LOAN DEFAULT RISK MODELING USING PYTHON-BASED ECONOMETRIC ANALYSIS FOR U.S. CREDIT UNIONS

1. Background

Credit unions in the United States increasingly face loan default risks as member profiles become more diverse and economic volatility rises. One regional credit union consortium sought to better understand which factors—across income, age, product type, and employment—were statistically linked to defaults on personal loans.

We were engaged to apply econometric modeling techniques in Python to derive a robust and interpretable loan default risk model. The focus was not only on prediction, but also on actionable insight for underwriting policy and member profiling.

2. Objective

- To identify statistically significant drivers of personal loan default using econometric tools
- To build a transparent, regression-based model with interpretable coefficients and policy implications
- To help credit unions refine credit scoring logic and establish risk-tiering rules based on model insights
- To deliver reproducible Python code, ready-to-use reporting tables, and APA-style documentation

3. Data Used

Source: Internal borrower-level data from three affiliated credit unions

Dataset: 10,500 member loan applications, including both performing and defaulted loans, from Jan 2020 to Dec 2022

Variables:

- Defaulted (1 = Defaulted, 0 = Repaid)
- Member_Age
- Loan_Amount

- Loan_Term (in months)
- Employment_Status (Employed / Self-Employed / Unemployed)
- Annual_Income
- Credit_Score
- Loan_Type (Personal / Auto / Line of Credit)
- Debt_To_Income_Ratio
- Previous_Default_History (Yes/No)

4. Methodology

4.1 Data Preparation

- Cleaned outliers and missing records using IQR filtering and mean/mode imputation
- Encoded categorical variables with dummy variables (Loan_Type, Employment_Status, Previous_Default_History)
- Created interaction terms (e.g., Credit_Score × Loan_Type) to capture segment-specific behaviors
- Standardized all continuous variables for comparability in coefficient interpretation

4.2 Econometric Model

- **Logistic Regression Model** estimated using statsmodels.Logit
- Model equation:

$$\text{logit}\big(P(Default = 1)\big) = \beta_0 + \beta_1 CreditScore + \beta_2 DTI + \beta_3 Income + \beta_4 Age + \beta_5 EmploymentStatus + \cdots +$$

- Included diagnostics:
  - Multicollinearity check (VIF < 5 for all variables)
  - Goodness-of-fit: Log-likelihood, McFadden R^2
  - ROC curve and AUC for predictive quality
  - Residual analysis and Hosmer-Lemeshow test

5. Key Results

Variable	Coefficient (Log-Odds)	Odds Ratio	P-Value
Credit_Score	–0.018	0.982	0.000
Debt_To_Income_Ratio	+0.039	1.039	0.001
Previous_Default_History (Yes)	+0.884	2.42	0.000
Employment_Status (Unemployed)	+0.542	1.72	0.015
Loan_Type (Auto)	–0.312	0.73	0.042
Age	–0.009	0.991	0.034

- **McFadden R^2**: 0.29
- **AUC-ROC**: 0.81 (good model discrimination)
- The model correctly predicted default classification for **78.3%** of test set records

6. Econometric Interpretation

- **Lower credit scores**, **higher DTI**, and **prior default history** were the strongest indicators of default
- **Younger and unemployed members** had a higher risk of defaulting, controlling for all else
- **Auto loans** showed significantly lower odds of default than unsecured personal loans
- The **odds of default more than doubled** for members with previous defaults
- The logistic model coefficients were stable and statistically significant, with no multicollinearity or misspecification issues

7. Report Output

- **PDF Report (APA format, 18 pages)**:
    - Executive summary, variable tables, odds ratio interpretation, AUC plot
    - Appendix with VIF table, Hosmer-Lemeshow statistic, and model assumptions
    - Clear commentary on what each variable means for underwriting teams
- **Python Notebook**:

    - Clean script divided into data prep, model build, diagnostics, and export sections
    - evaluate_model() function to repeat testing on future datasets
    - Output-ready tables formatted as Pandas DataFrames and exported to .csv
- **Loan Risk Tier Sheet**:
    - Excel sheet assigning “Low”, “Medium”, or “High Risk” tags based on model-predicted probabilities (>0.6 = High Risk)
    - Shared with client’s credit analysts for pilot rollout

8. Strategic Recommendations

- Use model outputs to assign **pre-screening risk flags** before full underwriting
- Reduce approval caps for high-risk profiles (e.g., low credit + prior default + high DTI)
- Offer safer loan types (e.g., collateral-backed auto loans) to mid-risk borrowers instead of personal loans
- Update and retrain model quarterly as macroeconomic conditions shift
- Test hybrid model combining logistic regression insights with machine learning (e.g., Random Forest) for deeper segmentation