Credit Scoring Report

Project Title: Retail Credit Scoring Model

Prepared by: Arsen Tagibekov

Date: April, 2025

This final report and credit scoring system were built with support from ChatGPT (OpenAI), used as a co-pilot for idea structuring, Python logic validation, and formatting refinement. The project reflects my own analytical judgment, assumptions, and execution, but benefited from AI-based structuring and iterative feedback throughout.

Project Overview:

This project simulates the design and implementation of a credit scoring system, replicating the workflow followed by banks and fintech lenders in retail credit underwriting. The primary objective is to build a model that predicts the Probability of Default (PD) of retail loan applicants using historical lending data and statistical learning techniques.

Tools Used:

- Python (Pandas, Scikit-learn, Seaborn Matplotlib)
- Jupyter Notebook

Dataset:

- Source: Kaggle's "Give Me Some Credit" competition dataset

https://www.kaggle.com/datasets/brycecf/give-me-some-credit-dataset

- Composition: 15,000+ anonymized loan applications (cs-training.csv)
- Variables: Credit utilization, monthly income, past delinquencies, number of open credit lines, age, dependents, and default indicator (target). More detailed description of variables is provided by the table below.

Table 1. Description of variables of "Give Me Credit" dataset

Source: https://www.kaggle.com/datasets/brycecf/give-me-some-credit-dataset

Variable Name	Description	Type
SeriousDlqin2yrs	**Person experienced 90 days past due delinquency or worse**	**Y/N**
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Data Preprocessing

- Removed outliers (e.g., age < 18, credit utilization > 150%)
- Imputed missing values using median (for income) and mode (for dependents)
- Created two separate datasets: one raw (for EDA), one scaled (for modeling)

EDA (Exploratory Data Analysis)

- Identified key correlations: higher delinquency and debt ratio associated with default
- Detected class imbalance: ~8% default rate
- Visualized distribution of income, age, delinquencies, and utilization

Figure 1. Default Distribution

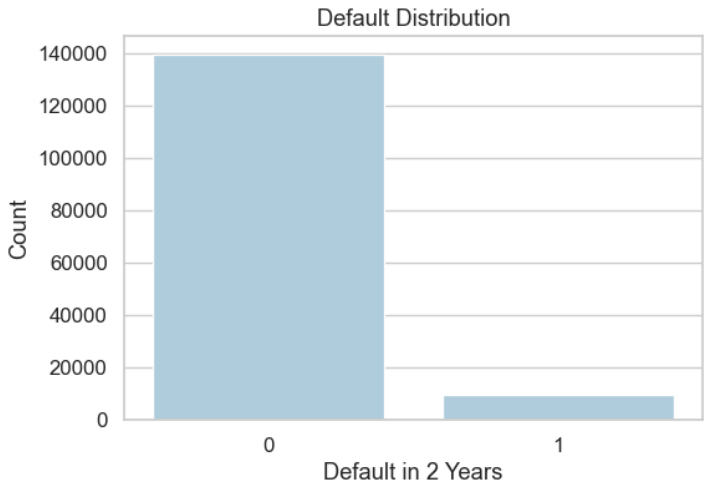

Figure 1 shows the distribution of the target variable, specifically, whether a borrower defaulted within the next two years. The dataset is highly imbalanced, with fewer than 10% of applicants defaulting, which is typical in consumer lending portfolios.

Figure 2. Confusion Matrix

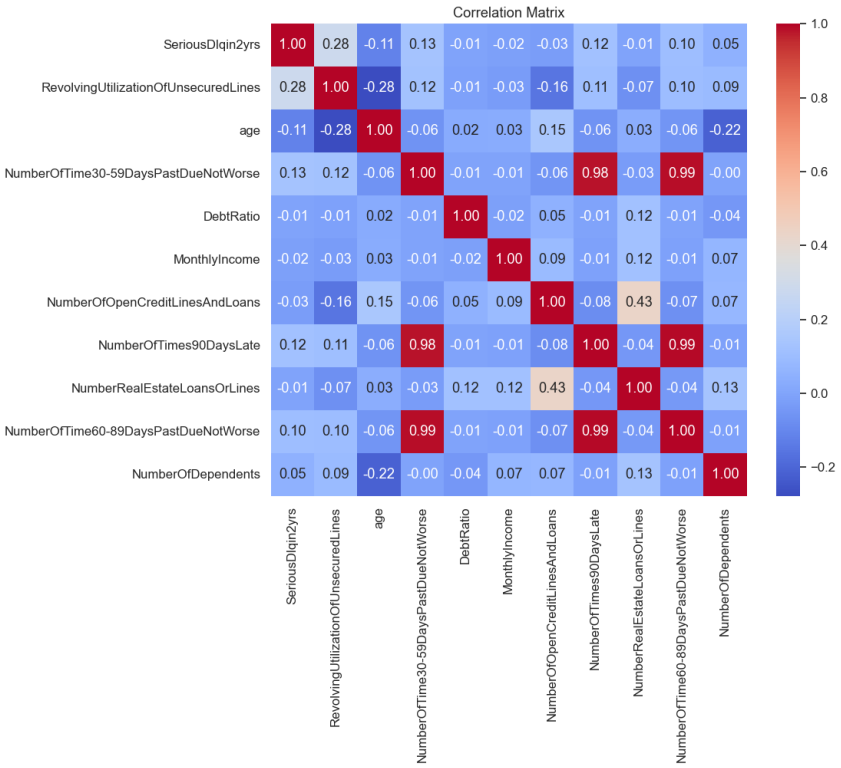

Figure 2 presents the correlation matrix of numeric features in the dataset. Default (SeriousDlqin2yrs) shows the strongest positive correlations with late payment features, particularly NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse, and NumberOfTimes90DaysLate, which confirms their predictive importance in modeling credit risk.

Figure 3: Revolving Utilization vs. Default

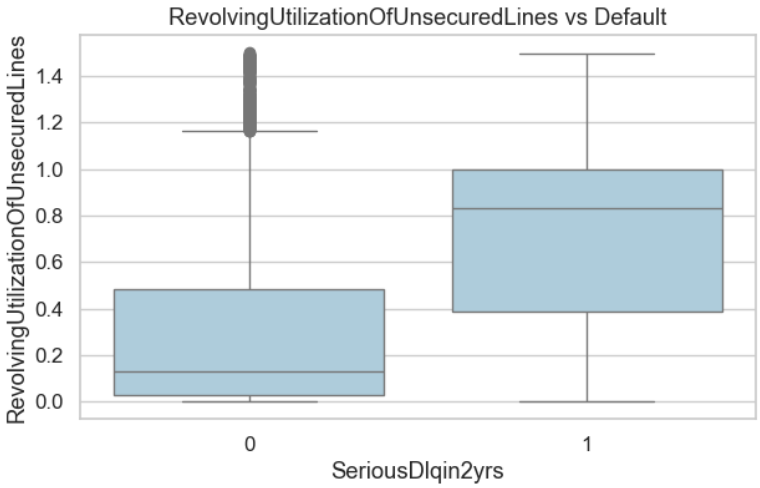

Figure 3 displays a boxplot of revolving credit utilization across defaulters (1) and non-defaulters (0). Borrowers who defaulted tend to have significantly higher credit utilization, suggesting that high dependency on unsecured credit is a strong predictor of financial distress.

Figure 4: Age vs. Default

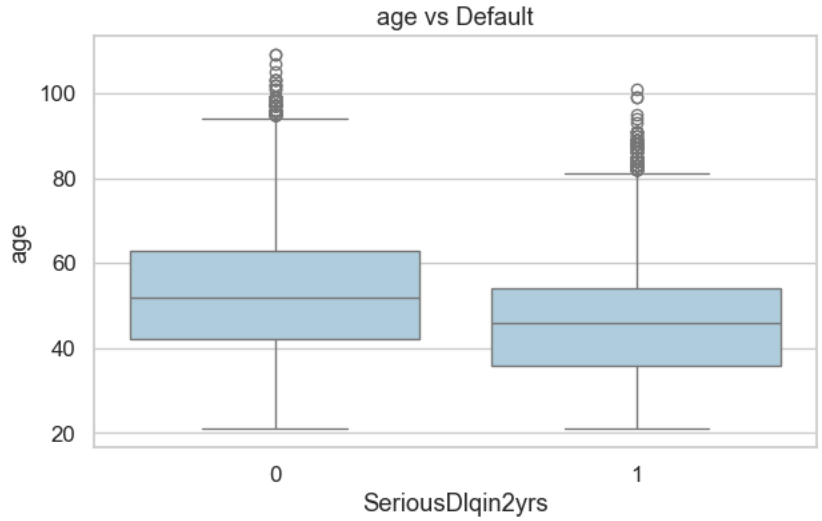

Figure 4 illustrates the distribution of borrower age for defaulters (1) and non-defaulters (0). Defaulting borrowers are generally younger, with a median age notably lower than that of non-defaulters (~45 vs. 50). This supports the inclusion of age as a predictive feature, as younger borrowers may have less financial stability or credit maturity.

Figure 5: Debt Ratio vs. Default

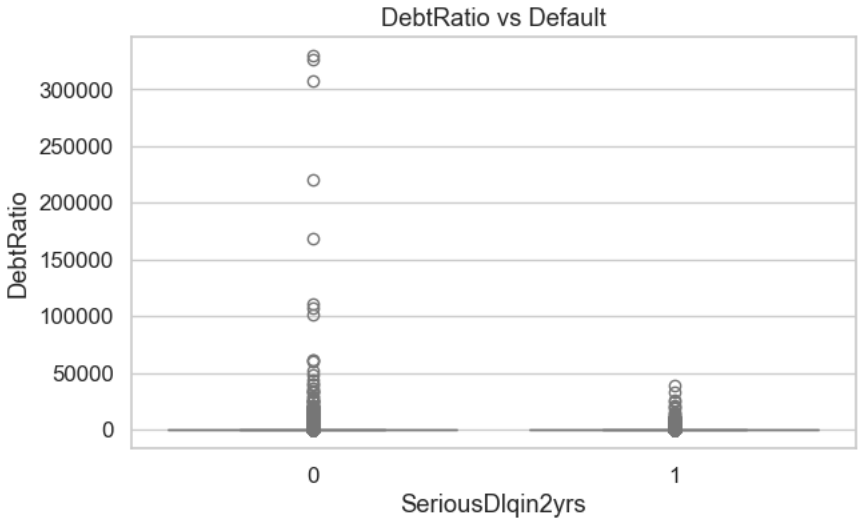

Figure 5 compares debt ratios between defaulters (1) and non-defaulters (0). While extreme outliers are present in both groups, the overall trend suggests that borrowers who default

often exhibit higher debt burdens relative to income. This reinforces the importance of incorporating Debt-to-Income (DTI) as a derived risk feature.

Figure 6: Monthly Income vs. Default

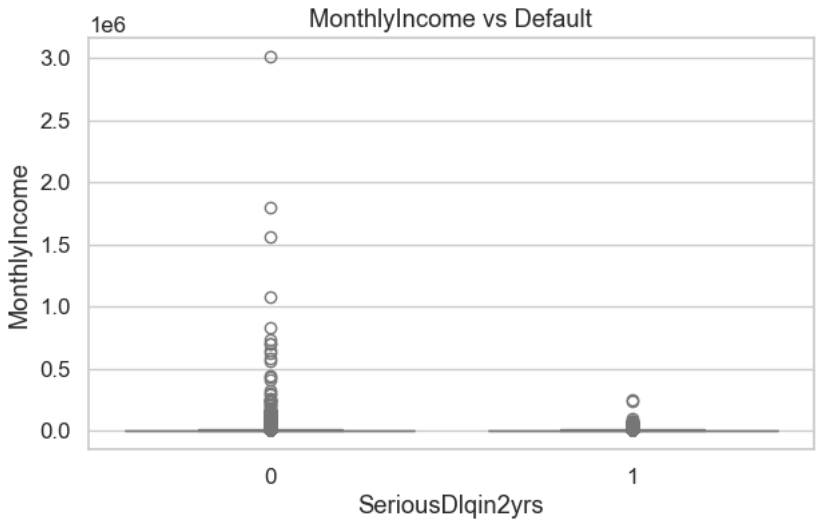

Figure 6 shows the distribution of monthly income among defaulters (1) and non-defaulters (0). While both groups exhibit extreme outliers, the chart indicates that defaulting borrowers generally report lower income levels. This supports income's role as a predictor and justifies log-transformation to reduce skew and improve model stability.

Figure 7: 30-59 Day Delinquency vs. Default

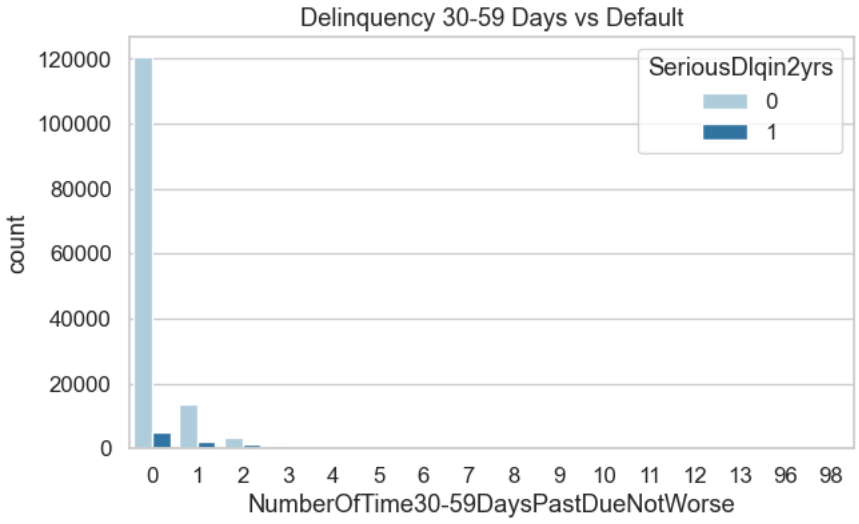

Figure 7 illustrates the relationship between short-term delinquencies and loan default. Borrowers with even one occurrence of being 30-59 days past due are disproportionately likely to default. This confirms that recent late payments are strong early warning indicators of credit risk and should be prioritized in modeling.

Figure 8: Default Rate by Income Quintile

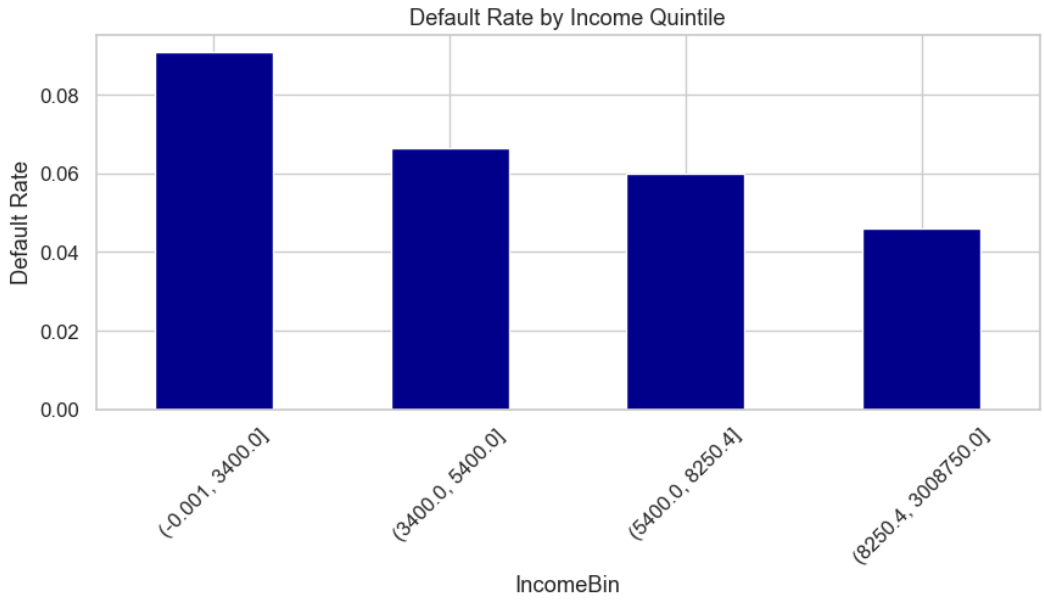

Figure 8 shows default rates across four income-based quintiles. A clear inverse relationship is observed: lower-income borrowers consistently exhibit higher default rates, with the lowest-income group facing nearly double the risk of the top earners. This justifies income segmentation as a key driver in retail credit policy design.

Feature Engineering

- Created Debt-to-Income Ratio (DTI)
- Engineered binary late payment flags (30+, 60+, 90+ days late)
- Derived credit activity score by summing delinquencies
- Created age and income bands
- Applied log transformation to monthly income

Modeling

- Trained two classifiers: Logistic Regression and Random Forest
- Evaluation Metrics:
    - Confusion Matrix
    - Classification Report (Precision, Recall, F1)
    - AUC-ROC Score
- Random Forest showed superior performance with strong separation between risk groups

Figure 9: ROC Curve Comparison of Logistic Regression and Random Forest

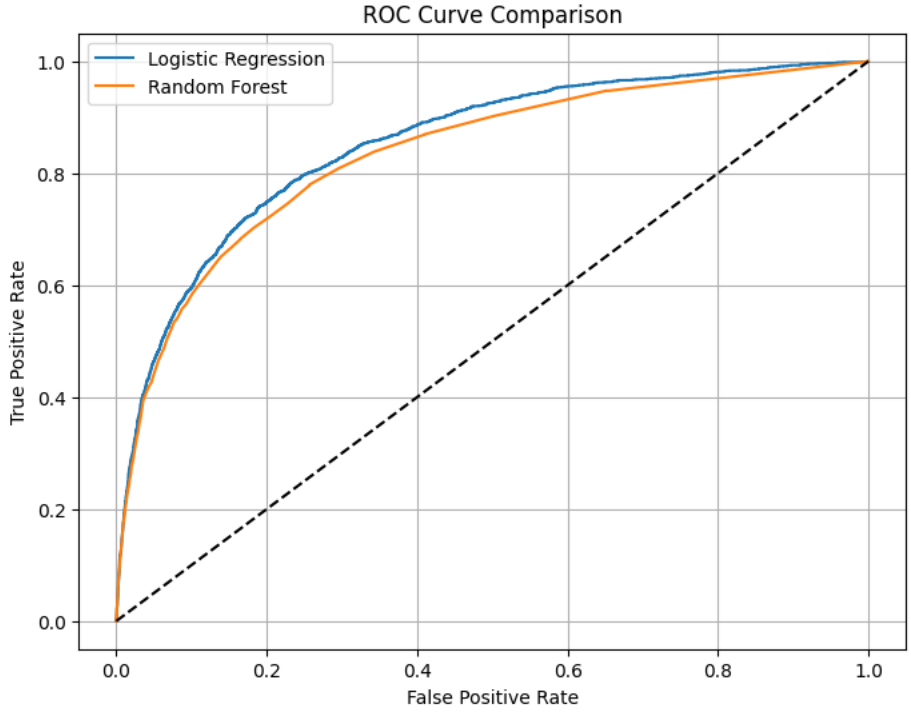

Figure 9 compares the ROC (Receiver Operating Characteristic) curves of Logistic Regression and Random Forest models. Both models perform well above the random baseline (dotted line), with Logistic Regression achieving slightly higher AUC across most thresholds. This validates both models' predictive power while highlighting Logistic Regression's robustness and explainability for risk segmentation.

Probability Segmentation

- Modeled Probability of Default for each applicant
- Grouped applicants into 5 Risk Bands:
    - A: PD 0.00-0.05
    - B: 0.05-0.10
    - C: 0.10-0.20
    - D: 0.20-0.35
    - E: 0.35-1.00
- Observed default rates by band confirmed monotonicity:
    - A: ~2% | E: ~50%+

Figure 10: Default Rate by Risk Band

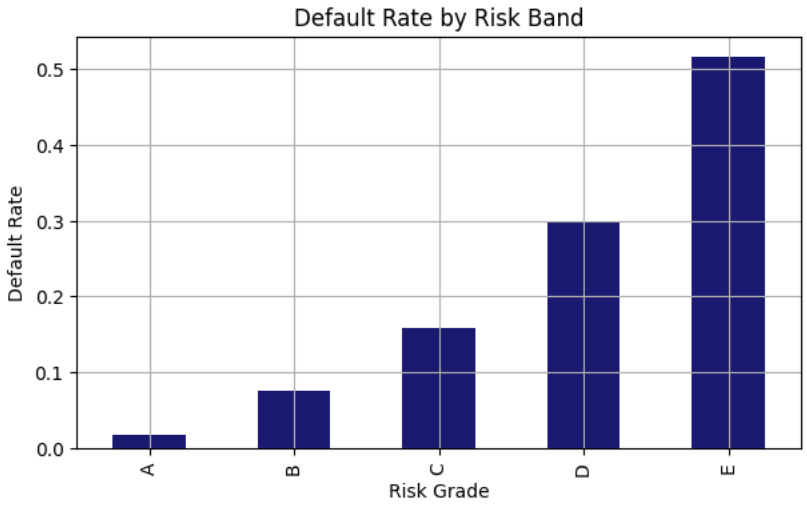

Figure 10 displays the actual default rates within five predicted risk bands (A to E) based on Probability of Default (PD). The clearly increasing trend confirms that the model is well-calibrated: higher PD groups experience significantly higher default rates, with Band E borowers defaulting over 50% of the time. This segmentation enables tiered lending decisions, such as approvals, rejections, or pricing adjustments.

Business Interpretation

- Bands A and B are strong candidates for automatic approval
- Band C may trigger manual review
- Band D requires stricter terms (e.g., higher APR, collateral)
- Band E is unbendable based on projected PD

This segmentation mimics real credit policy design, enabling automated, data-driven credit decisions.

Outcome:

This project delivers a complete credit scoring pipeline capable of powering underwriting workflows in real-world lending contexts. The model accurately differentiates between risk groups, while engineered features and PD-based segmentation support explainability and policy development.