**Credit Risk Scoring Model Using Logistic Regression**

**1. Introduction**

Credit risk refers to the possibility that a borrower will fail to meet their financial obligations, such as repaying a loan or credit card balance. Financial institutions use credit scoring models to estimate this risk and make informed lending decisions.

In this project, a Credit Risk Scoring Model is developed using Logistic Regression, a statistical method suitable for binary classification problems — here, predicting whether a borrower will default (1) or not default (0).

The model evaluates various borrower characteristics such as income, debt ratio, credit utilization, and payment history to estimate the probability of default. The final goal is to identify high-risk customers and minimize potential financial losses.

**2. Objective**

The main objectives of this project are:

1. To design a credit scoring model that predicts the likelihood of loan default.
2. To use logistic regression for binary classification of customers (default vs. non-default).
3. To evaluate the model’s accuracy and interpret the most influential financial factors.

**3. Methodology**

**3.1 Tools and Libraries**

The model was implemented in Python, using the following key libraries:

* Pandas – for data handling
* NumPy – for numerical computation
* Matplotlib – for visualization
* Scikit-learn (sklearn) – for model building and evaluation

**3.2 Data Generation**

A synthetic dataset was created for 1000 applicants, containing key financial and behavioral attributes:

* Age – applicant’s age
* Income – monthly/annual income
* Debt – outstanding debt
* Payment\_History – number of missed payments
* Credit\_Utilization – proportion of used credit limit
* Debt\_to\_Income – ratio of debt to income (derived feature)

A binary target variable Default was generated:

* 1 → Default
* 0 → No Default

**3.3 Model Design**

The project follows these major steps:

1. **Data Preparation:**
   * Combine features into a DataFrame.
   * Create a derived variable “Debt-to-Income”.
2. **Feature and Target Selection:**
   * Independent variables (X): Income, Debt\_to\_Income, Credit\_Utilization, Payment\_History
   * Dependent variable (y): Default
3. **Train-Test Split:**
   * 80% data used for training, 20% for testing.
4. **Model Training:**
   * Logistic Regression model trained using the training dataset.
5. **Prediction:**
   * The model predicts both class labels (default/no default) and probabilities of default.
6. **Evaluation:**
   * Performance evaluated using Confusion Matrix, Classification Report, and ROC-AUC Score.

**4. Results and Evaluation**

After training and testing, the following results were obtained.

**4.1 Confusion Matrix**

|  | **Predicted No Default (0)** | **Predicted Default (1)** |
| --- | --- | --- |
| **Actual No Default (0)** | 155 | 7 |
| **Actual Default (1)** | 18 | 20 |

**Interpretation:**

* True Negatives (155): Correctly predicted non-defaulters.
* True Positives (20): Correctly predicted defaulters.
* False Positives (7): Incorrectly labeled safe customers as risky.
* False Negatives (18): Missed 18 actual defaulters.

The model performs very well in identifying non-defaulters and reasonably well in identifying defaulters.

**4.2 Classification Report**

| **Metric** | **Class 0 (No Default)** | **Class 1 (Default)** | **Macro Avg** | **Weighted Avg** |
| --- | --- | --- | --- | --- |
| **Precision** | 0.90 | 0.74 | 0.82 | 0.87 |
| **Recall** | 0.96 | 0.53 | 0.74 | 0.88 |
| **F1-Score** | 0.93 | 0.62 | 0.77 | 0.87 |
| **Accuracy** |  |  |  | **0.88** |

**Interpretation:**

* Accuracy (0.88): The model correctly predicts 88% of all cases.
* Precision (Class 1 = 0.74): When predicting default, it’s correct 74% of the time.
* Recall (Class 1 = 0.53): It captures 53% of all true defaulters.
* F1-score (0.62): Balance between precision and recall.

The model performs better in detecting safe borrowers (non-defaulters) but still provides good discrimination for defaulters.

**4.3 ROC-AUC Score**

**ROC-AUC = 0.88**

**Interpretation:**

A ROC-AUC of 0.88 indicates excellent performance — the model can effectively differentiate between high-risk and low-risk applicants.

* A score of 1.0 → perfect separation
* A score of 0.5 → random guessing

Our model (0.88) → Excellent discrimination power.

**4.4 Feature Coefficients**

| **Feature** | **Coefficient** | **Effect** |
| --- | --- | --- |
| **Income** | 0.000008 | Very weak effect. Slightly increases default probability. |
| **Debt\_to\_Income** | 2.1239 | Strong positive effect. Higher debt ratio increases default chance. |
| **Credit\_Utilization** | 5.1438 | Very strong positive effect. High utilization → much higher default risk. |
| **Payment\_History** | 0.8040 | Positive effect. More missed payments increase the likelihood of default. |

**Interpretation:**

* Credit\_Utilization and Debt\_to\_Income are the most critical features influencing default.
* Payment\_History also contributes significantly.
* Income has minimal impact because risk depends more on ratios than absolute income.

**5. ROC Curve Visualization**

The ROC curve plots True Positive Rate vs. False Positive Rate, showing the trade-off between sensitivity and specificity.

The ROC curve for this model lies well above the diagonal baseline, confirming its strong predictive ability.

A graph of a credit failure

AI-generated content may be incorrect.

**6. Discussion**

The logistic regression model achieved:

* 88% overall accuracy
* Strong AUC score (0.88)
* Clear identification of key financial risk indicators

However, the model is slightly biased toward non-defaulters, which is common in credit data due to class imbalance (fewer defaults). Future improvements could include:

* Using more real-world credit data.
* Applying class-balancing techniques (e.g., SMOTE).
* Comparing performance with tree-based models like Random Forest or XGBoost.

**7. Conclusion**

The Credit Risk Scoring Model using Logistic Regression successfully predicts the probability of default based on customers’ financial behavior.

It demonstrates that:

* Borrowers with high credit utilization and high debt-to-income ratios are the most likely to default.
* The model provides reliable and interpretable results for preliminary credit risk analysis.

This approach can be used by financial institutions to assign risk scores, reduce loan defaults, and improve credit decision-making.

**8. Key Learnings**

Through this project, several important technical and analytical insights were gained:

1. **Understanding Credit Risk:**  
   Learned how financial institutions assess borrower risk using quantitative models instead of subjective judgment.
2. **Logistic Regression for Classification:**  
   Understood how logistic regression converts input features into probabilities between 0 and 1 — making it ideal for predicting defaults.
3. **Feature Engineering Importance:**  
   Realized that derived features like **Debt-to-Income Ratio** and **Credit Utilization** provide stronger predictive power than raw variables.
4. **Model Evaluation Metrics:**  
   Gained experience using **Confusion Matrix**, **Classification Report**, and **ROC-AUC Score** to assess model performance from multiple perspectives.
5. **Interpreting Coefficients:**  
   Learned to interpret regression coefficients to identify which factors most influence credit default risk.
6. **Handling Class Imbalance:**  
   Understood that default cases are fewer than non-defaults in real datasets, and this imbalance affects model recall for defaulters.
7. **Practical Credit Insights:**  
   Found that borrowers with **high utilization** and **high debt relative to income** are most prone to default, which matches real-world banking intuition.
8. **Data-Driven Decision Making:**  
   Experienced how machine learning supports better lending decisions, reducing human bias and financial loss.