**SBA Loan Default Prediction and Analysis**

1. Designing the Study

Research Question

The central question of this study is:

\*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.

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

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

Objective:

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.

Significance:

This study's 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).

2. Data Collection and Characteristics

Dataset Overview

The dataset (`SBAnational.csv`) contained \*\*899,164 observations\*\* and \*\*27 columns\*\*, representing SBA loan data from 1987 to 2014. Key fields included:

- Loan Status: `MIS\_Status` (Charged Off vs. Paid in Full)

- Loan Amounts: `DisbursementGross`, `BalanceGross`, `ChgOffPrinGr`

- Temporal Variables: `ApprovalDate`, `DisbursementDate`, `ChgOffDate`

- Borrower Information:`UrbanRural`, `FranchiseCode`, `RevLineCr`, `LowDoc`

Initial Data Challenges

- Missing Values: Several columns had incomplete data.

- Inconsistent Formats: Dates were stored as strings, while financial fields contained non-numeric symbols (e.g., `$`, `,`).

- Mixed Data Types: Key variables were improperly classified (e.g., financial data as objects).

Preprocessing Step

1. Date Conversion:

- Converted date columns (`ApprovalDate`, `DisbursementDate`, `ChgOffDate`) into datetime formats.

- Removed irrelevant columns after feature engineering.

2. Handling Missing Data:

- Imputed missing values or removed records with insufficient information to avoid bias in the analysis.

3. Encoding Categorical Variables:

- Encoded `MIS\_Status` as binary (1 = Default, 0 = Paid in Full).

- Transformed `FranchiseCode` into a binary variable (franchise vs. non-franchise).

- Label-encoded variables like `LowDoc` and `RevLineCr`.

4. Standardization of Numerical Features:

- Used `StandardScaler` to normalize features like `DisbursementGross` and `Term`.

Cleaned Dataset

The final dataset retained only meaningful variables:

- `ChgOffPrinGr` (Charge-Off Principal)

- `Term` (Loan Term)

- `LowDoc` (Low Documentation Indicator)

- `RevLineCr` (Revolving Credit Line Indicator)

- `UrbanRural` (Urban/Rural Location)

3. Description of Data

Exploratory Data Analysis (EDA)

1. Disbursement Trends:

- Average loan disbursement: \$350,000

- Default rate: ~16% of all loans

2. Key Predictors:

- Loans with longer terms were associated with higher default rates.

- Loans with higher `ChgOffPrinGr` (charge-off principal) showed stronger likelihood of default.

- `LowDoc` loans were more likely to default.

3. Correlation Analysis:

- `ChgOffPrinGr` had the highest correlation with loan defaults.

- `Term` showed moderate positive correlation with defaults.

Statistical Validation

- Chi-Square Test Results:

- Categorical variables (`LowDoc`, `RevLineCr`) were significantly associated with default outcomes.

- T-Test Results:

- Mean differences in numerical variables (e.g., `DisbursementGross`) between default and non-default groups were statistically significant.

4. Model Development and Evaluation

Predictive Models

1. Logistic Regression (Lasso, Ridge, ElasticNet):

- Pros: Easy to interpret, efficient.

- Cons: Lower accuracy compared to other models.

2. Decision Trees & Bagging Models (Random Forest):

- Pros: Captures complex interactions, handles non-linear data.

- Cons: Computationally intensive.

3. Gradient Boosting (XGBoost):

- Pros: Achieved the highest accuracy (~99.4%).

- Cons: Longer training time.

4. Neural Networks

- Pros: Handles high-dimensional data well.

- Cons: Requires careful tuning, less interpretable.

5. Discriminant Analysis (LDA, QDA):

- Pros: Effective for linear separability.

- Cons: Struggles with non-linear relationships.

Model Performance Metrics

| Model | Accuracy | Precision Recall | ROC-AUC|

|-------------------------|--------------|---------------|------------|-------------|

| Gradient Boosting | 99.4% | 98.7% | 99.0% | 0.996 |

| Random Forest | 99.1% | 97.8% | 98.4% | 0.994 |

| Logistic Regression | 92.3% | 85.1% | 87.4% | 0.903 |

Cost-Sensitive Analysis

- Profit for "Paid in Full": +5% of disbursed amount.

- Loss for "Default": -5× disbursed amount.

The Gradient Boosting model maximized net profit by minimizing false negatives (defaulted loans incorrectly classified as low risk).

5. Interpretation and Recommendations

Key Insight

1. Significant Predictors:

- `ChgOffPrinGr`: Loans with higher charge-off principals had higher default probabilities.

- `Term`: Longer-term loans posed higher risks.

- `LowDoc`: Loans with low documentation requirements were riskier.

2. Optimal Threshold

- Threshold of 0.25 maximized bank profitability by balancing sensitivity and specificity.

Business Implications

1. Deploy Gradient Boosting:

- Integrate Gradient Boosting into the loan approval pipeline for real-time decision-making.

2. Risk Classification:

- Flag high-risk loans for further scrutiny.

3. Policy Refinement:

- Use insights to reassess low documentation loan policies.

**Executive Summary**

Project Name: SBA Loan Default Prediction for Profit Maximization

Problem Statement:

Banks face challenges in mitigating risks associated with SBA-guaranteed loans. This project aimed to develop a predictive model to estimate the Probability of Default (PD), optimizing loan approval strategies to maximize profits.

Key Findings:

1. Gradient Boosting achieved the highest accuracy (99.4%) and net profitability.

2. Significant predictors included `ChgOffPrinGr`, `Term`, and `LowDoc`.

Business Recommendations:

- Implement Gradient Boosting in loan processing workflows.

- Regularly update the model with new data.

- Adjust policies for high-risk loans, especially those with long terms or low documentation requirements.

Next Steps:

1. Develop interactive dashboards for visualizing loan risks.

2. Extend analysis to include macroeconomic factors.

3. Provide training to staff on leveraging predictive models in decision-making.

This project empowers banks to make data-driven, cost-effective decisions while fostering small business growth.