PURBANCHAL UNIVERSITY

KANTIPUR CITY COLLEGE (KCC)

Putalisadak, Kathmandu, Nepal

Masters of Computer Application (M.C.A.)

Reg. No.

A Proposal draft on

"A Comparative Analysis of Classification Algorithms for Bank Loan Eligibility predictions system"

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF COMPUTER APPLICATION

Submitted by

Manoj Rai

2024

ABSTRACT

Due to substantial technological advancements, people's needs have expanded. Consequently, there has been an increase in the number of loan approval requests in the banking industry. Several criteria are considered while selecting a candidate for loan approval in order to ascertain the loan's status. Banks encounter major challenges in evaluating loan applications and mitigating the risks linked to prospective borrower defaults. Due to the need to thoroughly assess the eligibility of every borrower for a loan, banks consider this process as notably burdensome. To determine the most effective machine learning model for loan approval, various algorithms such as Logistic Regression, Naive Bayes and Support Vector Machines (SVM) will be compared. Each algorithm comes with its own set of strengths and weaknesses, making the choice dependent on the specific characteristics of the dataset and the goals of the lending institution.

Keywords: Loan, Logistic Regression, Explainable, ML, Naive Bayes, Prediction, Random Forest, SVM

Table of Contents

LIST OF FIGURES

Chapter 1: INTRODUCTION

1.1 Background

In an era marked by unprecedented technological progress, the banking industry is undergoing a transformative shift in response to the growing and diverse needs of its clientele. As individuals seek financial support for a myriad of purposes, from homeownership to education and entrepreneurship, the volume of loan approval requests has surged, necessitating a more sophisticated approach to credit risk assessment. Recognizing the limitations of traditional manual underwriting methods, financial institutions are increasingly turning to machine learning (ML) algorithms to enhance the efficiency and accuracy of their loan approval processes.

This research seeks to delve into the intricate landscape of ML applications within the banking sector, with a particular focus on evaluating and comparing three prominent algorithms – logistic regression, decision trees, and support vector machines (SVM) – in the context of loan approval. As the banking industry grapples with the complexities of assessing a myriad of factors influencing creditworthiness, from financial histories to debt-to-income ratios, the need for a robust and streamlined approach to credit risk management has never been more pressing.

The primary objective of this research is to identify the most effective ML algorithm for loan approval by analyzing and comparing the performance of Logistic regression, Naïve Bayes, and SVM. Each algorithm brings a unique set of strengths to the table, ranging from interpretability to the ability to handle complex, nonlinear relationships within large datasets. Through an in-depth exploration of these algorithms and their application in the banking domain, this research aims to provide valuable insights into optimizing loan approval processes for financial institutions.

By leveraging historical data or simulated scenarios, the study will assess the performance metrics of each algorithm, including accuracy, precision, recall, and F1 score. The comparative analysis will not only shed light on the strengths and weaknesses of each model but will also assist banks in making informed decisions regarding the adoption of ML algorithms based on the specific characteristics of their datasets and institutional goals.

1.2 Statement of Problem

The primary problem at hand is the need for financial institutions to identify the most suitable machine learning algorithm for loan approval. Logistic regression, Naïve Bayes, and Support Vector Machines (SVM) are prominent contenders, each with its own strengths and weaknesses. The challenge lies in determining which algorithm offers the optimal balance of accuracy, interpretability, and adaptability to the dynamic nature of financial markets.

Moreover, as the banking industry grapples with an increasing volume of diverse loan applicants, the current processes are perceived as notably burdensome. The time-consuming nature of manual evaluations and the potential for inaccuracies underscore the urgency for a more streamlined and automated approach to credit risk assessment. Consequently, there is a critical need for research that systematically compares the performance of different machine learning algorithms in the context of loan approval, aiming to provide insights into optimizing these processes and addressing the challenges faced by financial institutions

1.3 Research Objectives

- To compare different Machine learning algorithms for predicting the eligibility of loan.
- To build a robust model for predicting eligibility of loan according to past data

Chapter 2: LITERATURE REVIEW

This section presents a literature survey. Relevant literature from multiple sources is referred for analysis of loan prediction system:

Prabaljeet Singh Saini et al. proposed a comparative analysis of various machine learning algorithms for predicting loan approval. The explored algorithms include Random Forest Classifier, K-Nearest Neighbors Classifier, Support Vector Classifier, and Logistic Regression. The findings show that the Random Forest Classifier had the highest accuracy of 98.04%, followed by K-Nearest Neighbors Classifier (78.49%), Logistic Regression (79.60%), and Support Vector Classifier (68.71%). This study provides insights into the effectiveness of different machine learning algorithms for loan approval prediction, and can be useful for financial institutions in improving their decision-making process.[1]

Ugochukwu. E. Orji et al. presented six machine learning algorithms (Random Forest, Gradient Boost, Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression) for predicting loan eligibility. The authors compare the performance of these algorithms using metrics such as accuracy, precision, recall, and F1 score. The results show that the Random Forest algorithm outperforms the other algorithms in terms of accuracy and F1 score.[2]

Viswanatha V et al. used four different algorithms namely Random Forest, Naive Bayes, Decision Tree, and KNN to predict the accuracy of loan approval status for an applied person. By using these, the authors obtained better accuracy of 83.73% with the Naïve Bayes algorithm as the best one.[3]

Ritika Purswani et al provided a review of various machine learning models used for loan approval prediction. The authors discuss the advantages and limitations of different models and provide insights into the future research directions in this field.[4]

Chapter 3: RESEARCH METHODOLOGY

3.1 Methods of Data Collection

For this research work, datasets are collected from Kaggle, Loan approval analysis | Kaggle The size of the dataset is 4268 samples, which have nine fields, where 11 fields are for input charactertics and one field for an output field. are representing the input fields, while the output field pertains to the presence of heart attack (class), which is divided into two categories (negative and positive); negative refers to the rejection of a loan, while positive refers to the approval of a loan.

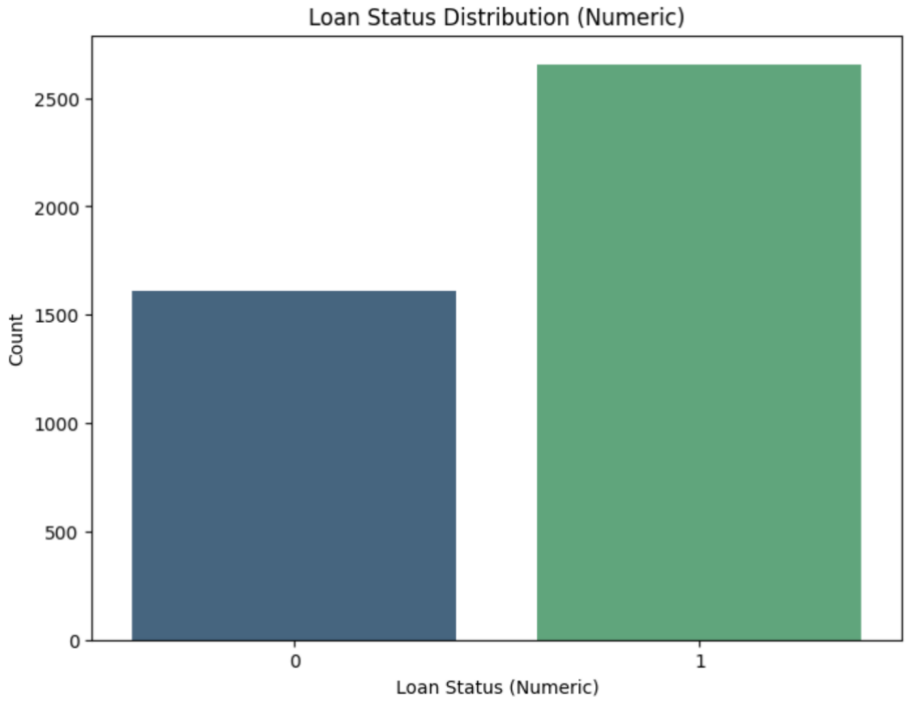

Figure 3-1: Dataset Visualization

3.2 Data Preprocessing

To prepare the data for modeling, the first step is to dropped the loan status column from the categorical data of each field and assign the binary labels for each values. The Correlation will be calculated to determine the similarities and dissimilarities between the categorical data. For the filling of missing values, the median will be calculated and assigned to those fields.

Once confirmed that there are no missing values, the entire dataset is split into two sets: training (4:5) and testing (1:5)

3.3 Model Development

For the development of model, various ML algorithm will be taken into consideration such as Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression.

3.3.1 Support Vector Machine

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression Support Vector Machine (SVM), in other words, is a classification and regression prediction tool that automatically detects over-fitting to the data while maximizing predictive accuracy using machine learning theory. Systems can be defined as Support Vector machines. This uses the linear function's hypotheses space in the high-dimensional feature space and was trained using an optimization theory-based learning algorithm that uses a learning bias obtained from theory of statistical learning.

The basic idea of SVM is to construct an optimal hyper plane, which can be used for classification, The optimal hyper plane is a hyper plane selected from the set of hyper planes for classifying patterns that maximizes the margin of the hyper planes.

- **Linear SVM**

  When the data is perfectly linearly separable only then we can use Linear SVM. Perfectly linearly separable means that the data points can be classified into 2 classes by using a single straight line(if two dimensional).

- **Non-Linear SVM**

  When the data is not linearly separable then we can use Non-Linear SVM, which means when the data points cannot be separated into 2 classes by using a straight line (if 2D) then we use some advanced techniques like kernel tricks to classify them. In most real-world applications we do not find linearly separable datapoints hence we use kernel trick to solve them.

The main two terms that are used in SVM are: -

Support Vectors: These are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.

Margin: it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin. There are two types of margins hard margin and soft margin.

The equation for hyperplane can be given as:

$$aX + bY = C \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

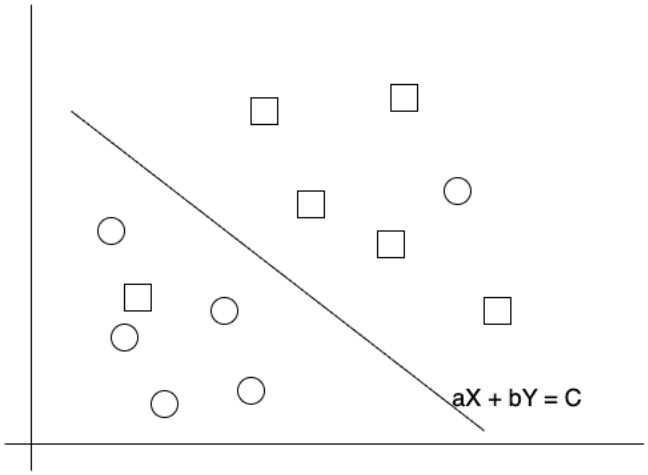

Figure 3-2: SVM Hyperplane

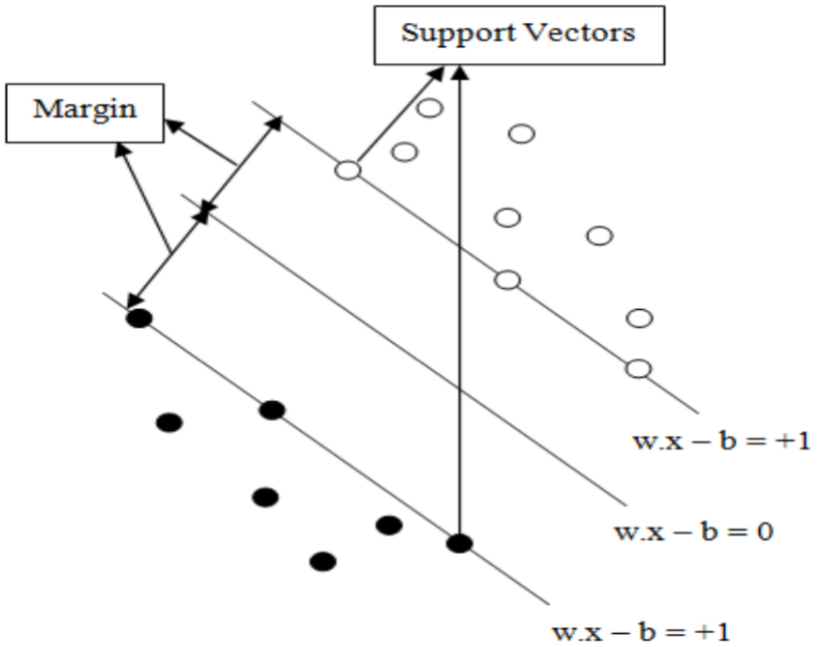

Figure 3-3: SVM Model

The figure 3.4 shows that the model consists of three different lines. These three lines construct the hyper plane that separates the given patterns and the pattern that lies on the edges of the hyper plane is called support vectors. The perpendicular distance between the line of margin and the edge of hyper plane is known as margin. The objective of support vectors is to optimize the margin so that it can classify the given problem.

3.3.2 Naïve Bayes

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem, named after the 18th-century mathematician and philosopher Thomas Bayes. It is particularly popular for classification tasks, such as spam detection or sentiment analysis, including its application in credit risk assessment in the banking industry.

3.3.2.1 Bayes' Theorem

At the core of Naive Bayes is Bayes' theorem, which mathematically describes the probability of an event based on prior knowledge of conditions that might be related to that event. The formula for Bayes' theorem is as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Where,

P(A|B) is the probability of event A given that event B has occurred.

P(B|A) is the probability of event B given that event A has occurred.

P(A) and P(B) are the probabilities of events A and B independently.

3.3.2.2 Naïve Assumption

The "naive" aspect of Naive Bayes comes from the assumption that features used to describe data are conditionally independent given the class label. In other words, the presence or absence of a particular feature does not influence the presence or absence of any other feature.

Naive Bayes is simple and computationally efficient. It requires less training data compared to other algorithms. It is robust to irrelevant features due to its conditional independence assumption.

3.3.3 Logistic Regression

Logistic Regression uses the sigmoid (logistic) function to model the probability of a binary outcome. The sigmoid function transforms any real-valued number into the range [0, 1]. The model is trained using a set of labeled data, where the features are used to predict the probability of belonging to the positive class. The parameters (coefficients) are adjusted during training to minimize the difference between predicted probabilities and actual class labels. Logistic Regression creates a decision boundary that separates the data into two classes. If the predicted probability is above a certain threshold (usually 0.5), the instance is classified as the positive class; otherwise, it is classified as the negative class.

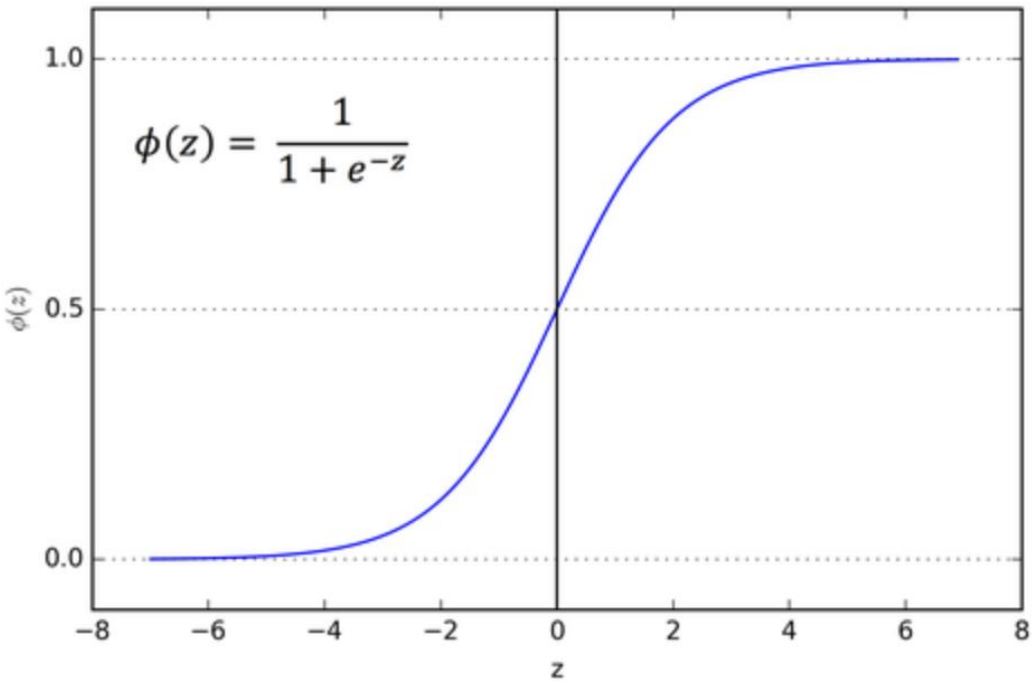

Figure 3-4: Sigmoid Function for Logistic Regression

Logistic Regression provides interpretable coefficients, allowing for a clear understanding of the impact of each feature on the predicted outcome. It performs well when the relationship between features and the log-odds of the outcome is approximately linear. Logistic Regression is less prone to overfitting, especially when the number of features is relatively small.

Chapter 4: References

[1] P. S. Saini, A. Bhatnagar, and L. Rani, "Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, May 2023, pp. 1821–1826. doi: 10.1109/ICACITE57410.2023.10182799.

[2] "Machine Learning Models for Predicting Bank Loan Eligibility | IEEE Conference Publication | IEEE Xplore." Accessed: Jan. 20, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9803172

[3] V. V, R. A. C, V. K N, and A. G, "Prediction of Loan Approval in Banks Using Machine Learning Approach." Rochester, NY, Aug. 04, 2023. Accessed: Jan. 20, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4532468

[4] R. Purswani, S. Verma, and Y. Jaiswal, "Loan Approval Prediction using Machine Learning: A Review," vol. 08, no. 06, 2021.