**Application of Decision Tree for Loan Default Prediction**

PROJECT REPORT

**By**

**AVIJIT BISWAS (MSDSM/01/007)**

**M.Sc. Data Science & Management, 2023-25**

**Under the Supervision of**

**Dr. MAHIMA GUPTA**

**Operations Management and Quantitative Methods**

**Indian Institute of Management Amritsar**

****

**Operations Management and Quantitative Methods**

**Indian Institute of Management Amritsar**

**Autumn 2024-25**

**ABSTRACT**

This report addresses the prediction of loan defaulters using a decision tree machine learning model applied to a dataset encompassing loan application and payment information. The dataset includes various loan attributes and borrower details, essential for assessing default risk. Decision trees are employed to classify loans based on their likelihood to default, leveraging these features to provide insights into risk factors. By analysing historical loan data, the models aim to enhance underwriting processes and improve the accuracy of default predictions. This study contributes to refining financial risk management strategies, ultimately supporting more effective decision-making in lending operations.

**Keywords** Machine Learning · Decision Tree · Loan Credit · Default Prediction · Risk Prediction

**CONTENTS**

1. Introduction
2. Literature Review
3. Method
4. Model

4.1 Decision Tree Classifier

1. Results
2. Conclusion
3. References
4. Appendices

**Chapter 1 – Introduction**

In the dynamic landscape of modern banking, the process of assessing creditworthiness is crucial for determining the feasibility of extending loans to customers. Traditionally, this assessment has relied heavily on manual evaluation methods, which are not only time-consuming but also prone to human error and subjectivity. This inefficiency underscores the need for more sophisticated and automated approaches to enhance the accuracy and efficiency of credit risk management.

The objective of this report is to present a comprehensive analysis and application of machine learning techniques, specifically focusing on the utilization of a decision tree model, to predict loan defaulters for our banking client. By leveraging a dataset enriched with diverse customer attributes and detailed loan application information, we seek to automate and optimize the process of identifying individuals at risk of defaulting on their loans. This approach not only streamlines decision-making but also enhances the overall reliability of credit assessments.

The rationale behind adopting machine learning, particularly decision trees, lies in their ability to effectively handle complex datasets and uncover intricate patterns that may not be readily apparent through traditional methods. Decision trees are intuitive, interpretable models that segment data based on hierarchical decision rules, thereby providing transparent insights into the factors influencing loan default probabilities. By harnessing these capabilities, we aim to empower our banking client with actionable insights that can facilitate proactive risk mitigation strategies.

Through an in-depth exploration of historical loan data, encompassing a diverse range of customer demographics, financial metrics, and loan characteristics, this report aims to showcase the transformative potential of machine learning in refining financial risk management practices. By automating the identification of credit loan defaulters, banks can not only optimize operational efficiency but also improve decision-making accuracy, thereby fostering a more robust and sustainable lending environment.

Furthermore, this study underscores the broader implications of leveraging advanced analytics in banking operations. By embracing machine learning models such as decision trees, banks can adapt to evolving market dynamics, mitigate potential risks more effectively, and ultimately enhance their competitive advantage in the financial services sector. This report thus serves as a foundational resource for understanding the strategic role of machine learning in modernizing credit risk assessment processes, paving the way for future innovations in financial services.

In summary, this report aims to provide a comprehensive overview of how machine learning, specifically decision trees, can revolutionize the identification and management of credit loan defaulters for our banking client. By bridging the gap between data-driven insights and actionable decision-making, we envision a future where predictive analytics empower financial institutions to make informed, timely, and strategic lending decisions that drive sustainable growth and customer satisfaction.

**Chapter 2 – Literature Review**

The rapid advancements in the banking sector and the growing tendency among individuals to apply for loans have resulted in a heightened need for effective risk assessment methodologies. As the incidence of loan defaults escalates, banks face significant challenges in accurately evaluating loan applications and managing associated risks. This literature review explores existing research focused on improving loan approval processes through machine learning techniques, emphasizing the comparative performance of different algorithms and methodologies.

A critical issue in the banking sector is the increasing rate of loan defaults, which complicates the assessment of loan requests and risk management. Various studies have sought to address two fundamental questions: (i) How risky is the borrower? and (ii) Given the borrower's risk, should the bank lend to them? Accurate prediction of loan defaults is essential for minimizing financial losses and maintaining the stability of financial institutions.

Machine learning models have gained prominence in recent years for their ability to enhance predictive accuracy in various applications, including loan default prediction and credit assessment. Research has focused on developing models that can assess borrower risk based on specific attributes, thereby streamlining the loan approval process for banking authorities.

A notable study proposes the use of two machine learning models—Random Forest and Decision Trees—to predict loan approval outcomes. By applying these models to the same dataset, the study provides a comparative analysis of their performance. Random Forest, an ensemble learning method, constructs multiple decision trees and merges their results to improve predictive accuracy. This method is known for its robustness and ability to handle large datasets with high dimensionality. On the other hand, Decision Trees are a non-parametric supervised learning method that uses a tree-like model of decisions. While simpler and more interpretable, decision trees can be prone to overfitting and may not perform as well as ensemble methods in complex scenarios.

With the advent of big data, financial institutions have access to massive datasets from various sources, enabling more comprehensive credit assessments. However, this also introduces challenges related to data processing and noise elimination. Another study addresses the complexity of credit assessment in a big data environment by proposing a Decision Tree Credit Assessment Approach (DTCAA). This method leverages the interpretability and rule-based structure of decision trees while incorporating data consolidation techniques to manage noisy datasets. DTCAA has been validated using a large dataset from a major car collateral loan company in Taiwan, demonstrating its efficiency and effectiveness in various scenarios. The approach combines the advantages of decision trees with enhanced data processing capabilities, making it a competitive tool for credit assessment in big data contexts.

The study comparing Random Forest and Decision Tree models reveals that the Random Forest algorithm significantly outperforms the Decision Tree algorithm in predicting loan approval, demonstrating higher accuracy in identifying suitable candidates for loans. This finding underscores the potential of ensemble learning techniques to enhance risk assessment processes in the banking sector. Similarly, the DTCAA study highlights the practical applicability of decision trees in big data environments, showing that decision tree models can achieve competitive performance while maintaining interpretability and ease of use. The combination of decision trees with data consolidation methods enhances their ability to handle complex and noisy datasets, further supporting their use in credit assessment practices.

**Chapter 3 – Method**

The dataset consists of two primary files, ‘loan.csv’ and ‘payment.csv’, which contain comprehensive information about loan applications and their subsequent payment details.

The ‘loan.csv’ file encompasses detailed data about loan applications, containing 18 columns. Each row in this file represents an accepted loan application or a successfully funded loan. The key identifiers in this file include loanId, which is a unique loan identifier used for joining with the payment.csv file, and ‘anon\_ssn’, a hashed version of the client's social security number, useful for identifying repeat customers. The file details the loan's repayment frequency through the ‘payFrequency’ column, with possible values indicating biweekly, irregular, monthly, semi-monthly, and weekly payments.

Financial specifics are covered by the ‘apr’ column, indicating the annual percentage rate, and the ‘loanAmount’ column, which shows the principal amount of the loan. The ‘applicationDate’ and ‘originatedDate’ columns provide timestamps for when the loan application was submitted and when the loan was originated, respectively. The dataset also includes various status indicators such as originated, approved, ‘isFunded’, and ‘loanStatus’, which track the loan's progression from application to funding and beyond. Notably, the ‘loanStatus’ column has specific statuses to denote returned items, rejected applications, withdrawn applications, and voided loans.

Additional columns like ‘nPaidOff’ count the number of previous loans paid off by the client, and ‘originallyScheduledPaymentAmount’ indicates the scheduled repayment amount. The client's state of residence is captured in the state column, while ‘Lead\_type’ and ‘Lead\_cost’ provide information on the type and cost of the lead that generated the loan application. The ‘fpStatus’ column describes the result of the first loan payment, and ‘clarityFraudId’ is a unique underwriting ID used for further data joins.

The ‘payment.csv’ file contains nine columns detailing the payments made on loans. Each payment is identified by ‘loanId’, which can be joined with the ‘loan.csv’ file. The ‘isCollection’ column indicates whether the payment is part of a custom collection plan designed for customers struggling with the original repayment schedule. The ‘installmentIndex’ counts the nth payment for each loan, resetting for collection payment plans. Payment specifics are provided by ‘paymentDate’, which records the effective date of the payment, and ‘paymentAmount’, which includes both the principal (principal) and the fee or interest amount (fees).

Payment statuses are tracked in the ‘paymentStatus’ column, which details whether a payment was successful (Checked), unsuccessful (Rejected), cancelled (Cancelled), pending (Pending), skipped (Skipped), not attempted yet (None), or awaiting retry after rejection (Rejected awaiting retry). Additionally, the ‘paymentReturnCode’ column provides ACH error codes explaining the reasons for any payment failures.

Together, these files provide a rich dataset that captures the lifecycle of loan applications from submission through funding and repayment, offering valuable insights for analysis and decision-making.

**Chapter 4 – Model**

The dataset, which consists of consumer data, was divided into training and testing subsets to ensure an unbiased evaluation of the model's performance. This initial partitioning is crucial for developing a reliable model that can generalize well to new, unseen data.

For the classification task, a decision tree classifier was employed. The training subset was used to build the decision tree model. To fine-tune the model's parameters and mitigate overfitting, a further train-test split was conducted within the training data. This involved dividing the training data into training and validation sets. The model was trained on the training set, and its performance was evaluated on the validation set to optimize the parameters.

Once the model was optimized, it was tested on the held-out testing dataset to assess its generalization capabilities. To further evaluate the model's robustness and reliability, k-fold cross-validation was performed. This technique involves dividing the dataset into k equal-sized folds, training the model on k-1 folds, and evaluating it on the remaining fold. This process is repeated k times (k = 10), with each fold serving as the validation set once, ensuring a thorough evaluation of the model's performance.

Performance metrics, such as accuracy, precision, recall, F1-score, and confusion matrix, were used to measure the effectiveness of the model. These metrics provide a comprehensive understanding of the model's predictive capabilities and its ability to accurately classify consumer data.

**4.1. Decision Tree Classifier**

The decision tree classifier, used as an alternative to traditional classification methods, was informed by insights from the literature. Cross-validation revealed that the decision tree effectively handles both categorical and numerical data, making it a versatile choice for classification tasks. The main equation guiding the decision tree classifier involves selecting features that maximize information gain, typically calculated using the formula for information entropy in (1)

Information Gain = H(*S*) - ) (1)

where H(*S*) is the entropy of the entire set, and ) is the entropy. However, despite its strengths, the decision tree classifier showed limitations in performance compared to more complex algorithms, often resulting in a tendency to overfit the training data.

**Chapter 5 – Results**

This section presents an overview of the paper's core findings and includes tables and charts to facilitate a thorough understanding of the methodological procedure. It also offers a critical analysis of the results based on the evaluation metrics used to assess modeling performance.

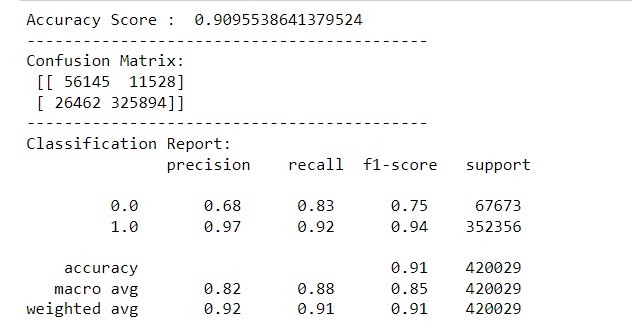


Fig. 1 Confusion Matrix when used with Gini Index

In this study, we evaluated several machine learning models to predict whether to allocate loans to a person or not. Using the data, we trained and tested decision tree algorithm, focusing on maximizing predictive accuracy and AUC-ROC performance. The model was tuned to find the optimal hyperparameters based on the AUC-ROC metric, enhancing predictive performance.

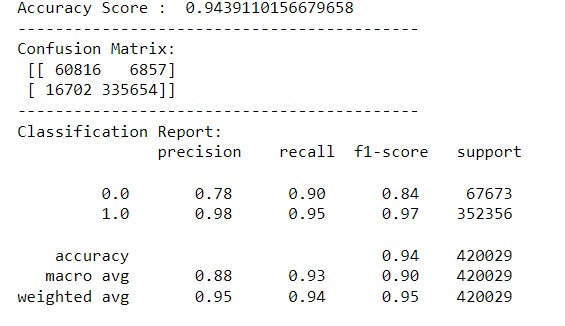


Fig. 2 Confusion Matrix when used with Entropy.

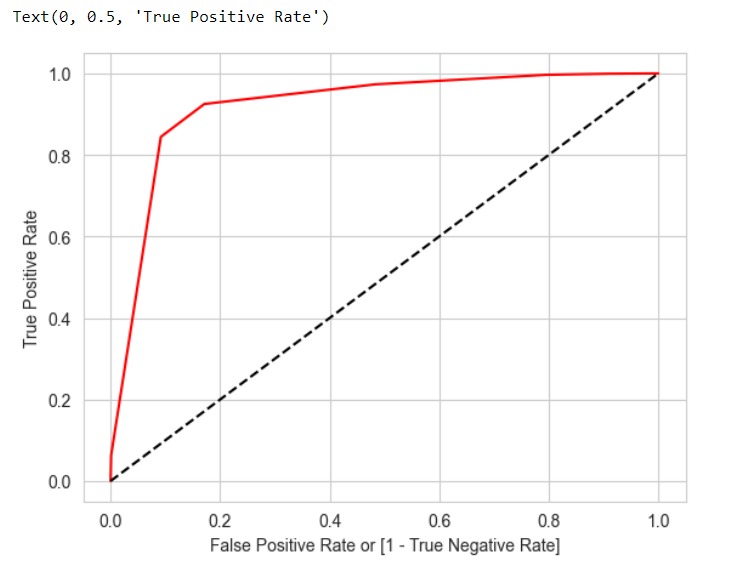


Fig. 3 AUC-ROC Curve when used with Gini Index.

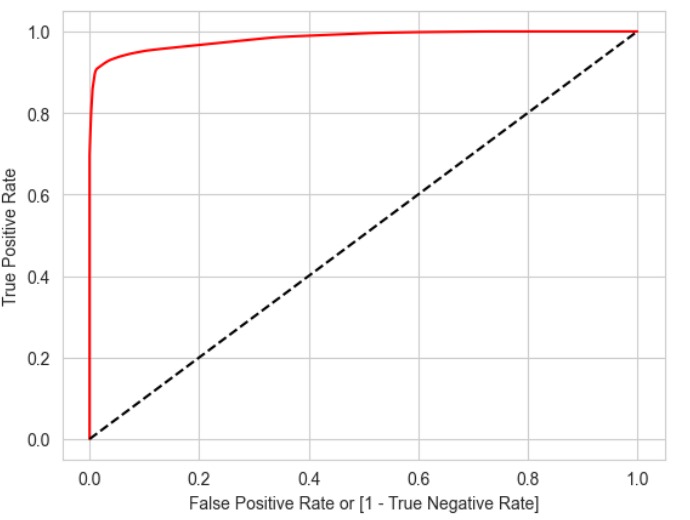


Fig. 4 AUC-ROC Curve when used with Entropy.

**Chapter 6 – Conclusion**

The decision tree classifier effectively predicted loan defaulters, demonstrating its capability in handling consumer loan data. By accurately classifying potential defaulters, the model can assist financial institutions in mitigating risks and making informed lending decisions. The use of cross-validation ensured the model's robustness and reliability, confirming its utility in real-world applications. Overall, the decision tree classifier proves to be a valuable tool in the proactive management of loan default risks.

**Chapter 7 – References**

1. Mehul Madaan *et al* 2021 “Loan default prediction using decision trees and random forest: A comparative study ” *IOP Conf. Ser.: Mater. Sci. Eng.* 1022 012042.
2. Chern, CC., Lei, WU., Huang, KL. *et al.* A decision tree classifier for credit assessment problems in big data environments. *Inf Syst E-Bus Manage* 19, 363–386 (2021). https://doi.org/10.1007/s10257-021-00511-w

**Chapter 8 – Appendices**

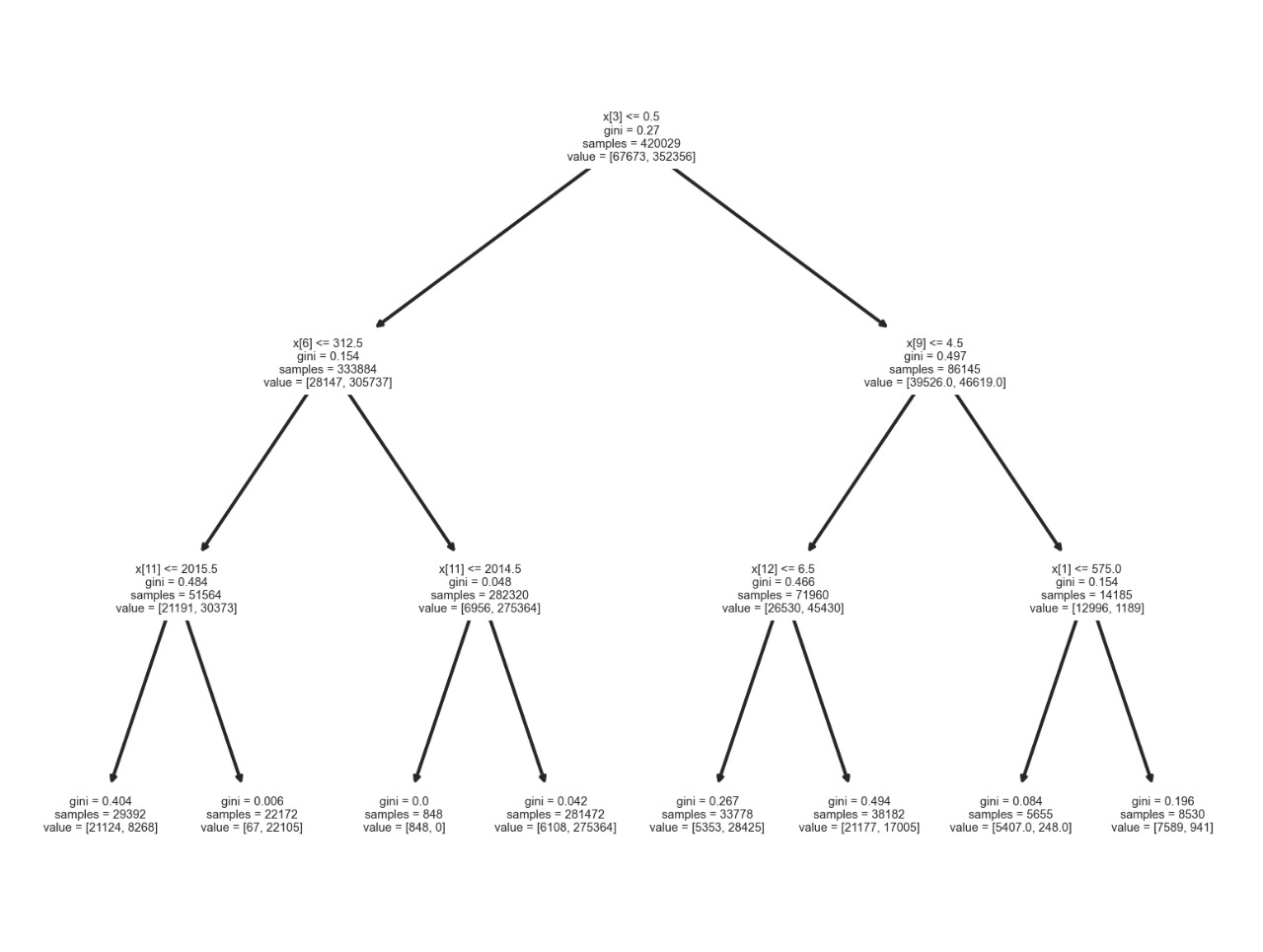


Fig. 5 Decision Tree Diagram for Gini Index.