TechnicalReport

February 19, 2026

Author: Eva Song and Eton Tackett

1 Background

A local credit union provides 12-month loans to small businesses that often lack strong credit histories, with repayment collected as a fixed percentage of the business's monthly credit card sales, including both principal and interest. Because payments depend on fluctuating sales, repayment progress can vary significantly over time, creating uncertainty in borrowers' ability to repay their loans. To better monitor repayment and identify borrowers at risk of falling behind, the credit union developed a metric called PRSM, defined as twice the proportion of the total loan amount repaid after six months. Due to the variability and uncertainty in repayment, the goal is to predict the PRSM score using available borrower and business features to improve risk assessment and decision-making. In this project, we were asked to build a multiple regression model to accurately predict credit risk for a credit union lending loans to small businesses.

2 Data preprocessing

We begin with data preprocessing and exploratory data analysis (EDA) to ensure data quality and suitability for subsequent modeling. No missing values were identified in the dataset.

We applied several data cleaning steps based on domain constraints rather than statistical outlier detection. Specifically, FICO scores were restricted to the valid range [300, 800], and observations with negative values of `PRSM` or `Stress` were removed. PRSM is defined as a repayment ratio and is therefore non-negative by construction, while Stress is also a ratio for which negative values are not meaningful. These filtering steps ensure that all retained observations are consistent with the underlying definitions of the variables.

To improve feature representation of credit risk, we constructed a ratio variable capturing the proportion of delinquent credit lines:

$$\text{Prop\_Delinquent} = \frac{\text{Num\_Delinquent}}{\text{Num\_CreditLines}}$$

This ratio provides a more informative measure than the raw count alone, as it accounts for the total number of credit lines. For example, one delinquent line out of ten total lines represents a substantially lower risk profile than four delinquent lines out of five.

We assessed multicollinearity using variance inflation factors (VIF). As shown in Tables 1a and 1b, `Num_Delinquent` exhibits high collinearity with the constructed ratio variable and was there-

fore excluded from the model. After removal, all remaining predictors have VIF values below 3 (maximum VIF = 2.51), indicating no serious multicollinearity concerns.

For categorical variables, we applied one-hot encoding to `CorpStructure` and `NAICS_industry`. To reduce dimensionality and avoid sparse categories, NAICS codes were aggregated to the sector level using the first two digits. In addition, we discretized `FICO` scores into five ordinal categories (0–4), where higher values indicate better credit quality, following standard credit score ranges (Experian, 2023). This transformation allows for potential nonlinear effects while maintaining interpretability.

Several continuous predictors, including `TotalAmtOwed`, `Months`, and `Volume`, exhibited substantial right skewness (see Figure 1). To mitigate skewness and stabilize variance, we applied log transformations to these variables. However, since such transformations may not necessarily improve the fit of linear models, we retain both raw-scale and log-transformed specifications. Subsequent model comparison evaluates whether the transformation leads to improved predictive performance.

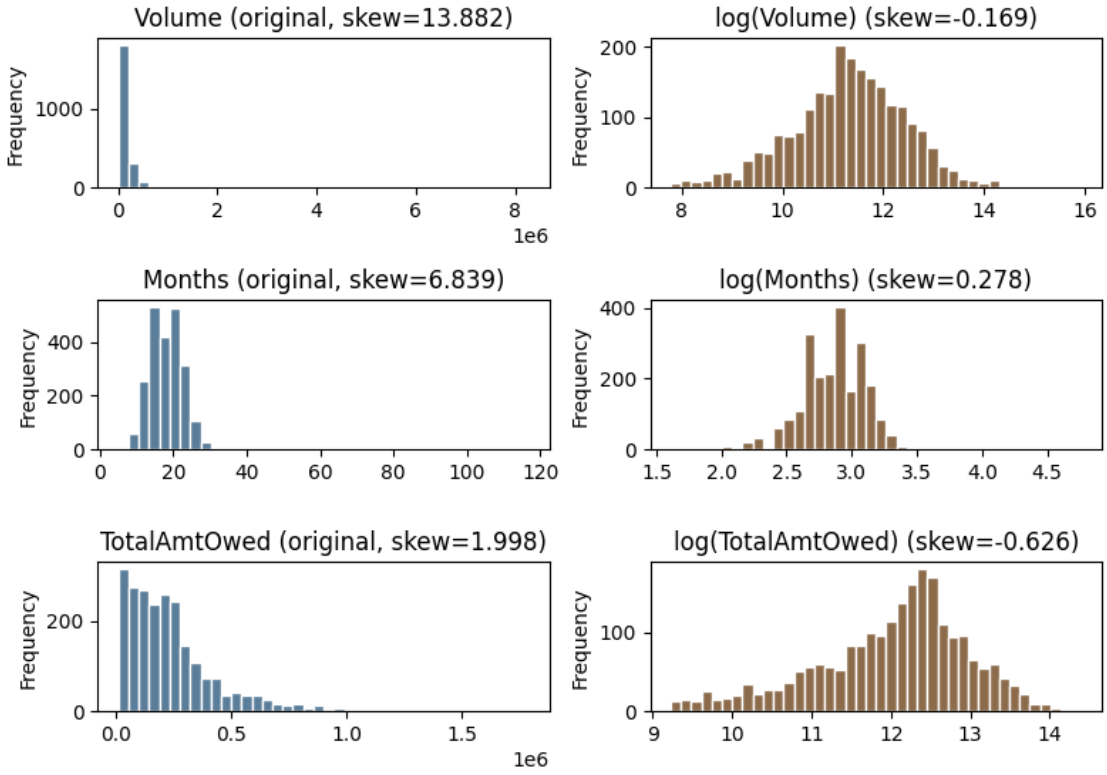
Table 1a. Variance Inflation Factor with Num_Delinquent

	Variable	VIF
5	Prop_Delinquent_Credit	44.630
3	Num_CreditLines	22.862
4	Num_Delinquent	18.078
1	Volume	1.358
0	TotalAmtOwed	1.224
2	Stress	1.153
6	Months	1.001

Table 1b. Variance Inflation Factor without Num_Delinquent

	Variable	VIF
4	Prop_Delinquent_Credit	2.511
3	Num_CreditLines	2.510
1	Volume	1.357
0	TotalAmtOwed	1.223
2	Stress	1.151
5	Months	1.001

Figure 1. Original vs log distributions for transform candidates



3 Modeling

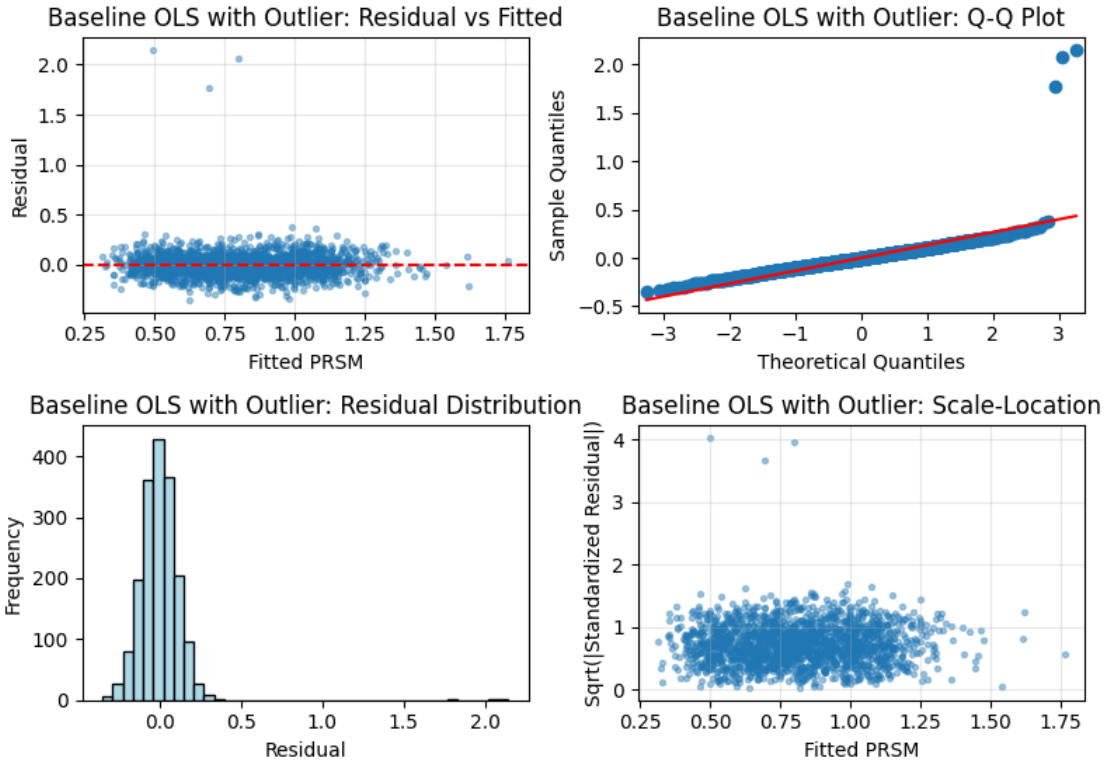
We split the dataset into training and validation sets using an 80:20 ratio. The training set is used for model fitting and variable selection, while the validation set is reserved for out-of-sample model comparison.

To evaluate modeling choices in a structured manner, we consider three regression methods—ordinary least squares (OLS), LASSO, and backward stepwise regression—under two feature specifications: (i) raw-scale predictors and (ii) log-transformed predictors for selected variables. This results in a 2×3 modeling design, allowing us to assess the impact of both model class and feature transformation on predictive performance.

As a baseline, we fit an OLS model using raw-scale predictors without log transformation. Diagnostic plots for this model (Figure 2) indicate the presence of observations with large standardized residuals.

Observations with standardized residuals exceeding 0.5 were identified as potential outliers and removed from the dataset. All subsequent models were refitted using the cleaned data. This procedure aims to reduce the influence of extreme observations on model estimation. However, we note that the threshold used for defining outliers is relatively conservative and primarily motivated by diagnostic inspection rather than a formal statistical rule.

Figure 2. Residual Diagnostics for Baseline OLS before Outlier Removal



The set of candidate predictors is fixed across all models to ensure comparability. For the raw-scale specification, the predictors include: `TotalAmtOwed`, `Volume`, `Months`, `Stress`, `Num_CreditLines`, `Prop_Delinquent_Credit`, `FICO_category`, `WomanOwned`, `CorpStructure_Corp`, `CorpStructure_LLC`, `CorpStructure_Partner`, and `NAICS_RetailTrade`. For the log-transformed specification, log transformations are applied to `TotalAmtOwed`, `Volume`, and `Months`, while all other variables remain unchanged. This parallel setup allows us to isolate the effect of transformation on model performance without confounding it with changes in the feature set.

Model performance is evaluated on the validation set using root mean squared error (RMSE) and prediction interval (PI) coverage. RMSE measures predictive accuracy, while PI coverage assesses the calibration of uncertainty estimates relative to the nominal 95% level.

As shown in Table 6, models using raw-scale predictors consistently outperform their log-transformed counterparts, achieving lower RMSE and coverage levels closer to the nominal target. This suggests that, in this dataset, log transformation does not improve predictive performance.

Among all models, the LASSO model without log transformation achieves the lowest RMSE (approximately 0.217) while maintaining prediction interval coverage near 95%, making it the selected model.

Table 2. Model comparison on validation set

	Model	Transformation	TrainR2	AdjR2	DevRMSE	DevMAE	PI Coverage
0	Baseline OLS	NoLog	0.8264	0.8253	0.2168	0.1029	94.9115
1	Baseline LASSO	NoLog	0.8263	0.8252	0.2166	0.1029	95.1327
2	Baseline Stepwise	NoLog	0.8264	0.8253	0.2168	0.1029	94.9115
3	OLS	Log	0.7843	0.7829	0.2224	0.1115	93.8053
4	LASSO	Log	0.7843	0.7829	0.2224	0.1116	93.8053
5	Stepwise	Log	0.7843	0.7829	0.2224	0.1115	93.8053

Compared to OLS and stepwise regression, LASSO applies regularization that shrinks coefficients toward zero, reducing model variance and improving generalization, particularly in the presence of correlated predictors. This is consistent with the observed performance gains.

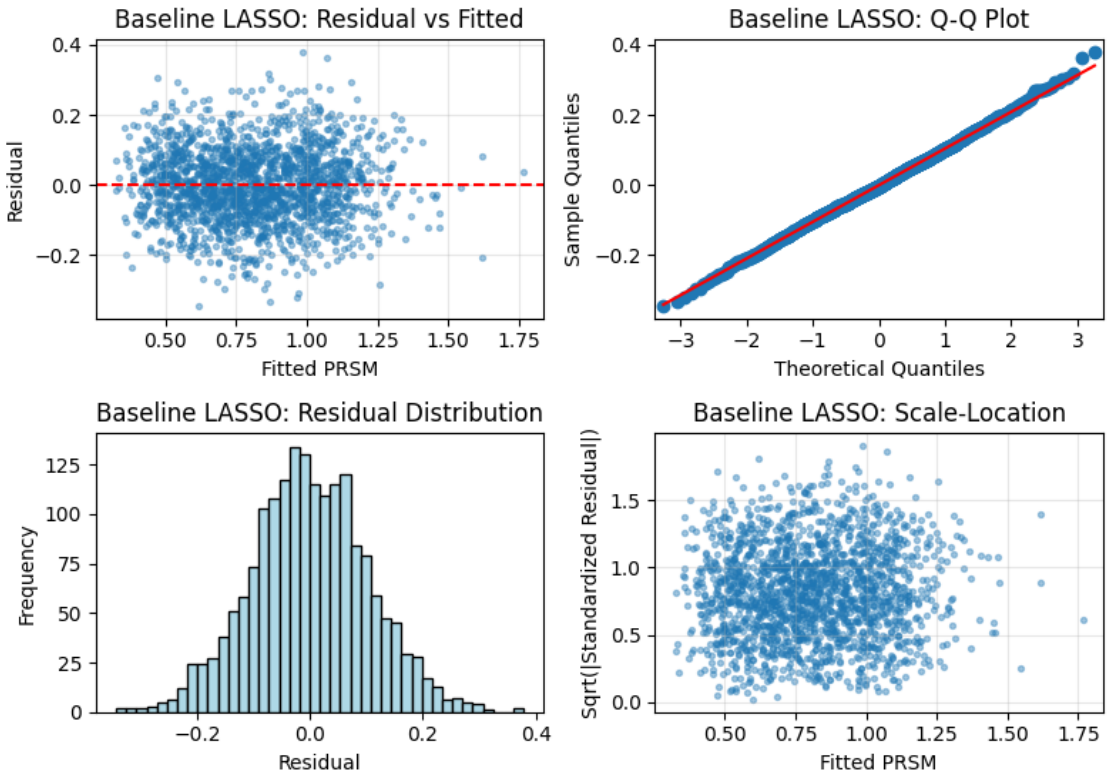
Overall, the results indicate that feature construction and preprocessing have a greater impact on predictive performance than the choice among the considered linear modeling approaches.

Table 3. Non-zero coefficients from Baseline LASSO Model

	Predictor	Coefficient	AbsCoefficient
6	WomanOwned	0.139	0.139
9	CorpStructure_LLC	0.104	0.104
1	TotalAmtOwed	0.100	0.100
10	CorpStructure_Partner	0.073	0.073
0	FICO_category	0.054	0.054
3	Stress	0.050	0.050
7	Months	0.013	0.013
8	CorpStructure_Corp	0.012	0.012
2	Volume	-0.003	0.003
11	NAICS_ind_Retail Trade	-0.002	0.002
5	Prop_Delinquent_Credit	-0.001	0.001
4	Num_CreditLines	0.000	0.000

Finally, diagnostic plots for the selected model (Figure 3) suggest that the main linear model assumptions—linearity, homoscedasticity, and approximate normality of residuals—are reasonably satisfied, supporting the validity of inference and prediction.

Figure 3. Residual Diagnostics for Baseline LASSO



4 Analysis of Results

To facilitate interpretation, we first define a baseline borrower profile (Table 4a) by fixing all predictors at representative values (means for continuous variables and reference levels for categorical variables). The corresponding predicted PRSM for this baseline profile is 0.547 (Table 4b), which serves as a reference point for evaluating marginal effects.

Table 5 reports predictors with practically important effects, defined as changes in PRSM exceeding a pre-specified practical-effect threshold. The results indicate that ownership and business structure are the dominant drivers. In particular, transitioning from non–woman-owned to woman-owned is associated with an increase of 0.279 in PRSM, while LLC and partnership structures are associated with increases of 0.243 and 0.172, respectively. These effect sizes are substantially larger than those of other predictors, suggesting that business characteristics play a central role in repayment performance.

Financial variables such as FICO category and total amount owed also exhibit meaningful effects. For example, a two-category increase in FICO score corresponds to an increase of 0.110 in PRSM, indicating improved repayment performance for borrowers with stronger credit profiles. Similarly, higher total amount owed and stress levels are associated with moderate changes in PRSM, although their magnitudes are smaller relative to ownership and structure variables.

Table 6 presents predictors that are statistically significant but have small practical impact. For instance, corporate structure (Corp) and months in operation have estimated effects of 0.028 and 0.013, respectively. While these effects are statistically detectable (p-values near zero), their magnitudes are negligible in practical terms and therefore unlikely to materially influence decision-making.

Notably, the effect of months in operation is relatively small, suggesting that business tenure alone is not a strong predictor of repayment performance in this dataset. This may reflect the heterogeneity of small businesses, where longevity does not necessarily imply financial stability.

Overall, these results highlight the distinction between statistical significance and practical relevance. Variables such as business ownership and organizational structure have substantial influence on predicted PRSM, whereas other statistically significant variables contribute little to prediction in practical terms.

Table 4a. Baseline borrower profile

	Predictor	BaselineValue
0	FICO_category	2.000
1	TotalAmtOwed	194674.000
2	Volume	84626.000
3	Stress	0.190
4	Num_CreditLines	10.000
5	Prop_Delinquent_Credit	0.400
6	WomanOwned	0.000
7	Months	18.000
8	CorpStructure_Corp	0.000
9	CorpStructure_LLC	0.000
10	CorpStructure_Partner	0.000
11	NAICS_ind_Retail Trade	1.000

Table 4b. Baseline prediction

	Metric	Value
0	BaselinePredictedPRSM	0.547

Table 5. Practically important drivers

	Predictor	ScenarioChange	EstimatedDeltaPRSM	PValue	StatSig5pct
6	WomanOwned	0 to 1	0.279	0.000	True
9	CorpStructure_LLC	0 to 1	0.243	0.000	True
10	CorpStructure_Partner	0 to 1	0.172	0.000	True
0	FICO_category	+2.0	0.110	0.000	True
1	TotalAmtOwed	+200985.0	0.100	0.000	True
3	Stress	+0.15	0.070	0.000	True

Table 6. Statistically detectable but practically small effects

	Predictor	ScenarioChange	EstimatedDeltaPRSM	PValue
8	CorpStructure_Corp	0 to 1	0.028	0.000
7	Months	+6.0	0.013	0.000