Logistic Regression for Credit Risk Analysis

By Ali Jomaa and **Sai Deepak Chandra**
University of Michigan -Dearborn
alijom@umich.edu
University of Michigan -Dearborn
csai@umich.edu

1 - INTRODUCTION & LITERATURE REVIEW

Credit risk analysis is a term in the financial industry that describes statistical methods used by financial institutions to assess the risk associated with a loan for a particular applicant. In this report, we will analyze the Loan Default Prediction Dataset from Kaggle and fit a logistic regression model to predict the likelihood that an applicant will default on a loan, using features from each loan application. These features include Age, Income, Loan amount, credit score, months employed, number of credit lines, Interest rate, loan term, DTI ratio, education level (High school, bachelor's, master's, PhD), Employment type (self-employed, full-time, part-time, unemployed), Marital status (single, married), Has Mortgage, Has dependents, loan purpose (business, education, home, other), and has co-signer. These are features that financial institutions consider in every loan application to assess risk.

As the credit industry grows, an increasing number of institutions will rely on risk management models, such as the one discussed in this report, to decide whether to grant credit to an individual. Over the past couple of decades, computer methods and statistical modeling have advanced significantly and proven themselves to be a vital resource for financial institutions such as banks. According to a Workday survey, 86% of respondents in financial services agree that machine learning methods are essential to keeping their businesses competitive and operating efficiently.[1]

Similar research has been conducted by the KTH Institute of Research and Technology [2] and by the SRM Institute of Science and Technology [3]. Both papers use Logistic Regression to predict the probability that an individual will default on a loan.

According to a study by JOEL NORLING & SAMI ABDU at the KTH Royal Institute of Technology, despite the increase in interest in artificial intelligence and more advanced machine learning methods, Logistic regression remains the most widely used approach due to its simplicity and lower implementation costs. Also, a study by West [4] indicated that, on average, logistic regression is slightly more accurate than neural network models in terms of credit risk. Overall, researchers have found that data-driven processes lead to more accurate credit risk assessments.

2 - DESCRIPTION OF DATASET

The data set used in our project contains detailed information about borrowers and their loan attributes, enabling the prediction of loan default. It includes a mix of demographic, financial, and behavioral variables that reflect different factors influencing credit risk. The dataset consists of a Loan ID, a unique identifier for each record, and a binary Default variable indicating the target outcome (0 = non-default, 1 = default).

Overall, the dataset includes 15 independent variables and one target variable, covering the following categories:

Attributes - These variables describe the borrower's personal background:

Age – The borrower's age in years, Education – Highest education level obtained (e.g., High School, Bachelor's, Master's, Ph.D.), Marital Status – Indicates whether the borrower is Married, Single, or Divorced.

Financial Attributes - These fields provide insight into the customer's economic capacity:

Income – Annual income of the borrower, Credit Score – Creditworthiness score associated with the borrower, DTI Ratio – Debt-to-Income Ratio, indicating how much of the borrower's income goes toward debt payments, Has Mortgage – Whether the borrower already has an existing mortgage, Has Dependents – Whether the borrower supports dependents.

Employment-Related Attributes - These variables describe the borrower's job profile:

Employment Type – Full-time, Part-time, Self-employed, or Unemployed, Months Employed – Number of months the borrower has been employed, Number of credit lines – Number of credit lines the borrower currently has open.

Loan-Specific Attributes - These variables describe the loan applied for:

Loan Amount – Total loan amount requested by the borrower, Interest Rate – Annual interest rate applied to the loan, Loan Term – Duration of the loan in months, Loan Purpose – Reason for the loan (e.g., Auto, Business, Education, Other), Has Cosigner – Indicates whether another individual has co-signed the loan.

Target Variable - Default:

Binary label indicating whether the borrower failed to repay the loan (1 = default, 0 = did not default)

Dataset Structure Overview:

Number of columns: 18.

Number of rows: 255347.

Variable types: Mix of numerical and categorical attributes.

3. EXPLORATORY DATA ANALYSIS

To understand the dataset structure, identify trends, spot exceptions, and compile essential features of both numerical and categorical variables, an exploratory data analysis (EDA) was conducted. The goal of this phase is to gain knowledge about borrower demographics, financial behavior, and loan features that could affect default risk. The basis for choosing these features and directing the modeling process was established. During this process, we also tested for interactions by analyzing the P values associated with them to determine Statistical Significance. During these tests, one interaction was found to be statistically significant ($(P = 0.003) < (\propto = 0.05)$). This interaction was between age and income. Despite the low P-value, this interaction was not included in the final model because it decreased the model's accuracy, recall, and ROC AUC scores. This is most likely due to Multicollinearity. These two features are often highly correlated, which leads to a misleading P-value and decreased model performance.

Distribution of Loan Default:

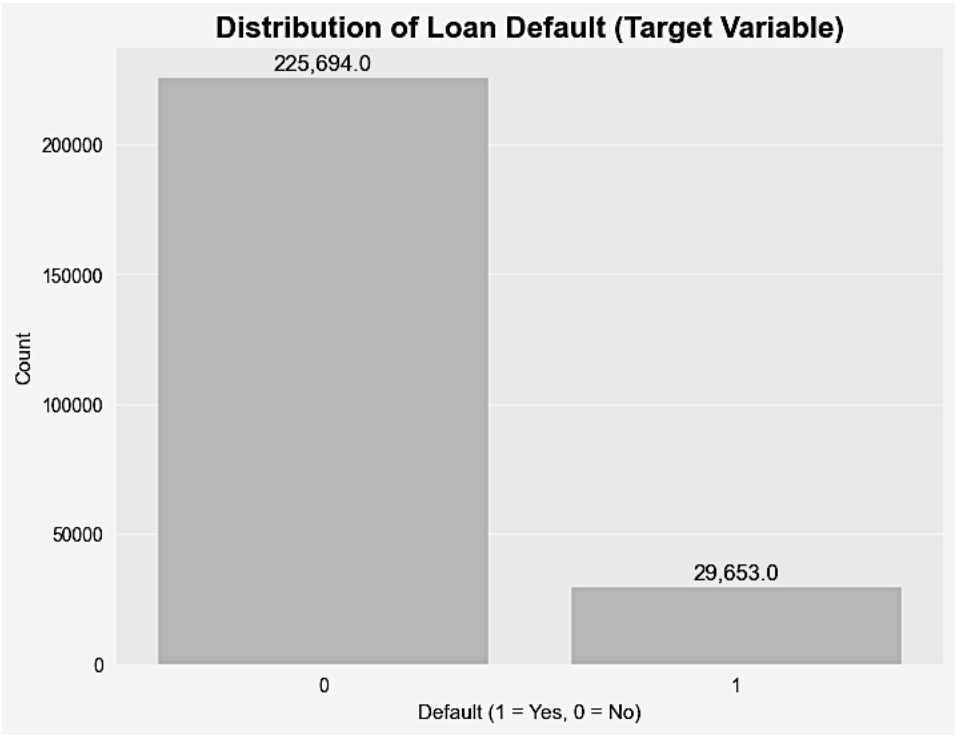

Figure 1

Figure 1 presents the class distribution of the dataset. Among all observations, 225,694 are non-default cases, and 29,653 are default cases, yielding an overall default rate of 11.61%. This magnitude is broadly consistent with real-world credit datasets, which typically exhibit default rates of approximately 5–10%.

Class Imbalance:

Class imbalance is a prevalent issue in credit risk datasets, where the number of non-default cases typically far exceeds the number of default cases. This imbalance reflects real-world conditions, as defaults are relatively rare compared to successful repayments. For example, default rates around 5% are commonly observed in consumer credit datasets.

To address class imbalance, practitioners often employ balancing techniques, such as reweighting, oversampling, or undersampling, and assess model performance using metrics that account for both classes, including ROC AUC, F1 score, Accuracy, Precision, and Recall.

In our analysis, we compared the performance of the original, unbalanced logistic regression model with that of a model trained with class balancing. We observed that the Precision, Recall, and F1 scores remained essentially unchanged, with the balanced model performing slightly worse overall. Specifically, the ROC AUC decreased from 0.7478 in the original model to 0.7411 in the balanced model. This reduction is expected: balancing can modify the predicted probability distribution, which, in turn, affects the ranking of positive and negative cases, slightly lowering the ROC AUC even when class-specific metrics remain similar. Factors contributing to this effect may include information loss due to undersampling, overfitting in oversampling methods, or shifts in the underlying data distribution.

Upon examining the class distributions before and after balancing, we found that the default rate in the original dataset was 11.71%, whereas after balancing, it increased to 36.20%. The balancing process reduced the number of non-default cases while increasing the representation of defaults, which likely explains the observed decrease in overall model accuracy.

Considering these findings and the largely unaffected class-specific performance metrics, we chose to proceed with the original unbalanced model for subsequent analysis.

Our observations regarding balancing techniques are consistent with the findings of many studies, including the 2024 study conducted at the University of Trento by Ahmed Almustfa Hussin, Adam Khatir, and Marco Bee, which reported that “in terms of accuracy, the results were worse for the NN and LR (Logistic Regression) and remained approximately the same for RF” [5]. Overall, this is a serious, unresolved issue in credit risk modeling.

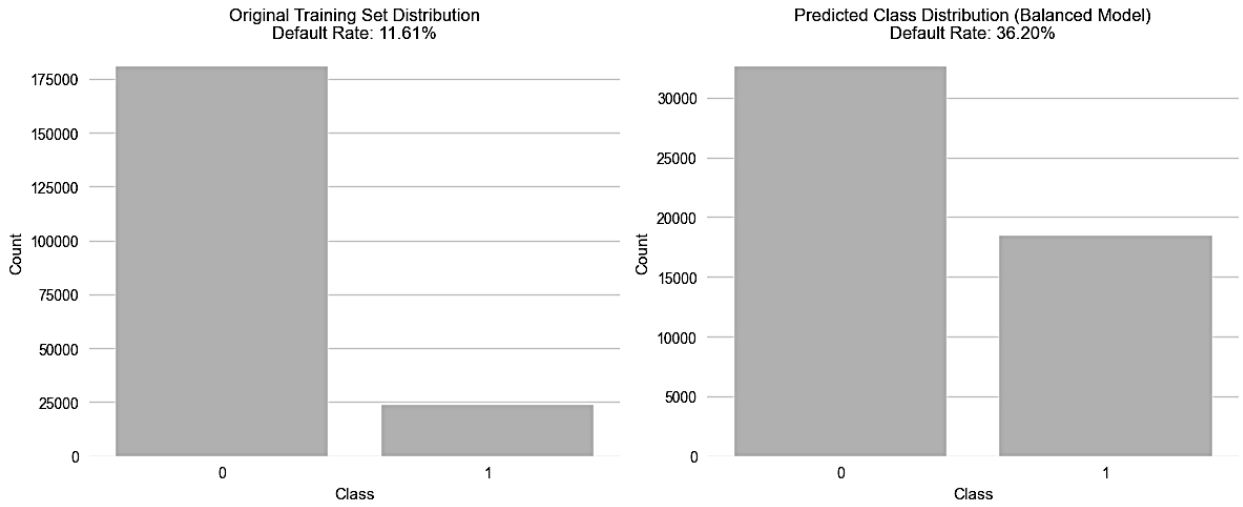

Figure 2

Data Transformation:

Since our dataset included multiple categorical variables, such as Education level, we used one-hot encoding to prepare our data for modeling. For example, for the Education level category, we have high school, bachelor's, master's, and Ph.D., one-hot encoding splits this one column into four separate binary columns. In each row, only one of these new columns will be set to "hot" (1), the others will be set to "cold" (0).

Default rates for categorical predictors

Figure 3 presents default rates for the seven categorical predictors. The patterns are clear, economically intuitive, and strongly support including all variables in the multivariate model.

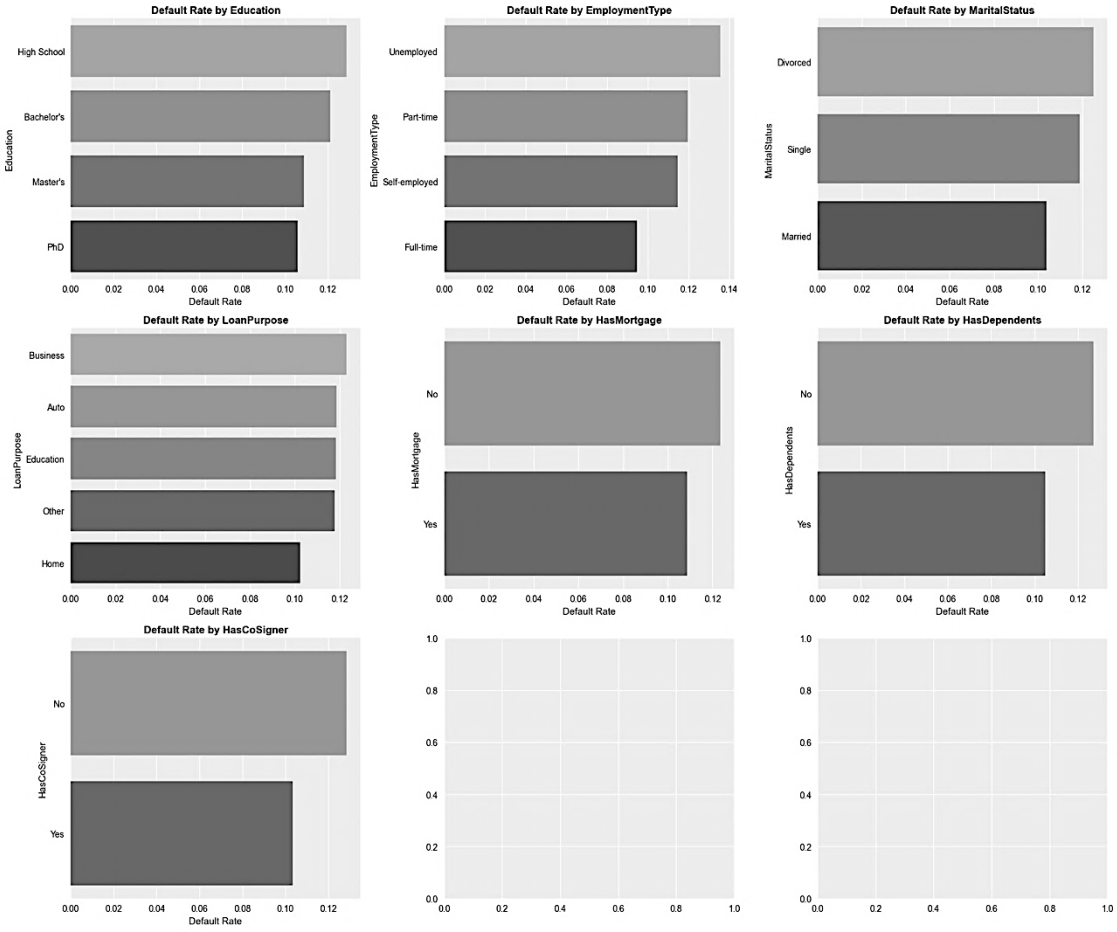

Figure 3

- Higher educational attainment is associated with progressively lower default risk. The most advanced degree category shows the lowest default rate.
- Employment status exhibits a dramatic gradient: full-time employees have markedly lower default rates than part-time employees, the self-employed, or the unemployed.
- Married individuals display noticeably lower default rates than single or divorced borrowers.
- Loan purpose reveals meaningful differences: loans taken for home purchase and education purposes are considerably safer than those for business or auto financing.
- Borrowers who already have a mortgage, who have dependents, or who have a co-signer consistently show lower default rates than their counterparts without these characteristics. The presence of a co-signer is one of the strongest protective factors in the univariate analysis.

This aligns closely with financial theory and credit-risk intuition: greater human capital, income stability, collateral, and third-party liability all serve as powerful buffers against default. The observed patterns provide strong justification for retaining all categorical variables in the logistic regression specification.

Feature Engineering:

To better capture the borrower's financial leverage, we engineered a new feature called 'Loan-to-Income Ratio' (LTI) by dividing the requested Loan Amount by the Annual Income. We investigate this variable's impact in the modeling section.

4 - DESCRIPTIVE STATISTICS

To give a preliminary understanding of data distribution, central tendency, and variability, descriptive statistics were calculated for the numerical variables.

	MINIMUM	Q1	MEDIAN	Q3	MEAN	STD DEV
Age	18	31	43	56	43.5	14.99
Income	15000	48825	82466	116219	82499.30	38963.01
LoanAmount	5000	66156	127556	188985	127578.87	70840.71
CreditScore	300	437	574	712	574.26	158.90
Months Employed	0	30	60	90	59.54	34.64
NumCreditLines	1	2	2	3	2.50	1.13
Interest Rate	2	7.77	13.46	19.25	13.49	6.64
Loan Term	12	24	36	48	36.03	16.97
DTI Ratio	0.1	0.3	0.5	0.7	0.50	0.23

Table 1

Table 1 presents the summary statistics for the numeric predictors in the loan default prediction dataset (N = 255,347). Borrowers range in age from 18 to 69 years, with a mean age of 43.5 years. Annual income displays considerable variation, ranging from $15,000 to nearly $150,000, with a median of $82,466 and a mean of $82,499. Loan amounts range from $5,000 to $250,000, with a median of $127,556. Credit scores follow an approximately uniform distribution between 300 and 849 (mean = 574.3), consistent with many public credit-risk datasets.

Employment tenure averages 59.5 months (about 5 years), while the number of existing credit lines is typically low (median = 2). Interest rates range from 2% to nearly 25%, with a median of 13.46%, reflecting a mixture of prime and subprime lending. The typical loan term is 36 months, and the median debt-to-income (DTI) ratio is 0.50,

indicating that half of the borrowers devote at least 50% of their gross income to debt obligations; a level generally considered high-risk by lending standards.

These descriptive patterns are consistent with a portfolio that includes both prime and subprime borrowers, providing substantial variation for modeling default risk. The relatively high median interest rate and DTI ratio suggest a meaningful proportion of higher-risk loans, reflected in the overall default rate of approximately 11.6%.

5 - MODEL CHOICE, FEATURE ENGINEERING, & PERFORMANCE

To rigorously assess credit risk, we employed a two-stage modeling approach. First, we established a baseline using standard logistic regression on the raw dataset. Second, we refined the model by introducing a feature-engineered variable, the **Loan-to-Income (LTI) Ratio**, to better capture borrower leverage.

5.1 Baseline Model:

For our analysis, we selected **logistic regression** as the baseline model due to its interpretability, efficiency, and firm performance in binary classification tasks, particularly in credit risk modeling. Logistic regression allows us to directly interpret feature coefficients as odds ratios, which is valuable for understanding the relative impact of predictors on default probability. Additionally, logistic regression is widely used in financial risk assessment, providing a benchmark for comparison with more complex models.

The performance of the logistic regression model on the test dataset is summarized below:

Class	Precision	Recall	F1-score	Support
0 (Non-default)	0.94	0.65	0.77	45,139
1 (Default)	0.21	0.71	0.33	5,931
Accuracy			0.66	51,070
Macro avg	0.58	0.68	0.55	51,070
Weighted avg	0.86	0.66	0.72	51,070

Table 2

Confusion Matrix:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} = \begin{bmatrix} 29377 & 15762 \\ 1717 & 4214 \end{bmatrix}$$

ROC-AUC Score: 0.7478

The model demonstrates a strong ability to identify non-default cases, reflected in high precision (0.94) for the majority class, but lower precision (0.21) for default cases due to the inherent class imbalance. Recall defaults are comparatively higher (0.71), indicating that the model correctly identifies most actual defaults, which is crucial in credit risk contexts where missing defaults carry a higher cost. The ROC-AUC score of 0.7478 further confirms that the model achieves a reasonable balance between sensitivity and specificity across different thresholds.

Overall, logistic regression provides a transparent, interpretable, and reasonably accurate model, making it suitable for both predictive performance and regulatory interpretability in credit risk applications. Given the modest improvements achieved with balancing techniques and the preservation of class-specific metrics, we elected to proceed with the unbalanced logistic regression model for subsequent analyses.

Examining the confusion matrix provides further insight into model behavior. Out of 45,139 non-default cases, 29,377 were correctly classified, while 15,762 were misclassified as defaults. For the 5,931 default cases, 4,214 were correctly identified, while 1,717 were misclassified as non-defaults. This highlights that the model is more effective at identifying non-defaults than defaults, consistent with the dataset's class imbalance. While the lower precision for defaults (0.21) indicates some false positives, the relatively high recall (0.71) shows that most actual defaults are detected, a critical factor in credit risk applications, where failing to flag potential defaulters has higher financial consequences than incorrectly flagging non-defaulters. These findings are consistent with Credit risk models used in the industry.

In fact, prior research has shown that class imbalance in credit risk datasets often leads to more false negatives**.** In these cases, actual defaulters are incorrectly classified as non-defaulters, which can have serious financial consequences for financial institutions. In our analysis, however, the logistic regression model demonstrates a relatively low number of false negatives compared to false positives. Specifically, while the model misclassifies a portion of defaults, the majority are correctly identified, resulting in a recall of 0.71 for the default class. This indicates that the model is effective at capturing most high-risk borrowers, even in the presence of class imbalance, while still generating more false positives (non-defaults predicted as defaults), which is a more manageable risk for financial firms than missing actual defaulters.

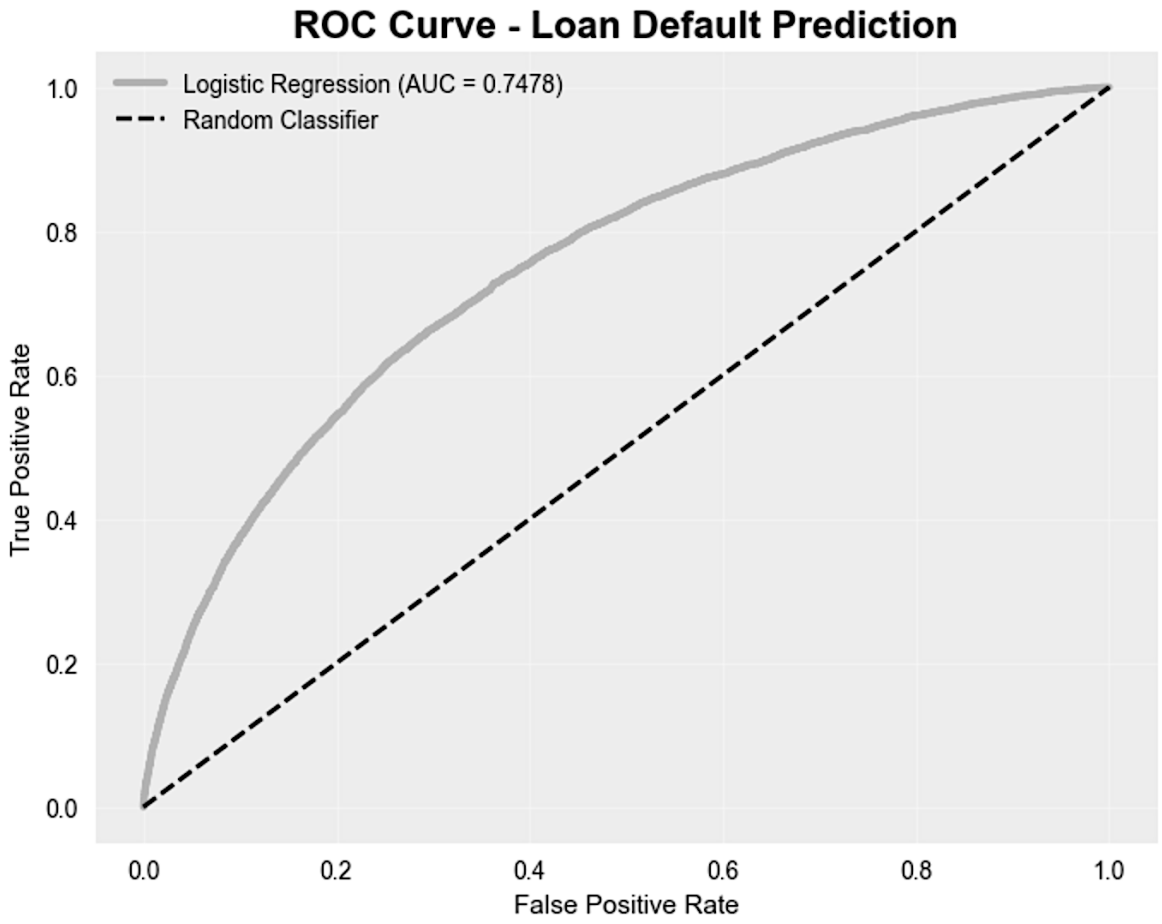

Figure 4

The ROC curve in Figure 3 evaluates the discriminatory ability of the final logistic regression model on the test set. The model achieves an area under the ROC curve (AUC) of 0.748, substantially higher than the 0.50 obtained by random guessing. An AUC of 0.748 is generally regarded as good in credit risk modeling. Specific to Credit Risk Modeling, the optimal AUC range is 0.70-0.8. If the score is below 0.70, the model is considered underfit; if it's over 0.8, the model is considered overfit. This indicates that the model has solid ability to rank borrowers by default risk. If one defaulter and one non-defaulter are randomly selected, there is a 74.8% chance that the model will assign a higher predicted probability of default to the actual defaulter. The performance is comparable to published benchmarks for scorecard models built on similar socioeconomic and loan-level variables.

Baseline Mathematical Model:

To predict loan default probability, we use **logistic regression**, which is appropriate for binary responses. The dependent variable is Default (1 = default, 0 = no default), and the independent variables include both continuous and categorical predictors, such as Age, Income, LoanAmount, CreditScore, Employment Type, Education, and Loan Purpose.

Compact model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{9} \beta_j X_{ij}^{\text{continuous}} + \sum_{k=1}^{15} \beta_{9+k} X_{ik}^{\text{categorical}}$$

All predictors were retained in the model to ensure a comprehensive assessment of factors influencing default. The large dataset (over 255,000 observations) supports estimation of 24 predictors without risk of overfitting. Categorical variables are represented using dummy variables, with reference categories selected to avoid multicollinearity. Variables with lower statistical significance were still included due to their conceptual importance in credit risk evaluation. A general rule of thumb for logistic regression is to have at least 10 "events" (cases in the less frequent outcome category) per predictor variable.

Model: 24 predictors.

Minimum required sample size (based on 10 events/predictor): 24* 10=240 events. With a total sample size of 255,000, this criterion is easily met, even if the outcomes are highly imbalanced. The model is unlikely to be underpowered, which would be a concern with smaller sample sizes.

Predicted probability of default:

$$\hat{p}_i = \frac{1}{1+\exp(-(\beta_0 + \mathbf{X}_i\boldsymbol{\beta}))} \quad \Rightarrow \quad \text{Predicted default chance} = \hat{p}_i \times 100\%$$

This produces a percentage risk score for each borrower, which can be used directly for credit decision-making.

Baseline Model Summary Table:

Variable	Coef	Std Err	z	P¿—z—	[0.025	0.975]
const	-0.4546	0.059	-7.692	0.000	-0.570	-0.339
Age	-0.0395	0.001	-76.226	0.000	-0.041	-0.039
Income	-8.672e-06	1.9e-07	-45.524	0.000	-9.05e-06	-8.3e-06
LoanAmount	4.248e-06	1.04e-07	40.658	0.000	4.04e-06	4.45e-06
CreditScore	-0.0008	4.59e-05	-16.650	0.000	-0.001	-0.001
MonthsEmployed	-0.0089	0.000	-46.255	0.000	-0.010	-0.010
NumCreditLines	0.0872	0.007	13.366	0.000	0.074	0.100
InterestRate	0.0685	0.001	60.100	0.000	0.066	0.071
LoanTerm	0.0002	0.000	0.569	0.569	-0.001	0.001
DTIRatio	0.2643	0.032	8.386	0.000	0.203	0.326
Education_High School	0.0774	0.020	3.888	0.000	0.038	0.116
Education_Master's	-0.1364	0.021	-6.602	0.000	-0.177	-0.096
Education_PhD	-0.1796	0.021	-8.654	0.000	-0.220	-0.139
EmploymentType_Part-time	0.2846	0.021	13.329	0.000	0.243	0.326
EmploymentType_Self-employed	0.2415	0.022	11.222	0.000	0.199	0.284
EmploymentType_Unemployed	0.4403	0.021	21.041	0.000	0.399	0.481
MaritalStatus_Married	-0.2349	0.018	-13.066	0.000	-0.270	-0.200
MaritalStatus_Single	-0.0607	0.017	-3.478	0.001	-0.095	-0.026
HasMortgage_Yes	-0.1435	0.015	-9.854	0.000	-0.172	-0.115
HasDependents_Yes	-0.2513	0.015	-17.218	0.000	-0.280	-0.223
LoanPurpose_Business	0.0347	0.023	1.533	0.125	-0.010	0.079
LoanPurpose_Education	-0.0089	0.023	-0.169	0.865	-0.049	0.041
LoanPurpose_Home	-0.1988	0.024	-8.456	0.000	-0.245	-0.153
LoanPurpose_Other	-0.0139	0.023	-0.607	0.544	-0.059	0.031
HasCoSigner_Yes	-0.2738	0.015	-18.744	0.000	-0.302	-0.245

Table 3

The logistic regression results confirm that most selected features are highly statistically significant predictors of loan default ($p < 0.001$), validating their inclusion in the model. The coefficients align with economic intuition: indicators of stability, such as Age, Income, Credit Score, and Education (Master's/PhD), possess negative coefficients, indicating that as these values increase, the probability of default decreases. In contrast, risk factors such as Interest Rate, DTI Ratio, and Unemployment exhibit positive coefficients, signifying a direct correlation with increased default risk. Notably, Unemployment shows the most substantial positive impact on risk (beta = 0.4403), while having a Co-Signer provides the most potent protective effect (beta = -0.2738). However, the variable Loan Term was found to be statistically insignificant ($p > 0.05$), suggesting that, controlling for other factors, the loan duration does not independently drive default rates in this dataset.

5.2 Refined Logistic Model with Engineered LTI Ratio

To improve predictive performance, we hypothesized that the absolute loan amount was less predictive than the loan's relative burden on the borrower. We engineered a new feature, the **Loan-to-Income (LTI) Ratio**, defined as:

$$\text{LTI Ratio} = \frac{\text{Loan Amount}}{\text{Annual Income}}$$

This derived variable directly measures financial leverage, offering a view of affordability that neither Income nor LoanAmount provides alone. We trained a second logistic regression model including this new term.

Statistical Significance:

The LTI Ratio proved highly significant (P-value = 0.000). The positive coefficient of 0.2112 indicates that as the loan-to-income ratio increases, the probability of default rises significantly, validating standard credit theory.

Metric	Baseline Model	Refined Model (with LTI)	Change
ROC-AUC Score	0.7478	0.7576	+0.0098
Accuracy	66.00%	67.42%	+1.42%
Recall (Defaulters)	0.71	0.7124	+0.0024%
F1-Score	0.33	0.3368	+0.0068%

Table 4

Both models were evaluated at a probability threshold of 0.11 to prioritize sensitivity to default risk. They were also tested at higher thresholds of 0.2, 0.3, 0.4, and 0.5, all of which showed significant drops in ROC AUC, indicating that 0.11 is the optimal threshold.

Removing the 'Loan Term' variable, which was statistically insignificant ($p > 0.05$), did not change our performance metrics like AUC; therefore, we chose to remove it due to its statistical insignificance. Also, after computing the Stepwise AIC, the AIC improved (was lowest) after the Loan Term was removed. Supporting the removal of this term.

Key Findings:

1. **Improved Discrimination:** The ROC-AUC score increased from **0.748 to 0.758**. Crossing the 0.75 threshold is a meaningful benchmark in credit scoring. The score is still within the optimal range of 0.70-0.80.
2. **Better Risk Capture:** The refined model slightly improved Recall (0.7124), indicating it identified more total defaulters than the baseline while simultaneously enhancing overall Accuracy (67.42%).
3. **Feature Value:** The statistical significance and performance gains confirm that **borrower leverage (LTI)** is a critical driver of default risk that was not fully captured by the original variables.

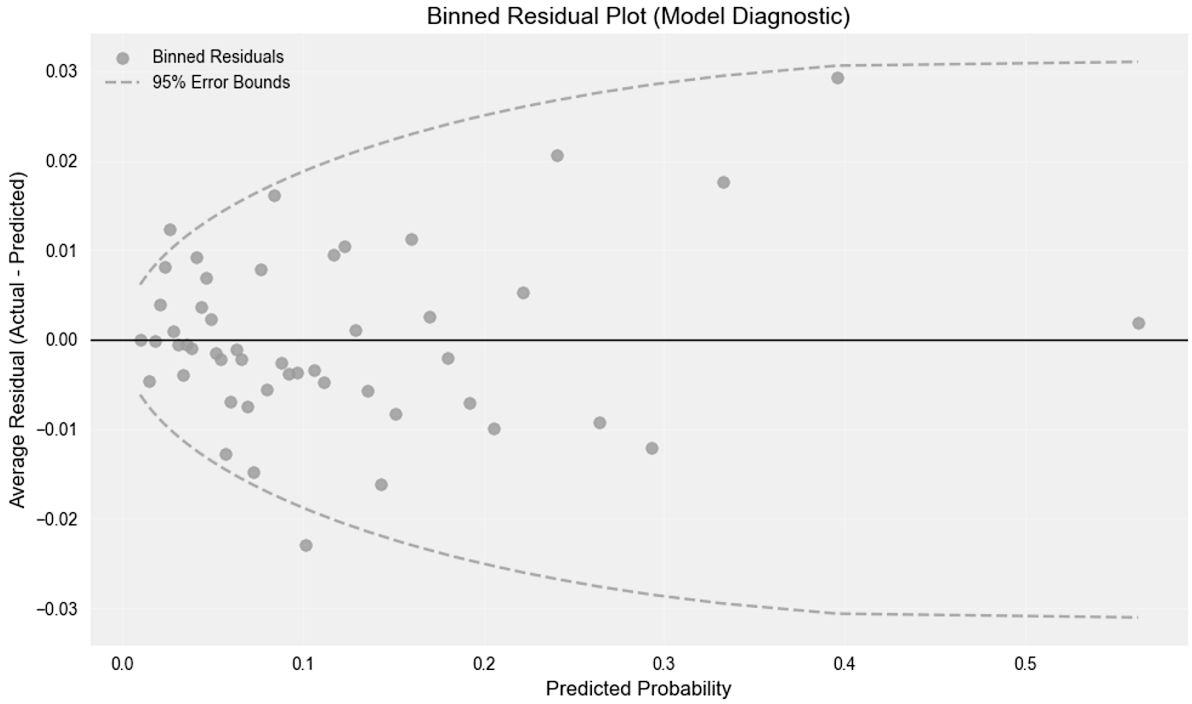

Figure 5:

Binned Residual Plot (Model Diagnostic). To verify the model's statistical assumptions, we plotted the binned residuals (Figure 5). This graph compares the error between observed defaults and our predicted probabilities. Ideally, we want to see random noise, and that is exactly what the plot shows: a random scatter around the zero line with no obvious nonlinear patterns, such as U-shaped or curved features. This supports the linearity assumption of the logit function.

Additionally, nearly all points stay within the red 95% error bounds. The one or two outliers are statistically expected given the large sample size and do not suggest systematic bias. Overall, the plot confirms the model is well-calibrated and that our risk predictions are accurate across the spectrum.

5.3 Random Forest Model

The Random Forest model was evaluated as a potential alternative to logistic regression to capture non-linear relationships. However, its performance metrics indicate that it did not outperform the linear models. With a ROC-AUC score of 0.7473, it fell short of both the baseline logistic regression (0.7478) and the refined LTI model (0.7576). At the decision threshold of 0.11, the Random Forest achieved a recall of 0.74, slightly higher than that of the two logistic models. The accuracy (63.48%) was lower than that of the other models. The lower AUC and accuracy suggest that the underlying risk factors in this dataset are predominantly linear, making the more straightforward, more interpretable logistic regression approach, specifically the refined model with the LTI ratio, the better choice for this credit risk analysis. Complexity should only be "paid for" with a significant boost in performance. Since the Random Forest failed to beat the linear models by a substantial amount, reverting to the simpler model is the better move

Model Conclusion:

The inclusion of the Loan-to-Income (LTI) ratio yielded the most significant predictive improvement, increasing the model's ROC-AUC from 0.7478 (Baseline) to 0.7576. By transforming raw inputs into a meaningful financial ratio, we enhanced the model's ability to rank order risk without increasing complexity. This performance also surpassed the Random Forest approach (0.7473), confirming that a linear model is sufficient for this dataset. Consequently, the Refined Model (LTI) is selected as the final model due to its superior performance and high interpretability, allowing stakeholders to easily justify credit decisions based on borrower repayment capacity.

6-DISCUSSION

The primary objective of this study was to determine if borrower demographic and financial attributes could effectively predict loan default probabilities using logistic regression. Our analysis confirms that these variables are indeed strong predictors, answering the research question proposed in the introduction. The Refined Logistic Regression Model (incorporating the Loan-to-Income ratio) proved to be the better approach, achieving a ROC-AUC of 0.7576, an accuracy of 67.42%, and a recall score of 0.7124.

Interpretation of Results

The study showed two critical insights into the nature of credit risk for this dataset. First, the predictive superiority of the linear model over the non-linear Random Forest (AUC 0.7473) suggests that the relationship between borrower attributes and default risk is mostly linear. This validates the industry-standard preference for logistic regression, not just for its interpretability, but also for its performance on standard credit data. Second, the significance of the engineered Loan-to-Income (LTI) ratio highlights that financial leverage is often a better predictor of default than absolute loan size or income alone. By explicitly modeling the loan size relative to income, we significantly improved the model's rank-ordering performance.

Addressing Model Limitations

While the model successfully identifies 71% of defaulters (Recall Score), the Precision remains low (0.21), a common trade-off in imbalanced datasets where risk aversion (catching bad loans) takes priority over minimizing false alarms.

Future Work and Extensions

To further improve the predictive power and utility of this model, future research could explore the following:

1. **Macroeconomic indicators:** The current dataset provides a snapshot of borrower characteristics. Future work could incorporate external macroeconomic indicators (e.g., national unemployment rates, inflation indices, or GDP growth) to test the model's performance under economic stress scenarios.
2. **Advanced Methods:** While Random Forest did not outperform Logistic Regression, other boosting algorithms, such as XGBoost and LightGBM, often handle class imbalance and subtle non-linearities more effectively than Random Forest.

References

[1] https://blog.workday.com/en-us/how-ai-ml-transforming-banking-capital-markets.html

[2] J. Norling and S. Abdu, "From Data to Decision: Using Logistic Regression to Determine Creditworthiness – A Study on Predictive Modeling: Exploring Logistic Regression for Creditworthiness Assessment," master's thesis, KTH Royal Institute of Technology, 2023. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1833653/FULLTEXT01.pdf

[3] S. Shakthipriya, "A Study on Credit Risk Modelling using Logistic Regression and Decision Trees," SSRN, Working Paper, Aug. 30, 2025. [Online]. Available: https://ssrn.com/abstract=5469667

[4] D. West, "Neural network credit scoring models," Department of Decision Sciences, College of Business Administration, East Carolina University, Greenville, NC 27836, USA, [online]. Available: https://www.researchgate.net/publication/223425357_Neural_Network_Credit_Scoring_Models

[5] D. J. Hand and W. E. Henley, "Statistical Classification Methods in Consumer Credit Scoring: A Review," *Journal of the Royal Statistical Society – Series A (Statistics in Society)*, vol. 160, no. 3, pp. 523–541, 1997.

[6] Z. Atodaria and S. Pentar, "Credit Risk Analysis Using Logistic Regression Modeling," *NIU International Journal of Human Rights*, vol. 9, no. 1, pp. 57–64, 2022.