Credit Scorecard Model Development

Elizabeth Mwania,

mwania.m.elizabeth@gmail.com

29 September 2025

Contents

1 Introduction

This report presents predictive credit risk model development for evaluating new applicants' creditworthiness in a finance ecosystems. It leverages transaction data from suppliers and marketplaces to analyse applicants' historical financial behaviours. The model's primary objectives are to assign appropriate credit limits based on applicants' risk profiles and to enhance credit evaluations by using merchant transaction data to predict the likelihood of future defaults. Logistic regression has been adopted upon comparing with other machine learning models including Random Forest, Gradient Boosting and Decision Tree, achieving 70.2% AUC through systematic application of statistical and machine learning methodologies. The project successfully transforms transactional behavioral data into actionable credit scoring models, enabling data-driven lending decisions while maintaining regulatory compliance and business interpretability. Through strategic feature engineering, algorithmic optimization, and adherence to fairness principles, the solution achieves optimal balance between predictive performance and operational requirements. This document details the end-to-end methodology, technical implementation decisions, and business implications of the behavioral credit scoring approach.

1.1 Project Summary

The modern financial services landscape demands sophisticated approaches to credit risk assessment, particularly in merchant lending where traditional credit bureau data may be insufficient or unavailable. This project addresses the critical challenge of evaluating small business creditworthiness using transactional behavioral data through statistical and machine learning techniques.
The project's success stems from three key innovations:

1. **Behavioral Feature Engineering:** Capturing comprehensive business operational patterns across multiple temporal dimensions

2. **Model Selection Framework:** Balancing predictive performance with interpretability requirements for regulatory compliance

3. **AI Governance Implementation:** Ensuring fairness and regulatory compliance throughout the model development lifecycle

1.2 Challenge

Traditional credit scoring models face significant limitations in the merchant lending space. Small businesses often lack extensive credit histories, making conventional bureau-based assessments inadequate for comprehensive risk evaluation. This creates a critical market gap where potentially viable businesses may be denied credit due to insufficient traditional credit data, while lenders struggle to accurately assess risk without comprehensive behavioral insights. The emergence of digital payment systems and transaction processing platforms has created opportunities to understand business behavior through transactional patterns. However, translating this rich behavioral data into reliable credit risk indicators requires sophisticated analytical approaches that can capture the nuanced relationships between operational patterns and repayment capacity.

1.3 Objectives

This project was designed to achieve four primary business objectives:

1. **Improved Risk Assessment:** Improve prediction accuracy for merchant default probability using comprehensive behavioral transaction data
2. **Operational Efficiency:** Enable automated credit decisions while maintaining appropriate risk controls and governance
3. **Portfolio Optimization:** Provide clear risk segmentation for differentiated pricing strategies and approval frameworks
4. **Regulatory Compliance:** Ensure adherence to fair lending regulations while maintaining complete model interpretability

1.4 Model Type

This project was designed to achieve four primary business objectives:

1. **Classification Model:** Binary classification for default prediction
2. **Algorithm:** Logistic Regression with Weight of Evidence (WoE) binning technique, after comparing with ML models
3. **Implementation:** Developed in Python with scikit-learn, Advanced feature selection using IV

2 Data Description

The dataset utilized in this project consists of transactional and lending performance data spanning multiple temporal dimensions- containing 67279 transactions with 1,290 customers. This represents typical merchant behavioral patterns, loan characteristics, and repayment outcomes, making it an ideal foundation for advanced risk modeling. Key features include:

1. Transactional volume and frequency analysis
2. Comprehensive income, spending behaviour assessment
3. Behavioural pattern recognition
4. Optimized Predictive Inputs based on Information Value (IV)

2.1 Data Characteristics

- The data had no missing values, though it contained notable zero-amount transactions which were treated as valid business activities.
- Numerical features exhibit right-skewed distributions where average face amounts hover around 9,600 KES with extremes reaching 298,000 KES, while borrowers owe an average of 11,500 KES, slightly higher than the face amount due to interest or fees, with a wide range up to 138,000 KES.

- Repayment performance is strong overall, with a mean repayment rate of 94.6% and median of 100%, indicating that most loans are fully repaid. However, a distinct 4% segment with poor repayment creates class imbalance that requires careful handling during model development.

- Analysis revealed distinct temporal patterns with peak activity mid-week and during specific months (May-July). Loan repayment performance was best on Tuesdays (96.55%) and Wednesdays (95.31%), and worst on Fridays (84.40%). Weekends showed minimal transaction activities.

2.2 Data Quality

- Ensured data completeness eliminating columns with more than 90% (threshold) missing values.

- Ensured data consistency by validating temporal relationships across all observation windows.

- Adhering to business rules i.e., remove some peak seasons which. might hype transaction patterns compared to expected merchant behavior cycles and seasonal variations

- Outlier management by binning values within 25th - 75th percentiles

3 Target Variable Definition

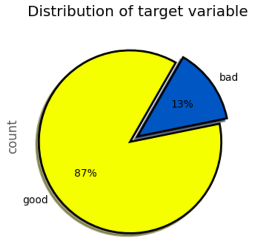
To define a target variable, a customer is classified as "bad" if they cannot repay loans in full within 10 days after the due date, as defined by the following business logic:

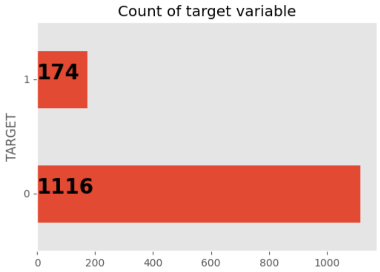
$$\text{TARGET} = 1 \text{ (Bad) if: } \begin{cases} \text{Loan Repayment Rate} < 100\% \quad \text{OR} \\ \text{Loan Repay Days} > 10 \end{cases}$$

Conversely, a "good" customer demonstrates full repayment (100%) within the 10-day grace period. This creates a binary target variable where 0 = GOOD and 1 = BAD, resulting in the following target distribution:

- **Good Customers (0):** 87% of population - demonstrating strong repayment discipline

- **Bad Customers (1):** 13% of population - requiring enhanced risk management





This design supports scalability and future adaptability, with configurable thresholds allowing business-driven adjustments as market conditions or risk appetite evolves. The 13% bad rate, while creating class imbalance, provides meaningful risk representation that was technically addressed using SMOTE (Synthetic Minority Oversampling Technique) to achieve balanced representation during model training.

4 Feature Engineering

Feature engineering represents the critical bridge between raw transactional data and actionable predictive insights. The approach systematically transforms merchant behavioral patterns into risk signals using recency analysis, time-windowed aggregations, statistical measures, and derived relational features that provide meaningful predictive power.

4.1 Recency Features

Recency features focus on the temporal relationship between transaction timing and loan origination. The relative timing analysis calculates days before loan by determining the difference between first funded date and transaction date, creating a timeline relative to loan origination where negative values indicate pre-loan transactions and positive values indicate post-loan activity. Key recency indicators engineered are:

- **Days Since Last Transaction:** Minimum `days_before_loan` per customer-merchant pair, indicating recent engagement levels
- **Days Since First Transaction:** Maximum `days_before_loan`, representing the depth of historical business relationship
- **Transaction Days Range:** Difference between first and last transaction days, representing the span of active business relationship

4.2 Time-Windowed Transaction Features

Merchants exhibit distinct temporal behavioral patterns that reflect business health, seasonal variations, and operational consistency. The following time-windowed strategy was used to capture both short-term volatility and long-term stability patterns.

- **7-day window:** Captures immediate business activity patterns and short-term volatility indicators
- **1-month window:** Reflects typical business cycle patterns and monthly operational performance
- **3-month window:** Indicates seasonal business trends and medium-term operational stability
- **6-month window:** Demonstrates long-term business viability and sustainable growth patterns

For each temporal window, statistical methods including transaction counts, sum aggregations, averages, maximum and minimum amounts, standard deviations, coefficients of variation, logarithmic transformations, and Z-score calculations were employed. This enables models to distinguish between temporary business fluctuations and concerning operational trends, providing nuanced risk assessment capabilities.The following behavioral metrics were derived:

- **Transaction Volume and Amount:** `transaction_count`, `sum_amount`
- **Central Tendency:** `average_amount` across all time windows
- **Variability and Consistency:** `max_amount`, `min_amount`, `std_amount`, `cv_amount`
- **Activity Patterns:** `active_days` representing business engagement frequency
- **Transaction Intensity:** Number of transactions per window indicating operational tempo
- **Growth and Trend Analysis:** Period-over-period growth calculations per window

Table 1: Sample Descriptive Statistics for Engineered Variables

Variable	**Null**	**Unique**	**Max**	**Mean**	**Std**
days_since_last_txn	0	77	112	14.58	17.21
txn_count_7d	0	65	48	6.34	5.89
sum_amount_7d	0	450	20500	1530.42	2380.55
avg_amount_7d	0	420	5125	245.80	315.17
max_amount_7d	0	430	5100	367.33	560.90
active_days_1m	0	30	30	8.14	7.01

To ensure feature quality and model stability, a threshold-based approach (90% completeness requirement) was implemented to eliminate columns with excessive missing values. Additionally, features exhibiting low variance (quasi-constant behavior) were systematically removed, along with duplicate features, ensuring optimal model input quality.

5 Binning, Feature Selection and Scorecard Development

Effective feature selection requires careful balancing of predictive power with business interpretability and model stability. This systematic approach incorporates configurable binning parameters, rigorous selection criteria, and business rules validation. Multiple statistical techniques are used to identify features providing genuine risk discrimination while maintaining operational relevance.

5.1 Binning Process

The binning process transforms raw behavioral data into structured, predictive formats suitable for scorecard development. Optimal binning maximizes information extraction while providing clear risk categories for business decision-making. Every variable receives specific binning configuration including data type classification, maximum bin count limitations, and monotonic trend expectations. The binning strategy includes:

- **Decision Tree Binning:** Utilizes classification trees to identify optimal split points that maximize separation between good and bad customers
- **Equal Frequency Binning:** Provides fallback methodology ensuring equal observation counts per bin for statistical stability
- **Monotonicity Enforcement:** Adjusts Weight of Evidence values to maintain expected risk trend relationships

5.2 Information Value (IV) Analysis

Information Value provides a quantitative measure of feature predictive power by assessing the strength of relationship between behavioral features and target variables. It helps to quantify how much each behavioral pattern contributes to risk prediction accuracy.

$$IV = \sum (\text{\% Good Customers} - \text{\% Bad Customers}) \times WoE$$

5.3 Weight of Evidence (WoE) Transformation

Weight of Evidence transformation provides a principled statistical approach to handling categorical variables and creating monotonic relationships with target variables. This maintains business interpretability and stability in score generation.

$$WoE = \ln\left(\frac{\text{Good Distribution}}{\text{Bad Distribution}}\right) \tag{1}$$

5.4 Feature Selection

The selection criteria follows three-stage evaluation process: Information Value between 0.05 and 0.08, maximum inter-variable correlation of 0.9, and a minimum of 95% data completeness, and compliance with defined business rules and constraints. After evaluating 98 features, 20 variables demonstrated strong predictive power while maintaining business relevance and statistical stability. The model regresses the target variable (default status) against the weight of evidence (WoE) of each predictor variable. This transformation ensures a consistent monotonic relationship between independent variables and the dependent variable, facilitating reliable predictions.

5.5 Scorecard Characteristics

This scorecard incorporates 20 carefully selected variables derived from the customer's transactional history with the partner, organized into the following key categories:

- **Transactional activity & Frequency (7 variables):** `txn_density_3m`, `txn_days_range`, `txn_count_slope_6m`, `txn_count_growth_1m_vs_3m`, `avg_txn_per_day_6m`, `log_avg_days_between_` `log_median_days_between_txn`

- **Transaction amounts & Spending behaviour (6 variables):** `max_amount_1m`, `log_sum_amount_1m`, `log_avg_amount_1m`,`std_amount_3m`, `cv_amount_3m`, `avg_amt_ratio_1m_6m`

- **Transaction growth & Trend features (3 variables):** `sum_amount_growth_3m_vs_6m`, `sum_amount_growth_1m_vs_3m`,`txn_count_growth_1m_vs_3m`

- **Recency & Dependency ratios (2 variables):** `rec_ratio_txn_1m_vs_3m`, `dependency_ratio_3`

- **Sector Indicators (2 variables):** `sector_FMCG`, `sector_AQUA_AGRI`

Logistic regression outputs were converted into scorecard points using the points-to-double-odds (PDO) method. Each bin receives point assignments based on corresponding Weight of Evidence values. Low-risk bins receive higher point allocations (indicating better credit scores), while high-risk bins receive lower point assignments (indicating poorer credit scores). This methodology successfully transforms raw behavioral features into actionable risk scores suitable for business decision-making.

5.6 Quality Assurance

To ensure quality throughout the binning process, comprehensive bin validation was implemented. Bins with fewer than 3 or more than 10 categories were systematically penalized. Additionally, bin distribution validation ensured minimum observation count requirements per bin. Monotonicity validation confirmed that Weight of Evidence patterns matched expected business logic and risk relationships.

6 Model Development and Selection

Selecting appropriate algorithms for credit scoring requires careful consideration of predictive performance, model interpretability, and regulatory compliance requirements. Our approach trained multiple models simultaneously, creating a competitive framework to identify the optimal risk assessment methodology. Various hyperparameter optimization techniques were employed to maximize predictive power using the binned feature set.

1. **Binning & WoE Transformation** – Continuous and categorical variables were optimally binned to maximize separation between good and bad loans. WoE values were calculated for each bin.

2. **Information Value (IV)** – Features with IV > 0.05 were retained, prioritizing those with strong and consistent predictive power.

3. **Bad Loan Definition** – Loans repaid more than 10 days past due date were defined as bad loans for modeling purposes.

6.1 Model Configuration

Four key algorithms were systematically trained and evaluated:

- **Logistic Regression:** High interpretability baseline model suitable for regulatory environments
- **Random Forest:** Strong performance characteristics with built-in feature importance analysis
- **Gradient Boosting:** Typically delivers superior predictive performance through sequential learning
- **Decision Tree:** Maximum interpretability with clear decision path visualization

Each model utilized predefined parameter spaces where hyperparameter grid search methodology identified optimal configurations.

Model	CV Score	Best Parameters
Logistic Regression	0.748	C = 0.01, class_weight = balanced, penalty = l2, solver = saga
Random Forest	0.948	class_weight = balanced, max_depth = None, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 200
Gradient Boosting	0.961	learning_rate = 0.2, max_depth = 7, min_samples_split = 10, n_estimators = 200, subsample = 0.9
Decision Tree	0.864	class_weight = balanced, criterion = entropy, max_depth = 15, min_samples_leaf = 10, min_samples_split = 2

6.2 Model Evaluation

The evaluation across multiple performance dimensions was conducted, resulting in the following:

Table 2: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1	AUC	Gini	KS	Log Loss
Random Forest	0.798	0.447	0.457	0.452	0.723	0.447	0.370	0.482
Gradient Boosting	0.806	0.465	0.435	0.449	0.719	0.438	0.343	0.964
Logistic Regression	0.701	0.691	0.662	0.684	0.73	0.403	0.343	0.617
Decision Tree	0.755	0.346	0.391	0.367	0.649	0.297	0.254	4.557

6.3 Model Selection Rationale

Following comprehensive evaluation, **Logistic Regression** was selected as the primary model despite Random Forest achieving marginally higher AUC performance. This decision was driven by several strategic considerations:

- **Regulatory Compliance:** Logistic regression provides complete interpretability required for regulatory submissions and audit requirements
- **Business Transparency:** Clear coefficient interpretation enables business stakeholders to understand risk drivers
- **Operational Simplicity:** Straightforward implementation and maintenance in production environments
- **Performance Adequacy:** 73% AUC represents acceptable performance for business deployment

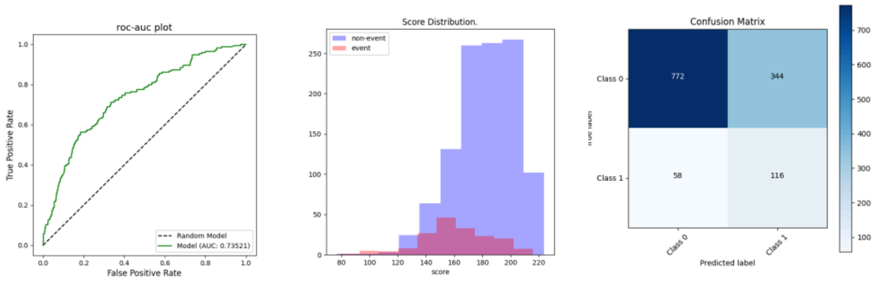


Figure 1: Performance charts of Logistic regression

6.4 Business Impact Assessment

The evaluation process employed multiple complementary metrics to assess performance across different business scenarios and operational use cases:

Table 3: Model Performance Core Metrics

Metric	Value	Business Interpretation
AUC	73%	Good discriminatory power for business decisions
Gini Coefficient	45%	Acceptable separation between risk segments
KS Statistic	37%	Strong risk differentiation capabilities
Precision	69%	Reasonable control of false positive rates
Recall	66%	Adequate identification of high-risk customers
F1-Score	68%	Balanced precision-recall trade-off

6.5 Business Risk Segmentation Analysis

The scorecard demonstrates robust risk segmentation capabilities, enabling clear differentiation across customer populations for strategic business decision-making:

Table 4: Customer Segmentation Framework Based on Model Output

Segment	Score Band	Risk Profile	Description & Recommended Business Action
Prime	Band 1	Very Low Risk	**Description:** Core, healthiest customer segment with very low default probability (12.3% bad rate). **Action:** Auto-approve with best rates and premium product offerings.
Preferred	Bands 2-8	Low-Medium Risk	**Description:** Mixed but generally favorable segment with manageable bad rates (0-33%). **Action:** Standard approval through automated underwriting processes.
Sub-Prime	Band 9	High Risk	**Description:** Critical segment requiring attention with elevated bad rate (33.3%). **Action:** Enhanced scrutiny with manual review requirements.
Decline	Band 10	Very High Risk	**Description:** Highest risk segment with substantial bad rate (54.5%). **Action:** Auto-decline to preserve portfolio quality and operational efficiency.

This segmentation framework provides clear business decision guidelines, with the majority of customers falling into low-risk categories suitable for automated processing and favorable terms.

6.6 Feature Importance and Business Insights

The model analysis identified key behavioral risk drivers that align with business intuition while revealing nuanced patterns previously unrecognized:

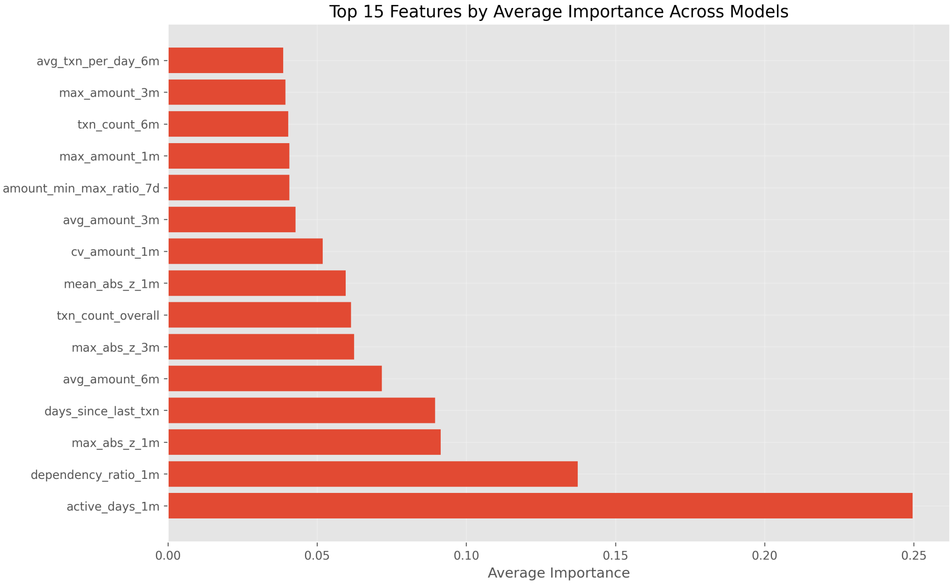


Figure 2: Top Risk Drivers in Merchant Credit Assessment

Key Business Insights from Feature Analysis:

- **Transaction Recency:** Recent activity patterns prove most predictive of repayment behavior
- **Activity Consistency:** Regular transaction patterns indicate business stability and lower risk
- **Amount Volatility:** High spending variability correlates with increased default risk
- **Dependency Ratios:** Concentrated transaction patterns suggest operational vulnerabilities

7 Fairness and Bias Prevention

Ensuring fairness and preventing discriminatory bias represents a fundamental requirement in responsible AI model development. This approach explicitly prevents utilization of protected characteristics throughout the modeling process. This model development explicitly excludes all protected characteristics from feature engineering and model training:

- **Demographic Characteristics:** Race, ethnicity, religion, national origin
- **Personal Characteristics:** Gender, sexual orientation, marital status
- **Age-Related:** Age-based variables (unless legally required for specific business purposes)
- **Geographic Proxies:** Location-based variables that might serve as proxies for protected classes
- **Disability Status:** Any variables related to physical or mental disabilities

8 Model Improvement Opportunities

While the current model demonstrates strong business performance, several improvement opportunities exist for future development phases:

- **Advanced Feature Engineering:** Exploration of additional behavioral patterns and external data sources to improve scorecard range to discriminate well between good and bad customers.
- **Dynamic Threshold Adjustment:** Implementation of business rules that adapt to changing market conditions
- **False Positive Reduction:** Focused efforts to minimize revenue loss from declined good customers
- **False Negative Prevention:** Enhanced detection capabilities to protect portfolio quality

- **Explainable AI:** Implementation of advanced model explainability techniques
- **Fairness Monitoring:** Enhanced bias detection and mitigation strategies
- **Stress Testing:** Economic scenario analysis and model resilience testing
- **Model Benchmarking:** Continuous comparison against industry best practices

9 Conclusion

This project successfully developed a credit scoring model that transforms transactional behavioral data into actionable risk assessment capabilities. The systematic approach employed from behavioral feature engineering through model validation ensures business applicability. Although Logistic regression achieves 73% AUC, there is room for improving the model's performance for better discrimination power between good and bad customers.