


## RESEARCH ARTICLE



# Predictive Analytics for Personalized Debt Management: Leveraging Machine Learning to Provide Actionable Financial Advice

Amir Ahmad Dar<sup>1,\*</sup> , Trushal Hirani<sup>1</sup>, Vanshita Arora<sup>1</sup> and Sukriti Kakkar<sup>1</sup>

<sup>1</sup>Department of Statistics, Lovely Professional University, India

**Abstract:** In today's complex financial environment, managing personal debt effectively has become a significant challenge, often leading to increased loan defaults. This study aims to develop a machine learning-based framework for personalized debt management by analyzing borrower data to identify risk levels and offer tailored financial advice. The scientific novelty of this research lies in its integration of both supervised and unsupervised learning techniques to gain deeper insights into the characteristics of defaulters and predict their risk levels. This methodology improves prediction accuracy and interpretability and applicability in real-world lending. The study offers actionable strategies for debt reduction, optimized spending, and personalized financial planning based on risk profiles. The findings can support financial institutions in refining credit risk assessment models, promoting responsible lending, and contributing to the achievement of broader sustainability goals through improved financial inclusion and stability. Unsupervised learning techniques, such as K-nearest neighbors (KNN), DBSCAN, and rule-based methods, were applied to cluster defaulters based on their risk profiles. These clustering methods allowed us to distinguish various groups of defaulters, providing a nuanced view of risk categories. Financial institutions can use these risk categories to design tailored financial products and adjust lending strategies and policies for lower-risk groups or offer guidance to higher-risk defaulters on areas needing improvement, such as increasing income or enhancing credit scores. Governments in developing countries could make the most use of this study, where most of the population lacks financial knowledge and struggles to get financial help from private institutions once they are categorized as defaulters. For instance, insurance companies have different policies for different age groups, and financial institutions can also make such policies for different risk levels that benefit both parties in the long term.

**Keywords:** personal finance, debt management, machine learning, financial advice, credit risk assessment

## 1. Introduction

The growing complexity of personal finance has created a need for innovative systems that identify loan default prospects. This study builds a reliable data-driven system to create borrower risk categories that improve credit authorization and debt management decisions [1]. Current statistics show that numerous people continue without receiving the specialized financial guidance needed to control their borrowing patterns and handling of funds. The foundation of financial stability depends on effective personal finance planning, which allows people to prepare investments and manage debt [2]. Machine learning (ML) algorithms evaluate extensive financial data sources to find hidden patterns, which lead to automated decisions that offer debt-related financial guidance for people struggling with debt, while appropriate advice enables them to create debt repayment strategies and improve their credit ratings and prevent defaults [3]. Our project seeks to generate actionable advice based on a person's financial status, leveraging data-driven insights derived from ML models. This advice may include recommendations on adjusting spending habits, optimizing loan repayments, or improving creditworthiness, all of which are essential for achieving long-term financial stability and

collateral management [4]. This is a multifaceted problem and involves scaling data, labeling data, selecting relevant features, and applying both supervised and unsupervised learning algorithms to create a predictive and cluster model. For this research, we use various ML techniques such as logistic regression and random forest for feature extraction and then apply KNN and DBSCAN (Density-Based Spatial Clustering of Applications with Noise), among others. These models are employed for loan takers based on their financial behavior and demographic characteristics [5]. The core purpose of this study is to create a scalable, explainable system that categorizes borrowers into risk groups using key financial and demographic indicators. This enables the generation of tailored advice aimed at reducing default risk, improving creditworthiness, and promoting responsible lending. By enhancing debt management strategies and aligning with the Sustainable Development Goals—specifically those related to financial inclusion—our research contributes to building a more equitable and resilient financial ecosystem.

## 2. Data and Preprocessing

### 2.1. Data sources

The dataset contains 255,347 observations and 16 variables, with “Default” as the target variable. The features encompass various borrower and loan characteristics relevant to predicting the

\*Corresponding author: Amir Ahmad Dar, Department of Statistics, Lovely Professional University, India. Email: [amir.30646@lpu.co.in](mailto:amir.30646@lpu.co.in)

**Table 1**  
**Feature variable and data description**

Variables	Data type	Description
Age	Integer	The age of the borrower.
Income	Integer	The annual income of the borrower.
Loan Amount	Integer	The amount of money being borrowed.
Credit Score	Integer	The credit score indicates creditworthiness.
Months Employed	Integer	The number of months the borrower has been employed.
Num Credit Lines	Integer	The number of credit lines the borrower has opened.
Interest Rate	Float	The interest rate for the loan.
Loan Term	Integer	The term length of the loan is in months.
DTI Ratio	Float	The debt-to-income ratio indicates the borrower's debt to their income.
Education	String	The highest education level (PhD, master's, bachelor's, high school).
Employment Type	String	The employment status (full-time, part-time, self-employed, unemployed)
Marital Status	String	The marital status of the borrower (single, married, divorced).
Has Mortgage	String	Whether the borrower has a mortgage (Yes or No).
Has Dependents	String	Whether the borrower has dependents (Yes or No).
Loan Purpose	String	The purpose of the loan (home, auto, education, business, other).
Has Cosigner	String	Whether the loan has a cosigner (Yes or No).
Default	Integer	The binary target variable indicating whether the loan defaulted (1) or not (0).

likelihood of loan default. These variables are designed to support the development of models aimed at understanding and forecasting default behavior. Table 1 presents the variables used in the study and their descriptions.

Table 1 shows the research features that include three types of variables to study demographic information and financial details and loan characteristics. All variables selected for this analysis exhibited specific potential influence on credit risks as well as loan payment behaviors. The variables Age and Months Employed measure financial maturity and job stability because these factors have a proven relationship with creditworthiness. Income, together with Loan Amount along with Interest Rate, demonstrates how well borrowers can repay their debt and how much it will cost them to borrow funds.

The assessment of borrower reliability heavily depends on Credit Score as an essential metric, together with DTI Ratio, which reveals how a borrower distributes their income between debts. The categorical variables Education, Employment Type, and Marital Status provide contextual knowledge to disclose social patterns that affect default behavior. The binary flags used for Has Mortgage, Has Dependents, and Has Cosigner provide additional information, which indirectly affects the financial pressure and financial support networks. Borrowers with comprehensive financial valuations based on statistical and descriptive data attributes allow ML models to generate more precise predictions.

In Figures 1 and 2, we can see uniformity of variables of several features (e.g., age, income, loan amount, credit score) exhibiting uniform distributions, which may indicate that the data was either preprocessed or collected in such a way that all ranges of values are evenly represented. Features such as marital status, employment type, education, and loan purpose are categorical, but their distributions are uneven (e.g., many more married borrowers or full-time employees). These categories might be important for clustering or risk modeling. The default feature shows a significant imbalance, which could lead to challenges in classification tasks, especially when modeling the likelihood of default. This imbalance may require re-sampling techniques or adjustments during model training. Some features, like the number of credit lines and loan purpose, have sparse distributions with distinct peaks. This could impact how these features are treated in a clustering or rule-based model.

The most notable differences between defaulters and non-defaulters are in credit score and interest rate, where defaulters tend to have lower credit scores and slightly higher interest rates. Many other features (such as income, loan amount, loan term, education, marital status, etc.) show very similar distributions between the two groups, indicating that these variables alone may not be strong predictors of default. Credit score and interest rate could be more predictive for identifying default risk, while other features may require more advanced modeling techniques or interaction effects to be useful in predicting default.

Figure 3 focuses on the relationship with the target variable, Default; we see that age has a weak negative correlation ( $-0.17$ ), suggesting that older individuals are slightly less likely to default on loans. Income also shows a slight negative correlation ( $-0.099$ ), indicating that those with higher incomes are somewhat less prone to default. Conversely, Loan Amount ( $0.087$ ), Interest Rate ( $0.13$ ), and DTI Ratio ( $0.19$ ) display mild positive correlations with Default, implying that larger loans, higher interest rates, and higher debt-to-income ratios are associated with an increased likelihood of default. However, no variable exhibits an overwhelmingly strong correlation with Default, signifying that the default risk is influenced by a combination of factors rather than being driven by any single variable. Beyond the relationships with the target, most features exhibit weak correlations with each other. For example, Income and Loan Amount have a weak negative correlation ( $-0.007$ ), suggesting that individuals with higher incomes may tend to borrow slightly smaller loan amounts. Similarly, Credit Score and Default have a weak negative correlation ( $-0.034$ ), indicating that individuals with lower credit scores have a marginally higher likelihood of defaulting, though this effect is not particularly strong. Overall, the heatmap suggests a relatively low degree of multicollinearity among features, as most feature-feature correlations are weak, which is advantageous for predictive modeling since it means the variables can contribute uniquely to the model. The moderate correlations observed between key variables and Default (such as Age, Income, Loan Amount, Interest Rate, and DTI Ratio) highlight their potential importance as predictors in a loan default prediction model, but it also suggests that no single factor dominates the prediction, necessitating a combination of features to effectively assess the

Figure 1  
Distribution of quantitative variables

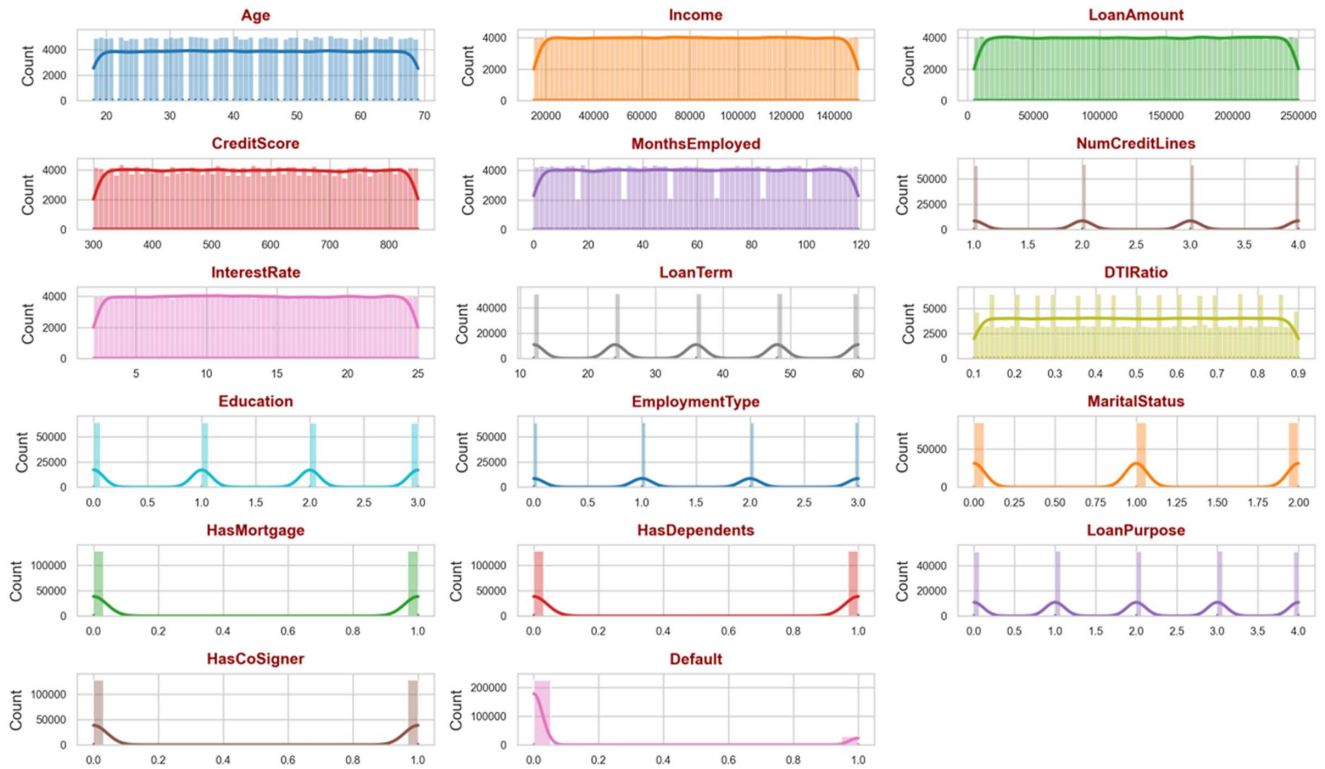


Figure 2  
Distribution of variables vs default

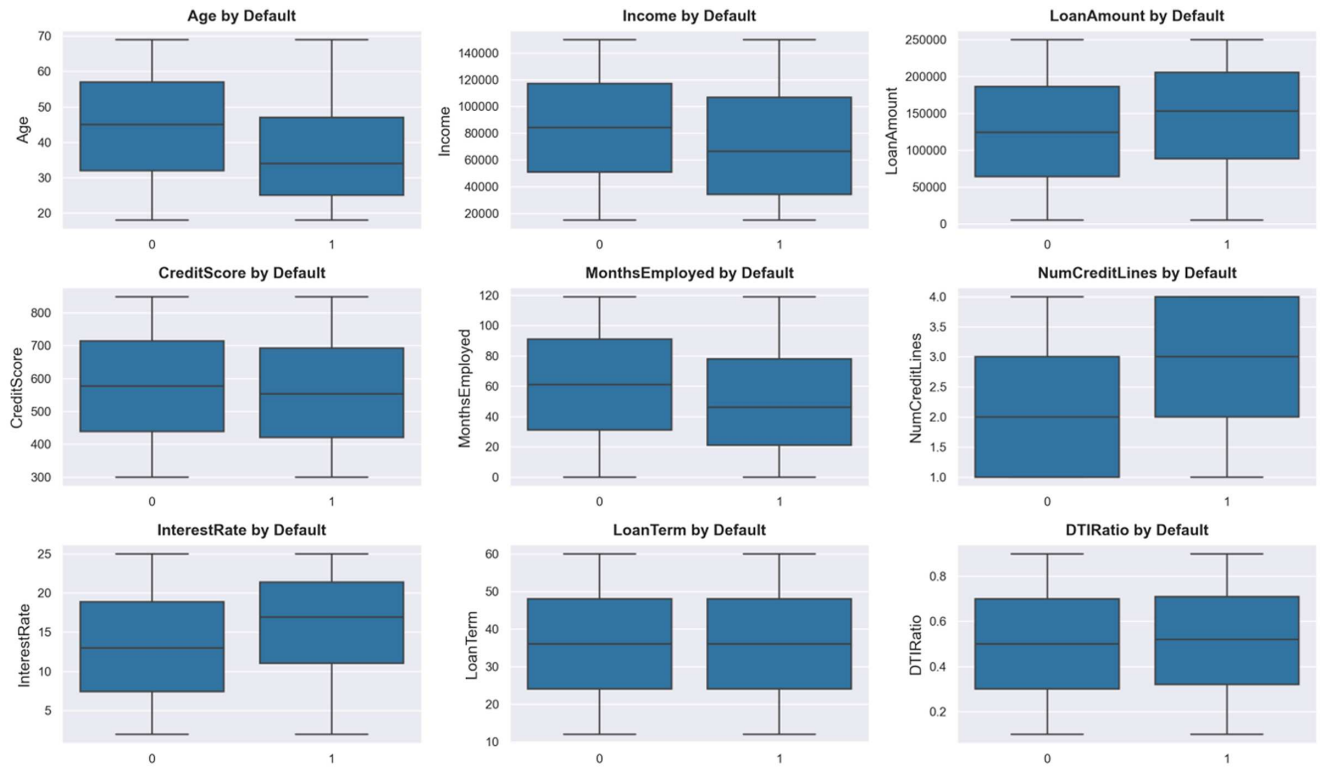


Figure 3  
Correlation coefficient heatmap between features

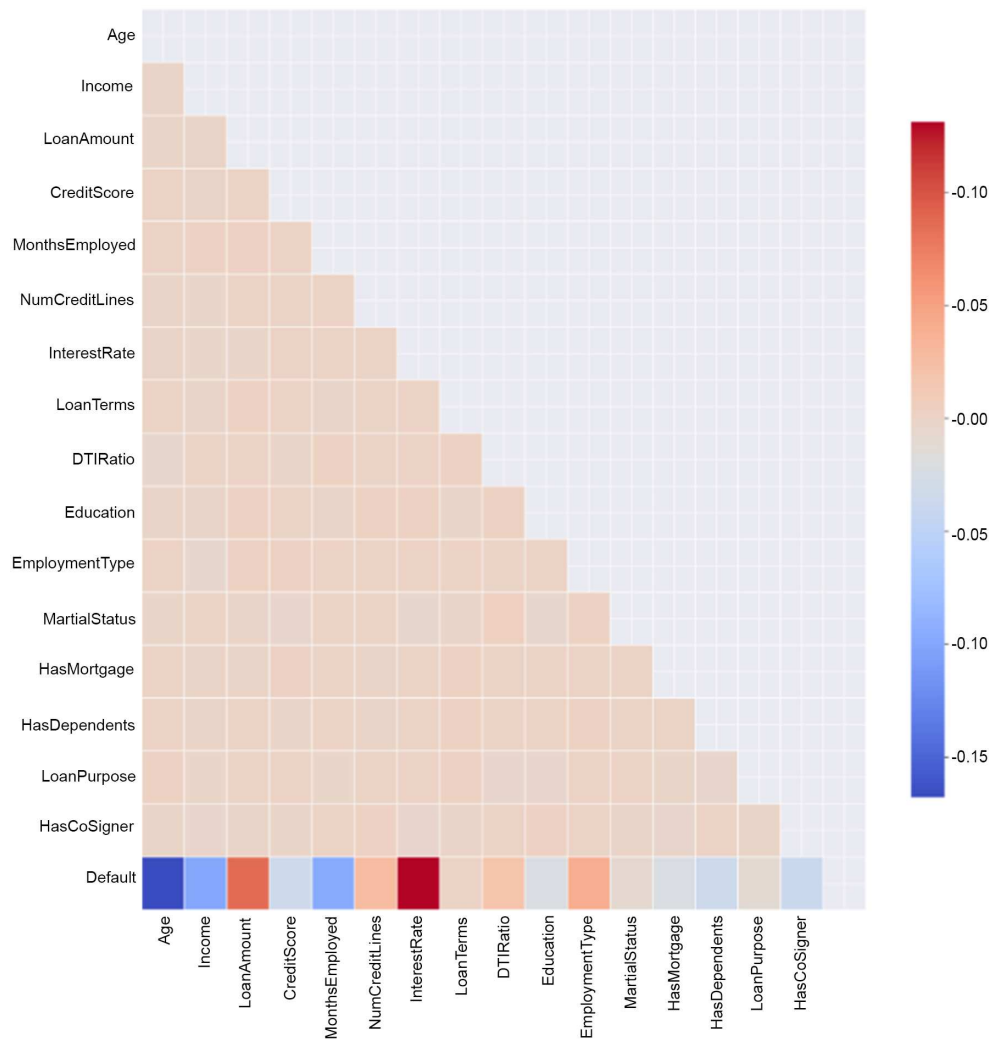


Table 2  
Summary table highlighting the observed trends in defaulter vs non-defaulter groups

Feature	Defaulters (Trend)	Non-Defaulters (Trend)
Credit Score	Lower	Higher
Interest Rate	Slightly higher	Slightly lower
Income	Similar	Similar
Loan Amount	Similar	Similar
Loan Term	Slight variation, no clear trend	Slight variation, no clear trend
Marital Status	Mixed distribution	More stable across categories
Employment Type	More full-time and self-employed	Similar with slightly more full-time

**Note:** Observations are based on visual inspection of Figures 1 and 2.

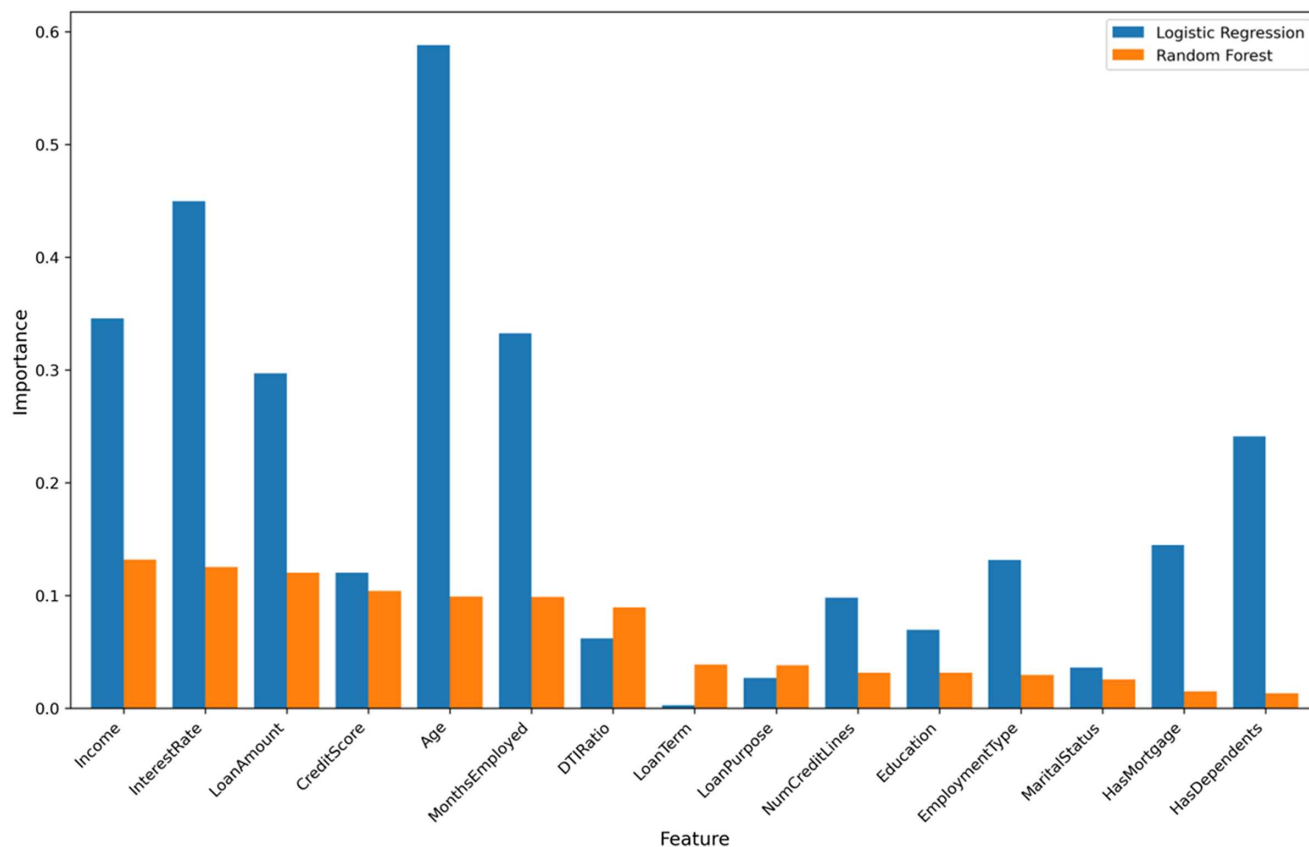
likelihood of default. Table 2 presents the trends in the defaulters and non-defaulters groups.

2.2. Feature extraction

The procedure of selecting and transforming raw data into important variables (features) for ML models remains essential for achieving effective predictive performance [6]. The selection of important features helps the model achieve improved accuracy

when predicting loan defaults and delivering customized financial guidance. The success of the analytical model depends on feature selection because it decreases dimensionality and boosts both model performance and understanding ability [7]. We implemented normalization and scaling with categorical encoding and correlation analysis to determine and pick the most significant variables in our evaluation [6, 7, 8, 9]. The validation of feature importance was accomplished through random forest modeling, which provided scores to confirm feature relevance [8]. Researchers computed

**Figure 4**  
**Comparison of top 10 important features (feature importance based on random forest and logistic regression)**



feature importance through the Gini importance metric, which serves as a mean decrease in impurity derived from random forest models [9]. Random Forest Gini importance tracks how much node impurity (based on the Gini index) declines from all trees in the ensemble because of each feature. The algorithm choice selection for Gini importance calculation was based on its computational speed and native integration within the random forest structure, which delivers a straightforward understanding of the predictor's impact on error reduction.

The assessment demonstrates that income, interest rate, and loan amount serve as the most influential elements for default prediction, yet credit score and employment stability play important roles. Such knowledge can lead to improved model development, together with the identification of crucial factors in financial lending risk, so financial institutions can develop targeted solutions for specific groups.

The model flags those persons with high debt-to-income ratios and unfavorable payment records as being high risk, causing it to deliver counseling that focuses on debt reduction and loan avoidance. A person with a solid income alongside a proper credit score could get advice about basic financial guidance alongside investment insights.

Figure 4 displays a ranked list of the ten most important features that predict default based on random forest model scores because random forest delivers more reliable and logical results than logistic regression (logistic regression identifies age as most significant, but random forest selects income). The three most important variables identified by the analysis are Income together with Interest Rate and Loan Amount, which confirms earlier identified correlation patterns. Together, the Credit Score and DTI Ratio and Months

Employed variables demonstrate a strong influence that helps businesses determine borrower repayment potential. The lesser-ranked variables of Employment Type and Loan Purpose retain potential value in combination with other contributing factors. The visual display demonstrates that credit risk involves multiple dimensions, which validates the selection of random forest ensemble methods to handle variable interrelationships.

### 3. Methodology

#### 3.1. K-means clustering

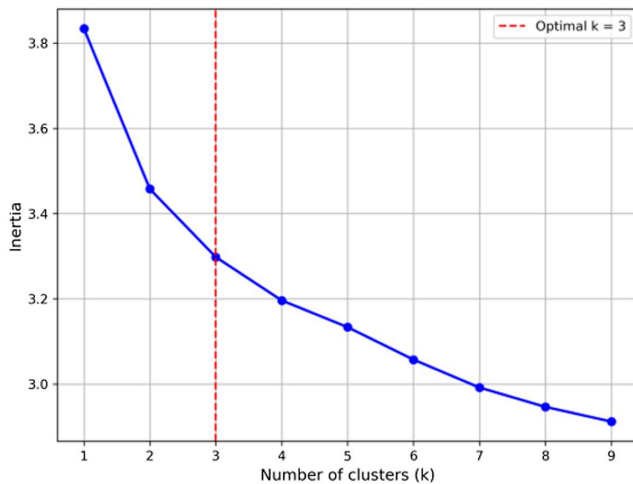
We evaluated loan defaulter risk categories using K-means clustering as an unsupervised learning approach [10]. The main purpose was to organize people according to essential financial attributes to help financial institutions discover potential patterns within default risk patterns. Financial institutions widely use K-means clustering because it selects appropriate clusters ( $k$ ) that produce relevant segmentation groups for intervention purposes [11].

We began by identifying the optimal number of clusters for the analysis. The elbow method helped us determine the value of  $k$  that allowed the within-cluster sum of squares (inertia) to decline at a steadily slowing rate according to [12]. Beyond the “elbow” point, the fit quality remains unchanged by adding additional clusters.

The elbow method was employed to evaluate inertia values between 1 and 9. Between values  $k = 1$  and  $k = 3$ , the inertia decreased sharply, yet the pace of decrease became moderate when  $k$  reached  $k = 3$ . The data analysis suggested that three distinct clusters should provide optimal results when interpreting the major distinctions in data without creating overlapping risk-grouping categories.



**Figure 5**  
Elbow method to determine optimal k



The determination of an appropriate cluster count relies on two evaluation metrics shown in Figure 5, which use inertia and silhouette score evaluation.

- 1) Inertia: The measure reflects the total squares of the separation distances between points and their assigned cluster center points. The distance between points and their cluster centroids becomes shorter when inertia values decrease, thus indicating more focused clustering. The value of inertia does not consistently select the optimal number of clusters because its measurement decreases with the addition of more clusters.
- 2) Silhouette score: The score evaluates the clustering gap between the different groups. It is the scoring method that arranges the distance calculations through a neat differentiation between neighborhood cluster average point distances and cluster internal point distances. The metric normalizes the difference between two averages, where one measures the nearest cluster distances and the other measures intra-cluster distances. The value of the silhouette score approaching 1 indicates effective cluster segregation. The clusters appear distinct from one another when the silhouette score approaches 1. Nevertheless, scores below or equal to 0 indicate clusters that display overlapping or indistinct definitions.

These metrics augment each other because they measure cluster compactness and separation from other clusters, respectively, to evaluate clustering quality.

We generated silhouette scores across various cluster number conditions. The cluster separation quality is measured by the silhouette score, which generates results between  $-1$  and  $1$  [13]. The cluster separation quality improves as the score increases. The maximum silhouette score that resulted from different parameter adaptations reached approximately  $0.1$ , showing that data separation worked partially, but cluster boundaries were not clearly defined. Our findings from the elbow method verify the selection of three clusters as most appropriate because the score matches our conclusions [14]. Three distinct risk levels exist within the loan defaulter population according to the clusters our analysis created. Financial cluster elements enable risk determination of different population segments by examining their income levels and credit together with their loan amounts.

The created clusters show promise for risk segmentation in financial institutions although the silhouette score remains low.

Financial institutions should consider supplementing these clusters with new variables and mixed analytical methods to enhance their effectiveness. These preliminary cluster groupings offer a preliminary identification system to distinguish high-risk from low-risk individuals even though additional improvements can optimize cluster separation.

### 3.2. Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is an alternative clustering technique, particularly effective for datasets with complex structures, as it clusters data points based on density, making it suitable for identifying irregularly shaped clusters and handling noise [15]. DBSCAN relies on two primary parameters: epsilon (eps), which defines the maximum distance between two points to be considered in the same neighborhood, and the minimum number of points (MinPts) required to form a dense region or cluster.

We explored a broad spectrum of eps values and MinPts combinations in an attempt to capture meaningful density-based clusters within the loan data; after experimenting with several parameter values, the best results were achieved with eps values of  $0.6$  and the best minimum samples of  $2$ . However, the outcome, as indicated by the silhouette score of  $-0.297$ , still pointed toward poor clustering performance. This negative score suggests that DBSCAN struggled to find defined clusters, even with the tuned parameters, and therefore highlights that this might not be suitable for segmenting this data, as it fails to capture meaningful separation between groups in this context [16].

### 3.3. Possible causes for overlapping

Based on the unsupervised learning attempts – particularly clustering methods like K-means and DBSCAN – the data appears to resist clear separation into distinct clusters. Here are the conclusions and potential causes related to customer behavior and data-specific issues that might explain these results:

#### 1) Complexity of customer behavior

Loan default risk as a spectrum: Risk isn't always neatly divided into clusters. Instead, it exists on a continuum. Customers can display a range of behaviors that gradually shift between low and high risk rather than fall into distinct categories. This complexity makes it difficult for clustering algorithms to identify hard boundaries between groups of customers. Borrowers may behave similarly in some respects (like income, loan amount, or credit score), but these factors alone might not differentiate them sufficiently. For example, two borrowers with similar incomes may default for entirely different reasons (one due to poor financial management and another due to sudden job loss). Clustering methods struggle to account for these hidden variables, which affect the overall risk of default [17].

#### 2) High-dimensional nature of financial data

Even after performing feature extraction to reduce the number of features, the data still seems to retain a high level of complexity. In high-dimensional data, many clustering algorithms become less effective because distances between points tend to be more uniform. This means clusters are less distinct. Many features in financial datasets (e.g., income, credit score, loan term) are not independent. These interactions may mask the natural structure of the data, leading to overlapping clusters and overfitting the model [18], especially if the relationships between these features are nonlinear.

Clustering methods like K-means, which assume linear separability, may underperform in such cases.

### 3) Lack of well-defined groups

Unlike other domains where customers or data points might naturally fall into clear segments (e.g., in customer segmentation for marketing), financial risk data has inherently blurred boundaries. Risk groups like “low-risk” or “high-risk” borrowers may not have sharply distinct characteristics. The transition from one group to another may be gradual, which prevents clustering algorithms from effectively distinguishing them. Features such as mortgage, income, and interest rate likely overlap among defaulters and non-defaulters. For example, a borrower with a high income might still default due to personal financial mismanagement, making income alone a poor clustering criterion. Similarly, employment type (whether a borrower is salaried or self-employed) might not be sufficient to classify someone as high or low risk.

### 4) Potential data-related issues

When most samples belong to the non-defaulter category (a frequent data distribution), the clustering methods often prioritize the majority group, which complicates cluster distinction among smaller populations. Unsupervised learning algorithms become tricked by unbalanced data when trying to cluster defaulters with non-defaulters because it eliminates the distinction between different risk levels. Several pieces of information in the data collection lack sufficient value in identifying defaulting customers. The inclusion of Loan Purpose and Dependents information in the analysis generates unimportant clusters because these variables show no relationship to default rates.

### 5) Economic or behavioral shifts

The financial decision-making patterns of borrowers undergo changes because of economic adjustments along with modifications in lending regulations. After the pandemic, people developed financial patterns that were not present in previous record data. Traditional clustering methods risk difficulty capturing stable patterns because their application occurs during times of continuous changes in variable relationships. Exogenous elements like inflation, together with unemployment rates and fiscal policies of the government, create an unpredictable effect on borrower conduct. The external factors create random variation that complicates cluster identification processes in the data.

### 6) Potential customer behavior hypotheses

Homogeneity in certain features: It is possible that borrowers across risk levels display similar characteristics in key features like income, credit score, or employment. For instance, customers in both the high- and low-risk categories might have similar credit scores, and their risk levels could be more influenced by less visible factors like personal spending habits, family support, or specific terms of their loan. It may be that financial risk, particularly default behavior, doesn't naturally form clusters in this dataset. Instead, it might be driven by complex, overlapping factors that defy clear categorization. Traditional clustering may not be the best fit for problems like this, where decision boundaries are complex, multidimensional, and fuzzy. The difficulty in clustering this dataset suggests that loan default risk is likely more complex than can be captured by unsupervised clustering. The behavior of borrowers is likely influenced by a range of interdependent factors that are not easily separable into distinct categories; hidden features like this can be revealed using a defined rule-based algorithm [19].

## 3.4. Rule-based algorithm

We applied two rule-based models for loan defaulter risk classification through percentiles and z-scores. Rule-based financial indicator categorization enables institutions to separate borrowers into risk categories through essential values such as income data alongside credit score and total loan amount (See Figure 6).

### 3.4.1. Percentile-based rule system

The percentile-based rule categorizes risk levels of borrowers through financial thresholds established from quartile points (25th, 50th, and 75th percentiles). The evaluation system uses three variables named Income and Credit Score together with Loan Amount to determine financial stability and repayment ability among borrowers. Financial standing determinations are possible through quartile division, resulting in borrower classifications based on their relative positions regarding finance and loans.

There are three distinct groups according to risk assessment – “low risk” describes borrowers whose income exceeds the 75th percentile and whose credit score exceeds the 75th percentile and whose loan amount falls below the 25th percentile. They show excellent credentials through their substantial earnings and strong credit record and minimal outstanding loans.

Borrowers who have their values lying between the 25th and the 75th percentiles are placed in the “medium risk” group. These subjects stand midway between the average scores, indicating their financial situation is moderate.

High-risk borrowers are defined as those who maintain a below-average income alongside a below-average credit score but utilize loan amounts that surpass average levels. The debt-to-income ratio matches a potentially significant chunk of earnings and borrowing capacity.

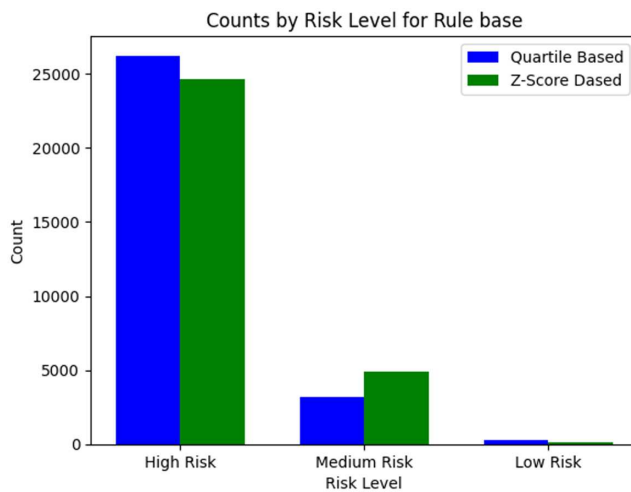
### 3.4.2. Z-score-based rule system

To further refine the categorization, we used a z-score-based rule system, which standardizes each variable based on its mean and standard deviation. Z-scores measure how far a given value is from the mean in terms of standard deviations, making it easier to identify values that deviate significantly from the average. The z-scores allow us to identify borrowers who are significantly above or below the average for each financial indicator. Borrowers with z-scores greater than 1 for both income and credit score (indicating high financial standing) and less than -1 for loan amount (indicating low debt burden) are classified as “low risk.” Borrowers with z-scores between -1 and 1 for all variables are considered “medium risk” as their financial indicators are close to the average. Borrowers with z-scores of less than -1 for income and credit score (indicating financial challenges) and greater than 1 for loan amount (indicating high debt burden) are categorized as “high risk.”

The rule-based approach implements percentile and z-score methods to produce a flexible system that allows managers to make transparent borrower risk assessments. Using percentiles gives risk assessments based on borrower group distribution patterns, but z-score normalization helps find borrowers whose financial metrics differ sharply from the mean. Both approaches work together as a system that allows institutions to categorize rigorously while offering a complete perspective on borrower risk abilities to focus loan distributions and generate actionable default responses.

While prior studies have relied heavily on logistic regression (e.g., [6]) or black-box ML models, our hybrid framework bridges predictive accuracy and interpretability. The rule-based classification offers transparency that traditional models lack, especially

**Figure 6**  
**Quartile-based vs z-score-based rules**



in financial contexts where decision accountability is essential. Recent studies (e.g., [20, 21]) have explored the financial behavior of individuals and institutional dynamics in cash management and financial integration. Integrating such behavioral insights could enhance the applicability of predictive models in real-world lending practices.

#### 4. Result and Conclusion

K-nearest neighbors (KNN) and DBSCAN were initially applied to segment borrowers into risk categories. Although both models have strengths, they struggled to perform adequately in this context due to several key limitations. KNN, being a distance-based algorithm, performed poorly due to the imbalanced nature of our financial data, where high-risk and low-risk borrowers were unevenly represented. This imbalance led to misclassifying a substantial number of borrowers, especially those on the fringe of typical high-risk and low-risk characteristics. Given the large number of features extracted from financial data.

The cluster discovery method DBSCAN needed detailed parameter adjustment of eps and min samples for obtaining adequate performance outcomes. The application of DBSCAN with its optimal parameter setting did not produce distinct groups among borrowers because it generated results that were inconsistent and noisy. The model suffered from over-categorization, classifying many borrowers as noise or grouping them without adequate precision for meaningful insights. We discontinued the use of KNN and DBSCAN as our solution options because of the limitations encountered. Our research led us to develop a different approach because KNN and DBSCAN failed to accommodate the complexities of risk profiles and did not meet financial institution priorities. We built a rule-based system that integrates expert-derived payment ratios and score indicators with financial thresholds to resolve the performance limitations of KNN and DBSCAN.

The rule-based model differs from ML methods because it enables the easy definition of risk categories without needing the same extensive training and parameter adjustment demands found in ML applications. Financial institutions apply flexible threshold adjustments based on their lending criteria and market conditions together with individual borrower behavior in order to achieve specific business needs. Under the rule-based method, financial institutions alongside borrowers have precise pathways to making

decisions. The classification system relies on explicit criteria so external stakeholders can clearly observe how the financial characteristics of borrowers affect their positioned risk group. Users build stronger trust in models because they can follow the operations that drive each recommendation. The rule-based model functions efficiently at the computation level, so it works optimally with large datasets alongside real-time applications. Financial organizations thrive using this system because it enables fast assessment of large numbers of loan applicants with high reliability for risk determination. The rule-based model accepts industry data and financial expertise because it operates according to established lending practices. The model alignment enables institutions to embed both historical business knowledge and domain field experience into the assessment system, leading to a stronger, more realistic evaluation process.

We develop several proposals to build better relationships between borrowers and financial institutions, which establish transparent and trusted deals beneficial to both parties.

- 1) Borrowers gain better debt management by learning about critical financial markets including their debt-to-income ratio and credit score. Financial institutions should provide educational materials about personal finance together with debt repayment methods to teach budgeting basics to their borrowers. Open dialogue between borrowers and lenders creates opportunities for debt-term negotiations during financial difficulties or allows borrowers to find temporary solutions that prevent default situations. Relationships with financial institutions improve the chances of obtaining better loan conditions along with individualized repayment solutions. Borrowers need to check their credit scores regularly because higher scores provide better loan conditions and make default less probable. Everyone should practice on-time bill payments while decreasing their debts and refraining from making many credit application requests to build stronger creditworthiness.
- 2) Financial institutions can strengthen borrower relationships by clearly disclosing loan terms, interest rates, and associated fees. Borrowers show more trust in lending practices when institutions reveal the methods used to determine borrower risk assessment such as rule-based models. The data from rule-based risk assessments permits institutions to provide specialized support and recommendations, which help borrowers better comprehend their finances along with ways they can improve their situation. The institution would offer specific guidance to high-risk borrowers about expense control and income growth, whereas low-risk borrowers would get recommendations about saving or investing. Financial organizations must teach responsible borrowing by creating appropriate credit limitations and offering loans, which stick to borrowers' payment abilities. Financial institutions that practice responsible lending reduce defaults while building customer relationships, which helps maintain continuous financial stability.
- 3) Bankers should build financial wellness tools through mobile apps, which provide customized guidance about credit score improvement and debt reduction and high-interest loan evasion together with fintech alliances for implementing these features into personal finance applications that generate instant user interaction.

While existing literature supports the importance of these themes (e.g., [22, 23, 24]), the above strategies are proposed independently by the authors and grounded in our empirical findings and model outputs. Through this model, we bridge the gap between predictive analytics and practical financial advice. Borrowers gain



a deeper understanding of their financial health and receive tailored recommendations for improvement, while financial institutions can make informed lending decisions that foster long-term, trust-based relationships. Ultimately, this collaborative approach benefits both parties: borrowers gain financial stability and trust in the lending process, and institutions reduce default rates and build loyal customer bases.

Looking ahead, this rule-based framework can serve as a foundation for more sophisticated debt management systems, potentially integrating real-time data and behavioral insights to further enhance predictive accuracy and personal financial guidance. By continuously refining our model based on empirical findings and borrower feedback, we can contribute to a financial ecosystem that is fair, transparent, and sustainable for all stakeholders involved.

In future studies, we intend to enhance model precision. Furthermore, we plan to explore the integration of natural language processing techniques to analyze borrower communication (e.g., loan applications, support chats) as supplementary risk indicators; however, this is a long-term goal, and the current proposed framework for rule-based algorithms has to be implemented to make a stronger dataset. Finally, testing the system in real-world settings (via pilot collaborations with fintech firms or microfinance organizations) will be crucial to validate its practical impact and refine the model for deployment.

By combining predictive analytics with practical financial recommendations, this research aims not only to reduce default rates but also to empower individuals to take proactive control of their financial health.

## Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

## Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

## Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

## Author Contribution Statement

**Amir Ahmad Dar:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration. **Trushal Hirani:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Vanshita Arora:** Writing – review & editing, Visualization. **Sukriti Kakkar:** Writing – review & editing, Visualization.

## References

- [1] Nwaimo, C. S., Adegbola, A. E., & Adegbola, M. D. (2024). Predictive analytics for financial inclusion: Using machine learning to improve credit access for under banked populations. *Computer Science & IT Research Journal*, 5(6), 1358–1373. <https://doi.org/10.51594/csitrj.v5i6.1201>
- [2] Reddy, Y. J. (2024). Data analytics in personal finance management: Optimizing decision-making in spending, savings, and investments. *International Journal of Research Publication and Reviews*, 5(2), 1826–1832.
- [3] Alma Çağrı, B., & Coşkun, E. (2021). A longitudinal systematic review of credit risk assessment and credit default predictors. *Sage Open*, 11(4), 21582440211061333. <https://doi.org/10.1177/21582440211061333>
- [4] Le, C. H. A., & Nguyen, H. L. (2019). Collateral quality and loan default risk: The case of Vietnam. *Comparative Economic Studies*, 61(1), 103–118. <https://doi.org/10.1057/s41294-018-0072-6>
- [5] Sai, Y. C. (2024). Human-centric machine learning for personal finance. *International Journal of Research Publication and Reviews*, 5(1), 2549–2556.
- [6] Visalakshi, S., & Radha, V. (2014). A literature review of feature selection techniques and applications: Review of feature selection in data mining. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 1–6. <https://doi.org/10.1109/ICCIC.2014.7238499>
- [7] Ruano-Ordás, D. (2024). Machine learning-based feature extraction and selection. *Applied Sciences*, 14(15), 6567. <https://doi.org/10.3390/app14156567>
- [8] Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems*, 66(3), 1575–1637. <https://doi.org/10.1007/s10115-023-02010-5>
- [9] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [10] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28(1), 100–108. <https://doi.org/10.2307/2346830>
- [11] Bo, S., & Xiao, M. (2024). Time-series k-means in causal inference and mechanism clustering for financial data. In *Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*, 515–520. <https://doi.org/10.1145/3703187.3703274>
- [12] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1), 31. <https://doi.org/10.1186/s13638-021-01910-w>
- [13] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [14] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2–3), 107–145. <https://doi.org/10.1023/A:1012801612483>
- [15] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.
- [16] Shutaywi, M., & Kachouie, N. N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, 23(6), 759. <https://doi.org/10.3390/e23060759>
- [17] Cai, F., Le-Khac, N.-A., & Kechadi, T. (2016). Clustering approaches for financial data analysis: A survey. *arXiv Preprint: 1609.08520*
- [18] Lai, L. (2020). Loan default prediction with machine learning techniques. In *2020 International Conference on Computer*

- Communication and Network Security, 5–9. <https://doi.org/10.1109/CCNS50731.2020.00009>
- [19] Soui, M., Gasmi, I., Smiti, S., & Ghédira, K. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. *Expert Systems with Applications*, 126, 144–157. <https://doi.org/10.1016/j.eswa.2019.01.078>
- [20] Tekin, B., & Gör, Y. (2022). Cash conversion cycle and its relationship with profitability as a cash management tool in companies: An application on companies trading in Borsa Istanbul. *Theoretical and Applied Economics*, 29(3), 113–130.
- [21] Tekin, B. (2021). Modeling the relation of financial integration-economic growth with GMM and QR methods. *Economic Studies*, 30(8), 32–47.
- [22] Lusardi, A., & Messy, F.-A. (2023). The importance of financial literacy and its impact on financial wellbeing. *Journal of Financial Literacy and Wellbeing*, 1(1), 1–11. <https://doi.org/10.1017/flw.2023.8>
- [23] Adrian, T., Shabsigh, G., & Khan, A. (2020). *Transparency makes central banks more effective and trusted*. Retrieved from: <https://www.tbsnews.net/world/global-economy/transparency-makes-central-banks-more-effective-and-trusted-113854>
- [24] Kinniry, F. M., Jaconetti, C. M., DiJoseph, M. A., Zilbering, Y., Bennyhoff, D. G., & Yarwood, G. (2020). *Putting a value on your value: Quantifying Vanguard Adviser's Alpha in the UK*. Vanguard Research. Retrieved from: <https://consilium-ifa.co.uk/wp-content/uploads/2024/10/quantifying-vanguards-advisers-alpha.pdf>

**How to Cite:** Dar, A. A., Hirani, T., Arora, V., & Kakkar, S. (2025). Predictive Analytics for Personalized Debt Management: Leveraging Machine Learning to Provide Actionable Financial Advice. *FinTech and Sustainable Innovation*. <https://doi.org/10.47852/bonviewFSI52025591>