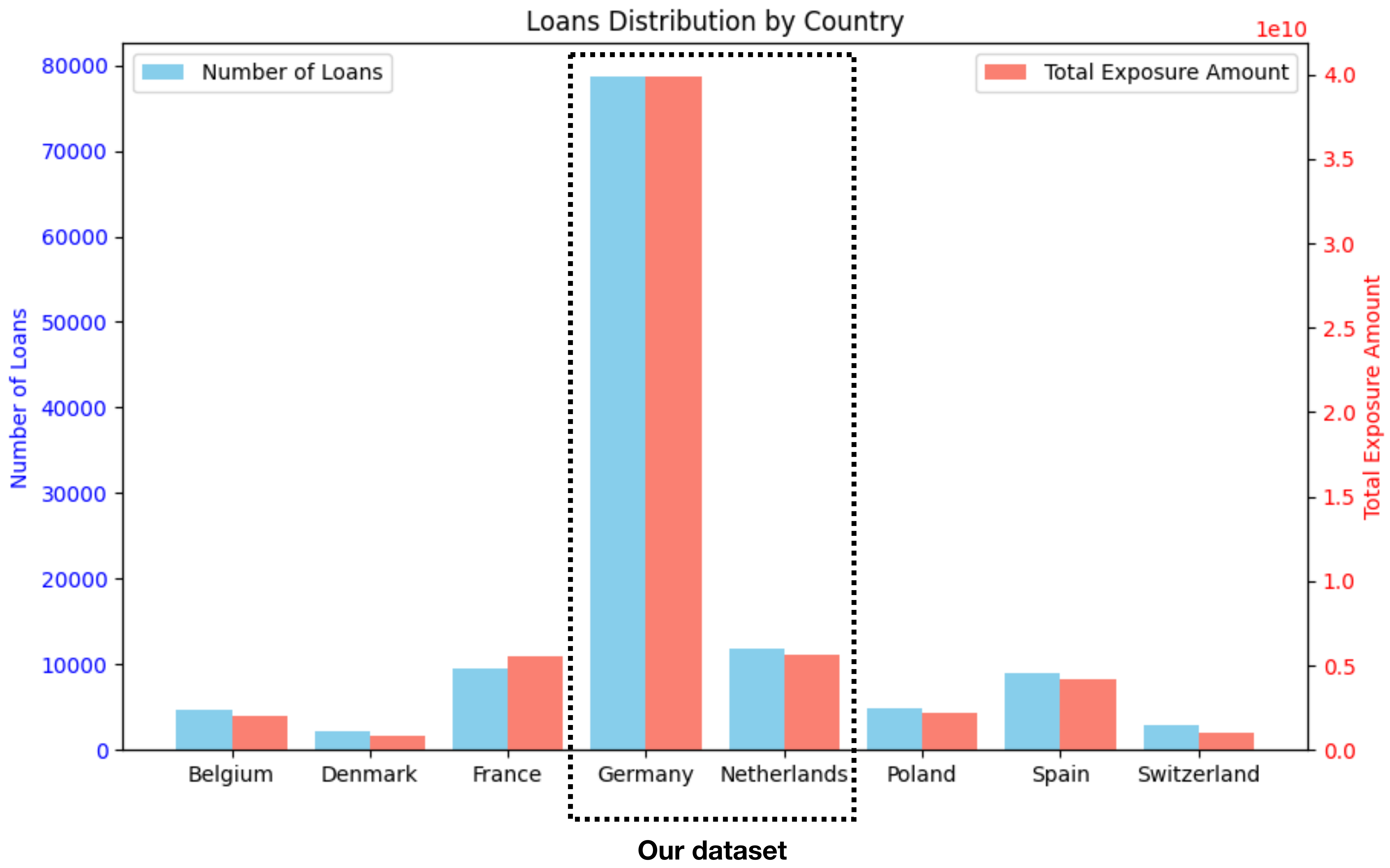
Predicting Default on Loan Data

Case Study

Konstantinos Kazanas

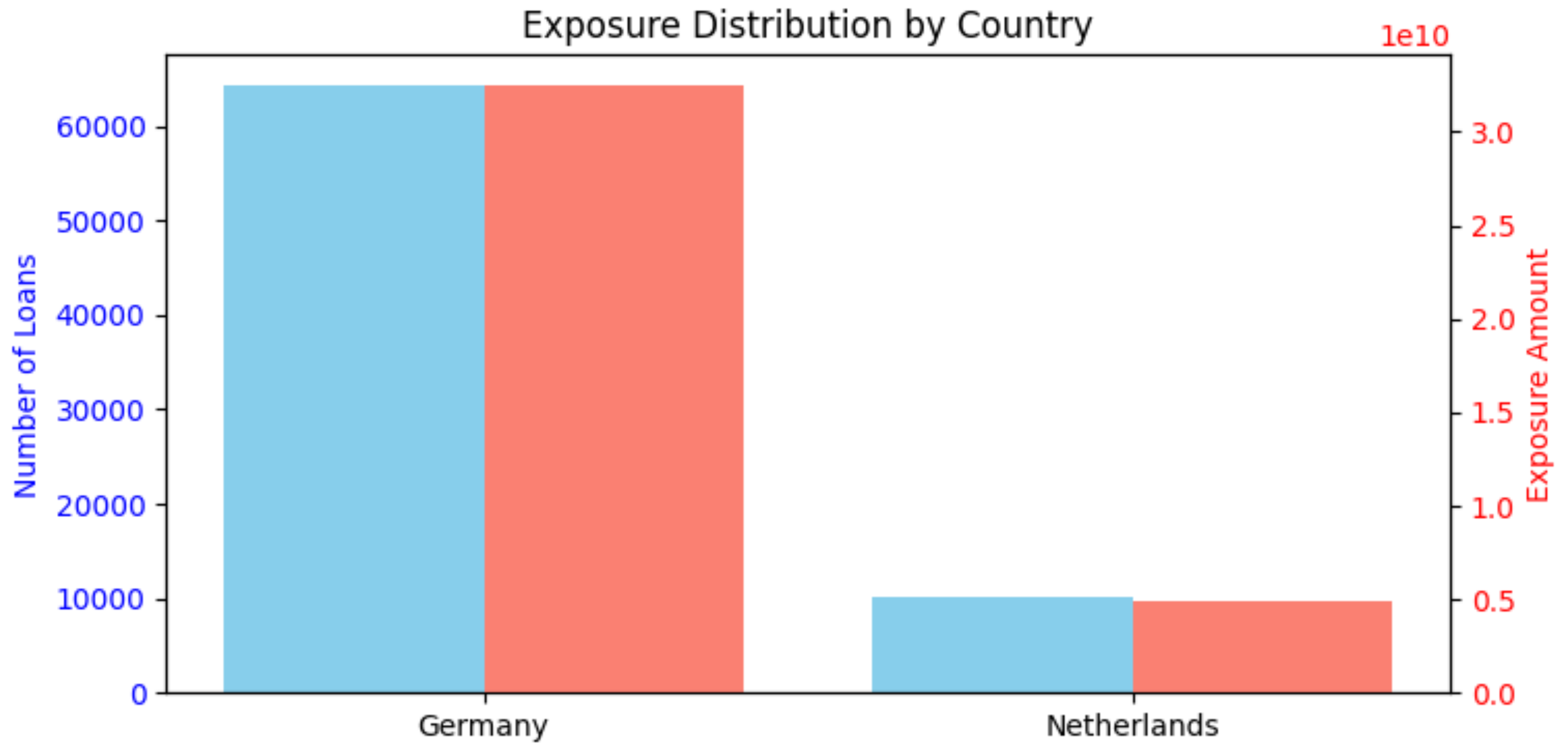
20th January 2026

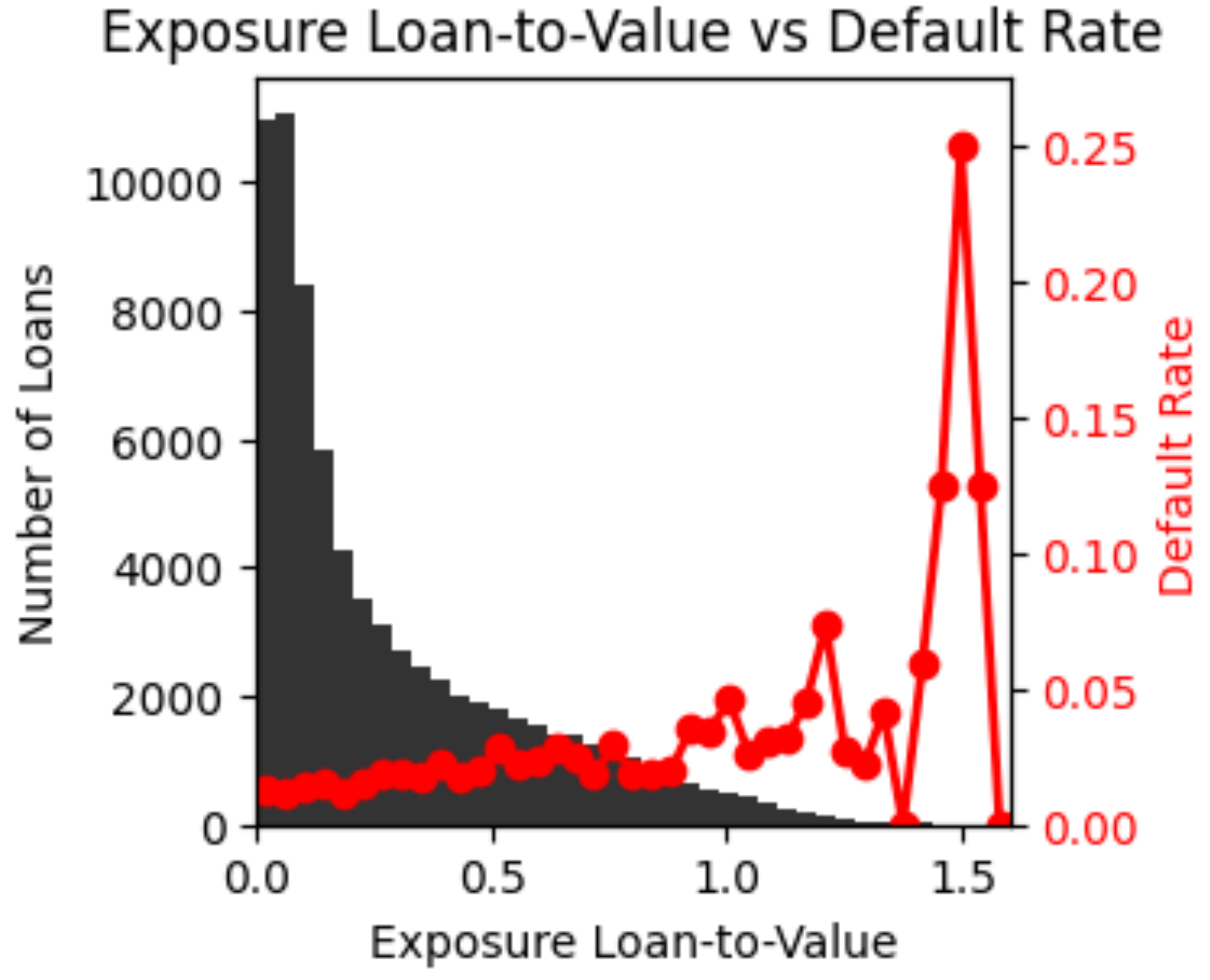
Dataset Overview

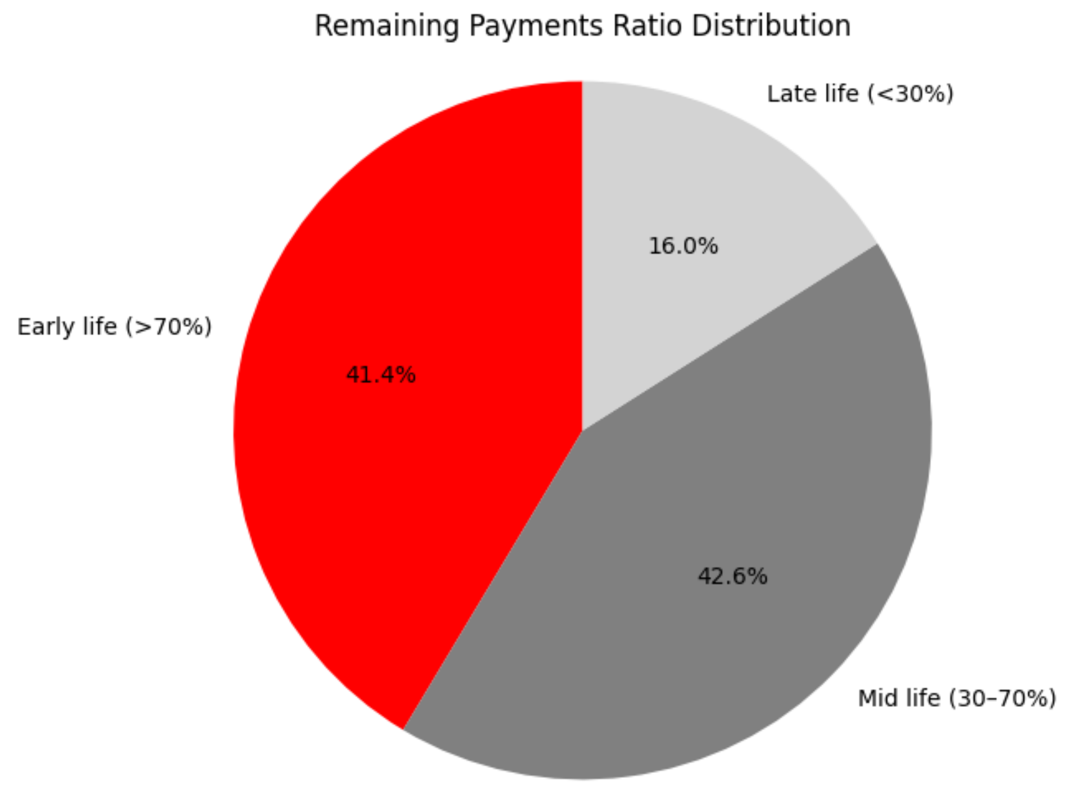


Dataset Overview

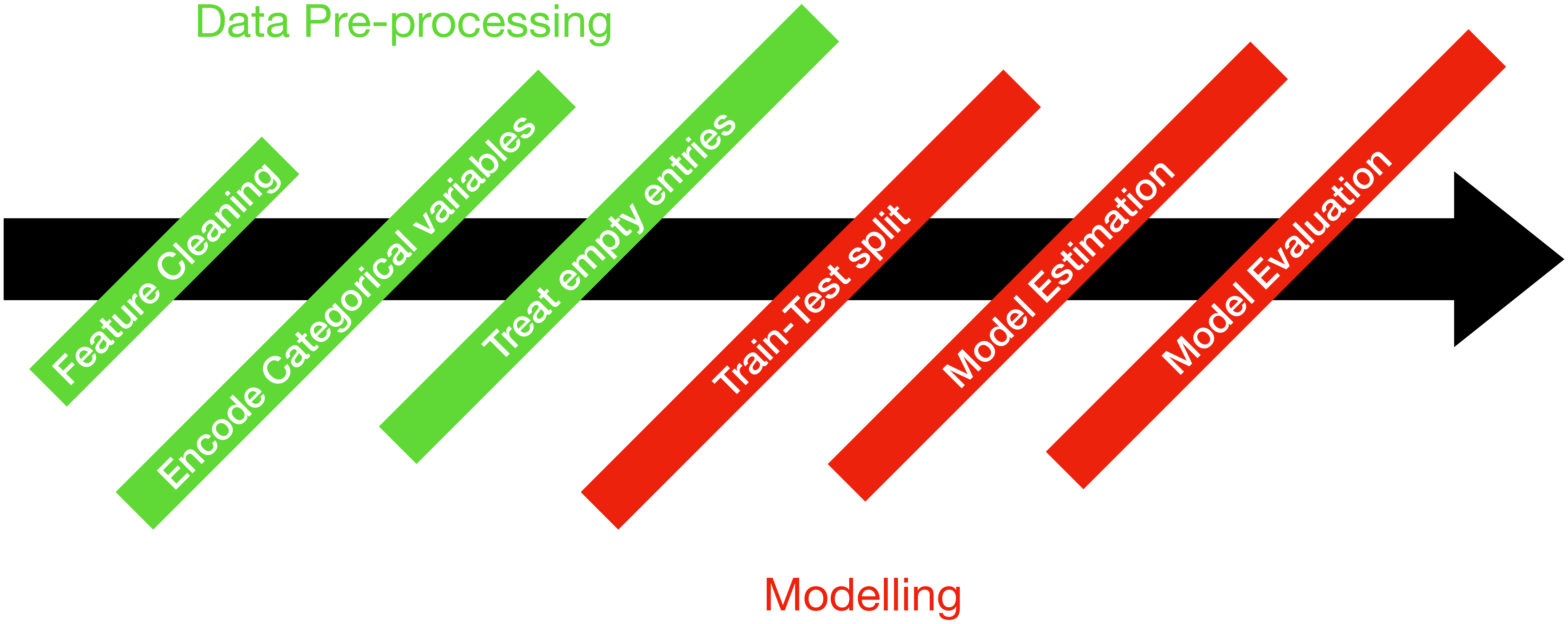
- The dataset contains general mortgage data that originated in Germany and the Netherlands with total exposures ~32.5bn and ~4.9bn respectively.
- Default Rate overall is 1.7% with 2,2% of the exposure on defaulted cases.
- High risk cases (LTV>80%) count ~5k with ~7bn exposure.
- Majority of cases (42.6%) is in mid-life stage, followed by the cases in early life stage (41.4%).







Modelling Approach



Model A

```
Optimization terminated successfully.
         Current function value: 0.061758
         Iterations 11
                           Logit Regression Results
==============================================================================
Dep. Variable:       num__DefaultFlag   No. Observations:                51988
Model:                          Logit   Df Residuals:                    51951
Method:                           MLE   Df Model:                           36
Date:                Sat, 24 Jan 2026   Pseudo R-squ.:                  0.2945
Time:                        17:04:31   Log-Likelihood:                -3210.7
converged:                       True   LL-Null:                       -4551.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

```
Log-Likelihood: -3210.656326143615
AIC: 6495.31265228723
BIC: 6823.087075750184
AUC: 0.8499031902943439
Accuracy Ratio (AR): 0.6998063805886878
```

Model B

```
Optimization terminated successfully.
         Current function value: 0.064888
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:       num__DefaultFlag   No. Observations:                51988
Model:                          Logit   Df Residuals:                    51979
Method:                           MLE   Df Model:                            8
Date:                Sat, 24 Jan 2026   Pseudo R-squ.:                  0.2588
Time:                        17:09:46   Log-Likelihood:                -3373.4
converged:                       True   LL-Null:                       -4551.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

```
Log-Likelihood: -3373.379190585881
AIC: 6764.758381171762
BIC: 6844.487294987076
AUC: 0.8413969023387677
Accuracy Ratio (AR): 0.6827938046775355
```

Model C

```
Optimization terminated successfully.
         Current function value: 0.063580
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:       num__DefaultFlag   No. Observations:                51988
Model:                          Logit   Df Residuals:                    51973
Method:                           MLE   Df Model:                           14
Date:                Sat, 24 Jan 2026   Pseudo R-squ.:                  0.2737
Time:                        17:31:46   Log-Likelihood:                -3305.4
converged:                       True   LL-Null:                       -4551.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

```
Log-Likelihood: -3305.4181105749612
AIC: 6640.8362211499225
BIC: 6773.717744175445
AUC: 0.8391915361013553
Accuracy Ratio (AR): 0.6783830722027107
```

Model D

```
Optimization terminated successfully.
         Current function value: 0.063581
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:       num__DefaultFlag   No. Observations:                51988
Model:                          Logit   Df Residuals:                    51974
Method:                           MLE   Df Model:                           13
Date:                Sat, 24 Jan 2026   Pseudo R-squ.:                  0.2737
Time:                        17:40:14   Log-Likelihood:                -3305.4
converged:                       True   LL-Null:                       -4551.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

```
Log-Likelihood: -3305.4429885950676
AIC: 6638.885977190135
BIC: 6762.908732013956
AUC: 0.8391688453525731
Accuracy Ratio (AR): 0.6783376907051462
```

Variable selection in models A-D is done by using L1 regularisation which penalises weak variables (loss=-LL+λΣβ_j). For c=1/λ=0.0009 we keep 13 variables (model D)

Model E

```
Optimization terminated successfully.
         Current function value: 0.063632
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:      num__DefaultFlag   No. Observations:                51988
Model:                         Logit   Df Residuals:                    51977
Method:                          MLE   Df Model:                           10
Date:               Sun, 25 Jan 2026   Pseudo R-squ.:                  0.2731
Time:                       15:26:15   Log-Likelihood:                -3308.1
converged:                      True   LL-Null:                       -4551.0
Covariance Type:           nonrobust   LLR p-value:                     0.000
=================================================================================================
                                      coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                              -4.6665      0.379    -12.321      0.000      -5.409      -3.924
num__PropertyValue              -7.868e-08   4.17e-08     -1.886      0.059    -1.6e-07    3.11e-09
num__NumberOfExposures              0.0923      0.031      2.973      0.003       0.031       0.153
num__ExposureAmount              2.092e-07   1.02e-07      2.053      0.040    9.48e-09    4.09e-07
num__TimeToMaturity                 0.0110      0.005      2.069      0.039       0.001       0.021
num__InterestRate                  -0.3678      0.124     -2.977      0.003      -0.610      -0.126
num__MonthsOnBook                   0.0011      0.001      1.670      0.095      -0.000       0.002
num__DelinquencyLast3Mon            0.3391      0.047      7.163      0.000       0.246       0.432
num__30PlusDelinquencyLast12Mon     0.1836      0.016     11.763      0.000       0.153       0.214
num__30_60DelinquencyLast12Mon     -0.2129      0.024     -8.765      0.000      -0.260      -0.165
num__DaysInDelinquency              0.0437      0.002     20.299      0.000       0.039       0.048
=================================================================================================
```

```
Log-Likelihood: -3308.1141151507054
AIC: 6638.228230301411
BIC: 6735.674680520127
AUC: 0.8411843324119893
Accuracy Ratio (AR): 0.6823686648239786
```

Model E occurs when removing one by one statistically insignificant variables (p-value) of model D:

num__PropertySize
num__60PlusDelinquencyLast12Mon
num__DelinquencyLast12Mon

Model results: performance

The model discriminates between good and bad borrowers relatively good, (better than random) with an **AUC** of 0.84 and an **AR** of 0.68.

Area Under Curve

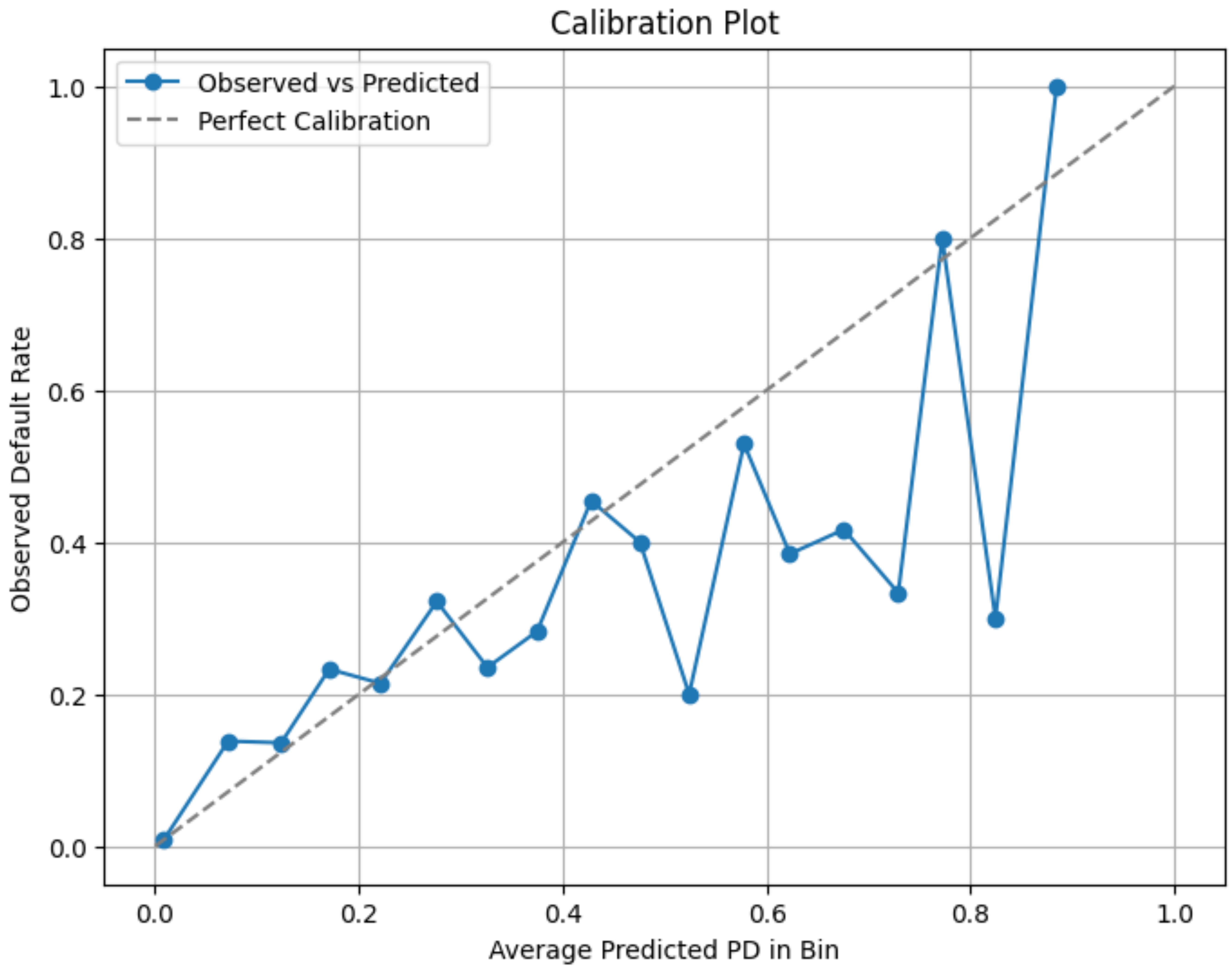
Accuracy Ratio

It behaves as expected for an unbalanced dataset with a confusion matrix as follows:

21868	47
332	34

Since **GINI** (=AR=2AUC-1) >0, the higher the model score the higher the positive outcomes (defaults).

Model results: calibration



risk_table

prob_bin_pc	Count	Defaults	Sum_PD	Avg PD	Default Rate	Share of Portfolio
(0,0000-0,0046]	3714	21	15,58593	0,00420	0,00565	0,05001
(0,0046-0,00502]	3713	14	17,89882	0,00482	0,00377	0,04999
(0,00502-0,00537]	3714	17	19,29553	0,00520	0,00458	0,05001
(0,00537-0,00568]	3713	18	20,52609	0,00553	0,00485	0,04999
(0,00568-0,00596]	3714	16	21,62759	0,00582	0,00431	0,05001
(0,00596-0,00619]	3713	16	22,54843	0,00607	0,00431	0,04999
(0,00619-0,00642]	3713	15	23,40174	0,00630	0,00404	0,04999
(0,00642-0,00661]	3714	13	24,20131	0,00652	0,00350	0,05001
(0,00661-0,0068]	3713	14	24,90959	0,00671	0,00377	0,04999
(0,0068-0,00701]	3714	14	25,63300	0,00690	0,00377	0,05001
(0,00701-0,00724]	3713	9	26,44126	0,00712	0,00242	0,04999
(0,00724-0,00748]	3713	30	27,32222	0,00736	0,00808	0,04999
(0,00748-0,00774]	3714	9	28,26509	0,00761	0,00242	0,05001
(0,00774-0,00805]	3713	26	29,30952	0,00789	0,00700	0,04999
(0,00805-0,00841]	3714	29	30,53448	0,00822	0,00781	0,05001
(0,00841-0,00895]	3713	24	32,12297	0,00865	0,00646	0,04999
(0,00895-0,0101]	3713	32	34,97635	0,00942	0,00862	0,04999
(0,0101-0,0158]	3714	72	45,64523	0,01229	0,01939	0,05001
(0,0158-0,0385]	3713	138	91,25413	0,02458	0,03717	0,04999
(0,0385-0,91]	3714	741	715,89848	0,19276	0,19952	0,05001

Recommendations

The current logistic regression model shows high accuracy for non-defaults but low recall for the default class (15%), which comes from class imbalance. To improve the model's performance, the following are recommended:

TP/AP

- **Balance dataset:** Using class_weight="balanced" or using SMOTE would give more importance to the minority class, helping the model detect defaults more effectively.
- **Increase Iterations:** Raising the max_iter value would ensure better convergence.
- **Feature Pre-processing:** Scaling numeric features, properly imputing missing values, and WOE transformed variables would provide a more consistent input for the model.

Appendix

Questions

A. How big is the complete dataset (rows and columns)?

- *123681 rows and 42 columns*

B. How many columns contain no data or NULL values?

- *6 columns*

C. What is the exposure amount of general mortgages linked to properties that have size greater than 300?

- *1.516.426.867*

D. How many customers have exactly three exposures and what is the total exposure amount of such clients?

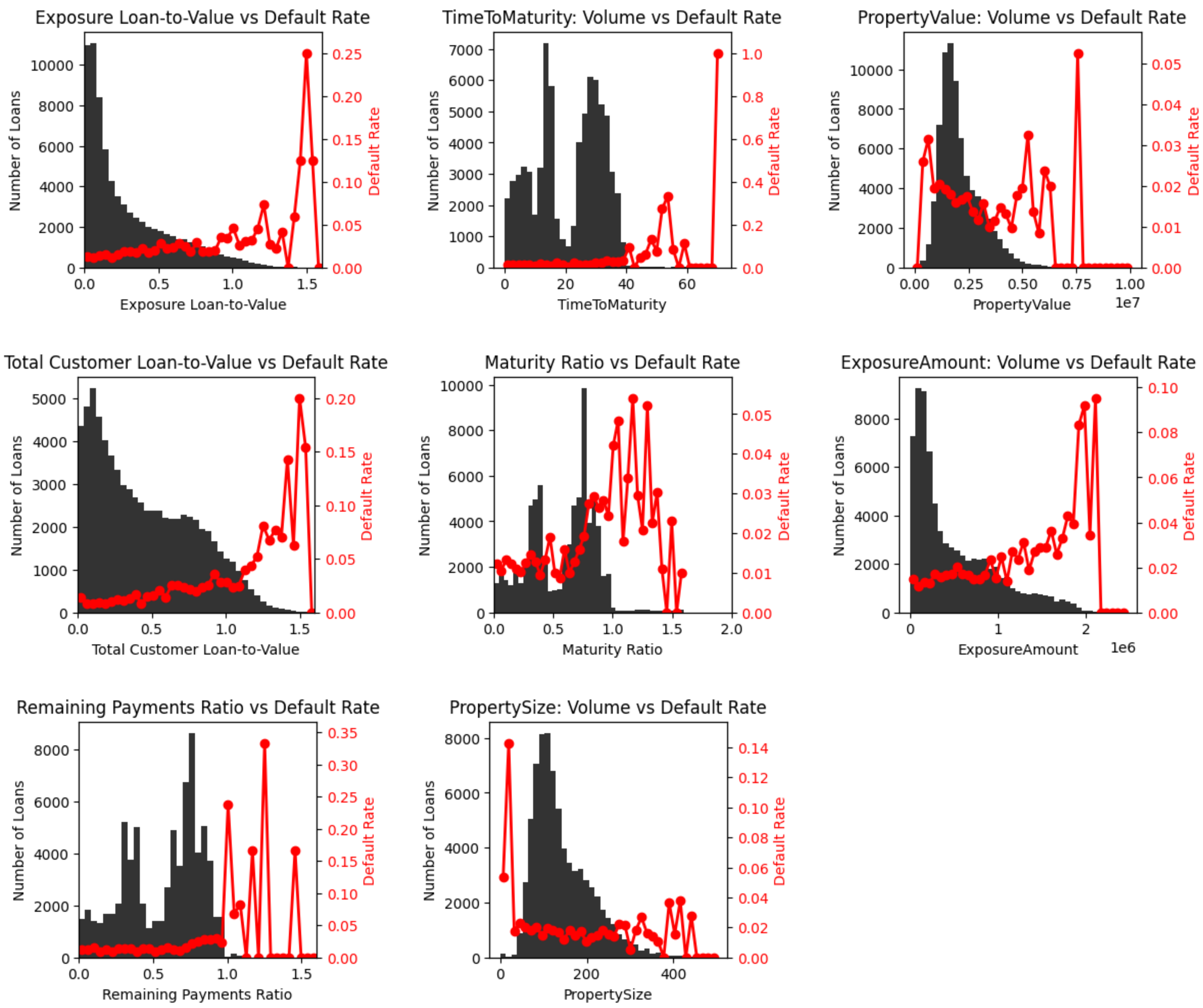
- \# *client 681991 with exposure amount 697.291*
- \# *client 736964 with exposure amount 2.327.310*

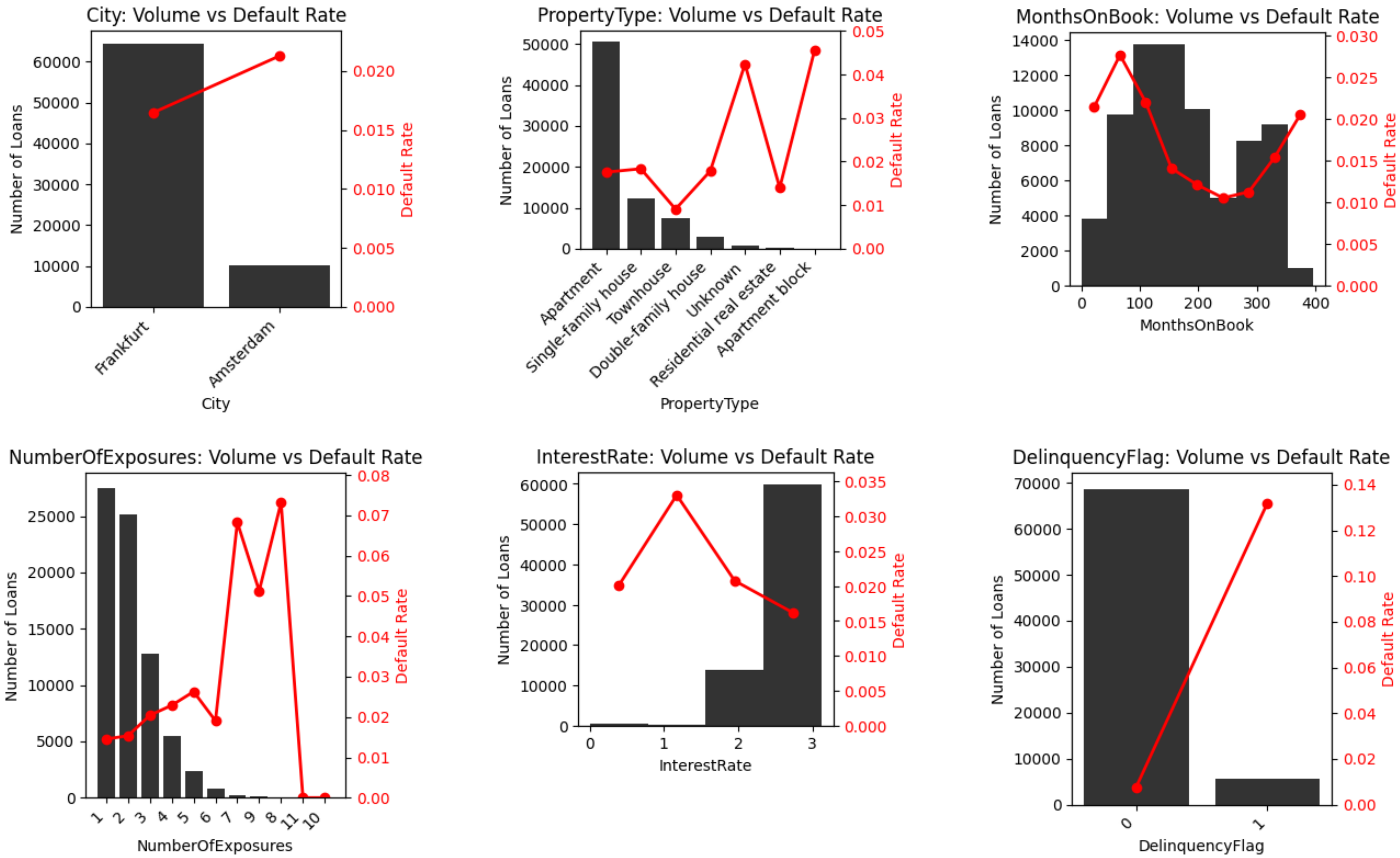
List of available variables

```
<class 'pandas.core.frame.DataFrame'>
Index: 74354 entries, 2 to 123680
Data columns (total 32 columns):
 #   Column                        Non-Null Count  Dtype
 0   DefaultFlag                   74354 non-null  int64
 1   PropertyType                  74265 non-null  object
 2   PropertyValue                 74354 non-null  int64
 3   PropertySize                  74354 non-null  float64
 4   ExposureLoanToValue           74354 non-null  float64
 5   TotalCustomerLoanToValue      74354 non-null  float64
 6   CountryOfOrigination          74354 non-null  object
 7   City                          74354 non-null  object
 8   NumberOfExposures             74354 non-null  int64
 9   ProductName                   74354 non-null  object
 10  ExposureAmount                74354 non-null  int64
 11  RemainingPaymentsRatio        74354 non-null  float64
 12  TimeToMaturity                74354 non-null  float64
 13  MaturityRatio                 74354 non-null  float64
 14  InterestRate                  74354 non-null  float64
 15  MonthsOnBook                  74354 non-null  int64
 16  ExposureDefaultFlagCount      74354 non-null  int64
 17  ClientDefaultFlagCount        74354 non-null  int64
 18  DelinquencyFlag               74354 non-null  int64
 19  DelinquencyLast3Mon           74269 non-null  float64
 20  DelinquencyLast12Mon          74270 non-null  float64
 21  30PlusDelinquencyLast3Mon     74269 non-null  float64
 22  30PlusDelinquencyLast12Mon    74270 non-null  float64
 23  60PlusDelinquencyLast3Mon     74269 non-null  float64
 24  60PlusDelinquencyLast12Mon    74270 non-null  float64
 25  0_30DelinquencyLast3Mon       74269 non-null  float64
 26  0_30DelinquencyLast12Mon      74270 non-null  float64
 27  30_60DelinquencyLast3Mon      74269 non-null  float64
 28  30_60DelinquencyLast12Mon     74270 non-null  float64
 29  60_90DelinquencyLast3Mon      74269 non-null  float64
 30  60_90DelinquencyLast12Mon     74270 non-null  float64
 31  DaysInDelinquency             74354 non-null  int64
```

Portion of assets remaining after debt

Missed payments


Exposure Loan-to-Value vs Default Rate
Number of Loans
Default Rate
Exposure Loan-to-Value
TimeToMaturity: Volume vs Default Rate
Number of Loans
Default Rate
TimeToMaturity
PropertyValue: Volume vs Default Rate
Number of Loans
Default Rate
PropertyValue
1e7
Total Customer Loan-to-Value vs Default Rate
Number of Loans
Default Rate
Total Customer Loan-to-Value
Maturity Ratio vs Default Rate
Number of Loans
Default Rate
Maturity Ratio
ExposureAmount: Volume vs Default Rate
Number of Loans
Default Rate
ExposureAmount
1e6
Remaining Payments Ratio vs Default Rate
Number of Loans
Default Rate
Remaining Payments Ratio
PropertySize: Volume vs Default Rate
Number of Loans
Default Rate
PropertySize


City: Volume vs Default Rate
Number of Loans
Default Rate
Frankfurt
Amsterdam
City
PropertyType: Volume vs Default Rate
Apartment
Single-family house
Townhouse
Double-family house
Unknown
Residential real estate
Apartment block
PropertyType
MonthsOnBook: Volume vs Default Rate
MonthsOnBook
NumberOfExposures: Volume vs Default Rate
NumberOfExposures
InterestRate: Volume vs Default Rate
InterestRate
DelinquencyFlag: Volume vs Default Rate
DelinquencyFlag