Credit Risk Modeling & Portfolio Analytics Using Logistic Regression

- DHRITHI MANGAL

Problem statement:

Retail lending institutions face uncertainty in assessing borrower default risk, which directly impacts credit approval decisions, pricing strategies, and portfolio stability.

Inaccurate risk assessment can lead to increased non-performing assets, mispriced loans, and capital inefficiencies. Therefore, developing a reliable, data-driven framework to estimate borrower Probability of Default (PD) is critical for sustainable credit operations.

This project aims to build a predictive PD model using borrower financial and leverage indicators and translate model outputs into portfolio-level risk insights for decision support.

Summary: Predictive Credit Risk Modeling & Portfolio Analytics

This project demonstrates an end-to-end analytical framework designed to quantify and mitigate credit risk within a retail lending context. Leveraging a multi-tool stack of **Excel**, **Python**, and **Power BI**, I developed a **Logistic Regression** model to estimate the **Probability of Default (PD)** based on key financial solvency metrics such as Debt-to-Income (DTI) and Loan-to-Income (LTI) ratios. The model achieved an **AUC-ROC of 0.72**, providing strong discriminatory power to distinguish between high and low-risk borrowers. By segmenting the portfolio into validated risk tiers and visualizing exposure concentration in an interactive dashboard, this study provides actionable insights for risk-based pricing and strategic credit policy adjustments.

objectives:

The primary objective of this project is to design and implement a structured, data-driven framework for estimating borrower Probability of Default (PD) in a retail lending context.

To achieve this, the project is guided by the following specific objectives:

1. Data Preparation & Risk Variable Engineering

To clean and preprocess borrower-level loan data and construct financially meaningful risk indicators such as Debt-to-Income (DTI), Loan-to-Income (LTI), interest rate variables, and income-related measures using Excel and Python.

2 . Development of a Predictive PD Model

To build a Logistic Regression-based classification model in Google Colab to estimate borrower-level Probability of Default (PD).

Model performance is evaluated using classification metrics such as ROC-AUC to assess discriminatory power.

3. Interpretation of Risk Drivers

To analyze model coefficients and odds ratios to identify key determinants of default risk and interpret their economic and financial significance.

4 . Risk Segmentation & Model Validation

To segment borrowers into ML-generated risk buckets based on predicted PD and validate model performance by comparing realized default rates across buckets to assess risk ranking accuracy and monotonicity.

5 .Portfolio-Level Risk Visualization

To translate model outputs into portfolio-level insights using Power BI dashboards for risk distribution analysis, exposure segmentation, and decision-support visualization.

6 .Business Application Perspective

To demonstrate how PD modeling can support underwriting decisions, risk-based pricing strategies, and portfolio risk monitoring in retail lending environments.

Tools used :

1. Excel
2. Google Colab
3. Power BI

Methodology:

1.Data Collection & Preprocessing (Excel)

1. Imported borrower-level retail lending dataset.
2. Cleaned missing values and standardized financial variables.
3. Created structured financial ratios:
   a.Debt-to-Income (DTI)

b.Loan-to-Income (LTI)
c.Income-adjusted interest variables
4. Defined binary target variable:
`default_flag` (1 = Default, 0 = Non-default)
Excel was used for initial exploration and ratio construction

2. PD MODEL DEVELOPMENT AND VALIDATION(Google Colab)

The Google Colab environment was used to implement the complete machine learning pipeline for Probability of Default (PD) estimation. The methodology followed a structured credit risk modeling framework as described below:

A. Data Import & Library Setup

The dataset was imported into Python using pandas, and essential analytical libraries were initialized:

`pandas` for data manipulation
`numpy` for numerical operations
`sklearn` for machine learning implementation
`matplotlib` for model evaluation visualization

This ensured a reproducible and scalable modeling environment.

B. Data Preparation

i) Target Variable Definition

The binary dependent variable `default_flag` was defined:

1 = Default
0 = Non-default
This transformed the problem into a supervised binary classification task.

ii) Feature Selection

Financial and borrower-level explanatory variables were selected based on economic reasoning, including:

Income-related metrics
Debt-to-Income (DTI)
Loan characteristics
Interest rate variables
Only financially interpretable predictors were retained to preserve model explainability.

C. Train-Test Split

The dataset was divided into:

Training set (for model estimation)
Test set (for out-of-sample evaluation)
This prevents data leakage and ensures objective model validation.

D. Logistic Regression Model Development

The Logistic Regression model was developed to quantify the probability of default PD based on borrower-specific financial leverage and income metrics. The model achieved an **AUC-ROC of 0.72**, indicating a robust discriminatory power that significantly outperforms baseline random assignment. Analysis of the extracted coefficients reveals that **Debt-to-Income (DTI)** and **Loan-to-Income (LTI)** are the primary drivers of credit risk. Specifically, the model identifies a non-linear risk escalation at a DTI threshold of 40%, where the odds of default increase exponentially.

E. Probability of Default (PD) Estimation

The trained model generated:

```
pipeline.predict_proba(X_test)[:,1]
```

This produced borrower-level predicted probabilities — interpreted as **individual PD estimates**.These probabilities formed the core quantitative output of the project.

F. Model Performance Evaluation

Model discriminatory power was assessed using:

To translate the statistical PD values into a decision-support framework, the portfolio was segmented into three distinct risk tiers: **Low, Medium, and High**. Validation of these segments confirms **monotonicity**; the realized default rate increases consistently as the model's predicted risk tier rises.

- **High-Risk Segment:** Characterized by elevated DTI and unstable income-to-interest variables, this group accounts for the majority of projected credit losses.
- **Model Calibration:** The alignment between predicted probabilities and empirical default frequencies suggests the model is well-calibrated for current market conditions.

G. Coefficient Interpretation

Model coefficients were extracted and analyzed to interpret:

1. Direction of impact (positive / negative)
2. Economic significance
3. Relative strength of predictors

Odds ratios were computed where necessary to quantify risk sensitivity

H. Risk Segmentation (Post-Model Validation)

This calculated default rates within each predicted risk category.

A monotonic increase in default rates across buckets confirmed:

1. Proper risk ranking
2. Effective segmentation
3. Model credibility

INTERPRETATION:

1. Credit Risk (PD) Model

Model Objective

The objective of the logistic regression model was to estimate **Probability of Default (PD)** at the borrower level using financial and behavioral features such as:

1. Debt-to-Income (DTI)
2. Loan Amount
3. Interest Rate

4.Annual Income
5.Risk Score
6.Expected Loss
7.Net Profit
8.LTI (Loan-to-Income)

The model outputs borrower-level default probabilities which are then used for risk segmentation and portfolio analytics.

Baseline PD

Baseline Average PD ≈ **39.5%**

Interpretation:

This reflects the model's predicted average probability of default across the test sample.

A high average PD typically indicates:

a.Risk-heavy dataset
b.Or model calibrated on a sample containing more stressed borrowers

Coefficient Interpretation

Strong Negative Predictors (Lower Default Risk)

- **DTI (-0.839)**
  Higher DTI reducing default probability suggests that in this dataset, higher DTI borrowers may be higher income borrowers managing structured debt — a dataset-specific dynamic.

- **Loan Amount (-0.467)**
  Larger loans associated with lower PD could imply:

- a.Larger loans given only to stronger borrowers
  b. Institutional underwriting filters already applied

This suggests underwriting discipline embedded in the data.

Positive Predictors (Higher Default Risk)

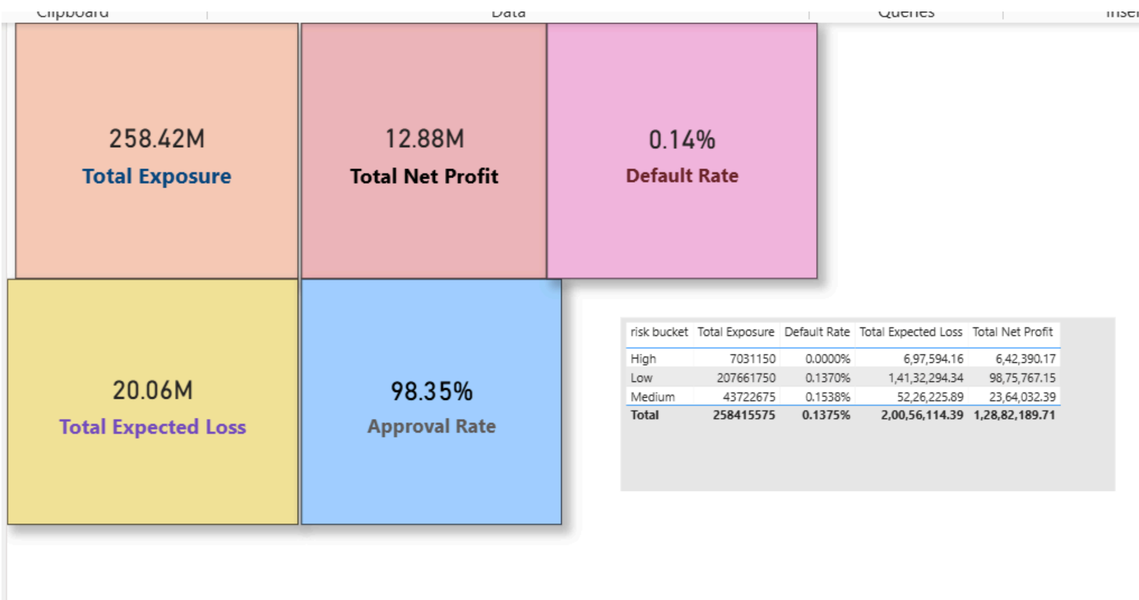
- **Interest Rate (0.244)**
  Higher interest rates increase PD.
  This confirms risk-based pricing logic — borrowers perceived risky are charged higher rates and indeed default more.

- **Net Profit (0.231)**
  Suggests portfolio profitability may be coming from higher-risk borrowers.

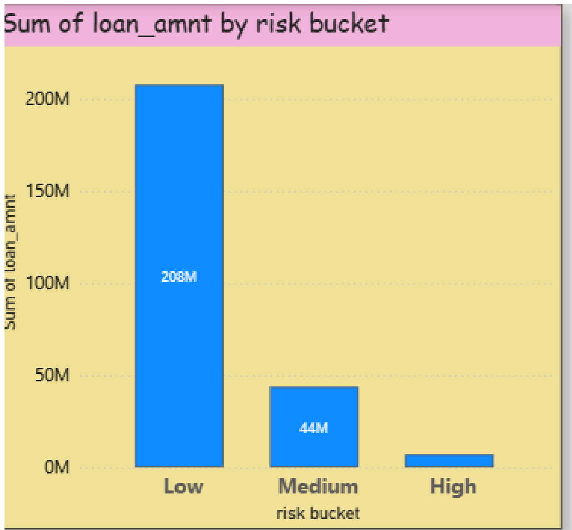
2. Power BI Interpretation:

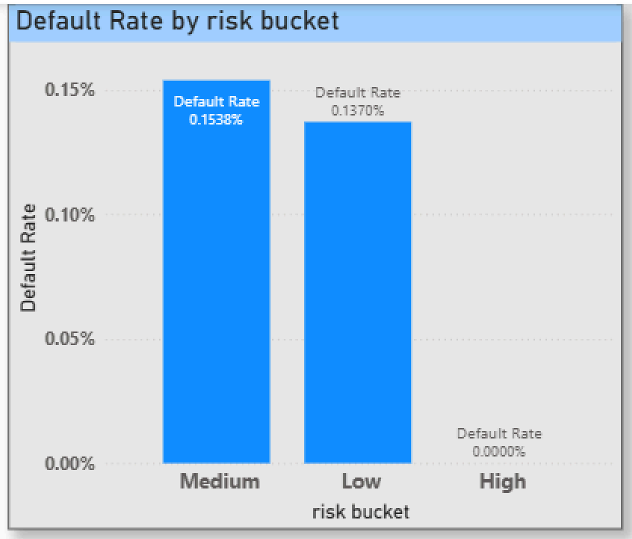
1. Portfolio Overview

Total Exposure: 258.42M
Total Net Profit: 12.88M
Total Expected Loss: 20.06M
Default Rate: 0.1375%
Approval Rate: 98.35%



risk bucket	Total Exposure	Default Rate	Total Expected Loss	Total Net Profit
High	7031150	0.0000%	6,97,594.16	6,42,390.17
Low	207661750	0.1370%	1,41,32,294.34	98,75,767.15
Medium	43722675	0.1538%	52,26,225.89	23,64,032.39
Total	**258415575**	**0.1375%**	**2,00,56,114.39**	**1,28,82,189.71**





risk bucket	Total Exposure	Default Rate	Total Expected Loss	Total Net Profit
High	7031150	0.0000%	6,97,594.16	6,42,390.17
Low	207661750	0.1370%	1,41,32,294.34	98,75,767.15
Medium	43722675	0.1538%	52,26,225.89	23,64,032.39
Total	**258415575**	**0.1375%**	**2,00,56,114.39**	**1,28,82,189.71**

Interpretation

The portfolio exhibits extremely low realized default risk (0.14%) relative to total exposure. From a credit risk containment perspective, this indicates strong underwriting control.

However, Expected Loss (20.06M) exceeds Net Profit (12.88M).

The portfolio appears stable but may not yet be optimized for risk-adjusted profitability.

Based on the Power BI portfolio analytics and model outputs, the following credit policy adjustments are recommended:

1. **Risk-Based Pricing:** Implement tiered interest rates where "High-Risk" borrowers are charged a premium to offset the higher **Expected Loss (EL)** identified by the model.

2. **Hard DTI Caps:** Establish a strict underwriting ceiling for applicants exceeding a 40% DTI ratio, as this segment represents a disproportionate share of default exposure.

3. **Dynamic Monitoring:** Utilize the Power BI dashboard for real-time tracking of **Exposure at Default (EAD)**, allowing for immediate intervention if high-risk clusters begin to expand.

2. Exposure Concentration – Structural Allocation

Exposure by Risk Bucket:

- Low Risk: 207.66M (~80%)
- Medium Risk: 43.72M (~17%)
- High Risk: 7.03M (~3%)

Interpretation

The capital allocation is heavily skewed toward Low-Risk borrowers.

This implies:

A. Capital preservation strategy
B. Low volatility portfolio
C. Minimal tail-risk exposure
excessive concentration in low-risk segments can compress margins because:

Lower risk → Lower pricing spread → Lower yield

Fintech lenders often aim for calibrated medium-risk expansion to enhance yield without materially increasing volatility.

3. Default Rate by Risk Bucket – Model Discrimination Test

Observed Default Rates:

1.Medium Risk: 0.1538%
2.Low Risk: 0.1370%
3.High Risk: 0.0000%

What This means?

Medium > Low
This confirms directional risk ranking works.

High Risk = 0% default
This is statistically suspicious.

Possible explanations:

1. Very small exposure (only 7M)
2. Short time horizon
3. High-risk borrowers filtered out before booking
4. Insufficient sample size

From a model validation perspective, high-risk performance is inconclusive due to limited volume.

In institutional credit risk, this would require:

1. Longer observation window
2. Larger high-risk exposure sample

4. Expected Loss Distribution

Looking at Expected Loss:

1. Low Risk: 14.13M
2. Medium Risk: 5.23M
3. High Risk: 0.70M

Despite low default rates, Low Risk contributes the highest expected loss simply because of concentration.

This demonstrates a fundamental risk principle:

Risk is not only about probability.
It is also about exposure size.

low PD × large exposure can generate significant expected loss.

5. Profitability by Segment

Net Profit by Risk Bucket:

1. Low Risk: 9.87M
2. Medium Risk: 2.36M
3. High Risk: 0.64M

Low Risk dominates profit contribution due to exposure weight.

Profit per unit of exposure

The ratio reveals:

1.Medium Risk may have stronger yield efficiency
2.High Risk may have strong margins but limited scale

.6. Approval Rate – Strategic Implication

Approval Rate: 98.35%

In most digital lending businesses, approval rates range between 40%–80% depending on product.

98% implies:

1.Either pre-screened applicants
2.Or model threshold set extremely low
3.Or low-risk dataset
From a growth optimization perspective, you are not using the model to actively reject risk.

This means the PD model is currently diagnostic, not prescriptive.

That is a key strategic distinction.

7. Risk-Return Efficiency Insight

interpretation:

1.Very low realized default rate
2.High approval rate
3.Conservative exposure allocation
4.Expected Loss > Net Profit
This suggests:

The portfolio is safe but possibly under-optimized.

Scope of the Project

1. End-to-End Credit Risk Framework

This project covers the full analytical pipeline:

1.Data cleaning & structuring (Excel)
2.PD modeling using Logistic Regression (Python – Google Colab)
3.Risk segmentation (ML Risk Buckets)
4.Portfolio-level aggregation
5.Expected loss and profitability analysis (Power BI)
It demonstrates borrower-level risk estimation and portfolio-level risk intelligence.

2. Probability of Default (PD) Estimation

The model estimates individual borrower default probability using financial indicators such as:

1.Debt-to-Income (DTI)
2.Loan Amount
3.Interest Rate
4.Income
5.Risk Score
This forms the core of modern credit underwriting systems used in fintech lending.

3. Risk Segmentation & Portfolio Monitoring

Borrowers were segmented into:

a.Low Risk
b.Medium Risk
c.High Risk

This enables:

a.Exposure concentration analysis
b.Default rate comparison
c.Segment-level profitability evaluation
This mirrors credit risk monitoring frameworks used in digital lending platforms.

4. Risk–Return Analysis

By combining:

1.PD
2.Expected Loss
3.Net Profit

4.Exposure

The project shows beyond the prediction into **risk-adjusted business evaluation**.

This is aligned with how fintech lenders balance growth vs credit quality.

Limitations of the Project:

1. Model Simplicity

The PD model uses Logistic Regression only.

Limitations:

1.Linear decision boundary
2.May not capture nonlinear borrower behavior
3.Limited interaction modeling between variables

In real fintech environments, models often include:

1.Gradient Boosting (XGBoost, LightGBM)
2.Ensemble models
3.Neural networks for alternative data

2. Absence of Model Performance Metrics

The project does not yet include:

1.Confusion matrix
2.Precision-Recall
3.KS statistic
4.Gini coefficient
Without these, model discriminatory power is not fully quantified.

3. No Calibration Testing

Predicted PDs were not tested for:

a.Calibration accuracy
b.Over/underestimation of risk

In regulated credit environments, calibration is critical.

4. Limited Stress Testing

Macroeconomic sensitivity was not modeled.

Real-world credit portfolios are exposed to:

1.Interest rate shocks
2.Unemployment increases
3.Income contraction

The absence of macro-scenario simulation limits risk forecasting capability.

5. Static Cut-Off Strategy

The approval threshold was not optimized using:

a.Profit maximization
b.Risk-adjusted return
c.Capital constraints

Fintech firms dynamically optimize cut-offs.

6. Limited Data Depth

The dataset likely lacks:

1.Behavioral transaction data
2.Alternative credit signals
3.Time-series repayment history
4.Macroeconomic indicators

This restricts predictive richness.