AMSTERDAM
DATA
COLLECTIVE

Scorecard Development

24-10-2022 Zening Zheng

Contents

1. Overview
2. Final Model + Features + Deployment
3. Approach
4. Conclusions & Recommendations

AMSTERDAM
DATA
COLLECTIVE

1 Overview

Goal

- Automate credit scoring process
  - Select important features
  - Built credit scorecard

Approach

- **Data exploration & cleaning**
  - Descriptive analysis
  - Data quality check
    - Convert data to right format, Imputation of missing value, outlier
- **Variable selection (statistics & expert-based)**
  - Information value
    - Binning – reduce dimensionality
    - Transform variables to WOE score
  - Multicollinearity – remove highly correlated variables
  - Lasso regression - penalize complexity
  - Expert
- **Testing**
  - Split data, oversampling technique
  - Modeling
    - Logistic regression
    - Decision tree
  - Gini
- **Scorecard development**

AMSTERDAM DATA COLLECTIVE

AMSTERDAM
DATA
COLLECTIVE

2 Final Model

Logistic regression

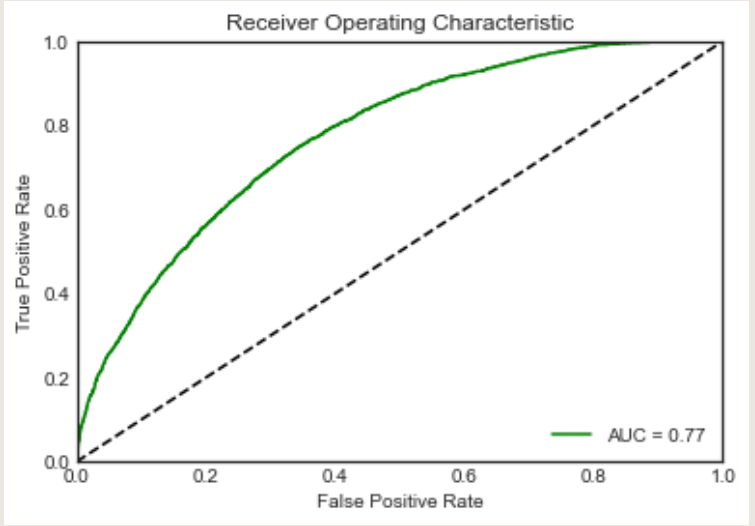
2 Final Model

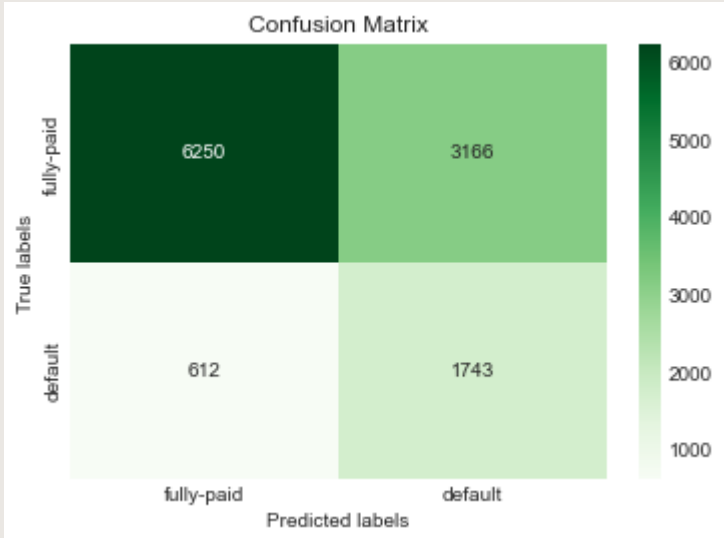
Model: logistic regression

Features : 13

['term', 'installment', 'int_rate', 'home_ownership','annual_inc', 'verification_status', 'addr_state', 'dti', 'fico_range_high', 'revol_util', 'mort_acc', 'age', 'pay_status', 'loan_amnt']

Criteria	Value
GINI	0.54
R squared	0.68
Precision	0.63
Recall	0.69
F-1 score	0.62





ERDAM

ECTIVE

2 Final Model - deployment

		Variable	Binning	Score
Variable				
addr_state	112	addr_state	VT	79
	73	addr_state	DC	38
	103	addr_state	OR	37
	87	addr_state	ME	34
	105	addr_state	RI	33
...	...	...	...	...
term	0	term	36	10
	1	term	60	-26
verification_status	63	verification_status	Not Verified	10
	64	verification_status	Source Verified	-2
	65	verification_status	Verified	-5

166 rows × 3 columns

AMSTERDAM DATA COLLECTIVE

AMSTERDAM
DATA
COLLECTIVE

3 Approach

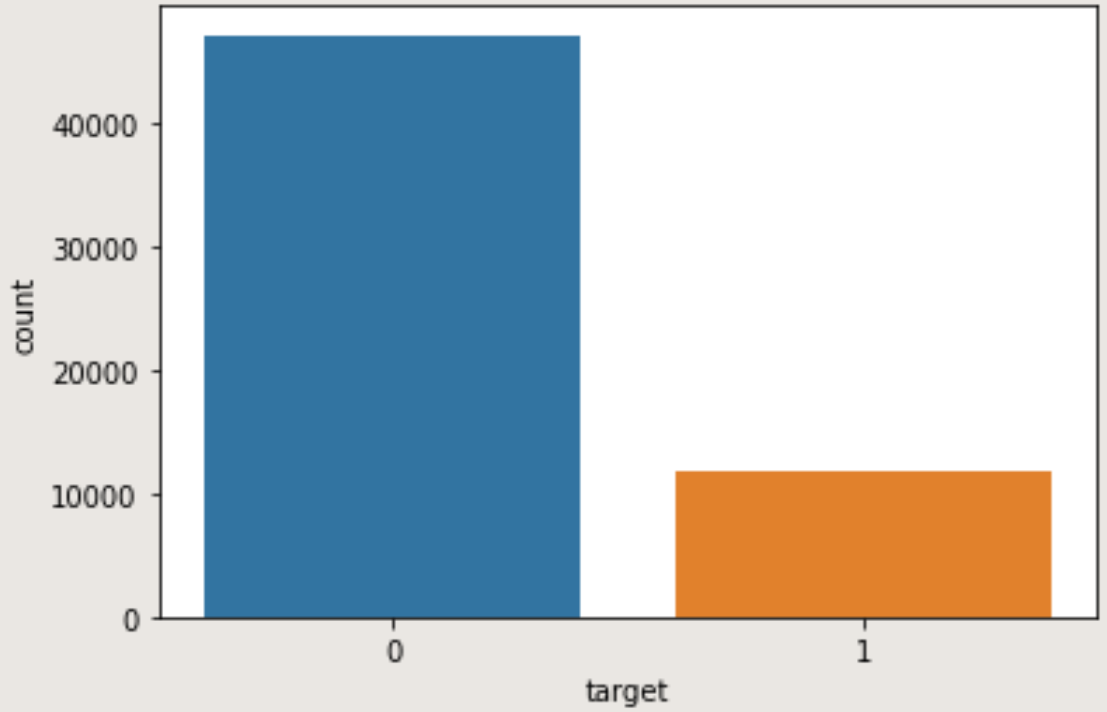
- **Data exploration & cleaning**
- **Variable selection (statistics & expert-based)**
- **Testing**
- **Scorecard development**

3 Approach - Data exploration & cleaning

Default rate: 0.2

number of default: 11772

number of non-default: 47080



Drop: `loan_status_2, issue_d`

Transform: `pay_status (-2,-1 = pay duly)`

AMSTERDAM DATA COLLECTIVE

3 Approach - Data exploration & cleaning

Descriptive analysis

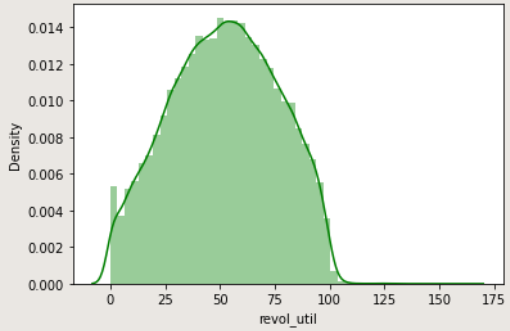
	target	loan_amnt	term	int_rate	installment	emp_length	annual_inc	dti	fico_range_low	fico_range_high	open_acc	pub_rec	revol_bal	revol_util	mort_acc	pub_rec_bankruptcies	age	pay_status
count	58852.0	58852.00	58852.00	58852.00	58852.00	55510.00	58852.00	58852.00	58852.00	58852.00	58852.00	58852.00	58852.00	58819.00	56813.00	58825.00	58852.00	58852.00
mean	0.2	14281.25	41.65	13.18	434.54	5.99	76732.13	17.99	695.97	699.97	11.62	0.22	16196.97	51.97	1.68	0.13	35.29	0.05
std	0.4	8617.25	10.18	4.75	258.65	3.68	73901.82	8.34	31.74	31.74	5.47	0.59	21133.72	24.43	2.02	0.38	9.36	1.14
min	0.0	1000.00	36.00	5.31	30.65	0.00	6695.00	0.00	660.00	664.00	1.00	0.00	0.00	0.00	0.00	0.00	20.00	-2.00
25%	0.0	7800.00	36.00	9.67	247.29	3.00	46000.00	11.77	670.00	674.00	8.00	0.00	6009.00	33.70	0.00	0.00	28.00	-1.00
50%	0.0	12000.00	36.00	12.73	373.22	6.00	65000.00	17.52	690.00	694.00	11.00	0.00	11110.50	52.50	1.00	0.00	34.00	0.00
75%	0.0	20000.00	36.00	15.99	572.60	10.00	90500.00	23.87	710.00	714.00	14.00	0.00	19836.25	70.80	3.00	0.00	41.00	1.00
max	1.0	40000.00	60.00	30.99	1607.80	10.00	6998721.00	49.94	845.00	850.00	67.00	21.00	1044210.00	162.00	24.00	8.00	78.00	9.00

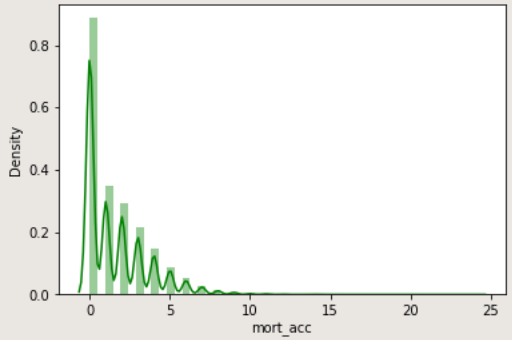
3 Approach - Data exploration & cleaning

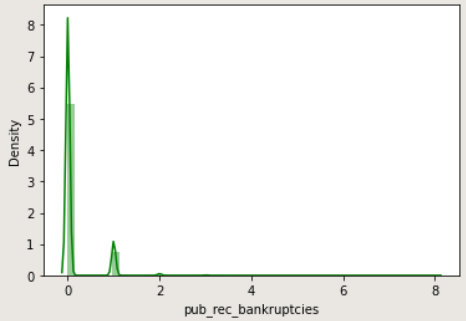
Missing value

- Emp_length –left for binning
- revol_util – mean
- mort_acc – 0
- pub_rec_bankruptcies

Outlier - binning







	column	number of null values	proportion
0	target	0	0.000000
1	loan_amnt	0	0.000000
2	term	0	0.000000
3	int_rate	0	0.000000
4	installment	0	0.000000
5	sub_grade	0	0.000000
6	emp_length	3342	0.056787
7	home_ownership	0	0.000000
8	annual_inc	0	0.000000
9	verification_status	0	0.000000
10	issue_d	0	0.000000
11	purpose	0	0.000000
12	addr_state	0	0.000000
13	dti	0	0.000000
14	fico_range_low	0	0.000000
15	fico_range_high	0	0.000000
16	open_acc	0	0.000000
17	pub_rec	0	0.000000
18	revol_bal	0	0.000000
19	revol_util	33	0.000561
20	mort_acc	2039	0.034646
21	pub_rec_bankruptcies	27	0.000459
22	age	0	0.000000
23	pay_status	0	0.000000

3 Approach - Variable selection

Benefits of Using WoE transformation

- **It can treat outliers**.
- It can **handle missing values** as missing values can be binned separately.
- Since WOE Transformation **handles categorical variable** so there is no need for dummy variables.
- WoE transformation helps you to **build strict linear relationship with log odds**. Otherwise it is not easy to accomplish linear relationship using other transformation methods such as log, square-root etc.

- Information Value (IV) comes from information theory, it measures the predictive power of independent variables

3 Approach - Variable selection

Binning: Numeric -> 10 bins

- Transform to Woe value
- Information value: >0.02
- Final variables: 16

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

```
['loan_amnt', 'term', 'int_rate', 'installment', 'sub_grade', 'home_ownership','annual_inc', 'verification_status',
'addr_state', 'dti','fico_range_low',  'fico_range_high', 'revol_util', 'mort_acc', 'age', 'pay_status' ]
```

AMSTERDAM DATA COLLECTIVE

3 Approach - Variable selection

Multicollinearity

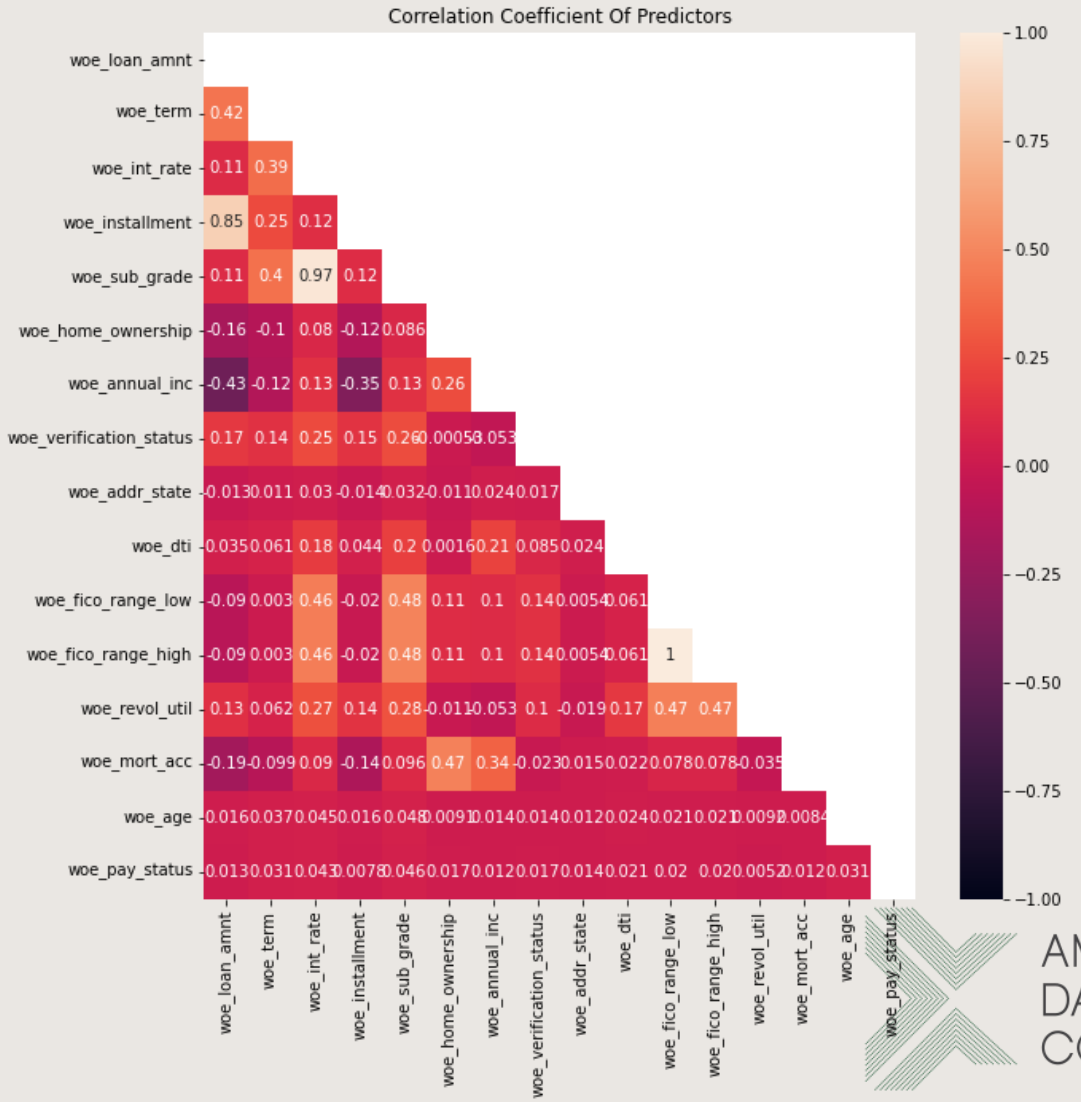
VIF>5

Variables with high correlation:

'installment', 'loan_amnt',

'sub_grade', 'int_rate',

'fico_range_high', 'fico_range_low'



AMSTERDAM DATA COLLECTIVE

3 Approach - Variable selection

Lasso - penalize complexity, coefficient equal to 0

No need to drop any variables

3 Approach - Variable selection

Final features: 13

['term', 'installment', 'int_rate', 'home_ownership','annual_inc', 'verification_status', 'addr_state', 'dti', 'fico_range_high', 'revol_util', 'mort_acc', 'age', 'pay_status', 'loan_amnt']

AMSTERDAM DATA COLLECTIVE

3 Approach – testing, model comparison

Split data (0.2-0.8)

Oversampling (SMOTE)

Modeling

- Logistic regression
- Decision tree

AMSTERDAM DATA COLLECTIVE

3 Approach – testing

Criteria	Logistic - Value	Decision tree value
GINI	**0.54**	**0.54**
R squared	0.68	0.72
Precision	**0.63**	**0.59**
Recall	0.69	0.58
F-1 score	0.62	0.58
Accuracy	0.54	0.72

AMSTERDAM DATA COLLECTIVE

3 Approach – scorecard development

Scorecard

Score = (β×WoE+ α/n)×Factor + Offset/n

odds of 1:1 at 650 points and the pdo of 50 (odds to double every 20 points)

		Variable	Binning	Score
Variable				
addr_state	112	addr_state	VT	79
	73	addr_state	DC	38
	103	addr_state	OR	37
	87	addr_state	ME	34
	105	addr_state	RI	33
...	...	...	...	...
term	0	term	36	10
	1	term	60	-26
verification_status	63	verification_status	Not Verified	10
	64	verification_status	Source Verified	-2
	65	verification_status	Verified	-5

166 rows × 3 columns

Need further validation of the model

AMSTERDAM
DATA
COLLECTIVE

4 Conclusions & Recommendations

4 Conclusions & Recommendations

Conclusions

- 13 final Variables
- Gini : 0.54

Recommendations

- Date variable, outlier
- Customized bin for each variable
- Combine categories with similar woe
- More models

AMSTERDAM
DATA
COLLECTIVE