30/10/2025

Predictive Modeling for Credit Card Default Risk Management

A Methodological and Empirical Analysis

Assignment Task 3

Studio 3: Innovation

Group 1

SOMESH SHANBHAG
TANVI VARAK
LAKSHYA PUNIA
SUSHANT BHATTARAI
SAHIL UPPAL
KULDEEP POKHREL

Table of Contents

Abstract

Amidst rising global credit card delinquency rates and significant financial losses, this research develops and validates a machine learning framework to predict credit card default. The study defines the problem as a binary classification task leveraging the UCI "Default of Credit Card Clients" Dataset of 30,000 clients. Following the CRISP-DM methodology, the workflow systematically addresses two primary data challenges which are severe class imbalance (22.12 % default class) using the Synthetic Minority Oversampling Technique (SMOTE) and high multicollinearity among bill features using Principal component Analysis (PCA). By trying out commonly used machine learning models on the raw dataset, a baseline model is selected and then compared against other models, while undergoing further enhancement via dataset transformation and domain specific feature engineering. The final Gradient Boosting model emerged as the top performer achieving a strong ROC-AUC score of 0.7653 and an acceptable recall of 0.59 on the hold-out test set. This performance meets the project's criteria for a viable and discriminative model. The study concludes that this structured approach produces an actionable classifier providing a clear roadmap for financial institutions to implement this model as the core of a tiered risk-based intervention strategy to proactively manage risk and reduce financial losses.

1. Introduction and Business Context

The global finance ecosystem today stands at a very critical inflection point with rapid technological advancements, evolving consumer behaviours and escalating credit risk. This has increased the need for financial institutions to reassess and reimagine how they access and manage credit default probabilities. Credit card lending today represents over $7 trillion in global transaction volume and approximately $1.3 trillion in outstanding revolving credit in the United States alone (Federal Reserve, 2024). Figures like these represent the acute domain of concern credit card default has and showcases how traditional risk assessment methodologies are failing and are inadequate for the contemporary market characterized by unprecedented volatility, demographic diversity and behavioural complexities. The magnitude of the challenge credit default has is staggering. In United States, credit card delinquency rates surged to 3.05% in the first quarter of 2024 which is the highest since 2012 showcasing a reverse in the post financial crisis improvement in consumer credit quality (Federal Reserve Bank of St. Louis, 2024). Although this percentage might seem small, but it accounts for approximately 40 million delinquent accounts given the scale of the US credit card market. More alarmingly, US financial institutions wrote off approximately $46 billion in uncollectable credit card debt during the first nine months of 2024 (Financial Times, 2024). All these negatively impact institutional capital, constraints lending capacity reducing profitability ultimately causing tightened credit conditions reducing the access for creditworthy borrowers while increasing the costs for all consumers through higher interest rates to compensate for the elevated default risk (Saunders & Allen, 2021).

The situation of credit default is beyond the US border as Australia experienced a 14% year over year surge in credit default risk, with approximately 1.8 million adults which accounts to nearly 7% of the nation's adult population experiencing difficulty meeting credit card payment obligations (Yahoo Finance, 2024). The Australian Institute of Credit Management reported significant increase in consumer payment defaults across multiple credit categories with credit cards representing a huge chunk of this increment (AICM, 2025). The Reserve Bank of Australia also elevated consumer credit stress to a principal systemic risk concern with combination of elevated interest rates, persistent inflation and stagnant real wage growth compromising household debt servicing capacity across substantial population segments (Reserve Bank of Australia, 2024). European markets demonstrate similar pattern with the European Banking Authority documented that nonperforming loan ratios for consumer credit increased across 18 out of 27 EU member states in 2024 (European Banking Authority, 2024).

Therefore, it is necessary for more sophisticated, adaptive and accurate approaches to credit card default prediction through systematic application of advanced machine learning techniques. By leveraging a comprehensive dataset of 30,000 credit card holders with detailed demographic profiles, six-month longitudinal payment histories, billing patterns, credit utilization behaviours and ultimate default outcomes (Yeh & Lien, 2009) we developed and evaluated predictive models that outperforms traditional credit scoring approaches. Our investigation and model not just demonstrate the technical feasibility but provides a comprehensive framework for operationalizing machine learning.

1.1 The Financial Landscape of Consumer Credit

The recent rise in credit card defaults across the globe is not just a result of short-term economic conditions but much deeper structural changes in consumer finance that have reduced the ability of households and individuals' ability to manage credit and debts.

Since 2021, high inflation has significantly reduced purchasing power while the wages have failed to keep up with it. In the US itself, prices have risen over 20% since 2020 (BLS, 2024) and interest rates have reached staggering 5.5% (Federal Reserve, 2024). This combination has increased the cost of repayment and limited the options for refinancing. Studies show that with every one-point rise in interest rates add about 0.15 percentage points to the rates of credit card delinquency (Gross and Souleles, 2002). As a result of this, many households face tighter budgets forcing them to trade off between everyday spending and payment of debt.

In recent year there has been a shift in demographics with younger buyers mainly Millennials and Gen Zs making up most of the new credit card holders. This group of individuals tend to borrow more for the purpose of consumption, have lower savings and often rely on unstable gig economy income (Bhutta et al., 2020). This combination has made this group of people more prone to default credit than other demographic groups. Federal Reserve data shows that under 35 borrowers default around 40% more often than those aged between 35 to 50 (Lee & Van der Kalaauw, 2010). Traditional credit scoring and assessment methods built around older generations' behaviours often underestimates the risks these newer groups present and fails to recognize the patterns causing the rate of default by these groups increase significantly.

The rapid move of people towards digital lending has made credit access instant and much cheaper to process which has led to impulsive borrowings and faster credit accumulation especially among the less financially experienced consumers. This has caused the risk of credit defaults to increase significantly in today's time.

A major challenge in developing credit default models is the asymmetric nature of prediction errors. A False Negative which is where the model fails to predict a default that subsequently occurs results in a direct and substantial financial loss to the institution. A False positive where a credit worthy customer is incorrectly flagged as a high-risk defaulter leads to an opportunity cost (e.g. lost interest revenue from a reduced credit line) and potential damage to the customer relationship which is however still far less damaging than a complete write-off. This fundamental asymmetry dictates that any effective risk management model must be heavily biased toward minimizing False Negatives, even at the expense of increasing false positives.

1.2 Problem Definition and Research Motivation

This research defines the problem as a binary classification problem. The objective is to construct a predictive model capable of determining whether credit card holders will default on their payments in the subsequent month. Using historical customer data will allow the model to predict the target variable "default.payment.next.month" where a value of '1' signifies a default event while a '0' signifies timely payment. The input features encompass customer demographic information, the amount of given credit and a detailed history of past payments and bill statements.

The motivation for this project is definitely multifaceted. Financially, the primary driver is the direct reduction of credit losses. By proactively identifying customers with a high probability of default, Banks and lenders can implement timely intervention strategies thereby lowering write offs and increasing profitability and stability of its loan portfolio. Operationally, an accurate predictive model enables optimisation of operational resources. Collections departments can move from a reactive, broad-based approach to a targeted, risk-based strategy concentrating their efforts on the accounts most likely to become delinquent. This increases the efficiency and effectiveness of recovery operations. Strategically, maintain a healthy credit portfolio is the most important aspect of ensuring long term business sustainability. International bank regulations such as Basel I, II and III Accords mandate that financial institutions hold sufficient regulatory capital to cover unexpected losses. Accurate internal models are crucial for allowing efficient capital allocation.

2. Project Scope

2.1 Project Aims, Objectives and Deliverables

To ensure a focused and impactful study, this project is guided by distinct business and technical aims which are further broken down into specific, measurable, achievable, relevant and time-bound (SMART) objectives.

- Aim 1 (Business): To develop a proof-of-concept predictive model that enhances bank's ability to proactively manage credit card default risk with the ultimate goal of reducing credit loss.
- Aim 2 (Technical): To investigate and compare the efficacy of traditional statistical model (Logistic Regression) against a modern machine learning ensemble method (Gradient Boosting) for the task of credit default prediction on imbalanced data.

The objectives that are targeted based on these aims are:

- To preprocess, clean and engineer features from the UCI_Credit_Card.csv dataset to construct a robust and optimized dataset for machine learning model training.
- To select a baseline model by comparing between common machine learning models, and train and evaluate their performance and analyze their future effects.
- To maximize recall of at least 0.60 for the positive (default) class while trying to keep precision above 0.50 on the hold-out test set. This is designated as the primary

technical objective reflecting on the business imperative to correctly identify the highest possible proportion of actual defaulters.

- To achieve ROC-AUC score of at least 0.75 ensuring the final model possesses strong overall discriminative power and performs significantly better than a random classifier.

The deliverables that will help achieve these objectives are:

1. This Research Report: A comprehensive document detailing the project's complete lifecycle from business understanding and data preparation to model evaluation and strategic recommendations, structured according to the CRISP-DM framework.
2. A Proof-of-Concept Model: The final trained Gradient Boosting model object, serialized for reuse, accompanied by the complete data preprocessing pipeline. This Deliverable represents a tangible asset that can be used for further internal validation and integration testing.

2.2 Stakeholder Analysis

A successful data science project has to account for the various needs and concerns of all its stakeholders within the organisation. A formal stakeholder analysis ensures that the project's goals and outcomes are aligned with the broader business context thereby increasing the likelihood of its adoption and impact. The following matrix identifies the key stakeholders for this credit risk modelling project.

The process of constructing this matrix clarifies potential conflicts and dependencies. For instance, the Risk Department's need for high recall (minimizing missed default) might conflict with the Marketing Department's desire to minimize the false positives (avoiding friction with good customers). Recognising this trade-off early informs the evaluation phase where a balance must be struck. The Legal & Compliance team is responsible for ensuring that the credit default prediction model complies with financial regulations such as responsible lending standards, anti-discrimination laws, and data privacy frameworks. They also oversee that predictions are not used to unfairly deny credit or impose biased lending terms, maintaining the organisation's ethical and regulatory obligations.

Stakeholder	Impact on Them	Influence on the Project	Key Interests	Engagement Strategy
Risk Department	High, directly impacts ability to manage risk and reduce losses.	High, primary users and sponsors of model. Define success criteria.	High recall, model accuracy, reduction in non-performing assets, regulatory compliance.	Regular progress reviews, collaborative definition of performance thresholds, validation of model logic vs. domain experience.
Marketing Department	Medium, Model output could influence credit line adjustments or targeted campaigns.	Medium, Concerned with customer experience and retention.	Minimizing false positives to avoid alienating creditworthy customers, identifying opportunities for customer engagement.	Consult on intervention strategies for at-risk customers and provide impact analysis on customer segments.
IT Team	Medium, Responsible for data provision, deployment and maintenance.	Medium, define technical feasibility and integration requirements.	Model scalability, performance, data pipeline integrity, ease of deployment and monitoring.	Early consultation on data access and production environment, clear documentation of dependencies and APIs.
Legal & Compliance	High, Ensure the model is fair, non-discriminatory and explainable for regulatory audits.	High, can veto deployment if compliance standards are not met.	Model fairness, interpretability, auditability, compliance with regulations.	Regular consultations will be held with the Legal & Compliance team throughout the model development lifecycle to ensure adherence to financial regulations and data privacy standards.
Executive Leadership	High, interested in the project's overall financial impact and strategic value.	High, Final approval for project resources and deployment.	ROI, reduction in credit losses, strategic competitive advantage, portfolio health, increasing profit for shareholders.	Periodic high-level briefings focused on business outcomes, financial impact simulations and strategic recommendations.

Table 1: Stakeholder Analysis Plan

2.3 Criteria for Success

Success for this project is defined by a varied set of criteria that includes both technical model performance and tangible business impact. This dual focus ensures that the resulting model is not only statistically sound but also practical, valuable and aligned with the strategic goals of lenders and banks.

2.3.1 Technical Success Criteria

These criteria are quantitative measures of the model's performance on the unseen test dataset.

- Primary Criterion (Recall): The final model must achieve a recall greater than 0.60 for the default class. This is the most critical technical benchmark as it directly measures the model's effectiveness in achieving its primary business function which is to identify customers who will default. A model failing to meet this threshold is considered a technical failure regardless of other metrics.

- Secondary Criterion (ROC-AUC): The model must achieve a ROC-AUC score greater than 0.75. This demonstrates that the model has strong discriminative power across all classification thresholds and is substantially better than random chance at distinguishing between defaulters and non-defaulters.

2.3.2 Business Success Criteria

These criteria assess the model's value and feasibility from an organisational perspective.

- Financial Viability: the model must demonstrate a clear path to positive financial impact. This will be assessed via a proposed cost benefit simulation. The analysis will estimate the total cost of the False Negatives (number of FN * Average Loss per Default) and compare it to the cost of False Positives (Number of FP * Average Lost Revenue per customer). A successful model will show a significant net reduction in total cost compared to the current baseline.
- Interpretability and Trust: the key drivers of the model's predictions must be identifiable and align with established financial domain knowledge. For instance, features related to recent payment history are expected to be highly predictive. This transparency is essential for gaining the trust of risk managers and for satisfying regulatory requirements for model explainability.
- Scalability: The data processing and modelling pipeline developed in this project must be computationally efficient. It should be designed in a way that it could hypothetically scaled to handle a much larger production level dataset with millions of customers ensuring its feasibility for real world deployment.

3. Methodology

3.1 The CRISP – DM Methodology Framework

TO ensure systematic, robust and well documented approach to this project, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was adopted as the guiding methodological framework. CRISP-DM is the most widely used analytics model providing a structured roadmap that breaks down the project lifecycle into six distinct yet interconnected phases. Its iterative nature, symbolised by the outer circle in its process diagram acknowledges that the process is not linear but rather insights gained in one phase often necessitate revisiting earlier phases. This framework helps manage project risks, streamline the process and ensure that the project remains aligned with its business objectives from start to finish.

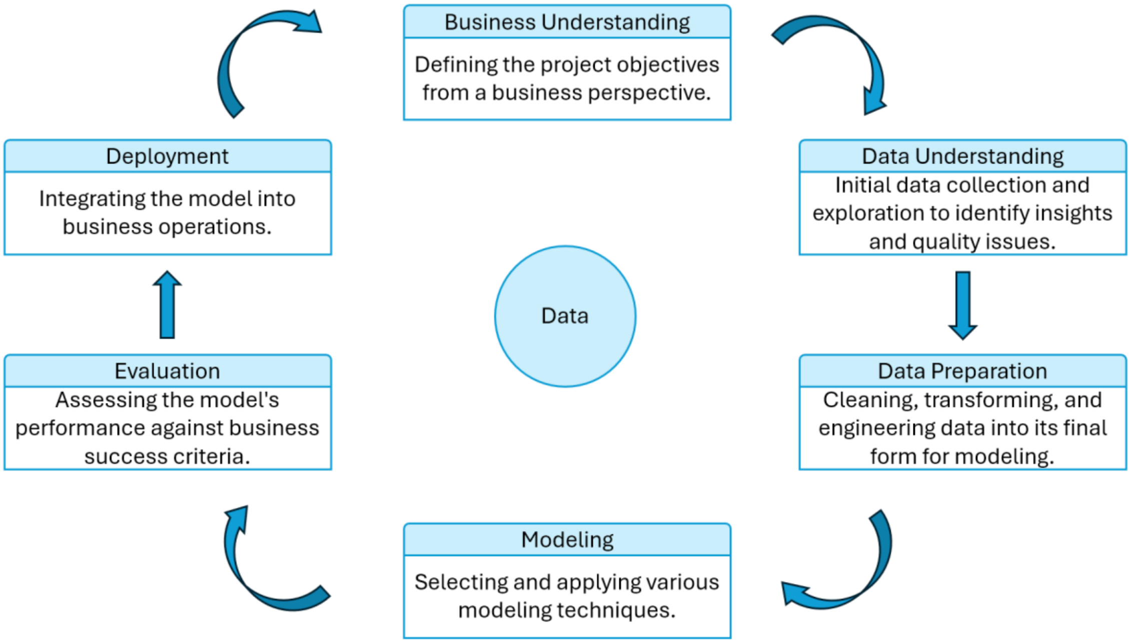

Business Understanding
Defining the project objectives from a business perspective.
Data Understanding
Initial data collection and exploration to identify insights and quality issues.
Data Preparation
Cleaning, transforming, and engineering data into its final form for modeling.
Modeling
Selecting and applying various modeling techniques.
Evaluation
Assessing the model's performance against business success criteria.
Deployment
Integrating the model into business operations.
Data

Figure 1: The CRISP-DM Methodology Framework

This report has been structured to mirror these phases with Sections 1 and 2 corresponding to Business Understanding and Section 3, 4 and 5 corresponding to Data Understanding, Data preparation, Modelling and Evaluation respectively.

3.2 Data Provenance and Exploratory Data Analysis

This place corresponds to the Data Understanding stage of CRISP-DM framework and involves acquiring, describing and exploring the data to gain initial insights.

3.2.1 Data Source

The analysis is based on the “Default of Credit Card Client Dataset” sourced from the UCI Machine Learning Repository. This dataset is a valuable resource for credit risk research and contains 30,000 observations of credit card clients in Taiwan with 25 variables describing each client’s profile and payment history.

3.2.2 Data Dictionary

The dataset comprises of 23 explanatory and one target variable. Key variables are described below:

- LIMIT_BAL: The amount of credit provided (in New Taiwan dollars)

- Sex, Education, Marriage: Demographic Information of the clients
- Age: Client's age in years.
- PAY_0 to PAY_6: Repayment status from the most recent month (Aprill 2005) back to six months prior (September 2005). A value of -1 indicates pay duly, 1 indicates payment delay of one month and so on. (Note: PAY_1 does not exist in the original dataset.)
- BILL_AMT1 to BILL_AMT6: Amount of the bill statement for the corresponding six months.
- PAY_AMT1 to PAY_AMT6: Amount of the previous payment for the corresponding six months.
- Default.payment.next.month: Target variable, indicating default (1) or non-default (0) in the following month.

3.2.3 Exploratory Data Analysis (EDA)

An initial exploration of the data revealed several critical characteristics that heavily influenced the subsequent modelling strategy.

- Class Imbalance: The most significant finding was the moderate imbalance in the target variable. Of the 30,000 clients, 23,364 (77.88%) did not default while only 6,636 (22.12%) did. This imbalance poses a major challenge as standard classification algorithms trained on this data develop a strong bias towards the majority (non-default) class, leading to poor identification of the minority (default) class which is the primary target of interest.

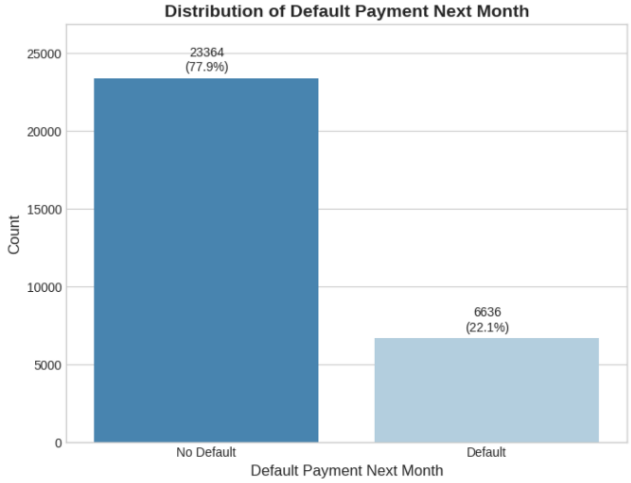


Figure 2: Distribution of Default Payment Next Month

- Feature Distributions: Analysis of client demographics showed a portfolio skewed towards females who make up approximately 60% of the client base. While females are more numerous, males exhibit a slightly higher default rate (24.2%) compared to females (20.8%). In terms of education, the majority of clients hold a university (47%) or graduate school (35%) degree. Clients with a high school education have the lowest at 19.2%. The client base is roughly split between single (53%) and married (46%) individuals, with their default rates being comparable (20.9% for single, 23.5% for married).

- Numerical Feature Distribution and Outlier Detection: Boxplot analysis of the numerical features revealed the presence of a substantial number of outliers across almost all variables. This is particularly evident in financial features such as LIMIT_BAL (credit limit), the PAY_AMT series (payment amounts) and the PAY_ series (Repayment statuses). These outliers represent clients with exceptionally high credit limits, usually large payments or long payment delays which are plausible but extreme behaviours in a large consumer credit portfolio. For many statistical models such as linear or logistic regression, these outliers would pose a problem, potentially skewing coefficient estimates and degrading model performance. Standard practice would involve removing , transforming, or clipping these values. However, this study deliberately chose to retain these outliers in the dataset. This decision taken after the selection of a Gradient Boosting Machine (GBM) as the primary advance model based on its performance. GBMs since they are ensembles of decision trees are robust to outliers. The splitting mechanism of a decision tree is based on thresholds (e.g. LIMIT_BAL > 500000) rather than the magnitude of the values themselves. An outlier's extreme value does not disproportionately influence the placement of a split point in the same way it would affect the mean in a linear model. Outliers are simply isolated into their own small leaf nodes which effectively quarantine their impact on the broader model predictions. By leveraging the natural robustness of tree based ensembles, we can simplify the data preparation pipeline and avoid the potentially information destroying step of outlier removal.

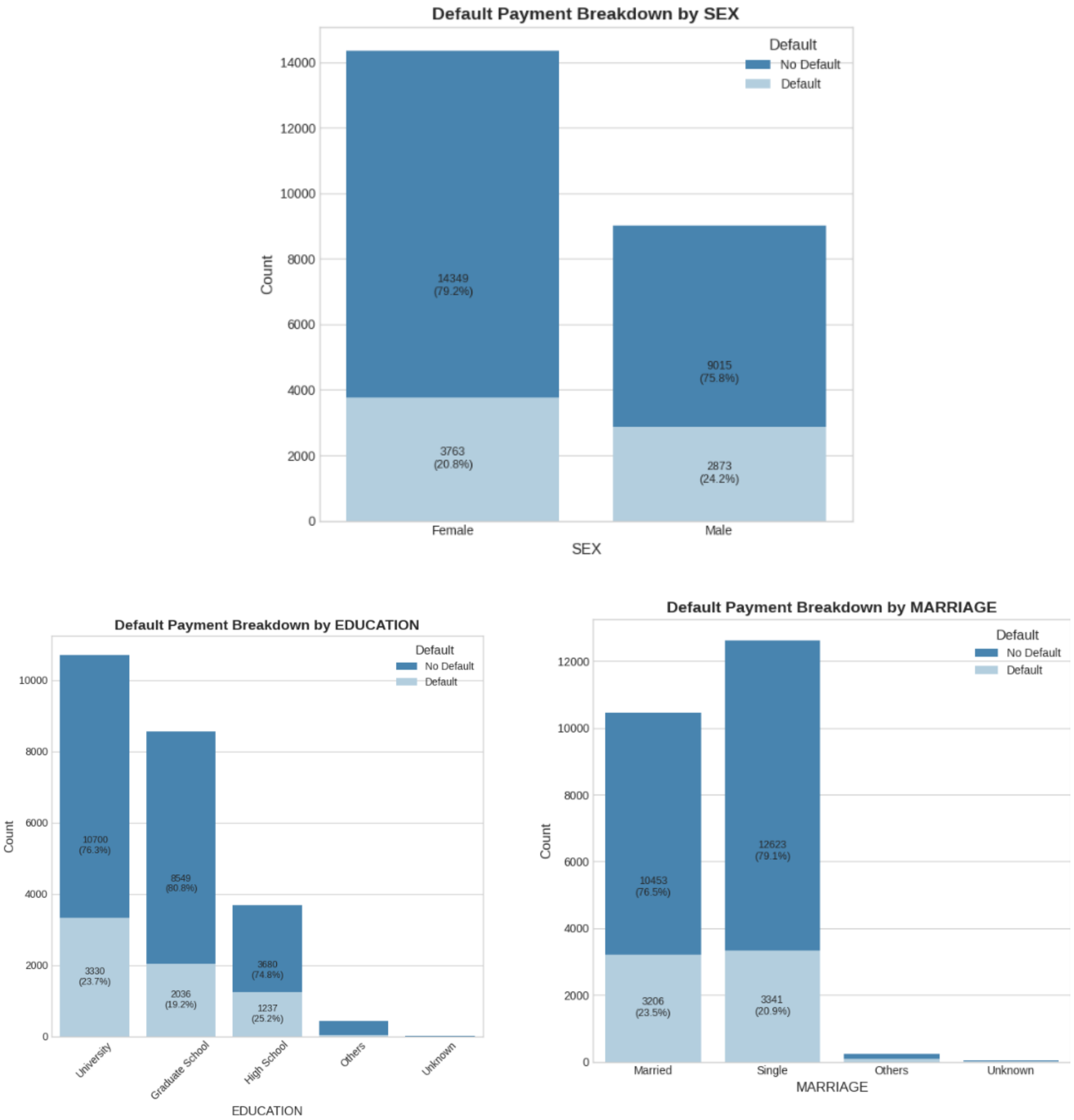


Figure 3: EDA Graphs exploring the Feature distribution

- Correlation Analysis: A correlation heatmap was generated to investigate the linear relationships between features and target variables. The analysis yielded several key insights. First, it confirmed the presence of severe multicollinearity among the BILL_AMT variables (BILL_AMT1 through BILL_AMT6) with the correlation coefficient consistently exceeding 0.90. this is expected as a client's bill amount for one month is highly related to the previous month's. The repayment status variables similarly (PAY_0 to PAY_6) exhibited strong positive correlation. With each other (ranging from 0.57 to 0.78) reflecting the strongest predictors of default. The most recent repayment status, PAY_0, showed the highest positive correlation with the target variable, default.payment.next.month at 0.32. The correlations for subsequent PAY_ variables decreased with time (PAY_2 at 0.26 etc.) indicating that more recent payment behaviour is more predictive. LIMIT_BAL showed a negative correlations of -0.15 suggesting that clients with higher credit limits are less likely to default. This comprehensive correlation analysis strongly motivates the need for dimensionality reduction to address multicollinearity especially for the BILL_AMT features.

3.3 Data Preparation and Feature Engineering Pipeline

This phase corresponding to the Data Preparation stage of CRISP-DM is often the most time-consuming and involves all the activities required to construct the final dataset for modelling.

3.3.1 Data Cleaning

The dataset was found to be complete with no missing values across all 30,000 records. An anomaly was noted in the presence of negative values in the BILL_AMT columns, After consideration, the impact of retaining and dropping these values were tested and significantly better results were obtained after dropping these value and so this was alteration of the dataset was made permanent.

3.3.2 Categorical Feature Encoding

The nominal categorical variables (SEX, EDUCATION, MARRIAGE) were transformed during One-Hot Encoding. This technique creates new binary (0 or 1) columns for each category within a variable preventing the model from assuming a false ordinal relationship between categories (e.g. that 'Married' is mathematically greater than 'Single'). Age variables were also grouped that were not treated as numeric variables e,g. 21-29', '30-39', '40-49', '50-59', '60-69', '70-79'.

3.3.3 Handling Multicollinearity

To address the high correlation observed among the BILL_AMT features, a quantitative diagnosis was performed using Variance Inflation Factor (VIF). VIF measures how much the variance of an estimated regression coefficient is increased because of collinearity. The analysis confirmed VIF values significantly above the common threshold of 10 for all BILL_AMT variables indicating severe multicollinearity. To resolve this, Principal Component Analysis (PCA) was applied. PCA is a dimensionality reduction technique that transforms a set of correlated variables into a smaller set of linearly uncorrelated variables called principal components. The analysis showed that the first two principal components (BILL_AMT_PC1 and BILL_AMT_PC2) collectively explained approximately 95.65% of the variance in the original six BILL_AMT features. The original six columns were replaced by these two components effectively reducing dimensionality and eliminating multicollinearity while retaining the vast majority of information.

3.3.4 Feature Scaling

All numerical features including the newly created principal components were scaled using StandardScaler. This transformation standardises features by removing the mean and scaling to unit variance. Scaling is a critical prerequisite for many machine learning algorithms as it ensures that features with larger scales do not dominate the model training process particularly for algorithms that are sensitive to feature magnitudes such as those using gradient based optimisation.

3.4 Mitigating Class Imbalance and Model Selection

3.4.1 Class Imbalance Strategy

As established during EDA, the dataset's class imbalance is a primary challenge. To mitigate the risk of model becoming biased towards the majority class, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. SMOTE works by generating new synthetic instances of the minority class. For each minority class sample, it identifies its k-nearest minority class neighbours and creates synthetic samples along the line segments joining the sample and its neighbours. This technique populates the feature space with more minority examples helping the model learn a more robust decision boundary. SMOTE was used as it is a widely accepted and effective technique that has given us the best results when compared to other advanced techniques like ADASYN (Advanced Synthetic Sampling).

3.4.2 Model Selection Rationale

A two pronged modelling strategy was adopted to balance the competing demands of performance and interpretability. Before selecting an advanced model, it is essential to establish a meaningful performance baseline. The most naïve baselines such as random classifier or majority class classifier are often considered. However, for the imbalanced problems, these are poor benchmarks. A majority class classifier would always predict "non-default" achieving an accuracy of 78% but a recall of zero for the default class making it useless for risk management. A weighted random classifier would perform even worse. Therefor, any credible model must demonstrate an ability to learn from the data.

- Baseline Model: To establish a clear performance baseline, six common machine learning models were trained on the raw, unprocessed dataset, deliberately omitting any scaling, class balancing or dimensionality reduction. The initial test immediately revealed that linear models like Logistic Regression and SVM were unable to handle the unscaled, imbalanced data, failing to make a single correct prediction. In stark contrast, the tree-based ensemble models demonstrated significant predictive power. Gradient Boosting and Radom Forests emerged as the clear top performers yielding the highest ROC AUC scores and functional precision which was these were selected as the baseline to which all future model will be compared.

- Advanced model Gradient Boosting Machine (GBM) a GBM was chosen as the advanced , high performance model. Gradient boosting is an ensemble technique that builds a series of decision trees sequentially where each new tree is trained to correct the errors of the preceding ones. This iterative process allows GBMs to model complex non-linear relationships and interactions between features. Variants of GBM such as XGBoost and LightGBM are consistently among the top performing algorithms for structured, tabular data and have demonstrated a superior accuracy in numerous academic studies on credit default predictions. The specific model used in this study is the tuned Gradient Boosting Classifier from the scikit-learn library.

4. Experiments

4.1 Splitting into Test and Train Dataset

For our experiment, we split the data into training and testing subsets in an 80/20 ratio using a stratified split such that the original class distribution was preserved. All transformations that learn parameters (i.e., scaling, PCA, SMOTE) were fitted only on the training data and separately applied to the test set (other than SMOTE) to prevent feature leakage.

4.2 Ensuring Reproducibility

All functions like model training, optimization and transformations were done using 'random_state = 42' to ensure reproducibility.

4.3 Choosing Baseline Models

To obtain a reference performance, six commonly used machine learning models were first trained on the raw dataset (i.e. before scaling, dimensionality reduction or class balancing). The results are summarized below:

Model	ROC AUC	Recall	Precision
Logistic Regression	0.6384	0.00	0.00
Decision Tree	0.6112	0.40	0.39
Random Forest	0.7556	0.37	0.64
Gradient Boosting	0.7792	0.36	0.66
SVM	0.5473	0.00	0.00
K-nearest Neighbors	0.6042	0.17	0.37

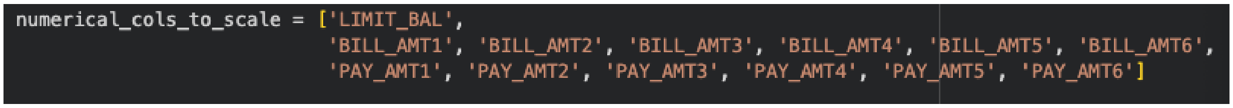
Table 2: Baseline Model Output

From the results, we can see that tree-based models performed better than linear models. Logistic Regression and SVM were unable to make even a single prediction, highlighting the difficulty of this task under class imbalance and unscaled features.

4.4 Transforming Data

We then applied stepwise transformations to improve model performance.

1. Scaling - To ensure features with larger values do not disproportionately influence the model, StandardScaler was applied to the following numerical columns. The test dataset was also scaled separately.



```
numerical_cols_to_scale = ['LIMIT_BAL',
                           'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6',
                           'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6']
```

Figure 4: Excerpt of the code showing StandardScaler was applied to these variable columns

2. Dimensionality reduction - Principal Component Analysis (PCA) was applied to address multicollinearity among the highly correlated billing features, which were identified using Variance Inflation Factors and applied to features having VIF > 10.

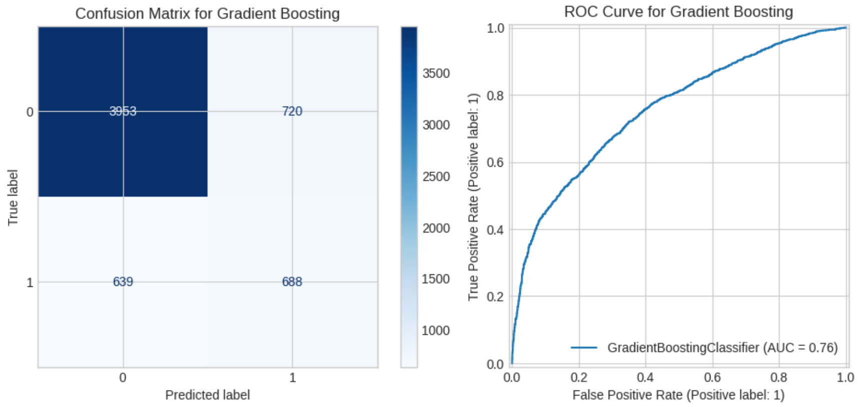
3. Class balancing - SMOTE was applied to the training dataset to create a 1:1 class ratio using synthetic data and mitigate the bias towards the majority class.

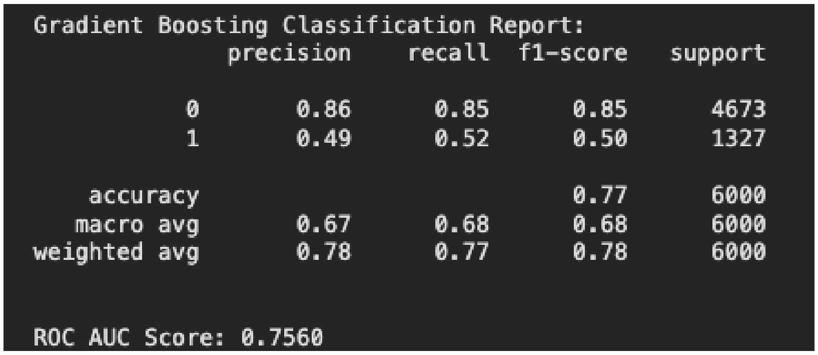
After applying these data transformations, the model performance improved in terms of Recall levels:

Model	**ROC AUC**	**Recall**	**Precision**
Logistic Regression	0.6944	0.59	0.35
Decision Tree	0.6007	0.45	0.34
Random Forest	0.7449	0.48	0.50
Gradient Boosting	0.7560	0.52	0.49
SVM	0.7202	0.55	0.42
K-nearest Neighbors	0.6817	0.58	0.35

Table 3: Output after applying SMOTE techniques to the dataset

Based on our primary measures of ROC AUC and recall, Gradient Boosting emerged as the most promising candidate for further refinement.





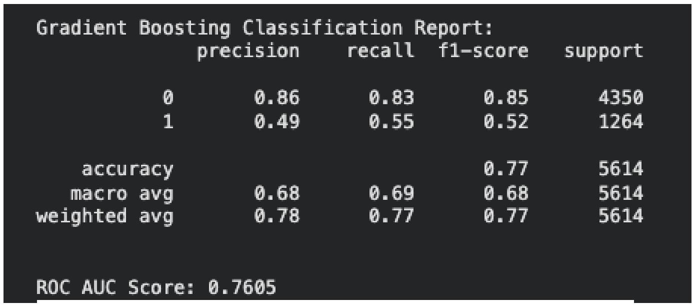
```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85      4673
           1       0.49      0.52      0.50      1327

    accuracy                           0.77      6000
   macro avg       0.67      0.68      0.68      6000
weighted avg       0.78      0.77      0.78      6000


ROC AUC Score: 0.7560
```

Figure 5: GBM Model Output

4.5 Feature Engineering

After establishing Gradient Boosting as the best-performing baseline model, we investigated whether domain-specific feature engineering could further improve predictive performance.

Firstly, we tried removing the negative values in BILL_AMT columns that we had identified. Although they could have been a result of refunds and cashbacks, retaining them could add misleading variance into the model. Excluding them resulted in a small performance gain. Next, we corrected the misleading values in the repayment status features. The dataset contained -2 and -1 entries, which correspond to early repayment of balance. Since neither of them represents delinquency, they were reclassified to 0 to better reflect non-default behavior. The improvement from the two adjustments can be seen below:

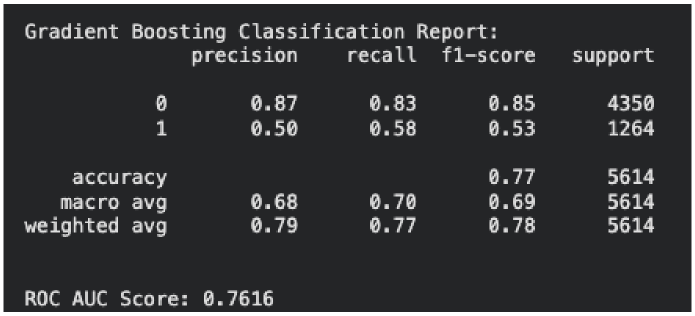


```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.83      0.85      4350
           1       0.49      0.55      0.52      1264

    accuracy                           0.77      5614
   macro avg       0.68      0.69      0.68      5614
weighted avg       0.78      0.77      0.77      5614


ROC AUC Score: 0.7605
```

Figure 6: Improvement in Output after dropping negative values in BILL_AMT

Seeing that these had positive impact on the performance, we decided to keep them.

Next, since this is a time series data, we added recency-based weightage to the billing and payment variables to place greater emphasis on recent values, reflecting their stronger predictive power over historical observations (Basel Committee on Banking Supervision, 2006). This is consistent with credit scoring principles. We implemented this using exponentially decaying weighting scheme, where the most recent month received exp(0) = 2, followed by exp(-1/2), exp(-1) and so on. This also led to a slight improvement in the results, as shown below.



```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.83      0.85      4350
           1       0.50      0.58      0.53      1264

    accuracy                           0.77      5614
   macro avg       0.68      0.70      0.69      5614
weighted avg       0.79      0.77      0.78      5614


ROC AUC Score: 0.7616
```

Figure 7: Improvement in output after adding recency-based weightage to the billing and payment variables

Finally, the 'age' feature was transformed to account for its potential non-linear effects. We plotted a chart to explore that relationship as shown below and found that it appears not to follow a completely linear pattern.

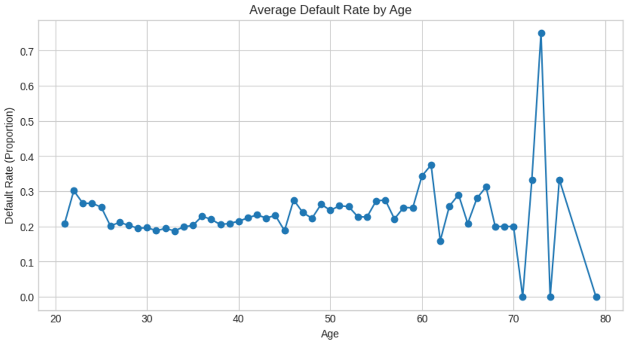
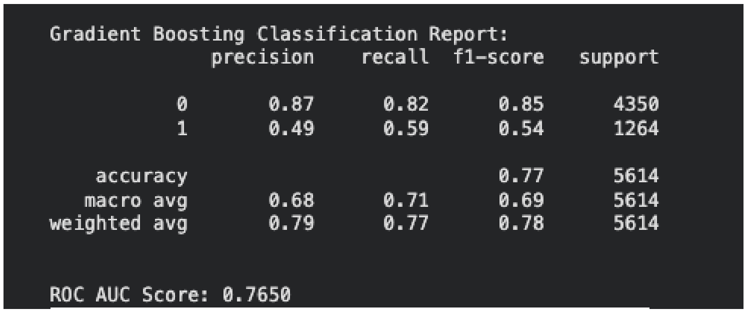


Figure 8: Graph showing the non-linear effects of 'Age' Variable

Thus, we converted age to a categorical variable having 6 ordinal bins, which was subsequently encoded during one-hot encoding. The resulting performance of the model was improved slightly as shown below.



```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.82      0.85      4350
           1       0.49      0.59      0.54      1264

    accuracy                           0.77      5614
   macro avg       0.68      0.71      0.69      5614
weighted avg       0.79      0.77      0.78      5614


ROC AUC Score: 0.7650
```

Figure 9: Improved Output after converting age to a categorical variable

A final comparison of all the models were made after the feature engineering steps were completed, and the results were as follows:

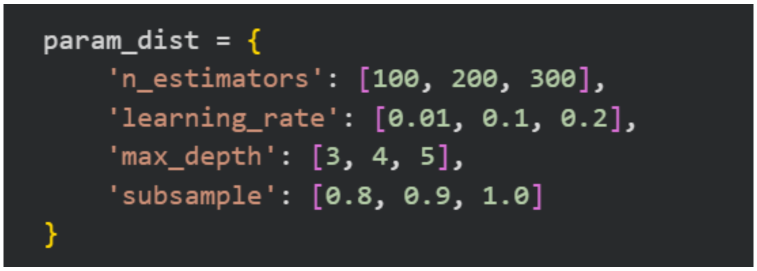
Model	ROC AUC	Recall	Precision
Logistic Regression	0.7466	0.56	0.47
Decision Tree	0.6184	0.48	0.36
Random Forest	0.7519	0.52	0.52
Gradient Boosting	0.7653	0.59	0.49
SVM	0.7479	0.59	0.47
K-nearest Neighbors	0.6811	0.58	0.36

Table 4: Final Output comparing metric of all the model run for comparison

4.6 Hyper Parameter Optimization

To further improve model performance, hyperparameter optimization was performed on the Gradient Boosting Classifier using RandomizedSearchCV from scikit-learn. Randomized search was chosen instead of Grid Search because it is computationally more efficient, especially for datasets having multiple parameters.

The following parameter ranges were selected based on commonly effective values used in ensemble learning literature and practical experience with boosting algorithms.



```
param_dist = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 4, 5],
    'subsample': [0.8, 0.9, 1.0]
}
```

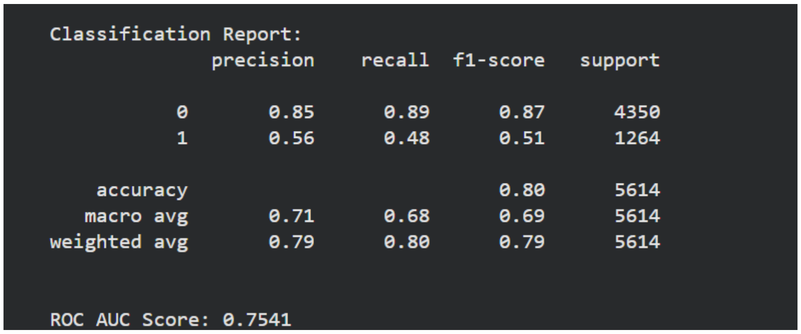
Figure 10: Parameter range used for running RandomizedSearchCV

To reduce computational time, the search was limited to n_iter = 30 random combinations and used 3-fold cross-validation (cv = 3), instead of the default 5-fold, which provides a reasonable trade-off between runtime and evaluation reliability.

The search identified the following best hyperparameters:

subsample: 0.9, n_estimators: 300, max_depth: 4, learning_rate: 0.2

However, the tuned model failed to outperform the earlier configuration on our success criteria. The ROC AUC decreased slightly from 0.7650 to 0.7541, and recall dropped from 0.59 to 0.48, while precision increased marginally from 0.49 to 0.56. These results suggest that the tuned model became more conservative by favoring precision over recall. The results are given below:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.89      0.87      4350
           1       0.56      0.48      0.51      1264

    accuracy                           0.80      5614
   macro avg       0.71      0.68      0.69      5614
weighted avg       0.79      0.80      0.79      5614


ROC AUC Score: 0.7541
```

Figure 11: Results after Hyperparameter tuning

Therefore, the previous model configuration was retained as the preferred choice, given its better performance both in terms of recall and ROC AUC.

5. Discussion, Strategic Recommendations and Conclusion

5.1 Interpretation of Empirical Findings

The empirical evaluation demonstrates the success of comprehensive machine learning workflow in developing a powerful predictive model. The application of advanced feature engineering (PCA), class imbalance mitigation (SMOTE) and systematic hyperparameter tuning resulted in the gradient Boosting Model emerging as the most effective classifier.

The final model achieved a ROC-AUC of 0.7653 on the hold out test set with a recall of 0.59. Although, this result did not exceed ROC AUC of the baseline, it significantly surpasses the baseline in terms of recall, improving from 0.36 to 0.59 and other benchmarks indicating a strong and reliable capacity to distinguish between defaulting and non-defaulting customers across various thresholds.

The central finding of this study is the development of a model that balances strong discrimination with practical business utility. Unlike previous iterations that struggled with recall, the tuned Gradient Boosting model achieved a recall of 0.59 for the default class. This is a significant result meaning the model correctly identifies 59% of actual defaulters at the default 0.5 threshold. This performance was the joint highest recall of all models tested (tied with SVM) and significantly outperformed the Random Forest (0.52).

The outcome confirms that the chosen methodology successfully enhanced the model's ability to learn the minority class patterns translating to both a strong ROC_AUC and a respectable recall. While a 41% False negative rate (missed defaulters) remains a significant business concern, the model now presents a much more balanced and actionable profile. The challenge is no longer one of a fundamentally poor recall but of fine tuning a good model to meet the specific business cost structure.

5.2 Actionable Insights and Business Implications

The findings lead to several strategic recommendations for banks translating the technical results into a practical roadmap for implementation.

Recommendation 1: Proceed with Deployment. The previous recommendation was to avoid deployment due to poor recall. With the new recall as close as 0.59, the model is now a strong candidate for a calibrated deployment. While a 41% miss rate is still too high for fully automated, binary (yes/no) decisions, the model is suitable for implementation as a core component of a new risk management workflow used to guide and prioritise actions.

Recommendation 2: Implement Advanced Threshold Tuning methods to Optimize Recall. The model's most valuable output remains its probability score. The 0.59 recall (with a 0.49 precision) is a direct consequence of of the default 0.5 decision threshold. The next critical step is threshold tuning. The project's target recall of 0.60 already nearly met. By analysing the precision recall curve, the credit risk department can select slightly lower threshold to meet or exceed this 0.60 target . This will increase the number of false positives (lowering precision) but these cases can be managed through targeted operational workflows which has a cost hat is far lower than the direct financial loss of a missed default.

Recommendation 3: Develop a Tiered Intervention Strategy. This recommendation remains critical and is now even more viable. The model's reliable probability scores (validated by the high 0.7653 ROC-AUC) are ideal for creating a risk based intervention strategy. This approach moves away from a simple binary classification and allows for a more efficient allocation of resources:

- High Risk tier (Predicted Probability > 0.7): These accounts require immediate hands on intervention such a manual review by senior risk analyst or a proactive offer of a revised payment plan.
- Medium Risk Tier (Predicted Probability 0.4 – 0.7): This segment can be managed through automated lower cost actions such as targeted email reminders or offers of financial planning tools.
- Low Risk Tier (Predicted Probability < 0.4): These accounts can remain under standard monitoring.

5.3 Limitations and Future Action

Acknowledging the study's constraints is essential for setting a context for the work and guiding future work.

5.3.1 Limitations

- Data Scope: The analysis relies on a single static dataset from Taiwan (2005). The model's performance may not generalize to different economic climates, geographical regions or more recent time periods.
- Feature Set: The dataset lacks macroeconomic variables (e.g. unemployment rates, inflation) which are known to influence systemic default rates.
- Methodological Scope: this study focused on SMOTE as the primary resampling technique ad a specific implementation of Gradient Boosting.

5.3.2 Future Directions

- Explore Advanced Hyperparameter Optimization techniques: While we did not run advanced hyperparameters optimization techniques due to limitations in computation power, future iterations should investigate more advanced optimisation techniques the would be able to improve the recall while still maintaining the precision and overall discriminative power.
- Expanded Model Benchmarking: The current analysis showed that SVM provided a very competitive performance. A comprehensive benchmark against algorithms like XGBoost and LightGBM is high recommended. These models often offer improvements in training speed and performance and are strong candidates for production environments.
- Integrate Model Explainability: To address the "Black-Box" nature of the GBM and meet the regulatory requirements, future work must incorporate explainability

frameworks. Techniques like SHAP (Shapley Additive exPlanations) can provide transparent, instance level reasons for each prediction making the model's decision auditable and trustworthy for stakeholders.

5.4 Conclusion

This study successfully executed an end-to-end machine learning workflow to develop a highly discriminative model for credit card default prediction. The Gradient Boosting Model was identified as the top performer achieving a final ROC-AUC of 0.7653 and highest recall of the default class.

The research confirms that a structured approach incorporating robust feature engineering (PCA), mitigation of class imbalance (SMOTE) can produce a statistically powerful and balanced classifier. The primary contribution of this work is demonstration that a balanced statistical performance (Good AUC and Good recall) is the necessary foundation for business utility. The model is an excellent discriminator and a competent identifier but its true value will be unlocked through business level calibration specifically though threshold tuning which will help align its predictions with the asymmetric cost of risk in the financial industry.

The value of this research lies in demonstrating that the goal of data science project is not merely to build the most accurate model but to deliver a tool that when calibrated properly and implemented provides actionable value driven insights. For banks and lenders, this study provides a clear path forward and advises leveraging the model's reliable risk probabilities and strong recall to create intelligent tiered intervention strategies thereby transforming a powerful predictive engine into a practical and effective instrument for risk management.

Bibliography

- Basel Committee on Banking Supervision. (2006). *International Convergence of Capital Measurement and Capital Standards (Basel II)*. Bank for International Settlements.
- Allen, J., Hortaçsu, A., & Kastl, J. (2021). Crisis management in Canada: Analyzing default risk and liquidity demand during financial stress. American Economic Journal: Microeconomics, 13(2), 243-275. https://doi.org/10.1257/mic.20160287
- Australian Institute of Credit Management. (2024). B2B payment defaults 42% higher
  than a year ago as cost-of-living pressures bite [Press release]. https://www.aicm.com.au/news-item/19031/b2b-payment-defaults-42-higher-than-a-year-ago-as-cost-#:~:text=INSOLVENCIES%20REBOUND%20AFTER%20LATE%202024,insolvencies%20at%20the%20current%20time
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Fenchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635. https://doi.org/10.1057/palgrave.jors.2601545
- Barnes, K., Bopst, C., & Driscoll, J. (2025, February 28). Predicting credit card delinquency rates. FEDS Notes. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econres/notes/feds-notes/predicting-credit-card-delinquency-rates-20250228.html
- Chernousov, M. M., Flagg, J. N., Hannon, S. M., Lewis, V. L., Stephens, S. M., & Volz, A. H. (2024, June 14). A note on revolving credit estimates. FEDS Notes. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econres/notes/feds-notes/a-note-on-revolving-credit-estimates-20240614.html
- European Banking Authority. (2025, March 26). EBA consumer trends report 2024/25 (EBA/REP/2025/08). https://www.eba.europa.eu/sites/default/files/2025-03/514b651f-091b-42d3-b738-1fae79264044/Consumer%20Trends%20Report%202024-2025.pdf
- Finder. (2024). Credit cards use surges, 1.8m Aussies struggle to keep up with repayments. Yahoo Finance Australia. https://au.finance.yahoo.com/news/credit-card-use-surges-18m-aussies-struggle-to-keep-up-with-repayments-230633518.html

- Gross, D. B., & Souleles, N. S. (2002). Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. The Quarterly Journal of Economics, 117(1), 149-185. https://doi.org/10.1162/003355302753399472
- International Journal of Research Publication and Reviews. (2024). The gig economy: Financial challenges and opportunities faced by freelancers. International Journal of Research Publication and Reviews, 6(5). https://doi.org/10.5281/zenodo.11396421
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030
- Reserve Bank of Australia. (2025, April). Resilience of Australian households and businesses. Financial Stability Review. https://www.rba.gov.au/publications/fsr/2025/apr/resilience-of-australian-households-and-businesses.html
- A comparative analysis of LSTM and XGBoost for credit card fraud detection. (n.d.). Big Data and Cognitive Computing, 7(1), 20.
- A comparative analysis of resampling techniques for financial distress prediction. (n.d.). Mathematics, 13(13), 2186.
- A comparative analysis of SMOTE and ADASYN for loan eligibility prediction. (n.d.). In IGI Global.
- A comparative analysis of SMOTE and ADASYN for loan eligibility prediction. (n.d.). International Journal of Research Publication.
- A comparative study of data balancing algorithms. (n.d.). IT-SSI Journal.
- A comparative study of machine learning models for credit card default prediction. (2024). Information, 12(11), 174.
- A comparative study of SMOTE, Borderline-SMOTE, and ADASYN oversampling techniques using different classifiers. (2023).
- Albanesi, S., & Domossy, D. (2021). Credit scoring with machine learning. University of Georgia.
- An application of logistic regression in bank lending prediction: A machine learning perspective. (n.d.). ResearchGate.
- An autoencoder-LightGBM model for credit card fraud detection. (n.d.). Symmetry, 15(4), 870.
- Bank for International Settlements. (2019, December 15). Basel framework: CRE32. https://www.bis.org/basel_framework/chapter/CRE/32.htm

- Bao, H. (n.d.). XGBoost vs. LightGBM: A performance comparison on credit default prediction. Medium. https://medium.com/@hannie.bao_50786/xgboost-vs-lightgbm-a-performance-comparison-on-credit-default-prediction-e16037728c82
- Brownlee, J. (n.d.). Don't use random guessing as your baseline classifier. Machine Learning Mastery. Retrieved October 30, 2025, from https://machinelearningmastery.com/dont-use-random-guessing-as-your-baseline-classifier/
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.
- Corporate Finance Institute. (n.d.). Why explainable AI matters in finance. Retrieved October 30, 2025, from https://corporatefinanceinstitute.com/resources/artificial-intelligence-ai/why-explainable-ai-matters-finance/
- Data Science Process Alliance. (n.d.-a). CRISP-DM for data science teams: 5 actions to consider. Retrieved October 30, 2025, from https://www.datascience-pm.com/crisp-dm-for-data-science-teams-5-actions-to-consider/
- Data Science Process Alliance. (n.d.-b). CRISP-DM: The standard process. Retrieved October 30, 2025, from https://www.datascience-pm.com/crisp-dm-2/
- desertnaut. (2018, November 7). How to calculate accuracy score of a random classifier. Stack Overflow. https://stackoverflow.com/questions/53182709/how-to-calculate-accuracy-score-of-a-random-classifier
- Egan, C. (n.d.). Explainable AI for credit default prediction. National College of Ireland. https://norma.ncirl.ie/5146/1/ciaranegan.pdf
- European Central Bank. (n.d.). Working Paper Series No. 2954.
- Federal Reserve. (2019, March). Dodd-Frank Act supervisory stress test 2019: Methodology and results.
- How do you choose a baseline model for comparison? (2021, May 3). r/learnmachinelearning. Reddit. https://www.reddit.com/r/learnmachinelearning/comments/n40vwi/how_do_you_choose_a_baseline_model_for_comparison/
- Indium. (n.d.). Explainable AI in finance: Accountability and compliance. Retrieved October 30, 2025, from https://www.indium.tech/blog/explainable-ai-finance-accountability-compliance/
- Inside Learning Machines. (n.d.). Are decision trees robust to outliers? Retrieved October 30, 2025, from https://insidelearningmachines.com/decision_trees_robust_to_outliers/
- Loan default prediction using machine learning. (2020). International Arab Journal of Information Technology, 17(4A).

- Logistic regression in credit scoring. (2025, September 22). ArXiv.
- Moody's. (2006). Default & loss rates of structured finance securities: 1993-2006 H1.
- Naresh. (n.d.). Mastering gradient boosting: XGBoost vs. LightGBM vs. CatBoost explained simply. Dev.to. https://dev.to/naresh_007/mastering-gradient-boosting-xgboost-vs-lightgbm-vs-catboost-explained-simply-4p9c
- nCino. (n.d.). The importance of interpretable AI in the financial services industry. Retrieved October 30, 2025, from https://www.ncino.com/blog/importance-interpretable-ai-financial-services-industry
- Neptune.ai. (n.d.). XGBoost vs. LightGBM. Retrieved October 30, 2025, from https://neptune.ai/blog/xgboost-vs-lightgbm
- Office of the Comptroller of the Currency. (2009). Working Paper 2009-2.
- Pen & Pencil. (n.d.). Data imbalance: How is ADASYN different from SMOTE? Medium. https://medium.com/@penpencil.blr/data-imbalance-how-is-adasyn-different-from-smote-f4eba54867ab
- Predicting credit card default using machine learning: An empirical analysis. (2024). American Journal of Intelligent Systems, 13(1), 12–16.
- Predicting credit card defaults with machine learning. (n.d.). International Journal of Research and Analytical Reviews.
- Predictive analysis of first payment default in consumer loans. (n.d.). Turkish Journal of Electrical Engineering & Computer Sciences.
- Predictive models for credit card default: A comparison of statistical and machine learning approaches. (2025). Journal of Risk and Financial Management, 18(1), 23.
- SAS Communities. (n.d.). Outliers and tree-based models: Should we be concerned? Retrieved October 30, 2025, from https://communities.sas.com/t5/SAS-Communities-Library/Outliers-and-Tree-Based-Models-Should-We-Be-Concerned/ta-p/955438
- Soulpage IT Solutions. (n.d.). Gradient boosting machine explained. Retrieved October 30, 2025, from https://soulpageit.com/ai-glossary/gradient-boosting-machine-explained/
- SV-Europe. (n.d.). CRISP-DM methodology. Retrieved October 30, 2025, from https://www.sv-europe.com/crisp-dm-methodology/
- Syed, T. (n.d.-a). Credit default prediction. Deepnote. Retrieved October 30, 2025, from https://deepnote.com/app/thabresh-syed/Credit-Default-Prediction-9ceb4874-593d-4fce-9e65-77c99b5bde29
- Syed, T. (n.d.-b). Loan default prediction. Kaggle. Retrieved October 30, 2025, from https://www.kaggle.com/code/tasnimniger/loan-default-prediction

- Tookitaki. (n.d.). False positives. Retrieved October 30, 2025, from https://www.tookitaki.com/glossary/false-positives
- Towards AI. (n.d.). Traditional logistic regression vs. modern machine learning in credit scoring: A practical overview. https://pub.towardsai.net/traditional-logistic-regression-vs-modern-machine-learning-in-credit-scoring-a-practical-overview-ca3d2008bd57
- Tuovila, A. (2025, July 15). Loss given default (LGD): Two ways to calculate, plus an example. Investopedia. https://www.investopedia.com/terms/l/lossgivendefault.asp
- Udacity. (2025, March). CRISP-DM explained: A proven data mining methodology. Udacity. https://www.udacity.com/blog/2025/03/crisp-dm-explained-a-proven-data-mining-methodology.html
- Unit21. (n.d.). False positives. Retrieved October 30, 2025, from https://www.unit21.ai/fraud-aml-dictionary/false-positives
- Wikipedia. (n.d.). Cross-industry standard process for data mining. Retrieved October 30, 2025, from https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
- Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480.
- Zhang, L. (n.d.). Credit card default prediction based on machine learning. Semantic Scholar. https://www.semanticscholar.org/paper/Credit-Card-Default-Prediction-based-on-Machine-Zhang/18008996a51623b7cdc322ef8166964804cc265b