International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

CUSTOMER CREDIT SCORING

[1] Siddamma C.M, [2] Nimmala Sainath, [3] K Poojitha, [4] Harsha Vardhan, [5] S Venkatesh
[1]AssistantProfessor, [2345]Students
Department of Computer Engineering(Internet Of Things)
Siddhartha Institute of Technology & Sciences, Narapally
siddmmacm@siddhartha.org.in , 23tq1a6908@siddhartha.co.in, 23tq1a6932@siddhartha.co.in, 23tq1a6949@siddhartha.co.in, 23tq1a6959@siddhartha.co.in

Abstract

Credit risk analysis is a critical component of the banking and financial services industry, as it helps minimize losses caused by borrower defaults and ensures financial stability. Traditional methods of credit assessment are often subjective, inconsistent, and time-consuming, which can lead to inaccurate lending decisions. This project focuses on developing a machine learning-based system to predict the likelihood of a customer defaulting on a loan.

The system utilizes a dataset containing demographic and financial information of loan applicants, including features such as age, education, employment years, income, and debt-related attributes. Data preprocessing techniques are applied to handle missing values and improve data quality. Additionally, new features such as total debt and debt ratio are engineered to enhance predictive performance. Exploratory Data Analysis (EDA) is performed to understand patterns and relationships within the data.

A Random Forest Classifier is used to build the predictive model, achieving high accuracy and effectively capturing complex relationships between variables. Model performance is evaluated using metrics such as accuracy, confusion matrix, and classification report. The results highlight the importance of factors like debt ratio, total debt, and income in determining credit risk. Overall, this project demonstrates how machine learning can provide a reliable, objective, and scalable solution for credit scoring, improving loan approval processes and reducing financial risk.

I. Introduction

In the modern financial ecosystem, credit risk assessment plays a vital role in ensuring the stability and profitability of financial institutions. Banks and lending organizations must evaluate whether a loan applicant is capable of repaying borrowed funds without defaulting. However, traditional credit scoring methods often rely on manual evaluation, expert judgment, or basic statistical models, which may lead to inconsistencies and inefficiencies. These methods are not always capable of capturing complex relationships between multiple financial and demographic factors.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

Financial institutions face significant challenges due to borrower defaults, which can result in substantial financial losses. Traditional risk assessment methods may either approve high-risk applicants or reject potential customers who are actually creditworthy. Therefore, there is a need for an automated and data-driven system that can accurately predict the likelihood of default based on applicant data.

This project aims to develop a machine learning-based credit scoring system using applicant information such as age, income, employment history, and debt ratios. By applying advanced algorithms like Random Forest, the system can learn patterns from historical data and make reliable predictions. This approach not only improves accuracy but also reduces manual effort and enhances decision-making in loan approval processes. Ultimately, the system helps financial institutions minimize risk, improve efficiency, and make fair and consistent lending decisions.

II. Literature Survey

A review of existing literature shows a significant evolution in credit risk assessment techniques, moving from traditional statistical approaches to advanced machine learning and explainable AI methods. Early research by Altman (1968) introduced the Z-score model based on discriminant analysis, which used financial ratios to predict corporate bankruptcy. This model laid the foundation for modern credit risk analysis by demonstrating how quantitative methods can be used for decision-making.

Later studies, such as Abdou and Pointon (2011), provided a comprehensive review of credit scoring techniques, highlighting the use of statistical models like logistic regression and discriminant analysis. Their work emphasized the importance of model evaluation criteria such as accuracy, sensitivity, and specificity in assessing performance. These traditional models were widely used due to their simplicity and interpretability but had limitations in capturing complex relationships in data.

With the advancement of technology, researchers began adopting machine learning approaches for credit risk prediction. Kou et al. (2019) demonstrated that machine learning models such as XGBoost outperform traditional models in terms of accuracy and efficiency. These models are capable of handling large datasets and identifying non-linear relationships between variables, making them more suitable for modern financial systems.

Furthermore, guidelines from the European Banking Authority (2021) emphasize the importance of robust methodologies for estimating probability of default (PD) and loss given default (LGD). These guidelines encourage the adoption of advanced analytical techniques while maintaining transparency and compliance with regulatory standards.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

III. System Analysis

System analysis for customer credit scoring focuses on understanding how financial and demographic data can be used to assess the creditworthiness of loan applicants. The system analyzes factors such as income, employment history, age, education, and debt-related attributes to predict the likelihood of default. It involves collecting applicant data from financial records and preprocessing it by handling missing values, normalization, and feature engineering. The system must identify key variables that significantly influence credit risk, such as debt ratio and total debt. Various machine learning algorithms are evaluated to determine the most suitable model for prediction. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess model effectiveness. The system should be scalable to handle large volumes of applicant data efficiently. It also needs to ensure data privacy, security, and compliance with financial regulations. Overall, system analysis helps design a reliable and efficient credit scoring model for better lending decisions.

Existing System

The existing system for credit scoring primarily relies on traditional statistical methods and manual evaluation processes. Financial institutions often use rule-based systems, basic scoring models, or expert judgment to assess the creditworthiness of applicants. These systems depend heavily on predefined criteria and may not consider complex relationships between multiple factors. Traditional methods such as credit bureau scores and simple logistic regression models are commonly used but have limitations in handling large and diverse datasets. The evaluation process can be time-consuming and prone to human bias. Additionally, these systems may not adapt well to changing financial behaviors or new data patterns. They often provide limited accuracy and lack the ability to capture non-linear relationships in the data. As a result, existing systems may lead to incorrect approvals or rejections, affecting both financial institutions and customers.

Disadvantages of Existing System

- Dependence on manual and rule-based evaluation
- Limited ability to handle complex and non-linear data
- Prone to human bias and inconsistency
- Lower prediction accuracy
- Time-consuming process
- Limited scalability for large datasets
- Inflexible to changing financial patterns

Proposed System

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

The proposed system utilizes machine learning techniques to improve the accuracy and efficiency of customer credit scoring. It collects and processes applicant data such as income, employment details, age, education, and debt-related features. Data preprocessing techniques are applied to clean and transform the data, including handling missing values and encoding categorical variables. Feature engineering is used to create meaningful attributes like total debt and debt ratio. Machine learning models such as Random Forest, Decision Tree, and Logistic Regression are trained on the processed dataset. The system evaluates model performance using metrics like accuracy, precision, recall, and confusion matrix to ensure reliability. The best-performing model is selected for predicting the probability of default. The system enables automated and real-time credit risk assessment. It is scalable and adaptable to new data patterns. Overall, the proposed system provides a more accurate, efficient, and data-driven solution for credit scoring.

Advantages of Proposed System

- Higher prediction accuracy using machine learning
- Automated and faster decision-making
- Reduces human bias and errors
- Handles complex and non-linear relationships
- Scalable for large datasets
- Adaptable to changing financial patterns
- Improves risk management and reduces losses

IV. Methodology

The methodology of the customer credit scoring system follows a structured approach to accurately predict the likelihood of loan default. Initially, data is collected from financial records, including attributes such as age, income, education, employment history, and debt-related features. The collected data is then preprocessed by handling missing values, removing inconsistencies, and encoding categorical variables into numerical form. Feature engineering techniques are applied to create meaningful attributes such as total debt and debt ratio, which improve model performance.

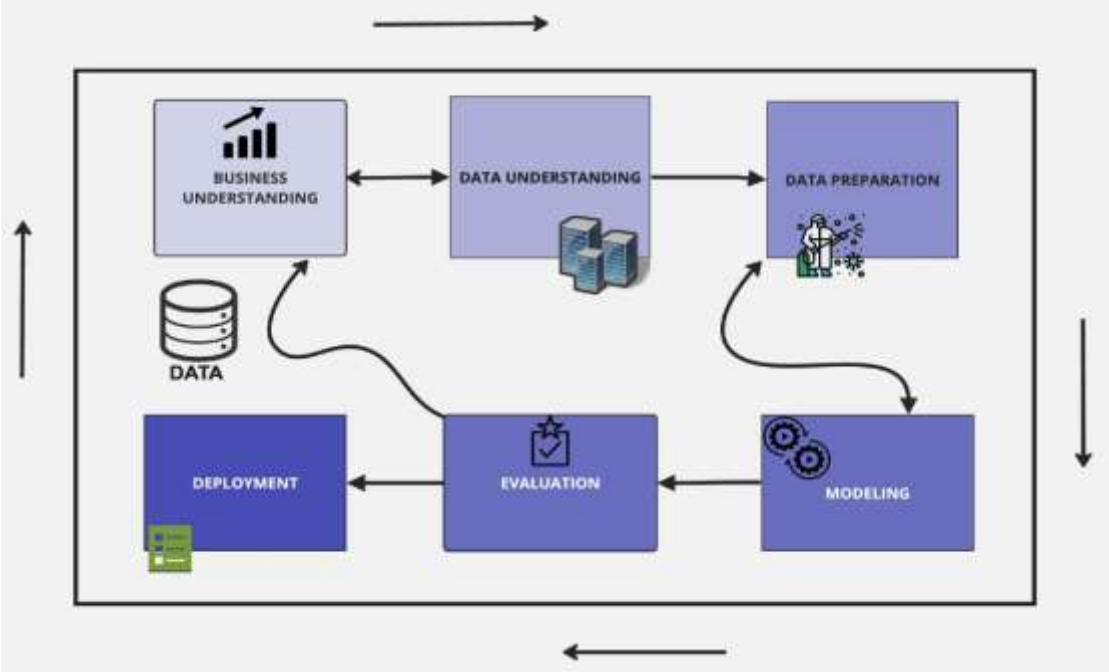
Next, exploratory data analysis (EDA) is performed to understand patterns, correlations, and distributions within the dataset. The dataset is then split into training and testing sets to evaluate model performance effectively. Machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest are applied to train predictive models. The performance of these models is evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrix. The best-performing model is selected and deployed for predicting whether a customer is

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X ww.ijdim.com Original Research Paper

likely to default. The system can be continuously improved by retraining the model with updated data to maintain accuracy and reliability.

System Architecture

The system architecture of the customer credit scoring system is designed as a multi-stage pipeline that processes applicant data from input to prediction. It begins with the data collection layer, where demographic and financial information such as income, employment history, age, and debt details are gathered. This data is then passed to the preprocessing layer, where it is cleaned, normalized, and transformed by handling missing values and encoding categorical variables. The feature engineering layer creates important attributes such as total debt and debt ratio to enhance predictive capability.



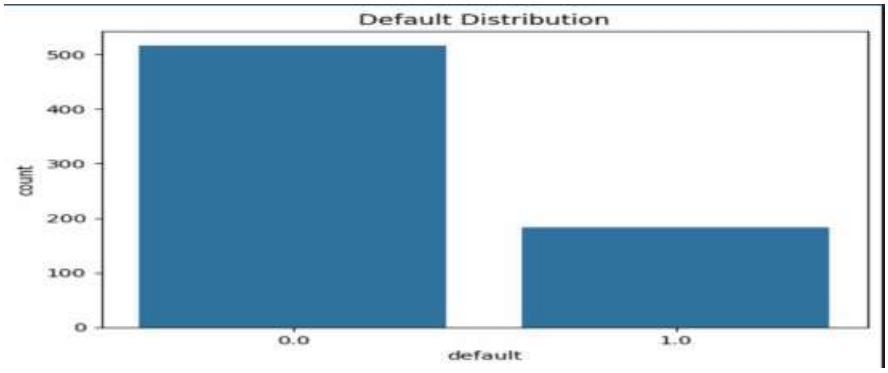
V. Result and Output

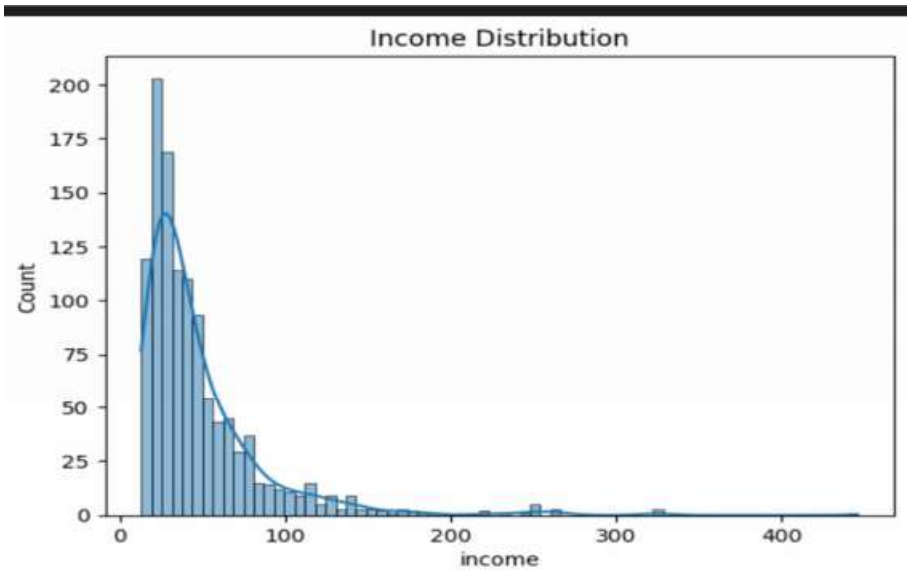
International Journal of

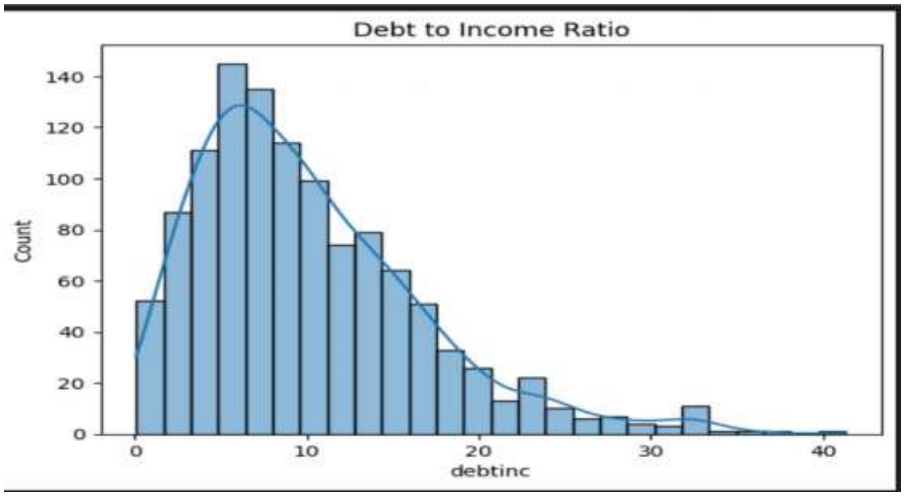
DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper





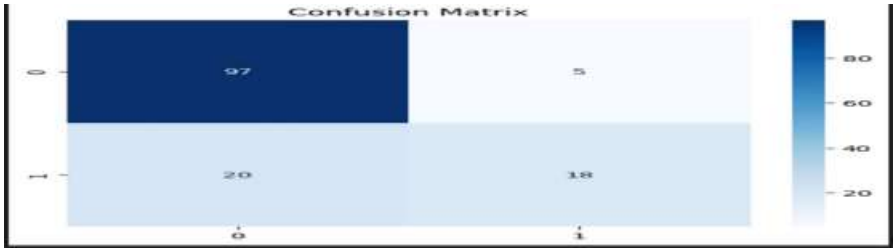


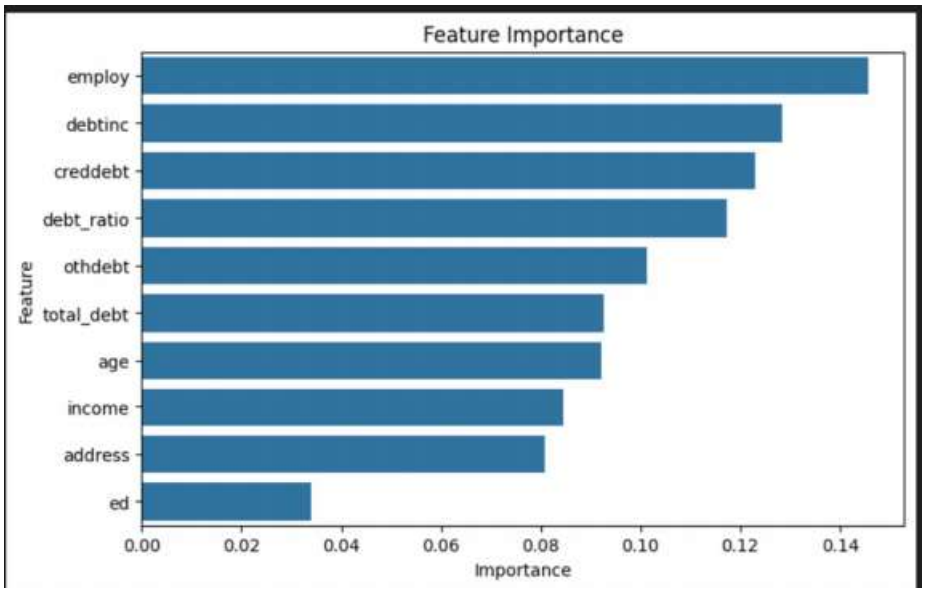
International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper





VI. Conclusion

In conclusion, the Customer Credit Scoring project demonstrates the effectiveness of machine learning techniques in predicting loan default risk using financial and demographic data. By analyzing key factors such as income, employment history, and debt ratios, the system is able to identify patterns that influence creditworthiness. The use of models like Random Forest, Decision Tree, and Logistic Regression improves prediction accuracy and reduces dependency on manual evaluation methods.

The project highlights the importance of data preprocessing, feature engineering, and model evaluation in building a reliable credit scoring system. The automated nature of the system ensures faster, unbiased, and consistent decision-making. Overall, this project proves that machine learning-based credit scoring systems can significantly enhance risk management, reduce financial losses, and improve the efficiency of loan approval processes in modern financial institutions.

International Journal of

DATA SCIENCE AND IOT MANAGEMENT SYSTEM

Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X ww.ijdim.com Original Research Paper

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of

Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.