

**SBA Loan Default Prediction and Analysis**

## Daivik Poonja (dpoonja@sfsu.edu)

**Amogh Ranganathaiah(aranganathaiah@sfsu.edu)**

**Ekta Singh(esingh@sfsu.edu)**

## Executive Summary

1. Introduction

This report investigates the factors influencing Small Business Administration (SBA) loan defaults and presents a data-driven framework to improve loan approval processes. By employing predictive modeling, the study identifies significant risk factors, enhances decision-making strategies, and aligns lending practices with regulatory and economic goals.

2. Objectives

* Determine the most significant predictors of SBA loan defaults.
* Build and validate a robust model to classify loan outcomes.
* Recommend strategies to improve profitability and reduce default risks while supporting small businesses.

3. Key Findings

Predictive Modeling Results for XGBoost Model:

- Accuracy: 99%

- Precision: 97%

- Recall: 99%

- F1-Score: 0.99

- Financial Impact: The model generated the highest profitability of $116M.

4. Influential Factors:

- Loan term (`Term`): Longer terms increase default risk due to prolonged exposure to economic volatility.

- Charged-off principal balance (`ChgOffPrinGr`): High charge-off balances correlate strongly with defaults.

- Geographic indicators (`UrbanRural`, `State`, `BankState`): Regional economic conditions and borrower locations significantly impact repayment behavior.

- Industry codes (`NAICS`): Industry-specific risks drive default probabilities.

- Temporal Features: Derived variables such as `DisbursementTime` and `ChgOffTime` provide insights into time-based patterns that influence defaults.

- Class Imbalance Handling: The application of SMOTE ensured balanced training data, improving model sensitivity to minority classes (defaults).

5. Financial Analysis:

A cost-benefit evaluation revealed that XGBoost not only minimized default-related losses but also maximized returns by targeting high-repayment borrowers, leading to sustainable lending practices.

6. Recommendations

a. Risk Mitigation:

1. Strategic Loan Approvals:

- Implement predictive models like XGBoost to flag high-risk loans for further review.

- Focus approvals on borrowers with high repayment potential.

2. Regional Risk Adjustments:

- Tailor lending policies to account for regional economic disparities, especially in rural and high-risk areas.

b. Regulatory Compliance:

1. Basel III Alignment:

- Use Probability of Default (PD) estimates to ensure adequate capital reserves and align with global banking standards.

2. Economic Stability:

- Incorporate industry-specific and temporal risk assessments to mitigate broader economic disruptions.

c. Promoting Small Business Growth:

1. Fair Lending Practices:

- Reduce biases in loan approvals to ensure equitable access to capital across industries and regions.

2. Data-Driven Insights:

- Leverage advanced analytics to enhance SBA’s dual mission of profitability and fostering small business growth.

7. Conclusion

This study provides actionable insights and a predictive framework to transform SBA’s lending strategy. By implementing XGBoost and addressing key risk factors, SBA can:

- Enhance accuracy in predicting loan defaults.

- Maximize financial returns while supporting small businesses.

- Ensure compliance with regulatory and macroeconomic goals.

Adopting these recommendations positions SBA to drive equitable and sustainable economic growth while safeguarding the financial stability of its loan programs.

## Introduction

**Purpose**

The report addresses the increasing challenges in managing SBA loan defaults by using advanced analytics to predict risk and optimize lending practices. The primary goal was to develop an accurate and interpretable model to predict SBA loan default probability (`MIS\_Status`). Key objectives included:

1. Identifying the most influential variables impacting loan default.

2. Optimizing a loan approval strategy using cost-sensitive metrics.

3. Providing actionable recommendations for banks to mitigate financial risks while supporting small businesses.

**Scope**

The study evaluates key predictors of loan outcomes, builds machine learning models to classify defaults, and develops recommendations for actionable strategies. The insights can help lenders reduce default-related losses and strategically allocate loan resources to improve financial returns, especially under economic constraints and regulatory frameworks (e.g., Basel III).

**Background**

As a critical source of funding for small businesses, SBA loans support economic growth and job creation. However, loan defaults impose financial risks. This study aims to mitigate these risks through data-driven strategies.

**Research Question**

What factors influence the likelihood of default for SBA loans, and how can these factors be utilized to predict loan outcomes while maximizing bank profitability?

**Hypotheses**

- Null Hypothesis (H₀): There is no significant relationship between variables such as loan disbursement times, borrower details, and loan features (e.g., term, low documentation) and the probability of loan defaults (MIS\_Status).

- Alternative Hypothesis (H₁): Specific factors such as loan disbursement times, borrower details, and loan features significantly impact loan default outcomes.

**Research Design**

The study utilized a combination of descriptive, correlation, and experimental designs to investigate SBA loans:

1.Descriptive Design: Analyzed key metrics (e.g., loan approval/disbursement times, charge-off trends) to understand basic patterns in the dataset.

2.Correlation Design: Explored relationships between numerical variables (e.g., loan amount, term) and outcomes (`MIS\_Status`).

**Experimental Design**

Employed machine learning classification models to predict loan statuses (`MIS\_Status`) and quantify the effect of variables on outcomes.

1. **Methodology**

Let us first start with Exploratory Data Analysis.

1. Loading and Inspecting the Dataset

The dataset was loaded using Python's pandas library and inspected for its shape, column names, and data types. It consists of 899,164 rows and 27 columns, representing information about loans, businesses, and loan statuses.

Missing values and data types were reviewed to identify cleaning and preprocessing needs.

1. Descriptive Statistics

Summary statistics such as mean, median, and standard deviation were calculated for numeric columns. These metrics provide insights into central tendencies, variability, and the presence of outliers in the data.

1. Visualization of Data Distributions

Histograms, bar plots, and boxplots were used to analyze the spread, skewness, and trends within the dataset.

Correlation Analysis

A correlation matrix and heatmap were generated to assess relationships among numeric variables and detect multicollinearity issues.

Category-Level Analysis

Bar plots and count plots were created for categorical variables to explore the frequency and distribution of values such as loan statuses and regional trends.

1. Key Observations and Insights from Visualizations

1. Histogram of Loan Amounts (DisbursementGross)

Purpose: To analyze the distribution of loan disbursement amounts.

Key Insights: Most loans are concentrated in lower ranges, indicating that smaller loans are more common.

The distribution is right-skewed, with a long tail representing high-value loans.

Significance: Helps in understanding the scale and skewness of loan amounts, which may require transformations for modeling.

Indicates the potential need to segment analysis based on loan size.

2. Bar Plot of Loan Status (MIS\_Status)

Purpose: To evaluate the distribution of loan outcomes (e.g., "Paid", "Default").

Key Insights: A majority of loans have been successfully paid, while a smaller portion is in default.

The imbalance in loan status is evident.

Significance: Highlights class imbalance in the target variable, necessitating techniques like resampling or cost-sensitive learning in predictive models.

Provides an initial understanding of success and risk rates.

3. Heatmap of Feature Correlations

a. Purpose: To examine relationships between numeric variables.

b. Key Insights:

* Strong Positive Correlation Between GrAppv and SBA\_Appv: These variables are closely related since SBA\_Appv represents the SBA-guaranteed portion of GrAppv. Including both in a model could introduce multicollinearity.
* Moderate Correlation Between DisbursementGross and GrAppv: Indicates that loan disbursements align with gross approved amounts, as expected.
* Weak Correlation Between Features and MIS\_Status: Suggests numeric features alone might not strongly predict loan outcomes, emphasizing the importance of categorical variables and feature interactions.

c. Significance: Guides feature selection by identifying redundant or correlated variables.Encourages further exploration of categorical features and engineering derived features for better predictive power.

4. Boxplot of Loan Amounts by Loan Status

a. Purpose: To compare loan amounts across different loan statuses.

b. Key Insights:

- Defaulted loans tend to have higher median loan amounts compared to loans marked as "Paid".

-Outliers are present, especially in defaulted loans.

c. Significance: Indicates that larger loans may carry a higher risk of default, suggesting loan size as a key factor in risk assessment. Provides insights for segmentation or risk stratification during modeling.

5. Count Plot of States

a. Purpose: To visualize the geographic distribution of loans.

b. Key Insights:

-States like California and Texas have the highest number of loans, reflecting their large economic activity and business density.

- Less populous states exhibit significantly lower loan activity.

c. Significance: Highlights regional variations, which may influence loan approval strategies and business opportunities.Provides a basis for analyzing localized factors affecting loan performance.

Correlation Matrix Observations:

The correlation matrix revealed the following critical insights about numeric variables:

1.Significant Features:

a. GrAppv (Gross Approval Amount) and SBA\_Appv (SBA-Guaranteed Amount) exhibit a strong positive correlation, suggesting redundancy.

b. DisbursementGross shows moderate correlation with GrAppv, indicating its potential as a predictor of loan success or risk.

2.Low Correlations with Target Variable:

a. Most numeric features have weak correlations with MIS\_Status (loan status), which emphasizes the need for:

-Including categorical variables like State, UrbanRural, and Industry.

-Exploring interactions or non-linear relationships.

b. Findings

-Loan amounts are heavily skewed, with larger loans showing a higher risk of default.

The target variable (MIS\_Status) exhibits significant class imbalance, requiring thoughtful handling during modeling.

-Strong correlations among certain numeric features indicate the potential for multicollinearity, while weak correlations with the target variable highlight the importance of categorical data and feature engineering.

-Regional and categorical insights underline the importance of geographic and industry- level analysis.

**C. Data Pre-processing**

1. Loading and Inspecting Data:

1. The dataset SBAnational.csv was loaded and inspected to understand its structure and content. Preliminary observations included:
2. Categorical columns: Variables like State, BankState, RevLineCr, UrbanRural, State, NAICS and LowDoc.
3. Numerical columns: ChgOffPrinGr, Terms, DisbursementGross and principal balance.
4. Target variable: MIS\_Status, indicating whether a loan was repaid or charged off.

2. Feature Engineering:

1. Temporal variables such as DisbursementTime, ChgOffTime, and DisbursementToChgOffTime were created. These features quantify the duration between key events like loan approval, disbursement, and charge-off, capturing important time dynamics affecting loan outcomes.
2. The newly calculated columns (DisbursementTime, ChgOffTime, DisbursementToChgOffTime) serve as derived features that provide additional context about loan behavior over time.
3. These features can significantly enhance the predictive power of models by capturing time-based patterns.
4. Time intervals between loan approval, disbursement, and charge-off can be key indicators of loan performance.
5. These time-based features are critical inputs for models predicting loan defaults (MIS\_Status), as they capture temporal dynamics that static features (e.g., loan amount, interest rate) cannot.

3. Data Cleaning:

1. LoanNr\_ChkDgt: Identifier for the loan number, which doesn't provide meaningful information for analysis.
2. Name, City, Zip, Bank: These columns might contain sensitive or irrelevant information for modeling (e.g., personal identifiers). Removing irrelevant or redundant columns reduces the dimensionality of the dataset, which can improve the performance of machine learning models and simplify exploratory data analysis.
3. ApprovalDate, ChgOffDate, DisbursementDate: These columns were used earlier for calculating derived features (DisbursementTime, ChgOffTime, DisbursementToChgOffTime). Since the raw dates are no longer required, they can be dropped to avoid redundancy.
4. The sba\_loans DataFrame is streamlined, containing only columns that are essential for further analysis or modeling.
5. Non-numeric categorical values (NewExist, RevLineCr, LowDoc) were converted into numeric formats for compatibility with machine learning algorithms.
6. Missing and invalid entries in critical columns were removed to ensure data quality.

4. Feature Scaling:

1. Machine learning models in scikit-learn cannot work directly with categorical string data. Label encoding converts these strings into numeric representations, making them suitable for modeling.
2. Encoding ensures all categories are uniformly represented as integers, avoiding issues that may arise from mixed data types.
3. Although LabelEncoder assigns random labels, it is sufficient for columns like State, BankState, and NAICS, where no natural order exists.

5. Statistical Analysis:

a. Correlation Analysis:

* Correlation analysis was performed to evaluate the linear relationship between predictors and the target variable. Key takeaways:
* High Correlation: Features like ChgOffPrinGr (Charged-Off Principal Balance) and Term exhibited significant correlations with MIS\_Status.
* A higher charge-off principal balance increases the likelihood of loan default.
* Longer loan terms are associated with a higher risk of default, reflecting potential economic volatility over extended periods.
* Weak Correlations: Features like SBA\_Appv, GrAppv, DisbursementGross, and DisbursementTime have weak correlations but might still be useful.
* Very Weak or No Correlation: Features like NoEmp, CreateJob, RetainedJob, BalanceGross, and ChgOffTime are weakly correlated and may have minimal impact on predicting MIS\_Status.

2. ANOVA (Analysis of Variance):

* ANOVA was used to assess the statistical significance of numerical features in distinguishing between the loan repayment statuses (MIS\_Status).
* Significant Features:DisbursementTime and ChgOffPrinGr were statistically significant, indicating their strong influence on loan outcomes.
* Interpretation:The time between loan disbursement and repayment/charge-off plays a crucial role in understanding default behavior.
* The charged-off principal highlights the financial burden of failed loans.

3. Chi-Square Test:

* Chi-Square tests evaluated the dependency between categorical variables and the target variable (MIS\_Status).
* Significant Variables:All listed features are statistically significant, with p-values below 0.05, indicating that there is a meaningful association with the target variable, MIS\_Status.
* The highest Chi2 statistics are seen with UrbanRural, RevLineCr, and NewExist, suggesting these may be particularly strong predictors of loan risk.
* Revolving credit and low documentation loans are inherently riskier, especially in volatile economic conditions.
* Urban versus rural location differences reflect regional economic disparities and their impact on loan performance.
* Most Important Feature: BankState has the highest F-statistic among these features, suggesting it might be the most significant predictor among them.
* Significant Features: State, BankState, and NAICS are all statistically significant, with p-values of 0.

**The features that are going to highly influence Probability of Default:**

1. ChgOffPrinGr (Num):A high charge-off principal balance is a direct measure of loan defaults and reflects broader economic conditions like unemployment, market volatility, or financial crises. During recessions or economic downturns, businesses often struggle to generate sufficient revenue to cover their liabilities, resulting in higher charge-off balances.

2. Term (Num)- By including loan duration in predictive models, we can capture the risk associated with long-term economic variability. Loans with longer terms are more likely to default due to their prolonged exposure to unfavorable economic events

3. UrbanRural (Cat)- This categorical variable provides insights into regional economic health and its impact on loan performance. Understanding whether the borrower is in an urban or rural area helps capture location-based risks in the prediction model.

4. RevLineCr (Cat)- Revolving credit signals borrowers’ liquidity management. During economic downturns, businesses relying heavily on revolving credit may face increased difficulty maintaining cash flow due to tightened credit conditions or rising interest rates.

5. LowDoc (Cat)- Low documentation loans, often granted during periods of economic optimism or credit booms, are inherently riskier due to minimal financial scrutiny, leaving borrowers vulnerable during economic contractions. This feature serves as a strong risk signal, especially when combined with factors like high loan amounts or weak industry performance, enhancing its predictive power for defaults.

6. BankState (Cat)- A bank’s location reflects state-specific economic conditions and policies, with stronger economies linked to lower default rates. This feature captures geographic variations, enhancing the model’s ability to predict defaults accurately.

7. State (Cat)- Borrower state reflects regional economic conditions that directly affect businesses. For instance, states reliant on oil or agriculture may experience higher default rates during commodity price downturns. Conversely, states with diversified economies may show better loan performance.

8. NAICS (Cat)- Understanding industry-specific risks allows a more nuanced assessment of default probability. The NAICS code is a valuable predictor as it enables the model to associate industry risk profiles with default behavior, especially under macroeconomic stressors like recessions or sectoral downturns.

9. We also need DisbursementGross as need that to calculate net profit of the model.

After data pre-processing we have created a file named **selected\_sba\_loans\_data.csv** in which we have included the final cleaned dataset which is ready for data modelling and interpretations.

**D. Data Modelling**

1. Data Loading and Initial Exploration

The first step involves loading the dataset using the `pandas` library. The dataset, named `selected\_sba\_loans\_data.csv`, contains information on loan details and statuses. Once the data is loaded, a quick review of its structure reveals that it has 613,723 rows and 10 columns. These columns represent features such as loan disbursement amounts, loan terms, and the target variable `MIS\_Status`, which indicates whether a loan was paid (`0`) or defaulted (`1`).

2. Feature Selection and Train-Test Split

* Feature selection involves choosing the most relevant columns for the machine learning task. In this case, the columns `ChgOffPrinGr`, `Term`, `UrbanRural`, `RevLineCr`, `LowDoc`, `BankState`, `State`, and `NAICS` are selected as predictors (`X`), while the column `MIS\_Status` serves as the target variable (`y`). The dataset is then split into training and testing subsets, using 80% of the data for training and 20% for testing.
* The column `DisbursementGross` is excluded from modeling but preserved for later evaluation of predictions on the test set.
* A balanced distribution between training and test sets is ensured by using random splitting with a fixed seed.

The separation of training and testing data ensures that models are evaluated on unseen data, preventing overfitting and providing a more realistic measure of model performance.

3. Addressing Class Imbalance

* The target variable `MIS\_Status` is imbalanced, with far fewer loans classified as defaults (`1`) compared to those paid off (`0`). To address this, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE generates synthetic examples of the minority class in the training dataset, making the classes balanced and improving model sensitivity to the minority class.
* The training data becomes balanced, with an equal number of samples for both `MIS\_Status = 0` and `MIS\_Status = 1`.
* Balancing the data ensures that machine learning models are not biased toward the majority class and are better equipped to identify defaults (`1`).

4. Models

**Model 1:** k-Nearest Neighbors (kNN)

kNN classifies loans based on similarity to their neighbors using a Manhattan distance metric. This distance metric calculates absolute differences between feature values, which makes it more robust to outliers compared to Euclidean distance. However, the performance of kNN is heavily influenced by the choice of the number of neighbors (k) and the scale of the data.

Output:

Accuracy: 92%

Precision (Default): 79%

Recall (Default): 79%

Weighted F1-score: 0.92

Net Profit: - $49,987,428.50

Observation:

While kNN achieves moderate accuracy, its reliance on local patterns rather than global trends leads to misclassification in high-dimensional data. This limitation results in a significant number of false positives, directly impacting the net profit. The simplicity of the model makes it less effective in capturing complex relationships within the dataset, thereby limiting its utility for SBA's financial risk assessment.

**Model 2:** Decision Tree Classifier

Decision trees split the data iteratively based on feature thresholds, creating an interpretable tree-like structure. This model is highly effective for datasets with strong feature interactions, as it can capture both categorical and continuous variable splits.

Output:

Accuracy: 99%

Precision (Default): 100%

Recall (Default): 99%

Weighted F1-score: 0.99

Net Profit: $116,307,604.50

Observation:

Decision trees provide excellent interpretability, allowing SBA stakeholders to clearly understand the factors driving high-risk loan classifications. For example, splits on features like UrbanRural or LowDoc can reveal critical insights into which loans are more likely to default, helping to refine lending criteria for different borrower profiles.

**Model 3:** Random Forest Classifier

Random forests build multiple decision trees and aggregate their predictions, reducing variance and improving generalization.

Output:

Accuracy: 99%

Precision (Default): 100%

Recall (Default): 99%

Weighted F1-score: 0.99

Net Profit: $113,016,941.00

Observation: Random forests provide consistent results but slightly lower profitability compared to decision trees.

**Model 4:** Bagging Classifier

Bagging, or Bootstrapped Aggregating, combines predictions from multiple models trained on different bootstrapped samples of the dataset. This technique helps reduce variance by ensuring each model sees slightly different data, which improves stability and generalization. Bagging classifiers are particularly useful in mitigating overfitting for decision tree-based models.

Output:

Accuracy: 99%

Precision (Default): 100%

Recall (Default): 99%

Weighted F1-score: 0.99

Net Profit: $116,307,604.50

Observation:

Bagging’s ensemble approach mirrors the need for diversified risk assessment in SBA’s portfolio. The stability provided by aggregating predictions across multiple models makes it highly effective for managing diverse loan scenarios under varying macroeconomic conditions.

**Model 5:** Gradient Boosting Classifier

Description: Sequentially builds models to reduce prediction errors iteratively.

Output:

Accuracy: 99%

Precision (Default): 100%

Recall (Default): 99%

Weighted F1-score: 0.99

Net Profit: $116,307,604.50

Observation: Effectively captures complex relationships in data.

**Model 6:** XGBoost

XGBoost is an advanced gradient boosting algorithm that incorporates regularization (L1 and L2) and efficient tree-pruning strategies. It employs parallel processing for faster computation and manages missing data effectively, making it highly adaptable to large datasets like SBA loans. These optimization techniques allow the model to capture nuanced relationships in the data while avoiding overfitting.

Output:

Accuracy: 99%

Precision (Default): 100%

Recall (Default): 99%

Weighted F1-score: 0.99

Net Profit: $116,372,293.10

Observation:

XGBoost’s ability to balance precision and recall while maximizing profitability makes it the best-performing model. Its regularization techniques ensure stability, while its speed and efficiency enable handling of complex patterns, such as the interaction between loan disbursement amounts, geographic disparities, and industry-specific risks. This adaptability aligns perfectly with SBA’s goal to support small businesses while minimizing financial losses.

**Model 7:** Logistic Regression

Logistic regression is used to model the probability of loan defaults as a linear function of predictors. In this analysis, three variations of logistic regression are employed:

* Ridge (L2 Regularization): Penalizes large coefficients to prevent overfitting, particularly useful for datasets with multicollinearity.
* Lasso (L1 Regularization): Shrinks less important coefficients to zero, effectively performing feature selection.
* Elastic Net (Combination of L1 and L2 Regularization): Balances feature selection and coefficient shrinkage for improved performance.

Output:

* Ridge Logistic Regression:

Accuracy: 99%

Precision (Default): 98%

Recall (Default): 95%

Weighted F1-score: 0.97

Net Profit: $115,629,050.45

* Lasso Logistic Regression:

Accuracy: 98%

Precision (Default): 98%

Recall (Default): 93%

Weighted F1-score: 0.96

Net Profit: $105,470,272.90

* Elastic Net Logistic Regression:

Accuracy: 98%

Precision (Default): 98%

Recall (Default): 93%

Weighted F1-score: 0.96

Net Profit: $105,065,693.35

Observation:

* Ridge regression provides the most robust performance among the three, balancing accuracy and financial outcomes.
* Lasso and Elastic Net regression offer insights into feature importance but slightly underperform in terms of net profit.

These models are simpler implement, making them suitable for baseline comparisons and quick evaluations.

**Model 8:** Neural Network

The neural network implemented here is a multi-layer perceptron (MLP) model designed to capture complex, non-linear patterns in the data. It includes multiple hidden layers with ReLU (Rectified Linear Unit) activation functions to introduce non-linearity. The model is trained using the Adam optimizer, which efficiently adjusts learning rates for better convergence, and employs a categorical cross-entropy loss function suitable for binary classification tasks.

Hyperparameter Tuning:

* Hidden layers: Adjusted for depth and complexity.
* Neurons per layer: Tuned to balance model capacity and training time.
* Learning rate: Optimized for faster convergence without overshooting.
* Batch size: Set to handle the dataset size efficiently without memory bottlenecks.

Output:

Accuracy: 99%

Precision (Default): 100%

Recall (Default): 98%

Weighted F1-score: 0.99

Net Profit: $116,232,116.80

Observation:

Neural networks excel in capturing non-linear interactions among features, such as complex relationships between loan amount, term, and geographic risk. Despite their computational intensity, their robust performance and adaptability make them valuable for nuanced analyses like SBA’s loan evaluations. However, the need for computational resources and expertise could limit their deployment in real-time decision-making.

**Model 9**: Discriminant Analysis

Linear Discriminant Analysis (LDA): LDA assumes that the data points for each class are normally distributed and shares the same covariance structure. It finds a linear combination of features that best separates the classes by maximizing the ratio of between-class variance to within-class variance.

Quadratic Discriminant Analysis (QDA): QDA is a more flexible version of LDA that allows each class to have its own covariance matrix, enabling it to model more complex decision boundaries for non-linear separability.

Output:

* LDA:

Accuracy: 80%

Precision (Default): 46%

Recall (Default): 90%

Weighted F1-score: 0.82

Net Profit: - $729,096,566.95

Observation:

LDA struggles with this dataset, primarily due to its linearity assumption, leading to significant misclassifications and financial losses.

* QDA:

Accuracy: 97%

Precision (Default): 91%

Recall (Default): 92%

Weighted F1-score: 0.97

Net Profit: $65,232,514.65

Observation:

By relaxing the linearity assumption, QDA achieves higher accuracy and financial performance, making it more suitable for this dataset compared to LDA.

Key Insights:

* LDA is less effective for datasets with non-linear separability due to its strict linearity assumption.
* QDA’s flexibility allows it to better capture the complexities of the data, though it is computationally more intensive and requires careful handling of covariance matrices.

5. Highest Net Profit

Net profit evaluates the financial impact of each model by combining the benefits of correctly identifying defaults (True Positives) with the costs of misclassifying loans (False Positives).

* XGBoost achieves the highest net profit of $116,372,293.10, making it the most effective model for optimizing loan approvals and minimizing losses.
* Gradient Boosting, Bagging, and Decision Trees closely follow, with profits of $116,307,604.50, while Neural Networks also perform well at $116,232,116.80.
* Logistic Regression (Ridge) provides $115,629,050.45, balancing simplicity and performance.
* However, simpler models like kNN (-$49,987,428.50) and Linear Discriminant Analysis (-$729,096,566.95) incur significant losses, demonstrating their unsuitability for this dataset.

These insights highlight the importance of selecting advanced ensemble methods to maximize profit.

6. Graph Analysis

a. Gains Chart

* The gains chart shows the cumulative percentage of true positives (correctly predicted defaults) as a function of the percentage of the population ranked by the model’s predicted probability of default.
* The steeper the curve, the more effective the model is at identifying defaults in the top percentages of high-risk loans.

Insights:

* Models like XGBoost and Gradient Boosting achieve sharp gains early in the curve, indicating their ability to capture a significant proportion of defaults within the top deciles of high-risk loans.
* Logistic regression produces a more gradual curve, reflecting less focus on prioritizing high-risk cases.
* The chart demonstrates that advanced models like XGBoost are highly effective in targeting the riskiest loans, a critical requirement for SBA’s financial risk management.

b. Lift Chart

The lift chart compares the model’s ability to identify defaults against a baseline of random selection. A higher lift value in the initial deciles indicates better prioritization of high-risk loans.

Insights:

* XGBoost and Gradient Boosting achieve the highest lift values in the top deciles, demonstrating superior performance in identifying high-risk loans compared to other models.
* Lift values diminish in the later deciles, indicating reduced returns as the model reaches lower-probability cases.
* This chart underscores the practical advantage of advanced models in focusing resources on high-risk loans.

c. Net Profits by Model

This bar chart illustrates the net profit achieved by each machine learning model. It evaluates the financial impact of model predictions by combining the benefits of correctly identifying defaults with the costs of false positives.

Insights:

* XGBoost delivers the highest net profit (
* LDA incurs the most significant financial loss, emphasizing its unsuitability for this dataset.
* The chart highlights the importance of selecting models that balance prediction accuracy with financial considerations.

d. Max Profits by Model

This bar chart shows the maximum profits achieved by various classification methods applied to a dataset.

Insights:

Ensemble-based methods (like Decision Tree, Random Forest, Bagging, and Boosting) are the most effective in maximizing profits for this classification task, while KNN and Linear Discriminant Analysis are the least effective. The results highlight the strength of advanced or ensemble learning techniques in optimizing outcomes.

**E. Conclusion**

**Key Achievements:**

The analysis of SBA loan defaults was systematically approached using a combination of exploratory data analysis, feature engineering, and predictive modeling. This comprehensive methodology addressed the challenges of identifying factors influencing loan defaults and developed a reliable framework to predict loan outcomes.

1. Addressing Class Imbalance:

The imbalance in loan statuses (paid vs. default) was mitigated using SMOTE, which ensured balanced training data. This step significantly improved the sensitivity of models to defaults, allowing for better prediction of high-risk loans.

2. Feature Selection and Engineering:

Key features such as ChgOffPrinGr, Term, and categorical variables like UrbanRural and NAICS were identified as critical predictors. Derived time-based features (e.g., DisbursementTime, ChgOffTime) enhanced the models by capturing temporal dynamics in loan performance.

These features allowed models to account for economic factors such as regional disparities, industry-specific risks, and temporal changes in loan behavior.

3. Model Selection and Optimization:

a. Multiple machine learning models were evaluated, including kNN, Decision Trees, Random Forest, Gradient Boosting, and XGBoost.

b. XGBoost emerged as the best-performing model, achieving a high accuracy of 99%, with precision (97%), recall (99%), and F1-score (0.99). It provided the most reliable predictions for defaults, balancing false positive and false negative rates.

4. Financial Impact Analysis:

A cost-benefit analysis quantified the financial implications of model predictions. XGBoost achieved the highest profitability of $116M by accurately classifying high-risk loans, reducing default-related losses while maximizing returns on approved loans.

**Business Implications:**

1. Enhanced Risk Management by Strategizing Loan Approvals:

By identifying high-risk loans with greater accuracy, the SBA can refine its loan approval process, focusing on borrowers with a higher probability of repayment. This reduces financial exposure to defaults.

2. Region-Specific Insights:

Variables like UrbanRural, State, and BankState enable the SBA to tailor its strategies based on geographic risk. For instance, rural areas with higher default probabilities may benefit from additional support mechanisms or stricter credit assessments.

3. Compliance with Basel III Framework:

The predictive model aligns with the Basel III guidelines by providing accurate Probability of Default (PD) estimates. This ensures adequate capital reserves are maintained, safeguarding the banking system under economic stress.

4. Macroeconomic Resilience:

Features like NAICS and Term help assess industry-specific risks and long-term loan exposures, aligning SBA policies with broader economic stability goals.

5. Support for Small Business Growth

By leveraging models to minimize biases, the SBA can ensure fair access to loans across diverse industries and regions, supporting small businesses while reducing systemic risks.

6. Data-Driven Decision Making:

The insights derived from the analysis empower SBA to make informed decisions that balance financial returns with its mission to promote small business growth and job creation.