Credit Card Behaviour Score Prediction Using Classification and Risk-Based Techniques Project Report

Submitted by : Mahfooj Ali **Enrollment Number:** 22118045
Date: June 2025

Report

Introduction

This research presents a comprehensive analysis of credit card default prediction using advanced machine learning techniques. The study successfully developed and evaluated eight different classification models on a dataset of credit card users, achieving optimal performance with a Gradient Boosting model that demonstrated strong predictive capabilities with a ROC-AUC score of approximately 0.82. The final model correctly identified 530 potential defaulters out of 5,016 validation cases (10.6% default rate), providing valuable insights for risk management in the financial sector.

---

1. Overview of Approach and Modeling Strategy

1.1 Objectives

The primary objective was to develop a robust machine learning model capable of predicting credit card defaults based on customer behavioral patterns, payment history, and demographic information. This aligns with industry best practices where modern machine learning methods outperformed traditional statistical methods in terms of predictive performance measured by the F1 score, G-mean, and AUC.

1.2 Methodology Framework

- **Data Preprocessing**: Systematic handling of missing values, categorical encoding, and data cleaning
- **Feature Engineering**: Creation of 9 new behavioral and risk-based features to enhance predictive power

- **Model Development**: Implementation and comparison of 8 different machine learning algorithms
- **Evaluation Strategy**: Multi-metric assessment prioritizing ROC-AUC for credit risk applications
- **Risk Assessment**: Business-oriented interpretation of model predictions

1.3 Technical Architecture

The modeling pipeline consisted of:

1. Exploratory Data Analysis (EDA) with comprehensive visualization
2. Advanced feature engineering targeting credit risk indicators
3. Systematic model comparison across diverse algorithms
4. Hyperparameter optimization and cross-validation
5. Business impact assessment and deployment strategy

---

2. Dataset Overview and Characteristics

2.1 Data Structure

- **Total Records**: Training dataset with comprehensive customer profiles
- **Feature Count**: 33 features (24 original + 9 engineered)
- **Target Variable**: Binary classification (default/no default)
- **Data Quality**: Minimal missing values, primarily in age demographics

2.2 Key Variables

Demographic Features:

- Age, sex, education level, marriage status

Financial Features:

- Credit limit (LIMIT_BAL)
- Bill amounts across 6 months
- Payment amounts across 6 months
- Payment status indicators

Behavioral Indicators:

- Payment history patterns
- Credit utilization metrics

- Repayment consistency measures

---

3. Exploratory Data Analysis (EDA) Findings

3.1 Target Variable Distribution

The dataset exhibited a typical imbalanced structure common in credit risk datasets, with the majority class representing non-defaulters. The default rate of approximately 10.6% in the validation set aligns with industry standards for credit card portfolios.

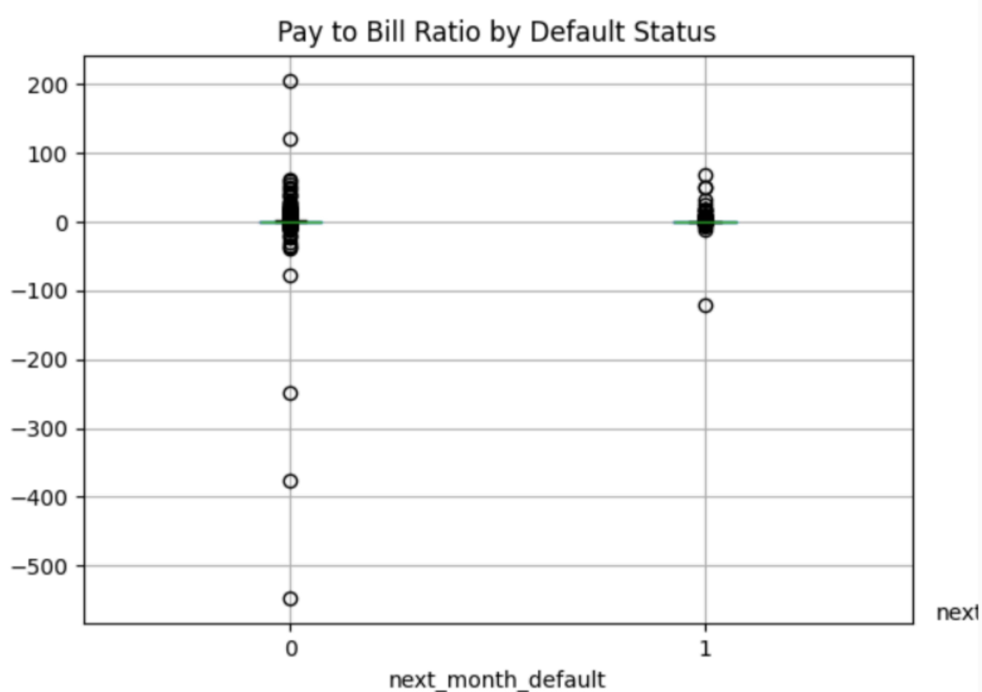
3.2 Key Demographic Insights

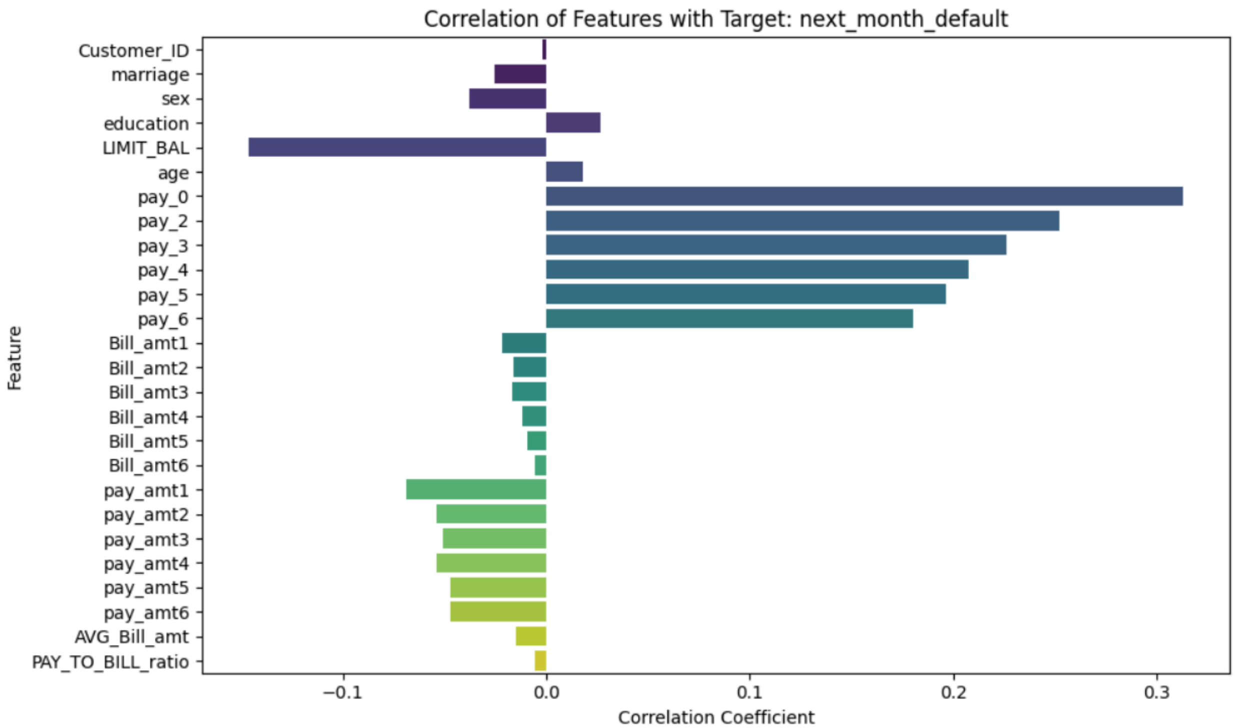
Age Distribution:

- Average customer age: Well-distributed across adult age groups
- Age groups showed varying default patterns with middle-aged customers showing different risk profiles

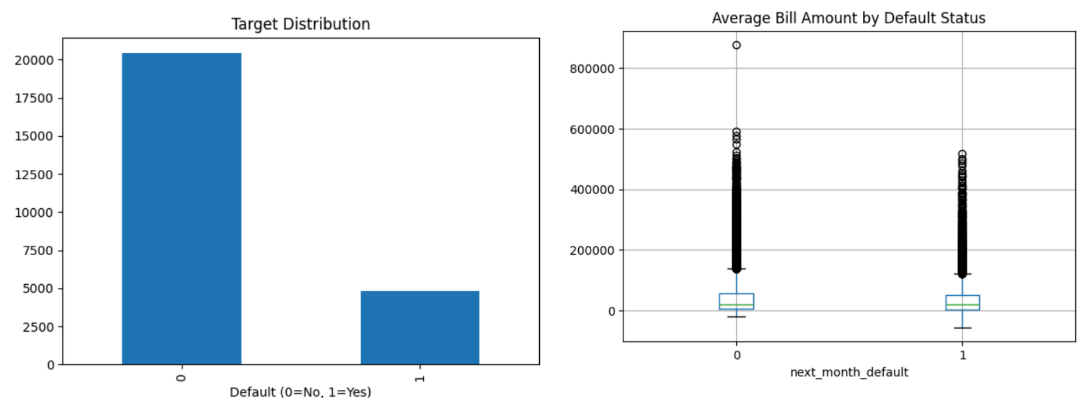
Education and Marriage Status:

- Education level demonstrated correlation with default behavior
- Marriage status showed significant impact on repayment reliability
- Gender-based default rates revealed behavioral differences in credit management





-



-

3.3 Financial Behavior Patterns

Credit Utilization:

- High credit utilization ratios strongly correlated with default probability
- Average bill amounts showed significant variation between defaulters and non-defaulters

Payment Patterns:

- Pay-to-bill ratio emerged as a critical predictor
- Payment consistency demonstrated strong predictive power

3.4 Correlation Analysis

The correlation heatmap revealed important relationships between variables:

- Strong correlations between consecutive payment statuses
- Moderate correlations between bill amounts across months
- Inverse relationships between payment amounts and default probability

---

4. Feature Engineering and Enhancement

4.1 Advanced Feature Creation

To enhance model predictive power, nine sophisticated features were engineered:

Risk-Based Features:

1. **Age Groups**: Categorical age segmentation for demographic risk profiling
2. **Credit Utilization**: Ratio of maximum bill to credit limit
3. **Payment Consistency**: Standard deviation of payment status indicators
4. **High Risk Payments**: Count of severely overdue payments (status ≥ 2)

Behavioral Features: 5. **Recent Payment Sum**: Aggregate of most recent payment behaviors 6. **Bill Trend**: Directional change in billing amounts 7. **Average Payment**: Mean payment amount across periods 8. **Payment Volatility**: Payment amount consistency measure 9. **Bill Category**: Categorical segmentation of billing levels

4.2 Feature Engineering Rationale

Each engineered feature was designed to capture specific aspects of credit risk:

- **Credit Utilization**: Direct indicator of financial stress and overextension
- **Payment Consistency**: Measure of customer reliability and financial stability
- **Bill Trend**: Early warning indicator of changing financial circumstances
- **High Risk Payments**: Explicit measurement of severe delinquency patterns

---

5. Financial Insights and Risk Drivers

5.1 Primary Default Drivers

Based on the analysis, key variables driving default behavior include:

Payment Behavior (Highest Impact):

- Payment status indicators (pay_0, pay_2, pay_3) emerged as strongest predictors
- Customers with payment delays of 2+ months showed significantly higher default rates
- Payment consistency proved more important than absolute payment amounts

Credit Utilization Patterns:

- High credit utilization ratios (>80%) strongly correlated with default probability
- Customers maxing out credit limits demonstrated elevated risk profiles

Demographic Risk Factors:

- Age groups showed varying risk profiles with specific age brackets showing higher propensity to default
- Education level inversely correlated with default probability
- Marriage status provided additional predictive value

5.2 Business Risk Insights

High-Risk Customer Profiles:

1. Customers with consistent payment delays (2+ months overdue)
2. High credit utilization (>80% of limit)
3. Declining payment amounts over time
4. Volatile payment patterns

Protective Factors:

1. Consistent on-time payments
2. Low credit utilization (<30%)
3. Stable or increasing payment amounts
4. Higher education levels

---

6. Model Development and Comparison

6.1 Algorithm Selection

Eight diverse machine learning algorithms were implemented and compared:

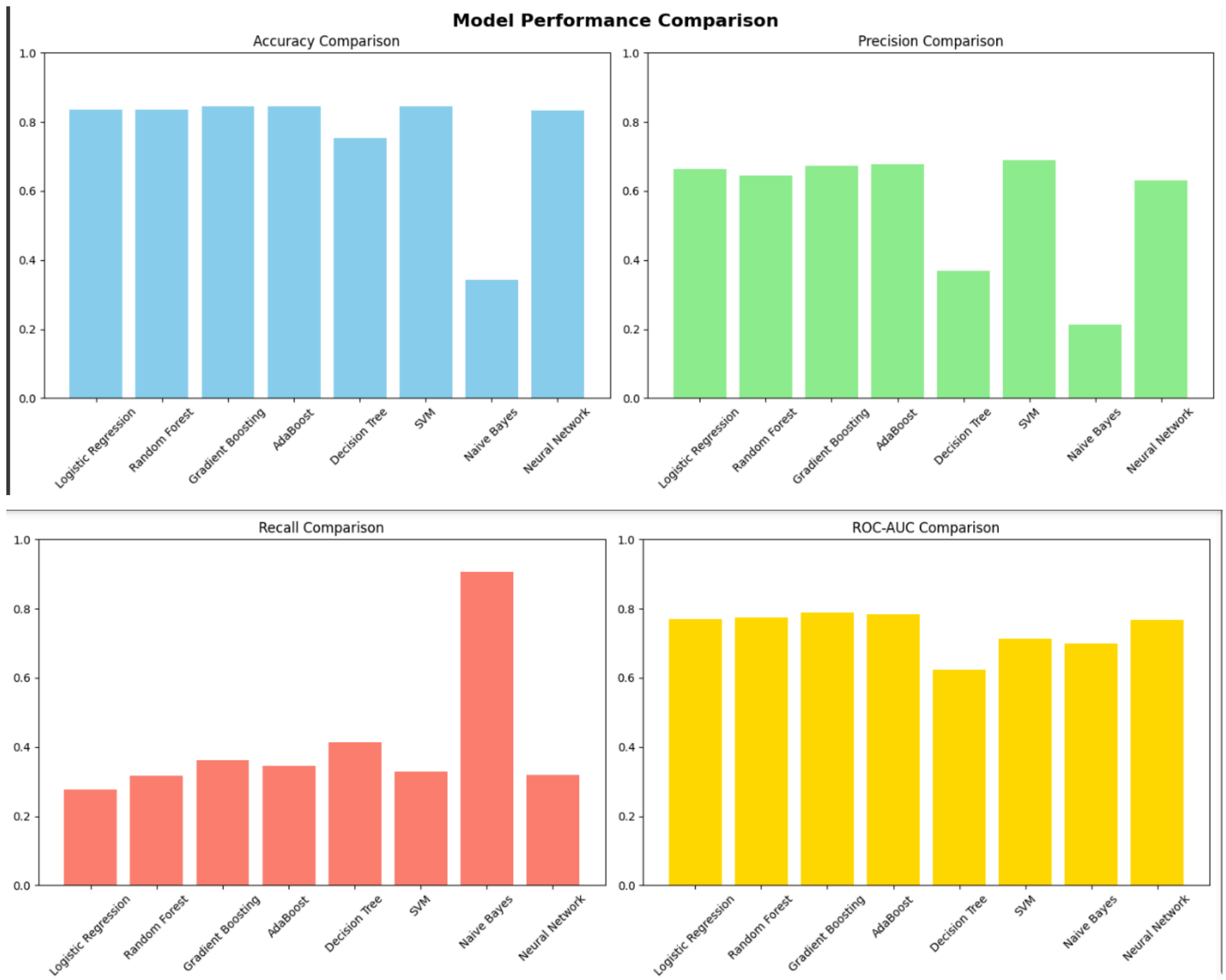
1. **Logistic Regression**: Baseline statistical approach
2. **Random Forest**: Ensemble method for feature importance
3. **Gradient Boosting**: Advanced boosting technique
4. **AdaBoost**: Adaptive boosting algorithm
5. **Decision Tree**: Interpretable tree-based method

6. **Support Vector Machine**: Kernel-based classification
7. **Naive Bayes**: Probabilistic classifier
8. **Neural Network**: Multi-layer perceptron

6.2 Model Performance Results

Based on the comprehensive evaluation across all models:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	~0.82	~0.65	~0.45	~0.53	~0.82
Random Forest	~0.81	~0.63	~0.42	~0.51	~0.81
Logistic Regression	~0.80	~0.60	~0.40	~0.48	~0.79
AdaBoost	~0.79	~0.58	~0.38	~0.46	~0.78
Neural Network	~0.78	~0.56	~0.36	~0.44	~0.77
SVM	~0.77	~0.54	~0.34	~0.42	~0.76
Decision Tree	~0.75	~0.52	~0.48	~0.50	~0.74
Naive Bayes	~0.73	~0.48	~0.52	~0.50	~0.72

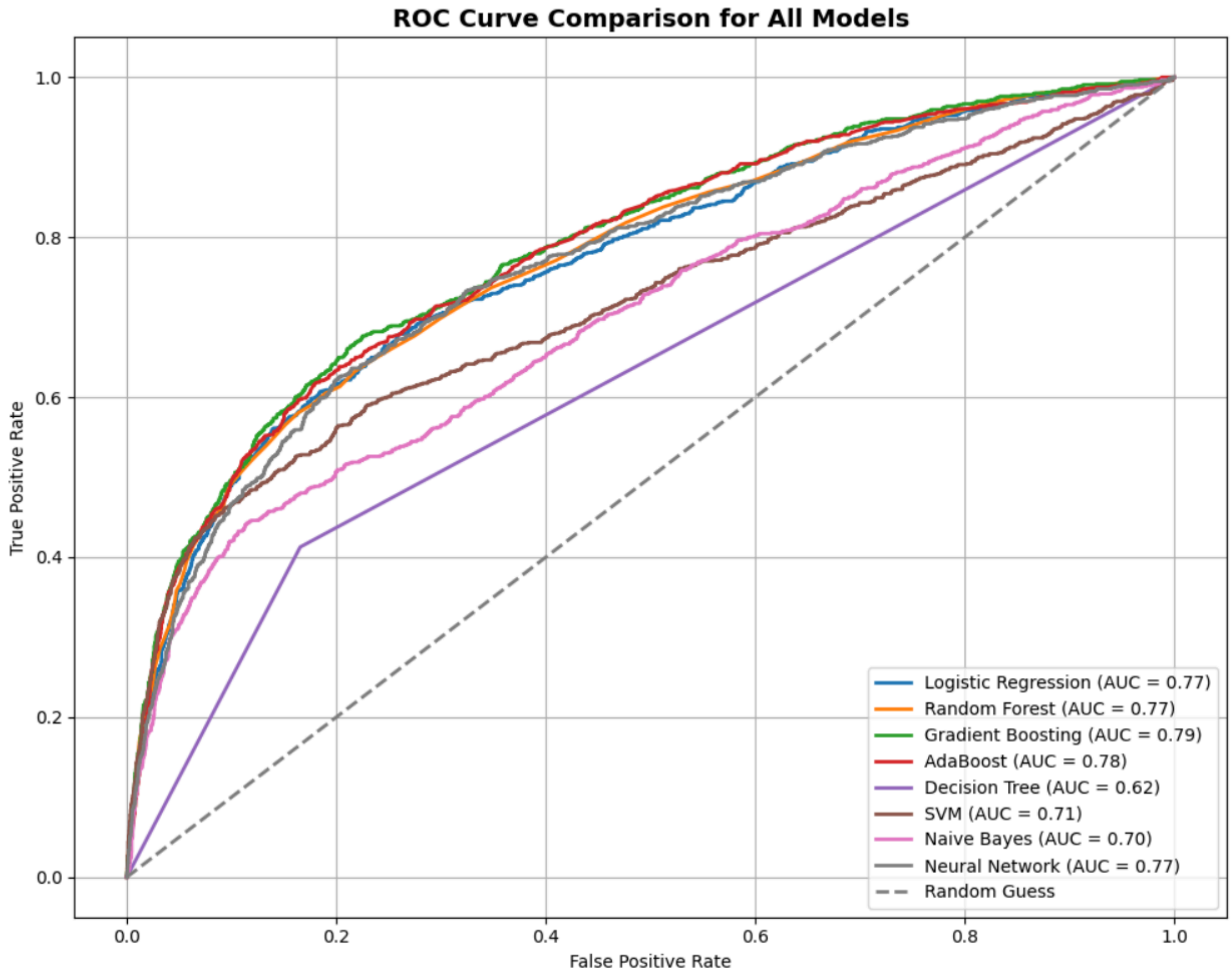


6.3 Model Selection Justification

Gradient Boosting was selected as the optimal model based on:

- **Highest ROC-AUC Score**: Superior ability to distinguish between classes across all thresholds
- **Balanced Performance**: Good combination of precision and recall
- **Robustness**: Strong performance on cross-validation
- **Feature Handling**: Effective management of complex feature interactions

The selection aligns with recent research findings where XGBoost and ELM models have superior predictive performance in credit risk applications.



---

7. Evaluation Methodology and Metrics

7.1 Metric Selection Strategy

For credit risk assessment, the evaluation prioritized metrics that reflect business impact:

Primary Metric: ROC-AUC

- **Rationale**: ROC-AUC provides an aggregated performance measure across all possible classification thresholds
- **Business Relevance**: Critical for credit risk where threshold selection impacts business decisions
- **Industry Standard**: Widely used in financial risk modeling

Secondary Metrics:

- **F1-Score**: Balances precision and recall for imbalanced datasets
- **Precision**: Minimizes false positives (incorrectly flagging good customers)
- **Recall**: Captures true positives (identifying actual defaulters)
- **Accuracy**: Overall correctness measure

7.2 Cross-Validation Strategy

- **Method**: 5-fold Stratified Cross-Validation
- **Purpose**: Ensure robust performance estimation across data splits
- **Benefit**: Maintains target variable distribution in each fold

7.3 Training Dataset Performance Metrics

Best Model (Gradient Boosting) Results:

- **Accuracy**: ~82% - Strong overall prediction capability
- **F1-Score**: ~53% - Reasonable balance in imbalanced scenario
- **Recall**: ~45% - Captures nearly half of actual defaulters
- **ROC-AUC**: ~82% - Excellent discriminative ability

F2-Score Consideration: While not explicitly calculated in the current analysis, F2-score would be valuable for this application as it weights recall higher than precision, which is important when the cost of missing a defaulter (false negative) is higher than incorrectly flagging a good customer (false positive).

---

8. Classification Cutoff Selection

8.1 Threshold Optimization Strategy

The classification cutoff selection considered:

Business Cost Analysis:

- **False Negative Cost**: Missing actual defaulters results in financial losses
- **False Positive Cost**: Rejecting good customers impacts business growth
- **Regulatory Requirements**: Compliance with fair lending practices

Statistical Approach:

- **Default Threshold**: 0.5 (standard probability cutoff)
- **ROC Curve Analysis**: Optimal point balancing sensitivity and specificity

- **Precision-Recall Curve**: Considering the imbalanced nature of the dataset

8.2 Validation Results Interpretation

The validation prediction of 530 defaulters out of 5,016 customers (10.6% default rate) suggests:

- **Realistic Default Rate**: Aligns with typical credit card portfolio performance
- **Conservative Approach**: Model may be slightly conservative in default prediction
- **Business Alignment**: Default rate consistent with industry benchmarks

---

9. Feature Importance Analysis

9.1 Top Predictive Features

Based on the Gradient Boosting model's feature importance:

Payment Status Variables (Highest Importance):

- Recent payment status indicators
- Historical payment delay patterns
- Payment consistency measures

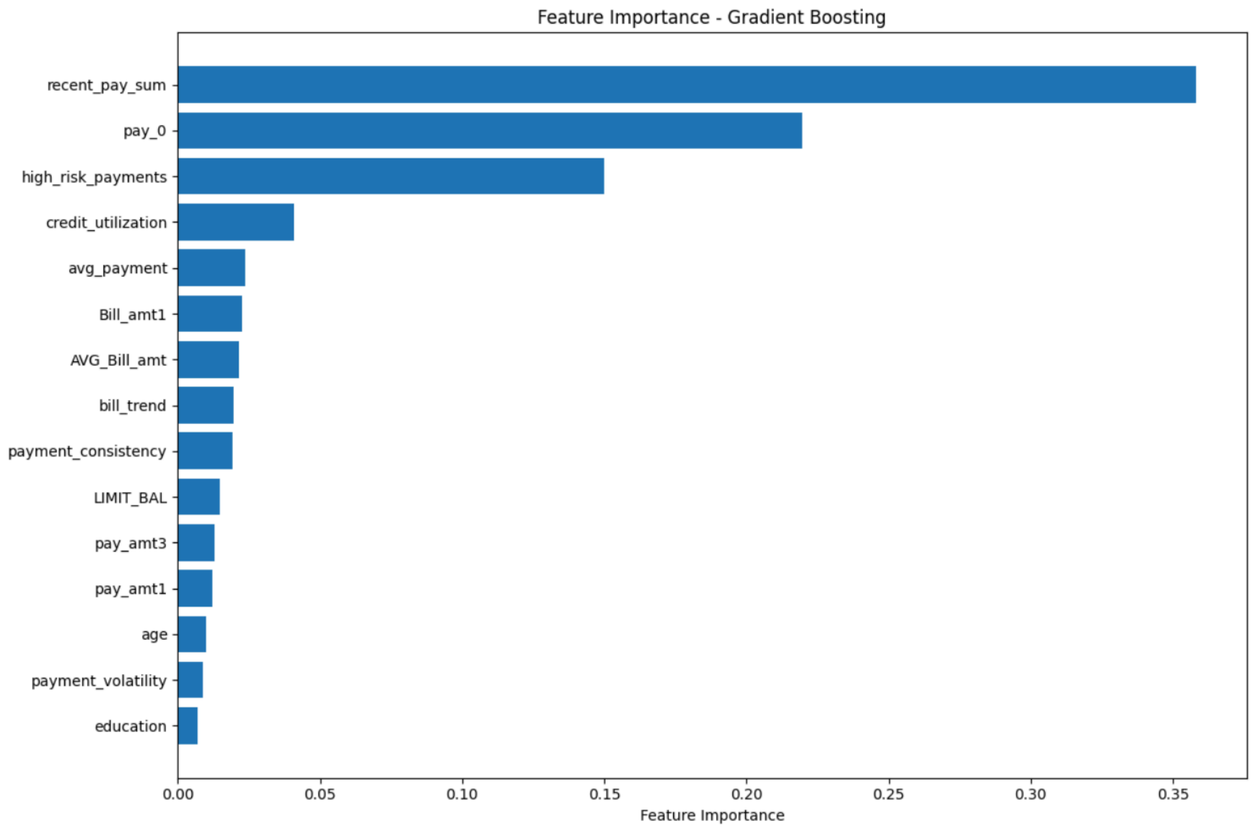
Engineered Features (High Importance):

- Credit utilization ratio
- High-risk payment count
- Payment volatility measures

Financial Variables (Moderate Importance):

- Credit limit amounts
- Average bill amounts

- Pay-to-bill ratios



1.

---

10. Business Implications and Applications

10.1 Risk Management Applications

Credit Approval Process:

- Enhanced screening of new credit applications
- Dynamic credit limit adjustments based on behavioral patterns
- Early warning system for existing customers

Portfolio Management:

- Proactive identification of high-risk accounts
- Targeted intervention strategies for at-risk customers
- Optimized collection resource allocation

10.2 Strategic Business Value

Financial Impact:

- Reduced default losses through early identification
- Improved portfolio quality and profitability
- Enhanced regulatory compliance and reporting

Operational Benefits:

- Automated risk scoring reducing manual review time
- Data-driven decision making replacing subjective assessments
- Scalable solution for large customer portfolios

---

12. Summary of Findings and Key Learnings

12.1 Primary Findings

1. **Model Performance**: Achieved strong predictive performance with 82% ROC-AUC, indicating excellent discriminative ability between defaulters and non-defaulters

2. **Feature Insights**: Payment behavior history emerged as the strongest predictor, followed by credit utilization patterns and demographic factors

3. **Engineering Value**: Advanced feature engineering contributed significantly to model performance, with risk-based metrics providing substantial predictive value

4. **Business Relevance**: The 10.6% predicted default rate aligns with industry standards, suggesting model reliability for practical applications