Loan Prediction on LendingClub Issued Loans Dataset

ALEX MARIN FELICES

Visual Analytics

20th December 2021
Final Project

Index

- Introduction and motivation
- Description, objectives and expected benefits
- Expected delivery/output
- Dataset and required data
- Visualizations
- Python Code in Jupyter Notebook
  - Dataset Exploratory Data Analysis (EDA)
  - Data Wrangling
  - Pandas Profiling
  - Classification Models
- Dashboard
- Results and Evaluation
- Bibliography

Introduction and Motivation

Description, objectives and expected benefits

DESCRIPTION

- Project applying good visualisation techniques to find good loans.

OBJECTIVES

- Develop a classification model
- Create a training dataset
- Create a cluster method
- Identify the most important causes

EXPECTED BENEFITS

- Assess whether or not a new customer is likely to pay back the loan.

LendingClub

LendingClub Bank, N.A.

DATASET

- Lending Club, a US peer-to-peer lending company.
- Issued Loans between 2007 and 2017.
- 1,646,717 rows
- 74 columns

REQUIRED DATA

- **Customer data:** Home ownership, annual income, loan purpose
- **Loan data:** Loan amount, term, interest rate, grade, loan status, ...
- **Location**

Visualization

Good visualisations of the data so that it is easy and fast to understand the underlying characteristics of the data

MATPLOTLIB/SEABORN

PANDAS PROFILING

TABLEAU DASHBOARD

Python Code in Jupyter Notebook

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1646717 entries, 0 to 1646716
Data columns (total 74 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   id                           1646717 non-null  int64
 1   member_id                    887379 non-null   float64
 2   loan_amnt                    1646717 non-null  float64
 3   funded_amnt                  1646717 non-null  float64
 4   funded_amnt_inv              1646717 non-null  float64
 5   term                         1646717 non-null  object
 6   int_rate                     1646717 non-null  float64
 7   installment                  1646717 non-null  float64
 8   grade                        1646717 non-null  object
 9   sub_grade                    1646717 non-null  object
 10  emp_title                    1544285 non-null  object
 11  emp_length                   1551529 non-null  object
 12  home_ownership               1646717 non-null  object
 13  annual_inc                   1646713 non-null  float64
 14  verification_status          1646717 non-null  object
 15  issue_d                      1646717 non-null  object
 16  loan_status                  1646717 non-null  object
 17  pymnt_plan                   1646717 non-null  object
 18  url                          887379 non-null   object
 19  desc                         126045 non-null   object
 20  purpose                      1646717 non-null  object
 21  title                        1623392 non-null  object
 22  zip_code                     1646716 non-null  object
 23  addr_state                   1646717 non-null  object
 24  dti                          1646362 non-null  float64
 25  delinq_2yrs                  1646688 non-null  float64
 26  earliest_cr_line             1646688 non-null  object
 27  inq_last_6mths               1646687 non-null  float64
 28  mths_since_last_delinq       829700 non-null   float64
 29  mths_since_last_record       278232 non-null   float64
 30  open_acc                     1646688 non-null  float64
 31  pub_rec                      1646688 non-null  float64
 32  revol_bal                    1646717 non-null  float64
 33  revol_util                   1645698 non-null  float64
 34  total_acc                    1646688 non-null  float64
 35  initial_list_status          1646717 non-null  object
 36  out_prncp                    1646717 non-null  float64
 37  out_prncp_inv                1646717 non-null  float64
 38  total_pymnt                  1646717 non-null  float64
 39  total_pymnt_inv              1646717 non-null  float64
 40  total_rec_prncp              1646717 non-null  float64
 41  total_rec_int                1646717 non-null  float64
 42  total_rec_late_fee           1646717 non-null  float64
 43  recoveries                   1646717 non-null  float64
 44  collection_recovery_fee      1646717 non-null  float64
 45  last_pymnt_d                 1628110 non-null  object
 46  last_pymnt_amnt              1646717 non-null  float64
 47  next_pymnt_d                 1225831 non-null  object
 48  last_credit_pull_d           1646646 non-null  object
 49  collections_12_mths_ex_med   1646572 non-null  float64
 50  mths_since_last_major_derog  436808 non-null   float64
 51  policy_code                  1646717 non-null  float64
 52  application_type             1646717 non-null  object
 53  annual_inc_joint             34514 non-null    float64
 54  dti_joint                    34510 non-null    float64
 55  verification_status_joint    34514 non-null    object
 56  acc_now_delinq               1646688 non-null  float64
 57  tot_coll_amt                 1576441 non-null  float64
 58  tot_cur_bal                  1576441 non-null  float64
 59  open_acc_6m                  780648 non-null   float64
 60  open_il_6m                   21372 non-null    float64
 61  open_il_12m                  780649 non-null   float64
 62  open_il_24m                  780649 non-null   float64
 63  mths_since_rcnt_il           759605 non-null   float64
 64  total_bal_il                 780649 non-null   float64
 65  il_util                      677360 non-null   float64
 66  open_rv_12m                  780649 non-null   float64
 67  open_rv_24m                  780649 non-null   float64
 68  max_bal_bc                   780649 non-null   float64
 69  all_util                     780596 non-null   float64
 70  total_rev_hi_lim             1576441 non-null  float64
 71  inq_fi                       780649 non-null   float64
 72  total_cu_tl                  780648 non-null   float64
 73  inq_last_12m                 780648 non-null   float64
dtypes: float64(50), int64(1), object(23)
memory usage: 929.7+ MB
```

Exploratory Data Analysis (EDA)

- Current loans
- NaN
- Non-relevant features
- Formatting
- Data leakage
- Bad/Good loan definition

Result:

- 1,646,717 → 370,702 rows
- 74 → 50 columns

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	total_bal_il	il_util	open_rv_12m	open_
0	1077501	1296599.0	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	B2	...	NaN	NaN	NaN	
1	1077430	1314167.0	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	C4	...	NaN	NaN	NaN	
2	1077175	1313524.0	2400.0	2400.0	2400.0	36 months	15.96	84.33	C	C5	...	NaN	NaN	NaN	
3	1076863	1277178.0	10000.0	10000.0	10000.0	36 months	13.49	339.31	C	C1	...	NaN	NaN	NaN	
4	1075358	1311748.0	3000.0	3000.0	3000.0	60 months	12.69	67.79	B	B5	...	NaN	NaN	NaN	

5 rows × 74 columns

Data Wrangling

Data pre-processing
- One Hot Encoding
- Label Encoding

Over-sampling
- SMOTE

Down-sampling
- Random under sampling

Correlation analysis
- Deletion of high correlated variables

Grade:
{A, B, C, D, E, F, G} → {0, 1, 2, 3, 4, 5, 6, 7}

Payment Plan:
{'y', 'n'} → {0,1}

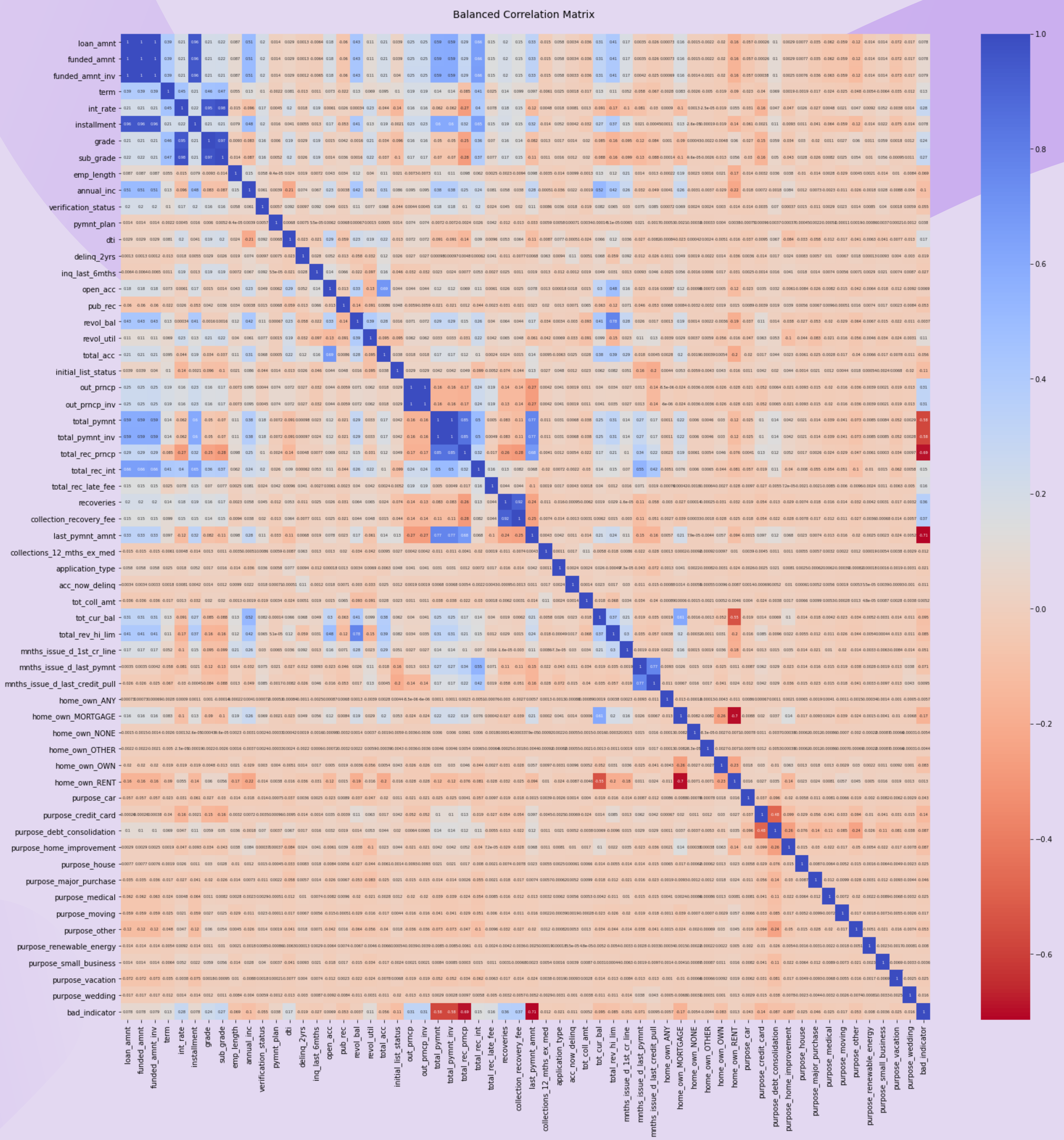
Before:
75% Good loans - 25% Bad loans

After:
52% Good loans - **48%** Bad loans

Data Wrangling

10 Most Correlated

loan_amnt	funded_amnt	1.000000
out_prncp	out_prncp_inv	1.000000
total_pymnt	total_pymnt_inv	0.999999
fndd_amnt	fndd_amnt_inv	0.999997
loan_amnt	fndd_amnt_inv	0.999997
grade	sub_grade	0.999029
int_rate	sub_grade	0.998434
int_grade	grade	0.997233
fndd_amnt_inv	installment	0.994888
fndd_amnt	installment	0.994856



Pandas profiling

- General dataset's overview
- Variable distribution
- Irrelevant variable (All values the same, 0, ...)

Overview

Overview | Alerts 40 | Reproduction

Dataset statistics

Number of variables	49
Number of observations	409072
Missing cells	0
Missing cells (%)	0.0%
Total size in memory	148.2 MiB
Average record size in memory	380.0 B

Variable types

Numeric	49

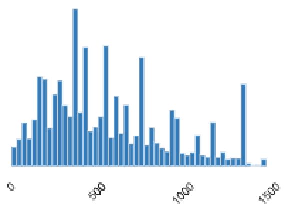
Variables

loan_amnt

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	1481
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	559.0586792

Minimum	0
Maximum	1480
Zeros	1258
Zeros (%)	0.3%
Negative	0
Negative (%)	0.0%
Memory size	3.1 MiB



Toggle details

Classification Modelling

Logistic Regression (x2)
Random Forest
AdaBoost

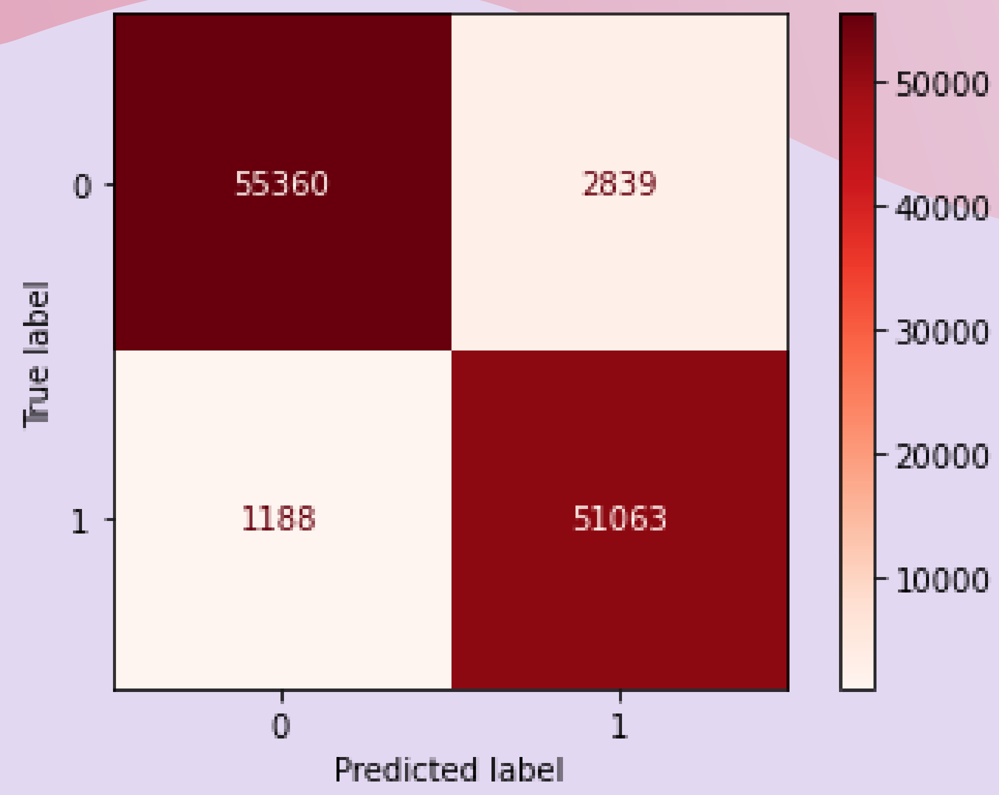
Classification Modelling

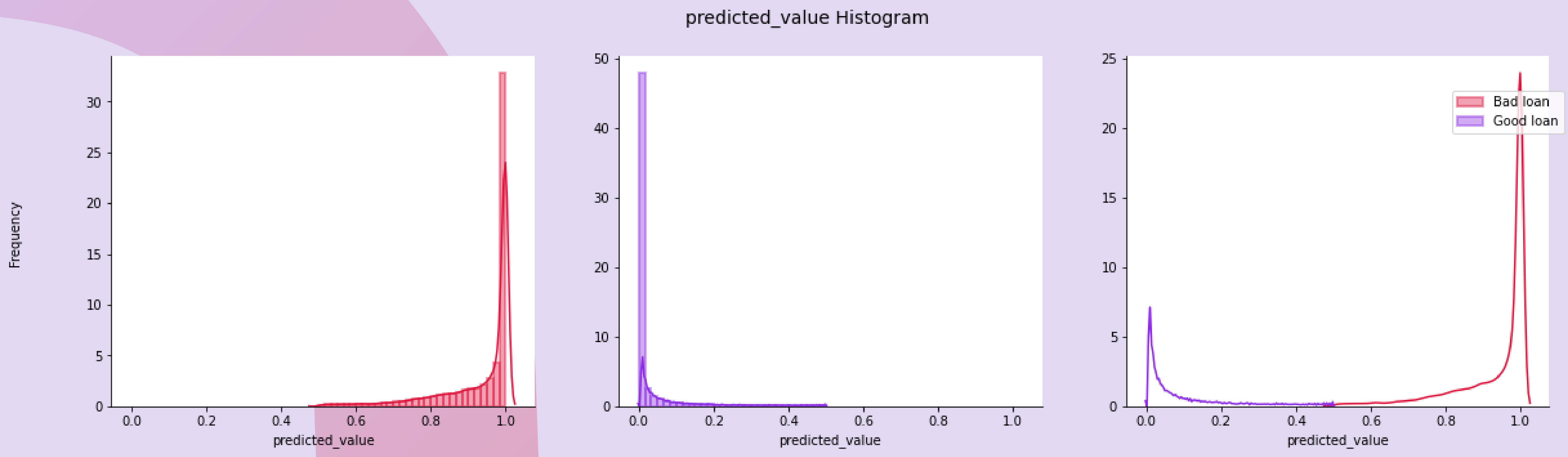
Logistic Regression 1

LogisticRegression(C=1.0, max_iter=100, solver='liblinear')

Classification report:

	precision	recall	f1-score	support
0	0.98	0.95	0.96	58199
1	0.95	0.98	0.96	52251
accuracy			0.96	110450
macro avg	0.96	0.96	0.96	110450
weighted avg	0.96	0.96	0.96	110450

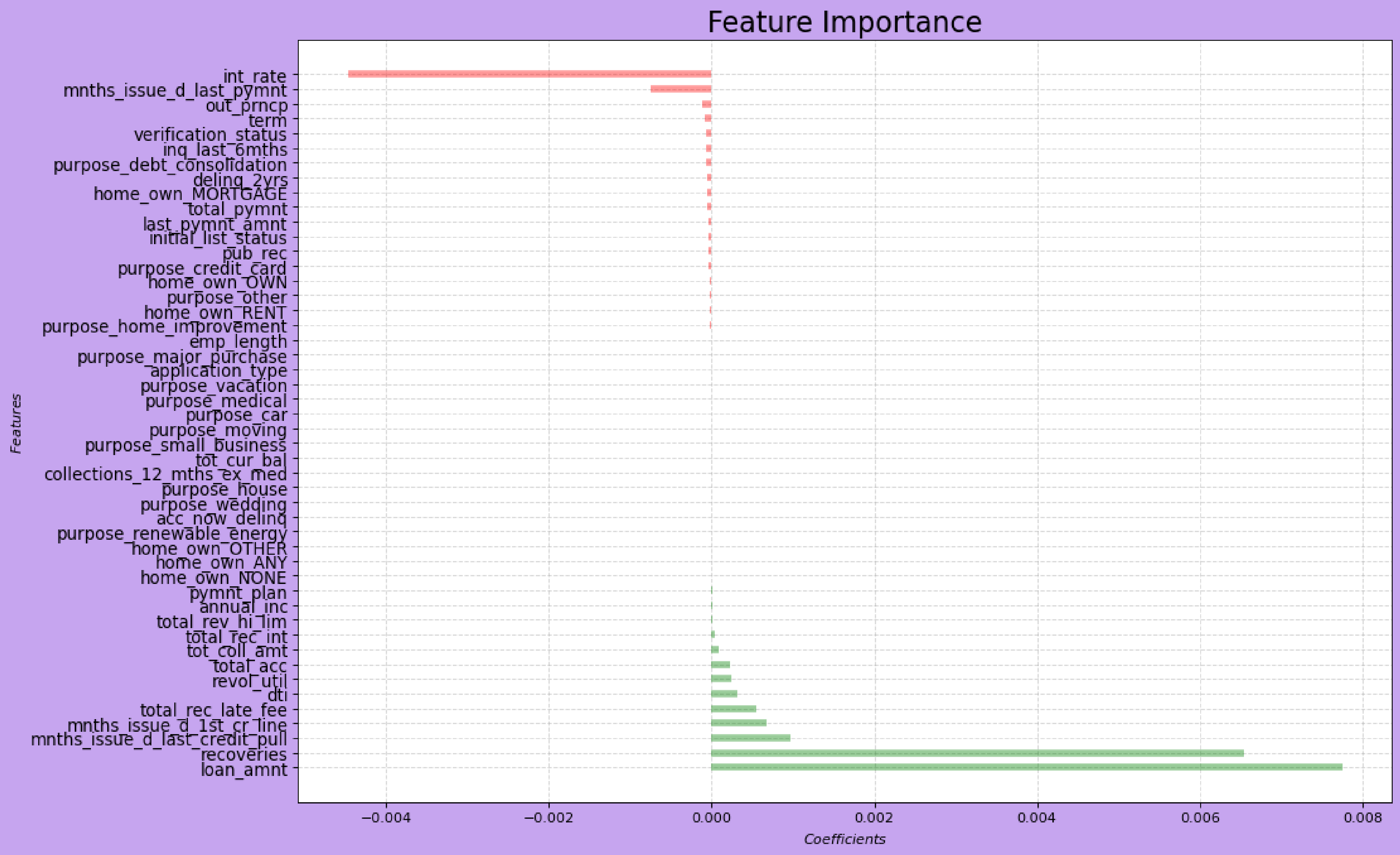




Classification Modelling

Logistic Regression 1

LogisticRegression(C=1.0, max_iter=100, solver='liblinear')



Top features:
- Loan amount
- Recoveries
- Months issue and credit pull
- Months issue and 1st credit
- Total recoveries late fee

Data Leakage:
- recoveries
- total_rec_late_fee
- total_rec_int
- total_pyment
- last_pymnt_d
- last_pymt_amnt
- mnths_issue_d_last_pymnt
- mnths_issue_d_last_credit_pull
- out_prncp

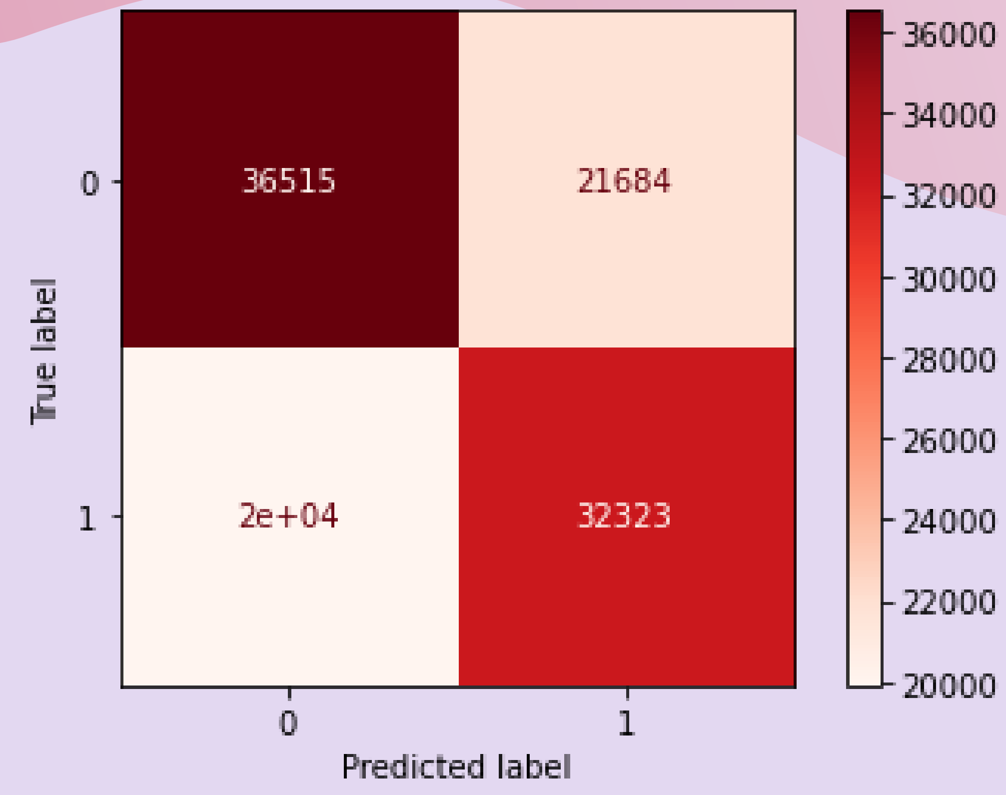
Classification Modelling

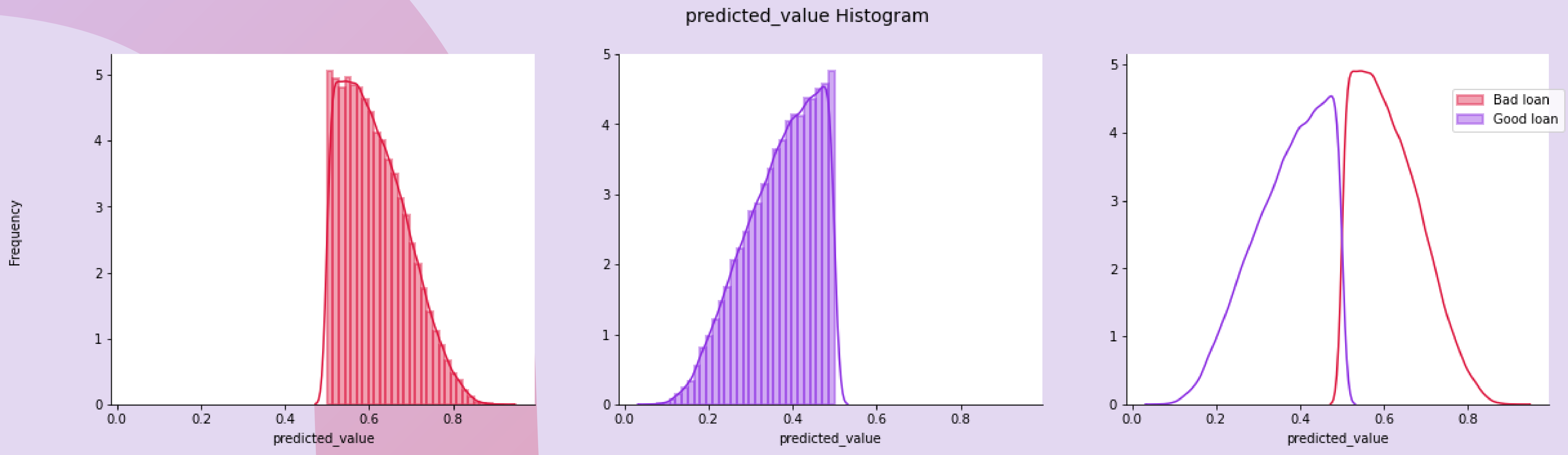
Logistic Regression 2

LogisticRegression(C=1.0, max_iter=100, solver='liblinear')

Classification report:

	precision	recall	f1-score	support
0	0.65	0.63	0.64	58199
1	0.60	0.62	0.61	52251
accuracy			0.62	110450
macro avg	0.62	0.62	0.62	110450
weighted avg	0.62	0.62	0.62	110450

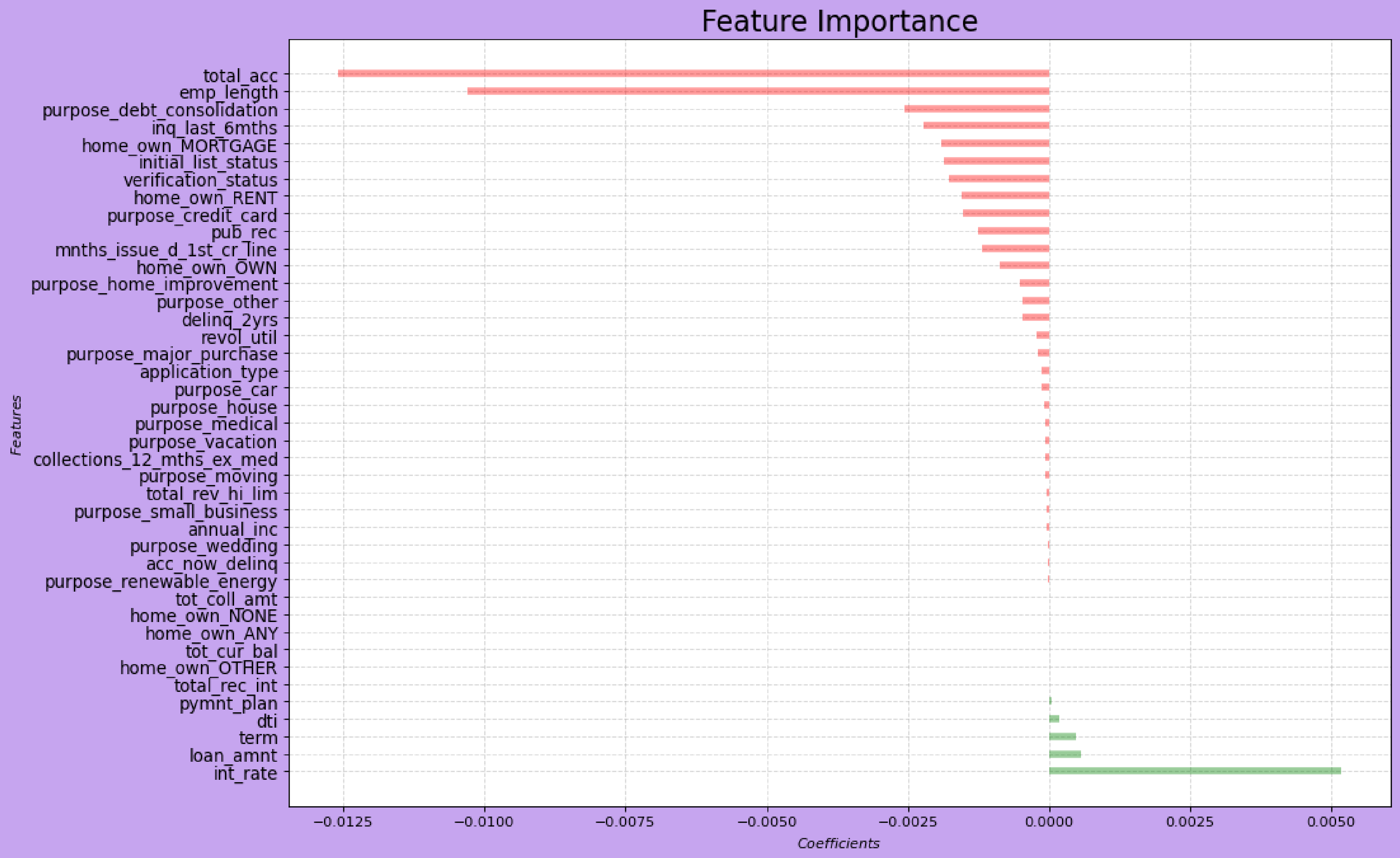




Classification Modelling

Logistic Regression 2

LogisticRegression(C=1.0, max_iter=100, solver='liblinear')



Top features:

- Interest rate
- Loan amount
- Term
- DTI

Classification Modelling

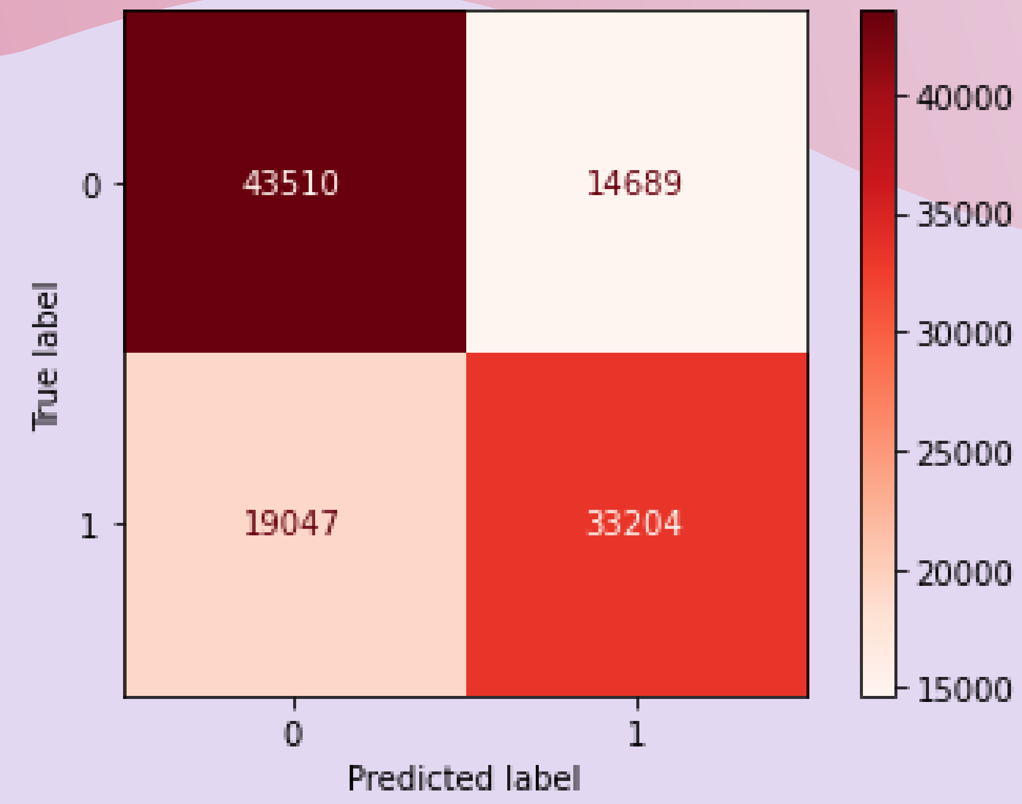
Random Forest

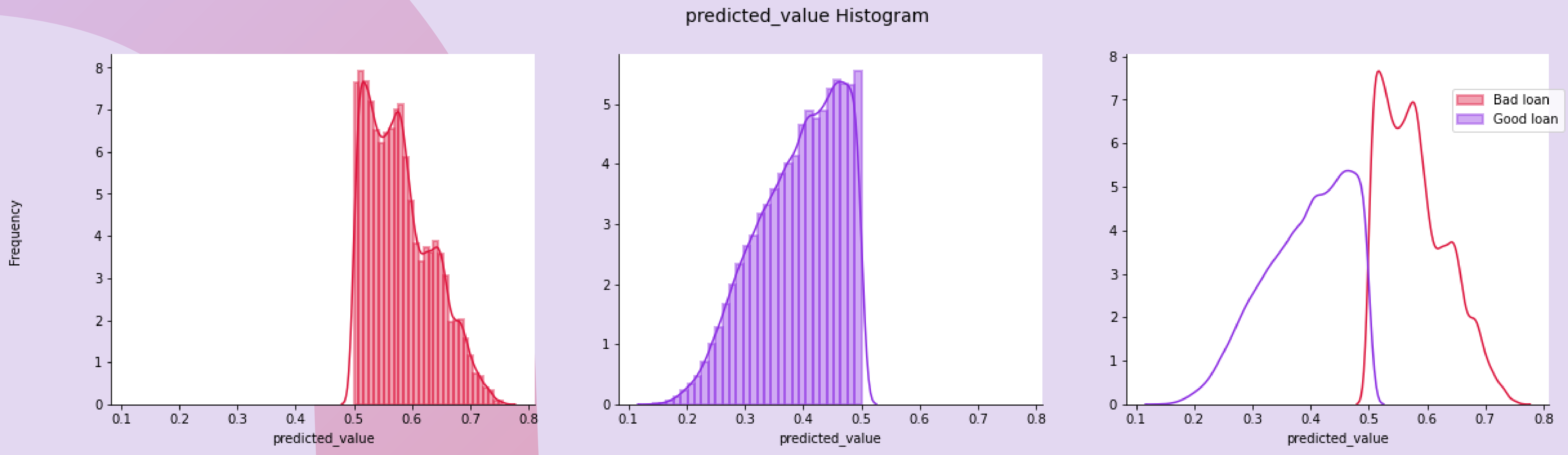
RandomForestClassifier(max_depth=5)

```
Classification report:
              precision    recall  f1-score   support

           0       0.70      0.75      0.72     58199
           1       0.69      0.64      0.66     52251

    accuracy                           0.69    110450
   macro avg       0.69      0.69      0.69    110450
weighted avg       0.69      0.69      0.69    110450
```

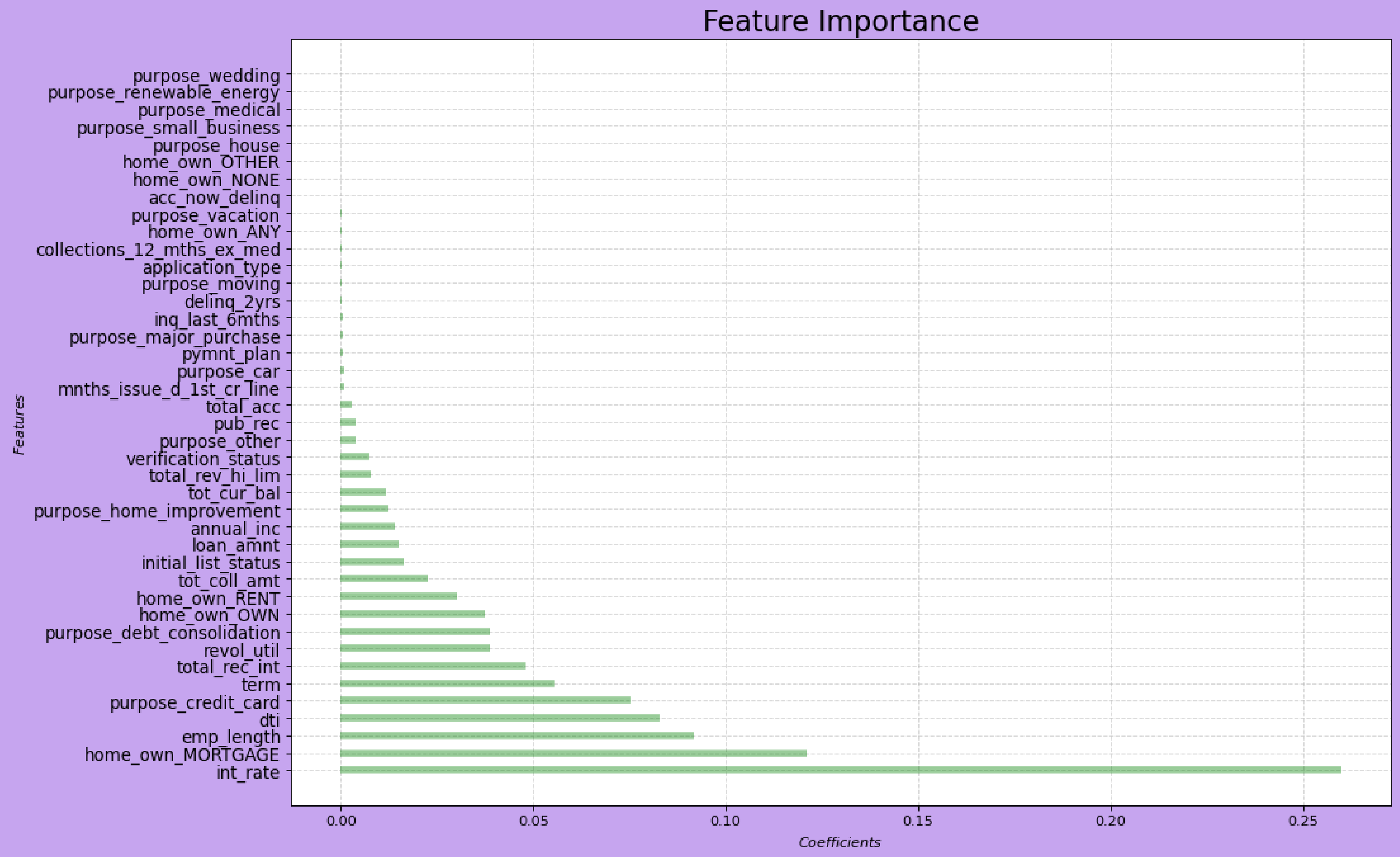




Classification Modelling

Random Forest

RandomForestClassifier(max_depth=5)



Top features:

- Interest rate
- Home ownership: Mortgage
- Employment length
- DTI
- Purpose: Credit card
- Term

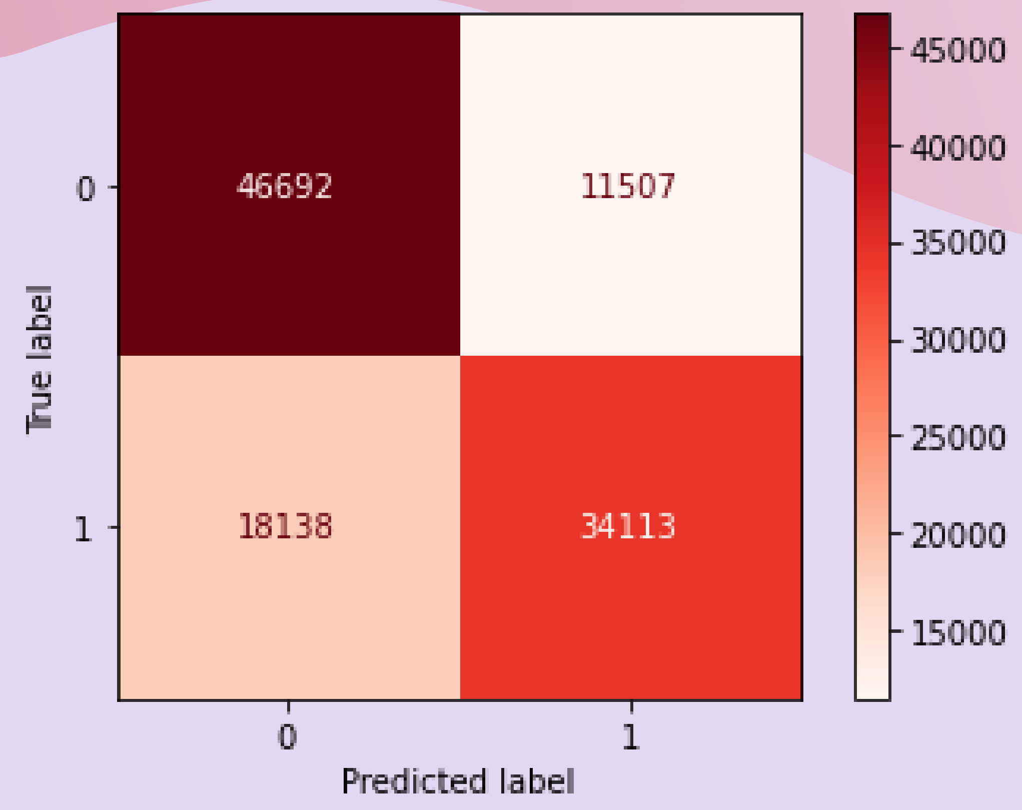
Classification Modelling

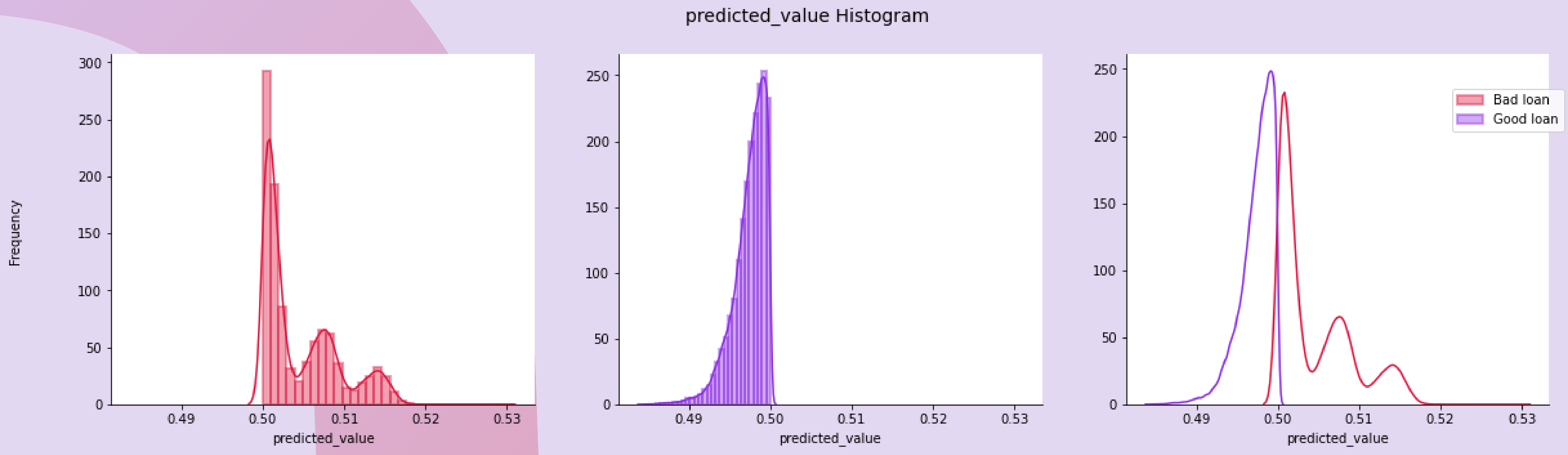
AdaBoost

AdaBoostClassifier(random_state=0, n_estimators=100)

Classification report:

	precision	recall	f1-score	support
0	0.72	0.80	0.76	58199
1	0.75	0.65	0.70	52251
accuracy			0.73	110450
macro avg	0.73	0.73	0.73	110450
weighted avg	0.73	0.73	0.73	110450

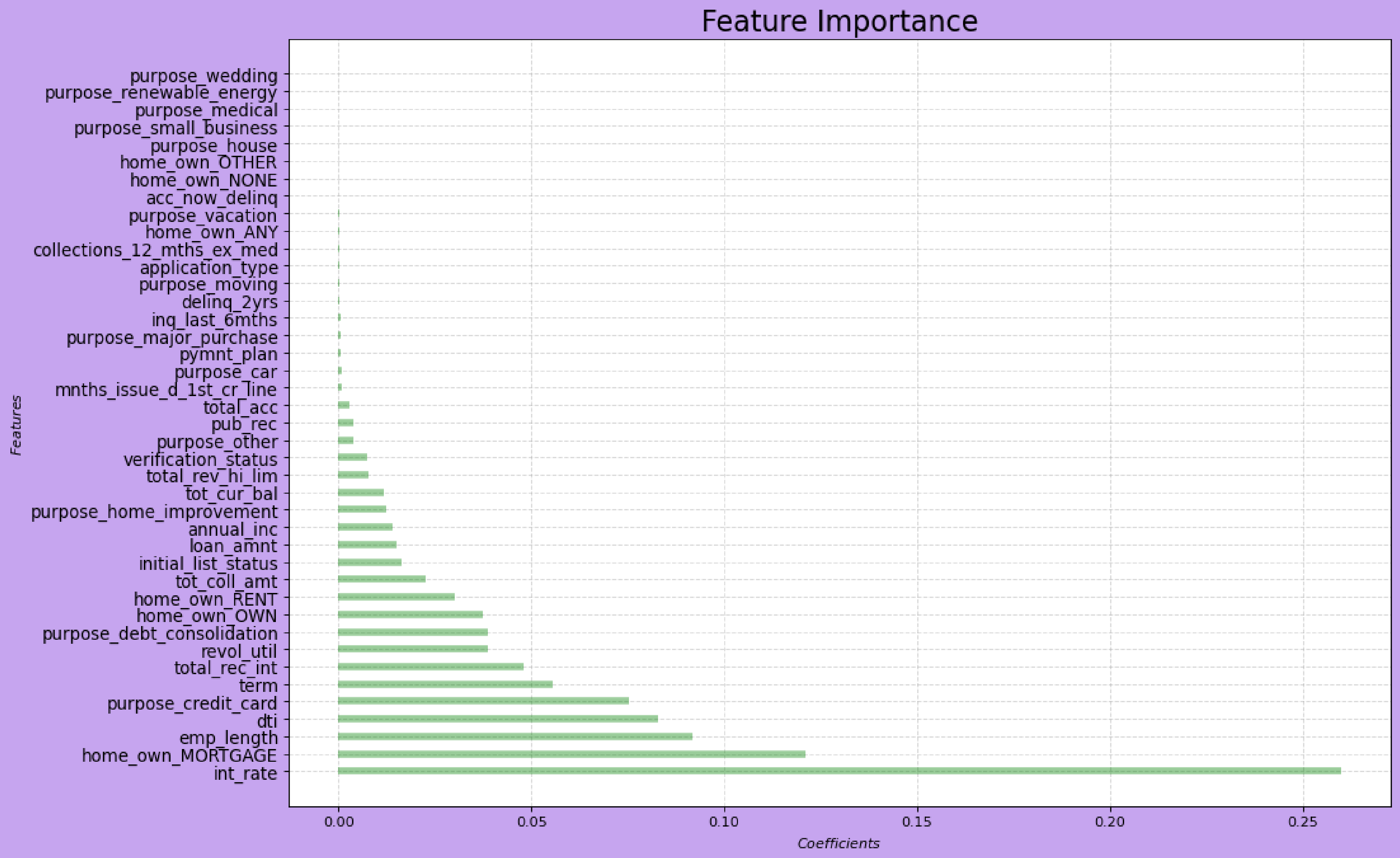




Classification Modelling

AdaBoost

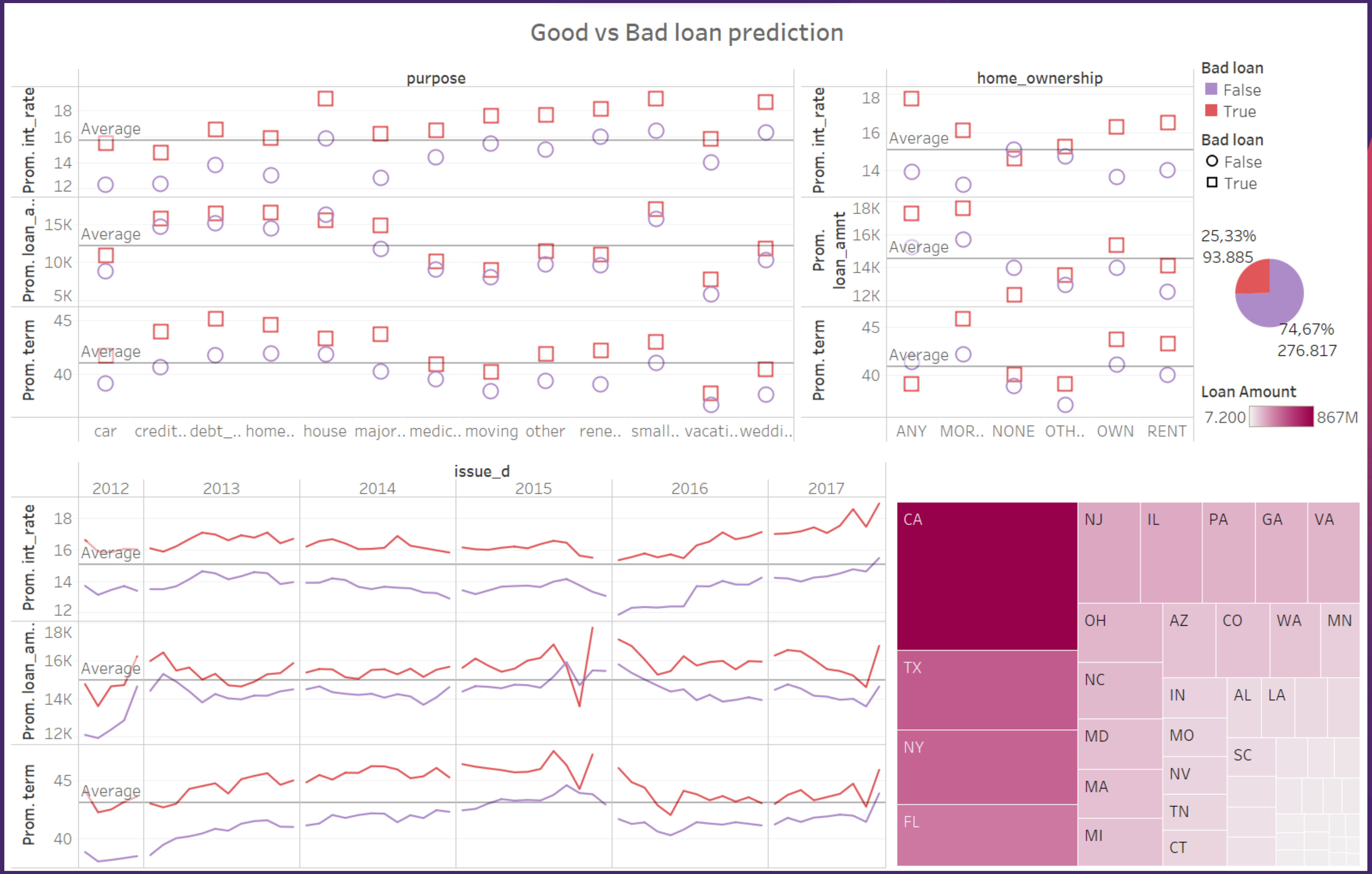
AdaBoostClassifier(random_state=0, n_estimators=100)



Top features:

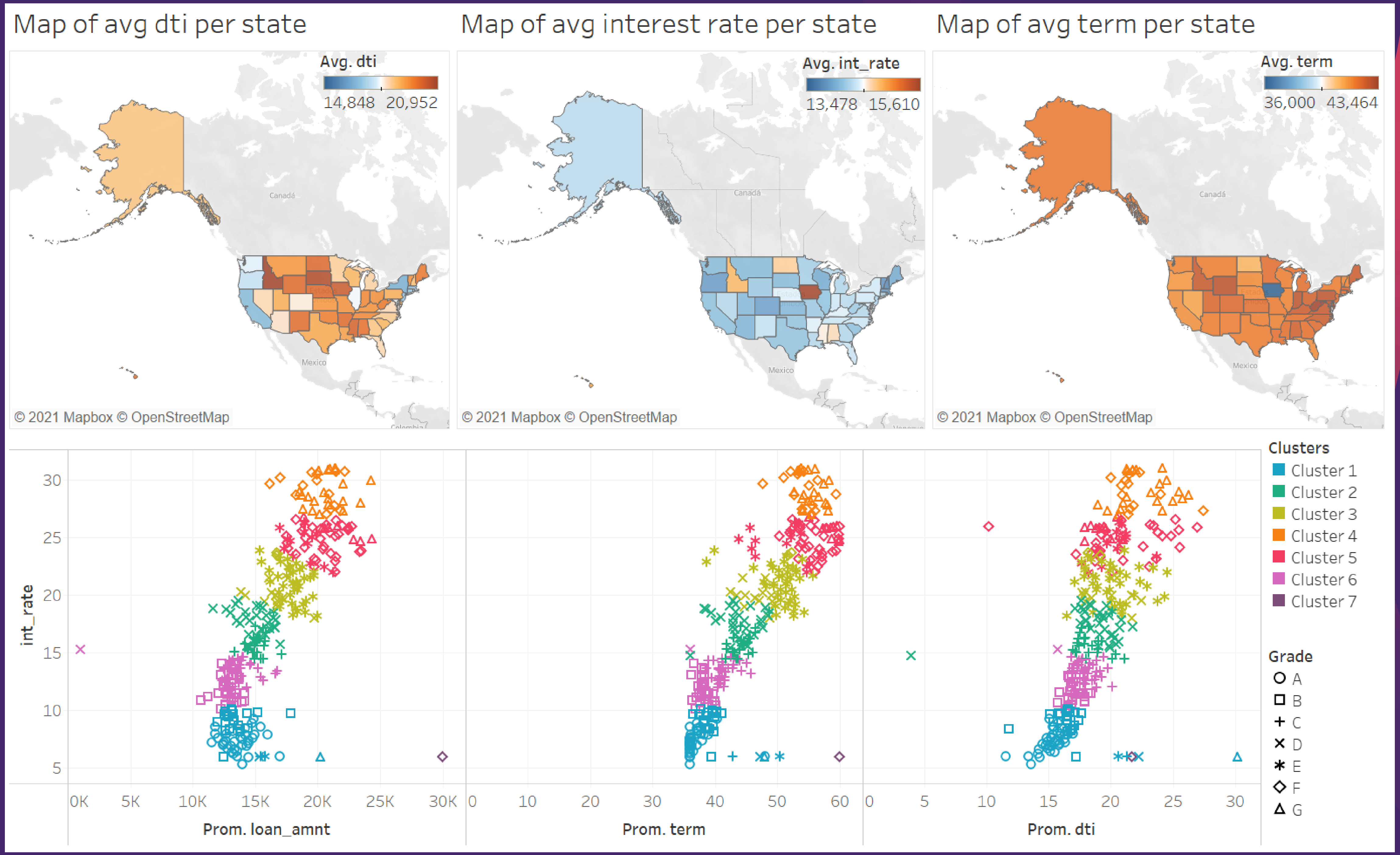
- Home ownership: Mortgage
- Home ownership: Rent
- Purpose: Debit consolidation
- Home ownership: Own
- Interest rate

Dashboard



Good vs Bad loan prediction
purpose
home_ownership
Prom. int_rate
Prom. loan_a..
Prom. term
Prom. loan_amnt
Average
car
credit..
debt_..
home..
house
major..
medic..
moving
other
rene..
small..
vacati..
weddi..
ANY
MOR..
NONE
OTH..
OWN
RENT
Bad loan
False
True
25,33%
93.885
74,67%
276.817
Loan Amount
7.200
867M
issue_d
2012
2013
2014
2015
2016
2017
Prom. loan_am..
CA
NJ
IL
PA
GA
VA
OH
AZ
CO
WA
MN
TX
NC
IN
AL
LA
NY
MD
MO
SC
NV
MA
TN
FL
MI
CT

Dashboard



Map of avg dti per state
Avg. dti
14,848 20,952
© 2021 Mapbox © OpenStreetMap
Map of avg interest rate per state
Avg. int_rate
13,478 15,610
© 2021 Mapbox © OpenStreetMap
Map of avg term per state
Avg. term
36,000 43,464
© 2021 Mapbox © OpenStreetMap
int_rate
5 10 15 20 25 30
0K 5K 10K 15K 20K 25K 30K
Prom. loan_amnt
0 10 20 30 40 50 60
Prom. term
0 5 10 15 20 25 30
Prom. dti
Clusters
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Grade
A
B
C
D
E
F
G

Results and Evaluation

Main features

- Interest Rate
- Loan Amount
- DTI
- Term
- Purpose
- Home ownership

Future work

- Cluster loans into ranges
- Parameter fine-tuning
- More/Different relevant features
- More/Different models

Bibliography

- https://www.kaggle.com/husainsb/lendingclub-issued-loans
- https://www.kaggle.com/ethon0426/lending-club-20072020q1

Thank you for your attention!

Any doubts?

VISUAL ANALYTICS FINAL PROJECT:
LOAN PREDICTION ON LENDINGCLUB ISSUED LOANS DATASET

Alex Marin Felices

20/12/2021