Table of Contents

Executive Summary

This report outlined the development of a data-driven credit-scoring model for unsecured loan applications, leveraging a dataset of 5,960 applicants with historical loan attributes. Banks are often faced with the trade-off between increasing market share through interest income and reduce the exposure of losses from defaults, therefore banks require a more precise analytics tool. The report applied predictive analytics to estimate each applicant's probability of default. The data analytics process was structured around the CRISP-DM framework, ensuring that data preparation, modelling, and performance evaluation remain relevant with the bank's business objectives.

During preprocessing, features engineering was performed to better reflect the business context and the data characteristics. A set of statistical and machine learning models was employed to classify the applicants into two classes, defaulted or not. The models were assessed under three different scenarios:

1. Goal 1: accept the maximum number of good customers if at least 85% of bad customers are correctly identified.
2. Goal 2: accept at least 70% of good customers while rejecting as many bad customers as possible.
3. No goal specified. Therefore, the focus was to correctly identify bad customers while minimising missed opportunities with good customers.

For each scenario, specific machine learning models were chosen. The selection process of the model considered both the predictive power and interpretability of the models. Some models are inherently simple to interpret, while some complex models can be explained using post-hoc methods.

Going forward, the bank could choose a strategy that aligns with its risk tolerance, macroeconomic conditions, and banks' business objectives:

- Selective growth strategy: adopt the first goal to expand the customer base and reducing missed opportunities, while ensuring the quality approvals with at least 85% of bad customers are correctly flagged.
- High risk-averse strategy: focus on the second goal to protect the bank against losses from default cases by capturing as many defaulters as possible, demonstrating conservative approach.
- Balanced growth strategy: opt for the "no goal" scenario when transparency and clarity of the result are crucial for the bank. This strategy will accept more applicants than the other two scenarios, unless strong default risk is detected.

Each model used threshold tuning to adjust the decision boundary for classifying applicants as default or non-default. This allowed for more flexibility in capturing bad customers or reducing false rejections, depending on the specific scenario and objectives.

The analysis revealed that debt-to-income ratio, and number of delinquent credit lines are two key factors in predicting the loan defaults, as they reflected both the applicants' ability and willingness to repay the loan. The absence of debt-to-income values also served as an additional warning signal. However, it is important to consider these variables alongside all other features for a more accurate and comprehensive prediction.

This analysis was based on several key assumptions, given the significant missing values and the imbalance between good and bad customers in the dataset. To enhance the analysis further, the bank could consider incorporating cost-sensitive analysis, which assign different weight to the consequences of accepting bad customers versus rejecting the good ones. Additionally, gathering more comprehensive customer data, including factors such as income would improve the accuracy of the predictions and support more informed lending decisions. These improvements could lead to better risk management and optimised the loan approval processes.

1. Introduction

Traditionally, credit risk assessments relied on heuristic methods and basic statistical model, often guided by the 5Cs framework, Character, Capacity, Capital, Conditions, and Collateral (Baiden, 2011). The goal of this project is to develop a data-driven model that could predict the likelihood of loan default based on historical customer and loan data. The project scope aligned with the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework introduced by Shearer (2000), an iterative-phases methodology, that uses continuous feedback loops to refine objectives, insights, and models aligning with specific business needs:

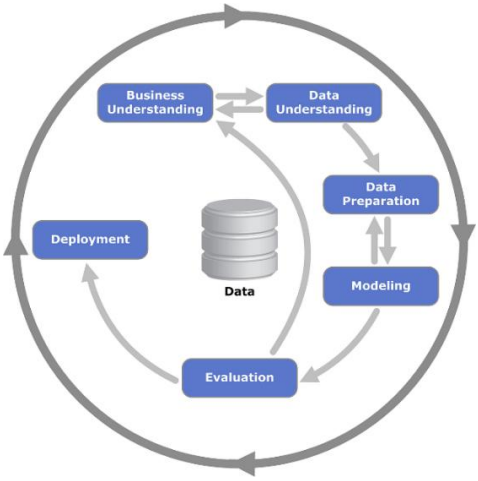


Figure 1. CRISP-DM framework

Business Understanding: In the banking industry, balancing profits from loan issuance and managing the risk of loan defaults is essential. One of the major risks faced by commercial banks is credit risk, which manifests itself as loan losses (Sinkey and Greenawalt, 1991). To minimise the losses, it is important for banks to accurately assess the default risk of the customers. The complexity of customers' behaviour could be captured by leveraging predictive models, which could lead to more informed and effective credit decision-making (Addy et al. 2024). Gaining insight into customer repayment behaviour is crucial to achieving balance between profitability and risk mitigation.

Data Understanding & **Data Preparation**: To gain a deeper understanding, the dataset was explored and visualised to uncover patterns, trends, and quality. To prepare for the modelling, data was also cleaned and transformed to enhance the data quality and predictive power. These two phases are covered in **Part 2. Exploratory Data Analysis**.

Modelling & **Evaluation**: Created multiple models to generate predictions aligned with goals and assessed the model's performance using metrics such as confusion matrix and PR AUC. These two phases are discussed in details in **Part 3. Modelling and Result Analysis**.

Deployment: The model will be applied to predict customer's behaviour, delivering actionable insights to make informed decision. This phase is explained in **Part 4. Business Recommendations, Assumptions, and Limitations**.

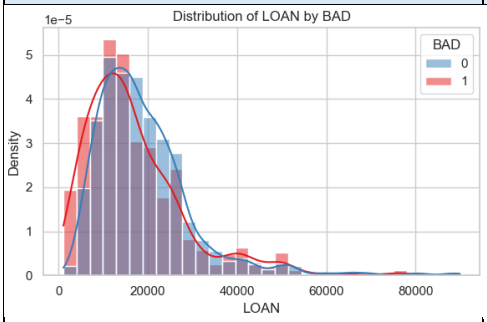
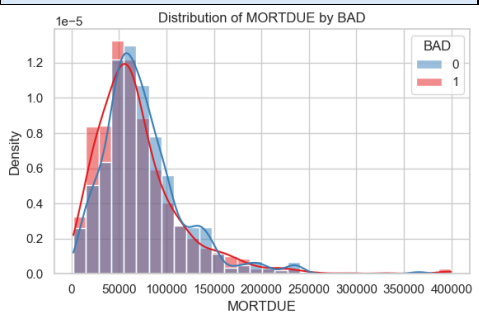
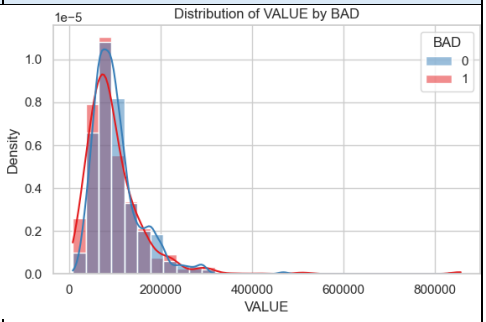
This project has practical implications in retail banking, particularly in credit risk management. According to Addy et al. (2024), the usage of analytical tools within the banking industry could lead to:

- Enhanced prediction accuracy on customers behaviour: the model offers enhanced accuracy and flexibility to handle more complex datasets and banks can predict the likelihood of loan defaults with greater accuracy based on the customers' historical data.
- Reshaping risk assessment and decision making: Data-driven models offer a deeper understanding of risk factors, facilitating more informed lending decisions compared to traditional methods.
- Effective credit risks management: Predictive analytics support more accurate creditworthiness assessments and contribute to reduced default rates.
- Operational efficiency: Automating complex activities, such as credit scoring and risk assessment, reduces processing time, optimises resource usage, minimises human error, and improves reliability.

2. Exploratory Data Analysis

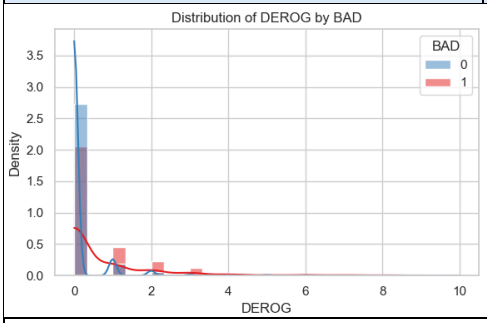
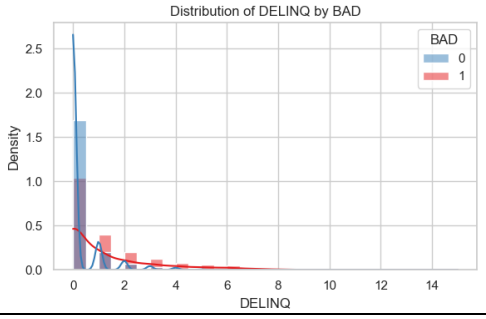
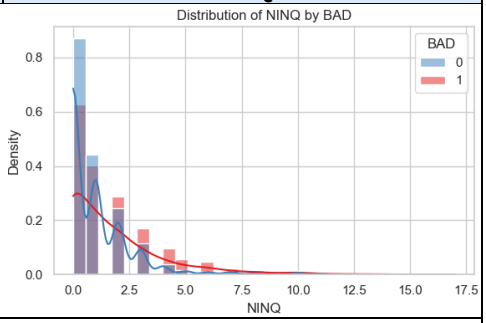
Step 1. Initial Data Assessment

A preliminary analysis of the data showed that 4,771 customers (80.05%) successfully repaid their loans (`BAD=0`), while the remaining 1,189 customers (19.95%) defaulted (`BAD=1`). This imbalance is common in real-world situation as bad customers typically account for only a small proportion of all customers (Chen et al., 2023). The below graphs depict the distributions of each variable to provide insights of good and bad customers characteristics, along with brief explanation on the variables.

LOAN	MORTDUE	VALUE
		

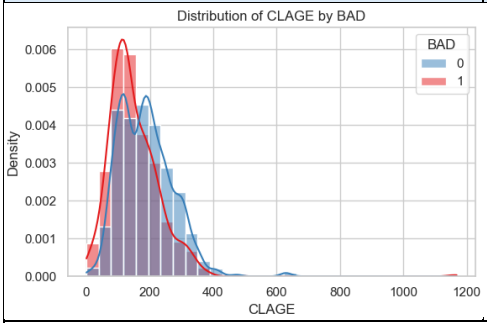
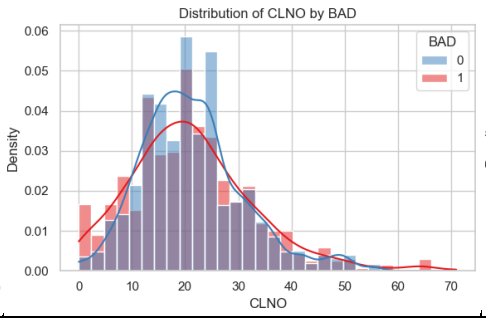
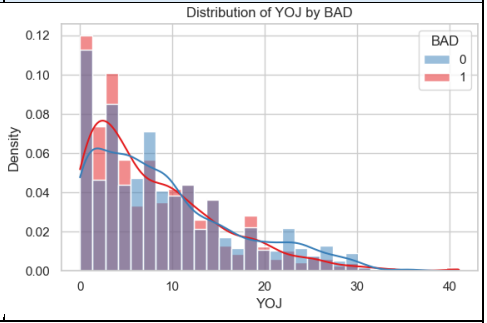
`LOAN`: The amount the applicant is requesting, risk may depend on the loan size relating to applicant's financial capacity.
`MORTDUE`: Remaining mortgage balance, reflecting the applicant's long-term financial liabilities.
`VALUE`: Estimated property value, higher value may indicate financial standing relevant to creditworthiness.

`LOAN`, `MORTDUE`, and `VALUE` distributions are right skewed. A slight shift in defaulted loans toward lower amounts suggested a higher default tendency at lower loan, mortgage, and property values.

DEROG	DELINQ	NINQ
		

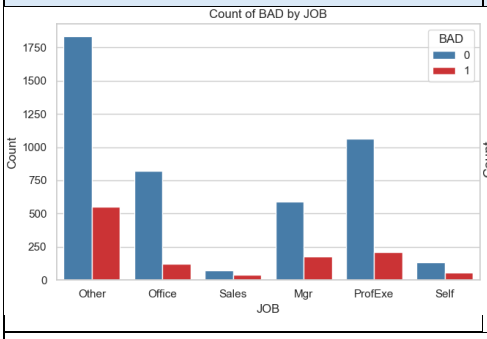
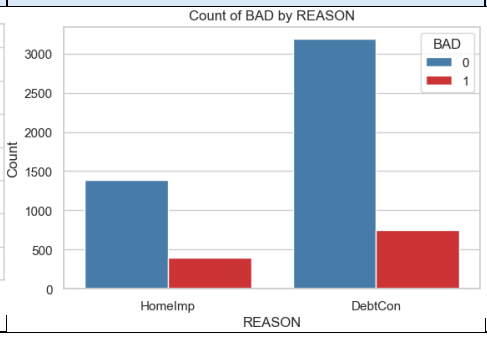
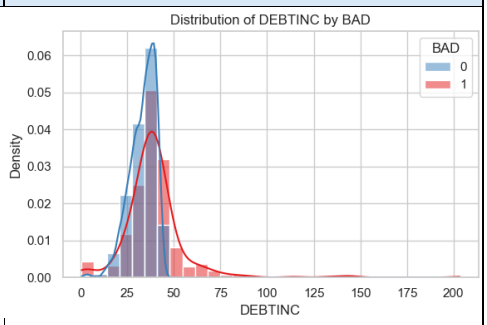
`DEROG`: Number of major derogatory records such as bankruptcy, reflecting serious issue in applicant credit history.
`DELINQ`: Number of past payments issues, reflecting repayment behaviour towards credit obligations.
`NINQ`: Number of recent credit inquiries. Higher value may indicate increased credit-seeking activity or financial distress.

`DEROG`, `DELINQ`, and `NINQ` distributions are right-skewed, with most borrowers at zero, and bad loans having higher densities at non-zero values.

CLAGE	CLNO	YOJ
		

`CLAGE`: Measures length of credit history, longer histories reflect higher financial experience.
`CLNO`: Number of credit lines. Low values may suggest limited credit use, while high values could imply overborrowing.
`YOJ`: Years at current job, longer tenure often correlates with financial stability.

`CLAGE` and `YOJ` distributions are right-skewed, with defaults common among borrowers with shorter credit histories and fewer years at their current job. `CLNO` distribution is more symmetric, but bad loans tend to have fewer credit lines or extremely high credit lines.

JOB	REASON	DEBTINC
		

JOB: Reflects occupation type, which may relate to income stability or employment risk.
REASON: Loan purpose (home improvement vs. debt consolidation), which may influence risk with slight default rate variation.
DEBTINC: Debt-to-income ratio, higher values indicate greater debt burden relative to income.

Good loans dominate across all `JOB` categories, though `"Sales"` and `"Self"` show relatively higher proportions of bad loans. For loan `REASON`, `"HomeImp"` has a slightly higher proportion of bad loans compared to `"DebtCon"`. The distribution of `DEBTINC` is right-skewed, with bad loans more concentrated at higher debt-to-income ratios.

Step 2. Missing Values Checking

As shown in Table 1, an initial data quality check was performed to assess missing values across the dataset. The results indicated that only around 3.21% of the rows had more than five missing values, while majority of the rows had missing values in five or less variables. To ensure the dataset was suitable for further analysis, 192 rows with more than five missing values were removed.

Table 1. Missing Variable Columns Counts

Num. of Missing Variable Columns	Count	Cumulative Count	Percentage	Cumulative Percentage
11	2	2	0.03	0.03
10	11	13	0.18	0.21
9	49	62	0.82	1.03
8	39	101	0.65	1.68
7	25	126	0.42	2.10
6	66	192	1.11	3.21
5	83	275	1.39	4.60
4	64	339	1.07	5.67
3	219	558	3.67	9.34
2	449	1007	7.53	16.88
1	1589	2596	26.66	43.54
0	3364	5960	56.44	100.00

Step 3. Feature Correlation Analysis and Feature Engineering

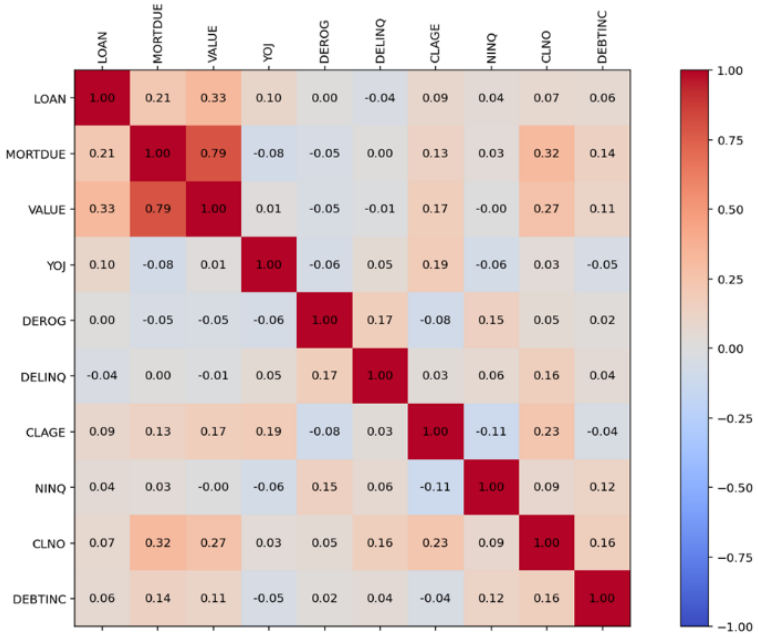


Figure 2. Correlation Matrix

Based on the correlation matrix, MORTDUE and VALUE were highly correlated, with a correlation coefficient of 0.79. Additionally, MORTDUE and LOAN are correlated because in a business term they both represent parts of a person's total debt if the loan is approved. To deal with this, a new feature was created: the Loan-to-Value (LTV) Ratio, calculated by adding LOAN and MORTDUE and dividing by VALUE. This approach mirrors practices in real-world lending where lenders commonly assess the LTV ratio to evaluate a borrower's likelihood of default (Saunders and Allen, 2010). The original LOAN, MORTDUE, and VALUE variables were then removed to streamline model and minimise redundancy.

Step 4. Analysis of Missing Value Informativeness

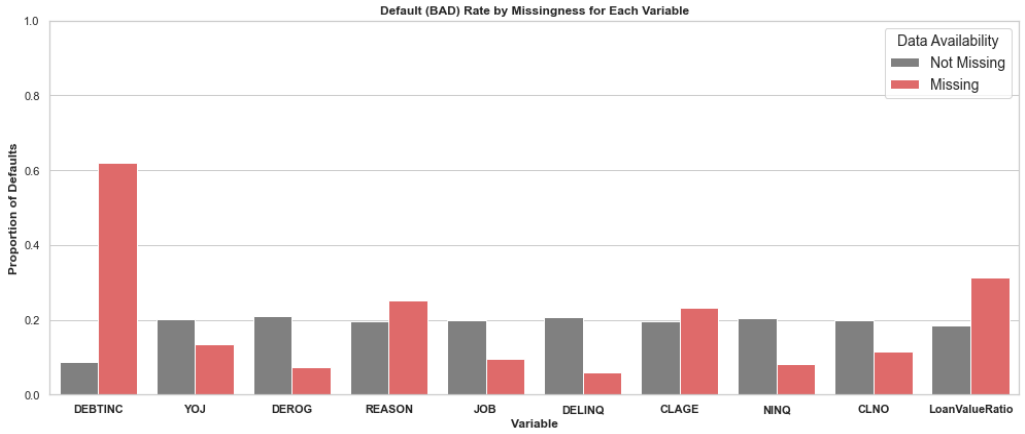


Figure 3. Default Rate by Missingness for Each Variable

Table 2. Results of Missingness Check

Features	Chi-Square Statistic	p-value	Missingness Informative?	Assumptions for Missing Variable
DEBTINC	1,700.89	$< 1 \times 10^{-15}$	**Yes**	• Applicants did not report their income, did not have regular income, or unverifiable income sources. • Could not be calculated due to missing income variable.
DEROG	54.2161	1.80×10^{-13}	**Yes**	• Applicants have no prior credit history, hence data is not available in the credit bureau. • Credit accounts are not open long enough to be reported.
DELINQ	49.4445	2.04×10^{-12}	**Yes**	
Loan Value Ratio	46.5162	9.09×10^{-12}	**Yes**	• Property values are not appraised or not reported (as this is an unsecured loan). • Applicant did not have any mortgage obligation. • Could not be calculated due to missing property value or mortgage variable.
NINQ	27.8847	1.29×10^{-7}	**Yes**	Applicants have not applied for credit recently.
YOJ	10.1527	1.44×10^{-3}	**Yes**	Applicants are unemployed, retired, students, or choose not to disclose.
JOB	9.1063	2.55×10^{-3}	**Yes**	
REASON	2.9762	0.0844	**No**	
CLNO	1.0548	0.3044	**No**	
CLAGE	0.7770	0.3780	**No**	

As shown in Figure and Table above, an analysis was performed to check whether missing values carried important information. Chi-square test was used to explore the relationship between missingness and loan outcomes (Dong and Peng, 2013). The results showed that missing values in several variables were statistically informative, with *p-values* below 0.05. Missing values in DEBTINC and LoanValueRatio were linked to much higher default rates as illustrated in the bar chart. New variables were created to flag the missing values to preserve important signals in the data that could be missed during standard missing value handling (Jakobsen et al., 2017). The new variables were applied to all columns, except for the non-informative ones, which are REASON, CLNO, and CLAGE.

Step 5. Handling Missing Values: Imputation Strategy

Different imputation methods were used to handle missing values based on the data characteristics. When missingness was linked to higher risk in the case of `LoanValueRatio` and `DEBTINC`, a large constant (9999) was used to highlight this risk in the model (Florez-Lopez, 2010). For highly skewed variables like `DEROG`, `DELINQ`, and `NINQ`, the mode was applied to reflect the most common value (Jerez et al., 2010). For other numerical variables with more even distribution such as `CLAGE`, `CLNO`, and `YOJ`, the median was used to preserve original skewness (Jerez et al., 2010). Categorical variables were imputed with logical replacements: `REASON` was filled with "`Not Provided`" and `JOB` with "`Other`," the mode. This tailored approach ensured data consistency while retaining important variance and avoiding bias in model predictions, and it also complements the missingness assumptions identified in Table 2.

Step 6. Outlier Identification and Handling

For preprocessing, it is essential to address both scale differences across features and the presence of outliers. The distributions of variables in Step 1 show the presence of outliers, which might represent valid but extreme financial behaviour. The distributions of variables in Step 1 show the presence of outliers, which might represent valid but extreme financial behaviour. While there are several methods to standardise data, such as min-max or standard scaling, Robust Scaling was opted. This method scales features using the median and interquartile range (IQR), making it less sensitive to the extreme values imputed (Röchner et al., 2024). Importantly, log transformation was not applied, as it alters the interpretability of the variables and reduces natural skewness that might hold predictive value in credit risk models. Robust scaling was particularly useful for skewed features like `DEBTINC` and `LoanValueRatio`, with the following formula:

$$X_{scaled} = \frac{X - X_{median}}{X_{IQR}}$$

X = features/variables

Step 7. Data Splitting Strategy

A stratified 75:25 train-test split was applied to maintain the proportion of the target variable (`BAD`) across both sets. This approach ensures balanced representation of classes, which is essential for reliable model evaluation in imbalanced datasets (Kuhn and Johnson, 2013).

Variable Importance and Ranking

To assess variable importances, L1-Regularised Logistic (Lasso) was applied for feature importances. Lasso penalises the absolute values of coefficients, shrinking irrelevant ones to zero (Muthukrishnan and Rohini, 2016). This makes it particularly effective for identifying key predictors while avoiding overfitting. Lasso importance scores were paired with Mutual Information (MI) to capture both linear and non-linear relationships (Verleysen et al., 2009). The combined Lasso and MI ranks provided a stable measure of feature relevance, to improve the accuracy of credit risk assessments.

Table 3. Variable Importance and Ranking

Variables	Mutual Information		Lasso		Average Rank
	MI Importance	MI Rank	L1 Logistic Importance	L1 Logistic Rank	
JOB_Office	0.0558	2	0.6988	2	2
JOB_Mgr	0.0317	4	0.4745	4	4
JOB_Other	0.0218	6	0.1338	7	6.5
JOB_ProfExe	0.0372	3	0.0061	10	6.5
REASON_HomeImp	0.1661	1	0.0003	12	6.5
NINQ	0.0068	11	0.5755	3	7
LoanValueRatio	0.0036	14	0.9276	1	7.5
JOB_Self	0.0198	7	0.0211	9	8
REASON_Not Provided	0.0224	5	5.12×10^{-5}	13	9
JOB_Sales	0.0163	8	0.0055	11	9.5
YOJ	0.0049	13	0.2150	6	9.5
REASON_DebtCon	7.02×10^{-5}	17	0.3584	5	11
DEBTINC	0.0014	16	0.1087	8	12
DELINQ	0.0111	9	0.0000	15.5	12.25
DEROG	0.0086	10	0.0000	15.5	12.75
CLAGE	0.0054	12	0.0000	15.5	13.75
CLNO	0.0036	15	0.0000	15.5	15.25

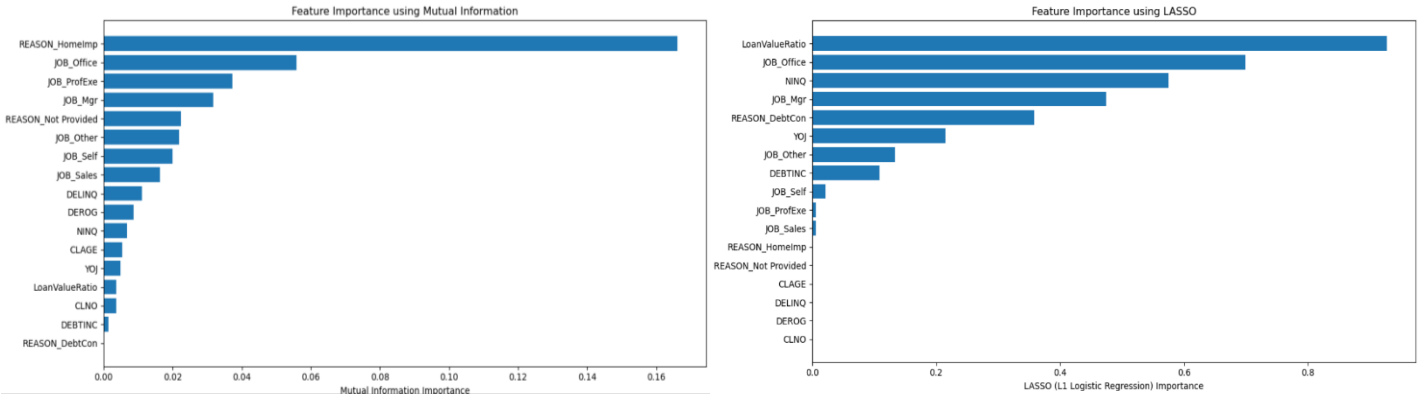


Figure 4. Feature Importance using Lasso and Mutual Information

`LoanValueRatio` and `JOB_Office` emerged as top predictors in Lasso with coefficients of 0.9276 and 0.6988 respectively, making them critical variables for default risk modelling. While `REASON_HomeImp` ranked first in MI (0.167) for its strong non-linear relationship, it ranked 12th in Lasso (0.0003), highlighting its limited linear predictiveness. All variables had non-zero importance in at least one method and were therefore retained and included in the modelling process.

3. Modelling and Results Analysis

3.1. Performance metrics

To analyse the performance of the models on goals, confusion matrix was used to identify how well the model performed and the percentage of the errors predicted.

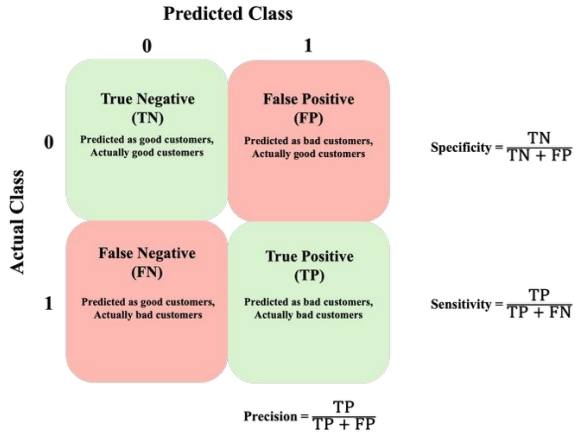


Figure 5. Confusion Matrix

Table 4. Model Goal Objectives and Metrics Focus

Goal	Objective	Metric Focus	Explanation
1	Accept maximum number of good customers if at least 85% of bad customers are correctly identified	Sensitivity (`BAD = 1`) ≥ 85% Maximise Specificity (`BAD = 0`)	Sensitivity measures the model's ability to correctly identify defaulters. Using it as a parameter helps minimises false negatives by focusing on accurately detecting bad customers. Specificity measures how well the model identifies actual good customers. Using it as a parameter helps minimises false positives by focusing on correctly classifying those who do not default.
2	Accept at least 70% of good customers while rejecting as many bad customers as possible	Specificity (`BAD = 0`) ≥ 70%) Maximise Sensitivity (`BAD = 1`)	

Goal	Objective	Metric Focus	Explanation
3	No specific goal	Maximise F1 Score, where $\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$	The F1 score balances precision and sensitivity, minimising both type of errors, false positives (rejecting good customers) & false negatives (accepting bad customers). It is especially useful for imbalanced datasets where default (`BAD = 1`) cases are rare.

The goals outlined in the table above highlight different objectives and focuses that will be applicated in the model. In banking, using a model that could accurately determine customers' probability to default could resulted in a decrease in their losses (Dash et al. 2021) and accurately identifying good customers allows banks to offer favourable terms, such as lower interest rates and easier loan approvals, which could lead to attracting and retaining good quality customers (Capital One, 2023).

3.2. Modelling Process

After splitting the dataset, all preprocessing steps were integrated into a pipeline to prevent data leakage during modelling validation and test. The modelling process was done using stratified cross-validation with 10-folds to get a stable evaluation while still maintaining the bad and good customers ratio in each fold (Prusty, Patnaik and Dash, 2022) combined with hyperparameter tuning to adjust each model setting to improve their performance (Agrawal, 2021). Initially, various data preprocessing techniques including variable reduction, binarisation, and clustering were explored, however these did not yield improved results, therefore the current preprocessing approach was proceeded as it also better aligned with the business context.

This is a supervised learning problem, as the target variable (default status) is known for all observation (Murphy, 2012). Accordingly, several classification models (Table 5) were tested and compared according to the project objectives.

Table 5. Machine Learning Models Description

Machine Learning Model	Model Description
Histogram-Based Gradient Boosting (HGBT) Classifier	As part of the ensemble boosting method, it sequentially fits trees, where each new tree focuses on correcting errors made by previous ones. Misclassified records are given more weight, improving accuracy, especially for rare classes (Shmueli, 2017).
Random Forest (RF) Classifier	Belonging to the ensemble bagging family, RF generates many decision trees with a random element, evaluates their performance, and then selects the best trees to form an ensemble (Bramer, 2013).
Support Vector Classifier (SVC)	A classifier that finds the optimal hyperplane separating classes with the maximum margin (Runkler, 2016).
Logistic Regression	Predicts the probability of a record belonging to a class and classifies it based on a cutoff value (Shmueli, 2017).
k-Nearest Neighbours (kNN) Classifier	Classifies a point based on the majority class among its k-nearest neighbors using a distance metric (Runkler, 2016).
Neural Network (NN)/ Multi-layer Perceptron (MLP) Classifier	A neural network that maps inputs to outputs through layers of neurons, each applying a weighted sum, bias, and sigmoid activation (Runkler, 2016).
Decision Tree (DT) Classifier	Splits data by attribute values until each branch leads to a single class label (Bramer, 2013).

Threshold tuning was also used during the CV process to meet the project objectives. Thresholds were used as a cut-off to the model probability prediction when classifying a customer as bad or good (Leevy et al., 2023), which helps the model be more strict or lenient. Lower thresholds could help capturing more bad customers with the risk of rejecting good customers and higher thresholds reduced false rejection but risk accepting some bad customers.

Each modelling result generated an area under the precision-recall (PR-AUC) curve to see the overall model performance. It focuses on precision (how many flagged bad customers are truly bad) and recalls (how many actual bad customers are correctly flagged). PR AUC was selected over ROC as PR curves are more informative for imbalanced dataset, providing better insight into potential signs of underfitting or overfitting on the test set (Saito and Rehmsmeier, 2015).

3.3. Results

The following table shows the top-performing model from each classifier selected for each goal. All classifiers achieved sensitivity level required in Goal 1 in both CV and test sets. However, for Goal 2, the Logistic and DT failed to meet the minimum specificity of 70% in the test set. Full results are available in Appendix A.

Table 6. Performance Metrics for Each Goal Across Classification Models

Goal	Metric	HGBT		RF		SVC		Logistic		kNN		NN		DT	
		CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
Goal 1	**Sensitivity**	**0.867**	**0.853**	**0.862**	**0.853**	**0.867**	**0.870**	**0.861**	**0.870**	**0.864**	**0.874**	**0.871**	**0.863**	**0.863**	**0.881**
	Specificity	**0.928**	**0.928**	**0.914**	**0.908**	**0.828**	**0.822**	**0.699**	**0.717**	**0.789**	**0.777**	**0.628**	**0.674**	**0.750**	**0.742**
	PR-AUC	0.894	0.898	0.830	0.846	0.653	0.644	0.747	0.754	0.680	0.707	0.542	0.586	0.707	0.707
	Threshold	0.147	0.147	0.253	0.253	0.157	0.157	0.313	0.313	0.087	0.087	0.090	0.090	0.293	0.293
Goal 2	**Sensitivity**	**0.964**	**0.975**	**0.966**	**0.990**	**0.926**	**0.919**	**0.858**	**0.877**	**0.888**	**0.909**	**0.820**	**0.835**	**0.875**	**0.891**
	Specificity	**0.732**	**0.728**	**0.741**	**0.726**	**0.706**	**0.701**	**0.706**	**0.695**	**0.731**	**0.700**	**0.734**	**0.709**	**0.718**	**0.666**
	PR-AUC	0.880	0.881	0.830	0.846	0.614	0.625	0.747	0.754	0.680	0.707	0.550	0.537	0.707	0.707
	Threshold	0.010	0.010	0.061	0.061	0.195	0.195	0.300	0.300	0.062	0.062	0.105	0.105	0.181	0.181
No goal	**F1 score**	**0.826**	**0.810**	**0.791**	**0.767**	**0.709**	**0.676**	**0.693**	**0.700**	**0.713**	**0.675**	**0.658**	**0.660**	**0.720**	**0.703**
	PR-AUC	0.893	0.902	0.833	0.842	0.653	0.644	0.747	0.754	0.643	0.690	0.542	0.586	0.712	0.725
	Threshold	0.267	0.267	0.251	0.251	0.227	0.227	0.691	0.691	0.240	0.240	0.323	0.323	0.679	0.679

Legends: **Goal achieved** **Goal not achieved**

The curves below illustrate the combined Precision-Recall (PR) Area Under the Curve (AUC) from all classifiers. In most scenarios, the SVC and NN curves showed unusual movement (i.e., sudden drops and spikes), indicating possible overfitting or underfitting. The curves from HGBT and RF results were more stable and dominated the other models.

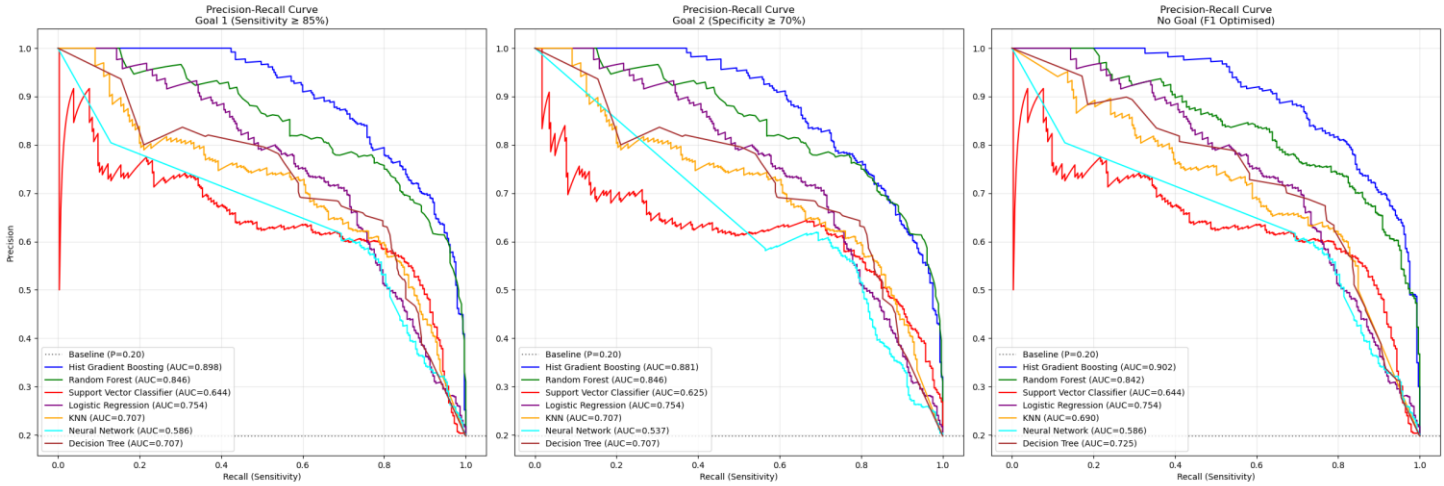


Figure 6. PR-AUC of the Test Results

3.4. Choosing the models

To select one most appropriate model for each case, the following criteria were applied:

- Chose models that meet the specified goals (for Goal 1 and Goal 2) in both the CV and test sets to ensure the model's generalisation on unseen data.
- Evaluated the PR AUC index and the curve to exclude models with a PR AUC index below 0.7, especially those showing signs of underfitting or overfitting (Çorbacıoğlu and Aksel, 2023).
- Ensured that the absolute differences between CV and test results for the relevant metrics in each case were less than 3% to further validate the robustness of the model (Murphy, 2012).

Table 7. Criteria Applied to Choose Models

Steps	Goal 1	Goal 2	No goal
1. Meet the specified goals in both CV and test sets	No model excluded	**Logistic and DT excluded** due to specificity on test set being <70%	No model excluded
2. Evaluation on PR AUC index and the curve	**SVC and NN excluded** due to consistently low PR AUC (< 0.7) in all scenarios and signs of underfitting/overfitting in the curve		
3. Absolute difference < 3% for relevant metrics	No further exclusion	**kNN excluded** due to difference in specificity (3.1%)	**kNN excluded** due to difference in F1 score (3.8%)
Remaining eligible options:	HGBT, RF, Logistic, kNN, DT	HGBT, RF	HGBT, RF, Logistic, DT

With the remaining eligible options for each scenario, the next step was to evaluate the trade-off between the predictive power and interpretability of each classifier (Arrieta et al., 2020). The following chart illustrates these trade-offs, with Logistic being the most interpretable classifier but having relatively low predictive power. In contrast, NN usually offers high predictive power, though it is more difficult to interpret.

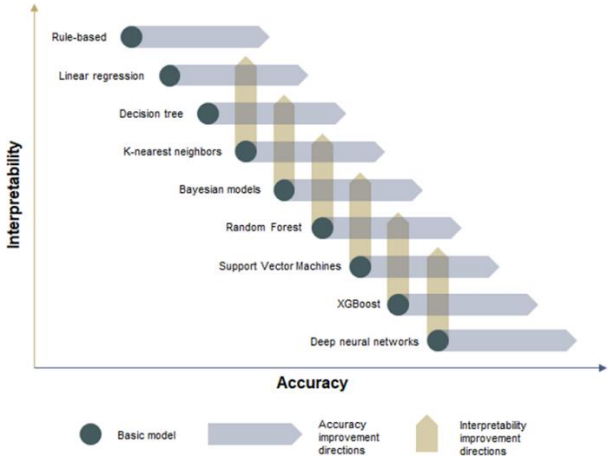


Figure 7. Accuracy vs Interpretability of Machine Learning Models

While boosting and ensemble classifiers are more complex and harder to interpret, they demonstrated strong performance in terms of the predictive power. In situations where achieving specific performance metrics is critical (in the case of Goal 1 and Goal 2), the predictive power is given more emphasis over the full interpretability, with post-hoc methods still available to aid in the interpretation (Nesvijevskaia et al., 2021).

- Goal 1: Maximising **specificity** if at least **85% sensitivity** is achieved. **HGBT** was chosen due to its highest specificity results.
- Goal 2: Meet at least **70% specificity** while maximising **sensitivity**. **RF** was chosen due to its highest sensitivity results.

As there were no specified goals in "No goal" scenario, it is recommended to choose the models that offers a balance predictive power and interpretability. For this reason, DT was selected. With good combination of F1 score, PR AUC index, and the interpretability of the tree, DT allows effective prediction of debt defaults while remaining relevant for addressing general business concerns. Logistic was not chosen as it assumes linear relationship between features and the target, which may oversimplify and tends to underestimate the likelihood of rare events (King and Zeng, 2001).

Table 8. Detailed Result of the Chosen Models

Results	**Goal 1**			**Goal 2**			**No goal**		
Chosen classifier	**HGBT**			**RF**			**DT**		
	CV	**Test**	**|Diff.|**	**CV**	**Test**	**|Diff.|**	**CV**	**Test**	**|Diff.|**
Sensitivity	**0.867**	**0.853**	**0.014**	**0.966**	**0.990**	**0.023**	0.754	0.790	0.035
Specificity	**0.928**	**0.928**	**0.000**	**0.741**	**0.726**	**0.015**	0.916	0.888	0.028
Precision	0.754	0.745	0.008	0.479	0.471	0.008	0.692	0.634	0.058
F1 score	0.805	0.795	0.009	0.640	0.638	0.002	**0.720**	**0.703**	**0.017**
PR AUC	0.894	0.898	0.004	0.830	0.846	0.016	0.712	0.725	0.013
Thresholds	0.147			0.061			0.679		
Parameters	'l2_regularization': 0.0, 'learning_rate': 0.1, 'max_bins': 128, 'max_depth': 50, 'max_iter': 500, 'min_samples_leaf': 1, 'class_weight' = 'balanced'			'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200, 'class_weight' = 'balanced'			'criterion': 'entropy', 'max_depth': 8, 'min_impurity_decrease': 0.001, 'min_samples_leaf': 5, 'min_samples_split': 10, 'class_weight' = 'balanced'		

The small variations between CV and test results for the chosen models confirmed the models' generalisation and robustness, with no significant overfitting present. The use of all variables, with attention to their importance, resulted in optimised prediction performance.

3.5. Interpreting the chosen models

Although the full trees of HGBT and RF are harder to visualise as they consist of multiple decision trees, post-hoc method such as SHAP can be employed. SHAP is a game-theoretic attribution method introduced by Lundberg and Lee (2017) that computes Shapley values for each feature in each sample. By averaging these local attributions across the dataset, SHAP captures robust global feature importances while retaining the sign of each contribution. For "No goal" scenario which utilised DT model, the result can be interpreted either directly from the decision tree or SHAP.

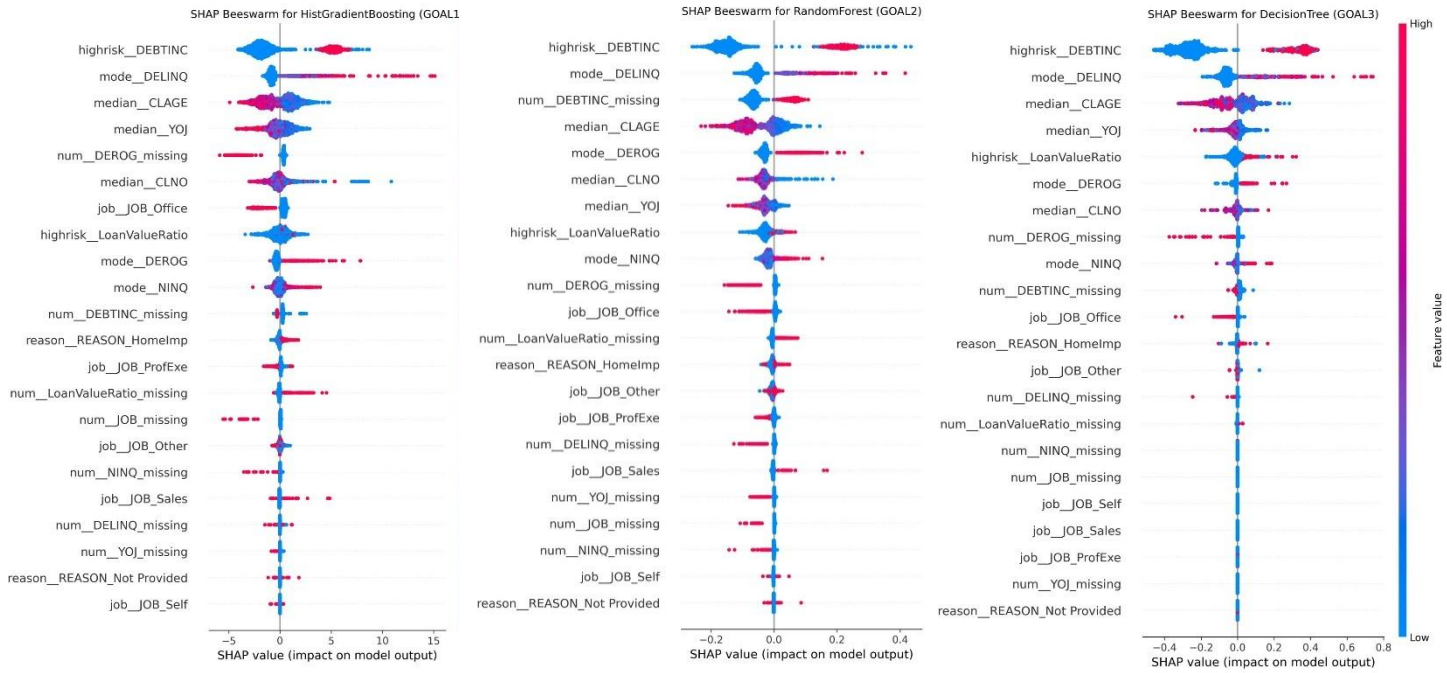


Figure 8. Beeswarm Plots of the Chosen Models

The beeswarm plots above which represented the average SHAP value, revealed concordance among models with respect to both feature importance rankings and the directional effects of key predictors, `DEBTINC` and `DELINQ`. `DEBTINC` and `DELINQ` showed dominance in driving model predictions with positive relationship, in which as those features increase, the higher probabilities of default predicted by the model. However, for some cases, low `DEBTINC` were accompanied by positive SHAP contribution, indicating that `DEBTINC`'s predictive power was maximised only when evaluated alongside other attributes.

In Goal 2, RF treated missing `DEBTINC` as red flag and customers without a recorded `DEBTINC` were assumed to carry higher default risk. Other findings were seen in `CLAGE` and `YOJ`, in which shorter credit histories and longer duration of current employment respectively corresponded to negative SHAP values (i.e., tenure stability and solid track record reduce default likelihood). The full importances from SHAP are shown in Appendix B.

Below is the tree excerpt from DT result for "No goal" scenario (detailed DT rule in Appendix D). The tree's root split occurred at a `DEBTINC` of 43.76%, such that any borrower exceeding this ratio is routed to the "default" leaf. The subsequent branches revealed that even when `DEBTINC` falls below this critical value, `DELINQ` above 4.5 (i.e., a proxy for more than four past missed payments) would likewise direct the observation into the default terminal node (in red box).

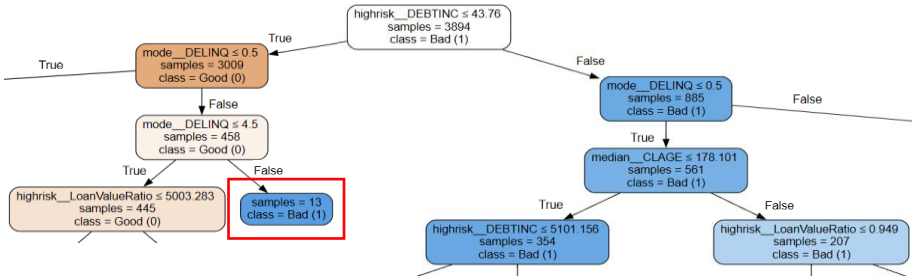


Figure 9. Excerpt of Decision Tree for "No goal" Scenario

4. Business Recommendations, Assumptions and Limitations

The bank operates in a competitive and regulated credit market, where it must balance the need to grow its loan book with maintaining asset quality. According to UK industry benchmarks, banks such as Lloyds and NatWest aim to keep portfolio-level default rates between 1% and 3%, depending on the asset class (Lloyds Annual Report, 2024; EY UK Banking Outlook 2025; UK Finance, 2025). Although approximately 20% of past applicants in the provided dataset defaulted, this may reflect sample limitations rather than true portfolio risk. Nevertheless, it reinforces the need for a robust credit risk model to help the bank manage default rates effectively and stay aligned with industry standards.

To accommodate varying risk appetites, evolving macroeconomic conditions, and alignment with banking industry best practices, multiple lending strategies are recommended, in accordance with model results in Table 8 and confusion matrix in Appendix C:

1. **Selective Growth Strategy (Expand customer base with quality approvals)**
   To responsibly expand the customer base in more competitive and retail-focused markets, such as unsecured personal loans and credit cards, the HGBT model (Goal 1) is recommended. This model successfully captured 85-87% of defaulters while maintaining a high approval rate of over 77% across validation and test sets (4,453 total approvals). With a relatively low decision threshold (~15%), the model is more cautious, flagging applicants as high risk at modest predicted probabilities. This cautious approach helps balance growth and risk by approving more applicants but still maintaining strong approval quality, evidenced by a high specificity of 93% (i.e., most accepted customers are genuinely low risk). This selective growth approach enables the bank to tap into new and underserved segments while keeping default risk within industry benchmarks, aligning well with ambitions for measured expansion without sacrificing portfolio quality.

2. **Highly Risk-Averse Strategy (Protect loan book quality)**
   To prioritise risk control and safeguard asset quality, especially for core lending segments such as mortgages and secured loans, the RF model (Goal 2) is recommended. This model captures 96-99% of defaulters, reflecting its strong focus on minimising risks. It operates at a very low decision threshold (~6%), meaning applicants are flagged as high risk even at low predicted probabilities of default. This results in the rejection of 2,421 applications (42%), demonstrating its conservative stance. While the specificity is lower at 72-74%, this trade-off is expected in a highly risk-averse strategy that favours rejecting borderline cases to ensure defaulters are not mistakenly approved. This approach aligns with regulatory expectations to maintain default rates below 3% (PRA, 2021), supporting long-term stability over rapid growth.

3. **Balanced Growth Strategy (Balance lending expansion and risk control)**
   To support portfolio growth while ensuring a reasonable level of risk control and high interpretability, the DT model (No goal scenario) is recommended. Its straightforward and transparent decision paths make it particularly suitable for contexts prioritising compliance with regulations such as fair lending and transparency under the FCA (2021) and the ‘right to be informed’ in the UK GDPR (2021).

   While its predictive performance is more modest compared to the HGBT and RF model, it offers a balanced trade-off between acceptance rate and risk control. The model captures 75-79% of defaulters and approves approximately 78% of applicants across validation and test sets (4,476 total approvals). Operating at a higher decision threshold (~68%), it adopts a more acceptance-tolerant approach, approving most applicants unless a strong default risk is indicated. Despite being less selective than first two strategies, it still achieves a strong specificity around 88%, ensuring that most good applicants are approved.

Key Risk Indicators for Immediate Action:

Based on the model findings and validation against industry standards, two customer attributes clearly drive default risk and should become top screening priorities:

- The debt-to-income (`DEBTINC`) ratio is a key predictor of a customer's ability to repay. The result from DT model identified a critical split point at 43.76%, meaning borrowers above this value are much more likely to default. This aligns with general benchmarks, where `DEBTINC` ratios above 40%-45% are viewed as higher risk (Bank of England, 2023; Fannie Mae, 2025). Limiting approvals or applying stricter conditions for applicants with `DEBTINC` exceeding ~44% is recommended, unless strong compensating factors are present (e.g., high income, excellent credit history, long-term employment stability).
- Number of past delinquencies (`DELINQ`) strongly reflects borrower reliability. The analysis shows that applicants with more than four or five delinquencies have a significantly higher likelihood of default. Academic research confirms that customers who become delinquent on one debt are 33%-56% more likely to default again in three years (Braga et al., 2019). Therefore, any applicants with four or more past delinquencies should undergo enhanced affordability checks or be priced accordingly to reflect their higher risk.

Limitations and Assumptions:

- The models assumed that past applicant behaviour remains a good predictor of future repayment patterns, despite potential macroeconomic changes (e.g., interest rate hikes, inflation). Therefore, stable economic condition was assumed, and the models were trained solely on historical behaviour.
- The sample dataset has around 80% non-defaulters and 20% defaulters, making models biased toward predicting non-default. Stratified train-test split and cross validation in all models were utilised to ensure defaulters received sufficient attention during training.
- A substantial portion of features exhibit missing values, with ~44% of applicants having at least one missing field. The statistical test revealed that missingness correlates significantly with `BAD` outcomes for most of the features, which were assumed to be informative (i.e., data not recorded or unavailable in credit bureau system). To preserve predictive signals, missing indicators were added instead of dropping rows or using blind imputation. Tailored imputation methods were performed, to match the characteristics and distribution of each variable.
- Complex models such as HGBT and RF offer better predictive performance but lower explainability compared to simpler models like DT. It is acknowledged that interpretation method for complex models such as SHAP values is available, which can still provide regulatory and operational transparency.

Scopes for Further Improvement:

- Future work could incorporate cost-sensitive learning to weigh false positives (rejecting good customers), and false negatives (accepting bad customers) differently based on business impact.
- Improve data collection processes by incorporating standardised credit scores (e.g., FICO, Experian), detailed employment information (income, permanent vs contract, industry sector), and demographic information (e.g., marital status, number of dependents, education level). These additional features are widely used in industry-standard credit scoring models (Thomas, 2000) and are encouraged under fair lending regulations (FCA, 2021). Better data quality would enhance risk prediction accuracy, customer segmentation, and regulatory compliance.
- Implement regular model monitoring, audit, and refresh cycles to maintain accuracy and fairness as customer profiles and economic conditions evolve.

References

1. Addy, W., Ugochukwu, C., Oyewole, A., Adeoye, O. and Okoye, C. (2024) 'Predictive analytics in credit risk management for banks: A comprehensive review', *GSC Advanced Research and Reviews*, 18(2), pp.434–449. doi:https://doi.org/10.30574/gscarr.2024.18.2.0077.
2. Agrawal, T. (2021) *Hyperparameter optimization in machine learning : Make your machine learning and deep learning models more efficient.* Berkeley: Apress.
3. Arrieta, A. B. et al. (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information fusion*, 58, pp.82–115.
4. Baiden, J. E. (2011) *The 5 C's of credit in the lending industry*. Available at: https://ssrn.com/abstract=1872804 or http://dx.doi.org/10.2139/ssrn.1872804
5. Bank of England (2023) *How are higher interest rates affecting new mortgage lending?* Available at: https://www.bankofengland.co.uk/bank-overground/2023/how-are-higher-interest-rates-affecting-new-mortgage-lending.
6. Braga, B., McKernan, S.-M. and Hassani, H. (2019) 'Delinquent debt decisions and their consequences over time', *Urban Institute*, Available at: https://www.urban.org/sites/default/files/publication/100005/delinquent_debt_decisions_and_their_consequences_over_time_0.pdf.
7. Bramer, M. (2013) *Principles of data mining*. 2nd edn. London: Springer London.
8. Capital One (2023) *Credit risk: How creditors are evaluating you*. Available at: https://www.capitalone.com/learn-grow/money-management/credit-risk.
9. Chen, L., Weng, J., Zhou, J., & Zhang, M. (2023) 'Interpretable machine learning for credit ccoring: A survey', *European Journal of Operational Research*.
10. Çorbacıoğlu, Ş.K. and Aksel, G. (2023). 'Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value', *Turkish Journal of Emergency Medicine*, 23(4), pp.195–198.
11. Dash, R., Kremer, A. and Petrov, A. (2021) 'Designing next-generation credit-decisioning models', *McKinsey & Company*. Available at: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/designing-next-generation-credit-decisioning-models.
12. Dong, Y., and Peng, C. -Y. J. (2013) 'Principled missing data methods for researchers', *SpringerPlus*, 2(1), 222. doi:https://doi.org/10.1186/2193-1801-2-222
13. Experian UK (2023) *M-Index: A new measure of consumer credit resilience*. Available at: https://www.experian.co.uk/blogs/latest-thinking/automated-credit-decisions/m-index.
14. EY (2025) *Growth in UK mortgage lending to double this year*. Available at: https://www.ey.com/en_uk/newsroom/2025/02/growth-in-uk-mortgage-lending-to-double-over-2025.
15. Fannie Mae (2024) *Debt-to-Income ratios. Selling guide*, Section B3-6-02. Available at: https://selling-guide.fanniemae.com/sel/b3-6-02/debt-income-ratios.
16. FCA (2021) *Guidance on fair treatment of customers*. Available at: https://www.fca.org.uk/publications/finalised-guidance/guidance-firms-fair-treatment-vulnerable-customers.
17. Florez-Lopez, R. (2010) 'Effects of missing data in credit risk scoring: A comparative analysis of methods to achieve robustness in the absence of sufficient data', *Journal of the Operational Research Society*, 61(3), 486–501.
18. Ilyasov, V. (2025) 'Identifying high-risk borrowers: Top strategies for digital lenders, identifying high-risk borrowers: Top strategies for digital lenders', RiskSeal, 4 February. Available at: https://riskseal.io/blog/how-to-spot-high-risk-borrowers-before-they-default.
19. Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017) 'When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts', *BMC Medical Research Methodology*, 17(1), 162.
20. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010) 'Missing data imputation using statistical and machine learning methods in a real breast cancer problem', *Artificial Intelligence in Medicine*, 50(2), 105–115.
21. King, G. and Zeng, L. (2001) 'Logistic regression in rare events data', *Political Analysis*, 9(2).
22. Kuhn, M., & Johnson, K. (2013) *Applied predictive modelling*. New York: Springer.

23. Leevy, J.L., Johnson, J.M., Hancock, J. and Khoshgoftaar, T.M. (2023) 'Threshold optimization and random undersampling for imbalanced credit card data', *Journal of Big Data*, 10(1). doi:https://doi.org/10.1186/s40537-023-00738-z.
24. Li, H. and Wu, W. (2024) 'Loan default predictability with explainable machine learning', *Finance Research Letters*, 60, p. 104867. doi:10.1016/j.frl.2023.104867.
25. Lloyds Banking Group (2024) *Annual Report and Accounts 2024*. Available at: https://www.lloydsbankinggroup.com/investors/annual-report.html.
26. Lundberg, S. and Lee, S.-I. (2017). 'A unified approach to interpreting model predictions'.
27. Murphy, K.P. (2012) *Machine learning: A probabilistic perspective*. Cambridge: The MIT Press.
28. Muthukrishnan, R. and Rohini, R. (2016) 'Lasso: A feature selection technique in predictive modeling for machine learning', *IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore, India, 2016, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.
29. Nesvijevskaia, A. *et al.* (2021) The accuracy versus interpretability trade-off in fraud detection model, *Data & Policy*, 3, p. e12. doi:10.1017/dap.2021.3.
30. Philipp Röchner, Marques, H.O., Ricardo, Zimek, A. and Franz Rothlauf (2024). Robust Statistical Scaling of Outlier Scores: Improving the Quality of Outlier Probabilities for Outliers. *Lecture notes in computer science*, pp.215–222. doi:https://doi.org/10.1007/978-3-031-75823-2_18.
31. Prusty, S., Patnaik, S. and Dash, S.K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4. doi:https://doi.org/10.3389/fnano.2022.972421.
32. Röchner, P., Marques, H. O., Campello, R. J. G. B., Zimek, A., & Rothlauf, F. (2024). *Robust statistical scaling of outlier scores: Improving the quality of outlier probabilities for outliers*. arXiv preprint arXiv:2408.15874.
33. Runkler, T.A. (2016) *Data analytics : Models and algorithms for intelligent data analysis*. 2nd edition. Wiesbaden: Springer.
34. Saito, T. and Rehmsmeier, M. (2015) 'The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PloS one*, 10(3), pp.e0118432–e0118432.
35. Saunders, A., & Allen, L. (2010) *Credit risk management in and out of the financial crisis: New approaches to value at risk and other paradigms.* 3rd edn. Wiley.
36. Shearer, C. (2000) 'The CRISP-DM model : The new blueprint for data mining', *Journal of Data Warehousing*, 5(4), pp. 13–22.
37. Shmueli, Galit. (2017) *Data mining for business intelligence concepts, techniques, and applications in R*. Newark: John Wiley & Sons, Incorporated.
38. Sinkey, J.F. and Greenawalt, M.B. (1991) 'Loan-loss experience and risk-taking behavior at large commercial banks', *Journal of Financial Services Research*, 5(1), pp. 43–59. doi:10.1007/bf00127083.
39. Thomas, L.C. (2000) 'A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers', *International Journal of Forecasting*.
40. UK Finance (2025) *Household finance review and consumer credit statistics*. Available at: https://www.ukfinance.org.uk/data-analysis/data.
41. Verleysen, M. et al. (2009). *Advances in feature selection with mutual Information in similarity-based clustering.* Germany: Springer, pp. 52–69.

Appendix A: Full Model Results

The following table shows the full results of the modelling trials using seven different classifiers:

Goal	Metric	HGBT		RF		SVC		Logistic		kNN		NN/MLP		DT	
		CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
Goal 1	Sensitivity	**0.867**	**0.853**	**0.862**	**0.853**	**0.867**	**0.870**	**0.861**	**0.870**	**0.864**	**0.874**	**0.871**	**0.863**	**0.863**	**0.881**
	Specificity	**0.928**	**0.928**	**0.914**	**0.908**	**0.828**	**0.822**	**0.699**	**0.717**	**0.789**	**0.777**	**0.628**	**0.674**	**0.750**	**0.742**
	Precision	0.754	0.745	0.714	0.696	0.559	0.546	0.429	0.431	0.507	0.491	0.369	0.395	0.466	0.456
	F1 score	0.805	0.795	0.781	0.767	0.678	0.671	0.567	0.576	0.638	0.629	0.517	0.542	0.603	0.601
	PR-AUC	0.894	0.898	0.830	0.846	0.653	0.644	0.747	0.754	0.680	0.707	0.542	0.586	0.707	0.707
	Threshold	0.147	0.147	0.253	0.253	0.157	0.157	0.313	0.313	0.087	0.087	0.090	0.090	0.293	0.293
	Parameters	'l2_regularization': 0.0, 'learning_rate': 0.1, 'max_bins': 128, 'max_depth': 50, 'max_iter': 500, 'min_samples_leaf': 1 'class_weight': 'balanced'		'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200 'class_weight': 'balanced'		'C': 1, 'gamma': 0.01, 'kernel': 'rbf' 'class_weight': 'balanced'		'C': 1, 'max_iter': 500, 'penalty': 'l1', 'solver': 'liblinear' 'class_weight': 'balanced'		algorithm': 'auto', 'n_neighbors': 15, 'p': 1		activation': 'logistic', 'alpha': 0.1, 'early_stopping': True, 'hidden_layer_sizes': (32,16), 'learning_rate_init': 0.01, 'max_iter': 300, 'solver': 'adam'		criterion': 'entropy', 'max_depth': 12, 'min_impurity_decrease': 0.001, 'min_samples_leaf': 10, 'min_samples_split': 10 'class_weight': 'balanced'	
Goal 2	Sensitivity	**0.964**	**0.975**	**0.966**	**0.990**	**0.926**	**0.919**	**0.858**	**0.877**	**0.888**	**0.909**	**0.820**	**0.835**	**0.875**	**0.891**
	Specificity	**0.732**	**0.728**	**0.741**	**0.726**	**0.706**	**0.701**	**0.706**	**0.695**	**0.731**	**0.700**	**0.734**	**0.709**	**0.718**	**0.666**
	Precision	0.470	0.469	0.479	0.471	0.437	0.431	0.418	0.415	0.448	0.427	0.432	0.414	0.433	0.396
	F1 score	0.632	0.633	0.640	0.638	0.593	0.587	0.562	0.563	0.596	0.581	0.565	0.554	0.579	0.549
	PR-AUC	0.880	0.881	0.830	0.846	0.614	0.625	0.747	0.754	0.680	0.707	0.550	0.537	0.707	0.707
	Threshold	0.010	0.010	0.061	0.061	0.195	0.195	0.300	0.300	0.062	0.062	0.105	0.105	0.181	0.181
	Parameters	'l2_regularization': 1.0, 'learning_rate': 0.1, 'max_bins': 128, 'max_depth': 50, 'max_iter': 500, 'min_samples_leaf': 10 'class_weight': 'balanced'		max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200 'class_weight': 'balanced'		'C': 1, 'gamma': 0.1, 'kernel': 'rbf' 'class_weight': 'balanced'		'C': 1, 'max_iter': 500, 'penalty': 'l1', 'solver': 'liblinear' 'class_weight': 'balanced'		algorithm': 'auto', 'n_neighbors': 15, 'p': 1		activation': 'logistic', 'alpha': 0.1, 'early_stopping': True, 'hidden_layer_sizes': (32,16), 'learning_rate_init': 0.001, 'max_iter': 300, 'solver': 'adam'		criterion': 'entropy', 'max_depth': 12, 'min_impurity_decrease': 0.001, 'min_samples_leaf': 10, 'min_samples_split': 10 'class_weight': 'balanced'	
No goal	Sensitivity	**0.835**	**0.835**	**0.875**	**0.860**	**0.832**	**0.818**	**0.686**	**0.716**	**0.754**	**0.695**	**0.707**	**0.740**	**0.754**	**0.790**
	Specificity	0.954	0.944	0.917	0.906	0.874	0.852	0.926	0.919	0.911	0.910	0.891	0.876	0.916	0.888
	Precision	**0.821**	**0.786**	**0.724**	**0.692**	**0.620**	**0.577**	**0.704**	**0.685**	**0.683**	**0.656**	**0.617**	**0.596**	**0.692**	**0.634**
	F1 score	**0.826**	**0.810**	**0.791**	**0.767**	**0.709**	**0.676**	**0.693**	**0.700**	**0.713**	**0.675**	**0.658**	**0.660**	**0.720**	**0.703**
	PR-AUC	0.893	0.902	0.833	0.842	0.653	0.644	0.747	0.754	0.643	0.690	0.542	0.586	0.712	0.725
	Threshold	0.267	0.267	0.251	0.251	0.227	0.227	0.691	0.691	0.240	0.240	0.323	0.323	0.679	0.679
	Parameters	'l2_regularization': 0.0, 'learning_rate': 0.1, 'max_bins': 128, 'max_depth': 50, 'max_iter': 500, 'min_samples_leaf': 5 'class_weight': 'balanced'		max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200 'class_weight': 'balanced'		'C': 1, 'gamma': 0.01, 'kernel': 'rbf' 'class_weight': 'balanced'		'C': 1, 'max_iter': 500, 'penalty': 'l1', 'solver': 'liblinear' 'class_weight': 'balanced'		algorithm': 'auto', 'n_neighbors': 5, 'p': 1		activation': 'logistic', 'alpha': 0.1, 'early_stopping': True, 'hidden_layer_sizes': (32,16), 'learning_rate_init': 0.01, 'max_iter': 300, 'solver': 'adam'		criterion': 'entropy', 'max_depth': 8, 'min_impurity_decrease': 0.001, 'min_samples_leaf': 5, 'min_samples_split': 10 'class_weight': 'balanced'	

Appendix B: SHAP Results

The below table shows the feature importances of all variables used in the modelling, based on SHAP values, with `DEBTINC` and `DELINQ` consistenly became the top two predictors in all scenarios and chosen models.

Feature importance based on SHAP					
Goal 1		**Goal 2**		**No goal**	
HGBT		**RF**		**DT**	
Variables	**Importances**	**Variables**	**Importances**	**Variables**	**Importances**
DEBTINC	2.6524	DEBTINC	0.1719	DEBTINC	0.2817
DELINQ	1.3976	DELINQ	0.0679	DELINQ	0.0889
CLAGE	1.3475	DEBTINC missing	0.0589	CLAGE	0.0740
YOJ	0.7875	CLAGE	0.0584	YOJ	0.0318
DEROG missing	0.7160	DEROG	0.0381	LoanValueRatio	0.0240
CLNO	0.6721	CLNO	0.0331	DEROG	0.0200
JOB Office	0.6387	YOJ	0.0278	CLNO	0.0182
LoanValueRatio	0.5856	LoanValueRatio	0.0267	DEROG missing	0.0151
DEROG	0.5200	NINQ	0.0221	NINQ	0.0114
NINQ	0.3785	JOB Office	0.0125	DEBTINC missing	0.0106
DEBTINC missing	0.2892	DEROG missing	0.0125	LoanValueRatio missing	0.0104
REASON HomeImp	0.2034	REASON HomeImp	0.0079	JOB Office	0.0083
JOB ProfExe	0.1786	LoanValueRatio missing	0.0078	REASON HomeImp	0.0058
LoanValueRatio missing	0.1749	JOB Other	0.0070	JOB Other	0.0026
JOB missing	0.1710	JOB ProfExe	0.0056	REASON Not Provided	0.0000
JOB Other	0.1310	DELINQ missing	0.0043	JOB ProfExe	0.0000
NINQ missing	0.0644	JOB Sales	0.0042	JOB Sales	0.0000
JOB Sales	0.0509	YOJ missing	0.0035	JOB Self	0.0000
DELINQ missing	0.0362	JOB missing	0.0032	NINQ missing	0.0000
YOJ missing	0.0353	NINQ missing	0.0024	DELINQ missing	0.0000
REASON Not Provided	0.0263	JOB Self	0.0020	JOB missing	0.0000

Appendix C: Confusion Matrices

Below are the confusion matrices of the chosen models for each scenario, from both validation and test sets which are the basis for metrics calculation (e.g., sensitivity, specificity, precision, F1 score, etc).

Goal 1 (HGBT Model)

Validation Set (75%)

Actual \ Predicted	0	1
0	TN 3,223	FP 249
1	FN 114	TP 740
Total	3,337	989

Test Set (25%)

Actual \ Predicted	0	1
0	TN 1,074	FP 83
1	FN 42	TP 243
Total	1,116	326

Predicted approvals: 4,453 (77%)
Predicted rejections: 1,315 (23%)

Goal 2 (RF Model)

Validation Set (75%)

Actual \ Predicted	0	1
0	TN 2,502	FP 970
1	FN 26	TP 828
Total	2,528	1798

Test Set (25%)

Actual \ Predicted	0	1
0	TN 815	FP 342
1	FN 4	TP 281
Total	819	623

Predicted approvals: 3,347 (58%)
Predicted rejections: 2,421 (42%)

Goal 3 (DT Model)

Validation Set (75%)

Actual \ Predicted	0	1
0	TN 3,179	FP 293
1	FN 210	TP 644
Total	3,389	937

Test Set (25%)

Actual \ Predicted	0	1
0	TN 1,027	FP 130
1	FN 60	TP 225
Total	1,087	355

Predicted approvals: 4,476 (78%)
Predicted rejections: 1,292 (22%)

Appendix D: Decision Tree Rule