TechnicalReport

February 18, 2026

1 Technical Report

Author: Eva Song and Eton Tackett

2 Data Preprocessing

To start our analysis, we first started pre-processing our data and applying the exploratory data analysis (EDA) process.

2.1 Data Snapshot

Table 1 summarizes variable types, ranges, and missingness for the training data; no missing values are present.

Table 1

	variable	dtype	missing	unique	min	max	sample_categories
0	PRSM	float64	0.000	2263.000	-0.975	2.977	
1	FICO	int64	0.000	299.000	482.000	850.000	
2	TotalAmtOwed	int64	0.000	2259.000	10136.000	1791524.000	
3	Volume	int64	0.000	2256.000	2393.000	8284497.000	
4	Stress	float64	0.000	2263.000	0.005	0.704	
5	Num_Delinquent	int64	0.000	6.000	3.000	8.000	
6	Num_CreditLines	int64	0.000	6.000	8.000	13.000	
7	WomanOwned	int64	0.000	2.000	0.000	1.000	
8	CorpStructure	object	0.000	4.000			LLC, Corp, Partner, Sole
9	NAICS	int64	0.000	21.000	441120.000	722514.000	
10	Months	int64	0.000	38.000	5.000	117.000	

9
10

2.2 Data Quality and Outlier Screening

We enforce basic validity rules before modeling. Table 1A shows four rows with PRSM < 0 removed; all other checks are zero. Figure 1 highlights right-tail outliers in TotalAmtOwed and Volume, motivating log transformation in the preprocessed track.

Table 1A

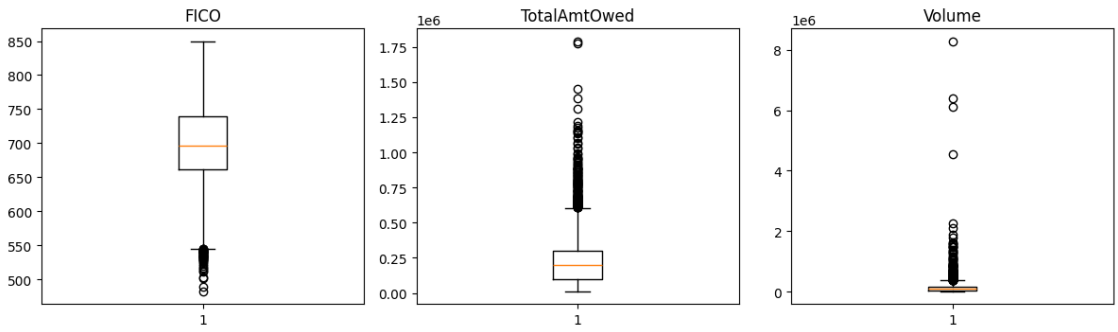
	Check	Count
0	FICO outside [300, 850]	0.000
1	PRSM below 0	4.000
2	Stress below 0	0.000
3	Num_Delinquent > Num_CreditLines	0.000
4	Rows removed due to invalid PRSM	4.000

Table 1A. PRSM<0 rows

	PRSM	FICO	TotalAmtOwed	Volume	Stress	Num_Delinquent
620	-0.813	698.000	51311.000	13042.000	0.328	4.000
931	-0.975	735.000	285143.000	260634.000	0.091	4.000
1513	-0.839	688.000	377793.000	165021.000	0.191	4.000
1741	-0.023	679.000	38315.000	16011.000	0.199	4.000

	Num_CreditLines	WomanOwned	CorpStructure	NAICS	Months
620	10.000	1.000	Corp	459210.000	15.000
931	12.000	0.000	Corp	722330.000	20.000
1513	11.000	1.000	Sole	722330.000	18.000
1741	9.000	0.000	Sole	445240.000	30.000

Figure 1. Outlier scan by variable



2.3 Feature Engineering

We create interpretable risk features. Table 1B confirms NAICS mapping coverage for train and evaluation sets (three unique 2-digit codes, zero missing). Prop_Delinquent_Credit, FICO_category, and NAICS_industry are retained for modeling.

Table 1B

	Dataset	UniqueNAICS2digitCount	MissingIndustryMappings
0	Training	3.000	0.000
1	Evaluation	3.000	0.000

2.4 Multicollinearity Diagnostics

Table 2a reports VIF values (maximum 2.511), indicating no multicollinearity concern. Table 2b shows low pairwise correlations, with the highest magnitude -0.775 between Prop_Delinquent_Credit and Num_CreditLines, acceptable for OLS.

Table 2a. Variance Inflation Factor

	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

Table 2b. Numeric predictor correlations

	TotalAmtOwed	Volume	Stress	Num_CreditLines	\
TotalAmtOwed	1.000	0.392	0.031	-0.008	
Volume	0.392	1.000	-0.318	0.003	
Stress	0.031	-0.318	1.000	-0.025	
Num_CreditLines	-0.008	0.003	-0.025	1.000	
Prop_Delinquent_Credit	-0.024	-0.015	0.022	-0.775	
Months	0.012	0.004	0.028	-0.013	

	Prop_Delinquent_Credit	Months
TotalAmtOwed	-0.024	0.012
Volume	-0.015	0.004
Stress	0.022	0.028
Num_CreditLines	-0.775	-0.013
Prop_Delinquent_Credit	1.000	0.007
Months	0.007	1.000

2.5 Distribution Check and Transform Plan

Table 3 shows strong right skew in TotalAmtOwed and Volume; Figure 2 confirms log transforms tighten their spread. We keep a baseline (raw-scale) track and a preprocessed (log-scale) track for comparison.

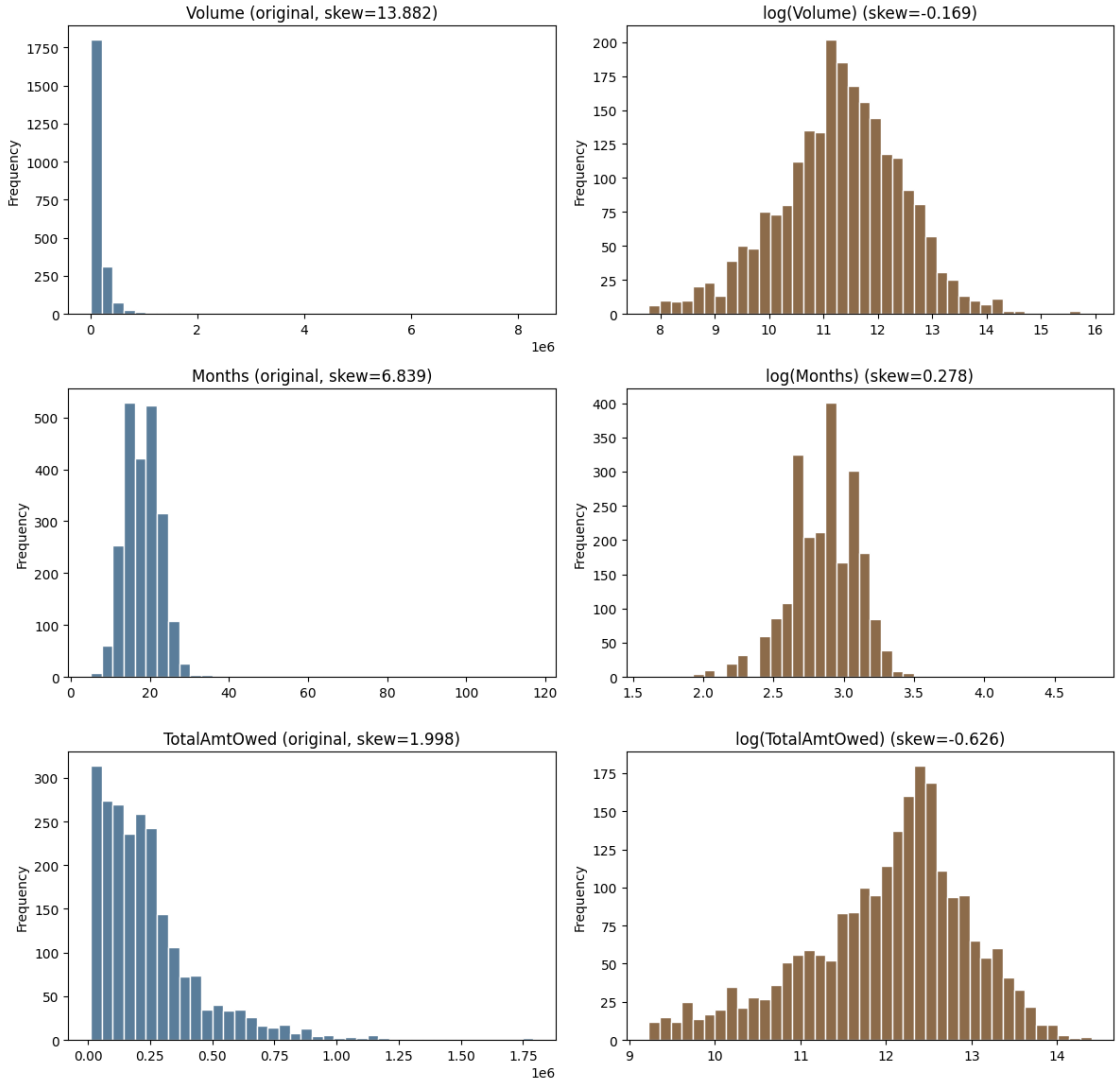
Table 3. Distribution summary for continuous predictors

	Variable	Skewness	Min	Max	Median	\
0	TotalAmtOwed	1.998	10136.000	1791524.000	196365.000	
1	Volume	13.882	2393.000	8284497.000	84414.000	
2	Stress	0.706	0.005	0.704	0.186	

3	Num_CreditLines	0.026	8.000	13.000	10.000
4	Prop_Delinquent_Credit	0.847	0.231	1.000	0.400
5	Months	6.839	5.000	117.000	18.000

	Mean
0	238990.766
1	157069.589
2	0.199
3	10.223
4	0.406
5	18.209

Figure 2. Original vs log distributions for transform candidates



2.6 Train/Validation Split

Table 4 records the split: 1,807 train rows, 452 dev rows, 2,500 evaluation rows, ensuring an unbiased dev set for tuning.

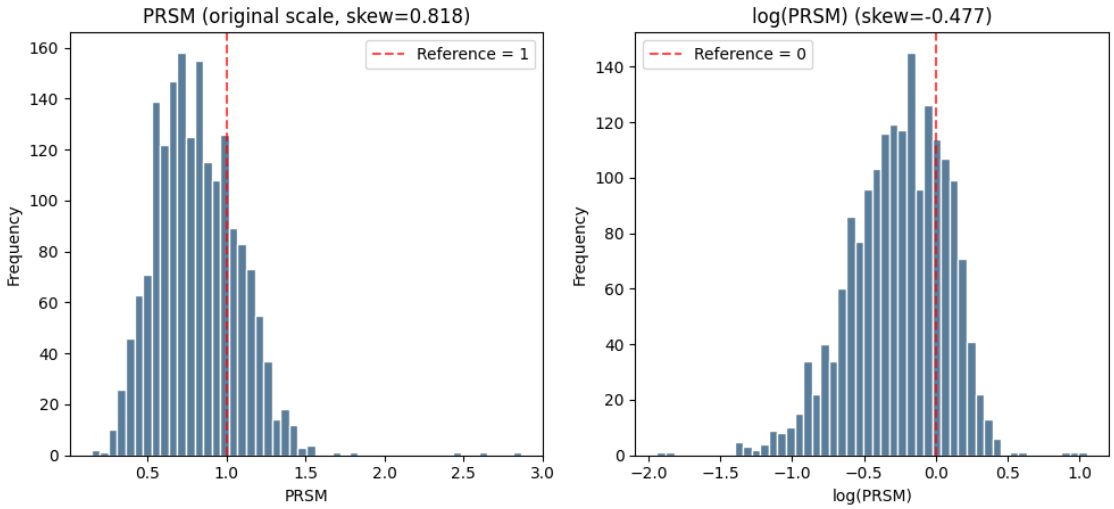
Table 4

	Dataset	Rows
0	Train	1807.000
1	Dev	452.000
2	Evaluation	2500.000

2.7 Response Distribution Check

Figure 3 shows raw and log PRSM distributions. PRSM is moderately right-skewed; we keep it on the original scale and rely on residual diagnostics after model fitting.

Figure 3. PRSM distribution on original and log scales

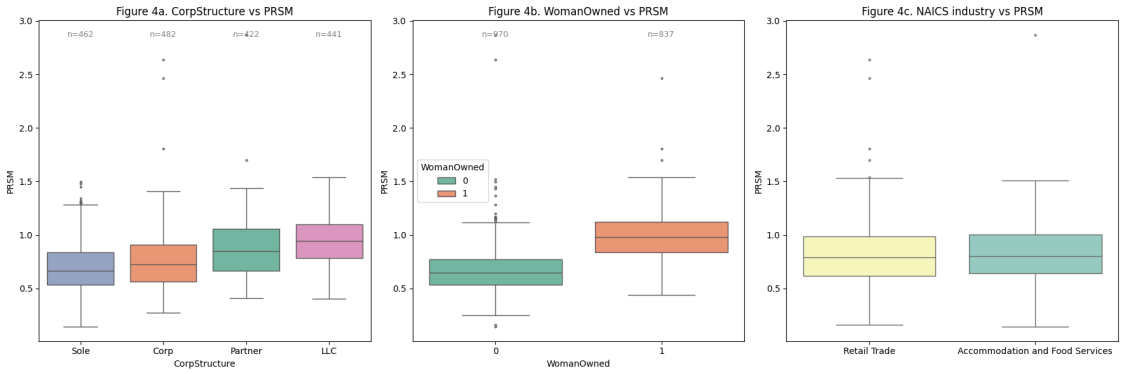


2.8 Categorical Predictors vs PRSM

Figure 4 compares PRSM across CorpStructure, WomanOwned, and NAICS_industry. Table A1 shows Retail (1,407 rows) and Accommodation/Food (400 rows) provide adequate support for dummies.

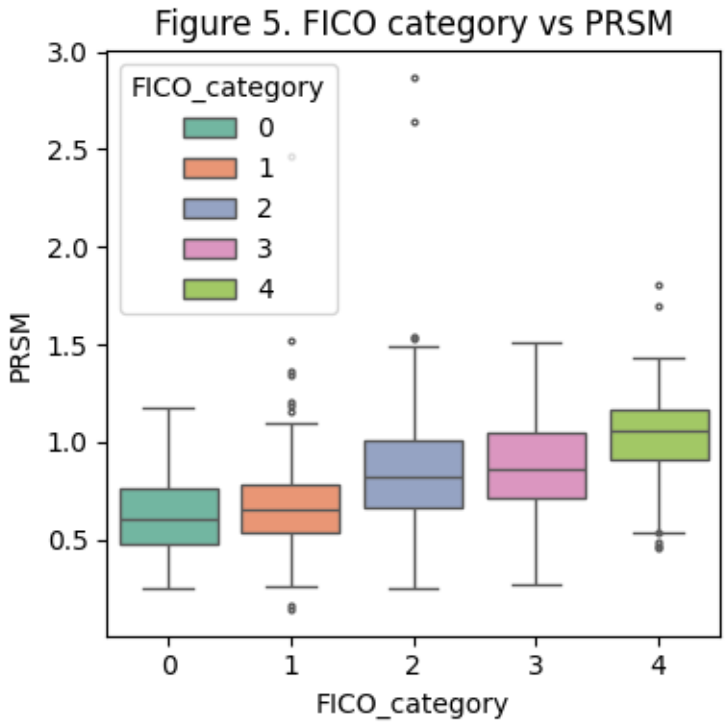
Table A1. Industry sample sizes used in Figure 4c

	NAICS_industry	count
0	Retail Trade	1407.000
1	Accommodation and Food Services	400.000



2.9 FICO Category vs PRSM

Figure 5 confirms higher FICO categories correspond to lower median PRSM, supporting categorical FICO representation.



2.10 Data Preparation Summary

Table 5 lists the final modeling design, including response, transforms, categorical handling, and dropped fields.

3 Data Export for Modeling

Processed train/dev/test datasets are saved for reproducible downstream modeling.

Table 5A

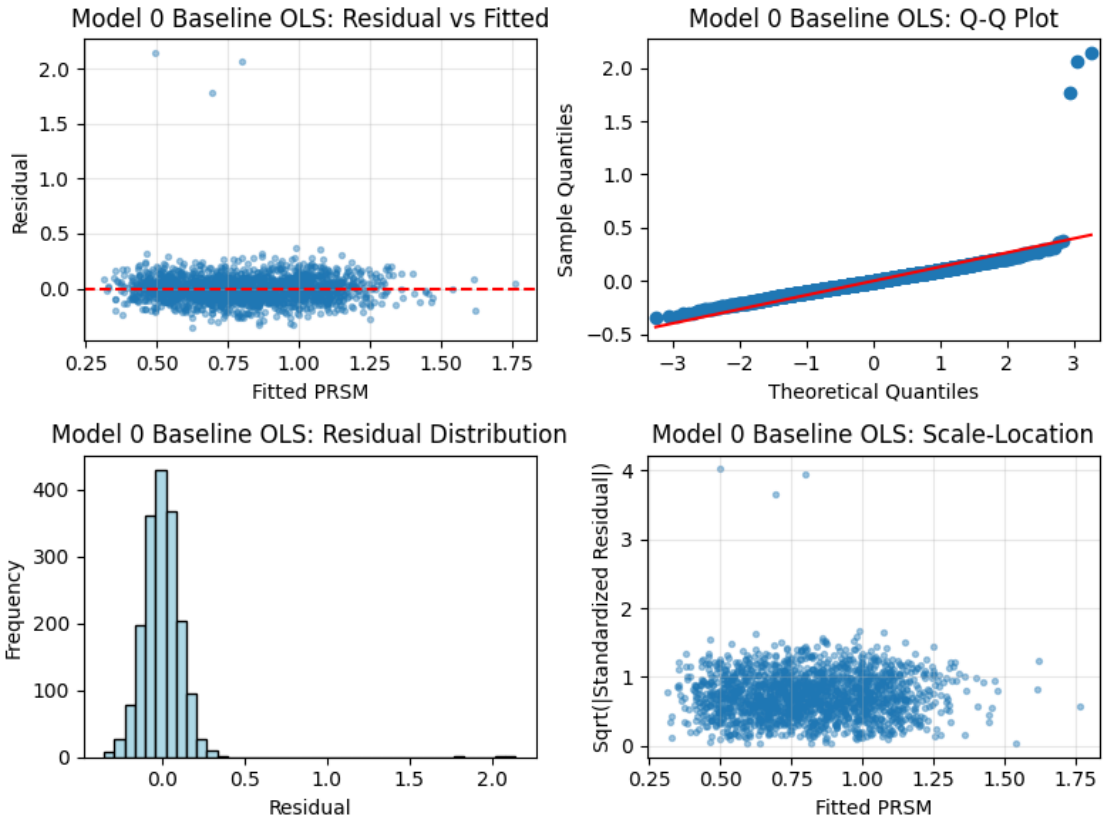
	Dataset	Shape	ColumnCount
0	Train	(1807, 45)	45.000
1	Dev	(452, 45)	45.000

4 Model Development

We evaluate six models: three baseline (no log) and three preprocessed (log). Each model block reports training fit, diagnostics, and dev-set accuracy; interpretations are summarized in markdown and in the final comparison.

Table M0A. Model 0 training metrics

	Model	TrainR2	TrainAdjR2	ParamCount
0	Model 0 Baseline OLS, raw predictors	0.743	0.741	12.000



4.0.1 Baseline OLS Review and Outlier Trimming

```
PRSM ~ FICO_category + TotalAmtOwed + Volume + Stress + Num_CreditLines + Prop_Delinquent_Credi
```

Diagnostics flagged large residuals; trimming |residual| > 0.5 (Table M0B/C) leaves 1,804 rows. Trimmed Model 0 achieves dev RMSE 0.217 and PI95 coverage 94.9% (Table M0E), and serves as the no-log benchmark. Residual plots show homoscedasticity improves after trimming.

Table M0B. Outlier screening summary

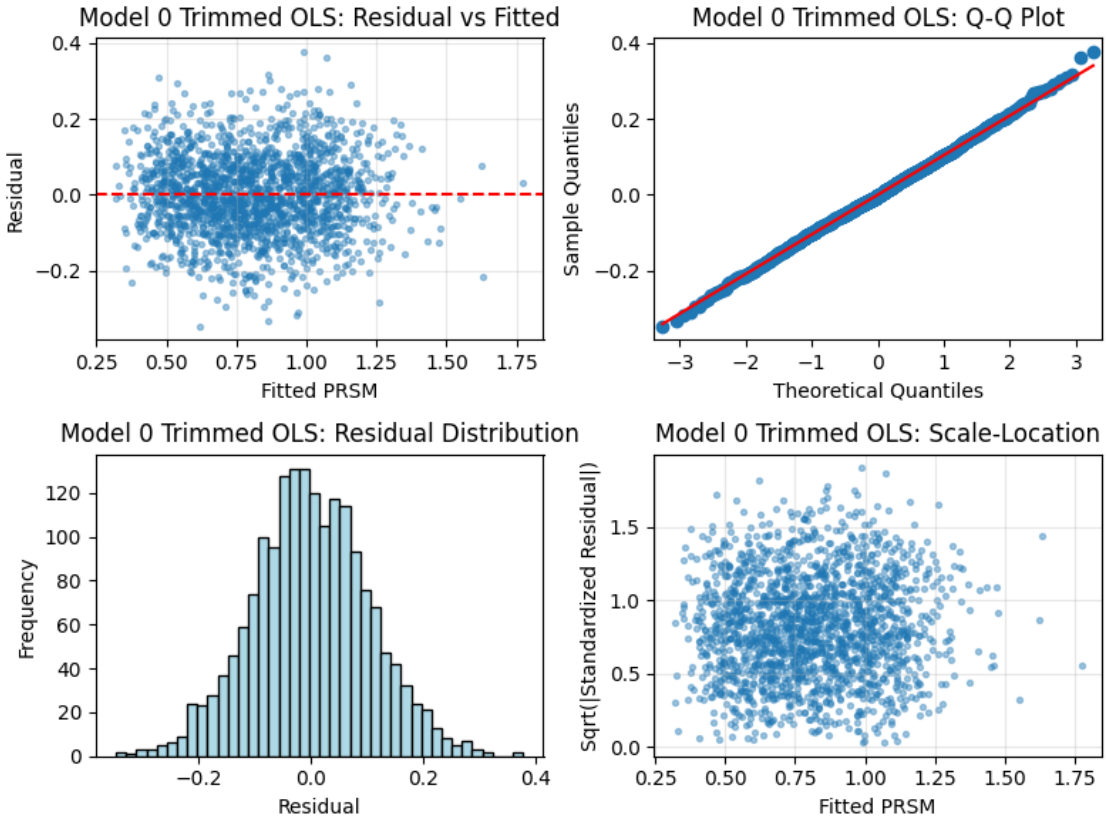
	Metric	Value
0	ResidualThreshold	0.500
1	OutlierCount	3.000
2	TrainRowsBeforeTrim	1807.000

Table M0C. Largest residual outliers (top 20)

	Index	Actual_PRSM	Predicted_PRSM	Residual	Abs_Residual
0	49.000	2.640	0.497	2.143	2.143
1	866.000	2.867	0.798	2.069	2.069
2	1152.000	2.467	0.693	1.774	1.774

Table M0D. Model 0 trimmed training metrics

	Model	TrainR2	TrainAdjR2	ParamCount
0	Model 0 Trimmed OLS	0.826	0.825	12.000



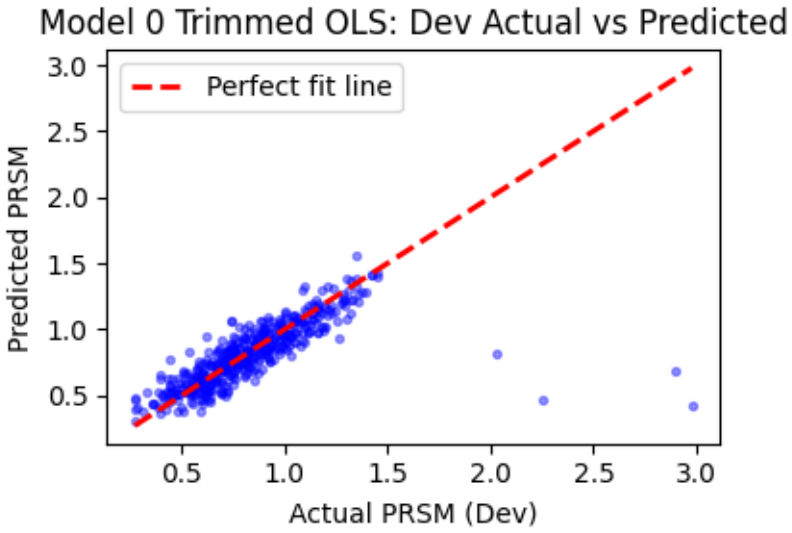


Table M0E. Model 0 dev metrics

	Model	DevRMSE	DevMAE	PI95Coverage	IntervalMethod
0	Model 0 Trimmed OLS	0.217	0.103	0.949	OLS observation interval

4.1 Model 1: Baseline LASSO

```
PRSM ~ FICO_category + TotalAmtOwed + Volume + Stress + Num_CreditLines + Prop_Delinquent_Credi
```

L1 shrinkage on raw predictors retains 12 non-zero terms (Table M1B). Dev RMSE 0.217 and PI95 coverage 95.1% (Table M1C) match Model 0; no gain in error but coefficients highlight dominant dummies (WomanOwned, CorpStructure). Use Model 1 when a sparse linear form is desired without log transforms.

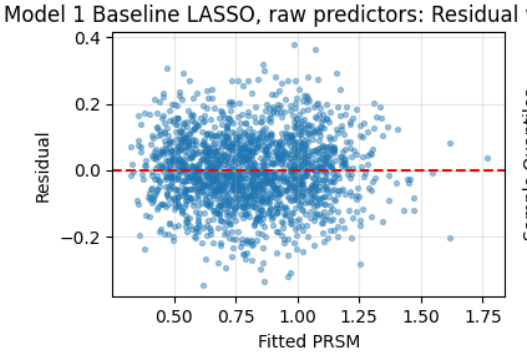
Table M1A. Model 1 training metrics

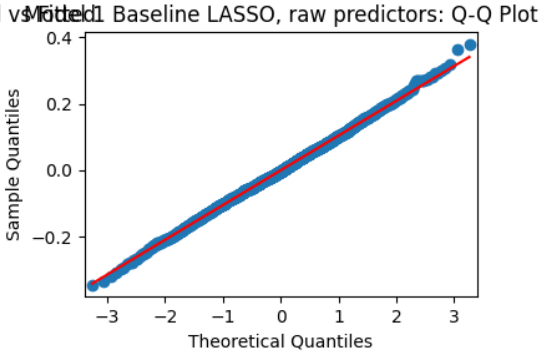
	Model	TrainR2	TrainAdjR2	ParamCount	OptimalAlpha
0	Model 1 Baseline LASSO, raw predictors	0.826	0.825	12.000	0.001

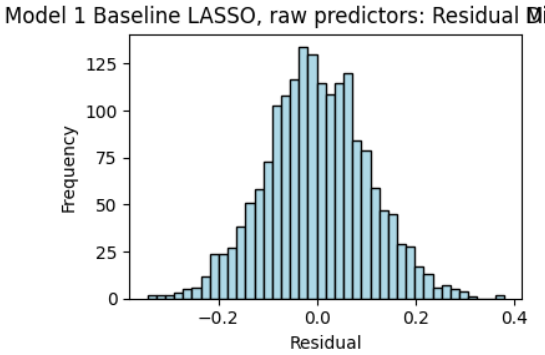
Table M1B. Model 1 non-zero coefficients

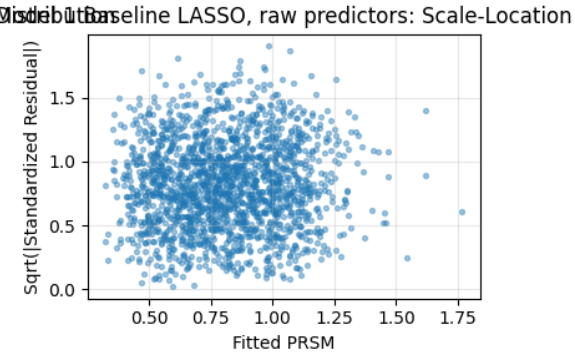
	Predictor	Coefficient	AbsCoefficient
6	WomanOwned	0.139	0.139
9	CorpStructure_LLC	0.104	0.104
1	TotalAmtOwed	0.100	0.100
10	CorpStructure_Partner	0.073	0.073
0	FICO_category	0.054	0.054
3	Stress	0.050	0.050

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









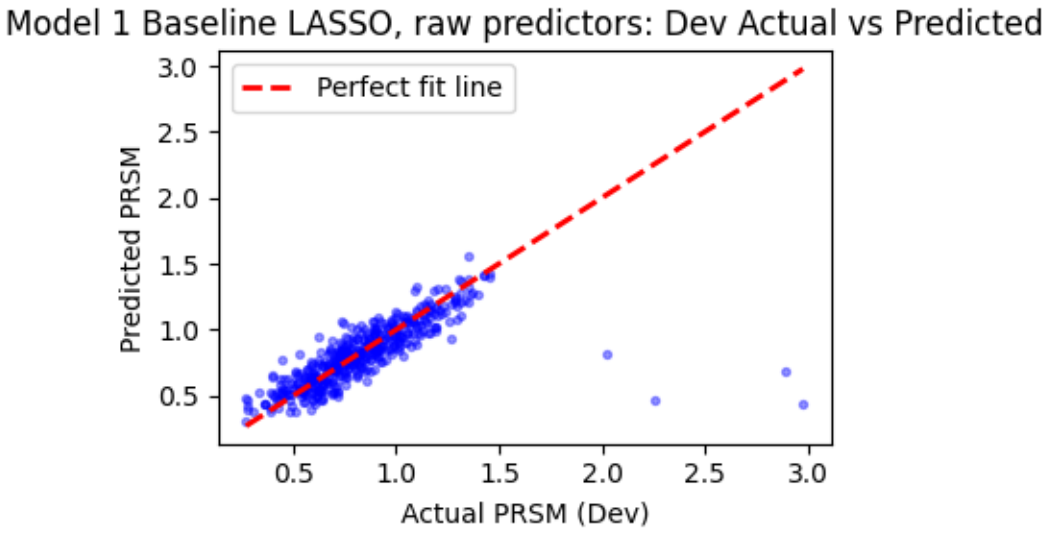


Table M1C. Model 1 dev metrics

```
                                    Model DevRMSE DevMAE PI95Coverage  \
0  Model 1 Baseline LASSO, raw predictors   0.217  0.103        0.951

               IntervalMethod
0  Residual-quantile interval
```

4.2 Model 2: Baseline Stepwise OLS

```
PRSM ~ FICO_category + TotalAmtOwed + Volume + Stress + Num_CreditLines + Prop_Delinquent_Credi
```

CV-based backward selection removed no predictors (Table M2B/C); dev RMSE 0.217 (Table M2D) equals Models 0?1. Conclusion: stepwise adds complexity without parsimony benefit on raw predictors.

Table M2A. Model 2 training metrics

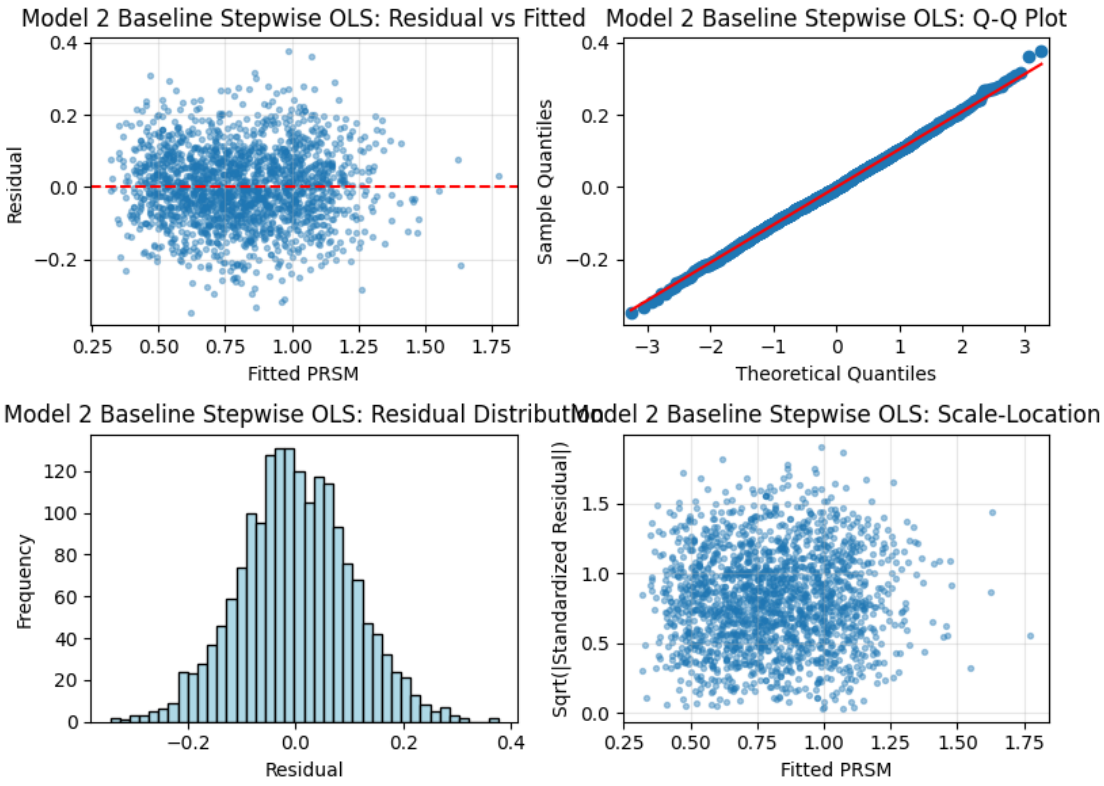
```
                          Model TrainR2 TrainAdjR2 ParamCount
0  Model 2 Baseline Stepwise OLS   0.826      0.825     12.000
```

Table M2B. Model 2 stepwise CV summary

```
  InitialPredictorCount FinalPredictorCount InitialCVRMSE FinalCVRMSE  \
0                12.000              12.000         0.105       0.105

  RemovedPredictors
0              None
```

Table M2C. Model 2 retained coefficients

```
                 predictor    coef
0            FICO_category   0.056
1             TotalAmtOwed   0.000
2                   Volume  -0.000
3                   Stress   0.471
4          Num_CreditLines   0.000
5   Prop_Delinquent_Credit  -0.011
6               WomanOwned   0.279
7                   Months   0.002
8       CorpStructure_Corp   0.032
9        CorpStructure_LLC   0.246
10   CorpStructure_Partner   0.175
11  NAICS_ind_Retail Trade  -0.006
```



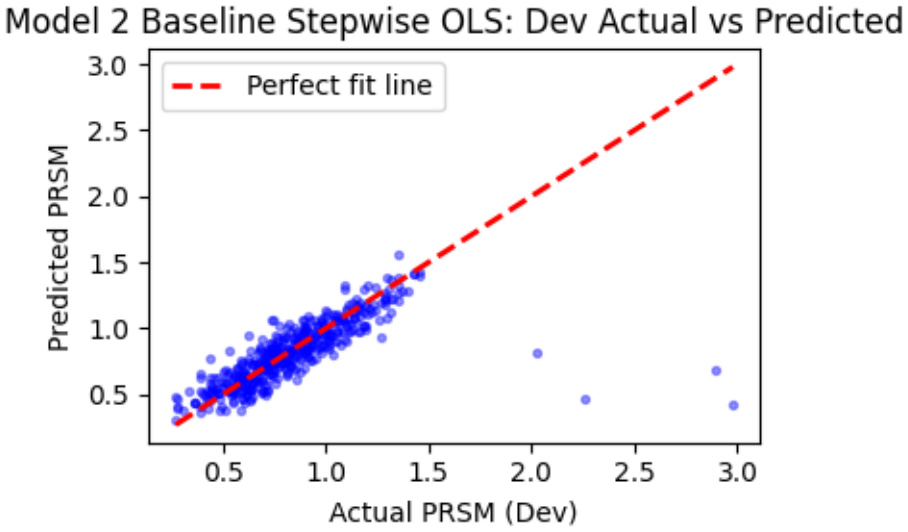


Table M2D. Model 2 dev metrics

	Model	DevRMSE	DevMAE	PI95Coverage
0	Model 2 Baseline Stepwise OLS	0.217	0.103	0.949

```
              IntervalMethod
0  OLS observation interval
```

4.3 Model 3: Preprocessed OLS

```
PRSM ~ FICO_category + log_TotalAmtOwed + log_Volume + Stress + log_Months +
        Num_CreditLines + Prop_Delinquent_Credit  + WomanOwned +
       CorpStructure_Corp + CorpStructure_LLC + CorpStructure_Partner + NAICS_Retail Trade
```

Using log-transformed scale variables, trimmed Model 3 reaches dev RMSE 0.222 (Table M3E), worse than baseline track. Residual plots improve shape but error rises; logs alone do not outperform raw-scale OLS here.

Table M3A. Model 3 training metrics

```
                                      Model  TrainR2  TrainAdjR2  ParamCount
0  Model 3 Preprocessed OLS, log predictors    0.700       0.698      12.000
```

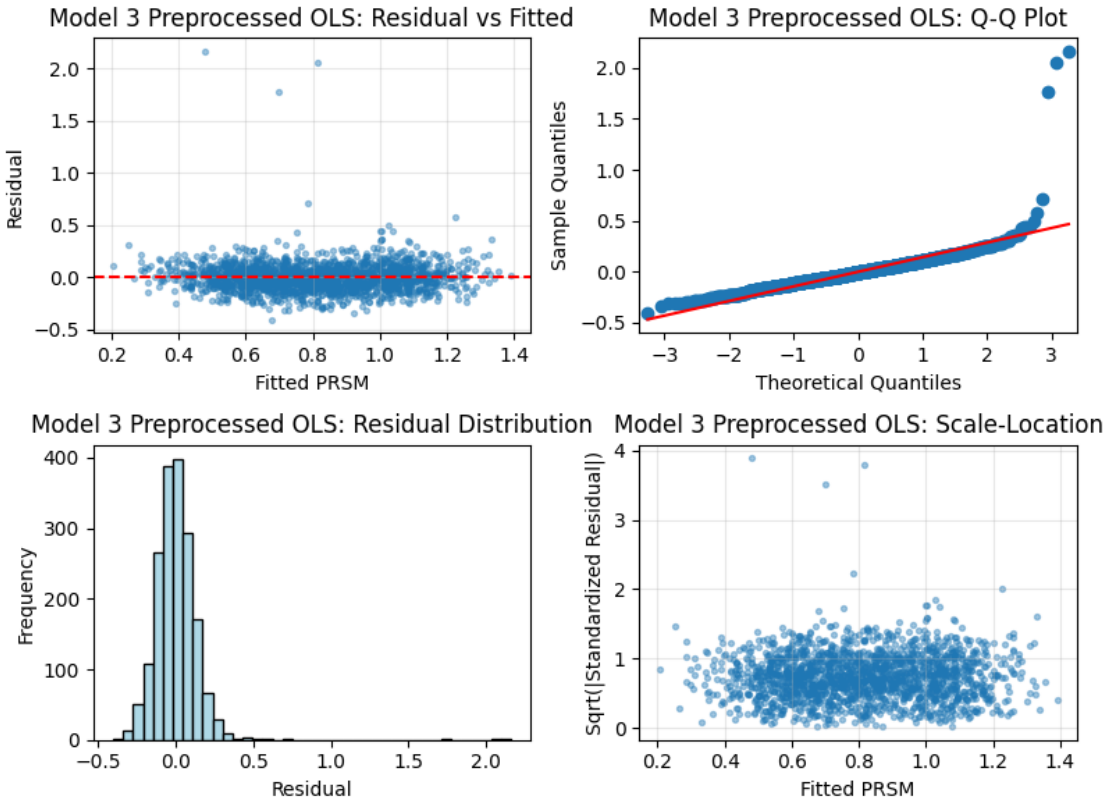


Table M3B

```
                Metric     Value
0    ResidualThreshold     0.500
1         OutlierCount     5.000
```

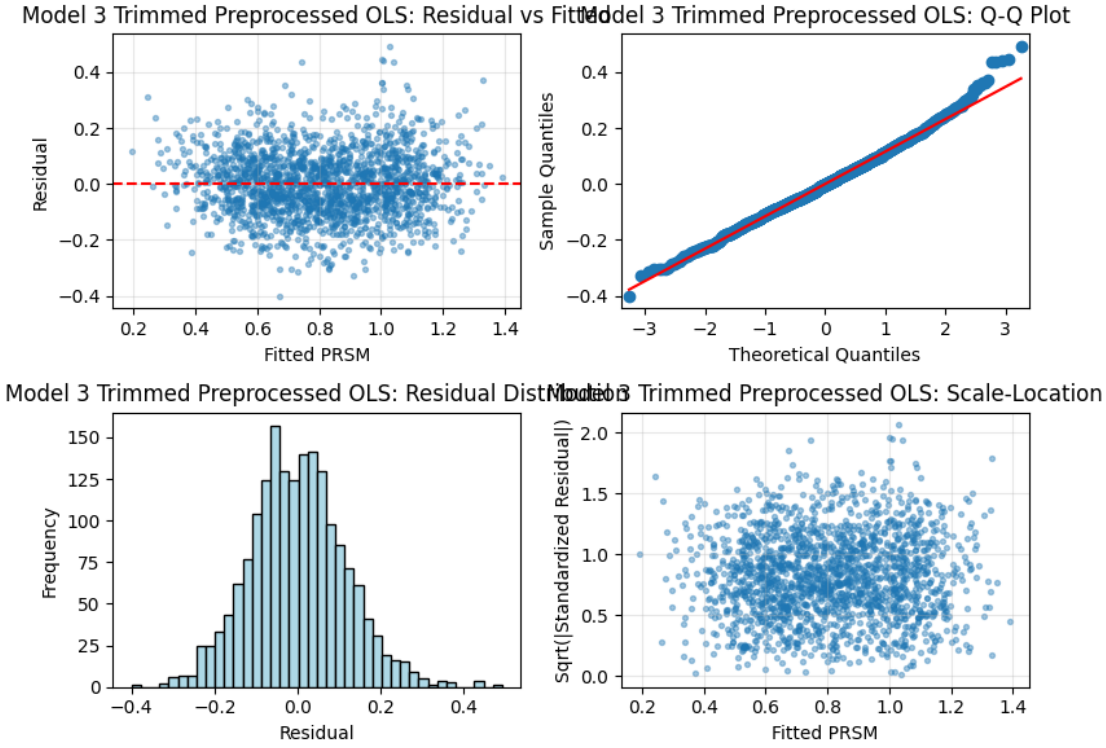
2	TrainRowsBeforeTrim	1807.000

Table M3C

	Index	Actual_PRSM	Predicted_PRSM	Residual	Abs_Residual
0	49.000	2.640	0.479	2.161	2.161
2	866.000	2.867	0.816	2.051	2.051
4	1152.000	2.467	0.699	1.768	1.768
3	952.000	1.495	0.784	0.711	0.711
1	134.000	1.806	1.225	0.580	0.580

Table M3D

	Model	TrainR2	TrainAdjR2	ParamCount
0	Model 3 Trimmed Preprocessed OLS	0.784	0.783	12.000



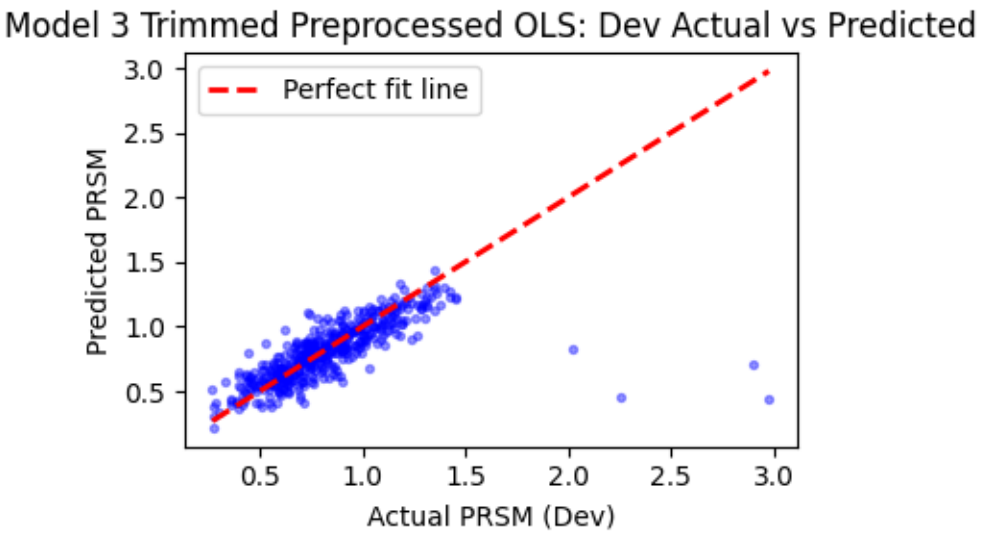


Table M3E

```
                                Model DevRMSE DevMAE PI95Coverage  \
0  Model 3 Trimmed Preprocessed OLS   0.222  0.112        0.938

             IntervalMethod
0  OLS observation interval
```

4.4 Model 4: Preprocessed Stepwise OLS

```
PRSM ~ FICO_category + log_TotalAmtOwed + log_Volume + Stress + log_Months +
        Num_CreditLines + Prop_Delinquent_Credit  + WomanOwned +
       CorpStructure_Corp + CorpStructure_LLC + CorpStructure_Partner + NAICS_Retail Trade
```

Stepwise on log predictors removes nothing (Table M4B/C); dev RMSE 0.222 (Table M4D) matches Model 3 and remains above baseline. No performance win relative to simpler Model 3.

Table M4A

```
                               Model TrainR2 TrainAdjR2 ParamCount
0  Model 4 Preprocessed Stepwise OLS   0.784      0.783     12.000
```

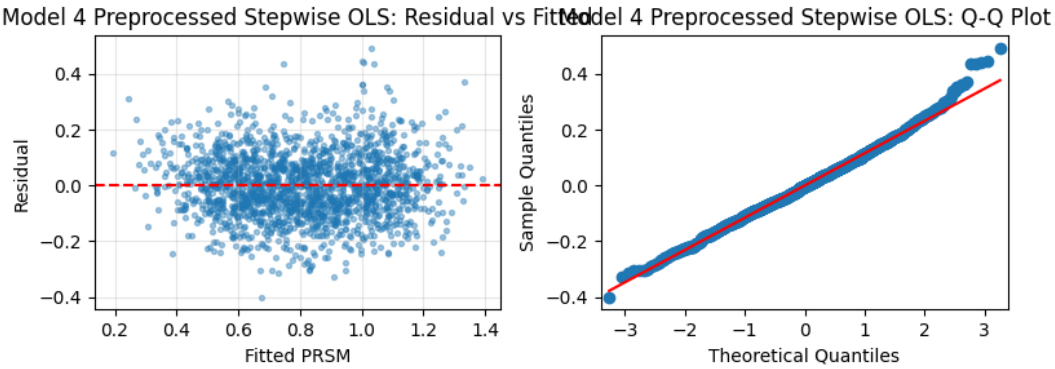
Table M4B

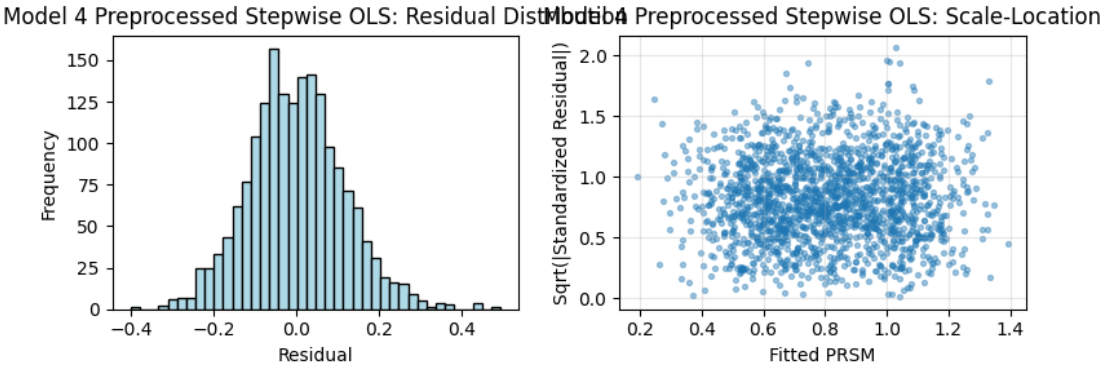
```
  InitialPredictorCount FinalPredictorCount InitialCVRMSE FinalCVRMSE  \
0                12.000              12.000         0.117       0.117

  RemovedPredictors
0              None
```

Table M4C

```
                predictor    coef
```

0	FICO_category	0.056
1	log_TotalAmtOwed	0.104
2	log_Volume	-0.018
3	Stress	0.388
4	log_Months	0.077
5	Num_CreditLines	-0.001
6	Prop_Delinquent_Credit	-0.028
7	WomanOwned	0.278
8	CorpStructure_Corp	0.033
9	CorpStructure_LLC	0.245
10	CorpStructure_Partner	0.170
11	NAICS_ind_Retail Trade	-0.004





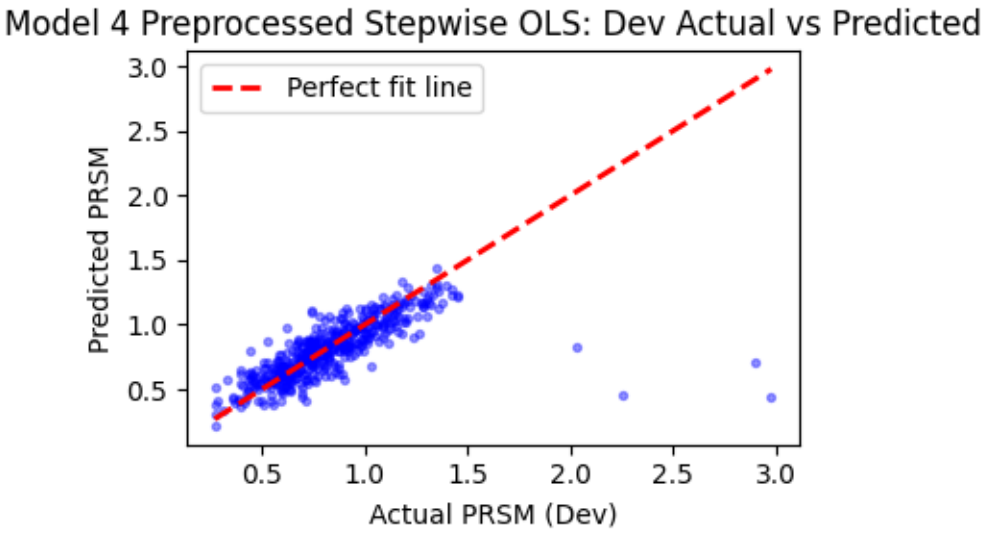


Table M4D

```
                                Model DevRMSE DevMAE PI95Coverage  \
0  Model 4 Preprocessed Stepwise OLS   0.222  0.112        0.938

             IntervalMethod
0  OLS observation interval
```

4.5 Model 5: Preprocessed LASSO

```
PRSM ~ FICO_category + log_TotalAmtOwed + log_Volume + Stress + log_Months +
        Num_CreditLines + Prop_Delinquent_Credit  + WomanOwned +
       CorpStructure_Corp + CorpStructure_LLC + CorpStructure_Partner + NAICS_Retail Trade
```

LASSO on log predictors keeps key dummies (Table M5B) but dev RMSE 0.222 (Table M5C) still lags baseline track. Conclusion: among log-track models, shrinkage does not close the gap to baseline OLS.

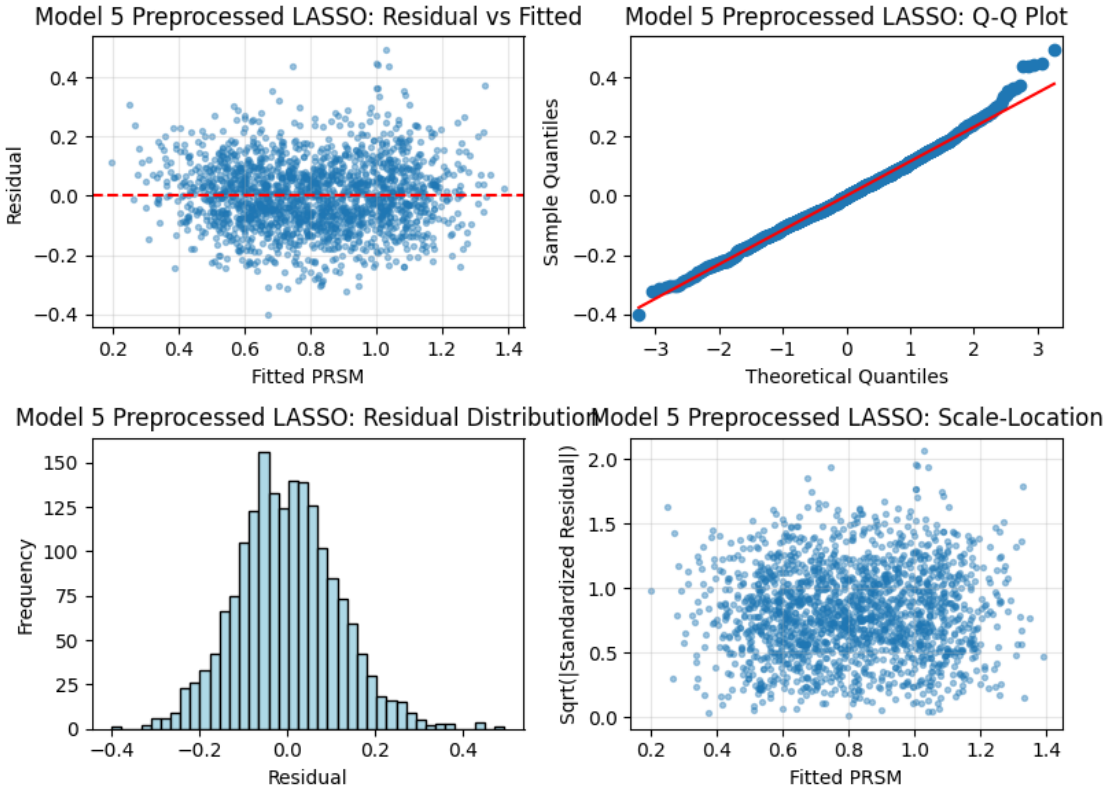
Table M5A

```
                                     Model TrainR2 TrainAdjR2 ParamCount  \
0  Model 5 Preprocessed LASSO, log predictors   0.784      0.783     12.000

  OptimalAlpha
0        0.000
```

Table M5B

```
                Predictor Coefficient AbsCoefficient
7              WomanOwned       0.139          0.139
9       CorpStructure_LLC       0.105          0.105
1        log_TotalAmtOwed       0.095          0.095
```

10	CorpStructure_Partner	0.071	0.071
0	FICO_category	0.055	0.055
3	Stress	0.044	0.044
4	log_Months	0.020	0.020
2	log_Volume	-0.017	0.017
8	CorpStructure_Corp	0.014	0.014
11	NAICS_ind_Retail Trade	-0.002	0.002
6	Prop_Delinquent_Credit	-0.001	0.001
5	Num_CreditLines	-0.000	0.000



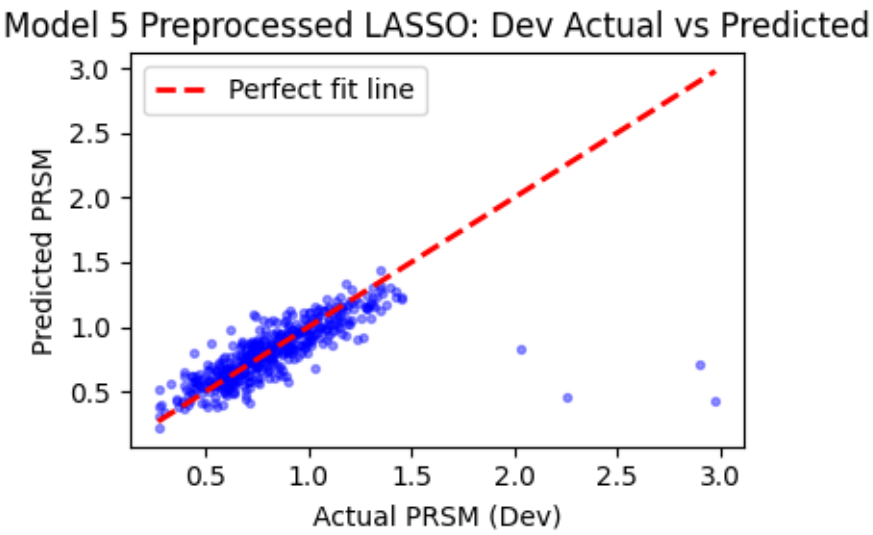


Table M5C

```
                        Model DevRMSE DevMAE PI95Coverage  \
0  Model 5 Preprocessed LASSO   0.222  0.112        0.938

               IntervalMethod
0  Residual-quantile interval
```

4.6 Model Comparison and Final Selection

Table 6 reports model metrics at full precision. The primary selection criterion is minimum dev RMSE. If dev RMSE is tied, Table 6A breaks the tie using the smallest absolute gap between PI95 coverage and the nominal 95% level.

Table 6. Model comparison (ranked by DevRMSE, then PI95GapAbs, then DevMAE)

```
                        Model  Track   TrainR2     AdjR2   DevRMSE    DevMAE  \
0             Baseline LASSO  NoLog  0.826344  0.825180  0.216622  0.102911
1      Baseline OLS (Trimmed)  NoLog  0.826416  0.825253  0.216799  0.102916
2       Baseline Stepwise OLS  NoLog  0.826416  0.825253  0.216799  0.102916
3  Preprocessed OLS (Trimmed)    Log  0.784348  0.782902  0.222413  0.111527
4   Preprocessed Stepwise OLS    Log  0.784348  0.782902  0.222413  0.111527
5          Preprocessed LASSO    Log  0.784320  0.782873  0.222426  0.111607

  PI95CovPct PI95GapAbs
0  95.132743   0.132743
1  94.911504   0.088496
2  94.911504   0.088496
3  93.805310   1.194690
4  93.805310   1.194690
5  93.805310   1.194690
```

Table 6A. Final model selection summary

	SelectedModel	PrimaryCriterion	TieBreaker1	TieBreaker2	\
0	Baseline LASSO	Minimum DevRMSE	Minimum |PI95CovPct - 95|	Minimum DevMAE	

	SelectedDevRMSE	SelectedPI95CovPct	SelectedDevMAE
0	0.216622	95.132743	0.102911

4.7 Prediction on Evaluation Set

Table 7A records the selected model and output file metadata. Table 7 presents a preview of predicted PRSM and 95% prediction intervals for evaluation records.

Table 7A

	SelectedModel	RowsPredicted	OutputFile
0	Baseline LASSO	2500.000	predictions.csv

Table 7

	pred	pi95_lo	pi95_hi
0	0.511	0.304	0.718
1	0.813	0.606	1.020
2	0.411	0.204	0.618
3	0.727	0.520	0.934
4	0.956	0.749	1.162
5	1.007	0.800	1.214
6	0.942	0.735	1.149
7	1.158	0.951	1.365
8	0.964	0.757	1.171
9	0.591	0.384	0.798

4.8 Interpretation for Decision Support

Table 7A reports baseline prediction level and practical-effect threshold. Table 8 lists baseline borrower values. Table 9 reports practically important drivers and directions. Table 10 reports statistically detectable but practically small effects.

Table 7A. Interpretation summary metrics

	Metric	Value
0	BaselinePredictedPRSM	0.547
1	PracticalEffectThreshold	0.030

Table 8. 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 9. Practically important drivers

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

Table 10. 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