LOAN DEFAULT RISK PREDICTION

Credit Risk Analytics using Machine Learning

Submitted by

Omkar Bhalekar

omkarbhalekar2003@gmail.com

February 2026

1. Overview of Credit Risk in Financial Institutions

Credit risk is one of the most significant risks faced by financial institutions worldwide. It refers to the probability that a borrower will fail to meet their contractual debt obligations, resulting in financial loss for the lending institution. Effective credit risk management is a foundational pillar of sustainable banking, directly influencing profitability, capital adequacy, and regulatory compliance.

Financial institutions encounter credit risk across a broad spectrum of products, including retail loans, mortgages, credit cards, corporate lending, and trade finance. When credit risk is poorly managed, it can lead to elevated levels of non-performing assets (NPAs), increased loan loss provisioning, reduced capital buffers, and, in severe cases, systemic financial instability.

1.1 Quantification of Credit Risk

The global banking industry, guided by the Basel II and Basel III regulatory frameworks, quantifies credit risk using the Expected Loss (EL) formula:

Expected Loss (EL) = PD x LGD x EAD

Where each component carries a specific business meaning:

- PD (Probability of Default): The likelihood that a borrower will default on their obligations within a specified time horizon, typically one year.
- LGD (Loss Given Default): The proportion of the outstanding exposure that is not recovered following a default event, expressed as a percentage.
- EAD (Exposure at Default): The total outstanding balance at the time of default, representing the bank's maximum financial exposure.

Under Basel regulations, banks must maintain Risk-Weighted Assets (RWA) calculations that account for credit risk, and are required to maintain a minimum Capital Adequacy Ratio (CAR) of 8% under Basel II and 10.5% under Basel III (including the conservation buffer). Accurate default prediction is therefore not merely an analytical exercise — it is a regulatory necessity.

1.2 Types of Credit Risk

Credit risk encompasses several interrelated sub-types that together define the full spectrum of lending risk:

Risk Type	Description
Default Risk	Borrower fails entirely to repay outstanding loan principal or interest obligations.
Credit Spread Risk	Risk arising from changes in borrower credit quality, reflected in changing risk premiums.
Downgrade Risk	Risk of deterioration in borrower credit rating, increasing cost of capital and reducing loan value.
Counterparty Risk	Risk associated with the failure of a financial counterparty in derivatives and inter-bank lending.
Concentration Risk	Excessive exposure to a single borrower, sector, or geographic region, amplifying systemic vulnerability.

Table 1: Classification of Credit Risk Types in Financial Institutions

2. Importance of Default Prediction and Risk Modeling

Accurate default prediction is a critical capability for financial institutions seeking to maintain healthy loan portfolios and sustainable profitability. The ability to identify high-risk borrowers before credit is extended enables institutions to adjust loan terms, increase collateral requirements, or decline applications — all of which directly reduce expected credit losses.

The consequences of inadequate default prediction cascade across multiple dimensions of institutional performance, including profitability, regulatory standing, investor confidence, and long-term solvency.

2.1 Financial Consequences of Prediction Errors

In machine learning-based credit risk models, two primary types of prediction error occur, each carrying distinct business consequences:

Error Type	Description	Business Impact
False Negative	A defaulting borrower is incorrectly predicted as creditworthy.	Direct capital loss, increased provisioning, NPA accumulation, regulatory scrutiny.
False Positive	A creditworthy borrower is incorrectly flagged as a default risk.	Opportunity cost, reduced loan book growth, potential customer dissatisfaction.

Table 2: Financial Consequences of Model Misclassification

> *In credit risk management, False Negatives are significantly more damaging than False Positives. Approving a loan to a borrower who subsequently defaults results in direct, quantifiable capital loss. Rejecting a creditworthy borrower represents only an opportunity cost — a foregone revenue opportunity rather than an actual loss.*

This asymmetry of consequences is why this project prioritizes Recall for the default class (class 1) over overall accuracy. The threshold optimization conducted in the modeling phase directly addresses this business requirement by maximizing the detection of actual defaulters.

2.2 Portfolio-Level Impact

Beyond individual loan decisions, the cumulative effect of default prediction quality shapes the overall health of a lending institution's portfolio. A model with low recall for defaults will allow high-risk loans to accumulate, leading to:

- Deterioration in loan portfolio quality over successive lending cycles
- Increased loan loss provisions reducing reported profitability
- Elevated risk-weighted assets increasing capital consumption
- Regulatory intervention and increased compliance scrutiny
- Reduced investor confidence and higher cost of wholesale funding

3. Role of Machine Learning in Credit Risk Assessment

Traditional credit risk assessment relied heavily on rule-based systems, manual underwriting procedures, and bureau-based credit scores such as FICO. While effective to a degree, these approaches are limited by their inability to capture non-linear relationships between variables, process large datasets efficiently, or adapt dynamically to changing borrower behavior patterns.

Machine learning introduces a fundamentally different paradigm for credit risk assessment — one that leverages historical data patterns to generate probabilistic default predictions at scale, with greater consistency and objectivity than manual underwriting.

3.1 Advantages of Machine Learning over Traditional Methods

The application of supervised learning algorithms to credit risk offers several concrete advantages:

- Pattern Recognition: ML algorithms can identify complex, non-linear relationships between borrower characteristics and default outcomes that traditional regression models may miss.
- Scalability: Automated scoring systems can evaluate thousands of loan applications per hour with consistent methodology.
- Continuous Learning: Models can be retrained periodically on new data to maintain predictive relevance as borrower behavior evolves.
- Segmentation Capability: Risk scoring allows for granular risk-based pricing strategies, enabling institutions to tailor loan terms to individual risk profiles.

3.2 Why Logistic Regression Was Selected

This project employs Logistic Regression as the primary classification algorithm, a choice well-aligned with both the technical requirements of binary classification and the practical requirements of credit risk modeling in regulated environments:

- Interpretability: Logistic Regression produces coefficient estimates that can be directly translated into business insights about feature importance.
- Probabilistic Output: The model generates probability scores P(Default), enabling flexible threshold-based decision-making.
- Regulatory Acceptance: Banking regulators (including the Basel Committee) favour models whose predictions can be explained and audited.
- Baseline Validity: Logistic Regression serves as a robust baseline against which more complex models can be benchmarked.

4. Dataset Description and Exploratory Data Analysis

4.1 Dataset Overview

The dataset used in this project is the Lending Club Loan Dataset (2007-2018), publicly available on Kaggle. Lending Club operates as a peer-to-peer lending marketplace in the United States, connecting borrowers with investors. The dataset contains rich borrower-level and loan-level information, making it ideal for credit risk modeling research.

Dataset Attribute	Value
Source	Lending Club (Kaggle) — Accepted Loans 2007-2018
Initial File Size	~1.6 GB (CSV format)
Sample Used for Analysis	50,000 rows (initial exploration)
Final Modeling Dataset	44,006 finalized loans (active loans removed)
Total Raw Features	151 columns
Features Used for Modeling	16 (after encoding)
Observed Default Rate	20.52% (9,028 defaults / 44,006 total)

Table 3: Dataset Overview and Characteristics

4.2 Target Variable Engineering

The original loan_status column contained multiple outcome categories including fully paid, charged off, current, default, and various late-payment stages. To construct a clean binary classification target variable, the following decision logic was applied:

- Default (Target = 1): Loans with status 'Charged Off' or 'Default' — representing confirmed financial losses.
- Non-Default (Target = 0): Loans with status 'Fully Paid' — representing successful repayment completion.

- Excluded: Loans with status 'Current', 'Late (16-30 days)', 'Late (31-120 days)', or 'In Grace Period' — these represent active or transitional loans with undetermined final outcomes.

Excluding active loans is critical to prevent data leakage and ensure the model is trained exclusively on confirmed outcomes. This produced a final modeling dataset of 44,006 records with a default rate of 20.52%, which closely reflects realistic consumer lending portfolio characteristics.

4.3 Class Distribution Analysis

Class	Count	Proportion
0 — Fully Paid	34,978	79.48%
1 — Default	9,028	20.52%
Total	**44,006**	**100%**

Table 4: Target Variable Class Distribution

While this imbalance is moderate (approximately 4:1 ratio), it is sufficient to cause standard Logistic Regression models to over-predict the majority class. This necessitates threshold optimization to improve recall for the minority (default) class — a core focus of the modeling phase described in Section 6.

4.4 Key Risk Driver Analysis

<u>4.4.1 Interest Rate vs. Default</u>

Statistical analysis reveals a strong positive relationship between loan interest rate and default probability. Loans that ultimately defaulted carried significantly higher interest rates compared to fully repaid loans.

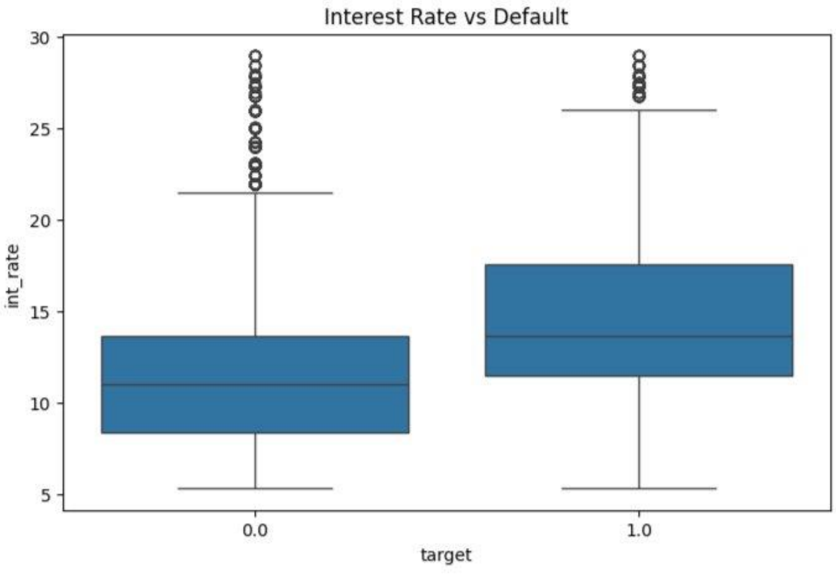

Figure 1: Boxplot of Interest Rate Distribution by Loan Default Status (0 = Fully Paid, 1 = Defaulted)

The boxplot clearly demonstrates that defaulted loans (target = 1) exhibit a higher median interest rate (~13-14%) compared to fully paid loans (~10-11%), with a statistically significant upward shift for defaults. This is consistent with risk-based pricing theory: higher-risk borrowers are charged higher rates, and those higher-risk borrowers default more frequently.

Statistical Measure	Value
T-Statistic	68.78
P-Value	< 0.001 (effectively 0)
Interpretation	Highly significant — interest rate is a strong predictor of default

Table 5: Hypothesis Test Results — Interest Rate vs. Default (Independent Samples t-test)

4.4.2 FICO Credit Score vs. Default

FICO credit scores, which represent borrower creditworthiness on a scale of 300-850, demonstrate a strong inverse relationship with default probability. Borrowers who subsequently defaulted had measurably lower FICO scores at loan origination.

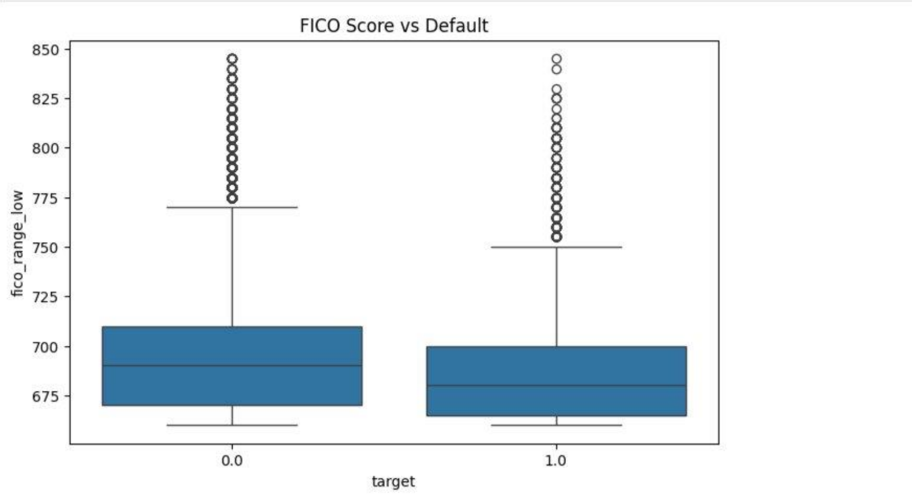

Figure 2: Boxplot of FICO Score Distribution by Loan Default Status (0 = Fully Paid, 1 = Defaulted)

The FICO score boxplot confirms that defaulted loans originate from borrowers with lower median credit scores (~675-680) compared to fully repaid borrowers (~690). The distributional separation is consistent and statistically robust.

Statistical Measure	Value
T-Statistic	-29.10 (negative = defaulters have lower FICO)
P-Value	2.17 x 10^(-184)
Interpretation	Extremely significant — FICO score is a powerful predictor of default risk

Table 6: Hypothesis Test Results — FICO Score vs. Default (Independent Samples t-test)

4.4.3 Loan Grade vs. Default Rate

Lending Club assigns borrowers a loan grade (A through G) based on their credit profile, with Grade A representing the highest creditworthiness and Grade G the lowest. Analysis of default rates by grade reveals a clear and monotonic relationship:

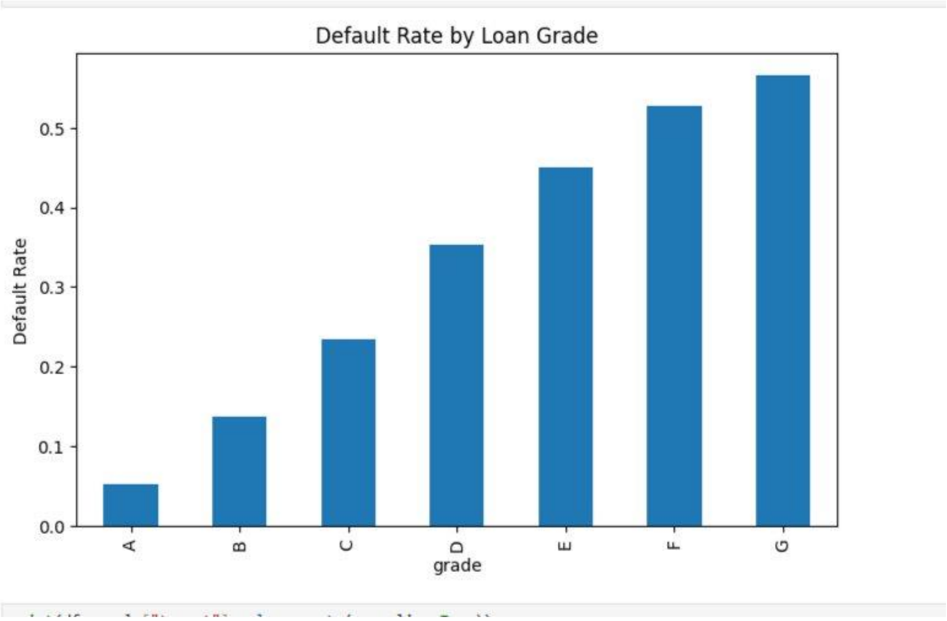

Figure 3: Default Rate by Loan Grade — Monotonic Increase from Grade A (Safest) to Grade G (Riskiest)

Loan Grade	Default Rate	Risk Interpretation
A	5.3%	Prime borrowers — very low risk
B	13.8%	Near-prime — moderate-low risk
C	23.5%	Sub-prime — moderate risk
D	35.2%	Elevated risk — requires careful screening
E	44.9%	High risk — significant provisions required
F	52.7%	Very high risk — majority will default
G	56.5%	Extreme risk — majority will default

Table 7: Default Rate by Loan Grade — Risk Segmentation Analysis

The monotonic increase in default rate from Grade A (5.3%) to Grade G (56.5%) validates the internal consistency of Lending Club's grading methodology. Grade G borrowers are over 10 times more likely to default than Grade A borrowers. Loan grade is one of the strongest predictors of default and is included as a key feature in the predictive model.

5. Key Financial Risk Metrics and Business Impact

5.1 Model Evaluation Metrics in Credit Risk Context

Standard machine learning evaluation metrics take on specific financial interpretations in the credit risk domain. Understanding these interpretations is critical to selecting an appropriate operating threshold and communicating model performance to business stakeholders.

Metric	Statistical Definition	Credit Risk Interpretation

Accuracy	(TP + TN) / Total	Misleading in imbalanced datasets; a model predicting all non-default achieves ~80% accuracy.
Recall (Sensitivity)	TP / (TP + FN)	Most critical metric — proportion of actual defaulters correctly identified. Low recall = missed defaults = capital loss.
Precision	TP / (TP + FP)	Proportion of predicted defaults that are truly defaults. Low precision = good borrowers unnecessarily rejected.
F1-Score	2 x (P x R) / (P + R)	Harmonic mean of Precision and Recall. Useful single metric for imbalanced classes.
ROC-AUC	Area under ROC curve	Probability that model ranks a random defaulter higher than a random non-defaulter. AUC = 0.72 is a solid baseline.

Table 8: Credit Risk Interpretation of Machine Learning Evaluation Metrics

5.2 Threshold Optimization and Business Alignment

In credit risk applications, the default probability threshold used for classification is a business decision, not a statistical one. The standard 0.50 threshold maximizes accuracy but is inappropriate for risk-sensitive applications. This project analyzed three operating points:

Model Configuration	Recall	Precision	F1	Accuracy	AUC
LR — Threshold 0.50 (Baseline)	0.15	0.56	0.23	80%	0.72
LR — Threshold 0.30	0.44	0.41	0.42	76%	0.72
LR — class_weight balanced	0.64	0.33	0.44	66%	0.72
LR — Threshold 0.26 (Final)	**0.52**	**0.38**	**0.44**	**73%**	**0.72**

Table 9: Comparative Model Performance Across Threshold Configurations

The final selected operating point — threshold = 0.26 — was determined through Precision-Recall curve analysis and represents a deliberate balance between risk sensitivity and operational feasibility. At this threshold, the model correctly identifies 52% of actual defaulters (940 out of 1,806), compared to only 15% (266) at the default threshold — a 3.5x improvement in default detection.

6. Model Building, Evaluation, and Results

6.1 Feature Selection and Engineering

Feature selection was guided by both domain knowledge and the requirement to avoid data leakage. Post-loan performance variables (e.g., total payments received, recovery amounts, last payment date) were deliberately excluded as they constitute look-ahead information unavailable at the time of a loan approval decision.

Feature	Type	Credit Risk Rationale
loan_amnt	Numerical	Higher loan amounts increase absolute loss exposure.
int_rate	Numerical	Strong positive correlation with default (validated statistically).
annual_inc	Numerical	Income level is a fundamental determinant of repayment capacity.
fico_range_low	Numerical	Credit score — strong negative correlation with default (validated statistically).
dti	Numerical	Debt-to-Income ratio measures debt burden relative to income.
installment	Numerical	Monthly payment burden relative to income capacity.
grade	Categorical (A-G)	Strongest categorical predictor — monotonic default rate increase.
term	Categorical	Loan term (36 or 60 months) affects repayment trajectory.
home_ownership	Categorical	Housing status reflects financial stability and collateral capacity.

Table 10: Selected Features and Credit Risk Rationale

6.2 Preprocessing Pipeline

A structured preprocessing pipeline was applied prior to model training:

- Missing Value Imputation: The debt-to-income ratio (dti) contained a single missing value, imputed using median substitution — a robust approach for financial ratios that resists the influence of outliers.
- Categorical Encoding: Grade, term, and home_ownership were one-hot encoded with drop_first=True to prevent multicollinearity (dummy variable trap), yielding 16 total features after encoding.
- Feature Scaling: StandardScaler was applied to normalize numerical features to zero mean and unit variance, ensuring Logistic Regression convergence stability.
- Train-Test Split: An 80/20 stratified split was applied, preserving the 20.52% default rate in both training (35,204 samples) and test (8,802 samples) sets.

6.3 Final Model Performance

The final Logistic Regression model with optimized threshold of 0.26 was evaluated on the held-out test set of 8,802 observations:

	Predicted: Non-Default (0)	Predicted: Default (1)
Actual: Non-Default (0)	5,455 (True Negatives)	1,541 (False Positives)
Actual: Default (1)	866 (False Negatives)	940 (True Positives)

Table 11: Confusion Matrix — Final Model (Logistic Regression, Threshold = 0.26)

6.4 ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve provides a threshold-independent assessment of model discrimination capability. The area under the ROC curve (AUC) summarizes the model's ability to distinguish between default and non-default cases across all possible thresholds.

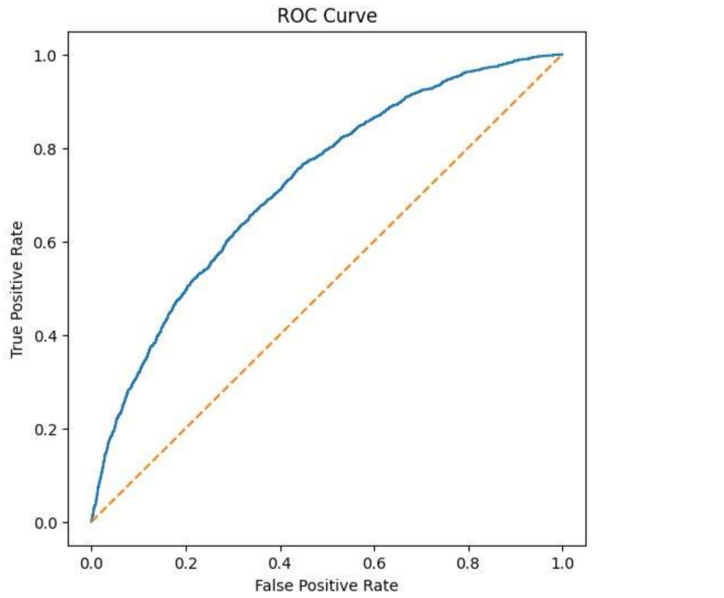

Figure 4: ROC Curve — Logistic Regression Model (AUC = 0.72)

The ROC curve demonstrates that the model performs substantially above the diagonal random classifier baseline across all threshold values. An AUC of 0.72 indicates that, given a randomly selected defaulter and non-defaulter, the model assigns a higher probability of default to the actual defaulter 72% of the time. In retail credit risk benchmarking, AUC values of 0.65-0.75 represent acceptable baseline performance for a Logistic Regression model on raw features.

7. Limitations of Predictive Models in Lending

Acknowledging the limitations of predictive models is as important as documenting their performance. No credit risk model is universally applicable, and the following constraints apply to the current implementation:

7.1 Technical Modeling Limitations

- Linear Assumption: Logistic Regression assumes a linear relationship between features and log-odds of default. Complex non-linear interactions between variables may be inadequately captured, limiting predictive power.
- Feature Scope: The model uses only nine input features. Real-world credit scoring systems typically incorporate hundreds of features including behavioral repayment history, employment stability, transaction data, and social indicators.
- No Cross-Validation: Model performance was evaluated on a single train-test split. Cross-validation would provide a more robust estimate of generalization performance and reduce variance in evaluation metrics.
- No Probability Calibration: While the model generates probability estimates, their calibration was not formally tested. Calibration is important for pricing applications where the exact probability of default is used to set interest rates.

- No Ensemble Methods Tested: More powerful models such as Random Forest, Gradient Boosted Trees (XGBoost/LightGBM), or neural networks were not evaluated, which may yield substantially higher predictive performance.

7.2 Data and Sampling Limitations

- Historical Bias: The training data reflects lending decisions and economic conditions from 2007-2018, a period that includes the Global Financial Crisis. Models trained on this period may not generalize well to different economic environments.
- Survivorship Bias: The dataset contains only loans that were originated — not applications that were rejected. This selection bias means the model has not been trained on the full distribution of applicant risk profiles.
- Sample Size Limitation: Analysis was conducted on a 50,000-row sample for computational efficiency. Training on the full 2.2+ million loan dataset could potentially improve model stability and performance.

8. Ethical Considerations and Regulatory Awareness

The deployment of machine learning models in credit lending decisions carries significant ethical responsibility. Credit decisions determine individuals' access to financial services, housing, education, and economic mobility. Errors in these decisions — particularly systematic ones caused by biased models — can perpetuate and amplify existing socioeconomic inequalities.

8.1 Risk of Algorithmic Bias

While protected attributes such as race, gender, religion, or national origin are not included in this model, proxy variables that correlate with these characteristics may still introduce indirect discrimination:

- Annual Income: May reflect historical wage disparities across demographic groups.
- Home Ownership: Ownership rates differ significantly across racial and ethnic groups in the United States due to historical redlining and discriminatory lending practices.
- Geographic Indicators: ZIP code or region can serve as a proxy for race in neighbourhoods with high racial segregation.

The phenomenon of algorithmic bias through proxy variables has been extensively documented and is a central concern in responsible AI for financial services. Even a model that never explicitly incorporates protected attributes can produce discriminatory outcomes through correlated proxies.

8.2 Regulatory Framework

Credit risk models deployed in regulated banking environments must comply with a range of legal and supervisory requirements:

- Equal Credit Opportunity Act (ECOA): Prohibits credit discrimination based on race, gender, religion, national origin, age, or marital status.
- Fair Housing Act (FHA): Prohibits discriminatory practices in mortgage lending.
- SR 11-7 (Model Risk Management): Federal Reserve guidance requiring model validation, documentation, and ongoing monitoring of all models used in bank decision-making.
- GDPR (Article 22): In European jurisdictions, requires that automated decisions affecting individuals be explainable and subject to human review.

8.3 Responsible AI Principles for Credit Models

This project incorporates and advocates for the following responsible AI principles in lending:

- Transparency: Model inputs, logic, and decision thresholds should be fully documented and auditable.
- Explainability: Individual loan decisions should be explainable to borrowers, specifying which factors contributed to an adverse outcome.
- Fairness Auditing: Before production deployment, models should be tested for disparate impact across demographic groups using metrics such as demographic parity, equalized odds, and calibration by group.
- Human Oversight: Algorithmic scores should serve as inputs to — not replacements for — credit officer judgment, particularly for borderline cases.
- Model Governance: Deployed models require continuous monitoring for performance drift, distributional shift, and potential bias amplification over time.

> *Models should support, not replace, human judgment in credit lending. The risk of unchecked algorithmic lending is the automation of historical inequalities at scale and speed. Responsible deployment requires active governance, regular auditing, and meaningful human accountability for all credit decisions.*

9. Conclusion

This research report presents a comprehensive end-to-end credit risk analytics project, spanning the full pipeline from problem definition and data exploration to statistical hypothesis testing, machine learning modeling, and ethical analysis.

The project successfully demonstrates that borrower default risk in peer-to-peer lending can be meaningfully predicted using structured loan-level features. Statistical analysis confirmed that interest rate, FICO credit score, and loan grade are strong, statistically significant predictors of default. The Logistic Regression model achieved an ROC-AUC of 0.72, and threshold optimization improved default recall from a baseline of 15% to 52% — representing a 3.5x improvement in high-risk borrower detection.

The project also critically examines the limitations of the modeling approach and acknowledges the ethical dimensions of deploying predictive models in credit decisions. The analysis concludes that while machine learning offers genuine value in risk quantification, responsible deployment requires fairness auditing, human oversight, and ongoing model governance.

Project Deliverable	Outcome
Risk Driver Analysis	Interest rate, FICO, and loan grade validated as strong predictors ($p < 0.001$)
Hypothesis Testing	Both hypotheses confirmed with extremely high statistical significance
Baseline Model	Logistic Regression — ROC-AUC = 0.72
Threshold Optimization	Threshold = 0.26 — Recall improved from 15% to 52%
Ethics and Governance	Proxy bias, regulatory compliance, and responsible AI principles addressed

Table 12: Summary of Project Achievements