Volume 10, Issue 4, April – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1924

Exploratory Data Analysis for Banking

Snehal Shingode[1*]; Saurabh Thaware[2]; Sakshi Katre[3]; Harshal Balpande[4]; Shravasti Gaikwad[5]; Adil Sheikh[6]; Harsh Dubey[7]

[1;2;3;4;5;6]Department of Robotics & Artificial Intelligent Priyadarshini College of Engineering, RTMNU University Nagpur, 440016, Maharashtra, India

Corresponding Author: Snehal Shingode[1*]

Contributing Authors: Harsh Dubey[7]; Sakshi Katre[3]

Publication Date: 2025/05/09

Abstract: In the banking and finance industry, exploratory data analysis, or EDA, is essential because it helps businesses extract insightful information from big, complicated datasets. EDA aids in the identification of patterns, the detection of anomalies, and the comprehension of underlying trends that influence decision-making in this field. EDA enables financial institutions to better understand market dynamics, risk factors, portfolio performance, and consumer behaviour by applying statistical and visualisation approaches. The uses of EDA in banking and finance are examined in this research, with a focus on how it might enhance investment strategies, fraud detection systems, and credit scoring models. It also emphasises how crucial feature selection, data preprocessing, and visualisation tools are to supporting efficient data-driven decision-making.

How to Cite: Snehal Shingode; Saurabh Thaware; Sakshi Katre; Harshal Balpande; Shravasti Gaikwad; Adil Sheikh; Harsh Dubey (2025). Exploratory Data Analysis for Banking. *International Journal of Innovative Science and Research Technology*, 10(4), 2919-2923. https://doi.org/10.38124/ijisrt/25apr1924

I. INTRODUCTION

Receiving and safeguarding funds deposited by individuals or organisations is referred to as banking. This also includes giving them loans that they will pay back within the allotted period. Since the banking industry plays a significant role in defining the nation's financial stability, it is regulated in the majority of nations. The Banking Regulation Act makes it possible for the general population to get loans. Loans are substantial sums of money that are taken out over time with the expectation that they will be repaid at a specific interest rate. Depending on the needs of the client, the loan may be used for any purpose. Open-ended and closed-ended loans are the two main categories of loans. Loans that the client has been approved for up to a certain amount are known as open-ended loans. Credit cards and home equity lines of credit (HELOCs) are two types of open-end loans. With every instalments, closed-ended debts get smaller. Stated differently, it is a legally binding agreement that the borrower cannot change. The most prevalent types of closed-ended loans are student loans, mortgages, auto loans, personal loans, and instalment loans. Loans that are backed by an asset are known as secured or collateral loans. The personal assets that are utilised to secure the loan include homes, cars, and savings accounts. Personal or signature loans are other names for unsecured loans. Based on the borrower's financial resources, the lender in this case thinks the borrower can repay the loan. The risk associated with an investment's inability to be bought or sold quickly enough to avoid or reduce a loss is known as liquidity risk.

Interest rate risk is the chance that loan interest rates will be too low to provide revenue for the bank. Ensuring that their wealth is in safer hands is the bank's main goal. Banks now approve loans after confirming and certifying the customer's submitted documentation. However, there is no assurance that the candidate is worthy. Customers are categorised in this document according to specific standards. Exploratory data analysis is used to classify the data. Exploratory Data Analysis (EDA) is a way of analysing datasets that uses visual techniques to highlight their key features. The goal of EDA is to use visualisation techniques to reveal the underlying structure of a comparatively bigger set of variables.

II. LITERATURE SURVEY

The researchers use data mining techniques to analyse the data set in [1]. Loan prediction systems benefit greatly from data mining procedures since they quickly identify borrowers who can repay the loan balance within a given time frame. The J48 algorithm, Bayes net, and Naive Bayes are among the algorithms that are employed. When these algorithms were applied to the datasets, it was discovered that the "J48 algorithm" had a high accuracy (correct percent) of 78.3784%, giving the banker the ability to determine whether or not the customer could receive the loan. The Tree model, Random Forest model, and SVM model were utilised in the study [2] titled "Loan prediction using Ensemble technique,"

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1924

and the three models were integrated to form an Ensemble model. In order for the banking industry to accept or reject the loan request from their clients, a prototype has been discussed in the paper [2]. Real coded genetic algorithms are the primary technique employed. Loan prediction is made easier with the help of the ensemble model's blended algorithms. The tree algorithm is determined to have accuracy81.25%. Since the probability of default (PD) is a crucial step for clients seeking a bank loan, an enhanced risk prediction clustering algorithm is employed in the research [3,4] utilising the R programming language to identify risky loan customers.

Thus, the data mining technique offers a framework for identifying PD in the data set. When there are missing values in the data set, the R-language's KNN (K-nearest neighbour) algorithm is utilised to execute multiple imputation calculations.

Using the decision tree induction technique, the paper [5] discovered that the algorithm determines the optimal method for assessing credit risk. Bankers use a strategy known as "credit score" to ward off credit risk. This technique helps lenders keep track of which applicants are likely to default or who can repay the amount owed. Civil score, WEKA software, and client data were the inputs used for credit evaluation. Understanding the problem and the data, filtering the data, modelling the system, and evaluating the system were the steps in the prediction system's technique. The bank's current dataset, which included 1140 records and 24 attributes, was used for this. . The technology has finally been tested and is able to assist bankers in making the right decision on the approval or rejection of loans.

In order to forecast bank loan approval, the study [6] employed both descriptive and predictive model techniques. Regression and classification were employed in the predictive model approach, while association and clustering were employed in the descriptive model technique. Classifiers also use a variety of methods, such as the R language's naive Bayes and KNN algorithms, whereas regressors use a variety of techniques, such as decision trees and neural networks. Out of all these algorithms, naive Bayes generates the most accurate classifier for this prediction analysis, while techniques such as decision trees, neural networks, and K-NN algorithms will be more accurate regressors. Predicting the loan categorisation based on the loan type, loan applicant, and assets (property) that the loan applicant possesses is the primary objective of the study.

It was found that the decision tree algorithm gave an improved accuracy of almost 85% on doing the analysis.

III. LOAN APPLICANT DATA ANALYSIS

The bank automatically exposes itself to a number of financial risks whenever it decides to lend money to any of its clients. The bank must be aware of the customers who are applying for loans. This issue encourages an EDA on the provided dataset, which analyses the customer's characteristics. Normalisation, missing value treatment, filtering to choose key columns, creating new columns, determining the target variables, and graphical data visualisation are all processes that are applied to the dataset that uses EDA. Python is used to process data quickly and easily. This study processed and extracted data from the provided dataset using the Python pandas module. To improve comprehension and visualisation of the results, the processed data is transformed into the proper graphs. To acquire the graph, the Matplot package is utilised.

➢ *Annual Income Compared to Loan Purpose*

The X axis in Figure 1 denotes the loan's purpose, or the reason it is being applied for. Among the goals are house repair and debt consolidation. The annual income of those who fall into the following range is represented by the terms high, moderate, and poor. People with an annual income between 10 lakhs and 25 lakhs are considered to have a low income, those with an annual income between 10 lakhs and 25 lakhs are considered to have a moderate income, and those over 25 have a high income. lakhs. By these criteria, a new column called Category is derived.

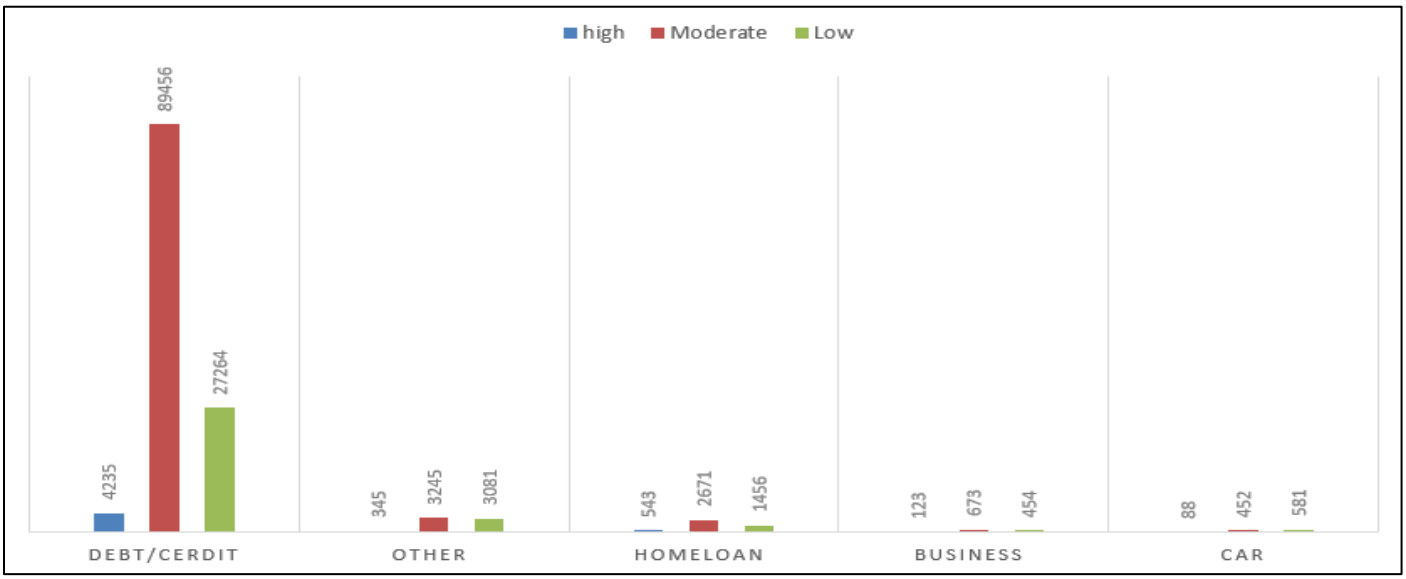

Fig 1 Annual Income vs Purpose Thus, grouping the Category that is High, Moderate and Low.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1924

- *Inference from the Figure 1 is as follows:*

✓ The field of debt consolidation exhibits the highest dispersion; more people in the moderate group apply for loans.

✓ Applicants in the low and moderate categories equally attempt to obtain auto loans and other purposes.

➢ *Trust Customer Classification*

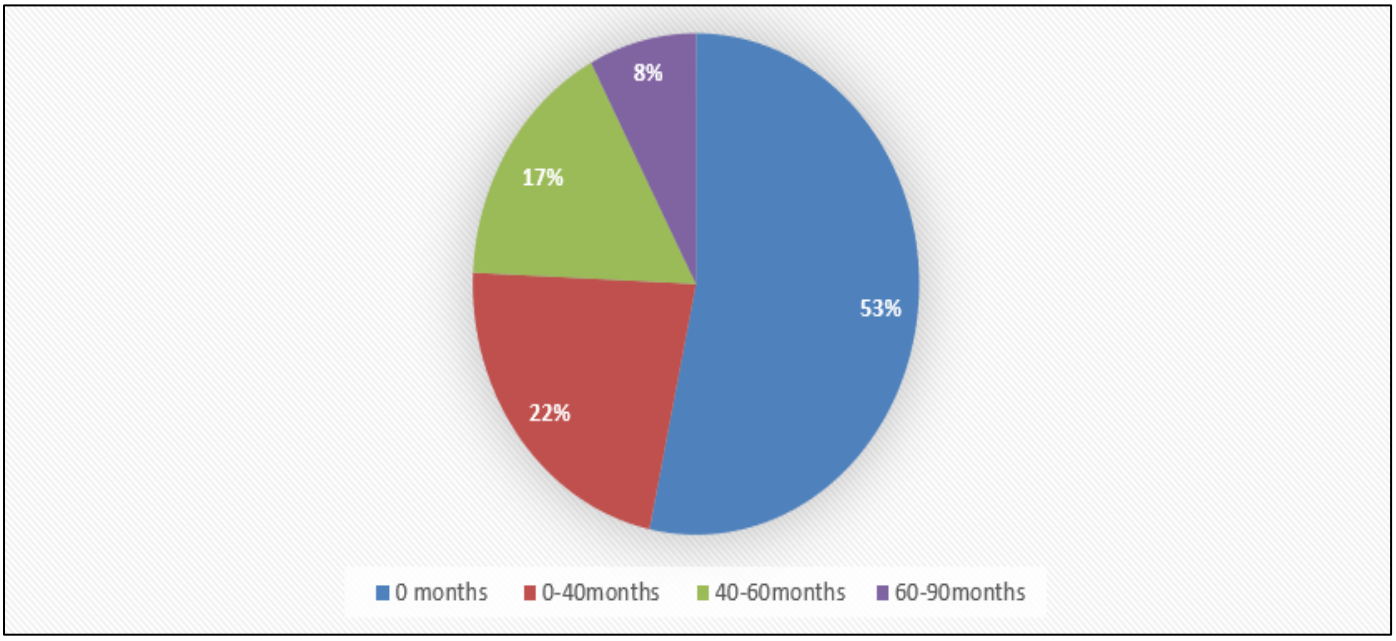

Fig 2 Trust Customers from the Figure 2 it is inferred as follows

- Many clients who have no delinquencies have sought for loans, implying or subtly suggesting that the applicant has a possibility of being approved because they are delinquent-free. Approximately 53.3% of applicants are the outcome.

- Additionally, it may be deduced that as the number of overdue months rises, fewer persons apply for loans. This indicates that the applicant's prospects of having their loan application approved are minimal.

➢ *Loan Term Vs Delinquent Months*

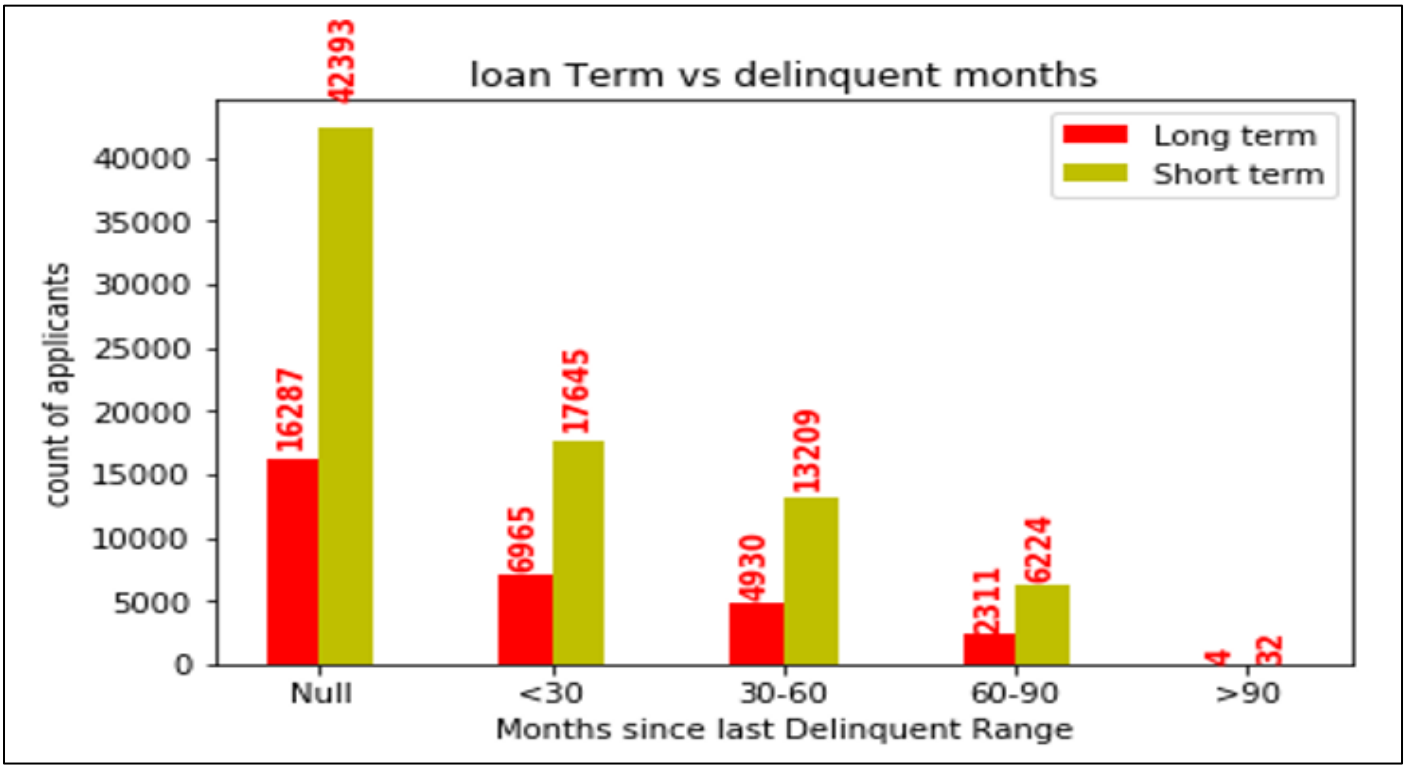

Fig 3 Loan Term vs Delinquent Months

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1924

- Figure 3 deals with the customers who can pay the loan within term period against customers who cannot repay their monthly depts within the particular term.
- From the Figure 3 it can be concluded that:
- This analysis can find a higher number of customers who are able to repay without deliquiates and for short term.
- Almost all applicants who are even delinquent more than 90 months prefers only short term.
- Applicants who delinquent more than 90 months are less in number and it indirectly conveys that their loan will never be sanctioned and if its yes, the applicant will not be able to pay it back

➢ *Loan Term Vs Credit Category*

Individuals with credit scores between 300 and 850 fall into one of the following categories: poor, fair, good, very good, and undefined. Applicants without a credit score fall into the undefined category. Those with scores between 300 and 579 fall into the poor category, those with scores between 580 and 669 into the fair category, and those with scores between 670 and 739 into the good category.

Fig 4 Loan Term vs Credit Category

By combining the generated column credit category and loan term, Figure 4 displays the loan repayment period against credit score under different categories. The following has been inferred from Figure 4:

First-time loan applicants prefer short-term loans because the lender does not run a credit check, making it easier for them to obtain loans; clients with good and very good credit scores favour short-term payback periods, in contrast to those with fair credit scores.

➢ *Loan Term Vs Years in Current Job*

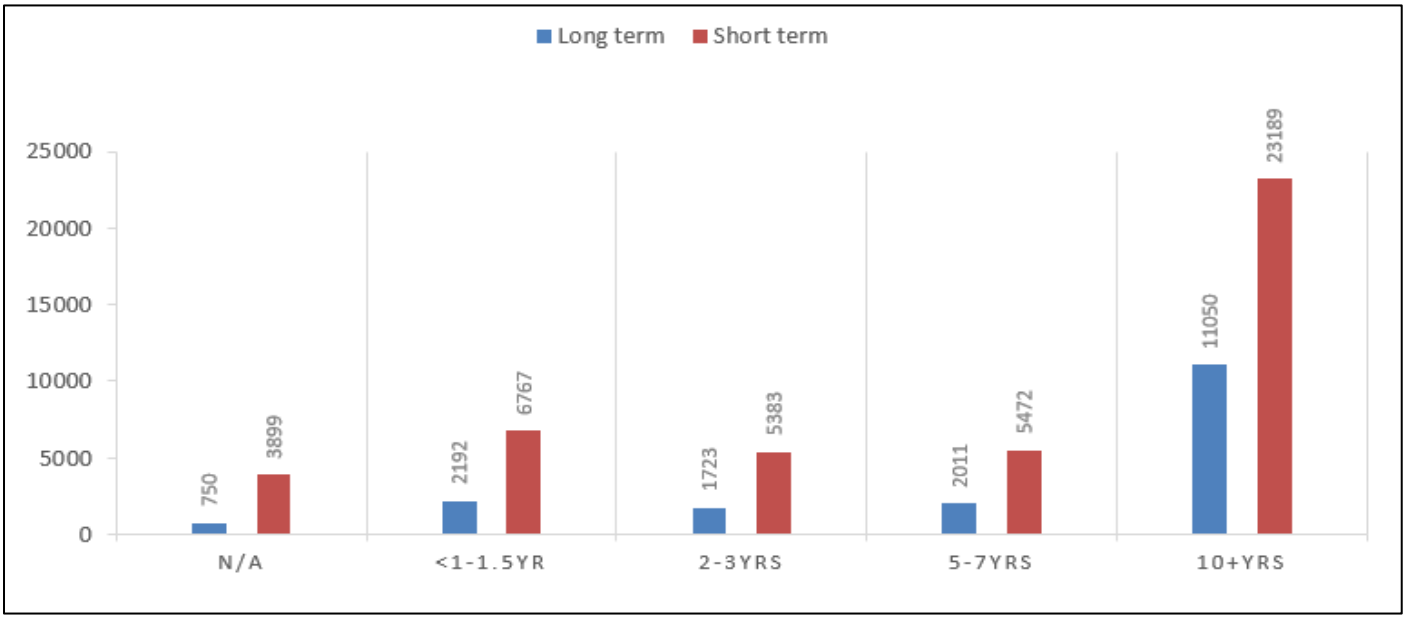

Fig 5 Loan term vs Years in Current Job

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1924

- *From the Figure 5 it is Concluded that:*

✓ The number of applicants with varying years of experience in their current position is plotted against the loan payback time in Figure 5.
✓ The following conclusions are drawn from Figure Applicants with different years of experience in the same job apply for loans for a brief period of time;
✓ Applicants who are new to the field also apply for loans for a brief period of time.
✓ This suggests that long-term loans are taken out by those who have not yet launched their own businesses and can only be repaid once those businesses are profitable.
✓ Long-term goals carry a higher risk to the money lenders and are therefore not reading.

➢ *Loan Payment Chances Vs Home Ownership*

There are three types of loan payment chances: can pay, may pay, and not payable. One can determine if a person would be able to repay a loan by deducting their monthly debt from their current credit balance. In other words, applicants with balances under 50,000 are classified as not payable, applicants with balances between 50,000 and 3 lakhs are classified as may pay, and applicants with balances over 50,000 are classified as may pay. Figure 6 leads to the conclusion that,

- Those who rent a residence are classified as not payable under ownership credentials.
- The largest number of loan applications are from applicants who have a mortgage on their home.

IV. CONCLUSION AND FUTURE WORK

➢ *Final thoughts and Future Work*

The classification and analysis of the loan applicants' characteristics is the paper's primary goal. Seven distinct graphs were created and displayed after a thorough examination of the data set and the banking industry's limitations. Numerous deductions and inferences have been drawn from the graphs, including the fact that most loan applicants chose short-term loans and that clients primarily applied for loans in order to consolidate their debt.

In the future, this paper work may be expanded to a higher level. machine learning algorithm-based predictive model for loans, where each paper graph's output can be interpreted as a separate criterion for the algorithm.

REFERENCES

[1]. A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.
[2]. A. J. Hamid and T. M. Ahmed, "Developing Prediction Model of Loan Risk in Banks using Data Mining".
[3]. G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".
[4]. A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".
[5]. M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model for Retail Bank Loan Applications Using Decision Tree Data Mining Technique", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5(3), pp. 705- 718, 2016.
[6]. Gerritsen, R. (1999). Assessing loan risks: a data mining case study. IT professional, 1(6), 16-21.
[7]. Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. Expert systems with Applications, 37(1), 534-545.
[8]. https://en.wikipedia.org/wiki/Exploratory_data_analysis
[9]. https://pandas.pydata.org/pandas-docs/stable/
[10]. https://www.experian.com/blogs/ask-experian/credit-education/score- basics/what-is-a-good-credit-score/