Credit Risk Project

- Introduction & Background
- Exploratory Data Analysis
- Data Preparation & Feature Engineering
- Model Assessment & Insights
- Key Findings & Future Directions

Jiapeng Wang

Introduction & Background

1 Problem Statement:

- Small business owners lack credit history, making them risky for banks.
- Credit union issues short-term loans (12 months, $10K - $2M) to small businesses.
- Sales fluctuations create repayment uncertainty, increasing credit risk.

2 Objective:

- Assess borrower creditworthiness using PRSM (Performance Ratio at Six Months).

$$\text{PRSM} = 2 \times \frac{\text{Amount repaid at 6 months}}{\text{Total amount owed}}.$$

3 Dataset Overview:

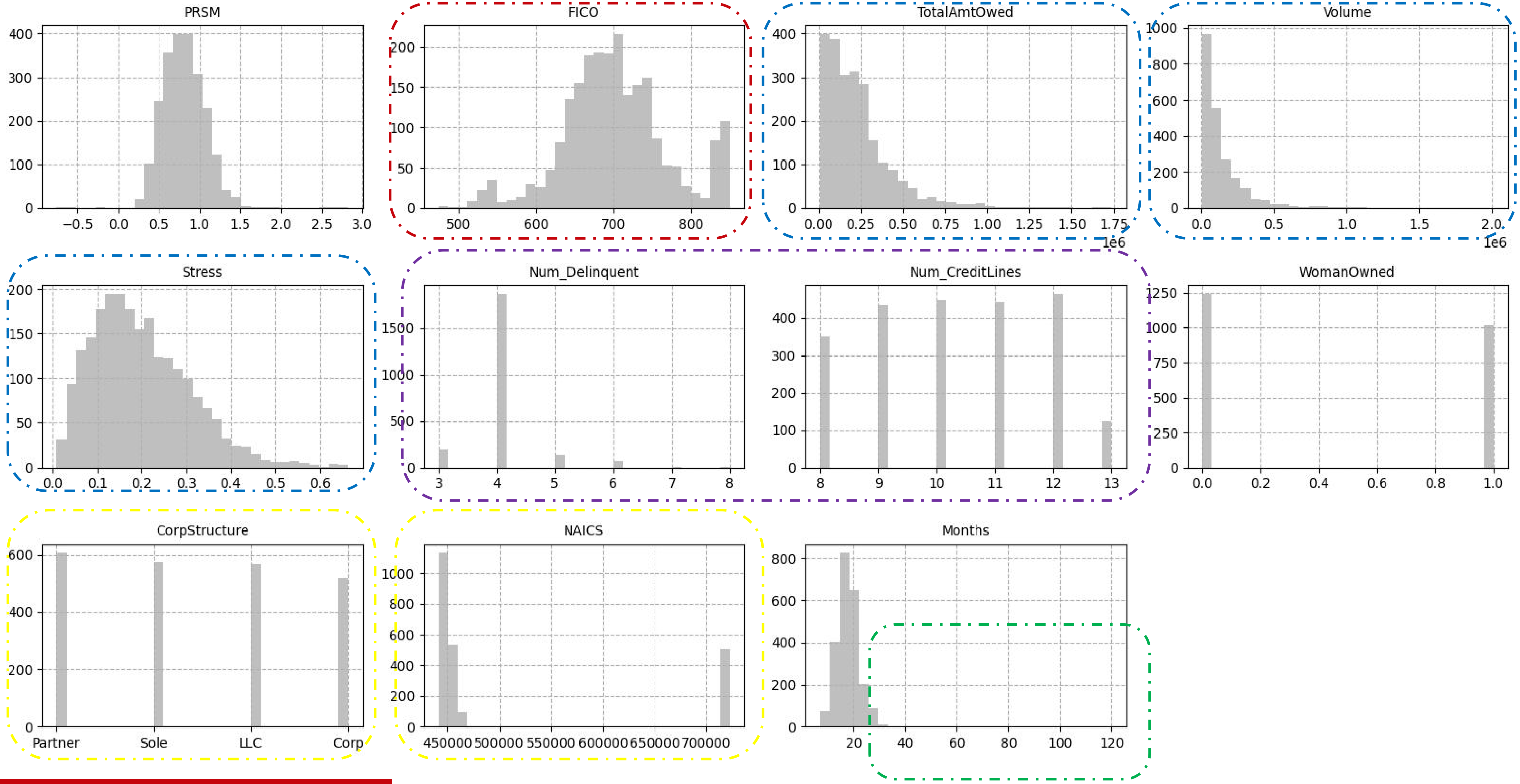
Key variables include:

- Binary features: WomanOwned
- Continuous features: TotalAmtOwed, etc.
- Categorical features: NAICS, CorpStructure, etc.

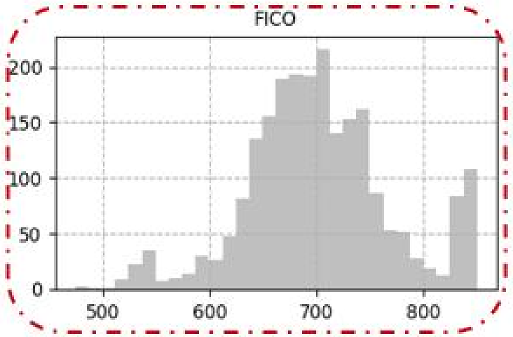
4 Approach:

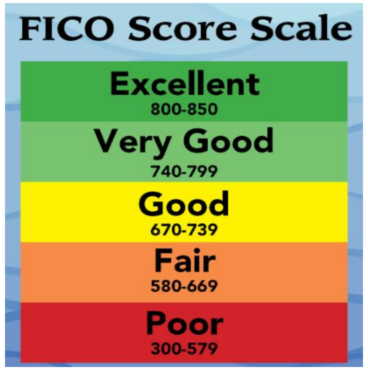
- Preprocess data to meet linear regression assumptions
- Train regression model to predict PRSM, identify key features.
- Use predictions to enhance loan approval decisions and mitigate risk.

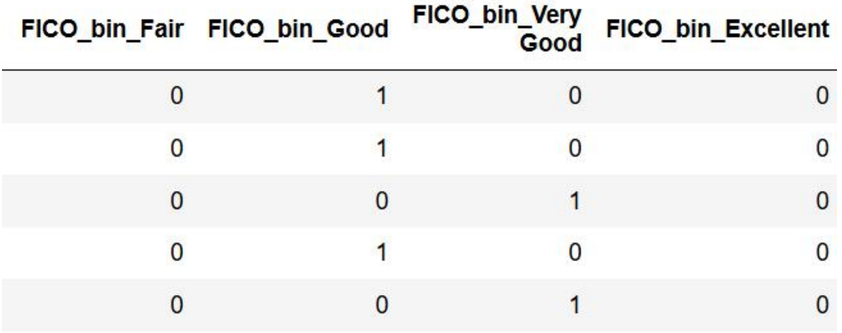
Exploratory Data Analysis



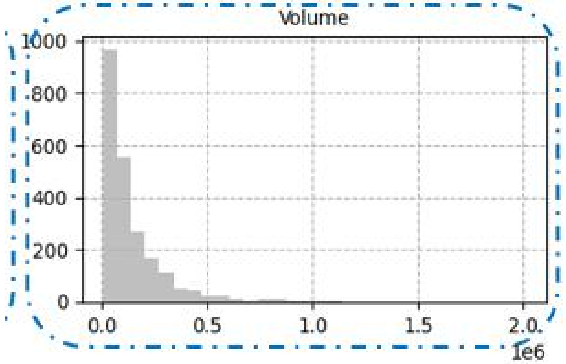
Data Preparation & Feature Engineering

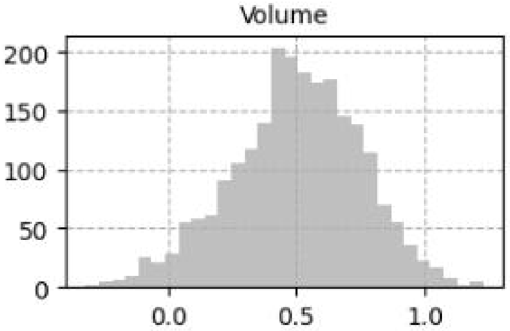


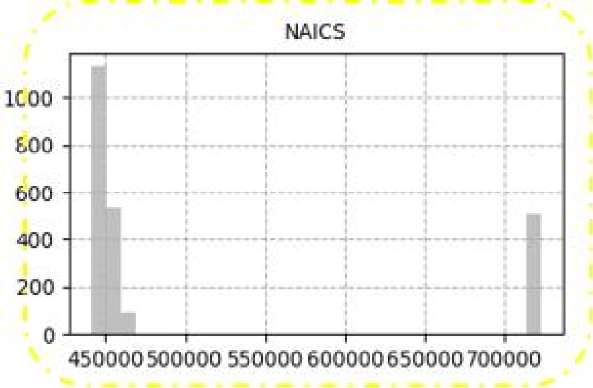


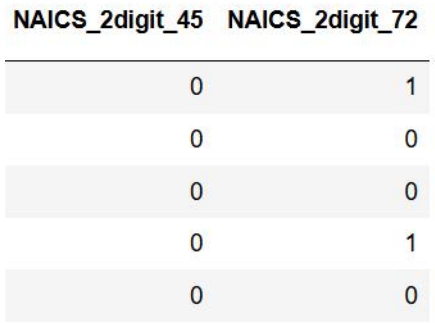


FICO_bin_Fair	FICO_bin_Good	FICO_bin_Very Good	FICO_bin_Excellent
0	1	0	0
0	1	0	0
0	0	1	0
0	1	0	0
0	0	1	0



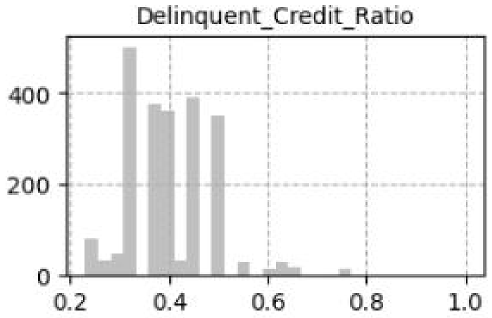


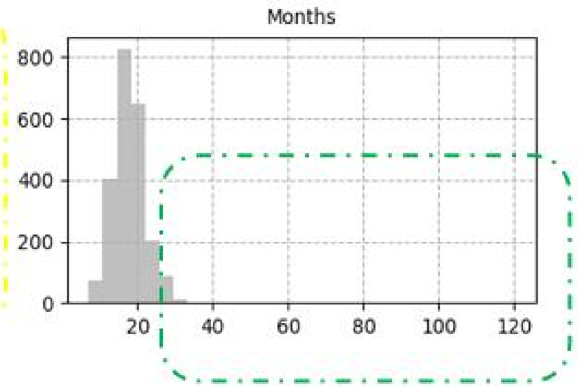


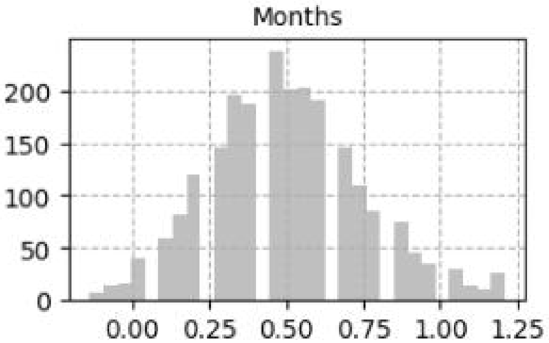


NAICS_2digit_45	NAICS_2digit_72
0	1
0	0
0	0
0	1
0	0

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

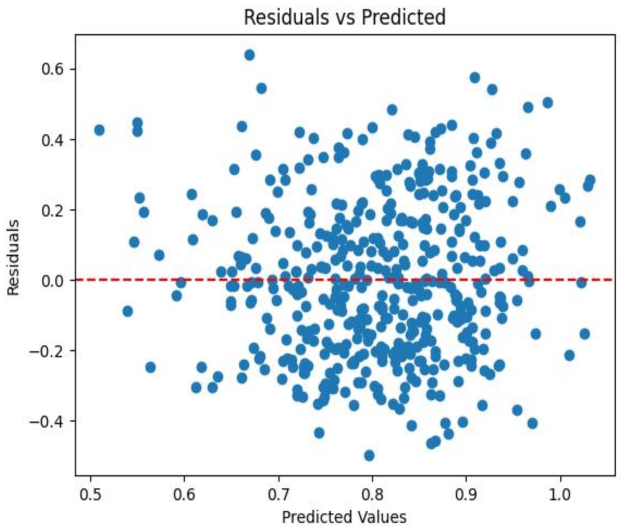


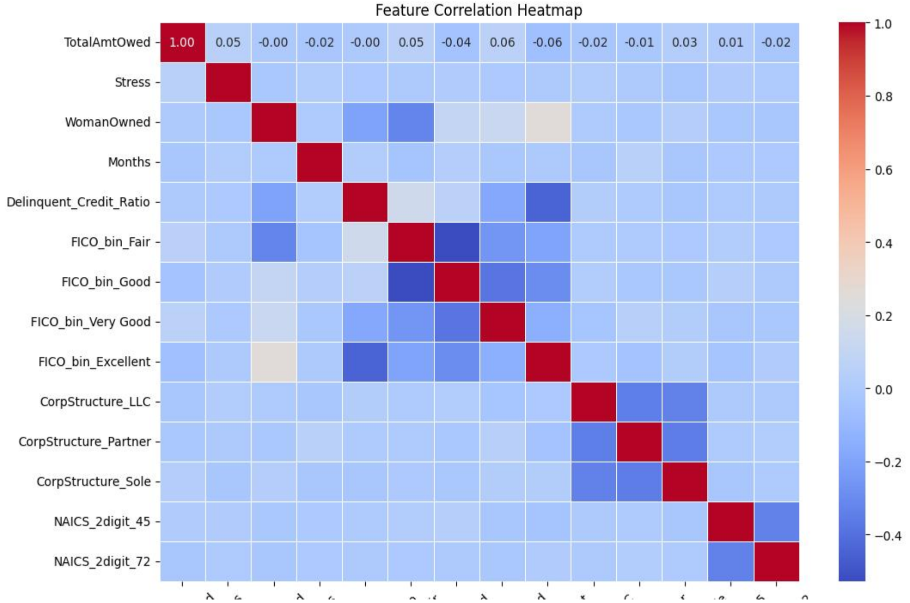




Model Assessment & Insights

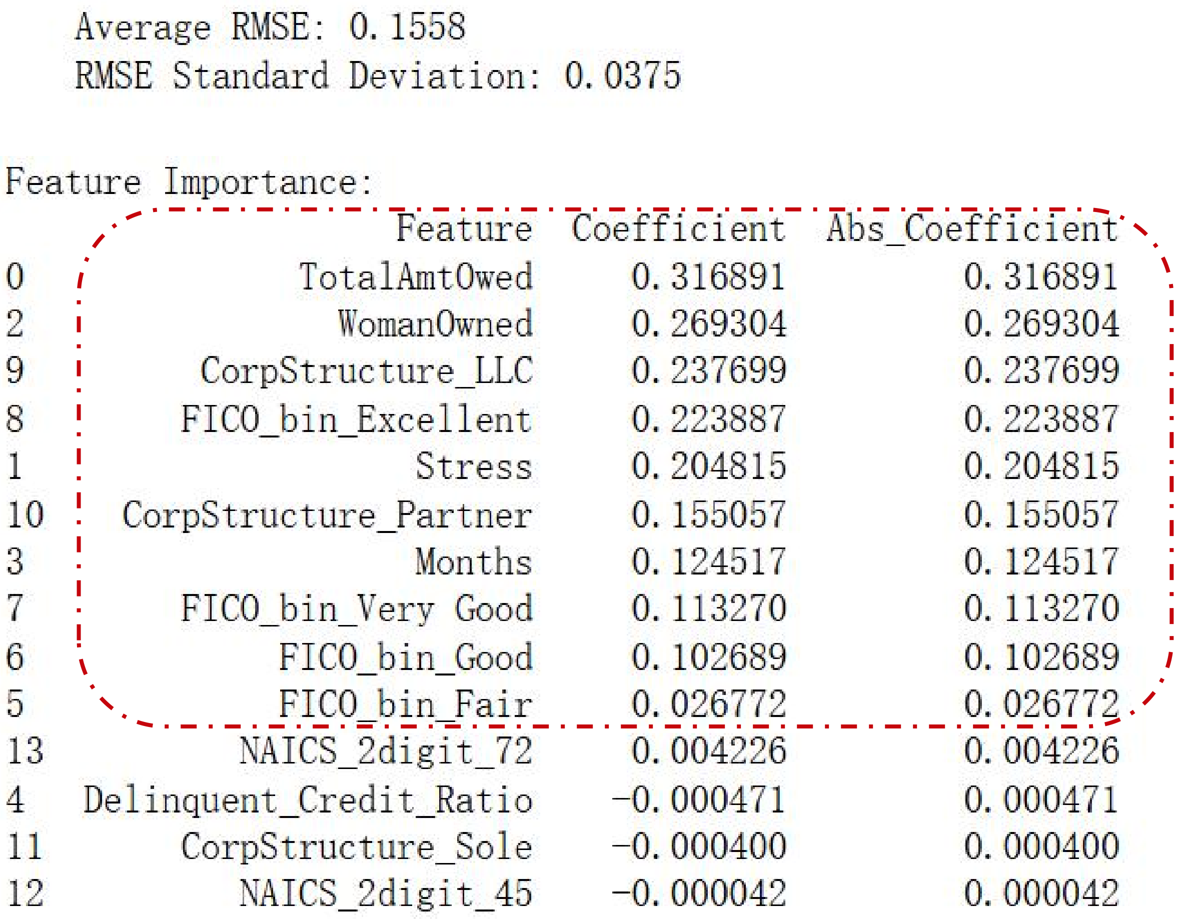
Checking Linear Regression Assumptions





Model Performance and Feature Importance

Under 5Fold cross validation:



```
Average RMSE: 0.1558
RMSE Standard Deviation: 0.0375

Feature Importance:
                    Feature  Coefficient  Abs_Coefficient
0              TotalAmtOwed     0.316891         0.316891
2                WomanOwned     0.269304         0.269304
9         CorpStructure_LLC     0.237699         0.237699
8        FICO_bin_Excellent     0.223887         0.223887
1                    Stress     0.204815         0.204815
10    CorpStructure_Partner     0.155057         0.155057
3                    Months     0.124517         0.124517
7        FICO_bin_Very Good     0.113270         0.113270
6             FICO_bin_Good     0.102689         0.102689
5             FICO_bin_Fair     0.026772         0.026772
13          NAICS_2digit_72     0.004226         0.004226
4   Delinquent_Credit_Ratio    -0.000471         0.000471
11       CorpStructure_Sole    -0.000400         0.000400
12          NAICS_2digit_45    -0.000042         0.000042
```

Key Findings & Future Directions

Key Drivers of Credit Risk

- **Loan Amount**: Larger loans link to higher PRSM scores, indicating stronger cash flow.
- **Women-Owned Businesses**: Higher PRSM scores, possibly due to financial management or industry focus.
- **Business Structure**: LLCs and partnerships score higher than sole proprietorships and corporations.
- **FICO Score**: Higher scores ("Excellent" or "Very Good") suggest better repayment performance.
- **Financial Stress**: A higher garnishment-to-volume ratio is associated with lower credit risk.
- **Months in Business**: Longer operation reduces default risk, as shown by a positive coefficient.

Future Work

Improving Data Analysis

- Explore feature interactions to uncover hidden patterns.
- Test advanced models for better accuracy and interpretability.
- Enhance weak features like NAICS for better insights.

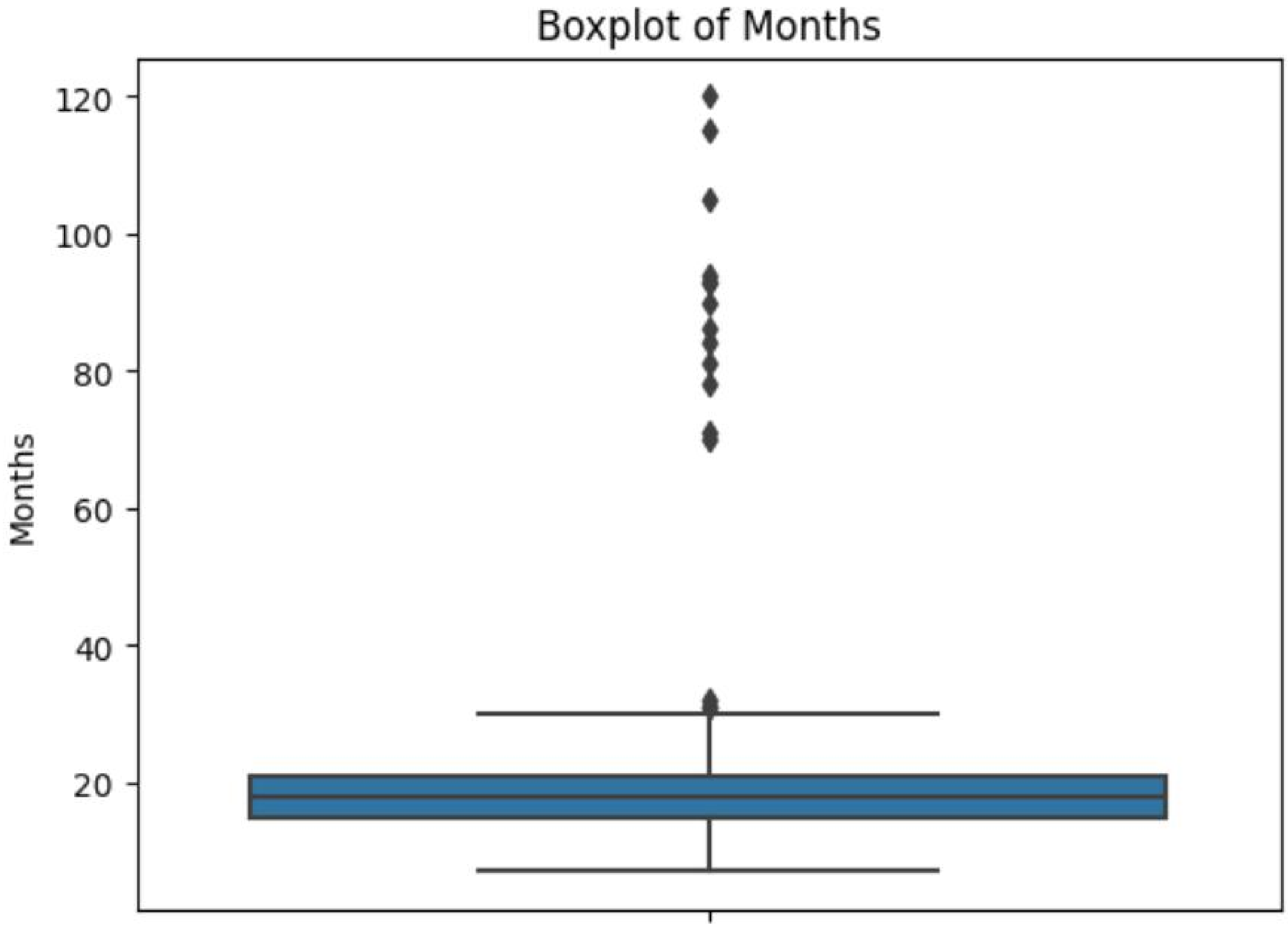
Next Steps for Business Impact

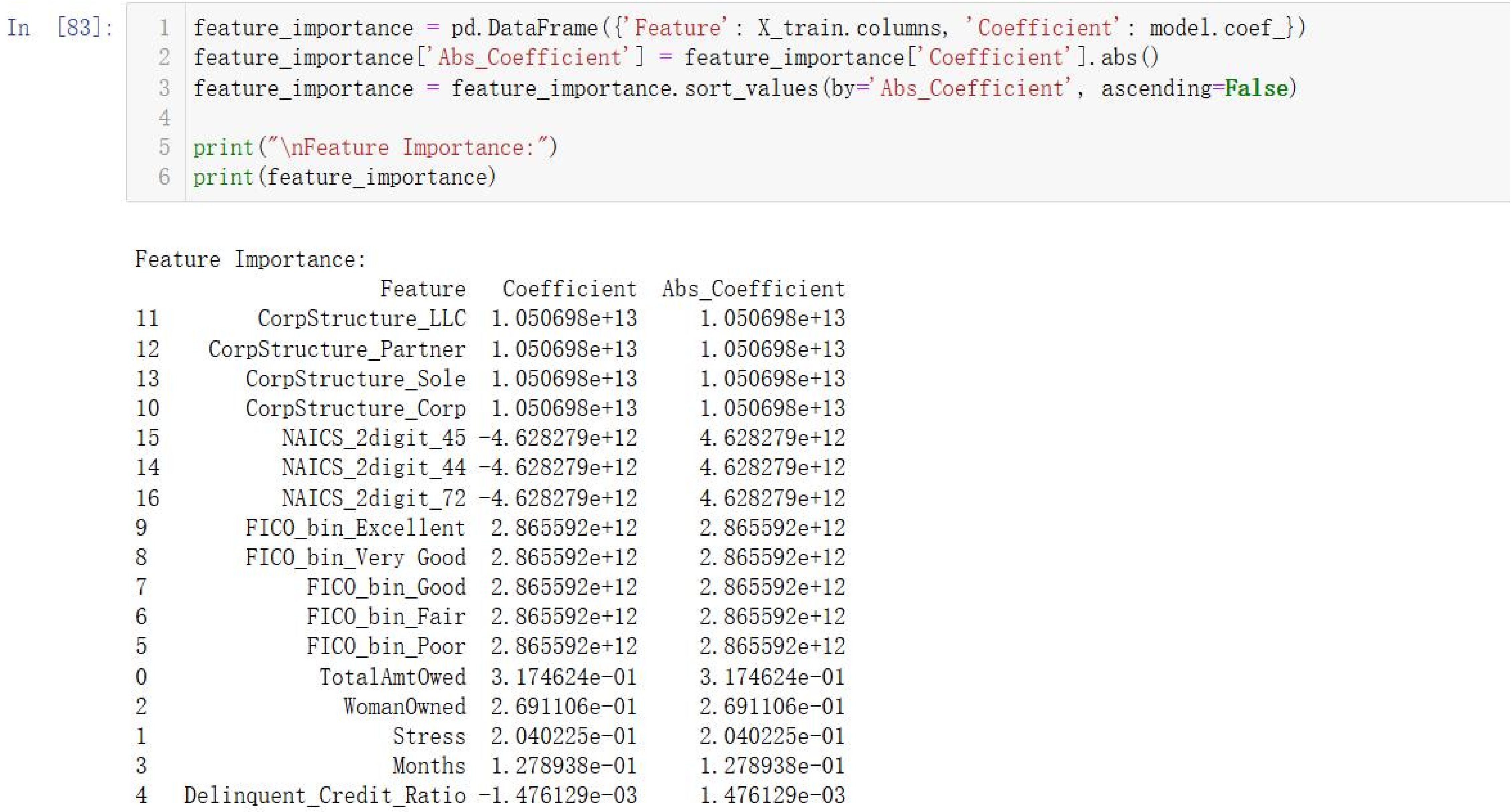
- Validate the model with real-world testing.
- Integrate it into decision-making processes.
- Monitor performance and refine based on new data.

Thank you

Q&A

Backup


Boxplot of Months
Months
120
100
80
60
40
20



In [83]:

```python
feature_importance = pd.DataFrame({'Feature': X_train.columns, 'Coefficient': model.coef_})
feature_importance['Abs_Coefficient'] = feature_importance['Coefficient'].abs()
feature_importance = feature_importance.sort_values(by='Abs_Coefficient', ascending=False)

print("\nFeature Importance:")
print(feature_importance)
```

```
Feature Importance:
                    Feature   Coefficient  Abs_Coefficient
11        CorpStructure_LLC  1.050698e+13     1.050698e+13
12    CorpStructure_Partner  1.050698e+13     1.050698e+13
13       CorpStructure_Sole  1.050698e+13     1.050698e+13
10       CorpStructure_Corp  1.050698e+13     1.050698e+13
15          NAICS_2digit_45 -4.628279e+12     4.628279e+12
14          NAICS_2digit_44 -4.628279e+12     4.628279e+12
16          NAICS_2digit_72 -4.628279e+12     4.628279e+12
9        FICO_bin_Excellent  2.865592e+12     2.865592e+12
8        FICO_bin_Very Good  2.865592e+12     2.865592e+12
7             FICO_bin_Good  2.865592e+12     2.865592e+12
6             FICO_bin_Fair  2.865592e+12     2.865592e+12
5             FICO_bin_Poor  2.865592e+12     2.865592e+12
0              TotalAmtOwed  3.174624e-01     3.174624e-01
2                WomanOwned  2.691106e-01     2.691106e-01
1                    Stress  2.040225e-01     2.040225e-01
3                    Months  1.278938e-01     1.278938e-01
4   Delinquent_Credit_Ratio -1.476129e-03     1.476129e-03
```

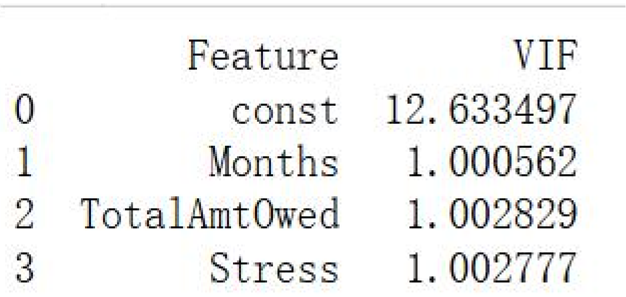
The **Variance Inflation Factor (VIF)** measures how much a predictor's variance is increased due to its correlation with other predictors in a regression model. A high VIF indicates multicollinearity, meaning the predictor is highly correlated with others, which can make the regression coefficients unstable.

Mathematically, the VIF for a predictor variable X_i is calculated as:

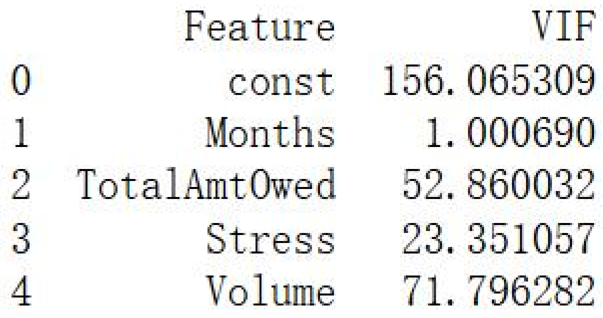
$$VIF(X_i) = \frac{1}{1 - R_i^2}$$

Where R_i^2 is the R-squared value obtained by regressing X_i on all other predictor variables in the model.

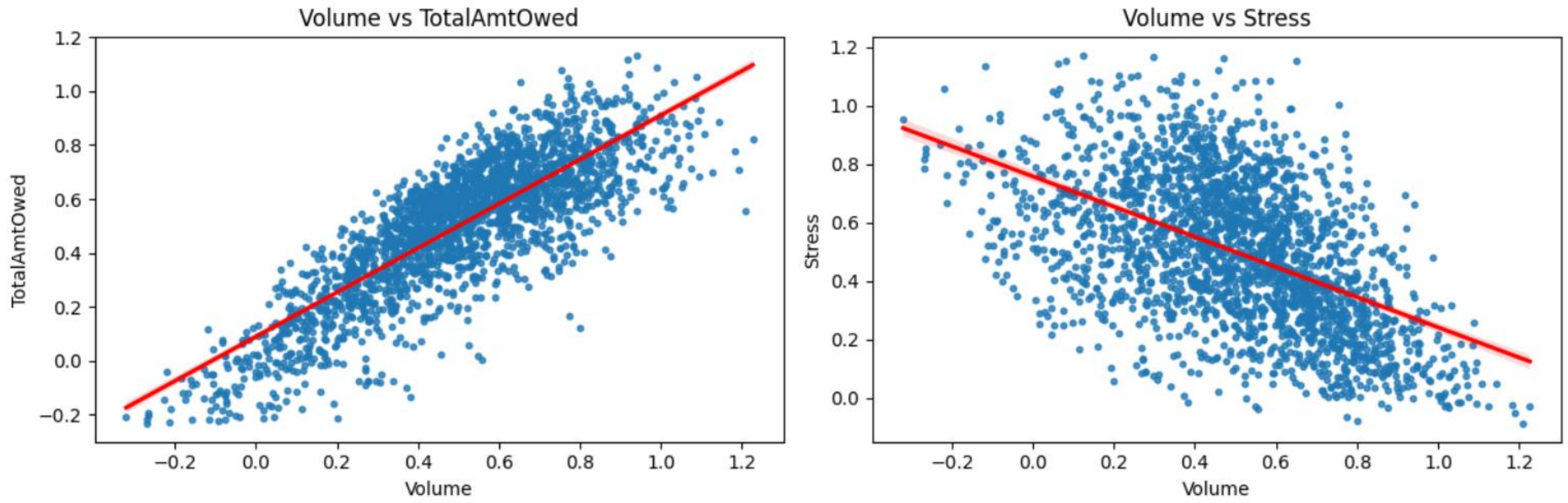
- **VIF > 10** suggests a high level of multicollinearity, which may require addressing by removing variables, combining them, or using regularization techniques.
- **VIF < 5** generally indicates that multicollinearity is not a concern.

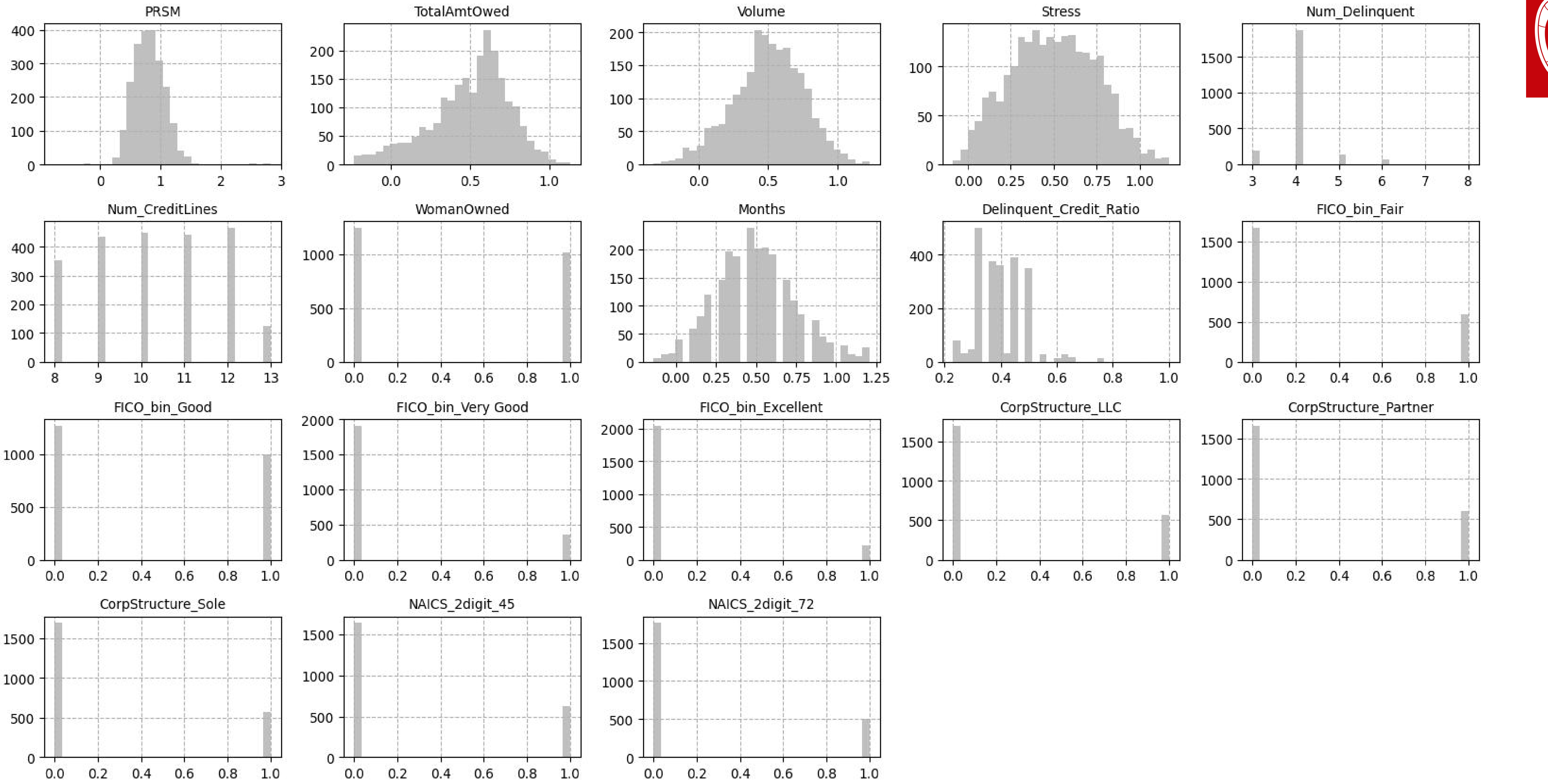


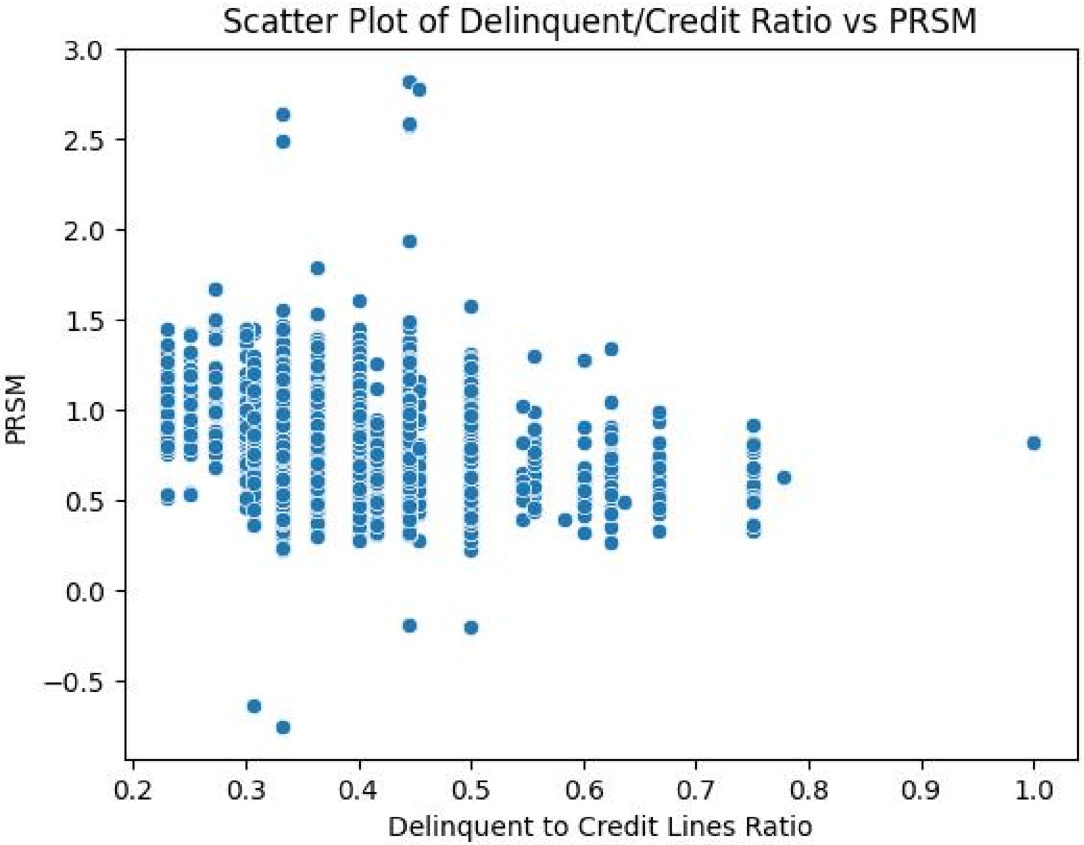
```
      Feature        VIF
0       const  12.633497
1      Months   1.000562
2 TotalAmtOwed   1.002829
3      Stress   1.002777
```



```
      Feature         VIF
0       const  156.065309
1      Months    1.000690
2 TotalAmtOwed  52.860032
3      Stress   23.351057
4      Volume   71.796282
```




PRSM
0
100
200
300
400
0
1
2
3
TotalAmtOwed
0
50
100
150
200
0.0
0.5
1.0
Volume
0
50
100
150
200
0.0
0.5
1.0
Stress
0
50
100
0.00
0.25
0.50
0.75
1.00
Num_Delinquent
0
500
1000
1500
3
4
5
6
7
8
Num_CreditLines
0
100
200
300
400
8
9
10
11
12
13
WomanOwned
0
500
1000
0.0
0.2
0.4
0.6
0.8
1.0
Months
0
50
100
150
200
0.00
0.25
0.50
0.75
1.00
1.25
Delinquent_Credit_Ratio
0
200
400
0.2
0.4
0.6
0.8
1.0
FICO_bin_Fair
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
FICO_bin_Good
0
500
1000
0.0
0.2
0.4
0.6
0.8
1.0
FICO_bin_Very Good
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
FICO_bin_Excellent
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
CorpStructure_LLC
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
CorpStructure_Partner
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
CorpStructure_Sole
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
NAICS_2digit_45
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
NAICS_2digit_72
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0


Scatter Plot of Delinquent/Credit Ratio vs PRSM
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
PRSM
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Delinquent to Credit Lines Ratio

Note: Only use red icons on white or light gray backgrounds