Enhancing Loan Approval Precision: A Machine Learning Approach to Identify and Mitigate Default Risks in Financial Institutions

By Group: CS9

Aditya Gupta 210101009 Swagat Sathwara 210101102 Gautam Juneja 210101041 Shivam Gupta 210101114

1. Introduction

1.1. Target problem and Motivation

This project aims to create a strong predictive model to identify potential loan defaulters. Financial institutions face a crucial need to reduce lending risks, given the widespread issue of loan defaults affecting their financial stability. The motivation behind this project is to use advanced machine learning techniques to improve the efficiency of loan approval processes. By doing so, we seek to empower financial institutions to make well-informed decisions and decrease the chances of approving loans for individuals with a higher risk of default.

1.2. Major challenges

The major challenge would be imbalanced data which may lead the model to be biased. Other challenges would be implementing the algorithms and integrating them to work on our proposed data set. Another challenge includes preprocessing the data and tuning the learning rate to achieve the best possible results.

1.3. Outline

Through comprehensive data exploration, preprocessing, and feature engineering, the project will implement and compare three models—Logistic Regression, Naive Bayes, and Random Forest evaluating their performance and interpretability.

2. Methods

This report focuses on the application of three main ML algorithms: Logistic Regression, Naive Bayes, and Random Forest, in predicting loan defaults. These models are chosen for their diverse approaches to learning and prediction - Logistic Regression for its robustness in binary outcomes, Naive Bayes for its efficiency with probabilistic assumptions, and Random Forest for their interpretability and handling of non-linear relationships.Additionally we may explore hyperparameter tuning such as grid search, hit and trial ,Bayesian Optimization etc .

2.1. Exploratory Data Analysis

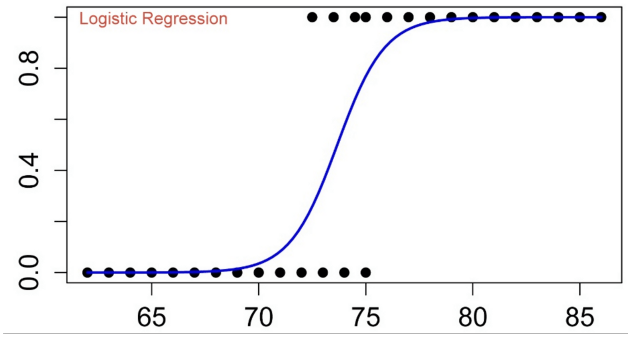
- Target Variable Analysis : Check for class imbalance and consider strategies for it.
- Feature Distributions : Analyze the distribution of each feature to understand their ranges and identify potential outliers.
- Data Visualization : Use visualizations such as correlation matrices to explore relationships between features.
- Data Preprocessing Decisions : Based on the insights gained during EDA, make decisions on data preprocessing steps, like encoding categorical variables, and scaling features.

2.2. Logistic Regression

Logistic Regression is a statistical model for predicting the probability of a binary outcome based on several predictor parameters.

$$P(y = 1|X) = \sigma(w^T X + b) = \frac{1}{1 + e^{-(w^T\ X\ +\ b)}} \quad (1)$$

In this equation : $P(y = 1|X)$ is the probability that the target variable y is 1, given the input features X.



2.3. Naive Bayes'

Naive Bayes is a probabilistic machine learning algorithm based on applying Bayes' theorem with a **"naive"** assumption of conditional independence between every pair of features given the class variable. The basic equation of Naive Bayes, derived from Bayes' theorem, is given by:

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|y) \times P(y)}{P(x_1, x_2, \ldots, x_n)} \quad (2)$$

Under the naive independence assumption, the equation simplifies to:

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(y) \times \prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \ldots, x_n)} \quad (3)$$

For classification purposes, since the denominator is constant given the input, we use the proportional relationship:

$$P(y|x_1, x_2, \ldots, x_n) \propto P(y) \times \prod_{i=1}^{n} P(x_i|y) \quad (4)$$

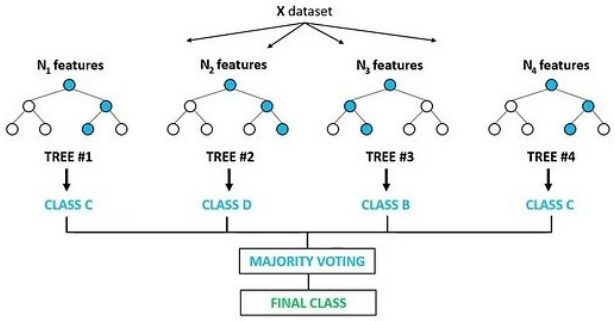
This formula calculates the posterior probability $P(y|x_1, x_2, \ldots, x_n)$ of a class y given independent features $x_1, x_2, \ldots, x_n$, where $P(y)$ is the prior probability of class y, and $P(x_i|y)$ is the likelihood of feature x_i given class y.

2.4. Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks by combining the predictions of individual trees through a majority voting mechanism. The class with the most votes becomes the final predicted class.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad (5)$$

Gini index is used for deciding how nodes on a decision tree branch. This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here, p_i represents the relative frequency of the class you are observing in the dataset and c represents the number of classes.



Random Forest Classifier

3. Intended Experiments

3.1. Encoding Parameters

Among various types of encoding like nominal encoding, ordinal encoding and one-hot encoding, we would use nominal encoding. One-hot encoding will create a lot of columns which will become difficult to handle and nominal encoding is preferred over ordinal encoding in this context since it will be difficult to assign ranks to such wide variety of dataset.

3.2. Why use these models?

- **Logistic Regression** : In the context of loan default prediction, Logistic Regression models the probability that a loan will default based on various attributes such as borrower's credit score, income level, loan amount, and profession. This method is useful due to its simplicity, efficiency, and the interpretability of its coefficients.
- **Naive Bayes'** : Naive Bayes provides a probabilistic framework for classification. It not only predicts the class label but also provides probability estimates for each class, which can be valuable in risk assessment scenarios like loan default prediction. It simplifies the modeling process and can capture certain patterns even if the features are not entirely independent. Naive Bayes can also handle a large number of features efficiently.
- **Random Forest** : Random Forest provides high accuracy in classification tasks. By combining the predictions of multiple trees, it reduces the risk of overfitting and increases the model's generalization performance. Loan default prediction problems often involve nonlinear relationships between features and the target variable. It is capable of capturing such nonlinearities, making it more flexible than simpler models. The combination of multiple trees in a Random Forest helps to balance bias and variance, resulting in a more stable and accurate model.

References

1. Kaggle dataset and paper:
   - Data set and paper: Loan Prediction Based on Customer Behavior.
2. References for models:
   - Logistic regression: Logistic Regression on Spiceworks.
   - Naive Bayes: Naive Bayes Classifiers on GeeksforGeeks.
   - Random Forest: Random Forest Algorithm on Medium.
3. Additional reference:
   - Categorical Data Encoding Techniques: Medium Article.
4. Research paper:
   - Research paper reference: Loan Default Prediction Model.