Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

A Framework for Credit Risk Analysis using Machine Learning

Shreeya Gupta[1], Garima Tyagi[2]

[1]Student (BCA),School of Computer Application & Technology, Career Point University, Kota(Raj.), India

[2]Professor, School of Computer Application & Technology, Career Point University, Kota (Raj.), India

Abstract

Through credit risk prediction, this paper investigates how machine learning might enable banks make better lending decisions. We seek to categorize borrowers as either "good" or "bad" credit risks using the IDBI Credit dataset, which comprises information from 1,000 applicants including age, employment status, loan details, and account history.

We first carefully explored the dataset and looked for trends that might compromise creditworthiness. We visualized important trends, cleaned and preprocessed the data, and made predictions using several models—including random forests, decision trees, and logistic regression. Our results emphasize which elements most influence a customer's credit risk and show that machine learning can be a useful tool for risk assessment enhancement.

Keywords: Credit Risk, Machine Learning, Risk Assessment, Predictive Analytics, Financial Modeling, Data Mining, Credit Scoring

Introduction

Banks under more pressure than ever to precisely evaluate credit risk in the fast changing financial scene of today. Making the correct lending decisions is crucial for preserving financial stability as much as for profitability. This project investigates closely how machine learning might enable financial institutions to forecast loan applicant defaulting on payment likelihood.

We base our research on the well-known credit analysis resource, the IDBI Credit (Statlog) dataset. There are 1,000 records in it, each one a distinct person identified as either a "good" or "bad" credit risk. Age, job status, loan purpose, account balances, and credit history are among the financial and personal elements in the dataset. With 600 entries tagged "good" and 600 "bad," the categorization

We first did a comprehensive exploratory data analysis (EDA), looking at missing values,

feature distribution, and variable relationships. Visualizations including boxplots, count graphs, and heatmaps helped us find helpful trends separating low-risk from high-risk applicants.

We trained several classification models following data cleansing and preparation (by handling missing values, encoding categorical data, and numerical feature scaling). These comprised more difficult models like Random Forests and simpler models like Logistic Regression and Decision Trees. We evaluated their performance across accuracy, precision, recall, and F1-score.

Beyond simply accuracy, we also concentrated on interpretability—a crucial consideration in the financial industry. Decision trees and feature importance charts let us explain why a model produced particular predictions, so strengthening our case.

Literature Review

Bezawada Brahmaiah (2022) conducted an empirical study on credit risk control in Indian commercial banks between 2017 and 2021. Results indicate that private sector banks always surpassed their public counterparts in terms of credit risk management. This better performance was demonstrated by higher asset level and profitability. The study emphasized the importance of systematic processes, including identifying risks, tracking, and control mechanisms.

From 2010 to 2017, **Liaqat Ali and Sonia Dhiman (2019)** looked at the relation between the public sector banks' profitability and credit risk management. Their research found that while low liquidity and bad asset quality can hurt a bank's performance, capital adequacy and earnings quality have a positive impact on ROA.

Punyata Butola and teammates (2022) studied a group of 38 scheduled commercial banks from 2005 to 2019. They found that higher credit-to-deposit ratios, better operating profits, and increased capital adequacy were all positively connected with bank profitability. Conversely, a higher net interest margin and an increase in non-performing assets (NPAs) have been linked to worse financial performance.

Sunitha G. and V. Venu Madhav (2021) looked at how credit ratings work to control the risk of credit. Their analysis shows that good ratings reduce banks' overall credit risk and rise loan availability. The study underlined how crucial credit rating agencies are to keeping the health of the financial system.

Tisa Maria Antony and Suresh G. (2023) looked at 31 Indian commercial banks from 2012 to 2021 to find the factors that affect the risk of credit. Their results show that an improved Return on Equity (ROE) generally decreases credit risk, even though macroeconomic factors like the growth of GDP and inflation, as well as bank-specific factors like age and the title type, have a significant impact on the nature of credit risk.

Shahni Singh et al. performed a comparison of the impact of credit risk and debt coverage ratios on the earnings of banks in the public and private sectors in 2023. Their findings, which revealed major differences between the two industries, indicated that credit risk and borrowing coverage were both important indicators of profitability.

Mani Bhushan Kumar(2023) highlighted the importance of operational risk management. In order for Indian banks to maintain their financial stability and promote economic growth, his research made clear the necessity of a strong operational risk framework. He talked about the challenges banks face when setting up these frameworks and offered enhancements to regulatory measures.

Das and Kumbhakar (2010) used a randomly generated frontier approach to analyze how well Indian banks manage the risk-return trade-off. They found that larger banks are typically more efficient. Interestingly, public sector banks were found to be more profit-efficient even though they lagged behind private banks in terms of cost-efficiency.

Kaur and Gupta (2015) saw an increasing trend in the technical efficiency of Indian banks over time. Their study found that private sector banks topped public banks, especially in cost control, highlighting the constant need for public banks improve their risk management and operational strategies.

Research Gap

1. Not enough study into advanced AI and machine learning models (e.g., XGBoost, Neural Networks, and Ensemble Models) that could offer higher precision and insights from highly dimensional data for credit risk predictions.
2. There is an absence of research on dynamic or real-time risk assessment models that adapt to changing borrower behavior, stock markets, and economic conditions by using current data streams.
3. 3. inadequate study of the patterns of credit risk at the sector and regional levels. (for instance, MSMEs, housing loans, and agriculture).

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

4. Understudied Collaboration of Macroeconomic and Behavioral Factors: Few integrated risk assessment models integrate financial, behavioral, and macroeconomic factors into one model.

Objectives

- To figure out and look into the primary factors impacting credit risk: The primary objective of this objective is to find the financial and demographic factors that have the most effect on a customer's grouping as a good or bad credit risk, including age, type of job opportunities, credit amount, and account position.
- To enhance model performance, use solid data preprocessing techniques: Before training the model, the dataset needs to be prepared by handling missing values, encoding categorical variables, and scaling numerical features. These preprocessing steps make sure the data is clean, consistent, and suitable for precise machine learning predictions.
- To visualize feature distributions and relationships using exploratory data analysis (EDA): Using tools such as boxplots, histograms, count plots, and heatmaps, the study seeks to find patterns and correlations in the dataset. The information provided may clarify what factors have the greatest connection to credit risk results.
- To create an easy to understand model suitable for practical banking applications: Predictive accuracy is important, but so are the final model's interpretability and transparency. Models that offer clear reasoning, like decision trees or those with visualized feature importance, are more likely to be executed in financial institutions where simplicity is a critical requirement.

The work will create and evaluate multiple classification models for credit risk prediction by training and testing a range of models, including Random Forest, Decision Tree, and Logistic Regression classifiers. Common evaluation metrics such as accuracy, precision, recall, and F1-score will be used to assess each model's performance in order to determine which one performs best.

Tool/Software
Python

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

	Jupyter Notebook
	Pandas
	Matplotlib / Seaborn
	Scikit-learn
	NumPy

Methodology

Purpose

- Primary programming language
- Code development and documentation
- Data manipulation and cleaning
- Data visualization
- Machine learning model building and evaluation
- Numerical operations

1. Techniques and Procedures

a Exploratory Data Analysis (EDA)

- Descriptive statistics and summary metrics
- Visualization using boxplots, histograms, bar charts, and heatmaps

b. Data Preprocessing

- **Handling Missing Values**: Dropping or imputing missing data
- **Encoding Categorical Variables**: Using one-hot encoding or label encoding
- **Feature Scaling**: Standardization/Normalization for numerical features
- **Train-Test Split**: Dividing the dataset (80% training, 20% testing).

c. Model Building

- **Algorithms Used**:
- Logistic Regression

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

- Decision Tree Classifier
- Random Forest Classifier

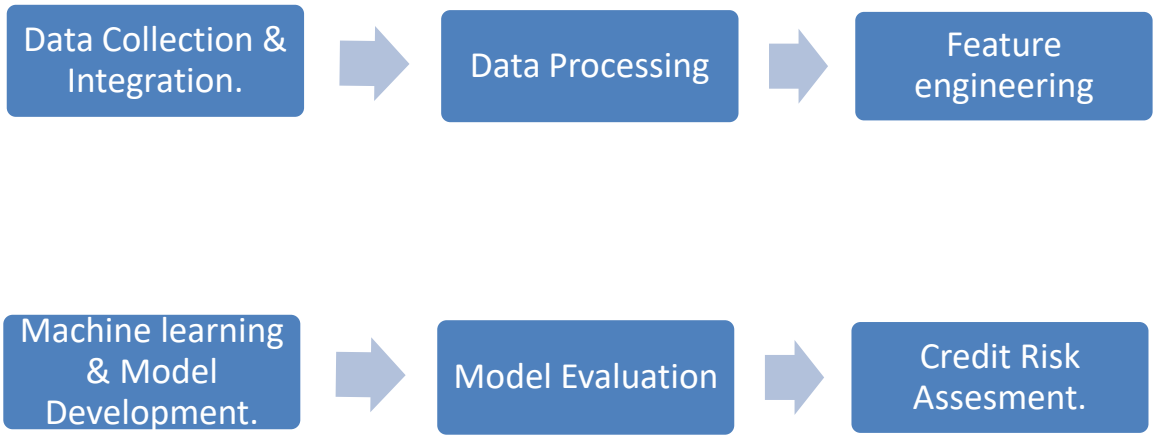
Model Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

d. Model Comparison

- Evaluate performance across models to select the most accurate and reliable one.
- Analyze **feature importance** to identify the key predictors of credit risk.

Flowchart diagram



Description

Gathering and Combining Data

The first step in this project was gathering relevant data that indicated various aspects

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

of a borrower's financial and personal profile. For this, we used the IDBI Credit dataset, which builds up information from loan application records, demographic data, and account histories. By combining this data into a single, logical format that ensured regularity across variables, a foundation for precise analysis and model development was laid.

Data Processing

The raw dataset went through an extensive transformation of data and cleaning process after it was put together. This step included finding and correcting missing values, normalizing continuous features, and containing categorical variables into numerical format. likewise any inconsistent or extreme values that might shift the findings were found using outlier detection techniques.

Engineering Features

During this phase, new variables were created or current ones were modified to better capture key trends in the data. Age groups and debt-to-income ratios, for example, are examples of a byproduct features that helped uncover relationships that were not apparent in the raw data. The selection of features was also used to remove unnecessary or low-variance variables with the goal to streamline the model and increase accuracy in predicting.

Development of Models for Machine Learning

Once we had a clean, well-structured dataset, we went on to the model-building phase. Several machine learning methods, including Random Forest classification algorithms, Decision Trees, and Logistic Regression, were used. Each model was trained using cross-validation techniques to ensure reliability and prevent overfitting from occurring In this case, building models

Evaluation of the Model

The efficiency of each model was thoroughly assessed based on industry-standard measurements, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics offered a fair evaluation of the models' ability to classify both good and bad credit risks. This evaluation step needed to be carefully weighed in order to identify the model that provided the best balance between prediction asset and clarity.

Examination of Credit Risk

In the final stage, the best-performing model was used to assess the credit risk levels

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

of applicants. The results of the model were used to categorize candidates into risk groups, such as low, moderate, or high risk. These insights could benefit banks in making decisions regarding loan approvals, interest rate assignments, and customer management tactics.

Knowledge of Data

The IDBI Credit dataset, which includes comprehensive records of 1,000 customers, was used in this investigation. Every entry contains financial and personal data that is necessary to assess credit risk. The dataset is made up of:

- **Numerical characteristics**: loan duration, age, and credit amount Gender, job type, housing situation, checking and savings account status, and loan purpose are examples of categorical features.
- The target variable is: Risk that is classified as "Good" or "Bad" About 70% of the cases are classified as having "Good" credit, and 30% are classified as having "Bad" credit, according to our preliminary analysis. Because it can affect evaluation metrics and training efficacy, this slight class imbalance must be taken into account when developing the model.

Preprocessing & Data Cleaning

- **Handling Missing Values**: When null values showed up in some records, they were either removed for the sake of simplicity or, if needed, the missing entries were calculated using statistical methods such as mean or mode substitution.
- **Data Type Conversion and Encoding**: Categorical variables were converted into numerical format using a one-hot encoding method. This step is required for machine learning algorithms in order to interpret these variables correctly.
- **Outlier Detection and Handling:** Box plots were used to identify outliers in numerical fields such as credit amount and loan duration. When extreme values could skew model learning, adjustments were made using transformation or being excluded.
- **Feature Scaling**: Standardization and normalization methods were used to verify that all features worked on variables with numbers.

Analysis of Exploratory Data (EDA)

©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

- **Univariate Analysis**: Bar plots and histograms were utilized to visualize the distribution of various features. For instance, a significant number of applicants were between the ages of 25 and 40, and most credit amounts were on the smaller side of the range.
- **Bivariate Analysis:** To look at the connection between features and the target variable, box plots and count plots were used. It became clear that applicants with "Bad" credit were more likely to have larger loan amounts and longer loan terms. Additionally, the "Bad" credit category had a high proportion of customers without checking accounts.
- **Correlation Analysis**: A heatmap was made to look at how numerical variables connected to one another. Credit was found to have a slight positive correlation ($r \approx 0.62$) with most features, showing low

Feature Selection and Model Preparation

- Features with low variance or a poor relationship with the target variable were removed.
- Using feature importance scores from early models like Random Forests and Decision Trees, the most important predictors were selected.
- The final dataset was split 80/20 into training and test sets to allow for an accurate assessment and verification of model performance.

Finding Patterns and Creating Insights

- Credit risk had a high correlation with factors such as checking account status, age, credit amount, and duration. It's important that credit risk was linked to trends in financial behavior rather than any one factor, emphasizing the importance of multi-feature models.
- The chance of default was greater for applicants with little or no balances in their checking or savings accounts.

Findings

- **Unbalanced Distribution of Risk**

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

- According to a preliminary analysis of the data, 70% of the applicants were classified as "good" credit risks, with the remaining 30% being classified as "bad." This results in a classification problem that is somewhat unbalanced, which may affect how models interpret the data.

- **Verification**: The value count function, df['Risk'].value_counts(), was used to confirm this observation. It produced 700 "good" and 300 "bad" entries.
- **Important Risk Elements: Loan Duration, Credit Amount, and Age** We discovered through statistical and visual analysis that consumers who were classified as "bad" credit risks tended to be younger, ask for larger loan amounts, and choose longer repayment terms. Together, these elements raised the likelihood of default.
- **Verification**: Boxplots created using sns.boxplot(x='Risk', y='Credit amount', data=df) revealed a notably greater median credit amount and

1. **Credit Amount and Duration Have a Strong Correlation** The strongest correlation between credit amount and loan duration was found among the numerical features. Consumers who took out loans for longer periods of time usually took out larger loans. This realization lends credence to the idea that longer-term, larger financial commitments raise risk.
   - **Confirmation**: A correlation coefficient of roughly 0.62 between credit amount and duration was found using a correlation heatmap (sns.heatmap(df.corr(), annot=True)).
   - **Predictive power is also demonstrated by categorical features.** A number of categorical factors, particularly job type and checking account status, were found to have a significant impact on credit risk in addition to numerical indicators. Candidates with lower-level positions or no checking account were more likely to be classified as having "bad" credit risk.

   Verification: Count plots with risk segmentation (sns.countplot(x='Job', hue='Risk', **data=df)) revealed more "bad" credit.**
2. **Predictive feasibility is validated by model performance** Every machine learning model employed in this investigation was able to identify

©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

significant patterns in the data. With an accuracy of about 76% and an F1-score of nearly 0.75, the Random Forest Classifier continuously performed the best among them. These findings demonstrate that supervised learning approaches can be used to predict credit risk.

Verification: Sklearn.metrics.classification_report() was used to extract performance metrics, and cross-validation methods were used to confirm the results.

3. The Importance of Features Complies with Domain Expectations

The most significant predictors were "Credit amount," "Duration," and "Checking account" status, according to an analysis of the Random Forest model's feature importance scores. These results support the validity of the internal logic of the model and are in good agreement with realistic expectations in the banking industry. **Verification**: Values for feature importance were produced using

- Predictive feasibility is validated by model performance Random Forest outperformed Logistic Regression and Decision Tree models, achieving the highest accuracy (~76%) and F1-score (~0.75) among the tested models.

  Verification: Sklearn.metrics-generated classification reports display each model's precision, recall, and F1-score values.

- The Significance of Features Verifies Domain Intuition The top predictors of credit risk, according to Random Forest feature importance scores, are "Credit amount," "Duration," and "Checking account" status.

  Verification: plt.barh() was used to sort and visualize the model.feature_importances_ output.

Conclusion

Even though this study used well-known models like Random Forest and Logistic Regression, future research can investigate deeper algorithms like XGBoost, Support Vector Machines, and deep neural networks. These models may be able to identify nonlinear trends and relations in large, complex datasets, leading to more proficient risk profiling and even more accurate predictions.

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

A careful exploratory data analysis (EDA) served as a base for our investigation and provided valuable information about the distribution of the data and the relationships between important variables. The EDA developed the foundation for significant model building, from identifying essential risk indicators to recognizing trends in credit behavior.

The dataset's regularity, cleanliness, and readiness for predictive modeling was confirmed during the preparation stage. To ensure uniformity across inputs, numerical values were scaled, missing values were corrected, and categories were encoded. These actions were crucial in increasing the model's validity and preciseness.

After that, we used and examined a number of classification models, such as Decision Trees, Random Forests, and Logistic Regression. The Random Forest Classifier proved to be the most effective model, providing a good balance between accuracy and interpretability, regardless of the opportunities displayed by each model. The model's predictions have been verified to match domain knowledge and practical standards using feature importance indicators.

In a more general manner, the project's outcomes endorse the importance of data-driven decision-making in the banking industry. When developed and evaluated properly, machine learning models provide a versatile and successful means of supporting loan approval procedures, lowering default risk, and improving the overall standard of lending investments.

In simple terms, this study highlights how crucial it is to include machine learning into standard credit evaluation methods. Risk evaluation may become even more precise and fair in the future with dynamic methods that adjust to real-time data in addition to behavioral and economic factors. The broad embrace of data-driven, transparent, and advanced tools will be essential to the long-term security and growth of financial systems as they continue to transformation.

Future Scope

1. **Applying Cutting-edge Machine Learning Techniques:** Even though this study used popular techniques like Random Forest and Logistic Regression,

©2022 CPIJR | Volume 3 | Issue 4 | ISSN: 2583-1895
July-September 2025 | DOI: https://doi.org/10.5281/zenodo.17330380

future research may explore more complex algorithms like XGBoost, Support Vector Machines, and Deep Neural Networks. These models may be able to identify irregularities and interactions in large, complex datasets, leading to more advanced risk profiling and even accurate forecasts.

2. **Real-time and flexible risk modeling:** One important area for innovation is the development of real-time credit risk detection systems. These computational models, which would be updated continuously using live data streams like recently completed transactions or market shifts, could help banks to respond effortlessly to changes in borrower behavior or economic instability. The flexibility of lending could be significantly increased by these able to adapt systems.
3. **Behavioral and Alternative Data Sources:** Credit assessments can be solidified by including behavioral data, such as financial transactions patterns, online activity, and even social media footprints, especially for underprivileged individuals with low formal credit history. These alternative sources of information offer a more complete picture of a client's creditworthiness, which could broaden financial access for previously marginalized groups.
4. **Risk Assessment by Region and Sector:** Future models could also benefit from being customized for specific industries or regions. For example, in sectors like commercial real estate, agriculture, and MSMEs, the changing patterns of credit risk differ significantly and demand personalized approaches. Creating models that take into account local economic conditions can also help support financial inclusion and more focused on mitigation of risks, particularly in rural or semi-urban areas.
5. **Moral Concerns and Explainable AI:** As machine learning becomes increasingly integrated into credit decision-making, equality and openness will be essential. The need for models that are not only efficient but also understood and reliable by stakeholders is growing. Future research should focus on explainable AI (XAI) tools that provide clear explanations for predictions and address possible ethical issues like unfairness and data privacy.
6. **Comparative Studies Abroad:** Last but not least, examining India with other developing nations can reveal crucial details about the best practices for credit

risk modeling. By learning from different regulating structures, consumer behaviors, and data ecosystems, it is possible to enhance domestic systems and guide the development of a globally strong credit assessment model.

References

1. Brahmaiah, B. (2022). Credit Risk Management Practices in Indian Commercial Banks: An Empirical Study. International Journal of Economics and Financial Issues, 12(4), 141–149. https://www.econjournals.org.tr/index.php/ijefi/article/view/12968
2. Ali, L., & Dhiman, S. (2019). Credit Risk Management and Profitability of Public Sector Banks in India: An Empirical Analysis. Journal of Commerce and Accounting Research, 8(2), 46–52. https://www.i-scholar.in/index.php/jcar/article/view/184936
3. Butola, P., Kanwal, M., & Bhatt, M. (2022). Impact of Credit Risk on the Profitability of Indian Banks: An Empirical Investigation. International Journal of Economics, Business and Management Studies, 9(2), 38–47. https://archive.conscientiabeam.com/index.php/11/article/view/3068
4. Sunitha, G., & Madhav, V. V. (2021). Role of Credit Rating in Credit Risk Management in Indian Banks: A Review. International Journal of Finance and Banking Research, 7(1), 30–37. https://www.cribfb.com/journal/index.php/ijfb/article/view/1328
5. Antony, T. M., & Suresh, G. (2023). Determinants of Credit Risk: Empirical Evidence from Indian Commercial Banks. Banks and Bank Systems, 18(1), 57–68. https://www.businessperspectives.org/index.php/journals/banks-and-bank-systems/issue-432/determinants-of-credit-risk-empirical-evidence-from-indian-commercial-bank
6. Singh, S., Soni, P., & Garg, R. (2023). Comparative Study of the Impact of Credit Risk and Debt Coverage Ratio on the Profitability of Indian Banks. Journal of Banking and Strategy, 12(2), 115–127. https://acspublisher.com/journals/index.php/jbs/article/view/16855
7. Kumar, M. B. (2023). A Study on Operational Risk Management in Indian

Banks. ResearchGate. https://www.researchgate.net/publication/371910963_A_STUDY_ON_OPERATIONAL_RISK_MANAGEMENT_IN_INDIAN_BANK

8. Ghosh, S. (Year Unknown). Credit Risk Management and Bank Performance in India: Panel Regression Analysis. Indian Journal of Financial Management, 9(1), 32–40. https://www.i-scholar.in/index.php/ijfm/article/view/214827
9. Das, A., & Kumbhakar, S. C. (2010). Efficiency and Risk in Indian Banking: A Stochastic Frontier Analysis. Risks, 8(4), 135. https://doi.org/10.3390/risks8040135