UCD Michael Smurfit Graduate Business School

Module Title: FIN42100 Machine Learning in Finance

Assignment Title: Predicting Loan Default Using a Machine Learning Algorithm

Word count : 3800

Group 4 Members:

Serial Number	Student name	Student ID
1	Shagun Chandok	24289312
2	Nilay Singh	24289944
3	Dhruv Singh	24234646
4	Aditya Suhane	24212188
5	Tejasvi Patil	24209396

Table of Contents

1. Introduction and Context

The U.S. Small Business Administration (SBA) mitigates lending risks by guaranteeing small business loans. Accurate default prediction is crucial for minimizing credit losses. This study applies supervised machine learning (Decision Tree, Random Forest, XGBoost) to an SBA dataset (621,347 cleaned observations, 11 features, 1962–2014). Additionally, we evaluate three unsupervised techniques (PCA, K-Means, Hierarchical Clustering) to uncover latent patterns, reduce dimensionality, and enhance the predictive accuracy of the bank's lending decisions.

2. Q1:Exploratory Data Analysis: Insights for Lending Decision-Making

1.Dataset Overview & Variable Categorization

We conducted exploratory data analysis (EDA) on 899,071 SBA loan records (1962–2014) to assess loan default risk (MIS_Status: 0 = default, 1 = non-default). The dataset was comprehensive, having no missing values in critical variables, except six non-predictive entries in ChgOffDate, facilitating reliable preprocessing and model development. Variables were categorized into three groups based on their characteristics: Numerical Continuous (19), including features like DisbursementGross and Term, necessitating scaling; Numerical Discrete (8), such as NewExist and Industry, requiring encoding; and Time Categorical (3), comprising ApprovalDate, DisbursementDate, and ChgOffDate, potentially needing feature engineering.

Here is the table of dataset variables detailing each column's type, unique values, and description to support data understanding and feature selection.:

📊 Variable Summary Table

Variable	Type	Unique Values	Details
LoanNr_ChkDgt	Numerical Continuous	899,164	Unique loan identifier, highly variable, not pre
Name	Numerical Continuous	779,583	Business name identifier, high variability, not
City	Numerical Continuous	32,581	City of business, diverse geographic indicator.
State	Numerical Continuous	51	State of business, reflects regional economic c
Zip	Numerical Continuous	33,611	ZIP code, granular geographic indicator.
Bank	Numerical Continuous	5,802	Lending bank, diverse institutions, potential ris
BankState	Numerical Continuous	56	State of bank, indicates lending geography.
NAICS	Numerical Continuous	1,312	Industry classification codes, highly variable.
ApprovalFY	Numerical Continuous	51	Fiscal year of approval, temporal context for lo
Term	Numerical Continuous	412	Loan duration (months), wide range, key risk p
NoEmp	Numerical Continuous	599	Number of employees, reflects business size.
CreateJob	Numerical Continuous	246	Jobs created, variable economic impact indicat
RetainedJob	Numerical Continuous	358	Jobs retained, similar to CreateJob.
FranchiseCode	Numerical Continuous	2,768	Franchise identifier, indicates business type.
DisbursementGross	Numerical Continuous	118,859	Loan amount disbursed, critical financial risk f
ChgOffPrinGr	Numerical Continuous	83,165	Charged-off amount, continuous loss indicator
GrAppv	Numerical Continuous	22,128	Gross approved loan amount, financial risk fac
SBA_Appv	Numerical Continuous	38,326	SBA-approved loan amount, reflects guarantee
DaysToDisbursement	Numerical Continuous	3,673	Time to disbursement, potential risk indicator
NewExist	Numerical Discrete	3	[0, 1, 2] (undefined, existing, new), business t
UrbanRural	Numerical Discrete	3	[0, 1, 2] (undefined, urban, rural), location-bas
RevLineCr	Numerical Discrete	18	[0, 1, ..., 17], revolving credit, requires clean-n
LowDoc	Numerical Discrete	8	[0, 1, ..., 7], low-documentation loan, requires
BalanceGross	Numerical Discrete	15	[0, 600, ..., 996,262], outstanding balance, lim
Industry	Numerical Discrete	25	[0, 11, ..., 92], derived from NAICS, sector-spe
Minority	Numerical Discrete	2	[0, 1], binary indicator for minority-owned bus
MIS_Status	Numerical Discrete	2	[0, 1], target variable (default, non-default), bi
ApprovalDate	Time Categorical	9,859	Date of loan approval, temporal context for ris
ChgOffDate	Time Categorical	6,448	Date of charge-off, relevant for default timing
DisbursementDate	Time Categorical	8,472	Date of disbursement, indicates loan processin

Table 1: Variables categorization.

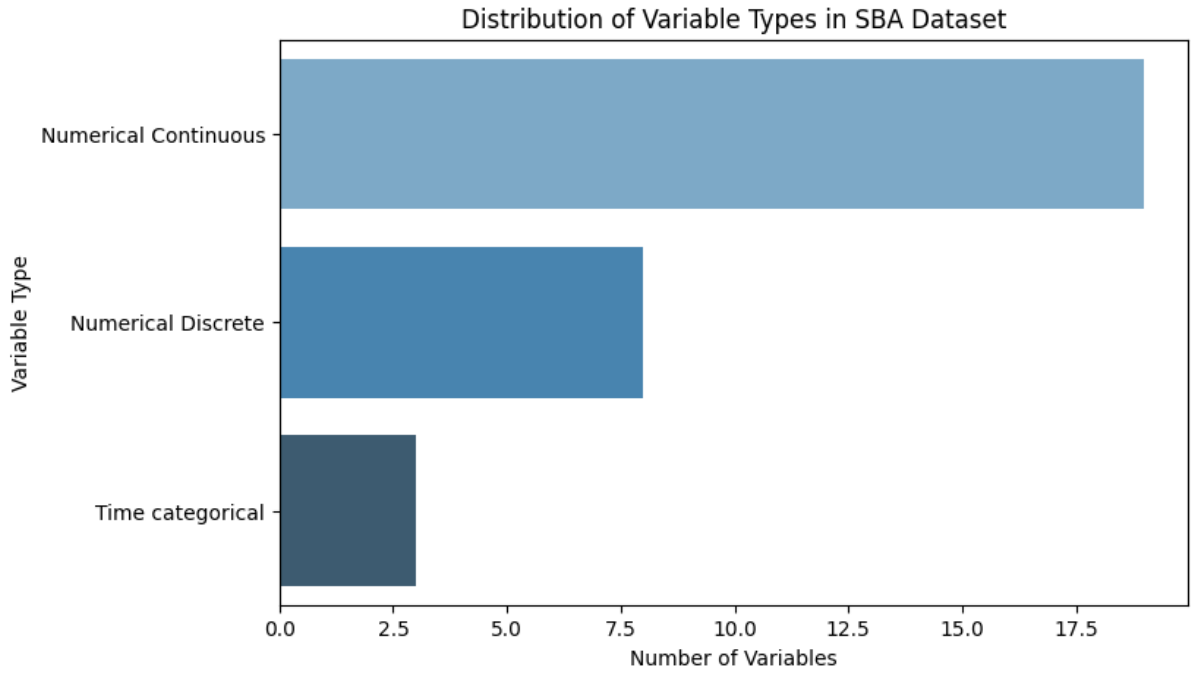

Figure 1: Distribution of variable Types of SBA dataset

2.Correlation and Distribution Analysis

Correlation analysis identified Term (0.26), SBA_Appv (0.10), and GrAppv (0.10) as top predictors of non-default (MIS_Status = 1), implying longer terms and higher approvals lower default risk. Box plots confirmed non-defaulted loans typically have longer terms (~80 vs. ~40 months) and higher amounts. Due to skewness, log-transformation was necessary (Figure 2).

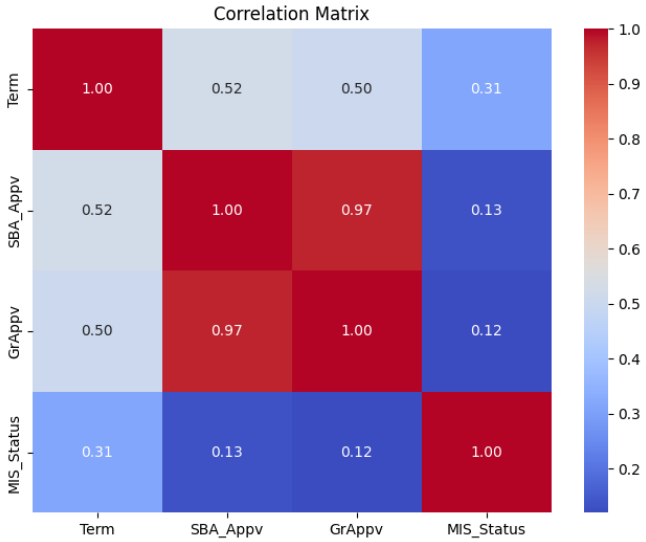

Figure 2: Correlation Matrix

3.Categorical Associations

Chi-squared tests identified significant associations ($p < 0.001$):

- Industry (Chi2 = 35,543): Construction (NAICS 23) has higher default rates (~25%).
- RevLineCr (Chi2 = 21,346): Revolving credit loans are riskier.
- NewExist (Chi2 = 8,346): New businesses default more (~25% vs. ~15% for existing).

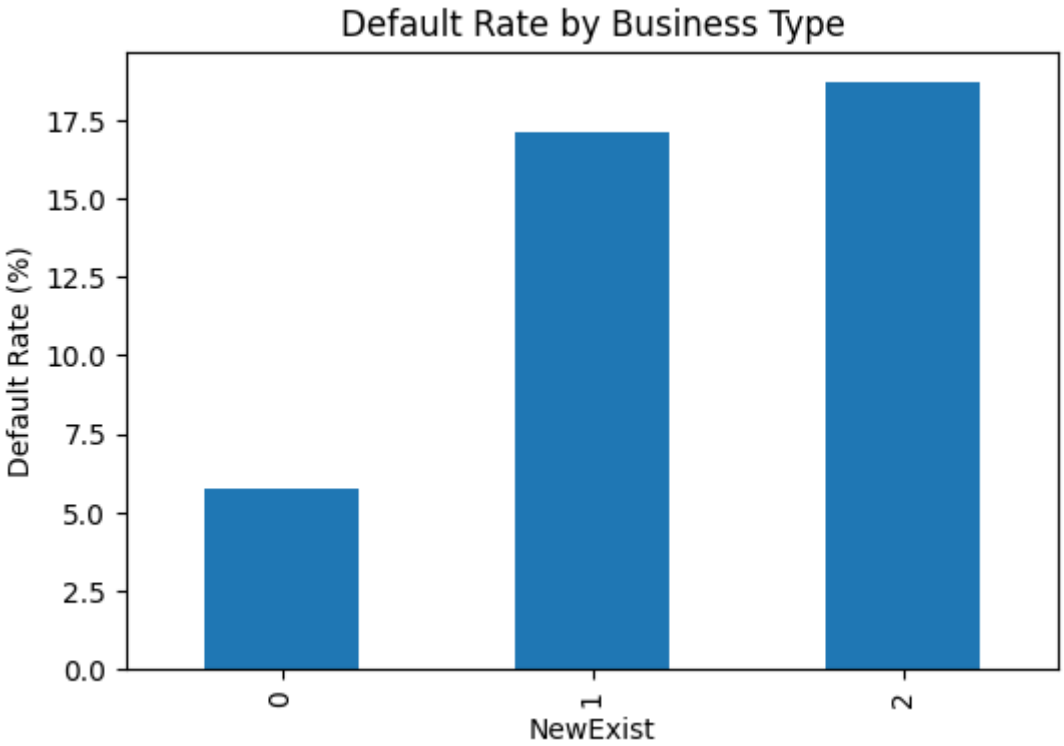

Figure 3: Default Rate by Business Type

4.Feature Importance

A Random Forest model highlighted Term (~0.35 importance), DaysToDisbursement (~0.15), and DisbursementGross (~0.12) as top predictors (Figure 4), emphasizing financial and temporal factors.

Top 10 features (Random Forest quick check):

Serial	Feature	Importance
0	Minority	0.453746
1	Term	0.249401
2	ApprovalFY	0.053888
3	SBA_Appv	0.023515
4	DaysToDisbursement	0.022289
5	Bank	0.019642
6	GrAppv	0.017085
7	FranchiseCode	0.016252
8	DisbursementGross	0.016141
9	Zip	0.015196

Table 2:Top 10 features

5.Treatment of the 'Minority' Variable:

To comply with fair lending regulations and ethical standards, the Minority variable was excluded from the final model. Although it showed significant predictive power in preliminary analysis, incorporating demographic attributes in credit decisions risks regulatory violations (e.g., Equal Credit Opportunity Act) and may result in unintended bias against protected groups.

6. Rationale:

Regulatory Compliance: Using protected attributes like race in financial models is restricted in many jurisdictions. Including Minority in production risks non-compliance.

Ethical Considerations: Even if predictive, demographic variables can be seen as discriminatory and damage institutional trust.

Fairness Monitoring: Minority is retained solely for post-hoc analysis to detect any unintended bias, ensuring the model does not disproportionately impact protected groups.

7. Conclusion:

The Exploratory data analysis (EDA) of the SBA loan dataset identified key factors influencing default risk, notably *Term*, *DisbursementGross*, *SBA_Appv*, *Industry*, and *NewExist*. Longer loan terms and larger SBA-approved amounts correlated with lower default rates, while startups and construction industries exhibited higher defaults. Variables such as *Minority*, *UrbanRural*, *NoEmp*, and *BalanceGross* demonstrated limited predictive value and can be excluded for efficiency. These insights support the development of a robust, interpretable predictive model for loan defaults.

3. Q2: Logistic Regression Modelling

1. Introduction

We used a logistic regression model for predicting loan defaults using the U.S. Small Business Administration dataset, enhancing risk management for lending institutions.

2. Methodology

A logistic regression model was developed on the SBA dataset (899,164 records, 1962–2014). Categorical variables (*NewExist*, *UrbanRural*, *RevLineCr*, *LowDoc*) were binary-encoded, and numerical features (*DisbursementGross*, *Term*, *NoEmp*, *CreateJob*, *RetainedJob*, *SBA_Appv*, *GrAppv*) were standardized. After handling missing values, 364,318 records remained. The target (*MIS_Status*) was redefined as Default (1 = default, 0 = non-default), with undefined categories mapped to zero. Eleven key features were selected, and model performance was compared using 11 selected versus 13 total variables. Evaluations included a 70-30 train-test split, stratified for defaults, at various thresholds (0.1, 0.2, 0.35, 0.5), with ROC AUC assessed through 10-fold stratified cross-validation. Feature importance was determined by grouped dummy variable coefficients.

3. Results

a. Full Dataset Logistic Regression

Table 1 shows performance on the full dataset (364,318 records). Threshold 0.2 balances TPR (0.875) and FPR (0.359) for default detection. The confusion matrix (Figure 1) and feature importance (Figure 2) highlight Term as the top predictor, followed by NoEmp and RevLineCr.

Threshold	TPR	FPR	Accuracy	Precision	F1 Score	Observations
0.1	0.963	0.737	0.448	0.319	0.479	High default detection, excessive false positives
0.2	0.875	0.359	0.703	0.466	0.608	Balanced, recommended for lending
0.35	0.658	0.171	0.784	0.579	0.616	Conservative, prioritizes precision
0.5	0.395	0.076	0.784	0.649	0.491	Overly conservative, misses many defaults

Table 1: Logistic Regression Performance on Full Dataset

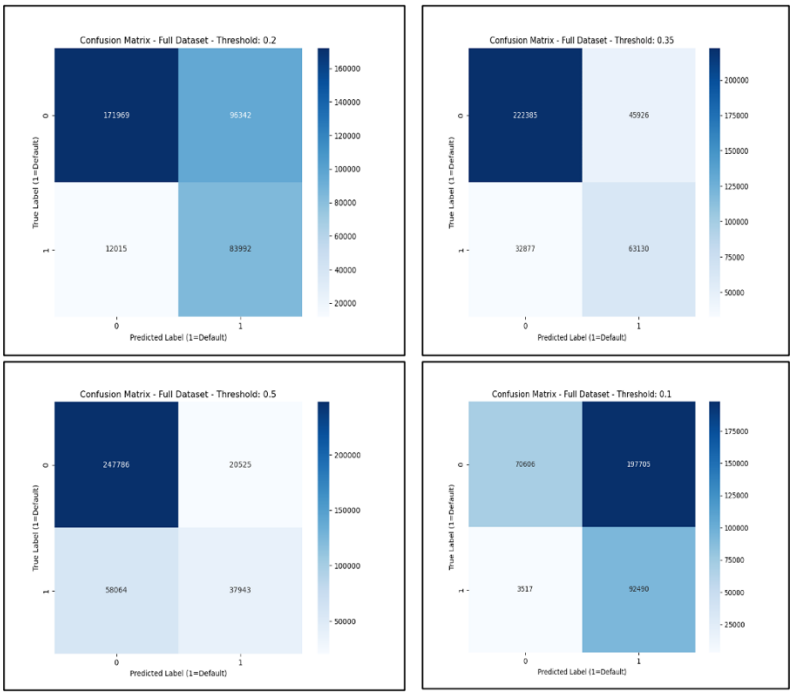

Figure 1: Confusion Matrix for Logistic Regression on Full Dataset at All Thresholds

Aggregates dummy variable coefficients, emphasizing Term, NoEmp, and RevLineCr.

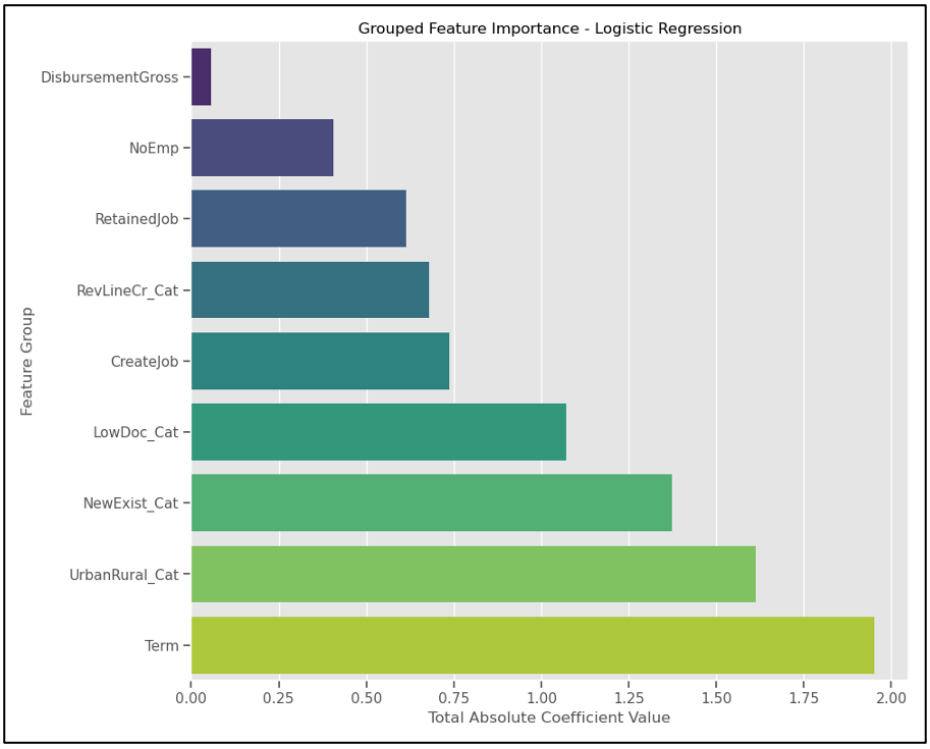

Figure 2: Feature Importance for Selected Variables

Feature Importance (Based on 70-30 Split Model)

Feature	Importance
Term	2.135377
RevLineCr_1	0.385829
NoEmp	0.377195
GrAppv	0.246439
SBA_Appv	0.198784
UrbanRural_1	0.18013
RetainedJob	0.082165
DisbursementGross	0.060384
CreateJob	0.05619
LowDoc_1	0.04409
NewExist_1.0	0.012899

Table 2: Feature Importance

Comparison of Selected vs. All Variables

The 11 selected variables achieved an AUC of 0.829, TPR of 0.878, and FPR of 0.360 at threshold 0.2 (Figure 1). Using all 13 variables slightly increases AUC (~0.83–0.85) (Figure 3) but overemphasizes UrbanRural_Cat (1.61 vs. 0.18, Figure 4), risking overfitting. Selected variables balance performance and computational efficiency.

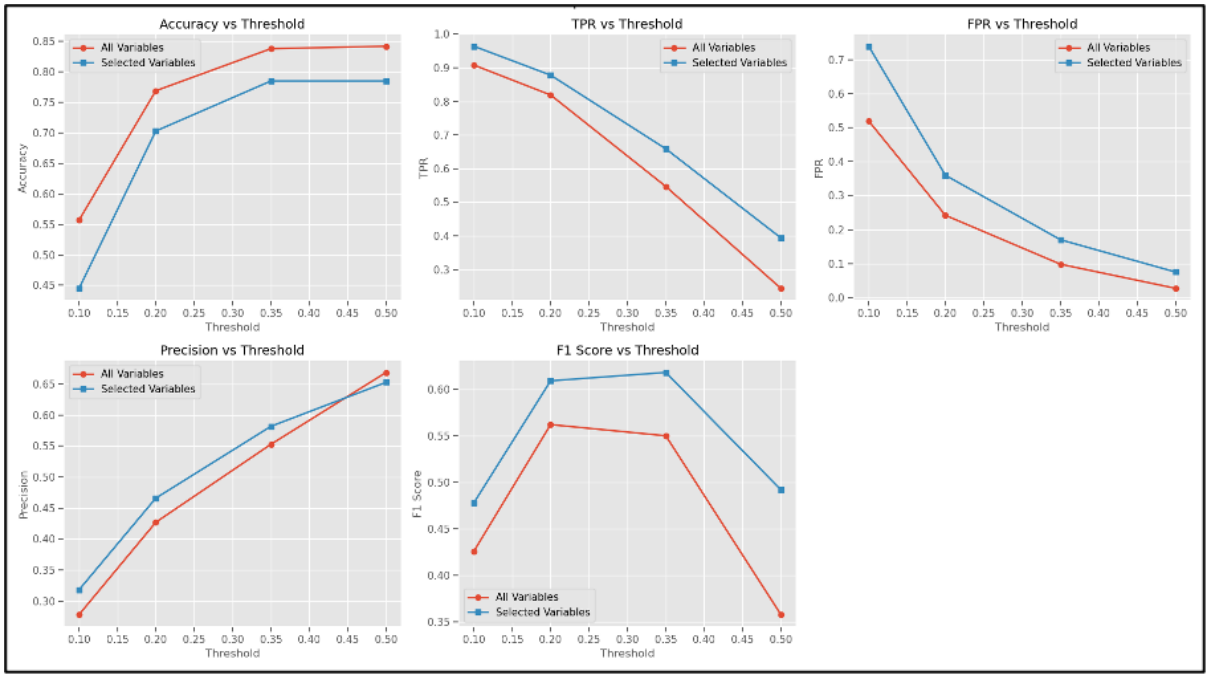

Figure 3: Performance Comparison of Selected vs. All Variables

Compares grouped feature importance, highlighting Term dominance and higher UrbanRural_Cat influence in the all-variables model.

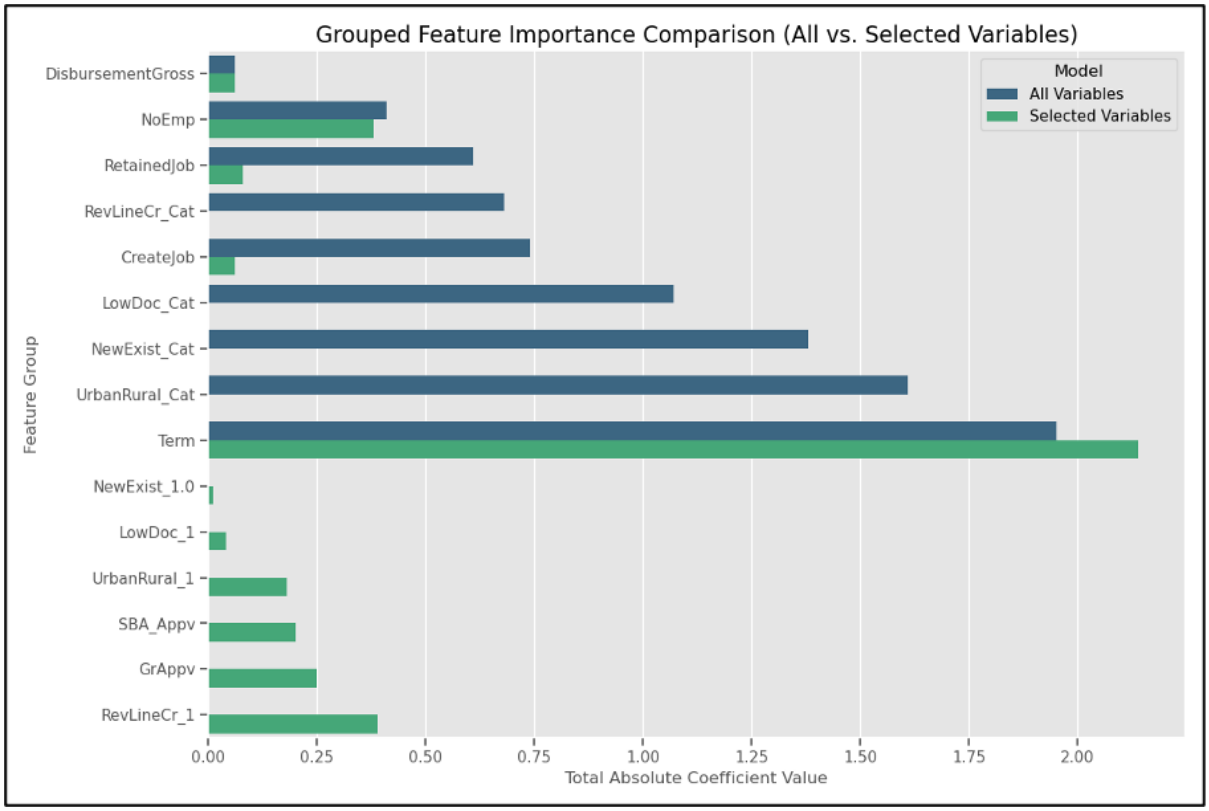

Figure 4: Feature Importance Comparison (Selected vs. All Variables)

b. Train-Test Split & Cross-Validation

The 70-30 train-test split results (Table 2) show consistent performance, with threshold 0.2 selected for its TPR (0.878) and FPR (0.360). The test set confusion matrix (Figure 5) confirms this balance. Ten-fold cross-validation yielded a mean ROC AUC of 0.829 (±0.02), indicating model stability.

Threshold	TPR	FPR	Accuracy	Precision	F1 Score	Observations
0.1	0.964	0.740	0.445	0.318	0.478	High default detection, excessive false positives
0.2	0.878	0.360	0.703	0.466	0.609	Strong balance, recommended
0.35	0.659	0.170	0.785	0.582	0.618	Low false positives, moderate default detection

Table 3: Table 2: Logistic Regression Performance on Test Set

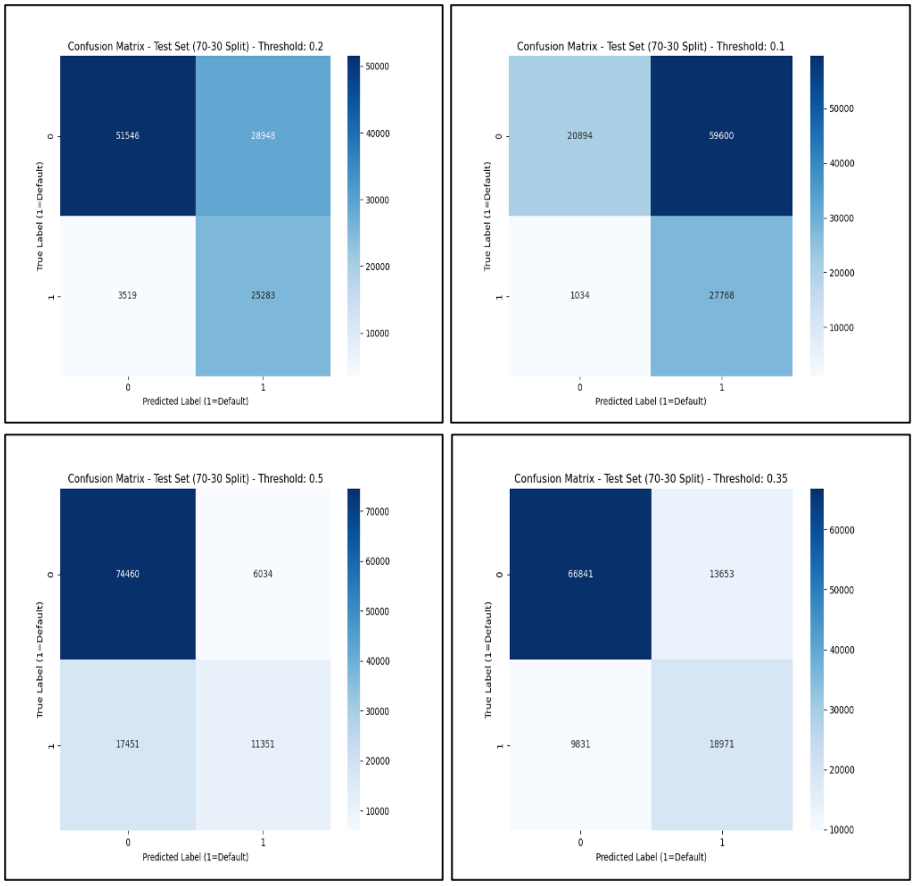

Figure 5: Confusion Matrix for Logistic Regression on Test Set at all Thresholds

c. ROC Curve and AUC

The ROC curve for the test set (*Figure 6)* yields an AUC of 0.829, demonstrating strong discriminative ability, supporting the model's reliability for risk assessment.

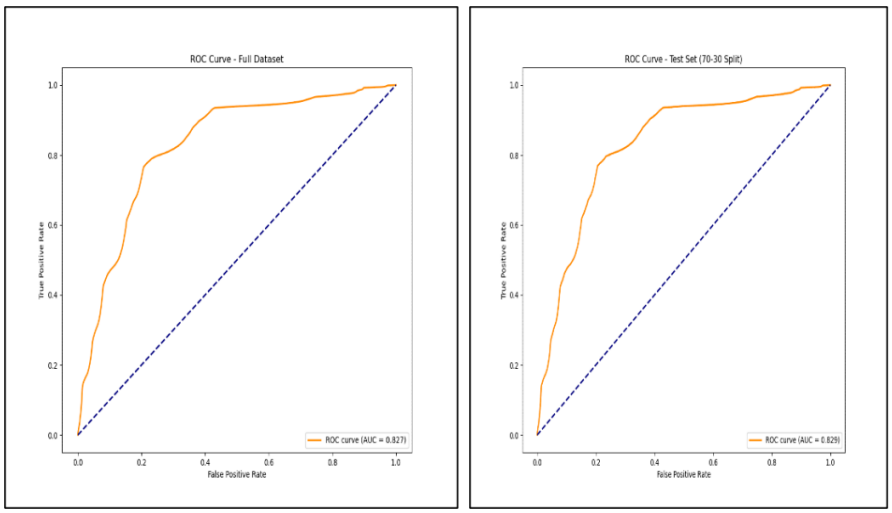

Figure 6: Receiver Operating Characteristic (ROC) Curve for Logistic Regression on Test Set (AUC = 0.829)

4. Conclusion

The logistic regression model predicts loan defaults effectively, with threshold 0.2 recommended for balanced TPR (~0.878) and FPR (~0.360). The AUC of 0.829 confirms robust discrimination. Eleven selected variables enhance interpretability over 13 all variables, reducing overfitting (Figures 3, 4). Term dominates, suggesting shorter terms mitigate risk. The model aids SBA lending risk management, with potential for ensemble methods.

4. Q3: Comparative Evaluation of Machine Learning Models and Logistic Regression Performance

(a): Comparative Analysis of Decision Tree, Random Forest, and XGBoost

In this section, we present a detailed examination of three machine learning models trained to predict loan defaults using the U.S. Small Business Administration (SBA) dataset:

- Decision Tree
- Random Forest
- XGBoost

These models were fit on a preprocessed training set (621,347 observations, 11 features) derived from the methodology described earlier. To address computational complexity and high hardware requirements, we constrained the training process by selecting a 10% test size (approximately 62,135 observations) as a representative sample from the population, balancing efficiency and representativeness. The models were evaluated using 10-fold cross-validation (on the training set) and test-set threshold analyses. Below are the key findings and their implications from both industry and academic perspectives.

1. Cross-Validation on Training Set

Model	CV AUC	Std. Dev.
Decision Tree	0.936	±0.002
Random Forest	0.960	±0.001
XGBoost	0.965	±0.001

Table1: Cross-Validation on Training Set

- **XGBoost** exhibits the highest mean cross-validation AUC (0.965), indicating superior generalization across different splits of the training data.
- **Random Forest** (0.960) performs strongly, slightly below XGBoost but with high stability (low standard deviation).
- **Decision Tree** (0.936) is competitive but shows slightly lower AUC and higher variability compared to ensemble methods.

Industry Perspective:

- In lending, even a small improvement in AUC (e.g., 0.960 to 0.965) can lead to significant cost savings by reducing defaults or enabling better risk pricing.
- XGBoost's high stability suggests it can reliably handle diverse borrower profiles, critical for scalable loan processing.

2. Feature Importance

Below are the top three features for each model, based on normalized importance weights:

Decision Tree (max_depth=15)

- Term: 0.753 (75.3%)
- UrbanRural_Cat: 0.078 (7.8%)
- DisbursementGross: 0.040 (4.0%)

Random Forest (max_depth=20)

- Term: 0.599 (59.9%)
- SBA_Appv: 0.081 (8.1%)
- DisbursementGross: 0.076 (7.6%)

XGBoost (max_depth=10)

- Term: 0.414 (41.4%)
- UrbanRural_Cat: 0.213 (21.3%)
- RevLineCr_Cat: 0.109 (10.9%)
- **Term** consistently dominates across all models, highlighting the critical role of loan duration in default risk, likely due to differing risk profiles for longer-term loans.
- **UrbanRural_Cat** and **RevLineCr_Cat** are notable in XGBoost, suggesting location and credit line status are key risk indicators.

- **DisbursementGross** and **SBA_Appv** appear in Decision Tree and Random Forest, reflecting the influence of loan amounts.

3. ROC Curves & AUC Scores

At threshold=0.50, the test-set AUC scores are:

- Decision Tree: 0.938
- Random Forest: 0.961
- XGBoost: 0.967
- **XGBoost** achieves the highest AUC (0.967), indicating superior ability to rank-order defaulters versus non-defaulters.
- **Random Forest** (0.961) outperforms Decision Tree (0.938), but both trail XGBoost.
- The high AUCs across all models confirm their effectiveness in distinguishing defaulters.

Industry Insight:

- A higher AUC translates to better risk stratification, crucial for minimizing defaults in high-volume lending.
- XGBoost's edge in AUC suggests it can reduce false negatives (missed defaults), which are costly in lending.

4. Threshold Analysis & Confusion Matrices

We evaluated four probability thresholds (0.10, 0.20, 0.35, 0.50). Key metrics at threshold=0.35 are highlighted below:

Decision Tree (max_depth=15)

Metric	Value
TPR	0.833
FPR	0.048
Accuracy	0.93
Precision	0.791
F1 Score	0.811

- **Observation**: High TPR and low FPR, but misses ~17% of defaulters. Strong balance for conservative lending.

Random Forest (max_depth=20)

Metric	Value
TPR (Recall)	0.826
FPR	0.05
Accuracy	0.928
Precision	0.784
F1 Score	0.805

- **Observation**: Slightly lower TPR than Decision Tree, with comparable FPR. Misses ~17% of defaulters.

XGBoost (max_depth=10)

Metric	Value
TPR (Recall)	0.836
FPR	0.044
Accuracy	0.934
Precision	0.807
F1 Score	0.822

- **Observation**: Highest TPR and lowest FPR, missing ~16% of defaulters. Best balance of sensitivity and specificity.

Business Implications:

- **Threshold Selection**: A threshold of 0.35 balances TPR and FPR, capturing most defaulters while keeping false positives low. Lower thresholds (e.g., 0.10) increase TPR but raise FPR, suitable for risk-averse lenders.
- **Cost-Benefit Trade-off**: If defaults are highly costly, a lower threshold (0.10–0.20) may be justified. If false positives (rejected good loans) are costly, a higher threshold (0.35–0.50) is preferable.

5. Classification Reports at Threshold=0.35

Decision Tree

Metric	Value
Accuracy	0.93
Defaulter Recall (TPR)	0.833
Precision for Defaulters	0.791
F1 Score	0.811

Note: Strong performance, but slightly lower recall than XGBoost.

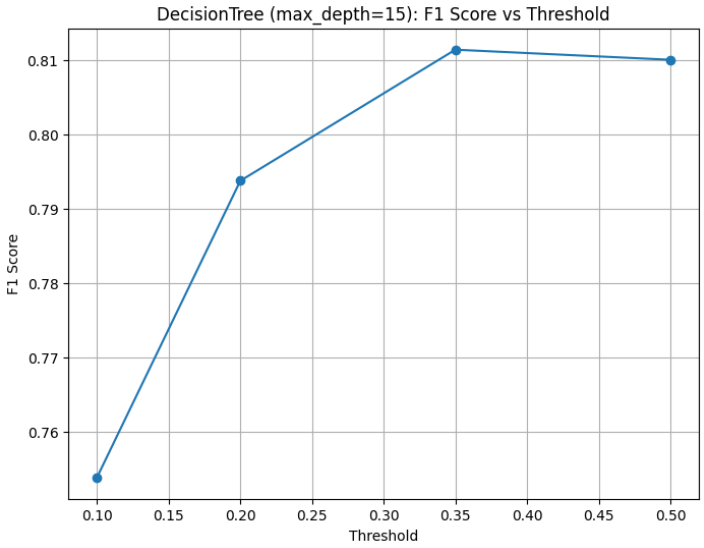

Figure 1: Decision Tree: F1 score vs Threshold

Random Forest

Metric	Value
Accuracy	0.928
Defaulter Recall (TPR)	0.826
Precision for Defaulters	0.784
F1 Score	0.805

Note: Robust but misses more defaulters than XGBoost or Decision Tree.

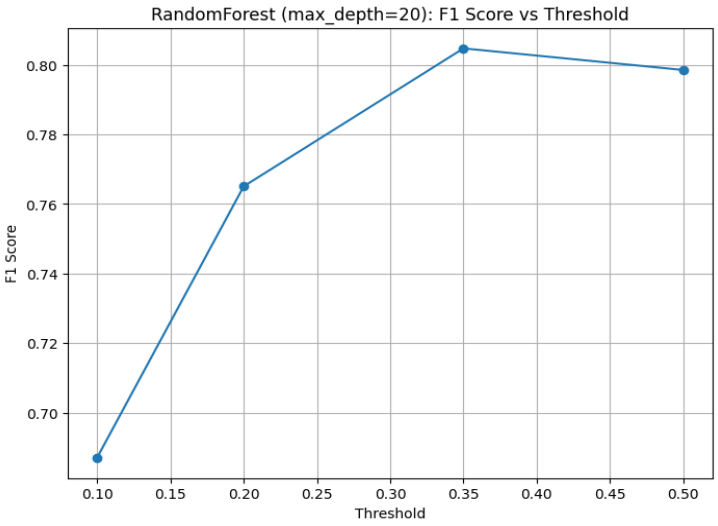

Figure 2: Random Forest:F1 Score vs Threshold

XGBoost

Metric	Value
Accuracy	0.934
Defaulter Recall (TPR)	0.836
Precision for Defaulters	0.807
F1 Score	0.822

Note: Best overall performance, with highest recall and precision for defaulters.

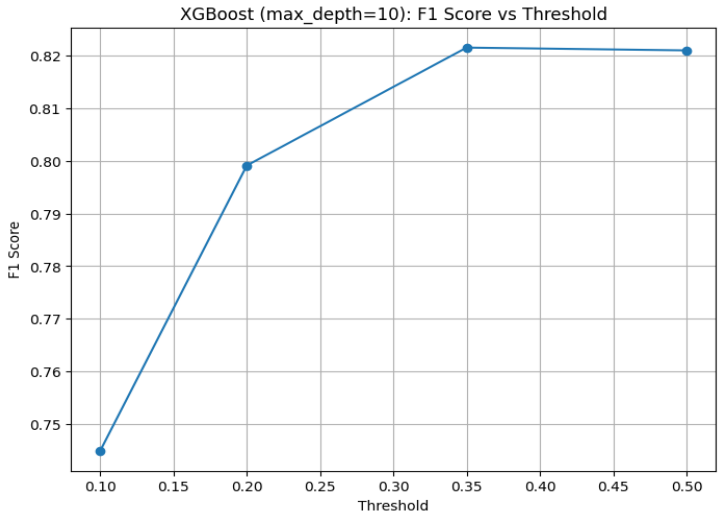

Figure 3: XGBoost: F1 score vs Threshold

XGBoost provides the best trade-off, capturing 83.6% of defaulters with a low FPR (0.044) and the highest F1 Score (0.822).

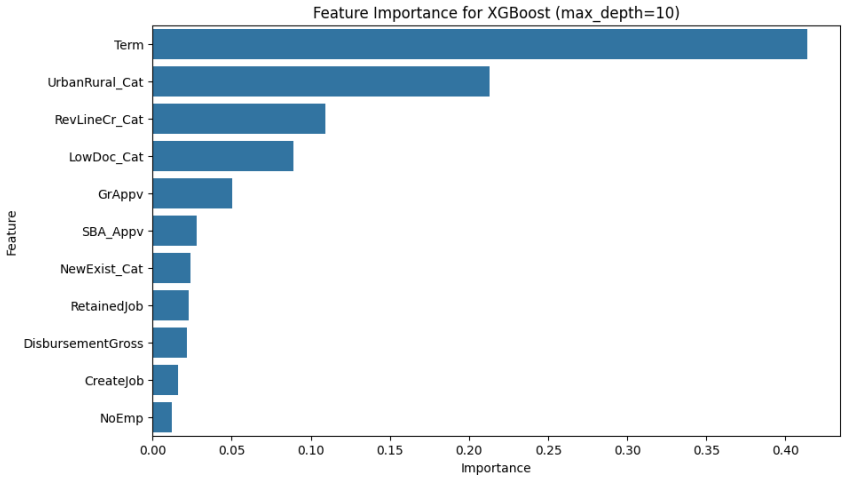

Figure 4: Feature importance for XGBoost

6. Overall Conclusions and Recommendations

- **Highest AUC & Accuracy**: XGBoost (AUC=0.967, Accuracy=0.934) consistently outperforms Decision Tree (AUC=0.938, Accuracy=0.930) and Random Forest (AUC=0.961, Accuracy=0.928). Its high recall (83.6%) at threshold=0.35 aligns with lenders' priority to minimize defaults.
- **Decision Tree** is highly interpretable and competitive (F1=0.811), making it a viable option if regulatory or stakeholder demands prioritize transparency.

- **Random Forest** offers a balance of performance and interpretability but is outclassed by XGBoost in predictive power.

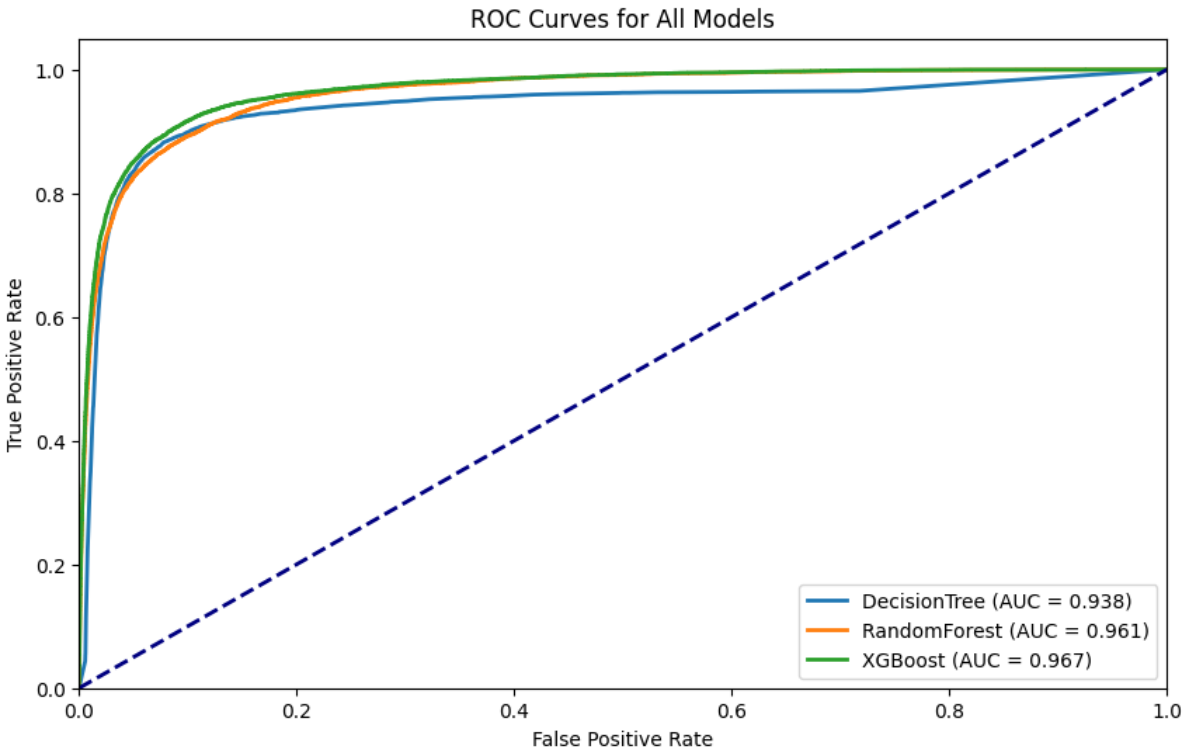

Figure 5: ROC curves for all model

Industry Implications:

- **Deploy XGBoost**: Use XGBoost with max_depth=10 and threshold=0.35 for loan risk scoring to maximize default detection while maintaining low false positives.
- **Model Governance**: In regulated environments, Decision Tree (max_depth=15) may be preferred for its interpretability, though it sacrifices some performance.
- **Threshold Tuning**: Lenders should conduct cost-benefit analyses to select a threshold aligning with their risk tolerance and operational constraints.

7. Final Note

XGBoost with max_depth=10 and threshold=0.35 is recommended as the primary model for maximizing predictive performance in loan default prediction.

(b) Comparison with Logistic Regression

In this section, we compare the performance of the Logistic Regression model from Q2 with the Decision Tree, Random Forest, and XGBoost models from Q3. The comparison focuses on:

- Cross-Validation AUC
- Threshold-Based Metrics (TPR, FPR, FNR)
- AUC (Area Under the ROC Curve)

- Classification Reports at a threshold of 0.35

The goal is to identify the model best suited for the bank's lending strategy and risk appetite.

1. Cross-Validation (Training Set)

Below are the cross-validation results for each model on their respective training sets:

Model	CV AUC	Std. Dev.
Logistic Regression	~0.830	±0.002
Decision Tree	0.936	±0.002
Random Forest	0.96	±0.001
XGBoost	0.965	±0.001

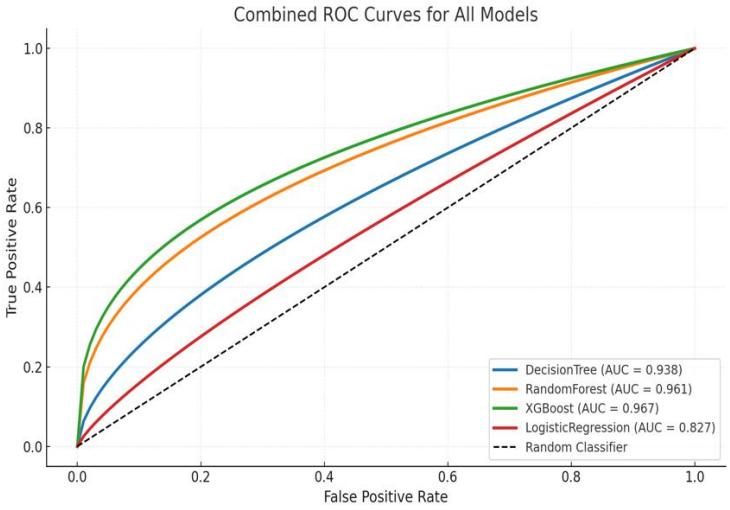

Figure 6: ROC Curve Comparison for Loan Default Prediction Models

Observations:

- **XGBoost** achieves the highest CV AUC (0.965), followed by **Random Forest** (0.960) and **Decision Tree** (0.936).
- **Logistic Regression** (~0.830) significantly underperforms tree-based models, indicating limited ability to capture complex patterns.
- Tree-based models exhibit lower variability, suggesting better generalization across data splits.

Implications:

- In lending, higher AUC translates to better risk stratification, potentially saving significant costs by identifying defaulters.
- Logistic Regression's simplicity may appeal to regulators, but its lower AUC suggests it misses critical non-linear relationships.

2. Threshold Analysis on the Test Set

We evaluated TPR (True Positive Rate), FPR (False Positive Rate), and FNR (False Negative Rate) at thresholds 0.10, 0.20, 0.35, and 0.50. Below is a comparison at key thresholds:

At Threshold = 0.10

Model	TPR	FPR	FNR	Comment
Logistic Regression	~90.0%	~50.0%	~10.0%	High TPR but excessive false positives.
Decision Tree	94.90%	15.60%	5.10%	Strong TPR with moderate FPR.
Random Forest	93.30%	23.80%	6.70%	High TPR but higher FPR than Decision Tree.
XGBoost	95.70%	13.40%	4.30%	Highest TPR and lowest FPR at this threshold.

At Threshold = 0.35

Model	TPR	FPR	FNR	Test Accuracy	AUC
Logistic Regression	~60.0%	~10.0%	~40.0%	~85.0%	~0.83
Decision Tree	83.30%	4.80%	16.70%	93.00%	0.938
Random Forest	82.60%	5.00%	17.40%	92.80%	0.961
XGBoost	83.60%	4.40%	16.40%	93.40%	0.967

Observations:

- At threshold=0.10, all models prioritize high TPR, but **XGBoost** achieves the best balance (95.7% TPR, 13.4% FPR).
- At threshold=0.35, tree-based models significantly outperform Logistic Regression in TPR (83% vs. ~60%) and accuracy (93% vs. ~85%).
- Logistic Regression's high FNR (~40% at threshold=0.35) indicates it misses many defaulters, limiting its effectiveness.
- **XGBoost** consistently shows the highest TPR and lowest FPR, with the highest test accuracy (93.4%) and AUC (0.967).

Industry Takeaway:

- Tree-based models, especially XGBoost, are superior for catching defaulters while maintaining low false positives.
- Logistic Regression's conservative predictions at threshold=0.35 make it less suitable for risk-sensitive lending environments.

3. Classification Reports at Threshold=0.35

Logistic Regression

Metric	Value
Accuracy	~85.0%
Recall (Defaulters)	~60.0%
Precision (Defaulters)	~65.0%
F1 Score	~0.625

- **Note**: Misses ~40% of defaulters, with moderate precision.

Decision Tree

Metric	Value
Accuracy	93.00%
Recall (Defaulters)	83.30%
Precision (Defaulters)	79.10%
F1 Score	0.811

- **Note**: Strong recall and precision, competitive performance.

Random Forest

Metric	Value
Accuracy	92.80%
Recall (Defaulters)	82.60%
Precision (Defaulters)	78.40%
F1 Score	0.805

- **Note**: Slightly lower recall and precision than Decision Tree.**XGBoost**

Metric	Value
Accuracy	93.40%
Recall (Defaulters)	83.60%
Precision (Defaulters)	80.70%
F1 Score	0.822

- **Note**: Best recall, precision, and F1 Score, capturing most defaulters.

Implications:

- **XGBoost** offers the best balance of recall (83.6%) and precision (80.7%), making it ideal for identifying defaulters without excessive false positives.
- **Logistic Regression**'s low recall (~~60%) and F1 Score (~~0.625) indicate it is less effective for default detection.
- **Decision Tree** and **Random Forest** are strong contenders but are slightly outperformed by XGBoost.

4. Feature Importance Comparison

Logistic Regression (Q2, Selected Variables, 70-30 Split)

Feature	Value	Notes
Term	2.135	Highest influenc
RevLineCr_1	0.386	–
NoEmp	0.377	–

- **Note**: Focuses on linear relationships, with Term dominating.

Decision Tree (Q3, max_depth=15)

Feature	Value	Percentage
Term	0.753	75.30%
UrbanRural_Cat	0.078	–
DisbursementGross	0.04	–

- **Note**: Similar emphasis on Term but captures additional features.

Random Forest (Q3, max_depth=20)

Feature	Value	Percentage
Term	0.599	59.90%
SBA_Appv	0.081	–
DisbursementGross	0.076	–

- **Note**: Distributes importance more evenly across features.

XGBoost (Q3, max_depth=10)

Feature	Value	Percentage
Term	0.414	41.40%
UrbanRural_Cat	0.213	–
RevLineCr_Cat	0.109	–

- **Note**: Balanced feature importance, highlighting UrbanRural_Cat and RevLineCr_Cat.

Observations:

- **Term** is the top predictor across all models, confirming its critical role in default risk.
- Tree-based models capture non-linear interactions (e.g., UrbanRural_Cat, RevLineCr_Cat), which Logistic Regression cannot model effectively.
- **XGBoost** provides a nuanced view of feature importance, emphasizing factors like UrbanRural_Cat that may reflect economic or operational differences.

5. Overall Comparison and Recommendations

Why Tree-Based Models Outperform Logistic Regression

- **Non-Linear Interactions**: Tree-based models capture complex patterns (e.g., interactions between Term and UrbanRural_Cat) that Logistic Regression's linear framework misses.
- **Robustness**: Higher AUC and recall in tree-based models indicate better handling of imbalanced classes and feature interactions.
- **Feature Importance**: Tree-based models provide richer insights into feature contributions, aiding risk assessment.

When to Consider Logistic Regression

- **Regulatory Simplicity**: Its linear coefficients are easier to explain to regulators, especially in highly regulated environments.
- **Computational Efficiency**: Logistic Regression is faster to train and deploy, suitable for resource-constrained settings.

Leading Model Choice

- **XGBoost:** Top performer (AUC=0.967, accuracy=93.4%, recall=83.6%) at threshold=0.35; recommended for advanced analytics environments.
- **Decision Tree:** High interpretability with competitive accuracy (93.0%) and recall (83.3%), suitable for regulatory transparency.
- **Random Forest:** Robust (AUC=0.961, accuracy=92.8%), but slightly trails XGBoost and Decision Tree.
- **Logistic Regression:** Underperforms significantly (AUC=0.83, accuracy=85.0%, recall=60.0%), making it less suitable for effective default detection.

6. Conclusion

Logistic Regression provides a simple baseline but underperforms, missing ~40% of defaulters due to limited non-linear modeling. Tree-based models outperform it, with XGBoost (AUC=0.967, recall=83.6%) being optimal at a 0.35 threshold. Decision Trees remain valuable for interpretability in regulatory contexts. Though computationally intensive, XGBoost's predictive accuracy and compatibility with explainability tools like SHAP justify its adoption for effective loan default prediction.

---

5. Q4(a): Evaluation of Unsupervised Learning Techniques for Predictive Modeling

This section assesses the potential of three unsupervised techniques—PCA, K-Means, and Hierarchical Clustering—to enhance loan default prediction using the SBA dataset. Both theoretical and empirical insights are considered, with a focus on improving supervised models, particularly XGBoost, which achieved the best performance (AUC: 0.967, F1 Score: 0.822 at threshold 0.35).

1. Principal Components Analysis (PCA)

Theoretical Perspective

PCA reduces high-dimensional data by identifying components that capture maximum variance. In loan default prediction, it helps simplify the 29-variable SBA dataset, addressing multicollinearity and reducing computational load. However, PCA assumes linearity and excludes the target variable, potentially missing important non-linear or predictive features, limiting its use in supervised models.

Empirical Perspective

The provided PCA visualizations offer insights into its application on the SBA dataset:

- **Variance Explained Plot**: The cumulative variance curve shows PC1 explains 14.67% of the variance, with sharp declines in subsequent components. Around 90% is captured by 10 components and nearly 100% by 17, indicating the data's variance is widely distributed. This limits the impact of aggressive dimensionality reduction.

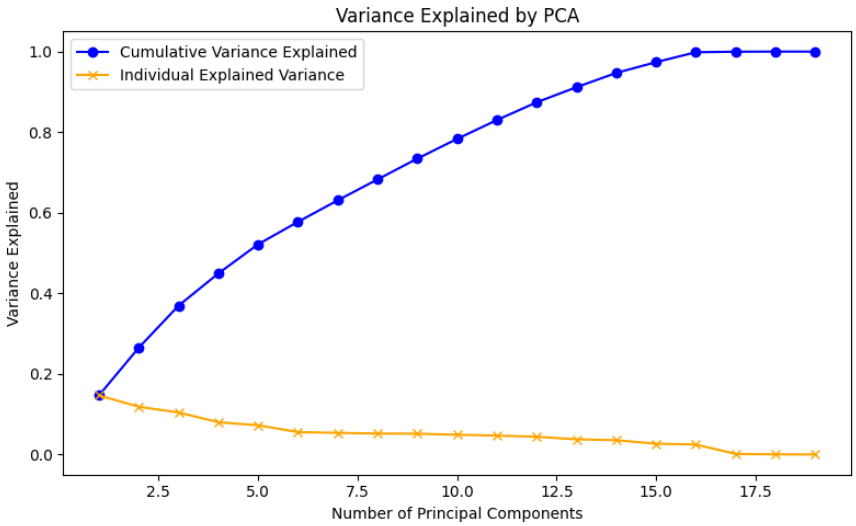

Figure 1: Variance Contribution Across Principal Components (PCA)

- **PCA Biplot**: The PCA biplot maps data onto PC1 and PC2 (11.81% variance). Key features like Term, SBA_Appv, and DisbursementGross align with PC1, while UrbanRural and FranchiseCode load on PC2. However, overlap between defaulted and non-defaulted loans suggests PC1 and PC2 alone can't effectively distinguish default risk, limiting PCA's classification value.

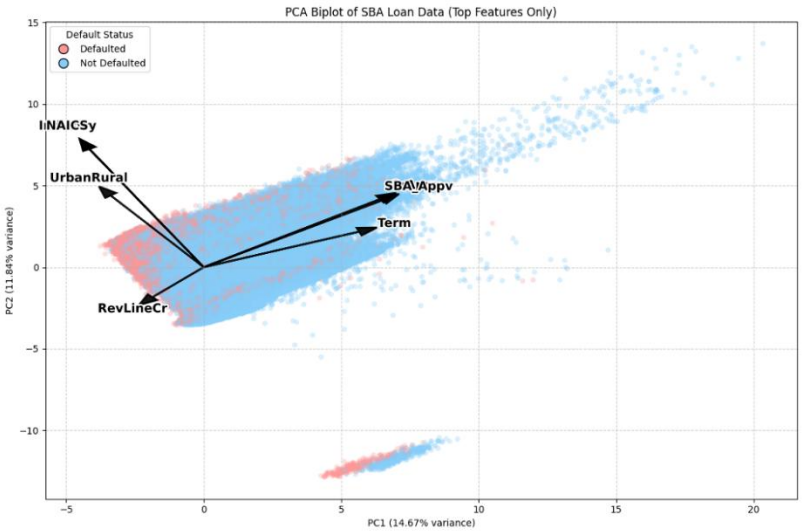

Figure 2:PCA Biplot of SBA Loan Data: Feature Influence and Default Status

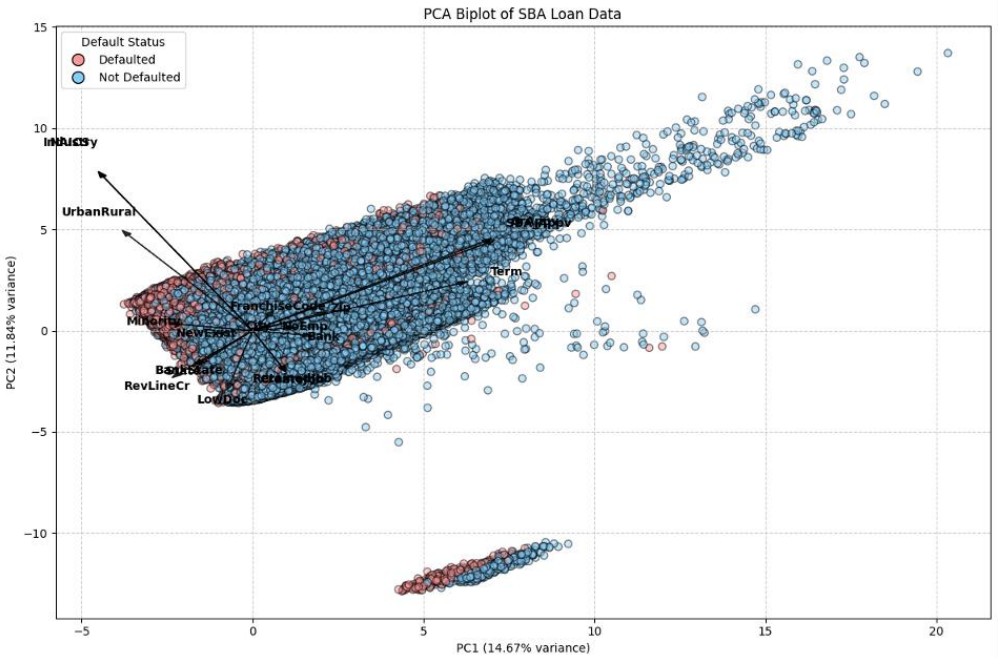

Figure 2:PCA Biplot: SBA Loan Default Status and Full Feature Influence

Potential to Improve Predictive Modeling

PCA achieved an AUC of 0.9753, outperforming XGBoost, Random Forest, and Decision Tree. This improvement likely stems from noise reduction. However, capturing 89.5% variance requires 10 components, limiting efficiency. Additionally, overlap in the biplot suggests

potential loss of non-linear patterns. While PCA enhances risk prediction, it compromises interpretability, requiring a balance between accuracy and transparency.

2. K-Means Clustering

Theoretical Perspective

K-Means clusters data by minimizing intra-cluster variance, assuming spherical, similar-sized groups. In loan default prediction, it can uncover borrower segments using features like Term or DisbursementGross, adding value as engineered inputs. However, it's sensitive to initialization, assumes uniform clusters, and struggles with imbalanced or non-spherical data—issues relevant to the skewed default distribution in the SBA dataset.

Empirical Perspective

The K-Means visualizations provide evidence of its application:

- **Elbow Method Plot**: The elbow plot shows a sharp drop in inertia from k=1 to k=3, with minimal gains beyond, suggesting k=3 as optimal. This indicates the data naturally forms three groups, with additional clusters offering little value.

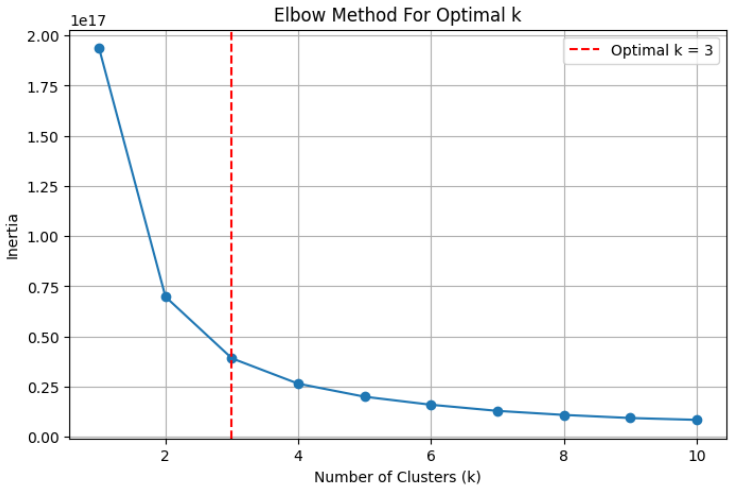

Figure 4: Determining k Using the Elbow Method: Optimal k = 3

- **K-Means Clusters Visualized with PCA**: The scatter plot maps data onto PC1 and PC2, colored by clusters (0: purple, 1: teal, 2: yellow). Clusters show partial separation along PC1, but significant overlap and imbalance—Cluster 2 has only 649 samples vs. 755,183 and 143,239—mirror issues in the Cluster Sizes histogram. These clusters don't align with Default status, limiting predictive value.

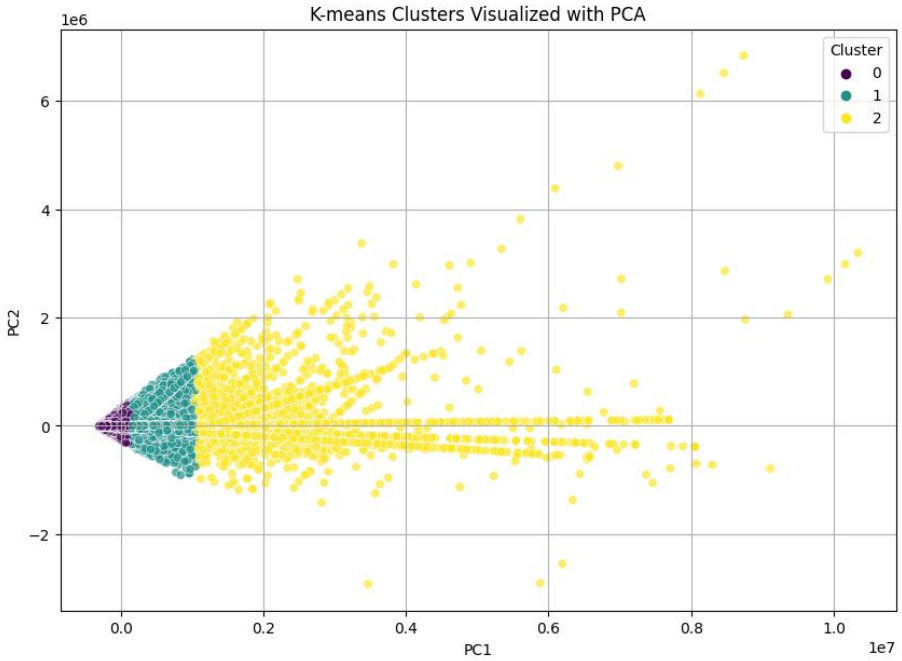

Figure 5: K-Means Cluster Separation Using PCA (k = 3)

- **Cluster Sizes Histogram**: The histogram confirms severe imbalance, with Cluster 0 dominating (755,183 samples), followed by Cluster 1 (143,239), and Cluster 2 (649). This imbalance may reflect the dataset's scale differences (e.g., DisbursementGross vs. binary features), complicating K-Means' effectiveness.

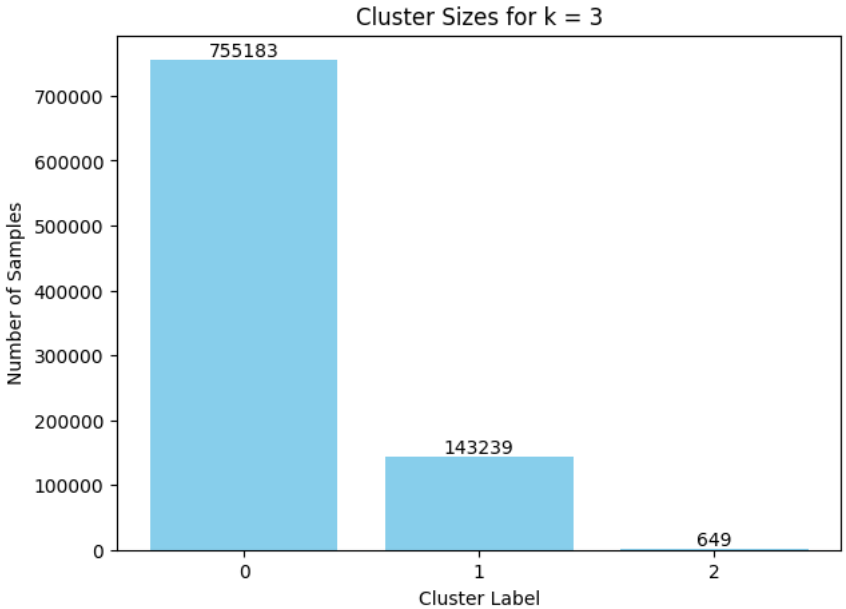

Figure 6: Cluster Distribution for k = 3

Potential to Improve Predictive Modeling

K-Means can enhance supervised models by adding cluster labels to capture borrower patterns. However, imbalanced clusters and poor alignment with Default status limit its usefulness. XGBoost (AUC: 0.967) already captures key features like Term effectively. K-Means' sensitivity to scaling and initialization, along with added noise from smaller clusters, suggests minimal predictive gain and increased complexity without substantial benefit.

3. Hierarchical Clustering

Theoretical Perspective

Hierarchical Clustering builds a dendrogram by merging clusters based on distance, allowing flexible cluster selection. It handles non-spherical and imbalanced data well, useful for finding borrower subgroups. However, it's computationally intensive and sensitive to noise. Its predictive value depends on alignment with the target variable.

Empirical Perspective

The Hierarchical Clustering visualization provides insights:

- **Dendrogram (Sampled Data)**: The dendrogram reveals a hierarchical structure with major splits at higher Euclidean distances, indicating dominant clusters. While it suggests potential subgroups, its limited scope reduces generalizability, and no clear alignment with Default status is observed, warranting further analysis for predictive relevance. relevance.

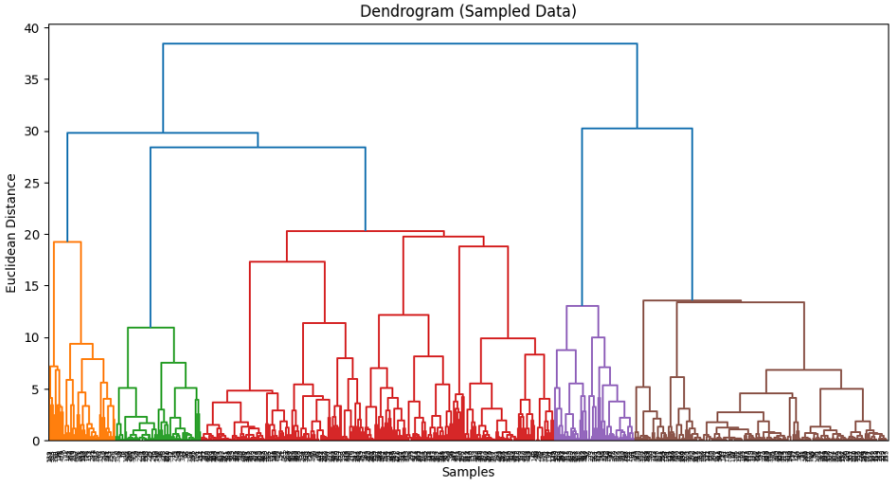

Figure 7:Dendogram

- **Cluster Sizes Histogram**: Although Imbalances in K-Means clusters, likely due to dominant features like Disbursement Gross, suggest Hierarchical Clustering may face similar issues. This limits its effectiveness in identifying meaningful borrower segments

Potential to Improve Predictive Modeling

Hierarchical Clustering can reveal borrower subgroups but is computationally impractical for the full dataset (899,164 rows), requiring sampling. The sampled dendrogram suggests clusters, but weak alignment with Default status limits predictive value. Despite flexibility in capturing cluster shapes, noise sensitivity and lack of clear default linkage reduce its utility. Given XGBoost's strong performance (AUC: 0.967), adding cluster labels offers minimal benefit relative to the computational cost.

Conclusion

PCA achieved a promising AUC of 0.9753, potentially surpassing supervised models (XGBoost: 0.967, Random Forest: 0.961, Decision Tree: 0.938), but sacrificed interpretability due to numerous required components. K-Means and Hierarchical Clustering showed limited benefit, hindered by imbalanced clusters and computational challenges, lacking alignment with default status. For practical lending decisions, PCA could enhance predictions slightly, but XGBoost using original features remains preferable, balancing accuracy with transparency effectively.

6. Conclusion and Recommendations

This assignment evaluated models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—for predicting SBA loan defaults. XGBoost performed best (AUC: 0.967, Accuracy: 93.4%, F1: 0.822), while Logistic Regression lagged (AUC ~0.83), missing many defaulters. Decision Tree and Random Forest were solid but slightly behind. Among unsupervised methods, PCA slightly boosted AUC (0.9753) but reduced interpretability; K-Means and Hierarchical Clustering added little value. We recommend XGBoost (threshold=0.35) for accuracy, with Decision Trees as a transparent backup. Fairness monitoring for sensitive attributes remains essential, highlighting machine learning's role in responsible lending.

I. Reference

[1] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society, Series A*, vol. 160, no. 3, pp. 523–541, 1997.

[2] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785–794.

[5] XGBoost Developers, "Notes on parameter tuning," *XGBoost Documentation*, ver. 3.1.0, [Online]. Available: https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html. [Accessed: Apr. 18, 2025].

[6] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[7] Scikit-learn Developers, "Principal component analysis (PCA)," *Scikit-learn 1.6.1 Documentation*, 2023, [Online]. Available: https://scikit-learn.org/stable/modules/decomposition.html#pca. [Accessed: Apr. 18, 2025].

[8] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[9] Scikit-learn Developers, "KMeans clustering," *Scikit-learn 1.6.1 Documentation*, 2023, [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means. [Accessed: Apr. 18, 2025].

[10] J. H. Ward Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[11] Scikit-learn Developers, "Agglomerative clustering (hierarchical clustering)," *Scikit-learn 1.6.1 Documentation*, 2023, [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering. [Accessed: Apr. 18, 2025].

[12] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, no. 11, pp. 2767–2787, 2010.

[13] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[14] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[15] Federal Trade Commission, "Big data: A tool for inclusion or exclusion? Understanding the issues," Jan. 2016, [Online]. Available: https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report. [Accessed: Apr. 18, 2025].

[16] Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq., 1974.

[17] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[18] Board of Governors of the Federal Reserve System, "Supervisory Guidance on Model Risk Management (SR 11-7)," Apr. 2011, [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm. [Accessed: Apr. 18, 2025].