빅데이터와 금융자료분석 프로젝트 (Team 4)

XGboost 알고리즘을 활용한 은행 대출의 부도 여부 예측 모델 구축

강상묵(20259013) 김형환(20249132) 유석호(20249264) 이현준(20249349) 최영서(20249430) 최재필(20249433)

1. 프로젝트 개요

본 프로젝트는 **빅데이터와 금융자료 분석 수업시간에 다룬 여러 데이터 전처리 기법과 머신러닝 알고리즘을 실제 금융데이터에 적용**해보고 시사점을 도출하기 위해 작성되었습니다.

이를 위해 미국 Lending Club의 P2P 대출 데이터를 사용하였으며, 전반적인 워크플로우는 아래와 같습니다.

1. 데이터의 구조, 특성 파악 (EDA)
2. 데이터의 결측치, 이상치 처리 등 전처리
3. 수치형, 범주형 변수의 처리 및 변수 선택 등 특성공학
4. 클래스 불균형 문제 처리 및 시각화를 위한 차원축소(T-SNE) 활용
5. XGBoost를 이용한 모델 구축 및 하이퍼파라미터 튜닝, 일반화 성능 평가

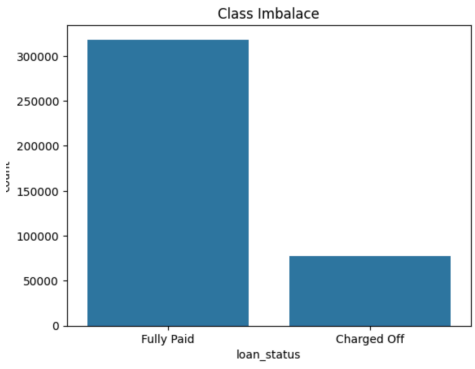
2. 데이터의 구조, 특성 (EDA)

데이터의 수집, 기본구조

미국 소재의 P2P 대출 전문은행인 Lending Club의 '07~'20년 대출 데이터를 사용하였습니다. (출처 : Kaggle)

약 40만개의 데이터로, 목적변수인 대출상태를 포함해 전체 27개의 칼럼(수치형 12 + 문자형 15)으로 이루어져있으며, 목적변수는 정상(상환, Fully paid) 및 부도(연체, Charged off)로 이진분류 문제입니다.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396030 entries, 0 to 396029
Data columns (total 27 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   loan_amnt             396030 non-null  float64
 1   term                  396030 non-null  object
 2   int_rate              396030 non-null  float64
 3   installment           396030 non-null  float64
 4   grade                 396030 non-null  object
 5   sub_grade             396030 non-null  object
 6   emp_title             373103 non-null  object
 7   emp_length            377729 non-null  object
 8   home_ownership        396030 non-null  object
 9   annual_inc            396030 non-null  float64
 10  verification_status   396030 non-null  object
 11  issue_d               396030 non-null  object
 12  loan_status           396030 non-null  object
 13  purpose               396030 non-null  object
 14  title                 394274 non-null  object
 15  dti                   396030 non-null  float64
 16  earliest_cr_line      396030 non-null  object
 17  open_acc              396030 non-null  float64
 18  pub_rec               396030 non-null  float64
 19  revol_bal             396030 non-null  float64
 20  revol_util            395754 non-null  float64
 21  total_acc             396030 non-null  float64
 22  initial_list_status   396030 non-null  object
 23  application_type      396030 non-null  object
 24  mort_acc              358235 non-null  float64
 25  pub_rec_bankruptcies  395495 non-null  float64
 26  address               396030 non-null  object
dtypes: float64(12), object(15)
memory usage: 81.6+ MB
```



데이터의 특성

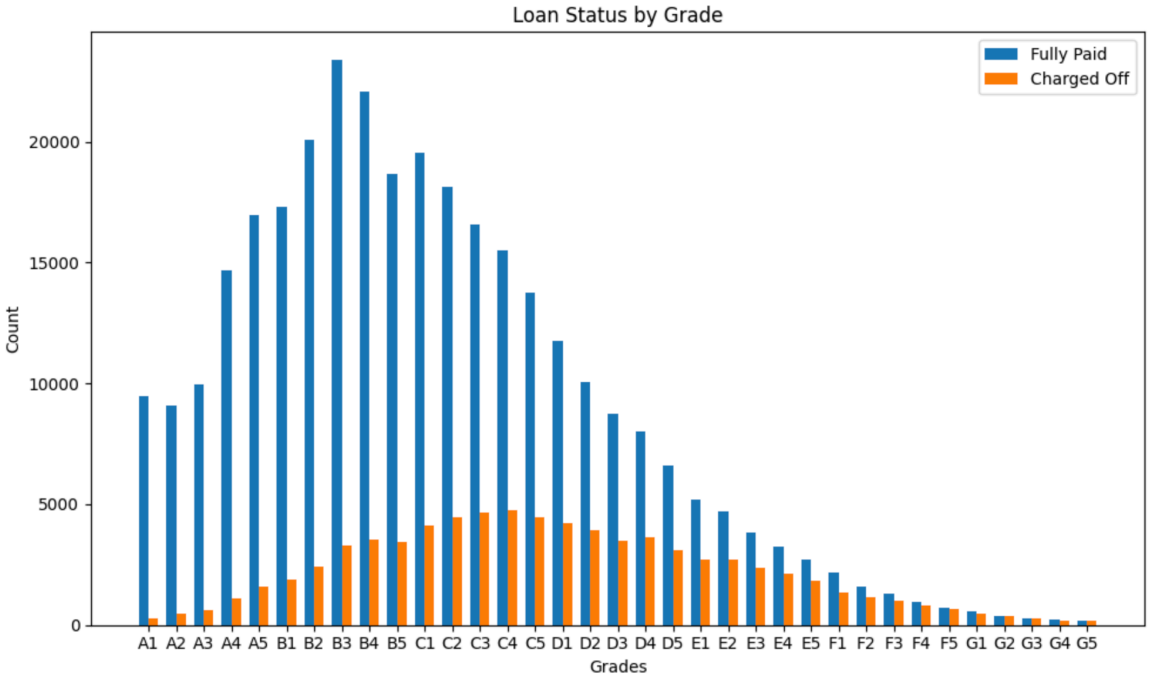
데이터의 특징과 주요 칼럼을 확인해보겠습니다.

수치형 변수들은 Heatmap을 이용해 상관관계를 간단히 알아보았습니다.

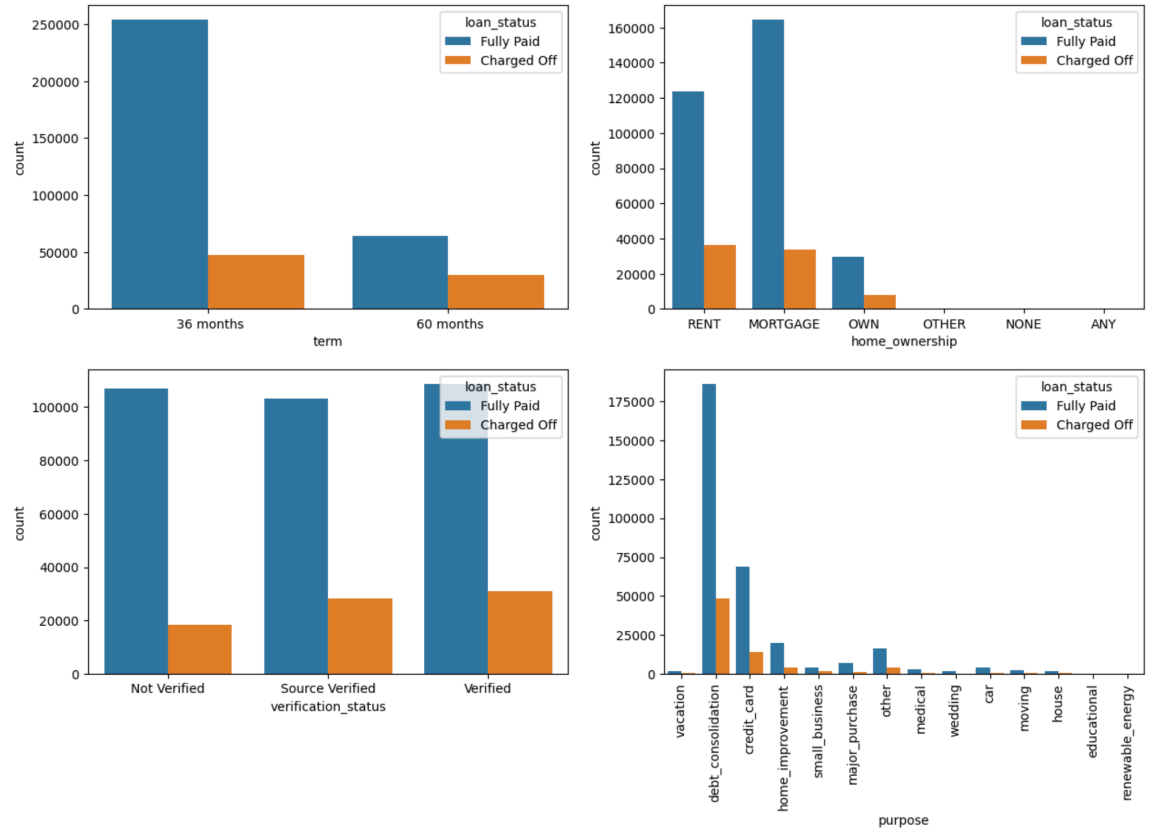
대체적으로 상관관계가 약해 다중공선성 문제는 없을 것으로 보이며, 일부 상관관계가 높은 변수가 있으나 따로 처리하지는 않고 이후에 특성공학에서 변수 선택으로 처리하도록 하겠습니다.

주요 **문자형 변수**를 살펴보겠습니다.

먼저, 신용등급이 하락함에 따라 부도율이 높아지는 추이를 보입니다.



이외의 주요 문자형 변수에 따른 목적변수의 분포를 살펴보겠습니다.



대출기간(term), 집보유형태(home_ownership), 대출목적(purpose)가 목적변수에 영향을 미치는 것으로 추정됩니다.

문자형 변수들의 고유값을 살펴본 결과, 일부 변수는 고유값이 과도하게 많아 분석에 제외해야할 필요성이 있어보입니다.

```
term                       2
grade                      7
sub_grade                 35
emp_title             173105
emp_length                11
home_ownership             6
verification_status        3
issue_d                  115
loan_status                2
purpose                   14
title                  48816
earliest_cr_line         684
initial_list_status        2
application_type           3
address               393700
dtype: int64
```

3. 데이터의 전처리

데이터의 전처리는 아래의 과정으로 실시하였습니다.

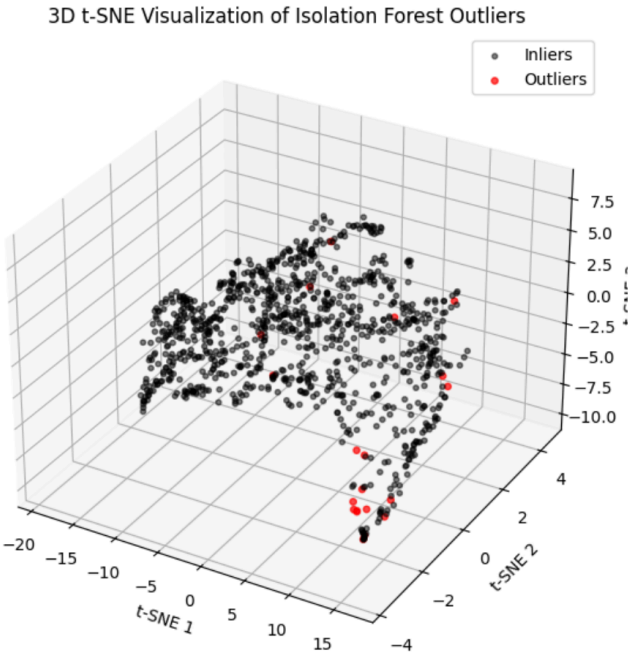
1. 일부 칼럼 가공
    - 주소는 우편번호만 추출, 대출기간 수치형 변환, 집 소유여부 극소수 값들 Other로 통합
2. 고유값이 너무 많거나, 다른 변수로부터 추출할 수 있는 칼럼 제거
    - 신용점수(대분류), 직업, 주소, 직업글자수, 최초연도, 발행일 등 7개
3. 문자형 변수 처리 : 라벨인코딩, 원핫인코딩 활용
    - 목적변수, 신용점수 : 라벨인코딩
    - 이외의 문자형 변수 : 원핫인코딩
4. 이상치 처리 : Isolation Forest을 이용해 수치형 변수의 1% 이상치 제거
    - T-SNE를 활용하여 3차원 축소 후 이상치 분류가 적합한지 시각화
5. Boruta 알고리즘을 이용한 변수선택 : 최종 데이터 가공
    - 원핫인코딩 대상을 제외한 변수들 중, 11개의 변수를 선택(1개 Tantative, 2개 reject)

4. 대출 연체여부 예측 모델 구축

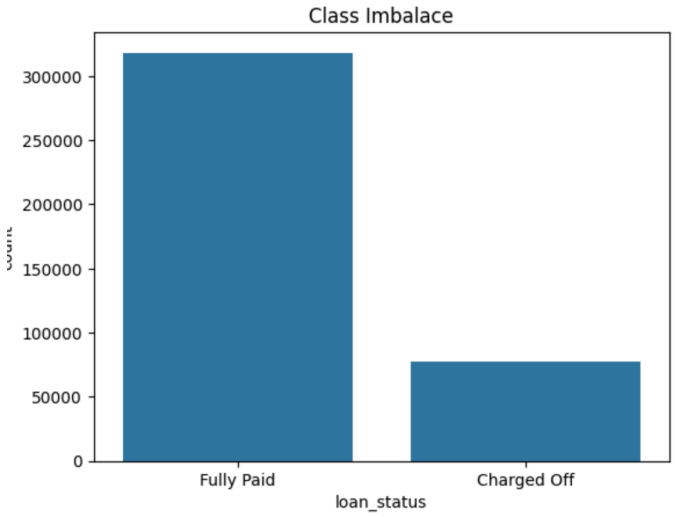
예측 모델은 XGBoost 알고리즘을 활용하여 구축하였으며, 모델링 이전에 클래스 불균형 문제를 해소하기 위해 **ADASYN을 이용하여 훈련데이터를 오버샘플링** 하였습니다.

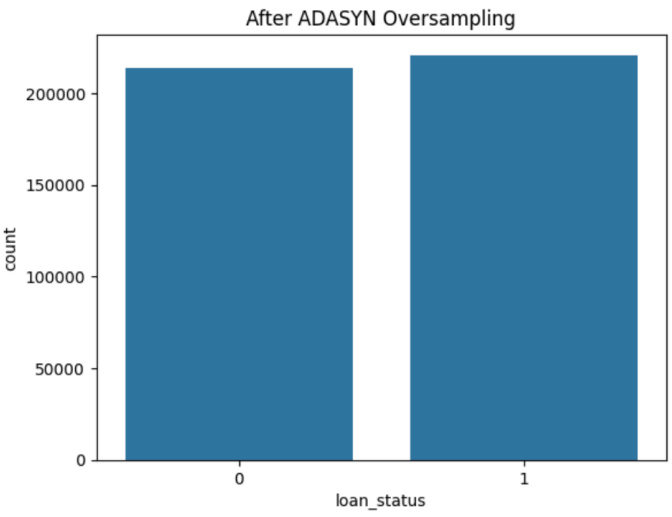
하이퍼파라미터 튜닝은 RandomizeCV를 이용하였으며, 최적 파라미터 및 성능은 아래와 같습니다.

```
Fitting 3 folds for each of 15 candidates, totalling 45 fits
Best Params: {'subsample': 0.8, 'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.05, 'colsample_bytree': 0.8}
Best ROC-AUC Score (CV): 0.923741343521181
```




```
<class 'pandas.core.frame.DataFrame'>
Index: 274448 entries, 3412 to 121958
Data columns (total 41 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   loan_amnt                             274448 non-null  float64
 1   term                                  274448 non-null  int64
 2   int_rate                              274448 non-null  float64
 3   installment                           274448 non-null  float64
 4   sub_grade                             274448 non-null  int64
 5   annual_inc                            274448 non-null  float64
 6   dti                                   274448 non-null  float64
 7   revol_bal                             274448 non-null  float64
 8   revol_util                            274448 non-null  float64
 9   total_acc                             274448 non-null  float64
 10  mort_acc                              274448 non-null  float64
 11  home_ownership_OTHER                  274448 non-null  bool
 12  home_ownership_OWN                    274448 non-null  bool
 13  home_ownership_RENT                   274448 non-null  bool
 14  verification_status_Source Verified   274448 non-null  bool
 15  verification_status_Verified          274448 non-null  bool
 16  purpose_credit_card                   274448 non-null  bool
 17  purpose_debt_consolidation            274448 non-null  bool
 18  purpose_educational                   274448 non-null  bool
 19  purpose_home_improvement              274448 non-null  bool
 20  purpose_house                         274448 non-null  bool
 21  purpose_major_purchase                274448 non-null  bool
 22  purpose_medical                       274448 non-null  bool
 23  purpose_moving                        274448 non-null  bool
 24  purpose_other                         274448 non-null  bool
 25  purpose_renewable_energy              274448 non-null  bool
 26  purpose_small_business                274448 non-null  bool
 27  purpose_vacation                      274448 non-null  bool
 28  purpose_wedding                       274448 non-null  bool
 29  initial_list_status_w                 274448 non-null  bool
 30  application_type_INDIVIDUAL           274448 non-null  bool
 31  application_type_JOINT                274448 non-null  bool
 32  zip_code_05113                        274448 non-null  bool
 33  zip_code_11650                        274448 non-null  bool
 34  zip_code_22690                        274448 non-null  bool
 35  zip_code_29597                        274448 non-null  bool
 36  zip_code_30723                        274448 non-null  bool
 37  zip_code_48052                        274448 non-null  bool
 38  zip_code_70466                        274448 non-null  bool
 39  zip_code_86630                        274448 non-null  bool
 40  zip_code_93700                        274448 non-null  bool
dtypes: bool(30), float64(9), int64(2)
memory usage: 33.0 MB
```



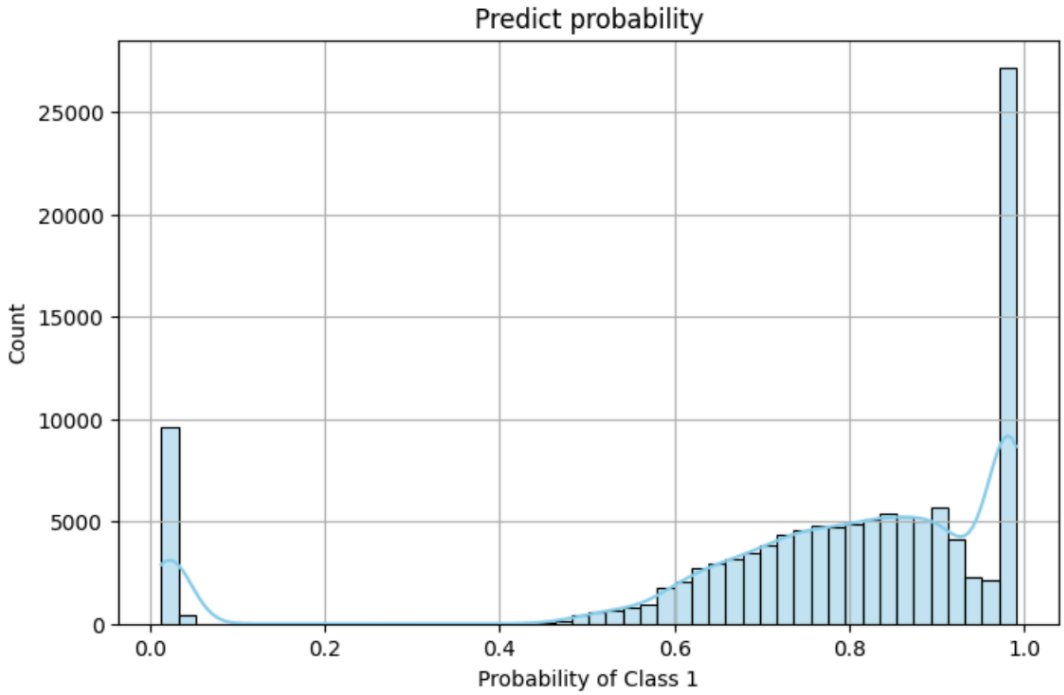


```
<< Best XGBoost model perfomance >>
F1 Score: 0.9344705841914274
ROC AUC: 0.8967291971591063

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.44      0.61     23370
           1       0.88      1.00      0.93     95439

    accuracy                           0.89    118809
   macro avg       0.93      0.72      0.77    118809
weighted avg       0.90      0.89      0.87    118809
```

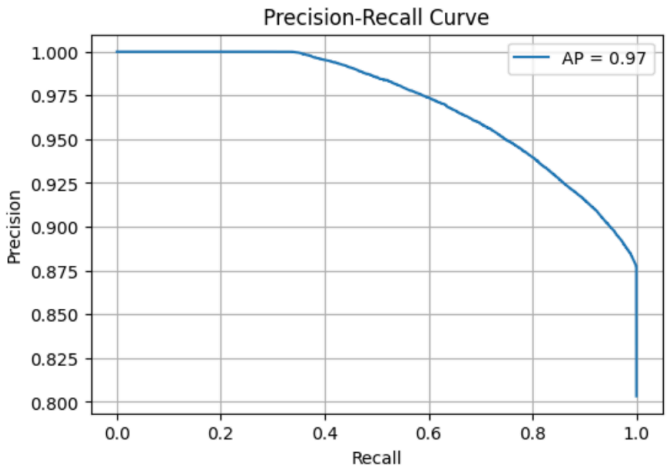
클래스 불균형을 고려하여 F1-score를 위주로 튜닝하였으며, 전반적으로 좋은 성능을 보여주었습니다.

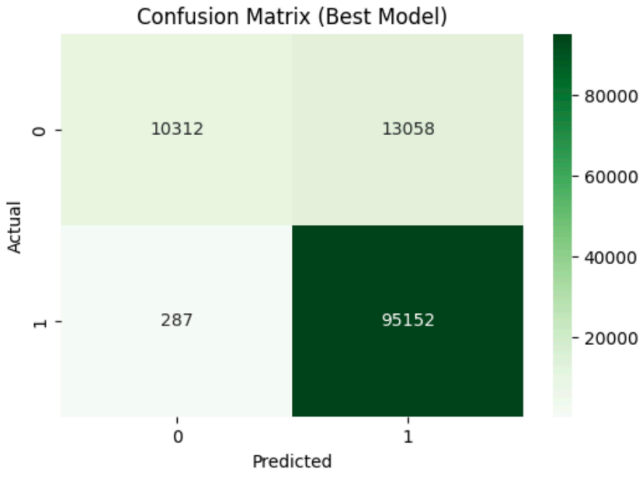
그러나, 여러 노력에도 불구하고 **샘플의 수가 상대적으로 적은 "부도" 경우에 대해 예측 성능이 다소 떨어지는 모습**이 관측되었습니다.

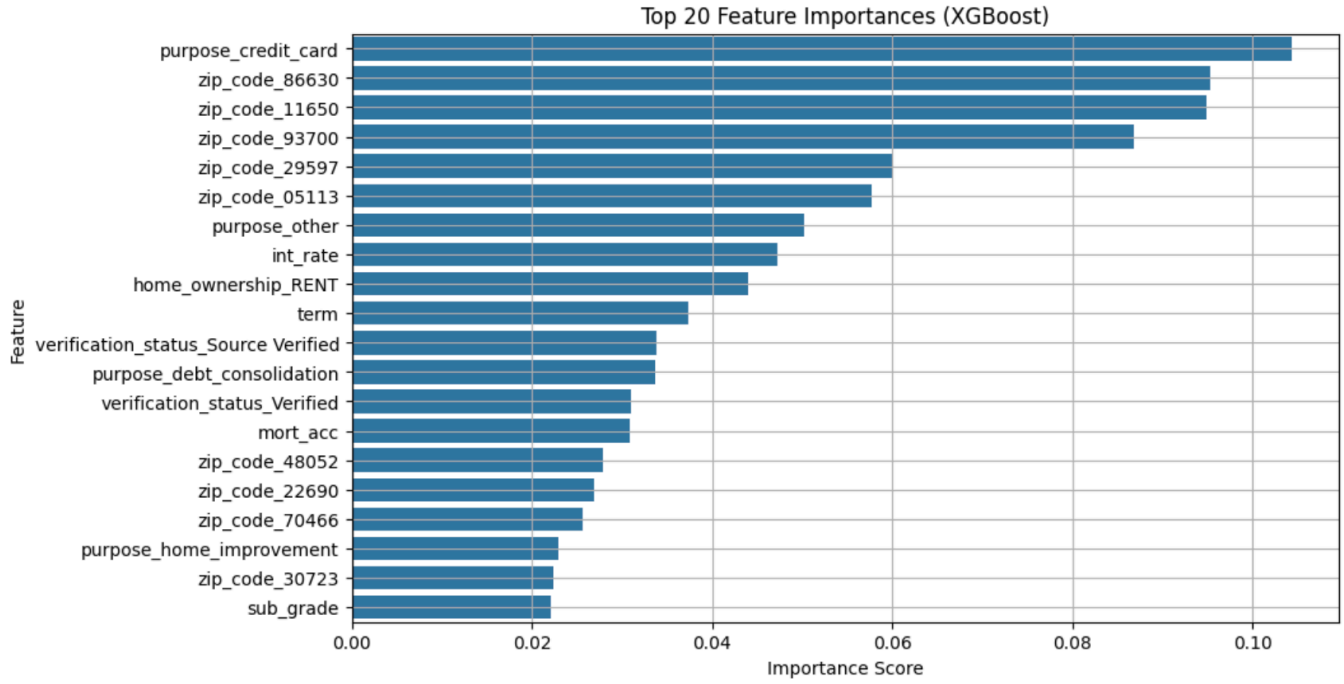
대부분의 정상 확률이 0.5보다 크게 산출되었으며, 이에 따라 실제 부도데이터를 정상데이터로 분류하는 것을 Confusion Matrix에서 확인할 수 있습니다.



마지막으로, 변수의 기여도에 대해 평가해보았습니다. 대출목적과 주거지가 큰 영향을 미치는 것을 확인할 수 있었으며, 이외에도 이자율(int_rate) 및 대출기간 순으로 분류에 영향을 미치는 것을 알 수 있었습니다.







5. 시사점

실제 대출 데이터를 이용하여 XGBoost 기반의 연체 여부 예측 모델을 구축한 결과, F1-score 및 ROC AUC 기준으로 높은 예측 성능을 보여주었습니다.

연체 여부에는 예상외로 소득이나 대출금액 이외에 대출목적 및 주거지, 이자율, 대출기간 등이 큰 영향을 미치는 것으로 나타났습니다.

그러나, 오버샘플링 및 모델의 여러 규제를 적용하였음에도 불구하고 샘플 비율이 다소 낮은 “부도”의 경우를 “정상”으로 예측하는 경우가 상당부분 발생하였습니다. (오버샘플링 기법을 바꾸거나, 언더샘플링을 하더라도 동일한 추이)

이는 데이터 자체의 한계이거나, Grediant Boost 계열의 알고리즘의 한계점으로 추정되며, 신경망 계열의 알고리즘을 사용한다면 모델의 성능이 보다 개선될 수 있을 것으로 추정됩니다.