VINUNIVERSITY

COMPUTATIONAL MACHINE LEARNING FOR BUSINESS ANALYTICS
BANA 4020

Should This Loan Be Approved or Denied ?

Predictive Modeling to Maximize Loan Prediction and Profit

Presented by Group 4

Doan Huong Thanh
V202301046
SBA Data Analyst

Hoang Nguyen Gia Huy
V202200853
SBA Data Analyst

Do Quang Hai
V202200822
SBA Data Analyst

TABLE OF CONTENTS

1. Introduction
2. Literature Review
3. Dataset Description
4. Methodology
5. Conclusion

About Small Business Administration (SBA)

The SBA supports entrepreneurship as small businesses drive job creation and help reduce unemployment.

SBA works to ignite change and spark action so small businesses can confidently start, grow, expand, or recover

Founding Date

1953

Mission

Promote and assist the growth of small enterprises within the American credit market

Flagship Initiatives

Loan guarantee program as a backstop for lenders, lowering their risk when lending to small businesses

Scale

10 Regional Offices **8000** Employees

In FY 2024, the SBA supported over 103,000 small-business financings, marking a 22% increase over FY 2023. Overall capital impact climbed to $56 billion

Figure 1: SBA's Loan Breakdown by Program

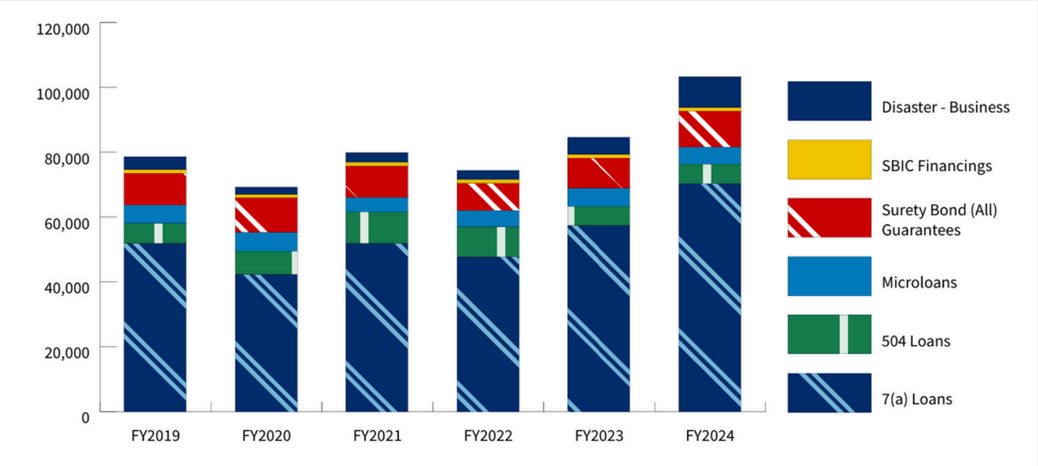
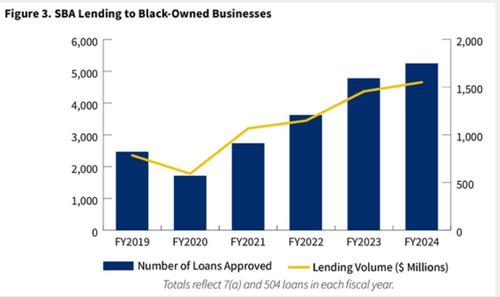
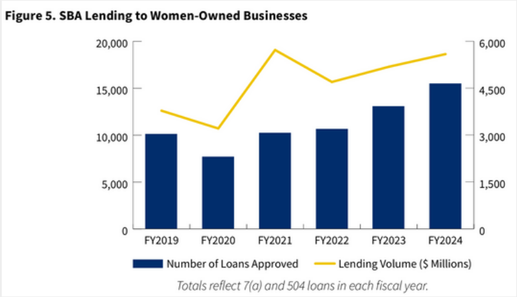


Figure 2: Lending to Underserved and Minority Entrepreneurs



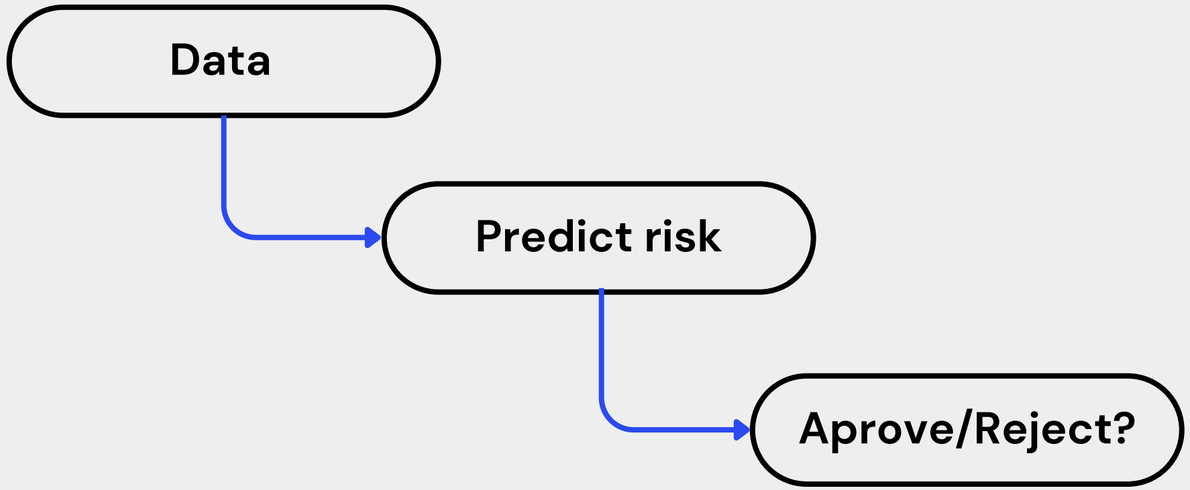


SBA Dataset: Existed applications and common modeling approaches

Real-world data for default prediction with top models are Logistic Regression, Random Forests, XGBoost.

The SBA dataset has been widely used for risk prediction and educational purposes:

- Predict default risk for small business loans.
- Simulate loan officer decision-making in real-world settings.
- Benchmark machine learning models for binary classification tasks.
- Feature engineering to improve prediction: recession, real estate, loan size.



There are several methods have been used, including:

Methods	Key purposes	Strengths
Logistic Regression	Model binary outcomes	Simple, interpretable, efficient on large data
Random Forests	Handle high-dimensional structured data	Reduces overfitting, highlights feature importance
XGBoost	Boosted trees for classification	High accuracy, handles missing data, regularization support

and other methods: Decision Tree, KNN Classifiers, Artificial Neural Networks, AdaBoost/Gradient Boost, etc.

Sources:

(1): "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines

(2): https://github.com/bagjasatia/Final_Project_SBA_Loan_Approval/blob/master/README.md

SBA Dataset: Key results summary

Gradient Boosting achieves 91% accuracy; industry and business age are key predictors

Default risk varies by business characteristics; Gradient Boosting delivers top prediction performance.

Default rates average 18–20% but **vary widely** across industries and business profiles (1)
Eg: <10% in healthcare to >30% in real estate and recessions.

Real estate-backed loans show **much lower** default risk **(~1.6%)** (1)

Strong predictors include: Industry type (NAICS), business age, loan size, recession periods. (1) & (2)

Gradient Boosting outperforms other models with **91% accuracy, AUC 0.971.** (2)

Feature engineering and **model tuning** significantly boost model performance.

Key limitations in current SBA loan default models:

- **Selection bias**: only approved loans are included.
- **Missing key financial features:** no credit score, income, or debt data.
- **Weak macroeconomic context:** limited recession adjustment only.
- **Low interpretability:** complex models lack transparency.
- **No cost-sensitive evaluation**: equal cost assigned to false positives and false negatives.

Sources:
(1): "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines
(2): https://github.com/bagjasatia/Final_Project_SBA_Loan_Approval/blob/master/README.md

SBAnational dataset: Collection & Structure

Real-world dataset from the U.S. Small Business Administration (SBA) — authentic historical loan data from 1987–2014.

How Was the Data Set Collected?

- Publicly available SBA 7(a) loan data (1987–2014) from the SBA's National Loan Database.
- EAggregated from participating lenders (banks, credit unions, and community lenders).
- Includes borrower details, loan terms, and final repayment status as reported by lenders

What Are the Variables?

ROWS (OBSERVATIONS)

Each row represents a single, unique SBA-guaranteed loan application

Total observations: approximately **899,000** individual loans

COLUMNS (VARIABLES)

Loan Identifiers & Dates

LoanNr_ChkDgt
ApprovalDate
DisbursementDate
ChgOffDate
ApprovalFY
ChgOffFiscalYear)

Loan Attributes & Financials

LoanStatus
MIS_Status
ChgOffPrinGr
BalanceGross
DisbursementGross
GrAppv
SBA_Appv
Term
RevLineCr
LowDoc

Borrower & Business Characteristics

Name
City, State, Zip,
CountyName
NAICS
SIC
NewExist
FranchiseCode
UrbanRural
NoEmp
CreateJob
RetainedJob

Other Descriptive Fields

BankName, BankState
BankZip

SBAnational dataset: Collection & Structure

WHAT DO THE VARIABLES DESCRIBE?

LOAN IDENTIFIERS & DATES

LoanNr_ChkDgt: Categorical (string; unique ID)

ApprovalDate, DisbursementDate, ChgOffDate: Temporal (date variables; not directly numeric but convertible to features like “loan age”)

ApprovalFY, ChgOffFiscalYear: Discrete integer (fiscal year)

BORROWER & BUSINESS CHARACTERISTICS

City, State, CountyName, Name: Categorical (string)

Zip, FranchiseCode, NAICS, SIC: Categorical once treated (though NAICS/SIC could be binned into industry groups)

NewExist, UrbanRural: Categorical (integer codes)

NoEmp, CreateJob, RetainedJob: Numerical (count data; integer)

LOAN ATTRIBUTES & FINANCIALS

DisbursementGross, GrAppv, SBA_Appv, BalanceGross, ChgOffPrinGr, Term: Numerical (continuous/dollar amounts or integer months)

RevLineCr, LowDoc: Categorical (binary “Y” or “N”)

MIS_Status: Categorical (binary labels “PIF” or “CHGOFF”)

LENDER INFORMATION

BankName, BankState, BankZip: Categorical (string or integer for ZIP)

SBAnational dataset: Collection & Structure

WHAT DO THE VARIABLES DESCRIBE?

NUMERICAL FEATURES

- Term
- NoEmp
- CreateJob
- RetainedJob
- FranchiseCode
- DisbursementGross
- BalanceGross
- ChgOffPrinGr
- GrAppv
- SBA_Appv
- LowDoc_Binary
- RevLineCr_Binary
- UrbanRural_Numeric
- MIS_Status_Binary

→ Standard Scaler

CATEGORICAL FEATURES

- FranchiseCode
- NewExist
- RevLineCr_Binary
- LowDoc_Binary
- UrbanRural_Numeric
- RealEstate
- Recession
- NAICS2
- State
- ApprovalFY

→ One-hot Encoding

SBAnational dataset: Collection & Structure

WHAT IS THE RESPONSE VARIABLE OF INTEREST

MIS_Status_Binary (derived from MIS_Status)

Definition

0 = Loan Paid In Full (MIS_Status = "PIF")
1 = Loan Defaulted/Charge-Off (MIS_Status = "CHGOFF")

Type: Discrete (binary)

Description:

Indicates whether the borrower satisfied the **loan obligation (0)** or **defaulted (1)**. This variable drives the classification models and reflects the ultimate credit risk outcome.

WHAT DO YOU WANT THE MODELS TO DO?

PRIMARY GOAL

Classify each new loan application as "Low Risk" (**Predict MIS_Status_Binary = 0**) or "High Risk" (**Predict MIS_Status_Binary = 1**)

BUSINESS OBJECTIVES

- Predict borrower's probability of paying in full vs. default
- Use cost-sensitive approach: penalize false negatives (approving a defaulter) more than false positives (denying a good loan)
- Determine cutoff ŧ that maximizes net profit on a validation set (balance interest income vs. default losses)
- Generate gain/lift charts to help credit officers target top x% of applicants for optimal returns

SBAnational dataset: Collection & Structure

Basic Exploration for the Variables

TOTAL OBSERVIATION

~ 899,000 loans

27 columns

(including identifiers, borrower attributes, loan attributes, geographical fields, and outcome)

MISSING-VALUE OVERVIEW

- Certain categorical fields (e.g., UrbanRural) have codes for "Undefined" (0) rather than explicit NaN.
- Some loans lack ChgOffDate because they never defaulted—this is expected and must be handled when computing "loan age."
- A small fraction of records have missing values for NoEmp, CreateJob, or RetainedJob (each < 5%); these will be imputed with median values.

TARGET DISTRIBUTION (MIS_STATUS_BINARY)

Approximately 82% "Paid In Full" (0) vs. 18% "Charge-Off" (1). (Exact percentages will be confirmed in the next step of EDA.)

EXAMPLE SUMMARIES OF KEY NUMERICAL FEATURES

DISBURSEMENTGROSS (LOAN SIZE)

Range: $1,000 – $5,000,000

Median: ~ $67,500

Skew: Right-skewed (a small subset of very large loans)

NOEMP (NUMBER OF EMPLOYEES)

Range: 0 – 3,000+ employees

(mode typically around 5–10 employees)

Many small businesses cluster below 20 employees.

CREATEJOB / RETAINEDJOB

Typical values:

0 – 100 jobs; many loans aimed at creating or retaining fewer than 10 jobs.

SBAnational dataset: Collection & Structure

Basic Exploration for the Variables

EXAMPLE SUMMARIES OF KEY CATEGORICAL FEATURES

NAICS (Industry Code)

- **1,000+** codes ➡ Grouped into 2-digit industries
- **Default rates vary:** High: Construction
  Low: Health Care

State

- **50** states + DC + territories
- CA, TX, FL = large loan volumes
- Higher defaults in recession-hit states (e.g., CA)

UrbanRural

70 %	25%	5%
Urban	Rural	Undersined

- **Urban loans** show slightly lower default rates

CORRELATIONS & MULTICOLLINEARITY (NUMERICAL)

Correlations & Multicollinearity (Numerical)

- CreateJob & RetainedJob: moderate correlation **($\rho \approx 0.62$)**
- GrAppv & DisbursementGross: highly correlated **(> 0.90)**
  ➡ use one or ratio
- BalanceGross = 0 for “PIF”; > 0 indicates default

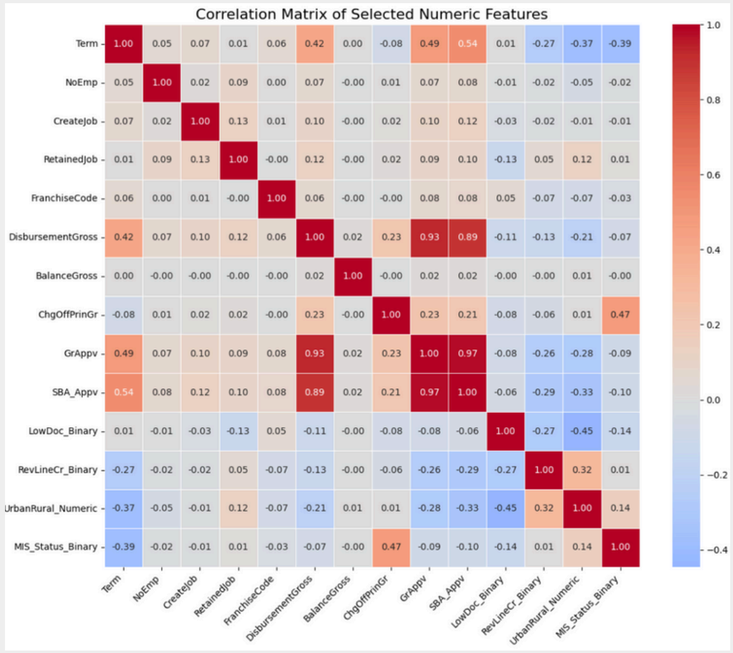
First Glimpse of Target vs. Key Predictors

- **Median loan:** PIF $75K > CHGOFF $50K
- **Default rates:**
  ~ **25%** in Accommodation & Food Services
  ~10% in Health Care
- **Small firms (<5 employees)**: **~22%** default
- **Larger firms (>20):** **~12%** default

Exploratory Data Analysis (EDA)

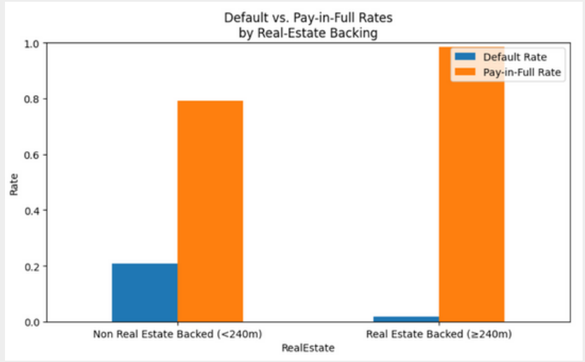
We conduct EDA to better understand underlying characteristics of the dataset

Correlation Matrix



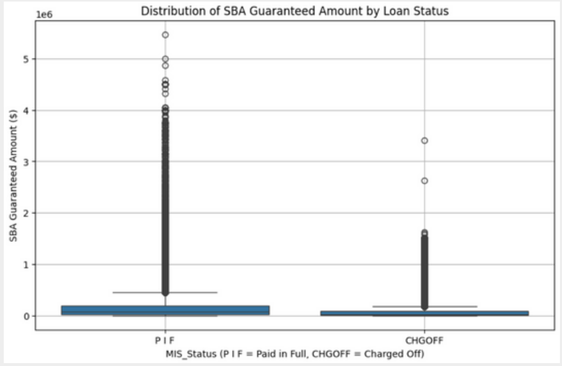
- **Term** DisbursementGross **0.420238**, GrAppv **0.486792**, SBA_Appv **0.542396**
- **DisbursementGross** GrAppv **0.931974**, SBA_Appv **0.891309**
- **ChgOffPrinGr** MIS_Status_Binary **0.474818**
- **GrAppv** SBA_Appv **0.972688**

SBA's Guaranteed Proportion of Loan



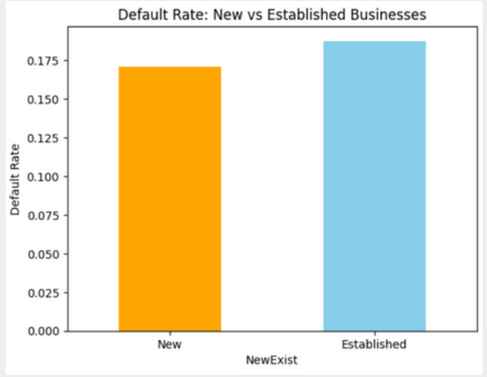
Long-term, real-estate-secured loans carry far less risk of charge-off

SBA's Guaranteed Proportion of Loan



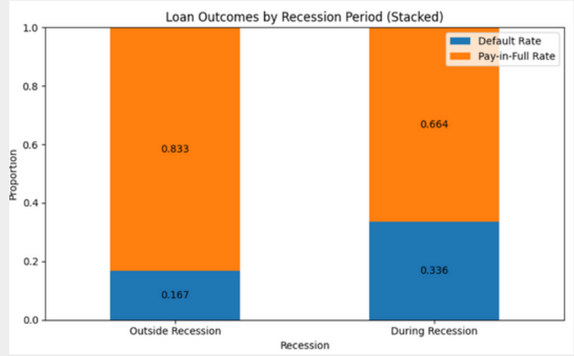
Loans with larger SBA guarantees are more likely to be repaid

Default Rate by New vs Established Business



There is not significant difference between New and Established Business

MIS Status in Recession Time



Businesses during recession have higher default rate and lower PIF

Exploratory Data Analysis (EDA)

We conduct EDA to better understand underlying characteristics of the dataset

Top 10 States by SBA Default Rate

	State	loan_count	default_rate
9	FL	41212	0.273694
7	DC	1613	0.239926
10	GA	22277	0.239628
33	NV	8024	0.232236
14	IL	29669	0.226701
22	MI	20545	0.225052
42	TN	9403	0.212128
3	AZ	17631	0.207501
40	SC	5597	0.204647
31	NJ	24035	0.201125

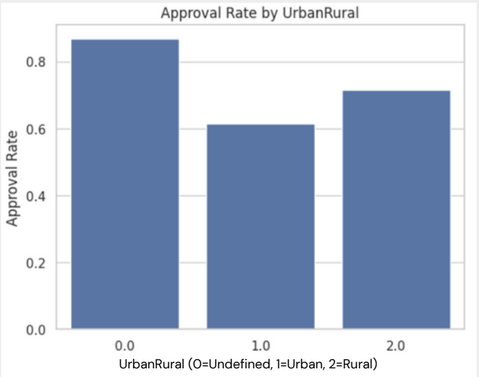
- Florida (FL) and DC have the highest default rates.
- High loan counts in FL and IL amplify risk exposure.
- Regional concentration may impact overall portfolio performance.

Top 10 NAICS by Default Rate

	NAICS2	loan_count	default_rate
15	53	13632	0.287312
14	52	9496	0.284266
11	48	20310	0.268888
13	51	11379	0.248284
19	61	6425	0.242462
18	56	32685	0.235513
10	45	42514	0.234154
4	23	66646	0.232554
12	49	2221	0.229864
9	44	84737	0.223941

- Florida (FL) and DC have the highest default rates.
- High loan counts in FL and IL amplify risk exposure.
- Regional concentration may impact overall portfolio performance.

Approval Rate by UrbanRural



- Loan approval rates are higher in urban areas than in suburban areas.

Business objective and problem formulation

Profit optimization requires balancing prediction accuracy with risk management.

Objective:

Maximize lender profit using historical loan data and predefined cost matrix.

Problem formulation:

- classify loans as PIF or default based on borrower and loan characteristics.
- minimize cost-sensitive misclassification losses.
- balance accuracy, F1-score, and recall to optimize profit.
- profit is calculated based on the disbursement loan amount.

Proposed modeling methods and execution roadmap

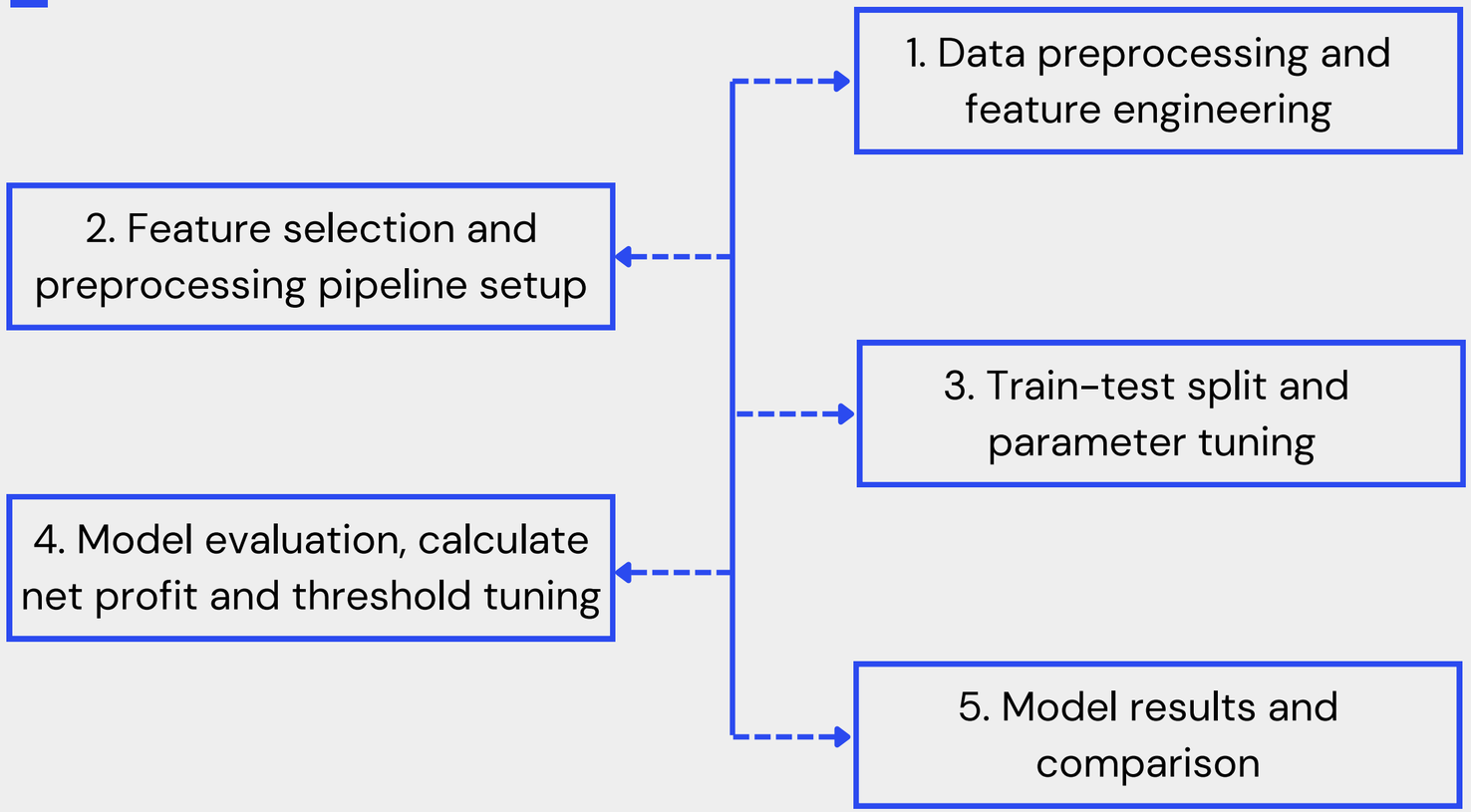
A diverse set of models is developed and tuned to maximize predictive performance and business profitability.

Proposed methods:

- KNN classifiers
- Random Forest
- XGBoost
- Neural Network
- Logistic Regression

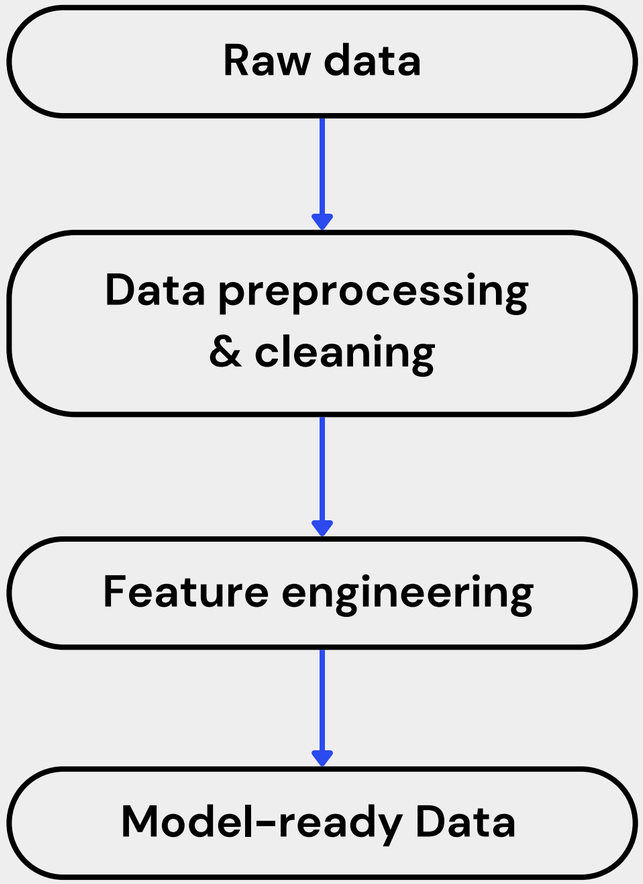
>> Selected models cover both **explainable** and **high-performance** algorithms for robust prediction.

Execution roadmap:



Data preprocessing, cleaning, and feature engineering

Robust preprocessing and feature engineering create a solid foundation for model performance.



Data preprocessing and cleaning

Step	Purpose
Convert currency columns to numeric	Enable numerical computation and modeling.
Create binary target variable (MIS_Status_Binary)	Transform loan status into ML format
Create binary features (LowDoc_Binary, RevLineCr_Binary, UrbanRural_Numeric)	Standardize categorical data for modeling

Feature engineering

- Create:
  - GuaranteeRatio = SBA_Appv / GrAppv
  - RealEstate (term ≥ 240 months → 1, else 0)
  - Recession indicator (approval date in 12/2007–06/2009 → 1)
- Extract:
  - NAICS2 (first 2 digits of NAICS)
  - ApprovalFY (approval year from approval date)

Feature selection

Effective feature selection improves model accuracy and interpretability.

The selected features capture key borrower attributes, loan details, and economic conditions.

Term	loan term in months
Disbursement	
NoEmp	
CreateJo	
RetainedJ	
GuaranteeRatio	percentage guaranteed by SBA
FranchiseCode	indicator for franchise affiliation
NewExist	indicator if business is new or existing

RevLineCr_Binary	indicator for revolving line of credit
	entation loan
	rban or rural
	backed loans
	ing recession
NAICS2	two-digit industry sector code
State	borrower's state
ApprovalFY	year of loan approval

To select feature, we:

- perform EDA (as above) to identify important features
- select features based on correlation strength and business relevance
- remove redundant and low-important features to enhance model efficiency

Feature selection

Effective feature selection improves model accuracy and interpretability.

The selected features capture key borrower attributes, loan details, and economic conditions.

Term	loan term in months
DisbursementGross	total amount disbursed to the borrower
NoEmp	number of employees
CreateJob	number of jobs created by the loan
RetainedJob	number of jobs retained due to the loan
GuaranteeRatio	percentage guaranteed by SBA
FranchiseCode	indicator for franchise affiliation
NewExist	indicator if business is new or existing

RevLineCr_Binary	indicator for revolving line of credit
LowDoc_Binary	indicator for low-documentation loan
UrbanRural_Numeric	location classification: urban or rural
RealEstate	indicator for real estate-backed loans
Recession	indicator for approval during recession
NAICS2	two-digit industry sector code
State	borrower's state
ApprovalFY	year of loan approval

Model Training Methods

Important Steps to ensure model accuracy (KNN, Random Forest, XGBoost, Logistic Regression, Neural Network)

1 - Create Pipeline & Split training and testing

Ensure that features like loan size can be correctly scaled, correlated, and fed into models.

X_train.shape = (61446, 16)
X_test.shape = (553 201, 16)

KNN, RF, XGB, LR

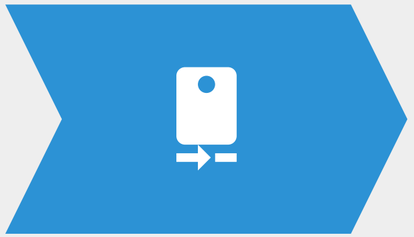
2 - Create Parameter Grid

Prepare to tune parameter including necessary factors in each model

3 - Perform Grid Search CV

Rigorously tune hyperparameters to balance bias vs. variance, prevent overfitting, and maximize predictive power

4 - Results



P(Default)
ROC-AUC
Classification Report
Confusion Matrix

NN

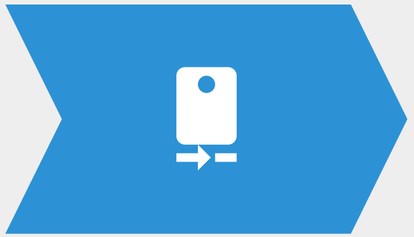
2 - Standalone Preprocessing

Use preprocessor.fit_transform for X_train and X_test

3 - Model Architect & Training

Set up 4 layers with nodes, activation relu and base layer use sigmoid to handle binary target → fit result

4 - Results

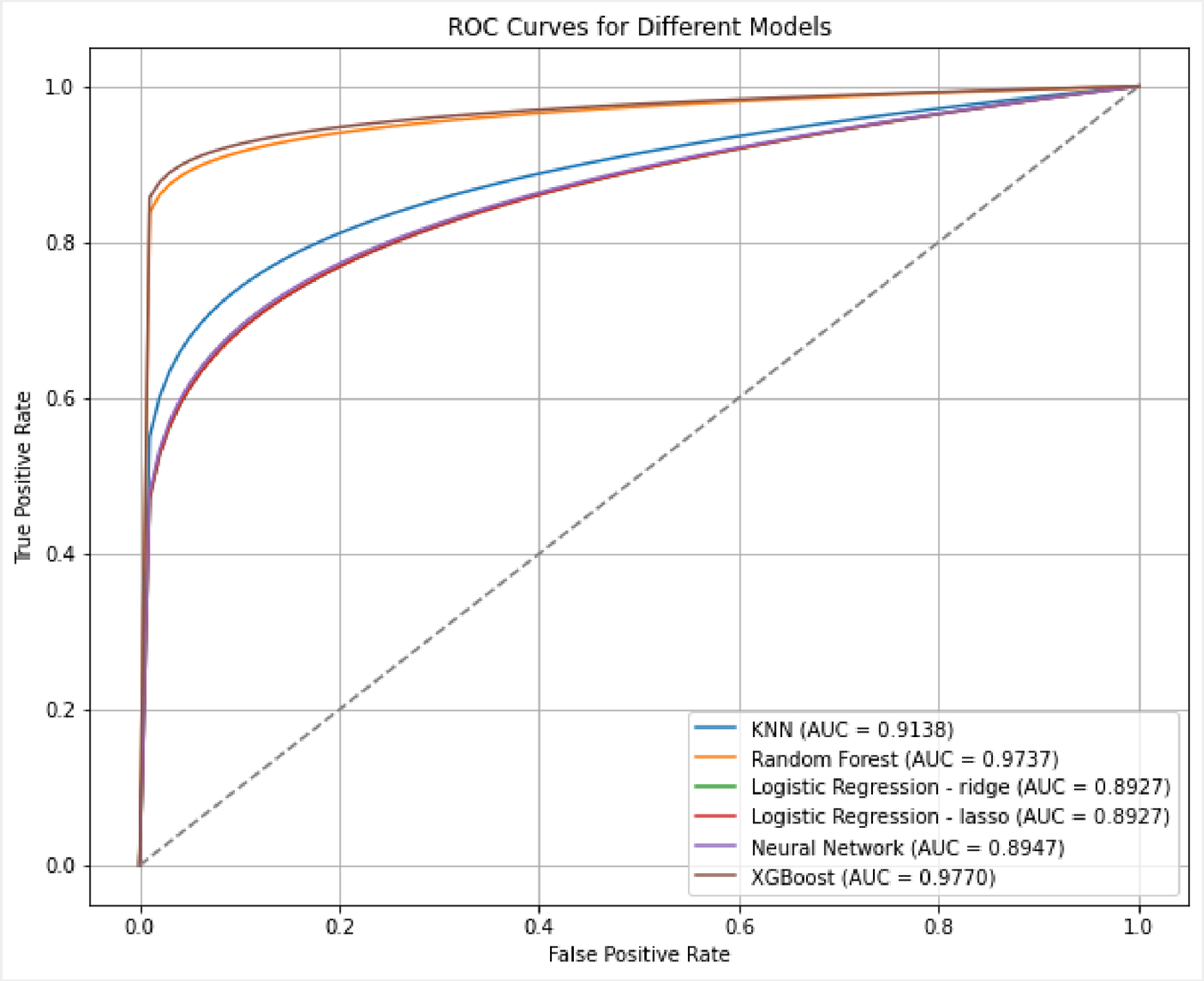


P(Default)
ROC-AUC
Classification Report
Confusion Matrix

Model evaluation results *(for default threshold = 0.5)*

Model	Best hyperparameters	Test ROC AUC	Accuracy	TPR	TNR	Error Rate
KNN	knn__n_neighbors': 31, 'knn__p': 1, 'knn__weights': 'distance	0.91381	0.8939	0.5498	0.9697	0.1061
Random Forest	rf__max_depth': 23, 'rf__max_features': None, 'rf__min_samples_leaf': 8, 'rf__min_samples_split': 5, 'rf__n_estimators': 153	**0.9737**	**0.9442**	0.8163	0.9723	0.0559
Logistic Regression (ridge)	clf__class_weight': None, 'clf__C': 1438.44988	0.8927	0.1155	0.4737	0.0365	0.4896
Logistic Regression (lasso)	clf__class_weight': None, 'clf__C': 1438.44988	0.8927	0.1155	0.4737	0.0365	0.4896
Neural Network	x	0.8947	0.8866	0.6957	0.9286	0.1135
XGBoost	'xgb__subsample': 0.7, 'xgb__n_estimators': 150, 'xgb__max_depth': 13, 'xgb__learning_rate': 0.10666666666666666, 'xgb__gamma': 1.0, 'xgb__colsample_bytree': 0.5	**0.9770**	**0.9478**	0.8254	0.9749	0.0522

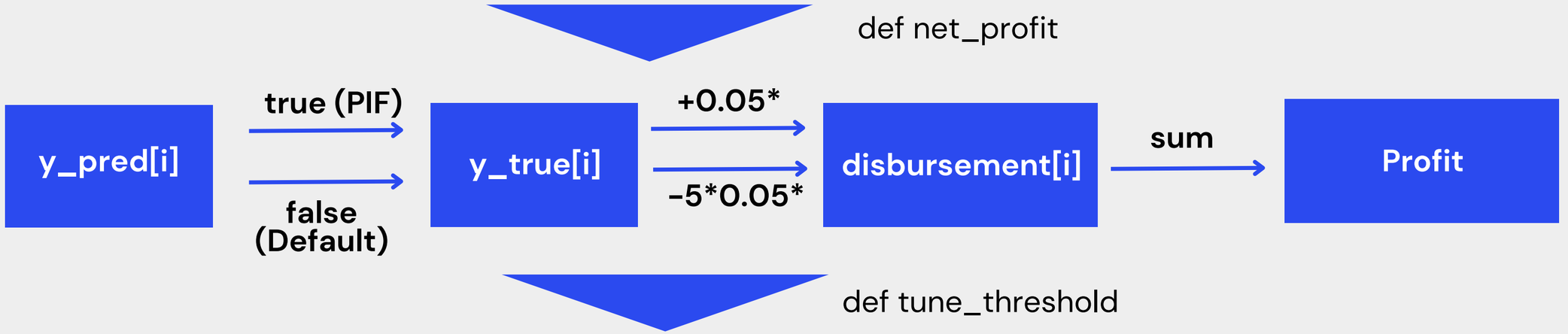
Model ROC AUC Plot *(for default threshold = 0.5)*



Set Up Cost Matrix

Average Net Profit (U.S. dollars)

Predicted (decision)	Actual: Paid in full	Actual: Default
Paid in full (grant the loan)	5% of DisbursementGross	-5 times 5% of DisbursementGross
Default (deny the loan)	0	0



Net Profit Result

Model	Optimal Threshold (P(Default)	Net Profit at optimal threshold	Net Profit at threshold = 0.5	Rate of Return
KNN	**0.21**	**$3.190B**		
Random Forest	**0.16**	**$3.764B**	**$3.436B**	**3.69 %**
Logistic Regression (ridge)	**0.24**	**$3.036B**	**$-1.048B**	3.57%
Logistic Regression (lasso)	**0.24**	**$3.036B**	**$-1,049B**	3.57%
Neural Network	**0.05**	**$3.118**	**$3.036B**	
XGBoost	**0.15**	**$3.808**	**$3.478**	

Model evaluation results *(for optimal threshold)*

Model	Accuracy
KNN	**0.8204**
Random Forest	**0.9151**
Logistic Regression (ridge)	**0.8399**
Logistic Regression (lasso)	**0.8399**
Neural Network	**0.8693**
XGBoost	**0.9244**

Accuracy Tradeoff for Profit

What can we learn from the results?

Primary Objective: Maximize Net Profit, Not Just Accuracy

Using the default threshold of 0.5 results in many missed high-risk loans, causing significant financial loss.

Threshold optimization based on expected profit leads to much better financial outcomes

Model Comparison Overview

1.  XGBOOST
2.  RANDOM FOREST
3.  NEURAL NETWORK
4.  KNN AND LR

Among all tested models, **XGBoost** delivers the best classification performance (ROC AUC ≈ 0.977) and generates the highest expected net profit (~$3.81B) when using an optimized threshold of 0.15

Complex, many trees, many hyperparameters, feature importance not always transparrent

SHAP (SHapley Additive exPlanations)

or Choose RF for better interpretation

Opportunities for further improvement

Refinements in modeling, feature engineering, and threshold strategies could further enhance profitability and predictive accuracy.

Modeling improvements

- **Tune** hyperparameters **more extensively**, especially tree depth, learning rate, and regularization parameters.
- Try **advanced ensemble methods** (e.g., stacking, blending) to leverage strengths of different models.
- Experiment with **cost-sensitive learning** by directly incorporate business profit-loss matrix into model training to better align with financial goals.
- **Adjust threshold optimization strategy** to achieve a better balance between false positives and false negatives.

Data and feature enhancements

- **Incorporate additional macroeconomic indicators** such as interest rates and unemployment rates to capture broader economic trends affecting loan defaults.
- **Apply more sophisticated feature selection techniques**: use methods like recursive feature elimination (RFE) or SHAP values to refine the feature set and enhance model interpretability.
- **Enrich dataset with rejected loans**: if available, add rejected loan applications to reduce selection bias and improve model generalization.

BANA4020

THANK YOU!

GROUP 4