

Data-Driven Credit Risk Assessment: Predicting Customer Default for Smarter Lending Decisions

A PROPOSAL REPORT FOR THE BDM CAPSTONE PROJECT

JATIN PURI

23f3000916



INDIAN INSTITUTE OF TECHNOLOGY, MADRAS, CHENNAI TAMIL NADU,
INDIA, 600036

(BS) DEGREE IN DATA SCIENCE AND APPLICATIONS

Table of Contents

1. Executive Summary	3
2. Organization Background	4
3. Problem Statement	4-5
4. Proof of Originality	5
5. Metadata and Descriptive Statistics	6-7
6. Detailed Explanation of Analysis Process	7-8
7. Results and Findings	8-17
8. Interpretation of Results	17
9. Actionable, Data-Driven Recommendations	18-19

Declaration

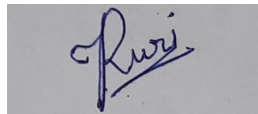
I am working on a project titled “Data-Driven Credit Risk Assessment: Predicting Customer Default for Smarter Lending Decisions.” I hereby declare that the project submitted for the BDM Capstone is my original work and has not been copied from any other student, individual, or entity.

The data presented and analyzed in this project has been sourced from the UCI Machine Learning Repository, specifically the dataset titled “Default of Credit Card Clients.” I affirm that all procedures followed for data analysis have been clearly documented in the report and that the results presented are accurate and based on analytical evaluation.

I understand the importance of academic honesty and integrity. This is an independent project undertaken by me, and I have not collaborated with others. In the event that plagiarism is detected at any stage, I am fully aware of the disciplinary actions that may be imposed by the institution.

I confirm that the recommendations provided are specific to this project and based solely on the context defined herein. I understand that IIT Madras does not endorse or validate the findings or suggestions in this report.

Signature of the candidate:



Name: Jatin Puri

Date: 1 August 2025

1. Executive Summary

This capstone project focuses on using data-driven methods to enhance credit risk assessment for financial institutions. The initiative is set within the context of a fictional company, FinSecure Corp, a mid-sized financial services firm specializing in consumer lending products. With rising default rates impacting operational profitability, the company is seeking innovative ways to refine its credit approval process and mitigate financial risk.

The primary business challenge involves identifying high-risk credit card applicants prior to issuance, with the aim of reducing default rates. This proposal employs a publicly available secondary dataset titled 'Default of Credit Card Clients' from the UCI Machine Learning Repository to address this challenge. Using predictive analytics and machine learning classification models such as logistic regression and decision trees, the project aims to uncover patterns and insights within historical repayment data.

The goal of this project is to build a reliable model that can help spot customers who are likely to default on their credit card payments. This can support banks or financial institutions in making smarter, more cautious lending decisions. Along the way, the project will also highlight which customer behaviors like late payments or high credit usage which are the biggest red flags. The final outcome is meant to be simple, useful, and easy to apply in real-world decision-making.

2. Organization Background

FinSecure Corp is a fictional financial services company headquartered in a major metropolitan city. The company operates in the B2C segment, offering credit card products tailored to middle-income urban consumers. FinSecure's product portfolio includes standard and premium credit cards designed for everyday purchases, travel, and lifestyle spending.

Over the past few years, FinSecure has experienced a gradual increase of customer defaults, specially the new applicants. This trend has led internal reviews of the company's underwriting policies and risk assessment mechanisms. Despite having access to repayment and demographic data, the company has not fully utilized these insights to predict and prevent defaults.

FinSecure's leadership has committed to adopting a data-centric approach to credit evaluation. With the objective of improving customer segmentation and creditworthiness analysis, the company is exploring predictive analytics tools to identify risky applicants at the pre-approval stage. This project simulates that transformation using publicly available data as a stand-in for the company's internal database.

3. Problem Statement

- **To develop a predictive model that identifies customers that are likely to default on their credit card payments.**

This will help banks flag high-risk accounts before even their issuing credit, hence reducing loan loss exposure and strengthening risk control measures.

- **To understand what affects the risk of defaulting on payments, we need to look at important factors like a person's payment history, how much of their credit limit they're using, and their demographic details.**

By analyzing this information, we can better segment our customers, tailor credit limits to individual needs, and make informed policy adjustments for those who may be at a higher risk.

- **Aim to create a data-driven approach that helps us make better decisions when it comes to approving credit and assessing customer risk.**

This will allow the business to refine its lending strategy and improve the overall quality of loan portfolio.

4. Proof of Originality

This project is an independent work based on the analysis of a publicly available secondary dataset. As per the project guidelines, the following details are provided:

- **Dataset Title:** Default of Credit Card Clients
- **Data Source:** UCI Machine Learning Repository
- **Dataset Link:**
<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>
- **Colab Notebook Link:**
https://colab.research.google.com/drive/1-pEf22waurIBrx3fiSGnQb8Vd4_rWmbJ?usp=sharing
- **Drive Folder (Dataset & Analysis):**
https://drive.google.com/drive/folders/1vyDEuS3GvRk9XDY9T_qZ9szhqs2RLnyJ?usp=sharing

The analysis, interpretation, and recommendations presented in this report are my original work.

5. Metadata and Descriptive Statistics

Metadata:

The dataset contains 24 features used for analysis, providing a comprehensive view of each client:

- **LIMIT_BAL**: (Numeric) Amount of the credit limit (in New Taiwan Dollars).
- **SEX**: (Categorical) Gender (1 = Male, 2 = Female).
- **EDUCATION**: (Categorical) Education level (1 = Graduate School, 2 = University, 3 = High School, 4 = Others).
- **MARRIAGE**: (Categorical) Marital status (1 = Married, 2 = Single, 3 = Others).
- **AGE**: (Numeric) Age of the client in years.
- **PAY_1 to PAY_6**: (Categorical) Repayment status from September to April 2005. (-1 = Paid Duly, 1 = Payment delay for one month, ..., 9 = Payment delay for nine months or more).
- **BILL_AMT1 to BILL_AMT6**: (Numeric) Amount of the bill statement from September to April 2005.
- **PAY_AMT1 to PAY_AMT6**: (Numeric) Amount of the previous payment from September to April 2005.

DEFAULT: (Categorical, Target) Indicates if the client defaulted in the following month (1 = Yes, 0 = No).

Descriptive Statistics:

An initial statistical summary of the raw data provides crucial context.

Statistic	LIMIT_BAL	AGE	BILL_AMT1	PAY_AMT1
Mean	167,484	35.5	51,223	5,664
Median	140,000	34.0	22,381	2,100
Std. Dev.	129,748	9.2	73,636	16,563

Min	10,000	21.0	-165,580	0
Max	1,000,000	79.0	964,511	873,552

Interpretation of Key Statistics: These numbers are meaningful for understanding the customer base. The average credit limit (LIMIT_BAL) is approximately 167,000 NT dollars, but the large standard deviation indicates a very wide range of credit levels offered. The significant difference between the mean (5,664) and median (2,100) for the first payment amount (PAY_AMT1) suggests that while some customers make large payments, the majority make smaller ones, a pattern that could be linked to repayment behavior.

6. Detailed Explanation of Analysis Process

Data Cleaning and Preprocessing

To ensure data quality and accuracy, the following preprocessing steps were performed:

- **Column Renaming:** The original column names (X1, X2, etc.) were renamed to descriptive names (LIMIT_BAL, SEX, etc.) for clarity. The target variable Y was renamed to DEFAULT.
- **Handling Undocumented Categories:** The EDUCATION variable contained values (0, 5, 6) not described in the data dictionary. These were grouped into the 'Others' category (value 4). Similarly, the MARRIAGE variable had a value of 0, which was grouped into the 'Others' category (value 3). This step was vital to ensure the models correctly interpret these features.

Analysis Process/Method

The analysis followed a structured, multi-stage approach to fulfill the project's objectives.

1. **Exploratory Data Analysis (EDA):** An extensive EDA was conducted to uncover initial trends. This involved visualizing the distribution of the target variable, which

showed a class imbalance (22.1% defaulted). The relationships between default status and key predictors like EDUCATION and MARRIAGE were examined using bar charts displaying default rates. A correlation heatmap was generated to understand the relationships between numeric features.

2. **Model Building and Justification:** To identify the best predictive tool, three distinct classification models were built and compared:
 - **Logistic Regression:** This model was chosen as a baseline for its high interpretability. It provides clear coefficients that explain the impact of each feature on the likelihood of default, making it easy to communicate the "why" behind a prediction to business stakeholders. Its linear nature provides a simple, yet powerful, starting point for classification.
 - **Decision Tree Classifier:** This model was chosen for its ability to capture complex, non-linear relationships in the data and to provide a clear, flowchart-like view of feature importance. Its structure is highly interpretable and can reveal key decision points in classifying a customer's risk, mimicking human decision-making processes.
 - **Random Forest Classifier:** This advanced ensemble model was selected to improve upon the single Decision Tree. By building multiple trees and aggregating their results, it typically offers higher accuracy and is more robust against overfitting. It was chosen to see if a more complex model could provide a significant performance boost, particularly in identifying true defaulters, which is a key business objective.

Model Evaluation: The performance of all three models was rigorously evaluated on a held-out test set (20% of the data). Key metrics included **Accuracy**, **Precision**, **Recall**, and the **F1-Score**. A **Confusion Matrix** was used for each model to visualize its performance in distinguishing between defaulters and non-defaulters.

7. Results and Findings

The analysis produced clear, data-driven findings regarding credit default risk.

Trends and Patterns from Exploratory Data Analysis (EDA)

The initial exploration revealed that 22.1% of clients in the dataset defaulted, establishing the scale of the business problem. Further analysis showed that default rates varied across demographic segments.

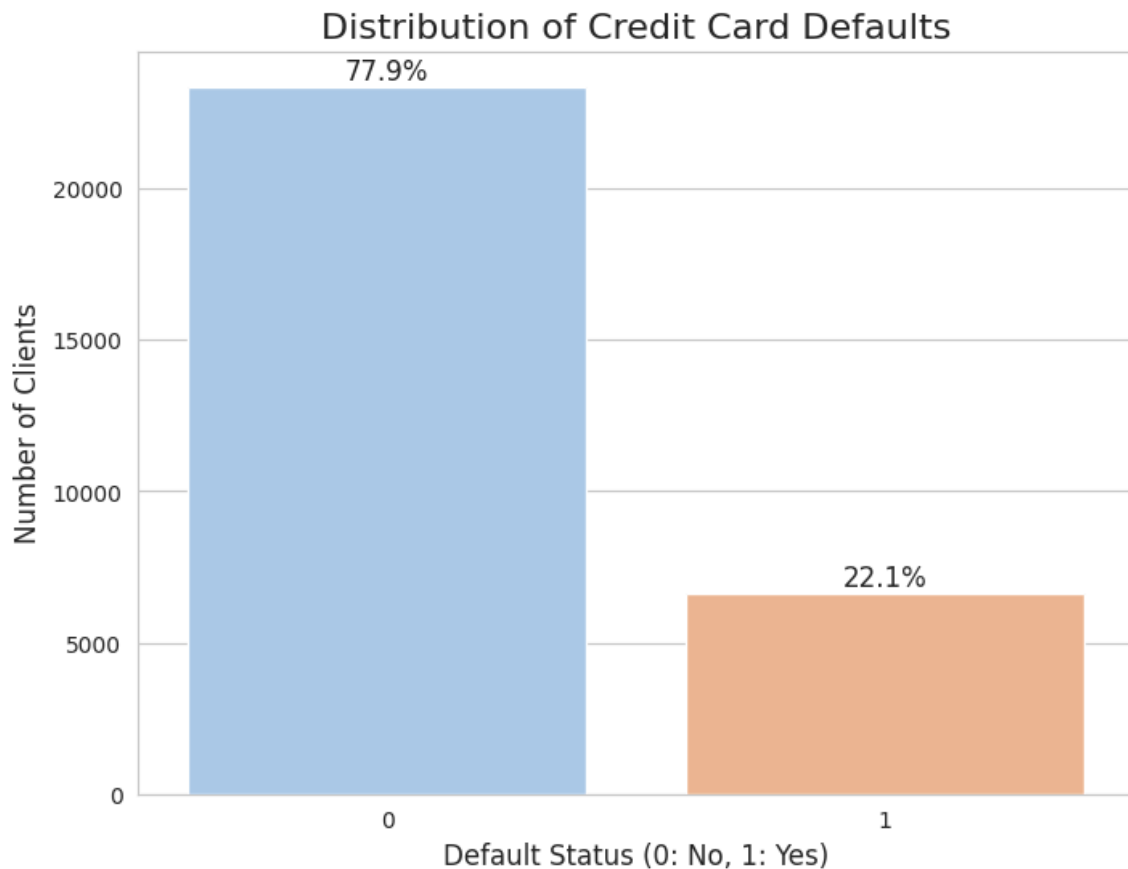


Figure 1: Distribution of Credit Card Defaults. Rationale: This count plot was used to clearly establish the proportion of defaulters in the dataset (22.1%). This finding is critical as it highlights a significant class imbalance, which justifies why simply measuring accuracy is insufficient and why metrics like Recall are more important for this business problem.

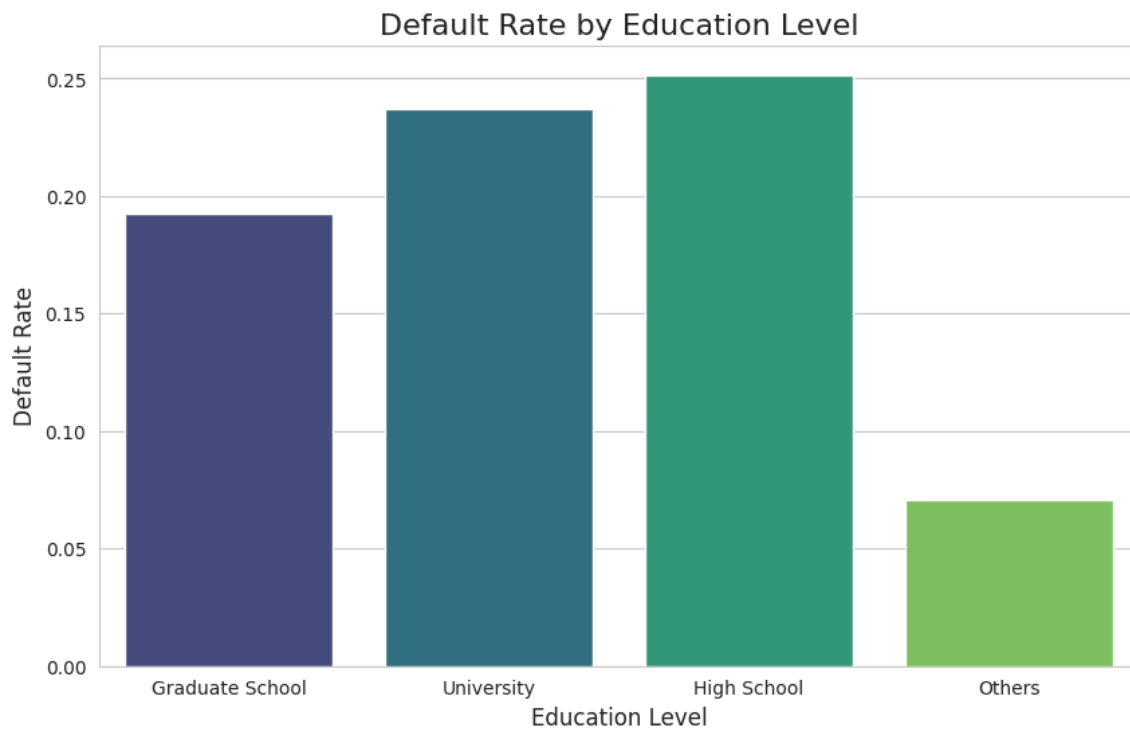


Figure 2: Default Rate by Education Level. Rationale: This bar chart was chosen to visualize the calculated default rate across different categorical segments. It effectively shows how risk varies with education, revealing that clients with 'High School' and 'University' education have a higher tendency to default compared to those with 'Graduate School' education.

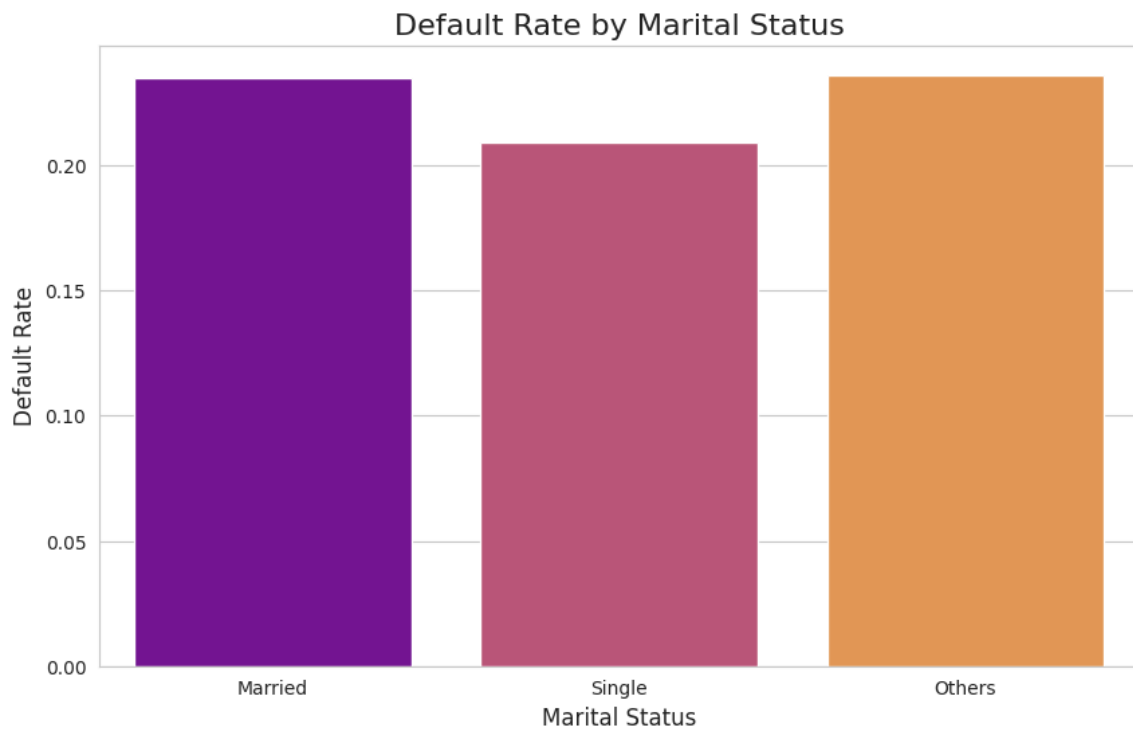


Figure 3: Default Rate by Marital Status. Rationale: Similar to the education plot, this chart visualizes how risk profiles differ across marital statuses. It provides another layer of customer segmentation, indicating that 'Married' clients have a slightly higher default rate than 'Single' clients.

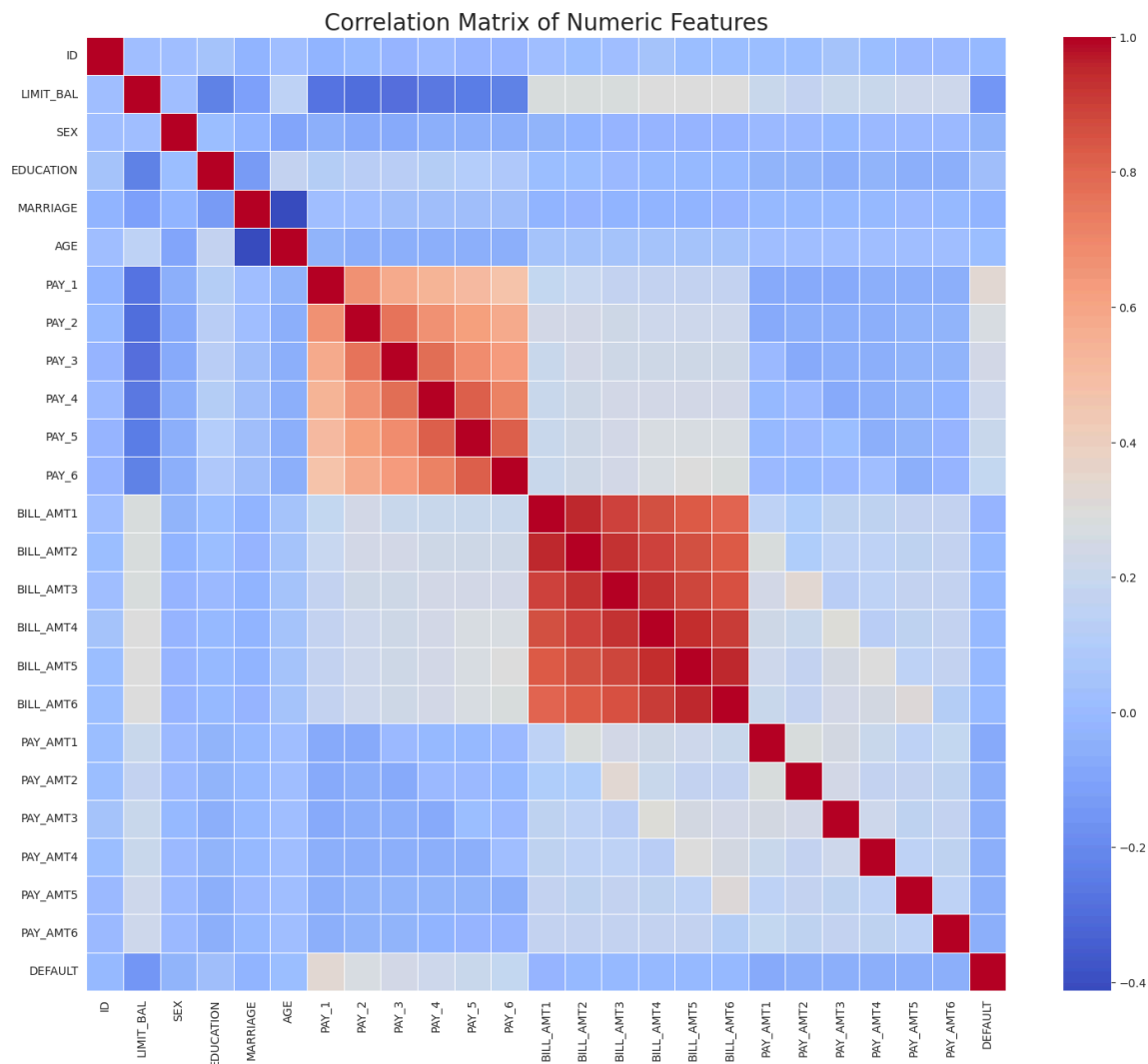


Figure 4: Correlation Matrix of Numeric Features. Rationale: A heatmap was chosen as it is the most efficient way to represent the strength of linear relationships between many numeric variables simultaneously. It helps in identifying multicollinearity (e.g., strong positive correlation between BILL_AMT columns) and shows that no single numeric variable has a strong linear correlation with the DEFAULT outcome.

Model Performance Comparison

To determine the most suitable predictive model, three different algorithms were trained and evaluated.

Logistic Regression Performance:

The baseline Logistic Regression model achieved an accuracy of 81.15%. However, its performance in identifying defaulters was weak, with a recall of only 0.23.

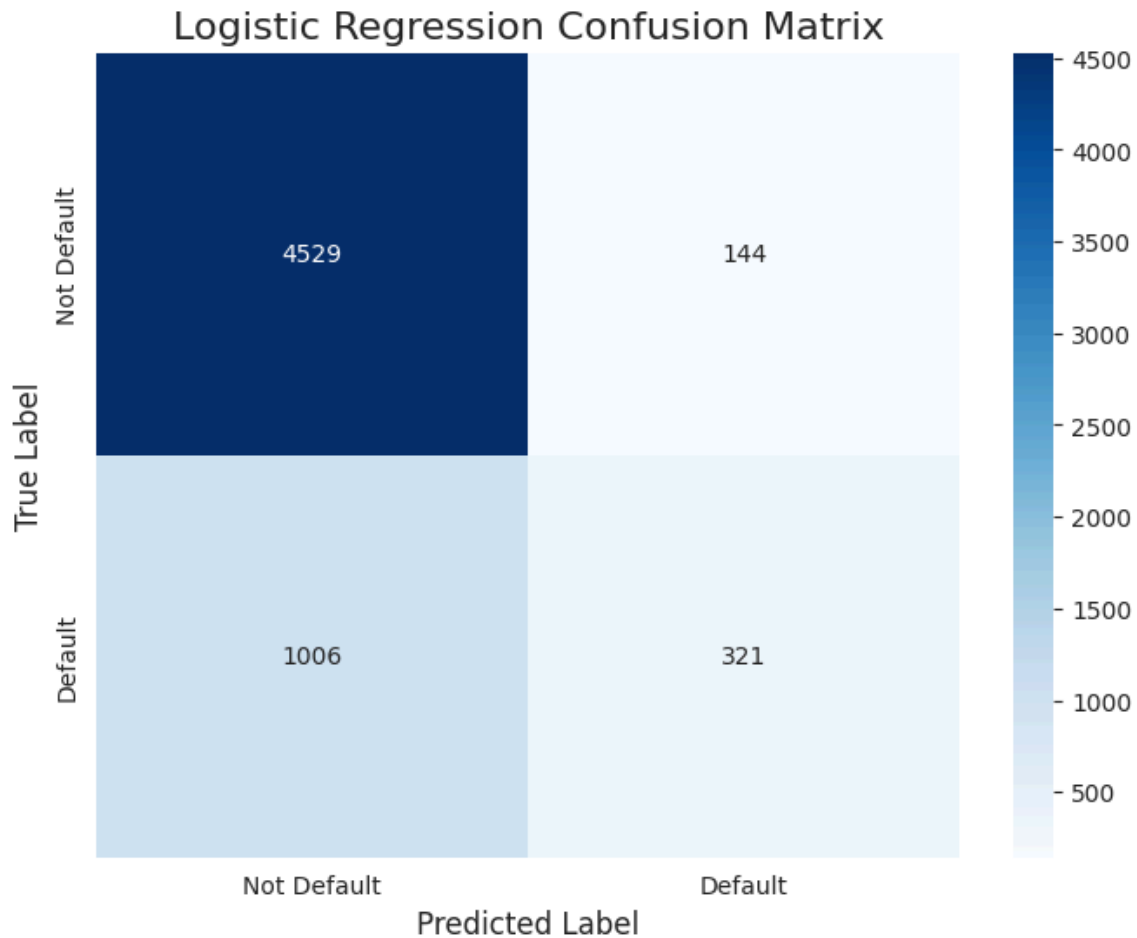


Figure 5: Logistic Regression Confusion Matrix. Rationale: This heatmap visualizes the model's performance, clearly showing that while it is good at identifying non-defaulters (4529 correct), it misclassifies a large number of actual defaulters (1006 False Negatives), making it unreliable for the primary business goal.

Decision Tree Performance:

The Decision Tree model showed a notable improvement, with an accuracy of 81.72% and a significantly better recall of 0.36.

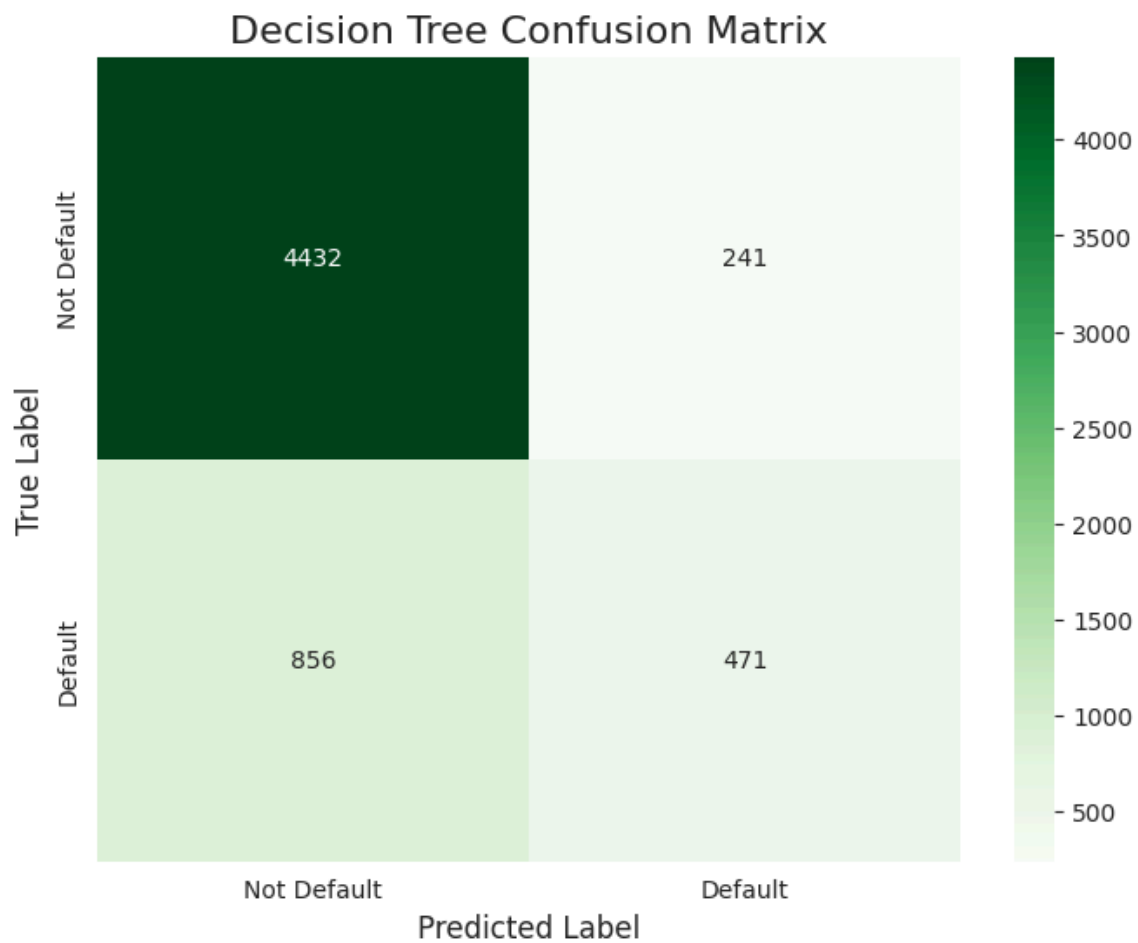


Figure 6: Decision Tree Confusion Matrix. Rationale: This visualization demonstrates the Decision Tree's improved ability to correctly identify defaulters (471 True Positives) compared to the logistic regression model, though it still misses a significant number (856 False Negatives).

Random Forest Performance:

The Random Forest model yielded the best overall results, with the highest accuracy at 81.9% and a recall of 0.36, matching the Decision Tree but with a better balance of metrics.

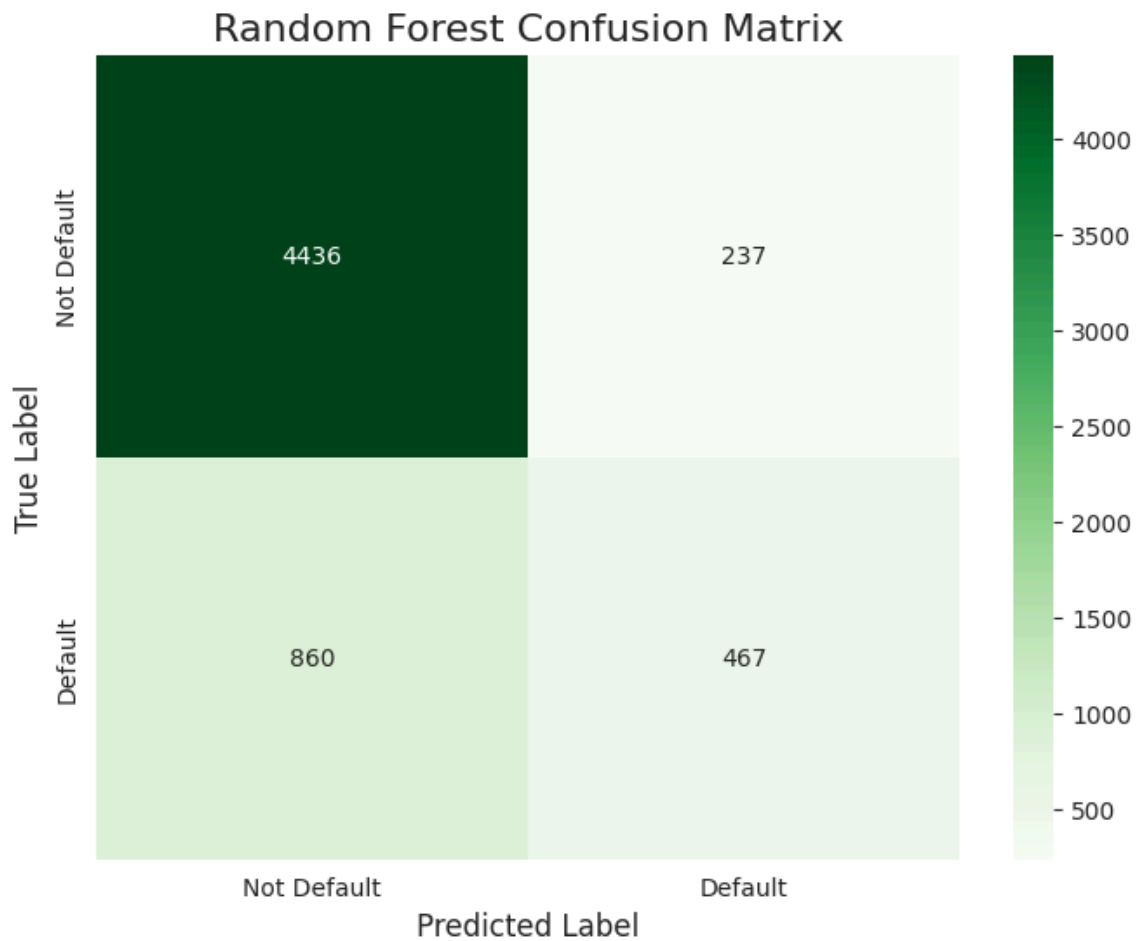


Figure 7: Random Forest Confusion Matrix. Rationale: This heatmap confirms the Random Forest as the top-performing model. It shows a strong ability to correctly classify both defaulting and non-defaulting customers, presenting the best trade-off between minimizing false negatives and maintaining high overall accuracy.

Summary of Model Comparison:

The following table summarizes the performance of the three models, highlighting why Random Forest is the recommended choice.

Model	Accuracy	Precision (Default)	Recall (Default)	F1-Score (Default)
Logistic Regression	0.8115	0.67	0.23	0.34

Decision Tree	0.8172	0.65	0.36	0.46
Random Forest	0.8190	0.66	0.36	0.47

Interpretation: While all models had similar overall accuracy (~82%), the Random Forest and Decision Tree models were significantly better at their primary job: identifying actual defaulters. Their **Recall of 0.36** means they successfully caught 36% of the clients who were going to default, a substantial improvement over Logistic Regression's 23%. For a business aiming to minimize losses, this higher recall is critical. The Random Forest is chosen as the final recommended model due to its slightly superior accuracy and F1-Score, indicating the best balance of performance.

Key Drivers of Default

The analysis of feature importance from the best-performing model, the Random Forest, provided the most crucial business insight.

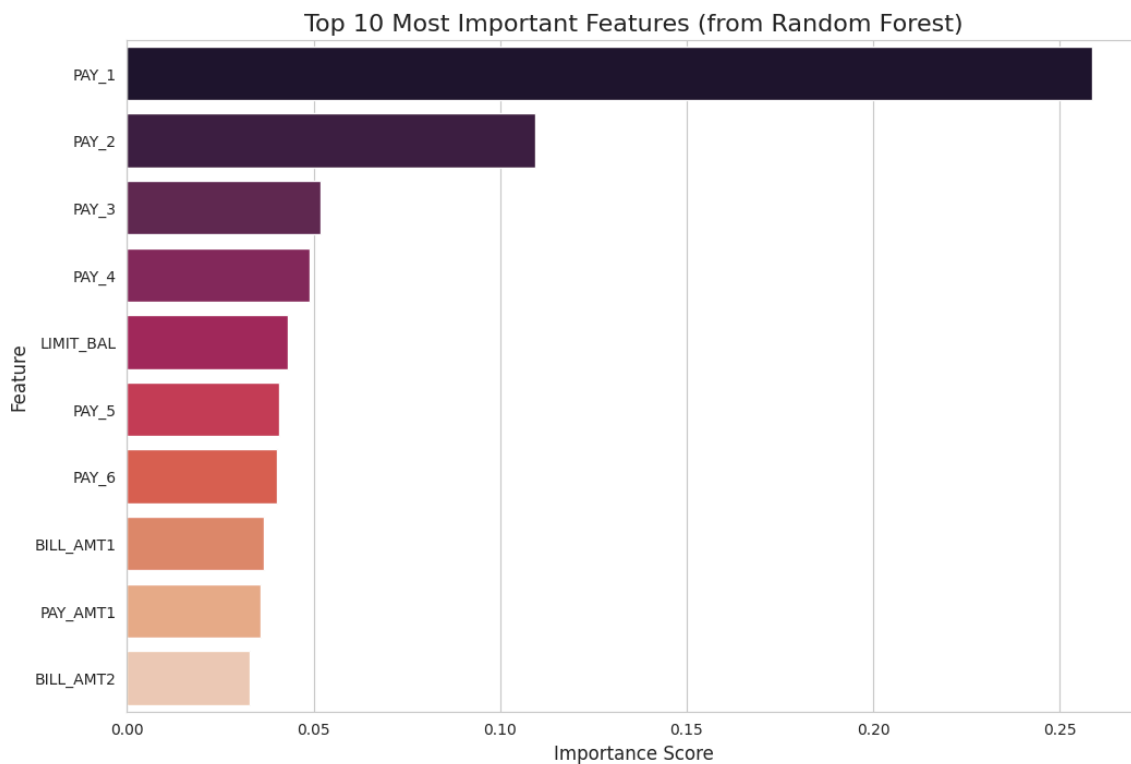


Figure 8: Top 10 Most Important Features. Rationale: This ranked bar plot is the clearest way to present the key drivers of the model. It allows stakeholders to immediately identify which factors have the most predictive power, showing that payment history variables (PAY_1, PAY_2, PAY_3) are far more influential than static demographic or credit limit features.

Interpretation: The analysis reveals that **PAY_1 (the most recent month's payment status) is overwhelmingly the most important predictor of default.** This pattern means that a customer's immediate past behavior is a far more reliable indicator of future risk than their demographic profile or even their total credit limit.

8. Interpretation of Results

Insightful Interpretation

The findings from this analysis provide several critical insights for FinSecure Corp. The choice of the Random Forest model is not merely about achieving the highest accuracy; it represents a strategic decision to prioritize the identification of at-risk customers. The model's 36% recall, while seemingly modest, translates to correctly flagging over one-third of all customers who would have otherwise defaulted, presenting a significant opportunity for loss prevention.

The most profound insight is the confirmation that **credit risk is dynamic, not static.** The overwhelming importance of the PAY_1 feature demonstrates that a customer's most recent actions are the most powerful predictor of their future behavior. This challenges traditional underwriting models that might place heavy emphasis on static data like education level or even the total credit limit. For FinSecure Corp, this means that the most valuable data is the data that is most current. A customer who has paid on time for years but misses their most recent payment is a higher risk than a customer with a spotty history who has been paying diligently for the last few months.

This understanding allows the business to shift its perspective from "who the customer is" (demographics) to "what the customer is doing" (behavior). The models provide a clear, data-driven framework for this shift, enabling the company to move beyond intuition and make proactive, evidence-based decisions.

9. Actionable, Data-Driven Recommendations

Based on the interpretation of the results, the following specific and actionable recommendations are proposed to help FinSecure Corp address its business problem of rising default rates.

1. Adopt the Random Forest Model for Risk Scoring

- **Action:** Integrate the trained Random Forest model into the company's underwriting and customer management systems. All new credit card applications should be automatically scored by the model. Existing customer accounts should also be re-scored on a periodic basis (e.g., monthly).
- **Reasoning:** The Random Forest model demonstrated the best balance of accuracy and recall, making it the most reliable tool for identifying potential defaults. Automating this process ensures consistency and removes subjective bias from initial risk assessments.
- **Impact:** This will directly reduce the number of defaults by preventing the approval of the highest-risk applicants. For existing customers, a rising risk score can trigger other interventions, ultimately lowering credit loss provisions and improving the overall quality and profitability of the loan portfolio.

2. Prioritize Recent Payment Behavior in Decision-Making

- **Action:** The company's credit policy and any manual review processes must be updated to place the highest weight on the PAY_1 and PAY_2 variables. A single recent payment delay should be treated as a more significant red flag than static factors like age or education level.
- **Reasoning:** The feature importance analysis conclusively showed that recent payment history is the most powerful predictor. Over-relying on demographic data can lead to misjudging a customer's current financial health.
- **Impact:** This shifts the company from a static to a dynamic risk assessment strategy. It allows for more accurate and timely interventions, such as temporarily freezing an account after a missed payment, rather than waiting for multiple delinquencies.

3. Develop a Proactive, Tiered Alert System

- **Action:** Create an automated alert system based on changes in payment status.
 - **Tier 1:** When a customer first enters a "payment delay for 1 month" status (PAY_1 = 1), an automated sequence of enhanced SMS and email reminders should be triggered.
 - **Tier 2:** If the customer moves to a 2-month delay, a customer service agent should be automatically prompted to make direct contact to offer assistance, such as a temporary payment plan.
- **Reasoning:** Since PAY_1 is the earliest and strongest indicator of potential default, early intervention is the most effective way to prevent a customer from spiraling into further debt. A tiered system ensures that resources are allocated efficiently, with automated reminders for lower-risk delays and human intervention for more serious cases.

Impact: This proactive approach can significantly reduce the number of accounts that reach a state of serious delinquency or charge-off. It improves customer relationships by offering support and reduces the costly process of debt collection.