**Report-SBA Loan Prediction Project**

Procedure:-

1.The first step involves loading data and performing an initial inspection. The process involves importing libraries (pandas and numpy) for data manipulation. The SBAnational dataset was loaded. through a DataFrame. Dot info() and .head() were used to examine the data in order to find missing values, column types, and dataset structure. The dataset, which includes fields like ApprovalDate, MIS\_Status, and DisbursementGross, has 899,164 rows and 27 columns. Disparities were noted, including dates stored as strings and numerical data stored as objects. Missing values in several columns were found. Significance: - Contributed to a fundamental comprehension of the dataset. - Draw attention to preprocessing requirements like handling missing data and type conversion.

2. In order to maintain consistency, the date columns (ApprovalDate, DisbursementDate, and ChgOffDate) were converted to datetime format in Step 2: Handling Date Variables. After determining that these columns were not directly related to the prediction, they were dropped. Consistent date formats that are prepared for time-based analyses when necessary are the outputs. Significance: Data was made simpler by eliminating unnecessary columns. Verified that temporal features could be extracted for upcoming analyses if necessary.

3.Label encoding was used for variables like State, MIS\_Status, LowDoc, and RevLineCr in step three, which involves encoding categorical variables. - Mapped particular categories: - MIS\_Status: converted 'P I F' to 0 (full payment) and 'CHGOFF' to 1 (default). The franchise code has been converted to binary (franchise vs. non-franchising. In order to facilitate modeling, categorical variables were encoded into numerical formats. The use of categorical data in machine learning algorithms was made easier. The essential data required to forecast loan results was preserved.

4.The fourth step, cleaning financial data involves eliminating non-numeric characters (e.g. G. , $,,) from financial fields such as ChgOffPrinGr, BalanceGross, and DisbursementGross. These fields were changed to floating-point values. The output drawn from financial metrics show that loan balances and amounts were cleaned. It improved data precision for modeling and analysis. Issues brought on by mixed data types in crucial fields were resolved.

5.In the next step, we standardized numerical features. The procedure is as follows: - Standardized numerical features (e.g. A. , Term, and DisbursementGross) with StandardScaler. As a result, characteristics normalized to a variance of 1 and a mean of 0. Better compatibility with machine learning models that are sensitive to feature scaling, like gradient boosting and neural networks, is significant.

6. Calculated correlations between the target variable (MIS\_Status) and numerical features in step six of the correlation analysis process. Strong predictors were found, including: -

 ChgOffPrinGr (Charge-Off Principal): This variable has a strong positive correlation with loan default. The longer the term, the greater the default risk. Importance: - The selection of features for modeling was guided by the highlighted influential predictors.

7. Statistical Tests:  T-tests: Numerical feature means were compared across default categories. Using chi-square tests, relationships between loan outcomes and categorical variables (LowDoc, RevLineCr) were evaluated. LowDoc and RevLineCr were significant predictors, showing strong correlations with loan status, according to the outputs. It offered statistical support for the selected attributes. This guaranteed significant inputs for predictive modeling.

8. Final Dataset Preparation: - Key features: UrbanRural, LowDoc, RevLineCr, Term, and ChgOffPrinGr were chosen. The cleaned dataset was saved with the name selected\_sba\_loans\_data.csv. The results include a small dataset that has been optimized for modeling. It enhanced computational efficiency due to reduced dimensionality. Only significant features for prediction were kept.

**Model Development and Evaluation:**

9: For model implementation, the following classification models were performed:

- kNN

- Decision Tree, Bagging, Random Forest, Boosting

- Logistic Regression (Lasso, Ridge, ElasticNet)

- Neural Networks

- Discriminant Analysis (LDA, QDA)

10.Cost-sensitive metrics were incorporated. The profit margin for "Paid in Full" loans was 5% of the total amount disbursed. The loss for "Default" loans is -5× the amount that was disbursed.

11.Calculated accuracy, precision, recall, F1-score, and ROC-AUC are the validation metrics . Performance was visualized through the use of lift and gains charts.

Key Findings and Recommendations:

1. Model Performance: Random Forest and Gradient Boosting had the best accuracy (~99.4%). Although it was less accurate, logistic regression was still interpretable.
2. Cost-Sensitive Insights: By reducing false negatives, gradient boosting produced the largest net profit.
3. The ideal threshold was set at 0 to 25 in order to maximize net profit while keeping sensitivity and specificity within acceptable bounds.

Implications for Business:

1. Best Model: The most successful model was found to be gradient boosting.
2. Important features included ChgOffPrinGr, Term, LowDoc, and RevLineCr, these were key predictors.
3. Incorporate the Gradient Boosting model into the loan approval pipeline.
4. Use the optimized threshold to classify loans as "high risk" or "low risk."
5. Regularly retrain the model with updated data to ensure reliability.