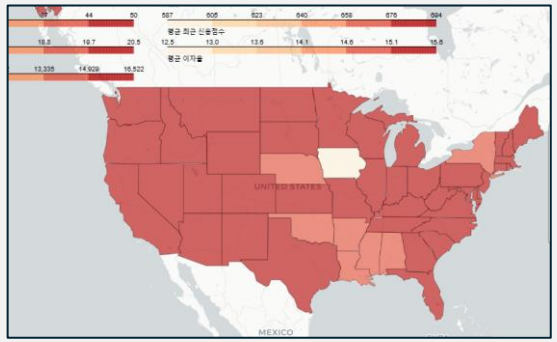


그림 2. 주별 최근 신용점수 평균



주별 미상환 비율의 차이는 있지만 모델 예측에 가장 큰 영향을 미치는
최근 신용점수(last_fico_score)의 주별 차이가 미미한 것으로 보입니다.

특성 별 제거 후 성능 확인

> 교차 검증의 검증세트 평균 평가지표 비교

Model	ROC-AUC	Recall
특성 제거 전	0.9462	0.9110
특성 제거 후	0.9462	0.9110

특성중요도가 0인 특성 제거 후 훈련을
진행해도 평가지표에 영향이 없습니다.

계산비용 절감 및 모델 간소화를 위해 제거하기로 결정했습니다.

모델 비교 및 최종 모델 선정

앙상블 모델과의 benchmark 비교를 통해, Decision Tree의 성능이 뒤떨어지지 않는다는 것을 확인했습니다.

Model		ROC-AUC	Recall
선정 모델	Decision Tree	0.9462	0.9110
앙상블 모델	Ada Boost	0.9495	0.9118
	Bagging	0.9484	0.9152
	CatBoost	0.9514	0.9136
	LGBM	0.9506	0.9141
	XGB	0.9503	0.9128
	GBM	0.9494	0.9142
	ET	0.9471	0.9072
	RF	0.9469	0.9116

최종 모델 : Decision Tree 로 선정

여러 앙상블 모델과 Decision Tree를 비교했을 때,
ROC-AUC 기준 모두 약 0.95정도로 **큰 성능의 차이가 없습니다**.

최종 모델 : Decision Tree

최종 Test set 의 평가지표를 확인했습니다.

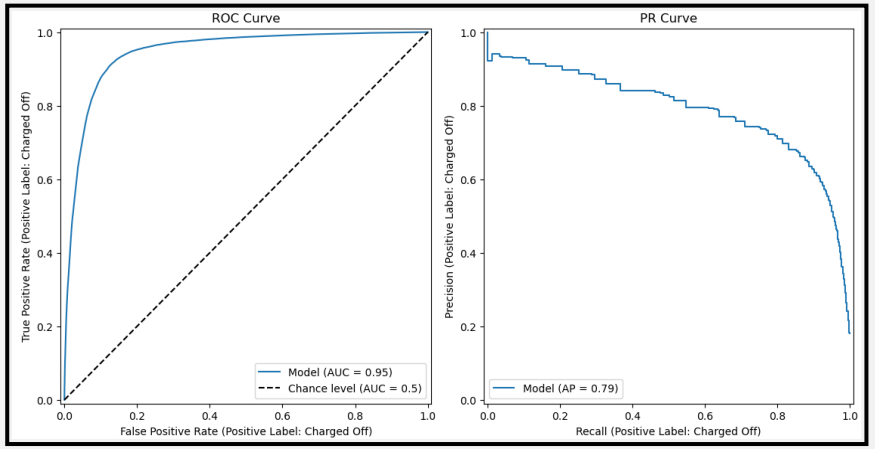
최종 모델의 Test set 평가지표 비교

	ROC-AUC	Recall
Best Model (Train)	0.9491	0.9161
Best Model (Test)	**0.9482**	**0.9142**

▽

새로운 데이터(Test set)에 대해서도
예측 성능이 우수한 것이 확인되었습니다.

최종 모델 Test set 모델 해석

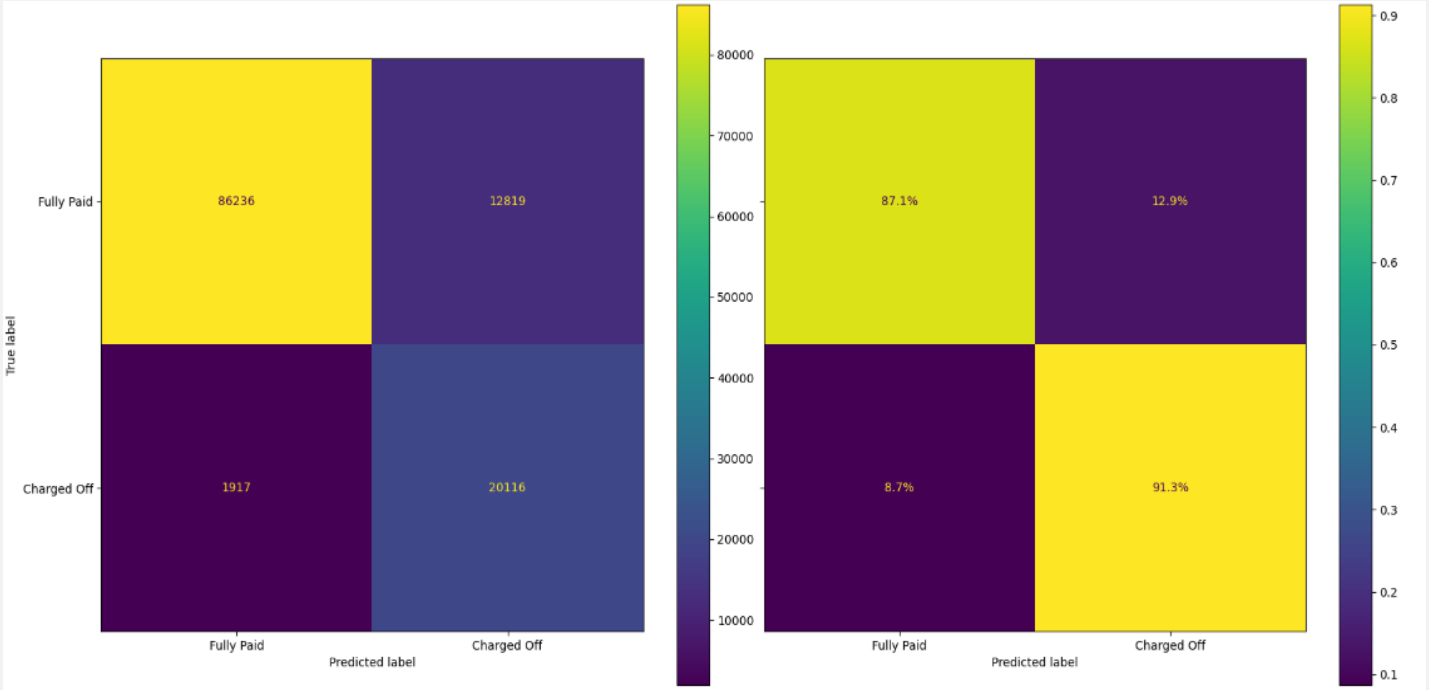


- 최종모델 성능이 약 0.95로 TPR이 높고 FPR이 낮은 정확한 예측 수행 능력을 확인할 수 있습니다.
- AP(Average Precision)가 약 0.8로 불균형한 데이터 셋을 잘 예측했다는 것을 확인할 수 있습니다.

최종 모델 평가

Test set에 대한 혼동행렬로 최종모델의 타겟 분류능력이 뛰어남을 확인했습니다.

혼동행렬(Confusion Matrix)을 통한 성능 확인



모델 평가

Charged off

91.3%

Fully paid

87.1%

- 각 타겟을 약 90% 구분하므로, **타겟을 분류하는 능력이 뛰어납니다.**
- 비즈니스 관점에 맞게 **미상환(Charged Off)을 구분하는 능력이 더 뛰어납니다.**

5장. 대시보드 : Streamlit

대시보드 구성

① '분석가'를 위한 **EDA** 페이지 – 데이터 현황과 특성을 빠르고 간결하게 확인할 수 있습니다.

② '투자자'를 위한 **예측** 페이지 – 대출자의 대출 상환 여부 예측을 확률과 함께 확인할 수 있습니다.

① 분석가(Analyst)

사용자

데이터 분석가

목적

데이터의 현황과 EDA 탐색 과정을 간결하고 직관적으로 파악할 수 있습니다.

기능

- 특성의 분포 및 타겟 비율 확인
- 특성간 상관 관계를 파악
- 지도를 통한 특성 분포 및 패턴 파악

▶ 아이디어 도출

② 투자자(Investor)

대출 투자자

투자 대상에 대한 대출 상환 여부 예측에 도움을 줄 수 있습니다.

- 각 특성의 값을 입력하여, 상환 여부를 예측
- 예측에 대한 확률 값을 제시
- 예측 결과에 대한 보조 설명

▶ 리스크 헷지

대시보드 시연

대시보드 시연

6장.아이디어 도출: P2P 비즈니스 방향성

아이디어 제안

현재 모델에 대한 고려사항

하나의 지표에 예측 의존성이 상당히 높습니다. 최근 신용점수(last_fico_score) 가 높으면 상환으로 분류되는 모델이 너무 단순하다는 문제가 될 수 있습니다.

LendingClub

아이디어 1. 비금융과 비정형 데이터를 새로운 변수로 활용 #정성적 평가 요소 추가

사용한 데이터의 특성들이 모두 금융과 신용관련에 집중하여 구성되어 있습니다.

리스크 관리의 전통적인 지표인 재무 특성과 더불어 **대출자의 개인 특성을 파악할 수 있는 비금융특성에 대한 개발을 제안합니다.**

동일한 직장에 동일한 연봉을 받고 있는 사람이라도 상환여부는 달라질 수 있습니다.

P2P 대출의 특성상 이미 1금융권에서 대출이 어려운 인원이 주 사용자로써 재무적 요소들로 고객 위험도를 세분화하기 어려울 수 있습니다.

따라서, 대출자별 고유의 특성이 반영된 정보를 발굴할 필요가 있습니다.

아이디어 제안

대출 거절, 투자를 받지 못한 대출자에 대한 고려사항

대출이 거절되거나, 투자를 받지 못한 대출자들에 대한 정성적인 관리가 필요합니다.

LendingClub

아이디어 2. 대출자를 위한 예측 해석 기반의 대출 전략 제시 # 대출 승인 # 투자 가능성 상승

모델의 투명성을 이용하여 대출이 거절되거나 투자를 받지 못한 대출자에게 **예측 해석 기반의 대출 전략을 제시**하여 대출자에게 대출 가능성에 부정적 영향을 미치는 요소들과, 리스크 관리에 대한 내용을 고지하고 알맞은 대출 조건을 이용하게 함으로써 **대출의 건전성과 건전한 대출 거래량을 상승**시킬 것으로 기대합니다.

ex) 특정 대출자의 예측 결과가 미상환(Charge off)으로 예측되지만, 대출기간 감소/리볼빙 사용률 향상을 통해 상환(Fully Paid)으로 예측되어 투자 가능성을 높일 수 있다는 점을 해당 대출자에게 제공

추후 분석 계획

1. settlement(재조정) / hardship(어려움) 특성 추가 분석

모델의 일반화 성능 문제로 제거했던 settlement / hardship 관련 특성(결측 90%이상) 을 따로 분리하고 분석하여, 해당 특성을 가진 고객의 특징을 알아볼 수 있습니다.

2. 최근신용점수(last_fico_score)별 고객 세분화 분석

가장 높은 중요도를 보인 최근신용점수(last_fico_score)를 중심으로 고객 세분화 분석을 진행하여 고객관점에서 최근신용점수(last_fico_score)를 높이기 위한 신용 관리 방향성을 제시할 수 있습니다.

3. 해석 가능성이 있는 다른 모델 생성 및 성능 비교

더욱 복잡하지만 해석 가능성이 높은 모델을 생성하여 이전 모델과 성능, 소요시간을 비교해볼 수 있습니다.

프로젝트 회고

하요한	프로젝트를 진행하면서 많은 시행착오와 어려움에 부딪혔지만, 팀원들 간의 긍정적인 협력과 의사소통이 있었기에 이러한 어려움들을 극복하고 프로젝트에 헌신한 덕분에 성공적으로 결과물을 도출할 수 있었다 생각합니다. 함께한 팀원들에게 감사의 말씀 전합니다.
곽용현	특성을 파악하여 선택하는데 많은 어려움을 겪었기 때문에 도메인 지식과 EDA과정의 중요성을 알게 되었고, 팀장님의 도움으로 프로젝트 기간동안 필요한 라이브러리를 사용법을 빠르게 배웠는데 향후 문제 해결 과정에 큰 도움이 될 것이라고 생각합니다.
박규리	여러 특성을 분석하고 처리방안을 고민하는 과정이 프로젝트의 핵심이라는 생각이 들었습니다. AUTOML인 Pycaret 사용법과 대시보드 구축법을 함께 공부하고 사용해 본 것이 추후에도 유용한 경험이 될 것이라고 생각합니다.
정혜영	다양한 EDA를 통해 특성 선택을 시도해 볼 수 있었던 프로젝트였습니다. 특성 의미 파악을 위해 데이터를 검증하고 특성간 관계 파악에 깊이 고민한 시간이었습니다. 모델 성능을 빠르게 비교할 수 있는 Pycaret 라이브러리를 사용해 볼 수 있어 AUTOML을 경험해볼 수 있어 좋았습니다.

참고 문헌

[1]Lending Club, (2022,01,26), "LendingClub Reports Fourth Quarter and Full Year 2021 Results", [https://ir.lendingclub.com/news/news-details/2022/LendingClub-Reports-Fourth-Quarter-and-Full-Year-2021-Results/default.aspx , [(accessed on 2024.03.04)].

[2]PMC, (2022,06,24), "P2P Lending Default Prediction Based on AI and Statistical Models ", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9222552/ , [(accessed on 2024.03.04)].

[3]Maureen Milliken (2023,12,21), "Lending Club", Debt.org, https://www.debt.org/credit/loans/personal/lending-club-review/ , [(accessed on 2024.03.04)].

[4]Shareholders Unite, (2022,10,14). "LendingClub Has A Killer Business Model", Seeking Alpha, https://seekingalpha.com/article/4546819-lendingclub-killer-business-model , [(accessed on 2024.03.04)].

[4]이현진, "딥러닝 기법을 이용한 P2P 소셜 대출 채무자 부도 예측모델에 관한 연구", 디지털콘텐츠학회논문지(J. DCS) Vol. 20, No. 7, pp. 1409-1416, 2019

[5] 최수만, 전동화, 오경주, "P2P 플랫폼에서의 대출자 신용분석 사례연구: 8퍼센트, 렌딧, 어니스트 펀드", 지식경영연구 제21권 제3호, 2020

[6]정재훈, (2023.05.04). " [기자수첩] 소셜임팩트 생태계에서 '관계형 금융' 꽃피우길", 소셜임팩트뉴스, https://www.socialimpactnews.net/news/articleView.html?idxno=519 , [(accessed on 2024.03.04)]

[7]강경훈, "핀테크 확산이 금융산업에 미치는 영향", 한국은행, (2017.10)

[8]강경훈, "관계형 금융의 활성화를 위한 과제", 하나금융경영연구소 제 5권 44호, (2015.11.16)

[9]김호현, (2024.01.09), "핀테크 주담대 갈아타기 출시, 카뱅·카카오페이·토스·핀크·네이버페이 참여", Business Post, https://www.businesspost.co.kr/BP?command=article_view&num=338738 , [(accessed on 2024.03.04)]

[10]김지혜, (2020.02.21), "'골리앗 무너뜨린 다윗'...美 렌딩클럽, 인터넷은행 인수", 전자신문, https://www.etnews.com/20200221000188 , [(accessed on 2024.03.05)]

[11]황춘매, 양철원, "온라인 P2P 대출 부도의 영향요인: 중국에 대한 실증분석", 산업연구 Journal of Industrial Studies(J.I.S) 47권 2호, 2023

[12]박재현, (2023.04.13), "[서경대 MFS] 핀테크 대출(Loan) 사례 – LendingClub", 서경TODAY, https://www.skuniv.ac.kr/227291 , [(accessed on 2024.03.05)]

[13] 천예은, 김세빈, 이자윤, 우지환, (2021.03.02), 설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구, 한국데이터정보과학회지 제32권 제2호, https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE10542538 , [(accessed on 2024.03.17)]

감사합니다.
긱고잉 이었습니다.

Final Project 2024.1.20~2024.03.20

하요한 . 곽용현. 박규리. 정혜영