IFA
Iranian Journal of Finance
Print ISSN 2676-6337
Online ISSN 2676-6345

# Predicting Corporate Loan Defaults Using Deep Learning Algorithms and a Comparative Analysis with Linear Models: A Case Study of a Major Commercial Bank

**Mohammad Ahmadi Azar**

PhD. Candidate, Department of Financial Management, Faculty of Financial Engineering, Kish campus, University of Tehran, Kish, Iran. (Email: ahmadiazar@ut.ac.ir)

**Reza Tehrani***

*Corresponding Author, Prof., Department of Financial and Insurance Management, Faculty of Management and Accounting, College of Management, University of Tehran, Tehran, Iran (Email: rtehtrani@ut.ac.ir)

**Seyed Mojtabi Mirlohi**

Assistant Prof., Department of Finance, Faculty of Management, Shahrood University, Shahrood, Iran. (Email: mirlohism@shahroodut.ac.ir)

Iranian Journal of Finance, 2026, Vol. 10, No.1, pp. 1-42.
Publisher: Iran Finance Association
doi: https://doi.org/10.30699/ijf.2025.444059.1460
Article Type: Original Article


Received: August 21, 2025
Received in revised form: October 24, 2025
Accepted: December 14, 2025
Published online: January 01, 2026

## Abstract

In today's complex economic landscape, accurately predicting events such as customer loan defaults presents a significant challenge for financial institutions. Traditional methods have shown limitations in accuracy, prompting the adoption of data-driven machine learning techniques for enhanced predictive capabilities. This study investigates the efficacy of novel machine-learning algorithms compared with linear models for predicting loan defaults at a major commercial bank. Data from over six thousand customer loan files spanning 2019 to 2022 were collected, cleaned, and clustered based on key loan indicators. The accuracy of predicting loan defaults was first evaluated using popular machine learning classification models, including LightGBM, XGBoost, Multilayer Perceptron, and Logistic Regression, and XGBoost performed best. After that, prediction accuracy was evaluated using various time-series machine learning algorithms, with a particular focus on a combined Gradient Boosting and Long Short-Term Memory (LSTM) approach. Results indicate that the combined algorithm outperforms traditional linear models, showing a substantial 40% improvement over the ARIMA algorithm in predicting loan default behavior. This study underscores the potential of advanced machine learning techniques to enhance predictive accuracy in the banking sector, offering valuable insights for risk assessment and financial decision-making.



**Keywords:** Loan Default, Financial Prediction, XGBoost, Long Short-Term Memory, Machine Learning

**JEL Classification:** G21, C53, C45, C38


## Introduction

The banking industry plays a vital role in the global economy by providing economic services to businesses and individuals. One of the key services banks offer is lending, enabling individuals and economic entities to finance their investments and activities through debt instruments. However, granting loans involves risks, the most significant being default. Default, in this context, refers to the borrower's inability to repay the principal and interest on the loan in accordance with the agreed terms. One of the crucial indicators used in calculating expected losses is the probability of default for a specific customer. Credit default occurs when a customer fails to meet their payment obligations under the contractual agreement between the bank and the customer. Therefore, a model that can predict customers with a high probability of default is highly attractive for banks (Granstroem, 2019). Various credit models, including traditional and AI-based systemic models, have been developed for customer credit scoring before providing loans. As banks expand their online service offerings and the demand for online loan services increases, the need for a monitoring system to oversee customers' credit behavior during the loan repayment period becomes crucial. Research consistently emphasizes the importance of monitoring customers and ensuring successful loan repayment throughout the loan tenure. For example, Branzoli and Fringuellotti (2022) highlight the importance of bank supervision in loan repayment, especially depending on the type of loan granted, demonstrating that with precise bank monitoring during this period, the default rate decreases. In previous studies, more non-systematic, linear methods were used to predict customer credit behavior and loan defaults. Methods such as linear regression, logistic, and Probit regression have been widely used to predict loan defaults.

To address the research Hypothesis, the following steps have been taken:

1. Data Preprocessing and Sorting:

Given that the data in the bank's loan system is maintained using an in-house model, the data needed to be preprocessed and sorted to be usable for the research. This involved refining and repairing the data, handling issues such as imbalanced data and outliers, and performing feature engineering. Furthermore, the data were normalized as required for use in machine learning algorithms.

2. Formatting Data for Machine Learning Algorithms:

To leverage machine learning algorithms, the data must be stored in a specific format. This step ensured that the data were not only cleaned and restored but also formatted for use in machine learning algorithms.

3. Algorithm Prediction Power Assessment:

After preparing the data, the predictive power of each deep learning algorithm was compared with that of others. This assessment evaluated the algorithms' predictive capabilities.

4. Evaluation of the results Using ROC-AUC:

The evaluation and measurement of results employed the "Area Under the

ROC Curve (ROC-AUC)" to assess the performance of the algorithms. This metric is widely used in classification problems to evaluate the tradeoff between the true positive rate and the false positive rate.

5. Optimizing Long Short-Term Memory (LSTM) Algorithm:

The final stage involves evaluating the impact of the deep learning algorithm, specifically LSTM, on loan default prediction. The LSTM's performance is then compared with that of traditional linear algorithms. The assessment and ranking of different methods are based on the "Mean Squared Error" metric. This comprehensive approach ensures a thorough investigation of the performance of various machine learning algorithms for predicting loan defaults, with a specific focus on optimizing the LSTM algorithm and benchmarking against linear models. Also, by improving the LSTM with the best classification model from item 3, all results are rechecked to find the best algorithm among linear models, LSTM, and the LSTM-XGBoost Model.

In the next section, a review of literature and theoretical foundations is presented. Subsequently, the research methodology is explained using the CRISP-DM method, a standard model for data mining research. Finally, a summary of the results is provided. The next section of the document reviews existing literature and the theoretical foundations relevant to the research topic. Following that, the research methodology is elucidated, with a specific reference to the CRISP-DM method, a widely used model in data mining research. The concluding Segment is the last part of the article.

## Theoretical Foundations

In banking-centered economies, especially in developing countries, the banking system plays a pivotal role in the economy and significantly influences economic growth, stability, and development. The central operation in banking is the management of loan portfolios, and if not handled properly, it can have severe financial consequences for both institutions and the broader economy. One of the critical aspects of the banking system is the management of the primary asset and function of banks, namely, loan portfolio management, its importance, challenges, and macroeconomic determinants affecting the credit risk of loan portfolios. The management of loan portfolios has garnered considerable attention, particularly given its direct implications for the financial health and stability of banks, especially amid economic uncertainties. Bahat et al. (2020) highlight the importance of loan portfolio management, especially amid economic uncertainties. Financial institutions always contend with the risk of counterparties not adhering to their commitments, which can jeopardize their performance. Bahat focuses on banks' loan processes, emphasising the tools and techniques they use to identify and manage default risk. Ultimately, it is essential to note that macroeconomic conditions significantly impact credit risk in banks. In this regard, Milleris (2013) emphasizes that the primary goal of credit risk management in banks is managing the bank's loan portfolio to operate within an acceptable risk threshold.

## Credit Default

Credit default in banking refers to a borrower's failure to fulfil their repayment obligations under the agreed terms and conditions of a loan. Credit defaults can take various forms, including non-payment, delayed payment, loan restructuring (forbearance), or bankruptcy. Non-payment occurs when the borrower fails to repay the loan in full. Delayed payment occurs when the borrower misses the payment deadline but eventually makes it within the specified deadline. Loan restructuring is a scenario in which the borrower and the lender agree to modify the loan terms to make them more manageable for the borrower. Bankruptcy means the borrower is unable to fulfil their financial commitments and legally declares their inability to repay their debt.

A non-performing loan is another term for a credit default. It is used when the loan maturity has passed for a certain period (60 days), and there is doubt about its complete repayment. When a bank encounters signs that a loan may not be fully repaid (e.g., due to some missed payments in the past), it must identify the loss and create reserves. In accounting terms, this operation leads to a reduction in the loan value (a negative value on the asset side) and to the incurring of expenses (a decrease in net profit and, consequently, a reduction in shareholders' equity). For this purpose, the bank estimates the expected loss. The expected loss is based on the loan value (i.e., the face value at the time of default) multiplied by the probability of default.

Furthermore, the bank considers that even if a default occurs, it may recover part of the loan (e.g., through bankruptcy proceedings). Therefore, the previous amount is multiplied by the estimated portion of the loan that may be recovered in the event of default, resulting in the loss due to default.

Calculation of Expected Loss (EL) in Credit Default:

The expected loss is calculated based on the following formula:

[EL = PD times LGD times EAD = PD times (1-RR) times EAD]

Formula 1: Expected Loss Formula

Where:

- (PD) is the Probability of Default.

- (LGD) is the Loss Given Default.

- (EAD) is the Exposure at Default.

- (RR) is the Recovery Rate ((RR = 1 - LGD)).

According to World Bank statistics, defaulted loans should be between 2% and 5%. However, among commercial banks in the country, this figure is significantly higher, as observed in the following chart. The chart, extracted from World Bank reports, illustrates the proportion of defaulted Loans to total Loans through 2022. As shown in the chart, the COVID-19 pandemic had a significant impact on this indicator, prompting governments to take corrective measures to mitigate its effects (World Bank Report, 2022).

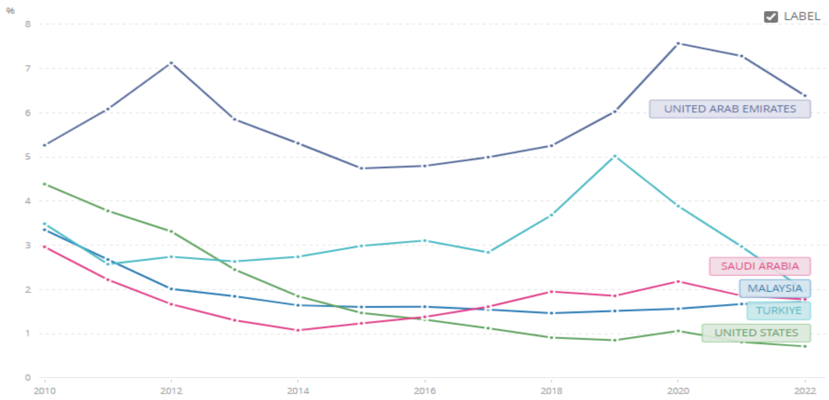


**Figure 1. The Total rate of Loan Default in Selected countries (World Bank Report, 2022)**

For comparison, the Non-Performing Loans Percentage in Iranian banks ranged from 4.52% to 25.65% in 2022. It is worth noting that government loans from banks for various government projects are considered current loans even if the repayment period has passed. As shown in the chart, various banks consistently adopt control policies to ensure proper Loan provisioning and prevent Loans from being transferred to non-current categories. Simultaneously, regulatory authorities employ control tools to minimize the proportion of banks' non-performing assets, ensuring that resources circulate as needed to support the country's macroeconomic plans (Report of the Banking and Financial Institute for the years 2018-2022). Predicting loan defaults has a rich history, dating back to the early 20th century. In 1909, John Moody took steps to classify and rank credit risks associated with promissory notes. Subsequently, researchers sought to measure the risk of loan defaults, drawing on Moody's work in promissory notes (Case, 2003). Another study, conducted by Fitzpatrick (1931) in 1932, focused on predicting a customer's inability to repay loans. One of the most significant research efforts in predicting loan default risk was Edward Altman's 1968 study. Altman introduced the Z-score to predict the financial distress of borrowing companies, laying the foundation for many classic studies in credit risk and customer prediction (Edward Altman, 1968). In the field of predicting loan defaults, researchers have used various statistical techniques. Numerous studies and applications have been conducted in credit scoring to identify good and bad customers for banks. Judgmental methods in credit scoring gradually gave way to parametric and non-parametric methods due to their errors and time-consuming nature (Thomas, 2000). Parametric methods such as logistic and Probit regression, discriminant analysis, and logistic regression were initially used in credit scoring. Later, non-parametric methods and data mining techniques, such as decision trees, neural networks, and expert systems, gained attention (Sabzevari et al., 2006).

## Machine Learning

Samuel (1959) defines machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed." Mitchell (1997) further expands the concept of machine learning, describing it as a computer program that learns from experience, represented by a task set T, based on performance indicators P, and a dataset E, if and only if overall performance improves by performing tasks defined in T and evaluated with indicators P. Research indicates that artificial intelligence has the potential to revolutionize the financial and banking industry, improving efficiency, accuracy, and customer experience. Konigstrofer (2020) suggests that AI can be used in commercial banking to reduce lending losses, enhance payment processing security, automate compliance-related tasks, and improve customer targeting. Malali (2020) also highlights the potential of artificial intelligence in the banking and financial industry, stating that it can redefine institutional operations, create innovative products and services, and enhance customer experience. Boukherouaa and colleagues (2021), in a report published by the International Monetary Fund, outline the primary areas of application for artificial intelligence and machine learning in finance and banking as follows:

- Prediction: Artificial intelligence and machine learning models, compared to traditional statistical and econometric models, provide flexibility and can help discover non-obvious relationships between variables, enhancing the tools used by institutions. Evidence suggests that machine learning methods often outperform linear regression-based methods in terms of prediction accuracy and robustness.

- Investment and Banking Services: In the financial sector, recent advances in artificial intelligence and machine learning have had the most significant impact on the asset management industry. AI and machine learning, along with related technologies such as introducing new participants to the market (e.g., product customization), improving customer interfaces (e.g., chatbots), enhancing analysis and decision-making methods, and reducing costs through automation, are transforming the industry.

- Risk Management and Compliance: Supervisory technology has gained greater importance in response to tightened regulatory regulations and increased compliance costs after the global financial crisis in 2008. Advances in artificial intelligence and machine learning in recent years have transformed risk management and compliance by leveraging extensive datasets, often in real time, and automating compliance decision-making, thereby improving compliance quality and reducing costs.

- Preventive Supervision: While decisions ultimately depend on the judgment of supervisors, artificial intelligence and machine learning technologies have been widely used in data collection and analysis. Many authorities of countries participating in the Financial Stability Board currently use machine learning and natural language processing tools to analyze, process, validate, and accept data (Financial Stability Board, 2020). With artificial intelligence, supervisors can gain deeper insights into diverse types of data and make more informed, data-driven decisions. It can identify patterns that humans cannot recognize, thereby improving the quality of supervision. It can also provide real-time anomaly marking to supervisors, making supervision more agile (European Central Bank, 2019).

In the McKinsey Management Consulting Group's 2021 report on "Building the Bank of the Future," the application of machine learning in the customer lifecycle is illustrated as follows:

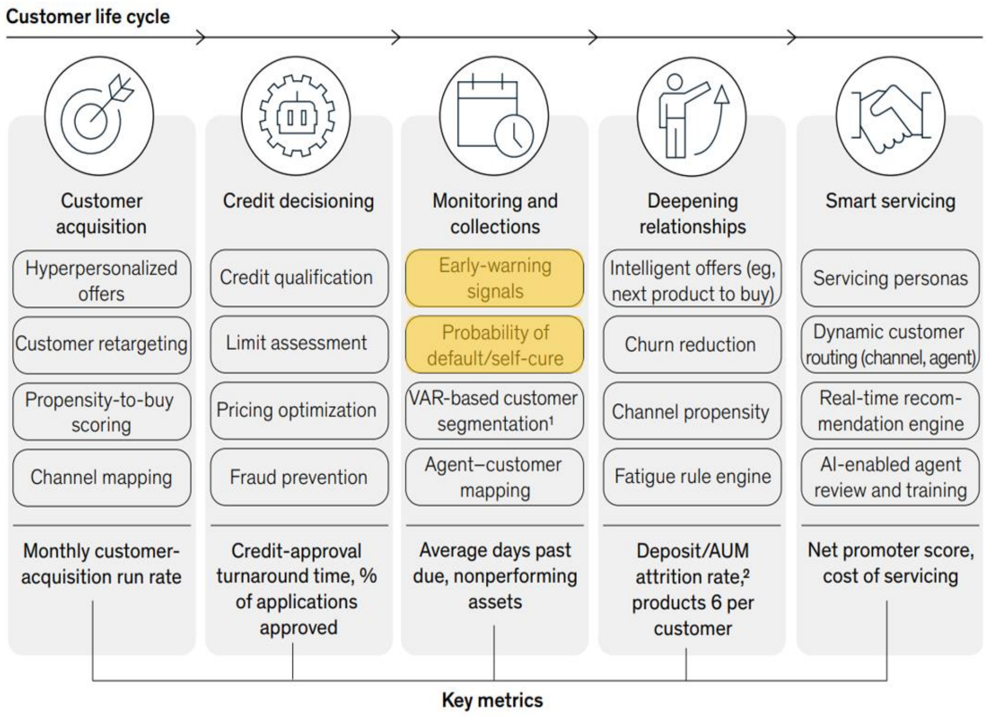


**Figure 2. The application of machine learning in the customer lifecycle**

As evident in the figure, one of the most crucial functions of machine learning in AI-equipped banks is the early detection of the likelihood of default and the customer's inability to repay Loans, a topic explored in this study.

The algorithms used in this research are introduced as follows:

### Gradient Boosted Decision Trees

Gradient Boosted Trees, an algorithm proposed by Chen and Guestrin, are widely used in text classification, product categorization, and customer behavior prediction. This algorithm has achieved state-of-the-art results in many machine learning competitions (Chen & Guestrin, 2016).

The prediction results of the model formed by K decision trees are as follows:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \tag{1}$$

In a context where xi represents the ith input sample, ŷi is the predicted value calculated through the mapping of the function fk, and F is a set of mapped relationships. The optimization objective and the loss function are defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2. \quad \mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{2}$$

Here, l is a differentiable convex loss function that measures the difference between the predicted value ŷi and the target value yi, and Ω is an additional regularization term that helps smooth the model weights to prevent overfitting. The mentioned optimization involves functions as parameters, which cannot be optimized in Euclidean space using traditional methods. ŷi (t) is defined as the i-th prediction at the t-th iteration. Then, the loss function is defined as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t). \tag{3}$$

For a specific tree structure q(x), the optimal weights (ωj*) for the leaf nodes can be calculated using the following formula:

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}. \tag{4}$$

The corresponding optimal objective function can be calculated as follows:

$$\tilde{\mathscr{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T. \tag{5}$$

LightGBM is a supervised learning algorithm that has demonstrated exceptional performance in predictive models since its introduction among researchers in 2016. It offers notable advantages in speed and computational power, distinguishing itself from other algorithms. Winners of many global data analysis competitions have often utilized LightGBM. Developed by Microsoft, LightGBM (Light Gradient Boosting Machine) is a free, open-source distributed gradient boosting framework for machine learning. This framework, based on decision tree algorithms, is employed for ranking, classification, and various other machine-learning tasks, with a primary focus on performance and scalability.

In contrast to other implementations, LightGBM uses a leaf-wise tree growth strategy rather than a row-wise one. It selects the leaf that offers the greatest reduction in loss, resulting in faster execution without compromising accuracy. Additionally, LightGBM incorporates two novel techniques: one-sided gradient-based sampling and exclusive feature bundling. These techniques enhance accuracy by selecting samples with large gradients for processing while aggregating exclusive features to reduce data dimensionality.

Long-Short Term Memory (LSTM), on the other hand, addresses the limitations of traditional neural networks in handling sequential or time series data. Unlike fully connected neural network models, LSTM networks exhibit recurrent connections within layers, enabling them to process sequential information effectively. The adaptability of LSTM lies in its ability to adjust its weights based on input thresholds dynamically, the forget gate, and the output, thereby preventing issues such as vanishing or exploding gradients.

The figure below illustrates the architecture of an LSTM unit, highlighting its unique ability to process sequential data.

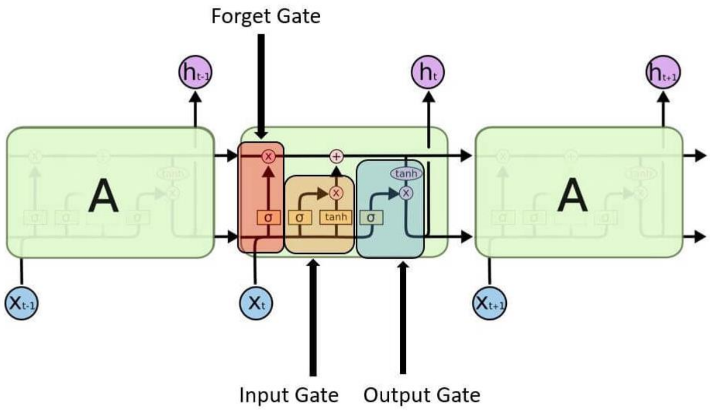


**Figure 3. LSTM Cell Schema, Source: Javatpoint.com**

## Literature Review

The exploration of customer default risk prediction models extends back approximately a century. In 1909, John Moody initiated credit risk classification and ranking concerning bonds, sparking subsequent research endeavors aimed at gauging loan default risk (Keyes, 2003). Fitzpatrick furthered this pursuit in 1932 by delving into studies aimed at predicting customers' inability to repay (Fitzpatrick, 1931). Notably, Edward Altman's pivotal research in the late 1960s introduced the Altman Z-score, revolutionizing debt default forecasting for borrowing companies and laying the groundwork for classical studies in credit and customer risk analysis (Altman, 1968). A comprehensive review of the literature reveals the use of various statistical techniques to predict loan defaults. Credit scoring has attracted numerous studies and applications, with judgmental methods gradually yielding to parametric and nonparametric approaches due to their errors and time-consuming nature (Thomas, 2000).

Parametric methods such as Logit and Probit regression, discriminant analysis, and logistic regression initially dominated credit scoring, later giving way to nonparametric methods and data mining techniques such as decision trees, neural networks, and expert systems (Sabzevari et al., 2006). Multiple discriminant analysis emerged in the late 1960s as a statistical method for categorizing observations into predefined groups, notably combining financial ratios for company analysis and performance assessment. Logit and Probit models, introduced in the late 1970s, were developed to predict the financial and credit status of companies, with logistic regression assigning weights to independent variables to estimate the likelihood of default (Jacobsen & Rosbach, 2003).In recent years, the advent of artificial intelligence applications, especially in the financial and banking sectors, has led to the widespread adoption of machine learning algorithms for risk management and portfolio optimization in banks. A case study scrutinizing domestic research journals over the past five years highlighted a significant increase in articles within the financial domain employing machine learning models, signaling a growing trend towards integrating advanced analytical techniques in banking practices.

**Table 1. Domains of Articles Presented Using Artificial Intelligence Tools in the Recent 5-Year Period**

| Financial Segments | Using AI Models |
|---|---|
| Currency | 8 |
| Bank | 31 |
| Stock Market | 211 |
| Insurance | 2 |
| Monetary | 2 |
| Inflation | 3 |
| Stock | 2 |
| Gold | 3 |
| Macro Economics | 11 |
| Financial | 36 |
| Tax | 12 |
| Oil | 3 |
| Sum | 328 |

Considerable research has been conducted in loan prediction, employing various models to predict loan defaults. Purkazemi et al. (2017) investigated factors influencing loan defaults and used neural network models to predict the likelihood of default for individual applicants at Pasargad Bank. The study examined the loan profiles of 470 customers from a statistical population of 25,342 customers in Pasargad Bank branches in Tehran from 2013 to 2014. The results indicated that the neural network approach achieved 92% prediction accuracy for loan default likelihood, highlighting the significant impact of variables such as financial history and collateral type on the prediction. In a study by Mohammadian-Haji Kurd et al. (2016) on Bank Tejarat, the credit risk of legal customers was examined using a Support Vector Machine (SVM) and a hybrid Genetic Algorithm model. The aim was to develop a credit risk assessment model for the bank's legal customers, employing SVM and a genetic algorithm to optimize financial variables for 282 companies that received loans from Bank Tejarat between 2008 and 2011. The results revealed that the Genetic Algorithm enhanced the Support Vector

Machine model's performance in identifying good and bad customers and predicting credit risk. Ameri Sayavashi et al. (2021) addressed potential challenges in loan repayment and their impact on fundamental issues, including liquidity risk, interest rate risk, and bankruptcy. They estimated the probability of default on granted loans to legal entities in non-bank deposit institutions using linear, non-linear Logit and Probit models, and the Z-score. Their research, involving qualitative and financial variables from 400 legal customers of banks and non-governmental credit institutions during 2016 to 2019, indicated that the variables used effectively explained credit status, with all models successfully predicting the likelihood of default by over 80%, with the logistic model exhibiting slightly higher performance.

Emekter et al. (2015) used logistic regression to predict loan default, identifying credit scoring features, renewable credit lines, debt-to-income ratio, and credit rating as important predictors. Buttaru et al. (2016) found that decision trees and random forests outperformed logistic regression in predicting credit card fraud. Fitzpatrick and Moyes (2016) demonstrated that boosted regression trees outperform logistic regression in predicting mortgage defaults, while Deng (2019) identified 20 variables with the greatest impact on loan default using logistic regression. Zhu et al. (2019) linearly weighted the predictions of gradient-boosted tree models, XGBoost, and LightGBM, concluding that they effectively handle imbalanced, high-dimensional, and sparse samples.

The table below summarizes other relevant research in this field, along with the algorithms used:

**Table 2. Summaries of Relevant Research Based on Prediction Default**

| Used Algorithm | Published Date | Authors | Subject |
|---|---|---|---|
| Deep Learning | 2011 | Ribeiro, B. and Lopes, N. | Deep belief networks for financial prediction |
| Deep Learning | 2016 | Giesecke, K., Sirignano, J. and Sadhwani, A. | Deep learning for mortgage risk. Working Paper |
| Deep Learning | 2017 | Luo, C., Wu, D. and Wu, D. | A deep learning approach for credit scoring using credit default swaps |
| Deep Learning | 2018 | Kvamme, H., Sellereite, N., Aas, K. and Sjursen, S., | Predicting mortgage default using convolutional neural networks |
| Deep Learning | 2018 | Hamori, S., Kawai, M., Kume, T., Murakami, | Ensemble Learning or Deep Learning? Application to Default |

| | | | |
|---|---|---|---|
| | | Y. and Watanabe, C | Risk Analysis |
| Logistic Regression | 2020 | Amaro, M.M. | Credit scoring: Statistical issues and evidence from credit-bureau files |
| Statistical Models | 2020 | Dastile, X.; Celik, T.; Potsane, M. | Statistical and machine learning models in credit scoring: A systematic literature survey |
| Naïve Bayse Model | 2015 | Lessmann, S.; Baesens, B.; Seow, H.V.; Thomas, L.C. | Benchmarking state-of-the-art classification algorithms for credit scoring: An updated research |
| Multicriteria Models | 2021 | Liu, W | Enterprise Credit Risk Management Using Multicriteria Decision-Making. |
| Multicriteria Models | 2021 | Pla-Santamaria, D.; Bravo, M.; Reig-Mullor, J.; Salas-Molina, F. | A multicriteria approach to manage credit risk under strict uncertainty |
| K-Nearest Clustering | 2016 | Louzada, F.; Ara, A.; Fernandes, G.B. | Classification methods applied to credit scoring: Systematic review and overall comparison |
| Decision Tree | 2020 | Teles, G.; Rodrigues, J.J.; Saleem, K.; Kozlov, S.; Rabêlo, R.A. | Machine learning and a decision support system for credit scoring |
| Genetic Algorithm | 2014 | Oreski, S.; Oreski, G. | Genetic algorithm-based heuristic for feature selection in credit risk assessment |
| Support Vector Machines | 2010 | Yu, L.; Yue, W.; Wang, S.; Lai, K.K. | Support vector machine-based multi-agent ensemble learning for credit risk evaluation. |
| Artificial Neural Networks | 1996 | Desai, V.S.; Crook, J.N.; Overstreet, G.A., Jr. | A comparison of neural networks and linear scoring models in the credit union Environment |
| Artificial Neural Networks | 2006 | Lai, K.K.; Yu, L.; Wang, S.; Zhou, L. | Credit risk analysis using a reliability-based neural network ensemble model |
| Bagging | 2010 | Zhang, D.; Zhou, X.; Leung, S.C.; Zheng, J. | Vertical bagging decision trees model for credit scoring |
| Random Forest | 2018 | Zhang, X.; Yang, Y.; Zhou, Z | A novel credit scoring model based on optimized random forest |
| XGBOOST | 2021 | Qin, C.; Zhang, Y.; Bao, F.; Zhang, C.; Liu, P.; Liu, P. | XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring |
| XGBOOST | 2020 | Li, H.; Cao, Y.; Li, S.; Zhao, J.; Sun, Y | XGBoost model and its application to personal credit evaluation |

In the research background, various methods have been employed to identify the most effective predictive model for financial issues, considering factors such as the problem's nature, available data volume, and other influential factors. This study aims to cluster loans and evaluate the prediction accuracy of novel machine learning algorithms compared to traditional, often linear methods. Ultimately, the study seeks to compare the performance of deep learning algorithms with traditional models, assessing whether deep learning algorithms enhance or degrade quality when combined with optimal prediction algorithms.

### Research Hypothesizes

This research endeavors to address the following Hypothesizes by analyzing customer behavioral data during the loan repayment period:

Hypothesis 1: Loan clustering improves the accuracy of loan default predictions.

Hypothesis 2: Among machine learning classification algorithms (XGBoost, Multilayer Perceptron, LightGBM, Logistic Regression), XGBoost provides the most accurate prediction of loan defaults.

Hypothesis 3: Deep learning algorithms, specifically Long Short-Term Memory (LSTM), outperform traditional linear algorithms like Logistic Regression in predicting loan defaults.

Hypothesis 4: Combining XGBoost and LSTM algorithms yields better prediction results than using XGBoost alone.

## Research Methodology

The data used in this study comprises loan tables, repayment records, documents, and other relevant datasets spanning 4 years from 2019 to 2022, obtained from one of the country's 5 major commercial banks. To select a research model, existing studies in the field were reviewed. The CRISP-DM process is the most widely recognized framework for developing data science solutions (Kuto & Dashtbandeh, 2019), although other methods such as DMAIC or SEMMA are also used. In this research, the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is employed for machine learning modeling. The overview of the CRISP-DM process is depicted in the figure below:

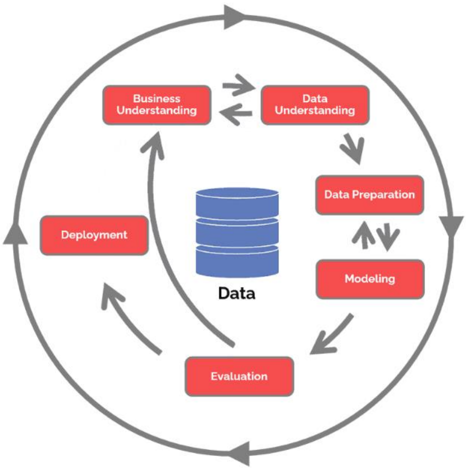


**Figure 4. Schematic CRISP-DM Model**

Based on the CRISP-DM model, the initial phase involves understanding the business domain and identifying and quantifying the necessary data for the study. Subsequently, data cleansing and purification are performed to prepare the data for use in the proposed model. Following this, the results obtained from machine learning algorithms are analyzed and compared. Based on the analysis of the results and the meeting of the research requirements, the algorithms are optimized, and the proposed hypotheses are scrutinized. The following stages of the work have been examined.

## Understanding the Business Domain

The primary challenge addressed in financial institutions, as defined in the problem statement, is optimizing loan portfolio management. Within this context, managing credit risk during the loan repayment period poses significant challenges for financial institutions. Therefore, solutions that enhance a financial institution's ability to systematically control loans, especially for banks dealing with tens of thousands of loans annually, are crucial. This study focuses on investigating the accuracy of predicting loan defaults using new machine learning algorithms.

## Understanding the Data

Informed by the literature review and previous studies, the required data for this research have been identified. Notably, since this research aims to examine customer credit behavior during the repayment period, less emphasis is placed on data used in the credit evaluation stage, which pertains to loan approval or rejection. Instead, the focus is on customer credit behavior during the repayment period. The data utilized in this research are broadly categorized into three groups: contractual data, collateral data, and repayment behavior data. Given that bank data is structured by internal systems, the necessary variables were extracted from the general schema of bank loan data for the years 1398 to 1401. This dataset encompasses information from approximately 17,000 loan files, 146,000 installment repayment records, and about 14,000 contract documents. Since the Murabaha contract is widely used in banking loans and its behavior is well-defined compared to other Islamic contracts, this research specifically focuses on data related to Murabaha contracts. Of the total registered loans, 394 do not have instalment records in the instalment table and are therefore excluded from the analysis. These loans account for 90,862 instalment records. To identify loan defaults, the time difference between the installment due date and the receipt date is examined. According to the literature, a delay of more than 60 days between the due date and receipt date is considered a default. Based on this criterion, the status of delayed payments is determined.

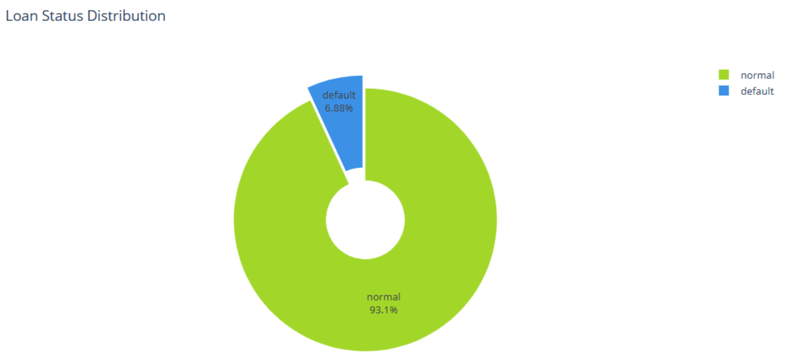


**Figure 5. Comparison of the Percentage of Defaulted Loans to Total**

One segment of the research model focuses on clustering loans to identify similar behaviors, regardless of repayment conditions. For this purpose, six Loan contract features are utilized for clustering. In the subsequent section, 22 variables are used to predict defaults. Ultimately, the dependent variables in this research are the delay time in the repayment of instalments by legal customers exceeding 60 days and the status change of loans at maturity.

## Data Preparation

Data preparation involved several steps after extracting the necessary data from the existing schema. Key data-cleansing operations included checking for incomplete data, identifying outliers, balancing imbalanced data, and modifying variables through feature engineering. Additionally, data were normalized and categorical data converted to numerical values. The following outlines these steps:

### Incomplete Data

Due to supervisory policies governing data entry in banking systems and ongoing monitoring to mitigate incomplete data in loan files, the data required for this research was complete, obviating the need for corrective actions. Moreover, the decision tree algorithms employed in this research exhibit robustness in handling incomplete data, a notable strength. However, after constructing the required variables for the research, certain rows that failed to meet the conditions for variable construction were assigned null values. Given the sensitivity of decision tree and other linear algorithms to null values, a remedial value of 0 was assigned to these values, with negligible impact on outcomes.

### Outlier Data

To identify and address outlier data, the Local Outlier Factor (LOF) method was employed. This method, introduced by Markus M. Breunig and colleagues in 2000, is a valuable machine learning technique for detecting anomalies and outlier data points in the data space by assessing the local deviation of each data point relative to its neighbors. The algorithm incorporates concepts from other algorithms, including DBSCAN, distance from the center, and reachability distance. A notable advantage of this method is its ability to comprehensively examine outlier data across multiple dimensions, enabling researchers to explore diverse facets of the data space. To evaluate the extent of outlier data in this research, the primary machine learning algorithms were assessed both with and without outlier removal, with the results presented in the subsequent table.

**Table 3. Comparison of the Effect of Outlier Removal on Output Quality**

| After Cleaning Data Outlier | Before Cleaning Data Outlier | |
|---|---|---|
| 0.90687 | 0.90424 | XGBoost |
| 0.90272 | 0.90429 | LightGBM |

From the table, it is evident that XGBoost's performance improves slightly with outlier removal. Data imbalance arises when the volume of data for one group significantly differs from that of other groups. To tackle data imbalance, several methods have been proposed, broadly classified into under-sampling, oversampling, and hybrid approaches. Douzas et al. (2018) proposed a method that first divides minority and majority classes separately using the k-means algorithm. It then performs oversampling by iteratively refining the resulting clusters to rebalance the class distribution and enlarge small clusters to address within-class imbalance via SMOTE. Werner Du Vargas and colleagues (2023) conducted a comprehensive analysis of preprocessing techniques for imbalanced data in machine learning, scrutinizing over 9,927 articles in this domain. Their study revealed that the financial sector accounted for approximately 23% of the data, ranking second in terms of imbalance. Among 55 identified sampling techniques, oversampling accounted for 55%, under-sampling for 27%, and hybrid methods for 17%. Notably, oversampling techniques performed better than other types, and hybrid sampling techniques also showed relatively favorable results.

In this investigation, four methods were deployed to handle imbalance, and their results were assessed alongside four selected algorithms, as shown in the table.

**Table 4. Comparison of the Effect of Data Imbalance Correction Using Different Methods and Its Impact on Algorithms**

| SMOTE Tomek | SMOTE | Near Miss | Tomek Link | Without Data Imbalance Correction | |
|---|---|---|---|---|---|
| 0.8689 | 0.8678 | 0.6448 | 0.8877 | 0.8872 | LightGBM |
| 0.8699 | 0.8720 | 0.6581 | 0.8895 | 0.8893 | XGBoost |
| 0.8470 | 0.8471 | 0.7618 | 0.8460 | 0.8460 | Logistic Regression |
| 0.8515 | 0.8508 | 0.7321 | 0.8641 | 0.8602 | MLP |

## Feature Engineering

A feature represents a quantitative characteristic unique to the observation process. However, not all features are essential for extracting relevant information from datasets. Some features may be extraneous or irrelevant for various machine learning approaches, deep learning, and data science methodologies, potentially leading to misleading results and a decline in model quality. Consequently, selecting a subset of essential features often results in enhanced performance. The research on proper feature selection has a long history, with early evidence dating back to 1924 when Fisher introduced a test for variable selection in regression during a discussion at the Royal Statistical Society. Feature engineering, which involves modifying the feature representation of a predictive modeling problem to suit a learning algorithm better, is crucial for optimizing model performance (Khorana, Smolovitz, & Toraga, 2017).

According to Nilchi et al.'s research, the key shaping features influencing customer behavior are as follows (Nilchi, Moghadam, Nasser Sadrabadi, & Farhadian, 2018):

**Table 5. Fundamental Features Shaping the Credit Behavior of Loan Recipients**

| Variable Name | Sources |
|---|---|
| Economic Sector (Purpose of Getting Credit) | Sohrabi et. Al. (2016), Alborzi et. Al. (2013), Gholamian, Sadat Rasoul & Haji Mohammadi (2012), Mirzaei, Nazarian & Bagheri (2011), Chi & Lee (2017) |
| Number of Installment | Alborzi et. Al. (2013), Khalegi far (2013), Kimiagari et. Al. (2012) |
| Total Period of Repayment | Mirzaei, Nazarian & Bagheri (2011) |
| Type of Contract | Kimiagari et. Al. (2012), Alborzi, Mohammad Pourzarandi & Khan Babaei (2013), Khaleghi Far(1392), Mirzaei, Nazarian & Bagheri (2011) |
| Type Of Collateral | Alborzi, M. Pourzarandi & Khan Babaei (2013). Shourvarzi, Masih Abadi & Ghiasy Shahraki (2012), Zamani & Vadiei (2012), Kimiagari et. Al. (2012) |
| Date of Loan Payment | Alborzi et. Al. (2013) |
| Collateral Amount | Tehrani et. Al. (2009) |
| Production Field | Taghavi et. Al. (2008) |
| Credit Confirming Unit | |
| Purpose of Activity (Factory Establishment, Developing Programs, Working Capital, Repairment, Purchase) | |
| Installment Period(month) | |

To conduct a more detailed examination of the variable space, a correlation matrix is generated to assess the correlation between variables.

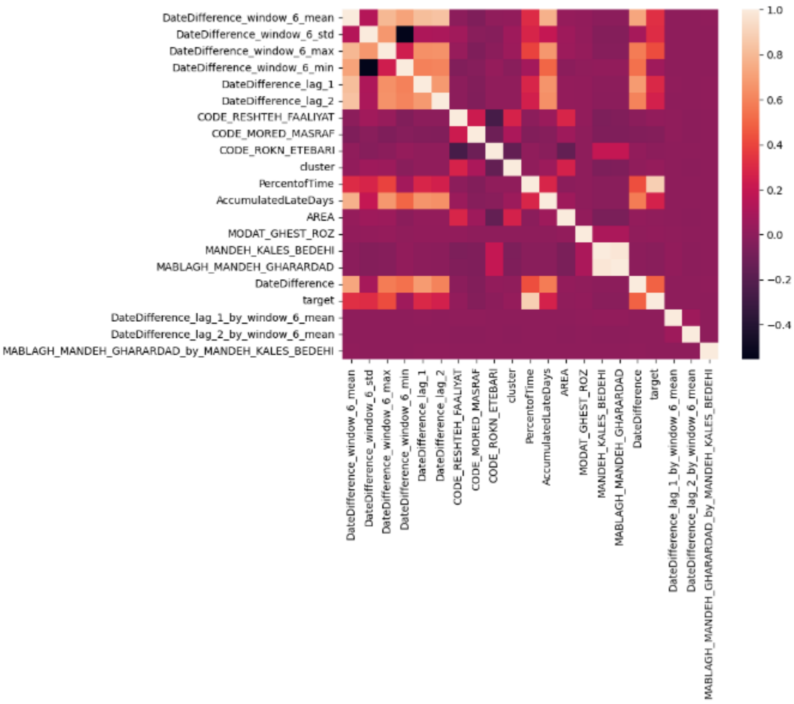


**Figure 6. Correlation Matrix of Research independent variables**

As observed in the correlation table, there is no significant strong correlation among the variables. Considering the similarity between the "Net Debt" and "Contract Remaining Amount" variables, where both refer to the same principal amount, one of them is removed from the variable space. Furthermore, based on expert opinions from Loan specialists, the following variables are added to the variable space:

1. Time to Default (percentage of the time of occurrence relative to the total loan period)

2. Cumulative Delay Days (total number of customer delay days during the repayment period)

The figure below illustrates the density of loan defaults plotted over the repayment period. It is evident from the figure that the highest level of loan defaults occurs within the initial 20% to 40% of the repayment period.

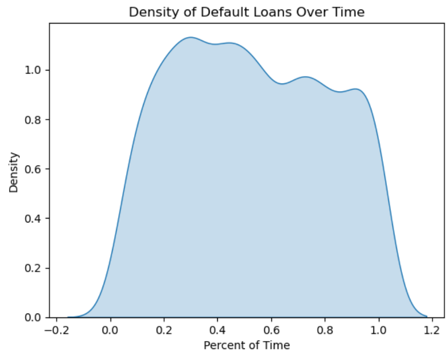


**Figure 7. Density of Loan Defaults in the Loan Receipt Period**

In the subsequent step, meaningful relationships among variables and the dependent variable were identified using a Python feature engineering library. Through this process, the following data points, which exhibit significant relationships with the dependent variable, were obtained:

1. (Mean, Standard Deviation, Minimum, Maximum) 6-Month Window of Delay Duration (Days)

2. Delay Duration in Period (-1) and Period (-2)

3. Period Duration Before / Average 6-Month Delay Duration Window (to calculate the slope of changes)

4. Period Duration Two Periods Before / Average 6-Month Delay Duration Window (to calculate the slope of changes)

5. Contract Balance Amount / Net Debt Balance

To mitigate redundant variables, various methods are available for selecting appropriate features. Rashidul Islam comprehensively classified all feature selection approaches employed in machine learning in a detailed study (Rashidul Islam et al., 2022). The figure below categorizes the primary feature selection methods.

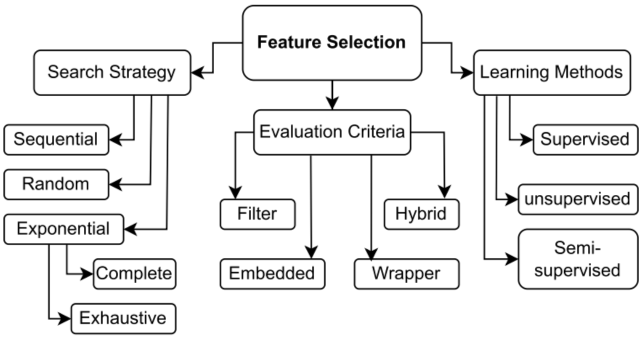


**Figure 8. Feature Space Exploration Methods and Optimal Feature Selection**

In this study, unsupervised learning models were utilized for optimal feature selection. To achieve this objective, the significance of variables concerning the dependent variable was evaluated using a random forest algorithm, and the outcome is illustrated in the following figure:

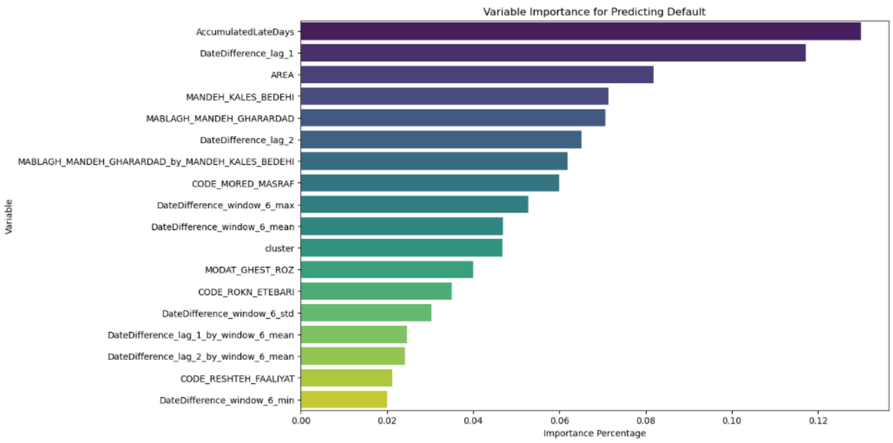


**Figure 9. Importance of variables used for predicting loan defaults using the Random Forest method**

As depicted in the figure, the minimum delay within the 6-month window has the least impact on predicting loan defaults, accounting for less than 2%. To assess the enhancement in results, this variable is eliminated from the variable space. Notably, the graph highlights the segmentation effect (Cluster Attribute) as a significant predictor of loan defaults, thereby affirming Hypothesis 1 of this study. The software used in this research includes a SQL database for organizing extracted data, Power BI for graphical representation of data, and Python for statistical analysis and machine learning, including scikit-learn and hypothesis testing.

The conceptual model of the research is presented in the figure below, delineating the steps of the work:

1. Data cleaning and preparation for use in machine learning algorithms
2. Clustering of Loans based on behavioral variables
3. Removal of redundant variables and addition of required variables based on feature engineering
4. Selection of optimal machine learning classification algorithm
5. Examination and comparison of the performance of the LSTM algorithm against traditional linear algorithms
6. Examination and comparison of the performance of regular LSTM algorithm and Hybrid LSTM with classification algorithm
7. Interpretation of results

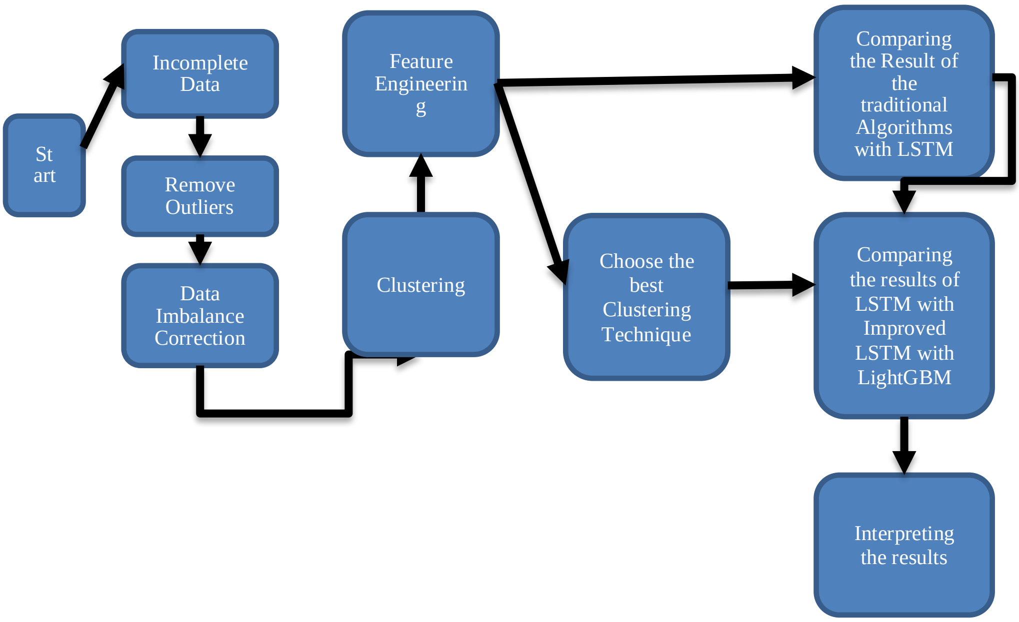


**Figure 10. Conceptual Model for Research Predictions of Loan Default**

## Loans Clustering

To cluster the loans using the K-Means method and the provided data in the table, the number of clusters was determined using the Silhouette score. The list of variables is given in annex A. For loan clustering, it is necessary to quantify all classification variables using a conventional method. In this study, two variables - the activity domain, the type of collateral and the purpose of the activity - were quantified using the well-known Ward method. After analyzing the output, the silhouette score was determined. The table showed that, considering all the variables, the loans should be divided into 60 clusters. However, upon examining the variables in each category, it was determined that this number of clusters is insufficient to distinguish between loan clusters properly.

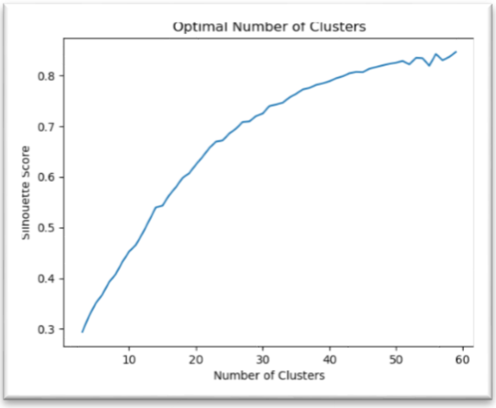


**Figure 11. Finding the Best Clustering Based on Silhouette Index Before Optimization**

By excluding the activity goal item from the clustering, we obtain a silhouette score of 14 for the clusters. Upon examining the information about the variables used for clustering, the distinctions between clusters become discernible. The final silhouette diagram is as follows:

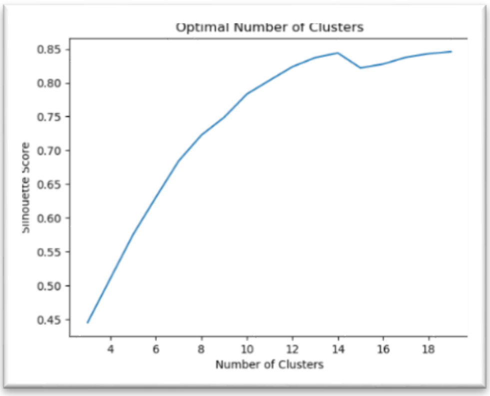


**Figure 12. Final Silhouette Diagram after Removing the Activity Goal Item**

## Selecting the Optimal Machine Learning Classification Algorithm

Upon considering the elimination of outliers, data imputation to address imbalance, and a thorough examination of the efficacy of the generated clustering feature, along with the removal of less significant features, the outputs of all four methods are re-evaluated. The summarized findings are delineated in the subsequent table:

**Table 6. Algorithm Results after Outlier Removal, Imbalanced Data Handling, and Feature Engineering**

| F1 Score | ROC_AUC Score | Algorithm Name |
|---|---|---|
| 0.450101 | 0.904937 | XGBoost |
| 0.399048 | 0.903515 | LightGBM |
| 0.419159 | 0.868384 | Logistic Regression |
| 0.386751 | 0.88605 | Multilayer Perceptron |

As shown in the table, the XGBoost algorithm achieves the highest performance in predicting loan defaults, with a score of 0.904937. Furthermore, the AUC-ROC chart for these algorithms, elucidated below, vividly illustrates the predictive prowess of the XGBoost algorithm in addressing the second research Hypothesis.

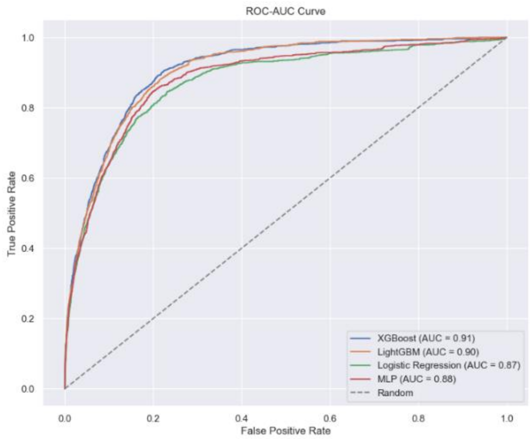


**Figure 13. ROC_AUC Chart for Examined Algorithms**

Upon scrutinizing the pivotal features of the XGBoost algorithm, it becomes evident that each feature is significant in shaping the final trained model.

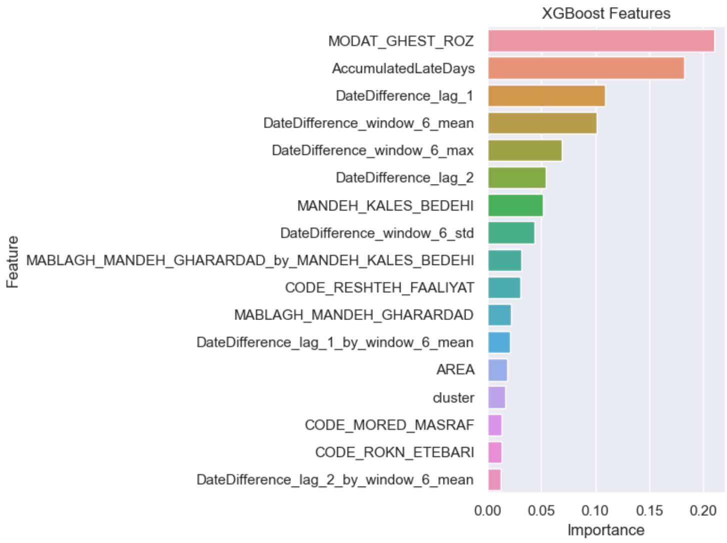


**Figure 14. Ranking Features in Order of Importance for XGBoost**

To enhance output quality, exploring the combination of multiple algorithms, known as ensemble algorithms, can be pursued. In this study, the output is reassessed by integrating the following algorithms:

pred = pred_proba_lgb[:,1]*0.25 + pred_proba_xgb[:,1]*0.7 + pred_proba_logit[:,1]*0.03 + pred_proba_mlp[:,1]*0.02 (1)

The output of the algorithm is compared with the output of XGBoost in the following table:

**Table 7. Comparison of the output of the ensemble algorithm with XGBOOST**

| ROC_AUC Score | Algorithm Name |
|---|---|
| 0.90731 | XGBoost |
| 0.90751 | Ensemble |

In the table, we observe a significant increase in output when using the ensemble method. This observation suggests that the combination approach could serve as a viable alternative to the individual XGBoost algorithm.

Comparison of Short-Term and Long-Term Memory Algorithms with Traditional Algorithms

In our investigation, we examine the behavior of time-series machine learning algorithms compared to traditional linear algorithms. Specifically, we explore the prediction results obtained from two distinct approaches: the Long Short-Term Memory (LSTM) neural network, a type of recurrent neural network, and the linear ARIMA algorithm. To construct the linear ARIMA model, we employ the ARIMA library in Python. This library serves as an optimization tool for ARIMA models, selecting the optimal state-space parameters based on input variables and guided by the Akaike Information Criterion (AIC). Once the optimal values for the p, d, and q variables are determined, predictions are generated. Subsequently, we turn our attention to the LSTM neural network and investigate both short- and long-term memory predictions. To effectively utilize the LSTM network, we reshape the data into a rolling window format, using a 3-month window in our research. Additionally, we select the mean squared error as the regression factor to evaluate the performance of the chosen algorithm based on average absolute prediction error.

The specifications of the LSTM neural network employed in our study are as follows:

**Table 8. Specifications of the Neural Network**

| Amount | Description |
|---|---|
| Adam | optimizer |
| 2 | Verbose |
| Mean Absolute Error (MAE) | Loss |
| 3 | Lookback Window |
| 50 | Epochs |

In our investigation, we meticulously compared the absolute prediction errors of the aforementioned algorithms. Additionally, we examined the output of the Prophet algorithm, another time series approach developed by Facebook. Notably, Prophet demonstrates a remarkable ability to analyse and predict time series data with seasonal behavior .The summarized results are meticulously presented in the following table:

**Table 9. Comparison of Outputs from Short-Term Memory, Long-Term Memory, ARIMA, and Prophet Algorithms**

| Mean Absolute Error (MAE) | Algorithm Name |
|---|---|
| 12.6060 | LSTM |
| 13.4310 | ARIMA |
| 18.3669 | Prophet |

In the table above, we observe that the Long Short-Term Memory (LSTM) algorithm achieves 6% higher predictive power than the linear ARIMA model. Interestingly, despite its development for time-series data with high seasonality, the Prophet algorithm did not perform well at predicting loan defaults. Next, we explore the combination effect of the XGBoost and Long-

Term Memory algorithms. Specifically, we use the XGBoost algorithm's predictive power as an input to the LSTM algorithm. Our goal is to determine whether this combination improves the performance of the LSTM time-series model. The results of this analysis are summarized in the following table:

**Table 10. Comparison of the Outputs of Long Short-Term Memory (LSTM), ARIMA, and LSTM-XGBoost Combination Algorithms**

| Mean Absolute Error (MAE) | Algorithm Name |
|---|---|
| 12.6060 | LSTM |
| 13.4310 | ARIMA |
| 8.0123 | LSTM-XGBOOST |

Our analysis reveals that incorporating XGBoost as an input to the Long Short-Term Memory (LSTM) algorithm results in a substantial reduction in the average absolute prediction error—up to 36%. This improvement represents a significant leap over the ARIMA model, achieving a remarkable 40% improvement. In light of these findings, we address the third and fourth research Hypothesis: deep learning algorithms outperform traditional and linear algorithms. Furthermore, the combination of XGBoost and LSTM shows potential to reduce the average absolute prediction error.significantly

## Model Validation

To validate the predictive accuracy and robustness of our model, we employed a comprehensive test-data validation approach. After preparing and cleaning the data, we divided it into training and test datasets. The model was initially trained on a dataset containing contractual, collateral, and repayment behavior data. Subsequently, we tested the model using the test dataset to assess its predictive performance. We utilized several machines learning algorithms, including XGBoost, LightGBM, and Long Short-Term Memory (LSTM) neural networks. The performance of these models was evaluated using metrics such as Mean Absolute Error (MAE), Receiver Operating Characteristic (ROC) curve, and the Area Under the ROC Curve (AUC-ROC). The XGBoost algorithm demonstrated superior predictive performance, with an AUC-ROC of 0.90751 in Ensemble mode, indicating high accuracy in distinguishing between defaulters and non-defaulters. Also, as mentioned above, incorporating XGBoost as an input to the Long Short-Term Memory (LSTM) algorithm results in a substantial reduction in the average absolute prediction error—up to 36%. This improvement represents a significant leap over the ARIMA model, achieving a remarkable 40% improvement. These results demonstrate the model's robustness and its potential applicability in real-world scenarios. However, areas for improvement were identified, including further hyperparameter tuning and the exploration of additional features to enhance predictive accuracy.

## Discussion

The output from our models provides valuable insights into the variables utilized in this research. In the table below, we present these variables alongside their respective importance rankings. Additionally, we provide explanations for the significance of each variable:

**Table 11. Research Variable Results Interpretation**

| Variable Name | Importance Level | Explanation |
|---|---|---|
| Cumulative Delay Days | 0.13 | Represents the total days a customer has delayed installment payments, serving as a vital predictor for loan default risk. For example, a borrower consistently delaying repayments indicates recurring financial incapacity, increasing the likelihood of default. |
| Late Payment Days in the Previous and 2 Previous Periods | 0.12 | Indicates payment delays in previous installment periods, emphasizing the continuity of delayed repayments over time. This variable highlights a recurring pattern of financial instability, raising the risk of default. |
| Province | 0.09 | Geographical dimension with the feature "Province" representing the customer's location. Indicates that borrowers residing in economically challenged areas may face higher financial challenges, potentially increasing the risk of default. |
| Net Debt Balance | 0.075 | Reflects the net outstanding balance and total loan amount. For instance, a borrower with a significant overdue balance, indicating high debt levels, may be more susceptible to default due to financial pressure from repaying a substantial loan. |
| Maximum and Average Late Payment Days in the Six Months | 0.055 | Indicates payment delays in the two previous periods, providing insights into historical payment trends. A realistic scenario involves a borrower consistently making delayed payments, with extensive historical data indicating an increased risk of default. |
| Net Debt-to-Total Contract Ratio | 0.065 | Provides insight into the proportion of overdue debts. A customer with a higher ratio is more vulnerable, as a significant portion of their loan remains unpaid. |
| Loan Usage Purpose | 0.061<br>0.021 | Indicates the primary area of customer activity in a high-risk industry or credit sector facing economic conditions such as inflation, currency supply, or economic downturn. This potentially increases the risk of default |
| Credit Approver | 0.038 | Indicates the approving entity for the credit Loan file. Depending on the accuracy of the conducted reviews, it |

| | | |
|---|---|---|
| | | can be effective in predicting the risk of Loan default. |
| Clustering | 0.05 | Customer clustering based on contractual features reveals commonalities among Loan borrowers, irrespective of conventional boundaries. The likelihood of default varies across clusters, and borrowers in economically unstable clusters may exhibit distinct credit behavior. |
| Installment Duration (Days) | 0.04 | Represents the repayment period (in days), considering longer repayment periods. Borrowers with more/less time for financial and installment repayment have different behaviors. |
| Maximum, Average, and Standard Deviation of Late Payment Days in the Six-Month Window | 0.055<br>0.025 | Temporal behavioral variables generally emphasize the importance of understanding payment timing, changes, trends, and averages. For instance, a borrower with high instability during the repayment period or an upward trend in late payment delays may be more likely to default. |

In predicting loan defaults within a bank, selecting the "Best" algorithm hinges on the bank's priorities and specific goals. To identify a more suitable algorithm, the following factors should be carefully considered:

Accuracy and Recall Tradeoff

If the bank's primary concern is minimizing false positives (i.e., incorrectly classifying non-defaulters as defaulters), accuracy becomes crucial. Higher accuracy reduces the risk of misclassification but may sacrifice the identification of genuine defaulters (lower recall). Conversely, if the bank aims to pinpoint true defaulters—even at the expense of some false positives—prioritizing recall is essential. Higher recall ensures that a greater proportion of actual defaulters are correctly identified, albeit at an increased false-positive rate.

Receiver Operating Characteristic (ROC) Curve for Overall Fitness

The area under the ROC curve (AUC-ROC) provides insight into the model's ability to differentiate between customers with and without defaults across various probability thresholds. If the goal is to have a model that performs consistently well at different decision thresholds, a higher AUC-ROC is a critical criterion.

Interpretability and Explainability

Considering the above factors, "XGBoost" emerges as a strong candidate for an effective predictive algorithm based on the available informationgiven its , .balanced accuracy and high AUC-ROC However, prudent decision-making requires further analysis, potentially involving additional model evaluation techniques and collaboration with domain experts. Especially in industries where model transparency is paramount, interpretability and explainability should be carefully weighed.

Our study revealed a significant performance advantage when combining the XGBoost algorithm with the Long Short-Term Memory (LSTM) neural network. Specifically, this hybrid approach outperformed both the standalone LSTM network and traditional linear models. As a result, our research hypotheses regarding the higher predictive accuracy of machine learning algorithms and the effectiveness of the XGBoost clustering algorithm in conjunction with LSTM time series models, as compared to linear approaches, have been confirmed.

## Conclusion

In today's dynamic business landscape, characterized by interconnectedness and transformative shifts, predicting customer credit behavior has become both challenging and crucial for banks and financial institutions. Minimizing the risk of loan defaults through accurate credit scoring is paramount. Additionally, post-loan approval, continuous monitoring of customer credit behavior—facilitated by intelligent systems—is essential for precise risk assessment. Over the past few decades, the rise of machine learning models, particularly in predicting financial events, has presented a unique opportunity for financial institutions. These institutions can harness the power of machine learning to detect and predict the behaviors of high-risk customers, thereby replacing traditional methods. In this research, we meticulously explored the necessary steps and tools for data cleaning and preparation in the context of machine learning algorithms. Subsequently, we delved into the behavior of each cluster regarding loan defaults, focusing on loan features.

Furthermore, we conducted a comprehensive evaluation by comparing the outputs of novel machine learning algorithms—such as XGBoost, LightGBM, and LSTM neural networks—with those of traditional models, such as ARIMA. In our study, we introduced a novel hybrid model by combining the XGBoost algorithm with the LSTM neural network. This hybrid approach was meticulously evaluated against both the standalone LSTM network and traditional ARIMA models. The results were compelling: our hybrid model outperformed the ARIMA model by up to 40%. This improvement underscores the potential of leveraging combined machine learning techniques for more accurate predictions.

One of the key advantages of our research is its ability to forecast the behavior of new loan recipients based on their clustering. By analyzing their behavior using our proposed algorithm, financial institutions can make informed decisions. However, we acknowledge challenges in extending machine learning models beyond loan contracts. The limited availability of bank data for research and modeling purposes poses a hurdle. Nevertheless, if loan-related information can be generated and stored beyond internal bank systems, the horizon for expanding machine learning models in this domain widens. In summary, our hybrid model represents a promising step toward enhancing credit risk assessment and shaping the future of predictive analytics in banking.

In this research, clustering models were implemented to classify loans into distinct clusters. The findings indicate that loans have the highest default probability in the initial stages compared to later periods. Clustering is a significant factor in predicting loan defaults. The study demonstrates that by creating intelligent loan categories and analyzing repayment trends, new loans can be monitored and predicted with high accuracy using company behavior and cluster assignments. This can significantly help banks calculate future liquidity risk and credit risk buffers. In the initial phase, outliers were removed using the Local Outlier Factor (LOF) method. It was observed that while removing outliers alone did not significantly improve the performance of machine learning algorithms, combining this step with subsequent phases optimized the model's outputs. This study considered an outlier removal rate of 0.05, indicating that, due to directive behaviors driven by policymakers' strategies and specific conditions such as earthquakes or currency crises, outliers exist in loan systems, and their removal improves the predictive power of machine learning algorithms .In the next phase, data imbalance was addressed using methods such as TOMEK and SMOTE.

Contrary to Douzas (2018), who suggested SMOTE for data imbalance correction, this study utilized TOMEK links, as seen in Werner de Vargas' (2023) analysis of financial models, for imbalance correction. Using feature engineering, several influential variables were created to analyze borrower behavior, demonstrating their effectiveness in predicting loan defaults in the feature table. As noted in the literature review, modern supervised machine learning methods such as XGBoost and LightGBM have higher predictive performance for customer behavior than linear models, consistent with Dong's (2022) findings. Abdoli et al. (2020) also showed that LSTM had higher predictive power than linear models for fluctuating data, a finding this study confirms. Additionally, Siami Namin et al. (2019) found that LSTM outperformed ARIMA by 37% in financial time-series data, a finding also corroborated by this research. Overall, loan defaults, especially with the increasing complexity of macroeconomic conditions and the influence of internal and external variables, cannot be predicted using traditional linear methods. Supervised machine learning models play a crucial role in improving loan monitoring and supervision systems. Furthermore, the performance of LSTM alone and LSTM-XGBoost combined was analyzed, with the latter showing superiority, consistent with Gao's (2021) study on credit card default prediction. Although other methods such as ADABOOST and SVM are also mentioned in the literature, the combination of XGBoost and LSTM has shown positive results, as discussed.

The study highlights that Iranian banks have a significantly higher non-performing loan ratio than Basel standards and global and regional averages. Although external and unsystematic factors, such as sanctions, contribute to this, banks lack systematic programs to control and monitor issued loans. In contrast, countries in the region have integrated financial infrastructures with regulatory bodies, such as the Quick system in the UAE, where all registered companies' financial systems are connected to government oversight systems and monitored using intelligent systems. Key variables such as the borrower's province, loan purpose, industry, and delay duration are critical. It is recommended that banks review the coefficients for these variables in their customer behavior.based credit scoring models-

This research used logistic regression, multilayer perceptron, XGBoost, LightGBM, and LSTM to develop accurate algorithms for loan default prediction. Given the growing trend toward learning AI models, it is suggested to use them in future research due to their continuous nature. The study focused on Murabaha contracts, which can be extended to other loan contracts. Identifying key variables in predicting loan defaults can also be used in smart credit scoring models. Expanding this system to other financial instruments, such as letters of credit, guarantees, and insurance, can provide more precise oversight, preventing future issues such as legal disputes, collateral liquidation, and, importantly, social tensions and damage to bank brands.

Practical and Strategic Suggestions for AI Adoption in Banking

## Practical Suggestions

1. Implementation of Hybrid Machine Learning Models:

Financial institutions should consider adopting hybrid machine learning models, such as the XGBoost-LSTM combination used in this study, to achieve more accurate and reliable predictions of loan defaults, especially given their extensive time-series and well-structured data. These models have demonstrated significant improvements in predictive performance compared to traditional methods. By implementing such advanced models, banks can better identify high-risk customers and take proactive measures to mitigate potential losses. For practical application, banks should invest in the necessary infrastructure to support these computational models, including high-performance computing resources and skilled data scientists to manage and interpret the results.

2. Enhanced Feature Engineering and Data Handling:

To maximize the effectiveness of machine learning models, banks should review their data infrastructure to minimize the ETL and data preprocessing. This involves integrating diverse data sources such as contractual data, collateral information, and customer repayment behavior to build a robust feature set. Additionally, addressing data imbalances through methods such as TOMEK links and removing outliers using techniques such as the Local Outlier Factor (LOF) can significantly enhance model accuracy. Banks should establish standardized procedures for data cleaning and feature engineering to ensure consistency and reliability in predictive analytics.

**Strategic Suggestions**

1. Developing a Data-Driven Culture:

To fully leverage AI solutions, banks need to foster a data-driven culture across all levels of the organization. This involves promoting the importance of data in decision-making processes and encouraging collaboration between IT, data science, and business units. Regular training and workshops can help employees understand the value of data and the role of AI in enhancing banking operations. By embedding data literacy into corporate culture, banks can ensure staff are equipped to support and use AI initiatives effectively.

2. Strategic Data Partnerships and Integration:

Banks should also consider forming strategic partnerships with technology providers and fintech companies to enhance their data capabilities. These partnerships can provide access to advanced data analytics tools, platforms, and expertise that may not be available in-house. Furthermore, integrating external data sources such as social media, market trends, and economic forecasts can enrich the data ecosystem and provide more comprehensive insights for AI models. This holistic approach to data integration can significantly improve the accuracy and relevance of AI-driven predictions and recommendations.

In conclusion, by investing in robust data infrastructure and advancing data maturity, banks can create a solid foundation for deploying AI and machine learning solutions. These steps are essential for unlocking the full potential of AI in credit risk assessment, customer behavior prediction, and other critical banking functions. As the financial landscape continues to evolve, a strategic focus on data excellence will enable banks to remain competitive and innovative.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest concerning the research, authorship and, or publication of this article.

## Funding

The authors received no financial support for the research, authorship and, or publication of this article.

## References


Abdoli, G. M. (2020). Comparing the Prediction Accuracy of LSTM and ARIMA Models for Time-Series with Permanent Fluctuation. *Periódico do Núcleo de Estudos e Pesquisas sobre Gênero e DireitovCentro de Ciências Jurídicas - Universidade Federal da Paraíba*, 9(02), 314-339.

Alborzi, M., Mohammad Pourzandi, M. E., & Khan Babaei, M. (2013). Application of Genetic Algorithm in Decision Tree Optimization for Customer Credit Scoring in Banks. *Journal of Information Technology Management, 23*(1), 23–38.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance, 23*(4), 589–609.

Ameri Sayavashi, S., Kurdlooei, H., & Abdollahi Keyvani, S. M. (2021). Estimating the Probability of Default of Loans Granted to Legal Entities in Non-Banking Deposit Institutions. *Financial Knowledge Securities Analysis Journal,* 11(1), 17–32.

Bhat, M. T. (2020). Scrutinize The Effectiveness of Loan Portfolio Management: Challenges and Remedial. *Studies in Indian Place Names(UGC Care Listed Journal) ISSN:2394-3114,* 40(59) , 303 – 314

Boukherouaa, E., Shabsigh, G., AlAjmi, K., Deodoro, J., Farias, A., Iskender, E., & Ravikumar, R. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance. *International Monetary Fund, 34*(2), 34.

Branzoli, N., & Fringuellotti, F. (2022, February 21). The Effect of Bank Monitoring on Loan Repayment. *FRB of New York Staff Report No. 923*.

Butaru, F. C. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 11*(3), 218-239.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Chi, Q., & Li, W. (2017). Economic Policy Uncertainty, Credit Risks, and Banks' Lending Decisions: Evidence from Chinese Commercial Banks. *China Journal of Accounting Research, 10*(1), 33-50.

Deng, T. (2019). Study of the Prediction of Micro-Loan Default Based on Logit Model. In *International Conference on Economic Management and Model Engineering (ICEMME)* (pp. 260-264). Malacca, Malaysia.

Dong, X. (2022). Loan Default Prediction based on Machine Learning (LightGBM). *BCP Business & Management*, 457–468.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information*

*Sciences*, 1-20.

Emekter, R., Tu, Y., & Jirasakuldech, B. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics, 12*(5), 54–70.

Financial Stability Board (FSB) Annual Financial Report. (2021). Retrieved from https://www.fsb.org/wp-content/uploads/P300821.pdf

FitzPatrick, P. J. (1931). *Symptoms of industrial failures as revealed by an analysis of the financial statements of failed companies, 1920-1929*. Catholic University of America, Pp. ix, 132.

Fitzpatrick, T., & Mues, C. (2016). An Empirical Comparison of Classification Algorithms for Mortgage Default Prediction: Evidence from a Distressed Mortgage Market. *European Journal of Operational Research*, *25*(2), 427–439.

Gao, J., Sun, W., & Sui, X. (2021). Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. *Discrete Dynamics in Nature and Society*, 1–13.

Gholamian, M. R., Sadat Rassoul, M., & Haji Mohammadi, Z. (2012). Association Rule Extraction for Credit Ranking Using Genetic Algorithm: A Case Study of Iranian Bank. In *Third National Industrial Engineering and Systems Conference*.

Jacobson, T., & Roszbach, K. (2003). Bank Lending Policy, Credit Scoring and Value-at-Risk. *Journal of Banking & Finance, 27*(4), 615–633.

Khaleghi Far, H. (2013). Presentation of an Effective Model on Credit Risk Using Logistic Regression in Bank Eghtesad Novin Kish Island Branch. In *First National Conference on Monetary and Banking Management Development*.

Kimyagari, M. A., & Amini, M. J. (2012). Credit Risk Model for Repayment of Bank Credit Loans (Case Study: Shahreza Area of Bank Melli Iran). *Ninth Industrial Engineering Conference*.

Königstorfer, F., & Thalmann, S. (2020). Applications of Artificial Intelligence in Commercial Banks – A Research Agenda for Behavioral Finance. *Journal of Behavioral and Experimental Finance, 39*(4), 55-60.

Lee, S., & Lundberg, S. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

Liu, J., Gao, Y., & Hu, F. (2021). A Fast Network Intrusion Detection System Using Adaptive Synthetic Oversampling and LightGBM. *Journal of Computer and Security, 106*, 102289.

Malali, A. B. (2020). Application of Artificial Intelligence and Its Powered Technologies in the Indian Banking and Financial Industry: An Overview. *IOSR Journal Of Humanities And Social Science, 25*(1), 55–60.

Mileris, R. (2013). Macroeconomic Determinants of Loan Portfolio Credit Risk in Banks. *Economics, 12* (3), 496–504.

Mirzaei, H., Nazarian, R., & Bagheri, R. (2011). Investigating Effective Factors on Credit Risk of Bank Customers. *Quarterly Journal of Economic Trends, 16*(3), 67–98.

Mitchell, T. M. (1997). Machine Learning. *New York: McGraw Hill, 2–5.*

Mohammadian Hajikord, A., Asgharzadeh Zaferani, M., & Emamdoust, M. (2016). Credit Risk Assessment of Legal Customers Using Support Vector Machine Model and Hybrid Model of Genetic Algorithm - A Case Study of Bank Tejarat. *Financial Engineering and Securities Management, 11* (2), 17–32.

Nilchi, M., Moghadam, K., Sadrabadi, A. N., & Farhadian, A. (2018). Prediction of Payable Loans Risk Using Data Mining Tools. *Journal of Monetary and Banking Research, 11* (4), 625–654.

Purkazemi, M., Sadeghatparast, A., & Dehpanah. (2017). Estimating the Probability of Real Customer Default in Banks Using Neural Network Method (Case Study: Pasargad Bank). *Journal of Studies in Banking and Islamic Banking Management, 10* (1), 1–23.

Rashedul Islam, M., Akter Lima, A., Chandra Das, S., Mridha, M., Rahman Prodeep, A., & Watanobe, Y. (2022). A Comprehensive Survey on the Process, Methods, Evaluation, and Challenges of Feature Selection. *IEEE Access, 10* (2), 34–51.

Sabzevari, H., Soleymani, M., & Noorbakhsh, E. (2007, November). A comparison between statistical and data mining methods for credit scoring in case of limited available data. In Proceedings of the 3rd CRC Credit Scoring Conference (pp. 1–5). Citeseer.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development,* 3 (3), 210–229.

Shoorvarzi, M. R., Masih Abadi, A., & Ghiasi Shahraki, M. (2012). Evaluation of Credit Risk of Legal Customers of Export Bank Based on Two Artificial Neural Network and Logistic Models and Their Comparison. In *National Conference on Accounting, Financial Management, and Investment.*

Siami-Namini, S. T. (2019). A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM. https://arxiv.org/abs/1911.09512.

Sohrabi, B., Raeisi Vanani, E., & Zare Mirak Abadi, F. (2016). Designing a Recommender System for Optimization and Management of Banking Loans Based on Clustering and Classification Algorithms. *Innovative Research in Decision Making, 13*(2), 53–76.

Taghavi, M., Lotfi, A. A., & Sohrabi, A. (2008). Credit Risk Model and Ranking of Agricultural Bank Corporate Customers. *Economic Research Journal, 21* (1), 99–

128.

Tehrani, R., Mohammadi, M., & Rahimi, A. M. (2009). Credit Scoring System and its Role in Improving the Financial Supply System. In *First International Conference on the Development of Financial Supply System in Iran, Sharif University of Technology*. https://sid.ir/paper/807684/fa

The World Bank Report. (2022). Retrieved from https://data.worldbank.org/indicator/FB.AST.NPER.ZS?contextual=default&end=2022&locations=MY-US-AE-SA-TR&start=2010&view=chart

Thomas, L. C. (2000). Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers. *International Journal of Forecasting, 16*(2), 149–172.

Werner de Vargas, V., Arthur Schneider Aranda, J., dos Santos Costa, R., Ricardo da Silva Pereira, P., & Luis Victória Barbosa, J. (2023). Imbalanced Data Preprocessing Techniques for Machine Learning: A Systematic Mapping Study. *Knowledge and Information Systems, 40* (3), 34-51.

Ye, A. (2020). Xgboost, lightgbm, and other kaggle competition favorites. Sep, 2020Towards Data Science. *10*(3), 785-794.

Zakavat, S. (2003). Credit Risk Models of Export Development Bank of Iran Customers. Master's Thesis in Banking Management, Banking Institute.

Zamani, S., & Abdipour, S. (2012). Credit Risk Prediction of Banking Customers Using Neural Network Model. In *National Accounting, Financial Management, and Investment Conference*.

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default Prediction in P2P Lending from High-Dimensional Data Based on Machine Learning. *Physica A: Statistical Mechanics and its Applications, 10* (2), 785–794.

Annex A: List of used variables for clustering the Loans.

| Variable Name | Variable Type | Comments |
|---|---|---|
| Loan Amount | Numerical | |
| Interest Rate | Numerical | |
| Installment Period(month) | Numerical | Based on the month Number between each installment |
| Number of installments | Numerical | |
| Collateral Type | Categorical | Promissory Note, Bonds, Bank Deposit, Letter of Guarantee, Stocks, Pledged Assets, Legal Contract |
| Field of Activity | Categorical | Industry and Mine, Service, Agriculture, Real State, Business, etc.. |
| Purpose of Activity | Categorical | Factory Establishment, Developing Programs, Working Purchase Capital, Repairment, |

**Bibliographic information of this paper for citing:**

Ahmadi Azar, Mohammad; Tehrani, Reza & Mirlohi, Seyed Mojtabi (2026). Predicting Corporate Loan Defaults Using Deep Learning Algorithms and a Comparative Analysis with Linear Models: A Case Study of a Major Commercial Bank. *Iranian Journal of Finance*, 10(1), 1-42.