Comparative Analysis of Machine Learning Models for Loan Default Prediction

By: T.Nethmi Sarandi Sumathipala

Introduction

In today's banking and financial services industry, one of the most significant challenges faced by institutions is the problem of credit risk. Credit risk arises when borrowers fail to meet their repayment obligations, leading to potential financial losses for the lender. With the rapid growth of consumer lending and the increasing volume of loan applications, banks are under pressure to design effective strategies for assessing the repayment capability of customers before extending credit. In this context, data driven risk analytics has become an essential tool for minimizing loan defaults and ensuring the stability of financial operations.

This project focuses on analyzing a loan dataset that contains demographic, financial, and behavioral attributes of customers, along with a target variable indicating repayment difficulties. The dataset is particularly valuable because it helps to identify which characteristics of clients are strong indicators of loan default. By exploring and modeling these patterns, we can provide meaningful insights into how banks can reduce financial risk and make better lending decisions.

The approach taken in this study involves not only exploratory data analysis (EDA) and handling of missing values but also the application of multiple machine learning models for predictive analysis. A total of seven models were trained and compared: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting (GB), and Artificial Neural Network (ANN). These models were evaluated using accuracy, classification metrics, and confusion matrices to assess their ability to correctly classify clients as reliable or at risk of default.

By comparing the performance of different algorithms, this study aims to identify the most effective model for predicting loan defaults in this dataset. Such predictive modeling not only supports risk minimization but also provides financial institutions with a systematic framework for customer evaluation, portfolio optimization, and long-term decision making.

Problem Statement

In the financial sector, lending institutions face significant risk from loan defaults, which can lead to substantial financial losses and affect overall portfolio stability. Traditional credit evaluation methods, such as rule based scoring systems or manual assessments, often fail to capture complex relationships between borrower characteristics and default risk. These approaches may also lack scalability, accuracy, and adaptability to changing financial behaviors.

With the increasing availability of structured financial data, machine learning (ML) techniques present a promising solution for improving the accuracy of default prediction. By leveraging algorithms capable of handling large datasets and identifying non-linear patterns, ML models can provide deeper insights into borrower risk profiles and enhance decision-making in credit management.

Therefore, this project focuses on developing and comparing multiple machine learning models including Logistic Regression, Decision Tree, Random Forest, KNN, SVM, Gradient Boosting, and Artificial Neural Network (ANN) to predict the likelihood of loan defaults. The goal is to determine the most effective model that balances accuracy, interpretability, and computational efficiency, ultimately providing a data driven approach to risk mitigation in financial services.

Machine Learning Models

In this study, seven machine learning algorithms were employed to build and evaluate predictive models for loan default detection. Each model has unique characteristics and strengths that make it suitable for risk analytics in financial services. A brief description of each model is provided below:

1. **Logistic Regression (LR):**

   A statistical model widely used for binary classification problems. Logistic Regression estimates the probability of default using a linear combination of input features and applies the logistic (sigmoid) function for prediction. It serves as a strong baseline model due to its simplicity and interpretability.

2. **Decision Tree (DT):**

   A tree based algorithm that splits data into branches based on feature conditions to predict outcomes. Decision Trees are easy to interpret and can capture non-linear relationships, but they may be prone to overfitting if not properly tuned.

3. **Random Forest (RF):**

   An ensemble method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Random Forest is robust, handles large datasets well, and provides feature importance, making it valuable for identifying key risk drivers.

4. **K-Nearest Neighbors (KNN):**

   A distance based algorithm that classifies new instances based on the majority class of their nearest neighbors. While simple and effective in certain cases, KNN can be computationally expensive for large datasets.

5. **Support Vector Machine (SVM):**

   A powerful classifier that finds the optimal hyperplane separating different classes in the feature space. With kernel functions, SVM can handle both linear and non-linear relationships, making it effective in complex risk prediction scenarios.

6. **Artificial Neural Network (ANN):**

   A deep learning model inspired by the human brain, consisting of interconnected layers of neurons. ANNs are capable of capturing complex non-linear patterns and interactions among features, making them suitable for advanced predictive tasks.

7. **Gradient Boosting (GB):**

   An ensemble technique that builds models sequentially, where each new model corrects the errors of the previous one. Gradient Boosting is highly effective in improving predictive accuracy and has become one of the most popular algorithms for classification tasks in finance.

Research methodology

This section describes the dataset used in the study and the end-to-end preprocessing, modelling and evaluation procedures applied to predict loan default.

A. Dataset

The dataset used in this study was obtained from Kaggle and is designed to provide insights into credit risk and loan default prediction. It contains 122 features in the application data, including demographic, socio-economic, and financial variables such as total income, loan application amount, and credit amount. These attributes capture important aspects of a client's financial profile and repayment capacity, making the dataset highly suitable for exploring the patterns and factors that contribute to loan default.

B. Data Preprocessing

Before training machine learning models, the dataset underwent several preprocessing steps to ensure data quality and reliability. First, missing values were removed to maintain consistency across the dataset, and irrelevant identifiers such as SK_ID_CURR were dropped as they do not contribute to prediction. Next, the features were separated into independent variables **(**X**)** and the target variable (**y**), where the target represents whether a client defaulted on a loan. To address the issue of outliers in numerical columns, the Interquartile Range (IQR) clipping method was applied, which adjusts extreme values within acceptable statistical boundaries without discarding data. For categorical variables, One-Hot Encoding was performed using a column transformer to convert them into numerical representations suitable for machine learning algorithms. Since the dataset was highly imbalanced with fewer defaulters compared to non-defaulters, the ADASYN (Adaptive Synthetic Sampling) technique was used to

generate synthetic samples of the minority class, thereby balancing the target distribution. Finally, the data was split into training (80%) and testing (20%) sets using stratified sampling to preserve the class ratio, ensuring fair evaluation of the models.

C. Feature Selection

Feature selection was carried out to identify the most influential variables that contribute to predicting loan defaults while reducing model complexity and preventing overfitting. A hybrid approach combining both filter and wrapper methods was applied. Initially, statistical analysis using SelectKBest with ANOVA F-test was employed to shortlist the most relevant features based on their relationship with the target variable. This was followed by a Random Forest based feature importance ranking using SelectFromModel, which further refined the set by retaining only the top predictors. This two-stage process ensured that only the strongest and most informative features were selected, allowing the models to focus on the key drivers of loan default. The final selected features are presented in the following table.

D. Model Training

After feature selection, the dataset was used to train the selected machine learning models. All seven models were trained on the processed training dataset using the 15 selected features, with hyperparameters tuned to achieve optimal performance. A stratified train-test split was applied to maintain class balance, and models were evaluated on both training and testing sets. For consistency, each model was assessed using multiple performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix, to capture both overall predictive

power and class specific performance. This ensured that the evaluation considered not only the ability to classify correctly but also the balance between false positives and false negatives.

- Accuracy (ACC), which expresses the number of correctly classified instances from all instances, indicating the overall performance of the model.
- Precision (PRE), which reflects the proportion of true positive predictions that were truly positive, indicating the model's ability to minimize false positives.
- Recall (REC) also known as sensitivity which expresses the proportion of positive instances detectable by the model of all actual positive instances, indicating the model's ability to identify positive instance.
- F1-Score (F1) which is defined as the harmonic mean of precision and recall; this score is especially useful when working with imbalanced datasets and indicates how well the model performs based on both precision and recall together.

Model Performance Summary

Table 1: Training and testing accuracies

Model	Training Accuracy	Testing Accuracy
Logistic Regression	0.697	0.694
Decision Tree	0.933	0.902
Random Forest	0.999	0.967
Gradient Boosting	0.968	0.967
Support Vector Machine	0.956	0.943
K-Nearest Neighbors	0.938	0.864

Artificial Neural Network	0.953	0.928

The training and testing accuracies of all models were compared to assess their performance and generalization ability. As shown in the table1, Gradient Boosting achieved the most balanced performance with both training and testing accuracies of 0.968 and 0.967, respectively, indicating excellent generalization with minimal overfitting. Random Forest also exhibited high testing accuracy (0.967), though its perfect training score (0.999) suggests a mild degree of overfitting.

The Support Vector Machine (SVM) and Artificial Neural Network (ANN) models demonstrated strong and consistent performance, with testing accuracies of 0.943 and 0.928, respectively. These models effectively captured nonlinear patterns within the data, showing reliable predictive power without significant overfitting.

The Decision Tree achieved a good level of accuracy (training: 0.933, testing: 0.902), but its performance gap indicates some overfitting to the training data. Similarly, K-Nearest Neighbors (KNN) showed the largest difference between training (0.938) and testing (0.864) accuracy, suggesting it was highly sensitive to variations in the data and less stable on unseen samples. Logistic Regression, while showing consistent training and testing accuracy (around 0.69), performed the weakest overall, likely due to its inability to model complex nonlinear relationships among the features.

Overall, Gradient Boosting emerged as the most effective model, achieving high accuracy and stable performance across both training and testing datasets, making it the best-suited algorithm for predicting credit card default risk in this study.

Table 2: Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.694	0.70	0.69	0.69
Decision Tree	0.902	0.91	0.90	0.90
Random Forest	0.967	0.97	0.96	0.96
Gradient Boosting	0.967	0.97	0.97	0.97
Support Vector Machine	0.943	0.95	0.94	0.94
K-Nearest Neighbors	0.864	0.87	0.86	0.86
Artificial Neural Network	0.928	0.93	0.93	0.93

The performance metrics presented in Table 2 provide a more detailed understanding of how effectively each model classified credit card defaulters and non-defaulters. Among all models, Gradient Boosting and Random Forest achieved the highest overall precision, recall, and F1-scores (around 0.97), confirming their superior balance between correctly identifying both classes and minimizing false predictions. The Support Vector Machine (SVM) also delivered strong and stable results, maintaining precision and recall values above 0.94, indicating reliable classification capability across different data distributions.

The Artificial Neural Network (ANN) achieved comparable results, with precision, recall, and F1-scores of 0.93, demonstrating its effectiveness in capturing complex data relationships. Meanwhile, the Decision Tree achieved reasonably good precision and recall (both 0.90), but slightly lower than the ensemble models, suggesting a minor trade-off between model complexity and generalization.

On the other hand, K-Nearest Neighbors (KNN) exhibited moderate performance (F1-score 0.86), while Logistic Regression produced the lowest precision and recall (around 0.69), indicating limited capability in distinguishing defaulting customers from non-defaulters. Overall, Gradient Boosting stood out as the most balanced and reliable model across all evaluation metrics, confirming its robustness for credit card default prediction.

Conclusion

This study demonstrated the effectiveness of various machine learning algorithms in predicting loan default risk using a comprehensive financial dataset obtained from Kaggle. By applying systematic preprocessing, feature selection, and balanced data sampling techniques, the models were trained on reliable and representative data, ensuring robust performance evaluation.

Among the seven models tested Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, and Artificial Neural Network Gradient Boosting emerged as the most effective and reliable algorithm. It achieved the highest accuracy (0.967) along with balanced precision, recall, and F1-score values, indicating strong predictive capability and minimal overfitting. Random Forest also delivered competitive performance, while SVM and ANN achieved stable and high accuracies, confirming their suitability for complex nonlinear classification tasks.

Traditional models such as Logistic Regression and K-Nearest Neighbors performed relatively weaker, primarily due to their limited capacity to capture complex relationships and sensitivity to data variations. These results highlight the importance of ensemble and advanced models for accurate risk assessment in financial analytics.