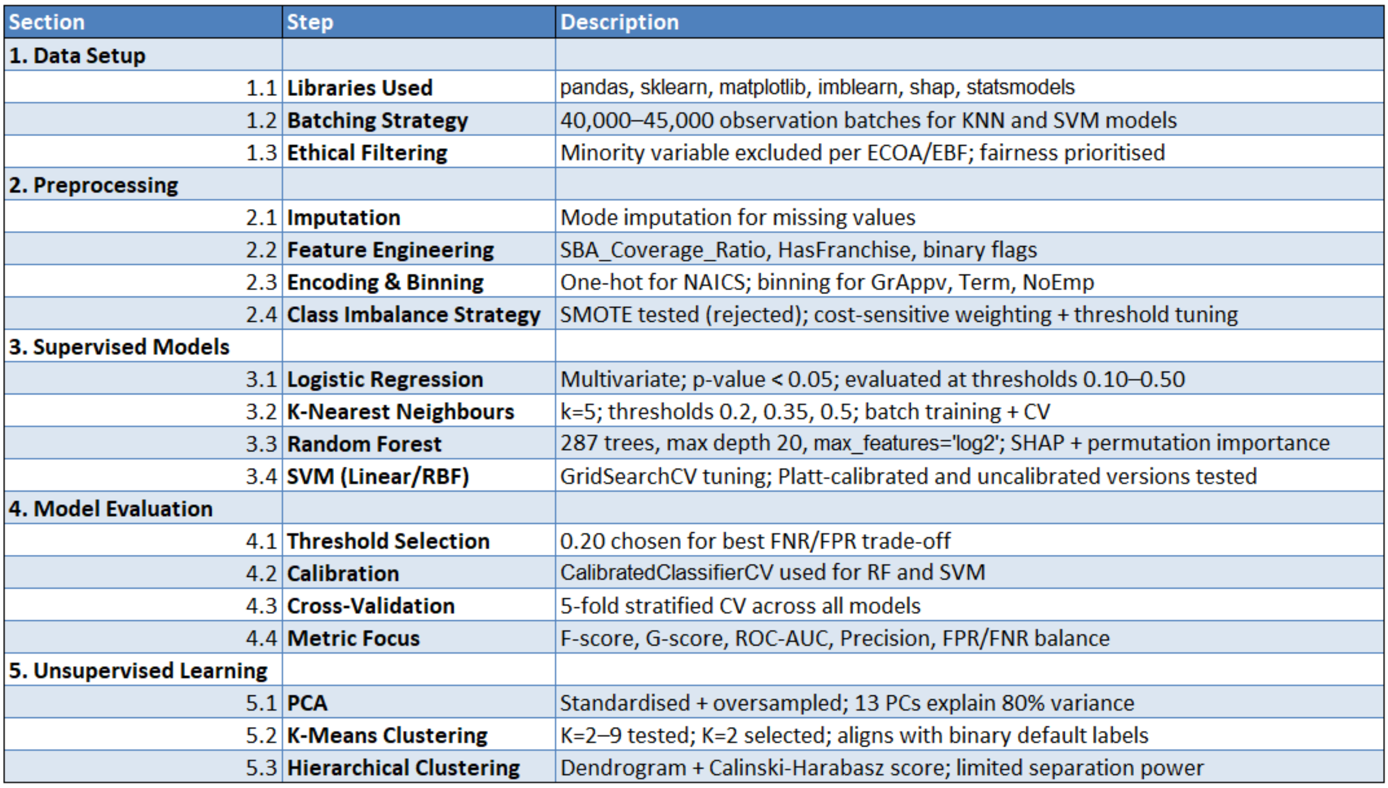
Predicting Loan Default Using a Machine Learning Algorithm

David Crowe | University College Dublin

Executive Summary

This project analyses ~ 900,000 SBA loans from 1962–2014 to develop a predictive model for loan default in support of commercial lending decisions. Following data cleaning, feature engineering, and class imbalance mitigation, we trained multiple classifiers: Logistic Regression, K-Nearest Neighbours, Random Forest, and Support Vector Machines. Emphasis was placed on minimising false negatives, reflecting the cost of undetected defaults. We also applied unsupervised learning to explore latent borrower structures.

Our final recommendation is a calibrated Random Forest model evaluated at a 20% threshold, which achieved the best trade-off between sensitivity, precision, and fairness. The table below outlines our modelling pipeline, ensuring transparency, reproducibility, and alignment with regulatory practices (EBA, 2023; EBF, 2019).

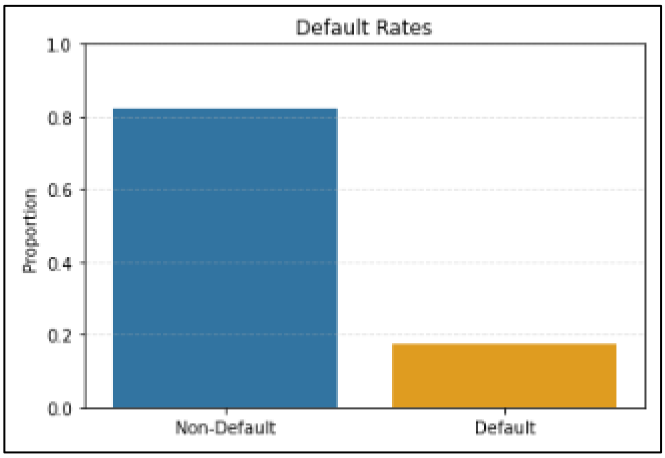


Section	Step	Description
1. Data Setup		
1.1	**Libraries Used**	pandas, sklearn, matplotlib, imblearn, shap, statsmodels
1.2	**Batching Strategy**	40,000–45,000 observation batches for KNN and SVM models
1.3	**Ethical Filtering**	Minority variable excluded per ECOA/EBF; fairness prioritised
2. Preprocessing		
2.1	**Imputation**	Mode imputation for missing values
2.2	**Feature Engineering**	SBA_Coverage_Ratio, HasFranchise, binary flags
2.3	**Encoding & Binning**	One-hot for NAICS; binning for GrAppv, Term, NoEmp
2.4	**Class Imbalance Strategy**	SMOTE tested (rejected); cost-sensitive weighting + threshold tuning
3. Supervised Models		
3.1	**Logistic Regression**	Multivariate; p-value < 0.05; evaluated at thresholds 0.10–0.50
3.2	**K-Nearest Neighbours**	k=5; thresholds 0.2, 0.35, 0.5; batch training + CV
3.3	**Random Forest**	287 trees, max depth 20, max_features='log2'; SHAP + permutation importance
3.4	**SVM (Linear/RBF)**	GridSearchCV tuning; Platt-calibrated and uncalibrated versions tested
4. Model Evaluation		
4.1	**Threshold Selection**	0.20 chosen for best FNR/FPR trade-off
4.2	**Calibration**	CalibratedClassifierCV used for RF and SVM
4.3	**Cross-Validation**	5-fold stratified CV across all models
4.4	**Metric Focus**	F-score, G-score, ROC-AUC, Precision, FPR/FNR balance
5. Unsupervised Learning		
5.1	**PCA**	Standardised + oversampled; 13 PCs explain 80% variance
5.2	**K-Means Clustering**	K=2–9 tested; K=2 selected; aligns with binary default labels
5.3	**Hierarchical Clustering**	Dendrogram + Calinski-Harabasz score; limited separation power

1. Data

We aim to predict business default by recoding **MIS_Status** into a binary Default variable (1=**Default**).

1.1 Class imbalance

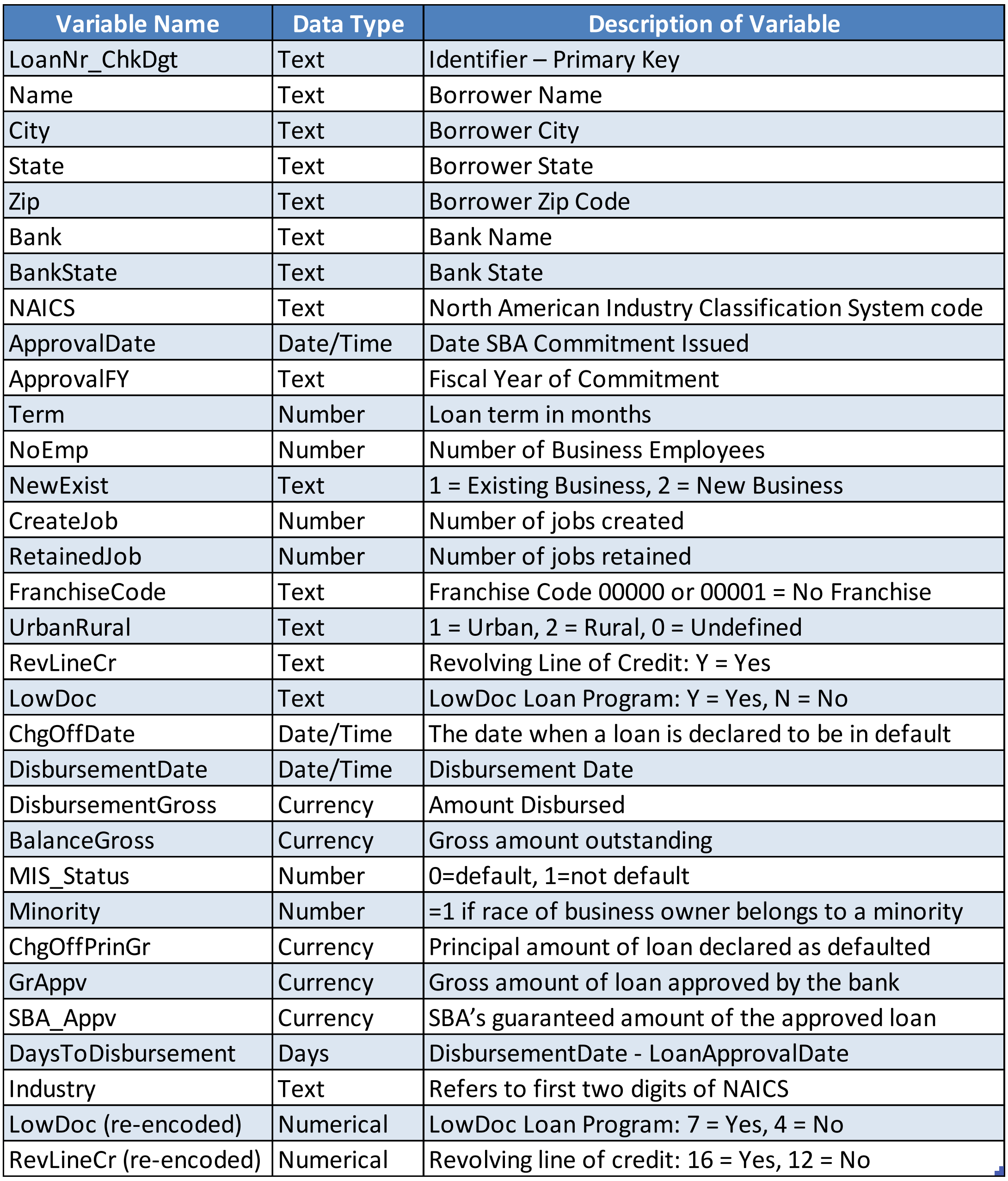


As shown above, fewer than 20% of loans defaulted, creating class imbalance, a challenge common in credit scoring (Johnson & Khoshgoftaar, 2019). Class imbalance biases models toward non-defaults and weakens detection of risky loans. Since false negatives are costlier, we tested multiple thresholds and applied mitigation techniques across all models.

Model	Balancing Technique
Logistic	Synthetic Minority Over-sampling Technique (SMOTE)
Random Forest	Class Weighting
SVM	Class Weighting
KNN	Synthetic Minority Over-sampling Technique (SMOTE)
PCA's	Synthetic Minority Over-sampling Technique (SMOTE)

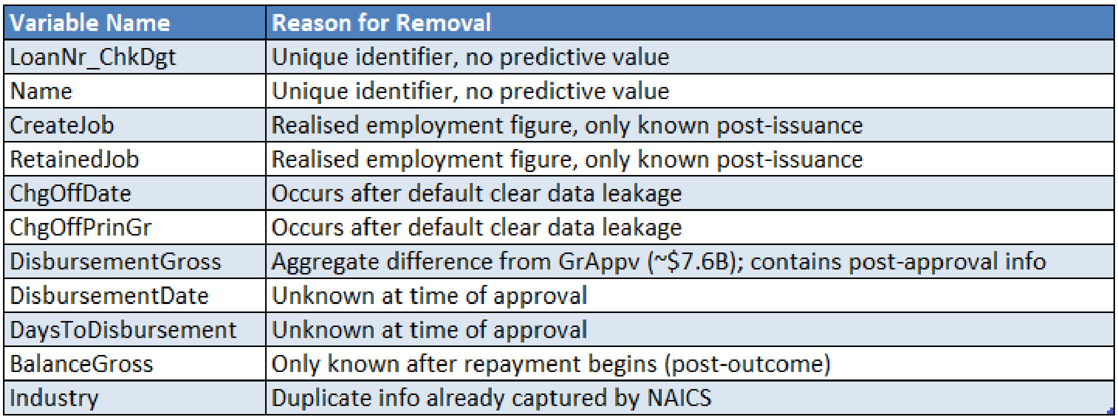
1.2 Data cleaning

We removed variables with data leakage, low predictive power, or ethical risks.



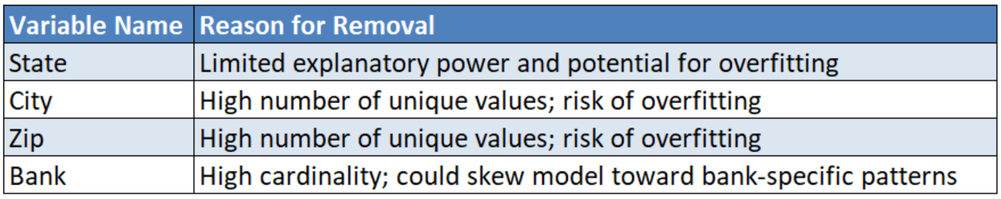
Variable Name	Data Type	Description of Variable
LoanNr_ChkDgt	Text	Identifier – Primary Key
Name	Text	Borrower Name
City	Text	Borrower City
State	Text	Borrower State
Zip	Text	Borrower Zip Code
Bank	Text	Bank Name
BankState	Text	Bank State
NAICS	Text	North American Industry Classification System code
ApprovalDate	Date/Time	Date SBA Commitment Issued
ApprovalFY	Text	Fiscal Year of Commitment
Term	Number	Loan term in months
NoEmp	Number	Number of Business Employees
NewExist	Text	1 = Existing Business, 2 = New Business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise Code 00000 or 00001 = No Franchise
UrbanRural	Text	1 = Urban, 2 = Rural, 0 = Undefined
RevLineCr	Text	Revolving Line of Credit: Y = Yes
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement Date
DisbursementGross	Currency	Amount Disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Number	0=default, 1=not default
Minority	Number	=1 if race of business owner belongs to a minority
ChgOffPrinGr	Currency	Principal amount of loan declared as defaulted
GrAppv	Currency	Gross amount of loan approved by the bank
SBA_Appv	Currency	SBA’s guaranteed amount of the approved loan
DaysToDisbursement	Days	DisbursementDate - LoanApprovalDate
Industry	Text	Refers to first two digits of NAICS
LowDoc (re-encoded)	Numerical	LowDoc Loan Program: 7 = Yes, 4 = No
RevLineCr (re-encoded)	Numerical	Revolving line of credit: 16 = Yes, 12 = No

The following variables were removed:

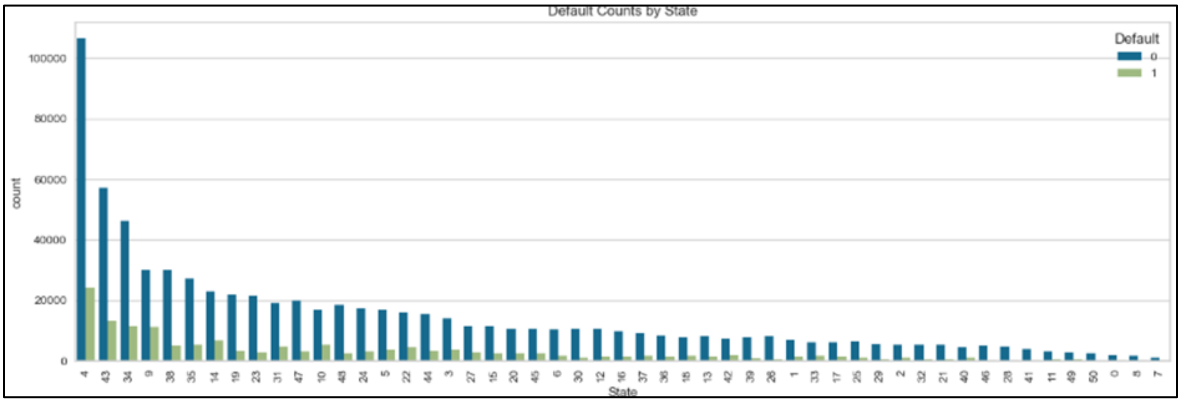


Variable Name	Reason for Removal
LoanNr_ChkDgt	Unique identifier, no predictive value
Name	Unique identifier, no predictive value
CreateJob	Realised employment figure, only known post-issuance
RetainedJob	Realised employment figure, only known post-issuance
ChgOffDate	Occurs after default clear data leakage
ChgOffPrinGr	Occurs after default clear data leakage
DisbursementGross	Aggregate difference from GrAppv (~$7.6B); contains post-approval info
DisbursementDate	Unknown at time of approval
DaysToDisbursement	Unknown at time of approval
BalanceGross	Only known after repayment begins (post-outcome)
Industry	Duplicate info already captured by NAICS

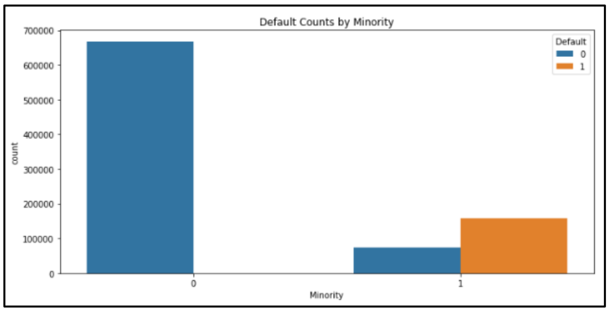
Despite statistical relevance in Chi-Squared tests, we dropped the variables below:



Variable Name	Reason for Removal
State	Limited explanatory power and potential for overfitting
City	High number of unique values; risk of overfitting
Zip	High number of unique values; risk of overfitting
Bank	High cardinality; could skew model toward bank-specific patterns



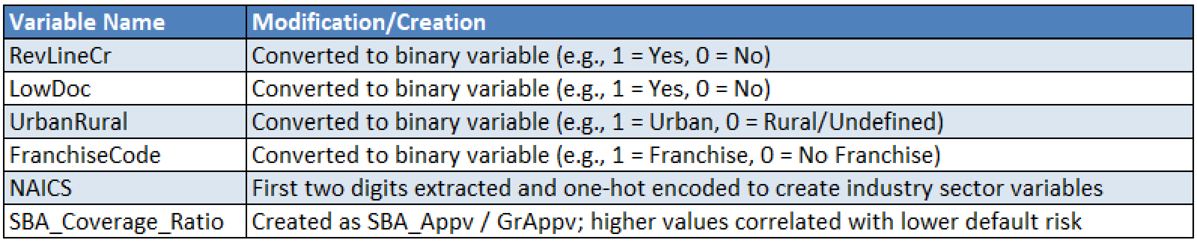
As shown above, while some states have higher loan volumes, the distribution of defaults closely mirrors overall loan counts, suggesting that state-level variation lacks predictive power and could lead to overfitting without adding meaningful signal.



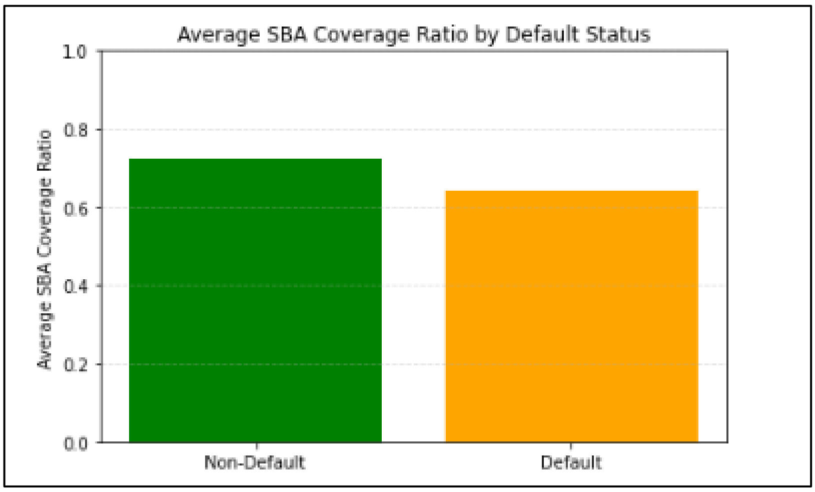
Despite statistical strength, **Minority** was excluded for compliance with ECOA and to avoid discriminatory bias (EBF, 2023).

1.3 Feature Engineering

To improve predictive signal, we engineered several features:



Variable Name	Modification/Creation
RevLineCr	Converted to binary variable (e.g., 1 = Yes, 0 = No)
LowDoc	Converted to binary variable (e.g., 1 = Yes, 0 = No)
UrbanRural	Converted to binary variable (e.g., 1 = Urban, 0 = Rural/Undefined)
FranchiseCode	Converted to binary variable (e.g., 1 = Franchise, 0 = No Franchise)
NAICS	First two digits extracted and one-hot encoded to create industry sector variables
SBA_Coverage_Ratio	Created as SBA_Appv / GrAppv; higher values correlated with lower default risk



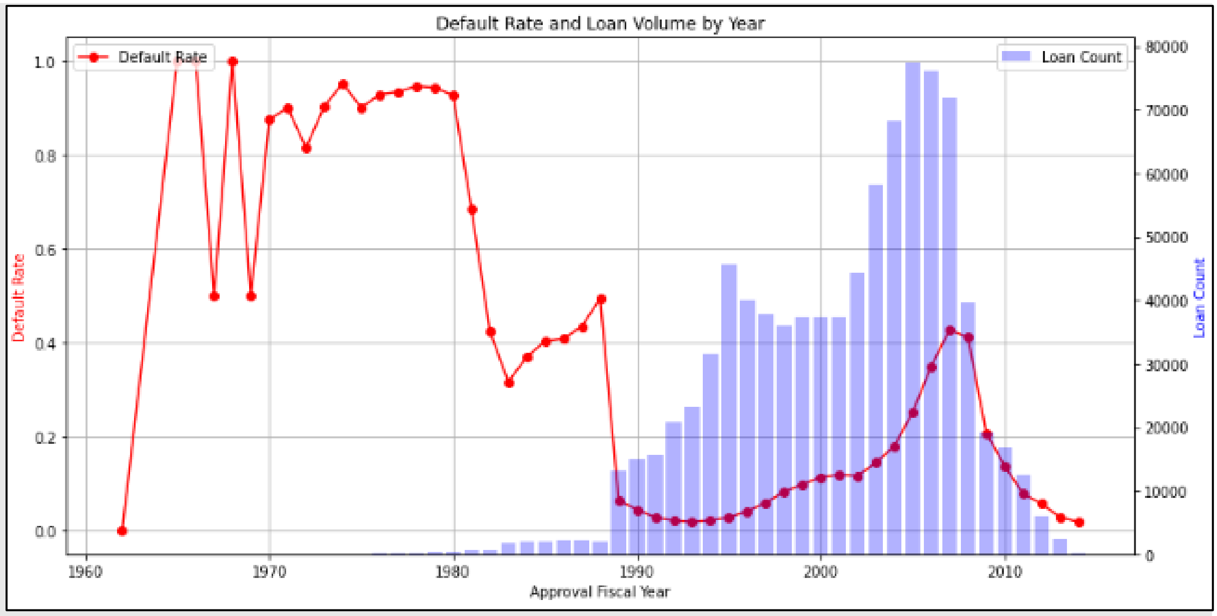
- **SBA_Coverage_Ratio**=**SBA_Appv** / **GrAppv**, capturing the proportion of the loan guaranteed by SBA. As shown above, non-defaulters had higher average guarantees, indicating its insightfulness as a potential feature.

- **NAICS_Sector**: Extracted first two digits of NAICS code to group firms into sectors (one-hot encoded).

- **LowDoc**, **FranchiseCode** (converted to **HasFranchise**), and **UrbanRural** were re-encoded as 0/1 binary flags.

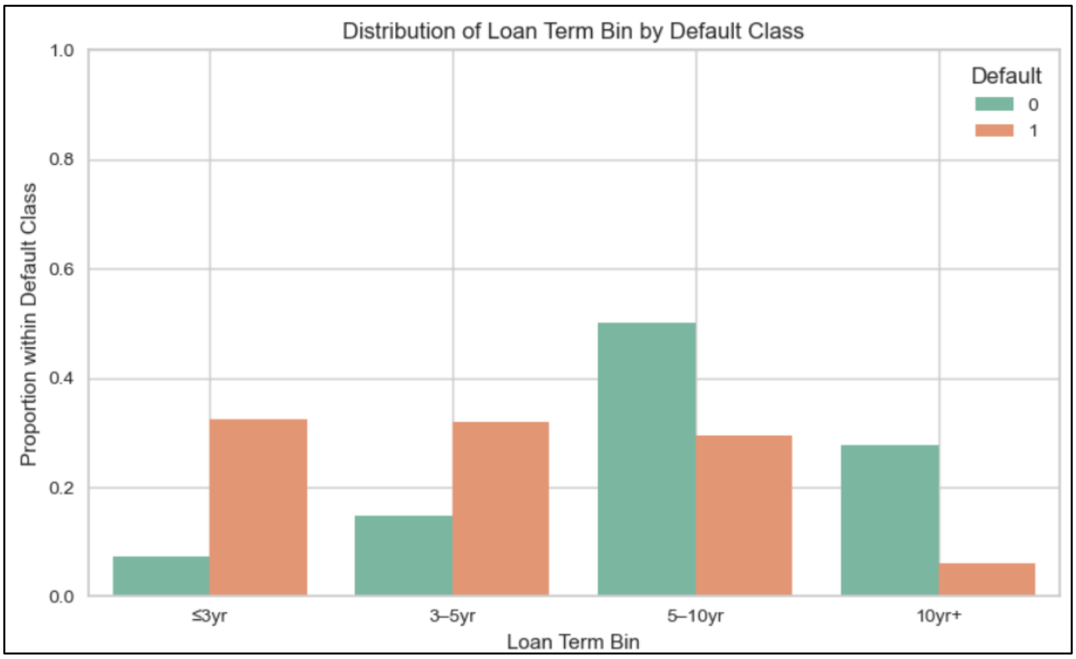
Missing values were imputed using mode. **RevLineCr** contained ~200,000 missing values and **RevLineCr_missing** was created to capture potential information.

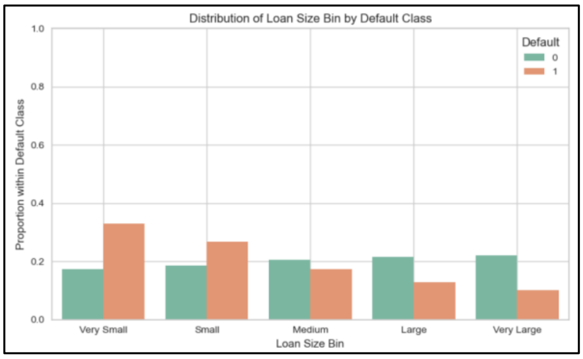
1.4 Exploratory & Visual Analysis

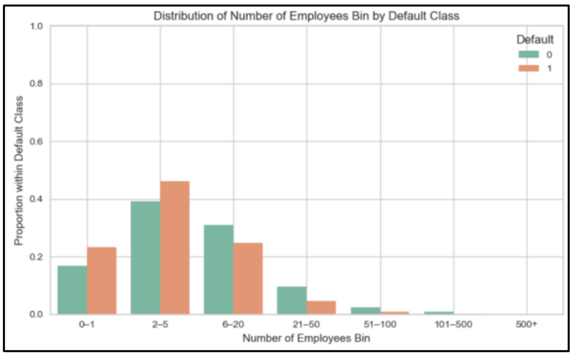


Pre-1985 default rates were volatile due to sparse data. From the late 1980s, patterns mirrored macro cycles, with a spike during the 2008 crisis. To reduce time-dependence, we dropped **ApprovalFY**.

To explore numerical effects, we binned key variables (**Term, GrAppv, NoEmp**) to visualise their relationship with default rates. Evidently, shorter terms, lower loan amounts and smaller firms were all strongly associated with higher default rates, validating their inclusion as features.

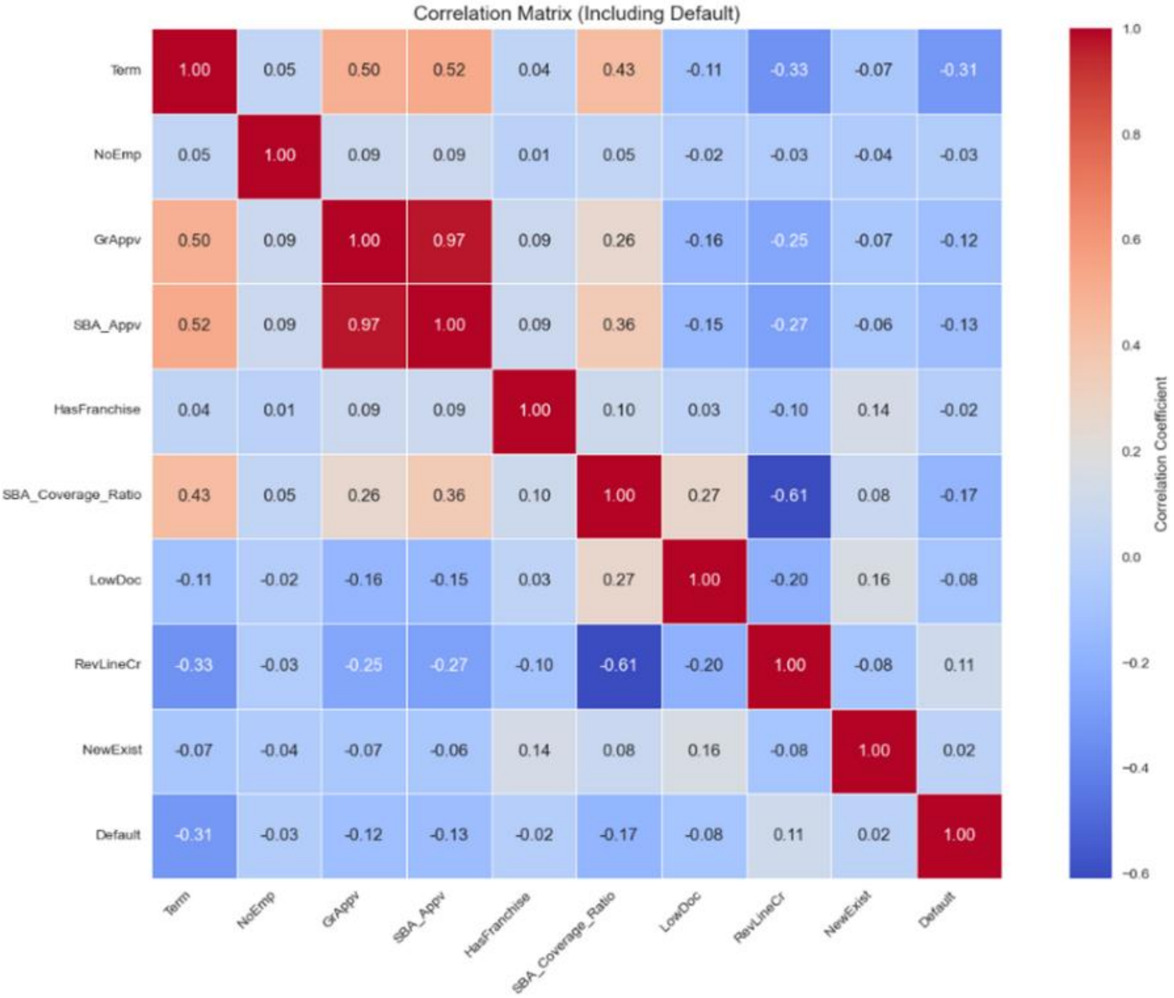






1.5 Multicollinearity Analysis

We assessed multicollinearity using a correlation matrix and VIF analysis to ensure model stability and interpretability



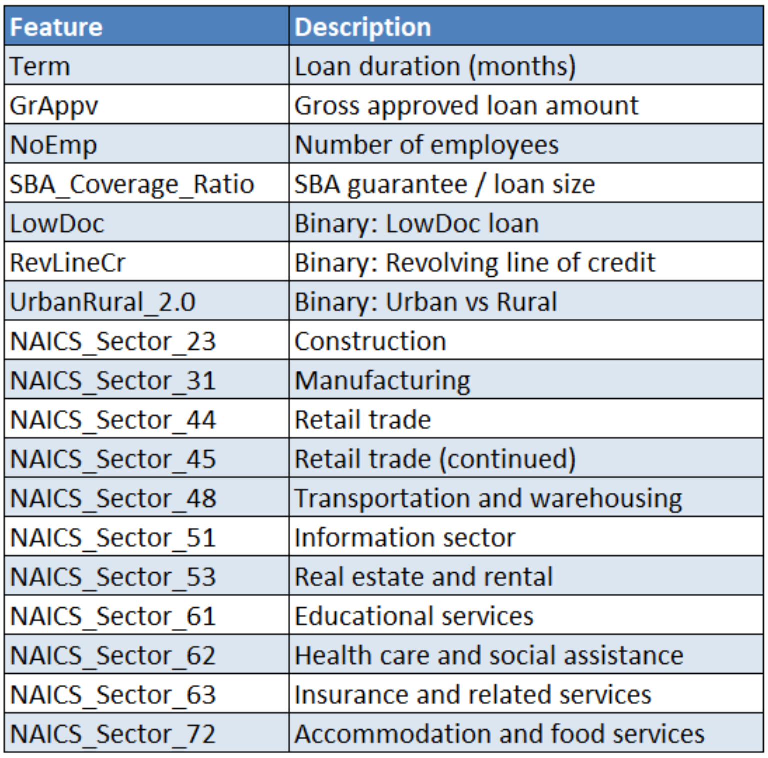
	Term	NoEmp	GrAppv	SBA_Appv	HasFranchise	SBA_Coverage_Ratio	LowDoc	RevLineCr	NewExist	Default
Term	1.00	0.05	0.50	0.52	0.04	0.43	-0.11	-0.33	-0.07	-0.31
NoEmp	0.05	1.00	0.09	0.09	0.01	0.05	-0.02	-0.03	-0.04	-0.03
GrAppv	0.50	0.09	1.00	0.97	0.09	0.26	-0.16	-0.25	-0.07	-0.12
SBA_Appv	0.52	0.09	0.97	1.00	0.09	0.36	-0.15	-0.27	-0.06	-0.13
HasFranchise	0.04	0.01	0.09	0.09	1.00	0.10	0.03	-0.10	0.14	-0.02
SBA_Coverage_Ratio	0.43	0.05	0.26	0.36	0.10	1.00	0.27	-0.61	0.08	-0.17
LowDoc	-0.11	-0.02	-0.16	-0.15	0.03	0.27	1.00	-0.20	0.16	-0.08
RevLineCr	-0.33	-0.03	-0.25	-0.27	-0.10	-0.61	-0.20	1.00	-0.08	0.11
NewExist	-0.07	-0.04	-0.07	-0.06	0.14	0.08	0.16	-0.08	1.00	0.02
Default	-0.31	-0.03	-0.12	-0.13	-0.02	-0.17	-0.08	0.11	0.02	1.00

A strong correlation (r=0.97) between **GrAppv** and **SBA_Appv** led us to retain **GrAppv**, drop **SBA_Appv**, and keep **SBA_Coverage_Ratio** (no multicollinearity).

1.6 Mutual Information Analysis

We used Mutual Information to identify non-linear links to default, excluding features with MI < 0.003, including **HasFranchise**, **NewExist**, and several NAICS sectors.

Our final model was trained on the below 18 features, each selected based on predictive relevance, interpretability and fairness.



Feature	Description
Term	Loan duration (months)
GrAppv	Gross approved loan amount
NoEmp	Number of employees
SBA_Coverage_Ratio	SBA guarantee / loan size
LowDoc	Binary: LowDoc loan
RevLineCr	Binary: Revolving line of credit
UrbanRural_2.0	Binary: Urban vs Rural
NAICS_Sector_23	Construction
NAICS_Sector_31	Manufacturing
NAICS_Sector_44	Retail trade
NAICS_Sector_45	Retail trade (continued)
NAICS_Sector_48	Transportation and warehousing
NAICS_Sector_51	Information sector
NAICS_Sector_53	Real estate and rental
NAICS_Sector_61	Educational services
NAICS_Sector_62	Health care and social assistance
NAICS_Sector_63	Insurance and related services
NAICS_Sector_72	Accommodation and food services

2. Feature Engineering

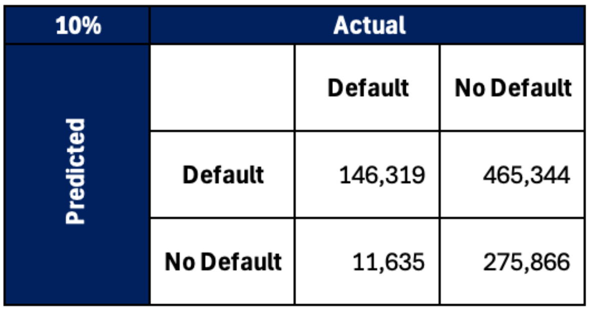
2.1 Logistic Regression

We applied a multivariate logistic regression model to estimate the probability of a borrower defaulting on a loan based on the features outlined in Question 1, using the formula below.

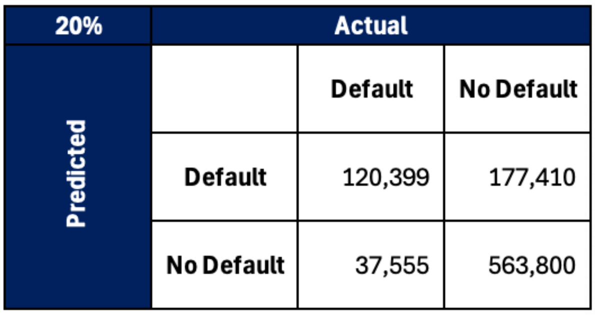
$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}$$

If *p(X),* exceeds the specified threshold, the model classifies the observation as a default (1); otherwise, as non-default (0). An initial model was run using the selected variables from Question 1. All had p-values of 0.000, except **SBA_Coverage_Ratio** (p=0.176), which was dropped due to statistical insignificance.

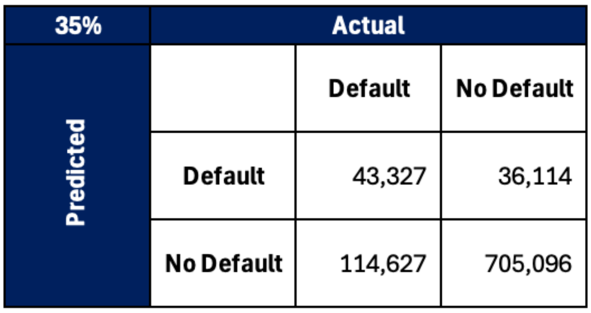
A summary of confusion matrices across thresholds is shown below:



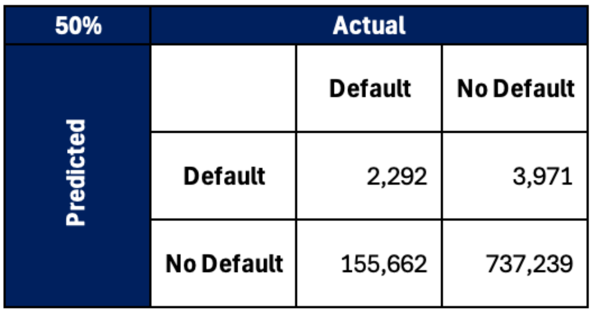
10%	Actual		
Predicted		Default	No Default
	Default	146,319	465,344
	No Default	11,635	275,866



20%	Actual		
Predicted		Default	No Default
	Default	120,399	177,410
	No Default	37,555	563,800



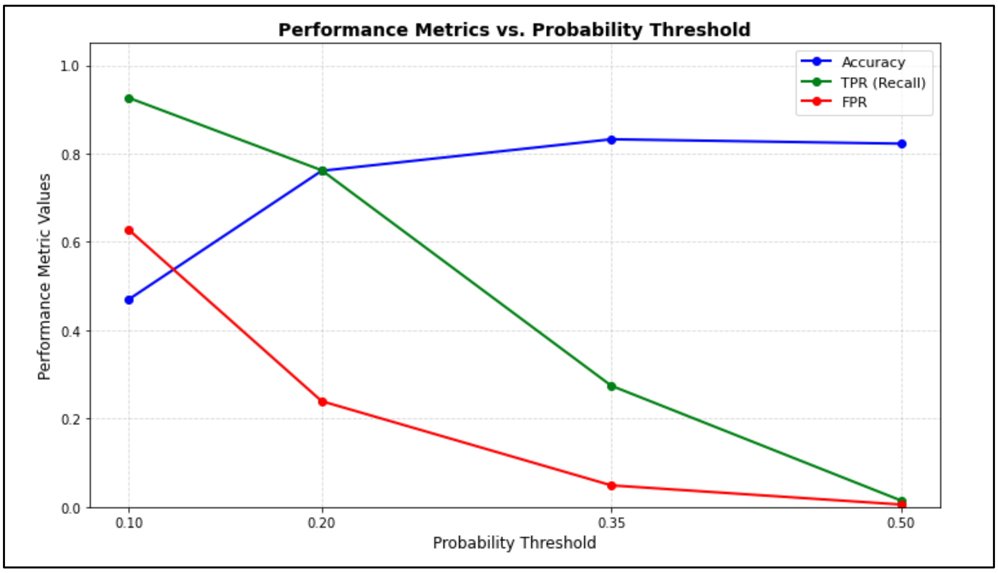
35%	Actual		
Predicted		Default	No Default
	Default	43,327	36,114
	No Default	114,627	705,096



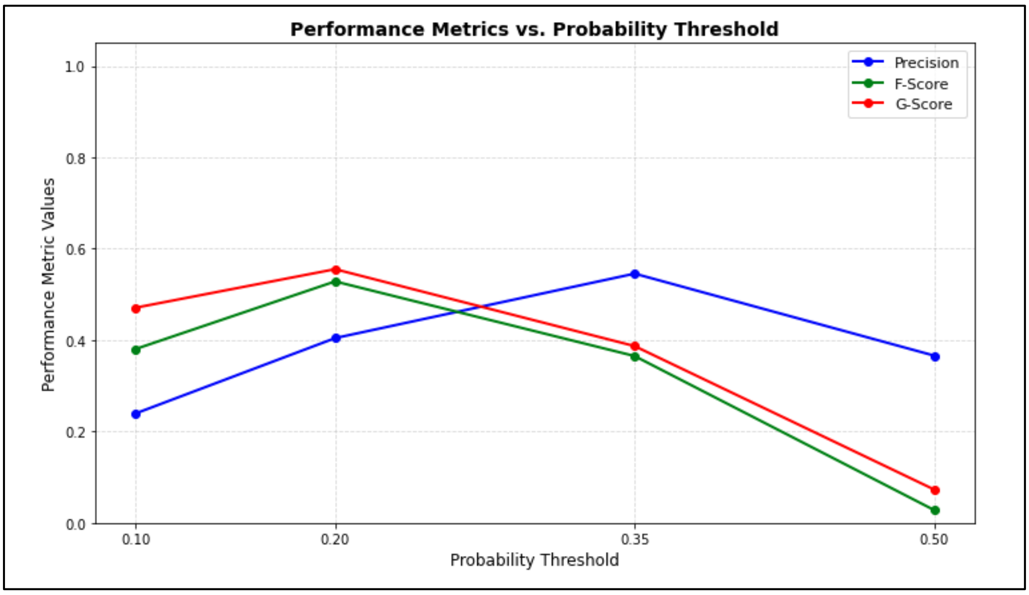
50%	Actual		
Predicted		Default	No Default
	Default	2,292	3,971
	No Default	155,662	737,239

Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score	G-Score
10%	92.63%	62.78%	37.22%	7.37%	46.95%	23.92%	38.02%	47.07%
20%	76.22%	23.94%	76.06%	23.78%	76.09%	40.43%	52.83%	55.51%
35%	27.43%	4.87%	95.13%	72.57%	83.24%	54.54%	36.50%	38.68%
50%	1.45%	0.54%	99.46%	98.55%	82.25%	36.60%	2.79%	7.29%

The confusion matrix at the 50% threshold reveals a clear class imbalance: 892,901 predicted negatives and only 6,263 predicted positives. This introduces cost asymmetry, making accuracy a misleading metric as it treats FNs and FPs equally. This is problematic as approving a defaulter is more costly than rejecting a non-defaulter in this context. Whilst the 0.10 threshold yields the lowest FNR, it produces an excessive FPR of 62.78%, making it commercially inviable. In contrast, 0.20 achieves a better balance between recall and precision, with improved F/G-scores. Although this entails a minor loss in accuracy, it significantly reduces the cost of undetected defaults.

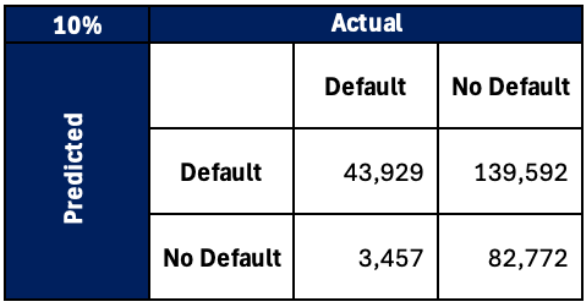


The graphs here highlight the movement of our performance metrics across thresholds. As the threshold rises, the accuracy rises and the TPR and FPR fall. Our F and G-scores peak at 0.20. Evidently, 0.10 and 0.20 are superior to our other thresholds, but again we confirm 0.20 as our selection, sacrificing a decrease in TPR of 16.41% from 0.1 in exchange for a severe improvement of 38.84% in FPR, as well as significantly higher accuracy and precision. We also achieve increases of 14.81% and 8.44% in F-scores and G-scores, respectively.

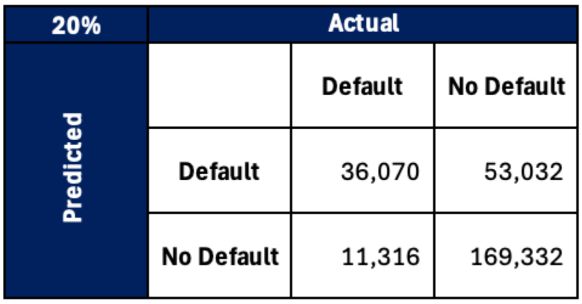


2.1 Splitting Dataset

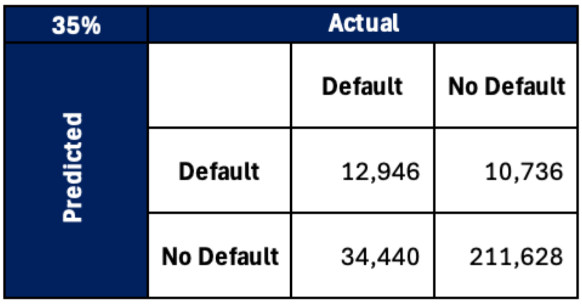
We divided our dataset into a training and test set to evaluate model performance on unseen data, using a 70/30 split. The results are summarised in the confusion matrices below, applied during the testing phase across our respective thresholds.



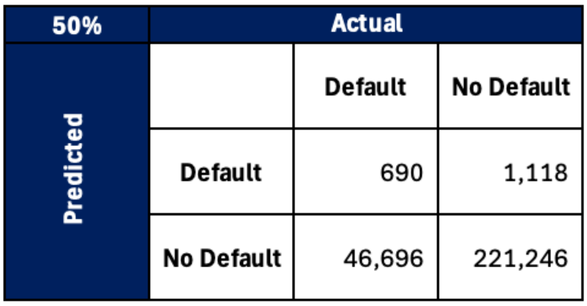
10%		Actual	
Predicted		Default	No Default
	Default	43,929	139,592
	No Default	3,457	82,772



20%		Actual	
Predicted		Default	No Default
	Default	36,070	53,032
	No Default	11,316	169,332



35%		Actual	
Predicted		Default	No Default
	Default	12,946	10,736
	No Default	34,440	211,628



50%		Actual	
Predicted		Default	No Default
	Default	690	1,118
	No Default	46,696	221,246

Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score	G-Score
10%	92.70%	62.78%	37.22%	7.30%	46.97%	23.94%	38.05%	47.11%
20%	76.12%	23.85%	76.15%	23.88%	76.15%	40.48%	52.85%	55.51%
35%	27.32%	4.83%	95.17%	72.68%	83.25%	54.67%	36.43%	38.65%
50%	1.46%	0.50%	99.50%	98.54%	82.27%	38.16%	2.81%	7.45%

We observe strong consistency between the full sample and test results, with similar performance metrics in our respective thresholds, indicating that the model generalises well to unseen data. This trade-off prioritises a commercially sensitive gain in FNR. This reinforces the 0.20 threshold as optimal for our test set, offering the best trade-off between FPR and FNR, and maximising both the F and G-score.

Synthetic Minority Oversampling Technique (SMOTE)

Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score	G-Score
10%	99.97%	99.97%	0.03%	0.03%	17.59%	17.57%	29.88%	41.91%
20%	98.94%	86.38%	13.62%	1.06%	28.60%	19.62%	32.74%	44.06%
35%	91.99%	64.75%	35.25%	8.01%	45.22%	23.24%	37.11%	46.24%
50%	6.90%	2.57%	97.43%	93.10%	81.53%	36.39%	11.60%	15.85%

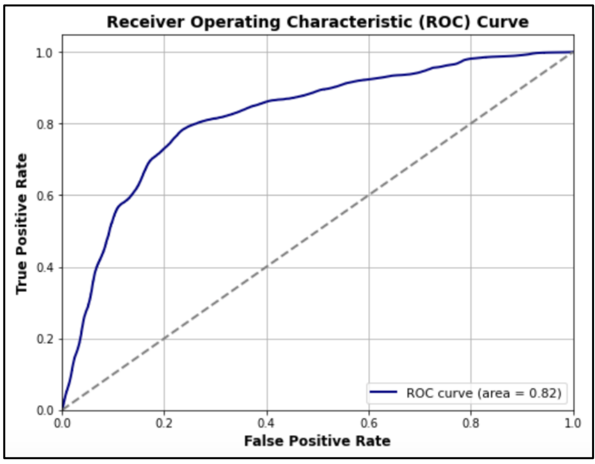
To evaluate whether class imbalance could be mitigated more directly, we applied SMOTE to the training portion of the 70/30 split. SMOTE artificially generates new examples of the minority class (defaults), aiming to increase sensitivity to rare events. Whilst this approach improved TPRs, it severely inflated FPRs and delivered weaker F and G-scores relative to the original model. Synthetic samples likely introduced noise, distorting boundaries. Therefore, SMOTE was not adopted further, retaining the original training setup as more reliable. Agarwal et al. (2023).

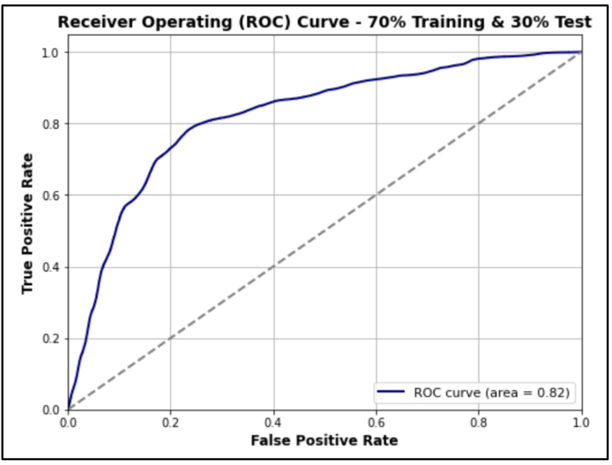
Cross-validation (5-fold)

Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score	G-Score
10%	92.68%	62.83%	37.17%	7.32%	46.92%	23.92%	38.02%	47.08%
20%	76.35%	23.68%	76.32%	23.65%	76.32%	40.72%	53.12%	55.76%
35%	27.09%	4.76%	95.24%	72.91%	83.27%	54.80%	36.25%	38.53%
50%	1.11%	0.48%	99.52%	98.89%	82.24%	33.15%	2.16%	6.08%

We applied 5-fold cross-validation, randomly dividing n observations into five non-overlapping groups, with each fold used once as a validation set whilst the others form the training set, with model performance averaged across the five folds. We use k=5 throughout this paper as it typically balances the bias-variance trade-off, rarely experiencing excessively high bias/variance. Our results closely mirror our previous results, with minimal deviations, reinforcing 0.20 once again as the optimal threshold to balance FNR and accuracy.

2.2 ROC & AUC





The ROC curve depicts TPR against FPR across thresholds, with the AUC ranging from 0 to 1, measuring the ability of a binary classifier to distinguish between classes. AUC≈1, which hugs the top left corner, is an ideal model, whilst AUC=0.5 (dashed lines) indicates that predictions are indistinguishable from random chance. Our full & split dataset both have AUC=0.82 (strong predictive power), indicating an 82% probability that a random positive example will be ranked above a random negative example. However, we note that AUC doesn't account for the asymmetric costs associated with FPs/FNs, limiting its use as a standalone performance metric in this context.

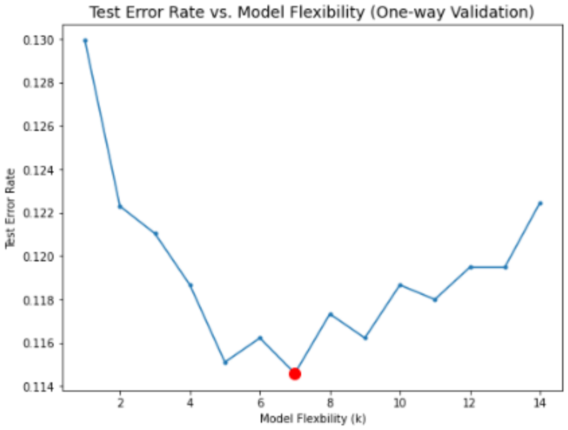
3. Modelling Approach

Due to the computational complexity of fitting KNN and SVM models on ~900,000 observations, we implemented a 'batching' approach, dividing the dataset randomly into equally sized batches of 40,000-45,000 observations. Within each batch, models were trained and evaluated independently, with performance metrics aggregated to assess overall effectiveness. This approach ensured scalability whilst maintaining representative performance across the entire dataset.

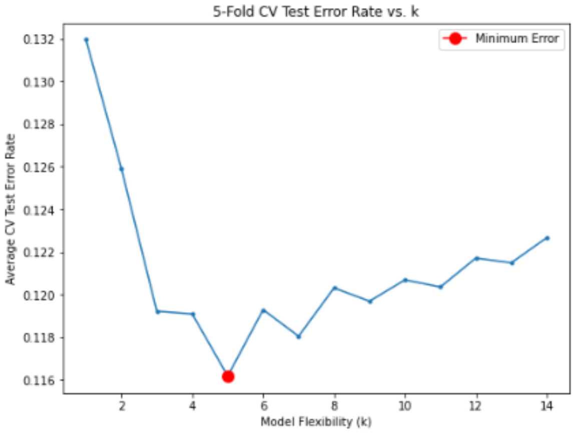
3.1 KNN

KNN classifies observations by majority vote among their K nearest neighbours, guided by a decision threshold. Due to runtime constraints, tuning (i.e. selecting k and threshold) on the full dataset was infeasible, so we used a single random batch of 45,000 observations.

To select *k*, we minimised test error using one-way and 5-fold cross-validation at a 0.5 threshold. Both produced U-shaped curves, reflecting the bias-variance trade-off between overfitting (low k) and underfitting (high k). While one-way suggested *k*=7, cross-validation was more robust, averaging errors across folds. We selected *k*=5 for its balance of flexibility and generalisability.



```
Best k = 7 with avg error = 0.1146
```



```
Best k = 5 with avg error = 0.1162
```

We evaluated a 5-NN model at thresholds of 0.2, 0.35 and 0.5 using cross-validation.

(NB: for k=5, thresholds 0.1 and 0.2 yield identical results, as one positive neighbour suffices).

20%	Actual		
Predicted		Default	No Default
	Default	6,778	8,481
	No Default	1,127	28,614

35%	Actual		
Predicted		Default	No Default
	Default	5,648	3,756
	No Default	2,257	33,339

50%	Actual		
Predicted		Default	No Default
	Default	4,457	1,780
	No Default	3,448	35,315

Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-score	AUC
0.2	85.74%	22.86%	77.14%	14.26%	78.65%	44.43%	58.53%	86.78%
0.35	71.45%	10.13%	89.87%	28.55%	86.64%	60.06%	65.26%	86.78%
0.5	56.38%	4.80%	95.20%	43.62%	88.38%	71.45%	63.03%	86.78%

The 0.5 threshold gave highest precision and accuracy but also the highest FNR. In line with our objective to minimise false negatives, we selected 0.2 despite a lower F-score and 8–10pp drop in accuracy.

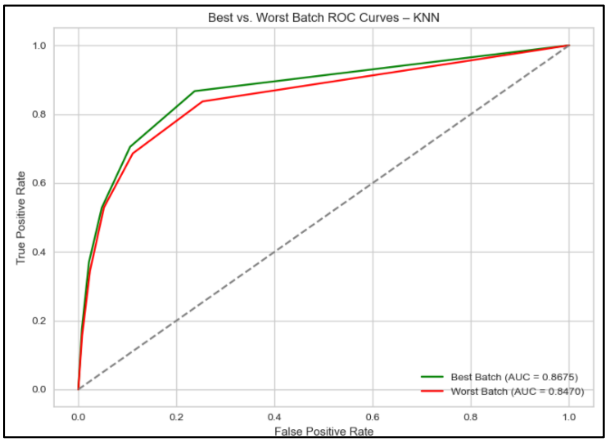
We applied *k*=5 with a 0.2 threshold across all batches using both validation methods, producing ROC curves and metrics per batch.

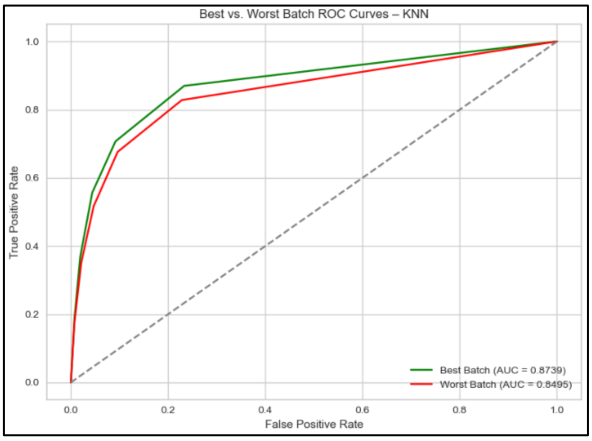
One-Way Validation

20%	Actual		
Predicted		Default	No Default
	Default	40,079	53,473
	No Default	7,297	168,851

5-Fold Cross-validation

20%	Actual		
Predicted		Default	No Default
	Default	134,145	175,312
	No Default	23,775	565,768



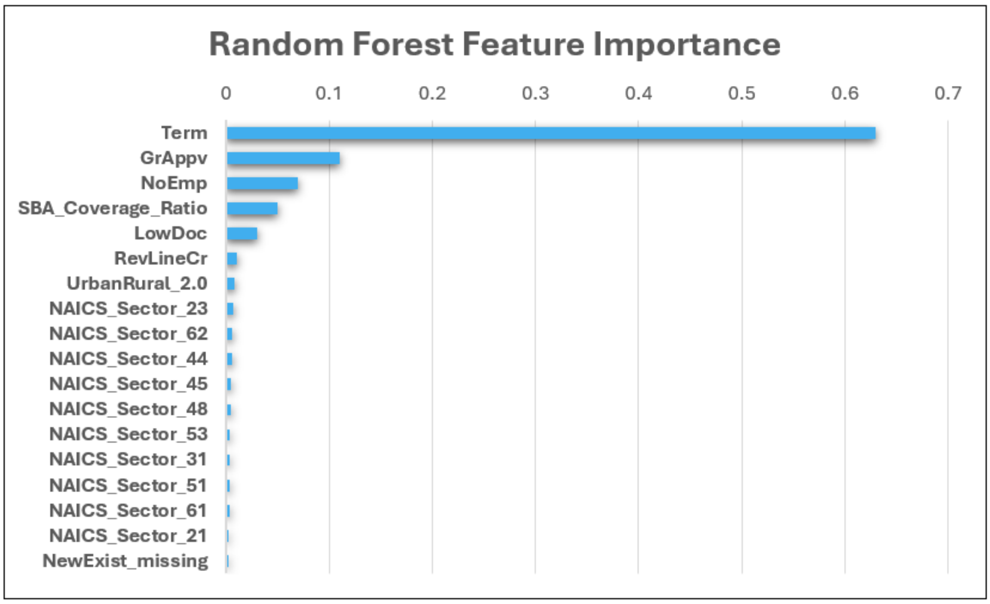


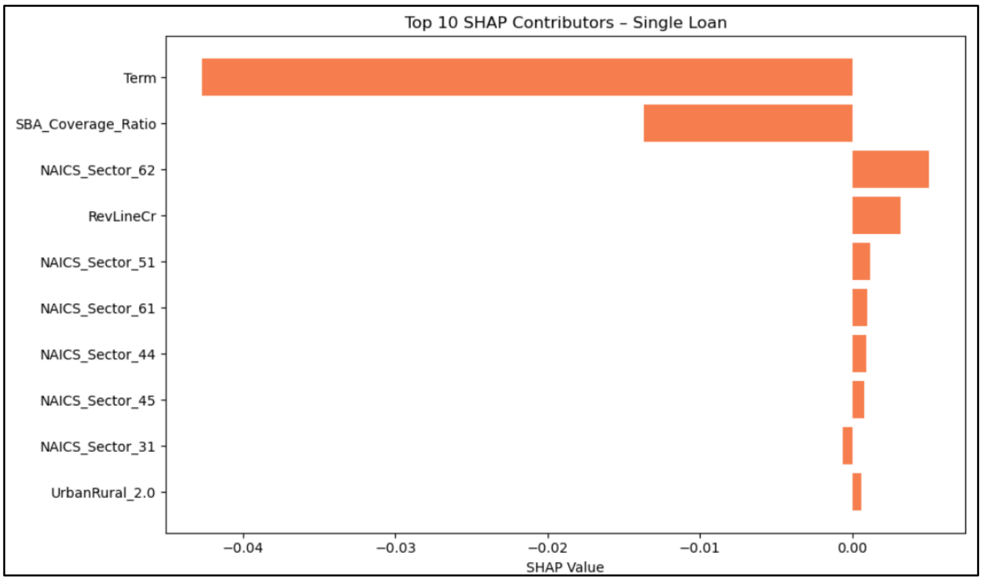
Threshold	Method	TPR	FPR	TNR	FNR	Accuracy	Precision	F-score	AUC
0.2	One-Way	84.60%	24.05%	75.95%	15.40%	77.47%	42.85%	56.88%	85.71%
	5-Fold CV	84.94%	23.66%	76.34%	15.06%	77.85%	43.37%	57.41%	86.12%

At the 0.2 threshold, 5-NN detected ~85% of defaulters while limiting false positives. Cross-validation yielded superior TPR, accuracy and AUC, confirming it as our optimal KNN model.

3.2 Random Forest

We trained a baseline RF using a 70/30 stratified train-test split. Next, hyperparameter tuning was performed on a 100,000-observation subset using **RandomizedSearchCV**, then retrained and evaluated on the full dataset using 5-Fold stratified CV. The final configuration used 287 estimators, max depth of 20, and log2 as the max feature criterion.





.

To support feature selection and transparency, we applied permutation importance and SHAP (SHapley Additive exPlanations) to the final calibrated model. SHAP identified **Term**, **SBA_Coverage_Ratio**, and **NAICS_Sector_62** as top features aligning with EBA (2023) guidance on IRB model auditability, and the BIS (2024) call for transparent AI in high-risk financial applications.

Calibrated RF

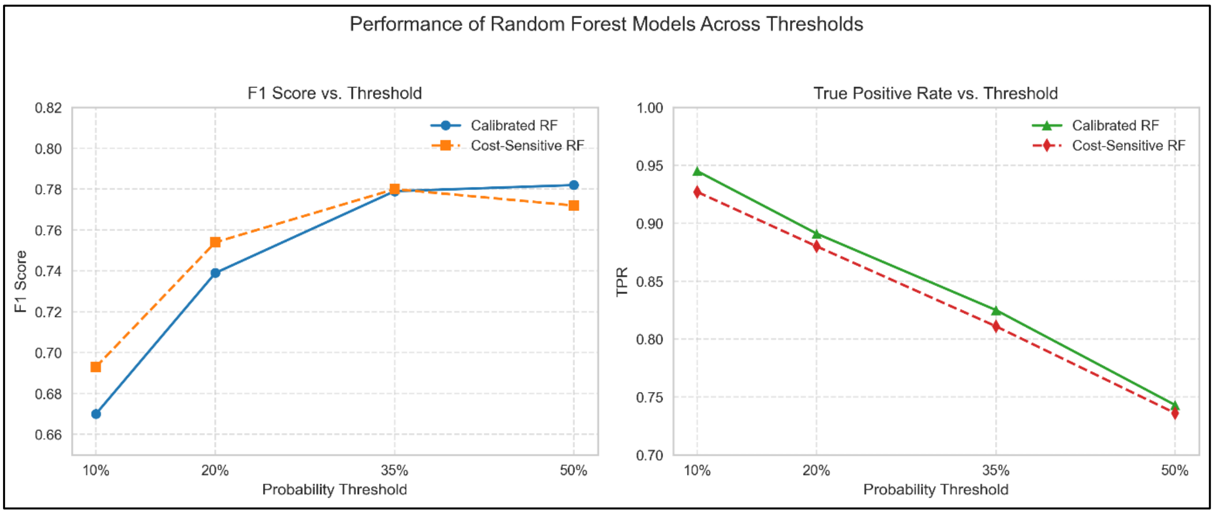
Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score
0.1	94.50%	18.70%	81.30%	5.50%	83.60%	51.80%	67.00%
0.2	89.10%	11.10%	88.90%	10.90%	88.90%	63.10%	73.90%
0.35	82.50%	6.20%	93.80%	17.50%	91.80%	73.80%	77.90%
0.5	74.30%	3.40%	96.60%	25.70%	92.70%	82.50%	78.20%

As RFs do not produce calibrated probabilities by default, we applied Platt scaling via **CalibratedClassifierCV** using 5-fold cross-validation. This yielded a probability-aligned model better suited to threshold-based lending decisions. The 20% threshold also aligns with real-world decision thresholds in lending contexts (Grant et al., 2019; Janssens & Martens, 2020).

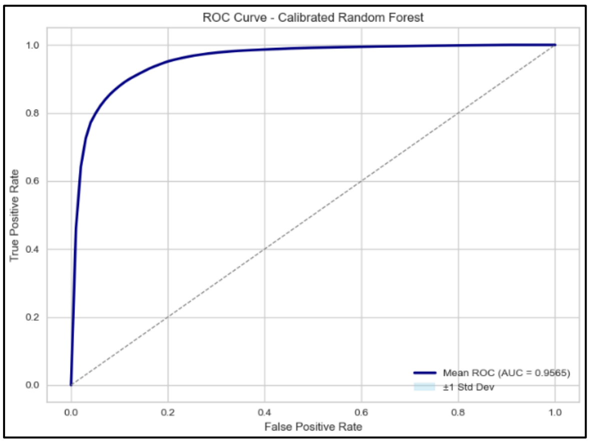
Cost-Sensitive RF

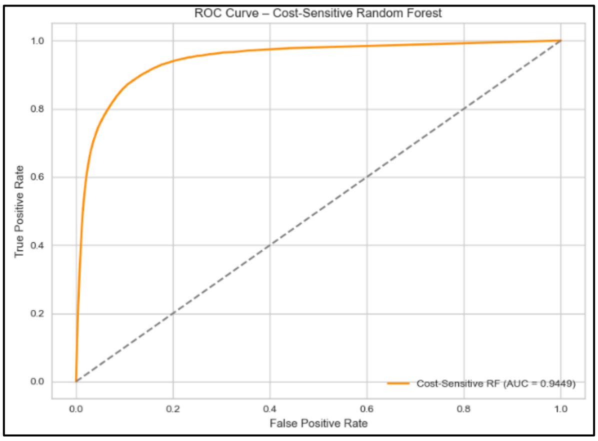
A cost-sensitive variant (**class_weight={0:1, 1:5}**) slightly boosted recall but hurt precision and F-Score.

Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score
0.1	92.70%	16.00%	84.00%	7.30%	85.60%	55.30%	71.60%
0.2	88.00%	9.70%	90.30%	12.00%	89.90%	65.90%	76.20%
0.35	81.10%	5.80%	94.20%	18.90%	91.90%	75.00%	78.00%
0.5	73.60%	3.70%	96.30%	26.40%	92.30%	80.90%	77.20%

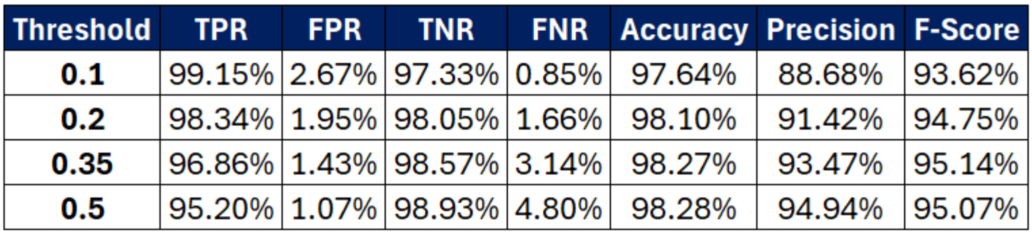


As seen below, Calibrated RF (AUC=0.9565) slightly outperformed the cost-sensitive model (AUC=0.9449).





Minority RF



Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score
0.1	99.15%	2.67%	97.33%	0.85%	97.64%	88.68%	93.62%
0.2	98.34%	1.95%	98.05%	1.66%	98.10%	91.42%	94.75%
0.35	96.86%	1.43%	98.57%	3.14%	98.27%	93.47%	95.14%
0.5	95.20%	1.07%	98.93%	4.80%	98.28%	94.94%	95.07%

To assess the ethical sensitivity of our model, we reintroduced the **Minority** variable, originally excluded due to regulatory and fairness concerns. Whilst this improved performance (AUC > 0.99), it posed ethical risks and violated principles outlined in the European Banking Federation (2023) and the U.S. Equal Credit Opportunity Act. This sensitivity test reinforced the trade-off between raw performance and fairness. FATF (2021) similarly cautions against bias amplification in opaque risk models.

We recommend deploying the calibrated RF at a 20% threshold. It combines strong performance, calibrated outputs, and SHAP-based interpretability, complying with EBA, BIS, and EBF standards for governance and fairness in AI credit scoring, echoing Agarwal et al. (2023).

3.3 SVM

We evaluated linear and RBF SVMs, with/without calibration tuned via **GridSearchCV** beside scaling and **class_weight='balanced'.** Polynomial kernels were dropped due to high runtime and marginal gains.

Kernel Type	Best Parameters	Best AUC
Linear	C=10	0.8165
RBF	C=10, gamma='scale'	0.8444
Polynomial	C=1, degree=3, coef0=1.0, gamma='scale'	0.8416

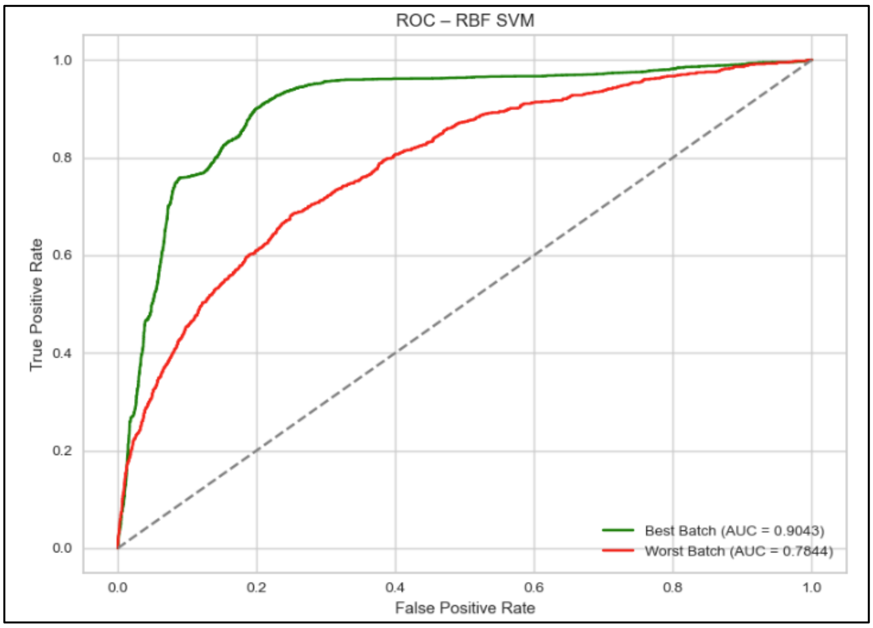
Calibrated SVMs, implemented using Platt scaling via **CalibratedClassifierCV,** reduced FPR to 3.7% but at the cost of high FNR, which grew above 60%. In credit lending contexts, where undetected defaults pose significant risk, this trade-off proved limiting.

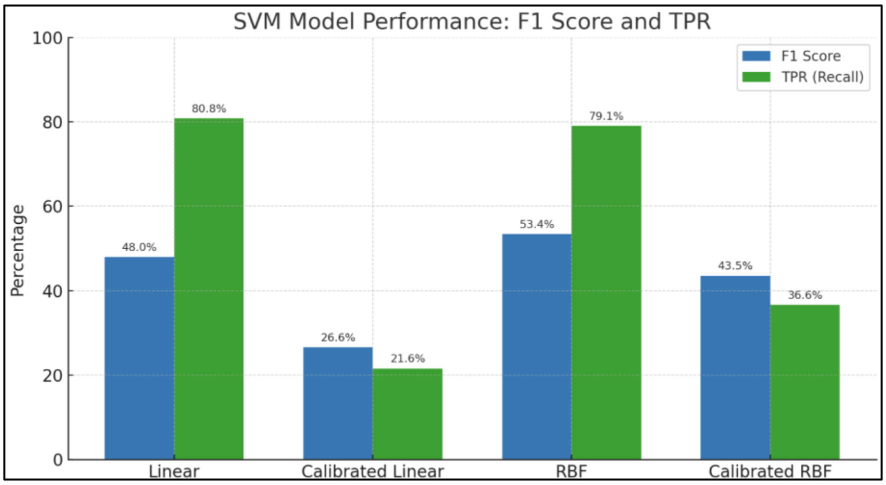
Model	Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score	AUC
Linear Kernel	0.5	80.80%	31.01%	68.99%	19.20%	70.85%	36.45%	48%	82.00%
Calibrated Linear Kernel	0.5	21.55%	3.69%	96.31%	78.45%	85.40%	44.22%	26.61%	82.05%
RBF Kernel	0.5	79.07%	21.79%	78.21%	20.93%	78.77%	42.00%	53.41%	85.40%
Calibrated RBF Kernel	0.5	36.61%	4.55%	95.45%	63.39%	87.35%	65.34%	43.52%	85.42%

The best model, uncalibrated RBF SVM, achieved TPR 79.1%, F1 53.4%, AUC 85.4%. Despite lacking calibration, its margins and rank-ordering made it scalable and competitive. For computational efficiency, all SVM variants were initially assessed at a 0.5 threshold. Based on trends in prior models, we re-evaluated the uncalibrated RBF SVM at a 0.2 threshold to prioritise sensitivity to defaulters.

Model	Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score	AUC
RBF Kernel	0.2	83.47%	16.57%	83.43%	16.53%	83.47%	47.50%	56.49%	85.40%

However, non-linear SVMs lack native interpretability. Due to computational limits, we couldn't apply post-hoc tools like SHAP, limiting explainability, a requirement per EBA (2023) and BIS (2024).





McKinsey (2024) underscores explainability as critical to responsible AI. Overall, the uncalibrated RBF SVM offers strong default prediction but needs more transparency safeguards underscored by FATF (2021) highlighting fairness-performance trade-offs in financial AI. Though less interpretable than RFs, it's a valuable part of an ethically governed ML toolkit.

3.4 Logistic Regression Comparison

Our final recommendation balances three objectives: (1) minimising undetected defaulters, (2) avoiding over-flagging safe applicants, and (3) preserving model interpretability and reliability.

	Threshold	TPR	FPR	TNR	FNR	Accuracy	Precision	F-score	AUC
Logistic Regression	0.2	76.22%	23.94%	76.06%	23.78%	76.09%	40.43%	52.83%	82.00%
KNN	0.2	84.94%	23.66%	76.34%	15.06%	77.85%	43.37%	57.41%	86.12%
SVM	0.2	83.47%	16.57%	83.43%	16.53%	83.47%	47.50%	56.49%	85.40%
Random Forest	0.2	89.10%	11.10%	88.90%	10.90%	88.90%	63.10%	73.90%	95.65%

The calibrated RF consistently outperformed alternatives, confirming prior findings (Wu, 2022; Bhargav & Malathi, 2023) on RFs in credit risk modelling.

While 5-NN had strong recall, its low precision and high FPR led to excessive rejections. SVMs offered balance but lacked transparency. Logistic regression underperformed due to its linearity and missed complex patterns. A 20% threshold yielded the best trade-off for RF, offering high accuracy and low misclassification costs.

We recommend deploying the calibrated RF at a 20% threshold. It combines predictive strength, fairness, and regulatory compliance, making it the most practical option. Not just selected for its superior accuracy and F-score, but also for its alignment with real-world requirements. It supports SHAP-based explainability, calibrated probabilities for threshold tuning, and meets governance standards outlined by the EBA (2023). Its high sensitivity at 20% substantially reduces undetected defaults, aligning with the commercial objective of minimising credit losses.

4. Unsupervised Analysis

4.1 Unsupervised Learning

Unsupervised learning uncovers hidden patterns and structure within unlabelled datasets. We applied PCA, K-Means Clustering (KMC), and Hierarchical Clustering (HC) to support our predictive modelling. While not predictive, these techniques enhance model understanding and subgroup segmentation (EBA, 2023).

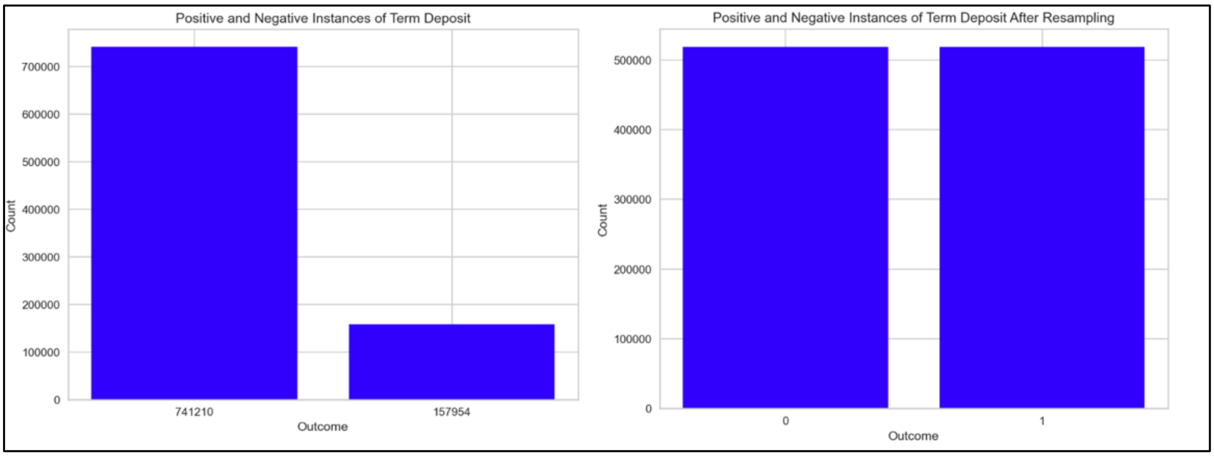
PCA reduces dimensionality by transforming correlated features into uncorrelated principal components. These components retain crucial variance whilst reducing the feature space.

KMC partitions data into K distinct, non-overlapping clusters by minimising within-cluster sum-of-squares. KMC uses the squared Euclidean distance to assign points to the nearest centroid before iteratively updating centroids and cluster assignments until convergence. Choosing a suitable K is imperative for meaningful interpretation.

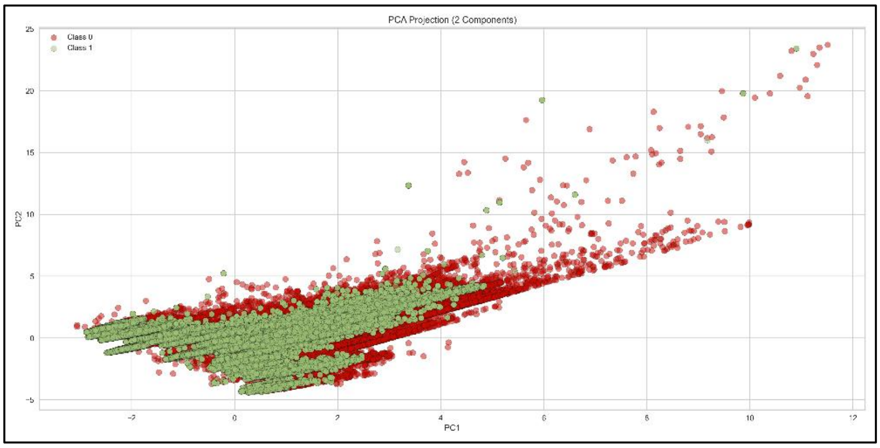
HC forms a tree-like structure of nested clusters visually illustrating how data points cluster. Due to its configuration and visual representation, this method is useful for analysing the underlying structure of the dataset and validates cluster patterns identified by KMC.

A key limitation is the absence of labels for direct cluster interpretation. However, these methods as a complementary step can identify latent structure and inform downstream tasks such as risk stratification and resource allocation. (EBA, 2023)

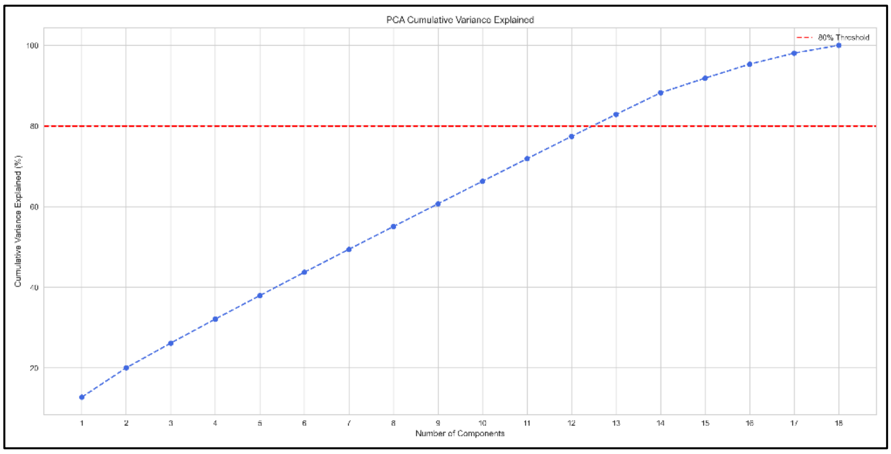
4.2 PCA



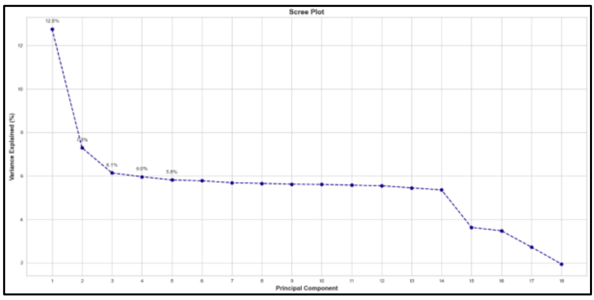
To reduce class imbalance, random oversampling was synthetically applied. The rebalanced dataset was then standardised and transformed using PCA to improve model-performance whilst retaining maximum variance.

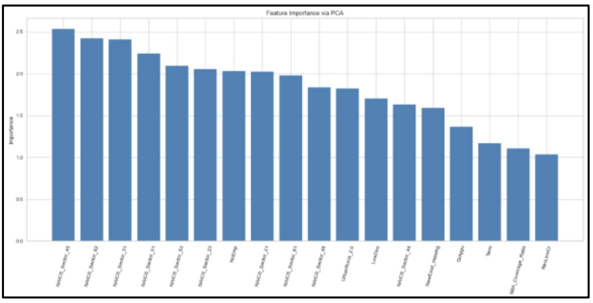


The first two PCs exhibit a dense linear structure, indicating high correlation among original features and discrete class separation. Class-distinction is more visible at higher values suggesting PCA captured variance relevant to default behaviour. (EBA, 2023)

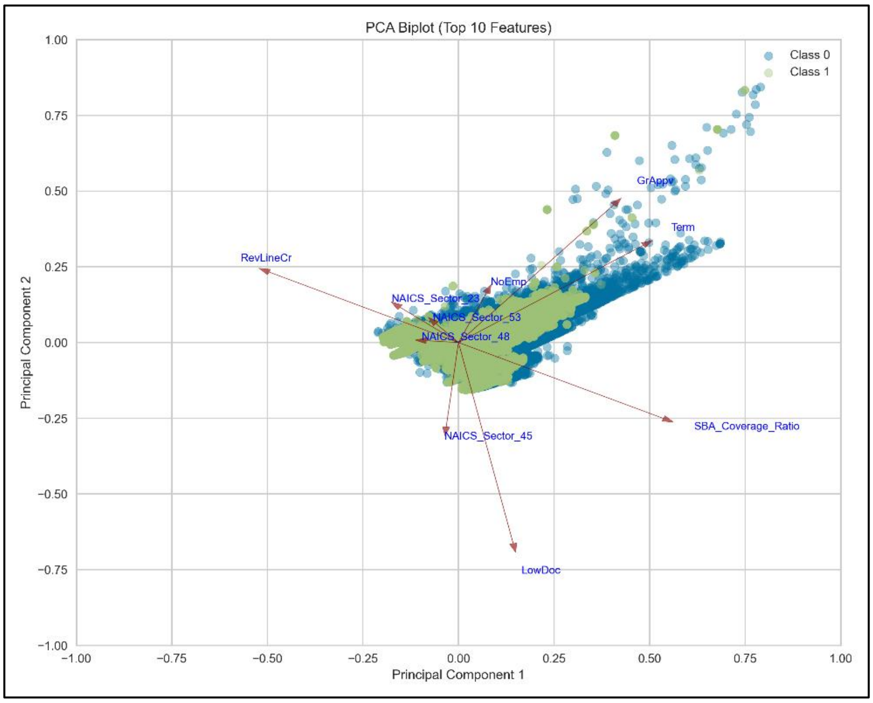


The cumulative variance graph displays the incremental variance explained by 18 PCA components. The red line represents the threshold for dimensionality at 13 components explaining 80% of the cumulative variance.



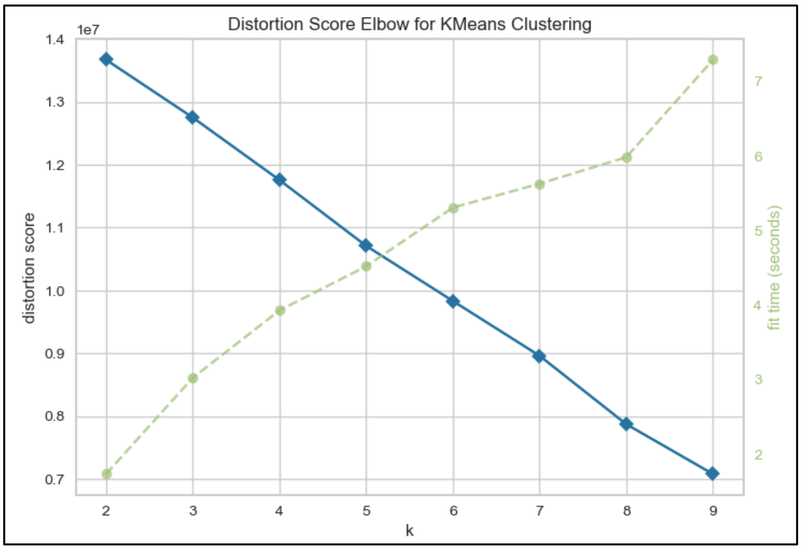


The Scree Plot shows an elbow at component 14, with the first two PCs explaining most variance. Retail trade, healthcare, and manufacturing industries dominated feature importance indicating sector-specific default risk patterns (EBF, 2019). Other variables like loan amount, number of employees and urban/rural classification also influenced clustering and classification tasks. This analysis highlights which real-world factors contribute to the dataset variance, making the reduced feature set efficient and interpretable.

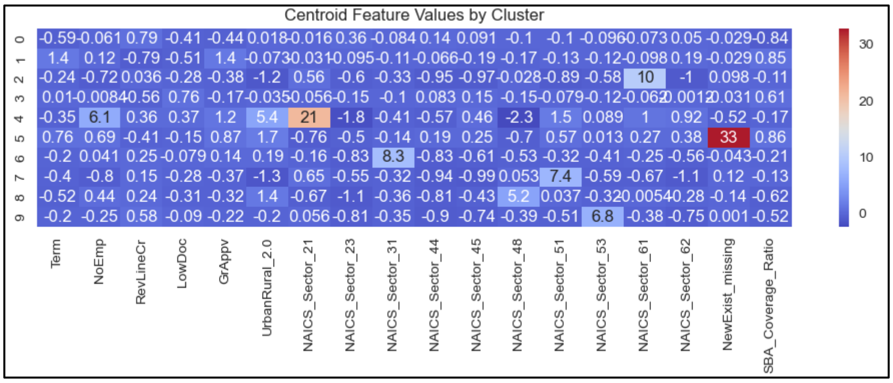


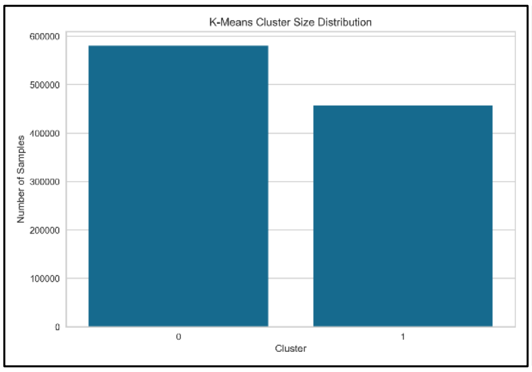
The PCA Biplot represents how the features influence the data's projection onto the first two PCs. Many sector features cluster near the origin suggesting their influence is more evenly distributed across multiple components, reiterating that specific sectors and operational loan features are key to understanding variance patterns in the data and distinguishing default risk.

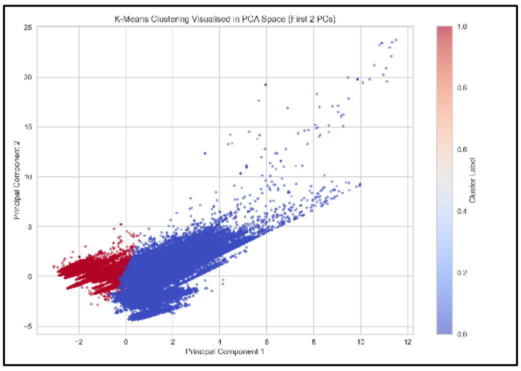
4.3 K-Means Clustering



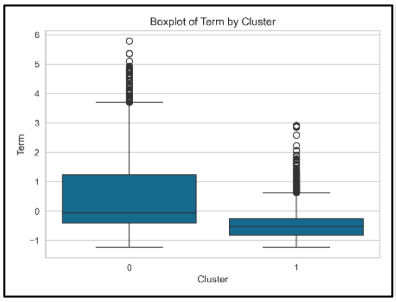
KMC with K=2 was applied to the PCA-transformed data to uncover natural groupings relating to loan default behaviour. Gains diminish after K=4 (Elbow Method, Distortion Score), suggesting 2–5 clusters are meaningful. However, the Silhouette Score displays a peak at K=9, suggesting that the data might support more complex partitioning. For supervised learning with a binary target, K=2 remains an efficient choice.

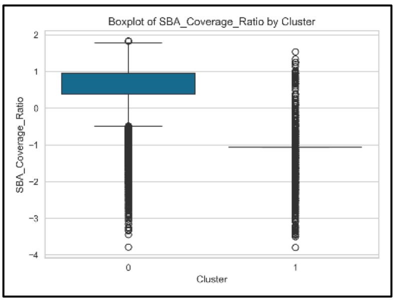


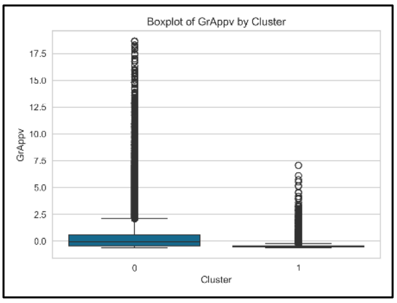




Crosstab comparison indicated meaningful segmentation, with one cluster having a high default rate. The scatter plot and centroid heatmap highlighted operational loan characteristics such as **GrAppv, SBA_Coverage_Ratio** and **Term** as central indicators of separation.





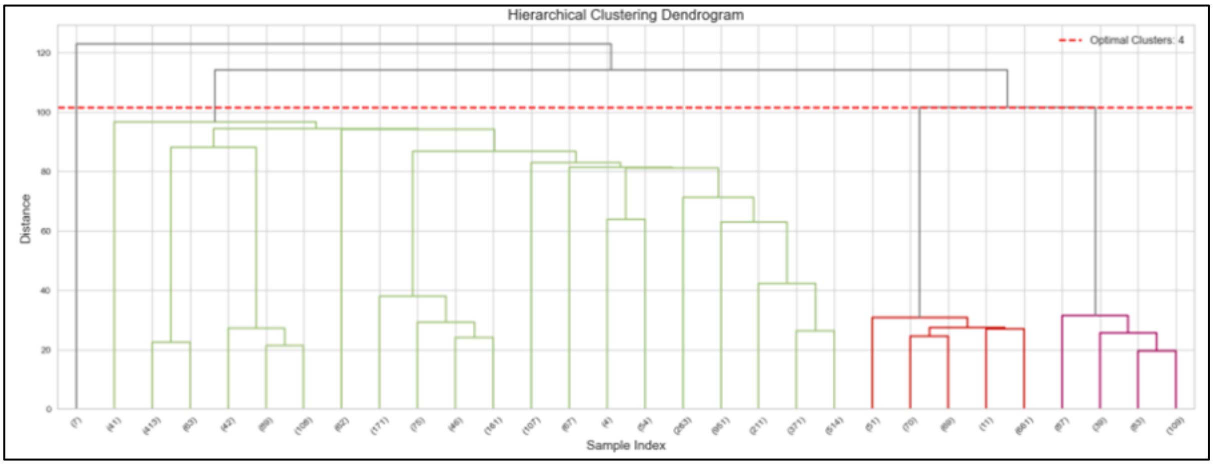


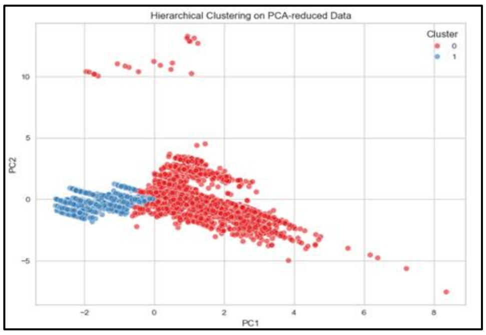
The cluster labels with default outcomes yielded 61.6% accuracy suggesting the previous feature-based patterns, (SBA backing and term) align with default behaviour.

Applying PCs to Supervised Models (K = 2)							
	TPR	FPR	TNR	FNR	Accuracy	Precision	F-Score
Logistic Regression	65.30%	35.90%	64.10%	34.70%	64.30%	27.90%	39.10%
KNN	71.90%	11.70%	88.30%	28.10%	85.50%	56.80%	63.50%
Random Forest	70.60%	6.60%	93.40%	29.40%	89.40%	69.40%	70.00%
SVM	60.80%	33.60%	66.40%	39.20%	65.40%	27.80%	38.20%

The RF and KNN models produce high accuracy rates whilst also producing a superior precision rating and F-Score. Compared to original supervised model trained on true default labels, the unsupervised model displayed worse performance, but the meaningful patterns depicted highlight the value of PCA and clustering for preprocessing and exploratory analysis. Models trained on actual default outcomes like calibrated RFs remain superior for operational prediction.

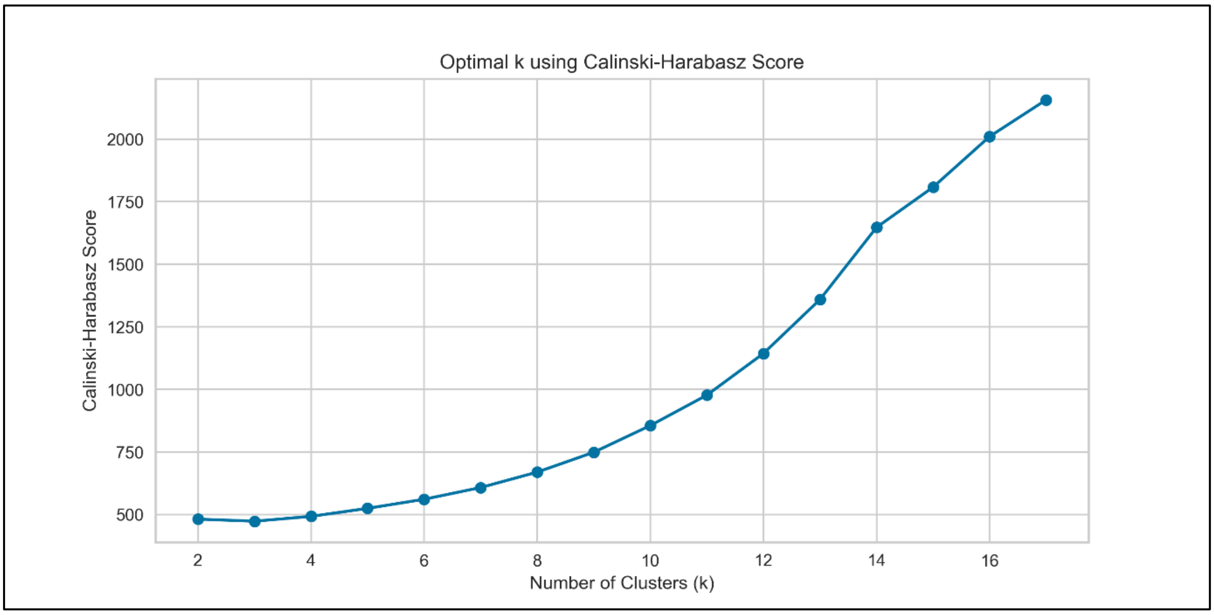
4.4 Hierarchical Clustering

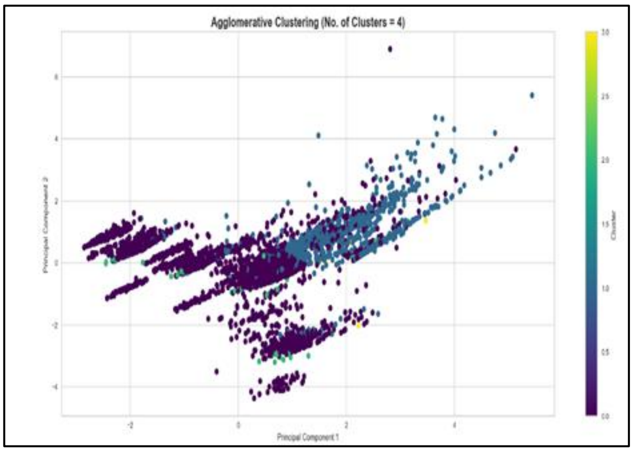


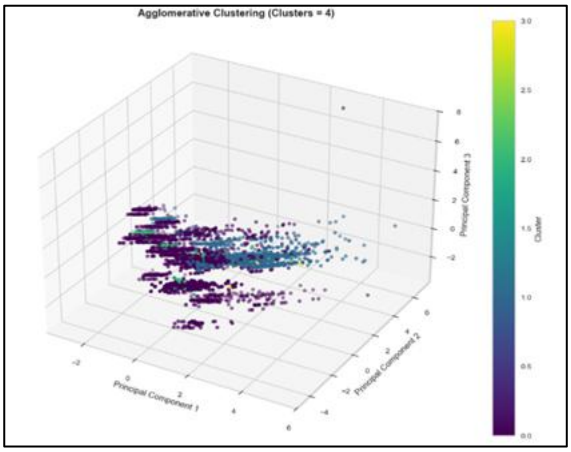


10%		Actual	
Predicted		Default	No Default
	Default	2,754	1,627
	No Default	2,246	3,373

Visual and quantitative assessments of HC indicate inability to separate default from non-default borrowers. HC skewed toward one dominant cluster, failing to capture borrower heterogeneity. The 50% accuracy of the confusion matrix struggles to align with actual risk categories. The unbalanced and overlapping dendrogram reduces practical value.







The Calinski-Harabasz score shows increasing clusters improves cohesion and separation, but the considerable overlap and high imbalance reinforce HC's struggles to uncover risk structure.

5. Conclusion

Theoretical strengths in dimensionality reduction and uncovering structure were partially realised through PCA and KMC, whilst HC offered limited utility. PCA highlighted key sectoral patterns, KMC moderately aligned with default behaviour whereas HC lacked clear separation between defaulters and non-defaulters. Unsupervised methods added exploratory insight and enriched data understanding but provided only modest gains in predictive accuracy. Integration into predictive modelling should be pursued cautiously and aligned with regulatory guidance (EBA, 2023; EBF, 2019)

6. Limitations

Despite strong results, several limitations should be considered prior to deployment:

Category	Limitation	Implication
Overfitting Risk	Historic data may not generalise to future lending patterns	Needs retraining under new economic conditions
Data & Imbalance	Missing values; defaults only ~11%	Impacts reliability; requires threshold calibration
Fairness & Compliance	Protected vars removed, but proxy bias may remain	Requires SHAP auditing and ongoing fairness checks
Interpretability	RF is non-linear and opaque	SHAP improves transparency; EBA/BIS governance still needed
Model Drift	Lending context may shift over time	Ongoing monitoring and validation required

7. References

Bhargav, P. & Malathi, K. (2023). *Using Machine Learning, the Random Forest Algorithm and Logistic Regression to Predict Default Loan Approval*. Journal of Fisheries Sciences.

EBA (2023). *Follow-up report on machine learning for IRB models*.

EBF (2019). *Artificial Intelligence in the Banking Sector*.

EU Agency for Fundamental Rights (2022). *Bias in Algorithms – Artificial Intelligence and Discrimination*.

Grant, S.W., Hickey, G.L., & Head, S.J. (2018). *Multivariable regression considerations and pitfalls*. European Journal of Cardio-Thoracic Surgery.

Janssens, A.C.J.W., & Martens, F.K. (2020). *Revisiting the AUC in predictive modelling*. International Journal of Epidemiology, 49(4), 1397–1403.

Johnson, J.M., & Khoshgoftaar, T.M. (2019). *Survey on deep learning with class imbalance*. Journal of Big Data, 6(1).

Wu, W. (2022). *Machine Learning Approaches to Predict Loan Default*. Open Journal of Statistics.

Agarwal, S., Muckley, C.B., & Neelakantan, P. (2023). Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods. *Economics Letters, 226*, 111117.