CREDIT RISK ANALYSIS:

Leveraging Advanced Analytics to Enhance Loan Default Detection

WORD COUNT: 3300
STUDENT ID: 251256517 | 251308573 | 251293304
COURSE CODE: FM9528 – BANKING ANALYTICS

TABLE OF CONTENTS

EXECUTIVE SUMMARY **2**

INTRODUCTION **2**

DATA PREPARATION **3**

DATA CLEANING AND PREPROCESSING 3
Missing Values Analysis (MVA) *3*
Outlier Treatment *3*
FEATURE ENGINEERING 3
Annual Income Joint *4*
Installment Account *4*
Collateral Coverage Ratio *4*
Bank Delinquency *4*
DATASET DIVISION & VARIABLE SELECTION 4
Splitting Data into Training and Testing Sets *4*
Weight of Evidence (WoE) Analysis and Variable Selection *5*
Correlation Analysis *5*

METHODS AND MODEL DEVELOPMENT **6**

METHODS 6
MODEL PERFORMANCE AND COMPARISON 6
VARIABLE IMPORTANCE 6

PERFORMANCE VARIABLES APPLICATION **7**

LOSS GIVEN DEFAULT (LGD) CALCULATION 7
LOAN PROFIT CALCULATION 7
OPTIMAL CUT-OFF POINT 7

PD CALIBRATION **7**

CONCLUSION AND DISCUSSION **7**

APPENDIX A: TABLES **9**

TABLE 1: LOGISTIC REGRESSION COEFFICIENTS 9
TABLE 2: MISSING VALUE IMPUTED WITH -1 10
TABLE 3: MISSING VALUE IMPUTED WITH MEDIAN 10
TABLE 4: DELETED VARIABLES 11
TABLE 5: VARIABLE CUT-OFF VALUES 11
TABLE 6: WEIGHT OF EVIDENCE VALUES 13
TABLE 7: MODEL METRICS 14
TABLE 8: CUT-OFF VALUES (OPTIMAL VALUE) 14
TABLE 9: VARIABLE DESCRIPTION 16
TABLE 10: PD CALIBRATION 20
TABLE 11: CUT-OFF DESCRIPTION 20

APPENDIX B: VISUALIZATION **22**

APPENDIX C: FORMULAS **29**

Executive Summary

This report presents a comprehensive analysis of a dataset containing 1.3 million loans issued between 2007 and 2018. The primary objective was to develop an advanced Internal Ratings-Based (IRB) model for origination scoring in credit risk analysis. We aim to improve the loan approval process, optimize capital requirements, and enhance the profitability of lending institutions.

Our analysis began with an exploratory data analysis that addressed missing values, outliers, and conducted feature engineering to extract additional insights. We carefully considered the potential of each variable for inclusion in our predictive models and performed Weight of Evidence (WoE) analysis and correlation analysis to identify the most relevant variables. This process led us to a subset of 17 variables from an initial set of 128 total variables to ensure our models focus on significant predictors for loan default.

In this work, we developed logistic regression and a XGBoost model for probability of default (defaulters or non-defaulters) predictions. The logistic regression model was optimized using elastic net regularization, while the XGBoost model was tuned using grid search. The most important factors identified by logistic regression model were term length, debt-to-income ratio, and total high credit limit. In contract, XGBoost ranked installment, funded amount invested and term as the top variables. We can see the differences between the two models, but the variable 'term' was highlighted by both as it proved to have significance in predicting probability of default.

Results showed that XGBoost is the optimal model in loan default prediction with an area under the curve (AUC) of 0.74 and as well as both precision and recall of 0.81 and 0.97 respectively. We calculated Loss Given Default (LGD) for loan defaults and estimated the loan profit of loan profit for non-defaulters of $2,235,250,118.2 by accounting for both the interest revenue generated from the loan and the associated costs. The optimal cut-off value was 0.475 for probability of default which maximized the profit of $58,870,390. The PD calibration allowed us to understand that CPI was the most significant macroeconomic variable.

Introduction

Peer-to-peer (P2P) lending has revolutionized the financial industry, offering an innovative alternative to traditional banking systems. Lending Club, a prominent P2P lending platform in North America, exemplifies this transformation by seamlessly connecting borrowers and investors. As a data-centric platform, Lending Club maintains a wealth of historical loan data, creating a unique opportunity to delve into the realm of credit risk analytics.

In this study, we embark on an exciting journey to develop an advanced Internal Ratings-Based (IRB) model for origination scoring, utilizing Lending Club's extensive dataset. Comprising information on around 1.3 million loans and 128 variables, this dataset provides fertile ground for exploration and analysis. Our primary objective is to estimate the capital requirements for Lending Club, offering insights into the potential implications of regulatory compliance, should it be introduced.

Data Preparation

DATA CLEANING AND PREPROCESSING

To create the most accurate prediction model based on a dataset of 1.3 million loans and 128 variables, we needed to carefully examine the complex dataset. This data includes loans that are either current or defaulted and contains important rows of information that are vital to our analysis.

During our in-depth exploration of the dataset, we discovered the presence of several duplicate variables that had minor differences in their names. We compared descriptions to identify and remove these redundant variables from our analysis. This comprehensive assessment of the dataset and its metadata was crucial in laying a solid foundation for the subsequent data preparation steps. By making sure that our analysis was based on an accurate understanding of the available information, we could confidently derive valuable and insightful results in our credit risk analysis. (Table 9)

Missing Values Analysis (MVA)

During the data preparation phase, one of the critical steps we undertook was addressing missing values in our dataset. Our primary objective was to ensure that our analysis would yield reliable results, which required a robust strategy for handling missing data. To achieve this, we began by evaluating the proportion of missing values for each variable in the dataset.

This preliminary assessment allowed us to classify the variables into two categories: those requiring deletion and those warranting further investigation. However, we acknowledged that a high percentage of missing values for some variables might not necessarily indicate the absence of recorded data, but rather an alternative interpretation of the data. Therefore, before finalizing the deletion of any variable, we carefully considered its potential for feature engineering, such as combining it with other variables or transforming it into another type. (Table 2 and 3)

Outlier Treatment

Outlier analysis was a critical step in refining our dataset to ensure its suitability for credit risk analysis. To achieve this, we used kernel density estimation (KDE) plots and boxplots as visualization tools to explore the presence of outliers in the data effectively.

We determined appropriate cut-off values that would enable us to maintain the essence of our data while reducing the effects of extreme outliers. To preserve as much information as possible, we chose not to remove the rows containing outliers but instead to cap the values beyond the established cut-off points. This decision allowed us to retain the original structure of our dataset while minimizing the influence of extreme values on our subsequent analysis. (Table 5)

Feature Engineering

We discuss the importance of feature engineering to transform existing variables and create new ones to enhance our analysis. By transforming and creating new variables, we can extract more information from the available data and better understand the relationships between the variables and the target outcome. Certain variables that fell into the deletion category earlier in section MVA required more investigation as their information involving missing values was miss leading. The

variables we created were Annual Income Joint, Installment Account, Collateral Coverage Ratio, and Bank Delinquency. (Appendix C)

Annual Income Joint

Annual Income Joint represents the combined income of the borrower and the co-borrower. Originally, the variable had many missing values, indicating that some loans did not have a co-borrower. To retain the valuable income information, we replaced the missing values with the borrower's income. This variable now considers the total income of a borrower or/and co-borrower, providing a more comprehensive view of the borrower's repayment capacity.

Installment Account

Installment Account is a categorical variable indicating whether a borrower has an installment account (1) or not (0). The original variable contained the number of months since the borrower last opened their installment account and had missing values for borrowers who never opened one. By transforming this variable into a binary output, we preserved its usability and ensured that we could include it in our analysis as a potential risk factor.

Collateral Coverage Ratio

The Collateral Coverage Ratio is calculated by dividing the total installment high credit and credit limit by the loan amount. This ratio indicates the proportion of collateral available relative to the loan amount. A higher ratio suggests more collateral to recover losses, potentially reducing the Loss Given Default (LGD). Including this variable in our analysis helps assess the borrower's collateral situation, which can be a critical factor in predicting the risk of default.

Bank Delinquency

Bank Delinquency is a categorical variable indicating whether the borrower associated with the loans has ever had a delinquency (1) or not (0). The original variable contained the number of months since the borrower's last delinquency, with missing values for those who never had one. By transforming this variable into a binary output, we preserved its importance in our analysis and made it more suitable for predicting loan defaults. Including this variable in our model helps identify borrowers with a history of delinquency, which may be an indicator of their likelihood to default on their loans.

Dataset Division & Variable Selection

Splitting Data into Training and Testing Sets

After completing the data cleaning, preprocessing, and feature engineering, we split the dataset into separate training and testing sets. This step was crucial in maintaining the integrity of our model evaluation process and preventing potential bias and leakage.

By reserving a testing set, we ensured that our variable selection and WoE analysis would be conducted solely on the training set. This mitigated the risk of incorporating patterns or relationships from the testing set into our model, which could lead to information leakage.

Additionally, keeping the testing set independent prevented biased estimates of the model's performance, as the selected variables were optimized solely for the training data.

Weight of Evidence (WoE) Analysis and Variable Selection

Weight of Evidence (WoE) analysis is a valuable technique in credit risk analysis, as it provides an effective method for transforming and encoding categorical variables into numerical values. WoE allows us to analyze and quantify the relationship between the predictor variables and the target variable (loan default), which is critical for selecting the right set of variables to include in the predictive model.

The use of WoE analysis in our variable selection process is beneficial because it considers the unique characteristics of borrowers and loans. By identifying the most relevant variables that have a strong influence on the likelihood of loan default, we can build a model that is both accurate and interpretable. This is crucial for making informed decisions in the banking industry, where the ability to understand the underlying factors contributing to credit risk is essential for effective risk management.

Having identified 40 variables through the WoE analysis, we ensured that our model would be able to capture the complex relationships between borrower and loan characteristics and the probability of default. We have a good number of variables to preserve enough information about the borrowers and loans, allowing our model to make accurate predictions while remaining interpretable. As we proceed with model diagnostics and fine-tuning, we may find that certain combinations of these 40 variables lead to even better model performance. This iterative process will enable us to refine our model further, ultimately enhancing its ability to predict loan defaults and support decision-making in the banking industry. (Table 6)

CORRELATION ANALYSIS

Correlation analysis is a crucial step in the variable selection process, particularly when dealing with many predictor variables, as it helps to identify and address issues of collinearity. Collinearity occurs when two or more predictor variables are highly correlated, causing potential problems in the estimation of model coefficients and the interpretation of results. Moreover, the presence of collinearity can result in overfitting, where the model is too sensitive to the training data and fails to generalize well to unseen data. This can adversely impact the model's interpretability and its ability to accurately predict loan defaults.

By performing correlation analysis on the 40 variables identified through WoE analysis, we can systematically assess the relationships between each variable and identify any potential issues of collinearity. By removing highly correlated variables or combining them into a single representative variable, we can further refine our variable selection and ensure that our final model is both accurate and interpretable. This step ultimately contributes to a more robust credit risk analysis and supports effective decision-making in the banking industry. This step essentially split our working variables in half to 17 variables going into modeling. (Figure 1 and 2)

Methods and Model Development

Methods

To model the probability of default for the loans, we constructed a scorecard using a logistic regression model. To optimize the model's performance, we implemented an elastic net penalty, which allowed us to find the ideal balance between ridge and lasso regularization methods. We then used cross-validation to determine the optimal regularization parameter and ridge regression was chosen as the optimal choice. The response variable was loan status where the outcomes were either default or not.

After fitting the data with a logistic regression model, we recommend using 15 variables as the coefficients demonstrated they would have significant impact. We left out 2 variables that would provide little predictive capability as ridge regression would shrink the coefficients. (Table 1)

We also trained an XGBoost model without using the WoE transformation on the data. This model is based on decision trees and known for handling large datasets. We can now test the difference between using a linear model with WoE transformation versus a non-linear model such as XGBoost. To optimize the model's performance, we used grid search to estimate the best parameters.

Model Performance and Comparison

We fairly assessed both models using performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. A comprehensive table of these metrics is provided in Appendix A. The Logistic Regression model achieved a recall of 0.62, precision of 0.87, F1-score of 0.72, and an AUC-ROC of 0.70. On the other hand, the XGBoost model demonstrated superior performance with a recall of 0.97, precision of 0.81, F1-score of 0.88, and an AUC of 0.74. The comparison of these metrics suggests that the XGBoost model outperforms the Logistic Regression model in terms of accuracy, precision, recall, F1-score, and AUC-ROC. The better performance of the XGBoost model can be attributed to its ability to capture complex nonlinear relationships in the data, which may be difficult for a logistic regression model to capture. (Table 7)

Variable Importance

The Logistic regression model indicated the most influential factors were term length, debt-to-income ratio, and total high credit limit. A longer-term length may indicate a higher likelihood of default due to the extended repayment period and potential for unforeseen circumstances. A higher debt-to-income ratio suggests that the borrower has a greater debt burden, which may affect their ability to repay the loan. Meanwhile, a higher total high credit limit may indicate a borrower's creditworthiness and their ability to manage larger amounts of debt.

Based on the SHAP values extracted from the XGBoost model, the variables that have the highest predictive capabilities were installment, funded amount invested, and term. Higher installment amounts may indicate a larger financial burden for the borrower, which could lead to a higher risk of default. A higher funded amount invested might imply a greater investment by the lender, potentially leading to increased scrutiny of the borrower's creditworthiness and a lower probability of default. Finally, the term variable suggests borrowers with shorter loan terms may have lower overall financial commitments and might be considered less risky by the model. (Figure 4)

Based on the top 3 variables from both models, the 'term' variable is found to be important in both models, indicating that the loan term significantly affects the probability of default across the two models. However, the logistic regression model ranked 'debt-to-income ratio' and 'total high credit limit' as the other top variables, while the XGBoost model identifies installment and funded amount invested as the other top variables. This difference in variable importance is due to the feature selection and the modeling of the data. The logistic regression model incorporates the WoE transformation, whereas the latter does not.

Performance Variables Application

Loss Given Default (LGD) Calculation

The Loss Given Default (LGD) is a measure of the potential loss a bank faces if a borrower defaults on their loan. In our analysis, we calculate LGD for loans in default by considering the original loan amount, the total amount of principal repaid by the borrower until the point of default, and the amount recovered post-charge-off through collection efforts. The LGD calculation considers the actual losses the bank has incurred after the borrower defaults, considering both the principal repayment and any recoveries made after the default.

Loan Profit Calculation

We calculated the loan profit by accounting for both the interest revenue generated from the loan and the associated costs. We first estimated the interest revenue by combining the total interest received so far and the expected interest profit from the remaining payments.

The cost of capital was 50% of the interest revenue, reflecting the lender's cost for providing the funds, and ensuring a proportional relationship between the revenue and the cost. The total late fees received were also considered in our calculation, as they represent additional income for the lender and accounts for the added risk associated with late payments. The total loan profit for all non-defaulters was $2,235,250,118.2.

Optimal Cut-off Point

We determined the optimal cut-off value to be 0.475 for the Probability of Default (PD). This optimal value allowed us to maximize the profit of approximately $58,870,390. This cut-off value allows to accept approximately 51.38% of the total applicants as the accuracy for correctly predicting total applications was 60.54%. While maintaining a higher accuracy of 71.84% for non-defaulters, the accuracy for total defaulters was 42.46%. This approach not only maximizes the financial gains but also supports prudent risk management practices in line with our overall strategy.

PD CALIBRATION

On performing PD calibration for long term values through seasonal effects for macroeconomic variables, we found out that CPI was significant (Table 10).

CONCLUSION AND DISCUSSION

In conclusion, this report presents a comprehensive credit risk analysis using advanced machine learning techniques to predict loan defaults. Our study demonstrates the value of incorporating logistic regression and XGBoost models, with the latter exhibiting superior performance in terms

of accuracy, precision, recall, F1-score, and AUC-ROC. By comparing these models, we have identified key variables that significantly impact the probability of default, enabling financial institutions to make more informed decisions regarding loan approvals.

In addition to model comparison, our analysis also focused on the practical applications of the findings, including Loss Given Default (LGD) and loan profit for non-defaulters of $2,235,250,118.2, as well as determining the optimal cut-off value of 0.475 for loan acceptance. These performance variables offer valuable insights for banks to enhance their risk management strategies and optimize profitability. It is crucial to consider the trade-offs between different models and select the most suitable one based on the specific goals and requirements of the analysis. For instance, if accuracy in default prediction is the primary concern, the XGBoost model would be the preferred choice, whereas the logistic regression model might be favored for its interpretability. Furthermore, our results may serve as a basis for future work, such as exploring alternative machine learning models or incorporating new variables to further improve the predictive power of the models.

It is important to note that this analysis was conducted on a dataset that may not be representative of all loan scenarios, and the results may vary depending on the specific context. Future studies could utilize larger and more diverse datasets to validate and extend the findings presented in this report. Additionally, the incorporation of novel variables or the application of more sophisticated models, such as deep learning techniques, could further enhance the prediction accuracy and offer new insights into credit risk management.

Source Code: https://github.com/preetmodi/Credit-Risk-Analytics

APPENDIX A: TABLES

TABLE 1: LOGISTIC REGRESSION COEFFICIENTS

	Variables	Coefficients
0	open_rv_24m_woe	0.420184
1	fico_range_high_woe	0.515041
2	il_util_woe	0.458955
3	bc_open_to_buy_woe	0.260928
4	acc_open_past_24mths_woe	0.513377
5	total_bc_limit_woe	0.285878
6	inq_last_6mths_woe	0.374993
7	percent_bc_gt_75_woe	0.346154
8	term_woe	1.102339
9	annual_inc_joint_woe	0.283621
10	bc_util_woe	0.031386
11	mo_sin_rcnt_tl_woe	0.092458
12	tot_hi_cred_lim_woe	0.738449
13	mths_since_recent_inq_woe	0.483171

	Variables	Coefficients
14	dti_woe	0.865521
15	dti_joint_woe	0.579215
16	mths_since_recent_bc_woe	0.334838

TABLE 2: MISSING VALUE IMPUTED WITH -1

	Variables	No. of Null Values
0	dti_joint	1215152
1	mths_since_recent_bc_dlq	982792
2	mths_since_rcnt_il	531540
3	total_bal_il	507026
4	il_util	618602
5	mo_sin_old_il_acct	78075
6	mo_sin_old_rev_tl_op	38816
7	mo_sin_rcnt_rev_tl_op	38816
8	mo_sin_rcnt_tl	38815
9	all_util	507145
12	open_il_12m	507026
13	open_acc_6m	507026
14	open_act_il	507026
15	open_il_24m	507026
16	open_il_12m	507026
17	open_rv_12m	507026
18	open_rv_24m	507026
19	max_bal_bc	507026
20	mths_since_recent_revol_delinq	857372
21	total_cu_tl	507026
22	inq_last_12m	507026
23	inq_fi	507026
24	mths_since_recent_inq	160464

TABLE 3: MISSING VALUE IMPUTED WITH MEDIAN

	Variable	No. of Null
0	chargeoff_within_12_mths	56
1	dti	939
2	revol_util	1008
3	tot_coll_amt	38815
4	tot_cur_bal	38815
5	acc_open_past_24mths	27008

6	avg_cur_bal	38851
7	bc_open_to_buy	41185
8	bc_util	41853
9	chargeoff_within_12_mths	56
10	mort_acc	27008
11	mths_since_recent_bc	40315
12	num_accts_ever_120_pd	38815
13	num_actv_bc_tl	38815
14	num_actv_rev_tl	38815
15	num_bc_sats	31968
16	num_bc_tl	38815
17	num_il_tl	38815
18	num_op_rev_tl	38815
19	num_rev_accts	38815
20	num_rev_tl_bal_gt_0	38815
21	num_sats	31968
22	pct_tl_nvr_dlq	38910
23	percent_bc_gt_75	41444
24	pub_rec_bankruptcies	435
25	tax_liens	46
26	tot_hi_cred_lim	38815
27	total_bal_ex_mort	27008
28	total_bc_limit	27008
29	total_il_high_credit_limit	38815

TABLE 4: DELETED VARIABLES

	Variable	**No. of Null**
0	annual_inc	31
1	mths_since_last_record	1072333
2	emp_length	85395
3	emp_title	96177
4	desc	1209498
5	revol_bal_joint	1222472
6	sec_app_earliest_cr_line	1222471
7	sec_app_inq_last_6mths	1222471
8	sec_app_mort_acc	1222471
9	sec_app_open_acc	1222471
10	sec_app_revol_util	1223480
11	sec_app_open_act_il	1222471
12	sec_app_num_rev_accts	1222471
13	sec_app_chargeoff_within_12_mths	1222471
14	sec_app_collections_12_mths_ex_med	1222471
15	sec_app_mths_since_last_major_derog	1260860
16	title	14319

TABLE 5: VARIABLE CUT-OFF VALUES

	Variable	**Cut off Value**

0	installment	1650.0
1	dti	900.0
2	delinq_2yrs	30.0
3	open_acc	75.0
4	pub_rec	40.0
5	revol_bal	1500000.0
6	revol_util	200.0
7	total_acc	125.0
8	annual_inc_joint	500000.0
9	dti_joint	40.0
10	tot_coll_amt	500000.0
11	tot_cur_bal	1000000.0
12	open_acc_6m	10.0
13	open_act_il	40.0
14	open_il_12m	10.0
15	open_il_24m	20.0
16	mths_since_rcnt_il	150.0
17	total_bal_il	300000.0
18	il_util	400.0
19	open_rv_12m	10.0
20	open_rv_24m	20.0
21	max_bal_bc	150000.0
22	all_util	200.0
23	inq_fi	15.0
24	total_cu_tl	40.0
25	inq_last_12m	13.0
26	acc_open_past_24mths	19.0
27	avg_cur_bal	110000.0
28	bc_open_to_buy	110000.0
29	bc_util	120.0
30	chargeoff_within_12_mths	1.0
31	mo_sin_old_il_acct	350.0
32	mo_sin_rcnt_rev_tl_op	110.0
33	mo_sin_rcnt_tl	60.0
34	mort_acc	11.0
35	mths_since_recent_bc	190.0
36	mths_since_recent_revol_delinq	80.0
37	num_accts_ever_120_pd	8.0
38	num_actv_bc_tl	15.0
39	num_actv_rev_tl	22.0
40	num_bc_sats	19.0
41	num_bc_tl	30.0
42	num_il_tl	50.0
43	num_op_rev_tl	35.0
44	num_rev_accts	60.0
45	num_rev_tl_bal_gt_0	30.0
46	num_sats	40.0
47	pct_tl_nvr_dlq	50.0
48	pub_rec_bankruptcies	2.0
49	tax_liens	15.0
50	tot_hi_cred_lim	1800000.0
51	total_bal_ex_mort	600000.0
52	total_bc_limit	200000.0

53	total_il_high_credit_limit	400000.0
54	payment_history	-0.3
55	collateral_coverage_ratio	50.0

TABLE 6: WEIGHT OF EVIDENCE VALUES

	variable	**info_value**
0	open_rv_24m_woe	0.20537237
1	inq_last_12m_woe	0.19893216
2	all_util_woe	0.19724753
3	open_rv_12m_woe	0.19029613
4	open_acc_6m_woe	0.18999127
5	open_il_12m_woe	0.18470927
6	open_il_24m_woe	0.18107621
7	max_bal_bc_woe	0.17389522
8	inq_fi_woe	0.17089161
9	mths_since_rcnt_il_woe	0.17078523
10	total_bal_il_woe	0.16438414
11	open_act_il_woe	0.16344332
12	fico_range_low_woe	0.16327031
13	fico_range_high_woe	0.16327031
14	total_cu_tl_woe	0.16256897
15	il_acc_woe	0.15105693
16	il_util_woe	0.12763499
17	bc_open_to_buy_woe	0.09895391
18	acc_open_past_24mths_woe	0.07957303
19	total_bc_limit_woe	0.06963993
20	inq_last_6mths_woe	0.06327448
21	percent_bc_gt_75_woe	0.06161792
22	term_woe	0.05918398
23	annual_inc_joint_woe	0.05496803
24	bc_util_woe	0.05334222
25	mo_sin_rcnt_tl_woe	0.04901634
26	tot_hi_cred_lim_woe	0.04889122
27	mo_sin_rcnt_rev_tl_op_woe	0.04574617
28	mths_since_recent_inq_woe	0.04509181
29	revol_util_woe	0.04492675
30	avg_cur_bal_woe	0.04279111
31	dti_woe	0.0422689
32	tot_cur_bal_woe	0.03587336
33	dti_joint_woe	0.03393333
34	application_type_woe	0.0324572
35	mths_since_recent_bc_woe	0.03184732
36	num_actv_rev_tl_woe	0.02860826
37	num_rev_tl_bal_gt_0_woe	0.02728286
38	earliest_cr_line	0.02650722
39	home_ownership_woe	0.02435886

40	installment_woe	0.02342356
41	mo_sin_old_rev_tl_op_woe	0.02115268
42	purpose_woe	0.02013935
43	funded_amnt_inv_woe	0.01967689
44	collateral_coverage_ratio_woe	0.01915295
45	funded_amnt_woe	0.01878853
46	loan_amnt_woe	0.01877161
47	mort_acc_woe	0.01558042
48	pub_rec_woe	0.01171103
49	num_op_rev_tl_woe	0.00982296
50	mo_sin_old_il_acct_woe	0.00941634
51	revol_bal_woe	0.00938574
52	num_sats_woe	0.00770166
53	total_bal_ex_mort_woe	0.00746133
54	num_rev_accts_woe	0.00737987
55	total_il_high_credit_limit_woe	0.00723505
56	num_actv_bc_tl_woe	0.00633616
57	open_acc_woe	0.00584911
58	pub_rec_bankruptcies_woe	0.0051378
59	delinq_2yrs_woe	0.00400871
60	num_accts_ever_120_pd_woe	0.00344673
61	num_bc_tl_woe	0.00338783
62	mths_since_recent_bc_dlq_woe	0.00336138
63	total_acc_woe	0.0032016
64	pct_tl_nvr_dlq_woe	0.00311274
65	payment_history_woe	0.00294863
66	mths_since_recent_revol_delinq_woe	0.00289326
67	tax_liens_woe	0.00270684
68	bank_dlq_woe	0.00258039
69	tot_coll_amt_woe	0.00231411
70	num_il_tl_woe	0.00203264
71	num_bc_sats_woe	0.00107558
72	chargeoff_within_12_mths_woe	0

TABLE 7: MODEL METRICS

Metrics	Logistic Regression	XGB
Precision	0.32	0.57
Recall	0.67	0.14
F1	0.43	0.23
AUC	0.702	0.743

TABLE 8: CUT-OFF VALUES (OPTIMAL VALUE)

cuts	accepted	non_def_acc	def_acc	tot_acc	avg_non_def	avg_def	tot_non_def_profit	tot_def_profit	total_profit

0.1	0.01282309	0.99840443	0.984197	0.22189686	1513.31328	-13873.363	3388308.44	-832401.78	2555906.66
0.115	0.01942706	0.99744708	0.97609435	0.22809924	1559.64405	-13879.497	5282514.4	-1332431.7	3950082.67
0.13	0.02691231	0.99606425	0.96698946	0.23500441	1583.04919	-13934.02	7403921.07	-2062234.9	5341686.13
0.145	0.03622703	0.99414956	0.95571068	0.24351595	1593.56479	-13698.714	9999619.06	-3013717.2	6985901.88
0.16	0.04655132	0.99202213	0.94321085	0.2529478	1628.0953	-13358.278	13099654.8	-4007483.3	9092171.48
0.175	0.05852102	0.98888416	0.92889711	0.26360117	1674.07705	-13101.411	16864652.2	-5476389.9	11388262.3
0.19	0.07127718	0.98582598	0.913567	0.27507446	1705.94799	-13432.643	20891039.1	-7159598.7	13731440.3
0.205	0.08542775	0.98247527	0.89655002	0.28781946	1746.8411	-13056.214	25603450	-8604044.9	16999405.1
0.22	0.10067713	0.97678438	0.87876371	0.30068159	1763.16312	-12696.895	30285852.8	-11084390	19201463.2
0.235	0.11700858	0.97069461	0.85971401	0.31445846	1795.66805	-12687.404	35690698.2	-13981520	21709178.5
0.25	0.13509142	0.96367408	0.83869511	0.32959629	1831.01453	-12314.606	41846006.1	-16821752	25024254.4
0.265	0.15454079	0.95625465	0.81605285	0.34593331	1865.24781	-12144.538	48612088.5	-19977764	28634324.3
0.28	0.17479892	0.94713328	0.79283889	0.36236516	1900.37153	-12025.835	55777804.7	-23907359	31870445.3
0.295	0.19660765	0.93718753	0.76788159	0.38000178	1922.24238	-11927.818	63216785	-28173507	35043278
0.31	0.21904109	0.92657696	0.74231024	0.39798423	1950.28816	-11804.687	71205020.6	-32592742	38612279
0.325	0.24246734	0.91463674	0.71583546	0.41640173	1960.93876	-11583.36	78949355.3	-37182586	41766768.9
0.34	0.26705934	0.90195192	0.68808317	0.43567261	1976.81682	-11533.2	87361465.6	-42522907	44838558.2
0.355	0.29204177	0.88690033	0.66046498	0.45434111	1987.84837	-11362.532	95627433.8	-48324849	47302584.4
0.37	0.31757081	0.87059887	0.63248684	0.47303192	2001.57555	-11260.315	104222039	-54792691	49429347.8
0.385	0.34455563	0.85251569	0.60313942	0.49243109	2014.37268	-11150.361	113264147	-61839902	51424244.8
0.4	0.37193646	0.8333422	0.57358027	0.51176891	2024.33912	-11051.864	122302472	-69262032	53040440.3
0.415	0.39912207	0.81448782	0.54418345	0.53104537	2041.46645	-10953.179	131839945	-76409376	55430568.7
0.43	0.42724474	0.79289437	0.51432786	0.55010988	2057.10334	-10823.009	141551338	-84289593	57261744.9
0.445	0.45608692	0.77013084	0.48387233	0.56940308	2067.58396	-10744.157	151194145	-92872493	58321651.9
0.46	0.48488449	0.74470801	0.45417908	0.58753612	2076.91056	-10611.193	160613724	-101867452	58746272.1
0.475	0.51379918	0.71840762	0.42457052	0.60541816	2087.65193	-10513.901	170202086	-111331698	58870388.4
0.49	0.54267483	0.69056483	0.39542073	0.62261415	2092.59022	-10403.401	179247093	-121053975	58193118.3
0.505	0.57182937	0.66117966	0.36632741	0.63944201	2096.89015	-10321.15	188258798	-131501769	56757028.2
0.52	0.60087793	0.63126263	0.33750935	0.65594079	2101.30651	-10235.197	197234933	-141921241	55313692.1
0.535	0.62950816	0.60065419	0.30940416	0.6717312	2103.23903	-10148.644	205791423	-152402189	53389233.9
0.55	0.65864596	0.56770556	0.28127779	0.68704751	2108.56485	-10078.526	214715159	-163836515	50878644.1
0.565	0.68684671	0.53481013	0.25432306	0.70144908	2119.81029	-10007.158	223955838	-175055216	48900622.1
0.58	0.71477974	0.50018615	0.22816589	0.71485782	2126.01501	-9961.7713	232490372	-187231491	45258881.2

0.595	0.74231117	0.46370067	0.20301097	0.7270841	2132.20118	-9927.9291	240766025	-200216546	40549479.8
0.61	0.76901152	0.42795979	0.17871007	0.73879165	2137.55002	-9863.9062	248729596	-212182485	36547110.2
0.625	0.79452383	0.39256462	0.15582078	0.74945618	2145.66686	-9816.566	256632485	-224230000	32402485.3
0.64	0.81885368	0.3551218	0.13497127	0.75807927	2148.45248	-9762.3741	263312188	-236737573	26574614.9
0.655	0.84191181	0.31847676	0.11551926	0.76576531	2154.89966	-9740.8075	270041251	-249637415	20403835.6
0.67	0.86370938	0.28326774	0.09728124	0.77279319	2162.55726	-9722.7394	276588911	-262047272	14541638.6
0.685	0.88420736	0.24795235	0.08071597	0.77847685	2167.43354	-9710.4021	282299549	-274610170	7689378.22
0.7	0.90300414	0.21425912	0.06587287	0.78313979	2173.60435	-9700.1496	287674363	-286610320	1064042.15
0.715	0.91974276	0.18373045	0.05279429	0.78707205	2179.58626	-9698.9037	292504836	-297707850	-5203014.5
0.73	0.93582879	0.15221785	0.04080264	0.78993898	2187.48389	-9697.8138	297281249	-309166304	-11885056
0.745	0.94943275	0.12331135	0.03126015	0.79141707	2194.35891	-9719.2943	301182343	-320415976	-19233632
0.76	0.96141361	0.09818104	0.0227693	0.79285611	2200.38371	-9731.9924	304656327	-330031326	-25374999
0.775	0.97148132	0.07533773	0.01609238	0.79334137	2205.53813	-9753.1659	307456426	-339127331	-31670905
0.79	0.97975302	0.05515371	0.01098234	0.79314615	2209.65581	-9775.2094	309630230	-347313189	-37682959
0.805	0.98644624	0.03824061	0.0070016	0.79274455	2213.16019	-9803.1738	311369508	-354541583	-43172076
0.82	0.9912542	0.0255292	0.0042913	0.79222025	2216.38368	-9820.6262	312674111	-359867026	-47192914
0.835	0.99511953	0.01433358	0.00237151	0.79138918	2218.58874	-9842.669	313588645	-364818526	-51229881
0.85	0.99760718	0.00741942	0.00105871	0.79097643	2220.09262	-9852.9356	314214149	-367760822	-53546673
0.865	0.99909642	0.00281885	0.00039525	0.79053579	2220.905	-9858.8224	314537892	-369686124	-55148232

TABLE 9: VARIABLE DESCRIPTION

	Variable	Description
1	collection_recovery_fee	post charge off collection fee
2	delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
3	issue_d	The month which the loan was funded
4	last_pymnt_amnt	Last total payment amount received
5	last_pymnt_d	Last month payment was received
6	loan_status	Current status of the loan
7	mths_since_last_delinq	The number of months since the borrower's last delinquency.
8	mths_since_last_major_derog	Months since most recent 90-day or worse rating
9	next_pymnt_d	Next scheduled payment date
10	pymnt_plan	Indicates if a payment plan has been put in place for the loan
11	recoveries	post charge off gross recovery
12	hardship_flag	Flags whether or not the borrower is on a hardship plan
13	hardship_type	Describes the hardship plan offering

14	hardship_reason	Describes the reason the hardship plan was offered
15	hardship_status	Describes if the hardship plan is active, pending, canceled, completed, or broken
16	deferral_term	Amount of months that the borrower is expected to pay less than the contractual monthly payment amount due to a hardship plan
17	hardship_amount	The interest payment that the borrower has committed to make each month while they are on a hardship plan
18	hardship_start_date	The start date of the hardship plan period
19	hardship_end_date	The end date of the hardship plan period
20	payment_plan_start_date	The day the first hardship plan payment is due. For example, if a borrower has a hardship plan period of 3 months, the start date is the start of the three-month period in which the borrower is allowed to make interest-only payments.
21	hardship_length	The number of months the borrower will make smaller payments than normally obligated due to a hardship plan
22	hardship_dpd	Account days past due as of the hardship plan start date
23	hardship_loan_status	Loan Status as of the hardship plan start date
24	orig_projected_additional_accrued_interest	The original projected additional interest amount that will accrue for the given hardship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan.
25	hardship_payoff_balance_amount	The payoff balance amount as of the hardship plan start date
26	hardship_last_payment_amount	The last payment amount as of the hardship plan start date
27	settlement_status	The status of the borrower's settlement plan. Possible values are COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT
28	settlement_date	The date that the borrower agrees to the settlement plan
29	settlement_amount	The loan amount that the borrower has agreed to settle for
30	settlement_percentage	The settlement amount as a percentage of the payoff balance amount on the loan
31	settlement_term	The number of months that the borrower will be on the settlement plan
32	int_rate	Interest Rate on the loan
33	total_pymnt	Payments received to date for total amount funded
34	total_rec_int	Interest received to date
35	total_rec_late_fee	Late fees received to date
36	total_rec_prncp	Principal received to date
37	total_rev_hi_lim	Total revolving high credit/credit limit
38	out_prncp	Remaining outstanding principal for total amount funded
39	out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
40	id	The loan ID
41	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.
42	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
43	acc_open_past_24mths	Number of trades opened in past 24 months.
44	addr_state	The state provided by the borrower in the loan application
45	all_util	Balance to credit limit on all trades

46	annual_inc	The self reported annual income provided by the borrower during registration.
47	annual_inc_joint	The combined self-reported annual income provided by the coborrowers during registration
48	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
49	avg_cur_bal	Average current balance of all accounts
50	bc_open_to_buy	Total open to buy on revolving bankcards.
51	bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
52	chargeoff_within_12_mths	Number of charge-offs within 12 months
53	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
54	desc	Loan description provided by the borrower
55	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.
56	dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co borrowers' combined self-reported monthly income
57	earliest_cr_line	The month the borrower's earliest reported credit line was opened
58	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.
59	emp_title	The job title supplied by the Borrower when applying for the loan.
60	fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
61	fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
62	funded_amnt	The total amount committed to that loan at that point in time.
63	funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
64	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
65	il_util	Ratio of total current balance to high credit/credit limit on all install acct
66	inq_fi	Number of personal finance inquiries
67	inq_last_12m	Number of credit inquiries in past 12 months
68	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
69	max_bal_bc	Maximum current balance owed on all revolving accounts
70	mo_sin_old_il_acct	Months since oldest bank installment account opened
71	mo_sin_old_rev_tl_op	Months since oldest revolving account opened
72	mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
73	mo_sin_rcnt_tl	Months since most recent account opened
74	mort_acc	Number of mortgage accounts.
75	mths_since_last_record	The number of months since the last public record.

76	mths_since_rcnt_il	Months since most recent installment accounts opened
77	mths_since_recent_bc	Months since most recent bankcard account opened.
78	mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
79	mths_since_recent_inq	Months since most recent inquiry.
80	mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
81	num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
82	num_actv_bc_tl	Number of currently active bankcard accounts
83	num_actv_rev_tl	Number of currently active revolving trades
84	num_bc_sats	Number of satisfactory bankcard accounts
85	num_bc_tl	Number of bankcard accounts
86	num_il_tl	Number of installment accounts
87	num_op_rev_tl	Number of open revolving accounts
88	num_rev_accts	Number of revolving accounts
89	num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
90	num_sats	Number of satisfactory accounts
91	open_acc	The number of open credit lines in the borrower's credit file.
92	open_acc_6m	Number of open trades in last 6 months
93	open_il_12m	Number of installment accounts opened in past 12 months
94	open_il_24m	Number of installment accounts opened in past 24 months
95	open_act_il	Number of currently active installment trades
96	open_rv_12m	Number of revolving trades opened in past 12 months
97	open_rv_24m	Number of revolving trades opened in past 24 months
98	pct_tl_nvr_dlq	Percent of trades never delinquent
99	percent_bc_gt_75	Percentage of all bankcard accounts > 75
100	pub_rec	Number of derogatory public records
101	pub_rec_bankruptcies	Number of public record bankruptcies
102	purpose	A category provided by the borrower for the loan request.
103	revol_bal	Total credit revolving balance
104	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
105	tax_liens	Number of tax liens
106	title	The loan title provided by the borrower
107	tot_coll_amt	Total collection amounts ever owed
108	tot_cur_bal	Total current balance of all accounts
109	tot_hi_cred_lim	Total high credit/credit limit
110	total_acc	The total number of credit lines currently in the borrower's credit file
111	total_bal_ex_mort	Total credit balance excluding mortgage
112	total_bal_il	Total current balance of all installment accounts
113	total_bc_limit	Total bankcard high credit/credit limit
114	total_cu_tl	Number of finance trades
115	total_il_high_credit_limit	Total installment high credit/credit limit
116	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
117	revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances
118	sec_app_earliest_cr_line	Earliest credit line at time of application for the secondary applicant
119	sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant

120	sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant
121	sec_app_open_acc	Number of open trades at time of application for the secondary applicant
122	sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts
123	sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant
124	sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant
125	sec_app_chargeoff_within_12_mths	Number of charge-ffs within last 12 months at time of application for the secondary applicant
126	sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant
127	sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant
128	annual_inc_joint	The combined self-reported annual income provided by the co borrowers during registration
129	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.
130	dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
131	sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant
132	installment	The monthly payment owed by the borrower if the loan originates.

TABLE 10: PD CALIBRATION

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
x1	0.0028	0.006	0.43	0.667	-0.01	0.015
x2	-0.0117	0.007	-1.691	0.091	-0.025	0.002
ar.L1	0.184	0.135	1.364	0.172	-0.08	0.448
sigma2	0.0004	8.24E-05	4.566	0	0	0.001

TABLE 11: CUT-OFF DESCRIPTION

Cuts	**cut-off value (The cut off we accept based on Probability of default)**
Accepted	Total percentage of people that applied we accept
Non defaulters accuracy	out of the total non-defaulters applied, how many of them were accepted
Defaulters accuracy	out of the total defaulters applied, how many of them were accepted

Total accuracy	Out of the total people applied, out of how many of them were correctly predicted
Average non defaulter	average profit for all accepted non-defaulters
Average defaulter	average profit for all accepted defaulters, for defaulters its loss (negative values)
total non defaulters profit	number of non-defaulters multiplied by average non defaulters
total defaulters profit	number of defaulters multiplied by average defaulters
total profit	sum of the total non defaulters profit and total defaulters profit

APPENDIX B: VISUALIZATION

FIGURE 1: TOTAL CORRELATION PLOT

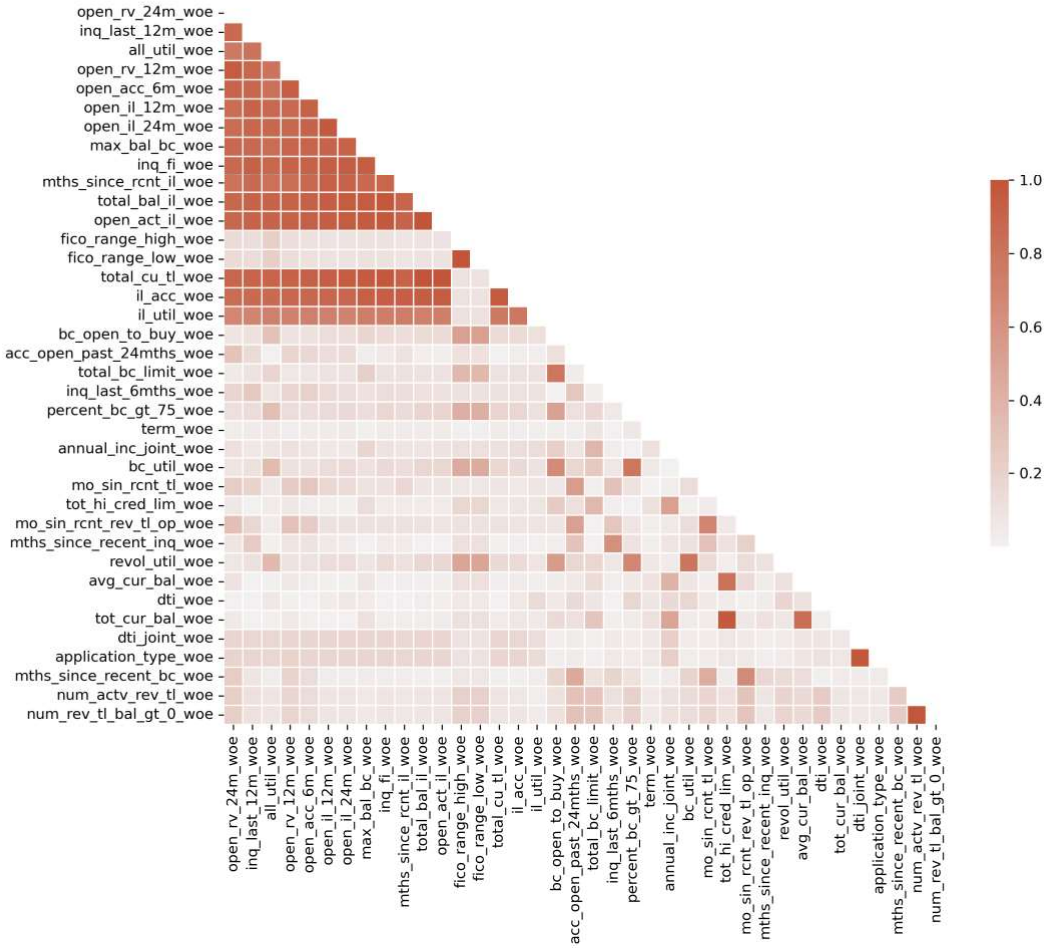


FIGURE 2: SUBSET CORRELATION PLOT

Figure 3: XGBOOST VARIABLE IMPORTANCE

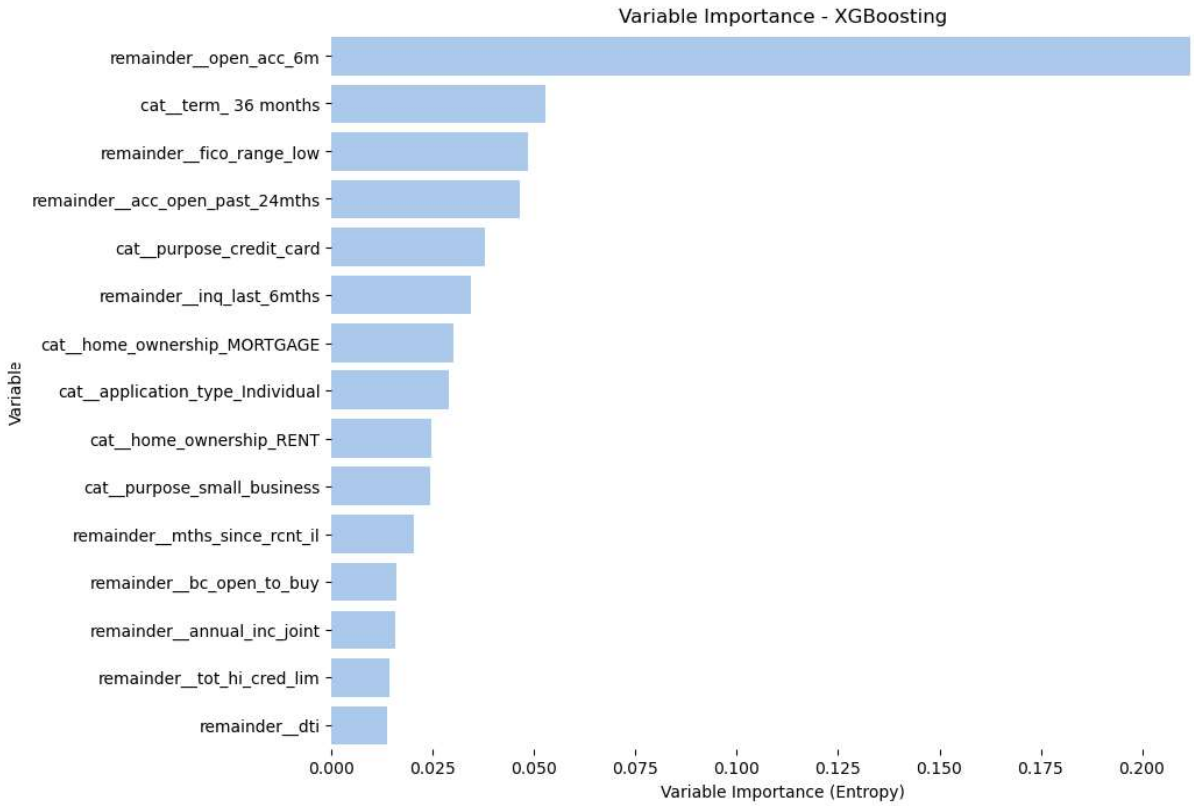
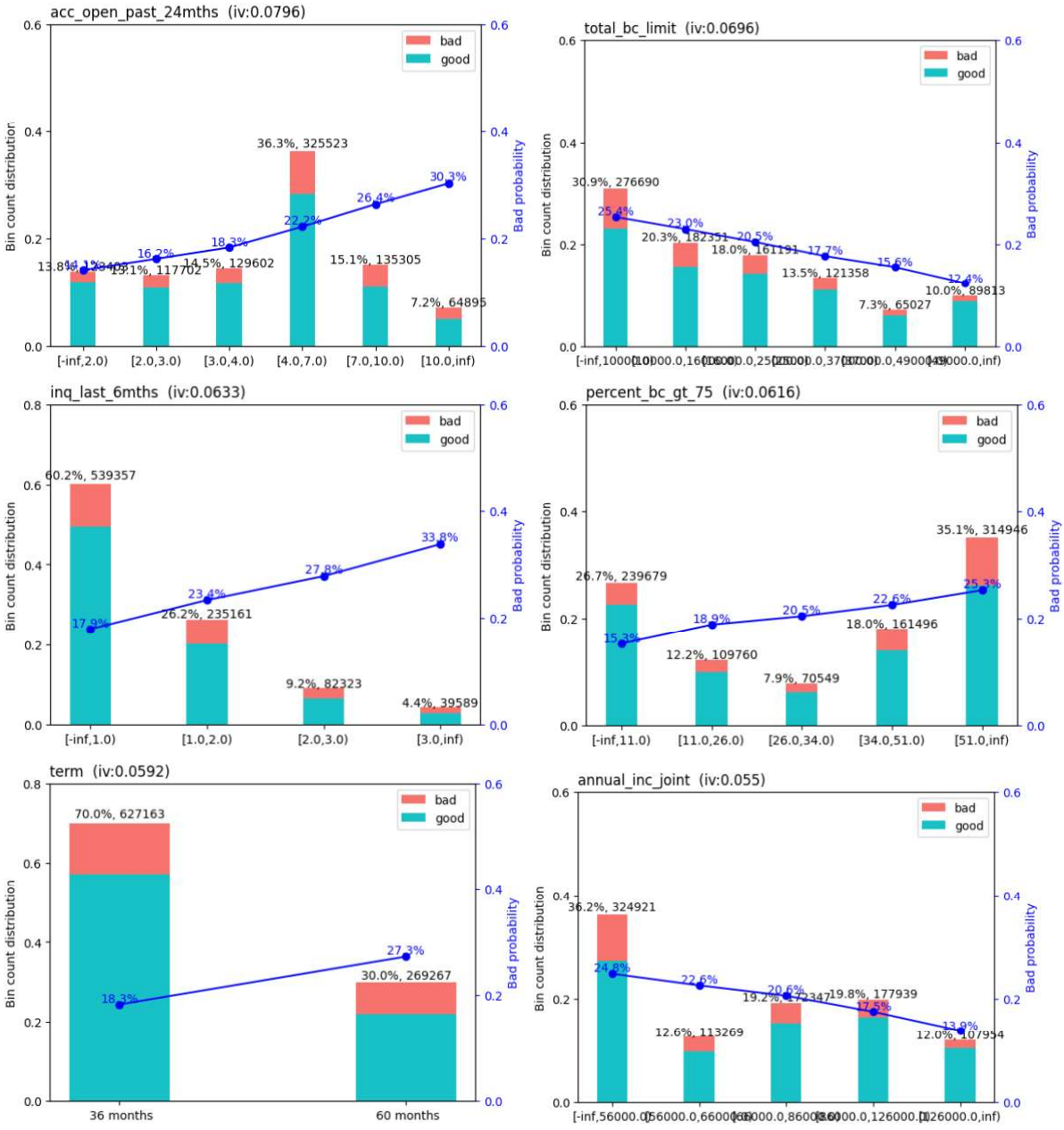
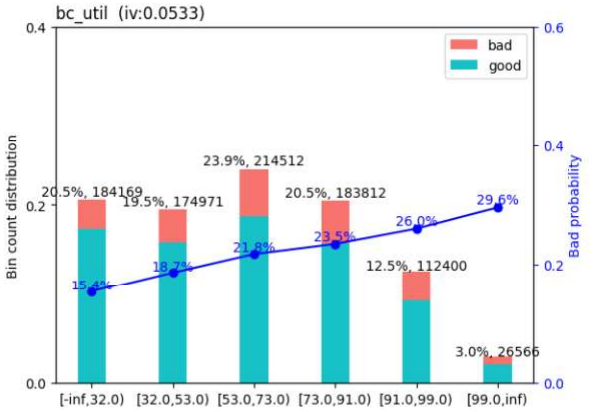
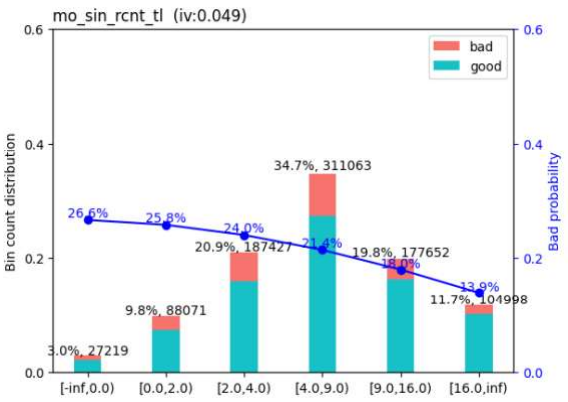
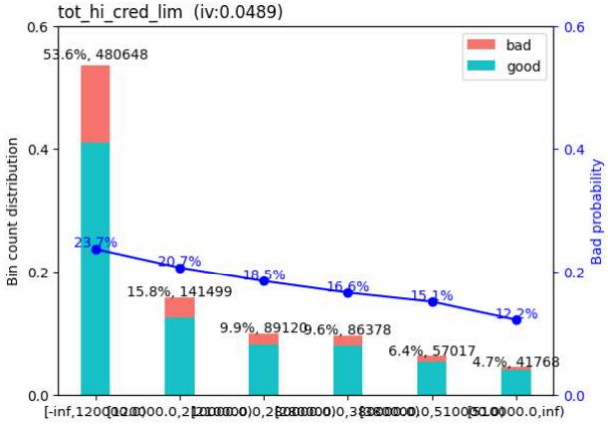


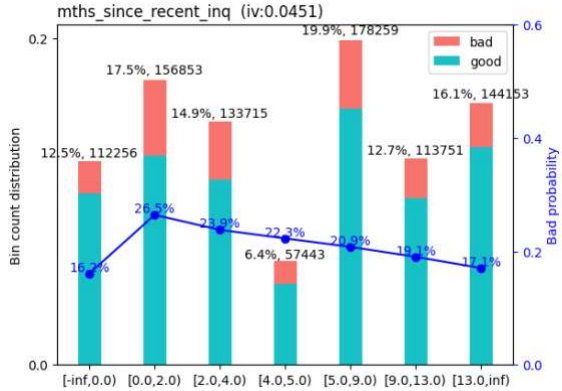
FIGURE 4: WOE PLOTS

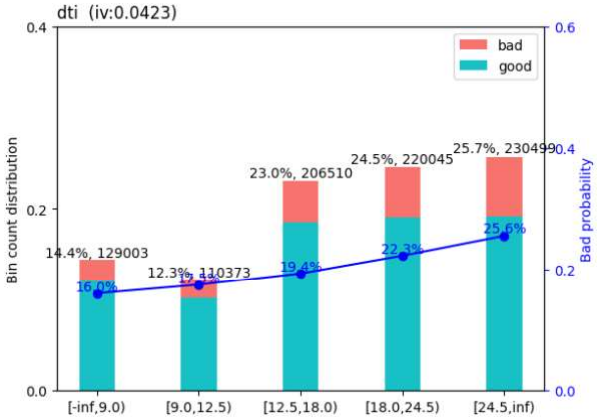


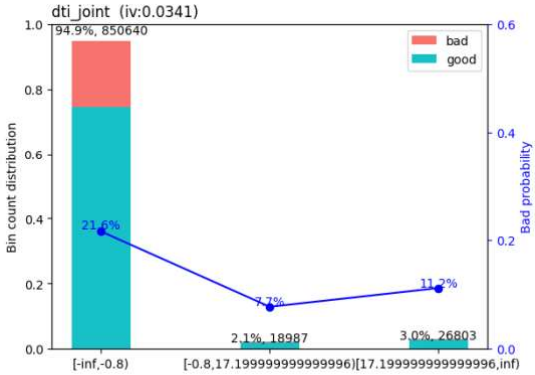

bc_util (iv:0.0533)
bad
good
Bin count distribution
Bad probability
20.5%, 184169
19.5%, 174971
23.9%, 214512
20.5%, 183812
12.5%, 112400
3.0%, 26566
15.4%
18.7%
21.8%
23.5%
26.0%
29.6%
[-inf,32.0)
[32.0,53.0)
[53.0,73.0)
[73.0,91.0)
[91.0,99.0)
[99.0,inf)

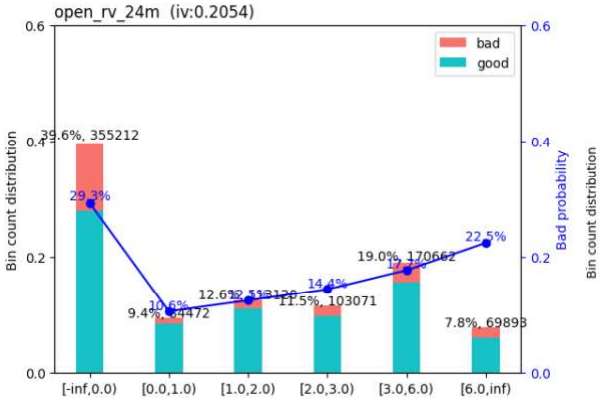

mo_sin_rcnt_tl (iv:0.049)
bad
good
Bin count distribution
Bad probability
3.0%, 27219
9.8%, 88071
20.9%, 187427
34.7%, 311063
19.8%, 177652
11.7%, 104998
26.6%
25.8%
24.0%
21.4%
18.0%
13.9%
[-inf,0.0)
[0.0,2.0)
[2.0,4.0)
[4.0,9.0)
[9.0,16.0)
[16.0,inf)

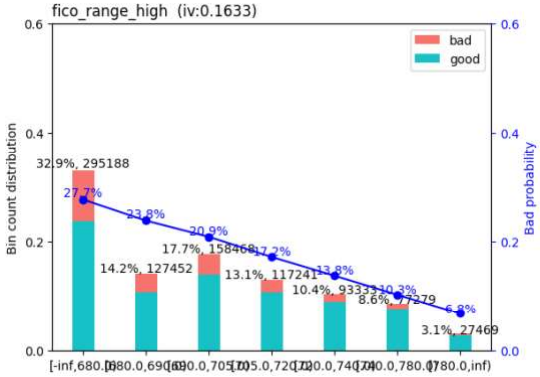

tot_hi_cred_lim (iv:0.0489)
bad
good
Bin count distribution
Bad probability
53.6%, 480648
15.8%, 141499
9.9%, 89120
9.6%, 86378
6.4%, 57017
4.7%, 41768
23.7%
20.7%
18.5%
16.6%
15.1%
12.2%

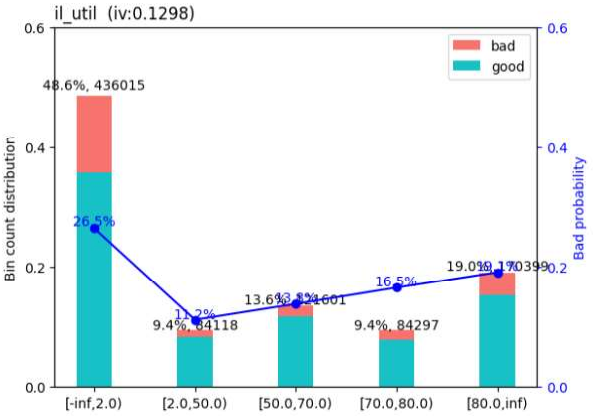

mths_since_recent_inq (iv:0.0451)
bad
good
Bin count distribution
Bad probability
12.5%, 112256
17.5%, 156853
14.9%, 133715
6.4%, 57443
19.9%, 178259
12.7%, 113751
16.1%, 144153
16.2%
26.5%
23.9%
22.3%
20.9%
19.1%
17.1%
[-inf,0.0)
[0.0,2.0)
[2.0,4.0)
[4.0,5.0)
[5.0,9.0)
[9.0,13.0)
[13.0,inf)

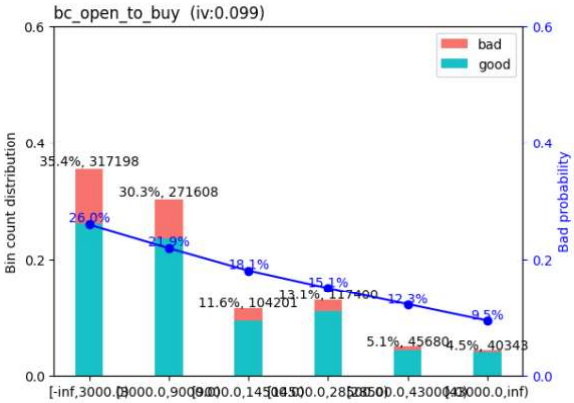

dti (iv:0.0423)
bad
good
Bin count distribution
Bad probability
14.4%, 129003
12.3%, 110373
23.0%, 206510
24.5%, 220045
25.7%, 230499
16.0%
19.4%
22.3%
25.6%
[-inf,9.0)
[9.0,12.5)
[12.5,18.0)
[18.0,24.5)
[24.5,inf)


dti_joint (iv:0.0341)
bad
good
Bin count distribution
Bad probability
94.9%, 850640
2.1%, 18987
3.0%, 26803
21.6%
7.7%
11.2%
[-inf,-0.8)
[-0.8,17.199999999999996)
[17.199999999999996,inf)


open_rv_24m (iv:0.2054)
bad
good
Bin count distribution
Bad probability
39.6%, 355212
9.4%, 84472
12.6%
11.5%, 103071
19.0%, 170662
7.8%, 69893
29.3%
10.6%
14.4%
22.5%
[-inf,0.0)
[0.0,1.0)
[1.0,2.0)
[2.0,3.0)
[3.0,6.0)
[6.0,inf)


fico_range_high (iv:0.1633)
bad
good
Bin count distribution
Bad probability
32.9%, 295188
14.2%, 127452
17.7%, 158468
13.1%, 117241
10.4%, 93333
8.6%, 77279
3.1%, 27469
27.7%
23.8%
20.8%
17.2%
13.8%
10.3%
6.8%
[-inf,680.0)
[680.0,69
[700.0,705.0)
[705.0,72
[720.0,740.0)
[740.0,780.0)
[780.0,inf)


il_util (iv:0.1298)
bad
good
Bin count distribution
Bad probability
48.6%, 436015
26.5%
9.4%, 84118
9.4%, 84297
16.5%
[-inf,2.0)
[2.0,50.0)
[50.0,70.0)
[70.0,80.0)
[80.0,inf)


bc_open_to_buy (iv:0.099)
bad
good
Bin count distribution
Bad probability
35.4%, 317198
30.3%, 271608
26.0%
21.9%
18.1%
11.6%, 104201
15.1%
13.1%, 117490
12.3%
5.1%, 45680
4.5%, 40343
9.5%

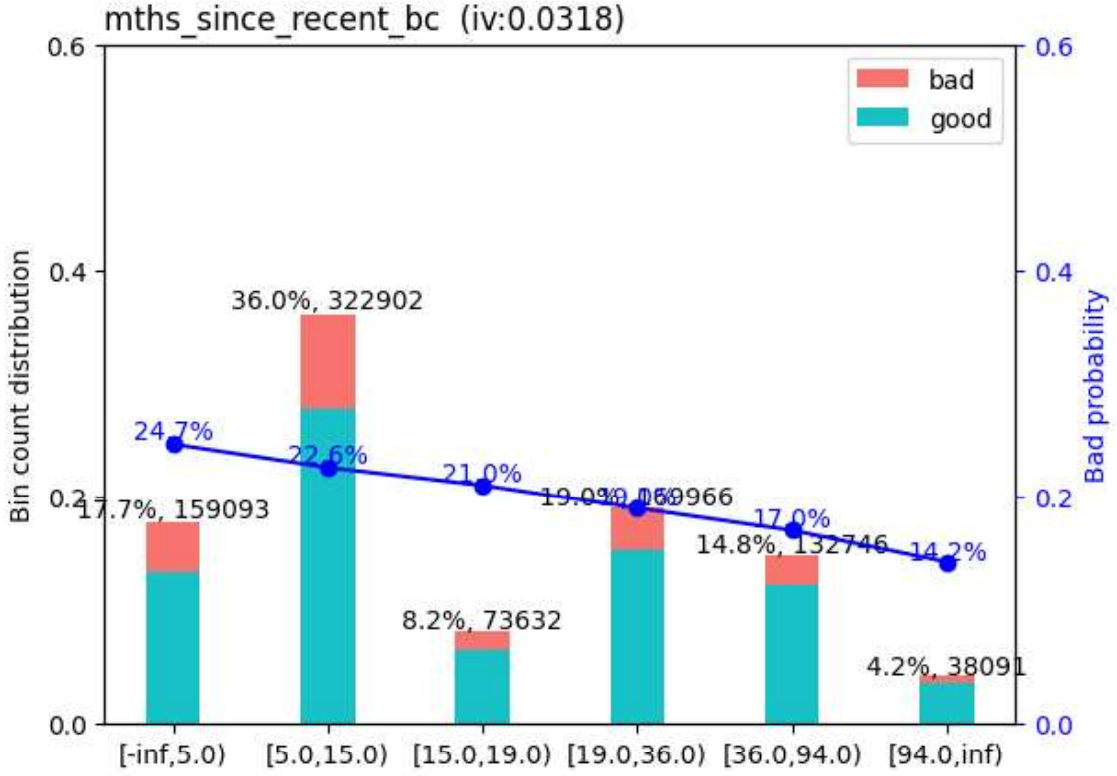
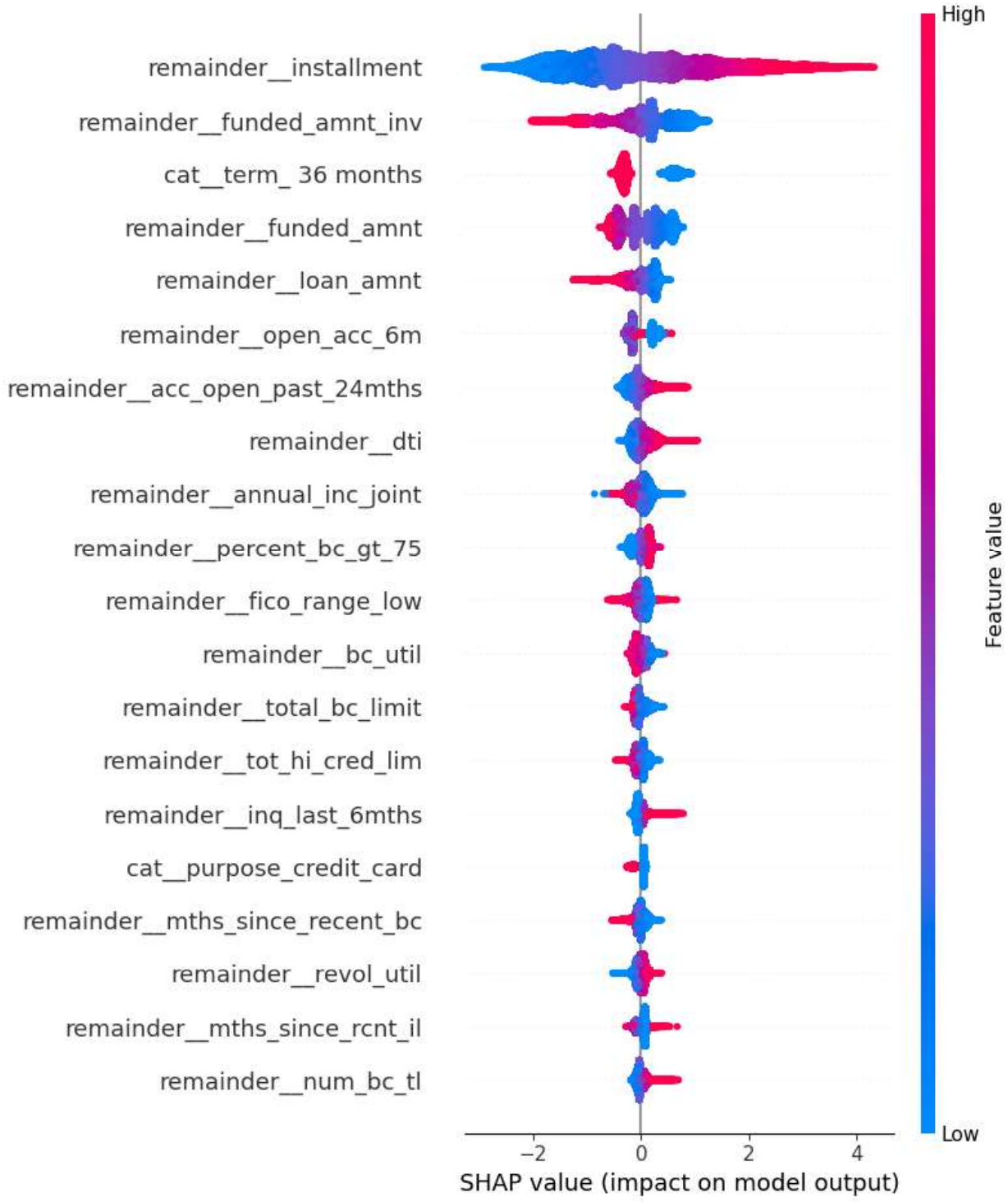

mths_since_recent_bc (iv:0.0318)
bad
good
Bin count distribution
Bad probability
36.0%, 322902
24.7%
22.6%
21.0%
17.7%, 159093
17.0%
14.8%, 132746
14.2%
8.2%, 73632
4.2%, 38091
[-inf,5.0)
[5.0,15.0)
[15.0,19.0)
[19.0,36.0)
[36.0,94.0)
[94.0,inf)

FIGURE 5: XGBOOST SHAP VALUES



Exposure at Default

$EAD = funded\_amnt - total\_rec\_prncp$

Loss Given Default

$LGD = 1 - (recoveries/EAD)$

Loan Profit:

Loan Profit for Non-Defaulters

$remaining\_terms = term - (last\_payment\_d - issue\_d)\ (in\ months)$

$expected\_int\_profit = int\_rate*out\_prncp*remaining\_terms/(12*100)$

$int\_revenue = total\_rec\_int + expected\_int\_profit$

$cost = 0.5*int\_revenue$

$loan\_profit = int\_revenue - cost + total\_rec\_late\_fee$

Loan Loss for Defaulters:

$loss = -1*mean\_LGD*EAD$

$Capital\ Requirements = PD*LGD*EAD$

New Variables:

Annual Income Joint:

$$\text{annual\_inc\_joint} = \begin{cases} annual\_inc\ ; & if\ annual\_inc\_joint\ = 0 \\ annual\_inc\_joint & \end{cases}$$

Installment Account:

$$\text{il\_acc} = \begin{cases} 0\ ; if\ installment\ acc\ present \\ 1; if\ installment\ acc\ not\ present \end{cases}$$

Collateral Coverage Ratio:

$collateral\_coverage\_ratio = total\_il\_high\_credit\_limit/loan\_amnt$

Bank Delinquency:

$$\text{bank\_dlq} = \begin{cases} 0; if\ never\ had\ dlq \\ 1; if\ record\ of\ dlq\ present \end{cases}$$