Loan Default Predictor

https://github.com/yxing12/data1030-load-default-prob

Yunfei (Cynthia) Xing
Data Science Institute

October 19, 2023

01 Introduction

Overview

Problem

Loan Default - situation where a borrower fails to repay a loan according to the terms agreed upon.

Negatively impacts the borrower (credit rating) and the financial institution (revenue, reputation).

We'll try to leverage ML techniques to predict loan defaults.

Importance

The model hopefully offers more insights to enhance:

- Risk & Reward management for institutions
- Financial planning for borrowers
- Economic health - fewer unexpected defaults

Dataset

- Lending Club's loan data from 2007-2017 adopted from Kaggle
  - Aggregated data from Lending Club website
- Feature description of all 150 columns adapted from another Kaggle source
  - borrower details, credit history, loan specs, financials, other misc info

- Full Dataset shape: (1,646,801, 150)
- Cleaned Dataset shape: (815,056, 32)
- I.I.D: Yes
- Missing Value: Yes

- Target variable: 'loan_status' ('defaulted' in cleaned dataset) - categorical
- **Classification Problem** - categorize the loan into one of the two class (fully paid/defaulted) based on loan's and borrower's financials

EDA & Data Preparation

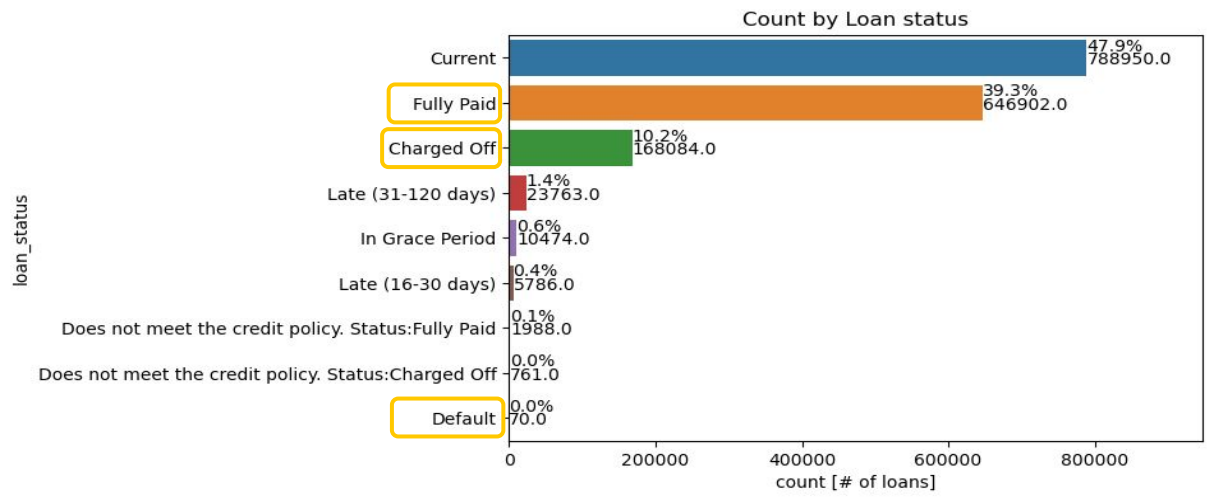
02

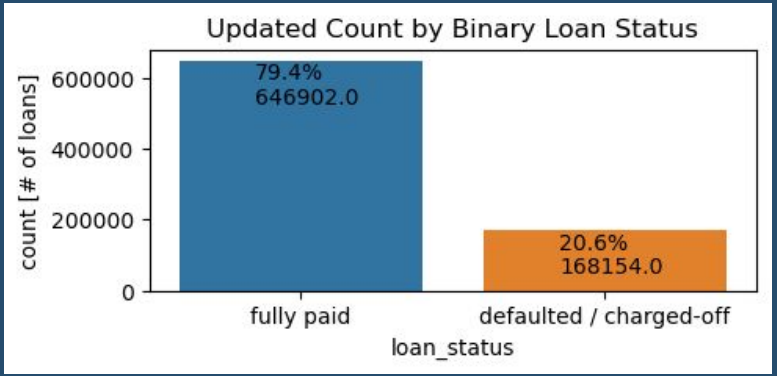
Target Variable

- Only focus on fully-paid or charged-off/defaulted completed loans

- updated dataset shape: (815056, 150)
- ~80% loans fully paid ~20% charged off / defaulted

→ a somewhat unbalanced classification problem





* A charge-off is a debt that a creditor has given up trying to collect on after borrower have missed payments for several months.

Feature Selection

We'll only keep the most relevant features to feed into our ML model.
Features will be selected based on the following criteria (determined during EDA):

Sparse Features

Eliminate features with > 20% missing values

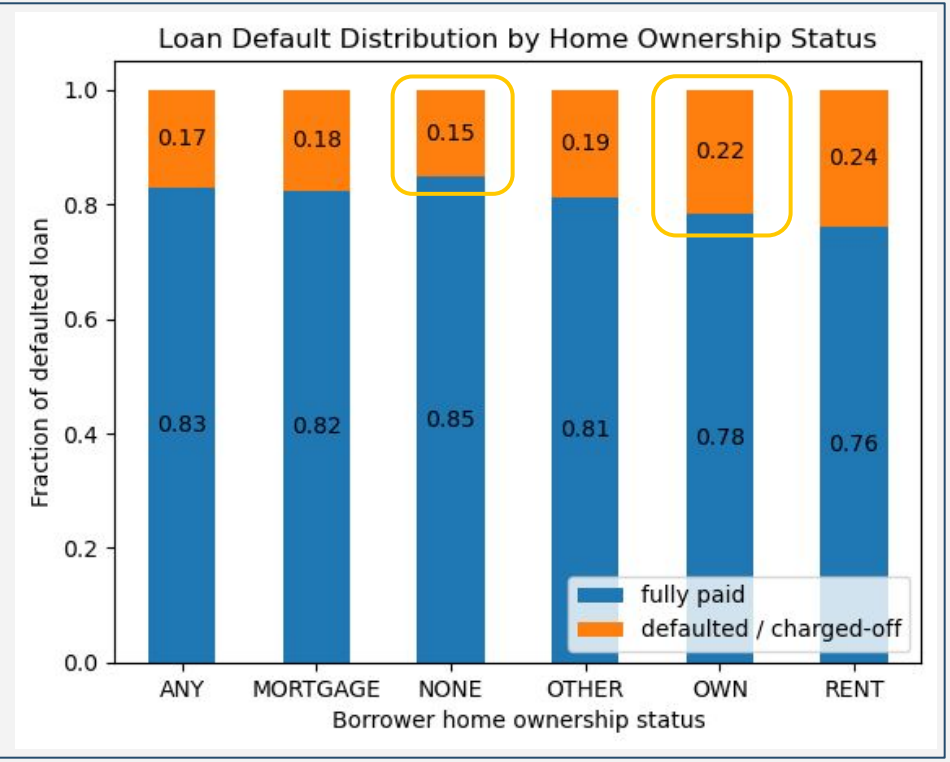
Non-disclosed Features

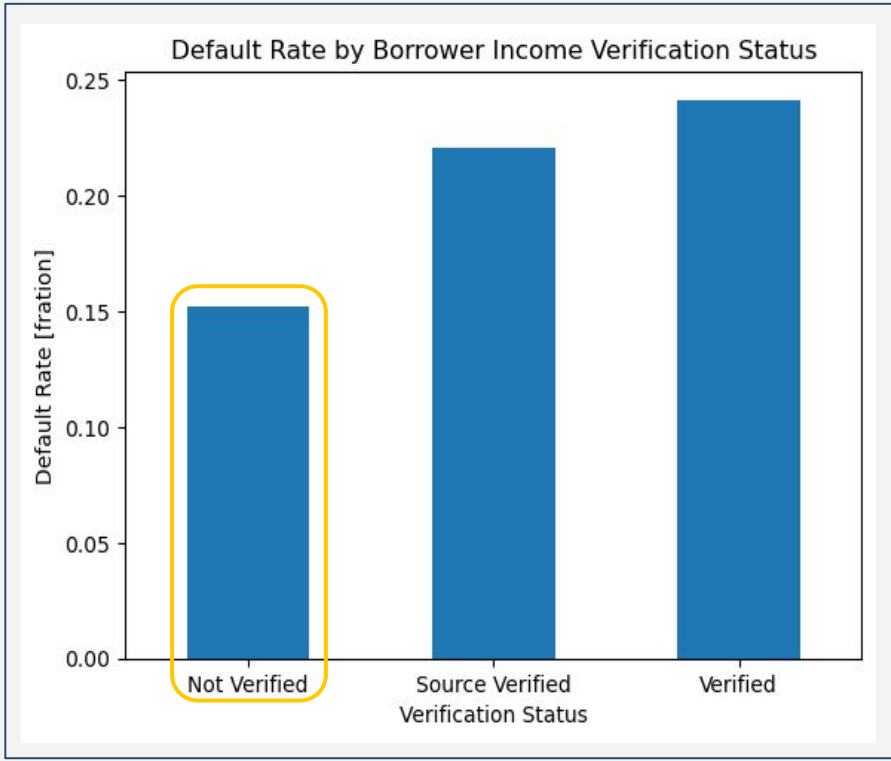
Eliminate features that are not available to borrowers when making investment decisions

Low-impact & Redundant Features

Eliminate features with low correlation with the target variable "defaulted", and one of the features fully correlated with each other

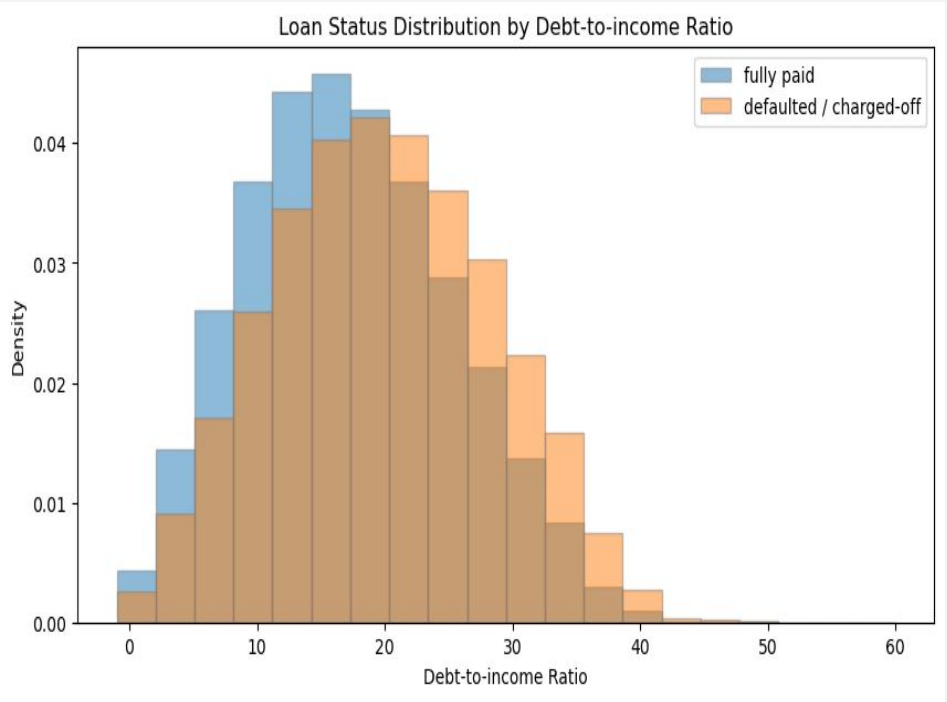
Interesting finds

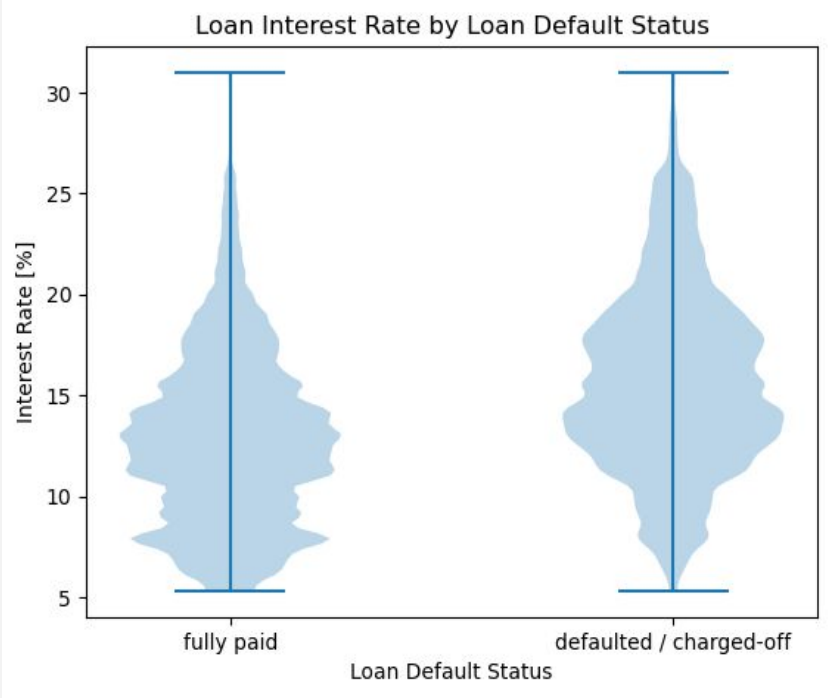




verification_status	
Not Verified	0.152264
Source Verified	0.220319
Verified	0.241354

Other observations





03 Splitting & Preprocessing

Missing Values in cleaned dataset

- Full dataset
  - 90/150 features with missing values
    - float & object
  - 100% rows contain missing values
- Cleaned dataset
  - 11/32 features with missing values
    - float only
  - 9.28% rows contain missing values

```
percentage of missing values features:
dti                       0.004417
revol_util                0.060119
acc_open_past_24mths      5.800951
avg_cur_bal               8.286424
bc_open_to_buy            6.769842
bc_util                   6.829837
mo_sin_old_rev_tl_op      8.285075
mo_sin_rcnt_rev_tl_op     8.285075
mort_acc                  5.800951
num_actv_rev_tl           8.284952
pub_rec_bankruptcies      0.085516
dtype: float64


 data types of the features with missing values:
dti                       float64
revol_util                float64
acc_open_past_24mths      float64
avg_cur_bal               float64
bc_open_to_buy            float64
bc_util                   float64
mo_sin_old_rev_tl_op      float64
mo_sin_rcnt_rev_tl_op     float64
mort_acc                  float64
num_actv_rev_tl           float64
pub_rec_bankruptcies      float64
dtype: object


percentage of data with missing values: 9.283411201193537
```

Splitting

Basic train_test_split

- Large i.i.d dataset (>800k rows)
- 98-1-1 train-val-test split

```
X_other shape (806905, 31)
--------------------------
train shape (798754, 31) (798754,)
val shape (8151, 31) (8151,)
test shape (8151, 31) (8151,)

y_train Distribution: Fully-Paid: 79.36%, Defaulted/Charged-off: 20.64%
y_val Distribution: Fully-Paid: 79.86%, Defaulted/Charged-off: 20.14%
y_test Distribution: Fully-Paid: 79.35%, Defaulted/Charged-off: 20.65%
```

K-fold for Cross Validation

- Not choosing StratifiedKFold because target variable distribution is NOT extreme
- Split out 1% test data
- Perform k-fold on X_other using n=5
- **Preprocess each train/val set during each iteration**

Preprocessing

OneHotEncoder

Use SimpleImputer to first replace Null value with a string

ex.
'home_ownership',
'verification_status',
'application_type'

Categorical

OrdinalEncoder

Use SimpleImputer to first replace Null value with a string

ex.
'grade',
'sub_grade'

Ordinal

MinMaxScaler

Continuous variables with a clear range

ex.
'loan_amnt',
'installment',
'fico_score'

Continuous

StandardScaler

All other continuous variables

ex.
'term',
'dti',
'earliest_cr_line',
'open_acc',
'total_pymnt'

Continuous

103 Features after preprocessing

Thank you!

Questions?

Appendix

Sparse Features

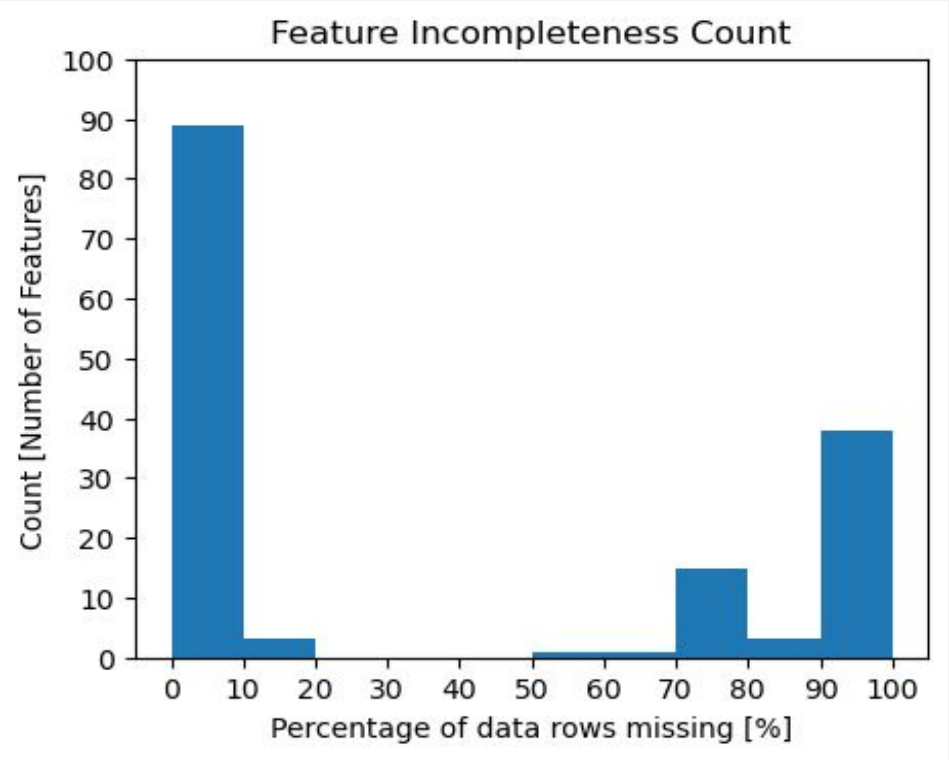
Eliminate features with > 20% missing values

percentage of data with missing values: 100.0

Features with < 20% missing data: 92
i.e. 61.33 % of total features

Features with 20%-50% missing data: 0
i.e. 0.0 % of total features

Features with > 50% missing data: 58
i.e. 38.67 % of total features



Selected Eliminated features:

['all_util', 'annual_inc_joint', 'debt_settlement_flag_date', 'deferral_term', 'desc', 'dti_joint', 'hardship_amount', 'hardship_dpd', 'hardship_end_date', 'hardship_last_payment_amount', 'hardship_length', 'hardship_loan_status', 'hardship_payoff_balance_amount', 'revol_bal_joint', 'sec_app_chargeoff_within_12_mths', 'sec_app_collections_12_mths_ex_med', 'sec_app_earliest_cr_line', 'sec_app_fico_range_high', 'sec_app_fico_range_low', 'verification_status_joint']

Non-disclosed Features

Eliminate features that are not available to investors when making investment decisions

Manually went through the remaining 92 features and selected 42 to keep

LoanStatNew	Description
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
defaulted	Current status of the loan (0 = fully paid, 1 = off-charged / defaulted)
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
id	A unique LC assigned ID for the loan listing.
initial_list_status	The initial listing status of the loan. Possible values are – W, F
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
last_pymnt_amnt	Last total payment amount received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mort_acc	Number of mortgage accounts.
num_actv_rev_tl	Number of currently active revolving trades
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_rec_int	Interest received to date
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

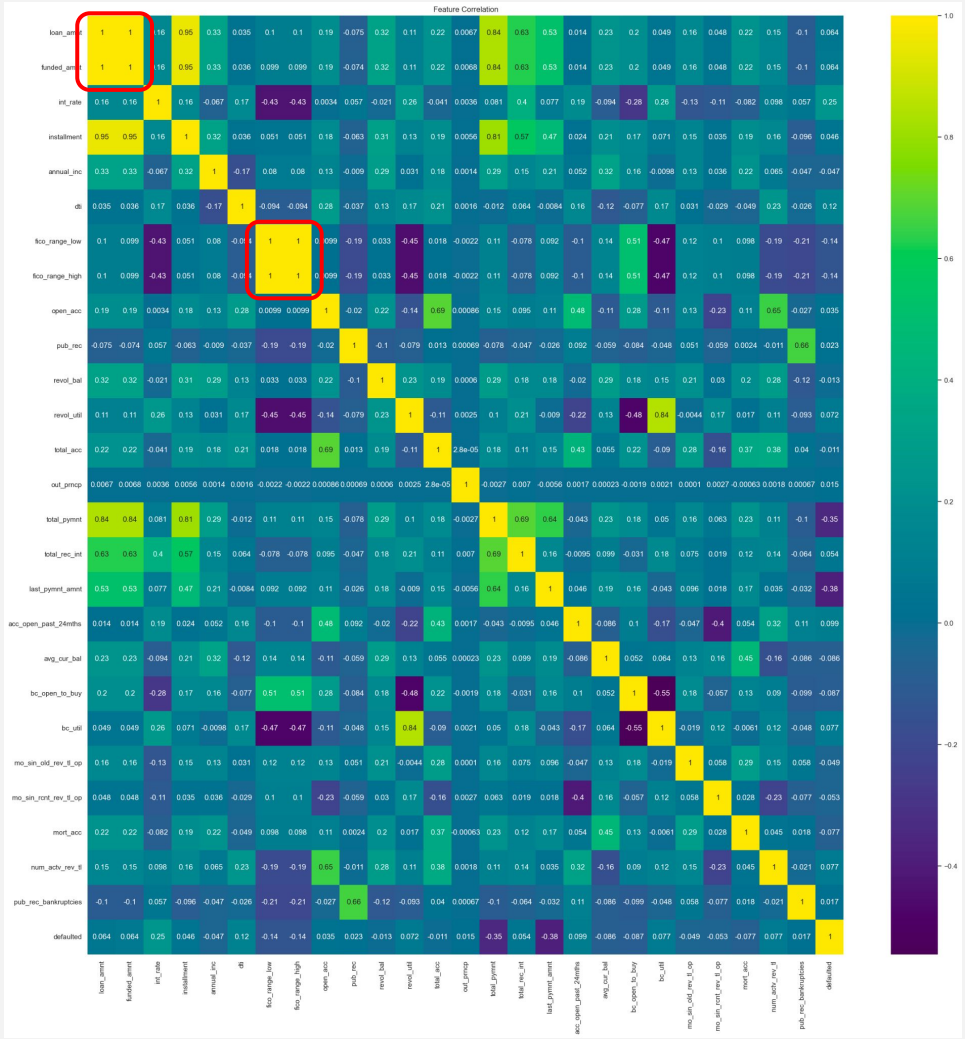
Low-impact & Redundant Features

Eliminate features with low correlation with the target variable "defaulted", and one of the features fully correlated with each other

Remove one of the fully correlated features to reduce the risk of overfitting and save computational resources

- loan_amnt, funded_amnt: remove funded_amnt

- fico_range_high and fico_range_low: create a new feature averaging the two



Low-impact & Redundant Features

Eliminate features with low correlation with the target variable "defaulted", and one of the features fully correlated with each other

Remove features not closely correlated with target variable ‘defaulted’

After this step, 37 features remain. Following further data cleaning and reformatting, the cleaned dataset retains 32 features.

```
correlation_defaulted = abs(correlation['defaulted']).sort_values(ascending=False)
print(correlation_defaulted)
```

```
defaulted                1.000000
last_pymnt_amnt          0.381411
total_pymnt              0.350998
int_rate                 0.247824
fico_score               0.139470
dti                      0.123057
acc_open_past_24mths     0.099007
bc_open_to_buy           0.086927
avg_cur_bal              0.085786
num_actv_rev_tl          0.077226
bc_util                  0.077168
mort_acc                 0.077101
revol_util               0.072213
loan_amnt                0.064161
total_rec_int            0.053620
mo_sin_rcnt_rev_tl_op    0.053411
mo_sin_old_rev_tl_op     0.048513
annual_inc               0.046680
installment              0.046310
open_acc                 0.034675
pub_rec                  0.023115
pub_rec_bankruptcies     0.017334
out_prncp                0.014772
revol_bal                0.013156
total_acc                0.011180
Name: defaulted, dtype: float64
```